Data Co-Ops Workshop

Executive Summary of a December 22, 2019 Workshop Hosted at the Hebrew University of Jerusalem

Organized by:

Katrina Ligett, Associate Professor of Computer Science, Hebrew University of Jerusalem

Kobbi Nissim, Professor of Computer Science, Georgetown University

Yuval Shany, Professor of Law, Hebrew University of Jerusalem

Participants:

Amnon Reichman, University of Haifa

Amy O'Hara, Georgetown University

Allison Whitmer, Georgetown University

Ayelet Gordon-Tapiero, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Benny Pinkas, Bar Ilan University

Bo Waggoner, University of Colorado

Daniel Goroff, Alfred P. Sloan Foundation

David Levi Faur, The Hebrew University of Jerusalem

Dima Epstein, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Efrat Daskal, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Elizabeth Edenberg, Georgetown University

Fabiana Di Porto, University of Salento and The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Frauke Kreuter, The University of Maryland

Gal Yona, Weizmann Institute

Jules Polonetsky, Future of Privacy Forum

Katrina Ligett, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Kobbi Nissim, Georgetown University

Limor Shmerling Magazanik, IsraelTech Policy Institute

Mark Hanin, Georgetown University

Michele Loi, University of Zurich

Nicolo Zingales, University of Sussex

Nimrod Karin, New York University and The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Noam Nisan, The Hebrew University of Jerusalem

Omer Tene, College of Management

Oz Levy, Israel Ministry of Health

Ron Hermon, Israel Ministry of Health

Ron Kupfer, The Hebrew University of Jerusalem

Rotem Medzini, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Roy Cohen, Israel Ministry of Health

Sharon Bassan, DePaul University

Tamar Berenblum, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Tomer Shadmy, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Yuval Shany, The Federmann Cyber Security Research Center, The Hebrew University of Jerusalem

Agenda

09:00-09:30	**Opening Remarks** Yuval Shany, Hebrew University of Jerusalem
09:30-10:15	**Introduction to Co-ops (Presentation)** Katrina Ligett, Hebrew University of Jerusalem Kobbi Nissim, Georgetown University
10:15-10:45	Coffee Break
10:45-11:15	**The Rise of Infomediaries: A Matter of Trust and Joint Responsibility** Nicolo Zingales (Presentation)
11:15-11:45	**Fairness in Statistical Data Analysis** Gal Yona (Presentation)
11:45-12:15	Discussion
12:15-13:30	Lunch
13:30-14:00	**Zero Knowledge Proofs** Aviv Zohar (Presentation)
14:00-14:30	**Data Cooperatives and the Wide Dispersal of Data Power** Michele Loi (Presentation)
14:30-15:00	Discussion
15:00-15:30	Coffee Break – Remote Presentation by Paul Romer
15:30-17:00	Working Groups
17:00-17:30	Report Back

17:30 Candle Lighting

Katrina Ligett – Introduction to Co-Ops

Technology has made it possible to measure and digitize almost every aspect of our existence: behavioral, social, financial, medical and much more. The data market is already a multi-million-dollar business, and claims have even been made that data has replaced oil at the world's most valuable commodity. Personal data is central to advertising, sales and strategic decision making, data brokers collect and sell our data, and private companies extract value from it. Our collective data has enormous potential to benefit society, but unfortunately, this potential is not being realized.

Additionally, in practice, individuals have very little knowledge or control regarding who gets which of their data and for what purpose. They don't get remunerated for it and don't know what data is being collected, or what is done with it. This raises concerns not only for individuals but for society as a whole – a small number of unelected people, who are not accountable to the public, have the ability and power to control this valuable resource, and to use it to open society to a wide range of harms and manipulations. This is disturbing from various perspectives, including freedom of expression, people's ability to freely make choices and decisions, and national security.

These considerations lead us to believe that it is time to rethink the data ecosystem. Luckily, technological advancements from the past decade or so can assist in doing so. Advancements in a wide range of decentralization technologies suggest a guiding principle for such a revision: we could eliminate the need for individuals to share their information for any form of centralized data gathering or centralized computation.

