Lund University Master of Science in
International Development and Management
June 2013

Replicating Data Mining Techniques for Development: A Case Study of Corruption

Author: James Ransom
Supervisor: Martin Andersson

Abstract

Data Mining has a reputation in social science for lacking statistical rigour. This study challenges this reputation and argues that, whilst such a method (as with any other) can be abused, it has particular promise as a tool to be used for monitoring and explorative research, especially by smaller development organisations. Drawing on recent advances in adapting commercial 'Big Data' techniques for use in international development, this study uses an example data set of global news reports to measure the level of discussion about corruption using a Text Mining methodology. The methodology outlined holds particular promise for tracking the dissemination of ideas and concepts, although it is heavily dependent on contextual interpretation and the quality of the data set used.

Key words: Data Mining, Big Data, Corruption, Bribery, Text Mining, Civil Society Organisations, Broadcast Media, Data Analysis

Word Count - 14,401

Acknowledgements

The idea behind this thesis was first born from discussions held in Vietnam in 2012 as part of my fieldwork. I would like to thank Sun Lei at UNESCO, Stephanie Chow at Transparency International and Jairo Acuna-Alfaro at UNDP for sharing their time, insights and knowledge, and inspiring me in my work.

Warm thanks to my supervisor, Martin Andersson, and my esteemed peer reviewers – Amadeus, Miriam and Szilvia – for their support and constructive criticism. And many thanks to Lisa Eklund, Sahar Valizadeh and Marie Wiman of the LUMID programme for their support and encouragement along the way. A special thanks also to Iina and my family for being consistently fantastic throughout.

This was my first foray into open-source data analysis software. This work would not have been possible without the team behind RapidMiner and the countless other individuals who contribute, document, and promote open-source software alternatives. Whilst the individuals dedicating hour after hour to such projects are often anonymous, their commitment to openness, innovation and accessibility – and the collective dynamism and strength of community that results – is slowly changing the world.

Contents

Introduction

Data Mining is the subject of considerable controversy in the field of social science. Deemed by some as the 'wild west' of data analysis, Data Mining is the process of looking for interesting patterns and information in a data set (Miner *et al* 2012: 30). Data Mining has been derided for being ineffective, producing misleading results, and, when applied in real-life, leading to serious confidentiality and privacy issues. In *World Development*, Howard White (2002: 511-13) writes that data miners know what they are looking for and simply search until they find it, and that is the story they tell. He argues that Data Mining is mining for the 'right result'; the role of researchers should be to look for patterns in the data, rather than force the data into a 'preconceived view of the world'. This study challenges the reputation of Data Mining and argues that, whilst such a method (as with any other) can be abused, it has particular promise as a tool to be used for monitoring and explorative research, especially by smaller development organisations.

Recent increases in technological capacity have introduced a new statistical tool for development: Big Data - the analysis of massive data sets. Much has been made of Big Data and the tremendous potential for yielding new insights into traditional development issues (Letouzé 2012: 4). However, limitations in technological infrastructure and technical expertise currently restrict Big Data to big organisations – a 'digital divide' exists with the vast majority of Big Data hardware existing in high income countries (Hilbert 2013: 17). Whilst Data Mining is no direct substitute for Big Data, it can be used for similar analyses on a smaller scale. The data set is necessarily smaller, but the hardware requirements are minimal and the software is open source. Formerly the preserve of online businesses, Big Data techniques have successfully shifted to the development sector. A similar shift for Data Mining would enable small organisations access to the kind of novel insights that Big Data currently affords.

One example application of Data Mining as a monitoring tool is explored in this study. Using a body of news and online media reports from the past five years, the levels of discussion about corruption are measured in 215 countries. The resulting Broadcast Discussion Index allows us to track how prominently corruption is featured in public discussion over time, and tangential analyses allow insight into the nature and form of discussion about corruption. It is envisaged that such a tool could be used by Civil Society Organisations as a complement to existing measures (such as Transparency International's Corruption Perceptions Index) to devise an appropriate, context- and country-specific strategy to fight corruption and increase transparency. However, the emphasis in

this study is on a transferable methodology that could be easily utilised by other development organisations.

This study is organised in six parts. First, the emergence of data mining is placed within the statistical narrative, in particular its divergence from 'traditional' statistics. Big Data emerges later in this narrative, and the role that massive data sets play in development research is briefly examined together with parallels between Big Data and Data Mining. Second, corruption as an example of an international development issue is explored, notably as a phenomenon that is notoriously difficult to measure. The Broadcast Discussion Index is introduced, and the relevance of such a measure for development organisations is discussed, together with existing measures of corruption. The third part sets out the methodology, including the *corpus* (collection of documents) to be analysed. This study employs one particular strand of Data Mining – *Text Mining* – and specifically *Information Extraction*; the methodology therefore has a linguistic leaning. In part four, the methodology is tested on the corpus for several other current development concerns – diseases, harmful drugs, and arms trafficking – and part five presents and discusses the results for corruption. Part six concludes, with an emphasis on directions the Data Mining methodology can be taken forward.

This study therefore answers the following research questions:

1. How can Data Mining – and specifically Text Mining – techniques be used by small development organisations with limited technical expertise to gain accurate, relevant insights into development issues?

2. Using BBC Monitoring Reports as an example data source, to what extent can a summary of global broadcast media be used to measure the extent of discussion about corruption? Can this measure be disaggregated by time and location, and be used to tell us more about the nature of corruption?

1. Data Mining in the Statistical Narrative

1.1. 'Traditional' Statistics and Data Mining

The foreword to *The History of Statistics, Their Development and Progress in Many Countries*, published at the close of World War One, states that

> the history of the development and organisation of official statistics is not a barren record of steps in a scientific process of dealing with facts, but of efforts to get a working knowledge about the fundamental elements in the life of a country – the population, its environment, and its manifold economic and social relations. By taking measure of these elements, statistics reveal the condition of growth and trend in every direction and set out the milestones for the guidance of the administrator and legislator. (Koren 1918: ix).

It is in the spirit of gaining a working knowledge of the life of a country that Data Mining may be best utilised. A central tenet of this methodology is that, whilst the principal process is a quantitative one, the results must be interpreted with a contextual awareness; this often means the quantitative process must be underlined by a qualitative one. This is particularly important for Data Mining, as many 'spurious' patterns will be 'discovered', and an expert in the field is needed to interpret the data, and determine whether the patterns are real and have meaning (Hand *et al* 2000: 117). Merging the data with existing knowledge is a key challenge, and is central to the success of the methodology (Baesens *et al* 2009: 22). Indeed, the most common accusation against Data Mining is that practitioners 'fish' the data for the pattern that fits their argument, an affront to 'traditional' statistics.

Traditional statistics is characterised by probabilities, hypotheses, and distributions; models are built that can then be used for predictions and classifications, such as the spread of disease. This method is mostly unchanged since the 1930s, yet it is not so useful when the data set is large and little understood (Franklin 2005: 83). As data sets increased in size, new statistical methods emerged to analyse them. Supporters of Data Mining have encouraged this divergence from traditional statistics. Breiman (2001: 202) writes that the traditional use of data models is 'deeply troubling', especially when they are so commonly used to draw quantitative conclusions and make policy decisions.

The aim of Data Mining is to find patterns and interesting information, especially those that

describe underlying structures in the data. Data Mining originated as early as 1941, with a set of 7,000 observations 'mined' for credit scoring (Baesens *et al* 2009: 16). The modern usage of Data Mining is predominantly commercial – for example, to analyse consumer shopping habits (if a customer purchases a book on gardening, what other products may they be interested in?) – and has its basis in the work of computer scientists in the 1980s (Franklin 2005: 84). Text Mining emerged as a subset of Data Mining with the need to catalogue text in books, and has evolved considerably since, including into the field of Artificial Intelligence. Text Mining is the central methodology of this study, and is defined as the 'discovery of previously unknown information in the text that is implicit but not immediately obvious' (Miner *et al* 2012: 4). Its use is also largely commercial – to 'mine' the attitude of consumers towards a product from online reviews, or strengthening spam filters – but also has increasing usage in molecular biology to classify genes and proteins (see Krallinger & Valencia 2005).

Data Mining flips the focus away from sampling and inference towards processing and tabulating data. Instead of using a sample to make inferences about a population, Data Mining often looks at the entire dataset. This leads to different statistical techniques, although the basic principle is unchanged – to reduce the 'noise' from data to produce information (Hand *et al* 2000: 112; Lambert 2003: 217-220). Nonetheless, the debate between 'traditional' statisticians and data miners is amongst the most fierce in the field (Franklin 2005: 84).

Hand *et al* (2000: 111-112) provide a astute explanation for the negative connotations of Data Mining. A researcher can always find apparent patterns in data sets, and whereas in traditional statistics a researcher may take a small set of interesting data points and see whether they can be attributed to chance, a data miner simply looks for the small set of interesting data points. The key to accurate and valid Data Mining is to look for 'systematic' patterns, rather than random ones. Within this study, for example, it is found that the words 'government' and 'anti' appear often in the text near the word 'corruption', yet not with each other. This pattern – that 'corruption' is associated with 'government' but 'anti-corruption' is not – appears consistently in the text, and is worthy of attention.

Other critics are concerned with the ethos of Data Mining. Solove (2008: 343) draws on an example of the US Government constructing the 'largest database in the world' and using Data Mining to 'predict' likely terrorists using health, financial and educational data. Such use of Data Mining has serious implications for privacy, confidentiality and data protection, and is an extreme example of

data being used to try to *prevent* an activity or action rather than *investigate* it. The adoption of Data Mining by the US Government was allegedly the result of aggressive marketing by commercial database companies; Solove states that the risks of a erroneous online book recommendation may be slight, but the risk of false-positive analyses in a security scan are serious indeed. More practically, he argues that Data Mining simply has not proved itself useful as a security measure (Solove 2008: 353).

For others, the issue is not Data Mining itself, but that the findings are not fully understood and effectively used (Kahn 2000: 127). The lessons for development are evident: ethical considerations of privacy and data protection are paramount, and a successful transition from the private to public sector is not as straightforward as simply transposing the technology from one sector to another; the merging of knowledge and data is crucial, and Data Mining should be but one tool in the effort to tackle development issues, and one part of the process in understanding them.

1.2. Big Data and Data Mining

Big Data refers to data sets so large they exceed the processing capacity of conventional database systems. It picks up the central strand of Data Mining - that data can be used for *predictive* as well as *descriptive* purposes – and applies it on a massive scale (Ohlhorst 2013: 4). Placed in historical context, the motivation behind the rise of Big Data in the past few years and Data Mining in the past few decades is the same - 'to leverage large quantities of cheap computing power to perform analyses previously unthinkable at scales previously intractable' (Leetaru 2012: 16). Similarly, Big Data and Data Mining adopt the same statistical philosophy – instead of analysing the larger picture from a smaller sample, individual data points and the relation between them take on a greater significance; put simply, these methods deviate from traditional statistics by 'zooming in' rather than 'zooming out'. The same disclaimers also apply – larger data sets are not necessarily 'better', and the quality of the data overrides the importance of statistical methods (Janert 2010: 7).

However, whilst both Big Data and Data Mining have open source software options, Big Data requires considerable technical expertise and access to a distributed computing system – a set of powerful networked computers (Ohlhorst 2013: 9). This prohibits the adoption of Big Data analysis for many organisations, especially those in low-income countries (Hilbert 2013: 17). The World Economic Forum (2012: 6) notes that large companies often lack access to suitably skilled

professionals for Data Mining analysis, let alone Big Data analysis. Similarly, information extraction (a branch of text mining, itself a branch of data mining, and the methodology used in this study) has been described as difficult to utilise without considerable effort and the application of specialist algorithms and software (Miner *et al* 2012: 37). It is hoped that this study, using a stripped-down methodology, will help to redress common concerns that Data Mining techniques are out of the technical reach of small organisations, and demonstrate that these organisations can replicate some of the novel insights that larger development organisations, using Big Data techniques, have been afforded.

1.3. Replicating Big Data Techniques for Development

In 2009, the UN Secretary General launched UN Global Pulse, an initiative aimed at 'analysing real-time digital data to detect early emerging vulnerabilities' (Letouzé 2012: 3). Early work from Global Pulse has used social media data to predict changes in unemployment levels and website data to model food prices – the premise of the initiative is that the same 'Big Data' analysis techniques used by online retailers to predict customer shopping preferences can be adapted to help with international development research and interventions (Kirkpatrick 2012: 6). Whilst Global Pulse acknowledges that Big Data is not a panacea for development problems, the potential is exciting: increases in both data processing capabilities and access to data sets may 'reveal remarkable insights into the collective behaviour of communities' (Letouzé 2012: 6); hidden within large data sets lie 'valuable patterns and information, previously hidden because of the amount of work required to extract them' (Dumbill 2012: 3). This study adopts the central tenet of Big Data for development – that large, fast-changing data sets hold great potential for providing new insight to existing development issues – and replicates the techniques on a smaller scale.

The UN is not the only organisation using big data for development analysis. Researchers have demonstrated how mobile phone use can model disease outbreaks and the movement of refugees, and patterns in mobile credit purchases can indicate imminent economic distress (World Economic Forum 2012: 2-4). Such analyses do not replace the need for effective interventions, but they can shorten the lag between a changing development situation and an appropriate reaction.

