A Novel Approach to User Involved Big Data Provenance Visualization

Ilkay Melek Yazici, Mehmet S. Aktas
Computer Engineering Department
Yildiz Technical University
Istanbul, Turkey
email:ilkay.yazici@std.yildiz.edu.tr,
mehmet@ce.yildiz.edu.tr

Mehmet Gokturk
Computer Engineering Department
Gebze Technical University
Koceli, Turkey
email: gokturk@bilmuh.gyte.edu.tr

Abstract—We are living in the "big data" age. In many ways, big data is equivalent to complexity or mess. Extracting relevant information from any complex environment is a challenging but necessary task required in every scientific field. An assortment of graphs, figures, and charts have been developed to visualize *n*-dimensional data since the early ages of science. In the recent past, the number of dimensions in a visualization were limited by computational factors. Visualizing *n*-dimension is difficult but achievable by using data projection and reduction methods. Unfortunately, these methods often introduce ambiguities and inaccuracies, which can subtly corrupt results. Data provenance chronicles the core life cycle of a data set, which includes data source and creation processes, accounts for many of the processing techniques that a data set is subject to, like debugging, auditing and quality control. Additionally, data protection mechanisms such as data access control and authenticity valuation methods are also tracked by provenance. In this paper, we introduce an effective method to visualize and analyze semantic provenance data by adhering to the Human Computer Interaction principles. Our proposed data provenance visualization system involves the user in the visualization process. By capturing and analyzing a user's attentiveness and perception level, we develop a provenance visualization system with specific visualization types and methodologies.

Keywords-big data; provenance; visualization; open provenance model (OPM); human machine interface(HMI).

I. INTRODUCTION

Data provenance records the journey of data from its creation to its application [1]. Data provenance is produced with complex transformations and processes like workflows [2]. Data provenance collection systems capture provenance on the fly. However, their collection mechanisms may be faulty and have dropped provenance notifications. Hence, provenance records may be partial, partitioned, or simply inaccurate [3]. Incompleteness and inconsistency of provenance records, if they exist, are a challenge for analyzing provenance datasets.

As more information technology (IT) systems are developed and implemented, the interpretation requirements for big provenance data also increase [4]. Data visualization is an important field, which offers many techniques to develop an intuitive interpretation of data, often by way of efficient visualization capabilities [5]. Therefore, data visualization is an important step in the process of maximizing perception efficiency [6].

Data visualization is considered visual communication by many experts. Many visual communication techniques create processes that improve visual data representation via diagrammatic displays that use data properties, variables, and information units. Effective visualization can assist users with analyzing and evaluating data by making complex data more accessible, clear and usable [7].

In this paper, our goal is to help users by generating an effective provenance visualization process that complies with Human Computer Interaction (HCI) principles and user navigation models. A mind map for an experimental protocol will be used to jointly explore provenance and user interaction. To achieve these goals, this study states the following objectives, which are briefly described below:

Objective 1: Achieve user-assisted visualization of Big Provenance. We will develop real-time and offline visualization techniques by capitalizing on existing visualization techniques from relevant fields. In our case, we concern ourselves with the social media domain. We will research existing provenance visualization methods for temporal provenance data and analyze the various visualization techniques and methodologies. A user's preference for customized visualizations, and their perception of these preferred methods, will be analyzed using HCI evaluation methods. The results will be used to improve the visualization system.

Objective 2: Design and create Big Provenance visualization methods, such as visualization layouts, and templates. Few research studies on provenance visualization exist in the literature. We will study and extend the existing methodologies and introduce novel visualization layouts that deal with temporal provenance data. We argue that new customized layouts and visualization methodologies can be developed. The techniques need to be based on both the provenance data domain and user requirements.

Objective 3: Develop domain-independent Big Provenance visualization. We will study existing domain-independent provenance specifications for data representation. Particularly, we will consider Open

 ISBN: 978-1-61208-558-6

Provenance Model (OPM) and Provenance Ontology (PROV-O) specifications.

The remainder of this paper is organized as follows: Section II provides the relevant literature review. Section III discusses various application scenarios to describe the scope of this research. Section IV reviews our proposed methodology and Section V describes the aspects of Human Machine Interface (HMI) that are important to this study. Finally, Section VI presents the conclusion and future work of our paper.

