Modeling the Right to Information Query Log: Learning Latent Parameters to Identify Amendment Scopes in the RTI Act

in partial fulfillment for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Nayantara Kotoky

Under the supervision of

Dr. Vijaya Saradhi Vedula

Department of Computer Science and Engineering

Indian Institute of Technology Guwahati

Guwahati - 781039 Assam India

September 2020

Dedicated to my family.

Acknowledgements

I am heartily thankful to my supervisor, Dr. Vijaya V. Saradhi for giving me the opportunity to work under him and providing me ample guidance and support through the course of this study. I thank my Doctoral Committee Members Dr. Sanasam Ranbir Singh, Dr. Hemangee Kapoor, Dr. Rashmi Dutta Baruah and Dr. Sawmya Ray for providing me valuable inputs with regards to my thesis work.

I am also thankful to my M.Tech thesis supervisor Dr. Shyamanta M. Hazarika, who gave me the first taste of research, and encouraged me to go for higher studies.

I would also like to thank all those who supported me in any manner during the course of my studies. My co-workers Praveen Kolla and Tejasri Yedulapuram helped me in my initial years during data collection and idea brainstorming. The data processing part was tough, and several people helped me during this time. I thank Pallabi Saikia, Rupam Sharma, Swarup Ranjan Behera, Saroj Snehal Shivagunde and couple of other friends who helped me in my data processing stage. It would be remiss of me to not mention the support I received from the department of Computer Science and Engineering, IITG, and their officials for providing me the necessary resources that directly or indirectly led to the completion of my thesis. Thank you Gauri, Monojit Da, Pranjit Da, Raktajit Da, Nanu Da, Bhrigu Da and Naba Da for lending me the required help in my time of need.

Lastly, I come to the list of people who made this place livable inspite of the violent ups and downs that any researcher goes through. Thank you my 'evening tea' friends, Shilpa di, Dipika, Hema, Basant and Sukarn for guaranteeing my happiness atleast once per day. A huge thank you to my seniors who provided me relevant counselling during the course of my doctoral studies, Shirshendu Da, Mayank Bhaiya, Shashi Bhaiya from whom I asked many academic materials and suggestions at random times, and they never refused. Some of my friends mentioned earlier have transitioned into seniors, but I shall keep you as friends because the most relevant and beautiful memories are when you were here with me in IITG.

To those whom I have not mentioned here, I shall thank you personally.

And to my friends-cum-sisters, who I met after I left home, Bandita, Jayshree and Uddipana (listed alphabetically!), you are an asset for a lifetime. The lush green campus of IITG became even more attractive to me because you three were present there. You guys are completely mine. Me entering this institute, going out with a job in my hand and living my life with you guys, no one can take that credit away from me.

30th September 2020 Nayantara Kotoky

Declaration

I certify that

- The work contained in this thesis is original and has been done by myself and the general supervision of my supervisor.
- The work has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- No part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.

30th September 2020 Nayantara Kotoky

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Dr. V Vijaya Saradhi
Associate Professor
+91-361-2582367
saradhi@iitg.ac.in

Certificate

This is to certify that the thesis entitled **Modeling the Right to Information Query Log: Learning Latent Parameters to Identify Amendment Scopes in the RTI Act** being submitted by **Nayantara Kotoky** to the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, is a record of bona fide research work under my supervision and is worthy of consideration for the award of the degree of Doctor of Philosophy of the Institute.

Date: 30th September 2020
Place: Guwahati

Dr. Vijaya Saradhi Vedula

Publications

- **Kotoky, Nayantara** and Vijaya V. Saradhi. (2019): Learning to Propose Amendments: Identifying Patterns in the Right to Information Query Log. *18th IEEE International Conference on Machine Learning and Applications - ICMLA 2019, Boca Raton, Florida, USA.*
- **Kotoky, Nayantara** and Vijaya V. Saradhi. (2019): Analysis of Right to Information Query-length. *Grace Hopper Celebration India 2019, Bangalore, India.*
- **Kotoky, Nayantara** and Vijaya V. Saradhi. (2018): Analysis of Right to Information Query-length. *3rd Indian Workshop on Machine Learning, IIT-BHU, Uttar Pradesh, India.*
- **Kotoky, Nayantara.** (2017): Predicting Amendments via Right to Information Query Log Analysis. *CSE Doctoral Symposium, NIIT University, Neemrana, Rajasthan.*
- **Kotoky, Nayantara** and Vijaya V. Saradhi. (2017): Modelling Right to Information Query-Reply Data to Uncover Latent Patterns. *Interactive Machine Learning Workshop@ICML 2017, Sydney, Australia.*
- **Kotoky, Nayantara.** (2017): Right to Information Query Analysis for Predicting Amendments. *Doctoral Consortium, Advances in Information Retrieval: 39th European Conference on IR Research, ECIR, Aberdeen, Scotland.*
- **Kotoky, Nayantara** and Vijaya V. Saradhi. (2016): Right to Information Query Modelling via Graded Response Model. *SoGood-Workshop on Data Science and Social Good@ECML-PKDD, Riva Del Garda, Italy.*
- **Kotoky, Nayantara** (2016): Predicting Amendments via Right to Information Query Analysis. *IDC-2016@IDRBT, Hyderabad, India.*
- **Kotoky, Nayantara**, Kolla, Praveen Kumar and Vijaya V. Saradhi. (2016): Modelling Right to Information Queries via Item Response Theory. *2nd Indian Workshop on Machine Learning, IIT Kanpur, Uttar Pradesh.*

Abstract

Amendments to laws are necessary to keep up with the changing needs of the society. Such a process is largely manual, and policy-makers take feedback from the society for the introduction of an amendment. These include statistics and other observations taken into account by law makers during the amendment process. The Right to information (RTI) Act 2005 gives Indian citizens the opportunity to interact with the government. Any Indian citizen can question public authorities and demand information from them. Evidence suggests that the RTI query-reply process has also sown seeds towards the proposal as well as the introduction of amendments to the RTI Act. This direct interaction between citizens and the government have *latent information* that leads towards potential amendment scopes. The presence of such pointers encourages us to think that tentative amendments can be proposed by a learning process, and that feedback for potential amendments can be identified by an extensive analysis of the database of RTI queries and their reply-statistics.

The objective of this thesis is to analyse RTI query and reply statistics data to mine latent patterns that act as feedback for potential amendments to the existing RTI laws. We have collected RTI applications containing queries of citizens. From this, relevant RTI properties are identified and their definitions and indicators are studied from the literature. These definitions are used for proposing data models, and learning models are used to quantify RTI parameters. The quantified RTI parameters obtained in the course of this thesis reveal patterns that suggest several scopes for potential amendments in the RTI laws.

In the first work, distributional analysis of RTI data is performed. Two properties, namely, RTI query-length (word count of an application) and query-reply-time (number of days in replying to an RTI query) are used. A methodical approach to fit power-law distribution is adopted. We observe that (i) RTI applications with larger word count are rare, yet they take large duration for getting replied to (ii) the RTI query-length data do not follow a single distribution, but the data follow a mixed model distribution for different portions of the query-length data. With this analysis *we uncover information that is present in one of the draft of the RTI Act 2012 amendment by the government.* With the query-reply-time analysis, we estimate the probability of getting a reply to RTI queries within the stipulated 30-day time-limit given a query-category. A query-category represents the different departments within a government educational institution. With this analysis, (i) we quantify 'transparency of query-categories' across India (ii) considering the threshold of 75% we observe that only three query-categories are transparent, that is, the probability of getting a reply within 30-days (time-limit set by the RTI Act) for only three query-categories falls above 75%. This shows that RTI reply durations are different based on the type of query citizens ask.

In the second work, RTI data is used to quantify three latent parameters from the RTI reply rates of institutions. For the first time we quantify 'transparency of institutions', 'discriminative power of query-categories that affect transparency of institutions' and 'difficulty of a query-category'. A new data model in the form of an institute-query-category matrix is proposed, consolidating several definitions of relevant RTI properties. The data matrix

undergoes psychometric analysis using Graded Response Model to estimate the values of the three parameters using Maximum Likelihood Estimation. The quantified values for institutions and query-categories reveal that (i) institutions have different transparency values across India (ii) different query-categories have different discriminating power influencing the transparency of institutions (iii) we quantify the 'difficulty' of a query-category in replying to RTI queries. Differing difficulty levels of query-categories indicate that that the provisions of the RTI Act are not implemented effectively throughout the different departments within the institutions. This leads us to a tentative amendment scope on the RTI Act where we propose to use the 'difficulty' value obtained from our analysis to compute the reply-deadlines for separate RTI query-categories.

The third work models the temporal variations in the RTI query-reply data. The objective is to capture time-intervals involving high variations in the reply-rates; these fluctuations are indicative of a corresponding change in the input parameters that affect the RTI performance. The inputs for the RTI system are facilities such as PIOs, and their resources are the precursors to having a smooth query-reply experience for citizens. Any change in the PIO appointments and/or associated rules and provisions bring significant changes in the reply-rates for institutions. This in turn will also bring changes to the quantified RTI parameters computed from these fluctuating reply-rates. We propose a three dimensional tensor representation of RTI data containing institutions, query-categories and time in the three dimensions respectively. Using tensor-CUR decomposition, two feature-matrices capturing high variations in the RTI reply rates are obtained and used for psychometric analysis to quantify the parameter 'transparency of institutions'. Comparison of this parameter with transparency computed without time-modeling reveals latent patterns that occurred in the input factors of the RTI system, thereby providing us a peek at the changing input facilities as viewed from the corresponding changes in the output reply rates data. We observe (i) there is a clear segregation between central and state institutions in the most fluctuating year, where all the state institutions have lower values of transparency (ii) this indicates that there is some change in the input rules governing the state institutions that collectively affected their transparency negatively. Identification of this changed input by the policy-makers that is the source of the low transparency of state institutions provides scope for amendments.

Contents

List of Figures

List of Tables

1

Introduction

Right to information (RTI) laws enable citizens to question government establishments and seek answers from the government officials. Known by different names such as Access To Information (ATI) or Freedom of Information Act (FOIA), these laws in varying forms are enacted in over 100 countries [3, 4, 5]. A large number of countries passed the RTI laws and brought them into enforcement between the years 2000 and 2009. Finland is the first country to pass this law in 1951 and Seychelles being the country passing this law as recently as 2018. This fact can be attributed to one among several factors namely information revolution that took place between these years. Seeking information from the government establishment is viewed as a fundamental right by various bodies: executive, judiciary and legislative. Also, one of the factors that account for good governance is achieving transparency. The genesis for the RTI Act is to achieve transparency in every government establishment in the country.

In India, the RTI Act was introduced in the year 2005. The central information commission (CIC) is responsible for monitoring the RTI Act's effective implementation across the country. The jurisdiction of the CIC extends to all the *central* public authorities (PAs[1]). In the recent reports produced by the CIC [6] it is noted that over 10 million RTI queries are answered by all government institutions in total by the end of the 2018-19 year spanning over 14 years. At the time of inception, the total number of RTI applications received is 24436. At the end of the year 2018-19, the number of RTI applications received by all the PAs is increased by a factor of 66. During the 2018-19 year, the highest number of RTI queries were received. The year-wise trend in the received applications is given in the Figure 1.1. Increase in the number of RTI applications indicates vibrant engagement of citizens with the government organizations for meeting their information needs. The number of RTI applications has witnessed a significant increase in every ministry or department. As an example, the ministry of human resource development (Ministry of HRD) under which all the centrally funded government educational institutions comes is presented in Figure 1.1. We focus on this ministry as the RTI applications data collection is focused on educational institutions alone, some of which comes within this ministry.

The voluminous queries surprisingly are read, analyzed for information need and answered

[1] any authority or institution of self government established

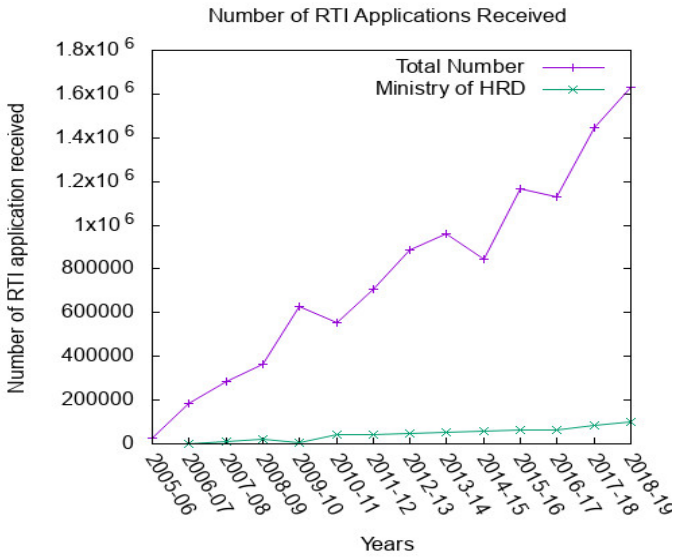


Figure 1.1: Year wise number of RTI applications received by PAs

by manual procedures. The PA to which the RTI queries are addressed is responsible for compiling the answers. PA in turn has an internal governing structure which replies to the applicant. The number of PAs also increased in the span of 14 years. This suggests the efforts made in complying with the RTI Act across India. This also indicates that the implementation is done in phases and in every phase the focus is on covering new PAs. Figure 1.2 shows an increase in the trend of the number of PAs. At the time of inception of the RTI Act, this number was 938; in the year 2018, this number is increased to 2145.

One of the responsibilities of the CIC is to ensure the *effective implementation of the RTI Act* in the PAs. The CIC is interested in a variety of statistics from each of the PAs. Towards this, CIC compiles an exhaustive list of 30 parameters using which the PA will be evaluated in achieving implementation rigour of the RTI Act. These are summary statistics of the RTI applications received by the PAs and associated reply statistics.

Also, CIC conducts what is known as *transparency audit*. In this audit, it has acknowledged that *transparency is a much broader and deeper concept and multidimensional too, which cannot be limited by any straitjacket*. The main objective of the audit is to make a (qualitative and quantitative) assessment on the degree of voluntary disclosure of information by PAs.

The audit carried out on the websites of the PAs and the content of the website was assessed on six broad dimensions (along with weightages) namely

1. Organization and Function (10%)
2. Budget and Programmes (30%)
3. Publicity and Public Interface (25%)

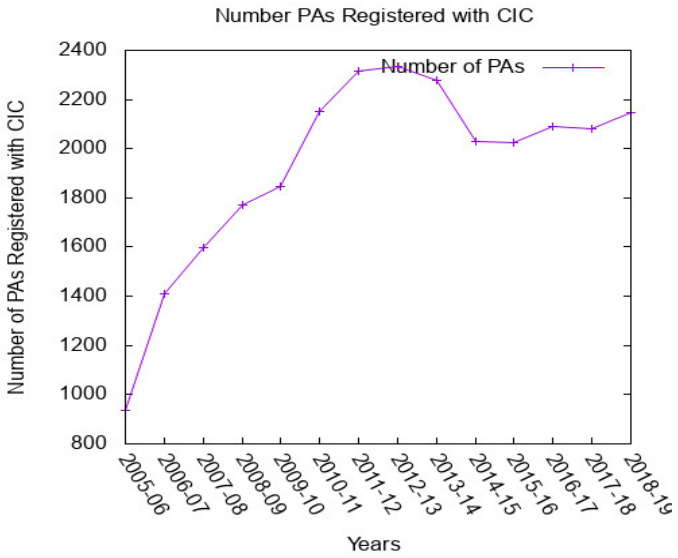


Figure 1.2: Year wise number of PAs registered with CIC

4. E-Governance/Digitization (20%)
5. Information as prescribed (10%)
6. Information disclosed on own initiative (5%)

The audit studied 838 PAs and categorized them into two classes, PAs *Meeting the requirement* and PAs *partially meeting the requirement*. Based on the levels of disclosure, each participating PA is scored and given a letter grade {A, B, C, D, E}. The transparency audit report and the details of the evaluation and scoring mechanism are available at [7]. Individual institutions are given transparency grades through this audit. In this audit, the Prime minister's office secured A grade (highest grade). The ministry of HRD received C grade. Two limitations with the transparency audit (among other limitations) are (i) the evaluation was confined to websites of the PAs and (ii) Only Section 4 of the RTI Act was subject to the audit.

In the literature, obtaining composite indexes for transparency is of value in motivating compliance [8, 9]. The international transparency policy index (ITPI) measured using 102 countries draw significant impact on investors to invest in other countries to take into account these indices. These ratings are computed for each country by measuring 61 indicators [10]. Indicator based computation for transparency index have been questioned by researchers in terms of generating representative results [9].

From the above discussion it is to be noted that

1. Measuring transparency is of importance.
2. There is no specified method that exists for measuring transparency.

3. Quantifying effectiveness of the RTI Act's implementation is required for achieving the very objectives of the RTI Act.

4. CIC focuses on the summary statistics alone. However, given the rich resource of RTI applications text data, analyzing the RTI query-log is of immense value. In this thesis work, we place efforts on analyzing the RTI applications text data.

It is to be noted that the above efforts are an informal way of obtaining transparency grade or transparency index that uses a fixed set of dimensions and assigning weights for those dimensions. In addition, these indexes are computed at a coarse level that is either at the ministry level or at the country level. There is no method that exists to compute at a fine grain level that is at an individual government institute level.

The focus of this thesis is on

1. Identifying a *computational model* for obtaining *individual government institutes* transparency values. The computational model should be independent of the fixed weightage given to obtain the transparency values. The model should consider multiple dimensions in obtaining a single value that accounts for transparency.

2. Quantitatively characterize *the effectiveness of RTI Act implementation*. The quantitative model once again should be free from user-provided input weightages.

3. Obtain transparency values at fine-grain level that is obtaining transparency values of individual institutes as opposed to a coarse level where transparency values are computed at the ministry/country level.

4. To compute these values, RTI applications text data received by the institutes are collected and examined instead of summary statistics.

5. Designing a data model for the RTI applications text data that is readily consumable for the computational model.

6. Computational model output when interpreted in the RTI Act context lead to identifying potential amendments to the RTI Act itself.

1.1 Contributions

Following are the key contributions of the thesis work:

1.1.1 RTI Applications Data Collection

To compute the transparency of individual institutes, we collect RTI application data from *government educational institutes* from the date of inception of the RTI Act till 01-Jan-2015. A detailed description of the data collection is presented in Chapter 3. The data collection process spanned over *1.5 years*. As the RTI query-reply procedure is manual the data collection took significant time and efforts.

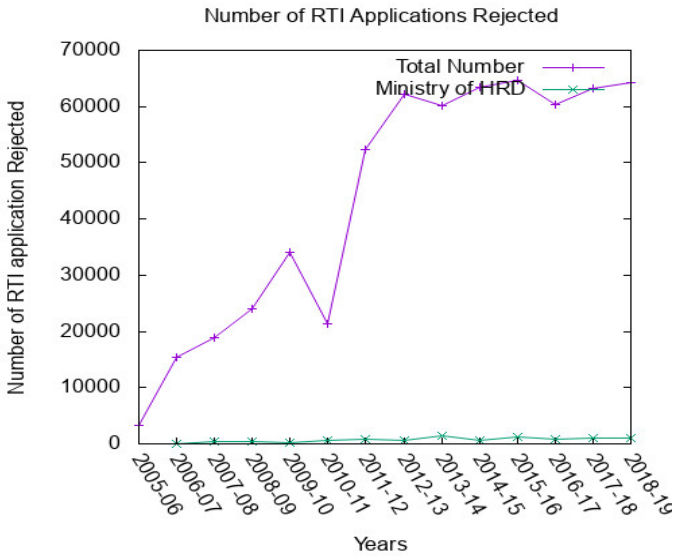


Figure 1.3: Year wise number of RTI applications rejected by PAs

The collected RTI application data is subject to analysis in three different perspectives namely

1. RTI query log distributional analysis.
2. Computational model for obtaining transparency of individual institutions and effectiveness of RTI Act's implementation.
3. Temporal variations of RTI query-reply data.
4. In the first two analysis potential amendments are identified.

1.1.2 RTI Query Log Distributional Analysis

Two query-log quantities are analyzed by fitting probability distribution. For the first time, we perform distributional analysis on RTI query log data which is viewed as *offline* and *paid queries* which are posed to the PAs (Information Retrieval system). Following research questions are investigated:

RQ1 Which distribution(s) empirically fits the query-length given the RTI query-log data? In addition, is the empirical distribution comparable to that of the web query-log based empirical distribution observed in the literature [11]?

The word-count of each RTI application is computed and the vector is used for distributional analysis. This is done to compute the probability of occurrence of RTI applications having large word counts.

RQ2 Which distribution(s) empirically fits the query-reply-time given the RTI query-log data?

RTI queries (individual sub-queries within an RTI application) are categorized into 26 query-categories, and the reply-times for each query is computed. These 26 vectors are subjected to distributional analysis and the estimated parameters are used to compute the *probability of getting a reply to an RTI query within the stipulated 30-day time-frame for a given query-category.*

RQ3 How the distributions and associated parameters are interpreted as amendments?

The distributional fit and information pertaining to the reply time are taken into account and interpreted to propose a potential amendment to limit the number of words in the RTI application to 500. This is validated using the ground truth. The government of India has suggested that RTI applications should not exceed 500 words.

Data Representation:

The query-length and query-category reply time data are represented as one-dimensional vectors. The vectors then undergo distributional analysis.

Learning Algorithm: Broadly, the steps listed below are followed to achieve the distributional analysis:

1. Estimate the parameters of power-law distribution given the dataset.
2. Examine the estimated parameters to ascertain the distributional fit.
3. When power-law distribution fits the RTI query log data, the fit is then compared to other alternate distributions to test whether the fitted power-law is the best fit.
4. In case power-law distribution is not accepted, seven other candidate distributions are fitted and their goodness-of-fit tested.

For the 26 query-category reply-time data vectors, the estimated parameters are used to quantify *transparency* of query-categories based on the probability of reply to a given query-category within 30-days.

Results and Observations:

1. The tail end of the RTI query-length data follows power-law. This means that as word-count of applications increase, the probability of occurrence of such applications decreases. It is also observed from the collected data that such applications with high word count, although few and far between, have high reply-time. With this finding, the query-length modeling is in favor of a potential amendment to *limit RTI word-count to 500 words per application.* Since large applications have very high reply time, avoiding the filing of such large applications will reduce the overall reply time, thus improving transparency.

2. A mixed-model distribution fit is obtained for RTI query-length data that is not at the tail of the distribution. It is experimentally observed that RTI word-count does not follow a single distribution for the entire query-length range. Different ranges of the data follow different distributions.

3. For the query-category reply time data, 19 query-categories follow power-law distribution, two query-categories follow Burr's distribution and five query-categories follow log-logistic distribution.

4. Only three query-categories have probability of reply above 75% within 30-days time to RTI queries. This indicates a gap in implementation of the RTI Act at individual institute level.

1.1.3 Computational Model for Quantifying Transparency of Individual Institutes

The quantity estimated in the query-category reply-time analysis is the *probability of replying within 30 days for a given query category*. A higher value of this probability is an indication of *transparency*. The definition of transparency is modeled in a limited way through the distributional analysis. The quantity of interest is the probability of getting a reply for a given query-category from a *particular institution*. To model the transparency, the above probability is expressed in terms of transparency and effectiveness of the RTI Act's implementation. The estimated values are used to obtain the above probability. In particular, the following research questions are investigated:

RQ1: How to quantify transparency and effectiveness of the RTI Act? Quantifying transparency is of immense value to the policymakers. In particular, such quantitative measures help in attracting foreign investments [9].

RQ2: How to validate the obtained estimates? The psychometric models provide the best estimates for the latent parameters given the RTI query reply statistics data. However, the goodness of the estimated parameters needs to be examined for further usage of the identified latent values.

To address the above research questions, the graded response model (GRM) is employed which estimates *the probability that an institution replying to a query-category within 30 days*. Using the GRM, three latent variables are estimated

1. *Transparency* of individual institutes.

2. *Discriminative power of query-categories* to quantify their influence on the transparency of institutions.

3. The *difficulty* of a query-category is computed to quantify the effectiveness of the RTI Act's implementation across the country.

Data representation: A two-dimensional matrix representation called the institute, query-category (IQC) matrix is proposed. The representation incorporates multiple definitions of transparency and effectiveness of implementation of the RTI Act as found in literature from the RTI applications data.

Learning Model: The above three latent parameters are estimated from the IQC data matrix using GRM. Given the IQC matrix as input, the GRM yields the above three latent parameters.

Results We highlight some of the obtained results here.

1. Using the IQC matrix as input to the GRM, transparency of individual institutes are computed. The method alleviates the limitations presented in the index-based transparency score computation.

2. Administration query-category (that is, queries posed to administration department/section) is a highly discriminatory query-category. It means that when queries are asked regarding this department, it leads to high deviation of the transparency values of institutions.

1.1.4 Modeling Temporal Fluctuations in the RTI data

The collectible RTI statistics are the *output-oriented* indicators which include the reply rates of institutions, reply duration. The output-oriented indicators are dependent on the *input factors* namely public information officer (PIO) who is responsible for replying to the RTI applications. In this experiment, we consider the influence of the *inputs* on the *output parameters*. The following two research questions are explored in this contribution:

RQ1: Which time of the year maximum variation in the reply rates are observed?

RQ2: Which of the query-categories exhibit maximum variation in reply rates?

Data Representation: IQC matrices are created for a specific time interval (1 year). Several IQC matrices are stacked together to form an RTI tensor, which is the key data representation. The tensor consists of institutions, query-categories and time (in years) as its three dimensions.

Time model: By stacking IQC matrices for different time-intervals, the RTI tensor in the third dimension represents the temporal trends in the RTI query-reply dynamics.

Learning Model: The RTI tensor is subjected to tensor decomposition to find "Which time of the year maximum variation in the reply rates are observed?". This decomposition yields an IQC_{year} matrix which contains maximum variation in reply rates. This matrix is subjected to psychometric modeling using GRM to obtain the transparency values when maximum variation in reply rates are observed in the obtained time interval.

In addition, the RTI tensor is subject to decomposition to find "Which of the query-categories exhibit maximum variation in reply rates?". This decomposition yields $IT_{query-category}$ matrix which contains maximum variation in reply rates across *query-categories*. This matrix

is subject to psychometric modeling using GRM to obtain the transparency values when there is maximum variation in the reply rate observed in a query category.

Results and Observations:

1. The year 2010 is identified as the most fluctuating year with respect to RTI reply rates. 'Administration' is the most fluctuating query-category across all times.

2. Without time-modeling, central and state institutions were mixed up on the transparency scale. However, transparency values with time modeling show that for the year 2010 the state institutions have low transparency values and are segregated from the central institutions, which have higher transparency values for that year.

The following are the highlighting points in this thesis:

- For the first time, we analyzed RTI Act data to *quantify* latent parameters namely, transparency and effectiveness of implementation.

- New data models which are tailored to capture the quantities of interest are proposed using RTI applications raw data in all the three contributions namely, distributional analysis, latent variable modeling and capturing temporal fluctuations.

- Psychometric analysis which is popular in the test-question analysis domain is adopted to model transparency of individual institutes. The applicability of this model in the RTI Act context is not a straight forward one. We made an intuitive choice in the data model in order to apply the psychometric model.

- Computation of quantities such as transparency of query-categories, transparency of institutions, the discriminative power of query-categories and quantification of effectiveness of the implementation of the Act across India have been performed for the first time.

- This thesis has computationally shown prevalent issues and opportunities for improvement in the RTI Act unlike the qualitative studies present in the literature.

- All the models identified and applied have the interpretability characteristic. This allowed us to identify scopes for potential amendments.

1.2 Thesis Organization

This thesis consists of eight Chapters.

Chapter 1: This Chapter introduces the research objective. This Chapter provides an overview of different analysis performed on the RTI dataset, the data model used for the analysis and the highlights of the obtained results are presented.

Chapter 2: This Chapter provides the history of the RTI Act 2005, the procedure for filing an RTI application, amendments that were included in the RTI Act due to analysis of RTI query log summary statistics and challenges working with the RTI data are presented.

Chapter 3: This Chapter exhaustively details the data collection drive performed. The collected data and its characteristics are presented.

Chapter 4: As the attempted problem is interdisciplinary, exhaustive literature is presented in the RTI Act and its analysis, FOIA and its analysis in the literature. Query-logs in various domains are examined and the analysis carried out for diverse query-logs are presented. Similarities and differences with the various queries and their analysis are drawn from the perspective of RTI query-log data. Lastly, we touch upon the role of technology in proposing policies known as policy 2.0.

Chapter 5: This Chapter presents a distributional analysis of the RTI data. Two quantities, namely, RTI query-length and query-category reply-time are analyzed by fitting probability distributions.

Chapter 6: In this Chapter, transparency of individual institutes is computed using GRM. The three latent parameters as stated above are obtained and associated results are presented.

Chapter 7: This Chapter presents a temporal analysis of the RTI query-reply data using tensor decomposition. Transparency of institutions is computed from temporal factors. The quantified values capture and demonstrate the effects of temporal variations in the RTI query-reply dynamics.

Chapter 8: This Chapter contains the concluding remarks of the thesis. The Chapter summarizes the main outcomes of the thesis and discusses the limitations and future scopes for the presented work.

2

Introduction to RTI

2.1 Introduction

In this Chapter, we detail the history of the RTI Act in Section 2.2. Description of the RTI process of applying to obtain information is presented in 2.3. RTI Act 2005 witnessed amendments which are discussed in Section 2.4 along with the genesis of amendments. Challenges in experimentation with RTI data collection and RTI data analysis are presented in Section 2.5. Summary of this Chapter is presented in Section 2.6.

2.2 Brief History of the RTI Act

Counterparts of the Indian RTI Act can be found in many countries of the world under various names such as Access to Information (Canada, Mexico) and Freedom of Information (USA, United Kingdom, Australia). As obvious as the existence of the Act might sound, the initiation of the RTI Act in India was not a smooth one. There have been hurdles at the beginning, and we owe it to the continued persistence of certain pioneers who led the translation of the citizens demands to information to a Bill and finally to an Act. It has been more than 20 years since the early struggle started, and the present RTI Act 2005 is almost 15 years old.

2.2.1 Initial Struggle for Information

The government records in India were originally classified under official secrets. This meant that there was no way for the people to have access to information about the workings of any PA. Out of the many deviations from the duties of the government and the actual scenario, perhaps the most severe was the non-payment of wages. Whenever citizens wanted to seek information regarding the same, they were stalled at every step and the officials refrained from divulging any information about funds and its use. This was a major provocation for the citizens. This led to the first-ever decision of making demands for access to the public

information and copying of official documents [12].

The Mazdoor Kisan Shakti Sangathan (MKSS) is an association of a group of people that worked with the poor people in Bhim Tehsil, Rajasthan, India. MKSS made a significant effort that led to the first-ever "Jan Sunwayi" a public hearing, in 1990 in the history of India. The execution saw the arrival of two decisions by the citizens:

1. Any citizen from the village should have the right to make photocopies of all bills, vouchers and muster rolls on payment on any work done by the government in their village.

2. Funds embezzled and misappropriated should be recovered from these village officials and politicians, their property attached, and assets frozen and publicly auctioned and that money recovered should be spent back in that same village. No departmental enquiry, no due process of law, no cases to be registered.

2.2.2 Official Acknowledgement of the Right to Information

In April 1995, the chief minister of Rajasthan declared in the state assembly that any citizen has the right to information. This was a pivotal moment in the struggle for the right to information as this was the first official acknowledgement of the people's fundamental right. It was stated that citizens were allowed to receive details of expenditure on work done over the last five years. Also, official documents could be photocopied as evidence.

Implementation delays in the above decision once again lead to impatience among the citizens. On July 13, 1997, the social work research centre (SWRC) Tilonia known as the *Barefoot College* in Ajmer district of Rajasthan organized a public hearing near Tilonia. It was the first time in the history of volunteerism in India where a voluntary agency made its accounts, books, vouchers, bills, muster rolls and other details of expenditure available to the rural community for public scrutiny. Many prominent people attended the hearing, and accounts of the last 10 years were displayed. This silenced critics in political and administrative circles, stopped from making false accusations, put accountability in the hands of government officials and prompted officials to maintain proper records. By this time, the power of the right to information to the citizens was more than evident. All these attempts and the persistence of the volunteer groups can be regarded as the earlier stage of the present Right to Information Act.

2.2.3 Freedom of Information Act

The Freedom of Information Act was passed by the Central Government of India in December 2002 [13]. It can be considered as the predecessor of the Indian Right to Information Act, 2005. Even after the presidential approval was achieved soon, the Act did not come into force after 18 months. This Act was deficient in a lot of ways. Some of them are:

1. **Title:** Information is a fundamental right, not freedom given. The title of the Act should clearly state the right.

2. **Scope:** Bill should cover all sections of society including private bodies, Non-Government organizations etc. and not just the public body.

3. **Clarity:** Disclosure should be made the norm and the officials should be trained for such an attitude.

The Freedom of Information Act eventually died out to give way to the new and improved Right to Information Act, 2005.

2.3 RTI Application Procedure

In this Section, we discuss the role of the PIO, mode of questioning, and grounds for rejection.

2.3.1 Role of the PIO

Each public institution must appoint a PIO who is responsible for matters related to the acceptance of RTI queries and replying to the applicants. The PIO accepts the queries, understand the information need and passes the individual sub-queries to department/section which holds the data. The PIO collect the replies from each department/section, complies the reply and replies to the applicant within a stipulated time. The main task of the PIO therefore is the identification of department/section within which the answer to the sub-query lies, coordinating with them to obtain the relevant information. Finally, the reply is complied for all sub-queries.

2.3.2 Mode of Questioning

Indian citizens can request information from any government institution by posting an application. A request is to be made using three distinct methods (i) By submitting a written application physically to the PIO (ii) By posting a written application to the PIO and (iii) By submitting the RTI application using the online interface. The RTI application can be written in English, Hindi or any of the official Indian languages. There is no prescribed format of application for seeking information. The application can be made on plain paper. The application should, however, have the name and complete postal address of the applicant. The information seeker is not required to give reasons for seeking the information. A fee of 10 rupees is associated with every RTI application. Given the three modes of filing an RTI application discussed above, this work relies on the second method of filing RTI requests. It is the responsibility of the institution to provide the answers to an application in the format specified by the applicant. A flow chart of the RTI application procedure is given in Figure 2.1. This flow chart is reproduced from the annual report 2007-08 of CIC [14].

2.3.3 Grounds for Rejection

PIO, after understanding the information need, assesses the sensitivities of the involved information based on certain grounds. When determined that the RTI application is seeking

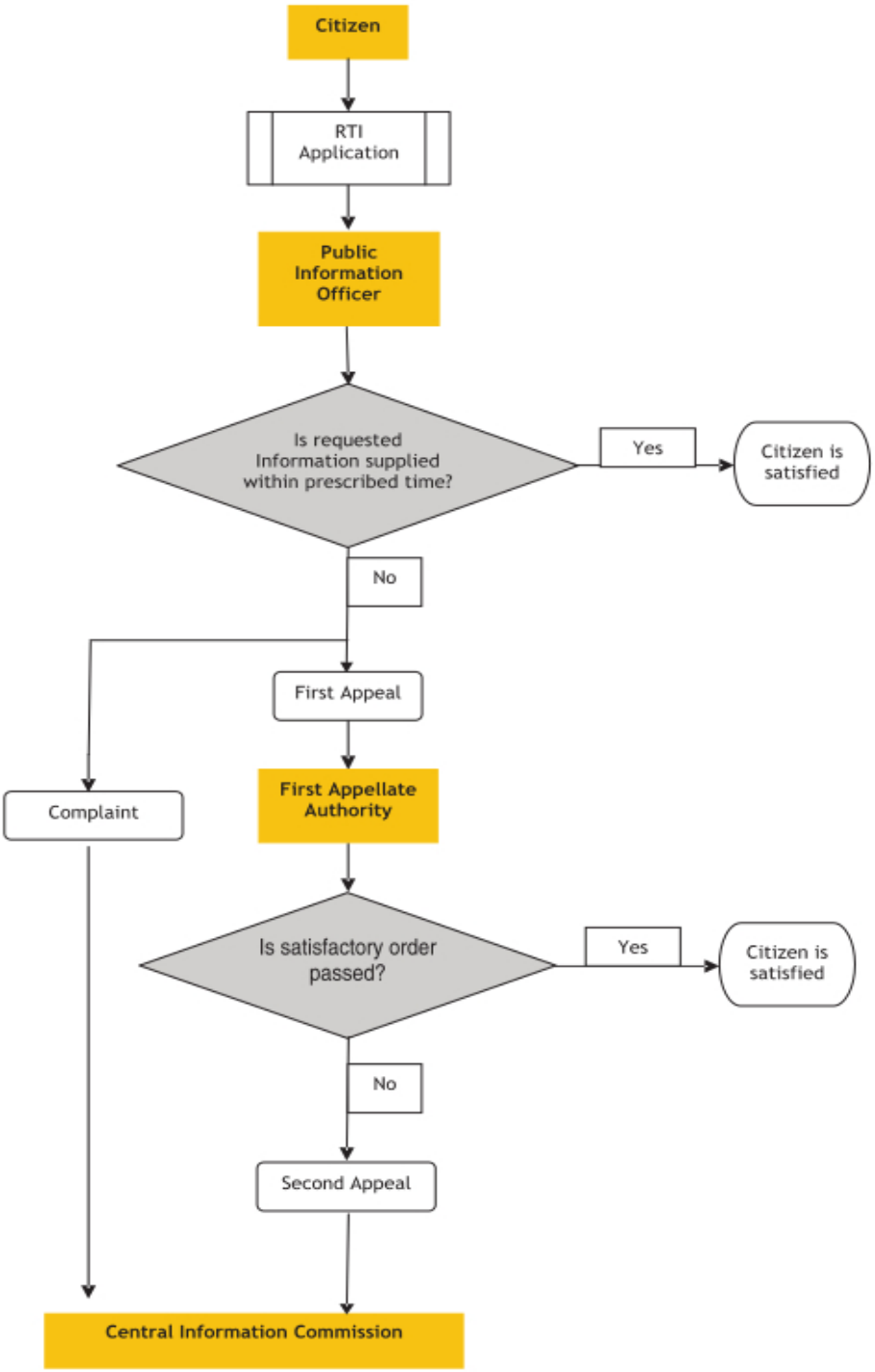


Figure 2.1: RTI application procedure.

sensitive information, the PIO sends a reply to the applicant stating that requested information is not provided by citing appropriate reasons. The reasons include: (i) seeking personal information of officials (ii) providing the information greatly diverts the resources of the institution (iii) Disclosure of information will endanger the physical safety of a person.

When the applicant is not satisfied with the reply, two appeals are at the applicant's disposal. These are approaching the first appellate authority and approaching CIC. The decision of the CIC is final and no further appeals can be made by the applicant.

2.4 Amendments to the RTI Act

The RTI Act has undergone two amendments (partly) due to the observed pattern on the grounds of rejection. When the volume of applications are rejected on identical grounds, it is subject to scrutiny by CIC and in turn by the government. The scrutiny *may* result in an amendment. We present two cases where amendments were introduced due to repeatedly rejecting to provide information.

1. Inclusion of Indian Postal Orders [15]:

   Every RTI application must associate with an application fee of rupees 10. Three modes of fee payment namely bankers cheque, demand draft or cash is listed in the RTI Act. All three modes had hidden costs associated. Demand draft and banker's cheque had their service charges attached from the bank. Payment by cash required visiting the public institution in person. Indian Postal Order (IPO) is another convenient mode of paying fees, with a nominal service charge of 10% of the total amount, which is Re.1. However, IPOs were not acceptable as a mode of RTI fee payment in the RTI Act 2005, because of which there were multiple rejections of RTI applications which were perfectly good with their content. This event was widespread enough to catch the government's eye, and there was discussion of its inclusion as a mode of payment. Ultimately a circular was passed on 26th of April, 2011 [15] citing IPOs as another mode of fee payment.

