A Graph Theoretic Approach for Spatial Analysis of Induced Fracture Networks

Deborah Glosser[1,2]* and Jennifer R. Bauer[3,4]

[1]*United States Department of Energy, National Energy Technology Laboratory, Pittsburgh, PA, USA*
[2]*Oak Ridge Institute for Science Education, Oak Ridge TN*
[3]*United States Department of Energy, National Energy Technology Laboratory, Albany, Oregon, USA*
[4]*United States Department of Energy, National Energy Technology Laboratory, AECOM, Albany, Oregon, USA*

Received May 27, 2016; Accepted August 26, 2016

Abstract: Drilling induced fractures are generated when excessive stresses around a borehole cause tensile failure of the wellbore wall. If stress concentrations are great enough, compressive failures can form in the region surrounding the wellbore, leading to wellbore breakout, and the potential compromise of wellbore integrity. Another category of induced fracture networks are hydraulically induced fractures, which are generated by the injection of pressurized fluids into the subsurface. Overlapping induced fracture networks between collocated wellbores may increase pathways in the subsurface, and create the potential for unwanted fluid leakage. The generation of induced fractures is greatly dependent upon the structural and geological characteristics. Probabilistic-based simulations are often used to model fracture systems. Several methods for modeling local fracture networks have been proposed in the literature. These models often involve the generation of randomly located fractures, and may have limited capabilities for honoring engineered fractures such as induced fracture networks. We present a graph theoretic approach for identifying geospatial regions and wellbores at increased risk for subsurface connectivity based on wellbore proximity and local lithologic characteristics. The algorithm is coded in Matlab, and transforms 3 dimensional geospatial data to graph form for rapid computation of pairwise and topological relationships between wellbores (nodes), and the spatial radius of induced fractures (edges). Induced fracture reaches are represented as cylinders with a radius *r*, based on literature derived ranges for fracture lengths for different lithologies (e.g. shale, sandstone).

**Corresponding author*: deborah.glosser@netl.doe.gov

The topological algorithm is compared to a standard graph-based k-nearest neighbor algorithm to demonstrate the value of incorporating lithologic attributes in graph-based fracture models. The algorithms are applied to two scenarios using Pennsylvania wellbore and lithologic data: a subset of data from the Bradford field, as well as a known leakage scenario in Armstrong County. The topological algorithm presented in this paper can be used to complement existing fracture models to better account for the reach of induced fractures, and to identify spatial extents at increased risk for unwanted subsurface connectivity. As a result, the method presented in this paper can be part of a cumulative strategy to reduce uncertainty inherent to combined geologic and engineered systems. The model output provides valuable information for industry to develop environmentally safe drilling and injection plans; and for regulators to identify specific wellbores at greater risk for leakage, and to develop targeted, science-based monitoring policies for higher risk regions.

Keywords: Graph theory, spatial analysis, hydraulic fracturing

1 Background and Rationale

Commercial hydrocarbon drilling began as early as the 1800s in West Virginia and Pennsylvania [1]. Since then, human engineering of the subsurface has expanded worldwide to include gas storage [2]; CO_2 injection [3]; unconventional resource exploration [4]; and injection of hazardous waste [5]. As the character and degree of subsurface activities has expanded, it has become increasingly important to develop methods and techniques capable of addressing the interactions between engineered features and local geology. Information provided by such techniques is of critical value to both industry and regulators: leakage via wellbores and fracture networks is a documented concern [6], and the data provided by such methods is imperative for the development of environmentally safe drilling and injection plans, as well as science-based monitoring and plugging plans for regions (or particular wellbores) at greater risk of connectivity and leakage.

Engineered and induced fractures, such as drilling induced fractures (DIFs) and hydraulically induced fractures (HIFs), are important phenomena that can occur when a wellbore is drilled. DIFs are generated when stresses around a borehole are in excess of those required to cause tensile failure of the wellbore wall [7]. If stress concentrations are great enough, compressive failures can form in the region surrounding the wellbore, leading to wellbore breakout, and the potential loss of wellbore integrity [8]. HIFs are pressure induced fractures that are generated when fluid is injected at high pressure into subsurface formations. If induced fracture networks around collocated wellbores intersect, there may be an increased likelihood of communication between wellbores, and the potential for unwanted

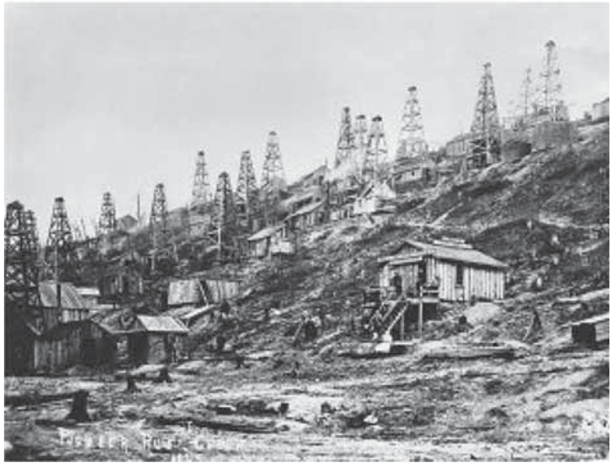

Figure 1 Historical photograph of spatially dense wellbores in TItusville, PA (http://www.acceity.org/2010/09/oil-in-them-thare-hills/).

fluid leakage between such networks. This is particularly of concern for older, structurally unstable wellbores, which may lack adequate casing or cementing necessary for zonal isolation [9]. Furthermore, wellbore spatial densities are likely to be relatively high in regions with historical drilling activity: Before regulations requiring minimum wellbore spacing were implemented, it was a common practice to drill multiple wellbores in close proximity (Figure 1). Furthermore, regions with extensive drilling histories likely have higher wellbore densities as a result of the multiple centuries of exploration of resources in those regions [1].