II. LITERATURE SUMMARY

Kunde's seminal work [12] on Provenance Visualization Components provides us with an important set of requirements, summarized below:

a) Process: a summary of the process as a sequence of data inspection steps;
b) Results: user-centric and includes the intermediate- and end-results of interactions;
c) Relationship: the relationship between actors or interactions;
d) Timeline: observations of time;
e) Participation: the accuracy of the participants
f) Compare: describes the distinction between two subjects
g) Interpretation: the individual visualization related to the end user's special questions

In his considerable research, Chen's visualization goal is to satisfy the audience or reader. He addresses Kunde's requirements as follows: a-c) Chen's visualization tool is based on a known provenance model called Open Provenance Model (OPM) for provenance representation. This model represents entities and relationships as nodes and edges on a graph [9]. OPM models can represent a full graph with process steps and process results, an abstract graph with any of them; d) OPM represents time information as edges and nodes; e) OPM represents participation by agents. Chen's tool enable users to evaluate the accuracy of the participations visually; f) a tool is used for comparing attributes of nodes. Chen extended the DePiero graph-matching algorithm to compare provenance graphs; g) Chen developed a customized layout algorithm and a visual style to interpret specific use cases.

In Chen's research, network application provenance is studied from large-scale distributed apps that run on large testbeds like the Planet Lab or in network simulations like the provenance data from NASA satellite imaginary.

Provenance visualization research is often restricted to small graphs, graph matching techniques and graph layouts. Taverna is a scientific workflow management system (SWMS) that benefits from Chen's visualization method. Taverna helps answer questions based on experimental results. VisTrails is another SWMS that can navigate workflows by using the users intuition to compare workflows, intermediate- or end- results, or to evaluate the results. Probe-It is a popular SWMS that allows scientists to focus on intermediate or final visualization results for back and forth provenance [8].

The Prototype Lineage Server (PLS) enables users to get lineage information by searching metadata groups that can provide helpful details about the workflow transformations and data products. Pedigree Graph is a tool in Multiscale Chemical Science (CMCS) and Multi-Scale Chemistry (MSC); it uses a portal to view multi-dimensional provenance. The My Grid tool displays graphs based representations of RDF-coded provenance by using Haystack. Provenance Explorer, reliable provenance visualization tool that creates customized dynamic views of scientific provenance data depending on the user requirements and access privileges.

In another study [8], Chen collected e-science provenance data, which yielded an OPM visualization Direct Acyclic Graph (DAG). The temporal representation of provenance graphs has generated partitions that maintain temporal order between node subsets by using the Logical-P algorithm [8]. Graph annotations and fully labeled graphs are visualization tools for representation. Temporal representation failures can be detected in workflow executions or the provenance capture.

III. APPLICATION FEATURES

Features of study like ontology model, layout components, use cases and visualization tool are presented in this section.

A. Ontology

Ontologies are at the heart of any semantic technology. An ontology is formally defined as a set of specifications associated with a concept. Many researchers use ontologies as mechanisms for sharing and reusing information. Ontologies can easily express relationships between identifiers. They share many qualities with knowledge representation systems.

PROV-O: The PROV ontology (PROV-O) is an OWL-based ontology that allows PROV-data models to use RDF mapping developed by W3C. The PROV-O terms are defined as classes and properties, and they are grouped into three categories: starting point terms, expansion terms and qualifying relationship terms to provide an incremental input to the ontology [11].

Starting point term classes and properties are used to creating simple provenance descriptions that can be detailed using the terms of other categories. Ultimately, starting point term classes are a small set of classes and properties that create simple, initial provenance descriptions. PROV-O categories are listed and defined below:

- **Entity**: An entity is a conceptual, digital, physical or virtual object with specific aspects. It can be real or imaginary.
- **Activity**: An activity is an action that repeats over a period-of-time and is acted upon by entities; actions may include processing, consuming, modifying, transforming, using, relocating or generating entities.

 ISBN: 978-1-61208-558-6

- **Agent**: An agent is an operation (or operator) that is responsible for activities, for the existence of an entity, and for the activities of other agents.