2. Exemption of political parties from being a public authority [16]:

   Seeking information on the source of funding for political parties is not uncommon. To understand the inner workings of the organization, political parties are often asked to cite their source of funding. With the advent of the RTI, multiple applications were filed requesting for their financial details. The political parties argued that they are not under the direct funding of the central or state government and hence are not liable to divulge such information. Such queries were repeatedly rejected, and a notice issued stating political parties as not being PAs. It was finally included in the RTI Amendment Bill 2013 [16].

In both of these cases, RTI applications were rejected by government officials stating a common reason. Such rejections caught the government's attention and amendments were made to address the issues. This is an indication that RTI query-log contains *latent patterns* that assist in potential amendments, thus motivating the need for us to understand RTI properties and analyze them.

2.5 Challenges

The challenges in data collection, identifying latent patterns, data representation and validation of the obtained results are briefly presented.

Data collection RTI data is not found in a centralized place but are distributed in individual government institutions across the country. The mode of acquisition of RTI data is offline, and the data have to be manually collected from each institution via post and/or personal visit. In addition, raw RTI data needs to be translated (India has 22 official languages) and digitized to be able to be used for further analysis.

Latent patterns Identifying latent patterns beyond the statistical summaries is one of the key challenges. The latent patterns are to be general and well accepted and well defined (when possible) in the literature. Particularly, the identified latent patterns should be quantifiable which is at the core of the present work. The identified latent patterns when interpreted in the RTI Act context should provide directions for proposing potential amendments. We identify latent patterns with the literature as the base.

Representation The RTI data consists of queries put forth by citizens, and other associated statistics like reply time and rejection grounds. The RTI data is predominantly text data. Traditional text modeling techniques to represent the RTI data are not amicable in achieving computable latent patterns discussed above. In particular, text models such as TFIDF [17] and its variants are most suitable for *effective storage* of the collection of documents and *efficient retrieval*. In the present work, neither storage nor retrieval is of concern. Therefore tailored data models are needed to achieve the computable latent patterns in addition to their interpretation.

Learning models The identified patterns are to be learned by employing suitable models. Traditional machine learning techniques such as supervised learning are of limited applicability in this context. We look for models that are widely used in the query-log analysis of web search engines, models that are well known in analyzing the queries, namely, psychometric modeling techniques.

Interpretable models Identifying *interpretable* learning models are at the heart of this work. When the estimated latent patterns are interpreted in the context of the RTI Act we hope to identify pointers that lead to potential amendments.

Validation Validating the experiments is one of the requirements for the learning models. The collected data include RTI applications and reply statistics. It does not have the true values of the latent patterns nor the data contains any interpreted future amendment. This pose challenges in validating the obtained latent pattern values and associated potential amendments. We make an effort to validate the presented learning model and obtained potential amendments.

2.6 Summary

This Chapter discussed the brief history of the RTI Act and the RTI process for filing an application. One pattern of *repeated rejection* and its role in paving way for amendment in the RTI Act is discussed. Additionally, challenges in the RTI data analysis is also presented. In the next Chapter, the process of RTI data collection is detailed.

3

Dataset

3.1 Introduction

In Chapter 2, a brief history of the RTI Act and its process is presented. In this Chapter, we present the efforts towards the collection of RTI applications. Collection of RTI applications which were posed to government institutions is at the heart of this thesis work. The objective is to create a repository of RTI applications and associated reply statistics. This includes collecting all the RTI applications filed to *each* government institution across the country from the date of inception of the RTI Act, that is, 12-Oct-2005 till 01-Jan-2015. Collection of this particular data from government institutions is fairly manual due to the following reasons:

1. Identifying the list of government institutions is the first step in filing the RTI application. An exhaustive list of government institutions is not readily available. Even if we restrict to government educational institutions, an exhaustive list of governmental educational institutions need to be gathered from different sources such as all central government universities, all the state government universities etc. manually.

2. To file an RTI application, the application must be addressed to the PIO of a government institute. This information about the PIO includes the name of the PIO, address and Pincode has to be collected manually.

3. Some details in government websites are provided in regional languages. Translation expert's help is needed in such instances.

4. To file identical RTI applications to two or more government institutions, one RTI application to each government institution needs to be filed separately. This is because government institutions are independent in their functioning and no hierarchy that exists across government institutions.

5. An online RTI interface at https://www.rtionline.gov.in/ exists for filing RTI applications to the central government institutions. This system was introduced in 2013 to have a single-window mechanism for filing RTI applications to any of the central

government institutions through this system. However, this system is only for the *central government institutions*. In addition, only one application at a time can be filed using this interface. No automated method exists for filing identical RTI application to multiple central government institutions using this interface.

6. State government institutions have their own online RTI systems and the application filing and reply mechanisms differ from that of http://www.rtionline.gov.in/request/request.php. These systems inherit similar limitation as the RTI interfaces of the central government.

Given a large number of government institutions across India and expected time to obtain the requested data, we have focused on collecting RTI query logs (RTI applications) from government *educational* institutes only. The reasons for restricting the domain of the RTI query log are the following:

- Certain government organizations have sensitive material. Collecting the data stated above has a chance of getting rejected in the light of the sensitivity of the information.
- The information requested above from educational institutes is relatively less sensitive.
- The education domain leads to less number of government institutes from where the data is to be collected.
- The collected data allows for comparison among similar peers in a meaningful way.

3.2 Educational Institutions

The first task in data collection is the identification of institutions to which the RTI application is to be filed. As the data to be collected is limited to educational institutions, we have broadly categorized the educational institutions into the following two categories of institutions from where the data was collected:

Class 10 and 10+2 educational boards: in this category, we have identified 64 boards, of which 22 are central educational boards.

Universities: in this category, we have identified 31 central universities, 197 state universities and 60 deemed universities.

For each of the above categories, a complete list of government institutions are identified. Addresses of the PIOs for all the institutes in each category are collected by visiting websites of the respective institutes and by collecting from known sources. Every state has a class 10 educational board and 10+2 educational board. In addition, the central school educational boards also are considered. For the second category, the list of central universities and state universities are obtained from the University Grants Commission (UGC) website.

The complete list of institutions from where data was collected is given in Table A.1 and Table A.2. The URL of the institutions are provided in Table A.1 and A.2. The PIO details

are obtained by visiting each of the websites of the institutions. Some institutions do not have the postal address of the institution in their official website. In that case, the address obtained from a popular search engine was used (for example, Nilamber-Pitamber University, Jharkhand).

3.3 Data Collection Process

3.3.1 Method Adopted for Data Collection

The government has directed all its ministries to display the RTI applications and their respective replies on the respective institute websites. On 21-Oct-2014, the Times of India published an article regarding this directive[1]. However, our data collection through online web crawling method has not been possible as the majority of the government educational institutions have not published the complete list of received RTI applications and associated replies on their websites. The RTI application system of the central government does adhere to this directive[2]. However, this system publishes the list of received RTI applications and the summary of *a few* RTI applications, but the actual queries are not available, in some cases, even the summary of the queries are not available.

A bootstrap approach is adopted to obtain the RTI query-reply statistics data by filing an RTI application seeking the desired data. An RTI application addressing to every PIO of the government institution given in Table A.1 and Table A.2 is posted. We prepared an RTI application consisting of three queries:

- All the RTI applications received by an institution during 12-Oct-2005 to 01-Jan-2015[3].
- Date, month and year of reply for each query.
- The number of queries that have been rejected, and their grounds of rejection.

An example RTI application filed to Tezpur Central University is presented below. Identical information requests were made to the educational institutes given in Table A.1 and Table A.2.

05-Oct-2015

To
The Public Information Officer,
Tezpur University,
Napaam, Tezpur-784028.
Subject: An application under the Right to Information Act, 2005.

[1] https://timesofindia.indiatimes.com/india/RTI-DoPTCPIO/articleshow/44900437.cms

[2] https://dsscic.nic.in/rti-request/

[3] The end date is different for different institutions depending on when the request for data was sent. For example, we sent the RTI application to Tezpur University on 05-10-2015, and the data was requested for the time 12-Oct-2005 to *31-July-2015*

Sir/Madam,
I would like to seek the following information under the Right to Information Act, 2005 from its inception in your institution.

1. The photocopies of the RTI applications that have been posted to your institution from 12-Oct-2005 to 31-July-2015.

2. Date, month & year of reply of each query.

3. Number of RTI queries that have been rejected, and the grounds of rejection.

A fee of Rs. 10 is being submitted by Indian Postal Order along with this application. A photocopy of my identity card is also attached as proof of identity. The above information is sought purely for research and academic purpose.

Thanking you,
Nayantara Kotoky,
Department of Computer Science and Engineering,
Indian Institute of Technology, Guwahati,
Amingaon, Assam, India-781039.
Email: nayantara.kotoky@gmail.com

It is to be noted that RTI replies are not a part of the collected data. The reasons are as follows:

1. RTI replies are huge in number and difficult to obtain.

2. RTI replies consist of a bunch of official documents instead of specific RTI reply, which often leads to a huge number of documents[4].

3. RTI replies have redundancy; for many RTI queries the same official documents are served as replies. With every document costing money, the collected data would be huge, consuming both time and money.

4. Commonality in expressed concerns is found in the question rather than the reply.

The above RTI request when processed successfully (that is when the state *Citizen is satisfied* is reached in the RTI flow chart given in Figure 2.1) resulted in a favorable reply from the PIO to share the data on payment of an additional fee for photocopying. We obtain 625 documents from Tezpur university. An example RTI application received by this university is presented in Figure 3.1. Time taken to receive the first reply is 38 days and the time taken to obtain the photocopies of all the RTI applications is 95 days. A similar effort is made to all the listed 352 educational institutions.

[4] https://timesofindia.indiatimes.com/india/After-2-year-delay-man-receives-RTI-reply-40000-pages-long/articleshow/46422555.cms

Mode of sharing	Cost
Photocopy	2.00 rupees per page
CD/DVD	50.00 rupees per CD/DVD
Online inspection	5.00 rupees per hour

Table 3.1: Mode of data sharing and cost associated with the respective mode

It is to be noted that obtaining data from government institutions requires persistent efforts. Of the 352 filed RTI applications, 341 RTI applications required follow-up. Follow-ups were required for multiple reasons. These are (i) when PIO agrees to share the requested data, an additional fee has to be paid. (ii) incomplete data received (iii) appeal in case of rejection (vi) there was no reply from the institute.

An instance in which our RTI application was rejected and first appeal to the appellate authority resulted in a favorable reply is presented here. Tumkur University rejected our RTI application citing that the required data does not fall under the purview of the RTI Act. In this instance, a follow-up was performed. The appeal letter is given in Figure 3.2. The appellate authority gave the decision favouring to share the requested data.

Table B.1 and Table B.2 provides a detailed list of institutes and their reply summary. It also enumerates which institutions required follow-up. The 11 institutions did not undergo follow-ups because of the following reasons: (i) Our RTI application was returned as there was *no such institute at the given address* (ii) The data was provided to us without additional payment (iii) Collected the data by visiting their office without any follow-up.

3.4 Cost Associated With Query And Reply

Each RTI application and the associated reply comes with a cost. There are two types of costs in the query reply process.

Direct cost Fee paid to the institute to file a query and to obtain documents associated with the reply.

1. At the time of filing the RTI application. Every application is associated with an application fee of Rs. 10.
2. At the time of sharing the requested information. Depending on the mode of data sharing, the associated fee has to be paid for receiving the requested documents. These are detailed in Table 3.1

Indirect cost in which the fee is not paid to the institute directly but spent in the process of obtaining the requested data.

A total expenditure of 85602.00 rupees is incurred in filing applications for 352 institutions and receiving data from 56 institutions.

3.5 Characteristics of the Collected Data

The total number of documents collected is 34,976 from 56 government educational institutes. Following are the observations made from the obtained data:

- RTI applications are found in multiple regional languages. In particular, we have noted 8 distinct official languages in the collected data.
- RTI applications are long. The maximum number of words found is 2362. The minimum number of words in an RTI application is 8. The word length distribution is given in Figure 3.3.
- RTI applications seek information from one or more departments/sections within an institution. An example RTI application given in Figure 3.4 contains three questions (subqueries) asking information regarding syllabus, affiliation, and exam. The requested information resides in different sections of the institution depending on its organisational hierarchy. The PIO is responsible for identifying the department/section in which the requested data is available within an institute, coordinate and obtain the information and finally compile the reply. Throughout this thesis, this is referred to as query-category. The query-categories and their frequency of occurrence is shown in Figure 3.5. It is to be noted that the RTI application contains only questions.
- The *reply given* by every institute for our RTI application is *incomplete*. The reply contains the answer to the first question alone. Other two questions were *unanswered*. We *derive* the answer for the second question by examining the stamp on the RTI application as when it was replied by the PIO.
- The average time required for obtaining the requested data is 124.8 days.

3.6 Data Processing

Processing the received data is an intensive task and involves manual work. Following are the main tasks involved.

Digitization: Majority of the obtained data is in the form of physical photocopies, JPEG, and PDF formats. Converting these different formats of data into ASCII text format is the first huge task.

Adopted strategy: We typed several RTI applications into the ASCII format. We employed commercial OCR software to obtain some data in the ASCII format and used PDF to text converter software to obtain data in the required format.

Translation: The received data is present in 8 official Indian languages. Translating the data in local languages into the English language is the second difficult and huge task.

Adopted strategy: We have translated part of the collected data pertaining to 5 local Indian languages namely Telugu, Assamese, Bodo, Kannada, and Hindi by hiring local translation experts.

Extracting date of reply: As noted in the previous Section, we have not received readily consumable information from the PIOs. The date of reply for each RTI application needs to be manually identified from the stamp present on the RTI application. Figure 3.6 shows that the RTI application was filed on 01-12-2014 and it is replied by the PIO on 09-01-2015 as stamped on the RTI application. The date *09-Jan-2015* was extracted manually.

Adopted strategy: We have read all the digitized RTI applications and for each of the RTI application we computed the time taken by the PIO to reply.

Query-Categorisation: Each sub-query within an RTI application needs to be categorised. This is because which section/department within an institute the query is addressing need to be identified manually.

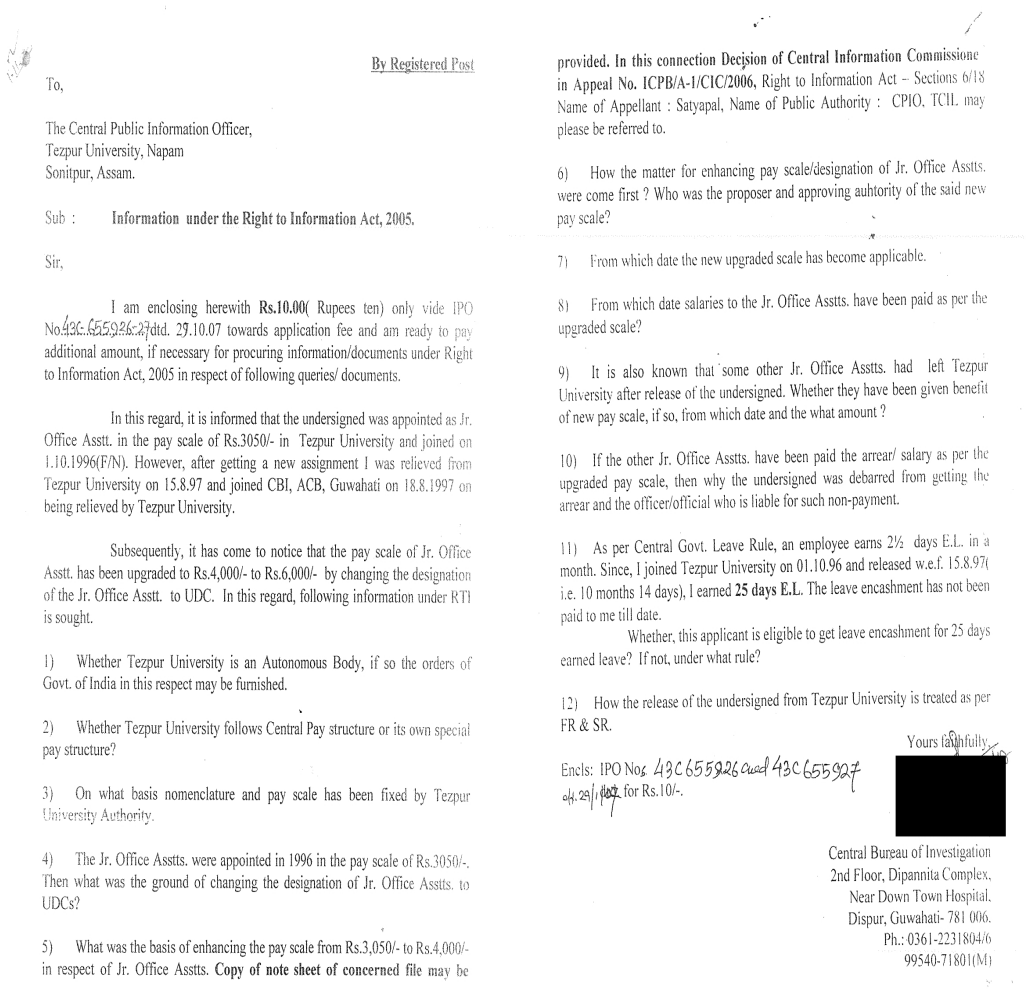
Adopted strategy: Three experts were asked to perform the categorisation of each query in every digitized RTI application. We assigned the query-category based on the majority decision of the experts.

Using the above strategies, we have digitised a total of 1614 RTI applications of the 34,976 RTI applications. Of these 1614 RTI applications, 1262 applications have reply duration information. The individual (sub)-queries in these 1262 applications have been categorised.

3.7 Summary

The data collection is a very slow process and involves manual efforts. Processing the obtained data once again draw manual efforts. We have placed *significant efforts* in obtaining data that is readily consumable by learning models.

The following Chapters present (i) literature survey (ii) RTI query log distributional analysis (iii) Computational model for obtaining transparency of individual institutions and (iv) Temporal variations of RTI query-reply data.



By Registered Post

To,

The Central Public Information Officer,
Tezpur University, Napam
Sonitpur, Assam.

Sub : **Information under the Right to Information Act, 2005.**

Sir,

I am enclosing herewith **Rs.10.00**(Rupees ten) only vide IPO No.43C.655926-27dtd. 29.10.07 towards application fee and am ready to pay additional amount, if necessary for procuring information/documents under Right to Information Act, 2005 in respect of following queries/ documents.

In this regard, it is informed that the undersigned was appointed as Jr. Office Asstt. in the pay scale of Rs.3050/- in Tezpur University and joined on 1.10.1996(F/N). However, after getting a new assignment I was relieved from Tezpur University on 15.8.97 and joined CBI, ACB, Guwahati on 18.8.1997 on being relieved by Tezpur University.

Subsequently, it has come to notice that the pay scale of Jr. Office Asstt. has been upgraded to Rs.4,000/- to Rs.6,000/- by changing the designation of the Jr. Office Asstt. to UDC. In this regard, following information under RTI is sought.

1) Whether Tezpur University is an Autonomous Body, if so the orders of Govt. of India in this respect may be furnished.

2) Whether Tezpur University follows Central Pay structure or its own special pay structure?

3) On what basis nomenclature and pay scale has been fixed by Tezpur University Authority.

4) The Jr. Office Asstts. were appointed in 1996 in the pay scale of Rs.3050/-. Then what was the ground of changing the designation of Jr. Office Asstts. to UDCs?

5) What was the basis of enhancing the pay scale from Rs.3,050/- to Rs.4,000/- in respect of Jr. Office Asstts. **Copy of note sheet of concerned file** may be **provided. In this connection Decision of Central Information Commissione in Appeal No. ICPB/A-1/CIC/2006,** Right to Information Act – Sections 6/18 Name of Appellant : Satyapal, Name of Public Authority : CPIO, TCIL may please be referred to.

6) How the matter for enhancing pay scale/designation of Jr. Office Asstts. were come first ? Who was the proposer and approving auhtority of the said new pay scale?

7) From which date the new upgraded scale has become applicable.

8) From which date salaries to the Jr. Office Asstts. have been paid as per the upgraded scale?

9) It is also known that some other Jr. Office Asstts. had left Tezpur University after release of the undersigned. Whether they have been given benefit of new pay scale, if so, from which date and the what amount ?

10) If the other Jr. Office Asstts. have been paid the arrear/ salary as per the upgraded pay scale, then why the undersigned was debarred from getting the arrear and the officer/official who is liable for such non-payment.

11) As per Central Govt. Leave Rule, an employee earns 2½ days E.L. in a month. Since, I joined Tezpur University on 01.10.96 and released w.e.f. 15.8.97(i.e. 10 months 14 days), I earned **25 days E.L.** The leave encashment has not been paid to me till date.

Whether, this applicant is eligible to get leave encashment for 25 days earned leave? If not, under what rule?

12) How the release of the undersigned from Tezpur University is treated as per FR & SR.

Yours faithfully,

Encls: IPO Nos 43C655926 and 43C655927 dt.29/10/07 for Rs.10/-.

[redacted]

Central Bureau of Investigation
2nd Floor, Dipannita Complex,
Near Down Town Hospital,
Dispur, Guwahati- 781 006.
Ph.: 0361-2231804/6
99540-71801(M)

Figure 3.1: An RTI application collected from Tezpur University

05.01.16

To
The Appellate Authority,
Tumkur University,
Vishwavidyanilaya Karyalaya, B.H Road,
Tumkur, Karnataka-572103.

Subject: Rejection of RTI application

Sir,

I had sent an RTI application dated 13-11-2015 seeking **all the RTI applications** that have been received by your institution, date of reply and the rejected applications. My application however was rejected citing applicants' information as private. In this regard, I would like to mention the following:

1. The Government of India has directed public institutions to post all RTI applications and replies on their respective websites from Oct 31, 2014. There is no such disclosure in your website.
2. You assume an applicant's information is private. This information, however, is in your possession in an official document (RTI application) and hence its privacy can be debated. Moreover, a few institutions have provided me with this information, which indicates that this interpretation of privacy is restricted to your institution.

My main aim in seeking the information is because I require **all the RTI questions** and the **location** of the questions that your institution has received from the RTI Act's inception. If you find it difficult to part with the applicants' personal information, then I request you to provide the RTI applications excluding the applicant's personal information. This is completely in keeping with the existing rules of the RTI Act.

I request you to provide me with:

1. Photocopies of the RTI applications excluding personal information of the applicants that your institution has received from 12-Oct-2005 till 31-July-2015.
2. Date, month & year of reply of each query.
3. Number of RTI queries that have been rejected, and the grounds of rejection.

I hope you consider my request this time. The above information is sought **purely for research and academic purpose**. I am ready to bear the cost of this information. Any decision regarding my request, please inform by email as well to the following email address: nayantara.kotoky@gmail.com.

Thanking you,

NayantaraKotoky,
Department of Computer Science and Engineering,
Indian Institute of Technology, Guwahati,
Amingaon, Assam, India-781039.

Figure 3.2: Appeal letter sent to Tumkur University

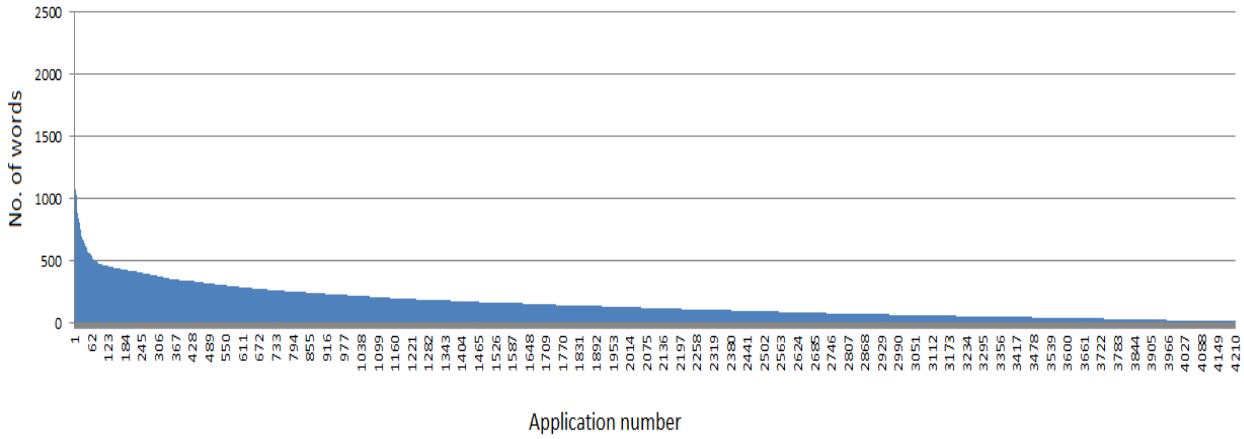
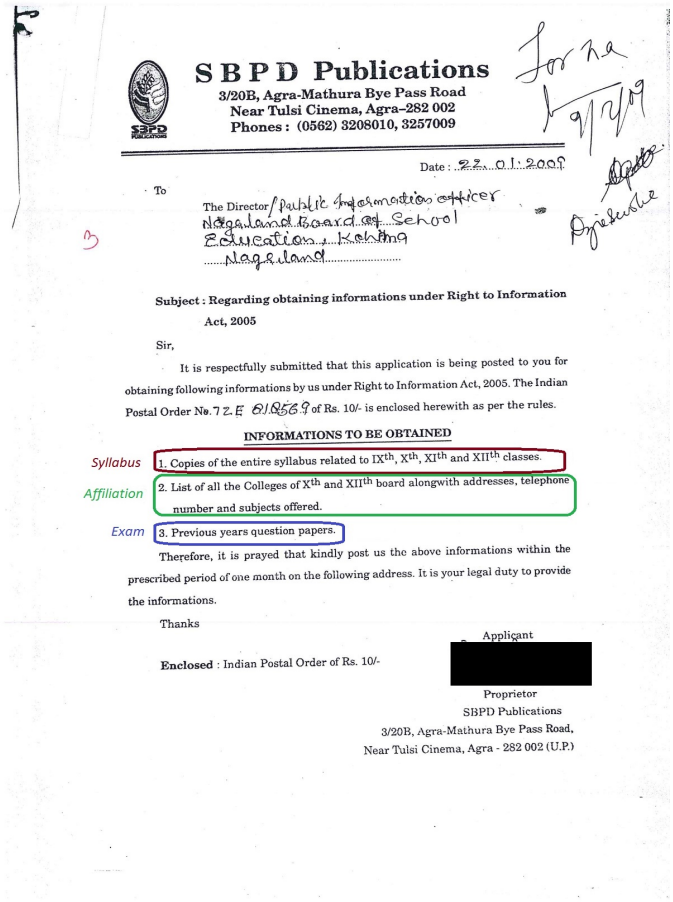


Figure 3.3: Word-length distribution for RTI applications



SBPD Publications
3/20B, Agra-Mathura Bye Pass Road
Near Tulsi Cinema, Agra–282 002
Phones : (0562) 3208010, 3257009

Date : 22.01.2009

To
The Director / Public Information officer
Nagaland Board of School Education, Kohima
Nagaland

Subject : Regarding obtaining informations under Right to Information Act, 2005

Sir,

It is respectfully submitted that this application is being posted to you for obtaining following informations by us under Right to Information Act, 2005. The Indian Postal Order No. 72E 618569 of Rs. 10/- is enclosed herewith as per the rules.

INFORMATIONS TO BE OBTAINED

Syllabus — 1. Copies of the entire syllabus related to IXth, Xth, XIth and XIIth classes.

Affiliation — 2. List of all the Colleges of Xth and XIIth board alongwith addresses, telephone number and subjects offered.

Exam — 3. Previous years question papers.

Therefore, it is prayed that kindly post us the above informations within the prescribed period of one month on the following address. It is your legal duty to provide the informations.

Thanks

Applicant

Enclosed : Indian Postal Order of Rs. 10/-

Proprietor
SBPD Publications
3/20B, Agra-Mathura Bye Pass Road,
Near Tulsi Cinema, Agra - 282 002 (U.P.)

Figure 3.4: An RTI application where each (sub-)query is of a different query-category

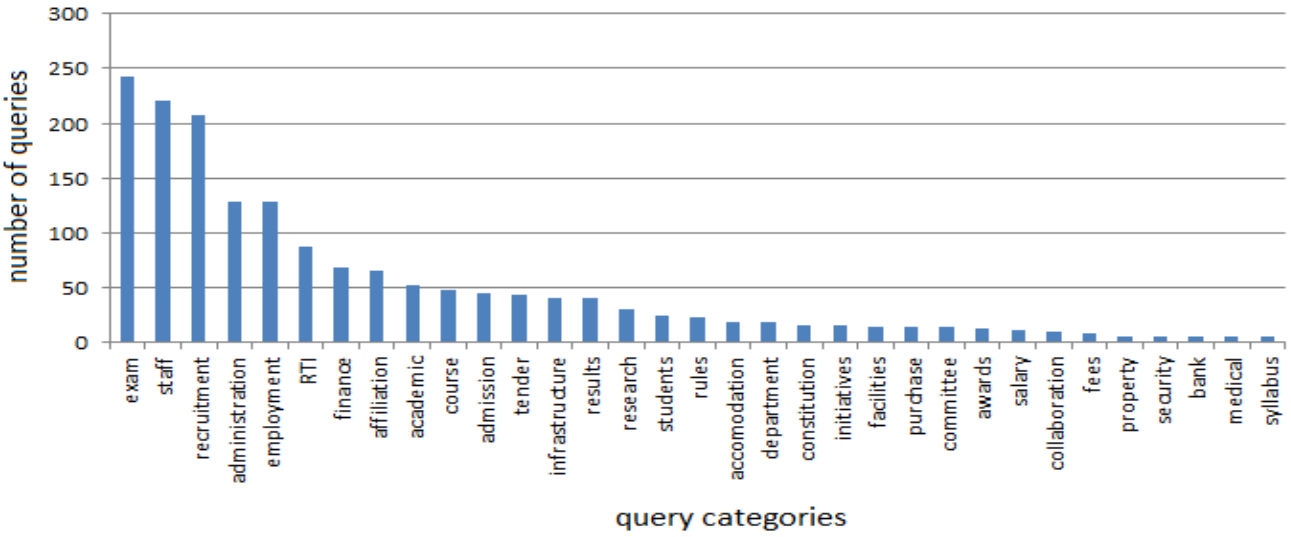
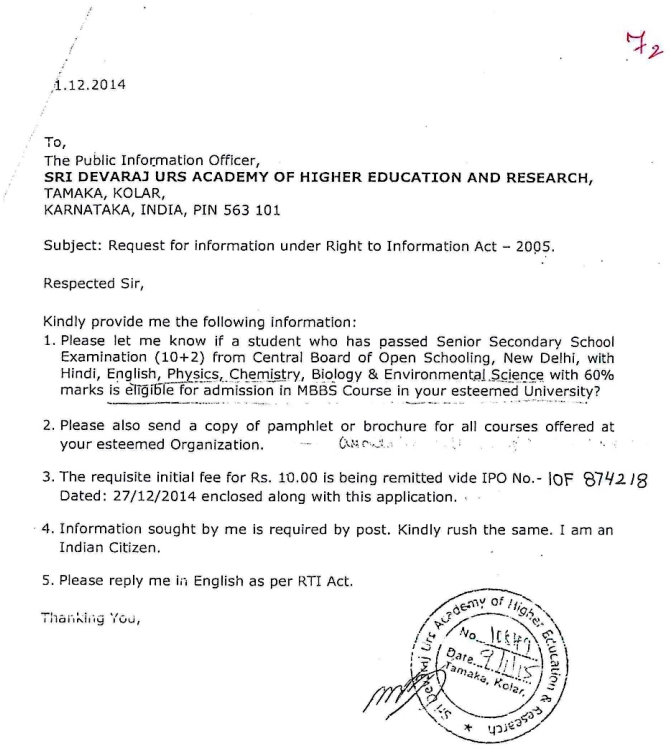


Figure 3.5: Number of queries for different categories of queries



1.12.2014

To,
The Public Information Officer,
SRI DEVARAJ URS ACADEMY OF HIGHER EDUCATION AND RESEARCH,
TAMAKA, KOLAR,
KARNATAKA, INDIA, PIN 563 101

Subject: Request for information under Right to Information Act – 2005.

Respected Sir,

Kindly provide me the following information:

1. Please let me know if a student who has passed Senior Secondary School Examination (10+2) from Central Board of Open Schooling, New Delhi, with Hindi, English, Physics, Chemistry, Biology & Environmental Science with 60% marks is eligible for admission in MBBS Course in your esteemed University?
2. Please also send a copy of pamphlet or brochure for all courses offered at your esteemed Organization.
3. The requisite initial fee for Rs. 10.00 is being remitted vide IPO No.- 10F 874218 Dated: 27/12/2014 enclosed along with this application.
4. Information sought by me is required by post. Kindly rush the same. I am an Indian Citizen.
5. Please reply me in English as per RTI Act.

Thanking You,

Sri Devaraj Urs Academy of Higher Education & Research * No. Date. Tamaka, Kolar.

Figure 3.6: An RTI application where the reply date is extracted from the stamped image on the application

4

Literature Survey

4.1 Introduction

The RTI query log data described in Chapter 3 is a new kind of data. For the first time, we attempt introducing computational methods on this special kind of data. Before discussing specific methods, literature work is presented in this Chapter. In particular, we present efforts by researchers on the RTI Act in the Indian context followed by the FOIA and its analysis at the global stage. As RTI applications predominantly comprise of queries, we examine three distinct query analysis systems namely (a) web query log (b) survey questions and (c) test questions in examination setting by students. Computational models that fit the RTI query log analysis are identified through this literature as the base. Finally, we present the use of technology in policymaking process to support the fact that computational methods making their mark in shaping policies.

4.2 Literature on the RTI Act

Initial FOI implementation The Indian Freedom of Information Act 2002 was the predecessor of the RTI Act 2005. Godbole [13] discussed the Indian FOI bill and compared it against a background of the state-level local RTI Acts of India. Several inadequacies were identified several modifications to the Bill were suggested.

The Indian RTI Act 2005 had predecessors in the state of Maharashtra (2002). The Maharashtra RTI Act has been used as a data source for collecting and analyzing RTI specific information. A survey experiment was performed by Ashwini Kulkarni [18] to assess how well this Act has worked in practice. The survey qualitatively assesses the implementation rigour and the outcome was that the Act was implemented poorly. In particular, the survey observes that the basic information about PIOs is not published. In addition, the replies do not specifically address the questions and are not in a consumable form.

In the Goa RTI Act, information requirement is not interpreted uniformly which lead

to a dialogue between citizens and the government officials [19]. This study observed that this Act has direct benefits in terms of increased awareness of citizens on social, environmental issues. It also identified the difficulty of the government officials to part with the data they own.

In its early stages of implementation, RTI applications filed to non-PAs raised questions on the very definition of what constitutes PA [20]. The bodies which are controlled by private entities with the help of substantial financial assistance by the government (central or state) are also considered public authorities. Examples are Public Sector Undertaking (PSUs), Non-Governmental Organization (NGOs) etc. However, the term "substantial funding" is not clearly defined in the Act.

Improving the Effectiveness of Implementation Niranjan [21] and Kejriwal [22] noted that the RTI Act rules were not executed to their full potential leading to not meeting the objectives of the RTI Act. To improve the effectiveness of the RTI Act, the authors suggest the following points (among others); (i) citizens should not put frivolous applications and (ii) government should use technology to ease the access to information. Gadgil [23] suggests the government documents be treated at par with the scientific documents. Results produced using the government data must be error-free and error, if any, in the official data and document be corrected on an immediate basis.

Proposed Amendments to the RTI Act [24] presents a critical analysis of the contents of the RTI Act in India and presents several pointers for amending the Act. These are: (i) Authorities should put as much information as possible on websites and notice boards which otherwise leads to frivolous applications. (ii) Rigorous monitoring to ensure proper implementation of all the provisions of the Act would be more effective. On this front, certain auditing activity was proposed to find loopholes and inconsistencies in the Act [25].

4.3 Transparency & Effectiveness of Implementation

4.3.1 Transparency

Transparency is one factor that attributes towards good governance. Michener [26] presented seven international transparency policy indexes that measure the transparencies of various policies pertaining to the government, like fund allocation, right to information by the citizens and budgetary disclosure provisions. In addition, a definition of transparency is provided for achieving *visibility* and *inferability*. Visibility refers to completeness and ease of access, and inferability refers to data leading to accurate conclusions.

Hollyer et al. [27] defines transparency as a full flow of information within a polity. This work is based on "the collection and dissemination of credible economic data by national statistical offices". This work quantifies transparency by using the information whether all the 149 countries are providing data related to 172 economic variables to the World Development Indicators and proposes a transparency index known as HRV index (HRV stand for the first name of authors that is: Hollyer, Rosendorff and Vreeland).

Heald [28] states that 'Transparency extends beyond openness to embrace simplicity and comprehensibility'. If information regarding an institution is opened to the public but the data is not comprehensible, then the institution cannot be termed as transparent. This work proposes *event transparency* and *process transparency*. Event transparency refers to *externally visible and measurable* quantities. Process transparency states how the externally visible and measurable quantities are connected and their working.

Oberoi [29] argues that information must be timely, relevant, accurate and complete for it to be used effectively. Any institution that complies with achieving transparency must adhere to these concepts. Following are the main points suggested for promoting transparency namely, maximum disclosure, obligation to publish, promotion of open government, the limited scope of exceptions and processes to facilitate access among a few others. Vishwanath and Kaufmann [30] defines transparency as the increased flow of timely and reliable economic, social and political information, which is accessible to all relevant stakeholders.

4.3.2 Effectiveness of Implementation of FOIA

Trapnell and Lemieux [31] studied the drivers for effectiveness in RTI implementation. Indicators for measuring the effectiveness of RTI implementation across countries are characterized. They, however, note that *"there is no quantifiable, reliable measurement of RTI implementation yet available to measure effectiveness"*. In particular, five key descriptors are proposed namely (a) enabling conditions, (b) demand for information, (c) institutional capacity, (d) oversight and (e) transformative factors. Though these indicators cover a broad spectrum and study the effectiveness at a greater length they still are qualitative. Neuman [32] observes that *several factors influence the enforcement model of access to information (ATI) laws*. The author notes that enforcement model-specific indicators need to be identified for evaluation of the effectiveness of the ATI laws.

Hazell and Worthy [33] studied the FOI systems in UK, Australia, New Zealand, Canada, and Ireland and addressed the question of how to measure the performance of FOI regimes. The authors enumerate the quantifiable measures across FOI regimes. The main data collected by the government are:

1. How much is the Act used, that is, how many FOI requests are there?
2. How many FOI requests are granted?
3. How many FOI requests are refused, and for what reasons?
4. How many refusals are taken to appeal?
5. How many appeals are successful?
6. How many FOI requests are subject to delay beyond the stipulated deadline?
7. How many appeals are delayed?

Additional measures of importance are the time taken to process requests, number of appeals undertaken, reduction of backlog and the attitude of the government towards transparency and FOI. The number of requests is considered as a performance measure. The

count of how many FOI requests are granted is another measure of performance. Better this proportion, more transparent is the institution in providing information to the citizens.

