Why Online Services Should Pay You for Your Data? The Arguments for a Human-Centric Data Economy

Nikolaos Laoutaris
IMDEA Networks Institute

***Abstract*—Data, and the economy around it, are said to be driving the fourth industrial revolution. Interestingly, the people—whose data are what moves the new economy—have a rather passive role in it as they are left outside the direct value flow that transforms raw data into huge monetary benefits. This is a consequence of a de facto understanding (or, one may say, misunderstanding) between people and companies that the former receive unpaid access to online services in exchange for the unpaid access to their personal data. This article argues in favor of an alternative *human-centric data economy* in which people will be paid whenever their data will be used by revenue-generating products and services. We discuss the benefits of such an economy, the main challenges for realizing it, and its feasibility in the view of existing technologies and business practices.**

Imagine a future in which a recommendation for a product at an e-commerce website for a hotel room at a booking website, or for a movie at a pay-per-view streaming service, would all redistribute a part of the resulting payment among the users whose previous shopping, travel, or viewing patterns were mined in order to produce the respective (successful) recommendations. The

Digital Object Identifier 10.1109/MIC.2019.2953764
Date of current version 18 December 2019.

same would apply for the services and products that, instead of individual payable transactions, accrue revenue through subscription fees in an all-you-can-eat flat or tiered manner. A car navigation system with live traffic information would, for example, redistribute part of its revenues to owners of cars, subscribers to the service or not, whose location data helped the service run an accurate and timely traffic/navigation service. We will call such an economy *Human Centric* in order to contrast it with today's data economy in which people are not compensated for their data.

Human centricity refers to the ultimate goal of upgrading the position of citizens in the data economy by allowing them to progress from the unpaid "data subjects" to the partners who participate actively in profit sharing in proportion to the value of the data that they contribute. This will be an important step toward realizing a new data-driven industrial revolution. The current practice of compensating the users *implicitly* (*in kind*) by giving them free online services is a half-measure at best that, as we will argue next, is harmful to both privacy and sustainability in the web. Fully realizing the data-driven revolution will require the creation of fair rules for providing *explicit monetary benefits* to the providers of data without whom the data-driven business models and their supporting machine learning algorithms have almost zero value.

WHY DO WE NEED A HUMAN-CENTRIC DATA ECONOMY?

Paying people for their data seems to be at least one step ahead from discussions about data in our times, which go mostly in the direction of the data protection and privacy. Still, the arguments in favor of alternative data economies are mounting fast. Next, we discuss why such economies can be beneficial for individuals, society, and even the companies that engage in data collection and monetization.

Paying for Data as an Economics Solution to Privacy Problems

Paying for data puts economic pressure on online services to apply *data minimization principles*, i.e., to collect and process only the minimum amount of data necessary for their operation. Data minimization is mentioned in the General Data Protection Regulation (GDPR) and other data protection laws but is seldom applied in practice. Indeed, since the data are by now a major asset, companies collect all they can, far and beyond their actual information needs, in the hope that future versions of their service or new business models will allow them to eventually monetize the extra amount collected. This in conjunction with the fact that the *collection and processing of data costs close to zero* are fueling greedy "all-you-can-eat" practices. Data collection costs close to zero for one part due to Moore's law that allows for the transmission, processing, and storage of data at very low cost, and for the other, because users are not compensated. Even if IT costs start deflating slower than Moore's law, it is quite certain that, in the foreseeable future, user tracking and data collection will remain cheap enough to allow the all-you-can-eat approach to persist uninhibited.

Coming to the data protection laws, the issue of applying them in practice by proactively monitoring and detecting violations remains a huge challenge.[6] If we fail to apply in practice the existing and forthcoming data protection laws, then a privacy-/trust-related tragedy of the commons on the web becomes a realistic possibility. The data minimization and sound data-driven business models can play a key role alongside data protection regulation in avoiding such a grim outcome. Ideally, online services should collect the minimum amount of data that they need and engage in the data collection, only when the benefits that they create for society outweigh, and can, therefore, recompensate for the damages they impose. In other words, in the same way that factories and private cars pay some type of tax or fine for polluting the environment, online services should pay for privacy risks and damages imposed on the people. This would institute a currently missing "economic signal" for pushing in the direction of replacing all-you-can-eat with more rational data collection practices. By having to pay for data, "parasitic" services, such as trackers that compile lists of anything from suspected alcoholics and HIV positive individuals, to active police officers (https://money.cnn.com/2013/12/18/pf/data-broker-lists/), would go out of business, whereas valuable services such as search and maps would proceed uninhibited and even benefit, as explained later. Data collection could be allowed for free as long as the total volume is small to protect the startups during their first steps. Currently, for a small unilateral profit, scores of abusive services are imposing huge privacy risks to society, while remaining economically viable largely because they can collect data almost for free.

