The Operational Framework: Institutional Controls

ABSTRACT

To realize the promise and prospects of big data and avoid its security and confidentiality perils, a balanced set of institutional controls are needed. These institutional controls must support and reflect greater user control over personal data and also large scale interoperability for data sharing between and among institutions. Core capabilities of these controls include responsive rules-based systems governance and fine grained authorizations for distributed rights management.

Basic drivers and inhibitors underlying the emergence of big data are discussed. Emergent characteristics and capabilities comprising larger big data trends are explored, including the emergence of location and geo-aware services, harbingers arising in urban and dense network environments, web 2.0 cross-system interoperability, federated identity for intensely personal data services and various major industry and governmental dependencies on sharing of or access to massive distributed sources of data.

This article discusses the role of personal data stores and user-centered identity for data sharing services as key components of a broader New Deal on Data approach. From an institutional perspective, the fit of open data components, extended network systems design and cross-boundary federated infrastructures are examined as part of strategic enterprise architecture.

Illustrating the nature of institutional controls, the article posits common business, legal and technical use cases in mobile communications, financial services, social networking and e-government from the vantage points of individual users, system providers and third party users or service providers. The current state of affairs is compared with near and longer term future big data scenarios to highlight the emerging dynamics in play.

The chapter concludes with a discussion of the problems and prospects for managing the transition to big data systems and identifies needed interdisciplinary research agendas in the computational social science, informatics, economic and political science fields.

ESSENTIAL ELEMENTS OF THE NEW DEAL ON DATA IN THE CONTEXT OF INSTITUTIONAL CONTROLS (AREK)

To realize the promise and prospects of Big Data, and to avoid its privacy perils, we need a balanced set of institutional controls. Theses controls must support and reflect a greater user control over personal data, as well as large scale interoperability for data sharing between and among institutions. The core capabilities of theses controls should include responsive rule-based systems governance and fine grained authorizations for distributed rights management.

Our lives are embedded within institutions. We are citizens of the countries and cities, receive services from telecom operators, search for things to buy in the online stores. All the activities we perform generate data, and those breadcrumbs of our lives are important part of the Big Data promise. The data that are not curated by us, but is collected as it is, reflecting our lives.

Today, all these data that we generate in the context of institutions, are closed in silos. The trace of our mobility is owned by our phone provider, our music tastes are stored and used by music services. For these data to be useful for society, it must be opened, must be used much more than it is now. If the access to the data for creating the value, either for the user or for the society, is very limited, it doesn't matter how big the data is. The value is not in the sheer data existence, it must be created from the data. Opening the data from multiple silos at once is even more challenging. Living under multiple jurisdictions, accessing the multi-faced data about one person may be prohibitively difficult, silos are hard to crack open. And such data, no just Big but Deep, covering multiple facets of person's life may be invaluable for research.

Recently, we have shown how challenging but also perfectly possible is to open such institutional Big Data. In the Data For Development (D4D) Challenge [1], the telecom operator Orange opened the access to a large dataset of records from Ivory Coast. Organized as a challenge, teams of researchers came up with life-changing insights based on the data. The privacy of the people was protected not only by the technical means, such as removal of the Personally Identifiable Information, but also by the legal means, with the researchers signing the agreement that they will not use the data for evil. As we have seen in several cases, such as Netflix Prize privacy disaster, true anonymization is extremely hard and some of the weight of the privacy must rest on the legal framework.

Opening the data from the silos, by publishing static datasets is important, but only the first step. We can do even more important things, when the data is available in real time and can become a part of a nervous system of a society. Epidemics and traffic congestions can be monitored and prevented in real time, underpferoming students can be helped, people with health risks can be treated before they get sick. The same data can be used for stalking, burglarizing my home, and as a reason to charge me more for an insurance.

In the Unique in the Crowd project [1], we have shown that, even though human beings are highly predictable [2], we are also very unique. Having access to one dataset, it is easy to uniquely fingerprint someone based on just few datapoints, and use this fingerprint to discover their true identity. The higher the resolution of the data, the better the data, the easier it gets.

