REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

FEATURE EXTRACTION METHODOLOGY FOR PROVENANCE DATA USING SNA METRICS

MEHMET GÜNGÖREN

MSc. THESIS
DEPARTMENT OF COMPUTER ENGINEERING
PROGRAM OF COMPUTER ENGINEERING

ADVISER
ASST. PROF. DR. MEHMET SIDDIK AKTAŞ

İSTANBUL, 2016

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

FEATURE EXTRACTION METHODOLOGY FOR PROVENANCE DATA USING SNA METRICS

A thesis submitted by Mehmet GÜNGÖREN in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 11.07.2016 in Department of Computer Engineering, Computer Engineering Program.

Thesis Adviser

Asst. Prof.Dr. Mehmet Sıddık AKTAŞ

Yıldız Technical University

Approved By the Examining Committee

Asst. Prof. Dr. Mehmet Sıddık AKTAŞ

Yıldız Technical University _____

Assoc. Prof. Dr. Songül ALBAYRAK, Member

Yıldız Technical University _____

Assoc. Prof. Dr. Ahmet SAYAR, Member

Kocaeli University _____

This study was supported by Scientific and Technological Research Council of Turkey (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).

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Mehmet Sıddık Aktaş, for his support, patience, and encouragement throughout my graduate studies. It is not often that one finds an advisor and colleague that always finds the time for listening to the little problems and roadblocks that unavoidably crop up in the course of performing research. His technical and editorial advice was essential to the completion of this dissertation and has taught me innumerable lessons and insights on the workings of academic research in general.

Finally, I must express my very profound gratitude to my parents and to my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

July, 2016

Mehmet GÜNGÖREN

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

D21	Data to Insight Center
NO	Network Overview
NMI	Normalized Mutual Information
OPM	Open Sources Provenance Model
SNA	Social Network Analysis
TR	Time Representation
TUBITAK	Scientific and Technological Research Council of Turkey
WCSS	Within-Cluster Sum of Squares
XML	Extensible Markup Language

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABSTRACT

FEATURE EXTRACTION METHODOLOGY FOR PROVENANCE DATA USING SNA METRICS

Mehmet GÜNGÖREN

Department of Computer Engineering

MSc. Thesis

Adviser: Asst. Prof. Dr. Mehmet Sıddık AKTAŞ

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.

Key words: Scientific workflows, scientific data provenance, complex structural information, data provenance, provenance

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ÖZET

PROVENANCE VERİSİ İÇİN YENİ BİR AZALTIMLI GÖSTERİM YÖNTEMİ

Mehmet GÜNGÖREN

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yard. Doç. Dr. Mehmet Sıddık AKTAŞ

Provenans (Veri kökü) verisinin yapısını ve konseptini öğrenmek için bir iş akışını takip eden iş akışı sistemine ihtiyaç vardır. Provenans (Veri kökü) verisi çok hızlı büyüdüğünden ayrıntıyı yakalamak gittikçe zorlamaktadır. Bu çalışma da kompleks yapılı bilgi tabanlı provenance göstermi üzerinde çalışılmıştır. Bilimsel iş akışlarının çalışma izlerinin takibi için yapılan boyut azaltımlı provenans (veri kökü) gösterimi yöntemi analizleri gizlenmiş olan yapıları başarılı bir şekilde gruplamayı başarmıştır. Bu çalışmada aynı zamanda kompleks yapıdaki veriler içerisinden en çok kullanışlı, faydalı olan yapıları da tespit etmektedir. Bu çalışmanın yapılan bir önceki çalışmalardan farkı önceki çalışmaların kararsız boyutlarda olması bu yöntemde ise sabit ve daha az boyutlu veriler ile çalışarak daha iyi başarım elde edilmesidir. Yapılan çalışma da yeni yöntemin veri madenciliği algoritmaları kullanılarak daha üstün başarım elde ettiği ortaya konulmuştur.

Anahtar Kelimeler: Bilimsel iş akışları, bilimsel provenans verisi, kompleks yapılı bilgi, provenans verisi, provenans

CHAPTER I

INTRODUCTION

1.1. Literature Review

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 lifecycle 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.

1.2. Objective of the Thesis

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.

The contributions of the thesis are as following. 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 thesis evaluates a large weather forecast scientific provenance datasetwith 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 thesis is organized as follows: Chapter II reviews related work. Chapter III introduces the complex structural information based representation approach. The methodology is explained in Chapter IV, followed by the experimental evaluation ofa large database of provenance in Chapter V. Chapter VI concludes the thesis and discusses future work.

1.3. Hypothesis

In this study, I have tried to address the following research questions: “Can we detect failed workflows in a provenance dataset, without the guidance of a workflow script?”, “Can we detect provenance variants, in which the cause of variants may either come from workflow execution failure or provenance capture failure?”. To answer these questions, I explore the usability of complex structural information based reduced provenance representations. Ihave focussed on finding variants that would help to detect faulty provenance data by checking cluster centroids in the case where correct and faulty provenance are naturally separated into different clusters.

In addition, “Can data mining tasks be used on provenance graph?” is an important question to evaluate the reduced representation algorithm results. On experimental evaluation section these question answer will be answered.

CHAPTER 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 modelsand 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.

CHAPTER III

METHODOLGY OF REDUCED REPRESENTATION: NETWORK METRICS BASED REDUCED PROVENANCE REPRESENTATION

This study addresseswhether 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. Table 9.2 is shown compared results of clustering algorithms on appendix section.

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.

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= {v1,…, vn} denotes all the nodes in the provenance graph, the function F: V → D1 x D2 x … x Dd is a feature function that assigns a feature vector to any node v ∈ V, and the set D = {D1,D2,D3, …, Dd} is called the feature space of N. Here, each feature is a network metric and has a numerical value. For example, 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: D1 x D2 x … x Di → Si applies average operation to feature Di∈ D of all nodes in V and the set S = {S1,S2,S3, …, Sd} is called the feature space of N. Here, for a provenance graph, set S becomes the reduced provenance representation.*

Figure 3.1 An example illustrating Temporal Partition [4]

1.4. 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. In this study, due to directed link structure of the provenance graphs, we investigate whether network overview metrics can be utilized as distinguishing features in provenance representation. For this investigation, we use six commonly used network properties as listed below on Table 3.1.

Table 3.1 Network Overview Based Provenance Representation Metrics

	Metric	Explanation
1	Average Degree	The degree of a node is the number of edges that are adjacent to the node.
2	Diameter	The maximal distance between all pairs of nodes.
3	Path Length	The graph-distance between all pairs of nodes.
4	Density	The measurement for how close the network is to complete.
5	Modularity	The measurement for how well a network decomposes into modular communities.
6	Connected Component Count	The measurement that determines the number of connected components in the network.
7	Giant Component	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 9.1, 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 3.1(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 3.1(a) is $D_{accessor}$ = {3.00, 2.00, 0.20, 0.11, 0.34, 0.00, 0.00}. After applying the average to D over all nodes belonging to Figure 3.1(a), the NO-metrics based reduced feature space is S = {1.00, 2.00, 1.20, 0.11, 0.34, 0.00, 0.00}. 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 usesthe average function.

1.5. SNA-Metric Based Reduced Provenance Representation

We also investigated the use of SNA-metrics in provenance representation. Social network analysis is the measuring 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. In this study, we conduct an experimental study to find out whether SNA-metrics can capture enough information from provenance graphs that can be used as feature space for reduced provenance representation. To do this, commonly used SNA metrics as described belowon Table 3.2 are utilized.

Table 3.1 SNA Based Provenance Representation Metrics

	Metric	**Explanation**
1	Degree Centrality	Measures the "importance" or "influence" of a particular node within a network.
2	Betweenness Centrality	Measures the influence over what flows in the network.
3	Closeness Centrality	Measures the visibility of nodes to monitor the information flow in the network.
4	Eccentricity	The measurement reflects how far, at most, is each node from every other node.
5	Proximity Prestige	Measures how close other actors are to a given actor.
6	Degree Prestige	Measures counts the number of inbound actors are to a given actor.

The pseudo code for the algorithm that creates the SNA-Metric based reduced representation is presented in Figure 9.2.

In this representation, SNA-metrics are considered the representative features of a given provenance graph, so the feature space for each node in Figure 3.1(a) will be D_i = {Degree Centrality, Betweenness Centrality, Closeness Centrality, Density, Eccentricity,Proximity Prestige, Degree Prestige}. For example, the resulting feature space values for “accessor” node in Figure 3.1(a) will be $D_{accessor}$ = {0.00, 0.00, 0.22, 2.00, 2.00, 0.00}. After applying the average to D over all nodes that belong to Figure 3.1(a), the SNA-metrics based reduced feature space is S = {0.01, 0.00, 0.05, 1.13, 1.09,

0.00}. 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 3.1 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.

