Preprint (to appear in 8th IEEE International Conference on eScience 2012, Chicago IL USA)

Provenance Analysis: Towards Quality Provenance

You-Wei Cheah and Beth Plale
School of Informatics and Computing
Indiana University
Bloomington, IN U.S.A.
yocheah@cs.indiana.edu, plale@cs.indiana.edu

***Abstract*—Data provenance, a key piece of metadata that describes the lifecycle of a data product, is crucial in aiding scientists to better understand and facilitate reproducibility and reuse of scientific results. Provenance collection systems often capture provenance on the fly and the protocol between application and provenance tool may not be reliable. As a result, data provenance can become ambiguous or simply inaccurate. In this paper, we identify likely quality issues in data provenance. We also establish crucial quality dimensions that are especially critical for the evaluation of provenance quality. We analyze synthetic and real-world provenance based on these quality dimensions and summarize our contributions to provenance quality.**

Keywords-Data Provenance, Provenance Quality, Scientific Workflows, Provenance Analysis

I. INTRODUCTION

Over the years, provenance has seen increasing use in various applications. It is especially crucial in the aspects of reproducing results of scientific experiments and enabling the sharing and reuse of knowledge in the scientific community [25]. Extensive research has been done in the aspects of provenance capture [13, 20, 21], query [15, 17], and management [16]. With provenance becoming more ubiquitous, research is now shifting to the applications and use of provenance. Data provenance, metadata that provides the lineage or history of how data objects are generated and transformed [12], is used, but not limited to the purposes of data preservation, data reproduction [19], data auditing [14], data quality assessment [6], and the assessment of data trustworthiness [3]. Provenance may be used across various domains where different standards are used, driving needs for interoperability which is currently satisfied by the Open Provenance Model (OPM) [11].

Provenance can often be incomplete with resulting gaps and errors in the provenance record. This may be a result of the unreliable protocols between application and provenance storage [5] or the act of stitching together provenance traces through time [22]. Semi-structured workflows, such as human-centric workflows are likely to introduce uncertainty, leading to incomplete or noisy provenance traces.

Since provenance is used to assess the quality of data [2, 3, 6], it is important to evaluate the quality of provenance to ensure that captured provenance traces can be used as intended. Drawing from Lee et al. work on data quality [9], we posit that the quality dimensions of 1.) correctness, 2.) completeness and 3.) relevancy are especially critical for the evaluation of provenance quality. Although there are many other information quality (IQ) dimensions, for example, timeliness, uniqueness, validity, and believability, many are applicable when data collection is being done manually. Since provenance is largely an automated data collection process, these additional IQ dimensions are less relevant. Other IQ dimensions involve evaluating the sources of data creation.

In this paper, we assume the data creation process to be automated, and focus our attention on assessing correctness and completeness of the provenance that is captured about the created data objects. The issue of relevancy gets to questions of whether the provenance gathered is the right provenance to begin with. Since our work is with representations of OPM, we do not ask questions in this paper about OPM's relevance. So we instead focus on correctness and completeness, and do so by partitioning the problem into contextual and structural analyses. Our initial investigations examine correctness through contextual provenance and completeness issues through the structural analysis.

We propose a methodology for evaluating the quality of provenance graphs. While other studies have used provenance as a means of assessing data quality [2, 3, 6], none has taken the approach of evaluating the quality of provenance itself. Our work draws on previous research in information and data quality (IQ/DQ), and through the analysis of provenance graphs both structurally and contextually, we examine quality issues that are associated with the correctness and completeness of provenance graphs. We also identify quality issues that we have observed in both synthetic and real-world provenance, and discuss approaches to addressing some of the issues that we have discovered. As our larger research goals are related to issues in provenance quality, we characterize our contribution to provenance quality. We test our analysis methodologies on synthetic and real-world provenance. In this paper we limit our study to errors in provenance that are introduced during the capture, storage, query, or stitching phases of provenance processing. We assume that a correct provenance trace may still contain errors that are reflected at the original data.

The remainder of this paper is structured as follows. Section II discusses related work. Section III discusses our motivation for the problem. In Section IV, we describe our application data sets. Section V introduces the contextual analysis of provenance followed by the structural analysis of provenance graphs in section VI. In Section VII we talk to the application of our contextual and structural analysis techniques discussed in Section V and VI. Finally, we end in section VIII with our conclusion and discuss open issues.

II. Related Work

Data quality (DQ) assessment is typically recognized as a difficult and multidimensional concept [4]. Over the years, the field known as information quality (IQ) (analogous to DQ) has proposed multiple approaches to thoroughly understand the problem. The traditional literature in IQ deals with the quality of data in organizations. Lee et al. [1] developed a methodology called AIMQ to assess IQ based on questionnaires and analysis techniques to interpret IQ measures. Lee et al. [9] looked into IQ from the perspective of quality during data collection, data custodianship, and data consumerism. Many traditional IQ approaches rely on a questionnaire approach, which is based on subjective user inputs. A more recent study by Stvilia et al. [10] implemented a general IQ assessment framework. Their framework is based on typologies of IQ problems and a comprehensive IQ taxonomy based on 22 dimensions. The framework was then validated using Simple Dublin Core records and Wikipedia articles.

The metadata community has also studied the use of provenance in measuring data quality. Bruce et al. [8] suggest that provenance is one of the seven most commonly recognized characteristics of quality metadata. They propose the concept of tiered quality indicators, containing a set of indicators that are considered basic indicators. Beyond the basic sets of indicators, quality is improved by more detailed information. We adapt this notion and suggest a set of criteria that weights "basic provenance" more highly over "detailed provenance".

