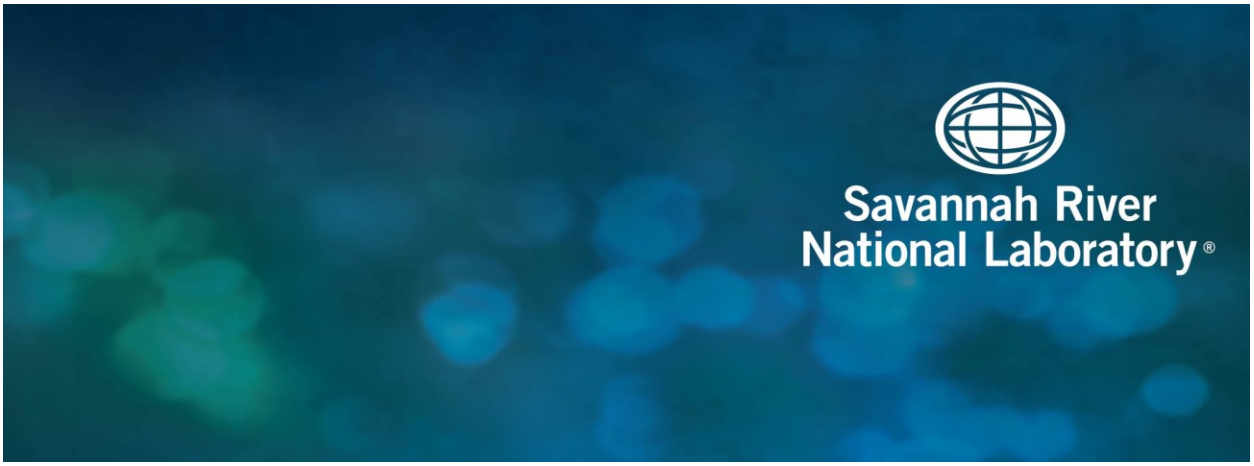
Contract No:

This document was prepared in conjunction with work accomplished under Contract No. 89303321CEM000080 with the U.S. Department of Energy (DOE) Office of Environmental Management (EM).

Disclaimer:



A U.S. DEPARTMENT OF ENERGY NATIONAL LAB • SAVANNAH RIVER SITE • AIKEN, SC • USA

Machine Learning Modeling Pipeline for Extracting Nuclear Proliferation Events of Interest from Open Data Sources (U)

T. L. Danielson

B. Mayer

N. Muralidhar

J. Miller

H. Dogan

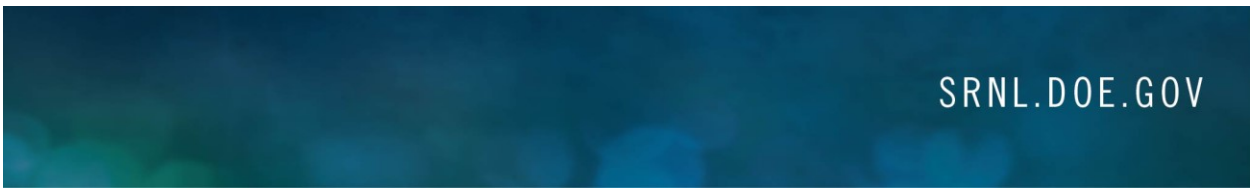
N. Self

P. Butler

F. Liu

January 2022

SRNL-STI-2022-00036, Revision 0



DISCLAIMER

Printed in the United States of America

Prepared for
U.S. Department of Energy

SRNL-STI-2022-00036
Revision 0

Keywords: *Aritificial Intelligence*
Machine Learning
Natural Language Processing
Proliferation

Retention: *Permanent*

Machine Learning Modeling Pipeline for Extracting Nuclear Proliferation Events of Interest from Open Data Sources (U)

T. L. Danielson
B. Mayer
N. Muralidhar
J. Miller
H. Dogan
N. Self
P. Butler
F. Liu

January 2022

Savannah River National Laboratory is operated by Battelle Savannah River Alliance for the U.S. Department of Energy under Contract No. 89303321CEM000080.

REVIEWS AND APPROVALS

AUTHORS:

T. L. Danielson, SRNL, Environmental Sciences & Dosimetry Date

B. Mayer, SCAIDA, Virginia Polytechnic Institute and State University Date

N. Muralidhar, SCAIDA, Virginia Polytechnic Institute and State University Date

J. Miller, SCAIDA, Virginia Polytechnic Institute and State University Date

H. Dogan, SCAIDA, Virginia Polytechnic Institute and State University Date

N. Self, SCAIDA, Virginia Polytechnic Institute and State University Date

P. Butler, SCAIDA, Virginia Polytechnic Institute and State University Date

F. Liu, SCAIDA, Virginia Polytechnic Institute and State University Date

APPROVAL:

M. Cofer, Manager Date
Environmental Sciences & Dosimetry

ACKNOWLEDGEMENTS

The team would like to acknowledge Jeff Pike who was a PI on the project until his retirement from SRNL in May of 2021. Jeff was an asset to the project team given his decades of subject matter expertise.

EXECUTIVE SUMMARY

In FY2020, the Savannah River National Laboratory (SRNL) and the Sanghani Center for Artificial Intelligence and Data Analytics at Virginia Polytechnic Institute and State University entered a collaboration funded by Department of Energy's (DOE) Office of Defense Nuclear Nonproliferation Research and Development. The project's mission was to take the first steps toward developing a demonstration prototype system that uses multiple machine learning and data analytics methods on large-scale open data sources to identify new, developing, and/or undeclared nuclear programs. Given the SRNL team's on-site perspective of events culminating in the DOE's decision to pursue the Savannah River Plutonium Processing Facility (SRPPF), the team targeted the identification of events and indicators in retrospective datasets that pointed to the activity of "fissile core fabrication at the Savannah River Site" prior to the official announcement in May of 2018.

A preliminary modeling pipeline was developed in FY20 that showed the datasets contained adequate signal for continuation of efforts. In FY21, a modular demonstration prototype modeling pipeline has continued in development for two text-based data sources: a broad internet archive (Webhose Ltd.) and a decahose Twitter database (i.e., a global sampling of one in every ten Tweets). The techniques that have been developed rely on graph theory and anomaly detection to identify contextual shifts in key words and phrases at various points in time such that indicators of events of interest could be identified and subsequently, events could be extracted from the corpuses. The foundational concept behind the approaches is that contextual shifts in key words and phrases can act as indicators of events of interest.

Both datasets have proven successful in extracting events of interest related to pit production at the Savannah River Site prior to the official announcement. In addition, the pipelines have generated a wide range of events broadly summarized as: the awarding of DOE contracts at major sites, DOE investments in various programs, accidents at DOE national laboratories, speculations about the fate of pit production in the DOE complex, domestic and international shipments and receipts of nuclear materials at DOE sites, termination of non-proliferation agreements with Russia, termination of MOX, new weapons development approvals/testing, nuclear posture reviews, major DOE cleanup/production milestones, political opinions, and nuclear watch groups' opinions, among many others

TABLE OF CONTENTS

LIST OF TABLES viii
LIST OF FIGURES viii
LIST OF ABBREVIATIONS x
1.0 Introduction 1
1.1 Background 1
1.1.1 System Overview 1
1.1.2 Conceptual Model Hierarchy 2
1.1.3 Glossary of Key Terms 3
1.1.4 FY20 Preliminary Modeling Pipeline 4
1.1.5 FY21 Demonstration Prototype Development 7
2.0 Twitter Dataset Modeling Pipeline 7
2.1 Graph-Based Approach for Identifying and Analyzing Contextual Shifts 8
2.1.1 Cosine Similarity-Based Network Graphs 8
2.1.2 Ensemble Approach for Identifying Time Periods with Contextual Shifts 17
2.1.3 Extracting Events of Interest 20
2.2 Performance Assessment of Twitter Modeling Pipeline – Sensitivity to a Reduced Dataset 27
3.0 News Article Aggregator Data Source – Webhose Ltd. 30
3.1 Entity Characterization for Anomaly Detection and Event Extraction 30
3.1.1 Anomaly Detection Using Similarity Metrics Over Time 30
3.1.2 Anomaly Detection by Comparing Entity Co-Occurrence Graph Edges Over Time 32
3.1.3 Event Extraction 33
3.2 Performance Assessment of Webhose.io Modeling Pipeline 36
4.0 Discussion 44
5.0 Conclusions 48
6.0 References 49

LIST OF TABLES

Table 2-1. Events of interest extracted for the key terms “pit production”, “plutonium disposition”, and “plutonium pit”. Inflection window denotes the inflection window at which the event was detected. 21

Table 2-2. Number of Tweets in each time interval searched. 26

Table 2-3. Number of Tweets found at different levels of similarity. 27

Table 2-4. Number of Tweets found at different levels of similarity when Tweets containing the term “national_security” are removed. 27

Table 4-1. Parameters in the modeling pipeline. 46

LIST OF FIGURES

Figure 1-1. Illustration of the long-term vision for the event detection and forecasting system. 2

Figure 1-2. Illustration of hierarchical conceptual model of nuclear activity domains. 3

Figure 1-3. Approach for curating a glossary of key terms. 4

Figure 1-4. Principal component analysis plot of the embedding model trained on the Webhose corpus and a query to the Twitter word embedding model for the top 10 most similar terms, based on the cosine similarity. 5

Figure 1-5. Similarity rank over time between the key terms “pit_production” and “savannah_river_site” in the preliminary modeling pipeline 5

Figure 1-6: Heterogeneous entity ego-network (*nrc*). 6

Figure 1-7: Heterogeneous entity ego-network (*sandia*). 7

Figure 2-1. Illustration of the network graph-based approach for a graph with three ranks and four nearest neighbors 8

Figure 2-2. Illustration of the graph-based approach for computing the weighted average embedding vector in a graph with two ranks and four nearest neighbors. 9

Figure 2-3. Graph-based similarity metrics for the graph of “pit_production”, rank 1 and 10 nearest neighbors 11

Figure 2-4. Graph (influence rank, total influence) for time intervals 0 through 9. 12

Figure 2-5. Graph (influence rank, total influence) for time intervals 10 through 18. 13

Figure 2-6. Graph (influence rank, total influence) for time intervals 19 through 27. 14

Figure 2-7. Graph (influence rank, total influence) for time intervals 28 through 36. 15

Figure 2-8. Graph (influence rank, total influence) for time intervals 37 through 44. 16

Figure 2-9. Ensemble approach for identifying important time windows of contextual change for the term "nuclear_security". Blue, yellow, green, and black lines represent the similarity metric, smoothed profile (where the Gaussian filter was applied), the second derivative of the smoothed profile, and the inflection point of the second derivative, respectively. 19

Figure 2-10. Matrix showing the time periods where inflection points were detected. Annotated numbers and colors indicate the number of key terms with an inflection point in that time interval. 20

Figure 2-11. Matrix showing the time periods where inflection points were detected for the reduced corpus. Annotated numbers indicate the number of key terms with an inflection point in that time interval. . 29

Figure 3-1: Similarity metrics for "savannah river" computed from analysis of documents returned by seed phrases classified as "political/diplomatic events". 32

Figure 3-2: Anomalous or outlier edges for "los alamos national lab" computed from analysis of documents returned by seed phrases classified as "political/diplomatic events". 33

Figure 3-3: Article encoding with entities shown in red. 33

Figure 3-4: Article encoding with anomalous edge list (in red) 33

Figure 3-5: Sample of event extraction output for the graph of "savannah river" at May of 2018. 35

Figure 3-6: Similarity metrics for "savannah river" computed from analysis of documents returned by seed phrases classified as general seed phrases. 36

Figure 3-7: Similarity measures for "sandia national lab" computed from analysis of documents returned by seed phrases classified as general seed phrases. 37

Figure 3-8: Similarity measures for "sandia national lab" using general seed phrases with increased articles and uncapped ego-networks. 38

Figure 3-9: Cumulative co-occurrences over the entire time period for "savannah river" using general seed phrases (left) and political seed phrases (right). 39

Figure 3-10: Anomalous edges for "savannah river" using general seed phrases with increased articles and uncapped ego-networks. 40

Figure 3-11 Extracted events and text for anomalous edges from "savannah river" during May 2018 using general seed phrases with increased articles and uncapped ego-networks. 41

Figure 3-12: Extracted article titles and event text for anomalous edges from "savannah river" during May 2018 using general seed phrases with increased articles and uncapped ego-networks. 42

Figure 3-13: Extracted article titles and event text for anomalous edges from "savannah river" from June 2016 to April 2018 using general seed phrases with increased articles and uncapped ego-networks. 43

LIST OF ABBREVIATIONS

DOE	Department of Energy
MOX	Mixed Oxide Fuel Fabrication Facility
SCAIDA	Sanghani Center for Artificial Intelligence and Data Analytics
SRNL	Savannah River National Laboratory
SRPPF	Savannah River Plutonium Processing Facility
SRS	Savannah River Site

1.0 Introduction

In FY2020, the Savannah River National Laboratory (SRNL) and the Sanghani Center for Artificial Intelligence and Data Analytics at Virginia Polytechnic Institute and State University entered a collaboration funded by Department of Energy's (DOE) Office of Defense Nuclear Nonproliferation Research and Development. The project's mission was to take the first steps toward developing a demonstration prototype system that uses multiple machine learning and data analytics methods on large-scale open data sources to identify new, developing, and/or undeclared nuclear programs. Given the SRNL team's on-site perspective of events culminating in the DOE's decision to pursue the Savannah River Plutonium Processing Facility (SRPPF), the team targeted the identification of events and indicators in retrospective datasets that pointed to the activity of "fissile core fabrication at the Savannah River Site" prior to the official announcement in May of 2018. After two years of development, this report will document and characterize the current capability of the demonstration prototype and outline the remaining challenges and future development needs.

1.1 Background

FY20 efforts were focused primarily on developing a preliminary prototype modeling pipeline by identifying machine learning methods that could successfully extract signal from two open-source datasets. In the preliminary modeling stages, this signal was demonstrated by identifying a connectedness between key words and phrases, which are primarily entities and/or select activities that are common across the DOE Complex. Two reports have documented the conceptual and algorithmic approaches that were developed during the preliminary prototype modeling stages [1,2]. The following subsections will provide a high-level summary of those efforts, which led to the continued development of the demonstration prototype in FY21.

1.1.1 System Overview

The prototype system has sought to adapt a modular architecture of the EMBERS forecasting system [3] with models developed specifically for detection of indicators of events related to weapons development facilities and programs. Here, two text-based data sources have been used: a broad internet archive (from Webhose Ltd.) and a decahose Twitter database (i.e., a global sampling of one in every 10 tweets). Natural language processing is the foundation from which several techniques such as word embedding models, entity extraction, event extraction, and geocoding have been selected to identify and extract events and indicators from the individual data sources. The primary working hypothesis is that *contextual shifts in key words and phrases over time can act as indicators of events of interest*. The long-term vision of the modeling pipeline is illustrated in Figure 1-1, where each data source has its own set of models geared toward the extraction of events that can be fused together to make an inferential forecast or provide relevant information to an analyst who is operating in a massive data environment. This modular architecture is amenable to the addition of new data types and/or sources and can be applied to a broad range of topical domains, assuming sufficient domain-specific data can be compiled.

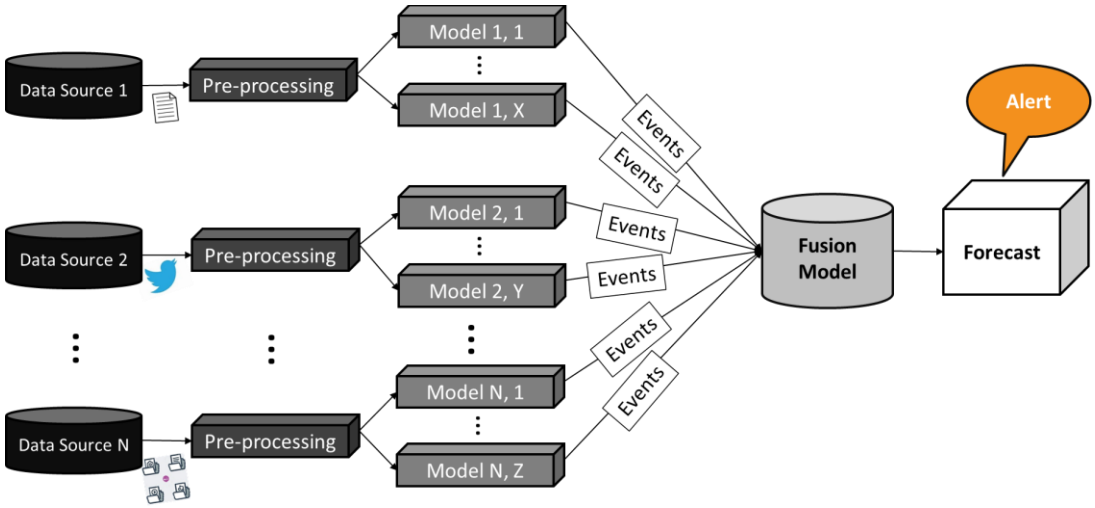


Figure 1-1. Illustration of the long-term vision for the event detection and forecasting system.

1.1.2 Conceptual Model Hierarchy

The prototype modeling pipeline relies on the pre-filtration of datasets through carefully formulated queries to reduce noise and feed primarily domain relevant information to the models. However, nuclear proliferation activities of interest are generally infrequently occurring and/or rare events and may have substantial overlap with other nuclear activities. To overcome these challenges, hierarchical conceptual models representing the activities and events of interest were created to facilitate the formation of queries that include relevant terminology that may be present in the overlapping domains. Additionally, such a conceptual model allows a baseline to be established such that models can be developed to capture change or "out of the ordinary" events.

The conceptual model hierarchy outlining the high-level "Nuclear Activities" domains of interest is illustrated in Figure 1-2. In this hierarchical structure, the specificity of the activity increases toward the bottom of the hierarchy and domains that fall along the same vertical alignment are presumed to have potential overlap in the open source. For example, "Weapons Component Fabrication" and "Nuclear Power" share several of the same steps within the nuclear fuel cycle and may become discernable only after some key activity is executed. In addition to the activity domains of interest, event domains were defined to establish the nature of the events of interest that may be discovered in the open source when a nuclear activity is discovered. The domains are defined using four primary categories:

- Acquisition Events
- Political/Diplomatic Events
- Economic Events
- Population/Personnel Events

By establishing this structure for the events of interest, the key words and phrases that are used to query the data sources can be pre-classified as belonging to the event type with which they are most heavily associated. Furthermore, documents containing those pre-classified terms would also be presumed to contain information related to that event type such that the context surrounding the key term could be extracted.

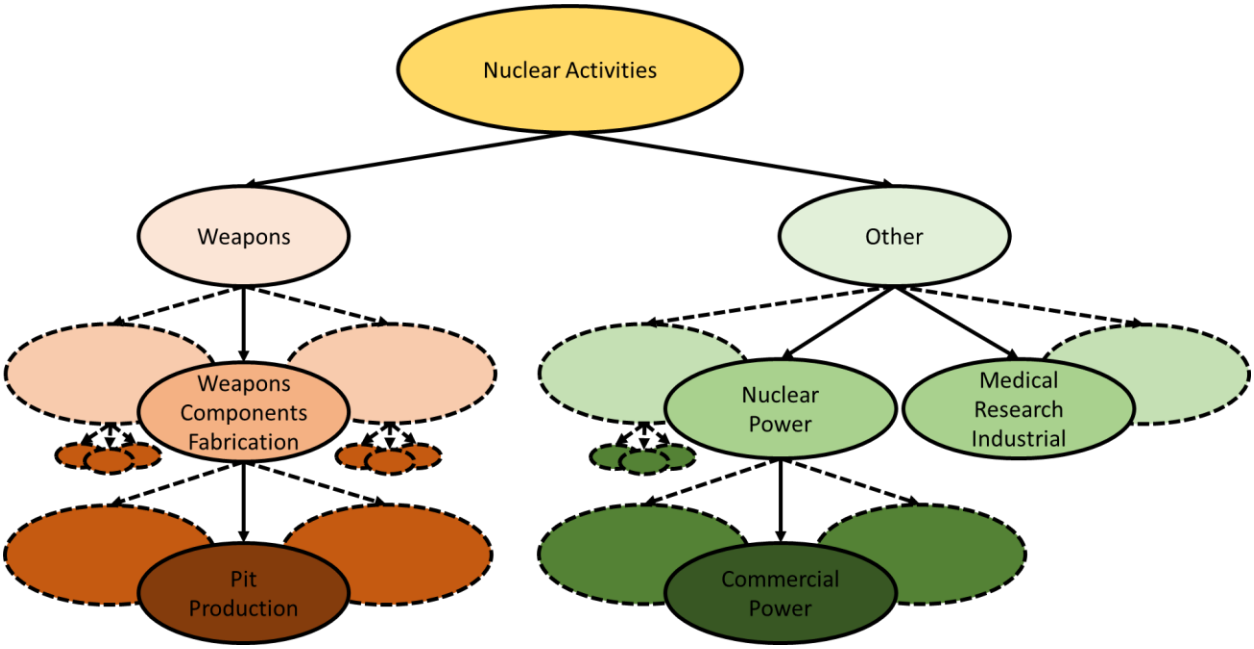


Figure 1-2. Illustration of hierarchical conceptual model of nuclear activity domains.

1.1.3 Glossary of Key Terms

Having a conceptual model, a semi-automated approach was developed (illustrated in Figure 1-3) to create a glossary of approximately 450 key terms that would be used to query the data sources. Each key term was pre-classified according to the event domain with which it was most closely associated. Having labeled key terms provides flexibility to formulate separate models for the different event types based on the documents returned by specific key terms, though is not a requirement within the pipeline.

Both datasets were queried using the list of key terms assembled using Boolean logic. While the Webhose data source was queried using the full list of key terms, the Twitter data source was queried using a smaller subset of only 90 key terms that were presumed to be most relevant to fissile core fabrication. From these queries, approximately 12 million webpages and 3 million Tweets were returned spanning a time period of August of 2014 to May of 2018.

