ENGELBERG CENTER
on Innovation Law & Policy
NYU School of Law

Data Portability and Platform Competition

Is User Data Exported From Facebook Actually Useful to Competitors?

Gabriel Nicholas and Michael Weinberg

November 2019

Executive Summary

In the Wild West internet of the 1990s and 2000s, only the scrappiest, most innovative companies could survive. Today, some of those that survived and thrived have grown into platforms used by billions, difficult to avoid and hard to leave. Regulators, policymakers, and the public at large worry that these large platforms may now have the ability to freeze out competitors and stifle innovation.

Data portability is often suggested as a tool to counteract the power of large platforms. In its simplest form, data portability is a user's ability to download her data from a platform in a format that allows her to use it somewhere else. At least in theory, this lets users bring their data to new services outside the control of the original platform and helps competitors jump-start new products. A robust data portability system might allow regulators to contain the power of large platforms without having to take the drastic step of breaking them up.

This theory is especially attractive in the context of services that rely on network effects, such as social networks. Users have years of conversations, shared photos, and connections with others on existing platforms. Being forced to leave that information behind would create a significant disincentive to jump to a competing platform, no matter how much better it is. Data portability allows users to bring their history somewhere new, even if they leave or delete their data from another platform.

The Key Question

However compelling in theory, few have investigated whether competitors can actually use ported data to create or grow competing platforms. This gap is particularly troublesome because we found no competitive products built on ported data, despite the fact that many large platforms have enabled users to export their own data for years. For example, Facebook has allowed users to download their data since 2010—well before current competition concerns emerged, and long enough ago for a competitor built on ported Facebook data to emerge. Still, no such competitor has emerged.

If data portability can fuel competition, and data portability has been available for almost a decade, where are the competitors built on data portability? And what does this absence mean for regulators considering using data portability as a competitive measure?

To understand the role that data portability can play in creating new, innovative services, we put real data in the hands of real competitors to see what they could do. We are starting with Facebook because we believe that the data related to social networks present some of the biggest challenges to a portability-based approach. We also focused our investigation on one-off data exports, as opposed to continuous integration via API, because of concerns related to the sustained availability of continuous integration. We hope to be able to examine other types of platforms in the data portability context in the future.

For this project, we exported and anonymized user data from Facebook's Download Your Information tool and brought it to individuals in the New York City tech community. We asked a range of people, from

junior engineers to serial C-level executives, how they would use this data to build new products to compete with Facebook. We asked them about the data's strengths and weaknesses, and how it might be improved to make it more useful for potential competitors. We asked them why they were not already using this data to build their services, and what kinds of changes might allow them to do so.

This exploration is important because data portability is such an attractive tool in the regulatory toolbox. If data portability really can allow new services to grow and coexist with today's large platforms, regulators, the public, and the platforms themselves could potentially avoid the dramatic process of breaking the platforms into smaller entities. However, if data portability is not a viable path for competition and innovation, debating the details of data portability schemes could serve as a distraction from other, more effective means of addressing concerns with large platforms.

What We Learned About the Data

In our discussions, interviewees struggled to come up with new, competitive products they could build from, or meaningfully grow with, ported Facebook data. This suggests that regulators should not assume that competitors will be able to use ported data to build innovative products and services. An over-reliance on data portability may distract from more effective tools for addressing concerns with large platforms. We came to this conclusion based on some key limitations our interviewees ran into about how they could use ported Facebook data create new products:

You cannot replicate Facebook with exported Facebook data. Facebook allows a user to export all of the data she explicitly shared with Facebook. That includes photos she uploaded, events she attended, and comments in group discussions she made. However, Facebook does not allow users to export the context that data was shared into. For example, while a user can download the posts she made in a group discussion, she cannot access the data required to reconstruct the full conversation or even the identity of other participants. She also cannot access the inferences Facebook has drawn from her data to build and improve its own service. As a result, trying to use exported user data to reproduce Facebook would be like trying to use furniture to reproduce the office building it came from.

Facebook data is best suited for building Facebook. The data Facebook collects is useful to a service like Facebook. That means it is best suited to build another social network that monetizes insights from user data, and ill suited to building a radically different service. The products that could come out of ported Facebook data will probably bear a striking resemblance to Facebook or one of its features, and are less likely to address a new need or be truly innovative.

Even if it were possible to build a competitor similar to Facebook, it may not be desirable. Ported data is simultaneously insufficient to replicate Facebook and too tailored to Facebook to be useful for much else. Even if neither of these observations were true, there may be reason for concern about the kind of innovation Facebook data might encourage. To the extent that exported data might be useful for building a new platform, that platform is mostly likely to be based on invasive, highly targeted advertising. Regulators and consumers are increasingly scrutinizing this type of business model. It seems unlikely

that a new surveillance-based advertising network would be welcomed by those expressing increasing concern about Facebook itself.

What This Suggests for Policymakers

None of this means that data portability should be abandoned as a regulatory tool in every case. It does, however, suggest that looking to data portability as the primary way to address competition concerns related to large social networking platforms would be a mistake, and it raises concerns about its use in the context of large platforms more generally. When considering implementing data portability regulations, policymakers should weigh these important factors:

Privacy and competition concerns are in tension when it comes to social network data. Social networks connect large numbers of users. When one of those users decides to export her data, a platform must define the frontiers of where her data ends and another user's begins. That decision can be heavily influenced by the priorities articulated by regulators. A data portability program designed to maximize competition would allow users to export data that includes entire comment threads (not merely the user's contribution), the identities of their friends, and data uploaded by others that relates to the exporting user (for example, a photo of the exporting user's face, taken by someone else). This would make it easier for the exporting user to replicate her experience and reconstruct her social network on a new platform.

Conversely, a data portability program designed to maximize user privacy would strictly limit the types of third-party data that she could export. Her friends did not necessarily consent to the data export, so she could not export their names, their photos, or even their sides of a private conversation. Such a regime would be much more respectful of the privacy of non-exporting users. It would also make the data much less useful for competitors.

While there are ways to balance these competing design priorities, in the context of social networks they appear to be fundamentally in tension. Policymakers need to understand which priority they are elevating, and the consequences of that decision.

