Analyzing the Gender Wage-gap in Ontario’s Public Sector

by

Andrew D’Angelo

A Thesis
Presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Computer Science

Guelph, Ontario, Canada
© Andrew D’Angelo, May, 2016
ABSTRACT

Analyzing the Gender Wage-gap in Ontario’s Public Sector

Andrew D’Angelo
University of Guelph, 2016

Advisors:
Dr. Gary Gréwal
Dr. Luiza Antonie

The disparity in wages between men and women is a well known fact; however, the contribution of each known factor is not fully understood. Leveraging the salary information provided by the Ontario Ministry of Finance could allow for a better understanding of the factors that contribute to gender wage disparity. The Ontario public salary data, also known as the ‘Sunshine List’, contains the salary information of individuals working in the public sector that earn $100,000 or more annually. Unfortunately, the Sunshine List data is not in a form that allows for direct analysis. The information must first be collected, cleaned, and standardized due to formatting inconsistencies within the Sunshine List data. Furthermore, although these salaries are publicly available, a key attribute is missing from the public data, the gender variable. A novel hybrid model is proposed to predict the gender based on an individual’s given name, and the original database is augmented with the new gender variable. With the new database created, the wage-gap is analyzed and the results are presented and discussed. The findings of this research are being used by Ontario’s provincial government to reassess and change current policies for pay equity.
Acknowledgments

I would first and foremost like to thank my advisors, Dr. Gary Gréwal and Dr. Luiza Antonie, for giving me the opportunity to work with them. This experience has opened many doors for me that I otherwise would have never discovered. I am very grateful for your endless support and encouragement.

I would also like to acknowledge Dr. Miana Plesca, for providing insightful knowledge throughout my graduate studies.

I would like to thank my family, who have provided me with endless love and support through all of my endeavours. Mom and dad, 10 years of school is a very long time, but your encouragement has allowed me to be where I am today. To the greatest siblings anyone could ask for; Sarah, Paul, Maria, Pierre and Issac, you have all taught me so much and I would not be who I am without each and every one of you.

Lastly, but not least, thank you to my wife, Danielle, for your continuous love, support, patience, and understanding. Without you this dissertation would not have been possible, and I can not thank you enough.
# Table of Contents

List of Tables vi

List of Figures vii

1 Introduction 1
   1.1 Motivation 1
   1.2 Thesis Objective 2
   1.3 Research Contributions 2
   1.4 Thesis Organization 3

2 Background 4
   2.1 Gender Wage-gap 4
      2.1.1 Current State of the Gender Wage-gap 4
   2.2 Gender Inference 7
      2.2.1 Gender Inference from a Given Name 7
   2.3 Longitudinal Data 11
      2.3.1 Longitudinal Linkage Techniques 11
   2.4 Summary 13

3 Data Collection and Cleaning 15
   3.1 Sunshine List Data 15
   3.2 Data Collection 17
   3.3 Data Cleaning 22
      3.3.1 Data Preparation 23
      3.3.2 Employer Name Changes 24
      3.3.3 Language Inconsistencies 25
      3.3.4 Special Characters 26
   3.4 Longitudinal Data 27
      3.4.1 Naive Approach 28
      3.4.2 Weighted Probability Approach 33
   3.5 Experimental Methodology 33
   3.6 Summary 35

4 Gender Inference 36
   4.1 Relative Name Frequencies 36
      4.1.1 Relative Name Frequencies Model Performance 41
   4.2 Heuristic Model 42
4.2.1 Heuristic Model Performance ........................................ 45
4.3 Hybrid Model ............................................................... 46
  4.3.1 Hybrid Model Performance ........................................ 47
4.4 Enhanced Database ....................................................... 48
4.5 Summary ................................................................. 49

5 Gender Wage-gap Analysis ................................................. 50
  5.1 Gender Shares in the Sunshine Data List Across Time ............ 50
  5.2 Gender Wage-gap in the Sunshine List .............................. 54
  5.3 Quantiles of the Income Distribution ............................... 57
  5.4 Summary ............................................................... 58

6 Conclusion and Future Work .............................................. 60
  6.1 Future Work ............................................................ 61

Bibliography ........................................................................ 62

A Weighted Probability Approach ......................................... 67

B Given Names Classified as ‘Unknown’ ................................... 73

C Gender Inference Summary Statistics ................................. 74
List of Tables

3.1 Sunshine list data summary [10]. ................................................. 16
3.2 Scraped data from the Crown sector in 1999. ............................... 20
3.3 Sunshine List data for a single real-world entity over 18-year span. . . 28
3.4 Two different real-world entities in the Sunshine List data. ............. 30
3.5 First stage results of naïve approach. ........................................... 30
3.6 Second stage results of naïve approach. ....................................... 32
3.7 Performance on the verified Guelph data. .................................... 34

4.1 Performance of relative name frequency model. ............................. 42
4.2 Performance of heuristic model. .................................................. 46
4.3 Performance of heuristic model on unknown University of Guelph names. 47
4.4 Combined performance of both gender models. .............................. 48
4.5 Subsection of the enhanced database. .......................................... 48

A.1 Sample records. ................................................................. 67

B.1 Unknown given names. ......................................................... 73

C.1 Gender inference summary statistics: All records ......................... 74
C.2 Gender inference summary statistics: Unique given names ............ 74
List of Figures

2.1 2015 wage equality leaders vs. G8 countries. .......................... 5
2.2 2014 provincial private sector gender wage-gap by province for new graduates. ......................................................... 6
3.1 The 2015 disclosure year table of contents. .............................. 17
3.2 1997 disclosure year showing 7 attributes, including sector. .... 18
3.3 1998 disclosure year showing 6 attributes, without sector. ..... 18
3.4 Ontario Public Service sector in 2003. ........................................ 19
3.5 Ministries sector in 2004. .......................................................... 19
3.6 Professional designation in given name field. ....................... 23
3.7 Employer name changes a) 2005 disclosure; b) 2006 disclosure; c) 2007 disclosure. .............................................................. 24
3.8 Language provided by the University of Ottawa a) 2012 disclosure; b) 2014 disclosure. ....................................................... 26
3.9 HTML source code accents. ....................................................... 27
4.1 Screenshot of the 2014 US Social Security data. ...................... 37
4.2 Screenshot of the file containing total male and female frequencies. .. 39
4.3 Hybrid model process flow. ....................................................... 46
5.1 Number of records in the Sunshine List. ................................. 51
5.2 Proportion of women relative to men on the Sunshine List. ........ 52
5.3 Percentage of women on the Sunshine List, by sector, nominal threshold. 53
5.4 Percentage of women on the Sunshine List, by sector, real threshold. . 53
5.5 Income differential, nominal threshold. ................................. 54
5.6 Income differential, real threshold. ............................................. 55
5.7 Income differential by sector, nominal threshold. .................... 56
5.8 Income differential by sector, real threshold. .......................... 56
5.9 Gender gap at quantiles of earnings distribution, nominal threshold. . 57
5.10 Gender gap at quantiles of earnings distribution, real threshold. . . 58
A.1 Full and partial life spans for the affiliation attribute. .......... 68
Chapter 1

Introduction

The *gender wage-gap* is defined as the earnings gap between men and women expressed as the percentage difference between male and female salaries on a per dollar basis [1]. While it is well known that a gender wage-gap exists globally [2], as well as in the province of Ontario [1], the underlying cause of this disparity is not known. To gain a better understanding of the role gender plays in wage disparity an interdisciplinary approach is taken to analyze Ontario’s public sector. The approach is interdisciplinary in the sense that methods from data mining and machine learning, including data collection, data cleaning and gender inference, are employed along side traditional economic analysis. This analysis will provide a better understanding of the causes of gender wage disparity and will allow policy makers and legislators to develop effective means for reducing or eliminating the gap.

1.1 Motivation

Over recent decades, the career expectations of Canadian women have evolved considerably. University enrolment statistics now show that female students account for approximately two thirds of the student body across Canada [3]. Labor force participation rates for women, and the number of hours worked by women, are also edging closer to those of men, especially in positions requiring higher levels of education [4]. However, despite these trends, a real gender wage-gap still persists in the Province of Ontario, Canada. The current wage-gap is 26%, meaning that on average for every $1.00 a male earns a female earns $0.74 for equivalent work [1].
While legislation [5] has been developed to try and mitigate or eliminate the
gender wage-gap, several studies have shown that a gender wage-gap continues to
persist in Ontario [1, 6, 7, 8]. The *Ontario Pay Equity Act* is arguably the most
extensive equal pay legislation in the world; however, since its passage in 1989, the
gender wage-gap has not been significantly reduced [7]. Moreover, the work in [9]
reveals that the gender wage-gap across income distribution in Canada’s public sector
increases for the high earners (top 20%). Fortunately, the salary information for
the top earners in Ontario’s public sector is made publicly available by the Ontario
Ministry of Finance, in the form of the ‘Sunshine List’.

1.2 Thesis Objective

In this thesis, we present the limitations of the information provided in the
Sunshine List; most notably, the gender of each individual is not provided. We
develop techniques to collect, clean and standardize the Sunshine List data and store
the information in a database. We also develop algorithms to identify links between
records that map to the same real-world entity for the purposes of correcting and
updating inaccurate or missing information. Furthermore, to prepare the data for
the gender wage-gap analysis we propose a novel hybrid gender inference model to
predict the gender of each individual with a high level of accuracy. Finally, we aim to
gain a better understanding of the current state of gender wage disparity in Ontario
by analyzing the gender wage-gap in Ontario’s public sector.

1.3 Research Contributions

In this section, the research contributions from the work performed as part of
this thesis are presented. These include:

- The creation of a database containing Sunshine List data that is cleaned, stan-
standardized and enhanced with a new gender variable. This database will be made publicly available to all researchers;

- Longitudinal links, which track the same record at different points in time, were created for 91% of the records;

- A novel hybrid gender predictor model is proposed, and is shown to infer gender with an overall accuracy rate of almost 90%; and,

- The first analysis of the gender wage-gap based on the Sunshine List data is performed.

1.4 Thesis Organization

The remainder of this thesis is organized as follows. In Chapter 2, the necessary background information is introduced. Chapter 3 provides a detailed overview of the Sunshine List data, the limitations of this data, and the steps taken to collect, clean, standardize and update missing or inaccurate information. In Chapter 4, the proposed hybrid gender inference model is introduced, analyzed, and compared against gender inference models found in literature. Chapter 5 presents the gender wage-gap analysis, and Chapter 6 concludes the thesis with closing remarks, as well as future research directions. Papers related to this thesis include [10].
Chapter 2

Background

In this chapter, the current state of the gender wage-gap in Canada and abroad is introduced, as well as the motivation for specifically analyzing Ontario’s public sector (Sunshine List). Due to the limitations of the Sunshine List data, as gender is not provided, a review of gender inference models that can be used to enhance the Sunshine List data with gender are reviewed. Moreover, techniques to track data points over time, longitudinal data, are discussed as a means of updating and correcting incomplete information.

2.1 Gender Wage-gap

The gender wage-gap is defined as the earnings gap between men and women for equivalent work and is expressed as the percentage difference between male and female salaries on a per dollar basis [1]. This section presents the current state of the gender wage-gap globally, in Canada, and in Ontario, as well as the known contributing factors.

2.1.1 Current State of the Gender Wage-gap

Since 2006, the gender wage-gap has been studied annually at an international level by the World Economic Forum (WEF). The findings of each study are released annually in the form of the Global Gender Gap Report [2]. This report ranks the top 145 countries using an overall global index score, which is comprised of four subindexes: Economic Participation and Opportunity, Educational Attainment,
Health and Survival, and Political Empowerment. Each subindex is scored using a variety of indicators, one of which is wage equality. The wage equality indicator is calculated through the Executive Opinion Survey (EOS), which is conducted by the WEF [11]. The EOS survey conducted by the WEF involves 13,000 executives in 144 countries and captures invaluable information on a broad range of socio-economic issues [11]. The wage equality indicator for each country is used to determine the difference in earnings between men and women for equal work (i.e., the gender wage-gap). From the 2014 Global Gender Gap Report, the findings show that the average gender wage-gap for the top 145 countries is 0.65 [2]. The 10 year-span of the Global Gender Gap Report reveals an overall decreasing trend in the gender wage-gap [2]. With that being said, the gender wage-gap is an increasing concern for many developed countries, such as Canada (0.72), the United Kingdom (0.69), Japan (0.68), and the United States of American (0.66) [2]. A comparison of the 2014 wage equality between the Group of Eight (G8) (striped bars) and the global wage equality leaders (solid black bars) is presented in Fig. 2.1.

![Figure 2.1: 2015 wage equality leaders vs. G8 countries.](image)

The statistics in Fig. 2.1 reveal that there is a significant gap between the
G8 countries and the gender wage equality leaders. While the gender wage-gap as a percent provides a standard measure of comparison, the gender wage-gap in real dollars can be a more telling indicator for the wage equality of a country. For example, Canada’s percentage gender wage-gap is 7 percentiles above the international gender wage-gap average; however, the gender wage-gap in real dollars is twice the global average [12]. The international gender wage disparity average, in terms of real dollars, is $4,000 annually, whereas in Canada the disparity is just over $8,000 annually [13].

The gender wage-gap in Canada, as reported by the WEF, is 0.72 (all workers) which is further confirmed by Statistics Canada and Canadian Women’s Foundation[14]. In contrast, Catalyst Canada reports a gender wage-gap of 0.82 for full time workers, and a combined gender wage-gap (full-time and part-time workers) of 0.753 [12]. A study performed by the Laurier Centre for Economic Research & Policy Analysis identifies the gender wage-gap by province for recent graduates (age 25-29) in the private sector [15]. The findings for each province (Ontario is highlighted with striped bars) are presented in Fig. 2.2.

