DATA AS LABOUR AND FIDUCIARIES

JONAS KGOMO

Date: January 7, 2022.
Key words and phrases. differential privacy, data as labour, passive work, MID, data divident
.

Abstract. We study the problem of passive data work that is prevalent on the internet, we look at the data dignity and how to properly represent individuals data. To this end, protocols and ideas have been presented to creating a user data-centric economy. We present the solutions to the current data conundrum that can be improved with data policies and a better semantic web. Alongside the research we also produce a working project that was developed to showcase the accompanying theories. We will look at how data can be considered to be a labour, how that value can be distributed. We also offer methods of dealing with privacy and data collection frameworks, we develop a two methods: CUCI(Comparative Universal Cookie Identifiers) and the Privacy Spectrum. We then provide a framework of how to create such Data Unions and Fiduciaries. We conclude that such changes will require both governmental and co-operate regulations to take place.

1. Introduction: Data As Labour

In 2018, Jaron Lanier and Glen Weyl wrote a manifesto on the case of *data dignity*, although not a novel concept, they presented a formalism to the already prevalent ideology. The Radical Market thesis, looked to conjure the history of labour movements and workers unions in the technological sphere and help create collective decentralised organisations. However the research overlooked the nature of data – in theory it is not only personal but interpersonal, and must be assumed as a social element, and the complexity accompanying it. In our framework we look at incorporating this aspect of data, we look at addressing the *monopsony* on data. There has also been a sound discussion on the topic of treating data as labour:

– *Should We Treat Data as Labor?*
– *How can we price private data?*

Jaron and Glen [4] asked the first question, and proposed a methodology of how we can treat data as labour through a so-called radical market. Chao and Daniel Li [5], posed the latter question and provided a framework of price theoretic allocation of value to noisy query answers. In our research, we will look at how to design a framework for data as labour and how to organise polity for representing that rights and labour laws.

1.1. **Data Dignity.**

Data dignity is the right of individuals to control their personal data and to be informed about how it is used. We are aware of how our location data on mobile phones can be used to create traffic flow, for instance with Google Maps. However there is no financial incentives for users in providing the data that makes this service improve. The privacy overriding effect of the data harvesting campaign deny the user the dignity and choice to give permission. Moreover, Acquisti showed how we can also measure the value that people place on privacy, [15].

1.2. **Labour Share.**

The global decline of the Labour Share[3], which is critical, is apparent and leads further disparities in how much of productivity growth goes to workers in the corporate sector. In her dissertation, Emile Durkheim [12] purported that the division of labour states that people are allocated

With our current imagination of data as labour, we theorise that division of labour is such that the participants may not have to specialise and separation of tasks is agnostic of the individual, everyone's data is equally valuable.

1.3. **Database Entries.** *A database is a collection of real-valued data items* $x = (x_1, x_2 \cdots x_n)$ *that form a vector*

Every data item x_i is some personal data that belongs to an individual. The price of each data point can be confined to the following factors:

(1) a query is made for specific micro-task
(2) privacy loss, leakage of some information ϵ_i
(3) compensation of each data for the loss of privacy

The payments are made using micro-finance, established by Muhammad Yunus [6], through micro-payments.

1.4. **Privacy Loss.** : Theory of Pricing Private Data, (Chao Li , 2012)[5]

Suppose we have a deterministic mechanism δ *(for every instance* x*, and* $\delta(x)$ *is a random variable on the data base), then* $\epsilon_i(\delta) = 0$ *if* δ *is independent of the data input* x_i*.*

If the asking price is set very high, then the privacy loss diminishes to zero. That means, in a privacy preserving mechanism, zero information loss occurs when the mechanism.

1.5. **Differential privacy.** Differential privacy shares data about clusters but not individual data, that is data is perturbed to hide certain properties of the data. We can look into other ways like homomorphic encryption and decentralised learning for masking producers' data.

Remark 1.1. It is not easy to formulate the pricing value from the market maker using price theory (Aperjis and Huberman, 2020), [7] proposed a way to offer different options of pricing for the valuation of data query. In this paper, we propose an option of making the price of the data, π dependent on the economic output, this is variable and similar to a stock value.

$$\pi = \frac{EG + KD}{K^2} \tag{1.1}$$

P is the value of the data, E is the company's earnings on that data, G is the companies growth rate and K the risk of privacy loss and D the companies compensation. This will ensure that the data extracted is always at par with the value created.

