

Network Text Analysis in Computer-Intensive Rapid Ethnography Retrieval: An Example from Political Networks of Sudan*

Laurent Tambayong

California State University at Fullerton; ltambayong@fullerton.edu

Kathleen M. Carley

Carnegie Mellon University; kathleen.carley@cs.cmu.edu

Abstract: Advances in text analysis, particularly the ability to extract network based information from texts, is enabling researches to conduct detailed socio-cultural ethnographies rapidly by retrieving characteristic descriptions from texts and fusing the results from varied sources. We describe this process and illustrate it in the context of conflict in the Sudan. We show how network information can be extracted from vast quantities of unstructured text-based information using computer assisted processes. This is illustrated by an examination of changes in the political networks in Sudan as extracted from the *Sudan Tribune*. We find that this approach enables rapid high level assessment of a socio-cultural environment, generates results that are viewed as accurate by subject matter experts, and match actual historical events. The relative value of this socio-cultural analysis approach is discussed.

Keywords: social networks; text mining; network text analysis; rapid ethnography retrieval (RER); Sudan

*This work is part of the Rapid Ethnographic Retrieval project at the center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (under a Multidisciplinary University Research Initiative (MURI) grant, number N00014-08-1-1186. Additional support was provided by CASOS, the center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, or the U.S. government. We are also grateful for generous advice and thoughts from Jeffrey C. Johnson and Richard A. Lobban.

Introduction

Without a doubt there has been tremendous growth in the amount of information available on the internet, largely in unstructured or text form. This form includes news articles, blogs, scholarly writings, and crowd-sourced information of web 2.0. Following a crisis, such information provides the analyst with the possibility of rapidly assessing socio-cultural facets of the region of concern and information about issues of relevance. These include changes in political elite, beliefs, attitudes and much more. Such information is one traditional source of data used by Subject Matter Experts (SMEs), in addition to interviews, participant observation, and other forms of first-hand experience. However, canvassing and analyzing this open-source text data is extremely meticulous and labor-intensive as it involves copious reading samples. The sheer volume of texts suggests the need for a new and more automated approach (Alexa and Zuell 2000; Batagelj, Mrvar and Zaveršnik 2002; Corman et al. 2002). Hence, the potential exists to use semi-automated text mining and assessment techniques to rapidly process vast quantities of text-based information. Such automation will provide the analyst with, at least, a more accurate first pass understanding of a socio-cultural system and the changes that have been described in these texts. The growth of computing power and advances in language technologies are making this potential a reality. These advances mean that more accurate socially and culturally nuanced data can now be extracted from texts using computational techniques. These methods are especially helpful and efficient when dealing with abundant quantities of unstructured texts that cannot be analyzed individually by socio-cultural SME analyst in a timely, unbiased, and systematic manner (Corman et al. 2002). From the analyst's perspective, this means that technological support will be available for the identification of big social issues such as conflicts, which involve specific agents (agents or organizations) and events. Thus, such automated approach will result not only in labor efficiency but also in significant increase in the accuracy of socio-cultural analysis.

There have been previous developments for text extraction for information gathering and communication (Chakrabarti 2002; Manning, Raghavan & Schütze 2008). Examples include various methodologies such as *entity extraction* (Chakrabarti 2002), *content analysis* (Holsti 1969; Krippendorff 2004), *topic modeling* (Hofmann 1999, Blei, Ng & Jordan 2003), *entity resolution and event discovery* (Roth & Yih 2007), *theme analysis* (Landauer, Foltz & Laham 1998), *information retrieval* (Manning, Raghavan & Schütze 2008). More specific developments related to the network approach include *role discovery* (Wang et al. 2010), *link extraction* (Ramakrishnan, Kochut & Sheth 2006), *multi-dimensional text-analysis* (Lin et al. 2010; Zhang, Zhai & Han 2009; Ding et al. 2010), *meta-network* (Carley et al. 2009; Diesner & Carley, 2008).

With roots in information retrieval and text mining, Network Text Analysis (NTA) is an analysis method that encodes a network based on the relationships of words in a text (Carley 1997; Popping 2000). NTA systematically extracts and analyzes the relationships of words in a set of texts under the assumption that the relationships of language and knowledge could represent the mental model (Carley and Palmquist 1992) of people and their social and organizational relationships and structure (Sowa 1984). Its methodology is grounded on established methodologies for indexing the relations of words, syntactic grouping of words, and the hierarchical/non-hierarchical linking of words (Kelle 1997). NTA, under various alternative names, has been an active field of research as it allows for a clearer and more

efficient analysis of a large collection of text data (Carley & Palmquist 1992; Corman, Kuhn, McPhee, and Dooley 2002; Danowski 1993; Popping 2000, 2003; Popping and Roberts 1997; Ryan and Bernard 2000). The key difference between NTA and most work in the text analysis area is that it enables the extraction of both concepts and relations, codes the text as a network, and, depending on the analytic technique, may even extract multiple networks with different types of nodes and relations from texts. At one level, NTA relies on entity, link, and theme extraction to build semantic networks or mental models; i.e., networks of concepts. NTA allows for not only an efficient extraction of a large collection of data but also an ontology of the network relations of entities involved in the context of social structure (Carley 2002, 2003).

More advanced NTA techniques utilize event, role, and ontological classification schemes to recast this semantic network as a meta-network (Carley et al. 2009), which result in an ontologically defined set of networks (Diesner & Carley 2008). A meta-network (Carley 2002, 2003) represents the socio-cultural environment in terms of an ecology of networks linking who, what, when, where, why and how. As a result, NTA can extract meta-network information and is often used for cultural assessment (Carley 1994). This leads to our vision to create Rapid Ethnography Retrieval (RER) technologies that enable analysts to rapidly extract from vast quantities of text detailed cultural information in both an efficient and accurate manner (Diesner et al. 2012). Advanced NTA techniques support this vision by enabling the identification of meta-network data; i.e., extraction of not just concepts but of people, organizations and the relations among them, locations, events, issues of concern, and so on.

In this paper, we describe a specific part of RER: NTA methodology that enables meta-network to be extracted from vast quantities of unstructured texts-based information in a highly-automated computer-intensive process. We will show how this semi-automated technology can be embedded in a workflow that moves from text collection to data extraction to analysis. This workflow, and its automation, is the foundation of the proposed RER system. Although highly-automated, RER is not a magical black box – a holy grail – that allows a layperson to analyze the subject of the text set without any ethnography expertise of SMEs. The key point of this method is the increased manageability and standardization of analysis processes in dealing with abundant numbers of texts: it will allow for not only increased efficiency and accuracy but also a better understanding of the analysis as a complement to the traditional ethnography method.

We will elaborate how the SMEs' knowledge, judgment, and ethnographical insights are still an integral part of the standardization process embedded in the RER. To illustrate this approach, and show the gains by enabling the analyst to make use of an RER system, we will explore the changing political structure in the Sudan in a later section. Sudan is an interesting case application for this methodology for a number of reasons. First, vast quantities of unstructured texts are available; e.g., we have close to 40,000 articles just from the *Sudan Tribune* online. This vast amount of data is important as it increases the reliability of the NTA-generated political networks. Second, there are Sudanist SMEs whom we can draw on for the RER process, such as a group led by Richard Lobban. Employing SMEs' ethnography expertise are important, as we will elaborate later, not only in construction process or this methodology but also in validation process of the conclusions derived from this methodology. Third, Sudan has undergone major domestic conflicts resulting in changes in its political map. These changes allow us to

test if the RER process is capable of capturing the political changes. Therefore, these three aforementioned reasons are important as it allows us to demonstrate the applicability of the RER process through a comparison with actual changes in historical events. In particular, they allow us to compare the extracted information, in the form of NTA-generated political networks, to the actual historical events such that, when they match, it will serve as a proof of concepts that validates the method we use.

