SI: DATA TO MODEL

# Rapid modeling and analyzing networks extracted from pre-structured news articles

Jürgen Pfeffer · Kathleen M. Carley

Published online: 19 June 2012

© Springer Science+Business Media, LLC 2012

**Abstract** In the face of uprisings and revolutions happening in several countries within short period of time (Arab Spring 2011), the need for fast network assessments is compelling. In this article we present a rapid network assessment approach which uses a vast amount of pre-indexed news data to provide up-to-date overview and orientation in emerging and ongoing incidents. We describe the fully automated process of preparing the data and creating the dynamic meta-networks. We also describe the network analytical measures that we are using to identify important topics, persons, organizations, and locations in these networks. With our rapid network modeling and analysis approach first results can be provided within hours. In the explorative study of this article we use 108,000+ articles from 600+ English written news sources discussing Egypt, Libya, and Sudan within a time period of 18 months to show an application scenario of our approach. In particular we are looking at the involvement of other countries and their politicians during time periods of major incidents.

**Keywords** Rapid network analysis · Rapid assessment · Network text analysis · Dynamic networks · Two mode networks · Weighted networks

#### 1 Introduction

News articles from a large number of newspapers, news agencies, and magazines provide daily information about incidents in every country of the world. In case of larger events, e.g. the protests and revolutions during the Arab Spring 2011, a single web archive (e.g. LexisNexis) makes available hundreds of news article for a

J. Pfeffer (⋈) · K.M. Carley

CASOS, ISR, SCS, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

e-mail: jpfeffer@cs.cmu.edu

K.M. Carley

e-mail: kathleen.carley@cs.cmu.edu



specific country by day. The analysis of the content of such written information has been a challenge over the last decades (Holsti 1969). Extracting semantic networks and meta-networks from texts is a widely-used approach to gain insights into complex socio-cultural systems encoded in unstructured text data. Network text analysis (Carley 1997) creates networks of words and underlying concepts connected by co-occurrence within the same sentence, paragraph, or document. The extraction of actors (people, organizations, etc.) from unstructured text data as well as the reconstruction of the social structure (Diesner and Carley 2005) of these actors is a key task of the data-to-model process of network text analysis (Diesner and Carley 2004). This process combines methods from text mining and machine learning (Diesner and Carley 2008) with the knowledge of subject matter experts.

In opposite to network text analysis which uses unstructured text data to extract networks, Batagelj et al. (2002) suggest to using structured texts to create semantic networks. A structured text is connected to keywords or tags describing the content of the text, e.g. scientific article are index with keywords by their authors or afterwards by human indexers (Bertrand and Cellier 1995). Pinning tags to texts in the web and sharing bookmarks are further manual ways to structure text data (Hotho et al. 2006) by crowdsourcing the work force of a large number of people (Howe 2008). In the context of news articles index items (Mayfield et al. 2002) are normally assigned automatically to predefined concepts through text categorization (Dumais et al. 1998) using decision rules (Apte et al. 1994), Bayesian networks (Chickering et al. 1997), or other machine learning approaches.

When applied to text data discussing a specific country, network text analysis (Carley 1997) can provide a rapid ethnographic assessment (Harris et al. 1997) of the structure and dynamics of this country (Diesner and Carley 2010). This rapid ethnographic assessment data-to-model process (Carley et al. 2011a, 2011b) can be accomplished in some weeks, which is a huge gain of time compared to traditional scientific assessment approaches. Nevertheless, in the face of the Arab Spring, when protests and revolutions start in a dozen countries within a short period of time, even this processing time turns out to be insufficient when scientists as well as decision makers want to get a first overview of a new incident and the pre-incident conditions.

In this article we introduce a data-to-model process to utilize automatically prestructured text data for a rapid network text analysis method. Using texts, which are already indexed or categorized, result in a tremendous saving in time. In combination with optimizing other steps of the data-to-model process, i.e. the network handling and the analysis of dynamic networks, a rapid network assessment can be offered. This rapid network assessment can provide first analysis of countries with emerging or ongoing conflicts or natural disasters within hours. We do not discuss text mining or text categorization here, but use pre-structured data from news article to extract and analyze networks and elaborate this data-to-model approach in detail. We focus on the index items and their relevance scores for articles provided by the LexisNexis Smartindexing system. Using these pre-structured index items allows us to create a fully automated data-to-model process which is a pre-condition for a rapid analysis of networks based on texts.

Creating networks based on co-occurrence of items in common article includes folding two-mode networks of articles and items to one-mode networks of items.



Even two-mode networks have a long tradition in social network analysis (Davis et al. 1941), the networks which we create in this article have a barely discussed characteristic—they are weighted with relevance scores. Therefore, we also discuss the implication of these weighted networks to the folding process and for the measures used to calculate important persons, organizations, and topics. We demonstrate the application of the data-to-model process and the network analysis in the explorative study of the article which covers three countries with different incidents in a similar time period: the revolution in Egypt, the war in Libya, and the separation into two countries in Sudan. By analyzing 108,000+ articles from 600+ news sources over a time period of 15 months analyze the look at the change during these incidents compared to the pre-incident situation. In particular we are interested in the involvement of foreign countries during the incidents.

In following section two the rapid network analysis approach is introduced and compared to the classic network analysis procedure. In section three we describe the data-to-model process of extracting the data and constructing the networks. The networks created in section three are used for the case studies in section four where we identify important topics, persons, and countries and analyze change over time. In the final section we discuss advantages and drawbacks of the rapid modeling and analyzing approach.

## 2 Rapid network assessment

Networks are relational representations of real world systems. A set of nodes N models agents, organizations, companies, etc. and a set of edges E describes the connection between the nodes. A single edge  $e \in E$  (also called link or connection) connects exactly two nodes  $u, v \in N$ . A network is usually stored in a matrix M and the cell m(u, v) holds information about whether the two nodes u and v are connected. If the links between all connected pairs of nodes are not discriminated by a link weight then the network is unweighted. In case of different link weights the network is weighted. We denote the weight of a link connecting the nodes u and v with w(u, v). A self-loop w(u, u) is a link from a node to itself. If for all u and v w(u, v) = w(v, u) then the network is symmetric, otherwise asymmetric. The networks discussed in this article are weighted, symmetric, with self-loops, and can consist of multiple components.