Data portability can be useful in select contexts. There may be domains entirely disconnected from social networking, such as music streaming or fitness tracking, where a well-designed portability regime could encourage competition. Data portability can also facilitate the concept of data ownership—a value that may have importance independent of competitive concerns.

Data portability may be a distraction in the competition debate. Data portability has been the subject of intense focus by both tech companies and policymakers. However, it may be that the type of data portability that is the focus of those discussions—and of this paper—is simply a poor mechanism to increase competition online. If that is the case, time spent debating specific aspects of a given data portability regime may be better spent considering different types of approaches to competition concerns.

Introduction[1]

In 2011, Google's then-CEO Eric Schmidt told Congress that on the internet, "competition is just a click away."[2] Eight years later, five of the six most valuable publicly traded companies would be American tech companies (Microsoft, Amazon, Apple, Alphabet, and Facebook).[3] Four of them would be the active subject of federal antitrust scrutiny.[4]

The issue of competition in the tech sector is having its day in the sun, but there is little consensus about how best to address it. A handful of Democratic presidential candidates want to break up the big tech companies. The European Commission wants to prevent them from abusing their market dominance.[5] Fox News host Tucker Carlson wants to regulate Google as a utility, and some academics agree with him.[6]

One widely discussed approach is to improve tech competition through data portability. Data portability is the principle that users should be able to take their data from one service and move it to another.[7] The theory underpinning this link is that ported data can form the raw material for the creation of a new, competitive service.[8]

Data portability has been lauded by the public and private sectors alike. Senators Richard Blumenthal (D-CT), Josh Hawley (R-MO), and Mark Warner (D-VA) introduced legislation requiring platforms with more than 100,000,000 active monthly users in the United States to make their data available to competing platforms.[9] Congressman David Cicilline (D-RI) of the House Judiciary Committee's Antitrust Subcommittee has said that pro-competitive tech policies should start by "taking on walled gardens that

---

[1] The authors wish to thank Scott Hemphill, Delon Lier, Kevin Qiao, Steve Weber, the Information Law Institute at NYU Law's Privacy Research Group, and everyone who participated in and helped pull together the interview groups.

[2] *The Power of Google: Serving Customers or Threatening Competition? Before the Subcomm. on Antitrust, Competition Policy and Consumer Rights*, 112th Cong 1 (2011) (statement of Eric Schmidt, Executive Chairman, Google. Inc), https://www.judiciary.senate.gov/imo/media/doc/11-9-21SchmidtTestimony.pdf

[3] As of September 30, 2019, the largest publicly traded companies by market capitalization were Microsoft ($1.062 trillion), Apple, Inc. ($1.012 trillion), Amazon ($0.859 trillion), Alphabet, Inc. ($0.838 trillion), Berkshire Hathaway ($0.509 trillion), and Facebook ($0.508 trillion).

[4] The fifth, Microsoft, was involved in antitrust disputes with the United States government for much of the 1990s.

[5] *See, e.g.* Natasha Lomas, *Google tweaks search ads after EU shopping antitrust ruling*, TechCrunch (Sept. 29, 2017), https://techcrunch.com/2017/09/28/google-tweaks-search-ads-after-eu-shopping-antitrust-ruling/

[6] David McCabe, *Why regulating Google and Facebook like utilities is a long shot*, Axios (Sept. 22, 2017), https://www.axios.com/why-regulating-google-and-facebook-like-utilities-is-a-long-shot-1513305664-9a388f01-f71a-4b45-8844-fec8b74d95d6.html

[7] *See, e.g.* Erin Egan, *Data Portability and Privacy*, Facebook Newsroom (Sept. 2019), https://fbnewsroomus.files.wordpress.com/2019/09/data-portability-privacy-white-paper.pdf [hereinafter "Facebook Portability White Paper"].

[8] *Id.*

[9] Augmenting Compatibility and Competition by Enabling Service Switching Act of 2019, S.2658, 116th Cong. (2019).

block startups and other competitors from entering the market through high switching costs."[10] Mark Zuckerberg mentioned data portability multiple times when he testified before the Senate.[11] It was also one of his "Four Ideas to Regulate the Internet." There, he noted how portability "gives people choice and enables developers to innovate and compete."[12] Europe's GDPR (General Data Protection Regulation) also places data portability front and center.[13]

The promise of data portability appears to be undermined by the fact that major platforms have facilitated data portability for years. Facebook introduced its Download Your Information tool in 2010,[14] and Google's project Takeout launched one year later.[15] If data portability is a key to increasing competition, and data portability has been available for years, how is it that we find ourselves in the current competitive situation?

The purpose of this paper is to examine the extent to which data portability actually allows competitors to create innovative, competitive products. While much work has been done on data portability from the perspective of regulators and incumbents, less has been done from the perspective of potential competitors. The work that has been done in this area is largely theoretical, focusing on network effects,[16] switching costs,[17] and the exponential effects of data aggregation.[18] We began to address the lack of information about data portability's competitive utility by putting real, ported data into the hands of people who would be expected to build new services with it.

This paper specifically focuses on the data Facebook users can download about themselves. We chose to focus on data export solutions rather than continuous data flow solutions, like API access. In general,

---

[10] Representative David Cicilline, Remarks at New America: A Deep Dive Into Data Portability (June 6, 2018), *available at* https://www.youtube.com/watch?v=uW6ONz0CLyc.