4.3.3 Performance Assessments

Performance assessment of policies such as FOIA is attempted in the literature. Understanding the performance of the FOIA is of importance as it influences the perceptions and may potentially impact the decisions. Examples include decisions on investment [34, 35, 36, 37]. Measuring the performance of FOIA is broadly classified into:

1. Index-based assessment: The index-based methods focus on obtaining aggregated quantity based on a set of parameters (indicators). Examples include the *global right to information rating* https://www.rti-rating.org. This index-based system uses 61 indicators[1], each of which corresponds to a *good* RTI regime. Gregory Michener [26] studied six such international policy indexes. He has identified pitfalls in computing the respective policy indexes and offered arguments on where to exercise caution in using these policy indexes.

2. Aggregates: These focus on the parameters measured in each of the government institutes. The aggregates include the number of requests received, number of requests replied etc. However, these aggregates are not aggregated further. That is, they do not summarize into a single number for each of the institute. Therefore analysing FOIA performance across institutes becomes difficult. Hazell and Worthy [33] studied five metrics across five countries and presented the FOIA performance results.

3. Quantitative efforts: These efforts quantitatively compare FOIA performances. These works rely on specific research task such as comparing the FOIA performance during the Bush administration and Obama administration [38]. As they compare several government institutions within two distinct administrations, they rely once again on aggregates discussed in the above point to make the comparison.

4. Performance measures by journalists: The fourth approach examines FOIA performance by individuals such as journalists.

Transparency and effectiveness of implementing the Act are the central qualities and there have been several qualitative studies on assessing what are the appropriate measures to understand these properties. [27] quantify transparency of countries using one definition of transparency, that is, economic data information. There are several definitions of transparency, providing various pointers of measurements. This thesis explores at least two of the definitions given above, represents them using RTI data and model the data to quantify transparency. Effectiveness has not been quantified so far in the literature.

[1] https://www.rti-rating.org/country-data/by-indicator/

4.4 Query Analysis

Three forms of questions are closely examined in the following subsections and computational techniques associated with each form of a question are presented. These are (a) web query logs (b) survey questions and (c) test questions in any examination.

4.4.1 Web Query Log Analysis

Web query logs are subject to diverse analysis to improve results obtained by search engines. A typical analysis of interest for web query logs include understanding *query length* and its complexity, how often queries are repeated, the average number of queries posed per second [39]. In this work, the authors performed an analysis of the AltaVista search requests over several weeks. They collected a dataset as a text file of search requests containing attributes associated with a search query. On the obtained data, first-order analysis and second-order analysis were performed. First-order analysis included counting of single items like the frequency of query terms, number of words per query etc. The second-order analysis requires joint counts to find a correlation between items.

Phan et al. [40] studied the 'narrowness' and 'broadness' of a search query and its correlation to query-length of users. 'Narrowness' and 'broadness' of a query refers to how much specific the query is regarding the information required by the user. The outcome of this study is (i) the intersection between broadness and narrowness is at 3 words (ii) as query-length increases, the information sought in the query is more likely to be perceived as narrow.

Weber et al. [41] analysed web search query log of Yahoo! search engine from July 4, 2011, to January 8, 2012, to understand the political leanings of the citizens of the USA. They analysed the queries to understand and isolate several important patterns: (i) assigned a political leaning to the searched queries and corresponding Wikipedia articles (ii) found trending issues in the political domain, and created a timeline of the issues by showing popular queries in multiple time-intervals arranged chronologically and (iii) false political statements made have more corresponding search queries than truer statements.

A study to identify user goals were performed by Strohmaier et al. [42]. The task was to identify whether the queries did indeed have a specific goal associated with it, or were they random queries whose goal could not be determined.

ATI query log analysis In addition to the web query log analysis, ATI query logs too were subject to analysis. Berliner et al. [43] investigate what kind of information citizens demand from government establishments. In particular, 1 million ATI requests made to the federal Mexican government by the citizens during the 2003-2015 period was analyzed. Unsupervised topic modeling using Latent Dirichlet Allocation (LDA) method [44] was employed to identify important topics which form the citizens' interest. Top 20 topics were analyzed to understand if the topics have potential for accountability or not. It is concluded that approximately 30% requests are of the nature of potential accountability.

4.4.2 Analysis of Survey Questions

Survey data comprises of questions and associated response choices. These survey questions are asked with a specific objective or product in mind, like product viability in a market survey analysis. More often than not, the answers are usually from a predefined set. For example, the answer can be a simple 'yes or a 'no, or it can be a rating from 1 to 5. Surveys may also be questions with multiple choice answers. Such responses are modelled using statistical modeling techniques such as linear regression models, logistic regression models, generalised linear regression, Generalized Linear Models for binary, Multinomial, Ordinal, and Count Variables etc. [45].

Information gathered through survey analysis is used for analysing the impact of policies in the society. Martini and Trivellato [46] insists that microdata, collected from individuals, rather than collective statistics taken from tabulated information is a better choice for analysing the distributional impact and the effectiveness of the policies. Usage of microdata also enables analysts to use micro-simulation models (MSMs), which are analytical tools that simulate the effects of policies, both aggregate and distributional effects. Collection of microdata is performed through a survey of the population.

Kreuter et al. [47] used Latent Class Analysis (LCA) to detect bad questions in a survey. Bad questions are those that do not meet criteria such as low error variance and freedom from bias. Normally, tests for the above criteria are computed by comparing the responses to some external measure, and we can have accurate measurements of errors if the external measure is error-free. This is often not true because no gold standard may exist for a variable that the survey is attempting to measure. The authors constrained the experimental setting so that the true value is known for the question beforehand. The authors evaluate three survey items where they ask candidates to respond to their past academic difficulties. In this setting, one of the questions was designed as an erroneous question to understand the utility of the proposed models in identifying the bad question. The true values are taken from the actual academic transcripts of the candidates.

4.4.3 Analysis of Test Questions

Test question response analysis is aimed to determine the qualification of individuals or behaviour of events. Typical examples are tests for students, diagnosis of illness, measuring emotional states of people etc. Models used for analysing test questions are psychometric models, of which Item Response Theory (IRT) and its variants are popular.

Dodd et al. [48] used the Graded Response Model (GRM) to provide guidelines for setting up computerized adaptive testing (CAT) systems. These are automated test-taking systems that can assess the underlying ability of test-takers and the system can adapt newer questions depending on the test-taker's ability. The authors investigated the influence of the size of items (questions) on the internal estimation techniques like stopping rule for convergence of maximum likelihood during the estimation of the ability parameter. The results show that a minimum 30 items are required to satisfactorily estimate the ability of the test-takers.

Preston et al. [49] used the Nominal Response Model to understand the relationships among family members. The authors construct a scale to test *positive family relationships*

and use longitudinal data to assess family relationships for people across ages 1 to 29 years old. Items consisted of questions that consider support, agreement, helpfulness, unity etc. as part of a positive relationship. Responses were ratings from 1 to 6, 1 meaning 'never' and 6 meaning 'always'. Thus, higher the response values, more positive is the family relationship. Families with people of different age groups were surveyed, and item's discrimination parameter from the analysis using Nominal Response Model was used to determine which items are most appropriate to judge family relationships depending on different ages of test-takers.

Of the three distinct types of query analysis, we consider two of them namely characterizing the distribution of the queries in the RTI query log analysis and employing psychometric analysis namely GRM using the RTI query log data. These two models are chosen as they can quantify transparency and effectiveness of RTI Act implementation. In addition, they retain interpretability of the estimated parameters.

4.5 Technology in Policymaking Process

Technology in several instances has been used to influence policy outcomes. From simulating policy effects to identifying issues in government tasks policymakers and researchers have identified computational methods influencing policies. The functions of technology in effective policymaking are multiple, which are summarized in two categories namely

1. Role of Technology in Policy 2.0
2. Role of Artificial Intelligence in Assisting Policymaking Process

Efforts in each of these categories are detailed in the following subsections.

4.5.1 Role of Technology in Policy 2.0

Policy 2.0 is a new methodology where policymakers use technology in creating effective laws. Using Internet and Communication Technology (ICT) tools, lawmakers understand the impact of policies in the society as well as teach the public about existing policies. The utility of using ICT is two-fold:

1. ICT is used for modelling complex processes that are not feasible by manual observation. Societal structure is complex, with multiple parameters working together that cumulates to the well-being of each individual. ICT enables modelling the intricacies of the society, thus assist in creating better policies that suit the needs of the society at large.
2. ICT also provide visualization techniques and simulation tools that lead to a user-friendly environment to deal with the cause and effects of policies, ultimately leading to a better understanding of the societal structure and the effectiveness of policies. This simplifies the decision-making process for creating policies.

There are several policy 2.0 case studies in different parts of the world.

The 2050 Pathways Calculator [50] is a game-like environment that treats users as a minister for Energy and Climate Change. The game allows the user to make various changes to the inputs of the society like demand for electricity, home insulation, and carbon emissions from transportation systems etc. With the change in these input variables, the objective is to understand the increase or decrease of carbon emission. The end goal is to reduce 80% emission reduction targets by 2050 in the UK, and the game creates a pathway for the user to simulate how sensible usage of the inputs can lead to a reduction in carbon emission. This is a great interface for novice public to understand which policies are required for achieving the end goal. In this example, ICT is providing a method for teaching the public as well as policymakers to understand what input parameters influence carbon emissions and what policies need to be created to achieve carbon emission reduction.

ϵCONOMIA The Monetary Policy Game [51] is another game-like interface that helps the public in understanding how monetary policies work. This is an initiative of the European Central Bank. The objective is to keep a stable and low inflation rate, at below 2%. The input parameter for users to handle is 'key interest rate'.

Global Epidemic and Mobility Model (GLEaM) [52] is a modelling of human population and their mobility to identify and track diseases. The system provides a simulation platform using stochastic models and enables preventive measures in the form of intervention systems so that the influence of globally devastating disease transmissions and the corresponding social and economic damage is minimized.

Koussouris et al. [53] puts ten propositions towards policymakers to take the full benefit of Policy 2.0. Two of these relevant propositions (in the context of RTI analysis) are listed:

1. Open data is required for achieving transparency, accuracy and effectiveness of policies.

2. Usage of computational methods and their visualization makes the process of understanding policy implications easier for all parties, citizens and the policymakers.

4.5.2 Role of Artificial Intelligence in Assisting Policymaking Process

Chun [54] discusses an XML-based AI framework for e-government form processing *which adheres to all the relevant rules*. The objective involves streamlining government form processing to deliver faster and more effective results so that forms are processed accurately and fairly *through rule compliance* and a higher quality of service is delivered in the face of rapidly increasing workloads.

Nabavi-Pelesaraei et al. [55] analyses energy consumption and greenhouse gas (GHG) emission in Kiwi fruit farming. Artificial neural networks (ANNs) are used for forecasting and sensitivity analysis of energy inputs and GHG emissions. The authors computed input and output energies for kiwifruit orchards, direct, indirect, renewable and non-renewable energies were calculated and GHG emissions of each input and the total inputs' values for kiwifruit production were calculated. Several ANNs were trained to predict kiwifruit yield and GHG emissions based on orchard size and other factors like machinery, fertilizers etc. The correlation coefficients for actual and predicted values were calculated as 0.993 and 0.996 for yield and GHG emissions, respectively. Sensitivity analysis revealed that diesel fuel and nitrogen

fertilizer were the most sensitive inputs for kiwifruit yield and greenhouse gas emissions. The authors at the end of the analysis suggest that newer policies need to be made to decrease the consumption of nitrogen fertilizer.

Zhang et al. [56] create a participatory ecosystem management model for the coal-mine region in China. The authors use Fuzzy Cognitive Mapping (FCM) method to conduct simulations of how coal-mining affects the environment and the ecosystem in the area of the mining. Simulations performed using Artificial Neural Networks revealed that protecting farmland, increasing vegetation coverage, reducing solid contamination and improving energy efficiency to decrease air pollution are some of the core variables to keep in mind to reduce the environmental impacts of a coal-mining operation. The analysis ultimately suggests that national policy to maintain these variables to sustain environment-friendly coal-mining in China.

4.6 Summary

We presented a comprehensive literature base in diverse areas namely, RTI Act in the Indian context, Freedom Of Information Act (FOAI) in the global context. We presented two important qualities of RTI Act/FOAI namely transparency and effectiveness of implementation and the efforts in qualitatively characterizing them. Three distinct types of queries are identified and associated analysis efforts in the literature are presented. Recent developments in policymaking well known as Policy 2.0 using computational methods is presented.

In particular, we identify the relevant works that are adopted in this thesis work. (1) Distributional analysis of web query log and in particular, the query length data (2) The utility of psychometric models namely GRM. These two distinct analyses are adopted in the context of RTI applications analysis. In addition, we make an informed effort in quantifying (a) transparency of individual institutes unlike global transparency indices (b) effectiveness of the RTI Act implementation through the use of GRM. In addition to achieving the task of quantifying transparency & effectiveness of the implementation of RTI Act, these models are interpretable. The interpretability of these models help in identifying potential amendments.

5

Distributional Analysis of RTI Queries

5.1 Introduction

Chapter 3 detailed the data collection process and described the obtained dataset. As the dataset contains only the RTI queries it is referred to as RTI query-log data. In this Chapter, we characterize the RTI query-log data. Specifically, we are interested in two important quantities of this data namely 'query-length' and 'query-reply time'. We examine whether RTI query length follows power-law distribution or not. Query-reply time is analyzed for obtaining the *probability of getting a reply to a given query-category*. These two quantities play an important role in describing the RTI query-log data.

We aim to fit probability distributions empirically to RTI query-log data for the above two quantities. From the empirical distribution, we find a pattern using which an amendment to the RTI Act in terms of RTI query-length is proposed. The proposed amendment is then validated with the ground truth, namely, the tentative amendment put forth by the Indian government. It is to be noted that this tentative amendment was eventually not incorporated as an amendment in the RTI Act. In addition to this, the estimated distributional parameters obtained using the RTI query-length data are compared with the conventional web-query log distributional parameters and draw similarities between these two estimated parameters. This effort is made to highlight that despite the differences in both the systems, they exhibit similarity in empirical distributional fit. Following research questions are addressed in this Chapter:

RQ1 Which distribution(s) empirically fits the query-length given the RTI query-log data? In addition, is the empirical distribution comparable to that of the web query-log based empirical distribution observed in the literature [11]?

RQ2 Which distribution(s) empirically fits the query-reply-time given the RTI query-log data?

RQ3 How the distributions and associated parameters are interpreted as amendments?

Following are the main contributions of this Chapter:

1. Analysis of data distribution of Freedom of Information Act (FOIA) system across the world has not been attempted till now. For the first time we analyze query-log of an RTI system.

2. This is the first analysis of a *paid and offline* information retrieval system. We present similarity and/or dissimilarity of query-length analysis of the RTI query-log with that of the traditional web query logs [11].

3. By empirically fitting the power-law distribution to RTI query-length data, we infer that RTI queries rarely have query-length more than 500 words. Based on this observation, we propose an amendment in terms of RTI query-length. The proposed amendment is validated by citing the original proposal that involves query-length.

4. RTI studies until now have been qualitative in nature. With the distributional analysis of RTI query-reply properties, we for the first time quantify *transparency* by estimating the *probability of getting a reply to a given query-category*. We observe that query-categories such as administration, recruitment and infrastructure are some of the least transparent query-categories across India.

The rest of the Chapter is organized as follows: Data representation is presented in Section 5.2. Distribution modeling methodology is presented in Section 5.3. Section 5.4 and 5.5 present experimental results for query-length and query-reply-time data respectively. Section 5.6 presents a discussion on the obtained results and their interpretation in terms of potential amendments. Summary of this Chapter is presented in Section 5.7.

5.2 Dataset

To address the research questions RQ1 and RQ2, we describe the data that is obtained from the RTI query log. A one-dimensional vector data model is proposed. This data is given as input to the distribution model algorithm to fit the query-length data and query-reply time data. We discuss each of these below:

5.2.1 Query Length

The query-length is measured in terms of the number of words an RTI application contains. It is to be noted that an RTI application may contain multiple sub-queries. For example, the RTI application in Figure 3.4 consists of three sub-queries. In the present analysis, the query-length is measured as the sum of the words of *all the sub-queries* in an RTI application. This choice is because the PIO processes an RTI application as a whole when replying.

From the collected RTI applications, we construct a dataset that is of the form: $\{(nw, f_{nw})\}_{nw=1}^{2362}$ where nw denote the *number of words* and f_{nw} denote the number of RTI applications which contains exactly nw words. The maximum number of words found in the RTI application dataset is 2362. We fit the distributions empirically using this dataset. In Section 3.6, we have digitized a total of 1262 RTI applications. In addition, we obtain 2948 RTI applications from the Ministry of HRD, India. The data is available at the URL https://www.mhrd.gov.in/

rti-higher-education?field_requestreg_no_value=&field_request_recvd_date_value=&amp%3Bpage=21. A total of 4210 RTI applications belonging to 23 different institutions is used for the *query-length distributional analysis*. The institutions from where the query-length data is used are enumerated in the Table 5.1.

Sl. No.	Institution Name	# Applications	Details
23	Council of Higher Secondary Education Manipur	28	English, Printed
24	Mizoram Board of School Education	29	English
25	Nagaland Board of School Education	33	English
34	Telangana Open School Society	8	English
55	CBSE, Panjabari, Guwahati	100	English, Assamese, Hindi
65	University of Hyderabad	216	English, Printed
67	The English and Foreign Language University Shillong	1	English
78	Krishna University	23	English, Printed
95	Jawaharlal Nehru Architecture & Fine Arts University	77	English, Printed
96	A.P.Horticultural University	33	English, Printed
102	Vignan University	7	English
103	Koneru Lakshmaiah Education Foundation (K L University)	40	English
112	Tezpur University	262	English, Printed
119	Krishna Kanta Handique State Open University	59	English, Printed
125	Chanakya National Law University	47	English, Printed
149	Pt. Sundarlal Sharma (Open) University	92	Hindi
224	Rani Channamma University, Belgaum	82	English, Printed
226	Karnataka Janapada Vishwavidyalaya	19	English, Printed
232	Jawaharlal Nehru Centre for Advanced Scientific Research	56	English, Printed
237	Sri Devaraj Urs Academy of Higher Education and Research	7	English, Printed
252	Maharshi Panini Sanskrit & Vedic Vishwavidyalaya	8	English, Printed
286	Central Institute of Fisheries Education	35	English
Online	Ministry of Human Resource Development	2948	English

Table 5.1: Institutions from where data is used for query-length analysis.

5.2.2 Category-wise Query Reply Time

We have identified a total of 26 sections/departments within the educational institutions[1]. Every institute contains a subset of these departments. We assign each sub-query to a category which is termed as *Query-Category* denoting which section/department within an institute is responsible for answering the sub-query.

For the query-reply-time analysis, each sub-query in an RTI application is considered separately and reply time is associated with the sub-queries. The following information is extracted from each of the 26 query-category reply data: {Number of queries replied in X days | X $\geq$ 1}. The dataset is of the form: $\{(nd, nq_{nd})_{nd=1}^{30}\}$ where nq_{nd} denote the number of sub-queries replied and nd denotes the number of days taken by PIO to reply to those sub-queries. In other words, we count: how many RTI sub-queries were replied in one day? How many RTI sub-queries were replied in two days? and so on. A separate dataset is obtained for every query-category. A probability distribution is fitted to every query-category dataset. The dataset for every query-category is presented as follows:

[1] academics, accommodation, administration, affiliated, etc

Query-Category	Dataset
Academic	$\{(1, nq_1), (2, nq_2), (3, nq_3) \cdots, (30, nq_{30})\}$
Accommodation	$\{(1, nq_1), (2, nq_2), (3, nq_3) \cdots, (30, nq_{30})\}$
Administration	$\{(1, nq_1), (2, nq_2), (3, nq_3) \cdots, (30, nq_{30})\}$
$\vdots$	$\vdots$
Tender	$\{(1, nq_1), (2, nq_2), (3, nq_3) \cdots, (30, nq_{30})\}$

Table 5.2: RTI query-reply time data representation for every query-category

Sl. No.	Name	Type
23	Central Institute of Fisheries Education	University
24	Mizoram Board of School Education	Board
34	Telangana Open School Society	Board
95	Jawaharlal Nehru Architecture & Fine Arts University	University
103	Koneru Lakshmaiah Education Foundation	University
115	Assam Agricultural University	University
118	Bodoland University	University
231	Jagadguru Sri Shivarathreeswara University	University
301	Tata Institute of Fundamental Research	University
10	Board of Secondary Education Assam	Board

Table 5.3: List of institutions for query-reply-time data.

5.3 Modeling Methodology

Given the query-length and query-category reply time data, the task is to

- Find a probability distribution that best describes the given data.
- Interpret the obtained distribution, its parameters and auxiliary information that leads to identifying potential amendment.

To achieve the first task, we start with power-law distribution and estimate the parameters of this distribution given the RTI data. Alternate distributions are explored to test if power-law is the best fit for the RTI data. RTI query-log analysis is performed along the similar lines of web query log analysis [11]. A methodological approach to fit alternate probability distributions is adopted along similar lines presented in Clauset et al. [57].

5.3.1 Power-Law Distribution

A power-law is a relationship between two distinct variables such that change in one variable causes a relative change in the other variable. In other words, one variable varies as a power of the other variable. Let x be a variable said to follow power-law if it is drawn from the

probability distribution given as:

$$p(x) \propto x^{-\phi} \tag{5.1}$$

ϕ is a positive parameter (> 0) that is used to denote the power of the variable x by which the probability changes. Estimating the parameter ϕ that explains the given data is about empirically fitting the distribution to the given data. However in practice the estimated parameter ϕ do not explain the given data *entirely*. Only a portion of the data may follow power-law. Hence a lower threshold x_{min} that denotes the minimum value of the variable x beyond which power-law is applicable for the given data is of interest in the estimation process. Accordingly equation 5.1 is modified to reflect this notion and is given as

$$p(x) = \frac{\phi - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\phi} \tag{5.2}$$

Fitting power-law distribution to the given RTI data requires estimating the values of ϕ and x_{min}. There are four prominent methods for power-law detection. These are given below:

Graphical Method: The graphical method includes detecting a power-law distribution visually. The logarithm of equation 5.1 yields a straight line. This property is used to plot a distribution of the given data in a double-logarithmic scale; if this plot resembles a straight line, then the data follows a power-law distribution.

Straight-line Approximation Method: A straight line is fitted to the log-log plot using additional methods, for example, using Linear Least Squares (LLS) [58].

Generative Models: A generative process is applied to generate data following power-law distribution when empirical data is not available. An example of such a generative process is the Pitman-Yor process [59].

Statistical Model Selection Method: A methodological process of model selection (estimating x_{min} and ϕ) to a given data and compute relative goodness-of-fit to test the quality of fit. Clauset et al. [57] executes this method by

1. Selecting the best model among a set of models by estimating parameters of each model using the Maximum Likelihood Estimation (MLE) method.
2. Use statistical significance test to identify if the model minimizes some divergent measure (for example, Kullback-Leibler (KL) divergence), to compare to the true underlying (but unknown) model. If the divergence is minimum (p-value$>$ 0.1), the model is a good fit for the data.

We adopt the statistical model selection method as it provides a *principled and rigorous approach* of fitting power-law distribution to data as compared to other existing methods. Clauset et. al. [57] suggests a series of steps to conclusively fit power-law distribution by the statistical method given a dataset. If power-law is a good fit, the data is also used to fit alternate distributions to test if power-law is *the best fit*. This method is adopted as part of the

distribution fit to estimate parameters algorithmically for query-length and query-category reply time analysis.

The algorithm followed for the distributional fit to the RTI data is given in Figure 5.1. The algorithm contains three major steps:

1. Fit power-law distribution to the data.
2. Compare power-law fit to alternate distributions to check if power-law is *the best fit*.
3. Fit other candidate distributions if power-law is *not the best fit*.

Since the experiment requires finding the *best-fit model*, the method from Clauset et al. [57] is extended to test the fit for other candidate distributions. The following subsection describes the steps towards distributional fit for RTI query-length and query-category reply time data.

5.3.2 Modeling RTI Query Length

Word-count for each of the RTI applications is computed to create the query-length data as explained in 5.2.1. Fitting power-law distribution requires the estimation of two parameters: x_{min} and ϕ. The x_{min} is estimated via Kolmogorov Smirnoff (KS) statistic [60]. Using the estimated x_{min} , the value of ϕ is computed via Maximum Likelihood Estimation (MLE) method. The goodness-of-fit (GOF) for power-law to the data is tested via the KS-test, and p-value is calculated considering the threshold as 0.1 [57]. If p-value > 0.1, power-law is a good fit for the query-length data, else power-law is not a good fit. The GOF test only tells if power-law is a good fit, but fails to tell if there are other related distributions which can better approximate the given data. The power-law fit is further compared with two other *closely related distributions* [61], namely, log-normal and exponential distributions, using Vuong's test [57]. After comparison with these two alternate distributions, one can be fairly confident whether power-law is the best fit among the closely-related distributions. Figure 5.1(a) depicts the flowchart for the steps in the distributional fit for RTI query-length.

There are two instances where power-law does not turn out to be a good fit:

1. During GOF using KS-test when $p < 0.1$
2. During comparison with alternate distributions using Vuongs test when one or both of the alternate distributions are indicated to be a better fit than power-law.

In these two instances, we proceed to a pool of other candidate distributions given in Table 5.4. We select seven common continuous probability distribution functions and fit each distribution to the given query-length data and test the GOF using three statistical tests, namely, Kolmogorov-Smirnov (KS) statistic, Anderson-Darling (AD) Test and Cramer-von Mises (CVM) Test.

From Table 5.4 the distributions that are 'not rejected' by at least one of the GOF tests is passed on to the next step. After this step, we have a collection of models that are a good

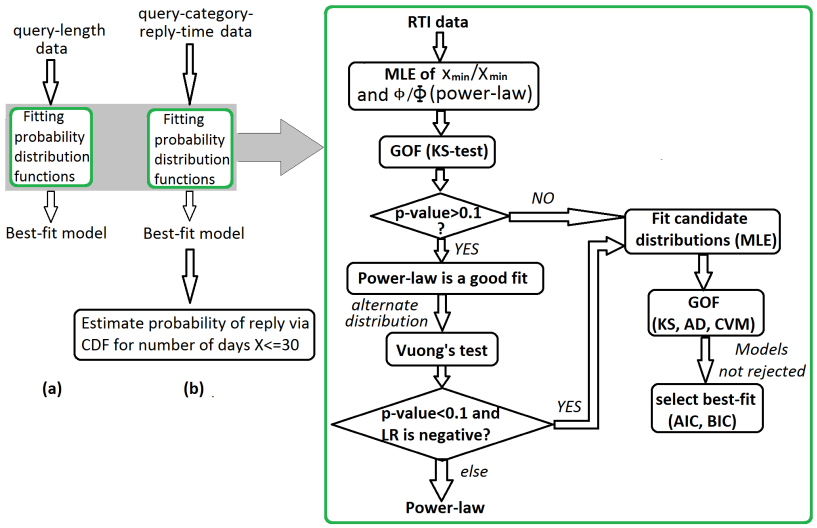


Figure 5.1: Flow-chart representing the steps to model (a) query-length data (b) query-category reply time data. MLE=Maximum Likelihood Estimation, CDF=Cumulative Distribution Function, LR=Likelihood Ratio, KS=Kolmogorov Smirnoff test, AD=Anderson-Darling Test, CVM=Cramer-Von-Mises test, AIC=Akaike's Information Criteria, BIC=Bayesian Information Criteria.

Model	Probability Distribution function	Conditions
Exponential	$f(x\|\mu) = \frac{1}{\mu}exp(\frac{-x}{\mu})$	$x\epsilon\mathbb{R}_0^+, \mu\epsilon\mathbb{R}^+$
Gamma	$f(x\|a,b) = \frac{x^{a-1}exp(\frac{-x}{b})}{b^a\Gamma(a)}$	$x\epsilon\mathbb{R}_0^+, a, b\epsilon\mathbb{R}^+$
Log-logistic	$f(x\|\mu,\sigma) = \frac{exp(\frac{x-\mu}{\sigma})}{\sigma(1+exp(\frac{x-\mu}{\sigma}))^2}$	$x\epsilon\mathbb{R}, \mu\epsilon\mathbb{R}, \sigma\epsilon\mathbb{R}^+$
Log-normal	$f(x\|\mu,\sigma^2) = \frac{exp(\frac{-(\ln x-\mu)^2}{2\sigma^2})}{x\sigma\sqrt{2\pi}}$	$x, \sigma^2\epsilon\mathbb{R}^+, \mu\epsilon\mathbb{R}$
Pareto	$f(x\|s,\sigma,\Omega) = (\frac{1}{\sigma})(1+s\frac{(x-\Omega)}{\sigma})^{-1-\frac{1}{s}}$	$x \geq \Omega : s > 0$
		$\Omega \leq x \leq \Omega - \frac{\sigma}{s} : s < 0$
Weibull	$f(x\|a,b) = ba^{-b}x^{b-1}exp(-(\frac{x}{a})^b)$	$x\epsilon\mathbb{R}_0^+, a, b\epsilon\mathbb{R}^+$
Burr	$f(x\|s_1,s_2,\sigma) = \sigma s_1 s_2 \frac{\sigma x^{s_1-1}}{1+(\sigma x)^{s_2+1}}$	$x, s_1, s_2 > 0$

Table 5.4: Candidate Probability Distribution Functions used to fit to RTI query-length and query-category-reply data.

fit for the data. To select the best-fit model we use Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). The model with the lowest value of AIC and BIC is selected as the best-fit model. We finally have either power-law or one of the seven-candidate distributions as the best fit model for RTI query-length data.

In the steps followed for fitting the candidate distributions to the data (when power-law is not a good fit), three statistical tests for testing the GOF are used. Studies have been made to understand the power of these tests [62, 63], and the conclusion is that all of them

are powerful under different characteristics of the data. For example, the AD test exhibits better results in data having fatter tails, and the KS test shows less bias against deviations in the centre of the data. In previous works that analyzed online search queries [11, 64, 65], the data size is often in millions. This is not the case for RTI data since getting access to this amount of RTI data is difficult and time-consuming due to difficulty in data collection and pre-processing. Certain probability distributions are sensitive to the size of the data. Since we are fitting a variety of probability distributions to our data, we use three GOF tests to work around the 'rejection' by some distribution models due to data constraint. Any distribution model that is not rejected by any of the three GOF tests shall be considered as a potential model for the data and shall undergo the next step of the analysis.

5.3.3 Modeling Query-Reply Time

For query-reply-time analysis, the distributional analysis is undertaken for two purposes:

1. To model transparency of an institution, we take into account a simple definition of transparency, namely, *swiftness in replying to the requested information*. The 30 days time frame is considered to be the threshold for the swiftness. We compute the probability of reply of RTI queries within 30 days based on query categories. In particular, we quantify transparency of query-categories and find the most and the least transparent query-category across India.
2. To examine whether power-law is a good fit for the query-reply-time data. If power-law is not a good fit then which alternate distribution best describes the RTI query-reply-time data?

The steps followed are shown as a flow-chart in Figure 5.1(b). Similar to power-law fit to query-length data, we first estimate the power-law parameters for query-category reply time data via MLE and test the power-law fit via KS-test. This is followed by the Vuong's test [57] to measure if power-law is the best fit. In case power-law is not a good fit, we test the fits of the candidate distributions. After finding the best fit distribution (power-law or otherwise), the parameters of the fitted distribution are used to estimate the probability of reply for that query-category at the 30-day point in time via Cumulative Probability Distribution. This is used to compute the probability of reply to an RTI query in that query-category within 30 days. The entire procedure is performed 26 times, once for each query-category.

5.3.4 Symbols Used

In the experiments performed, the following symbols are used. x represents query-length. x_{min} represents the minimum value of query-length after which power-law distribution is followed. ϕ is the slope of the power-law curve fitted to RTI query-length data. For the query-category reply time experiment, reply-duration (in days) is represented by X, where $X \in \{X_1, X_2, ..., X_{26}\}$ for all 26 query-categories. The corresponding power-law parameters are represented by $X_{min} \in \{X_{min1}, X_{min2}, ..., X_{min26}\}$ as the minimum value from which power-law fits, and the slopes are $\Phi \in \{\Phi_1, \Phi_2, ..., \Phi_{26}\}$ respectively. n_{tail} is the number of data-points (query-length) that follow power-law distribution.

5.4 Experimental Results Using Query-Length

The power-law distribution parameter x_{min} is estimated via Kolmogorov Smirnoff (KS) statistic and ϕ is estimated using MLE procedure. For the goodness-of-fit test, the p-value is calculated by generating a number of datasets that have similar distributions to the given data distribution below x_{min}, but follow the fitted power-law distribution above x_{min}. The number of datasets generated is as follows: considering the required accuracy of p-value up to two decimal digits, we select $\epsilon = 0.01$; the number of synthetic datasets generated is $\frac{1}{4}\epsilon^{-2}$, that is, 2500 [57].

5.4.1 Power-law Fit on Longer Query-Length

The estimated values of the power-law distribution are presented in Table 5.5. The $p = 0.33$ is comfortably above the threshold of 0.1, indicating that power-law is not ruled out for the query-length data. It is also observed that the value of $\phi = 4.39$, which is outside the typical range of $2 < \phi < 3$ for the web-log queries [57]. This indicates that the curve is very steep, and the majority of the queries (RTI applications) belongs to a short-range of smaller query-length. In other words, applications with large word-count (that is, high query-length) are very infrequent in number. It is further observed that the value of $x_{min} = 476$, which indicates that power-law is a good fit for the longer queries and does not fit the shorter length queries. This is in agreement with the experimental findings of Arampatzis and Kamps [11] where they analyze query-length of web queries and conclude that power-law fits well for longer queries but not for the shorter ones, thus showing similarities in patterns between retrieval systems where every query has a cost associated and retrieval systems where queries have no direct cost associated.

total	**min**	**max**	x_{min}	n_{tail}	ϕ	p-value
4210	8	2362	476	88	4.39	0.33

Table 5.5: Query-length data values and estimated parameters for power-law fit and goodness-of-fit using KS-test.

5.4.2 Comparison with Alternate Distributions

To ascertain that distributions other than power-law do not explain the RTI query-length data, the power-law fit is further compared with two other closely related distributions [61], namely, log-normal and exponential distributions. Vuong's test is a method to determine the closeness of a model to the truth. It uses the Kullback Leibler Information Criterion to determine which of the two models is a closer fit to a given data distribution.

Vuong's test is performed following the work of Clauset et al. [57]. A *negative sign* in the log-likelihood ratio (LR) indicates that the alternate distribution is a closer fit to the data as compared to power-law. However, the confidence of the LR is determined by the associated p-value. If $p < 0.1$, the LR sign is a reliable indicator of the comparison test.

Power-law p-value	Log-normal		Exponential		Support
	LR	p-value	LR	p-value	
0.33	0.83	0.2	1.19	0.12	moderate

Table 5.6: Normalized log likelihood ratios (LR) and p-values obtained with log-normal and exponential distribution for query-length data.

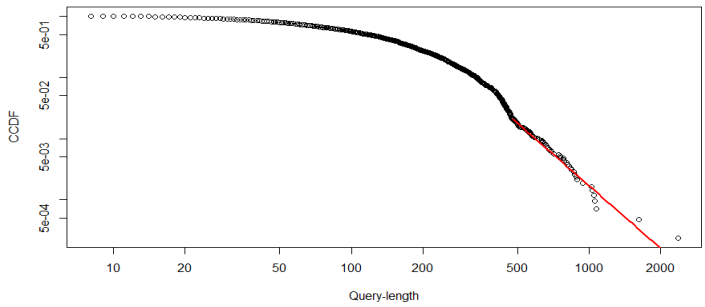


Figure 5.2: Complementary Cumulative Distribution plot (log-log plot).

The obtained result is presented in Table 5.6. From this table, for log-normal distribution, we note that the sign of LR is positive but $p > 0.1$, indicating that the LR value is not reliable. In other words, there is no evidence that log-normal distribution is a closer fit than power-law distribution to the query-length data. A similar comparison is made between an exponential distribution and power-law distribution. From Table 5.6, the sign of LR is positive and $p > 0.1$ indicating that exponential distribution is not a closer fit than power-law distribution to the query-length data. This comparison analysis concludes that power-law distribution is *the best fit* model for query-length data with moderate support.

5.4.3 Query-Length: Interpretation of Power-law Fit

The power-law fit to the RTI query-length data is presented in Figure 5.2. The red colored line depicts the fit. From this figure, we observe that the power law fit starts from word count 476 (x_{min}). This observation means that the number of documents having word count equal to $x \in \{x_{min}, ..., \max\}$; where `max` = 2362 words (refer Table 5.5) decreases as x increases. That is, *the number of documents having higher word count has less probability of occurrence*. In addition, the rate of decrease in the probability value is governed by the parameter $\phi = 4.39$ and so is the decline in its probability value. The average reply time for longer query lengths having count equal to or greater than x_{min} is 168 days whereas according to the RTI Act, every RTI application should be replied within 30 days by the PIO. Table 5.7 shows four example queries having long reply times. We observe two important points about query-lengths:

1. There are very few RTI applications posed to the Educational institutions with query length 500 or more words across India.

number of words	reply time (in days)
436	38
461	213
571	305
886	116

Table 5.7: Reply time (in days) for RTI applications whose word-count approaches 500 words.

2. The reply times for RTI application having longer lengths (around 500 or above) is unusually high (168 days on average).

Though there are *only a few occurrences* of RTI applications with large query-length, the time taken to reply to such queries is *beyond the stipulated 30-days* time limit. These observations from the obtained results provide us with a pointer to a potential amendment involving RTI query-length property: *limit the number of words in the RTI application to 500.*

We observe the following limitations with the presented experimentation: (i) The number of RTI applications collected is very small (4210) (ii) The number of institutions from which the RTI applications collected is very small (23). However, the adopted distributional analysis is a methodical one and the results obtained using three diverse GOF tests provide us with confidence on the small scale data employed.

5.4.4 Validation

We have noted that the identified pointer to the amendment based on the above experiments was indeed a candidate item for amending the RTI Act relating to the number of words in the RTI application. Many leading newspapers reported that this change in a word limit is being considered for amendment. `The Hindu` published an article on 12-Aug-2012[2] stating that *The government has put a word limit of 500 words for filing an application under the Right to Information (RTI) Act.*

The press information bureau government of India ministry of personnel, public grievances & pensions, however, has refuted the claims stating that the facts are "The existing RTI Rules 2012 notified on 31st July 2012 specifically provide in Section 3 that an application shall ordinarily not contain more than 500 words excluding annexure. It further provides that no application shall be rejected only on the grounds that it contains more than 500 words. There is no change proposed in these provisions in the new rules." However "legality of the CIC (Management) regulations of 2007 were challenged before the Delhi High Court and these were quashed. The matter has been pending before the Supreme Court". Given this context, the query-length property takes prominence.

Taking into account the feedback on introducing word length limit, in 2017, the then Union Minister has stated that *There is no change "even in a comma or a full stop" in the proposed amendment to the RTI rules relating to a word limit and fee* https://economictimes.indiatimes.com/news/politics-and-nation/no-change-in-word-limit-fee-in-proposed-rti-rules-v

[2] http://www.thehindu.com/news/national/govt-puts-word-limit-on-rti-pleas-defines-format/article3757532.ece

articleshow/58031850.cms?from=mdr. Accordingly, in the recent amendment introduced to the RTI Act 2019 the rule for limiting word count to 500 in RTI applications is **not** introduced.

We acknowledge the proposal on the word limit to the RTI applications is a very sensitive issue and has been widely discussed. This work must be viewed from the prism of scientific understanding and efficiency improvisation of the RTI Act 2005. The intention is to gain insights from the observed data and what the data offers in terms of efficiency/inefficiencies within the RTI implication systems and possible directions for improvements.

