EXPERIMENT OF MAPPER ALGORITHM ON HIGH-DIMENSIONAL DATA IN MICROSEISMIC MONITORING

A Thesis

by

WEIHAO DING

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	John Killough
Committee Members,	Eduardo Gildin
	Maria Barrufet
Head of Department,	A. Dan Hill

August 2017

Major Subject: Petroleum Engineering

ABSTRACT

The objective of this research to utilize data driven methods to analyze microseismic monitoring, especially using Topological data analysis (TDA) with limited physically based approaches. Python Mapper (PM) is the tool of TDA for this study. Microseismic data has great characteristics of big data. Previous studies suggesting stage-by-stage microseismic analysis also avoid the limitation of current software, which can only process slightly over 10,000 data points. During this study, more TDA packages are constantly evolving to handle larger and more complex data such as Betti Mapper by Spark.

PM is a tool by combining topology principles and machine learning methods into an integrated data analytic implementation. The high-dimensionality of microseismic data practically limits what classical statistical analyses can achieve. Machine learning techniques such as dimensionality reduction are required for such datasets. Where PM stands out is its ability to retain the raw feature of data set when machine-learning algorithm is applied.

The first portion of the study is to observe the data point relation of microseismic data entirely and stage-by-stage. Dividing attributes into location and signal data reveals the relation within and between two different data types.

The main discovery from location data of network is the high density areas are tend to be earlier events and could locate where high pressure start to build up, or the origins of the fracture networks. Origins that are far apart in the beginning grow into

each other to result in one (most of the time) or more (rarely more than two) networks. The fracture growth with complex directions of extensions can be represented with a much simpler, single-directional network. Signal data reveals location-specific data quality trends. These trends are hardly visible if attributes are investigated in pairs but obvious when mapped altogether. Locational and geological characteristics may be an explanation, but this needs further information to prove the observations. In fracture growth softwares, these trends will allow researchers to ignore the location of the wellbore and focuses at the actual origins of the fracture network. An override including discontinuity of the network and confidence of stimulated reservoir volume could be manually added to improve the accuracy of the fracture simulation.

A sensitivity analysis to PM parameters is carried out to test the robustness of the method and comparing raw data clustering method to prove the effectiveness and benefits of using TDA. TDA is a great method for data preprocesses, analyses, and has virtually infinite possibility, but should never be the end of a project. The results from PM could be used as input for many other studies.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my research and thesis advisor, Dr. John Killough for his guidance throughout the study. His insight and vision in data science enable me to approach application of petroleum engineering from a different perspective.

I would also like to thank my committee member, Dr. Eduardo Gildin and Dr. Maria Barrufet for their time and advice on my work. Their reviews further solidify my focus point of the research by providing more directions on attacking the problems.

Many thanks to my research team, Dr. Yan Bicheng, Dr. He Jie, Dr Wang Ying, Cao Yang, Chai Zhi, An Cheng, Masoud Alfi, Tang Hewei, Guo Xuyang, Li Ning, and Yu Jinhui for their help during my time in and out of the office.

I have enjoyed the amazing two years of graduate studies and my previous experience at Texas A&M University always reminds me of being a loud and proud member of Fighting Texas Aggies Class of 2015!

To my parents, thank you for your unconditional love and support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor John Killough and Professor Eduardo Gildin of the Department of Petroleum Engineering and Professor Maria Barrufet of the Department of Chemical Engineering.

All work for the thesis was completed by the student, under the advisement of Professor John Killough of the Department of Petroleum Engineering.

Funding Sources

Graduate study was partly supported by graduate student scholarship from the Department of Petroleum Engineering, Texas A&M University.

NOMENCLATURE

TDA	Topological data analysis
PM	Python Mapper
PCA	Principal component analysis
MA	Mapper algorithm
LDA	Linear discriminant analysis
SVD	Singular value decomposition
MDS	Multidimensional scaling
SRV	Stimulated reservoir volume
SNR	Signal-to-noise ratio

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

The current world is driven by the data. Harvard Business Review regards data scientist the hottest vocation in the 21st century. Big data has become the standard issue in many industries and the oil and gas is no exception. Big data is referred to data sets having size and complexity that are beyond traditional data processing to deal with. Many different methods and techniques have been developed to analyze these data. A quick search in SPE's OnePetro reveals over 10,000 (a relative small number compared to other areas) technical papers in the topic of big data. One of the emerging methods is topological data analysis (TDA), which has been widely utilized in the field of health and medical science, national securities, and finance. However, the application of TDA is essentially untouched in traditional exploration and production business. There are very few research papers about this topic, but showing great potential in many areas. The motivation of the research is to further develop and discover more ways to process data with TDA. During the shale boom, many data became available obtained during exploration, drilling, completion, and production, but were allocated to little time to be analyzed due to different priority for companies at that time. Low oil price slowed down field activities, and the industry started to realize the presence of large amount of undigested information and the hidden meaning behind it. As a result, it is a great time to look back what we have but not understand in our data arsenal. This could provide great insight for the next cycle for oil and gas industry.

CHAPTER II

BACKGROUND

2.1 Data Science in Oil and Gas Industry

While data science and analytics has been widely adapted in many different industries, oil and gas is not one of them. Halsey et al. (2017) pointed out that for years there has been a cultural gap between data science and technical petroleum professionals. This industry is widely considered conservative, and most of time, researches follow physical based approaches and are ambivalent to data-driven results. Furthermore, the ultimate reason of introducing a technology is to provide value. The value is hard to judge considering a project will experience all departments in real world such as geology and geoscience, information technology, engineering, business, data, and etc.

Data like in most elements on the planet shows a behavior of a pyramid model. Datasets that are expensive usually provide more information and potential values and vice versa. For instance, microseismic monitoring, extensive core measurement and high-resolution gyro survey are likely to be only performed on limited number of wells. Combining the previous points, the industry is having a difficult time to justify developments of data-driven methods because the uplifts from them are hard to validate.

2.2 Data Preprocessing

The “garbage in, garbage out” principle needs critical attention in any data analysis. The quality of the data depends on various factors such as sources, accuracy,

complexity, reliability, etc. Table 1 (Famili, A. et al., 1997) displayed some of the major problems the data scientists and data analytics are facing every day.

Table 1. Different types of data problems in the real world

Too much Data	Too Little data	Fracture data
- Corrupt and noisy data - Feature extraction - Irrelevant data - Very large data size	- Missing attributes - Missing values - Featureless/Discrete data - Small amount of data	- Incompatible data - Multiple data sources - Data from multiple levels of granularity

Additionally, in order to separate relevant data to specific projects, it is essential to prioritize certain attributes in high-dimensional point cloud data for effective analysis. In this study, TDA algorithm needs normalized data for either equal weight among all attributes or intentional bias towards certain data type. Data transformation techniques such as filtering, ordering, editing, noise modeling can be achieved by traditional statistics methods, machine learning algorithm or even TDA.

2.3 Key Methodology, Topological Data Analysis

Data, especially in great size, has shape. Shape matters, which in most cases is proved by experience or examples rather than theories. Since shape is a nebulous notion, we need a way to properly and formally measure and represent it. Topology, in the most general term, is the study of shape, and shapes in topological space present two

important properties, connectedness and compactness. The first practical application of topology originated in the 1700s when the problem of the Seven Bridges of Konigsberg was raised. In the problem, lands/islands and bridges were represented in a compressed manner by dots and connections respectively shown in Figure 1. The entire geographic features are equivalent to the topological network. This is based on the three concepts of topology so that the information can be analyzed in such way.

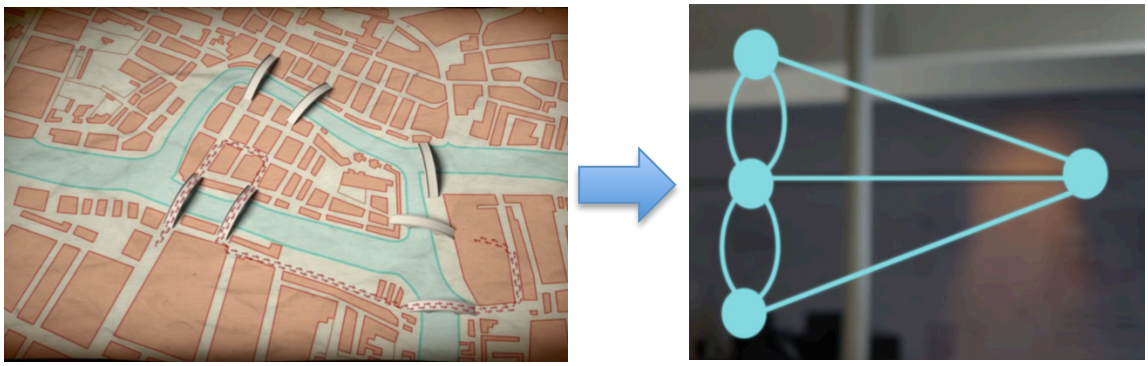


Figure 1. The actual geographic of the Seven Bridges of Konigsberg can be represent with a simple network with useless information neglected. Two pictures are equivalent in topological space after deformation (Reprinted from Carlsson, 2013)

1. Coordinate invariance: the properties of shape measured in topology will not change if the coordinate is different; this means rotating the shape in a certain coordinate system without changing its properties
2. Deformation invariance: the shape can be "stretched" and "squashed" as long as the feature is kept. One example is the homeomorphism between A and R, where

we can see both letters contain a loop with to flare attached. As a result, these two letters have no differences.

3. Compressed representation: In the Seven Bridges of Konigsberg problem, many details of a single land piece are compressed into a single point, similarly bridges into connections. Although information such as the area of the island, lengths of bridges, and etc. are not included, the represented plot helped us to solve the problem by retaining the most essential features

Point cloud from data will show features such as clusters, flares, or loop when applying topological techniques, such as Mapper algorithm (MA). This algorithm developed by Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson from Stanford University, can extract simple description of high-dimensional datasets in the form of simplicial complexes. This process is also known as dimensionality reduction. Apart from TDA methods, dimensionality reduction can also be achieved by machine-learning techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA). One of the most successful implementation is Ayasdi, where there are many different commercial versions for different industries. Python Mapper (PM) written by Daniel Müllner and Aravindakshan Babu (Danifold) is our choice of TDA due to the advantage of being open-source. The detailed algorithm used in the software is described in the next section.

2.4 Mapper Algorithm

The algorithm starts with the selection of distance metric, preparing for the distance matrix. The distance matrix is used to capture the similarity among the data points. In the PM, three metrics are included: Euclidean, Chebyshev, Minkowski. Euclidean metric is the least computational expensive and it will be a good starting point for the current research. Tools that are more powerful such as Ayasdi core include hamming, cosine, and correlation metrics for different applications. The next step is computing filter functions, also known as lenses in TDA. This will map out the multi-dimensional data points for each entry into a single value on a real line (1D representation). The filter function can be based on raw, statistics, or geometric features. In PM, we focus on geometric features such as eccentricity, kernel density, and distance to the nearest point especially considering many datasets encountered in petroleum engineering are location sensitive. Detailed calculations are available in the algorithm inventing paper by Singh et al. (2007). PM also has detailed formulas for each lens in the implementation.

For a point cloud data X with N points, and if the distance between two points $x, y \in X$ is $d(x, y)$, the eccentricity function is

$$E_p(x) = \left(\frac{\sum_{y \in X} d(x, y)^p}{N} \right)^{\frac{1}{p}}$$

where the range of exponent meets the condition $1 \leq p < +\infty$.

The kernel density function is

$$f_\varepsilon(x) = C_\varepsilon \sum_y \exp\left(\frac{-d(x,y)^2}{\varepsilon}\right)$$

where C_ε is the normalizing constant. $\varepsilon > 0$ is known as the bandwidth and it determines the smoothness of the density function. A larger bandwidth will result in a smoother version of one with a smaller ε (Silverman, 1986).

Two distance filters, kNN-distance and distance to measure are included in PM. The distance to the k-th nearest neighbor is an inverse measure of density. The distance to measure function is

$$f(x) = \sqrt{\frac{1}{k}\sum_{j=1}^{k} d\left(x, v_j(x)\right)^2}$$

where $v_1(x), \ldots, v_k(x)$ are the k nearest neighbors of x in the data set (Chazal, 2010).

Distance matrix eigenvector lens in PM returns *k*-th eigenvector of the distance matrix. This function is the key step of singular value decomposition (SVD), suggesting PM has a primitive combined utilization of TDA and machine learning. The experiment carried in the study only involved these three functions, but many other lenses can be applied to the established MA. In the case of a Gaussian distribution, eccentricity function is negatively correlated with kernel density. This property is useful when different datasets require different lenses but the same or similar feature output. PM also provides build-in distance to *k*-th nearest neighbors and graph Laplacian with extra dependency named `cmappertool`.

After the filter values are calculated, they can be presented with histogram in the PM software, allowing for modification or evaluation on the filter before the complete

algorithm is carried out. Next, we can apply cover (also known as resolution) and overlap to the filtered values on the real line. The number of covers and the percentage of overlap are the variables to be altered to obtain the desired output i.e. to show the most wanted features in the shape. For instance, a large degree of overlap combines more similarity between each overlap, the output will tend to have fewer clusters, and vice versa, a small degree of overlap results in more different overlaps, hence more clusters and more discreteness among clusters. If multiple filter functions are used, a higher dimensional parameter space is needed for covering and overlapping. In most cases and in the PM, only single lens method and 1D-cover is considered and it has been proved sufficient to solve complex problems in many fields of study. Most applications and demonstrations in the ongoing research in various industries predominately use 1D covering; some of the examples are uniform 1D cover and balanced 1D cover.

Performing clustering is the next step. Each data point forms a single cluster in the beginning in hierarchical clustering and clusters form larger clusters when the most similar pairs are combined. PM can implement single-linkage, complete-linkage, average-linkage, weighted-linkage, median-linkage, centroid-linkage, and Ward-linkage clustering with the assist of Python package `fastercluster` (Müllner, 2013). The time saving with `fastercluster` is significant, ranging from 50 to 100 times faster than calculation without the package in this study. In the end, we can build the entire TDA network, joining them into shapes, different colors and etc. In this network, nodes represent collections of similar data points from the input point cloud and edges connect the nodes that share the same data points. Coloring (Figure 2) by different raw attributes

or the filter function can help us to localize behaviors and derive hidden information. Problem localization can greatly improve the efficiency when working on complex data.

Low value High value

Figure 2. The node color scheme used in PM output for the entire study, with a scaling value from 1 to 10 used in summary. The coloring scheme applies to all attributes values and filter/lens value.

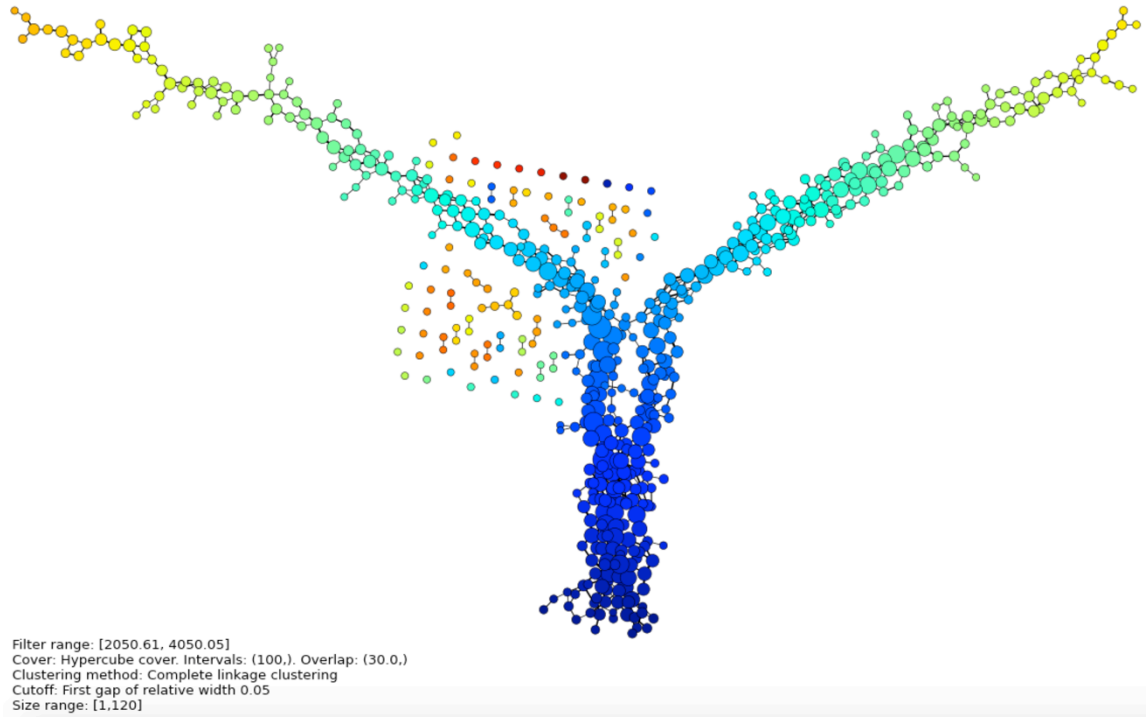


Figure 3. The first PM output generated with raw microseismic data in this study as an example network.