[11] *Facebook, Social Media Privacy, and the Use and Abuse of Data Before the Senate Committee on the Judiciary, Senate Committee on Commerce, Science, and Transportation*, 115th Congress (2019) https://www.judiciary.senate.gov/meetings/facebook-social-media-privacy-and-the-use-and-abuse-of-data

[12] Mark Zuckerberg, *Four Ideas to Regulate the Internet*, Facebook Newsroom (March 30, 2019), https://newsroom.fb.com/news/2019/03/four-ideas-regulate-internet/

[13] *See*, *e.g.* Parliament and Council Regulation 2016/679 of April 27, 2016, on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 at art. 20, http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[14] Alexia Tsotsis, *Facebook Now Allows You To "Download Your Information,"* TechCrunch (Oct. 6, 2010), https://techcrunch.com/2010/10/06/facebook-now-allows-you-to-download-your-information/

[15] *The Data Liberation Front Delivers Google Takeout*, Google Data Liberation Blog (June 28, 2011), http://dataliberation.blogspot.com/2011/06/data-liberation-front-delivers-google.html

[16] *See* Gus Rossi & Charlotte Slaiman, *Interoperability = Privacy + Competition* Public Knowledge (Apr. 29, 2019), https://www.publicknowledge.org/blog/interoperability-privacy-competition/ Bennett Cyphers & Danny O'Brien, *Facing Facebook: Data Portability and Interoperability Are Anti-Monopoly Medicine* Electronic Frontier Foundation (July 24, 2018), https://www.eff.org/deeplinks/2018/07/facing-facebook-data-portability-and-interoperability-are-anti-monopoly-medicine

[17] *See* Peter Swire & Yianni Lagos, *Why the Right to Data Portability Likely Reduces Consumer Welfare: Antitrust and Privacy Critique*, 72 Maryland Law Review 335 (2013).

[18] *See* Daniel L. Rubinfeld & Michal Gal, *Access Barriers to Big Data*, 59 Arizona Law Review 339 (2017).

we focused on data exports because they give users full control over their data, as opposed to API access, where incumbents retain significant control over who can access what data, and how often.[19] We expand on our concerns with the limitations of API-based access in the Background section and Appendix A.

We brought examples of a data set exported from Facebook to developers, product managers, and executives from innovative tech companies in NYC and asked them a simple question: what new products or features could you create with this data? These individuals had a range of experiences, from tiny startups to name-brand tech companies, and junior employees to senior managers. Between them, they had experience both in conceiving products and features that could appeal to users and in doing the technical work required to turn those ideas into real-world products.

For reasons detailed below, our cohort found surprisingly limited value in the data Facebook allowed users to download. Often, the data provided too little context or was too closely tied to the design of Facebook itself to build new products with or use to bootstrap growth. While the exported data might suggest a specific new product or feature, upon further discussion it usually became clear that the success or failure of that product or feature was not actually related to access to Facebook data.

We chose to analyze Facebook data in this paper as a first step in understanding the real-world value of data portability more broadly. Facebook is the focus of wide ranging competitive scrutiny and will likely continue to be so in the future. Although this makes Facebook a reasonable place to start our investigation, we also recognize that the usefulness of Facebook's data may be somewhat idiosyncratic. Other types of applications may be more or less conducive to the pro-competitive effects of data portability. That is why we see this paper as a first step in our analysis, and hope to expand this research to other platforms in the future.

[19] It is possible that regulating these API connections could significantly reduce incumbent's control over how competitors might manipulate them to block competitors. However, other attempts to regulate integration in a rapidly evolving technical environment suggest that this type of regulation can be highly challenging to maintain and enforce. *See, e.g.* Harold Feld, *My Insanely Long Field Guide To The War On CableCARD - Part I: More Background Than You Can Possibly Imagine*, Tales of the Sausage Factory (October 19, 2014), https://wetmachine.com/tales-of-the-sausage-factory/my-insanely-long-field-guide-to-the-war-on-cablecard-part-i-more-background-than-you-can-possibly-imagine/ for a partial history of the Federal Communication Commission's multi-decade attempt to bring interoperability and competition to television set-top boxes and the cable industry's attempt to thwart it. In contrast, the interoperability mandate imposed upon the AOL Instant Messenger (AIM) platform by the Federal Communications Commission as part of the AOL - Time Warner merger in the early 2000s provides an example of a successful attempt to mandate technical compatibility, at least for a period of time. *See In the Matter of Applications for Consent to the Transfer of Control of Licenses and Section 214 Authorizations by Time Warner, Inc. and America Online, Inc. Transferors, to AOL Time Warner, Inc., Transferee*, Memorandum Opinion and Order, 16 FCC Rcd 6547 at 6627-28 ¶¶ 191-95 (2001). The relatively static technical nature of real-time text-based chat may have contributed to that success. Nonetheless, AIM was eventually surpassed by richer real-time interactive technologies. It is unclear what impact the interoperability mandate had on that dynamic, or how if the interoperability mandate would have been able to successfully incorporate new features.

This paper will begin by contextualizing Facebook's approach to data portability, both in terms of legal requirements and the history of what data Facebook has made available to developers over the years. Then, we will review our findings from the interviews, which ended up highlighting some of the shortcomings of the competition-by-portability approach. We did not come into this investigation with a presupposition about how useful the Facebook data would be to our cohort—the focus on shortcomings is the result of the primary conclusions raised by the cohort itself.

Finally, we make suggestions about how platforms can create competition-friendly portability offerings and how policymakers can bring nuance to their evaluation of data portability as a regulatory tool.

While our study has made us skeptical that data portability alone can be the primary driver of increased competition, data portability requirements may still be useful in specific contexts. Their success will hinge on a nuanced understanding of data portability's strengths and limitations.

Background

There are two primary ways platforms can implement data portability: public-facing APIs (application programming interfaces) and one-off exports.

With a public-facing API, an incumbent exposes part of its backend functionality to third-party applications, usually with a user's permission. The user does not act as an intermediary between platforms in the sending and receiving of data, besides giving initial permission to do so. This allows for continuous integration between the incumbent and the third-party application.