We begin our description by describing the RER process and the NTA methodology we employ. The role of SMEs' ethnography expert judgment in the text extraction process and the network generation process is discussed. Then, we describe the extracted data. We illustrate the value of the approach by presenting the network analytic results that we derive about the important political agents in Sudan as a longitudinal process parallel to actual historical events, results that are validated by ethnographer SMEs. Finally, we will conclude the results and discuss the impacts of NTA part or RER.

Methodology

RER: A Synergy of Human and Machine

The Rapid Ethnographic Retrieval (RER) system is an integrated data-to-model process system serving as an analytical tool in assessing change in socio-cultural systems. The RER system is envisioned as a human-in-the-loop system that enables ethnographers and those engaged in field studies to work with modelers and to reduce the time and barriers in moving from the detailed qualitative historical realm to the predictive quantitative forecasting realm. We envision that the RER system will eventually employ a number of distinct methodologies from web-scrapers and sensor feeds to open-source text-mining techniques to network analysis tools and possibly to agent-based simulations. Thus, RER will allow for not only increased efficiency and accuracy but also a better understanding and potential prediction, of the analysis on the regions of interests to complement the traditional ethnography method.

In this paper, we are describing only the first part of the RER system that we have developed. However, the methodology that we explain in this paper is not preliminary: it is the very foundation of the whole RER system. Here, we describe the central text mining and surrounding data cleaning methodology using Network Text Analysis (NTA): how over-arching corpus meta-network information is extracted from open-source text data. Other methodologies, such as more advanced network change detection methods and agent-based models are not described here and are subjects for forthcoming papers. We focus here on the data encoding step and the way in which Subject Matter Experts (SMEs) input plays an integral role in both the construction and validation processes. In this part of development, our goal is to identify the preferred coding in creating a more automated system and a more detailed workflow. These increase efficiency by reducing analysis time and increase accuracy by omitting subjective human errors, which in turn results in a better understanding of the region of interests.

Automated Data Extraction

At an abstract and very high level, Network Text Analysis follows a five-step procedure in extracting information from unstructured texts (Diesner and Carley, 2005):

1. *Concept Identification*: This step identifies concepts of interests and deletes irrelevant or meaningless terms.
2. *Entity Identification*: This step identifies the ontology; i.e., the set of defined entities into which the data will be coded. Note that step 1 and 2 are temporally interchangeable.
3. *Concept Classification*: This step maps the identified concepts of interest into the relevant entities using another type of thesaurus: meta-network thesaurus or ontology thesaurus.
4. *Meta-network extraction*: This step applies the aforementioned thesauri and delete lists and processes the texts to extract concepts, cross-classifies them into their ontological category, and then extract the relation among those concepts
5. *Graph and Analyze Data*: At this point, the text has been turned into a structured data set, specifically a meta-network that can be graphed and analyzed using available Social Network Analysis (SNA) measures.

Concept identification is a very involved process that includes deletion of irrelevant/meaningless terms and generalization of relevant concepts into the set of interest. For example, we transform typos into correct forms, remove plurals, resolve anaphors, delete common-yet-meaningless words (“a,” “the,” “and,” “of,” etc.), locate known and common n-grams, and employ thesauri listing common aliases for known political agents. Abbreviations are also attached to the full concept term in these thesauri. These thesauri are referred to as generalization thesauri. A generalization thesaurus is a two-columned collection that associates a concept with a corresponding higher-level concept (Burkart 2004:141-154; Klein 1997:255-261). An example would be associating *Bashir or General Bashir* with *Omar Hassan Al-Bashir*. Thesauri can be used to handle aliases, reduce specific concepts to more general concepts, and combine similar concepts.

We went through multiple iterations in constructing the generalization thesauri and identifying the sets of relevant concepts with help of Sudanist SMEs. One of the key difficulties, from a cultural inference perspective is handling g-grams. Consider the bi-gram *Ali Abdul*. Clearly not all instances of *Ali* and *Abdul* can be converted to as a meaningful bi-gram as they are too-common first names in Sudan even as a bi-gram *Abdul Ali*. We found that simply finding all bi-grams in the regional corpus, even only those that occur with a reasonably high frequency, tended to result in a large number of “meaningless” pairs that were simply a function of literary style (for example, *Ali Abdul* does not refer to a specific agent). Different writing styles result in different “noise” bi-grams, an artifact of automated and indiscriminating extraction process. To resolve this issue and validate the extracted bi-grams, we created a common n-grams list by a) processing many documents, b) looking at external resources, and employing SMEs to identify specific concepts for this region. Consequently, SME judgment and cross-reference to the current and historical events on subject-oriented ethnography are critical to identify relevant n-gram concepts that can be used in the generalization thesauri. This identification is comprehensive as construction of generalization thesaurus includes possible variations of the names of the agents,

including nicknames, aliases, and abbreviations. Meticulous consultations with SME ethnographers, Sudanist SMEs in this case, ensure the accuracy of the final lists and helped identify points whereas computer assistance could reduce the burden on the SME in abundantly repetitive tasks (i.e. repetitively identifying the exact same bi-gram noise in abundant different documents).¹ Although it is labor-intensive in the front end, this process ensures a systematically standardized and reliable way to repetitively extract information from abundant data such that it is not prone to errors caused by inconsistencies in human judgment: once the noises of interests are defined, they will be eliminated as their search is a highly-automated computer process.

Entity Identification is a theoretical-based step; i.e., it depends on the research questions of interests. For this paper, we use only one entity class (agents) as our purpose is to demonstrate the use of NTA within an RER system in the most direct and simplistic way. However, it is possible to include other entities such as organizations, tasks, locations, and knowledge. The complexity of the data processing and analysis increases nonlinearly as the number of entities increases. Thus, there is a trade-off that one needs to consider when deciding the number of entity classes. Since each entity class defines the number of modes, if there is a potential for a minimum of N-choose-2 networks that can be extracted where N is the number of entity classes. In our case, since our goal is to illustrate how the methodology works in the simplest way possible, we use N=1 and the networks of interest is only agent-by-agent (AA). Even only employing one entity, the complexity can increase: within each entity class there may be sub-groups. For example, in our cases our class of entity is agents, which consist of people. People are then further sub-divided into subclasses of Sudanese and Other. A second complication is that, in texts, entities may be referred to in the general or in the specific form; e.g., *president* is general and can refer to any president personas while *President Bashir* refers to only one specific persona. Although we coded for both, in this paper we utilize only the specific named entities that refer only to specific identifiable personas.

