

TRUST IN THE DATA

EXTERNAL DATA USE BY THE SCOTTISH THIRD SECTOR

A thesis submitted for the degree of:
Doctor of Philosophy (Sociology and Social Policy)
Faculty of Social Sciences
University of Stirling

Tom Wallace
August 2018

ACKNOWLEDGMENTS

The list of people I am indebted to for helping me to produce this thesis is long. I would like to start by thanking my supervision team and my principal supervisor Professor Alasdair Rutherford, for helping me refine my ideas. Reacting to Alasdair's criticisms is what helped shape the project into the thesis which is presented here. Dr Vikki McCall deserves thanks for lending her expertise and providing a different perspective on the research; I often went to Vikki to talk out the big conceptual issues. Claire Wainwright and Mairead Wood were crucial for providing the Scottish Government's point of view on my work and helping organise access within the Government. Finally, I am indebted to Grant Gibson for the support and advice he provided during a crucial period when Vikki was on maternity leave. Without the continuing guidance and feedback of those mentioned above this thesis would not have been possible.

I would also like to thank several other academics who, while not directly connected to the project, were crucial to its success. Professor Paul Lambert for encouraging me to apply for a PhD at Stirling in the first place, for providing continuous methodological support, and for convincing me to attend training which helped me develop skills I will use beyond the PhD. Dr David Griffiths, for giving me feedback during my annual reviews which shaped the direction of the thesis, for his support around network analysis methods, and for helping me prepare by conducting my mock-viva.

On the topic of the viva, I am deeply grateful to my examination team; Professor Mark Tranmer and Professor Paul Lambert (again!). Professors Tranmer and Lambert gave me a chance to defend my work and asked some tough, but eminently sensible questions about the relative strengths and weaknesses of my work. I thank them for being fair, for accepting the value of my work, and for facilitating a stimulating discussion throughout the viva.

My research would not have been possible without the engagement of my respondents. Whether they allowed me to interview them or engaged with my surveys, they contributed directly to the success of the thesis. I hope that the findings and recommendations drawn from this work, and the work I plan to undertake in this area in the future, translate into tangible improvements for charitable organisations.

I also owe a lot to the friends I have made during my, more than five years, in the department (my MSc is included in that five!). Particularly my officemates in 4S24, Kane, who is still figuring out his research philosophy, Paul, who's probably on another internship, and Diarmuid, who was a year ahead of the rest of us and therefore was constantly pestered for hints and tips on the PhD process. My time at Stirling would also not have been the same without the other PhD students with which I have formed friendships; Jen, Camilla, Will, Alana, and Matthew.

Finally, I want to thank my family. My parents for their continued support, both emotional and financial. And my sister for her friendship, I wish her luck during her own PhD.

ABSTRACT

This thesis investigates how the Scottish third sector engages with external data resources. The research is broken into two main strands. The first strand determines factors which affect the level of data use by charities. The second concerns the role of trust in third sector data use, in particular, what effect third sector use of data has on other users of data and the role data plays in society more generally. While the first of these strands is more pertinent to charities themselves, the second is particularly important in the wake of 'Fake news' and a general decline in trust for experts driven by misinformation online. Findings from this research show that there is a relatively low ability to use data among Scottish charities and many are resorting to pre-analysed, aggregate findings. Factors related to a low ability to use data include being a small charity, an old charity and a charity with a narrow focus. Analysis of a series of barriers and enablers found that barriers tend to inhibit data use altogether, where enablers tend to determine level of use. Support to help facilitate data use was then considered, finding that Twitter acts as a forum where support relationships develop between charities and infrastructure bodies, who share widely aimed, one-to-many tweets to support data use. The importance of infrastructure organisation became even more apparent when issues of trust were considered and found that, while charities trust data, they are less trusting of the interpretation which is laid over data and therefore they invest their trust in infrastructure organisations. Infrastructure organisations were found to have a healthy distrust of government data and are invested in feedback mechanisms where they correct mistakes in data and help increase the quality of, and trust in, Scottish Government data more generally.

TABLE OF CONTENTS

Acknowledgments.....	2
Abstract.....	4
Table of Contents.....	5
Table of Figures.....	9
Table of Tables.....	10
Table of Sociograms.....	11
Table of Exponential random graph models.....	12
Chapter 1: Introduction.....	13
Chapter 2: Literature review.....	17
2.1 Introduction.....	17
2.2 Definitions.....	17
2.2.1 Charity and the third sector.....	17
2.2.2 Data and statistics.....	18
2.2.3 The Scottish Government and Government data.....	19
2.2.4 Barriers and Enablers.....	19
2.3 Data use in the third sector.....	20
2.3.1 Impact/accountability reporting and staff, skill, cost, and size.....	21
2.3.2 Fundraising and data security.....	23
2.3.3 Other enablers.....	25
2.3.4 Other barriers.....	25
2.3.5 Summary of barriers and enablers.....	26
2.4 Support for data use.....	27
2.4.1 Defining infrastructure and data organisations.....	27
2.4.2 Social media and Twitter.....	28
2.4.3 Twitter use by charities.....	29
2.4.4 Types of support for data use and Twitter as a forum for support.....	30
2.5 Trust in data.....	31
2.5.1 The concept of trust.....	32

2.5.2	Levels of trust.....	32
2.5.3	Distrust	33
2.5.4	‘Manufactured Risk’	34
2.6	Conclusion	35
Chapter 3: Methodology		37
3.1	Introduction.....	37
3.2	Research philosophy	37
3.2.1	Social Constructivism	37
3.2.2	Structure and agency	38
3.2.3	Mixed methods research	39
3.3	Ethics.....	40
3.4	Survey statistics: methods for studying charity data use.....	41
3.4.1	Software for survey methods	41
3.4.2	Sampling and data	42
3.4.3	Methods of analysis.....	44
3.5	Network analysis: methods for studying support for charity data use	48
3.5.1	Software for network analysis.....	48
3.5.2	Sampling and data	49
3.5.3	Methods of analysis.....	52
3.6	Semi-structured interviews: methods for studying trust	59
3.6.1	Software for qualitative analysis	60
3.6.2	Sampling and data	60
3.6.3	Methods of analysis.....	61
3.7	Conclusion	62
Chapter 4: Data use		63
4.1	Introduction.....	63
4.1.1	Course of analysis	63
4.2	Analytical question 1.a.....	64
4.2.1	Level of use.....	64
4.3	Analytical question 2.a.....	66

4.3.1	Bivariate analysis	67
4.3.2	Combined modelling.....	72
4.4	Analytical question 2.b.....	77
4.4.1	Bivariate analysis	78
4.4.2	Factor analysis.....	81
4.4.3	Multivariate modelling.....	83
4.5	Conclusion	86
Chapter 5: Support		89
5.1	Introduction.....	89
5.1.1	Course of analysis	89
5.2	Analytical question 4.a.....	90
5.2.1	Twitter usage by group.....	90
5.2.2	Centrality by group	92
5.2.3	Sociogram	95
5.3	Analytical question 4.b.....	97
5.3.1	Summary content	98
5.3.2	Tweet-level content analysis	100
5.4	Analytical question 4.c.....	102
5.4.1	Data mentions sociogram.....	103
5.4.2	Data mentions group connection.....	104
5.4.3	Data mentions ERGM	105
5.5	Analytical question 4.d.....	110
5.5.1	Following links sociogram.....	111
5.5.2	Data mentions and following ERGM.....	112
5.6	Conclusion	115
Chapter 6: Trust		117
6.1	Introduction.....	117
6.1.1	Course of analysis	117
6.2	Analytical question 3.a.....	118
6.2.1	Is external data trusted?.....	118

6.2.2	Feeding back	119
6.2.3	Interpretation.....	121
6.2.4	Inter-charity networking and trust.....	123
6.3	Analytical question 2.b.....	124
6.3.1	Budget and staffing	124
6.3.2	Collaboration and formatting	125
6.4	Analytical question 5.a.....	127
6.4.1	Twitter for inter-organisational networking.....	127
6.4.2	Is Twitter a ‘network of trust’?	128
6.5	Conclusion	130
Chapter 7: Discussion		133
7.1	Summary of findings to research questions	133
7.2	Discussion	137
Chapter 8: Conclusion.....		141
8.1	Weaknesses	141
8.2	Future work.....	142
8.3	Recommendations	142
8.4	Concluding remarks	144
Bibliography.....		146
Appendices.....		156
Appendix I. Interview topic guides.....		156
Appendix II. Interview information sheet for participants.....		159
Appendix III. Interview consent form.....		161
Appendix IV. Ethics form.....		162
Appendix V. Clarification of ethics form after review by ethics committee		180
Appendix VI. Network diagrams for interview participants (anonymised example)		184
Appendix VII. Full network group connections table.....		185
Appendix VIII. Link to GitHub		185

TABLE OF FIGURES

Figure 3-1 Sequence of analytical methods	40
Figure 4-1 Distribution of external data use variable.....	64
Figure 5-1 Data content dissemination on Twitter diagram.....	110

TABLE OF TABLES

Table 2-1 Barriers to data use from secondary survey (ordered alphabetically).....	26
Table 2-2 Enablers of data use from secondary survey (ordered alphabetically)	27
Table 3-1 Dependent variable from secondary survey: External data use	44
Table 3-2 Sample of Twitter handles split by group	50
Table 3-3 Summary of mentions and following data.....	51
Table 3-4 Example of a group connection table	57
Table 4-1 Data sourcing responses from primary survey	65
Table 4-2 Organisational characteristics (ordered alphabetically).....	68
Table 4-3 Beneficiary groups bivariate association and significance tests	70
Table 4-4 Financial characteristics bivariate regression results	71
Table 4-5 Aggregate regression results for the combined models	73
Table 4-6 Full model results for the combined models.....	74
Table 4-7 Variable coding for barriers and enablers.....	77
Table 4-8 Summary of barriers bivariate association and significance tests	79
Table 4-9 Summary of enablers bivariate association and significance tests	80
Table 4-10 Aggregate regression results for barriers	84
Table 4-11 Aggregate regression results for enablers.....	86
Table 5-1 Twitter usage quantile regression results with quasi-standard errors	91
Table 5-2 Median degree centrality scores by group	93
Table 5-3 Median centrality scores by group.....	94
Table 5-4 Group Twitter activity statistics summary	98
Table 5-5 Group Twitter content statistics summary	99
Table 5-6 Group connection table for data related content	104
Table 7-1 Main findings and contributions to the literature (sorted by importance)	139
Table 8-1 Recommendations for stakeholders based on findings.....	142

TABLE OF SOCIOGRAMS

Sociogram 5-1 Mentions network sociogram with nodes split by group.....	96
Sociogram 5-2 Data mentions network sociogram with nodes split by group.....	103
Sociogram 5-3 Following network sociogram with nodes split by group	111
Sociogram 6-1. Sociogram of only retweets: a ‘network of trust’	130

TABLE OF EXPONENTIAL RANDOM GRAPH MODELS

Model 5-1 and 5-2 Basic ERGM and Group ERGM for data content	106
Model 5-3 Categorical group ERGM for data content.....	108
Model 5-4 and 5-5 Basic and Data mentions ERGM for charity-support following links	113
Model 5-6 and 5-7 Basic and Data mentions ERGM for support-data following links.....	114

CHAPTER 1: INTRODUCTION

In our modern world, data matters. Data is integrated into how society functions; we interact with, use, and create data just by performing day-to-day actions. The amount of new data created every day is staggering, 2.5 million terabytes in 2013 (Jacobson 2013), but with the emergence of widespread big data and the continued growth of social media since then, this number is almost certainly far higher at the time of writing. This thesis takes the position that data is good for society in general and in a data-saturated world there will be a polarisation between people and organisations with the skills and ability to make use of data, and those who do not. With data so heavily integrated into society, those with the ability to obtain and process this resource are at a clear advantage. However, this ability to use data need not have to mean live processing of vast big data resources, something as simple as being able to find and understand relevant information online to solve a problem or inform a decision could set an individual or organisation apart in our evidence-driven society.

Where do Scottish charities fit into this data paradigm? Registered charities in Scotland, whether they are aware of it or not, necessarily use data; this may be as simple as collecting information on their service users' needs, or feedback on their charitable activities. It could even be as basic as the data they use to maintain their accounts or the information they supply to the regulator. This thesis takes the stance that data is good for charities, and that engagement with data resources helps third sector organisations to better understand their performance, their users, and their organisational practices; ultimately it enables charities to make better informed decisions (Burt and Otto 2017; Hoare and Noble 2016; Steiner et al 2015).

There is a spectrum of data use in the Scottish third sector, with some organisations simply meeting the baseline and others making much better use of the available resources to enhance the way they operate. In general, however, charities are viewed as being poor users of data and a UK Government report in 2003 found that a lack of data capacity reduced the effectiveness of charities to deliver their services (The Comptroller and Auditor General 2009). Even less common for charities is using external, or secondary, data resources. External data is data which has not been collected by the user but rather another person or organisation and then made available for further research. The Scottish Government is one of the biggest publishers of secondary data which would be relevant to the Scottish third sector and their data is, therefore, of particular interest to this project, along with other external data resources charities may use. This PhD is a collaborative studentship funded by the Scottish Government and Economic and Social Research Council. The financial involvement of the Scottish Government in this project shows their desire to investigate this topic and it is hoped that the conclusions of this research will inform future policy in this area. The choice to focus on external data reflects the view that charities are generally not using these

resources to the fullest and that the efficacy of the sector may be, as a consequence, negatively affected.

This thesis will investigate the spectrum of external data use by charities in two ways. The first is to understand what factors lead to this stratification of use. Is there something notable about charities which use lots of external data? Or those that use little to none? And what is the average level of data use in the Scottish third sector? Once level of data use and organisational features which may affect it have been studied, the impact of a series of barriers and enablers to data use, drawn from the literature, will be assessed. This analysis should provide a good overview of levels of data use and what factors may enhance or inhibit data use. This leads to the question of what support there is available to charities to help them overcome barriers, build on enablers, and ultimately increase data use. This thesis chooses to focus on social media, particularly Twitter, as a forum for support for charity data use. Twitter is a particularly relevant place to study charity support because it is widely used throughout the sector and is relatively understudied (Guo and Saxton 2014). The common use of Twitter in the third sector means that it is a natural place for support relationships to form between charities and infrastructure organisations with the latter facilitating the former's access to data related content on the platform.

The second strand of this investigation will look at what effect charity use of data has on other users of external data. Collins Dictionary selected the term 'Fake news' as its Word of the Year for 2017 and this reflects a notable decrease in the trust which the public place in facts and experts (British Broadcasting Corporation 2017). This decrease in trust comes in the wake of a flourishing of wilfully inaccurate information online; particularly on social media which has been a forum for government sponsored campaigns of misinformation aimed at subverting democratic elections (Shane and Goel 2017). Charities could be an important part of pushing back against this trend by both correcting inaccurate data, sharing data with their peers after adding context to it, and disseminating robust findings drawn from external data to the public. Each charity alone will not be able to push back against 'Fake news' but, in congress, the sector may play a vital role in increasing the trust in, and availability of, good quality data and research which helps counter wilfully wrong information. Trust is at the very root of this issue; for charities to increase public trust they have to have trust in the data themselves but trust has been left out of the discussion of charity data use so far and does not feature notably in any of the literature reviewed during this project. This thesis will show that the issue of trust in data is absolutely essential for understanding how charities interact with data and help improve it for other users and the public.

These two main strands of analysis feed into the project's research questions, each of which is broken down into more detailed analytical questions:

1. What level of external data usage is there in the Scottish third sector?

- 1.a. *What level of external data usage is there among charities in Scotland?*
2. What barriers, enablers, and organisational features affect the ability of third sector organisations to make use of external data?
 - 2.a. *Which organisational features best predict differences in levels of use?*
 - 2.b. *What other factors enable or inhibit use of external data in the Scottish third sector?*
3. To what extent is external data trusted by third sector organisations and what effect does this trust have on other users of data?
 - 3.a. *What level of trust is there for Scottish Government data and other external data among charities and do charities help increase the quality of and trust in external data for other charities and other users of data?*
4. What evidence is there of support for data usage among charities, support organisations, and data organisations on Twitter and how does this support manifest?
 - 4.a. *To what extent do charities, infrastructure organisations, and data organisations use Twitter differently and what implications does this have for how they are networked?*
 - 4.b. *What content is actually exchanged on Twitter between charities, infrastructure organisations, and data organisations, how much of this content is data related, and what form of support do these data related tweets embody?*
 - 4.c. *What are the dynamics of data related tweets between the groups and what does this reveal about support for data use on Twitter?*
 - 4.d. *Is there evidence of support for data use disseminating through following links?*
5. To what extent is the network of charities on Twitter a ‘network of trust’?
 - 5.a. *To what extent do links on Twitter embody trust for other organisations and what does this reveal about trust for data?*

These questions map onto the three analysis chapters which make up the core of this thesis. To comprehensively answer both strands of analysis, and all of the research questions, different methods were chosen for each chapter of analysis. The first analysis chapter, Chapter 4, uses survey methods and statistical techniques to determine levels of external data use in the third sector and which organisational factors, enablers, and barriers most strongly predict levels of use. The intention of this chapter is to set the context for the rest of the analysis and identify factors which may be helping or hindering third sector data use. Chapter 5 uses network analysis techniques and data gathered from Twitter to map a sample of charities, infrastructure organisations, and data related organisations. This chapter will determine the dynamics of support for data use; if and how support is being delivered to front-line charities on Twitter which may help them use data better. Data from this chapter will also be used to investigate the second strand of analysis, but this

discussion will take place in Chapter 6 which is the final chapter of analysis. Chapter 6 looks directly at the issue of trust, firstly as a barrier or enabler to data use, but then more importantly with regard to the second strand of analysis looking at what effect charities have on other users of data; do charities help improve the quality of and trust in government data?

Following from this introduction, a literature review will underpin each chapter of analysis with relevant previous research. A methodology chapter will then cover, in detail, the methods of analysis used in each of the separate chapters and the philosophy used to unite these disparate methods. The analysis chapters follow and are then summarised in a discussion chapter and final conclusion.

Scottish charities' ability to use data is generally regarded as quite low and previous research suggests that this may have negative consequences for charitable outcomes (De Las Casas et al 2013). This thesis, therefore, intends to determine what factors most pertinently affect a charity's ability to use data and if adequate support is in place to help increase use. An intended outcome of this research is that support for data use could be focused on those factors which have the biggest impeding effect and thereby the ability of the sector to use data could be increased. Running parallel to this, there is also an assumption that a data capable charitable sector plays an important part in increasing trust in data resources overall and, therefore, increasing charity use of data may result in greater trust for data resources from other users and, ultimately, the public. Ascertaining the extent to which any of the above holds true is the *raison d'être* of this research.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

The relationship between charities and external data resources is complex. Some elements of this relationship have featured in academic literature, while others are best covered by grey literature, and still others have had very little attention. This review, therefore, will not endeavour to provide a panoptic summary of charity data use according to previous research, as this would be far from comprehensive, rather it will draw together disparate strands of literature to contextualise and underpin the research carried out in this thesis. This primarily involves establishing a foundation for each chapter of analysis. Chapter 4 involves looking at external data use directly, and looking at what factors may inhibit or enhance a charity's data use. Chapter 5 looks at what support charities have available to help them overcome barriers and increase their external data use. Chapter 6 looks at the role trust plays in how charities relate to data and support for data.

It should be noted that while this thesis is focused on the Scottish third sector and secondary data produced, mainly, by the Scottish Government, the literature is more broadly focused. The bulk of research in this literature review concerns the UK third sector which may or may not include Scottish charities depending on the specific details of each piece of research. The distinction between charities from different parts of the UK is therefore difficult to study and will not feature as a pertinent part of this review. Additionally, much of the literature concerning charity data use either focuses on the data collected by charities themselves, or does not differentiate between internal and external data. The analysis in this thesis focuses on the latter, but literature concerning the former is often still relevant as many issues which affect external data use also affect internal use. It is theorised that external use develops from internal use and therefore barriers affecting the initial development of data skills are relevant to later external data use. The distinction between external and internal data is reflected upon where appropriate.

With these gaps in the literature, and other difficulties studying Scottish charity's use of external data, it is sensible to defend the choice to focus on charities in the first place; this topic is of particular interest precisely because of the gaps in our current understanding of the field. Charities are not easy to study but it is in the difficult topics where the impact of good research will be most deeply felt.

2.2 DEFINITIONS

2.2.1 Charity and the third sector

There are varying definitions of what constitutes a charity, and varying terms which can be applied to these sorts of organisations, for example: charity, non-profit, third sector organisation, or voluntary organisation. Salamon and Anheier (1992) proposed a foundational definition of

charitable organisations which moved beyond simply describing their structure or sources of income, which previous definitions had focused on. Salamon and Anheier (1992) instead define five critical features of a charity: formal, private, non-profit-distributing, self-governing, and voluntary. Formal means a charity has to be institutionalised to some extent and cannot be purely *ad hoc*. Private simply means a charity must be separate from the government or state. Non-profit-distributing means funds must be used to achieve charitable purpose. Self-governing means that a charity has internal organisational agency and is not controlled from without. Finally, voluntary does not mean that a charity can only employ volunteers but rather that a charity must have a significant voluntary component.

In Scotland, from a research point of view, the task of defining charity and the wider charitable sector is made much easier by the Office of the Scottish Charity Regulator (OSCR). OSCR stipulates that a charity must; have only charitable purpose, provide public benefit, use their funds and property only for charitable purposes, allow fair access to the benefit they provide, and not exist to advance a political party (Office of the Scottish Charity Regulator 2018). Organisations which meet these requirements are then allowed to register and the Scottish charitable sector is defined as active registered organisations. In Scotland, an organisation not on the Scottish Charity Register is not a charity (Office of the Scottish Charity Regulator 2018). There are organisations similar to formal charities, though usually on a smaller scale, which are not formally part of the sector but which Soteri-Proctor and Alcock (2011; 2012) consider part of the wider charitable sphere. These micro organisations often fly '*beneath the radar*' of charities research which makes defining and studying them even more problematic (Soteri-Proctor and Alcock 2012: 379). This study does not include these organisations.

A phrase used throughout this thesis is 'frontline charities'. This phrase has no formal definition and is simply used to differentiate infrastructure organisations from the larger group of charities who directly deliver a charitable benefit to the public; charities at the coalface. Infrastructure organisations are defined in section 2.4.1.

2.2.2 Data and statistics

Data is quantifiable information, that is information or facts which can be expressed in terms of quantities or numbers; that which can be counted (Bryman 2012). For the purposes of this thesis, 'data' does not include qualitative information. Statistics are quantitative facts collected, aggregated, structured, and analysed to facilitate understanding; it is the organisation and analysis of data into forms that can be readily understood (Hauser 1973). Datasets or data resources fall somewhere between these two definitions; they are structured collections of data, often presented in a way which facilitates analysis (MELODA (Metric for Releasing Open Data) 2018).

2.2.2.1 Primary data vs secondary data

Primary data is data that has been collected close to the source by the person, organisation, or entities which will go on to process the data (Walliman 2006). Secondary data is when this data is processed by someone other than the entity that collected it. This distinction sounds minor, but, within social science, secondary data, generally, refers to large-scale, professionally collected and organised data which is specifically designed to aid research. This includes the census and large surveys carried out by, or on behalf of, governments (Scottish Government 2017). This thesis focuses on government published secondary data, but many issues relevant to primary data also affect secondary data so both will be considered.

It should be noted that while these definitions of data make sense in an academic context, they may be understood differently by charities. This is mainly a problem for the definition of secondary data which some respondents to the surveys and interviews understood as aggregate findings or reports based on data sets. While to a quantitative researcher the distinction between data and results is very clear, this does not appear to be as readily understood outside of academia and research. This confused distinction for secondary data is discussed in Chapter 4 and is why the terms ‘external’ (secondary data) and ‘internal’ (primary data) have been used throughout this thesis to avoid confusion.

2.2.2.2 Data use

This thesis refers to ‘data use’ throughout. This has no formalised definition and is used in a very wide sense to refer to any engagement with or use of data resources.

2.2.3 The Scottish Government and Government data

The *state* is a complex interconnected web of institutions and bodies which are organised hierarchically (Hall and Taylor 1996). It is difficult to break down the concept of the state into its individual components, firstly because they are legion, and secondly because the concept of the *state* derives from the interaction of its subcomponents rather than simply their summation (Cairney 2012); the state is irreducibly complex. *Government*, more simplistically, refers to the structures and agents of the incumbent political party in their role as the executive. The Scottish Executive was renamed the ‘Scottish Government’ in 2012. Therefore, in Scotland, ‘government data’ refers to any data produced by or for the state under the authority of the Scottish Government.

2.2.4 Barriers and Enablers

Barriers and enablers to charity data are not well defined concepts in the existing literature despite many of them being widely discussed, as reviewed in Section 2.3. Definitions for these concepts will, therefore, not draw on the literature but instead discuss working definitions of barrier and enablers to clarify how the terms will be used in this thesis.

Barriers are factors which inhibit, block, or impede a charity from collecting, processing, analysing, or engaging with data resources.

Enablers are factors which enhance, facilitate, or permit a charity in the process of collecting, processing, analysing, or engaging with data resources.

Lists of the barriers and enablers used in this thesis can be found in Table 2-1 and Table 2-2 and they are discussed at length in section 2.3.

2.3 DATA USE IN THE THIRD SECTOR

Data use is good for charities. As discussed in the introduction, the ability to engage with data helps charities to make better decisions and better understand their users (Burt and Otto 2017; Hoare and Noble 2016; Steiner et al 2015). Effectively used data can act as a ‘*multiplier*’ for charity effectiveness, providing significant returns on investment in the skills and resources required to engage (Ellis and Gregory 2008). A review by the UK Government into charities’ relationships with data, as far back as 2003, found that a lack of capacity for data across the third sector was a significant barrier to the ability of charities to effectively deliver public services (The Comptroller and Auditor General 2009). This being the case, why is effective use of any data, internal or external, by charities relatively rare¹? A report by Steiner et al. for JustGiving (2015: 5) found that only 24% of the charities they surveyed were employing data as part of their strategic decision making. A report for NTEN by Andrei et al. (2012) found that a similar proportion, 26%, of charities were using data effectively, while 99% of charities were collecting some form of data; every charity is collecting data, but only a minority use it effectively. De Las Casas et al. (2013) concurred with this conclusion and argued that while many charities are nominally using data, the vast majority are only engaging with it to a very limited extent. Exploring why this is the case makes up the bulk of this segment of literature review, but what is clear from the outset is that, perhaps ironically, there is a lack of data concerning charity use of data. A recent report into third sector skills by Broomhead and Lam (2017) runs for 22 pages without mentioning data once, implying a lack of literature concerning this issue and a lack of attention in general. An older report by Macmillan et al. (2014) concurs that there is a lack of evidence on third sector capacity building; though it does not specify data use. This provides the lacuna for this thesis, but also curtails the scope of this literature review to some extent.

¹ It should be noted that, while this emphatic endorsement of data’s utility to charitable organisations is well situated in the literature, this is not universally accepted. The most notable rejection of the utility of data in the third sector has been to the Social Return on Investment (SROI) and other accountability frameworks which can be onerous and require charities to redirect resources which some feel would be better spent on charitable activity (Harlock 2013). A more detailed discussion on accountability can be found in Section 2.3.1.

There are two dominant factors which, the literature suggests, drive charity data use: impact/accountability reporting, and fundraising. Impact and accountability reporting is discussed as a driver of data use and then consideration is given to the effect which staffing, skills, costs, and size have on this form of data use. Fundraising, likewise, is discussed as a factor which can drive charity data use, and data security is then discussed as a barrier which can curtail this enabler. Both of these enablers are most relevant to internal charity data use, but the barriers which constrain them are relevant to all data use. There are also several lesser enablers and barriers discussed in Sections 2.3.3 and 2.3.4 respectively. There is one strand of literature which is not reviewed in this chapter as it concerns enablers or uses of charity data which are too advanced. As described above, the majority of charities struggle to engage with data beyond very limited uses. While there are trailblazers, literature promoting charity use of real-time data (Amar 2017), automation, accessing APIs, or programming languages (Steiner et al 2015) is not particularly useful for this research as these enablers are only likely to be accessible to a tiny minority of, already skilled and capable, charities.

2.3.1 Impact/accountability reporting and staff, skill, cost, and size

Impact and accountability reporting is probably the most powerful driver of charity data use, even if it tends to concern internal data, that is, data collected by charities rather than external sources of information (Harlock 2013). Data on impact allows charities to better understand how they operate and improve their effectiveness (Curvers et al 2016; De Las Casas et al 2013). It also allows them to better understand their users (Burt and Otto 2017). However, the reason impact and accountability is such a powerful enabler is not because it is beneficial to charities, but rather because many funders and charity support organisations require this sort of reporting (Amar 2017). Charities are often required to collect data on their activities but, as described above, most do not make very good use of this data, only doing what they must to satisfy their minimum requirements (Burt and Otto 2017). As there are many benefits to analysing impact data, it is unlikely that charities are choosing to disengage, and are instead being blocked from fully making use of this data by barriers.

The first issue for some charities is that there are a variety of impact and evaluation tools and some of the most well-known, notably the Social Return on Investment but also the Social Accounting Audit, are quite demanding on the charity's time and resources (Harlock 2013). Low capacity charities, therefore, tend to opt for less involved forms of impact measurement but some funders require certain forms of reporting and if a charity is supported by several organisations they often have multiple reporting requirements (Harlock 2013). This lack of coordination amongst funders and support organisations is a barrier to effective impact evaluation by lower capability organisations. This barrier, specific to reporting, also reveals more fundamental impediments to

charity data use: skills, staffing, and funds, all of which are linked to organisational size (Millar and Hall 2013) and all of which had a negative effect on both internal and external data use.

Staffing and skills are interrelated to the point of being somewhat indistinguishable as barriers to data use. Only a small minority of charities have dedicated analysts (De Las Casas et al 2013) and a majority of charities, across two separate studies, reported not having the right skills or staff to analyse data (Boswell et al 2016; Ógáin et al 2012). This being the case, data use skills can either be obtained by up-skilling existing staff or by recruiting staff with pre-existing skills (Steiner et al 2015). However, a survey by Lloyds Bank (2016: 38) found that fewer than 20% of charities were able to recruit to bring in digital skills, mostly due to financial constraints. Despite this report specifying ‘digital skills’ rather than data analysis, the Lloyds report still suggests that up-skilling and training are the focus for most charities over recruitment. Training and support for analysis skills are investigated in more depth in Section 2.4 but an implication of training staff over recruiting them is that charities are investing in their staff, which makes them vulnerable to staff turnover which is notably common in the sector (Cunningham 2001). This issue is particularly acute in charities with a small number of key data literate personnel and can threaten the sustainability of the charity’s ability to use data (Burt and Otto 2017). This instability and uncertainty can make it hard for charities to fully embrace data and embed it within how the organisation operates, which is how the full benefits of data are realised (Burt and Otto 2017).

A second issue, raised by the Lloyds report (2016), which is related to staffing, is cost and budget. Hiring staff is obviously expensive, but training can also be resource intensive. Harlock (2013) reports that some charities feel that training staff with advanced skills diverts resources from other areas, particularly from core charitable activity. Harlock provides no evidence for how widespread this view is, but Ógáin et al. (2012: 46) report that 79% of their sample of 1,000 charities felt funding or resources was their greatest barrier while Lloyds (2016: 6) found a strikingly similar proportion, 78%, of their more than 200,000 charity sample were investing nothing in digital skills. While ‘digital skills’ are not exactly the same as data use skills², this still shows a general resistance from charities to divert limited resources to training which could increase their data use skills. McCabe and Phillmore (2012) pointed out that this forms a vicious cycle; as expressed in Section 2.3, data analysis skills appear to be extremely valuable for charities and can help them save money and operate more effectively. However, if a charity does not have the resources to invest in developing these skills then they never see the benefits and remain uncompetitive. This cycle is particularly hard to break because data use appears to be something that charities have to

² Generally ‘digital skills’ has a wider definition than data use skills; including ability to create and engage with digital content such as the charity’s website. The definition used by Lloyds Bank (2016) also includes the ability to source and manage information which has significant overlap with data use skills. IT skills are also a critical antecedent to developing data skills (Burt and Otto 2017; Ellis and Gregory 2008; Steiner et al 2015).

go all in on; the value of data needs to be embedded throughout the organisation for the full benefits to be apparent (Burt and Otto 2017).

It should be obvious, with staffing and skills being constrained by resources and the size of charities usually being measured by their income (National Council for Voluntary Organisations 2014), that smaller organisations will have less access to training and staff and there is some direct evidence which bears this out. Ógáin et al. (2012: 2) found that while 25% of all charities did not do any impact analysis, the figure for organisations with income below £100,000 was nearly 50%; though their study did under-sample these smaller organisations. Ellis and Gregory (2008) found, in a Scottish pilot study, that the skills burden was particularly acute for small charities, though their study is somewhat dated. The Lloyds Digital Index (2016: 39-42), which defined small charities as those with between zero and nine employees, found that these organisations were less likely to invest in digital skills and consequently also reported being less proficient in these skills which are related to and somewhat underpin data use. However, the 0 to 9 employees used in the Lloyds report is a very broad category; a charity employing no one is significantly different from one employing nine people and this undermines their findings to some extent. One possible route for smaller organisations, which is not covered in the literature, is hiring external consultants to provide data analysis. This avoids the risks of losing the skilled member of staff, but the problem with this as a solution is twofold. Firstly, statistical contractors can be expensive and so are constrained by the financial barrier previously discussed, and secondly, as Burt and Otto (2017) discuss for data to really enhance a charity it needs to be embedded in the day to day workings of the organisation. Therefore, consultancy is clearly not a solution for low-capacity charities, but rather a useful stopgap for charities who already have some ability to use data.

It is clear that the most pertinent factor holding most charities back from making more use of data, internal or external, is staffing and skills. However, the root of this barrier is cost which prevents charities from being able to hire or properly train analysts who are in high demand and are therefore expensive (Big Lottery Fund 2015; Ellis and Gregory 2008; Ógáin et al 2012). With these factors in mind it is not surprising that small organisations struggle the most to develop data analysis skills. It is these smaller charities who are most likely to get caught in the cycle of being uncompetitive and inefficient through lack of engagement with data and therefore never having the resources to build up skills which would ultimately benefit them (Hoare and Noble 2016).

2.3.2 Fundraising and data security

The second notable driver of charity data use is fundraising. Data is used to aid fundraising in two ways; it helps charities better understand their donors and it helps them put together funding bids using data to highlight areas where they can improve society (Steiner et al 2015). An Institute of Fundraising report (2016) argued that if charities understand their donors better they can raise

funds more cost-effectively by targeting the individuals most likely to give. The literature focuses on the former of these uses, which mostly involves internal data, but Macmillan et al. (2014) suggest that capacity building focused on fundraising, driven by funders, could help develop charity skills more generally. Although Macmillan et al. provide no evidence to back up their assertion, a working paper by Dayson and Sanderson (2014) found that charities that rely on the public sector, or Big Lottery Fund, as a key funder are more likely to access support; suggesting that funders play an important role in driving charities to develop their skills in general. Similarly to impact and accountability, this means that strategic coordination amongst funders and third sector support organisations could have a big impact on charity skills.

However, there is one significant impediment to charity data use which is particularly relevant to charity fundraising data. Data collected to inform charities about their donors is usually internal and is invariably personal human data³. Clark (2018) found, from a survey carried out by the Technology Trust, that 82% of charities were holding personal data of this sort. This can be problematic because, as shown previously, many charities do not have the skills to fully engage with data and data security can have legal implications if something goes wrong (Curvers et al 2016; W. Hall et al 2012). The combination of personal data, which charities are often compelled to collect and hold, and a lack of skilled staff results in a powerful barrier to data use for low skilled organisations that are fearful of breaking the law (W. Hall et al 2012; Lloyds Bank 2016). This may be particularly relevant at the time of writing and for several years into the future due to the new General Data Protection Regulation (GDPR) laws, which update the Data Protection Act (DPA) and came into effect in May 2018 (Institute of Fundraising 2017). As the Institute of Fundraising (2017: 1) notes:

“[GDPR] isn’t a ‘nice to have’, it’s a fundamental legal responsibility of every charity to ensure that they have the right policies and procedures in place so that they are being run properly and are taking individuals’ rights seriously.”

As Burt and Otto (2017) discussed in reference to organisational strategy, GDPR compliant analysis is not something that can be built up; charities have to go all the way or risk prosecution and with limited skills and resources it isn’t surprising that many are choosing to disengage and simply hold data rather than risking doing anything with it.

The barrier presented by data protection and security concerns appears to be particularly broad; GDPR, and DPA before it, should not have much impact on secondary analysis as external data sets are usually anonymised and made safe by providers. However, it appears, as mentioned in the introduction, that charities begin building data skills by analysing their internal data and anything

³ Personal human data is data about or identifying a living individual such as; address, name, contact details or date of birth. Some charities may even hold more sensitive personal data such as medical information.

which obstructs this consequently also acts as a barrier for external data use. Fears over data security are, therefore, a major barrier to all forms of charity data use. Data protection as a barrier is also related to the theory of 'Manufactured Risk' which is discussed in Section 2.5.4.

Compounding these issues, and related to staffing, a study by Big Lottery Fund found that smaller charities reported fundraising staff being the most difficult vacancy to fill in terms of recruitment (Leat 2011). Some charities may, therefore, be stuck; collecting data for fundraising purposes but without access to the staff to properly and legally process it.

2.3.3 Other enablers

This section briefly reviews two important enablers which are not directly constrained by, or related to, particular barriers. The first of these enablers is leadership and organisational strategy. Case studies carried out by Burt and Otto (2017) suggested that the background of a charity's CEO was a common theme in very data capable organisations; CEOs with backgrounds in research seemed to understand the value of data and prioritise its use. The importance of leadership was also highlighted by Clark (2018) who reported on a survey of more than 1,200 charities. Curvers et al. (2016) argued that there is a clear link between the priorities of charity leaders and organisational strategy, with charities led by those who do not place value in data not developing organisational strategies which emphasise data use.

A second key enabler is information technology (IT) systems. IT, or digital infrastructure more widely, is the foundation of data use by charities; if client records are not digitised then they cannot be effectively analysed. Even something as simple as moving data from paper to a computerised spreadsheet could significantly increase a charity's ability to use the data they hold (Burt and Otto 2017; Ellis and Gregory 2008; Steiner et al 2015). This enabler is also pertinent to external data as this is invariably sourced from the internet and analysed with specialist software. Limitations to this enabler come down to staffing, Clark (2018) found that only around half of charities have staff allocated to IT and almost two-thirds rely on volunteers to help with IT. This means charities could be vulnerable to IT trained staff and volunteers moving on just as with data analysis staff. Additionally, and related to leadership above, Clark (2018) found that 58% of charities don't have a digital strategy which is part of making sure that the IT infrastructure, which allows for data use, is in place. Therefore, it seems that leadership/strategy and IT systems are important enablers to charity data use, both internal and external.

2.3.4 Other barriers

Organisational age is only explicitly covered in the extant literature by the Lloyds Digital Index (2016) which concerns digital skills rather than data use specifically. This report found that younger charities tended to have more mature digital skills than older organisations. This may

suggest that younger charities are also better users of data, as IT has been shown to be an enabler of data use. However, the effects of size will need to be controlled for when considering age because surviving older charities are likely to be larger than younger charities and the Lloyds report does not account for this. Although, the direction of effect they found would suggest age is not masking size as previous literature has shown larger size to be an indicator of more ability to use data.

2.3.5 Summary of barriers and enablers

Because the analysis of barriers and enablers in Chapter 4 uses data drawn from a secondary survey, it was not possible to perfectly translate the factors discussed in the literature into the analysis as the wording of the questions respondents were asked, and the resulting variables, were predetermined. Therefore, the tables below set out the barriers and enablers which are used in the forthcoming analysis.

Table 2-1 Barriers to data use from secondary survey (ordered alphabetically)

Barriers in the survey
Agreeing information standards
Agreeing shared information standards
Cost
Data privacy concerns
Data security concerns
Ethical issues
Integrating IT with partner
Integrating IT within charity
Lack of analytical skills
Legal/regulatory constraints
Seeing data informed performance
Seeing innovation in data
Time

Source: Secondary survey

Table 2-2 Enablers of data use from secondary survey (ordered alphabetically)

Enablers in the survey
Better use of resources
Competitive advantage
Data can inform strategy and operation
Leadership
Reporting and accountability requirements

Source: Secondary survey

2.4 SUPPORT FOR DATA USE

This section reviews literature which underpins the second analysis chapter of this thesis. Chapter 5 concerns support for charity data use, primarily for external data use from infrastructure organisations, but also data organisations, and uses Twitter as a case study for the relationships between charities and support, as well as a new, underutilised, forum to seek and provide support. Underpinning this analysis with literature involves first defining support organisations, data organisations, and Twitter before looking at how charities are engaging with Twitter. The review then covers some of the types of support currently provided to charities by infrastructure organisations and how greater use of Twitter could, potentially, resolve some of the weaknesses in the current support paradigm.

2.4.1 Defining infrastructure and data organisations

When looking at charity support, Ellis and Gregory (2008) make the point that there are many suppliers of support to charities. These include organisations from the public and private sectors and, in Scotland, local Third Sector Interfaces (TSIs) (Scottish Council for Voluntary Organisations 2016). However, this study focuses on the support provided by third sector infrastructure organisations; known variously as infrastructure, intermediary, support, or umbrella bodies (Walton and Macmillan 2014). These are, generally, charities which provide infrastructure for the rest of the third sector and may be broad umbrella organisations, such as the Scottish Council for Voluntary Organisations (SCVO), subsector specific bodies, such as The National Association for Mental Health, or focused on a particular type of support, such as Evaluation Support Scotland (Charity Commission 2013). This thesis chooses to focus on these organisations because, as pointed out by Dayson (2011), supporting charities, particularly smaller charities, is the core purpose of most infrastructure organisations. Dayson and Sanderson (2014) detail in one article that, by the number of organisations supported, infrastructure organisations provide 95% of the support which charities receive. While their definition of support only includes formal aid, this still speaks to how dominant infrastructure organisations are in supporting charities and this point has been borne out by other research (Macmillan et al 2014). Therefore, any drive for increasing the use of data by charities will necessarily involve infrastructure organisations and the support they provide.

Data organisations are not a formalised, predefined, group like support organisations and are only grouped for the analysis in this thesis. The group is comprised of data producers and advocacy organisations that are theorised to be important in supplying support specifically for data use either directly to charities, or vicariously through infrastructure organisations. There is no body of literature concerning how these groups may support or facilitate third sector data use, particularly on Twitter. There is a discussion around how these organisations were sourced in the methodology in Section 3.5.2.1.

2.4.2 Social media and Twitter

Twitter is not a new platform, having been founded in 2006 and popular since 2009 (Arceneaux and Schmitz Weiss 2010). However, academic literature focused on the platform is still patchy and owes much to the body of literature covering Facebook and older forms of interaction on the internet.

The foundations for social media research were laid by Barry Wellman around the turn of the millennium; before social media had even been established. Wellman was interested in how email and personal web pages replicated and supplemented offline relationships (Wellman and Hampton 1999; Wellman et al 2002; Wellman 2001). He found that online links followed offline friendships remarkably closely but his evidence was small scale and anecdotal. Adamic and Adar (2003) provided more robust evidence when they looked at personal homepages, which linked to each other, on university campuses and found that online networks replicated those offline so well that they could infer offline connections from the data they collected online. Adamic and Adar were working just prior to the advent of large-scale social networking sites and their study used personal web pages as proto-social networking profiles, with hyperlinks standing in for ‘friends’. Given its age, their study has strikingly similar conclusions to studies of more modern social networking, as detailed below, and it forms a link between the dial-up dominated world Wellman was writing in and Web 2.0⁴.

Facebook launched in 2004, but didn’t go fully public until 2006, therefore Golder et al. (2007) studied it relatively early in its life. Their study was similar to Adamic and Adar’s, focusing on university campuses, though, on a larger scale, and build on Wellman’s use of email as a proxy for the strength of a relationship by looking at messaging links (Golder et al 2007). They used time zones to estimate geographic proximity and found that networks were clustered into schools suggesting that Facebook was replicating offline networks. They collected their data at various different times and periods in the year and most interestingly found that activity on the social

⁴ Web 2.0 refers to the age of user-driven content, such as social media.

network spiked over holidays suggesting that students were ‘making up’ for the loss of face to face contact with messaging on Facebook (Golder et al 2007).