If the set of nodes in a given network consists of a single group (e.g. people), the network is called a *one-mode* network. When the nodes of a network are concepts extracted from texts and the edges are co-occurrences of concepts, e.g. within a single sentence or document we use the term semantic network (Diesner and Carley 2005). If two different sets of nodes are connected by edges (e.g. people participating at events or items occurring in articles) the network is called *two-mode* (Borgatti and Everett 1997; Faust 1997). A *meta-network* (Carley 2002a, 2002b) is created by two or more different node sets and is therefore composed of several connected one-mode and two-mode networks. If network data is available for more than one time point the term *dynamic* network is used. Usually dynamic networks are created by aggregating data within a specific time period (e.g. one week, one month, etc.) into one *key-frame*. A key-frame of a dynamic meta-network includes the node classes and several networks describing the connections between the nodes.



	TI		
Analysis	Method	Result	Time
High Overview	Rapid Network Assessment, Data Mining Selected Network Analytics	High Level Insight Some Analysis & Comparison	A few hours to a day
Detailed	Rapid Ethnographic Assessment—Text Mining Machine Learning (Data to Model) Network Analytics Simple Simulation	Deep Dive	A few days to a month
Scientific	Multi-Method Simulation	Conclusions Forecasting	Several months

**Table 1** Network assessment approaches

Data Mining		Dynamic Algorithms	Visual Analytics	Modeling
Text Mining	Network Model	Change Detection	Geo Analysis	Simulation

Server Based Multi-User Processes

Fig. 1 Rapid network assessment workflow to analyze dynamic socio-cultural systems

In the context of semantic networks *thesauri* (Carley and Columbus 2011) are used in different contexts. A generalization thesaurus standardizes different phrases with the same meaning (e.g. USA and United States of America). A meta-network thesaurus groups the concepts to different node classes (e.g. agents, organizations). The classification of the concepts into node classes is an essential task to transform the semantic network into a meta-network.

We distinguish three network assessment approaches. Table 1 gives an overview of these approaches. This enumeration can also be seen as an analysis pyramid. With every additional line of analysis, the results get broader but also the time needed for analysis increases. In this article we focus on this first network assessment approach. We discuss the methods and characteristics of this approach in the following.

Figure 1 shows the workflow of the rapid network assessment approach for analyzing dynamic socio-cultural systems. In the following enumeration the challenges are listed where automation and optimization are necessary to make the workflow actually rapidly.

These following nine points can be seen as an agenda towards a rapid network assessment of dynamic socio-cultural systems. In this article we discuss essential parts of this workflow covering the gathering and processing of structured text data, the data-to-network modeling, and the analysis of the dynamic meta-networks.

1. Mining of structured and semi-structured data to gather a large amount of diverse information about a specific socio-cultural system. 2. Text mining of unstructured text data to extract agents, organizations, etc., but also to identify tasks, knowledge, and resources and to find opinions and beliefs (a.k.a. sentiment analysis). 3. Fully automated data-to-model processes to generate dynamic meta-networks. 4. Developing new network analytical algorithms for dynamic meta-networks and optimize existing algorithms to get results within an adequate time period. 5. Efficient change detection algorithms for an up-to-date identification of forerunners of change



in upcoming or ongoing incidents. 6. Innovative visualizations and visual analytic approaches to optimize the computer/analyst interaction and to utilize the human power to capture complex interrelations. 7. Connecting dynamic socio-cultural data with geographical information to identify patterns (trails) in time and space. 8. Modeling and simulation of real world socio-cultural systems based on gathered data and analyzed networks to better understand the underlying dynamics of the systems. Immediate impact analysis to explore possible strategic interventions (e.g. remove one or more nodes or links). 9. Server based multi-user systems to handle the processes and store the data (e.g. in MySQL data bases) efficiently.

# 3 From news articles to the analysis of dynamic meta-networks

In the following sub-sections the different steps of the data-to-model process and the analysis of the networks are described. To illustrate these steps, data from Egypt, Libya, and Sudan is used. We will use the data gathered and processed in this section for the illustrative study of Sect. 4. First, the news articles are gathered from the LexisNexis news archive. Second, the index items which describe the content of the articles are extracted. Third, the data is cleaned to improve quality of the following analysis. Forth, dynamic meta-networks are created based on the co-occurrences of items. Fifth, we describe some network measures which we have used for the analysis of the explorative study in Sect. 4. Finally, the technical environment for the data-to-model process and the network analysis is described at the end of Sect. 3.

#### 3.1 LexisNexis news article archive

The news data used for this article is part of the CASOS/CMU Arab Spring data covering 18 Northern African and Middle East countries. The data is gathered from the LexisNexis Academic<sup>1</sup> news archive. Selecting "Major World Publications" includes about 600 English written news sources (newspapers, magazines, press agencies). The time period of the data includes July 2010 to September 2011. This time period covers the revolution in Egypt, the uprisings and the following war in Libya, and the referendum of independence of South Sudan as well as the separation into two countries. The first ten months of the data were gathered in April and May 2011, the following months of data where extracted five to ten days after the end of each month.

The *LexisNexis Smartindexing* system (which we discuss more in details later in this article) allows for searching for articles discussing specific topics. Adding the geographical index terms *Egypt*, Sudan, and *Libyan Arab Jamahiriya* results in all articles indexed with at least one of the three country items. As the maximum number of articles to search for and to export is limited to 3,000 articles, the news articles are selected and exported by day. The 457 days of the observation period include 108,505 articles for the three countries of interest.

International media coverage often connects incidents in different countries with each other e.g. the beginnings of the uprisings in Libya are often compared to the incidents happened in Egypt some weeks before. Articles describing the war situation

<sup>&</sup>lt;sup>1</sup>http://www.lexisnexis.com/hottopics/lnacademic/?



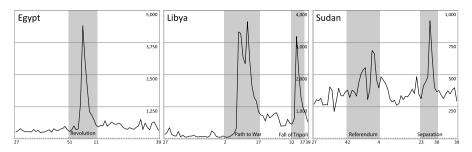


Fig. 2 Articles per country by week for Egypt, Libya, and Sudan

in Libya and discussing differences to the situation in Egypt connect the conflict topics with Egypt. This leads to a high importance of conflict and war topics in articles discussing Egypt even the situation was relatively peaceful in Egypt at this time.