2.5 TDA and Machine Learning

In a high level of generality, many people will agree TDA is an example of machine learning technique. The foundation of TDA is algebraic topology, to some extent, pure math. The method and general framework do not need to "understand" the data. A TDA method such as MA and persistent homology can be generally used in any type of data. However, machine learning is more specific, with detailed and isolated algorithm so that the method actually studies the data and obtains useful information. This requires stronger evidence for the method to read and interpret. Furthermore, a wrong algorithm may fail to recognize important insights or even worse to derive invalid results. Because of all these points, selecting the optimal algorithm is crucial but improbable when thousands of alternatives are available currently and requires complete evaluation. MA is a single computational tool that belongs to TDA. It models a data network output in order to be interpreted by human or machine. At last, the lens in the algorithm can be a machine learning function such as PCA and SVD. Ayasdi claims that the generality is one of the biggest advantages over machine learning.

The generality of TDA means the output from its methods may still need to be interpreted for better understanding. As a result, the combination of TDA and machine learning can be much more powerful than apart. For instance, the MA utilizes clustering to group similar data points into a single node. TDA makes machine learning methods much more effective, comparing with clustering raw data rather than data filtered by

TDA lenses. As a core step in building topological network, clustering is performed to segments data and then combines all the portions into the final network representation.

Dimensionality reduction algorithm such as PCA and multidimensional scaling (MDS) usually leaves projection loss issue, wherein data points that are well apart in original dimensions overlap in reduced dimensions eventually (Ayasdi, 2016). TDA has the ability to perform clustering at the original dimensions, and the feature of clusters will be carried out after running the algorithm.

2.6 Microseismic Monitoring

Microseismic mapping during hydraulic fracturing is a common practice to measure the location, complexity, and strength of underground movement. Measurements from the technology not only provide reservoir information, but also help enhance production ultimately. A microseismic event triggered by hydraulic fracturing from the treatment well is "heard" by the geophones installed on the monitoring well. The data acquired by the geophones will tell us when and where the event occurs, and how big the event is. Microseismic monitoring is used to interpret geometries of fractures and stimulated reservoir volume (SRV) (Maxwell et al., 2002). The calculation of SRV is critical as it is a direct correlation parameter for well performance. This means that it is important to derive the dependency between SRV and fracture network growth by studying raw microseismic data.

Many microseismic service companies have in-house software to interpret raw data to obtain necessary stimulation parameters such as fracture azimuth, fracture

network length, height and volume and etc. The smallest unit of microseismic data study is stage, where there may be multiple sources accounting for all microseismic events within each stage. There have been many studies on determining locations of hypocenters with rock mechanic along with the microseismic density volume, which can be created by counting the number of events in established given grid (Maxwell, 2011). Most microseismic analyses are based on geophysics or geomechanical model. There is little information of interpreting by statistical or data driven methods.

Microseismic monitoring has been a good production indication statistically, although without very strong foundation of physical model. Many researches showed more than 90% correlation between two data types. With a small number of data points located in Horn River, production correlation has an approximate R squared value of 0.98 with total events, 0.76 with SRV, 0.96 fracture area (Snelling, 2014).

2.7 Previous Application of TDA

Not only did the algorithm inventing paper detailed the implementation of the MA and filter functions within, which are mentioned in the earlier section, but it also presented a simple application to synthetic data as a reference for future studies. One of the great examples in the paper to demonstrate the three key concepts in topological space is that a running horse and static horse will share virtually the same network representation. Lum et al. (2012) utilizing their own Ayasdi package successfully analyzed three datasets in three very different industries, namely genetics, political science, and sport performance. With their result, they proved that TDA is more

sensitive than PCA, MDS, and cluster analysis in detecting both large and small-scale patterns, resulting more refined conclusions than those drew from standard procedures.

TDA is a great tool to recognize similarity and dissimilarity. This capability has great potentials in reservoir engineering and geology. Alfaleh et al. (2014) showed that it could compartmentalize reservoir later validated by a conventional reservoir-modeling simulator in two separated cases (Brillig and Norne). Selecting the correct combination of input data, metrics and lenses, is the key to successfully extract the meaningful information and often requires extensive experiments. The study also successfully implemented pressure difference, a physical petroleum engineering function, as a lens in MA. On the geology side, TDA was able to identify lithofacies in a Marcellus shale gas formation for geoscientist (Cortis, 2015). TDA revealed that the traditional three geological groups were insufficient to capture all the information and the differences that may be extracted with more logging data. Node network features such as flares could be visually captured and provided more guidance in grouping (an improvement to 12 definite groups). Four geological layers of studies and four groups of high total organic content regions were clearly identified after apply TDA. Some structures could not be captured by classical classification methods but TDA in this formation. This new classification was also applicable to more geological structures other than shale-gas unconventional reservoir. In summary, although both studies did not aim to be universal, the implementation suggested virtually infinite possibilities in E&P researches.

In additional to the published paper, Kraft (2016) combined the conventional statistic methods and MA to analyze marketing research and work wage datasets in his

thesis. The same classification technique was applied to both raw data directly and the shape of the data generated by MA. It showed that the combination of two methods could preprocess the data and localized the problems especially when the large data size made the initial analysis confusing and difficult. MA was also able to recognize shape features despite the presence of noise. Ayasdi mentioned the possibility and potential benefit to combine machine learning with TDA and Choi's (2016) thesis utilized the numerical result from the TDA and combined the power of support vector machines to predict well production by raw microseismic data. It was also possible to apply MA to newly generated variables in additional to raw data. Although some statistical problems were present such as result-distorting multicollinearity, this also pioneered another way to quantitatively analyze the features of the TDA result. Choi pointed out that it was possible to utilize MA to scrutinize microseismic events stage by stage, as it might reveal information on how the fracture network grow over time during the stimulation.

In summary, TDA is a great modeling framework that many of its possibilities are being discovered and utilized in many industries (Carlsson, 2017):

1. Geometric features in the dataset usually represent groups of data set for future application to subset.
2. The TDA network can identify the most explanatory or the driven factors for the group
3. It is possible to investigate relations among different attributes, distributions of data value, and behavior of lenses by different coloring schemes for data matrix

(row = data points, column = features). These also allow a broad hot spot analysis.

4. Using different lenses creates multiple network models simultaneously to observe how the same group of data locates or behaves differently in different models.
5. Refined linear regression can be achieved to solve prediction or optimization problems.

CHAPTER III

METHODOLOGY AND PROCEDURE

3.1 Objectives

Based on the previous researches on TDA and their suggestions, the purposes of this study include listed below:

1. Map stage-by-stage microseismic data to observe any features that are obvious using TDA. TDA results show features that can be further graphed using TDA or analyzed by conventional statistical method. TDA reveals affiliations of microseismics events within stages, creating multiple possible networks in each stage. Future fracture growth modeling could utilize observations from the PM networks as additional reference.
2. Develop a framework for typical TDA used in petroleum engineering when high dimensional data is available. In this study, there is a hierarchy in the data obtaining, meaning the workflow will be different between datasets with limited varieties and ones with extensive categories.
3. Identify locational and geological characteristics in order to be proven and justified by physical model when more information is available.
4. Calculate and calibrate regression model with TDA to predict future production according to initial production and feedback from PM outputs.

3.2 Area of Interest

In this study, the area of interest is Permian Basin, specifically the University Lands wells located in the Midland Basins. Some of the reasons for this selection include:

1. Permian basin is a major and will be the largest contributor of oil and gas in the nation as well as the world. There are more than 200 operators active in the area.
2. There are approximately 9,000 producing wells currently, and 21,000 drilling location identified. This suggests good data availability and value of the research for future endeavor.
3. A great portion of the data is public information to avoid any confidentiality issues.

Three wells with microseismic data were picked as the subjects of study, for these wells were monitored in much longer intervals compared to other cases (13 to 15 stages versus three to four stages). All three wells are located in the Spraberry Trend area, and completed in the Wolfcamp formation. Table 2 lists the common attributes that are available for all three wells (Here denoted as A, B, and C). Well A has the largest number of events, and monitored exclusively. Well B, C and two other wells are monitored by the same setup. Only four stages are monitored in the two extra wells each well, thus eliminated them for PM analysis, due to smaller size of datasets. A complete study on Well A was carried out to start the experiment, and a sound explanatory result

is discussed in this chapter. More summarized results from other wells are completed in the next chapter without details.

Table 2. Available attributes for Well A

Measurement	Description
MS_LOC_SNR	Location signal to noise ratio (SNR)
NOISE_LEVEL	Average RMS of noise level
PSH_AMPL_RATIO	Compressional amplitude/Shear amplitude (P/Sh)
QC_DISTANCE	Distance from center of sensor array to microseismic event
QC_LOC_X	GIS location longitude, ft.
QC_LOC_Y	GIS location latitude, ft.
QC_LOC_Z	Vertical depth, mean sea level corrected, ft.
SP_MAGNITUDE	Moment magnitude of microseismic event
SP_MOMENT	Seismic moment, the size of an earthquake

3.3 Overall Information

Well A was completed in Wolfcamp B formation, at a maximum vertical depth of 8365 ft. and kelly bushing of 2881 ft. The monitoring setup was shown in Figure 4, generated by the service company along with Figure 5. Stage 21 to 33 out of 33 stages were monitored. It is preferred to keep the raw data for analysis to observe any dependencies in the first place. However, data preprocessing is required to extract hidden information. Raw microseismic data is different depend on service companies provided and the parameters that are being measured.

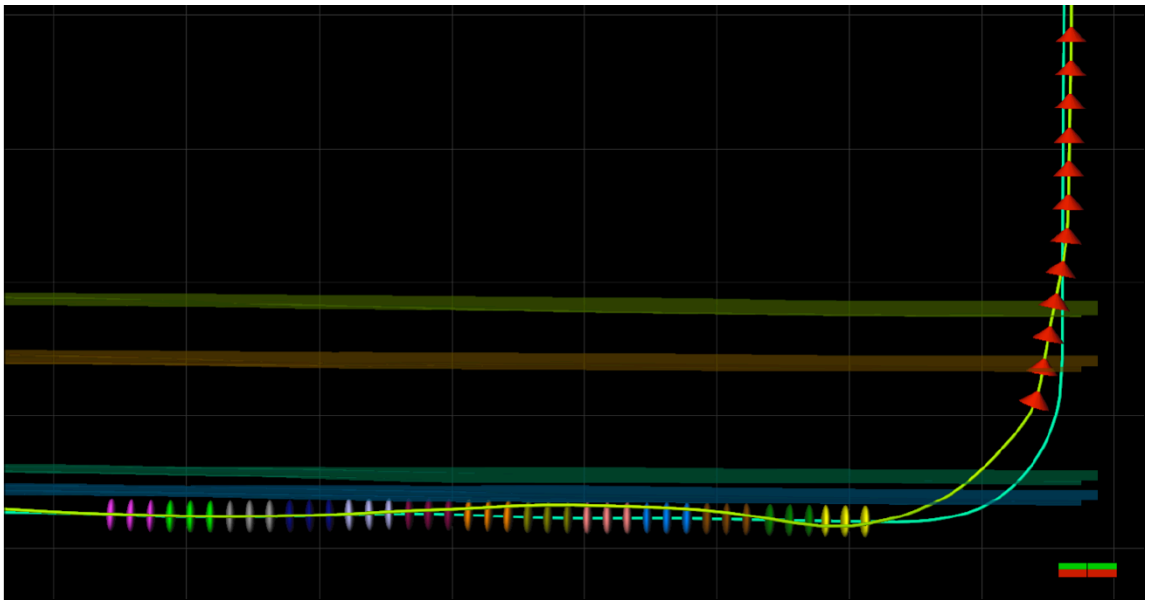


Figure 4. Side view of Well A setup monitoring

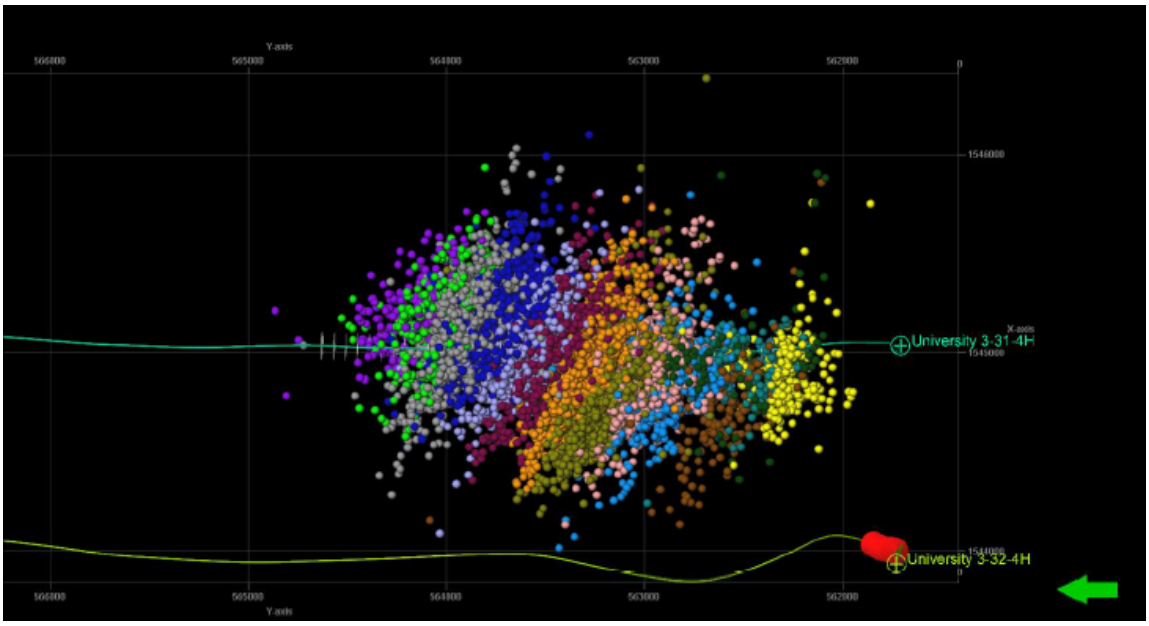


Figure 5. Map view of Well A microseismic events

Previous TDA study on microseismic did not apply the algorithm directly to raw data. To begin with, event ID is ignored due to lack of physical meaning other than the timely order of events. For entire well data analysis, stage number is also deleted while single-stage data will be investigated separately. Source attributes magnitude and moment are highly correlated and as a result, moment is left out of the TDA input to eliminate problem of multicollinearity (Here multicollinearity is used loosely as the relationship between two attributes is not linear). The slightly prepared dataset contains eight attributes/dimensions in 7162 cases. Since visualization of the problem is obvious by human perception, it is easy to notice that data points with large noise are located in the last few stages, where contains a smaller Y location value. The quality of the data in these stages is questionable. The entire study shares the same color scheme shown in Figure 2 to represent low or high values of colored lenses or attributes.

After visualizing all eight dimensions in colors, we observe another highly correlated input: QC_LOC_Y and QC_DISTANCE, due to the monitoring location, see illustration in Figure 6 (Figure label from 1 to 8 are MS_LOC_SNR, NOISE_LEVEL, PSH_AMPL_RATIO, QC_DISTANCE, QC_LOC_X, QC_LOC_Y, QC_LOC_Z, SP_MAGNITUDE respectively).

