Evaluating Provenance-based Trust for Scientific Workflows

Shrija Rajbhandari, Ian Wootten, Ali Shaikh Ali and Omer F. Rana
School of Computer Science
Cardiff University, UK
{S.Rajbhandari, I.M.Wootten, Ali.Shaikhali, O.F.Rana}@cs.cardiff.ac.uk

Abstract

Provenance is the documentation concerning the origin of a result generated by a process, and provides explanations about who, how, what resources were used in a process, and the processing steps that occurred to produce the result. Such provenance information is important to improve a scientist's ability to judge and place certain amount of trust on the generated data. We illustrate how provenance information associated with a workflow can be used to evaluate trust. This work is based on several use cases from a Bio-Diversity application. We also propose a simple architecture to illustrate our trust framework.

1. Introduction

Computational scientists in recent years have been increasingly relying on distributed computing technologies as an essential part of their everyday research. Although the concept of sharing distributed resources amongst geographically distributed groups is not new, increasing advancement in Service Oriented Architectures (SOA) in Grid and Web Services makes the vision more realistic. Amongst the consequences of the progress toward SOA in scientific domain is an increased emphasis on provenance data, and the need for mechanisms to acquire, use and manage such data [7] [4]. A key focus of this paper is to understand how much trust a scientist can place in results that have been produced through a distributed workflow session. Generally, such a session will involve use of resources that are not owned by a single site or administrator.

SOA has made feasible the use of distributed and heterogeneous resources for scientific domains such as Bio-Diversity [5]. It is apparent that with time, there will be an increase in the quantity of such resources available to a scientist. As resources such as algorithm implementations and data increase, so does the variety in their quality, and thereby the level of trust that can be placed in them. Many research scientists will make use of such resources in their experimental workflow, and at some point they may wish to share the produced results with their fellow researchers. Provenance is described as the documentation of a process (workflow) that led to a particular result, and will specify who, how, what and which resources were used in the process. Such provenance information along with the result improves a user's ability to judge the validity of a result. Although provenance provides justification for results, the notion of how much trust can be placed on the result is completely implicit – to the extent that such concern has not been fully addressed in existing workflow systems. The proposed framework outlines the ability for the user to use the provenance of result and compute the degree of trust the user places on that result. The framework is developed with the help of motivating use cases from a Bio-Diversity workflow scenario.

Related work on trust models is presented in Section 2. In Section 3 a brief overview of the Bio-Diversity workflow and identified use cases are presented. Section 4 presents the trust architecture. Section 5 provides discussions on the decision tree that is incorporated and used to formulate our trust framework and Section 5.2 illustrate how trust is calculated.

2. Related Work in Trust

Trust issues in Grid computing are viewed from two perspective: (i) security aspect dealing with confidentiality and authentication of the parties involved (digital signature, private keys and credentials); (ii) subjective aspect like credibility of source/service or actor – based on work in Peer-2-Peer systems [3] [6]. We focus on aspect (ii).

2.1. Trust

There are two main approaches to trust introduced in literature for evaluating credibility issue. Firstly, to allow actors to trust each other, there is a need to endow them with the ability to reason about the reliability, or honesty of their counterparts. This ability is captured through trust models.

Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID'06)
0-7695-2585-7/06 $20.00 © 2006 IEEE

The latter aims to enable actors to calculate the amount of trust they can place in their interaction partners. A high degree of trust in an actor would mean it is likely to be chosen as an interaction partner. Hence, trust models aim to guide an agent in deciding how, when, and who to interact with. However, in order to do so, trust models initially require actors to gather some knowledge about the characteristics of their counterparts. Based on existing work, this may be achieved as follows:

1. **A presumption drawn from the actor's own experience:** Trust is computed as a rating of the level of performance of the actor. The actor's performance is assessed over multiple interactions checking how good and consistent it is at doing what it says it does. To this end, Witkowski et al. [11] propose a model whereby the trust in an actor is calculated based on its performance in past interactions. Similarly, Sabater et al. [9] propose a similar model but do not just limit the overall performance to the actor's direct perception, but they also evaluate its behavior with other actors in the system.

