Open data and digital identity: Lessons for Aadhaar[1]

December 3, 2017

Amba Kak

Mozilla Foundation
ambakak@gmail.com

Smriti Parsheera

National Institute of Public Finance and Policy
smriti.parsheera@gmail.com

Vinod Kotwal

Department of Telecommunications
vinod.kotwal@gov.in

Aadhaar, the largest national biometric system in the world, has been lauded for its promise to bring efficiencies to government service delivery, and the stimulus to private sector innovation. Yet it is contested and criticised for the vulnerabilities created by biometric data, potential threats to privacy and exclusion. However, in all of this, there has been relatively less exploration of the 'open data' possibilities from the Aadhaar ecosystem.

Every day, large volumes of data are being generated through the use of Aadhaar-enabled authentication and eKYC systems, both by government and private entities. The challenge now is to find ways to nudge the UIDAI and all users of Aadhaar towards greater sharing of data, in privacy-protecting ways that do not create risks for Aadhaar-number holders. We propose an implementation framework that can achieve these goals by leveraging the existing provisions of the Aadhaar Act to create an open data ecosystem that balances the needs of openness and privacy.

[1] An earlier version of this paper was presented at the International Telecommunication Union (ITU) Kaleidoscope Conference, 2017 held in Nanjing, China.

Contents

1 Introduction

Aadhaar, meaning foundation, refers to a 12-digit random identification number issued by the Unique Identification Authority of India (UIDAI). Originally established under an executive order in January, 2009, UIDAI came to become a statutory body under the Aadhaar (Targeted Delivery of Financial and Other Subsidies, Benefits and Services) Act, 2016 ("Aadhaar Act"). The project currently holds a biometric database of more than 1.18 billion individuals. Covering over 85 percent of India's population, it is the largest national biometric database in the world.

From its inception, Aadhaar was a unique government project - in part due to its collaboration with technologists and entrepreneurs, and a focus on the potential applications or 'use-cases' the Aadhaar could lend itself to. This is also reflected in its API-based architecture, that allows private companies to query the database for authenticating users.

Its ability to 'uniquely' identify individuals based on their biometric / demographic information and Aadhaar numbers is the stated basis for the government's push to link Aadhaar across (and even beyond) government services. Over the years, the government has linked, and made mandatory, the use of Aadhaar numbers for various welfare schemes like the transfer of direct cash benefits under public distribution of food grains, employment guarantee benefits, mid-day meals in schools, LPG subsidies, etc. It is also increasingly used as identification proof for availing services like banking and finance, digital payments and utility connections, among others.

Despite this rapid proliferation, the goals and architecture of the project have met with growing resistance. The Supreme Court of India is currently hearing a series of petitions challenging the constitutionality of Aadhaar, its compulsory linkage for the delivery of government benefits, potential for exclusion of beneficiaries; and impact on privacy, among others. These hearings recently led to a pronouncement by a nine judge bench of the Indian Supreme Court that there exists a fundamental right to privacy in India, which cannot be denied except through a fair, just and reasonable procedure established by law. The Court also spoke of other tests to question the existence of a legitimate state aim and proportionality of the measure to achieve that aim (Bhandari et al, 2017 [20]). These tests will now be applied for testing the constitutionality of Aadhaar.

While the judicial determination of these issues remains pending, the Aadhaar database continues to grow as the focal point of a rapidly evolving digital ecosystem. Hence there is a need to examine the data emanating from the Aadhaar system, and its varied uses. Aadhaar is a publicly funded resource, and as such, there is a strong case for promoting the disclosure of data points that can facilitate more informed research, policy making, business decisions, as well strengthen the accountability of the UIDAI itself.

In this paper, we (i) identify the various streams of data generated both by the Aadhaar system, as well as its varied applications across sectors; (ii) identify the existing incentives for public and private sector to create open data; and (iii) suggest privacy principles and an implementation framework to guide the release of more open data through Aadhaar.

2 Sources and potential of Aadhaar data

Open data is defined as *"data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike"*. Therefore, the most important features of open data are - availability and access; re-use and redistribution; and universal participation (Open Knowledge International [9]). In case of Aadhaar, its open data potential is closely linked to its characteristic design, features and functionalities. We therefore begin by examining the architecture of the Aadhaar project and then proceed to identify the categories of data that can emanate from its different processes.

