



UNIVERSITÀ DEGLI STUDI DI TRENTO

MASTER'S THESIS

Using Simpson's Paradox as a guide for the Assessment of the Quality of Integrated Data

Author:
Barlow KASULE

Supervisor:
Prof. Enrico BLANZIERI

*A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the*

Department of
Information Engineering and Computer Science

October 5, 2018

Declaration of Authorship

I, Barlow KASULE, declare that this thesis titled, “Using Simpson’s Paradox as a guide for the Assessment of the Quality of Integrated Data” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master’s degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“If you want to live a good life, hard work must be your best friend.”

Barlow Kasule

Abstract

Data quality is such an important aspect in determining how ethical, legitimate or trustworthy a certain data segment is. In a world where the amounts of data keep on rising every single day, the need to integrate this data that originates from several sources has become a necessity. The integration depending on the used method, does not come free of errors and this circumstance has created the need for data integration techniques that eliminate as much as possible any conflicts that arise during the process. These data integration techniques have been developed under accepted standard procedures, by firms and individuals well respected in the world of software development. They are also developed by teams or committees that usually comprise of data analytic professionals and other stakeholders. However, these data integration techniques and the integrated data as the end product, are not one hundred percent error proof and this is where we come in with our assessment criteria that is intended to check whether no matter the used data integration technique, there might be fault or contradictions in the conclusions drawn from the integrated data when compared to its sources.

The aim of this thesis is to propose a possible guidance assessment criteria, that can be applied after the data integration process while using some of the common data integration techniques, to see whether paradoxical results appear that would compromise the quality of conclusions drawn from the integrated data hence creating a possible quality issue of the data as a whole.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Enrico Blanzieri for his motivation and guidance during this research undertaking. I would also like to thank the Almighty God for His grace towards me, and my family and friends for their support through this entire process. I will be forever grateful. . .

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	ix
1 Introduction	1
1.1 Introduction	1
1.2 Problem	2
1.3 Motivation	2
1.4 Objective	3
2 Literature Review	5
2.1 Simpsons paradox	5
2.1.1 Example 1: Gender bias at UC Berkeley.	5
2.1.2 Example 2: Death sentence in Southern USA	6
2.1.3 Simpsons paradox and causality	7
2.1.4 Simpsons paradox and collapsibility	8
2.1.5 Simpson's paradox and Confounders	8
2.1.6 Short conclusion	8
2.2 Data Integration	9
2.2.1 The need for Data Integration	9
2.2.2 Data Integration Techniques	10
2.3 Data Quality Assessment Metrics / Dimensions	11
2.3.1 State of the Art of Data Quality Assessment Metrics	11
Further discussions:	11
Core Quality Metrics	12
Other Data Quality Metrics / Dimensions	16
3 Research Baseline	17
3.1 Understanding the paradox in relation to Data Mining	17
3.2 Algorithm for discovering instances of Simpson's Paradox in Flat Data	18
4 Over View	21
4.1 Over view of some Integrated Data Quality Assessment Metrics	21
5 Methodology	23
5.1 Proposed Simpson's Paradox based Data Quality Assessment Metric	23
5.2 Our Algorithm	24
5.3 Experiments	24
5.3.1 The Synthetic Data set experiment:	24

6 Case Study	27
6.1 The Fatal Force Data	27
6.2 Experimentation process	28
6.2.1 Running Algorithm 2 on the (Post, 2018) datasets:	28
6.2.2 Results:	29
The other 52 Models' results	29
6.3 Validation	31
6.3.1 Measure of acceptance	32
6.3.2 Conclusion	32
7 Conclusion	33
7.1 What we have done so far	33
7.2 The way forward	33
7.3 Conclusion	34
Bibliography	35

List of Tables

2.1	Showing number of applicants admitted	6
2.2	Showing number of applicants admitted broken down by department . . .	6
2.3	Showing percentage of death sentence w.r.t to victim's race	7
2.4	Table for those cases involving white victims :	7
2.5	Table for those cases involving black victims :	7
5.1	Statistics of the Synthetic dataset	25
5.2	Showing the α , β_a and β_b variables from the synthetic dataset	25
5.3	Linear regression model results of the α and β_a variable combinations	25
5.4	Linear regression model results of the α , β_a and β_b variable combinations from the synthetic dataset	26
6.1	Showing the number of α , β_a and β_b variables	28
6.2	Linear regression model results from the Post dataset	29
6.3	Algorithm results for the $\beta_b = \text{threat_level}$	30
6.4	Algorithm results for the $\beta_b = \text{flee}$	30
6.5	Algorithm results for the $\beta_b = \text{race}$	31

Chapter 1

Introduction

1.1 Introduction

Quality of data in any organisation is of great significance. With a well-structured, trustworthy and well-consolidated data structure, organisations find it easy to do efficient and effective work basing on well-organised and well-sampled data. The world is currently in what is termed as the **information age** also known as the **Computer age** or **Digital age**, which is a period in human history characterized by the shift from traditional industry that the Industrial Revolution brought through industrialization, to an economy based on information technology (Contributors, 2018).

With large amounts of data accumulating each single day in form of **Big Data**, organisations have resorted to combining accumulated data from several sectors into a one integrated database creating what we call **Data Warehouses**. These are central repositories of integrated data from one or more disparate sources. This accumulation and combination of data from the different sources has come with both benefits and challenges. Social data analysis presents multi-faceted challenges that existing data mining approaches are not well-equipped to handle. These are in form of irregularities that arise due to differing statistics exhibited between the accumulated integrated data and the sparse sources that the integrated data was obtained from. Real-world data is noisy and sparse. To uncover hidden patterns, scientists aggregate data over the population, which tends to improve statistics and signal-to-noise ratio. However, real-world data is also heterogeneous, i.e., composed of subgroups that vary widely in size and behaviour. This heterogeneity can severely bias analysis of social data (Alipourfard, Fennell, and Lerman, 2018).

The Simpsons paradox is one example that presents such a bias. Observed associations or attributes in data that has been aggregated over an entire population may be quite different from, and even opposite to, those of the underlying subgroups. Failure to take this effect into account can distort conclusions drawn from data. Simpsons paradox also affects analysis of trends. When measuring how an outcome changes as a function of an independent variable, the characteristics of the population over which the trend is measured may change as a function of the independent variable due to survivor bias. As a result, the data may appear to exhibit a trend, which disappears or reverses when the data is disaggregated (Alipourfard, Fennell, and Lerman, 2018). Taking for example from literature (Romero, Meeder, and Kleinberg, 2011), it may appear that often exposures to information or hashtags on social media make an individual less likely to share it with their followers. However, this is far from the truth: the more trending a hashtag, the more it will be shared among circles of different people and their followers.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

However, those who follow many others (and are likely to be exposed to a meme or a hashtag multiple times) are less responsive overall, due to the high volume of information they receive (Hodas and Lerman, 2012). Their reduced susceptibility to exposures biases the aggregate response, leading to wrong conclusions about behaviour. Once disaggregated based on the volume of information received, a clearer pattern of exposure response emerges, one that is more predictive of the actual response (Alipourfard, Fennell, and Lerman, 2018)(Hodas and Lerman, 2014). The unfortunate truth is that despite the overwhelming evidence accumulated of Simpsons paradox affecting inference of trends in social and behavioural data, researchers do not routinely test for it in their studies.

Existence of a Simpsons paradox in a given data set could signify noticeable patterns, including important behavioural differences. The paradox shows that subgroups within the aggregated portion in contrast, precisely diverge in their characteristics, and these differences are large enough to affect aggregate trends. In such cases the trends discovered in disaggregated data are more likely to describe and predict individual behaviour than the trends found in aggregated data (Alipourfard, Fennell, and Lerman, 2018).

Thus, to build more robust models of behaviour, data scientists need to identify the subgroups within their data or risk drawing wrong conclusions.

