IGI PUBLISHING
WWW.IGI-GLOBAL.COM

International Journal of Applied Geospatial Research

April-June 2015, Vol. 6, No. 2

Table of Contents

Editorial Preface

Research Articles

The *International Journal of Applied Geospatial Research* is indexed or listed in the following: ACM Digital Library; Bacon's Media Directory; DBLP; Google Scholar; INSPEC; JournalTOCs; Library & Information Science Abstracts (LISA); MediaFinder; SCOPUS; The Standard Periodical Directory; Ulrich's Periodicals Directory

Spatiotemporal Network Analysis and Visualization

Judith Gelernter, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Kathleen M. Carley, Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Spatiotemporal social network analysis shows relationships among people at a particular time and location. This paper presents an algorithm that mines text for person and location words and creates connections among words. It shows how this algorithm output, when chunked by time intervals, may be visualized by third-party social network analysis software in the form of standard network pin diagrams or geographic maps. Its data sample comes from newspaper articles concerning the 2006 Darfur crisis in Sudan. Given an immense data sample, it would be possible to use the algorithm to detect trends that would predict the next geographic center(s) of influence and types of actors (foreign dignitaries or domestic leaders, for example). This algorithm should be widely generalizable to many text domains as long as the external resources are modified accordingly.

Keywords: *Data Mining, Social Network Analysis, Spatiotemporal Network Analysis, Text Mining Knowledge Discovery, Visualization*

1. INTRODUCTION

1.1. Social Networks and Assumptions about Place and Time

A social network describes a group of people who share some sort of social connections, whether through work, or friendship, or otherwise. The social network concept stems from mid 20th century sociology. Alignment of social network studies with computer science in about the 1970s allowed the connections among individuals to be weighted and computed mathematically on a large scale, with weights indicating, for example, strength of the relationship.

An analysis of a social network generally focuses on the groupings of people. The people might be employees of an organization, for example, or colleagues in a discipline, sportsmen on teams, or characters in a novel. Questions that could be answered by the analysis include: Which individuals are in what group? Who leads each group? Who is second to the leader? Who is between two groups? Are group relations friendly or antagonistic? At the foundation of the social network literature are Scott (1991 [2004]), and Wasserman and Faust (2004).

DOI: 10.4018/ijagr.2015040105

Social networks have been diagrammed with people as points and their social relations as lines. The points are called nodes; the relations are called ties or edges. The groups are cliques. The aggregate of cliques form a network at some time. Most social network studies and the standard social network diagrams account for neither space nor time. This may be because it is assumed that the network is spatially and temporally persistent and so does not change, or because a network snapshot is good enough. Whichever assumption is made, often space and time are treated as irrelevant factors.

Our social networks in our research are formed from connections among people mentioned in news articles. The activities of people and their co-relationships in space and time come from the context of the news story. When we extract names of peoples from these texts and build links among them using the proximity of their names in the article, we are in essence attributing a relationship among the people. This context can be characterized by the spatio-temporal setting of the news article.

Social network analysis of vast amounts of text through data mining, as described in this research, affords an overview of events. No reading of the text is necessary. This does not simply save an analyst much time and effort; it allows assimilation of text on a scale that would otherwise require many people to analyze. Even though errors occur because network nodes are only *inferred,* automated node extraction saves times and offers insight that is valuable. We suggest that the accuracy of the extracted network can be enhanced through improved extraction of spatio-temporal information that describes the network membership and relations among members.

1.2. Data Mining For Social Networks

Data mining techniques are used to extract social network data from text. The way it works is that the text is submitted to a series of filters until the sought-after information remains. In the early phases of processing, grammatical articles (a, an, the) are filtered as noise. Sometimes numerals and symbols in the text are removed, and the text is normalized to lower case. Remaining words may be reduced to their stems so that noun plurals, verb past tenses, gerund endings and so forth are removed. Then external lists will be more effective in finding relevant network words in the text.

How do we mine for people and location nodes? To identify person and place, external sources as well as language processing methods play a role. In Named Entity Recognition, an entity is a proper name, an organization, an event, or a location (Giuliano, Lavelli, & Romano, 2007). To simplify the data mining required in our case, we restricted the entities to names on a match list, and to places in a world gazetteer. We mine date from the header information that accompanies each news article. A bibliography for the mining of network data for spatio-temporal information is found in Roddick and Spiliopoulou (1999).

How do we determine edges that are between nodes? Edges are created according to node word proximity as dictated by grammar and syntax, and within a word-window size set by the text-processing software user. As an example, person node and location node found anywhere in the same sentence ordinarily will receive an edge. If that same person node ends one sentence and the location begins the following sentence, they might not receive an edge if the proximity window size set by the user had been small, although they would receive an edge if the proximity window set had been large. Thus, the process of setting node edges differs widely. Edges are determined irrespective of the meaning of the sentence from which they are extracted. Extracting network edges that are meaningful is substantially harder than extracting nodes by entity recognition, and state-of-the-art systems perform less well on this task than on the recognition task.

In short, the nature of the relationship between entities connected by an edge cannot be understood automatically. No algorithm that extracts network edges as yet provides complete semantic understanding. For example, if one

node represents a person and another node represents a place, the edge algorithm does not include information from the text that reveals whether the person comes from that place, is in that place temporarily, speaks about that place, or writes a legal document concerning that place. All the algorithm shows is some relationship between that person-node and location-node.

1.3. The Significance of This Research

Spatiotemporal information visualization has been called "a key research issue" (Klamma, Cao, Spaniol, & Leng, 2007) in spatiotemporal knowledge reasoning. Here we provide two visualizations using third-party software to show the data mined by our program. We visualize the networks as temporal pin diagrams, and as spatiotemporal maps.

Our maps make use of latitude longitude coordinates, as distinct from a recent trend called socio-mapping, which shows nodes in 2D or 3D space mapped by height and distance among nodes to represent not location but proportional strength of ties (Jenček et al., 2009). Our maps which show network information in time clusters are valuable because they reveal patterns. These patterns can be used to make inferences about what information is missing, and potentially to predict how the next network in the sequence might look.

Pre-processing software such as AutoMap may be used with external sources to prepare text for social network analysis (Carley, 2010). Our contributions are in the areas of finding location data, both by adapting external sources and by heuristics, and in the method of discerning strong or weak ties between person and location. These are elaborated below.

1.4. Our Research

Our research questions consider the ability of our algorithm to mine text for data that can be used to make a social network.

Can we find locations in text and associate these with geographic coordinates? How does our algorithm compare to a standard location-mining algorithm? Compared to a manual standard? [Research question set 1]

Can we find person-location-date tuples? How can we visualize this data in a spatio-temporal frame? [Research question set 2]

Below we describe related projects in mining for social networks. We then describe our text files, external resources, and some heuristics we devised to identify locations in the text. Then we present our GeoRef algorithm. We follow this with two experiments that demonstrate the utility of GeoRef. Our first experiment evaluates the accuracy of the algorithm's location mining component by comparison to a manual standard and to a freely-available location-mining program by Yahoo! called Placemaker. The second experiment uses our algorithm's spatio-temporal output in pin diagram and map visualizations made with the third-party social network analysis software, *ORA. Extended discussion of the second experiment follows. We conclude with future directions for research, and restate our main contributions in summary.

2. RELATED WORK ON DATA MINING

Extracting network nodes from text is accomplished though data mining. It uses techniques such as Natural Language Processing and other semantic methods (Carley 1993; Carley 1997), and resources such as controlled vocabulary in the form of dictionaries, thesauri, ontologies or gazetteers to extract knowledge from texts (Kodratoff, 1999). Workshops such as the Data Mining WebKDD/SNAKDD 2007 (Zhang et al., 2007) and conference presentations (Srivastava, 2008) have been devoted specifically to mining data for social network analysis. Here we briefly discuss others' research in finding locations in text, in finding temporal data in text, and in linking people to location.