1.6. 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 3.1. For example, the resulting provenance partition of Figure 3.1(a) is represented as: S={<1.00, 2.00, 0.20, 0.11, 0.34, 0.00, 0.00>, <3.00, 2.00, 0.20, 0.11, 0.34, 0.00, 0.00>, <2.00, 0.00, 0.00, 0.11, 0.34, 0.00, 0.00>, <2.00, 1.00, 0.10, 0.11, 0.34, 0.00, 0.00>, <2.00, 1.00, 0.10, 0.11, 0.34, 0.00, 0.00>, <3.00, 2.00, 0.20, 0.11, 0.34, 0.00, 0.00>, <1.00, 2.00, 0.20, 0.11, 0.34, 0.00, 0.00>}. 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 3.1. The resulting provenance partition of Figure 3.1(a) is represented as: S={<0.11, 0.00, 0.11, 1.00, 1.00, 0.00>, <0.00, 0.00, 0.22, 2.00, 2.00, 0.00>, <0.00, 0.00, 0.00, 0.00, 0.00, 0.00>, <0.00,

0.00, 0.11, 1.00, 1.00, 0.00>, <0.00, 0.00, 0.11, 1.00, 1.00, 0.00>, <0.00, 0.00, 0.22, 2.00, 2.00, 0.00>, <0.11, 0.00, 0.11, 1.00, 1.00, 0.00>}.

To see how to distinguish groups of inaccurate work process examples, Figure 3.2 demonstrates the provenance charts of a few centroids with istatistical feature results from “Temporal Representation for Mining Scientific Data Provenance” by P. Chen, and B. Plale, and M. Aktas. We purposely pick provenance charts from the LEAD North American Mesoscale (NAM) forecast workflow in light of the fact that it best delineates failures in provenance catch. Things being what they are the NAM provenance diagram with 10 subsets is a finished chart. While difficult to perceive, this is prove by a relic (circle) at base of diagram. The NAM provenance diagrams with under 10 subsets parcel the chart, all variants of which are inadequate and brought about by disappointments or dropped notifications. The NAM provenance chart with 2 subsets comprises of a few units of a complete provenance diagram, which is likely the aftereffect of disappointments. OPM XML contents of Figure 3.2 is attached on appendix as Figure 9.3.

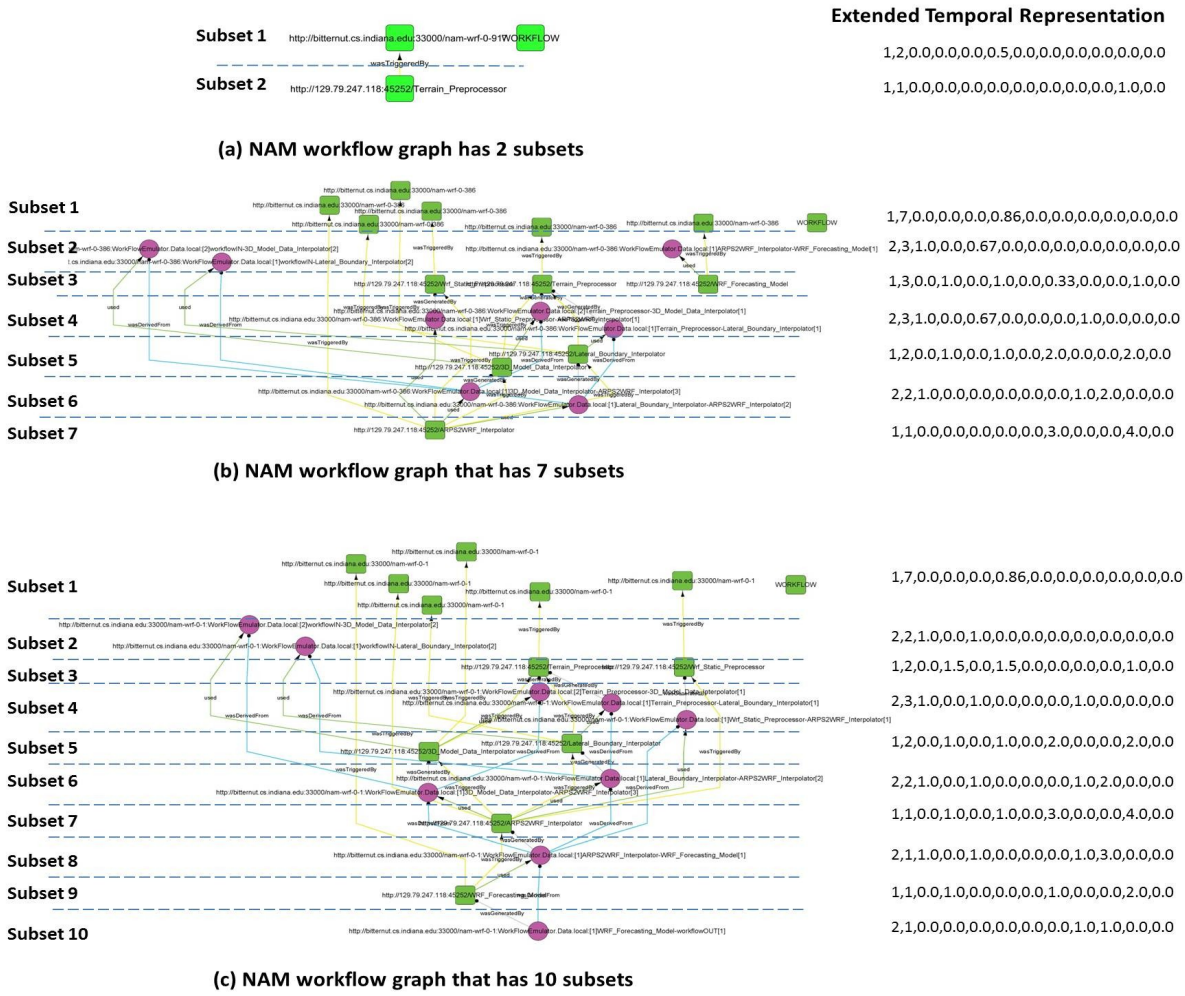

Figure 3.1 Provenance graphs of several centroids. Square nodes represent processes, and circles represent artifacts [6]

1.7. 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 3.1(a) and Figure 3.1(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 is shown in Table 3.3. 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 3.1(a) and Figure 3.1(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 3.1 Euclidean Distance For Differing Feature Sets Amongst Different Representations

	Figures 3.1(a) – 3.1(b)
Distance in Simple Structural Feature Set (from [5],[6])	0.74
Distance in NO-Metric Based Structural Feature Set	0.54
Distance in SNA-Metric Based Structural Feature Set	**0.44**

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.

CHAPTER IV

EXPERIMENTAL DATASETS

The trial assessment is done utilizing a substantial 10GB semi-engineered database and a genuine provenance dataset. The 10GB provenance database, produced from genuine work processes utilizing the WORKEM emulator[13], has known defect designs [7]. The database is populated with the provenance of roughly 48,000 workow execution examples, demonstrated on six real workflows: NCFS (ocean), LEAD North American Mesocale (LEAD NAM) forecast (weather), Animation (CS), SCOOP ADCIRC (coastline), Gene2Life (bio) and MotifNetwork (bio). A portion of the workflows are little, having a couple of nodes and edges, while others like the Motif workflow have a couple of hundred nodes and edges. Each workflow type has around 2000 occurrences for each defectnode, with defectnodes including arbitrary dropped messages (an assignment finishes yet the warning is not effectively transmitted) and workflows that come up short. I utilize the Karma provenance instrument [9] to send out the 10GB provenance dataset into OPM agreeable XML strings. This outcomes in a 2.1GB record with 47912 lines, each an OPM graph. This outcomes transferred to node and edge format for using this on calculation process of metrics.

Since we need to apply the association rules mining onto the 10GB provenance dataset to check whether it can find decides that reflect significant varieties, there is an issue that must be managed previously. Regardless of failed workows and dropped messages demonstrated into the semi-synthetic database, there are couple of huge basic varieties amongst the workflow examples. Along these lines, I physically present two variations of NAM weatherforecast workflow.

The real dataset that I try different things with is a NASA AMSR-E provenance document dataset [14]. The University of Alabama in Huntsville forms information from the NASA AMSR-E instrument, and the Karma venture instrumented the ingest handling framework and caught provenance for 3,890 keeps running for period Sept 2 -

Oct 4 2011. There are six types of day by day workflows, one five-day workflow, one week by week workflow, and three month to month workows in that dataset. The littlest provenance diagram is the Daily Rain grapgh that when spoke to as a XML record is 11KB, and the biggest is the Monthly Rain diagram that is 9MB in size. The aggregate size of AMSR-E dataset put away as XML files is 83.0MB.

CHAPTER 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 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 systemis 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 followsthe 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 results are written on this section, are mostly about NAM workflow. The results about the rest groups are located on appendix section.

3.1. Evaluation Metrics

Purity and the Normalized Mutual Information (NMI) are utilized to look at the execution of different grouping procedures. Purity accept that all specimens of a bunch are anticipated to be individuals from the genuine prevailing class for that group. Be that as it may, high purity can be accomplished when the quantity of groups is substantial. For instance, purity would be one when every diagram has a place with its own particular bunch. We can't utilize immaculateness to exchange of the nature of the grouping against the quantity of bunches, so we additionally utilize NMI bunching assessment metric. NMI is defined as the shared data between the bunch assignments and a prior naming of the dataset standardized by the number juggling mean of the most extreme conceivable entropies of the exact marginals. k-implies execution is assessed utilizing Within-Cluster Sum of Squares (WCSS), notwithstanding purity and Normalized Mutual Information(NMI) measurements.