The generation of induced fractures is greatly dependent upon the structural and lithological characteristics of local geology, which is often difficult to accurately characterize in the absence of expensive geophysical surveys. Consequently, probabilistic-based simulations are often used to model such fracture systems. Several methods for modeling local fracture networks have been proposed in the literature [10–12]. These models often involve the generation of randomly located fractures, with varying degrees of user defined connectivity controls. Because of the importance of wellbore locations; spatial densities; and the potential for overlapping induced fracture networks to create fluid flow pathways, it is important to account for the probable radius of influence of induced fractures around a wellbore, as determined by the local geologic and geospatial attributes. A model capable of incorporating these factors would complement advanced fracture and fracture flow modeling methods. The data provided by such an approach would allow for improved identification of spatial areas at higher risk for communication between wellbores and geologic networks, and allow for the identification of specific regions and wellbores at increased risk for unwanted leakage. Additionally, such data would provide critical information needed for finer scale study or simulation of these spatial extents and wellbores. The data provided by such an approach supports a range of cumulative risk reduction and modeling strategies due to improved uncertainty constrains regarding subsurface engineered-geologic system characteristics. The information provided by this data is of critical importance to both industry and regulators. Due diligence for wellbore construction and injection plans – as well as monitoring and plugging of wellbores – requires data on the relative likelihood of leakage potential in particular areas and through specific wellbores. The model presented in this paper supports these uses.

Because fracture and fracture flow models tend to cross into the realm of big data, computational efficiency is an important consideration for the development of an integrated method for modeling the relationship between geologic and engineered systems. Traditional geospatial models often require access to financially costly methods such as a geographic information system (GIS). Such traditional geospatial models tend to be either cell or raster based, and hence potentially computationally expensive when applied to big data sets. The developing field of graph theory allows for the representation, storage, and manipulation of geospatial data in the form of graphs, which can provide improved computational efficiency when certain graph theoretic data structures are employed [13]. Broadly speaking, graph theory is a field of mathematics and computer science which involves the study of graphs [14]. Graphs are mathematical structures that can represent real world data, and may be used to model pairwise relationships between objects. A graph structure consists of vertices (sometimes called "nodes") and edges. Or, stated more rigorously, a graph G, consists of two discrete sets, V (vertices) and E (edges). The present work specifically uses what is called an "undirected graph," in which the elements of E are unordered pairs of vertices. The vertex set of a graph G is denoted by $V(G)$, and the edge set as $E(G)$ [13]. Graphs are naturally suited for visual representation of spatial relationships: Although the graph structure itself can be either list or matrix based, graph diagrams – such as the example shown in Figure 2 – show how elements in a graph lend themselves to visualization. The vertices in a graph can represent almost any type of data (both abstract and discrete), and likewise, the edges can represent multiple types of relationships

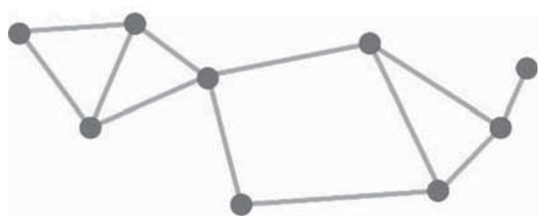

Figure 2 A drawing of an example graph structure showing vertices (blue circles) and edges (orange lines).

between the vertices. Edges of a graph are often associated with a weight function, $w(e)$, that maps each edge e in E to a number. The types of relationships represented by the edges and their associated weights can be as simple as Euclidean distance [14], or as complex as spatio-temporal interactions in complex networks [15, 16].

The graph-based "topological" model presented in this paper (coded in Matlab) employs such graph structures in a novel method for characterizing the spatial radius of influence of induced fractures around a wellbore, and the spatial extents potentially at greater risk for unwanted fluid migration. The algorithm is further compared and contrasted to a standard graph-based k-nearest neighbor algorithm, to demonstrate the importance of incorporating lithologic factors into induced fracture and wellbore connectivity models. The topological model can be used to complement existing fracture models to better account for the reach of induced fractures around a wellbore, and to identify potentially connected wellbores and spatial extents for additional investigation as part of a cumulative strategy to reduce uncertainty inherent to combined engineered and geologic systems.

2 Graph-Based Spatial Analysis

The general workflow for the model is shown in Figure 3, and referenced and described in detail in the forthcoming sections. The model algorithm is coded in Matlab and is developed for easy integration with other commercially available software.

2.1 *Acquire Geologic Data and Define Regional Bounding Lithology*

Subsurface geology – particularly deeper lithology and structure – is impossible to accurately characterize. Even with access to expensive geophysical surveys or map databases, it is often the case that only limited information on the rock type or in situ structural characteristics of strata at certain depth intervals is known [17]. However, regional-scale geological databases (maps, surveys, core logs,

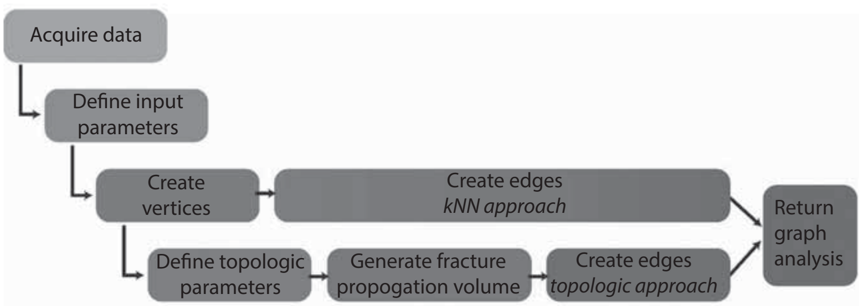

Figure 3 Workflow for preforming a graph analysis to spatially assess induced fractures.