2. **Information gathered from other actors:** Trust in this approach is drawn indirectly from recommendations provided by others. As the recommendations could be unreliable, the actor must be able to reason about the recommendations gathered from the other actors. The latter is achieved in different ways: (1) deploying rules to enable the actors to decide which other actors' recommendation they trust more [1]; (2) weighting the recommendation by the trust the actor has in the recommender – EigenTrust [6] and PageRank [8] are examples of this approach.

3. **Socio-Cognitive Trust:** Trust is drawn by characterizing the known motivations of the other actors. This involves forming coherent beliefs about different characteristics of these actors and reasoning about these beliefs in order to decide how much trust should be put in them [3].

Refer to [10] for more details on trust and reputation approaches. Such existing work is either the subjective information or objective opinions formed on the basis of factual evidence or recommendation by arbitrating authorities; or the different combinations of all. Our framework differs from such models as the concern is towards trustworthiness of an outcome that is the result of a scientific experiment – performed in a distributed, service oriented environment. We recognize the importance of provenance data and exploit this in our trust framework. Thus, apart from provenance data providing the explanation about how a result came to be, it also provide a way to formulate trustworthiness to place some judgement on the result. As a whole, our trust framework provides a basis for subjective opinions on the trust that may be placed on a particular outcome of a workflow process, and the actors and data involved in the process. These factors thus lead to the measure of "result trustworthiness".

3. Bioclimatic Modelling

We present a bioclimatic modelling workflow scenario from the BioDiversityWorld (BDW) project at Cardiff [5]. [12] describes the workflow in more detail. In figure 1, given a set of species, climate and locality data the model allows predictions of anticipated effects of climate change upon biodiversity by projecting this upon an image of the world map. Below we present some use cases based on the BDW project.

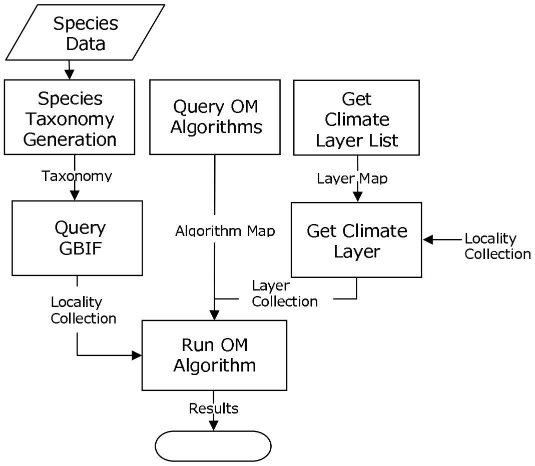

Figure 1. Bioclimatic Modelling within BDW

Use Case 1: Workflow and Result Accuracy: *A bioinformatitian runs the bioclimatic modelling experiment presented in figure 1. Later, another bioinformatition B wants to use the result of this experiment to do comparative studies. B determines whether the resultant projection image is one which is accurate and can be relied upon.*

In this case, B could simply view the order in which the experiment was carried out and make a judgement on the process to place a degree of trust on the result. It can be speculated that after viewing the services/algorithm involved and the way the process was constructed, B can also determine a certain level of result accuracy. Thus, given that each accuracy at each intermediate stage of the workflow is known, B is able to determine the overall accuracy of the projection image.

Use Case 2: Execution Bottleneck *A bioinformatition B, downloads some locality information for a particular species from the GBIF database and runs the bioclimatic* *modelling experiment upon it. A later run of the same process yields an overall execution time which is far greater than the earlier run. B determines which of the processes involved caused the extension in execution time in this experiment.*

Through inspection of the execution times maintained by an actor (part of actor provenance data) managing a service, B is able to determine which service(s) caused the increased time for this particular process. Thus, any major increase in the total execution time would make B conclude that the service(s) that are causing this have shortcomings and could not be completely trusted.

