A Novel Reduced Representation Methodology for Provenance Data

Mehmet Gungoren[1], Mehmet S. Aktas[2]
Computer Engineering
Yildiz Technical University
İstanbul, Turkey
e-mail: [1]mehmet.gungoren@std.yildiz.edu.tr, [2]aktas@yildiz.edu.tr

***Abstract*— Learning structure and concepts in provenance data have created a need for monitoring scientific workflow systems. Provenance data is capable of expanding quickly due to the catch level of granularity, which can be quite high. This study examines complex structural information based provenance representations, such as Network Overview and Social Network Analysis. Further examination includes whether such reduced provenance representation approaches achieve clustering effective for understanding the hidden structures within the execution traces of scientific workflows. The study applies clustering on a scientific dataset from a weather forecast to determine its usefulness, compares the proposed provenance representations against prior studies on reduced provenance representation, and analyzes the quality of clustering on different types of reduced provenance representations. The results show that, compared to prior studies on representation, the Social Network Analysis based representation is more capable of completing data mining tasks like clustering while maintaining more reduced provenance feature space.**

Keywords- scientific workflows; scientific data provenance; complex structural information; data provenance; provenance.

I. INTRODUCTION

Provenance can be used as a ground basis for various applications and use cases, such as identifying trust values for data or data fragments [1]. The scientific data provenance collected from the life cycle of a data product is a record of the actions that contribute to the existence of the product. In other words, it identifies the object: the measures that have been implemented, and how, where, and by whom these actions have been implemented. Data provenance determines the extent to which a data product results from raw data. Recording the lineage of a data product is the latest series of activities (or "workflow") applied [2].

Scientific digital data is an important component of metadata for a data object. It can be used to determine the allocation, to identify relationships between objects, and to trace the differences in similar results [3]. Furthermore, in a broader purpose, digital data can help a researcher determine whether a given acquired object can be reused in its work by providing lineage information to support the quality of the data set.

One model that represents such entities and relationships is the Open Provenance Model (OPM) [4]. OPM defines the historical dependencies between entities. The source may be very large, and the catch may be carried out at a high level of granularity. This may occur, for instance, in a workflow system that encourages grained nodes (ex: at a mathematical operation) instead of coarse grains (i.e., at a great work parallel computing.). Moreover, XML-based OPM representation makes it difficult to conduct data analytics tasks on data provenance.

Chen, Plale, and Aktas introduced an approach to deal with large volumes of OPM-based provenance by assuming that the volumes would be large, and then selectively reducing the feature space while simultaneously preserving interesting features so that data mining on the reduced space will yield useful information [5][6]. To do this, they used statistical feature space integrated with a temporal representation of provenance data. Simple structural features (such as the number of in-degree/out-degree) and attribute features (such as number of characters in node name) were also used.

This study takes a slightly different approach in providing a reduced provenance representation of scientific datasets by investigating various complex structural information based representations for scientific data provenance. Algorithms for Network Overview (NO)-metric and Social Network Analysis (SNA)-metric representations of provenance data are introduced. Similar to the work of Chen, Plale, and Aktas, the present study also uses data mining tasks such as clustering to evaluate the usability of the NO-metric and SNA-metric representations of the datasets [5][6]. Such clustering tasks include understanding structures that describe and distinguish the general properties of the datasets in provenance databases to help with detecting any defective provenance data.

This paper provides several new factors to the scientific community. First, it introduces algorithms that convert OPM compatible provenance graphs to Network Overview metrics and Social Network Analysis based reduced provenance representations. It also assesses these complex structural information based representations by using data mining techniques on scientific provenance datasets. The paper evaluates a large weather forecast scientific provenance dataset with provenance traces generated from a real-life workflow [7]. The results demonstrate that, compared with other representation approaches, the SNA-metric representation is more capable of achieving data mining tasks like clustering while maintaining more reduced provenance feature space without any information loss.

The remainder of the paper is organized as follows: Section II reviews related work. Section III introduces the

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-486-2

complex structural information based representation approach. The methodology is explained in Section IV, followed by the experimental evaluation of a large database of provenance in Section V. Section VI concludes the paper and discusses future work.