Whilst social media data is often 'open', confidentiality concerns remain, and analysis of mobile phone usage also raises concerns over privacy and data protection (World Economic Forum 2012:

5). A more traditional data source, avoiding the confidentiality issues of social media and mobile phone data, is news media. Researchers at the University of Illinois have used news reports to retrospectively predict the Arab Spring in Tunisia, Libya and Egypt, and the location of Osama Bin Laden to within 200 kilometres (Leetaru 2011: 1). The study used a '30–year translated archive of news reports from nearly every country of the world, applying a range of computational content analysis approaches including tone mining, geocoding, and network analysis'. This approach – named 'Culturomics 2.0' by the authors - is an 'intriguing new approach to modelling the behaviour of global society itself' (Leetaru 2011: 1-2). This archive – the reports of BBC Monitoring – is also used as the data source for this study, with a specific emphasis on one particular development issue: corruption.

2. Corruption as a Development Issue

This study uses corruption as an example of a development issue and a social phenomenon that Data Mining can be used to investigate. The reasons for this are twofold: first, corruption is regarded as intangible and difficult to measure, and is therefore a challenging test of the merits and drawbacks of Data Mining. Second, a measure of the level of discussion about corruption in public broadcasts may be a useful complement to existing measures of corruption for Civil Society and Non-Governmental Organisations. This section examines corruption in the context of international development before discussing existing measurements of corruption, and concludes by introducing the theoretical basis of the Broadcast Discussion Index.

2.1. Corruption and Development

Corruption is commonly defined as the abuse of public office for private gain, and is distinct from *illicit* or *immoral* activities (Bardhan 1997: 1320). This may include bribes, embezzlement of public funds, kickbacks in public procurement, or sale of government property. Corruption is an outcome of political, cultural, economic and legal institutions, and a growing body of evidence suggests that corruption in low-income countries constrains economic growth (Pande 2007: 3; Svensson 2005: 20).

There are two common schools of thought on the impact of corruption – that it can 'grease the wheels', keeping the economy moving and is a necessary cost of doing business, or that it represents 'sand in the wheels', increasing inefficiency in the system and damaging institutional effectiveness. A study by Aidt (2009) reviewed both macro- and micro-level evidence and found little support that corruption 'greases' the wheels; instead corruption is a source of unsustainable development and may lead to increases in GDP per capita, but not genuine *wealth* per capita. There is, however, considerable academic debate on the effect corruption has on development, with some studies refuting Aidt's findings. Campos *et al* (2010) reviewed 41 similar empirical studies, and find that, whilst the negative effect of corruption can be dampened by including institutions and trade openness in the quantitative models, corruption still has a genuine negative effect on growth.

International development, of course, is concerned with more than just growth. Transparency International, a global civil society organisation with chapters in more than 90 countries, stresses that corruption hits the poorest hardest, preventing access to basic healthcare and education and trapping the poorest people in a cycle of inequality (Transparency International 2012a: 39). Corruption also adversely affects incentive structures – widespread corruption may support inefficient firms, can drive entrepreneurs away from the private sector (in extreme cases, it can encourage them to become corrupt officials themselves), and re-allocates resources away from their most socially-productive use (Svensson 2005: 37). Effective regulations and policies to curb corruption are a facet of good governance, which itself is central to efforts to reduce poverty and increase social equality. However, whilst the literature comprehensively covers quantification of poverty and governance, measuring corruption is a considerably more challenging area.

2.2. Existing Measurements of Corruption

Corruption has been cited as 'the prime example of an observable social phenomenon that is not quantifiable since there cannot be statistics on a phenomenon which by its very nature is concealed' (Dogan & Kazacigil in Galtung 2006: 102). However, this has not prevented attempts to measure it.

The World Bank lists three broad ways in which corruption is being measured: first, by gathering the informed views of relevant stakeholders, second, by careful audits of specific projects, and third, by tracking countries' 'institutional features' (Kaufmann *et al* 2006:2). The most well-known measure, Transparency International's Corruption Perceptions Index (CPI), falls into the first

category. One benefit of using multiple measures and perspectives, aggregated into an index, is that an approximate measure of corruption can be gained with incomplete data. The CPI compiles data from 17 surveys from 13 sources – a combination of sources measuring the same phenomenon is more reliable than each source taken separately (Transparency International 2011b: 1).

The CPI is deliberately broad: 'the data sources used to compile the index include questions relating to the abuse of public power and focus on bribery of public officials, kickbacks in public procurement, embezzlement of public funds, and on questions that probe the strength and effectiveness of anti-corruption efforts in the public sector. As such, it covers both the administrative and political aspects of corruption' (Transparency International 2011a).

This and other indices of corruption tend to rely on perception or experience surveys – for example, how widespread the respondent believes corruption is (based on opinions, anecdotes or experience), or recording corruption encountered first-hand. They are popular with good reason – such a method is easily replicable in other countries thus allowing easy comparison and ranking, does not rely on specialist equipment or expertise, and can be tracked over time with the main costs being the survey administration. However, people often perceive greater corruption than they experience, and it is difficult to establish whether the corruption experienced is representative (Hussmann 2011: 36). Despite this, perceptions are important as they affect how citizens engage with government services, and perception surveys tend to reflect the observations of development banks and risk-rating agencies – two other sources of measurements of corruption (Kaufmann *et al* 2006: 2).

The literature on identifying and measuring corruption is still in its infancy (Svensson 2005: 31); a measure of the level of public discussion about corruption can complement existing measures.

2.3. The Broadcast Discussion Index (BDI)

The Broadcast Discussion Index is a ranking of places and other items of interest by the number of mentions they receive in a corpus of news reports. The places can be countries, regions or cities, and other items of interest can be sectors (such as the education, health, or legal sector), institutions and organisations (banks, terrorist organisations or government ministries, for example) or individuals. The corpus used in this example is the BBC Monitoring dataset of reports concerning corruption.

For example, an influential blog that exposes corruption in Japan's railways sector will both increase the score that Japan receives in the country corruption rankings, and the count of 'railways' in the 'sector' index (perhaps under the header of 'infrastructure'). A local newspaper article praising the Afghan government's tough stance on corruption will increase the score of both 'government' and 'Afghanistan'. Notably, the index does not record that the latter example has a 'positive' slant – that efforts are being made to decrease corruption. As a result, this index explicitly measures *discussion* about corruption in broadcast media; it does not measure the *incidence* or *prevalence* of corruption.

Without detailed text network analysis or sentiment mining (see appendix 4 for definitions), the 'tone' of each article is difficult to determine. Section 5.4 looks at the semantic usage of the word 'corruption', although this is disconnected from specific countries. Future iterations of this work can draw upon text network analysis, which could connect frequent usage of, for example, 'anti-corruption', 'Afghan' and 'government' (an example is provided in appendix 3).

2.3.1. Why Broadcast Media is Important for Development

The World Bank (2009: 5-8) explains how the media plays an instrumental role in development by storing and sharing knowledge, and lists several further contributions that the media can make to development. Two of these – influencing behaviour and increasing transparency – are relevant to discussions on the media and corruption. Using a broad archive such as BBC Monitoring Reports allows an insight into behaviour and attitudes. Yet the media does not just *reflect* behaviour, it can also *influence* it. Whilst a significant level of discussion of corruption does not necessarily equal an influence on behaviour (an article may state, "corruption is rife, refuse to pay bribes!" or, "this country has no corruption!"), it does increase *awareness* of corruption as a concept and as a possible issue. Whilst awareness is not enough by itself to tackle corruption – an empowered judiciary, active civil society and receptive government are all necessary (Rønning 2009: 169) – it is a necessary first step.

Analysing domestic media affords researchers a uniquely grounded, 'local' perspective of development issues. Analysing domestic media may also present a quite different set of issues to those deemed important by multilateral bodies. In a wider sense, broadcast media – especially with the inclusion of social media such as blogs - may be seen as an embodiment of public discussion.

As Castells (2010: 5) remarks, the media represents the 'public mind', and 'what does not exist in the media does not exist in the public mind, even if it may have a fragmented presence in individual minds'. The link between corruption being discussed in the media and the media influencing the public mind is, in essence, why discussion of corruption in broadcast media is important.

2.3.2. Theoretical and Empirical Considerations of the BDI

The BDI measures public discussion of corruption, and several theoretical assumptions are made in this regard. First, countries are considered actors, rather than geographic entities. Whilst most articles in the corpus will be discussing national affairs, external countries are also referenced. For example, a Nigerian radio report on alleged corruption in the oil sector will increase the count of 'Nigeria' in the country index, and by several increments if 'Nigeria' is mentioned several times. If the oil companies in question are owned by European countries, these countries too will see their score in the country index rise. There is therefore no distinction made between 'internal' and 'external' actors.

Second, discussion of corruption is taken as a proxy for corruption being *considered an issue.* This assumes that in a country where corruption is rampant but accepted, perhaps as 'the cost of doing business', it would most likely fail to score highly in the index. If, as with the Afghan newspaper article, the government is praised for fighting corruption, this confirms that corruption is a current topic of interest, and therefore warrants a higher score on the BDI. We can also make the theoretical assumption that anti-corruption efforts would not make headlines without any instance of corruption to tackle. In this regard, anti-corruption mirrors corruption. After all, a country with a decade free from corruption is unlikely to be heralded in the media for renewed anti-corruption efforts.

The Broadcast Discussion Index differs from the Corruption Perceptions Index in that a country may be held accountable in the scoring to its actions abroad. If, for example, Sweden owned a refinery in the oil sector under scrutiny in the Nigerian radio broadcast, Sweden's score would increase. There is a risk that international actors assisting in the fight against corruption would see their score increase – for example, SIDA launching a programme to increase accountability in the oil sector. A scan of the corpus suggests that local media, as picked up by BBC Monitoring, rarely features stories of this kind. Nonetheless, future iterations of the methodology should be able to make such distinctions (see section 3.5); the results presented here are a preliminary guide to what

insights are possible from Data Mining.

Lastly, it is worth considering factors - beyond an increased awareness of corruption - that could contribute to a rise in the level of discussion about corruption.

- A sudden, unusual scandal could increase coverage (the Member of Parliament expense scandal in the UK in 2009 is such an example). Most likely, this would lead to a surge in just one or two years. As a result, the country rankings for the past five years have been disaggregated year-on-year.

- A greater level of media freedom could lead to greater discussion of corruption. A powerful, dissatisfied media could try to destabilise the government with corruption claims; similarly, there may be widespread acknowledgement of corruption but a weakened, censored media may be unable to report it. Media freedom is explored further in section 5.7.2; very little correlation between a free press and the BDI score is found – if anything, there is an inverse relationship between press freedom and discussion of corruption. Furthermore, the number and breadth of sources that BBC Monitoring utilises is intended to avoid such issues – for example, by tracking online blogs such as those that predicted the Russian election protests in 2012 (BBC Monitoring 2012: 1).

- Costa (2012: 3) finds that the introduction of Freedom of Information laws leads to higher perceptions of corruption – and that these perceptions remain higher over time. Freedom of Information is assumed to increase asymmetry of knowledge (Islam 2003: 12), and this may mean previously 'hidden' corruption becomes exposed. We may therefore expect a greater level of media broadcasts referring to corruption when awareness of corruption increases.

2.4. Relevance of the BDI for Development Organisations

The BDI, as an example of how Data Mining can be applied to real-world problems, presents potential benefits for Civil Society Organisations (CSOs). Civil Society consists of organisations that exist outside the formal state apparatus, and, as they form 'cooperative, but independent and critical' partnerships with governments, they are seen as instrumental in long-term efforts to tackle and prevent corruption (Pope 2000: 134).

The BDI is a complement to existing corruption indices; it does not measure corruption itself, but discussion of corruption. As such, a small CSO working in a country regarded by the Corruption Perceptions Index as being highly corrupt may wish to plan a country strategy to increase public awareness of corruption and corrupt practices. The BDI would allow the CSO to determine the level of public discussion of corruption – if levels are high, then perhaps popular knowledge about corruption is not the issue, rather there are institutional blockages that can be identified and work should be focussed on reducing these. If levels of discussion are low, perhaps a media strategy or public information campaign to highlight the negative effect of corruption is needed. Discussion may focus on a particular sector, department or individual – Data Mining can identify these, and the CSO can respond accordingly. Expert interpretation of the Data Mining data is essential to its success, and local CSOs are often the most knowledgeable on local context, and have experience of how to leverage new knowledge to increase the impact of their work; Data Mining is simply the tool.

Other CSO organisations exist to increase accountability and transparency. Corruption can 'flourish' without accountability, yet accountability needs to be accompanied by effective monitoring (Costa 2012: 1-2). Gauging the content of the 'public mind' is one way to do this. The BDI can identify trends in discussion – for example, can a sudden increase in discussion be attributed to an event or intervention? The dissemination of new concepts of governance, and new incidences of corruption, can be monitored. Of course, corruption and governance is but one example application of the BDI. A human rights organisation may track the coverage of sexual violence in the media and, depending on the context, the *lack* of broadcast coverage may be of equal concern as a high level of discussion.