5.4.5 Mixed Model Distributional Fit for Shorter Queries

The result obtained from Section 5.4 indicates that power-law distribution fits the query-length data for queries *having a length greater than or equal to* 476 (that is $x_{min} = 476$). For queries whose length is less than x_{min}, the present power-law model (fitting query-length above x_{min}) is not a good fit. In order to understand which distributions fit for shorter-length queries, additional experimentation is performed. In this, we consider all those queries whose length is less than 476 and perform *a separate distributional fit* experiment in accordance with the steps enumerated in the Figure 5.1(a).

The experiment is done for query-length data D in the interval $[8, 475]$. We proceed according to the steps in Figure 5.1(a). In the end, we arrive at the best-fit distribution. If no distribution is a good fit among power-law or the other candidate distributions, we divide the data D into two parts D_1 and D_2 using a cut-off value x_{cut}. The value x_{cut} divides the data $D = [8, 475]$ by incrementally increasing starting from $x_{cut} = 9$. So in the first division, D_1 includes the shorter query-length data, that is, $[8, x_{cut}]$ and D_2 contains the longer query-length portion of D, that is, $[x_{cut} + 1, 475]$. If we find a best-fit model for D_1 and D_2 in accordance with the steps in flow-chart 5.1(a), then our experiment is complete. Else, the x_{cut} is incremented by one query-length, that is, $x_{cut} = 10$. So the new datasets are $D_1 = [8, 10]$ and $D_2 = [11, 475]$. In this way, we divide the data D into two parts by increasing the value of x_{cut} until a best-fit model is obtained for D_1 or D_2. At this point, if there are any data left without a best-fit model, that becomes the new D.

Three points regarding this method are:

1. The increase in x_{cut} is stopped whenever any distribution fits either D_1 or D_2. Once the parameters are obtained for the fit, the remaining data is considered as D and the process starts again, starting from the lowest possible value of x_{cut}. For example, if a best-fit model is found for the range $[8, 100]$, then in the next round we have $D = [101, 475]$ and x_{cut} starts from 102 until a portion of the data D provides a best-fit model.

2. The maximum possible value of x_{cut} is 474. If x_{cut} reaches 474 and there is no distribution that fits either D_1 or D_2, it means that there is no best-fit model for the shorter query-length data below x_{min}.

3. For any dataset D, D_1 and D_2, the distribution fit is not executed if the number of data points are below 100.

In other words, the x_{cut} value incrementally divides the dataset when the *entire* data $D = [8, 475]$ do not yield a best-fit model. Using the divisions, *a portion of* the data is

tested for power-law, and then the other candidate distributions as shown in the flow-chart. Whenever power-law is found to be not the best-fit model, we test the fit for seven candidate distributions. The algorithm that is followed is shown below for the dataset D.

Input: Query-length data D and seven distribution models.

Output: Best-fit model for query-length data D.

1. Fit all seven models to D and estimate their parameters.
2. For each of the seven model-fit, check goodness-of-fit via (i) Kolmogorov-Smirnov (KS) statistic, (ii) Anderson-Darling (AD) Test and (iii) Cramer-von Mises (CVM) Test.
3. Choose the model(s) that have at least one out of three 'not-rejected' goodness-of-fit from the above step.
4. For those models that have been 'not-rejected', calculate Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC).
5. The model with the smallest AIC and BIC value is the best-fit model for the given data D.

With the described procedure above, we find the following distribution fit, shown in Table 5.8. The obtained parameters are provided in Table 5.9. We arrive at the best fit distributions for portions of the query-length data and achieve a mixed model distributional fit; given the entire query-length data, different portions of the data (specific range of query-length) follow different underlying distributions in their probability of occurrence. From Table 5.8 we observe that there is no single model that fits the entire shorter query length range $[8, 475]$, we have identified 22 different models that fit 22 different fragments of the shorter query-length range. It is also observed that power-law is the distribution that fits the query-length data and no candidate distribution has emerged to be the best fit for any range of the query-length data.

5.4.6 RTI Query Log Vs Web Query Log

Petersen et al. [65] fits two distributions namely Negative Binomial for query length in the range $[1, 10]$ and Inverse Gaussian for query length $[11, 20]$ to web queries and demonstrate that neither model provides a good approximation to the tail with queries having length $x > 40$. A similar query-length analysis by Arampatzis and Kamps [11], fits Poisson's distribution to shorter length web queries and power-law to the longer-length web queries. Our analysis of RTI query-length data also demonstrates the necessity of a **mixed model** to describe the data distribution in the entire query-length range; no single probability distribution has a good fit for the entire data. For RTI queries, power-law provides a good approximation to the tail (x_{min}=476) of the distribution. However, shorter length queries follow a different power-law model with different parameter values for different ranges of query-length. It is to be noted that what is longer query-length for web queries is actually short length for RTI queries. The average query-length for RTI queries is 154.12 words, and for web queries it is between two and five [11]. However, the common pattern observed is that in both the web

Dataset	**Distribution**	**Range of fit**		**Size of fitted data**	**Support**
		min	max		
data from 336 to 475	power law	402	475	160	high
data from 250 to 401	power law	343	401	137	high
data from 155 to 342	power law	303	342	153	high
data from 8 to 302	power law	258	302	210	high
data from 8 to 257	power law	211	257	244	high
data from 23 to 210	power law	173	210	329	moderate
data from 8 to 172	power law	148	172	310	moderate
data from 8 to 147	power law	131	147	226	high
data from 25 to 130	power law	117	130	172	moderate
data from 8 to 116	power law	111	116	73	high
data from 26 to 110	power law	100	110	183	high
data from 32 to 99	power law	86	99	227	high
data from 19 to 85	power law	80	85	106	high
data from 8 to 79	power law	73	79	122	moderate
data from 8 to 72	power law	66	72	141	high
data from 10 to 65	power law	59	65	162	high
data from 35 to 58	power law	50	58	202	high
data from 40 to 49	power law	45	49	114	high
data from 8 to 44	power law	41	44	91	high
data from 8 to 40	power law	36	40	113	moderate
data from 9 to 35	power law	32	35	94	high
data from 26 to 31	power law	28	31	92	high

Table 5.8: Fitted distribution for shorter length queries below $x_{min} = 476$.

Dataset	x_{min}	ϕ	n_{tail}	$p-value$
data from 336 to 475	402	6.47	160	0.934
data from 250 to 401	343	6.79	137	0.736
data from 155 to 342	303	6.9	153	0.112
data from 8 to 302	258	6.85	210	0.288
data from 8 to 257	211	7.14	244	0.184
data from 23 to 210	173	7.24	329	0.182
data from 8 to 172	148	7.46	310	0.154
data from 8 to 147	131	7.89	226	0.824
data from 25 to 130	117	7.77	172	0.288
data from 8 to 116	111	8.02	73	0.55
data from 26 to 110	100	8.29	183	0.533
data from 32 to 99	86	8.43	227	0.295
data from 19 to 85	80	8.43	106	0.762
data from 8 to 79	73	8.62	122	0.966
data from 8 to 72	66	8.81	141	0.546
data from 10 to 65	59	9.06	162	0.999
data from 35 to 58	50	8.98	202	0.75
data from 40 to 49	45	9.48	114	0.717
data from 8 to 44	41	10	91	0.998
data from 8 to 40	36	9.8	113	0.5
data from 9 to 35	32	10.3	94	0.585
data from 26 to 31	28	10.5	92	0.646

Table 5.9: Parameters obtained from the fitted distributions to shorter query-length data. All the data follow power-law distribution.

and RTI queries, queries having longer length follow distributions that are different from the rest of the data.

Cost associated with RTI applications has a major influence on the length of queries in comparison to the web queries which have no direct cost associated. RTI queries are much longer since users (citizens) carefully construct the queries to minimize the ambiguity and convey specific information need as ambiguity in query formulation result in additional costs.

The highest RTI query-length in the collected data is 2362 whereas for web queries it is 245 for the AOL search engine [11]. For both web and RTI queries, the longer query-length data follow a power-law distribution. For shorter-length queries, the queries with cost associated (RTI) follow multiple power-law models. If we observe the ϕ-values from Table 5.9, we see that the slopes are very high compared to the typical value of $2 < \phi < 3$ for the web-log queries [57]; the minimum value is 6.47 and the maximum goes as high as 10.5. This shows that the fitted power-law models have a very steep slope. For each fragment of the data that follows power-law (for example, the last data in Table 5.9 from 26 to 31), the power-law distribution quickly decays within a short-range (in the range $[28, 31]$ from Table 5.8). In this experiment, we have aggressively pursued to find distribution fits for any range of the query-length data. The fitting of multiple power-law models over the entire query-length data indicates that the data distribution is very non-uniform in nature.

5.5 Experimental Results Using Query-Reply Time

In the previous section, query-length distribution was examined. This section presents a distributional analysis of reply time of sub-queries within an RTI application belonging to a given query-category. In particular, the power-law distribution is fitted to every query-category reply-time data. Power-law parameters X_{min} and Φ are estimated and the KS-test is performed. The estimated parameters and the goodness-of-fit test result (p-value) for the 26 query-categories are given in Table 5.10.

The query-categories whose p-value is 0.1 or less are presented in bold letters in the table, and denotes query-categories *for which power-law is not a good fit*. It is observed that the *Awards* query-category has least X_{min} value indicating that majority of the data follow power-law for this category, and all the other query-categories that follow power-law have $X_{min} \leq 30$ except 'Administration'. The value of X_{min} for 'Administration' is 44, meaning that only the tail part of the data fits power-law; reply duration of 44 days and above follows a power-law distribution. This observation is crucial as we want to estimate the probability of reply at 30 days. For 'Administration', power-law does not fit for data at reply day $X = 30$, hence using power-law for this query-category to estimate reply probability within 30 days is not useful.

A large value of Φ indicates a steeper slope. A steeper slope signifies a small window of reply time compared to a broader slope. Such curves decay quickly and the probability of reply becomes negligible as the number of days increases. In the context of RTI replies, such a query-category with high Φ has a short range of time-period in which most of the queries are replied, hence all institutions have more or less similar reply-durations for queries in such query-categories. From Table 5.10 we observe that the query-category 'Rules' has the steepest

Query categories	X_{min}	Φ	$p - value$
Academic	25	2.55	0.34
Accommodation	26	3.81	0.24
Administration	44	3.05	0.53
Admission	25	2.57	**0.03**
Affiliated	14	2.12	0.57
Awards	3	1.58	0.26
Collaboration	21	1.97	0.78
Committee	18	1.90	0.18
Constitution	30	2.43	0.42
Course	29	2.49	0.80
Department	22	5.02	0.25
Employment	23	2.51	**0.04**
Exam	28	2.14	0.50
Facilities	24	3.39	0.37
Finance	20	2.41	0.30
Infrastructure	28	4.34	**0.01**
Initiatives	26	2.85	0.66
Recruitment	19	2.34	**0.10**
Research	17	2.14	0.26
Results	18	2.15	**0.06**
RTI	30	2.27	0.67
Rules	28	6.42	0.71
Salary	16	2.34	0.81
Staff	24	2.42	**0.02**
Student	25	3.45	0.14
Tender	11	1.86	0.17

Table 5.10: Estimated parameters and goodness of fit value of each query-category after fitting power-law distribution.

slope with $\Phi = 6.42$; institutions across India have less deviation in reply-duration for queries regarding 'Rules'.

A smaller value indicates a broader, lower slope. A query-category that has a broader slope has a larger window of reply durations, meaning that there is a large deviation between the reply durations of institutions for that query-category. For example, at one extreme an institution may reply as early as within five days whereas another institution may take 50 days or more to reply to such queries in query-categories with low Φ value.

The transparent institutions provide replies swiftly and have small reply-durations. Non-transparent institutions on the other hand have high reply-durations for queries in the same query-category. 'Tender' has the smallest slope with $\Phi = 1.86$. Queries seeking information on 'Tender' have a probability of reply over a large time-frame. The implication of such an observation is that (i) the differences in the reply-time for 'Tender' related queries across the country is significant (ii) the reply-time is dependent on the transparency of the individual

institutions. The impact of such query-categories with small Φ is that if more queries on Tender are asked to less transparent institutions, such institutions will result in more instances of *high reply time* for the RTI queries. This would lead to ineffective implementation of the RTI Act, which states that all RTI queries irrespective of the query-category should be replied within 30 days. This demonstrates an important underlying pattern in the RTI system: the overall efficiency of this system is largely dependent on the *categories of queries* that citizens ask.

5.5.1 Comparison with Alternate Distributions

For the 20 query-categories for which power-law is accepted as a good fit, Vuong's test is used to examine if power-law is the best fit. The Log-Normal and the Exponential distributions are compared with power-law distribution. The result of Vuong's test is given in Table 5.11. It is observed that for 'Administration' query-category, the p-value is less than 0.1, and the LR shows that power-law is *not* a good fit for that query-category. This works in our favor since it was already concluded in the previous result that the power-law fit is not useful for estimation at reply-time $X = 30$ days for this query-category. For the remaining query-categories, the LR value for none of the query-categories is significant to indicate if any of the alternate models are a better fit. Since GOF p-value for power-law is above the threshold of 0.1, power-law is understood to be the best fit for the 19 query-categories with 'moderate' support. At the end of this experiment, we have a total of seven query-categories where power-law is not the best fit. They are *Administration, Admission, Employment, Infrastructure, Recruitment, Results*, and *Staff*.

5.5.2 Candidate Distributions

For the seven query-categories that do not follow power law, we consider seven alternate probability distributions (Table 5.4) for testing and follow a methodical approach to discard at each step distributions that are not a good fit, to ultimately arrive at the best fit. This is done for all the seven query-categories. We perform the steps as enumerated in subsection 5.4.5; with input as query-category reply time data. The final selected best-fit distributions are enumerated in Table 5.12. Both 'Staff' and 'Administration' follow Burr's distribution, and for the remaining five query-categories log-logistic is observed to be the best fit model. At the end of this step, all 26 query-categories are now assigned a distribution model that best fits the given data.

5.5.3 Quantifying Transparency via Cumulative Probability Distribution

Section 5.5 characterized the query-category reply time data using various probability distributions. The parameters obtained in the distributional fit are used to quantify the probability of getting a reply in a specified number of days. In this section we are interested in the quantity P($X \leq 30$), that is, what is the probability of getting a reply within 30 days when the query belongs to a given query-category. We term this quantity as 'transparency' as it fits the RTI Act 2005 time frame given for the PIO to reply to RTI queries. The more this probability, the more transparent is the given query category when measured across institutions

Categories	Log-normal		Exponential	
	LR	p-value	LR	p-value
Academic	0.15	0.88	1.52	0.13
Accommodation	0.17	0.87	1.11	0.27
Administration	-7.04	1.91E-12	-8.25	1.54E-16
Affiliated	-0.54	0.59	1.58	0.11
Awards	-0.90	0.37	0.39	0.70
Collaboration	-0.32	0.75	0.86	0.39
Committee	-0.28	0.78	0.16	0.87
Constitution	-0.40	0.69	-0.52	0.60
Course	-0.39	0.70	0.42	0.68
Department	-1.06	0.29	2.21	0.03
Exam	-0.31	0.76	1.58	0.11
Facilities	-0.86	0.39	-1.49	0.14
Finance	0.29	0.78	1.67	0.09
Initiatives	-0.13	0.90	0.65	0.51
Research	-0.28	0.78	1.11	0.27
RTI	-0.20	0.84	1.04	0.30
Rules	0.07	0.95	0.87	0.38
Salary	-0.05	0.96	1.32	0.19
Student	0.05	0.96	1.37	0.17
Tender	-0.69	0.49	0.72	0.47

Table 5.11: Normalized log likelihood ratios (LR) and p-values obtained with log-normal and exponential distribution for query-categories for which power law is a good fit.

in the country. In particular, we consider 0.75 (or 75%) as the threshold for transparency and examine all query-categories whose $P(X \leq 30) > 0.75$. The government of India does not discuss on what percentage of reply is considered as transparent. This value of 75% or above as 'transparency threshold' is not an established value but is taken as a threshold value in this experiment.

For estimating the $P(X \leq 30)$ for each query-category, we take the parameters of the respective fitted distribution and estimate the Cumulative Distribution Function (CDF) for each query-category. This gives us the cumulative probability of reply at the 30-day point in the time axis. The final probabilities estimated via Cumulative Distribution Functions are enumerated in Table 5.14.

Out of the 26 query-categories, 19 query-categories follow a power-law distribution, five query-categories follow a log-logistic distribution and two query-categories follow Burr distribution. In addition, the probability of reply for each query-category is different (as observed from the variation in the tabulated values), indicating that there is a variation in getting a reply to an RTI query based on the *category of query* asked. The overall probability of reply to RTI queries across India stands at 0.6, which is below the threshold of 'transparency'. For the query categories that do not follow power-law distribution the probability of a query

Query-category	Best fit
Administration	Burr
Admission	log-logistic
Employment	log-logistic
Infrastructure	log-logistic
Recruitment	log-logistic
Results	log-logistic
Staff	Burr

Table 5.12: Best-fit distribution for the seven query-categories that do not follow power-law.

Query-Category	P($X \leq 30$)
Department	0.92
Exam	0.81
Research	0.77

Table 5.13: List of query-categories where P(X $\leq$ 30) $>$ 0.75

getting replied within 30 days is 0.48. Query categories for which P($X \leq 30$) is observed as above 0.75 are given in Table 5.13

Only three query-categories can be termed as *transparent*, and have more than 75% reply rate within 30 days across the country. From Table 5.14 we observe that the query-category 'Admission' is one of the least-transparent query-category in terms of getting a reply to RTI queries. Since our collected RTI data belongs to educational institutions, having low transparency to Admission related RTI queries is a counter-intuitive point. Another non-transparent query-category to notice is 'Recruitment'. Despite having specified deadlines for the recruitment process across all the institutions, such time-sensitive RTI queries belonging to this category has low probability in getting a reply within the stipulated time-frame. In addition, queries seeking information regarding 'Administration' and 'Infrastructure' also have a low probability of reply across India. The ten least-transparent query-categories are enumerated in descending order in Table 5.15.

5.5.4 Summary of the Results

The experiment performed on query-category reply time data has given insights into several new information, and can be summarized as follows: (1) RTI reply-times are not uniform but have different distributional characteristics for different query-categories. 19 query-categories follow a power-law distribution and the rest follow log-logistic and Burr distribution. (2) The quantity P($X \leq 30$) is different for the different query-categories, thereby indicating that *depth of access* is not allowed by individual institutes across the country making them less

Query-Category	$P(X \leq 30)$
Academic	0.63
Accommodation	0.55
Administration	0.56
Admission	0.52
Affiliated	0.69
Awards	0.69
Collaboration	0.40
Committee	0.60
Constitution	0.46
Course	0.60
Department	0.92
Employment	0.54
Exam	0.81
Facilities	0.62
Finance	0.62
Infrastructure	0.41
Initiatives	0.27
Recruitment	0.52
Research	0.77
Results	0.60
RTI	0.70
Rules	0.65
Salary	0.67
Staff	0.58
Student	0.61
Tender	0.50

Table 5.14: Probability of getting a reply within 30 days for all 26 query-categories.

transparent. By computing the probability of reply, we have quantified 'transparency of query-categories' (3) The query-categories with low Φ-value are potential reasons for influencing the transparency of institutions. Such query-categories increase the deviation between reply-times for transparent and non-transparent institutions.

5.6 Discussion on Amendments

As discussed in Chapter 2, the RTI Act 2005 is brought to increase transparency in the governmental institutions. By quantifying 'transparency' in this Chapter, we get to observe patterns in the quantified parameter across query-categories. This leads us to understand the disparity between the practical execution of the RTI Act and the actual RTI Act that exist on paper. This disparity is where amendment scopes can be found.

Query-Category	$P(X \leq 30)$
Administration	0.56
Accommodation	0.55
Employment	0.54
Admission	0.52
Recruitment	0.52
Tender	0.5
Constitution	0.46
Infrastructure	0.41
Collaboration	0.4
Initiatives	0.27

Table 5.15: Ten least transparent query-categories across India with their reply-probabilities within 30 days in descending order.

From the quantified values as presented in the previous sections, there is a large variation in the transparency of the query-categories, and some query-categories are highly non-transparent. From the view-point of the citizens who wants transparency of government affairs, this experiment reveals that the extraction of timely information about the government is visibly based on the query-categories. Thus, any proposed amendment is in the light of improving transparency of the query-categories. The uncertainty in the reply durations for some queries is what makes the RTI system inefficient for the citizens.

5.7 Summary

This Chapter analyzed two important parameters of the RTI query log data namely (i) RTI query length and (ii) RTI query-category reply time. For both of these parameters, power-law distribution moderately fits the given RTI data.

For the query-length distribution analysis, it is observed that RTI applications with longer query-length are rare, and they also have very high reply-time. This observation is along similar lines of the proposed amendment by the government of India to restrict the RTI query-length to 500 words. For the query-category reply time analysis, it is observed that time-bound categories like Recruitment, Tender have low P($X \leq 30$). The large variation in the P($X \leq 30$) across query-categories indicate a serious gap in the transparency definition *depth of access* of the information. We take forward this notion further, and perform in-depth experiment in the subsequent Chapter.

6

Latent Variable Modeling Using RTI Query Logs

6.1 Introduction

Query-category-reply time is one of the RTI properties analyzed in Chapter 5. The quantity that is estimated is the *probability of replying within 30 days for a given query category*. A higher value of this probability is an indication of *transparency*. The definition of transparency is modeled by considering whether the RTI application has been replied within 30 days period for a given query-category. That is:

1. The estimated quantity is the probability of getting a reply within 30 days for a specific query-category. The way this is achieved is by filtering the RTI query log data using the specified query-category. The resulting RTI query log data is used in estimating the probability.
2. The estimated probability is the value obtained across all the educational institutions for a specified query-category.

The quantity of interest, however, is the probability of getting a reply for a given query-category from a *particular institution*. In this Chapter, we propose the use of modeling techniques that estimates the probability methodically as given in Equation 6.1.

The probability that an institution replying to a query-category within 30 days (6.1)

This quantity is of immense value and will strengthen the quantification of transparency. From this point of view, this Chapter advances the performance assessment of FOIA discussed in [33]. The index-based methods and aggregates are computed at a coarse level (typically national level) [7, 8, 9]. The proposed method fully quantifies transparency at a fine-grain level, namely, an institute level.

In addition, this Chapter proposes a *new data model* that incorporates multiple definitions of transparency [66, 67, 68] using the collected RTI application data. We identify a suitable modeling technique to estimate the probability given in Equation 6.1. The probability given in Equation 6.1 in turn depends on two quantities namely, transparency and effectiveness of implementation which are to be estimated from the RTI dataset. These parameters are interpreted in the social context and provide pointers to amendments.

The World Bank has undertaken an extensive study on the properties of FOIA/ RTI Acts across the globe to understand the relevant aspects of these Acts [69]. Specific to the Indian context, the CIC has conducted transparency audit [7] and acknowledge that obtaining transparency values by taking multidimensional factors into account is of importance. CIC relies on summary statistics obtained from each of the PA over 30 parameters and publishes them every year [6].

Both these studies identify two important properties namely "transparency" of institutions and the "effectiveness of implementation" of the RTI Act. Transparency parameter is captured *qualitatively* by several studies that examine the FOIA/RTI Act [66, 67, 68]. There is no *quantitative* approach proposed for capturing transparency of individual institutions. In addition, the effectiveness of FOIA/RTI Act implementation is once again *qualitatively* analyzed for very few countries [69]. In this thesis, for the first time, we quantitatively capture the transparency and effectiveness of RTI Act implementation. In addition, we identify potential pointers (which are technical in nature) for amendments to the RTI Act. In particular we answer the following research questions:

RQ1: How to quantify transparency and effectiveness of the RTI Act? Quantifying transparency is of immense value to the policymakers. In particular, such quantitative measures help in attracting foreign investments[9].

RQ2: How to validate the obtained estimates? The psychometric model provides best estimates for the latent parameters given the RTI query reply statistics data. However, the goodness of the estimated parameters needs to be examined for further usage of the identified latent values.

The core contributions of this work are:

1. Demonstrate the utility of psychometric models in the context of RTI application log data.
2. Propose the applicability of the well known *Item Characteristic Curve (ICC)* in psychometric analysis to model the probability given in Equation 6.1. The ICC model is then extended in the RTI Act context and is referred to as *Query-Category Characteristic Curve* (QCCC).
3. For the first time, a model to **quantify** transparency, discriminative factors that affect transparency and effectiveness of the implementation of the RTI Act is presented. These parameters are in turn used in the QCCC model.
4. A **data model** of the RTI query, reply-statistics is proposed in the form of a two-dimensional matrix namely the institute query-category (IQC) matrix. The proposed

IQC matrix takes into account multiple definitions of transparency as suggested in the literature.

5. Validate the estimated parameters.
6. Identify pointers for amendments to the RTI Act through the interpretation of the three estimated latent parameters.

6.2 RTI Properties in Literature

In the literature varying definitions exist for properties relevant to the RTI Act/FOIA. In order to model the RTI properties, the indicators for measuring the properties are studied. These are:

1. **Effectiveness of implementation:** In order to understand the effectiveness of implementation of the RTI Act, several RTI specific indicators were suggested [69]. These are broadly categorized into the following:

   Input oriented indicators: This refers to the facilities and the internal provisions within the public institutions. Proper facilities lead to better efficiency in the RTI system hence better reply rates. Examples include the appointment of a dedicated PIO, proper infrastructures, the establishment of records management system for efficient retrieval of all information within the institution. Such indicators can also refer to institute-specific properties for a group of institutions [69].

   Output oriented indicators: These indicators measure the operational features of the RTI Act. Example measures are the number of applications received, number of applications replied, number of applications rejected (on varying grounds), the swiftness with which the replies are served within the stipulated time etc. The collectible RTI statistics like RTI applications, date of reply to those applications etc. mainly consist of these output-oriented indicators. These indicators are limited in ways such as they do not reveal whether an applicant is satisfied with the received information or not. Hence, these indicators do not offer an estimation of the quality of the information retrieved, rather it tells about the execution of the RTI Act by different PAs [69].

   Outcome-oriented indicators: This refers to the impact of the collected RTI information on the society at large, and whether it has helped bring any positive changes. These indicators are the most revealing in terms of the effectiveness of the implementation of the RTI Act, yet are the most difficult to model and quantify [69].

2. **Transparency** in general has been discussed and defined in a variety of ways [27, 28]. However, these studies do not discuss transparency specific to RTI Act or FOIA. Three distinct definitions of transparency with respect to the RTI Act given in the literature are as follows:

   2.1. Mitchel [66] defines transparency as "dissemination of regular and accurate information".

2.2. Kopits and Craig [67] define transparency as "ready access to reliable, comprehensive, timely, understandable information".

2.3. Meijer [68] states individual organization's transparency as "depth of access individual institutions allow".

3. **Discriminative factors** For the 'discriminative factors' property, there are no definitions in the literature regarding what factors influence the transparency of institutions and differentiate the transparent from the non-transparent ones. From the analysis in the previous Chapter, we have identified 'query-category' as a potential indicator that affects RTI query-reply dynamics. However, we have not come across any direct indicator(s) for measuring this quantity in the literature. We address this and quantify this parameter through the use of GRM.

6.3 RTI Data Model

In order to represent a suitable data model to quantify transparency, the data model should incorporate the attributes of transparency. To achieve this, we first examine the definition of transparency presented in the previous section. In particular, we highlight the following two definitions that are incorporated in the data model.

- Easy and *timely access* to understandable, consumable data [67].

  Timely access: We consider only those queries that are replied within 30 days time frame as per the RTI Act. When an RTI applicant gets a reply within the specified time frame of 30 days, the institute is understood to share the requested information in a timely fashion. From the citizen's perspective, the government is being transparent. From the government's perspective, the RTI Act is implemented rigorously. We note that we are not examining the relevance of the reply given the query in this definition. The only measure used is the time duration in replying.

- The ability to have *access to all types of information*, which is regarded as *depth of access* [68].

  Access to all types of information/depth of access: We examine every RTI query and categorize it into a department/section from which information is sought. We examine if the queries across all the departments are answered by the institute. Whenever an institution is allowing citizens to access information from any of its departments, we say that the institute is allowing *depth of access*.

Considering the above discussion, the RTI query-reply data is represented as a two dimensional institute, query-category (IQC) matrix. Here rows of the IQC matrix correspond to institutions and columns correspond to query-categories (departments/sections) on which questions were posed to individual institutions. An entry ji in this matrix correspond to the percentage of replies institution j has given against a query category i within the stipulated time of 30 days. Table 6.1 shows the IQC matrix obtained from the RTI data. Each element ji in the matrix is computed as a percentage $(n_1/n_2) * 100$; where n_2 represents the number

Inst. no.	administration	admission	affiliation	course	exam	finance	recruitment	RTI	staff	students
1	66.67	28.57	16.67	0	33.33	100	11.76	0	21.05	100
2	71.43	60	90.91	100	91.3	100	80	14.29	25	66.67
3	54.54	42.86	33.33	33.33	33.33	62.5	57.63	50	61.97	0
4	57.14	62.5	20	52.94	50	15	44.83	88	7.89	83.33
5	0	0	40	0	75	60	0	100	50	0
6	33.33	62.5	100	100	66.67	100	0	100	75	100
7	20	0	57.14	0	80	0	47.37	77.27	18.75	33.33
8	100	100	66.67	0	22.2	100	62.22	0	80.95	0
9	66.67	25	50	100	50	46.67	51.97	80	50	0
10	100	0	95.24	100	59.66	100	33.33	25	100	50

Table 6.1: Institute-query category (IQC) matrix with reply percentages for ten institutions and ten query categories

	administration	admission	affiliation	$\cdots$	students
I_j	96.00	98.00	100.00	$\cdots$	100.00

Table 6.2: A row vector representation of an institution I_j and its reply rate (in percentage) in multiple query-categories

	administration	admission	affiliation	$\cdots$	students
I_k	26.00	18.00	10.00	$\cdots$	19.00

Table 6.3: A row vector representation of an institution I_k and its reply rate (in percentage) in multiple query-categories

of queries received in the i^{th} query-category by institution j and n_1 represents the number of queries replied within 30 days out of the total n_2 queries received ($n_1 \leq n_2$). Thus, each value in the IQC is between 0-100%.

In addition, the IQC takes into account the three *output-oriented indicators* for measuring the performance of the RTI act [33, 69], namely the *number of queries received* by the institutions, the *number of queries replied* by the institutions and whether queries are *replied within the stipulated time limit* of 30 days.

6.3.1 Interpretation of The Data Model

We interpret the IQC matrix from the RQ1 perspective and examine how the two latent parameters transparency and effectiveness of the RTI Act implementation are captured in this data model.

Rows of IQC Matrix and Transparency

Consider two rows of an IQC matrix as shown in Table 6.2 and Table 6.3. The reply percentages presented here are only for the purpose of an example. Columns in each of these tables denote the query-categories. The values in any of these two rows indicate the reply

Institution	administration	admission
I_1	100	10
I_2	96.00	19.60
$\vdots$	$\vdots$	$\vdots$
I_N	98	28.27

Table 6.4: Columns represent query-categories and entries consisting of reply rates for that query-category for all N institutions

percentage. In this example, the institute I_j has high reply percentages (Table 6.2). It also has high reply percentages in *every* query-category. This implies that the institute I_j allows citizens to access information from *any of the departments/sections*, in turn, adhering to the transparency definition (allowing for *depth of access* and in *timely* manner).

On the other hand institute I_k (Table 6.3) has low reply percentages. It also has low reply percentages in the *majority* of the query-categories. This implies that the institute I_k does not allow citizens to access information from *majority of the departments/sections*, in turn, leaning towards a less transparent institute.

Columns of IQC Matrix And Effectiveness of Implementation

Consider two distinct columns of the IQC matrix as given in Table 6.4. The reply rates presented in this table are only for the purpose of an example. By examining every row (institution) under the column *administration* of Table 6.4, it is noted that every institute has high reply rates to this query category. This implies that RTI Act is rigorously followed in this query category across all the institutes in India. On the contrary, the query category *admission* has low reply rates across all the institutes. From this, it is inferred that the implementation of the RTI Act across Indian institutes is not effective. The columns of the IQC matrix (that is, query categories) indicate the *effectiveness* with which the RTI Act is implemented at the ground level.

Thus the IQC matrix data model captures both features of transparency and effectiveness of the RTI Act. In addition to these two latent parameters, we also model a very important latent parameter known as the *difficulty* of the query category. Queries for some departments are inherently difficult to reply compared to other departments. We explain this parameter in the following section.

6.4 Learning Model

In this section, we discuss Item Response Theory (IRT), the basic method of the psychometric modeling. We then discuss the Graded Response Model (GRM), a variant of IRT, that models non-binary responses. These two psychometric models are chosen to model (i) transparency of individual institutions and (ii) effectiveness of the RTI Act's implementation.

6.4.1 Item Response Theory

Item Response Theory (IRT) was first proposed in the literature to assess test takers' aptitude based on responses to a given set of questions [1]. Assessment is based on latent variables, namely, the *ability* of the test taker and *difficulty* of the test questions. In assessing the test taker's aptitude, the following parameters are associated with (i) each test taker and (ii) every question:

Test taker One parameter namely *ability* (θ) is associated with each test taker. The probability that a test taker gives correct response to a test question depends on test taker's (latent) *ability* (θ). The more the value of ability a test taker has the higher is the probability of giving correct response.

Test question Two parameters are associated with each test question.

The difficulty of the question β is the point on the ability scale at which the probability of correct response is 0.5. In other words, this parameter states that for a group of test-takers having *same ability value*, half of the group have correctly given the answer and half of the group have incorrectly answered the question.

Discrimination power of the question α The question which differentiates high ability test taker with low ability test taker. From Equation 6.2 we note that α is just a scaling parameter.

The quantity of interest in IRT is the *probability of correct response at a given ability to a test question* which depends on θ, β and α. This is denoted as $P(\theta) = P(\beta, \alpha, \theta)$. This probability is modeled as

$$P(\theta) = P(\beta, \alpha, \theta) = \frac{1}{1 + e^{-\alpha(\theta - \beta)}} \tag{6.2}$$

The Equation 6.2 is referred to as the item characteristic curve (ICC). Figure 6.1 depicts a sample ICC curve. This curve shows, as the ability of the test taker increases the probability of correct response to a test question increases.

Given a matrix of test-takers and test questions in which each element represents whether a test taker responded correctly or incorrectly to a test question, IRT aims to model Equation 6.2 and estimate the parameters θ (ability) of every test taker, β and α for every test question.

Estimating β and α: Assume that we know the ability values (θ's) of the test takers. Let there be k distinct ability values. For each ability value θ_j, let there be f_j test takers. That is, f_j test takers possess ability value θ_j. Of these f_j test takers, let r_j gave correct response to the test question and $(f_j - r_j)$ gave incorrect response. The response vector for all the k distinct ability values is given by: $\mathrm{R} = (r_1, r_2, \cdots, r_k)$. Given the response data of test takers on the given test questions, the *likelihood of observing the response vector* R (the probability of observing the response) is given by

$$Prob(R) = \Pi_{j=1}^{k} \frac{f_j!}{r_j!(f_j - r_j)!} P_j^{r_j} Q_j^{(f_j - r_j)} \tag{6.3}$$

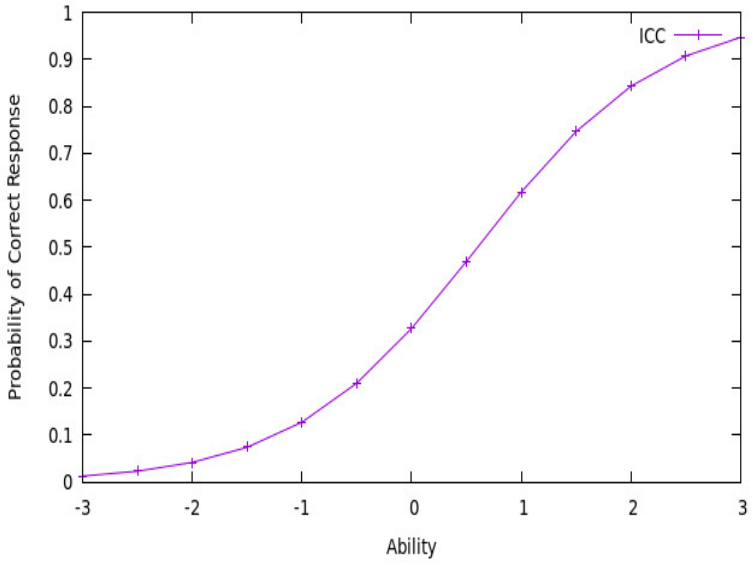


Figure 6.1: An example item characteristic curve $\beta = 0.6$ and $\alpha = 1.2$ reconstructed using the data given in [1].

Here P_j denotes the probability of correct response (whose equation is given in 6.2) and $Q_j = 1 - P_j$. Taking logarithm on both sides of Equation 6.3 we have:

$$L = \log(Prob(R)) = constant + \sum_{j=1}^{k} r_j \log P_j + \sum_{j=1}^{k} (f_j - r_j) \log Q_j \tag{6.4}$$

In order to estimate the parameters β and α, the log likelihood given in Equation 6.4 is maximized. Maximizing the log likelihood optimization formulation is solved using an iterative approach namely Newton Raphson method. Detailed derivation of parameter estimation is given in [1]. The parameters β and α are estimated for every test question by repeatedly solving Equation 6.4 for every test question.

Estimating θ_j $\forall$ test takers: Let n denote the number of questions in the test. Assume that every question's response is binary (0 or 1). Let there be N test takers. For the j^{th} test taker, let U_j denote the response vector given by the j^{th} test taker for all the n questions. That is $U_j = (u_{1j}, u_{2j}, \cdots, u_{nj}|\theta_j)$, where

$$u_{ij} = \begin{cases} 0 & \text{if answer to } i^{th} \text{ question is incorrect by } j^{th} \text{ test taker} \\ 1 & \text{otherwise} \end{cases}$$

The likelihood of observing the response vector U_j is given by

$$Prob(U_j|\theta_j) = \Pi_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{(1-u_{ij})} \tag{6.5}$$

Where $P_{ij} = P_i(\theta_j)$ and $Q_{ij} = Q_i(\theta_j)$. The equation assumes that u_{ij}'s are statistically

independent. Taking the logarithm on both sides of Equation 6.5 yields the log likelihood equation:

$$L = \log(Prob(U_j|\theta_j)) = \sum_{i=1}^{n} (u_{ij} \log P_{ij} + (1 - u_{ij}) \log Q_{ij}) \tag{6.6}$$

In order to estimate the parameter θ_j the log likelihood given in Equation 6.6 is maximized. The optimization problem is solved using the Newton Raphson iterative approach. Detailed derivation of parameter estimation is presented in [1]. The Equation 6.6 is solved for every test taker to estimate the ability parameter.

6.4.2 Graded Response Model

In the IRT, every test question takes binary response 0 or 1. For some tests however, each questions accommodate multiple categories of response such as *strongly disagree, disagree, agree, strongly agree* or *category 1, category 2, category 3, category 4*. The extension of binary response to graded response is quite natural. In the binary response case we have two quantities namely probability of correct response P_j as given in Equation 6.3 and probability of incorrect response Q_j such that $P_j + Q_j = 1$. Associated with these two quantities, only one difficulty parameter exist as the incorrect response probability curve is computed using the equation $Q_j = 1 - P_j$ therefore parameters associated with this curve need not be estimated explicitly.