The UIDAI is tasked with three key functional processes: enrolment, identification and verification (MeitY, 2017 [14]) Through an extensive network of enrolment agencies, UIDAI collects the demographic (name, date of birth, gender, address) and biometric (fingerprints, iris scan and photograph) information of individuals for the purposes of enrolling them into the Aadhaar system. All the collected information is housed in, and managed by, the UIDAI Central Identities Data Repository. The next step of “identification” refers to the de-duplication of biometric data in the UIDAI database. In this de-duplication process the Aadhaar system performs a check of the information collected for each new enrolment against all the enrolled data to ensure “uniqueness”. This results in the issuance of a unique Aadhaar number to the individual, which is meant to be a random number with no built-in intelligence.

Finally, it is the verification process that is employed in a variety of use-cases. This verification can be of two kinds - authentication and eKYC. The authentication services respond with a “yes” or “no” answer to the Aadhaar number holder’s claim of identity and no personal information is shared in the process with the querying entity. On the other hand, electronic know-your-customer functionality or eKYC allows authorised users to seek a person's identity information (but not their biometric information) from the Aadhaar database. The UIDAI rules allow the authorised eKYC agencies to keep the collected data in their records and use it for the purpose of delivering their services.

The list of agencies that have already adopted Aadhaar-based authentication systems includes Government benefit transfers and e-governance initiatives, banks and financial service providers, telecom companies, and digital certifying agencies. As of mid November 2017, UIDAI reported over 13 billion cumulative authentication transactions and over 3.5 billion eKYC transactions. This represents a drastic increase over the 4.5 billion authentications and 665 million eKYCs reported as of December, 2016 (UIDAI [16]). A number of factors have contributed to this increase, particularly the encouragement of eKYC driven financial inclusion and its use by telecom service providers pursuant to directions issued by the Government.

As more and more Government and private agencies move towards Aadhaar-based authentication systems, we see two primary sources of data emanating from the Aadhaar ecosystem:

1. statistics of Aadhaar enrolment and usage of the database available with UIDAI; and
2. data generated through government and private uses of Aadhaar.

Each of these categories of data comes with a unique set of challenges pertaining to the ownership of the information, the extent to which it can and should be made public and the incentives that might drive such disclosure. Before turning to these issues in the next section, we first identify the types of information that can emerge from Aadhaar and its uses, and the potential value of such data.

2.1 Release of open data by UIDAI

The decision and the responsibility of creating open data vests upon the owner or manager of the database. This right is exercised within the bounds of legally permissible disclosures. We therefore begin this section by examining the extent to which the Aadhaar Act permits (or, at the least, does not prohibit) UIDAI from making any Aadhaar related data publicly available.

The Aadhaar Act does not expressly vest the ownership of the collected demographic and biometric data with the UIDAI. However, the UIDAI claims to hold the data pertaining to residents as a trustee/custodian. UIDAI's control over the collected data is also exemplified by the fact that the individual providing her information does not have the option to exit from the system (although she can request access to her information).

Irrespective of the issue of ownership, the sensitivity of the information and scope for its misuse demands that UIDAI, as its custodian, deal with this data in a highly controlled manner. Privacy and data protection concerns demand that an individual's Aadhaar number; the demographic or biometric information collected during the enrollment process; or authentication records of a person should not be released publicly, by UIDAI, its enrolment partners or the authorised users of its authentication and eKYC systems.

Keeping this in mind, the Aadhaar Act casts an obligation on the UIDAI to ensure the confidentiality of the identity information and authentication records of individuals. Subject to certain exceptions, the law also specifically bars UIDAI from revealing any information stored in its database or authentication records to any person. The authority is also restricted from collecting or maintaining any information about the purpose of authentication. These provisions put some basic restrictions on the information that can legitimately and legally be released in the public domain by UIDAI. However, in discharge of its daily functions, the UIDAI also gains access to a number of other data points that would not be captured by the confidentiality restrictions in the Aadhaar Act. Many aspects of this information are already being released as open data.

For instance, the Authority currently maintains an online dashboard that offers data about the State-wise status of enrolments, including by age and gender and the entities involved in the process. Similarly, monthly information is also being made available regarding the usage of the UIDAI authentication / eKYC architecture by its approved agencies for the period post December, 2016. This is accompanied by daily transaction figures, name of the authorised entity making the request and type of authentication (biometric, demographic or using one-time password) for the last one month. While these are notable developments, the system could gradually evolve to offer more and more granular data on a daily basis, including historical data

In comparison, almost negligible amounts of information is available regarding the number of failed transactions in the Aadhaar ecosystem, in terms of generation of Aadhaar number, enrolment rejections (and reasons for the same), failure of authentication and eKYC requests, etc. Transparency demands that these and other process statistics should also be made available publicly by the UIDAI. Access to this information will guide the users of Aadhaar, researchers and other third parties in assessing the extent of its adoption, the purposes for which it is being deployed and the failure rates. The last of these elements can serve a legitimate basis for conducting a systematic audit of the extent and cost of the potential exclusion from the benefits that have been linked to Aadhaar. This is a prerequisite for an open and informed debate on issues relating to Aadhaar, including in the context of the ongoing litigations on the project. At the same time, this data can also be used as a basis to make improvements in the system, including enabling more effective grievance redress.