A Data Mining model has several other advantages. First, data can be processed in real time, without the lag associated with survey data – this means organisations can respond faster, and may be able to act pre-emptively (for example, investigating a mention of corruption in a public procurement project before the contract is signed). Second, the costs are considerably lower. Third, whilst the BDI does not measure corruption itself, the model is infinitely customisable and future work may lead to a measure more directly comparable to the CPI. Of course, the BDI and Data Mining is subject to limitations and constraints too. In addition to the methodological limitations identified in section 3.4, the quality and usefulness of the information is limited by the quality of the data set used, and findings may be subject to incorrect interpretation and false conclusions.

3. Methodology

The Data Mining process is, upon first inspection, largely a technical one. However, there are two non-technical factors that are essential for a successful and useful project. The first has been discussed in section 1.1 and is the need for contextual understanding of the results. Although the principal aim of Data Mining is to find new and interesting patterns and information, patterns that can be explained are more likely to be 'real' than those for which a convincing explanation cannot be found (Hand *et al* 2000: 117). Second, the data source should be of the highest quality before the process begins. Some preprocessing can (and should) 'clean' the data prior to analysis, but the most effective path to more useful results is to keep the model simple, and improve the data that goes into it (Baesens *et al* 2009: 17).

The process outlined here is for Text Mining, which is a subset of Data Mining but for which the same principles and methodological steps apply. Text Mining has much in common with Content Analysis, and more complex studies with a deeper analysis of the results will borrow heavily from Computational Linguistics. Put simply, however, Text Mining is a means of analysing the communications of other people (Miner *et al* 2012: 1008), and the analysis that follows can be reproduced by organisations without the technical expertise needed for more complex statistical analysis.

This section sets out the methodological process of Text Mining for the example of creating an index of discussion on corruption. First, the data source is discussed, followed by a description of the corpus. The Text Mining process is then explained; this methodology is a basic foundation and should be transferable to other contexts and purposes. Last, the limitations of the methodology are discussed.

3.1. Source Material and Limitations: BBC Monitoring Reports

Data quality is often a troublesome issue in Data Mining analyses, and emphasis should be placed on securing the highest quality data set. One should expect any large data set to have imperfections, but a suspect data set can result in patterns and information appearing in results simply because of errors and oddities in the data collection process (Hand *et al* 2000: 111-119). At the very least, imperfections in the data set should be borne in mind at all stages of the data analysis. This study

uses BBC Monitoring Reports as the data source.

BBC Monitoring Reports, formerly known as the Summary of World Broadcasts, follow and transcribe news reports daily from over 150 countries in over 100 languages (BBC Monitoring 2013). The service was established together with the Foreign Broadcast Information Service in the US (now the Open Source Center under the auspices of the US Government) in the lead up to the Second World War to collect 'open intelligence' on the opinions and reactions of local media throughout the world. Together, these organisations monitor up to 32,000 sources, and particular emphasis is placed on a unique 'iterative translation process' that preserves the 'minute nuances of vernacular content' and captures the 'subtleties of domestic reaction' (Leetaru 2011: 2). This allows effective analysis of *domestic* discussions, as English-print news in non-English speaking countries is often Western-focussed, and global news sources do not maintain geographically-even coverage. Although a product of British and American intelligence organisations (and still widely used by both), this 'open source intelligence' serves as a 'strategic resource, maintaining relatively even monitoring volume across the globe on a broad range of topics' (Leetaru 2010: 34; Leetaru 2011: 2).

BBC Monitoring Reports also represent the changing medium of broadcast discussion. Reports are increasingly being sourced from online, with over 45 percent of content gathered from online sources in 2010, including influential blogs and online-only news sources (Leetaru 2011: 3). Other sources include printed newspaper articles, television programmes, and radio shows. These sources are collectively referred to as 'broadcast media' in this thesis.

Although BBC Monitoring Reports are used here as an example data source (a civil society organisation working at the national level may prefer an complete archive of national press, for example), awareness of the source limitations is central to the methodology. Accordingly, the reports are a convenient and broad source, but they cannot be considered representative – priority is given to countries deemed politically important, such as Afghanistan, Iraq and Pakistan (Leetaru 2010: 30). This has significant implications for the impartiality of the source and its applicability for development organisations; these political 'weightings' need to frame the interpretation of the results. Furthermore, there is no information provided on country breakdown by number of reporters or number of articles; the source covers a large geographic area but the balance and distribution is uncertain. These limitations are considered further in section 3.4, in particular issues surrounding editorial bias and the translation process.

3.2. Sample Data: Corpus on Corruption and Bribery

Whilst the full BBC Monitoring archive dates back to 1979, this research focusses on a five-year period to February 2013. The corpus contains every document from BBC Monitoring that mentions 'corruption' within this time frame – a total of 34,666 articles. Each article can range from 300 to 3,000 words; the corpus contains just over 140,000 different words, with a total of over 31 million words. The term 'corruption' appears 77,892 times within the corpus. A smaller analysis was also run on the term 'bribery' within the same five year period – this produced just under 4,000 articles.

Table 1 disaggregates the articles featuring corruption by year and geographic area. Four areas were used for search parameters: Africa and the Middle East, Asia-Pacific, Europe, and South America. The majority of articles in the corpus are from Asia-Pacific, and Africa and the Middle East. Monitoring reports from other areas, such as the Caucasus and former Soviet Union, are excluded from the main analysis (but are included in the accuracy tests in section 4). Countries in these excluded regions still appear in the index, as they are referenced in articles from other regions, but are under-represented. The UK and the US are not included in BBC Monitoring coverage and are removed from the index – the former is covered by regular BBC news reporting, and the latter is restricted by CIA regulations on monitoring American press (Leetaru 2011: 2).[1]

Table 1: BBC Monitoring articles in 'corruption' corpus by region and year

Year	Africa and Middle East	Asia Pacific	Europe	South America	Total
2012-13	2148	3980	1756	124	8008
2011-12	3230	3199	1625	161	8215
2010-11	2264	2542	1156	55	6017
2009-10	2524	2535	985	102	6146
2008-09	2430	2571	1151	128	6280
Total	12596	14827	6673	570	34666

3.3. The Text Mining Process

The Text Mining process mirrors that of a typical Data Mining methodology (such as Han *et al* 2012: 6). The data is preprocessed – cleaned, selected and transformed. It is then mined, evaluated and presented. The form of Text Mining employed in this methodology is known as *information*

1 There are other reasons for excluding the UK and US. There may be a skew towards UK interests, as this is a British source. The US is complicated by technical elements of the methodology, which necessitate lower case and thus conflates the nation 'US' with the pronoun 'us'.

extraction, and, as the name implies, involves extracting specific information which can then be analysed; this analysis can result in new information that is predictive in nature, as opposed to simply describing the original text (Miner *et al* 2012: 12-13). Indeed, analysis of text can be particularly valuable for international development – words can predict actions or activity, and analysing online text can produce near-identical results as household surveys and polls (Hilbert 2013: 7). The process begins by sourcing the data.

BBC Monitoring Reports were accessed from LexisNexis, an electronic database for legal and public records related information, and all articles meeting the search parameters were stored as PDF files for offline analysis. LexisNexis has a built in semantic search: for example, if searching for 'bribery', an article describing the payment of bribes in Algeria but using the word 'kickback' rather than 'bribe' will still be featured in the search results. Each article is categorised by LexisNexis and topics (such as 'bribery' and 'corruption') assigned; however, these index terms are not saved in the final PDF data – first, to ensure that only the original news text is analysed, and second, to prevent interference with the 'token region' analysis (explained below).

RapidMiner, an open-source data mining and analysis program, was used to sort and clean the data. First, articles were grouped by year for time-series analysis. All articles were then 'tokenised', whereby each unit of analysis was specified – in this instance, each separate word was defined as a token.[2] Third, all words were made lower case to avoid duplication (combining 'Afghanistan' and 'afghanistan', for example).

The underlying assumption of Text Mining is that the 'meaning' of a text can be represented by a frequency list of the words used in that text (Miner *et al* 2012: 80). At this stage, RapidMiner was used to perform specific data analyses:

1. For the primary analysis, a word vector was created of all words in the corpus. This is an alphabetical list of every different word used, and a count of the number of times it appears. The unifying theme of all Text Mining operations is to transform text to numbers, so data analysis techniques can then be applied (Miner *et al* 2012: 30).

2. For the secondary analysis, the proximity of words to each other within the corpus was

2 Unfortunately, additional RapidMiner functionality, such as N-Grams, could not be utilised for the primary analysis. N-Grams look for combinations of two (or more) words commonly together, such as 'Prime Minister'. A consequence of this is that 'South Sudan' cannot be isolated from 'Sudan', nor 'South Africa' from 'Africa' (although it is still possible to include Sri Lanka owing to its unique constituent words). N-Grams are commonly used on smaller bodies of text (such as in section 5.5); the main corruption corpus here would require considerably more computing resources.

examined, known as a 'token region' analysis. To do this, a central word (or 'token') is specified – 'corruption' – and significant words that co-exist within a set range of words either side of 'corruption' were included in the word vector. Depending on the specific token region analysis conducted, some commonly used words[3] were then stripped from the word vector, as were single characters (such as 'I' or ''s'), leaving the significant words most commonly found in close proximity to 'corruption'.[4]

With large data sets, a certain degree of tailoring is needed for the mining process. Flexibility is needed to model the complicated structure of data, but it also needs to be 'smooth' to avoid 'over-fitting'; it is beneficial to 'learn the idiosyncrasies of the particular data set in a way that will not generalise to other sets of the same kind' (Franklin 2005: 84). In this study, further cleaning was required for the word vectors.

Word vectors were analysed in Microsoft Excel. This study uses the United Nations Secretariat Statistics Division's list of countries (UN 2013). This list of countries was adapted to include different spellings of the same country (Cameroon and Cameroun), and geopolitical variations (Myanmar and Burma). Multiple-word countries were reduced to single search terms where possible ('Sri Lanka' to 'Lanka', for example). Further word lists were selected from another open-source application, the General Architecture for Text Engineering (GATE). Using Excel's *vlookup* functionality (open source spreadsheet applications have similar functions), frequency counts for words on each of the lists were drawn from the larger word vector.

Owing to the N-Gram issue (see footnote 2), Congo and DR Congo, South Africa, Central African Republic and North and South Korea were removed (along with some smaller territories, such as Channel Islands and Cook Islands), and Sudan and South Sudan are amalgamated, as much of the reporting is dated prior to South Sudan's independence in 2011. 215 countries were included in the final analysis.

3 The most frequently used words are known as 'stopwords'. These words – such as articles, conjunctions and auxiliary verbs – do not directly add meaning to the content (Paranyushkin 2011). However, they are still included in the primary analysis, where proximity is not modelled.

4 For example, stopwords and single characters were stripped *before* running the 'token region' process for words immediately either side of 'corruption', as we are looking to determine closest significant words, but were stripped *afterwards* (from the vector) for the sentence analysis in order to reproduce an average sentence length.

3.4. Methodological Limitations

As an experimental study, this research has several notable limitations. Notwithstanding the promising results that this study provides, the methodology will require a more gradated understanding of the *tone* and *sentiment* of articles in future iterations. Technical suggestions for implementing these are provided in section 3.5. At present, an article may included in the corpus with just one mention of the word 'corruption', and, accordingly, corruption may be tangential to the article. To counter this, the large sample size (nearly 35,000 articles) should reduce the number of anomalies, and the token region analysis focuses more closely on words surrounding 'corruption' to avoid other 'noise' in the article (in a minority of cases, articles are a summary of news broadcasts on radio or television, and cover multiple, wide-ranging topics).

The BBC Monitoring archive is comprehensive, but it nonetheless represents the selective choices of country analysts and is thus subjective. The corpus represents a selection of an already curated selection of world news. Rather than being seen as representative of domestic media broadcasting in 215 countries of the world or as a pool of untouched primary source material, it should be seen as a reflection of notable discussion topics and news items. However, there is no indication of whether an article on corruption in the Ugandan education sector is notable as an exceptional example of investigative journalism, or an example of many similar articles and simply chosen to highlight an emerging trend in editorial focus. Neither can the dissemination of articles easily be analysed without more detailed analysis – whilst each article states the source, this methodology does not currently assign any weight to a newspaper read by an elite minority, or a television programme enjoyed by the masses; these are areas for a future study.

Furthermore, this study only captures discussion in Castells' 'public mind', rather than private discussion; broadcast media, rather than social media (such as Facebook or Twitter or local variations), is the focus of analysis. Neither is this study able (through access issues) to corroborate the findings from the BBC archive with its US counterpart, the Open Source Center. Whilst these limitations may be surmountable in time, here they help to delineate the boundaries of the research.

The advantages of using open source intelligence, such as BBC Monitoring, it that it can be used to 'peer into closed societies, to predict major events and to offer real-time updates' (Leetaru 2010: 20). However, despite the use of skilled translators, source material is sometimes revised and re-translated. For example, a headline is changed from 'inaugurated' to 'set up', and the article is re-added to the database; this duplication of articles creates issues for content analysis, although

duplication appears to be concentrated between 1998 and 2002, outside the range of this study (Leetaru 2010: 32). A strength of this source, and news broadcasts more generally, is the ability to gauge opinion and public discussion, and a weakness is the according lack of objectivity. It it thus highly suited for modelling behaviour and attitudes, but less useful for determining root causes and structural changes. News broadcasts are dynamic and subjective, reflecting behaviour and attitudes, issues and concerns – they are the frame through which corruption as a phenomenon in society can be captured and measured.