The question of privacy in this context effectively becomes the question of control. Who can release the data of my movements? To whom? How much and how often? The data are collected by the institution. The data are about me and do not belong to me, I may not even be aware that they exist. I cannot decide upon them, I cannot check them out. I cannot delete them. And very few can use such data, even if I wanted them to.

It does not have to be like that. Within existing legal frameworks, it is possible to change the vantage point of the data ownership and put the user, the entity about whom the data are, in control. It may be the copy of the data living in the great silo, that is being given to the user. The user becoming the owner of their copy of the data, or where possible the original, owner in the old Common Law sense: the right to use, transfer, and remove. An example of such mechanism is Blue Button initiative [2], where the patients can get the copy of their health records. Once the copy is with them, they can do with it as they wish: give it to someone, make it public, do research on it, destroy.

The user can accumulate data about herself from multiple places. Healthcare records, mobility patterns, favorite movies, all this information belongs to the user and can be accessed based on this user authorization. This changes how and what data that can be obtained for research and providing services. Rather than gaining access to the movements of millions of people from a telcom operator, one can potentially gain access to a smaller number but much richer datasets describing the users from the mobility, health, shopping etc. perspectives. New startups do not have to build the user profile from scratch, but can jump in offering competitive services based on the user's collected data.

The first, operational challenge of moving towards the end-user data ownership on a large scale, is to create an ecosystem where such user-owned data is noticed and accessed. We are currently within a feudal framework: Facebook owns the data generated by you and about you, and will provide the access to it to the 3rd parties that you might or might have not authorized. It is reasonably easy to download all your data from Facebook. It is reasonably easy to put it on Dropbox or even create myself-API, becoming a self-hosted API to one's own personal data. The challenge is to have clients to talk to this API and provide services, rather than going to Facebook for your data. Today, virtually no-one is ready to access your data directly from you. We have done slightly better on the Internet scale with identity: you can deploy your own OpenID server fairly easily, and many services will allow you to sign in. We should be heading in the same direction with data.

The way the user grants the authorizations to the data he owns, is not a trivial matter. Think who you have autho-

[1] http://www.d4d.orange.com/home

[2] http://www.healthit.gov/bluebutton

rized to know what city you are in today. The 'Yes' you have clicked many times, gave access to your location to multiple services. Every tweet, every geo-tagged picture, every checkin provide your location not only to the primary service you are using, but also to all the applications that you have authorized to access this data. Take a look at your applications page on Twitter, Facebook, Google...

Increasing the amount of data the user controls and increasing the granularity of the control, is meaningless if this control cannot be exercised in an informed way. The EULA-catastrophe, where you may be just as well giving up your soul when signing up without reading, will not bring us closer to the New Deal on Data. In the end, it must the be user that makes the informed decision about who will get the access to the data and for what purpose. Make the authorization interface too complex and you will fail. Make it too simple, and you will also fail, as you will not convey the complexity of privacy-related decisions. Write it in legal complex language, and you cannot claim that the user expresses informed consent. Start asking the user for authorization every 5 minutes, and you will only teach her to press 'Yes' every time a pop-up is presented.

In addition to the data ownership, we need a better way for the end-user to control what happens with them. Will user realize that clicking this single 'Yes' provides a service with a second-resolution location data? And what can be inferred from such data, regarding alcohol abuse (we see you a lot in a liquor store), driving habits, not enough exercise. This gap between the interface and the effect, can render the data ownership meaningless. There is a need for Living Informed Consent, where the interface for the user to grant the authorizations is created to give the user understanding of the consequences of the granted authorizations. This understanding will never be perfect, but aligning this user's understanding with the reality is the goal of the Living Informed Consent concept.