2.2 Data generated by Aadhaar users

Authentication: Every day, large volumes of data are being generated through the use of UIDAI's authentication and eKYC systems, both by government as well as private entities. In case of an authentication query, the Aadhaar repository offers only a positive or negative response to confirm whether the submitted information matches with the information recorded in UIDAI's database. None of the Aadhaar information is shared with the requesting entity although the process of authentication in itself leads to the creation of new data. For instance, a bank that uses Aadhaar authentication to verify the identity of a customer prior to authorising the transfer of funds from her account is creating new data in the process. The bank is then in a position to use the fact of Aadhaar authentication along with customer data already available with it to generate daily details of the number of persons of different age groups who used Aadhaar authentication to carry out fund transfers of different denominations.

The Aadhaar Act and the regulations framed under it circumscribe the manner in which information collected through Aadhaar can be used by such requesting agencies. As per Section 8(2), a requesting entity can use the identity information of an individual only for submission to the UIDAI repository for authentication purposes. In the above example, the bank would not need to (or be able to) use the customer's identity information collected by UIDAI, although it would already have similar information in its records. The bank would, however, need to utilise the authentication logs generated through Aadhaar. The current regulatory framework may constrain such use due to the requirement that the authentication logs can only be used for certain identified purposes. This includes sharing of the logs for grievance redress, dispute resolution and audits by UIDAI. The regulations may therefore need to be revisited to clarify that the generation of open data, within the framework specified by UIDAI, would be regarded as one of the permitted uses of authentication logs.

eKYC: There is marked difference, however, when it comes to the amount of data made available to and generated by authorised eKYC partners. The Aadhaar (Authentication) Regulations, 2016 allow the requesting entity to gain access to the person's demographic information that is filed with UIDAI and printed on the person’s Aadhaar card. This information can be used by it “for its own purpose”, i.e. for the purposes of its business. It may also share the e-KYC data with other agencies for a specified purpose, with the consent of the individual.

With eKYC agencies, there is scope for release of valuable data points. We illustrate this using an example from the telecommunications sector. In September, 2016, a new telecom player, Reliance Jio, entered the Indian market employing Aadhaar eKYC as its primary mode of verifying and enrolling new subscribers. It is estimated to have added approximately 600 thousand new users per day in its first six months. More recently, the Department of Telecommunications has issued a direction to all telecom service providers to re-identify their mobile subscribers through the eKYC process by February, 2018. Based on current figures, this move would cover a telecom subscriber base of about 1.2 billion connections.

While the aggregate number of mobile users is significant, reports suggest that there exists a vast gender divide in the adoption of technology in India (Aneja and Mishra, 2017 [19]). Yet, we do not have any official statistics on the ratio of men and women among telecom users in India, either at the country-wide level or in local areas. The move towards eKYC verification of all telecom subscribers in India, means that telecom operators will soon have a Aadhaar-verified (private) database of telecom users in the country. This would include the gender and geographic information of each operator’s user base. Supporters of the Aadhaar-mobile number linkage see the re-identification process as an opportunity for improving trust in the existing customer information held by telecom providers.

Aggregated together, the verified database of each provider’s telecom users can serve to find out the total number of female telecom users in each geographic location, including rural-urban variations. Further, periodic disclosure of such data by all telecom operators will also allow the trends to be tracked over a period of time. It may be noted that most of this information is already available with the companies today also, however, no systematic measures have been taken from the perspective of aggregating this data and exploiting its open data potential.

The online registration system (ORS), a framework that links various government hospitals across the country to an Aadhaar based online registration and appointment system, can be another use case. The ORS facilitates eKYC of the patient, which is then used for providing appointments at various departments of different hospitals. Using the appointments database along with the Aadhaar identification information, ORS will be in a position to disclose aggregated data about the age and gender profiles of the patients visiting different departments. This information can be sewn together to gain insights into the broad categories of health problems faced by different groups, the burden on different departments and the variations based on the location of the hospital. All of this can contribute towards evidence-based research and policymaking in the field of healthcare.