2. Passive AND Active Data Work

Economists such as Krutel and Ohanian [2], have suggested the importance of the impact different types of skill have on the wage of the workers. The dichotomy of unskilled vs skilled labour imply a different *skill-premium (the wage of skilled relative to unskilled labour)*. In our study we look at how a new class of passive labour diminishes the duality of unskilled vs skilled labour, resulting in a mutually inclusive subset of workers. We understand *data* to include any digital activity such as entertainment data, surveillance or biological sensory data.

2.1. **Passive Work.**

Definition 2.1. Passive Work consists of any activity or non-activity that result in aggregation of data without the user's knowledge or intention.

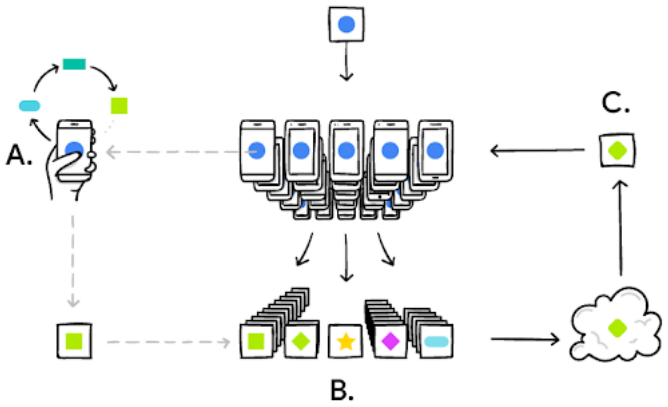
Remark 2.2. *Work here is following the physics definition, that is, work performed by a system is the energy transferred by the system to its surroundings. For instance, wearing a smart-watch that measures your body temperature is passive work.*

Definition 2.3. *Active work consists of any activity that result in aggregation of data with the consciousness and knowledge of the user*

2.2. **Analogues of Active Work.**

Mechanism like mining bitcoin directly pays you for your resources, you compute a hash and perform an algorithmic computation to prove a specific mathematical logic, and get rewarded as a miner. A miner is conscious of their resources and allocation, they are performing *active work* , however if a mining process is performed on a browser of an unassuming user, then its passive work. Another example is health care providers who pay clients for medical trials to improve medication and experimentation which can have varied privacy risks (Latanya Sweeney, 2020) [14].

A non-exhaustive list of active data-work is platforms Amazon Mechanical Turk, Mining Bitcoin and Factory Farms. These platforms generally do not effectively reward those with the greatest competence and they do not address those who are producing the data. There is difficulty in collecting data and rewarding individuals for the work they do. GDPR and other data laws forbid collection and use of data in different places, this leads to data fragmentation and isolation. However, with Federated Learning[1], the concept of the active data-work is more feasible , unlike standard Machine Learning approaches that have a centralised training centre



A model showing how federated learning can assist passive data work by allowing mobile phones to process fragmented machine learning models

3. MIDS: Mediators of Individual Data

3.1. **Mediation.** A MID is an organization with intermediate size and structure that represent the data producer and mediates the transfer of data between third parties. MIDs are created with similar features to commons, as proposed first by Garrett in *the tragedy of the commons*, (Garett, 1968) [9] and strongly supported by Yochai [10], in his study of how it is evident that *information is a public good* in an economic sense and can be treated as so called information commons. We build on this idea that *information is costly to produce but cheap to reproduce* [11] to elicit that mediation will be fundamentally looking at data as property and assign economic value to it.

Some of the features of a MID are :

- perform accounting and legal duties
- payment of data royalities
- allocate data servers and permissions
- represent individuals on a per-case scenario
- promote standards for data producers

A MID should be open-source and transparent in nature. The composition of labour unions and State intervention, increases wages and also positively affects human capital in significant ways (Beatrice and Sidney Webb, 2004) [18].

4. Data Divident

The Berggruen Institute [12] announced a plan to request multinational data-centric companies to pay for the users' personal information. The initiative is to be designed for the California's state. The article produces payment and mediator structure and artifacts. The MID structure implemented in this policy brief is called Data Relations Board, it is also a public-private cooperation and the Public Data Trusts , which are state administered data banks that collect public data and make it commercially available, either for free or for a fee, to any qualified firm. However in the study, the incentives are given towards firms to behave in a more prosocial manner. Our study looks at rewarding users for their contribution to public goods.