Table 1 Literature derived values for the average fracture radii for a bounding lithology type [17, 31, 32–34].

	Shale	Sand
Induced fracture radius (m)	67	113

stratigraphic columns, or well logs) representing depth-dependent lithology and stresses are generally freely available from state agencies [18] and other public sources. The information provided by such resources can provide sufficient information from which to determine the most geologically brittle and/or over-stressed lithology in a given region. Such a lithology can be considered to be the "bounding lithology", or the regional geologic media that provides the physical bounding conditions for the maximum fracture radius in that region. Arguably, these freely available data sources provide the best information for determining the bounding lithology, given the inherent ambiguities in interpretation of geophysical surveys (such as gravity or magnetic anomalies). Furthermore, given the technological limitations of resolving small scale fractures in geologic media (as well as the changes to fracture networks caused by drilling), finer scale data sources would not necessarily yield much more useful information regarding the present state of the geology.

Once the bounding lithology is identified from the map surveys, a range for the probable fracture radii can be assigned based off existing literature (Table 1). One of the key features of the topological model described in this paper is the adaptability of the analysis for user defined needs: Allowing the user to select the range of fracture radii based on the bounding lithology and associated values reported in Table 1 (or other user defined values), gives the model flexibility for a range of geologic conditions.

2.2 *Details of the Topological Algorithm*

The topological algorithm developed in this model is based on the principle of cylindrical intersection. Wellbore point data are treated as graph nodes, and are imported by the user. The wellbore points (x, y, z) are converted to graph structure by the program. This is accomplished by transforming the geographic coordinates of the wellbore data to an orthogonal, earth fixed frame of reference and Cartesian coordinate system, so that curvature effects can be simplified, and Euclidean distances can be calculated in common units such as meters [19]. Each wellbore node has an associated radius of influence (r), based on the literature derived bounding lithology values (as described in the previous section). The bounding lithology for each wellbore is represented as a numeral in the Matlab program (1 for shale, 2 for sandstone, etc), and the data array for each wellbore node is tagged with its bounding lithology value (alternatively, this numeral value can be entered manually for each wellbore within Matlab, or it can be entered as a user defined radius length value within the program). The bounding lithology represents the most fracture prone strata that the wellbore penetrates, and hence approximates the physical boundary of the potential fracture reach of a wellbore intersecting that strata. Computationally, this is represented as a cylinder's radius. The model computes the associated radius of influence around each wellbore node in 3-dimensions. This results in a series of finite cylinders of radius *r*, associated with each graph node. The algorithm computes the intersection of the cylinders on the graph, and when such intersections are identified, an edge is drawn between the originating nodes. This edge indicates that these originating wellbore nodes fall within the potential fracture reach zone of each other.

2.2.1 Data Acquisition, Conditioning and Quanta

The non-zero graph entries for the topological algorithm are stored in the form of a sparse matrix, making storage and manipulation of large data quanta computationally feasible via compressed sparse row indexing. However, certain practical data acquisition and conditioning procedures are necessary to apply the algorithm to the geologic and geospatial data that the algorithm is designed for. As described in prior sections, it is necessary for the user to have *a priori* knowledge of the predominant geologic characteristics in the subsurface. These data should be derived from geologic maps, surveys, or well logs. In certain cases – particularly where a user is running the algorithm over large spatial extents with multiple wellbores – it may be necessary for the user to perform an initial GIS-based analysis overlaying wellbore data with geologic layers to determine the bounding lithology. The bounding lithology of each wellbore must then be converted to a numeric representation (e.g. "1" for shale, "2" for sandstone), and stored as an element in the row associated with the wellbore points, before it is imported for analysis into the Matlab program. The topological algorithm is also capable of allowing the user to define by hand, both the length of the radius of influence around each wellbore, or simply select by hand the bounding lithology from the pre-populated values within Matlab, when running the algorithm for smaller datasets.

2.2.2 Details of the k-nearest Neighbor Algorithm

Many traditional geospatial models rely on nearest-neighbor associations to assess spatial relationships between geospatial features. For this reason, a standard k-nearest neighbor algorithm is presented and applied to the same data as the topological algorithm, as a means to demonstrate the value added by the latter. Details of the *knn* algorithm can be found in multiple literature sources [20–22]. The basic principle behind the *knn* approach is that the data points (here, wellbores) exist in a metric feature space. The algorithm is only distance based, and does not integrate any information on lithology or wellbore spatial densities. The user must define the *k*, or number of nearest neighbors around each wellbore, and the algorithm makes a distance-based selection.

2.3 *The Value of the Topological Approach Algorithm*

To demonstrate the value added by the newly developed topological algorithm, it is compared and contrasted to the standard *knn* algorithm. The *knn* algorithm establishes an edge between each node and its *k* closest neighbors (in Euclidean distance) for some user-specified integer *k*. The topological approach – which is the primary output of the method – considers both geologic and topologic relationships by computing edges based on bounding lithologies and associated radii of influence. Comparing and contrasting the outputs from two algorithms serves to highlight the importance of accounting for subsurface geologic features in representing real world geospatial data in graph form. A real world example is demonstrated in the sections below.

3 Real World Applications of the Algorithm

3.1 *Bradford Field: Contrasting the Graph-based Approaches; k Sensitivity*

Commercial hydrocarbon exploration has been occurring in the Bradford field in Pennsylvania since the 1800s [9]. The extensive drilling history in the Bradford field results in a high spatial density of wellbores; particularly older "legacy" wellbores which are likely to be structurally unsound and poorly sealed [1] (Figure 4). The model is demonstrated using a synthetic subset of spatially continuous wellbore data in this region.