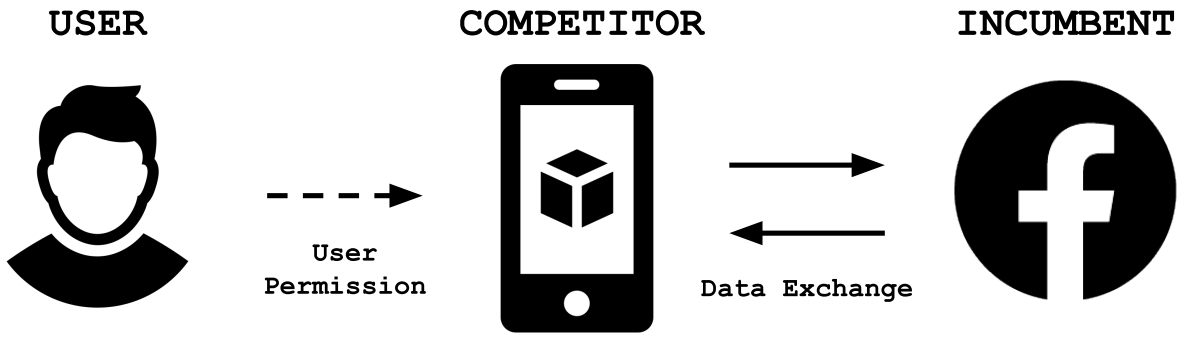

Figure 1: Data flow for a public-facing API[20]

With a one-off export, a user downloads her own data locally and then, if she so chooses, can upload it to a third-party application. Here, the user does act as an intermediary, initiating and controlling the data transfer process between the two platforms. The requirement for the user to manually transfer the data

[20] Images, from left to right: "User" by Gregor Cresnar from The Noun Project / CC BY 3.0; "App" by Adrien Coquet from The Noun Project / CC BY 3.0; Facebook logo from the Facebook Brand Resource Center.

makes this model a poor fit for continuous integration between the incumbent and the third-party application.

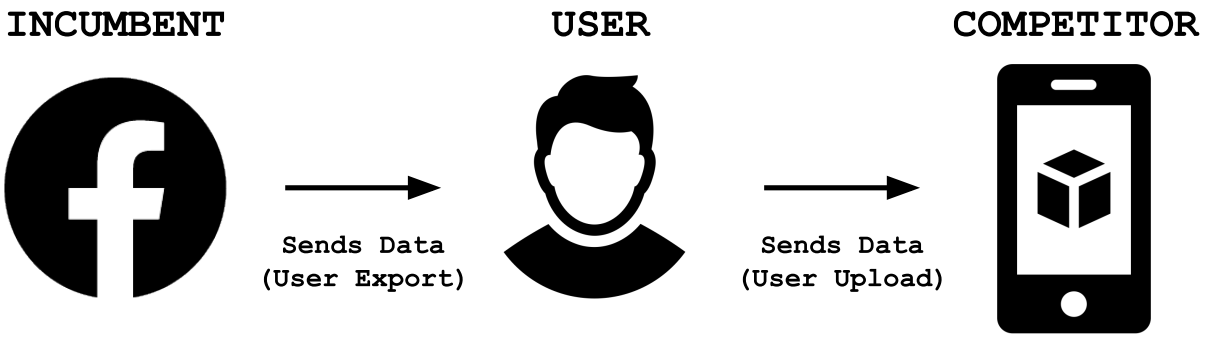

Figure 2: Data flow for a one-off export

This paper will consider only one-off exports because public-facing APIs have multiple shortcomings as a competitive measure. First, incumbents can monitor how competitors use their APIs and potentially use that information to copy them. Second, incumbents can limit or cut off API access to competitors as they see fit. Third, incumbents might change the data or the structure of the data they make available, creating technical overhead or even destroying the business model for competitors. Facebook competitors have faced all of these issues with Facebook's Graph API (see Appendix A). Although there may be governance mechanisms to address these issues, they are outside the scope of this paper.

One-off data exports have none of these issues. When users download their data locally, they gain complete control over it. The original platform cannot track what users do with this data, nor prevent them from uploading it to another platform. They also cannot track how or how often competitors use the data. And even though the original platform may change the data it makes available in its portability offering, the changes do not retroactively affect users who have already downloaded their data.

Even with a one-off data export, though, users do not gain access to all the information a platform may have on them. Portability offerings, including Facebook's, usually exclude non-personal data (for example, if a user likes the page for the band U2, Facebook will not export the fact that it knows U2 is a band) and inferences the platform has made (for example, Facebook will not export a user's opinion on gun control if it has determined it from links they clicked).

Facebook offers both a public-facing API and a one-off data export, with the Graph API and Download Your Information (DYI) tool respectively.

The Graph API is open to any competitor that allows its users to sign into their application through Facebook. The competitor embeds a snippet of code into their product that links the user account to

their Facebook account. In return, the competitor can access certain data from that user's Facebook account, such as their name, profile picture, and friends, among other things.[21]

Figure 3: Sign in with Facebook button[22]

In contrast, DYI lets users download a snapshot of essentially all the data they have ever entered into Facebook, including photos, videos, and text. Users can request this file either in an HTML format, for personal viewing, or a JSON format, for computers to read.[23] Because we are putting API portability to the side for the reasons mentioned above, this paper will mostly focus on DYI.

Data portability efforts on social networks, like the Graph API and DYI, come with their own set of distinct issues. Most importantly, they face a tradeoff between privacy and usefulness to competitors in their implementations.[24] The data on social networks is inherently relational, and platforms must create a perimeter between where one user's data ends and another's begins. The smaller that perimeter, the less contextual data exporting users and competing platforms have, and the less ability third parties have to create competing platforms.

It is a difficult balance to strike between privacy and competitive utility questions. For example, if hypothetical Facebook user Alan likes Brandi's photo, should that photo be available in Alan's data export? Should Alan's like be available in Brandi's data export? What if both users upload their data to a new service? Should that new service be able to re-link the two sides of the interaction? This paper will not make suggestions on how to make this privacy/competitive utility trade-off, but the findings section will showcase how this tension might affect competitors.[25]

[21] *Overview - Graph API*, Facebook for Developers, https://developers.facebook.com/docs/graph-api/overview (last visited Oct. 18, 2019).
[22] *Login Button - Facebook Login*, Facebook for Developers, https://developers.facebook.com/docs/facebook-login/web/login-button/ (last visited Oct. 18, 2019).
[23] Facebook has offered this feature since 2010, but it was mostly for personal archiving purposes. They introduced the more computer-readable JSON format in 2018, shortly after Europe's General Data Protection Regulation came into effect. *See* Facebook Portability White Paper, *supra* note 7.
[24] *See* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 University of Colorado Law Review 1117 (2013).
[25] Regulators should be mindful that incumbents could advocate for privacy-prioritizing regulations that result in higher barriers of entry for challengers.