3.5. Extending the Methodology

This study presents a basic Data Mining methodology. There are exciting areas in which this work can be expanded and the model further refined, although in all applications the emphasis should remain on a 'clean' and simple methodology tailored for the purpose and that recognises the limitations of the source data. Conceptually, different themes and issues can be explored - those similar to corruption, for example fraud and election rigging, and unrelated issues such as immigration, urbanisation or HIV/AIDS. Different source materials such as social media could be used, or even transcripts from public speeches and parliament debates, or archives of all print media from a particular region.

There are also several directions that technical modifications can be made. For example, the text network analysis and N-Grams briefly explored in section 5.5 and appendix 3 can be applied to a wider corpus. Sentiment mining can capture the *tone* of individual articles, and this tone can be tracked over time, as with the 'Culturomics 2.0' work. However, Data Mining, and in particular Text Mining, is still an emerging field, and considerable advances can be expected in a relatively short period of time. A greater convergence with the field of computational linguistics will allow 'deep discovery' of features such as sarcasm, innuendo and idiomatic language (Miner *et al* 2012: 12). Real-time Data Mining is also an emerging area, whereby data is analysed and interesting patterns found on a continuous basis (Baesens *et al* 2009: 22). Last, and perhaps of greatest relevance for smaller organisations working in development, will be more advanced trend detection, and better multilingual analysis (Miner *et al* 2012: 24). Coupled with the recent increase in availability of large open datasets and the concept of 'data philanthropy' – large organisations stripping personal information from data records and releasing them for analysis (Letouzé 2012: 25) – the potential for gaining innovative, accessible insights is greater than ever before.

4. Testing the accuracy of the BDI

The Text Mining methodology was tested on the BBC Monitoring Reports data set for several international development issues to test the accuracy of the BDI. As the primary existing measure of corruption is a ranking of the level of perceived corruption, it is difficult to verify with empirical evidence whether a measure of the discussion of corruption is reflecting bias in the source material. As a result, several other current issues with more easily verifiable real-world phenomena were analysed using BBC Monitoring Reports. To explore the theoretical assumption that countries are *actors* rather than *geographic entities*, acting both within their own territory and interacting with others, two issues involving complex international movements were tested – arms trafficking and cocaine. To test against development priorities with well-established reporting practices and reliable national statistics, malaria and tuberculosis were also tested.

4.1. International Movements: Arms Dealers and Cocaine

A search for 'arms dealer' yielded a relatively low total of 519 articles for the past five years, worldwide. The words 'Russia' and 'Russian' appear over 2,500 times in total; China is the second ranked country by word occurrences, although USA also appears often (despite 'US' being filtered; see footnote 1). The Stockholm International Peace Research Institute (SIPRI) estimates the US and Russia to be the two largest arms exporters; China ranks fifth (SIPRI 2013). Viktor Bout, arguably the world's most infamous arms dealer (Austin 2002: 203), is mentioned over 750 times. His arrest in Bangkok in 2008 explains his prominence in the BDI over this time period; 'Thailand' and 'Thai' also rank highly.

The BDI appears to be stronger at locating arms-exporting countries than arms-importing ones. Closer semantic analysis would be required to determine where arms may be headed. After the aforementioned countries, Yemen and Iran are mentioned most frequently – 340 and 294 times, respectively. Discussion of these countries in the corpus is noteworthy because of their absence from the top of SIPRI's arms import and export lists; instead, both have been implicated with shipping arms to al Shabaab militants in Somalia (Charbonneau 2013).

A different analysis was conducted on the corpus, using 'Cocaine' as the search term. As with arms transfers, cocaine is internationally trafficked and difficult to quantify and hence is a useful test for the BDI. The top mentioned country in the 'cocaine' corpus is the Russian Federation, followed by

Serbia, Montenegro, Colombia and Ukraine. As the world's leading producer of cocaine, it is unsurprising to see Colombia in fourth place (Bolivia and Peru, also noted for production, are number six and 11 respectively). The UN office on Drugs and Crime noted an increase of seizures of cocaine of up to 30% during this time period in the Russian Federation and Ukraine as cocaine may be entering Europe through new routes; this may also explain the high ranking of Serbia and Montenegro (UNODC 2011: 112).

4.2. National Statistics: Malaria and Tuberculosis

Information on global disease trends is relatively easy to obtain and is therefore straightforward to cross-reference to the source material. Recent work, such as Google Flu Trends, is already using human-created data – in this case, search engine trends – to predict and map disease spread (Letouzé 2012: 20). This test analysed the countries associated with malaria in the corpus. China tops the list, in part due to public health interventions in other countries. Consequently, a BDI ranking for malaria, for example, would not solely reflect prevalence of malaria, but also countries involved in fighting it.

Over the past five years, the top ranking countries for Malaria in the source material are China, Pakistan, Eritrea, Nigeria and Uganda. Whilst none of these five countries are in the top five countries by number of reported malaria cases – Uganda is closest in sixth place, and Nigeria is number 17 (WHO 2013) – all with the exception of China (where reported cases have sharply dropped) have experienced a significant rise in the number of reported cases: Pakistan from 104,334 in 2008 to 240,591 in 2010, Eritrea from 8,764 to 35,982, Nigeria from 143,079 to 551,187 and Uganda from 979,298 to 1,581,160. The BDI is therefore able to track emerging trends and important actors, although contextual interpretation is required.

Tuberculosis (TB) was also tested. The World Health Organisation lists 22 High Burden Countries, accounting for 80 percent of all new TB cases each year (WHO 2012: 8). Four of the top five countries in the BDI for tuberculosis are High Burden Countries; Ukraine, in third place, is not a High Burden Country but does have high levels of drug resistance, and unusually high public expenditure to tackle TB (WHO 2012: 64).

4.3. Summary of Tests

Such tests are useful to triangulate the information in the corpus with real-world events. The robustness tests demonstrate that the findings of the Text Mining analysis do reflect real-world events, although there may be over-representation of some countries – Afghanistan, for example, features prominently in all of the above analyses. The tests also demonstrate that the BDI reflects both ongoing concerns (for example, Colombia manufacturing cocaine), and 'spikes' in news, or departures from the norm (such as reported malaria cases increasing in Nigeria by over 380 percent). This has implications for the measurement of discussion about corruption – if a country is hit by an unexpected corruption scandal, this is likely to be reflected in the index, whereas it would not be reflected in, for example, the Corruption Perceptions Index, unless the corruption scandal is representative of widespread corruption.

Lastly, it is worth noting the total number of articles analysed for the four issues above represent under 15 percent of the total articles for corruption. Whilst malaria, TB, cocaine and arms trafficking are undeniably pressing issues, corruption – an invisible phenomenon – commands considerably more broadcasting attention.

5. Results

Using the BBC Monitoring Reports as an example source and corruption as an example topic, this section presents the results of a possible application of Data Mining. First, an overall Broadcast Discussion Index is presented – a five-year ranking of 215 countries – followed by more detailed analysis of the context and semantic structure. Such analysis is essential to correctly frame the results – it is 'dangerous' to analyse the data without knowledge of the context (Lambert 2003: 222). Kenya is then examined as a case study of the data, and bribery is explored as a subset of corruption. Lastly, the BDI is compared to existing datasets.

5.1. Five Year Corruption Broadcast Discussion Index

Table 2 shows the Broadcast Discussion Index for corruption for 2008 to 2013. This is a straightforward total of the mentions of each country in the articles that discuss corruption. The index places Afghanistan in first position in each year, by a considerable margin and with a mean average of 12,829 mentions per year. A full table for all 215 countries is presented in appendix 1.

Table 2: Top 20 countries in BDI for 'corruption', per year, 2008 – 2013

	2012-13	2011-12	2010-11	2009-10	2008-09	Mean Average
1	Afghanistan	Afghanistan	Afghanistan	Afghanistan	Afghanistan	Afghanistan
2	Pakistan	Pakistan	Pakistan	Iraq	Iran (Islamic Republic of)	Pakistan
3	Iran (Islamic Republic of)	Iran (Islamic Republic of)	Iran (Islamic Republic of)	Pakistan	Iraq	Iran (Islamic Republic of)
4	China	Iraq	China	Iran (Islamic Republic of)	China	Iraq
5	Syrian Arab Republic	Syrian Arab Republic	Iraq	China	Pakistan	China
6	Iraq	China	Egypt	Kenya	Israel	Syrian Arab Republic
7	Serbia	Libya	Israel	Sudan	Bangladesh	Egypt
8	Egypt	Egypt	Sudan	Yemen	Zimbabwe	Israel
9	Russian Federation	Jordan	Nigeria	Zimbabwe	Serbia	Serbia

10	Turkey	Israel	Russian Federation	Croatia	Russian Federation	Russian Federation
11	India	Serbia	India	Bosnia and Herzegovina	Turkey	Turkey
12	Somalia	Russian Federation	Zimbabwe	Nigeria	Kenya	India
13	Bosnia and Herzegovina	Turkey	Yemen	Serbia	Sudan	Sudan
14	Israel	India	Serbia	Israel	Czech Republic	Bosnia and Herzegovina
15	Czech Republic	Sudan	Tunisia	Lebanon	Lebanon	Czech Republic
16	Jordan	Czech Republic	Bosnia and Herzegovina	Turkey	Albania	Croatia
17	Saudi Arabia	Croatia	Croatia	Czech Republic	Egypt	Kenya
18	Nigeria	Yemen	Turkey	India	Syrian Arab Republic	Jordan
19	Libya	Lebanon	Czech Republic	Russian Federation	Bosnia and Herzegovina	Zimbabwe
20	Yemen	Tunisia	Kenya	Egypt	Croatia	Yemen

The top ranking countries – Afghanistan, Pakistan, Iran and Iraq – are also those that capture much of the political interest of the West, and may confirm the suspected source bias explored in section 3.1. The top ranking cities also reflect these countries – Kabul heads the list, followed by Baghdad, Islamabad, Karachi and Tehran. Similarly, provinces and regions are overwhelmingly Afghan – Herat and Kandahar first and second respectively and Helmand fifth – with Arbil (Iraq) third and Punjab (India and Pakistan) fourth. This possible bias does not necessarily invalidate the rankings – Afghanistan, for example, is ranked as the most corrupt country (joint with Somalia and North Korea) in the 2012 Corruption Perceptions Index (Transparency International 2012b: 3), and the Afghan population rated corruption a more pressing issue than poverty, external influence and the performance of the government in a recent UN survey (UNODC 2012: 3). The source material may, however, amplify the levels of discussion in select countries – pushing the ranking of Iran, for example, ahead of Somalia and weakening the value of the model for performing direct country comparisons.

With such a large data set and with potential bias in the source material, it can be beneficial to limit

the geographic scope; indeed, the most insight is likely to be gained from a smaller, 'micro' level focus, rather than a 'macro' one, and from monitoring trends over time. Table 3 ranks the top five sub-Saharan African countries by mean, and demonstrates how the scores can vary year-on-year. For example, Kenya shows a significant peak in 2009-10; this is explored in more depth as a case study in section 5.5.

Table 3: Top 5 sub-Saharan Africa countries in BDI, by mean

	Country	Mean Average	2012-13	2011-12	2010-11	2009-10	2008-09
1	Sudan	1018.2	508	1332	1181	1261	809
2	Kenya	836.6	489	565	701	1530	898
3	Zimbabwe	820.6	216	601	971	1115	1200
4	Nigeria	786.8	777	910	1084	1009	154
5	Somalia	518.8	1302	548	313	356	75

5.2. Change in Corruption Index Scores 2008 – 2013

Calculation of the percentage change between the total score each country received in 2008 and in 2013 allows a crude measurement of change in the emphasis placed on corruption in broadcast media over time. To ensure a robust sample, only countries scoring a five year total of at least 100 mentions of corruption were included in the percentage calculations; this eliminates many small island states that, with a total of, for example, six citations in five years, may record a 500% increase in discussion about corruption.

Table 4 shows the top five increases in Broadcast Discussion; all are African countries, with Cameroon recording an increase of over 4,000 percent. This could be due to changes within Cameroon, or changes in the source material (closer investigation, similar to that in section 5.5, could help determine this). Table 5 shows the top five decreases in Broadcast Discussion, with Namibia experiencing the sharpest decline. 87 countries saw their score increase over five years, and 105 countries had fewer mentions in 2013 than 2008 (the remainder experienced no change, or had no mentions in one of the years).