Another notable feature of the Aadhaar database is that it was among the first government-issued identifications in the country to recognise “transgender” as a separate category (Nilekani and Shah, 2014 [7]). The release of aggregated data related to use of banking, payments, telecom, health, education and other Aadhaar linked services by members of the transgender community offers a unique opportunity to study the extent of their exclusion from the mainstream discourse. This however remains subject to concerns about the targeting of individuals and possibility of re-identification from aggregated data, given the small size of the total data set. These issues will need to be addressed through careful thinking about the principles that should govern the sharing of Aadhaar linked open data, as discussed further in Section 4.

3 Incentives to "open"

The case for promoting disclosures of open data emanating from Aadhaar applies equally to all authorised users of Aadhaar. However, the incentives for public and private users to disclose this data are very different. Unlike the public sector, where legal requirements and policy initiatives compel and encourage government agencies towards proactive disclosures, private companies are outside the purview of this legal framework. They also typically view data as a source of competitive advantage, and would be reluctant to disclose data points voluntarily. The challenge therefore is to find ways to nudge all users of Aadhaar towards greater sharing of data, in the interests of transparency for accountability, research and more sound policy making.

3.1 For public bodies

The legal basis for the government to open up datasets to the public comes from the ‘right to information’ (known in some jurisdictions as freedom of information) regime. The idea of open government data presupposes willingness of governments to proactively disclose information to its citizens, and has been a hard fought battle in many countries. In India, this right of access to information held by public authorities has been codified through the Right to Information Act, 2005 (RTI Act). The passage of the law emanated from a grassroots movement that insisted on “people's’ audit” of government services to address corruption.

There is a comprehensive proactive disclosure provision in Section 4 of the RTI

Act, which puts a general duty on every public authority to provide "as much information suo moto to the public at regular intervals through various means of communication, including the internet". This puts the onus on public authorities to release data, so that the public has to minimally resort to the use of the law to obtain information. The provision also states that all public authorities shall routinely disclose a varied list of information including about its functions, decision-making norms, documents held, employee contracts, budgets – along with a catch-all direction to release "such other information as may be prescribed". Some studies however suggest that the promise of Section 4 has been watered down significantly in practice due to insufficient proactive disclosures (RaaG & SNS, 2017 [13]).

Outside of the RTI Act, there have been a few other measures to encourage disclosures. The President of India, in her address to the Parliament in June 2009, voiced the need for "A public data policy to place all information covering non-strategic areas in the public domain. It would help citizens to challenge the data and engage directly in governance reform". In March 2012, the Indian Government brought out the National Data Sharing and Accessibility Policy (National Data Policy). It remains the only official policy document on open data, with the stated objective of increasing accessibility and easier sharing of "government-owned", "non-sensitive" data amongst registered users particularly for scientific, economic and social development purposes. Pertinently, the policy rationale for open data is the investment of public funds that goes into collecting and processing such data. The emphasis on government ownership and the use of public funds is also reflected in the scope of the policy, which defines data to be limited to that generated "using public funds by various ministries/ departments/ organisations and agencies of the Government of India". The policy however has not been operationalised in the form of binding legal rules.

Specifically in the context of Aadhaar, Nandan Nilekani, founding Chair of the UIDAI, made a speech in 2010 stating that "*Aadhaar enabled applications the UIDAI envisions can turbo-charge the enforcement of Section 4 provisions (of the RTI) across our subsidy and welfare schemes*". He further said that the "availability of electronic records within such programmes" would be a "natural outcome" of its linkage with Aadhaar.

The digitisation of records, however, on its own has not led to proactive disclosure. As discussed earlier, UIDAI has uploaded some heads of information on its Aadhaar dashboard, yet there remain several gaps in the publicly available data emerging from the usage of Aadhaar. This is particularly true in respect of its various applications, or "use cases". Research group IDinsight identifies "transaction or beneficiary-level data" as one area which would benefit those doing data-driven studies of the efficacy of the project (IDinsight [5]). However, such granular disclosures could raise privacy concerns as a result of which the law itself restricts UIDAI and its related agencies from gathering and disclosing certain types of user-level data. Where there has been proactive disclosure of government databases seeded with Aadhaar, there has been significant controversy around the disclosure of Aadhaar numbers in the process, which is not permitted under the Aadhaar Act. A report by a civil society group found that government portals using Aadhaar for making payments had uploaded the bank account numbers, and Aadhaar numbers of 13 crore people, raising serious data protection concerns (Amber Sinha & Srinivas Kodali, [1]). These proactive disclosures on the disbursement of welfare schemes serve as a means to ensure accountability in the disbursement of social welfare benefits. It is therefore essential to devise an acceptable mechanism of disclosures without compromising on the confidentiality requirements of Aadhaar or disclosing other personally identifiable information.