A few studies have integrated provenance as a criterion for DQ assessment. Simmhan et al. [2] developed a data quality assessment model that incorporates provenance as part of the evaluation criteria, alongside social perception, data accessibility and intrinsic metadata. Dai et al. [3] propose a model that uses provenance to evaluate the trustworthiness of data. Hartig et al. [6] used web data provenance to assess the trustworthiness and quality of the data of the Web through the use of annotating provenance graphs with impact values. Their quality model also takes into account incomplete provenance information through the use of alternative impact values and through the representation of uncertainty.

Many have used provenance as a means of assessing the quality of data, but none have taken the approach to evaluate the quality of provenance itself. Our work draws on previous IQ/DQ research to assess the quality of data provenance.

Zhao et al. [23] propose an approach that uses semantic associations for predicting missing provenance in reservoir engineering. In our case, we assume 100% confidence for the prediction of provenance and focus on what should be filled in for an incomplete provenance trace. We also provide an algorithm that scores entities in a provenance trace.

III. Problem Motivation

Our research has the assumption that provenance traces are not perfect and may have missing data or erroneous data. In this section, we discuss potential quality issues that a provenance trace may contain.

We assume provenance traces follow the OPM v1.1 standard. Provenance traces in OPM are directed acyclic graphs of causal dependencies. OPM nodes can be one of three types, namely *Processes*, *Artifacts* or *Agents*. OPM edges can be one of five types: *wasDerivedFrom*, *used*, *wasTriggeredBy*, *wasControlledBy*, and *wasGeneratedBy*. In this version of OPM, annotations can be added to any node or edge and are used to add extra information to OPM entities. These annotations are essentially name value pairs, with the name being a subject under the OPM specification and the value being a typed value with an associating namespace.

Provenance traces can be incomplete at a structural level, and contain missing nodes or edges. This may be the result of dropped messages during provenance capture or it could be a result of failed workflow executions. The identification of a failed workflow execution can be complicated. For simplicity, we define a failed provenance trace as a trace that does not contain the final process or data object of a workflow execution. Incomplete provenance, on the other hand, at the contextual level is a result of incomplete instrumentation during provenance capture. Incomplete provenance reduces the richness of provenance but does not affect the overall lineage trace.

A provenance trace may contain errors related to the accuracy of provenance capture. These errors may be as simple as numerical rounding errors, or it may be a complex problem with the provenance capture mechanism introducing errors in a random fashion that may require domain specific knowledge for fixing. Inaccurate provenance may also arise when duplicate and conflicting provenance records are captured. Since provenance typically passes through a capture, store and query phase before being returned to a user, the inaccuracy of provenance could stem from errors in the system at any one of these phases. Provenance traces that are stitched together have an additional phase and errors may also be introduced here.

Consistency issues can also be another source of problems in provenance traces. The problem of consistency falls under the data quality dimension of correctness. When two different provenance traces are stitched together to form a single provenance trace, the combined provenance trace may be inconsistent, since provenance may have been captured differently according to different standards. A good example of this is inconsistencies in timestamps. Timestamps are represented in many ways internationally. Two of the usual formats are MM-DD-YYYY versus DD-MM-YYYY and it is apparent how inconsistencies in a trace may lead to confusion for a date such as 02-01-2012. Timestamps can also become inconsistent with the causal dependencies in a provenance graph due to a variety of reasons. One reason for this could be because of time drifts in the provenance capture system.

Traditional data cleaning techniques may be able to solve some of the issues in accuracy and consistency. However, some of the problems require stitching provenance traces together, and this can involve multiple provenance databases having different standards and schemas.

Our model of analysis is shown in Fig. 1. Correctness is assessed primarily through contextual analysis while completeness is assessed through structural analysis. Both types of analysis are done at graph level (G), while multi-graph analysis (M-G) is used for completeness and node/edge analysis (N-E) for correctness approach.