Twitter is slightly different from previous social networks. If Facebook is made up of discrete, interconnected, private networks (originally stemming from university campus networks) then Twitter is more diffuse, less focused, and yet more externally accessible due to its relative lack of privacy settings. Kwak et al. (2010) found that Twitter has a very low rate of reciprocation among follower/following pairs and exhibits non-power law follower distributions, both of which are not consistent with normal patterns of human interaction and those replicated by the other networks covered previously. Twitter is, therefore, more of an information spreading medium than a friendship network and is less likely to replicate offline links (Kwak et al 2010). Research on Twitter, therefore, needs to be aware that the platform is somewhat different from previous forms of social media. In the case of this thesis, the information spreading nature of Twitter appears perfectly suited to the provision of support to charities.

2.4.3 Twitter use by charities

Social media use is common among charities. The Lloyds Digital Index (2016: 47) found that while only 31% of charities professed to using social media in 2015, this number was up to 44% in 2016. This estimate is far lower than the 90% reported by Clark (2018: 1) two years later but social media use is clearly prevalent in the sector despite varying estimates. For Twitter specifically, Guo and Saxton (2014) found 80% of their sample was using Twitter in 2014, but they did acknowledge that their sample only contained large organisations who may be more likely to invest in social media. Sampling differences between the Lloyds study and that reported on by Clark may therefore explain their very different results. What is clear is that charity use of social media and Twitter in particular is common and growing, if not yet universal.

But what do charities use Twitter for? Auger (2013) argues that charities use Twitter for public engagement as many rely heavily on individual donations or networks of public support to sustain them. Charities attempt to engage with their publics, several authors claim, primarily through broadcasting one-to-many information messages (Lovejoy and Saxton 2012; Phethean et al 2015; Waters and Jamal 2011). The aim of these tweets is usually to grab people’s attention and promote the charity’s cause. Phethean et al. (2013) lament this trend as a missed opportunity but Lovejoy and Saxton, who performed a much deeper content analysis, were impressed with how rich the information being broadcast was, invariably using URL shortening services to connect followers to full resources, studies, and data. The literature seems to concur that social media is part of what makes a successful charity (Lloyds Bank 2016; McCabe and Phillimore 2012). However, there is almost no literature looking at charity to charity interactions, or charity to infrastructure interactions; the latter being the crux of Chapter 5 of this thesis and one way in which it seeks to

contribute to the literature. Therefore, this review will move on to summarising the types of support infrastructure organisations currently provide to charities, relevant to data use, and how Twitter may aid or augment this support.

2.4.4 Types of support for data use and Twitter as a forum for support

This section covers some of the types of support infrastructure organisations provide to charities, the drawbacks or limitations of these forms of support, and where Twitter may augment the current support paradigm.

One of the most notable variations in the support provided by infrastructure organisations is how deep it is versus how wide it is. Generally, support is either intense but provided to only a few recipients, one-to-one support, or is shallower but provided to a much wider audience; one-to-many support (Wells and Dayson 2010). Web-based resources tend to be at the wide-but-shallow end of this spectrum; they have impressive reach but limited actual utility as support tools (Leat 2011). Bubb and Michell (2009) argued that capacity building has not resulted in sector-wide benefits because narrowly focused support has only applied to a small number of organisations. There is, therefore, clearly the desire for a form of support which is both detailed and wide-ranging. Twitter may have the potential to fill this gap, it is certainly wide-ranging and used by many charities already, and previous literature has suggested that the information shared on the platform can be surprisingly deep. Twitter can also be used for direct communication besides simply broadcasting information. How Twitter is used to support charities and the extent to which it can overcome the wide-shallow paradigm will be assessed in Chapter 5.

A second factor affecting the support given by infrastructure organisations is cost. The cost of support provided by the private sector is often prohibitively high, but support from infrastructure organisations can also cost charities (Macmillan et al 2014). This cost may be monetary, but it can also be the cost of having staff out of the office or travelling to training or workshops. These costs have resulted in what Harlock (2013: 18) described as *'inequitable and variable take up'* with time and resource-poor organisations effectively being disadvantaged and entrenching skills imbalances across the sector. In the light of these costs it is unsurprising that the Lloyds Digital Index (2016) found that charities were most likely to seek out informal and cheap sources of support such as the internet or peers. This is a clear opportunity for social media, although previous research by the author has found differences in Twitter take-up based on charity income, the platform remains a free and comparatively low impact investment which has the potential to host direct or indirect support from infrastructure organisations. The peer-to-peer aspect of support in particular is one which the literature suggests is valuable but lacking; with charities saying that they want to speak to similar organisations (Ógáin et al 2012) or that they have learned by *'seeing and doing'* rather than formal training (McCabe and Phillimore 2012: 11). The role of infrastructure organisations

may be to coordinate and facilitate this peer-to-peer interaction on social media, though there is no evidence in the literature of infrastructure organisations actually performing this role. Relevant to everything above, there is also the potential for social media to help infrastructure organisations coordinate amongst themselves and ensure they are providing complementary support to charities (The Comptroller and Auditor General 2009). Again, there is no evidence of this actually taking place in the extant literature, only a suggestion.

Given Twitter is primarily an information sharing platform, one form of support which is expected to be particularly prevalent on the platform is making secondary data resources, that is data sets or aggregate results, more available. Ógáin et al. (2012) report that, when asked what would help them progress, charities identified better access to analytical resources, such as data set, among other factors. This feeds into a common theme in the literature that charity awareness of data resources is generally poor (De Las Casas et al 2013; Macmillan et al 2014) and charities struggle to access data (Gyateng 2017). This is not solely due to weaknesses within charities however, on the supply side data access systems and procedures are often unclear or confusing (De Las Casas et al 2013; Gyateng et al 2013) particularly if a charity is trying to use data from multiple providers which have different procedures for accessing data (Gyateng et al 2013). Dayson (2010: 30) argues that it is part of the role of infrastructure organisations to *'raise standards by providing access to information... and establish forums of networking where they can share good practice'*. Though Dayson was not specifically referring to Twitter, it is easy to see how Twitter is the perfect platform for the provision of this form of support; many charities are already using it and it is inherently an information sharing tool. McCabe and Phillimore (2012: 3), writing two years after Dayson, acknowledge the role *'social networks'* should be playing in facilitating the provision of resources and knowledge to help charities work better. The same year Hall et al. (2012: 5) discussed charities sharing their own data to form a *'web of data'*. More recently Curvers et al. (2016: 21) advocated for the same approach, arguing that if funders encouraged their applicants to share their data as part of their funding agreement there would be an appreciation of value of the individual data as part of a *'network effect'*. Although these papers do not directly address the value of secondary data (that is data external to the third sector), they acknowledge the power of networks in making data more accessible, and therefore more valuable, and Chapter 5 will determine if this is true of secondary data in the Scottish third sector.

2.5 TRUST IN DATA

The final analytical chapter of this thesis concerns issues of trust in data, mainly external data, and this is the chapter where the distinction between primary and secondary data is most pertinent. The issue of trust in data has not been studied from the charity's perspective before, with previous studies generally concerning public trust in data (Cate 2008; Choldin 1988; Holt 2008) or public trust in how charities are using, primary, personal data (Morris 2005). Trust in data from the

charities perspective has the potential to have a significant impact both on how charities use data, as lack of trust could be a barrier, and on the quality of data overall, as charities may play a role in feeding back mistakes to the data producers.

Trust is integral to studying charity use of data, but it is not mentioned in the vast majority of the literature; the Lloyds Bank UK Business Digital Index, for example, is a well-funded large scale study with a sample of over 200,000 charities and does not mention issues of trust at all (Lloyds Bank 2016). Ellis and Gregory (2008) mention in passing that trust between charities and funders is part of what makes a successful relationship, but they do not elaborate. This is an important omission in the current body of literature and is one area where this thesis seeks to contribute. It is, therefore, difficult to integrate trust into a study of charity external data use because the foundation in the literature does not exist. Therefore, this section breaks down the concept of trust to determine different types and levels of trust so that charity trust in data can be suitably framed. Following from this, a discussion around the role of distrust leads to a consideration of the concept of 'Manufactured Risk' which will allow for a discussion in Chapter 6 of the role the third sector plays in maintaining wider trust in data.

2.5.1 The concept of trust

Trust is a capricious concept (Coulson 1998). Scholarship concerning trust dates back to, at least, Ancient Greece and has a complex history of competing dichotomies (Coulson 1998). More recently, however, there has been a focus on holism; that is moving beyond dichotomies by uniting the, principally economic, form of trust underpinned by risk and the more sociological conceptions of trust based on the work of David Hume (Lewicki et al 1998; Nooteboom 2002; Shapiro et al 1992; Tolbert and Mossberger 2006). Humean trust is focused on emotions and relationships between people who are seen as naturally trusting (Coulson 1998). Economic conceptions of trust, alternatively, assume rational calculations of risk; Coulson (1998) argues that trusting is to make oneself vulnerable, and risk that the other party will betray the trust placed in them. Together, these two conceptualisations lead to a more holistic view of trust; one which can have different levels and where distrust is not necessarily negative.

2.5.2 Levels of trust

The economic form of trust described above, one involving calculation of risk, sits at the midpoint of a three-part continuum of trust. This continuum is of indeterminate origin with several scholars converging on commonalities from different areas and at different times. The most common versions of the continuum are synthesised below:

Calculation/process/deterrence is the idea that it is safe to engage in trust because of an external authority which punishes breaking or betraying trust (Lewicki and Bunker 1995). The laws

imposed by governments and institutions are the most obvious illustration of this type of trust (Nooteboom 2002).

Experience/knowledge/characteristic is trust based on knowledge of the other party's characteristics which can be used to make decisions about to what extent to trust them (Shapiro et al 1992). The 'experience' tag used by Lewicki and Bunker (1995) reflects that this knowledge is usually derived from repeated engagement with the other party. This espouses the economic view of trust based on rational choices guided by information on the possible risks (Coulson 1998). Several scholars posit that this form of trust will develop from *calculation/process/deterrence* based trust after the two parties have sufficient experience of each other (Lewicki and Bunker 1995; Shapiro et al 1992).

Instinctive/identification based trust is the Humean view of trust. It reflects the trust between a dog and its master or a parent and child which is without calculation and is instinctive and tacit (Coulson 1998). This is the strongest form of trust.

2.5.3 Distrust

Only a few scholars have studied the concept of distrust in detail; often it is simply assumed to be the antithesis of trust and is posited as a destructive force (Kramer and Tyler 1995). McKnight and Chervany (2001) suggest this is an unhelpful way to classify distrust. They argue that it is possible to have both high trust and high distrust of the same person or party concurrently. The duopoly of simultaneous trust and distrust appears logically incongruous; however, distrust is not necessarily a lack of trust, but rather a distinct concept which is antithetical, rather than opposed (Lewicki et al 1998; McKnight and Chervany 2001). More than 100 years prior to McKnight and Chervany's discussion of distrust, John Stuart Mill's summarised the concept perfectly:

"democratic political culture is characterized by a vigilant skepticism [sic] (or realistic cynicism) rather than an unquestioning faith in the motives and abilities of political authority." (Citrin 1974: 998)

Mill's postulate espouses the idea that unquestioning trust, in this case in political authority, is bad for society and that a healthy distrust is key. This view is echoed in the modern literature (Chanley et al 2000; Cook and Gronke 2005; McKnight and Chervany 2001). A vivid modern example of a failure of distrust and 'vigilant scepticism' is the scandal surrounding Cambridge Analytica. Facebook management admitted not reading the terms and conditions of the app which Cambridge Analytica used to unethically collect Facebook user data (Romm 2018). If Facebook were more distrusting they may have taken steps which revealed the danger the app posed to their users before any damage was done. Facebook's failure in this instance, also relates to Giddens's theory of 'Manufactured Risk'.

2.5.4 ‘Manufactured Risk’

‘Manufactured Risks’ are risks induced by the advancement of science and technology in modern society, they are distinct from traditional risks in that it is difficult for the public to quantify them without access to esoteric and advanced knowledge (Beck 1992; Giddens 1990; Giddens 1999). Because the layperson is ignorant of most of the knowledge which allows experts to create and manage the risks of modern society, they are compelled to place trust in those with specialist knowledge such as scientists or academics; experts in their own areas (Giddens 1990; Giddens 1999). However, when trust is invested in experts and something goes wrong, that trust is damaged (Holt 2008). An example of this was the decrease in trust that occurred following the 2008 financial crisis which experts largely failed to anticipate (Krugman 2009). Therefore, as the specialist knowledge to effectively manage risk grows increasingly esoteric, a healthy distrust from experts in each other and in authority is paramount to maintaining wider trust (Beck 1992; Doyle 2007).

In the context of this thesis, the ‘Manufactured Risk’ is the possibility that data is inaccurate in a way which misguides policy decisions and causes harm. This data is produced by experts internal to the Scottish Government but, it is theorised that key to maintaining trust in government data is the knowledge of external experts who, with a healthy distrust or ‘vigilant scepticism’, can notify the government of errors in data before they impact on wider trust. These experts may be among the government’s stakeholders, especially charities and third sector infrastructure organisations who are often experts in their own areas and are independent of the state but also pro-social in a way which may make them inclined to feedback. This, potentially, important role in maintaining public trust in data means that charity data use may be more broadly important than might be initially assumed; charities may be both consumers of data and conduits through which data is spread. Charities may add their own meaning or trust to data as they process it for others to use further. The extent to which this is the case is examined in Chapters 5 and 6; the web of data and adding of meaning will be investigated in Chapter 5 while more direct issues of trust as discussed in Chapter 6.

It is also possible that the government has structures or procedures in place to manage these risks or to enhance the role of the experts who manage them, such as formalised feedback mechanisms. This would reflect Giddens’s (1999) and Beck’s (1992) conceptualisation of ‘risk society’; that is the way in which modern society organises itself to manage risk.

As mentioned in Section 2.3.2, data protection regulations, including the new GDPR, are related to ‘Manufactured Risks’ in that they are an attempt by the state to regulate the experts who process and control data resources; data protection is a safety net for public trust. In the general case, this is a good thing, trust which took years to build up can be destroyed in seconds and data protection

helps mitigate 'Manufactured Risks'. However, for charities who lack data skills these regulations can be difficult and onerous to navigate. There is a necessary trade-off in data protection between protecting public trust and making data easy for charities to access and analyse.

2.6 CONCLUSION

Establishing a foundation for this thesis entailed reviewing literature spread across several disciplines and areas. After defining several key concepts, it was established that data usage is likely beneficial to charities and the third sector as a whole and therefore it is desirable to seek to increase charity use of data, though this is not universally accepted. This led to a discussion around enablers of charity data use; factors which drive charities to use more data. The discussion of impact reporting inevitably resulted in the scrutiny of several interrelated barriers to this enabler and wider data use; staffing, skills, cost and size. The literature suggests that many organisations, particularly smaller charities, do not have the resources to employ a dedicated analyst or hire a consultant and must rely on up-skilling existing employees or volunteers. This makes it difficult for smaller charities to fully embrace data, creating a vicious cycle. The literature concerning fundraising makes the point that many charities are obliged to collect data by funders but that many struggle to make use of this data. The most pertinent barrier to fundraising was data security concerns, particularly in the wake of the introduction of new data protection legislation which appears to have the potential to have a notably negative effect on charity use of data. Several individual enablers and barriers were then discussed including the importance of charity leadership and strategic planning, the requirement for good IT infrastructure to underpin data use, and charity age.

Having reviewed enablers and barriers which affect individual charities' ability to use data, reviewing the literature for the second analysis chapter involved looking at what support is available for charities to help enhance their data use and overcome barriers. This thesis focuses on third sector infrastructure organisations and data organisations so these were first defined with the literature suggesting infrastructure organisations provide the majority of support to the third sector. Twitter, and charity use of Twitter, was then reviewed. The majority of charities are using Twitter and the platform seems to be more of an information exchange platform than a friendship network. Both of these factors make Twitter a perfect forum for the dissemination of support for charity data use. The final section covered the current paradigm of charity support and how Twitter may be able to improve it. The most obvious issue which Twitter may help resolve is access to data; many charities are unaware of what resources are available and data access procedures can be complicated. Given the information sharing nature of Twitter this may be how support is most effectively delivered on the platform. There was also a discussion around the cost of support, with Twitter being free and low impact though still requiring time to be spent, and around the breadth of support against its depth - Twitter may be able to provide both wide and deep support. Twitter may

also provide an example of how meaning and trust interact with data as it is passed around a network and this is addressed in Chapter 6.

The final part of this review covered issues of trust in data from the charities' perspective. This was a topic not directly covered by the existing body of literature and so the review attempted to construct a theoretical frame to underpin the discussions in Chapter 6. This involved conceptualising trust, which has a complex history, and defining a scale of trust made up of hierarchical levels. The most important part of this final section was the discussion around distrust and 'Manufactured Risk' which is, potentially, how charities aid in maintaining public trust in secondary data; a healthy scepticism of the data lets third sector organisations, who have specialist knowledge, feedback on errors and add context to data resources as discussed in Chapter 5.

CHAPTER 3: METHODOLOGY

3.1 INTRODUCTION

From the review of the literature in the previous chapter, it is clear that charity use of external data is understudied and, therefore, there is little established process or method for doing research on this topic. This presents an opportunity for this thesis. Therefore, guided by research philosophy, this project employs a series of methods, some conventional, some niche, to answer the given research questions. The first part of this chapter will briefly cover the research philosophy and how it underpins the mixed methods design. There will then be a discussion of the ethical challenges which had to be overcome, followed by details of the methods selected for each chapter and why these methods were selected. Chapter 4 uses survey data and mostly conventional statistical analysis. Chapter 5 uses Twitter data and network analysis including network modelling. Chapter 6 uses interview data and qualitative analysis. The sections covering each analysis chapter begin with a discussion of why the particular methods and data were selected for that particular analysis, they then detail the sampling and data collection, followed by a discussion around each individual method.

3.2 RESEARCH PHILOSOPHY

Research philosophy defines how data generated during research is understood and informs decisions on how data is to be collected and analysed. There is a philosophical divide in the academic community. The traditional view of social science methodologies is that quantitative and qualitative methods exist in an irreconcilable dichotomy; each underpinned by an idiosyncratic and, mutually exclusive, philosophy. This has been termed the *'incompatibility thesis'* and could be a potential issue for a mixed methods research project which seeks to understand both sorts of data (Teddlie and Tashakkori 2009). The *'incompatibility thesis'* is a valid perspective, however, this section will posit an opposing view; that quantitative and qualitative techniques are not necessarily in opposition and, provided they are underpinned by a suitable philosophy, they can be utilised within the same research project.

3.2.1 Social Constructivism

Social constructivism⁵ is a meta-theoretical orientation. Meta-theory, which is synonymous with philosophy, is a set of ontological and epistemological assumptions which determine lower level theory (Jachtenfuchs 2002). Social constructivism is a meta-theory with some idiosyncratic characteristics relating to its ontology and epistemology which make it ideally suited to mixed methods research.

⁵ There are several semantic variations on social constructivism. Firstly, the 'social' may be dropped without implying any distinction. Secondly, it can be spelled variously constructivism/constructionism (Burr 2003). The former will be used in this thesis but this implies no distinction from the latter.

3.2.1.1 *Ontology*

Ontology concerns the nature of being and the structure of reality (Crotty 1998). Constructivism is perceived by some areas of social science, notably sociology, as aligning with interpretivism; that is, largely qualitative approaches (Creswell and Plano-Clark 2007). This is a misconception; constructivism unequivocally shares its ontology with positivism:

“The manner in which the material world shapes and is shaped by human action and interaction depends on dynamic normative and epistemic interpretations of the material world.” (Smith 2006)

The material world plays a central role in constructivist theory. Ontologically constructivism is positivist and views the real world as separate from human cognition (Adler 1997; Burr 2003; Crotty 1998; Jackson et al 2006; McMahon 1997). This ontology makes constructivism compatible with quantitative methods.

3.2.1.2 *Epistemology*

Epistemology concerns how a meta-theory generates knowledge (Crotty 1998). Several scholars have equated constructivist epistemology directly to interpretivism and posited the idea that knowledge is created *ex nihilo* by the human mind and its experiences, or constructed by a network of human minds (Kukla 2000; Lincoln et al 2011). While not entirely erroneous, this assertion is incompatible with constructivism’s, previously discussed, positivist ontology. Therefore, constructivism does not share an epistemology with interpretivism, but it is equally distinct from positivism (Crotty 1998). Constructivism’s epistemology does not even lie along the spectrum created by the conflict of interpretivism and positivism as suggested by Christiansen et al (1999); it is unique, idiosyncratic, and transcends the traditional dichotomy (Crotty 1998). The key to constructivist epistemology is interaction:

“Individuals create meaning through their interactions with each other and with the environment they live in.” (Jackson et al 2006)

Centrally, constructivism concerns the interaction between human consciousness and the extant material world (Smith 2006). This epistemology does not have an official name, but, based on the work of Smith (2006), *interactionism* is perhaps the best descriptor.

3.2.2 **Structure and agency**

An example of *interactionism* can be seen in how constructivism approaches the structure and agency dichotomy. The structure and agency dichotomy concerns the relative importance which theoretical frameworks place on individual action versus rules and constraints (Giddens 1990). The traditional view is that positivist quantitative research focuses on the primacy of structure, that is recurrent patterns of constraints on individual action (Adler 1997), while interpretivist qualitative

research is better at understanding the agency of individuals (Smith 2006). Constructivism prioritises neither of these, instead taking a dialectical approach known as *structuration*. *Structuration*, first posited by Anthony Giddens (1974), introduces cyclicity to the structure/agency dichotomy. Wendt (1994) argues that as actors interact with the constraints provided by structures, they redefine those constraints for future agency. An example of this would be a judge referring to legal precedent, which is a structure, to come to a decision during a particular case; his decision is then an expression of agency. During the course of this expression of agency, new legal precedent is created which would constrain any further agency during a similar case in the future.

Given the above example, it is clear that *interactionism*, combined with positivist ontology, makes constructivism uniquely suited to mixed methods research.

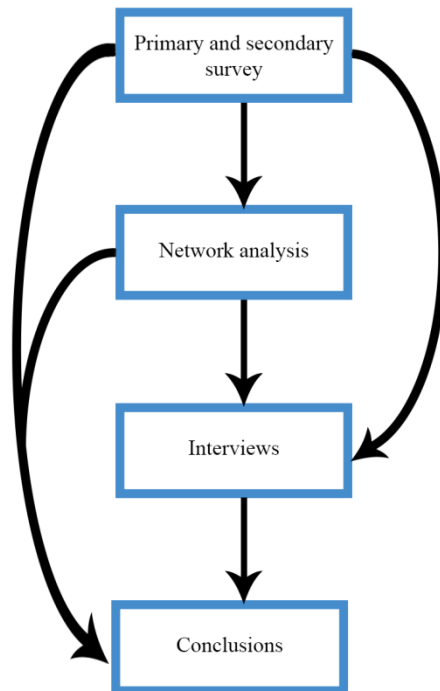
3.2.3 Mixed methods research

Beyond the philosophy which underpins mixed methods, it is desirable, Creswell and Plano-Clark (2007) would argue it is necessary, to justify why mixed methods have been employed. In this thesis, mixed methods have been chosen primarily because they are the best way to examine the research questions and topic at hand. Qualitative and quantitative methods answer different types of question and provide different insights (Creswell and Plano-Clark 2007). Mixed methods research can, therefore, provide stronger inferences, reconcile divergent findings (Teddlie and Tashakkori 2009), and can answer questions the individual methods cannot tackle alone (Creswell and Plano-Clark 2007). In this thesis, this is of particular use because the issue of charity use of external data breaks down into a number of issues which are best measured quantitatively and others which are best addressed qualitatively.

Constructivist research philosophy allows for the findings of mixed methods research to be integrated to form greater conclusions, however, philosophy does not provide any guidance on the sequence or order in which methods are used. This thesis has chosen to employ an explanatory sequential design where quantitative findings are followed by qualitative analysis to help contextualise and explain; the interviews play a corroboratory role to the thesis as a whole as well as providing insights on specific facets of the research questions (Creswell and Plano-Clark 2007).

The explanatory sequential design used in this research is complicated, to some extent, by the social network analysis of Chapter 5 which is primarily quantitative but also includes a partly qualitative content analysis. As shown in Figure 3-1, this means that each chapter of analysis is informed by the previous chapter and all three chapters of analysis come together to inform the final discussion and conclusions.

Figure 3-1 Sequence of analytical methods



3.3 ETHICS

Ethical issues for this research were complicated by the number of different techniques and data sources used, but simplified by the uncontentious nature of the topic of study. Informed consent was an interesting issue which varied across the types of data collected; for the secondary survey consent had already been obtained during the original project to use the data for further research and in the primary survey consent was gained by use of a preamble and tick box. For the interviews, the participants read an information sheet and signed a consent form, also giving consent for audio recording, but in the network analysis chapter, consent was more complex. In a strict sense, informed consent was not necessary for this chapter because the data used was entirely public and Twitter’s privacy policy informs users that their tweets may be used in academic research (Fiesler and Proferes 2018). Despite this Fiesler and Proferes (2018) found that most users were unaware that their content may be used in research and generally had not read the privacy policy. This presents a difficult situation ethically as there are no norms and many open questions about how to deal with consent issues in social media research; generally for large scale social networking, it is unfeasible to obtain direct consent from each participant (Schechter and Bravo-Lillo 2014). These issues are becoming more pertinent as social media research becomes more common but this research should be largely unaffected, primarily because of its uncontentious nature; the accounts under study are organisational accounts rather than personal. There was a focus throughout this research on maintaining the anonymity and confidentiality of the respondents and this is particularly relevant in this chapter to quell issues of consent. This meant not labelling

sociograms, not discussing who important nodes were, and paraphrasing quotes in the content analysis. The Twitter data was also stored as if it were sensitive research data.

All of the various research data generated or obtained during the project was stored in accordance with data protection principles as if it were sensitive and personal. This meant keeping the data on password protected computers and university network space. This thesis will also adhere to data lodging requirements set by the University of Stirling and Economic and Social Research Council (ESRC). The university requires that the research data generated by the project is held for a period of ten years from the cessation of funding in March 2018 and that it is then destroyed (University of Stirling 2015). The ESRC require that a fully anonymised version of any quantitative data be offered to the UK Data Archive within three months of the end of the project (ESRC 2014).

Full details of all ethical procedures and documentation can be found in the Appendix (Appendix I, Appendix II, Appendix III, Appendix IV, and Appendix V).

3.4 SURVEY STATISTICS: METHODS FOR STUDYING CHARITY DATA USE

The first chapter of analysis looks at levels of external data use as well as factors which enhance or impede charity data use. It was decided at the beginning of the project that a survey instrument would be the best way to collect data on external data use.

The first chapter of analysis sets the context for the following chapters; determining the level of external data use among charities before the following chapters addressed issues of support and trust which affect use. Therefore, a wide-but-shallow approach, a survey, was deemed to be the most appropriate method to set the context for the following chapters to build upon. This is in line with the project's explanatory sequential mixed methods design.

3.4.1 Software for survey methods

Stata was the primary software used in the first chapter of analysis. Stata is a general statistical analysis package and, along with user-written extension modules, is capable of performing every method and technique used in this part of the thesis. The only other software used in Chapter 4 was Excel for downloading the primary survey data from Bristol Online Survey in CSV format. This was converted into Stata format for encoding and analysis.

Stata is a syntax based environment which has advantages for the transparency and replicability of this research. Syntax files for this thesis are available on the author's GitHub (tomwallace1990/charity_data_PhD).

3.4.2 Sampling and data

The survey data breaks into two; a primary survey collected by the researcher and a secondary survey collected by Rutherford and Burt on a previous project. The intention at the outset was to use only a bespoke primary survey but, as discussed below, this instrument did not achieve the number of responses required and so the secondary survey became the main data source for Chapter 4 with the primary survey only being used in the first analytical question.

3.4.2.1 *Primary survey*

The primary survey was collected online between January and May 2016 using Bristol Online Survey. There were two versions of the survey initially, one focused on charity data and the other on immigration data; the intention at this early stage was to focus on the users of specific types of data, rather than necessarily charities, although charities were heavily implicated as users of this data and were heavily targeted. A second version of each survey, which was shorter and easier to complete, was launched in May 2016 and attempted to boost the number of responses. Respondents were sourced in two ways. The first was through direct email targeting of charities, organisations, and individuals who the researcher determined may be users of charity or immigration data based on online research. This included charities with a publically stated aim related to immigration or other third sector organisations and academics who research immigration or the third sector. The second sourcing method was dissemination through the email lists and newsletters of infrastructure organisations and the Scottish Government. Access to these lists was gained through contacts sourced from the supervision team and government sponsor, and direct email contact by the researcher⁶.

All versions of the survey included a name generator question which asked respondents to list the contact details of any of their peers who may be interested in participating in the research. The intention was that the survey would spread through the network of data users as responses snowballed. Unfortunately, this was not the case as response rates were low and many respondents did not answer the name generator question. The final number of valid responses for the primary surveys in total was 42. The project specification then changed to focus specifically on charities and this reduced the case numbers to 20. As the survey was partly distributed through email lists, the full sampling frame is not known, and therefore the overall response rate cannot be calculated. However, based on the directly targeted emails, the rate was around 8%. This is not unusually low for internet-based survey research (Hill 1998) and on reflection, choosing to focus only on two data topics, charity data and immigration data, was too restrictive.

⁶ All versions of the primary survey instrument can be found on the author's GitHub (https://github.com/tomwallace1990/charity_data_PhD)

It is clear that the primary survey data is not suitable for detailed statistical analysis. Therefore, the secondary survey was employed as the main data source for Chapter 4 while the primary survey was used descriptively in the first analytical question as it asked for information on where respondents sourced data which was not available in the secondary survey or anywhere else. Consideration was given to linking the primary survey data with the secondary survey data, but harmonisation of the main dependent variable, external data use, was deemed too statistically noisy for the minor increase in cases and several other variables could not be harmonised at all as they had no direct equivalents.

3.4.2.2 Secondary Survey

The secondary survey was collected by Rutherford and Burt in 2014. The survey was piloted with a group of five executive managers from the Scottish third sector. A sample of 1,000 charities with income in excess of £1 million in 2013 was drawn from the Office of the Scottish Charity Regulator's (OSCR) Charity Register. Organisations which register with OSCR but are not conventionally considered part of the third sector were manually removed from this sample giving the survey a final sampling frame of 704. The questionnaire was posted in hardcopy form to the chief executive of these 704 charities. 161 responses were received giving a response rate of 23% but once the data was cleaned the final number of charities dropped to 154 with a good spread of organisational sizes within that data.

This survey covers charity use of data in a more general sense than external data use which is the focus of this project. It does, however, include a relevant external data use question (*'What data sources are most useful to your charity: external data?'*) which was employed as a dependent variable (discussed in detail Section 3.4.2.3 below) alongside a series of enablers and barriers to data use as well as various organisational characteristics. Most of the barriers and enablers covered previously in the literature are present in this data. 154 cases is suitable, if not ideal, for statistical analysis. The, relatively, low case numbers obtained by both the primary and secondary survey instruments reflects the difficulties of sampling charitable organisations; the survey may not reach the right person within the organisation, the staff may feel too busy to give time to research, and there are a finite number of charities to sample from.

In light of these data collection issues it was decided to enhance the secondary survey data by linking it to data produced by the Office of the Scottish Charity Regulator (OSCR) which contained annual returns information for the year 2014. This linkage provided access to a range of, primarily financial, indicators for each charity such as their gross income and which sources they receive their funding from. This linkage was facilitated by the inclusion of charity numbers in both data sets and was not onerous. The Rutherford and Burt data was already encoded, clean, and ready for analysis in Stata format so very little preparation was required to utilise it in this project beyond the

aforementioned linkage and the preparation of the dependent variable as described in Section 3.4.2.3.

3.4.2.3 *Dependent variable*

The primary dependent variable for the analysis of charity use of external data in Chapter 4 is derived from a question in the secondary survey and is displayed in Table 3-1.

Table 3-1 Dependent variable from secondary survey: External data use

What data sources are most useful to your charity: external data?	Frequency	Percent
0. Not using (missing)	31	20%
1. Not useful	17	11%
2. Extremely limited use	23	15%
3. Moderately useful	53	34%
4. Extremely useful	30	19%
Total	154	100%

Source: Secondary survey

This variable will be treated as a close proxy for the concept of ‘use of external data’; be it ability to use, volume of use, or quality of use. The wording of the question is not conducive to specific interpretations and so it is only appropriate to treat it as representing some form of ‘use’ rather than trying to enforce a specific meaning.

This variable also includes a distinction between whether data is being used at all (the difference between category 0 and any of the other categories) and a scale of use (categories 1 through 4). A transformation was, therefore, applied to the dependent variable to aid in the analysis. This involved creating new configurations of the variable reflecting the binary information and the scale information contained in the variable. For the binary variable, all of the non-0 categories were coded into category 1 to create a variable which distinguished between using in any sense and not using. The second configuration involved removing category 0 which left only a scale of responses which indicate the use of data to some extent. The three different configurations of the dependent variable (full, binary, and scale) should allow for factors which affect external data use to be assessed in terms of how they affect use; does size affect the scale of use while lack of skills affects using at all?

3.4.3 **Methods of analysis**

A variety of statistical methods are applied to the survey data in Chapter 4; generally, the analysis begins descriptively, proceeds to bivariate testing and exploration of relationships, and finishes

with multivariate statistical modelling. However, there are several other methods, such as graphing and factor analysis, which are applied where relevant. All of the methods utilised in Chapter 4 are discussed below.

3.4.3.1 Descriptives

Analytical question 1, which sets up the rest of the analysis in the chapter, utilizes only descriptive methods. These methods involve tabulating responses from both surveys so they can be explored and discussed in terms of the number of respondents which selected each category. Although this process is not analytical, it is an important antecedent for the analysis which follows.

3.4.3.2 Graphing

A histogram is the only graph used in the first chapter of analysis. A histogram is a univariate visualisation which simply plots the values of a variable against their frequency (Gomm 2009). The histogram used in Chapter 4 is generated with Stata's inbuilt graphing tools and is used to visualise the distribution of the dependent variable when it is first described. This is important because the following analysis relies heavily on describing how other variables explain variation in this dependent variable.

3.4.3.3 Bivariate analysis

The bivariate analysis in this chapter involves calculating a measure of association for a given relationship along with an associated hypothesis test. Goodman and Kruskal's gamma is used throughout the first chapter of analysis as a measure of rank correlation. This is appropriate because, regardless of the distribution of the independent variable, the dependent variable - external data use - is ranked and therefore gamma is an appropriate statistic (Bernard 2012). Maintaining the use of a single test also aids comparability. Paired with each of these gamma tests is a chi-squared hypothesis test to determine the statistical significance of each relationship and advise on the acceptance, or rejection, of the null hypothesis (Bernard 2012). As case numbers are low in this chapter, chi-squared results below 0.1 were considered marginally insignificant but noteworthy, while results below 0.05 were considered significant in line with established social science thresholds (Gorard 2003). The bivariate analysis does not bear particular weight in this chapter, mostly acting as an exploration and primer for the proceeding regression analysis which is more robust and insightful.

3.4.3.4 Regression

Regression analysis is one of the primary methods used in Chapter 4. Primarily this chapter uses logistic and ordered logistic regression because the dependent variable, external data use, is categorical or binary, dependent on its configuration as discussed in Section 3.4.2.3.

Where linear regression assesses the joint impact of a series of indicators (independent variables) on the magnitude of an outcome (dependent variable), logistic regression predicts the likelihood of a binary outcome (Long and Freese 2014). This makes it particularly useful for assessing which factor best predicts binary data use; that is using or not using data without regard for the level of use. Ordered logistic regression uses similar foundations to predict on a scale of ordered outcomes rather than a binary outcome. This form of regression is particularly useful for assessing what factors predict the level of data use for those using it to some extent; variations in the scale of external data use (Long and Freese 2014).

There are several tests and statistics generated by regression, used throughout this analysis, which warrant further discussion. Regression generates several important variable-level statistics which are used in combination to assess individual factors within a model. The first of these is the coefficient (often shown as '*coef.*' or ' β ') which reflects the marginal impact that each independent variable has on the outcome (Gelman and Hill 2006). In the case of logistic style regressions, this number is the increase or decrease each variable contributes to the log-odds of selecting binary category 1 (Long and Freese 2014). This coefficient is an estimate and is, therefore, accompanied by a standard error which is used to calculate a p-value (often shown as ' $P > |z|$ ') which helps determine which coefficients are likely to be observed by random chance and which are of sufficient magnitude to be considered 'significant' (Gelman and Hill 2006). The standard errors are also used to calculate confidence intervals for each effect which can be useful when comparing the relative impact of different coefficients.

Moving to model-level statistics, just as each coefficient has a P-value, the model itself is given a P-value (shown as '*Prob > Chi2*') which is the chance of obtaining the given likelihood ratio chi-squared value if the null hypothesis is true. The given likelihood ratio chi-squared value is equal to two times the difference between the log likelihood of the null model and converged model (Idre 2017). In other words, the model P-value tests the chances of seeing the total effects of the independent variables in combination (the likelihood ratio chi-squared value) if there is no effect. If the model is significant, $\text{Prob} > \text{Chi}^2$ is below 0.05, there is a low chance of observing an effect of the combined independent variables of equal magnitude to that observed by random chance. This P-value may be significant where all of the model point estimates are insignificant as it measures their cumulative impact on the outcome. The second major model-level statistic is pseudo-R-squared which is an approximation of the R-squared statistic found in linear regression. R-squared measures the proportion of variance in the dependent variable which is explained by the combination of the independent variables (Andersen 2008). Conventionally, this is considered a goodness-of-fit measure as it can be considered a measure of improvement from the null model to the fitted model (Idre 2011). R-squared is not compatible with logistic regression and so Stata, by default, displays McFadden's pseudo-R-squared which attempts to approximate both the variance

explained and goodness-of-fit aspects of R-squared by adapting the principles to fit with the different procedures used in fitting logistic models. Pseudo-R-squared is not as robust as conventional R-squared but it is still useful for reviewing and comparing models. Pseudo-R-squared is not, however, ideal for comparing between models of very different specifications and, therefore, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are also used for comparing models. Both of these methods are penalised-likelihood criteria and can be used to assess non-nested models in terms of how close they are to ‘*truth*’ (Claeskens and Hjort 2008). In other words, they assess the explanatory power of a model against its complexity and therefore determine the relative parsimony of models. BIC and AIC attempt to achieve similar goals based on differing underlying assumptions and have different strengths and weaknesses. BIC, for example, becomes increasingly less likely to choose too big a model as n grows and while AIC does not gain this additional precision, it is generally less likely to choose too small a model. Generally, it is good practice to use both BIC and AIC simultaneously as part of model selection (Claeskens and Hjort 2008).

3.4.3.5 *Factor analysis*

Factor analysis is a method which assesses groups of variables to determine if they can be represented by a smaller number of factors; it searches for latent variables, that is, variables which are not present in the data but can be revealed by a combination of other variables (Thompson 2004). Factor analysis can be used to explore variables for factors or to confirm hypothesised factors in groups of similar variables. Latent variables are also often used to simplify analysis when a researcher wishes to avoid analysing a large number of similar variables (Thompson 2004). However, in this analysis, factor analysis is used for two primary purposes.

Firstly, it is used to support the hypothesised interpretation of several interrelated survey questions; due to survey measurement error, it is not always clear how respondents have interpreted the wordings of particular questions and conflation of questions by respondents is possible. The most striking example of this from the analysis in Chapter 4 are the barriers ‘Data security concerns’, ‘Data privacy concerns’, ‘Ethical concerns’, and ‘Legal/regulatory constraints’. These all have similar interpretations and appeared, in bivariate tests, to vary together and therefore it was hypothesised that they have, to some extent, been conflated by respondents and are actually components of a ‘Data protection’ latent variable. Factor analysis helps confirm this interpretation by measuring the shared variation.

The second use of factor analysis in this thesis is related to modelling; where variables are very similar and express similar variation they lead to significant collinearity when modelled together. Collinearity occurs when independent variables in a regression predict each other with a high degree of accuracy and this situation can not only bias the individual results for those variables but

also notably inflate the pseudo-R-Squared value (Long and Freese 2014). Factors can help quell this collinearity by combining the shared variation into a single independent variable. Even where factor models are weaker than full models in terms of their significance they can provide a useful reference for pseudo-R-Squared values where collinearity has been controlled for.

3.5 NETWORK ANALYSIS: METHODS FOR STUDYING SUPPORT FOR CHARITY DATA USE

Support is a concept which could be measured with both survey and interview methods, as used in the other analysis chapters, but both of these methods source data based on the opinions of respondents. Information sourced directly from respondents is, of course, highly valuable, but measuring the concept of support by direct observation was deemed to be the best possible source of data and new administrative data has made direct observation possible on a large scale.

Direct observation has advantages in its epistemological concision; there is no distortion brought about by perspective or opinion as the subjects of study are directly observed. This form of data collection has traditionally been the reserve of qualitative anthropology but the increasing importance of the internet in society and data available on internet activity has made this method available to other areas of research. For the purposes of this study, the ubiquity and openness of social media is crucial for sourcing data on support for charity data use; Twitter was chosen as a case study because it is particularly easy to collect data from and is commonly used by third sector organisations (Guo and Saxton 2014).

3.5.1 Software for network analysis

This chapter of analysis uses the largest variety of software packages to achieve its aims. The primary software, used for both data collection and the majority of analysis, is NodeXL which is an extension of Excel. The majority of network analysis and graphing can be performed with this software and it supports easy handling of node attributes, such as organisational group, which other software is less well optimised for. NodeXL has some stability issues, bugs and incompatibilities, but was still deemed the best tool for the collection, storage and processing of the majority of this chapter's data.

The ERGM analysis is where more specialist software had to be employed, both in the preparation of data and the analysis itself. ERGM can be run in several statistical packages but PNet was chosen for this analysis due to researcher training. PNet can only read from raw matrix text files and, as NodeXL holds network data in edge list format⁷, a third piece of software had to be

⁷ An edge list is simply a two column width list of senders and receivers. It is an intuitive network data structure but generally more complex network analysis software will hold the data in matrix format. In the

employed to transform the data between these formats. UCINET was chosen for this purpose due to its ability to read and write many different network data formats. UCINET was also used to perform several analyses which were not used in the final version of this research, including; Quadratic Assignment Procedure, E-I index, K-Cores, and Triad Census. Results of these methods can be found on the GitHub repository for this project (tomwallace1990/charity_data_PhD).

3.5.2 Sampling and data

3.5.2.1 Sampling

Before any data could be collected, a sample of accounts (Twitter handles) had to be obtained. The core of the Twitter sample is made up of organisations who responded to the secondary survey used in chapter 4. The original intention was to link social network data to the survey for combined analysis, but as the methods used to analyse network data are so distinct from survey methods, this section became a separate chapter and the linkage was dropped while these charities were kept as a core sample. This resulted in a sample of 125 charity handles, with the remaining 29 charities not owning Twitter accounts. A group of 79 infrastructure organisations was added to this. This group comprised of funders and organisations identified by the SCVO as support organisations (Scottish Council for Voluntary Organisations 2016). Handles for both the charities and support organisations were sourced by searching Twitter for the organisation's name and, if that failed, searching for Twitter information on the charity's own website.

Finally, a set of data organisations was added to the sample which involved searching Twitter for the keywords 'data' and 'statistics' and identifying handles which used these terms frequently – resulting in 160,104 accounts being identified. Networks were then constructed for this extremely large group alone and with the previously collected handles to identify the most influential and well-connected accounts based on centrality metrics, which are discussed below in Section 3.5.3.3. Accounts with a combined degree score of 6 or more in the influence network and any connection to the charities network were retained. Personal accounts were then removed, this left 28 active and connected data organisations. This 28 included several Scottish Government data-related accounts. Finally, the Scottish Government's primary handle was also added, resulting in 233 accounts in the final sample as shown below. The charities sample only contained Scottish organisations while the support and data groups were limited to the UK; this reflects that support for external data use in the Scottish third sector is not necessarily limited to Scottish organisations.

raw matrix format, each node is represented along the columns and rows and contact between them is represented as a binary 1 or 0.

Table 3-2 Sample of Twitter handles split by group

Group	Number of handles in sample	Geographical scope
Charities	125	Scotland
Support organisations	79	UK
Data organisations	28	UK
Scottish Government	1	Scotland
Total	233	

Source: Primary Twitter data

3.5.2.2 *Data collection*

The data was collected through the Twitter Application Programming Interface (API) using NodeXL. This collects up to 3,200 tweets from each user from newest to oldest and creates a link between users where one mentions the other using the '@' symbol. The users become *nodes* or *vertices* while the links between them are referred to as *edges* (Knoke and Yang 2008). The initial data imported is mostly tweets which contain no linking information and which need to be removed. This becomes problematic when trying to collect data over a period of time as accounts which post very frequently may exceed the 3,200 limit within only a few months while less active accounts may have tweets stretching back years accessible through the API. The inability of the API to only supply tweets with linking information means that the 3,200 tweet limit is problematic for data collection. For this reason, data was collected once a month, for twelve months, which ensured that every tweet, sent by every account, was captured in the period between 13/09/2016 and 13/09/2017. Once duplicates and tweets with no links had been removed this left 12,756 tweets between 211 accounts. 22 handles did not feature in the network over this year of data collection. Of these 22, 13 were charities, 8 were support organisations, and 1 was a data organisation and these are therefore excluded from the analysis. Because this data is based on mentions between accounts it will be referred to as the *mentions data* or *mentions network*.