Table 4: Top 5 Increases in Broadcast Discussion, 2008 - 2013							
	Country	2013 as % of 2008	2012-13	2011-12	2010-11	2009-10	2008-09
1	Cameroon	4022.2	362	340	14	11	9
2	Mali	2992.3	389	68	23	9	13
3	Niger	2475.0	198	184	184	280	8
4	Tunisia	2325.0	558	966	855	18	24
5	Somalia	1736.0	1302	548	313	356	75

Table 5: Top 5 Decreases in Broadcast Discussion, 2008 - 2013							
	Country	2013 as % of 2008	2012-13	2011-12	2010-11	2009-10	2008-09
1	Namibia	3.3	5	7	19	83	150
2	Papua New Guinea	3.9	3	21	88	48	77
3	Fiji	4.9	4	16	50	81	81
4	Botswana	5.8	5	77	66	77	86
5	Paraguay	6.3	7	28	6	7	112

5.3. Context Analysis: Individuals and Institutions

In the context of corruption, an individual may be a reformer (seeking to curb corruption), a victim (perhaps an individual alleging corruption) or a perpetrator (an individual using public office for private gain). Distinguishing between these roles is part of the contextual understanding required as part of the Data Mining process. Additionally, the limitations of the source material also need to be borne in mind: in monitoring global broadcasts, certain very prominent individuals are likely to permeate many stories, even where their involvement is tangential. This is likely to be the case with Barack Obama, who is the second ranked individual in the corruption corpus with 7,149 mentions. The highest ranked individual with 18,295 mentions is Hamid Karzai, President of Afghanistan since 2004. Nouri al-Maliki, Prime Minister of Iraq since 2006, has 6,218.

The prevalence of Obama could be due to US involvement in Afghanistan and Iraq and his name may be cited with reference to more general policy concerns. However, American authorities have mounted ‘increasingly confrontational anti-corruption investigations’ of Karzai’s inner circle, and Obama entered office with a list of priorities for Karzai to address, including nepotism and corruption (Rashid 2011: 2-5). Karzai himself has admitted that ‘there is corruption in the whole

system' (Ghufran 2007: 90); as such, it is unsurprising to see such a high level of discussion in the corruption corpus for Karzai and Afghanistan. Whilst the source material may have a disproportionate focus on Afghanistan, it is also clear that there is awareness of corruption and that it features heavily in the local broadcast media.

The data can also be analysed for organisations, institutions and sectors. For example, Airways and Airlines appear a total of 431 times, Brewery 20 times, Circus 49 times and Drilling 97 times. Each can be disaggregated with minimal effort by year and country, or whichever parameter is required. A rough approximation of discussion by sector can also be constructed, by aggregating the industries and institutions within, for example, Finance. Here, 'banks', 'savings' and 'treasury' (using the singular and plural of each) are measured together with 'finance' and 'financial'. As a result, Finance records a total of just under 32,000 mentions, Military just under 25,000, Infrastructure 20,000, Education 15,000 and Health 12,000. Legal, which is likely to cover many other sectors (as part of a prosecution process, for example) records over 40,000. Full tables with breakdowns are presented in table 1 in appendix 2.

This example is somewhat crude and simply illustrates an example application of Data Mining; an organisation wishing to analyse corruption by sector would draw up a far more comprehensive list, would analyse the results in the appropriate context, and, as with all analyses, would consider the limitations of the source material. Furthermore, as with the country scores, the sector information is most accurately interpreted with a more specific question or goal in mind – such as whether there is public discussion of corruption in the education and health sectors – rather than comparing sectors directly, where the relative values are misleading to compare. Arguably, Data Mining is more useful and accurate when the scope is as narrow as possible – for example, to track the discussion about one particular company in a particular country, or allegations about bribery or vote rigging.

5.4. Semantic Analysis: Corruption

Whilst the country index allows us to see the frequency of geographic references in the data, and context analysis affords an insight into people and institutions, a closer look at the words immediately surrounding the word 'corruption' can help us to understand the usage of the term in broadcast media. In this particular methodology, a 'token region' analysis is a surrogate for a sentiment analysis – instead of analysing the tone of a text (using a weighted dictionary to assign

values to ‘positive’ and ‘negative’ words) – we see which words are being used in conjunction with corruption. The advantage of sentiment analysis is that the whole article can be labelled with a particular tone, and thus tones can be tracked over time; however, this requires specialist technical expertise and a weighted dictionary appropriate to the corpus. Token region analysis examines the text ‘as it is’, without recourse to external references and letting the inherent word order tell the story.

The word ‘corruption’ appears 77,892 times in the five-year corpus. The first analysis looked at the words immediately either side of the word ‘corruption’. Table 6 shows a breakdown of which words appear in each year by percentage; most notable is the consistency of ‘anti-corruption’ and ‘fight corruption’, together forming approximately 18-19 percent of corruption word combinations in each year – we can draw the tentative conclusion that a fifth of broadcast reporting in the corpus is concerned with combating corruption. Also noteworthy is the lack of ‘political’ or ‘economic’ corruption; instead, more references are made to administrative corruption. An examination of the two words preceding ‘corruption’ (see table 2 in appendix 2) found ‘administrative’ preceding ‘corruption’ in 5.8% of instances, ‘government’ 2.4%, ‘financial’ 1.8%, ‘economic’ 1.4% and ‘political’ 0.7%. The overall theme of each article may, of course, be related to corruption within economic or political areas, but analysis suggests the *nature* of the corruption itself is often administrative.

Table 6: Ranking of word occurrences immediately before or after 'corruption', percent

	2012-13	% of Corruption (17062)	2011-12	% of Corruption (16407)	2010-11	% of Corruption (14189)	2009-10	% of Corruption (15632)	2008-09	% of Corruption (14602)
1	anti	10.8	anti	11.8	anti	9.9	anti	9.7	anti	11.1
2	fight	8.2	fight	7.0	fight	8.8	fight	8.0	fight	6.9
3	administrative	6.0	administrative	5.6	administrative	5.1	administrative	4.9	administrative	4.7
4	cases	3.2	cases	3.3	fighting	3.5	fighting	3.4	commission	3.2
5	fighting	3.1	fighting	3.1	cases	2.8	cases	2.6	fighting	2.8
6	crime	2.7	combating	2.2	combating	2.6	commission	2.5	charges	2.8
7	government	2.2	country	2.2	crime	2.4	government	2.4	cases	2.6
8	involved	2.1	government	2.1	government	2.0	charges	1.9	economic	2.4
9	case	2.0	crime	1.8	commission	1.8	anticorruption	1.9	crime	2.4
10	country	1.8	involved	1.7	said	1.8	crime	1.9	government	2.3

The second analysis looked at words in the same sentence as the word corruption. The median length of a journalistic sentence is 15 words (Kornai 2008: 187) – therefore, to gain an approximation of a sentence, the analysis looked at seven words either side of the word 'corruption'. The semantic sentence findings largely mirrored the immediate word analysis, although 'Afghan' and 'president' feature more prominently, and 'government' overtakes 'administrative'. 'Government' appears in the same sentence as 'corruption' in 11.6% of instances, and 'anti' appears in the same sentence as 'corruption' in 11% of instances. This suggests strong linkages between corruption and the state and, at first glance, perhaps between anti-corruption and the state. However, when 'government' replaces 'corruption' in the same semantic sentence analysis (also using the corruption corpus), 'anti' appears in the same sentence as 'government' in only 1.1% of occurrences ('fight' appears even fewer times; 'corruption' appears 6.6%). Therefore, whilst 'government' and 'anti' often appear in the same sentence as 'corruption', they do not appear together. This suggests that, in most of the broadcast material in the corpus, the government is not associated with anti-corruption efforts. However, such a conclusion is tentative – the words 'official' and 'said' also appear frequently with 'government' and the lack of specific words together does not necessarily imply a lack of specific meaning.

Table 7 presents a summary of word frequencies seven words either side of 'corruption' and 'government'.

Table 7: Comparison of top 5 rankings of word occurrences within 7 words either side of 'corruption' and 'government', 2008 – 2013 (from 'corruption' corpus)

	corruption	77892	government	144419
1	government	9071	afghan	11324
2	anti	8552	corruption	9175
3	fight	7856	said	8746
4	words	5187	says	8262
5	said	4944	words	7779

5.5. A Case Study of Kenya and Corruption in 2009

Kenya experienced a peak in discussion about corruption in the BBC Monitoring Reports in 2009 – 2010. As an example of how Data Mining can be particularly useful on a smaller, more defined data

set, a new corpus of 186 articles featuring both ‘Kenya’ and ‘corruption’ was analysed. These cover a six month period from February 2009, and allow a deeper, more detailed application of the Text Mining methodology.

Because the corpus is considerably smaller than the primary one used in this study, it was feasible to include N-Grams (words that frequently appear together) in the analysis – for example, ‘president_kibaki’ features often. An experimental text network analysis was also conducted; because this is an extension to the basic methodology the process and findings are analysed separately in appendix 3.

Whilst the list of most frequently used words contains many that we would expect to see - ‘africa’, ‘government’ (as this features prominently in the larger corruption corpus), ‘shillings’ (the Kenyan currency) – there are other words that may give a clue as to the spike in discussion in 2009. ‘Maize’ features 247 times, and this can be linked to a high profile scandal in Kenya in 2009 that implicated several government ministries and the illegal selling of maize. This was investigated by the Kenya Anti-Corruption Commission, leading to the suspension of the Minister of Agriculture, William Ruto (Transparency International Kenya 2013; Ross 2010). ‘Ruto’ and ‘minister’ also feature prominently. Former Minister of Justice Martha Karua is mentioned 150 times; she resigned her post during this period, but also had a high-profile clash with Ruto over the maize scandal in February 2009 (Namunane & Mugonyi 2009).

Last, a previous, high-profile corruption incident – the so-called ‘Anglo-Leasing’ scandal – is also featured over 100 times, despite being exposed in 2002. This is likely due to ‘callback’ references, where articles refer to a noteworthy precedent – the Anglo-Leasing scandal was the centrepiece of the work of anti-corruption official John Githongo, who was forced into exile after exposing the involvement of senior officials in the scandal (Rønning 2009: 163).

5.6. Five Year Bribery Broadcast Discussion Index

Using the smaller corpus of 3,939 articles on bribery – also drawn from BBC Monitoring Reports – a BDI for bribery was constructed. The purpose of a separate corpus and index was to examine one particular manifestation of corruption to see whether the same countries emerged at the top of the list. Using firm and household survey data, the World Bank Institute estimates that global bribes total one trillion US dollars per year (Rose-Ackerman in Svensson 2005: 20). Table 8 presents the

top 20 countries in the index; over the five year period Afghanistan heads both the corruption and bribery indexes, although the Czech Republic climbs from 15 in the corruption index to 5 in the bribery index and Hong Kong and Kazakhstan are new entries.

The OECD (2013: 8) notes that there have been several high profile bribery cases in the Czech Republic since the country enacted an Anti-Bribery Convention in 2000; this may explain the high ranking. Similarly, there have been high profile cases of arrests for bribery in Hong Kong (BBC 2012). In Kazakhstan, legislation criminalising bribe-taking by officials was enacted in this time period (OECD 2011: 4).

Table 8: Top 20 countries in BDI for 'bribery', 2008 – 2013

	Country	Count
1	Afghanistan	4113
2	China	3472
3	Pakistan	1747
4	Iran (Islamic Republic of)	1423
5	Czech Republic	1040
6	Iraq	1001
7	Syrian Arab Republic	614
8	Kazakhstan	490
9	Kenya	472
10	Hong Kong	448
11	Israel	417
12	Nigeria	415
13	India	370
14	Russian Federation	357
15	Sudan	356
16	Egypt	343
17	Lebanon	336
18	Somalia	314
19	Turkey	312
20	Libya	276

5.7. Measures of correlation: the BDI and External Datasets

This section compares the Broadcast Discussion Index to two existing datasets. The most widely referenced ranking of corruption is Transparency International's Corruption Perceptions Index (CPI)

– this is compared to the BDI to see whether the level of discussion of corruption in broadcast media acts as a proxy for the perceived level of corruption. Do countries with higher perceived levels of corruption also discuss corruption more in broadcast media? The BDI is then compared to the Freedom of the Press rankings produced by Freedom House: does a not-free press restrict discussion of corruption?

5.7.1. The CPI and BDI on Corruption

CPI results were analysed from 2009 and 2012, and comparable data was found for 164 countries across these two datasets and the BDI. Table 9 shows the top ranked countries for each index; BDI data was taken from 2011-12 rather than 2012-13 owing to a time lag between the CPI data being collected and reported – the year under analysis in each index should be approximately comparable. Across all countries the correlation coefficient between BDI and CPI in 2011 is 0.15 by ranking and 0.22 by absolute score – there is almost no correlation between the data sets. Narrowing the selection to sub-Saharan Africa gives a slightly higher correlation coefficient of 0.43, indicating a positive relationship.

Table 9: 10 'most corrupt' countries in CPI and 10 'most discussed' countries in BDI, 2011 data

CPI (BDI)	Country	BDI (CPI)	Country
1 (1)	Afghanistan	1 (1)	Afghanistan
2 (31)	Somalia	2 (34)	Pakistan
3 (15)	Sudan	3 (38)	Iran (Islamic Republic of)
4 (43)	Myanmar	4 (7)	Iraq
5 (72)	Turkmenistan	5 (28)	Syrian Arab Republic
6 (60)	Uzbekistan	6 (90)	China
7 (4)	Iraq	7 (15)	Libya
8 (119)	Burundi	8 (53)	Egypt
9 (100)	Chad	9 (111)	Jordan
10 (105)	Haiti	10 (130)	Israel

Source: Transparency International 2012b: 3

A comparison of changes in score over a comparable three year period (2008-2011) yields a correlation coefficient of 0.16; the degree of movement in rankings for a country in the BDI bears little relation to movements in the CPI over a corresponding time frame. Of the 107 countries whose

scores increased over three years in the BDI, only 26 saw the perception of corruption increase in the CPI. However, of the 54 BDI scores that went down over three years, 43 of these saw their CPI score improve. This may suggest – albeit tentatively (the CPI only registers small changes over three years) – that when a country sees a reduction in the number of mentions in the BDI for corruption, the level of perceived corruption also decreases (Transparency International 2009 & 2012b: 3).