II. Related Work

Extraction and representation of information about the data-sources has been a subject of research for many years. Many studies have been conducted to represent data sources with reduced representation models and to provide extensive survey studies on data representation methods [1][8]. Agrawal et al. provides one of the first surveys in the context of applied scientific data processing [8]. Antunes et al. offers, in a more general context, a taxonomy for understanding and comparing various data representation techniques [1]. Simmhan et al. first suggested the value that provenance brings to e-Science applications [9]. Davidson et al. introduced the problem of mining and extracting information from provenance for the first time [10].

Santos et al. use clustering techniques to organize collections of workflow graphs [11]. They discuss reduced representations using labeled graphs and multidimensional vectors. However, their representation becomes too large when the workflow is big, and the structural information is lost when using multidimensional vectors.

Bose and Frew introduce a comprehensive survey of lineage retrieval for scientific data processing [12]. In this study, they also introduce a meta-model to identify and assess the components of lineage retrieval systems.

Dealing with temporal data dependencies is yet another problem in discovering hidden information. The goal of temporal data mining is to find hidden relationships between sequences and subsequences [1]. Chen, Plale and Aktas investigate the use of statistical features in order to represent provenance graphs [5][6]. Their study uses non-structural features, such as the number of characters in node labels, and structural features, such as the number of in-degree/out-degree of a node. Chen et al. [5] proposed a temporal graph partitioning algorithm as the basis for an abstract provenance representation. Based on this approach, the non-structural and structural features for each node within each partition are calculated, processed (with statistical operations (average)), and converted into a reduced abstract provenance representation. Chen et al. [5] address the problem of extraction and knowledge discovery from graphs of origin while overcoming the problem of scalability by reducing the large graphic source to a small sequence of temporal representation.

The present study differs from previous work by investigating the use of complex structural information, such as network overview metrics or social network metrics, for the reduced representation of provenance datasets. With the use of temporal representation, the representation sequences of provenance graphs may not be the same length, as the number of partitions will differ between provenance graphs. For example, in large provenance graphs, the number of partitions is high. In return, this increases the size of the reduced temporal provenance representation. However, this study explores the use of network metrics based representation in which the representation sequence is always the same length, regardless of the size of graphs.

III. Network Metrics Based Reduced Provenance Representation

This study defines the complex structural information (network-metrics) feature space vector of a provenance graph. Then, a function that creates the feature vector of the provenance graph based on the network-metrics feature space is defined. In addition two different categories of network metrics are introduced: Network Overview Metrics and Social Network Analysis Metrics. The following definitions work with both categories.

Definition 1. *For a feature space (vector) $N = (V, F, D)$, $V = \{v_1, \ldots, v_n\}$ denotes all the nodes in the provenance graph, the function $F: V \rightarrow D_1 \times D_2 \times \ldots \times D_d$ is a feature function that assigns a feature vector to any node $v \in V$, and the set $D = \{D_1, D_2, D_3, \ldots, D_d\}$ is called the feature space of N. Here, each feature is a network metric and has a numerical value. For example, the diameter of a node within a provenance graph is a feature. For each node in the V, D needs to be calculated.*

Definition 2. *For a network metric based feature space (vector) $N = (V, F, G, D, S)$, a representation function $G: D_1 \times D_2 \times \ldots \times D_i \rightarrow Si$ applies average operation to feature $Di \in D$ of all nodes in V and the set $S = \{S_1, S_2, S_3, \ldots, S_d\}$ is called the feature space of N. Here, for a provenance graph, set S becomes the reduced provenance representation.*

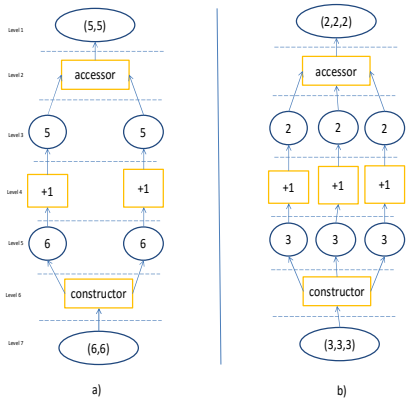

Figure 1. An example illustrating Temporal Partition [4].