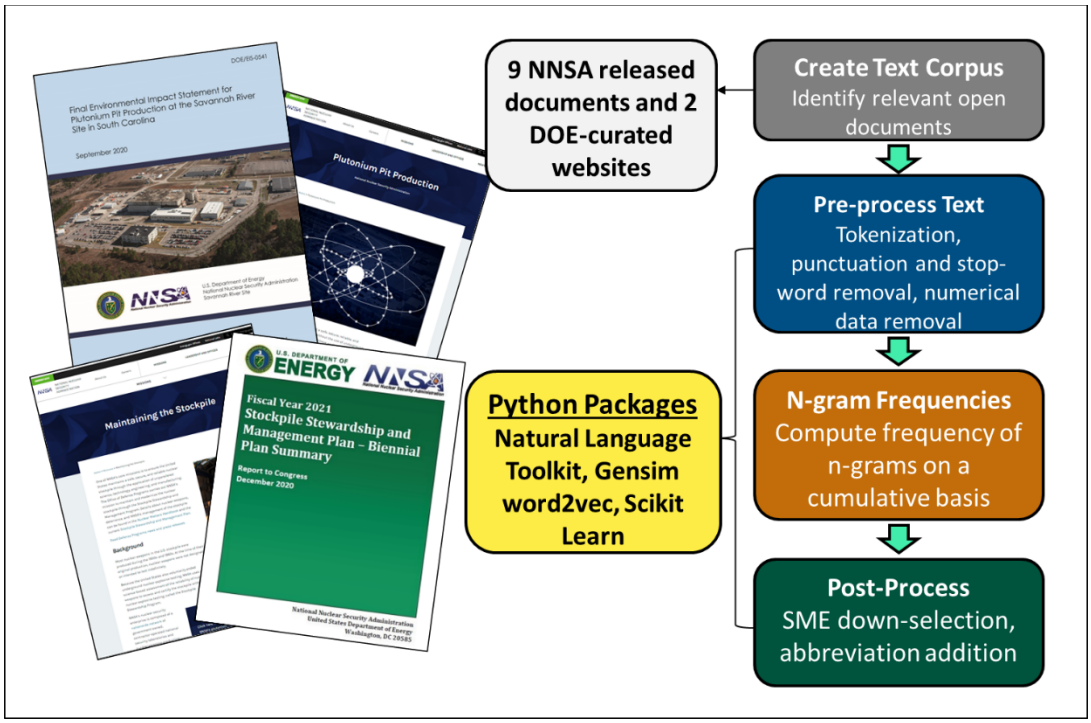


Figure 1-3. Approach for curating a glossary of key terms.

1.1.4 FY20 Preliminary Modeling Pipeline

In FY20, the team developed a preliminary modeling pipeline based on word embedding models as implemented in the Word2Vec formalism. Initially, individual embedding models were trained on the entire corpuses and exploratory data analysis techniques, such as principal component analysis, were applied to ensure that domain specific datasets had been retrieved. Notably, both vectorized corpuses demonstrated that domain-specific datasets had been obtained and the proximity of terms in the vocabularies were representative of the real world. This structure is illustrated in the principal component analysis plot and cosine similarity results in Figure 1-4.

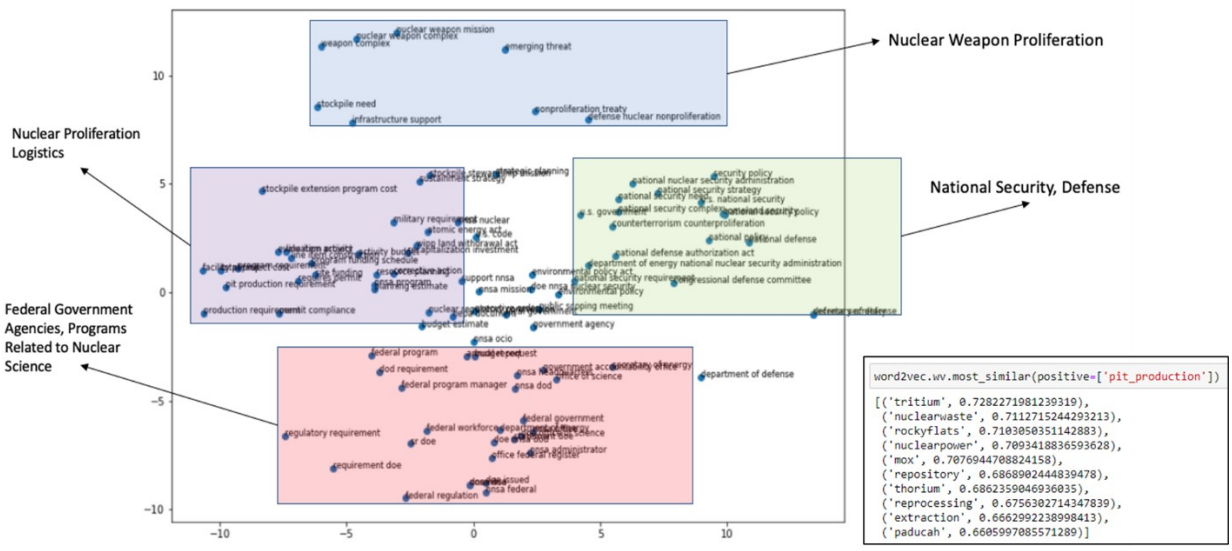


Figure 1-4. Principal component analysis plot of the embedding model trained on the Webhose corpus and a query to the Twitter word embedding model for the top 10 most similar terms, based on the cosine similarity.

Having retrieved domain-specific datasets, preliminary time dependent word embedding models were developed to demonstrate that a measurable level of "signal" could be achieved by analyzing contextual shifts in key words and phrases over time. In the preliminary modeling stages, successful identification of "signal" involved creating models that showed a valid connectedness of keywords from the glossary of key terms, as discovered in the open source. A key result in the preliminary analysis performed on the Twitter dataset was the identification of a connection between the Savannah River Site and the key term "pit production" prior to the official announcement, as shown in Figure 1-5.

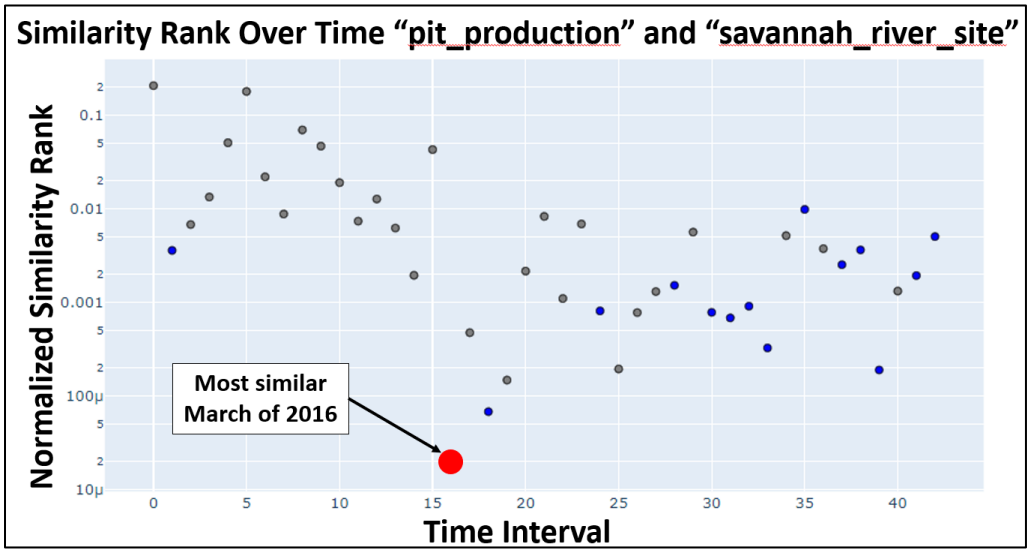


Figure 1-5. Similarity rank over time between the key terms "pit_production" and "savannah_river_site" in the preliminary modeling pipeline.

The analysis of heterogenous weighted temporal entity co-occurrence graphs built from articles returned by the news aggregator data source demonstrated that information regarding specific organizations,

geopolitical entities, and specific products could be extracted. This is illustrated by two sub-graphs shown in Figure 1-6 and Figure 1-7 for the key terms "*nrc*" and "*sandia*", constructed using entities extracted from the articles published between June and August of 2017. Notably, both graphs are comprised of intuitively acceptable connections. In addition, Figure 1-7 revealed a notable connection between Sandia National Laboratory and the terms "cyanobacteria" and "heliobiosys inc.". Upon further exploration, this was found to be representative of a research effort centered around cyanobacteria for biofuels production and a collaboration between Heliobiosys, Inc. (a company extensively dealing with cyanobacteria), Lawrence Berkeley National Laboratory, and Sandia National Laboratory.

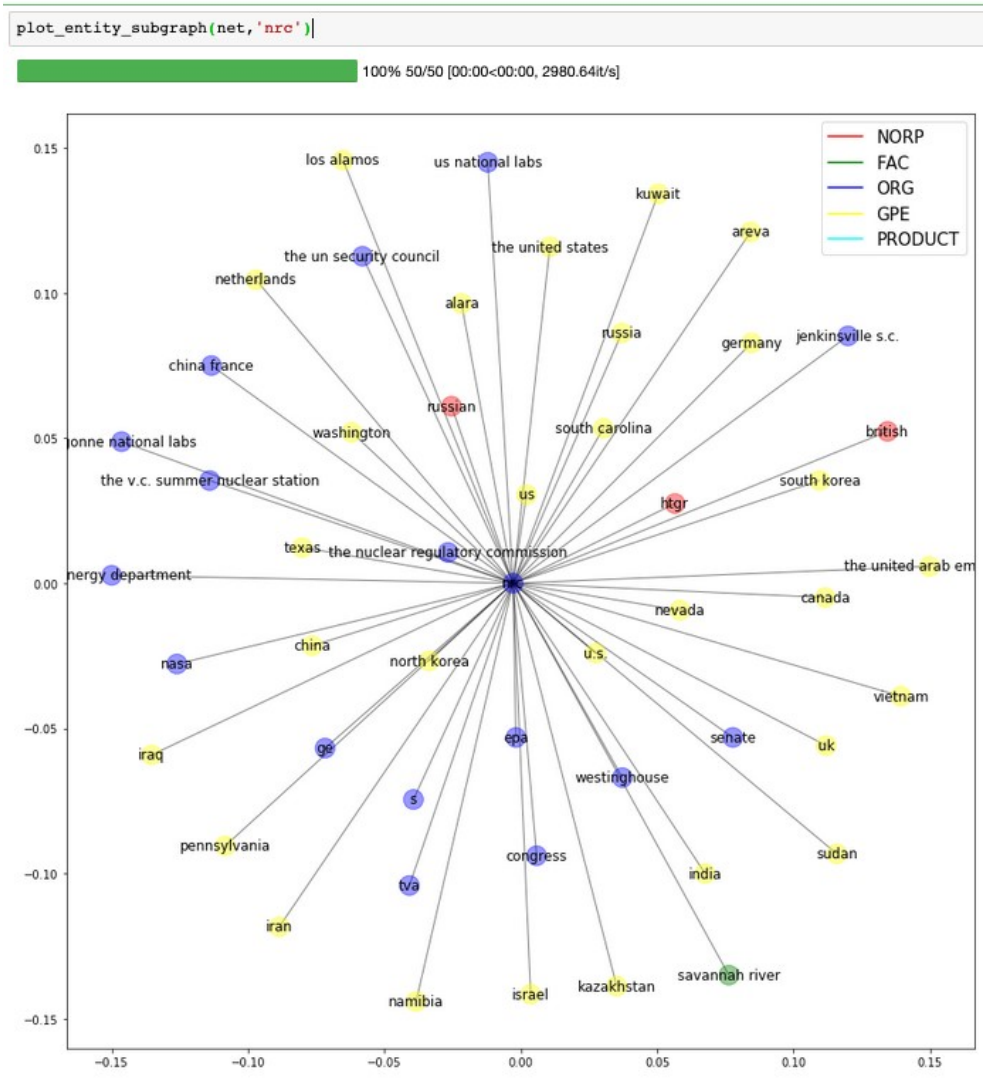


Figure 1-6: Heterogeneous entity ego-network (*nrc*).

The distances and positions of nodes are arbitrary and have been chosen only for visual clarity.

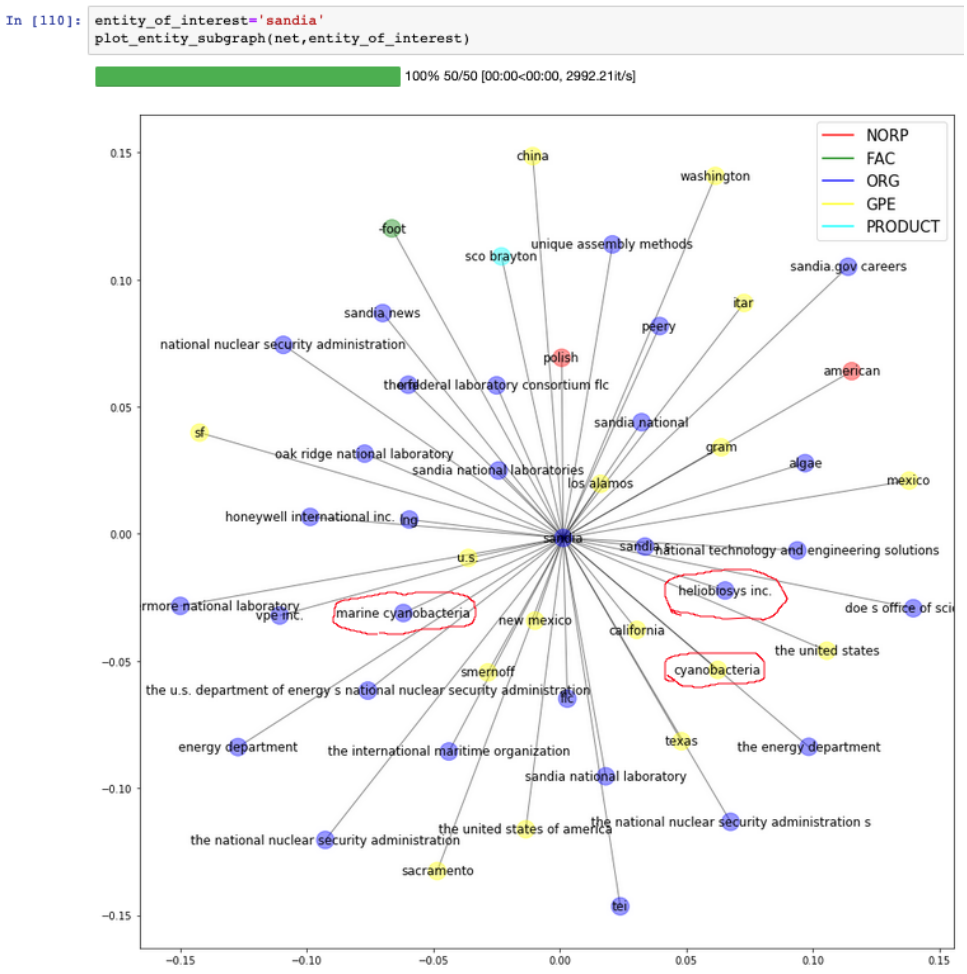


Figure 1-7: Heterogeneous entity ego-network (*sandia*).

The distances and positions of nodes are arbitrary and have been chosen only for visual clarity.

Complete details regarding the preliminary steps of the modeling pipeline can be found in [2].

1.1.5 FY21 Demonstration Prototype Development

In FY21, the team has continued development moving toward a demonstration prototype modeling pipeline. The time dependent word embedding models that were developed in the preliminary prototype continue to be the foundation of the modeling pipeline. However, while the preliminary effort showed a valid connectedness between keywords and phrases, it did not provide additional context about the events contained in the corpus that led to the results. Therefore, in FY21, the team has incorporated a number of new techniques to further enrich the datasets and extract events based on the analysis of the time dependent word embedding models. The following sections will outline the analysis pipeline and offer a performance assessment of the capabilities when the datasets are modified.

2.0 Twitter Dataset Modeling Pipeline

FY20 preliminary modeling efforts explored the use of time dependent word embedding models trained on the Twitter corpus. Positive signal from the time dependent word embedding approach has led to further development of the modeling pipeline. In the following subsections, all time dependent word embedding models use the static, growing window method with a growth rate of 30 days (refer to [2] report for the

training algorithm). This method was found to be more robust in that fewer terms were dropped from the vocabulary over time when compared to a rolling window approach.

After training the time dependent word embedding models, the modeling pipeline is outlined by the following steps:

1. Create cosine similarity-based network graphs for all keywords and all time intervals
2. Compute the weighted averaged embedding vectors of network graphs from Step 1
3. Compute time dependent similarity metrics (e.g., Jaccard similarity, cosine similarity, Euclidean distance, Hamming distance, Damerau-Levenshtein distance) comparing network graphs between successive time periods
4. Compute inflection points of the similarity metric profiles to identify time intervals where a contextual shift occurs for all keywords
5. Explore the matrix of contextual shifts across time to identify time periods of interest
6. Search the corpus for events of interest based on similarity comparisons of the Tweets, keywords, and network graphs

In the following subsections, the steps of the modeling pipeline will be described in detail and results from the approach will be presented.

2.1 Graph-Based Approach for Identifying and Analyzing Contextual Shifts

2.1.1 Cosine Similarity-Based Network Graphs

A network graph-based approach was developed to identify and analyze contextual shifts in key words and phrases over time. The approach leverages the time dependent word embedding models at each time interval by identifying the most similar words (i.e., based on the cosine similarity metric) to a user-selected keyword and is iterated based on the user-specified rank and number of nearest neighbors. For example, a network graph with only one rank and a user-specified ten nearest neighbors would contain the ten words with the highest cosine similarity to the keyword. A network graph with two ranks and a user-specified ten nearest neighbors would contain the ten words with the highest cosine similarity to the key word in the first rank and the ten most similar words to each first rank word in the second rank (i.e., 110 total nodes). In other words, each successive rank continues adding the N terms with the highest cosine similarity to its parent node in the preceding rank. This concept is illustrated in Figure 2-1. Note that in this approach, duplicate terms may appear at different nodes in the network graph unless prevented by the user.

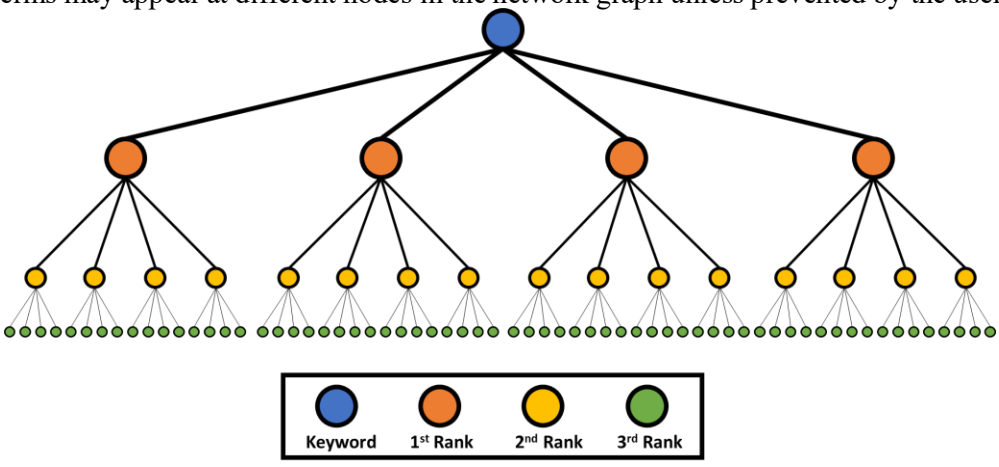


Figure 2-1. Illustration of the network graph-based approach for a graph with three ranks and four nearest neighbors.

The graph-based approach allows a user to compare successive graphs over time as ordered sets using metrics such as the Jaccard similarity, which only accounts for the presence of key words in two different

graphs and not the ordering. Because the graphs are ordered with respect to the similarity rank, additional metrics such as the Hamming distance (i.e., the number of positions in the ordered set which are different) or Damerau-Levenshtein distance (i.e., the number of insertions, deletions, substitutions, or transpositions required to make the sets the same) can be used to identify the number of operations required to transform the graph at time *t* to the graph at time *t+1*, indicating the degree to which a graph is evolving over time. Note that the computation of the Damerau-Levenshtein distance is more computationally expensive than the Hamming distance, but can provide more information about the sets of terms in the graph.

In addition to treating the graphs as ordered sets, edges of the graph are used to compute a weighted average embedding vector, as illustrated in Figure 2-2. Here, the edge weight is computed as:

$$EdgeWeight = w_n = \left(\frac{R_{W_n} f_{W_n}}{\sum_{i=1}^{N} f_{W_i}}\right)^{-1}$$

R_{W_n}: Nearest neighbor similarity rank
f_{W_n}: Nearest neighbor term frequency
$\sum_{i=1}^{N} f_{W_i}$: Cumulative term frequency for all terms in the vocabulary

This definition applies higher weights to less frequently occurring terms in the corpus and to the terms that are most similar to the keyword (i.e., the first most similar term has a higher weight than the tenth most similar term) and allows exploration of the nodes that have the highest influence on the keyword's contextual representation.

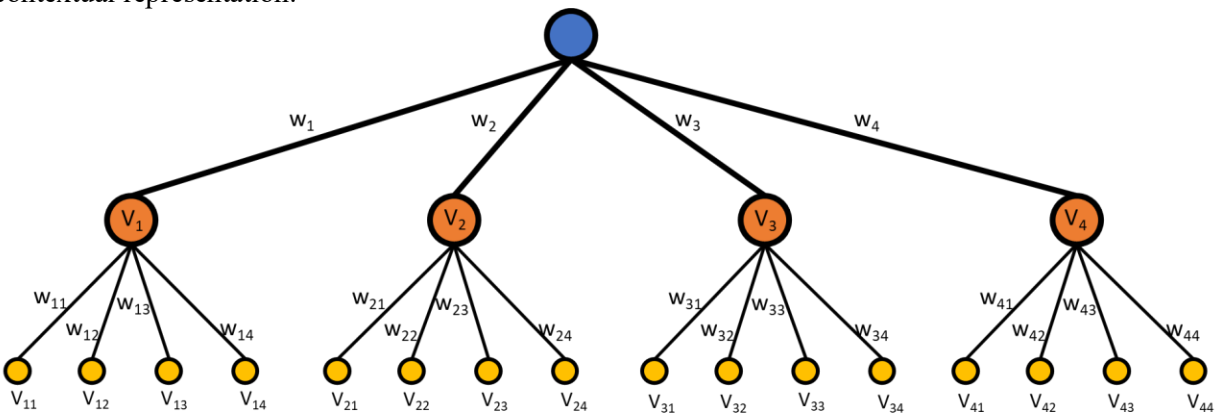


Figure 2-2. Illustration of the graph-based approach for computing the weighted average embedding vector in a graph with two ranks and four nearest neighbors.

By computing this new vector representation, two additional similarity metrics can be computed to identify contextual shifts in keywords and phrases: 1) the Euclidean distance between the weighted average embedding vector at time *t* and time *t+1* and 2) the cosine similarity (i.e., the normalized dot product) of the weighted average embedding vector at time *t* and time *t+1*. These two metrics capture changes that occur to both the magnitude and angle of the vector, which can indicate a growing contextual representation (i.e., increasingly broad context or high influence from a term in the graph that has a broad context) and/or a high degree of change in the surrounding terms. Note that when using the dot product similarity metrics to compare the weighted embedding vector of two successive time windows, the embedding vectors at time

t+1 are translated into the same space as the embedding vectors at time *t* using a modified version of reference 4.