Concept Classification: This step involves creation of another type of thesaurus, the meta-network or ontology thesaurus. These meta-network thesauri are used to assign concepts to their corresponding entity class. High quality classification requires employing many classification venues to create these thesauri. Our multiple sources include historical meta-network thesauri developed on other projects including world-leader lists, machine learning techniques, parts-of-speech tagging, and specialized meta-network thesauri developed by Sudanist SMEs. In this paper, people are identified as agents. The concept classification in this case is clear as there is generally a clear cut that classify entities as agent. For a more complex ontological scheme with multiple entity classes, this proves to be more difficult as it is possible for concept to be classifiable into more than one entity class (Carley 2003). This problem is rectified by using meta-network attribute thesauri that have more than two columns such that it is possible to identify concepts into multiple classes of entity (Diesner and Carley 2005). Machine learning techniques are actually quite valuable for ontological classification; e.g., conditional random fields'

¹ In this paper, our SME is a team led by Richard A. Lobban who is a Sudanist SME, a faculty member of Anthropology and African Study at Rhode Island College, and an adjunct faculty at Carnegie Mellon University. For a comprehensive historical background of Sudan, consult Fluehr-Lobban, et. al. (2002).

techniques work well particularly for people, organizations and locations (Diesner and Carley 2008). In the end, however, we find that for between 5 and 10% of the concepts, entity classification is dependent on expert knowledge and ethnographic judgment. Future research should determine whether this fraction is a critical fraction or not. Similar to concept identification, this process ensures a systematically standardized and reliable way to extract information such that it is not prone to errors caused by inconsistencies in human judgment.

Meta-network extraction: This step is very computer-intensive and results in the extracted networks. We use AutoMap (Carley et.al. 2009a) to process the set of texts, applying the thesauri, and generate out a meta-network.² A large number of procedures are used as part of this process including proximity based mapping, parts-of-speech categorization, and so on. AutoMap searches the whole text set for the concepts previously defined in the generalization thesauri. AutoMap then builds a semantic network based on the generalization thesauri, delete lists etc., then it cross-classifies the concepts into their ontological categories using the meta-network thesauri, and stores the resulting system as a meta-network. This process causes all the predefined concepts in a text to be linked.

Graph and Analyze Data: At this point, the text has been turned into a structured data set, specifically a meta-network, which can then be graphed and analyzed. AutoMap generates one such meta-network per text. Since there are thousands of texts, there are thousands of networks. The next step is to create a frame to combine these results in a meaningful way. This frame is dependent on the research question. As we are interested in how the socio-cultural environment in Sudan has changed over time, we frame the results by year: one combined meta-network per year. There are many ways to fuse network data; we chose to employ a unioning procedure. A yearly compression enabled us to take a longitudinal perspective. Next, we use *ORA (Carley et al. 2009b) for graphing and analyzing the data.³

In addition to the aforementioned steps, we also identified a number of preliminary preprocessing steps to standardize the extraction process and to eliminate non-bearing concepts from across the texts (Carley 1993). First, we ran standard cleaning processes to clean the data from unwanted characters such as numbers and punctuations. Next, we applied a set of pre-processing thesauri containing pre-made standard and commonly-used delete list in English language (i.e. who, what, the, etc.). We also applied the stemmer from (Porter 1980), a process that converted each concept into its related morpheme to eliminate redundancy of concepts (Jurafsky & Martin 2000: 83, 654). All of these preliminary processing steps are done automatically using AutoMap. We tracked the order of processing, experimented with different orders, and identified the best order for reducing error in both what concepts were extracted and the connections among them.

² AutoMap is available for download at <http://www.casos.cs.cmu.edu/projects/automap/software.php>

³ *ORA is available for download at <http://www.casos.cs.cmu.edu/projects/ora/software.php> . Output from AutoMap is stored as dynetml files (an xml format for graphs of which graphml is a subset). Dynetml is directly readable by *ORA (Carley, et.al. 2007).

Degrees of Generalization and Connection

On the surface this processing sequence sounds straight forward. It turns out, however, that there are many coding choices that the analyst must make along the way that influence the results (Carley 1993). There are a number of cases where SMEs need to be involved in the initial run-in extractions for verification in the construction processes. We now describe key issues that arose and how we resolved these to create the RER system.

The first overarching key issue was degree of generalization. Coding choices can result in the extracted concept set being over- or under-generalized relative to the analysts needs. For example, imagine we are trying to extract all references to a specific persona. A *low extraction frequency* for this agent of interest occurs when the variations of its names, including nicknames and abbreviations, are not properly identified. Misspelling, another source of error, could also lower the extraction frequency. This under-extraction, in turn, results in missing links among agents, which in turn affects the characteristics of the network generated. To reduce this, we found it necessary to include common variation and misspelling of names of the agents (and organizations, locations or other named entities of interests in general case), as well as known pseudonyms or aliases. However, this does not reduce another under-extraction problem: pronouns (he or she) are often used to describe the agent previously discussed. Yet, it is also fair to assume that this kind of under-extraction problem is evenly distributed among extracted agents such that it will not affect the relative frequency. It is also possible to have a *falsely higher extraction frequency*. This over-extraction occurs, for example, when a generalization thesaurus includes very common nicknames or aliases for a particular agent who is equally applicable to other agents in the corpus. For example, the aliases of “William F. Clinton” in generalization thesaurus should include “Bill Clinton” but should not include only “Bill” as “Bill” is a very common nickname and might apply to other Bill’s such Bill Hancock or even Sudan Referendum Bill. If “Bill” were included as an alias of “William F. Clinton,” then *all* “Bill” unidentifiable to *any* specific agent or entity would then be associated with “William F. Clinton.” These false associations would cause “William F. Clinton” to have not only a falsely higher extraction frequency but also false links to other agents, affecting the characteristics of the network generated. We find, in general, for named entities that are n-grams, that single concepts should generally not be associated by generalization thesauri. However, sometimes, advanced machine learning processes can cross-catalog these terms during a second pass (e.g., Diesner & Carley 2008).

The second overarching key issue was the *degree of connection*. We found that link extraction was improved by using multiple link-extraction techniques and fusing the results. One technique is *proximity based*. This technique is widely used in the social sciences for text analysis. Proximity based techniques place a link between two concepts just in case they occur within some distance of each other; e.g., number of words, sentences, paragraphs. This process is highly automated and is not dependent on the context and substantive meaning of the sentence, paragraph, and article. We find that the media impacts this choice of window size e.g.: email, tweets, and email headers can be treated with the window size equal to the text; PowerPoint slides with window size equal to the bullet; and for newspaper data, scholarly articles, and blogs the most accurate connections are at the sentence level. However, this needs to be supplemented with secondary processing to handle lists. Further, it appears for named entities, such as people, organizations and locations, specialized paragraph level processing is

needed after anaphora resolution is employed. In general, the longer the window (more concepts) the higher the number of links (denser networks); whereas, the smaller the window the fewer the number of links (see Diesner & Carley 2005). Thus, the validity of the extraction is intrinsic to the window choice itself and nothing else.

A window size of 7 that we use link concepts within 7 concepts to one another. For example on window size, consider this sentence: "President Omar Hassan Al Bashir and Salva Kiir shake hands after a funeral service for John Garang in Juba." There are 19 words in this sentence. Based on the thesauri we earlier created, there are 10 identified concepts in this sentence: "President Omar Hassan Al Bashir," "and," "Salva Kiir," "shake," "hands," "after," "a," "funeral service," "for," "John Garang." However, "and," "after," "a," and "for" are automatically eliminated by the standard delete list. The concept "hands" is merged into its singular form, "hand," in the stemming process.⁴ Thus, after these processes, we have this row of concepts: "President Omar Hassan Al Bashir" "Salva Kiir" "shake" "hand" "funeral service" "John Garang." For our research here, we are only interested in concepts that are aliases of political agents. Thus, the concept "President Omar Hassan Al Bashir" will be linked to "Salva Kiir" and "John Garang" as they are within 7 concepts to one another.

