SANDIA REPORT
SAND2016-9834
Unlimited Release
Printed Month and Year

Geospatial-Temporal Semantic Graph Evaluation for Induced Seismicity Analysis

Diane Myung-kyung Woodbridge, and Randy C. Brost

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: http://www.ntis.gov/search

SAND2016-9834
Unlimited Release
Printed Month Year

Geospatial-Temporal Semantic Graph Evaluation for Induced Seismicity Analysis

Diane Myung-kyung Woodbridge, Randy C. Brost
Sandia National Laboratories
P.O. Box 5800, MS 1326
Albuquerque, New Mexico 87185-MS1326
dwoodbr, rcbrost@sandia.gov

Abstract

We assess how geospatial-temporal semantic graphs [1] and our GeoGraphy code implementation might contribute to induced seismicity analysis. We focus on evaluating strengths and weaknesses of both 1) the fundamental concept of semantic graphs and 2) our current code implementation. With extensions and research effort, code implementation limitations can be overcome. The paper also describes relevance including possible data input types, expected analytical outcomes and how it can pair with other approaches and fit into a workflow.

ACKNOWLEDGMENTS

We gratefully thank Kelly Rose, Jennifer Bauer, Devin Justman and Deborah Glosser at National Energy Technology Laboratory (NETL) for providing geoscience expert knowledge and insights. We also thank Kristina Czuchlewski and David Stracuzzi at Sandia National Laboratories for their efforts on scoping the project and examples.
This work was funded by the SubTER Fossil Energy Program.

CONTENTS

1. Introduction.....7

2. Description.....9
   2.1. Fundamental Concept.....9
   2.2. Implementation.....9

3. STRENGTH.....11
   3.1. Fundamental concept strengths.....11
   3.2. Implementation strengths.....12

4. Weaknesses.....13
   4.1. Fundamental concept weakness.....13
   4.2. Implementation weakness.....13

5. Possible Extensions.....15

6. Relevance.....17
   6.1. What data would it use?.....17
   6.2. Expected analytical outcomes?.....17
   6.3. Would it fit into a workflow?.....17
   6.4. Does it pair with other approaches well?.....18

7. References.....19

Appendix A: GeoGraphy Algorithm and workflow.....21

Appendix B: Evaluation Slides.....25

Appendix C: Preliminary Results Slides.....29

FIGURES

Figure 1. Example Hierarchical Spatiotemporal Search.....11
Figure 2. Example star graph search definition and outputs: An earthquake (hub) with preceding injection events (spoke) along a common fault (spoke).....21
Figure 3. The path found by path-connected search is shown by the dark path that connects the Parking Lot node to the House node using road network segments.....22
Figure 4. Example connected components search: connected pixels of faults.....22
Figure 5. Change detection example : A complex of new buildings, replacing a previous complex.....23

1. INTRODUCTION

This document will describe our assessment of how geospatial-temporal semantic graphs, and our GeoGraphy code implementation, might contribute to SubTER project objectives. Note that there are two distinct issues here: the fundamental graph approach, and our code implementation. Code implementation shortcomings can be overcome with additional programming effort, while limitations of the fundamental approach require deeper research to resolve, if even possible. Thus in the discussion that follows, we will attempt to distinguish these issues.

In the rest of the paper, we will describe the geospatial-temporal graph method and its strengths and weaknesses, focusing on both fundamental and implementation aspects. Then we will discuss possible extensions to resolve weaknesses and relevance with the induced seismicity approach. The appendix will provide additional details including algorithm description and possible applications for SubTER.

2. DESCRIPTION

2.1. Fundamental Concept

Geospatial-temporal semantic graphs are data structures that represent discrete world objects and events as nodes, and relationships between those objects and/or events as edges. This fundamentally requires delineating objects and events in space and time, so that graph nodes can be identified. At the most basic level, once discrete objects and events are identified, the graph structure enables search for patterns specifying pre-specified relationships including both spatiotemporal relationships and other semantic relationships. This approach has a great deal of flexibility in representing a wide range of problems. However, it also incurs fundamental challenges that must be appreciated by SubTER; these will be explained below.

2.2. Implementation

We have implemented a software code base, named GeoGraphy, which implements geospatial-temporal semantic graphs and several associated search algorithms. This code has demonstrated the application of geospatial-temporal semantic graphs to a fairly broad set of example problems, ranging from wide-area search for facilities in overhead image data to analysis of trajectories within geospatial context to a simple analysis of induced seismicity. While the code has successfully solved these examples, it remains a research code with a number of limitations, explained below.

3. STRENGTH

3.1. Fundamental concept strengths

The geospatial-temporal semantic graph representation provides semantic encoding and search of data. Geospatial-temporal semantic graphs encapsulate attributes of regions, events, and physical and temporal relationships into nodes and edges, which enables data from different sources to be represented in a single uniform search representation.

As the graph representation allows relationships to be represented as edges, it allows flexible search designs and efficient searches compared to relational databases, which require several join operations over tables which are complicated and expensive. Additionally, many graph algorithms including connected components, matching, path search, edge and node connectivity, transitive closure, etc, can be applied to geospatial-temporal semantic graphs.

The graphs enable identification and analysis of aggregates that correspond to semantically important multi-object groups, which taken together constitute important spatiotemporal features for subsequent analysis or review. For instance, the geospatial-temporal semantic graph representation allows finding an earthquake with preceding injection events along a common fault, using both spatial and temporal relationships (Figure 1). The resulting injection-fault-earthquake ensembles then become higher-level semantic objects available for further analysis. One example subsequent analysis might be cascading earthquake analysis, which could also be supported by the graph representation.

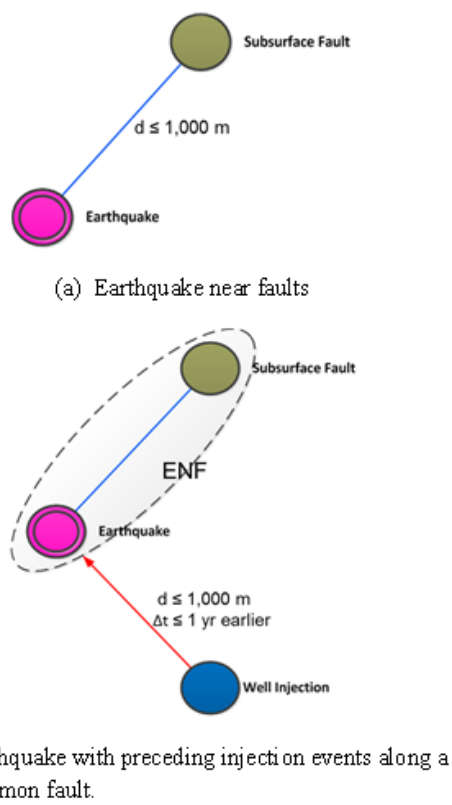


Figure 1. Example Hierarchical Spatiotemporal Search

3.2. Implementation strengths

The GeoGraphy software allows multi-modality data as input including preprocessed image data and point data. The GeoGraphy graph construction program employs a streaming algorithm that generates this type of geospatial-temporal semantic graph incrementally; image data is scanned line-by-line, enabling the processing of very large images. The program also accepts formatted text files as a source of point data.

GeoGraphy allows search results to be written back to the graph, which in turn enables multi-step searches and cumulatively enhances the graph with hierarchical and growing semantics. For instance, an earthquake with associated preceding events can be found via a multi-step search (Figure 1):

1. Search for earthquake events near subsurface faults, producing "earthquake near fault" (ENF) matches.
2. Search for injections near the resulting ENF matches, subject to the temporal constraint that the injections must precede the earthquake in the ENF.