Figure 2.2: 2014 provincial private sector gender wage-gap by province for new graduates.
From Fig. 2.2, the average gender wage-gap among recent graduates in the private sector is 0.79, and the gender wage-gap among recent graduates in the private sector in Ontario is 0.83. These statistics are confirmed by the work in [6], which show that for new graduates, women on average earn 6-14% less than males during the period of two to fives years after graduation [6]. The gender wage-gap has further been studied in [9], which compares Canada’s and Ontario’s private and public sectors. The findings in [9] reveal that Canada’s and Ontario’s public sectors experience an increase in the gender wage-gap for the top 20% of earners, and the individuals in Ontario’s public sector that comprise the top 20% are represented on the Sunshine List. *The cause of this increase is not known; however, an analysis of the gender wage-gap of Ontario’s public sector can offer insight.* Unfortunately, while the salary information for the high earners in Ontario’s public sector are publicly available, the gender information for each individual is not provided and, therefore, the gender for each individual must be inferred from a record’s given name attribute.

### 2.2 Gender Inference

Gender inference is the process of predicting gender using various features. For example, gender prediction models have been developed to use features extracted from an individual’s iris [16], speech [17], web-activity [18, 19] and even from their clothing and facial cues [20]. However, in the absence of these and other types of high-level features, an individual’s gender can be inferred from a single, fundamental attribute, their given name.

#### 2.2.1 Gender Inference from a Given Name

Gender inference models that predict the gender of an individual based solely on given name fall into two categories: i) models that utilize *relative name frequencies*, and ii) models that use machine learning and various classification techniques. The
work in [19, 21, 22] utilize *relative name frequencies* and employ a historical list of given names labelled male or female and a frequency of occurrence within a population. In other words, each record in the historical list contains a given name, a corresponding gender (i.e., ‘M’ for male, and ‘F’ for female), and a count (frequency) of the number of times the given name-gender pair occur in a population. These techniques use both the male and female frequencies of a given name to assign a gender probability, as shown in Equations 2.1 and 2.2.

\[ P(\text{Male}|\text{Name}) = \frac{\text{MaleFrequency}(\text{Name})}{\text{MaleFrequency}(\text{Name}) + \text{FemaleFrequency}(\text{Name})} \] (2.1)

\[ P(\text{Female}|\text{Name}) = \frac{\text{FemaleFrequency}(\text{Name})}{\text{MaleFrequency}(\text{Name}) + \text{FemaleFrequency}(\text{Name})} \] (2.2)

The advantage of the *relative name frequencies* approach is its simplicity and low computational demand. However, despite its simplicity, this approach has been shown to perform effectively in a variety of contexts. A limitation of this approach is that given names that do not appear in the historical data cannot be assigned a gender probability. In contrast, models that utilize machine learning and classification [23, 24] can overcome this limitation and assign a gender in all cases.

Machine learning is a technique that gives computers (algorithms) the ability to learn without being explicitly programmed [25]. These algorithms contain two stages: training and classification. The training process for each of the algorithms investigated in this thesis relies on *supervised learning*, which involves processing *labelled data* (i.e., data that contains an input and its expected result) to learn patterns or features that correspond to an expected result (output) [26]. In other words, supervised learning is learning by example. For instance, the training data for an
algorithm designed to detect faces in photographs would contain an image and its associated content (e.g., [image1, ‘face’] or [image2, ‘tree’]) [27]. Once the model has been trained, the classification stage is employed to predict the expected results of unlabelled data, which is data that only contains inputs, using the features extracted during the training stage [28]. There are various supervised classification models that can be utilized; however, these models can be categorized into two groups: parametric and non-parametric [29]. Parametric classifiers involve algorithms that contain a fixed number of parameters, whereas non-parametric algorithms use a flexible number of parameters that may grow in number as the model learns from more data [30]. Moreover, parametric classifiers assume that the data has a specific structure (i.e., normal distribution) or that the data is linearly separable, while non-parametric classifiers make no assumptions about the data [31].

Within the category of parametric classifiers, the simplest to implement and least computationally demanding are probabilistic models, such as Naïve Bayes, which uses conditional probability to predict the expected result of unlabelled data [32]. For example, the work in [23, 24] attempt to exploit the subtle morphological and phonetical differences in structure between male and female names using n-grams which are n-character length sequences \((n = 1, 2 \text{ or } 3)\) that appear in a given name. For instance, for the given name ‘Paul’ the model would generate the following n-grams.

\[
\text{Paul} = \{p, a, u, l, pa, au, ul, pau, aul\}
\]

This model calculates the conditional probability of a given name being male or female given that specific n-grams appear in the given name. The advantage of this model is the simplicity of implementation and interpretation of results and the speed at which convergence is reached, which means less training data is required [33]. A disadvantage is that probabilistic models make strong assumptions about data, most notably, that the expected results of one input are independent of another.
which means that these types of classifiers cannot learn the interaction between features. The major issue with all probabilistic models is that if the initial assumptions made on the data are incorrect, the performance suffers greatly [31]. In contrast, Support Vector Machines (SVM) can map the data to a high-dimensional feature space, meaning that feature interaction and non-linearity in the data can be accounted for [34]. The tradeoff, however, is high computational demand and a more memory intensive model [35]. In addition, the output of an SVM may be more difficult to interpret and the initial tuning parameters may be difficult to optimize [35].

Recall that non-parametric models do not make any assumptions on the data and, therefore, the models are not limited to data that is linearly separable [35]. Models such as decision trees have been employed extensively in part-of-speech tagging, which build trees of $n$-grams to classify various parts of speech, such as nouns, adjectives, verbs, and adverbs [36]. The structure of a decision tree allows for fast and scalable feature extraction, including feature interactions; however, it prevents the support for online learning, which means that the tree must be rebuilt when new examples are to be used [35]. Another disadvantage of decision trees is that they are susceptible to overfitting, meaning that they will not generalize well to new data [35]. Conversely, decision forests, which are a collection of multiple decisions trees within a single model, have been developed to prevent overfitting [37]. While these models are more computationally expensive than single decision trees, variations of decisions forest have been developed [38] to maintain a relatively quick (hours) and scalable training and classification run times.

In regards to inferring the gender for an individual on the Sunshine List, the major benefit of using a classification technique, as opposed to relative name frequencies, is that a gender can be assigned in all cases. However, the performance of a classification model is highly dependent on the contents of the training data [23]. Furthermore, each gender inference model, regardless of methodology, rely on
the assumption that a full given name has been provided. Because the Sunshine List data spans 18 years, an individual can appear on the Sunshine List for multiple years. In the cases where given names are incomplete (i.e., only an individual’s initials have been provided), longitudinal data for an individual may be used to update and complete this information.

2.3 Longitudinal Data

Longitudinal data, also referred to as panel data, is the tracking of an entity at different points in time [39]. The advantages of longitudinal data is that this allows for the assessment of an entity over time. For example, in the case of an individual, longitudinal data can track the changes in various attributes of an individual such as, given name, employer, and salary.

2.3.1 Longitudinal Linkage Techniques

In general, longitudinal linkage techniques are designed to create links between samples and group them into a single cluster, where a cluster represents all the samples that map to the same entity at different points in time. These models consist of two major components: one that addresses longitudinal disagreement and another for longitudinal agreement [40]. The work in [41] defines longitudinal disagreement as the probability that a sample changes its attribute value within a time interval $\Delta t$, whereas longitudinal agreement is the probability that two different samples share the same attribute within a time interval $\Delta t$ [41]. These values are determined for each attribute field by processing training data and are used to calculate a similarity score between a sample and an already created cluster. If the similarity score between a sample and a cluster is sufficiently high, the sample is inserted into the cluster; if not, a new cluster is created for the sample. The work in [41] defines three approaches: i) make eager cluster decisions in increasing time order, ii) make clustering decisions at the
end utilizing all evidence, and iii) perform multiple phases of clustering and compare all samples regardless of occurrence in time. The advantage of this model is most evident in the accuracy performance of the third approach; albeit, at the expense of high computational demand.

In contrast, the work in [42] argues that using only a temporal model may not produce an acceptable matching result and, therefore, employs a two-stage approach to link temporal records. The first stage is to collect evidence assuming that the attributes within a sample do not change by clustering samples based purely on exact matches and not using a temporal model. The second stage of the model further clusters the original groupings found in stage one by determining if it is possible for an entity to evolve from one cluster to another [42]. The benefit of the two-stage approach is that both the static evidence and dynamic evidence are considered when attempting to cluster samples; however, at high computational cost. Other works have introduced the concept of mutation, which is the probability that an attribute value will reappear after the time interval $\Delta t$ [40]. The idea behind mutation is that a change in an attribute of a sample may be dependent on its past values [40]. The advantage that this model presents in comparison to the models in [41], which make clustering using pairs of samples, is that the mutation model makes clustering decisions based on the entire history of a sample; or, in other words, compares sets of samples as opposed to pairs of samples [40]. Comparing the full history of a sample can increase the accuracy of clustering decisions; however, as the volume of history of a sample increases, so to does the computational demand.

The work in [43] has revealed that algorithms that employ longitudinal agreement and disagreement or mutation are limited as they do not consider the complex value transitions that may occur for an attribute over time. The mutation model fails to capture the fact that different values for an attribute could have different reappearance probabilities, and neither can discriminate between the various values that an attribute may change to [43]. The work in [43] focusses on an attributes tran-
sition probability, which answers the question “if an attribute has the value \( v \), what is the probability for this value to be \( \hat{v} \) after the time interval \( \Delta t \)?” Furthermore, the algorithm considers the characteristics of the data source, including whether or not the data is up to date, by expanding on the work in [44] and [45]. The model in [43] iteratively links a cluster to the sample based on the transition probability, the support for the data sources, and a more fine-grained cluster signature that describes the states of each attribute; however, by doing so, the model becomes more complex than previously discussed models.

The creation of longitudinal links can create a history of the attribute values for a sample, which means that if information is missing or incomplete at various points in time for a specific sample, the longitudinal links can be used to correct or complete this information. In regards to the Sunshine List data, longitudinal links may be used to update the given name field to allow for gender inference.

2.4 Summary

In this chapter, we have shown that the gender wage-gap is a global issue and that although the gender wage-gap in Canada, as a percent, is above the international average, the real gender wage-gap is twice the global average. Moreover, the gender wage-gap in Ontario’s public sector was shown to increase for high earners, the cause of which is unknown; however, an analysis of the gender wage-gap on Sunshine List can offer insight. Gender inference models were investigated to overcome the fact that gender is not provided in the Sunshine List data. Furthermore, the Sunshine List data contains incomplete information and, therefore, algorithms to create longitudinal links were presented as a solution to this problem.

In this thesis, we work towards the creation of a database containing clean, correct and enhanced (with gender) Sunshine List data. To do so we propose a novel hybrid gender inference model that combines the relative name frequency and
classification techniques into a single model. In addition, we develop an algorithm based on longitudinal linkage techniques to update incorrect or missing information within the dataset. Finally, we work towards a better understanding of the factors impacting the gender wage-gap in Ontario’s public sector.
Chapter 3

Data Collection and Cleaning

Although the Sunshine List data is publicly available, the data is not in a form that allows for direct analysis. In particular, the data is spread over 18 disclosure years, each of which is located at a unique URL. For each disclosure year, the information is further divided into a varying number of subsections or sectors, each of which are available in both HTML and PDF format. The first main contribution of this work is to i) collect and clean the Sunshine List data to create a database to store the information in a single location, ii) standardize the format of each field for consistency, and iii) update missing or incorrect information. This newly created database will be made available to all researchers who may wish to perform work related to the gender wage-gap in Ontario’s public sector.

The remainder of this Chapter is organized as follows: Section 3.1 discusses the contents of the Sunshine List data and the limitations of this data with regards to analyzing the gender wage-gap. Sections 3.2 and 3.3 discuss the techniques employed for collecting and cleaning the Sunshine List data, respectively. Section 3.4 discusses the methods used to update incorrect or missing values in the Sunshine List data.

3.1 Sunshine List Data

The Sunshine list data is made publicly available annually by the Ontario Ministry of Finance in the form of the Yearly Public Sector Salary Disclosure [46]. This list was first introduced in 1996 with the passing of the Public Salary Discloser Act [46] in an attempt to make Ontario’s public sector more transparent and accountable
to taxpayers. Because the act is not retro-active, the available data dates back only to 1996. It is important to note that the Sunshine List provides information for salaries earned in the previous year, meaning that the 2010 disclosure provides information for salaries earned in 2009.

The Sunshine List data contains 652,804 unique records spanning a period of 18 years, 1997 to 2014 inclusive, and is divided into 11 sectors: college, crown agencies (crown), hospitals and boards of public health (hospitals), hydro one and Ontario power generation (electric), judiciary, legislative assembly, ministries, municipalities, school boards, university, and other public service sector employers. Within each sector, the following attributes are provided for each record: employer, surname, given name, position and department, salary paid, and taxable benefits. In order for an individual to appear on the Sunshine List, they must be working in the public sector and earn an annual salary of $100,000 or more. The information contained in the Sunshine List data is summarized in Table 3.1.