On the last day of data collection (13/09/2017) a '*following*' network was also collected. Rather than linking accounts by direct mentions, this network connects them by following behaviour. This data did not need to be collected longitudinally as following networks are much more stable and the whole network can be downloaded through the API in a single day. The following network is slightly bigger than the mentions network with 218 accounts and a similar pattern of missing. The mentions and following data are summarised in Table 3-3 and then discussed individually below.

Table 3-3 Summary of mentions and following data

	Mentions data	Following data
Number of accounts	211	218
Number of total links	12,756	13,312
Number of unique links	2300	13,312
Density	0.052	0.141
Type of tie	Events	Relationship
Ties are duplicable?	Yes	No
Mode	One mode	One mode
Directed?	Yes	Yes

Source: Primary Twitter data

3.5.2.3 *Mentions data*

The mentions data is one-mode, meaning that all of the nodes were collected in a uniform way and are considered to be of a uniform type (Prell 2011; Wasserman and Faust 1994). This is despite the different groups of organisations in the data; a two-mode network would comprise two distinct types of nodes, such as individuals and events which link them. Ties in the mentions network are duplicable, meaning that contact between two given organisations can occur more than once and each instance is recorded as a separate tie. This allows the network to be ‘valued’ by counting and removing the duplicated edges and applying an ‘edge weight’ to represent how many times the contact took place over a particular dyad (that is any given pairing of two nodes). With duplicates, the network has 12,756 links, with these removed there are 2,300 links which gives the network a density of 0.052, which means that just over 5% of possible ties are present. The data is directed, meaning ties are directional from one account to another; there is a sender and receiver for every link. Directed networks can have mutual, or, reciprocal ties where the link exists in both directions. The directional information can also be used to investigate where particular accounts, or groups, tend to send out more links than they receive or vice-versa; reflecting popularity or activity of Twitter use.

3.5.2.4 *Following data*

The following network is one-mode and directed, like the mentions network, but ties are non-duplicable as accounts cannot follow one another more than once. This means that the network cannot be ‘valued’. The following data is much denser than the mentions network with 13,312 links between the 218 accounts, giving the network a density of 0.141. This relatively higher density is likely due to the maintenance free nature of following links; mentions are events whereas follows are relationships which, once established, persist maintenance free. It is important to collect both

mentions and following data for the Twitter network because information can spread by either type of link.

3.5.3 Methods of analysis

Despite being collected over a period of time, the network data is analysed with cross-sectional network methods. While longitudinal methods, such as relational event models, would yield very interesting results, they would not directly answer any of this project's research questions and would, therefore, be extraneous. The majority of methods used in this chapter can be classified as network analysis methods. Generally, conventional statistical techniques cannot be applied to network data because the data violates the assumption of independence of observations (Berry 1993). This assumption is that the cases in the data are independent and not connected to each other and, while this is often not the case in conventional data, for example in household surveys, cases in network data are explicitly interrelated. Indeed, one of the primary aims of network analysis is to study the patterns of this interrelation and so conventional statistics are inappropriate and specific networking methods are used instead. However, there are a few sections of analysis which use more conventional methods. The first is an analysis of Twitter usage metrics, which are not network data but rather measures of use of the Twitter platform such as counts of numbers of tweets. Analysing this data involves using quantile regression, bootstrapping, and quasi-variance. The second non-network analysis is the content analysis, used in answering question 4.b. These methods for non-network data are reviewed first, before discussing the network analysis methods.

3.5.3.1 Quantile regression, bootstrapping, and quasi-variance

Quantile regression is a rare but relatively simplistic form of linear regression which estimates on the specified conditional quantile rather than on the mean as with a conventional ordinary least squares regression. In this analysis, the 50th percentile, or median, is used and the intention is to quell the effect of highly skewed Twitter usage data. Analogous to income data, many Twitter metrics often exhibit extreme positive outliers and so are more usefully analysed with medians rather than means. Other than this distinction quantile regression is similar to conventional linear regression in that it predicts the joint effects of a series of predictors on the variation in an outcome.

The intention of the quantile regression in the analysis of support for charity data use is to differentiate between the Twitter metrics of the different groups of organisations to determine if any of them use Twitter more or less than the others. This is important context for a later examination of the interaction of the groups on Twitter. Estimating on the conditional median goes some way to accounting for the skew in this data but a decision was made to employ bootstrapping to further quell any issues present in the data. Bootstrapping is a resampling method which allows for the more accurate calculation of statistics for non-normally distributed samples. Many statistical methods make assumptions about the normality of their input data and the relationship between a

sample and distribution of the population. Bootstrapping helps resolve issues where a sample is not normal by randomly resampling, with replacement, from the sample and computing a new median. This resampling is repeated many times over and the medians from each stage are formed into a new distribution which helps determine how much the median varies. The result of this, in most cases, is more accurate statistics and, in a regression framework, smaller standard errors and tighter confidence intervals which should help identify any difference between the groups. The only disadvantage of bootstrapping is that it is computationally intensive which was not problematic for this application.

Using bootstrapped quantile regressions gives the best chances of observing differences in Twitter metrics among the different groups of organisations, but, when comparing the categories of a variable in a regression analysis in this way, the reference category problem becomes apparent (Gayle and Lambert 2007). Regression analysis manages categorical variables by excluding one category as the reference; all other categories are compared to that excluded category. When trying to compare idiosyncratic and non-ordinal categories, as with the groups of organisations in this application, this is problematic for two reasons. Firstly, in a strict sense categories cannot be easily compared; each category is only comparable to the base category. For example, if data organisations were the base category, charities and support organisations would not be comparable, defeating the purpose of the analysis. Secondly, and more significantly, the coefficient for the base category is set to 0 and the other categories are said to be significantly different from the base category if their coefficient is significantly different from 0 which is determined by their standard errors. The issue is that the base category, being set to 0, does not have standard errors calculated for it; there is uncertainty around its estimate but this is not calculated and this could potentially lead to a type I error in terms of the categories being significantly different. The solution to both of these issues is quasi-variance, a niche method which calculates a quasi-standard error for each of the categories including the base category (Firth and De Menezes 2004; Gayle and Lambert 2007). This is achieved by proportionally redistributing the total uncertainty to derive quasi-standard errors for each category. These quasi-standard errors are also inter-comparable in that any category can be compared to any other category which solves both of the issues of standard regression and allows for comparisons between the groups in this application.

3.5.3.2 *Content analysis*

Content analysis, alternative to the other methods in this chapter, is not wholly quantitative. However, its use in this thesis does not constitute fully qualitative content analysis either; it is a hybrid. This method involves the analysis of written language through a mixture of count-based and interpretive techniques; reflecting its hybrid nature. Compared to many content analyses found in the literature (Gálvez-Rodríguez et al 2016; Saxton and Guo 2014) the method's application in this thesis is relatively simple; the intention of the content analysis is to determine what is actually

discussed on Twitter between the groups and what type of support for data use takes place. This means that tweets can be analysed through aggregate counts and textual examples rather than any search for deeper meaning or subtext which is common in other applications, particularly in the sociological tradition of semiotics (Bryman 2012).

The count part of the content analysis begins by reiterating how active each of the groups of organisations is on Twitter by counting the average number of tweets per group. This section then details each group's use of retweets, URLs, and hashtags; URLs being of particular interest as they reflect sharing of information which is a significant form of support. This sets the context for a count of data related words used by the groups; the specific words to be counted are determined inductively from the data. This section will reveal any differences in the overall usage of data related terms by each group, but it also acts as a primer for the primary tweet-level content analysis.

This final part of the content analysis at tweet-level is divided in two. The first part analyses examples of commonly used general words and aims to explore the dynamics of general interaction between the groups not related to data. The second part, the crux of the content analysis, examines examples of the data related words which were counted previously and attempts to determine if there is evidence of support for data use on Twitter and what form this takes. This section of analysis does not differentiate between the groups in terms of their use of data related words, as this is covered more thoroughly in the following analytical question using a group connection table as discussed in Section 3.5.3.5. The discussion of methods now turns to more specific network analysis approaches.

3.5.3.3 *Median centrality*

Centrality is a node-level network analysis method which refers to how central a node is in a given network (Carolan 2014). There are several ways to conceptualise 'central' and, therefore, several types of network centrality. This analysis uses the four primary types of centrality; degree, closeness, betweenness, and eigenvector.

The simplest of the four types of centrality is degree. In a directed network, degree breaks into two components: in-degree and out-degree. In-degree is simply a count of how many other accounts send a link to a given account. This takes no account of how many links are sent, only how many sending partners there are (Robins et al 2009). Out-degree is the inverse; it is how many other accounts the given account sends links out to.

The first of the three more complex forms of centrality is closeness, which is a measure of centralisation. A given node's closeness score is equal to the inverse of the sum of all shortest paths

between the given node and every other node (Carolan 2014). Paths in this definition are links to other nodes, a direct connection between two nodes is a one-step path, while a connection that requires going through another intermediating node is a two-step path. In other words, closeness is a measure of the length of the links between the given node and every other node in the network, assuming the links are optimised to go by the shortest route. A node with shorter links will have a higher closeness centrality and be more central to the network.

Betweenness centrality is a measure of intermediation. Similarly, to closeness centrality, it is defined by shortest paths and, in this case, the number of shortest paths which pass through the given node. Paths allude to how information flows around a network and the intention of betweenness centrality is to capture brokerage or gatekeeping; having shortest paths pass through a node means that that node is linking other nodes together (Knoke and Yang 2008). In extreme cases, a few nodes will link entire components of a network together and will have very high betweenness centrality scores. Alternatively, nodes which are on the peripheries of a network and only connected by one link have no shortest paths passing through them and a betweenness centrality of zero.

Finally, eigenvector centrality is a measure of popularity similar to in-degree centrality but instead of a simple count, it weights connections based on the relative prominence of the nodes. This form of centrality works by assuming that nodes which have many links are more prominent than those with fewer and therefore values connections to these nodes more highly (McCulloh et al 2013). Burris (2004) argues that this weighting of relations is similar to Bourdieu's notion of social capital in that it acknowledges that not all connections are equally strong or valuable. PageRank, the system used by Google to sort its search results is conceptually very similar to eigenvector centrality.

Having described each type of centrality, it is important to note how they will be collated and displayed in the analysis. Centrality metrics are node-level which means they must be gathered into distributions to be analysed at the group level. Given these distributions can be quite skewed, a decision was made to employ medians to descriptively explore differences in centrality across the groups. The choice to employ medians meant that standard deviations, to help identify the spread of the data, could not be used and so Median Absolute Deviations (MAD) were employed. Median Absolute Deviations simply sum the absolute deviations of each data point from the median and then take the median of the distribution formed by those deviations (Everitt and Howell 2005). Median and MAD results for the group centrality scores are displayed in tables.

3.5.3.4 *Sociograms*

Sociograms are network visualisations. In a sociogram, each node is depicted as a point of some type, usually a shape to help differentiate different groups of nodes, and links between the nodes are drawn as arrow headed lines (Scott 2012). Where each node is placed is determined by an algorithm chosen by the researcher. These algorithms are generally ‘force-directed’ and attempt to keep nodes an equal distance apart and minimise the number of edges crossing other edges which makes graphs cleaner and easier to read. Within this class, there are many individual algorithms for drawing sociograms and this thesis makes exclusive use of the Harel Koren fast multi-scale method (Harel and Koren 2000). This algorithm was developed in the early 2000s with the intention of creating a faster way of drawing large graphs. While Harel and Koren’s speed optimisation is still appreciated, advancements in computing power have reduced the need to select an algorithm for its efficiency. The Harel Koren algorithm is still useful, however, because as a by-product of efficiency optimisation it is very good at identifying clustering and network structures (Harel and Koren 2000). This makes the features of networks easier to view and discuss, improving the visual presentation of the network.

3.5.3.5 *Group connection tables*

Group connection tables are a simple but powerful network descriptive which is particularly useful for studying networks which feature distinct groups of nodes, as in this research. A group connection table is simply a square matrix with each group laid out along the rows and down the columns (see Table 3-4). The columns represent links being sent, while the rows track where they are received (Wang et al 2009). This means that, rather than simply looking at how many links a group sends out or receives, this table can track which other group links are sent to and where they are received from. This is particularly useful when groups exhibit a significant degree of homophily in communication, that is, they send most of their links internally. If a method simply counts links sent and received then lots of internal contact and makes a group appear popular and active when this is mostly due to internal activity. By comparing the diagonal, homophilic connections, and off-diagonal, heterophilic connections, a group’s true popularity and activity can be compared. This makes this method particularly useful for this thesis, it is used in the exploration of the data-mentions network to track which groups send and receive data related tweets to one another. Without this matrix-style method, it would be difficult to determine where tweets supporting charity data use originated from.

Table 3-4 Example of a group connection table

		Receiving			
		1. Charities	2. Support	3. Data	
Sending	1. Charities	#	#	#	#
	2. Support	#	#	#	#
	3. Data	#	#	#	#
		#	#	#	
		In total			

3.5.3.6 *ERGM*

Exponential Random Graph Modelling (ERGM) is an advanced network analysis method which goes beyond descriptive or simple bivariate statistics by modelling network structures. ERGM is a tie-based method in that it estimates the likelihood of the presence, or absence, of ties which combine to form network structures. The goal of ERGM is to explain underlying processes which lead to network formation by examining network ties and structures (Robins and Lusher 2013). ERGMs find their beginning in Markov Models as described by Frank and Strauss (1986) and later P* models (Wasserman and Pattison 1996). However, these antecedent techniques tended to use maximum pseudolikelihood estimation which, while relatively fast, is biased. In the last fifteen years, both unbiased algorithms and the computing power to apply them to real world data have developed leading to an increasing popularity of ERGM as a method of exploring and describing networks (Robins and Lusher 2013).

As discussed previously, network data violates the assumption of independence of observations which makes it incompatible with standard regression analysis (Shields 2016). ERGM overcomes this issue by using a completely different approach to inference based on random simulation. ERGM uses Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE) which generates random graphs of the same number of nodes and density as the observed graph (Robins et al 2007). These graphs are generated in a Markov Chain in accordance with ‘theta’ values which the computer generates to manipulate the graph formation process; in each step of the chain a random dyad is selected and a tie is either added or removed based on the theta values. If there is a high theta for the triangle effect, ties will tend to form where they complete triangles for example. These random graphs are then sampled at set intervals along the Markov Chain to ensure they are independent of one another and compared to the original graph, if they have similar metrics on the factors which the researcher selected then the model has guessed the correct theta values for each factor and the model converges (Wang et al 2009). The theta values then become the main model output. If the computer had to set a large triangle theta to generate a random graph with a similar number of triangles to the original graph then this large theta implies that the original graph has

more triangles than would be expected by chance, controlling for other factors, and therefore triangles have a large effect size. This is how inference is gained in ERGM modelling; if an effect is notably larger or smaller than would be expected from a totally random network, net of the other effects, then it is reasonable to ascribe the factor to underlying and external forces (Robins et al 2007). In other words, if a graph has more triangles than random it is fair to assume that there is some external reason why the nodes are tending to form into triangles; ERGM cannot provide that reason and therefore the modelling must be underpinned by theory or other information on the network. This is what makes ERGM a hypothesis testing method and not a panacea for network analysis.

Besides the theta value, which becomes known as an ‘estimate’ once the model converges, and its associated standard error, the model also calculates a ‘T-ratio’ for each effect. The T-ratio is defined as the observed value for each effect, minus the mean of the sample of simulated networks, divided by the standard error. Effectively it is a goodness of fit for each effect and it must be within the range -0.1 to 0.1 for all effects for a model to be deemed a good enough fit to converge (Robins et al 2009).

The main complication of MCMCMLE is that the theta values are not independent and a change in one may have an effect on others. If the computer cannot find theta values which recreate the original graph, then the model will fail to converge. This must be countered by giving the model more time to search for the right values, or a greater number of short runs, which increases the estimation time and some ERGMs can take multiple days to converge.

ERGMs can estimate two distinct type of covariate: endogenous and exogenous. Endogenous covariates are network structures which are selected from a set list. These include: reciprocity, triangle effects, and star effects, among others (please see the PNet user manual for a full description of available endogenous effects (Wang et al 2009)). All of the effects are net of each other and the intercept, which is referred to as ‘arc’ and is the network density. The ability of ERGM to estimate effects net of one another is crucial when consideration is given to how many structures a single tie can be part of in a network; a tie always affects the density, and it may be part of a reciprocal pair, a triangle, and a star effect all simultaneously. This complexity is why simulation-based estimation is so important for the robustness of the method but also, as discussed above, why it is so computationally intensive.

Exogenous covariates, alternatively, are not factors internal to the network but external factors which may affect the formation of network ties; a classic example is a network of families linked by business relations which may be heavily dependent on the relative wealth of each family. A complicating factor with exogenous covariates, however, is that they must be defined across dyads

or in relation to ties and not nodes (Shields 2016). There may be interesting data available on nodes such as, in this case, the organisation group a node belongs to (charity, support organisation, data organisation) but this cannot be directly applied as a covariate because the units of analysis in ERGM are dyads; pairs of nodes. This means that, for the group variable, a dyadic effect must be defined. The simplest of these, for categorical data, is matching which is a simple binary dyadic effect which is defined as 1 where a dyad are of the same group and 0 where they differ. Two versions of this effect are specified simultaneously, one reflecting one-way ties and the other reciprocal ties. In the results, this effect will help describe if nodes in matching groups tend to be homophilous; in other words, does being matching in group make any given pair of nodes more likely to share ties or communicate with each other (Wang et al 2009). This effect, though simple and limited in explanatory power, also controls for the effects of matching in group for the other effects in the model and so may be included as a control.

A more complex application of exogenous covariates for groups of nodes is to use binary effects rather than matching. The logic of this approach is similar to the handling of categorical variables in conventional regression; the group variable is split into separate binary variables and these are included as individual binary exogenous covariates. One group must be left out to serve as the reference category. The advantage of this more complex approach to including the group data is twofold; firstly, more complex effects can be measured, and secondly, the individual groups can be compared. The latter advantage is important with groups, such as the organisations in this thesis, because there is no reason why a pair of charities matching in group should have the same relationship as infrastructure organisations matching in group; the simple matching covariate cannot distinguish this. The former advantage becomes apparent when viewing the results of a model which includes binary exogenous covariates; five effects can be specified for each group rather than two for the simple covariate. The first two of these, interaction and interaction reciprocity, are analogous to the matching one-way and matching reciprocity of the endogenous covariate but now give a result for each individual group compared to the reference group. For the three other effects, sender and receiver measure the activity and popularity of each group, while activity reciprocity gives the overall tendency of each group to reciprocate. Both the simple and complex approach to obtaining effects for organisational group are utilised in the analysis as they both give insights.

3.6 SEMI-STRUCTURED INTERVIEWS: METHODS FOR STUDYING TRUST

This final section covers the methods used in the final analytical chapter, which concerns issues of trust in data. Methodologically this was the most conventional chapter, employing face-to-face semi-structured interviews to discuss issues of trust in data as well as data use and support for data use through networking.

Interviews were selected as the best way of gaining rich qualitative data on issues of trust in data from respondents. Trust is a very personal concept which is hard to measure with quantitative methods and therefore a semi-structured discussion was deemed to be more appropriate than a questionnaire (Bryman 2012). The qualitative nature of the final analysis chapter also lent itself to playing a corroborative role with issues discussed, but not fully explored, in the preceding analysis chapters. For example, network analysis was the best method for studying Twitter support networks, but the final chapter allows for the views of the respondents to be added to the, mostly, quantitative evidence previously presented to add meaning, context, and richness. Similarly, barriers and enablers to use of data are covered by the survey evidence in the first analytical chapter but some of the factors, such as staffing, are more usefully explored in face-to-face discussions with respondents.

3.6.1 Software for qualitative analysis

As with the other chapter of analysis, a variety of software was employed in performing and analysing the interviews. Besides Word, which was used to store the transcribed data, two primary pieces of software were used to help organise the insights of the qualitative analysis. The main software aid was NVivo, a Qualitative Data Analysis (QDA) package which handled the initial coding of the transcripts. With the codes established, the quotes were then moved into Excel to be sifted and sorted to form the structure of the chapter.

3.6.2 Sampling and data

Twelve interviews were carried out by the researcher between the 18th of April 2017 and the 30th of June 2017. Three types of respondents were interviewed: frontline charities, infrastructure organisations, and individuals working within the Scottish Government. This spread of respondents allowed for trust in data and other issues related to charity use of data to be explored from different angles and perspectives. It was determined that this spread of responses was more valuable to the project than focusing on one type of respondent for all twelve interviews. Breaking down the twelve interviews; four were frontline charities, five support organisations, and three staff from the Scottish Government. One weakness of this section of analysis is the relatively small number of respondents who were interviewed; this reflected constraints on the research in terms of time, budget, and the positioning of the interviews as, roughly, one-third of the analysis in this mixed methods project.

Each type of respondent was interviewed with a different topic guide (Appendix I) which reflected their differing knowledge and experience. Frontline charities were asked about how they used and interacted with data, challenges or barriers to use, their trust in the data, and networking or support related to issues of sourcing, sharing, or analysing data. Infrastructure organisations were asked similar questions to charities but were also asked about what support they provided to frontline

charities for data use. This reflects the unique position of infrastructure organisations as both users of data and support for other charities use. The topic guide for the government respondents was more substantially different and was subdivided into two: one for government staff managing data resources and one for respondents concerned with the third sector more directly. For the data and survey respondents (two of the three government interviewees), the guide covered how aware they were of charities among their stakeholders, what support they provided, how they contacted their stakeholders, their use of social media, and direct issues of trust. For the single charity-focused government respondent, the guide covered their view on the capacity of the third sector to engage with data, the government's role in this and support for data use, use of Twitter and trust in data as a broad issue. The participants were also shown anonymised network maps (see discussion on sociograms in Section 3.5.3.4 and Appendix VI) to spur discussion around the use of Twitter. These proved quite successful with many respondents requesting to keep the map.

Respondents were recruited, mainly, through the primary survey instrument (detailed in Section 3.4.2 above) which asked respondents if they wished to be contacted for a follow-up interview. The Scottish Government respondents were selected by directly emailing the most appropriate people, which was facilitated by the researcher embarking on an internship at the Scottish Government which allowed access to the staff directory and made arranging interviews much easier. One infrastructure interview, which was with a particularly important organisation who had been mentioned in other interviews, was also directly emailed and asked to participate. Sourcing interviewees from the survey, combined with the openness and willingness of the Scottish Government (a cosponsor of this research) meant that access to respondents was not an issue for the research.

The interviews were recorded on two recording devices, with the permission of the respondents, and the researcher made some notes as well as recording reflections post-interview. The audio recordings for each interview were processed by a professional transcription service and all of the data was stored securely, in line with ethical guidelines.

3.6.3 Methods of analysis

With all of the interviews transcribed they were imported into the software NVivo for coding. Codes were drawn from the research questions before the coding process began and then updated and merged during several rounds of coding as similarities and distinctions began to appear (Harrell and Bradley 2009). This is a mixture of inductive and deductive coding. Part of these processes involved discussing the codes and coding with the supervision team which helped refocus and sharpen the coding.

With quotes from the interviews coded, they were then mapped onto each research question and sifted into groups based on similarities in content in Excel; if two respondents mentioned that they trust government data then these quotes were grouped together. This process is where the structure of the chapter began to emerge, and, by rearranging different groups of quotes, a narrative began to form within each research question. This narrative developed as the chapter was written, edited and restructured upon reflection and feedback.

3.7 CONCLUSION

Reflecting on the methods used throughout this thesis, each of the chapters has a distinct set of analytical questions, data source, and consequently methods. These disparate methods are brought together by the chosen research philosophy, which, with its unique epistemology and positivist ontology, can suitably unite and underpin both qualitative and quantitative analysis. This was essential with the first chapter of analysis being wholly quantitative, the third being qualitative and the second being a mix of the two, though leaning more towards the quantitative.

The first chapter mostly sets the context for the rest of the analysis by firstly determining the level of external data use among charities, and then discussing what organisational features, barriers, and enablers may affect levels of data use. This part of the analysis ties heavily into the existing literature and uses mostly conventional statistical methods, generally moving from descriptives to bivariate analysis in preparation for multivariate modelling. A few more advanced techniques, such as factor analysis, are used where appropriate.

The second chapter uses the most unconventional data source and methods but both clearly match up well with the aims of the chapter, to study support relationships. Though this could have been achieved in a survey or interview format, direct observation and analysis of support interactions was deemed to be the best way to gain insights into these relationships. Much of the network analysis revolves around describing the network of charities, support organisations, and data organisations which was collected, and determining what the links between them actually represent. Once the dynamic of support for external data use on Twitter has been identified, this subnetwork can be isolated and analysed with network modelling to determine how well this style of support is received by charities.

The final chapter is the simplest methodologically, primarily because answering the analytical questions around trust in data did not require complexity; face-to-face interviews were the best way to access data and insights on trust in data. The interviews also played a corroborative role, covering topics which the other methods had not fully elucidated; such as issues around staffing or how trust played into relationships on Twitter.

CHAPTER 4: DATA USE

4.1 INTRODUCTION

Data use is the first of the three components which are investigated in this thesis. In comparison to the other two components, support and trust, use is relatively easily measured by quantitative surveys which, therefore, form the data source for this initial analysis chapter. The chapter breaks roughly into two with the first part attempting to describe level of use by the third sector and the second attempting to determine what factors cause variations in use among third sector organisations. The analysis of external data use in the third sector acts as a springboard for further analysis by determining what level of data use there is in the sector, and therefore contextualising the support relationships, which are studied in Chapter 5, and highlighting particular factors or types of organisation which need particular support. The analysis also primes the discussions around trust in Chapter 6 by determining factors which may hold charities back from using data, and thereby affect their ability to engage with and trust data.

4.1.1 Course of analysis

This chapter hopes to answer, or partially answer, the following research and analytical questions:

1. What level of external data usage is there in the Scottish third sector?

1.a. What level of external data usage is there among charities in Scotland?

2. What barriers, enablers, and organisational features affect the ability of third sector organisations to make use of external data?

2.a. Which organisational features best predict differences in levels of use?

2.b. What other factors enable or inhibit use of external data in the Scottish third sector?

Analytical question 1.a. will attempt to describe the level of data use among organisations in the secondary survey and where charities are sourcing their data from. Analytical question 2.a. investigates which organisational features best predict variations in charity data use. This section includes a bivariate analysis and a multivariate modelling section. Analytical question 2.b. follows and, in a similar format to question 2.a., looks at what other, non-organisational, factors may inhibit or enhance charity data use. This comprises a set of barriers and enablers to data use which were put to respondents in the secondary survey and were reviewed in Chapter 2. This section also includes a factor analysis. The chapter ends with an overall conclusion which highlights how the findings from this analysis lead into the forthcoming chapters.

4.2 ANALYTICAL QUESTION 1.A.

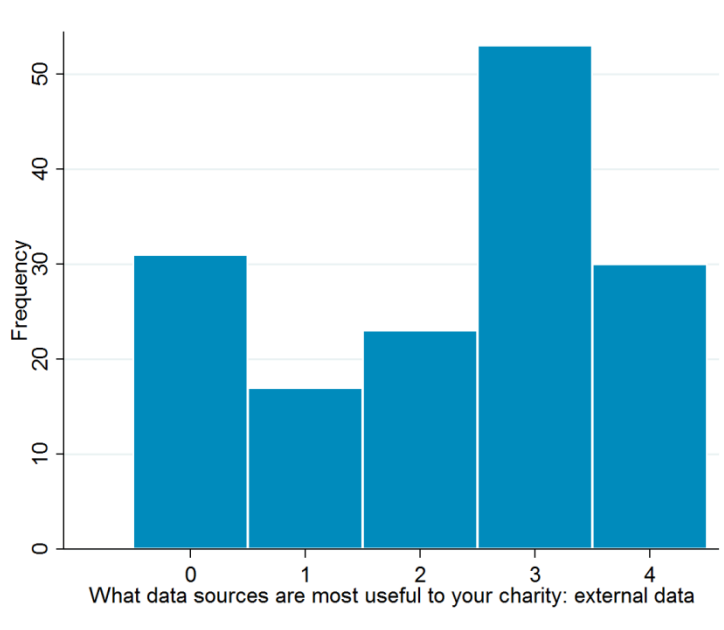
“1.a. What level of external data usage is there among charities in Scotland?”

Answering analytical question 1.a. involves describing results from the primary and secondary surveys. This process is not analytical, but it is an important outcome of the surveys and it sets the context for the project as a whole; are charities even using external data? The variable, which tracks the level of use, is also the dependent variable for the analysis of research question 2 and is discussed in more detail in the methodology. The primary survey is a detailed survey concerning secondary data use by the third sector which 20 charities gave full responses to. Due to its low response rate it is only used in this analysis for data sourcing information. The secondary survey, which is drawn upon for the rest of the analysis in this chapter, was collected by Rutherford and Burt in 2014 and has 154 valid responses.

4.2.1 Level of use

In the secondary survey, the level of use question was worded ‘*What data sources are most useful to your charity: external data?*’ and coded into four categories plus a non-applicable option. The categories were worded as follows; ‘*1. External data sets not useful, 2. External data sets extremely limited usefulness, 3. External data sets moderately useful, and 4. External data sets extremely useful*’. This data is visualised as a distribution in Figure 4-1.

Figure 4-1 Distribution of external data use variable



n=154. Source: Secondary survey

This distribution is negatively skewed, especially if category 0 (which records not responding to the question; interpreted as not using external data at all) is removed. This suggests that those who

do use external data make good use of it as they mostly select categories 3 and 4. However, these responses seem to be at odds with evidence from the interviews in Chapter 6 where respondents discussed lacking skilled staff and working with tight budgets. This may suggest some form of selection bias or sampling bias affecting the secondary survey. The survey was only sent to larger organisations, but it was not specifically about data use so self-selection on data related issues should be limited. There may, therefore, be bias around size, or income, which will be tested later in this chapter in question 2.a. It is also possible that there is some form of social desirability bias in the secondary survey, with respondents claiming they use more data than they do, either consciously or unconsciously (Fisher 1993). Another way to test for level of data use is to study the dataset sourcing responses from the primary survey as shown in Table 4-1.

Table 4-1 Data sourcing responses from primary survey

Data sources	Number of times selected
Local Authority	13
Scottish Government	11
OSCR	8
Social Enterprise Scotland	8
Volunteer Scotland	8
SCVO	7
UK Data Archive	4
NCVO	3
TSRC	2
Home Office	1
National Records Scotland	1

n=20. Source: Primary survey⁸. N.B. respondents could only select each source once.

These responses from the primary survey are from a different sample of organisations, which also had a negatively skewed distribution for a simple data use question, though with an even more limited response rate. In the data sourcing question, the charities were asked to select where they sourced large-scale secondary survey data from⁹. This question should be less subject to bias as the respondents were not asked to put themselves on a subjective scale but rather give a statement of fact (Fisher 1993). This is an indirect way to obtain information on data use.

⁸ https://github.com/tomwallace1990/charity_data_PhD/tree/master/Primary%20survey%20questionnaire

⁹ This question was answered in the context of sourcing the following data sets: Annual Population Survey, OSCR Charity Register, Scottish Household Survey, British Household Panel Survey/ Understanding Society, Labour Force Survey, Scottish Health Survey, Scottish Social Attitudes Survey, UK Census.

In the results, the public sector is heavily represented at the top of the table (Local Authorities and the Scottish Government), with the primary infrastructure organisations coming next (OSCR, SES, Volunteer Scotland, and the SCVO), and a miscellaneous group at the bottom including research (TSRC) and data organisations (UK Data Archive, National Records Scotland).

Though the groupings above are somewhat debatable with organisations, such as OSCR and NRS, being both public sector and data related organisations, what is clear is that many of the organisations at the top of the table are not direct publishers of the data sets which were specified. Most of the datasets listed are controlled by the Scottish or UK Governments, but in raw format they are usually lodged with the UK Data Service, National Records Scotland or the Office for National Statistics (OSCR's charity register notwithstanding), sources who were listed near the bottom of the table, or not at all in the case of ONS. With respondents tending to favour the government, councils, and third sector infrastructure organisations, this suggests that most respondents are not accessing the full raw data sets for analysis. Rather they may be involved in more complicated patterns of interaction with data providers and infrastructure organisations to gain access to data in more easily digested formats. This may include aggregate findings from government reports or output produced by support organisations. These support relationships and data sourcing dynamics are studied in much greater detail in Chapter 5.

For the discussion around data use, the selections made by the respondents suggest a lower level of data usage than responses to the secondary survey would suggest. Data from the primary survey suggests that most charitable organisations are making use of data while perhaps not really analysing it. These findings present an opportunity for infrastructure organisations to provide more easily digested data, which they may be doing already given how often they are listed as a source. The sort of support is studied in Chapter 5. The next set of questions in this chapter will look more closely at what features and other factors may lead to the stratification seen in the data use question; if some charities are able to analyse data, what makes them different from those who cannot?

4.3 ANALYTICAL QUESTION 2.A.

“2.a. Which organisational features best predict differences in levels of use?”

This question attempts to determine how a charity's features are related to their level of data use. These features are split into three subcategories: organisational characteristics, beneficiary groups, and financial characteristics. These three groupings have been chosen for ease of analysis, presentation, and discussion and not because of any implied theoretical division; they are all equally impingent on the research question in a theoretical sense and will be analysed together in

the final combined modelling section. The coding of these variables varies; ‘beneficiary groups’ are binary indicators, and ‘financial characteristics’ are all continuous, but ‘organisational characteristics’ have more variation. These differences in coding have an effect on how each set of variables is analysed.

The dependent variable for the analysis, charity data use, will be assessed in terms of its variation. The analysis of question 1.a. suggested that this variable may be biased, with organisations over-estimating their ability to use data, and so the forthcoming analysis will assess relative variations in this variable; factors which enhance or inhibit data use should have measurable effects on this variable regardless of its skew. The data use variable was transformed into three different configurations to represent different facets of data use. Firstly a normal, unaltered, dependent provided general insights into each relationship. Secondly category 0 of the dependent variable (interpreted as ‘not using external data at all’) was removed, this helped determine how a given variable affected the scope or variability of data use for those who are using it. Thirdly, and inverse to the scope dependent, a binary version of the dependent was generated where category 0 remains and categories 1-4 are all coded into 1. This represents using or not using data without regard for the level of use. These three dependent formats should allow the analysis to, not only measure what factors affect charity data use, but also determine something about the dynamics of how they affect use; does a factor predict if data is used at all or determine the scope of use?

The analysis of question 2.a. is broken into two parts. First, bivariate cross-tabulations will be used to find any individual factors related to data use. Secondly, a multivariate regression analysis will attempt to summarise the overall effects of the organisational features on data use and re-examine the individual results in a modelling framework. Revealing how organisational factors impact on data use will help add meaning to the discussion around charity data use, as well as suggest ways in which use could be enhanced.

4.3.1 Bivariate analysis

4.3.1.1 Organisational characteristics

‘Organisational characteristics’ is a catch-all term for questions asked in the secondary survey which may act as a predictor of data use, but that do not fit neatly into benefactor group or financial indicators which are analysed below. Each categorical variable was cross-tabulated against the three dependent configurations and tested with the gamma statistic. Pearson’s chi-square was used to determine significance; P-values of 0.1 or less were considered marginally insignificant but noteworthy, while less than 0.05 were considered significant. Continuous variables were assessed with appropriate bivariate regression models. The results of these bivariate tests are shown in Table 4-2 and discussed below.

Table 4-2 Organisational characteristics (ordered alphabetically)

Characteristic	Test	General	Scope	Binary
Age	Logistic Regression	-0.01 (0.08)	-0.00 (0.59)	-0.01 (0.07)
Archetype	Gamma/Chi ²	0.11	-0.02	0.35
Constitutional form	Gamma/Chi ²	-0.02	-0.11	0.10
Geographical spread	Gamma/Chi ²	0.06	-0.01	0.12**
Has a data management strategy	Gamma/Chi ²	-0.23	-0.26*	-0.19
Length of time respondent has had a data management strategy	Gamma/Chi ²	0.14	0.17	0.09
Length of time the respondent has had a website	Gamma/Chi ²	0.08**	0.10***	0.05
Number of trustees	Logistic Regression	-0.01 (0.84)	0.01 (0.72)	-0.02 (0.58)
Plan to personalise web services	Gamma/Chi ²	-0.15	-0.16	-0.13
Primary area of activity	Gamma/Chi ²	0.21	0.15	0.29
User profile	Gamma/Chi ²	-0.02	0.02	-0.06

Significance: * < 0.1 ** < 0.05 *** < 0.01 based on Chi². Source: Secondary survey

'Gamma/Chi²' rows show gamma result to two decimal places. 'Regression' rows show coefficient of characteristic with model significance shown in parenthesis.

The following variables were insignificant in all configurations: 'length of time respondent has had a data management strategy', 'age', 'primary area of activity', 'constitutional form', 'user profile', 'number of trustees', 'plan to personalise web services', and 'archetype'. Most of these results are not surprising as there was no theoretical indication that they would be correlated with data use. The final result is surprising, however, 'archetype' is a four option question asking respondents to indicate which data archetype best describes their organisation. Responses range from 'struggling' to 'pushing the boundaries'. It seems natural to assume that organisations placing themselves higher on this scale would also be making more use of external data, but the bivariate results did not bear this out. This may be because the dependent variable refers specifically to external data, as is the focus of this project, while the archetype question referred to data in general, but it could also simply be a result of low case numbers.

Turning to results, which were statistically significant enough to reasonably interpret, 'length of time the organisation has had a website' was significantly associated with the general dependent and even more so with the scope dependent, returning a probability value of 0.003. Despite this, the relationship was weak with a gamma of around 0.1. The interpretation of this result is that the longer an organisation has had a website the more it is using data and, with the result being

strongest without category 0, the variation appears to be within data use rather than determining use or not. This may suggest a link between IT proficiency or digital infrastructure and the scope of data use but, as the question also concerns length of time, the relationship could be masking an age effect; older organisations are likely to have had websites longer. This possibility was examined by looking at the results for ‘age’, despite none of them being significant. The association between ‘age’ and data use was closest to significant with the binary dependent ($Pr=0.065$) and the point estimate was small and negative (-0.01). Considering uncertainty, this suggests no substantial relationship between organisational age and data use, which is counter to the theory posed above. Therefore there seems to be an association between ‘length of time a charity has had a website’, a proxy for IT proficiency or infrastructure, and the scope of data use.

Another significant variable was ‘geographical spread’. This returned a significant result ($Pr=0.015$) for the binary dependent with a small positive gamma (0.12). Given the way the variable was coded, this suggests that more wide-ranging charities are more likely to use data, but the relationship was not particularly strong.

A final interesting result was ‘data management strategy’. This variable, which would be expected to correlate with data use, was marginally insignificant ($Pr=0.084$) when category 0 of the dependent variable was removed and was insignificant in all other configurations. Its gamma, however, was one of the largest observed at -0.26 . The negative sign of this relationship indicates those who have a data management strategy make better use of data and this relationship reflected variation within data use rather than binary use. The marginal insignificance of this expected relationship reflects the difficulties of working with a small dataset; all that can be concluded from this first section of analysis is that the scope of data use seems to be predicted, at least partly, by how long the charity has had a website and whether the charity has a data management strategy. Binary use was predicted by geographic spread.

4.3.1.2 Beneficiary groups

The beneficiary groups are a series of binary variables which attempted to capture a charity’s target service demographic. Analysing these variables will reveal if the focus of a charity has any notable effect on its use of external data. The beneficiary groups and their bivariate results are displayed in Table 4-3.

Table 4-3 Beneficiary groups bivariate association and significance tests

Benefactor	General	Scope	Binary
Children and young people	-0.02	-0.14	0.15
Ethnic/racial	0.35	0.08	0.26
No specific group	0.24**	0.23**	0.25
Older people	0.31*	0.28*	0.35
Other charities	0.18	-0.01	0.56
Other defined group	-0.08	0.03	-0.21
The disabled	-0.05	-0.01	-0.11

Significance: * < 0.1 ** < 0.05 *** < 0.01 based on Chi². Cells show the result of a gamma test.

Source: Secondary survey

The following groups were not significant in any configuration; ‘children and young people’, ‘other defined group’, ‘ethnic/racial’, ‘other charities’, and ‘the disabled’. ‘Other charities’ is the only surprise from these results, as the second chapter of this thesis focuses on the role of infrastructure organisations in providing support for other charities’ data use and this relationship has already been demonstrated to an extent in question 1.a. This null result is likely due to the low number of support organisations in the secondary data set; only 14 out of 159 said they were a benefactor of another charity. In the cross tabulations, this 14 did make more use of data than the remaining 145, but the relationship was not statistically distinct from random variation at a defensible level in this data set and this relationship is more usefully studied in the following chapter.

Two benefactor groups returned significant results. The first was ‘no specific group’, a general category selected by 89 of the 159 organisations. This was significant for both the general dependent (Pr=0.018) and when category 0 was removed to reflect scope (Pr=0.015). The gamma results for these tests were higher than expected at ~0.23 which suggested this general group finds data significantly more useful than the more focused groups. This may be due, in part, to higher case numbers, but it could also reflect a greater need for external data among more diverse charities. This is counter to more focused charities that may make more use of their own data resources and databases rather than needing to make use of external data.

The second significant group is ‘older people’. The results for this group were marginally insignificant (Pr=0.066) with the general dependent with a gamma result of 0.31, suggesting charities intending to benefit older people tend to make more use of external data. This may be due to a medical focus of charities concerned with geriatrics. This result is still insignificant at a 0.05 level however and should not bear particular weight. Overall, the only strong suggestion from this section of analysis is that more widely focused charities may make more use of data.

4.3.1.3 *Financial characteristics*

This final section of organisational characteristics comprises a set of financial indicators derived from OSCAR annual returns data. Being continuous, the financial characteristics were tested using appropriate regression techniques rather than association techniques. The gross and total financial indicators were measured in natural units while the specific income variables, which track where a charity gains their income from, were proportionally normalised against the charities total income as shown along with the results in Table 4-4.

Table 4-4 Financial characteristics bivariate regression results

Variable	Units	General	Scope	Binary
Donations	Normalised	-0.00 (0.73)	-0.00 (0.63)	-0.00 (1.0)
Gross expenditure	Natural units	0.00** (0.04)	0.00** (0.01)	0.00 (0.74)
Gross income	Natural units	0.00* (0.08)	0.00** (0.03)	0.00 (0.75)
Income from charitable activity	Normalised	-0.00 (0.10)	0.00 (0.90)	0.00* (0.04)
Income from interest	Normalised	-0.00 (0.76)	-0.00 (0.55)	0.00 (0.84)
Income from the government	Normalised	-0.00 (0.76)	-0.00 (0.55)	0.00 (0.84)
Net assets	Natural units	-0.00 (0.49)	-0.00 (0.76)	-0.00 (0.53)
Total funds	Natural units	-0.00 (0.74)	-0.00 (0.75)	-0.00 (0.88)
Trading income	Normalised	0.00 (0.22)	0.00 (0.70)	0.00 (0.09)

Significance: * < 0.1 ** < 0.05 *** < 0.01. Cells show coefficient of each bivariate model. Model significance shown in parenthesis. Normalised units are the proportion of a charities gross income which this variable represents.

Source: Secondary survey

As with the previous sections, there were several results which do not warrant interpretation, these include; ‘donations’, ‘income from interest’, ‘income from the government’, ‘net assets’, and ‘total funds’.

A variable which encapsulates many of these insignificant income factors, but which was significant, is ‘gross income’. ‘Gross income’ was most significant against the scope dependent with category 0 removed, in that configuration the ordered logistic regression was significant at a model level (Prob>Chi2=0.03) and a point estimate level (P>|z|=0.03). The coefficient was small relative to the cut-points, but ‘gross income’ in its natural units ranges over nine orders of magnitude and so a small coefficient was expected. The coefficient was positive which indicates higher income charities tend to use more data. This could also be interpreted as larger size correlating with more use of data as income is a common proxy for charity size (Quinton and Fennemore 2013).

Accounting for the other half of charity finances, 'gross expenditure' was also significant. Like income, expenditure was most significant with the dependant excluding category 0 and reflecting the scope of use, where the model had a significance of 0.01. The point estimate significance was 0.02 and the coefficient was positive and slightly larger than that for income. This suggests that data use correlates with higher expenditure as well as higher income, which make sense given expenditure and income are highly correlated with each other. Overall it seems fair to conclude that more money relates to a greater scope of data use, whether the measure reflects income, expenditure, or size.