This will produce matches that include an earthquake with associated injections that are spatially close to a common fault, and where the injections are guaranteed to precede the earthquake within a user-specified time delay. Such a search could support exploration of a hypothesis "induced earthquakes are preceded by injections that propagate along a fault."

In general, a hierarchical search with search writeback helps reuse searches, run smaller and modular searches, develop hierarchical semantics and expand the scope of possible searches.

GeoGraphy supports match quality ranking based on a user-defined metric. The search metric can be defined using properties in the graph including attributes on nodes, edges, cardinality, etc. For instance, users can sort the matches based on the size of the earthquake, injection volumes, distances between injection and earthquake, etc. The quality metric method sorts best matches to the front, while poor false positives are shifted to the back of the list.

4. WEAKNESSES

4.1. Fundamental concept weakness

Geospatial-temporal graphs are a discrete representation that fundamentally requires identifying discrete objects in the world. Thus any given object in the world must have a well-defined boundary. (Note that this does not necessarily require partitioning the world; objects may have overlapping boundaries. See below.) For SubTER, this raises these important considerations:

- *Uncertainty in subterranean features.* At some locations, much is known about subterranean features (for example, where well core samples are obtained), while at other locations little is known. For areas where little is known, it may be difficult to define meaningful nodes, and in particular the separation between nodes.
- *Lack of clearly defined boundaries.* Some object boundaries are crisp and well-defined, such as the differentiation between a built structure and a surrounding lawn. Meanwhile other boundaries might be diffuse and ill-defined, such as the transition between thick forest, thinning trees at the edge of a meadow, and an open grassy plain. Where does the forest end and the plain begin? The lack of a clear boundary transition might be common in some subterranean domains.
- *Object boundaries might not be semantically relevant.* Even if subterranean features are well-known and can be clearly delineated, they might not match what is relevant for a given analysis. For example, imagine an underground petroleum deposit in permeable rock. Imagine that extraction in one area has led to reduced petroleum concentration and/or pressure, and this effect appears as a gradient along the extent of the deposit. Treating this as one single monolithic feature might overlook the important distinction that one end of the deposit has less petroleum content and pressure than the entire deposit.
- *Some objects might not have boundaries at all.* Consider an example well injection. While the injection's geospatial location is discrete and well-defined, suppose that it is operated continuously in support of a nearby process. Considering time, where is the temporal boundary?
- *3-d representation.* Full three-dimensional representation of subsurface features further complicates these issues. It also makes partitioning connected features more difficult.

In short, the discretization required by geospatial-temporal semantic graphs imposes requirements and models of SubTER-relevant objects, and what is known about them. It seems that quakes and injections submit to discrete representations fairly easily, but effective discretization of subsurface features is less clear to us at this time. However, the "Possible Extensions" section below will outline how additional research might develop a technique for resolving these issues.

4.2. Implementation weakness

The GeoGraphy software is research code. It does not accept the wide range of input formats one would expect from a commercial tool, it does not have a full-featured friendly user interface, and it is not yet optimized for speed. Further, some desired node-to-node relationships that may be important to SubTER problems might require code extensions to support.

Temporal analysis capabilities are currently limited. The code already does support change detection, temporal clusters via connected component analysis and short-sequence temporal ordering algorithms (up to three steps). But the code does not support duration-limited search and multi-step sequence search. Please see more details in the appendix. The range of possible temporal analysis approaches and algorithms is huge, and it seems unrealistic to expect any implementation to fully cover all potential temporal analysis capabilities that might be desired.

Modeling the interaction between subterranean features might require representation of parameters such as subsurface permeability, and a desired analysis might seek the likely speed and volume of flow through a subsurface region, accumulating terms such as pressure drop, total volume, or aggregate flow rate across sets of nodes and edges. While the fundamental geospatial-temporal graph approach could support this sort of information, is not yet supported in the GeoGraphy implementation.

The current code supports two-dimensional data, plus height above ground. Three-dimensional subsurface features are not supported.

5. POSSIBLE EXTENSIONS

Above we noted that geospatial-temporal semantic graphs require discretization of the domain, and that this can lead to challenges due to uncertainty, diffuse boundaries, and a mismatch between object boundaries and what's semantically relevant. In this passage, we will describe an envisioned extension of the concept to resolve these issues.

The basic approach is to define "cells" underground that incorporate regions where a given description can hold. These are reminiscent of finite-element meshes, but differ in two important ways: (1) They are not of regular geometric shape, and (2) they do not partition space. Instead, they have shapes that vary based on available data, and can overlap.

- Using this approach, we can resolve the diffuse boundary issue by either defining intermediate cells, with varying levels of mixed content, having cells overlap where boundaries are not clear, or some combination of both.
- We can address the mismatch between object boundaries and semantics of interest by splitting cells into smaller cells with the desired semantic properties. For example, for the extended petroleum deposit with a pressure gradient described above, we could break the petroleum deposit object into pieces, each with a nearly constant pressure within it. This is similar to techniques we have employed for past problems where object boundaries did not exactly match the semantics of interest, although the computational details would be different.
- We can address the uncertainty issue by defining cell contents to be ambiguous where data are uncertain, and adjusting cell density and size to match the specificity of what is known in a given region. For example, we might represent a larger number of fine-grain high-confidence cells in the vicinity of a core sample location, with broader, ambiguous cells in the wide area far from well bores.
- This cell approach, with overlap and uncertainty, could be used in three dimensions also, including the presence of oblique fault boundaries.

The graph analysis would remain similar, with attributes of cells used to guide search, and relationships between cells calculated to support search. All of these envisioned new capabilities would require significant further development, both conceptually and in code implementation.

A second extension would be to include weights on nodes and edges (e.g., permeability). This would enable potential future graph search algorithms to accumulate flow and pressure drop along sequences of cells to account for flow distance and resistance effects. Again this would require significant development. Note that other database approaches do not support the aggregate chain analysis that would be enabled by the graph approach.

A third extension would be to augment the graph search with a temporal causality analysis. We have a current LDRD project in progress which is developing techniques for analyzing multi-modality data streams and events of interest, seeking precursors that are causally related to the events in a statistically significant way. These analysis methods consider both temporal statistical causality relationships and spatiotemporal relationships explicitly represented in the geospatial-temporal graph. We believe that the analysis techniques under development are an

excellent match to the SubTER analysis problem, and feel that extending GeoGraphy to perform temporal causality analysis for SubTER data sets might be very fruitful.

6. RELEVANCE

6.1. What data would it use?

GeoGraphy currently supports image and point data. For the SubTER domain, the code can utilize projected two-dimensional subsurface fault data, earthquake data including location, magnitude, timestamp, etc., human activity data such as injection location, volume, timestamp, type, etc. Two-dimensional maps of fault locations can currently be input; with significant extension, 3-d fault models could be input. In addition, landcover image data could be utilized to search aspects not included in the point data sets such as construction, nearby facilities, population density of nearby locations, etc.

6.2. Expected analytical outcomes?

If we envision a system with the extensions listed above, we would anticipate producing an analysis of induced seismicity that would shed light on the conditions that preceded induced earthquakes, demonstrating a statistically significant causal relationship. This would help quantify the possible causal link between injection or other human activities and later earthquakes, and characterize the conditions that are most closely correlated with induced earthquake events. If successful, these results might also be used to identify regions most prone to future induced seismic events, or the human activities most likely to cause further seismicity.

Here's how this might work: First we would extend GeoGraphy to support three-dimensional subsurface feature models. This might be accomplished using the cell approach described above, or perhaps via a simpler approach of representing faults as three-dimensional surfaces approximating the fault location. A graph decomposition process analogous to our current path partitioning capability would delineate these fault models into individual segments suitable for detailed graph analysis. Then a multi-step graph search would identify (1) earthquakes near fault structures, (2) well injections and other human activity near fault structures, and (3) activity*→fault* → earthquake ensembles (where * denotes potentially multiple entities). These ensembles would then inform a temporal analysis that would determine whether the aggregate of many ensembles across Oklahoma is consistent with a statistically significant indication of injections causing seismicity, and if so, where and under what injection conditions. Then interpretation of this output would yield the results envisioned above. Note that this envisioned outcome would require substantial extensions of our current capability, developed under future SubTER funding.