To avoid artifacts based on these cross references, we make use of a feature of the LexisNexis Smartindexing system; every index item of an article is scored with a relevance value which rates the importance of the item for that specific article. We use these relevance scores to assign every article to a single country depending on the highest country value. Doing so, we gathered 38,228 articles for Egypt, 45,152 for Libya, and 25,125 for Sudan. Figure 2 gives a first overview of the amount of articles over time. The aggregation level is by week (Monday to Sunday) excluding the first four days of the observation period to create 66 seven day weeks. The scaling of the y-axes is different for the three countries; the peaks of high media interest are much higher in Egypt and Libya than in Sudan. In the Egypt chart the escalation of the incidents resulting in the Egypt revolution in January and February 2011 are well documented with a huge peak in media coverage. In Libya we can see two large peaks of media coverage. The first peak is reached during the escalation of the demonstrations in the second half of February and lasts until the start of the war in Libya in the second half of March ("Path to War"). The second peak in Libya's media coverage marks the second half of August when the rebels took over Tripoli and the last part of the Libyan war started ("Fall of Tripoli"). The picture for Sudan looks different as the change of media awareness between high and low media coverage is less extreme. The first increase in media coverage towards the referendum of independence of South Sudan ("Referendum") is paused the Christmas holidays of western journalists and media. This effect is also visible in the other charts. The second, higher but shorter, peak is caused by the actual independence of South Sudan which took place on July 9th 2011 ("Separation"). These media peak periods (marked gray in the charts in Fig. 2 are in the focus of the further analysis in Sect. 4.

## 3.2 Index items and their relevance scores

In the previous sub-section we mentioned the LexisNexis Smartindexing items and their relevance scores. In this sub-section we take a closer look at these index items. The index items are provided to help users of the media archive to better find their interested articles. After exporting the full articles with the index items these index



J. Pfeffer, K.M. Carley

**Table 2** Number of items per category for the countries of interest

Category	Egypt	Libya	Sudan	All
Subject (S)	3,690	3,317	2,810	4,029
Geographic (G)	761	707	630	794
Person (P)	693	688	314	979
Organization (O)	553	421	337	807
Company (C)	1,887	1,611	559	2,993
Industry (I)	882	837	465	1,097
Ticker (T)	1,279	1,163	336	1,872
All Categories	9,745	8,744	5,451	12,571

items can be extracted from the news articles. Index items have the following representation:

SUBJECT: PROTESTS & DEMONSTRATIONS (90 %) GEOGRAPHIC: EGYPT (95 %); YEMEN (80 %)

The 108,505 articles gathered from Egypt, Sudan, and Libya are described by 16.4 index items on average resulting in 1,778,849 index items. The index items provide two additional features which are very useful for the data-to-network process described in this article. First, every index item is connected with a relevance score to an article. Looking at the three index items enumerated above shows a 95 % relevance of the index item *Egypt* for this specific article, while the relevance score for *Yemen* is just 80 %. Without reading the article these index items indicate an article discussing the protests in Egypt and referring to incidents in Yemen. The relevance scores provided by LexisNexis are within the range of 50 % and 99 %. The second feature of index items is the classification of all items into seven different groups. We will use these groups later as different modes of our meta-networks.

At this point we take a closer look at the distribution of the items of the seven categories. Table 2 lists the number of items per category for Egypt, Libya, and Sudan. *Subject* items are topics and describe the content of the article (e.g. conflicts, economic news, human rights). *Geographic* items include countries, cities, or regions (e.g. Middle East). *Persons* are primarily politician but also other well-known people (e.g. movie actors, athletes). International organizations (e.g. NATO), terror groups (e.g. Al-Qaeda), and other associations are groups to the category *Organization*. *Company* describes business organizations, *Ticker* represents the stock exchange abbreviations of companies, and *Industry* assigns business related articles to an industry classification system (NAICS). Table 2 shows a high proportion of business related index items (the last three categories) presuming a large number of articles discussing business topics.

## 3.3 Data cleaning

The LexisNexis Smartindexing system is based on automatically assigning predefined index items to news articles. The decision whether an item connected with an article or not is rule based. This top-down approach results in very *clean* data. The



index items can be almost immediately used for statistical analysis or for the further network generation process. Nevertheless, we add some data cleaning steps to our data-to-model process to increase the quality of the analysis outcome. The first step of data cleaning is a meta-network thesaurus. The LexisNexis Smartindexing system groups every item into one of the seven categories enumerated in Table 1. This categorization is not always stable, i.e. one item can be put in two different categories in different articles (e.g. *United Nations* is an *organization*, but sometimes it is labeled as *company*). To automatically decide which classification is correct, we suggest two simple rules.

- 1. The category from the younger article is more likely to be correct than from the older article.
- Classifications which are assigned more often are more likely to be correct than those which are assigned rarely.

The second cleaning step is a generalization thesaurus. This kind of thesaurus is used to convert different ways of referring to the same concept to a common form. The process of standardizing writing is a very essential part of text analysis data-to-model processes because insufficient standardization has very negative impact on subsequent analysis. In case of using pre-structure texts with index items provided by a third party, this cleaning step can be neglected. As the indexing process happens through a top-down approach with pre-defined index items, different writings of the same concept are the result of an inaccurate item list rather than of the underlying text data. When adding a second source of pre-structured text data, this issue is again important because different indexing systems use different index items. As the data used for this article uses a single source, the generalization thesaurus consists of a very small number of items. Finally, deleting multiple index items within one article, which can be a result of the applied thesauri, is a last and easy to process data cleaning step.

Even we do not apply additional data cleaning procedures to the data used for this article, we want to point the attention of the reader to another possible data cleaning step when dealing with data from news archives. As in news archives one article can occur more than once, e.g. an article written by a press agency is used by different newspapers, a de-duplication of identical articles is a widely used method to avoid overrepresentation of specific topics. Normally, identifying *identical* articles is a nontrivial and fuzzy task because changing or adding some words results technically in a new article while actually it is just another copy of the same article. In case of indexed articles it is possible to easily and rapidly identify identical articles. The index items (and their relevance scores) can be seen as lower dimensional representations of the articles and there is very high chance that articles, which are mapped to the same 10 to 15 index items with identical relevance scores, are actually identical.