Within this overarching issue, one can also question what type of link is produced. For example, it is a valid question to ask whether a link produced between two agents is positive or negative. Although it is theoretically possible, we do not distinguish between positive and negative link in this article. For our research in studying the dynamics of political agents of Sudan, the differentiation between positive and negative link is not critical. A well-known quote from *The Art of War* by Sun Tzu says "keep your friends close, and your enemies closer." The dynamics of political networks often result in changeable, often ambiguous, ties: a friend today might be an enemy tomorrow. Thus, it is not critical to define a link as positive or negative. As long as such link exists, it gives value to the linked agents because it enables the one to monitor to the other thus gives values.

Applying the RER: Illustrative Results

As we mentioned in the previous section, our focus in this article is on how corpus meta-network information is extracted from open-source text data. We have explained in the previous section how SME is integral in the construction process. We will now present illustrative results as a proof of concepts on how RER gives rapid high level assessment of a socio-cultural environment by generating results that are validated as accurate by ethnographer SME and match actual historical events. To provide for the most direct link of the corpus meta-network and its extracted open source text data, we will use the simplest example and focus only on the networks of the extracted agents. Presenting an example in terms of networks of agents allow to present readers with uncomplicated and intuitive analysis that can be substantiated with real historical events.

⁴ Although irrelevant in this paper as our focus is in agents, this process is relevant if, for different research, one wishes to link verbs to create a different kind of network of concepts.

SMEs need to validate the results using a frame that they understand. Our yearly meta-networks display results in forms of: a) visualization by year, b) top agents by year on various metrics, and c) over-all graph-level metrics and change in these. These are not in forms familiar to Sudanists SMEs as they are ethnographers and not social network analysts. Hence, we then create narrations that incorporate external documents, in which yearly data is elaborated longitudinally in term of rank order (ordinal). Sudanist SMEs then can validate if these narrations match real historical events. Before proceeding to the next section, we suggest readers who are unfamiliar with the recent developments in Sudan to see Appendix A for a brief background.

Data

We began with more than 40,000 newspaper articles in English language from the *Sudan Tribune* online for years 2003-2008. We chose articles in English due to structural issues in language. Built from our previous research, we have a library of thesauri required for general preliminary processes in English language. Different languages are likely to have different structures thus requiring different sets of thesauri for the extraction processes. This is especially critical for the preliminary processes in which different delete lists and stemming processes are required for different languages. Thus, English, and not Sudanese, is our language of choice for this research.

The number of articles varies by year: 2003: 2,932, 2004: 6,943, 2005: 3,828, 2006: 3,828, 2007: 5,815, and 2008: 9,266 articles. Applying the semi-automated RER process, the numbers of agents for the corresponding years are 28, 38, 33, 39, 41, and 39. These are political agents who have been reported in engaging in some political activities in Sudan.

Agent-Level Analysis

Herein, we describe and use the RER system and associated NTA process to examine political change in Sudan. Analysts in looking at data such as extracted by the RER system would ask, "who is key?" From the extracted agents, we select those who have had consistent and/or important roles in this conflict. We ask, across time, who stands out? We use four categories of SNA measures most relevant to power structure to determine the important agents: *Total Degree Centrality* (Wasserman and Faust, 1994), *Cognitive Demand* (Carley, et.al., 2009b), *Eigenvector Centrality* (Bonacich, 1987), and *Betweenness Centrality* (Freeman, 1979). These measures are most relevant to power structure because:

- (i) *Total Degree Centrality* is the most obvious and frequently used measured of power structure. It measures the total number of links of an agent in a network, which means that he is in the center of what is happening. The higher the score the more likely is an agent to receive and potentially pass on critical information that flows through the network (Wasserman and Faust, 1994:199). This measure identifies the formal leaders of the network.
- (ii) *Cognitive Demand* measures the total amount of cognitive effort expended by each agent to connect to other agents. The expended cognitive effort is inferred based on the agent's position in the network. Such agents may never become the formal leader of an organization yet they hold

very important positions in their networks as the de facto leaders. Thus, this measure identifies the informal/underground leaders of the network.

- (iii) *Eigenvector Centrality* reflects an agent's links to other well-linked agents. If an agent is linked to isolated agents without many links, such agent will score low on this measure although it might have many links itself. This is an important power measure as an agent linked to well-linked agents has a better access to spread and obtain information in a network. Agents with higher scores of Eigenvector Centrality could be critical when rapid communication is needed. This means that such agents are powerful due to their associations to those who are well-connected and thus powerful. Thus, this measure is about the power-through-association of an agent,
- (iv) *Betweenness Centrality* measures how well an agent is connected to other parts of a network. This measure tells us which agent in a network is the most central to the network as a whole. Betweenness Centrality measures the frequency of links passing through a single agent to be connected. Thus, it measures how an agent is a broker of indirect connections among all other agents in network, making it to be a powerful gatekeeper of information flow. Agents who occur on many shortest paths among other agents have highest Betweenness Centrality value. This measure identifies the agents who importantly broker/bridge factions.

The four categories of network-level measures (figure 1-4) show four important agents who stand out: Bashir, Taha, Mayardit, and Garang. Once identified, we are now interested in analyzing their networks of relationships to other agents. We focus on ego-centered longitudinal (across the years) roles of these agents because:

- (i) It shows us agents whose roles have been consistent in all these years (long-term political agents) or have been increasingly important in the recent years (emergent agents).
- (ii) It provides for consistent relative network measures for such agents compared to their peers.
- (iii) And most importantly, it allows us to draw a parallel of the data-driven agents' consistent roles to the actual chronological historical events. This is very important as we need to validate our results using the actual historical events and the SMEs' understanding.

Such narrations enable the "story" of political change to unfold from this computer assisted process, transforming them into a familiar ethnography form that is eligible to be validated by Sudanist SMEs.

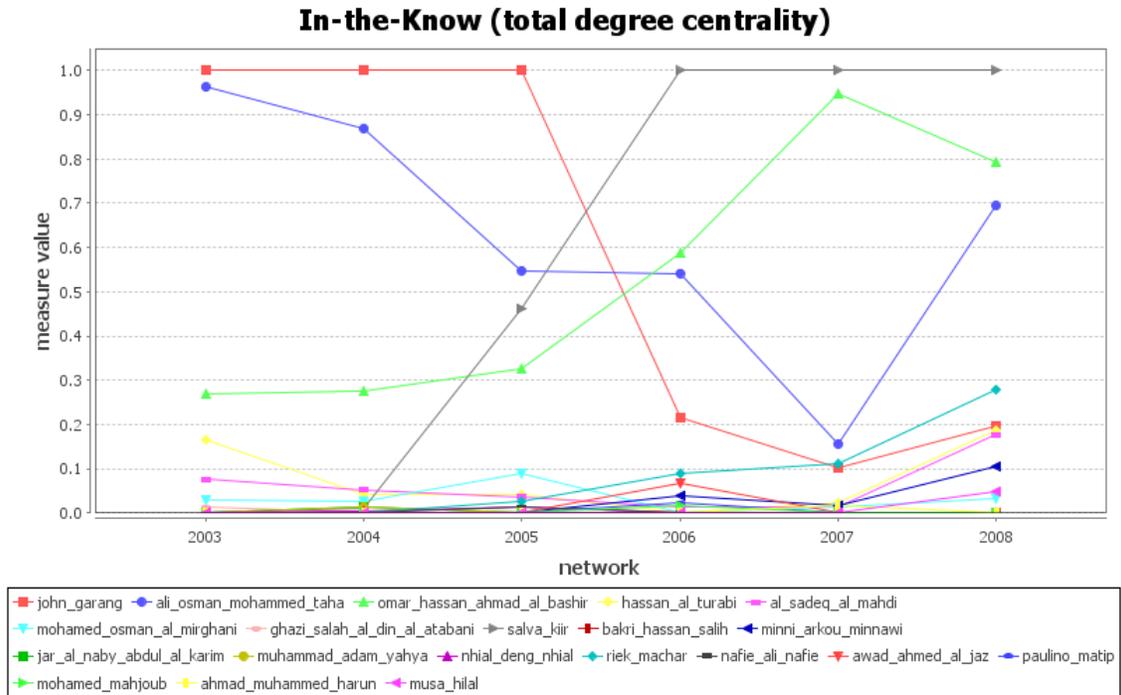


Figure 1: Total Degree Centrality for political agents of Sudan in 2003-2008. This figure shows how Garang’s score sharply declines after 2005. It also shows how Taha’s score declines in 2003-2007 yet increases again in 2008. In contrast to Garang’s, Bashir’s and Mayardit’s scores sharply increase beginning in 2004 and 2005.