2.1. Identifying Locations in a Text

What words in the text can be used to find locations? Mining location from text is a complex problem. A first step is toponym resolution, or

attaching a location to a place named in a text. The difficulty is that not all location words refer to actual locations, in what is called non-geo/geo ambiguity ("mobile" may describe a phone rather than a town in Alabama). The other problem is geo/geo ambiguity, introduced when there are several places with the same name. Location and time mining techniques are described by Roddick and Lees (2009), and location techniques by Buttenfield et al., (2001). A group at the University of Edinburgh has developed their own geoparser[1] (Tobin et al., 2010). Mining for location words in news, in text data similar to what we use here, has achieved up to 78.5% accuracy using unsupervised machine learning to develop disambiguation rules (Garbin & Main, 2005). Location-mining software has gone commercial. MetaCarta[2] will locate places named in a document or text stream. Yahoo! Placemaker[3] has a web service to do the same.

What external references can be used find locations? Reliance on a gazetteer improves an algorithm's ability to recognize locations. Gazetteers differ in scope, coverage, balance, accuracy, and entry specificity. Choice of gazetteer influences data mining results. Some researchers have used the National Geospatial Intelligence Agency gazetteer[4] or GeoNames,[5] while others have generated them automatically (Kozareva, 2006) or derived them from Wikipedia (Popescu & Grefenstette, 2010).

How can one understand location in a social network? Geographic proximity in a network has been called propinquity. Particular measures such as spatial closeness centrality and spatial betweenness centrality have been developed to analyze networks with location nodes (Olson, Malloy, & Carley, 2009). Location may be taken to be either a separate location node or as a location attribute of the person associated with that location.

2.2. Identifying Temporal Information in Text

What words in a text can be used to identify temporal information? Time may be discerned from seasonal words or holidays (such as Christmas) or time-centered events (such as the Beijing Olympics). It has also been construed as finding the order of events (Alvarez et al., 2010). Mining for time data may be as straightforward as scraping the time stamp from usage logs (Lauw, Lim, Pang, & Tan, 2010), or taking the metadata from news articles, as we have done here.

What external resources can be used to identify temporal information? One preliminary study created a time period directory that connects person or event to a date (Petras et al., 2006). In our study, we did not need an external reference for information because we mined the date from the article metadata.

How can one understand time in a social network? Article date has been used to "time-slice" a network into different intervals (Danowski, 2009). Events might occur across intervals as well as within intervals, so a mix of interval sizes might be preferable when the data are vast. Xu & Zheng (2009) cluster nodes at different time intervals and add ties that make sense among individual nodes afterward. Those time intervals that do not have ties among individual nodes are discarded, so that only linked clusters remain.

2.3. Identifying Network Edges

What methods can be used to find network edges? Main methods for extracting relations between entities are to discover verb relations (Pazienza, Pennacchiotti, & Zanzotto, 2006), construct concept graphs based on rules (Xu, Mete, & Yuruk, 2005), or use proximity to find relations within a sentence using a "word window" (Carley, et al., 2010). In data unambiguously associated with a social network such as the usage log from an online social network site, links can be given weights to show association by degree (Lauw, Lim, Pang, & Tan, 2010), or the links might be based on frequency of contact, so that a person associated with a particular location multiple times would have a weightier link (Danowski, 2009). Our algorithm only attaches two degrees to a relationship—strong or weak—and it associates

only one person with one location at one time. Others use soft clustering. For example, Lin, Chi, Zhu, Sundaram, & Tseng (2009) count the same person in more than one group by allowing soft community membership, and by proposing a probabilistic model that distributes individuals among communities.

3. DATA AND RESOURCES FOR DATA PROCESSING THAT PRECEDS SOCIAL NETWORK ANALYSIS

3.1. Data

We mined news articles to determine who the actors were, in what locations they were associated, and whether they changed location over time or disappeared, to be replaced by other actors. Social network analysis based on a very large number of articles from the Sudan Tribune from say, 2003 (a separate government had formed in southern Sudan by 2005) and 2011 (when a vote favored an independent south) might predict Sudan's political split. A series of social network diagrams from 2003 to 2011 might begin with a more homogeneous network that progressively split into dense cliques in the north that were distinct from dense cliques in the south, with relatively fewer edges between north and south. The algorithms described in this paper could be used to improve the accuracy of such a large scale assessment by improving the identification of who was where, when.

3.2. Resources for Data Processing

We used external lists for people's names and for place names in order to identify people and places mentioned in the text. A more general list of political officials and foreign dignitaries found almost none of the people mentioned in our data sample, so we created a match list for people manually by extracting names from our text corpus. For place names, we used the GeoNames gazetteer. Its advantages are that it is reasonably comprehensive in local towns which are the lower levels of the spatial hierarchy. Also, GeoNames is useful of the purposes of this study because it includes numerous alternate spellings for the same place, and some places found in our text corpora have no standard spelling because they are transliterated from Arabic, an official language of the Sudan. GeoNames, however, proved too comprehensive for our purposes. Its enormous size slowed processing greatly (the main download file in Nov. 2010 was 878 MB). We therefore used entries for upper levels of the global spatial hierarchy only.[6] Only for Sudan did we retain all feature classes and lower levels of the spatial hierarchy. In this way, we modified the GeoNames gazetteer to conform to our data set.

3.3. Heuristics for Resolving Location

Geoparsing is the process of identifying place names in text, and it is a core function of our GeoRef algorithm. Recall that difficulties in resolving locations are of two types: place names that are not recognized as places, and non-place names that are taken to be place names incorrectly, as mentioned above. Heuristics are described in more detail in Gelernter, Cao and Carley (2011). Below are examples of potential difficulties in resolving locations from our Sudan Tribune text. We have referenced examples below by download file name that corresponds to date, so that for example, the file 2006_wk19_21p contains news articles from the 19th week of 2006.

3.4. Place Names Not Identified As Places

Places not found in a standard gazetteer, such as regions that extend across countries, or are local, have ill-defined borders or multiple spellings, will not be identified correctly. This section gives examples and suggestions solutions.

3.4.1. Large Places

Regions which correspond to geographical areas larger than a country, such as the "Middle East",

do not appear in most gazetteers. An example from our corpus is "I want to appeal especially to those donors that have contributed much less so far than last year, as well as donors in the *Gulf region*." [2006_wk19_21p] Some of these could be added to a gazetteer manually, but additions would be done on a case-by-case basis.

3.4.2. Small Places

Neighborhoods and small villages often do not appear in world gazetteers. An example from our corpus is "saw in the *Gereida* area in South Darfur: massive displacement, constant violence and attacks against civilians…". [2006_wk19_21p] To supplement a gazetteer, one might find small places in large scale, local maps and guidebooks, and in mining crowd-sourced geography data such as in the OpenStreetMap project.[7] This would be done once the domain is known, however, otherwise this level of detail for the world makes a gazetteer too unwieldy for efficient processing.

3.4.3. Imprecise Regions

Regions that do not correspond to a precise geographic area do not appear in gazetteers. Recurring in our corpora, for example, are "Sudan's north," and "in the north." Another example in context is "the news of America's alleged willingness to set up a military base in *south Sudan* comes not as a surprise because the timing is premature … " [Wk2006_wk9_aol].[8] We consulted a range of Sudan maps, newspaper articles, and an expert in our research group on northern Africa to determine the conventional boundaries for each such imprecise region, and added them to the gazetteer.

