THE ECONOMICS OF SOCIAL DATA: AN INTROCUCTION

By

Dirk Bergemann and Alessandro Bonatti

March 2019

COWLES FOUNDATION DISCUSSION PAPER NO. 2171

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

The Economics of Social Data: An Introduction

Dirk Bergemann[†] Alessandro Bonatti[‡]

March 26, 2019

Abstract

Large internet platforms collect data from individual users in almost every interaction on the internet. Whenever an individual browses a news website, searches for a medical term or for a travel recommendation, or simply checks the weather forecast on an app, that individual generates data.

A central feature of the data collected from the individuals is its social aspect. Namely, the data captured from an individual user is not only informative about this specific individual, but also about users in some metric similar to the individual. Thus, the *individual data* is really *social data*. The social nature of the data generates an *informational externality* that we investigate in this note.

Keywords: Individual Data, Social Data, Informational Externality, Internet Platforms, Data Collection, Data Markup.

JEL Classification: D80, D82, D83.

[*] We thank Steve Berry and Roger McNamee for helpful discussions.

[†] Douglass and Marion Campbell Professor of Economics and Chair, Department of Economics, Yale University, New Haven, CT 06511, dirk.bergemann@yale.edu

[‡] Associate Professor of Applied Economics, Sloan School of Management, MIT, Cambridge, MA 02142, bonatti@mit.edu.

1 Individual Data and the Internet

The rise of large internet platforms, such as Facebook, Google, and Amazon in the US, and similar large entities in China, such as JD, Tencent and Alibaba is leading to an unprecedented collection and use of individual data. The ever increasing user base of these platforms generates massive amounts of data about individual consumers, their preferences, their locations, their friends, their political views and almost all other facets of their lives. In turn, many of the services provided by Facebook, Google, and the large majority of the internet companies rely critically on these individual data. The availability of individual-level data allows the companies to offer refined search results, personalized product recommendations, informative ratings, timely traffic data, and of course, targeted advertisements (see Bergemann and Bonatti (2019) for a recent introduction).

A central feature of the data collected from the individuals is its social aspect. Namely, the data captured from an individual user is not only informative about this specific individual, but also about users in some metric similar to the individual. Thus, the *individual data* is really *social data*. It almost always conveys insights about users similar to the individual who provides the information. The social nature of the data generates an *informational externality*: the individual user is producing information about himself and those similar to him in some relevant metric. In the context of geolocation data, an individual user conveys information about the traffic conditions for those nearby drivers. In the context of shopping data, an individual's purchases convey information about the appeal of a given product or service for consumers with similar purchase histories. The informational externality is a simply a by-product of the social data in the sense that the individual data also conveys information about (appropriately defined) nearby individuals.

The recent disclosures on the use and misuse of social data by the internet platforms, in particular Facebook, indicate the need to reflect about the largely unsupervised and unregulated use of individual data by these companies. To the extent that individual users provide most of the original data in their interaction with these platforms, it is important to understand the nature of the trade between large internet platforms and their users, and in particular, whether individual users are receiving the appropriate compensation for their data.

This question gains importance in the presence of the informational externality generated by the social data. To some extent, the informational externality is just another instance of a significant economic externality. And we know from the study of other externalities, such as the environmental externality of carbon emissions, that in the presence of economic externalities, the market by itself rarely guarantees the socially efficient outcome.

In this note, we discuss the value and the price of social data and how this defines a clear "data markup". We outline social welfare consequences and provide a preliminary discussion of regulatory responses. However, before turning to regulatory implications, it is necessary to understand two critical aspects of the economics of the Internet that explain the seemingly unlimited thirst of these companies for data. First, how the possession of individual data changes the terms of trade between consumers, advertisers, and large internet platforms. Second, how the social dimension of the data magnifies the value of individual data for the platforms.

2 How Data is Used

Large internet platforms collect data from individual users in almost every interaction on the internet. Whenever an individual browses a news website, searches for a medical term or for a travel recommendation, or simply checks the weather forecast on an app, that individual generates useful data. Information about an individual can impact the volume of trade and the level of surplus generated on a platform (see Bergemann and Morris (2019)). In particular, targeted advertising can increase the probability of a match between a buyer and a seller. Similarly, the use of data from multiple individuals leads to significant improvements in the matching of personalized services (e.g., Google Maps) to an individual's needs.

A central issue here is that the very same information that is valuable to form a good match will typically also impact the way the surplus is distributed. In particular, an important driver of the value of information for sellers is the ability to segment the market. For example, even if a seller had just partial information about some buyers, it could either offer different varieties of the product to different subsets of those consumers, or simply charge different prices to different segments of the population on the basis of some observable characteristics. In all of these cases, it is information that allows the seller to change the terms at which he offers a trade to the consumer. Thus, information not only affects the amount of surplus generated online; it can significantly change the way in which that surplus is shared between buyers and sellers, see Bergemann, Brooks, and Morris (2015) for a general statement and result.