1.2 Problem

Among the many reasons for doing Data Integration in an organisation, one of the most important of them all is to make meaningful conclusions out of it. We need to make **accurate** and **precise** conclusions that are consistent in both the disaggregated [sub-groups] data and the integrated [grouped] data. This could be simple if there were no paradoxical tendencies that can easily appear in one way or another. As the world of Data keeps growing each day that passes by, this explosion of data availability and the advances in computing power and processing tools and software have paved the way for statistical modelling to be at the front and centre of decision making not just in business, but everywhere. Anomalies and misinterpretations lurk behind even the simplest of statistical operations, providing implications that operations in business analytics can produce deceptive results if not performed properly. Simpsons Paradox, which regularly crops up in statistical research, business analytics, and public policy, is a prime example of why statistical analysis is useful as a corrective for the many ways in which humans intuit false patterns in complex datasets (Berman et al., 2012).

Therefore, there is need for a solution that could explain in any case if such fallacies would be encountered in a certain data integration and hence guide an organise on how to draw conclusions from the data at hand.

1.3 Motivation

Considering Moore's Law (stating that computing power is increasing rapidly) and the technologies to look through vast quantities of data is improving each year, our analytical capabilities and in turn insights are starting to grow exponentially and will soon change organizations to become more data driven and less "business instinct" driven (Macsin and Abate, 2017). Data is going to become the next global currency and is already being globally monetized by corporations. This turn of events mean more integration of the several data segments that organisations and enterprises will have accumulated in pursue of making meaningful conclusions out of these chunks of data.

Unofficial sources claim that some experts say that big data is dead. They argue that businesses do not even use a small portion of data they have access to and big does not always mean better. Sooner rather than later, big data will be replaced by fast and actionable data, which will help businesses, take the right decisions at the right time. Having tremendous amounts of data will not give you a competitive advantage over your competitors but how effectively and quickly you analyse the data and extract actionable information from it (Ahmed, 2017). Hence the motivation of providing a criteria of making quick analyses on data as it is being integrated and after, to make quick actionable decisions and conclusions from it, that are consistent in both data segments i.e. the disaggregate sources and the aggregated result.

1.4 Objective

The main objective of this thesis is to provide a criteria for assessing the quality of integrated data basing on the conclusions drawn from the aggregated data comparing them with the disaggregated sources, to check for occurrences of Simpson's paradoxical behaviours that would compromise the decisions taken, basing on the drawn conclusions from a given data set or data sets.

The method presented in this paper provides a simple framework for identifying such interesting subgroups by systematically searching for Simpson's paradox in trend data.

Chapter 2

Literature Review

Introduction

This chapter introduces the notions of Simpson's Paradox and Data Integration. It also illustrates details of what these two concepts mean and also presents examples that better explain the two phenomena. In details, we give an explanation of what we based on in order to carry out our research work in these fields. Furthermore, we discuss works that have already been done on data quality assessment. We consider the current state of data quality assessment metrics that we take as the foundation of our research study.

2.1 Simpsons paradox

Simpsons paradox, also known as the Yule-Simpson effect and the reversal paradox, is a phenomenon arising from multivariate statistical analyses. It is called a paradox because it often leads to paradoxical conclusions. Such conclusions may lead to the development of theories that incorporate causal effects disconnected from reality, based on empirical findings distorted by Simpsons paradox (Kock, 2015). It is sometimes given the descriptive title reversal paradox or amalgamation paradox.

Several occurrences of this effect have been recorded in social-science and medical-science statistics and is particularly problematic when frequency data is unduly given causal interpretations, "the Simpsons paradox occurs when an observed association between two random variables, say X and Y , gets reversed after considering the third variable W , called a covariate or a background variable. The situation of having two contradictory conclusions makes this phenomenon paradoxical" (Vellaisamy, 2014). In this case, the variable W is a summation or aggregation of the two previous ones i.e. X and Y .

If we dissect the data into sub-groups, each representing a specific value of the third variable, the phenomenon appears as a sign reversal between the associations measured in the disaggregated sub-groups relative to the aggregated data, which describes the Summed Group W . Many scholars however argue that this phenomenon is exclusively causal. The question is: "What causal dependencies can produce such a case, properly known as the Simpson's paradox?"

To give a good explanation of what the Simpson's paradox is, we give several examples below.

2.1.1 Example 1: Gender bias at UC Berkeley.

One of the most famous real-life examples is that of University of California Berkeley. In 1973, the University of California-Berkeley was sued for sex discrimination. The numbers

looked incriminating: the graduate schools had just accepted 44% of male applicants but only 35% of female applicants. When researchers looked at the evidence, though, they uncovered something surprising.

The admission figures categorically showed that men applying were more likely than women to be admitted, and the difference was so evident and large enough to dismiss any suggestions that it was due to chance.

	Applicants Admitted	
Men	8442	44%
Women	4321	35%

TABLE 2.1: Showing number of applicants admitted

When it came down to individual departments analysis, evidence showed that six out of the 85 departments were ideally biased against men, while just four were significantly biased against women. Truth be told, the combined and analysed data showed a "small but statistically significant bias in favour of women. The data from the six largest departments is listed below. Real department names have been altered.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

TABLE 2.2: Showing number of applicants admitted broken down by department

The research paper by (Bickel, Hammel, and O'Connell, 1975) concluded that women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry).

2.1.2 Example 2: Death sentence in Southern USA

We give another example got from the data from Radelet, 1981 (Witmer, 1992) on the relationship between race and the imposition of the death penalty for convicted first-degree murderers. Is the application of the death sentence racially motivated?

In 1978, Warren McClesky, a black man, was convicted of killing a white police officer and was sentenced to death in Georgia. In an appeal before the U.S. Supreme Court, McCleskys lawyers argued that imposition of the death penalty in Georgia was racially biased. We look here at data taken from all homicide convictions in the state of Florida between 1976 and 1980 in which the suspect was black or white and the victim was black or white. (Georgias data are very similar.)

People from the same race tend to kill each other more than the opposite race. So, in the

first table we see the stats showing that white suspects received the death sentence more than the black suspects. This is due to the fact that there were higher numbers of white suspects killing whites than there were for black suspects killing blacks. In that case, the death sentence was given more when the victim was white than when was black.

The first table shows an apparent bias, but in favour of black suspects. The lurking variable (confounder variable) here is the victim's race, as the subsequent tables show.

		YES	NO	TOTAL	(%YES)
Race of Suspect	White	39	308	347	11.2
	Black	32	345	377	8.5
	Total	71	653	724	9.8

TABLE 2.3: Showing percentage of death sentence w.r.t to victim's race

Analysing this table will give an impression that the death sentence was given more to white suspects when looking at the percentages. In the subsequent tables we take into consideration the confounding variable (victim's race) to show that the statistics were misleading to the wrong conclusion and thus bringing about the Simpsons paradox.

		YES	NO	TOTAL	(%YES)
Race of Suspect	White	39	279	318	12.3
	Black	29	121	150	19.3
	Total	68	400	468	14.5

TABLE 2.4: Table for those cases involving **white victims**:

		YES	NO	TOTAL	(%YES)
Race of Suspect	White	0	29	29	0.0
	Black	3	224	227	1.3
	Total	3	253	256	1.2

TABLE 2.5: Table for those cases involving **black victims**:

In both the above tables we see that in reality the death sentence was given more to the black suspects than the white suspects. Here, for both black and white victims, a higher percentage of black convicted murderers are sentenced to death than are white murderers (1.3% vs. 0% and 19.3% vs 12.3% for black and white victims respectively). There is a third variable here that is hidden in these statistics. If we look at likelihood of being sentenced to death for killing whites and blacks, we see that the death sentence is more likely if the victim was white than if the victim was black.

