How Public Should Public Data Be? Privacy & E-governance in India

Chintan Vaishnav
Sloan School, MIT
chintanv@mit.edu

Karen Sollins
CSAIL, MIT
sollins@csail.mit.edu

Nikita Kodali
EECS, MIT
nkkodali@mit.edu

Abstract

India is at a critical juncture as cities and governments in general move into the digital age in order not only to provide enhanced services, but to improve transparency and efficiency, and as this development is taking place at a time when the Supreme Court bench has decided in August, 2017 that privacy is a constitutional right. It is in this context that this research asks the following question: how can cities (a state actor) use citizen data to maximize the governance while protecting the citizens fundamental right to privacy? We examine how this balance and tradeoffs arise due to the way Urban Local Bodies (ULBs) collect, use, and disclose citizen data. For the investigation, we collect ULB metadata by examining the architecture of the products of eGovernments Foundation, one of the leading providers of digital tools for ULBs, and by directly by interacting with ULBs. Based on this investigation, we define two new indices, a Governance Efficiency Index (GEI) and a Information Privacy Index (IPI), that allows for measuring city's performance on the two dimensions, understanding where tensions arise in simultaneously improving performance on both, and where innovation will overcome such tensions. In addition, to being able to make specific observations and recommendations for the particular cases we examined, this work is a proof of concept that such analysis can and should be done more widely, in order to understand the tradeoffs between government transparency and efficiency on one hand, and privacy on the other.

1 Introduction

India as a nation is at dual inflection points. On the one hand there are many initiatives to move various aspects of Indian civil life into the digital age, including the work of the eGovernments Foundation beginning in 2003 [5], the work of the Unique Identification Authority of India (UIDAI) [22] beginning in 2009 and the realization of its work in Aadhaar [1] and the Smart Cities initiative [12] begun in 2015. Concurrently, the Supreme Court in its decision of August, 2017 [13] and further expanded and explored by the SriKantha panel of experts [25] has defined privacy as a constitutional right, which will have tectonic societal and operational implications.

At the juxtaposition of these two forces in this paper we analyze the increased the roles, rights and responsibilities of citizens, businesses, and municipal governments in simultaneously providing increased transparency and efficiency of municipal government operation and privacy as defined by the Supreme Court decision as well as the subsequent Srikrishna experts white paper. This paper takes concrete steps in analyzing the types of data being collected by municipalities, the uses of that data to achieve governmental operations, and tradeoffs between governmental effectiveness and privacy driven constraints on data access.

Because this research is focused on those two major transformations in Indian municipal governance, it is important to note that until recently all government information was collected on paper forms. The first step in automating that has been to collect the same data as on the paper forms, but electronically. We begin by observing that the paper-based data collection, provided significant privacy for the citizen in conjunction with much more limited transparency and efficiency of those operations. With a mandate to improve transparency and efficiency, cities are moving toward digitization of data collection, especially as enabled by the work of the eGovernments Foundation. This is occurring while separately the legal situation is being transformed with the Supreme Court decision on privacy and the succeeding committee of experts' white paper. The final element in the evolving context is the explosive technological growth in BigData and inference. It is at the cross roads of these that we proceed with our study.

As discussed by Barocas and Nissenbaum [15], questions of privacy can valuably be separated from questions of anonymity. There is increasingly sophisticated work on anonymity in large datasets, beginning with the early work on k-anonymity [27, 28] and thence to differential privacy [18] and searchable encryption [16] to provide anonymity. But as Barocas and Nissenbaum point out much of the challenge of privacy has to do with privacy violations or infringements based on inference. With respect to inference one interesting further path of research is to consider the basis for any inference, because that basis is often key to definitions of privacy-violating or unfair discrimination, such as the work by Datta [17] on discovering the underlying data used for inference. These advances in technology and thinking allow us to be careful in defining the scope of our study.

The context for this study, which is that municipal government operation, is divided into more than two dozen primary areas of operation, ranging from grievances, to several kinds of taxes such as property and water, to government services such as water and sewage hookups, to record keeping such as births, deaths, and marriages. These areas are organized into four primary foci of operation: revenue, expenditure, citizen services, and administration. Each area has a defined set of data it collects, a rich record for each transaction, ranging from identification of the submitter of the record, to information in order to perform the transactions. Each of the primary areas performs a number of functions, based on their internal organizational structure using some or all of the data and possibly data from other functions. Our objective is to define a framework for making choices between privacy and the governmental operations defined by transparency, efficiency and effectiveness at doing the required tasks. To understand and evaluate this tradeoff, we define two new indices: a Government Efficiency Endex (GEI) and an Information pPivacy Index (IPI). The central question we want to ask for each piece of data is about its utility in the context of its governmental area of operation, and the impact of either constraining or eliminating it. This will allow us to better understand the implications of privacy policies on governance.Our analysis enables us to answer questions about each data field collected as well as aggregate information, its use, importance in either efficiency or transparency and risk with respect to privacy. Furthermore, with more in-depth analysis we can understand the cultural, financial and functional exposure impact.

The paper proceeds as follows. In Section 2, we expand on the research questions considering evaluation in the context of both existing, widely held privacy principles and our two indices. Because the context of this work is so tumultuous, Section 3 provides extensive background separately on the situation on the ground in India (technical and cultural) in Section 3.1 and separately on the legal constitutional decisions with respect to privacy, including a discussion about Aadhaar, in India, in Section 3.2. For the reader knowledgeable in one or the other of these areas, these sections may be skipped. Section 4 explains our and data collection processes. Analysis and results are discussed in Section 5, and further synthesis with additional discussion of the use and impact of our two new indices occurs in Section 6. We conclude in Section 7 with a summary of our observations and lessons, a discussion of limitations of the work, and finally directions in which this work will move forward.

2 The Research Questions and Relevant Privacy Principles

In order to analyze the privacy implications inherent in the move to digitized Urban Local Bodies (ULBs) governance, we first formulate the research questions of interest. Then, we evaluate the questions from the perspective of the four key privacy principles, derived from a subset of those specified by the Supreme Court judgment, in Section 2.2. Finally, we mention the two indices that are proposed as a result of this research, a Governance Efficiency Index (GEI) and an Information Privacy Index (IPI), in order to help the reader evaluate tradeoffs in the collection, use, disclosure, and possibly opportunities for innovation, as they read the rest of the paper.

2.1 Research Questions

The overarching question of interest to this paper is this:

> *How can cities (a state actor) use citizen data to maximize the governance while protecting the citizens fundamental right to privacy?*

The Srikrishna Committee white paper [**?**] identifies several privacy principles to be of importance to India's context. Given the focus of this research on cities, we focus on the following subset: *Data Collection, Data Use, Data Disclosure, Data Security* and *Data Anonymity*. These are principles that affect both governance efficiency and privacy, and where cities, acting as a *Data Controller*, must determine *what* their policies ought to be. Below, these principles are presented as subquestions serving the above overarching question:

1. How does limiting the collection of citizen data affect governance efficiency and information privacy?

2. How does loss of data integrity affect data use?

3. How to disclose or anonymize citizen data without affecting privacy in undesirable way?

Answering these questions would then clarify city-level action on the rest of the principles that have more to do with *how to implement* privacy protection, and are also identified by Srikrishna Committee as being relevant to India, such as *notice* and *consent* for data collection and use, and *openness* with respect to publication of these policies. These can be dealt with once we answer the above questions, and are therefore not focused upon in this paper.

2.2 Evaluation in the context of privacy principles

The Supreme Court decision on privacy identified nine privacy principles, as the basis for its decision.[2] In this work, we concentrate on four: collection limitation, purpose limitation, disclosure, and anonymity, the last two of which fall under the security principle of the Supreme Court decision. We quote the decision on these three principles:

> (iii) Collection Limitation: A data controller shall only collect personal information from data subjects as is necessary for the purposes identified for such collection, regarding which notice has been provided and consent of the inidividual taken. Such collection shall be through lawful and fair means;
>
> (iv) Purpose limitation: Personal data collected and processed by data controllers should be adequate and relevant to the purposes for which it is processed. A data controller shall collect, process, disclose, make available, or otherwise use personal information only for the purposes as stated in the notice after taking consent of individuals. If there is a change of purpose, this must be notified to the individual. After personal information has been used in accordance with the identified purpose it should be destroyed as per the identified procedures. Data retention mandates by the government should be in compliance with the National Privacy Principles.
>
> ...
>
> (vii) Security: A data controller shall secure personal information that they have either collected or have in their custody, by reasonable security safeguards against loss, unauthorised access, destruction, use, processing, storage, modification, deanonymization, unauthorised disclosure [either accidental or incidental] or other reasonably foreseeable risks.

In this context, our first topic of concern is the question of what data is being collected, its uses, necessity, and requirements. As we have discovered in our detailed examination of the data that is collected and its degree of necessity and importance, the current practice is to collect exactly the same data that was originally collected on paper. We began, for instance, by noticing that mobile phone numbers are collected everywhere. This led to a series of interesting questions, ranging from ”Is this necessary?” to ”Is this valuable?” to questions of how ’identifying’ mobile phone numbers are, to whether mobile phone numbers could *not* be collected, since an increasing number of the submissions are from mobile phones. As we will see below, we made some interesting discoveries in this area, and the answers are not necessarily what we had expected. In addition, since we will be analyzing our results for two types of potential violations, data integrity and data confidentiality, understanding the value and importance of the data is a component of evaluating the risks as traded against the value. One component of our data collection was to identify all the kinds of data that are being collected.

Second, we focus on analyzing the types and nature of data collection from the perspective of the second principle above, purpose or use limitation. This principle focuses on the context in which the data is intended for use, which is closely related to Nissenbaum’s concept of *privacy in context*. [23] One of the important observations we make in our contextual analysis of the data is that in analyzing for purpose limitation the results are generally not binary. It is not the case that a piece of data is either absolutely critical or irrelevant. There is middle ground, where the presence of a particular data field will improve some aspect of the purpose and functionality. The mobile phone numbers above are an example. If the function in question is the assignment of a value to a piece of property for tax purposes, having the owner’s phone number simplifies setting up an appointment to inspect the property. Thus the assessment in the end may be both more efficient (one of the objectives in improving city services) and more accurate, because the inspector can get onto the property. In contrast, one might ask about the necessity or value of including mobile phone numbers in marriage records, as is current practice. Thus, a component of our research has been to collect information about which data is used for which purposes in which modules, and its degree of importance or necessity in achieving government functions. For this, we consider not only which data is needed, but exactly who, within the governmental organization needs that access.

Our last two principles both derived from the description of security, above, in particular disclosure and anonymity. Our focus with respect to both unwanted or unauthorized disclosure and opportunities for deanonymization have led to an analysis of the data with respect to a combination of two subjects of risks, the source of the record and, if there is a subject, then the subject. Thus, in a request for a water connection, the only person involved is the requester. In a grievance, there is the submitter of the grievance as well as the target. We notice that targets may be the ULB itself as in a pothole or street light being out, an individual, as in a complaint about nighttime noise, or a business, as in a complaint about unlicensed selling of food. We examine the exposure and opportunities for deanonymization along three types of impact: cultural, functional, and financial, both with respect to data integrity violations and data confidentiality violations. For each, we also consider the probability of it occurring as well.

This collection of data about the data, both the fields of data collection and analysis of the use and implications of violations of either integrity or confidentiality forms the basis for our analysis.

2.3 Evaluation indices

This work arrives at two indices, a Governance Efficiency Index (GEI) and an Information Privacy Index (IPI), that together show how exactly does the data collection, use, disclosure, and anonymity affect cities performance on the dimensions of governance efficiency and privacy. More importantly, collection of what type of data creates a tension between increasing performance along one index while lowering it on the other, thereby identifying a space where innovation can help in keeping both of the indices high.

The Governance Efficiency Index (GEI) arises from multiplication of two components: Timeliness of Service, and Accuracy of Service. A service is timely when delivered on or before the time limit published by the city. A service is accurate when the right service is delivered to the right citizen without any rework.

The second index, Information Privacy Index (IPI), arises from multiplication of three components: Right Collection, Right Use, Right Disclosure of the citizen data. The Data Collection is right when it is minimalist in terms of limiting it to data absolutely necessary for providing the requested service. The Data Use is right when the data is accessible to only those functionaries at the city who need to deliver the requested service. The Data Disclosure is right when no data that makes a citizen personally identifiable nor allows for any 'undesirable' inference about them. In section 6 we define and formulate these indices precisely, discuss how to measure them, and their various implications.