As a consequence, the value of information can be positive for one side of the trade, and zero, or even negative for the other side of the trade relationship. Indeed, the ability to tailor the terms of the trade clearly benefits the sellers, who can reach more customers and offer them tailored products at prices that are closer to their willingness-to-pay. In contrast, the impact of the additional data on consumer surplus is less clear cut. While perfect price discrimination is clearly harmful for the consumer, whether other forms of market segmentation are harmful or beneficial is typically an empirical matter. The theoretical argument establishes that the sign can either be positive or negative.

3 Individual versus Social Data

To understand the significance of the social aspect of the data—whereby an individual's data is also informative (i.e., possibly predictive) of the behavior of others—it is perhaps easiest to consider the counterfactual world of independent individual traits. For just a moment, let us entertain the hypothesis that the data of every individual had no meaning or content beyond this specific individual. That is, nobody could learn anything from the data of one individual about any other individual, let alone a group of individuals. In such an, admittedly hypothetical, world, two things surely would be true. First, we would see much less data gathering by the internet platforms. Second, every individual user would display heightened concerns about any data collected about him. This is because every user would be able to control the amount of information about them that can be used in any kind of transaction. In particular, if the disclosure of information had negative consequences, then users would ask for an appropriate compensation, or else simply not disclose the information.

But this is not the world we live in, and in many respects fortunately so: the traffic data gathered from one individual is informative about the road conditions for other drivers nearby; and the "likes" of an individual in a social network presumably express a sentiment held by others in the network as well. This has two important consequences for the economics of the data: any piece of individual data is simultaneously conveying information about a much larger group of users; and consequently, the data gathered from many individuals together is very informative about any specific consumer, even one who did not disclose much information himself. These two interlocking aspects of the social data render the economic externality generated by the individual data particularly significant.

4 The Value and the Price of Social Data

The business models of large internet platforms share some important structural similarities. To a first approximation, Internet companies such as Amazon, Facebook, and Google are technology platforms that facilitate matches. A match can between individuals, or between individuals and products in a broad sense.

For example, Google matches consumers who enter a search term with sellers of a related product. An individual browsing his Facebook page is presented with a newsfeed that provides him with an algorithmic update about his social network, news items, and sponsored stories.

Critical for the success of the platform is the collection of data about the matching partners. A larger data base about the characteristics of the matching partners, one that contains more information about the relevant aspect for the matches increases the quality, and possibly the quantity of matches (see Bergemann, Bonatti, and Smolin (2018)).

Amazon, Facebook, and Google monetize the value of their matching services mostly through advertising revenues. We can now begin to understand how individual information and social data are the two central elements in their business strategy. Advertisers value the information that the internet platforms collect, as it enables to tailor their advertising and pricing decisions, see Bergemann and Bonatti (2015). Each consumer anticipates that the information available about them will affect the choices (products, recommendations, prices) he is offered, and may demand compensation (e.g., through the quality of the services received) if revealing information has any negative consequences. Indeed, Google encourages the transmission of personal information through the provision of an instructive search algorithm, and Facebook does so through the establishment of a social network.

However, the consumer's choice to provide information is guided only by his private benefits and costs, i.e., the informational externality generated by the individual data he provides is not part of his decision making. It follows that the internet platform has to compensate the individual consumer only to the extent that the disclosed information changes the welfare of that individual. Conversely, the platform does not have to compensate the individual consumer for any change he causes in the welfare of others, nor for any changes in his welfare caused by information revealed by others. In consequence, the cost of acquiring the individual data can be substantially below the gain that the platform (and the advertisers) can achieve with respect to all other consumers.

The resulting difference between the possible revenue gain in the interaction with many consumers and the small compensation necessary to receive the information likely drives the extraordinary appetite of the internet platforms to gather information. We discussed earlier how data matters for the distribution of the surplus. We can now see how the informational externality creates a gap between the gains from information that accrue to the platforms and the marginal compensation received by the individuals. The presence of an informational externality also indicates that the standard argument for competitive prices to establish efficient trade does not necessarily operate in these markets.

On May 22, 2018, The Economist wrote: "Facebook seems to think it only needs to tweak its approach. In fact it, and other firms that hoover up consumer data, should assume that their entire business model is at risk. As users become better informed, the alchemy of taking their data without paying and manipulating them for profit may die. Firms may need to compensate people for their data or let them pay to use platforms ad-free." This argument has only limited applicability. Indeed, it would be entirely correct in the absence of the social aspect of the individual. But with social data and market power, the firms only have to compensate the consumer for a small part of the value of his information. Moreover, the dual role of social information means that the internet platforms already know so much about each individual, that the required compensation, is in fact very small relative to the aggregate value of the information obtained indirectly through other individuals. It should thus not surprise us that in most cases, the compensation is in form of small services, rather than monetary.

