Unsupervised Event Detection in Long Horizon Timeseries Data

Gabriel P. Andrade[1], Jose Cadena[2], Goran Konjevod[2]
[1]University of Colorado Boulder, [2]Lawrence Livermore National Laboratory

DATA SCIENCE SUMMER INSTITUTE

Lawrence Livermore National Laboratory

MOTIVATION

- In many domains we have an abundance of extremely long timeseries data
- Often this data is unlabeled, inconsistently labeled, or improperly labeled
- Often contains subtle and noisy information about recurring events
- Examples abound in medicine, seismology, the stock market, and cosmology

RESEARCH QUESTIONS

- Can we detect distinct subsequences in timeseries of a repeating event?
- Can we do this when events are temporally disparate?
- Is this possible using an unsupervised algorithm?
- If so, how do we know a detected event is meaningful?

DATA SET

- Waveforms collected from Geophones at 500Hz continuously over one year
- Three temporally distinct DAG experiments recorded at NNSS
- Data collection along three axes at 6 spatially separated stations
- Timeseries of roughly 15 billion samples per year per station per axis
- Potential events for detection are repeated activity surrounding experiments

05 / 20 / 2018 05 / 29 / 2018

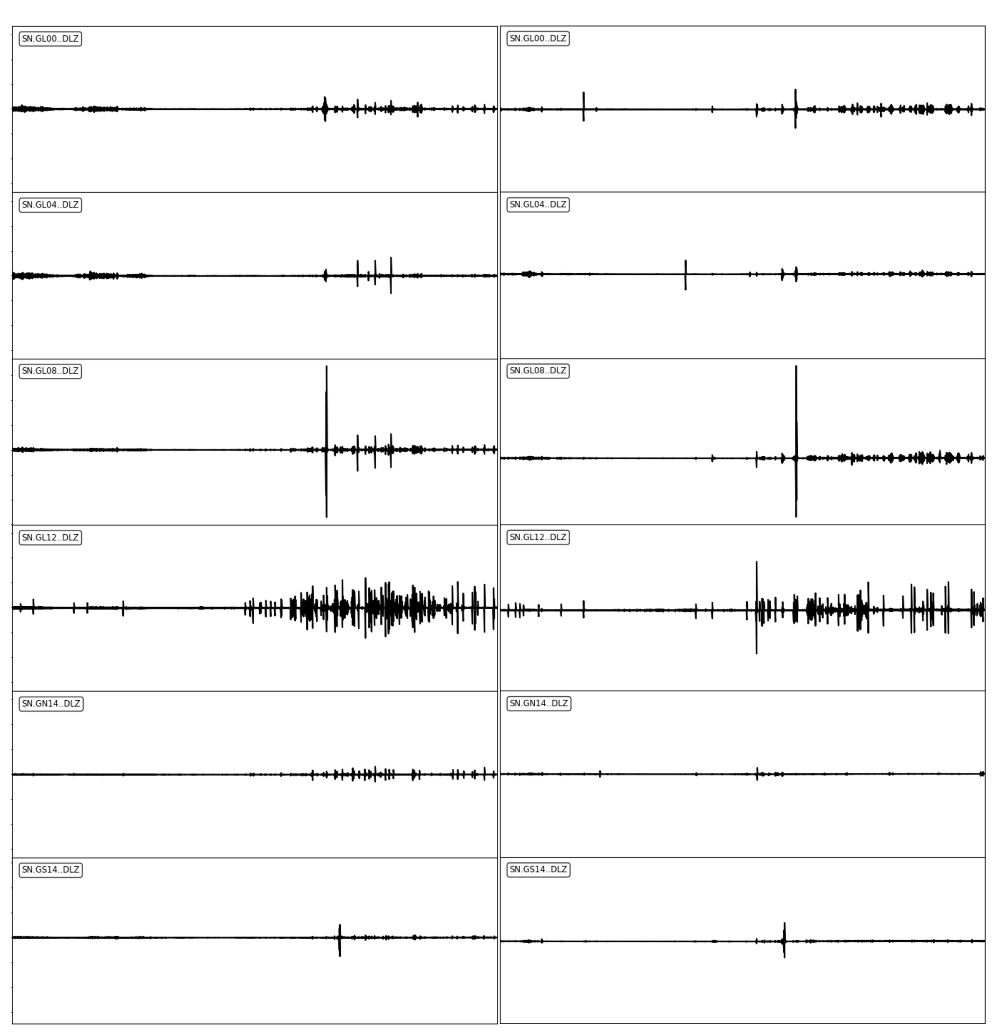
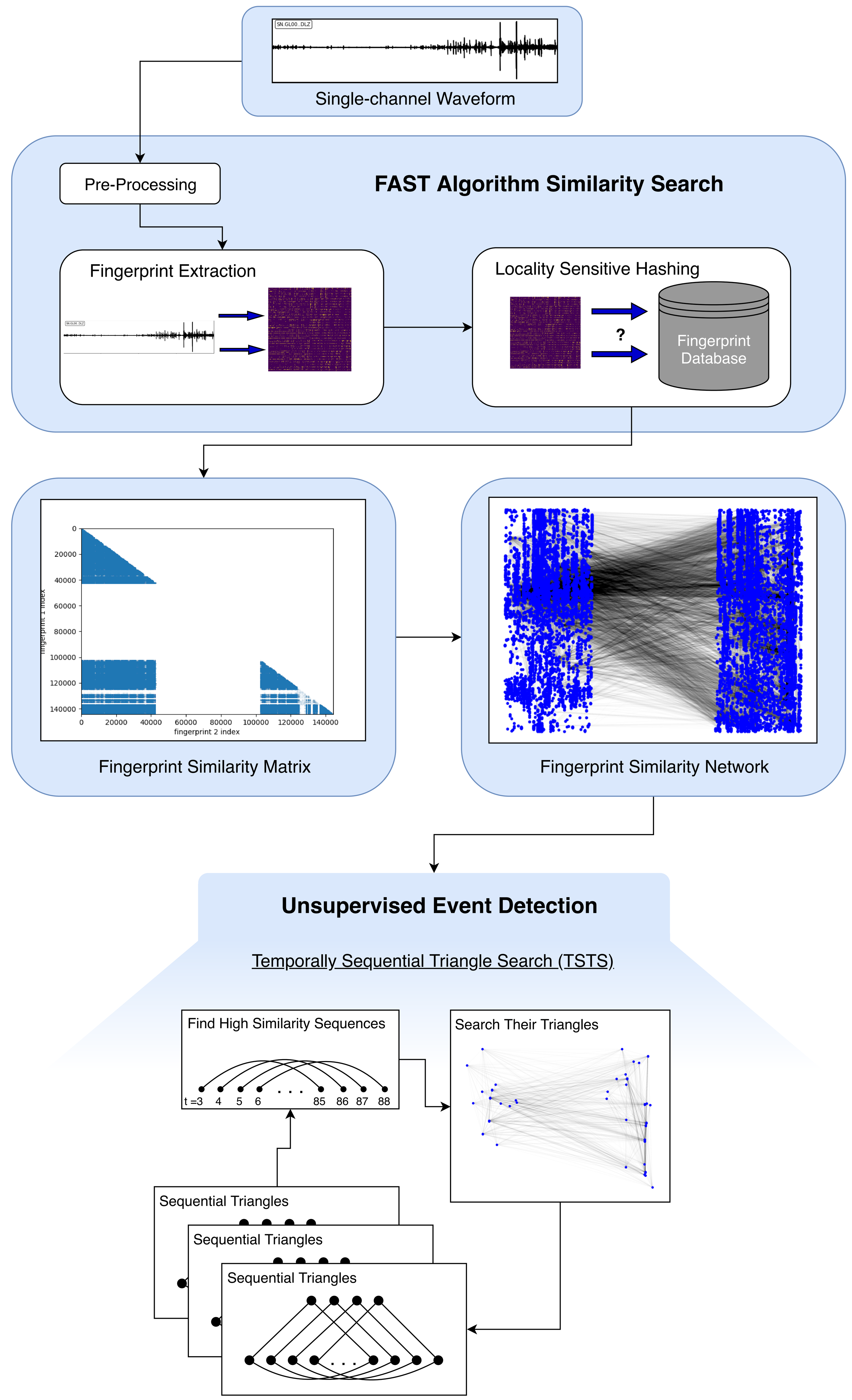