Decentralization on its own does not resolve all concerns with the current data ecosystem. For example, decentralization does not eliminate the ability to manipulate individuals. Decentralizing a computation also does not automatically give individuals better negotiating power. Decentralization additionally does not automatically increase

the societally beneficial uses of data. However, decentralization technologies, along with other technological and legal advances, such as differential privacy, homomorphic encryption, and modern data protection legislation (a la GDPR) together invite us to build a new and better data reality.

We propose data co-ops as a solution to many of the concerns raised with regard to the status quo data ecosystem. The idea of a data co-op is to create a new layer that sits between individuals and platforms who wish to use their data. Individuals would choose to join a co-op, and the co-op would potentially provide both technical and legal interventions involved in collecting data, computing over it, and negotiating the terms and conditions of its use. The co-op would thus be positioned to help preserve privacy and security, and to redistribute value back to the co-op members.

Several principles would guide the development of co-ops:

1. Security and privacy first.
    a. Moving away from a binary model of data access.
    b. Modeling data usage in a concrete and rigorous manner.
2. Individual control. Individuals being able to make informed choices and rely on delegated decision-making.
3. Value creation. This model will succeed only if people want to join it and use it.
4. Collective governance. Each individual has no power against any of the platforms, but together we have power in our collective data. The co-op will have to collectively decide what data is collected and used, for what purposes, under what conditions, and for whose benefit.

In this context, many questions arise, which require further academic research. A few example questions include:

1. What should be the design principles guiding data co-ops?
2. How should decisions be made within data co-ops?
3. What are the new risks created by data co-ops?

4. In the past, others have tried creating data co-ops or similar entities – what was it that caused them to fail, and how could the present initiative overcome those difficulties?

Nicolo Zingales – Infomediaries' Second Chance: a Matter of Trust and Responsibility

With the data economy, personal information has become a valuable asset not only for businesses offering customized goods and services or trading consumer data to third parties, but also and increasingly, for consumers themselves. Consumer awareness about the value of personal information is on the rise, as various options to trade, compare and extract benefits from that information become available. Regulators around the world, but particularly in the EU, have taken notice of that and have been tinkering with existing rules, not only in data protection law but also in the context of consumer protection and contract law, to enhance the protection of individuals who take part in such value exchange. The mere existence of additional tools, however, does not necessarily translate into a general empowerment of data subjects in their daily transactions: in the absence of procedural mechanisms facilitating the exercise of individual rights, dispersed and unsophisticated consumers are unable to make the most of these opportunities, leaving a significant gap between law on the books and law on the ground. In this light, the GDPR is to be welcomed for its introduction of the principle of accountability and the possibility to delegate to a non-profit body, organization or association the right to lodge complaints with the relevant supervisory authority, as well as representation before courts. In addition to these procedural facilitations, the GDPR empowers data subjects through new substantive provisions, such as the right to data portability and the so-called right to explanation. Although the breadth of these provisions will need to be fleshed out through interpretation, the incentives appear to be lined up for the rise of intermediary entities specializing in the management of personal data and the enforcement of data subject rights. This creates the conditions for a perfect storm in data protection law, pitting data-driven businesses against this new powerful force of "infomediaries", who are likely to alter the

competitive dynamics in the online ecosystem by providing a centralized avenue for personal information management and collective empowerment. This talk was part of a larger project aimed to identify challenges and opportunities associated with the reliance on these third parties for the exercise of data protection rights, and focuses on the roles and responsibilities of these infomediaries with regard to data processing. Different models will be examined to assess the different responsibilities they involve, and identify key governance and policy challenges.

Since the 1990s, when the notion of infomediaries was introduced in order to allow individuals to gain value from their data by controlling it, the relevant technical and legal frameworks have changed, resulting in an environment that may be more conducive to the development and success of co-ops.

The GDPR may usher in a new wave of infomediaries, since it creates rights to access, explanation and portability, along with the possibility to delegate to a non-profit body, organization or association, the right to lodge complaints with the relevant supervisory authority, as well as representation before courts.