3.1.1 Data Sources

The Bradford PA wellbore data used in this analysis were obtained from the National Energy Technology Laboratory, United States Department of Energy (NETL/DOE), and were part of an aggregated dataset produced using the methodology described by Dilmore *et al.* [9] and Glosser *et al.* [1]. Initial data provenance for the dataset includes, aeromagnetic surveys [23], digital databases [24, 25], and historical maps and minerals reports [26, 27]. Bounding lithologies were determined either by the individual wellbore records (where available), or by performing a geospatial overlay of the wellbore locations with a freely available geologic map [28].

3.1.2 Results

The two algorithms – the *knn* algorithm and the topological algorithm, were executed on these data. 50 wellbore locations from the Bradford field were subsampled, and associated bounding lithology values for the wellbores were chosen. The *knn* algorithm was run for three nearest neighbor scenarios: $k = 1$ (Figure 5f); $k = 2$ (Figure 5e); and $k = 3$ (Figure 5d). For the topological algorithm, the results are presented in 3 dimensions with the induced fracture radius of influence for

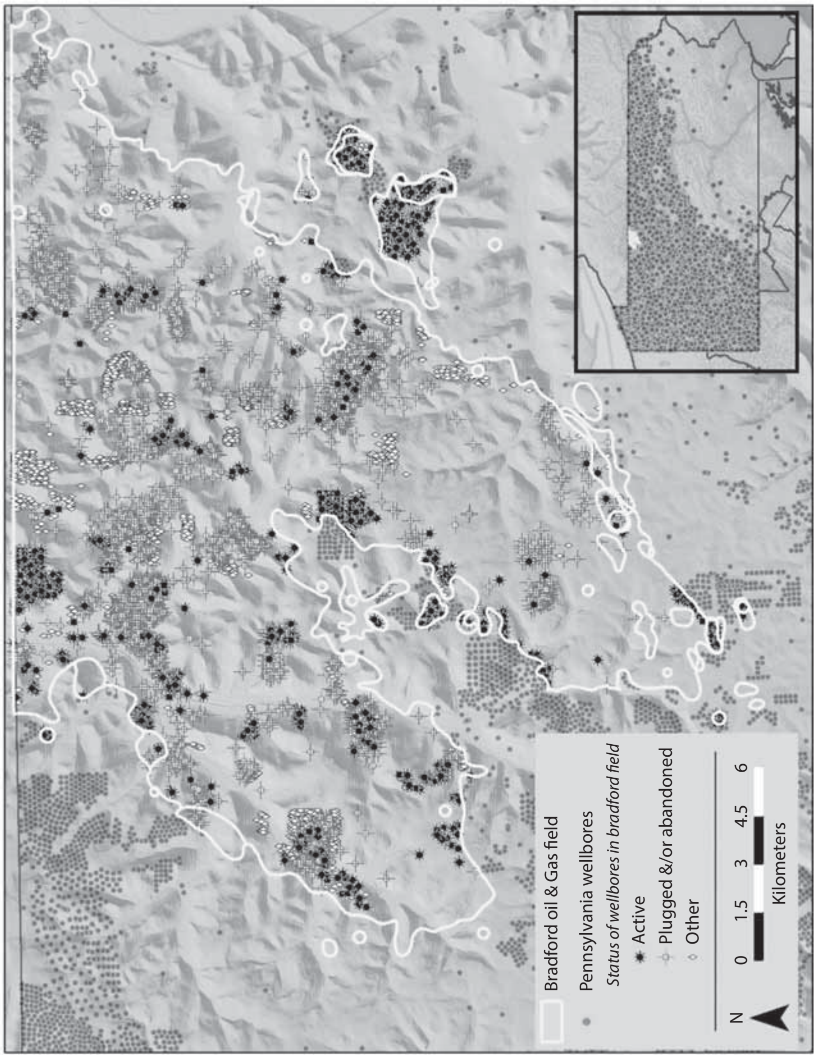

Figure 4 Distribution and known status of wellbores within the Bradford Oil & Gas field, Pennsylvania.