From the above given examples, we draw a practical conclusion: When presented with results, one always must ask themselves if there were other important variables that were omitted from the analysis. Their inclusion could change the initial obvious conclusions and thus eliminate paradoxical confusion.

2.1.3 Simpsons paradox and causality

As mentioned earlier, when we look at the above examples, we can give a general uneducated conclusion that indeed for the case of the death sentence, whites were given the

death penalty more than the blacks. This would be true if we did not have more information about the whole issue. Many scholars have argued that Simpsons paradoxes appearances have causal factors, and some say not. In (Bandyopadhyay et al., 2015), they give an account that unless someone uses the notion of causation trivially, for example, believes that $2+2$ *causes* 4, there is no reason to assume that there are causal intuitions lurking in the background. But we have seen that from the examples above, if it were not for some lurking variables, the initial results would have been indisputable. More so in the death sentence example, one can argue that the lurking variable in this case would be **race of both victim and perpetrator**, that led to the sentencing of more black convicts than their white counter parts.

Therefore, one must take notice that in some cases there are paradoxical results when certain types of data are combined and hence bringing about the Simpsons paradox. And yet in other cases there are not such variables. So, for one to conclude that the Simpsons paradox is causal, must look closely at what type of data they are dealing with, its source, the way it was collected, the way it was integrated, the motivation behind its collection and so on. In other words, it is not wise to conclude that Simpsons paradox is either causal or not without taking into consideration many other factors around that dataset that leads to paradoxical results.

2.1.4 Simpsons paradox and collapsibility

The concept of collapsibility goes hand in hand with that of Simpsons paradox. Where when one occurs, the other does not and vice versa. (Vellaisamy, 2014), the examples and the applications discussed in their paper clearly demonstrated that Simpsons paradox is a crucial aspect in the data analysis and the issue of collapsibility should be looked into only after ascertaining the non-occurrence of Simpsons paradox. They further go on to mention, that generally, whenever Simpsons paradox does not occur, the collapsibility issue arises naturally as a dimension reduction problem in the context of data analysis.

2.1.5 Simpson's paradox and Confounders

As seen in the previous sections, correlations, in which two variables appear to have [that is to say] a positive correlation towards one another, when in fact they have a negative correlation, the reversal having been brought about by a "lurking" variable or variables termed as *confounders*, could lead to the appearance of the paradox. (Berman et al., 2012) give an example from economics, where a dataset suggests overall demand is positively correlated with price, in contradiction of expectation. Analysis reveals time to be the confounding variable: plotting both price and demand against time reveals the expected negative correlation over various periods, which then reverses to become positive if the influence of time is ignored by simply plotting demand against price.

2.1.6 Short conclusion

Having seen how deceiving statistical measures can be, in any given dataset, at times producing paradoxical conclusions, there is need to further study this phenomenon and come up with guidelines that can help warn a data analyst if it happens or even avoid it entirely.

2.2 Data Integration

Data Integration by definition, is the process of retrieving data from multiple source systems and combining it in such a way that it can yield consistent, comprehensive, current and correct information for business reporting and analysis. The source systems may be several types of devices and the data may be in a variety of formats.

According to (Goodhue, et al.,1992), data integration will increase along one or both of two dimensions: (1) the number of fields with common definitions and codes, or (2) the number of systems or databases adhering to these standards. Data integration is an example of a highly formalized language for describing the events occurring in an organization's domain. The scope of data integration is the extent to which that formal language is used across multiple organizations or sub-units of the same organization. The objective of data integration is to bring together data from multiple data sources that have relevant information contributing to the achievement of the users' goals (AFT, 1997).

We can go ahead and identify some of the factors which must be or should be considered to carry out a successful and proper data integration:

- **Having excellent quality data**
Any Data Integration project or process should possess a data quality plan that focuses on quality before (i.e. selecting ideal subgroupings of the available data sources for integration), during (i.e. estimation of noise and distortion levels that could arise due to sensory, processing, and environmental surroundings during its collection) and after (i.e. the computational efficiency of the integrated data sets to achieve the goals of the users) the integration has been implemented. Without this, the project will inevitably and most likely fail. Conclusively, good data quality is what will ensure user-adoption and therefore leading to proper accomplishment of any Data Integration project.
- **Adopting a centralised approach**
Data must be harmonized in an automated and reliable way across all platforms for an organisation to have one version of them all. Errors caused by inconsistent data and manual data entry can prove very costly and could distort proceedings.
- **The formats of the data, the archive systems, and the data storage and retrieval**
It should be noted that for a smooth Data integration process, the subset sources should have a harmonious format that is general to all, be accessible with ease while also being easy to store and retrieve at any given instance.
- **Taking future changes into consideration**
It is very important to cast an eye in the future when it comes to integration of several data sources. Why? Most integration approaches are rarely developed with a solid long-term future-plan available. Take an instance when the integrated systems get upgraded, and an organisation wishes to extend the use of their integration tool and integrate with other systems. What would happen then? Any chosen Data Integration solution, should have a future-proofing mechanism and can keep operating when the integration circumstances change.

2.2.1 The need for Data Integration

Data Integration is literally done in all fields of specialization, from Mathematica to Biological to statistics and several others. Arguably, the most well-known implementation of

data integration is building an enterprise's Data Warehouse.

Several other well-known implementations include: Data migration, Enterprise application or information integration, Master data management, among many others.

There are several benefits that come with Data Integration from whichever field it is applied to. There are some obvious advantages in integrating information from multiple data sources. Such integration alleviates the burden of duplicating data gathering efforts, and enables the extraction of information that would otherwise be impossible (Subrahmanian, et al., 1996)

These benefits be listed as:

- Minimizing inconsistencies in data from the various sources it is acquired
- Allowing for effective and efficient segmentation of the data into actionable solicitation groups
- Integrated decision making for an organisation; to: Improving decision making, user experience and streamlining operations in an organisation.

To mention but a few, and many other benefits depending on the field in which application of the Integration of Data is done or needed.

2.2.2 Data Integration Techniques

Presently, there are three main techniques used to carry out data integration, These include: *consolidation*, *federation*, and *propagation*. One can argue that there are others, for example the *Hybrid* approach, but this comes from the three above, as well as all the other ones that have been developed.

We give a detailed explanation of these techniques, adopted from (White, 2018)

- **Data Consolidation**

Captures data from multiple source systems and integrates it into a single persistent data store. This data store may be used for reporting and analysis as in data warehousing, or it can act as a source of data for downstream applications as in an operational data store.

With data consolidation, there is usually a delay, or latency, between the time updates occur in source systems and the time those updates appear in the target store. Depending on business needs, this latency may be a few seconds, several hours, or many days. The term near real time is often used to describe target data that has a low latency of a few seconds, minutes, or hours. Data with zero latency is known as real-time data, but this is difficult to achieve using data consolidation.

- **Data Federation**

Provides a single virtual view of one or more source data files. When a business application issues a query against this virtual view, a data federation engine retrieves data from the appropriate source data stores, integrates it to match the virtual view and query definition, and sends the results to the requesting business application. By definition, data federation always pulls data from source systems on an on-demand basis. Any required data transformation is done as the data is retrieved from the source data files. Enterprise information integration (EII) is an example of a technology that supports a federated approach to data integration.

- **Data Propagation**

Applications copy data from one location to another. These applications usually operate online and push data to the target location; i.e., they are event-driven. Updates to a source system may be propagated asynchronously or synchronously to the target system. Synchronous propagation requires that updates to both source and target systems occur in the same physical transaction. Regardless of the type of synchronization used, propagation guarantees the delivery of the data to the target. This guarantee is a key distinguishing feature of data propagation. Most synchronous data propagation technologies support a two-way exchange of data between a data source and a data target. Enterprise application integration (EAI) and enterprise data replication (EDR) are examples of technologies that support data propagation.