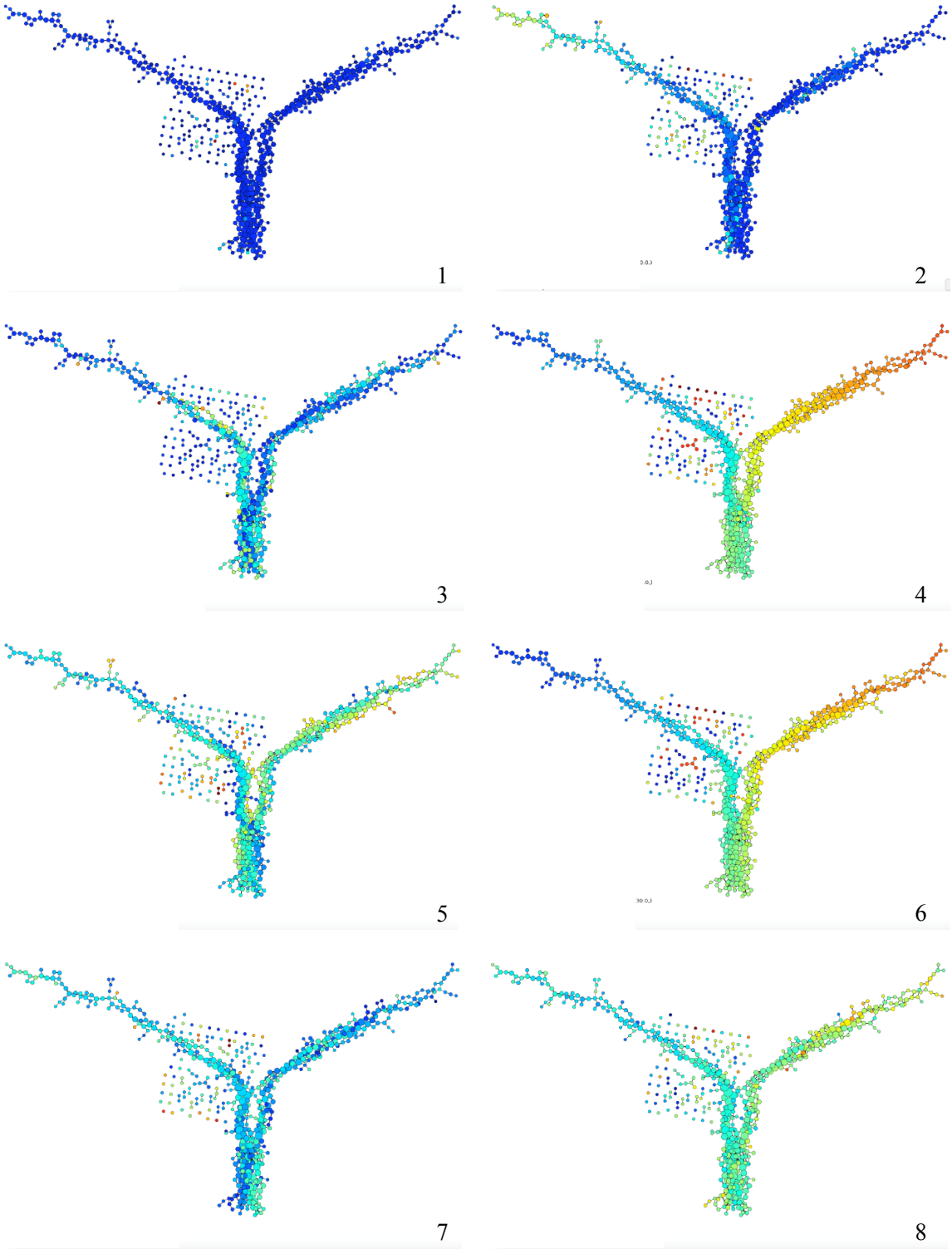


Figure 6. Well A PM output colored by 8 attributes

When distance is also omitted, the TDA result still presents the similar features of three flares, with one flare having a small loop feature. We also realized due to the large raw value from the location attributes, the shape features were largely driven by those attributes, also known as explanatory attributes.

3.4 Separating Location and Signal Data

The raw microseismic data contains two major types of data, location data (X, Y, Z coordinate) and event signal data (SNR, noise, P/Sh ratio, and magnitude). Since these two types of data value do not have the same unit, or are not in the same order of magnitude, we are able to utilize this feature to emphasize either type of data. It is possible to observe the features we need by magnifying the respective data values. Another method is to divide them to investigate separately in the first place, then combine and couple the results from each study to obtain a comprehensive conclusion.

3.5 Location Data Observation

Location data was interpreted first to observe any features. When stage-by-stage data was run by the PM, the obvious problem was large variance of the number of data points across different stages. This required adjustment of PM parameters to obtain the most revealing shape features. It became obvious after a few trials that a network build on density lens worked better with greater data number and eccentricity lens worked better with smaller data number. High microseismic frequency regions can be

recognized with nodes either with high filter value (red nodes) in density lenses or low filter value (blue nodes) in eccentricity functions depicted in Figure 7. For all 13 stages monitored, the network presents the trend of stimulation volume growth. In location analysis, 3D data point cloud is easily perceived. For all stage, data points are clearly clustered in two different depths. This suggests that the geology in these two depths were favorable to hydraulic fracturing.

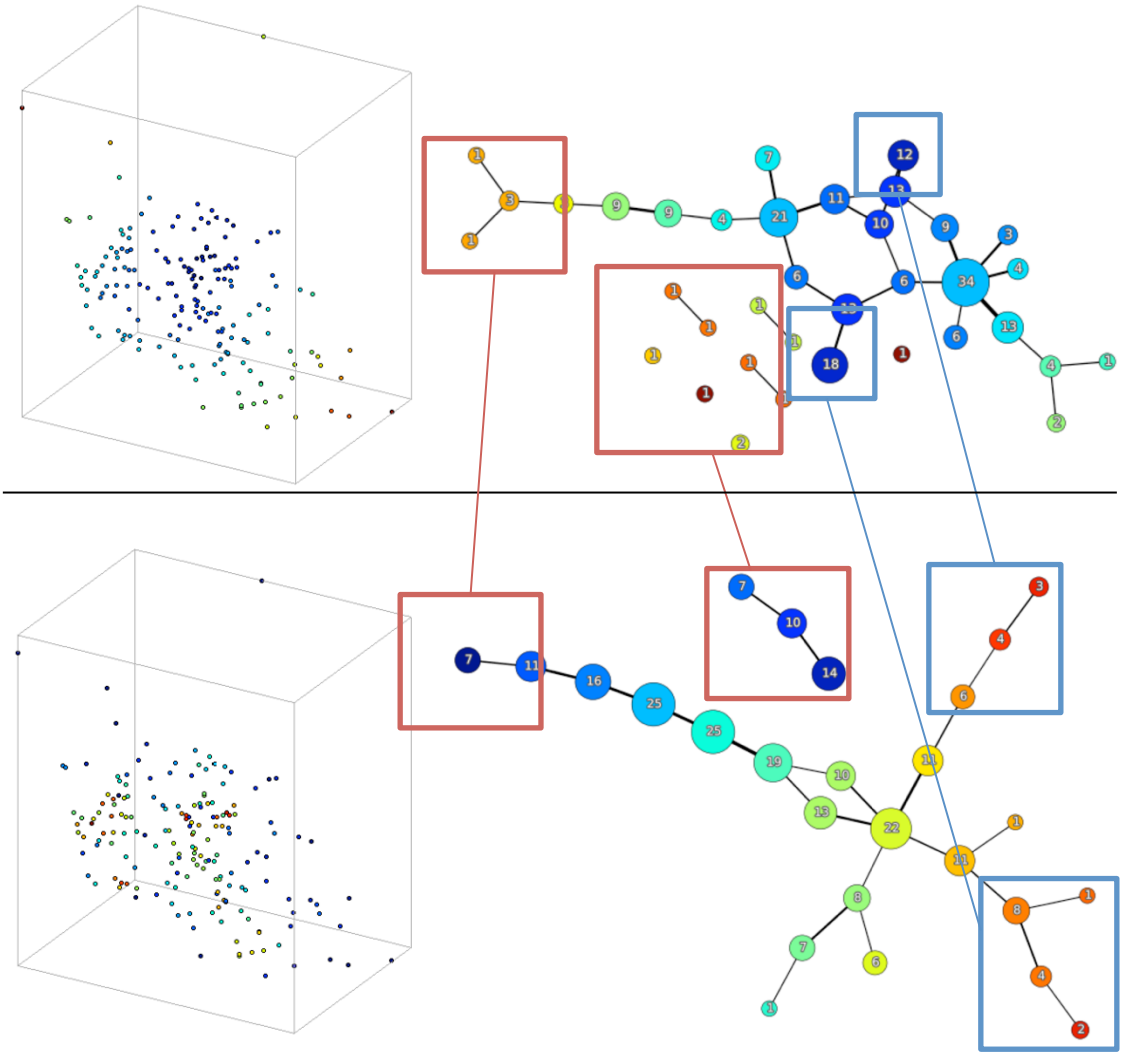


Figure 7. Stage 21 event networks with eccentricity lens (top) and kernel density lens (bottom). Rectangles with the same color point out the corresponding nodes that represent the similar data points. The color representation in the networks mapping is the same as in the event location plots

Distance lenses calculate local density depend on the number of neighbors that are chosen. For a synthetic shape, it is easy to perceive that the torus in Figure 8 has denser data points at inner radius and sparser at outer radius. However, without the

adjustment of smoothing (kernel density has this capability), it is difficult to cluster high filter value shown as in Figure 9. As a result, two distance lenses are not considered for location data.

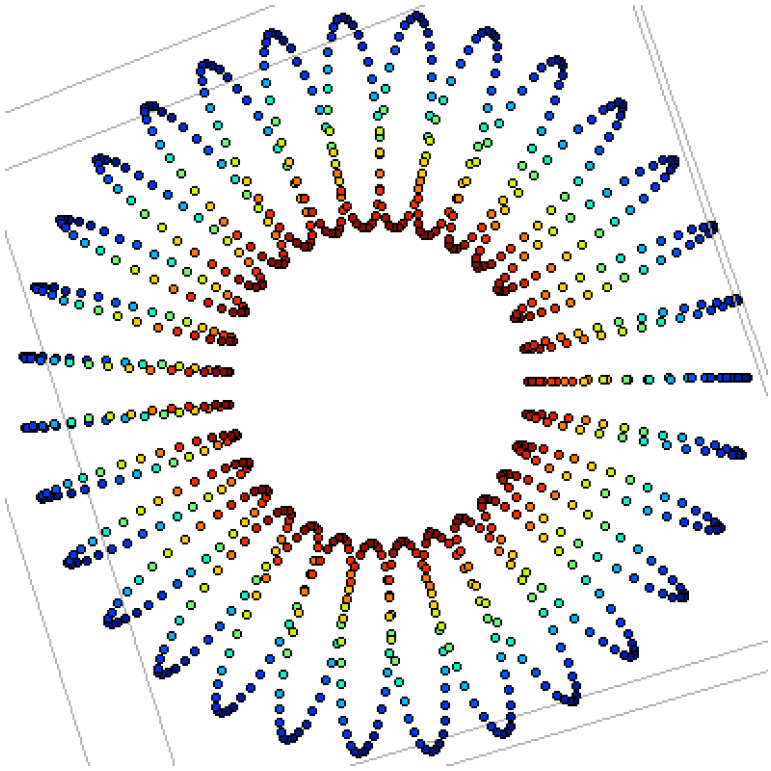


Figure 8. The torus has high filter value or high point density at inner radius, and low filter value at outer radius.

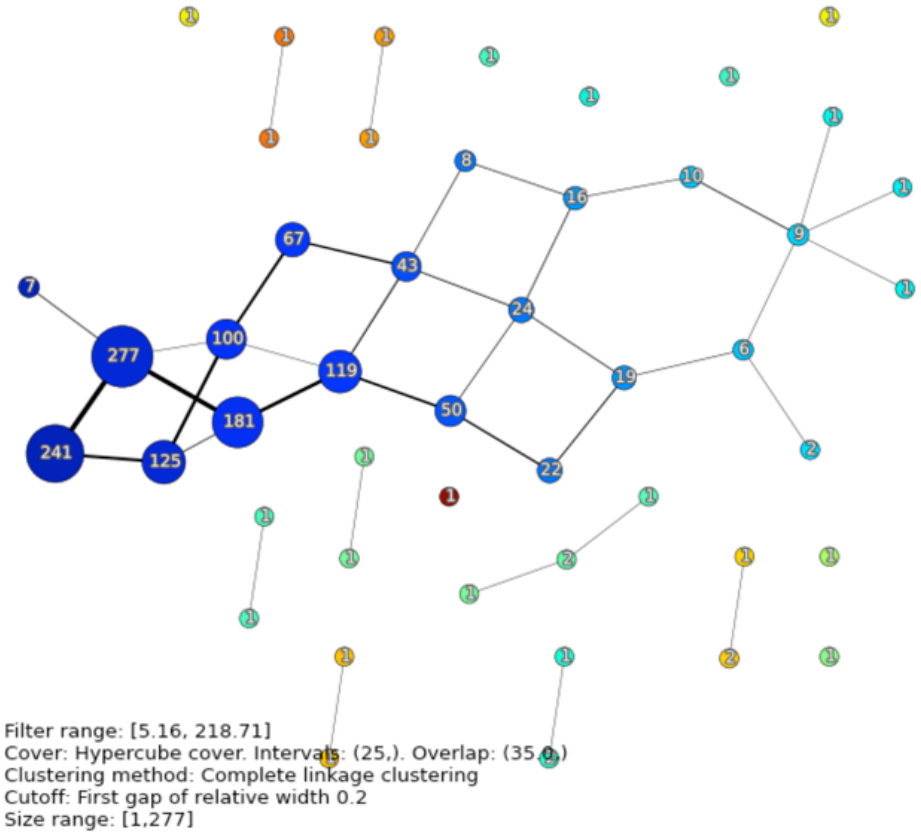


Figure 9. The high filter value points are not clustered due to high degree of difference.

Other than the filter value, it is anticipated that the color of the node would represent hidden insights that required some physical background to be extracted. The time and frequency of event occurrence in relation of the location may reveal the sequence of the stimulation, as one of the main purpose of microseismic is to evaluate the hydraulic fracturing.

According to Maxwell (2011), it is possible to quantitatively describe the high data point density area, obtaining the similar results from gridding method to establish

microseismic active volume. The activeness normally suggests the fracture development, although this may overestimate or underestimate due to a few reasons:

1. The stress from hydraulic fracturing may activate events that occurred outside of the connected fracture network or in isolated fractures, resulting in overestimation.
2. A generally low SNR (lower than 6) decreases the confidences; microseismic volume needs to include some uncertainty in the calculation, resulting in a larger point cloud volume (Figure 10).
3. It is widely written that microseismic monitoring is better picking up shear opening signal than compressional opening signal, causing some omissions in fracture activities, thus underestimation.

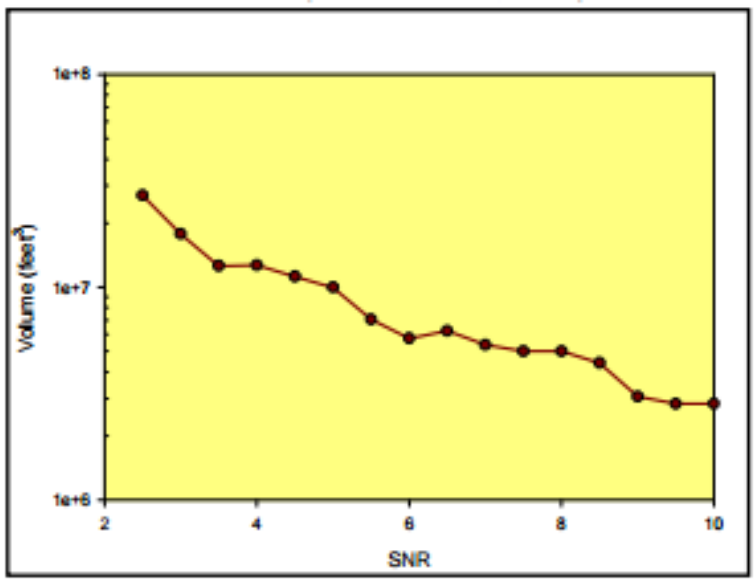


Figure 10. SNR and stimulated volume relation from previous research (Reprinted from Maxwell, 2011)

In general, more data means more stable features. When data points become discrete, it is difficult to locate the clustered area and this makes the PM less useful. After reviewing all 13 stages, it became clearly stage with more data points, in our cases more than 600 data points, exhibited more robust network behavior. Stage 25 was randomly chosen to be indirectly verified with traditional statistical analysis, and stage 27 was randomly chosen to be detailed in microseismic event progression.

In the network of stage 25 (655 total events), most nodes representing high event density area from PM had smaller event IDs, which mean earlier events. Main data point body has a uniform distribution and nodes at the flares and highly skewed to the late time, shown in Figure 11. This indirectly proved that the PM network could be used as a representation of network fracture. High event frequency areas depicted with high kernel density filter value (red) localize where the pressure first built up, or the "true" origins of fracture network. Low event frequency area spreading out across the entire volume shares the late time commonality.