In classical text analysis procedures (Diesner and Carley 2005) cleaning the data, creating the list of interesting concepts, and assigning the concepts to meta-network classes is very time consuming and requires the knowledge of subject matter experts. When using pre-structured text data which is indexed by a third party, we outsource these steps. This results on the one hand in a tremendous gain of time for the data-to-model process but, on the other hand, puts us into a position of dependency on the quality of the applied indexing systems.



#### 3.4 Semantic networks and meta-networks

In the first part of the data-to-model process we gathered the news articles and prepared the index items. These index items, together with their relevance scores, can be used for statistical content analysis to show important topics for selected countries (e.g. concept frequency) or to show the change of topics over time. We will show some statistical analysis in the explorative study in Sect. 4. In the following we use the index items to create semantic networks and meta-networks.

Persons, Organization, etc. which occur in news articles can be used to construct networks. The underlying assumption of this process is that items co-occurring together in the same articles are somehow connected with each other. The networks created in this way have the items described previously as the set of nodes and the co-occurrences of items in articles generate the set of edges. The articles and their enumerated items can be interpreted as a two-mode network. The process of creating networks by looking at co-occurrences of items within the same articles is equivalent to the process of transforming two-mode networks to one-mode networks and is called network folding. When folding the document/item network, the resulting one-mode network is a *semantic network* of index items. The links of the semantic network describe the co-occurrence of items within the same articles. We discuss the calculation of the link weights of this network later. Prior to that, another transformation step is introduced. As the items are grouped into seven node classes, we are able to create a lot of different networks which are sub-networks of the semantic network. Table 3 shows the different one-mode and two-mode networks which are combined to a meta-network, e.g. the SP-network represents the co-occurrences of subject and person items while the CC-network holds the co-occurrences of companies.

When folding two-mode networks to one-mode networks the link weights of the new one-mode networks have to be calculated. In case of two-mode data based on people and their participation at the same events (Davis et al. 1941) the link weight between two persons is the number of events they participated in together (Borgatti and Everett 1997). This approach could also be applied to items and co-occurrence in news articles. But, the news data which creates the item/document connections of our data is weighted. Every item is connected to an article with a relevance value. The higher the value is the more relevant is an item for a specific article. When counting the number of item co-occurrences we lose this relevance information. To illustrate

Table 3	Converting a two-mode item/document network to a seven-mode meta-network
---------	--

	Document (D)			S	G	P	О	С	I	T
Subject (S)	DS		Subject	SS	SG	SP	SO	SC	SI	ST
Geographic (G)	DG		Geographic		GG	GP	GO	GC	GI	GT
Person (P)	DP	$\rightarrow$	Person			PP	PO	PC	PΙ	PT
Organization (O)	DO		Organization				OO	OC	OI	OT
Company (C)	DC		Company					CC	CI	CT
Industry (I)	DI		Industry						II	IT
Ticker (T)	DT		Ticker							TT



Table 4	Two exam	ple articles
---------	----------	--------------

Article 1	Article 2
0.95 Libya (Geographic)	0.99 Egypt (Geographic)
0.95 Human Rights (Subject)	0.95 Protests & Demonstrations (Subject)
0.55 Protests & Demonstrations (Subject)	0.55 Human Rights (Subject)



Fig. 3 Visualization of the co-occurrences of the two example articles, left: number of co-occurring articles, right: weighted folding

this problem and our suggested solution we assume two articles, each connected to three items (see Table 4).

The different relevance scores suggest one article discussing primarily the topic of human rights in Libya while also protests and demonstrations are mentioned and another article where Egypt as well as protests and demonstrations are in focus while human rights is connected by a lower relevance score. Counting the number of co-occurrences leads to the left picture in Fig. 3; both subject items are connected to each of the country items with the same link weight and the two subject items are connected with a link weight of 2.0 because they co-occur in both articles. The right visualization of Fig. 3 shows the same folded network but the link weights are calculated based on a weighted folding approach. Instead of counting every co-occurrence equally (with the values 1.0), the relevance scores are taken into account by using the minimum value of the items for every co-occurring article, e.g. for *Human Rights* and *Protests & Demonstration* the new link weight is  $\min(0.95, 0.55) + \min(0.55, 0.95) = 1.1$ . The differences of the relevance scores remain in the folded network.

Taking the minimum of the relevance scores views an article as a collection of overlapping topic areas where the maximum size of the intersecting area is limited to the size of the smaller area. Following this logic we could also multiply the two values. The generalization of this approach can be found in Eq. (1). If a one-mode item/item network is created from a two-mode article/item network with |A| articles, then the link weights of the folded one-mode network is the average value of a link weight function f of the two-mode network weights. For the purpose of this article we use *minimum* for the link weight function. Of course, every possible other function can be used (e.g. maximum, average, product). The selection of the link weight function is dependent on the content of the networks, e.g. when folding two-mode networks of people and knowledge the minimum could be reasonable if the content of the person/person interaction is expert communication, in cases where the



person/person interaction is used to solve problems, a maximum folding function is more accurate.

$$\omega_{ab} = \frac{\sum_{i=1}^{|A|} f(\omega_{a_i}, \omega_{b_i})}{|A|} \tag{1}$$

Folding the networks using Eq. (1) offers another feature when looking at the case a = b when applying minimum, maximum, or average link weight functions; we can use the diagonal elements of the one-mode networks of our dynamic meta-networks to store the average relevance score of an item itself. Therefore, the statistical analysis of the relevance of items over time (see Sect. 4.1) can also be achieved with the dynamic meta-networks.

## 3.5 Network measures in weighted networks

To identify important nodes in networks, Freeman's (1979) centrality measures, degree centrality, betweenness centrality, and closeness centrality, are widely-used. Freeman defined these measures for un-weighted, symmetric, and connected networks which do not have self-loops. The networks created through the data-to-model process described in this article are weighted, and can be unconnected. Wei et al. (2011) enumerate the different adjustments for the most commonly used social network measures in case of weighted, asymmetric, self-linked, and disconnected networks. In the following paragraphs we focus on the implications for weighted, but symmetric networks when calculating degree centrality (Freeman 1979), betweenness centrality (Freeman 1977), and eigenvector centrality (Bonacich 1972). For calculating these measures we ignore self-loops in the networks. We use these measures in the following section for the illustrative study. Applying weighted measures provides another advantage. We do not need to discuss the issue of deleting lines with a weight lower a certain threshold because these links are automatically assigned little importance by the weighted measure algorithm.