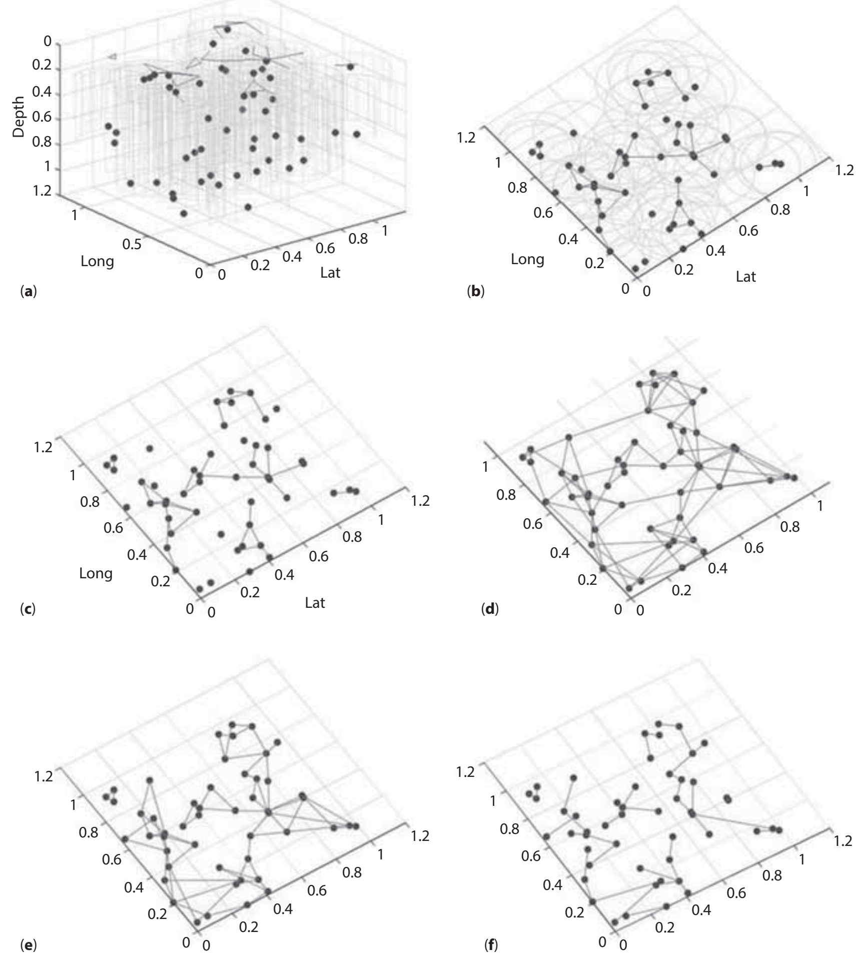

Figure 5 Representative graph output: Graph representation of wellbore points (n = 50). Blue dots represent wellbores (vertices), red lines represent edges, green ellipsoids represent radii of influence (**a**) topological approach in 3D; (**b**) topological approach in 2D form; (**c**) topological approach in 2D form (**d**) knn approach (k = 3) in 2D form (**e**) knn approach (k = 2); (**f**) knn approach (k = 1). Graph axes are represented non-dimensionally.

each wellbore (Figure 5a); in 2D form with the radius of influence (Figure 5b); and in simple graph form showing only the edge connections (Figure 5c). Unlike the *knn* approach – where edges are drawn to each wellbore (node's) k nearest neighbors, in the topological approach, the algorithm draws an edge if and only if a wellbore (node) is within the induced fracture radius of influence of another wellbore (node). That is – if the induced fracture radius of influence of wellbore nodes overlap, then an edge is drawn. It is apparent from the results that the *knn* algorithm is, by definition, sensitive to the number of neighbors, with the edge connections and overall graph connectivity varying greatly based on this value. It is further apparent that even when only one nearest neighbor is selected, the geometry of the edge connections is considerably different than the geometry of the edges in the topologic approach.

3.2 *Armstrong PA: Testing the Algorithms Against a Known Leakage Scenario*

In March 2008, a pressurization of surface casing in a newly drilled oil and gas well in Armstrong County, PA resulted in migration of fluids through two other producing wells [29], mediated by connectivity in the subsurface fracture network. Like the Bradford field, Armstrong County has an extensive regional history of oil and gas exploration. To test the performance of the graph-based spatial analysis on a real world leakage scenario, both the *knn* algorithm ($k = 1$) and the topological algorithm were applied to a subset of wellbores near the pressurized well that caused the leakage event.

3.2.1 Data Sources

The leakage event ("Dayton Investigation") was identified from a Pennsylvania Department of Environmental Protection (PADEP) report on oil and gas well stray gas cases [29], and the wellbore associated with the event was located in the PADEP Office of Oil and Gas Management Compliance Report [30]. A subset of 77 Armstrong County wells from the NETL/DOE Pennsylvania wellbore dataset were spatially selected using ArcGIS, and exported to a .csv file for import to the Matlab program.

3.2.2 Results

Both the *knn* ($k = 1$) algorithm and the topological algorithm were applied to the subset of Armstrong County data. Results are shown in 2D form in Figure 6. Both graph-based analyses identified a cluster of wellbores as associated with the leakage event: However, the *knn* algorithm erroneously identified subsurface connectivity between several more wellbores than were reported as being associated with the stray gas leakage. Overall, the *knn* algorithm (A) results in a qualitatively well connected graph, suggesting that there is extensive subsurface connectivity in the region, even with a k of 1. The wellbore associated with the leakage (black arrow) is shown to be connected to at least 6 neighboring wells via overlapping induced fracture networks. In contrast, in the topological algorithm (B and C), the graph is far less connected; and the affected wellbore is connected to two other affected producing wellbores. The topological algorithm also identified three other clusters of connected wellbores and fracture networks in the region.