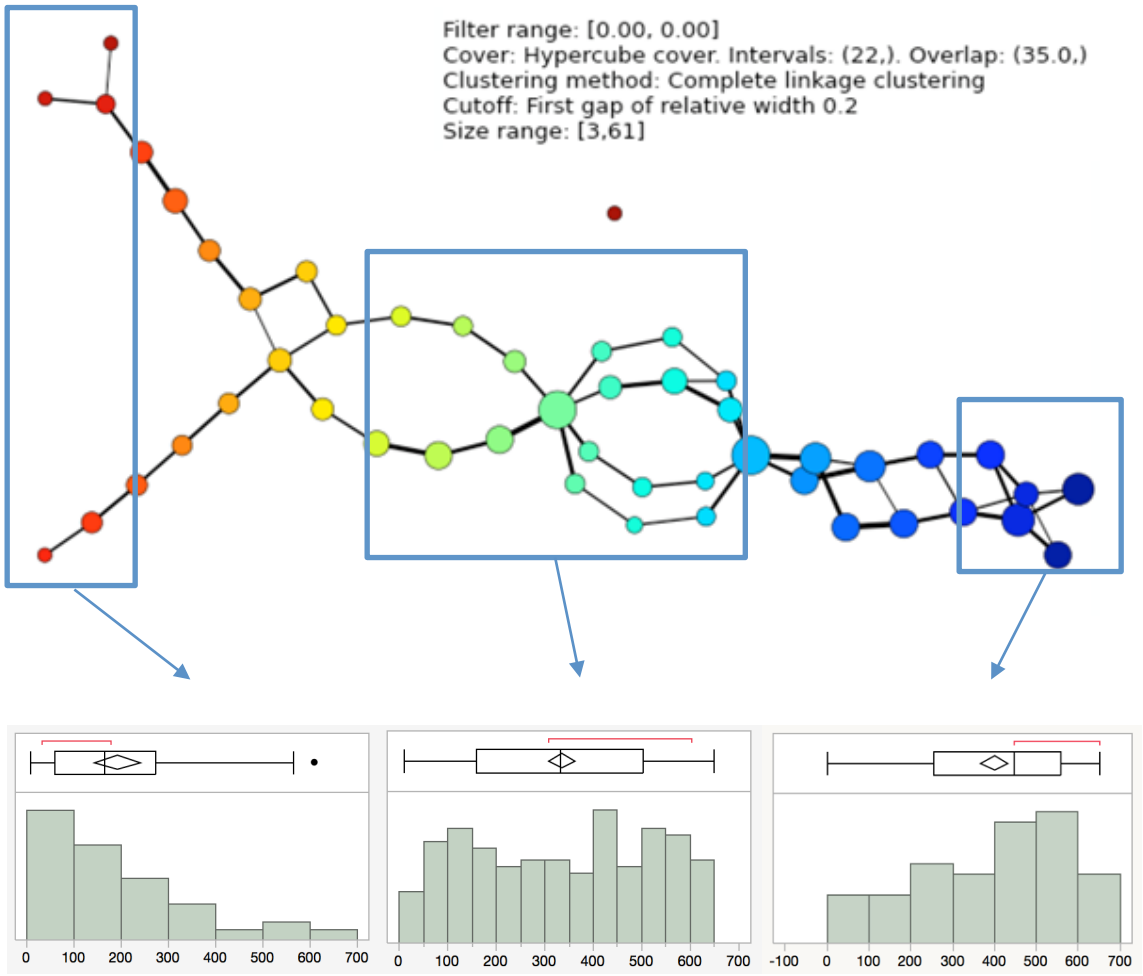


Figure 11. Early (mean: 192, median: 164) middle (mean: 334, median: 333), and late (mean: 399, median: 447) event ID distribution of stage 25, and their respective nodes representation

Stage 27 has 905 events, 25 intervals with an overlap of 35%. This stage originated from three relatively high microseismic density locations, then fracture network grew into each other to form a more complex fracture network, suggesting better connections from each fracture origins, and the flares are representing opposing directions of stimulation orientation. Isolated high filter value nodes suggested that the

pressure built up in the localized region but failed to make connections into the main fracture network. Figure 12 shows the map of event locations and corresponding network. Disconnections between networks indicate poor connection between networks or isolated networks within the stage such as in stage 28 and stage 32 (Figure 13). These disconnections could occur both in the early or middle phase of fracture development. Controlling the degree of overlap also reveals the degree or extent of disconnections. If the disconnections persist in even at a very high overlap, the probability of isolated fractures is also high. Loop features suggest similar density distribution in various directions with equal probability to be originated from different high-density area, signaling the interactions among smaller original fracture networks. As a result, this leads to a larger and more complex network benefiting production.

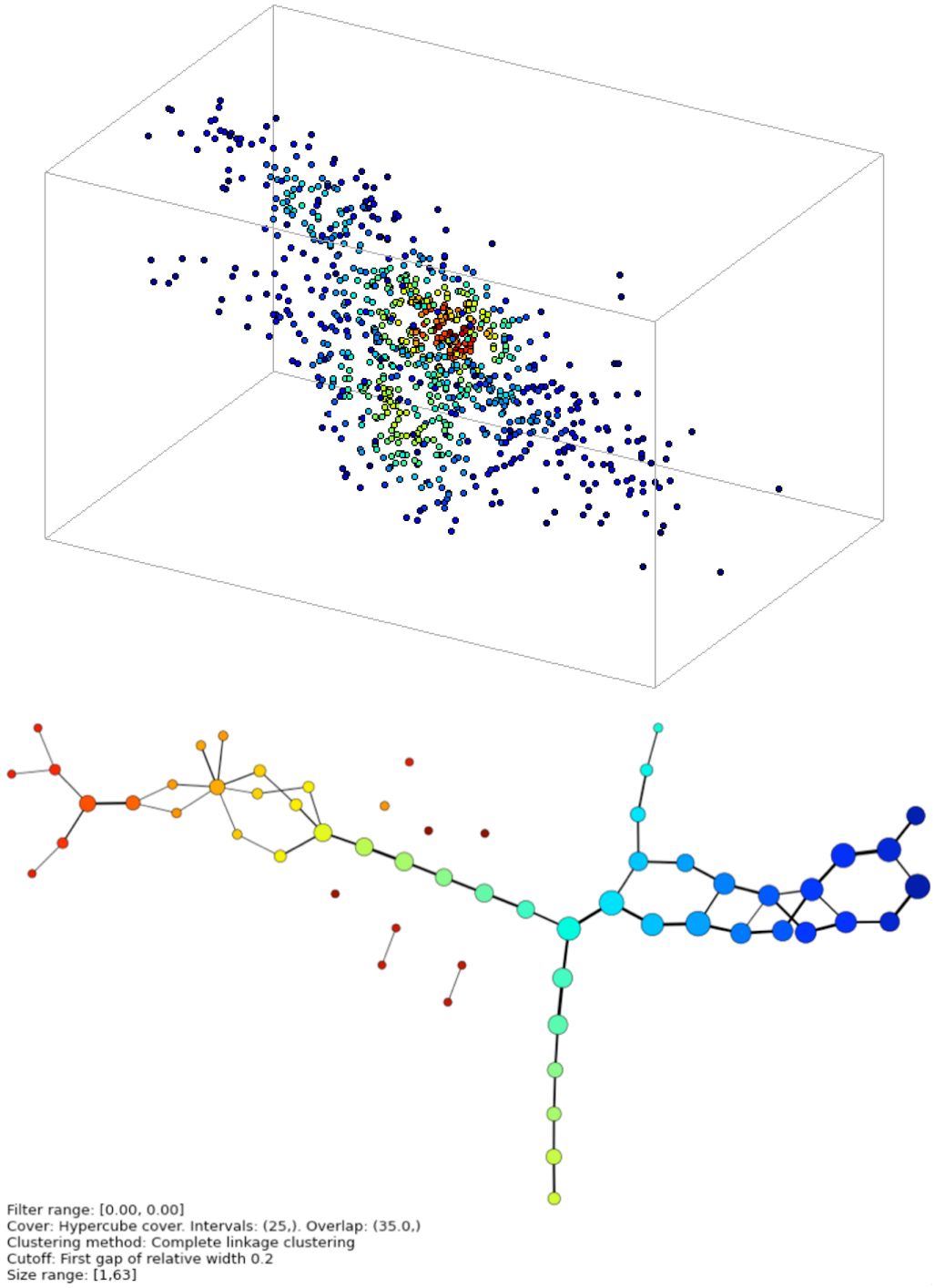


Figure 12. Network of stage 27 started with three flares, and the growing network to the middle section where networks joined each other to form a single complex network, while preserve two different flares. Origins in stage 27 do not have similar filter value.

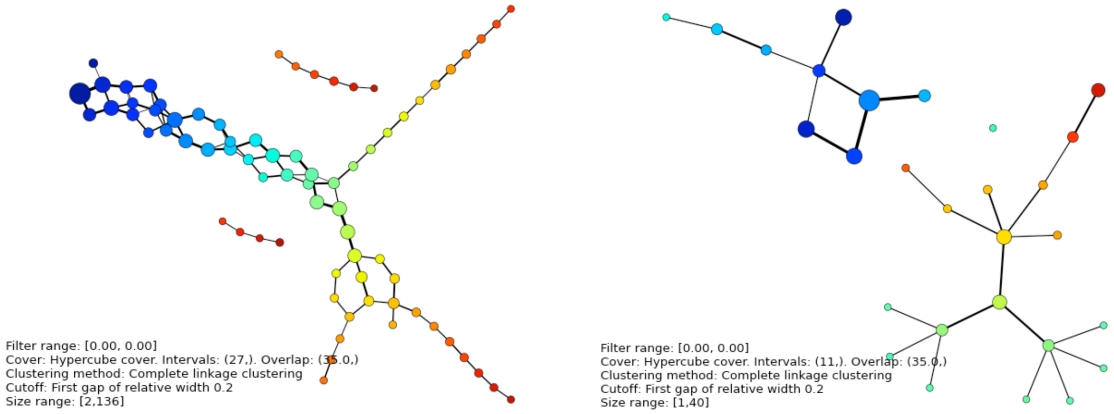


Figure 13. Both stage 28 and 32 have clearly network discontinuities.

Another observation from other stages is that density of each origin could be different or relatively similar. Higher density origins having more activity by common sense would contribute more at growing the network. In stage 27, it is clearly that there was single main origin that provided most of the driving force of the fracture development in theory. In stage 23 and 33, all the origins shared similar densities suggesting that each origin contributed relatively equal amount of energy in growing the network (Figure 14).

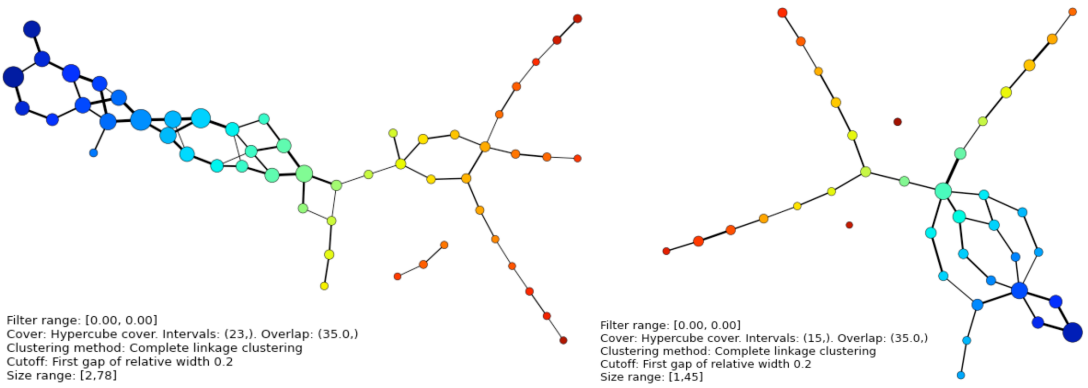


Figure 14. Origins share the similar filter value in stage 23 and 33.

Location attributes analysis reveals the origins, connectivity, and complexity of network simply from the relations of the coordination and order of occurrences of all events. Considering that some of the nodes representing early data points are more than 300 feet apart and each stage only spanned for about 110 feet, it is safe to say that some of the origins of the fracture networks are not located right next to wellbore, suggesting there were existed channels between layers before the stimulation.

The same approach can also be applied to the entire monitored section of the lateral. The number of high-density area will be easily captured on a larger scale.

3.6 Signal Data Observation

Next, four attributes of signal data were analyzed using PM. One critical difference between location and signal data is its additional dimension, which eliminates the possibility of raw data visualization. Coloring the final network output by different attributes indirectly visualizes effect of attributes on data network features. Values for all four attributes share the order of magnitude so that normalization is not necessary. Due to the limitation of the software, PM is currently able to process fewer than approximately 200 data points with distance to eigenvector lens, ruling the SVD out of the equation for large scale network mapping. Loop features were largely observed for relatively fewer intervals when kernel density lens was applied. When more intervals were used, the loops were “stretched” due to a smaller coverage of data points for each node and the probability of sharing the same data point diminished. A featureless

network model does not provide as much information; as a result, a rule of thumb when selecting intervals is $8\ln(n) - 30$, where n is the number of data points after multiple experiments. The overlap should be a value that creates the most features without isolating too many nodes outside of networks.

Signal attributes were mapped with all four coloring schemes. Two nodes representing 16 outliers with very high noise level was observed and eliminated, partly shown in Figure 15. Original PM output with filter value displayed does not provide additional information other than two clearly separated networks with significant gap in the filter value. After applying different coloring scheme to the signal data, it is obvious that in general the top network is relatively featureless as shown in Figure 17, where the bottom network shows some relations among attributes, especially for nodes whose filter values are the lowest.

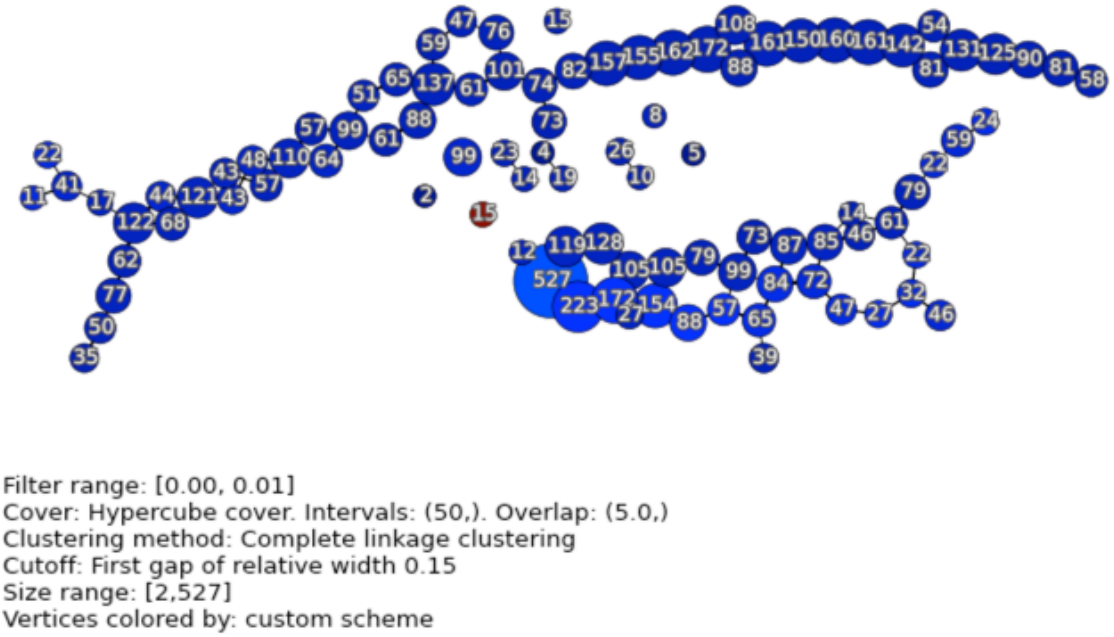
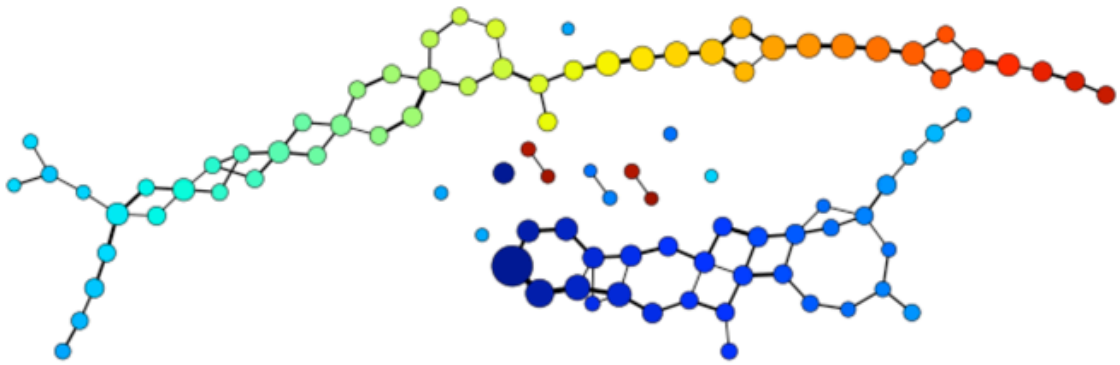


Figure 15. The red nodes represent 15 high noise data points, the other high noise data point was taken out earlier as the largest outlier.



Filter range: [0.00, 0.01]
Cover: Hypercube cover. Intervals: (50,). Overlap: (5.0,)
Clustering method: Complete linkage clustering
Cutoff: First gap of relative width 0.15
Size range: [2,527]
Vertices colored by: custom scheme

Figure 16. PM output after high noise data points were taken out. Two distinctive networks are present, colored by filter value.

Analyzing four different attributes simultaneously allows the data to show its feature without bias towards single specific attributes. The left portion of the lower networks contains around 1800 data points or approximately one quarter of all data points represented by 12 large nodes. These nodes stand out from the crowd and dependencies are likely and especially in the boxed region in Figure 17, we are able to see a positive correlation between SNR and magnitude, and a positive correlation between noise and P/Sh ratio in that specific area. These correlations are difficult to observe by simply plotting two attributes due to the large size of the datasets and frequent noise.