The user contributes data to public goods and private services, the MID then allocates relevant data infrastructure and laws apparent to the user, the firm using the data then returns revenue to the user.

In this report we look at both creating a MID based on both technology standards and traditional legal methods. The technological side handles subscriptions and user data while the legal side handles taxing and terms and conditions applying to each service. The legal side also handles the right of the data provider – we also note that this can be automated as well.

5. SOLID: Semantic Web

SOLID (Social Linked Data), is a set of protocols and conventions for a decentralised social applications that can be processed by machines, a Project proposed by Tim Berners-Lee, the inventor of the World Wide Web protocol. The project offers true data ownership (by deciding where their data resides), re-using data and ability to switching data. The foundation of the HTTP Web, is that nodes on the network are hyperlinks , while the Semantic Web uses data as nodes using the Resource Description Framework. Using these properties of the semantic web and desired outcomes of a MID, we can therefore create applications that respect *"data autonomy" (the amount of freedom a user has on their data).*

Semantic Web	Traditional Web
Decentralised	Centralised
Pods for Storage	Centralised Servers
Permissioned	No Permissions
Linked Data	Unstructured Data
Web Of Data	Web Of Links

The table shows the difference between the tradition web versus the semantic web, the semantic web bends more towards the user's autonomy on where to host their data and who should access it. Regular web applications do not afford such flexibility and comes at a technical cost to engineer.

6. Data Fiduciaries & Unions

6.1. **Data Protection vs Data Dignity.** The European Union's General Data Protection Regulation(GDPR) has many restrictions on the use of anonymous data versus personal data . Article 15 GDPR, states that users have *right of access* of the personal data an organisation has collected. The latest GDPR standards applied to pseudonymous data imply that data without identifiers, yet de-anonymisation techniques have proved that it is possible to re-identify, (Luc Rocher) [13] these sparse datasets. This means that MIDs will be able to report back to the user about the data collected for each user.

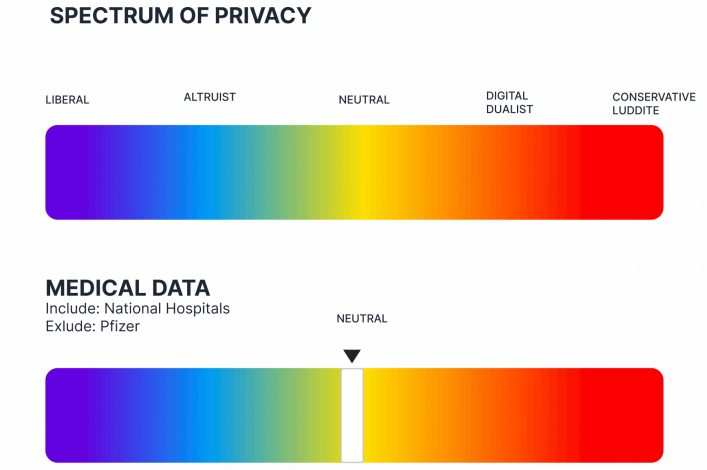
6.2. **Cookies: Data Sirens.** Cookies are small blocks of data that resides on a user's computer while a user browses the internet, these are essential to enable service providers to log stateful information like browsing an online store and user's activity. Cookies have become so prevalent that they deter the overall user-experience of the web. European law requires websites targeting members to acquire permission from user, however these have become obfuscated as the mechanism for accepting or denying cookies is masked with long legal compliance documents that one cannot possibly finish just to access a web page for a few seconds. The French legislation, CNIL recently fined a combined 238 million euros Google and Meta Platforms, for not offering a button enabling the internet user to easily refuse all cookies versus several clicks.

7. CUCI: Comparative Universal Cookie Identifiers

To solve these problems related with cookie collection we have designed a framework called the CUCI(Comparative Universal Cookie Identifiers), to allow users to accept tracking cookies through a MID, since cookies lack integrity [16] . The motivation to accept or reject cookies is influenced by the technical and legal knowledge a user has (Joanna, 2021)[17], in our method we relegate the knowledge to experts who in the MID organisation. In this section, we present a method for specifying how much digital data one wants to offer to an internet service. The user will specify to their MID what level of detail they want to share, what aspects of their data do they want to share and with which organisations.