2.3 Data Quality Assessment Metrics / Dimensions

First, what do we mean by Data Quality Assessment? We can define this as *"the process of scientifically and statistically evaluating data in order to determine whether they meet the quality required for projects or business processes and are of the right type and quantity to be able to actually support their intended use"* (Techopedia, 2018). A metric or dimension on the other hand is a standard of measurement that serve as an integral component of accountability.

When we combine the two together, we can get an understanding that Data Quality metrics or dimensions are important because they enable people to understand why data is being measured. Metrics are needed for quantifying Data Quality in order to answer questions like: Which measure improves Data Quality the most and which one has the best benefit-to-costs ratio?

2.3.1 State of the Art of Data Quality Assessment Metrics

Here we exhaustively assess existing data quality metrics, which are in line with the objectives of the thesis.

In this section, the current state of the metrics that are used in Data Quality assessment is evaluated and discussed. In addition, the requirements for good quality assessment metrics are also discussed. (Höchtel et al., 2015) gives us a lot of detailed explanations of the different metrics and their categorises. What is mentioned here helps to provide a basis for the development and implementation of our proposed data quality assessment metric.

Further discussions:

Data quality metrics or in other cases dimensions, is a term that has been widely used for a number of years to describe the measure of the quality of data. However, the key quality metrics or dimensions are still up to-date debatable even amongst data quality professionals. The key data quality dimensions are not universally agreed upon, but there are those that have appeared almost in every search result for data quality metrics or dimensions. This state of affairs has led to much confusion within the data quality community and is even more confusing for those who are less familiar in the field let alone business stakeholders.

We examine a set of standard indicators that are routinely reported through out organisations' information systems and quantify problems of data completeness, timeliness, consistency and accuracy. Those mentioned, we can say, are the core parameters on which all

the other quality metrics are built. We explain and detail these core metrics and also give brief definitions and descriptions on the others in the following sections.

Core Quality Metrics

Organisations select the data quality dimensions and associated dimension thresholds based on their business context, requirements, levels of risk etc. Note that each dimension is likely to have a different weighting and in order to obtain an accurate measure of the quality of data, the organisation will need to determine how much each dimension contributes to the data quality as a whole (Group, 2013).

The six core metrics or dimensions of data are:



- **Completeness Metric / Dimension:**

Title	Completeness
Definition	The proportion of stored data against the potential of "100% complete"
Reference	Business rules which define what "100% complete" represents.
Measure	A measure of the absence of blank (null or empty string) values or the presence of non-blank values.
Scope	0-100% of critical data to be measured in any data item, record, data set or database
Unit of Measure	Percentage
Type of Measure: Assessment, Continuous, Discrete	Assessment only
Related dimension	Validity and Accuracy
Optionality	If a data item is mandatory, 100% completeness will be achieved, however validity and accuracy checks would need to be performed to determine if the data item has been completed correctly

- **Uniqueness Metric / Dimension:**

Title	Uniqueness
Definition	No thing will be recorded more than once based upon how that thing is identified.
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of the number of things as assessed in the 'real world' compared to the number of records of things in the data set. The real world number of things could be either determined from a different and perhaps more reliable data set or a relevant external comparator.
Scope	Measured against all records within a single data set
Unit of Measure	Percentage
Type of Measure: Assessment, Continuous, Discrete	Discrete
Related dimension	Consistency
Optionality	Dependent on circumstances

- **Timeliness Metric / Dimension:**

Title	Timeliness
Definition	The degree to which data represent reality from the required point in time.
Reference	The time the real world event being recorded occurred.
Measure	Time difference
Scope	Any data item, record, data set or database.
Unit of Measure	Time
Type of Measure: Assessment, Continuous, Discrete	Assessment and Continuous
Related dimension	Accuracy because it inevitably decays with time.
Optionality	Optional dependent upon the needs of the business.

- **Validity Metric / Dimension:**

Title	Validity
Definition	Data are valid if it conforms to the syntax (format, type, range) of its definition.
Reference	Database, metadata or documentation rules as to the allowable types (string, integer, floating point etc.), the format (length, number of digits etc.) and range (minimum, maximum or contained within a set of allowable values).
Measure	Comparison between the data and the metadata or documentation for the data item.
Scope	All data can typically be measured for Validity. Validity applies at the data item level and record level (for combinations of valid values).
Unit of Measure	Percentage of data items deemed Valid to Invalid.
Type of Measure: Assessment, Continuous, Discrete	Assessment, Continuous and Discrete
Related dimension	Accuracy, Completeness, Consistency and Uniqueness
Optionality	Mandatory

- **Accuracy Metric / Dimension:**

Title	Accuracy
Definition	The degree to which data correctly describes the "real world" object or event being described.
Reference	Ideally the "real world" truth is established through primary research. However, as this is often not practical, it is common to use 3rd party reference data from sources which are deemed trustworthy and of the same chronology.
Measure	The degree to which the data mirrors the characteristics of the real world object or objects it represents.
Scope	Any "real world" object or objects that may be characterised or described by data, held as data item, record, data set or database
Unit of Measure	The percentage of data entries that pass the data accuracy rules.
Type of Measure: Assessment, Continuous, Discrete	Assessment, e.g. primary research or reference against trusted data. Continuous Measurement, e.g. age of students derived from the relationship between the students dates of birth and the current date. Discrete Measurement, e.g. date of birth recorded.
Related dimension	Validity is a related dimension because, in order to be accurate, values must be valid, the right value and in the correct representation.
Optionality	Mandatory because - when inaccurate - data may not be fit for use.

- **Consistency Metric / Dimension:**

Title	Consistency
Definition	The absence of difference, when comparing two or more representations of a thing against a definition
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of pattern and/or value frequency.
Scope	Assessment of things across multiple data sets and/or assessment of values or formats across data items, records, data sets and databases. Processes including: people based, automated, electronic or paper.
Unit of Measure	Percentage.
Type of Measure: Assessment, Continuous, Discrete	Assessment and Discrete.
Related dimension	Validity, Accuracy and Uniqueness
Optionality	It is possible to have consistency without validity or accuracy.

Other Data Quality Metrics / Dimensions

Considering the fore mentioned metrics, most of the other metrics are derived from the core ones in the section above. Data Quality Metrics can be formed according to an organisation's needs or generically. Most of the metrics out there are put into place due some need and below, is a table that includes some of them:

Metric	Description
Believability	Degree to which data is seen as credible and true
Relevancy	How usable, applicable, or interesting the data is
Reputation	Data contents or source are kept in high consideration
Appropriate amount of data	How suitable is the amount of data
Ease of manipulation	The extent to which data is easy to manipulate and apply to different tasks
Objectivity	How impartial the data is
Value-added	Data provides a competitive advantage
Interpretability	The extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Definition conformance / Documentation / Usage	The extent to which the data meaning is explained
(Access) Security	Access is secure or can be restricted
Conciseness	Data is compactly represented
Accuracy	Data is correct (free of errors) and reliable
Accessibility / Retrievability	The degree to which the data is retrievable
Identifiability	ease by which data can be found
Contactability	Possibility to give feedback on data The extent to which the data publisher provides contact information

Conclusion: As seen from the above literature, the Simpson's Paradox with all that goes with it, is an important phenomenon that has been extensively studied in statistics but still confusing for those that are out of the statistical world, up to present day. More sensitisation needs to be made about this phenomenon, more so in the world of data. Data being a driving force for almost every organisation today, it means that more and more people are bound to encounter such phenomenon, more especially majority of whom might not be aware of such statistical paradoxes. Integration of data being an inevitability, organisations and businesses are finding it impossible to keep their data distributed and most importantly, keep it of good quality. More research and studies are needed in areas that combine both data integration and Simpson's paradox, and this thesis is among those that are tackling this.

Chapter 3

Research Baseline

Introduction

In this section we present our research basis. We talk about the known and available data quality assessment metrics. We give an explanation to help show the relationship that is there between Simpson's paradox and Data Mining and we look at an algorithm that was developed to detect instances of Simpson's paradox and relate it to our research.