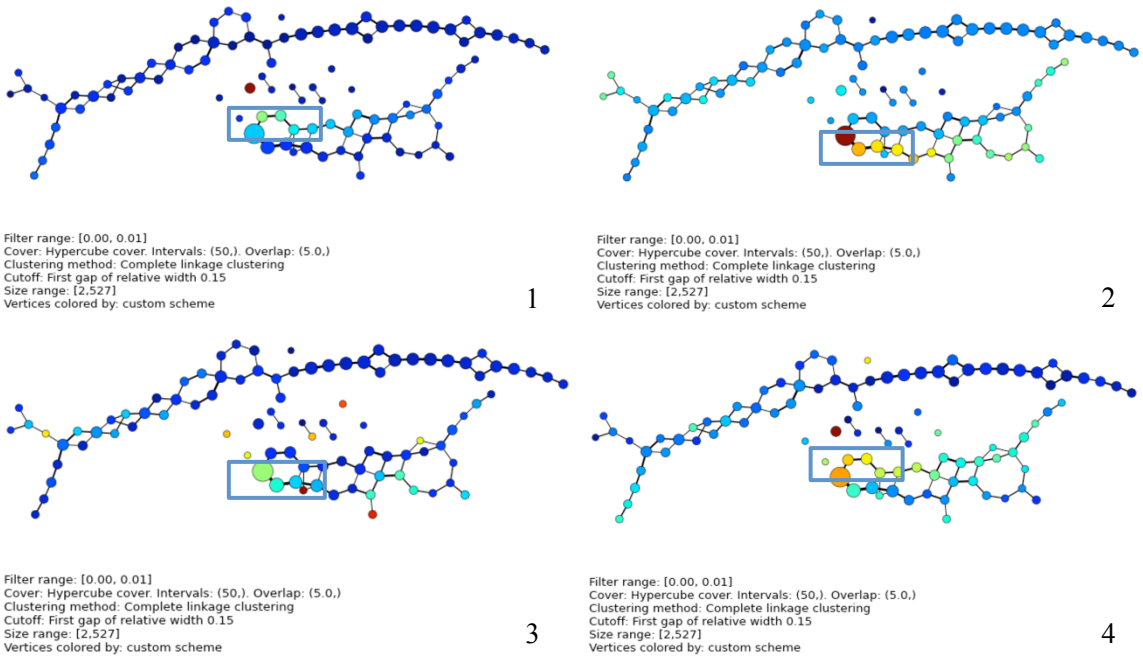


Figure 17. PM output colored by 1, SNR. 2, noise level. 3, P/Sh ratio. 4, Magnitude. High SNR nodes correspond high magnitude nodes and high noise level nodes correspond high P/Sh ratio node, shown in boxed area.

3.7 Combining Location and Signal Data

PM is great at localizing problem. In this problem, it is important to determine the location of the data points that are not generic. Tracing back to the location of the boxed node from the previous analysis enables researchers focus on studying more specific regions rather than the entire microseismic volume. First, 240 data points with high SNR and magnitude represented by two nodes located at top right of the largest node in Figure 15 were positioned. High SNR normally suggests better data quality, and in terms of location of the data, these events are highly clustered in the core region of the

microseismic cloud (Figure 18). The uncertainty of event locations in this region is low, and the probability of extensive stimulation is high.

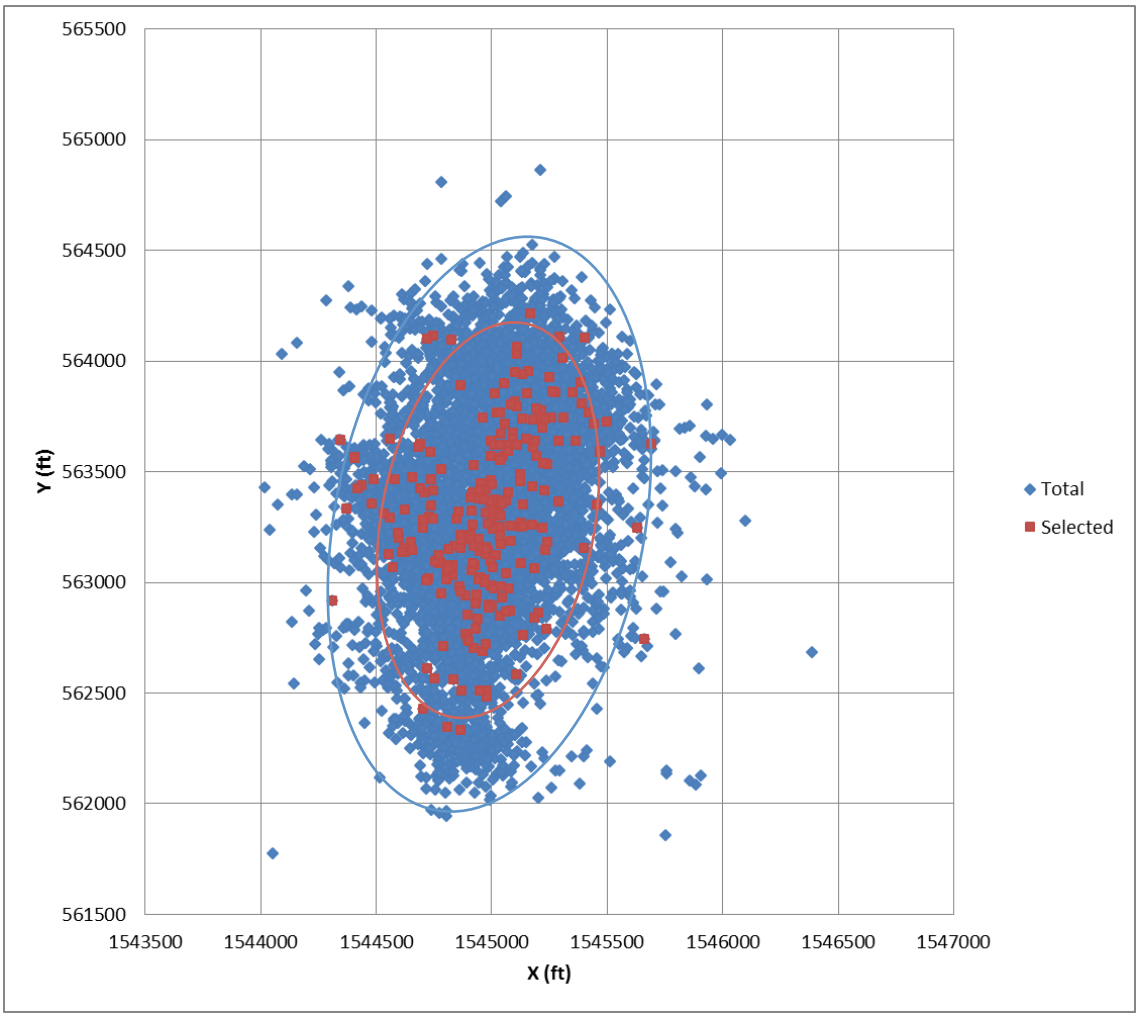


Figure 18. Map view of all Well A events and high SNR, high magnitude events

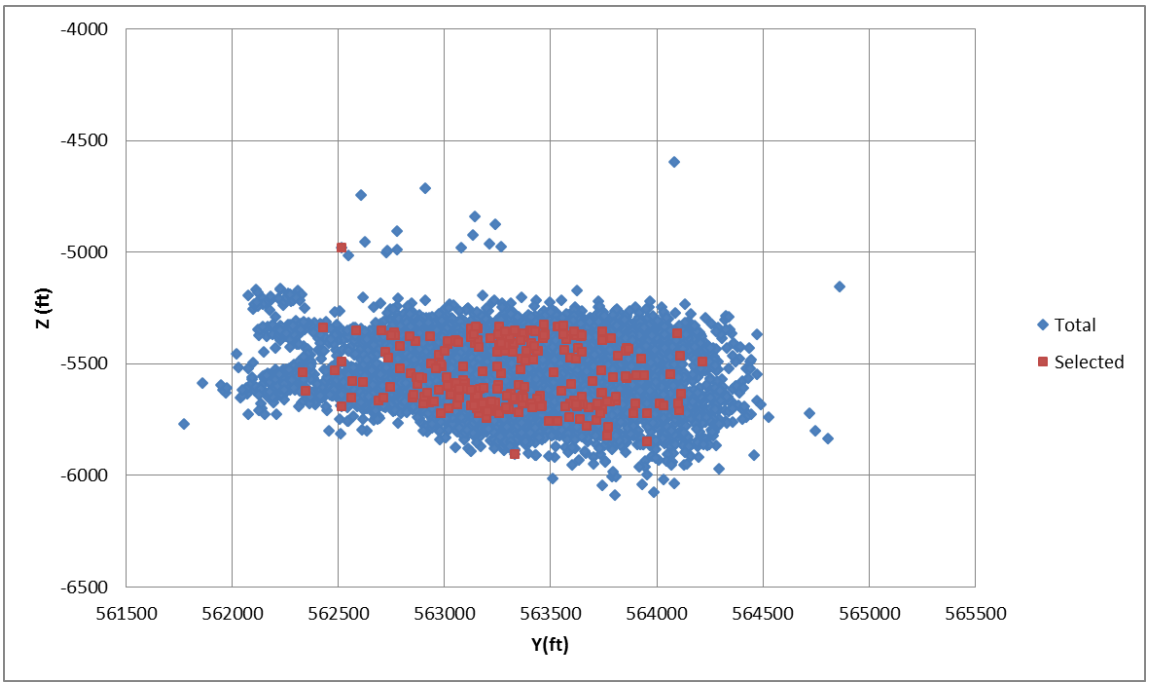


Figure 19. Side view of all Well A events and high SNR, high magnitude events

The largest node containing 527 points has the most distinctive features with the least similarity to most generic data. These data points are all represented by the lowest filter value due to their least similarities. Although these data points only account for less than 10% of the total points, the locations of these data points are highly skewed to low X coordinate. Approximately 79% of these data points (comparing to 54% of total events) are located west of the wellbore, suggesting poorer locational or geological environment for microseismic monitoring (Figure 20) in the area. If we also include two additional similar nodes at bottom right, the trend does not subside. However, a definite conclusion cannot be drawn without dedicated study on the geology of the specific area.

In both cases, the trend in Z direction generally follows the density of total event density; no significant irregularity is observed (Figure 19 and Figure 21).

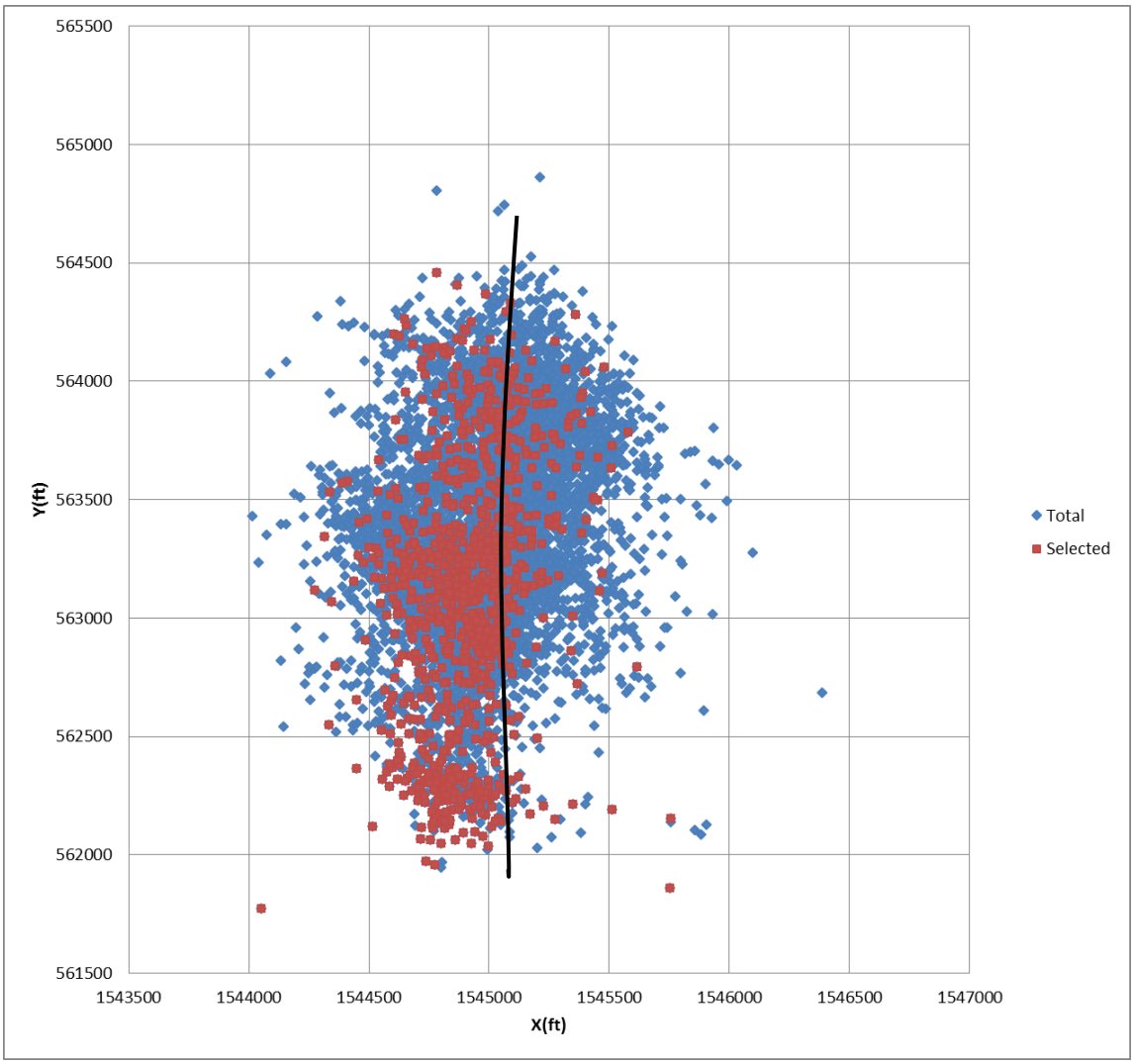


Figure 20. Map view of all Well A events and high noise, high P/Sh ratio events. The black line indicates the location of the wellbore.

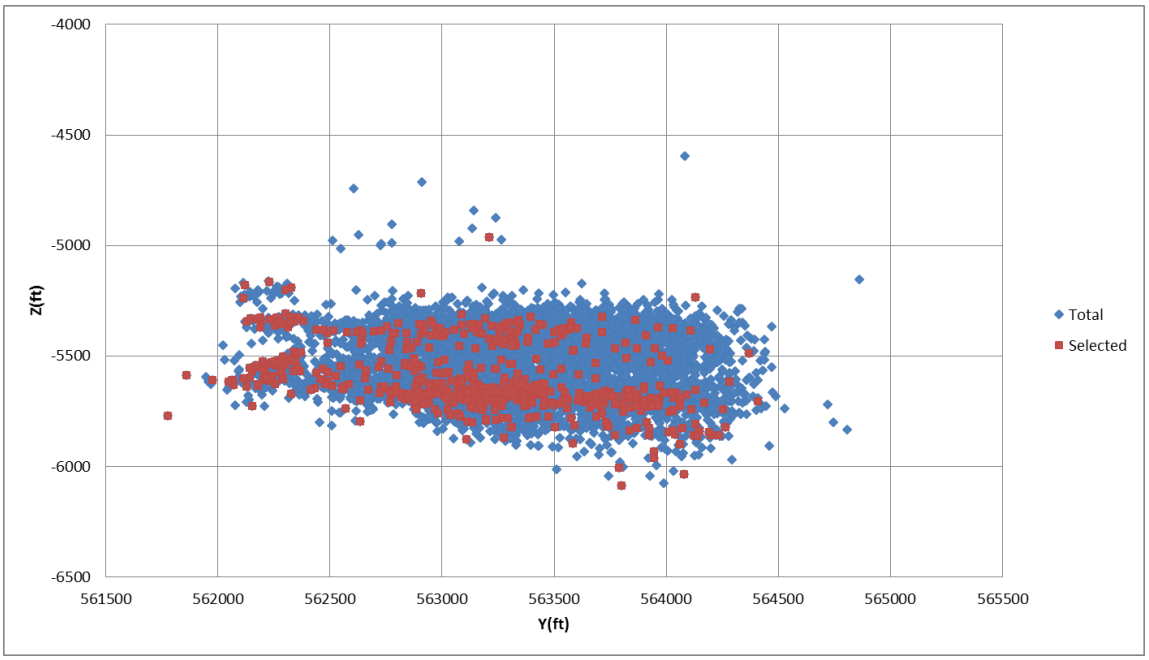


Figure 21. Side view of all Well A events and high noise, high P/Sh ratio events

Signal data controls the quality of data. In microseismic interpretation, the locational uncertainty of each microseismic event is calculated and these location uncertainties will also be carried while calculating the SRV. Utilizing the PM network generated by signal data provides another approach in applying uncertainty to divide the entire dataset into many subgroups according to their attributes, and an uncertainty for each subgroup can be calculated based on the average of its four attributes while determining SRV. Each subgroup could be a collection of similar nodes or a single node when interval is set small enough.