8. The Privacy Spectrum

Our first proposal is a practical privacy moderator. Suppose we have a user browsing a website, they only need to specify their data libertarianism , the MID will then reviews the users preferences with experts and legal representatives then regulates the users 'taste' in privacy over time based on frequency of visits and level of autonomy they accept, meaning the user **accepts only once** and a machine learning model applies to all other website the user will visit in the future.



An instance of the Privacy Spectrum, where a user can adjust their benevolence on which data to share with which party. The MID will be the mediator responsible for allocating data rights.

9. Conclusion

In this paper, we presented a method for managing cookies and creating organisations to represent individuals with their data and created a framework for data as labour. To this end, we have discovered the mechanics of such organisations but also the macroeconomics. related with it. We also developed frameworks for universal cookie moderators, the framework can be further developed to have technical motivations and applications. We can research how the algorithm for creating robust user-specific data autonoy and making a browser fingerprint-free algorithm. In principle, the foundations of the idea is based on economics and business, the methods presented can be further researched in social sciences, and computer science to understand the bigger scope of data as labour.

9.1. **Outlook on further research.** Throughout this paper, we assumed for the process of data as simple point, however the reality of data as collected is complex and introduces other factors like noise. Clearly, the idea of MIDs can be created with software alone, however a more legal and governmental entity can regulate this process faster and more reliably. Further research can build on how to form a hybrid of both.

References

1. P. Billingsley, *Federated learning:*, Collaborative machine learning without centralized training data, Brendan McMahan & Daniel Ramage, Google, 2017.
2. Krusell, Ohanian, Rios-Rull, and Violante *Capital-Skill Complementarity*, Econometrica, Volume **68**
3. Loukas Karabarbounis, Brent Neiman, *The Global Decline of the Labour Share*, The Quarterly Journal of Economics
4. Diego Hernandez, Jaron Lanier, E. Glen Weyl *Should We Treat Data as Labor?*, American Economic Association Papers Proceedings
5. Chao Li, Daniel Yang Li, Gerome Miklau, Dan Suciu, *A Theory of Pricing Private Data*, Statisical Databases
6. Muhammad Yunus, *Banker to the poor : micro-lending*, 2007
7. Christina Aperjis, Bernardo Huberman , *A market for unbiased private data: Paying individuals according to their privacy attitudes*, Hewlett Packard Laboratories
8. Data Relations Board *A Data Dividend that Works*
9. Hardin, Garrett , *The Tragedy of the Commons* Science. 162 (3859)
10. Yochai Benkler, *Towards a Political Economy of Information*, 52 Duke Law Journal 1245-1276 (2003)
11. Cal Shapiro and Hal R.Varian , *Information Rules*, p21
12. Durkheim , Emile, *The Division of Labour*, New York: Free Press
13. Luc Rocher, Julien M Hendrickx, and Yves-Alexandre de Montjoye , *Estimating the success of re-identifications in incomplete datasets using generative models*
14. Boronow, Katherine E., Laura J. Perovich, Latanya Sweeney, Ji Su Yoo, Ruthann A. Rudel, Phil Brown, and Julia Green Brody. *"Privacy Risks of Sharing Data from Environmental Health Studies."* , Environmental Health Perspectives 128.1 (January 2020).
15. Alessandro Acquisti, Leslie John, George Loewenstein, *What is privacy worth?* , The Journal of Legal Studies, Vol. 42, No. 2 (June 2013), pp. 249-274
16. Xiaofeng Zheng, Jian Jiang, Jinjin Liang, Haixin Duan, Shuo Chen, Tao Wan, and Nicholas Weaver *Cookie Lack Integrity*, 24th USENIX Security Symposium
17. Joanna Strycharz, Edith Smit , Natali Helberger, Guda van Noort, *No to cookies*, 2021, Computers in Human Behaviour
18. Guglielmo Forges Davanzati and Andrea Pacella, *SIDNEY AND BEATRICE WEBB: Towards an ethical foundation of the operation of the labour market*, Vol. 12, No. 3 (2004), pp. 25-49 (25 pages)