Methodology

Interviews

The roughly 12 interview candidates for this study came from the Engelberg Center's network in the NYC tech community, including product managers, developers, business leaders, and representatives of venture capital. Interviewees were affiliated with a wide range of companies, from large, household-name tech companies to small, seed-round startups. The group was not a statistically representative sample of the tech industry, nor did we intend for it to be. Our goal was not to poll the industry but to identify different ways ported data could or could not be used to innovate. To this end, our group was designed to represent diverse perspectives from different parts of the industry.

Interviewees were brought together in small group cohorts to discuss the possibilities presented by the Facebook data set. Discussions lasted about two hours and operated under the Chatham House rules in order to allow participants to discuss opportunities and challenges freely without fear of negatively impacting their company or their own professional opportunities.

Cohorts participated in an open-ended exercise to examine the data made available when a Facebook user exports her data via DYI. They first discussed the sub-components of the Facebook platform (chat, contacts, marketplace, etc.) in order to consider the types of data that Facebook might have about an individual. They then examined exported Facebook data to identify what was available and how it was structured. Specifically, we asked the cohort how they might use the data in innovative applications or features that could compete with Facebook.

Although we were not strict about our definitions, we guided the discussion toward certain conceptions of both "innovation" and "competition." "Innovation" was considered broadly—would the data allow an entity to create some sort of new value for its end users? This "new value" could be as small as a feature or as large as an entirely new offering.

"Competition" was conceived in an equally broad context. Compete did not mean "drive Facebook out of business," or even "draw users away from Facebook." It also did not need to be part of a for-profit company. We considered competitive products simply to be ones that were viable in a world where Facebook exists. For a business, viability means profitability. Other types of services could be sustainable on their own terms.

Data

Facebook provides minimal public documentation on the structure of their personal data export.[26] We therefore had to reverse-engineer the structure ourselves by downloading the Facebook data of one of

[26] There is no public documentation for the data structure that Facebook uses for user data exports. In May of 2019 we placed an initial request to Facebook for a complete example of the data structure that could be exported by users. Our request was for the data structure itself (categories and subcategories of the types of data that could be

our researchers. The data was pseudonymized, to protect the privacy of our researcher, and shortened, to make it easier for participants to read. Details on how this was done and the code used to do it are available on GitHub.[27]

The data, as downloaded from Facebook and as presented to study participants, was in a file tree structured as follows:

- about_you/
    - your_address_books.json
- ads/
    - ads_interests.json
    - advertisers_who_uploaded_a_contact_list_with_your_information.json
    - advertisers_you've_interacted_with.json
- apps_and_websites/
    - apps_and_websites.json
- comments/
    - comments.json
- events/
    - event_invitations.json
    - your_event_responses.json
    - your_events.json
- following_and_followers/
    - followed_pages.json
    - following.json
    - unfollowed_pages.json
- friends/
    - friends.json
    - rejected_friend_requests.json
    - removed_friends.json
    - sent_friend_requests.json
- groups/
    - your_group_membership_activity.json
    - your_groups.json
    - your_posts_and_comments_in_groups.json
- likes_and_reactions/
    - pages.json
    - posts_and_comments.json
- marketplace/
    - items_bought.json
    - items_sold.json
- payment_history/
    - payment_history.json
- posts/
    - other_people's_posts_to_your_timeline.json
    - your_posts.json

exported by a user), not for any individual or aggregate user data. Although Facebook was willing to discuss our request on numerous occasions, they were ultimately unwilling or unable to provide us with the data structure we requested. While we have made an effort to reconstruct significant portions of the data structure for the purposes of this investigation, the absence of any sort of available documentation of the data that can be exported by users presents a significant challenge to a new service trying to make use of exported data.

[27] *Portability-Project*, GitHub, https://github.com/gajeam/Portability-Project (last commit Sept. 13, 2019).

- profile_information/
    - profile_information.json
    - profile_update_history.json
- saved_items_and_collections/
    - saved_items_and_collections.json
- search_history/
    - your_search_history.json

This data represents most, but not all possible, data a user can export. For instance, the researcher whose data this came from never used Facebook's check-in feature. Therefore the data set is missing a folder titled *location*, which would include that information. Certain values within these files may also be missing based on the user's settings or behaviors on Facebook. For example, if the researcher used Facebook payments but never cancelled a payment, no data on how that action is represented would be available here. Finally, highly sensitive personal data, namely photo files, video files, and private messages, were removed entirely from this data set.

Findings

Finding #1: Interviewees were mostly interested in using Facebook data for growth and targeting but were underwhelmed with what was available.

When looking at the exported data, participants were most interested in information about the user's interests and her social graph. However, participants were largely underwhelmed with the data made available to them in these areas, and found them insufficient to build competitors.

User interest data was represented in the export as a list of pages that a user had "liked." Participants inspired by this list came up with product and feature ideas, including one to use interest categories to match users with communities on a new platform. Other participants debated how well Facebook's interest categories could be mapped onto a new platform, or how stable those categories would be over time.

Participants were also interested in leveraging Facebook's social graph, but here too they found themselves unable to do what they wanted. DYI contains two types of social graph data: first, it has a list of a user's friends, their names, and the time they became friends.[28] Second, if a user ever connected Facebook with their phone contact list, it makes those contacts available.[29] In general, there was a lack of external contact information on connections. This made it difficult for participants to find ways to use

[28] For some users, in our data set about 2%, it also had their phone number or email address. The settings that allowed for this are unclear, but participants said that was not enough to be useful.
[29] Competitors may not be able to map from the social graph data onto the contact list data because users may not always use someone's real name in their contact list.

this data to bootstrap growth by contacting the user's connections directly, the way LinkedIn did through Gmail contacts[30] or Instagram did through sharing posts on other social networks.[31]

Participants often concluded that the social graph could largely be reconstructed with a sufficiently large percentage of exported data from all Facebook users. This created a sort of chicken-and-egg situation: Facebook's social graph data could be useful once a service reached sufficient size to create a meaningful network, but once a service reached sufficient size, the social graph could largely be created directly from users without relying on exported Facebook data. In other words, social graph data itself may not be useful in catalyzing the initial set of users.

Participants were largely uninterested in data related to comments, statuses, groups, and events, even when we specifically asked them to consider this information. This may reflect a shortcoming in the data made available in these areas (see Finding #2 below) or a general lack of confidence in a new business built on these features.