With respect to the classification, I utilize NaiveBayes as the assessing classifier for Clas-sifierSubsetEval, and tested different classification strategies on the components chose by the CfsSubsetEval, including the NaiveBayes and the Random forest. The precision is assessed utilizing the 10-fold cross-validation strategy.

To assess the execution of affiliation principles mining, we check if there are coming about standards that can uncover the known variations.

3.2. Workflow Based Tests (10GB)

5.2.1 Unsupervised clustering

The grouping of the NAM provenance dataset (based on the temporal length defined by Chen et al.) is shown in Figure 5.1. 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.

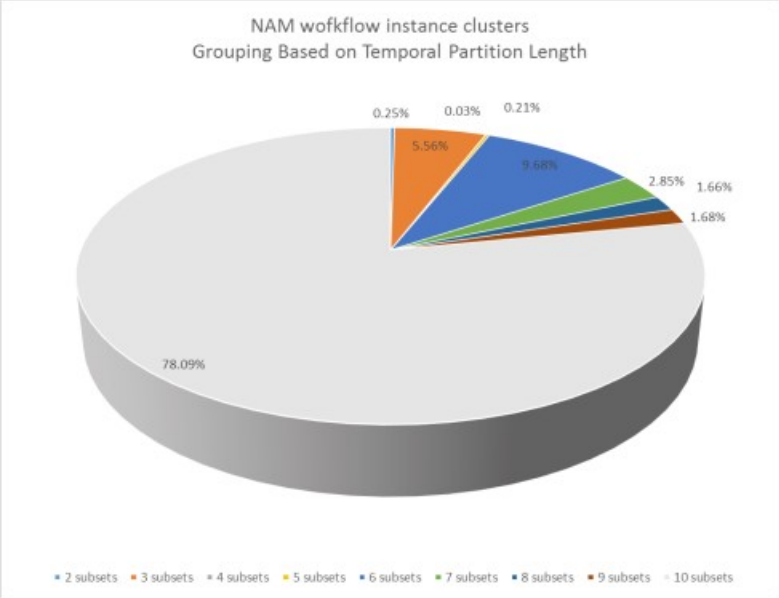

Figure 5.1 NAM workflow clusters - Grouping Based on Temporal Representation

To test the clustering on the reduced representations, the grouping results (shown in Figure 5.1) 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 5.1 Nam Workflow Clustering Results for K=9 and for Varying Reduced

	Purity	NMI	WCCS
Network Overview	0.80	0.50	171.62
SNA	**0.92**	0.51	**43.17**
TR+Istatistical Feature	0.91	0.51	664.23
TR+SNA	**0.93**	**0.55**	561.37

Table 5.1 gives the summary of the quality of clustering results when we chose 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 5.1 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.

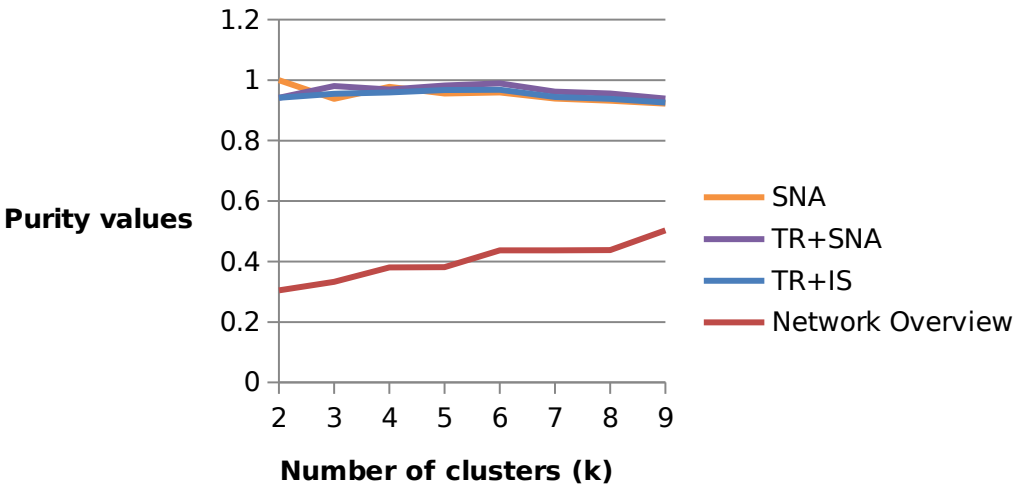

Figure 5.1 NAM - Purity results

Figure 5.2 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.

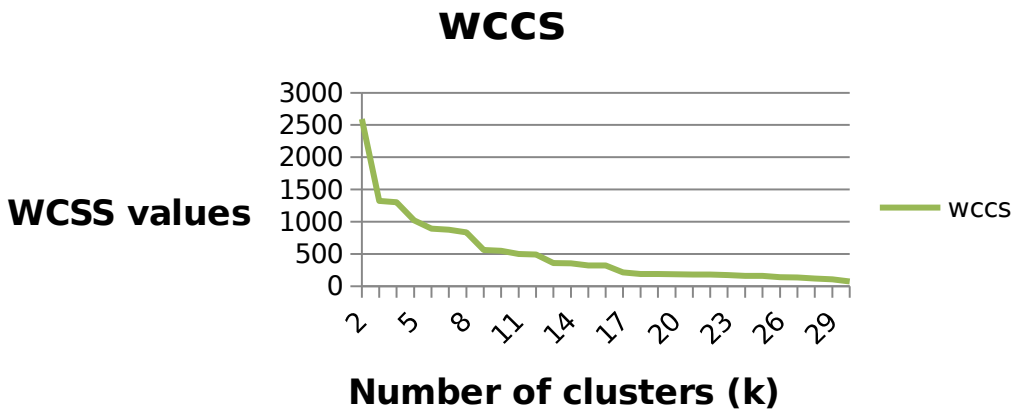

Figure 5.1 NAM – SNA k = [2,30] WCSS

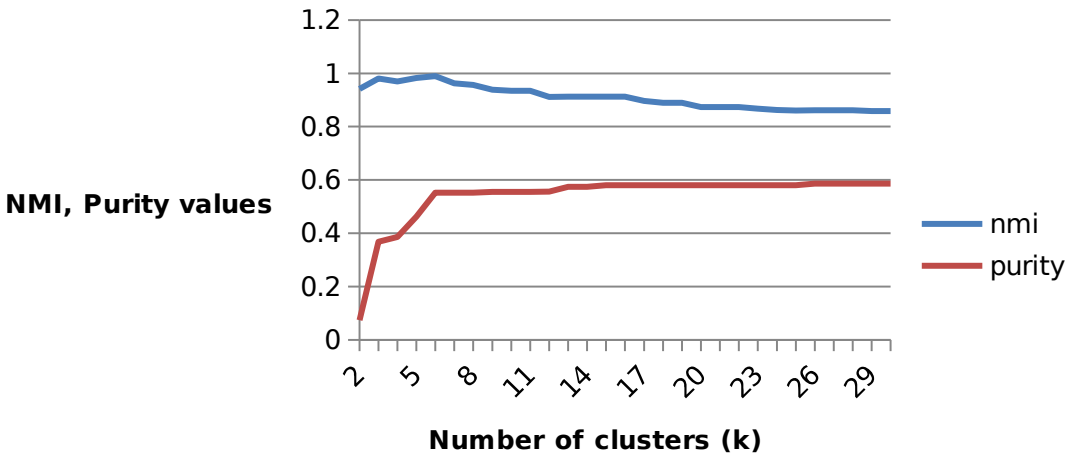

Figure 5.1 NAM - SNA k=[2,30] NMI Purity

Figure 5.3 and Figure 5.4 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.

5.2.2 Classification of Provenance Workflow

Recall that we select seven features out of 120 features by using the feature selection algorithm ClassifferSubsetEval, and select nine features by using CfsSubsetEval. The NavieBayes model trained from the seven features has a high correctness of 99.83% in its10-fold cross validation (Table 5.2). On the nine features, we test different classification methods andmany of them, such as NaiveBayes and Random forests can achieve a correct ratio more than 97.5-99.8% (Table 5.2). This indicates the feature selection in our temporal representations is independent of learning algorithm, and the resulting classiffer also has good performance.

Table 5.1 Classification model trained on features selected by ClassifierSubsetEval and CfsSubsetEval

Feature Selector	**Classifier**	**Selected Features**	**Results**
ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegreePrestige	7807/8000 97.58 %
CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	7807/8000 97.58 %
	weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	7987/8000 99.83 %

Other workflow results are shown on Table 9.1 on appendix section in detail.

Table 5.3 shows the confusion matrix of nam workflow.

Table 5.1 Confusion Matrix of Nam Workflow

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	20	0	0	0	0	0	0	0	0
3	0	445	0	0	0	0	0	0	0
4	0	0	0	2	0	0	0	0	0
5	0	0	0	11	6	0	0	0	0
6	0	0	0	0	774	0	0	0	0
7	0	0	0	0	1	227	0	0	0
8	0	0	0	0	0	1	110	22	0
9	0	0	0	0	0	0	7	127	0
10	0	0	0	0	0	0	183	27	6037

5.2.3 Association Rule Mining

Table 5.4 shows the Scheme of the Weka method i applied and the resulting association rules that can reflect the variants we introduced. 10 rules are found by Apriori any of them can be used.