6.3. Would it fit into a workflow?

We don't have a clear understanding of this question. By "workflow," do we mean work flow in a production user environment, or workflow in a research environment? The following paragraphs will offer a few remarks on each of these.

A production user scenario might correspond to a situation where one or more users perform induced seismicity analysis on a recurring basis, in response to the arrival of new data. In such a scenario, we would envision a growing geospatial-temporal semantic graph, encoding past seismic events and human activity, and as new data arrives, it is added to the graph. Searches for correlations of interest could then be repeated, to identify new occurrences of patterns of interest, and temporal causality analysis could also be updated. We note that changes to the model of

subsurface features could also be modeled, using the geospatial-temporal graph's durable change analysis representation. (The feasibility of this will depend on the nature of the changes to the subsurface model, for example its geometric representation and whether it obeys certain monotonicity requirements. More study is needed here.)

For a research work flow, we could envision data provided by data providers, passed to researchers constructing graphs to support inquiry. Meanwhile, information describing desired questions flows from other researchers to those constructing the graph. Then the questions would be translated to graph search queries, fed as input to graph search software to produce ensemble matches. The resulting matches are then output and passed along to interested researchers for further analysis. The injection-fault-earthquake analysis described in the Appendix is a simple example, where the output of the graph search could be passed to subsequent temporal analysis.

6.4. Does it pair with other approaches well?

The geospatial-temporal graph technique might fit well with some other methods explored in the SubTER sapling phase. For example, outcomes from tensor decomposition would narrow down data that is relevant to seismic events and could potentially be used as GeoGraphy input, yielding the benefit of reducing complexity of graphs. As another example, various uncertainty models could be incorporated with geospatial-temporal semantic graphs to represent uncertainty of graph attributes and elements, which would then propagate into estimates of confidence in the output search and analysis results.

In addition, we feel that there is a potential for pairing this approach with analysis techniques being developed at other laboratories, such as the nearest-neighbor work being developed by Veronika Vasylkivska at NETL. As another example, one could imagine pairing geospatial-temporal graphs with graph-based uncertainty analysis techniques including Bayesian networks [4], which are being explored by Jen Bauer at NETL. But to date we have not put sufficient effort into exploring possible synergistic connections.

The high-performance computing work that is part of SubTER Task 1 could inform further extensions of GeoGraphy to run on parallel platforms to improve computation speed. In our work so far and also envisioned in the near term, we have emphasized study of how to produce a correct, useful computation result, and have not focused on speed. Once high-value computation designs have been achieved, further work could optimize performance and integrate with SubTER Task 1 results. We have not yet engaged this topic in detail.

7. REFERENCES

[1] Brost, Randy C., William C. McLendon III, Ojas Parekh, Mark D. Rintoul, David R. Strip, and Diane Myung-kyung Woodbridge. "A computational framework for ontologically storing and analyzing very large overhead image sets." In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, pp. 1-10. ACM, 2014.

[2] Brost, Randolph C., David N. Perkins, Kristina Czuchlewski, "Activity Analysis with Geospatial-Temporal Semantic Graphs." No. SAND2015-10455. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2015.

[3] McLendon, William C., and Randolph Brost. "Path Network Recovery Using Remote Sensing Data and Geospatial-Temporal Semantic Graphs." No. SAND2016-4557. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2016.

[4] Stracuzzi, David J., Randy C. Brost, Cynthia A. Phillips, David G. Robinson, Alyson G. Wilson, and Diane M-K. Woodbridge. "Computing quality scores and uncertainty for approximate pattern matching in geospatial semantic graphs." Statistical Analysis and Data Mining: The ASA Data Science Journal 8, no. 5-6 (2015): 340-352.

APPENDIX A: GEOGRAPHY ALGORITHM AND WORKFLOW

Search Algorithms Supported by GeoGraphy

- Star Graph Search – A hub and spokes, with various constraints.

  A star-graph search query is defined with respect to a hub node of a given type and potentially heterogeneous spoke nodes connected to the hub node. The star graph algorithm also includes the notion of a spoke-to-spoke constraint, which is a constraint which must be satisfied across certain spokes. This is required to solve some problems of interest, due to actual inter-spoke requirements[1].

Figure 2. Example star graph search definition and outputs: An earthquake (hub) with preceding injection events (spoke) along a common fault (spoke).

- Transitive Closure Search – Unlike other geographic information systems, GeoGraphy provides transitive closure searches via the interrupted star algorithm which allows a user to define ordinary nodes and interrupt nodes along with edges connecting them [2].

- Path Search – Some nodes from image data including roads and subsurface faults are highly connected to a large portion of a graph. This does not match how a human would typically describe a road—we tend to think of roads in smaller units such as "blocks."

  Further, if we wish to reason about the area or relative direction to a path from some other node such as a building, we must break these large road networks into more manageable pieces. GeoGraphy provides path the network recovery algorithm to segment path network and find paths from one node to the other [3].

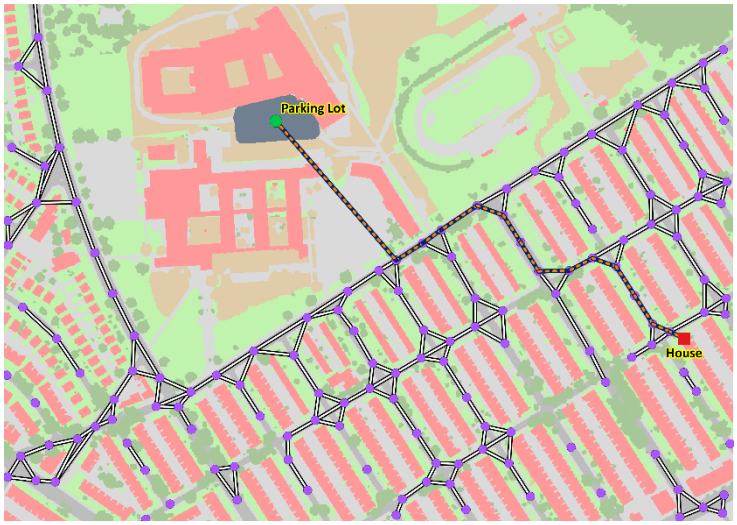


Figure 3. The path found by path-connected search is shown by the dark path that connects the Parking Lot node to the House node using road network segments.

- Connected Component Search- A connected-component search finds connected components of nodes in the SearchGraph, whose edges are controlled by the user-defined query specification. For example, a user may specify a connected-component search including edges only where two nodes are within a specified distance threshold.

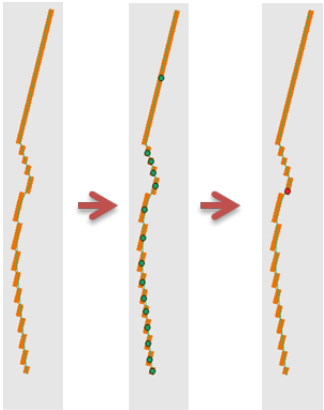


Figure 4. Example connected components search: connected pixels of faults.

- Change Detection: Calculating the magnitude of change, GeoGraphy can detect true changes. Additionally, our graph-based change analysis approach enables a more accurate depiction of complex change events than purely geometric approaches [1].