3.4.4. Multiple Spellings

Place names transliterated from other languages sometimes are given spellings unrecognizable to standard gazetteers. For example, "Aradipe," the small town in Chad mentioned in the news article below, appears in GeoNames as "Aradip". From our text: "According to Mr Abdurasset, whose village is four miles east of Koukou-Angarana, two columns of Arabs made their first attack on *Aradipe* on Friday morning." [Wk51_etv] One method to lessen mistakes from transliteration differences would be to employ a Soundex phonetic algorithm that indexes names by sound as pronounced in English. The algorithm would allow different spellings which produce approximately the same-sounding place name to be equated, thereby improving identification of place names in text. Our data were not extensive enough to warrant that a Soundex module be added to our place identification algorithm.

3.5. Non-Place Names Mistaken For Places

Non-place names in a text may be misidentified as locations in cases when the names double as common terms, when they are found in titles, and when they occur in metonymy (when one concept stands for another). Each is described below.

3.5.1. Common Words

A standard gazetteer contains thousands of place names that are also common words. Some of this is happenstance ("Shirley" is a girl's name as well as a town in Limpopo, South Africa, and Illinois, U.S.A.). Place names become common names also as a result of the name being transliterated into English, as "Nor" and "Both" name places in Sudan's Upper Nile. The question is how we can prevent such place names from mining common words from a text that are not places. Others have tried to automate this filtering with limited success (Amitay et al., 2004). We surmounted this problem by manually reviewing all words in the gazetteer of 7 characters or less, and creating a filter list of 1169 places that are also common words. We permit words on the filter list to serve as place names only if they are preceded or followed immediately in the sentence by another place name (Mobile, Alabama, U.S.A.) or (Beijing, China).

3.5.2. Named Entities

Some titles of reports, organizations, corporations or books contain geographical names that do not refer to actual locations. For example, take the sentence "[t]he *Princeton* Project composed of eminent international jurists has contributed significantly to the foundation of the International Criminal Court (ICC) and its enabling act famously called the *Rome* Statute." [Wk2006_wk9_aol] The majorities of these are sources of error for our GeoRef algorithm, although we do provide a short filter list of a number of commonly-appearing newspapers with locations in their titles that tend to appear often in news data, such as "The New York Times," or "Washington Post."

3.5.3. Metonymy

Metonymy is a literary conceit in which one concept is substituted for another with which it is associated. Metonymy is fairly common in news when the name of a capital city or of a country is used to refer to the government of that country. One group of researchers found that is occurs about 17% of the time in geographic information retrieval (Leveling and Hartrumpf, 2008). Our GeoRef algorithm mistakenly identifies many of these instances of metonymy as places. For example, "Human Rights Watch has offered the most authoritative and detailed accounts of how the Janjaweed and *Khartoum* have coordinated, particularly in its December 2005 report." [2006_wk4_av] To prevent a few of the more common locutions, we hard coded [*fill in a capital city*] regime, and [*fill in a capital city*] government as not to refer to the named city.

3.6. Which Is The Correct Match In The Gazetteer?

Difficulties arise in parsing text in determining which is the correct location match when there are two or more places in the gazetteer with the same name. Leidner (2007) lists some rules that have been used by different researchers to resolve this problem. For example, one rule is to select the place that is higher in the geographical hierarchy (country above city), another rule is to select the place that is more populous, and another rule is to select the place that is in the geographic domain of the text, or that is closer in geographic distance to other non-ambiguous places named in the text. This problem is rare in our text corpus due to the limited number of repetitive names in the Sudan, although a prominent instance of repetition is the province of Kassala and the city of Kassala. Our rule when this arises is to use the place higher in the geographical hierarchy.

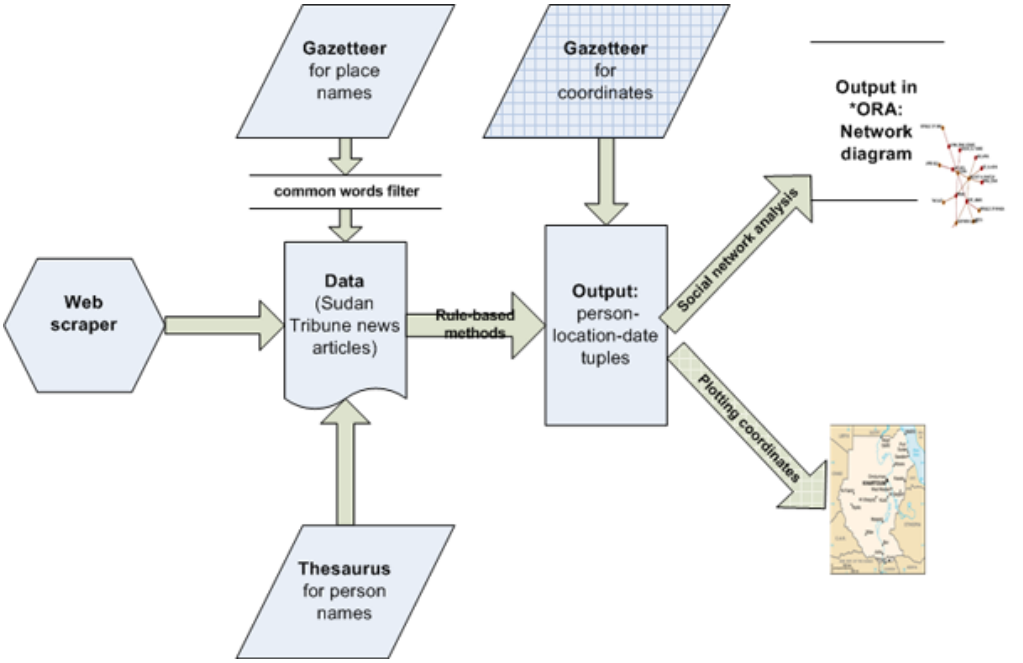
4. OUR GEOREF ALGORITHM

Our algorithm identifies person—location—date tuples, and it geo-codes each tuple with the latitude and longitude coordinates associated with that location's centroid. Figure 1 presents the algorithm's tasks schematically. Persons' names are mined through comparison to the thesaurus, and the place names are mined by comparison to the gazetteer. The algorithm output then is fed into a social network software to make diagrams or other sorts of visualizations.

4.1. Explanation for How Social Network Analysis Was Performed

1. **Download Data:** News articles were harvested from the Sudan Tribune website. We drew our data sample from these articles. We were constrained in the size of our data sample by the experiment that used human coders to find locations.
2. **Create Or Modify External Resources:** We first mined person's names from the text by using named entity extraction, but this list was inadequate for our small data sample since too many names from the text were missing from the list. Our solution on this small scale was to extract names manually from our text. However, a Named Entity Recognition algorithm should perform adequately for a much larger data sample. We altered the Geo-

Figure 1. Stages of the GeoRef data mining and output visualization. (Figure adapted from Figure 2 of Gelernter, Cao & Carley, 2011). The Sudan map from Travellerspoint.com isin the public domain



Names gazetteer to retain upper levels of the spatial hierarchy worldwide, along with local place names for the Sudan.

3. **Devise an Algorithm to Find Locations:** Our GeoRef algorithm uses string matching to identify place names in the text with matches in the gazetteer. The algorithm also employs some heuristics for resolving location as mentioned in the section above. Upper levels of the geospatial hierarchy are added to mined locations (so that state and country are added to city, for example).
4. **Create Person-Place-Date Tuples For The Social Network:**
   a. Identify the article date
   b. Identify persons named in the text
   c. Use GeoRef algorithm to identify locations
   d. Determine ties between person and location. We assign strong or weak ties based on the distance between the node words. Strong ties are assigned node words that were near each other in the text; weak ties are assigned when the location word was relatively farther from the person's name, or when it is necessary to return to the article title for a location. When no location is found to associate with the person, the tuple is dropped. Here it is in pseudocode:
      i. If geo-word occurs in same paragraph = 2 (strong link)
      ii. Else if geo-word occurs in title or first paragraph of article = 1 (weak link),
      iii. Else = 0
   e. Tuples are then assigned latitude, longitude coordinates by lookup of place name in the gazetteer.
5. **Visualize the Network:** To create social network diagrams and geographic maps, we input the tuples into the *ORA (Organization Risk Analysis) software, which can be downloaded freely.[9]

4.2. Generalizability of Our GeoRef Algorithm

Use of other text domains and other external match lists extend the utility of our algorithm. Many sorts of spatiotemporal networks could be created, for example, to chart businesses, indicate economic factors or distribution of goods, illustrate historical events, follow candidates on the election trail or the spread of a disease, or map crime.