<table>
<thead>
<tr>
<th>Records</th>
<th>652,804</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sectors</td>
<td>college, crown agencies (crown), hospitals and boards of public health (hospitals), hydro one and Ontario power generation (electric), judiciary, legislative assembly, ministries, municipalities, school boards, university, and other public service sector employers</td>
</tr>
<tr>
<td>Years</td>
<td>1997 to 2014</td>
</tr>
<tr>
<td>Variables</td>
<td>employer, position, surname, given name, salary, taxable benefits</td>
</tr>
</tbody>
</table>

While the information contained in the Sunshine List data is useful when assess-
ing the wage-gap as a whole, a key attribute, gender, is not provided. To prepare the data for the gender wage-gap analysis in Chapter 5, the gender for each record must first be inferred using an individual’s given name (Chapter 4).

3.2 Data Collection

The Sunshine List data is provided by the Ontario Ministry of Finance and is publicly available at the Ontario Ministry of Finance’s website\(^1\). The landing page for each disclosure year acts as a table of contents with the information being separated by sector, and available in both HTML and PDF formats. A snapshot of the landing page for the 2015 disclosure year is shown in Fig. 3.1.

![Public Sector Salary Disclosure 2015 (Disclosure for 2014)](image)

- Introduction [HTML], [PDF]
- Ministries [HTML], [PDF]
  - Individuals Seconded to Ministries from Public Sector Organizations [HTML], [PDF]
- Legislative Assembly and Offices [HTML], [PDF]
- Judiciary [HTML], [PDF]
- Crown Agencies [HTML], [PDF]
- Hydro One and Ontario Power Generation [HTML], [PDF]
- Municipalities and Services [HTML], [PDF]
- School Boards [HTML], [PDF]
- Universities [HTML], [PDF]
- Colleges [HTML], [PDF]
- Hospitals and Boards of Public Health [HTML], [PDF]
- Other Public Sector Employers [HTML], [PDF]

Figure 3.1: The 2015 disclosure year table of contents.

The first step performed when collecting the Sunshine List data was to manually inspect the landing page and HTML source code for each disclosure year to identify

\(^1\)Ontario Ministry of Finance URL: https://www.ontario.ca/page/public-sector-salary-disclosure
changes or formatting inconsistencies that may impact the scraping process. Several changes were identified, such as an increase in the number of subsections or sectors, a reduction in the number of attributes provided for each record, and the introduction of paged HTML source files. The first of these changes occurred in 1998, when the number of attributes provided for each record decreased from 7 to 6. This difference can be seen in Figs. 3.2 and 3.3, which show the data provided in the 1997 and the 1998 disclosures, respectively. Notice that the ‘HOSPITALS’ attribute is not present in the 1998 disclosure.

### Hospitals

**Public Sector Salary Disclosure 1997 (Disclosure for 1996) : Hospitals**

<table>
<thead>
<tr>
<th>Sector</th>
<th>Employer</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
<th>Salary Paid</th>
<th>Taxable Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOSPITALS</td>
<td>Ajax and Pickering General Hospital</td>
<td>Cliff</td>
<td>Bruce</td>
<td>President &amp; CEO</td>
<td>$110,309.00</td>
<td>$3,157.00</td>
</tr>
<tr>
<td>HOSPITALS</td>
<td>Belleville General Hospital</td>
<td>Cassidy</td>
<td>Patrick</td>
<td>Chief Pathologist</td>
<td>$102,694.00</td>
<td>$477.00</td>
</tr>
<tr>
<td>HOSPITALS</td>
<td>Belleville General Hospital</td>
<td>Gaber</td>
<td>M. L.</td>
<td>Associate Pathologist I</td>
<td>$124,960.00</td>
<td>$741.00</td>
</tr>
<tr>
<td>HOSPITALS</td>
<td>Belleville General Hospital</td>
<td>Steinberg</td>
<td>Brian</td>
<td>President</td>
<td>$122,130.00</td>
<td>$4,925.00</td>
</tr>
<tr>
<td>HOSPITALS</td>
<td>Belleville General Hospital</td>
<td>Twemlow</td>
<td>Greg D.</td>
<td>Associate Pathologist II</td>
<td>$119,905.00</td>
<td>$712.00</td>
</tr>
<tr>
<td>HOSPITALS</td>
<td>Bloorview MacMillan Centre</td>
<td>Curtis</td>
<td>Rosalind</td>
<td>Chief of Staff, Bloorview Site</td>
<td>$151,500.40</td>
<td>$505.56</td>
</tr>
</tbody>
</table>

Figure 3.2: 1997 disclosure year showing 7 attributes, including sector.

### Hospitals

**Public Sector Salary Disclosure 1998 (Disclosure for 1997) : Hospitals**

<table>
<thead>
<tr>
<th>Employer</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
<th>Salary Paid</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>Ade</td>
<td>Clare</td>
<td>V.P. Professional Services</td>
<td>$107,996.21</td>
<td>$2,136.84</td>
</tr>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>Gordon</td>
<td>Michael</td>
<td>V.P. Medical Services</td>
<td>$110,386.36</td>
<td>$1,434.42</td>
</tr>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>GreenGarten</td>
<td>Martin</td>
<td>V.P. Public &amp; Community Affairs</td>
<td>$128,653.96</td>
<td>$1,264.56</td>
</tr>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>Harrison</td>
<td>Laurie</td>
<td>V.P. Finance</td>
<td>$144,093.16</td>
<td>$1,242.56</td>
</tr>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>Herbert</td>
<td>Stephen</td>
<td>President &amp; CEO</td>
<td>$2,27,000.84</td>
<td>$21,947.49</td>
</tr>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>Lambert</td>
<td>Victor</td>
<td>V.P. Information &amp; Support Services</td>
<td>$131,147.36</td>
<td>$1,125.51</td>
</tr>
<tr>
<td>Baycrest Centre for Geriatric Care</td>
<td>MacAdam</td>
<td>Margaret</td>
<td>Senior V.P. &amp; V.P., Social Services</td>
<td>$146,903.48</td>
<td>$1,280.26</td>
</tr>
</tbody>
</table>

Figure 3.3: 1998 disclosure year showing 6 attributes, without sector.

After 1997, the sector attribute was no longer provided within the HTML source code and, therefore, had to be obtained from the URL for each sector using regular
expressions.

The second change occurred in 2000 with the introduction of the Electric sector, which increased the number of sectors from 8 to 9. Prior to 2000, the employers in this sector were part of the Crown sector, meaning that for years prior to 2000 the sector attribute for these employers was simply ‘Crown’. From 2000 onwards the sector attribute became ‘Electric’. Similarly, in 2004 the Judiciary sector was introduced, and the the Ontario Public Service sector was replaced with the Legislative Assembly and Ministries sectors bringing the total number of sectors to 11. An example of this can be seen in Figs. 3.4 and 3.5, which show the employer Agriculture and Food in the Ontario Public Service sector in 2003 and the Ministries sector in 2004, respectively.

**Salary Disclosure 2003 (Disclosure for 2002)**
Ontario Public Service

<table>
<thead>
<tr>
<th>Employer / Employer</th>
<th>Surname / Nom de famille</th>
<th>Given Name / Prénom</th>
<th>Position / Poste</th>
<th>Salary Paid / Traitement</th>
<th>Taxable Benefits / Avant. impos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture and Food</td>
<td>ANTE</td>
<td>DAVID</td>
<td>Director, Strategic Policy Branch</td>
<td>$123,629.97</td>
<td>$223.52</td>
</tr>
<tr>
<td>Agriculture and Food</td>
<td>ARCHIBALD</td>
<td>BRUCE</td>
<td>ADM, Policy &amp; Farm Finance Division</td>
<td>$149,108.65</td>
<td>$290.28</td>
</tr>
<tr>
<td>Agriculture and Food</td>
<td>BAKER</td>
<td>THOMAS</td>
<td>Director, Food Inspection Branch</td>
<td>$102,295.52</td>
<td>$211.08</td>
</tr>
<tr>
<td>Agriculture and Food</td>
<td>CHAPMAN</td>
<td>J. DOUGLAS</td>
<td>Director, Food Industry Competitiveness</td>
<td>$112,381.67</td>
<td>$223.52</td>
</tr>
<tr>
<td>Agriculture and Food</td>
<td>CLARKE</td>
<td>DAVID</td>
<td>Director, Market Development Branch</td>
<td>$109,331.03</td>
<td>$223.52</td>
</tr>
</tbody>
</table>

Figure 3.4: Ontario Public Service sector in 2003.

**Salary Disclosure 2004 (Disclosure for 2003)**
Government of Ontario: Ministries

<table>
<thead>
<tr>
<th>Employer / Employer</th>
<th>Surname / Nom de famille</th>
<th>Given Name / Prénom</th>
<th>Position / Poste</th>
<th>Salary Paid / Traitement</th>
<th>Taxable Benefits / Avant. impos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture &amp; Food</td>
<td>ANTE</td>
<td>DAVID</td>
<td>Director, Strategic Policy Branch</td>
<td>$112,785.31</td>
<td>$207.09</td>
</tr>
<tr>
<td>Agriculture &amp; Food</td>
<td>ARCHIBALD</td>
<td>BRUCE</td>
<td>ADM, Policy &amp; Farm Finance Division</td>
<td>$161,194.88</td>
<td>$299.99</td>
</tr>
<tr>
<td>Agriculture &amp; Food</td>
<td>BAKER</td>
<td>THOMAS</td>
<td>Director, Food Inspection Branch</td>
<td>$111,379.15</td>
<td>$206.45</td>
</tr>
<tr>
<td>Agriculture &amp; Food</td>
<td>BITRAN</td>
<td>MAURICE</td>
<td>Director, Innovation &amp; Risk Mgm</td>
<td>$104,986.79</td>
<td>$187.11</td>
</tr>
<tr>
<td>Agriculture &amp; Food</td>
<td>CHAPMAN</td>
<td>J. DOUGLAS</td>
<td>Director, Food Industry Competitiveness</td>
<td>$111,379.15</td>
<td>$206.45</td>
</tr>
</tbody>
</table>

Figure 3.5: Ministries sector in 2004.

To maintain consistency in the sector attribute for each employer, manual inspection of the Crown and Ontario Public Service sectors was performed to identify the employers impacted by this change. Specific Python scripts were then written for
the Crown and Ontario Public Service sectors to assign the correct sector attribute during the collection process. For example, the result of these scripts can be seen in Table 3.2, which shows records scraped from the Crown sector in 1999.

Table 3.2: Scraped data from the Crown sector in 1999.

<table>
<thead>
<tr>
<th>Sector</th>
<th>Employer</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
<th>Salary Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crown</td>
<td>Ontario Financing Authority</td>
<td>Allain</td>
<td>Charles A</td>
<td>Manager</td>
<td>$131,808.14</td>
</tr>
<tr>
<td>Crown</td>
<td>Ontario Financing Authority</td>
<td>Cheunh</td>
<td>Morris GM</td>
<td>Proj. Manager</td>
<td>$105,816.04</td>
</tr>
<tr>
<td>Electric</td>
<td>Ontario Hydro</td>
<td>Acchione</td>
<td>PN</td>
<td>Manager</td>
<td>$107,527.00</td>
</tr>
<tr>
<td>Electric</td>
<td>Ontario Hydro</td>
<td>Agostino</td>
<td>J</td>
<td>Solicitor</td>
<td>$111,778.00</td>
</tr>
<tr>
<td>Crown</td>
<td>Ontario International Trade Corp.</td>
<td>Crispino</td>
<td>Leonard</td>
<td>CEO</td>
<td>$116,238.88</td>
</tr>
<tr>
<td>Crown</td>
<td>Ontario Lottery Corporation</td>
<td>Barnett</td>
<td>Ken</td>
<td>Relationship Mgr.</td>
<td>$106,396.53</td>
</tr>
</tbody>
</table>

Even though the data was scraped from the Crown sector, Table 3.2 shows that records for the employer Ontario Hydro are augmented with the sector attribute ‘Electric’, which reflects the introduction of the Electric sector in the 2000 disclosure year.

Once these changes were identified, the collection process was performed. To collect the Sunshine List data, the HTML versions for each sector were crawled and scraped using Python scripts created as part of this thesis. While the algorithms varied from year-to-year, Algorithm 1 shows the general procedure that was used when scraping the HTML source code.

In Algorithm 1, the URL of the sector to be scraped is passed to the function
Algorithm 1 Data Collection

1: procedure COLLECTDATA (URL)
2: BeautifulSoup(url)
3: for each ‘table’ do
4: for each ‘tr’ do
5: for each ‘td’ in ‘tr’ do
6: if lang="en" in cell then
7: td ← English Value
8: else
9: td ← Value
10: end if
11: end for
12: sector ← URL[‘ =(.+?)&’].group(1)
13: employer ← td[0]
14: surname ← td[1]
15: given name ← td[2]
16: position ← td[3]
17: salary ← td[4]
18: benefits ← td[5]
19: end for
20: end for
21: Write to file (sector, employer, surname, given name, position, salary, benefits)
22: end procedure

‘CollectData’ (line 1). The URL is then passed to a method within the BeautifulSoup² library to extract data within various HTML tags, such as table, tr, and td (line 2).

In lines 3-5, the algorithm loops through each table, each table row, and then each cell within that row. Lines 6-10 check if the language contained in each cell is provided. If the language tag for English is provided, only the English value is extracted from the cell; otherwise, the entire value is extracted. Line 12 extracts the sector name from the URL using regular expressions, lines 15-20 store the cell values into the correctly named variable; and line 21 writes the data to a Comma Separated Value (CSV) file.

In total, 15 different versions of Algorithm 1 were created to maintain correct scraping functionality through the 18-year span of the Sunshine List disclosure.