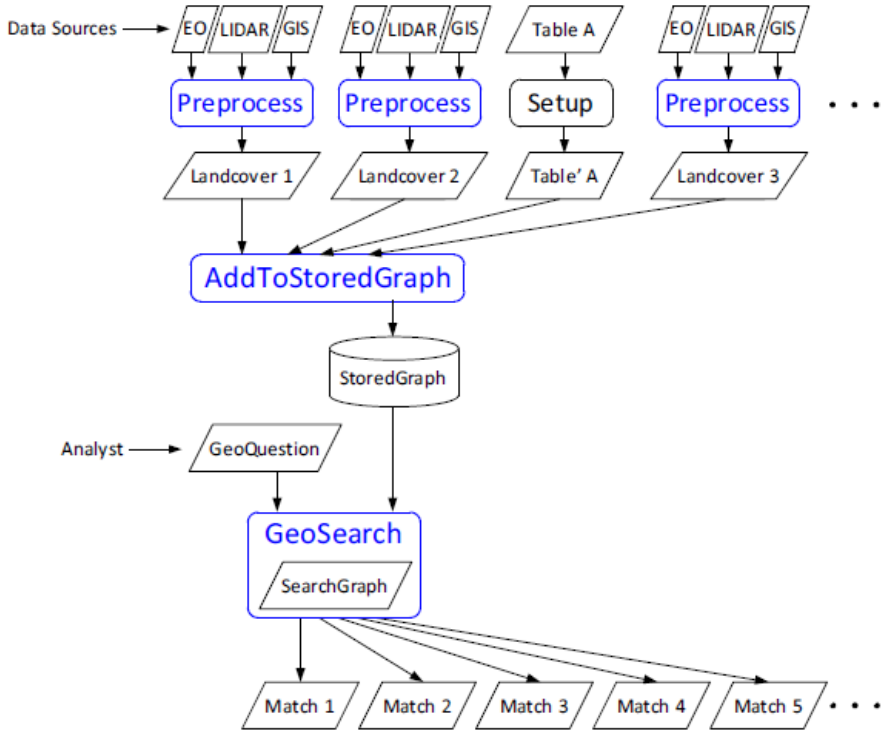
Figure 5. Change detection example : A complex of new buildings, replacing a previous complex.

- Temporal Groups: Searches with edges of temporal proximity including temporal overlap, adjacency and distance are supported.
- Short-Sequence Temporal Ordering: Searches with less than a three step time-sequence could be designed using the star graph search algorithm.

Example Search Algorithms Not Yet Supported by GeoGraphy

- Duration-limited Search - The basic algorithm procedure of the duration-limited search is to first execute a standard GeoSearch, and then visit each of the resulting search results in turn. For each search result, the algorithm sorts the ephemeral nodes of the result in order of increasing start time, and then walks from the beginning of the list toward the end, considering each ephemeral node as a possible match start node. For each candidate start node, the algorithm marches forward in the list (and thus in time), accumulating nodes until the maximum duration limit is exceeded. The resulting node sequence is a maximal-length match beginning at the start node and obeying the maximum duration limit [2]. Pseudo code and algorithms have been developed, but the code is not implemented.
- Multi-step Sequence: Algorithms to look for time-sequence matches with more than three steps are not yet developed.

GeoGraphy Workflow Diagram



APPENDIX B: EVALUATION SLIDES

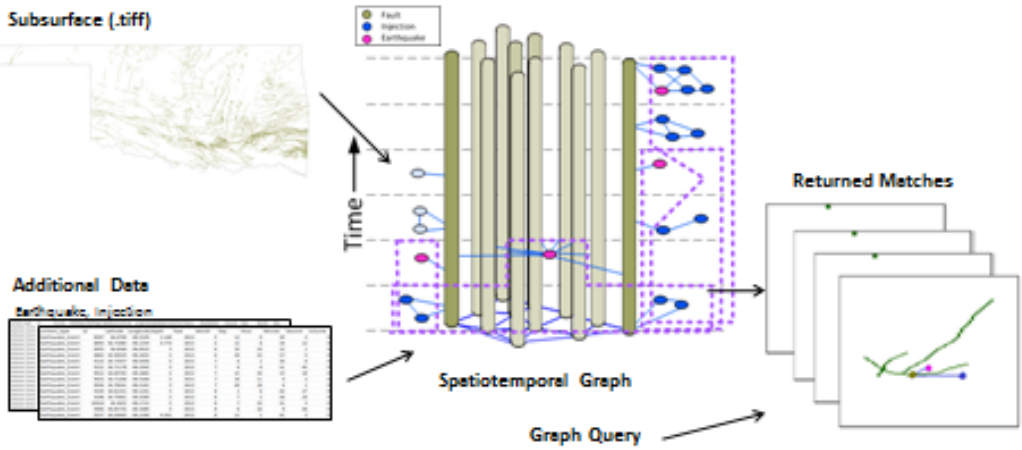
Geospatial-Temporal Semantic Graph Evaluation for Induced Seismicity Analysis

Diane Woodbridge and Randy Brost

June 20, 2016

GeoGraphy Overview

- Geospatial-temporal Semantic Graph Construction
- Search
  - Star Graph – Hub and spokes, with various constraints.
  - Connected Components- Find groups of related items.



: Search earthquake events with preceding injection events along a common fault.

2

Geospatial-Temporal Semantic Graph

- **Strengths:**
  - **Fundamental Strengths**
    - Semantic encoding and search of data.
    - Flexible search design.
    - Identification of higher level semantic ensembles.
  - **Implementation Strengths**
    - Multi-modality input data (image and formatted text file).
    - Search writeback enables hierarchical semantic searches.
    - Match quality ranking.
- **Weaknesses:**
  - **Fundamental Weaknesses**
    - Challenges in discretizing objects and modeling boundary uncertainty.
  - **Implementation Weaknesses**
    - Research code.
    - 3-d analysis not implemented.
    - Limited temporal analysis capabilities.
    - Subterranean feature interactions (e.g., permeability) not implemented.
- **Possible Extensions:**
  - "Cells" to model subsurface uncertainty and boundary uncertainty.
  - Edge weights encoding permeability, etc. and associated algorithms.
  - Temporal causality analysis.

3

Geospatial-Temporal Semantic Graph

- **Related work:**
  - Facility searches [1] : Power plants, tank complex searches.
  - Change analysis [1] : Complex of new buildings.
  - Activity analysis [2] : Finding active businesses.
  - Path analysis [3] : Path partitioning and path searches.
  - Match quality ranking [4] : Elicitation-based, Naive Bayes, and quality metric-based approaches.
- **Relevance (ways can be used for induced seismicity approach)**
  - Semantic representation and searches of big data in the induced seismicity domain.
  - Potential analytical outcomes : Statistical analysis of induced seismicity causal relationships.
- **Software/data input type:**
  - Software : GeoGraphy (SNL Research Code)
  - Input data type
    - Image data : Projected subsurface fault geometry. Multiple images could enable graph modeling of fault change (2-d) over time.
    - Point data : Formatted text file indicating location and time of seismic or human activity events.
- **Workflow (pair with other approaches as well?)**
  - Tensor decomposition : Reduce complexity of graphs.
  - Uncertainty models : Multi-modality data integration, Bayesian
  - Other graphical techniques : Nearest Neighbor condenses input to graphs.
  - Other graphical techniques : Graphs provides input to Neural Networks.
  - Distributed computing/HPC (Task 1) : Improve computational speed.

4

References

GeoGraphy Algorithm

1. Brost, Randy C., William C. McLendon III, Ojas Parekh, Mark D. Rintoul, David R. Strip, and Diane Myung-kyung Woodbridge. "A computational framework for ontologically storing and analyzing very large overhead image sets." In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, pp. 1-10. ACM, 2014.
2. Brost, Randolph C., David N. Perkins, Kristina Czuchlewski, "Activity Analysis with Geospatial-Temporal Semantic Graphs." No. SAND2015-10455. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2015.
3. McLendon, William C., and Randolph Brost. "Path Network Recovery Using Remote Sensing Data and Geospatial-Temporal Semantic Graphs." No. SAND2016-4557. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2016.