Thus, the platforms are able to acquire scores of data at a very low cost. Still, the question remains, whether the use of this data on large internet platforms is beneficial to consumers and to society?

5 The Data Markup

The internet platforms make most of their profits by collecting data from users and then using that information to match consumers with advertisers, who pay for the offered access. In the popular press and elsewhere, many have questioned whether the large size of these firms create antitrust concerns. One argument against such concerns is the fact that Google and Facebook provide valuable services (internet search, access to a social network) for free. How can a price of "free" be so high as to be of concern to the antitrust authorities?

Recent theoretical work on markets for information helps to clarify this issue, see Bergemann, Bonatti, and Gan (2019). In informational markets, large firms trade consumer services for access to the consumers' data. As we discussed above the *externality of the social data* leads to result where consumers do not receive the full value of their data. This discrepancy implies a clear definition of the markup received by data intermediaries. This markup, in dollars, is the difference between the value of the information to the firm minus the value of the services provided to the consumer. In recent work we show that this markup is increasing in the number of consumers using the services and in the number of different services offered by the firms. In equilibrium, the firms can offer relatively small benefits to the consumers that do not reflect the data economies of scale that are being exploited by the firms. That is, private benefits of increasing firms size, which are considerable, do not translate into increasing consumer benefits.

Furthermore, entry into informational markets may be very difficult. In particular, we show that small firms may fail to make money in such markets, as they receive insufficient benefits from data aggregation. Only a firm of sufficient scale will be profitable, and profits then increase in further scale.

These theoretical advances, then, place data markets into a clear antitrust context. Markups are potentially large, indicating market power, and entry is difficult, so the market power may not be naturally eroded. To be clear, though, the existing theory (by itself) does not immediately suggest a antitrust solution to the problem. This will require further analysis of the facts of competition in the industry together with the appropriate application of legal and economic analysis.

Thus there are two distinct but related concerns regarding price and markup in data markets:

1. The first one is "classic antitrust" concern and it is dynamic in nature. If markups are high because of the externality, then the platform size (as represented by the numbers of users and the number of tied-in services) creates a barrier to entry that limits innovation and reduces future benefits to consumers.

2. The second one is static. Even assuming that the market structure is accepted, then the platform-optimal use of the information fails to compensate consumers for harmful uses if those are enabled by the externality.

6 Welfare Consequences

At first sight, the externality imposed by one consumer's individual information on other consumers is positive rather than negative. Unlike the environmental externality caused by carbon emissions, where the private use of fossil fuel generates an environmental cost for society, the private provision of data generally appears to generate an informational benefit for society. A driver who shares his travelling information with a traffic app, including location, velocity, and direction, helps other drivers on the road to find the best navigation route. Similar, a consumer who allows a merchant to record his purchase history and use it to predict the next purchase of a similar customer can only improve the predictive ability of the merchant. However, we shall see that it is not the sign, but the difference between the private cost (or benefit) and the social cost (or benefit) that matters for the economic consequences of social data. Indeed, the informational externality generated by the social data can now lead to a divergence between the private value and the social value of information. Let us then explain what we mean.

We have seen earlier that many possible uses of social information benefit consumers, while some may hurt them. (For example, the quality of Google Maps helps consumers, but price-setting and disguised political advertising may hurt them). However, all uses benefit advertisers, who consequently are willing to pay for every bit of information generated on a platform. Furthermore, thanks to their market power, the platforms are able to extract most of this value from the advertisers. Therefore, the value of information for the platform is given by the incremental producer surplus associated with any use of that information.

What is the cost of acquiring information for the platform? A rational consumer would be willing to disclose private information to a platform in exchange for monetary or non-monetary compensation. As we saw earlier, the compensation is frequently offered in terms of service rather than money. The consumer may be willing to anticipate some form of allocative distortion or rent extraction that comes with the disclosure of information if the value of the services rendered is sufficiently high. If the platform had to compensate consumers for all the possible uses, it would internalize the cost to consumers of "excessive" information use, and would simply not allow the corresponding "harmful" uses.

But importantly, and at the core of the informational externality, the platform doesn't have to pay "full price" for those uses of the information. These uses are enabled in large part by the information provided by *other* consumers with correlated preferences. Each consumer is then willing to reveal their data for a fraction of the true cost to them. This motivates the platform to acquire and use excessive data, from a social perspective. Of course, this does not mean the consumer is worse off with Facebook than without; just that on the margin the platform reduces consumer and even social surplus.