3.1 Understanding the paradox in relation to Data Mining

The past several years have seen decision making in many areas of business steadily evolve from judgement-driven domains into scientific domains in which the analysis of data and careful consideration of evidence are more prominent than ever before. Additionally, mainstream books, movies, alternative media and newspapers have covered many topics describing how fact and metric driven analysis and subsequent action can exceed results previously achieved through less rigorous methods. This trend has been driven in part by the explosive growth of data availability resulting from Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) applications and the Internet and eCommerce more generally. There are estimates that predict that more data will be created in the next four years than in the history of the planet (Berman et al., 2012).

There are several issues that could arise during a data integration process and among which, one might encounter contradictions in the data disaggregates compared to the data aggregates. In other words, differing results in form of descriptive statistics of these data segments.

We need to understand that: (i) **Reversal paradox** (ii) **Simpson's Paradox** and (iii) **Amalgamation paradox**, all mean the same thing. If we are to make claims of findings of a paradox, more so this one in question, we must meet certain criteria. First, we have to make sure that the findings explain which parameters in the dataset are leading to these conclusions and why a user considers the phenomenon surprising or unbelievable. Second, the solution must identify the class of scenarios in which the paradox may surface and distinguish it from scenarios where it will surely not surface. Finally, in those scenarios where the paradox leads to indecision, we must identify the correct answer, explain the features of the scenario that lead to that choice, and prove mathematically that the answer chosen is indeed correct.

It is important to note that, strictly speaking, Simpson's paradox is not really a paradox in the context of probability theory, because it does not involve a real contradiction, just an apparent contradiction that is eliminated by providing an explanation for the occurrence

of the paradox, as just discussed in the case of the paradox occurrence (Freitas, McGarry, and Correa, 2007)

A second important and related point is that Simpsons paradox is well known by statisticians in general. Despite these two points, it should be noted that the apparent contradiction associated with Simpsons paradox looks very puzzling (actually, it looks like a real contradiction) to most users, and, as pointed out by (Fabris and Freitas, 2006), in practice Simpsons paradox is usually very surprising to users, who typically have no formal statistical training. Hence, the discovery of Simpsons paradox instances is a valid approach for discovering surprising patterns from data, since the goal of data mining is to discover patterns that are surprising to users, rather than to statisticians or data analysts

3.2 Algorithm for discovering instances of Simpson's Paradox in Flat Data

(Fabris and Freitas, 2006) suggested and came up with two algorithms that detect surprising occurrences of Simpson's paradox in a flat data sample and in a hierarchical multidimensional data sample. A flat file database is a database that stores data in a plain text file. Each line of the text file holds one record, with fields separated by delimiters, such as commas or tabs. Multidimensional Data on the other hand can be defined as that which is comprised of multiple dimensions, such as space and time. For example, a temperature dataset could have dimensions of latitude, longitude, altitude, and time.

This section briefly describes an algorithm for discovering instances of Simpsons paradox in "flat data", stored in a single relation (corresponding to the universal relation of a database system), which is the typical kind of data mined by most data mining algorithms. This algorithm was further adopted by (Alipourfard, Fennell, and Lerman, 2018), where they investigated how Simpsons paradox affects analysis of trends in social data. According to them, the trends observed in data that has been aggregated over an entire population may be different from, and even opposite to, those of the underlying subgroups. It is very important to take such records into account because if not, they can lead analysis to wrong conclusions, which is our main goal in this thesis.

Short description of the algorithm: As in put, we have a list **LG**, which is a list containing user defined binary goals showing whether or not a certain situation occurred i.e. in our test data case, we have a situation where by a *fatality* occurred from a shooting or not. The algorithm there on proceeds to identify all the binary attributes or boolean attributes (potential "first partitioning attributes") in the data and puts them into the list **L1**. It then identifies all categorical attributes (potential "second partitioning attributes") in the data and puts them into the list **L2**.

The list **L1** can be a field in the data set that a user chooses, same for the list **L2** but condition being that the records in each of these lists must be different from each other's. The algorithm enforces the constraint that any goal attribute included in **LG** cannot be included in **L1** nor in **L2**. The output of Algorithm 1 consists of all instances of Simpsons paradox found by the algorithm.

These instances can be ranked in decreasing degree of surprisingly-ness to the user, so that a user with little available time can focus on the analyses of only the most surprising instances of the paradox.

The algorithm in question is as shown below:

Algorithm 1 Search for instances of Simpsons paradox in flat data

```

1: INPUT: list of user-defined goal attributes, denoted LG
2: Begin
3: Identify attributes that can be used as 1stPartAtt and put them in list L1
4: Identify attributes that can be used as 2ndPartAtt and put them in list L2
5: for EACH goal attribute G in LG do
6:   for EACH first partitioning attribute A in L1 do
7:     Partition population into Pop1 and Pop2 according to values of A
8:      $P(G1) = P(G = \text{"yes"} \mid A = 1)$ 
9:      $P(G2) = P(G = \text{"yes"} \mid A = 2)$ 
10:    for EACH second partitioning attribute B in L2 such that  $B \neq A$  do
11:      for  $i = 1, 2$  do
12:        Partition Popi into  $m$  new populations Popi1 ... Popim, according to the values
        of B
13:        for  $j = 1, \dots, m$  do
14:           $P(G_{ij}) = P(G = \text{"yes"} \mid A = i, B = j)$ 
15:          if (  $P(G1) > P(G2) \ \& \ P(G1_j) \leq P(G2_j), j = 1, \dots, m$  ) or (  $P(G1) < P(G2) \ \& \ P(G1_j) \geq P(G2_j), j = 1, \dots, m$  ) then
16:            OUTPUT the occurrence of the paradox to the user
17:          end if
18:        end for
19:      end for
20:    end for
21:  end for
22: end for
23: End

```

The algorithm 1 (Fabris and Freitas, 2006) above searches for surprising occurrences of Simpson's Paradox in flat data sets. It is meant to return results showing that between selected fields in a data set, taking them as Simpson pairs, there exists varying statistics from when the data is analysed as a whole to when the data is separated into subgroups.

Chapter 4

Over View

4.1 Over view of some Integrated Data Quality Assessment Metrics

Data quality is measured against a list of criteria called 'data quality dimensions', like timeliness, for instance. While ideally all of these will be of high quality, depending on what you want out of your data you may decide to prioritise some dimensions over others. So what metrics should you measure data quality on?

Organisations must deal with both the subjective perceptions of the individuals involved with the data, and the objective measurements based on the data set in question. Subjective data quality assessments reflect the needs and experiences of stakeholders: the collectors, custodians, and consumers of data products. If stakeholders assess the quality of data as poor, their behaviour towards the collected data will be influenced by this assessment. Some of the most common data quality assessment metrics as mentioned in the previous chapter are: **Completeness, Uniqueness, Timeliness, Validity, Accuracy, Consistency**, to mention but a few.

We have to know and keep in mind that our data is rarely going to be perfect, and that we have to juggle managing our data quality with actually using the data. If we were to take too much time ensuring its quality, then there would be no point analysing it, because it would be far past its sell-by date. However, we should perform regular data quality audits especially as we're probably regularly collecting new data sets, to ensure it's as clean and useful as we can make it. We need to have good data so as to rely upon it produce meaningful essential results and conclusions.

We are proposing in the following chapter, a novel metric based on the quality measure of Granularity, that will check for instances of the Simpson's paradox in an integrated data set of which set has been merged together using any of the data integration methods out there, and return warnings to the user that there are some inconsistencies in the integrated dataset [if any], hence leaving the final decision to the analyst or user to decide on what to do or not.