Long-Term Sustainability

Beyond its negative effects on privacy, the current economic model around data has lead to market failures in the form of large data monopolies and oligopolies, and may even become a threat to employment in the future due to job loss from data-driven automation. Paying people for their data could, therefore, be an alternative to labor-based compensation in the future in which most work will be done by machines. Indeed, it was estimated recently that if fair remuneration algorithms are set in place, a family of four could earn up to \$20 000 per year from their data.[10] The above figure may seem small to be a full alternative to labor-based compensation but can only increase as more and more sectors become catalyzed by automation.

The idea of paying people for their data is a fairly recent one. Its origins can also be traced to the recent idea of (universal) guaranteed minimum income. The latter is discussed, and has even been tested in Finland and Switzerland, as a potential remedy for the modern societal ailments such as increased income disparity, increased unemployment, and the other labor-related challenges emerging in the context of machine learning automation, robots, three-dimensional printing, self-driving cars, and other employment-threatening technologies (see the work presented by Harari[4] for a general treatment of the above). Technology philosopher and author Jaron Lanier[5] has taken the idea of guaranteed minimum income and transformed it by stripping away the negative aspects raised by its critics, such as that it is a nonsustainable form of charity. Instead, Lanier has argued that paying people for their data is an altruism-free idea compatible with the modern capitalism for achieving the positive objectives of guaranteed minimum income without harming but instead benefiting the market, innovation, and investment in technology. The fundamental argument behind this position is a simple one: the business models and machine learning algorithms have zero value without the data and, therefore, paying for those data is not charity but rather neoclassical economics.

Win–Win for People and Online Services

Going a step further, paying for data should not be seen as harmful to business for the simple reason that the online services market is certainly not a zero-sum market—increasing the profit of the users does not have to harm the profits of online services. In fact, by providing compensations for data, online services can acquire more data, and of higher quality, than the data they collected intrusively today, and, by doing so, increase their revenues and the utility for their users. For example, instead of just collecting the product pages that the users visit in an e-commerce site, in a human-centric data economy (HCDE), the users that have opted-in to receive micropayments may further release the amount of time spent at each page as well as the local interaction patterns, such as scrolling up and down, that may not be visible through the standard cookies and other mechanisms used today. It is not surprising, therefore, that the idea of paying or being taxed for the data has been positively received by many, including the industry leaders, such as Elon Musk, Mark Zuckerberg, and Bill Gates (https://qz.com/911968/bill-gates-the-robot-that-takes-your-job-should-pay-taxes/).

CHALLENGES

Next, we review some of the fundamental challenges for realizing a human-centric data economy.

Horizontal Value Split

The first challenge is to devise a fair way to split the revenue among the main actors of the data value chain, i.e., among online services, human data providers, and any other collaborating third parties (see Figure 1). For example, in the

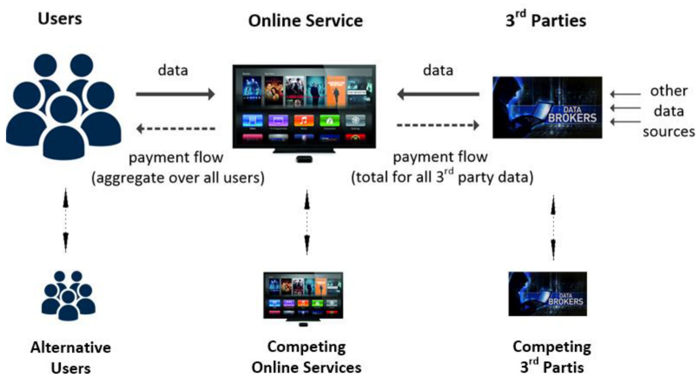

Figure 1. Data value chain with three entity types: users, online services, and third parties. Side payments flow horizontally between entity types. The bargaining power of each entity type depends, among others, on the amount of competition within its sector.

