Provenance: Current Directions and Future Challenges for Service Oriented Computing

Colin C. Venters, Paul Townend, Lydia Lau, Karim Djemame, Vania Dimitrova, Alison Marshall, and Jie Xu
University of Leeds
Leeds, United Kingdom
e-mail: {c.venters; p.m.townend; l.m.s.lau; k.djemame; v.g.dimitrova; a.marshall; j.xu} @leeds.ac.uk

Charlie Dibsdale and Nick Taylor
Rolls Royce PLC
P.O. BOX 31 Derby, United Kingdom
e-mail: {charlie.e.dibsdale; nick.taylor}@o-sys.com

Jim Austin, John McAvoy, Martyn Fletcher, and Stephen Hobson
Cybula Ltd.
Computer Science Building, York, United Kingdom
e-mail: {austin; mcavoy; martyn; stephen}@cybula.com

Abstract—Modern businesses increasingly depend heavily on information stored and processed in distributed, heterogeneous data sources and services to make critical, high-value business decisions. Service-oriented systems are dynamic in nature and are becoming even more complex systems of systems. In such systems, knowing how a data set was derived is of significant importance in determining its validity and reliability. To address this, a number of advocates and theorists postulate that provenance is critical to building trust in and the reuse of data and services that generated it as it provides evidence for data consumers to judge the integrity of the results. This paper provides an overview of current provenance research with an emphasis on its application in the domain of service-oriented computing. The goal of this paper is not to provide an exhaustive survey of the provenance literature but rather to highlight emerging areas, themes and issues related to the use of provenance as a mechanism for improving trust in data utilized in distributed computing environments.

Keywords-component; provenance; trust; service-oriented computing.

I. INTRODUCTION

The emergence of service-oriented computing represents a paradigm shift in the way that software and hardware are not only developed but also utilized by end-users [1]. The adoption of distributed computing technology in the scientific domain in particular, which relies heavily on the dissemination and exchange of data sets, has increased the need for a mechanism to engender trust in the results of scientific data sets [2, 3]. It is argued that the ability to reproduce results is one of the cornerstones of the scientific method and the inability to do so can significantly undermine the ‘confidence’ in the results [4]. Tsai et. al., suggest that due to the dynamic nature of service-oriented systems, it is critical to consider not only the security and integrity of the data but also its trustworthiness. To address this, a number of advocates and theorists postulate that provenance is critical to building trust in and the reuse of data and services that generated it as it provides evidence for data consumers to judge the reliability and validity of the results. Golbeck and Hendler [5] and Davidson and Freire [6] suggest that provenance is of critical importance for scientific workflow systems, as it allows users to share and verify data, repeat experiments, and discover dependencies. This is a view shared by Moreau et. al., [7] who argue that provenance is crucial for end users of data generated by computer-based systems in order that the data’s provenance may be retrieved, analyzed and reasoned over, advocating that all computer-based systems should be transformed so that they become provenance-aware. In addition, Cheney et. al., [8] argue that without pervasive adoption of provenance technology, critical systems and networks can be exposed to serious vulnerabilities, which can result in what they term provenance failures; a negative outcome caused by a failure to record, manage, interpret, or control access to data and its provenance.

The decision to trust is based on evidence to believe or to be confident in someone or something [9]. The Oxford English Dictionary [10] defines trust as the ‘confidence in or reliance
on some quality or attribute of a person or thing, or the truth of a statement.’ Related to the concept of trust are the terms distrust and uncertainty where the former is defined as ‘the absence or want of trust; lack of confidence, faith, or reliance; doubt, suspicion’ [11] and the latter is defined as ‘the state or character of being uncertain in mind; a state of doubt; want of assurance or confidence; hesitation, irresolution’ [12]. A common theme running through all of the definitions is the concept of confidence which is defined as ‘the belief in, trustworthiness, or reliability of a person or thing’ [13]. Mayer, Davis, and Schoorman [14] state that trust is crucial wherever risk, uncertainty, or interdependence exist. To compound the issue further, Huang and Fox [15] suggest that trust is context specific and that the context can have very complex structures. A wealth of literature has been devoted to the subject of trust from a variety of different discipline perspectives including computer science, business management, sociology and psychology and readers are directed to these for comprehensive reviews of the topic [16, 17]. This paper is focuses on the concept of provenance as a mechanism for engendering trust in data used and generated in a service-oriented computing environment and is organized as follows. Section II examines the concept of provenance. Section III highlights a number of theoretical models of trust. Section IV considers emerging approaches to develop trusted architectures and frameworks. Section V highlights attempts to understand the relationship between provenance and data consumers. Section VI outlines issues, challenges and future directions.