In the original microseismic report, the filtering process of the data points was only limited to magnitude alone. The grouping from PM provided additional information

to the data points containing lowest magnitudes. These data points may have other significantly different attributes that may change the filtering process. Since the highest kernel density pointed out the most irregular data points, these data were considered outliers in one or more attributes, and by probability, did not have the best validity. These data may be presented as isolated nodes or isolate networks. For instance, a collection of data points suggests high stimulation in an isolated volume, but poor connection to the main network preventing the best production. Filtering out irregularities may improve the regression model of production prediction. Still, this needs multiple cases for any consistency.

CHAPTER IV

EVALUATION AND DISCUSSION

4.1 Sensitivity Analysis on PM Parameters Setup

Sensitivity analysis was conducted to test the robustness of the procedure and requirement of accuracy in PM parameters. Subject of test is Stage 23 by random selection. Selecting an appropriate smoothing factor or bandwidth is the main factor to extract insight needed. Too small bandwidth will create discreteness at high filter values, resulting in network output similar to the one generated with distance lenses. Too large bandwidth smooths out too much information, making the result featureless and meaningless. Cover and overlap have been discussed in the previous chapter and the selection of clustering method does not cause any significant differences in the final network (Figure 22). It is safe to extend the consistency of the PM network output into the robustness of the implications concluded in the earlier chapter.

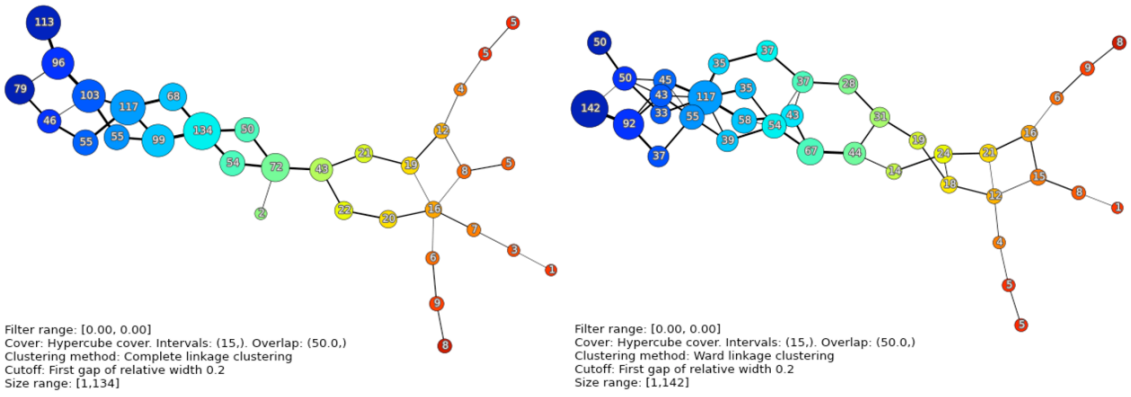


Figure 22. PM output using complete linkage clustering method (Left), using Ward linkage clustering method (Right).

4.2 Comparing TDA with Clustering Method Alone

TDA methods have always been juxtaposed with machine learning algorithm in many research cases. In this study where the grouping played great roles in drawing possible microseismic monitoring conclusion, clustering raw data was chosen to be compared with MA. In the comparison, it is observed that clustering raw data can pick up data points that are significantly different from the rest. In a simplified network representation of stage 21, two nodes were study closely to identify the difference from both methods. These two nodes were selected such that one represents data points that were significantly different from the rest of the data points and the other one represented a collection of data points that contained average values of attributes. Where PM stands out is that when some data point are sharing similar attributes, but do not have significant difference than more generic data points, it is still be able to locate and group them into a single node. Clustering method is less sensitive to small variations thus needs strong signatures to locate the similarity of data. Figure 23 shows the PM output with two nodes of study, and the clustering result and detailed explanation can be found in Figure 24 and Figure 25.

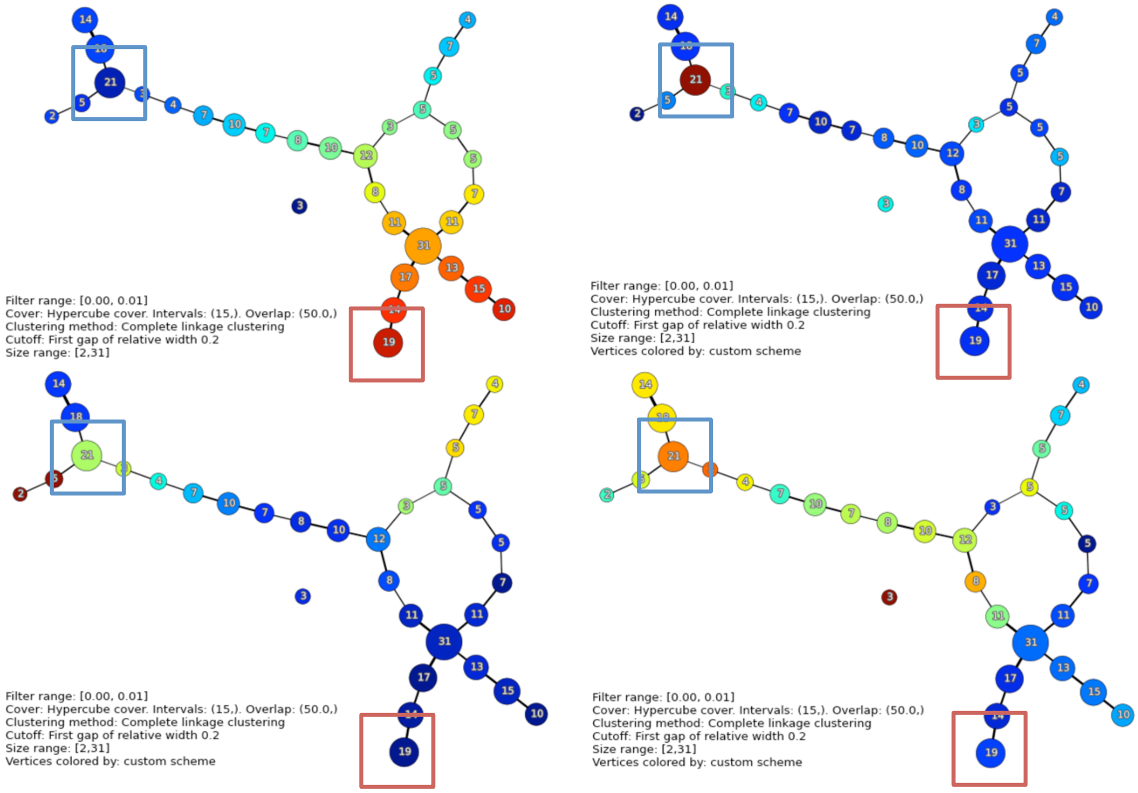


Figure 23. A PM example output to be compared with raw data hierarchy clustering below (in Figure 24 and Figure 25)

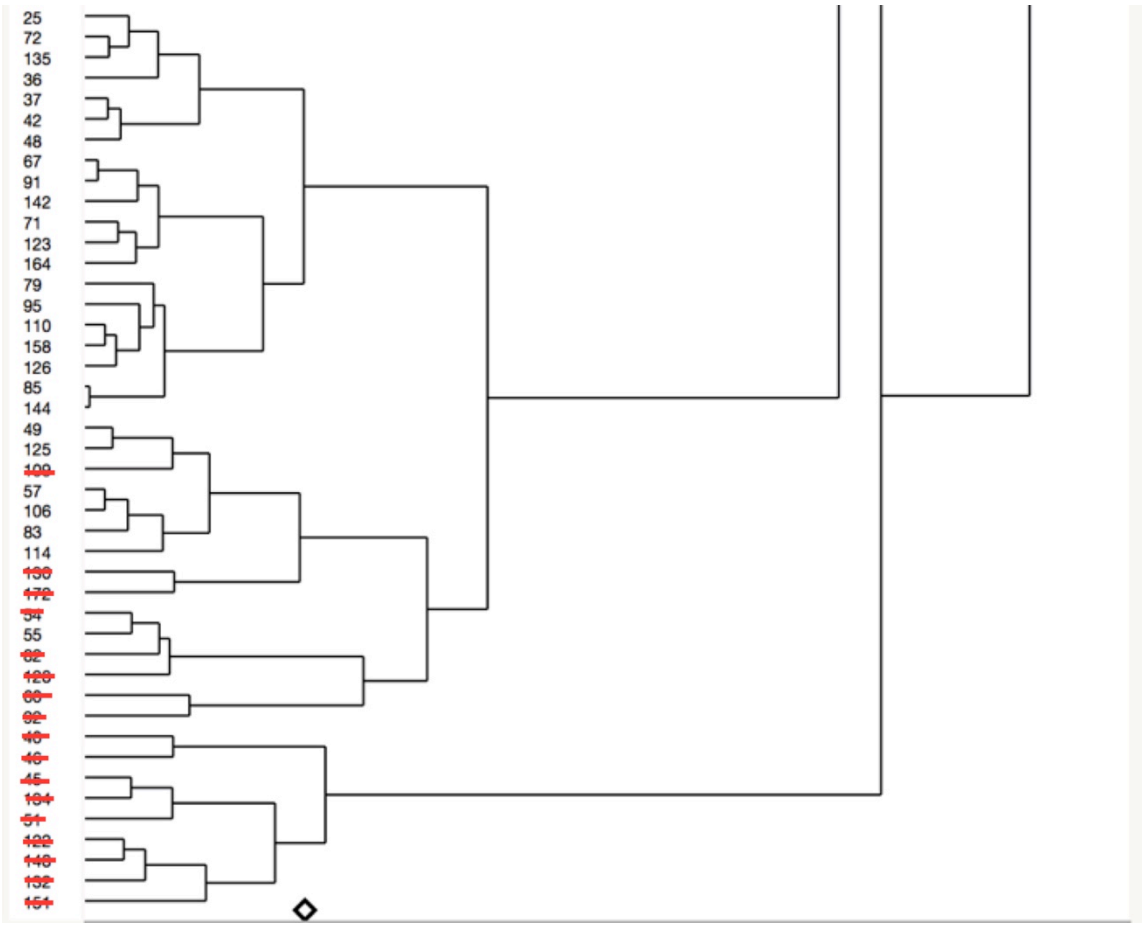


Figure 24. PM and clustering method shared similar results when the node represent data points that were vastly different. Marked data points are in the represented by the blue-boxed node.

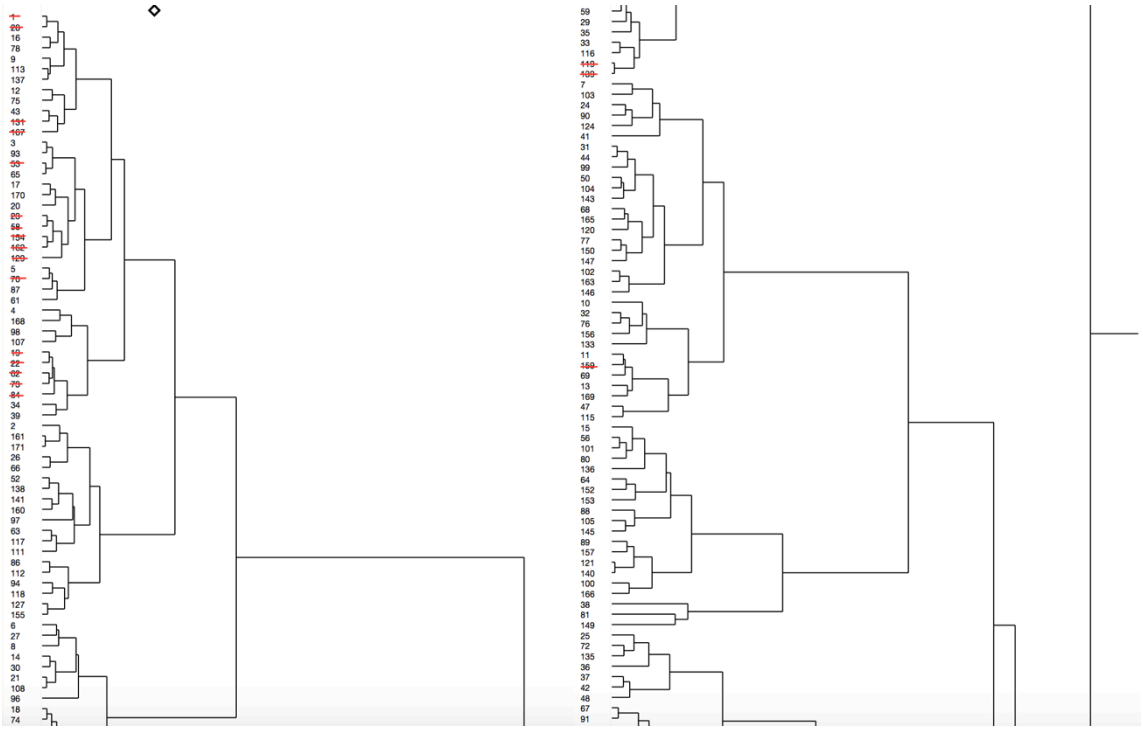


Figure 25. Results from PM and clustering method were different for a more generic data points. Marked data points are in the red-boxed node.

4.3 Preparing Data Points for Future Correlation

Summary of PM containing information such as the number of network nodes and isolated nodes, number of fracture networks, connection quality which can be derived from the number of intermediate-filter-value nodes provides supplementary information to raw data summary like number of untreated data points. When additional zonal production information such as distributed temperatures sensing is available, it is possible to derive correlations between network data and production of each stage. An example of fracture network summary after considering PM output is shown in Table 3.

Table 3. The stage by stage summary of all the findings from PM to Well A

Stage	Data Pts	Origins	Type 1	Type 2	Magn. sum	Special Note
21	172	2	169	3	-354.17	
22	271	4	262	9	-556.984	1 isolated origin
23	744	4	697	47	-1537.998	1 isolated origin
24	725	4	698	27	-1579.237	
25	655	2	610	45	-1414.102	
26	727	3	692	35	-1610.509	
27	905	4	829	76	-1999.278	
28	1248	5	1143	105	-2809.385	2 isolated origins
29	679	4	630	49	-1552.989	
30	347	3	322	25	-811.212	
31	257	4	233	24	-607.439	
32	177	2	164	13	-416.547	Disconnected networks
33	255	3	186	69	-595.295	

*Type 1 data points are events within network; Type 2 data points are isolated events. Shaded columns contain raw information

4.4 Result Summary

PM outputs from different views exhibit several intriguing behaviors of raw microseismic data. These observations are listed below. Although only part of the implications raised here have solid physical foundation, these results are purely data driven on a very small scale. Moreover, these points were tested in two more wells, which are discussed in the next section.

1. Location data reveals true origins of fracture network, which neglect the bias from the position of wellbore in fracture growth modeling. Discontinuity of

fracture within a single stage can be observed. Stage to stage difference in size, number, events number, disconnections within stage, density distribution of the fracture needs to be correlated with drilling information and logging data. This will help the operator predicting the fracture network growth from cheaper and easier monitoring techniques.

2. Signal data displays dependencies among microseismic attributes at relatively large values but not in smaller, more generic data points.
3. Combining information obtained from both location and signal data can assess data quality by localized regions. High quality data points are normally located at the higher density regions, or around origins. Low quality data points are generally location specific, and require more information for sound explanation.

4.5 Two Additional Wells for More Evidence

Well B is completed in Wolfcamp B formation and Well C is completed in Wolfcamp C. The similar method can be applied to these two wells monitored by the same geophones. It is common for the service companies to use single well to monitor multiple wells. Two overall PM networks were created, where they showed data relations, irregularities, and similarities. The outputs of the two wells exhibit very similar behaviors described in the previous section. It also eliminated some of the doubts raised in the original study such as the influence of location of monitoring well on data noise.

To elaborate, the noisier events occurred closer to the monitoring well in Well A, but farther away from the monitoring well in Well B and C.

CHAPTER V

CONCLUSION

PM proved to be a great tool for data analysis that can group and correlate multi-dimensional data from microseismic monitoring. After testing with four different lenses, it is noted that all lenses will create networks that show informative shape features with different emphasis, and it is recommended to test with all available lenses if possible. In this study, density lenses stood out as the best performing selection to determine similarities in both location and signal data.

PM results from location data showed that within stages, there were multiple stimulation origins and the PM network could indicate growth (or lack of) of the fracture network. For a complete microseismic analysis, data points that represented by isolated nodes should not be considered when calculating microseismic volume or part of connected fracture. PM is an excellent spotter as it localizes problems, irregularities, and characteristics. By processing multiple attributes simultaneously, PM could find dependencies among them that were difficult to conclude from other bivariate methods. Signal data revealed the quality of data, location-sensitive attributes. More detailed results were concluded and listed in the previous chapter. Once the cover and overlap were optimized, different clustering method would not create significant difference in networks and nodes as the representation of the dataset. However, clustering by raw data resulted in different grouping when data points were relative generic in the data set.