Besides these aggregate measures, there were two other variables which warrant interpretation. The first is 'income from charitable activity' which was significant with the binary format dependent ($\text{Prob} > \chi^2 = 0.04$), but had a very small negative coefficient and was marginally insignificant in terms of the point estimate ($P > |z| = 0.07$). The second is 'trading income' which was nearly significant with the binary dependent in terms of the model ($\text{Prob} > \chi^2 = 0.09$) but had an insignificant point estimate ($P > |z| = 0.24$). Its coefficient is also small but is positive in contrast with 'charitable activity'. These factors seem to affect binary use rather than the level of use like the gross measures, but with sporadic significance and small coefficients, the interpretation is difficult to justify as meaningful in these bivariate tests. Overall the top level characteristics, 'gross income' and 'gross expenditure', appear to be the best financial indicators of the level of charity data use. There is a limited utility to analysing these groups of variables independently, however, and the final section of this analysis will use a multivariate regression framework to assess the joint impact of organisational features of charity data use.

4.3.2 Combined modelling

The results from the bivariate analysis would suggest that the organisational features are largely idiosyncratic in terms of their relationship with data use; the organisational, benefactor, and financial categories are not cohesive and so the modelling will be carried out with all of the organisational feature variables combined together. Results which were significant in the bivariate tests will be reassessed in this more robust framework and the results from this analysis will give an overall impression of how well charity features can be used to predict differences in charity data use.

The models below include the following variables; 'data strategy', 'length of time had a data strategy', 'structure', 'primary area of activity', 'length of time had a website', 'plan to personalise web services', 'archetype', 'current constitutional form', 'trustee numbers', 'geographic spread', all seven benefactor groups, 'age', 'gross income', and 'gross expenditure'. 'Data strategy', 'structure', and 'user profile' were excluded from the logistic model for causing non-convergence.

The models have varying case numbers; some of the variables have missing data and with 21 independent variables in the analysis (18 for the logistic model), these missing patterns overlap and cause a drop in case numbers. The scope dependent variable lowers cases further as it removes cases which selected category 0. The binary model, which has fewer cases still, has less variation in its dependent variable which, combined with a number of independent categorical variables which have empty cells, leads to perfect prediction and inhibits estimation of the model (Long and Freese 2014). These perfectly predicting cases are automatically dropped by the model and therefore case numbers for the binary model are reduced. It could be argued that, for comparability purposes, the regressions should all be limited to the lowest case number but, with so few cases to draw on, this was found to unduly affect the analysis.

Table 4-5 Aggregate regression results for the combined models

	n	Model P-value	Pseudo R2	BIC	AIC
General ordered logistic	113	0.001	0.24 (0.06)	499.6	367.2
Scope ordered logistic	94	0.002	0.33 (0.06)	384.9	258.8
Binary logistic	80	0.007	0.58 (0.05)	173.6	102.8

Numbers in parenthesis are R-squared standard errors using Olkin and Finn's approximation. Source: Secondary survey

The probability values of the combined models were all highly significant which suggests that, in combination, these factors have some form of impact on charity data use. With such low case numbers, this is encouraging. The R-squared values are also all reasonable, though not particularly impressive given how many variables are present in the models. The higher R-squared result from the binary model is likely due to the lower variation of the binary dependent. The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are very different across the models; clearly the binary logistic model has the lowest value for both criterion, but this model has fewer variables. Of the two ordered logistic regressions, the scope model has the lower BIC and AIC results, which suggests it is the more parsimonious, and superior model. This implies that charity features may be better predictors of scope of use than they are of binary charity use, but, as the binary model is not directly comparable, this is not certain from these overall results.

As the above conclusion is not very insightful without considering individual factors which are detailed in Table 4-6.

Table 4-6 Full model results for the combined models

Variable	General model	Scope model	Binary model
Has a data strategy	-0.29 (0.873)	-3.42 (0.108)	[omitted]
How long had a data strategy: Don't	-0.02 (0.991)	-3.46 (0.123)	4.59 (0.236)
How long had a data strategy: <1 year	0.02 (0.988)	0.77 (0.686)	3.67 (0.393)
How long had a data strategy: 1-3 years	[base]	[base]	[base]
How long had a data strategy: 3-5 years	1.48 (0.138)	1.17 (0.319)	6.17 (0.104)
How long had a data strategy: 5 years +	1.29 (0.155)	0.90 (0.403)	1.80 (0.598)
Structure: Centralised	0.02 (0.986)	0.58 (0.675)	[omitted]
Structure: Decentralised/ federated	[base]	[base]	[omitted]
Structure: Other	-2.12 (0.181)	-1.53 (0.524)	[omitted]
Gross income	0.00 (0.061)	0.00 (0.32)	0.00 (0.271)
Area of activity: Culture/recreation	-0.094 (0.470)	-3.56 (0.055)	-7.73 (0.136)
Area of activity: Education	[base]	[base]	[base]
Area of activity: Health/social service	-0.01 (0.996)	-2.13 (0.194)	-5.53 (0.198)
Area of activity: Environment	1.59 (0.309)	-2.46 (0.270)	[omitted]
Area of activity: Housing	-0.32 (0.853)	-4.43 (0.069)	14.13 (0.995)
Area of activity: Law/advocacy	1.07 (0.667)	-2.00 (0.494)	[omitted]
Area of activity: Religion	-3.60 (0.087)	-9.94 (0.001)	-10.54 (0.215)
Area of activity: Other	1.46 (0.270)	-0.85 (0.626)	0.39 (0.934)
User profile: Children/young people	0.81 (0.408)	-2.41 (0.031)	[omitted]
User profile: Adults	[base]	[base]	[omitted]
User profile: Other	0.43 (0.488)	-0.41 (0.593)	[omitted]
How long had website: Don't have one	-3.42 (0.218)	-22.97 (0.989)	[omitted]
How long had website: up to 5 years	[base]	[base]	[base]
How long had website: 5-10 years	0.16 (0.831)	0.16 (0.864)	1.36 (0.407)
How long had website: 10 years +	0.41 (0.621)	1.05 (0.306)	-0.07 (0.975)
Plan to personalise web services: Yes	0.57 (0.254)	0.86 (0.181)	3.76 (0.033)
Plan to personalise web services: No	[base]	[base]	[base]
Data archetype: Struggling	1.40 (0.081)	1.56 (0.157)	-0.06 (0.967)
Data archetype: Just about managing	[base]	[base]	[base]
Data archetype: Managing	0.09 (0.874)	-1.06 (0.102)	4.86 (0.072)
Data archetype: Pushing the boundaries	1.29 (0.298)	1.71 (0.260)	[omitted]
Constitutional form: Company	-5.26 (0.003)	-5.42 (0.026)	-0.39 (0.944)
Constitutional form: Endowment	[base]	[base]	[base]
Constitutional form: Other	1.37 (0.565)	3.23 (0.260)	[omitted]
Constitutional form: Registered society	-5.16 (0.13)	-4.86 (0.080)	[omitted]

Constitutional form: SCIO	-9.88 (0.000)	-27.77 (0.987)	-19.34 (0.993)
Constitutional form: Corporation	-8.40 (0.000)	-8.51 (0.003)	[omitted]
Constitutional form: Trust	-7.69 (0.001)	-9.24 (0.003)	-8.89 (0.244)
Number of trustees	0.04 (0.513)	0.13 (0.055)	-0.37 (0.054)
Geographic spread: All of Scotland	1.03 (0.418)	0.25 (0.869)	[omitted]
Geographic spread: Local	[base]	[base]	[base]
Geographic spread: More than one LA	1.98 (0.190)	2.42 (0.168)	[omitted]
Geographic spread: Broad area	2.26 (0.001)	1.22 (0.170)	10.14 (0.064)
Geographic spread: Overseas	0.33 (0.766)	-0.89 (0.551)	-1.12 (0.621)
Geographic spread: Scotland and UK	2.78 (0.007)	3.95 (0.002)	0.14 (0.959)
Geographic spread: Wide	0.34 (0.654)	-0.13 (0.887)	0.53 (0.740)
Age	-0.04 (0.005)	-0.04 (0.034)	-0.09 (0.015)
Benefactor: Children	-1.46 (0.026)	-0.59 (0.435)	-3.83 (0.188)
Benefactor: No specific group	0.93 (0.140)	2.02 (0.018)	-0.03 (0.988)
Benefactor: Older people	1.81 (0.025)	1.99 (0.053)	3.70 (0.268)
Benefactor: Other charities	-0.04 (0.964)	-1.01 (0.348)	[omitted]
Benefactor: Other defined groups	0.16 (0.779)	0.79 (0.273)	-0.70 (0.591)
Benefactor: Particular race/religion	1.34 (0.272)	0.97 (0.523)	[omitted]
Benefactor: Disabilities/health	-0.75 (0.208)	-1.08 (0.164)	-3.29 (0.212)

Cell values are coefficients. Significance shown in parenthesis. Significant ($P < 0.05$) and near significant ($P < 0.10$) results are highlighted in bold. Source: Secondary survey

Starting with the binary use model, ‘Age’ is significant and, as in the bivariate analysis, it is negative suggesting younger charities are more likely to be using data than older charities. Age is also significant in both the scope and general models however, which suggests it has a general effect on data use. ‘Plan to personalise web services’ is significant and strongly positive, suggesting charities which do plan to personalise their website are more likely to use data than those who don’t; this may reflect IT skills and infrastructure. One category of the ‘Data archetype’ variable is significant, ‘managing’ against the base ‘just about managing’ the direction of this effect is as expected, with those saying they were managing more likely to use data. There is a near significant result from the binary model for the ‘Geographic spread’ variable with charities spread over a broad area more likely to use data in a binary sense than local charities. This variable is also significant in the general model, though with a greatly reduced effect size.

For the scope model there were more significant or near significant estimates. Two measures of the breadth of a charity’s activity were significant: ‘Geographic spread: Serving Scotland and UK’ in comparison to ‘Serving local’ and ‘Benefactor: No specific group’. Both of these variables were strongly positive and highly significant, suggesting that charities, which serve a wider area and

have a less focused benefactor demographic, make use of more data. The geographic spread result for 'Serving Scotland and UK' is also replicated in the general model, albeit with a slightly smaller effect. This combines with 'Geographic spread: Broad area', which is a measure of spread or breath, which was found to be significant in the binary model. This, as suggested previously, may be because charities which only serve a narrow niche may be able to rely on their own databases and knowledge of their benefactors and area. Charities which are more broadly focused cannot be experts in every area they serve and so utilize more external data resources.

'Number of trustees' was a near significant result from the scope model, but it has a small coefficient. It is also not clear what the interpretation of this result should be given size is controlled for through gross income and it was not significant in the bivariate analysis. 'Older people', another benefactor group, is positive in the scope model and the general model, which is the same result from the bivariate analysis, suggesting charities focused on old people are more likely to use data and use more of it. Contrasting the significance of the results for charities focused on older people, the user profile variable was significant for the category 'Children/young people' in the scope model with a negative effect direction suggesting charities focused on children make less use of data. Similarly, the benefactor group 'Children' was significant and negative in the general model.

Two 'Area of activity' categories were significant for the scope model: 'Culture/recreation' and 'Religion', both negative against the base category 'Education'. The implication of these findings is unclear, but as 'Religion' is also negative and significant in the general model it seems that these charities do make less use of data. Several 'Constitutional form' categories were also significant for the scope and general models, but these results are based on low cell values and are not very insightful.

Two variables, which are not significant in the scope model but were in the bivariate tests, are 'length of time the charity has had a website' and 'charity has a data management strategy'. The website result is a reversal of previous findings, with 'age' being significant in all of the models, and 'length of time had a website' not being significant in any. It may be, therefore, that 'length of time had a website' was masking an 'age' effect. 'Data management strategy' is more difficult to explain, it is a factor which would be expected to correlate with data use in some form but in the multivariate modelling, it is entirely insignificant.

Finally, to turn to the general model itself, only 'gross income' is significant and has not been covered by one of the other models. As expected, 'gross income' is positive in the general model, suggesting that charity size has a general effect on both facets of data use. One of the data archetype categories 'struggling' is near significant in the general model, but as it is positive in

comparison to the base category, ‘just about managing’. This result is in contrast with the result from the binary model, so it seems fair to conclude that data archetype is not a reliable predictor of charity data usage, at least in this sample.

The implication of this section of analysis is that a charity’s ‘user profile’, ‘age’, and ‘breadth of service area’ seem to be the biggest factors determining their scope of external data use. While some of these variables were also significant with the general model, they were insignificant with the binary specification suggesting they, mostly, reflect scope of use. ‘Gross income’ had an effect on overall data use but neither of the facets alone suggesting it has a general effect on data use. For the binary model only, ‘Plan to personalise web services’, a proxy for IT skills and infrastructure, was best predicted. The next section of analysis moves beyond the features of the respondent charities and investigates how important a series of barriers and enablers are in determining data use.

4.4 ANALYTICAL QUESTION 2.B.

“2.b. What other factors enable or inhibit use of external data in the Scottish third sector?”

This section has similar objectives as analytical question 2.a. but looks at factors which may affect data use but are not characteristics of the respondent charity. These factors comprise a series of barriers and enablers to data usage which were presented to respondents in the secondary survey. The intention of this section is to further explore what may inhibit or enhance a charity’s data use so that support relationships can be formed with those who need it the most and the context of charities’ trust in data can be fully understood. All of the barriers and enablers were coded similarly, as shown in Table 4-7. A full list of the barriers and enablers is available in the bivariate analysis of each below; Table 4-8 and Table 4-9 respectively.

Table 4-7 Variable coding for barriers and enablers

Barriers	Enablers
1. ‘Not a barrier’	1. ‘Extremely important’
2. ‘Minor barrier’	2. ‘Important’
3. ‘Considerable barrier’	3. ‘Minor importance’
4. ‘Very considerable barrier’	4. ‘Not important’

N.B. Missing categories removed. Source: Secondary survey

As with the previous analysis, the first section comprises a series of bivariate tabulations with each barrier and enabler against each of the three configurations of dependent. Once the pertinent results

from this analysis have been summarised, a factor analysis will be presented which attempts to determine any collinearity present in the linked explanatory variables and to what extent they are unique from one another. The factor analysis will influence the model building of the following section which comprises a series of multivariate models, which attempt to assess the joint impact of barriers and enablers individually. It should be noted that the barriers and enablers were not presented to the respondents as being specifically about external data but rather charity data use in general.

4.4.1 Bivariate analysis

Given the ordinal nature of the independent variables (and the dependent), gamma was used to determine correlation, while a Pearson's chi-squared test determined statistical significance. P-values of 0.1 or less were considered insignificant but noteworthy, while less than 0.05 were considered significant. Given the coding of the barriers, a negative gamma suggested higher data use correlates with the factor being less of a barrier. In other words, a negative gamma indicates that respondents with high data use did not find the factor a barrier, while those with low data use ranked the factor as a barrier to their data use. This relationship was interpreted as the factor being an important barrier to charity data use. For the enablers, a negative gamma indicated the enabler was ranked as more important by more data active charities and is therefore considered an effective enabler.

4.4.1.1 *Barriers*

There were thirteen barriers to data usage presented in the secondary survey. Table 4-8 summarises each barrier and its results against each version of the dependent variable.

Table 4-8 Summary of barriers bivariate association and significance tests

Barrier	General	Scope	Binary
Agreeing information standards	0.15	0.28	-0.04
Agreeing shared information standards	0.23	0.23	0.23
Cost	0.09	0.16	-0.02
Data privacy concerns	-0.11**	0.01	-0.27***
Data security concerns	-0.15***	-0.07	-0.29***
Ethical issues	0.04	0.26	-0.26**
Integrating IT with partner	0.10	0.10	0.10
Integrating IT within charity	0.09	0.15	0.00
Lack of analytical skills	-0.06	0.07	-0.24
Legal/regulatory constraints	-0.04	-0.04	-0.04
Seeing data informed performance	0.01	0.04	-0.03
Seeing innovation in data	-0.01	0.16	-0.25
Time	-0.20	-0.08	-0.37

Significance: * < 0.1 ** < 0.05 *** < 0.01 based on Chi². Cells show the result of a gamma test. Source: Secondary survey

The following barriers were not found to be significant in any configuration: ‘Cost’, ‘Time’, ‘Integrating IT within charity’, ‘Integrating IT with a partner’, ‘Agreeing information standards’, ‘Agreeing shared information standards’, ‘Lack of analytical skills’, ‘Legal/regulatory constraints’, ‘Seeking innovation in data’, and ‘Seeing data informed performance’. The extent to which these would be significant if case numbers were higher is undeterminable. Three of these insignificant factors were theorised to be important, and interrelated: ‘Cost’, ‘Time’, and ‘Lack of analytical skills’. Discussions in the interviews identified these factors as important barriers to data use in the third sector; why they are not informative in this analysis is difficult to determine. These factors are readdressed in the modelling stage.

Of the thirteen barriers, this leaves three which were informative to some extent. The first is ‘Data security concerns’ which, against the general dependent, returned a gamma of -0.15 and a probability value of 0.034 suggesting respondents who find external data more useful find ‘Data security concerns’ less of a barrier (and those who use it less find ‘Data security concerns’ more of a barrier). This was insignificant when category 0 of the dependent was removed but when the binary dependent was substituted, the gamma increased to -0.29 and the P-value shrunk to 0.009. This indicates that ‘Data security concerns’ may be a factor in which charities use data at all (rather than the variability of use) in that those who do not use data indicate that they find it a significant barrier.

‘Data privacy concerns’ gave similar results; the overall relationship is near significant and negative, which became significant and more negative when a binary dependent was used. The results were so strikingly similar that the variables may not be usefully different; respondents may have conflated these variables when answering the survey as ‘security’ and ‘privacy’ are conceptually similar. This possibility is assessed in the factor analysis in Section 4.4.2.

The final significant barrier was ‘Ethical concerns’. ‘Ethical concerns’ was insignificant and had a small gamma (0.04) when presented in the general tabulation. When the binary dependent was used however the gamma changed to -0.26 and the probability value became significant at 0.044. This brought ‘Ethical concerns’ in line with the ‘privacy’ and ‘security’ findings which may represent further conflation.

The significance of these three variables as barriers to data use may allude to data protection legislation, which was found to be a significant barrier to charity data use in the literature. Interestingly, responses to this survey were given in 2014 before the new General Data Protection Regulations (GDPR) had been publically announced, which suggests that this barrier could be even more important at the time of writing and into the future as charities struggle to adjust to the new rules. A further finding from this set of tests was that all three of the data protection barriers were most strongly associated with the binary dependent variable, which suggests that they have an effect on using data or not using it rather than on the variation of levels of use. This suggests that fears over data protection could be stopping some organisations from using data altogether which would represent a significant barrier to data use. This combination of variables will be tested in the factor analysis and the modelling sections.

4.4.1.2 *Enablers*

In contrast to the thirteen barriers, there are only five enablers, summarised in Table 4-9.

Table 4-9 Summary of enablers bivariate association and significance tests

Enabler	General	Scope	Binary
Better use of resources	-0.20*	-0.24*	-0.15**
Competitive advantage	-0.13	-0.05	-0.23
Data can inform strategy and operation	-0.27	-0.24	-0.30
Leadership	-0.29***	-0.36***	-0.20
Reporting and accountability requirements	-0.27**	-0.37***	-0.14

Significance: * < 0.1 ** < 0.05 *** < 0.01 based on Chi². Cells show the result of a gamma test. Source: Secondary survey

The enablers were formatted similarly to the barriers, but with differently worded categories as shown in Table 4-7. The analysis was therefore undertaken in the same way; tabulating each enabler against the three different configurations of the dependent.

Every gamma result from all five enablers against all three dependents was negative suggesting that, where significant, enablers were more highly valued by respondents making more use of data. This is the expected direction of this relationship. Enabler 1 'Data can inform strategy and operation' and enabler 5 'Competitive advantage' were insignificant in all cases so there is little utility in examining their gamma results. The remaining three enablers were all significant to varying degrees.

'Reporting and accountability requirements' and 'Leadership' returned strikingly similar results; both with strong and highly significant gammas within one-hundredth of a unit of each other (confidence intervals notwithstanding). They also both increased in significance and relationship strength when category 0 of the dependent was removed. This reflects that both 'Reporting and accountability requirements' and 'Leadership' are better predictors of variations in scope within data use than they are of binary use (the difference between the scope and binary dependents). This assertion is given further validity by the insignificant results both of these factors return with the binary dependent, suggesting they do not predict use of data versus not using data, or at least there is no evidence that they do.

'Better use of resources' was different; it was significant or near significant against all three dependent configurations, being most significant with the binary dependent, but it had the largest gamma against the scope dependent. This suggests that as an enabler 'Better use of resources' had an effect on both the use of data and to what extent it is used - but more strongly affected the latter.

With the three significant enablers correlating most with the scope dependent, the results are in contrast with the barriers which affected binary use but not internal variation. This difference will be substantiated in the modelling section after any potential collinearity is diagnosed in the factor analysis.

4.4.2 Factor analysis

4.4.2.1 *Barriers*

From the bivariate analysis, barriers 8 'Data security concerns', 9 'Data privacy concerns', and 13 'Ethical concerns' appeared to be collinear. This was initially apparent from their very similar results in the bivariate tabulations and their similar interpretation; all concerning safe handling of data in some form. They may, therefore, represent the same underlying latent concept. To fully explore this possibility, a factor analysis was conducted on all thirteen of the barriers.

The analysis discovered nine non-negative factors. Of these factors, 1 was dominant, representing 58% of the total variance in all thirteen barriers. A *screeplot* determined that the top three factors should be retained for further analysis. These three factors represented 84% of the variance in the barriers.

Oblique rotation factor loadings revealed that barriers 1 through 7 ('Cost', 'Time', 'Integrating IT within charity', 'Integrating IT with a partner', 'Agreeing information standards', 'Agreeing shared information standards', 'Lack of analytical skills') all loaded on factor 1. This suggested that factor 1 is some form of 'general barrier' factor and that barriers 1-7 are all largely independent of each other, only having the fact they are barriers to data usage in common. Their relatively high uniqueness scores, all above 0.3, were testament to this, suggesting a significant proportion of their variation was not captured by the common factor.

The results for factor 2 were different. Four barriers loaded strongest on this factor with three of them being those suspected of relating to data protection in the bivariate analysis: 'Data security concerns', 'Data privacy concerns', and 'Ethical issues'. The fourth was 'Legal/regulatory constraints' which was not obviously collinear in the tabular analysis but is substantively similar to the other barriers, particularly if the concern around data is partly down to data protection legislation.

Factor 3 was somewhat simpler than the other factors, only two barriers 'Seeing innovation in data' and 'Seeing data informed performance' loaded highly on this factor, but they both loaded very highly, with correlations in excess of 0.9. This relationship was not identified during the tabular analysis as these variables were insignificant and therefore this factor will be considered during the regression analysis. Why these factors are collinear is less clear than for factor 2.

The results for factors 2 and 3 suggest that in the modelling phase the barriers may be best represented as factors rather than the original variables, however a Kaiser-Meyer-Olkin test of sampling adjacency returned a result of 0.46 for the overall factor analysis, which suggested that the barriers are not suitable for representation as factors. The results for factor 2 and 3 in isolation were higher at 0.77 and 0.50 respectively, but these are still below the ideal threshold for inclusion in modelling (Crane et al 1991). Despite this, it is clear that there will be some degree of collinearity present in the regression models based on the results of this factor analysis and therefore models including the factors will be presented alongside the full models to aid in the diagnosis of collinearity.

4.4.2.2 *Enablers*

The only suspected collinearity in the enablers is between 2 ‘Reporting and accountability’ and 3 ‘Leadership’. The conflation of these enablers appears to be in high-level direction and strategy.

A factor analysis was run on all five enablers. This resulted in only two non-negative factors with factor 1 accounting for, effectively, 100% of the observed variation. This meant that the enablers were all collinear to a degree and there was no suggestion that enablers 2 and 3 were any more collinear. Additionally, a Kaiser-Meyer-Olkin test returned a high result of 0.84 which suggests the enablers are suitable for aggregation into a factor. Different configurations of variables and factors will, therefore, be utilized in the modelling.

4.4.3 **Multivariate modelling**

The bivariate analysis of the barriers and enablers suggested that they may represent different facets of charity data use; barriers predicting binary use and enablers reflecting the scope of use. The factor analysis has shown that these groups are largely homogenous, unlike the charity characteristics studied in Section 4.3, and therefore there is a utility to studying each group separately in regression frameworks.

4.4.3.1 *Barriers*

Similar to the tabular analysis, the modelling for each barrier was split into three to reflect the different dependent configurations; ordered logistic regression (general models), ordered logistic regression constrained by an if-statement to remove dependent category 0 (scope models), and logistic regression using the binary version of the dependent variable (binary models). For each of these models, the barriers were treated as categorical variables and split into categories. The reference category was set to 3 ‘Considerable barrier’, primarily as this was a commonly selected category and allowed for good comparability.

The analysis resulted in tables too big to present or review meaningfully, but with all of the independent variables broken into categories, coefficient point estimates were generally not useful in any case. Having discussed the bivariate relationships of each variable, the models are more useful for determining the cumulative effect of the barriers and so the aggregate outputs are the focus of this analysis. Aggregate results for the main models are shown in Table 4-10. Each of the three types of model (general, scope, binary) has four specifications; one with all of the barriers; one with barriers 8, 9, 10, and 13 replaced with factor 2; one with barriers 11 and 12 replaced by factor 3; and a final model including both of the factors.

Table 4-10 Aggregate regression results for barriers

	Model type	Factors	n	Model P-value	Pseudo R2	BIC	AIC
General	Ordered logistic	No factors	124	0.015	0.16 (0.05)	511.9	393.4
General	Ordered logistic	Factor 2	124	0.617	0.07 (0.04)	493.9	406.4
General	Ordered logistic	Factor 3	124	0.036	0.14 (0.05)	501.5	394.4
General	Ordered logistic	Factors 2 & 3	124	0.534	0.06 (0.04)	477	400.9
Scope	Ordered logistic	No factors	101	0.117	0.19 (0.06)	390.4	283
Scope	Ordered logistic	Factor 2	101	0.505	0.11 (0.05)	361.9	283.5
Scope	Ordered logistic	Factor 3	101	†	†	†	†
Scope	Ordered logistic	Factors 2 & 3	101	0.539	0.09 (0.05)	348	348
Binary	Logistic	No factors	100	0.022	0.52 (0.06)	192.8	109.4
Binary	Logistic	Factor 2	124	0.205	0.28 (0.06)	221.2	142.2
Binary	Logistic	Factor 3	100	†	†	†	†
Binary	Logistic	Factors 2 & 3	100	0.568	0.18 (0.06)	213.4	145.7

Numbers in parenthesis are R-squared standard errors using Olkin and Finn's approximation. † indicates model did not converge. Source: Secondary survey

Discussing the general models first, the model with no factors is significant with a P-value of 0.015 and an R-squared value of 0.16. This is an encouraging result as even though the direction of the effect is not calculable for these aggregate measures, as a group, these barriers seem to have some notable level of influence on external data usage. Briefly reviewing the point estimates (which are not presented for brevity), only 'Data security concerns' has P-values under 0.1 for all three of its categories. The coefficients for this variable are as expected in terms of direction, higher categories leading to more negative results; as in the bivariate analysis. The factor 2 model, which included a factor for the four 'data safety' barriers, was insignificant (0.617), as was the model including both of the factors (0.534). Both of these models had lower BIC values than the full model but higher AIC results. The model which replaced barriers 11 ('Seeing innovation in data') and 12 ('Seeing data informed performance') with factor 3 was significant but less so than the full model (0.036 > 0.015). This model had a slightly reduced R-squared point estimate but clearly overlapped with the full model when uncertainty around this estimate was considered. However, this model also had a

lower BIC result than the no factors model and a very similar AIC. The inclusion of factor 2 in the insignificant models had a greater reductive effect on the R-squared value but, given the relatively high standard errors, these results also overlapped with the 0.16 point estimate from the full model. Therefore, there does appear to be some collinearity in the barriers, which is unsurprising given the factor analysis, but the factors do not appear particularly suitable for modelling. Considering the BIC and AIC results, the model not including any factors is likely the best result for the general dependent; provided caution is taken over its R-squared value in the light of the other, insignificant, factor models.

In the next set of models, the scope models, the dependent was constrained by an if-statement to remove category 0 and represent scope in data use by excluding those who don't use external data. This specification was not found to be significantly associated with the barriers in the bivariate analysis, and the same is true of the regression models. The model which resulted from fitting all thirteen barriers had a decent R-squared value of 0.19 but was insignificant (0.117). No other specifications of this model were significant and so there is no evidence that barriers determine the scope of charity data use.

The final set of models were logistic regressions using the binary dependent variable; representing using or not using data. The results were intriguing; the full model was significant (0.022) and had a sizable pseudo R-squared value of 0.52. More coefficient point estimates were significant in this model than any of the previous models, but many categories were still insignificant and the significant categories were spread evenly across the thirteen variables (no variables contained all significant coefficients). The high pseudo R-squared value of this model is partially explainable by the dependent having fewer categories and therefore less variation, but there is also likely to be collinearity as both of the factor models, though insignificant, returned notably lower R-squared point estimates. What is clear is that the binary model not using any factors best reflects the relationship between barriers and external data usage as it is the only significant model and has the lowest BIC and AIC results. This suggests that these barriers tend to prevent data from being used at all and that once they are overcome they do not have a strong effect on the level of use.

4.4.3.2 *Enablers*

The modelling strategy for the enablers followed from the barriers; specifying a regression for each dependent variable using, either, all five enablers or the single factor, which was found to encapsulate all of the enablers. The model results are displayed in Table 4-11.

Table 4-11 Aggregate regression results for enablers

	Model type	Factors	n	Model P-value	Pseudo R2	BIC	AIC
General	Ordered logistic	No factors	147	0.087	0.04 (0.03)	508.6	460.7
	Ordered logistic	Factor 1	147	0.004	0.02 (0.02)	464.3	449.4
Scope	Ordered logistic	No factors	118	0.001	0.11 (0.05)	340.2	298.6
	Ordered logistic	Factor 1	118	0.004	0.03 (0.03)	312.4	301.3
Binary	Logistic	No factors	145	0.473	0.07 (0.04)	194.2	158.5
	Logistic	Factor 1	145	0.170	0.01 (0.02)	154.1	148.1

Numbers in parenthesis are R-squared standard errors using Olkin and Finn's approximation. Source: Secondary survey

It is immediately obvious that both logistic regressions are insignificant (0.473 and 0.170) which suggests that the enablers, in combination, have no effect on whether data is used or not, or at least there is no evidence that they do. The full general regression is only marginally insignificant (0.087), but with a very low R-squared value (0.04) and the factor general regression is significant (0.004), but has an even lower R-squared point estimate which is not discernible from zero (0.02).

The most interesting results are for the scope regressions which are both highly significant. The model without the factor has a reasonable R-squared value (0.11), but in the factor model, this is reduced to a basically negligible point estimate (0.03). This suggests, as the bivariate analysis inferred, that enablers have an effect on data usage levels rather than determining if data is used or not. However, these five enablers do not appear to explain much of the variation in the scope of data use. Despite being less significant, the BIC and AIC results suggest that the scope models are more parsimonious than the full specification version. The primary finding, therefore, is that in general factors which positively enable data use tend to affect the scope of use. Considering the modelling of the barriers, these analyses together suggest a complementary relationship between barriers and enablers; barriers appear to inhibit charities using data altogether but once they are overcome do not seem to affect how much data is used, which is predicted best by enablers which drive the scope of use.

4.5 CONCLUSION

The first part of this chapter explored charity data use descriptively. The core of analytical question 1.a., the level of charity data use, is a key outcome. However, the secondary survey data was not conducive to exploring this robustly in a descriptive sense, with the average organisation ranking

themselves as using more data than the literature and evidence from the interviews, detailed in Chapter 6, would suggest. As respondents appeared to be over-estimating their data use when asked directly, data sourcing information from the primary survey was consulted. This revealed that some of the most popular data sources respondents selected do not generally host raw data sets and it appeared that what respondents may be referring to when they are asked about external data use is aggregate findings, reports, or other pre-processed forms of data. This is more aligned with evidence from the interviews and suggests a generally low average ability to use data in the third sector.

With this finding in mind, the second part of the chapter expanded on the descriptive exploration of data use by examining what factors might predict, affect or determine variations in charity data use. For the organisational features, factors drawn from the literature such as age and income, a measure of size, had the strongest effects on overall charity data use; positively predicting both binary use and the scope of use and suggesting that smaller and older organisations are struggling to use data. Factors which only affected the scope of data use, that is to what extent it is used, tended to reflect the charities spread or focus; narrower organisations in both service demographic and geographic location made less use of external data resources. This is, perhaps, due to wider-ranging organisations not being able to rely as heavily on their own databases or tacit knowledge and, thereby, being driven to use more external data to fill in the gaps. While organisations themselves are unlikely to be able to change these factors to affect more data use, these findings could be useful for infrastructure organisations, data providers, and the Scottish Government to aid in targeting support for data use at those who need it the most. The dynamics of this sort of support are studied in Chapter 5.

The final part of the chapter looked at the influence of barriers and enablers to data use; pertinent factors not related directly to the features of each organisation but discussed in the previous literature. A few individual factors which the literature predicted would have an impact on charity data use were found to be individually significant in the bivariate analysis. The strongest of these was a set of variables which the factor analysis suggested jointly represent fear over data protection regulations, which was one of the most notable barriers mentioned in the literature (W. Hall et al 2012; Lloyds Bank 2016). The bivariate analysis also returned significant results for reporting and accountability and leadership as enablers. The literature argued that the former was one of the most important drivers of charity data use (Burt and Otto 2017; Curvers et al 2016; De Las Casas et al 2013). The later factor, leadership, was not nearly as prominent in the literature (Burt and Otto 2017; Clark 2018). The wider findings from this analytical question concern the analysis of the barriers and enablers in multivariate models. These models found that barriers tend to predict whether data is used or not used in a binary sense while enablers reflect variation in levels of use. This suggests two ways in which charitable organisations may be constrained in their use of data;

barriers such as concerns over data protection and security could inhibit data use altogether while, even if these barriers are overcome, appropriate enablers, such as leadership, need to be in place to drive data use so it becomes integrated into how a charity functions. The previous literature does not address barriers or enablers jointly, nor does it discuss the different dynamics which each group appears to have on charity data use; barriers predicting binary use and enablers predicting the level of use. As the barriers and enablers were not as informative individually as they were when aggregated into models, the most important individual factors which limit data use will be reassessed in greater detail during the interviews in Chapter 6.

Clearly, there are inhibitors to charity data use to be overcome, but many of these factors are either beyond the control of charities themselves, such as age, or rely on access to specific knowledge or skills, such as information on data protection law. This makes these barriers largely structural. With variation in levels of use in the sector and some charities already sourcing data through infrastructure organisations as shown in question 1.a., there is a clear place for intra-organisational support for external data use in the third sector. Support for data use could aid, not only in sourcing or pre-processing data, but in disseminating the knowledge and skills needed to overcome the barriers and embrace enablers. This could be through direct support, advertising training or events, disseminating support materials, or sharing data itself. All of these forms of support will be investigated in the next chapter which will use Twitter data as a case study for support for data use in the third sector.

CHAPTER 5: SUPPORT

5.1 INTRODUCTION

Chapter 4 found that charity data use is limited by a number of factors, many of them beyond the control of the charities themselves. This chapter looks at charity support relationships with other organisations, which could help mitigate these factors and enhance data use in other ways. This analysis focuses on support from third sector infrastructure organisations¹⁰ (also referred to as support organisations) on Twitter. Twitter is a platform which many charities are already engaged with and it has been shown in the previous literature to have the potential to be an information sharing medium (Kwak et al 2010). This chapter investigates what volume and type of support infrastructure organisations supply to charities through Twitter networks. Data organisations¹¹ are also reviewed in the analysis. These organisations are less involved with the third sector but are theorised to be the source of much of the information about data and data use in online networks. The Scottish Government's role in the Twitter network is also reviewed. Charities were found to be sourcing data from infrastructure and data related organisations in the previous chapter, to varying degrees. Despite this, little is known about the relationships between these organisations and charities on social media. Charities do not operate in vacuums and their relationships and interactions with other charities and organisations could be important to the support they receive, and ultimately their ability to use data effectively.

5.1.1 Course of analysis

This chapter concerns research question 4: *'What evidence is there of support for data usage among charities, support organisations, and data organisations on Twitter and how does this support manifest?'* This broad question breaks down into a tetrad of analytical questions which form the four main sections of this chapter:

¹⁰ Support organisations include umbrella bodies, funders, and other infrastructure organisations that support the third sector.

¹¹ Data organisations include data producers, custodians, advocates and other organisations primarily concerned with data or statistics.

4.a. To what extent do charities, infrastructure organisations, and data organisations use Twitter differently and what implications does this have for how they are networked?

4.b. What content is actually exchanged on Twitter between charities, infrastructure organisations, and data organisations, how much of this content is data related, and what form of support do these data related tweets embody?

4.c. What are the dynamics of data related tweets between the groups and what does this reveal about support for data use on Twitter?

4.d. Is there evidence of support for data use disseminating through following links?

Question 4.a. sets up the rest of the analysis by investigating how charities, infrastructure organisations, data organisations, and the Scottish Government use Twitter to different extents. Metrics are then reviewed, which describe how the organisations are networked, culminating in a discussion around a sociogram, which visualises how the organisations come together to form a network on Twitter. This discussion sets up the content analysis in question 4.b. which reveals what content the network is actually built on, how much of this content concerns data, and how this content embodies support for data use. This data related subnetwork is further explored in question 4.c., which uses a sociogram, group interconnection table, and Exponential Random Graph Modelling (ERGM) to determine which accounts are sending and receiving this data related content and the dynamics of support for data use on Twitter. Finally, question 4.d. uses the ERGM framework to investigate if the previously theorised dynamics of support can be substantiated by investigating a network of following links.

5.2 ANALYTICAL QUESTION 4.A.

4.a. To what extent do charities, infrastructure organisations, and data organisations use Twitter differently and what implications does this have for how they are networked?

Question 4.a. begins by investigating how the different groups of organisations (charities, support organisations, data organisations, and the Scottish Government) vary in their use of Twitter. This feeds into an exploration of how the different groups come together to form a network and how the groups differ in their networking metrics. Differences in the use of Twitter and network position between the groups is important context for studying their interactions in later analysis.

5.2.1 Twitter usage by group

The Twitter API supplies several metrics for each account which encapsulate the use of the platform; the number of followers, the number of following, the number of tweets, and when the account was created. These metrics provide context for how the different groups are using Twitter

without considering their role in the network. However, with a relatively small sample size, especially when the data is broken down into the individual groups, it is prudent to check that the groups are statistically distinguishable from each other before comparing them. This task is complicated, somewhat, by the distributions of the Twitter metrics; they are prone to extreme outliers at right side of the distributions which significantly skew their means. Using medians should suppress this issue as medians are commonly-used in analyses of outlier heavy, skewed, data sources such as income data (Johnson 2000). In this case, however, medians present a problem; confidence intervals are not usually calculated for them and, though this could be achieved manually, this is unconventional.

Therefore, a decision was made to employ a regression framework to aid in the differentiation of the groups. A quantile regression was chosen to eliminate the problem of outliers by estimating the conditional median. Furthermore, a decision was made to bootstrap the quantile regressions to fully account for the skewed distribution and quell any other violated assumptions. Finally, quasi-variance was employed to aid in the robust comparison of the groups by calculating a standard error for the base category (Gayle and Lambert 2007). This nonparametric approach to inference should result in the tightest possible confidence intervals for the medians of the Twitter metrics.

Each of the metrics was regressed against the group variable with the metric as the dependent and the group reference category set to 'Charities'. The Scottish Government was excluded from this analysis as it is in a group by itself and it represents the population values. The bootstrapping was found to be stable at 1,000 repetitions. Summary results for these models are shown in Table 5-1.

Table 5-1 Twitter usage quantile regression results with quasi-standard errors

	Following		Followers		Tweets		Days on Twitter	
	Coef.	Q-SE	Coef.	Q-SE	Coef.	Q-SE	Coef.	Q-SE
Charities	[base]	149	[base]	344	[base]	263	[base]	103
Support organisations	541	204	1873	592	3522	1064	178	109
Data organisations	-151	123	-194	648	-1407	745	-629	271

n=207. Multiplier=1.97. Note: Q-SE is the standard error calculated from quasi-variance. Source: Primary Twitter data

Looking at the coefficients, it appears that support organisations are more active, more popular and have been on Twitter for longer than charities. Data organisations appear to have lower scores on all of these metrics than the charities, however when the quasi-standard errors are considered these results are less certain. For the number of accounts each of the sample accounts are following, representing networking activity, only the data organisations and support organisations can be statistically differentiated, uncertainty around the estimate for the charities overlaps with both of

these other groups. For the number of followers, representing networking popularity, the support organisations can be said to be more popular than the charities, but the data organisations overlap with the charities and support organisations. The results are similar for the number of tweets, which represents activity, with the support organisations sending more than both of the other groups and the charities and data organisations confidence intervals overlapping with each other. Finally, for the length of time the accounts have been on Twitter, the support organisations and charities overlap with only the data organisations being significantly younger. Overall, these results are not definitive and are hampered by high standard errors. It appears that support organisations are making more use of Twitter than the other groups, but the data organisations and charities are difficult to separate, apart from the data organisations having been on Twitter for less time. The forthcoming analysis will, therefore, attempt to measure differences between the groups in terms of networking metrics rather than platform-wide metrics.

5.2.2 Centrality by group

On first inspection the network centrality metrics may appear similar to the Twitter usage metrics previously discussed; they are metric, node-level measures of interaction on Twitter. However, the centrality measures relate directly to the network which was collected for the sample of charities and not simply overall Twitter use; they are numerical representations of Sociogram 5-1 presented below. The mentions network, as detailed in the methodology, contains 211 accounts with 2,300 unique connections and 12,756 total connections between them.

Centrality measures are directly interrelated; a change in the centrality of one organisation will have a direct knock-on effect on others connected to them, which will have further knock-on effects. These complicated patterns of interrelation violate the assumptions of frequentist inferential statistics, which makes regressions and several other techniques inappropriate. The literature does, however, include examples of descriptive techniques being applied to centrality metrics (Guo and Saxton 2014; Jung and Valero 2016; Phethean et al 2015; Zhou and Pan 2016). Svensson et.al. (2015: 1099), in particular, compared charities by the mean number of tweets they had sent of a particular sort determined by a content analysis. This section of the analysis will attempt to discern differences in the organisational groups in terms of their centrality to the network, which will begin to reveal the relationships at play between the groups of organisations.

The centrality scores are not all normally distributed, therefore medians were used to compare between the groups. Table 5-2 below details the simple 'degree' scores.

Table 5-2 Median degree centrality scores by group

Group	In-degree	Out-degree
1. Charity	5 (3)	4 (3)
2. Support	13 (8)	16 (9)
3. Data	6 (3)	8 (5.5)
4. Scottish Government	112	8
Overall median	7 (5)	7 (6)

Figures in parenthesis show the median absolute deviation (MAD) by group. The Scottish Government is a group of one and has no deviation. Source: Primary Twitter data

Degree is a measure which simply tracks how many unique inbound or outbound connections each vertex has within the network. In-degree is how many other organisations in the network mention the given node at least once; this represents popularity in the network (Robins et al 2009). Out-degree is how many other organisations a node tweets within the network; this represents activity from the sending node. Degree does not consider edge weight and is, therefore, a function of the 2,300 unique links in the network rather than the 12,756 total links.

Focussing on in-degree first, charities and data organisations appear to have very similar median scores suggesting they are comparably popular for others in the network to mention. Support organisations are roughly twice as popular, though this figure has a higher median absolute deviation. The Scottish Government is a population figure and so comparisons with the medians of the other groups should be cautious, but it is clear the government has a substantially higher in-degree score. One note of caution should be heeded; degree scores do not take into account which groups the connections are coming from or going to, it is possible for a group to have high degree scores by talking internally amongst themselves. This, therefore, means that a group having a high popularity (in-degree) does not mean as a group it is popular with other groups; it means that the handles within that group tend to be more popular with other handles.

For out-degree, the story looks similar for the charities and data organisations; though with data organisations slightly more active than charities. Support organisations, once again have a notably higher score. The Scottish Government's out-degree, it is perhaps, surprisingly low. These being population figures, it can be said with certainty that while the Scottish Government received in links from 112 other organisations in the network, it only sent out to 8 organisations. It seems that,

for this network in particular, the Scottish Government is highly popular, but not very active. It should be stressed that this is only counting contact within the network and not the overall activity of the handles; where the Scottish Government would be far more active. Table 5-3 details a set of more complex centrality metrics.