Participants were curious about the personal profile data made available, such as hometown and religion, but did not think importing that data into a new service would help users overcome any significant barrier to trying new services. They felt that a sufficiently compelling service would not have trouble convincing users to re-enter that information as part of their profile, and that the requirement of adding that information would not present a disincentive for new users to try a service. They viewed the primary challenge of starting a new service to be attracting users in the first place, not obtaining the type of profile information available in a DYI export.

Perhaps of more interest, few participants were drawn to even conceive of new services that made meaningful use of that type of basic demographic information. To the extent they found the information useful, it was for creating a new targeted advertising product, not in creating the core functionality that would initially draw users to the platform. As we discuss below (Finding #3), using exported Facebook data to create a new interest-based advertising network may not be an optimal outcome of a data portability regime.

[30] *See* Linda Sandler, *LinkedIn Customers Allege Company Hacked E-Mail Addresses*, Bloomberg (Sept. 21, 2013), https://www.bloomberg.com/news/articles/2013-09-20/linkedin-customers-say-company-hacked-their-e-mail-address-books

[31] Today, Facebook is employing the strategy yet again, to let users cross-post Instagram stories as Facebook stories. *See How do I share my Instagram story to Facebook?*, Instagram Help, https://help.instagram.com/1936968516554161 (last visited Oct. 18, 2019).

Finding #2: Interviewees found the data Facebook made available to be an insufficient foundation for recreating or directly competing with specific Facebook features.

Participants found the data from the DYI tool insufficient to form the basis for products similar to Facebook or any of its features. They found much of the data to be too decontextualized to use, and the context necessary to make it useful would cross the line into what Facebook considers someone else's data. For instance, participants saw this sample data of a fictional Alan Aaronson commenting on fellow fictional Facebook user Brandi Barnacle's photo:


```
{
	"timestamp": 1477442502,
	"data": [{
		"comment": {
			"timestamp": 1477442502,
			"comment": "What a beautiful picture that is!",
			"author": "Alan Aaronson"
		}
	}],
	"title": "Alan Aaronson commented on Brandi Barnacle's photo."
}
```


Figure 4: Comment data example

Facebook provides the text of Alan's comment, the time it was posted, and the fact that it was on Brandi Barnacle's photo. However, there is no data about the photo itself, who responded to his comment, who liked it, or which "Brandi Barnacle" the data refers to. Furthermore, even if Brandi were to upload her data to the same third-party platform as Alan, it would be impossible to match which one of Brandi's photos he was commenting on.[32] When participants were asked whether Facebook provided sufficient information on comments, groups, or statuses to recreate even rough versions of what Facebook offers, the answer was consistently no.

Participants similarly found the data insufficient to recreate or compete with less personal features, like Facebook's events platform. This deficiency highlights the discrepancy between data Facebook makes available about events a user has attended:


```
{
	"events_joined": [{
		"name": "Beer Enthusiast - February Meet Up",
		"end_timestamp": 1518584000,
		"start_timestamp": 1518573200
	}]
}
```


Figure 5: Attended event data example

[32] One could make a decent, though imperfect, estimate about which photo was being commented on if Facebook provided the timestamp for when Brandi's photo was posted. However, they do not—the two timestamps in the JSON are the same, and both represent when the comment was posted.

and ones they have hosted:


```
{
	"your_events": [{
		"name": "Alan's Big Halloween Party",
		"start_timestamp": 1446350400,
		"end_timestamp": 0,
		"place": {
			"name": "123 Cherry Street, Chattanooga, TN",
			"coordinate": {
				"latitude": 35.0553176,
				"longitude": -85.3087483
			}
		},
		"description": "It's the creepiest party in town! Come on down."
		"create_timestamp": 1444176035
	}]
}
```


Figure 6: Hosted event data example

For events he attended, the user can see only the name, start time, and end time. For events he hosted, he has a much richer set of information available, although, notably, nothing about the attendees. This event data may be enough for a competitor to create a calendar, but participants found it insufficient for transferring to a new events platform. Once again, entity matching was an issue: if another user uploaded data about the "Beer Enthusiast - February Meet Up," there is no surefire way to link it to Alan's data. Nor is there a way to distinguish between multiple, unrelated "Beer Enthusiast - February Meet Up" events coordinated via Facebook. Facebook does not make available any sort of unique identifier to reconcile multiple events with this same name and start time. This may not be an issue for "Beer Enthusiast - February Meet Up" but could be a real issue for an event just called "Birthday" hosted on a Saturday at 8:00 p.m.

Participants repeatedly pointed to missing data that would prevent them from using the Facebook data export to bootstrap new products. This may be intentional on the part of Facebook in an effort to protect the privacy of other users. Nonetheless, in this case what helps privacy, hurts competitive utility.

Finding #3: Participants mostly came up with products that were so similar to Facebook itself that they may struggle to compete.

When participants brainstormed new products and features to build with Facebook's portability offering, their ideas often shared key traits with the Facebook product itself. They usually offered users social connections or data-driven insights, and earned revenue with data-driven ads.

In some ways, this result should be unsurprising. Facebook is a social network connected to an advertising platform, and any product built with its data is likely to conform at least somewhat to the source material. Facebook is the dominant player in this area and offers its product free of charge—it is possible that users may anchor to this price for social networking, making it difficult for a competitor to

adopt a revenue model that requires users to pay directly (e.g., subscription, freemium).[33] Inflexibility in product and revenue model may make the products that come out of data portability less innovative.

At least in part, this is because the Facebook product has an advantage over any competitor hoping to use data-driven ads, or indeed, to use their data for machine learning at all.[34] Machine learning models are developed by training the algorithms on large data sets, and often larger data sets allow for more accurate and precise algorithms.[35] As long as a competitor is relying on a subset of Facebook data to get started, Facebook's massive number of users gives it a uniquely broad and deep data set for any number of applications.

Despite coming up with a number of ideas that revolved around data-driven insights and machine learning, participants were skeptical that user-driven data portability would lead to significant competition in this space. They identified two primary obstacles to success.