Chapter 5

Methodology

Introduction

In this Methodology section of the thesis, we are presenting our proposed metric that we are developing for the assessment of the Quality of Integrated Data. Detailed explanations of the whole idea are mentioned in this section and experimental results shown too.

5.1 Proposed Simpson's Paradox based Data Quality Assessment Metric

We propose an Ad Hoc method to systematically check for Simpsons paradox instances in a given dataset, considering column combinations and carrying out statistical calculations on these combinations. We are looking for possible "*trend*" or "*relationship*" reversals in the aggregated datasets comparing them with the disaggregated ones.

There are variables that we have decided to highlight of which are to be used as Simpson's pairs. Taking a variable α as the dependent variables in the dataset [that desired outcome being measured], and we choose another variable $\beta = (\beta_1, \beta_2, \dots, \beta_m)$ the set of m independent variables or features.

The objective of this is to identify pairs of variables (β_a, β_b) , in our case columns from the dataset, which we term as "Simpsons pairs", such that an observation between our dependent variable α as a function of β_a reverses or ceases to exist when the dataset is disaggregated by conditioning on β_b .

$$\begin{aligned} \frac{d}{d\beta_a} \mathbb{E}[\alpha|\beta_a] &> 0 : \forall \beta_a; \\ \frac{d}{d\beta_a} \mathbb{E}[\alpha|\beta_a, \beta_b] &\leq 0 : \forall \beta_a, \beta_b; \end{aligned}$$

The above condition is the focal point in our pursuit for the desired The Simpson's Paradox based assessment metric.

5.2 Our Algorithm

Algorithm 2 The Algorithm

```

1: Get Simpson's Pairs  $[\alpha, \beta_a]$ 
2: Make linear models of the pairs
3: Check Gradient and P-value of the model:
4: if Gradient is positive or negative & Pvalue > 0.05 then
5:   There is NO correlation, Proceed
6: else
7:   if Gradient is negative & Pvalue > 0.05 then
8:     There is A correlation, Stop and Exit!
9:   end if
10: end if
11: Disaggregate dataset w.r.t Simpson's Pairs and  $\beta_b$  a confounding variable from the set

12: Make linear models for all combinations got
13: Detect Paradox instances:
14: if (Gradient  $\beta_b < \alpha$  intercept) and (Gradient  $\beta_b > 0$ ) and (Pvalue  $\beta_b \leq 0.05$ ) then
15:   Report instance of Simpson's paradox and count it
16: else
17:   Report absence of instance of Simpson's Paradox
18: end if
19: Return Reports counts

```

5.3 Experiments

We performed several experiments in order to help validate our metric. Below are the detailed explanations of them:

5.3.1 The Synthetic Data set experiment:

Our first experimentation data was engineered in a way that it describes how Simpson's Paradox accounts for a highly surprising observation in a fatal shootings study by police in a certain part of the world.

In this experimentation, we chose a sample size of 100 people with statistics indicating three particular races i.e. Black, White and Hispanic. The idea here was to check which particular race of people commits the most crimes and then make conclusions that that particular group of people would also be the most likely to face a fatal shooting from the police of that area. In simple terms, it would mean that the higher the number of crime offenders, the higher the likeliness of them being shot.

The role of this dataset is to help validate our algorithm which is intended to trace out Simpson's Paradox instances in a given dataset. On observation, we see that the most number of people to commit crimes were whites with 37%, followed by Blacks at 35% and finally Hispanics at 28%. One would be compelled to believe that for sure the above hypothesis is valid. The initial observations of the datasets while still in disaggregates shows the following statistics:

Race	Fatality	Crime committed	Total out of 100 samples
Blacks	29	12	35
Whites	11	3	37
Hispanics	24	14	28

TABLE 5.1: Statistics of the Synthetic dataset

Explanation of the dataset: We see that there are a total sample of 100 people in the three data groups of races; Black, White and Hispanic. The total number of Blacks with Fatality are 29, Whites 11 and Hispanics 24. Then we see that the total number of Crime committed is 12 for Blacks, 3 for Whites and 14 for Hispanics. Finally, we see that the total number of all out of the sample space for Blacks is 35, Whites 37 while 28 for Hispanics.

The dataset was engineered to have a bias towards the races of Black and Hispanic in order to try and develop an algorithm to detect whether really there traces of the Simpson's paradox when the data is aggregated. From the algorithm, formulating the α variable, β_a variable and β_b variable was as follows:

α variable	β_a variables	β_b variables
Fatality:	Crime:	Race:
Yes	Yes No	Black, White and Hispanic

TABLE 5.2: Showing the α , β_a and β_b variables from the synthetic dataset

We formulated linear models for the α and β_a variables and got the results reported in the table below:

$\alpha = \text{Fatality}$	$\beta_a = \text{Crime}$
Yes	Crime = Yes gradient
0.673077	-0.06891
No	Crime = No gradient
0.6042	0.0689

TABLE 5.3: Linear regression model results of the α and β_a variable combinations

Making analysis of the results, we can see that when the fatality is YES and crime is YES too, there is no correlation at all between the two of them since we get varying signs of both values. The other side shows that there is a correlation.

Creating models as required from the above algorithm 2, basing on the results in the above table of NO correlation, the regression results reported the following outcomes:

$\alpha = \text{Fatality}$	$\beta_b = \text{Races}$	
Yes 0.5577	Black gradient 0.2856	Black P-value 0.005
Yes 0.8606	White gradient -0.5417	White P-value 0.000
Yes 0.5960	Hispanic gradient 0.3084	Hispanic P-value 0.004

TABLE 5.4: Linear regression model results of the α , β_a and β_b variable combinations from the synthetic dataset

Analysing the results, we can see that the α variable intercept in the models of Black and Hispanics races is positive. The gradients of these two races are also positive and finally their P-values are below our threshold of 0.05. This can give us a clear conclusion that there are Simpson's Paradox instances in the subgroups of Black race and Hispanic race, which lies in line with our intention that led to the engineering of the data in the first place. So this renders our algorithm 2 operational and valid to progress with in the next chapters of our work.

Chapter 6

Case Study

Introduction

In this study chapter, we take real life datasets collected by [The Washington Post in the United States of America](#). We use them as our test case scenario for the algorithm and assessment of the integration method that we used to combine them.

6.1 The Fatal Force Data

The Washington database (Post, 2018), contains records of every fatal shooting in the United States by a police officer in the line of duty since Jan. 1, 2015.

In 2015, The Post began tracking more than a dozen details about each killing including the **race** of the deceased, the circumstances of the shooting, whether the **person was armed** and whether the **person was experiencing a mental-health crisis** by culling local news reports, law enforcement websites and social media, and by monitoring independent databases such as Killed by Police and Fatal Encounters. The Post conducted additional reporting in many cases.

The Post is documenting only those shootings in which a police officer, in the line of duty, shoots and kills a civilian the circumstances that most closely parallel the 2014 killing of Michael Brown in Ferguson, Mo., which began the protest movement culminating in Black Lives Matter and an increased focus on police accountability nationwide. The Post is not tracking deaths of people in police custody, fatal shootings by off-duty officers or non-shooting deaths.

The Federal Bureau of Investigation and the Centres for Disease Control and Prevention log fatal shootings by police, but officials acknowledge that their data is incomplete. Since 2015, The Post has documented more than twice as many fatal shootings by police as recorded on average annually.

The Posts database is updated regularly as fatal shootings are reported and as facts emerge about individual cases. The Post seeks to make the database as comprehensive as possible.

About the datasets The datasets have several fields each and these include:

1. Race, Which includes W: White, B: Black, A: Asian, N: Native American, H: Hispanic, O: Other and None: unknown
2. Gender, M: Male and F: Female and None: unknown
3. The threat_level, this column was used to flag incidents for the story by Amy Brittain in October 2015. As described in the story, the general criteria for the attack label

was that there was the most direct and immediate threat to life. That would include incidents where officers or others were shot at, threatened with a gun, attacked with other weapons or physical force, etc.