A. Network Overview Based Reduced Provenance Representation

Networks have certain attributes that can be calculated to analyze the network's properties or characteristics. These attributes are often called network overview metrics. Due to the directed link structure of the provenance graphs, the present study investigates whether network overview metrics can be utilized as distinguishing features in provenance representation. For this investigation, six commonly used network properties are used: average degree is the degree of a node, or the number of edges that are adjacent to the node;

diameter is the maximal distance between all pairs of nodes; path length is the graph-distance between all pairs of nodes; density is the measurement of how completeness of the network; modularity is the measurement for how well a network decomposes into modular communities; connected component count is the measurement that determines the number of connected components in the network; and finally, giant component is a connected component of a given random graph that contains a constant fraction of the entire graph's vertices.

The pseudo code for the algorithm that creates the network overview metric based reduced representation is presented in Figure 6, and the process through which the NO-metric based representation of a provenance graph is generated is illustrated with an example. Each provenance graph has a link structure. Structural features such as NO-metrics are used as the representative features of a given provenance graph. To facilitate testing the use of NO-metrics, the commonly used NO-metrics were chosen as described above. Therefore, the feature space for each node in Figure 1(a) is D_i = {Average Degree, Diameter, Path Length, Density, Modularity, Connected Component Count, Giant}. For example, the resulting feature space values for the "accessor" node in Figure 1(a) is $D_{accessor}$ = {3.0, 2.0, 0.2, 0.111, 0.34, 0.0, 0.0}. After applying the average to D over all nodes belonging to Figure 1(a), the NO-metrics based reduced feature space is S = {1.0, 2.0, 1.2, 0.111, 0.34, 0.0, 0.0}. Note that, to facilitate testing the representation power of the NO-metrics, this study uses a statistical operation (average) to calculate the signature-representation of a given provenance graph. Other statistical operations may also be tested. Since this study's focus was mainly on the features, it only uses the average function.

B. SNA-Metric based Reduced Provenance Representation

This study also investigates the use of SNA-metrics in provenance representation. Social network analysis is the measurement of relationships between participating entities in a network. In general, the nodes in the network are the people and groups, while the links show relationships or flows between the nodes. To understand networks and their participants, SNA provides metrics to evaluate the location of participating actors in the network. The present research was aimed to find out whether SNA-metrics can capture enough information from provenance graphs to use them as feature space for reduced provenance representation. To do this, commonly used SNA metrics as described below are utilized. Degree centrality measures the "importance" or "influence" of a particular node within a network; betweenness centrality measures the influence over what flows in the network; and closeness centrality measures the visibility of nodes to monitor the information flow in the network. Eccentricity is a measurement that reflects how far, at most, each node is from every other node, and proximity prestige measures how close other actors are to a given actor. The pseudo code for the algorithm that creates the SNA-Metric based reduced representation is presented in Figure 7.

In this representation, SNA-metrics are considered the representative features of a given provenance graph, so the feature space for each node in Figure 1(a) will be D_i = {Degree Centrality, Betweenness Centrality, Closeness Centrality, Density, Eccentricity}. For example, the resulting feature space values for "accessor" node in Figure 1(a) will be $D_{accessor}$ = {0.0, 0.0, 0.222, 2.0, 2.0, 0.0}. After applying the average to D over all nodes that belong to Figure 1(a), the SNA-metrics based reduced feature space is S = {0.012, 0.0, 0.049, 1.13, 1.086, 0.0}. Similar to the NO-Metric representation, to test the representation power of the SNA-metrics, a statistical operation (average) is used to calculate the signature-representation of a given provenance graph.

To further test the use of network-metrics based feature space as a representation, the researchers also apply the Temporal Representation approach introduced by Chen et al. [5][6]. Temporal Representation defines a strict, totally ordered partition that divides a provenance graph into a list of non-empty subsets. Given any provenance graph, Chen's Temporal Representation algorithm (Logical-P algorithm) generates a unique strict totally ordered partition. Figure 1 shows the temporal partitions obtained from three different provenance graphs. To test the usability of the Temporal Representation approach for feature space, simple structural feature sets were used in previous studies, including the node's in-degree and out-degree amounts [5][6]. In this study complex structural information based feature space is used for structural features.