Figure 2-3 shows an example of the graph-based similarity metrics for a graph of rank 1 with ten nearest neighbors for the term "pit_production" across 46 total time windows. Here, each data point represents a comparison between two successive embedding models using the graph-based similarity metrics. The ten-node graph for each time window is shown in Figure 2-4 through Figure 2-8, where the nodes are listed in order of highest to lowest cosine similarity from left to right. The influence rank captures the influence each term has on the weighted average embedding vector of the keyword (i.e., 1 = highest influence rank and 10 = lowest influence rank) relative to other terms in the graph, and the influence captures the total influence of the term on the embedding vector, computed as:

$$TermInfleunce = \|V_K - V_i\| \frac{1}{R_i * TF_i/TF_{Total}}$$

$\|V_K - V_I\|$: The magnitude of the difference of the keyword vector and the vector of the term at node i
V_K: The keyword vector
V_i: The vector for the term at node i
R_i: The similarity rank of the term at node i
TF_i/TF_{Total}: The frequency of the term at node i divided by the total count of all terms in the embedding vocabulary

Note that up to time interval 21, the Jaccard similarity, hamming distance, and Damerau-Levenshtein distance are all close to zero. Inspection of the networks shows that there are few common nodes in the graphs during those time periods. At time interval 21, the Jaccard similarity shows that there is an increase in repetitive terms between successive time periods, though the Hamming distance and the Damerau-Levenshtein distance show that the terms are reordered, with the highest similarity in ordering being in time interval 35. Inspecting the two vector similarity metrics shows similar results, where early time periods have a relatively small Euclidean distance between the two vectors and a cosine similarity close to 1, indicating little contextual shift has occurred due to a low number of mentions within the corpus. Notably, time interval 11 represents an inflection point, where the cosine similarity and Euclidean distance begin to undergo more significant change, reaching a peak for both metrics around time interval 20 and 21, respectively. Inspecting the networks reveals that it is around those time periods that similar nodes begin to repeat in the networks and a contextual representation for "pit_production" begins to exist in the embedding models that is closely related to DOE Weapons Complex nuclear activities, as well as with pit production (e.g., "savannah_river_site", "srsnews", "wcmonitor" [Weapons Complex Monitor], "rockyflats", "wipp", and "lanl").

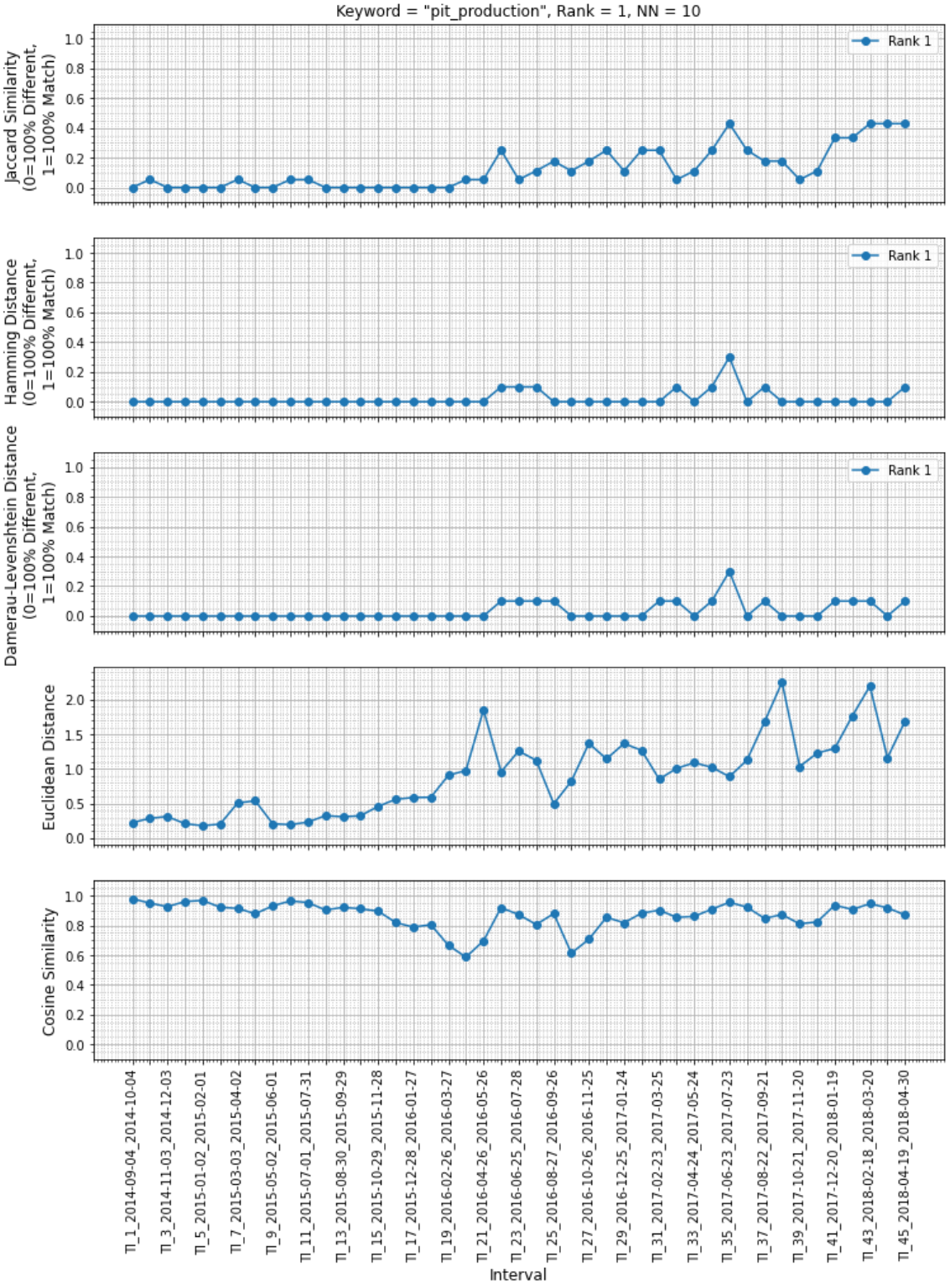


Figure 2-3. Graph-based similarity metrics for the graph of "pit_production", rank 1 and 10 nearest neighbors.

Time Interval	*Top N Most Similar Terms (Influence Rank, Influence)*									
0	groundwater (1, 13903.1)	referred (2, 5582.7)	minor (3, 5526.5)	lour (4, 3577.4)	shahbaz (5, 2495.9)	nrc (6, 1909.9)	fate (7, 1727.9)	spending (10, 928.8)	victim (9, 1072.3)	issued (8, 1396.8)
1	reserve (1, 15618.2)	exelis (2, 11772.5)	driven (3, 5761.7)	boiler (7, 3273.1)	input (6, 3990.1)	recommendation (10, 2635.1)	superb (9, 2688.7)	algae (4, 4181.8)	vandal (5, 4141.2)	institution (8, 2790.3)
2	lawbc (3, 20060.8)	abengoa (2, 21301.3)	algae (1, 22817.7)	biomass (4, 14088.7)	thatarcher (6, 5605.1)	sierra (7, 5396.9)	webinars (9, 4545.5)	explorer (8, 5368.2)	solicitation (5, 6196.3)	std (10, 2666.1)
3	each (2, 29609.3)	causing (1, 31372.0)	driven (3, 24332.8)	legalize (6, 10088.2)	shipbuilding (4, 15309.5)	wms1 (5, 10586.6)	fomento (9, 5757.0)	whitney (7, 8771.0)	licensing (10, 4049.0)	denial (8, 6671.1)
4	impulsa (1, 63715.6)	thrilled (2, 41608.3)	algae (3, 33478.0)	educate (4, 14078.4)	hostile (8, 8301.2)	workshop (10, 8028.7)	incentive (7, 8774.2)	journalistic (5, 9626.0)	governmental (9, 8064.2)	juba (6, 9258.2)
5	thunderclapit (1, 88052.3)	tribe (3, 34829.1)	publicadas (2, 40062.4)	huis (4, 24146.1)	centroamericano (6, 14033.5)	footprint (5, 20057.4)	ottnews (8, 10349.3)	discourse (9, 9239.3)	carib (7, 13195.8)	anger (10, 6200.5)
6	refrigeration (1, 113584.2)	default (2, 79402.1)	jgi (3, 51124.5)	bln (5, 25035.2)	improving (6, 21220.6)	ames (4, 27224.6)	katespade (7, 19493.8)	desired (8, 15610.3)	konsultasi (10, 6532.8)	fiap (9, 11335.8)
7	charging (1, 118680.3)	portion (2, 57126.1)	methane (4, 41596.9)	ames (3, 52399.4)	census (6, 20838.6)	ecological (5, 31403.9)	penalty (10, 13619.5)	generating (7, 20510.7)	shale (8, 17148.7)	congestion (9, 15936.1)
8	commish (1, 168028.2)	theguardian (2, 74458.8)	akhbar (4, 41268.3)	eia (5, 32679.5)	intact (9, 21849.5)	alaskan (3, 45621.5)	depts (7, 24219.2)	allonew (6, 27085.1)	desired (10, 21034.0)	scgsr (8, 22468.0)
9	thunderclapit (1, 283454.0)	jongen (2, 70473.0)	standardization (3, 67878.5)	bomberosgye (5, 34850.8)	beyondscaredstraight (4, 44145.5)	wewanthollandotratour (10, 18584.8)	cndores (6, 26384.6)	swaggie (7, 23264.8)	declarado (9, 18959.5)	kwf (8, 22611.1)

Figure 2-4. Graph (influence rank, total influence) for time intervals 0 through 9.

Time Interval	*Top N Most Similar Terms (Influence Rank, Influence)*									
10	eerily (1, 262502.0)	awar (3, 59616.4)	comenz (5, 49660.9)	wewanthollandotratour (9, 26280.4)	educativa (2, 71356.7)	terraformars (4, 50085.7)	dissemination (8, 35164.3)	leugenaar (7, 36043.4)	trabaja (10, 13502.1)	preventivas (6, 38674.2)
11	profitable (1, 188189.6)	harnessing (2, 129008.9)	konsultasi (8, 33347.7)	sharylattkisson (3, 69273.0)	thrilled (7, 38162.1)	oranje (4, 55629.5)	lancashire (5, 41030.2)	minmin (6, 38191.8)	cqnow (10, 12453.4)	trabaja (9, 15769.7)
12	ecology (1, 226499.3)	refrigeration (2, 97035.4)	nnsas (5, 68068.7)	westinghouse (3, 81512.9)	inl (7, 31834.9)	vpn (4, 69813.2)	scramble (9, 26486.6)	density (8, 30829.4)	nersc (10, 23638.2)	erp (6, 38701.6)
13	stricken (1, 249145.7)	dfars (2, 224055.3)	optimization (3, 81457.0)	advancing (7, 41728.1)	velocity (4, 79130.2)	ians (5, 77653.1)	national_nuclear_security_administration (10, 26109.2)	icp (6, 58923.9)	nuclear_security_enterprise (8, 40863.9)	nuclear_weapon_component (9, 33886.1)
14	westinghouse (1, 272899.0)	algal (2, 183421.2)	esri (4, 90645.8)	pedestrian (3, 116950.6)	sorghum (5, 60904.2)	floated (6, 48044.1)	towing (8, 40916.6)	aggregate (7, 41387.3)	methodology (9, 39531.3)	netl (10, 26863.4)
15	paducah (1, 261572.1)	congestion (2, 144903.2)	evaluating (5, 86742.2)	input (8, 41880.3)	icp (3, 116638.0)	projecting (4, 90360.3)	exascale (6, 45112.8)	solicitation (9, 28693.9)	genomics (7, 43086.1)	commercialization (10, 28515.4)
16	landscaping (1, 721313.0)	corpus (3, 160205.5)	salesforce (4, 154845.9)	elsegundo (2, 179393.3)	llvm (10, 66935.0)	assembler (6, 91611.6)	elearning (5, 108336.6)	systmes (7, 90335.3)	energization (8, 80301.0)	requi (9, 77984.7)
17	rathlin (1, 859952.1)	energypresssec (3, 158778.5)	petaflops (2, 248805.4)	paducah (5, 87067.4)	solarcity (4, 134811.9)	offshorewind (7, 70810.5)	levy (8, 54727.4)	compute (6, 73676.7)	solicitation (9, 30514.9)	nuclearenergy (10, 27183.8)
18	graphene (1, 319333.3)	wegen (2, 314830.1)	departmentofenergy (4, 135616.3)	coyote (7, 73463.3)	dock (8, 48816.8)	lush (5, 115283.3)	berweisungszweck (3, 144697.5)	transformational (10, 40096.3)	pixel (9, 41922.4)	koat (6, 81760.3)

Figure 2-5. Graph (influence rank, total influence) for time intervals 10 through 18.

Time Interval	*Top N Most Similar Terms (Influence Rank, Influence)*									
19	savannah_river_site (3, 170969.8)	llvm (1, 264079.4)	fortran (2, 202033.1)	compiler (5, 126536.4)	compute (4, 129222.2)	solicitation (8, 50431.2)	cleanup (10, 26579.6)	srsnews (6, 62277.2)	wcmonitor (9, 45881.9)	kurion (7, 53797.8)
20	parkhomov (1, 1261856.9)	radiology (2, 538167.5)	graphene (4, 121831.5)	chariot (3, 325437.3)	nrc (7, 40661.5)	cleanup (10, 30757.3)	southcarolina (6, 58293.7)	brine (5, 78639.0)	tritium (9, 36125.6)	carlsbad (8, 37738.0)
21	cleanup (1, 182673.0)	inl (2, 162817.3)	ornl (4, 77399.8)	savannah_river_site (6, 46281.1)	wcmonitor (3, 94727.3)	milestone (9, 27810.0)	repository (5, 53183.6)	energypresssec (7, 45622.3)	idaho (10, 27096.0)	separation (8, 41791.3)
22	cleanup (1, 186555.7)	tritium (2, 159314.1)	wipp (5, 85234.0)	mox (7, 66377.0)	repository (6, 79897.8)	cooled (3, 100923.7)	gaseous (4, 90674.5)	inl (8, 40577.9)	savannah_river_site (10, 20291.3)	fission (9, 37323.5)
23	cleanup (1, 200301.1)	graphene (2, 195812.9)	lanl (4, 85846.8)	abq (5, 85332.5)	solicitation (7, 68306.9)	yucca (6, 69310.2)	energypresssec (8, 52527.6)	ornl (10, 31256.4)	emitting (3, 186641.5)	licensing (9, 37956.5)
24	cleanup (2, 211338.7)	parkhomov (1, 894109.7)	separation (3, 140202.5)	finishing (7, 69689.2)	yucca (4, 85639.8)	srsnews (6, 80215.9)	priced (5, 84258.6)	isotope (10, 43498.7)	radium (8, 60911.8)	forge (9, 48088.1)
25	rockyflats (2, 569172.7)	cleanup (6, 110577.4)	parkhomov (1, 633687.6)	wipp (8, 77437.6)	sciencenews (5, 128010.9)	berkeleylab (7, 99814.8)	isotope (9, 51493.5)	lanl (10, 35014.8)	decorating (4, 137403.4)	zedman (3, 159327.7)
26	fining (1, 1955554.0)	solicitation (3, 194622.5)	berkeleylab (2, 224081.3)	hpc (6, 84122.5)	greenway (4, 190606.0)	lanl (10, 52034.2)	inl (9, 61048.6)	syngas (8, 64016.8)	fortran (5, 102301.4)	conditional (7, 75261.1)
27	isotope (1, 410602.6)	idaho (4, 162727.9)	rockyflats (2, 193415.7)	inl (6, 111016.9)	beryllium (3, 179436.6)	solicitation (8, 72947.7)	hpc (9, 53328.7)	neutrino (7, 74449.9)	emitting (5, 155844.8)	cleanup (10, 17826.3)

Figure 2-6. Graph (influence rank, total influence) for time intervals 19 through 27.

Time Interval	*Top N Most Similar Terms (Influence Rank, Influence)*									
28	yucca (1, 513402.6)	srsnews (2, 278974.0)	solicitation (4, 147296.9)	rockyflats (3, 155173.8)	wipp (6, 68854.3)	nrc (8, 54548.8)	inl (7, 68553.8)	paducah (5, 73646.9)	cleanup (10, 20499.2)	savannah (9, 38751.4)
29	processing (2, 372750.5)	decaying (1, 403019.3)	lanl (3, 116338.0)	wipp (6, 93770.1)	isotope (5, 95062.3)	recycling (7, 90549.9)	eis (8, 77763.3)	mapping (10, 71063.6)	rockyflats (9, 75047.3)	beryllium (4, 102641.4)
30	ecology (1, 679504.9)	srsnews (2, 365692.5)	rockyflats (3, 278633.5)	disposition (4, 200843.2)	processing (8, 89930.0)	isotope (7, 94451.7)	nmpol (5, 164167.4)	nnsanews (9, 76440.6)	ucberkeley (6, 114415.4)	lanl (10, 42460.8)
31	rockyflats (1, 926061.4)	lanl (4, 230100.9)	powering (3, 271759.7)	processing (5, 124147.7)	isotope (7, 114745.6)	proton (2, 332899.7)	pnnlab (6, 123992.3)	hpc (9, 62905.8)	berkeleylab (8, 97887.4)	packaging (10, 53134.9)
32	srsnews (1, 734355.1)	generates (3, 473520.1)	carlsbad (4, 207016.1)	yucca (5, 191875.5)	isotope (8, 123176.1)	usnistgov (2, 490429.5)	levy (7, 129651.5)	inhaled (6, 178994.0)	plymouth (10, 71157.2)	tritium (9, 72097.7)
33	processing (1, 521036.5)	dounreay (2, 493543.4)	transuranic_waste (3, 340273.2)	isotope (6, 165924.5)	wippnews (4, 235400.3)	geological (7, 152826.2)	gaseous (5, 219558.3)	yucca (8, 104295.5)	wipp (10, 56971.7)	nuclearwaste (9, 71840.7)
34	processing (1, 516549.0)	isotope (3, 334102.9)	dounreay (2, 341026.1)	nuclearwaste (5, 187173.7)	uraniu (4, 269242.3)	nrc (8, 81179.4)	mox (10, 69996.2)	srsnews (6, 93973.0)	plymouth (7, 89029.7)	recycling (9, 72887.2)
35	processing (1, 528734.8)	isotope (2, 337052.4)	lanl (5, 172733.6)	nrc (7, 123954.8)	disposition (4, 186935.2)	abq (8, 112246.0)	uraniu (3, 198834.2)	dounreay (6, 138873.5)	yucca (9, 86233.5)	recycling (10, 68986.8)
36	lanl (1, 523682.7)	recycling (2, 381898.8)	geological (3, 342799.2)	injection (5, 231224.7)	mox (7, 100152.5)	wcmonitor (4, 266619.6)	isotope (6, 100971.2)	abq (8, 82597.6)	liquid (10, 68914.3)	ecology (9, 79521.1)

Figure 2-7. Graph (influence rank, total influence) for time intervals 28 through 36.

Time Interval	*Top N Most Similar Terms (Influence Rank, Influence)*									
37	methane (1, 815160.3)	rockyflats (2, 511494.8)	tritium (4, 217896.3)	injection (3, 246417.2)	liquid (6, 130139.2)	isotope (7, 124223.3)	repository (8, 105616.5)	processing (10, 74544.4)	generate (9, 79098.7)	cartridge (5, 161496.9)
38	isotope (3, 722310.9)	chyronhego (1, 1801158.8)	recycling (5, 282002.5)	rockyflats (6, 244826.9)	bringhimhome (2, 1063102.2)	tritium (9, 104592.5)	nuclearpower (10, 95301.1)	unshielded (4, 375582.4)	powering (7, 110417.5)	disposition (8, 105304.9)
39	cooled (1, 1530262.8)	isotope (2, 391379.9)	lanl (6, 190270.7)	processing (7, 149038.6)	sludge (3, 309708.3)	uraniu (4, 239308.3)	yucca (8, 130920.4)	portsmouth (9, 93823.3)	beryllium (5, 211027.5)	particle (10, 68859.5)
40	tritium (1, 565170.1)	rockyflats (3, 487845.3)	nuclearpower (4, 275920.1)	nuclearwaste (6, 223473.2)	particle (8, 140787.9)	isotope (9, 103677.5)	ludwig (2, 559651.6)	srsnews (7, 142149.5)	hexafluoride (5, 270717.3)	mox (10, 51778.6)
41	rockyflats (1, 943006.7)	nuclearwaste (2, 364308.8)	tritium (5, 194437.4)	paducah (4, 264742.2)	electron (3, 279629.3)	recycling (7, 138210.8)	disposing (6, 154519.7)	nuclearpower (10, 102281.9)	srsnews (8, 119282.4)	byproduct (9, 113021.8)
42	tritium (1, 536741.7)	processing (2, 305685.6)	particle (5, 211072.8)	recycling (4, 213115.6)	disposing (3, 216220.0)	isotope (8, 115359.6)	rockyflats (6, 145272.8)	mox (10, 66462.5)	nuclearwaste (9, 91874.3)	dounreay (7, 134986.5)
43	mox (2, 513099.5)	tritium (4, 301340.9)	rockyflats (3, 354437.3)	disposing (5, 284189.0)	nuclearwaste (7, 214379.7)	shipment (9, 102507.0)	nuclearpower (6, 220253.5)	reprocessing (10, 82411.1)	dounreay (8, 201654.0)	subatomic (1, 1927334.7)
44	nuclearwaste (1, 1089915.4)	dounreay (2, 947508.5)	nuclearpower (3, 562277.3)	tritium (5, 159029.2)	repository (6, 144443.7)	mox (9, 86110.8)	isotope (8, 98932.2)	rockyflats (7, 137144.2)	byproduct (4, 210189.3)	wipp (10, 63884.8)

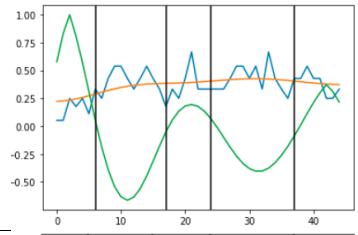
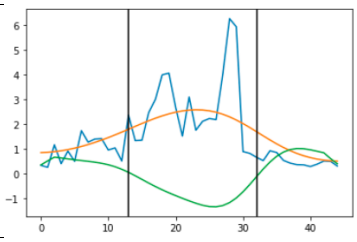
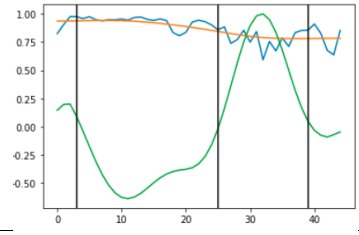
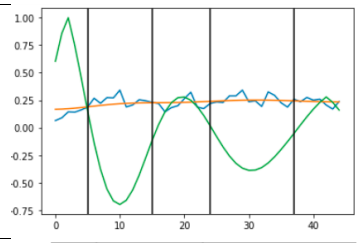
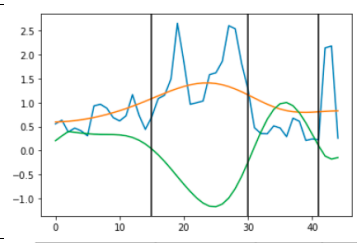
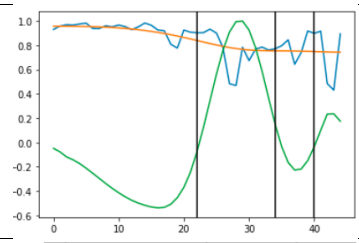
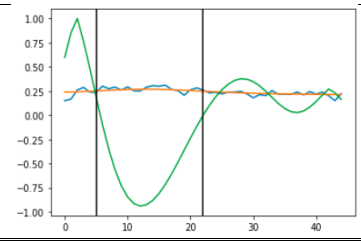
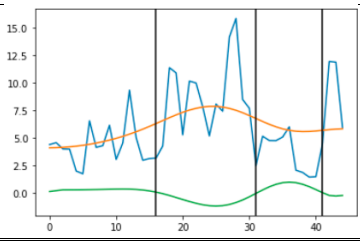
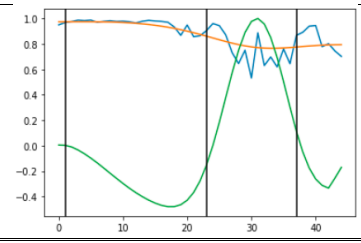
Figure 2-8. Graph (influence rank, total influence) for time intervals 37 through 44.