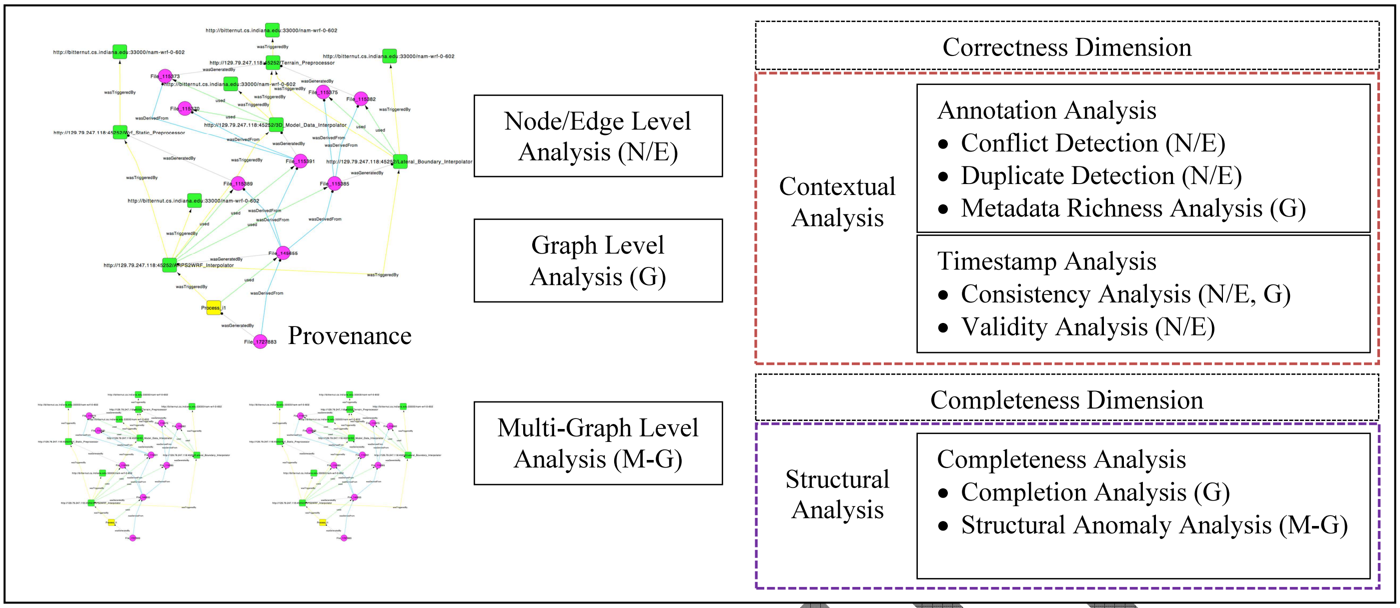

Provenance
Node/Edge Level Analysis (N/E)
Graph Level Analysis (G)
Multi-Graph Level Analysis (M-G)
Correctness Dimension
Contextual Analysis
Annotation Analysis
• Conflict Detection (N/E)
• Duplicate Detection (N/E)
• Metadata Richness Analysis (G)
Timestamp Analysis
• Consistency Analysis (N/E, G)
• Validity Analysis (N/E)
Completeness Dimension
Structural Analysis
Completeness Analysis
• Completion Analysis (G)
• Structural Anomaly Analysis (M-G)

Fig. 1. Provenance quality analysis overview. Correctness is assessed through primarily contextual analysis while completeness through structural analysis. Both types of analysis are done at graph level (G), while multi-graph analysis (M-G) is used for completeness and node/edge analysis (N-E) for correctness.

Our goal is to detect ambiguities and conflicts in real and synthetic provenance traces. Moreover, we hope to complete portions of missing provenance for workflow traces that have successfully executed, but have dropped messages. The severity of dropped messages in a workflow execution may impair our ability for us to complete the missing picture of provenance traces. Since we are dealing with a homogenous set of workflows, we assume a priori provenance of a complete workflow and attempt to repair provenance traces based on this knowledge. We also propose a provenance quality evaluation mechanism that scores and validates provenance traces. This scoring mechanism will serve the purpose of providing a useful comparison between provenance traces before and after the repair of provenance traces.

IV. Application Data Sets

We apply our methodology to both synthetic provenance and provenance from a NASA production satellite ingest processing pipeline. The synthetic dataset consists of provenance graphs for six types of workflows. We sampled 500 provenance graphs from a 10GB provenance database with known error characteristics [5], for a total of 3000 provenance traces, a total that amounts to approximately 630MB of provenance data. For each type of provenance graph, the distribution of the number of provenance graphs with successful workflow runs, failure runs, runs with dropped messages, and runs with both failure and dropped messages is maintained. The original distribution of these provenance graphs consists of 55-60% of graphs without failures and without dropped messages. Provenance of workflows with dropped messages consists of approximately 20%, and the remainder consists of provenance that involves failures, with and without dropped messages. The exception to this distribution is the larger provenance graphs of MotifNetwork and Animation workflows, where the provenance of failed workflows constitutes 50% of the total. The other 50% is split between provenance of successful workflow runs with or without dropped messages.

The synthetic provenance data was further manipulated by introducing additional errors that we have observed in actual applications, these errors include:

i) Duplicate and conflicting annotations
ii) Manipulation of timestamps (altering of timestamps, swapping of begin and end times)
iii) Reordering of timestamps between edges
iv) Duplicate edges

The real-world application dataset that we use comes from NASA's Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) [27] ingest processing workflows. AMSR-E is a passive microwave radiometer aboard a polar orbiting Aqua satellite that generates data about the poles. Provenance was captured for approximately 1 month of data for different scientific data products including sea-ice (L3), rain (L2B), snow (L3), land (L2B, L3), ocean (L2B, L3), drift (L3), where L2B and L3 refer to the data processing levels defined by NASA [28]. The daily L2B data consists of 905 provenance traces for each data product, while the L3 data consists of about 33 traces for each data product. Six 5-day traces were available for snow (L3), 5 for weekly ocean (L3) and a single monthly trace for each of ocean (L3), snow (L3), and rain (L3) constitute the rest of our sample. This total of 2890 provenance graphs amounted to approximately 60MB of provenance data.

V. Contextual Analysis

Contextual analysis addresses the correctness of provenance. Errors are a result of a failed or incomplete workflow execution or could be a result of the unreliability of the provenance capture mechanisms. We employ a number of methods of analysis that are explained in this section.

A. Methodology

The execution of workflows may contain anomalies. An anomaly is a deviation from the norm, where the norm is established within the context of a single provenance graph. We assume that the provenance capture process is sufficiently reliable and regular. Provenance captures of the workflow execution can reflect these anomalies in a number of ways. We check for anomalies in provenance by analyzing different properties of nodes and edges in a provenance graph.

One such method examines provenance through analysis of annotations in a provenance graph. Our analysis is performed through algorithms that look for duplicate and conflicting annotations. We also use a simple clustering algorithm that clusters annotations based on the count of annotations for each parent types to identify anomalies in the annotation of provenance graphs.

The goal of our methodology is to expose correctness and consistency issues that affect the quality of a provenance graph. Our contextual analysis techniques focus on two aspects in a provenance graph: annotation analysis and timestamp analysis.

Annotation Analysis - We assume provenance contains duplicate events. Duplicates are far more likely to occur in the annotations than in the structural aspect of a provenance graphs. Duplicate detection is used to detect exact replicas of annotations and also potential conflicting annotations within a single provenance graph. To the latter, we identify annotations under the same node or edge that have the same name but different values to be one that may be potentially conflicting.