GeoGraphy and Uncertainty Analysis

4. Stracuzzi, David J., Randy C. Brost, Cynthia A. Phillips, David G. Robinson, Alyson G. Wilson, and Diane M-K. Woodbridge. "Computing quality scores and uncertainty for approximate pattern matching in geospatial semantic graphs." Statistical Analysis

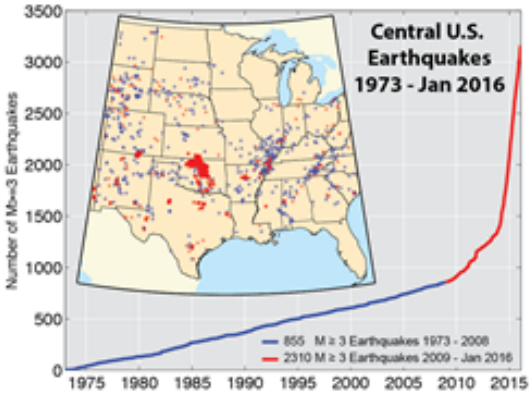
GeoGraphy SubTER Use Case

5. Woodbridge, Diane M., Randy C. Brost, SNL SubTER Graph Analytics, No. SAND2016-5384 W. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2016. https://edx.netl.doe.gov/dataset/snl-subter-graph-analytics

APPENDIX C: PRELIMINARY RESULTS SLIDES

Project Overview

- Evaluating Induced Seismicity in Oklahoma with Geoscience Computing & Big Data (PI : Kelly Rose, NETL)



- Are they natural, or man-made?
- Causes, consequences, risks?

USGS, Earthquake Hazards Program, http://earthquake.usgs.gov/research/induced/ 2

Project Overview

- Evaluating Induced Seismicity in Oklahoma with Geoscience Computing & Big Data (PI : Kelly Rose, NETL)
  - Objectives
    - Develop probabilistic approach that can be utilized to assess potential for induced seismicity impacts.
    - Develop stochastic approaches to reduce uncertainty and constrain subsurface trends.
    - Improve joint analysis of multiple datasets, focus on advancing "Big Data" mining and integration techniques.
  - Tasks
    - Geoscience computing advances for more efficient data management, fusion and accessibility. (NETL, PNNL, ANL)
    - **Development of probabilistic approaches to analyze likelihood of induced seismicity. (SNL, LANL, NETL)**

3

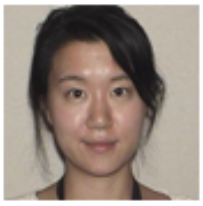
Motivation

- Today's Talk
  - Demonstrate how to use geospatial-temporal semantic graph in the geoscience domain.*
  - Overview tensor decomposition and uncertainty integration in the geoscience domain.

*Caveat: we are not geoscientists and the graph construction and search criteria are chosen to demonstrate GeoGraphy capabilities.

Team Members

Kristina Czuchlewski (PM)



Diane Woodbridge (PI)

Randy Brost

Danny Dunlavy

David Stracuzzi

Point of Contact

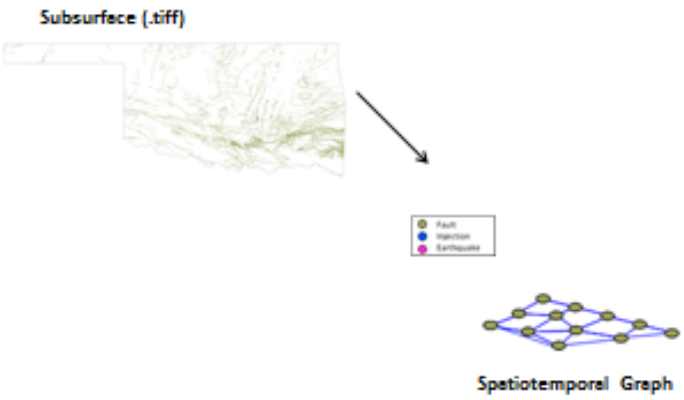
- Overall Projects (Kristina Czuchlewski)
- Geospatial-temporal Semantic Graphs (Randy Brost)
- Tensor Decomposition (Danny Dunlavy)
- Uncertainty Analysis (David Stracuzzi)

Overview

- GeoGraphy Concept
  - Graph Construction
  - Search
- GeoGraphy Results
- Additional GeoGraphy Capabilities
- Tensor Decomposition
- Uncertainty Analysis

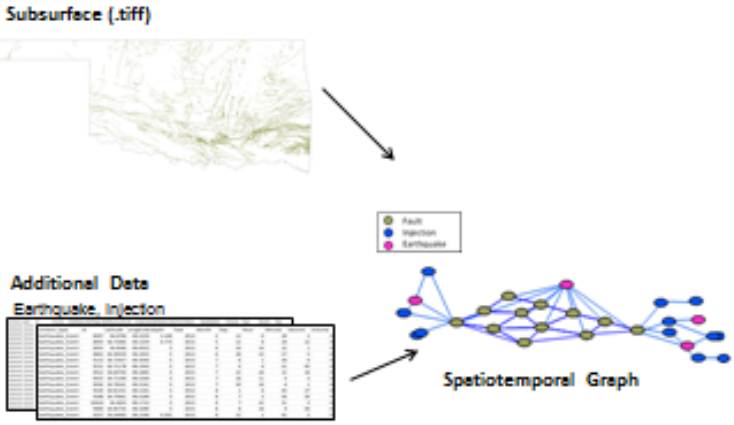
GeoGraphy Overview

- Geospatial-temporal Semantic Graph Construction



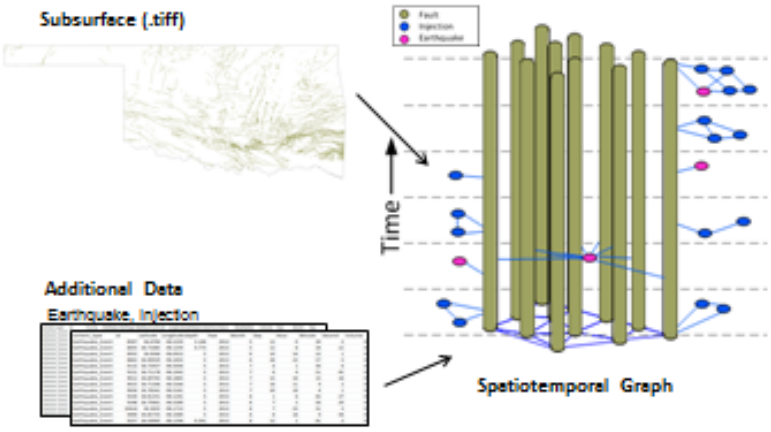
GeoGraphy Overview

- Geospatial-temporal Semantic Graph Construction



GeoGraphy Overview

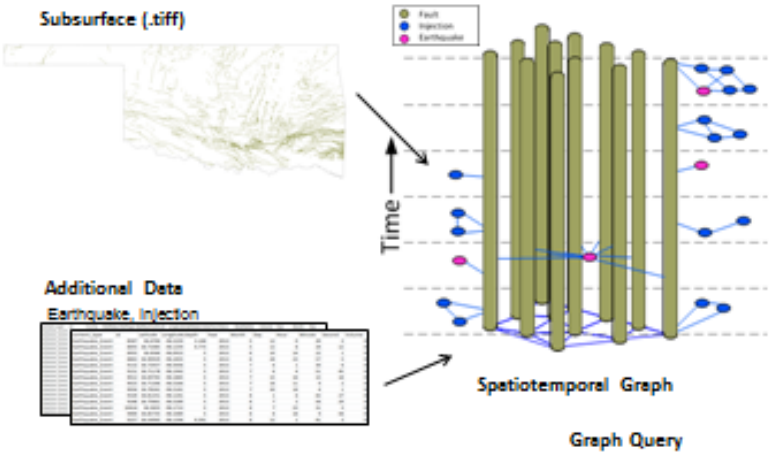
- Geospatial-temporal Semantic Graph Construction