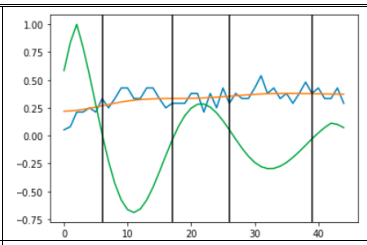
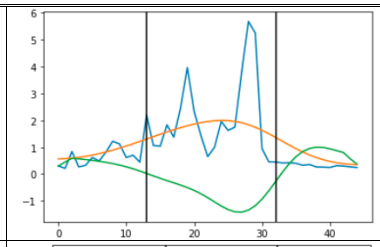
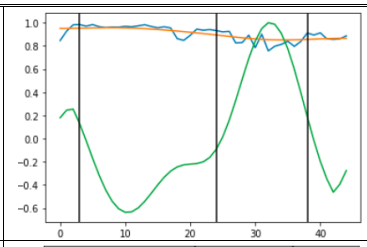
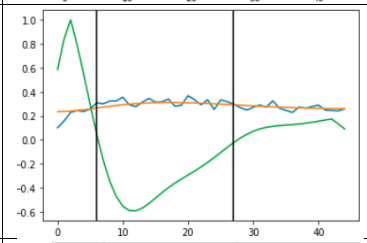
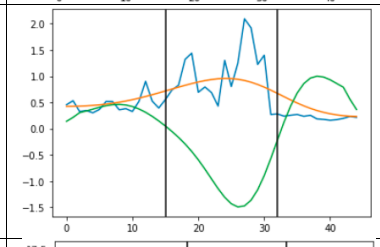
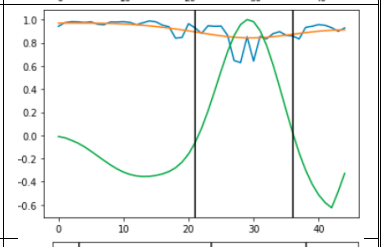
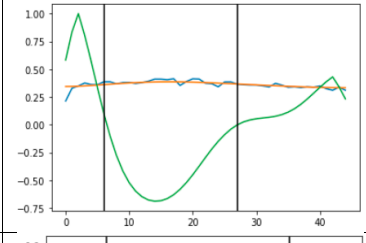
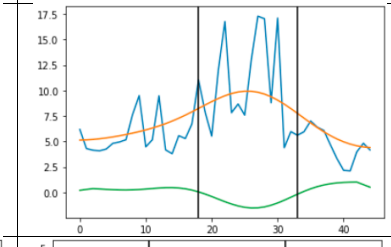
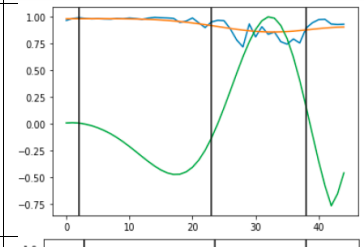
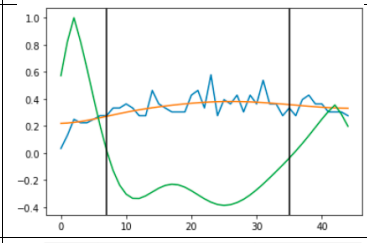
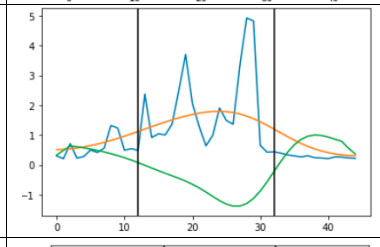
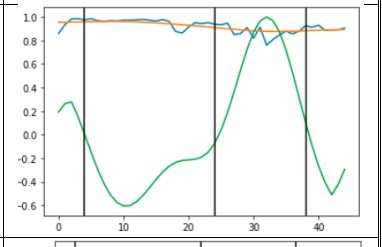
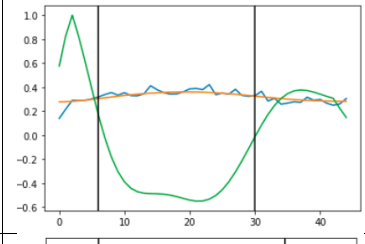
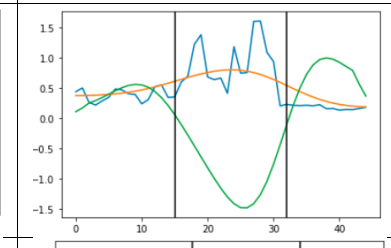
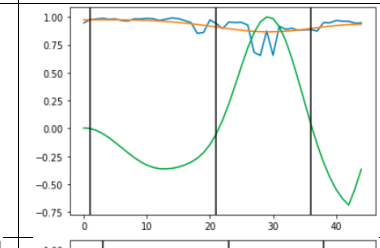
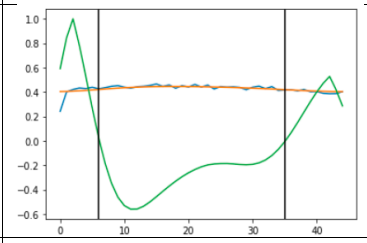
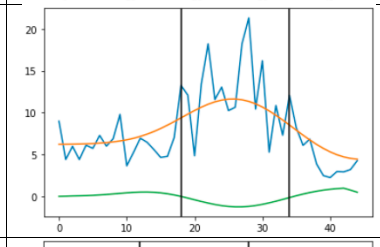
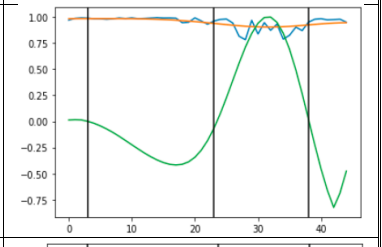
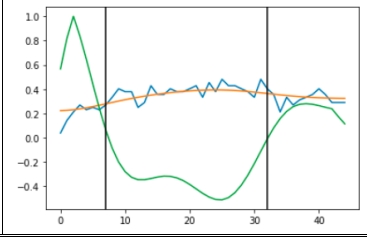
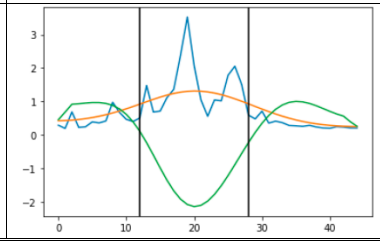
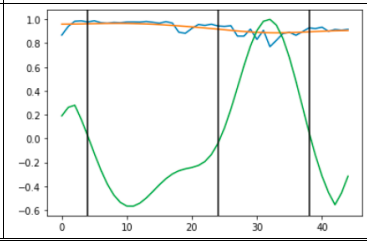
2.1.2 Ensemble Approach for Identifying Time Periods with Contextual Shifts

The example provided in Section 2.1.1 illustrates how even a small graph network of ten nodes can provide insights about the potential occurrence of an activity or event of interest. However, this approach alone does not provide context or real events. Additionally, each graph (i.e., with different numbers of ranks and nearest neighbors) may have a different temporal profile across the similarity metrics. The degree to which each graph's similarity metrics differ over time depends on the breadth of the contextual representation and is not straightforward to determine *a priori*. Therefore, to extract events of interest, an ensemble approach is used, whereby the similarity metrics for several different network graphs for the same keyword are analyzed.

Because the temporal profile of the similarity metrics can have significant oscillations, a Gaussian filter is applied to smooth the temporal profile and the second derivative of the smoothed profile is calculated to identify inflection points. Subsequently, the inflection points computed across all similarity metrics are combined to create time windows of interest, where further analysis (i.e., of the graphs and Tweets within the time interval) can be performed to extract events. (Note: Because the Hamming distance and Damerau-Levenshtein distance metrics have primarily abrupt step-like changes, inflection points here are only computed on the Jaccard similarity, Euclidean distance, and cosine similarity temporal profiles.)

The ensemble approach is illustrated in Figure 2-9 for the term "national_security", where the vertical black lines represent the computed inflection points of the second derivative, which are the time intervals where a contextual shift is presumed to have occurred for the key word using that network graph and similarity metric. When combining the time windows across the ensemble, if two immediately successive time intervals are captured, the upper bound length of the window is taken such that the time window is never smaller than two time intervals. Using this approach, the final time windows for the term "nuclear_security" are marked at intervals 7, 12, 15, 18, 20, and 41.

Rank/NN	Jaccard Similarity	Euclidean Distance	Cosine Similarity
1 / 10			
2 / 10			
3 / 10			

1 / 20			
2 / 20			
3 / 20			
1 / 30			
2 / 30			
3 / 30			
1 / 40			

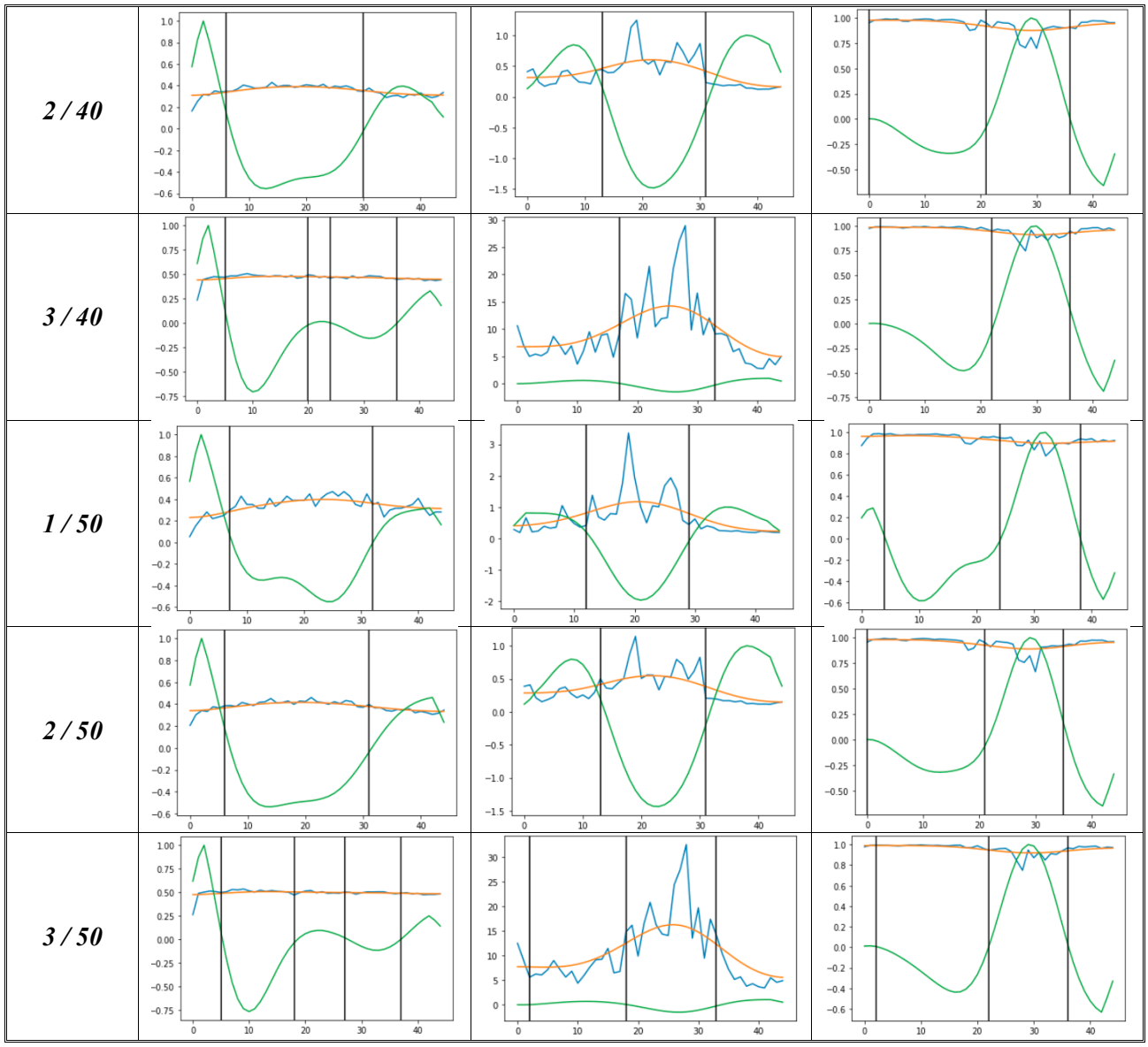


Figure 2-9. Ensemble approach for identifying important time windows of contextual change for the term "nuclear_security". Blue, yellow, green, and black lines represent the similarity metric, smoothed profile (where the Gaussian filter was applied), the second derivative of the smoothed profile, and the inflection point of the second derivative, respectively.

The ensemble approach is used to compute the inflection point time windows across all keywords to identify overlapping time intervals where contextual shifts in several keywords and phrases are detected. The matrix of overlapping time windows for all keywords in the vocabulary is shown in Figure 2-10. Here, the grid cells are colored based on the number of keywords with an inflection point falling at that interval. Using this approach, the Twitter corpus can be explored across a single keyword of interest's time intervals, or for clusters of keywords and time intervals where inflection points are detected. Note that it is not required that a new appearance of a keyword occurs at a given time interval for the embedding vector to change. Rather, each time the embedding models are updated, a shift in the embedding vector will occur. Thus, this ensemble approach captures terms that may have led to an indirect contextual shift (i.e., contextual shifts resulting from the contextual shift of a closely related term) of a specific keyword of interest.

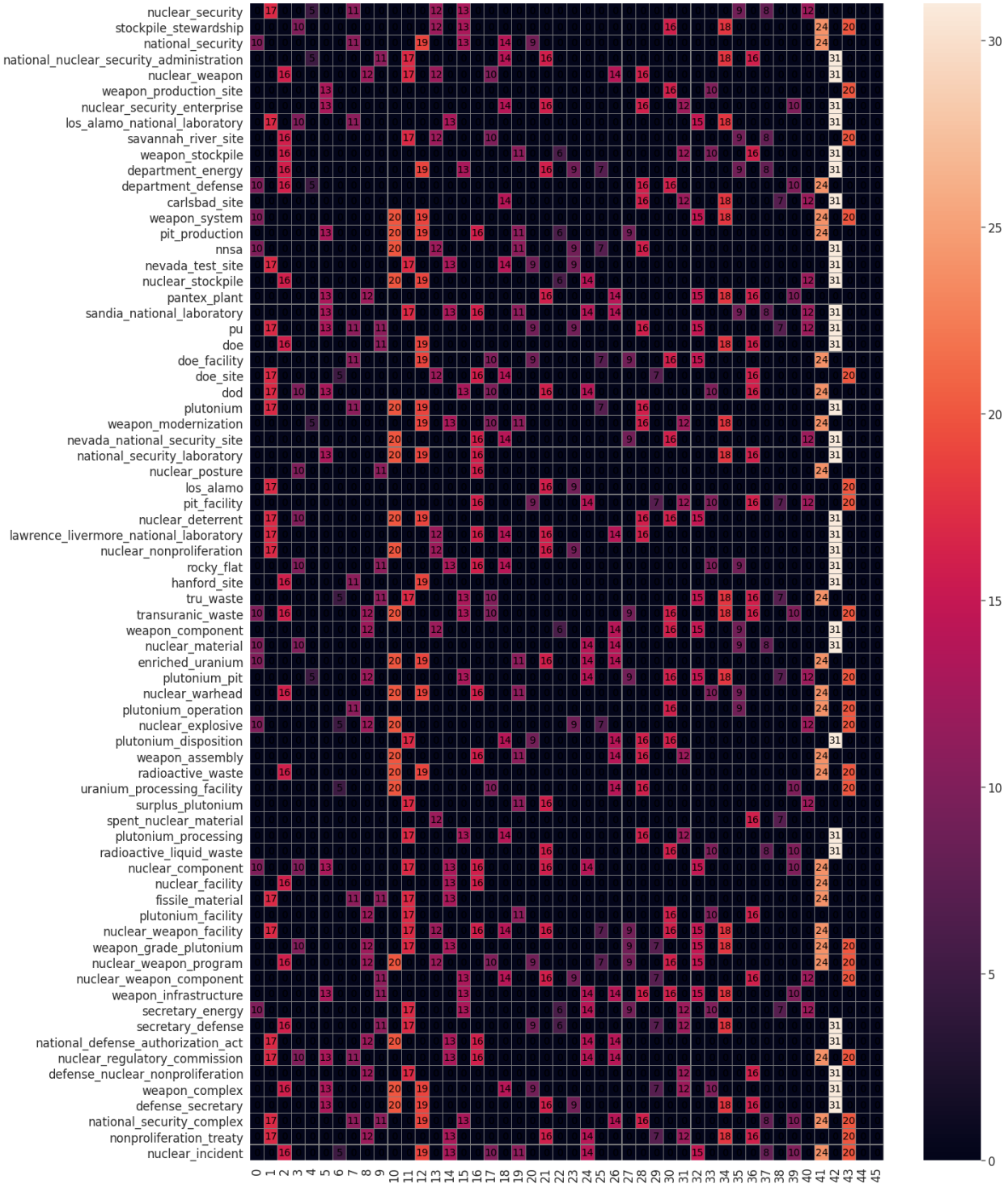


Figure 2-10. Matrix showing the time periods where inflection points were detected. Annotated numbers and colors indicate the number of key terms with an inflection point in that time interval.

2.1.3 Extracting Events of Interest

The matrix of inflection points in Figure 2-10 acts as a guide in the modeling pipeline to explore how a keyword or group of keywords evolve over time by searching for Tweets containing events of interest around/within these inflection point intervals. Three metrics are used to extract relevant Tweets from the

windows: 1) the cosine similarity of the mean embedding vector of all tokens in a Tweet and the weighted average graph embedding vector, 2) the cosine similarity of the base keyword embedding vector and the average embedding vector of all tokens in a Tweet, and 3) a direct mention of the keyword within a Tweet. A threshold similarity value is applied such that any Tweet that does not have a similarity greater than that threshold and does not contain a keyword is filtered out.

These filtration methods were applied with a threshold similarity of 0.75 for three terms relevant to the development of fissile core fabrication at the Savannah River Site: "pit production", "plutonium disposition" (i.e., as related to the Mixed Oxide fuel fabrication facility, or MOX, which was repurposed to be the Savannah River Plutonium Processing Facility), and "plutonium pit". Table 2-1 shows the Tweets that were extracted for these three terms, reduced from a list of duplicates (i.e., slight variations) and with "noisy", or unrelated, Tweets removed. Notably, from the earliest time period of the Twitter corpus, Tweets were discovered that captured the evolving conversation about pit production including speculation (before the official announcement) that pit production would be carried out at the Savannah River Site in addition to Los Alamos.

Table 2-1. Events of interest extracted for the key terms "pit production", "plutonium disposition", and "plutonium pit". Inflection window denotes the inflection window at which the event was detected.

Date	*Inflection Window*	*RawText*
Pit Production (Inflection at Intervals 5, 10, 12, 16, 19, 22, 27, 41)		
8/26/2014	5	Why Do DOE And LANL Refuse To Do A Pit Production Study?: A recent Congressional Research Service (CRS) Report... http://t.co/ngcEaVuVB1
9/24/2014	5	.@Westinghouse acquires #Italian #nuclear component manufacturer http://t.co/LvkoXiCXzW
12/17/2014	5	Cancel the "prescribed burn" at Rocky Flats http://t.co/6lAfQjtg00 @moveon
1/29/2015	5	NRC: DOE has met regulatory requirements for #Yucca repository
6/9/2015	10	Nuclear Weapon Pit Production: Options to Help Meet a Congressional Requirement R44033 http://t.co/1om1lwSu06
7/29/2015	12	@WMN4SRVL Another age old cry there r insuff funds for cleanup but plenty of cash for pit production
1/18/2016	19	LANL poised to ramp up plutonium pit production: report https://t.co/VR7nw8j3Ug
1/22/2016	19	Steps toward pit production made at Los Alamos - The Albuquerque Journal https://t.co/TJAzrrqH7a
2/20/2016	19	Obama's big nuke budget is a bowl of cherries for pit production at #LosAlamos. My latest: https://t.co/FXQHkMz9Sq https://t.co/DdTtkv3gcF
3/20/2016	19	Reader View: Accelerated plutonium pit production wrong for #NewMexico https://t.co/3QGoBT4OPV
4/5/2016	22	History of Plutonium Pit Production https://t.co/QRgiEPYJOu
7/2/2017	41	Amarillo's @PantexPlant could take on plutonium pit production. https://t.co/sFntSDToDJ

7/4/2017	41	DoE eyes terminating Savannah River MoX to use site for plutonium pit production & possible tritium work https://t.co/QoGRu1TNVk
8/9/2017	41	@GerryDales Leslie Groves would be turning over in his grave if he could see the Hanford vit plant
8/22/2017	41	Proposals to establish opposition to LANL pit production & Trump nuke agenda delayed; councilors question efficacy: https://t.co/zNumTxUxH6
10/6/2017	41	And very active nuclear enterprise tactical & strategic with R&D on new warhead types with ~1000 pit production ann https://t.co/mEilsrqYcy
12/4/2017	41	Questions swirl about plutonium pit production at Los Alamos CLICK BELOW FOR FULL STORY... https://t.co/rqaAvnhioY
12/5/2017	41	The Latest: US: More study needed on nuclear pit production https://t.co/xQT7HByJRP
12/8/2017	41	More study needed on #nuclear pit production https://t.co/C5f5GNcgCA #NoNukes #NoWar #LosAlamos
12/8/2017	41	@WilliamJBroad I now NNSA's 9-page presentation
12/15/2017	41	Trump signs FY18 NDAA that would keep @SRSNews MOX program alive and help keep pit production at @LosAlamosNatLab. https://t.co/4xtWZczBI3
12/15/2017	41	What does the newly signed FY18 NDAA says about MOX
2/4/2018	45	Questions Swirl About Plutonium Pit Production at US Lab | New Mexico News | US News https://t.co/21ZbAty2Zf // Tot https://t.co/Q9G46z1vxA
2/5/2018	45	U.S. @RepJoeWilson: Pit production at SRS would be complementary to #MOXproject. https://t.co/iDCFcg41Fa
2/9/2018	45	Trump's nominee to lead NNSA defends pit production priority
2/17/2018	45	Trump's pro-pit production nominee to lead NNSA confirmed by U.S. Senate https://t.co/hcNqChYqxe
2/23/2018	45	US takes steps to resume plutonium pit production for nukes-Call to increase plutonium storage at New Mexico facil https://t.co/MwYHoVc9y0
2/23/2018	45	US takes steps to resume plutonium pit production for nukes | Via AJENews https://t.co/iqRIL4zQXd
3/16/2018	45	The Los Alamos Study Group also contends the risks of pit production at Los Alamos are significant and should be di https://t.co/r7mIlp16Ys
3/22/2018	45	NNSA chief Lisa Gordon-Hagerty (@LGHNNSA) tells House Armed Services subcomm. pit production locations are narrowed https://t.co/fgm0hksv4s
4/11/2018	45	CODA: And no update about the @NNSANews' progress with its pit-production decision process. @LGHNNSA is supposed to https://t.co/dKp6CDgX7I
Plutonium Disposition (Inflection at Intervals 11, 18, 20, 26, 28, 30, 42)		