5. EXPERIMENTS CONCERNING THE GEOREF ALGORITHM

5.1. Experiment 1: Location Mining From Text

- **Objective:** We will test the location data mining feature of the GeoRef algorithm in comparison to the location-mining program Yahoo! Placemaker.
- **Data:** 11 files of a total of 101 KB of text, randomly selected from the 2006 *Sudan Tribune.*
- **Method:** Each file was to be coded manually for location by participants who volunteered for the study. We used these manually-found locations as a benchmark to compare to the GeoRef and Placemaker algorithms.
- **Procedure:** The annotators were presented with the 11 text files. They were asked to find location words in the files and enter them in a spreadsheet, given that a place name is a noun, it is not part of an organization name, and it is not an instance of metonymy (where a place name stands for the government of the place). Only one of the two participants finished the study, so we used these results as the gold standard.

We entered the same files into the GeoRef and Placemaker algorithm, and used the manual annotations as a gold standard to score the algorithm output.[10] We illustrate the procedure in Table 1 with a segment of text and the location words selected by an annotator, and the GeoRef and Placemaker algorithms.

- **Scoring:** We counted the number of location words found by GeoRef and Placemaker that (1) agree with the manual coding, (2) do not agree with the manual coding (type II error) and (3) are missing (type I error). We scored a location word found by either software to be correct if it matched the place found by the manual coder, if it was lower in the hierarchy, or if it was higher in the hierarchy. So for example, if the manually-found location were Jonglei (a state in Sudan), and algorithm found Bor, Jonglei (where Bor is a city in the Jonglei state of Sudan), the algorithm output which is lower but in the same hierarchy, was considered correct. Or if the manually-found location were Southern Sudan and the algorithm output was Sudan which is higher in the spatial hierarchy, it was considered correct.
- **Results:** Our goal is the creation of a spatiotemporal network from texts. How well an algorithm finds places is thus a key determinant of network accuracy, and a critical metric for evaluation. We evaluated the algorithms on the basis of their location-finding accuracy.

Accuracy involves an aggregate of statistics for true and false positives and negatives. To estimate accuracy at the corpus level, we used all the document-level statistics in the formula:

$$accuracy = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

The equation shows that the accuracy statistic is comprised of true and false positives and negatives. A true positive (TP) is a correctly

Table 1. For the given paragraph of text, the table shows what a person, GeoRef and placemaker identified as locations

Khartoum escalates conflict in Eastern, Southern Sudan, and Darfur
Sunday 15 January 2006 01:30. [wk_2]
Escalates Conflict in Eastern Sudan, Southern Sudan, and Darfur; Kofi Annan belatedly acknowledges the need for robust international intervention to replace AU force in Darfur Jan 14, 2006 A wide range of recent news and policy reports clearly reveal the consequences of ongoing international failure to confront Khartoum s National Islamic Front, the dominant force in Sudan's nominal "Government of National Unity." For the NIF continues to escalate a series of militarily-driven crises in Africa s largest country, all of which imperil the widely heralded north/south peace agreement of a year ago. Physicians for Human Rights and the International Crisis Group have released particularly important reports: on the aftermath of genocidal violence in Darfur; on the growing military confrontation in eastern Sudan; and on Khartoum's continuing support for the destabilizing Lord's Resistance Army in southern Sudan and northern Uganda. Yet other reports suggest that a border war between Chad and Sudan, in areas that are filled with desperate refugees and internally displaced persons, may break out at any time.

Manually found places	GeoRef			Placemaker		
Geo-word in text	**Geo-word in text**	**State**	**Country**	**Geo-word in text**	**State**	**Country**
Eastern (Sudan)	Khartoum	Khartoum	SUDAN	Sudan		SUDAN
Southern Sudan	Southern Sudan		SUDAN	Africa and / Khartoum	Khartoum	SUDAN
Darfur	Darfur	Darfur Wilayat	SUDAN	Khartoum	Khartoum	SUDAN
Darfur	Khartoum	Khartoum	SUDAN	Al Khartum	Khartoum	SUDAN
Sudan	Eastern Sudan		SUDAN	Sudan		SUDAN
Africa	Southern Sudan		SUDAN	Darfur		SUDAN
Darfur	Darfur	Darfur Wilayat	SUDAN	Sudan		SUDAN
Eastern Sudan	Darfur	Darfur Wilayat	SUDAN	Khartoum	Khartoum	SUDAN
Southern Sudan	Khartoum	Khartoum	SUDAN	Chad		CHAD
Northern Uganda	Sudan		SUDAN	Sudan		SUDAN
Chad	Africa		Africa	Uganda		UGANDA
Sudan	Darfur	Darfur Wilayat	SUDAN	Africa		Africa
	Eastern Sudan		SUDAN	Darfur		SUDAN
	Khartoum	Khartoum	SUDAN			
	Southern Sudan		SUDAN			
	Uganda		UGANDA			
	Chad		CHAD			
	Sudan		SUDAN			

Table 2. Summary statistics for the location mining algorithms

	GeoRef	Placemaker
correct (TP)	289	235
missing (FN)—type I error	55	118
incorrect (FP)—type II error	176	121

identified location, that is, a location found by the algorithm that was also found manually. A false positive (FP) is an example incorrectly identified as positive (saying it is Cairo when actually Cairo is not represented in the annotated standard). A true negative (TN) is a negative example correctly identified (recognizing that Cairo is *not* in the data when it also not in the standard), and a false negative (FN) is mistaking a negative (omitting Cairo when it is in fact represented in the data and appears in the standard). When we calculated scores for the TP, FP and FN for the respective algorithms as shown in Table 2 and entered them in the above formula, we found that the GeoRef algorithm yielded 56% accuracy and Yahoo! Placemaker yielded 50% accuracy.

- **Discussion:** Refer again to Table 2 for values for type I and type II errors from both algorithms. Type II error which involves understanding of word usage in text is the source of more inaccuracy than is type I error, which is somewhat correctable with a broader resource for place name resolution. Both types of error bedeviled both algorithms, but GeoRef performed better of the two. For the GeoRef algorithm, type II error is high. This is mostly the result of the algorithm's counting place words in the names of titles or organizations, and cases of metonymy. It is likely the result of Placemaker's more effective rules determining whether a location word is a location that explains why type II error is lower than GeoRef's type II error. Type I error is lower for the GeoRef algorithm than Placemaker, probably because the GeoRef gazetteer was adapted to the text domain by enriching it with local regions in the Sudan.

Random error is introduced by the nature of the sample, so that some texts will have more or simpler place names than other texts. Systematic error is introduced by the fact that the manually-wrought gold standard was flawed. In the case of this particular annotation set, a few locations were missed as a result of fatigue or carelessness, but we used this answer key anyway to evaluate both GeoRef and Placemaker. That means that a few locations found correctly by the algorithms were scored wrong.

The sample size is adequate to compare the algorithms' performance. Even so, the experiment should be repeated on at least one different data set from a different set of news or other sources. Our method is wholly generalizable to any country or region, and can be fit to a different domain with gazetteer adjustment. Whether Placemaker would perform as well in another domain, we do not know.