Section 8(1)(j) of the RTI Act provides that personal information which does not relate to any public activity or interest, or could cause unwarranted invasion of an individual's privacy should not be , unless there is a compelling public interest reason to do so. Further, Section 6 of the Aadhaar data security regulations also lay down a requirement that no government agency should publish Aadhaar numbers, unless they are redacted or blacked out "through appropriate means". Absent clear specifications about these means, governments could err on the side of caution by removing entire datasets. In the next section we explore how best to achieve the balance between the goals of open data for research and transparency for accountability on one hand, and privacy concerns on the other.

3.2 For private bodies

As discussed, Aadhaar is a public infrastructure being used by various private companies for authentication (through seeding) and verification (through eKYC). These companies, like telecom operators or banks, are custodians of several useful demographic data points, some of which have been identified above. We argue that there is scope to encourage and facilitate disclosure of information held by entities that use Aadhaar.

This could be done through various means. In the next section we propose a proactive disclosure regime, akin to the one in the RTI Act, which will be enforced through the UIDAI's contracts with such entities. Other options could include encouraging disclosures by way of non-enforceable but enabling government policies. This could be coupled with ongoing guidance on kinds of data that would be a priority for disclosure, along with the necessary safeguards.

Specific disclosures might also be mandated by particular government agencies or sector regulators. For instance, continuing with the earlier example of telecom subscriber data, the Department of Telecommunication or the Telecom Regulatory Authority of India (TRAI) could mandate that each telecom operator must share district, rural/urban, and gender-wise information of its subscriber base on a periodic basis, which could then be released as open data either by the government or the regulator.

This debate also needs to be situated within a broader global push to encourage private companies to contribute more to publicly available data, particularly for research and policy making. Although, the term open data is usually used in the context of government or government funded data, some like the Open for Business Report, 2014 (Gruen et al, 2014 [8]) suggest that the term would also encompass private sector data. For private sector data, the challenge is to incentivise the companies to release non-strategic data that would contribute to research and development.

The UK government has an innovative model of a voluntary programme (called Midata) for private sector disclosures that are made to particular consumers, rather than to the public at large. Established in 2011, Midata invites signatories to provide consumers with "increasing access to their personal data in a portable, electronic format" subject to certain principles (BIS, 2014 [17]). UK's Enterprise and Regulatory Reform Act (ERR) allows mandating private sector disclosures and empowered consumers to enforce their data access rights in court. In this way, the ERR Act serves as a way to incentivise companies to make voluntary disclosures, through the looming threat of enforcement of its punitive powers (Out-law, 2014 [10]).

The International Open Data Charter (a collaboration between more than 70 governments and stakeholders) also questions the boundaries of the data that a typical policy should cover. They state that while the focus has been primarily on "government owned data" - "often the datasets that most matter, and that could have the most impact if they were open, do not belong to governments" (Davies & Tennison, 2017 [15]). In fact, it goes further to recommend that governments should have the power to mandate open data publication as part of giving licences to run a register, or negotiating directly with private providers to secure access to data which can then be shared as open data.

Apart from government facilitated or enforced disclosures, the coinage of "data philanthropy" has been used to describe the trend of companies volunteering anonymised and aggregated data with (usually select) third party users who might use this for research or policy purposes. Facebook's decision to share data on disaster maps, including valuable location information shared by users, with trusted organisations like UNICEF and Red Cross (Facebook Research, 2017 [4]) and ‘data grants’ by the Mastercard Centre for Inclusive Growth (Randy Bean, 2017 [14]) offer some examples.

We also find similar instances from the telecommunications sector. Orange Telecom’s Data for Development challenge encouraged researchers to use aggregate data in pursuit of development goals like health, transport and agriculture (Orange Telecom, 2015 [11]). They also rewarded best practices of anonymisation and cross-referencing of data. In 2014, it was reported that South African telecom operator MTN made anonymised call records available to researchers through a data analytics firm that provides predictive solutions (UN Global Pulse, 2014 [18]).