C. Temporal Representation with NO and SNA metrics Based Feature Spaces

Chen et al. [5] define the feature space for a node subset and the statistical feature function that converts the provenance graphs, partitioned into subsets using Logical-P algorithm, into statistical feature space. Based on these definitions, the present study captures the following features for NO-metrics feature space from each subset Vi: <Average Degree, Diameter, Path Length, Density, Modularity, Connected Component Count, Giant>. The Temporal Representation with NO-metric based Feature Space is tested on the partitioned provenance graph shown in Figure 1. For example, the resulting provenance partition of Figure 1(a) is represented as: S={<1.0, 2.0, 0.2, 0.111, 0.34, 0.0, 0.0>, <3.0, 2.0, 0.2, 0.111, 0.34, 0.0, 0.0>, <2.0, 0.0, 0.0, 0.111, 0.34, 0.0, 0.0>, <2.0, 1.0, 0.1, 0.111, 0.34, 0.0, 0.0>, <2.0, 1.0, 0.1, 0.111, 0.34, 0.0, 0.0>, <3.0, 2.0, 0.2, 0.111, 0.34, 0.0, 0.0>, <1.0, 2.0, 0.2, 0.111, 0.34, 0.0, 0.0>}. Likewise, the following features for SNA-metrics feature space from each subset Vi are: <Degree Centrality, Betweenness Centrality, Closeness Centrality, Density, Eccentricity>. The Temporal Representation with SNA-Metric based Feature Space was tested on the partitioned provenance graph shown in Figure 1. The resulting provenance partition of Figure 1(a) is represented as: S={<0.111, 0.0, 0.111, 1.0, 1.0, 0.0>, <0.0, 0.0, 0.222, 2.0, 2.0, 0.0>, <0.0, 0.0, 0.0, 0.0, 0.0, 0.0>, <0.0,

0.0, 0.111, 1.0, 1.0, 0.0>, <0.0, 0.0, 0.111, 1.0, 1.0, 0.0>, <0.0, 0.0, 0.222, 2.0, 2.0, 0.0>, <0.111, 0.0, 0.111, 1.0, 1.0, 0.0>}.

D. Feature Selection Methodology

The selection of an optimal feature set depends upon both the mining targets and the nature of the provenance [5][6]. In this study, the target in unsupervised clustering is to group together provenance instances based on their original experiment. Therefore, the aim is to select a feature set that can discriminate between provenance instances of different experiments. In other words, the distance between two representations of provenance derived from the same experiment should be smaller than the distance between representations of provenance derived from different experiments. More features result in more distinguishing power while adding irrelevant features to a dataset often decreases the accuracy of the unsupervised clustering approaches. This study investigates whether network overview metrics or social network analysis metrics have enough discriminating power for unsupervised clustering tasks in scientific provenance datasets.

This study assumes that provenance graphs from related experiments have similar structure and similar attribute information, while provenance graphs from different experiments are either different in attribute information or in structural information. To this end, while using any feature set, Figure 1(a) and Figure 1(b) should be clustered together. To test this assumption, the Euclidean distance, calculated from the simple structural feature set (proposed by [5][6]), NO-metrics, and SNA-metrics based complex structural feature sets were investigated. The results of this investigation are shown in Table I. These distances show whether two graphs are relatively closer to each other for differing metrics. The results indicated that the distance between the provenance graphs in Figure 1(a) and Figure 1(b) turned out to be closer to each other for certain features more than others, meaning the distances calculated from complex structural information based feature space representations had more distinguishing power compared to other representations. It is important to note that the values of features were normalized before calculating the Euclidean Distance to make sure that all metrics contribute equally to the results. The normalized feature values scaled between 0 and 1.