The first is that the relative value of the ported data is likely much higher for the new platform than for the user bringing her Facebook data to that platform. A service would need a significant number of users to port their data into the service in order to build an algorithmic product.[36] In aggregate, that data would be valuable to the service, but any individual user's data would be of limited value. Thus, the value that the user receives for bringing data to the service would be limited, especially in the early stages of the service's development. Participants struggled to construct a service they believed would be compelling enough to convince users to individually share their Facebook data.

A potential exception to this concern is an app that engages users by letting them use their data in a novel way, and earns money through algorithmic insights about that data. This model is often discussed in the context of "selfie transformation" apps that allow users to upload and alter their pictures (for example, convincingly aging the subject of the photo), or dating apps that allow users to upload pictures.

---

[33] *See* Amos Tversky & Daniel Kahneman, *Judgment and Decision Making: An Interdisciplinary Reader*, 35 (1986); Katherine J. Strandburg, *Free Fall: The Online Market's Consumer Preference Disconnect*, 2013 University of Chicago Legal Forum 95.

[34] Some commentators have suggested that the advantages inherent in having access to such large quantities of data would justify requiring some entities to share important data sets. *See* Samuel Himel & Robert Seamans, *Artificial Intelligence, Incentives To Innovate, And Competition Policy*, Competition Policy International (Dec. 19, 2017), https://www.competitionpolicyinternational.com/wp-content/uploads/2017/12/CPI-Himel-Seamans.pdf Others have raised concerns that forcing data sharing could result in platforms using data to train a machine learning model and then deleting the data in the name of user privacy, thus effectively preventing a competitor from following the same path. *See* C. Scott Hemphill, *Disruptive Incumbents: Platform Competition in an Age of Machine Learning*, Colum. L. Rev. (forthcoming 2019).

[35] Pedro Domingos, *A few useful things to know about machine learning*, 55 Communications of the ACM, no. 10, 2012, at 78.

[36] Although the amount of data required to build machine learning models is substantial, there is evidence that at least in some domains, the value of additional data decreases at some point. Other algorithm design factors such as feature engineering and model selection may be more important. *See, e.g.* Xinran He et al., *Practical Lessons from Predicting Clicks on Ads at Facebook*, Proceedings of the Eighth International Workshop on Data Mining for Online Advertising (Aug. 24, 2014).

These apps then aggregate the images collected from users in order to train and sell machine learning models, such as facial recognition. There has been an increasingly strong public backlash against this type of model,[37] which suggests that regulators may wish to proceed with caution if they hope to rely on services based on these models to address competition issues.

The second obstacle is that relying on individual data transfers is an inefficient way to collect large amounts of data, especially with the one-off export mode of portability. A one-user-at-a-time model is unlikely ever to allow a new service to accumulate enough data to meaningfully compete with Facebook.[38]

Recommendations

By talking with the technologists and entrepreneurs who could be expected to use ported data to create new, innovative products, we found that data portability alone, particularly as facilitated by Facebook with its DYI tool, is not up to the task of increasing online competition. This suggests that data portability should not be the primary tool that regulators use to address platform competition concerns. If incumbent platforms want to argue otherwise, the burden is on them to prove that competitors have use for the data made available.

Nonetheless, data portability may be able to play a supporting role. Below, we outline some regulatory and technical recommendations to make data portability more effective in improving competition.

Regulatory Recommendations

The most important step that policymakers can make in designing a data portability regime is to clearly express its intended purpose.

Data portability regimes can differ significantly depending on the purposes they intend to achieve. A competition-maximizing regime will likely include fundamentally different data than a privacy-maximizing one. In light of this, policymakers cannot simply mandate abstract data portability requirements. Instead, they must clearly articulate the specific goals and purposes of such a regime.

Implementing any data portability process will require significant tradeoffs between competing public policy goals. Without a clear articulation of the intended purpose of the regime, regulators will be unable to accurately tailor their requirements to those goals. In articulating those goals, policymakers would be well served to understand goals that may be achievable through a data portability regime and those that data portability is likely poorly designed to address.

---

[37] *See, e.g.* Sidney Fussell, *FaceApp Is Everyone's Problem*, The Atlantic (July 19, 2019), https://www.theatlantic.com/technology/archive/2019/07/faceapp-mess/594361/

[38] This inefficiency is what drives trusted third-party data repository proposals. *See* Himel & Seamans, *supra* note 34.

In the context of competition policy specifically, data portability—whether mandated by regulation or created by industry consensus—is most effective when it is applied to the following types of data:

- **Data that can be consistently structured and used in specific, easy-to-anticipate applications.** One example of this type of application is financial data. Financial data has a relatively limited and stable number of elements (amount of money, payment/transfer source, payment/transfer recipient, etc.). It also has a relatively limited number of uses (make a payment, receive a payment). That makes it easier to build a standard set of data structures for the financial industry than for something as dynamic and complex as all social network sites.
- **Data that does not depend on context from other, private data for value.** Exported Facebook data relies on links to other users' private data for value. In contrast, playlists on music streaming apps, like Spotify or Apple Music, have value irrespective of other users.
- **Data with clear ownership.** Facebook data raises many questions about ownership. For example, if a user posts a poll, should she be able to download the results of the poll? Or is that data owned by those who responded? (Facebook implies the latter.) In contrast, fitness tracking data, like on Runkeeper or MyFitnessPal, is more clearly owned by the person who generated that data.

Beyond what types of data are amenable to data portability, there are open questions about what kind of problems data portability has the potential to address, especially given the lack of clear success stories. For example, it is unclear if data portability can meaningfully mitigate advantages based on network effects. The technical challenge of porting data to a new service may be relatively small compared to the business challenge of building awareness of the competitor among potential users.

Similarly, portability may not help small platforms overcome incumbent advantages based on the volume of data. Beyond more data for better machine learning algorithms, usage data can allow large platforms to anticipate and capitalize on emerging trends.[39] Allowing users to port data from the platform to a competitor may be an inefficient mechanism to mitigate this advantage, although again, more research is necessary here.