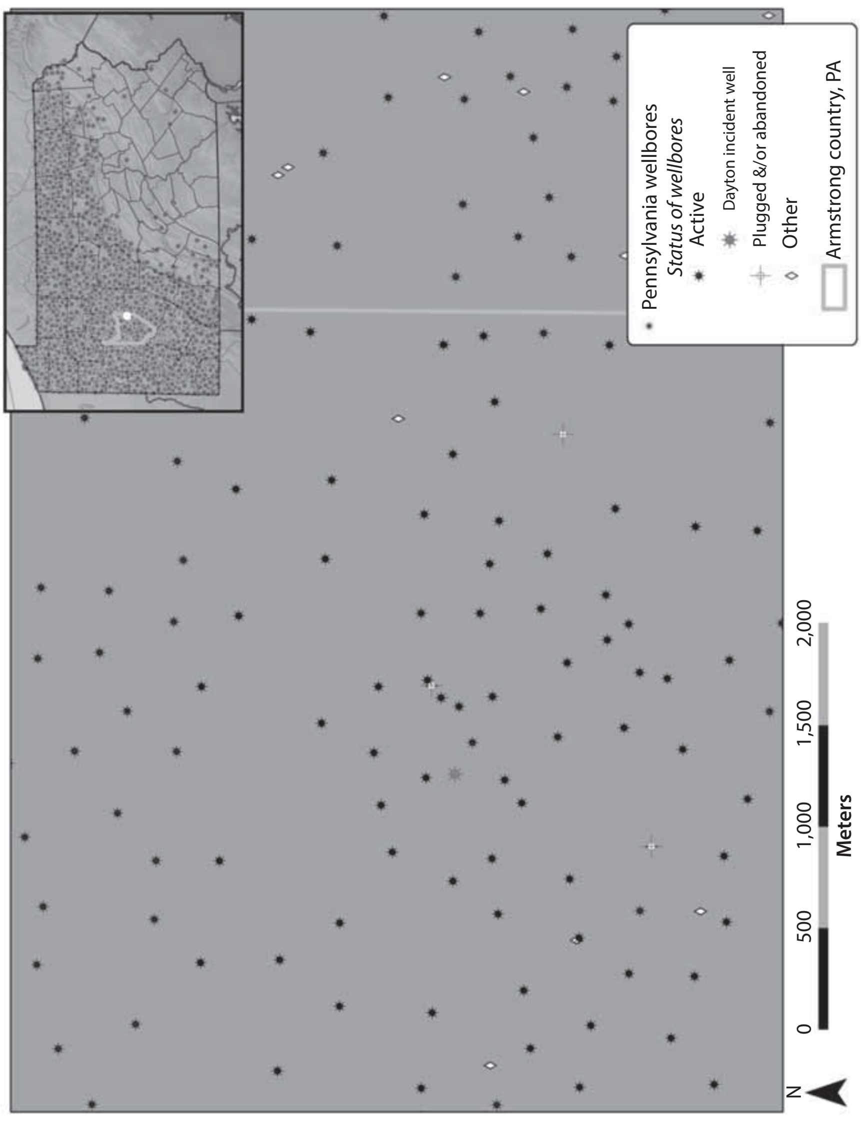

Figure 6 Location of Dayton Incident Well.

4 Discussion

The spatial locations of wellbores, as well as the geologic characteristics of the reservoir, place physical controls on the formation of induced fractures. Many fracture and fracture flow models do not honor wellbore locations and related engineered fractures; or, they consider only the spatial or geologic attributes of the wellbores or reservoir in stochastically generating random fractures. The importance of considering both spatial and geologic attributes in identifying areas at greater risk for overlapping induced fracture networks is highlighted by comparing the topological to a simple distance-based nearest neighbor algorithm. Once identified, these regions can be targeted for finer scale modelling, and used as part of a cumulative modeling strategy to constrain uncertainty in the subsurface. The results of the model are also of value as a standalone piece of data for the development of science-based wellbore drilling, injection, and risk management plans.

As shown in Figure 5, a nearest neighbor approach is by nature sensitive to the selected *k* number of neighbors. The subgraphs produced by the *knn* algorithm contrast considerably from the subgraphs produced by the topological approach. In particular, the *knn* approach suggests far greater connectivity between nodes than does the topologic approach, particularly when higher values of "*k*" are chosen. Conversely, when low values of "*k*" are chosen, several potential connections between nodes may be missed, since the algorithm chooses the nearest neighbor of a node, and not its range of influence on other nodes.

Since the purpose of this model is to identify probable regions of subsurface connectivity – and wellbores and wellbore clusters at greater risk for unwanted fluid migration – the *knn* approach gives a substantially less accurate representation of the spatial extents that are likely to contain overlapping induced fracture networks. First – the *k* is user defined, and the algorithm output is greatly dependent on the user selection. Second – the *knn* algorithm does not take into account the geologically-based induced fracture radius in drawing edges: This results in a connected graph, which is falsely suggestive of subsurface connectivity, and hence, higher risk for fluid leakage through the subsurface and through wellbores. In neglecting to consider geophysical attributes, *knn* relies solely on wellbore neighborhoods, and not actual induced fracture networks in defining connectivity.

In contrast, the topological approach considers both geology and wellbore spatial locations in assessing the probability of induced fracture network overlaps. Edges are drawn if and only if a wellbore falls within the radius of influence of a nearby wellbore. Hence, the subgraphs are representative of both geologic and wellbore networks, and give a more realistic (and in a sense, more conservative) representation of probable connectivity. Furthermore, unlike *knn*, the topological approach is not susceptible to user bias in the selection of neighbors: the algorithm itself automatically identifies the induced fracture radius and number of associated connections (edges), based on literature supplied values for lithologic fracture reach.

When tested against a known leakage scenario, the topological algorithm identifies two producing wellbores associated with the leakage event (Figure 7b and 7c). Although the *knn* algorithm also identifies affected wellbores, it also falsely identifies several other nearby wellbores as being connected (Figure 7a). In addition, the overall *knn* graph is qualitatively connected, resulting in a likely erroneous (overly connected) representation of the subsurface induced fracture network connectivity, even with a k value of 1. In contrast, the topological algorithm identifies 4 clusters of connected wellbores (Figure 7b), including the wellbores associated with the leakage event. The overall topological graph is qualitatively less connected than the *knn* graph, and appears to provide a more realistic representation of the connectivity of wellbores and induced fracture networks. The example application serves to highlight the importance of the

