Turning Time Outside-In:
Networks Among Time Slices as Nodes with Semantic Similarity as Links

James A. Danowski, Ph.D.
Dept. of Communication
University of Illinois at Chicago[1]
Chicago, USA
jdanowski@gmail.com

***Abstract*—This research demonstrated a new approach to time-series analysis of semantic network data. Three years of the Obama administrations coverage in the *New York Time*s and *Washington Post* was extracted from Lexis-Nexis. The text was sliced into 74 two-week time intervals corresponding to frequency of Gallup polls. Words appearing in three word positions on either side of each word were tabulated, one for the aggregate text file and 74 times for each of the time slice data. A matrix of word pairs by time slices created the basis for computing a time slice similarity score for each pair of slices. These time units then became nodes in a network analysis with link strengths defined by semantic similarity. Network analysis procedures were applied to explore the meaning of the structures identified.**

Keywords-Semantic Network Analysis; Time-Series; Obama administration; time node centrality

I. INTRODUCTION

Goals

A goal of this paper is to investigate a non-sequential conception of time in which time slices are nodes and links are the semantic similarities among these nodes. This bases time similarity not on chronological contiguity of time slices, instead on their content similarities to other time slices. This is the basis for a network of time slices. This paper explores what such an approach reveals about events and frames for political domains.

Time Concepts in Social Research

Linear, chronological time conceptions and time-series analysis are deeply rooted in the modernist social scientific world view. Classic time-series analysis gathers data from 120 points in time or more and removes serial autocorrelation within variables as they change over time, then examine the associations among these. Sometimes to the goal is to forecast future levels of the variables, while other times a variety of hypotheses are tested. Yet, anthropological studies reveal large variations in cultural time conceptions other than the modernist perspective. Taking a different view of time in social network studies can reveal information not as readily available in traditional time-series analysis.

Social network analysis typically analyzes data without regard for time. An aggregate network is analyzed. There are increasing exceptions. Valente [1] reviews studies that examine change over time in networks. Snjders [2] has developed the Siena software for examining changes in relatively small social networks. Carley's ORA [3] handles large social network data over time.

In the particular area of semantic network analyses based on mining text data typically aggregate data into one set, ignoring time. The mining of text data, however, is often done from sources that are time-stamped and enable time-sensitive analysis. For example, Danowski [4] mined text about jihad from Muslim majority countries before, during, and after the uprising in Tunisia and Egypt and found support for the hypothesis that countries that became more central in the network increased the number of web pages devoted to jihad.

For the approach this paper uses, the key idea is rather than treating time as an exogenous variable external to the system, time is made an endogenous variable internal to the system of analysis. This can be seen as turning time outside-in. Network analysis is central to this method, but in a new way. Traditionally, nodes in relevant network would be events or event frames and these nodes would be linked according to their cooccurrences over time. Problems remain in how to use text mining to effectively identify these events and frame in a systematic manner. Instead of events and frames being nodes in the network, however, one makes time intervals the nodes. This links among nodes vary in strength based on how similar the totality of text content is for pairs of time nodes. Network analysis of time intervals then reveals particular network structures that further inform the analysis. A key variable is the centrality of time nodes in the network. Because these connect a large number of relatively diversely linked time nodes, by opening up the central time nodes and examining the word pairs that are most frequent, this reveals the key propaganda frames that cut across time. Examining the different most central time nodes, one has a systematic method for seeing as well the key sub-themes of the over-time discourse.

II. Overview of Methods

[1]Retired.

I. For a particular discourse domain, for example, jihadist discourse about terrorist activities, one conducts a search of a text database that contains translations of jihadist web pages, television news, radio broadcasts, and newspapers. BBC International Monitoring service provides such transcripts and they are available in the Lexis-Nexis database.

II. Full-text documents over approximately 100 time intervals are extracted. These time intervals could be daily if the domain is highly active, or more likely is weekly, which would cover an approximate two-year time frame.

III. The first step is to segment the file containing all of the documents into time intervals, such as weeks. WORDij's [5] TimeSlice program does such time segmentation and inserts codes into the large text file representing the time segments.

IV. Next a semantic network analysis procedure is run to produce a file for each time segment that contains each word pair occurring within a window that is three words wide on either side of each word in the text. There is a unique word pair file for each time segment. It contains for each word pair found the frequency within the time segment.