The three main classes are correlated and use the properties that are illustrated in Figure 1.

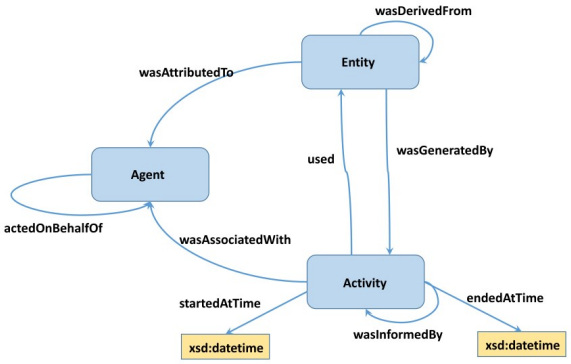

Figure 1. Starting Point classes and their properties. This figure is obtained from the PROV-O Ontology document in [11].

Extended classes and properties provide additional actions that may be used to relate classes in the Starting Point category. Qualified classes and properties provide elaborate information about relations using Starting Point and Expanded features.

B. Layouts

A layout is a main component in data provenance visualization; it improves the comprehension of a user. A layout depends on user requirements and the origin of provenance data [9]. A researcher often needs to see multiple layouts and determine the layout , which is the most meaningful and relevant. As an example, the hierarchical provenance visualization layout is a provenance graph that is separated into layers according to relationships; the most important relationships appear at the top layer and the final results appear at the bottom of visualization.

Figures 2, 3 and 4 shows customized layout examples. As mentioned earlier, in provenance visualization a circle denotes a process while a square denotes an artifact [15].

C. Use Cases

Provenance visualization concepts are difficult to develop. To have a general visualization technique, care must be taken to prevent the loss of connection between an application domain's specific requirements and data provenance interpretation.

The main purpose of this study is to develop several visualization concepts and evaluate them based on concrete requirements. The provenance community supports the development of different visualization techniques and evaluates them for possible application domains.

In the proposed research, to define scope, we outline potential application areas and their associated Interactive Provenance Visualization solution requirements.

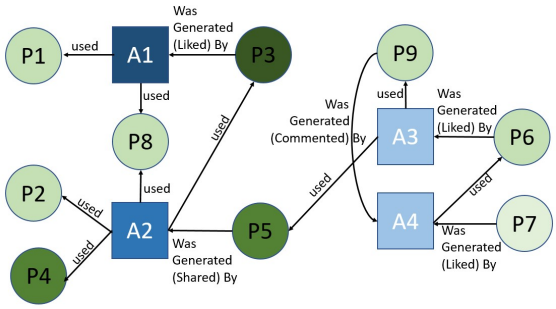

Figure 2. The time-to-complete process and artifacts

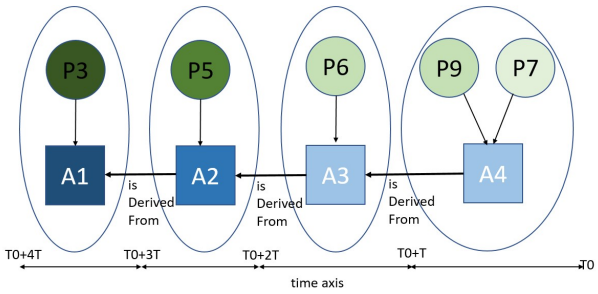

Figure 3. The time-to-complete individual processes and artifact relationships

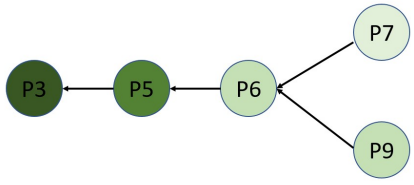

Figure 4. Lineage or process order

1) Social Media Provenance Visualization:

We have witnessed social media growing at an unprecedented rate in the last decade. Social media has inherent characteristics and pathways that enable nefarious activities, which could hinder its growth. Social media provide users with a platform to communicate with a large audience in a simple way. The unprecedented scale by, which users can communicate and the rate of information transfer is virtually unseen in conventional media like newspaper, radio, T.V. Social media data is user-generated, massive, distributed, noisy, dynamic, and unstructured-in-nature.