Rest of other workflow results are listed on Table 9.3 in detail.

Table 5.1 Sampling of association rules mined by Apriori

Workflow	**Weka Scheme**	**Sample of association rules found**

Nam	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageProximityPrestige='(-inf-0.035855]' ==> averageDegreePrestige='All' 2. class=10 ==> averageDegreePrestige='All' 3. averageClosenessCentrality='(6.987879-7.77803]' ==> averageDegreePrestige='All' 4. averageProximityPrestige='(-inf-0.035855]' class=10 ==> averageDegreePrestige='All' 5. averageClosenessCentrality='(6.987879-7.77803]' averageProximityPrestige='(-inf-0.035855]' ==> averageDegreePrestige='All' 6. averageClosenessCentrality='(6.987879-7.77803]' class=10 ==> averageDegreePrestige='All' 7. averageEccentricity='(2.219048-inf)' ==> averageDegreePrestige='All' 8. averageEccentricity='(2.219048-inf)' class=10 ==> averageDegreePrestige='All' 9. averageClosenessCentrality='(6.987879-7.77803]' averageProximityPrestige='(-inf-0.035855]' class=10 ==> averageDegreePrestige='All' 10. averageDegreeCentrality='(0.09-inf)' ==> averageDegreePrestige='All'

3.3. Subset Based Tests (10GB)

5.3.1 Unsupervised clustering

To test the clustering on the reduced representations based on subset, the grouping results 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 for subset based representation. 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 5.1 10 Subset Clustering Results for K=6 and for Varying Reduced Provenance Representations

	Purity	**NMI**	**WCCS**
Network Overview	0.86	0.93	**118.84**
SNA	**0.97**	0.91	**145.24**
TR+Istatistical Feature	0.95	0.93	1453.88
TR+SNA	0.95	**0.94**	1443.44

Table 5.5 gives the summary of the quality of clustering results when we chose k = 6. 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 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.

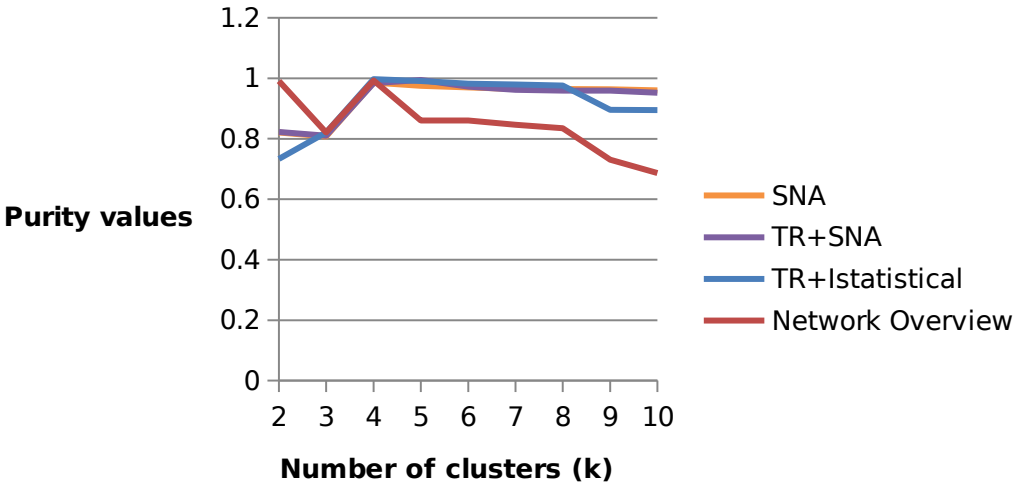

Figure 5.1 Workflows has 10 subsets - Purity results

Figure 5.5 shows that SNA-metrics based reduced provenance representation on subsets 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.

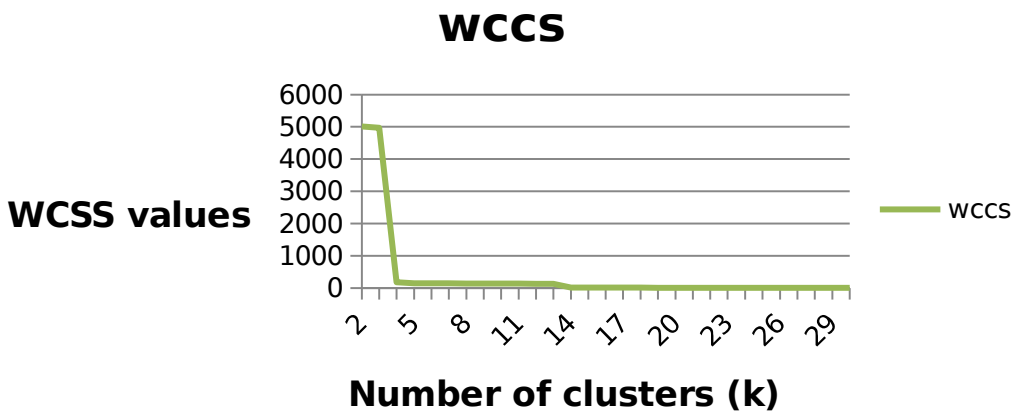

Figure 5.1 Workflows has 10 subsets – SNA k = [2,30] WCSS

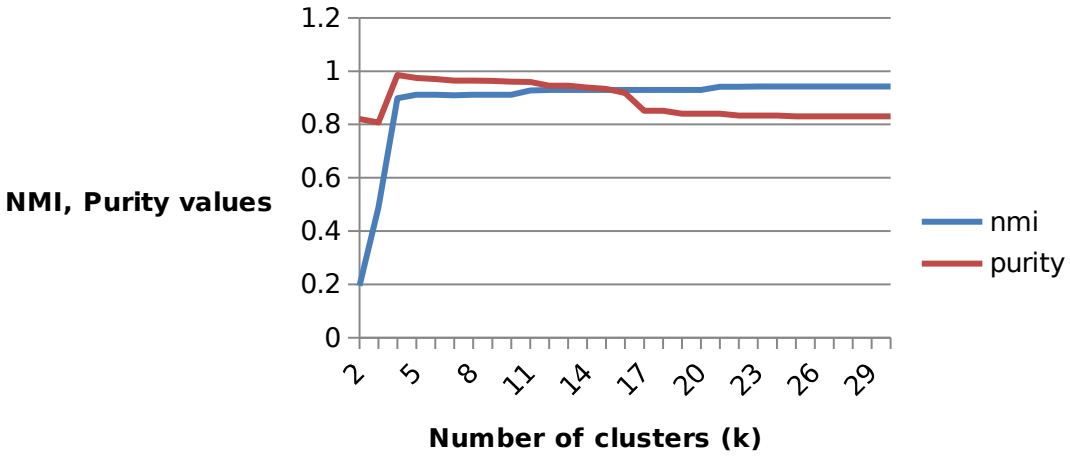

Figure 5.1 Workflows has 10 subsets - SNA k=[2,30] NMI Purity

Figure 5.6 and Figure 5.7 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.

5.3.2 Classification of Provenance Workflow

Recall that we select seven features out of 120 features by using the feature selection algorithm ClassifferSubsetEval, and select seven features by using CfsSubsetEval. The NavieBayes model trained from the seven features has a high correctness of 98.39% in its 10-fold cross validation (Table 5.6). On the seven features, I test different classification methods and many of them, such as NaiveBayes and Random forests can achieve a correct ratio more than 97.1-99.8% (Table 5.6). This indicates the feature selection in new representations are independent of learning algorithm, and the resulting classiffer also has good performance.

Table 5.1 Classification model trained on features selected by ClassifierSubsetEval and CfsSubsetEval

Feature Selector	**Classifier**	**Selected Features**	**Results**
ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageDegreePrestige	7871/8000 98.39 %
CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	7787/8000 97.33 %
	weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	7772/8000 97.15 %

Table 5.7 illustrates workflows have 10 subsets confustion matrix distrubition.

Table 5.1 Confusion Matrix of Workflows have 10 subset

	animation	**gene2life**	**nam**	**motif**	**ncfs**	**scoop**
animation	0	0	0	0	0	0
gene2life	0	5882	0	1	0	0
nam	0	0	6247	0	0	0
motif	0	2	1	3918	0	0
ncfs	0	0	0	0	5838	0
scoop	0	0	0	0	0	0

5.3.3 Association Rule Mining

Table 5.8 shows the Scheme of the Weka method we applied and the resulting association rules that can reflect the variants we introduced. 10 rules are found by Apriori any of them can be used. Rule 1 says that if the eccentricity in subset 10 (which are the data inputs for the last processing step) is between 2.2636 and 3.0445 (including 3.0445), then closeness centrality in subset 10 (which are the final dataoutputs) will be smaller than 32.137218. Rule 2 says that if the eccentricity in subset 10 is between 2.2636 and 3.0445 (including 3.0445) and degree prestige is 0, then closeness centrality in subset 10 will be smallerthan 32.137218. Because the SNA-metric can only be integer, all rules mean oneintermediate data input for the final processing step will lead to one final data output, while more data inputs lead to more final data outputs, which reveals exactly the firstvariant we introduced. This single example shows that SNA-metric based provenance representation with selected number of features supports the apriori algorithm well: the association rules canshow variants during execution.