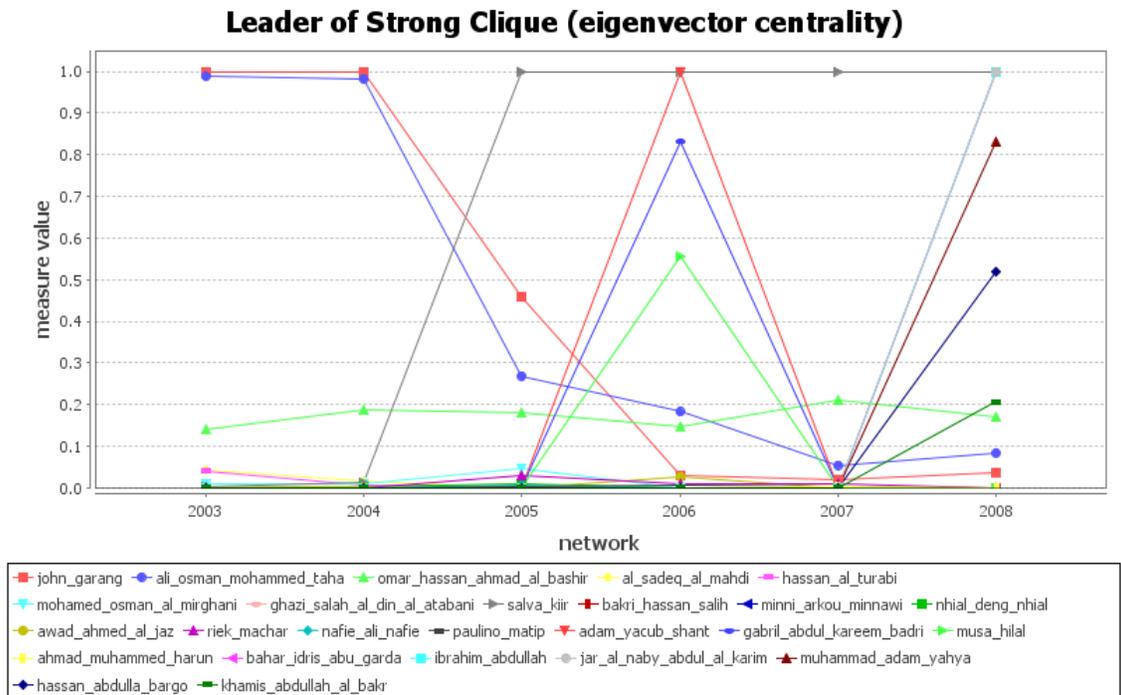


Figure 2: Eigenvector Centrality for political agents of Sudan in 2003-2008. Figure 2 shows the Eigenvector Centrality of the political agents of Sudan. This figure shows that Taha’s Eigenvector Centrality measure sharply

Figure 4: Betweenness Centrality for political agents of Sudan in 2003-2008. This figure shows the Betweenness Centrality measures. It shows how Garang's score steadily decreases after 2005. It also shows how Mayardit's score starts to steadily increase in 2005. This measure also shows how Taha's score decreases sharply in 2005 yet he is able to regain it such that he has the second position in 2008. However, this measure most importantly shows how Bashir's power has increased sharply after it temporarily decreases in 2005 such that his score as the first rank is significantly higher than the other agents in 2007-2008.

Narrated Results

Our results show that the current president of Sudan, Omar Hassan Ahmad Al Bashir, holds top ranks in 2008 (table 1). He ranks first in Cognitive Demand and Betweenness Centrality, second in Total Degree Centrality, and eight in Eigenvector Centrality. Furthermore, his ranks in every category have consistently been high in 2003-2008. This shows that he has been the most powerful (and even getting more and more powerful) person in Sudan. Our SME easily validates this result as it is historically true that Bashir has been in power as a president since 1989 after he successfully led a military coup that ousted the government of Sadeq Al Mahdi (Cowell 1989). Bashir is exceptionally strong in Total Degree Centrality, Cognitive Demand and Betweenness Centrality, meaning that he is not only a formal leader but is also a de facto leader who connects factions.

Table 1: Measures for Omar Hassan Ahmad Al Bashir

Year	2003		2004		2005		2006		2007		2008	
	Rank	Value										
Centrality, Total Degree	3	0.27	3	0.27	4	0.33	2	0.59	2	0.95	2	0.79
Cognitive Demand	1	0.21	3	0.16	2	0.18	3	0.13	1	0.22	1	0.26
Centrality, Eigenvector	3	0.14	3	0.18	4	0.18	6	0.15	2	0.21	8	0.17
Centrality, Betweenness	1	0.04	1	0.08	2	0.05	5	0.02	1	0.07	1	0.13

Salva Kiir Mayardit's important role is shown by our result as he tops Bashir in Total Degree Centrality from 2006 and in Eigenvector Centrality since 2005 (table 2). This means that he is powerful not only because of his own important position but also because of his position relative to other important agents in the political networks of Sudan. This is another straight forward validation by our SME as it is the fact that Mayardit is the current President of Autonomous Government of Southern Sudan and the First Vice President of Sudan. He has become a more important figure in recent years agreeing with the fact that he has been strongly supported by the people of Southern Sudan. His ranks are first significant in 2005 when he not only succeeds Garang but also starts his office as the First Vice President of Sudan, a position that makes him the second most powerful person in Sudan; a position referred as "the most powerful Vice President in the world" (Tombe 2008). Mayardit is indeed the second most powerful person in Sudan, not far off from Bashir.

Table 2: Measures for Salva Kiir Mayardit

Measure Name	2003		2004		2005		2006		2007		2008	
	Rank	Value										
Centrality, Total Degree	8	0.00	9	0.01	3	0.36	1	1	1	1	1	1
Cognitive Demand	7	0.07	7	0.05	3	0.18	1	0.23	3	0.15	3	0.18
Centrality, Eigenvector	8	0.00	5	0.01	1	1	1	1	1	1	1	1
Centrality, Betweenness	6	0.02	-	-	5	0.03	8	0.02	6	0.02	5	0.06

Ali Osman Mohammed Taha has top ranks in 2003 and 2004, especially in Total Degree and Eigenvector Centralities where he ranks second (table 3). He also ranks third and first in 2003 and 2004 for Cognitive Demand. These mean that he is an important agent in 2003-2004 due to his de facto leadership obtained by not only his own important position but also his proximate position relative to other important agents, not unlike Mayardit in later years. Our SME validates this result because it is a historical fact that those are the years when he has an important proxy role for Bashir in negotiating with John Garang and the Southern Sudanese for the handling of the Darfur crisis (Sudan Tribune 2004). However, following his trial and relinquishment from the office of the First Vice President into the Second Vice President of Sudan in 2005, his role and power have decreased as shown in his declining ranks in the subsequent years.