Lastly, the divergence was measured. This examines the difference in the relative rankings between countries in each index, in 2011. A score of zero means the country was equally ranked in each index; for example, Afghanistan, which heads both indexes, scored zero. Table 10 shows countries with the closest divergence; however, many of the rankings were wildly different – for example, the BDI ranks China 84 places higher than in the CPI. This may hint at reporting bias within the BBC Monitoring Reports, or a comprehensive fight against corruption in Chinese society (leading to a higher level of media discussion and a relatively better corruption perception score), or perhaps that the CPI is undervaluing the extent of corruption in China – we cannot be sure. We can, however, conclude that corruption is a topic of discussion within China, and accordingly it is within the 'public mind'.

Table 10: Closest 'divergence' between BDI and CPI rankings, 2011 data

Country	Divergence	BDI Ranking	CPI Ranking
Panama	-10	98	88
Cape Verde	-9	138	129
Uganda	-8	51	43
Liberia	-8	103	95
Uruguay	-5	152	147
Trinidad and Tobago	-4	96	92
Afghanistan	0	1	1
Ethiopia	0	57	57
Yemen	1	18	19
Kenya	1	30	31
Niger	1	58	59
Iraq	3	4	7
Libya	8	7	15
Indonesia	8	46	54
Luxembourg	10	143	153

Source: Transparency International 2012b: 3

The lack of significant correlation between the CPI and BDI is not surprising – after all, the two indexes are measuring different factors. However, the comparison is useful because both are attempting to capture different facets of the same phenomenon – perceptions of how corrupt a country is, and the level of media discussion about corruption. The comparison suggests that countries that are perceived as being more corrupt do not necessarily have an equally high level of public discussion about corruption.

5.7.2. Freedom of the Press and BDI on Corruption

The Freedom of the Press rankings assign a label of 'free', 'partly free' and 'not free' to 197 countries and territories. The most recent rankings for the Freedom of the Press index and the BDI were compared; all but two of the ten highest ranked BDI countries are designated 'not free' (Egypt and Turkey, in eighth and tenth place respectively, were designated 'partly free') (Freedom House 2012).

It is somewhat problematic to analyse the relationship between these two indexes. An initial prediction may anticipate the highest level of discussion of corruption in countries with, one, the greatest actual level of corruption and, two, the greatest level of press freedom. However, we must remember that discussion can be both positively and negatively frame the state, and thus greater media freedom should not necessarily be expected to be correlated with, or cause, discussion of corruption.

It could be (and is perhaps likely) that the countries with low press freedom are limited from broadcasting *even more* on corruption; in other words, these countries still record high levels of discussion on corruption *despite* restrictions on press freedom. Alternatively, the sources used by BBC Monitoring may largely circumvent the restrictions on mainstream press by, for example, including blogs and underground, opposition newspapers.

The relationship between press freedom and discussion of corruption is thus ambiguous in this study – the BDI methodology does not discriminate between what may be popularised as propaganda, and objective reporting. Irrespective of tone, mentions of corruption are taken as corruption being in the public mind and being considered an issue. We can see that a restricted press does not lead to substantive restrictions on discussion of corruption.

In summary, the comparison of the Broadcast Discussion Index to other datasets demonstrates that the BDI is a unique measure in its own right; it does not act as a proxy for perceptions of corruption or levels of media freedom, but a complement to these measures to gain a more multi-faceted understanding of corruption as a development issue.

6. Conclusions

The purpose of this research was to explore the potential for small civil society organisations to adapt Data Mining techniques for international development applications. There has recently been a significant increase in the importance placed upon Big Data for development work, yet Big Data techniques are not readily accessible to smaller organisations lacking technical expertise and hardware. Big Data has successfully transitioned from a primarily commercial usage to a development-focussed one, being applied to real-world social, health and economic issues. This research has examined whether a similar transition is viable for Data Mining – that is, can a form of data analysis primarily used for email filtering and analysing product reviews on websites be adapted as a tool for development organisations needing accurate insights into real-world problems but lacking technical resources.

This research developed a straightforward Data Mining methodology using open-source software. Corruption was selected as an example of a development issue – a difficult phenomenon to measure and a challenging application of the methodology. Close scrutiny of the data source is an essential component of the process, and the BBC Monitoring Reports archive proved to be expansive but had a questionable editorial bias – future applications may prefer a more 'narrow but deep' archive, such as the complete broadcast media of a specific region. Contextual interpretation is a second essential component, and this requires extensive cross-referencing and application of theory (in this example, broadcast media was treated as a manifestation of the 'public mind').

The results demonstrated a workable implementation of a Data Mining process. The quantitative methodology underpins a constant process of 'discovering apparent structure and interpreting that structure', and the results were framed within the limitations of the data source (Hand 2000: 177). In this example – using the BBC Monitoring Reports to measure the level of discussion about corruption – possible bias in the source material towards countries of political interest (such as Afghanistan, Pakistan and Iran) meant that cross-country comparisons were less trustworthy, and the most valuable information was found from closer analysis of countries, regions and concepts. A particular strength of the application is the degree of customisation possible – for example, disaggregating the data set by time, location or any other parameter useful for analysis enables a richer understanding of the issue. This does, however, necessitate careful planning and delineation of the analysis beforehand, and a clear understanding of the purpose and research questions –

without these, Data Mining simply becomes ‘fishing’, a reputation that this study has challenged as unfair.

The insights possible are relevant and accurate. For example, the finding that anti-corruption efforts are rarely associated with governments may be relevant to an organisation lobbying for an anti-corruption commission to be established, or the finding that corruption is often seen as administrative could inspire a campaign to increase the accountability and spot-checking of bureaucratic processes. The accuracy was confirmed by the tests in section 4, and the case study of Kenya, where several corruption scandals and the key individuals concerned were identified. A key conclusion is that the most relevant and accurate insights are most likely to arise when the methodology is used (and adapted) by experts in a particular field – a sexual discrimination campaigner, or a food security researcher – who can both frame the study questions and interpret the results; a sound technical methodology is a complement rather than a substitute for specialist local knowledge.

These results are encouraging, but the applicability for Data Mining as a tool for small organisations to gain novel insights needs to be tested through extensive, real-world application. Different development issues need to be tested with different data sets. Training guides and case studies should be created. The previous barriers for gaining such novel insights were technical expertise and technical capacity – these are now mostly obsolete for smaller analyses. Furthermore, never before have access to cheap computing power (this case study was completed on a laptop) and the availability of open data sets on the internet (boosted further by the rise in ‘data philanthropy’) evolved and converged to the point today where financial barriers have also mostly disappeared.

Tim Berners-Lee, founder of the World Wide Web, has stated that efforts are needed in both engineering and language development so that Data Mining can be made accessible to all without requiring computing resources that only the largest companies can afford (Smith *et al* 2006: 1682). Yet a stripped down, basic Data Mining methodology outlined in this study can afford smaller organisations novel insights without the technical expertise or large data servers needed for larger scale analyses. Success is dependent on choosing the appropriate technique, applying contextual interpretation and, most importantly, careful selection of the data set. The data sets are necessarily smaller than those used for ‘Big Data’ analysis, but the techniques mirror those used by large companies and facilitate the discovery of new knowledge – the process of turning words to numbers and querying and interrogating these numbers can uncover useful information and hidden patterns

that are otherwise inaccessible. The data becomes information, and with interpretation this information becomes knowledge: Data Mining can allow a data set to become more than the sum of its parts.

7. Bibliography

Aidt, T. (2009). Corruption, institutions, and economic development. *Oxford Review of Economic Policy* 25 (2), 271-291.

Austin, K. (2002). Illicit Arms Brokers: Aiding and Abetting Atrocities. *The Brown Journal of World Affairs* 9 (1), 203-216.

Baesens, B., Mues, C., Martens D., and Vanthienen, J. (2009). 50 Years of Data Mining and OR: Upcoming Trends and Challenges . *The Journal of the Operational Research Society* 60 (Supplement 1), 16-23.

Bardhan, P. (1997). Corruption and Development: A Review of Issues. *Journal of Economic Literature* 35 (3), 1310-1346.

BBC (2012). Kwok brothers arrested in Hong Kong on bribery charges. *BBC News Website*, 29 March.

BBC Monitoring (2012). *Case Study: Predicting Russian election protests.* London: BBC Monitoring.

BBC Monitoring (2013). Welcome to BBC Monitoring Online. www.bbcmonitoringonline.com/mmu/. Accessed 2013-05-16.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science* 16 (3), 199–231.

Campos, N., Dimova , R., and Saleh, A. (2010). Whither Corruption? A Quantitative Survey of the Literature on Corruption and Growth. *Forschungsinstitut zur Zukunft der Arbeit Discussion paper series* 5334. Bonn: IZA.

Castells, M. (2010). Communication power: mass communication, mass self-communications, and power relations in the network society. In Curran, J. (Ed.) *Media and Society.* London: Bloomsbury, pp: 3-17.

Charbonneau, L. (2013). Exclusive: U.N. monitors see arms reaching Somalia from Yemen, Iran. *Reuters,* 10 February.

Costa, S. (2012). Do Freedom of Information Laws Decrease Corruption? *The Journal of Law, Economics, and Organization* (Advance Access), 1-27.

Dumbill, E. (2012). *Big Data Now: 2012 Edition.* California: O'Reilly Media, Inc.

Franklin, J. (2005). Review: The Elements of Statistical Learning: Data Mining, Inference and Prediction, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Mathematical Intelligencer* 27 (2), 83-85.

Freedom House (2012). *Freedom of the Press 2012: Global Press Freedom Rankings*. Washington, D.C.: Freedom House.

Galtung, F. (2006). Measuring the immeasurable: boundaries and functions of (macro) corruption indices. In Sampford, C., Shacklock, A., Connors, C. and Galtung, F. *Measuring corruption.* London: Ashgate, pp: 101-130.

Ghufran, N. (2007). Afghanistan in 2006: The Complications of Post-Conflict Transition . *Asian Survey* 47 (1), 87-98.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques.* Massachusetts: Elsevier.

Hand, D., Blunt, G., Kelly, M., and Adams , N. (2000). Data Mining for Fun and Profit. *Statistical Science* 15 (2), 111-126.

Hilbert, M. (2013). *Big Data for Development: From Information to Knowledge Societies.* Pre-published version. Accessed 2013-05-16 from: http://ssrn.com/abstract=2205145.

Hussmann, K. (2011). Addressing corruption in the health sector: Securing equitable access to health care for everyone. *U4 Issue* 2011 (1).

Islam, R. (2003). Do more Transparent Governments Govern Better? *World Bank Policy Research Working Paper* 3077. Washington, D.C: The World Bank.

Janert, P. (2010). *Data Analysis with Open Source Tools.* California: O'Reilly Media, Inc.

Jeske, D. and Liu, R. (2007). Mining and Tracking Massive Text Data: Classification, Construction of Tracking Statistics, and Inference under Misclassification. *Technometrics* 49 (2), 116-128.

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2006). Measuring Corruption: Myths and Realities. *Development Outreach.* Washington, D.C: The World Bank.

Khan, W. (2000). Data Mining for Fun and Profit: Comment. *Statistical Science* 15 (2), 127-130.

Kirkpatrick, R. (2012). Big Data for Development. *Big Data: Preview of articles from premier issue*, 6-7.

Koren, J. (Ed.) (1918).The History of Statistics, Their Development and Progress in Many Countries. New York: Macmillan.

Kornai, A. (2008). *Mathematical Linguistics.* London: Springer.

Krallinger, M. and Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. *Genome Biology* 6 (224), 1-8.

Lambert, D. (2003). What Use Is Statistics for Massive Data? *Lecture Notes-Monograph Series. Crossing Boundaries: Statistical Essays in Honor of Jack Hall* 43, 217-228.

Leetaru, K. (2010). The scope of FBIS and BBC open source media coverage, 1979–2008. *Studies in Intelligence* 54 (1), 51-71.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large–scale human behavior using global news media tone in time and space. *First Monday* 16 (9).

Leetaru, K. (2012). *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content.* New York: Routledge.

Letouzé, E. (2012). *Big Data for Development: Challenges & Opportunities.* New York: UN Global Pulse.

Locksley, G. (2009). The Media and Development: What's the Story? *World Bank Working Paper* 158. Washington, D.C: The World Bank.

Miner, G., John, E., Hill, T., Nisbet, R., Delen, D., Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Massachusetts: Elsevier.

Namunane, B. and Mugonyi, D. (2009). Chaos in cabinet as Karua and Ruto clash. *Daily Nation [Kenya].* 17 February.

OECD (2011). *Second Round of Monitoring: Kazakhstan Monitoring Report.* Paris: OECD.

OECD (2013). *Phase 3 report on implementing the OECD Anti-Bribery Convention in the Czech Republic*. Paris: OECD.

Ohlhorst, F. (2013). *Big Data Analytics: Turning Big Data into Big Money.* New Jersey: John Wiley & Sons, Inc.