3 Background: Digitization and Constitutional Privacy

In this section we examine how India arrived at the point where we could and should seek answers to the above questions. We consider advances in technology and the legal and social basis for attention to privacy. With respect to technology, there are two thrusts that bear review, the increased accessibility to the Internet in the hands of individuals and the evolution of computerized systems in urban governments, including the technical work on Aadhaar. With respect to privacy, the development of Aadhaar and its biometric-identity basis has raised increasing numbers of legal cases on questions of privacy, which in the longer term led to the Supreme Court Bench decision in August, 2017 on the constitutionality of privacy in India. We will examine these two topics separately below.

Again, for the reader familiar with technology and digitization developments in India, we recommend skipping the next section. For the reader familiar with the constitutional developments in India with respect to privacy, we recommend skipping Section 3.2 and going directly to Section 4. .

3.1 Advances in Digitization of Cities

Over the last two decades or so, there has been a concerted effort to transform urban local bodies (ULBs), the legal term for municipal governments, with the intention of making them more agile in addressing citizens' concerns, to make them scalable, transparent, and efficient. We review this development here, because it is culminating in digitizing and automating as much of the ULBs' activities as possible. Although the ULBs themselves are responsible for their own governance, the federal government has also enhanced its support and encouragement of these advances.

At the federal level, the current ministry responsible for urban governance is the Ministry of Housing and Urban Affairs (MOHUA). Its current set of responsibilities [7] is the apex organization at the national level addressing issues of housing and urban issues, formulating policy, coordinating and sponsoring state level programs, generally executed by state and local governments. In terms of ULB governance, from the perspective of this project one of the particularly interesting programs formulated by the MOHUA was a program to reorganize ULBs to develop and support satellite ULBs around the seven largest cities, [9]. This is only one of their many programs and efforts.

In terms of the cities themselves, there are two primary trajectories that are important to note here. The first is the growth in urban population. The 2011 census [1] reported a total population of 1.2B people. There are three mega-cities of over 10M each and another 43 with populations over 1M. Recognizing that the metropolitan areas of the top seven largest cities in India had serious organizational and service-provision issues, [9] the government created the Jawaharlal Nehru National Urban Renewal Mission in 2005 [6] to create and improve governance in 63 cities across India includ these satellite areas, , especially to improve service delivery to poorer citizens. Part of this effort, as can be seen by the checklists in the satellite town efforts is to move their work to a digital rather than paper-based approach. It is important to note that it was during this time that the eGovernments Foundation, which had been founded in 2003, grew in importance, building their open source technologies for ULBs to use in providing citizen interfaces, both in terms of initial requests for services and in tracking the progress of their requests.Two interesting recent developments (missions) of the government of India are the Atal Mission for Rejuvenation and Urban Transformation (AMRUT) to provide many improved city services especially for poorer citizens, including among other things water, sewerage, transportation and park services [2] in their homes and neighborhoods across the country, and the Smart Cities initiative [10] to make city governance ”smart”, both begun in 2015. The latter is bringing a cadre of cities into its program each, in order to grow its base. There are also parallel programs to improve urban transportation, although that is less related to our particular study.

At the same time both mobile/smart phone use and literacy continue to grow, making smart phone access to the sorts of government services addressed in this paper increasingly available to more of the population. Overall, according to the 2011 Census, literacy grew from over 64% in 2001 to over 74% in 2011 although the average number of years of schooling across the country is only 5.1 years. [8] Data tells us [11] that in 2013 there were 524.9M mobile phone users and predicts that by 2019 that number will be 813,2M and a growth between 2015 and 2019 of smart phone ownership going from 199M to 383.9M. This is important because the two methods of citizen interaction in the egovernment types of initiatives are through the citizen using a smart phone to directly submit electronic forms or through direct contact in which a government representative will enter an electronic form on behalf of the citizen. The intention is that smart phone interaction comprise the vast majority of these, which becomes increasingly viable with increased literacy and increased smart phone availability.

Another significant development in this same time period is the work of the Unique Identification Authority of India (UIDAI) and what has come to be called Aadhaar, as mentioned above, the move to biometric identity for all Indian residents. In an effort to create a comprehensive database of its citizens, in 2009 the Government of India created Aadhaar, the worlds largest biometric ID system, collected by the UIDAI. We note here that as Nilekani and Shah [22] report that between 2009 and 2014 they had registered 900 million people.

3.2 The Constitutionality of Privacy

In 2009, India began down the path of registering every citizen with their information and certain biometric data under Aadhaar. While Aadhaar was used to verify identification and document citizens, there occurred numerous cases where sensitive personal information had been leaked or hacked and privacy was compromised. As a result petitions questioned the need to collect biometric information, data security, and personal privacy, the central government issued a Supreme Court bench to decide on whether under the Indian Constitution, if privacy is a guaranteed fundamental right.

3.2.1 Supreme Court Decision

At this point, the central government issued a nine-judge bench decision to reflecting how the Constitution makers envisioned the nature of privacy:

- Is privacy a guaranteed fundamental right in the Constitution?
- What is privacy defined as?
- Is the right to privacy embedded in the right to liberty and personal dignity, or other guarantees of protected fundamental rights?
- In what parts of a citizen’s life is privacy guaranteed?
- How much should the government regulate privacy (nature of regulatory power)?
- What are the different aspects of privacy and does the Constitution cover some but not the others?

On August 24, 2017, the Bench unanimously decided that under the Indian Constitution, privacy is a fundamental right other than for reasons of national security, protection against crime, and protection of revenue. Observing that the Indian Constitution is a dignitarian constitution focused on upholding every citizens personal dignity, the Bench outlined several reasons why privacy is important for ordered liberty: (1) privacy is a form of dignity; (2) privacy provides a limit on the government’s power as well as a limit on private sector entities’ power; (3) privacy is key for freedom of thought and opinion; (4) it provides the right to control personal information as well as provides incentive for development of personality; (5) a guarantee of privacy prevents unreasonable intrusions by malicious public, private, or individual actors. It was determined that privacy is intrinsic to the values of Article 21 which gives citizens the right to life and personal liberty. Furthermore, privacy should apply to both physical forms and to technological forms of information; rights to enter the home should be up to the individual, excepting security reasons listed in Article 14. Lastly, privacy serves eternal values and guarantees as well as foundation of ordered liberty. Consequently, the Bench formulated a three-fold requirement for a valid law on privacy:

1. A law stating the privacy is a fundamental right according to Article 21 should exist.

2. To guard against arbitrary state action, the restrictions imposed on the nature and content of the law should abide by Article 14's exceptions to reasonableness.

3. The legislature must be proportional to the object and needs sought to be fulfilled by the law.

The Bench, recognizing that data protection and data privacy are complex issues that require expert opinion, mandated that the government create a Committee of Experts under the Chairmanship of Justice BN Srikrishna, a former judge of the Indian Supreme Court, to deliberate on a data protection framework for the country. While the constitutionality of the right to privacy was decided upon, the complexity of regulating privacy derives from the context-dependent economics of privacy. To better understand existing models of privacy protection and enforcement, an understanding of the transforming definition and value of privacy depending on contexts is important. The committee's report was delivered in late November, 2017 to solicit feedback, with a more final report in July, 2018.

3.2.2 Understanding the Value of Privacy in the Indian Context

The combination of big data and machine learning techniques can inform significantly about society and have the potential to bring about positive societal change and digital records can allow for greater efficiency and accountability. Policymakers have to reconsider how open should open data be, and where the fine line lies between keeping information private yet taking the most advantage of the large scale of digitized information. Often, data collected with informed consent for a particular purpose can be repurposed and analyzed for a different subset of insights. In these cases, the economic value of the data changes and especially in the big data realm, privacy regulation grapples with problems of unpredictability, externalities, probabilistic harms, and valuation difficulties.[26]

The economics of privacy concerns the trade-offs associated with the balance of public and private spheres between individuals, organizations, and governments with respect to personal privacy, as discussed by Acquisti et al.[14] In the Indian situation, the data generated by the citizens, who are often the data subjects or providers, is passed sequentially to the data collector, data holder, and data users who may be private or public entities providing a particular service to the citizens. The data collectors for e-governance data are the municipal governments, but the data holders in the backend differ from state to state, the e-Governments Foundation [5] being one of these hired by individual states independently. For true economic analysis, one would need to consider the full set of participants in the data including: (1) the data providers, often the subjects of the data; (2) the data collectors, generally the ULBs themselves; (3) the data storage managers, which may be the ULBs or private contractors; (4) the data analyzers and users, who again may be the ULBs themselves, or third party contractors; and, finally, (5) infrastructure providers such as cloud services and network providers. All will have access to the data in one way or another and all will have potential economic incentives in handling the data and possibly in the privacy of it (or not).

While citizens derive individual benefits and enjoy any common public goods produced using the assembled data, three key themes emerge from the flow and use of information about individuals by firms or governmental organizations as discussed by Acquisti et al.. First, a single unifying practice of privacy is difficult to formulate, as privacy issues of economic relevance arise in a wide variety of contexts and a variety of markets for personal information. Although the Smart Cities Mission mentions only security in the e-governance context, the Srikrishna Committee seeks to create an overarching data privacy protection framework that may minimizes costs from standardizing across sectors, even if it may not be able to minimize trade-offs.

Second, it is difficult to conclude whether privacy protection entails a net positive or negative change in economic terms, because of the tradeoffs between protecting against fraud and identity theft, and the costs of anonymizing data, securing the storage of data, and so on. For instance, while revealing mobile location information can be beneficial in improving traffic conditions or transportation efficiency, it may be considered an intrusion upon privacy if the government continuously monitors citizens' locations with the intent of surveillance.

Lastly, especially in a country like India where its 1.3 billion citizens lie on a broad spectrum of levels of education and income, a large number of poor or poorly educated people are at a disadvantage in accurately assessing the benefits or consequences from the sharing or protecting of personal information. Even the most educated citizens may not necessarily understand the power of analytics or apply machine learning. Furthermore even the most conscientious organizations may not be able to lay out all of the scenarios in which the information will be used. Disclosing data causes a reversal of information asymmetries: before the information is released, the data subject, holds greater knowledge about the information than the data holder. Afterwards, the data subject may not know what the data holder can do with the data and the consequences associated with sharing the data. While giving up privacy may allow a citizen to receive tangible benefits such as welfare approval, revealing the data may also incur intangible consequences, such as the loss of autonomy and possibility of increased surveillance .The market cannot respond appropriately to information gaps where users cannot express their true preferences for privacy protection. [19] As a result of this information asymmetry, designing even specific privacy regulations cannot necessarily cover unknown use cases or account for under-informed citizens.

In addition to the caveats that exist with creating privacy legislation, there are two basic tradeoffs that the government sees with the sharing of personal data, as discussed by Acquisti et al. First, individuals and communities can economically benefit from sharing data. One particular case in which the sharing of data is undoubtedly beneficial is India's Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA), the world's largest social welfare scheme [24] to alleviate poverty and provide benefits to impoverished and marginalized sections of society. At the same time, when Aadhaar is linked to welfare schemes and education scholarships, inappropriate access to the information could compromise personally identifiable information like banking information or culturally sensitive information such as caste.

Second, certain positive and negative externalities arise through data creation and transmission. The obvious positive externality is that specific aggregate and individual analysis of the data may lead to correlation between particular events in for example the health or education sector. For example, researchers with access to education data were able to discover that student attendance in low-income areas increased when in-school meals were provided. Negative externalities, however, can include intrusive surveillance by the government or targeted pricing by corporations arising from a comfort of sharing one's information. As a result, the economic value of one's personal data continuously changes depending on context and how willing other people are to share their personal information.

3.2.3 Existing Models of Privacy

The constitutionality and contextual dependence of privacy provide a considerable challenge in formulating one set of standardized regulations on the conditions under which personal information can be shared and the methods by which to share and monitor the data. The two main components of international privacy regulations are guidelines first, on how to protect the data and, second, on how to enforce privacy protection. Internationally, the three common models of privacy protection can be described as i) the ”Command and Control” Model, ii) the Self-Regulation/Sectoral Model, and iii) the Co-Regulatory Model.[25] The Srikrishna Committee assessed the three models and concluded that the Co-Regulatory Model was appropriate for India as its varying levels of government involvement and industry participation can be molded to the Indian context. The Command and Control Model, also known as the Comprehensive Model,[3] includes a general law that regulates the collection, use and dissemination of personal information in the private and public sectors, governed by an oversight body.