In the graded response, consider $k = 1, 2, \cdots, m$ categories exist. The quantity of interest to compute P_k (where $k = 1, 2, \cdots m$) needs to be obtained such that the following equation holds $\sum_{k=1}^{m} P_k = 1$. In addition, for every question, we have $m - 1$ difficulty parameters namely $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_{m-1})$ (in the case of binary response we have m = 2 and hence only one β value is to be estimated). Associated with these m categories, we have m curves in the GRM ICC. An example ICC curve in the graded response model is given in Figure 6.2.

In the Figure 6.2 the probability of responding to category 1 is given by the curve labeled 'category 1'. In a similar manner the category 2, category 3 and category 4 curves are interpreted. For binary categories, Figure 6.1 is re-produced in Figure 6.3 by including the derivable curve associated with the probability of incorrect response Q_j (cyan color curve).

Given a matrix of test takers and test questions in which each element presents test taker response (from the set of categories $\{1, 2, \cdots, m\}$) to a test question, GRM aims to estimate parameters θ (ability) of every test taker, and β_k where $k = \{1, 2, \cdots, (m-1)\}$ and α for every test question.

Estimating $\boldsymbol{\beta}$ and α Assume that we know the ability values (θ's) of the test takers. Let there the G distinct ability values. For each ability value θ_g, let there be f_g test takers. That is, f_g test takers possess ability value θ_g. Of these f_g test takers, let r_{g1} gave response category 1, r_{g2} gave response category 2 and so on, r_{gm} gave response category m to the test question. Note that $\sum_{k=1}^{m} r_{gk} = f_g$ which ensures g^{th} group has exactly f_g test takers. The *response matrix*

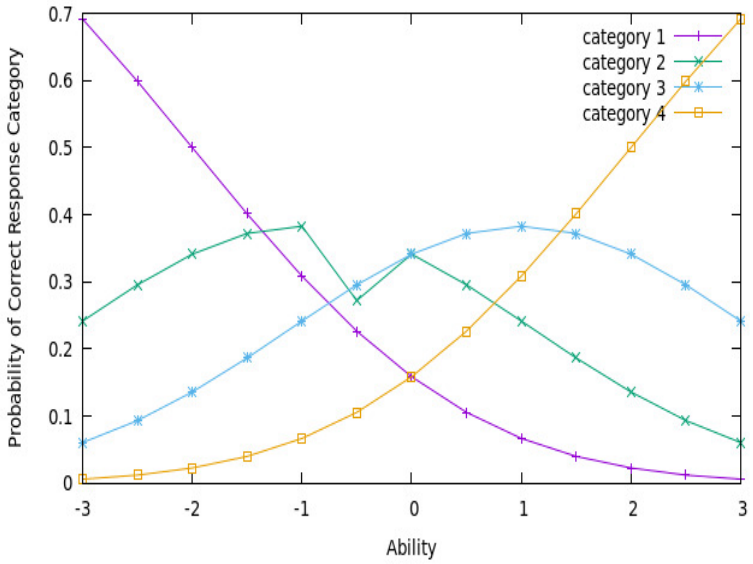


Figure 6.2: An example item characteristic curve. Reconstructed using the data given in [1].

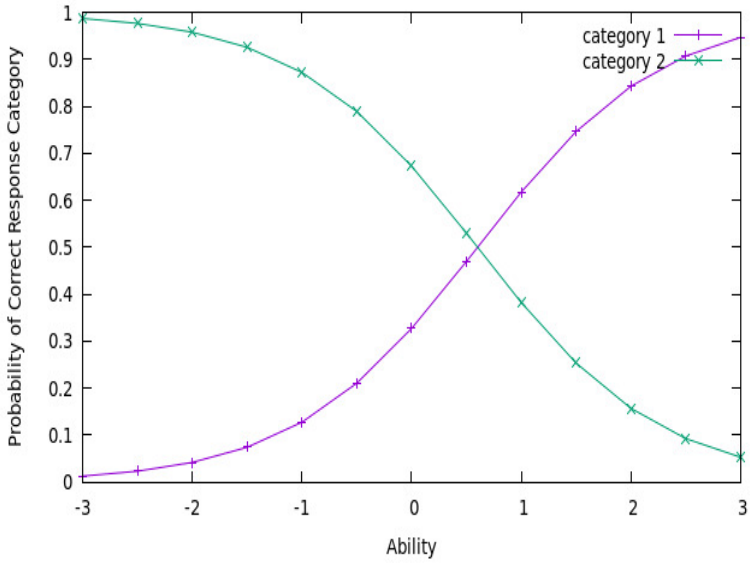


Figure 6.3: Item characteristic curve for two categories as given in 6.1. Reconstructed using the data given in [1].

for all G ability values is given by: R = $(r_{g1}, r_{g2}, \cdots, r_{gk})$ for all $g = 1, 2, \cdots, G$.

Given the test takers and test questions response matrix above, the likelihood of observing the response matrix R (the probability of observing the response) is given by:

$$Prob(R|\boldsymbol{\beta}, \alpha, \theta) = \Pi_{g=1}^{G} \frac{f_g!}{r_{g1}! r_{g2}! \cdots r_{gm}!} P_{g1}^{r_{g1}} P_{g2}^{r_{g2}} \cdots P_{gm}^{r_{gm}} \tag{6.7}$$

The logarithm on both sides of the Equation 6.7 we have the log likelihood equation:

$$L = \log(Prob(R|\boldsymbol{\beta}, \alpha, \theta)) = \text{constant} \sum_{g=1}^{G} \sum_{k=1}^{m} r_{gk} \log P_{gk} \tag{6.8}$$

The parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_{(m-1)})$ and α are solved by maximizing the log likelihood equation given in 6.8 by using Newton Raphson iterative optimization method. Details of the derivations are given in [1]. These parameters are estimated for every test question.

Estimating θ_j $\forall$ test takers Let n denote the number of questions in the test. We assume that every question's response is from the set $\{1, 2, \cdots, m\}$. Let there be N test takers. For the j^{th} test taker, let U_j denote the response vector given by the j^{th} test taker for all the n questions. That is $U_j = (u_{1k}, u_{2k}, \cdots, u_{nk})$. Where

$$u_{ik} = \begin{cases} 1 & \text{if } j^{th} \text{ test taker choose response category } k \text{ for the } i^{th} \text{ question} \\ 0 & \text{otherwise} \end{cases}$$

The likelihood of observing the response vector U_j is given by

$$Prob(U_j|\boldsymbol{\beta}, \alpha, \theta_j) = \Pi_{i=1}^{n} \Pi_{k=1}^{m} P_{ik}^{u_{ik}} \tag{6.9}$$

Taking the logarithm on both sides of Equation 6.9 yields the log likelihood equation:

$$L = \log(Prob(U_j|\boldsymbol{\beta}, \alpha, \theta_j)) = \sum_{i=1}^{n} \sum_{k=1}^{m} u_{ik} \log P_{ik} \tag{6.10}$$

In order to estimate the parameter θ_j the log likelihood given in Equation 6.10 is maximized. The optimization problem is solved using Newton Raphson iterative approach. Detailed derivation of parameter estimation is presented in [1]. The Equation 6.10 is solved for every test taker to estimate the ability parameter.

6.5 Use of GRM in the RTI Context

The psychometric model discussed in the previous section namely GRM is applied in the RTI context. In particular, every governmental educational institute is viewed as a test taker. Every query-category is viewed as a test question. We are interested in modeling

Category	Meaning	Reply rate values
1	Low reply rate	[0, 25)
2	Average reply rate	[25, 50)
3	Good reply rate	[50, 75)
4	Very good reply rate	[75, 100]

Table 6.5: Mapping of the percentage of replies to graded response or category. Every element of the IQC matrix is replaced with a category value between 1 and 4. For example, category 1 in IQC matrix suggest that reply rates are between [0, 25).

Inst. no.	administration	admission	affiliation	course	exam	finance	recruitment	RTI	staff	students
1	3	2	1	1	2	4	1	1	1	4
2	3	3	4	4	4	4	4	1	2	3
3	3	2	2	2	2	4	3	3	3	1
4	3	3	1	3	3	1	2	4	1	4
5	1	1	2	1	4	3	1	4	3	1
6	2	3	4	4	3	4	1	4	4	4
7	1	1	3	1	4	1	2	4	1	2
8	4	4	3	1	1	4	3	1	4	1
9	3	2	3	4	3	2	3	4	3	1
10	4	1	4	4	3	4	2	2	4	3

Table 6.6: Transformed IQC matrix 6.1 containing reply percentages for ten institutions and ten query categories

Equation 6.1 Following are the differences in mapping (institute, query-category) as (test taker, test question):

1. In the case of IRT, every test-taker is presented with a fixed set of *test questions*. In case of RTI, institutes respond to a diverse set of RTI queries.

2. In the case of IRT, one test-taker responds to one test question. In the case of RTI, one institute responds to several RTI questions belonging one query-category. In the IQC data matrix representation, this fact is represented in terms of percentage of replies given by the institute in a given query-category.

To apply IRT/GRM in the RTI context, we address the above two differences in the following manner:

1. Every institute replies to a fixed set of *query categories*.

2. The percentage of queries replied in the IQC matrix given in section 6.3 is converted as a graded response as given in Table 6.5.

   The IQC matrix presented in Table 6.1 is converted as a graded response as given in Table 6.6

3. The graded IQC matrix given in Table 6.6 is interpreted as an institute replying to a query-category with one of the response from the set (1, 2, 3, 4).

The above two changes will allow the mapping (test taker, test question) to (institute, query-category) and in turn, allow the use of GRM in the RTI context to model. In addition, we interpret the parameters of test-taker and test question as given in section 6.4 in the context of RTI as follows:

Institute One parameter namely *transparency* is associated with each institute. The probability that an institute gives *timely* response in a query-category depends on the institute's (latent) *transparency* (θ) value. The more transparent an institute is, the higher is the probability of giving a *timely* response. Through this, we, for the first time model and computationally *quantify* transparency. That is, Equation 6.1 is expressed in terms of transparency as a parameter to estimate the required probability.

Query-category The parameters $\boldsymbol{\beta}$ and α are associated with each query-category.

Effectiveness of implementation $\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ (as there are four categories as given in Table 6.5) is the point on the transparency scale at which the probability of *timely* response is 0.5. In other words, this parameter states that for a group of institutions having the *same transparency value*, half of the group have given *timely* response and half of the group have not given *timely* response.

Discrimination power of query-category α The query-category which differentiates highly transparent institute with the less transparent institute. From Equation 6.2 we note that α is just a scaling parameter.

The above two interpretations address the RQ1 presented in section 6.1. In order to compute the probability 6.1 we use the ICC equation in terms of transparency, effectiveness and discrimination power of query-category parameters and estimate the parameters $\boldsymbol{\beta}$, α and θ as described in section 6.4. Equation 6.2 in the context of RTI is referred to as query-category characteristic curve (QCCC).

6.6 Algorithm for Computing the Parameters

Algorithm 1 presents the estimation of parameters namely θ, α and $\boldsymbol{\beta}$ via the Maximum Likelihood Estimation (MLE) using the IQC matrix. The detailed solution for the equations can be found in Baker and Kim [1].

6.7 Validation of the Estimated Parameters

In order to address the second research question given in section 6.1, the obtained transparency value of each institute is to be compared with the ground truth transparency value of the respective institute. However, there is no published data by any government establishment with respect to individual institute's transparency scores. In the absence of the ground truth information on individual institutes transparency values, we assess the *quality of the estimated*

Algorithm 1 Algorithm for estimating the RTI parameters

Input: IQC matrix

Output: Estimates of θ, α, $\boldsymbol{\beta}$ and $P_i(\theta_j)$ using Maximum Likelihood Estimation procedure

1: For each threshold between categories (or grades) $k-1$ and k, divide the data in the IQC into binary form: reply-rates above the threshold are assigned '1', below the threshold are assigned '0'.
2: **for** Each threshold between categories (or grades) $k-1$ and k **do**
3: Initialize $\hat{\theta}$
4: Initialize $\hat{\boldsymbol{\beta}}$
5: Initialize $\hat{\alpha}$
6: **for** Each query-category **do**
7: Maximize the log likelihood given in Equation 6.8 using Newton Raphson method to obtain $\boldsymbol{\beta}$ and α
8: **end for**
9: **for** Each Institute **do**
10: Maximize the log likelihood given in Equation 6.10 using Newton Raphson method to obtain θ
11: **end for**
12: **end for**
13: Obtain the estimates $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$
14: Use the final estimates to compute the probability an institute replying to an RTI query in the given query-category within 30 days time.

parameters. This is achieved by measuring the variance of the estimated parameter(s) given as

$$S^2_{\hat{\theta}_j} = \frac{1}{-E[\frac{\partial^2 L}{\partial \theta_j^2}]} \tag{6.11}$$

Limitation with the Equation 6.11 is that the computed variance depends only on the values of the parameters of the n query-categories. The variance of the estimated parameters for each of the category is computed in an alternate way given as: $\sigma^2 = \frac{1}{I_i(\theta)}$ where $I_i(\theta)$ is the information function for a query-category and is given by:

$$I_i(\theta) = \frac{\left[\frac{\partial P_i(\theta)}{\partial \theta}\right]^2}{P_i(\theta)Q_i(\theta)} \tag{6.12}$$

The total information function (IF) is given by $\sum_{i=1}^{n} I_i(\theta)$.

Inst. no	Name	Operation Level	Type	Location	No. of queries
1	Central Institute of Fisheries Education	Central	University	Middle	126
2	Mizoram Board of School Education	State	Board	North-East	107
3	Tata Institute of Fundamental Research	Deemed	University	South	425
4	Jawaharlal Nehru Architecture & Fine Arts University	State	University	South	324
5	Telangana Open School Society	State	Board	South	48
6	Jagadguru Sri Shivarathreeswara University	Deemed	University	South	29
7	Assam Agricultural University	State	University	North-East	174
8	Bodoland University	State	University	North-East	130
9	Koneru Lakshmaiah Education Foundation	Deemed	University	South	38
10	Central Board of Secondary Education Assam	Central	Board	North-East	275

Table 6.7: List of ten institutions

6.7.1 Interpreting the Information Share

Let us consider $\hat{\theta}$ to be the maximum likelihood estimator of the underlying transparency value θ. The estimated $\hat{\theta}$ has mean $\bar{\theta}$ $(=\theta)$ and variance as ρ^2. According to the concept derived from Fisher Information, the information is given by the reciprocal of the variance, that is, $\rho^2 = 1/I(\theta)$. Smaller the variance of the estimated $\hat{\theta}$, more precise is the estimate of the unknown transparency θ. This means that greater the amount of information at a given transparency level θ, more closely the maximum likelihood estimates of the transparency $(\hat{\theta})$ cluster around the true but unknown transparency level θ, hence more precise is the estimate $\hat{\theta}$.

Evaluation using information function is not done by the absolute value of the information obtained as we do not have any benchmark for comparing the estimation quality of the RTI parameters. Instead, the percentage of the information obtained as compared to the maximum information possible for the given data distribution is used. Higher the percentage, better is the estimation hence more accurate are the estimated parameters. This percentage value is used to understand the quality of the estimated parameters after the experimental results are obtained.

6.8 Experimental Results

Data from a total of 10 institutions belonging to 10 distinct query-categories and comprising of 1676 queries have been processed and used to build the IQC data matrix (Table 6.1). The IQC matrix representation and analysis via GRM require that all institutions have some queries received in all query-categories. The raw RTI data have been categorized into 26 query-categories. However, not all the institutions (considered in the IQC matrix) have received queries in all the query-categories. Thus we have selected those institutions that have data in the query-categories considered; IRT/GRM considers that all institutions have replies in all query-categories in the IQC column *(every test taker provides responses to all the test questions)*. Query-categories where reply data is not available for all institutions have not been considered. The IQC matrix has 5 state institutions, 3 deemed institutions and 2 central institutions. The institutions considered for experimentation are listed in Table 6.7. In order to estimate the parameters, the IQC matrix has been transformed by replacing the reply rates with their class values and is shown in Table 6.6.

6.8.1 Transparency (θ)

The obtained transparency values (in descending order of transparency) are presented in Table 6.8. The variance of each estimated transparency value is also shown in the table. Theoretically, θ ranges from -∞ to +∞. However, practical values lie within -4 to +4. It is observed that even the most transparent institution (institution 8) has a low value of $\theta = 1.952$, indicating that reply-rates to Indian institutions is on the lower end. Even if the institution has high reply-rates to most RTI queries, certain query-categories have extremely poor reply rates bringing down the overall transparency of that institution. An additional pattern that is observable from the quantified transparency values is that all the southern Indian institutions have a low transparency rate compared to the rest of India, given the data that we have used for the experiments.

Inst. no.	Transparency (θ)	Variance	Location
8	1.952	0.093	N-East
1	0.661	0.127	Middle
10	0.659	0.244	N-East
2	0.593	0.252	N-East
3	0.46	0.077	South
9	-0.063	0.061	South
4	-0.134	0.076	South
6	-0.141	0.075	South
5	-1.036	0.063	South
7	-1.457	0.084	N-East

Table 6.8: Transparency (θ) parameter of each institution sorted decreasingly

6.8.2 Discriminating Factors (α)

This parameter separates the institutions from most transparent to least transparent given a query-category. The higher the value of α, the higher is the separation between reply-rates

Query-categories	β_1	β_2	β_3	α
administration	-0.791	-0.304	0.941	4.179
admission	-0.543	0.468	1.931	1.663
affiliation	-8.8	-2.51	5.473	0.157
course	-1.598	0.242	2.065	0.240
exam	1.637	0.645	-0.403	-2.848
finance	-1.013	-0.615	-0.226	2.049
recruitment	-1.276	0.849	3.777	0.672
RTI	0.697	0.483	0.209	-4.325
staff	-1.039	-0.456	1.235	0.827
students	10.42	-0.167	21.277	-0.040

Table 6.9: Query-category parameters after running the Graded Response Model on RTI data

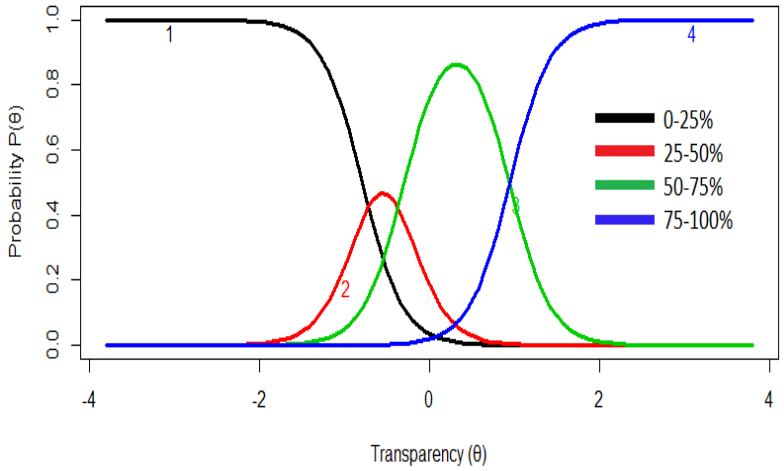


Figure 6.4: Query-Category Characteristic Curve for *Administration*

of institutions. This implies that query-categories with a high α value serves as a decisive factor in segregating transparent institutions from non-transparent ones. Table 6.9 lists the α parameter for all the 10 query-categories. Three different kinds of α values are observed: high α, low α and negative α. We discuss each kind and the implication of the values in the RTI context.

From the table, it is observed that *administration* query-category has the **highest** value of 4.179. Meaning of this parameter is that the 'administration' category largely affects and differentiates the transparency of institutions. The QCCC for 'administration' is shown in Figure 6.4. The curves are high and steep. Let us consider the blue curve in the QCCC (reply probability above 75%). Two institutions in the θ-axis close to one another have a sharp difference in the corresponding $P(\theta)$ in the vertical axis. In comparison, let us take the query-category 'recruitment' with a **low** discriminating power having $\alpha = 0.672$, whose QCCC is shown in Figure 6.5. It is visually detectable that the curves are flatter in comparison to 'administration'. The inference is that institutions having different θ values in the x-axis do not register a major difference in the y-axis ($P(\theta)$). In other words, *transparent and non-transparent institutions have similar behavior* in terms of reply rates for such query-categories having low α values. The difference in the α values for the query-categories indicates that *query-category* indeed is a discriminating factor that affects the transparency of institutions.

When α parameter takes **negative** value for certain query-category, it implies that the non-transparent institutions have a high probability of replying to queries belonging to that query-category; at the same time, highly transparent institutions have a low probability of replying in the same query category. This anomalous behavior is demonstrated by the QCCC of the 'RTI' query-category shown in Figure 6.6, where the blue line indicating RTI reply-rates in the 75-100% range is towards lower transparency (θ) values. From Table 6.9, we note that RTI category has the least α value of -4.325; for this query-category, the least transparent institution has high probability (≈ 1.0) of reply above 75%. Institutions with overall high transparency have a strikingly low probability for high reply rates in this query-category. Such areas of the government that break intuitive patterns are cues for potential amendment scope. The QCCC figures for all the 10 query-categories are presented in appendix C.

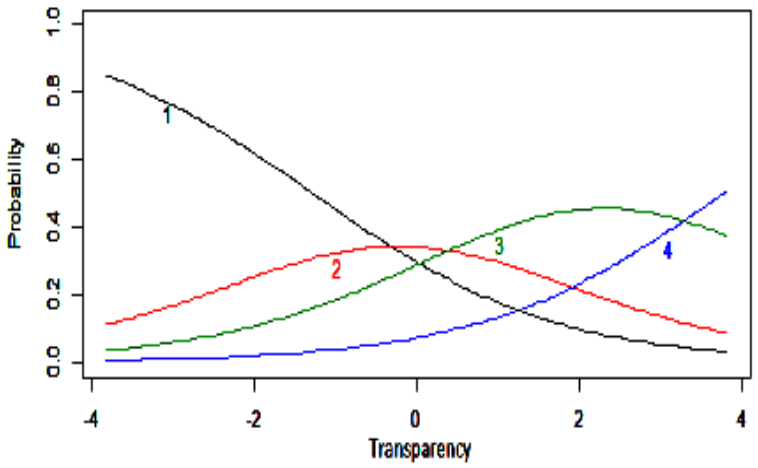


Figure 6.5: Query-Category Characteristic Curve for *Recruitment*

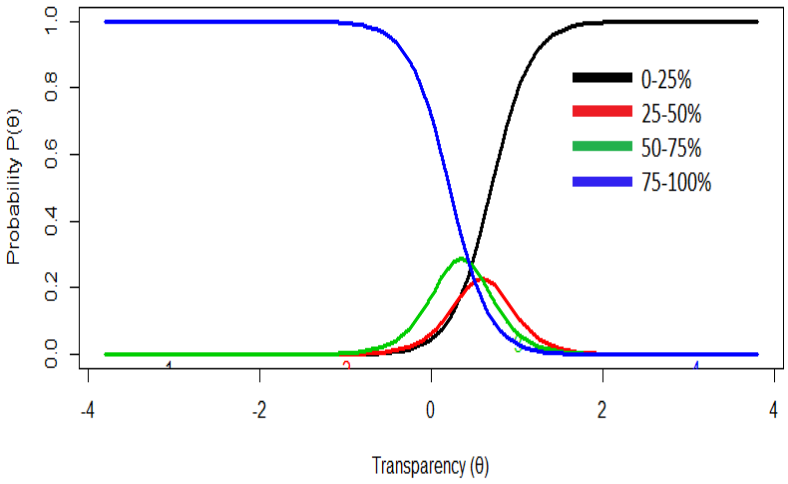


Figure 6.6: Query-Category Characteristic Curve for *RTI*

6.8.3 Effectiveness of Implementation of the RTI Act (β_3)

As discussed in section 6.3.1, we consider [75,100%] reply-rate by institutions as a measure of effective implementation of the RTI law. This is the β_3 value denoting the 'difficulty' level of a query-category at the threshold of 75%, enumerated in Table 6.9. Table 6.9 also presents β_1 and β_2 values. For the interpretation, we use β_3 alone.

High β_3 value: (Please refer to Table 6.9) We observe that the categories *affiliation, recruitment* and *courses* have high threshold values of 5.47, 3.78, 2.065 respectively. This means that an institution should have a transparency value equal to or greater than 5.47 (3.78 or 2.065) in order to reply to queries belonging to affiliation (recruitment or courses) query-category for reply rates to be between [75, 100]%. We also observe that the maximum transparency value any institution has achieved is 1.952, that is, even the most transparent institution has a very low probability of replying (reply rates between [75, 100]%) (approxi-

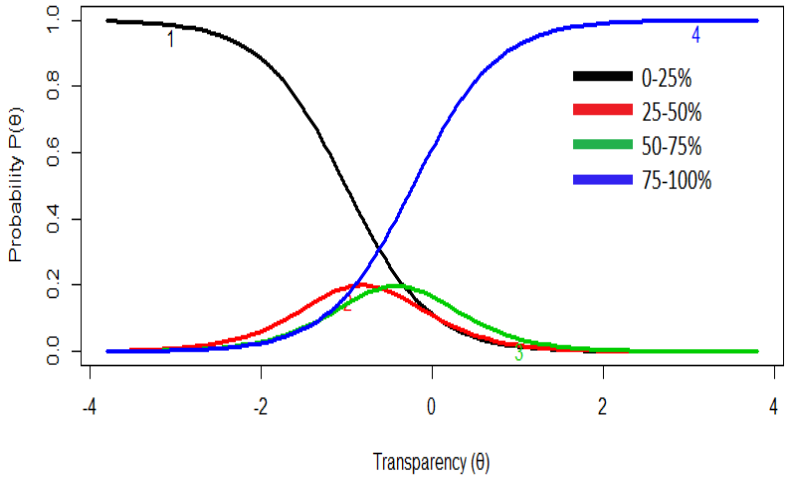


Figure 6.7: Query-Category Characteristic Curve for *Finance*

mately 0.4, 0.2 and 0.1 respectively). This is visually verified from the QCCC for 'recruitment' in Figure 6.5, where the blue line (denoting reply-rates between [75, 100]%) is quite flat and low. It is to be noted that considering the pointers provided in the literature, the present model takes into account whether a query has been *replied or not* and does not examine whether the reply has satisfied the citizen's information need. In this context, it is surprising to note that queries belonging to `courses` category have a low probability of getting a reply from all the *educational* institutions, indicating *poor effectiveness of the implementation* of the RTI Act in several sections of government educational institutions.

Low β_3 value: (Please refer to Table 6.9) The least β_3 value observed is -0.226 for the category finance (only positive α are considered). Eight out of ten institutions have transparency value more than -0.226 suggesting that queries belonging to the finance category has high probability of getting a reply (probability value close to 1.0 as seen in Figure 6.7) from institutions across India. 'Finance' is thus an easy query-category with regards to getting RTI replies, indicating that the RTI Act is implemented effectively for this section of government institutions across the country. Three out of ten categories have threshold value less than 1.952 (maximum transparency value achieved by any of the 10 institutions). These categories are administration (0.941), admissions (1.931) and staff (1.235). Queries belonging to these categories have a high probability of getting a reply from the most transparent institution namely Bodoland university. For the rest of the categories, the probability of getting a reply even from the most transparent institution is very low. However, the difference observed in the high and low β_3 values indicate that reply to RTI queries largely depend on the query-categories. An institution might be overall highly transparent but have a poor reply rate for a specific query-category, suggesting a need for some remedy query-category-wise.

6.9 Quality of the Estimated Parameters

We calculate the Information Function (IF) for the entire experiment using Equation 6.13 (in this equation the number of query-categories n is equal to 10) and calculate the percentage of

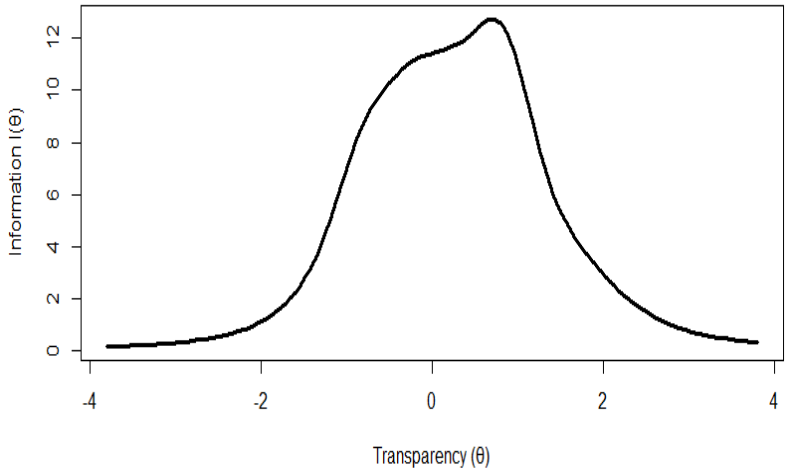


Figure 6.8: Information Curve against transparency values for all query-categories combined

information share for the range of the estimated parameters using equations. The Information curve for our experiment is shown in Figure 6.8. The information, hence the accuracy of the estimation process, is maximum at the centre and gradually decreases towards the end. The QCCCs for the query-categories depicted in the previous figures are shown within the range of -4 to +4. This range is calculated from the IF and covers 97.66% of the maximum information that the query-categories can have. Out of this, the range of -2 to +2 accounts for 89.3%. It is observed from Table 6.8 that all the institutions' transparency estimates fall within -2 to +2, suggesting that the transparencies for the institutions have been estimated with reasonable accuracy.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{6.13}$$

6.10 Interpretation Of the Parameters - Scope for a Potential Amendment

With the help of the quantified parameters we observe the following patterns:

1. Institutions have varying and low θ (transparency) values.
2. Using the α values we infer that transparency of institutions is largely dependent on query-categories.
3. The difficulty parameter β_3 shows that certain query-categories are very easy in terms of reply rates whereas some have a very high threshold, making them 'difficult' query-categories.

With the above three observed patterns, we draw the following conclusion in the form of a tentative amendment proposal: "***Reply duration should be based on the difficulty (β_3)***

of the query-category instead of 30-days irrespective of the query-category". This particular one is suggested to strengthen the effectiveness of the RTI Act's implementation. Query-category consideration already exists in the RTI Act, albeit to a limited extent. For queries pertaining to life or liberty of a person, the specified reply time is 48 hours. We propose to extend this consideration and relax the reply time for queries in the 'difficult' query-categories. Presently, queries are not being replied within the 30-day limit (as seen by the poor reply rates in our collected data). This has also been quantified in the previous Chapter as 'transparency of query-categories' by computing the probability of reply at the 30-day time limit.

6.11 Conclusion

In this Chapter, three latent patterns that can have an impact in suggesting amendments to laws have been identified and quantified. The IQC matrix representation for RTI query reply-statistics data have been included in a way that incorporates the definitions of these latent patterns as stated in the literature. Quantification has been done using GRM via maximum likelihood estimation of the data matrix. The proposed quantification of transparency is at the level of individual institutions as opposed to transparency indexes computed by various organizations which are at the ministry/department level or at the country level.

From the given RTI data, the most transparent institution and least transparent institution have been identified. Query-categories belonging to 'administration' serve as a discriminating factor in identifying transparent institution from the non-transparent institution. We have limited the data collection to educational institutions only, and observe that *courses* query-category has a high 'difficulty' value, meaning that queries belonging to *courses* have low probability of getting replies despite being educational institutions. Given the results, we have identified and proposed a tentative amendment to the RTI Act. Our experiment and analysis demonstrate that there is a significant gap between the theoretical RTI Act and its execution at the institution level across India. This establishes that reply to RTI queries belonging to certain query-categories can be very hard to receive. These are evidence of inefficient implementation.

7

Identifying Temporal Fluctuations in the RTI query-log

7.1 Introduction

In Chapter 5, RTI data was subjected to query-log analysis by fitting distributions to query-length and query-category reply time data. A one-dimensional vector representation of RTI query-category reply time data are analyzed to quantify *the probability of getting a reply within 30 days in a given query-categories*. In Chapter 6, a two-dimensional IQC representation was proposed to quantify *transparency of institutions*, *the discriminating factors that affect transparency of institutions* and *effectiveness of the implementation* of the RTI Act was presented. In the one-dimensional and two-dimensional RTI data representation, the data is static. All the RTI applications irrespective of the year of receiving them by the PIO are taken to form the data model and analysis was performed on the modeled data. In this Chapter, we incorporate time information into the RTI data model to analyse the RTI temporal dynamics.

Implementation of the RTI Act at the grass-root level is measured using two specific indicators: *inputs* and *outputs* [69]. In order to implement the RTI Act, several inputs are needed to deliver the outputs. Inputs include the appointment of the PIOs, their numbers, and office allocation. Outputs include the number of queries received and the number of queries replied by public institutions [33, 69]. However, the inputs (for example, the number of officers responsible for replying to the queries) are not taken into account in these studies.

In the present work we propose a model for the RTI query-reply data by focusing on both the input and output indicators. The RTI inputs and outputs are a function of time. The output-oriented indicators are directly related to the input namely RTI officials like PIOs interfacing the citizens and the government who are affected by the regime changes. As inputs change over time (PIO appointment, rule changes etc.), outputs have a direct bearing on the time as well. We therefore explicitly incorporate time as a feature in the RTI data model. RTI data (queries and their reply rates) is represented as a three-dimensional tensor in which government educational institutions, query-categories and time are the dimensions of the tensor. This tensor is subject to decomposition using the tensor-CUR method [70]

to identify the year in which maximum variation in reply rates are observed. The obtained factors are subject to GRM modeling to quantify *the transparency of institutions*. Following two research questions are addressed in this Chapter:

RQ1: Which time of the year maximum variation in the reply rates are observed? The RTI tensor representation consists of reply percentages. With time incorporated as the third-dimension, the tensor captures the fluctuations of the RTI query-reply data, unlike the IQC matrix where time encoding is not present (refer to Section 6.3). By decomposing the RTI-tensor, the year in which maximum variation in the reply rates is identified.

RQ2: Which of the query-categories exhibit maximum variation in reply rates? From the analysis in the previous Chapters, it is evident that *query-category* is a decisive factor in influencing transparency of institutions. A temporal analysis of query-categories is further performed to understand additional patterns that are otherwise hidden from the static analysis. We want to understand which is the most fluctuating query-category across all times; this is the query-category that impacts the transparency of institutions the most.

In addressing the above two research questions the following contributions are made in this Chapter:

A New Data Model: RTI query-reply statistics data is modeled as a three-dimensional RTI tensor in which government institutions, query-categories and time are along the first, second and third dimension respectively. Input and output definitions as present in literature are captured in this representation. This RTI tensor is composed of several IQC matrices, each IQC is constructed on RTI applications received during a specific year by the institution. To understand the RQ2, an RTI tensor with dimensions institutions, time and query category is constructed. In the second tensor, the third dimension is the query category along which the temporal variation analysis is performed.

Feature extraction: RTI tensors are subjected to tensor-CUR decomposition to capture the variations in reply-rates. The obtained factors after decomposition are employed as features in the learning model. These factors contain the highest variation in the data, thus capturing maximum information.

Learning Model: GRM is chosen to quantify the 'transparency' of individual institutions. This quantification is done on the factor matrices obtained after tensor-CUR decomposition.

7.2 Assumption

In this section the main assumption for the input oriented analysis is presented. As stated in section 6.2, the input-oriented measures explicitly take into account the appointment of PIO. The PIO duration as per the RTI Act is three years from the date of appointment.

PIO	institute	administration	admission	affiliation	$\cdots$	students
1	I_j	96.00	98.00	100.00	$\cdots$	100.00

Table 7.1: PIO_1 reply rate (in percentage) in multiple query-categories

PIO	institute	administration	admission	affiliation	$\cdots$	students
2	I_j	43.00	48.00	40.00	$\cdots$	49.00

Table 7.2: PIO_2 reply rate (in percentage) in multiple query-categories

The assumption we make is that the reply time patterns depends on the PIO. This in turn affects the way one PIO functions differs from another PIO. This functioning has a direct bearing on the replies constructed by the individual PIOs. When we segregate year-wise reply time patterns the pattern would be evident. Consider the synthetic example: assume that a PIO was appointed in the year 2008 whose duration is 3 years, that is, till 2010. During the 2010 period, this PIO produces the reply-rates as shown in Table 7.1. After 2010, another distinct PIO takes up the charge. This PIO's working characteristic would be different (if there is any) to that of the first PIO. This results in the reply-rates produced by PIO_2 is shown in Table 7.2. The tensor decomposition brings out the year in which maximum variation in the reply rates are observed in the RTI-tensor. The maximum variation in reply rates is attributed to internal changes in the input-oriented indicators at the respective institutes.

7.3 Tensors: Notations, Definitions & Literature

Following notations and definitions are provided which will be used in the subsequent sections and in the algorithm for the tensor CUR decomposition [2].

Tensor is a multidimensional array. Tensor is represented as a script letter $\mathcal{A}$. A three dimensional tensor $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$ is given in the Figure 7.1. The first dimension ($i = 1, 2, \cdots, I$) is along the vertical axis. Second dimension is along the horizontal axis ($j = 1, 2, \cdots, J$) and the third dimension is along the direction towards the Z-axis ($z = 1, 2, \cdots, K$).

Order or Mode Number of dimensions of the given tensor. The tensor in the Figure 7.1 is of order 3 (or mode 3).

Fibers Every index of the tensor is fixed except one. In a two dimensional tensor, columns denote mode-1 fiber. Rows denote mode-2 fiber. In Figure 7.2, the mode-1, mode-2 and mode-3 fibers are depicted. These vectors are represented as mode-1 as $\mathbf{a}_{:jk}$, mode-2 as $\mathbf{a}_{i:k}$ and mode-3 as $\mathbf{a}_{ij:}$.

Slices are the matrices obtained by fixing all indexes except two index. Figure 7.3 show horizontal slices, lateral slices and frontal slices of a three dimensional tensor.

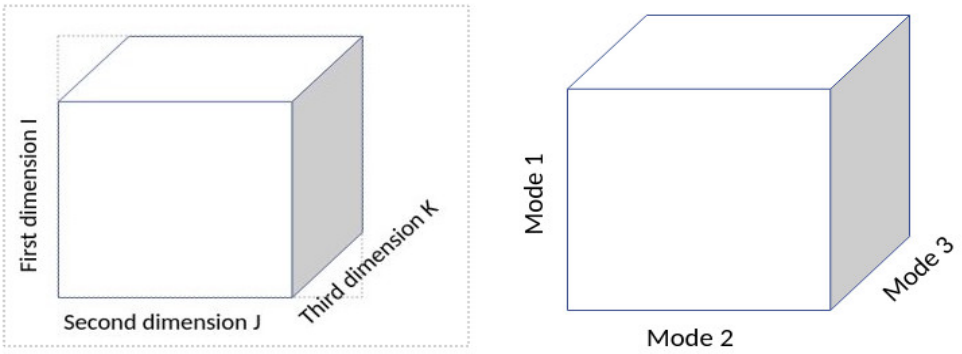


Figure 7.1: A three dimensional tensor and associated notation.