Pande, R. (2007). Understanding Political Corruption in Low Income Countries. *CID Working Paper No. 145*. Massachusetts: Harvard University Press.

Paranyushkin, D. (2011). *Identifying the Pathways for Meaning Circulation using Text Network Analysis.* Berlin: Nodus Labs.

Pope, J (2000). *Confronting Corruption: the elements of a National Integrity System - TI Source Book*. Berlin: Transparency International.

Rashid, A. (2011). How Obama Lost Karzai. *Foreign Policy* March/April 2011.

Rønning, H. (2009). The politics of corruption and the media in Africa. *Journal of African Media Studies* 1 (1), 155-171.

Ross, W. (2010). Kenya officials step down amid corruption inquiries. *BBC News Website,* 13 February.

SIPRI (2013). SIPRI Arms Transfers Database. http://armstrade.sipri.org/armstrade/page/toplist.php. Accessed 2013-05-16.

Smith, A., Gerstein, M., Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Weitzner, D. (2006). Letters: Data Mining on the Web [with Response]. *Science* 314 (5806), 1682.

Solove, D. (2008). Data Mining and the Security-Liberty Debate. *The University of Chicago Law Review* 75 (1), 343-362.

Svensson, J. (2005). Eight Questions about Corruption. *Journal of Economic Perspectives* 19 (3), 19-42.

Transparency International (2009). Corruption Perceptions Index 2009. http://archive.transparency.org/policy_research/surveys_indices/cpi/2009/cpi_2009_table. Accessed 2013-05-16.

Transparency International (2011a). 2011 Corruption Perceptions Index: In Detail. http://www.transparency.org/cpi2011/in_detail. Accessed 2013-05-16.

Transparency International (2011b). Long Methodological Brief - Corruption Perceptions Index 2011. Berlin: Transparency International.

Transparency International (2012a). Annual Report 2011. Berlin: Transparency International.

Transparency International (2012b). Corruption Perceptions Index 2012. Berlin: Transparency International.

Transparency International Kenya (2013). Lest we forget past corruption cases. tikenya.wordpress.com/lest-we-forget-past-corruption-cases/. Accessed 2013-05-16.

UN (2013). United Nations Statistics Division - Standard Country and Area Codes Classifications. unstats.un.org/unsd/methods/m49/m49alpha.htm. Accessed 2013-05-16.

UNODC (2011). *World Drug Report 2011*. Vienna: UNODC.

UNODC (2012). *Corruption in Afghanistan: Recent Patterns and Trends.* Vienna: UNODC.

White, H. (2002). Combining Quantitative and Qualitative Approaches in Poverty Analysis. *World Development* 30 (3), 511-522.

WHO (2012). *Global Tuberculosis Report 2012*. Geneva: World Health Organisation.

WHO (2013). Global Health Observatory Data Repository: Reported Malaria Cases by country. http://apps.who.int/gho/data/node.main.A1364?lang=en#.UZTbmiorxpQ.mailto. Accessed 2013-05-16.

World Economic Forum (2012). *Big Data, Big Impact: New Possibilities for International Development.* Geneva: World Economic Forum.

8.1. Appendix One – The complete BDI: Corruption

Complete Broadcast Discussion Index for Corruption 2008 – 2013							
Country or area name	2012-13	2011-12	2010-11	2009-10	2008-09	Percentage Change 2008-2013	Mean Average
Afghanistan	18240	19438	7816	10453	8196	222.55	12828.60
Åland Islands	0	1	0	1	4	N/A	1.20
Albania	489	344	253	477	679	72.02	448.40
Algeria	338	259	219	95	148	228.38	211.80
American Samoa	0	0	0	2	1	N/A	0.60
Andorra	0	1	0	0	1	N/A	0.40
Angola	29	60	38	80	57	50.88	52.80
Anguilla	10	20	1	6	1	1000.00	7.60
Antigua and Barbuda	10	1	0	11	6	166.67	5.60
Argentina	34	78	66	16	21	161.90	43.00
Armenia	21	24	21	12	12	175.00	18.00
Australia	101	120	225	190	90	112.22	145.20
Austria	119	132	83	63	51	233.33	89.60
Azerbaijan	95	107	37	16	49	193.88	60.80
Bahamas	4	1	1	9	8	50.00	4.60
Bahrain	236	853	104	22	19	1242.11	246.80
Bangladesh	369	247	241	491	1556	23.71	580.80
Barbados	2	2	3	3	5	40.00	3.00
Belarus	72	87	83	28	11	654.55	56.20
Belgium	28	35	28	31	29	96.55	30.20
Belize	2	15	3	16	6	33.33	8.40
Benin	34	22	19	17	30	113.33	24.40
Bermuda	2	12	14	24	21	9.52	14.60
Bhutan	1	7	15	23	30	3.33	15.20
Bolivia	37	34	28	15	22	168.18	27.20

Bosnia and Herzegovina	1258	703	792	1058	581	216.52	878.40
Botswana	5	77	66	77	86	5.81	62.20
Brazil	139	138	139	31	59	235.59	101.20
Brunei Darussalam	11	16	9	3	8	137.50	9.40
Bulgaria	252	144	169	126	293	86.01	196.80
Burkina Faso	7	16	1	4	16	43.75	8.80
Burundi	12	29	53	76	26	46.15	39.20
Cambodia	29	37	70	163	78	37.18	75.40
Cameroon	362	340	14	11	9	4022.22	147.20
Canada	163	118	117	87	127	128.35	122.40
Cape Verde	1	13	6	4	4	25.00	5.60
Cayman Islands	21	0	2	11	15	140.00	9.80
Chad	17	48	18	57	34	50.00	34.80
Chile	34	31	6	20	37	91.89	25.60
China	3794	3040	3390	2548	2480	152.98	3050.40
China, Hong Kong	578	420	376	416	321	180.06	422.20
China, Macao	21	39	41	69	88	23.86	51.60
Colombia	54	78	94	130	86	62.79	88.40
Comoros	1	4	1	4	1	100.00	2.20
Costa Rica	5	1	13	1	1	500.00	4.20
Côte d'Ivoire	28	95	21	47	65	43.08	51.20
Croatia	626	1213	768	1060	564	110.99	846.20
Cuba	60	108	36	36	71	84.51	62.20
Cyprus	188	51	11	35	27	696.30	62.40
Czech Republic	1009	1209	696	608	779	129.53	860.20
Denmark	31	30	37	44	37	83.78	35.80
Djibouti	35	27	29	37	3	1166.67	26.20
Dominica	2	10	9	15	7	28.57	8.60
Dominican Republic	8	8	9	7	10	80.00	8.40
Ecuador	28	27	25	24	29	96.55	26.60

Egypt	1972	2184	2468	531	602	327.57	1551.40
El Salvador	5	6	5	9	1	500.00	5.20
Equatorial Guinea	13	12	13	1	1	1300.00	8.00
Eritrea	51	154	77	241	18	283.33	108.20
Estonia	116	72	35	12	39	297.44	54.80
Ethiopia	195	201	115	96	56	348.21	132.60
Faeroe Islands	1	0	0	0	0	N/A	0.20
Falkland Islands (Malvinas)	0	0	0	1	0	N/A	0.20
Fiji	4	16	50	81	81	4.94	46.40
Finland	21	79	33	39	27	77.78	39.80
France	531	595	348	327	346	153.47	429.40
French Guiana	0	2	0	0	0	N/A	0.40
French Polynesia	1	0	0	1	0	N/A	0.40
Gabon	11	10	6	7	16	68.75	10.00
Gambia	1	2	13	4	1	100.00	4.20
Georgia	189	84	80	128	168	112.50	129.80
Germany	644	688	328	285	251	256.57	439.20
Ghana	50	119	100	215	231	21.65	143.00
Gibraltar	1	3	3	12	1	100.00	4.00
Greece	285	458	115	115	93	306.45	213.20
Greenland	2	1	0	2	0	N/A	1.00
Grenada	5	7	6	4	30	16.67	10.40
Guam	0	1	0	0	0	N/A	0.20
Guatemala	10	3	3	5	1	1000.00	4.40
Guernsey	0	0	0	1	5	N/A	1.20
Guinea	40	75	89	134	173	23.12	102.20
Guinea-Bissau	2	2	8	9	3	66.67	4.80
Guyana	7	29	1	15	17	41.18	13.80
Haiti	69	43	70	51	52	132.69	57.00
Holy See	54	21	18	8	19	284.21	24.00

Honduras	6	26	2	11	4	150.00	9.80
Hungary	233	262	105	83	69	337.68	150.40
Iceland	1	14	5	29	4	25.00	10.60
India	1542	1562	978	557	507	304.14	1029.20
Indonesia	112	312	379	203	234	47.86	248.00
Iran (Islamic Republic of)	6446	6865	4996	2966	6146	104.88	5483.80
Iraq	3209	5139	3277	4876	4867	65.93	4273.60
Ireland	34	41	29	18	37	91.89	31.80
Israel	1256	1938	1447	886	1732	72.52	1451.80
Italy	373	402	141	91	138	270.29	229.00
Jamaica	57	38	14	35	69	82.61	42.60
Japan	421	450	251	267	323	130.34	342.40
Jersey	4	24	5	10	5	80.00	9.60
Jordan	977	1959	640	251	345	283.19	834.40
Kazakhstan	424	802	272	344	458	92.58	460.00
Kenya	489	565	701	1530	898	54.45	836.60
Kiribati	0	0	1	1	0	N/A	0.40
Kuwait	215	476	141	177	244	88.11	250.60
Kyrgyzstan	300	402	300	154	369	81.30	305.00
Lao People's Democratic Republic	22	20	23	4	8	275.00	15.40
Latvia	353	381	423	238	415	85.06	362.00
Lebanon	518	1116	601	799	695	74.53	745.80
Lesotho	1	3	3	3	22	4.55	6.40
Liberia	15	46	31	42	42	35.71	35.20
Libya	710	2430	376	83	77	922.08	735.20
Liechtenstein	0	3	3	10	10	N/A	5.20
Lithuania	259	126	175	48	48	539.58	131.20
Luxembourg	5	11	16	4	15	33.33	10.20
Madagascar	4	4	8	23	3	133.33	8.40

Malawi	40	36	49	32	41	97.56	39.60
Malaysia	74	56	124	107	227	32.60	117.60
Maldives	126	67	152	98	70	180.00	102.60
Mali	389	68	23	9	13	2992.31	100.40
Malta	2	12	8	4	5	40.00	6.20
Martinique	1	0	0	0	2	50.00	0.60
Mauritania	56	39	59	111	100	56.00	73.00
Mauritius	2	7	0	4	31	6.45	8.80
Mayotte	1	1	0	0	0	N/A	0.40
Mexico	55	67	29	12	35	157.14	39.60
Micronesia (Federated States of)	0	0	1	1	1	N/A	0.60
Monaco	2	3	3	0	5	40.00	2.60
Mongolia	35	29	45	26	39	89.74	34.80
Montenegro	265	500	496	253	464	57.11	395.60
Montserrat	0	5	1	6	3	N/A	3.00
Morocco	200	303	89	19	45	444.44	131.20
Mozambique	17	30	26	38	35	48.57	29.20
Myanmar	308	376	185	106	282	109.22	251.40
Namibia	5	7	19	83	150	3.33	52.80
Nauru	1	0	2	0	1	100.00	0.80
Nepal	52	30	54	57	538	9.67	146.20
Netherlands	62	75	43	93	78	79.49	70.20
New Caledonia	1	0	1	0	0	N/A	0.40
New Zealand	20	28	35	25	19	105.26	25.40
Nicaragua	6	16	20	14	5	120.00	12.20
Niger	198	184	184	280	8	2475.00	170.80
Nigeria	777	910	1084	1009	154	504.55	786.80
Niue	0	0	2	0	0	N/A	0.40
Norfolk Island	0	1	0	0	0	N/A	0.20
Northern Mariana Islands	1	0	2	0	1	100.00	0.80

Norway	107	105	69	63	54	198.15	79.60
Oman	37	207	19	7	18	205.56	57.60
Pakistan	12404	13147	6340	3780	2373	522.71	7608.80
Palau	1	0	0	21	0	N/A	4.40
Panama	110	51	21	24	30	366.67	47.20
Papua New Guinea	3	21	88	48	77	3.90	47.40
Paraguay	7	28	6	7	112	6.25	32.00
Peru	34	449	9	4	65	52.31	112.20
Philippines	120	248	332	195	187	64.17	216.40
Pitcairn	0	0	0	1	0	N/A	0.20
Poland	247	277	290	293	272	90.81	275.80
Portugal	24	79	22	33	11	218.18	33.80
Puerto Rico	1	4	2	2	4	25.00	2.60
Qatar	526	868	89	83	72	730.56	327.60
Republic of Moldova	14	48	45	27	10	140.00	28.80
Romania	631	453	141	117	123	513.01	293.00
Russian Federation	1929	1688	994	541	945	204.13	1219.40
Rwanda	50	44	120	219	148	33.78	116.20
Saint Helena	2	1	1	3	3	66.67	2.00
Saint Kitts and Nevis	2	12	0	8	6	33.33	5.60
Saint Lucia	12	26	25	9	15	80.00	17.40
Saint Vincent and the Grenadines	0	0	1	2	2	N/A	1.00
Samoa	0	0	0	2	1	N/A	0.60
San Marino	1	2	1	0	0	N/A	0.80
Sao Tome and Principe	4	6	8	2	2	200.00	4.40
Saudi Arabia	850	876	439	371	329	258.36	573.00
Senegal	28	17	26	14	41	68.29	25.20
Serbia	2658	1692	862	922	1092	243.41	1445.20
Seychelles	5	12	3	2	19	26.32	8.20