Clustering annotations based on the number of annotations for each node or edge is useful to observe whether a particular node or edge has richer or poorer annotations than the norm. This is first done by grouping nodes or edges by type, where the type of the node refers to the kinds of nodes in OPM: *Process*, *Artifact*, and *Agent*. The type of edges refers to the 5 different causal dependencies under the OPM specification: *wasDerivedFrom*, *used*, *wasTriggeredBy*, *wasControlledBy*, and *wasGeneratedBy*. For each type group, we do a second grouping based on the number of annotations for each group. Finally, we use a relative threshold to determine whether a particular group of nodes or edges have richer or poorer annotations than usual. For each type group, we select the number group with the highest occurrence and then compare the ratio of each number group with the group of the number group with the highest occurrence. A preset threshold is used to determine whether a certain group is considered an anomalous group. For example, nodes of the OPM *Artifact* type are grouped into a single group, and nodes of the OPM *Process* type are grouped into a single types group. For the OPM *Artifact* group, we will then group OPM *Artifacts* with 4 annotations into a single group, while OPM *Artifacts* with 6 annotations will be grouped into another group and so forth. If the OPM *Artifacts* with 6 annotations group has 126 occurrences while the group with 4 annotations only has 3 occurrences, we will select 126 as the denominator for the comparison. Using a preset threshold of 5%, we will find that the OPM *Artifacts* group with 4 annotations yields only a composition of 2.38%, hence rendering this group as an anomalous group.

Timestamp Analysis - Even though timestamps are optional in the OPM v1.1 format, they provide extra context and information to the transformation of data when present. We analyze timestamps in provenance graphs when present to ensure that consistency for timestamps is preserved throughout the provenance graph, i.e. that the timestamps for events are such that the timestamp of events are not in conflict with the causal order of events. For example, a certain process P1 was triggered by P2 and P2 was triggered by P3. If the timestamps of P2 was triggered by P3 were such that it implies that it occur before P1 was triggered by P2, then we will have inconsistent timestamps.

In addition, we examine timestamps to ensure that they are valid in the sense that they are well formed. The OPM definition does not restrict an event to possess an exact timestamp, but allows for a time range for an event. We take this into consideration and also check to see if the time range is a valid time range, i.e. that the begin time of the time range is before the end time of the time range.

As proof of concept, we apply our analysis techniques on a number of provenance datasets. These datasets are discussed in more depth in the following subsection.

B. Evaluation

In application of the proposed methodologies, we have been successful in detecting duplicates in annotations for synthetic provenance and AMSR-E workflows.

AMSR-E provenance was collected using a scavenging approach, where provenance is aggressively mined from log files. Duplicate annotations occur in the AMSR-E provenance as a result of log files not being preprocessed and cleaned beforehand, resulting in the possibility of logs containing multiple entries of a certain event. As a result, duplicates are easily picked up. Our results indicate that the magnitude of duplicates can be large so we implemented a cleanup tool that can scour the Karma provenance system's database for exact duplicates; this can improve the performance of provenance graph queries. In addition to hindering scalability, duplicate annotations could suggest other problems, for example, a leakage of data from an unintended source.

An example of duplicate annotations is provided through an AMSR-E monthly ocean provenance graph. For this graph, we observe a high number of duplicate annotations as depicted in Fig. 2. The provenance graph itself contains 875 artifacts, which are all files. Approximately 90% of these files contain duplicate annotations that make up half of the total annotations for each file. Only about 35 files (or 4% of the total files) have no duplicate annotations. 29 files have 27.27% of annotations consisting of duplicates and another 29 files have 33.33% of their total annotations containing duplicates. Based on our observations, duplicate annotations could possibly double the amount of storage required if not handled properly. These duplicate annotations are a result of the processing of provenance from log files and this once again confirms that log files are noisy and either requires pre-processing to

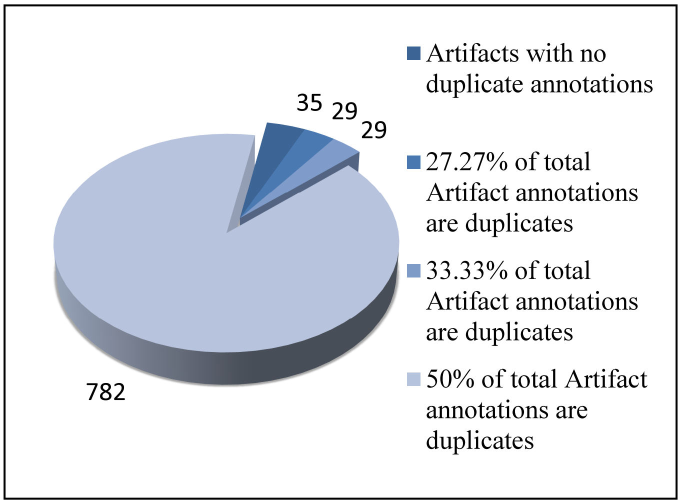


Fig. 2. Distribution of duplicate annotations in *Artifacts* of AMSR-E Monthly Ocean Provenance Graph. The composition of duplicate annotations for each *Artifact* node is shown in the pie chart in terms of percentages.

cleanup log files or post-processing to eliminate duplicates from provenance traces.

We are also able to identify nodes and edges in a provenance graph that has an unusually high or low number of annotations through clustering. Our results for this performed against an AMSR-E monthly ocean provenance graph is provided in Table I below. There are two anomalies that we observe from Table I, namely the *wasDerivedFrom* edges with 2 annotations, and the *Artifacts* that have 16 and 22 annotations.