Table 3: Measures for Ali Osman Mohammed Taha

Measure Name	2003		2004		2005		2006		2007		2008	
	Rank	Value										
Centrality, Total Degree	2	0.96	2	0.87	2	0.55	3	0.54	3	0.16	3	0.7
Cognitive Demand	3	0.21	1	0.21	5	0.06	-	-	4	0.15	2	0.23
Centrality, Eigenvector	2	0.99	2	0.98	3	0.27	5	0.18	3	0.06	9	0.08
Centrality, Betweenness	3	0.02	4	0.05	8	0.00	10	0.01	9	0.01	10	0.01

John Garang has very high ranks in some important categories in 2003-2005 (table 4). He ranks top two, mostly the first, in Cognitive Demand, Total Degree Centrality and Eigenvector Centrality. He also ranks top four in Betweenness Centrality. His ranks peak in 2005 when he tops all but one category. This result is validated by our SME as Garang had been historically the leader of Southern Sudan until 2005. Negotiating against Taha, Garang has an important in the Darfur Crisis in 2003-2004. This negotiation brings his political career to a peak in 2005 when he holds the office of the first vice president, making him the second most powerful person in Sudan. However, his legacy ends due to his death in 2005. Yet, there could be bias in the increments in his ranks in 2005 due to the amount of news published related to his legacy and the controversies of his death (Young 2005). This publication bias is evident as Garang still ranks top ten in most categories, although they are subsequently declining, in 2006-2008 although he has passed.

Table 4: Measures for John Garang

Year	2003		2004		2005		2006		2007		2008	
Measure Name	Rank	Value										
Centrality, Total Degree	1	1	1	1	1	1	4	0.22	5	0.1	5	0.2
Cognitive Demand	2	0.21	2	0.18	1	0.27	2	0.13	2	0.15	6	0.13
Centrality, Eigenvector	1	1	1	1	2	0.46	7	0.03	4	0.02	10	0.04
Centrality, Betweenness	4	0.02	3	0.06	1	0.06	1	0.06	3	0.04	-	-

Conclusions and Discussions

Conclusions

Our results also show that there is almost an even distribution of power between the central government (Bashir and Taha) and the Southern Sudan government (Mayardit and Garang). Consistent with the actual historical events, our analysis identifies that President Al Bashir is still a very powerful agent despite the allegations of war crimes and genocide against him (Gray 2009). This conclusion on Al Bashir extends beyond our available data as he has been reelected again as the President of Sudan in 2010 (BBC World Service 2010). On the other hand, our analysis indicates that Ali Osman Mohammed Taha loses some of his influence after his relinquishment from the office of the First Vice President into the Second Vice President in 2005 although he is still powerful. The importance of the office of vice presidency in Sudan is also shown by the increase in Salva Kiir Mayardit's influence after he replaces Taha following the death of the powerful and influential John Garang. Again, our result is validated by a historical development of event that is beyond the scope of our data: Mayardit is the the first president of Southern Sudan after the split of the country into two independent countries in July 2011. Our analysis was done before a referendum was passed in January 2011, a referendum that results in the independence of Southern Sudan in July 2011 (BBC News 2011). This shows that our method works even though we only employ a simple analysis of narrated rank order.⁵

Using the Rapid Ethnography Retrieval system we are able to explain the longitudinal changes of political agents of Sudan. The changes we identify are parallel to historical events and the analysis has been validated by not only Sudanist Subject Matter Experts but also by actual historical events beyond the scope of our data. It is not surprising that most of our results were known by Sudanist SMEs as we were processing historical data. While most Sudanist SMEs would have not likely discovered much new and unexpected results in the relations shown in this paper, the fact that independent non-Sudanist-SME researchers could accurately detect these relations through network measures extracted from Network Text Analysis method using the computer RER system processing the data in much efficient short time is important.

⁵ Readers interested in the further dynamics of this process should consult forthcoming papers. In these papers, the longitudinal progression is further investigated in which the significance of change is quantified (Tambayong, 2013a) and its stability is measured (Tambayong, 2013b).

We also want to note that the results obtained through this data-to-model process with the RER did lead to two new findings that surprised many of the Sudanists. In this paper we did not report all results generated by using the RER. Yet, many others, such as those involving emergent leaders were also confirmed by SMEs. They agreed they made sense in retrospect but they just had not thought of it. The first is that our analysis (figure 1-4) showed Minni Arkou Minnawi as someone to watch; i.e., he was becoming increasingly influential and consistently ranks in top 10 in all categories. For most Sudanists, they had dismissed him as an assistant of Bashir; whereas, we find that his powerbase and influence is increasing which could make him a key person in the near future. Time elapsed after we drew such conclusion when we read the news that Minnawi had abandoned his allegiance to the central government (i.e. Bashir) and fled to Southern Sudan (Lavallee 2010). Second, our analysis showed that as you moved through time there was great volatility in an agent served as the best broker or intermediary. This suggests that the intermediary may depend on the specific sub-conflict, which we plan to investigate. The Sudanist SMEs, however, tended to lock onto a particular person and did not see change as quickly as our techniques did. This suggests that RER if used in conjunction with SMEs, or by SMEs, may serve as an early indicator system to predict what who to watch.

Discussions

These illustrative results are important as it is a solid proof of concepts of principle and practice that underscores the validity of RER system, especially founded in open source NTA and dynamic network analysis. We do not view the RER system as a replacement for SMEs. On the contrary, SME's ethnography expertise and knowledge are an integral part of RER. SMEs employing RER can actually work faster and spend more of their effort on assessment and understanding in both construction and validation processes; tasks that may be better suited to people than machines. Built on SMEs' expertise, NTA-utilized RER system excels as an efficient and accurate methodology as it provides for a systematically standardized, reliable, and accurate way to automatically extract information from a vast amount of texts such that it is not prone to errors caused by inconsistencies in human judgment. In turn, this allows for a better understanding of the region of interests.

This method is especially helpful and more efficient and accurate when one deals with vast amount of texts that are beyond practicality, and possibly beyond physical capacity, to be analyzed individually by SMEs. It is also helpful when one is exploring a new region of interest and there are few SMEs and too little time to support doing an in-depth multi-year study, which is what most Sudanist SMEs have done. A working RER system would streamline this process and facilitate the iterative cycle so as to minimize the time and effort required by the analyst. What we did was to a) run through this entire process with detailed human-in-the-loop processing, b) identified all points that could be automated, c) automated those points, d) identified all thesauri that had value beyond just this datasets, e) cleaned and formatted those, and f) then re-processed all the data with the more fully automated and streamlined the system. The result was a much reduced processing time for this data (which in our case only took 2 weeks); and, a system that can be re-used on other datasets.