4 The Data and Data Collection

In this section we review the sources of our data, both in terms of use of the current tools for collecting data, through eGov, and our decisions about both site and data type selection.

4.1 e-Governments Foundation, Current installations and Digital Services

The eGovernments Foundation (eGov) develops platforms that enable city and state governments to improve accountability, transparency, and efficiency of the delivery of citizen services. The eGov platform is designed to aid in the management of four categories of government information: administration, revenue, expenditure, and citizen services. While administration and expenditure modules account for employee management, legal case management, payroll and pensions, assets, and so on, revenue and citizen service modules mainly include tax evaluations and registrations filed by citizens. Revenue sources include collection of property tax, water tax, trade licenses, advertisement tax and fees from government land and estates while citizen services include birth and death registrations, marriage registrations, an online citizen portal, public grievance registrations, and building plan approvals. The platform allows municipal officials to enter information and view individual and cumulative data on quantitative and geospatial dashboards. The digital actions of each employee are logged in order to monitor performance and accountability. It also promotes citizen engagement by interfacing with an online citizen portal and mobile app where people can submit and view the status of their applications and registration, improving transparency and accessibility. EGovs clients include but are not limited to the state of Andhra Pradesh, the state of Punjab, Greater Chennai Corporation, and the state of Maharashtra.

4.2 Site and Module Selection

For this study, we chose research sites located in a single state. For confidentiality reasons, the identity of the state is kept private. Within these states there are over 325 of ULBs. In the state we studied, the ULBs we chose are now equipped with the eGovs platform. ULBs are classified into three types by population: Nagar Panchayats have a population of less than 100,000 people, municipalities have populations greater than 100,000 people, and municipal corporations have populations of greater than one million people. For this study, we chose two municipal corporations and one municipality to construct a representative sample. Administratively, the Director of Municipal Administration (DMA) oversees the eGov implementations in all ULBs.

Within a state, the DMA, who provides state level oversight for support services in the municipalities, manages the Additional Director, Joint Directors, and Assistant Directors who oversee various aspects of all municipalities. Then, each municipal corporation houses a commissioner who, with the ULB mayor, provides administration and governance of the operations of each district. Each ULB is assigned a commissioner depending on which district it resides in. The Commissioner defines access controls for employees and can monitor employee performance. Within every ULB, there exist the Administration, Revenue, Accounts, Public Health and Sanitation, Engineering, Town Planning and Poverty Alleviation departments. Each department is responsible for processing certain modules classified under Expenditure and Revenue. All departments, however, are responsible for the Public Grievance module depending on factors discussed in Section 4.2.3.

4.2.1 Classification of eGov Modules

Of the four categories of eGov modules, the Revenue and and Citizen Services modules are public-facing and relevant to citizen data collection and citizen service delivery. These subset of modules can be grouped into four types based on what kind of information they may reveal about a citizen. First, modules may be revealing of personal identity, as they contain highly sensitive personal information. Birth and Death Registration, Marriage Licenses, and the Citizen Portal fall into this category. Second, modules such as Water Charges, Property Taxes, and Building Approval are revealing of personal assets. Third, the Trade License and Advertisement Tax modules are examples of modules that are revealing of a citizen's commercial assets. Lastly, the Public Grievances module forms its own category, as its function does not necessarily require citizens to reveal sensitive personal information.

4.2.2 Modules Selected for This Study

The Water Charges Module, Property Tax Module, and Public Grievance Module (PGR) were chosen for this study. They are some of the earliest implemented eGov modules at the sites we visited, and so produce a large volume of transactions.

4.3 Data Available for Analysis

The Property Tax (PT) and Water Charges modules offer various services. For example, in the Water Charges Module, citizens can apply for New Connection, Re-Connection, Closure of Connection, and so on. For both modules, the new property tax assessment and new water connection applications and workflows are fairly representative of and the most comprehensive of all of the services in their respective modules. As such, we refer to the new property tax assessment and new water connection workflows as the generalized Property Tax Module and the Water Charges Module, respectively.

Each of the three selected modules have their own workflow. Once a citizen submits a form or a request, all of the information that they have submitted is passed through various levels of hierarchy in the appropriate department.

4.3.1 Property Tax Module

The Property Tax Module includes services to evaluate or change property tax. While the representative service is New Property Assessment, the module also includes services like Transfer of Title, Bifurcation, Addition/Alterations, Revision Petitions, Demolitions, and so on. The module requires the applicant to give owner details, property address details, assessment details, amenities, construction details, floor details, details of surrounding boundaries of the properties, court documents, and vacant land details if applicable.

The quantitative evaluation of property tax payment depends on Usage, Classification, Zone, Age, and Occupancy Type data fields. Application particulars, such as contact details and address are important in verifying personal identity and assets. Once a citizen submits an evaluation request, the data is verified by a Junior Senior Assistant, then send to a Bill Collector and Revenue Inspector who verify details and conduct site visits. A revenue officer validates the evaluation, at which point the application must be approved at the Commissioner Level in order to be complete. In smaller ULBs, two or more of these functions may be completed by the same official. In larger ULBs, the process may be less uniform so that work is spread across multiple officials in the same level of hierarchy.

4.3.2 Water Charges Module

The Water Tax module, which the Engineering department manages, includes services to evaluate or change water tax payments. The fields that are essential for the evaluation of water tax are Zone, Uses Type, Water Source, Pipe Size and where it is applicable the White Ration Card. If the resident holds a White Ration Card, that means they are eligible for subsidies. In that case, the name and address become important for verification purposes, that the person holding the white ration card is the one living at the property. The Property Assessment ID must also be provided in the application, where all of the information from that Property Tax (PT) assessment is available to the officials in the Water Charges workflow. Other services in the module include Change of Usage, Closure of Connection, Reconnection Service, and Additional Water Tap Connection. Similar to the PT module, application particulars are necessary as well for verification.

The workflow generally looks similar to the PT module, where a Junior/Senior Assistant verifies application details, Assistant Engineer does a field verification and feasibility testing, Deputy Executive Engineer/Executive Engineer/Superintendent Engineer scrutiny the estimation details, and the Commissioner approves the evaluation.

4.3.3 Public Grievances Module

The Public Grievance Module allows citizens to submit a complaint to the municipality about complaints like sanitation issues, stray animals, illegal businesses, non-functioning of street lights, complaints regarding schools, voter lists, and so on. The module maps to a municipal administration department depending on the type of grievance submitted. This module requires the citizen to input contact details of the citizen and grievance details including location and photos.Once the complaint is submitted and reaches an official in the relevant department, the official has a certain number of days by which he must address the issue. These number of days, called Service Level Agreements or SLAs, are unique to the Indian state where we carried out our site visits. If an official does not complete his task within the given SLA, then the task will be escalated to the next level in the hierarchy. This accountability model promotes transparency and improves efficiency.

5 Analysis and Results

In this section we discuss both our analysis of the data and the results we derive from it, with respect separately on collection minimization, loss of data integrity, and data disclosure.

5.1 Understanding Data Use

While every data field collected in each of the three modules is visible to every official involved in the workflow of the module, officials do not necessarily need all the data available to them in order to complete their task. To understand the data viewable by each official, we began by creating a binary matrix to reflect which official actually used each piece of data. That let us conclude which officials actually needed access to each piece of data, and in particular what percentage of officials needed that data. We realized that there was more to learn in a matrix of data types and officials. In particular, our next question of the data was to label each cell with one of four characteristics: (1) whether the data field mandatory for this official to do their job; (2) whether the field is used by this official for operational use; (3) whether the field is used by this official for administrative use; or (4) it is unused.

A data field can be utilized operationally or administratively, or not utilized at all. An official uses a data field operationally if his role cannot be completed without access to that data field. On the other hand, an official uses a data field administratively if he must legally or in some circumstances be able to view the data field, even if he does not directly need it to complete his role. This allowed us to aggregate the information into Figures 1, 3[3] and 6, for the public grievances, property tax, and water modules, by computing the percentage of officials for whom each of categories (1) through (4) are true. This analysis leads to the observation that data collection could be reduced without any impact of operations and responsibilities of the officials.

The matrix also enabled us to summarize the data along the other axis by collecting the percentage of data items that fall into each of our four categories above for each official. Thus, for example in the Water charges module, the officials are Junior/Senior Assistant, Assistant Engineer, Deputy Executive/Executive/Superintendent Engineer, and Commissioner. A Junior/Senior assistant, for example, in the water charges module only verifies application particulars, such as checking that the address details match those in the property assessment ID or that the name on the application matches the name on the white ration card, if submitted. According to our site visits, connection details are irrelevant to the Junior/Senior assistants. In this case, only the data fields included in Application Particulars and the white ration card are used operationally. The rest of the data fields are not used at all to complete the role designated to the Junior/Senior Assistant. The Deputy Executive/Executive/Superintendent Engineer is the main decision-making authority in an evaluation and so reviews all details of the application. Hence, we could compute the percentage of fields that are mandatory, used for operation use, administrative use, or not used at all.

In the property tax module, at the Commissioner level, however, most of the data is used administratively. The Commissioner must legally be able to view all of the data, as he is the final approver for all evaluations. In most cases though, he does not check application details, as the details have been verified multiple times during the work flow, and mistakes in evaluations of small properties do not significantly affect the municipal revenue. If an evaluation for a large commercial complex is submitted, then he may check that the evaluated tax value correlates with the size or geographic location of the property. In this case, he will operationally use information about the building details, photo of the property, and owner details. We then graphed this data in Figures 2, 5 and 7 for the public grievances, property tax and water modules respectively. From this analysis, one is led to the observation that different officials only need access to subsets of the data, thus leading to reduce purpose driven access for each official to only those fields they need.

The above led to determining which data fields are necessary, collected for efficiency or accuracy, or unnecessary as indicated in Figure 8. This analysis in conjunction with the results above led to our final graph, Figure 9, in which we summarize the data utility of specific data to specific officials, overall. It is this final graph led us to summarize the overall utility of the data collected and used across all three of our modules, water charges, property tax, and public grievances.

5.2 Understanding Implications of Collection Minimization

The current online forms were transcribed from the previously existing paper forms. Pre-digitization, as much information as possible was collected from the citizen, even if some fields seemed unnecessary. Privacy was still conserved as data was not easily searchable, and citizens were fine with giving up more information so that they would not have to spend more time and money to return to the municipal office again to give more details. In the digital era where data is searchable and more easily accessible, collecting unnecessary data jeopardizes personal privacy.

From talking with various officials and gaining an indepth understanding of the workflow, we understand that all citizen data collected for each module can be categorized broadly into three categories: necessary for completing the function of the module, collected for efficiency/accuracy of the workflow, and unnecessary to the module (see Figure 8). Necessary use means that a data field is either required for a quantitative valuation or for personal identity or asset verification purposes. In Water Charges for example, the attributes Property Type, Usage Type, Pipe Size, Water Source Type, Connection Type, and Address are the only factors used to quantitatively calculate the water tax. Attributes such as Property Assessment number, White Ration Card, and Name of Applicant are used to verify details of the request, and so are important to validate the request. Data fields collected for efficiency/accuracy are not necessarily needed in order to complete the function of the module, but give officials a clearer picture of the application. Some of these attributes include information that can be observable on the site, such as amenities or details of surrounding areas for the property tax module. Other information makes establishing contact with the applicant easier, such as the name and phone number of the citizen filing a public grievance. Lastly, there are data fields that are unnecessary to complete a function, but are still collected. In PGR for example, the address of the citizen complaining is not necessary to either contact him or address the grievance. For Water Charges and Property Tax module, email is not used to contact or give updates to the citizen, as the status of an application is communicated verbally or through the citizen portal or app.

5.3 Understanding Implications of Loss of Data Integrity

In this section we focus on several sorts of implications that derive from loss of data integrity. We being by considering privacy implication and the consider the functional and financial implications of loss of data integrity.

The reasons for loss of integrity are left to different work.