- **Limitations:** The experiment compares result output of our GeoRef and Placemaker. Without the ability to hold either the external sources or the rules constant between GeoRef and Placemaker, we cannot look into exactly what is going on. In terms of external sources, Placemaker uses the Yahoo! GeoPlanet web service which stores about six million named places, including administrative areas, variant names, and points of interest.[11] GeoRef uses upper hierarchical levels of GeoNames for its backbone, supplemented with specific local place names for the Sudan.

5.2. Experiment 2: Name-Location Pairing

- **Objectives:** We use person-location pairs in the form of standard network pin diagrams and geographic maps to show network change over time.
- **Data:** The data consisted of the same 11 files from the 2006 Sudan Tribune that were used in Experiment 1.
- **Method:** We followed the method set forth in the "explanation for how social network analysis was performed" as outlined above.
- **Procedure:** After the new articles were downloaded, we adapted external sources of personal name list and a gazetteer of place names to find person and place names within the text. We coded the GeoRef algorithm with rules to help identify place names and remove common words likely to be misidentified as places. The algorithm created person-location-date tuples, and assigned geographic coordinates to the tuples so that each may be plotted on a map. Finally, the GeoRef results were input into the third party social network software, *ORA, the Organizational Risk Analyzer (Carley et al., 2010b) to create two types of visualizations. ORA converted the data into DyNetML files. DyNetML is XML-derived language for social network data that facilitates data interchange among data gathering, analysis and visualization tools (Tsvetovat, Reminga, & Carley, 2004).

The tuples were ordered automatically by date and then separated into three groups to balance the number of nodes per group. The result was three sub-networks. The same sub-network is shown twice in the Results section, once as network diagram and again from the ORA geospatial visualizer as a map. The full output is then 3 sub-network diagrams, and 3 sub-network maps.

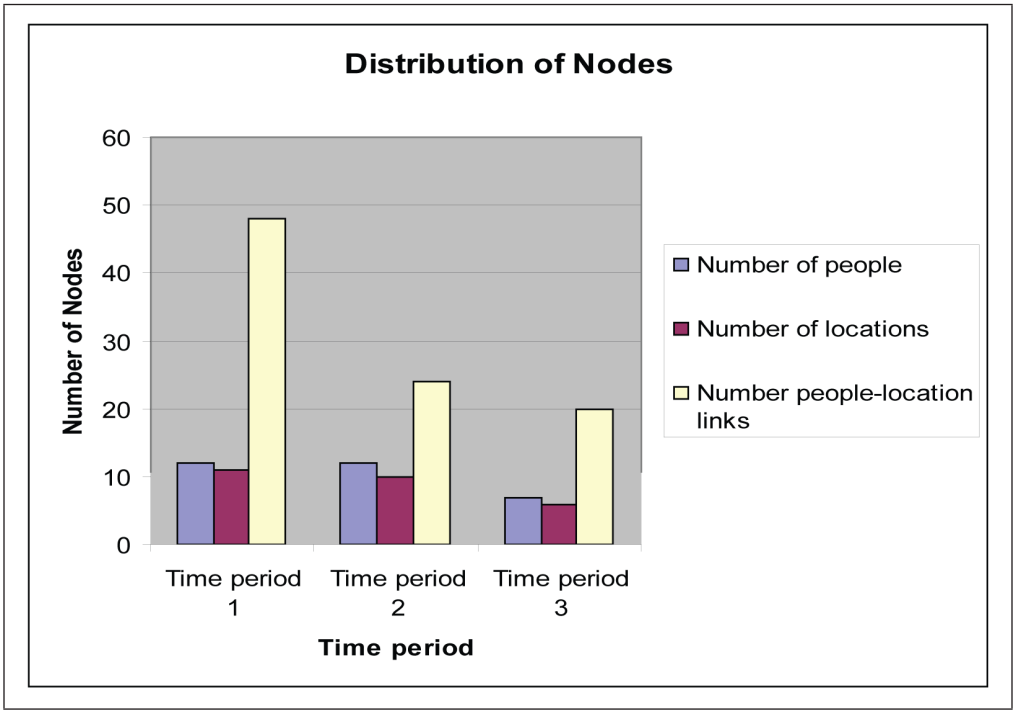
- **Results:** Table 3 compares sub-networks for the three time periods.

The networks display different densities, as represented by the varying number of people-location links. More links suggest a denser network, which in turn suggest more group activities in the same place. Table 3 reflects the division of the news articles into three time-sequential sub-networks. We divided the data to make about the same number of persons per group. This division by node left each network unequal in duration (network 1 and 2 each cover a period of about one month, whereas network 3 covers about two months). The amount of time elapsed between network visualizations also is unequal, with there being about six weeks between the first two networks, and about two weeks between the second two networks. Were we to divide the data into different time period, different sub-networks would result.

We balanced data such that the three networks have comparable *person* node density. The reason we divided the data according to the number of actors rather than, say, number of person-location links, is that actors are the social network core. For social network analysis purposes, it matters less that the time period divisions are unequal than that the networks are *sequential.* We could use time sequence to infer missing data, or to conjecture about what the next network might look like. The larger the data sample, the more sure we would be of our inferences.

- **Network Diagrams:** The network diagrams help answer our research questions and demonstrate the viability of our method. The "a" figures (Figure 2a, Figure 3a, and Figure 4a) resemble standard network diagrams. They show nodes for people (circles) as distinct from nodes for location (hexagons). The network in Figure 2a is more connected than those in Figures 3a and 4a. In our context, connectedness usually comes from the fact that more than one person is tied to the same place, although in a few cases in our data, it means that a person is tied to more than one place. Our notation of Agent (6) indicates that 6 people are tied to the same notation. As an

Table 3. Node distribution in each of the three time periods, with network 1 corresponding to time period 1, network 2 to time 2 and network 3 to time 3



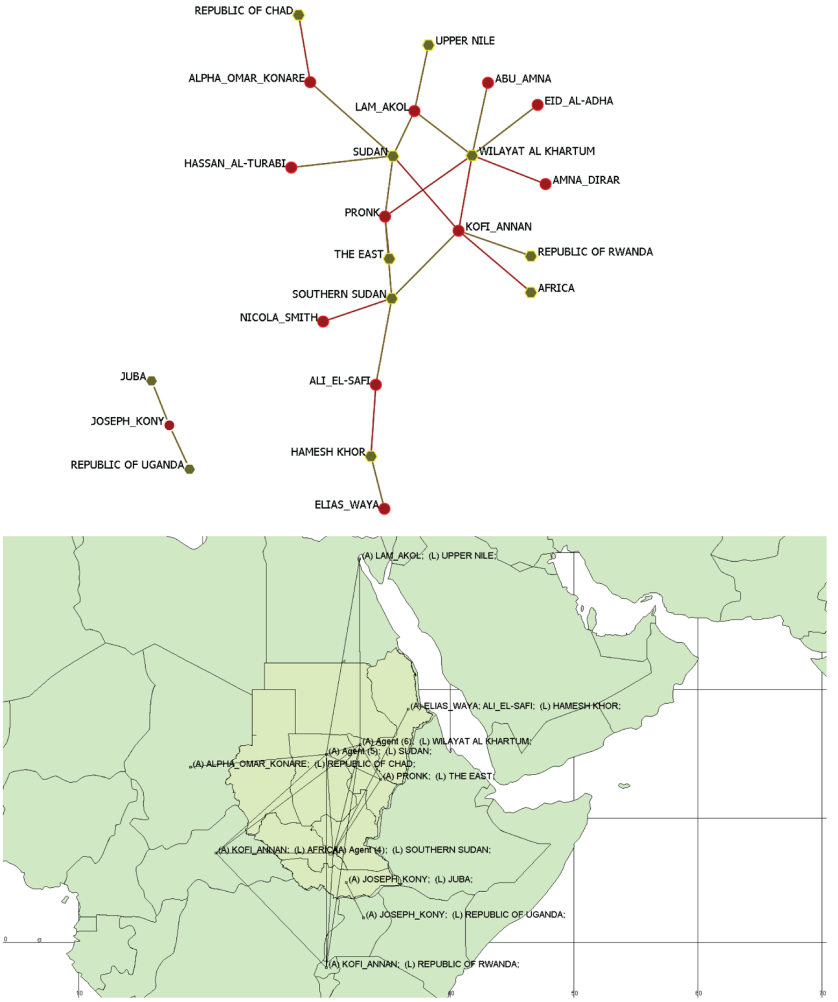
alternative to the lines between person and location on the geographic maps, we could have put a dot labeled with the person's name above the correct location on the map. Our representation makes location and person nodes separate.