II. PROVENANCE

The concept of provenance is well understood in the study of fine art where it refers to the documented history of an art object and plays a major factor in establishing authenticity. Derived from the French word ‘provenir’, to come from, provenance means the origin or the source of a ‘thing’. The Oxford English Dictionary [18] defines provenance as:

i. ‘The fact of coming from some particular source or quarter; origin, derivation’;

Data origin, its derivation in terms of how it has been transformed from the original source and the recording of this process are fundamental to providing provenance data. In addition, provenance has also been defined in a number of ways depending on the context in which it is being applied or the application domain in which it is being used such as scientific workflows, geographical information systems, GRID and Cloud computing infrastructures [6, 19, 20]. Ram and Liu [21] highlight that provenance not only provides the background information about data but enables it to be interpreted and used correctly within context. Nevertheless, Simmhan, Plale and Gannon [22] argue that regardless of the context or application domain that provenance is being used in there are two important features: the ancestral data and the process of the transformations by which the data product evolved.

Groth, Luck and Moreau [23] also distinguish between two types of provenance that exist in a service-oriented architecture: interaction and actor provenance. Interaction provenance is the documentation of interactions between actors that led to the data. They argue that interactions in a SOA are fundamentally a client invoking a service. As a result, interaction provenance can be obtained by recording the inputs and outputs of the various services involved in generating a result. Actor provenance is documentation that can only be provided by a particular actor pertaining to the process that led to the data; actors are either client’s who invoke services or are services that receive invocations and return results. Similarly, Huang and Fox [24] extend the basic definition of provenance postulating that it is an approach not only to determining the origin and validity of data but also a means of modeling and maintaining data sources, their dependences, and their trust structures. The basic unit of information is a proposition, which has a truth value of ‘true’, ‘false’ or ‘unknown’. From this they suggest that there are four levels of provenance:
- Static (Level 1);
- Dynamic (Level 2);
- Uncertainty-oriented (Level 3);
- Judgment-based (Level 4).

where static provenance is provenance related to fixed and unchanging data; dynamic provenance is concerned with the validity of data over time; uncertainty-oriented provenance considers the validity of data which is inherently uncertain; and judgment-based provenance focuses on social processes required to support knowledge provenance. They argue that at Levels 1 and 2, a creator or provider of information is either ‘trusted’ or ‘distrusted’ although they acknowledge that a data consumer is highly likely to trust information providers by degrees rather than completely ‘trust’ or ‘distrust’ them. da Silva, McGuinness and McCool [25] make a distinction between the terms knowledge provenance and knowledge process information where the former includes source meta-data, which is a description of the origin of a piece of knowledge, and the latter is a description of the reasoning process used to generate the answer. They also draw a distinction between the term data provenance, which they argue can be viewed as the analog to knowledge provenance aimed at the database community that typically includes a description of the origin of the information and the process by which it arrived in the database. By their definition, knowledge provenance is essentially the same except that it includes proof-like information about the process by which knowledge arrives in the knowledge base, which may include extensive reasoning used to generate deductive closure information. In this sense, knowledge provenance broadens the notion of data derivation that can be performed before data is inserted into a database or after data is retrieved from a database. This is a view supported by Vasquez, Gomadam, and Patterson [26] who suggest that in addition to provenance being focused merely on the history, processes and the transformations of data sources, it should provide a qualitative and quantitative metric to analyze the quality and the dependability of the data, based on the data consumers trust of the source of creation and the sources that were responsible for its modification. Clifford et. al., [20] argue that there are two distinct forms of provenance:

- Prospective;
- Retrospective.