²A Python package for parsing HTML documents. This library provides methods for creating parse trees and extracting data from HTML documents, and is available from: http://www.crummy.com/software/BeautifulSoup/
The previous algorithm operates with a complexity of $O(nmj)$, where $n$ is the number of tables in the HTML source code, $m$ is the total number of table rows, and $j$ is the number of cells per table row. Each script takes approximately 20 to 40 seconds to run\textsuperscript{3}. Once all 18 CSV files were collected, the data contained in the files was imported in an database. This was done to provide an efficient means of processing bulk data through the use of SQL queries.

3.3 Data Cleaning

Data cleaning is the process of detecting, correcting and/or removing inaccurate records from a data set [47]. This can be done through manual inspection of the data, using a database framework, or as batch process by writing tailored scripts [47]. The cleaning performed on the Sunshine List data was a lengthy process that incorporated each of these methods to varying degrees. A large portion of the cleaning process was filtering the data within the database and then manually inspecting the data to understand the specific formatting for each disclosure year. It was observed that the formatting of the Sunshine List varied from year-to-year with substantial changes occurring in years 2004, 2009, 2011 and 2013. Moreover, the formatting for a given sector within a given year varied as well and, in fact, the formatting for some sectors was observed to be employer dependent.

The following subsections of this chapter describe the formatting, language, and attribute inconsistencies encountered during the cleaning process, as well as the techniques used to correct and complete inaccurate or missing information, and the steps taken to standardize the Sunshine List data.

\textsuperscript{3}Mac OS X 10.10.5, 2.6GHz Intel Core i5, 8GB 1600 MHz DDR3 ram
3.3.1 Data Preparation

The first stage of cleaning was to inspect the collected data to ensure that the content in each column matched their respective column headings. This was done specifically for the given name and salary fields through batch processing. For the given name field, scripts were written to extract only the given name and then regular expressions were used to determine if non-alphabetic characters were present, such as ‘.’ or ‘-’. Through this process it was observed that the information provided in the given name field was inconsistent. For example, in some instances only initials were provided and in others additional information, such as a professional designation, prefaced an individual’s given name. The most prevalent case was the title ‘Dr.’, as seen in Fig. 3.6.

![Salary Disclosure 2003 (Disclosure for 2002 Hospitals)]

<table>
<thead>
<tr>
<th>Employer / Employeur</th>
<th>Surname / Nom de famille</th>
<th>Given Name / Prénom</th>
<th>Position / Poste</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>CHAN</td>
<td>WINNIE</td>
<td>Pathologist</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>FRAPE</td>
<td>NORI</td>
<td>Oncology Associate</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>GROULX</td>
<td>ANNE</td>
<td>Program Manager, Finance</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>GUTHRIE</td>
<td>STUART</td>
<td>Program Manager, Property and Materials</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>KAIGAS</td>
<td>TINA</td>
<td>Medical Director</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>MAY</td>
<td>DOUG</td>
<td>Oncology Associate</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>MORRIS</td>
<td>ANNA</td>
<td>Pathologist</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>RUSSELL</td>
<td>SHARON</td>
<td>Oncology Associate</td>
</tr>
<tr>
<td>Cambridge Memorial Hospital</td>
<td>SNIDER</td>
<td>JAMES</td>
<td>Chief of Staff</td>
</tr>
</tbody>
</table>

Figure 3.6: Professional designation in given name field.

A high level of clarity in the given name field is crucial for the gender inference described in Chapter 4. As such, the format of the given name field was standardized. To accomplish this, Python scripts, using regular expressions, were created to append professional designations to the end of the given name field, and replace uncommon characters with spaces. For example, the given name ‘Dr. Jean-Paul’ was replaced with ‘Jean Paul Dr.’ to match the expected formatting of the inference model dis-
cussed in Chapter 4. This standardization was performed for approximately 1.0% of the 652,804 records.

The cleaning of the salary field involved removing dollar signs ($) and commas. While the dollar sign and comma characters provide context, they are not usable when processing the salary as a numerical (integer) value. Dollar signs and commas were removed from all 652,804 records.

3.3.2 Employer Name Changes

The formatting of the employer attribute for each record was observed to be inconsistent for each disclosure year. This inconsistency was most prevalent in the University and School Board sectors, and was based on whether or not an employer provided their full name or an abbreviated version. For example, in the School Board sector some employers abbreviate the terms ‘Catholic District School Board’ and ‘District School Board’ to ‘CDSB’ and ‘DSB’, respectively, while others did not. This inconsistency was also from year-to-year for the same employer as shown in Fig. 3.7, which highlights the evolution of the Algoma District School Board’s employer name over time.

<table>
<thead>
<tr>
<th>Employer / Employeur</th>
<th>Employer / Employeur</th>
<th>Employer / Employeur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
<tr>
<td>Algoma District School Board</td>
<td>Algoma DSB</td>
<td>Algoma District School Board</td>
</tr>
</tbody>
</table>

(a) (b) (c)

Figure 3.7: Employer name changes a) 2005 disclosure; b) 2006 disclosure; c) 2007 disclosure.
Filtering the data in the database using SQL queries revealed that for the years 1997-2004 and 2006 the abbreviated versions were most common; however, in 2005 and 2007-2014 the long forms were heavily utilized by employers. The inconsistency of abbreviating the term ‘District School Board’ accounted for 6862 records alone. Similarly, the University sector experienced the shortening of employer names that incorporated the terms ‘University’, ‘School of’, and ‘College’. For example, the University of Guelph is referred to as the ‘University of Guelph’ or simply as ‘Guelph’ depending on the disclosure year, whereas ‘Niagara College’ can be identified as ‘Niagara College’ or ‘Niagara’. The employer fields were standardized based on their most recent long-form name. This means that if an employer’s most recent disclosure was in 2014, but the most recent long-form name occurred in 2013, all records provided by this employer would have their employer name field updated to share the 2013 value, including the 2014 disclosure year. The standardization of employer name was required for 72,150 records (11%). This was done to assist with creating longitudinal information for the Sunshine List data, as explained in Section 3.4.

3.3.3 Language Inconsistencies

The language requirement for the information provided by each employer is simply that it must be in at least one of the two official languages of Canada: English or French. This flexibility leads to an inconsistency with regards to the language provided by each employer. For a given year, some employers provide English only, French only, or both. Moreover, this reporting was not consistent through the years, even for the same employer. Fig. 3.8 demonstrates the inconsistency in the language provided by the University of Ottawa.

Language inconsistencies varied from year-to-year, sector-to-sector, and in some cases for each attribute field within a record. As 74% of the records in the Sunshine List data are exclusively English, if information was provided in both languages, English was kept as the standard. This affected 9% of the records. If only French
Figure 3.8: Language provided by the University of Ottawa a) 2012 disclosure; b) 2014 disclosure.

was provided, common words (specifically employers) that could be identified were translated from French to English. This was done to standardize the language of the cleaned data, and to assist with comparing employers (e.g., University of Ottawa vs. Université d’Ottawa) and positions (e.g., Professor vs. Professeur) when creating longitudinal data links between records. This standardization affected 4% of the records.

3.3.4 Special Characters

The occurrence of special characters within the attribute of a record also caused issues when trying to scrape the HTML source code. The problem is how these characters are represented in HTML as opposed to rich text applications. For example, the character ‘é’ is represented by the character code ‘&eacute;’, whereas the character ‘ô’ is referenced by ‘&ocirc;’. The representation of these characters in the HTML character code can be seen in Fig. 3.9, along with several other special characters prefixed with ‘&’.

For each record, regardless of the language provided by the employer, a special
character could exist in any one of a records attribute fields. While this was most common in employers providing French only, this also occurred in the given name field of records when an employer was providing English only for the remaining attributes. To standardize the language of the cleaned data the special characters were replaced by the characters they are derived from. For example, for the given name ‘Adèle’, the resulting given name would be recorded as ‘Adele’. This standardization was applied to 13% of the records; however, the total number of replacements is a significantly higher number due to records containing multiple special characters. This standardization was performed to simplify the information storage processes, and to ensure that a record’s given name was correctly processed by the gender inference model discussed in Chapter 4.

### 3.4 Longitudinal Data

*Longitudinal* data tracks the same sample at different points in time [39]. In this thesis, we also considered linking records across the 18-year span of disclosures. The purpose of linking records is to create a foundation for updating incomplete and missing values found in the original data set. For example, if the same individual appears on the Sunshine List from 1997-2008, these links offer a way of tracking the changes of this record over time to ensure that each record has the most complete information. For example, Table 3.3 shows the original Sunshine List for a single real-world entity over an 18-year span.
Using these links in Table 3.3, incomplete or missing information for a given disclosure year can be corrected. For example, only an initial is provided for the given name field for disclosure years 1997 and 1999. Once a longitudinal link is created for these records, the initial can be converted to the correct given name of ‘Joseph’. In this thesis we propose a 2-stage naïve approach for linking records based on exact matches (Section 3.4.1). In addition, we investigate a second method that allows for inexact records to be matched (Section 3.4.2).

### 3.4.1 Naïve Approach

The naïve approach is a 2-stage process. The first stage creates links only if there are exact matches for the sector, employer, surname, and given name fields. A caveat of this approach is that the given name must be longer than 1 character in length and not contain a ‘.’ character to avoid comparing only initials. In addition,
the disclosure years of the records being compared must be different. The purpose of the first stage is to safeguard against records with subtle differences that could lead to the linking of multiple real-world entities. The linking process is described in Algorithm 2.

**Algorithm 2 First stage of naive approach**

1: allRecords[] ← loadRecords()
2: procedure naiveStageOne (allRecords)
3: for each record[i] in allRecords do
4:   for each record[j] in allRecords do
5:     if record[i].year != record[j].year and record[i] == record[j] then
6:       record[i].ID ← uniqueID
7:       record[j].ID ← uniqueID
8:     end if
9:   end for
10: end for
11: end procedure

In Algorithm 2, prior to executing the procedure *naiveStageOne*, the routine *loadRecords* loads all of the cleaned and standardized records into memory (*allRecords*). The outer *for* loop iterates over each record[i] and the inner *for* loop iterates over each record[j] where i and j denote the index position of a record. During each iteration of the inner *for* loop, the algorithm checks if record[i] and record[j] are from different disclosure years (each individual only appears once per disclosure year) and if the records share exact matches for sector, employer, surname and given name. The complexity of Algorithm 2 is $O(n^2)$, where $n$ is the number of records in *allRecords*. If the conditional at line 5 is met, record[i] and record[j] are assigned the same unique ID. For example, the records in Table 3.4 map to a different real-world entity; however, aside from the given name, the remaining attributes are the same.

The result of the first stage of this algorithm would link the records in Table 3.4 as shown in Table 3.5 by assigning a unique ID number to the linked records.

Table 3.5 shows that the records for 1997 and 1998 were not linked due to only initials being provided. In addition, the record for 2002 was not linked because the
Table 3.4: Two different real-world entities in the Sunshine List data.

<table>
<thead>
<tr>
<th>Sector</th>
<th>Year</th>
<th>Employer</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>1997</td>
<td>McMaster University</td>
<td>Childs</td>
<td>R. F.</td>
<td>Professor</td>
</tr>
<tr>
<td>University</td>
<td>1998</td>
<td>McMaster University</td>
<td>Childs</td>
<td>R. F.</td>
<td>Professor</td>
</tr>
<tr>
<td>University</td>
<td>1999</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald F.</td>
<td>Professor</td>
</tr>
<tr>
<td>University</td>
<td>2001</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald F.</td>
<td>Professor</td>
</tr>
<tr>
<td>University</td>
<td>2002</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>University</td>
<td>2014</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Robert</td>
<td>Professor</td>
</tr>
<tr>
<td>University</td>
<td>2015</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Robert</td>
<td>Professor</td>
</tr>
</tbody>
</table>

Table 3.5: First stage results of naïve approach.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sector</th>
<th>Year</th>
<th>Employer</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>University</td>
<td>1997</td>
<td>McMaster University</td>
<td>Childs</td>
<td>R. F.</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>1998</td>
<td>McMaster University</td>
<td>Childs</td>
<td>R. F.</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>1999</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald F.</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>2001</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald F.</td>
<td>Professor</td>
</tr>
<tr>
<td>10001</td>
<td>University</td>
<td>2014</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>10001</td>
<td>University</td>
<td>2015</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Robert</td>
<td>Professor</td>
</tr>
</tbody>
</table>

given name field did not match any other records. The records for 1999 and 2001 were linked, as were the records for 2014 and 2015 due to exact matches in the required fields.

The second stage creates links in a similar fashion to stage 1; however, when comparing given names during stage 2, only the first initial must be an exact match. A caveat of this is that at least one of the compared records must contain a full given name. In Table 3.5, although the 1997 and 1998 records are an exact match, these records would still not be linked as neither provides a full given name. Moreover, if there is ambiguity between different records, the algorithm will also check middle names (if provided) to decide which link would be most correct. If no distinction between conflicting records can be made, the algorithm will err on the side of caution and not create a link. In the case that a record is linked to a record with a unique ID (link created in the first stage), the unique ID of the newly linked record will be that of the record it was compared to. In contrast, if neither of the newly linked records
received an ID during the first stage, the newly linked records are assigned a new unique ID. The process of linking records in stage 2 is shown in Algorithm 3.