While such voluntary initiatives, which focus on disclosures to certain trusted intermediaries, are very valuable and should be encouraged for the many benefits that they generate, it is relevant to distinguish them from actual “open data”. The goal of open data initiatives is to create unrestricted public access to the underlying information. It is therefore important to think about additional frameworks that enable the release of data points publicly making it accessible to a larger and growing pool of researchers and policy makers.

Another variation could be the use of interactive techniques. Here, the data administrator (say, in this case, UIDAI, government departments, banks, telecom companies) answers specific questions about the dataset without releasing the underlying dataset. For example, if priority areas for Aadhaar related open data were identified in advance, then this could act as a guide for the disclosures to be made subsequently. While the interactive method can prove to be instructive, we regard it to be only small part of the overall open data solution for the following reasons. Firstly, the RTI Act allows individuals to make such queries to public authorities, but the onus here would once again fall on individuals or research groups, taking away from the principle of open data altogether. Secondly, private companies are not included in its scope leaving any interactive disclosures on their part to be a voluntary exercise. Thirdly, the implementation of such a mechanism would still require a mechanism to scrutinise the data being released so as to prevent against privacy harms.

Taking into account these factors we proceed to identify the contours of what could be an Aadhaar-specific open data framework and the privacy and other challenges that may be encountered in that process.

4 Privacy and implementation framework

As we make a case for responsible data disclosures by the UIDAI and other government and private users of Aadhaar, the manner of implementation of this responsibility also needs to be spelt out. First and foremost, is the concern that any open data disclosures should not threaten the privacy of the individual data subjects, leaving them vulnerable to a host of harms, including financial fraud. In this section we propose an Aadhaar-centric open data privacy framework that must be supplemented by principles of interoperability, accessibility and comparability in the creation of open data.

4.1 Privacy framework for open data

Most data protection regimes today afford legal protection only to personal data or "personally identifiable information" (PII). The ability of this information to be traced to a particular individual or to an object associated or used by an individual is what creates the potential for harming the person's privacy. It is therefore unsurprising that anonymisation, which refers to the process by which information is manipulated to make it difficult to identify data subjects, has come to be adopted as safeguard to privacy concerns. As a result, anonymised data is often carved out as an exception to privacy principles. Recital 26 of the European Data Protection Directive, which is arguably one of the more comprehensive legal regimes on this subject, states that the principles of data protection shall not apply to "data rendered anonymous in such a way that the data subject is no longer identifiable".

However, in the last few years, there is mounting evidence that traditional anonymisation techniques do not adequately prevent the risk of re-identification of the data subject, thus leaving them vulnerable to similar threats as though they were explicitly identified. For instance, a study in United States found that 87.1 percent of the people were uniquely identified by their combined five-digit ZIP code, birthdate and sex (Sweeney, 2010 [6]). Another study re-identified data subjects based purely on their movie preferences on Netflix (Arvind Narayanan et al, 2008 [2]). Thus, the science of what data fields might lead to re-identification when combined with other fields (and even other available databases) is an evolving one.

Accordingly, in proposing a framework for open data related to Aadhaar and its uses, we begin with the foundational principle that a person's Aadhaar number or other PII can never constitute a part of an open dataset. Even when such data is sought to be anonymised, it is critical to assess the risks of re-identification, and propose privacy principles that minimise these risks. We do not attempt a granular analysis of the re-identification risk in the sharing of raw data possibilities from Aadhaar (although such an exercise would also be valuable). Instead, we attempt to provide a heuristic by which to understand these risks, and recommend some approaches versus others. A similar study was done recently, by the Berkman Klein Centre at Harvard, which provided a risk-benefit framework to analyse open data emanating from municipal governments in the US (Green et.al.[4]).

Paul Ohm offers a sobering conclusion in his research on anonymisation and re-identification - "Data can be either useful or perfectly anonymous but never both" (Ohm, 2012 [12]). In doing so, the author highlights a necessary tension between the usefulness of data disclosures and privacy interests. In the following section we look at two methods by which anonymisation might be attempted, and identify possible points of tension:

1. *Redacting "identifying information"*: This is the process of redacting fields of information that are typically understood to identify individuals. In the case of, say, the telecom subscriber database, this might include name, phone number and legally mandated confidential categories like Aadhaar number. For a researcher it might well be that the existence of a unique identifier would allow far greater linkages and insights, particularly when comparing several telecom companies' datasets. However, it is precisely this that would make individuals identifiable and vulnerable to privacy threats, including from firms that seek to utilise this data for various purposes like marketing or promotions. An alternate mechanism is to hash/ transform the identifying information before it is used. Other techniques like adding "noise" - variations at random to the dataset - are also being explored as potential solutions.