Table 5.1 Sampling of association rules mined by Apriori

Workflow	Weka Scheme	Sample of association rules found
10 subset	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	**1.** averageEccentricity='(2.263649-3.04454]' 11832 ==> averageClosenessCentrality='(-inf-32.137218]' 11832 **2.** averageEccentricity='(2.263649-3.04454]' averageDegreePrestige=0 11831 ==> averageClosenessCentrality='(-inf-32.137218]' 11831 **3.** averageProximityPrestige='(0.027379-0.031633]' 11316 ==> averageClosenessCentrality='(-inf-32.137218]' 11316 **4.** averageProximityPrestige='(0.027379-0.031633]' averageDegreePrestige=0 11315 ==> averageClosenessCentrality='(-inf-32.137218]' 11315 **5.** averageClosenessCentrality='(-inf-32.137218]' 17971 ==> averageDegreePrestige=0 17970 **6.** averageEccentricity='(2.263649-3.04454]' 11832 ==> averageDegreePrestige=0 11831 **7.** averageClosenessCentrality='(-inf-32.137218]' averageEccentricity='(2.263649-3.04454]' 11832 ==> averageDegreePrestige=0 11831 **8.** averageEccentricity='(2.263649-3.04454]' 11832 ==> averageClosenessCentrality='(-inf-32.137218]' averageDegreePrestige=0 11831 **9.** averageProximityPrestige='(0.027379-0.031633]' 11316 ==> averageDegreePrestige=0 11315 **10.** averageClosenessCentrality='(-inf-32.137218]' averageProximityPrestige='(0.027379-0.031633]' 11316 ==> averageDegreePrestige=0 11315

3.4. Tests on real-life dataset (AMSR-E)

5.4.1 Unsupervised clustering

On the AMSR-E dataset, we likewise utilize the k-means (SimpleKMeans in Weka) and repeat the same clustering tests. The clustering execution on pre-gathered

representations is additionally great. We isolate representations into a 4-subset assemble, a 5-subset group and a 6-subset group. It is no more the case that 4-subset gathering is the generation of failures, and truth be told the 4-subset group has the dominant part of all cases (91%). That implies the AMSR-E dataset has less failures than the 10GB semi-syntehetic dataset. However, we are still ready to find variants many AMSR-E work process examples have a place with the same work process sort however are isolated into different groups.

Figure 5.8 and Figure 5.9 show the results of experiment on real AMSR-E dataset.

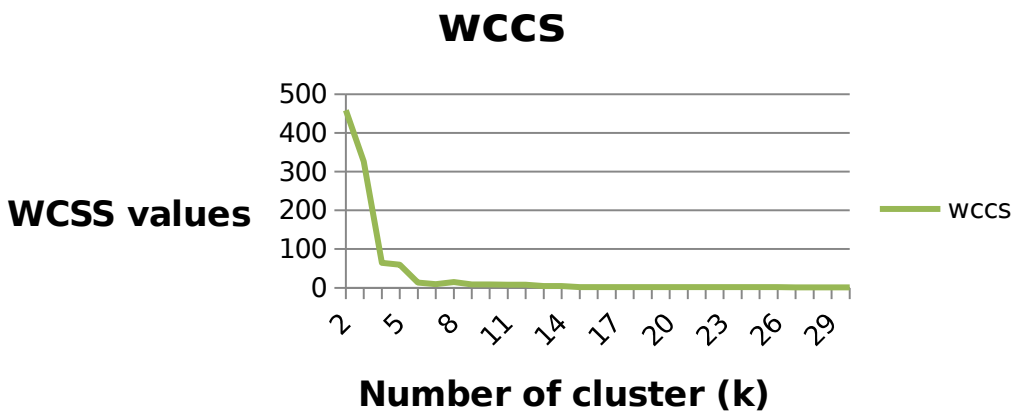

Figure 5.1 AMSR-E – SNA k = [2,30] WCSS

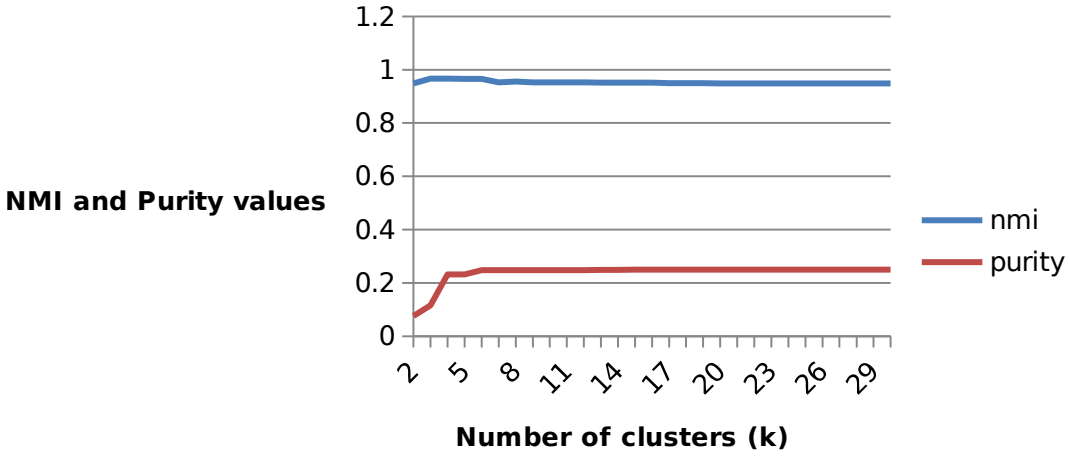

Figure 5.1 AMSR-E - SNA k=[2,30] NMI Purity

5.4.2 Classification of Provenance Workflow

For classification on the AMSR-E dataset, recall that we also use the feature selection method CfsSubsetEval, which selects two attributes from the initial feature set average closeness centralityand average eccentricity. Using these two attributes to train a Naive Bayes classiffer can achieve a high correctness of 99.93% (Table 5.9), which indicates that they are particular important in determining different types of workflow.

Table 5.1 AMSR-E classification model trained on features selected by CfsSubsetEval and ClassifierSubsetEval

Feature Selector	**Classifier**	**Selected Features**	**Results**
ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegreePrestige	2850/2891 98.58 %
CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	2850/2891 98.58 %
	weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	2889/2891 99.93 %

5.4.3 Association Rule Mining

Table 5.10 shows the Scheme of the Weka method we applied and the resulting association rules that can reflect the variants we introduced. 10 rules are found by Apriori any of them can be used.

Table 5.1 Sampling of association rules mined by Apriori

Workflow	Weka Scheme	Sample of association rules found
AMSR-E	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	**1.** averageDegreePrestige='All' ==> averageBetweennessCentrality='All' **2.** averageBetweennessCentrality='All' ==> averageDegreePrestige='All' **3.** averageProximityPrestige='(0.138333-inf)' ==> averageBetweennessCentrality='All' **4.** averageProximityPrestige='(0.138333-inf)' ==> averageDegreePrestige='All' **5.** averageProximityPrestige='(0.138333-inf)' averageDegreePrestige='All' ==> averageBetweennessCentrality='All' **6.** averageBetweennessCentrality='All' averageProximityPrestige='(0.138333-inf)' ==> averageDegreePrestige='All' **7.** averageProximityPrestige='(0.138333-inf)' ==> averageBetweennessCentrality='All' averageDegreePrestige='All' **8.** class=4 2628 ==> averageBetweennessCentrality='All' **9.** class=4 ==> averageDegreePrestige='All' **10.** averageDegreePrestige='All' class=4 ==> averageBetweennessCentrality='All'

CHAPTER VI

DISCUSSION

Experiments on semi-synthetic and real life provenance datasets show that the new representation methodolgy can detect failed work predict the type of new workflow instance or new workflow subset which is belonged to workflow, and can find descriptive knowledge about a cluster.

Despite the fact that there is data misfortune when selecting highlights from the factual element space, reduced representation is appropriate to information mining tasks, for example, unsupervised clustering, classification and associationmining rules, which are impractical on straightforward OPM diagrams. The execution assessment indicates great purity and NMI in unsupervised clustering and high correctness proportion in managed classification. The adaptability study shows that information mining on proposed worldly representations are versatile in view of the size and dimensionality of dataset are incredibly diminished in our reduction procedure, and the scalability of representing to process can be enhanced utilizing multi processor processing.

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.

On first methodology (Network Overview) results show that this new approach is not suitable for provenance representation. However, the SNA strategy demonstrates that the SNA results are significantly more effective than others strategies.

CHAPTER VII

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. I also plan to extend this work to combine both Network Overview metric and Social Network Analysis metric representations in one vector. The researcher plan to adapt state-of-the-art approaches for dimensionality reduction and high-contrast feature selection in future work.

REFERENCES

[1] C. M. Antunes, and A. L. Oliveira, (2011). "Temporal data mining: An overview" KDD Workshop on Temporal Data Mining, 21-24 August 2011, San Diego.

[2] M. Aktas, B. Plale, D. Leake and N. K. Mukhi, (2013). "Unmanaged Workflows: Their Provenance and Use", Data Provenance and Data Management in eScience, Studies in Computational Intelligence series, 426:59-81.

[3] S. Bechhofer, D. D. Roure, M. Gamble, C. Goble and L. Buchan, (2010). "Research objects: Towards exchange and reuse of digital knowledge", The Future of the Web for Collaborative Science, 26 April 2010, NC, USA.