References

- Alexa ,Melina and Cornelia Zuell. 2000. "Text Analysis Software: Commonalities, Differences, and Limitations: The Results of a Review." *Quality and Quantity* 34: 299-321.
- "Computer-Assisted Text Analysis Methodology in the Social Sciences." *ZUMA-Arbeitsbericht* 97/07.
- Batagelj, Vladimir, Andrej Mrvary, and Matjaz Zaveršnik. 2002. "Network Analysis of Texts." In *Proceedings of the 5th International Multi-Conference Information Society: Language Technologies, Ljubljana, Jezikovne tehnologije / Language Technologies*, edited by T. Erjavec and J. Gros.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2004. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.
- Bonacich, Phillip. 1987. "Power and Centrality: A Family of Measures." *The American Journal of Sociology* 92: 1170-1182.
- Burkart, Margarete. 2004. "Thesaurus." In *Grundlagen der Praktischen Information und Dokumentation: Ein Handbuch zur Einführung in die Fachliche Informationswissenschaft und praxis*, edited by R. Kuhlen, T. Seeger, and D. Strauch. Munich, Germany: Saur.
- Carley, Kathleen M. 1993. "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." *Sociological Methodology* 23: 75-126.
- Carley, Kathleen M. 1997. "Network Text Analysis: the Network Position of Concepts." In *Text Analysis for the Social Sciences*, edited by C.W. Roberts. Mahwah, NJ: Lawrence Erlbaum.
- Carley, Kathleen M. 2002. "Smart Agents and Organizations of the Future." In *The Handbook of New Media*, edited by L. Lievrouw L and S. Livingstone. Thousand Oaks, CA: Sage.
- Carley, Kathleen M. 2006. "Destabilization of Covert Networks." *Computational and Mathematical Organization Theory* 12: 51-66.
- Carley, Kathleen M., Jana Diesner, Jeffrey Reminga, and Maksim Tsvetovat. 2007. "Toward an Interoperable Dynamic Network Analysis Toolkit." *DSS Special Issue on Cyberinfrastructure for Homeland Security: Advances in Information Sharing, Data Mining, and Collaboration Systems* 43: 1324-1347.
- Carley, Kathleen M., Dave Columbus, Matthew DeReno, Michael Bigrigg, Jana Diesner, and Frank Kunkel. 2009a. *AutoMap User's Guide 2009*. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-09-114.
- Carley, Kathleen M. and Michael Palmquist. 1992. "Extracting, Representing, and Analyzing Mental Models." *Social Forces* 70: 601-636.
- Carley, Kathleen M., Jeffrey Reminga, Jonathon Storricks, and Matthew DeReno. 2009b. *ORA User's Guide 2009*. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-09-115.
- Carley, Kathleen M. 1994. "Extracting Culture through Textual Analysis," *Poetics* 22: 291-312.

- Chakrabarti, Soumen. 2002. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann.
- Corman, Steven R., Timothy Kuhn, Robert D. McPhee, and Kevin J. Dooley. 2002. "Studying Complex Discursive Systems: Centering Resonance Analysis of Communication." *Human Communication* 28: 157-206.
- Danowski, James. 1993. "Network Analysis of Message Content." *Progress in Communication Science XII*, edited by W.D. Richards and G.A. Barnett. Norwood, NJ: Ablex.
- Diesner, Jana and Kathleen M. Carley. 2005. "Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis." *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*. Harrisburg, PA: Idea Group Publishing.
- Diesner, Jana and Kathleen M. Carley. 2008. "Conditional Random Fields for Entity Extraction and Ontological Text Coding." *Journal of Computational and Mathematical Organization Theory* 13: 248-262.
- Diesner, J., Carley, K.M., and Tambayong, L. 2012. "Mapping Socio-Cultural Networks of Sudan from Open-Source, Large-Scale Text Data." *Computational and Mathematical Organization Theory* 18, 3; *Special Issue: Data to Model*.
- Ding, Bolin, Bo Zhao, Cindy Xide Lin, Jiawei Han, and Chengxiang Zhai. 2010. "TopCells: Keyword-based Search of Top-k Aggregated Documents in Text Cube." *Proceedings of 2010 International Conference on Data Engineering (ICDE'10)*.
- Fluehr-Lobban, Carolyn, Richard A. Lobban, and Robert S. Kramer. 2002. *Historical Dictionary of the Sudan*. Lanham, MD: The Scarecrow Press.
- Freeman, Linton C. 1979. "Centrality in Social Networks I: Conceptual Clarification." *Social Networks* 1:215-239.
- Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Analysis." *Proceedings of Uncertainty in Artificial Intelligence*.
- Holsti, Ole R. 1969. *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.
- Jurafsky, Daniel and James H. Marton. 2000. *Speech and Language Processing*. Upper Saddle River, NJ: Prentice-Hall.
- Kelle, Udo. 1997. "Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data." *Sociological Research Online* 2, 2.
- Klein, Harald. 1997. "Classification of Text Analysis Software." In *Classification and Knowledge Organization: Proceedings of the 20th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Freiburg, Berlin*, edited by R. Klar and O. Opitz. New York, NY: Springer.
- Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology, 2nd edition*. Thousand Oaks, CA: Sage.

- Landauer, Thomas, Peter W. Foltz, and Darrell Laham. 1998. "Introduction to Latent Semantic Analysis" *Discourse Processes* 25: 259-284.
- Lin, Cindy X., Bo Zhao, Qiaozhu Mei, and Jiawei Han. 2010. "A Statistical Model for Popular Event Tracking in Social Communities." *Proceedings of 2010 ACM International Conference on Knowledge Discovery and Data Mining (KDD'10)*.
- Manning, Christopher D. Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.
- Niblock, Tim. 1987. *Class and Power in Sudan*. Albany, NY: SUNY Press.
- Popping, Roel. 2003. "Knowledge Graphs and Network Text Analysis." *Social Science Information* 42:91-106.
- _____. 2000. *Computer-Assisted Text Analysis*. Thousand Oaks, CA:Sage.
- Popping, Roel and Carl W. Roberts. 1997. "Network Approaches in Text Analysis." In *Classification and Knowledge Organization: Proceedings of the 20th Annual Conference of the Gesellschaft für Klassifikation*, edited by R. Klar and O. Opitz. Berlin, Germany: Springer Berlin.
- Porter, Martin F. 1980. "An Algorithm for Suffix Stripping." *Program* 14, 3: 130-137.
- Ramakrishnan, Cartic, Krys. J. Kochut, and Amit. P. Sheth. 2006. "A Framework for Schema-Driven Relationship Discovery from Unstructured Text." *Proceedings of the International Semantic Web Conference*.
- Roth, Dan and W. Yih, 2007. "Global Inference for Entity and Relation Identification via a Linear Programming formulation." In *Introduction to Statistical Relational Learning*, edited by L. Getoor, L. and B. Taskar, B. Cambridge, MA: MIT Press.
- Ryan, Gery W. and H. Russell Bernard. 2000. "Data Management and Analysis Methods." In *Handbook of Qualitative Research (2nd edition)*, edited by N. Denzin and Y. Lincoln. Thousand Oaks, CA:Sage
- Sowa, John F. 1984. *Concept Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.
- Tambayong, Laurent. 2013a. "Change Detection in Dynamic Political Networks of Sudan." In *Modelling and Simulation of Complex Social Systems*. Eds. Vahid Dabbaghian and Vijay Mago. New York. Springer.
- Tambayong, Laurent. 2013b. "Stability and Dynamics in Political Networks of Sudan." *Journal of Artificial Societies and Social Simulation* (forthcoming).
- Wang, Chi, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. 2010. "Mining Advisor-Advisee Relationships from Research Publication Networks." *Proceedings of the 2010 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'10)*.
- Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK.

Young, John. 2005. "John Garang's Legacy to the Peace Process, the SPLM/A & the South." *Review of African Political Economy* 32: 535-548.

Zhang, Duo, Cheng Xiang Zhai, Jiawei Han, Ashok Srivastava, and Nikunj Oza. 2009. "Topic Modeling for OLAP on Multidimensional Text Databases: Topic Cube and its Applications." *Statistical Analysis and Data Mining* 2:378-395.