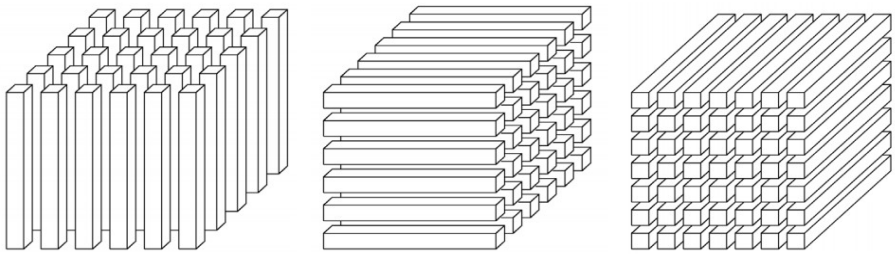


Figure 7.2: Fibers in a three dimensional tensor. Reproduced from [2]

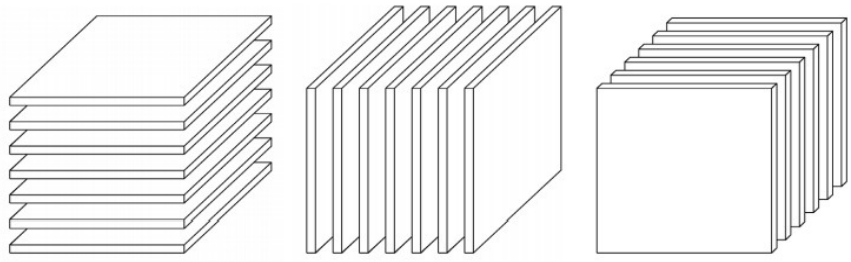


Figure 7.3: Slices in a three dimensional tensor. Reproduced from [2]

Matricization is the method of transforming a tensor into a matrix given a specific mode.

Mode-1 Matricization Represented as $\mathbf{A}_{(1)}$: A matrix is formed by joining all the frontal slices. Let $\mathcal{A}$ be a tensor with size $I_1 \times I_2 \times I_3$. Every frontal slice is of dimension $I_1 \times I_2$. These matrices are joined to form a matrix $I_1 \times (I_2 \times I_3)$. Let an example tensor has dimensions $3 \times 4 \times 2$ then the Mode-1 matricization result in a matrix of size 3×8.

Mode-2 Matricization Represented as $\mathbf{A}_{(2)}$: A matrix is formed by transposing the frontal slices and joining them. Size of the resulting matrix is $I_2 \times (I_1 \times I_3)$. Let an example tensor has dimensions $3 \times 4 \times 2$ then the Mode-2 matricization result in a matrix of size 4×6.

Mode-3 Matricization Represented as $\mathbf{A}_{(3)}$: Each frontal slice is transformed into a vector of size $(I_1 \times I_2)$. The obtained vector forms one row of $\mathbf{A}_3$ matrix. Size of the resulting matrix is $I_3 \times (I_1 \times I_2)$. Let an example tensor has dimensions $3 \times 4 \times 2$ then the Mode-3 matricization result in a matrix of size 2×12.

As tensor-CUR factorization centers uses mode-3 matricization, we re-produce the example given in [2]. Let tensor $\mathcal{A} \in \mathbb{R}^{3\times4\times2}$ contains two frontal slices each of size 3×4. Let these frontal slices be as given below:

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix} \tag{7.1}$$

Then mode-3 matricization of $\mathcal{A}$ denoted as $\mathbf{A}_{(3)}$ is given by:

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \end{bmatrix} \tag{7.2}$$

First row of $\mathbf{A}_{(3)}$ contains first frontal slice of $\mathcal{A}$ denoted by $\mathbf{A}_1$. Second row of $\mathbf{A}_{(3)}$ contains second frontal slice $\mathbf{A}_2$. Columns of $\mathbf{A}_{(3)}$ contains tubes of $\mathcal{A}$.

Applications of Tensors: Tensor representation is effective in certain applications for multi-dimensional data. Its decomposition is an efficient tool for analysis when the data contains multiple features. Acar et al. [71] used tensor to represent Internet chat room communications data with user, keyword and time as the dimensions. They used Tucker and PARAFAC decomposition [2] on the tensor data to extract (user,keyword) and (user,time) matrices, and performed comparative analysis of the tensor decompositions as well as Singular Value Decomposition (SVD) on the extracted matrices. Sun et al. [72] used tensors to represent web search queries with the dimensions as user, query and web page, and analyze the data by decomposing the tensor data by higher-order SVD (HOSVD, the tensor counterpart of matrix SVD). In yet another application, Vasilescu and Terzopoulos [73] used tensors to represent facial images, and used a methodology called TensorFaces to analyze the faces as its principal components across different dimensions of the data.

In the RTI applications data context, the inclusion of time in the RTI data model paves way for addressing the two research questions stated in Section 7.1.

7.4 Dataset & Data Model

To address RQ1 and RQ2, two RTI tensors are constructed using the RTI applications data. Towards this, the RTI data is grouped by year. The number of RTI queries received by an institute in each year is very low. In addition, these queries need to be categorized as per the query-category. This results in sparsity of the constructed RTI tensor. The data collected for 10 institutes and 1676 queries are not sufficient to construct a non-sparse RTI tensor. In addition, some institutions are new and have no RTI queries received for several years. Thus, using the collected RTI data for representing several time-intervals separately has been challenging.

To overcome the issue of sparsity in the RTI data model and subject the proposed methodology for decomposition, data from 8 institutions and 7 query-categories were considered. An year-wise (for the years {2010, 2011, 2012, 2013, 2014, 2015}) IQC matrices are created using the collected data. Each element in the matrix is *(number of queries replied within 30 days/number of queries received in a year)*100%*. This is analogous to the creation of the IQC matrix for all the RTI applications irrespective of the year of filing the RTI application in Chapter 6. The missing elements where data for a year is not present are filled with values from the previous year's IQC matrix. This occurs when an institution has not received any query belonging to a specific query-category for a year. This is done for creating six such institute×query-category matrices from 2010 to 2015. In order to address the two research questions stated above, two RTI tensors are constructed as described below:

RTI Tensor Addressing RQ1 : In order to answer the 'year' having maximum variation, we construct an RTI tensor $\mathcal{A}_{RQ1}$ having six frontal slices corresponding to the years {2010, 2011, 2012, 2013, 2014, 2015}. Each frontal slice is a matrix having dimensions (number of institutes × number of query-categories) referred to as IQC_{RQ1}. Construction of IQC_{RQ1} adopts the procedure described in Section 6.3. The difference is that in IQC_{RQ1} the RTI queries received in a specified year from the set given above are considered. The six frontal slices of $\mathcal{A}_{RQ1}$ are described below:

- 1^{st} **frontal slice** The IQC_{RQ1} is constructed by considering all the RTI queries received in the year 2010.
- 2^{nd} **frontal slice** The IQC_{RQ1} is constructed by considering all the RTI queries received in the year 2011.
- 3^{rd} **frontal slice** The IQC_{RQ1} is constructed by considering all the RTI queries received in the year 2012.
- 4^{th} **frontal slice** The IQC_{RQ1} is constructed by considering all the RTI queries received in the year 2013.
- 5^{th} **frontal slice** The IQC_{RQ1} is constructed by considering all the RTI queries received in the year 2014.
- 6^{th} **frontal slice** The IQC_{RQ1} is constructed by considering all the RTI queries received in the year 2015.

RTI Tensor For Addressing RQ2 : In order to answer the 'query-category' having maximum variation, we construct an RTI tensor $\mathcal{A}_{RQ2}$ having seven frontal slices corre-

sponding to the query-categories {administration, affiliation, course, employment, finance, medical, students}. Each frontal slice is a matrix having dimensions (number of institutes × number of years) referred to as IT_{RQ2} matrix. Construction of IT_{RQ2} is similar to the procedure described in Section 6.3. The difference is that the columns in the IT_{RQ2} matrix correspond to the years {2010, 2011, 2012, 2013, 2014, 2015}. In constructing the IT_{RQ2} frontal slice, all the RTI queries belonging to the query-category from the above-given query-category set are considered. The seven frontal slices of $\mathcal{A}_{RQ2}$ are described below:

1^{st} **frontal slice** The IT_{RQ2} is constructed by considering all the RTI queries belonging to query-category 'administration'.

2^{nd} **frontal slice** The IT_{RQ2} is constructed by considering all the RTI queries belonging to the query-category 'affiliation'.

3^{rd} **frontal slice** IT_{RQ2} is constructed by considering all the RTI queries belonging to the query-category 'course'.

4^{th} **frontal slice** The IT_{RQ2} is constructed by considering all the RTI queries belonging to the query-category 'employment'.

5^{th} **frontal slice** The IT_{RQ2} is constructed by considering all the RTI queries belonging to the query-category 'finance'.

6^{th} **frontal slice** The IT_{RQ2} is constructed by considering all the RTI queries belonging to the query-category 'medical'.

7^{th} **frontal slice** The IT_{RQ2} is constructed by considering all the RTI queries belonging to the query-category 'students'.

We have considered a total of 8 institutions, 7 query-categories and 6 time-intervals. The dimension of $\mathcal{A}_{RQ1}$ is $8 \times 7 \times 6$. The dimension of $\mathcal{A}_{RQ2}$ is $8 \times 6 \times 7$. The institutions are named as numbers 1 to 8; institutions 1 to 4 are state institutions and 5 to 8 are central institutions.

7.5 Feature Extraction and Learning Model

The RTI tensors explained in the previous section are subject to decomposition using tensor-CUR decomposition. From the obtained factors, two factor-matrices are selected. These factor-matrices are used as input to the GRM. The matrix-CUR decomposition is first discussed. This is then extended to the three-dimensional RTI tensor-CUR decomposition.

7.5.1 Matrix-CUR Decomposition

Matrix-CUR decomposition [74] is a lowrank matrix decomposition technique. This technique offers interpretability of the obtained factor matrices. In particular, this method decomposes the given input matrix ($\mathbf{M}$) into three factor matrices namely **C, U** and **R** such that the product of the three factor matrices results in the original input matrix. Interpretability of this decomposition comes from the fact that the factor matrices **C** and **R** are **subsets** of

the original matrix $\mathbf{M}$. The matrix $\mathbf{C}$ is constructed by selecting c columns from $\mathbf{M}$. $\mathbf{R}$ is constructed by selecting r rows from $\mathbf{M}$. The matrix $\mathbf{U}$ is computed such that $\mathbf{M} \approx \mathbf{C} \times \mathbf{U} \times \mathbf{R}$. In order to select c columns from $\mathbf{M}$, the following strategy is adopted [74].

Column Selection Strategy The input matrix $\mathbf{M}$ is subject to Singular Value Decomposition (SVD) [75] in order to *select c columns* of $\mathbf{M}$. That is

$$\mathbf{M} = \mathbf{U} \times \mathbf{\Sigma} \times \mathbf{V}^T \tag{7.3}$$

$\mathbf{U}$ Consists of n columns given as $\mathbf{U} = [\mathbf{u}_1|\mathbf{u}_2|\cdots|\mathbf{u}_n]$ where each $\{\mathbf{u}_i\}_{i=1}^n$ denotes left singular vector. The left singular vectors are the eigenvectors of $\mathbf{M}\mathbf{M}^T$.

$\mathbf{V}$ Consists of n columns given as $\mathbf{V} = [\mathbf{v}_1|\mathbf{v}_2|\cdots|\mathbf{v}_n]$ where each $\{\mathbf{v}_i\}_{i=1}^n$ denotes right singular vector. The right singular vectors are the eigenvectors of $\mathbf{M}^T\mathbf{M}$.

$\mathbf{\Sigma}$ is a diagonal matrix containing singular values $\{\sigma_i\}_{i=1}^n$. The diagonal elements of $\mathbf{\Sigma}$ are arranged in the descending order such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

When all the right singular vectors are used for reconstruction, the original input matrix $\mathbf{M}$ can be obtained without loss. That is:

$$\hat{\mathbf{M}} = \sum_{i=1}^{n} (\sigma_i \mathbf{u}_i)\mathbf{v}_i^T = \mathbf{M} \tag{7.4}$$

However, when first $k << n$ right singular vectors corresponding to the largest σ_i values are considered then the resulting matrix $\mathbf{M}_k$ is a rank-k approximation of $\mathbf{M}$ and is given below.

$$\mathbf{M}_k = \sum_{i=1}^{k} (\sigma_i \mathbf{u}_i)\mathbf{v}_i^T \approx \mathbf{M} \tag{7.5}$$

The right singular vectors are used in the selection of columns of $\mathbf{M}$ through the *normalized statistical leverage score* given in Equation 7.6 [74].

$$\pi_i = \frac{1}{k} \sum_{j=1}^{k} v_{ij}^2 \tag{7.6}$$

where π_i denote the normalized statistical leverage score of the i^{th} columns of $\mathbf{M}$. $\pi_i \geq 0$ and $\sum_{i=1}^{k} \pi_i = 1$. The quantity π_i is used to select columns of $\mathbf{M}$ to form matrix $\mathbf{C}$ as given in algorithm 2.

Row Selection Strategy The row selection is performed by providing $\mathbf{M}^T$ as input to the column selection strategy procedure given above. Input for Algorithm 2 is $\mathbf{M}^T$, number of rows to be selected r, c and ϵ. The columns of $\mathbf{M}^T$ will allow us to select rows in $\mathbf{M}$.

Algorithm 2 Construction of $\mathbf{C}$ by selecting columns of $\mathbf{M}$

Input: $\mathbf{M}$, k, c', ϵ.
Output: $\mathbf{C}$ such that $E[c] \leq c'$.

$\mathbf{C} = \{\}$ // start with empty set of columns
Compute top n right singular vectors $\{\mathbf{v}_i\}_{i=1}^{n}$ by employing SVD on $\mathbf{M}$
for j = 1 to n **do**
    Compute normalized statistical leverage score as given in Equation 7.6
    Let $p_j = min\{1, c'\pi_j\}$ for $c' = O\left(\frac{klogk}{\epsilon^2}\right)$
    Generate a random number ℓ between 0 and 1
    if $\ell \leq p_j$ **then**
        $\mathbf{C} \leftarrow \mathbf{C} \cup \mathbf{M}(:, j)$
    end if
end for
    return the set of columns $\mathbf{C}$ as a matrix.

Matrix U Computation After obtaining matrices $\mathbf{C}$ and $\mathbf{R}$ the matrix $\mathbf{U}$ is computed as:

$$\mathbf{U} = \mathbf{C}^{+} \times \mathbf{M} \times \mathbf{R}^{+} \tag{7.7}$$

where the operator $^{+}$ denotes the Moore-Penrose generalized inverse of matrix [76].

7.5.2 Tensor CUR-Decomposition

The matrix-CUR decomposition is extended to a multi-dimensional tensor setting. Given a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ this decomposition method constructs a tensor $\mathcal{C}$ and two matrices $\mathbf{R}$ and $\mathbf{U}$. Tensor $\mathcal{C}$ consists of c frontal slices from $\mathcal{A}$. Columns of the matrix $\mathbf{R}$ are constructed with r tubes of $\mathcal{A}$. The matrix $\mathbf{U}$ is constructed such that $\mathcal{A} \approx \mathcal{C} \times_3 \mathbf{U} \times_3 \mathbf{R}$ satisfies. Following four steps are performed to obtain factored tensors and matrices.

Mode-3 matricization As described above, the tensor $\mathcal{C}$ contains c frontal slices of the input $\mathcal{A}$. In order to obtain the frontal slices, $\mathcal{A}$ is subject to mode-3 matricization as explained in Section 7.3. The resulting matrix $\mathbf{A}_{(3)}$ is of size $I_3 \times (I_1 \times I_2)$ where each row is a frontal slice of $\mathcal{A}$.

Selection of frontal slices of $\mathcal{A}$ We note that frontal slices of $\mathcal{A}$ are the rows of $\mathbf{A}_{(3)}$. Therefore selection of frontal slices is nothing but selection of rows of $\mathbf{A}_{(3)}$. Row selection is performed using matrix-CUR decomposition. Following the procedure described in Section 7.5.1 by providing $\mathbf{A}_{(3)}^{T}, k, c', \epsilon$ as inputs to algorithm 2 yields c rows of $\mathbf{A}_{(3)}$. The obtained matrix is rearranged as a three dimensional tensor $\mathcal{C}$ of size $I_1 \times I_2 \times c$.

Selection of tubes of $\mathcal{A}$ Once again we note that columns of $\mathbf{A}_{(3)}$ contains tubes of $\mathcal{A}$. Therefore selection of tubes is nothing but selection of columns of $\mathbf{A}_{(3)}$ using matrix-CUR decomposition. Following the procedure described in Section 7.5.1 by providing $\mathbf{A}_{(3)}, k, c', \epsilon$ to Algorithm 2 yields r rows of $\mathbf{A}_{(3)}$ of size $I_3 \times r$.

Computing U Matrix $\mathbf{U}$ is obtained given as $\mathbf{U} = \mathbf{R}^{+} \times \mathbf{A}_{(3)} \times \mathbf{C}_{(3)}^{+}$

Inst. no.	administration	affiliation	course	employment	finance	medical	students
1	20	32	50	0	30	10	100
2	16	12	45	61	100	35	95
3	99	84	15	36	69	24	10
4	20	8	95	20	35	72	10
5	98	32	53	96	9	44	55
6	65	75	45	84	79	23	40
7	39	59	80	46	95	90	16
8	18	42	71	92	5	96	89

Table 7.3: IQC_{year} matrix obtained after tensor-CUR decomposition of $\mathcal{A}_{RQ1}$

Inst. no.	administration	affiliation	course	employment	finance	medical	students
1	1	2	3	1	2	1	4
2	1	1	2	3	4	2	4
3	4	4	1	2	3	1	1
4	1	1	4	1	2	3	1
5	4	2	3	4	1	2	3
6	3	4	2	4	4	1	2
7	2	3	4	2	4	4	1
8	1	2	3	4	1	4	4

Table 7.4: Transformed IQC_{year} matrix obtained after replacing reply-rates with reply classes

7.6 Experimental Results

In order to answer the two research questions, we make use of $\mathcal{A}_{RQ1}$ and $\mathcal{A}_{RQ2}$ RTI tensors as the input to tensor-CUR decomposition. Tensor-CUR decomposition is implemented in MATLAB®. The features obtained after the decomposition are used for quantifying 'transparency' of institutions using GRM, which is implemented in R with the 'ltm' package [77].

To answer **RQ1** the RTI tensor $\mathcal{A}_{RQ1}$ is subject to tensor-CUR decomposition to obtain three factors $\mathcal{C}_{RQ1}$, $\mathbf{R}_{RQ1}$ and $\mathbf{U}_{RQ1}$. $\mathcal{C}_{RQ1}$ is a tensor having dimension $8 * 7 * c$ that consists of a c number of frontal slices (institute×query-category matrix) of $\mathcal{A}_{RQ1}$, $\mathbf{R}_{RQ1}$ consists of vectors of dimension $6 * r$. From $\mathcal{C}_{RQ1}$, we take that slice (institute×query-category matrix) which has highest the variance in reply rates, and call it the IQC_{year} matrix, shown in Table 7.3. The time interval corresponding to this slice reveals the answer to **RQ1**. This is the time-interval that has the highest variance in the RTI output-oriented indicators.

The IQC_{year} matrix is transformed where each of the reply rates is replaced by the category value (1 to 4). The transformed matrix is given in Table 7.4. This is used as input to GRM following algorithm 1 to quantify $\theta_{IQC_{year}}$. The parameters obtained through GRM

Inst. no.	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
1	20	99	100	90	92	53
2	16	25	5	10	60	27
3	99	30	94	72	42	14
4	20	78	55	45	5	86
5	98	72	46	24	99	95
6	65	35	10	81	90	30
7	39	16	81	72	38	97
8	18	90	79	40	100	66

Table 7.5: $IT_{query-category}$ matrix obtained after tensor-CUR decomposition of $\mathcal{A}_{RQ2}$

Inst. no.	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
1	1	4	4	4	4	3
2	1	2	1	1	3	2
3	4	2	4	3	2	1
4	1	4	3	2	1	4
5	4	3	2	1	4	4
6	3	2	1	4	4	2
7	2	1	4	3	2	4
8	1	4	4	2	4	3

Table 7.6: Transformed $IT_{query-category}$ matrix obtained after replacing reply-rates with reply classes

Inst. no.	$\theta_{avg-year}$	Year 2010 $\theta_{IQ_{year}}$	Administration $\theta_{IT_{query-category}}$
1	2.225	-0.770	-1.120
2	2.622	-1.147	-0.076
3	2.552	-0.416	0.479
4	2.551	-0.574	0.93
5	2.550	0.079	0.440
6	2.551	0.849	-0.906
7	-0.237	0.013	0.366
8	2.553	1.560	-0.515

Table 7.7: Transparency of each institution

are depicted in Table 7.7.

To understand what temporal fluctuations are captured, GRM has been applied on the whole RTI data when no time dimension is introduced in the RTI query data. For this, the reply-rates across the time-dimension in $\mathcal{A}_{RQ1}$ are combined and averaged and we get a static matrix from the RTI tensor, containing institutions in the rows and query-categories in the columns. We call it the $IQC_{avg-year}$ matrix, shown in Table 7.8. Transparency estimated via GRM using the $IQC_{avg-year}$ is denoted as $\theta_{avg-year}$, presented in Table 7.7. We observe the following:

1. There are significant changes in the transparency values observed for the $IQC_{avg-year}$ and the IQC_{year} matrix in the year 2010 which happens to have maximum statistical leverage score.

2. We attribute the change in the transparency values to the changes in the inputs in the year 2010. If there is no change in the inputs, we expect no change in the reply rates trend.

Inst. no.	administration	affiliation	course	employment	finance	medical	students
1	75.67	36.67	23.67	18.67	47.17	23.17	42.83
2	23.83	40.17	34.67	44	56.5	47.67	70.17
3	58.5	36.33	39.17	44.33	41.33	42	18.17
4	48.17	23.83	44.17	45.5	34.67	42.17	45.83
5	72.33	45	35.67	81.67	39	56.17	53
6	51.83	57.5	55.17	53.17	41.17	63.17	60.17
7	57.17	45.5	60.83	56	60.67	70.33	62.33
8	65.5	48.83	53.67	64.17	20.67	53.33	75.17

Table 7.8: $IQC_{avg-year}$ matrix

3. Given the changes in the transparency values $\theta_{avg-year}$ and $\theta_{IQC_{year}}$, we observe an important hidden pattern which is presented in Figure 7.4. In this figure, the x-axis denotes the transparency values. The top-most line corresponds to $\theta_{avg-year}$ values presented in Table 7.7. The second line from top corresponds to $\theta_{IQC_{year}}$ values. The blue dots correspond to central institutions and red dots correspond to state institutions. In the first line, there is no segregation of these transparency values. However, in the second line, we see a clear separation trend in the transparency values where all central institutions have high transparency values and are towards the right side of the x-axis. This is a hidden pattern that is invisible in the static RTI data representation. A time-interval that has segregated transparencies between state and central institutions indicates that there has been a notable influence of input-oriented factor affecting the state institutions negatively, leading to their reduced transparencies for that particular year. Identification of these hidden trends is very useful for policymakers to associate such fluctuations in the outputs to the relevant input factor that caused them. The significance of this is that finding the cause of such patterns has great policy implications and provides an opportunity of improvement to the system for the law-makers to correct the input factors, creating a scope for effective amendment.

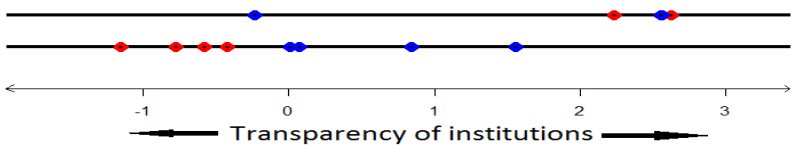


Figure 7.4: Transparency with and without tensor-CUR decomposition

To answer **RQ2**, the RTI tensor $\mathcal{A}_{RQ2}$ is subject to tensor-CUR decomposition to obtain three factors $\mathcal{C}_{RQ2}$, $\mathbf{R}_{RQ2}$ and $\mathbf{U}_{RQ2}$. From $\mathcal{C}_{RQ2}$, we take that slice which has highest the statistical leverage score, called the $IT_{query-category}$ matrix. The $IT_{query-category}$ matrix is given in Table 7.5.

The query-category corresponding to this slice is the *administration* query-category. This means that when RTI queries correspond to the *administration* category, there is a large variation in reply rates across all the years and all the institutions. In other words, *administration* contributes to most of the variations in the reply rates across all times. In the previous Chapter, we had quantified this query-category which saw the most fluctuations among the institutions. The interpretation of the outcome of the present experiment is that for this query-category, fluctuations across *all times* in RTI query-reply statistics are high. Once again identification of this latent pattern brings value to the policymakers in understanding the underlying cause for these spikes and finding improvements through relevant amendments.

With the obtained results, our RTI time model and tensor-CUR decomposition has demonstrated its capability to capture hidden patterns that are relevant in the social context, thereby providing important information that acts as guidance to policymakers for proposing amendments.

7.7 Conclusion

Understanding temporal dynamics of the RTI query-reply data is important to obtain a deeper understanding of the government system, thereby uncovering further scope for potential amendments to the existing laws. This Chapter deals with observing the temporal fluctuations of the RTI data by modeling both the inputs and outputs of the RTI system. Since outputs are a function of the input parameters and time, capturing changes in the output properties over a time-period provides a way to observe the influence of the input-oriented factors.

In this Chapter we (i) propose two RTI tensors capturing 'time' as the third dimension/'query-category' as the third dimension (ii) performed tensor-CUR decomposition on these two tensors to obtain an IQC_{year} containing maximum variation in reply time, $IT_{query-category}$ containing maximum variation in the query-category reply time. (iii) These two matrices are subject to GRM to obtain transparency of institutions. The most fluctuating time-period and the most fluctuating query-category across all times are identified.

The year 2010 is identified as the most fluctuating year; in this year, all state institutions have very low transparency compared to the central institutions. This information is potentially significant in identifying the cause of this pattern and compel policy-makers to propose relevant amendments to avoid such spikes in the government system. In addition, the query-category 'administration' is having most variations in reply rates. This is in agreement with the previous two Chapters.

8

Conclusions and Future Work

Wagstaff [78] has presented six impacting challenges that can be considered as examples of 'machine learning that matters'. The first challenge among them is stated as *A law passed or legal decision made that relies on the result of an ML analysis*. We take this as the motivation for the presented work. In addition, the *transparency audit* by CIC [7] presented a pragmatic view on the difficulty in computing the quantity transparency as it is inherently multidimensional. When accurate information is disseminated regularly when information is readily accessible in a timely fashion and when information is provided by every department, section and body of government institution it leads to achieving transparency. The RTI Act is viewed as a tool for the citizens to seek information which is a *fundamental right*. In this thesis, we study the RTI Act by collecting the RTI applications received by government educational institutions across India. A computational model for quantifying transparency was carried out. Interpreting the internal parameters of the model yielded potential amendments to the existing RTI Act whose validation was performed in a limited way.

8.1 Outcome of the Thesis

Chapter 3 This Chapter discussed the data collection efforts, the obtained data and their characteristics. Collection of data from government institutions needs persistent efforts and patience. It also throws a challenge in terms of processing the obtained data in a form that is readily consumable by the learning algorithms. It took 18 months and several postal communications with the government educational institutions to collect this dataset. Despite filing RTI applications from 352 institutions across the country, we could get *replies* from 50 institutions. The replies obtained in some instances were well over the 30 day period that the RTI Act mandates. Though *replying to RTI application*, in theory, is mandatory, we witnessed a huge gap in its execution. All the analysis carried out in this work are limited with our ability in processing and obtaining *data that is readily consumable* by the learning models that were identified.

Chapter 4 This work, being interdisciplinary, the literature from multiple domains were studied and presented. These include (i) the subtopics of humanities and social sciences

(ii) Machine Learning (iii) Psychometric analysis and (iv) query analysis (web search engine queries or survey questions or test questions).

Chapter 5 For the first time the distributional characteristics of RTI applications data was carried out and presented. We visualize the RTI query and reply process as a *paid* and *offline* information retrieval system. In this system, (i) querying is associated with cost (time as well as money) and (ii) retrieving the documents is also associated with cost. We compared the estimated distributional parameters of this query log with that of the web query-log (which is a free and online system) system [11]. We observed similarities in terms of the fitted distributions and dissimilarities in terms of the obtained parameter ranges (sharpness of the curves). The parameters estimated using RTI query log data are outside the typical range of the estimated parameters of the web query logs.

A potential amendment was identified using the obtained results of query length distribution analysis namely limiting the number of words in the RTI application to 500. This was proposed by the government of India but later not implemented due to resistance from various sections of the society. Through query-reply-time, transparency is quantified through the estimation of the probability of replying within 30 days in a given query-category. The query reply time analysis stresses the importance of query-category in achieving high probability of reply.

Chapter 6 Similarities between test questions and the RTI applications were presented. Similarities between the ability of test takers and transparency of institutes were presented. The quantity, probability of getting a reply for a specified institute and specified query category within 30 days is modeled in terms of an institute's latent parameter, namely, *transparency*, and query-categories' latent parameters namely *effectiveness of the implementation of the RTI Act* and *discriminative power of the query-category*. These parameters were estimated using GRM on the IQC matrix data. Individual institute's transparency values are obtained and validated using the variance of estimated parameters. Through the interpretation of the parameters, we propose a potential amendment that *Reply duration should be based on the difficulty of the query-category instead of 30-days irrespective of the query-category*".

Chapter 7 The transparency model proposed in Chapter 6 take into account *output-oriented indicators alone*. In addition, the data model does not consider the time of receiving the RTI application by the PIO. This limitation was addressed by including an *input-oriented indicator, namely, time of receiving the RTI application*. A three-dimensional RTI tensor representation was proposed to take into account the time of filing the RTI application. Year in which maximum variation in reply rates observed was identified and the query-category in which maximum variation in the reply rates observed was identified. The obtained results are consistent with respect to the query-category observation provided in Chapter 5 and 6. That is, Administration query-category captures the maximum variation in the reply-rates.

8.1.1 Discussion

The analysis presented in Chapters 5, 6 and 7 help both Government and citizens.

Benefits of the presented analysis for government Quantification of transparency is attempted in this thesis has multiple uses (1) Government make use of the transparency values to rank institutes (2) Government institute compete for improvisation of transparency ranking (3) Use obtained transparency values for bringing foreign investments (4) Help understand the overall effectiveness of implementation of the RTI act

Benefits of this analysis for citizens So far, there is no specific information about individual institute's transparency values to citizens. Publishing and publicizing these values will bring awareness among citizens about the level of government engagement with citizens. In addition, the obtained latent parameters help citizens understand what can be expected out of a query from specific section within government establishment.

8.2 Limitations

Following are the limitations of the work carried out

Chapter 3 Though we have obtained 34,976 RTI applications from 56 institutions, all the received applications could not be processed due to (i) diverse languages in which applications were written (ii) Categorizing each sub-query within an RTI application is an intensive effort (iii) Computing the time of reply for each RTI application is once again a manual effort. (iv) Digitization of obtained applications was another bottleneck. Partially automation was put in place in terms of using OCR software. However, the use of such tools is limited to printed RTI applications written in English. The experimental set up was constrained by the size of the data.

Chapter 6 The number of query categories (10) is very less due to the low volume of the processed RTI applications. Typically a large number of query categories help in obtaining a better GRM. In addition, the number of institutes is also constrained due to the processed data. Despite these limitations, the variance of the obtained parameters has low values. However, increasing the number of query-categories and number institutes will strengthen the GRM.

Chapter 7 The input-oriented factors such as PIOs, their appointment dates and duration dates potentially helped refine the RTI tensor model. In the absence of these, the time dimension is incorporated in modeling the inputs of the RTI applications. On the technical front, the decomposed tensor is of very small dimensions ($8 \times 7 \times 6$). The built tensor's dimensions should have been larger for effectively understanding the temporal fluctuations.

8.3 Future Directions

The work done in this thesis opens up a number of directions for future research. During the research, there have also been a few limitations in the work done. These are additional directions for future work.

Chapter 3 discusses the RTI data collection process. RTI data was collected from educational institutions across India. On this front, we attempted to collect data from 352 institutions, but have been successful in collecting data from 50 institutions. As future work, data collection to be performed comprehensively, collecting the entire query log (RTI applications, date of reply and rejection grounds) from more institutions including primary and high schools. Data collection shall be extended to all the 67 ministries and independent departments within the government. Analysis and quantification of the RTI properties from diverse query-log shall provide diverse patterns that can be interpreted for amendments.

The public authorities reply to the received question. However, it is not guaranteed that the given reply satisfied the information need of the citizen. Citizen's feedback was never taken by the institutes after serving the reply. This information is not assessed by CIC in their analysis of the annual reports. It is therefore of importance to compute and quantify the degree of relevance of the answer to the given RTI query.

RTI query-categorization in this thesis has been manual. As the objective of categorization was to segregate the queries based on government institutional structure, using automatic topic modeling algorithms did not provide us with the required categories. On this front, selective topic modeling techniques leading to topics based on sections of institutions can be explored.

Chapter 5, 6 and 7 saw data models that incorporated various definitions of RTI properties. Analysis of the data models and the corresponding results are interpreted based on the specific properties represented in the data models. There are additional definitions of transparency and effectiveness of implementation that can be modeled. This thesis has considered modeling only a subset of the provided definitions. In addition, other performance measures of the RTI Act like 'rejection statistics, 'delay in appeals etc. have not been analyzed. These form a good perspective to propose new data models, and quantification of these properties will provide more information as to the experience of citizens with regards to the RTI Act, thus enabling further scopes for amendment.

The tools used in this thesis are of immense importance in the context of machine learning. In particular classification methods. Item response theory (IRT) is a popular model for analyzing students test questions. Recently, IRT's utility in machine learning domain has been demonstrated. In particular Fernando Martnez-Plumed et al[79] employed IRT model for analyzing classification algorithms where data point correspond to items and respondents correspond to classification algorithms. The objective was to understand what the IRT parameters such as discrimination, difficulty and guessing mean for classification instances.

IRT and GRM model's utility in explainable AI, explainable ML is a potential area. In particular, these methods have to ability to explain hardness of data point through discrimination parameter. In addition, ability/transparency parameter is used for characterizing classification instance. From this point of view, IRT/GRM are of importance in understanding fairness of classifiers [80]. A further investigation is needed in this direction.

Appendix A

Government Educational Institutes' Details

List of Class 10 and Class 10 + 2 boards across all states of India to whom our RTI application was filed.

Sl. No	URL	**Board Name**	**Location**	**Type**
1	http://bseape.org/	Board of Secondary Education Andhra Pradesh	Andhra Pradesh	State
2	https://bie.ap.gov.in/	Board of Intermediate Education Andhra Pradesh	Andhra Pradesh	State
3	https://sebaonline.org/	Secondary Education Board of Assam	Assam	State
4	http://www.ahsec.nic.in/	Assam Higher Secondary Education Council	Assam	State
5	http://biharboardonline.bihar.gov.in/	Bihar School Examination Board	Bihar	State
6	http://www.bbose.org/	Bihar Board of Open Schooling	Bihar	State
7	Not Available	Bihar Intermediate Education Council	Bihar	State
8	https://cgbse.nic.in/	Chhattisgarh Board of Secondary Education	Chhattisgarh	State
9	http://www.cgsos.co.in/	Chattisgarh State Open School	Chhattisgarh	State
10	https://gbshse.gov.in/	Goa Board of Secondary and Higher Secondary Education	Goa	State
11	http://www.gseb.org/	Gujarat Secondary and Higher Secondary Education Board	Gujarat	State
12	https://bseh.org.in/	Haryana Board of School Education	Haryana	State
13	https://www.hpbose.org/	H.P. Board of School Education	Himachal	State

14	https://jac.jharkhand.gov.in/jac/	Jharkhand Academic Council	Jharkhand	State
15	http://kseeb.kar.nic.in/	Karnataka Secondary Education Examination Board	Karnataka	State
16	http://pue.kar.nic.in/	Government of Karnataka Dept. of Pre-University Education	Karnataka	State
17	http://www.dhsekerala.gov.in/	Kerala Board of Higher Secondary Education	Kerala	State
18	https://kbpe.org/	Kerala Board of Public Examinations	Kerala	State
19	http://www.mpsos.nic.in/	M.P. State Open School Board of Secondary Education Campus	Madhya Pradesh	State
20	http://mpbse.nic.in/	Madhya Pradesh Board of Secondary Education	Madhya Pradesh	State
21	http://www.mahahsscboard.in/	Maharashtra State Board of Secondary and Higher Secondary Education	Maharashtra	State
22	https://bsem.nic.in/	Manipur Board of Secondary Education	Manipur	State
23	https://cohsem.nic.in/	Council of Higher Secondary Education Manipur	Manipur	State
24	http://www.mbse.edu.in/	Mizoram Board of School Education	Mizoram	State
25	http://www.nbsenagaland.com/	Nagaland Board of School Education	Nagaland	State
26	http://chseodisha.nic.in/	Council of Higher Secondary Education, Odisha	Odisha	State
27	http://www.bseodisha.nic.in/	Board of Secondary Education Odisha	Orissa	State
28	https://www.pseb.ac.in/	Punjab School Education Board	Punjab	State
29	http://rajeduboard.rajasthan.gov.in/	Board of Secondary Education Rajasthan	Rajasthan	State
30	http://sikkim-hrdd.gov.in/directorate_school_education_administration.htm	Directorate of School Education and Administration	Sikkim	State

31	https://www.tn.gov.in/department/28	School Education Department	Tamil nadu	State
32	http://bie.tg.nic.in/	Telangana Board of Intermediate Education	Telangana	State
33	https://www.bse.telangana.gov.in/	Telangana Board of Secondary Education	Telangana	State
34	https://www.telanganaopenschool.org/	Telangana Open School Society	Telangana	State
35	http://tbse.in/new/welcome.html	Tripura Board of Secondary Education	Tripura	State
36	http://bseup.org/	Board of Secondary Education Kant Shahjahanpur Uttar Pradesh	Uttar Pradesh	State
37	https://upmsp.edu.in/	U.P. Board of High School & Intermediate Education	Uttar Pradesh	State
38	https://ubse.uk.gov.in/	Uttranchal Shiksha Evm Pariksha Prishad	Uttrakhand	State
39	https://wbchse.nic.in/html/index.html	West Bengal Council of Higher Secondary Education	West Bengal	State
40	http://www.wbbpe.org/	West Bengal Board of Primary Education	West Bengal	State
41	http://wbbse.org/	West Bengal Board of Secondary Education	West Bengal	State
42	http://wbcros.in/	Ravindra Mukta Vidyalaya	West Bengal	State
43	postal address provided by main CBSE PIO	Assistant Secretary (Admn.II)/PIO	CBSE, Preet Vihar, Delhi	Central
44	postal address provided by main CBSE PIO	Assistant Secretary (Co-ord)/PIO	CBSE, Preet Vihar, Delhi	Central
45	postal address provided by main CBSE PIO	Joint Director (Academic & Vocational)/PIO	CBSE, Delhi	Central
46	postal address provided by main CBSE PIO	Assistant Secretary (CTET)/PIO	CBSE, Delhi	Central
47	postal address provided by main CBSE PIO	Assistant Secretary (AIPMT)/PIO	CBSE, Preet Vihar, Delhi	Central

48	postal address provided by main CBSE PIO	Desk Officer (Affln)/PIO	CBSE, Preet Vihar, Delhi	Central
49	postal address provided by main CBSE PIO	Deputy Secretary (JEE)/PIO	CBSE,Noida, UP	Central
50	postal address provided by main CBSE PIO	Assistant Public Relation Officer/PIO	CBSE, Delhi	Central
51	postal address provided by main CBSE PIO	Assistant Secretary/PIO	CBSE, Haryana	Central
52	postal address provided by main CBSE PIO	Assistant Secretary/PIO	CBSE, Alla-habad	Central
53	postal address provided by main CBSE PIO	Deputy Director (Vocational)/PIO	CBSE, Delhi	Central
54	postal address provided by main CBSE PIO	Assistant Secretary/PIO	CBSE, Ajmer	Central
55	postal address provided by main CBSE PIO	Assistant Secretary/PIO	CBSE, Pan-jabari, Guwa-hati	Central
56	postal address provided by main CBSE PIO	Deputy Secretary/PIO	CBSE, Chennai	Central
57	postal address provided by main CBSE PIO	Assiatant Secretary/PIO	CBSE, Patna	Central
58	postal address provided by main CBSE PIO	Deputy Secretary/PIO	CBSE, Bhubaneswar	Central
59	postal address provided by main CBSE PIO	Assistant Secretary/PIO	CBSE, Thiru-vanan-thapu-ram	Central
60	postal address provided by main CBSE PIO	Assistant Secretary/PIO	CBSE, Dehradun	Central
61	https://www.bhsedelhiboard.net/	Board of Higher Secondary Education, Delhi	New Delhi	Central

62	https://www.nbyeindia.co.in/	Board of Youth Education India	West Bengal	Central
63	https://www.nios.ac.in/	National Institute of Open Schooling	New Delhi	Central
64	http://www.sanskrit.nic.in/	Rashtriya Sanskrit Sansthan, Delhi	New Delhi	Central

Table A.1: List of class 10, class 10 + 2 boards for data collection

List of state, central and deemed universities to whom our RTI application was filed.