5.3.1 Privacy Analysis I: Loss of Data Integrity

According to NIST,[21] the loss of data integrity is defined as data being altered in an unauthorized manner during storage, processing, or in transit. In order to understand the implications if data integrity is lost for each data field in the three modules, we built an Implications matrix. This matrix is too large to present in a paper, so we discuss it here. We determined three types of implications that the loss of integrity may have: cultural, financial, and functional. Cultural implications relate to what social inferences can be made about a person from a particular data field. There would be a financial implication if the citizen is affected financially, and a functional implication if the function of the module is unfulfilled.

5.3.2 Functional and Financial Implications of Loss of Data Integrity

For each data field in each of the modules, we evaluate what kind of implication may result if only that particular data field was altered in an unauthorized manner. For example, if only a person's name in Water Charges module was changed, and the name on the application does not match the name in the property assessment anymore, then the water tax evaluation would be stalled. As a result, the loss of integrity of the Name attribute in the Water Charges module would have a functional implication. In the case of Water Charges and Property Tax, financial implications and functional implications go hand in hand. If a tax evaluation is incorrect, then the citizen is financially effected.

From the Implications Matrix, we understand that the loss of integrity is likely to have functional and financial implications. The loss of integrity can affect the verification of personal identity or assets, the quantitative valuation of the water or property tax, or hinder efficiency. As described in Section 6.2, necessary data fields can be separated into those required for verification and those required for the quantitative calculation that is the output of a module. If the integrity of either type of necessary field is lost, then by definition the function of the module cannot be completed. There may be an over-valuation or under-valuation of a property or water tax if the factors determining them are changed. This poses a financial implication. A functional implication is the function being stalled if the identifying information of a person, his assets, or a public grievance are inaccurate.

If the data fields collected for efficiency/accuracy are corrupted, then the function of the module may not be stalled, but may be severely hindered. If a Grievance Photo was changed to a different image, a functionary would still be able to find the location of the grievance by the Grievance Details or contacting the complainer, but his job would likely be severely sidetracked by an irrelevant image. Therefore, we understand that protecting the integrity of citizen data fields is important in order to protect against harmful functional and financial implications.

5.4 Understanding the Implications of Data Disclosure on Privacy

In this section we consider the effects of data disclosure on privacy. We begin by considering who will be impacted, and the relative severity of that. We then discuss the financial and cultural implication, and conclude this section with a discussion of the challenges of inference over the grievance data.

5.4.1 Privacy Analysis II: Loss of Data Confidentiality

According to NIST, the loss of data confidentiality means that protected data is accessed by or disclosed to an unauthorized party. Similar to the Implications matrix for the loss of integrity, we built an Implications matrix for the loss of data confidentiality. For each data field, we evaluated what kind of implication may arise from the data field being exposed. At a high level, we can separate two dimensions of a citizen service: the type of service being requested/offered; and who is the service being offered to/offered by. The implications of the loss of confidentiality are the gravest when both aspects are leaked; meaning an unauthorized person knows not only what the service being offered is, but also who is being served. The type of implications can be functional or cultural.

5.4.2 Financial and Cultural Implications of Data Voluntary or Involuntary Disclosure

The next question we asked of the data was whether we could understand and categorize the types of implications from loss of confidentiality of the data. When we analyze the types of grievances and their implications, we find that some are solely financial. Figure 10 presents a subset of these that we found among the data. Typically, these complaints were made either by individual citizens or business, but about other businesses, but we note their financial implications. If a restaurant has complaints about incorrect slaughtering or garbage disposal, whether or not it is true, the restaurant may suffer financially. In contrast, we also found some kinds of grievances that were solely cultural. If a citizen complained about noise at night, it was likely to be about another citizen at home or walking down the street. As a third category, we found some complaints that had both financial and cultural implications. Examples of these can be found in Figure 11. Noting that generally, financial only losses would derive from a complaint against a business, a citizen complaining about another citizen would lead to a cultural loss, and a business complaining about a citizen would lead to a combination of cultural and financial loss, we tabulated the number of each type of relationship, as shown in Figure **??**. In addition, in that figure we noted the number of complaints where the respondent to the complaint would be the ULB, noting that that exposure there would have neither cultural or financial implications directly.

5.4.3 Grievance and Inference

For the PGR module, the types of inferences that can be made from the loss of confidentiality can be determined through an understanding of the actors and their roles. For a public grievance, there is a complainer, and the entity that must take action to fix the complaint. The complainer can either be a citizen or a business, and the entity who must take action to fix the problem can be another citizen, business, or the municipality. For example, a citizen may submit a complaint about the non-functioning of street lights. In that case, the municipal administration must take action. However, if the complaint is about the illegal slaughtering of animals, then the complainer could have been a citizen who may have been affected by the business or a legal competing business. The entity that must fix the issue is the illegal business.

From analyzing the two actors for the 110 types of public grievances published by the on-site's government, we found that we can assign certain types of implications depending on who the two actors are. We notice that citizen on citizen complaint has cultural implications, citizen or business on business has functional implications, and business on citizen has cultural and financial implications. As shown in Figure 12, most complaints must be fixed by the ULBs, likely regarding public infrastructure and health and sanitation issues. The majority of complaints that have implications come from complaining about a business, resulting in financial implications. If a business is affected by a complaint and the entity who has complained is exposed, the complainer may endure financial consequences due to bias against them from the business. In contrast, complaints against citizens tend to result in cultural implications, as the person against whom the complaint has been lodged may develop political, social, or other types of cultural biases against the complaining entity.

5.5 Risks from combining data

As background, we also reflect here on the inherent risks not only from individual data items, but also from combining data in various ways. To do that, we first consider the risks to privacy that derive from combining data within a database in order to expose a richer picture of the individual and then consider briefly the risks associated with combining data from different sources. We consider these in light of two important documents, the collection of papers edited by Lane et al. [20] and the US National Institutes of Standards and Technology's report on guidance for protecting the confidentiality of personally identifiable information. [21] The first includes a number of papers on the issues of privacy in the context of Big Data across the board, and the second, by focusing on a framework, processes and procedures for improving the protection of confidentiality, highlights a number of the concerns we raise here.

5.5.1 Risks from data within a single database

In order to make our discussion of privacy risks more concrete, we begin by focusing on the public grievance database, the data collected there and the risks associated with that data. In just this one database, the data can be divided into three categories: data about the person submitting the grievance, data about the location of the subject of the grievances, and data about the actual subject of the grievance. The submitter of a grievance includes several personal pieces of information: name, mobile phone number, and email. The person's name itself carries significant information about the individual beyond being just a name for the person, including language, religion, and caste. Thus simply including a name indicates significant cultural information about the person. Email and mobile phone numbers are used across many databases both inside these ULB operations and in other non-governmental settings, such as help desks, blogs, and so forth. In addition to the exposure of personal and possibly private cultural information, the combination of these three pieces of information can lead to creating a significantly richer picture of the individual. Furthermore, integrity violations of this data can lead to incorrect and therefore perhaps additional risky assignment of attributes to the individual. If an email address is corrupted into one that was used by someone else for posting on a blog, the original person may be incorrectly inferred as having made the posts, as an example.

When one considers location information such as locality, address and zone/ward/block the subject may be at risk for other sorts of inferences that may be privacy violations. First, location information can lead to inferences, about political perspectives by knowing the locale in which the person presumably votes, other personal cultural assumptions based on the location, and financial assumptions because people often live near other people of similar financial status. Second, inclusion of location information with a name can further identify the individual. Names may not be unique, but names in locations become increasingly unique. Thus, location information both alone and in conjunction with name can lead to yet further privacy violations. Finally, if a person submits a grievance about, for example, a street light being out at a particular location, that is a clear indication that that person was at that location after dark, leading perhaps to exposure of private information if the data is made public. In addition, integrity violations, as mentioned above, can lead to incorrect and perhaps unwanted or even dangerous inferences about the individual.

Loss of confidentiality or integrity in the data about the subject of a grievance poses yet further risks, in this case, to the subject of a grievance. Consider a grievance about sewage overflow at a location. This may lead to inferences about the types of activities happening inside the location, water usage, with possible significant implications in areas of water shortage, and so forth. Such information could lead to denial of water access, leading to both functional and financial implications based on either loss of confidentiality or integrity of the data.

5.5.2 Cross data-base implications: Implications of privacy in context

We must also briefly touch on some of the risks posed by the kind of data ULBs are collecting across the databases they support. First, consider that births, deaths and marriages must include names, for births parents names, locations, and that that information in general is public, there is an issue of combining things like name and address from these records with various fields from the grievances or property modules (property tax, water connection, water tax, etc.) to build a richer profile of an individual than any one database provides. Some of this information is necessary to be public (birth, death, marriage records) and some is needed only to achieve the function but need not be public (identify of grievance submitter). Furthermore, one must consider combining this data with other databases such as voter roles, bank information, presence and activity on social networks. And the list goes on.

What we note here is that governments that are collecting data about their citizens either as sources of information or subjects of those submissions are at risk with respect to the data both from loss of confidentiality and loss of integrity in four key dimensions of cultural, functional, financial, and privacy itself, especially as inference techniques become increasingly sophisticated.

6 Synthesis

We now synthesize the analysis above into two indices that help measure the efficiency of governance and information privacy. Defining these indices helps in highlighting the tension that arises when trying to maximize both indices simultaneously. This tension then carves out a space where innovation can help achieve a high level of governance efficiency, information privacy, and transparency.

6.1 Governance Efficiency Index (GEI)

We define Governance Efficiency Index (GEI) as follows:

*Governance Efficiency Index (GEI) = Timeliness of Service * Accuracy of Service*

GEI is constructed such that it ranges from 0-1, where a value of 1 denotes highest level of governance efficiency.

The definition of *Timeliness of Service* rests upon when a service is considered timely. We consider a service timely when it is delivered on or before the desired Service Level Agreement (SLA). The ULBs in India publish an SLA for each service they offer, as promised by the Citizen's Charter. [4][4] The *Timeliness of Service* component is measured as follows: for a given service, it is measured by the fraction of times the service is delivered on or before the SLA over a given unit of time (i.e., hour, day, month, etc.). For a given group (a division within ULB, the ULB as a whole), Timeliness of Service is measured by averaging the timeliness of the services delivered by the group over a given unit of time.

The definition of *Accuracy of Service* rests upon when a service is considered accurate. We consider a service accurate when right service is delivered to the right person without any rework. The *Accuracy of Service* component is measured as follows: for a given service, it is measured by the fraction of times the service is delivered without rework over a given unit of time (i.e., hour, day, month, etc.). For a given group (a division within ULB, the ULB as a whole), Accuracy of Service is measured by averaging the accuracy of the services delivered by the group over a given unit of time.

GEI is calculated empirically and in real time using the ULB level performance data

6.2 Information Privacy Index (IPI)

We define Information Privacy Index (IPI) as follows:

*Information Privacy Index (IPI) = Right Collection * Right Use * Right Disclosure*

IPI is constructed such that it ranges from 0-1, where a value of 1 denotes highest level of information privacy.

We define *Right Collection* as collection of those data fields that are necessary for delivering the service. In other words, without collecting these data fields, the requested service cannot be delivered. *Right Collection* is measured for a given service or for services offered by a given group as *Necessary Data Fields/Total Data Fields Collected.*

We define *Right Use* as access of data fields to only those (in the ULB) who need it for delivering the service. *Right Use* is measured for a given service or for services offered by a given group as *Number of Data Field To Which Access Is Necessary / Number of Data Fields To Which Access Is Granted.*

We define *Right Disclosure* as public disclosure data fields that protects personal identity and undesirable inference. *Right Disclosure* is measured for a given service or for services offered by a given group as *(1 - (Number of Data fields With PII or Undesirable Inference Disclosed / Total Number of Fields with PII or Undesirable Inference)).*

IPI is determined based on the analysis of data collection, use, and disclosure policies of ULBs. The real-time value of IPI will rest upon the frequency and types of service requests a ULB serves.

6.3 Trade-offs in maximizing GEI and IPI and a Space for Innovation

Setting up the two indices and understanding how data collection, use, and disclosure influence them reveals areas where trade off exist when maximizing the two indices.