5/4/2015	11	DOE: MOX No Longer Preferred Option For Plutonium Disposition | http://t.co/mQ0p0vSNLt
8/22/2015	18	13 Aug Final Report of the Plutonium Disposition Red Team Oak Ridge TN http://t.co/vbx2lZaeUn #nonukes #climate #security #policy #tech #law
8/30/2015	18	Department of Energy Oak Ridge National Laboratory Plutonium Disposition Red Team Report: The Plutonium Manage… http://t.co/hE59ye4bdV
9/10/2015	18	If youre attending @WorldNuclear #Sympo2015 you can learn more about plutonium disposition and PRISM at our booth. http://t.co/R7Vr6RqHSL
9/21/2015	18	GEHs David Powell discusses plutonium disposition & PRISM at @IMechE nuclear operations summit today in Manchester. http://t.co/xaMvd6bTpG
10/7/2015	18	LIVE NOW: Strategic Forces Subcomm - Plutonium Disposition and the MOX Project #HASC Watch Here - https://t.co/qTwltFkntc
10/13/2015	18	Russian Official: Govt Open to Non-MOX U.S. Plutonium Disposition http://t.co/FDITQH0yyM #NSDMonitor
12/8/2015	18	Release of Aerospace report on MOX plutonium disposition costs https://t.co/9nCOse0UAr raises critics' hackles https://t.co/I8aRp8vG6M
12/30/2015	18	South Carolina Gov. Haley skeptical of federal plutonium disposition plan https://t.co/VhVxQWS8iC
2/10/2016	20	Obama's budget would kill costly plutonium disposition project https://t.co/sPLT65dGwt
4/2/2016	26	US Plutonium Disposition Abides by Agreement With Russia - Security Council: The technology used b... https://t.co/9W2lTlnx6H #nieuws
4/7/2016	26	Russia raises concerns about changes in U.S. plutonium disposition plan https://t.co/3MHdya3D56
4/12/2016	26	#notonukes #manw US Plutonium Disposition Abides by Agreement With Russia Security Council https://t.co/f4TSKBCSpz
4/18/2016	26	Rayment: Also looking at reactors for plutonium disposition in UK. #ANSmeeting
4/2/2016	26	US Plutonium Disposition Abides by Agreement With Russia - Security Council https://t.co/jvmsxxBW1U
4/28/2016	26	Can the US-Russia plutonium disposition agreement be saved? by Pavel Podvig https://t.co/Ittk1PeQeD
10/3/2016	28	This is the plutonium disposition program (PMDA)
10/3/2016	28	Russia suspended the Plutonium Disposition agreement with United States https://t.co/XhIGmH0kci
10/3/2016	28	Re-upping my column on #PMDA: Can the US-Russia plutonium disposition agreement be saved? https://t.co/FHD5HMfbmc Now we know answer is NO

10/3/2016	28	Putin has suspended the agreement with the United States on plutonium disposition: The https://t.co/Rwjp6lkN3Z
10/3/2016	28	Russia Cannot Unilaterally Fulfil Plutonium Disposition Deal With US Any Longer https://t.co/0n52pxT3TK
10/3/2016	28	Russia suspends implementation of plutonium disposition agreement https://t.co/EnmVTlBgXl
10/3/2016	28	PT Remarkable list of conditions from Putin's decree on suspension of plutonium disposition agreement w US https://t.co/KHBBFpTIth
10/3/2016	28	.@realDonaldTrump promises action on cybersecurity to protect Americans from cyber-crime & national security threats https://t.co/ZVMFKHW8M7
10/18/2016	28	How The U.S. Failed In Excess Weapons Plutonium Disposition https://t.co/pGQnnv5cUl
10/18/2016	28	NuclearWatchNM: #MOX How The US Failed In Excess Weapons Plutonium Disposition... A Nuclear Sputnik Moment? | Los https://t.co/79TvOlJb89
11/23/2016	28	Lets Get a Better Deal on Plutonium Disposition - All Things Nuclear: https://t.co/RyCPQ0iAaP
5/3/2017	42	@LawDavF There's also this - Russia seems to have dropped its excessive demands on the plutonium disposition agreem https://t.co/ibNITKkB0R
5/4/2017	42	Russia wants to re-start plutonium disposition agreement with US if MOX funded https://t.co/s984oxdBSs @CherylRofer
6/30/2017	42	A #360photo inside the @XMaSBeam hutch @esrfsynchrotron: investigating ceramics for @NDAgovuk #plutonium disposition https://t.co/A5MRTH51OO
9/5/2017	42	Plutonium Disposition: Proposed Dilute and Dispose Approach Highlights Need for More Work at the Waste Isolation.. https://t.co/sdzLeRZJkW
11/15/2017	42	Plutonium Disposition: Observations on DOE and Army Corps Assessments of the Mixed Oxide Fuel Fabrication Facility.. https://t.co/mHyHA3Eo6Z
12/4/2017	42	Non-MOX plutonium disposition study by NAS gets under way
12/20/2017	42	New at https://t.co/1UvwuQ7zSC: Mixed-Oxide Fuel Fabrication Plant and Plutonium Disposition: Management and Policy Issues
2/25/2018	45	Time to Re-examine Alternatives for Plutonium Disposition - Dr. Peter Lyons explains why dilute and dispose is... https://t.co/3cezOrEztB
4/24/2018	45	What went wrong with US #plutonium disposition
4/25/2018	45	Russian MFA: main reason Russia stopped implementing the Plutonium Disposition Agreement are US unfriendly actions https://t.co/Z31inNjyY7
Plutonium Pit (Inflection at Intervals 4, 8, 15, 24, 27, 30, 32, 34, 38, 40, 43)		

12/16/2014	4	ReTw tribalidentity: RT Cirincione: "Does America really need more plutonium pits? The nuclear money pit " Thank... http://t.co/Tm1RhwOzaI
4/30/2015	8	4th amdt seeks info on why NNSA wants to make 50-80 new plutonium pits a year. Current rate 5-10. Why do we fear info?
6/3/2015	15	Incredible: a lucite hemisphere that was forged in the original mold used for the plutonium pit
1/18/2016	24	LANL poised to ramp up plutonium pit production: report https://t.co/VR7nw8j3Ug
2/11/2016	24	LANL would get $2.1 billion in proposed budget; officials talk plans for making plutonium pits https://t.co/bI3lf9bywT #abq via @abqjournal
3/20/2016	24	Reader View: Accelerated plutonium pit production wrong for #NewMexico https://t.co/3QGoBT4OPV
4/5/2016	24	History of Plutonium Pit Production https://t.co/QRgiEPYJOu
11/14/2016	24	One of the oddest things about Putin's recent decision re nuke cooperation is now the US has to keep a larger stockpile of Plutonium Pits.
1/5/2017	30	LANL on track to build two test plutonium pits fission triggers that ignite nuclear bombs within next 9 months: https://t.co/YajByQAx87
6/16/2017	34	Will production of plutonium pits used in nuclear warheads shift from New Mexico to the Savannah River Site? https://t.co/J6aMHkJfun
6/21/2017	34	#Plutonium pits at core of new Savannah River Site debate @SRSNews https://t.co/Wip45izY5p
7/2/2017	38	Amarillo's @PantexPlant could take on plutonium pit production. https://t.co/sFntSDToDJ
7/4/2017	38	DoE eyes terminating Savannah River MoX to use site for plutonium pit production & possible tritium work https://t.co/QoGRu1TNVk
7/16/2017	38	Manufacturing Plutonium Pits for Nuclear Weapons by Hardcover Book (English) https://t.co/RMaVryAi8p
9/29/2017	38	Workers at @LosAlamosNatLab caused a "criticality safety event" on Aug. 18 during production of new plutonium pits. https://t.co/T70mydkzKO
11/18/2017	40	#Defense bill on Trump's desk contains Heinrich/Udall amendment seeking fresh mandate for plutonium pit factory at https://t.co/z3ZMt8XsMb
12/6/2017	40	Leaked DOE doc outlines plutonium pit production at SRS
12/8/2017	40	@WilliamJBroad I now NNSA's 9-page presentation
12/4/2017	40	Questions swirl about plutonium pit production at Los Alamos CLICK BELOW FOR FULL STORY... https://t.co/rqaAvnhioY
1/3/2018	43	Letter: South Carolina governor should fight "plutonium pits" at already heavily contaminated Savannah River Site. https://t.co/HIj6mJRHpy

1/15/2018	43	.@ABQJournal: "Debate over #plutonium pits at LANL may be getting real" https://t.co/cwaY3Ux74J
1/19/2018	43	Per the lifespan of plutonium pits
2/8/2018	43	War and peace in the nuclear age https://t.co/kE67WUpxf3 Proposal for scientists to create detente. Thinking outside the plutonium pits.
2/12/2018	43	Political scuffle brewing between #NM and #SC over manufacture of plutonium pits for nuke weapons. From the Journal https://t.co/zhxOKLFAcq
2/14/2018	43	Producing plutonium pits too dangerous for New Mexico https://t.co/TaUA6OxrGj https://t.co/8s7OJDCQAM
2/23/2018	43	US takes steps to resume plutonium pit production for nukes- Call to increase plutonium storage at New Mexico facil https://t.co/MwYHoVc9y0
3/20/2018	43	Hearing adjourned. No one asked about moving plutonium pits to SC (briefing to Congress due May 11). No cost/schedu https://t.co/qVNrUOD2ib
3/22/2018	45	Plutonium "pits" last more than 100 years. No need yet for new ones--unless you are developing new weapons designs. https://t.co/0fUFpwRRUX
4/26/2018	45	its #thepits Plutonium pit misplaced at LANL https://t.co/1KZ1Jwo6M4

Notably, time intervals 41, 42, and 43 (in Figure 2-10) captured 66 total key terms that had inflection points in a 3-month time span. Such a change across a large list of key terms indicates the occurrence of a significant event that caused a contextual shift in the key terms in those intervals. The Twitter corpus has been searched from time interval 38 (embedding model through October of 2017) through 43 (embedding model through March of 2018) capturing 510,805 Tweets (Note: roughly $1/6^{th}$ of the total Tweets occurred in these six time windows). The list of Tweets is reduced by removing duplicate tweets (i.e., retweets) to a list of 135,112 Tweets and the total number of Tweets per time interval is shown in Table 2-2.

Table 2-2. Number of Tweets in each time interval searched.

Time Interval	Total Number of Tweets	Number of Unique Tweets
38	76,627	21,652
39	72,580	21,006
40	101,774	24,944
41	66,206	18,836
42	104,758	26,007
43	88,860	22,667

Using the aforementioned similarity metrics to compare Tweets to the key terms that had inflection points at these intervals, the list is reduced further to 22,393 tweets. Increasing the threshold similarity value would reduce the set further, as illustrated by the numbers of tweets that were captured above different similarity thresholds shown in Table 2-3. Note that once the threshold is met by a similarity metric, it is added to the list and therefore the similarity is not re-computed by the network graph. Therefore, while there is overlap in what would be discovered by any of the metrics, Table 2-3 demonstrates that the network graph approach identifies additional unique tweets compared to simply using the base embedding vector.

Table 2-3. Number of Tweets found at different levels of similarity.

Cosine Similarity	Total Number of Tweets	Found with Base Embedding	Found with Network Graph
0.75	16,164	2,141	14,023
0.80	4,869	926	3,943
0.85	966	412	554
0.90	207	202	5
0.95	187	187	0

Inspection of the Tweet corpus reveals that this time period contained a significant number of Tweets surrounding national security and various investigations occurring within high levels of the United States Government at the time. This high influx of Tweet traffic is the primary cause for the inflection points that occurred at this time as a significant shift in a perhaps a small subset of the terms can cascade through the remaining terms. Therefore, the Tweets were further filtered, removing any Tweet that contained the term "national_security". Table 2-4 shows the number of Tweets that remained and that were discovered with the network graphs versus the base embedding vector of the key term.

Table 2-4. Number of Tweets found at different levels of similarity when Tweets containing the term "national_security" are removed.

Cosine Similarity	Total Number of Tweets	Found with Base Embedding	Found with Network Graph
0.75	2,389	1,189	1,200
0.80	722	693	29
0.85	317	316	1
0.90	171	171	0
0.95	121	121	0

In addition to the events related specifically to pit production (e.g., such as the sample in Table 2-1), the pipeline captures events broadly related to new technology creation, domestic and international weapons testing, political and diplomatic decisions/opinions/announcements, contamination events in the DOE Complex, personal opinions and/or speculations related to nuclear weapons/materials in the US and around the world, government agency/lab job postings, DOE funding announcements/contract awards, among many others, were discovered in the inflection point windows.

2.2 Performance Assessment of Twitter Modeling Pipeline – Sensitivity to a Reduced Dataset

Perhaps the most important step in the modeling pipeline is the collection of data using the glossary of key terms as the embedding models rely on the presence of sufficient domain specific data to capture contextual shifts. The glossary of key terms is made up of words/phrases that relate to specific activities and events, as well as names of entities (e.g., national laboratories, government agencies, etc.). In the FY20 preliminary modeling efforts, a modified dataset was created that removed any Tweets that contained only an entity key term to emulate a query where entity names were unknown. Ultimately, this reduced the number of Tweets in the corpus by about one-half. Notably, the connection between pit production and the SRS was still made prior to the official announcement using the metric shown in Figure 1-5.

To explore the impact on the inflection windows when entity keywords are removed from the query, the analysis that was described in Section 2.1 was performed on the reduced corpus. The inflection point windows matrix is shown in Figure 2-11. Comparing this matrix to Figure 2-10 shows that the matrices are qualitatively similar, demonstrating that even without explicitly querying for Tweets containing entity

keywords, a sufficient contextual representation can still be obtained that captures important intervals around similar points in time. Note that the vast majority of Tweets that were captured in Table 2-1 contained the co-occurrence of the non-entity key term and an entity. The analysis demonstrates that the activity and event keywords broadly capture the activities of interest that are performed by the entities in the keyword list. Therefore, the exclusion of entities may in fact reduce some "noise" (i.e., unrelated activities performed by entities) while still capturing activities and events of highest interest due to the inclusion of the proper non-entity key terms. However, the inclusion of the entity key terms allows a contextual representation to be formed earlier for each entity, given sufficient Tweet data returned from the query. For example, if the query includes Tweets that contain the term "Savannah River Site", the embedding models can begin to build a contextual representation around activities pertaining nuclear materials processing. Therefore, having that contextual representation in place, contextual similarity between the known activities at the SRS and an activity like fissile core fabrication may be inferred at an earlier time because of the similarity in the activities. Thus, having all entities is not a necessity, but is presumed to help with earlier detection.

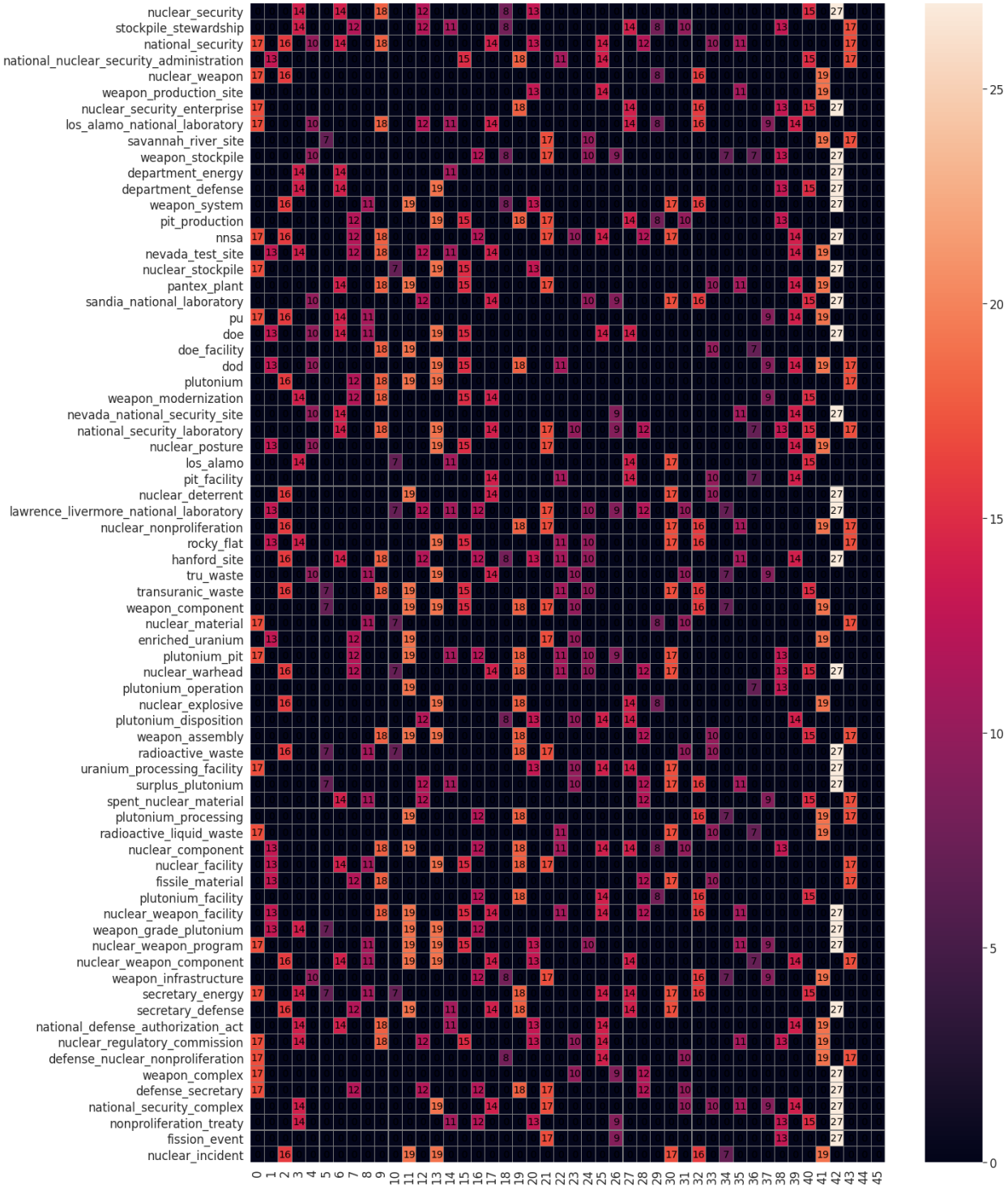


Figure 2-11. Matrix showing the time periods where inflection points were detected for the reduced corpus. Annotated numbers indicate the number of key terms with an inflection point in that time interval.

3.0 News Article Aggregator Data Source – Webhose Ltd.

The FY20 preliminary modeling efforts developed a blueprint for the modeling pipeline operating on the news aggregator data source. Once again, the pipeline relies on the use of time dependent word embedding models trained on the corpus. As discussed in Section 1.1.3, a more extensive list of key terms was used to query the Webhose data source, where each key term was labeled based on the event domain with which the term is most closely associated. Therefore, in the following sections, each event domain has been treated as a separate corpus containing documents returned by the query of the terms falling within that domain. The modeling pipeline can be summarized by the following steps:

1. Create Word Embeddings: Transfer articles and phrases into a common space to enable distance calculations between vectorized textual data and enable inference of the content of each article and phrase.
2. Search/Document Ranking: Identify the most similar documents with respect to specific phrases, i.e., ranking the most relevant documents.
3. Entity Extraction: Extract entities of interest from each document (i.e., from Step 2) in order to identify more apparent themes and topics. (Note that in the extraction of entities, an entity disambiguation approach was applied for key terms to attempt to decrease the overall number of entities by equating different variations of the same entity – e.g., "SRS" == "Savannah River Site".)
4. Anomaly Detection: Create heterogeneous weighted temporal entity graphs to evaluate the temporal evolution and explore the changing relationships between entities.
5. Event Extraction: Extract events from articles and text based on anomalous entities (and their connections) from step 4.

Steps 1 through 3 are described in detail in Section 4 of reference 2 and provide the means to explore the dynamic relationship of entities within the corpus such that the potential occurrence of events of interest can be identified. Steps 4 and 5 have continued in development throughout FY21 and will be described in the following sections to characterize the capability and performance of the pipeline.

3.1 Entity Characterization for Anomaly Detection and Event Extraction

To characterize the current state and evolution of entities of interest, a temporal graph representation has been developed that links co-occurring entities (i.e., entities occurring in the same document) with a weighted edge. The weight of each edge is governed by the number of times a pair of entities co-occur in a specified time period. The weighted entity graph G is constructed such that each graph in the time interval G^t considers articles published only from time $t - w$ to t, where w is the user-specified length of the time interval. Here, the entity co-occurrence sub-graphs are constructed using the top 50 co-occurring entities within a time interval of three months. This ego-network based entity representation focuses explicitly on key entities of interest to yield a multi-faceted view in terms of the "GPE" (i.e., geopolitical entities), "FAC" (i.e., facilities), "NORP" (i.e., nationalities/religious/political groups), and "ORG" (i.e., organizations) that are in the neighborhood (and hence highly related) to the entity of interest at the selected time slot.

3.1.1 Anomaly Detection Using Similarity Metrics Over Time

Anomaly detection is executed on the ego-network of a corpus of interest (i.e., documents returned by key terms labeled as a specific event type) for different time periods. The first approach that was developed compares ego networks of a specific entity of interest based on well-defined similarity metrics over time to identify how similar (or dissimilar) the networks are in one time period relative to another. The similarity metrics include the commonly used cosine similarity and Jaccard similarity, as well as customized metrics such as edge similarity. In general, the similarity metrics use edges and their corresponding weights in two networks (the same entity of interest in two different time periods) to obtain a single overall similarity measure between the two networks (i.e., time periods). Note that when a node exists in one time period but not another, the node must be created and given a weight of zero in the network it doesn't already appear. Detailed explanations of the customized similarity metrics used are the following:

- **Normalized Jaccard Similarity:** Jaccard similarity divided by total number of unique edges in both networks (i.e., union of edges). Therefore, the metric is always between 0 and 1. The higher the normalized Jaccard similarity, the more similar the two networks are to each other.
- **Weighted Jaccard Similarity:** Jaccard similarity and normalized Jaccard similarity both increase only when an identical edge is detected (the edge and its corresponding weight). Therefore, they do not capture common edges with different weights in two networks. The weighted Jaccard similarity, however, evaluates common edges and their weights even when the weights are different. It compares all pairs of edges in two networks, sums up the minimum of the weights in each pair, and divides by the sum of the maximum weights in each pair. In doing so, the similarity captures both common edges and weights such that it increases when weights of a common edge are close to each other. By definition, it is between 0 and 1.
- **Edge Similarity:** Edge similarity is the number of common edges between two networks (regardless of the weights of those common edges). The more common edges between two networks the higher the edge similarity.
- **Normalized Edge Similarity:** The number of common edges between two networks (regardless of the weights of those common edges) divided by total number of unique edges in both networks (i.e., union of edges). By definition, it can be between 0 and 1. The more common edges between two networks the higher the normalized edge similarity.