Degree centrality (Freeman 1979) reports the number of connection of a node. In case of weighted networks, the sum of the line weights is calculated rather than the number of lines. Equation (2) shows the modified formula for the normalized degree centrality for weighted networks. As our networks are symmetric and exclude self-loops, just one half of the matrix without the diagonal elements is taken into account. The divisor of Eq. (2) is the normalization to bring the result of every node within the range of 0 and 1. Normally,  $w^*$  is the value the maximum link weight in a network matrix (Wei et al. 2011). For the networks created with the folding process described above we know the maximum possible link weight in the matrix;  $w^* = 0.99$  in case two articles occur together in every article with the maximum possible relevance values of 0.99.

$$C_D'(k) = \frac{\sum_{v=1, v>k}^{|N|} w_{v,k}}{(|N|-1) \cdot w^*}$$
 (2)

Betweenness centrality (Freeman 1977) is used to measure the extent a node is between other nodes. To calculate betweenness centrality the shortest paths between all pairs of nodes are calculated and nodes being on these shortest paths are awarded with



betweenness scores. Equation (3) shows the formula for the normalized betweenness centrality in symmetric networks. The formula remains unchanged from Freeman (1977) but the breath-first search algorithm used to actually calculate the shortest paths between the nodes (Brandes 2001) is altered to a weighted breath-first search procedure (Wei et al. 2011). The weights calculated in our networks are similarity weights, i.e. the higher the value the stronger is the connection between two nodes. On the opposite, the weighted breath-first search procedure uses weights as distances, i.e. links with smaller weights are used earlier than links with higher weights. Therefore, the link weights of the dynamic meta-networks have to be inverted by subtracting from a fixed value which is slightly larger than the highest possible value: w(a, b) = 1.0 - w(a, b) if w(a, b) > 0.0.

$$C_B'(k) = \sum_{u=1}^{|V|} \sum_{v=u+1}^{|V|} \frac{2[g_{u,v}(k)]}{g_{u,v}[|V|^2 - 3|V| + 2]}$$
(3)

Eigenvector centrality (Bonacich 1972) is a measure to identify powerful nodes in a given network. Eigenvector centrality takes weighted networks into account but lack to cover another network characteristic. Wei et al. (2011) point to the artifacts in the context of eigenvector centrality in case of unconnected networks. As we cannot guarantee that our networks are connected, we calculate eigenvector centrality per component normalized with the component size  $N_k$  of the component including node k to identify powerful nodes in our networks (Wei et al. 2011). The additional normalization of Eq. (4) with  $\sqrt{0.5}$  describes the maximum possible result of unscaled eigenvector centrality for a single node.

$$C_E'(k) = \frac{\frac{1}{\lambda} \sum_{v=1}^{|V|} w_{k,v} C_E(K)}{\sqrt{0.5}} \frac{N_k}{N}$$
 (4)

## 3.6 Data handling and measure calculations

The news articles are exported from the LexisNexis Academic web site using the data export procedure offered by the website. The fully automated pre-processes steps of cleaning and preparing the data are organized within SORASCS (Service-Oriented Architectures for Socio-Cultural Analysis) introduced by Garlan et al. (2009). The news articles and the index items are stored in a MySQL database. The dynamic metanetworks which are extracted from the data base are stored in the DyNetML data format (Tsvetovat et al. 2004). The calculations of the network measures over time are accomplished with ORA (Carley et al. 2011a, 2011b). Exporting the data, preprocessing the index items, generating the dynamic meta-networks and calculating the measures one can find in the following illustrative study takes about one hour for one month of data.

### 4 Explorative Study on Egypt, Libya, and Sudan

In Sect. 3 we discussed the data-to-model process of transforming the index items of news articles to dynamic meta-networks of Egypt, Libya, and Sudan. We already



292 J. Pfeffer, K.M. Carley

showed the number of articles over time in Fig. 2 as well as the number of nodes in the seven node categories for the three countries in Table 2. In this section we take a closer look at the networks over time. We first provide statistics over time for the most relevant subject items per country. Second, we use the dynamic meta-networks to identify important persons, organizations, and geographic items over time. The common focus of our analysis will be the involvement of other countries during the incidents happened in the three countries. For the following calculations we constructed three (one for each country) dynamic meta-networks based on the procedure described in Sect. 3. These dynamic meta-networks include 9,745 (Egypt), 8,744 (Libya), and 5,451 (Sudan) nodes. The seven node categories create 28 networks (7 one-mode and 21 two-mode networks). The 66 weekly time periods therefore result in 5,544 networks for the three countries. When analyzing the networks, we focus on the specific time periods of interest of the three countries. In Fig. 2 the interesting time periods as well as some weeks before these time periods are marked gray. These are the time periods we will focus in the following analysis to show the change in topics and in the networks happened during the major incidents happened within the 18 months observation period. In particular we the following weeks are analyzed: Egypt 51/2010-11/2011, Libya 2/2011-17/2011 and 32/2011-37/2011, Sudan 42/2010-4/2011, 23/2011-30/2011.

#### 4.1 Topics that dominated media coverage

We first look at the change of the most relevant items over time. The aggregation level of this analysis is by week. The relevance of an item for a specific country by week is calculated by summing up the relevance scores for an item in all articles in which this item occurs divided by the overall number of articles of the country. This first analysis of the explorative study of this article is statistical. Nevertheless, we use the dynamic-meta networks created in Sect. 3 for this analysis as we stored the average relevance scores for every time period in the diagonal of the one-mode networks of the metanetworks. From the list of the most relevant items before and during the incidents we manually select a couple of the subject items of the three countries. In the list of relevant topics of all countries the index item *International Relations* fits our need to examine the involvement of other countries. *War & Conflict* and *Religion* are also selected for all three countries. In addition *Terrorism* and *Protests & Demonstrations* are selected for Egypt and Libya while *Referendums* and *Separatism & Secession* are selected for Sudan.

The charts in Fig. 4 visualize the relevance of these subject items for the selected time periods. The x-axes represent the number of the week starting in 2010 and changing to 2011 with week number 1. The y-axes are the average relevance values of subject items by day. For Libya and Sudan the analyzed weeks are from two separated time periods. Therefore, the charts are interrupted. The gray areas on bottom of all three charts show the relevance of *War & Conflict* indicating very peaceful incidents in Egypt (left picture), a medium conflicting situation in Sudan, and a very high conflict level in Libya caused by the war in this country.