Outputs of PM can be further analyzed using various data analysis, and they provide insights for future study. Many wells may have other type data that can be correlated to (unfortunately not in this case. Lateral core data are not taken and log data are incomplete). When analyzing network stage-by-stage, it neglected the interference among stages. One recommendation is to analyze segments of stages with overlapping to observe any features of interference.

As mentioned in the very beginning of the paper, TDA should never be the end of a study as it is a tool for extract features and relations without providing further information. Furthermore, the interpretation to visualization may be different from researcher to researcher. In most cases, it is used to point out a direction for future studies. In this case, the PM network needed validation from current fracture growth model, and even production correlation if production log is available. However, we live in an imperfect world where not everything is available.

The research on TDA or any *Mapper* related topic in petroleum engineering is just beginning, while this paper provided some results, comparisons, and instructions. However, at the finishing stage of the study, a few points was raised and listed here. With the PM code available, additions to the filter functions will definitely power up its ability. It is also possible to incorporate flow functions as a lens to discover different features from topological networks.

REFERENCES

Alfaleh, A., Wang, Y., Yan, B., Killough, J., Song, H., Wei, C., 2015. Topological Data Analysis to Solve Big Data Problem in Reservoir Engineering: Application to Inverted 4D Seismic Data. Society of Petroleum Engineers. doi:10.2118/174985-MS

Ayasdi, 2016. TDA and Machine Learning: Better Together (Whitepaper), Ayasdi, Inc.

Carlsson, G., 2017. The Shape of Biomedical Data. Current Opinion in Systems Biology 2017, 1:109–113

Carlsson, G., 2009. Topology and Data. Bulletin of the American Mathematical Society. vol. 46, no. 2, April 2009, pp. 255-308

Carlsson, G., 2011. The Shape of Data. Retrieved from University of Chicago Partha Niyogi Memorial Conference: Computer Science

Frédéric Chazal, David Cohen-Steiner and Quentin Mérigot, Geometric Inference for Measures based on Distance Functions, INRIA Rapport de recherche 6930 Second version, 2010, http://hal.inria.fr/inria-00383685/.

Choi, J., 2016. Well Performance Predictive Modeling Using Topological Data Analysis (Thesis), Texas A&M University.

Cortis, A., 2015. Topological Data Analysis of Marcellus Play Lithofacies. Society of Petroleum Engineers. doi:10.2118/178627-MS

Famili, A., Shen, W., Weber, R., Simoudis, E., 1997. Data Preprocessing and Intelligent Data Analysis. Intelligent Data Analysis 1 (1997), pp. 3-23

Halsey, T., Priyadarshy, S., Collins, E., Heilmann, M., Franchek, M., Husain, A., 2017. Big Data and Data Analytics, presented at Offshore Technology Conference in Houston, Texas, USA

Kraft, R., 2016. Illustrations of Data Analysis Using the Mapper Algorithm and Persistent Homology (Thesis), KTH Royal Institute of Technology.

Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G., 2013. Extracting insights from the shape of complex data using topology. Scientific reports 3. doi:10.1038/srep01236

Maxwell, S., 2011, What Does Microseismicity tells Us About Hydraulic Fractures, presented at the Society of Exploration Geophysicist Annual Meeting in San Antonio, Texas, USA

Maxwell, S., Urbancic, T, Demerling, T., Prince, M., 2002, Real Time 4D Seismic Imaging of Hydraulic Fractures, Society of Petroleum Engineers. doi:10.2118/78191-MS

Mishra, A., Spracklen, L., 2016. Enterprise-scale Topological Data Analysis Using Spark. Alpine Data, presented at Spark Summit in San Francisco, California, USA

Müllner, D., Babu, A., 2013. Python Mapper: An open-source toolchain for data exploration, analysis and visualization, URL http://danifold.net/mapper

Müllner, D., fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python, Journal of Statistical Software 53 (2013), no. 9, 1-18

Singh, G., Mémoli, F., Carlsson, G., 2007. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. Eurographics Symposium on Point-Based Graphics, pp. 91-100

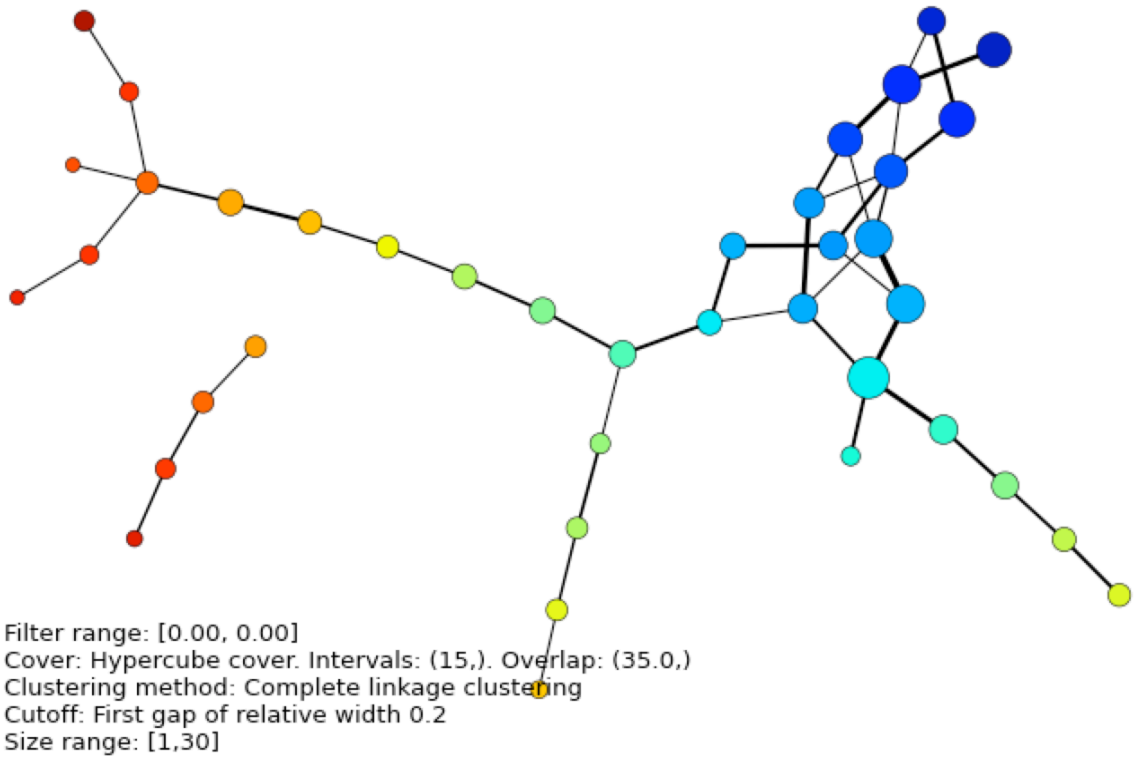
Silverman B., Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986

Snelling, P., Rahimi Zeynal, A., de Groot, M., & Hwang, K. 2014. Microseismic-Derived Correlations to Production in the Horn River Basin. Society of Petroleum Engineers. doi:10.2118/170627-MS

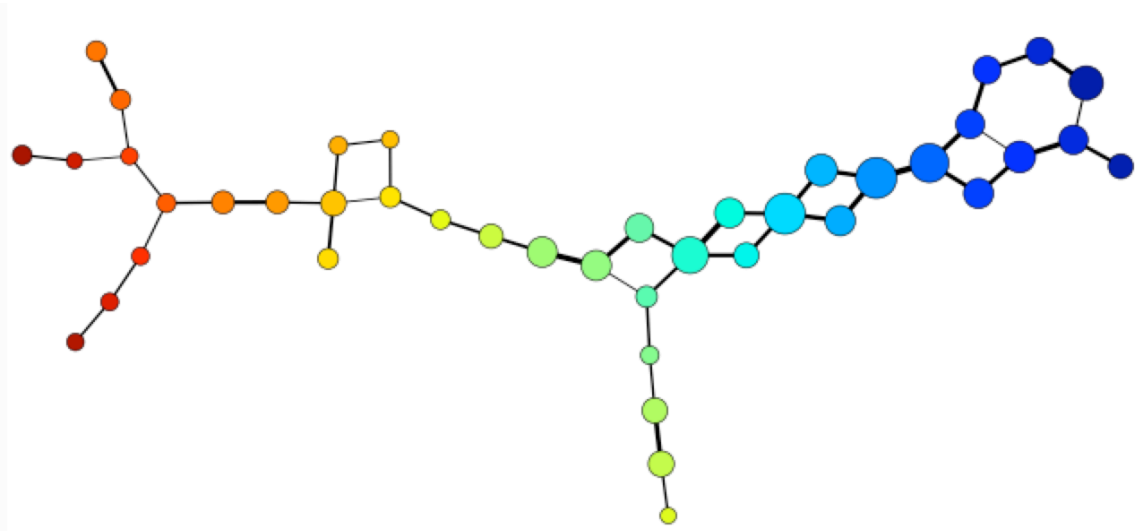
Stovner, R., 2012. On the Mapper Algorithm (Thesis), Norwegian University of Science and Technology.

APPENDIX

All location networks from Well A have been generated and listed below for stages that are not listed or detailed in the main body of the study. As a result, these plots are listed in the appendix for further reference and comparison with networks analyzed. The rule of thumb of choosing intervals and overlap is the same for all stage-by-stage network, and the complete linkage clustering is chosen for all networks

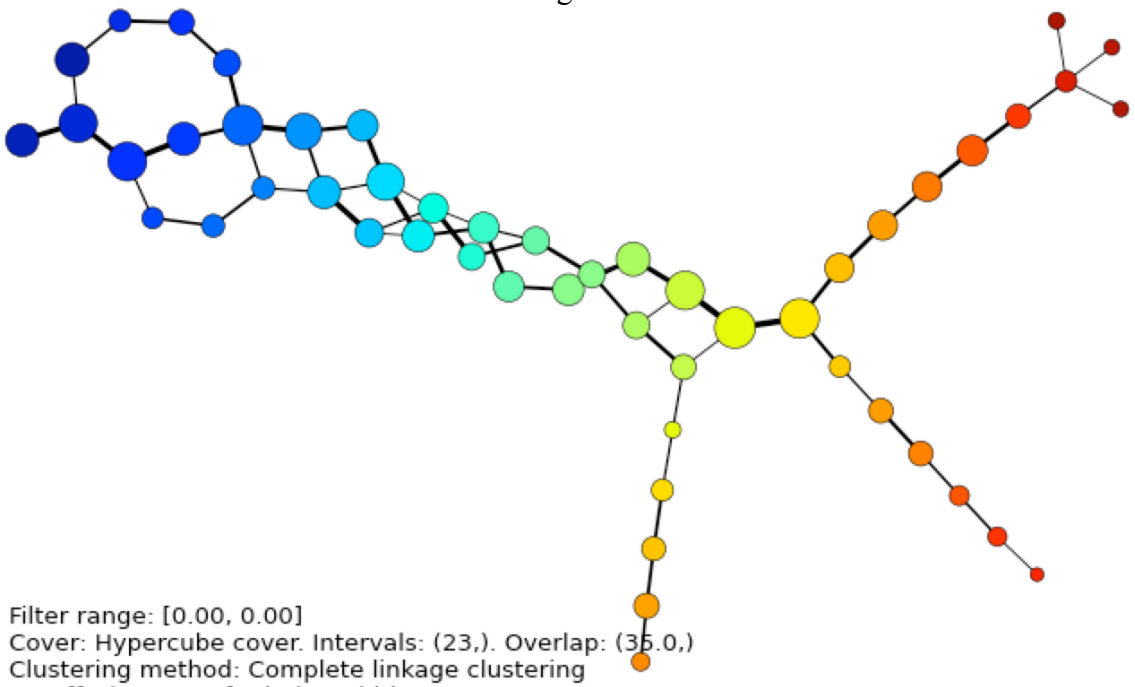


Stage 22



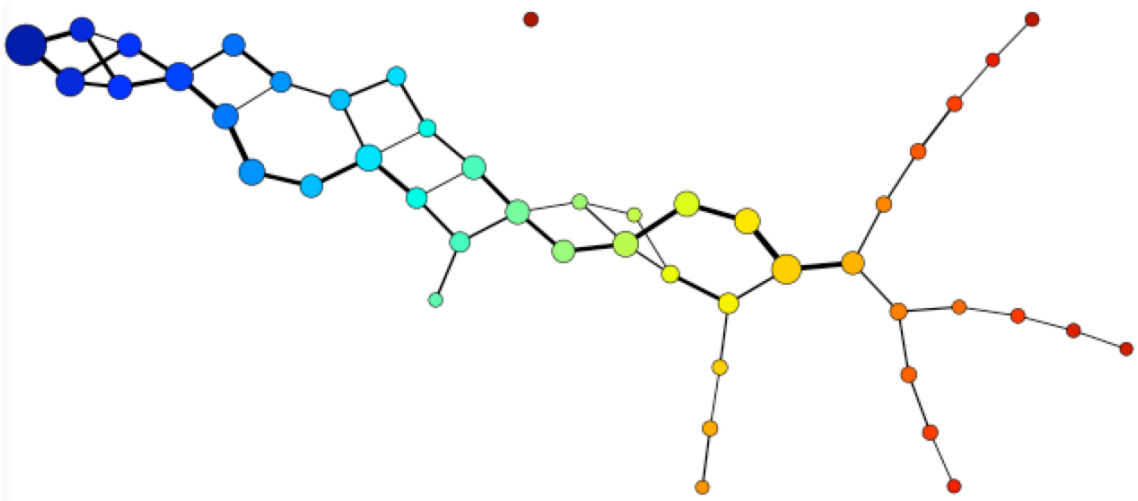
Filter range: [0.00, 0.00]
Cover: Hypercube cover. Intervals: (23,). Overlap: (35.0,)
Clustering method: Complete linkage clustering
Cutoff: First gap of relative width 0.2
Size range: [5,73]

Stage 24



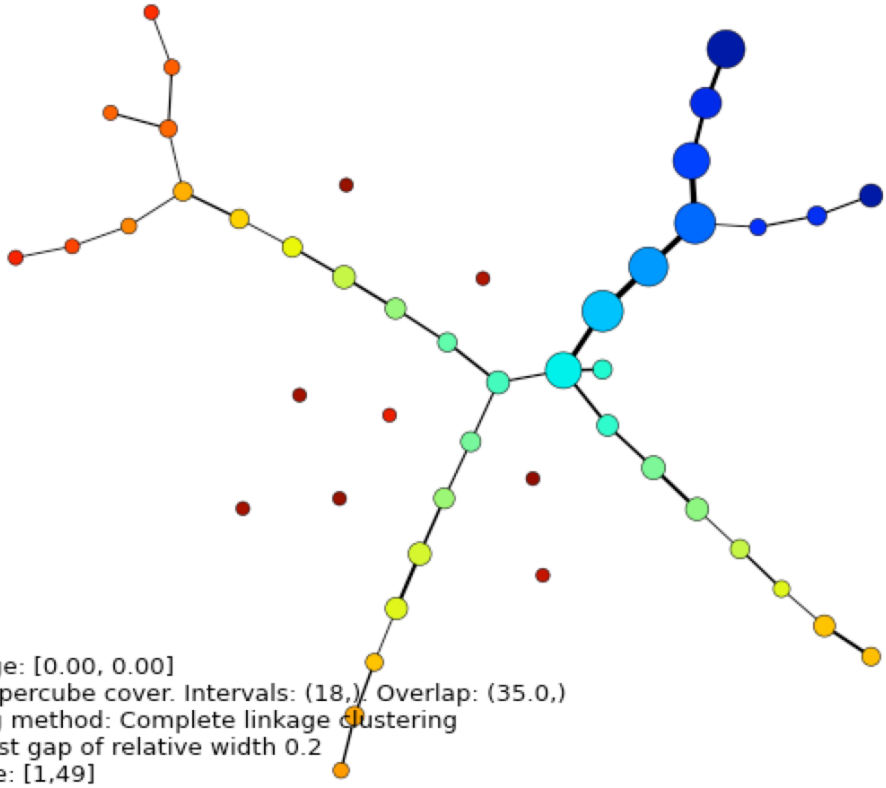
Filter range: [0.00, 0.00]
Cover: Hypercube cover. Intervals: (23,). Overlap: (35.0,)
Clustering method: Complete linkage clustering
Cutoff: First gap of relative width 0.2
Size range: [1,50]