10

GeoGraphy Overview

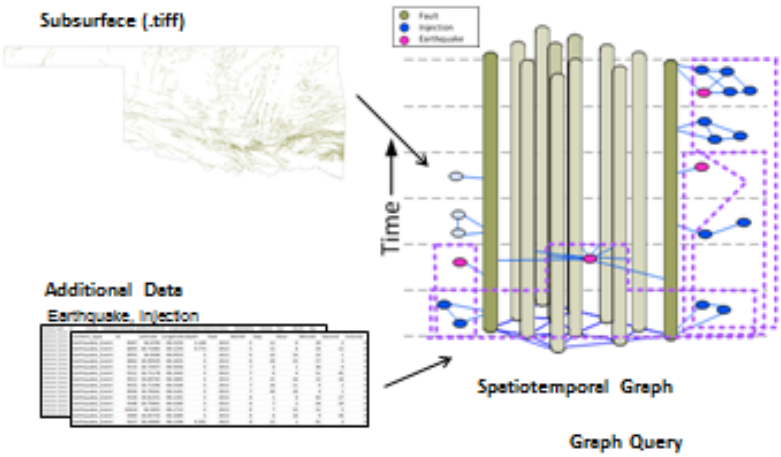
- Geospatial-temporal Semantic Graph Construction
- Search
  - Star Graph – Hub and spokes, with various constraints.
  - Connected Components- Find groups of related items.



11

GeoGraphy Overview

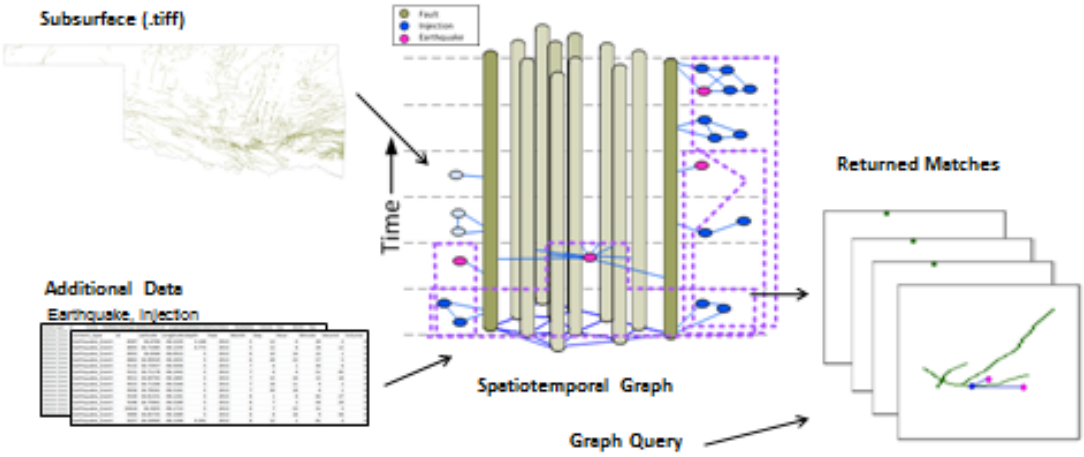
- Geospatial-temporal Semantic Graph Construction
- Search
  - Star Graph – Hub and spokes, with various constraints.
  - Connected Components- Find groups of related items.



12

GeoGraphy Overview

- Geospatial-temporal Semantic Graph Construction
- Search
  - Star Graph – Hub and spokes, with various constraints.
  - Connected Components- Find groups of related items.



13

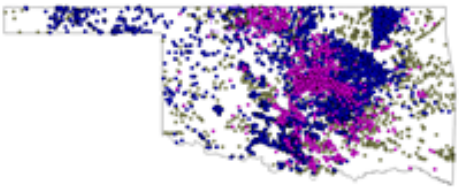
GeoGraphy Capabilities

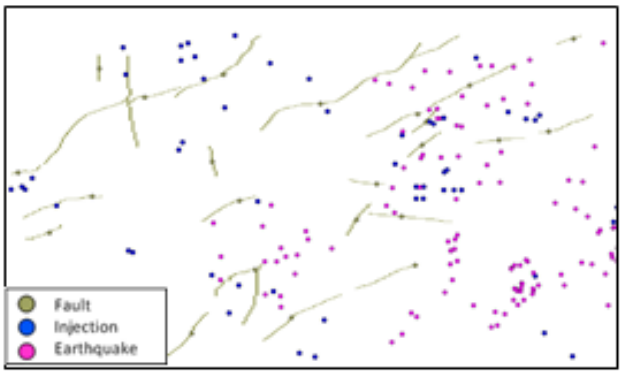
- Efficient data representation using spatial-temporal semantic graph.
- Multi-modality in a single uniform search representation.
- Search for semantic concepts, in image data, ensembles of point data, etc.
- Allow search goal variation in topology, shape, metric properties.
- Graph search enables multiple algorithms: star, transitive closure, path search, duration-limited, clusters, etc.
- Search across space and time: Change analysis, temporal clusters, short-sequence temporal ordering, duration-limited, multi-step sequence.
- Match quality ranking
- Foundation for temporal analysis.

Graph Construction using OGS data

- Utilized Data
  - Subsurface fault – x, y location only.
  - Injection – Timestamp, location, volume, well type/status, etc.
  - Earthquakes – Timestamp, location, depth, etc.

Graph Construction using OGS data



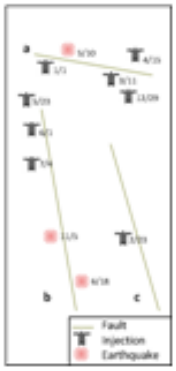


- Distance edges are implicit (calculated on demand)

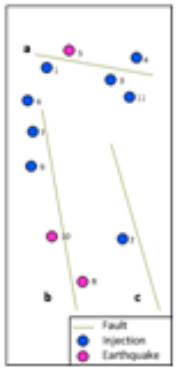
GeoSearch - Hypothesis Scenario

- Do injection events near faults cause earthquakes?*
- Seek earthquakes happening after injection events, along a common fault.