TABLE I. EUCLIDEAN DISTANCE FOR DIFFERING FEATURE SETS AMONGST DIFFERENT REPRESENTATIONS

	Distance in Simple Structural Feature Set (from [5][6])
Figures 1(a) – 1(b)	0.7454
	Distance in NO-Metric Based Structural Feature Set
Figures 1(a) – 1(b)	0.5408
	Distance in SNA-Metric Based Structural Feature Set
Figures 1(a) – 1(b)	0.4429

Chen reported that the disadvantage of the simple structural feature set (i.e., amount of in-degree/out-degree) is that if two provenance graphs have the same structure but different node/edge information, it would be impossible to distinguish between the two through the structural feature set alone. To this end, Chen proposed an extension to the set by further splitting the edges into different types in OPM (i.e., used, wasGeneratedBy, wasControlledBy, wasTriggeredBy, wasDerivedFrom), so that one can discriminate graphs that have similar structure but are semantically different. The present study follows similar methodology, but assigns differing weights to each semantically different edge. In distinguishing the semantically different but structurally the same provenance graphs, the approach works with both network overview and social network analysis metrics.

IV. METHODOLOGY

This study addresses whether it is possible to detect failed workflows in a provenance dataset without the guidance of a workflow script or to detect provenance variants caused by either workflow execution failure or provenance capture failure. To answer these questions, the usability of complex structural information based reduced provenance representations is explored with a focus on finding variants to help detect faulty provenance data by checking cluster centroids in the case where correct and faulty provenances are naturally separated into different clusters.

Much like the study by Chen et al., the present study investigates the best unsupervised algorithm for the graph structure based provenance representation from several popular clustering algorithms: centroid-based (k-means), distribution-based (DBScan), and density-based (EM algorithm) [5][6]. Results indicated that the k-means algorithm produced the highest quality clusters. Hence, in this study, the k-means algorithm was selected to show the usefulness of the proposed representations.

Weka libraries and SimpleKmeans were used in the k-means algorithm. Using Euclidean distance as the similarity function in k-means limited the application of k-means to same-length representation. Since both NO-metric and SNA-metric based representations provide same-length representation for all provenance graphs, this was not problematic. However, when testing the temporal representation with network-metric based representation, this issue was limiting. To overcome this, the researchers followed the same approach as Chen et al., filling missing features with a special value of 0 to provide good performance in clustering results.

V. EXPERIMENTAL EVALUATION

To prove that the provenance representations using the graph partitioning approach can support scalable analysis while being resilient to errors in provenance data, the experiment is conducted using a 10GB provenance database with known failure patterns [7]. This 10GB database of provenance is populated from a workload of roughly 48,000 workflow instances that are modeled based on six real

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-486-2

workflows. The LEAD NAM, SCOOP, and NCSF are weather and ocean modeling workflows, Gene2Life and MOTIF are bioinformatics and biomedical workflows, and the Animation workflow carries out computer animation rendering. Some of the workflows are small, having few nodes and edges, while others like Motif have a few hundred nodes and edges. In the 10GB database, each of the six workflow types has 2000 instances per failure mode, with the failure modes as following: No failures and dropped notifications (success case), 1% failure rate, 1% dropped notification rate, 1% failure rate, and 1% dropped notification rate.

The Karma provenance system is used to store the 10GB provenance dataset and to export the provenance in the form of OPM graphs [9]. From the provenance graphs, the adjacency matrix is generated. Then the complex network metrics and social network analysis metrics are calculated and stored.

The evaluation strategy used here follows the methodological analysis first described by Chen, Plale and Aktas [5][6]. No structural information is assumed in the representation of the provenance datasets within the 10GB database.

In order to help understand how the graph-structure based representations identify clusters, NAM workflow provenance datasets from weather forecast workflow were chosen, as Chen et al. identified that this is the best illustration of provenance capture from scientific workflows [5][6]. The temporal representation of the NAM provenance datasets has shown that NAM datasets include provenance graphs with varying numbers of partitions, ranging from 2 to 10. It turns out that a NAM provenance graph with 10 subsets is a complete graph, while provenance graphs with less than 10 partitions are incomplete and caused by dropped notifications. To test the usefulness of the graph structure based reduced representation approaches, a k-means clustering algorithm was applied to the provenance representations of NAM provenance datasets. Purity, Normalized Mutual Information (NMI), and Within-Cluster Sum of Squares (WCSS) were used to compare the performance of different clustering techniques.