V. To obtain a benchmark for setting up the next phase of the analysis, the same word-pair extraction is run on the large text file but ignoring the time segments. This produces a master list of word pairs for the entire corpus. For example, for a file containing two years of documents, one could expect 750,000 unique word pairs.

VI. The next step is to create a matrix of the master word pairs (approximately 750,000) as row labels and the time intervals (approximately 100). A program rearranges each of the time segment columns so that its individual word pairs appear in the same order as in the master file. If a word pair does not occur in the particular time segment is cell entry is zero, otherwise its entry is the frequency of that word pair.

VII. With such a rectangular matrix of time interval word pairs, one then computes the similarities of each pair of time intervals across the master word pairs. In standard network analysis this would be equivalent to converting a two-mode network of word pairs by time intervals into a one-mode time-interval network. The size of the matrix may be too large for some network analysis programs, so as an alternative the analysis was run in SPSS v. 20. The similarity scores for each pair of time intervals was computed with the analysis procedure "proximities" located under the "correlation" tab. Pearson correlation similarity coefficients were computed and the resulting matrix of time by time similarities exported for analysis in a network analysis program. Because these time pairs were not directed, actually only the lower or upper triangle of the matrix of time by time is in play. Each cell of the matrix contains the single similarity coefficient for how highly correlated the frequencies of word pairs of each time slice were across the 750,000 master word pairs.

VIII. Network analysis is then conducted where the nodes are time intervals and their link strengths are their similarity scores. The programs UCINET [6] and NetDraw [7] were used.

IX. Centrality scores for the time nodes are examined and the nodes with the highest centralities are further analyzed by opening up their respective word-pair files and examining the word pairs sorted by frequency. The highest frequency word pairs indicate the key content of the time mode and therefore reveal that these particular word pairs were most instrumental in creating high similarities with other time nodes. Central nodes typically have high degree centrality, which is the number of links to other nodes, and high Freeman betweenness centrality [8], which indexes on how many shortest paths between all other nodes the particular node lays.

X. Now the analyst studies the most central time nodes' key word pairs and the structure of the overall time node network to interpret the meaning in light of the original goals.

IV. Empirical Example: The Obama Administration

This paper uses as an example the first three years of the Obama administration. Using Lexis-Nexis we captured all news in the *New York Times* and *Washington Post* mentioning Obama and/or cabinet member over the first three years of the administration.

WORDij's TimeSlice program segments text into two week intervals (chosen based on the Gallup poll interval), resulting in 74 time slices. WORDij's WordLink program extracts from the aggregate file all word pairs occurring within three word positions on either side of each word. A small stop list is used. Here we dropped frequencies of word pairs less than 10.This produces a master list of word pairs for the creation of a matrix of word pair occurrences time slice by slice for the 74 intervals. WordLink extracts all word pairs, as in step two, for each time slice separately. For the individual time slice network analysis we dropped frequencies less than 3.

The master word pair file and measures of frequencies for these pairs in each of the 74 time slices was used to create a rectangular matrix of word pairs by time slices. We dropped word pairs from each time slice not appearing in the master file. For master list pairs for which a time slice had no occurrences the matrix cell was 0. There were 778,444 master pairs. We input this 74 by 778,444 matrix input to SPSS v. 20 and computed proximity scores for each pair of time slices based on similarities of distributions of word pairs. The Pearson correlation coefficient is the appropriate measure for proximity for scalar data such as frequencies. We exported the 74 by 74 matrix of correlations as an Excel file which we then imported into three programs for comparative analysis UCINET, NetDraw, and VOSviewer [9].

The big difference in our analysis is now the nodes in the network are time slices and the links among time units are the similarity of occurrences across the master word pairs. In this sense time has been turned outside-in. It is no longer an

exogenous variable but endogenous to the system. Time has been transformed into network nodes. In general the approach is one of converting a two-mode network, for example, the number of people at this conference as rows with columns being the sessions available. Analysts create a one-mode conversion, either creating a network with nodes being the people at the conference or being the sessions. The sessions are analogous to our time slices.

VOSviewer uses a metric-based clustering algorithm on network data. It produced the graph in Figure 1. It found two large clusters of time nodes.