4. Attack, this category is meant to flag the highest level of threat.
5. The other and undetermined categories represent all remaining cases. Other includes many incidents where officers or others faced significant threats.
6. Body_camera, News reports have indicated an officer was wearing a body camera and it may have recorded some portion of the incident.
7. Fleeing, indicating whether a victim was running away from the officials or not

The threat column and the fleeing column are not necessarily related. For example, there is an incident in which the suspect is moving away from officers (fleeing) and at the same time turns to fire at gun at the officer. Also, attacks represent a status immediately before fatal shots by police; while fleeing could begin slightly earlier and involve a chase.

6.2 Experimentation process

The datasets from (Post, 2018) are collected from each year since 2015 to 2018. These datasets like explained in the previous section, we got them and integrated them using **Field Mapping** or **Data Mapping** method.

In it's simplest terms, Data Mapping *is about relationships*. In particular, it is the process of specifying how one information set relates, or maps, to another. In our case, there were four sets of data and we joined them together basing on the fields in each set as explain in the previous section of the chapter.

6.2.1 Running Algorithm 2 on the (Post, 2018) datasets:

As requirement by the algorithm, we had to pick out Simpson's Pair variables i.e. α and β_a variables, and also the confounding variables β_b . The table below can help give a clear overview of the pairs that were chosen for the test case:

Algorithm variables	Corresponding Dataset columns
α variables	2
β_a variables	2
β_b variables	4

TABLE 6.1: Showing the number of α , β_a and β_b variables

From the above table, we formulated several column combinations for our linear models of the algorithm. These combinations are the ones that are fit in the models explained in the previous chapter 5.

Due to the reason that whenever a disaggregation of a column occurs, the number of resulting columns is exponential, we had to make a selection of those most relevant columns and dropped off others. Mainly, this is due to dimensionality issues that arise when the categorical data we have in our datasets is transformed into numerical.

Let us take an example of the model having its confounding variable (β_b) as **ARMED**:

- Having $\alpha = 2$ possibilities;
- $\beta_a = 2$ possibilities;
- $\beta_b = 75$ possibilities.

From the above list, the resultant total models would be equal to 300. This in turn, a time consuming process. The other consequent models from the other columns are: 12, 16 and 24.

6.2.2 Results:

Regression model results from the models of the first equations i.e. for the $\alpha = \text{shot}$ and shot_and_tasered , and $\beta_a = \text{signs_of_mental_illness}$ are recorded in the table below:

α	β_a	intercept values
Shot	Signs of mental illness = YES	
Gradient	P-value	
-0.032313	0.001456	0.947178
Shot	Signs of mental illness = NO	
Gradient	P-value	
0.032313	0.001456	0.914865
Shot and Tasered	Signs of mental illness = YES	
Gradient	P-value	
0.032313	0.001456	0.052822
Shot and Tasered	Signs of mental illness = NO	
Gradient	P-value	
-0.032313	0.001456	0.085135

TABLE 6.2: Linear regression model results from the Post dataset

The results in the table give us an understanding that there is a correlation between getting shot while signs of mental illness ARE NOT evident and also a correlation between getting shot and tasered while signs of mental illness ARE evident. The other two scenarios hold no correlation at all. These four models gave us the basis for creating the others which have the confounding variable β_b . In total, there are 52 models to run.

The other 52 Models' results

We want to determine whether the previously obtained model results still stand when we add in the confounding variables to the models and conclude that the initial assumptions made still stand.

1. Confounding variable $\beta_b = \text{"threat_level"}:$

Created 12 models and the algorithm returned the following results instances of the paradox:

α	β_a (mental illness)	β_b (threat_level)	SP instances discovered
SHOT	TRUE	attack	YES
SHOT	TRUE	other	NO
SHOT	TRUE	undetermined	NO
SHOT	FALSE	attack	YES
SHOT	FALSE	other	NO
SHOT	FALSE	undetermined	NO
SHOT_tasered	TRUE	attack	NO
SHOT_tasered	TRUE	other	NO
SHOT_tasered	TRUE	undetermined	NO
SHOT_tasered	FALSE	attack	NO
SHOT_tasered	FALSE	other	YES
SHOT_tasered	FALSE	undetermined	NO

TABLE 6.3: Algorithm results for the $\beta_b = \text{threat_level}$

2. Confounding variable $\beta_b = \text{"flee"}$:

Created 16 models and the algorithm returned the following results instances of the paradox:

α	β_a (mental illness)	β_b (flee)	SP instances discovered
SHOT	TRUE	car	YES
SHOT	TRUE	foot	NO
SHOT	TRUE	not_fleeing	NO
SHOT	TRUE	other	NO
SHOT	FALSE	car	YES
SHOT	FALSE	foot	NO
SHOT	FALSE	not_fleeing	NO
SHOT	FALSE	other	NO
SHOT_tasered	TRUE	car	NO
SHOT_tasered	TRUE	foot	NO
SHOT_tasered	TRUE	not_fleeing	YES
SHOT_tasered	TRUE	other	NO
SHOT_tasered	FALSE	car	NO
SHOT_tasered	FALSE	foot	NO
SHOT_tasered	FALSE	not_fleeing	YES
SHOT_tasered	FALSE	other	NO

TABLE 6.4: Algorithm results for the $\beta_b = \text{flee}$

3. Confounding variable $\beta_b = \text{"race"}$

Created 24 models and the algorithm returned the following results instances of the paradox:

α	β_a (mental illness)	β_b (race)	SP instances discovered
SHOT	TRUE	asian	NO
SHOT	TRUE	black	NO
SHOT	TRUE	hispanic	NO
SHOT	TRUE	native	NO
SHOT	TRUE	other	NO
SHOT	TRUE	white	NO
SHOT	FALSE	asian	NO
SHOT	FALSE	black	NO
SHOT	FALSE	hispanic	NO
SHOT	FALSE	native	NO
SHOT	FALSE	other	NO
SHOT	FALSE	white	NO
SHOT_tasered	TRUE	asian	NO
SHOT_tasered	TRUE	black	NO
SHOT_tasered	TRUE	hispanic	NO
SHOT_tasered	TRUE	native	NO
SHOT_tasered	TRUE	other	NO
SHOT_tasered	TRUE	white	NO
SHOT_tasered	FALSE	asian	NO
SHOT_tasered	FALSE	black	NO
SHOT_tasered	FALSE	hispanic	NO
SHOT_tasered	FALSE	native	NO
SHOT_tasered	FALSE	other	NO
SHOT_tasered	FALSE	white	NO

TABLE 6.5: Algorithm results for the $\beta_b = \text{race}$

The results from the algorithm displayed in the tables above can help us decide as to whether the method we used to do our integration was done well or not depending on the initial conclusions that were made on it or not.

6.3 Validation

The goal of data integration is to gather data from different sources, combine it and present it in such a way that it appears to be a unified whole.

The data integration process mainly focuses on new information. What actually data integration processes do is it create a way between a new data model and an old data model and make them connected to each other. By connecting both of the new and old data models it gets easier to access data from both of the sources. You can say it creates an argument between the new and existing data. Our assessment metric was developed for the sole purpose of making sure that this reason is kept in check and upheld.

6.3.1 Measure of acceptance

To make valid claims of the results that the algorithm in our metric produced in the tables above, we have to formulate a way of acceptance or measure of quality of the metric as compared to the data on which we are running it. To do this, we have to consider the number of instances found as a percentage of the total columns or fields on which the assessment was based. That is to say:

$$\frac{\text{Number_of_Simpson's_Paradox_instances_seen}}{\text{Total_number_of_column_combinantions_made}} \times 100 \quad (6.1)$$

The final decision is left to the user to put a limit to the percentage seen or acceptable depending on their organisation's needs and goals.