The grouping of the NAM provenance dataset (based on the temporal length defined by Chen et al.) is shown in Figure 2. The grouping results indicate that 78% of the NAM provenance graphs have the largest possible number of partitions, and 6% of the graphs have small partitions ranging from 2 to 4. Small-partitioned provenance graphs indicate dropped notifications or early failures that might happen in the NAM workflow execution.

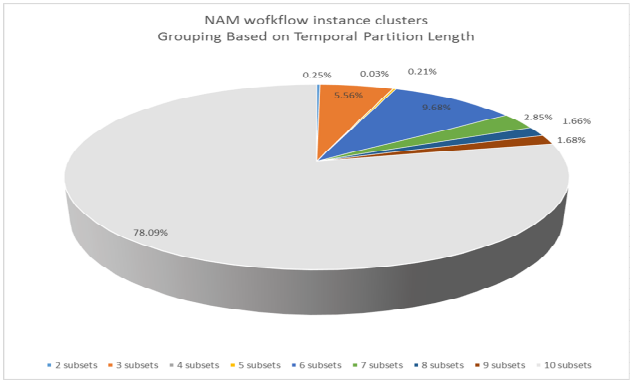

Figure 2. NAM workflow clusters - Grouping Based on Temporal Representation

To test the clustering on the reduced representations, the grouping results (shown in Figure 2) are used as the golden standard for Purity and NMI metrics. The clustering is evaluated on both NO-metric and SNA-metric based reduced provenance representations. The SimpleKMeans clustering algorithm with Euclidean Distance measurement is then applied to these representations. Unlike the temporal representation, the graph structure based representation has representation sequences of uniform length. Thus, the k-means clustering algorithm is applied without any limitations.

TABLE II. NAM WORKFLOW CLUSTERING RESULTS FOR K=9 AND VARYING REDUCED PROVENANCE REPRESENTATIONS

	Purity	NMI	WCCS
Network Overview	0.798375	0.503209	171.6229
SNA	0.922	0.516648	43.1794
TR+SNA	0.938625	0.555053	561.3776

Table II gives the summary of the quality of clustering results when k = 9. The results indicate that SNA-metric based representation and Temporal Partitioning with SNA representation lead to high quality clustering results. SNA metrics were better than NO metrics in capturing the complex structural information as features. Purity and NMI metrics were computed by calculating the correctly assigned workflow instances. To do this, the grouping results shown in Figure 2 were used as the golden standard. To further understand the behavior of clustering for varying reduced representation sizes, different k values were tested. To choose the number of cluster k, the quality of resulting clusters was plotted by computing Purity as an external evaluation criterion.

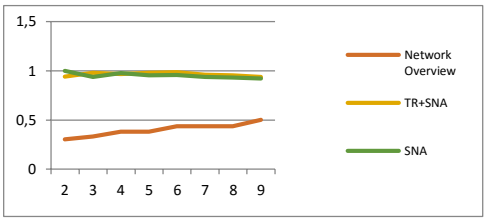

Figure 3. NAM - Purity results

Figure 3 shows that SNA-metrics based reduced provenance representation produced high quality clustering results for Purity Metric. This is because the link structure of the directed graphs contains enough information that can be used as features to differentiate the features and produce good clustering results. For example, an SNA-metric like prestige captures the popularity information of the nodes within a highly connected graph. The overall popularity value is expected to be higher in large graphs compared to small graphs. Similarly, the number of central nodes in large-scale linear provenance graphs is expected to be higher than in small-scale ones. Hence, the overall centrality value is expected to be high for large-scale graphs. The present investigation has shown that graph-structure based metrics can produce high quality clustering results while maintaining reduced provenance representation. The results also indicate that SNA metrics (such as centrality and prestige) capture the directed link structure of given provenance graphs better than network overview metrics based representation.