For Egypt, the high level of the topics *Religion* and *Terrorism* at the turn of the year is the result of a suicide bombing outside a Coptic Christians church in Alexandria on



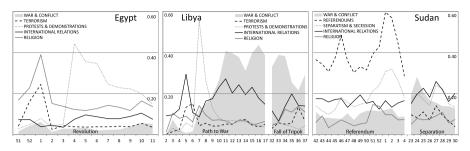


Fig. 4 Important subject items over time for Egypt, Libya, and Sudan

New Year's morning where 23 people died.<sup>2</sup> But these two topics stay on a stable (an in case of *Terrorism* very low) level during the Egypt revolution indicating that neither religious topics nor terrorism are important topics during the Egypt revolution. Looking at the overall data of 15 month from July 2010 to September 2011 shows that *Religion* is constant in a relevance range between 0.10 and 0.15 and *Terrorism* is below 0.05 except for short intervals after incidents like the New Year's bombing. The *Protests & Demonstrations* curve for Egypt traces the escalation of the demonstrations at the end of January, the ongoing protests until the Hosni Mubarak resigned from being president of Egypt, and the following relaxation of the situation. Comparing the evolution of Protests & Demonstrations and International Relations suggests evidence that the international involvement in Egypt is a result of the protests and focuses on the demission of Mubarak.

The situation in Libya, which is illustrated in the second picture of Fig. 4, looks different. In Libya the conflict level is much higher during the protest peak in week 7 and increases constantly towards war. The values of *International Relations* are a sign of that the conflict escalation, which started at the peak of the protests, happened parallel to the increasing involvement of foreign countries. The media coverage of *International Relations* decreased with the progress of the war implying incidents which are more and more considered as civil war by international media. The media coverage of War & Conflict starts to decrease with the collapse of the Gaddafi regime in August 2011. In parallel, the interest in *International Relations* increases again.

Finally, the third picture shows the topic dynamics of Sudan which is dominated by the referendum of independence of South Sudan. Compared to incidents in Sudan during recent year, the importance of conflict topics stood stable on a relatively low level in the weeks before and during the referendum. Also religious topics are little discussed during this time which is also different from the situation in previous years. The separation itself, which took place during the second part of our selected time period, was of much smaller media dominance than the referendum in January 2011. During this time the media coverage of conflicting incidents increased, but also discussion about *International Relations*. The constant, high level of media coverage on International Relations can be seen as an indicator for an increased international awareness of the critical overall situation in Sudan.



<sup>&</sup>lt;sup>2</sup>As week 1 starts on Monday, January 3rd, New Year's Day is part of week 52.

The charts in Fig. 4 show that the average relevance scores by time period are good representations of incidents in specific countries. To analyze these items over time is therefore a good way to retrieve a summary of a large number of news articles. As media coverage reacts immediately to major events, e.g. bombings, protests, etc. also the occurrence of related topic items changes. Hence, Identifying change of single topics or of the overall media coverage of a country is possible with this data. In addition, it is also possible to analyze the pre-incidents situation a posteriori.

# 4.2 The involvement of international acting politicians

To examine the involvement of internationally acting politicians we use betweenness centrality calculations on the one-mode person network. Betweenness centrality scores high for nodes which connect different parts of the network with each other. In the context of networks composed of people and their co-occurrence in media article, betweenness centrality measures the extent people are mentioned with various other people who are not connected with each other. The analysis covers the political leaders of the United States (*Barak Obama*), the United Kingdom (*David Cameron*), France (*Nicolas Sarkozy*), and the Russian Federation (*Vladimir Putin*<sup>3</sup>) as well as the leader of the three countries during the incidents, i.e. *Hosni Mubarak* for Egypt, *Muammar Gaddafi* for Libya, and *Omar Al-Bashir* for Sudan. The betweenness centrality score of each of the heads of the three states is visualized with gray areas in the charts of Fig. 5.

The first and most obvious result when looking at Fig. 5 is the dominant presence of the head of state in Egypt and Libya in contrast to Sudan. This provides evidence that the conflicts in Egypt and Libya were focused at the head of states by media while in Sudan the incidents were not personalized in that way. In general, all three countries show an absence of international politicians during the weeks before the beginning of the different incidents in the countries. *Nicolas Sarkozy*, *David Cameron*, and *Vladimir Putin* are not mentioned at all in these time periods. This is indicated in the charts by an absence of their curves. Having *Barack Obama* as the most between person item is a common characteristic of all three country networks, but in Egypt

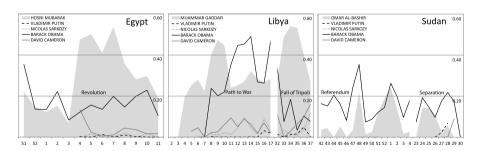


Fig. 5 Important persons over time (betweenness centrality)

<sup>&</sup>lt;sup>3</sup>For Russia Vladimir Putin and Dmitry Medvedev are both players in international media but Putin is more important.



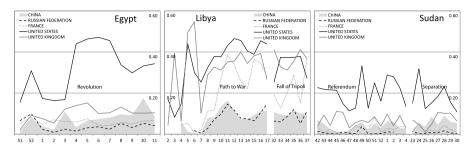


Fig. 6 The importance of other countries (eigenvector centrality)

and Sudan *Barack Obama* is in a central position before the incidents and stays on a rather constant high level during the incidents.

The second time period of interest in Libya is again dominated by the person of *Muammar Gaddafi* who is in the center of media coverage in the final stage of the war in Libya. In contrast, the centrality of *Barack Obama* is dramatically reduced during this time period. In week 37 the item of the US president ranked 4th in our selection of state leaders.

The large difference in the betweenness centrality analysis between different time points is amplified by the measure itself. A small number of added or removed links in a network can result in dramatic structural change which leads to completely different betweenness scores for the central nodes. Having a betweenness centrality of 0.0 does not imply that an agent is without connections, but none of its connections is an exclusive connection between other agents.