6.3.1 Data Collection Trade-offs and Need for Innovation

In general, the less is the data collected, used, and disclosed, the better the privacy protection is. Less data, however, does not always help improve the governance efficiency as the data collected for each service falls in three categories to greater or lesser extent: Data Necessary for Offering Service *Xn*; Data Collected for Efficiency or Accuracy Purposes *Xea*; and Data Unnecessary for Offering Service *Xu*, which may be collected for legacy reasons. If we could collect just *Xn*, it would maximizing both GEI and IPI; however, presently, service provisioning occasionally require using *Xea*. The need for using *Xea* is arguably different in different sizes of ULBs (e.g., larger municipal corporation vs. smaller Nagar Palikas) because their environments are at a different levels of maturity in terms of adoption of communications technologies, and digitization of identity and assets. Here then is one space for innovation in answering a question: can we innovate to offer requested service in a timely and accurate manner without collecting *Xea* and *Xu*?

6.3.2 Data Disclosure Trade-offs and Need for Innovation

As discussed above, disclosing data could lead to generating various forms inferences. One may argue that any inference about individuals and their action is potentially ”bad” from the perspective of privacy protection. Ironically, however, being able to generate inference about underserved needs is critical to innovating and serving them. So, not all inference is ”bad”.

From this perspective, we must distinguish between inference that is *desirable* vs. one that is *undesirable*. At the outset, an undesirable inference may arise when someone can be socially, religiously, or in other ways discriminated such that their rights are either denied or deferred. Ultimately, it is the legal framework that is to protect the use of data for unlawful purposes; however, simply having easy access to undesirable inference increases the ability to accelerate unlawful behavior, especially the access to and the precision with which it can be done.

One form of desirable inference arises when unmet or underserved needs of a society or segments of a society can be understood without revealing individual identities. Here then is another space for innovation. To realize these innovation, one must answer the following question: How does one disclose citizen data for desirable inference about unmet needs to be met by innovations by non-state actors?

6.4 Transparency of Governance and Its Implications for Efficiency and Privacy

The analyses of the two indices, and the concern for transparency of governance that emerged during our field interviews, indicate that Transparency of Governance, when conceived as separate from data disclosure, is not affected by the pursuit of higher efficiency and privacy. Transperancy of Governance may be required at three levels:

1. Transparency to the one requesting service.
2. Transparency to gauge government’s performance
3. Transparency from the perspective of Use Limitation, i.e., is the data being used for what it was collected for.

As such, for the person requesting service, transparency of governance should be maximum, meaning, they should be able to access the status of the service requested and know how much time is left for service completion, who to contact in case of any questions. Such controlled access does not hamper efficiency or privacy.

Next, for gauging ULBs performance, it is possible to disclose data that will show how efficient the government is: the number of services fulfilled, the fraction of services fulfilled within SLA, etc. Doing so does not require disclosing any personal information or information that leads to undesirable inference.

The final element of transparency about actual data use seems difficult to deliver without incorporating the necessary engineering in the digital platform to track data use, and set and enforce policies.

While we believe level 2 of transparency for gauging ULB performance is easier to achieve, further research is required to understand how to avoid unintended disclosure of data when offering transparency at levels 2 and 3.

7 Conclusions, Limitations and Next Steps

7.1 Conclusions

In this paper, we examine the question of how cities (a state actor) can use citizen data to maximize the governance while protecting the citizens fundamental right to privacy. We examine this question in the context of three cities in India. Our analysis comes at a critical juncture as privacy has been declared as a fundamental right of every Indian by the Supreme Court of India in August of 2017, and a highest level committee of Central Government led by Retd. Justice Srikrisha is in the process of seeking public comments on the proposed data protection bill. Based on field research performed at eGovernments Foundation and thee of their client cities, we have analyzed the present data flow, use, and disclosure in these cities to define two indices that help us answer the above question: Government Efficiency Index (GEI) and Information Privacy Index (IPI).

With the help of metadata analysis, we demonstrate that both efficiency and privacy are measurable concepts in the context of urban governance. Our analysis shows how exactly to measure them. More importantly, we argue that there does exist a tension when trying to maximize both governance efficiency and privacy simultaneously, and identify the regions of data where these tensions are manifest, making this observation more than just a philosophical argument. The way to reduce these tension is to promote innovations that could improve governance services and privacy simultaneously. Our analysis identifies issues with data collection and data disclosure, where we must experiment further in promoting innovation by non-state actors.

7.2 Limitations

There are several areas where this work requires further refinement. First, the recognition that inference generated from data can be undesirable and desirable, and this has relationship to innovation requires refinement. In particular, defining what is undesirable or desirable inference more rigorously, and then reexamining where such inferences get generated.

Second, the current definition of GEI incorporates timeliness and accuracy, but does not say anything about at what cost. This is something we need to examine.

Third, the idea of Operational and Administrative Uses of data may need further classification because both operations and administration have multiple different purposes. We need to examine whether addming this additional granularity teaches us any new lessons.

We will address these limitations as our immediate next steps.

7.3 Next Steps

We believe the indices defined in this paper may be generalizeable beyond Indian cities and also beyond any particular product for digital governance such as the eGovernments product suite. The basis for this hypothesis is the belief that the component of both indices rest upon what is universally acceptable in the realms of urban governance (timely and accurate service delivery) and privacy protection (right collection, use, and disclosure of data). The measurement of these indices, however, may experience different levels of difficulty in different cities depending upon how mature is their e-governance paradigm. Anecdotally, we believe the three Indian cities we studied may be ahead on e-governance as compared to most other cities of the world, including many in the developed world.

Our next step is to measure these indices with real data, and at different levels of aggregation: for individual service, for sub-divisions and divisions of ULBs, for ULBs, and for the whole state. We expect that this analysis will be revealing in new ways. Computing these indices at multiple levels would bring in both the time and space dimensions. For example, we may now find out that there is seasonality to how efficient and privacy sensitive are the service requests in, say, summer vs. monsoon. Similarly, one can analyze why a particular service is more efficient in one ULB and not another, or how the IPI gets impacted in one ULB because a privacy sensitive service gets requested more often. Along similar lines, it could reveal how a particular division (e.g., Engineering) may have more efficiency and privacy sensitive requests as compared to other divisions. At the ULB level also one might see how two ULBs that have the same policy for collection, use, and disclosure of data, and hence the same IPI if all else was equal, may differ in IPI because of different frequency of service requests given their locations.

Next, we hypothesize here that the goal of providing transparent governance, construed in a specific way, need not be compromized by the pursuit of higher governance efficiency and privacy. One way to examine this hypothesis is to construct a third index, Governance Transparency Index (GTI) and systematically study how to define it and what are the tensions between transparency, efficiency, and privacy. Finally, the most exciting next step is to now take on the study of how to innovate to simultaneously keep efficiency and privacy high. This paper identifies contours of such a study that we will be undertaking next.

Acknowledgements

This work would not have been possible without the work and contributions of Gautham Ravichander and his colleagues at the eGovernments Foundation. In addition, we wish to acknowledge research funding as well as many probing questions from the MIT Internet Policy Research Initiative.

References

[1] 2011 Census, Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Indi. Available at http://www.censusindia.gov.in/2011census/.

[2] Atal Mission for Rejuvenation and Urban Transformation (AMRUT), Ministry of Housing and Urban Affaris, Government of India. Available at http://mohua.gov.in/cms/amrut.php.

[3] CIPP Guide: Comparing the Co-Regulatory Model, Comprehensive Laws and the Sectoral Approach. Available at by Privacy Commissioner of Canada at https://www.cippguide.org/2010/06/01/comparing-the-co-regulatory-model-comprehensive-laws-and-the-sectoral-approach/.

[4] Citizen's Charterin Government of India, Dept. of Administrative Reforms and Public Grievances, Ministry of Personal Public Grievances and Pensions, Government of India. Available at https://goicharters.nic.in.

[5] eGovernments Foundation. Accessed at https://www.egovernments.org on August 12, 2018.

[6] Jawaharlal Nehru National Urban Renewal Mission, Ministry of Housing and Urban Affairs, Government of India. Available at http://mohua.gov.in/cms/jawaharlal-nehru-national-urban-renewal-mission.php.

[7] Mandate, Ministry of Housing and Urban Affairs, Government of India. Available at http://mohua.gov.in/cms/mandate.php.

[8] NationMaster: India vs United States Education Stats Compared. Available at http://www.nationmaster.com/country-info/compare/India/United-States/Education.

[9] Scheme for satellite towns around seven megacities, Ministry of Housing and Urban Affairs, Government of India. Available at http://mohua.gov.in/cms/scheme-for-satellite-towns-around-seven-megacities.php.

[10] Smart Cities, Ministry of Housing and Urban Affairs, Government of India. Available at http://mohua.gov.in/cms/smart-cities.php.

[11] Statista: Statistics and market data on telecommunications. Available at https://www.statista.com/markets/418/topic/481 /telecommunications/. Accessed Aug. 14, 2018.

[12] Smart cities mission statement and guidelines, June 2015. Available at: http://smartcities.gov.in/upload/uploadfiles/files/ SmartCityGuidelines(1).pdf.

[13] Writ petition (civil) no 494 of 2012. Supreme Court of India, August 2018. Available at https://www.thehindu.com/news/national/article19551816.ece/ BINARY/RightToPrivacyVerdict.

[14] Acquisti, A., Taylor, C. R., and Wagman, L. The economics of privacy. *Journal of Economic Literature 52*, 1 (2016). Sloan Foundation Economics Research Paper No. 2580411. Available at SSRN: https://ssrn.com/abstract=2580411 or http://dx.doi.org/10.2139/ssrn.2580411. 2.

[15] Barocas, S., and Nissenbaum, H. Big Data's End Run around Anonymity and Consent. In *Privacy, Big Data, and the Public Good*, J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, Eds.

[16] Curtmola, R., Garay, J., Kamara, S., and Ostrovsky, R. Searchable symmetric encryption: Improved definitions and efficient constructions. *Journal of Computer Security 19*, 5 (Nov. 2011), 895–934.

[17] Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2017), CCS '17, ACM, pp. 1193–1210. Available at http://doi.acm.org/10.1145/3133956.3134097.

[18] Dwork, C. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II* (Berlin, Heidelberg, 2006), ICALP'06, Springer-Verlag, pp. 1–12.

[19] Hirsch, D. D. The law and policy of online privacy: Regulation, self-regulation, or co-regulation? *Seattle University Law Review 439*, 440-441 (2011).

[20] Lane, J., Stodden, V., Bender, S., and Nissenbaum, H., Eds. *Privacy, Big Data, and the Public Good*. Cambridge University Press, New York, NY, 2014.

[21] McCallister, E., Grance, T., and Sarfone, K. Guide to protecting the confidentiality of personally identified information (pii). Special Publication 800-122, National Institute of Standards and Technology, Computer Security Division, Information Technology Laboratory, NIST, Gaithersburg, MD,20899-8930, April 2010. Available at https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-122.pdf.

[22] Nilekani, N., and Shah, V. *Rebooting India: Realizing a billions aspirations*. Allen Lane, 2015.

[23] Nissenbaum, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[24] Singh, S. Evaluation of worlds largest social welfare scheme: An assessment using non-parametric approach. *Evaluation and Program Planning 57* (August 2016), 16–29.

[25] SriKrishna, B. N., Sundararajan, A., Pandey, A. B., Kumar, A., Moona, R., Rai, G., Krishnan, R., Sengupta, A., and Vedashree, R. White paper of the committee of experts on a data protection framework for india, Nov 2017. Available at http://meity.gov.in/writereaddata/files/ white_paper_on_data_protection_in_india_18122017_final_v2.1.pdf.

[26] STRANDBURG, K. J. Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context. In *Privacy, Big Data, and the Public Good*, J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, Eds.

[27] SWEENEY, L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-base Systems 10*, 5 (2002), 571–588.

[28] SWEENEY, L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-base Systems 10*, 5 (2002), 557–570.

Notes

[1] The book of Nilekani and Shah [22] is one of many references on this work, but provides a solid, deeply knowledgeable and introspective review of these developments.

[2] The nine privacy principles identified by the Supreme Court are: Notice, Choice and Consent, Collection Limitation, Purpose Limitation, Access and Correction, Disclosure of Information, Security, Openness, and Accountability.

[3] Figure 4 contains the complete list of fields represented vertically in Figure 3.

[4] Not all ULBs provide such SLAs.