Stage 26



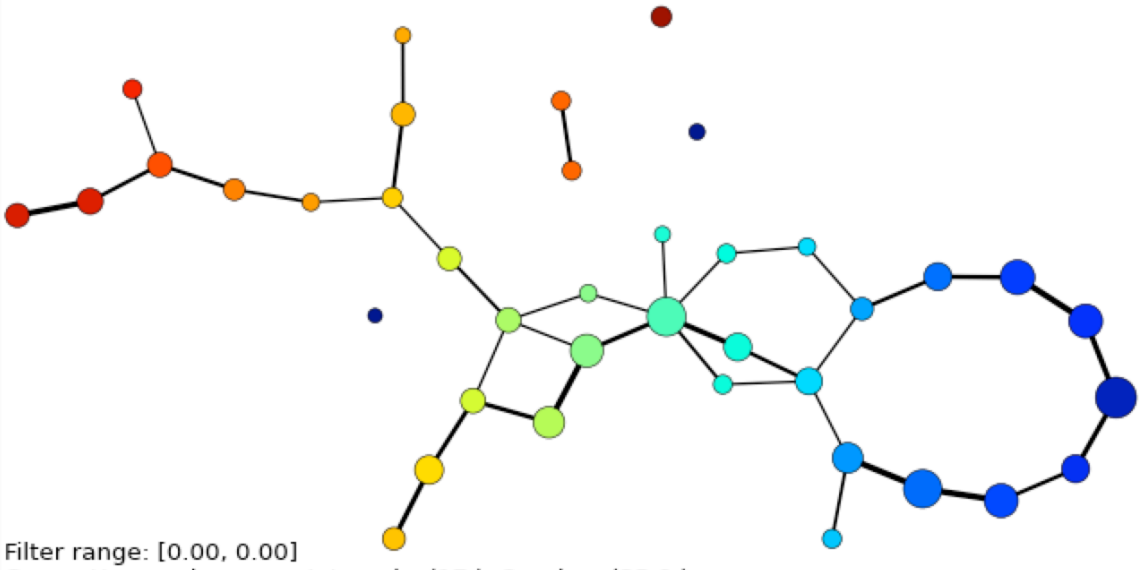
Filter range: [0.00, 0.00]
Cover: Hypercube cover. Intervals: (22,). Overlap: (35.0,)
Clustering method: Complete linkage clustering
Cutoff: First gap of relative width 0.2
Size range: [1,114]

Stage 29



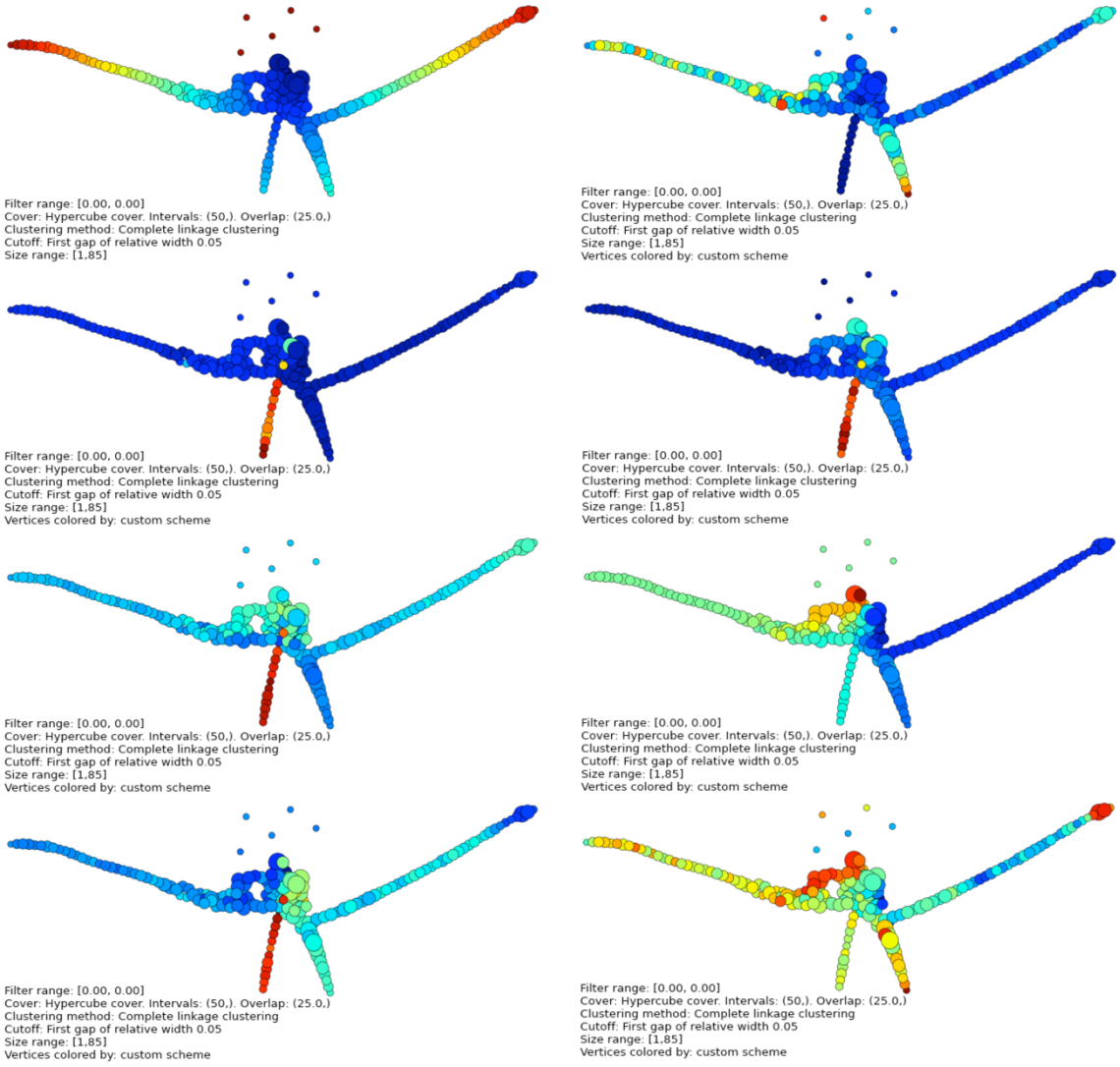
Filter range: [0.00, 0.00]
Cover: Hypercube cover. Intervals: (18,). Overlap: (35.0,)
Clustering method: Complete linkage clustering
Cutoff: First gap of relative width 0.2
Size range: [1,49]

Stage 30

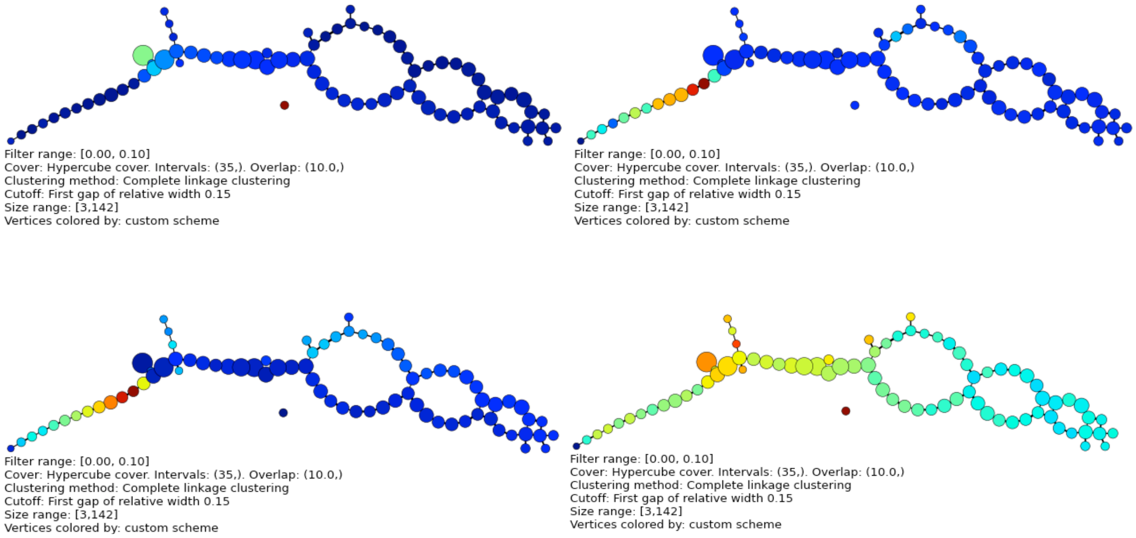

Filter range: [0.00, 0.00]
Cover: Hypercube cover. Intervals: (17,). Overlap: (35.0,)
Clustering method: Complete linkage clustering
Cutoff: First gap of relative width 0.2
Size range: [1,29]

Stage 31

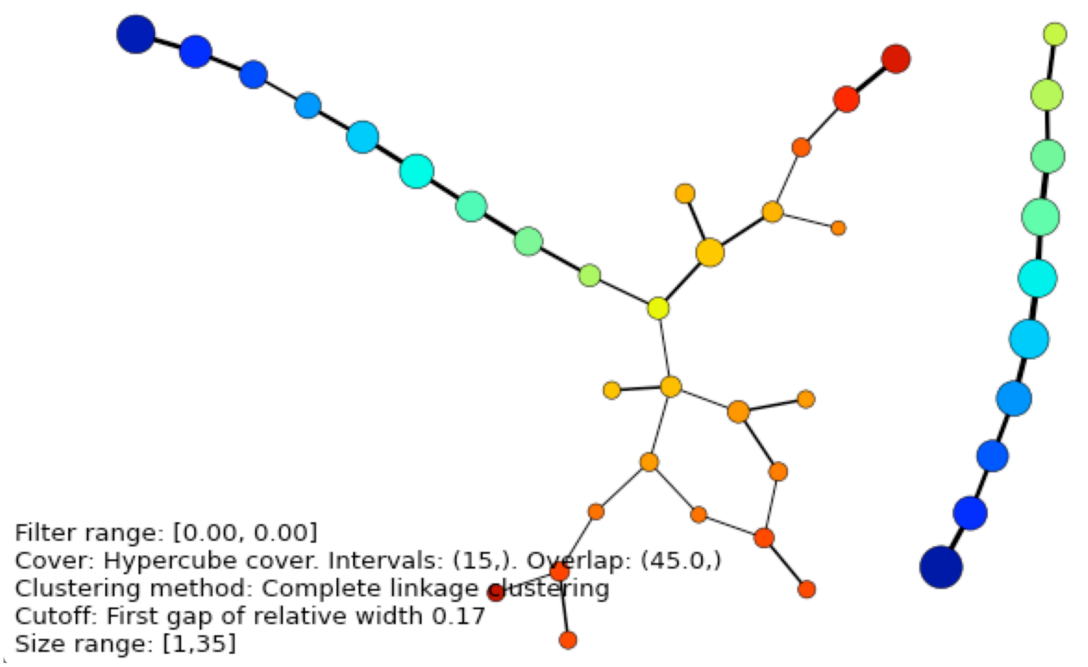
Well B summary output by filter values and all 7 attributes are listed below along with the signal attribute output. Location and signal irregularity (left bottom flare) is back traced to stage 8 containing an isolated cluster of events. The trend of high-SNR, high-magnitude and high-noise, high P/Sh is confirmed. Stage 8 location output is also shown following the signal data output.



Well B PM output colored by filter value and 7 attributes (SNR, noise, P/Sh, X, Y, Z, magnitude)

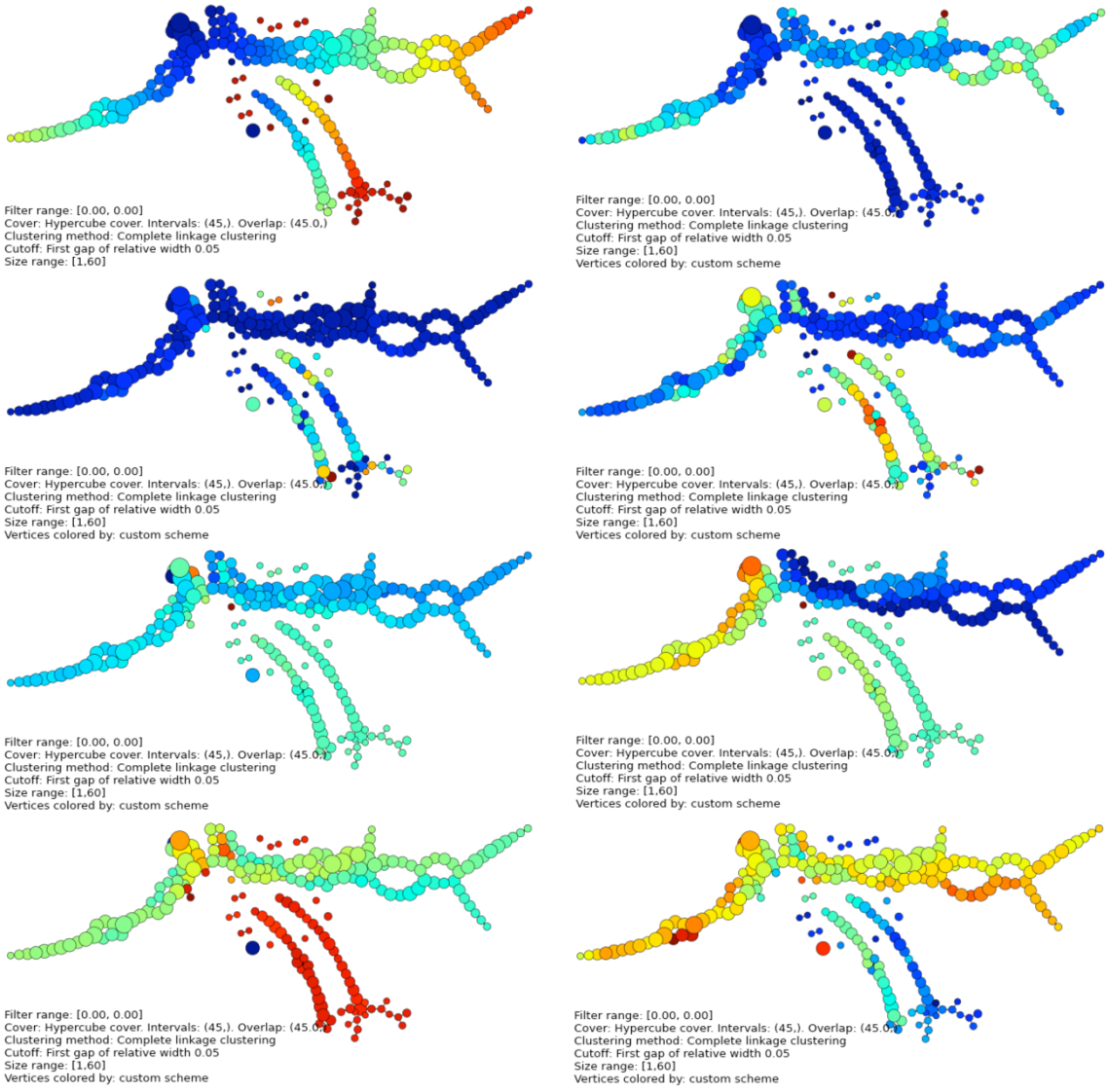


Well B signal data outputs colored by 4 attributes (SNR, noise, P/Sh, magnitude)

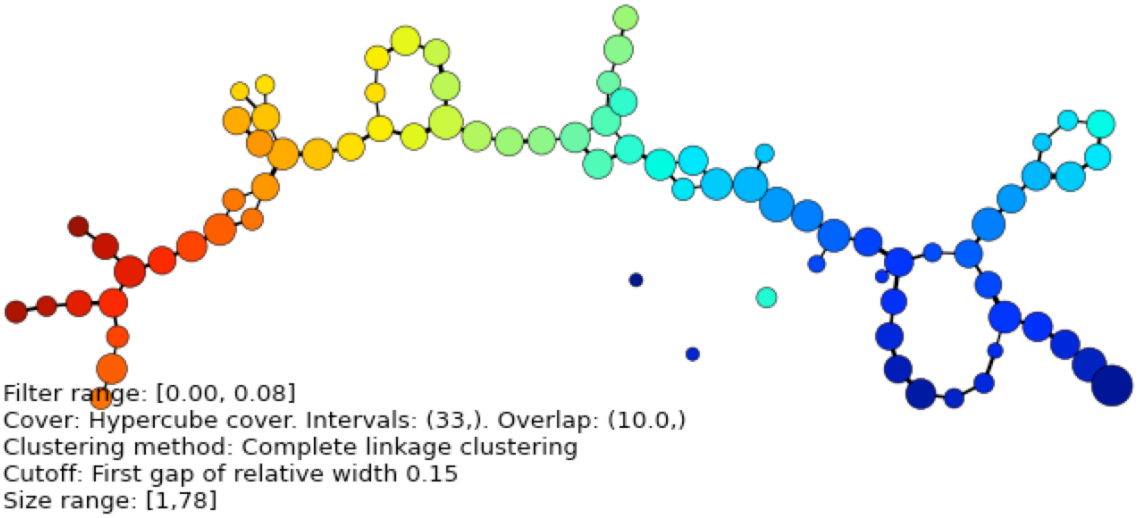


Stage 8 possessing two distinctive networks or SRV

Similarly, Well C outputs are listed below. The second plot is the signal data output color by filtered value. Since we understand events or data points the lowest filter values are the least common data points among the entire dataset, these nodes are likely to represent the correlated portion similarly to Well A and Well B



Well C PM output colored by filter value and 7 attributes



Well C signal data outputs colored by filter values