For our timestamp analysis, we observe that the start and end timestamps in a time range may be swapped for a causal relationship in provenance graphs, resulting in an invalid time range. Although this phenomenon is rarely observed, checks of this kind are important to ensure that provenance graphs have high data quality. This is illustrated in Fig. 3, where Process_{6250} was triggered by Process_{6251} with an invalid time range. In this case, our Karma provenance system accurately captured provenance. However, due to the logs being erroneous, ambiguity is introduced into the provenance trace. Another error that we have observed is that the temporal data conflicts with the structural causalities (Fig. 4). Under the OPM specification the structural causalities takes precedence over temporal metadata since temporal data is optional. Nevertheless, the presence of both forms of data adds conflicting information to the provenance trace and should be rectified.

TABLE I. CLUSTER RESULTS OF ANNOTATIONS IN AMSR-E MONTHLY OCEAN PROVENANCE GRAPH

Type	Number of Annotations	Occurrences	Percentage of occurrence for type
wasGeneratedBy	1	1	50%
	3	1	50%
wasTriggeredBy	1	2	50%
	2	1	25%
	3	1	25%
used	1	874	100%
wasDerivedFrom	2	4	0.46%
	4	869	99.54%
Process	5	4	100%
Artifact	16	29	3.31%
	4	63	7.2%
	22	1	0.11%
	12	782	89.37%

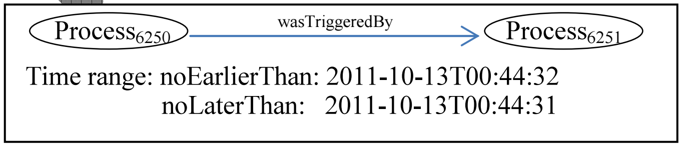
a. Annotations considered here have been stripped of duplicates

VI. STRUCTURAL ANALYSIS

The completeness of a provenance graph is determined through structural analysis by comparing nodes and edges of a graph to a template of generation, such as a workflow. Provenance graphs may not carry a generation template, in which case completeness would need to be approximated such as through machine learning, a subject of further investigation. Structural flaws occur in provenance graphs due to errors in the execution of a workflow or the dropping of event recordings during a workflow execution. Completeness analysis assumes that provenance graphs are directed acyclic graphs, and adhere to an execution template as stated earlier.

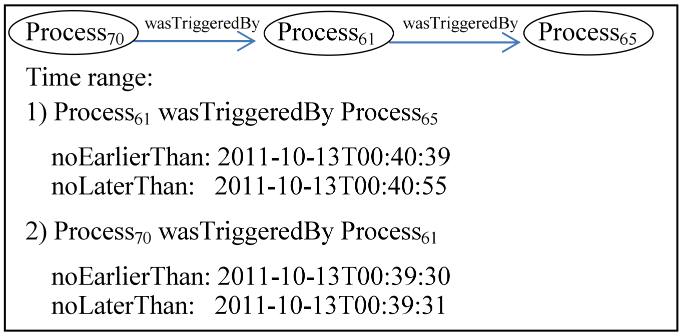
A. Methodology

Completeness analysis is a technique to evaluate the quality of each node by considering the number of errors that are detected for a node. The errors that we suggest that are relevant structurally are the number of input and output edges for each node and also other contextual errors that have been mentioned in the previous section.



Time range: noEarlierThan: 2011-10-13T00:44:32
noLaterThan: 2011-10-13T00:44:31

Fig. 3. An example of an invalid time range in a single OPM edge.



Time range:

1) Process_{61} wasTriggeredBy Process_{65}

noEarlierThan: 2011-10-13T00:40:39
noLaterThan: 2011-10-13T00:40:55

2) Process_{70} wasTriggeredBy Process_{61}

noEarlierThan: 2011-10-13T00:39:30
noLaterThan: 2011-10-13T00:39:31

Fig. 4. An example of temporal data conflicting with structural data.

We use a 2-pass algorithm that evaluates the structure of a provenance graph and structurally repair provenance graphs. The first-pass uses a depth-first traversal to identify disconnects within a provenance graph. This algorithm begins with a list of initial nodes or nodes that have no parents. The initial quality of each node and edge is evaluated based on the number of errors associated with each node or edge. A second pass is done to repair the provenance graph and reevaluate the quality score of the entire graph after reparation. It is for the repair of the graph structure that we assume knowledge of the execution of a complete workflow graph. Through direct comparison we are able to infer perfectly missing nodes and edges in the provenance graph.

We also built into our framework a method for identifying graphs with anomalous number of nodes or edges. For the identification of these structural anomalies, we use the statistical method of outlier detection based on interquartile ranges. This is the same method employed in box-and-whisker diagrams. In this method, we establish upper and lower fences by using the default formulas listed below. (The multipliers of 1.5 can be fine-tuned for different datasets.) Values that fall outside these fences are considered as outliers.

Interquartile Range = Upper Quartile – Lower Quartile
Upper Fence = Upper Quartile + 1.5 * Interquartile Range
Lower Fence = Lower Quartile - 1.5 * Interquartile Range

B. Structural Anomaly Analysis

Structural anomaly analysis identifies graphs that are incomplete or over-complete relative to other graphs of the same kind. It does so by highlighting graphs with nodes or edges that are more or less than the norm. Figure 5 gives an example for how to identify outliers. We sampled 450 SCOOP workflows from our synthetic dataset and analyzed the number of edges and nodes for the provenance graphs of these workflows. The upper and lower fences for nodes are 23.5 and 19.5 respectively, while the upper and lower fences for edges are 35.5 and 23.5. Using this method, we are able to identify 76.03% of incomplete or over complete graphs by analyzing the node counts and we are able to identify 81.20% of incomplete or over complete graphs through the analysis of the edge counts.