The excessive incentives to collect and transmit information can therefore lead to a number of social welfare decreasing activities, that are nonetheless privately optimal to pursue and enabled by the platform. We already mentioned that third degree price discrimination, that can induce allocative distortions, is likely to emerge from abundant social data. But there are many other forms of welfare-diminishing activities that can be explained through the lens of the informational externality. For example, a consumer's engagement with a platform simultaneously generates more social data and increases the chance of a finding a revenue-generating match. Since the social data is more valuable for the platform than the individual, in the absence of regulation, the platform may create excessive incentives to the individual to spend time on the platform than is individually optimal.

7 Policy Instruments

Here we argue why several “classic” policy interventions are likely to have limited impact on welfare in markets for information, if at all.

7.1 Platform Competition

Competition among platforms alone is unlikely to bring about an efficient use of information. First, and this may be part of the solution, platforms such as Amazon and Facebook do not compete for a user’s exclusive information. Rather, different platforms compete for shares of a consumer’s time (e.g., according to CEO Reed Hastings, Netflix “actually competes with sleep”). Because the consumer’s response is then continuous in the platform’s strategic choices, a small number of non-exclusive competitors with differentiated services will not restore full efficiency.

Of course, Google tries to strengthen the consumer’s involvement with the platform by providing additional services, such as Gmail, Youtube, Google Maps. Similarly, Facebook seeks to increase consumer engagement and the consequent disclosure of information through a portfolio of services and activities offered on the platform. But this only helps create barriers to entry into some of these services. In the medium run, this reduces the number of competitors and may well reduce the competitiveness of the market for individual information.

7.2 Data Portability

According to the European Union’s GDPR, data subjects are entitled to obtain data that a data controller holds on them, and to reuse it for their own purposes. It may be argued that data portability is akin to allowing individuals to select which pieces of their information to share with specific third parties. This type of intervention does not address the informational externality. In other words, it is likely that individuals will keep disclosing their information as long as platforms keep learning most of what they need from other, similar individuals’ actions.

Data portability could however be a first and important element in introducing competition among platforms. Without it, any individual who is contemplating to leave one platform for another one, currently leaves his data on the previous platform, and starts from fresh at the new platform. With data portability, he could remove the data from the old platform and immediately transfer it to new platform. This would strengthen the bargaining position of the individual user, both vis-a-vis the established as well as the newly entering platform.

Data portability could also allow for the entry of a new platform whose business model is to bundle the interests of the data users. By directly negotiating on behalf of a large group of users whose data the new platform administers as a gatekeeper, the new platform could to some extent internalize the informational externality for the users. With data portability, the users could coordinate in establishing a new platform. Implicitly, these gatekeepers are already present. For example, when Apple negotiates with Google to use google search engine as default, Apple is in a strong bargaining position with the collective data of its users.

7.3 Taxes

A common instrument to regulate externalities is taxation, or a close substitute thereof like a cap-and-trade mechanism. The goal is simply to raise the price of an activity with negative economic externalities so to reduce its equilibrium quantity. The theoretical difficulty (enforcement and implementation issues aside) with taxation in the market for information is that taxes, by construction, have an impact on the margin. Just like a carbon tax will eliminate the least profitable (marginal) current use of carbon, a tax on the use of information will eliminate the least profitable use of individual data for platforms and advertisers.

However, in the case of environmental regulation, the externality is easily quantifiable in the amount of carbon emissions. In contrast, in the case of information, the marginally profitable use of individual data might well be beneficial to consumers! In other words, the uses of individual and social data are heterogeneous and at least two-dimensional (how much they improve platform revenues, how much they benefit consumers). Taxation impacts their equilibrium provision along the first dimension, which might well be negatively correlated with the latter (e.g., if price discrimination is highly profitable for advertisers). In that case, taxes would make matters even worse.

References

Bergemann, D., and A. Bonatti (2015): “Selling Cookies,” *American Economic Journal: Microeconomics*, 7, 259–294.

——— (2019): “Markets for Information: An Introduction,” *Annual Review of Economics*.

Bergemann, D., A. Bonatti, and T. Gan (2019): “Markets for Information,” Discussion paper, Yale University.

Bergemann, D., A. Bonatti, and A. Smolin (2018): “The Design and Price of Information,” *American Economic Review*, 108, 1–45.

Bergemann, D., B. Brooks, and S. Morris (2015): “The Limits of Price Discrimination,” *American Economic Review*, 105, 921–957.

Bergemann, D., and S. Morris (2019): “Information Design: A Unified Perspective,” *Journal of Economic Literature*, 57, 1–52.