The same data of the "a" figures is shown geographically in the "b" figures (Figure 2b, Figure 3b, and Figure 4b). No chronology is shown within each sub-network, but each three diagram set the sub-networks in sequence. Each of the three sub-network diagrams and sub-network maps manifests a different network centrality. The diagrams depict *network centrality* (who is the leader of the social group) while the maps show *geographic centrality* (who is centrally located). The fewer the connections among people nodes, the less the group resembles a social network. Lack of connection among people is shown as a lack of lines among entities in the Figure 3b and Figure 4b geographic diagrams.

Period 1 (Figure 2a, 2b) represents a time interval within the month of January 2006. It shows a fairly well-connected network, as demonstrated by ties among entities. Compare the two types of representations for the first sub-network. In Figure 2a, foreign diplomats Kofi Annan and Jan Pronk are socially important. In Figure 2b, with its geographic emphasis, we see Kofi Annan's influence is important in the south, whereas Jan Pronk's influence is important in the east.

Period 2 in Figure 3a and Figure 3b represents a time interval between mid-March and mid-April 2006. There are almost no relationships among people, which is shown in the diagrams as almost no connecting lines. These are not networks in the true sense, since a network is a set of nodes tied to one another. A

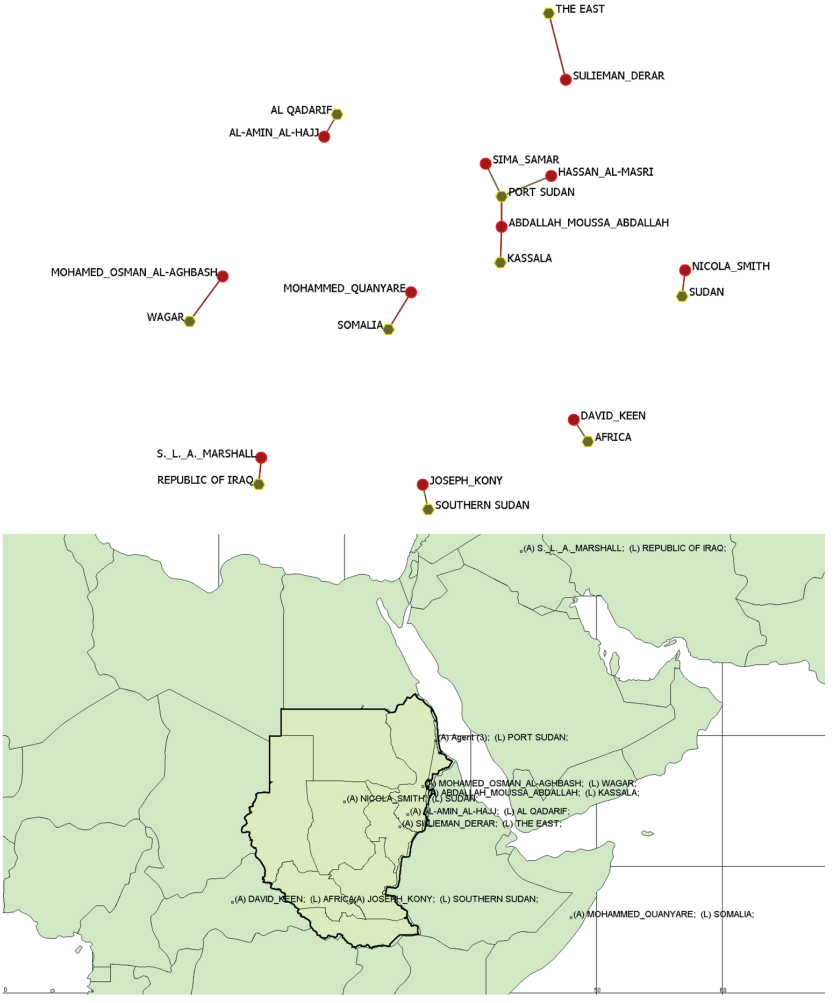
Figure 2. a) Temporal network diagram for period 1. People are represented by circles, places by hexagons. b) Spatiotemporal diagram for period 1. People are presented by points ("A" is for Agent) situated atop their associated locations ("L" is for Location); the darkened region is the Sudan



larger time slice or a lot more data likely would have been better to show networks.

Period 3 represents data harvested in a two-month interval from early May to early July 2006. It is a fairly unconnected network as in the middle period, although there is some activity in Sudan's east. Another presentation might simply omit the locations in Turkey and Great Britain which do not belong to the group. "Igdir" is an actual location in Turkey, but it appears to be an error in that it is an extreme outlier on the geographic map. In fact, "Igdir" must have been an error since we did not find this location in any of the texts. We show it here nonetheless because it was the output of GeoRef.

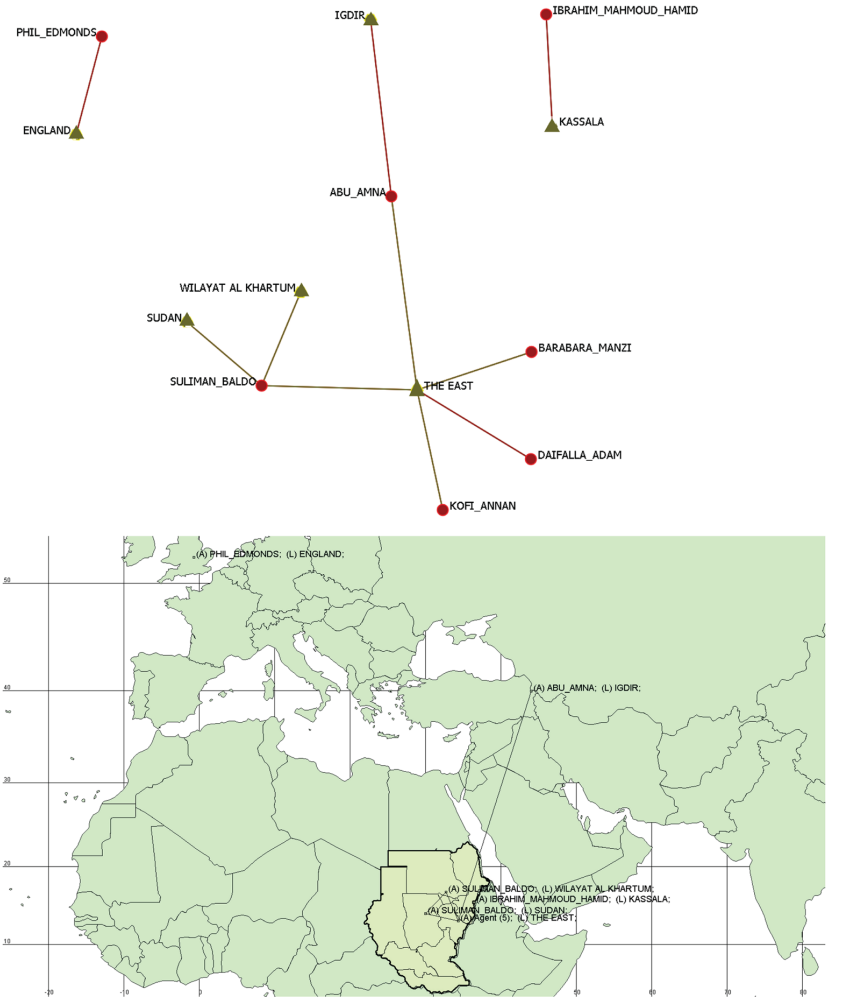
Figure 3. a) Temporal diagram for period 2. People are represented by circles, and places by hexagons. b) Spatiotemporal network diagram for period 2. People are presented by points ("A" is for Agent) situated atop their associated locations ("L" is for Location) within the Sudan