   We propose that re-identification risk in any Aadhaar linked dataset, including that of telecom subscribers, even where only licensed service area, gender and age are being used as parameters, should undergo rigorous assessments to mitigate against such risks. The use of appropriate masking techniques and their effectiveness should constitute a critical element of the dataset designing process.

2. *Releasing aggregate statistics*: Ohm points to another critical lesson - when PII is actually redacted from the dataset, with minimal risk of re-identification, then the release of the dataset on its own has little value for research. In the telecom dataset example, the primary insights would be aggregate statistics about total number of male/female/transgender, as well

> as statistics relating to age and licensed service area, and a combination of the three. Therefore, the release of summary statistics, without underlying full datasets, could be seen as a good starting point for facilitating more accountability, research and policy making.
>
> Accordingly, we propose in the next section that the immediate focus could be on the release of aggregated summary statistics generated through the use of Aadhaar. As discussed earlier, there could be various granular statistics, like authentication volumes and error rates, about the operation of the Aadhaar system that would help to evaluate the various programmes it is linked to and the operation of the system itself. Similarly, crucial information about the demography is held by multiple entities, and remains unknown to both government and the public – we discussed gender-base split up of telecom subscribers and health care disbursements as some examples.

The full benefits of open data will however accrue over time, as we develop a shared understanding of Aadhaar-specific principles of anonymisation and disclosures which is then used for putting out complete datasets in the public domain, while accounting for privacy protections. Interestingly, there can also be some other innovative uses of the Aadhaar database, which can be adopted even now without disclosing sensitive personal information. For instance, the list of Aadhaar holders could be used to create a dictionary of Indian names (with frequency) and this can be tracked over time to trace the periodic shifts in the popularity of particular names.

4.2 Monitoring and enforcement framework

Drawing from the above discussions, we propose the need for an independent implementation structure that can leverage the existing provisions of the Aadhaar Act to create a robust open data framework. We suggest that this can be done through the creation of a multi-stakeholder "open data committee" by UIDAI. Section 23(2)(p) of the Aadhaar Act entitles UIDAI to "appoint such committees as may be necessary to assist the Authority in discharge of its functions for the purposes of this Act".

The preamble to the Aadhaar Act recognises the importance of good governance and efficiency, particularly in the context of use of public resources. Further, the Aadhaar Act also lays down a number of requirements that are to be implemented by UIDAI through regulations framed by it and through the agreements that it enters into with authorised authentication and eKYC agencies. Accordingly, the creation of a committee that can assist the UIDAI in the discharge of these activities would fall within the scope of the Aadhaar Act.

We recommend that this committee should be multi-stakeholder in character to bring in technical expertise and viewpoints from a wide range of actors. This would include representatives from the Government and UIDAI, civil society groups, open data and privacy experts and various authentication and eKYC agencies.

We propose the following steps in this regard:

Step I: Recognising the importance of transparency and accountability as critical tools of good governance, the government and UIDAI should agree on the key priority areas around which Aadhaar related open data needs to be be built. Given the nature of data collected by UIDAI, gender, age and geographic location, would appear to be the logical choices.

Step II: UIDAI should formulate a new set of regulations to implement the Aadhaar open data policy. This would include the creation of a multi-stakeholder open data committee with representation from the Government, UIDAI, civil society, authorised authentication and eKYC agencies and other experts. The regulations will encode principles and processes for generating Aadhaar related open data. This process should be accompanied by a review and amendment of existing regulations that might constrain such use. For instance, the Aadhaar authentication regulations would need to be amended to allow the authentication records to be used for the purpose of generating aggregated statistics for the release of open data.

Step III: The open data committee should identify the types of aggregate statistics that may be generated by (i) UIDAI; and (ii) different categories of agencies that use Aadhaar for authentication and eKYC. To the extent that disclosures are sought to be enforced through UIDAI contracts, the committee would also recommend the appropriate provisions to be incorporated in the agreements between UIDAI and the relevant agencies. This step becomes particularly important in light of the fact that the information generated by each entity would vary based on the nature of its business and the likely purpose of its linkage with Aadhaar. For instance, an e-governance programme will have very different uses of Aadhaar compared to a payments service provider or a telecom company.