Table 5-3 Median centrality scores by group

Group	Closeness centrality	Betweenness centrality	Eigenvector centrality
1. Charity	0.0020 (0.000187)	23 (24)	0.0022 (0.0017)
2. Support	0.0023 (0.000148)	202 (191)	0.0076 (0.0039)
3. Data	0.0019 (0.000193)	50 (50)	0.0014 (0.0082)
4. Scottish Government	0.0032	9500	0.0256
Overall median	0.0021 (0.000206)	53 (53)	0.0031 (0.0023)

Figures in parenthesis show the median absolute deviation (MAD) by group. The Scottish Government is a group of one and has no deviation. Source: Primary Twitter data

The centrality scores in Table 5-3 are less intuitive than the degree scores. They are also in varying units, but their absolute values are not as important as comparison between the groups.

Closeness centrality is a measure of centralisation; it is the inverse of the sum of all shortest paths between a given node and all other nodes. In other words, nodes with short paths to many other nodes will be more central to the network than nodes which have long paths to most other nodes. This metric can indicate who is in the core of a network and who is peripheral. The highest closeness of the groups is the Scottish Government, which is unsurprising given its high popularity and therefore many short paths as many other nodes have contacted it directly. The next highest closeness of the groups is the support organisations who are notably more central than the charities or data organisations. The data organisations have the lowest closeness which, counter to their degree scores, suggest they are not as central to the network as the charities or other groups.

Betweenness centrality summarises how information flows around a network. It is defined as the number of shortest paths passing through each node. In this metric, the charities have a notably low score, which suggests that while they are relatively central and have reasonable degree scores, they do not inhabit the key linking positions in the network between the other groups. Data

organisations have a higher score, mostly due to a few high scoring organisations which link the data handles to the rest of the network (see the Sociogram 5-1 below), but it is support organisations who excel in this metric. With a score some ten times that of the charities, support organisations are key to linking the network together and perhaps this is what would be expected if the support organisations were doing their job; they should be performing a linking role between the more numerous charities and between the charities and the data organisations. The government has an extremely high betweenness score but, again, this is unsurprising considering how central and popular it is.

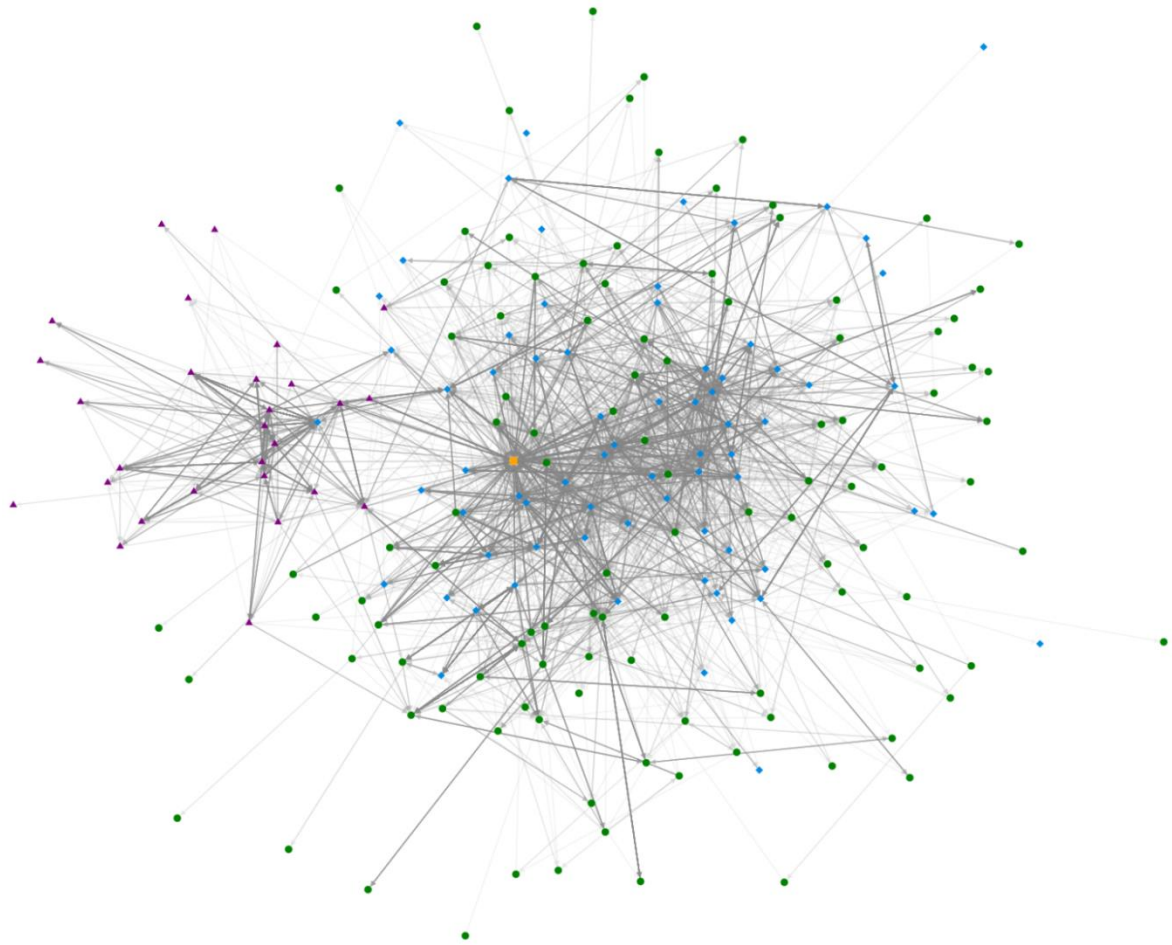
Eigenvector centrality is the last and least intuitive of the measures; it is similar to degree centrality but instead of counting links to all other accounts equally it weights the links so that connecting to prominent accounts results in a higher score than connecting to less prominent accounts. Burris (2004) links this idea of weighting to Bourdieu's conceptualisation of social capital where relationships are not of equal value. As the most prominent single account in the network, the Scottish Government, has by far the highest eigenvector centrality. Support organisations have the highest of any of the multimember groups, likely due to connections to the government and internally to each other, and charities have a higher eigenvector result than data organisations which, again, suggests that data organisations are somewhat separate from the network and not connected to its key players.

To summarise briefly on all of the centrality scores, charities appear to be the least active and least popular group to connect to, but their reasonable closeness and eigenvector centralities suggest they are in the main body of the network. Support organisations tend to dominate the network with high scores across all six metrics only being bested by the Scottish Government. This may suggest they are doing their job by holding the network together and facilitating links between accounts. The data organisations have the most unusual scores from this analysis, they have activity and popularity levels similar to charities and even lower closeness and eigenvector scores, but their middling betweenness suggests that they may be in a cluster separate from the main network. This is most easily assessed in Sociogram 5-1.

5.2.3 Sociogram

Sociograms are visual representations of networks where nodes are represented by points and edges as lines between them terminating in an arrow to indicate the direction of the tie. These graphs are laid out by an algorithm which determines where each node should be placed in the graph. This thesis exclusively uses the Harel Koren fast multi-scale algorithm, because of its ability to clearly show clustering and network structures (Harel and Koren 2000).

Sociogram 5-1 Mentions network sociogram with nodes split by group



Node groups: Charities (green disks), support organisations (blue diamonds), data organisations (purple triangles), The Scottish Government (yellow square). Boldness of edges represents intensity of repeated contact. Source: Primary Twitter data

In Sociogram 5-1, each handle is represented by a shape and colour as shown in the key. The links between these accounts are mentions, and all 12,756 mentions are drawn as a line between the two accounts involved with an arrow pointing towards the receiver; this line is an edge. Repeated contact results in layers of lines being drawn and by lowering the opacity of the edges the graph is able to highlight where the contact between accounts was fleeting and where it was sustained and intense.

Looking at the overall patterns in the graph, the most striking feature is the cluster of purple data organisations. All but one of the data organisations are linked to this cluster and most of their contact appears to be internal to the cluster. There are a handful of bridging organisations which sit between the data organisations and the rest of the network and they consequently have high individual betweenness centrality scores. These linking organisations tend to blur the lines between support organisations and data organisations. There appears to be very limited direct contact between data organisations and the overwhelming majority of charities.

The next feature of note is the Scottish Government which, as the centrality scores would suggest, is in the centre of the network and is highly connected; though as previously shown it does not tend to link out to others and so most of its links are inbound. Surrounding the government is a core of charities and support organisations. This core looks evenly split between these two groups with the outer cloud being mostly comprised of charities and accounting for their observed lower median centrality scores. This suggests, beyond the centrality analysis presented previously, that many of the charities are just as well connected and central as the support organisations but a large number of less well-connected charities brings down their median. It appears that while the support organisations tend to be more homogenous in their centrality, charities exhibit a greater range of connectedness.

One unmeasured factor which plays a role in how charitable organisations use Twitter and network with each other is size. An under review paper by Wallace & Rutherford details research using a random sample of UK charities. This paper shows that larger charities tend to be markedly more active and more popular on social media than their smaller counterparts. For the sample used in this thesis, this effect explains at least part of the dominant position of infrastructure organisations in the networks at discussed above; these organisations tend to be larger than the average charity.

What is not immediately obvious from Sociogram 5-1 is that each of the 12,756 edges is more than just a tie between two accounts; each is also a tweet, a piece of content, and what is actually being said is key to contextualising the network. For the purpose of this thesis how much of this content is related to support for data use is particularly pertinent and is considered in the following analysis.

5.3 ANALYTICAL QUESTION 4.B.

4.b. What content is actually exchanged on Twitter between charities, infrastructure organisations, and data organisations, how much of this content is data related, and what form of support do these data related tweets embody?

This section begins with a summary overview of the content tweeted by the different groups and how they differ in their use of certain data related words. The analysis then moves to tweet-level content analysis, which attempts to understand the usage and meaning of individual words in context. The first part of the tweet-level analysis investigates general commonly-used words to gain an understanding of how the organisations tend to interact on Twitter. The second part of the tweet-level analysis looks specifically at data related words and what the content of these data-related tweets implies about support for data use by charities. The content of relationships between and

within the different groups helps contextualise the rest of the chapter and the thesis as a whole; to what extent is there support for charity data use on Twitter?

5.3.1 Summary content

The mentions network contains 12,756 tweets and 203,066 words total. It is important to note that this network does not contain any tweets made by the sample which do not mention another organisation in the network. Therefore the findings presented here are for a subset of charity Twitter use only. Table 5-4 below details differing activity levels among the groups in terms of the number of tweets they have sent within the network.

Table 5-4 Group Twitter activity statistics summary

Group	Number of tweets	Percentage of tweets	Ave. tweets per account
1. Charities	2,972	23.3%	26.5
2. Support	7,528	59%	109.1
3. Data	2,247	17.6%	86.4
4. Scottish Government	9	0.1%	9
Total	12,756	100%	60.5

Source: Primary Twitter data

It is clear that support organisations are the dominant force in terms of numbers of tweets. Even considering the average number of tweets per accounts and the smaller number of data organisations in the sample, as shown in the final column, support organisations are the most active and numerous accounts in the network. In the group results below, which details the percentage of each group's tweets that contain specific features, this activity imbalance has been controlled for by using percentages of the group's tweets. It should, therefore, be noted that 50% of the support group's tweets is a much higher absolute figure than 50% of any other group's tweets.

Table 5-5 Group Twitter content statistics summary

Feature	Charities	Support	Data organisations	Scottish Government	Whole network
Retweets	70.4%	68.8%	82.1%	66.7%	71.5%
URLs	66.9%	64%	84.3%	66.7%	68.2%
Hashtags	47.2%	51.2%	54.4%	55.6%	50.8%
Data related words					
‘Data’	1%	1%	76.2%	0%	14.3%
‘%’	1.1%	0.9%	2.1%	0%	1.2%
‘Research’	0.8%	1.2%	12.6%	0%	3.1%
‘Stats/statistics’	4.2%	2.4%	34.7%	11.1%	8.4%
‘Findings’	0.2%	0.4%	0.7%	0%	0.4%
‘Evidence’	0.5%	0.7%	0.8%	0%	0.7%
‘Report’	1.1%	2.3%	2.4%	0%	2.1%

Source: Primary Twitter data

The first finding of note is that the network is primarily comprised of retweets. Retweets appear in the network as links sent from the retweeting account to the original author of the content. All of the groups retweet heavily, with the data organisations retweeting the most. The stand out in terms of sharing URLs is also the data group, which makes sense given the focus of these organisations. Perhaps more impressive is the number of URLs shared by the support organisations and charities; there is a lot of information flowing around the network which is a finding which corroborates Lovejoy and Saxton (2012), who were also surprised by the depth of information shared by URLs in charity Twitter networks.

Turning to summary statistics for the individual words, these seven words were selected inductively from a content analysis of the individual tweets to represent data related content and form a thematic framework for analysis. Unsurprisingly, the data organisations account for most of this content in the network in terms of volume, but the charities and support organisations are also using some of these words, particularly ‘data’, ‘%’, and ‘statistics’ in small but notable quantities. In proportional terms, charities appear to be using some of these words more than support organisations (‘statistics’, ‘%’) while for other words the opposite is true (‘research’, ‘report’). The Scottish Government only makes one data related tweet in the network and so from this point on it will be classified as a support organisation and not reviewed separately. Although the Government is the publisher of much of the data, which is discussed in the network, it appears to act vicariously through infrastructure and data organisations, including its own data focused accounts such as ScotStat. This vicarious relationship is explored in more depth in the interviews in Chapter 6. To

explore the usage of the data related words and their implications in more depth, this analysis will proceed to the tweet-level results.

5.3.2 Tweet-level content analysis

The raw tweet-level results contain many uninteresting common words ('to', 'and', 'is') and technical artefacts ('https', 'RT', '...'). Therefore, meaningful words were identified by the same inductive process as above, and are discussed below, along with examples of their use. The words are split into general content and data related content; forming separate thematic frameworks. The examples presented in parenthesis are paraphrased to protect the anonymity of the participants.

5.3.2.1 General content

General content comprises words which are not related to data or research, but which are still important for exploring how charities, support organisations and data organisations interact on Twitter. *Great* (638 uses) is the most commonly-used of these words being the 25th most used word overall. *Great* is a pure networking word, perhaps more than would be expected. It is used to signify face to face links between organisations ('great 2 meet', 'great presentation') or to advocate for another organisation ('great to see @X championing', 'follow these great organisations'). *Event* (395 uses), is used very similarly to *great*, tending to signal face to face interactions between organisations ('we are co-hosting this event with @X'), or to publicise an event ('welcomes all to #X event'). Usage of these words implies that charities are actively linking on Twitter and that Twitter is acting as a proxy for offline links to some extent; charities are not just tweeting about their own activities.

Support (407 uses) is an interesting word to be ranked so highly, 52nd overall. This chapter focuses on support of charities by infrastructure organisations and this is somewhat reflected in the use of the word *support* ('with the support of', 'coming to support collaboration'). However, the word is also used more widely to signify opportunity ('fund to support involvement') or simply charitable activity ('support some of Scotland's most at-risk'). *Help* (371 uses), is used very much like *support*, it sometimes signifies inter-organisational links ('with help from @X'), but is also commonly used to advertise wider charitable activity ('help fight poverty'). More public facing tweets, like 'help fight poverty', are a sign of support in the network as they have usually been retweeted by either an infrastructure organisation or a peer charity to widen the tweet's reach, which is clearly a form of support, if not necessarily support for data usage and therefore not the focus of this analysis. The general content analysis suggests that the charity network on Twitter is more connected and supportive than may be assumed; charities discuss off-line networking, share each other's content, and do not simply broadcast information on their own activities and campaigns. This suggests a fertile environment for support for data use which is discussed next.

5.3.2.2 *Data related content*

The following words are more directly related to statistics or research and elucidate the dynamics of support for data use on Twitter. *Data* (552 uses) is the most pertinent of these words. *Data* tends to be used when discussing new findings or resources being published and is almost always accompanied by a URL to the resource ('data linking project emerging findings @X [https...](https://)'). When it is used otherwise it tends to be to discuss data directly in terms of networking ('importance of data investments @X @X') or directly in terms of support ('if you work with data you might be interested in our Basic Statistics course'). The vast majority of the uses of the word *data* could be described as supporting data use, whether by advertising direct support or, more commonly, publicising data resources. *Data* is often used interchangeably with, or in a very similar fashion to, several other related terms as described below.

Research (243 uses) tends to be used in very similar tweets to *data*; it usually signals new findings and is often followed by a link and sometimes a hashtag to signify the subject of the findings ('new #X research'). *Research* can also be observed in tweets which more directly link to the organisations which have carried out the research ('new #X research by @X'), which suggests that the target of these data-related tweets, who they are tweeted at, may not always be the intended audience of the actual content. *Stats/statistics* (188 uses) tweets tend to be similar to *research*; sharing small snippets of findings and linking to a resource. In this case, these words appear to be used more by data organisations where *research* was used more by support organisations. *Findings*, *evidence*, *report*, and the % symbol all tend to be similar to *research* and are mostly used by support organisations to share information and link to data resources, or in the case of the % symbol to share snippets of statistics.

Mean (63 uses) is not a common word in the network, and it is not obviously data related¹², but it does hint at the sender adding interpretation to statistics, data sources or other information ('what does it mean for research?', 'what does this mean for Scotland?'). In particular it appears that when support organisations share data resources they will include a snippet of the data which is relevant to the third sector. This role of adding meaning to data could be crucial in showing charities how larger data sets may be relevant to them and will be discussed further in the interviews in Chapter 6.

Finally, there was some evidence in the Twitter network of support directed to overcoming the barriers discussed in Chapter 4; primarily issues around data protection which was the most pertinent barrier and appeared to have the potential to inhibit data use altogether. The phrase *data protection* (14 uses) appeared in only a few tweets, but they universally shared links to guides on how to manage data protection rules ('...get data protection right. Use our checklist [https...](https://)').

¹² In this instance 'mean' is synonymous with 'signify' or 'convey' and not statistical average.

Several of these tweets also stressed how important data protection was or showed sympathy for how big a barrier it can be for charities ('Data protection: too worrying, too complex, too terrifying? [https...](#)'). Others advertised that new regulations were coming into force and charities should be aware of the change ('NEWS – new data protection law – how secure are you? [https...](#)').

Clearly there is supportive content being shared in the network to help charities overcome barriers to data use, be it the ability to access data or issues around data protection. However, the use of data-related words in the network tend to be supportive in an indirect sense; the content is very rarely directed at the account being mentioned and seems intended to be more broadly accessible. In Burton and Soboleva's (2011) typology, the messages are one-to-many rather than one-to-one. This suggests an indirect form of support where the intended recipients may be followers of the sending account and not the mentioned account itself. This fits with the actual content of the data related tweets which tends to link to resources, training, or events, which could be useful to a range of charities; they are not usually specifically targeted. With the potential for content to spread via Twitter's following system, it is important to understand which organisations tend to send these sorts of tweets, and, if they are not the intended recipients of the support, which organisations receive them. This will elucidate the supply side of support and is explored in analytical question 4.c.

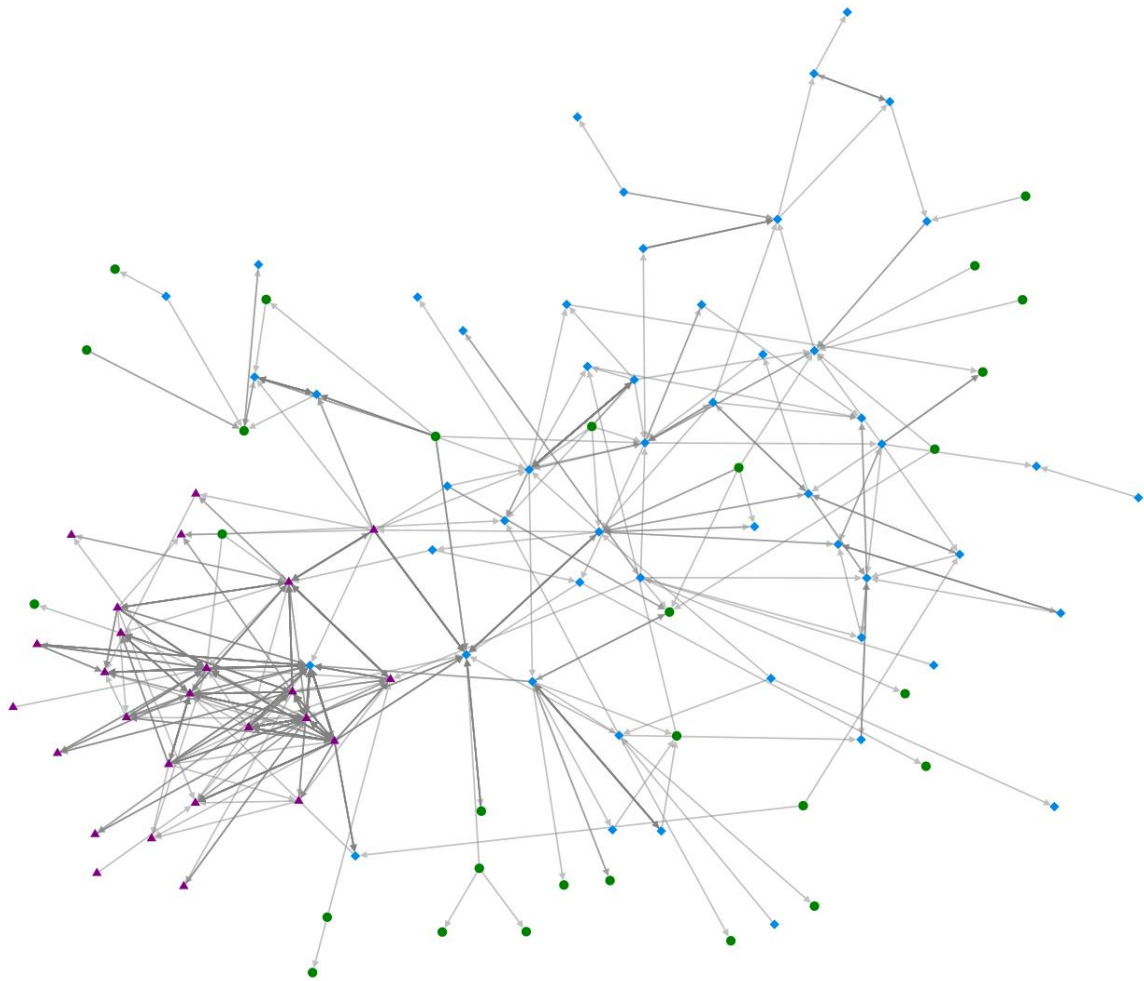
5.4 ANALYTICAL QUESTION 4.C.

4.c. What are the dynamics of data related tweets between the groups and what does this reveal about support for data use on Twitter?

Analytical question 4.c. expands on the data related content investigated in 4.b. by exploring a subnetwork which only features tweets including at least one data related word (data, research, stats, statistics, findings, evidence, report, %) and a URL. This pairing should provide enough tweets to meaningfully review but limit the analysis to only tweets which concern data related content and share a link to that content; this should broadly capture tweets which support data use as discussed in the content analysis. This subnetwork of the full mentions network has 107 nodes with 937 edges between them, comprising 7.35% of the tweets from the full network. This section of analysis will map and analyse this subnetwork using a sociogram, group connection table, and ERGM analysis to determine the dynamics of support for data use on twitter; investigating the supply side of support. This question does not look at the receiving side of support, which is analysed subsequently in question 4.d.

5.4.1 Data mentions sociogram

Sociogram 5-2 Data mentions network sociogram with nodes split by group



Node groups: Charities (green disks), support organisations (blue diamonds), data organisations (purple triangles). Boldness of edges represents intensity of repeated contact. Source: Primary Twitter data

Comparing this sociogram of data mentions to the full network sociogram (Sociogram 5-1) in question 4.a., the data organisations look reasonably similar, which suggests that many of their tweets in the original sample were data related. They also exhibit a strong internal bias to communication with most of their links being to other data organisations. It may be that support organisations are following the data organisations and pick up on these internally sent data tweets, which they then retweet and disseminate to their own charity followers. However when data organisations do have outside contact it appears to be with support organisations so there are bridges for information to flow between these two groups in the form of mentions.

There are far fewer support organisations and charities in this subnetwork than in the full network. The links between these accounts have also thinned out substantially from the full network, though there still is something of a core and a periphery. There appear to be more support organisations present than charities, which makes sense if charities are largely engaging by following and not

being mentioned or mentioning. What is harder to disentangle from Sociogram 5-2 are the exact group dynamics; clearly, the data organisations send mostly internally, but this is less clear for the support organisations and who is actually sending more of this content? Are support organisations retweeting data organisations? Table 5-6 below explores these dynamics in more detail.

5.4.2 Data mentions group connection

Table 5-6 breaks down and tracks all 937 tweets sent in the subnetwork and where they were received. An analysis of the full network (shown in Appendix VII) found a similar pattern of tweeting between the groups in terms of where tweets were sent and received. Therefore the analysis below is a descriptive of the data related subnetwork and does not imply anything causal about the patterns of tweets which is largely a feature of general tweeting between the groups.

Table 5-6 Group connection table for data related content

		Receiving				Out total
		1. Charities	2. Support	3. Data		
Sending	1. Charities	11 (15%)	43 (60%)	18 (25%)	72 (100%)	
	2. Support	28 (13%)	168 (79%)	18 (8%)	214 (100%)	
	3. Data	18 (3%)	159 (24%)	474 (73%)	651 (100%)	
		57	370	510		
		In total				

Cell values are the number of tweets sent between given pairings of groups, parenthesis values are the percentage of outgoing ties to each recipient. Light shading is contact internal to the groups. Dark shading is contact between data and support organisations. Source: Primary Twitter data

As Table 5-6 makes obvious data organisations send the majority of data related tweets (651) in the subnetwork, with support organisations second (214); clearly data organisations are the source of much of the data related content in the network. Charities send less of this sort of content but when they do it tends to be to support organisations – which aligns with previous findings that they do not communicate much with the data organisations. Support organisations do not often mention charities when tweeting this type of content, preferring to mention other support organisations which fits with the broad, widely directed, content of these tweets discussed previously; support is not one-to-one. This form of support from infrastructure organisations, and the distinction between one-to-one and one-to-many support, was noted by Wells and Dayson (2010) as also taking place offline.

Looking more closely at the contact between support and data organisations (the darker blue cells), it does not appear to be the case that support organisations often mention data organisations, with

only 18 instances. It is much more common for this to occur the other way round with 159 instances of data organisations mentioning support organisations. Data organisations may, therefore, be the active partner in disseminating content to the charities network by actively mentioning support organisations to make them directly aware of data related content.



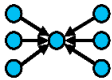
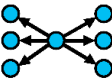


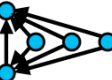

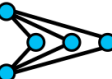




In the general case, data support tweets tend to be internal to the groups, with 70% of tweets being intra-group (the light blue cells). This within-group bias aligns with previous findings; it is not necessary for two accounts to have direct interaction for information to flow between them through following links, which is analysed in question 4.d. The final analysis of this section uses an ERGM framework to diagnose more complex patterns of interaction between the groups in the data support subnetwork.

5.4.3 Data mentions ERGM

Exponential random graph modelling is a method which uses Markov chain Monte Carlo estimation to overcome the violation of the assumption of independence which would occur if a network was modelled with a conventional regression (Shields 2016). This allows for inference in describing the network; it is a way of determining if the observed features are statistically significantly different from random. The goal of ERGM is to determine unusual structures in the observed network, which imply underlying processes, and then, where appropriate, to explain these processes with exogenous covariates (Robins et al 2007). Standard ERGMs have a major limitation in that they cannot accept valued data; that is a network which has an edge weighting to reflect the strength of relationships. This limitation requires the input data to be dichotomised; where this edge weighting is removed. Although this is a common limitation with several network analysis methods, care should still be taken when interpreting the results of methods using dichotomised data. It is possible to set a threshold when dichotomising so that only relationships of a certain intensity (as measured by edge weight) are retained. This was experimented with during the initial ERGM analysis but it was decided to not set a dichotomisation threshold for any of the models presented in the final thesis.

The input data for ERGM is the subnetwork itself and the intercept is the chance of a tie existing, shown as 'Arc'. At a network level, this corresponds to density.

Model 5-1 and 5-2 Basic ERGM and Group ERGM for data content

Parameters		Basic model		Group model	
		Estimates	SE	Estimates	SE
Arc		-5.25	0.23*	-5.43	0.27*
Reciprocity		2.95	0.27*	3.56	0.40*
Alternating in star		0.46	0.16*	0.41	0.16*
Alternating out star		0.41	0.16*	0.34	0.17*
Alternating triangle – transitive		0.60	0.32	0.60	0.27*
Alternating triangle – cyclical		-0.44	0.11*	-0.48	0.09*
Alternating triangle – down		0.59	0.23*	0.52	0.20*
Alternating triangle – up		0.13	0.21	0.14	0.18
Alternating two path – TDU		-0.17	0.04*	-0.16	0.04*
Group matching		-	-	0.73	0.16*
Group matching reciprocity		-	-	-0.83	0.46
Key: Any given node:  Specified group: 					

t-ratios not shown, <0.1 for all effects (see methods). Significance symbol: P<0.05 = * Source: Primary Twitter data



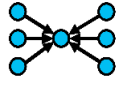
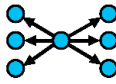























Reviewing the basic model, which only included endogenous variables, the first result is for arc which is the intercept and is not usually interpreted (Shields 2016). The result for reciprocity is positive and significant, suggesting that there are more reciprocal ties in the network than would be expected by random chance, given the density of the network. This suggests conversational relationships between organisations in the network with mentions going back and forth between accounts. The two star effects, which capture popularity and activity respectively, are both significant but small, suggesting moderate popularity and activity stratification with some organisations sending or receiving lots of data related content and others less so. Two of the four triangle effects are significant, cyclical and down, though both are quite small effects. The down effect is positive suggesting that there is a somewhat higher degree of closure or clustering than

would be expected by chance. The cyclical triadic effect is negative which suggests less non-hierarchical closure than random chance. These results together imply that there is a hierarchy in the network; some organisations do not form links to others because of imbalances in prominence or some other inhibiting factor (Robins and Daraganova 2013). The combined two path effect is significant and negative, though small. This represents non-direct contact which can have several interpretations including animosity and brokerage; the two path is an unclosed triad which implies there is some reason for it not closing and becoming a full triangle. This logic is similar to Granovetter's (1973) 'forbidden triangle', the 'forbidden' denoting that this situation is unusual in normal social relations as triangles tend to close without an external factor inhibiting them. In this case, there appears to be slightly less non-direct contact than expected by chance as the effect is negative. The most important finding from this initial model is the reciprocal effect which suggests support tweets are being sent back and forth between pairs of organisations in sustained relationships. This is further explored in the forthcoming models.

The second model expands on the first by including the group each node belongs to as an exogenous covariate; be it charity, support organisations, or data organisations. The covariate is defined as 1 if a dyad pair is of the same group and 0 if they are of differing groups. This covariate reflects a generalised group matching or not matching effect and does not look at the individual groups. In this case, the covariate meets the assumption put forward by Robins and Daraganova (2013) that the network should not be able to affect the exogenous covariate (ties on Twitter do not determine which group an organisation belongs to). This means that any effect predicted for the covariate can be ascribed to its influence on the formation of network ties and not vice-versa.

Two versions of the exogenous effect are included, one for one-way ties and one for reciprocal ties. The latter of these is insignificant, suggesting that being in the same group does not increase the chances of reciprocation of data related content, which is already highly reciprocal due to the base effect. The former, one-way, effect is significant and positive, though small, suggesting a slight tendency for organisations from the same group to send more one way, unreciprocated ties, to one another. This effect, though hinting at within group bias, is not particularly impactful and the overall results from the second model suggest that matching in group is not actually that determinant of the dynamics of data related tweets. This effect is counter to that seen in Table 5-6, likely because the data has been dichotomised, but also because of differences between the groups; the groups may have very different patterns of internal communication which this combined effect cannot differentiate. Model 5-3 below resolves the latter issue by breaking down the groups into binary categories.

Model 5-3 Categorical group ERGM for data content

Parameters		Estimates	SE
Arc		-5.20	0.34*
Reciprocity		2.54	0.92*
Alternating in star		0.33	0.18
Alternating out star		0.25	0.18
Alternating triangle – transitive		0.64	0.29*
Alternating triangle – cyclical		-0.46	0.10*
Alternating triangle – down		0.45	0.21*
Alternating triangle – up		0.08	0.19
Alternating two-path – TDU		-0.07	0.05
Charity sender		[base]	[base]
Support sender		-0.04	0.34
Data sender		-0.66	0.30*
Charity receiver		[base]	[base]
Support receiver		0.22	0.31
Data receiver		-0.95	0.41*
Charity interaction		[base]	[base]
Support interaction		0.23	0.38
Data interaction		2.56	0.50*
Charity activity reciprocity		[base]	[base]
Support activity reciprocity		-0.11	0.87
Data activity reciprocity		2.19	0.80*
Charity interaction reciprocity		[base]	[base]
Support interaction reciprocity		0.51	0.73
Data interaction reciprocity		-2.55	0.88*
Key: Any given node:  Specified group:  Other group: 			

t-ratios not shown, <0.1 for all effects (see methods). Significance symbol: P<0.05 = * Source: Primary Twitter data

In this model, instead of the groups simply matching or not, they are broken down into binary categories. The endogenous effects in this model are present as controls and will not be reviewed. There are five exogenous effects for each group, with charities serving as the reference category. The first two effects are sender and receiver, expressing activity and popularity respectively. Only the data organisations have significant results for these variables, with both of them being negative coefficients. This suggests that, given the number of data organisations in the subnetwork, they actually send and receive less data related content than would be expected. However, as previous results have shown they send by far the most in the non-dichotomised network data, so this result should not bear particular weight as it is likely a result of the dichotomisation process.

The next result is interaction, or how likely the groups are to send one-way ties internally to themselves. This was the significant result from the previous group model and here it is obvious that it is the data organisations who were responsible for this previous result; data organisations have a strong positive result suggesting they send far more one-way ties to each other than would be expected by chance; reflecting their strong internal bias.

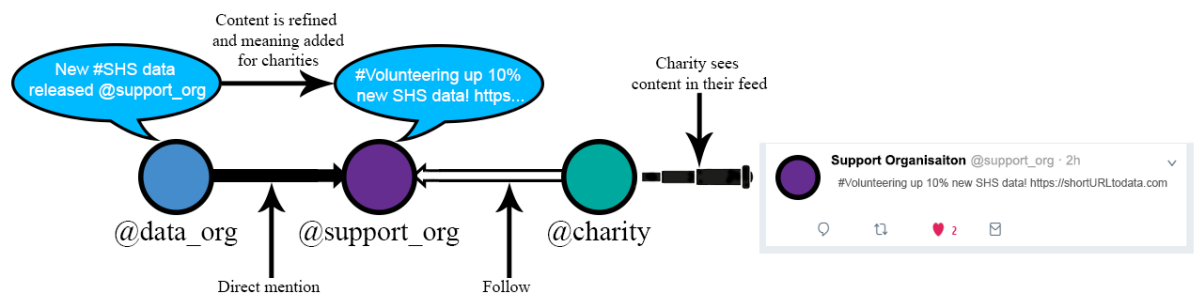
The final two variables are related to reciprocal contact. The first is the baseline reciprocity for each group over and above the generic reciprocity for the model which is strong and positive (generic reciprocity = 2.54). Again, in comparison to the charities, data organisations have the only significant result which is strong and positive. Added to the model's generic reciprocity rate this means that data organisations have a very strong tendency to reciprocate but, as the final result for interaction reciprocity shows, this tendency drops back to the model baseline when data organisations are tweeting to each other. This suggests a stronger link between the data organisations and the other organisations (primarily support organisations) than previously assumed; although data organisations mostly tweet amongst themselves, they are very likely to reciprocate to support organisations or charities when contact is external. Given there are only a few of these sorts of links in the network, as shown in Sociogram 5-2, there is an opportunity for support organisations to form new links with data organisations, who would then be more likely to reciprocate. Although the order of these links cannot be substantiated from the data. With previous analysis finding that data organisations mention support organisations far more than the other way around, this finding from the ERGM analysis further suggests that data organisations are more active in disseminating data related content in the charities network than their portrayal in Sociogram 5-2 would suggest. With data organisations tweeting data related content internally and externally, and strong reciprocity throughout the network, there is a web of supportive content on Twitter. This content is not usually tweeted directly to charities and, may therefore, be spreading to charities by following links which the final question of this chapter investigates.

5.5 ANALYTICAL QUESTION 4.D.

4.d. Is there evidence of support for data use disseminating through following links?

Previous sections have looked at both the content of data related tweets and the dynamics of the network formed of data related tweets. Both of these previous sections have suggested that support for data use on Twitter is not directly targeted at the intended recipient, rather the target of a support tweet is often another support or data organisation who then, often, reciprocates to form a web of data related content. The implication of these previous findings is that support for data use is spreading to charities by following links; support organisations modify and tweet content sent to them by data organisations, and this charity specific content is then viewed by their followers as it is retweeted. This dynamic is shown in Figure 5-1.

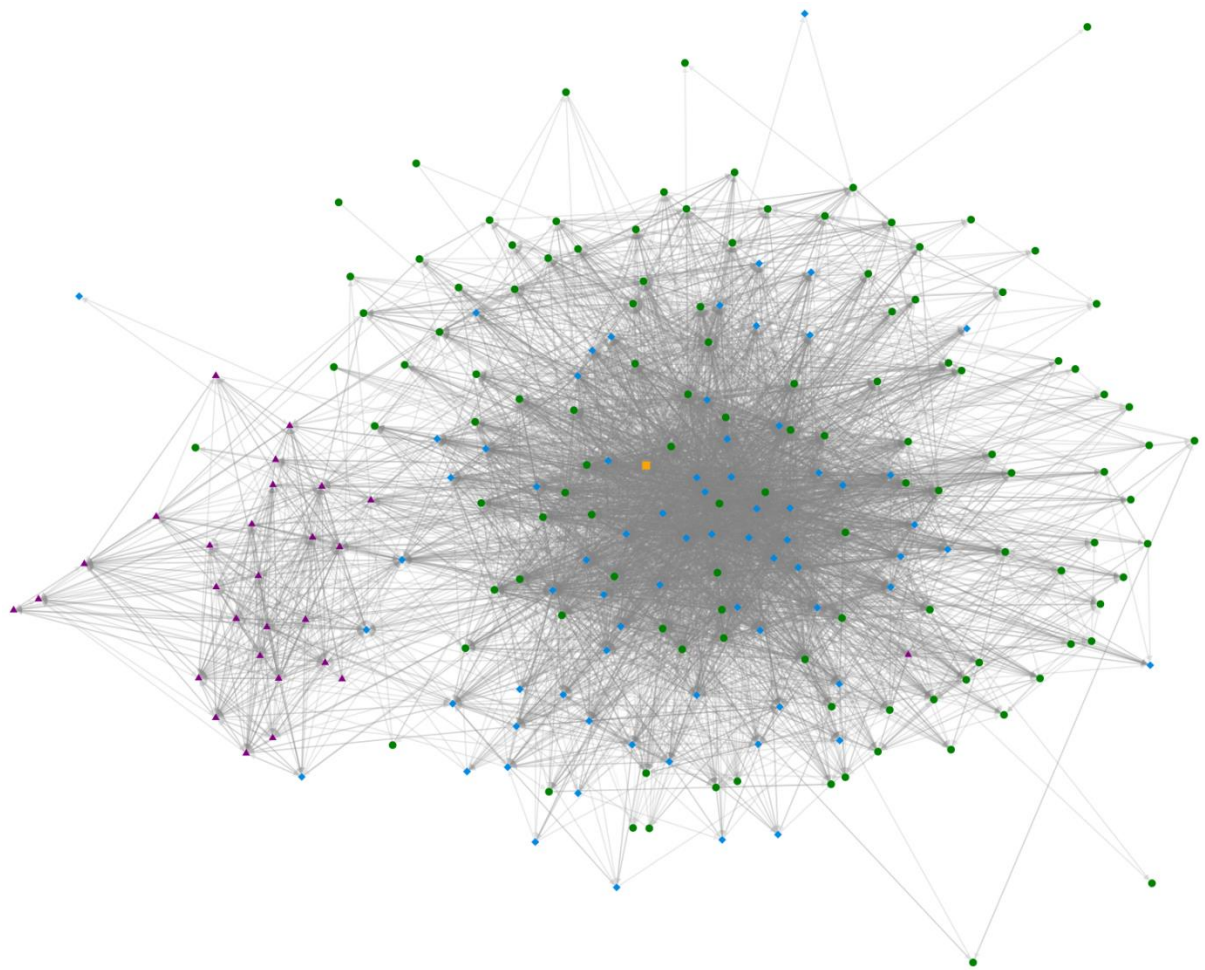
Figure 5-1 Data content dissemination on Twitter diagram



This final section of question 4 attempts to test this dynamic by examining the network of following links, which was collected alongside the mentions network analysed previously (see the methods chapter for details on both networks and how they differ). This network is first described by a sociogram and then modelled with ERGM.

5.5.1 Following links sociogram

Sociogram 5-3 Following network sociogram with nodes split by group



Node groups: Charities (green disks), support organisations (blue diamonds), data organisations (purple triangles), The Scottish Government (yellow square). Source: Primary Twitter data

Sociogram 5-3, laid out in the same way as those discussed previously, shows following links rather than mentions which has some important implications. Firstly, follows are non-duplicable as an account cannot follow another more than once. Secondly, follows are maintenance free once established; unlike mentions, they are not events which happen at a particular time, but relationships which persist until manually removed (Golder et al 2007). This means that the follower's network is extremely dense as can be seen in the sociogram. This can make analysing patterns in following behaviour difficult as there is a lack of variation in the dyadic dependent. Steps are taken to mitigate this issue in the ERGM analysis below, but what is clear from the sociogram is that the data organisations are still separate from the main body of the network. This implies that they have a different, perhaps more complex, relationship with the other organisations than the support organisations do with charities. This could be important because much of the data related content in the network originates from data organisations and how this content disseminates and eventually makes its way to charities is key to understanding support for data use in the third

sector on Twitter. There is direct contact between the support and data organisations, but is there also a following dynamic.

5.5.2 Data mentions and following ERGM

Because data organisations and support organisations, the two key groups in data related content on Twitter, have such different positions in the following network they are assessed separately below. From the sociogram, and previous analysis, it is obvious that there is limited contact between charities and data organisations and so this relationship will not be tested; charities only send 74 follows to data organisations in total. For comparison charities send 3,087 follows to support organisations. Therefore the first section of analysis looks at following behaviour between support organisations and charities, with the second section looking at following between data organisations and support organisations.


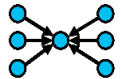
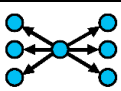
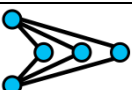

5.5.2.1 Support organisations and charities

The answer to the question: ‘are charities following support organisations who share data related content?’ is a resounding yes. From Sociogram 5-3 above it is clear that charities are following support organisations in large numbers and from the previous content analysis it is known that support organisations do tweet a small but notable volume of data support content. However, this is weak evidence of effective support for data use as there are many reasons why charities might follow or not follow specific support organisations. It is also difficult to prove charities are actually engaging with data related content in their feeds. Therefore, this section seeks to test the hypothesis that charities will be more likely to follow support organisations if they tweet more data related content. This hypothesis is not perfect; it is possible that charity engagement with support is not related to volume of tweets because charities are engaging for some other reason, such as quality of support or a factor not related to data support. Accepting the null hypothesis, therefore, will reveal nothing, but should the analysis allow for the rejection of the null hypothesis, it would provide some evidence of support for data use on Twitter and engagement of charities, at least in a descriptive sense.

Testing this hypothesis with ERGM requires altering the following data. The following network is very dense which inhibits the convergence of ERGM. Therefore, controlling for group and the direction of relationships was carried out externally using data management to create smaller, less dense networks rather than using control variables. For the first ERGM below, the data organisations were removed entirely as the hypothesis concerns charities and support organisations. Follows internal to the groups (charity to charity or support to support) and follows from support organisations to charities were then removed as these ties were extraneous to testing the hypothesis. This left only following ties from charities to support organisations which is the relationship of interest.

The exogenous covariate effect ‘data mentions receiver’ was only specified for the support organisations and is the number of tweets they have sent featuring a data related term (data, research, stats, statistics, findings, evidence, report, %) combined with a URL. The effect specified for this in the model was ‘receiver’ which measures the relationship between the support organisation receiving ties, which are exclusively follows from charities, and them sending data related tweets. The specification of these models meant that only a few endogenous effects could be included; the intercept, in and out star, and two path; neither triangles nor reciprocity are possible in this cut-down data. The key endogenous effect is in star, which reflects the popularity of support organisations, the other effects are structural controls.