From the plots, it is obvious what the norms are for both the number of nodes and edges due to the homogeneity of the nature of the provenance trace. However, in cases where the norm is evenly spread out among a range of values, this method would still be applicable since it is by nature statistically robust and does not make any assumptions about the underlying statistical distribution.

C. Completion Analysis

The completion of a provenance graph is essential to help improve the quality of provenance graphs. Structural completion of provenance graphs addresses the completeness aspect of quality. It is important to note that the idea of determining whether a provenance graph is complete requires a complete graph template or sample to be compared to. For provenance graphs of scientific workflows, this is easy to obtain. Since scientific workflows are often generated from a workflow template and executed through a workflow engine, a complete or ideal workflow template is available. Moreover, if the provenance graphs are homogeneous, a machine learning algorithm can be applied to a large sample of provenance graphs to obtain a complete provenance template. There are cases where provenance graphs are unique and do not possess a similar provenance graph or template that can be compared to. For these cases, the structural completeness of a provenance graph will be difficult to determine.

Through the use of configuration files that contain the number of input edges, the number of output edges, and also

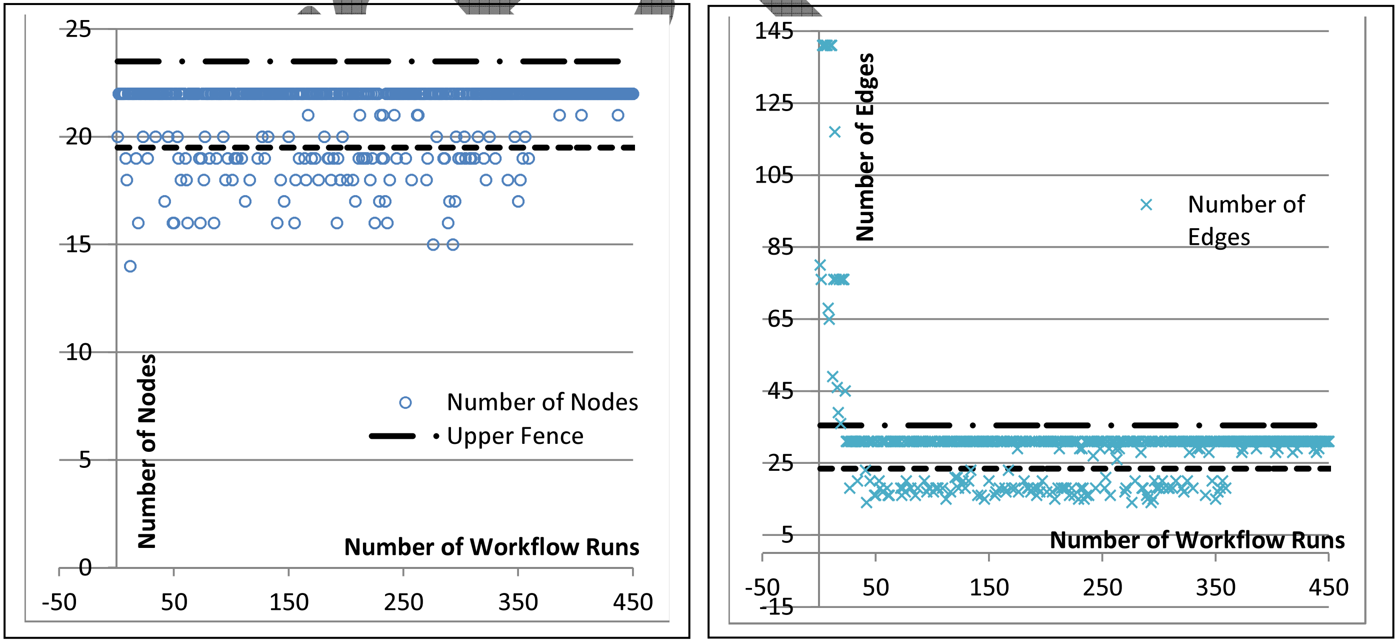
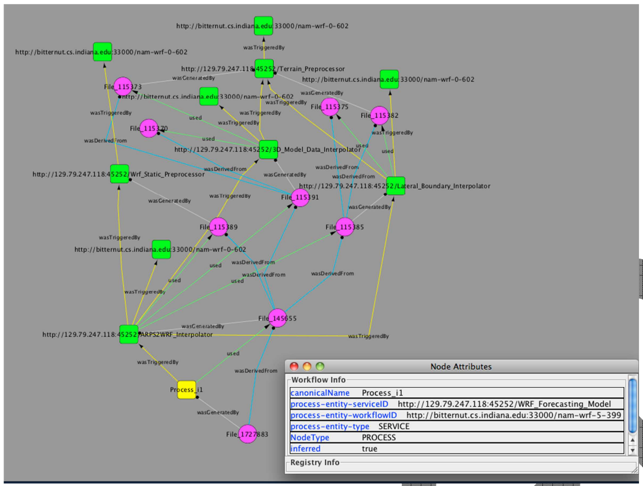


Fig. 5. Identification of outliers for (a) node count and (b) edge count in SCOOP provenance graph.

the connectivity between nodes, we are able to make inferences to reconnect missing nodes and edges in a provenance graph. Although we have managed to complete provenance graphs to perfection, this does not always yield the actual provenance of events. The restoration of a provenance graph should only correct provenance graphs to the extent that it accurately reflects provenance events. For instance, the restoration of a failed provenance graph should not proceed beyond the node that fails, but any disconnects in the graph before the failure point should be restored. The identification of a failure point in a workflow or a process is not an easy task and is ongoing research [24]. For simplicity, we simply dropped all of the inferred nodes/edges if they do not connect to an existing terminating node in a provenance graph. A terminating node or nodes are used to signal the end of a workflow execution and can be read into the framework through the use of a configuration file.