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

[5] P. Chen, and B. Plale, and M. Aktas, (2012). "Temporal representation for scientific data provenance", The 8th Int. IEEE Conference on e-Science, 21 Oct 2012, Chicago.

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

[7] Y. Cheah, B. Plale, J. Kendall-Morwick, D. Leake and L. Ramakrishnan, (2011). "A Noisy 10GB Provenance Database", 2nd Int'l Workshop on Traceability and Compliance of Semi-Structured Processes (TC4SP), 10 August 2011, Clermont-Ferrand, France.

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

[9] S. B. Davidson and J. Freire, (2008). "Provenance and scientific workflows: challenges and opportunities" SIGMOD Conference, 9-12 June 2008, Vancouver, BC, Canada.

[10] S. B. Davidson and J. Freire, (2008) "Provenance and scientific workflows: challenges and opportunities" SIGMOD Conference, 9-12 June 2008, Vancouver, BC, Canada.

[11] E. Santos, L. Lins, J. P. Ahrens, J. Freire, and C. T. Silva, (2008) "A first study on clustering collections of workflow graphs" IPAW, 17-18 June 2008, Salt Lake City, UT, USA.

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

[13] Ramakrishnan L, Gannon D, Plale B., (2010). “Workem: Representing and emulating distributed scientificwork flow execution state”, 10th IEEE/ACM International Conferenceon Cluster, Cloud and Grid Computing. IEEE Computer Society,17-20 May 2010, Melbourne, Australia.

[14] Luo Y, Plale B, Jensen S, Cheah YW, Conover H. (2011), “Provenance of amsr-e data from the national snow and ice data center (nsidc)”, http://dx.doi.org/10.5967/M0F47M2D , 14 July 2016.

[1]
[2]

PS

```
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
```

<v1:opmGraph xmlns:v1="http://openprovenance.org/model/v1.1.a"><v1:accounts><v1:account id="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:accounts><v1:processes><v1:process id="Process_34359"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:process><v1:process id="Process_264016"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:process><v1:process id="Process_2076803"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:process></v1:processes><v1:artifacts><v1:artifact id="File_1815326"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815301"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815338"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815290"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_358631"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815318"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815307"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815296"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815334"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815322"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815330"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact><v1:artifact id="File_1815312"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:artifact></v1:artifacts><v1:agents><v1:agent id="Agent_001"><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:agent></v1:agents><v1:causalDependencies><v1:wasTriggeredBy><v1:effect ref="Process_34359"/><v1:cause ref="Agent_001"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_264016"/><v1:cause ref="Process_34359"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2009-02-19T18:32:00.000-05:00" noLaterThan="2009-02-19T18:32:00.000-05:00"/></v1:wasTriggeredBy><v1:used><v1:effect ref="Process_264016"/><v1:role value="Input"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:28:53.000-04:00"/></v1:used><v1:wasDerivedFrom><v1:effect ref="File_1815290"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815296"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815301"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-

03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815307"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815312"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815318"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815322"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815326"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815330"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815334"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1815338"/><v1:cause ref="File_358631"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:53.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasDerivedFrom><v1:wasTriggeredBy><v1:effect ref="Process_264016"/><v1:cause ref="Process_2076803"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:28:52.000-04:00" noLaterThan="2010-03-15T03:28:52.000-04:00"/></v1:wasTriggeredBy><v1:wasGeneratedBy><v1:effect ref="File_1815290"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815296"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815301"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815307"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815312"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815318"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815322"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815326"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815330"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815334"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1815338"/><v1:role value="Output"/><v1:cause ref="Process_264016"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/animation-2-817"/><v1:time noEarlierThan="2010-03-15T03:31:56.000-04:00" noLaterThan="2010-03-15T03:31:56.000-04:00"/></v1:wasGeneratedBy></v1:causalDependencies></v1:opmGraph>

2- subset - Nam different subset OPM files as xml

<v1:opmGraph xmlns:v1="http://openprovenance.org/model/v1.1.a"><v1:accounts><v1:account

id="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:accounts><v1:processes><v1:process id="Process_320"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_49045"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_49054"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_49177"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_49184"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1042550"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1042554"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1070261"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1070267"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1070272"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1070280"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1070287"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process><v1:process id="Process_1070295"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:process></v1:processes><v1:artifacts><v1:artifact id="File_2252"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2524"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2263"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2258"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2247"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2508"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2530"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact><v1:artifact id="File_2516"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/></v1:artifact></v1:artifacts><v1:agents/><v1:causalDependencies><v1:used><v1:effect ref="Process_49045"/><v1:role value="Input"/><v1:cause ref="File_2247"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:48:06.000-05:00" noLaterThan="2010-02-19T08:49:44.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49045"/><v1:role value="Input"/><v1:cause ref="File_2252"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:40:42.000-05:00" noLaterThan="2010-02-19T08:49:47.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49045"/><v1:role value="Input"/><v1:cause ref="File_2258"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:43:31.000-05:00" noLaterThan="2010-02-19T08:50:34.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49054"/><v1:role value="Input"/><v1:cause ref="File_2263"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:53:18.000-05:00" noLaterThan="2010-02-19T08:53:18.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49177"/><v1:role value="Input"/><v1:cause ref="File_2508"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:41:13.000-05:00" noLaterThan="2010-02-19T08:41:13.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49177"/><v1:role value="Input"/><v1:cause ref="File_2516"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:55.000-05:00" noLaterThan="2010-02-19T08:41:13.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49184"/><v1:role value="Input"/><v1:cause ref="File_2524"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:44:02.000-05:00" noLaterThan="2010-02-19T08:44:02.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49184"/><v1:role value="Input"/><v1:cause ref="File_2530"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:55.000-05:00" noLaterThan="2010-02-19T08:44:02.000-05:00"/></v1:used><v1:wasDerivedFrom><v1:effect ref="File_2258"/><v1:cause ref="File_2508"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:41:13.000-05:00" noLaterThan="2010-02-19T08:43:31.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2258"/><v1:cause ref="File_2516"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:41:13.000-05:00" noLaterThan="2010-02-19T08:43:31.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2247"/><v1:cause ref="File_2524"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:44:02.000-05:00" noLaterThan="2010-02-19T08:48:06.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2247"/><v1:cause ref="File_2530"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:44:02.000-05:00" noLaterThan="2010-02-19T08:48:06.000-05:00"/></v1:wasDerivedFrom><v1:wasTriggeredBy><v1:effect ref="Process_49184"/><v1:cause ref="Process_1042550"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:55.000-05:00" noLaterThan="2010-02-19T08:44:02.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49177"/><v1:cause ref="Process_1042550"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:55.000-05:00" noLaterThan="2010-02-19T08:41:13.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49045"/><v1:cause ref="Process_1042554"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:40:42.000-05:00" noLaterThan="2010-02-19T08:49:47.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49045"/><v1:cause ref="Process_49177"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-

02-19T08:43:31.000-05:00" noLaterThan="2010-02-19T08:50:34.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49045"/><v1:cause ref="Process_49184"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:48:06.000-05:00" noLaterThan="2010-02-19T08:49:44.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49045"/><v1:cause ref="Process_1070287"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:52:36.000-05:00" noLaterThan="2010-02-19T08:52:36.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49054"/><v1:cause ref="Process_1070295"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T10:17:25.000-05:00" noLaterThan="2010-02-19T10:17:25.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49177"/><v1:cause ref="Process_1070272"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:43:32.000-05:00" noLaterThan="2010-02-19T08:43:32.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49184"/><v1:cause ref="Process_1070280"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:48:07.000-05:00" noLaterThan="2010-02-19T08:48:07.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_1042550"/><v1:cause ref="Process_1070261"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:56.000-05:00" noLaterThan="2010-02-19T08:34:56.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_1042554"/><v1:cause ref="Process_1070267"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:40:43.000-05:00" noLaterThan="2010-02-19T08:40:43.000-05:00"/></v1:wasTriggeredBy><v1:wasGeneratedBy><v1:effect ref="File_2258"/><v1:role value="Output"/><v1:cause ref="Process_49177"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:43:31.000-05:00" noLaterThan="2010-02-19T08:50:34.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2247"/><v1:role value="Output"/><v1:cause ref="Process_49184"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:48:06.000-05:00" noLaterThan="2010-02-19T08:49:44.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2530"/><v1:role value="Output"/><v1:cause ref="Process_1042550"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:55.000-05:00" noLaterThan="2010-02-19T08:44:02.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2516"/><v1:role value="Output"/><v1:cause ref="Process_1042550"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:34:55.000-05:00" noLaterThan="2010-02-19T08:41:13.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2252"/><v1:role value="Output"/><v1:cause ref="Process_1042554"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-386"/><v1:time noEarlierThan="2010-02-19T08:40:42.000-05:00" noLaterThan="2010-02-19T08:49:47.000-05:00"/></v1:wasGeneratedBy></v1:causalDependencies></v1:opmGraph>

7-subset - Nam different subset OPM files as xml

<v1:opmGraph xmlns:v1="http://openprovenance.org/model/v1.1.a"><v1:accounts><v1:account id="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:accounts><v1:processes><v1:process id="Process_321"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_49148"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_49156"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_49164"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_49174"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1042536"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1042538"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1070216"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1070224"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1070230"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1070236"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1070245"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process><v1:process id="Process_1070253"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:process></v1:processes><v1:artifacts><v1:artifact id="File_2497"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2470"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2485"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2453"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2477"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2461"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2491"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_1723586"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/></v1:artifact><v1:artifact id="File_2503"><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-