Social media data is primarily available in the form of an individual users' attributes, user-user connections (links), or user-generated content such as text, photos, and videos. The visualization requirements of age-based social life data from creation to archiving is one of the motivations of this work [10].

2) E-Science Provenance Visualization:

e-Science is computational, science-intensive data found in highly distributed network environments. Research into e-Science provenance focuses on the capturing, modeling, and storage of provenance data. In e-Science, considerably larger volumes of data with higher complexity exists because their systems are continuously running for extended periods to support scientific experimentation [8]. One of the motivations of this research is to find methods to improve visualization data provenance in the e-Science domain.

D. CYTOSCAPE

Cytoscape is an open source software platform that provides visualization tools using the interaction between networks and biologic pathways to integrate multiple networks that have expression profiles, annotations, and state data. Cytoscape was developed and designed for molecular biological research, but later became a general platform for complex visualization and analysis. Cytoscape's basic set of features include analysis, integration and visualization; they are provided by the core distribution of Cytoscape.

One of the strongest features of Cytoscape is that it allows the use of plugins to develop new visualization techniques. By supporting detailed and overlaying visualization with supplementary tools, Cytoscape is a suitable development environment for general provenance visualization. Provenance graph visualization that can interact with a Karma provenance server to extract provenance in XML form, can be created via a Cytoscape plugin [9].

IV. METHODOLOGY

In this study, temporal representation and stream data provenance visualization will be analyzed. We will research how to achieve provenance visualization of temporal provenance data by developing a set of visualization techniques and methodologies. Real-time visualization and offline visualization will be implemented. We apply developed methods on test case scenarios in social media domains.

Chen's provenance visualization steps will be used as a base for our test case scenarios [9].

A. Incremental Loading

Provenance data can be very large and dense. The lineage records or PROV-O annotations alone provide an opportunity to capture additional information about data execution or creation. To support visualization over large graphs, a system must be able to read XML-formatted provenance graphs with and without annotations. A KOMADU system will be the basis of this research. A KOMADU system is a data capture and visualization system. It was developed for scientific data provenance in the Data to Insight Center at Indiana University. KOMADU generates PROV-O compliant XML files that doesn't have annotations for Processes or Artifacts.

For scalability purposes, we will first investigate how to manage scalability in our developed solution. Provenance data features, i.e., lifecycle metadata on activities, will be reduced to create a reduced-dimension visualization. We will eliminate redundant data and generate provenance data partitions to group different visualization items to increase user's perception [9]. Data scalability for the proposed temporal representation process can be improved by using MapReduce programming paradigm [16].

B. Customized Layouts

Customized layouts developed in Chen's study include extending the hierarchical layout algorithm that sorts sibling nodes, grouping layout algorithms, creating concatenated string-embedded layout algorithms for provenance data like a history chain [9], which will be referenced in our study. Research on the improvement of existing layouts and the development of new customized layouts will be realized by considering user navigation models, HCI principles and provenance data nature.

The Bilkent University data visualization research group [13] has several projects on compound graph visualization based on several different customized layout displays. Their layout model handles the followings:

- levels of nesting
- inter-graph edges span multiple levels of nesting
- non-leaf nodes links in the nesting hierarchy

C. Visual Style

In data provenance visualization, Chen has created a default visual style for provenance graphs, using magenta for artifacts, different predefined colors for a different type of edges [9].

New visualization styles, according to HCI studies on related attributes like size, color, etc., and Gestalt rules parameters, are subject to the study. We will research how to achieve provenance visualization of temporal provenance data that is associated with different existing visualization techniques and methodologies. User attraction and perception towards the customized visualizations will be handled by user HCI evaluations and will be used to improve the visualization system. Gestalt rules will be examined to find adaptation-related principles to increase usability of the system [9].

D. Abstract View

Provenance relationships are complex graphs that overwhelm researchers. To eliminate and summarize provenance visualization, Chen's two approaches are:

- Clustering neighbor nodes
- Eliminating process/artifact

The process of clustering neighboring nodes into a single node was introduced using a plug-in in Cystoscape. This

approach is helpful while exploring a provenance graph and dealing with graphs that have many nodes [8].