Appendix: Figures

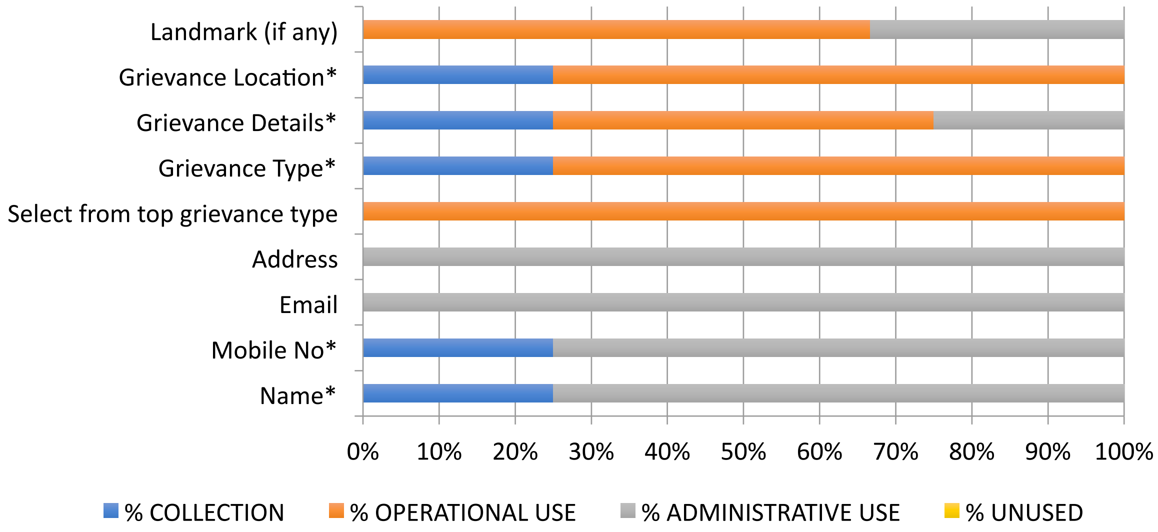


Figure 1: Public Grievance Data Usage

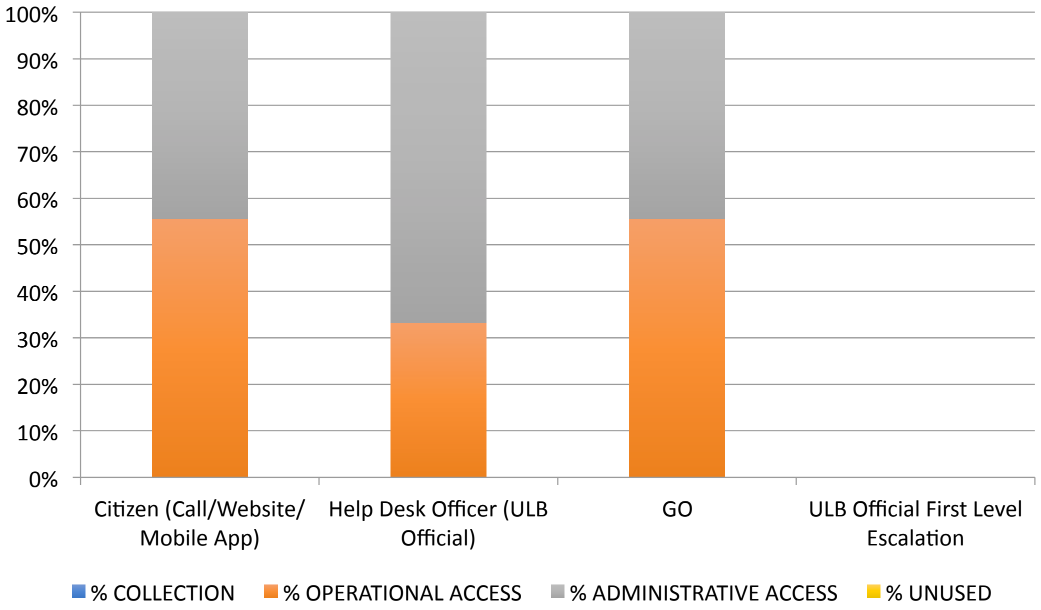


Figure 2: Public Grievance Data Access

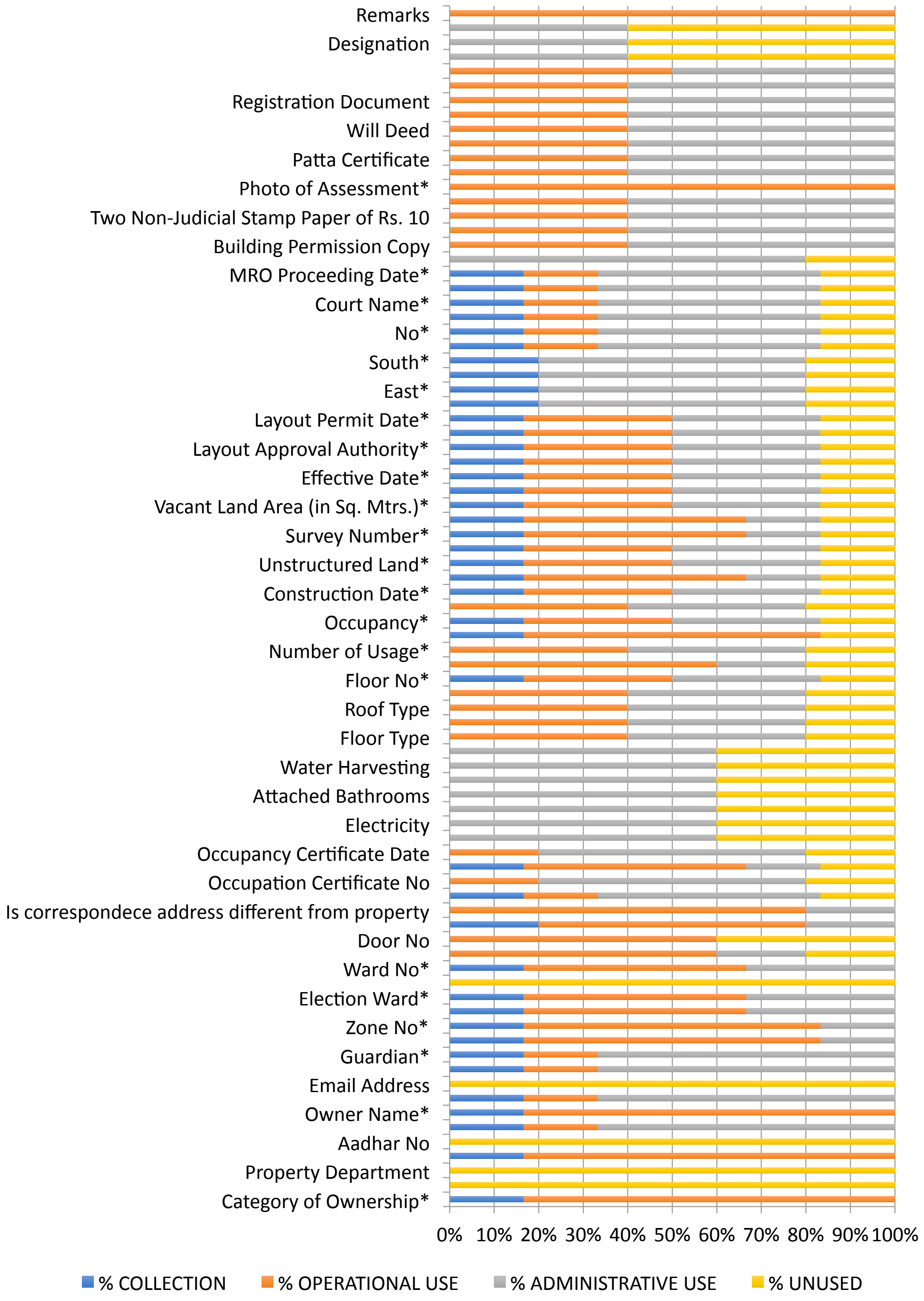
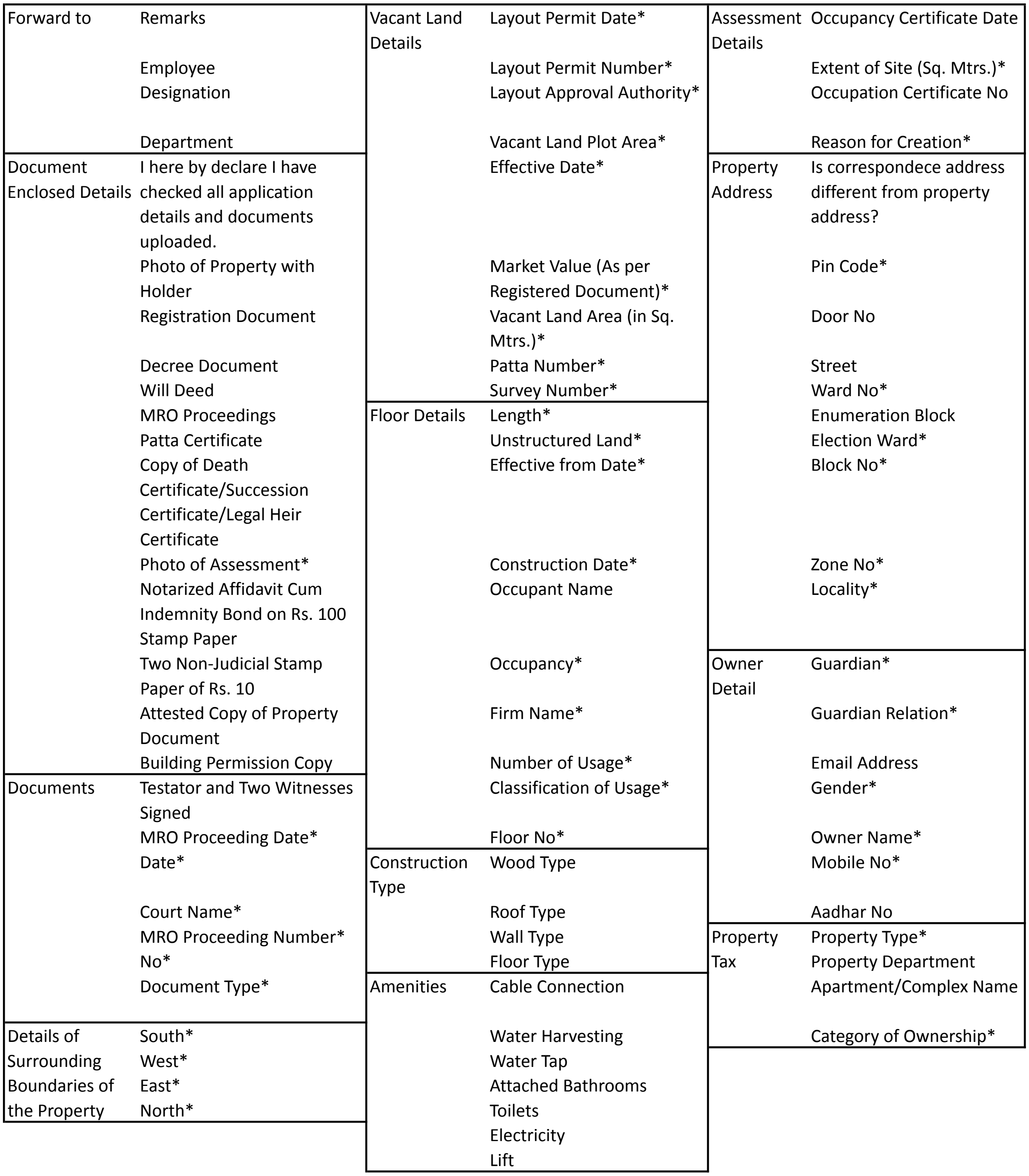


Figure 3: Property Tax Data Usage: Complete Y-axis labels in Figure 4



Forward to	Remarks
	Employee
	Designation
	Department
Document Enclosed Details	I here by declare I have checked all application details and documents uploaded.
	Photo of Property with Holder
	Registration Document
	Decree Document
	Will Deed
	MRO Proceedings
	Patta Certificate
	Copy of Death Certificate/Succession Certificate/Legal Heir Certificate
	Photo of Assessment*
	Notarized Affidavit Cum Indemnity Bond on Rs. 100 Stamp Paper
	Two Non-Judicial Stamp Paper of Rs. 10
	Attested Copy of Property Document
	Building Permission Copy
Documents	Testator and Two Witnesses Signed
	MRO Proceeding Date*
	Date*
	Court Name*
	MRO Proceeding Number*
	No*
	Document Type*
Details of Surrounding Boundaries of the Property	South*
	West*
	East*
	North*