A simple example:



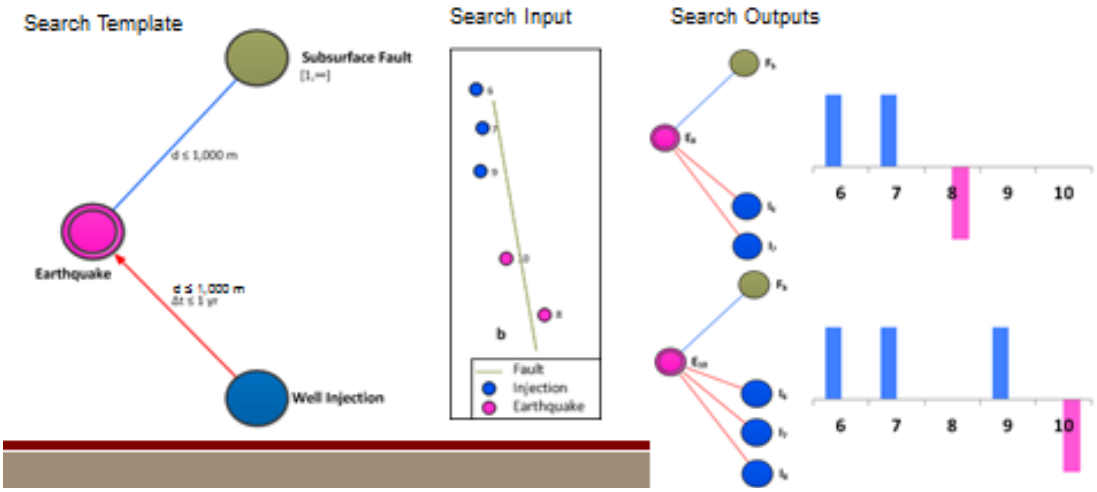
Simplified rendering :
(sequence instead of timestamps)



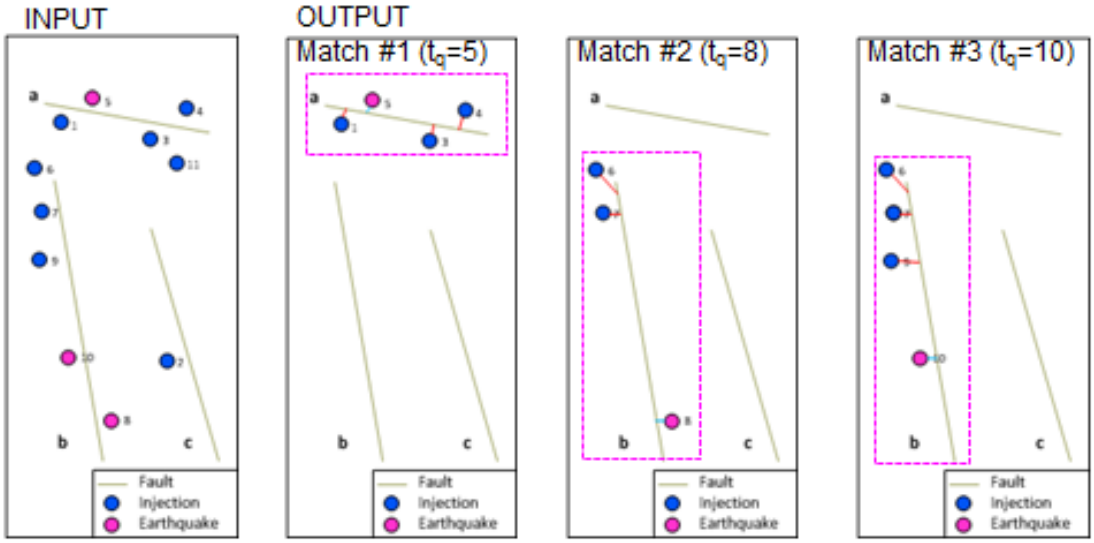
*Caveat: we are not geoscientists and the search criteria are chosen to demonstrate GeoGraphy capabilities.

Query Design

- Find an earthquake with preceding injection events along common fault(s).
  - Star Graph Search
    - Hub : Earthquake



Example Search Outcomes



Overview

- GeoGraphy Concept
  - Graph Construction
  - Search
- GeoGraphy Results
- Additional GeoGraphy Capabilities
- Tensor Decomposition
- Uncertainty Analysis

20

Graph Construction using SubTER Data

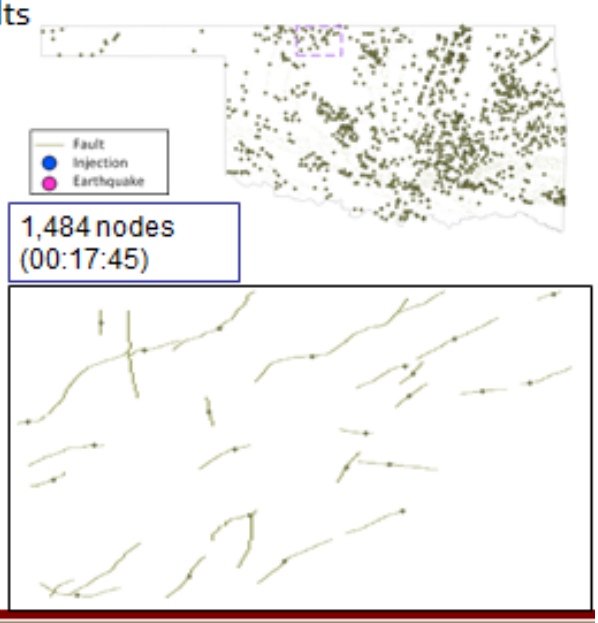
- Subsurface Faults (.tiff)

*Assumption : No 3D data for now.

21

Graph Construction using SubTER Data

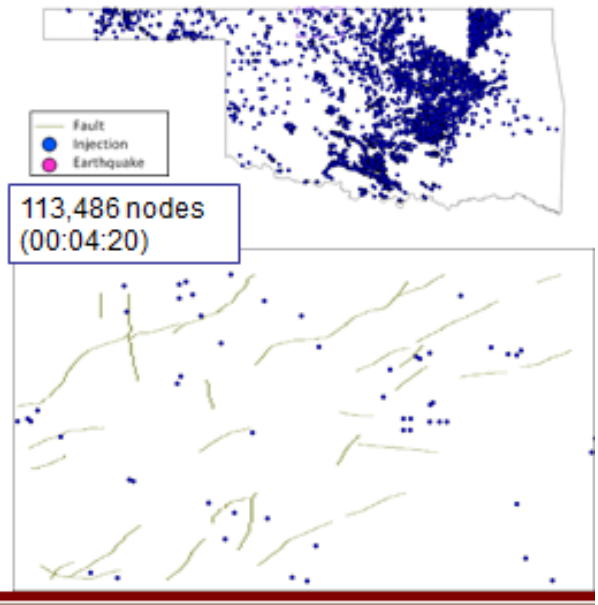
- Connected Subsurface Faults



22

Graph Construction using SubTER Data

- Point Data
  - Injection Data (2011)

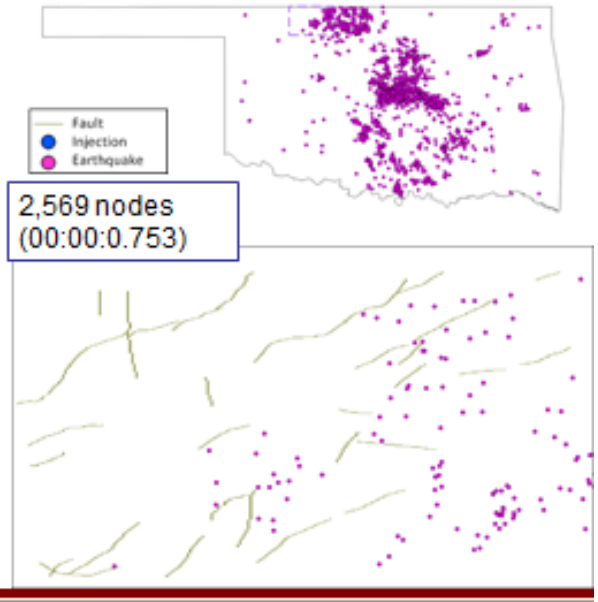


*Assumption : No 3D data for now.