#### 4.3 Influential countries

To calculate the impact of other countries to the three countries of interest, we calculate eigenvector centrality (Bonacich 1972) using the GG networks of the dynamic meta-networks created in Sect. 3. But before working with the geographical networks, the items of a country is deleted from the specific networks, e.g. the geographical item *Egypt* is removed from the Egypt network because this item was the criteria for creating the networks and therefore this item would be part of every article resulting in co-occurrences with all other items. The charts in Fig. 6 represent the eigenvector centralities of the five veto-wielded countries of the United Nations Security Council. For the *United States*, the *United Kingdom*, the *Russian Federation*, and *France*, the curve texture is identical to curves of the heads of states of these countries in Fig. 5. As the countries themselves are not part of the analysis, we use the gray area in the charts of Fig. 6 to visualize the centrality score of China.

For Egypt, the eigenvector centrality of the *United States* goes parallel with the *Protests & Demonstration* curve in Fig. 3. But after the peak of protests, the importance of the *United States* remain on high level. We interpret this as a result of the pressure of the US Government on Mubarak to step back from his presidency immediately. When Mubarak final declares his resignation, the importance of the *United States* decreases again. As the betweenness centrality of *Barack Obama* does not increase during this time period (Fig. 5), there is strong evidence that the importance



of the United States is created through different acting politicians (e.g. Secretary of State *Hilary Clinton*, Secretary of Defense *Robert Gates*).

In Libya the eigenvector centrality scores are higher for all five veto-countries during the incidents than in the other two countries. This is a result of the involvement of the United Nations Security Council and its power-wielded countries in the escalation of the conflict in Libya. The no-fly zone resolution of the United Nations lead directly to the war in Libya. In the Sudan network, on the other hand, the eigenvector centrality scores of the five selected countries are lower. In addition, the scores are relatively stable. This is another indicator that the transition in Sudan was a rather calm process without larger incidents.

In literature, eigenvector centrality (Bonacich 1972) is often connected to power. When applied to networks extracted from news articles an item scores high in eigenvector centrality if it co-occurs in articles with a lot of other well connected items. Hence, the *United States* is not only named very often in the context of the Egypt revolution but also the *United States is* often named when other countries are discussed. Therefore, we can use the term *media dominance* to describe this phenomenon.

## 4.4 Complexity of incidents

The complexity of a meta-network can be calculated by looking at the density of all its networks. The density of a network is the number of links divided by the number of possible links, i.e. the proportion of non-zero cells in the matrix (Wasserman and Faust 1995). Consequently, the density of a meta-network is the number of links in all networks divided by the number of possible links in all networks. When looking at density calculations there is one drawback. Larger networks tend to lead to lower density score than smaller networks because the number of neighbors of a node normally does not automatically increase proportionally in larger networks. For that reason we use a second aspect of complexity—the average meta-network degree (the number of edges in the meta-network divided by the squared number of nodes in the meta-network). When analyzing networks based on news articles the average degree reflects the extent of how different topics, people, locations, and organization are mixed together in the same news articles. Figure 7 visualizes the average degree and the density of the meta-networks of Egypt, Libya, and Sudan over time. As we are mainly interested in political activities, we removed the three economic node classes

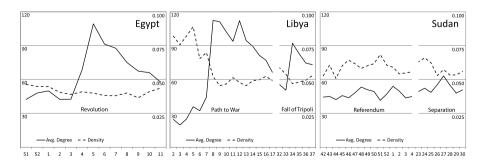


Fig. 7 Meta-network density and average degree for Egypt, Libya, and Sudan



(company, industry, and ticker) before we accomplished the calculations. Once again, the results for Sudan are very stable; the complexity of the networks in Sudan did not change during the referendum or the separation. When comparing the curves of Egypt and Libya on the other hand, interesting similarities of the average degree can be identified. At the peak of escalation during the revolution in Egypt and "Path to War" in Libya the average degree in the meta-network is at about at the same level. The slightly higher density at this time in Libya indicates a slightly smaller number of different people, organizations, locations, and topics involved in media coverage. A Similar number of nodes in the meta-networks of Egypt and Libya and a similar distribution of these nodes within the four node classes are grounds for the assumption that there could be a common structure for the overall extent and the complexity of media coverage of conflicting incidents. But this question has to be part of future analysis.

#### 5 Conclusions

In this article we offered the detailed description of a rapid modeling and analyzing approach for networks extracted from pre-structured news articles. This rapid network analysis approach uses large amounts of semi-structured data (e.g. news articles), fully automated data processing procedures, and algorithms optimized for this data to provide researchers and decision makers with rapid assessment of incidents and change in large scale complex social systems. Based on dozens or hundreds of thousands of publically available information we build up multi-mode networks consisting of topics, persons, organizations, companies, and countries. To analyze these networks over time gives us the ability to detect and describe change. As the modeling and analysis can be accomplished within hours, an up-to-date analysis is possible. The rapid network text analysis approach is therefore an adequate method to gain first insights in countries of emerging uprisings or other conflicts.

The ability to get quick first results implies two important advantages for the research process. First, the rapid results can be utilized to support the planning of the subsequent analysis strategy. Second, researchers obtain a validation of the data and the data-to-model process at an early stage of the analysis. Identifying data or manipulation errors at later date of the analysis process can result in tedious and costly error corrections. Rapid network assessment is closely associated with the ideas of rapid prototyping in software development processes. As first results can be provided within hours, these results can be used to collect feedback from subject matter experts. The data-to-model process as well as the data sources can be modified in case of unsatisfactory results.

Making use of pre-structured text data is rather data mining than text mining. The data-to-model process therefore focuses on gathering and pre-processing the data and neglects other procedures known from text mining data-to-model processes. To extend this approach and to reduce the dependency on a single indexing system, we propose to include additional news indexing systems but also to cover blogs and probably social media sources. Another way to increase the validity of the outcome and to reduce the bias of Western media in the analysis is to add non-English news



sources. But, combining different sources leads to new challenges of data cleaning and generalization. A possibility to avoid these time consuming steps would be to analyze different sources parallel but separately and to compare just the high level result of the analysis.

In this article we focused on three countries and the CMU/CASOS Arab Spring data includes 18 countries. The data-to-model process which we described in this article is not limited to this number of states and can be extended easily to other countries, e.g. at risk countries in Africa or Asia. Consequently, a world-wide monitoring of emerging conflicts of other incidents (e.g. natural disaster, dearth) can be implemented. But also applications in other research areas are conceivable, e.g. possible marketing questions could be: Which topics are connected to which company and does a media campaign of a specific company change the stories which are written about its products?