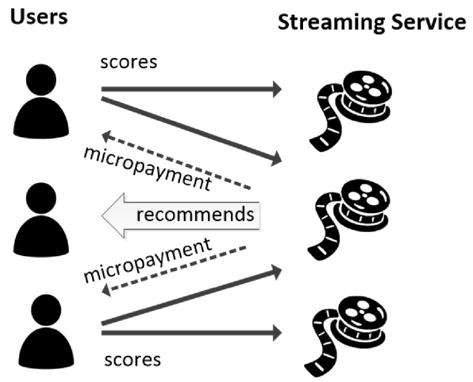

Figure 2. Micropayments received for offering movie scores used by the recommender system of an online movie streaming service. See the work presented by Paraschiv[9] for the formulation and results.

case of the successful sale of a TV from an online electronics store following a recommendation based on the purchase data from the other users of the store, one would need to decide what part of the price of the TV to return as a payback to those users. If the recommendation also used browsing and purchasing data of other users without accounts in the e-commerce site [such data can be bought online from the so-called data management platforms (DMPs)], then those users should also receive a part of the payback, as should the DMP that made their data available to the e-commerce site.

Notions from game theory, such as the Nash Bargaining and the Shapley Value,[8] can be used for defining a fair split of revenue among the main entities of the data value chain. Several modeling challenges ensue for capturing things, such as how the size of the user base impacts the ability of the recommenders to make better recommendations; how the existence of the additional data from the third-party brokers further improves the effectiveness of recommendations; how competition within the service provider or data provider segments increases or reduces the bargaining power of players; and how customer loyalty to a service or provider loyalty to a data provider impacts payments. Also, unlike most economics work in the value-chain modeling,[7] we would have to develop quantitative rather than the qualitative models, i.e., models that can be driven by real data to produce actual eurovalues for the amount of payback to be returned to the data providers for different online service types. See the work presented by Courcoubetis[2] and references therein for an example of such quantitative versus qualitative value chain modeling in the context of paid peering negotiation between Internet Service Provider (ISPs) and online service providers.

Vertical Value Split

Let us assume that the horizontal value chain modeling indicates that a percentage P, say 10%, of payable transactions revenue should be redistributed to users. The next question is how to find a way to split this P among the different individual users whose data were used in the abovementioned successful transactions, as depicted in Figure 2. In some cases, the data of different people make more or less equal contributions to a successful recommendation. That would be the case, for example, of a recommendation for a blockbuster movie produced by counting the number of people who chose that blockbuster movie versus another one airing at around the same time. In that case, most people who typically watch blockbuster movies would be assigned an almost even share of the amount to be redistributed. On the other hand, if the recommendation were for an experimental Maltese film of the 1950s, then any fans of Maltese cinema should be assigned a much higher proportion than that assigned to the mainstream viewers watching mostly blockbuster movies. Similarly, a driver who reports location data that lead to finding an uncongested route during a peak commute time should receive a higher return than the other users who report mobility data on already-known congested routes.

Notions of fairness such as the Shapley value can be used for this task as well, but they will have to be approximated by computationally efficient alternatives. Computing the Shapley value directly involves *O(N!)* computations, where N is the set of "players" participating in the value creation. In the vertical splitting mentioned before, the size of the player set N is inherently small. For example, if we consider three options for user-base size (small, medium, and large), three competing streaming providers, and two data brokers covering a region, then we will have a player set of size $|N| = 8$. For such a number, we can compute the Shapley value exactly by performing a number of computations in the order of $8! = 40\,320$. This would produce the total payoff to be returned to the users. However, to break this total payoff among the individual users who have provided data to a successful movie recommendation, we will need to consider $|N|$ is the order of thousands, if not millions. Even a mere user population of $|N| = 10\,000$ would, thus, lead to a computation in the order of factorial 10 000, which is a 35 658-digits-long number that is far beyond what our technology can deal with. Luckily, efforts are being made to simplify Shapley, in cases that the value function to be computed allows it[1] or approximate it by the fast polynomial algorithms or heuristics.[9]

Notice that the above split of value, first horizontally among entity types (users versus service versus third parties) and then vertically among one entity type (users), represents a radical deviation from the auction-based approach proposed today by most data marketplaces[11] and personal information management systems.[3] Auctions attempt to use the market forces in order to assign an *a priori* value to data, often in a context (application) unaware fashion. This is challenging, if not impossible, for a variety of reasons.