Prospective provenance captures the specification of a computational task, which corresponds to the steps that need to be followed to generate a data product or class of data products. Retrospective provenance captures the steps that were executed as well as information about the execution environment used to derive a specific data product. Tsai et. al., [27] consider the level of granularity of the provenance data to be captured, arguing that it is not necessary to collect all. As a result, they suggest that data provenance systems need to classify data into different provenance categories and proposed a classification of provenance:

- Maximum;
- Time-based;
- Event-based;
- Actor-related;
- Scenario-based;
- Minimum;
- No provenance.

where minimum provenance is priority data and includes a complete history of transformations and processes; time-based provenance covers specific periods of time; event-based provenance is associated with specific events; actor-related provenance is related to specific actors or agents which can either by human or systems that initiate actions; scenario-based provenance is related to specific scenarios; minimum provenance is for data that is considered of low priority where only certain aspects of the data will be tracked; and no provenance is non-critical data that has no value for data provenance tracking.

III. Trust Models & Provenance

A number of models have been developed for computing trust in the provenance data captured by a system. Dai et. al., [28] state that without high-assurance data integrity, information extracted from available data cannot be trusted. To
address this, they propose a provenance data trust model, which takes into account four factors that affect trustworthiness and assign trust scores to the data and the data providers. Trust scores range from 0 to 1 with higher scores indicating higher levels of trust. The trust score is computed by taking into account four factors:

- Data similarity;
- Path similarity;
- Data conflict;
- Data deduction.

where data similarity refers to the ‘likeness’ of different items \( r \) as defined as:

\[
sim(r) = e^{-\frac{\phi_c}{N_c}} \tag{1}
\]

where \( N_c \) is the number of records in cluster \( C \) and \( \phi_c \) is the diameter of cluster \( C \). The purpose of clustering is to eliminate minor errors such as typos. The effect of data similarity \( \sim(r) \) on trust scores is determined by the number of items in the same cluster \( C \) and the size of the cluster. Path similarity refers to how similar the paths followed by two data items from the sources of the destination. Path similarity is important in establishing the independence of two or more data items:

\[
\text{pathsim}(r_1,r_2) = \frac{\max\{|P_1|,|P_2|\} - \text{Idist}}{\max\{|P_1|,|P_2|\}} \tag{2}
\]

where \( r_1 \) and \( r_2 \) are two data items, \( P_1 \) and \( P_2 \) are their paths, \( \max\{|P_1|,|P_2|\} \) is the maximum number of identifiers in the two sequences, and \( \text{Idist} \) is the difference between identifiers. Data conflict refers to inconsistent descriptions or information about the same entity or event. In order to determine if data conflicts, data users are required to define the exact meaning of conflict, which they call prior knowledge. They also suggest that the trustworthiness of a knowledge item is dependent on the trustworthiness of information used to generate it and the trustworthiness of parties that handle it. Data deduction measures the effect of this process and can be defined as:

\[
\text{Ded}_k(t(a), t(r_1), \ldots, t(r_n)) \tag{3}
\]

where \( a \) is an intermediate agent, \( k \) is a knowledge item generated by \( a \) based in items \( r_1, \ldots, r_n \). The data deduction of \( k \) indicates the impact of the trustworthiness of \( r_1, \ldots, r_n \) and \( a \) on the trustworthiness of \( k \). The effect of trustworthiness of a source provider or an agent on its resulting data can be obtained using a weighted function to compute the trust score of the output information by taking into account the actions it took on this data; types of actions include passing data to another agent or data user or producing a knowledge item by combining the input information with some local knowledge:

\[
t(k) = \frac{w_i \cdot t(a) + \sum_{j=1}^{n} t(r_j)}{2} \tag{4}
\]

where \( w_i \) is a parameter based on the operation the intermediate agent takes and its impact in the trustworthiness of knowledge \( k \).