Algorithm 3 Second stage of naive approach

```plaintext
1: procedure naiveStageTwo (allRecords)
2:     for each record[i] in allRecords do
3:         matchedRecords[]
4:             if record[i].ID == None then
5:                 for each record[j] in allRecords do
6:                     if record[i].year != record[j].year and record[i] == record[j] then
7:                         matchedRecords.append(record[j])
8:                 end if
9:             end for
10:             if matchedRecords is not empty then
11:                 matchedRecords.append(record[i])
12:                 ID ← None
13:                 (matchedRecords, ID) ← checkAmbiguity (matchedRecords)
14:                 if ID == None then
15:                     (matchedRecords, ID) ← tryMiddleInitial(matchedRecords, ID)
16:                 end if
17:                 if ID != None then
18:                     matchedRecords ← updateRecords(matchedRecords, ID)
19:                 end if
20:             end if
21:     end for
22:     WriteToFile (record[i])
23: end procedure
```

In Algorithm 3, `allRecords` (created in `naiveStageOne`) is passed to the procedure `naiveStageTwo`. The outer `for` loop iterates over each record in `allRecords`. If `record[i]`, where `i` denotes the index position of the record in `allRecords`, was not assigned an ID in stage 1, the inner `for` loop is executed. `Record[i]` and `record[j]`, where `j` denotes the index position of the record in `allRecords`, are compared, and if they qualify for a match, `record[j]` is inserted into `matchedRecords`. Upon exiting the inner `for` loop, if at least one match was found (`matchedRecords` is not empty), `record[i]` is inserted into `matchedRecords`. To determine ambiguity the subroutine `checkAmbiguity` is called. If the subroutine is successful (resulting ID is not ‘None’), the subroutine `updateRecords`
is called. However, if an ambiguity is found, the subroutine \textit{tryMiddleInitial} is called, which uses middle initials in a record’s given name field, if provided, to try and clarify the ambiguity. If the subroutine \textit{tryMiddleInitial} is successful, the subroutine \textit{updateRecords} is called. The complexity of Algorithm 3 is $O(n^2 mj)$, where $n$ is the number of records in $allRecords$, $m$ is the number of records in $matchedRecords$ and $j$ is the number of subroutine calls.

In contrast to stage 1, stage 2 also updates or corrects incomplete or missing information (subroutine \textit{updateRecords}). This is done for all records that share the same unique ID, and is specifically designed to update the given name field of the linked records to the most recent long form given name. For example, the results of the second stage of this algorithm are shown in Table 3.6.

Table 3.6: Second stage results of naïve approach.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sector</th>
<th>Year</th>
<th>Employer</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>University</td>
<td>1997</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>1998</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>1999</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>2001</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>10000</td>
<td>University</td>
<td>2002</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Ronald Frank</td>
<td>Professor</td>
</tr>
<tr>
<td>10001</td>
<td>University</td>
<td>2014</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Robert</td>
<td>Professor</td>
</tr>
<tr>
<td>10001</td>
<td>University</td>
<td>2015</td>
<td>McMaster University</td>
<td>Childs</td>
<td>Robert</td>
<td>Professor</td>
</tr>
</tbody>
</table>

In comparison to stage 1, which was not able to find links for the 1997, 1998 and 2002 records, the results in Table 3.6, show that these records were linked in the the second stage (\textit{naiveStageTwo}). Executing Algorithm 3 would have revealed an ambiguity for the 1997, 1998 and 1999 records, as these records could be linked to the 1999 and 2000 records, or the 2014 and 2015 records. Because middle initials are provided, the subroutine \textit{tryMiddleInitial} clarifies the ambiguity and the correct records are linked and updated. Table 3.6 also shows that the information for each of the linked records (ID 10000) is updated to the most recent long form given name (`Ronald Frank`). Performance results of the naïve approach are presented in Section 3.5.
3.4.2 Weighted Probability Approach

In contrast to the previous algorithm, the work in [41] describes a method that takes into account differences in an entity over time. In other words, it provides a way to link records over time, even if the records do not have exact matches. The authors of [41] create longitudinal links through similarity computations of the records in increasing time order, and propose three different longitudinal record matching techniques: i) *Early binding* which merges records with an already created cluster if the similarity comparison is sufficiently high, or creates a new cluster if not, ii) *Late binding* performs soft clustering, which keeps all evidence and makes clustering decisions at the end, and iii) *Adjusted binding* which performs multiple phases of soft clustering and compares records with clusters that are created for records with later time stamps. The adjusted binding technique outperforms both the early and late binding techniques; however, is more complex [41]. The performance results of the weighted probability approach are presented in Section 3.5. A description of the Algorithm can be found in Appendix A.

3.5 Experimental Methodology

To verify the performance of the naïve and weighted probability approaches, a subsection of the Sunshine List, for which longitudinal links could be manually verified, was used. The records used were from the University of Guelph records (labelled data). The naïve and weighted probability approach were then employed separately to recreate the longitudinal links in the labelled data. The results of both algorithms were then compared to the labelled data to determine the performance of each algorithm. It should be noted that for this evaluation only the first stage of the naïve approach was used. This was done because the weighted probability approach does not update incomplete or missing information, and records with only initials in the given name field were not eligible to be linked. In contrast, the second stage of
the naïve approach may be able to update the given name field and link additional records. The performance of the naïve and weighted probability approaches, with regards to the manually verified labelled data (1154 real-world entities exist in the labelled data), is summarized in Table 3.7.

Table 3.7: Performance on the verified Guelph data.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Real-world Entities</th>
<th>Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>1154</td>
<td>67.0s</td>
</tr>
<tr>
<td>Weighted Probability</td>
<td>1154</td>
<td>232.0s</td>
</tr>
</tbody>
</table>

The results in Table 3.7 show that the naïve and weighted probability approaches successfully recreated the same number of longitudinal links in the labelled data. In addition, the real-world entities found with both models were the same. In other words, each set of records assigned a Unique ID by the naïve approach mapped to a single cluster generated by the weighted probability approach. Due to the extensive cleaning and standardization performed in Chapter 3 there was minimal variability in the attribute fields. Because of this, the number of available full life spans, utilized by the weighted probability approach, is limited, and essentially reduces the weighted probability approach to a computationally demanding naïve model (exact matches). The results are confirmed by the work in [43], which show that the weighted probability approach fails to take into consideration the characteristics of the data source. The performance disparity between these models is the run time. The run time of the naïve approach is 67.0s and the run time of the weighted probability approach is 232.0s (including training time).

Based on the fact that the algorithms produce identical links for the labelled data, but the naïve approach has smaller run times, we use the naïve approach for this thesis. Longitudinal links were created for 91% of the 652,804 records, which map

---

4The naïve approach was implemented in Python and the weighted probability approach was implemented in Java (as provided by the authors of [41]). These algorithms were executed using a laptop running Mac OS X 10.10.5, with a 2.6GHz Intel Core i5 processor, and 8GB 1600 MHz DDR3 ram.
to 155,114 unique real-world entities. Moreover, with this longitudinal information, the naïve approach was used to update the given name attribute for 11,929 records (72% of the records that had incomplete information in that field).

3.6 Summary

In this chapter, the inconsistencies, issues, and limitations of the Sunshine List data, as provided by the Ministry of Finance, were discussed and while the Sunshine List data contains useful information, the data is not readily available in a single location as a single data set. The available data first had to be scraped and loaded into a single location and loaded into a database management system. Once collected, the data underwent extensive cleaning and standardization and through this it was observed that records contained incomplete or missing information, specifically in the given name field. Because the given name field is a crucial attribute for the gender wage-gap analysis discussed in Chapter 5, the performance of two techniques for creating longitudinal links were compared. We discussed the performance of each and justified the selection of a naïve approach for this thesis. Using the naïve technique we created longitudinal links for 91% of the records and completed the given name field for 72% of the records with incomplete information (only initials) in that field. Finally, the cleaned and standardized data publicly will be made available to all researchers who wish to continue work in this area.
Chapter 4

Gender Inference

The information provided in the Sunshine List is not sufficient for analyzing and understanding the role of gender in Ontario’s wage-gap, as it lacks information about the gender of an individual. In this chapter, two techniques found in literature for inferring gender are presented and discussed, along with our own proposed hybrid model. This section also describes augmenting the cleaned and standardized data from Chapter 3 with the newly created gender variable to create the enhanced database used in the wage-gap analysis presented later in Chapter 5. Of the attributes provided in the original data, the most insightful with regards to the gender of an individual is the given name field. There are various techniques [17, 21, 24, 20, 22, 48] to infer gender form a given name; however, the methods studied as part of this thesis are a model that uses relative name frequencies, a heuristic model, and a hybrid model that overlays these techniques.

4.1 Relative Name Frequencies

The relative name frequency model uses historical data provided by the US Social Security Bureau. It’s important to note that US data is used as a surrogate because Canadian data is not available due to privacy concerns. The US Social Security data is available from the Official Social Security website\(^1\). This information is provided annually, spans the years 1880 to 2014, and in total contains 1,825,433 records. The records in this data represent all of the given names of US citizens with the exclusion

\(^1\)http://www.ssa.gov/oact/babynames/limits.html
of given names that occur less than 5 times within the US population. The records for each year are provided in a single text file, where each line in the file represents one record, and each record contains three values separated by a comma: *given name*, *gender*, and *frequency* (Fig. 4.1).

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emma</td>
<td>F</td>
<td>20799</td>
</tr>
<tr>
<td>Olivia</td>
<td>F</td>
<td>19674</td>
</tr>
<tr>
<td>Sophia</td>
<td>F</td>
<td>18490</td>
</tr>
<tr>
<td>Andrew</td>
<td>F</td>
<td>11</td>
</tr>
<tr>
<td>Andrew</td>
<td>M</td>
<td>11069</td>
</tr>
<tr>
<td>Andrew</td>
<td>M</td>
<td>11069</td>
</tr>
<tr>
<td>Riley</td>
<td>F</td>
<td>4761</td>
</tr>
<tr>
<td>Riley</td>
<td>M</td>
<td>2429</td>
</tr>
</tbody>
</table>

Figure 4.1: Screenshot of the 2014 US Social Security data.

The frequency of each record represents the number of times the given name occurs with the gender, ‘M’ for male or ‘F’ for female for that particular year. The first record in Fig. 4.1, *Emma,F,20799*, means that in the year 2014, there are 20,799 females with the given name Emma. It should be noted that each given name found within the file is not necessarily unique and can appear either as male, female or both. For example, in Fig. 4.1, the name ‘Andrew’ appears twice, and has a female frequency of 11 and a male frequency of 11,069.

The frequency of a given name will vary from year-to-year due to deaths and births within the US population. Over time, this can result in the removal of given names (i.e., the given name falls below the threshold of 5) or the introduction of new given names (i.e., given names with a frequency of at least 5, that were below the threshold of 5 in previous years). To maximize the number of given names available for the gender inference models discussed later in this chapter, the frequencies for each given name and specific gender were summed for the entire span of the data using Algorithm 4.

In Algorithm 4, before the procedure *sumFrequencies* is executed, the routine *concatenateSort* loads all of the records from the entire 134-year span into memory.
Algorithm 4 Sum Given Name Frequencies

1: SSdata ← ConcatenateSort()
2: procedure sumFrequencies (SSdata)
3: name ← Unknown
4: first ← True
5: for each record in SSdata do
6:     if name != record[0] and first == False then
7:         WriteToFile (name,maleTot,femaleTot)
8:     first ← True
9: end if
10: if first == True then
11:     name ← record[0]
12:     if record[1] == 'M' then
13:         maleTot = record[2]
14:         femaleTot = 0
15:     else
16:         maleTot = 0
17:         femaleTot = record[2]
18: end if
19: first ← False
20: else
21:     if record[1] == 'M' then
22:         maleTot += record[2]
23:     else
24:         femaleTot += record[2]
25: end if
26: end if
27: end for
28: end procedure

and sorts the records in ascending order by given name and then by gender (SSdata). Procedure sumFrequencies is then executed and the variables name (given name being searched for in SSdata), and first (Boolean value to determine if the current record is the first occurrence of a given name) are assigned initial values. For each iteration through the for loop, the given name of the current record (the index 0 in record[0] denotes the position of the given name within a record) is compared to the value in the variable name, and the Boolean value first is checked (Line 6). Because the records have been sorted, this conditional is used to determine if the given name of
the current record differs from the previous record and whether or not this is the first occurrence of the given name. If this condition is satisfied, every occurrence of a given name has been summed and a single record containing the given name, total male frequency, and total female frequency is written to file. If the condition in line 6 is not satisfied, the condition at line 12 checks to see if the current record contains the first occurrence of a given name. If first is true, the values of name, maleTot, and femaleTot are set equal to that of the current record, where record[0], record[1], and record[2] refer to the record’s given name, gender, and frequency, respectively. If the final condition on Line 17 is reached, the current record and previous record have matching given names, meaning that the frequency of the current record is added to the correct gender frequency total. The complexity of this algorithm is \( O(n) \), where \( n \) is the total number of records. The output of this algorithm is a single file, where each line represents a single record containing three values separated by commas: given name, total male frequency, and total female frequency (Fig. 4.2).

<table>
<thead>
<tr>
<th>Given Name</th>
<th>Male Frequency</th>
<th>Female Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaban</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>Aabha</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Aabid</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Aabriella</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Aadam</td>
<td>161</td>
<td>5</td>
</tr>
<tr>
<td>Aadan</td>
<td>91</td>
<td>0</td>
</tr>
<tr>
<td>Aadarsh</td>
<td>128</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.2: Screenshot of the file containing total male and female frequencies.

In contrast to the files provided by the US Social Security Bureau, each given name in the file created by Algorithm 4 (totalFrequencies) is unique, as each record (line) contains both the total male and total female frequencies.