From the results in the tables above, we can calculate a measure of acceptance as follows:

- We saw 7 instances of the paradox from a total possibility of 52. So...

$$\frac{7}{52} \times 100 \quad (6.2)$$

- Hence the percentage of instances seen as is **13.5%**
- The percentages of acceptance should be chosen depending on the organisation's needs or goals

6.3.2 Conclusion

The level of detail at which data is collected is important, because confusion and inaccurate decisions can otherwise occur. Aggregated, summarized and manipulated collections of data could offer a different meaning than the data implied at a lower level. An appropriate level of granularity must be defined to provide sufficient uniqueness and distinctive properties to become visible. It is therefore important to have a way of assessing this and making sure that the initially intended outcomes and conclusions about the data still hold meaning.

Chapter 7

Conclusion

Introduction

In this last chapter, we present the way forward, future possible works on our assessment metric and our conclusive remarks, to make it worthwhile for implementation

7.1 What we have done so far

We can give a recap of the work that we did in our thesis and draw some conclusions from there on:

- Our goal was to come up with a way to assess Integrated Data while looking for traces of the Simpson's Paradox in the data segments. The amount of found instances of the paradox in turn would give us grounds of assessing whether the integration was done properly or needed more modifications
- After creating the algorithm and running it on the synthetic dataset that we generated, results showed that indeed the initial intentions of the algorithm and one of the goals of our work was achieved since it returned instances of the Simpson's paradox that were intentionally included in the synthetic dataset.
- Using the field mapping integration method on our case study datasets, we later on run the algorithm on the resulting integrated dataset. Results showed that in some cases there were no instances and in a few of them, there were.
- The results returned a small percentage of instances of the Simpson's paradox compared to the total number of granularly segmented sets of the whole dataset.
- We can conclude that our metric traced and found instances of the Simpson's paradox in the case study dataset, while having used one of the most famous integration techniques. But the instances found do not give a strong persuasion to declare that the integration was done insufficiently.
- We therefore urge that more refinement need be done on our algorithm and more research needed to develop a more sophisticated metric that will address any issues concerning our topic

7.2 The way forward

Surely as long as the world has not ended yet, there are always better ways of doing something and our metric is not an exception to this. There is still a lot of work to be done

on it and make it more robust and reliant. For example, in future works, we are planning on using different more sophisticated machine learning techniques of models like Support Vector Machines, Kernels and many others to help in the creation of better models for the metric.

There is still a long way to go in order to better the metric and more research work is needed to make this a reality.

7.3 Conclusion

There are many elements that determine data quality, and each can be prioritized differently by different organizations. The prioritization could change depending on the stage of growth of an organization or even its current business cycle. The key is to remember you must define what is most important for your organization when evaluating data. Then, use these characteristics to define the criteria for high-quality, accurate data. Once defined, you can be assured of a better understanding and are better positioned to achieve your goals.

Bibliography

- Ahmed, Irfan (Sept. 2017). *The Future of Big Data: 10 Predictions You Should Be Aware Of*. Online. URL: <https://www.smartdatacollective.com/future-big-data-predictions/>.
- Alipourfard, Nazanin, Peter G. Fennell, and Kristina Lerman (Feb. 2018). "Can you Trust the Trend? Discovering Simpsons Paradoxes in Social Data". In: *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining*, Marina Del Rey, p. 9. DOI: [10.1145/3159652.3159684](https://doi.org/10.1145/3159652.3159684).
- Bandyopadhyay, Prasanta S. et al. (2015). "Simpsons Paradox and Causality". In: *American Philosophical Quarterly*, pp. 13–25. URL: <https://philpapers.org/rec/BANSPA>.
- Berman, Steve et al. (Sept. 2012). "Simpsons Paradox: a cautionary tale in advanced analytics". In: *The Statistics Dictionary*. URL: <https://www.statlife.org.uk/the-statistics-dictionary/2012-simpson-s-paradox-a-cautionary-tale-in-advanced-analytics>.
- Bickel, P. J., E. A. Hammel, and J. W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data from Berkeley". In: *Science* 187.4175, pp. 398–404. ISSN: 0036-8075. DOI: [10.1126/science.187.4175.398](https://doi.org/10.1126/science.187.4175.398). eprint: <http://science.sciencemag.org/content/187/4175/398.full.pdf>. URL: <http://science.sciencemag.org/content/187/4175/398>.
- Contributors, Wikipedia (Jan. 2018). *Information Age*. Wikipedia. URL: https://en.wikipedia.org/wiki/Information_Age.
- Fabris, Carem C. and Alex Alves Freitas (2006). "Discovering Surprising Instances of Simpson's Paradox in Hierarchical Multidimensional Data". In: *IJDWM 2*, pp. 27–49. URL: <https://www.semanticscholar.org/paper/Discovering-Surprising-Instances-of-Simpson-%27s-in-Fabris-Freitas/a57bedd707afb8b9590e630a9fb2d43066df3809>.
- Freitas, Alex A., Ken McGarry, and Elon Correa (2007). "Integrating Bayesian Networks and Simpson's Paradox in Data Mining". In: *Causality and Probability in the Sciences*, pp. 43–62. URL: <https://pdfs.semanticscholar.org/793b/8403b5aa8e108d4790ddeb0890be3df64e5f.pdf>.
- Group, DAMA UK Working (Oct. 2013). "THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT". In: *Whitepapers*, p. 17. URL: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf.
- Höchtel, Johann et al. (Mar. 2015). *State of the Art Analysis of Open Data Quality Assessment Metrics and Algorithms*. Research rep. DUK and SWC. URL: <https://www.adequate.at/sites/default/files/2018-01/D1.3State-of-the-ArtAnalysisofOpenDataQualityAssessmentMetricsandAlgorithms.pdf>.
- Hodas, Nathan O. and Kristina Lerman (Mar. 2014). "The Simple Rules of Social Contagion". In: *Scientific Reports* PY - 2014/03/11/online 4343. DOI: [10.1038/srep04343](https://doi.org/10.1038/srep04343).

- Hodas, Nathan Oken and Kristina Lerman (Sept. 2012). "How Visibility and Divided Attention Constrain Social Contagion". In: *ASE/IEEE International Conference on Social Computing (SocialCom-2012)*. DOI: [10.1109/SocialCom-PASSAT.2012.129](https://doi.org/10.1109/SocialCom-PASSAT.2012.129).
- Kock, Ned (2015). "How Likely is Simpson's Paradox in Path Models?" In: *IJeC* 11, pp. 1–7.
- Macsin, Andrei and Robert J. Abate (Feb. 2017). *25 Predictions About The Future Of Big Data*. Online. URL: <https://www.hadoop360.datasciencecentral.com/blog/25-predictions-about-the-future-of-big-data>.
- Post, Washington (July 2018). *Fatal Force*. Online website. URL: https://www.washingtonpost.com/graphics/2018/national/police-shootings-2018/?noredirect=on&utm_term=.fe4196e6ab46.
- Romero, Daniel M., Brendan Meeder, and Jon Kleinberg (Mar. 2011). "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter". In: *International World Wide Web Conference Committee (IW3C2)*. P. 10.
- Techopedia (Feb. 2018). *Data Quality Assessment (DQA)*. Online. URL: <https://www.techopedia.com/definition/30610/data-quality-assessment-dqa>.
- Vellaisamy, P. (2014). "Simpsons Paradox and Collapsibility". In: *Journal of Indian Statistical Association*, 50 (2012), 297–317, pp. 1–19.
- White, Colin (2018). *Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise (Report Excerpt)*. Online in Business Intelligence best practises. URL: <http://www.bi-bestpractices.com/view-articles/4737>.
- Witmer, J. A. (1992). "Data Analysis: An Introduction, Englewood Cliffs, NJ: Prentice-Hall, pp. 21–24." In: *Journal of Statistics Education* 14.1, pp. 110–112.