Huang and Fox [15] propose a trust judgment model for data provenance as an approach to determining the origin and validity of data by means of modeling and maintaining information sources, information dependencies and trust structures from a social networking perspective where people will accept the opinion of individuals close to them. They argue that minimum, maximum and weighted average methods are too simplistic for representing trust judgments and introduce aggregated compatibility and aggregated consistency, which are defined formally as:

\[
c(x_1,x_2) = \begin{cases} 
1 - d(x_1,x_2) / r, & d(x_1,x_2) \leq r \\
0, & \text{otherwise} 
\end{cases} \tag{5}
\]

where \( d(x_1,x_2) = |x_1 - x_2| \) and \( r=0.5 \). This proposed method attempts to find a solution that is most compatible to all friends’ opinions by a weighted voting process. Underlying the aggregation method is that the trustor request friends one-by-
one until a set of consistent opinions are found. Consistency rate is defined as:

$$\tau = \frac{\left( \sum_{b_i \in S} td(a, b_i) \right)}{\left( \sum_{b_j \in S} td(a, b_j) \right)}$$

(6)

In the aggregation process, trustor $a$ request friends $b_1$, $b_2$, $b_n$ one-by-one until the consistency rate $\tau$ is not less than a predefined threshold $\tau_0$ where $S$ is the set of requested friends who have answered their degrees of trust in trustee $c$, $S_i$ is the largest subset of friends whose opinions are consistent with each other, and $td(a, b_k)$ is the voting weight.

Chapman, Blaustein and Elsaeser [29] argue that current proposal to assess trust based solely on provenance is insufficient for rigorous decision making and proposed a provenance-based belief model that supported normative inference and employs propositional semantics. The underlying premise to the model is that ‘belief’ is not static and is influenced by evidence based on an individuals subjective probability that a proposition is true:

$$p(C \mid E) = \frac{p(C \wedge E)}{p(E)}$$

(7)

where $C$ and $E$ are propositions and $p(C \mid E)$ is the conditional probability of $C$ given $E$. They suggest that main requirement to improve believability is that the provenance system captures the accuracy of the data sources. As a result, the computations required to support decision-making can be efficiently calculated using off-the-shelf Bayesian network algorithms. The model relies on information in the provenance store about how that information is propagated through a lineage graph and how accurate each source is. This information is then used to compute a belief in the derived data items.

While the model proposed by Dai et. al., [28] has been evaluated to measures its performance in calculating trust scores on a range of datasets none of the models have been evaluated with end users to assess the effectiveness of their approaches.

IV. TRUSTED ARCHITECTURES, FRAMEWORKS & PROVENANCE

In addition to a number of theoretical models for calculating trust in various ways, a number of architectures and frameworks have also been proposed. Bertino, Dai and Kantarcioglu [30] proposed an architectural framework based on policy-driven data trustworthiness composed of four elements:

- Confidence values;
- Confidence policy;
- Computation of confidence values;
- A set of strategies for incrementing the confidence at query processing time.

A confidence value is a numeric value ranging from 0-1 and indicates the trustworthiness of the data. The confidence policy indicates a confidence level required for certain data used in a specific task. The computation of the confidence values are computed from the confidence values of each data item using lineage propagation techniques based on the work of Dalvi and Suciu [31]. They argue that the notion of a confidence policy is the novelty of their proposed approach as it allows a confidence level to be specified for a given data item in a certain task by a certain subject. However, they acknowledge that a critical issue in enforcing confidence policies is that end users may not receive enough data to make an informed decision and that the development of their approach requires addressing several issues including cost models, heuristics for confidence increases, and query-based data validation strategies. Tsai et. al., [32] propose a dynamic policy-based framework for the agile collection and classification of data provenance in a SOA. The framework allows two data collection strategies: actor-based or message-based. The actor-based strategy acts as a monitoring service for tracking communicated data. The message-based strategy carries its provenance, which can be updated at runtime by the services that interact with it using digital signatures. With both approaches they acknowledge that there is a
significant processing overhead. The data classification module includes two components: Data Classification Service (DCS) and a Data Provenance Manager (DPM). They argue that not all data needs to be recorded and analyzed. As a result, the DCD classifies provenance into a range of different categories outlined in Section III of this paper. Using the DCS, the DPM can decide which data to track in the SOA system. The framework also provides a number of analysis services, which can be used to enhance the system: Security Policy Checking Service (SPCS) which allows security policy editing and runtime enforcement for individual and integrated services; an Integrity Estimation Service (IES) which provides a structured-based integrity estimation model for individual and integrated services; and a Reliability Analysis Service (RAS) which provides reliability analysis of the SOA system.