387"/></v1:artifact></v1:artifacts><v1:agents/><v1:causalDependencies><v1:used><v1:effect ref="Process_49148"/><v1:role value="Input"/><v1:cause ref="File_2453"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:44:22.000-05:00" noLaterThan="2010-02-19T08:44:22.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49148"/><v1:role value="Input"/><v1:cause ref="File_2461"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:51.000-05:00" noLaterThan="2010-02-19T08:44:22.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49156"/><v1:role value="Input"/><v1:cause ref="File_2470"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:57.000-05:00" noLaterThan="2010-02-19T08:48:57.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49156"/><v1:role value="Input"/><v1:cause ref="File_2477"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:51.000-05:00" noLaterThan="2010-02-19T08:48:57.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49164"/><v1:role value="Input"/><v1:cause ref="File_2485"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:26.000-05:00" noLaterThan="2010-02-19T08:52:53.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49164"/><v1:role value="Input"/><v1:cause ref="File_2491"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:41.000-05:00" noLaterThan="2010-02-19T08:52:56.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49164"/><v1:role value="Input"/><v1:cause ref="File_2497"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:51:15.000-05:00" noLaterThan="2010-02-19T08:53:43.000-05:00"/></v1:used><v1:used><v1:effect ref="Process_49174"/><v1:role value="Input"/><v1:cause ref="File_2503"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:55:44.000-05:00" noLaterThan="2010-02-19T08:56:27.000-05:00"/></v1:used><v1:wasDerivedFrom><v1:effect ref="File_2485"/><v1:cause ref="File_2453"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:44:22.000-05:00" noLaterThan="2010-02-19T08:48:26.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2485"/><v1:cause ref="File_2461"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:44:22.000-05:00" noLaterThan="2010-02-19T08:48:26.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2497"/><v1:cause ref="File_2470"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:57.000-05:00" noLaterThan="2010-02-19T08:51:15.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2497"/><v1:cause ref="File_2477"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:57.000-05:00" noLaterThan="2010-02-19T08:51:15.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2503"/><v1:cause ref="File_2485"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:52:53.000-05:00" noLaterThan="2010-02-19T08:55:44.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2503"/><v1:cause ref="File_2491"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:52:56.000-05:00" noLaterThan="2010-02-19T08:55:44.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_2503"/><v1:cause ref="File_2497"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:53:43.000-05:00" noLaterThan="2010-02-19T08:55:44.000-05:00"/></v1:wasDerivedFrom><v1:wasDerivedFrom><v1:effect ref="File_1723586"/><v1:cause ref="File_2503"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:56:27.000-05:00" noLaterThan="2010-02-19T10:20:33.000-05:00"/></v1:wasDerivedFrom><v1:wasTriggeredBy><v1:effect ref="Process_49164"/><v1:cause ref="Process_1042536"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:41.000-05:00" noLaterThan="2010-02-19T08:52:56.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49148"/><v1:cause ref="Process_1042538"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:51.000-05:00" noLaterThan="2010-02-19T08:44:22.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49156"/><v1:cause ref="Process_1042538"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:51.000-05:00" noLaterThan="2010-02-19T08:48:57.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49164"/><v1:cause ref="Process_49148"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:26.000-05:00" noLaterThan="2010-02-19T08:52:53.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49164"/><v1:cause ref="Process_49156"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:51:15.000-05:00" noLaterThan="2010-02-19T08:53:43.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49174"/><v1:cause ref="Process_49164"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:55:44.000-05:00" noLaterThan="2010-02-19T08:56:27.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49148"/><v1:cause ref="Process_1070230"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:27.000-05:00" noLaterThan="2010-02-19T08:48:27.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49156"/><v1:cause ref="Process_1070236"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:51:16.000-05:00" noLaterThan="2010-02-19T08:51:16.000-

05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49164"/><v1:cause ref="Process_1070245"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:55:45.000-05:00" noLaterThan="2010-02-19T08:55:45.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_49174"/><v1:cause ref="Process_1070253"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T10:20:34.000-05:00" noLaterThan="2010-02-19T10:20:34.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_1042536"/><v1:cause ref="Process_1070216"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:42.000-05:00" noLaterThan="2010-02-19T08:43:42.000-05:00"/></v1:wasTriggeredBy><v1:wasTriggeredBy><v1:effect ref="Process_1042538"/><v1:cause ref="Process_1070224"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:52.000-05:00" noLaterThan="2010-02-19T08:43:52.000-05:00"/></v1:wasTriggeredBy><v1:wasGeneratedBy><v1:effect ref="File_2485"/><v1:role value="Output"/><v1:cause ref="Process_49148"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:48:26.000-05:00" noLaterThan="2010-02-19T08:52:53.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2497"/><v1:role value="Output"/><v1:cause ref="Process_49156"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:51:15.000-05:00" noLaterThan="2010-02-19T08:53:43.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2503"/><v1:role value="Output"/><v1:cause ref="Process_49164"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:55:44.000-05:00" noLaterThan="2010-02-19T08:56:27.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_1723586"/><v1:role value="Output"/><v1:cause ref="Process_49174"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T10:20:33.000-05:00" noLaterThan="2010-02-19T10:20:33.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2491"/><v1:role value="Output"/><v1:cause ref="Process_1042536"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:41.000-05:00" noLaterThan="2010-02-19T08:52:56.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2461"/><v1:role value="Output"/><v1:cause ref="Process_1042538"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:51.000-05:00" noLaterThan="2010-02-19T08:44:22.000-05:00"/></v1:wasGeneratedBy><v1:wasGeneratedBy><v1:effect ref="File_2477"/><v1:role value="Output"/><v1:cause ref="Process_1042538"/><v1:account ref="http://bitternut.cs.indiana.edu:33000/nam-wrf-0-387"/><v1:time noEarlierThan="2010-02-19T08:43:51.000-05:00" noLaterThan="2010-02-19T08:48:57.000-05:00"/></v1:wasGeneratedBy></v1:causalDependencies></v1:opmGraph>

10-subset - Nam different subset OPM files as xml

Figure 9.1 Nam different subset OPM files as xml

Table 9.1 Classification results of provenance workflow are selected by SNA

	Feature selection algorithm	**Weka Scheme**	**Selected Attributes**	**Result of 10-fold cross-validation**
Animation	ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	7909/7950 99.48 %
	CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality	7900/7950 99.37 %

			averageEccentricity averageDegree Prestige	
		weka.classifiers.trees.RandomForest	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegree Prestige	7940/7950 99.87 %

Table 9.1 Classification results of provenance workflow are selected by SNA (cont'd)

Gene2Life	ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageEccentricity averageProximityPrestige	7731/8000 96.63 %
	CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	7670/8000 95.87 %
		weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	7971/8000 99.63 %
Nam	ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegree Prestige	7807/8000 97.58 %
	CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	7807/8000 97.58 %
		weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	7987/8000 99.83 %

Scoop	ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegree Prestige	7956/7999 99.46 %
	CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageEccentricity	7956/7999 99.46 %
		weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	7998/7999 99.98 %
Motif	ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegree Prestige	7327/7963 92.01 %
	CfsSubsetEval	weka.classifiers.bayes.NaiveBayes	averageClosenessCentrality averageProximityPrestige	7330/7963 92.05 %
		weka.classifiers.trees.RandomForest	averageClosenessCentrality averageProximityPrestige	7308/7963 91.77 %

Table 9.1 Classification results of provenance workflow are selected by SNA (cont'd)

Ncfs	ClassifierSubsetEval	weka.classifiers.bayes.NaiveBayes	averageBetweennessCentrality averageClosenessCentrality averageEccentricity averageDegree Prestige	7558/8001 94.46 %
	CfsSubsetEval	weka.classifiers.bayes.NaiveBa	averageClosenessCentrality	7558/8001

		yes	averageEccentricity	94.46 %
		weka.classifiers.trees.RandomForest	averageClosenessCentrality averageEccentricity	7940/8001 99.23 %

Table 9. 1 Comparision of clustering algorithms based on Purity and NMI value

	Purity	NMI
EM	0.77	0.54
DBScan	0.79	**0.85**
Kmeans	**0.90**	0.82

Table 9. 1 Sampling of association rules mined by Apriori for all workflow

	Weka Scheme	**Sample of association rules found**
Animation	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageProximityPrestige='(-inf-0.024749]' 7030 ==> averageDegreePrestige='All' 7030 conf:(1) 2. averageClosenessCentrality='(31.837963-35.770833]' 4817 ==> averageProximityPrestige='(-inf-0.024749]' 4817 conf:(1) 3. averageClosenessCentrality='(31.837963-35.770833]' 4817 ==> averageDegreePrestige='All' 4817 conf:(1) 4. averageClosenessCentrality='(31.837963-35.770833]' averageDegreePrestige='All' 4817 ==> averageProximityPrestige='(-inf-0.024749]' 4817 conf:(1) 5. averageClosenessCentrality='(31.837963-35.770833]' averageProximityPrestige='(-inf-0.024749]' 4817 ==> averageDegreePrestige='All' 4817 conf:(1) 6. averageClosenessCentrality='(31.837963-35.770833]' 4817 ==> averageProximityPrestige='(-inf-0.024749]' averageDegreePrestige='All' 4817 conf:(1) 7. class=8 4153 ==> averageProximityPrestige='(-inf-0.024749]' 4153 conf:(1) 8. class=8 4153 ==> averageDegreePrestige='All' 4153 conf:(1) 9. averageDegreePrestige='All' class=8 4153 ==> averageProximityPrestige='(-inf-0.024749]' 4153 conf:(1) 10. averageProximityPrestige='(-inf-0.024749]' class=8 4153 ==> averageDegreePrestige='All' 4153 conf:(1)