Sl. No	URL	University Name	Location	Type
65	https://www.uohyd.ac.in/	University of Hyderabad	Andhra Pradesh	State
66	http://manuu.edu.in/ur	Maulana Azad National Urdu University	Andhra Pradesh	State
67	https://www.efluniversity.ac.in/	THE ENGLISH AND FOREIGN LANGUAGES UNIVERSITY	Andhra Pradesh	State
68	https://angrau.ac.in/angrau/	Acharya N G Ranga Agricultural University	Andhra Pradesh	State
69	https://aknu.edu.in/	Adikavi Nannaya University	Andhra Pradesh	State
70	https://www.andhrauniversity.edu.in/	Andhra University	Andhra Pradesh	State
71	http://www.nagarjunauniversity.ac.in/	Acharya Nagarjuna University	Andhra Pradesh	State
72	http://www.dravidianuniversity.ac.in/	Dravidian University	Andhra Pradesh	State
73	https://braou.ac.in/	Dr B R Ambedkar Open University	Andhra Pradesh	State
74	http://ntruhs.ap.nic.in/index.html	Dr. NTR University of Health Sciences	Andhra Pradesh	State
75	https://jntuh.ac.in/	Jawaharlal Nehru Technological University	Andhra Pradesh	State
76	https://www.jntuk.edu.in/	Jawaharlal Nehru Technological University Kakinada	Andhra Pradesh	State
77	https://www.kakatiya.ac.in/	Kakatiya University	Andhra Pradesh	State
78	http://www.krishnauniversity.ac.in/	Krishna University	Andhra Pradesh	State
79	https://mguniversity.ac.in/home.php	Mahatma Gandhi University	Andhra Pradesh	State
80	http://nalsar.ac.in/	National Academy of Legal Studies & Research University	Andhra Pradesh	State
81	https://www.osmania.ac.in/	Osmania University	Andhra Pradesh	State
82	http://palamuruuniversity.ac.in/	Palamuru University	Andhra Pradesh	State

83	http://teluguuniversity.ac.in/	Potti Sreeramulu Telugu University	Andhra Pradesh	State
84	https://www.ruk.ac.in/index.php	Rayalaseema University	Andhra Pradesh	State
85	http://www.satavahana.ac.in/	Satavahana University	Andhra Pradesh	State
86	http://www.skuniversity.ac.in/	Sri Krishnadevaraya University	Andhra Pradesh	State
87	http://www.spmvv.ac.in/	Sri Padmavati Mahila Vishwavidyalayam	Andhra Pradesh	State
88	https://svvu.edu.in/	Sri Venkateswara Veterinary University	Andhra Pradesh	State
89	https://www.svuniversity.edu.in/	Sri Venkateswara University	Andhra Pradesh	State
90	http://www.svvedicuniversity.ac.in/	Sri Venkateswara Vedic University	Andhra Pradesh	State
91	http://www.telanganauniversity.ac.in/	Telangana University	Andhra Pradesh	State
92	http://www.simhapuriuniv.ac.in/	Vikrama Simhapuri University	Andhra Pradesh	State
93	http://www.yogivemanauniversity.ac.in/ysr/	Yogi Vemana University	Andhra Pradesh	State
94	https://www.rgukt.ac.in/	Rajiv Gandhi University of Knowledge Technologies	Andhra Pradesh	State
95	https://www.jnafau.ac.in/	Jawaharlal Nehru Architecture & Fine Arts University	Andhra Pradesh	State
96	http://www.drysrhu.edu.in/	A.P.Horticultural University	Andhra Pradesh	State
97	https://www.gitam.edu/	Gandhi Institute of Technology and Management	Andhra Pradesh	State
98	https://www.iiit.ac.in/	International Institute of Information Technology, Hyderabad	Andhra Pradesh	State
99	http://rsvidyapeetha.ac.in/	Rashtriya Sanskrit Vidyapeeth	Andhra Pradesh	State
100	http://sssihl.edu.in/	Sri Sathya Sai Institute of Higher Learning	Andhra Pradesh	State
101	https://www.ifheindia.org/	ICFAI Foundation for Higher Education	Andhra Pradesh	State
102	http://www.vignan.ac.in/	Vignan University	Andhra Pradesh	State

103	https://www.kluniversity.in/	Koneru Lakshmaiah Education Foundation (K L University)	Andhra Pradesh	State
104	https://www.rgu.ac.in/	Rajiv Gandhi University	Arunachal Pradesh	State
105	https://nerist.ac.in/	North Eastern Regional Institute of Science & Technology	Arunachal Pradesh	State
106	http://igtamsu.ac.in/	Indira Gandhi Technological and Medical Sciences University	Arunachal Pradesh	State
107	https://www.arunachaluniversity.ac.in/	Arunachal University of Studies	Arunachal Pradesh	Central
108	http://vou.ac.in/	Venkateshwara Open University	Arunachal Pradesh	Central
109	https://www.apexuniversity.edu.in/	Apex Professional University	Arunachal Pradesh	Central
110	https://www.himalayanuniversity.com/	Himalayan University	Arunachal Pradesh	Central
111	http://www.aus.ac.in/	Assam University	Assam	Central
112	http://www.tezu.ernet.in/	Tezpur University	Assam	Central
113	http://nluassam.ac.in/	National Law University and Judicial Academy	Assam	Central
114	http://astu.ac.in/	Assam Science & Technology University	Assam	Central
115	http://www.aau.ac.in/	Assam Agricultural University	Assam	Central
116	https://www.dibru.ac.in/	Dibrugarh University	Assam	Central
117	https://www.gauhati.ac.in/	Gauhati University	Assam	Central
118	http://bodolanduniversity.ac.in/	Bodoland University	Assam	Central
119	http://www.kkhsou.in/web_new/index.php	Krishna Kanta Handique State Open University	Assam	Central
120	https://cottonuniversity.ac.in/	Cotton College State University	Assam	Central
121	http://ssuhs.in/	Srimanta Sankaradeva University of Health Sciences	Assam	Central
122	http://bnmu.ac.in/	Bhupendra Narayan Mandal University	Bihar	Central
123	https://www.brabu.net/	Babasaheb Bhimrao Ambedkar Bihar University	Bihar	Central

124	http://akubihar.ac.in/	Aryabhatta Knowledge University	Bihar	Central
125	http://cnlu.ac.in/	Chanakya National Law University	Bihar	Central
126	http://jpv.bih.nic.in/	Jai Prakash University	Bihar	Central
127	http://www.ksdsu.edu.in/home.htm	Kameshwar Singh Darbhanga Sanskrit University	Bihar	Central
128	http://www.lnmuuniversity.in/login	Lalit Narayan Mithila University	Bihar	Central
129	https://magadhuniversity.ac.in/	Magadh University	Bihar	state
130	http://mmhapu.bih.nic.in/results.htm	Maulana Mazharul Haque Arabic and Persian University	Bihar	state
131	http://www.nou.ac.in/	Nalanda Open University	Bihar	state
132	http://www.patnauniversity.ac.in/	Patna University	Bihar	state
133	http://tmbuniv.ac.in/	Tilka Manjhi Bhagalpur University	Bihar	state
134	https://vksu.ac.in/	Veer Kunwar Singh University	Bihar	state
135	http://www.bausabour.ac.in/	Bihar Agricultural University	Bihar	state
136	https://www.rpcau.ac.in/	Rajendra Agricultural University	Bihar	Central
137	http://www.cusb.ac.in/	Central University of South Bihar	Bihar	Central
138	https://www.amu.ac.in/	Aligarh Muslim University	Bihar	Central
139	https://nalandauniv.edu.in/	Nalanda International University	Bihar	Central
140	https://www.nnm.ac.in/	Nava Nalanda Mahavihara	Bihar	Deemed
141	https://www.biharyoga.net/bihar-yoga-bharati.php	Bihar Yoga Bharati	Bihar	Deemed
142	https://bvvjdp.ac.in/	Bastar Vishwavidyalaya	Chhattisgarh	state
143	https://csvtu.ac.in/ew/	Chhattisgarh Swami Vivekanand Technical University	Chhattisgarh	state
144	https://www.hnlu.ac.in/	Hidayatullah National Law University	Chhattisgarh	state

145	http://www.igau.edu.in/	Indira Gandhi Krishi Vishwavidyalaya	Chhattisgarh	state
146	http://www.iksv.ac.in/	Indira Kala Sangeet Vishwavidyalaya	Chhattisgarh	state
147	http://www.ktujm.ac.in/	Kushabhau Thakre Patrakarita Avam Jansanchar Vishwavidyalaya	Chhattisgarh	state
148	http://www.prsu.ac.in/	Pt. Ravishankar Shukla University	Chhattisgarh	state
149	http://pssou.ac.in/	Pt. Sundarlal Sharma (Open) University	Chhattisgarh	state
150	http://www.sargujauniversity.in/	Sarguja University	Chhattisgarh	state
151	http://ggu.ac.in/	Guru Ghasidas Vishwavidyalaya	Chhattisgarh	Central
152	https://www.cvru.ac.in/	C.V.Raman University	Chhattisgarh	Deemed
153	https://www.unigoa.ac.in/	Goa University	Goa	State
154	https://www.gujaratuniversity.ac.in/	Gujarat University	Gujarat	state
155	http://www.spuvvn.edu/	Sardar Patel University	Gujarat	state
156	http://www.vnsgu.ac.in/	Veer Narmad South Gujarat University	Gujarat	state
157	https://msubaroda.ac.in/	The Maharaja Sayajirao University	Gujarat	state
158	http://www.saurashtrauniversity.edu/	Saurashtra University	Gujarat	state
159	https://www.mkbhavuni.edu.in/mkbhavuniweb/	Bhavnagar University	Gujarat	state
160	https://www.ngu.ac.in/	Hemchandracharya North Gujarat University	Gujarat	state
161	http://kskvku.digitaluniversity.ac/	Krantiguru Shyamji Krishna Verma Kachchh University	Gujarat	state
162	http://www.baou.edu.in/	Dr. Babasaheb Ambedkar Open University	Gujarat	state
163	https://sssu.ac.in/	Shree Somnath Sanskrit University	Gujarat	state
164	http://www.cugujarat.ac.in/	Children's University Gujarat	Gujarat	state
165	https://www.gfsu.edu.in/	Gujarat Forensic Sciences University	Gujarat	state
166	http://rsu.ac.in/	Raksha Shakti University	Gujarat	state

167	http://ku-guj.org/Kamdhenu-University	Kamdhenu University	Gujarat	state
168	https://www.iite.ac.in/	The Indian Institute of Teacher Education	Gujarat	state
169	https://www.gtu.ac.in/	Gujarat Technological University	Gujarat	state
170	http://ayurveduniversity.edu.in/	Gujarat Ayurved University	Gujarat	state
171	http://www.aau.in/	Anand Agricultural University	Gujarat	state
172	https://nau.in/index	Navsari Agricultural University	Gujarat	state
173	http://www.sdau.edu.in/	Sardarkrushinagar Dantiwada Agricultural University	Gujarat	state
174	http://www.jau.in/	Junagadh Agricultural University	Gujarat	state
175	http://www.gujaratvidyapith.org/	Gujarat Vidyapith	Gujarat	Deemed
176	http://www.cug.ac.in/	Central University of Gujarat	Gujarat	Central
177	https://sumandeepvidyapeethdu.edu.in/	Sumandeep Vidyapith	Gujarat	Central
178	http://www.bpswomenuniversity.ac.in/	Bhagat Phool Singh Mahila Vishwavidyalaya	Haryana	state
179	http://cdlu.ac.in/	Chaudhary Devi Lal University	Haryana	state
180	https://www.hau.ac.in/	Chaudhary Charan Singh Haryana Agricultural University	Haryana	state
181	http://www.dcrustm.ac.in/	Deen Bandhu Chhotu Ram University of Science & Technology	Haryana	state
182	http://www.gjust.ac.in/	Guru Jambeshwar University of Science and Technology	Haryana	state
183	https://kuk.ac.in/	Kurukshetra University	Haryana	state
184	http://mdu.ac.in/	Maharshi Dayanand University	Haryana	state
185	http://www.uhsr.ac.in/	Pt. Bhagwat Dayal Sharma University of Health Sciences	Haryana	state
186	http://www.cuh.ac.in/	Central University of Haryana	Haryana	Central
187	https://www.mmumullana.org/	Maharishi Markandeshwar University	Haryana	Deemed
188	http://www.nbrc.ac.in/newweb/	National Brain Research Centre	Haryana	Deemed
189	http://www.ndri.res.in/	National Dairy Research Institute	Haryana	Deemed

190	https://manavrachna.edu.in/	Manav Rachna International University	Haryana	Deemed
191	https://www.lingayasuniversity.edu.in/	Lingaya's University	Haryana	Deemed
192	http://www.yspuniversity.ac.in/	Dr. Y.S.Parmar University of Horticulture & Forestry	Himachal Pradesh	state
193	http://www.hpuniv.ac.in/	Himachal Pradesh University	Himachal Pradesh	state
194	http://www.hillagric.ac.in/	Himachal Pradesh Agriculture University	Himachal Pradesh	state
195	https://www.himtu.ac.in/	Himachal Pradesh Technical University	Himachal Pradesh	state
196	http://cuhimachal.ac.in/	Central University of Himachal Pradesh	Himachal Pradesh	Central
197	https://www.bauranchi.org/	Birsa Agricultural University	Jharkhand	State
198	http://www.ranchiuniversity.ac.in/	Ranchi University	Jharkhand	State
199	http://skmu.ac.in/	Sido Kanhu Murmu University	Jharkhand	State
200	http://www.kolhanuniversity.ac.in/	Kolhan University	Jharkhand	State
201	https://npu.ac.in/	Nilamber-Pitamber University	Jharkhand	State
202	http://vbu.ac.in/	Vinoba Bhave University	Jharkhand	State
203	http://bangaloreuniversity.ac.in/	Bangalore University	Karnataka	state
204	http://davangereuniversity.ac.in/	Davangere University	Karnataka	state
205	https://gug.ac.in/	Gulbarga University	Karnataka	state
206	http://www.kannadauniversity.org/kannada/	Kannada University	Karnataka	state
207	http://www.kud.ac.in/	Karnatak University	Karnataka	state
208	http://www.kswu.ac.in/	Karnataka State Women University	Karnataka	state
209	http://www.kuvempu.ac.in/eng/index.php	Kuvempu University	Karnataka	state
210	https://kvafsu.edu.in/	Karnataka Veterinary, Animal & Fisheries Science University	Karnataka	state
211	http://www.kslu.ac.in/	Karnataka State Law University	Karnataka	state

212	https://www.ksoumysuru.ac.in/	Karnataka State Open University	Karnataka	state
213	https://mangaloreuniversity.ac.in/	Mangalore University	Karnataka	state
214	http://www.uni-mysore.ac.in/	Mysore University	Karnataka	state
215	https://www.nls.ac.in/	National law School of India University	Karnataka	state
216	http://www.rguhs.ac.in/	Rajiv Gandhi University of Health Sciences	Karnataka	state
217	http://tumkuruniversity.ac.in/	Tumkur University	Karnataka	state
218	https://www.uasbangalore.edu.in/	University of Agricultural Sciences, Bangalore	Karnataka	state
219	http://www.uasd.edu/	University of Agricultural Sciences, Dharwad	Karnataka	state
220	https://vtu.ac.in/	Visvesvaraya Technological University	Karnataka	state
221	http://uhsbagalkot.edu.in/	University of Horticultural Sciences, Bagalkot	Karnataka	state
222	https://uasraichur.edu.in/kannada/	University of Agricultural Sciences, Raichur	Karnataka	state
223	http://vskub.ac.in/	Vijayanagara Sri Krishnadevaraya University	Karnataka	state
224	http://www.rcub.ac.in/	Rani Channamma University, Belgaum	Karnataka	state
225	https://www.musicuniversity.ac.in/	Karnataka State Music University, Mysore	Karnataka	state
226	http://www.janapadauniversity.ac.in/	Karnataka Janapada Vishwavidyalaya	Karnataka	state
227	https://www.cuk.ac.in/	Central University of Karnataka	Karnataka	Central
228	https://bldedu.ac.in/	B.L.D.E. University	Karnataka	Deemed
229	https://www.iisc.ac.in/	Indian Institute of Science Bangalore	Karnataka	Deemed
230	https://www.iiitb.ac.in/	International Institute of Information Technology	Karnataka	Deemed
231	https://jssuni.edu.in/JSSWeb/WebHome.aspx	Jagadguru Sri Shivarathreeswara University	Karnataka	Deemed
232	http://www.jncasr.ac.in/	Jawaharlal Nehru Centre for Advanced Scientific Research	Karnataka	Deemed

233	http://kledeemeduniversity.edu.in/	K.L.E. Academy of Higher Education and Research	Karnataka	Deemed
234	https://manipal.edu/mu.html	Manipal Academy of Higher Education	Karnataka	Deemed
235	https://nimhans.ac.in/	National Institute of Mental Health & Neuro Sciences	Karnataka	Deemed
236	http://www.nitte.edu.in/	NITTE University	Karnataka	Deemed
237	http://sduu.ac.in/	Sri Devaraj Urs Academy of Higher Education and Research	Karnataka	Deemed
238	https://www.sahe.in/	Sri Siddhartha Academy of Higher Education	Karnataka	Deemed
239	https://svyasa.edu.in/	Swami Vivekananda Yoga Anusandhana Samsthana	Karnataka	Deemed
240	https://www.yenepoya.edu.in/	Yenepoya University	Karnataka	Deemed
241	https://christuniversity.in//	Christ University	Karnataka	Deemed
242	https://www.jainuniversity.ac.in/	Jain University	Karnataka	Deemed
243	http://apsurewa.ac.in/	Awadhesh Pratap Singh University	Madhya Pradesh	state
244	http://www.bubhopal.ac.in/1068/Home	Barkatullah University	Madhya Pradesh	state
245	https://www.dauniv.ac.in/	Devi Ahilya Vishwavidyalaya	Madhya Pradesh	state
246	http://www.jnkvv.org/	Jawaharlal Nehru Krishi Vishwavidyalaya	Madhya Pradesh	state
247	http://www.jiwaji.edu/	Jiwaji University	Madhya Pradesh	state
248	https://www.mgcgvchitrakoot.com/	Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya	Madhya Pradesh	state
249	http://mmyvv.com/index.htm	Maharishi Mahesh Yogi Vedic Vishwavidyalaya	Madhya Pradesh	state
250	http://www.bhojvirtualuniversity.com/	M.P.Bhoj (open) University	Madhya Pradesh	state
251	http://www.mcu.ac.in/	Makhanlal Chaturvedi Rashtriya Patrakarita National University of Journalism	Madhya Pradesh	state
252	https://mpsvvujjain.org/	Maharshi Panini Sanskrit & Vedic Vishwavidyalaya	Madhya Pradesh	state
253	https://www.nliu.ac.in/	National Law Institute University	Madhya Pradesh	state

254	https://www.rgpv.ac.in/	Rajiv Gandhi Proudyogiki Vishwavidyalaya	Madhya Pradesh	state
255	http://www.rdunijbpin.org/1068/Home	Rani Durgavati Vishwavidyalaya	Madhya Pradesh	state
256	http://vikramuniv.ac.in/	Vikram University	Madhya Pradesh	state
257	http://www.ndvsu.org/	Nanaji Deshmukh Veterinary Science University	Madhya Pradesh	state
258	http://www.rvskvv.net/	Rajmata Vijayaraje Scindia Krishi Vishwa Vidyalaya	Madhya Pradesh	state
259	http://www.mpmsu.edu.in/	Madhya Pradesh Medical University	Madhya Pradesh	state
260	http://www.igntu.ac.in/	The Indira Gandhi National Tribal University	Madhya Pradesh	Central
261	http://www.dhsgsu.ac.in/	Dr. Harisingh Gour Vishwavidyalaya	Madhya Pradesh	Central
262	http://rcbhopal.ignou.ac.in/	IGNOU Regional Center Bhopal	Madhya Pradesh	Central
263	http://www.iiitm.ac.in/index.php/en/	Indian Institute of Information Technology and Management	Madhya Pradesh	Deemed
264	http://www.lnipe.edu.in/wordpress/	Lakshmibai National Institute of Physical Education	Madhya Pradesh	Deemed
265	https://www.iiitdmj.ac.in/	Pandit Dwarka Prasad Mishra Indian Institute of Information Technology	Madhya Pradesh	Deemed
266	http://www.bamu.ac.in/	Dr. Babasaheb Ambedkar Marathwada University	Maharashtra	state
267	http://iiecdbatu.com/	Dr. Babasaheb Ambedkar Technological University	Maharashtra	state
268	https://www.pdkv.ac.in/	Dr. Panjabrao Deshmukh Krishi Vidyapeeth	Maharashtra	state
269	http://kksanskrituni.digitaluniversity.ac/	Kavi Kulguru Kalidas Sanskrit Vishwavidyalaya	Maharashtra	state
270	http://www.dbskkv.org/	Dr . Balasaheb Sawant Konkan Krishi Vidyapeeth	Maharashtra	state
271	http://www.mafsu.in/	Maharashtra Animal & Fishery Sciences University	Maharashtra	state
272	https://www.muhs.ac.in/Default.aspx	Maharashtra University of Health Sciences	Maharashtra	state
273	http://mpkv.ac.in/	Mahatma Phule Krishi Vidyapeeth	Maharashtra	state
274	http://vnmkv.ac.in/	Marathwada Agricultural University	Maharashtra	state
275	https://mu.ac.in/	University of Mumbai	Maharashtra	state

276	https://www.nagpuruniversity.ac.in/v2/	Rashtrasant Tukadoji Maharaj Nagpur University	Maharashtra	state
277	http://nmu.ac.in/	North Maharashtra University	Maharashtra	state
278	http://www.unipune.ac.in/	Savitribai Phule Pune University	Maharashtra	state
279	https://www.sgbau.ac.in/	Sant Gadge Baba Amravati University	Maharashtra	state
280	https://sndt.ac.in/	Smt. Nathibai Damodar Thackersey Women's University	Maharashtra	state
281	http://su.digitaluniversity.ac/	Solapur University	Maharashtra	state
282	http://ycmou.digitaluniversity.ac/	Yashwantrao Chavan Maharashtra Open University	Maharashtra	state
283	http://www.srtmun.ac.in/en/	Swami Ramanand Teerth Marathwada University	Maharashtra	state
284	https://hindivishwa.org/	Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya	Maharashtra	Central
285	https://bvuniversity.edu.in/	Bharati Vidyapeeth	Maharashtra	Deemed
286	http://www.cift.res.in/	Central Institute of Fisheries Education	Maharashtra	Deemed
287	http://www.dypatil.edu/	D.Y. Patil Educational Society	Maharashtra	Deemed
288	https://www.dmimsu.edu.in/	Datta Meghe Institute of Medical Sciences	Maharashtra	Deemed
289	http://www.dcpune.ac.in/	Deccan College Postgraduate & Research Institute	Maharashtra	Deemed
290	https://dpu.edu.in/	Dr. D.Y. Patil Vidyapeeth	Maharashtra	Deemed
291	http://www.gipe.ac.in/	Gokhale Institute of Politics & Economics	Maharashtra	Deemed
292	http://www.hbni.ac.in/	Homi Bhabha National Institute	Maharashtra	Deemed
293	http://www.igidr.ac.in/	Indira Gandhi Institute of Development Research	Maharashtra	Deemed
294	https://www.diat.ac.in/	Defence Institute Of Advanced Technology	Maharashtra	Deemed
295	https://iipsindia.ac.in/	International Institute for Population Sciences	Maharashtra	Deemed
296	http://www.kimskarad.in/	Krishna Institute of Medical Sciences	Maharashtra	Deemed
297	https://www.mgmuhs.com/	MGM Institute of Health Sciences	Maharashtra	Deemed
298	https://nmims.edu/	SVKM's Narsee Monjee Institute of Management Studies	Maharashtra	Deemed

299	https://www.pravara.com/	Pravara Institute of Medical Sciences	Maharashtra	Deemed
300	https://siu.edu.in/	Symbiosis International University	Maharashtra	Deemed
301	https://www.tifr.res.in/	Tata Institute of Fundamental Research	Maharashtra	Deemed
302	https://www.tiss.edu/	Tata Institute of Social Sciences	Maharashtra	Deemed
303	http://www.tmv.edu.in/	Tilak Maharashtra Vidyapeeth	Maharashtra	Deemed
304	https://www.ictmumbai.edu.in/	Institute of Chemical Technology	Maharashtra	Deemed
305	https://www.cau.ac.in/	Central Agricultural University	Manipur	Central
306	https://www.manipuruniv.ac.in/	Manipur University	Manipur	Central
307	https://www.nehu.ac.in/	North Eastern Hill University	Meghalaya	Central
308	http://eflushc.ac.in/	The English and Foreign Language University Shillong	Meghalaya	Central
309	http://rcshillong.ignou.ac.in/	Indira Gandhi National Open University, Shillong campus	Meghalaya	Central
310	https://www.ustm.ac.in/	Meghalaya State Technical University	Meghalaya	state
311	http://mzu.edu.in/	Mizoram University	Mizoram	Central
312	https://mzu.edu.in/	Nagaland University	Nagaland	Central
313	https://www.iunagaland.edu.in/	ICFAI University Nagaland	Nagaland	Central
314	https://tgounagaland.com/	The Global Open University Nagaland	Nagaland	state
315	http://bamu.nic.in/	Berhampur University	Odisha	state
316	http://www.bput.ac.in/	Biju Patnaik University of Technology	Odisha	state
317	http://www.fmuniversity.nic.in/	Fakir Mohan University	Odisha	state
318	http://www.nou.nic.in/	North Orissa University	Odisha	state
319	http://www.ouat.nic.in/	Orissa University of Agriculture & Technology	Odisha	state
320	https://www.ravenshawuniversity.ac.in/Home.php	Ravenshaw University	Odisha	state
321	http://www.suniv.ac.in/	Sambalpur University	Odisha	state
322	http://www.sjsv.nic.in/	Shri Jagannath Sanskrit Vishwavidyalaya	Odisha	state

323	https://www.utkaluniversity.nic.in/	Utkal University	Odisha	state
324	http://uuc.ac.in/	Utkal University of Culture	Odisha	state
325	https://www.nluo.ac.in/	National Law University, Odisha	Odisha	state
326	http://vssut.ac.in/	Veer Surendra Sai University of Technology	Odisha	state
327	http://cuo.ac.in/	Central University of Orissa	Odisha	Central
328	https://kiit.ac.in/	Kalinga Insitute of Industrial Technology	Odisha	Deemed
329	https://www.soa.ac.in/	Shiksha 'O' Anusandhan University	Odisha	Deemed
330	http://www.pondiuni.edu.in/	Pondicherry University	Pondicherry	Central
331	http://sbvu.ac.in/	Sri Balaji Vidyapeeth University	Pondicherry	Deemed
332	http://www.bfuhs.ac.in/	Baba Farid University of Health & Medical Sciences	Punjab	state
333	http://online.gndu.ac.in/	Guru Nanak Dev University	Punjab	state
334	https://www.gadvasu.in/	Guru Angad Dev Veterinary & Animal Sciences University	Punjab	state
335	https://www.pau.edu/	Punjab Agricultural University	Punjab	state
336	http://www.ptu.ac.in/	I. K. Gujral Punjab Technical University	Punjab	state
337	https://puchd.ac.in/	Panjab University, Chandigarh	Punjab	state
338	http://www.punjabiuniversity.ac.in/	Punjabi University, Patiala	Punjab	state
339	http://www.rgnul.ac.in/	The Rajiv Gandhi National University of Law	Punjab	state
340	http://cup.edu.in/	Central University of Punjab	Punjab	Central
341	http://sliet.ac.in/	Sant Longowal Institute of Engineering and Technology	Punjab	Deemed
342	http://www.thapar.edu/	Thapar Institute of Engineering & Technology	Punjab	Deemed
343	http://www.jnvu.edu.in/	Jai Narain Vyas University	Rajasthan	state
344	http://www.jrrsanskrituniversity.ac.in/	Jagadguru Ramanandacharya Sanskrit University	Rajasthan	state
345	https://www.mpuat.ac.in/	Maharana Pratap University of Agriculture & Technology	Rajasthan	state
346	http://www.mdsuajmer.ac.in/	Maharshi Dayanand Saraswati University	Rajasthan	state

347	https://www.mlsu.ac.in/	Mohanlal Sukhadia University	Rajasthan	state
348	http://nlujodhpur.ac.in/index-main.php	National Law University jodhpur	Rajasthan	state
349	http://raubikaner.org/	Swami Keshwanand Rajasthan Agriculture University	Rajasthan	state
350	https://education.rajasthan.gov.in/content/raj/education/en/home.html	Dr. Sarvepalli Radhakrishnan Rajasthan Ayurved University	Rajasthan	state
351	https://www.uniraj.ac.in/	University of Rajasthan	Rajasthan	state
352	http://www.rtu.ac.in/RTU/	Rajasthan Technical University	Rajasthan	state

Table A.2: List of Universities for data collection

Appendix B

Reply Summary

Table B.1 provides reply summary by individual boards to the RTI application filed for obtaining the data. This table also records whether a followup is required or not based on the reply summary presented. We note that only four out of 64 boards provided data in the first attempt. Rest of the boards required to appeals to the respective institutes. About 20% of boards did not reply to the RTI query.

Sl. No	**Board Name**	**Reply Summary**	**Follow-up Required?**
1	Board of Secondary Education Andhra Pradesh	Make payment for the RTI applications, partial data present	yes
2	Board of Intermediate Education Andhra Pradesh	Data available only after 2015, make payment	yes
3	Secondary Education Board of Assam	Reply statistics provided, visit office for queries	yes
4	Assam Higher Secondary Education Council	NO REPLY	yes
5	Bihar School Examination Board	Cannot give non-existent information	yes
6	Bihar Board of Open Schooling	Data provided	no
7	Bihar Intermediate Education Council	NO REPLY	yes
8	Chhattisgarh Board of Secondary Education	Only summary statistics are provided, visit office for the remaining data	yes
9	Chattisgarh State Open School	Make payment	yes

10	Goa Board of Secondary and Higher Secondary Education	Make payment for the RTI applications	yes
11	Gujarat Secondary and Higher Secondary Education Board	NO REPLY	yes
12	Haryana Board of School Education	Make payment	yes
13	H.P. Board of School Education	Rejected. Appeal won by us. Data not received yet	yes
14	Jharkhand Academic Council	Cannot provide personal information of applicants	yes
15	Karnataka Secondary Education Examination Board	Information has been disposed off	yes
16	Government of Karnataka Dept. of Pre-University Education	Only summary statistics are provided, visit office for the data	yes
17	Kerala Board of Higher Secondary Education	Huge data, visit office	yes
18	Kerala Board of Public Examinations	Data is not present, only summary statistics are provided	yes
19	M.P. State Open School Board of Secondary Education Campus	Visit office	yes
20	Madhya Pradesh Board of Secondary Education	Incomplete information in the RTI application	yes
21	Maharashtra State Board of Secondary and Higher Secondary Education	NO REPLY	yes
22	Manipur Board of Secondary Education	NO REPLY	yes
23	Council of Higher Secondary Education Manipur	Cannot give name of applicants. Data received after appeal won.	yes
24	Mizoram Board of School Education	Make payment. Data provided	yes
25	Nagaland Board of School Education	Data provided	no

26	Council of Higher Secondary Education, Odisha	Make payment	yes
27	Board of Secondary Education Odisha	Make payment for the RTI applications	yes
28	Punjab School Education Board	NO REPLY	yes
29	Board of Secondary Education Rajasthan	Cannot provide personal information of applicants, diverts resources	yes
30	Directorate of School Education and Administration	Make payment for RTI queries	yes
31	School Education Department	NO REPLY	yes
32	Telangana Board of Intermediate Education	Make payment (but amount not mentioned)	yes
33	Telangana Board of Secondary Education	Huge data, only summary statistics are provided	yes
34	Telangana Open School Society	Make payment. Registers with abridged queries provided	yes
35	Tripura Board of Secondary Education	NO REPLY	yes
36	Board of Secondary Education Kant Shahjahanpur Uttar Pradesh	Only summary statistics are provided, private organization	yes
37	U.P. Board of High School	NO REPLY	yes
38	Uttranchal Shiksha Evm Pariksha Prishad	Make payment	yes
39	West Bengal Council of Higher Secondary Education	Queries do not come under RTI act, personal information	yes
40	West Bengal Board of Primary Education	No record as soft copy	yes
41	West Bengal Board of Secondary Education	NO REPLY	yes
42	Ravindra Mukta Vidyalaya	NO REPLY	yes
43	Assistant Secretary (Admn.II)/PIO	Voluminous information, diverts resources, visit office on 15-01-2016	yes

44	Assistant Secretary (Co-ord)/PIO	Voluminous information, diverts resources, visit office	yes
45	Joint Director (Academic	Cannot provide personal information of applicants	yes
46	Assistant Secretary (CTET)/PIO	Cannot provide personal information of applicants	yes
47	Assistant Secretary (AIPMT)/PIO	Voluminous information, diverts resources,	yes
48	Desk Officer (Affln)/PIO	Data not available	yes
49	Deputy Secretary (JEE)/PIO	Cannot provide personal information of applicants	yes
50	Assistant Public Relation Officer/PIO	Send RTI fee in favour of "The Secretary, CBSE"	yes
51	Assistant Secretary/PIO	Voluminous information, diverts resources, visit office	yes
52	Assistant Secretary/PIO	Voluminous information, diverts resources, visit office on 16-01-2016	yes
53	Deputy Director (Vocational)/PIO	A few RTI applications provided, no reply statistics	yes
54	Assistant Secretary/PIO	Cannot provide personal information of applicants	yes
55	Assistant Secretary/PIO	Data not stored in required format, diverts resource, visit office on 21.12.2015 & 22.12.2015	no
56	Deputy Secretary/PIO	Voluminous information, diverts resources, visit office	yes
57	Assiatant Secretary/PIO	Make payment. Only RTI applications provided, reply statistics are not provided	yes
58	Deputy Secretary/PIO	Visit office within one month	yes
59	Assistant Secretary/PIO	Required information is not related to me, so cannot provide information	yes
60	Assistant Secretary/PIO	Huge data, visit office	yes
61	Board of Higher Secondary Education, Delhi	NO REPLY	yes
62	Board of Youth Education India	RTI application returned, address not found	no
63	National Institute of Open Schooling	Replied that information is present in the website. But no such information is present.	yes
64	Rashtriya Sanskrit Sansthan, Delhi	NO REPLY	yes

Table B.1: Reply summary from class 10, class 10 + 2 boards

Table B.2 provides reply summary by individual universities to the RTI application filed

for obtaining the data. This table also records whether a followup is required or not based on the reply summary presented. About 23% of universities did not reply to our filed RTI query.