Model 5-4 and 5-5 Basic and Data mentions ERGM for charity-support following links

Parameters		Basic model		Data mentions model	
		Estimates	SE	Estimates	SE
Arc		-9.39	0.22*	-9.29	1.34*
Alternating in star		2.83	0.22*	2.35	0.33*
Alternating out star		1.78	0.15*	2.05	0.37*
Alternating two path – TDU		-0.34	0.05*	-0.30	0.11*
Data mentions receiver		-	-	0.06	0.01*

t-ratios not shown, <0.1 for all effects (see methods). Significance symbol: P<0.05 = * Source: Primary Twitter data


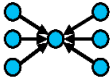
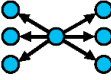
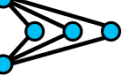

In the results, the exogenous covariate is small but positive and highly significant suggesting that support organisations, which send a greater volume of data related tweets, receive more follows from charities. Supporting this small effect is a reduction of the generic popularity effect from 2.83 in the base model to 2.35 suggesting a small but notable portion of support organisations’ attractiveness as a follow on Twitter is determined by how many data related tweets they send. As the exogenous covariate does not meet Robins and Daraganova’s (2013) causality assumption, this result could also be interpreted more robustly, without assuming any causation, by suggesting that the more popular support organisations send more data related content. There is also the possibility that the data related covariate is a proxy for sending tweets in general and more active support organisations are simply more popular. Even with this more timid conclusion however, this result is still evidence of support for data usage disseminating through Twitter’s following system as the organisations disseminating support for data use tend to be the popular ones. This is true regardless

of whether this is causal or simply a proxy for overall tweeting and is positive finding for data support on Twitter.

5.5.2.2 *Data organisations and support organisations*

The second set of models perform the same analysis but with support organisations following data organisations. This meant that the following data was modified to remove the charities, links internal to the groups and follows from the data organisations to support organisations; leaving only following links from the support organisations to the data organisations.

Model 5-6 and 5-7 Basic and Data mentions ERGM for support-data following links

Parameters		Basic model		Data mentions model	
		Estimates	SE	Estimates	SE
Arc		-5.20	0.46*	-5.39	0.35*
Alternating in star		0.70	0.36	0.69	0.38
Alternating out star		1.54	0.39*	1.53	0.35*
Alternating two path – TDU		-0.10	0.27	-0.06	0.23
Data mentions receiver		-	-	0.01	0.00

t-ratios not shown, <0.1 for all effects (see methods). Significance symbol: P<0.05 = * Source: Primary Twitter data

For this set of models, there is a negligible effect for data organisations receiving more follows if they send more data related tweets. There is also no appreciable decline in the in star effect suggesting that data mentions are not a good indicator of the following dynamics between data and support organisations.

There are several reasons why this might be the case for support and data organisations where there was an effect for charities and support organisations. Firstly, the group connection table has shown that data organisations are more active in sending data mentions to support organisations and so they have a more direct form of contact. Secondly, the data accounts are linked to the network primarily through a small number of bridging support accounts and so they have a more structurally complex following dynamic where the few bridging accounts relay to other infrastructure organisations. This contrasts with charities and support organisations who have a much more simplistic following dynamic, as shown in Sociogram 5-3.

5.6 CONCLUSION

There is evidence of support for external data use on Twitter among charities, support organisations, and data organisations, but revealing and understanding it required context.

The first analytical question provided part of this context by showing differences between the groups in terms of usage of Twitter and network position. It found that support organisations were the most central and active, charities were relatively central, but not nearly as active, and data organisations were active but seemed to form a separate cluster. These findings were corroborated by a discussion around the sociogram visualisation of the network, which showed data organisations in their own cluster linked to the network through a small number of bridging accounts and more charities than support organisations in the outer cloud, explaining their lower metrics.

The second analytical question dug deeper into the meaning underpinning the network by looking at what content and communication is actually shared between the different groups on Twitter. This content analysis found that many types of networking are occurring on Twitter and there was a notable volume of data related content being shared around the network. This data content could largely be described as supportive, but tended to be formatted quite generally and focused on sharing data resources or links to supportive documents such as those around data protection. The latter showing the network attempting to overcome some of the barriers discussed in the previous chapter. In these tweets, the account which was mentioned was almost never the intended recipient of the actual content. These findings suggested a one-to-many style of support for data use where support and data organisations tweet generally supportive content to each other, while the real recipients, charities, see this content on their feed if they follow the sending organisation.

To corroborate this suggestion, the third analytical question isolated a subnetwork of these data support tweets and sought to determine the dynamics of the sending side of support for data use on Twitter. Through the use of a sociogram and group connection table it became obvious that data organisations sent the majority of data related tweets, followed by support organisations, and that both of these groups had a strong internal bias; tending to send these sorts of tweets to other organisations within their group as suggested in the content analysis. The relationship between the data organisations and support organisations was also not as presumed, the data organisations were much more likely to mention the support organisations in data related tweets than vice versa. Finally, an ERGM analysis found that there was a strong tendency for reciprocation in the data support network meaning that the accounts which are mentioned in support tweets also tend to send this sort of tweet back. This builds relationships between support organisations and forms webs of support where a charity following any of these accounts would be informed of other organisations they may want to follow as well as seeing data related content. The ERGM also found that the data

organisations had a notably high rate of reciprocation with the support organisations and charities, meaning that while this form of contact is not as common as within-group tweeting, the few that do exist tend to be reciprocal. This suggests there is an opportunity for, particularly, support organisations to forge new links with data organisations and spread even more data related content to charities.

The final question attempted to provide evidence that the support content, described previously, was actually being engaged with by charities. With the data available from the network being limited this was difficult, but it was hypothesised that charities would be more likely to follow support organisations who tweeted more data related content. The first model found some evidence for this hypothesis but causality was difficult to determine and it was safer to conclude that the more popular support organisations tend to tweet more data related content; whether any of their popularity derives from this sort of content is less important than knowing there are followings relationships between charities and data-tweeting support organisations. The same was not true for support organisations following data organisations, likely due to their previously described, more complex interrelation.

To summarise overall and suggest where improvements could be made, there is evidence of support for data use on Twitter and this manifests as broad one-to-many messages tweeted between support organisations which are seen by charities who follow these support organisations. Much of the data content originates from data organisations, who are quite active in sharing content with third sector support organisations. There was some evidence that barriers identified in the previous analysis, primarily data protection, were being addressed by support on Twitter, but there were only a few instances of this form of support. There may, therefore, be a case for third sector infrastructure organisations to set up a more formal feedback loop by explicitly asking charities what is stopping them from using more data (or using the findings from this thesis) and then sourcing support for these barriers from data organisations. This form of feedback may already exist off of Twitter but having such a mechanism on Twitter would surely increase its reach.

It was not possible, with the data available, to test whether support from Twitter is having a direct effect on charity data usage, but it is clear that support for data usage is part of charity networking on Twitter.

CHAPTER 6: TRUST

6.1 INTRODUCTION

The final analysis chapter of this thesis discusses and analyses the results of twelve semi-structured interviews. The primary theme of this chapter concerns issues around trust which are difficult to measure using survey instruments or other quantitative methods. Trust is an element which cannot be ignored when discussing use of data and this chapter will discuss the issue in depth as well as relating back to previous analysis. This chapter attempts to answer analytical questions from three separate research questions and so synthesises issues of trust from across the thesis. This chapter concerns both of the strands analysis discussed in the introduction, looking at trust as a barrier or enabler and discussing what affect charity use of trust may have on other users of data.

6.1.1 Course of analysis

Chapter 6 is broken into three sections, each answering a single research question:

3. To what extent is external data trusted by third sector organisations and what effect does this trust have on other users of data?

3.a. What level of trust is there for Scottish Government data and other external data among charities and do charities help increase the quality of and trust in external data for other charities and the public?

2. What barriers, enablers, and organisational features affect the ability of third sector organisations to make use of external data?

2.b. What other factors enable or inhibit use of external data in the Scottish third sector?

5. To what extent is the network of charities on Twitter a ‘network of trust’?

5.a. To what extent do links on Twitter embody trust for other organisations and what does this reveal about trust for data?

The first of these sections directly discusses the trust charities place in government data and the mechanisms and structures which relate to this trust. The second section builds on the discussion of trust by extending analysis from Chapter 4, which concerns what factors affect the usage of and, therefore, the trust charities place in government data. Factors discussed in previous chapters will be reconsidered and new factors which respondents mentioned in the interviews are discussed. The final section of the chapter refers back to the social network analysis of Chapter 5 and assesses to what extent the network of charities on Twitter constitutes a ‘network of trust’, recontextualising the previous analysis while also providing an example of charity trust in government.

6.2 ANALYTICAL QUESTION 3.A.

3.a. What level of trust is there for Scottish Government data and other external data among charities and do charities help increase the quality of and trust in external data for other charities and the public?

The first part of this question is the crux of this chapter: is Scottish Government data trusted by the third sector? This section will examine the root sources of trust in government data and how differences in these root sources result in different forms of trust for different organisations. This develops into a discussion around how the different types of organisations feed into building trust in data for other users. There will also be a discussion of how networking affects trust which bridges to the final segment of this chapter discussing the Twitter network.

6.2.1 Is external data trusted?

The picture that builds up from the interviews as a whole is that Scottish Government data is overwhelmingly trusted, as are most other sources of public external data. Frontline charities gave several reasons for this emphatic trust:

“...we take on face value what the government is putting out for SIMD stuff because we reckon that’s big enough....We would probably trust it if it came from Scottish Government.” (Charity 1)

Charity 1 above refers to size, which could have several interpretations but appears to be implying that the government is a large enough organisation to produce trustworthy data. This logic is, perhaps, an extension of charity usage limitations discussed in Chapter 4; the respondent feels that their ability to produce robust data is curtailed by their limited funds and access to skilled staff and that an organisation the size of the Scottish Government will be less limited in this sense and thus able to produce more robust data. This directly links trust to ability to engage with data and it may also be the case that the respondent feels that they don’t have the skills to distrust the government data as alluded to in the following quote:

“Without question, because it’s the government, or it’s the man in the white coat has said it.” (Charity 2)

This respondent felt that the skill imbalance between their charity and the Scottish Government was so great that government produced data is unassailable; going so far as to hyperbolically evoke “*the man in the white coat*”. This form of empathic trust, a projection of the organisation’s own constraints seems particularly tacit. Corroborating this interpretation, this respondent later noted this tacit form of trust directly:

“Trust. Trust is, it's the big one, isn't it, more than anything. There's a tacit trust placed upon government statistics.” (Charity 2)

While not stated, charity 2's use of the phrase 'government statistics' implies that the source of this tacit trust is the official position of these statistics rather than the size of the Scottish Government. Several other respondents commented on this 'official' nature of government statistics and one explained further:

“...the people that collect it, because there's no agenda. You know, it's more of a, I suppose, it's a public, or I see it as a public service. Whereas, with so many other things that you see, there's an agenda, you know, a question's asked in a particular way, or something like that.” (Charity 3)

This quotation shows that the perceived impartiality of government data is part of the reason for the widespread tacit trust placed in it by frontline charities. This impartiality may have been what other respondents referred to as government data being 'official'. This form of trust derived from the official status of the government is similar to the *deterrence* based trust espoused by Shapiro et al. (1992) where trust is maintained by the threat of repercussions if it is broken; the government is bound by law and therefore is trustworthy. This is the weakest form of trust in the scale discussed in the literature review (Kramer and Tyler 1995). However, the tacit and overwhelming nature of charity trust in government data also evokes, to an extent, *instinctive* trust, the strongest form of trust (Kramer and Tyler 1995). It is clear, therefore, that there is widespread trust of government data from frontline charities, though in differing forms. The next section examines the perspective of support organisations.

6.2.2 Feeding back

Before looking at the views of the infrastructure organisations, the quote below is from a respondent in one of the Scottish Government survey teams and seems to be in opposition to the instinctive trust held by the charities:

“...there are some very naive people out there that think it's, you know, whatever the government publish is gospel kind of thing. But we would hope that if people kind of looked and thought that doesn't look right for my area they would actually contact us and say, okay, can you explain those figures to me? How have you worked them out?” (Government 1)

This respondent raises two issues: the interpretation of the data, discussed in the next section, and feeding back, or contact with the producers of the data. Government respondent 1 later elaborated on the latter point:

“But the external stakeholders are really important... it’s the external stakeholders that are really helping us introduce new questions and log what kind of information they use, and how we can re-word questions because some of them aren’t worded very well...” (Government 1)

This quotation notes that feedback and contact are not only important for highlighting and correcting mistakes, but also at the front-end for designing the wording of the questions and surveys in the first place. This respondent uses the term ‘*stakeholders*’ which may imply infrastructure organisations (among other organisations outwith the scope of this study such as local authorities), who also noted that they take part in this feedback loop:

“...they estimated the size of the social enterprise sector to be about a 1,000 but they based it on a, sort of, misinterpretation of data so we were able to, sort of, say, no, based on our database which isn’t complete we’d still estimate there’s at least 4,000 you haven’t found.” (Infrastructure 1)

In this quotation, infrastructure organisation 1 provides an example of feedback they gave from a position of expertise using their own data and experience. Other infrastructure organisations also professed to have fed back in this way; fixing mistakes and helping design questionnaires. As opposed to the *deterrence* based or *instinctive* trust the frontline charities tended to have in government data, the feedback infrastructure organisations are engaged in is an example of *experience* based trust which sits between *deterrence* and *instinctive* trust in the hierarchy discussed in the literature review. Lewicki and Bunker (1995) argue that *experience* based trust can evolve from *deterrence* based trust through repeated contact between the agents, such as feedback mechanisms. This may mean that there would be potential for charities to develop this form of trust if they partook in feeding back for a sustained period of time, though they are of course constrained by available skills and resources.

Moving to the second part of the analytical question, the feedback infrastructure organisations give could also be indicative of healthy distrust or ‘vigilant scepticism’, an idea originally posited by J.S. Mill (Citrin 1974: 988), where the infrastructure organisations feel an obligation to check government data in their area of expertise because they know it sometimes contains mistakes. Government respondent 1, above, expressed the government’s desire for users of their data to engage in these mechanisms. According to ‘Manufactured Risk’ theory, which posits that when experts get things wrong trust is damaged (Giddens 1990), this sort of distrust on behalf of those feeding back is key to maintaining wider trust, including from frontline charities and the public. As well as showing the active role of the infrastructure organisations, the feedback discussed above is also evidence that the Scottish Government has mechanisms in place to attempt to mitigate ‘Manufactured Risk’ by leveraging the expertise of infrastructure organisations. Burt (2007)

discusses how third sector organisations interrelate with the government to build public trust and these feedback mechanisms are an example of this trust building.

Building on the discussion in the introduction to this thesis of the impact of ‘Fake News’ and the general prevalence of untrustworthy sources of news on social media, it is clear that these ‘vigilant sceptical’ infrastructure organisations are a key part of the societal mechanism which can push back against biased and misleading information. Paradoxically the healthy distrust which these skilled and data-literate organisations use to improve the quality of government data is key to countering the unhealthy distrust which results from the permeation of low quality and intentionally deceptive content. The feedback provided by third sector infrastructure organisations doesn’t just help improve trust and use of data in the third sector; it is a cog in a much larger mechanism which helps increase trust and access to good quality data for the whole of society.

This form of feedback was less common among frontline charities, a few of whom claimed to have been to a feedback session or contacted the government but only when prompted and it did not appear to be a major part of their activities. Of course, charities have a lower ability to use data so it may be harder for them to engage with these mechanisms. It may seem like a negative that infrastructure organisations have a lower form of trust in Scottish Government data than charities (*experience* based being lower than *instinctive* in the typology set out in the literature review), but in the light of ‘Manufactured Risks’ it is actually a positive as ‘vigilant scepticism’ is incompatible with *instinctive* trust. Society needs healthy distrust from experts to mitigate ‘Manufactured Risks’. The lack of engagement from frontline charities may also represent a weakness in the government’s feedback system, however, a government respondent noted how they use infrastructure organisations as intermediaries, which can also be seen in the social network analysis of Chapter 5. Therefore, the government may be able to access charity knowledge indirectly through support organisations while not being able to contact or involve every frontline charity directly.

6.2.3 Interpretation

The first quote from the previous section alluded to another issue with trust in data; interpretation:

“And I mean that doesn’t mean that we don’t get asked to contribute to lines to take, and that’s kind of the government’s, kind of, like, you know, line on the data and will comment on policy lines. Or sometimes you do even produce them yourselves, but that is distinct from the actual, you know, report that’s produced which should be a completely objective account of the data.” (Government 2)

This quote, from a respondent with one of the Scottish Government’s survey teams, appears to undermine the impartiality which the charities, partly, base their trust in government data on. However, on closer inspection Government respondent 2 is saying that the data, and report based

on it, are still as objective as possible, it is the interpretation laid on top of the data which is not objective. This issue seems to be lucid for the charities and support organisations, despite the former's trust in the objective nature of the data. Charity 2 summed this issue up perfectly: *"But sometimes, it's actually, it's as simple as that, as how it's put"*. In other words the framing around the data can be manipulated and distorted, or in a more positive sense, meaning can be added. Another charity respondent provided an example:

"It's a difficult one because it's not mistrust in the quality of the data as much as occasional mistrust in the way that the data's interpreted and the way that the data's framed. The prime example for me is the homeless statistics that the Scottish Government publish. The way that it's published and the way that it's presented makes it look like they're presenting figures on homeless individuals. They're not, they're presenting information on advice provided to people self-categorising as homeless. Big difference." (Charity 4)

In this example the respondent alludes to 'vigilant scepticism' as discussed above when they refer to 'mistrust'. In this case the misinterpretation, though having the potential to be quite impactful on research, appears to be a misunderstanding, or perhaps an issue with the meta-data or how the data was framed; in any case it was not an intentional deception. This is an issue which infrastructure organisations could be key in elucidating for frontline charities and some of the support organisations acknowledged this:

"So, sometimes our role isn't necessarily to try and analyse the data it's to try and explain what the data actually means... so it's, kind of, adding a layer of richness to the data as well." (Infrastructure 1)

In a sector with such a shortage of skilled staff to provide this interpretation, infrastructure organisations are key to adding meaning, clarifying and therefore facilitating robust use of data in the sector as a whole and, consequently, increasing trust. This was exemplified in the Twitter analysis in Chapter 5 which found evidence of support organisations retweeting and sharing snippets of data which were designed to be appealing to frontline charities. There were even some direct examples on Twitter of support organisations discussing what a particular source of data meant for the third sector.

Having informed and expert organisations to validate knowledge is key to maintaining trust according to 'Manufactured Risk' theory, as described above, but misinterpretation of data can go beyond poor meta-data or misunderstanding into wilful misinterpretation which can damage trust. One of the government respondents noted:

“I was in criminal justice research and I sat in a few times with the criminal justice stats people when they were releasing the annual figures and talking to the media. And you could sit there all day and say to the media, this is what the statistic says, this is what the statistic says. They still got it wrong the next day. There’s always going to be a wilful misinterpretation.” (Government 3)

Misinterpretation of data is one way which general trust in data and experts is damaged (Doyle 2007). The analysis undertaken by charities and infrastructure organisations is key to generating robust interpretation and to pushing back against the wilful misinterpretation noted above to maintain trust in data.

6.2.4 Inter-charity networking and trust

As discussed throughout this analytical question, trust is not something which exists in isolation, feeding back on data and interpretation both involve contact between organisations and there are a few more specific mentions of networking which reveals another aspect to the interpretation which overlays data. One frontline charity respondent specifically discussed networking and trust:

“But I think, as matters of trust go, I think it's important. That if you don't know the source, that you've at least got somebody else's views that you trust, to talk you through it.” (Charity 2)

This quotation is interesting because it raises several previously discussed issues and ties them to networking; the respondent argues that if the source is unknown then the data is less trustworthy and the user should have a trusted party who can vouch for the data. In a world of completely objective data this should not be the case, but *‘the source’* will affect the interpretation and context of the data and so outside knowledge is required to contextualise. This quotation also hints at networking for increasing use with an external partner to *‘talk you through it’*. This is a more direct form of data use networking than was discovered on Twitter and so these links may exist through more private forms of communication. Finally, the respondent implies that the relationship is where the trust is really invested, not in structures or the data itself. This issue is discussed in more detail in section 6.4.1 which looks at trust and networking in more detail.

Charity 2 also noted how networking is linked to the interpretation of the data:

“...moderating that data before you pass it on, and put, surrounding it with your own context, and meaning...” (Charity 2)

This takes the discussion of interpretation, discussed previously, a step further, the meaning laid over the data is not only provided by the original publisher of the data, but by agents who analyse or reinterpret the data for their own purposes. This reinterpretation is crucial for making the data amenable to particular audiences; infrastructure organisations add context which will make the data

easier for charities to digest, charities add meaning which makes the data suitable for their stakeholders or the public, as do news organisations. This adding of interpretation may be direct, such as a news organisation or charity going straight to the source of data, but as the respondent above discusses, there is often a chain of contact involved. This mediation of meaning is an example of agency predominating over structures. A visible example of this is when the media report a finding from a trusted charity, The Joseph Rowntree Foundation for example, based on government data. The data is seen as objective, but the trust really comes from the reputation of the organisation analysing it and adding meaning.

6.3 ANALYTICAL QUESTION 2.B.

2.b. What other factors enable or inhibit use of external data in the Scottish third sector?

This question refers back to the survey analysis in Chapter 4 which looked at factors which affect charities' data use, but the analysis was limited to factors which appeared on the secondary survey. This section of qualitative analysis will corroborate and extend the discussion around these factors while also considering enabling or limiting factors which the respondents generated from free recall. Given how important skills and level of use appeared to be in question 3.a. for determining what form of trust an organisation places in government data, this section of analysis acts as a supplementary discussion of issues which may directly affect trust in data.

6.3.1 Budget and staffing

The first factor the interviewees noted when discussing limitations to data use was lacking funds and investment:

"We, the third sector, work on very tight budgets, ever decreasing budgets"
(Charity 2)

This limitation, which was reviewed in Chapter 4 but, led directly to the most common limiting factor mentioned in the interviews; issues around staff with the right skills:

"...because I'm the only person here with that skillset, there's nobody I can delegate that to." (Charity 1)

Staffing was not measurable in the survey analysis but the situation articulated by the quote above seemed to be common across the charity respondents; some charities are lucky enough to have someone with skills to engage with data, and others are not. This ties into the discussions around level of use from Chapter 4. The view above, from a frontline charity, is also shared by several infrastructure organisations in the sample who, despite their role and position within the sector, often rely on the skills of just one or two members of staff for data analysis. Support organisations seem to place greater weight on having these skills and are more likely to ensure they have a capable member of staff.

With low numbers of capable staff, support organisations could be vulnerable to staff leaving, which was a barrier to data use noted by several interviewees which seems to be separate from not being able to afford data literate staff altogether:

“We used to have two members of staff who knew how to do that, but over the last three or four years, they’ve both left just to go to other posts and one of them retired.” (Infrastructure 2)

The quote above is a prominent support organisation noting the loss of capacity when capable staff moved on. The fluidity of staffing in the sector seems to make it difficult for charities to rely on data analysis and so they cannot make it a core part of how they operate which suggests staffing is a major factor inhibiting the use of data in a binary sense. This issue can be even more disruptive however, as noted by another infrastructure organisation below:

“...the challenge that people faced, was around databases and client records and [Microsoft] Access databases that have been built by volunteers years ago that have long since left and they have no idea how to amend it, how to get the information out of it.” (Infrastructure 3)

In this quotation a support organisation is referring to frontline charities and, though not discussing external data, makes the point that staff leaving can damage the ability of others in the organisation to use certain resources. If the data analyst on the team is a facilitator for other activities or staff then their loss could be damaging beyond their official remit. They also make the point that if the data literate staff members leave, resources could become inaccessible, further damaging data use ability and potentially trust.

6.3.2 Collaboration and formatting

Budget and staffing were inhibiting factors which were expected to be important in determining charity data use and trust in data; they were prompted for in the interviews. The factors discussed below were not prompted for in the interviews; they derived naturally from the semi-structured nature of the discussions and on the respondents’ own initiatives.

The first of these factors was collaboration:

*“there’s not a huge amount of capacity so partnership is the way it has to be.”
(Infrastructure 1)*

This factor ties directly into the network analysis of Chapter 5. Charities do not exist in isolation and many are very well linked to their peers and support organisations. It therefore makes sense for organisations to work on initiatives together, share capacity for data analysis, share findings, or share data. Twitter may provide a perfect forum for this sort of collaboration and peer-support.

Indeed, there was some evidence of this sort of peer collaboration noted in the content analysis in section 5.3. However, one respondent noted a problem with this enabler:

“Because there’s big open data, open data this, open data that. Well, actually it’s not very open. It’s still very: it’s mine, you’re not playing with it.” (Charity 4)

This frontline charity, which had a relatively high level of data use and a lot of internal data, noted that they used lots of government and public data, but that it was difficult to get other organisations to share resources. The respondent implied that this was for selfish reasons, perhaps based on competition for limited funding, and this may imply an agency based barrier, but several infrastructure organisations noted the, perceived, threat posed by data protection laws:

“Our IT team go into organisations and there is very serious, kind of, data protection risks for organisations that I think small organisations don’t understand or have the capacity to manage effectively, client records on USB sticks given to volunteers...” (Infrastructure 1)

As the support organisation above notes, data protection can be lacking in organisations which do not have staff with the skills to properly manage data. Another support organisation noted that their stakeholders were concerned about data protection, especially in the wake of several notable data protection scandals in the past few years. Fears over data protection could easily lead to organisations shutting down access to their own data and stifling collaboration which would make this barrier to data use and collaboration more structural than the initial quote suggested. Analysis from Chapter 4 found that ‘data security concerns’ and ‘data privacy concerns’, similar to data protection, were significant in determining whether an organisation used data at all suggesting this barrier could be quite impactful on both use and trust. Worries about data protection could also be particularly pertinent at the time of writing and for several years hence due to new data protection legislation which came into force on the 25th of May 2018 (European Commission 2018). Perhaps ironically, a potential solution to concerns over data protection is through collaboration; several support organisations noted that they were providing training for these new regulations and the content analysis from Chapter 5 found that this form of support was present on Twitter, in small quantities. Charities should be more open and accepting of help rather than hiding poor data in fear of GDPR.

The second major barrier which came out of the interviews was formatting and categorisation of data:

“...biggest frustration, probably the only real true frustration is the way that data’s categorised, because everybody puts things into different types of categories, so the age categories are always different. You’re never sure if the, like, household’s data, that kind of thing, if they’re using the same definitions that you are and that drives us crazy...” (Charity 4)

This quotation came from a frontline charity, as did every other mention of formatting or categories in the data being a barrier. Charity 4 alluded that this is particularly problematic when using a range of external data sources, as the formatting keeps changing between them. Another respondent noted a very similar problem and gave a specific example:

“So you get one that says, Edinburgh, one says Edinburgh City, and one says Midlothian. But the files on the website didn't match the names on the shape file. So you couldn't actually link...” (Charity 3)

Charity 1 also gave an example of government and council areas not matching up and this barrier seemed particularly apparent not just for charities using many different data sets, but those using many different sources of data, such as the Scottish Government, UK Government, and councils; all of whom have different ways of formatting data. This issue was alluded to in Chapter 4, where wider-ranging charities were found to be making more use of data. Given complaints over formatting in the interviews came from the more data literature respondents, this factor appears to be a barrier to the upper end of use, limiting those already making some use of data. Adding to this problem is poor meta-data which, while not named directly, is implied by several respondents who felt confused by external data resources, and this is also a feature which varies by source.

6.4 ANALYTICAL QUESTION 5.A.

5.a. To what extent do links on Twitter embody trust for other organisations and what does this reveal about trust for data?

Having looked at issues of trust in data and expanding on the discussion around factors which may, partly, determine levels of trust, this final question explores an example of charity trust. Building on the Twitter analysis of Chapter 5, this final section attempts to determine if the Twitter network formed by charities and infrastructure organisations constitutes a network of trust. This is achieved by establishing that the respondents see Twitter as a place for inter-organisational contact and then discussing whether these links constitute a network of trust.

6.4.1 Twitter for inter-organisational networking

Chapter 5 proved that Twitter is used for charity to charity contact to some extent but, because of the sampling frame it used, it was not possible to determine the balance of this form of contact to other contact, such as charities talking to their own public, as only inter-organisational links were

captured. Therefore, despite the volume of contact uncovered previously, this first section will investigate what sort of contact Twitter is really for:

“...we don't really know why we need to have it, and it's kind of organically generated itself to be our organisational arm where we interact with organisations rather than communities.” (Charity 1)

The above quote is from a frontline charity who strongly felt that Twitter was for contact with other organisations over communication with the public. This view was echoed by several other charities and by infrastructure organisations: *“...particularly useful for that sector to sector or business to business type communication.” (Infrastructure 1)*. Respondents also contrasted Twitter with Facebook which was noted as the platform for charity to public interaction:

“...we don't have Facebook because our support is to third sector organisations rather than to members of the public.” (Infrastructure 4)

It seems likely, given the number of contacts charities often have on Twitter (see Chapter 5), that there is crossover between pre-existing offline inter-organisational networks and Twitter networks. This means that Twitter may, in some way replicate offline networks, but it also implies that Twitter connects charities with each other in ways which are not replicated offline; Twitter allows contact between charities who would not otherwise communicate. This is important context for the dynamics of charity relationships on Twitter.

An issue which arises from this dynamic was how respondents viewed the accounts they were connecting to. One respondent felt that Twitter was for connecting to organisations rather than ‘real people’, while another felt that their Twitter links with organisations were interpersonal:

“...conversations seem to happen more between the people in organisations...”
(Infrastructure 3)

This alludes to the contention between structure and agency and aligns with similar findings by McCabe and Phillimore (2012) who argued this interpersonal aspect of social networks mean organisations functioned in networks of agency. However, there is limited data from the interviews to explore this distinction further. The next section will look at what respondents used Twitter links for.

6.4.2 Is Twitter a ‘network of trust’?

The first issue, which came from the interviews concerning Twitter as a network of trust, relates to retweets, which form the majority of the network. An infrastructure organisation said this about retweets:

“But I don’t know if it’s really us saying, oh yes, we’re [Infrastructure 2], we’ve checked this...” (Infrastructure 2)

This fits with a common statement on Twitter that retweets are not endorsements, but others noted that they were very careful with what they retweeted in spite of this:

“We have got on the profile that retweets are not endorsements, like anybody else. We don't even really trust that the people will believe that.” (Infrastructure 4)

This seemed to be a more common view, that retweets are, in some way, going to be seen as endorsements and so organisations must be careful what they retweet. Charity 4 was more forceful in this assertion, claiming that they would not retweet something they had no confidence in and so retweets could be seen as endorsements. This alludes to the second contention between the respondents concerning Twitter as a network of trust: trust in content versus trust in source.

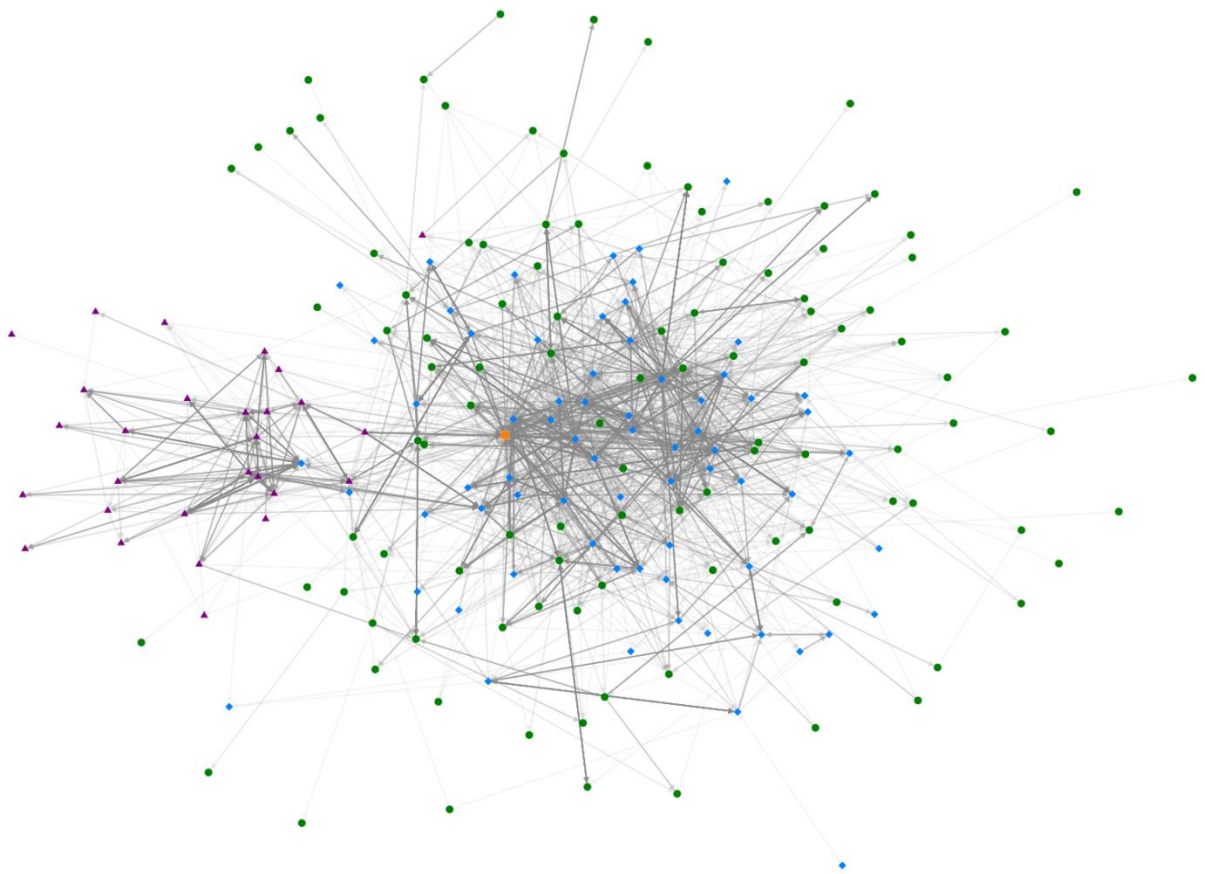
Infrastructure 4 mentioned that they only retweeted ‘things’ that they trusted. Their use of ‘things’ suggests that they trust the content that they are reposting, but the majority of respondents were opposed to this, instead placing trust in the source of the content:

“...but by and large, we know who it is that we can trust... And you can't check absolutely everything.” (Charity 2)

The quote from charity 2 typifies the more common view, that it is the organisation they are retweeting whom they place trust in and not the actual content. This makes sense for two reasons: firstly, although data is generally trusted, as previously discussed, the context and interpretation around data convolutes this trust and so the source, who is responsible for the interpretation, is a greater determinant of trust than the content. Secondly, as the quote above alludes to, it would be time consuming to check every piece of content and, as shown previously, many organisations lack the skills and time to do this checking and therefore put trust in their peers and support organisations as sources. This is in conflict with Section 6.2.4 where a respondent argued that each user adds interpretation to data before passing it on. While this may still be true, it appears that most charities place trust in sources rather than content and therefore trust is mostly peer-to-peer and cannot really be incorporated into data itself and passed down the chain; it is dyadic.

The reciprocal nature of Twitter use may also tie into a feature of *instinctive* trust, the strongest form of trust, which Kramer and Tyler (1995) argued was reciprocal by nature. While reciprocity on Twitter does not necessarily imply *instinctive* trust, combined with the views of the respondents above that they know who they can trust, this is evidence of strong bonds between third sector organisations. Therefore, for most charities retweets embody trust in some form and a network of retweets, as shown in Sociogram 6-1 below, could be viewed as a ‘network of trust’.

Sociogram 6-1. Sociogram of only retweets: a ‘network of trust’



Node groups: Charities (green disks), support organisations (blue diamonds), data organisations (purple triangles), The Scottish Government (yellow square). Source: Primary Twitter data

This is a subnetwork of the mentions network, examined in Chapter 5 (Sociogram 5-1), which was made up of 71.5% retweets (see table 5-5). This ‘network of trust’ therefore, has a very similar structure and interpretation to Sociogram 5-1. This includes the infrastructure organisations in the core, the charities in the peripheries, and the data organisations in their own cluster.

6.5 CONCLUSION

Scottish Government data is trusted by the third sector. This point was borne out overwhelmingly in the interviews with both charities and infrastructure organisations agreeing. For charities, reasons for this trust included the size and capacity of the government, its official status, and its perceived impartiality. This led to both weak *deterrence* based trust and strong *instinctive* trust. For support organisations, who tended to be more engaged with feedback mechanisms for data, trust was mostly derived from these mechanisms in the form of *experience* based trust; they know that the government responds when they feedback mistakes and they are often involved in designing the questions in the first place. This made support organisations the perfect ‘vigilant sceptics’ of government data; at the same time trusting and distrusting in a way which allows them to mitigate

'Manufactured Risks' and counteract 'Fake news'. Scottish Government feedback mechanisms are structural enhancers of trust. The issue of interpretation and context complicated the overwhelmingly positive trust of charities, with organisations trusting the underlying data but not always how it had been interpreted, especially in the light of wilful misinterpretation. Frontline charities dealt with this issue by placing trust in support organisations and peers they networked with who had the skills to engage with data and contextualise it in a way which is palatable for charities. Alternative to the structural feedback mechanisms above, this mediator of trust was agency based with individual organisations or people within them playing key roles in interpretation. This also linked to support and networking on Twitter as discussed in Chapter 5.

The second section of the chapter expanded on the discussion of barriers and enablers to data use first analysed in Chapter 4. Given the ability to engage with data appears to be key to what level of trust an organisation places in data, these factors have the potential to affect both trust and usage. The semi-structured interviews gave respondents a chance to lead the discussion in the selection of factors and the most pertinent factor in these discussions was staffing, a variable which it was not possible to analyse in the surveys but has been discussed in the literature. Respondents with the ability to engage with data generally felt privileged that they had the right people, or more often person, in post. Respondents without noted the lack of funds and the fluidity of staffing in charitable organisations. The fluidity of staffing in the third sector also meant it was difficult to retain skills and so it was difficult to rely on it and integrate data use into the core of the organisation. Collaboration was mentioned by several organisations as an enabler for data use and a way around staffing issues, which aligns with findings of support for data use on Twitter from Chapter 5, but one respondent noted that sharing of data is often limited, perhaps by fears over data protection and other legal issues. This led to a discussion of data protection concerns which were a notable barrier to data use, as discussed in Chapter 4, and could have a blocking effect on sharing of data or could even scare organisations out of analysing non-sensitive data; without data literate staff it can be hard to determine what is safe and what is not. Infrastructure organisations were found to be providing some support for data protection concerns but given how detrimental it appears to be for charity data usage they could be doing more to advertise their support. A final barrier mentioned by several frontline charities was the disparate formatting and categories in different data sets which made it difficult to synthesise evidence from several sources. Using multiple sources seemed to be quite common in the sector, particularly among wider ranging organisations, and with limited skills this barrier was particularly problematic for frontline charities. These issues could be mitigated by improving the quality and accessibility of meta-data as well as by more support from infrastructure organisations in interpreting what data actually means, which they appear to do at present on Twitter to some extent.

The final section of analysis helped corroborate and expand on the network analysis in Chapter 5. Findings from the interviews would seem to corroborate the framing and findings of Chapter 5, at least in terms of Twitter being for inter-organisational contact; respondents overwhelmingly agreed with this assessment, some even saying they only had Twitter because of the need to contact other organisations. This led to the final section which examined whether links on Twitter could be considered proxies for trust. This came down to a discussion around retweets and if retweets implied trust. Although one respondent felt uncomfortable with this, the majority had confidence in what they were retweeting. Crucially, this confidence tended to stem from trust in the organisation they were retweeting and not in the content of the tweets; similar to the trust placed in other organisations to interpret data as discussed in question 3.a. This dynamic suggests an actor-based form of trust rather than a structural one and meant that, in general, retweets between charities on Twitter can be considered a sign of trust and therefore the Twitter network, once limited to retweets only, as show in Sociogram 6-1, can be viewed as a ‘network of trust’.

CHAPTER 7: DISCUSSION

This chapter will bring together the findings from the preceding analytical chapters and integrate them into a larger discussion which will refer back to the previous literature. The chapter begins with a summary of how each research question has been answered and highlights where the findings align with the existing literature and where they represent new contributions. The second part of the chapter is a more narrative discussion which integrates the findings into the two strands of analysis discussed in the introduction; the first being factors which affect use while the second focuses on the critical role of trust.

7.1 SUMMARY OF FINDINGS TO RESEARCH QUESTIONS

1.a. What level of external data usage is there among charities in Scotland?

Charities struggle with data. Findings from both the statistical analysis in Chapter 4 and the interviews in Chapter 6 showed that there is a spectrum of data use in the Scottish third sector but, in general, engagement with raw data from frontline charities is low. This generally low ability to use data aligns with previous research (Andrei et al 2012; De Las Casas et al 2013; Steiner et al 2015). Interview data indicated that infrastructure organisations were more data literate in general than frontline charities. Chapter 4 revealed that many charities rely on previously analysed data and aggregate findings rather than using raw data themselves which is not a distinction which is clear in the existing literature but is important for the other research questions in this thesis.

2.a. Which organisational features best predict differences in levels of use?

Size, age, and scope were the strongest organisational factors which came out of the statistical analysis and interviews. Generally smaller, older, and narrower organisations were making poorer use of available data resources. Size appeared to be the most important of these factors, as it determines what resources a charity has to invest in data as discussed in previous literature (Ellis and Gregory 2008; Ógáin et al 2012). Age was only discussed, briefly, by one report in the literature (Lloyds Bank 2016) while scope or spread of focus is a new finding and contribution to the wider literature.

2.b. What other factors enable or inhibit use of external data in the Scottish third sector?

From the statistical analysis, data protection concerns arose as the most important barrier and this reflects this factor's prominence in the existing literature (Curvers et al 2016; W. Hall et al 2012; Lloyds Bank 2016). Staffing was also identified as an important barrier, though this became apparent in the interviews as staffing levels could not be accurately measured in the statistical analysis. Staffing also featured prominently in the previous literature with many previous authors discussing its links to size and available resources (Boswell et al 2016; De Las Casas et al 2013;

Ógáin et al 2012; Steiner et al 2015). The respondents also stressed that high staff turnover in the sector is problematic for building data capacity which is a limiting factor discussed by some existing literature (Burt and Otto 2017; Cunningham 2001).

For enablers, accountability requirements and leadership emerged from the statistical analysis, both of which were discussed in previous literature, particularly accountability and impact reporting (Amar 2017; Harlock 2013). Factors which featured in the literature but which did not emerge from the analysis included fundraising, which had no easy proxy in the statistical analysis and was not mentioned in any of the interviews, and IT systems which was simply insignificant in the statistical analysis.

The statistical analysis also found that barriers, when aggregated, appear to determine binary usage of data by charities; if they are using data or not. Enablers, alternatively, seem to best predict variations in levels of use. This dynamic is not covered in any of the literature which was reviewed during this thesis and is a new contribution of this research, though further research would be required to fully explore this.

3.a. What level of trust is there for Scottish Government data and other external data among charities and do charities help increase the quality of and trust in external data for other charities and other users of data?

There is a high level of trust in Scottish Government data from both frontline charities and infrastructure organisations. These groups trust for different reasons which results in different forms of trust. Charities seem to innately trust Scottish Government data based on the government's official position, skills, and perceived impartiality which is identified in the literature as *instinctive* trust (Coulson 1998). Charities also appeared to place this level of trust in other official sources of external data. Infrastructure organisations, alternatively, trust based on interactions with the government through feedback mechanisms which leads to *experience* based trust (Lewicki and Bunker 1995). The infrastructure organisations also have a healthy distrust of the data which allows them to engage with these feedback mechanisms as 'vigilant sceptics' to increase the quality of the data and mitigate 'Manufactured Risks' as discussed by a small section of the literature, notably Beck (1992) and Giddens (1990; 1999).

This, extremely positive, picture of trust in external data was complicated somewhat by the issue of interpretation; it was clear from the interviews that while the respondent charities place a great amount of trust in the data, they are not as trusting of the interpretation which can overlay data. This is a particularly pertinent issue for frontline charities because, as previously discussed in question 1.a., many of these organisations make greater use of aggregate findings rather than raw

data and this is where interpretation could prove most distorting. This is particularly problematic because charities who do not have the skills to use raw data are unlikely to have the skills to tell poor or biased interpretation from good analysis. Charities appear to mitigate this issue by placing their trust in particular sources of interpretation which includes the government themselves but also infrastructure organisations. This relates to the support which infrastructure organisations provide on Twitter which is discussed in question 5.a. below. The issue of trust in interpretation, and trust as a factor in charity data use as a whole, does not feature in the existing body of literature. This thesis addresses that gap in the literature.