Vacant Land Details	Layout Permit Date*
	Layout Permit Number*
	Layout Approval Authority*
	Vacant Land Plot Area*
	Effective Date*
	Market Value (As per Registered Document)*
	Vacant Land Area (in Sq. Mtrs.)*
	Patta Number*
	Survey Number*
Floor Details	Length*
	Unstructured Land*
	Effective from Date*
	Construction Date*
	Occupant Name
	Occupancy*
	Firm Name*
	Number of Usage*
	Classification of Usage*
	Floor No*
Construction Type	Wood Type
	Roof Type
	Wall Type
	Floor Type
Amenities	Cable Connection
	Water Harvesting
	Water Tap
	Attached Bathrooms
	Toilets
	Electricity
	Lift

Assessment Details	Occupancy Certificate Date
	Extent of Site (Sq. Mtrs.)*
	Occupation Certificate No
	Reason for Creation*
Property Address	Is correspondece address different from property address?
	Pin Code*
	Door No
	Street
	Ward No*
	Enumeration Block
	Election Ward*
	Block No*
	Zone No*
	Locality*
Owner Detail	Guardian*
	Guardian Relation*
	Email Address
	Gender*
	Owner Name*
	Mobile No*
	Aadhar No
Property Tax	Property Type*
	Property Department
	Apartment/Complex Name
	Category of Ownership*

Figure 4: Property Tax Data Categories and Fields: The Y-axis labels for Figure 3

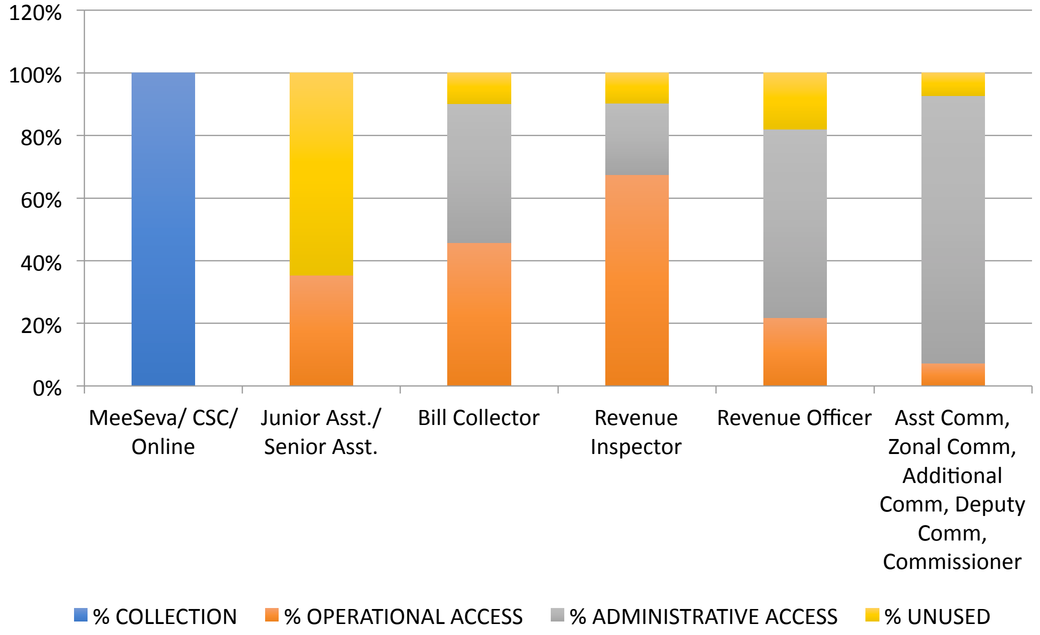


Figure 5: Property Tax Data Access

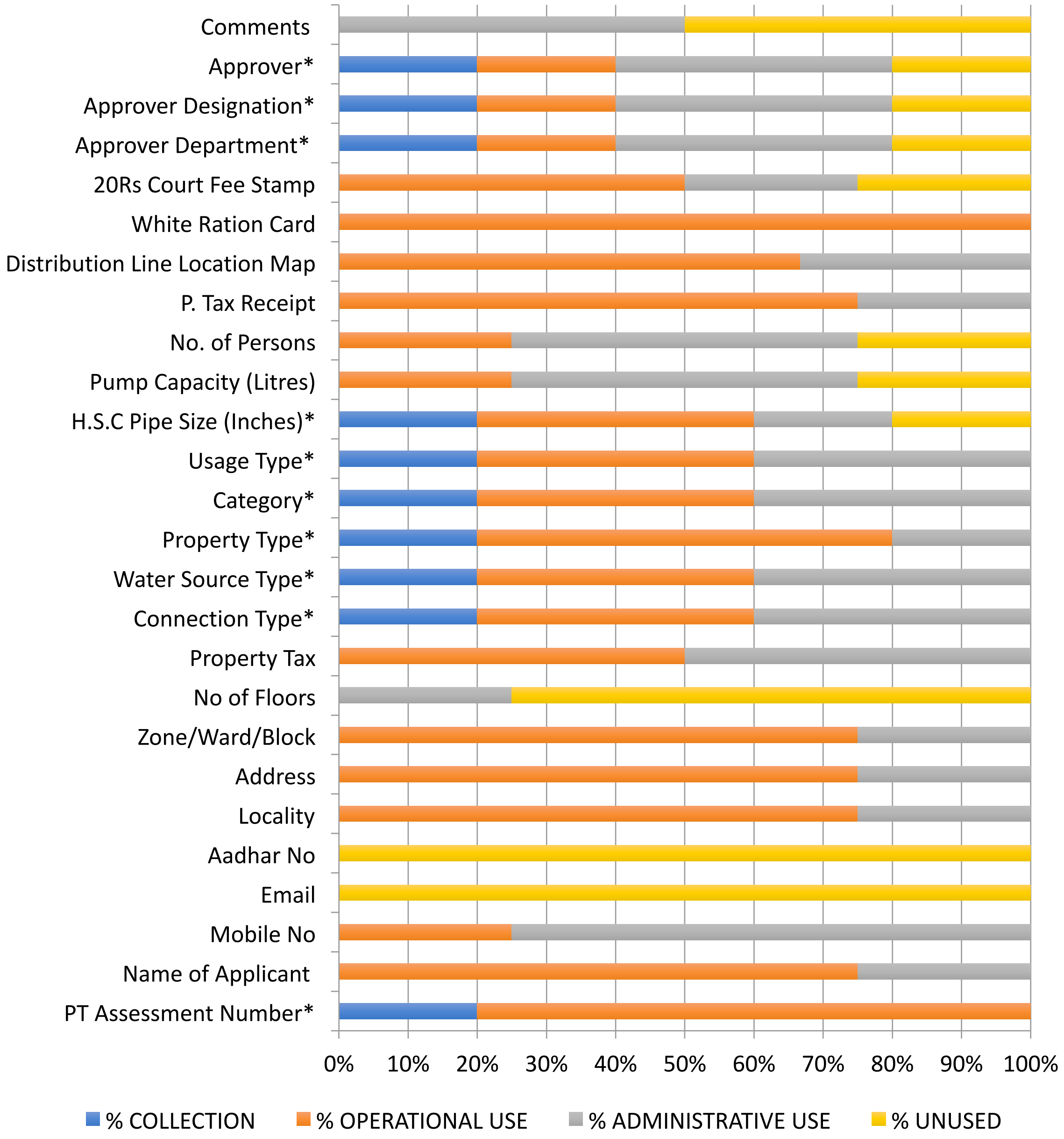


Figure 6: Water Data Usage

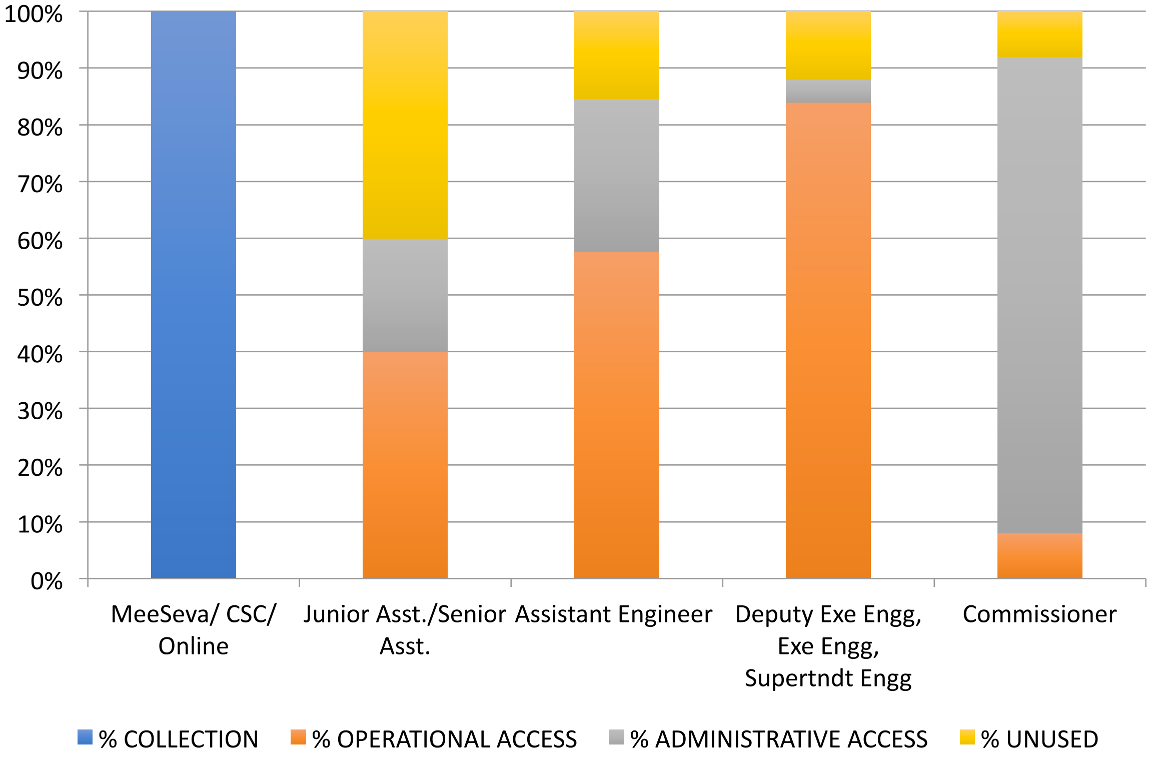
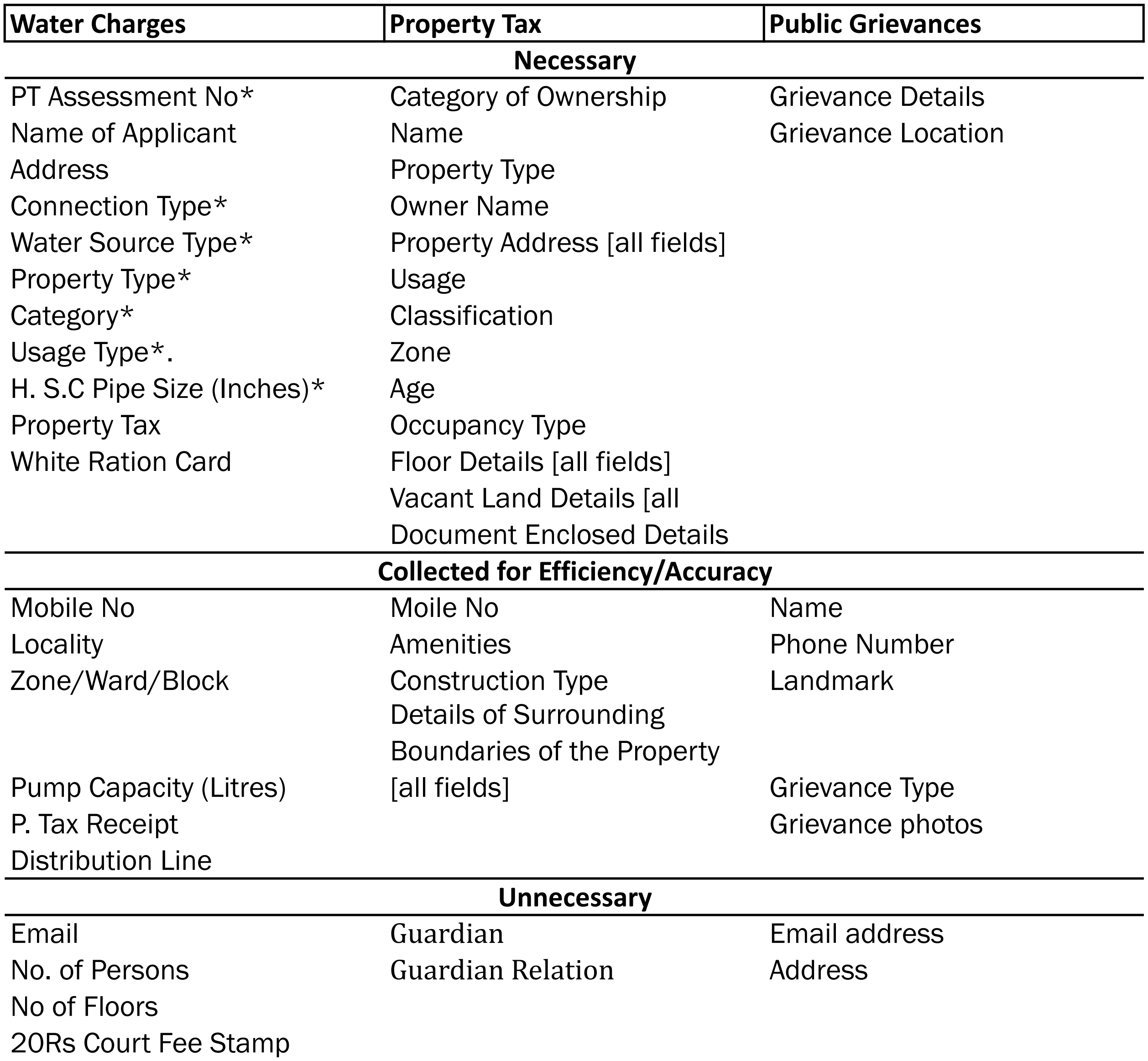