6. DISCUSSION

- **Nodes:** When social network locations are mined from a text, we are unlikely to have locations at the same specificity for all nodes. Some locations mined will be the names of neighborhoods, some states and some continents. Geographers call this mismatch of scale the modifiable areal unit problem. We have decided to include location specifics when known for the sake of helping the analyst who will use the data, rather than making all the location areas larger (all at the level of states or even countries) which would be more consistent.
- **Ties:** The connections that tie different nodes are called edges or links, as mentioned in this paper's introduction. The length of the edges stretches to accommodate geographic distance. Also as mentioned above, the difficulty in data

Figure 4. a) Temporal diagram for period 3. People are represented by circles; places are represented by hexagons.b) Spatiotemporal network map for time period 3 showing Sudan and a few distant nodes in Turkey and Great Britain



mining is determining what these links represent. For example, take the title sentence from the sample text in Table 1, "Kofi Annan belatedly acknowledges the need for robust international intervention to replace AU force in Darfur." The mapped representation shows the politician in the center of the Darfur region. But the nature of the link cannot be known between the mined data of Annan—Darfur, as in this case, the text reveals that politician is not *in* Darfur, he is just referring to it.

- **Visualizations:** Both the pin diagrams (the "a" figures) and the map diagrams (the "b" figures) depict node density. Only the pin diagrams show social network centrality clearly (Figures 2a, 3a, 4a). Network centrality is the number of links each node has. The geographic visualizations (Figures 2b, 3b, 4b) show network centrality, but in a roundabout way, because instead of

each individual shown in a cluster with short connecting lines, each individual is mapped onto his region with connections between individuals stretched to accommodate the geography.

- **Sub-Network Independence:** Each of the sub-networks in the pin diagram set is discrete and there is no time overlap; the same for the map diagram set. This is because our node selection was binary, and each person-location-date tuple was associated with one network only. Should entities on the boundaries of two groups be associated with both in what has been called soft community membership (Lin et al, 2009). The advantage of associating entities with both networks is that it would allow these nodes' associations to be felt within both prior and following network periods, with the disadvantage that it would give these nodes greater voice than others because they would appear twice.
- **Sub-Networks in Series:** Time periods for the sub-networks were arrived at by balancing the number of person nodes, as explained above. It would be more meaningful in the chronological interpretation of the network had the time periods been selected to reflect some natural break in the re-arrangement of the actors due to changing circumstances. There might be some ideal number of nodes or time periods that would make the sub-networks most helpful to data analysts. Our data sample was small enough that we did not do any cluster size optimization.
- **Knowledge Discovery:** The algorithm extracts data from news articles which, when input into third party software, allows at-a-glance visualization of a social network to a degree of accuracy that increases as the size of the data sample increases. We are able to detect who the main actors are at a particular time, with whom they interact, and where they exert their influence. Showing networks in series may suggest patterns to help fill in data that is missing, or help predict a future network. Network diagrams in series have the potential to suggest patterns that will help fill in data that is missing, or help predict what the next stage might show given the present picture. All of this we are able to discover automatically using the algorithm output and visualizations, and without reading a word of the articles.

6.1. Future Directions for Research

- **Data:** Our geo-referencing algorithm, with only minor adjustments, might be applied to either structured or unstructured texts. This is because the rules rest less on strict grammar which may be at odds with unstructured texts than on the basic English language syntax. We recommend that the algorithm be evaluated on text corpora in other domains and other levels of formality. The amount of data to be absorbed into any sub-network in a series, or the relative scarcity or density of the clustering, should be determined by the analyst.
- **Heuristics:** All of our heuristics should be tested with another data set. We are particularly concerned about the rule that 'when an area has more than one place with the same name, the default is to take the place higher in the hierarchy,' and suggest that this should be evaluated further.
- **Location:** To increase the number of location nodes and perhaps thereby the geographic specificity of person-place-date tuples, we could re-define what constitutes a location. We could augment the gazetteer with buildings or landmarks or even national societies along with their locations (United Auto Workers in Detroit, Michigan), or events associated with a region, if such corresponded to the corpus domain. Or we could create a controlled vocabulary in the form of a thesaurus of generic locales, such as room, office, plaza, patio that would pair to named places for further specificity, so that we could pinpoint the swimming pool of the King George II Hotel in Athens, Greece, for example,

when formerly we could only located the tuple to somewhere in Athens.

- **Time:** Presently, time data is mined from the date that introduces each news article. Further research would include mining for time words within the narrative context, whether for season, or historical event or relative time (such as two weeks afterward) that would allow a greater level of time specificity.

7. CONCLUSION

This paper describes the creation and evaluation of our GeoRef algorithm that extracts spatial information from text. The paper describes also the temporal division of algorithm output for social network analysis. Our algorithm enriches mined location data with other levels in the spatial hierarchy (city found in a text will appear in the output along with state or province and country, for example). We illustrate the output both with pin diagrams in series and with geographic maps in temporal series, created with the third-party social network analysis software *ORA. We propose that chunking time in different ways to create alternate sub-networks will provide additional insight into network evolution and will possibly allow prediction. We propose also that, with appropriate modifications in external resources such as person name lists, the GeoRef algorithm could be run productively over a range of domain text to create a social network.

ACKNOWLEDGMENT

The initial Java coding of GeoRef is thanks to Sammy Yelin, with contributions from Dong Cao, Dan Chieffallo and Mike Bigrigg. Thanks are due also to Jon Storrick, Frank Kunkel, Mike Bigrigg, Jana Diesner, Peter Landwehr and all others on the team who spoke with me about this project and answered questions that aided research, and to Kyle Figgatt and Joshua Wilder who helped with data coding. This work was supported in part by the Air Force Office of Sponsored Research (MURI: Computational Modeling of Cultural Dimensions in Adversary Organizations, FA9550-05-1-0388), the Army Research Institute W91WAW07C0063, and the Army Research Office ERDC-TEC W911NF0710317. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) 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 Air Force Office of Sponsored Research, Army Research Institute, the Army Research Office, or the U.S. government.

REFERENCES

Álvarez, M. R., Félix, P., Carinena, P., & Otero, A. (2010). A data mining algorithm for inducing temporal constraint networks. In Proceedings of Computational Intelligence for Knowledge-Based Systems Design (LNCS), (Vol. 6178, pp. 300-309). Berlin: Springer.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging web content. In *Proceedings of SIGIR'04*. New York: ACM.

Buttenfield, B., Gahegan, M., Miller, H., & Yuan, M. (2001). *Geospatial data mining and knowledge discovery*. Retrieved April 7, 2010, from http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emergin g/gkd.pdf

Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. In P. Marsden (Ed.), *Sociological methodology,* (vol. 23, pp. 75–126). Oxford, UK: Blackwell. doi:10.2307/271007

Carley, K. M. (1997). Network text analysis: The network position of concepts. In C. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 79–100). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carley, K. M., Columbus, D., Bigrigg, M., & Kunkel, F. (2010). *AutoMap user's guide* (Technical Report CMU-ISR-10-121) Pittsburgh, PA: Carnegie Mellon University, CASOS Center for Computational Analysis of Social and Organizational Systems. Retrieved October 14, 2010, from http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-10-121.pdf

Carley, K. M., Reminga, J., Storrick, J., & Columbus, D. (2010b). *ORA user's guide 2010* (Technical Report, CMU-ISR-10-120). Carnegie Mellon University, School of Computer Science, Institute for Software Research. Retrieved October 14, 2010, from http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-10-120.pdf

Danowski, J. A. (2009). Automatic mapping of social networks of actors from text corpora: Time series analysis. *Advances in Social Network Analysis and Mining*, 137-142.