Formally, the GDPR gives individuals the right to have their information transported to another infomediary in a commonly used, machine-readable format. This still leaves several unresolved issues:

1. The GDPR applies only to information collected on the basis of individual consent, but much data is collected in other contexts.
2. It only concerns raw data provided by an individual. Individuals may not have rights vis-à-vis information of theirs that was computed on by platforms.
3. The transfer of information can only be done inasmuch as it doesn't affect the rights and freedoms of third parties, and this includes the data controller's rights to trade secrets.
4. The relational character of the data could serve as a basis for data controller's refusing to disclose information. This renders information received from social media platforms almost worthless.

5. What does 'commonly used machine-readable' mean? This depends on who you ask.
6. What does the right for information in Article 22 entail?
7. The GDPR creates exemptions for the fulfillment of the aforementioned rights, which makes it even more difficult to fully understand and implement.

The GDPR also creates enforcement mechanisms such as substantial fines for violations, the creation of data protection officers and perhaps most importantly, data subjects' right to representation when interacting with a data protection authority. This enables intermediaries to accompany the data subject through all stages – asking for information, for portability, requesting an explanation if there is a problem, and finally representation before a data protection authority and even before courts.

Some questions that require attention:

1. Who shoulders the cost of infomediaries? How can we be sure they do not abuse their position?
2. Who sets the rules by which an infomediary operates?
3. Will infomediaries further process data, thus potentially enhancing its value?
4. What anti-trust issues arise in this context?
5. What type of data portability requirements do we think data co-ops should have? What data should be subject to portability requirements?

Gal Yona – Algorithmic Fairness

Statistical analysis of data is increasingly used to drive predictions and inform consequential decisions about individuals; examples range from estimating a felon's recidivism risk to determining whether a patient is a good candidate for a medical treatment. There is, however, a growing concern that these tools may inadvertently (or not) discriminate against individuals or groups. This talk provided a non-technical introduction to recent work in the field of computer science on defining and preventing discrimination in machine learning contexts.

The prevalence of algorithmic decision-making raises several questions, among them:

1. What are the sources of unfairness that exists in a system?
2. How do we define fairness?
3. How can we mitigate unfairness in the system?

In order to make sure an algorithm achieves group fairness we must first ask for whom are we trying to achieve fairness? Often the answer is for groups of individuals protected by law or ethics. One must specify the attributes we wish to protect---for example, gender, or race. We say that a model fulfills group notions of fairness if the chance that a member of the minority group will be harmed is about the same as the chance that a member of the majority group will be harmed, for various formal definitions of "harm" (one example of such a harm is being rejected from a job for which the applicant was qualified).

However, even if group fairness is achieved, this does not necessarily mean that a system is fair. For example, even if a system appears to treat women similarly to men and blacks similarly to whites, that does not mean that it will treat black women similarly to white men. In order to expose unfairness, we may need to consider complex layers of overlapping subgroups.

Since an algorithm must make future predictions based on learning from past data, sometimes the algorithm will naturally perpetuate past injustices that are reflected in that data. For example, suppose black people from a certain area have a lower than average chance of receiving a bank loan, because a model predicts they have a lower than average chance of repaying these loans. This then means that black people from this area will have less opportunity to build up their credit scores, and perpetuates a cycle of unequal access to credit.

One role for a data co-op might be in auditing and ensuring that data-driven recommendations, scores, and tailored services offered to its members obey certain standards of algorithmic fairness.

Aviv Zohar – Zero Knowledge Proofs and Blockchain

Zero Knowledge Proofs are on the bleeding edge of computer science research. They act as a double-edged sword: on the one hand, allowing modern cryptocurrencies to impenetrably mask transactions, but on the other hand opening the door to new forms of collaboration between companies, individuals and governments. This talk introduced what zero-knowledge proofs are, and how they can be used to build systems that are both private and resilient to abuse.

Blockchain technologies are a tool to reliably duplicate data between many computers, allowing people around the world to reach the same conclusions. Zero knowledge proofs allow the sender to convince the receiver that a certain property holds, without needing to reveal additional information.