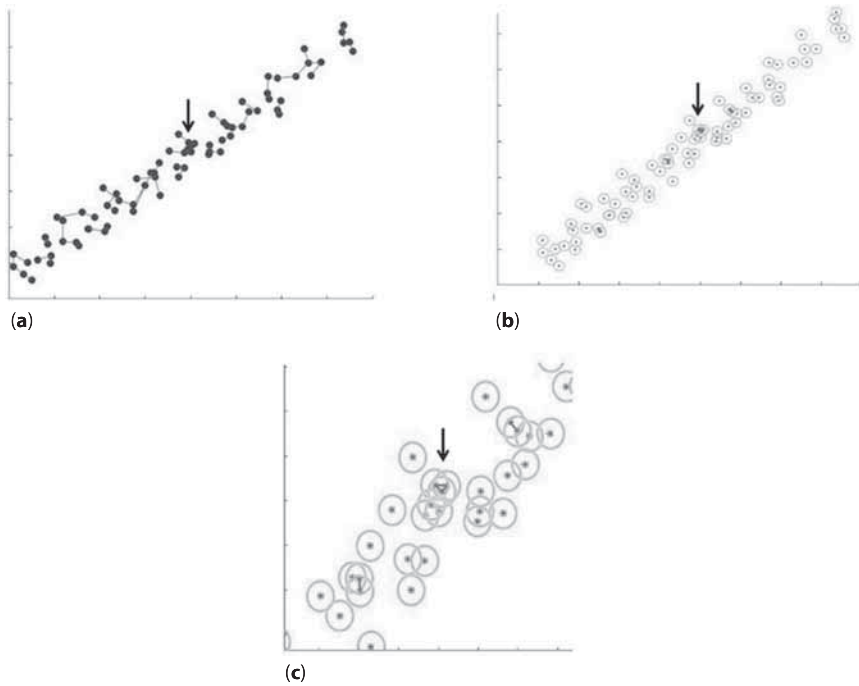

Figure 7 (**a**) Results of knn algorithm on Dayton PA area wellbore ($k = 1$) ($n = 77$); (**b**) Results of topological algorithm on Dayton PA wellbores (n = 77); (**c**) Results of topological algorithm on Dayton PA wellbores, zoomed in on wellbores associated with leakage. Black arrows point to overpressurized wellbore. All results are shown in 2D form for clearer visual representation of results.

geologic factors in assessing regions with probable natural and engineered flow pathways.

4.1 Uses for Industry and Regulators

There is an ongoing need in both industry and regulatory domains for science-based tools from which to develop risk management programs. This is particularly true in complex systems such as subsurface geologic systems, especially those impacted by hydraulic fracturing. The potential for subsurface leakage through induced fracture networks and existing wellbores is a concern to regulators, operators, and public stakeholders. The ability to provide better predictive tools for spatial regions or wellbores at risk for such events meets a critical need for the development of sound risk management strategies. The method presented in this paper can provide valuable information to stakeholders, and helps to reduce uncertainty inherent to these complex systems.

5 Conclusions

Overlapping induced fracture networks between collocated wellbores may increase communication in the subsurface, and create the potential for unwanted fluid flow. The generation of induced fractures is greatly dependent upon the structural and lithological characteristics of local geology, which is often difficult to accurately characterize in the absence of costly geophysical surveys. A robust, adaptable method for analyzing the spatial regions and wellbores at higher risk for subsurface induced fracture connectivity has been developed and presented. The result produced by the method is based on geologic data, and provides a sound basis for reduction of uncertainty inherent in subsurface systems. It is shown that the topological graph theory algorithm is a potentially powerful tool for rapid characterization of subsurface geospatial data. The topologic graph algorithm has several advantages over the *knn* algorithm: it is not susceptible to user error in the selection of a "*k*". And, in accounting for geologic factors, it provides a physically based, and hence more realistic, assessment of probable subsurface connectivity. The algorithm has been successfully demonstrated using a real world leakage scenario. Because subsurface modeling efforts tend to occupy the realm of "big data," the method increases modeling efficiency in two ways: First, the graph structures employed by the method allow for rapid computations involving big data sets; and second, the method can be used to identify spatial extents at greater risk for induced fracture network communication, and hence targeted fracture and fracture flow modeling. The method output can be transformed back into a geographic coordinate system, and/or integrated into existing fracture or fracture flow modeling software, as part of a cumulative modeling strategy for risk mitigation. The information provided by this approach can be used by regulators and industry in developing sound risk management plans related to hydraulic fracturing operations.

Acknowledgements

This work was completed as part of National Energy Technology Laboratory (NETL) research for the Department of Energy's Complementary Research Program under Section 999 of the Energy Policy Act of 2005. This research was supported in part by an appointment to the National Energy Technology Laboratory Research Participation Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education. The authors wish to acknowledge the technical feedback from Kelly Rose, Russell Schwartz, Circe Verba, and Grant Bromhal. Deborah also thanks her family for their support, with special recognition to Miriam and Michael Miller.