http://bitternut.cs.indiana.edu:33000/nam-wrf-0-602
http://129.79.247.118:45252/Terrain_Preprocessor
http://129.79.247.118:45252/3D_Model_Data_Interpolator
http://129.79.247.118:45252/Wrf_Static_Preprocessor
http://129.79.247.118:45252/Lateral_Boundary_Interpolator
http://129.79.247.118:45252/ARPS2WRF_Interpolator
File_115373
File_115370
File_115375
File_115382
File_115391
File_115389
File_115385
File_145655
File_1727883
Process_i1
wasTriggeredBy
wasGeneratedBy
wasDerivedFrom
used
Node Attributes

Workflow Info

canonicalName	Process_i1
process-entity-serviceID	http://129.79.247.118:45252/WRF_Forecasting_Model
process-entity-workflowID	http://bitternut.cs.indiana.edu:33000/nam-wrf-5-399
process-entity-type	SERVICE
NodeType	PROCESS
inferred	true

Registry Info

Fig. 6. Visualization of a NAM-WRF provenance trace through the Cytoscape Karma visualization plugin. Inferred node is marked in yellow.

D. Evaluation of Quality

We propose a scoring mechanism for assessing the overall quality of a graph along the dimensions of correctness and completeness. Scoring is done through the valuation at both graph and node/edge levels. Evaluation at the node level has both structural and contextual aspects. Specifically, the structural errors include the missing number of input/output edges and the contextual errors include duplicate or conflicting annotations. For edges, the evaluation of quality is focused on the contextual aspect of when an event happened between two nodes. The evaluation of timestamps and their consistencies are the primary aspect here.

We record the number of errors that are associated with each node/edge and calculate the quality for a node/edge using the following:

$$Quality = U - (\, n_{errors} / C \,)$$

where U and C are constants. U gives the maximum score for a node/edge without any detected errors. We assign $U = 1$ for simplicity, so that all nodes/edges without errors have a score of 1. C refers to a cutoff, where the number of errors that exceeds this cutoff would yield a negative score. If the cutoff value is used as an indicator for the number of acceptable errors for a node/edge, one can easily identify nodes/edges that have the number of errors that fall outside the cutoff threshold. We propose this equation since the lower bound for the number of errors is not usually known, but we are able to identify the case where a node has no errors.

The graph level information quality assessment takes into consideration the completeness of nodes and edges and also examines the detection of cycles and the consistency of timestamps throughout the provenance trace. Errors at the graph level are errors that belong to a graph "node" so the score using the same equation. The overall quality of a provenance graph is obtained by averaging the total quality score for edges and nodes over the sum of the expected number of nodes and edges.

$$Graph_{Quality} = \frac{(\sum NodeScore + \sum EdgeScore + GraphNodeScore)}{\sum ExpectedEdges + \sum ExpectedNodes + 1}$$

VII. Provenance Quality Application

We apply the analysis techniques of Sections V and VI to the real-world NASA AMSR-E dataset, and summarize what we found here. We do not carry out completion analysis since the application lacked a template for comparison at this time. Our analysis is applied to the daily provenance traces (both L2B and L3) data since they are sufficiently large (for L2B) and moderate (L3) in size. We discuss our high-level findings below.

1) We observed inconsistencies between the temporal data and the causal dependencies for every NASA AMSR-E provenance trace. Further investigation has uncovered that all timestamps do not have their hour field in their timestamps set. As a result, all timestamps fall within the range of 00:00:00 to 00:59:59. Invalid time ranges between edges were also observed in 183 of the provenance traces. One other interesting thing is that for one of the drift (L3) provenance trace, the time range between its edges is greater than a week. This seems to be an obvious error since the provenance trace in question is a daily provenance trace.
2) No structural anomalies were detected for the provenance traces of ocean (L2B), land (L2B), snow (L3), drift (L3). We did however find anomalies for the rain (L2B) (120 anomalies), land (L3) (7 anomalies), seaice (L3) (8 anomalies), and snow (L3) (8 anomalies). Since the provenance traces generally have very strong norms, subtle variations from the norm result as an anomaly. In reality, only a single anomaly can be classified as an outlier for each of the L3 traces (land, seaice and snow), this may be due to the strong norm that we observed. A histogram approach may be a good tool to combine with our current approach to further segregate the true outlier from the other anomalies.

3) For all the provenance traces, the majority of duplicate annotations occur in *Artifacts*. Out of the total 4919 *Artifacts*, 3677 of them contain duplicate annotations. There are however cases where *wasTriggeredBy* edges contain duplicate annotations (414 of a total of 3641 edges). All of the potentially conflicting annotations are annotated under the *wasTriggeredBy* edges.

Our results demonstrate the usefulness of our proposed techniques. We see these techniques as part of an auditing and validation suite that aids the user in isolating and identifying problems in provenance traces. Since the task of correction often requires domain knowledge, we leave this to the user.

VIII. Conclusion and Open Issues

In this paper we establish the quality dimensions of correctness and completeness as measures of provenance quality. Our motivation for evaluating the quality of provenance is directly tied to the quality of the data itself, about which provenance describes. Based on the quality dimensions that we have established, we set out to evaluate the quality of provenance traces through partitioning the problem into a contextual one and a structural one. We also summarized our contributions to provenance quality.