Web References

Associated Press. 2006a. "Arab League Nations Offer Peacekeeping Troops for Darfur." *The Washington Post*. Available: <http://www.washingtonpost.com/wp-dyn/content/article/2006/10/08/AR2006100800668.html> (October 8, 2012).

Associated Press. 2006b. "Ceasefire Deal Tabled for Sudan." *Sudanese Online*. Available: <http://www.sudaneseonline.com/enews2006/mar13-27631.shtml> (October 8, 2012)

BBC News. 2011. "South Sudan Backs Independence: Results." Available: <http://www.bbc.co.uk/news/world-africa-12379431> (October 8, 2012).

BBC News. 2009. "Q&A: Sudan's Darfur crisis." *BBC News*. Available: <http://news.bbc.co.uk/2/hi/africa/3496731.stm> (October 8, 2012).

BBC News. 2004. "Sudan Denies Darfur Militia Ties." *BBC News*. Available: <http://news.bbc.co.uk/2/hi/africa/3908645.stm> (October 8, 2012).

BBC World Service. 2010. "President Omar Al-Bashir Re-elected in Sudan Elections." *BBC World Service*. Available: http://www.bbc.co.uk/worldservice/news/2010/04/100426_sudan_elections_hs.shtml (October 8, 2012).

CNN. 2008. "Sudanese President Charged with Genocide." *CNN*. Available: <http://www.cnn.com/2008/WORLD/africa/07/14/darfur.charges/> (October 8, 2012).

Cowell, Alan. 1989. "Military Coup In Sudan Ousts Civilian Regime." *NY Times*. Available: <http://www.nytimes.com/1989/07/01/world/military-coup-in-sudan-ousts-civilian-regime.html> (October 8, 2012).

Goodman, Peter S. 2004. "China Invests Heavily in Sudan's Oil Industry." *The Washington Post*. Available: <http://www.washingtonpost.com/wp-dyn/articles/A21143-2004Dec22.html> (October 8, 2012).

Gray, Melissa. 2009. "Al-Bashir Prosecutor Pushes for Genocide Charge." *CNN*. Available: <http://edition.cnn.com/2009/WORLD/africa/07/08/sudan.bashir.war.crimes/index.html> (October 8, 2012).

Lavallee, Guillaume. 2010. "Sudan Rebel Chief Ready to Battle Khartoum." Available: <http://www.reliefweb.int/rw/rwb.nsf/db900SID/LSGZ-8C3MCC?OpenDocument> (October 8, 2012).

- Malek, Cate. 2005. "The Darfur Region of Sudan." Available:
http://www.beyondintractability.org/case_studies/Darfur.jsp?=5101 (October 8, 2012).
- McCrummen, Stephanie. 2009. "A Town Constantly On Brink of Chaos." *The Washington Post*. Available:
<http://www.washingtonpost.com/wp-dyn/content/article/2009/04/24/AR2009042403746.html>
(October 8, 2012).
- Nwazota, Kristina. 2008. "The Darfur Crisis: African Union's Effort." *PBS Newshour*. Available:
http://www.pbs.org/newshour/indepth_coverage/africa/darfur/union.html (October 8, 2012).
- Sudan Tribune*. 2004. "Taha, Garang Affirm Readiness to Settle Outstanding Issues." *Sudan Tribune*.
Available: <http://sudantribune.com/spip.php?article5877> (October 8, 2012).
- Tombe, Wani. 2008. "Critical Analysis of the Sudanese First Vice President's Speech in Juba on January 9, 2007." *Sudan Vision Daily*. Available:
<http://www.sudanvisiondaily.com//modules.php?name=News&file=article&sid=17520> (October 8, 2012).
- United Nations. 2008. *Security Council Resolution 1828*. Available:
http://www.un.org/ga/search/view_doc.asp?symbol=S/RES/1828%282008%29 (October 8, 2012).
- United Nations. 2004. *Security Council Demands Sudan Disarm Militias in Darfur, Adopting Resolution 1556 (2004) by Vote 13-0-2*. Available: <http://www.un.org/News/Press/docs/2004/sc8160.doc.htm>
(October 8, 2012).
- U.S. Department of State. 2009b. "Sudan: a Critical Moment, a Comprehensive Approach." *U.S. Department of State, Office of the Spokesman*. Available:
<http://www.state.gov/r/pa/prs/ps/2009/oct/130672.htm> (October 8, 2012).
- Walter, Peter and James Sturcke. 2008. "Darfur Genocide Charges for Sudanese President Omar al-Bashir." *Guardian*. Available: <http://www.guardian.co.uk/world/2008/jul/14/sudan.warcrimes1>
(October 8, 2012).
- de Waal, Alex. 2006. "Sudan: Disarming the Janjaweed and Armed Militia." Available:
<http://allafrica.com/stories/200607140742.html> (October 8, 2012).
- WYDA. 2008. "Dinka Tribe." *Werkok Youth Development Association*. Available:
<http://www.wydasudan.org/dinka-tribe> (October 8, 2012).

Appendix A: A Background on the Violent Conflict in Sudan

Since 2003, the Darfur rebellion has ignited a series of important events that lead to a change in the landscape of Sudan's political networks. This has resulted in an estimate of 300,000 casualties and 2.7 million refugees (BBC News 2009). Masked as an allegation of oppression of black Sudanese in favor of Arab Sudanese, the conflict has been centered in the issues of land and grazing rights (Malek 2005). The prolonged conflict is complex as Sudan is a home to many different tribes and semi-autonomous communities loyal to local leaderships (McCrummen 2009). The conflict has become more complicated when the government of Sudan allegedly encourages militias as a self-defense measure (BBC News 2004). In 2004, the government of Sudan is accused of genocide by the international world following a series of violent and brutal incidents. Following the genocide accusation, the United Nations sanctions the government of Sudan. This results in major changes in the political landscape of Sudan. In 2005, there is an international recognition of an autonomous government of Southern Sudan under John Garang, who also serves as the First Vice President of Sudan replacing Ali Osman Mohammed Taha. Taha then becomes the Second Vice President of Sudan. Only a few months in office, Garang dies in a helicopter crash. Garang's office as the First Vice President is then succeeded by Salva Kiir Mayardit. Year 2005 also marks the beginning of a series of monumental signings of peace agreements. The first one is the Comprehensive Peace Agreement (CPA) signed in January 2005 by the Government of Sudan and the Sudan People's Liberation Movement/Army (SPLM/A). Following that, Darfur Peace Agreement is signed by the Government of Sudan and the Sudan Liberation Movement/Army (not to be confused with SPLM/A) in May 2006. Later, the Eastern Sudan Peace Agreement (ESPA) is signed in October 2006. ESPA ends comparatively lower-intensity conflicts in the eastern part of Sudan. However, violence seems to intensify again in the following year. In 2007, President Omar Hassan Ahmad Al-Bashir receives war crime charges from the International Court. This leads to an increased UN presence in 2008. These series of events are summarized in table 5.

Table 5: Summary of historical events in Sudan 2003-2008

	2003	2004	2005	2006	2007	2008
Event	Darfur rebellion	International recognition of genocide	UN Sanctions Autonomous Southern Sudan CPA signed	DPA signed ESPA signed	War crime charges	Increased UN presence Sudan-Chad accord
Bashir	President of Sudan	President of Sudan	President of Sudan	President of Sudan	President of Sudan	President of Sudan
Garang			Becomes VP of Sudan & Southern Sudan President Dies in a plane crash			
Taha			Demoted to Second VP from First VP of Sudan			
Mayardit			Becomes First VP of Sudan			