The similarity metrics are calculated between a network at time t_n relative to four other networks: t_{n-1} (network for the previous time period), t_{n-2}, t_{n-5}, and t_0 (the first network in the time series) for 21 consecutive time periods. The similarity measures for "savannah river" (i.e., capturing all instances of SRS and SRNL) are shown in Figure 3-1.

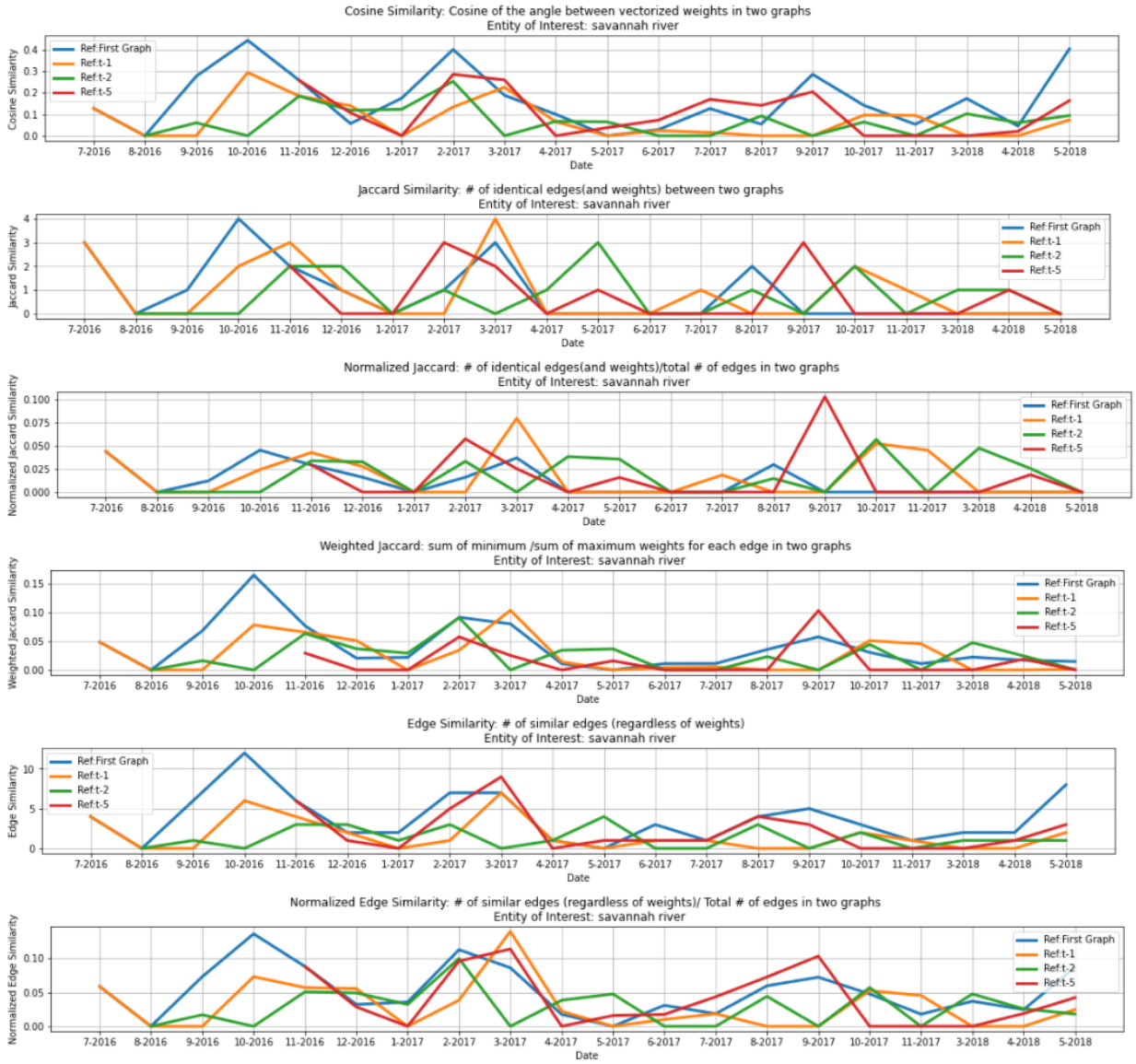


Figure 3-1: Similarity metrics for "savannah river" computed from analysis of documents returned by seed phrases classified as "political/diplomatic events".

Similar to the analysis performed on the Twitter dataset, the temporal profile of the similarity metrics can allow detection of significant changes in the overall ego-network (edges and weights) from one time-period to another. However, because of the high dimensionality (at least 50), low co-occurrences, and common yet regularly accruing changes, these similarity measures often tend to be low, indicating significant change. It is therefore difficult to use these metrics alone to identify an anomalous network or time interval for a specific entity. However, it is important to note that each similarity measure is an aggregate measure of many changes in a network. Thus, even when an anomaly is detected, specific edges that cause such an anomaly may not be identified using these aggregate metrics.

3.1.2 Anomaly Detection by Comparing Entity Co-Occurrence Graph Edges Over Time

To overcome the challenges associated with the similarity metrics, an additional approach was developed to compare the networks and identify anomalies. In this approach, like edges (i.e., entity co-occurrences) in two consecutive networks are compared. Again, if a node exists in one network, but not another, it is

created and assigned a weight of 0 in the network in which it does not appear. The difference between the weights of edges in each network (*weight_dif*) is calculated. If a node has more co-occurrences in the most recent time period, it has a negative weight difference and if it has fewer, it has a positive weight difference. Each weight difference is standardized (standard Z-scores) by subtracting the population mean of weight differences and then dividing by the standard deviation. In other words, the weight differences are transformed into a standard normal distribution with a standard deviation of 1. Therefore, each of these scores (*z_weight_diff*) show how many standard deviations a weight difference is away from the mean of the population weight difference. When a weight difference is far from the mean it is less likely to be an expected change (in other words, they are outliers or anomalies). Anomalies (or outliers) between two networks are determined when the weight difference for an edge is greater than a critical Z-score (2 to capture outliers that are different from 95% of the population, 3 to capture outliers that are different than 99% of the population). Outlier edges for Los Alamos National Laboratory are shown in Figure 3-2.

	node_from	node_to	weight_1	weight_2	weight_dif	Z_weight_dif	date_from	date_to
46	los alamos national laboratory	democrat	17.0	0.0	17.0	2.953498	6-2016	7-2016
65	los alamos national laboratory	california	0.0	11.0	-11.0	-2.271891	7-2016	8-2016
22	los alamos national laboratory	centrifuge hecker	0.0	37.0	-37.0	-2.585451	8-2016	9-2016
27	los alamos national laboratory	united states	8.0	42.0	-34.0	-2.347798	8-2016	9-2016
29	los alamos national laboratory	middlebury institute of international studies	0.0	38.0	-38.0	-2.664668	8-2016	9-2016

Figure 3-2: Anomalous or outlier edges for "los alamos national lab" computed from analysis of documents returned by seed phrases classified as "political/diplomatic events".

3.1.3 Event Extraction

The event extraction pipeline is used to extract events from articles based on anomalous edges, where each article has been encoded with a list of all entities that occur within (Figure 3-3).

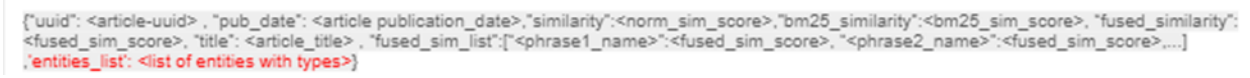


Figure 3-3: Article encoding with entities shown in red.

After encoding each article with a list of entities, a network graph with anomalous edges from a specific time interval is input to the event extraction pipeline from the anomaly detection module. Because each ego-network represents co-occurrences of entities in the articles of a specific time interval, events are only extracted from the articles that appeared in the time periods that were used to identify the anomalous edges. Notably, the graph may identify edges between the entity of interest and multiple adjacent nodes if more than one edge was anomalous during that time period. Subsequently, the event extraction process searches each article encoding to identify if the anomalous node pair exists. If the node pair is found, the edge is added to an edge list that is also encoded into the article as shown in Figure 3-4. The pipeline will then move on to the next article and repeat this process for each article in the specified time interval.

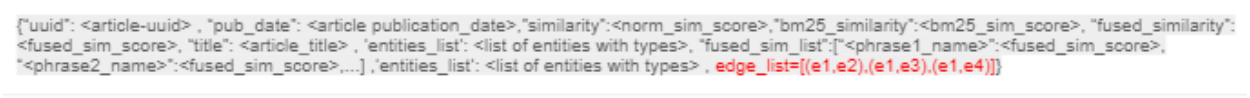


Figure 3-4: Article encoding with anomalous edge list (in red)

Once each article has an edge list, the sentences within the article that contain the edge are extracted into a list of sentences to highlight potential events. The sentence is extracted if both nodes of the edge are present and if the number of words between the nodes is below a certain parameter, *word_dist*. This process is repeated for each of the articles that contain the edge. The output (as seen in Figure 3-5) of the event extraction provides the article ID, publication date, title, edge entities, extracted text and full text. This

allows a reviewer or analyst to search the extracted events by time and/or entities and then review the extracted text, if available, for insight into the correlation between the two edges. If more detail is needed or an event is not extracted, then the full text can be reviewed.

uuid	pub_date	title	adjacent_node	extracted_text	article text
e0c13306	2018-05-01:02:56	(USA-SC-Aiken) Executive Assistant/Coor	['south carolina']	[('savannah river', 'america', ['aecom executive assistan	aecom executive assistant coordinator in aiken south carolina business line goverr
cb195007	2018-05-01:02:56	(USA-SC-Aiken) First Line Manager, Radio	['south carolina']	[('savannah river', 'america', ['aecom first line manager	aecom first line manager radiological control in aiken south carolina business line
f0abfe946	2018-05-01:02:56	(USA-SC-Aiken) Manager, Construction Er	['south carolina']	[('savannah river', 'america', ['aecom manager construct	aecom manager construction engineering in aiken south carolina business line gov
4825b612	2018-05-01:02:56	(USA-SC-Aiken) Manager, Construction O	['south carolina']	[('savannah river', 'america', ['aecom manager construct	aecom manager construction operations site b in aiken south carolina business line
5afeb36cb	2018-05-01:19:01	Key sites proposed for nuclear bomb proc	['new mexico', 'south carolina', 'los ala	[('savannah river', 'new mexico', ['some experts are wo	the department of energy is scheduled to decide within days where plutonium pa
e440a7c02	2018-05-02:03:00	wzzm13.com | Safety concerns plague ke	['new mexico', 'south carolina', 'los ala	[('savannah river', 'new mexico', ['some experts are wo	photo jeff blake for usa today safety concerns plague key sites proposed for nuclea
ea1b12fb2	2018-05-03:02:21	Safety concerns plague key sites propose	['south carolina']	[('savannah river', 'south carolina', ['to find out more ab	let friends in your social network know what you are reading about facebook emai
5265cc4df	2018-05-05:02:55	(USA-SC-Aiken) Work Control 2018 Stude	['south carolina']	[('savannah river', 'america', ['aecom work control stude	aecom work control student summer intern lse in aiken south carolina business lin
9dfc7582c	2018-05-07:18:30	The US is set to ramp up nuclear warhead	['south carolina']	[('savannah river', 'south carolina', ['a worker at the sav	a worker at the savannah river site in south carolina uses a glovebox to handle haz
039d2a55	2018-05-08:14:53	Talk to Us About Los Alamos National Lab	['south carolina']	[('savannah river', 'south carolina', ['and albany oregon	propublica and the santa fe new mexican are investigating health and safety condi
b4ce53da	2018-05-09:05:45	Decision on site for plutonium pit produc	['department of energy', 'south carolir	[('savannah river', 'department of energy', ['a similar sa	decision on site for plutonium pit production expected by friday by patrick malone
6fafa6eea	2018-05-10:00:31	US to decide best site option for nuclear v	['south carolina']	[('savannah river', 'south carolina', ['the other option w	us to decide best site option for nuclear weapons production los alamos national l
d977b40a	2018-05-10:01:02	Feds to announce best site option for nuc	['south carolina']	[('savannah river', 'america', ['the other site under cons	by karen graham hours ago in politics washington - the federal agency that oversee
3ce575c19	2018-05-10:02:56	(USA-SC-Aiken) Limited Service Associate	['south carolina']	[('savannah river', 'america', ['aecom limited service ass	aecom limited service associate design specialist - electrical in aiken south carolin
2c86eaa17	2018-05-10:03:00	NNSA announces decision on pit producti	['nnsa', 'new mexico', 'national nuclea	[('savannah river', 'nnsa', ['to achieve dod s pits per yea	los alamos national laboratory will share production of plutonium pits with the sav
dcd49ebe	2018-05-10:03:00	US plans to split work for producing nucle	['center for public integrity', 'henry mc	[('savannah river', 'center for public integrity', ['interna	albuquerque n.m. ap the federal agency that oversees the nation s nuclear weapo
17a0d121	2018-05-10:03:21	USC Aiken Celebrates Class of 2018	['south carolina']	[('savannah river', 'south carolina', ['charles munns a sp	usc aiken celebrates class of related media no description given. no description giv
1fe9a7ae0	2018-05-10:08:17	ALBUQUERQUE, N.M. (AP) \u2014 The fed	['center for public integrity']	[('savannah river', 'center for public integrity', ['n interr	albuquerque n.m. ap u the federal agency that oversees the nation s nuclear weap
702c4baa3	2018-05-11:00:18	Joint Statement from Ellen M. Lord and Li	['nnsa', 'new mexico', 'south carolina',	[('savannah river', 'nnsa', ['to achieve dod s pits per yea	joint statement from ellen m. lord and lisa e. gordon-hagerty on recapitalization o
70eeffb3c	2018-05-11:00:37	Breaking: Feds decide to split production	['henry mcmaster', 'department of ene	[('savannah river', 'henry mcmaster', ['henry mcmaster	copyright albuquerque journal washington d.c. the majority of the nation s produc
00f149aab	2018-05-11:03:16	LANL Officials Welcome Plutonium Pit Pr	['nnsa', 'south carolina']	[('savannah river', 'nnsa', ['to achieve dod s pits per yea	lanl officials welcome plutonium pit production announcement submitted by caro
3e67924b	2018-05-12:05:58	At least 50 plutonium pits will be used to	['national nuclear security administrat	[('savannah river', 'national nuclear security administra	savannah river site sc wfxg - the savannah river site will be producing plutonium pi
a99a81fd5	2018-05-13:01:51	Nuke agency pitches plutonium pits for S	['national nuclear security administrat	[('savannah river', 'national nuclear security administra	home national news nuke agency pitches plutonium nuke agency pitches plutoniu
90c8f5881	2018-05-13:09:02	New â€˜pitâ€™ plan may mean more was	['new mexico', 'south carolina', 'los ala	[('savannah river', 'new mexico', ['advertisement now t	copyright albuquerque journal santa fe the feds can taketh plutonium away even a
97327d7f1	2018-05-13:19:56	Trump Administration Axes Project To Ge	['department of energy', 'south carolir	[('savannah river', 'department of energy', ['news alerts	news alerts trump administration axes project to generate power from plutonium
b663bb89	2018-05-14:03:00	NNSAâ€™s pit decision restores confiden	['national nuclear security administrat	[('savannah river', 'national nuclear security administra	nnsa s pit decision restores confidence in local economy local and laboratory offici
776f3f96e	2018-05-14:16:43	Made in the USA: Department of Energy l	['nnsa', 'los alamos national laboratory	[('savannah river', 'nnsa', ['nnsa also provides funding f	made in the usa department of energy labs help advance technology to ensure sup
d7f8733fb	2018-05-14:17:27	Construction halted on MOX nuclear facil	['south carolina']	[('savannah river', 'south carolina', ['perry executed a w	home news news briefs construction halted on mox nuclear facility construction h
ddcdf30ac	2018-05-15:03:00	DOE Announces $72 Million to Advance H	['new mexico', 'south carolina']	[('savannah river', 'new mexico', ['million oak ridge nati	doe announces million to advance high-temperature concentrating solar power sy
b1baa1d4	2018-05-15:15:50	Billions of dollars later, Energy Departme	['congress', 'nnsa', 'new mexico', 'sout	[('savannah river', 'congress', ['the south carolina lawm	us doe has cancelled the always-questionable plan to build a mox fuel fabrication
434de2aa	2018-05-15:18:35	Correction: Plutonium Pits-SRS story	['national nuclear security administrat	[('savannah river', 'national nuclear security administra	comment off may. am edt aiken s.c. ap in a story may about the proposed producti
528e5b75	2018-05-15:21:54	Department of Energy Announces $72 Mil	['new mexico', 'south carolina']	[('savannah river', 'new mexico', ['million oak ridge nati	washington d.c. realestaterama the u.s. department of energy doe announced mill
62a3a2f80	2018-05-16:00:01	Department of Energy Announces $72 Mil	['new mexico']	[('savannah river', 'new mexico', ['million oak ridge nati	department of energy announces million to advance high-temperature concentrat
01b5e2da	2018-05-16:07:54	ORNL Collaboration Results in 3D Printed	['los alamos national laboratory']	[('savannah river', 'los alamos national laboratory', ['to	ornl collaboration results in d printed stainless steel target for medical isotope ste
46b0caad9	2018-05-16:15:41	New Standard for Toxic Dust Exposure Tal	['south carolina']	[('savannah river', 'south carolina', ['the agency has rec	read more of this story here from dcreport.org by david crook. water-downed safe
1ab0df9ff	2018-05-16:22:16	DOE Seeks to Halt MOX Construction, Con	['south carolina']	[('savannah river', 'south carolina', ['photo courtesy of h	doe seeks to halt mox construction convert facility to nuclear weapons plant the tr
0351a6f63	2018-05-17:00:45	House NNSA Budget Would Fund Low-Yie	['department of energy', 'new mexico	[('savannah river', 'department of energy', ['dilute-and-	house nnsa budget would fund low-yield warhead but not pit production in s.c. ho
16ac3112c	2018-05-17:23:13	R&D S&E, Mechanical Engineere (Early/M	['new mexico']	[('savannah river', 'new mexico', ['this position requires	job detail for r d s e mechanical engineere early mid-career job location sandia nat
812b6528	2018-05-18:06:00	Radioactivity in the Heart	['washington', 'south carolina', 'georgi	[('savannah river', 'washington', ['www.washingtonpos	the government s response to radioactivity in the heartland of america shows how
5ea43c1b3	2018-05-18:07:58	(USA-SC-Aiken) General Engineer	['department of energy']	[('savannah river', 'department of energy', ['specialized	you must be a united states citizen. this employer participates in the e-verify prog
452c80923	2018-05-22:10:34	America expands is nuclear arsenal as it d	['south carolina']	[('savannah river', 'south carolina', ['on thursday evenin	pages on nuclear issues america expands is nuclear arsenal as it demands that iran
abbc7f261	2018-05-23:19:55	A win for Allendale students through STE	['south carolina']	[('savannah river', 'south carolina', ['in the allendale tea	a win for allendale students through stem coalition challenge special to the t d hrs
88c6d6a7c	2018-05-25:11:00	(USA-SC-Aiken) Payroll Accountant	['america', 'south carolina']	[('savannah river', 'america', ['aecom payroll accountant	aecom payroll accountant in aiken south carolina business line government united
92e5eef4a	2018-05-26:00:14	Alan Wilson Files Lawsuit To Block MOX S	['south carolina']	[('savannah river', 'south carolina', ['mox program at the	tweet south carolina attorney general alan wilson has filed a lawsuit aimed at bloc
65c90021a	2018-05-27:01:13	S Carolina sues feds over end of nuclear f	['henry mcmaster']	[('savannah river', 'henry mcmaster', ['henry mcmaster	print by aiken s.c. ap - south carolina is suing the federal government after the ene
a3199ee6	2018-05-29:08:29	Radiation Hot Spot: New Mexicoâ€™s Nu	['department of energy', 'new mexico	[('savannah river', 'department of energy', ['this treache	browse home opinion radiation hot spot new mexico s nuclear graveyard radiation

Figure 3-5: Sample of event extraction output for the graph of "savannah river" at May of 2018.

3.2 Performance Assessment of Webhose.io Modeling Pipeline

The objective of the anomaly detection module is to identify anomalous entities at a specific time period for further analysis by the event detection module. The similarity metrics are calculated to find time periods where there is a significant change in the ego-network. As discussed in Section 3.1, analysis of the similarity metrics for several entities of interest using different seed phrase lists shows that low similarity scores through our entire time period are typical. In general, there are only occasional peaks of higher similarity scores, rendering it difficult to identify times of higher change because the model is detecting constant change with occasional periods of stagnation. Using these metrics, it actually appears anomalous to have a higher similarity score. See the similarity metrics for "savannah river", and "sandia national laboratory" (all using general terms) in Figure 3-6 and Figure 3-7, respectively, as well as Figure 3-1, presented earlier.