Our future work will expand on the structural completion of provenance traces to account for the importance of nodes and edges towards provenance quality. In addition, we are seeking to apply the methodology to broader and appropriate types of application provenance. We are also refining our current approach by adding additional analysis tools that are targeted towards the multiple provenance graph levels. The assumptions of provenance collection being a largely automated process and the appropriateness of the captured provenance with respect to the data can be relaxed to open up interesting questions for evaluating provenance quality.

Acknowledgments

This work was funded in part by NASA under grant number NNX10AM03G. We thank our fellow Data to Insight center colleagues Yiming Sun, Yuan Luo, Peng Chen, and Devarshi Ghoshal for their discussions and contributions to this paper. We also thank Lavanya Ramakrishnan and Elif Dede for suggestions towards the writing of this paper.

References

[1] Y. W. Lee, D.M. Strong, B. K. Kahn, and R. Y. Wang. AIMQ: A methodology for information quality assessment. *Information & Management, 40*, 2 (2002), 133-146.

[2] Y. Simmhan and B. Plale, Using Provenance for Personalized Quality Ranking of Scientific Datasets, *Intl. Journal of Computers and Their Applications (IJCA)*, 18 (3), 2011, pp. 180-195, ISCA.

[3] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An approach to evaluate data trustworthiness based on data provenance," in *SDM '08: Proceedings of the 5th VLDB workshop on Secure Data Management*, pp. 82–98, Springer-Verlag, 2008.

[4] F. Naumann, *Quality-driven query answering for integrated information systems*. Springer Verlag, 2002.

[5] Y.-W. Cheah, B. Plale, J. Kendall-Morwick, D. Leake, L. Ramakrishnan, A Noisy 10GB Provenance Database, *2nd Intl. Workshop on Traceability and Compliance of Semi-Structured Processes (TC4SP2011)*, Clermont-Ferrand, France, 2011.

[6] O. Hartig and J. Zhao. Using web data provenance for quality assessment. *Proceedings of the 1st Intl. Workshop on the Role of Semantic Web in Provenance Management, ISWC*, 2009.

[7] F. Curbera, Y. N. Doganata, A. Martens, N. Mukhi, and A. Slominski. Business Provenance - A Technology to Increase Traceability of End-to-End Operations. In *16th Intl. Conference on Cooperative Information Systems (CoopIS'08)*, pages 100–119. Springer, 2008.

[8] T. R. Bruce, D. I. Hillmann. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In *Metadata in Practice*, Chicago, ALA Editions, 2004.

[9] Y. W. Lee, D. M. Strong, Knowing-why about data processes and data quality. *Journal of Management Information Systems, 20*(3), (2003-2004), 13-39.

[10] B. Stvilia, L. Gasser, M.B. Twidale and L.C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720-1733, 2007.

[11] L. Moreau, B. Clifford, et al. The Open Provenance Model core specification (v1.1), *Future Generation Computer Systems (FGCS)*, 27, 2011, pp. 743-756, Elsevier.

[12] Y. Simmhan, B. Plale, D. Gannon. A survey of data provenance in e-science, *SIGMOD Record* 34(3): 31-36, 2005.

[13] Y. Simmhan, B. Plale, and D. Gannon, Karma2: Provenance Management for Data Driven Workflows, *Intl. Journal of Web Services Research*, IGI Publishing, vol. 5, no.2, 2008.

[14] S. Miles, P. Groth, M. Branco, and L. Moreau, The requirements of recording and using provenance in e-science experiments. *Journal of Grid Computing*, 2006.

[15] T. Heinis and G. Alonso, Efficient lineage tracking for scientic workows, in *SIGMOD*, 2008.

[16] M. K. Anand, S. Bowers, T. McPhillips, and B. Ludascher, Efficient provenance storage over nested data collections, in *EDBT*, 2009.

[17] M. K. Anand, S. Bowers, and B. Ludscher, Techniques for efficiently querying scientific workflow provenance graphs, in *EDBT*, 2010.

[18] A. Chapman, H. Jagadish, and P. Ramanan. Efficient provenance storage. In *Proceedings of the ACM SIGMOD/PODS Conf.*, Vancouver, Canada, 2008.

[19] C. Silva, J. Freire, and S. P. Callahan. Provenance for visualizations: Reproducibility and beyond. Computing in Science & Engineering, 9(5):82–89, 2007.

[20] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Vistrails: visualization meets data management. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD Intl.Conf. on Management of data*, New York, NY, USA, 2006.

[21] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer. Provenance aware storage systems. In *Proceedings of the 2006 USENIX Annual Technical Conf.*, June 2006.

[22] P. Missier, B. Ludaescher, et al. Golden-Trail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository, *7th Intl. Digital Curation Conf. (IDCC)*, Bristol, UK, 2011.

[23] J. Zhao, K. Gomadam, V. Prasanna: Predicting Missing Provenance Using Semantic Associations in Reservoir Engineering. In *Proceedings of the 5th IEEE Intl. Conf. on Semantic Computing (ICSC)*, 2011.

[24] N. Russell, W.M.P. van der Aalst, and A.H.M. ter Hofstede. Workflow Exception Patterns. In *E. Dubois and K. Pohl, editors, Proceedings of the 18th Intl. Conf. on Advanced Information Systems Engineering (CAiSE'06)*, volume 4001 of Lecture Notes in Computer Science, pages 288–302. Springer-Verlag, Berlin, 2006.

[25] S. B. Davidson and J. Freire. Provenance and Scientific Workflows: Challenges and Opportunities. In *SIGMOD Conference*, 2008.

[26] A community white paper developed by leading researchers across the United States. Challenges and Opportunities in Big Data. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[27] NASA AMSR-E. http://aqua.nasa.gov/about/instrument_amsr_dp.php

[28] NASA Earth Observing System Data and Information System Data Processing Levels. http://science1.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/