In the second-phase, a process of elimination removes process nodes that connect two artifact nodes with a "was generated by" incoming edge and "used" outgoing edge. The process node is changed by a new "was derived from" edge. This process of graph abstraction and pruning removes any unnecessary info and limits the graph size

In this section, we will study new approaches for summarization and elimination processes of provenance data. Using MapReduce programming paradigm, data abstraction on the proposed temporal representation process can be enhanced.

E. Graph Comparison

To satisfy provenance analysis, graph comparison is a key process. In Chen's study, Direct Clustering Algorithm (DCA) is used to compare two provenance graphs by finding the best matched and unmatched nodes [9]. The DCA algorithm basically depends on the order of inputs. For example, B to A is not same as A to B. In this research, an improved DCA algorithm will be developed and implemented to compare graphs. The comparison of more than two graphs is also an important task, for reasons below.

- Provenance data can exist in a high-dimensional space. This causes graphs to have thousands of nodes and attributes, which makes clustering of such data tremendously difficult
- Difficult to locate both structural and nonstructural information and combine it into a single uniform attribute space

V. HUMAN MACHINE INTERACTION ASPECTS

In this study, we propose an improved data provenance visualization system that involve user in the visualization process by capturing and analyzing user's attention and perception level towards specific visualization types and methodologies.

A. User Involved Visualization

The proposed system involves the user; the user can improve usability and increase the system's effectiveness. The system will support user-based navigation by capturing a user's attraction to and perception of a specific set of layouts/methodologies. At first, the distinct existing layouts will be studied to visualize big provenance data. If necessary, depending on the characterization of the provenance data and user requirements, new customized layouts will be introduced.

B. HCI Model Construction

We will measure the user's perception and analyze the level of provenance visualization for a set of abstract layouts by capturing and analyzing a user's attention, attraction and perception level. After that, the correlation between the attraction/perception level and type of provenance data visualization will be determined to find suitable layouts/visual styles for specific kind of requirements or specific data domains. A layout/visual style HCI model will be constructed to verify this process. In the user perception level, the following 2 methods are proposed:

1. Questionnaire: a set of questions based on experimental protocols will be asked to analyze attraction/perception level
2. Eye Tracking: gaze parameters are used to extract perception information based on literature parameters.

VI. CONCLUSION AND FUTURE WORK

We have discussed provenance and techniques for provenance visualization. We are interested in providing an effective solution to visualize and analyze semantic provenance data by adhering to HCI principles. Efficient visualization helps users to analyze and understand data using valid evidence. Visualization makes complex data more accessible, understandable and usable.

According to provenance data interpretation requirements, the development of provenance visualization is difficult to create. One of the most enduring challenges is to maintain the relationship between the data and its application.

Our improved data provenance visualization system that places the user in the loop, captures and analyzes the user's attention and perception level while he/she reviews specific visualization types and methodologies suggested by the system. The system then analyzes this information to determine the best type of visualization. In this way, the user determines the visualization process autonomously.

The goal of the system is to help the user navigate exploration provenance visualization. A user's mind map typically determines what is going on in an experiment. A mind map model will be used to interact with the user and explore visualization activities.

The major contributions of this research include the following:

- We develop an effective user-navigated provenance visualization system. We will introduce a temporal provenance data visualization solution with aspects of HCI research. Our approach will be based on user preferences and perception levels after reviewing several visualization layouts and types of provenance visualization. To increase the effectiveness of visualization, a user's attraction/perception level will be captured and analyzed.
- Domain independent visualization: We will abide by the recommendations of W3C's PROV-O specifications for provenance representation. Since

 ISBN: 978-1-61208-558-6

PROV-O specifications are domain-independent, the representation does not provide domain-specific vocabulary for our case study domains (e.g. social media data). Ontologies that define related domains will be introduced for specific domain representations.