Garbin, E., & Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of Human Language Technology Conference, and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP), (pp. 363-370). Academic Press. doi:10.3115/1220575.1220621

Gelernter, J., Cao, D., & Carley, K. M. (2011). (in press). Extraction of spatio-temporal data for social networks. In *Advanced social networking analysis and mining. Springer*.

Giuliano, C., Lavelli, A., & Romano, L. (2007). Relation extraction and the influence of automatic named-entity recognition. *ACM Transactions on Speech and Language Processing, 5*(1), 2:1-2:26.

Jenček, P., Vojtáš, P., Kopecký, M., & Hösch, C. (2009). Sociomapping in text retrieval systems. In T. Andreasen et al. (Eds.), *FQAS 2009, (LNAI), (vol. 5822*, pp. 122–133). Berlin: Springer.

Klamma, R., Cao, Y., Spaniol, M., & Leng, Y. (2007). Spatiotemporal knowledge visualization and discovery in dynamic social networks. In K. Tochtermann & M. Maurer (Eds.), *Proceedings of the I-KNOW*, (pp. 384-391). Graz: Know-Center.

Kodratoff, Y. (1999). Knowledge discovery in texts: A definition, and applications. In *Proceedings of the Foundations of Intelligent Systems* (*LNCS)*, (vol. *1609*, pp. 16–29). Berlin: Springer.

Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*. (pp. 15-21). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/1609039.1609041

Lauw, H. W., Lim, E-P, Pang, H, & Tan, T. T. (2010). STEvent: Spatio-temporal event model for social network discovery. *ACM Transactions on Information Systems, 28*(3), 15:1—15:32.

Leidner, J. L. (2007). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh, UK. Retrieved January 8, 2008, from http://hdl.handle.net/1842/1849

Leveling, J., & Hartrumpf, S. (2008). On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, *22*(3), 289–299. doi:10.1080/13658810701626244

Lin, Y-R, Chi, Y., Zhu, S., Sundaram, H. & Tseng, B. L. (2009). Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery form Data, 3*(2), 8:1-8:31.

Olson, J., Malloy, E., & Carley, K. M. (2009). *Surprised by propinquity: New centrality measures for spatially embedded networks (Technical Report, CMU-ISR-09-xxxx)*. Pittsburgh, PA: Carnegie Mellon University, CASOS Center for Computational Analysis of Social and Organizational Systems.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2006). Discovering verb relations in corpora: Distributional versus non-distributional approaches. In A. Ali & R. Dapoigny (Eds.), IEA/AIE 2006, (LNAI), (vol. 4031, pp. 1042-1052). Berlin: Springer.

Petras, V., Larson, R. R., & Buckland, M. (2006). Time period directories: A metadata infrastructure for placing events in temporal and geographic context. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries,* (pp. 151-160). New York: ACM.

Popescu, A., & Grefenstette, G. (2010). Spatiotemporal mapping of Wikipedia concepts. In *Proceedings of the JCDL '10,* (pp. 129-138). New York: ACM.

Roddick, J. F., & Lees, B. G. (2009). Spatio-temporal data mining paradigms and methodologies. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (2nd ed., pp. 27–44). New York: CRC Press.

Roddick, J. F. & Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research *SIGKDD Explorations Newsletter, 1*(1), 34-38.

Scott, J. (1991 [2004]). Social network analysis: A handbook (2nd ed.). London: Sage.

Srivastava, J. (2008). Data mining for social network analysis. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, (pp. xxxiii-xxxiv). IEEE.

Tobin, R., Grover, C., Bryne, K., Reid, J., & Walsh, J. (2010). Evaluation of georeferencing. In *Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR'10*. doi:10.1145/1722080.1722089

Tsvetovat, M., Reminga, J., & Carley, C. (2004). *DyNetML: Interchange format for rich social network data* (Technical Report, CMU-ISRI-04-105). Pittsburgh, PA: Carnegie Mellon University, CASOS Center for Computational Analysis of Social and Organizational Systems. Retrieved January 21, 2011, from http://www.casos.cs.cmu.edu/publications/papers/CMU-ISRI-04-105.pdf

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511815478

Xu, A., & Zheng, X. (2009). Dynamic social network analysis using latent space model and an integrated clustering algorithm. In YangB.ZhuW. DaiY.YangL.T.MaJ. (Eds.), *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, DASC 2009*. (pp. 620-626). IEEE. doi:10.1109/DASC.2009.59

Xu, X., Mete, M., & Yuruk, N. (2005). Mining concept associations for knowledge discovery in large textual databases. In LiebrockL. M. (Ed.), *Symposium on applied computing*, (pp. 549-550). New York: ACM. doi:10.1145/1066677.1066802

Zhang, H., Yen, J., Giles, C., Mombaster, B., Spiliopoulou, M., & Srivastava, J. et al. (2007). WebKDD/SNAKDD 2007 - Web mining and social network analysis post-workshop report. *SIGKDD Explorations*, *9*(2), 87–92. doi:10.1145/1345448.1345468

ENDNOTES

1 http://unlock.edina.ac.uk/
2 http://www.metacarta.com/
3 http://developer.yahoo.com/geo/placemaker/guide
4 National Geospatial Intelligence Agency gazetteer for download at http://earthinfo.nga.mil/gns/html/
5 http://www.geonames.org
6 In GeoNames, these upper levels of the spatial hierarchy are feature classes of independent and dependent political entities, territories and zones, first- and second-order administrative divisions, and seats of first-order administrative divisions.
7 http://www.openstreetmap.org/
8 The west is typically referred to as Darfur or the Darfur region, which corresponds to states named Darfur, so there is no ambiguity.
9 Download **ORA* from the Carnegie Mellon Computational Analysis of Social and Organizational Systems website as of June 12, 2011 from http://www.casos.cs.cmu.edu/projects/ora/
10 A Yahoo Placemaker web service application, as of January 2011, at http://lerdorf.com/pl.php
11 At http://developer.yahoo.com/geo/geoplanet/guide/concepts.html#woeids on May 12, 2011

Kathleen M. Carley is a professor in the School of Computer Science in the Institute for Software Research International and the director of the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. Her undergraduate degrees are from the Massachusetts Institute of Technology and her PhD in Sociology in 1984 is from Harvard University. Her research combines cognitive science, social networks and computer science to address complex social and organizational problems. Her specific research areas are dynamic network analysis, computational social and organization theory, adaptation and evolution, text mining and the impact of telecommunication technologies and policy on communication, information diffusion, disease contagion and response within and among groups particularly in disaster or crisis situations. She has co-edited several books in the computational organizations and dynamic network area and over 200 articles and developed the widely used ORA tool for dynamic network analysis.

Judith Gelernter is a project scientist in the Language Technologies Institute of Carnegie Mellon University. She received her BA from Yale University, her AM in Fine Arts from Harvard University, and her PhD in 2008 in information science from Rutgers University. She has applied her visual arts background to cartographic information visualization and geo-information retrieval. Her dissertation was published as a book Intelligent Information Retrieval for Maps (2009) Saarbrüken, Germany: VDM [ISBN: 978-3-639-14359-1]. She has published also on aspects of data mining and geoparsing, and has created a new course in Information Systems Department of Carnegie Mellon University which is an introduction to geoinformatics.