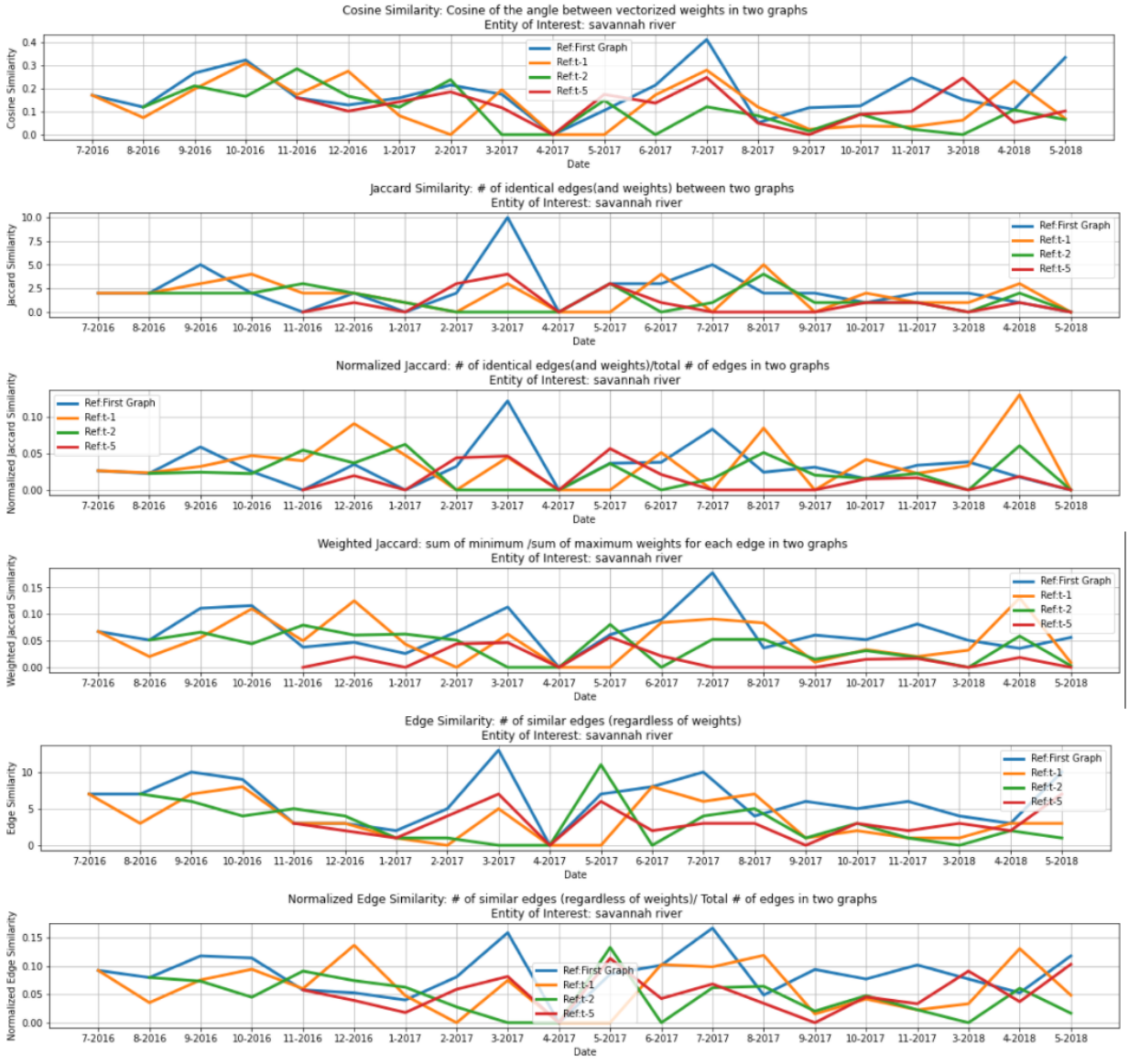


Figure 3-6: Similarity metrics for "savannah river" computed from analysis of documents returned by seed phrases classified as general seed phrases.

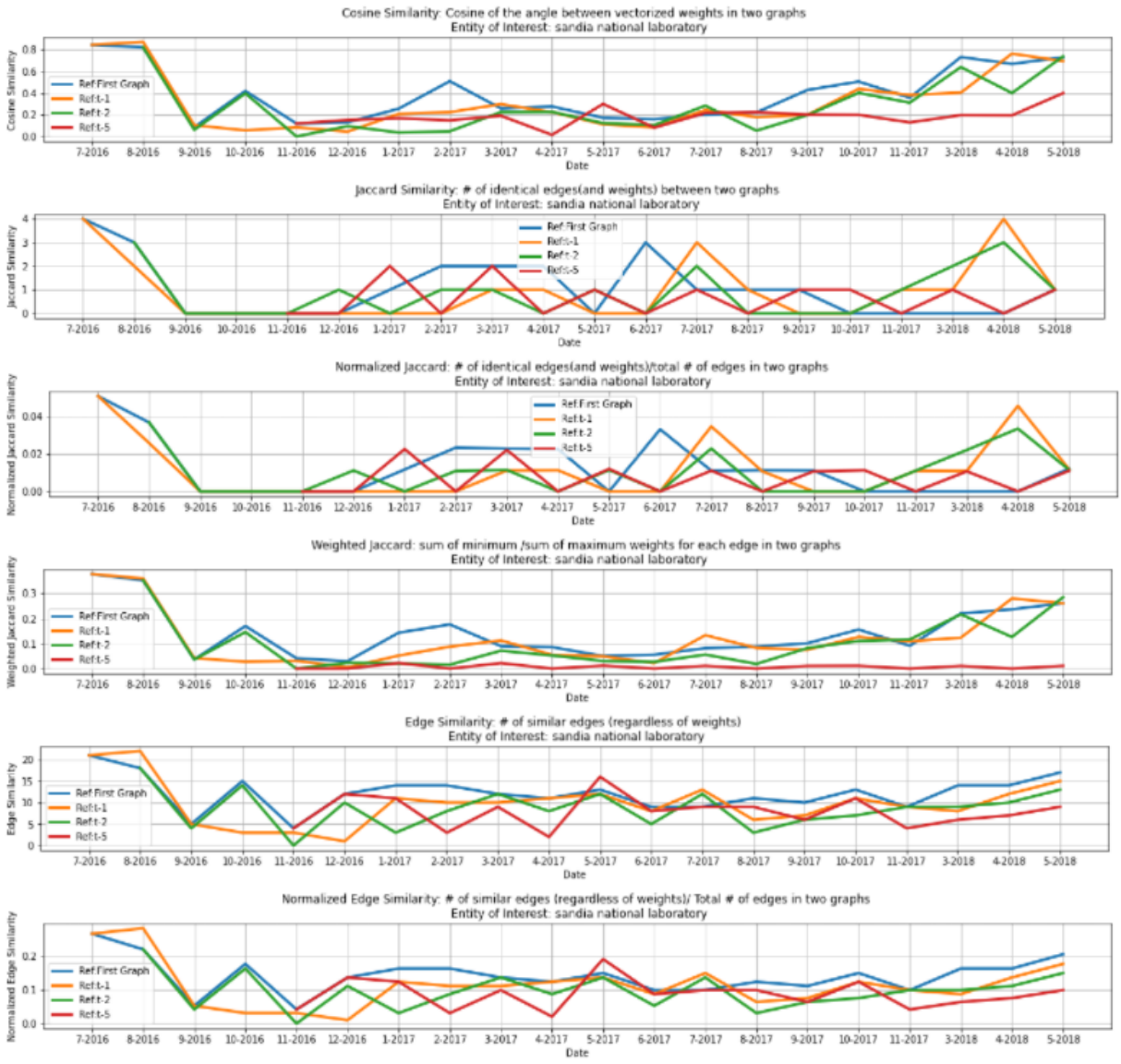


Figure 3-7: Similarity measures for "sandia national lab" computed from analysis of documents returned by seed phrases classified as general seed phrases.

It was identified that one of the contributing factors to this consistently low similarity score was that the number of co-occurrences is not higher. Therefore, the pipeline parameters were modified to allow more articles (from 30,000 to 50,000) into the entity extraction process (i.e., using the ranking algorithm), which could allow for more co-occurrences. This change did not greatly impact the co-occurrences in the ego-networks or the similarity metrics. Additionally, the cap of 50 nodes per ego-network was removed. Again, this change did not make any noteworthy difference for entities of interest like "savannah river", which only had four ego-networks exceed the 50-node cap. Alternatively, this did significantly increase the network size of other entities, such as "sandia national laboratory" and "los alamos national laboratory", where the average number of nodes per ego-network were 219 and 553, respectively. This increased network size correlated to higher similarity measures as shown in Figure 3-8. However, the temporal profile remains qualitatively consistent, with occasional peaks instead of dips.

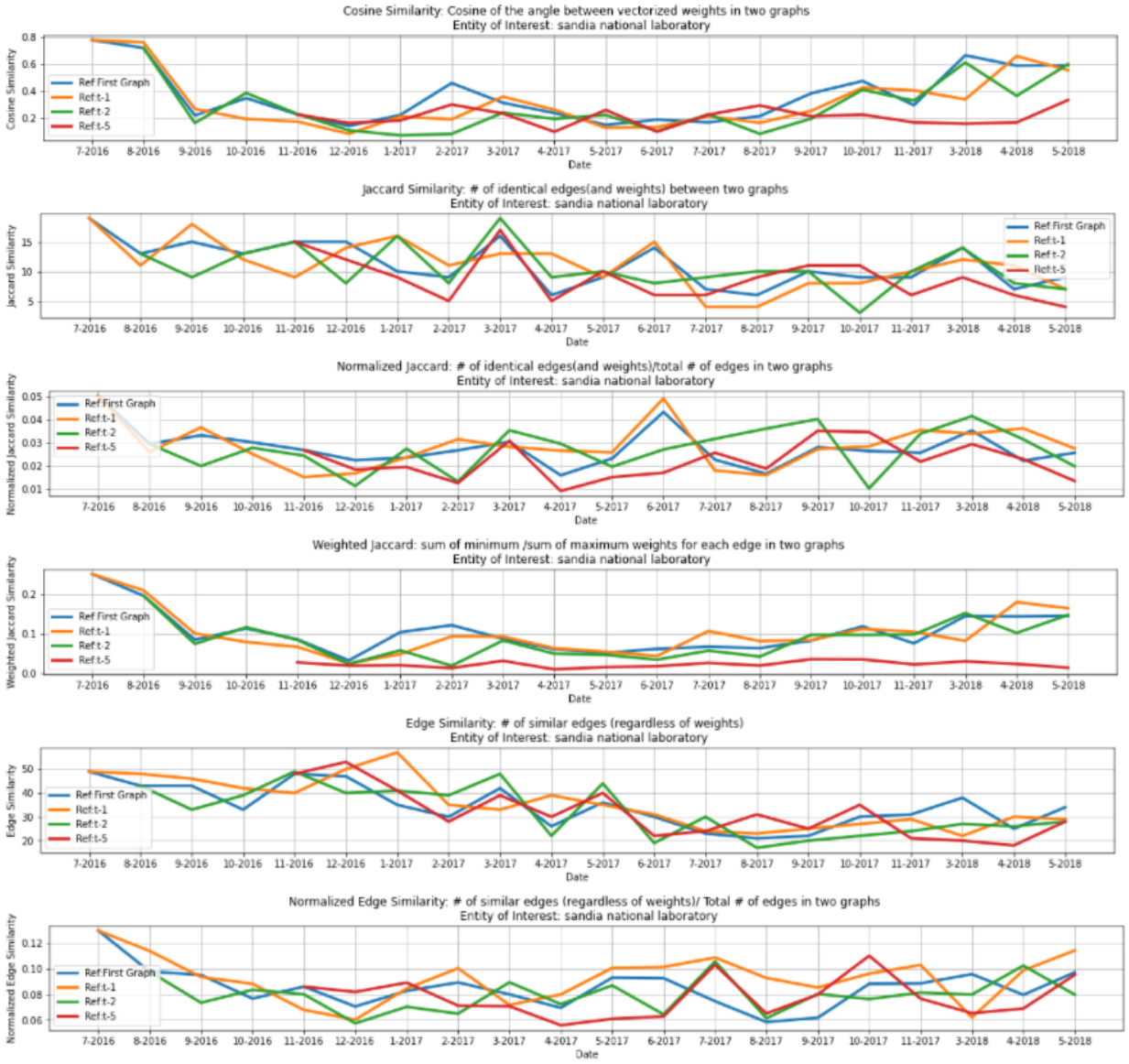


Figure 3-8: Similarity measures for "sandia national lab" using general seed phrases with increased articles and uncapped ego-networks.

As shown by comparing Figure 3-1 and Figure 3-6, different seed phrase lists were analyzed to explore the impact of the analysis on different sets of articles. This sometimes resulted in a significant change in the average size of the networks. The *general* seed phrase list typically resulted in the largest ego-networks and could decrease by as much as half for the other seed phrase lists (see Figure 3-9). Comparing these results across three seed phrase lists, it is observed that the general trend of the similarity metrics does not change. This, combined with the attempt to increase the number of documents allowed for creating the networks, indicates that the technique is not a result having too few documents, but rather the metrics as applied to the graphs. Therefore, to increase the capability of applying the similarity metrics for anomaly detection, it is recommended that the ego-networks be treated as weighted average embeddings, similar to the approach that was applied on the Twitter dataset in Section 2.1.

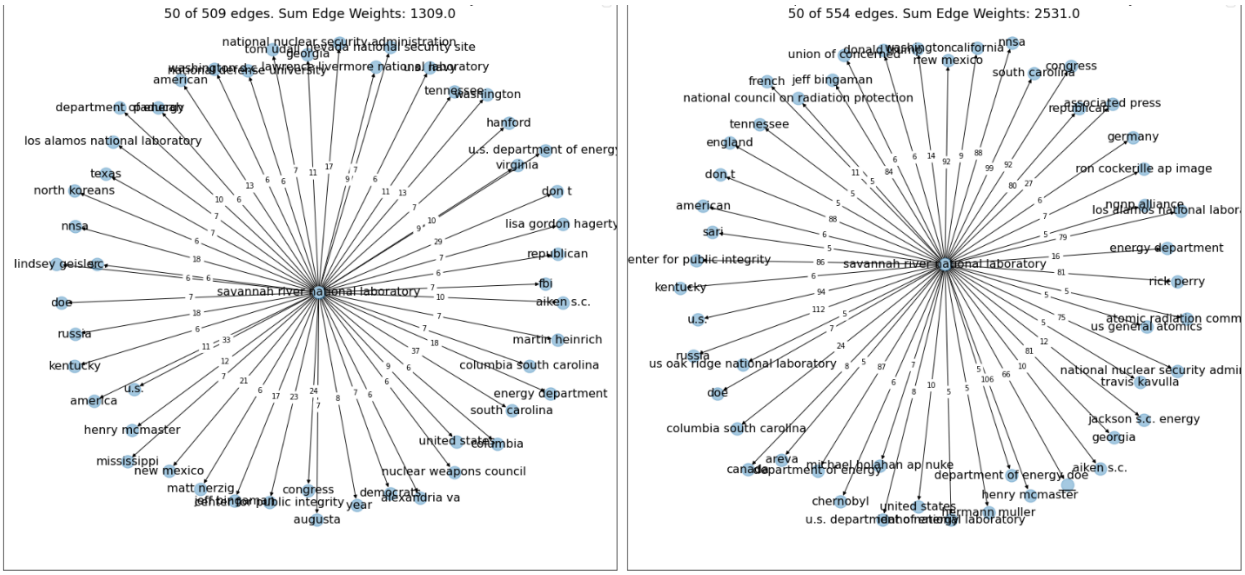


Figure 3-9: Cumulative co-occurrences over the entire time period for "savannah river" using general seed phrases (left) and political seed phrases (right).

Because it was not feasible to identify anomalous time periods using the entire ego-networks and similarity measures, it became necessary to rely on the process of identifying anomalous edges. The advantage to this approach is that both the entity of interest and the time period are identified, therefore providing a more targeted search for events. Thus, the results of this approach can be fed directly into the event extraction module for more detailed results and analysis.

The full list of anomalous edges for "savannah river" using the general seed phrases with uncapped ego-networks is shown in Figure 3-10. Exploring this list reveals the anomalous edges capture potential events of interest between the SRS and other DOE sites (e.g., Paducah, Piketon, Sandia, Nevada National Security Site, etc.). Likewise, many of these entities are connected to "savannah river" and "pit production" (e.g., New Mexico, Lisa Gordon Haggerty, NNSA, etc.). Notably, over one-third of the anomalous edges in Figure 3-10 are during the time period of April 2018 to May 2018 which is just before the official announcement of pit production at Savannah River Site was made.

Parameter tuning is a key aspect of identifying anomalous edges. The z-score used to determine if an edge is anomalous needs to be indicated. As can be seen in Figure 3-10 only 24 of the 70 anomalous edges have a z-score above 3.0 or below -3.0. If the z-score parameter of the anomaly detection module was changed to 3.0 from 2.0, 65% fewer anomalies would be captured. Conversely, if the z-score metric was reduced to 1.5 or 1.0 it would significantly increase the number of anomalies identified. When limiting the ego-networks to 50 nodes a z-score of 2.0 tends to provide a reasonable number of edges to explore. If the average size of the ego-networks increased or decreased it was necessary to increase or decrease, respectively, the z-score parameter to identify a reasonable number of anomalous edges.

date_from	date_to	node_to	weight_1	weight_2	weight_dif	n_weight_dif
Jun-16	Jul-16	national nuclear security administration	4	0	4	2.117141663
Jun-16	Jul-16	new mexico	5	0	5	2.619110305
Jul-16	Aug-16	year	0	6	-6	-2.902544753
Jul-16	Aug-16	new orleans	0	5	-5	-2.400576112
Jul-16	Aug-16	america	1	7	-6	-2.902544753
Jul-16	Aug-16	kentucky	0	5	-5	-2.400576112
Jul-16	Aug-16	piketon ohio	0	5	-5	-2.400576112
Jul-16	Aug-16	u.s.	1	7	-6	-2.902544753
Jul-16	Aug-16	paducah	0	5	-5	-2.400576112
Jul-16	Aug-16	mississippi	0	6	-6	-2.902544753
Jul-16	Aug-16	virginia	0	6	-6	-2.902544753
Jul-16	Aug-16	hanford	0	6	-6	-2.902544753
Aug-16	Sep-16	year	6	0	6	3.121078946
Aug-16	Sep-16	piketon ohio	5	0	5	2.619110305
Aug-16	Sep-16	new orleans	5	0	5	2.619110305
Aug-16	Sep-16	america	7	0	7	3.623047588
Aug-16	Sep-16	kentucky	5	1	4	2.117141663
Aug-16	Sep-16	u.s.	7	3	4	2.117141663
Aug-16	Sep-16	paducah	5	0	5	2.619110305
Aug-16	Sep-16	mississippi	6	0	6	3.121078946
Aug-16	Sep-16	virginia	6	0	6	3.121078946
Aug-16	Sep-16	hanford	6	0	6	3.121078946
Feb-17	Mar-17	aiken s.c.	4	0	4	2.117141663
Feb-17	Mar-17	foot long	4	0	4	2.117141663
Feb-17	Mar-17	pigs	4	0	4	2.117141663
Feb-17	Mar-17	florida	4	0	4	2.117141663
Feb-17	Mar-17	louisiana	4	0	4	2.117141663
Feb-17	Mar-17	javelina	4	0	4	2.117141663
Feb-17	Mar-17	u.s. department of agriculture s	4	0	4	2.117141663
Feb-17	Mar-17	boar	4	0	4	2.117141663
Feb-17	Mar-17	eurasian	4	0	4	2.117141663
Mar-17	Apr-17	bit.ly	4	0	4	2.117141663
Jul-17	Aug-17	washington d.c.	0	5	-5	-2.400576112
Jul-17	Aug-17	nevada national security site	0	7	-7	-3.404513395
Jul-17	Aug-17	sandia national laboratories	0	5	-5	-2.400576112
Jul-17	Aug-17	internal energy department	0	6	-6	-2.902544753
Aug-17	Sep-17	alexandria va	0	6	-6	-2.902544753
Aug-17	Sep-17	washington d.c.	5	0	5	2.619110305
Aug-17	Sep-17	nevada national security site	7	0	7	3.623047588
Aug-17	Sep-17	sandia national laboratories	5	0	5	2.619110305
Aug-17	Sep-17	internal energy department	6	0	6	3.121078946
Sep-17	Oct-17	alexandria va	6	0	6	3.121078946
Sep-17	Oct-17	businesswire.com	4	0	4	2.117141663
Apr-18	May-18	tom udall	0	7	-7	-3.404513395
Apr-18	May-18	democrats	0	7	-7	-3.404513395
Apr-18	May-18	russia	0	8	-8	-3.906482037
Apr-18	May-18	los alamos national laboratory	0	6	-6	-2.902544753
Apr-18	May-18	nuclear weapons council	0	6	-6	-2.902544753
Apr-18	May-18	lawrence livermore national laboratory	0	8	-8	-3.906482037
Apr-18	May-18	nnsa	0	11	-11	-5.412387961
Apr-18	May-18	new mexico	0	8	-8	-3.906482037
Apr-18	May-18	center for public integrity	0	15	-15	-7.420262528
Apr-18	May-18	lindsey geisler	0	6	-6	-2.902544753
Apr-18	May-18	national nuclear security administration	0	9	-9	-4.408450678
Apr-18	May-18	lisa gordon hagerty	0	7	-7	-3.404513395
Apr-18	May-18	don t	0	20	-20	-9.930105736
Apr-18	May-18	fbi	0	6	-6	-2.902544753
Apr-18	May-18	congress	0	10	-10	-4.91041932
Apr-18	May-18	energy department	0	6	-6	-2.902544753
Apr-18	May-18	martin heinrich	0	7	-7	-3.404513395
Apr-18	May-18	colorado	0	6	-6	-2.902544753
Apr-18	May-18	henry mcmaster	0	6	-6	-2.902544753
Apr-18	May-18	defense philip calbos	0	5	-5	-2.400576112
Apr-18	May-18	washington	0	7	-7	-3.404513395
Apr-18	May-18	national defense university	0	6	-6	-2.902544753
Apr-18	May-18	matt nerzig	0	6	-6	-2.902544753
Apr-18	May-18	u.s.	0	7	-7	-3.404513395
Apr-18	May-18	north koreans	0	6	-6	-2.902544753
Apr-18	May-18	south carolina	1	9	-8	-3.906482037
Apr-18	May-18	john e. hyten	0	6	-6	-2.902544753

Figure 3-10: Anomalous edges for "savannah river" using general seed phrases with increased articles and uncapped ego-networks.

The extreme increase in co-occurrences and large number of anomalous edges in May of 2018 necessitated the review of events extracted from these edges. Figure 3-11 shows many of the events extracted that specifically reference the announcements regarding pit production and MOX.