- Customized Visualization Layouts and methodologies: In this paper, to understand large-scale data provenance, new abstract layouts, based on various studies, will be developed to provide customized layouts to the user.
- A highly scalable and high-performance visualization system will be used in our test case scenarios. We will deal with critically large scale provenance data. To make our system scalable, we will use techniques like incremental loading and compression. The effectiveness of visualization, graph matching, and partitioning will be improved to provide faster query times in provenance graph data.
- Real-time support for the continuous and real-time analysis of Big Provenance data and stream data visualization will be provided via a series of stream processing techniques. User-assisted real-time visualization, anomaly detection, and root cause tracing will be analyzed in terms of provenance graphs.
- A platform-independent PROV-O Cystoscope plug-in will be developed to visualize provenance based on different layouts and visual style elements.

Our future work will be focusing on the details of the methodology that we briefly outlined in this discussion paper. Our work remains in applying our research to the use cases and conducting experiments based on our research methodology.

ACKNOWLEDGMENT

This study is being supported by the TUBITAK-3501-Career Development Program (CAREER) with the Project ID: 114E781. We would like to thank Yildiz Technical University Software Quality Laboratory for supporting this research and allowing us to use their computer facilities for this study. As always, we are grateful for the help of the extended team of our department.

REFERENCES

[1] B. Glavic, “Big Data Provenance: Challenges and Implications for Benchmarking”, WBDB 2012 - 2nd Workshop on Big Data Benchmarking, pp. 72–80.

[2] Y. L. Simmhan, B. Plale, and D. Gannon, “A Survey of Data Provenance Techniques”, Technical Report IUB-CS-TR618

[3] Oxford English Dictionary (OED), “The fact of coming from some particular source or quarter, source, derivation”, (2017, April 20) Retrieved from http://en.wikipedia.org/wiki/ Provenance.

[4] R. Hasan, R. Sion, and M. Winslett, M., “Preventing History Forgery with Secure Provenance”, ACM Transactions on Storage, 5(12), May 2009.

[5] E. Olshannikova, A. Ometov, Y. Koucheryavy, and T. Olsson, “Visualizing Big Data with Augmented and Virtual Reality: Challenges and Research Agenda”, Journal of Big Data, 2015.

[6] R. Tard´io, A. Mat´e, and J., Trujillo, “An Iterative Methodology for Big Data Management”,2015 IEEE International Conference on Big Data (Big Data), 2015, pp 545-550.

[7] J. Steele, and N. Illinsky, “Beautiful Visualization, Looking at data Through the Eyes of Experts”, O’Reilly Media, 2010.

[8] P. Chen, and B. A. Plale, “Big Data Provenance Analysis and Visualization”, IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 15th, 2015.

[9] P. Chen, B. Plale, Y. W. Cheah, D. Ghoskal, S. Jensen, and Y. Luo, “Visualization of Network Data Provenance”, 978-1-4673-2371-0/12/$31.00, IEEE, 2012.

[10] G. Barbier, Z. Feng, P. Gundecha, and H. Liu, “Provenance Data in Social Media” , Synthesis Lectures on Data Mining and Knowledge Discovery, 2013.

[11] T. Lebo, S. Sahoo, and D. McGuinness, “W3C PROV-O: The PROV Ontology Proposed Recommendation 12.03.2013”, Technical report, W3C, March 2013.

[12] M. Kunde, H. Bergmeyer and A. Schreiber, “Requirements for a Provenance Visualization Component”, IPAW, 2008.

[13] U. Dogruoz, E. Giral, A. Cetintas, A. Civril, and E. Demir, “A Layout Algorithm for Undirected Compound Graphs”, Information Sciences, 2009, pp 980-994.

[14] P. Chen, and B. Plale, “Visualizing Large Scale Scientific Data Provenance”, 2012 Super Computing Conference, High Performance Computing, Networking, Storage and Analysis (SCC) Workshop, USA, 2012.

[15] P. Chen, B. Plale, and M. S. Aktas, “Temporal Representation for Scientific Data Provenance”, 978-1-4673-4466-1/12/$31.00, IEEE, 2012, pp 1-8.

[16] J. Dean, and S. Ghemawat, “MapReduce: A Flexible Data Processing Tool.” Commun. ACM 53(1), (2010), pp 72-77.

 ISBN: 978-1-61208-558-6