Toy Example of a Zero Knowledge Proof:

Peter says he has solved a Sudoku problem that Sara wants, but he will only reveal the answer to her after she pays him. Sara wants to make sure that Peter really does know the answer before she transfers the money. In order to prove that he knows the answers, Peter will complete the Sudoku problem (arranging the tiles face-down so the numbers 1-9 each appear exactly once in each row, column, and 3x3 square) and allow Sara to randomly pick whether she wants to be shown the rows, the columns, or the squares. Say Sara chooses the rows. Peter then collects the face-down tiles from each row and puts them in a bag. If Peter did indeed know the solution, each bag should contain the numbers 1-9. If Sara checks all bags and sees that they all contain the numbers 1-9, she has not learned the solution, but she has gained some confidence that Peter did indeed know the solution. This exercise can be repeated to increase her confidence in Peter's knowledge.

More generally, zero knowledge proofs give us 3 properties:

1. Completeness. If Peter knows the solution, he will always pass the test.
2. Soundness. If Peter tries to cheat, there is at least one row or box or column that is wrong, and therefore Sara has a chance of catching him.

3. Zero knowledge. Sara learns nothing about the solution, except that Peter knows it.

Zero knowledge proofs also can provide desirable properties for a cryptocurrency system, allowing money to be transferred without publicly revealing how much money each person has, or who transfers what to whom, but ensuring that all transactions are valid---for example, no one can spend the same money twice.

Similarly, zero knowledge proofs could be useful in establishing credit-worthiness. Suppose Bob wants to borrow money from Alice, but before agreeing, Alice wants to know whether Bob will pay her back. One cannot always trust Bob's promise to pay back, but Bob does not want to disclose all his financial data to every potential lender. In the offline world, various institutions have created credit scoring systems, wherein a third party has access to Bob's financial data, and the third party potentially tells Alice how likely Bob is to repay the loan without revealing further financial data. This can also be done online, but we are concerned about data leakages and privacy. So now, when Bob approaches Alice online and asks for a loan, perhaps he will not give her a credit score to prove his creditworthiness, but perhaps he will give her zero knowledge proof of his credit score, and perhaps that score will be established without need for any centralized, trusted third party to gain access to his financial data.

This introduction has demonstrated how zero knowledge proofs facilitate new kinds of privacy guarantees, in new settings.

Michele Loi – Data Cooperatives and the Wide Dispersal of Data Power

Data cooperatives could change the competition landscape, promoting the wide dispersal of data assets, and making these assets available to smaller actors. This talk argued that in an ideal world, this may mitigate opportunity inequalities in society and in the economy. The presentation covered 1) the political principles supporting data cooperatives; 2) some unsolved challenges and problems of the data cooperative model.

What is a data co-op?

Here is one possible definition: A data co-op is a cooperative legal forum where users have democratic control over management, each data subject has one vote, the organization is a non-for-profit, and ownership of physical capital does not translate into voting rights. The members define the co-op's goals and values, they control management of the co-op, and they provide their personal data.

There are different possible models for co-ops (mixed regimes are also possible):

1. Socialist

   All members share all anonymized data and the management decides what to do with it.
2. Liberal

   Each member decides what data to share with whom and for what purposes.
3. Libertarian

   Each individual uses the platform to transact data for any purpose (with minimal control for unlawful use). The co-op charges fees for shared personal data management/online marketplace services and infrastructure (including cybersecurity).

Some of the central challenges of data co-ops:

1. A Co-op will only be able to generate significant economic value after a critical mass of users have joined it.
2. There is a trade-off between creating economic value for the co-op's users and allowing them to remain in control of their data.
3. Inasmuch as users don't disclose all data, but keep some of their data off the co-op, this would create biased data and would reduce the value of the co-op's data for many applications.
4. Decision making in the co-op. Can the majority decide that all anonymized information should be shared? Usually in co-ops, decisions are reached democratically, however we know that even stockholders are often not invested enough in a company to be able to make an informed vote on most issues.
5. Protection of the data in the co-op.