Center Node	Edge Node(s)	Extracted event text
savannah river	*congress*	the south carolina lawmakers expressed support for production of plutonium pits at the savannah river site but said they were concerned that congress would be skeptical of the move.
savannah river	*department of energy*	news alerts trump administration axes project to generate power from plutonium the energy department document estimated that diluting the plutonium would require jobs at savannah river
savannah river	*department of energy, lanl, south carolina*	the majority of the nation s production of plutonium cores for nuclear weapons would take place at the department of energy s savannah river site in south carolina under a plan certified by the nuclear weapons council and announced thursday but a lesser number of plutonium pits would still be made at los alamos national laboratory
savannah river	*department of energy, south carolina*	the department of energy submitted a document on may to the senate appropriations committee saying that the mixed oxide mox project at the savannah river site in south carolina would cost about billion more than
savannah river	*henry mcmaster*	henry mcmaster have been pushing to keep the mox mission even as local officials in the savannah river area near augusta ga. have been wooing the pit work with positive resolutions passed by local governments.
savannah river	*henry mcmaster*	henry mcmaster has expressed support for moving plutonium core production to the state as recently as an appearance in a city near savannah river earlier this week
savannah river	*lanl, new mexico, nnsa, south carolina*	with mox being discontinued the national nuclear security administration has proposed installing pits to store plutonium waste - per year at the savannah river site and per year at the los alamos national laboratory in new mexico
savannah river	*lanl, new mexico, nnsa, south carolina*	to achieve dod s pits per year requirement by nnsa s recommended alternative repurposes the mixed oxide fuel fabrication facility at the savannah river site in south carolina to produce plutonium pits while also maximizing pit production activities at los alamos national laboratory in new mexico according to thursday s release.
savannah river	*lanl, south carolina*	some experts are worried about the safety records of either choice los alamos national laboratory where plutonium parts have historically been assembled and the savannah river site in south carolina where other nuclear materials for america s bombs have been made
savannah river	*lanl, south carolina*	the agency has recommended that the pits be produced at the los alamos lab and at savannah river site in western south carolina which is already one of the most contaminated places on earth because of our country s nuclear program
savannah river	*new mexico*	according to nuclear watch new mexico a new doe deal was firmed up just this month for both los alamos and savannah river installations to divvy up the manufacture of new plutonium pits for bombs a year in georgia and a year in new mexico.
savannah river	*new mexico*	the contenders are the los alamos national laboratory in new mexico and the savannah river site in south carolina
savannah river	*new mexico, nnsa, south carolina*	the pentagon and the nnsa are now proposing that both los alamos in new mexico and savannah river in south carolina produce plutonium pits arguing that the military should not rely on a single facility for production.
savannah river	*nnsa*	the council has accepted a national nuclear security administration recommendation to repurpose a facility at savannah river to make pits
savannah river	*nnsa*	nuke agency pitches plutonium nuke agency pitches plutonium pits for savannah river site
savannah river	*nnsa*	the savannah river site will be producing plutonium pits for nuclear weapons according to the national nuclear security administration
savannah river	*nnsa, south carolina*	national nuclear security administration has officially proposed producing plutonium pits at two locations including the savannah river site in south carolina
savannah river	*south carolina*	the south carolina lawmakers expressed support for production of plutonium pits at the savannah river site but said they were concerned that congress would be skeptical of the move.
savannah river	*south carolina*	mox program at the savannah river site srs near aiken south carolina
savannah river	*south carolina*	on thursday evening hours after mr. trump announced that his meeting with kim jong-un the north korean leader would take place on june in singapore the pentagon and the energy department announced plans to begin building critical components for next-generation nuclear weapons at the savannah river site in south carolina
savannah river	*south carolina*	perry executed a waiver on may to terminate construction of the mixed oxide fuel fabrication facility at the savannah river site in south carolina

Figure 3-11 Extracted events and text for anomalous edges from "savannah river" during May 2018 using general seed phrases with increased articles and uncapped ego-networks.

Some events were extracted from a unique type of article. For example, one entity that was identified as anomalous was "south carolina". In Figure 3-12, observing the article titles it is determined that this is partially due to several job listings being released. Because the models use the word embedding rankings, these jobs are to some extent correlated to the seed phrases and therefore potentially related to pit production. Identifying a spike in related job postings is an observation worth noting.

title	adjacent_node	extracted_text
(USA-SC-Aiken) Executive Assistant/Coordinator	['america', 'south carolina']	[('savannah river', 'america', ['aecom executive assistant coordinator in aiken south carolina business line government united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking an executive assistant coordinator to be based in our aiken sc location.']), ('savannah river', 'south carolina', ['aecom executive assistant coordinator in aiken south carolina business line government united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking an executive assistant coordinator to be based in our aiken sc location.'])]
(USA-SC-Aiken) First Line Manager, Radiological Control	['america', 'south carolina']	[('savannah river', 'america', ['aecom first line manager radiological control in aiken south carolina business line government position title first line manager radiological control united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking a first line manager radiological control to be based in our aiken sc location.']), ('savannah river', 'south carolina', ['aecom first line manager radiological control in aiken south carolina business line government position title first line manager radiological control united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking a first line manager radiological control to be based in our aiken sc location.'])]
(USA-SC-Aiken) Limited Service Associate Design Specialist - Electrical	['america', 'south carolina']	[('savannah river', 'america', ['aecom limited service associate design specialist - electrical in aiken south carolina business line government position title limited service associate design specialist - electrical united states of america - south carolina aiken job summary savannah river remediation an llc of aecom is seeking a limited service employee lse associate design specialist - electrical to be based in our aiken sc location.']), ('savannah river', 'south carolina', ['aecom limited service associate design specialist - electrical in aiken south carolina business line government position title limited service associate design specialist - electrical united states of america - south carolina aiken job summary savannah river remediation an llc of aecom is seeking a limited service employee lse associate design specialist - electrical to be based in our aiken sc location.'])]
(USA-SC-Aiken) Manager, Construction Engineering	['america', 'south carolina']	[('savannah river', 'america', ['aecom manager construction engineering in aiken south carolina business line government position title manager construction engineering united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking a manager construction engineering to be based in our aiken sc location.']), ('savannah river', 'south carolina', ['aecom manager construction engineering in aiken south carolina business line government position title manager construction engineering united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking a manager construction engineering to be based in our aiken sc location.'])]
(USA-SC-Aiken) Manager, Construction Operations Site B	['america', 'south carolina']	[('savannah river', 'america', ['aecom manager construction operations site b in aiken south carolina business line government position title manager construction operations site b united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking a manager construction operations site b to be based in our aiken sc location.']), ('savannah river', 'south carolina', ['aecom manager construction operations site b in aiken south carolina business line government position title manager construction operations site b united states of america - south carolina aiken savannah river remediation an llc of aecom is seeking a manager construction operations site b to be based in our aiken sc location.'])]

Figure 3-12: Extracted article titles and event text for anomalous edges from "savannah river" during May 2018 using general seed phrases with increased articles and uncapped ego-networks.

Simply reviewing the pipeline's capability to identify the event one month prior to the official announcement does not equate to complete success. It is also necessary to evaluate the capability of the model to provide insight with a longer lead time. A search of additional events that were extracted from the database reveals that the modeling pipeline discovered events related to the awarding of DOE contracts at major sites, DOE investments in various programs, accidents at DOE national laboratories related to pit production, speculations about the fate of pit production in the DOE complex, domestic and international shipments and receipts of nuclear materials at DOE sites, termination of non-proliferation agreements with Russia, termination of MOX, new weapons development approvals/testing, nuclear posture reviews, DOE cleanup/production milestones, political opinions, and many examples of nuclear weapons watch groups' opinions, among many others. A sample of these articles and events can be found in Figure 3-13. It should be noted that events are not always extracted from articles (based on the parameter settings), but the article title, entities, and text can still be reviewed by an analyst for insight.

Date	Title	Edge Node(s)	Extracted Event Text
06/06/2016	NNSA Announces Arrival of Plutonium and Uranium from Japan's Fast Critical Assembly at Savannah River Site and Y-12 National Security Complex	['new mexico']	the rod the plutonium material recovered from japan will be prepared for disposition using the unique capabilities of the savannah river site srs for eventual disposal at doe s waste isolation pilot plant wipp near carlsbad new mexico.
06/07/2016	After plutonium arrives from Japan, Haley, Wilson still want nuclear materials out of S.C. - Aiken Standard	['new mexico']	according to that decision metric tons is expected to be processed at savannah river site in preparation for long-term disposal at the waste isolation pilot plant near carlsbad new mexico.
06/15/2016	Commission completes assessment of nuclear shipment at Savannah River Site - Aiken Standard	['new mexico']	the materials at savannah river site or srs are headed for interment at the waste isolation pilot plant or wipp in new mexico.
08/02/2016	Nuclear agency reports violations with construction, inspections at MOX plant - South Carolina Radio Network	[' year']	
09/01/2016	Fwd: EM Update | Vol. 8, Issue 16 | Aug. 31, 2016	['piketon ohio', 'u.s.', 'paducah', ' year', 'hanford']	as cleanup of the inactive plutonium fuel form puff facility in building enters its second year savannah river nuclear solutions risk reduction approach continues to bring success.
09/09/2016	Energy Department continues push to stop MOX plant	['u.s.', 'america', ' year']	
09/15/2016	Aiken official: Savannah River Site's MOX purposefully left out of NNSA discussion	['u.s.', ' year', 'virginia']	the speaker from the national nuclear security administration at the energy communities alliance meeting in arlington virginia this week intentionally snubbed the mixed oxide fuel fabrication facility or mox under construction at the savannah river site one aiken official said.
09/24/2016	Savannah River Remediation progress is topic of American Nuclear Society meeting	['america', ' year']	speaker mark schmitz chief operating officer and deputy project manager at savannah river remediation touts srr s safety technology utilization track records at tuesday s american nuclear society meeting in augusta.
09/27/2016	Groups oppose plan to store nuclear fuel near SRS	['u.s.']	
10/01/2016	Plans for new nuclear reactors to go before NRC for final step of licensing process	['georgia']	currently the are four nuclear power reactors under construction in the united states two at vc summer in jenkinsville and another two just across the savannah river at plant vogtle near waynesboro georgia.
10/03/2016	Russia Withdraws From Plutonium Disposal Treaty	['american']	but glitches and cost overruns in the mox plant at savannah river s.c. delayed the american program.
10/04/2016	Liquid waste shipments from Canada to Savannah River Site suspended	['energy department']	
10/04/2016	Feds agree to delay possible nuclear waste shipments across Peace Bridge - City & Region - The Buffalo News	['energy department']	current plans call for the repatriation of the nuclear material from chalk river laboratories an ontario nuclear facility to the savannah river site in aiken s.c. environmental groups argue transporting liquid nuclear waste in the manner proposed by the energy department is unprecedented and poses potential for catastrophic harm to public highways waters and the general population.
10/13/2016	Irreversible Arms Reductions Finds Reverse Gear: Mr. Putins Russia Sends A Signal Analysis	['new mexico', 'energy department', 'american']	the energy department s savannah river site s mixed oxide fuel fabrication facility in south carolina has been plagued with delays and cost overruns leading president obama to order its closure and to shift disposition to the immobilization method at a department of energy facility in new mexico.
10/24/2016	Energy Department Launches $10 Million Effort to Develop Advanced Water Splitting Materials	['energy department', 'american']	
11/03/2016	Security Event at Department of Energy's SRS prompts emergency response	['energy department']	
07/21/2017	Savannah River Site Reaches Milestone In Supplying Tritium For National Defense	['washington d.c.']	the national nuclear security administration s savannah river tritium enterprise srte in aiken s.c. has conducted three tritium extractions in fiscal year marking the first time the tritium extraction facility has performed more than one extraction in a year.
08/01/2017	Nuclear weapons contractors repeatedly violate shipping rules for dangerous materials	['internal energy department', 'nevada national security site', 'center for public integrity']	
10/14/2017	BWXT-Led Team Awarded $4.7 Billion Contract For DOE's Savannah River Site Liquid Waste Services	['businesswire.com']	
11/08/2017	Department of Energy Cites Savannah River Nuclear Solutions, LLC for Worker Safety and Health Program Violation	['washington d.c.']	department of energy cites savannah river nuclear solutions llc for worker safety and health program violation the u.s. department of energy doe today issued a preliminary notice of violation to savannah river nuclear solutions llc srns for a violation of worker safety and health requirements.

Figure 3-13: Extracted article titles and event text for anomalous edges from "savannah river" from June 2016 to April 2018 using general seed phrases with increased articles and uncapped ego-networks.

4.0 Discussion

Notably, even without a direct mention of the SRPPF, the contextual time dependent word embedding models as trained on the Twitter dataset were able to make a strong connection between the key terms "pit_production" and "savannah_river_site" as early as 2016, as shown in Figure 1-5. Exploring the Twitter corpus further revealed that this is a result of an accurate, and adequate, contextual representation of the activities that occur at the SRS (i.e., by capturing Tweets discussing nuclear materials reprocessing, plutonium, etc.) and their similarity to fissile core fabrication. As discussed in the sections above, the metrics used in Figure 1-5 do not supply context as to why the connection has been made. However, as also noted, a direct mention of a specific event is not necessary for the contextual relationship to grow stronger in the embedding models. This demonstrates the utility of having several metrics working together that show the relationship of key terms and phrases, as well as identifying key points in time where contextual shifts have occurred to extract explicit events of potential interest.

Additionally, despite the fact that the lead time for detecting the targeted event of interest (i.e., fissile core fabrication at the SRS) using the Webhose dataset and corresponding pipeline was only one month, the capability to detect these events is present. This is demonstrated in Figure 3-11 which shows that events indicating the targeted event of interest were extracted by the pipeline from the data, regardless of lead time. Furthermore, exploring the extracted events in earlier time periods for Savannah River, such as those shown in Figure 3-13, demonstrated that although "fissile core fabrication at the SRS" isn't explicitly detected in earlier time periods, the pipeline extracts many events that, in hindsight, have a close relationship to the targeted event of interest (e.g., articles about MOX and termination of MOX, questions about pit production at Los Alamos, the ending of the treaty between Russia and the US, etc.) and that build a strong contextual representation of the SRS and its activities. Therefore, future work should place emphasis on parameter tuning for known entities of interest that have apparently low signal, where perhaps fusion of the indicators extracted from the multiple document sets (i.e., from different key phrases), across closely related entities, or from the two datasets (i.e., including Twitter) over time, might allow enrichment for successive time periods and potentially increase the lead time.

The modeling pipeline as applied to both datasets captures a broad range of events that are governed by the keywords and phrases used to query the data sources. As expected, higher profile events were captured with relative ease and with large numbers of events, whereas lower profile (i.e., less widely publicized) events were captured, but with a much lower signal. One example of this is the overall signal that is produced by DOE sites/laboratories such as Los Alamos and Sandia versus Savannah River. The former two laboratories are larger and frequently have more widely publicized events occurring at their sites, whereas Savannah River is smaller and has fewer. This reality is reflected in the output of both datasets, where significantly larger volumes of Tweet traffic and news articles are produced by the larger entities. Notably, however, even with a relatively general set of key terms, Savannah River (i.e., a smaller entity relative to others) still produced signal of the targeted event. Therefore, if a user had a growing interest into a smaller entity, perhaps a new keyword set with finer contextual resolution of the activities occurring at the site would produce even more signal while reducing noise from less germane activities at the larger sites. Thus, in future research efforts, the ability to dynamically evolve the keyword sets over time could be of use (i.e., using dynamic query expansion techniques). During the course of this project, this extra step was not taken as a new data query would have been required at additional cost.

With such a large volume of data (i.e., over 3 million Tweets and 12 million long-form news articles), it is difficult to quantify accuracy metrics such as precision and recall that describe the quality of events that have been extracted. However, the research team's subject matter experts explored the events returned by many key entities of interest in the DOE Complex to provide a qualitative assessment. In this effort, it was identified that, in general, greater than 50% of the events that are extracted are germane to either pit production or other nuclear activities (i.e., as defined in the conceptual model) that would have been

expected to be discovered in the timeline of interest (e.g., SRS involvement at Paducah/Portsmouth, waste shipments to WIPP, material receipts, material processing/production milestones, etc.). Of the remaining 50% of the events extracted, it is estimated that approximately 25% consists of events that are not readily identifiable as related to an event of interest, but that are closely related to an entity of interest, and 25% consists of noise. Note that the former might still contain information that is relevant for building a strong contextual representation of the entities or activities being represented by the language models or contain information that becomes more apparently useful at later times. The extracted events from the Webhose pipeline contained a lower amount of noise relative to the Twitter dataset. The explanation for this is outlined as:

- Queries to Webhose filtered for English only articles. The Twitter dataset was not language filtered (language filtration was tested and found to cause a loss of some signal), though the vast majority of Tweets are in English.
- Tweet data generally contains more use of slang, allowing wholly unrelated topics (e.g., sports-related Tweets) to enter into the vocabulary.
- Because non-English Tweets and out-of-domain Tweets are relatively infrequent, they act as noise. However, some activities, such as pit production, are also infrequently occurring in the language model and therefore can have high similarity with noise. Thus, the embedding vector similarity-based approach for event extraction that is used in the Twitter pipeline can, at times, allow more noise to be extracted as events.
- The Webhose pipeline requires the co-occurrence of entities within articles to occur, and therefore has a higher likelihood of reducing noise as an entirely out of context co-occurrence is unlikely.

The fusion of insights extracted from the two models remains manual at this time. However, corroborating articles and Tweets were discovered. Therefore, future modeling efforts should seek to combine the output from the two datasets and perhaps provide some indication of events that are found in the multiple data streams, while also enhancing both pipelines in successive time periods with all indicators that have been extracted. While the two pipelines differ in the techniques applied, both seek to identify points in time where key words and phrases that are presumed to be anomalous, or have changed. Thus, the similarity in the event extraction output from the two pipelines will allow for seamless fusion in future development efforts.

The modeling pipelines that have been developed for both data sources have a wide parameter space that can potentially greatly impact the ability to identify indicators of events of interest. A summary of important parameters and their impact is provided in Table 4-1. Perhaps the most important parametrization step is in training the embedding models such that a rich contextual representation is obtained. Therefore, it is important to perform exploratory data analysis (e.g., using the techniques developed in the FY20 preliminary modeling) early in the pipeline to ensure that the key terms have intuitively accurate contextual representations. Exploring the impact of the corpus size in both pipelines revealed that erring on the side of too much data provides enhanced results, though greater noise is produced. Similarly, later stages of the pipeline should be explored with varying parameters to ensure that the models are not prohibitively exclusive of information. With this in mind, the use of ensemble techniques, such as those applied to the Twitter dataset, implicitly allows exploration of a wider range of parameters where post-processing and filtration of noisy data can be performed.

Table 4-1. Parameters in the modeling pipeline.

Pipeline Step	Parameters	Impact
Sampling Database	Number of key terms used	Number of documents/Tweets returned. Too many documents/Tweets may produce excess noise and too few documents/Tweets may produce little/no signal.
Text Pre-Processing	Tokenization/Lemmatization of key phrases	Embedding models are trained with tokenized and lemmatized multi-word phrases, potentially enriching the contextual model through a reduction in noise (e.g., "Savannah" "River" and "Site" trained alone versus "Savannah_River_Site").
	Entity disambiguation	The same entity may have multiple appearances in the contextual model (e.g., "SRS" versus "Savannah River Site"), but each occurrence may have a different contextual representation. Entity disambiguation therefore decreases the vocabulary size and gives exactly the same contextual representation for each occurrence of the same entity.
Time Dependent Word Embedding Models	Minimum occurrence of words in corpus to be in vocabulary	Words that occur fewer than the threshold minimum are excluded from the embedding vocabulary, but may have importance. Including all terms that appear may produce too much noise.
	Window size	Number of documents/Tweets per time period captured with each new embedding model.
	Static Growing versus Rolling Windows	Static windows maintain all text through time, while rolling windows drop text from embedding models over time. Static windows may become prohibitively large over time, while rolling windows may lose important information.
	Window Growth Rate	Number of documents/Tweets and the amount of time added to each successive embedding model.
	Rolling Length	Number of documents/Tweets that are dropped from the embedding model in each successive time period.

	Word2Vec Formalism for Embeddings	Embedding models are trained with the Word2Vec formalism. The use of other embedding model formalisms (e.g., BERT) may enrich the models further.
	Number of Articles per Seed Phrase (Webhose dataset only)	Word embedding models are used to generate phrase and document embeddings that rely on filtration of the top N articles that are most similar (i.e., cosine similarity) to each seed phrase. This parameter helps to reduce noise by filtering irrelevant articles. Allowing more articles may capture more information, but may increase noise. Allowing fewer articles may miss information, but will reduce noise.
Weighted Average Embedding Vector (Twitter dataset only)	Number of nearest neighbors	Number of terms from the embedding model that impact the weighted average embedding vector.
	Number of ranks	The inclusion of higher order interactions on the root word's weighted average embedding vector.
	Ensemble model versus single model	The number of different models that are used to capture important time periods where significant contextual shifts are detected. Including higher order interactions may capture more obscure connections.
Similarity Metrics	Jaccard Similarity, Cosine Similarity, Other graph similarity metrics	The similarity metrics that are used to compare embedding models over time may perform differently based on the breadth of the contextual representation of the key term being explored.
Inflection Window Detection (Twitter dataset only)	Gaussian Filter Smoothing Parameters	Modifying the smoothing parameters of the Guassian filter will change the timing and width of the inflection point windows, potentially resulting in shorter or longer windows.
Anomaly Detection (Webhose dataset only)	Z-Score	A lower z-score captures more edges as anomalous.
	Entity Co-Occurrence Graph Size	A higher number of entities in the co-occurrence graph allows wider exploration of the interrelationship of key terms, but may also introduce more noise.

Event Extraction (Webhose dataset only)	Word Adjacency	A higher word adjacency increases the chance that text might be extracted from an article that contains both terms of an anomalous edge, but also increases the length of the extracted text.

5.0 Conclusions

A demonstration prototype modeling pipeline has been developed to identify indicators of events of interest by detecting contextual shifts in key words and phrases using time dependent word embedding models and subsequent analyses that apply graph theory and anomaly detection. Two separate modeling pipelines have been developed for the two datasets that were explored in this work, but the techniques used in each can be applied to either dataset. The modularity of the approach allows for the inclusion of new techniques and the incorporation of several techniques working together, as has been demonstrated above. Notably, the techniques that have been developed for each dataset were shown to extract a broad range of events of interest, including the target event, "fissile core fabrication at the Savannah River Site", prior to its official announcement. Future work will seek to apply these techniques to new subject domains in new data environments and fuse information extracted by each pipeline.

6.0 References

[1] T. L. Danielson, A. A. Kail, and J. A. Pike. April 2020. *Event Definition for the Automated Detection of Nuclear Proliferation Activities.* Aiken, SC: Savannah River National Laboratory, SRNL-STI-2020-00155, Rev. 0.

[2] T. L. Danielson, J. A. Pike, T. Whiteside, B. Mayer, N. Muralidhar, N. Self, P. Butler. January 2021. *Machine Learning Using Open Data Sources for Detection of Nuclear Proliferation Activities (U)*. Aiken, SC : Savannah River National Laboratory, SRNL-STI-2021-00047.

[3] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, et al. 2014. *'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators*. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 1799–1808. arxiv.org/abs/1402.7035.

[4] W. L. Hamilton, J. Leskovec, D. Jurafsky. August 2016. *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. ACL. 1489-1501.

Distribution:

M. J. Barnes, mark.barnes@srnl.doe.gov
J. S. Bollinger, james02.bollinger@srnl.doe.gov
N. J. Bridges, Nicholas.Bridges@srnl.doe.gov
P. Butler, pabutler@vt.edu
G. R. Cefus, gregory.cefus@srnl.doe.gov
T. L. Danielson, Thomas.Danielson@srnl.doe.gov
C. C. Herman, connie.herman@srnl.doe.gov
C. M. Gregory, clint.gregory@srnl.doe.gov
R. B. James, Ralph.James@srnl.doe.gov
R. D. Jeffcoat, ron.jeffcoat@srnl.doe.gov
B. Mayer, bmayer@cs.vt.edu
N. Muralidhar, nik90@vt.edu
P. R. Nuessle, patterson.nuessle@srnl.doe.gov
N. Self, nwself@vt.edu
D. L. Wilson, david.wilson@srnl.doe.gov
M. Taylor, marc.taylor@srnl.doe.gov
T. P. Taylor, tammy.taylor@srnl.doe.gov
S. J. Branney, sean.branney@srnl.doe.gov
M. Cofer, marion.cofer@srnl.doe.gov
L. T. Brown, linday.brown@srnl.doe.gov
B. Lee, brady.lee@srnl.doe.gov
Records Administration (EDWS)