Lyle and Martin [33] argue that if provenance records are not protected, they cannot provide convincing evidence of the quality of the data itself. To address this they proposed an architectural framework for provenance based on trusted computing technology and standards; a hardware-based method where the integrity of data storage and program execution can be assessed and enforced remotely by identifying a complete ‘chain of trust’ in all the hardware and software that has been used [17]. Centered around the Trusted Platform Module (TPM), a basic implementation in a PC implementation connecting the CPU to the TPM chi. This provides isolated storage of encryption keys and 16 Platform Configuration Registers (PCR) which can be used to hold integrity measurements in the form of 20byte SHA-1 hash values. They argue that integrity measurements are applicable to two requirements of provenance: identifying which results have been affected by known software or hardware errors, and for accurately reproducing results. They suggest that there are a number of advantages to utilizing trusted computing techniques including secure storage, traceability of interactions, and tamper proof. However the acknowledge there are a number of challenges including system performance, the reliance on a public key infrastructure, and the different directions of researchers in the field of trusted computing. Lu et. al [34] proposed a secure provenance scheme based on the bilinear pairing techniques, which employs provable security in the standard model as an alternative approach to trusted computing. They argue that secure provenance should satisfy the two basic requirements:

- Unforgeability;
- Conditional privacy preservation.

Their scheme for secure provenance includes the following parts:

- Setup;
- KGen;
- AnonyAuth;
- AuthAccess;
- ProveTrack.

where Setup is a probabilistic algorithm run by the system model; KGen is key generation algorithm run by the system model which is can either be probabilistic or deterministic; AnonyAuth is an anonymous authentication algorithm run between the user $U_i$ and the service provider $SP$; AuthAccess is an authorized access algorithm between user $U_i$ and the service provider $SP$ to achieve fine-grained authorized access after a successful anonymous authentication; and ProveTrack is a provenance tracking algorithm which tracks the real identify of the user.

An alternative approach to that of policy-driven architectures and the utilization of trusted computing platforms is the use of semantic web technology and intelligent agents to make associations. Ding et. al., [35] proposed a provenance and trust aware inference framework which employs an ontology for representing associations of trust and provenance, a mechanisms to evaluate the trustworthiness of semantic association or any collection of statements obtained from multiple sources, and a trust based knowledge expansion mechanism that
incrementally outsources knowledge from peers to bound the size of knowledge base for inference. Their approach is based on the assumption that provenance and trust are associations amongst statements created and used by agents, which can be serialized in RDF. A set of statements are trusted by an agent only when they are consistent and believed to be true. Here they draw a distinction between trust and belief where the former is related to another agents trust in another agent’s knowledge and the latter is related to an agents trust in a set of statements. Trust confidence ranges from 0-1 where 0=distrust, 0.5=ignorance, and 1=trust and belief confidence ranges from -1,1 where -1=disbelief, 0=non-belief and 1=belief. They argue that their approach provides a multi-granularity method to group statements but has yet to be demonstrated empirically. A similar approach is proposed by Xu and Wang [36] who presented a data provenance architecture based on semantic web services for managing data provenance using OWL, which can record and query semantic provenance information.

Groth, Luck and Moreau [23] suggest that a SOA can be broken down into two types of actors: clients who invoke services and services that receive invocations and return results. Within this context, they introduce a provenance store as a third type of actor, which they believe is the key to fulfilling the non-functional requirements that a provenance system should support:

- Verifiability;
- Accountability;
- Reproducibility;
- Preservation;
- Scalability;
- Generality;
- Customizability.