Sl. No	**University Name**	**Reply Summary**	**Follow up Required?**
65	University of Hyderabad	data not in specified format, diverts resources, visit office	yes
66	Maulana Azad National Urdu University	Huge data, diverts resources	yes
67	THE ENGLISH AND FOREIGN LANGUAGES UNIVERSITY	Cannot provide personal data of applicants, visit office	yes
68	Acharya N G Ranga Agricultural University	NO REPLY	yes
69	Adikavi Nannaya University	Make payment. Data provided.	yes
70	Andhra University	Cannot provide personal data of applicants, visit office, abridged queries sent	yes
71	Acharya Nagarjuna University	Cannot provide personal data of applicants	yes
72	Dravidian University	Huge data, diverts resources, visit office	yes
73	Dr B R Ambedkar Open University	Only reply statistics provided, no RTI queries	yes
74	Dr. NTR University of Health Sciences	Huge data so visit office, reply statistics provided	yes
75	Jawaharlal Nehru Technological University	Huge data, visit office	yes
76	Jawaharlal Nehru Technological University Kakinada	NO REPLY	yes
77	Kakatiya University	Huge data so visit office, reply statistics provided	yes
78	Krishna University	Make payment. RTI applications provided bu tnot reply statistics	yes
79	Mahatma Gandhi University	Make payment. Visit office for reply statistics	yes
80	National Academy of Legal Studies	Visit office	yes
81	Osmania University	Huge data, visit office	yes
82	Palamuru University	NO REPLY	yes
83	Potti Sreeramulu Telugu University	Cannot provide personal data of applicants	yes
84	Rayalaseema University	Make payment	yes
85	Satavahana University	Data provided	yes
86	Sri Krishnadevaraya University	Make payment	yes
87	Sri Padmavati Mahila Vishwavidyalayam	Huge data, visit office. Some statistics provided	yes

88	Sri Venkateswara Veterinary University	Data provided	no
89	Sri Venkateswara University	Data provided partially	yes
90	Sri Venkateswara Vedic University	Make payment. Data provided.	yes
91	Telangana University	Make payment. Data provided.	yes
92	Vikrama Simhapuri University	Cannot provide personal data of applicants	yes
93	Yogi Vemana University	NO REPLY	yes
94	Rajiv Gandhi University of Knowledge Technologies	NO REPLY	yes
95	Jawaharlal Nehru Architecture	Make payment. Data provided.	yes
96	A.P.Horticultural University	Make payment	yes
97	Gandhi Institute of Technology and Management	NO REPLY	yes
98	International Institute of Information Technology, Hyderabad	NO REPLY	yes
99	Rashtriya Sanskrit Vidyapeeth	NO REPLY	yes
100	Sri Sathya Sai Institute of Higher Learning	Some statistics are provided. No queries and reply dates.	yes
101	ICFAI Foundation for Higher Education	NO REPLY	yes
102	Vignan University	Data provided	yes
103	Koneru Lakshmaiah Education Foundation (K L University)	Data provided	no
104	Rajiv Gandhi University	NO REPLY	yes
105	North Eastern Regional Institute of Science	Visit office	yes
106	Indira Gandhi Technological and Medical Sciences University	NO REPLY	yes
107	Arunachal University of Studies	NO REPLY	yes
108	Venkateshwara Open University	NO REPLY	yes
109	Apex Professional University	Make payment	yes
110	Himalayan University	RTI application returned, address not found	
111	Assam University	Cannot provide personal data of applicants. Make payment for reply statistics.	yes
112	Tezpur University	Data provided	yes

113	National Law University and Judicial Academy	Visit office	yes
114	Assam Science	RTI application returned, address not found	no
115	Assam Agricultural University	Make payment. Data provided.	yes
116	Dibrugarh University	Huge data, visit office	yes
117	Gauhati University	Make payment	yes
118	Bodoland University	Make payment. Data provided.	yes
119	Krishna Kanta Handique State Open University	Make payment. Data provided.	yes
120	Cotton College State University	Cannot provide personal data of applicants. Make payment for reply statistics.	yes
121	Srimanta Sankaradeva University of Health Sciences	Huge data, application rejected	yes
122	Bhupendra Narayan Mandal University	NO REPLY	yes
123	Babasaheb Bhimrao Ambedkar Bihar University	NO REPLY	yes
124	Aryabhatta Knowledge University	Required information is not under the RTI Act	yes
125	Chanakya National Law University	Make payment	yes
126	Jai Prakash University	NO REPLY	yes
127	Kameshwar Singh Darbhanga Sanskrit University	Summary statistics provided, make payment for the data	yes
128	Lalit Narayan Mithila University	Cannot provide information. No reason for rejection is stated.	yes
129	Magadh University	Visit office	yes
130	Maulana Mazharul Haque Arabic and Persian University	NO REPLY	yes
131	Nalanda Open University	NO REPLY	yes
132	Patna University	NO REPLY	yes
133	Tilka Manjhi Bhagalpur University	RTI application is not clear	yes
134	Veer Kunwar Singh University	NO REPLY	yes
135	Bihar Agricultural University	Summary statistics provided. Huge data, visit office	yes
136	Rajendra Agricultural University	Make payment. Data received.	yes
137	Central University of South Bihar	NO REPLY	yes
138	Aligarh Muslim University	Make payment	yes

139	Nalanda International University	NO REPLY	yes
140	Nava Nalanda Mahavihara	NO REPLY	yes
141	Bihar Yoga Bharati	Private university, does not come under the RTI Act	yes
142	Bastar Vishwavidyalaya	NO REPLY	yes
143	Chhattisgarh Swami Vivekanand Technical University	Visit office	yes
144	Hidayatullah National Law University	Make payment	yes
145	Indira Gandhi Krishi Vishwavidyalaya	Huge data, visit office	yes
146	Indira Kala Sangeet Vishwavidyalaya	Cannot provide RTI applications	yes
147	Kushabhau Thakre Patrakarita Avam Jansanchar Vishwavidyalaya	Make payment. RTI applications received but no reply dates	yes
148	Pt. Ravishankar Shukla University	Visit office	yes
149	Pt. Sundarlal Sharma (Open) University	Make payment. RTI applications received but no reply dates	yes
150	Sarguja University	NO REPLY	yes
151	Guru Ghasidas Vishwavidyalaya	Huge data, visit office. Make payment for summary statistics	yes
152	C.V.Raman University	Private university, does not come under the RTI Act. Visit office to observe the information	yes
153	Goa University	Make payment	yes
154	Gujarat University	Replied in Gujarati, has not been translated	yes
155	Sardar Patel University	Make payment	yes
156	Veer Narmad South Gujarat University	Summary statistics provided. Make payment for the data	yes
157	The Maharaja Sayajirao University	Information does not come under the RTI Act	yes
158	Saurashtra University	RTI application fee is Rs. 20	yes
159	Bhavnagar University	RTI application fee is Rs. 20. Cannot provide personal data of applicants	yes
160	Hemchandracharya North Gujarat University	NO REPLY	yes
161	Krantiguru Shyamji Krishna Verma Kachchh University	NO REPLY	yes
162	Dr. Babasaheb Ambedkar Open University	RTI application is Rs. 20	yes
163	Shree Somnath Sanskrit University	RTI application is Rs. 20	yes

164	Children's University Gujarat	RTI application is Rs. 20	yes
165	Gujarat Forensic Sciences University	Make payment. Data received.	yes
166	Raksha Shakti University	Make payment. Data received.	yes
167	Kamdhenu University	Make payment for reply dates, cannot provide RTI applications	yes
168	The Indian Institute of Teacher Education	Make payment. Data received.	yes
169	Gujarat Technological University	Huge data, visit office	yes
170	Gujarat Ayurved University	Make payment	yes
171	Anand Agricultural University	Make payment. Data received.	yes
172	Navsari Agricultural University	Make payment, reply received from multiple sections separately	yes
173	Sardarkrushinagar Dantiwada Agricultural University	Visit office	yes
174	Junagadh Agricultural University	Make payment	yes
175	Gujarat Vidyapith	Partial data sent	yes
176	Central University of Gujarat	Make payment	yes
177	Sumandeep Vidyapith	Deemed university, not under the RTI Act, rejected	yes
178	Bhagat Phool Singh Mahila Vishwavidyalaya	Application fee is Rs. 50, make payment for data	yes
179	Chaudhary Devi Lal University	Application fee is Rs. 50, make payment for data	yes
180	Chaudhary Charan Singh Haryana Agricultural University	Application fee is Rs. 50, huge data and no public interest, visit office	yes
181	Deen Bandhu Chhotu Ram University of Science	Application fee is Rs. 50. Data not present in any form	yes
182	Guru Jambeshwar University of Science and Technology	Application fee is Rs. 50, make payment for data within 15 days	yes
183	Kurukshetra University	NO REPLY	yes
184	Maharshi Dayanand University	Application fee is Rs. 50. Huge data so ask for specifc information	yes
185	Pt. Bhagwat Dayal Sharma University of Health Sciences	Make payment. After payment replied huge data so visit office.	yes
186	Central University of Haryana	Huge data, visit office	yes

187	Maharishi Markandeshwar University	Private university, not under the RTI Act, rejected	yes
188	National Brain Research Centre	Huge data, rejected	yes
189	National Dairy Research Institute	Make payment	yes
190	Manav Rachna International University	NO REPLY	yes
191	Lingaya's University	RTI application returned, address not found	no
192	Dr. Y.S.Parmar University of Horticulture	Send separate applications for each year and each PIO	yes
193	Himachal Pradesh University	Cannot provide personal data of applicants. Make payment for annual statistics	yes
194	Himachal Pradesh Agriculture University	Make payment, reply received from multiple sections separately	yes
195	Himachal Pradesh Technical University	Make payment	yes
196	Central University of Himachal Pradesh	Make payment	yes
197	Birsa Agricultural University	NO REPLY	yes
198	Ranchi University	Huge data, rejected	yes
199	Sido Kanhu Murmu University	Cannot provide personal data of applicants	yes
200	Kolhan University	Make payment within 15 days	yes
201	Nilamber-Pitamber University	Make payment. Data not received yet	yes
202	Vinoba Bhave University	Make payment. Data not received yet	yes
203	Bangalore University	Replied in Kannada, has not been translated	yes
204	Davangere University	Cannot provide personal data of applicants	yes
205	Gulbarga University	Make payment	yes
206	Kannada University	NO REPLY	yes
207	Karnatak University	Make payment. Only RTI applications provided.	yes
208	Karnataka State Women University	Make payment. Only RTI reply dates provided	yes
209	Kuvempu University	Partial data sent. Make payment for the remaining data.	yes
210	Karnataka Veterinary, Animal	Visit office on 19-12-2015	yes
211	Karnataka State Law University	Make payment	yes
212	Karnataka State Open University	NO REPLY	yes
213	Mangalore University	Make payment. Data received	yes
214	Mysore University	Could not understand their reply	yes

215	National law School of India University	Cannot provide personal data of applicants	yes
216	Rajiv Gandhi University of Health Sciences	NO REPLY	yes
217	Tumkur University	Information does not come under the RTI Act. Appeal accepted, make payment.	yes
218	University of Agricultural Sciences, Bangalore	Make payment, reply received from multiple sections separately	yes
219	University of Agricultural Sciences, Dharwad	Make payment. Data received	yes
220	Visvesvaraya Technological University	Cannot provide information. No reason for rejection is stated.	yes
221	University of Horticultural Sciences, Bagalkot	Make payment, reply received from multiple sections separately	yes
222	University of Agricultural Sciences, Raichur	Make payment. RTI applications received, but not reply dates	yes
223	Vijayanagara Sri Krishnadevaraya University	Replied in Kannada, has not been translated	yes
224	Rani Channamma University, Belgaum	Make payment. Data received	yes
225	Karnataka State Music University, Mysore	NO REPLY	yes
226	Karnataka Janapada Vishwavidyalaya	Make payment. RTI applications received, but not reply dates	yes
227	Central University of Karnataka	NO REPLY	yes
228	B.L.D.E. University	NO REPLY	yes
229	Indian Institute of Science Bangalore	Visit office	yes
230	International Institute of Information Technology	Make payment	yes
231	Jagadguru Sri Shivarathreeswara University	Reply dates sent, but no RTI applications	yes
232	Jawaharlal Nehru Centre for Advanced Scientific Research	Partial data sent	yes
233	K.L.E. Academy of Higher Education and Research	Deemed university, not under the RTI Act, rejected	yes
234	Manipal Academy of Higher Education	Deemed university, not under the RTI Act, rejected	yes
235	National Institute of Mental Health	NO REPLY	yes
236	NITTE University	Deemed university, not under the RTI Act, rejected	yes
237	Sri Devaraj Urs Academy of Higher Education and Research	Data received	no
238	Sri Siddhartha Academy of Higher Education	NO REPLY	yes

239	Swami Vivekananda Yoga Anusandhana Samsthana	NO REPLY	yes
240	Yenepoya University	Deemed university, not under the RTI Act, rejected	yes
241	Christ University	Deemed university, not under the RTI Act, rejected	yes
242	Jain University	NO REPLY	yes
243	Awadhesh Pratap Singh University	Make payment	yes
244	Barkatullah University	Data not stored in required format, visit office	yes
245	Devi Ahilya Vishwavidyalaya	Send RTI fee to correct account, no reply after sending the fee	yes
246	Jawaharlal Nehru Krishi Vishwavidyalaya	NO REPLY	yes
247	Jiwaji University	Cannot provide personal data of applicants	yes
248	Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya	Data is being prepared, but have not received yet	yes
249	Maharishi Mahesh Yogi Vedic Vishwavidyalaya	Could not understand their reply	yes
250	M.P.Bhoj (open) University	Huge data, visit office, or ask for specific information	yes
251	Makhanlal Chaturvedi Rashtriya Patrakarita National University of Journalism	NO REPLY	yes
252	Maharshi Panini Sanskrit	Make payment. Data received	yes
253	National Law Institute University	Ask for specific information. After follow-up replied huge data so visit office	yes
254	Rajiv Gandhi Proudyogiki Vishwavidyalaya	Visit office within one day	yes
255	Rani Durgavati Vishwavidyalaya	Send RTI fee to correct account. After correction, huge data so visit office	yes
256	Vikram University	NO REPLY	yes
257	Nanaji Deshmukh Veterinary Science University	NO REPLY	yes
258	Rajmata Vijayaraje Scindia Krishi Vishwa Vidyalaya	NO REPLY	yes
259	Madhya Pradesh Medical University	Make payment. Data not received yet	yes
260	The Indira Gandhi National Tribal University	Payment required. Did not mention the amount after follow-up application.	yes
261	Dr. Harisingh Gour Vishwavidyalaya	NO REPLY	yes
262	IGNOU Regional Center Bhopal	NO REPLY	yes

263	Indian Institute of Information Technology and Management	Make payment. Data received	yes
264	Lakshmibai National Institute of Physical Education	Make payment. Data received. Only register is available	yes
265	Pandit Dwarka Prasad Mishra Indian Institute of Information Technology	Make payment. Data not received yet	yes
266	Dr. Babasaheb Ambedkar Marathwada University	Letter returned, asked to send to correct address	yes
267	Dr. Babasaheb Ambedkar Technological University	NO REPLY	yes
268	Dr. Panjabrao Deshmukh Krishi Vidyapeeth	Cannot provide personal data of applicants	yes
269	Kavi Kulguru Kalidas Sanskrit Vishwavidyalaya	Huge data, visit office	yes
270	Dr . Balasaheb Sawant Konkan Krishi Vidyapeeth	Make payment. Partial data received.	yes
271	Maharashtra Animal	Replied in Marathi, has not been translated	yes
272	Maharashtra University of Health Sciences	Cannot provide personal data of applicants. Appeal rejected.	yes
273	Mahatma Phule Krishi Vidyapeeth	Send separate applications to each PIO	yes
274	Marathwada Agricultural University	Could not understand their reply	yes
275	University of Mumbai	Make payment, and collect data by visiting office	yes
276	Rashtrasant Tukadoji Maharaj Nagpur University	Replied in Marathi, has not been translated	yes
277	North Maharashtra University	Data not present in desired format	yes
278	Savitribai Phule Pune University	Visit office	yes
279	Sant Gadge Baba Amravati University	Huge data, diverts resources, visit office	yes
280	Smt. Nathibai Damodar Thackersey Women's University	Partial summary statistics provided. No reply after follow-up	yes
281	Solapur University	Send separate applications to each PIO, huge data so visit office	yes
282	Yashwantrao Chavan Maharashtra Open University	Make payment for RTI applications. Cannot create other required information	yes
283	Swami Ramanand Teerth Marathwada University	Cannot provide data	yes

284	Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya	Make payment	yes
285	Bharati Vidyapeeth	Make payment	yes
286	Central Institute of Fisheries Education	Make payment. Data received	yes
287	D.Y. Patil Educational Society	Deemed university, not under the RTI Act, rejected	yes
288	Datta Meghe Institute of Medical Sciences	NO REPLY	yes
289	Deccan College Postgraduate	Cannot provide personal data of applicants. Partial data sent	yes
290	Dr. D.Y. Patil Vidyapeeth	Deemed university, not under the RTI Act, rejected	yes
291	Gokhale Institute of Politics	Visit office	yes
292	Homi Bhabha National Institute	Make payment. Data received	yes
293	Indira Gandhi Institute of Development Research	Summary statistics and abridge queries sent	yes
294	Defence Institute Of Advanced Technology	Make payment. Data not received yet	yes
295	International Institute for Population Sciences	Cannot provide required data, ask for specific information	yes
296	Krishna Institute of Medical Sciences	NO REPLY	yes
297	MGM Institute of Health Sciences	Deemed university, not under the RTI Act, rejected	yes
298	SVKM's Narsee Monjee Institute of Management Studies	Deemed university, not under the RTI Act, rejected	yes
299	Pravara Institute of Medical Sciences	NO REPLY	yes
300	Symbiosis International University	Deemed university, not under the RTI Act, rejected	yes
301	Tata Institute of Fundamental Research	Make payment. Register with abridged queries received	yes
302	Tata Institute of Social Sciences	Make payment	yes
303	Tilak Maharashtra Vidyapeeth	Replied in Marathi. Summary statistics provided	yes
304	Institute of Chemical Technology	NO REPLY	yes
305	Central Agricultural University	Visit office	yes
306	Manipur University	NO REPLY	yes
307	North Eastern Hill University	Make payment	yes

308	The English and Foreign Language University Shillong	Data received	no
309	Indira Gandhi National Open University, Shillong campus	Data received	yes
310	Meghalaya State Technical University	NO REPLY	yes
311	Mizoram University	NO REPLY	yes
312	Nagaland University	NO REPLY	yes
313	ICFAI University Nagaland	NO REPLY	yes
314	The Global Open University Nagaland	NO REPLY	yes
315	Berhampur University	Visit office	yes
316	Biju Patnaik University of Technology	Cannot provide personal data of applicants. Cannot compile reply and rejection statistics	yes
317	Fakir Mohan University	RTI information is not meant for research	yes
318	North Orissa University	NO REPLY	yes
319	Orissa University of Agriculture	Make payment within 15 days	yes
320	Ravenshaw University	Huge data, visit office	yes
321	Sambalpur University	Huge data, seek specific information	yes
322	Shri Jagannath Sanskrit Vishwavidyalaya	NO REPLY	yes
323	Utkal University	NO REPLY	yes
324	Utkal University of Culture	Register with abridged queries received	yes
325	National Law University, Odisha	NO REPLY	yes
326	Veer Surendra Sai University of Technology	NO REPLY	yes
327	Central University of Orissa	Huge data, visit office within 15 days	yes
328	Kalinga Insitute of Industrial Technology	Letter returned, asked to send to correct address	yes
329	Shiksha 'O' Anusandhan University	NO REPLY	yes
330	Pondicherry University	Huge data, visit office	yes
331	Sri Balaji Vidyapeeth University	Deemed university, not under the RTI Act, rejected	yes
332	Baba Farid University of Health	Huge data, visit office	yes
333	Guru Nanak Dev University	Huge data, not related to public interest. Appeal rejected	yes
334	Guru Angad Dev Veterinary	Cannot provide personal data of applicants. Appeal rejected	yes

335	Punjab Agricultural University	Make payment	yes
336	I. K. Gujral Punjab Technical University	NO REPLY	yes
337	Panjab University, Chandigarh	Visit office	yes
338	Punjabi University, Patiala	Huge data. Rejected	yes
339	The Rajiv Gandhi National University of Law	Huge data, visit office	yes
340	Central University of Punjab	Make payment	yes
341	Sant Longowal Institute of Engineering and Technology	Huge data and cannot provide personal information of applicants, visit office	yes
342	Thapar Institute of Engineering	Deemed university, not under the RTI Act, rejected	yes
343	Jai Narain Vyas University	Huge data, visit office	yes
344	Jagadguru Ramanandacharya Sanskrit University	Data not present in required format, visit office	yes
345	Maharana Pratap University of Agriculture	NO REPLY	yes
346	Maharshi Dayanand Saraswati University	Huge data, diverts resources.	yes
347	Mohanlal Sukhadia University	Huge data, diverts resources, visit office	yes
348	National Law University jodhpur	Register with abridged queries received	yes
349	Swami Keshwanand Rajasthan Agriculture University	Make payment	yes
350	Dr. Sarvepalli Radhakrishnan Rajasthan Ayurved University	Ask for specific information	yes
351	University of Rajasthan	Asked to specify what public interest is there in the data required by us	yes
352	Rajasthan Technical University	Huge data, diverts resources.	yes

Table B.2: Reply summary from class 10, class 10 + 2 boards

Appendix C
QCCC Plots

QCCC curves for all the 10 query categories are presented in this appendix. These plots are arranged in the sorted order of query category names.

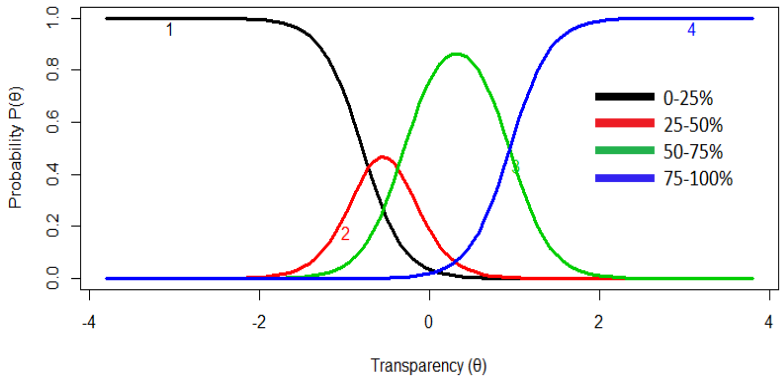


Figure C.1: Query-Category Characteristic Curve for *Administration*

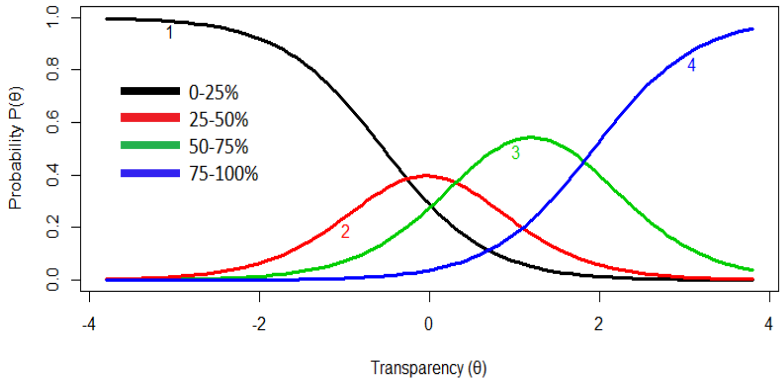


Figure C.2: Query-Category Characteristic Curve for *Admission*

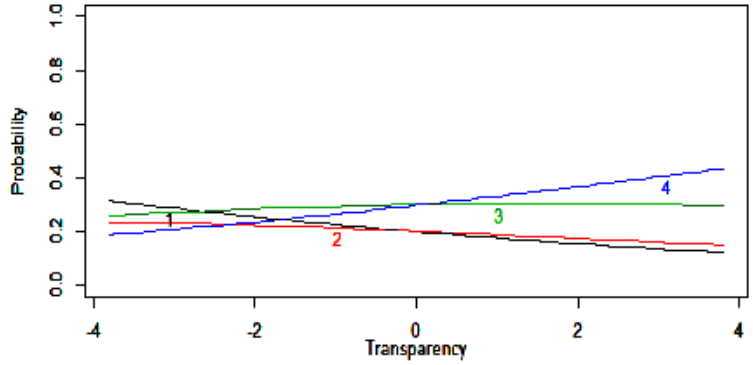


Figure C.3: Query-Category Characteristic Curve for *Affiliation*

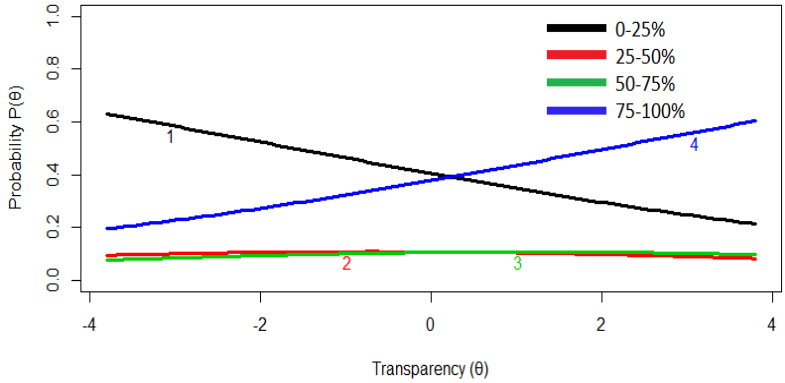


Figure C.4: Query-Category Characteristic Curve for *Course*

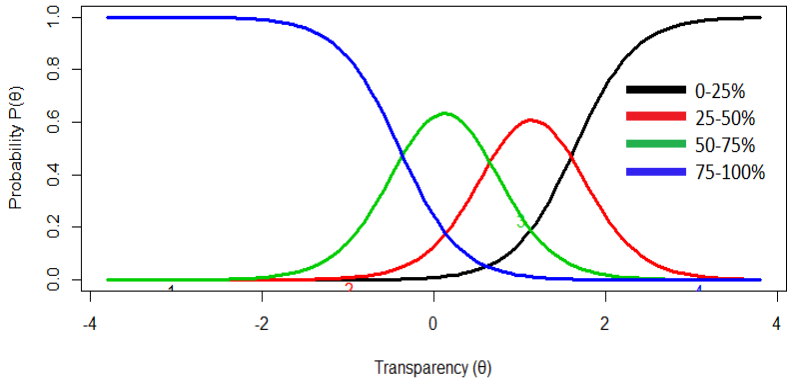


Figure C.5: Query-Category Characteristic Curve for *Exam*

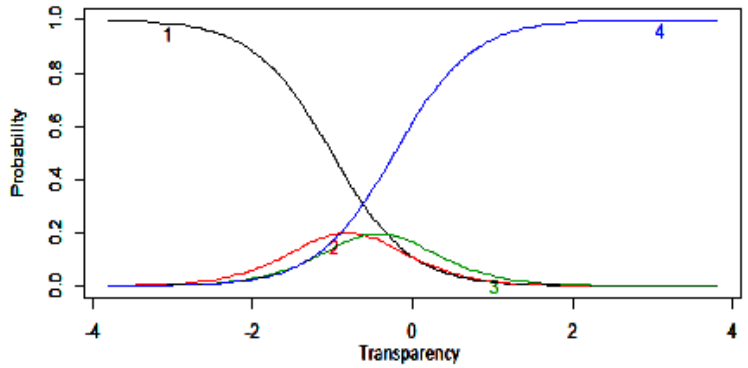


Figure C.6: Query-Category Characteristic Curve for *Finance*

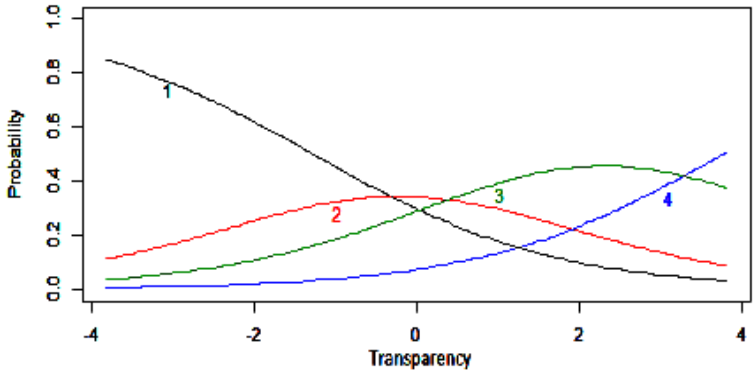


Figure C.7: Query-Category Characteristic Curve for *Recruitment*

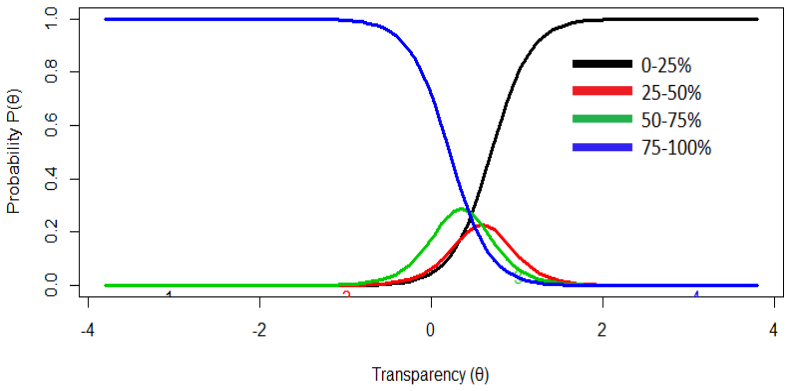


Figure C.8: Query-Category Characteristic Curve for *RTI*

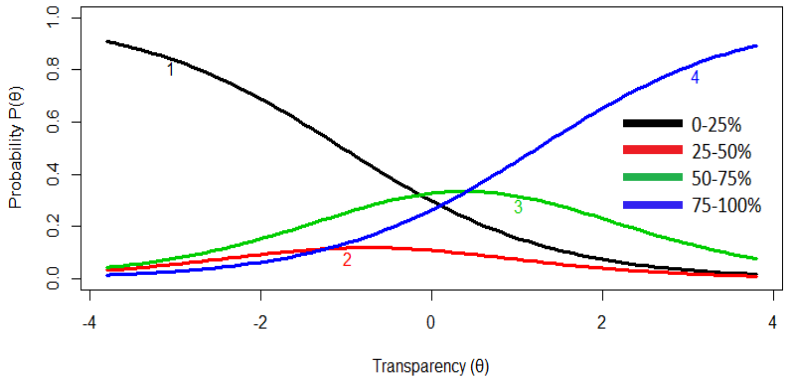


Figure C.9: Query-Category Characteristic Curve for *Staff*

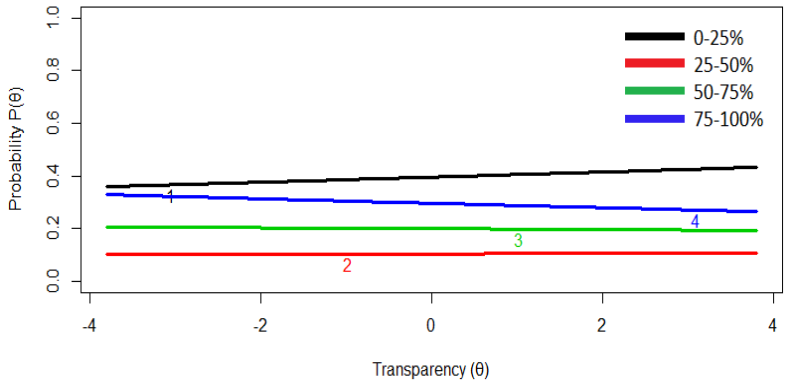


Figure C.10: Query-Category Characteristic Curve for *Students*

Vitae

Name: *Nayantara Kotoky*

Email: *nayantara@iitg.ac.in*
nayantara.kotoky@gmail.com

Phone number: *+91-9854239057*

LinkedIn ID: *linkedin.com/in/nayantara-kotoky-1b422668*

CURRENT POSITION

Assistant Professor,
Dept. of Computer Science and Engineering,
Pandit Deendayal Petroleum University,
Gandhinagar, Gujarat-382355.

EDUCATION

Doctor of Philosophy (continuing)

Specialization	:	Computer Science and Engineering
Institution	:	Indian Institute of Technology Guwahati

Master of Technology

Specialization	:	Information Technology
Institution	:	Tezpur Central University
Year	:	2013

Bachelor of Engineering

Specialization	:	Computer Science and Engineering
Institution	:	Assam Engineering College
Year	:	2011

AREA OF INTEREST

- Machine Learning (ML)/Artificial Intelligence (AI) for Social Good
- ML/AI applications in Policy-Making
- ML/AI Applications in Cognitive Psychology

Bibliography

[1] Frank B Baker and Seock-Ho Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.

[2] Brett W Bader profile imageBrett W. Bader Tamara G. Kolda. Tensor decompositions and applications. *SIAM Review*, 51, 2019.

[3] J Ackerman and I Sandoval-Ballesteros. The global explosion of freedom of information laws. *Administrative Law Review*, 58(1), 2006.

[4] A Roberts. *Blacked out: Government secrecy in the information age*. Cambridge University Press, 2006.

[5] A Florini. *The Right to Know: Transparency for an Open World*. Columbia University Press, 2007.

[6] Central Information Commissioner. CIC Annual Report 2018 - 19, 2018.

[7] A N Tiwari and M M Ansari. Transparency audit of disclosures u/s 4 of the right to information act by the public authorities, 2018.

[8] J G Kelly and B A Simmons. Politics by number: Indicators as social pressure in international relations. *American Journal of Political Science*, pages 55 – 70, 2015. http://dx.doi.org/10.1111/ajps.12119.

[9] Gregory Michener. Policy evaluation via composite indexes: Qualitative lessons from international transparency policy indexes. *World Development*, 74:184–196, 2015.

[10] https://www.rti-rating.org/ accessed: 2020-09-12, 2020.

[11] Avi Arampatzis and Jaap Kamps. A study of query length. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812. ACM, 2008.

[12] Bunker Roy. Right to information: Profile of a grass roots struggle. *Economic and Political Weekly*, pages 1120–1121, 1996.

[13] Madhav Godbole. Right to information: Write the law right. *Economic and Political Weekly*, pages 1423–1428, 2000.

[14] Central Information Commissioner. CIC Annual Report 2007 - 08, 2007.

[15] https://dopt.gov.in/sites/default/files/COMPENDIUM_Final_0.pdf accessed: 2020-01-20, 2011.

[16] The right to information (amendment) bill, 2013. http://www.prsindia.org/uploads/media/RTI%20%28A%29/RTI%20%28A%29%20Bill,%202013.pdf, Accessed: 2020-01-20, 2013.

[17] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[18] Ashwini Kulkarni. Governance and the right to information in maharashtra. *Economic and Political Weekly*, pages 15–17, 2008.

[19] Frederick Noronha. Perils of knowing: Right to information act, 1997. *Economic and Political Weekly*, pages 3021–3023, 2001.

[20] Prabodh Saxena. Public authority and the rti. *Economic and Political Weekly*, pages 13–16, 2009.

[21] Oulac Niranjan. Right to information and the road to heaven. *Economic and political weekly*, pages 4870–4872, 2005.

[22] OP Kejriwal. Loopholes and road ahead. *Economic and Political Weekly*, pages 940–941, 2006.

[23] Madhav Gadgil. Science and the right to information. *Economic and Political Weekly*, pages 1895–1902, 2006.

[24] Editorial. The right to know. *Economic and Political Weekly*, 44(48):6–6, 2009.

[25] Not Available. Auditing the right to information act. *Economic and Political Weekly*, Vol. 43(Issue No. 18), 03 2008.

[26] Gregory Michener. Policy evaluation via composite indexes: Qualitative lessons from international transparency policy indexes. *World Development*, 74:184–196, 2015.

[27] James R Hollyer, B Peter Rosendorff, and James Raymond Vreeland. Measuring transparency. *Political analysis*, 22(4):413–434, 2014.

[28] David Albert Heald. Varieties of transparency. In *Transparency: The Key to Better Governance?: Proceedings of the British Academy 135*, pages 25–43. Oxford University Press, 2006.

[29] Roopinder Oberoi. Institutionalizing transparency and accountability in indian governance: Understanding the impact of right to information. *Journal of Humanities and Social Science*, 11:41–53, 2013.

[30] Tara Vishwanath and Daniel Kaufmann. Towards transparency in finance and governance. *Available at SSRN 258978*, 1999.

[31] Stephanie E Trapnell and Victoria L Lemieux. Right to information: Identifying drivers of effectiveness in implementation. In *Working Paper No. 2. Available at SSRN: https://ssrn.com/abstract=2795127*, 2014.

[32] Laura Neuman. Enforcement models: Content and context. *Communication for Governance and Accountability Programme (CommGAP)*, 2009.

[33] Robert Hazell and Ben Worthy. Assessing the performance of freedom of information. *Government Information Quarterly*, 27(4):352–359, 2010.

[34] Daniel Berliner. The political origins of transparency. *The journal of Politics*, 76(2):479–491, 2014.

[35] Sarah Blodgett Bermeo. Foreign aid and regime change: A role for donor intent. *World Development*, 39(11):2021–2031, 2011.

[36] Gregory Michener. How cabinet size and legislative control shape the strength of transparency laws. *Governance*, 28(1):77–94, 2015.

[37] Jan Seifert, Ruth Carlitz, and Elena Mondo. The open budget index (obi) as a comparative statistical tool. *Journal of Comparative Policy Analysis: Research and Practice*, 15(1):87–101, 2013.

[38] Ben Wasike. Foia in the age of open. gov: An analysis of the performance of the freedom of information act under the obama and bush administrations. *Government information quarterly*, 33(3):417–426, 2016.

[39] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. In *ACm SIGIR Forum*, volume 33, pages 6–12. ACM New York, NY, USA, 1999.

[40] Nina Phan, Peter Bailey, and Ross Wilkinson. Understanding the relationship of information need specificity to search query length. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 709–710, 2007.

[41] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. Mining web query logs to analyze political issues. In *Proceedings of the 4th annual acm web science conference*, pages 330–334, 2012.

[42] Markus Strohmaier, Peter Prettenhofer, and Mark Kröll. Acquiring explicit user goals from search query logs. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 602–605. IEEE, 2008.

[43] Daniel Berliner, Benjamin E Bagozzi, and Brian Palmer-Rubin. What information do citizens want? evidence from one million information requests in mexico. *World Development*, 109:222–235, 2018.

[44] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[45] Steven G Heeringa, Brady T West, and Patricia A Berglund. *Applied survey data analysis*. CRC press, 2017.

[46] Alberto Martini and Ugo Trivellato. The role of survey data in microsimulation models for social policy analysis. *Labour*, 11(1):83–112, 1997.

[47] Frauke Kreuter, Ting Yan, and Roger Tourangeau. Good item or badcan latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3):723–738, 2008.

[48] Barbara G Dodd, William R Koch, and Ralph J De Ayala. Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement*, 13(2):129–143, 1989.

[49] Kathleen Suzanne Johnson Preston, Skye N Parral, Allen W Gottfried, Pamella H Oliver, Adele Eskeles Gottfried, Sirena M Ibrahim, and Danielle Delany. Applying the nominal response model within a longitudinal framework to construct the positive family relationships scale. *Educational and psychological measurement*, 75(6):901–930, 2015.

[50] http://2050-calculator-tool.decc.gov.uk/#/home accessed: 01-aug-2020.

[51] https://www.ecb.europa.eu/ecb/educational/educational-games/economia/html/economia.en.html accessed: 01-aug-2020, 2010.

[52] http://www.gleamviz.org/ accessed: 01-aug-2020, 2011.

[53] Sotirios Koussouris, Fenareti Lampathaki, Panagiotis Kokkinakos, Dimitrios Askounis, and Gianluca Misuraca. Accelerating policy making 2.0: Innovation directions and research perspectives as distilled from four standout cases. *Government Information Quarterly*, 32(2):142–153, 2015.

[54] Andy Hon Wai Chun. An ai framework for the automatic assessment of e-government forms. *AI Magazine*, 29(1):52–52, 2008.

[55] Ashkan Nabavi-Pelesaraei, Shahin Rafiee, Homa Hosseinzadeh-Bandbafha, and Shahaboddin Shamshirband. Modeling energy consumption and greenhouse gas emissions for kiwifruit production using artificial neural networks. *Journal of Cleaner Production*, 133:924–931, 2016.

[56] Hong Zhang, Jie Song, Chao Su, and Mengying He. Human attitudes in environmental management: Fuzzy cognitive maps and policy option simulations analysis for a coal-mine ecosystem in china. *Journal of environmental management*, 115:227–234, 2013.

[57] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[58] Staša Milojević. Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology*, 61(12):2417–2425, 2010.

[59] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[60] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[61] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM computing surveys (CSUR)*, 38(1):2, 2006.

[62] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[63] Francesco Laio. Cramer–von mises and anderson-darling goodness of fit tests for extreme value distributions with unknown parameters. *Water Resources Research*, 40(9), 2004.

[64] Steven M Beitzel, Eric C Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–328. ACM, 2004.

[65] Casper Petersen, Jakob Grue Simonsen, and Christina Lioma. Power law distributions in information retrieval. *ACM Transactions on Information Systems (TOIS)*, 34(2):8, 2016.

[66] Ronald B Mitchell. Sources of transparency: Information systems in international regimes. *International Studies Quarterly*, 42(1):109–130, 1998.

[67] Mr George Kopits and Mr JD Craig. *Transparency in government operations*. International monetary fund, 1998. 158.

[68] Deirdre Curtin and Albert Jacob Meijer. Does transparency strengthen legitimacy? *Information polity*, 11(2):109–122, 2006.

[69] Victoria Louise Lemieux and Stephanie E Trapnell. Public access to information for development: a guide to effective implementation of right to information laws. *Directions in Development. Washington, D.C*, 2016.

[70] Michael W Mahoney, Mauro Maggioni, and Petros Drineas. Tensor-cur decompositions for tensor-based data. *SIAM Journal on Matrix Analysis and Applications*, 30(3):957–987, 2008.

[71] Evrim Acar, Seyit A Çamtepe, Mukkai S Krishnamoorthy, and Bülent Yener. Modeling and multiway analysis of chatroom tensors. In *International Conference on Intelligence and Security Informatics*, pages 256–268. Springer, 2005.

[72] Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. Cubesvd: a novel approach to personalized web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 382–390. ACM, 2005.

[73] M Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. *Computer VisionECCV 2002*, pages 447–460, 2002.

[74] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, pages pnas–0803205106, 2009.

[75] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Hindustan Book Agency, 2015.

[76] R. Penrose. A generalized inverse for matrices. *Mathematical proceedings of the Cambridge philosophical society*, 51, 1955.

[77] Package 'ltm', 2018. https://cran.r-project.org/web/packages/ltm/ltm.pdf Accessed: 2020-01-20.

[78] Kiri Wagstaff. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.

[79] Fernando Martnez-Plumed, Ricardo B.C. Prudncio, Adolfo Martnez-Us, and Jos Hernndez-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019.

[80] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.