Sierra Leone	19	11	41	66	46	41.30	36.60
Singapore	47	74	154	108	142	33.10	105.00
Sint Maarten (Dutch part)	3	0	0	0	0	N/A	0.60
Slovakia	160	221	193	147	80	200.00	160.20
Slovenia	89	82	97	174	85	104.71	105.40
Solomon Islands	14	21	15	25	27	51.85	20.40
Somalia	1302	548	313	356	75	1736.00	518.80
Spain	146	121	83	76	87	167.82	102.60
Sri Lanka	31	35	27	67	58	53.45	43.60
State of Palestine	397	505	324	294	412	96.36	386.40
Sudan	508	1332	1181	1261	809	62.79	1018.20
Suriname	1	2	8	2	14	7.14	5.40
Swaziland	17	32	13	47	27	62.96	27.20
Sweden	111	47	63	72	67	165.67	72.00
Switzerland	145	79	77	52	77	188.31	86.00
Syrian Arab Republic	3688	4150	370	507	597	617.76	1862.40
Tajikistan	361	416	367	148	200	180.50	298.40
Thailand	65	175	144	229	296	21.96	181.80
The former Yugoslav Republic of Macedonia	348	553	312	326	361	96.40	380.00
Timor-Leste	2	3	21	6	7	28.57	7.80
Togo	25	14	10	11	15	166.67	15.00
Tokelau	0	0	1	0	0	N/A	0.20
Tonga	8	0	1	4	15	53.33	5.60
Trinidad and Tobago	25	59	20	53	30	83.33	37.40
Tunisia	558	966	855	18	24	2325.00	484.20
Turkey	1668	1686	764	699	922	180.91	1147.80
Turkmenistan	117	114	47	17	32	365.63	65.40
Turks and Caicos Islands	43	6	9	145	24	179.17	45.40
Tuvalu	0	0	1	0	0	N/A	0.20

Uganda	301	254	352	312	339	88.79	311.60
Ukraine	208	128	119	91	116	179.31	132.40
United Arab Emirates	57	67	42	45	64	89.06	55.00
United Republic of Tanzania	82	60	130	138	215	38.14	125.00
Uruguay	2	6	7	3	4	50.00	4.40
Uzbekistan	146	165	71	56	43	339.53	96.20
Vanuatu	0	0	1	2	2	N/A	1.00
Venezuela (Bolivarian Republic of)	56	87	39	43	57	98.25	56.40
Viet Nam	126	116	194	274	295	42.71	201.00
Wallis and Futuna Islands	0	0	1	0	0	N/A	0.20
Yemen	675	1133	922	1164	156	432.69	810.00
Zambia	93	245	129	430	487	19.10	276.80
Zimbabwe	216	601	971	1115	1200	18.00	820.60

8.2. Appendix Two – Other tables

Table 1: Sector Totals, 2008 - 2013			
Infrastructure	2284	**Health**	6181
Road	3589	Hospitals	1001
Roads	2085	Hospital	2287
Highway	852	Doctors	1066
Highways	354	Surgery	169
Train	971	Medicine	415
Trains	107	Doctor	462
Transport	1872	*Total*	11581
Construction	5631	**Education**	9158
Housing	2018	School	3453
Total	19763	Classroom	38
Legal	9629	Teacher	450
Courts	3409	Teachers	1745
Court	27367	*Total*	14844
Total	40405	**Finance**	6150
Military	21783	Financial	11551
Fighter	372	Banks	2205
Bomb	1856	Bank	10748
Jet	186	Savings	396
Total	24197	Treasury	843
		Total	31893

Source: BBC Monitoring Reports - corruption corpus

Table 2: Ranking of word occurrences in 2 words before “corruption”, 2008 – 2013			
	Word	**Total**	**Percentage**
	corruption	77892	100
1	anti	8328	10.7
2	fight	6977	9.0
3	administrative	4482	5.8
4	fighting	2876	3.7
5	government	1906	2.4
6	involved	1642	2.1
7	combating	1569	2.0
8	crime	1500	1.9
9	financial	1375	1.8
10	said	1096	1.4
11	economic	1070	1.4
12	anticorruption	1016	1.3
13	combat	1013	1.3
14	says	928	1.2
15	officials	902	1.2
16	widespread	798	1.0
17	cases	747	1.0
18	accused	733	0.9
19	organized	694	0.9
20	alleged	651	0.8
21	people	594	0.8
22	political	583	0.7
23	oversight	551	0.7
24	charges	548	0.7
25	country	543	0.7

Source: BBC Monitoring Reports - corruption corpus

8.3. Appendix Three – Text Network Analysis Example

Section 5.5 presented a case study of Kenya during a six month period from February 2009, using a corpus of 186 articles discussing corruption in Kenya. This appendix provides a brief illustration of how Text Network Analysis can be applied to the same set of articles, and the potential insights it can provide.

Text Network Analysis establishes what words are connected to which other words, and how often. Each word forms a *node* in the network, and the connections between words form *edges*. Analysis can indicate which words are most important to the network as a whole, as well as the number of different words each word is connected to.[1] As with the Token Region analysis in the Text Mining methodology, the size of the region surrounding each word can be specified – usually two to five words. Such analysis allows increasingly advanced interrogation of text. However, the overall concept is quite intuitive:

> Network analysis is about understanding patterns in the relationships that connect objects and using those relationships to characterize actors based on the roles they play in the overall network environment. Two characters in a play might appear in an equal number of scenes, but one may spend the majority of their appearances with just two other characters, while the other has one scene each with every character in the play. A traditional vocabulary analysis would be hard-pressed to differentiate between these two characters, since their differences lie in their relationships with others, rather than their own actions. (Leetaru 2012: 87).

Text Network Analysis methodology initially mirrors that of Text Mining: texts are preprocessed, common and redundant words (stopwords) are removed, and the text is made lower case. This can all be done using the open source software AutoMap, which is also used to create the network – essentially a list of the *nodes* and *edges*, known as a *graph*. This graph can be visualised using another piece of open source software, Gephi. However, in this example the process was completed using a free online software tool from Nodus Labs – http://textexture.com/, which prepossesses and converts the text into a graph and uses Gephi to instantly visualise it.

The most important words (those with the highest 'betweenness centrality'), are larger in the resulting visualisation – these are not necessarily the words that occur most often, but those that are important junctions for 'circulating meaning' in the text. Words are then coloured by the

1 A word connected to many other different words is said to have a high 'degree centrality'. A word connected to many other words that themselves have many connections is said to have a high 'betweenness centrality'. The latter is particularly important and shows that the word is important in the text to *connect* different concepts. An excellent applied example (using the Unabomber Manifesto) is available at: http://bit.ly/YVdpHO

'community' they belong to – these are words that are densely connected together. The result is 'insight into the hidden agendas present within a text and better understanding of its narrative structure', and is particularly applicable to media monitoring (Paranyushkin 2011: 3).

Figure 1 is a screenshot of the resulting Text Network Analysis for the Kenya corruption corpus. Whilst a thorough analysis would require investigation of individual nodes (clicking on a node reveals the most important connections specific to that word) and exploration of communities (the application allows 'zooming in' for closer analysis), we can see that the most influential terms and concepts are 'Kenya', 'July' and 'monitoring'. These (apart from 'July', which would need further analysis to determine the significance) are unremarkable; 'Kenya' is self-explanatory, and 'British' and 'Monitoring' refers to the source headers. The words 'cabinet' and 'minister' reflect the findings from the Text Mining analysis. However, there appears to be a heavy emphasis on the law – 'tribunal', 'police', 'justice', 'court' – together with 'violence' and 'crime' that was not readily apparent from the Text Mining analysis. A process of cross-referencing and deeper research would now be needed to produce valid findings from the data; perhaps a local civil society organisation would focus further work on the interface between violent crime and corruption.

It is important to note that, as with Data Mining, contextual interpretation and scrutiny of the data source is important to the accuracy and usefulness of Text Network Analysis.

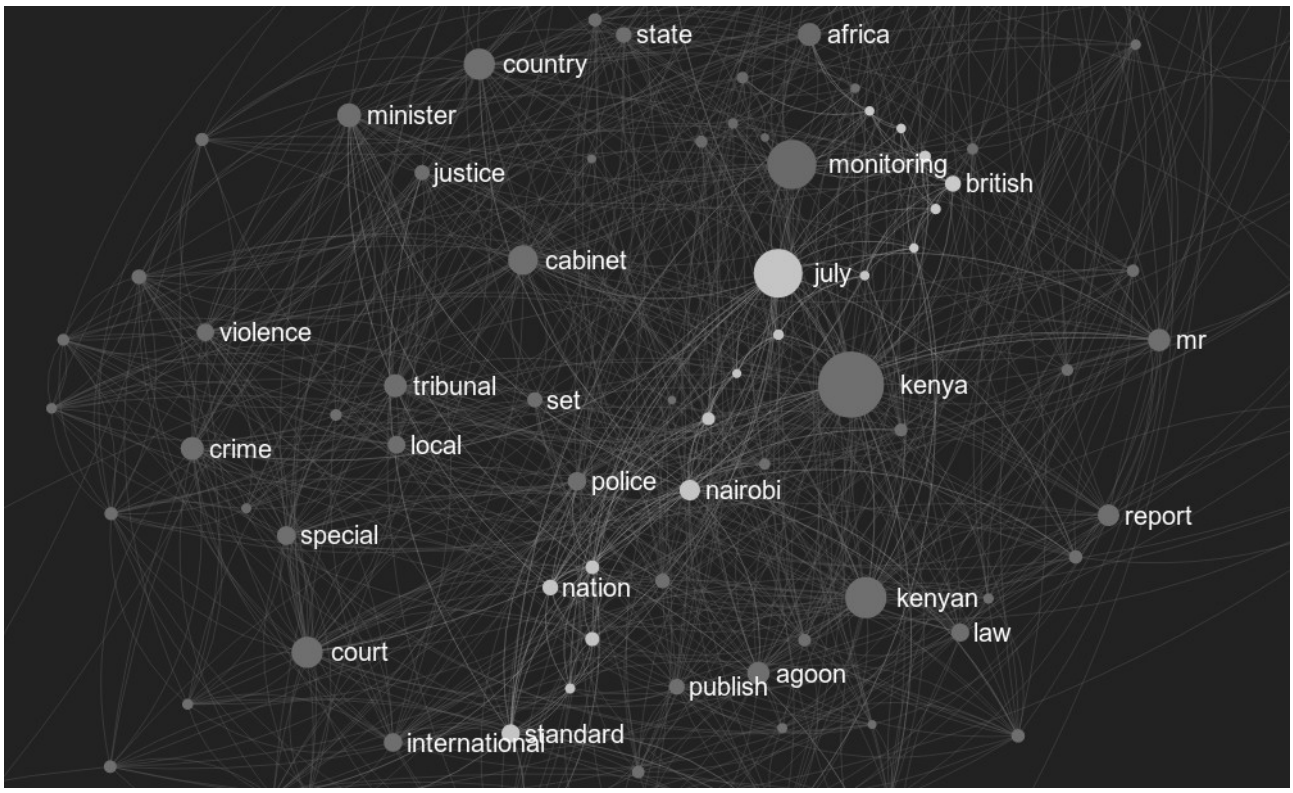

Figure 1: Text Network Analysis for Kenya corruption corpus, February–August 2009 (http://textexture.com/)

8.4. Appendix Four – Brief Glossary

BDI	The Broadcast Discussion Index is a ranking of places and other items of interest by the number of mentions they receive in a corpus of news reports
Big Data	Data sets so large they exceed the processing capacity of conventional database systems
Corpus	A collection of documents, used in this methodology as the data source
Corruption	The abuse of public office for private gain. This may include bribes, embezzlement of public funds, kickbacks in public procurement, or sale of government property
Data Mining	A process used to find patterns and interesting information within a data set, especially those that describe underlying structures in the data
Data Philanthropy	Private sector organisations stripping personal information from data records and sharing them for public analysis
Information Extraction	A strand of Text Mining which involves extracting specific information which can then be analysed; this analysis can result in new information that is predictive in nature
N-Grams	Combinations of two (or more) words commonly together, such as 'Prime Minister' or 'Sierra Leone'
Sentiment Mining	The analysis of the tone of a text using a weighted dictionary to assign values to, for example, 'positive' and 'negative' words
Stopwords	Words – such as articles, conjunctions and auxiliary verbs – that do not directly add meaning to the content and are commonly removed prior to analysis
Text Mining	The discovery of previously unknown information in the text that is not immediately obvious, by converting the text to numerical data. Commonly a means of analysing the communications of other people
Text Network Analysis	The analysis of which words are connected to which other words in a text. Can indicate the importance of certain words and concepts in the text
Token Region Analysis	Filtering significant words that exist within a certain proximity of a target word. Useful for examining the words that are commonly used in conjunction with a particular term