From this they proposed a conceptual architecture for recording provenance and an independent protocol for recording provenance data in the context of service-oriented architectures, Provenance Recording Protocol (PReP). PReP is a four-phase protocol consisting of:

- Negotiation;
- Invocation;
- Recording;
- Termination.

where negotiation is the process by which a client and service agree on which provenance store to use, invocation is related to service invocation, recording is the key phase of the protocol capturing interaction provenance, and termination is the final phase of the protocol which is invoked when the provenance store has received all expected messages from both the client and the service. Vasquez, Gomadam, and Patterson [26] argue that provenance information, both at the data and functional levels, is an important metric in measuring the quality of service. However, they suggest that current SOA infrastructure lacks a framework to capture this information. To address this they propose a framework to capture provenance information in web services, which includes a mechanism for discovering services based on their data and functional provenance information. Chen et al., [38] suggest that there is no commonly agreed conception of provenance in the context of service-oriented computing. To explore this they designed and implemented a prototype system to in order to illustrates the various facets of provenance that are integral to the functionality of service-oriented computing and in the process identified some of the issues underlying the design of a generic provenance system. Muniswamy-Reddy, Macko, and Seltzer [39] consider the problem from a Cloud computing perspective and identify the properties of provenance that enable its utility in a Cloud computing environment. They propose three protocols for maintaining provenance in current cloud stores. They suggest that there are four primary properties of provenance systems:

- Data coupling;
- Multi-object causal ordering;
- Data independent persistence;
- Efficient querying.
V. PROVENANCE AND DATA CONSUMERS

A number of theoretical provenance models, frameworks and architectures have been proposed for building trust in data. However, end-users have not been part of the validation process to assess whether these approaches improve trust. An extremely small number of studies have investigated the relationship between provenance and trust from an end-users perspective [40, 41]. These two studies suggest that provenance, encoded as metadata, is often the only mechanism in place to allow end-users to assess that the data they are interacting with is fit for use and can be trusted. In a recent paper, Donaldson and Fear [42] argued that there is little empirical evidence to support the premise that provenance improves trust and reuse of data. To investigate this, they conducted seventeen interviews with users of ProteomeCommon.org; a major data repository in the field of proteomics research. The results of their investigation suggest that while provenance does help end-users gauge the trustworthiness of data and builds confidence in its reuse, provenance also needs to be accompanied by other types of information including the original data, associated data quality information, and information regarding the authors reputation.

VI. ISSUES, CHALLENGES, & FUTURE DIRECTIONS

It is argued that provenance can increase trust in heterogeneous data and services. This paper provided an overview of the literature related to the relationship between trust and provenance with a focus on distributed and service oriented computing. Current directions and approaches have focused on the development of models, frameworks, and architectures, which it is suggested can be utilized for determining trust in this computing environment. How this can be best achieved is still largely unclear as the approaches lack any empirical evaluation and are purely theoretical. Emerging evidence suggests that understanding trust from an end-users perspective is essential and that they should play a pivotal role in the validation process to assess whether these approaches improve trust.

An area that has largely been ignored is related to provenance attributes. The Open Provenance Model provides a representation of processes that have led to data being produced or transformed into a new state [43]. While provenance data describes the history of an object and can be represented in different formats there is a need to investigate the attributes that are being captured. As far as the authors can determine no generic attribute model exits and it is unclear whether such an approach is feasible given the diverse nature of data in different domains. However, some commonality will exist between domains. Future work will focus on understanding the relationship between decision theory, choice models and models of trust, information quality attributes and provenance, and issues related to provenance access, storage, and persistence.

REFERENCES


Vasquez, I., K. Gomadam, and S. Patterson, Framework for representing provenance for web services and processes.


Dai, C., et al., Trust evaluation of data provenance. 2008, Center for Education and Research Information Assurance and Security, Purdue University.