23

Graph Construction using SubTER Data

- Point Data
  - Earthquake (2011-2012)

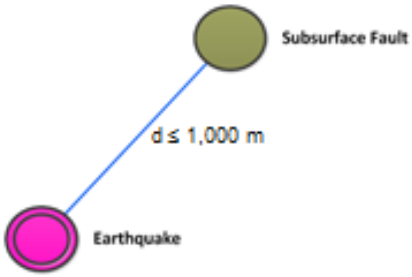


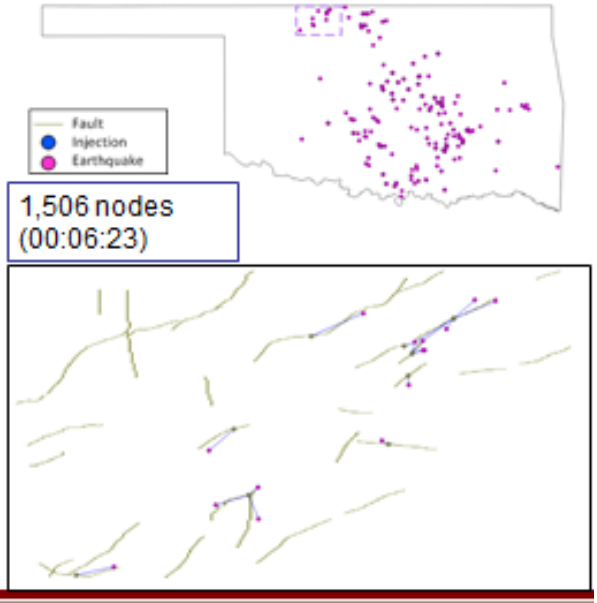
*Assumption : No 3D data for now.

24

Graph Search

- Earthquake Near Faults

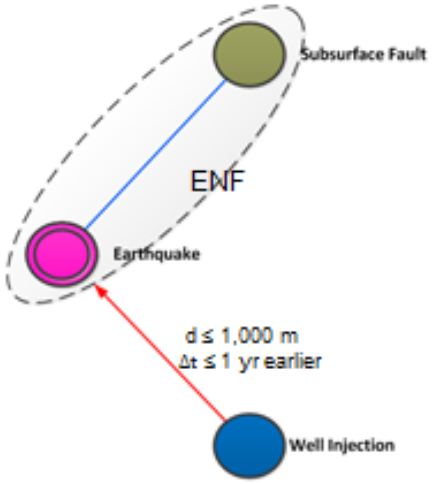


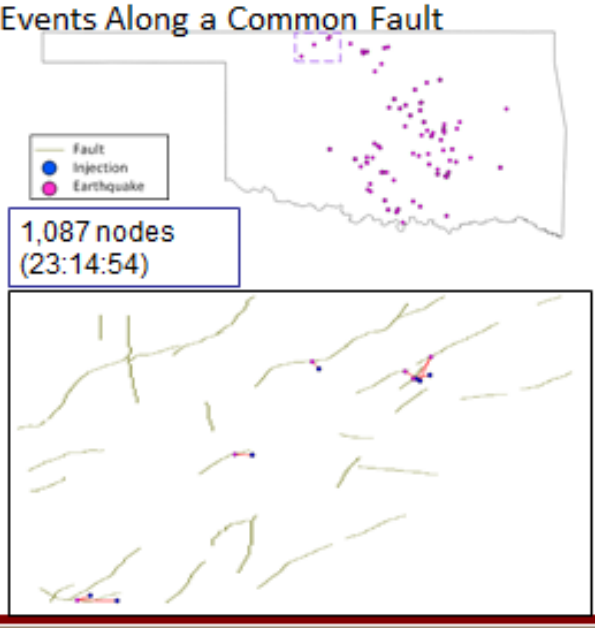


25

Graph Search

- Earthquake Near Injection Events Along a Common Fault





Overview

- GeoGraphy Concept
  - Graph Construction
  - Search
- GeoGraphy Results
- Additional GeoGraphy Capabilities
- Tensor Decomposition
- Uncertainty Analysis

Additional GeoGraphy Capability

- Partitioning connected faults
- Cascading induced seismicity detection
- GeoGraphy UI trial
- Metric criterion for sorting matches
  - Example
    - Largest quake
    - Largest injection
    - Largest injection closest to quake
    - Most permeable fault, etc.

Requires Extensions

- Add the z dimension
- Weight on nodes/edges, based on permeability, etc.
- Temporal Analysis (Randy's Event Prediction LDRD)
- Representing subsurface uncertainty

28

Overview

- GeoGraphy Concept
  - Graph Construction
  - Search
- GeoGraphy Results
- GeoGraphy Possible Extension
- Tensor Decomposition
- Uncertainty Analysis

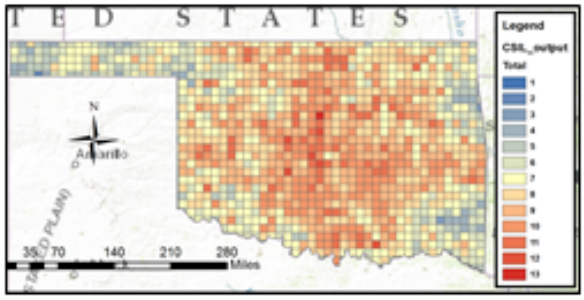
29

Analysis of Temporal Data using Tensor Decompositions

- **Tensor decompositions:** Higher order analog of Principal Component Analysis (PCA)
  - Models count or continuous sparse and dense data.
  - New research focuses on combinations of data types.

- **OK CSIL Data Analysis:** count of features contributing to induced seismicity
  - Aquifers, urban areas, rivers, wells, mines, dams, lakes, heavy rain, earthquakes, etc.

- **Goal:** Incorporate temporal information and perform sensitivity analysis.



Cumulative Spatial Impact Layers. Jennifer R. Bauer, et al., NETL-TRS-2-2015.

30

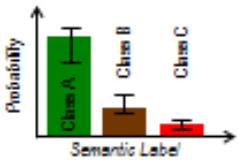
Overview

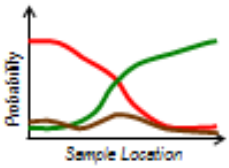
- GeoGraphy Concept
  - Graph Construction
  - Search
- GeoGraphy Results
- GeoGraphy Possible Extension
- Tensor Decomposition
- Uncertainty Analysis

31

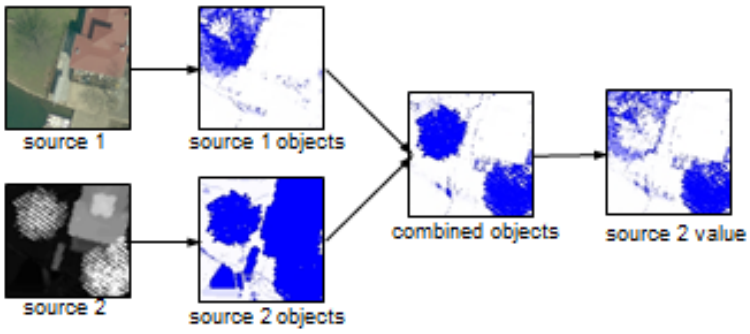
Multimodal Data Integration Under Uncertainty

1. Evaluate uncertainty associated with "objects" in data
   - Labels
   - Spatial extent
   - Temporal extent





2. Combine data from multiple sources
   - Analyze uncertainty associated with each source
   - Combine uncertainties across sources
   - Derive the value or impact of each source to some target question



Distribution

4 National Energy Technology Laboratory
Attn: Kelly Rose
1450 Queen Ave. SW
Albany, OR 97321

1	MS0519	Kristina Czuchlewski	Org. 05346
1	MS0735	Erik Webb	Org. 06910
1	MS0750	Hunter Knox	Org. 06911
1	MS0754	Susan Altman	Org. 06915
1	MS1326	Diane Woodbridge	Org. 01461
1	MS1326	Randy Brost	Org. 01462
1	MS1327	David Stracuzzi	Org. 01462
1	MS1327	Danny Dunlavy	Org. 01461
1	MS1327	John Wagner	Org. 01462
1	MS1327	Ron Oldfield	Org. 01461
1	MS0899	Technical Library	9536 (electronic copy)

Sandia National Laboratories