We envision several ways the Living Informed Consent can improve user's understanding of the authorization she is granting. The underlaying principle is that the status of the authorizations expressed via the interface (website, application) is the contract. By pressing the buttons, the user initiates technical actions (for example creation of OAuth2 tokens), but also changes her business and legal relation with the service. Such single screen, with a timestamped log constitutes a history of the consent. The granularity of offered control may differ, and some actions may or may not be permitted. Still, at any point in time, the user is in certain relation with the service, in the Business, Legal, and Technical domains. The consent only makes sense when the user understands what she is consenting to. Why even bother asking otherwise. Part of the gestalt is to provide concise authorizations description written in plain English. They will not always be trivial and may sometime turn into paragraphs of text, still the goal should be to provide a description easily understandable for the target audience. Additionally, the goal should be not only to ask for the access to data, but also include the purpose of the access. Location is a type of data. Using location to provide personalized music and using location to increase my insurance for careless driving are two very different authorizations. Currently, the widely used OAuth2 framework does not support the notion of purpose, focusing only on the data being accessed.

One possibility to make it easier for the user to understand what is happening with her data, is to reduce the dimensionality of the data already in the user-controlled domain, and only send high-level answers to the service requesting them. A lot can be inferred from a raw location trace, this fact is foundational for the concept of Big Data. The moment the raw data leaves user-controlled domain, it can be used for many things, some of them the user may have never thought about, and could not possibly have expressed informed consent. Extracting the high-level features of the data on the user side, as described in the openPDS framework, should allow for more informed decisions regarding the data access. All the raw data should not run wildly with every service providing a minuscule service to the user, exactly because a lot can be done with these data. It is much easier to control what can happen and thus what are the consequences of disclosing the city you live in versus all your location updates from the last year. It is not a perfect solution; even low-dimensionality data can still be used for evil and can be correlated with different sources. It is however a big step in the right direction, for the user to decide upon disclosing how much liqueur she buys per week versus this information being inferred from the GPS trace provided to a service in exchange for personalized music.

In addition, the information about data access and usage needs to be an integral part of Living Informed Consent. How often do services sample my location? Are they tracking me in real-time, or do they access the data on weekly basis? Am I singled out by them in how much they query about me, or is it the same for all the users? Being able for the user to answers those questions in a simple, even casual way, is a crucial part for the user to remain in the state of Living Informed Consent. Authorizations should not be of 'fire and forget' kind, instead they should be re-evaluated in some orderly fashion. How often depends on many factors, including the sensitivity of the data accessed, reputation of the service, user preferences, the balance between control and annoyance.

Giving the data ownership to the end-user makes it easier for the institutions to facilitate the data use. As the user is the fully-empowered party to make decisions about authorizing access to all here data, multiple silos do not have to be visited and contracts between them made. It is sufficient to talk to the end-user to gain access to all the data about her.

A crucial component for realizing this vision is the identity. It must be possible for the multiple institutions to find the user to give the data to. It must be possible for the user to identify multiple institutions and see where the data is coming from. All the questions that we have asked so far, who owns, who controls, who decides, who accesses, must have the 'who' component addressed.

Just as the data of the user should live under single control

of the user, the identity should also be brought closer to user control. It does not necessary mean that every user should be their identity provider, but rather than having hundreds of accounts in multiple services that do not interoperate, the identity of the user should be build on the principle of federated identity, where the services allow the user to choose their identity provider. In addition, just like data, certain attributes of identity need to be protected. Service does not need to know my email address to be able to log me in, a pseudonym is sufficient. If such service has a valid reason for asking for my address, it should be based upon my grant of authorization. Authorization that can be revoked and monitored.

In the existing system it is often hard to introduce the user data ownership. It may be for technical reasons, building infrastructure providing the space for the data. It may be for business or legal reasons where the data is considered not suitable for sharing. It may be for the lack of a clear incentive why to do it, and how to interact with the users. We feel that the first step in introducing more privacy into such system, is the notion that the user must be entitled to know at least about the existence of the data about her. The right to know about the data existence is hard to deny. It can be realized in an easier way and with less friction than transfer of the actual data. It can be the first step, enforced by legal framework.

REFERENCES

1. de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports 3* (2013).
2. Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of predictability in human mobility. *Science 327*, 5968 (2010), 1018–1021.