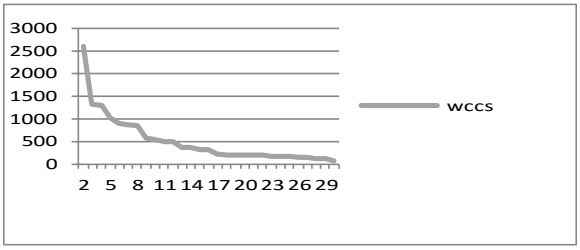

Figure 4. NAM – SNA k = [2,30] WCSS

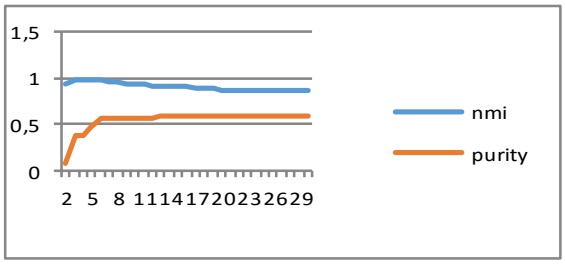

Figure 5. NAM - SNA k=[2,30] NMI Purity

Figure 4 and Figure 5 show the results of an experiment evaluating a k-means clustering algorithm on SNA-metric based representation by plotting the within cluster sum of squares and computing NMI and Purity for increasing values of k. The results indicated that after K reaches a value of 9, the Purity value is high and stable. This shows that the reduced representation in SNA-metric domain can lead to efficient unsupervised clustering.

As mentioned earlier, provenance graphs with few partitions indicate small provenance graphs. These graphs are often incomplete and may be caused by early failures in workflow execution or failures in provenance capture. Since the size of such graphs is small, their link structure information will not be enough to provide accurate clustering. Hence, if a provenance database contains a high number of small-size provenance graphs, graph structured based feature space is not expected to be effective in clustering. However, for scientific workflows where the number of nodes is high, such as NAM whether forecast workflow, such provenance representation can be useful. The present experiments indicated that, for a noisy large-scale scientific workflow dataset such as a NAM dataset, SNA-metric based representations provided high quality clustering.

VI. Conclusions and Future Work

This study investigates various graph structure based representations, such as Network Overview and Social Network Analysis metric representations for scientific data provenance. It also investigates whether such reduced provenance representation approaches lead to effective clustering on scientific data provenance for understanding the hidden structures within the execution traces of scientific workflows. Clustering was applied to the graph structure based representations on 10 GB scientific dataset to determine their usefulness. The graph structure based provenance representations were compared against other reduced provenance representation approaches. The quality of clustering on different types of reduced provenance representations was analyzed, and the results were reported. The results show that, compared with other representation approaches, the SNA-metric representation is more capable of data mining tasks like clustering while maintaining more reduced provenance feature space. In future work, the researchers plan to test the network-metrics based representations with a real-life dataset obtained from the AMSR-E satellite. They also plan to extend their work to combine both Network Overview metric and Social Network Analysis metric representations in one vector. Work remains to test whether complex structural information based provenance representation is useful in other data mining tasks, such as classification and association rule mining. The researchers plan to adapt state-of-the-art approaches for dimensionality reduction and high-contrast feature selection in future work, and to expand tests on the other scientific workflow datasets that are available in the 10GB provenance database introduced by Cheah et al. [7].

Acknowledgments

The authors would like to thank Dr. Beth Plale, the director of the Data to Insight Center (D2I) from Indiana University, for helping us throughout our studies, including but not limited to the use of Karma Service implementation, the computational facilities of the D2I Center, and the 10GB scientific provenance database. The authors also thank Dr. Peng Chen for his valuable insights on the use of temporal provenance representation. This study was supported by TUBITAK's (3501) National Young Researchers Career Development Program (Project No: 114E781, Project Title: Provenance Use in Social Media Software to Develop Methodologies for Detection of Information Pollution and Violation of Copyrights).

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-486-2

REFERENCES

[1] C. M. Antunes, and A. L. Oliveira, "Temporal data mining: An overview" KDD Workshop on Temporal Data Mining, 2001, pp. 1–13.

[2] M. Aktas, B. Plale, D. Leake and N. K. Mukhi, "Unmanaged Workflows: Their Provenance and Use", Q. Bai, Q. Liu eds. Data Provenance and Data Management in eScience, Studies in Computational Intelligence series, Springer, Vol 426, 2013, pp. 59-81.

[3] S. Bechhofer, D. D. Roure, M. Gamble, C. Goble and L. Buchan, "Research objects: Towards exchange and reuse of digital knowledge" The Future of the Web for Collaborative Science, 2010.