4.a. To what extent do charities, infrastructure organisations, and data organisations use Twitter differently and what implications does this have for how they are networked?

All three of these groups have a notable presence on Twitter. Support organisations appeared to be the most active group both in terms of tweeting and networking behaviour. While the charities and infrastructure organisations formed a network of mentions, the data organisations appeared to be somewhat separated and linked to the main body of the network by several bridging accounts. These findings serve to set up the following network analysis.

4.b. What content is actually exchanged on Twitter between charities, infrastructure organisations, and data organisations, how much of this content is data related, and what form of support do these data related tweets embody?

There is an impressive depth to the content shared on Twitter between these groups and often hyperlinks are employed to share more details which previous literature has shown (Lovejoy and Saxton 2012). General content in the network collected for this research tended to discuss events or charitable campaigns and utilise the network to spread awareness. Data related content only made up a fraction of tweets sent in this network but this sort of content was shared in appreciable volumes. In general these data related tweets either shared links to training events/resources or, more commonly, directly to data or findings with some contextualising information. This reflects Twitter as an information sharing platform which is discussed in previous literature (Kwak et al 2010), but without regard for data related content and charities specifically. The findings here show a complex and nuanced relationship between data, charities, and social media.

4.c. What are the dynamics of data related tweets between the groups and what does this reveal about support for data use on Twitter?

Perhaps surprisingly, this section of analysis found that there is only limited direct data related tweeting between the support organisations and charities. Direct tweeting was assumed to be the primary method of disseminating data related tweets in a one-to-one format (Wells and Dayson

2010) but the network formed by this content was dominated by intra-group tweeting; support and data organisations tweeting to each other in reciprocal relationships and not to charities. As there is no literature relevant to data related tweeting by charities, a network of following links was consulted to fully understand the dynamics of data content sharing between these groups, as discussed below.

4.d. Is there evidence of support for data use disseminating through following links?

Given the findings of question 4.c. it appeared that support for data use may be spreading via following links rather than direct mentions. This claim was difficult to test as the following network was very dense; there is a very high level of following between charities and support organisations. This means that data related tweets between support organisations will be appearing on the Twitter feeds of charities as it is common for charities to follow support organisations. This basic finding was expanded upon with an ERGM analysis which suggested that support organisations who tweet more data related content were more likely to be followed by charities. This analysis did not control for the overall number of tweets sent or other factors affecting the popularity of support organisations so is not causal but still describes a situation where charities are more likely follow support organisations who tweet a higher level of data related content. Regardless of why they tend to follow these accounts this is a positive situation for delivering support for data use through Twitter relationships. This section of analysis, combined with 4.c. above, suggests that support for data use on Twitter is one-to-many in format from the support organisation's perspective, but still based on peer-to-peer relationships between charities and trusted support organisations from the charity's perspective. This is a new finding not covered by the existing literature.

5.a. To what extent do links on Twitter embody trust for other organisations and what does this reveal about trust for data?

This question involved synthesising evidence from the interviews and networking analysis around retweeting. The majority of respondents, both charities and infrastructure organisations, expressed that their retweets could be seen as expressions of trust, but it was clear that this trust was placed in the organisation they were retweeting and not in the content. Many of the charities professed to not having the time or skills to check everything they were retweeting diligently, which aligns with findings from question 1.a. and 3.a., but they knew which accounts they could trust. This forms a 'network of trust' as charities retweet content from those they trust, but more than that it suggests that trust is not something which can be added to or embedded in data; it is peer to peer and based on idiosyncratic relationships individual to each charity. Content or data shared by an account trusted by one charity might not be trusted, and therefore is not retweeted or shared, by another. This dynamic around Twitter and trust has not featured in the previous literature.

7.2 DISCUSSION

Charities are struggling to effectively use data and it is clear that more needs to be done to support them. Organisational factors related to generally low levels of data use included small organisational size, organisational age and a narrow scope of focus. While there is no way to change or mitigate these factors, knowing that these types of organisations tend to struggle could allow infrastructure organisations to more effectively target support at those most in need. Targeted tweeting is not a form of support they appear to currently be engaged in to any notable degree and this may represent an opportunity for the future. Previous literature has covered both, generally low levels of charity data use, and factors such as size and age having an impact on levels of use (Ellis and Gregory 2008; Lloyds Bank 2016; Ógáin et al 2012). The findings from this first section of analysis are, therefore, not entirely new, but the aggregate effects of barriers and enablers on data use has not featured in any previous study. The analysis found that, in general, barriers, most notably fears over data protection and problems hiring or training staff, inhibited data from being used at all. Enablers, alternatively, were what drove data use once it had been established and included leadership and IT infrastructure, though IT was insignificant in this analysis. The distinction between binary and scope effects on data use has not been described before and could be important for how support is supplied to charities; it may be ineffective to provide support to enhance enablers if there are barriers still to be overcome. Trust in data was not found to be a notable barrier as there were high levels of trust in external data, though this was somewhat complicated by issues of interpretation as discussed below.

Current support to help overcome these barriers and enhance enablers, at least on Twitter, tends to focus on sharing resources and advertising training events; both quite wide and general forms of support which are discussed in the existing literature (Macmillan et al 2014; Ógáin et al 2012). The most notable topic of support in this study, beyond sharing direct links to data, was GDPR data protection regulations which are one of the biggest current barriers to charity data use, having just come into force at the time of writing. Although previous literature had touched on the sort of content used to support charity data use on social media, no study to date has looked at the dynamics of support for charity data use on social media. The network analysis in this thesis addressed that gap and found that support on Twitter tended not to be directly tweeted at charities, as may be expected. Rather, support was tweeted reciprocally between infrastructure organisations where it formed a network of support and disseminated to charities by their following behaviour; this one-to-many style of support fits in with the wide and general format of the supportive content. This means that, beyond a simple broadcasting platform, Twitter is a forum for relationships between charities and infrastructure organisations. For infrastructure organisations support is one-to-many, but each charity receives support based on idiosyncratic relationships with support organisations; they choose which infrastructure organisations to follow and engage with. These

relationships are how support is then disseminated. There was evidence that infrastructure organisations retweeting more supportive content tended to be followed by more charities but this finding was descriptive and not particularly robust. There is very little previous literature covering charities and social media and even less which looks at the interaction of charities with infrastructure organisations on social media so the dynamics of the support relationships described in this thesis represent a significant contribution to that emerging body of literature.

The second strand of analysis looked at the effects which charity trust in, and use of, data had on other users of external data, the first conclusion was that charities and infrastructure organisations overwhelmingly trusted external data, particularly Scottish Government data. Surprisingly, this finding is not mirrored in the extant body of literature as the issue of trust has been notably absent from charities research. Although both charities and infrastructure organisations trusted government and external data, it was very apparent in the interviews that these two types of organisation trusted for very different reasons. Charities tended to trust in the underlying data based on instinct and the official position of the government. Their trust in the interpretation placed on top of data stemmed from relationships with other organisations, notably on Twitter which acted as a platform for supportive relationships as previously discussed. This means that the root of charity trust in data and in interpretation of data was in relationships with other organisations and, ultimately, people rather than being in data or interpretation in itself which they did not have the skills to properly assess. This means that, in general, frontline charity interaction with data does not appear to have a notable effect on the quality of, or trust in, data because their trust is based on peer-to-peer relationships and cannot be embedded for other users. For infrastructure organisations the picture was very different. Infrastructure organisations tended to trust based on experiences with the producers of data; particularly participating in feedback mechanisms and being involved in designing survey instruments and questionnaires. This meant that, in absolute terms, these organisations were less trusting of the data than charities which allows them to act as ‘vigilant sceptics’; able to use their knowledge of their own subject area to catch mistakes and help improve the quality of data resources. An assumption made at the start of this research was that it would be frontline charities who took up this role but it appears that without the skills to work with data they are generally unsuited to this and instead any feedback or expertise they lend is vicarious through infrastructure organisations. This finding relates to the small body of literature formed around the work of Giddens (1999) concerning ‘Manufactured Risks’. Although the Scottish Government is likely aware of the engagement of infrastructure organisations in its feedback mechanisms, this has not been borne out by any previous research. This finding therefore contributes to the ‘Manufactured Risk’ literature as a theoretical development through a real world case study; Giddens focused on science and technology as ‘Manufactured Risks’ in his work, but there has been little acknowledgment of the danger posed by data in this body of literature. In our modern, data driven, society where concerns over the misuse of data, such as the Cambridge Analytica

scandal, are ever increasing, this is a new branch of the ‘Manufactured Risks’ theory which needs to be addressed. This thesis, not only, highlights the importance and potential danger of data but also how it is managed as a ‘Manufactured Risk’, through the engagement and feedback of infrastructure organisations. The findings of this research around trust also contribute to the more general body of charities literature which has ignored the concept despite how intertwined it appears to be with how charities engage with data; previous literature concerning charity data use is missing an essential part of the picture.

Third sector infrastructure organisations have a very privileged role in society; they not only help design and correct government produced data, but also help disseminate that data to frontline charities in an amenable format which charities trust. Support organisations, therefore, directly control how data will be interpreted, trusted, accessed, and to a large extent used by charities. Charities and support organisations approach data from different directions, charities are very trusting but lack the skills to engage, while infrastructure organisations generally have more skills which they use to both aid charities and improve the quality of data resources based on a healthy distrust. Twitter appears to be at least one forum where charities and support organisations meet in the middle and form relationships which aid the former in engaging with data. Infrastructure organisations appear to be largely unconscious of how important they are in every aspect of charity use of data and the importance of Twitter in how they deliver support. This unconsciousness ties into a general lack of focus from the existing literature on the role of infrastructure organisations in support for data use on social media, only Dayson (2010) and McCabe and Phillimore (2012) can lay claim to having broached elements of these findings before. Additionally, trust is entirely ignored by the existing literature and the findings of this thesis, therefore, represent a substantial contribution to that literature and an example of why trust should not be ignored in the future. The main findings and contributions made by this research are summarised in Table 7-1 below.

Table 7-1 Main findings and contributions to the literature (sorted by importance)

Finding number	Finding/contribution	Research question(s)	Previous literature
I	Infrastructure organisations are extremely important, both in supporting charities’ use of data through one-to-many relationships on Twitter and in improving the quality of data resources for all users by participating in feedback mechanisms and maintaining a healthy scepticism in the data.	4.b., 4.c., 4.d., 3.a.	(Giddens 1999; Macmillan et al 2014)
II	Trust is a crucial component of understanding charity data use. Although there is a high level of trust in raw data from charities, a lack of skills mean	1.a., 2.b., 3.a., 5.a.	-

	they are not engaging with this raw data and are instead relying on aggregate findings and pre-analysed data where trust is less universal and is based on peer-to-peer relationships with trusted partners, most notably infrastructure organisations.		
III	Twitter is a forum for support relationships which facilitate charity data use. For the infrastructure organisations, the broadcast nature of Twitter allows them to share data and supportive content on a one-to-many basis and support a large number of charities. For the recipient charities Twitter is a place of idiosyncratic relationships where they put trust in certain support organisations and the content they Tweet.	4.a., 4.b., 4.c., 4.d., 5.a.	-
IV	Barriers to data use seem to determine if data is used or not in a binary sense while enablers have a greater effect on the level or scope of data use.	2.b.	-
V	In general, smaller, older, and narrower charities tend to struggle the most with data.	2.a.	(Ellis and Gregory 2008; Lloyds Bank 2016; Ógáin et al 2012)
VI	There is a generally low ability to use data among frontline charities.	1.a.	(Andrei et al 2012; De Las Casas et al 2013)

CHAPTER 8: CONCLUSION

Having discussed the findings and implications of this research, and where it contributes to the literature, this final section will cover a few final topics before an ultimate summary. These topics include the weaknesses of the research, what direction future research could take and what recommendations for charities, infrastructure organisations, and the Scottish Government emerge from the findings of this research.

8.1 WEAKNESSES

As with any topic which has had little attention in previous research, this research has weaknesses. There was a desire to provide a rounded and panoptic overview of the topic, but this required employing various methods which meant that each section of analysis was necessarily limited within the time constraints of a doctoral thesis. Had the analysis focused entirely on trust, more interviews could have been performed, a focus on Twitter and more network analysis would have benefited the discussion around support, and more survey data would have increased the power of the analysis of usage, allowing for individual barriers and enablers to be better assessed. However, because of the interrelation between data use, social media relationships, and trust, it was deemed necessary to cover all three aspects of this dynamic even if it meant that each individual aspect was not explored as fully as it would have been in a narrower thesis.

This leads to a discussion of the more specific limitations and weaknesses of the research. A weakness apparent throughout the first analysis chapter was the difficulties of surveying charities, the bespoke survey designed by the researcher did not garner many responses and even the secondary survey sourced from a previous project had limited power for multivariate analysis. The root of this problem is that, even with a good response rate, accessing charities on a large scale is difficult. Email distribution lists proved largely ineffective as a method for reaching charities; perhaps reflecting some of the findings of this thesis, charities preferred to be approached peer-to-peer which takes far more researcher time and proved unfeasible for this research.

The final part of the network analysis, which looked at the popularity of infrastructure organisations who share data related content, would have benefitted from more controls for the overall popularity and activity of the support organisations as this would have increased the robustness of the findings. It would also have been insightful to model the effect of viewing support on Twitter on levels of charity data use but due to the surveying problems discussed above the data was not amenable to this form of analysis.

8.2 FUTURE WORK

Although this research answered many questions, it also generated and posed new ones. The role of trust in how charities engage with data was one of the key findings, but as this is entirely new to the third sector literature there are still many unanswered questions concerning this aspect of data use and charity relationships more generally. It may be that different types of charity have different levels of trust, and therefore engagement with relationships which facilitate data use. In general, future work needs to acknowledge the role of trust in how charities interact and use data.

This thesis focused on charity relationships with infrastructure organisations and on Twitter, but there are other bodies and platforms where trust may manifest differently or to different extents which warrant study. The Third Sector Interface (TSI) bodies, another type of infrastructure organisation which were not covered in this research for sampling reasons¹³ may also be important facilitators of data use and charity trust and contact with them would be an interesting avenue for further study. The other major social media platform charities use is Facebook but it may be more interesting for future work to attempt to glean insights into the more private forums of communication charities use – such as email or face-to-face networking at events.

Given the critical importance of infrastructure organisations in the sector, there is scope for further studies which focus on these organisations. In particular the support which infrastructure organisations receive from the government, each other, or elsewhere to keep their data use skills sharp is an important piece of the puzzle which is currently under-researched. Any diminution in the ability of infrastructure organisations to engage with data could have profound negative impacts on third sector data use and trust in data more generally, so it is important to know how these organisations are developing skills.

8.3 RECOMMENDATIONS

Table 8-1 Recommendations for stakeholders based on findings

Recipient	Research question	Finding number	Recommendation
Scottish Government	3.a.	I	The importance of third sector infrastructure organisations being as it is, the Scottish Government should continue to actively involve them in all parts of the data creation process, from design of the questionnaires to feeding back errors in the published data.

¹³ There is a single TSI for each of Scotland's 32 council regions which makes them difficult to study without focusing on one region.

Scottish Government	3.a., 4.c., 4.d., 5.a.	I, IV, VI	Following on from the point above, the Scottish Government should assess what support it is giving to infrastructure organisations and make sure that they are adequately equipped to perform the important role that this thesis has highlighted. It may be prudent to focus skills development on infrastructure organisations rather than frontline charities as support organisations are likely to make the biggest difference with these skills and act as a force multiplier for data use in the third sector as a whole.
Infrastructure organisations	3.a., 4.c., 4.d., 5.a.	I, II, III	Infrastructure organisations should be aware of the extremely important role they play in how data is used in the third sector, acting as both support and conduit for frontline charities while having a role in data creation as described above.
Infrastructure organisations	2.a., 2.b.	I, IV, V	Given their importance to the sector and improving the quality of data resources in general, infrastructure organisations should ensure that they invest in the staff and skills to maintain their ability to interact with data.
Infrastructure organisations	4.a., 4.b., 4.c., 4.d.	III	Infrastructure organisations should be aware that Twitter is an important forum for support relationships with frontline charities and not a trivial activity, the content they post and support they provide should therefore be of high quality. They should also ensure that they have the staff to adequately engage with Twitter and that these staff are engaged in dialogue with the staff who deal with data and analysis.
Infrastructure organisations	2.a., 4.b., 4.c.	III, V	With Twitter being so important to infrastructure organisations may want to consider more directly targeting or tailoring support tweets to those charities which tend to struggle the most. This includes; small, old, and narrowly focused organisations. Twitter is a network built on reciprocal relationships so a few direct tweets could turn into a longstanding relationship.
Charities	2.a., 2.b., 4.c., 4.d.	III, IV, VI	Throughout this thesis it has been apparent that charities, in general, do the best they can with the resources they have available and therefore they need to be smarter

			about how they access and use data rather than simply being urged to spend more money. One of the most obvious routes is increasing use of social media, particularly Twitter. While most charities already use Twitter, it is often not seen as a core part of their day-to-day activities and this research has shown that it can be an extremely cost effective way to access support and resources.
Charities	4.a.	II, III	Twitter is an important forum for charity relationships and information dissemination, but accessing social media resources requires an investment of time to build networks and engage with the platform. Charities should start by mirroring the connections they already have offline by following organisations they know and trust.
Charities	4.b.	III	The ERGM analysis revealed that reciprocity is a powerful force in charity Twitter networks and therefore charities should contact infrastructure organisations publically on Twitter to solicit support if they feel the one-to-many content in the network is not fulfilling their needs.
Charities	4.a.	II, III	Charities with more ability to analyse data should consider sharing data and findings with their peers and building up networks to share analysis, potentially on Twitter. Though in the light of GDPR and competition over funding, this may be an unrealistic request.

8.4 CONCLUDING REMARKS

Charities trust data, but they don't have access to the skills to fully utilise the available resources. With our world becoming increasingly data-centric this is a missed opportunity which will only grow with time. There is no panacea to increase the sector's ability to use data; money and resources are at the root of the problem but there is no reason to believe these will be any less restricted in the future. The sector has, therefore, developed cost and time effective ways of accessing quantitative information; generally, by sourcing aggregate findings or interpretation from trusted infrastructure organisations rather than performing analysis in-house on raw data; effectively short-cutting data analysis. This has placed a strong emphasis on the role of trust as, while the data itself is seen as trustworthy, charities know that interpretation laid over data can be open to manipulation and therefore they must trust those they choose to source data and interpretation from. This gives infrastructure organisations an extremely privileged role in the

sector as they effectively act as a mediator between raw data and frontline charities; even if they are not performing the analysis themselves, they are the trusted source who shares and adds credibility to interpretation and findings. Additionally, because trust appears to be something which is peer-to-peer and cannot be embedded within data, infrastructure organisations must perform their role for a massive number of frontline charities. Therefore, they mostly use one-to-many forms of support, such as Twitter, where they tweet each other to build up a network of support which can be widely accessed rather than directly targeting particular charities.

This arrangement is working to some extent, but if third sector data use is to be increased and improved, all of the participants need to be aware of their roles. Charities with a lack of resources to perform full analysis should embrace the use of pre-analysed data and focus on sourcing this information through networking with more able charities and infrastructure organisations. More able charities should consider making their data and skills available to their less able peers; beyond worries over data protection and competition, this form of inter-charity sharing is likely to have benefits for both parties. Infrastructure organisations need to be aware that they are the fulcrum around which third sector data use turns and make sure that they are fulfilling their obligations in the data creation process by supporting frontline charities. At present infrastructure organisations are likely unaware of how important the role they perform is to the third sector and to wider users of data. The Scottish Government should focus on facilitating the infrastructure organisations by involving them in the data creation and feedback process and providing support to help develop their skills when necessary; they tend to already be quite data literate, but the data use of many charities rests on the shoulders of relatively few infrastructure organisations. Charities should trust that data will improve the way they function and put their trust in infrastructure organisations and other charities to obtain the data they need in a format they can use. They should trust in the data.

BIBLIOGRAPHY

- Adamic, L.A. and Adar, E. (2003) Friends and neighbors on the web. *Social Networks*, 25 (3), pp. 211-230.
- Adler, E. (1997) Seizing the middle ground: constructivism in world politics. *European Journal of International Relations*, 3 (3), pp. 319-363.
- Amar, Z. (2017) *Three ways charities are using data creatively*. <http://blog.justgiving.com/three-ways-charities-are-using-data-creatively/>.
- Andersen, R. (2008) *Modern methods for robust regression*. London: Sage.
- Andrei, K., Pope, E., Hart, A. and Quinn, L.S. (2012) *The State of Nonprofit Data*. NTEN (The Nonprofit Technology Enterprise Network). https://www.nten.org/NTEN_images/reports/2012%20Data_Report_FINAL_b.pdf.
- Arceneaux, N. and Schmitz Weiss, A. (2010) Seems stupid until you try it: Press coverage of Twitter, 2006-9. *New Media & Society*, 12 (8), pp. 1262-1279.
- Auger, G.A. (2013) Fostering democracy through social media: Evaluating diametrically opposed nonprofit advocacy organizations' use of Facebook, Twitter, and YouTube. *Public Relations Review*, 39 (4), pp. 369-376.
- Beck, U. (1992) *Risk society: Towards a new modernity*. London: Sage.
- Bernard, H.R. (2012) *Social research methods: Qualitative and quantitative approaches*. London: Sage.
- Berry, W.D. (1993) *Understanding regression assumptions*. London: Sage.
- Big Lottery Fund (2015) *Reaching Communities: Additional funding for organisations funded through Reaching Communities and Reaching Communities buildings*. Big Lottery Fund. <https://www.tnlcommunityfund.org.uk/-/media/Files/Programme%20Documents/Reaching%20Communities%20England/RC%20June15%20-%20stage%20one%20form%20LIVE.pdf>.
- Borgatti, S.P. (2013) *Analyzing social networks*. London: Sage.
- Boswell, K., Gyateng, T. and Noble, J. (2016) *Freeing up health analysis*. NPC (New Philanthropy Capital). <https://www.thinknpc.org/resource-hub/freeing-up-health-analysis/>.
- British Broadcasting Corporation (2017) *What is 2017's word of the year?* <https://www.bbc.co.uk/news/uk-41838386>.
- Broomhead, P. and Lam, O. (2017) *Skills Survey 2016-2017*. The FSI (The Foundation For Social Improvement). <http://www.thefsi.org/wp-content/uploads/2017/03/Skills-Survey-16-17.pdf>.
- Bryman, A. (2012) *Social research methods*. Oxford: Oxford University Press.
- Bubb, S. and Michell, R. (2009) Investing in third-sector capacity. In: P. Hunter, ed. *Social Enterprise for Public Service: How Does the Third Sector Deliver?* London: Smith Institute, pp. 74-81.

- Burr, V. (2003) *Social constructionism*. London: Psychology Press.
- Burris, V. (2004) The academic caste system: Prestige hierarchies in PhD exchange networks. *American Sociological Review*, 69 (2), pp. 239-264.
- Burt, E. (2007) Voluntary organizations in the democratic polity: Managing legitimacy, accountability and trust. *Public Money and Management*, 27 (2), pp. 157-160.
- Burt, E. and Otto, S. (2017) *Scottish Charities: Engaging with Data – An Introduction to the Organisational Case Studies series*. Scottish Third Sector Data Network. <https://www.thinkdata.org.uk/media/ThinkData/Documents/CaseStudies/ScottishCharitiesEngagingWithDataFin.pdf>.
- Burton, S. and Soboleva, A. (2011) Interactive or reactive? Marketing with Twitter. *Journal of Consumer Marketing*, 28 (7), pp. 491-499.
- Cairney, P. (2012) Complexity theory in political science and public policy. *Political Studies Review*, 10 (3), pp. 346-358.
- Carolan, B. (2014) *Social Network Analysis and Education: Theory, Methods & Applications*. Thousand Oaks: Sage.
- Cate, F.H. (2008) Government data mining: The need for a legal framework. *Harvard Civil Rights-Civil Liberties Law Review (CR-CL)*, 43 (2).
- Chanley, V.A., Rudolph, T.J. and Rahn, W.M. (2000) The origins and consequences of public trust in government: a time series analysis. *Public Opinion Quarterly*, 64 (3), pp. 239-256.
- Charity Commission (2013) *Charity Commission: current partners*. Government Digital Service. <https://www.gov.uk/government/publications/charity-commission-partnership-strategy/charity-commission-current-partners>.
- Choldin, H.M. (1988) Government Statistics: The conflict between research and privacy. *Demography*, 25 (1), pp. 145-154.
- Christiansen, T., Jorgensen, K.E. and Wiener, A. (1999) The social construction of Europe. *Journal of European Public Policy*, 6 (4), pp. 528-544.
- Citrin, J. (1974) The political relevance of trust in government. *The American Political Science Review*, 68 (3), pp. 973-988.
- Claeskens, G. and Hjort, N.L. (2008) *Model selection and model averaging*. Cambridge: Cambridge University Press.
- Clark, A. (2018) *Key findings from Tech Trust's Digital Charity Survey 2018*. Charity Digital News. <https://www.charitydigitalnews.co.uk/2018/01/12/key-findings-from-tech-trusts-digital-charity-survey-2018-infographic/>.
- Contractor, N.S., Wasserman, S. and Faust, K. (2006) Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example. *Academy of Management Review*, 31 (3), pp. 681-703.
- Cook, T.E. and Gronke, P. (2005) The skeptical American: Revisiting the meanings of trust in government and confidence in institutions. *Journal of Politics*, 67 (3), pp. 784-803.

- Coulson, A. ed. (1998) *Trust and Contracts: Relations in local government, health and public services*. Southampton: Hobbs the Printers.
- Couper, M.P. (2000) Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, , pp. 464-494.
- Crane, D.R., Busby, D.M. and Larson, J.H. (1991) A factor analysis of the Dyadic Adjustment Scale with distressed and nondistressed couples. *American Journal of Family Therapy*, 19 (1), pp. 60-66.
- Creswell, J.W. and Plano-Clark, V.L. (2007) *Designing and conducting mixed methods research*. 2nd ed. London: Sage.
- Crotty, M. (1998) *The foundations of social research: Meaning and perspective in the research process*. London: Sage.
- Cunningham, I. (2001) Sweet charity! Managing employee commitment in the UK voluntary sector. *Employee Relations*, 23 (3), pp. 226-240.
- Curvers, S., Gripper, R. and Lomax, P. (2016) *Valuing Data: How to use it in your grant-making*. NPC (New Philanthropy Capital). <https://www.thinknpc.org/resource-hub/valuing-data-how-to-use-it-in-your-grant-making/>.
- Dayson, C. (2010) *Understanding personalisation: Implications for third sector infrastructure and their work with organisations on the frontline*. Sheffield Hallam University. <https://www4.shu.ac.uk/assets/pdf/cresr-understanding-personalisation-mainreport.pdf>.
- Dayson, C. (2011) The personalisation agenda: implications for organisational development and capacity building in the voluntary sector. *Voluntary Sector Review*, 2 (1), pp. 97-105.
- Dayson, C. and Sanderson, E. (2014) *Working Paper 127: Building capabilities in the voluntary sector: A review of the market*. Birmingham: Third Sector Research Centre. <https://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/working-papers/working-paper-127.pdf>.
- De Las Casas, L., Gyateng, T. and Pritchard, D. (2013) *The power of data: is the charity sector ready to plug in?* NPC (New Philanthropy Capital). <https://www.thinknpc.org/wp-content/uploads/2018/07/The-power-of-data.pdf>.
- Doyle, A. (2007) *Trust, citizenship and exclusion in the risk society*. Law Commission of Canada. http://brunswickbooks.ca/website_pdfs/riskandtrust.pdf.
- Ellis, J. and Gregory, T. (2008) *Developing monitoring and evaluation in the third sector: Research report*. Charities Evaluation Service. <https://www.homelesshub.ca/resource/accountability-and-learning-developing-monitoring-and-evaluation-third-sector-research>.
- ESRC (2014) *Research Data Policy*. <https://esrc.ukri.org/files/about-us/policies-and-standards/esrc-research-data-policy/>.
- ESRC (2015) *Framework for research ethics*. https://www.gla.ac.uk/media/media_326706_en.pdf.
- European Commission (2018) *EU General Data Protection Regulation (GDPR)*. <https://www.eugdpr.org/>.

- Everitt, B.S. and Howell, D.C. eds. (2005) *Encyclopedia of Statistics in Behavioral Science*. New York: John Wiley & Sons, Ltd.
- Fiesler, C. and Proferes, N. (2018) "Participant" Perceptions of Twitter Research Ethics. *Social Media Society*, 4 (1).
- Firth, D. and De Menezes, R.X. (2004) Quasi-variances. *Biometrika*, 91 (1), pp. 65-80.
- Fisher, R.J. (1993) Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20 (2), pp. 303-315.
- Frank, O. and Strauss, D. (1986) Markov graphs. *Journal of the American Statistical Association*, 81 (395), pp. 832-842.
- Gálvez-Rodríguez, M.d.M., Gálvez-Rodríguez, M.d.M., Caba-Pérez, C., Caba-Pérez, C., López-Godoy, M. and López-Godoy, M. (2016) Drivers of Twitter as a strategic communication tool for non-profit organizations. *Internet Research*, 26 (5), pp. 1052-1071.
- Gayle, V. and Lambert, P.S. (2007) Using quasi-variance to communicate sociological results from statistical models. *Sociology*, 41 (6), pp. 1191-1208.
- Gelman, A. and Hill, J. (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Giddens, A. (1974) *Positivism and sociology*. London: Heinemann.
- Giddens, A. (1990) *The Consequences of Modernity*. Cambridge: Polity Press.
- Giddens, A. (1999) Risk and responsibility. *The Modern Law Review*, 62 (1), pp. 1-10.
- Golder, S.A., Wilkinson, D.M. and Huberman, B.A. (2007) Rhythms of social interaction: Messaging within a massive online network. *Communities and technologies 2007*. Springer, pp. 41-66.
- Gomm, R. (2009) *Key concepts in social research methods*. Basingstoke: Palgrave Macmillan.
- Gorard, S. (2003) *Quantitative methods in social science research*. London: A&C Black.
- Granovetter, M.S. (1973) The strength of weak ties. *American Journal of Sociology*, 78 (6), pp. 1360-1380.
- Guo, C. and Saxton, G.D. (2014) Tweeting social change: How social media are changing nonprofit advocacy. *Nonprofit and Voluntary Sector Quarterly*, 43 (1), pp. 57-79.
- Gyateng, T. (2017) *How to create an impact data lab*. NPC (New Philanthropy Capital). <https://www.thinknpc.org/wp-content/uploads/2018/07/How-to-make-an-impact-data-lab-Final-1.pdf>.
- Gyateng, T., Pritchard, D. and de Las Casas, L. (2013) *Creating a 'Data lab'*. NPC (New Philanthropy Capital). <https://www.thinknpc.org/wp-content/uploads/2018/07/Creating-a-Data-Lab.pdf>.
- Hall, W., Shadbolt, N., Tiropanis, T., O'Hara, K. and Davies, T. (2012) *Open data and charities*. Nominet Trust.

- Hall and Taylor. (1996) Political science and the three new institutionalisms. *Political Studies*, 44 (5), pp. 936-957.
- Harel, D. and Koren, Y. (2000) A Fast Multi-Scale Method for Drawing Large Graphs. Proceedings of the Working Conference on Advanced Visual Interfaces.
- Harlock, J. (2013) *Impact measurement practice in the UK third sector: a review of emerging evidence*. Third Sector Research Centre.
<https://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/working-papers/working-paper-106.pdf>.
- Harrell, M.C. and Bradley, M.A. (2009) *Data collection methods: Semi-structured interviews and focus groups*. National Defense Research Institute.
https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR718.pdf.
- Hauser, P.M. (1973) Statistics and politics. *The American Statistician*, 27 (2), pp. 68-71.
- Hill, R. (1998) What sample size is "enough" in internet survey research. *Interpersonal Computing and Technology: An Electronic Journal for the 21st Century*, 6 (3-4), pp. 1-12.
- Hoare, G. and Noble, J. (2016) *How to make your data more meaningful*. NPC (New Philanthropy Capital). <https://www.thinknpc.org/resource-hub/how-to-make-your-data-more-meaningful/>.
- Holt, D.T. (2008) Official statistics, public policy and public trust. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171 (2), pp. 323-346.
- Idre (2011) *What are pseudo R-squareds*. <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/>.
- Idre (2017) *Logistic regression analysis stata annotated output*.
<https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/>.
- Institute of Fundraising (2016) *How and why do charities use personal data?* Institute of Fundraising. <https://www.institute-of-fundraising.org.uk/library/how-and-why-do-charities-use-personal-data/>.
- Institute of Fundraising (2017) *GDPR & Charitable Fundraising; Introduction*. Fundraising Regulator. <https://www.fundraisingregulator.org.uk/sites/default/files/2018-07/GDPR-briefings-intro.pdf>.
- Jachtenfuchs, M. (2002) Deepening and widening integration theory. *Journal of European Public Policy*, 9 (4), pp. 650-657.
- Jackson, R., Karp, J., Patrick, E. and Thrower, A. (2006) *Social Constructivism Vignette*. Athens: University of Georgia.
<https://pdfs.semanticscholar.org/328b/5b784b72f3f0f5c084b3eabf001ccc90c07c.pdf>.
- Jacobson, R. (2013) *2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?* <https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>.
- Johnson, P.A. (2000) A nonparametric analysis of income convergence across the US states. *Economics Letters*, 69 (2), pp. 219-223.

- Jung, K. and Valero, J.N. (2016) Assessing the evolutionary structure of homeless network: Social media use, keywords, and influential stakeholders. *Technological Forecasting and Social Change*, 110, pp. 51-60.
- Knoke, D. and Yang, S. (2008) *Social network analysis*. London: Sage.
- Kramer, R.M. and Tyler, T.R. (1995) *Trust in organizations: Frontiers of theory and research*. London: Sage.
- Krugman, P. (2009) How did economists get it so wrong? *New York Times*, 2 (9), pp. 2009.
- Kukla, A. (2000) *Social constructivism and the philosophy of science*. London: Psychology Press.
- Kwak, H., Lee, C., Park, H. and Moon, S. (2010) What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web.* , pp. 591-600.
- Leat, D. (2011) *New tools for a new world (or why we need to rethink capacity-building)*. Big Lottery Fund. No longer available.
- Lewicki, R.J. and Bunker, B.B. (1995) Trust in Relationships: A Model of Trust Development and Decline. In: B.B. Bunker and J.Z. Rubin, eds. *Conflict, Cooperation and Justice*. San Francisco: Jossey-Bass, pp. 133-173.
- Lewicki, R.J., McAllister, D.J. and Bies, R.J. (1998) Trust and distrust: New relationships and realities. *Academy of Management Review*, 23 (3), pp. 438-458.
- Lincoln, Y.S., Lynham, S.A. and Guba, E.G. (2011) Paradigmatic controversies, contradictions, and emerging confluences, revisited. *The Sage Handbook of Qualitative Research*, 4, pp. 97-128.
- Lloyds Bank (2016) *UK Business Digital Index 2016*. Lloyds Bank.
https://www.tpdegrees.com/globalassets/pdfs/research-2016/ukbusinessdigitalindex_oct16.pdf.
- Long, J.S. and Freese, J. (2014) *Regression models for categorical dependent variables using Stata*. College Station: Stata press.
- Lovejoy, K. and Saxton, G.D. (2012) Information, community, and action: How nonprofit organizations use social media. *Journal of Computer-Mediated Communication*, 17 (3), pp. 337-353.
- Macmillan, R., Paine, A.E., Wells, P., Kara, H., Dayson, C. and Sanderson, E. (2014) *Building Capabilities in the Voluntary Sector: What the evidence tells us*. 125. Third Sector Research Centre. <https://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/reports/research-report-125-building-capabilities.pdf>.
- McCabe, A. and Phillimore, J. (2012) *Seeing and doing: learning, resources and social networks below the radar*. Third Sector Research Centre.
<https://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/working-papers/working-paper-91.pdf>.
- McCulloh, I., Armstrong, H. and Johnson, A. (2013) *Social Network Analysis with Applications*. Somerset: John Wiley & Sons, Incorporated.
- McKnight, D.H. and Chervany, N.L. (2001) Trust and distrust definitions: One bite at a time. *Trust in Cyber-societies*. London: Springer, pp. 27-54.

- McMahon, M. (1997) Social constructivism and the World Wide Web-A paradigm for learning. *Australasian Society for Computers in Learning in Tertiary Education (ASCILITE) Conference*. Perth, Australia.
- MELODA (Metric for Releasing Open Data) (2018) *Dataset definition*. <http://www.meloda.org/dataset-definition/>.
- Millar, R. and Hall, K. (2013) Social return on investment (SROI) and performance measurement: The opportunities and barriers for social enterprises in health and social care. *Public Management Review*, 15 (6), pp. 923-941.
- Morris, D. (2005) New Charity Regulation Proposals for England and Wales: Overdue or Overdone. *Chicago-Kent Law Review*, 80 (2), pp. 779-802.
- National Council for Voluntary Organisations (2014) *How many voluntary organisations are active in the UK?* <https://data.ncvo.org.uk/a/almanac14/how-many-voluntary-organisations-are-active-in-the-uk-3/>.
- Neumayer, E. (2002) Do we trust the data? On the validity and reliability of cross-national environmental surveys. *Social Science Quarterly*, 83 (1), pp. 332-340.
- Nooteboom, B. (2002) *Trust: Forms, foundations, functions, failures and figures*. Cheltenham: Edward Elgar Publishing.
- Office of the Scottish Charity Regulator (2018) *Overview of Scottish charities*. <https://www.oscr.org.uk/about-charities/overview-of-scottish-charities>.
- Ógáin, E., Lumley, T. and Pritchard, D. (2012) *Making an impact: Impact measurement among charities and social enterprises in the UK*. New Philanthropy Capital.
- Phethean, C., Tiropanis, T. and Harris, L. (2013) *Automated analysis of charities' communication styles on Twitter*. University of Southampton. <https://eprints.soton.ac.uk/359848/1/DE-TechDemo-AutomatedProfiling-CameraReady.pdf>.
- Phethean, C., Tiropanis, T. and Harris, L. (2015) Engaging with Charities on Social Media: Comparing Interaction on Facebook and Twitter. *International Conference on Internet Science*. Berlin: Springer, pp. 15-29. <https://link.springer.com/content/pdf/10.1007%2F978-3-319-18609-2.pdf>.
- Prell, C. (2011) *Social network analysis: History, theory and methodology*. London: Sage.
- Quinton, S. and Fennemore, P. (2013) Missing a strategic marketing trick? The use of online social networks by UK charities. *International Journal of Nonprofit and Voluntary Sector Marketing*, 18 (1), pp. 36-51.
- Robins, G. and Daraganova, G. (2013) Social Selection, Dyadic Covariates and Geospatial Effects. In: D. Lusher et al., ed. *Exponential Random Graph Models for Social Networks*. New York: Cambridge University Press, pp. 91.
- Robins, G. and Lusher, D. (2013) What are exponential random graph models? In: D. Lusher et al., ed. *Exponential Random Graph Models for Social Networks*. New York: Cambridge University Press, pp. 16.
- Robins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007) An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29 (2), pp. 173-191.

- Robins, G., Pattison, P. and Wang, P. (2009) Closure, connectivity and degree distributions: Exponential random graph (p^*) models for directed social networks. *Social Networks*, 31 (2), pp. 105-117.
- Romm, T. (2018) *Facebook didn't read the terms and conditions for the app behind Cambridge Analytica*. https://www.washingtonpost.com/news/the-switch/wp/2018/04/26/facebook-didnt-read-the-terms-and-conditions-for-the-app-behind-cambridge-analytica/?noredirect=on&utm_term=.2083d96a1d04.
- Salamon, L.M. and Anheier, H.K. (1992) In search of the non-profit sector. I: The question of definitions. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 3 (2), pp. 125-151.
- Saxton, G.D. and Guo, C. (2014) Online stakeholder targeting and the acquisition of social media capital. *International Journal of Nonprofit and Voluntary Sector Marketing*, 19 (4), pp. 286-300.
- Schechter, S. and Bravo-Lillo, C. (2014) *Using Ethical-Response Surveys to Identify Sources of Disapproval and Concern with Facebook's Emotional Contagion Experiment and Other Controversial Studies*. Microsoft. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/CURRENT20DRAFT20-20Ethical-Response20Survey.pdf>.
- School of Applied Social Science (2014) *Ethics Handbook*. No longer available.
- Scott, J. (2012) *Social network analysis*. London: Sage.
- Scottish Council for Voluntary Organisations (2016) *An introduction to Scotland's national third sector umbrella bodies and intermediaries*. Edinburgh: SCVO Information Service. <https://scvo.org.uk/wp-content/uploads/2016/08/SCVO-Intermediaries-Report-Aug2016.pdf>.
- Scottish Government (2017) *Access Our Data*. <http://www.gov.scot/Topics/Statistics/About/DataAccess>.
- Shane, S. and Goel, V. (2017) *Fake Russian Facebook accounts bought \$100,000 in political ads*. The New York Times. <https://www.nytimes.com/2017/09/06/technology/facebook-russian-political-ads.html>.
- Shapiro, D.L., Sheppard, B.H. and Cheraskin, L. (1992) Business on a handshake. *Negotiation Journal*, 8 (4), pp. 365-377.
- Shields, R. (2016) Following the leader? Network models of "world-class" universities on Twitter. *Higher Education*, 71 (2), pp. 253-268.
- Smith, S. (2006) International Theory and European Integration. In: M. Kelstrup and M. Williams, eds. *International relations theory and the politics of European integration: power, security and community*. London: Routledge.
- Soteri-Proctor, A. (2011) *Little big societies: micro-mapping of organisations operating below the radar*. University of Birmingham. <https://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/working-papers/working-paper-71.pdf>.
- Soteri-Proctor, A. and Alcock, P. (2012) Micro-mapping: what lies beneath the third sector radar? *Voluntary Sector Review*, 3 (3), pp. 379-398.
- Steiner, A., Miranda, C. and Longhurst, E. (2015) *What's Data Got To Do With It*. JustGiving. <https://pages.justgiving.com/whats-data-got-to-do-with-it>.

- Svensson, P.G., Mahoney, T.Q. and Hambrick, M.E. (2015) Twitter as a communication tool for nonprofits: A study of sport-for-development organizations. *Nonprofit and Voluntary Sector Quarterly*, 44 (6), pp. 1086-1106.
- Teddlie, C. and Tashakkori, A. (2009) *Foundations of Mixed Methods Research*. London: Sage.
- The Comptroller and Auditor General (2009) *Building the Capacity of the Third Sector*. The Stationery Office. <https://www.nao.org.uk/wp-content/uploads/2009/02/0809132.pdf>.
- Thompson, B. (2004) *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington: American Psychological Association.
- Tolbert, C.J. and Mossberger, K. (2006) The Effects of E-Government on Trust and Confidence in Government. *Public Administration Review*, 66 (3), pp. 354-369.
- Data Protection Act 1998*. Act of UK Parliament. <http://www.legislation.gov.uk/ukpga/1998/29/part/IV>.
- University of Stirling (2015) *Information Services: Data Archiving*. <https://www.stir.ac.uk/about/faculties-and-services/information-services-and-library/current-students-and-staff/researchers/research-data/after-your-research/archiving/>.
- Van Duijn, M.A., Van Busschbach, J.T. and Snijders, T.A. (1999) Multilevel analysis of personal networks as dependent variables. *Social Networks*, 21 (2), pp. 187-210.
- Walliman, N. (2006) The nature of data. In: N. Walliman, ed. *Social Research Methods*. London: SAGE Publications, pp. 50-66.
- Walton, C. and Macmillan, R. (2014) *Working Paper 118: A brave new world for voluntary sector infrastructure? Vouchers, markets and demand led capacity building*. Birmingham: Third Sector Research Centre. <https://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/working-papers/working-paper-118.pdf>.
- Wang, P., Robins, G. and Pattison, P. (2009) *PNet: Program for the Simulation and Estimation of Exponential Random Graph (P*) Models*. Melbourne: University of Melbourne.
- Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge university press.
- Wasserman, S. and Pattison, P. (1996) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and P*. *Psychometrika*, 61 (3), pp. 401-425.
- Waters, R.D. and Jamal, J.Y. (2011) Tweet, tweet, tweet: A content analysis of nonprofit organizations' Twitter updates. *Public Relations Review*, 37 (3), pp. 321-324.
- Wellman, B. and Hampton, K. (1999) Living networked on and offline. *Contemporary Sociology*, 28 (6), pp. 648-654.
- Wellman, B., Quan-Haase, A., Boase, J. and Chen, W. (2002) Examining the Internet in everyday life. *Keynote Address to the Euricom Conference on E-Democracy, Nijmegen, Netherlands (October)*.
- Wellman, B. (2001) Computer networks as social networks. *Science (New York, N.Y.)*, 293 (5537), pp. 2031-2034.