With rapid network analysis we will be able to monitor and evaluate social changes. Rapid network analysis helps us to describe and *understand* the pre-incident situation which is a precondition for successful interventions. A rapid network analysis system relies on a large amount of data which is available in real time. Adding more data and adding the data faster to the system can help to create analytical results faster. Because of an up-to-date analysis, emerging conflicts or changes in ongoing incidents can be detected immediately. A real-time network analysis which is able to detect forerunners of incidents can help to predict and perhaps prevent specific events.

Acknowledgements This work is supported in part by the Office of Naval Research (ONR), United States Navy (ONR MURI N000140811186, ONR MMV N00014060104). 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. The authors wish to acknowledge Jeff Reminga who has been instrumental in developing much of the related technology and Bradley Schmerl for his endeavor to include the data-to-network process of this article into SORASCS.

## References

Apte C, Damerau F, Weiss S (1994) Automated learning of decision rules for text categorization. ACM Trans Inf Syst 12(3):233–251

Batagelj V, Mrvar A, Zaversnik M (2002) Network analysis of texts. In: Proceedings B of the 5th international multi-conference information science, pp 143–148

Bertrand A, Cellier JM (1995) Psychological approach to indexing: effects of the operator's expertise upon indexing behaviour. J Inf Sci 21(6):459–472

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Sociol 2:113–120

Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. Soc Netw 19(3):243-269

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163-177

Carley KM (1997) Network text analysis: the network position of concepts. In: Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts. Roberts CW (ed) Lawrence Erlbaum Associates, Mahwah

Carley KM (2002a) Summary of key network measures for characterizing organizational architectures. Unpublished document, CMU 2002, Carnegie Mellon University, SCS/ISR

Carley KM (2002b) Smart agents and organizations of the future. In: Lievrouw LA, Livingstone S (eds) The handbook of new media. Sage, Thousand Oaks

Carley KM, Columbus D (2011) Basic lessons in ORA and AutoMap 2011. Technical report, CMU-ISR-11-109, Carnegie Mellon University, SCS/ISR



- Carley KM, Bigrigg M, Papageorgiou C, Johnson J, Kunkel F, Lanham M, Martin M, Morgan G, Schmerl B, van Holt T (2011a) Rapid ethnographic assessment: data-to-model. In: Proceedings of HSCB focus 2011: integrating social science theory and analytic methods for operational use
- Carley KM, Reminga J, Storrick J, Columbus D (2011b) ORA user's guide 2011. Technical report, CMU-ISR-11-107, Carnegie Mellon University, SCS/ISR
- Chickering D, Heckerman D, Meek C (1997) A Bayesian approach for learning Bayesian networks with local structure. In: Proceedings of thirteenth conference on uncertainty in artificial intelligence
- Davis A, Gardner B, Gardner M (1941) Deep south: a social anthropological study of caste and class. University of Chicago Press, Chicago
- Diesner J, Carley KM (2004) Using network text analysis to detect the organizational structure of covert networks. In: Proceedings of the North American association for computational social and organizational science NAACSOS
- Diesner J, Carley KM (2005) Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In: Narayanan VK, Armstrong DJ (eds) Causal mapping for information systems and technology research. Idea Group Publishing, Harrisburg
- Diesner J, Carley KM (2008) Conditional random fields for entity extraction and ontological text coding. J Comput Math Organ Theory 14:248–262
- Diesner J, Carley KM (2010) Mapping socio-cultural networks of Sudan from open-source, large-scale text data. In: Proceedings of the 29th annual conference of the Sudan studies association
- Dumais S, Platt J, Heckerman D, Sahami M (1998) Inductive learning algorithms and representations for text categorization. In: Proceedings of the 7th international conference on information and knowledge management
- Faust K (1997) Centrality in affiliation networks. Soc Netw 19(2):157-191
- Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35-41
- Freeman LC (1979) Centrality in social networks: conceptual clarification. Soc Netw 1:215-239
- Garlan D, Carley KM, Schmerl B, Bigrigg M, Celiku O (2009) Using service-oriented architectures for socio-cultural analysis. In: Proceedings of the 21st international conference on software engineering and knowledge engineering
- Harris KJ, Jerome NW, Fawcett SB (1997) Rapid assessment procedures: a review and critique. Human Organ 56(3):375–378
- Holsti OR (1969) Content analysis for the social sciences and humanities. Addison-Wesley, Reading
- Hotho A, Jäschke R, Schmitz C, Stumme G (2006) BibSonomy: a social bookmark and publication sharing system. In: Proceedings of the conceptual structures tool interoperability workshop at the 14th international conference on conceptual structures, pp 87–102
- Howe J (2008) Crowdsourcing: why the power of the crowd is driving the future of business. Three Rivers, New York
- Mayfield J, McNamee P, Costello C, Piatko C, Banerjee A (2002) JHU/APL at TREC 2001: experiments in filtering and in Arabic, video, and web retrieval. In: Proceedings of the tenth text retrieval conference, NIST special publication
- Tsvetovat M, Reminga J, Carley KM (2004) DyNetML: interchange format for rich social network data. Technical report, CMU-ISRI-04-105, Carnegie Mellon University, SCS/ISR
- Wasserman S, Faust K (1995) Social network analysis, methods and applications. Cambridge University Press, Cambridge
- Wei W, Pfeffer J, Reminga J, Carley KM (2011) Handling weighted, asymmetric, self-looped, and disconnected networks in ORA. Technical report, CMU-ISR-11-113, Carnegie Mellon University, SCS/ISR

**Jürgen Pfeffer** earned a B.S. degree in Computer Science, an M.S. degree in Computer Science Management, and a Ph.D. degree in Business Informatics from Vienna University of Technology (Austria). He was working in industry and non-university research institutes for ten years and he is currently a Post Doctoral Associate at the Institute for Software Research at Carnegie Mellon University. His research focuses on the dynamic analysis of large scale real world networks.

Kathleen M. Carley is a professor in the School of Computer Science in the department—Institute for Software Research—at Carnegie Mellon University. Carley studied for an S.B. in Economics and an S.B. in Political Science from the Massachusetts Institute of Technology in 1978. She received her Ph.D. in Sociology from Harvard University in 1984. She is the director of the Center for Computational Analysis of Social and Organizational Systems (CASOS), a university wide interdisciplinary center that brings together network analysis, computer science and organization science (http://www.casos.cs.cmu.edu/) and has an associated NSF funded training program for Ph.D. students.