- Trading new commodities, be it precious metals, oil, or the latest cryptocurrency, is known to lead to wild price volatility for the extended amounts of time while people are trying to appreciate the real value of the new commodity.
- Calling data a commodity is more of an abuse of the term. Unlike two different liters of gasoline, which are indistinguishable (at least when coming from the same pump), the mobility or browsing pattern of different people may have radically different values for a service (think of the value for a stock recommendation service of having access to the browsing behavior of a famous investor versus an ordinary person).
- Even two identically valuable pieces of information can generate vastly different benefits for two information buyers. This is because, unlike a liter of gasoline that can be consumed only once, a mobility or browsing pattern can be "mined" (processed) infinitely without losing its value or decaying. A fundamental shortcoming of auctions is that they cannot know in advance, or limit in postauction time, the number of times that a piece of information will be reused. This makes deciding, or accepting a bid, very difficult.

In contrast to the auction-based valuation of the data in which data are sold *a priori* as a commodity *without a specific context and use case*, in our proposal, payments are made *a posteriori* for a *particular use case* under a *clear revenue model*.

Implementation

Laying the foundations of the new economy and dealing with the scalability challenges in payoff computation are only the tip of the iceberg on the road toward making a human-centric data economy a reality. For HCDE to have a realistic path toward the adoption, establishing *trust* is a key. Users need to be able to trust that payments in exchange for their data are fair and in accordance with the payment schemes offered by different services. To achieve this, payments need to be *transparent* and *attestable*. Transparency and attestability mean that the users need to have a way to understand how much and why they were paid for their data; moreover, they need to be able to do that remotely and in a trustworthy manner that guarantees that the service provider has faithfully followed the advertised compensation plans. Compensation plans can range from very simple ones, such as paying in proportion to the number of data

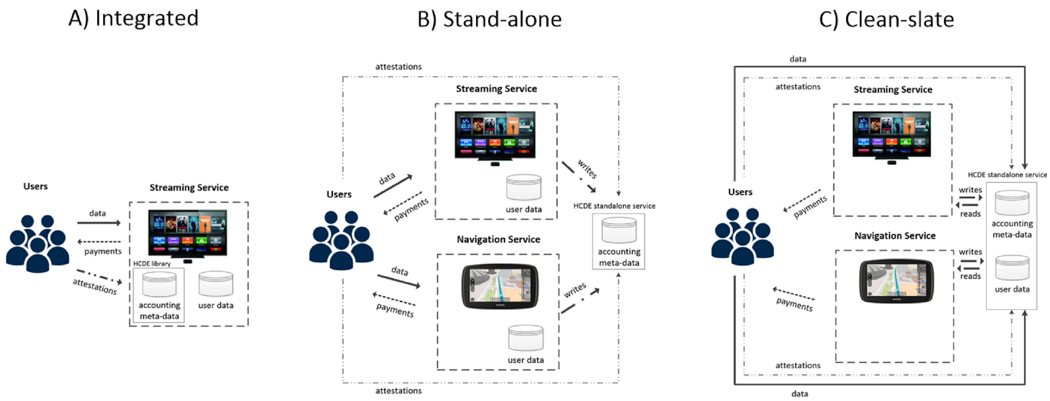

Figure 3. Alternative architectures for storing user data as well as accounting metadata for computing micropayments. In an integrated case, accounting metadata is integrated with existing services. In the standalone version, a separate service exists for storing accounting metadata across online services. In the clean-slate case, the standalone service also stores the user data to facilitate user control and improved privacy.

points provided, to more complex ones, such as paying in accordance with the actual importance of the data provided in the context of a specific computation, e.g., a recommendation. A simple way to achieve these objectives is to allow the users to see the actual data used in all the transactions carried out by a service and, thus, verify for themselves that payments have been fair. Obviously, such an approach is unrealistic. A service has many reasons and constraints that make opening up its data impractical, even illegal, in some cases, because—beyond jeopardizing its intellectual property and competitiveness in its market—it can also harm the privacy of its users. (Note here that when we write about opening up data, we are not referring to allowing a user to see and even retract his/her data record. We are referring to allowing a user to see the data of other users and thus verify that he/she was compensated a fair amount.)