Figure 1: Example of raw signals along an axis. Rows correspond to individual stations and columns correspond to a specific day (~43 million points per station). Notice how different they are.

DATA PROCESSING PIPELINE



SINGLE-CHANNEL TSTS

Input:

- Fingerprint Similarity Network (G)
- Minimum sized sequence to constitute event (T)
- Threshold on initial "high similarity" matches (θ)
- Threshold on candidate triangles (τ)

1. Form thresholded adjacency list of G with a node's neighbor included if and only if their similarity is greater than θ
2. Recurrently search the resultant adjacency list for maximal length temporal sequences with greater than T "high similarity" adjacent node pairs. All such sequences are stored as a "potential event"
3. **for** each "potential event" **do**
4. **for** each "high similarity" pair in the "potential event" **do**
5. Create a list of nodes forming triangles in G with said pair. Only include nodes with similarity greater than τ to *both* nodes in the "high similarity" pair
6. Recurrently search the resultant lists for maximal length temporal sequences with greater than T triangle forming nodes. All such sequences of triangles are stored as a detected event

Output: Set of detected events and the ordered fingerprints found to be in each event

EXTENSIONS

- Multi-Station TSTS
- Ground Truth Query TSTS
- Temporally Sequential N-Clique Search

FUTURE DIRECTIONS

- Considering the multi-scale setting
- Incorporating network statistics to mitigate false detections
- Improvements by incorporating weighted variant of the "Onion Decomposition"

REFERENCES

1. Bergen, Karianne & Yoon, Clara & C. Beroza, Gregory. (2016). Scalable Similarity Search in Seismology: A New Approach to Large-Scale Earthquake Detection. 301-308. 10.1007/978-3-319-46759-7 23
2. Bergen, Karianne & C Beroza, Gregory. (2018). Detecting Earthquakes over a Seismic Network using Single-Station Similarity Measures. Geophysical Journal International. 213. 10.1093/gji/ggy100
3. Hébert-Dufresne, Laurent & Grochow, Joshua & Allard, Antoine. (2016). Multi-scale structure and topological anomaly detection via a new network statistic: The onion decomposition. Scientific Reports. 6. 31708. 10.1038/srep31708

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-784739