References

1. D. Glosser, K. Rose, and J. Bauer, Spatio-temporal analysis to constrain uncertainty in wellbore datasets: an adaptable analytical approach in support of science-based decision making. *Journal of Sustainable Energy Engineering*, in press (2016).
2. State Oil and Gas Groundwater Investigations and Their Role in Advancing Regulatory Reforms a Two-State Review: Ohio and Texas, Groundwater Protection Council (2011).
3. K.M Keranen, H.M. Savage, G.A. Abers, and E.S. Cochran, Potentially induced earthquakes in Oklahoma, USA: Links between wastewater injection and the 2011 Mw 5.7 earthquake sequence. *Geology* **41**(6), 699–702 (2013).
4. D. Soeder, S. Sharma, N. Pekney, L. Hopkinson R. Dilmore, B. Kutchko, *et al.*, An approach for assessing engineering risk from shale gas wells in the United States. In International Journal of Coal Geology, Special Issue: Unconventional Natural Gas and the Environment (2014) (in press).
5. W. Ellsworth, Injection induced earthquakes. *Science*. **341**(6142), 1225942 (2013).
6. J. A. Montague and G. F. Pinder, Potential of hydraulically induced fractures to communicate with existing wellbores. *Water Resources Research* **51**(10), 8303–8315 (2015).
7. M. TIngay, J. Reinicker, and B. Muller, Borehole breakout and drilling-induced fracture analysis from image logs. *World Stress Map Project* (2008).
8. M. D. Zoback, D. Moos, and L. Mastin, Well Bore Breakouts and in Situ Stress, *J. Geophys. Res.*, USGS Staff Published Paper, **90**, B7, pp. 5523–5530 (1985).
9. R.M Dilmore, J.I. Sams, D.B. Glosser, and K.M. Carter, Spatial and temporal characteristics of historical oil and gas wells in Pennsylvania: implications for new shale gas resources. *Environmental Science and Technology* **49,** 12015–12023 (2015).
10. National Energy Technology Laboratory, FracGen and NNFlow software v. 14.9, https://edx.netl.doe.gov/dataset/fracgen-and-nfflow-version-14-9 accessed 4/11/2016.
11. P. Valko and M.J. Economides, Propagation of Hydraulically Induced Fractures – A Continuum Damage Mechanics Approach. *International Journal of Rock Mechanics, Mineral Science, and Geochemistry* **31**(3) 221–229 (1994).
12. C. Zeeb, E. Gomez-Rivez, P. Bons, S. Virgo, and P. Blum, Fracture network evaluation program (FraNEP): A software for analyzing 2D fracture trace-line maps. *Computers and Geoscience* **60**, 11–22 (2013).

13. J. Pach, Geometric Graph Theory, *Mathematical Institute of Hungarian Academy of Sciences*, http://dcg.epfl.ch/page-117408-en.html accessed 4/11/2016.
14. L. D. Cohen and R. Kimmel, Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision* **24**(1), 57–78 (1997).
15. J. Pereira-Leal, A. Enright, and C. Ouzounis. Detection of functional modules from protein interaction networks. *PROTEINS: Structure, Function and Bioinformatics* **54**, 49–57 (2004).
16. S. Yook, Z. Oltvai, and A. Barabasi, Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942 (2004).
17. M. Alizadeh, Z. Movahed, R. Junin, R. Mohsin, M. Alizadeh, and M. Alizadeh, Finding the Drilling Induced Fractures and Borehole Breakouts Using Image Logs. *Journal of Advanced Research* **10**(1), 9–30 (2015).
18. West Virginia Geologic and Environmental Survey, Interactive Mapping http://ims.wvgs.wvnet.edu/index.html (2015).
19. N.-N.G. Intellingence Agency of EUA, *World Geodetic System 1984*, WGS-84. p. 3 (1984).
20. W. Dong, C. Moses, and K. Li, Efficient k-nearest neighbor graph construction for generic similarity measures. *WWW* 577–586 (2011).
21. V. Garcia, E. Debreuve, and M. Barlaud, Fast k nearest neighbor search using GPU. *2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.* **2**, 1–6 (2008).
22. Y. P. Mack and M. Rosenblatt, Multivariate k-nearest neighbor density estimates. *J. Multivar. Anal.* **9**(1), 1–15 (1979).
23. R. Hammack, G. Veloski, and J. Sams, Rapid Methods for Locating Existing Well Penetrations in Unconventional Well Development Areas of Pennsylvania, Unconventional Resources Technology Conference, 20–22 July, San Antonio, SPE-178558-MS (2015).
24. PA DCNR. The Pennsylvania Internet Record Imaging System/Wells Information System (PA*IRIS/WIS): Gateway to Detailed Oil and Gas Well Information Data http://www.pairis.state.pa.us accessed during 2013.
25. Pennsylvania Spatial Data Access Center, Oil and Gas Wells in Pennsylvania, http://www.pasda.psu.edu/ accessed during 2013.
26. R. Arnold and W.J. Kemnitzer, Petroleum in the United States and possessions: a presentation and interpretation of the salient data of geology, technology, and economics of petroleum in each state and possession treated according to the conventional major field divisions; Harper & brothers (1931).
27. US Geological Survey. Minerals Yearbook Digital Collection http://digital.library.wisc.edu/1711.dl/EcoNatRes.MineralsYearBk.
28. Pennsylvania Spatial Data Access Center, Geology of Pennsylvania, http://www.pasda.psu.edu/ accessed during 2015.
29. Department of Environmental Protection, Stray Natural Gas Migration Associated with Oil and Gas Wells http://www.dep.state.pa.us/dep/subject/advcoun/oil_gas/2009/Stray%20Gas%20Migration%20Cases.pdf Accessed during 2016.
30. DEP Office of Oil and Gas Management Compliance Report, inspection ID 1692786 http://www.depreportingservices.state.pa.us/ReportServer/Pages/ReportViewer.aspx?/Oil_Gas/OG_Compliance accessed during 2016.
31. M. Brudy and M.D. Zoback, Drilling induced tensile well fractures: implications for determination of *in situ* stress orientation and magnitude. *International Journal of Rock Mechanics and Mining Sciences* **36**, 191–215 (1999).

32. J.R. de Druezy, P. Davy, and O. Bour, Hydraulic properties of two dimensional random fracture networks following a power law length distribution. *Water Resources Research* **37**(8), 2065–2078 (2001).
33. N.E. Odling, Scaling and connectivity of joint systems in sandstones from western Norway. *J. Struct. Geol.* **19**, 1257–1271 (1997).
34. O. Bour and P. Davy, Connectivity of random fault networks following a power law fault length distribution. *Water Resour. Res.* **33**, 1567–1583 (1997).