[4] L. Moreau, and et al., "The open provenance model core specification (v1. 1)" Future Generation Computer Systems. 27, 2011, pp. 743–756.

[5] P. Chen, and B. Plale, and M. Aktas, "Temporal representation for scientific data provenance" eScience 2012, pp. 1-8.

[6] P. Chen, and B. Plale, and M. Aktas, "Temporal Representation for Mining Scientific Data Provenance" Future Generation Comp. Syst. 36, 2014, pp. 363-378.

[7] Y. Cheah, B. Plale, J. Kendall-Morwick, D. Leake and L. Ramakrishnan, "A Noisy 10GB Provenance Database" 2nd Int'l Workshop on Traceability and Compliance of Semi-Structured Processes (TC4SP), co-located with Business Process Management (BPM), 2011, pp. 370-381.

[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules" 20th Int. Conf. Very Large Data Bases, VLDB. 1215, pp. 487-499, 1994.

[9] Y. L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science" ACM SIGMOD Record. 34, 2005, pp. 31–36.

[10] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities" SIGMOD Conf., 2008, pp. 1345–1350.

[11] E. Santos, L. Lins, J. P. Ahrens, J. Freire, and C. T. Silva, "A first study on clustering collections of workflow graphs" IPAW, 2008, pp. 160-173.

[12] R. Bose and J. Frew., "Lineage retrieval for scientific data processing: a survey" ACM Comput. Surv. 37(1), March 2005, pp. 1-28.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-486-2

```
1: assign 0 to total_avg, total_dia, total_pLen, total_dns, total_mdl, total_ccc, total_gia
2: for all nodes n in D do
3:          assign convertOPMtoAdjacencyMatrix(n) to adjacency_matrix
4:          assign averageDegree(adjacency _matrix) to avg
5:          assign diameter(adjacency _matrix) to dia
6:          assign pathLength(adjacency _matrix) to pLen
7:          assign density(adjacency_matrix) to dns
8:          assign modularity(adjacency_matrix) to mdl
9:          assign connectedComponentCount(adjacency_matrix) to ccc
10:         assign giant (adjacency_matrix) to gia
11:         assign total_avg + avg to total_avg; total_dia + dia to total_dia; total_pLen + pLen to  total_pLen;
total_dns + dns to total_dns; total_mdl + mdl to total_mdl; total_ccc + ccc to total_ccc; total_gia + gia to total_gia
12: end for
13: assign {total_avg/n, total_dia/n, total_pLen/n, total_dns/n, total_mdl/n, total_ccc/n, total_gia/n} to Feature_Space
```

Figure 6. Pseudo code for the algorithm CN metrics based reduced provenance representation

```
1:T <- set of all node in G
2:for all node k in T do
3:          assign empty to Stack(S)
4:          assign empty to LinkedList(Q)
5:          addLast k to LinkedListQ)
6:          while LinkedList(Q) is not empty do
7:                    assign removeFirst from LinkedList(Q) to v
8:                    add Stack(S) to v
9:                    for all edge of v do
10:                             add opposite node to Linked(Q)
11:                   end for
12:         end while
13:         for all nodes in T do
14:                   if count of neighbour of s > 0 do
15:                             add count of neighbour of s to closenessCentrality
16:                             add count of neighbour of s to proximityPrestige
17:                             add max in count of neighbour of s or eccentricity of s to eccenricity
18:                   end if
19:         end for
20:         for all out going edges from s do
21:                   if count of neighbour of s > 0 do
22:                             add count of neighbour of s to degreeCenrality
23:                   end if
24:         end for
25:         for all incoming edges from s do
26:                   if count of neighbour of s > 0 do
27:                             add count of neighbour of s to degreePrestige
28:                   end if
29:         end for
30:         if s is reachable from other nodes
31:                   closenessCentrality /= reachableCount
32:                   proximityPrestige /= reachableCount
33:         end if
34:         closenessCentrality /= All nodes count - 1
35:         degreeCentrality /= All nodes count - 1
36:         degreePrestige /= All nodes count - 1
37:end for
38:do normalization for all attributes of SNA
```

Figure 7. Pseudo code for the algorithm SNA metrics based reduced provenance representation

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-486-2