Step IV: The committee should also drive the process of developing Aadhaar-specific principles of open data, including on issues such as anonymisation, masking techniques, interoperability, etc. This should be accompanied by an open, consultative process to test the robustness of the proposed principles and solicit feedback on the same from experts and the public. Based on the inputs received through this process, the committee should make final recommendations to UIDAI, which should also be made available publicly.

Step V UIDAI should review the final recommendations of the open data committee and incorporate appropriate open data standards and provisions in the agreements entered into with different categories of authentication and eKYC agencies. In case the UIDAI does not agree with any of the recommendations of the committee, the reasons for the same should be indicated.

Step VI: The open data committee should also assist the UIDAI in the implementation of the open data principles adopted. They can do so by identifying potential violations and notifying UIDAI for the purposes of initiating necessary actions against any breach. It can also play a key role in adopting a communications strategy for sensitising Aadhaar users about the principles and value of Aadhaar related open data.

The proposed model will ensure multi-stakeholder participation in the Aadhaar open data framework. Further, a narrow focus on anonymised aggregate statistics in the initial phases will minimise privacy risks, while still contributing valuable data points to the public domain. The full benefits of open data will, however, accrue over time as we develop a shared understanding of Aadhaar-specific principles of anonymisation and disclosures. All of this will contribute towards better research, informed policy making, enhanced public accountability and design improvements in the Aadhaar ecosystem.

References

[1] Amber Sinha and Srinivas Kodali, "Information Security Practices of Aadhaar (or lack thereof)", The Centre for Internet and Society, https://cis-india.org/internet-governance/information-security-practices-of-aadhaar-or-lack-thereof/view, 16 May 2017.

[2] Arvind Narayanan and Vitaly Shmatikov, Robust De-Anonymization of Large Sparse Datasets, Proceedings of the 2008 IEEE Symposium on security and privacy, 2008.

[3] Facebook Research, Disaster maps methodology, https://research.fb.com/facebook-disaster-maps-methodolog, 2017.

[4] Green, Ben, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer and Susan Crawford, Open Data Privacy (2017) Berkman Klein Centre for Internet and Society Research at Harvard University, https://cyber.harvard.edu/publications/2017/02/opendataprivacyplaybook

[5] IDinsight, State of Aadhaar Report 2016-17, http://stateofaadhaar.in/wp-content/uploads/State-of-Aadhaar-Full-Report-2016-17-IDinsight.pdf, May, 2017.

[6] Latanya Sweeney, Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, Data Privacy Working Paper 3, 2000.

[7] Nandan Nilekani and Viral Shah, Rebooting India: Realizing a Billion Aspirations, Penguin, 2016.

[8] Nicholas Gruen, Houghton and Tooth, "Open for Business: How Open Data Can Help Achieve the G20 Growth Target", Lateral Economics, June 2014.

[9] Open Knowledge International, What is open?, https://okfn.org/opendata/.

[10] Out-law, Government steps back from threat to legislate on midata, goo.gl/jx9Emg, 2014.

[11] Orange Telecom, Data for Development, www.d4d.orange.com, 2015.

[12] Paul Ohm, Broken Promises of privacy: Responding to the surprising failure of anonymization, 57 UCLA L. Rev. 1701, 2010.

[13] RaaG and SNS, Tilting the balance of power: Adjudicating the RTI Act, http://snsindia.org/wp-content/uploads/2017/07/Adjudicating-the-RTI-Act-2nd-edition-2017.pdf, January 2017.

[14] Randy Bean, Mastercard's Big Data For Good Initiative, Forbes, August 7, 2017.

[15] Tim Davies and Jeni Tennison, “More than one way to open some data: government owned and government influenced”, Open Data Charter, http://opendatacharter.net/one-way-open-data-government-owned-government-influenced/, 2017.

[16] UIDAI, Aadhaar dashboard, https://www.uidai.gov.in/aadhaar_dashboard, November, 2017.

[17] UK Department of Business and Skills, Review of midata voluntary programme, July 2014.

[18] UN Global Pulse, Mapping the Next Frontier of Open Data, http://www.unglobalpulse.org/mapping-corporate-data-sharing, Sep 17, 2014.

[19] Urvashi Aneja and Vidisha Mishra, “Digital India Is No Country for Women. Here’s Why”, The Wire, https://thewire.in/139810/digital-india-women-technology/, 25 May 2017.

[20] Vrinda Bhandari, Amba Kak, Smriti Parsheera and Faiza Rahman, An analysis of Puttaswamy: the Supreme Court’s privacy verdict, https://ajayshahblog.blogspot.in/2017/09/an-analysis-of-puttaswamy-supreme.html, September 20, 2017.