Using the data in the totalFrequencies file, the probability that an individual is male or female, based on the frequency for each gender, can be calculated using Equations 4.1 and 4.2.
\[ P(\text{Male}|\text{Name}) = \frac{\text{MaleFrequency}(\text{Name})}{\text{MaleFrequency}(\text{Name}) + \text{FemaleFrequency}(\text{Name})} \] (4.1)

\[ P(\text{Female}|\text{Name}) = \frac{\text{FemaleFrequency}(\text{Name})}{\text{MaleFrequency}(\text{Name}) + \text{FemaleFrequency}(\text{Name})} \] (4.2)

For instance, using the information in Fig. 4.2, the given name ‘Aaden’ has a male frequency of 3474 and a female frequency of 5, which means that when the given name ‘Aaden’ is encountered in the dataset there is a 0.999 (i.e., 3474 / 3479) probability that the name belongs to a male. Because Equations 4.1 and 4.2 are complements of one another (\( P(\text{Female}|\text{Name}) = 1 - P(\text{Male}|\text{Name}) \)), only one is required to determine the gender probability. Equation 4.1 is employed in Algorithm 5 to assign gender to the records contained in the cleaned and standardized data generated in Chapter 3.

**Algorithm 5** Gender Assignment

1: Load totalFrequencies
2: for each record in the cleaned and standardized data do
3:     Gender = UNKNOWN
4:     if Given Name in totalFrequencies then
5:         Get total MALE frequency from totalFrequencies
6:         Get total FEMALE frequency from totalFrequencies
7:         Gender ← P(Male|Given Name)
8:     end if
9: end for

In Algorithm 5, the file *totalFrequencies* is loaded into memory and then the algorithm iterates over every record in the cleaned and standardized data generated in Chapter 3. If the given name of a record matches a given name in *totalFrequencies*, the total male and female frequencies are extracted, the probability of an individual
being male is calculated using Equation 4.2, and the record is augmented with this
gender value. The one caveat of this technique is that a gender can only
be assigned if the name appears in the historical data, otherwise the gender
remains as ‘Unknown’. The complexity of Algorithm 5 is \( O(nm) \), where \( n \) is the
number of records in the cleaned and standardized data, and \( m \) is the number of
given names in the totalFrequencies file that have the same first letter as the given
name in the record.

It should be noted that the relative name frequency model can produce one of
four results for each given name. The first is the correct gender assignment which
means that the gender probability assigned to a given name was greater than or equal
to 0.95 and matches the gender associated with the given name in the labelled data.
For example, if the given name ‘Andrew’ is assigned a male probability of 0.95
(or greater) and the gender in the labelled data is male, this is considered a correct
gender assignment. The second result is a gender probability that is less than 0.95%
(e.g., the given name ‘Nickie’ results in a female probability of 0.68). For this case
the records are assigned a gender probability; however, the records are not studied
in the gender wage-gap analysis presented later in Chapter 5. The third result is
an incorrect gender assignment, meaning that the gender probability assigned to a
given name does not match the gender in the labelled data. The final result is an
‘Unknown’ gender assignment, meaning that the given name was not found in the US
Social Security Data. In Section 4.3 we present a hybrid gender predictor model to
deal with the given names that result in an ‘Unknown’ gender assignment.

4.1.1 Relative Name Frequencies Model Performance

To assess the performance the relative name frequency, the gender for a subset
of the records in the cleaned and standardized data was manually verified (labelled
data). The subset use is for the University of Guelph, which accounts for 6616 records.
Among these, there are 590 unique (given) names. It is important to note that the
records that share one of the 590 given names, also share a common gender, meaning that each unique name given is used exclusively for male or female. The results of the relative name frequency model are presented in Table 4.1.

<table>
<thead>
<tr>
<th>Experimental Result</th>
<th>Number of (Unique) Names</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Gender Assigned</td>
<td>481</td>
<td>81.53</td>
</tr>
<tr>
<td>Less than 95% (Threshold)</td>
<td>48</td>
<td>8.14</td>
</tr>
<tr>
<td>Incorrect</td>
<td>3</td>
<td>0.51</td>
</tr>
<tr>
<td>Unknown</td>
<td>58</td>
<td>9.83</td>
</tr>
</tbody>
</table>

From Table 4.1, the relative name frequency model is able to assign the correct gender nearly 82% of the time, a gender probability less than 95% occurred for 8% of the the unique given names, and the incorrect gender is only assigned in 0.51% of the cases. Names that do not appear in the historical data list are not assigned a gender and are classified as ‘Unknown’ (9.83% of cases, Appendix B). The run time\(^2\) of this model for the labelled data is 301.79s.

### 4.2 Heuristic Model

The heuristic gender predictor model described in [24] utilizes a Naïve Bayes classifier, which is a conditional probability model based on applying Bayes’ theorem [32], to infer the gender from a given name. For this model to classify a given name as either male or female, the model must first learn the features of a given name that are most commonly associated with each gender (training phase). The data used for this model is the original data provided by the US Social Security Bureau [49]. Recall that this data is provided annually, spans the years 1880-2014, and contains three values per record: given name, gender, and frequency. As described in [24], the algorithm sums the data for every year into a single dataset, resulting in a dataset

\(^2\)Mac OS X 10.10.5, 2.6GHz Intel Core i5, 8GB 1600 MHz DDR3 ram
similar to totalFrequencies. It then randomly splits the data into two parts: training data (20%) and validation data (80%). It is important to note that one weakness of this model is that Algorithm 6 does not perform cross validation and, therefore, is potentially susceptible to Type III errors (assigning the correct gender for the wrong reasons) [50]. During the training process, the given names within the training data are decomposed to n-grams, which are character sequences of length \( n \), where \( n = 1, 2 \) or 3 that appear in a given name. For example, the given name ‘Chris’ would be decomposed into the following n-grams:

\[
\text{Chris} = \{ c, h, r, i, s, ch, hr, ri, is, chr, hri, ris \}
\]

The frequency of each n-gram is then used used to train the model. The training process utilizes conditional probability, which is the probability that a given name is male or female, given that it contains specific n-grams. The concept of conditional probability is presented in Equation 4.3.

\[
P(C_k | x) = \frac{p(C_k)p(x|C_k)}{p(x)}
\]  \hspace{1cm} (4.3)

In Equation 4.3, \( C_k \) denotes the classification \( k \), which is either male or female, and \( x \) is a feature vector that contains the various n-grams extracted from the given name. For this calculation, the probability of \( x \) occurring, given a gender \( P(x|C) \), is determined by the male and female frequencies of the n-grams in the training data. For example, if a feature vector contains the n-gram ‘lle’, which occurs in 990 female names, and in 10 male names, then \( P(x|C) \) would be equal to 0.99 (990 / 1000). This calculation is similar to that of the relative name frequency model, when using the male and female frequencies of a given name. The process of learning these conditional probabilities, and in essence training the heuristic model, is presented in Algorithm 6.

The algorithm first loads all of the training data into memory and then proceeds to iterate over each record. Prior to calculating the conditional probability, the routine
**Algorithm 6 Heuristic Training**

1: LoadTrainingData()
2: CondProbability[][]
3: for each record in trainingData do
4: nGrams ← makeNGrams(record[0])
5: gender ← record[1]
6: frequency ← record[2]
7: for each nGram in nGrams do
8: if nGram not in CondProbability[][] then
9: CondProbability[][] ← Equation 4.3
10: else
11: CondProbability[][] ← Update(CondProbability[][])
12: end if
13: end for
14: end for

*makeNGrams* creates all possible *n-grams* of length 3 or less, and the gender and frequency of the record are stored. The inner *for* loop iterates over each *n-gram* and checks to see if it has been seen before (line 8). If this is the first occurrence of the n-gram, the conditional probability is calculated and a new entry is added to the *CondProbability* matrix, where the first column stores *n-gram* itself, and the second column stores the conditional probability. If the *n-gram* has appeared before, the conditional probability of the *n-gram* is updated. The complexity of this algorithm is $O(nm)$, where $n$ is the number of records in *trainingData* and $m$ is the number of *n-grams* within a given name.

After the model has been trained, the classification of new data can be performed. This involves using the conditional probabilities calculated during the training phase and a maximum a posteriori decision rule, which is to pick the hypothesis that is most probable [51]. The classification of a given name is calculated as follows:

$$C_k = \arg\max_{k \in \{1, ..., K\}} p(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (4.4)$$

Equation 4.4 assigns the gender of a given name based on the conditional probabilities of the *n-grams* found within it. The *n-grams* in a given name may have
a conditional probability for both male and female; however, the given name is assigned a gender based on the features (n-grams) that are most probable, as shown in Algorithm 7.

**Algorithm 7 Heuristic Classification**

```plaintext
1: TrainModel ()
2: procedure classifyGivenName (givenName)
3:   nGrams ← makeNgrams(record[0])
4:   ProbabilityMale ← Equation 4.4, where x == nGrams and \( C_k = \text{male} \)
5:   ProbabilityFemale ← Equation 4.4, where x == nGrams and \( C_k = \text{female} \)
6:   if ProbabilityMale > ProbabilityFemale then
7:     Gender ← Male
8:   else
9:     Gender ← Female
10: end if
11: end procedure
```

Prior to classification, the model must be trained to learn the conditional probabilities of the male and female features. The given name of a record is passed to the procedure `classifyGivenName`, and then all possible n-grams of length 3 or less are created (nGramps). Using Equation 4.4, the probabilities of the features classifying the given name as male or female are calculated. The given name is then classified based on the gender with the highest probability. The procedure `classifyGivenName` is called once per record and has a linear time complexity, \( O(n) \), where \( n \) is the number of records. In contrast to the relative name frequency model, the heuristic model always assigns a gender so long as a full given name is available.

### 4.2.1 Heuristic Model Performance

The performance of the heuristic model is assessed with the same manually verified University of Guelph gender data used to assess the relative name frequency model in Section 4.1.1. The results of the heuristic model are presented in Table 4.2.

From Table 4.2, the heuristic model is able to assign the correct gender in 77% of
Table 4.2: Performance of heuristic model.

<table>
<thead>
<tr>
<th>Experimental Result</th>
<th>Number of (Unique) Names</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Gender Assigned</td>
<td>456</td>
<td>77.3</td>
</tr>
<tr>
<td>Incorrect Gender Assigned</td>
<td>134</td>
<td>22.7</td>
</tr>
</tbody>
</table>

the cases; however, due to the nature of this model (always assigning a gender), the model has an error rate of 22.7%, which is 45 times higher than that of the relative name frequency model. The run time\(^3\) of the the heuristic model, including training, for the labelled data is 28.31s.

4.3 Hybrid Model

The accuracy of the inference model is crucial in determining the role gender plays in Ontario’s wage-gap. This caused us to ask the question “can the gender inference accuracy be further improved by using both of the previous algorithms together?” In this section we present a novel hybrid algorithm based on the flow in Fig. 4.3.

![Figure 4.3: Hybrid model process flow.](image)

The proposed hybrid model is a combination of the relative name frequency and heuristic models. During the first stage, the model attempts to assign gender using the relative name frequencies from the totalFrequencies file created in Section 4.1.

\(^3\)Mac OS X 10.10.5, 2.6GHz Intel Core i5, 8GB 1600 MHz DDR3 ram
If the result of the first stage is a gender assignment of ‘Unknown’, than clearly the record is unavailable for the gender wage-gap analysis (outlined in Chapter 5). In order to maximize the amount of data available for analyzing the gender wage-gap, the number of records without gender must be minimized. Therefore, if the relative name frequency model fails to assign a gender to an individual, the heuristic model is employed [24] to these cases.

4.3.1 Hybrid Model Performance

Recall that in Table 4.1 the results of the relative name frequency model revealed an accuracy rate of 81%, an unknown gender rate of 10%, and an error rate of 0.51%. To understand the performance of the hybrid model, only the unique names that resulted in an ‘Unknown’ gender were passed through the heuristic model. The results of this analysis is summarized in Table 4.3.

Table 4.3: Performance of heuristic model on unknown University of Guelph names.

<table>
<thead>
<tr>
<th>Experimental Result</th>
<th>Number of Unknown Names</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Gender Assigned</td>
<td>46</td>
<td>79.31</td>
</tr>
<tr>
<td>Incorrect Gender Assigned</td>
<td>12</td>
<td>20.69</td>
</tr>
</tbody>
</table>

The ‘Unknown’ gender rate of approximately 10% produced by the relative name frequency model is a result of a unique given name not appearing in the total frequencies dataset. However, when these unknown given names are forwarded to the heuristic model, the heuristic model is able to assign the correct gender 79% of the time (Table 4.3). The run time\(^4\), including training, of the heuristic model for the unknown given names is 23.83s. When both gender prediction models are used together, as illustrated in Fig. 4.3, an overall accuracy of 89% is achieved, with only 2.5% of the original given names assigned an incorrect gender (Table 4.4).

\(^4\) Mac OS X 10.10.5, 2.6GHz Intel Core i5, 8GB 1600 MHz DDR3 ram
Table 4.4: Combined performance of both gender models.

<table>
<thead>
<tr>
<th>Experimental Result</th>
<th>Number of (Unique) Names</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Gender Assigned</td>
<td>527</td>
<td>89.30</td>
</tr>
<tr>
<td>Incorrect Gender Assigned</td>
<td>15</td>
<td>2.54</td>
</tr>
</tbody>
</table>

These results show that by combining both prediction models into a unified model it is possible to achieve a 7.7% improvement in accuracy over the relative name frequency model, and a 12% improvement in accuracy over the heuristic model, when these are employed in isolation. The run time\(^5\) of the hybrid model, including training of the heuristic model, is 325.66s.