Figure 7: Water Data Access



Water Charges	**Property Tax**	**Public Grievances**
	Necessary	
PT Assessment No*	Category of Ownership	Grievance Details
Name of Applicant	Name	Grievance Location
Address	Property Type	
Connection Type*	Owner Name	
Water Source Type*	Property Address [all fields]	
Property Type*	Usage	
Category*	Classification	
Usage Type*.	Zone	
H. S.C Pipe Size (Inches)*	Age	
Property Tax	Occupancy Type	
White Ration Card	Floor Details [all fields]	
	Vacant Land Details [all	
	Document Enclosed Details	
	Collected for Efficiency/Accuracy	
Mobile No	Moile No	Name
Locality	Amenities	Phone Number
Zone/Ward/Block	Construction Type	Landmark
Pump Capacity (Litres)	Details of Surrounding Boundaries of the Property [all fields]	Grievance Type
P. Tax Receipt		Grievance photos
Distribution Line		
	Unnecessary	
Email	Guardian	Email address
No. of Persons	Guardian Relation	Address
No of Floors		
20Rs Court Fee Stamp		

Figure 8: Data fields sorted by utility for Water Charges, Property Tax, and Public Grievence modules

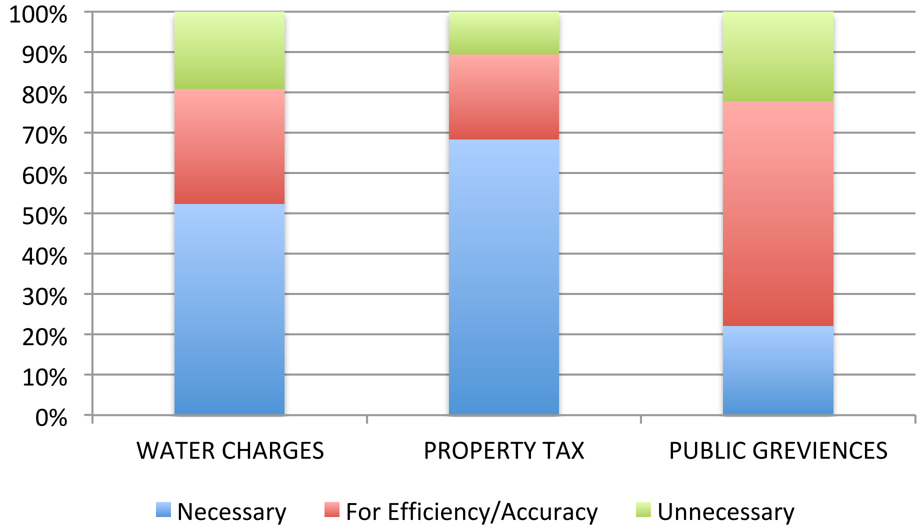
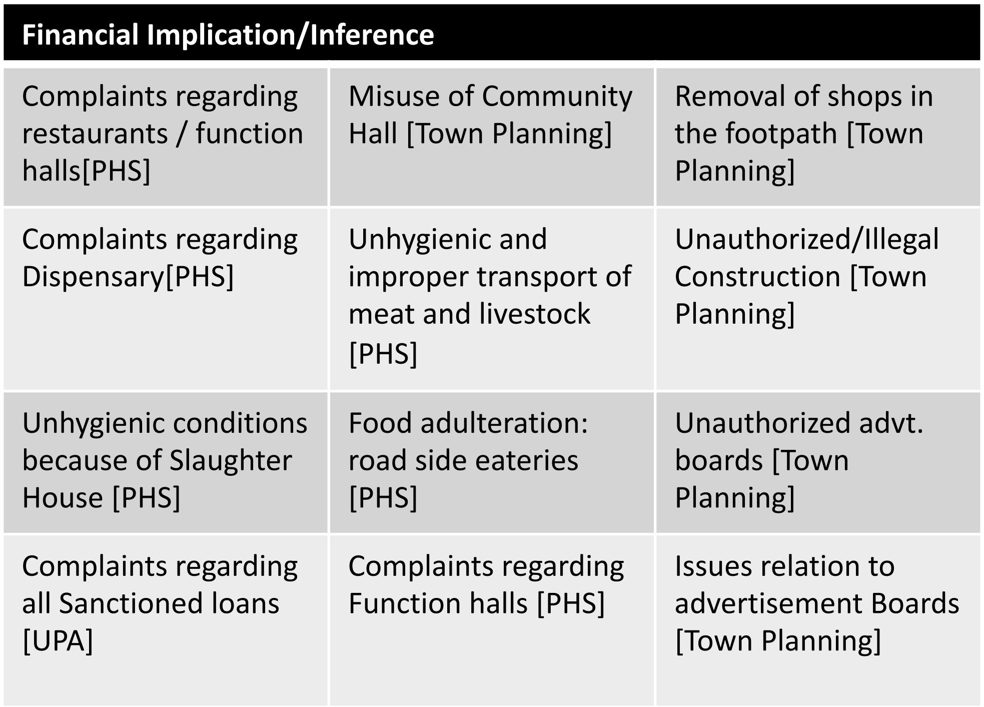
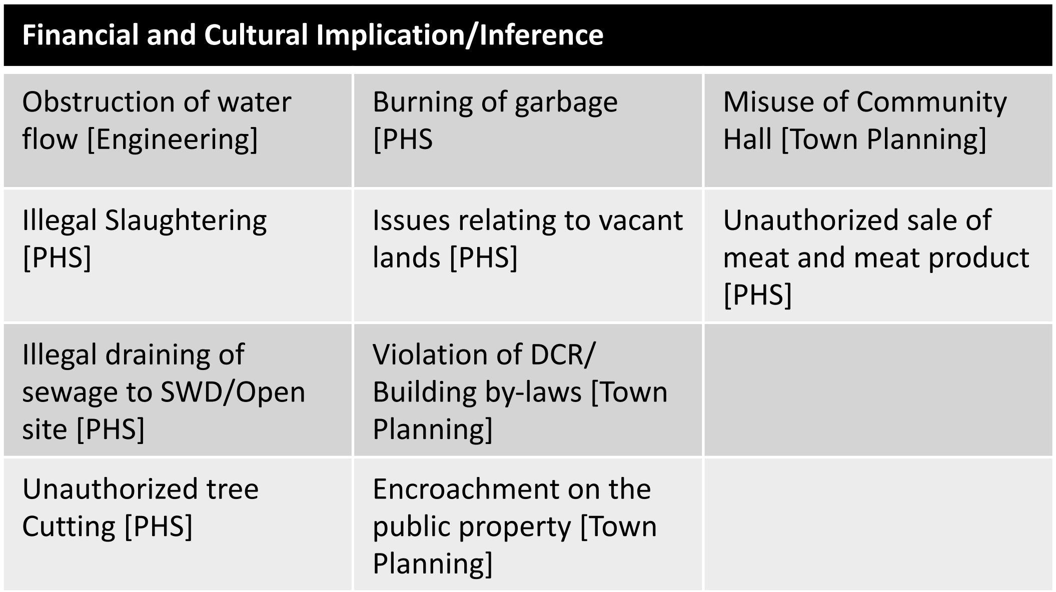


Figure 9: Overall Data Utility across Water, Property Tax and Public Grievances



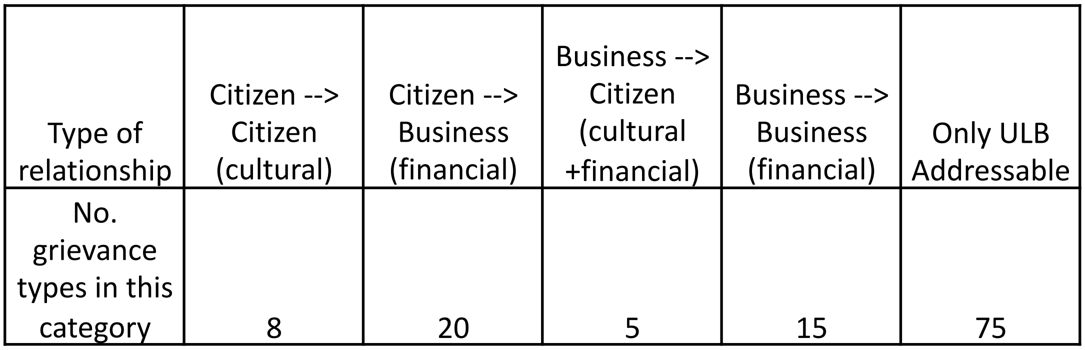
Financial Implication/Inference		
Complaints regarding restaurants / function halls[PHS]	Misuse of Community Hall [Town Planning]	Removal of shops in the footpath [Town Planning]
Complaints regarding Dispensary[PHS]	Unhygienic and improper transport of meat and livestock [PHS]	Unauthorized/Illegal Construction [Town Planning]
Unhygienic conditions because of Slaughter House [PHS]	Food adulteration: road side eateries [PHS]	Unauthorized advt. boards [Town Planning]
Complaints regarding all Sanctioned loans [UPA]	Complaints regarding Function halls [PHS]	Issues relation to advertisement Boards [Town Planning]

Figure 10: Examples of Public Grievances types for which loss of confidentiality leads to financial loss



Financial and Cultural Implication/Inference		
Obstruction of water flow [Engineering]	Burning of garbage [PHS	Misuse of Community Hall [Town Planning]
Illegal Slaughtering [PHS]	Issues relating to vacant lands [PHS]	Unauthorized sale of meat and meat product [PHS]
Illegal draining of sewage to SWD/Open site [PHS]	Violation of DCR/ Building by-laws [Town Planning]	
Unauthorized tree Cutting [PHS]	Encroachment on the public property [Town Planning]	

Figure 11: Examples of Public Grievances types for which loss of confidentiality leads to financial and cultural loss



Type of relationship	Citizen --> Citizen (cultural)	Citizen --> Business (financial)	Business --> Citizen (cultural +financial)	Business --> Business (financial)	Only ULB Addressable
No. grievance types in this category	8	20	5	15	75

Figure 12: Loss of confidentiality can be to either the complainer or the subject; the relationship is related to whether the loss is financial or cultural.