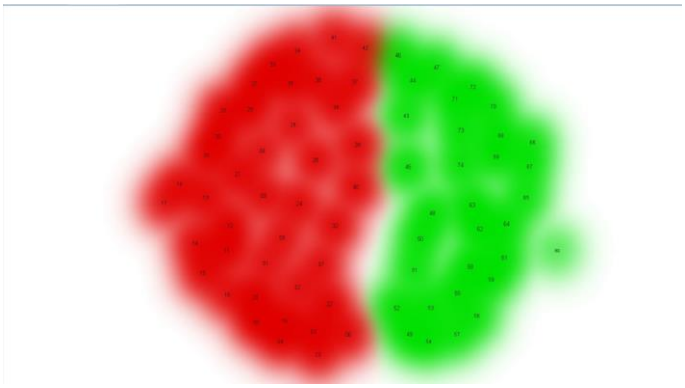

Figure 1. Vos Two Cluster Solution

In UCINET we used Newman's [ ] hierarchical community detection alrgorithm. It two found two main components in the network, as shown in Figure 2.

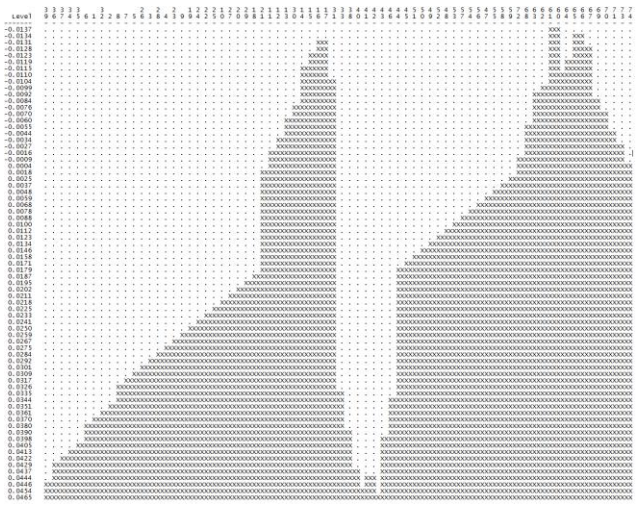

Figure 2. Newman Community Detection

To see the meaning of the most central time nodes for each component we opened the word pair frequency files and observed the most frequent word pairs. Table 1 shows these to be associated with Obamacare.

Table 1. Top Word Pairs for Time Node 16

WORD 1	WORD 2	FREQUENCY
health	care	881
public	option	314
health	reform	194
health	insurance	177
healthcare	reform	165
public	plan	143

The second components most central time node was 65. Table 2 shows this node to be associated with the debt crisis deliberations.

Table 2. Frequent Word Pairs for Time Node 65

WORD 1	WORD 2	FREQUENCY
debt	ceiling	459
debt	limit	320
spending	cuts	267
social	security	262
tax	increases	192
raise	debt	179
health	care	164
tax	cuts	136
increase	debt	118
raise	ceiling	109
raising	debt	106
federal	debt	95
raise	limit	93
entitlement	programs	91
increase	limit	82
deficit	reduction	82
big	deal	80
grand	bargain	79
raising	ceiling	78
nations	debt	74
tax	code	73
medicare	medicaid	69
tax	breaks	68
tax	increase	67
budget	deal	65
cuts	medicare	65
trillion	cuts	64
social	medicare	63

Figure 3 shows the graph theoretic spring embedding representation. Compared to look at the Vos and Newman figures, this is analogous to looking at the inner structure of an organism. Again there are two main components. Several key nodes like them together. The bottom component is from the first half of the Obama administration. Notice the higher density among time nodes, indicating more similarity than for the second component. The key to understanding what appears to have caused this bifurcated network structure of time is to look at the nodes at the top edge of the bottom structure. Time node 41 was the

Mid-Term election in which the Republications became the majority party in the house. The general tenor of the electorate building up to this election was a high level of dissatisfaction with Obama. The Tea Party exerted considerable influence over public opinion.

The size of the nodes reflects their Freeman Betweenness centrality. Node the largest most central node, time interval 48. 2010 Midterm Elections are pivotal. They bifurcate the network into two clusters (time interval, 41).

This shift requires 6 weeks. Time interval 48 is the most central node, configurational and foreshadowing, and also a bridge between the two clusters. Nearly all major initiatives to follow in the Obama administration are discussed in node 48. This is the launch of the Re-Election campaign in April 2011. By opening the most central time node to look at its word pairs one sees the content that results in this node being so centrally connected to other nodes. It is like opening a concentrated box of content. All of the issues seen in time node 48 were quite remarkably discussed in this one two-week time interval.