Wells, P. and Dayson, C. (2010) *Measuring the impact of third sector infrastructure organisations: A report to the NCVO Funding Commission*. Centre for Regional Economic and Social Research.

Wendt, A. (1994) Collective identity formation and the international state. *American Political Science Review*, 88 (02), pp. 384-396.

Zhou, H. and Pan, Q. (2016) Information, community, and action on Sina-Weibo: How Chinese philanthropic NGOs use social media. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 27 (5), pp. 2433-2457.

APPENDICES

Appendix I. Interview topic guides

Topic guide for third sector organisations

Introduction

My name is Tom, thank you for agreeing to speak with me today. I am looking to explore the relationships and issues around charities' use of external data, the role of infrastructure organisations in these relationships, and the issue of trust. As a third sector organisation I am particularly interested in hearing about your use of external data and social media, how these are linked, and how issues of trust play into both. Thank you for agreeing to the recorder and we can stop or pause the interview/recorder at any time.

Part 1

- What is data?
- Which external data sources does your charity use (prompt: Scottish Government, specific sets)?
 - Why do you use these sets?
- What are the challenges to using or analysing data within your charity (prompt: cost, time, analytical skills)?
- What quality are the data sets you use? (prompt: trust)
 - What makes these data sets trustworthy/untrustworthy (prompt: provider)?

Part 2 (general prompt: prompt: network maps)

- Do you share external data or findings stemming from it? If so who do you share this data with (prompt: other charities, infrastructure, public)?
 - How do you share this data (prompt: twitter)?
 - Does sharing this data imply trust?
 - Do you verify the info first (prompt: do you have the capacity to verify, do you have a duty to verify)?
 - What about rebroadcasting for other organisations? (prompt: does this imply trust, do you verify it?)
- Who do you interact with to obtain the data you use in the first place (prompt: help from infrastructure, do you feed back issues, go to consultation events)?
 - How about analysing the data (prompt: help from infrastructure)?
- Do you have contact with organisations you speak to on twitter is in other ways (prompt: offline, email, face to face, events)?
 - What role or position does twitter have in your communication with organisations or the public?

Topic guide for infrastructure sector organisations

Introduction

My name is Tom, thank you for agreeing to speak with me today. I am looking to explore the relationships and issues around charities' use of external data, the role of infrastructure organisations in these relationships, and the issue of trust. As an infrastructure organisation I am particularly interested in hearing about your support for the data use of other charities, as well as

your own data use, and how these interactions affect or are affected by trust. Thank you for agreeing to the recorder and we can stop or pause the interview/recorder at any time.

Part 1

- What is data?
- Which external data sources does your organisation use (prompt: Scottish Government, specific sets)?
 - Why do you use these sets?
- What are the challenges to using data within your organisation (prompt: cost, time, analytical skills)?
- What quality are the data sets you use? (prompt: trust)
 - What makes these data sets trustworthy/untrustworthy (prompt: provider)?

Part 2 (general prompt: prompt: network maps)

- Do you share external data or findings stemming from it? If so who do you share this data with (prompt: charities, other infrastructure, public)?
 - How do you share (prompt: twitter)?
 - Does sharing this data imply trust?
 - Do you verify the info first (prompt: do you have the capacity to verify, do you have a duty to verify)?
 - What about rebroadcasting for other organisations? (prompt: does this imply trust, do you verify it?)
- Who do you interact with to obtain the data you use (prompt: help from other infrastructure, do you feed back issues, go to consultation events)?
 - How about analysing the data (prompt: help from other infrastructure)?
- Do you provide support to other organisations to obtain or analyse data?
- Do you have contact with organisations you speak to on twitter is in other ways (prompt: offline, email, face to face)?
 - What role or position does twitter have in your communication with organisations or the public?

Topic guide for Scottish Government respondents

Introduction

My name is Tom, thank you for agreeing to speak with me today. I am looking to explore the relationships and issues around charities' use of external data, the role of infrastructure organisations in these relationships, and the issue of trust. As the primary provider of external data for the sector and also acting as an infrastructure organisation, I'm very keen to get the views of the Scottish Government. Thank you for agreeing to the recorder and we can stop or pause the interview/recorder at any time.

Charity team

- What is your view of the third sector's capacity to use external data?
- What is the government's role in the Scottish third sector concerning data issues? (prompt: facilitating access or analysis)
 - Do you support charities using external data, particularly Scottish government data?
 - How do you support them (prompt: supply of data, analysis)?
- Is the issue of trust in the government or its data ever apparent?
 - Do you seek feedback from them on issues with the data?
- How do you communicate with charities (prompt: twitter)?
 - What sort of contact do you have with charities on twitter?
 - Does twitter replace or augment other forms of communication with charities?

Social media team

- Why does the government use twitter so much?
- The government has lots of contact with charities on twitter, what is the government's role in the Scottish third sector? (prompt: build networks? Rebroadcaster?)
- What sort of contact do you have with charities on twitter? (prompt: conversation, rebroadcasting)
- Is it mostly frontline charities you have contact with or infrastructure bodies (prompt: like the SCVO)?
- Does retweeting information imply trust in it?
- Do you check information before retweeting it?
 - Does the government have a duty to check what is in the tweets it is retweeting (prompt: stats, capacity to check/right links to other departments)?

Data team

- What is data?
- Are charities among your stakeholders?
- How do you support charities making more use of your data?
- How do you speak to your stakeholders/charities (prompt: twitter)?
 - Does social media replace or augment other forms of contact?
- Is there ever an issue with trust in your data?
 - What do you do to build or ensure trust?

- How do you get feedback from charities/stakeholders (prompt: consultation events)?

Appendix II. Interview information sheet for participants

Information sheet for charities and infrastructure organisations

Charity use of external data

Introduction

This project hopes to explore how charities use external data (particularly Scottish Government data), what impact trust has on this, and what role infrastructure organisations play in this relationship.

You have been asked to take part in this interview either because of a survey you completed earlier in the project or because of the organisation you represent. If you agree to take part you will meet with the researcher and have a discussion lasting around an hour. The discussion will be about your views, in whatever role you work in, on charities and external data. Participation in these interviews is entirely voluntary and you may withdraw at any time.

Ideally the interviews will take place face to face at a location convenient for you but they could be carried out over the phone or a program such as Skype or Hangouts if a face to face meeting is impractical.

What happens to the information I give?

With your permission, the researcher will record audio from your interview and may also take notes; just to make sure nothing is missed. This information will be treated in strict confidence and only accessed by the researcher, project supervisors, and potentially a professional transcription service. When writing the findings this information may be drawn upon or directly quoted but pseudonyms will be used to ensure no one can be personally identified. Your organisation will not be referred to by name in the published work to ensure anonymity and will only be discussed by general type (i.e. a member of staff at an infrastructure organisation said...).

Once the project is finished, if you give permission, this information will be lodged for reuse. If you chose to withdraw from the research your data will be destroyed.

Funding

Funding for this research has been provided in part by the Scottish Government (<http://www.gov.scot/>) and in part by the Economic and Social Research Council (<http://www.esrc.ac.uk/>) in collaboration.

The Scottish Government may wish to public some aspects of the final thesis this research relates to - but this will follow the same confidentiality guidelines as outlined above.

Contact information

The research will be solely carried out by Tom Wallace, a PhD student at the University of Stirling: tom.wallace@stir.ac.uk

The project is being carried out under the supervision of Dr Alasdair Rutherford, University of Stirling: alasdair.rutherford@stir.ac.uk | Phone: 01786 466409

Information sheet for government participants

Charity use of external data

Introduction

This project hopes to explore how charities use external data (particularly Scottish Government data), what impact trust has on this, and what role infrastructure organisations play in this relationship.

You have been asked to take part in this interview either because of a survey you completed earlier in the project or because of the organisation you represent. If you agree to take part you will meet with the researcher and have a discussion lasting around an hour. The discussion will be about your views, in whatever role you work in, on charities and external data. Participation in these interviews is entirely voluntary and you may withdraw at any time.

Ideally the interviews will take place face to face at a location convenient for you but they could be carried out over the phone or a program such as Skype or Hangouts if a face to face meeting is impractical.

What happens to the information I give?

With your permission, the researcher will record audio from your interview and may also take notes; just to make sure nothing is missed. This information will be treated in strict confidence and only accessed by the researcher, project supervisors, and potentially a professional transcription service. When writing the findings this information may be drawn upon or directly quoted but pseudonyms will be used to ensure no one can be personally identified. Your organisation will not be referred to by name in the published work to ensure anonymity and will only be discussed by general type (i.e. a member of staff at an infrastructure organisation said...).

Once the project is finished, if you give permission, this information will be lodged for reuse. If you chose to withdraw from the research your data will be destroyed.

Funding

Funding for this research has been provided in part by the Scottish Government (<http://www.gov.scot/>) and in part by the Economic and Social Research Council (<http://www.esrc.ac.uk/>) in collaboration.

The Scottish Government may wish to public some aspects of the final thesis this research relates to - but this will follow the same confidentiality guidelines as outlined above.

These interviews are not related to my internship in any capacity. I am speaking to you entirely on behalf of my PhD research.

Contact information

The research will be solely carried out by Tom Wallace, a PhD student at the University of Stirling: tom.wallace@stir.ac.uk

The project is being carried out under the supervision of Dr Alasdair Rutherford, University of Stirling: alasdair.rutherford@stir.ac.uk | Phone: 01786 466409

Appendix III. Interview consent form

Consent sheet

Charity use of external data

Contact information

The research will be solely carried out by Tom Wallace, a PhD student at the University of Stirling: tom.wallace@stir.ac.uk

The project is being carried out under the supervision of Dr Alasdair Rutherford, University of Stirling: alasdair.rutherford@stir.ac.uk | Phone: 01786 466409

	Tick
I confirm that I understand the information provided about the research project. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily	
I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason.	
I understand that interviews will be audio recorded (and transcribed), but that these will not contain my name or any other identifiable information. I give permission for interviews to be recorded.	

I understand that my data will be kept confidential and stored securely under password, not disclosed to the third parties without my prior consent and used exclusively for the purposes of this project.	
I understand that some statements I give may be included in a report on the study and publications originating from it, but I will have the opportunity to see drafts of these.	
I agree that my data can be lodged and potentially used for other projects (optional)	
I agree to take part in the above study.	

Participant name:

Signature:

Date:

Researcher name:

Signature:

Date:

Appendix IV. Ethics form

Project Protocol

Principle Investigator: Tom Wallace

Postgraduate research student in Sociology and Social Policy

University of Stirling

September 2015

Project funding: ESRC and Scottish Government collaborative

Full Review Coversheet

Project Title	Trust in government data: Users and Stakeholders
Date of Review Meeting	17/12/15
Duration of Project (Estimate if unknown)	Oct 2014-Oct 2017
Funding Source	ESRC & Scottish Government collaborative

The following papers are attached (tick box or add comment if item is not included):

1	Research protocol	Please indicate where the following items may be found in the protocol (see ESRC REF 1.8.1)	
		Tick	Comment
Aims		✓	
Scientific background		✓	
Study design		✓	
Participants (who, how many, identification and recruitment, comment on vulnerable groups)		✓	
Methods of data collection		✓	
Methods of data analysis		✓	
Response to any conditions of use set by secondary data providers		N/A	No secondary data
Principal investigator's summary of potential ethical issues and how they will be addressed		✓	
Benefits to research participants or third parties		✓	
		Tick	Comment

Risks to researchers Please indicate if the risk assessment (Appendix 1) form has been completed		✓	
Procedures for informed consent (including information provided and methods of documenting initial and continuing consent)		✓	
Expected outcomes, impacts and benefits of research		✓	
Dissemination (and feedback to participants where appropriate)		✓	
Measures taken to ensure confidentiality, privacy and data protection		✓	
			Tick
2	Information sheets for participants		✓
3	Questionnaire, topic guides or other research instruments		✓
4	Comment on any need for further scrutiny at a later date e.g. where the research design is emergent		✓
5	Comment on any involvement of external contractors, and their compliance with ethical requirements (e.g. transcription services, interpreters, fieldworkers)		✓
6	Any other relevant material (please indicate below what it is) Survey preamble and afterword		

--	--

Signature of Principal Investigator

..... TOM WALLACE Date: 09/12/15.....

Name of Principal Investigator

Tom Wallace..... Date:
09/12/15.....

Risk Assessment

School of Applied Social Science, University of Stirling

All those doing the work must be involved in the completion of this form. Complete all sections, marking clearly those that are not applicable. The form must be signed by all involved, and copies made for each person. Hard copies of the completed form, with original signatures, must be sent by the principal investigator to the School Administrator within 3 months of the start date of the project, or prior to the commencement of fieldwork, whichever is the sooner.

Head of School	<i>Professor Alison Bowes</i>
School Administrator	<i>Mrs. Morag Crawford</i>
University Safety Advisor	<i>Mr. David Duckett</i>
Completed by	<i>Tom Wallace</i>
Date	

Contact in Emergency, name & telephone number	<i>Tom Wallace, 01786842239 or 07717565642</i>
----------------------------------------------------------	------------------------------------------------

Research Activity

Dates of activity:	Winter 2015 – Oct 2017
Activity: Give title and briefly summarise	Online survey followed by qualitative interviews
People involved: Give individual name(s)	Tom Wallace
Location(s) of the activity: Give specific locations, e.g. name of hospital, or town	To be confirmed at the start of the interview process. Most likely at the university or interviewees places of business but all will be within Scotland.

<p>Working in a dangerous area: e.g. high crime area, area of civil unrest. Give contact details and measures in case of emergency</p>			<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>
<p>Working in an isolated geographical area: Give contact details and measures in case of emergency</p>	<p>Interviews could take place in a rural location depending on the interviewees place of work and the logistics of meeting.</p>	<p>Rural interviewees will be invited to meet at their place of work, the university or a neutral ground in a populated area just like non-rural participants. Should all of these options prove impractical telephone or Skype contact will be employed as a backup option.</p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>
<p>Lone working: Give contact details and measures in case of emergency</p>	<p>Certain – research will be conducted by the principal investigator only</p>	<p>The researcher will always carry a mobile phone and will inform friends/family of his location when performing interviews</p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input checked="" type="checkbox"/></p>
<p>Working with Equipment: Risks associated</p>	<p>N/A</p>	<p>N/A</p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input type="checkbox"/></p>	<p>High <input type="checkbox"/></p> <p>Med <input type="checkbox"/></p> <p>Low <input type="checkbox"/></p>

Environmental hazards: e.g. weather, terrain, animals, plants, earthquake, water quality	N/A	N/A	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>
Chemical & Biological Hazards: e.g. laboratory chemicals, crop spraying, diseases	N/A	N/A	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>
Manual Handling: e.g. loading and unloading equipment	N/A	N/A	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input type="checkbox"/>
Emotional Risks: e.g. Sensitive research	Unlikely	The research does not target vulnerable individuals and does not cover any issues which could be considered personally sensitive	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input checked="" type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input checked="" type="checkbox"/>	High <input type="checkbox"/> Med <input type="checkbox"/> Low <input checked="" type="checkbox"/>

Legal compliance: Are there any specific standards relevant to the research activities?	Data protection, confidentiality – Data protection act
------------------------------------------------------------------------------------------------	--------------------------------------------------------

Training: Has special training been given for fieldwork activities in relation to safety?	Qualitative training to be confirmed at a later date
Supervision: What level of supervision is required, and are there sufficient supervisors for research?	Both primary and secondary supervisors are fully engaged with the project and provide a good level of supervision
Medical conditions/allergies: This information is to be kept confidential.	N/A
First Aid: Will a First Aid box be available? If research involves a group, name the First Aider(s)	In most places of business and public locations
Disabled persons: Detail any special arrangements required	N/A
Insurance: Are all activities covered by University insurance? Provide confirmation that this has been checked and approved. Give details of any additional personal insurance.	

Risk assessment: Overall **LOW** **MEDIUM** **HIGH**

Safe system of work procedure (to be completed by research team on basis of above information. Continue on separate sheet if necessary)

Carryout interviews in safe locations only

Primary researcher will inform his family of the location and times of interviews and will carry a mobile phone

Supervision team will be kept up to date with progress of interviews and notified of any relevant issues

Date:.....

Agreed date for review:...

17/12/15.....

Signed	Full Name	Role in Work
--------	-----------	--------------

1. 	...Tom Wallace.....	Principal Investigator
---------------------------------------------------------------------------------------------	---------------------	------------------------

Summary of research

Scottish government data resources are utilized by many different types of organisations. Academics, charities, non-governmental bodies, local authorities, and public sector institutions all make use of data which the government commissions, collates, and publishes. These users and consumers interact with government data in various ways, and to varying degrees, but what influence does that interaction have on building trust in government data? What are the users' capabilities to do so? Do they trust government data themselves? What do the different networks of stakeholders look like?

This project seeks to answer those questions, among others, through a mix of methods. Literature has already been collected and reviewed on data policy, data laws, the theory of trust, and other key concepts. The first stage of the research will use a survey to analyse the capabilities of users and stakeholders to interact with data resources and perhaps contribute to building trust in them. Analysis of the first stage will comprise quantitative methods and also social network analysis (SNA) to map networks of data users. The second stage will expand upon the first with a set of semi-structured interviews which will help tease out more complicated questions, such as the users own trust in government data. The primary aim of the project is to help the Scottish Government improve the quality of its data resources and its links within the network of users. This will be achieved by revealing the dynamic of the interplay between stakeholders and data and the factors which inhibit trust and factors which contribute to it in this relationship. A secondary aim is to map and compare several networks of users of government data. It is hypothesised at this point that users' critical engagement with data will add value to that data.

Research Questions

The project is based on a number of research questions which will be answered by the various methods discussed later.

	Social Network Analysis	Statistical Analysis	Qualitative Analysis
1. Who are the users and consumers of government data?	X	X	
2. What are the different capacities of users to critically engage with Government data?	X	X	
3. What are the barriers to further engagement and how might these be overcome?		X	
4. To what extent do consumers differentiate, in terms of trust and usage, between government data and		X	

administrative data that may be used for research?			
5. Is there a lack of trust in Government data amongst users of data?			X
6. What are the factors that most influence data users' trust in the Government's data and statistics?			X
7. How do the users of research data weigh up the value of Government data as compared to alternative primary or secondary research evidence?			X
8. How does the balance of individual privacy against making data available for independent analysis affect stakeholders' trust in government data?			X

These questions will be investigated through an explanatory sequential mixed methods design; the first methods will be expanded upon and explained further by the second. The first method comprises a combination of statistical analysis and social network analysis (SNA). The information for both of these methods will be collected on the same survey instrument but analysed mostly separately. There will also be a twitter analysis using SNA methods which will expand on the findings of the main SNA analysis and reach users which the survey may not have. The second main method is a series of semi-structured interviews following from the responses and insights of the first method. The study makes use of a social constructivist philosophy which allows for the use and analysis of both quantitative and qualitative data.

Literature

A literature review has been collected and written to support the background of this study. It includes an examination of topics such as the definition and nature of trust, the role legislation and laws have on data, and how the concept of trust interacts with data. It also extensively details the research philosophy.

Participants

Participants for the study are people working in professional roles involving data, including; academics, third sector personnel and, public sector workers. The first stage of the research which involves direct contact with participants is the survey instrument. This survey will be primarily targeted. Potential respondents will be selected from pre-existing knowledge of who uses data, publically available information on who uses data, and networks bringing together users of data, such as government working groups and email lists. The survey will also include a snowball element where respondents are asked to suggest peers who they think might be suitable for the research and a social networking element (detailed later) which should also provide more potential respondents. Both the snowball and SNA questions will be researcher driven so that all distribution

of the survey is done by the researcher and respondents are not encouraged to spread the survey among themselves. This allows the researcher to collect more information on the spread of the survey and also avoids issues relating to inter-respondent relationships as the nominator will be made anonymous by the researcher. It is impossible to put a number on how many surveys will be required to be distributed and with the snowball component the sampling frame will be fluid.

The secondary SNA analysis uses public admin data from twitter to attempt to map the end consumers of government data. The data will be gathered through the NodeXL software and will largely comprise the twitter accounts of organisations.

The interview stage uses participants from the survey stage who self-select into the interview process via a question on the survey. Not all participants who select-in will necessarily be interviewed. This method of selection is non-random but it should ensure that interview participants are those who genuinely want to be part of the study. Additionally, the researcher will select a mix of respondents based on answers from the survey and the findings of the SNA with the potential to further explore emerging themes. Should not enough respondents select in, or a different type be needed, the researcher may contact organisations and individuals directly to seek an interview. The non-random nature of these selection methods will be kept in mind when writing up the findings of the interviews.

Stage 1: Survey & Social Networking Analysis

Methods of collection

The survey instrument is the primary method of collection for this stage. It comprises the questionnaire which asks participants about their interaction with with Scottish Government data and the social networking questions which will be used to build a picture of the network of data users. Non-academic participants will act on behalf of their respective organisations and will not be analysed at the individual level. Academic participants will be analysed through their professional roles. While this is at the level of the individual they will not be personally identifiable.

Organisation names will be published – individual level names and information will not. The survey will be hosted online via Bristol Online Survey and will be distributed primarily through email. The questionnaire will comprise different questions depending on the target respondent, for example, academics will be asked which institution they work within whereas charity workers will instead get a question about what field their charity works within. This will be achieved with a series of initial questions to identify the type of respondent and then routing options will give the respondent the correct set of questions. Different groups and networks of respondents will receive different surveys; one relates to charities data, and one to migration data. These surveys are almost identical but refer to different government data sets and are kept separate for simplicity and clarity (see questionnaires). The survey will be distributed in multiple waves. The initial wave will be

targeted at known organisations and distributed through relevant mailing lists. Once this initial wave is returned the SNA and snowball questions will be used to contact further participants in a second wave. This will continue as long as is necessary, feasible, and fruitful. As previously discussed this process will be researcher controlled and directed so that participants are not encouraged to contact each other. To aid in this process two copies of each survey will be created which are identical in all respects apart from the internal survey name and the survey ID, which are only visible to the researcher. One version will be sent to respondents directly while the other will be published on lists and through indirect dissemination. This will allow a check to be performed on the source of a particular response; unsolicited responses to the directly disseminated survey mean a respondent has passed the survey on independently without the researcher control.

Methods of analysis

The survey will be analysed with two primary methods. The results of the questionnaire will be investigated quantitatively with data analysis software (Stata). This will primarily consist of multilevel regression modelling using different types of users for the levels. The primary dependant variable for this analysis will be the level of use of government data, primarily measured through a self-selected set of archetypes (trail blazer, utilizer, coping, struggling). Independent variables in this analysis will include a measure of the capabilities of data users, an index of inhibiting factors to further use of government data, a measure of centrality from the SNA analysis, and other more minor questions from the survey. The questionnaire includes several other measures of government data usage, besides the archetype question, which will be used to check internal validity (Neumayer 2002). Other quantitative methods include bivariate correlation and significance testing, and univariate distributions. Visualisations will be used where appropriate. The precise specification of the quantitative analysis will depend on the data received from the survey.

The second primary method for analysing the survey is social network analysis using the software Pajek, and, potentially UCINET. Prior to this analysis organisations will be linked by their responses to the name generator questions on the survey. This social networking study is using whole or complete networks (Knoke and Yang 2008). This does not mean that every possible actor is reached but that the network consist of multiple ties between various actors; there are no egos and alters. The purpose of this type of network is to determine the overall dynamics within and among all actors. Besides mapping the overall shape and dynamic of the network of data users, findings from the SNA study may be fed back into the quantitative analysis as explanatory factors for a stakeholders' use of government data. This feedback could reveal that being more closely associated with the government in the network leads to higher capabilities, or it could show that working groups are key to high capability stakeholders. Alternatively it could find that network position is irrelevant, which would be an important discovery alone.

Stage 2: Interviews

Methods of collection

Stage two comprises a series of semi-structured interviews; most likely between 20 and 25. As previously stated, respondents for this stage will be self-selected from the previous stage. These interviews will comprise a deeper investigation of issues raised initially in the survey and will be carried out after the initial results of the survey are known so that these findings can be expanded upon, contextualised, and explained. The interviews will also cover topics which were not appropriate to ask during the previous stage, such as the respondents trust in government data which is subjective and not easily captured by a survey. A topic guide for the interviews will be written once the initial results from the survey are analysed (this will be submitted for ethical clearance in due course) but the semi-structured nature of the survey allows for deviation from this guide where it is deemed fruitful. The surveys will be audio recorded, provided informed consent is granted, and these recordings will then be transcribed for analysis.

Methods of analysis

The interviews will be transcribed mainly by the principal investigator but depending on the final number conducted and time constraints of the project it may be necessary to have some externally transcribed by a professional service. Once the interviews are transcribed they will be coded and analysed through the use of NVivo. The analysis and findings of the interviews will then be considered alongside the findings from the previous survey stage during write up. This should result in a mix of methods and analysis which help explore the research questions fully and from multiple angles.

Ethical issues

Informed Consent

Participants in all stages of the research will have the capacity and information to give informed consent in line with the ESRC framework for research ethics principle 2 (ESRC 2015). For the survey stage this will be achieved by a preamble at the beginning of the survey detailing the purpose of the research, data protection procedures including logging data with the archive, voluntary participation, funding details and the contact details of the research team and head of school in line with SASS requirements (School of Applied Social Science 2014). The participant agrees to these stipulations with an opt-in check box and the survey cannot proceed without this box being checked, which is generally considered good practice for online surveys (Couper 2000). The details of funding are to ensure compliance with ESRC Ethics Principle 6 (ESRC 2015) and the SASS Ethics Handbook (2014) in which all potential conflicts of interest must be disclosed. This preamble can be found in appendix A.

The survey also makes use of an afterword which is shown as the last page of the survey. This document is similar to the preamble but is more concise and focuses only on procedures relating to the information participants have provided, issues of confidentiality, and contact details. This can be found in appendix B.

The survey can be stopped at any time and no data will be captured. Data is not stored until a final confirmation is received following the afterword. Respondents also have the right to not answer particular questions (ESRC 2015). A few questions on the survey require a response for technical reasons related to the routing options used but they will always include a ‘don’t know’ option, as with all other questions, to accommodate this requirement.

The interview stage makes use of a consent form which respondents will be presented with and asked to sign before the interview commences. This sheet covers similar topics to the qualitative form but also includes provision for the researcher to make audio recordings of the interview. Participants will be given a fair amount of time to read this form and the interview will not proceed without their signature. This document also makes clear that the interviewee may stop the interview at any time and may request that all data collected on them up to that point is destroyed. This document can be found in appendix C.

As the interviews will follow on from the findings of the survey stage, no discussion guide has been written yet. This document will be produced when the results of the survey are known and will be sent to the ethics committee for clearance. Additionally, any other ethical concerns which arise during the study will be referred back to the committee for guidance, as per ESRC principle 3 (ESRC 2015).

Participation in all stages of the research is entirely voluntary in line with the ESRC framework for research ethics principle 4 (ESRC 2015) and this is clearly stated in both the survey and interview preambles.

Anonymity and confidentiality

Conforming to ESRC ethics principle 3 (ESRC 2015), the anonymity and confidentiality of participants will be guaranteed and will be ensured via several methods. Participants’ names will be recorded in both stages of the research; this is primarily to allow the survey stage to link properly to the interview stage in the analysis and for record keeping. This information will not be published in any form. Published results will make use of public organisations’ names, pseudonyms, codes, or aggregate level data. This will help the research conform to ESRC principle 5 (ESRC 2015) which states that research must avoid harm to participants, and researchers, in all instances.

An anonymity issue to note concerning the SNA method is that it and the other quantitative data collection occur on the same survey instrument. They do not have much cross over in terms of analysis however. The social networking component will only make use of the basic demographic information (name of organisation, ect) from the survey and not use any of the actual questions. For example, the SNA analysis may show charity A linked to charity B but it will not reveal what those respondents answered on the capabilities or barriers sections of the survey. This means that the SNA study will be non-disclosive in relation to the rest of the survey. The quantitative analysis of the survey may use more detailed results from the SNA study but the result of that analysis will be at the aggregate level and non disclosive.

Furthermore the SNA study will use data which is public or could be obtained through a freedom of information request; the survey is just a better way to consistently collect this information. This should mean that even if the SNA study is potentially disclosive in itself, the information revealed should be publically available through other sources and non-sensitive (ESRC 2015). For example it may reveal that a charity is sourcing most of its data from another third sector organisation rather than direct from the government; this information is not sensitive, commercially or personally, and may be available at several other sources; the charities website, the data providers website, the funding agent, or publications of any of the above. Despite this added layer of protection, the SNA study will make the same efforts to avoid disclosure as all other aspects of the study. There are several examples in the literature of SNA being used successfully and non-disclosively alongside other quantitative methods, including multilevel modelling (Contractor et al 2006; Van Duijn et al 1999).

A second issue concerning SNA and anonymity is related to the relational sampling method, where respondents pass on information about other potential respondents to the researcher. The researcher then contacts these potential respondents (Scott 2012). This project will perform anonymisation so respondents who are recommended by others for contact will not know who has recommended them. This should help eliminate bias if the two respondents have a particular dynamic to their relationship; for example if a boss suggests their employees, those employees may get different responses if they know their boss is involved.

A final ethical concern related to the SNA method and anonymity is the possibility of a respondent refusing to give consent to be part of the project and what effect this would have on the relationships they have with other actors. The respondents own information will be removed but what if another actor then nominates them in a relationship? This issue relates to who ultimately 'owns' relationships but a sensible and ethical way to deal with this would be to include the actor who has withdrawn consent only based on information given by other actors as if they had never

been contacted. This is the most common solution employed by other researchers using social networking analysis (Borgatti 2013; Prell 2011).

During the interviews, the researcher will endeavour to provide a safe space where a confidential conversation can take place.

Data protection

All information collected or created during the research will be stored on secure password protected university network space. Transcripts and audio files from interviews will also be stored in this way with audio being transferred into secure space as soon as possible after recording. The one exception to this is the results from the survey which will remain hosted online with Bristol Online Survey in a password protected, university associated, account.

In line with data protection law, all data generated during this research will be held as long as required for any publications resulting from the research to be published (Data Protection Act 1998). However, also conforming to University of Stirling's records management policy this period will be at least 10 years from the cessation of funding in 2017 (University of Stirling 2015). After this period all data will be securely destroyed except for a fully anonymised version of any quantitative data which will have been lodged with the UK Data Archive within 3 months of the end of the project in 2017 in line with the ESRC Research Data Policy (ESRC 2014). Also in line with this policy, the anonymised data will be provided with high quality meta-data which will allow it to be used for further research.

Much of the data collected during the first stage of the project, particularly that used for the Social Network Analysis, is not considered 'human data' by the ESRC. It is also not classified as 'personal data' by the DPA (Data Protection Act 1998). The vast majority of this data is available online or elsewhere in the public domain and is not considered sensitive (ESRC 2015). Despite this, this data will be protected and stored as if it were sensitive human data. This also applies to the secondary analysis of twitter which only uses public admin data but will be treated the same as all other data in the project.

The project may use the services of a transcriber after the interview stage. The aim is for the principal researcher to transcribe all interviews but should this prove impractical a professional service may be used. In this case the transcriber will not receive any contextual information such as the participants name and will not be able to identify individuals from the data.

A copy of all the participants of the research will be kept separate from the research data for data protection purposes (Data Protection Act 1998).

Researcher safety & Risks

The survey stage does not involve any direct, face to face, contact but the interview stage will involve the researcher traveling to meet participants. The location for these interviews has not been confirmed and will most likely be arranged ad hoc with each respondent. Common locations are likely to comprise the participant's place of work, the university, or a neutral public place such as a coffee shop or café. The researcher will endeavour to provide a safe and private space. The researcher will carry a mobile phone and make family members aware of his location and intended time of return.

There should be no foreseeable or avoidable risks to respondents at any stage of the research beyond data protection issues already discussed.

Dissemination

The government has specified that a dissemination workshop be conducted, by the lead researcher, at the end of the project. The purpose of this workshop is to concisely feedback the main findings of the research to the government statistics teams. An executive summary of the findings and major outcomes of the research will be written at the end of the process and presented in non-academic language. It is also hoped that the findings of the research will be rewritten into one, or several, journal articles to be submitted to peer reviewed academic journals.

Reflective Research

Throughout this project the researcher will maintain an awareness of their place within the research process. Particular care will be maintained in any contact with the government or participants which is not directly related to the research. This situation may arise, for example, during an internship. In this case the researcher will record no data and will take particular care to not record anything shared in confidence. The researcher will also maintain an awareness of potential conflicts of interest when conducting a part-government funded project concerning a product of the Scottish Government.

Appendix V. Clarification of ethics form after review by ethics committee

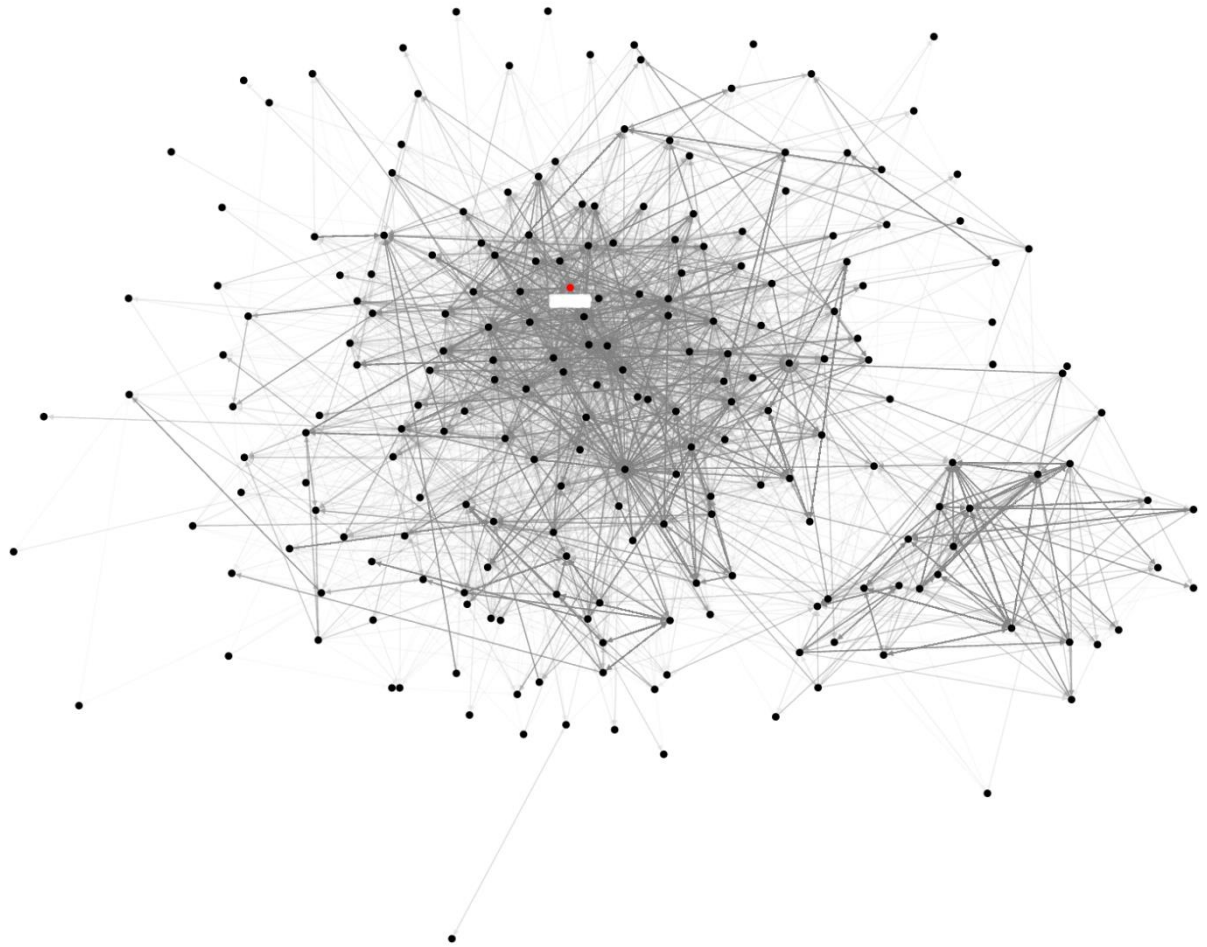
pp.	Old text	New text
5	Rural interviewees will be invited to meet at the university or a neutral ground in a populated area.	Rural interviewees will be invited to meet at their place of work, the university or a neutral ground in a populated area just like non-rural participants. Should all of these

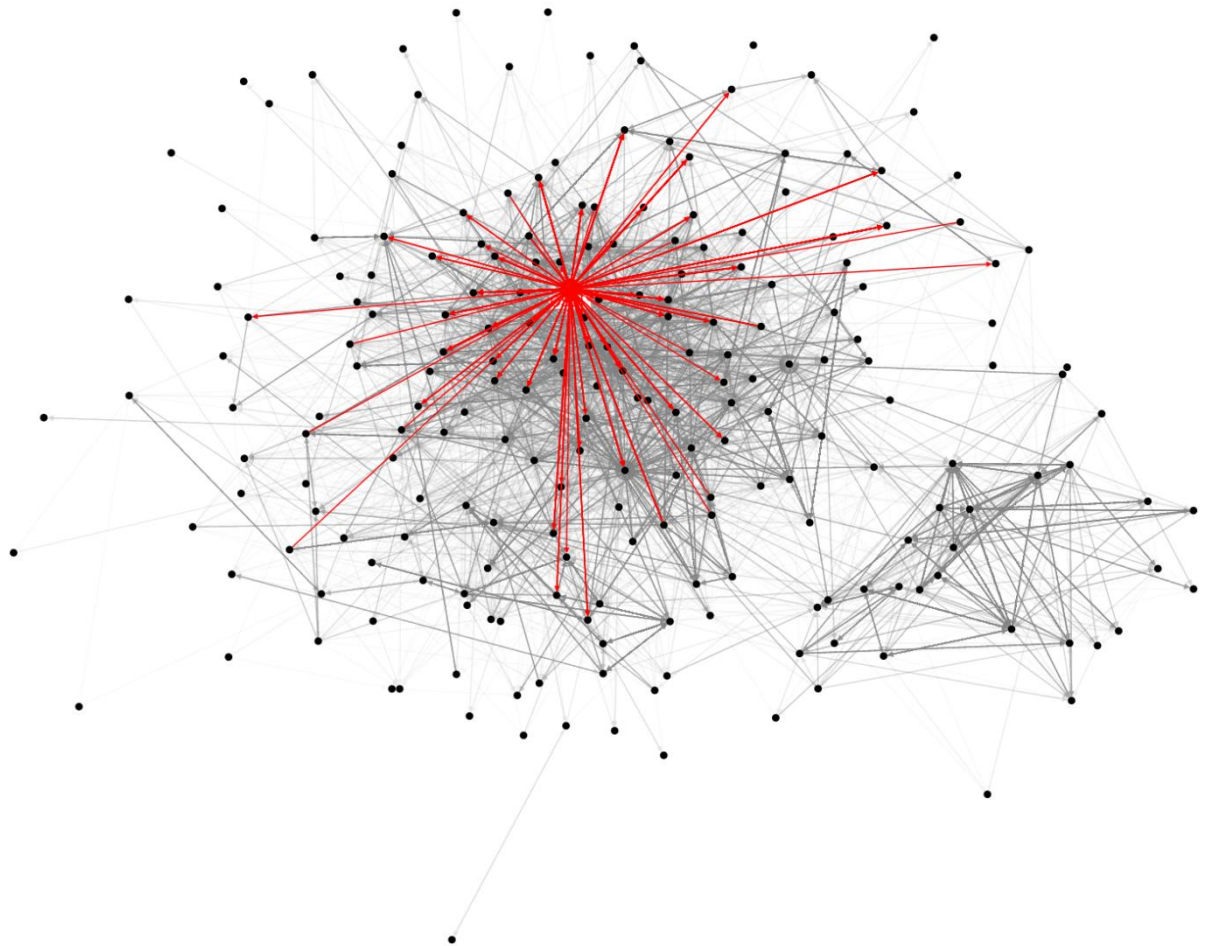
		options prove impractical telephone or Skype contact will be employed as a backup option.
10	-	Added: Non-academic participants will act on behalf of their respective organisations and will not be analysed at the individual level. Academic participants will be analysed through their professional roles. While this is at the level of the individual they will not be personally identifiable. Organisation names will be published – individual level names and information will not.
11	Removed: The secondary SNA analysis will collect public data live from twitter.	-
11		Added: To aid in this process two copies of each survey will be created which are identical in all respects apart from the internal survey name and the survey ID, which are only visible to the researcher. One version will be sent to respondents directly while the other will be published on lists and through indirect dissemination. This will allow a check to be performed on the source of a particular response; unsolicited responses to the directly disseminated survey mean a respondent has passed the survey on independently without the researcher control.
11	-	Added: Prior to this analysis organisations will be linked by their responses to the name generator questions on the survey.
	Removed: A secondary method of this stage is the use of Twitter admin	--

11	data collected and analysed with the software NodeXL. This method will produce a large scale, low detail, network map which primarily comprises that end consumers of data; a group which is not well captured by the primary survey instrument.	
20	Ideally the interviews will take place face to face at a location convenient for you but they could be carried out over the phone or a videotelephony program such as Skype or Hangouts if a face to face meeting was impractical.	Ideally the interviews will take place face to face at a location convenient for you but they could be carried out over the phone or a videotelephony program such as Skype or Hangouts if a face to face meeting was impractical – this is a backup option.
Survey: Immigration: Preamble	Your responses will help us better understand how researchers interact with data sources	Your responses will help us better understand how researchers interact with immigration and migration data sources
Survey: 3 rd sector: Preamble	Your responses will help us better understand how researchers interact with data sources	Your responses will help us better understand how researchers interact with 3 rd sector data sources
Surveys: Preamble	The survey comprises 3 sections as below and should take ~10 minutes to complete; <ul style="list-style-type: none"> • Where you source data from and how data is shared between you and other organisations • Your/your organisation’s capacities to engage with government data • Difficulties in engaging with this data 	The survey comprises 3 sections and should take ~10 minutes to complete; <ul style="list-style-type: none"> • Your/your organisation’s capacities to engage with government data • Difficulties in engaging with this data • Where you source data from and how data is shared between you and other organisations
Surveys: 4		Question 3 optionality changed to required
Surveys: 5 & 8		Question 12 and 13 & 26 and 27 swapped order
Surveys: 5 & 8	Which of the following data sources does your organisation make use of (in any capacity). Select as many as	Which of the following data sources does your organisation make use of (in any capacity). Select as many as apply.

	apply.	'Sources' means organisations, websites or other locations where you obtain data.
Surveys: 5 & 8	Which of the following data sets does your organisation use (in any capacity and from any source). Select as many as apply.	Which of the following data sets does your organisation use (in any capacity and from any source). Select as many as apply. 'Sets' means data files or resources
Surveys: 5 & 8	-	How often to you use external data? <ul style="list-style-type: none"> • Continuously • On a monthly basis • On a yearly basis • Less often than once a year • Don't know
Surveys: 7	Removed: question 20	
Survey: Immigration: 8	<ul style="list-style-type: none"> • The third sector itself • Immigration/refugees 	<ul style="list-style-type: none"> • Immigration/refugees • The third sector itself
Surveys: 10	For each data set and source you said you or your organisation uses (or any others not mentioned) could you provide a link to where you get the data from?	For each data set and source you said you or your organisation uses (or any others not mentioned) could you name the data and, if possible, provide a link to where you get the data from?
Surveys: 11	What is your email address? (if you don't have one leave blank)	What is your email address? (your work or organisational email is preferred)
Surveys: 11	-	Question 34 d optionality changed to required
Surveys: 11	If you selected yes could you provide an email address to contact you even if it is the same as the one you previously supplied)	If you selected yes could you provide an email address to contact you (if it is different from the one you previously supplied)
Surveys: 12	Removed: The information you have provided will remain confidential; the survey responses will not be shared with any other parties.	
Surveys: Throughout	Minor spelling corrections	Minor spelling corrections

Appendix VI. Network diagrams for interview participants (anonymised example)





Appendix VII. Full network group connections table

		Receiving			Out total
		1. Charities	2. Support	3. Data	
Sending	1. Charities	1139 (42%)	1522 (56%)	45 (2%)	2706
	2. Support	1933 (29%)	4748 (70%)	73 (1%)	6754
	3. Data	103 (5%)	586 (27%)	1462 (68%)	2151
		3175	6856	1580	
		In total			

Cell values are the number of tweets sent between given pairings of groups, parenthesis values are the percentage of outgoing ties to each recipient. Light shading is contact internal to the groups. Dark shading is contact between data and support organisations.

Appendix VIII. Link to GitHub

Further supplementary materials, syntax files, and data can be found on the author's GitHub:

https://github.com/tomwallace1990/charity_data_PhD