Going beyond transparency, attestability, and privacy challenges, we still face open important architectural, systems and interface definition problems that are crucial if one wants to produce a technology that can be implemented and adopted in practice. For example, the accounting metadata layer needed for the micropayment computation can be integrated within each particular service, or alternatively can exist as a standalone external service upon which different services will write and store accounting metadata, as depicted in Figure 3. In both cases, but especially in the second, additional architectural and design questions need to be answered, such as how will the users be identified? what communication interfaces will be used for communication among users, services, and third parties? where will the actual user data be stored?

FEASIBILITY

Achieving the final objective of HCDE, which is nothing other than disruption for the benefit of everyone—first and foremost the people but also the companies and the way that the data economy works–is a high risk, very high return bet. Clearly, designing and prototyping a fully working version of the necessary technology is something very difficult but, in our opinion, feasible. Economics provides the starting point by offering rigorous tools for defining value sharing among the groups of collaborating entities. Computer science offers tools for taking complex metrics and approximating them with fast but accurate heuristics. In terms of engineering feasibility, the proliferation of big data platforms and optimized machine learning implementations provide the tools needed to implement the necessary computations for deriving the data-related micropayments. The blockchain and trusted execution environments provide the tools for storing and accumulating such micropayments during a payment period (day, week, month) before actually transferring them to human data providers. Finally, the data protection regulations such as GDPR provide the framework and point toward the direction of making more rational use of data, which is what the micropayments can achieve using market, instead of regulatory pressure.

Last but not the least, even if it is not happening yet, it is certainly feasible to bootstrap the shift toward a new trend of paying the users for their data. All that this shift requires is a small set of visionary online services that will realize the benefits of the approach (smooth the privacy versus utility tussles, induce users to share more data, etc.) and use it as a differentiator and a competitive advantage over their competitors. If successful, they can convince more companies to adopt this practice and eventually make it a commonplace.

REFERENCES

1. S. Cabello and T. M. Chan, "Computing Shapley values in the plane," 2018. [Online]. Available: https://arxiv.org/abs/1804.03894
2. C. Courcoubetis, L. Gyarmati, N. Laoutaris, P. Rodriguez, and K. Sdrolias, "Negotiating premium peering prices: A quantitative model with applications," *ACM Trans. Internet Technol.,* vol. 23, no. 50, 2016, Art. no. 14.
3. Digital Single Market, "An emerging offer of "personal information management services"–Current state of service offers and challenges", Nov. 23, 2016. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/emerging-offer-personal-information-management-services-current-state-service-offers-and
4. Y. N. Harari, *Homo Deus: A Brief History of Tomorrow.* New York, NY, USA: Harper, 2017.
5. J. Lanier, *Who Owns the Future.* New York, NY, USA: Simon and Schuster, 2013.
6. N. Laoutaris, "Data transparency: Concerns and prospects," *Proc. IEEE*, vol. 106, no. 11, pp. 1867–1871, Nov. 2018.
7. R. T. B. Ma, D.-M. Chiu, J. C.-S. Lui, V. Misra, and D. Rubenstein, "Internet economics: The use of Shapley value for ISP settlement," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 775–787, Jun. 2010.
8. M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. Cambridge, MA, USA: MIT Press, 1994.
9. M. A. Paraschiv and N. Laoutaris, "Valuating user data in a human-centric data economy," 2019. [Online]. Available: https://arxiv.org/abs/1909.01137
10. E. A. Posner and E. G. Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton, NJ, USA: Princeton Univ. Press, 2018.
11. F. Stahl, F. Schomm, and G. Vossen, "The data marketplace survey revisited," Eur. Res. Center Inf. Syst., Working article, Münster, Germany, 2014.
12. K. Zhao, S. H. Mahboobi, and S. R. Bagheri, "Shapley value methods for attribution modeling in online advertising," 2018. [Online]. Available: https://arxiv.org/abs/1804.05327

Nikolaos Laoutaris received the Ph.D. degree in computer science from the University of Athens, Athens, Greece. He has been a Research Professor with IMDEA Networks Institute, Madrid, since December 2018. Prior to that, he was the Director of data science with Eurecat and the Chief Scientist of the Data Transparency Lab which he cofounded in 2014 during his 10 year tenure as a Researcher in Telefonica. Before that, he was a Postdoctoral Fellow with Harvard University and a Marie Curie Postdoctoral Fellow with Boston University. His main research interests include privacy, transparency, data protection, economics of networks and information, intelligent transportation, distributed systems, protocols, and network measurements. Contact him at nikolaos.laoutaris@imdea.org.