Table 3. Frequent Word Pairs for Time Node 48

WORD 1	WORD 2	FREQ
president	barack_obama	564
barack_obama	administration	314
tax	cuts	228
prime	minister	217
south	korea	212
national	security	204
healthcare		169
tea	party	167
social	security	148
lameduck	session	141
nuclear	weapons	131
start	treaty	109
justice	department	109
midterm	elections	107
foreign	policy	107
west	bank	105
middleeast		90
wall	street	83
arms	treaty	80
north	korea	75
secretservice		75
unitednations		70
east	jerusalem	65
missile	defense	63
security	forces	62
bush	tax	61
nuclear	program	61
tax	breaks	61
tax	rates	61
summit	meeting	60
repeal	don't	59
treaty	russia	59
tax	increases	58
human	rights	57
economic	growth	56
interest	rates	56
highspeed	rail	56
global	economy	54
federal	reserve	54
deficit	reduction	54
peace	talks	50
palestinian	state	50
arms	control	49
world	war	48
afghan	forces	47
spending	cuts	46
test	scores	46
foreign	relations	45
seoul	south	44
security	council	43
defense	bill	43
seoul	korea	43
monetary	policy	39
supreme	court	39
nato	summit	39
climate	change	38
second	term	37
guantanamo	bay	36
deficit	commission	34
midterm	election	33

Observations on Events and Frames

While events can be laid upon a chronological time series, events and event frames may often be of a different character than merely time-dependent. Some events/frames operate in spurts, perhaps with considerable time between them that may result in the analyst missing the wholeness of the event/frames. Some events/frames retrench and add new elements as they "move back in time, then forward again." Some events/frames contain arcs in which an alternative framing is offered by opposing groups. Some events/frames are pivotal, changing the fundamental dynamics of the overall event/frame space. Some events mark epics that may begin after a pivotal time.

Depending on socio-political perspectives some events/frames may be treated from a common semantic domain, while the same such events/frames may from a different socio-political be treated quite differently. Methods that foster parsimony in treating events/frames may be preferred in some circumstances.

The method used here, analyzing time as nodes in a network based on their similarity of word pair frequencies appears to offer the most flexibility for observing these various kinds of events and frames. Turning time outside in, that is treating time slices as nodes and their semantic similarities as links, yields some unique value. The event/time space is analyzed for its underlying network structure. Community detection reveals a simplified structure. Particular time nodes serve specialized roles in the network. Exploration with other networks can elaborate a

conceptual framework for event/time spaces for this new kind of semantic network analysis.

V. COMPETING INTERESTS

The author declares that he has no competing interest.

VI. ACKNOWLEDGMENT

The author is grateful Noah Cepela for text collection and to Richard Weeks for programming assistance.

VII. REFERENCES

[1] Valente, T.W. (2005). Network models and methods for studying the diffusion of innovations. In Peter J. Carrington, John Scott, Stanley Wassermann (eds). Models and methods in social network analysis. Cambridge University Press.

[2] Snjders, TAB (2005). Models for longitudinal network data. In Peter J. Carrington, John Scott, Stanley Wassermann (eds). Models and methods in social network analysis. Cambridge University Press.

[3] Carley, K.M. (2006). Destabilization of covert networks. Computational & Mathematical Organization Theory, 12(1), 51-66, DOI: 10.1007/s10588-006-7083-y

[4] Danowski, J.A. (2011). Changes in Muslim Nations' Centrality Mined from Open-Source World Jihad News: A Comparison of Networks in Late 2010, Early 2011, and Post-Bin Laden. IEEE Proceedings of the International Symposium on Open-Source Intelligence and Web Mining, Athens, Greece, September 12-14, 2011.

[5] Danowski, J.A. (2010). WORDij Software for Semantic Network Analysis [computer program]. Chicago: University of Illinois at Chicago.

[6] S. P. Borgatti, M. E. Everett, and L. C. Freeman, "Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies, 2002.

[7] S. P. Borgatti, "NetDraw: Graph visualization software. Harvard, MA: Analytic Technologies, 2002.

[8] L. C. Freeman, "Centrality in social networks conceptual clarification," Social Networks, vol 1., no. 3, pp. 215-239, 1978-1979.

[9] N.J.van Eck and L. Waltman1 (2010). Software survey: VOSviewer, a computer program for bibliometric mapping

[Scientometrics. 2010 August; 84(2): 523–538.