Technical Recommendations

Should regulators choose to use data portability to improve sector competition, they should make sure that incumbents design their portability offerings to maximize usefulness to competitors. Below are some technical principles incumbents should follow (and regulators may want to enforce) in designing their data exports for use by competitors:

- **Document the structure.** Competitors hoping to build products with ported data need to be able to understand what the data they will encounter might look like. Facebook's Graph API has clear, well-structured documentation explaining all possible data a developer might encounter. The DYI

[39] *See, e.g.* Evelyn M. Rusli, *Facebook Buys Instagram for $1 Billion*, New York Times (Apr. 9, 2012); Julie Creswell, *How Amazon Steers Shoppers to Its Own Products*, New York Times (June 23, 2018).

tool has far less documentation,[40] and developers hoping to integrate with it can only learn by trial and error.[41]

- **Focus on stability.** Some of our interviewees said they would not create a new product that used Facebook's Graph API because they feared it would change over time, introducing technical debt and potentially breaking their product entirely. This challenge is even greater for ported data, because users may have downloaded their data at any time and developers importing that data onto a new platform will have to continually support older versions. Platforms should change the structure of their data exports as little as possible and guarantee that certain aspects will not be deprecated without significant notice.
- **Cut out the middleman.** Downloading and uploading data is a cumbersome process for users that requires some technical savvy. To facilitate competition, incumbent platforms can allow their data downloads to be sent directly to competitors, ideally in a straightforward, no-strings-attached manner.
- **Provide unique identifiers.** Competitors can better create new products if they can match data points between users. With Facebook's DYI tool, if User A has commented on User B's photo and both upload their data to a new platform, the data exports are insufficient to reconstruct that connection. Providing unique identifiers for every data object allows for these connections to be made.
- **Let users take their network with them.** Participants expressed interest in bootstrapping their growth with Facebook's social graph. However, they found the DYI tool to be insufficient, because it only shared the names of friends and the timestamp they were friended at.[42] Facebook can encourage growth by sharing friends' other contact information. However, exposing user emails and phone numbers without permission raises privacy concerns. There have been a number of proposals about privacy-friendly ways to export the social graph.[43] The real-life utility of a hashed social graph is unclear, because no major platform has offered one.

The Future of Data Portability

The policy concerns raised by large platforms do not lend themselves to simple, one-size-fits-all solutions. While we believe it is unlikely that data portability can serve as the primary mechanism to the majority of these concerns, especially for social networks, we do not believe policymakers should abandon it as a potential regulatory tool altogether.

---

[40] *See Accessing & Downloading Your Information*, Facebook Help, https://www.facebook.com/help/1701730696756992 (last visited Oct. 18, 2019).

[41] The Data Transfer Project launched by major platforms may be a step toward creating this type of documentation, https://datatransferproject.dev/

[42] It is actually possible for users to create a rough social graph, by hashing the timestamp and the names of both friends. This creates what is called an "edge-first" graph, and allows two users to find each other if both have uploaded friends lists containing the other. *See* Facebook Portability White Paper *supra* note 7, for the possibilities and limitations of this approach.

[43] *See, e.g.* Josh Constine, *Facebook shouldn't block you from finding friends on competitors*, TechCrunch (Apr. 13, 2018), https://techcrunch.com/2018/04/13/free-the-social-graph/; Facebook Portability White Paper *supra* note 7.

The type of data portability considered in this paper may be especially fraught in the context of large platforms based on social networks. Solutions that are effective in the context of competition policy may be deeply counterproductive in the context of privacy policy. As a result, it would be easy for incumbents to use detailed debates about the form of data portability regulation to distract regulators from more effective interventions.

Nonetheless, there is value in and of itself in giving people a level of control over the data they create and contribute to a platform. The European approach recognizes that, and uses it as the foundation for the GDPR's data portability requirements. Policymakers should not feel compelled to rely on competitive concerns in order to feel comfortable recognizing this as a legitimate and independent benefit of data portability.

Similarly, there can be great innovative value in making it easier to port data between services and platforms. Even if that movement does not significantly address the competitive landscape defined by large platforms, it can fuel a wide range of applications and uses with inherent value.

Ultimately, data portability is a tool like any other. If it is used with precision in the context of a nuanced understanding of its capabilities, it can be effective. However, if deployed in a blunt attempt to address a wide range of complex concerns, it will likely fail.

Appendix A

We outline three problems with public-facing APIs as a means to increase competition. Below, we discuss how Facebook has manifested these problems with their implementation and administration of its Graph API:

- **Incumbents can monitor API usage and use that information to undercut competitors.** Facebook's Platform Policy explicitly states that for apps that integrate with the Graph API, Facebook reserves the right to "create apps or products that offer features and services similar to your app."[44] Facebook has cloned many popular apps in the past—compare TikTok, Snapchat, and Houseparty to Facebook's LASSO, Slingshot, and Bonfire. It has also imitated popular features from other apps, like stories from Snapchat, check-ins from Foursquare, and live video from Meerkat. Apps that request certain sensitive data, such as age range, birthday, friends, events, and gender, can require a potentially invasive app review from Facebook. This can include screencasts of the app, tax forms, or some form of government ID.
- **Incumbents can limit or cut off API access to competitors as they see fit.** Facebook's Platform Policy explicitly reserves this right.[45] Controversially, Facebook exercised this right against Vine, Twitter's now-defunct short-form video app. The day it launched, Mark Zuckerberg himself approved cutting Vine off from the Graph API, which Vine was using to help users find friends who had signed up.[46]
- **Incumbents might change the data or the structure of the data they make available to competitors.** Facebook has updated its Graph API numerous times, introducing breaking changes along the way. Zynga experienced this when Facebook deprecated the Graph API functionalities that originally hypercharged their growth. First, Facebook limited users' ability to share their in-game progress to Facebook's newsfeed. Then, Facebook effectively removed users' ability to invite friends to play. Zynga's reduced API access contributed to the company's valuation plummeting from almost $15 billion to $3 billion over six months in 2012.

[44] *Facebook Platform Policy*, https://developers.facebook.com/policy/ (last visited Oct. 18, 2019), Rule § 7.10.
[45] *Id.* at Rule § 7.16
[46] Adi Robertson, *Mark Zuckerberg personally approved cutting off Vine's friend-finding feature*, The Verge (Dec. 5, 2018), https://www.theverge.com/2018/12/5/18127202/mark-zuckerberg-facebook-vine-friends-api-block-parliament-documents