Table 9. 1 Sampling of association rules mined by Apriori for all workflow (cont'd)

Gene2Life	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageProximityPrestige='(-inf-0.034023]' 6287 ==> averageDegreePrestige='All' 6287 conf:(1) 2. averageDegreeCentrality='(0.034314-0.041176]' 6218 ==> averageDegreePrestige='All' 6218 conf:(1) 3. averageClosenessCentrality='(9.181183-10.439209]' 5992 ==> averageDegreePrestige='All' 5992 conf:(1) 4. averageDegreeCentrality='(0.034314-0.041176]' averageProximityPrestige='(-inf-0.034023]' 5981 ==> averageDegreePrestige='All' 5981 conf:(1) 5. averageClosenessCentrality='(9.181183-10.439209]' averageProximityPrestige='(-inf-0.034023]' 5967 ==> averageDegreePrestige='All' 5967 conf:(1) 6. class=10 5883 ==> averageDegreePrestige='All' 5883 conf:(1) 7. averageDegreeCentrality='(0.034314-0.041176]' averageClosenessCentrality='(9.181183-10.439209]' 5868 ==> averageDegreePrestige='All' 5868 conf:(1) 8. averageProximityPrestige='(-inf-0.034023]' class=10 5867 ==> averageDegreePrestige='All' 5867 conf:(1) 9. averageDegreeCentrality='(0.034314-0.041176]' averageClosenessCentrality='(9.181183-10.439209]' averageProximityPrestige='(-inf-0.034023]' 5865 ==> averageDegreePrestige='All' 5865 conf:(1) 10. averageEccentricity='(1.672619-inf)' 5862 ==> averageDegreePrestige='All' 5862 conf:(1)
Nam	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageProximityPrestige='(-inf-0.035855]' 7246 ==> averageDegreePrestige='All' 7246 conf:(1) 2. class=10 6247 ==> averageDegreePrestige='All' 6247 conf:(1) 3. averageClosenessCentrality='(6.987879-7.77803]' 6188 ==> averageDegreePrestige='All' 6188 conf:(1) 4. averageProximityPrestige='(-inf-0.035855]' class=10 6180 ==> averageDegreePrestige='All' 6180 conf:(1) 5. averageClosenessCentrality='(6.987879-7.77803]' averageProximityPrestige='(-inf-0.035855]' 6123 ==> averageDegreePrestige='All' 6123 conf:(1) 6. averageClosenessCentrality='(6.987879-7.77803]' class=10 6112 ==> averageDegreePrestige='All' 6112 conf:(1) 7. averageEccentricity='(2.219048-inf)' 6073 ==> averageDegreePrestige='All' 6073 conf:(1) 8. averageEccentricity='(2.219048-inf)' class=10 6072 ==> averageDegreePrestige='All' 6072 conf:(1) 9. averageClosenessCentrality='(6.987879-7.77803]' averageProximityPrestige='(-inf-

		0.035855]' class=10 6049 ==> averageDegreePrestige='All' 6049 conf:(1) 10. averageDegreeCentrality='(0.09-inf)' 6031 ==> averageDegreePrestige='All' 6031 conf:(1)

Table 9. 1 Sampling of association rules mined by Apriori for all workflow (cont'd)

Scoop	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageDegreePrestige='All' 7999 ==> averageBetweennessCentrality='All' 7999 conf:(1) 2. averageBetweennessCentrality='All' 7999 ==> averageDegreePrestige='All' 7999 conf:(1) 3. class=5 6510 ==> averageBetweennessCentrality='All' 6510 conf:(1) 4. class=5 6510 ==> averageDegreePrestige='All' 6510 conf:(1) 5. averageDegreePrestige='All' class=5 6510 ==> averageBetweennessCentrality='All' 6510 conf:(1) 6. averageBetweennessCentrality='All' class=5 6510 ==> averageDegreePrestige='All' 6510 conf:(1) 7. class=5 6510 ==> averageBetweennessCentrality='All' averageDegreePrestige='All' 6510 conf:(1) 8. averageProximityPrestige='(0.021789-0.023892]' 6488 ==> averageBetweennessCentrality='All' 6488 conf:(1) 9. averageProximityPrestige='(0.021789-0.023892]' 6488 ==> averageDegreePrestige='All' 6488 conf:(1) 10. averageProximityPrestige='(0.021789-0.023892]' averageDegreePrestige='All' 6488 ==> averageBetweennessCentrality='All' 6488 conf:(1)

Motif	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageProximityPrestige='(-inf-0.017965]' 5927 ==> averageDegreePrestige='All' 5927 conf:(1) 2. averageDegreeCentrality='(0.007143-0.014286]' 5006 ==> averageDegreePrestige='All' 5006 conf:(1) 3. averageDegreeCentrality='(0.007143-0.014286]' averageProximityPrestige='(-inf-0.017965]' 4910 ==> averageDegreePrestige='All' 4910 conf:(1) 4. averageClosenessCentrality='(212.941401-239.512201]' 4578 ==> averageProximityPrestige='(-inf-0.017965]' 4578 conf:(1) 5. averageClosenessCentrality='(212.941401-239.512201]' 4578 ==> averageDegreePrestige='All' 4578 conf:(1) 6. averageClosenessCentrality='(212.941401-239.512201]' averageDegreePrestige='All' 4578 ==> averageProximityPrestige='(-inf-0.017965]' 4578 conf:(1) 7. averageClosenessCentrality='(212.941401-239.512201]' averageProximityPrestige='(-inf-0.017965]' 4578 ==> averageDegreePrestige='All' 4578 conf:(1) 8. averageClosenessCentrality='(212.941401-239.512201]' 4578 ==> averageProximityPrestige='(-inf-0.017965]' averageDegreePrestige='All' 4578 conf:(1) 9. averageDegreeCentrality='(0.007143-0.014286]' 5006 ==> averageProximityPrestige='(-inf-0.017965]' 4910 conf:(0.98) 10. averageDegreeCentrality='(0.007143-0.014286]' averageDegreePrestige='All' 5006 ==> averageProximityPrestige='(-inf-0.017965]' 4910 conf:(0.98)

Table 9. 1 Sampling of association rules mined by Apriori for all workflow (cont'd)

Ncfs	weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1	1. averageProximityPrestige='(-inf-0.032765]' 6371 ==> averageDegreePrestige='All' 6371 conf:(1) 2. class=10 5838 ==> averageDegreePrestige='All' 5838 conf:(1) 3. averageClosenessCentrality='(7.653494-8.563305]' 5805 ==> averageDegreePrestige='All' 5805 conf:(1) 4. averageProximityPrestige='(-inf-0.032765]' class=10 5736 ==> averageDegreePrestige='All' 5736 conf:(1) 5. averageClosenessCentrality='(7.653494-8.563305]' averageProximityPrestige='(-inf-0.032765]' 5687 ==> averageDegreePrestige='All' 5687 conf:(1) 6. class=10 5838 ==> averageProximityPrestige='(-inf-0.032765]' 5736 conf:(0.98) 7. averageDegreePrestige='All' class=10 5838 ==> averageProximityPrestige='(-inf-0.032765]' 5736 conf:(0.98) 8. class=10 5838 ==> averageProximityPrestige='(-inf-0.032765]' averageDegreePrestige='All' 5736 conf:(0.98) 9. averageClosenessCentrality='(7.653494-8.563305]' 5805 ==> averageProximityPrestige='(-inf-0.032765]' 5687 conf:(0.98) 10. averageClosenessCentrality='(7.653494-8.563305]' averageDegreePrestige='All' 5805 ==> averageProximityPrestige='(-inf-0.032765]' 5687 conf:(0.98)

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname : Mehmet GÜNGÖREN

Date of birth and place : 14/07/1990 – ADIYAMAN

Foreign Languages : ENGLISH

E-mail :mehgungoren@gmail.com

EDUCATION

Degree	Department	University	Date of Graduation
Graduate	Computer Engineering	Yıldız Technical University	2016
Undergraduate	Computer Engineering	Yıldız Technical University	2012
High School	Natural Sciences	Rekabet Kurumu High School (YDA) - Adıyaman	2008

WORK EXPERIENCE

Year	Corporation/Institute	Enrollment
2012	İstanbul Takas ve Saklama Bankası AŞ	Software Specialist

PUBLISHMENTS

Conference Papers

1. Gungoren M. and Aktas M., (2016). “A Novel Reduced Representation Methodology for Provenance Data”, The Eighth International Conference on Advances in Databases, Knowledge, and Data Applications, 26-30 June 2016, Lizbon, Portugal.