### 4.4 Enhanced Database

The hybrid model was used to infer gender for the cleaned and standardized data created in Chapter 3 due to its improved accuracy when compared to the relative name frequencies and heuristic models used in isolation. Table 4.5 shows a few records for the University of Guelph.

Table 4.5: Subsection of the enhanced database.

<table>
<thead>
<tr>
<th>Sector</th>
<th>Year</th>
<th>Employer</th>
<th>% Male</th>
<th>Surname</th>
<th>Given Name</th>
<th>Position</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>2014</td>
<td>University Of Guelph</td>
<td>0.984598</td>
<td>Griswold</td>
<td>Cortland</td>
<td>Assistant Professor</td>
<td>100895.57</td>
</tr>
<tr>
<td>University</td>
<td>2014</td>
<td>University Of Guelph</td>
<td>0.995057</td>
<td>Grodzinski</td>
<td>Bernard</td>
<td>Professor</td>
<td>161127.06</td>
</tr>
<tr>
<td>University</td>
<td>2014</td>
<td>University Of Guelph</td>
<td>0.006563</td>
<td>Grogan</td>
<td>Louise</td>
<td>Associate Professor</td>
<td>137451.75</td>
</tr>
<tr>
<td>University</td>
<td>2014</td>
<td>University Of Guelph</td>
<td>0.011709</td>
<td>Guerin</td>
<td>Michele</td>
<td>Associate Professor</td>
<td>115815.04</td>
</tr>
</tbody>
</table>

\(^5\)Mac OS X 10.10.5, 2.6GHz Intel Core i5, 8GB 1600 MHz DDR3 ram
Overall, the hybrid gender inference model was used to augment 99.3% of the all records (i.e., all records with a full given name) with a new gender variable (summary statistics for the gender inference process are available in Appendix C). This enhanced database will be used in the gender wage-gap analysis presented in Chapter 5. Moreover, the enhanced database will be made publicly available to all researchers who wish to continue work in this area. It should be noted that each record is augmented with a gender probability, as opposed to a strict male or female assignment. This was done to allow flexibility in the types of analyses that other researchers may wish to perform.

4.5 Summary

In this chapter, two models from the literature, relative name frequencies and heuristic, were presented for inferring gender for a given name, along with a novel hybrid model that combines these techniques into a single model. The performance of each individual model was assessed and compared to the hybrid model using manually verified data. The results of this experimental results showed that the hybrid model outperformed the relative name frequencies and heuristic models when used in isolation by 7% and 12%, respectively. The hybrid model was then used to enhance 99.3% of the records in the cleaned and standardized data created in Chapter 3 with a new gender variable. Maximizing the number of records with a gender assignment was necessary to strengthen the results of the wage-gap analysis presented in Chapter 5.
Chapter 5

Gender Wage-gap Analysis

The cleaned, standardized and enhanced (with gender information) data generated in Chapters 3 and 4 provides the required information for performing an analysis of the gender wage-gap in Ontario’s public sector. In this chapter, we present our gender wage-gap analysis of Ontario’s public sector and discuss the conclusions of our findings. Note that in this analysis, the year corresponds to the year the Sunshine List was released, not the year of the reported income, which would be the previous calendar year (i.e., the year 2010 refers to salaries earned during 2009). Furthermore, the results are reported from two types of analysis: one taking into account all 652,804 records with a nominal income threshold of $100,000 for each year, and a separate analysis where inflation is accounted for and a threshold in real dollars is imposed by deflating the nominal $100,000 by the Consumer Price Index in Ontario for each year in the data. By doing so, the income cutoff increases over the 18-year span, up to $139,456 in 2014 (referring to salary information from 2013).

The remainder of this Chapter is organized as follows: Section 5.1 discusses the proportion of women on the Sunshine List relative to men and its impact on the gender wage-gap. Sections 5.2 and 5.3 present the gender wage-gap analysis on the Sunshine List as a whole and by quantiles of the income distribution, respectively.

5.1 Gender Shares in the Sunshine Data List Across Time

The number of records in the Sunshine List have increased each year since its inception in 1997, and as of 2014, there are a total of 652,804 records. From Fig.
5.1 it can be seen that the number of individuals on the Sunshine List has increased tremendously over the last 15 years.

![Figure 5.1: Number of records in the Sunshine List.](image)

In general, it can be seen that the growth of the number of records, with regards to the nominal threshold of $100,000, is exponential; however, the growth when inflation is accounted for is significantly less. The main reason for this disparity is the depreciation in the value of the nominal threshold relative to the average salary in the economy. In other words, part of the gap is due to inflation, which is captured by the dashed-line in Fig. 5.1, which plots the number of records above the inflation-adjusted threshold. Even then, while there is some growth in the Sunshine List, this is observed to match the growth in the real income in the economy, and not a relative growth in the public sector per se. With that being said, the more interesting observation is the change in gender shares, which is the percentage of the total records represented by each gender, across time. When the Sunshine List was first introduced the percentage of women was 22%; however, between the years 2000 and 2013, the percentage of women increased to 35%. The change in gender shares across time is presented in Fig. 5.2.

While the increase in the proportion of women on the Sunshine List is a positive
Figure 5.2: Proportion of women relative to men on the Sunshine List.

sign, women still only account for a third of the observations, and the 2014 data reveals a regression in the proportion of women. Moreover, the increase is slightly smaller if the focus is only on the subset of the Sunshine List earners who would have been on the list even if the threshold had moved at the pace of inflation; in that case, the percentage of women would increase from 22% to 31%.

The increase in gender shares seems to indicate that, as the nominal threshold on the Sunshine List continues to depreciate in value, more women relative to men make their way onto the list. To understand this further, an analysis by sector was performed; however, in order to maintain a consistency in the relative shares of sectors across time, as some sectors are introduced in different years, some sectors were combined and the original eleven sectors were reduced to five. The five sectors created from this process are: i) *Universities*, which combines the University and College sectors, ii) *Utilities* (Hydro One and Ontario Power Generation), iii) *Hospitals*, iv) *School Board*, and v) *Government and Judicial*, which combines the Crown, Judiciary, Ministries, Municipalities, and Other Public Service sectors. Fig. 5.3 and Fig. 5.4 show the percentage of women in each of these sectors across time for the Sunshine List using the nominal cut-off at $100,000 every year and the inflation-adjusted cut-off
in real dollars, respectively.

![Figure 5.3](image1.png)  
**Figure 5.3**: Percentage of women on the Sunshine List, by sector, nominal threshold.

![Figure 5.4](image2.png)  
**Figure 5.4**: Percentage of women on the Sunshine List, by sector, real threshold.

From Fig. 5.3 and Fig. 5.4 it is revealed that the *Hospitals* and *School Board* have the highest percentage of women relative to men, and currently women outnumber men in these two sectors. Conversely, women are significantly underrepresented in the *Utilities* sector, and the proportion of women relative to men is not increasing over time. An interesting observation is revealed for the *Government and Judicial*
sector as the percentage of women is higher for the inflation-adjusted threshold than
the nominal threshold. This indicates that women in this sector are relatively more
likely to have incomes above the inflation-adjusted threshold compared to men with
incomes between the two thresholds. The main conclusion from this analysis is that,
while the percentage of women on the Sunshine List increases over time, they are still
underrepresented among the top earners in the public sector.

5.2 Gender Wage-gap in the Sunshine List

The male-female wage-gap from the observations in the cleaned, standardized,
and enhanced database, using both the nominal threshold and real dollar threshold
is discussed in this section. The nominal gender wage-gap is presented in Fig. 5.5.

The results from the nominal threshold analysis reveal that, overall, the gender
wage-gap in Ontario’s public sector is minimal. Using the nominal threshold the
gender wage gap was observed to increase from zero in the late nineties to about
3% or 4% in current times. While an increasing trend was observed from 1997-
2014, the gap is relatively small compared to Ontario as a whole (26%). In contrast,
when the gender-wage gap was assessed using the inflation-adjusted threshold the gap disappears completely (Fig. 5.6).

![Image of graph showing income differential over years]

Figure 5.6: Income differential, real threshold.

The findings from this suggest that a factor contributing to the gender wage-gap is the percentage of women, rather than their earnings. The evidence shows that under the nominal threshold women who appear on the Sunshine List are paid almost as much as men, and when the records are restricted to those who would have qualified under the original terms (i.e., inflation-adjusted threshold), women are paid exactly as much as men. Moreover, the marginal increase in the gender wage-gap for the nominal threshold is consistent with the findings that individuals between the two thresholds are more likely to be female.

It should be noted that while the percentage of women was shown to be a contributing factor, it is not the only cause of the gender wage-gap. Furthermore, the marginal gender-wage gap identified for the Ontario public sector as a whole is not consistent with the findings for each sector. An analysis across the five sectors defined in Section 5.1 for the nominal and inflation-adjusted thresholds is presented Figs. 5.7 and 5.8, respectively.

The “by sector” analysis of the gender wage gap reveals that most of the average
gap, regardless of the threshold being studied, is coming from the Hospitals sector. Furthermore, an increasing trend in the gender wage-gap, for both the nominal and inflation-adjusted threshold analyses, is observed for the Universities sector. The main conclusion from these analyses is that aside from the Hospitals sector, women that do make their way onto the Sunshine List tend to get paid as much as men do. Further work will need to be conducted to identify where the pay gap of 20% is coming from in the Hospitals sector and the cause of the increasing trend for the

Figure 5.7: Income differential by sector, nominal threshold.

Figure 5.8: Income differential by sector, real threshold.
Universities sector.

5.3 Quantiles of the Income Distribution

To better understand the glass ceiling phenomena (i.e., unseen barrier preventing women from rising to the upper rungs of the cooperate ladder) and gender wage-gap revealed in [9] for the high earners in Ontario’s public sector, the differences at quantiles of the earnings distributions for men and women were analyzed. In other words, for each gender the records from the Sunshine List were cut into 5 equal sized groups based on increasing salary, meaning that the salaries of the bottom 20% females can be compared to the salaries of the bottom 20% males, and so on until the top 20%. The comparison of quantiles are shown in for the nominal and inflation-adjusted thresholds are shown in Figs. 5.9 and 5.10, respectively.

![Figure 5.9: Gender gap at quantiles of earnings distribution, nominal threshold.](image)

The findings from the nominal threshold analysis reveal that a majority of the gender wage-gap is coming from the top 20%, which is consistent with the findings
in [9]. Moreover, for the top 40% an increasing trend was observed in recent years. The findings of the inflation-adjusted threshold analysis (Fig. 5.10) show that while there is a gap for the top 40%, it is significantly smaller than the observed gap under the nominal threshold.

![Figure 5.10: Gender gap at quantiles of earnings distribution, real threshold.](image)

The findings from the differences at quantiles of the earnings distributions for men and women confirms that the glass-ceiling phenomena holds true for women in Ontario’s public sector. Moreover, there is an increasing trend in the gender wage-gap for the high earners in Ontario’s public sector.

### 5.4 Summary

In this chapter, two separate analyses of the gender wage-gap in Ontario’s public sector were presented: one that maintains a nominal threshold of $100,000, and another that accounts for inflation by deflating the nominal $100,000 threshold by the Consumer Price Index in Ontario for each year. Our analysis of the Sunshine List so
far supports four main conclusions: i) women are underrepresented on the Sunshine List by a ratio of two to one; ii) women who do make it onto the Sunshine List earn, by and large, comparable incomes with those of men; iii) while, in general, the gender pay gap does not seem to be an issue, some sectors on the Sunshine List, such as the Hospitals sector, still show a substantive pay gap; and iv) an analysis across the wage distribution of men and women seems to indicate a glass-ceiling type of gap at the higher end of the respective earning distributions of men and women.
Chapter 6

Conclusion and Future Work

In this thesis, we collected, cleaned and standardized the Sunshine List data into a form that allows for direct analysis. Once collected, longitudinal links were created to establish the history of an individual across the 18-year span of the Sunshine List data. The longitudinal links were used to update missing or incomplete information within a record, specifically the given name field for which 11,929 records were updated. Furthermore, we proposed a novel hybrid gender inference model that combines relative name frequency and heuristic techniques into a single model. Our hybrid model exhibits almost a 90% accuracy resulting in a 7.7% improvement in accuracy over the relative name frequency model and a 12% improvement in accuracy over the heuristic model, when these are employed in isolation. The hybrid gender inference model was used to enhance 99.3% of the records in the cleaned and standardized data with a new gender variable and this newly created database will be made publicly available to all researchers who wish to continue work in this area.

The findings from the gender wage-gap analysis, under the nominal and inflation-adjusted thresholds, reveal that the gender wage-gap in Ontario’s public sector is minimal, and significantly less than Ontario as a whole. The evidence identifies the underrepresentation of women on the Sunshine List as a contributing factor to the gender wage-gap; however, it was determined that women that do make it onto the Sunshine List tend to earn as much as men do. The analysis shows a substantive gap in the Hospitals sector and an increasing trend in the gender wage-gap for the Universities sector. Moreover, while the average gender wage-gap is small, the analysis
confirmed the presence of a glass-ceiling for women at the higher end of the earnings distribution.

6.1 Future Work

In the future, the longitudinal links can be used to track the progress of an individual over time, which could be a means of identifying the contribution known factors have on the gender wage-gap, or uncover additional factors. In addition, further enhancement of the database with performance markers, such as the H-Index, which measures both the productivity and citation impact of an individual, could help in understanding the increasing gender wage-gap trend in the Universities sector.

The application scope of the hybrid gender inference model extends far beyond analyzing the gender-wage gap. The hybrid gender inference model infers gender from a fundamental attribute, a given name, which can be used to enhance a database with gender, especially when high-level features, such as speech and web-activity, are not available. Gender-based analyses, while not limited to the following, can be used to examine the impacts policies, services, and programs have on each gender, as well as various personality and behaviour tendencies, such as bargaining power. Furthermore, the newly created database containing cleaned, standardized, and gender enhanced Sunshine List data will be made publicly available to all researchers who may wish to study Ontario’s public sector further.
Bibliography


[22] J. Merrill, “Guess a persons gender by their first name.”


Appendix A

Weighted Probability Approach

The weighted probability approach in [41] describes a method that takes into account the evolution of an entity over time. In other words, it provides a way to link records over time, even if the records do not have exact matches. To create longitudinal links using this approach, the model must be trained. For example, the sample records in Table A.1 refer to three real-world entities (labelled data): i) Entity 1 (E1) is represented by r1, ii) Entity 2 (E2) is represented by r2, r3, r4, r5, and r6, and iii) Entity 3 (E3) is represented by r7, r8, r9, r10, and r11.

The term training refers to calculating the full life spans and partial life spans for each attribute within each record found in the labelled data. The full and partial life spans are calculated using Equations A.1 and A.2.

<table>
<thead>
<tr>
<th>Entity</th>
<th>ID</th>
<th>Name</th>
<th>Affiliation</th>
<th>Co-authors</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>r1</td>
<td>Xin Dong</td>
<td>R. Polytechnic Institute</td>
<td>Wozny</td>
<td>1991</td>
</tr>
<tr>
<td>E2</td>
<td>r2</td>
<td>Xin Dong</td>
<td>Univ of Washington</td>
<td>Halevy, Tatarinov</td>
<td>2004</td>
</tr>
<tr>
<td>E2</td>
<td>r3</td>
<td>Xin Dong</td>
<td>Univ of Washington</td>
<td>Halevy</td>
<td>2005</td>
</tr>
<tr>
<td>E2</td>
<td>r4</td>
<td>Xin Luna Dong</td>
<td>Univ of Washington</td>
<td>Halevy, Yu</td>
<td>2006</td>
</tr>
<tr>
<td>E2</td>
<td>r5</td>
<td>Xin Luna Dong</td>
<td>AT&amp;T Labs-Research</td>
<td>Das Sarma, Halevy</td>
<td>2009</td>
</tr>
<tr>
<td>E2</td>
<td>r6</td>
<td>Xin Luna Dong</td>
<td>AT&amp;T Labs-Research</td>
<td>Naumann</td>
<td>2010</td>
</tr>
<tr>
<td>E3</td>
<td>r7</td>
<td>Dong Xin</td>
<td>Univ of Illinois</td>
<td>Han, Wah</td>
<td>2004</td>
</tr>
<tr>
<td>E3</td>
<td>r8</td>
<td>Dong Xin</td>
<td>Univ of Illinois</td>
<td>Wah</td>
<td>2007</td>
</tr>
<tr>
<td>E3</td>
<td>r9</td>
<td>Dong Xin</td>
<td>Microsoft Research</td>
<td>Wu, Han</td>
<td>2008</td>
</tr>
<tr>
<td>E3</td>
<td>r10</td>
<td>Dong Xin</td>
<td>Microsoft Research</td>
<td>Chaudhuri</td>
<td>2009</td>
</tr>
<tr>
<td>E3</td>
<td>r11</td>
<td>Dong Xin</td>
<td>Microsoft Research</td>
<td>Ganti</td>
<td>2010</td>
</tr>
</tbody>
</table>
\[ \text{Full lifespan} = \text{Attribute.endYear} - \text{Attribute.startYear} \quad (A.1) \]

\[ \text{Partial lifespan} = \text{Attribute.endYear} - \text{Attribute.startYear} + \delta, \quad (A.2) \]

where \( \delta = \) one time unit (1)

The full and partial life spans of an attribute are the lengths of time that we can say a specific value for an attribute has a known end and start year and the lengths of time that a specific value for an attribute has a known start year but not end year, respectively. Fig. A.1 presents the evolution of the \textit{affiliation} attribute for the entities listed in Table A.1.

Figure A.1: Full and partial life spans for the affiliation attribute [41].

From Fig. A.1, entity E1 has an \textit{affiliation} value of ‘R. Polytechnic Institute’, which has a known start year, 1991; however, unknown end year. Therefore, this
is a partial life span of 1 (i.e., 1991-1991 + δ(1) = 1). Similarly, entity E2 has an affiliation value of ‘Univ of Washington’, a start year of 2004 and an end year of 2009. Because there is a known end year, this is a full life span of 5 (i.e., 2009 - 2004). The process of calculating life spans is presented in Algorithm 8.

**Algorithm 8 Calculate full and partial life spans**

1: $L_f = [], L_p = []$
2: sort records in C in increasing time order $r_1,...,r_{|C|}$
3: for $C \in \text{labelledData}$ do
4:   start = 0
5:   while start $\leq |C|$ do
6:     end = start + 1
7:     while $record_{start}.\text{Attribute} = r_{end}.\text{Attribute}$ and end $\leq |C|$ do
8:       end += 1
9:     end while
10:    if end $> |C|$ then
11:       insert $record_{end}.\text{time} - record_{start}.\text{time} + \delta$ into $\bar{L}_p$
12:    else
13:       insert $record_{end}.\text{time} - record_{start}.\text{time}$ into $\bar{L}_f$
14:    end if
15:   start = end
16: end while
17: end for

In practice, the labelled data used in Algorithm 8 contains clusters of records, where each cluster represents a single real world entity. The records in each cluster are first sorted in increasing time order with respect to each record’s year attribute. The for loop iterates over each cluster in the labelled data. From lines 4 and 6, variables $start$ and $end$ are index positions within a cluster, and refer to the start and end years of an attribute value, respectively. The inner while loop iterates over each record within a cluster in increasing time order and increases the value of $end$ by 1. If the values in record[start].Attribute and record[end].Attribute are different, or if the value of $end$ is greater than the number of records in the cluster (i.e., no observed change in attribute value), the algorithm breaks out from the the inner while loop. If
the value in end is greater than the number of records, a partial life span is calculated (i.e., record[end].year - record[start].year + δ), otherwise a full life span is calculated (i.e., record[end].year - record[start].year). It should be noted that different full and partial life spans are calculated for each attribute. The complexity of Algorithm 8 is $O(nm_j)$, where $n$ is the number of clusters, $m$ is the number of records in each cluster, and $j$ is the number of attributes.

Using the full and partial life spans, two probabilities are calculated for each attribute. The first is the probability of another record sharing the same value for the attribute, known as agreement decay, and the second is the probability of an attribute changing its value, known as disagreement decay [41]. These two probabilities are used to determine if two records map to the same real-world entity and are defined in [41] by Equations A.3 and A.4, respectively, where $L_f$ and $L_p$ denote full and partial life spans.

$$AgreementDecay(\text{Attribute}, \Delta t) = \frac{|\{l \in L_f | l \leq \Delta t\}|}{|L_f| + |\{l \in L_p | l \geq \Delta t\}|}$$ (A.3)

$$DisagreementDecay(\text{Attribute}, \Delta t) = \frac{\Sigma_{l \in L_f, l \leq \Delta t} p(l)}{\Sigma_{l \in L_f} p(l) + |\{l \in L_p | l \geq \Delta t\}|}$$ (A.4)

When calculating the agreement and disagreement decay for each attribute, the reference to full and partial life spans in Equations A.3 and A.4, refers to the full and partial life spans previously calculated with Algorithm 8. The agreement and disagreement decay are combined with various similarity functions to cluster new unobserved data. If the similarity results are greater or equal to a similarity threshold, the records are clustered together and are assumed to map to the same real-world
entity [41]. Clustering similar records using the weighted probability approach is presented in Algorithm 9.

Algorithm 9 Weighted probability approach clustering

1: allRecords[] ← loadRecords()
2: allClusters[]
3: maxSim ← (0, 0)
4: for record[i] in allRecords do
5:     for cluster[j] in allClusters do
6:         //Computes record cluster similarity
7:         similarity ← sim(record[i], cluster[j])
8:         if similarity > maxSim then
9:             maxSim ← (similarity, j)
10:     end if
11: end for
12: if maxSim[0]) ≥ Θ then
13:     insert record into cluster[maxSim[1]]
14: else
15:     //Create new cluster with record[i]
16:     insert cluster{record[i]} into clusters
17: end if
18: end for

In Algorithm 9, the routine loadRecords loads all of the unclustered data into memory (allRecords) and initializes maxSim to 0. The outer for loop iterates over each record in allRecords and the inner for loop iterates over each cluster in allClusters. During each iteration of the inner for loop, the record-cluster similarity (sim(record, cluster)\(^1\)) is computed for record[i] and each cluster[j], where i and j denote the index position of a record and cluster, respectively. The maximum similarity and the index of the cluster involved in generating the maximum similarity are stored in memory (maxSim). If the maximum similarity (maxSim[0]) is greater than the similarity threshold (Θ), record[i] is inserted into cluster[maxSim[1]]; otherwise, a new cluster is created with record [i] and is inserted into allClusters. The complex-

\(^1\)The similarity calculated at line 5 involves the agreement and disagreement decay for each attribute, and various other similarity functions [41]
ity of Algorithm 9 is $O(nmjk)$, where $n$ is the number of records in $allRecords$, $m$ is the number of clusters (real-world entities found), $j$ is the number of records in each cluster, and $k$ is the number of attributes.
Appendix B

Given Names Classified as ‘Unknown’

Table B.1 contains the given names within the evaluation data (University of Guelph) that were classified as unknown by the Relative Name Frequency technique.

Table B.1: Given names classified as ‘Unknown’ (Relative name frequency technique).

<table>
<thead>
<tr>
<th>Given Name</th>
<th>Given Name</th>
<th>Given Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdollah</td>
<td>Hwan-suk</td>
<td>Rakhal</td>
</tr>
<tr>
<td>Adronie</td>
<td>Jayasankar</td>
<td>Ruediger</td>
</tr>
<tr>
<td>Altaf</td>
<td>Jinzhong</td>
<td>Shau</td>
</tr>
<tr>
<td>Alun</td>
<td>Jnanankur</td>
<td>Smaro</td>
</tr>
<tr>
<td>Bahram</td>
<td>Jnanankhur</td>
<td>Sujeevan</td>
</tr>
<tr>
<td>Baozhong</td>
<td>Joubert</td>
<td>Sukhpal</td>
</tr>
<tr>
<td>Boyer</td>
<td>Julang</td>
<td>Sunghwan</td>
</tr>
<tr>
<td>Byram</td>
<td>Khosrow</td>
<td>Thanasis</td>
</tr>
<tr>
<td>De-tong</td>
<td>Krassimir</td>
<td>Theodorus</td>
</tr>
<tr>
<td>Devakanand</td>
<td>Krista-britt</td>
<td>Towhidul</td>
</tr>
<tr>
<td>Dilip</td>
<td>Loong-tak</td>
<td>Wanhong</td>
</tr>
<tr>
<td>Doraiswamy</td>
<td>Madhur</td>
<td>Wlodzimier</td>
</tr>
<tr>
<td>Durda</td>
<td>Manjusri</td>
<td>Xiao-rong</td>
</tr>
<tr>
<td>Fangju</td>
<td>Matthijs</td>
<td>Xining</td>
</tr>
<tr>
<td>Fei</td>
<td>Medhat</td>
<td>Yiguo</td>
</tr>
<tr>
<td>France-Isabelle</td>
<td>Mieso</td>
<td>Yong-mei</td>
</tr>
<tr>
<td>Getu</td>
<td>Nonita</td>
<td>Yongbo</td>
</tr>
<tr>
<td>Gopinadhan</td>
<td>Pevneesh</td>
<td>Youbin</td>
</tr>
<tr>
<td>Harjinder</td>
<td>Radhey</td>
<td>Zvonimir</td>
</tr>
<tr>
<td>Hongde</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C

Gender Inference Summary Statistics

Tables C.1 and C.2 contain the gender inference summary statistics for all the records on the Sunshine List and the unique given names found on the Sunshine List, respectively.

Table C.1: Gender inference statistics for all records in the Sunshine List

<table>
<thead>
<tr>
<th>Statistic Name</th>
<th>Statistic Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of records</td>
<td>652,804</td>
<td>—</td>
</tr>
<tr>
<td>Total number of records with initials only</td>
<td>4,570</td>
<td>0.7</td>
</tr>
<tr>
<td>Total number of records with a full given name</td>
<td>648,234</td>
<td>99.3</td>
</tr>
<tr>
<td>Total number of records classified by the relative name frequency technique</td>
<td>616,194</td>
<td>94.4</td>
</tr>
<tr>
<td>Total number of records classified by the heuristic technique</td>
<td>32,040</td>
<td>4.9</td>
</tr>
<tr>
<td>Total number of records classified by the hybrid system (probability threshold set at 0.95)</td>
<td>617,016</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Table C.2: Gender inference statistics for unique given names in the Sunshine List

<table>
<thead>
<tr>
<th>Statistic Name</th>
<th>Statistic Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of unique given names</td>
<td>14,967</td>
<td>—</td>
</tr>
<tr>
<td>Total number of unique given names classified by the relative name frequency technique</td>
<td>8618</td>
<td>57.6</td>
</tr>
<tr>
<td>Total number of unique given names classified by the heuristic technique</td>
<td>6349</td>
<td>42.4</td>
</tr>
<tr>
<td>Total number of unique given names classified by the hybrid system (probability threshold set at 0.95)</td>
<td>13,946</td>
<td>93.2</td>
</tr>
</tbody>
</table>