Use Case 3: Input Parameter Requirement: *A bioinformatition runs the BDW experiment presented in figure 1. Later a reviewer assessing the workflow determines that the climate data that was used includes attributes such as temperature and rainfall. Based on this information, the reviewer could conclude that in order for her to trust the result (so as to serve a particular purpose), she requires humidity data also.*

Although the reviewer places a lesser degree of trust on the input data, the process as a whole could still be meaningful – to place some trust on its result.

Use Case 4: Data Consistency: *A bioinformatition runs the BDW experiment explained in figure 1. A reviewer wants to confirm whether the data that is passed between the services are consistent in terms of their type and value, for example whether the locality data output from the GBIF query match the input received by the OM Algorithm within the workflow.*

By examining provenance data, in this case the I/O of each actor for that particular process run, the reviewer is able to determine if a data an actor has generated is the same as the data received by another actor during their interaction.

Use Case 5: Data Schema Completeness: *A bioinformatition runs the BDW experiment explained in figure 1. Later a reviewer determines if all the data instances generated or consumed by the actors involved in this process run are complete in terms of its current (updated) schema so that it can be relied upon.*

Investigating a particular instance from recorded provenance data and comparing this with its current schema, the reviewer can check the presence of all the elements in the schema instance. This assume that there are predefined schemas for inputs and outputs for each node (actors) within the workflow. Note that the schema of data can change over time to reflect any changes or updates in the algorithms/actors which consumes and generates data. Thus, such validation is crucial to place an overall degree of trust on the result data.

Use Case 6: Data Updates: *In the BDW experiment shown in figure 1, GBIF database is queried for a particular species to get the locality information. Over time, the data used for the experiment might be updated or corrected in the database. This updating will result in making new corrected and/or updated data available for use. In such a case, in order for a reviewer to trust the locality data used for this experiment, there is a requirement that the update frequency of the data source (GBIF database) be defined.*

By investigating the update frequency information of the GBIF database present in the provenance record of this particular process run, the reviewer can place a degree of trust on the locality data used. The trust on the locality data would vary depending on when the workflow is reviewed and the frequency of updates. For example, if the workflow was reviewed following a recent update, locality data used in the workflow would not be completely trusted.

4. Provenance Trust Framework

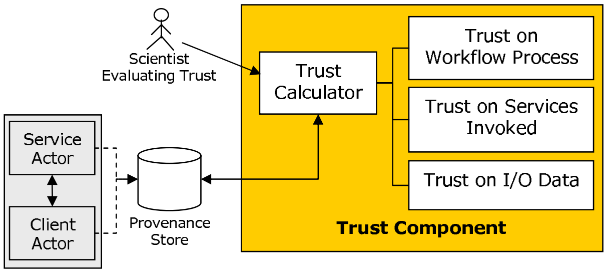

Figure 2. Trust Architecture

We focus on generating a trust measure for data that is an outcome of a workflow. This is achieved by combining different types of trust (process, service and data) to provide a "trust assessment" (see figure 2). The use cases in section 3 are analyzed in detail to identify the three main trust types (identified above), forming the core of our trust framework. It is assumed that both process and actor provenance are recorded in a repository which is refereed as Provenance Store.

Our trust framework consists of co-dependent parts shown in figure 2. The trust calculator allows users to pose queries to the Provenance Store for retrieving provenance data of past workflows, for example biodiversity experiments. A user may query data for each stage of the experiment. This decision process is described later in section 5 that incorporate the three stages of trust assessment. Provenance information utilized for trust evaluation may be categorized as follows: (1) Process Provenance: corresponds to the steps involved in the workflow that lead to a result. It also include the inputs and outputs for each actor involved in the process; (2) Actor Provenance: records the state of the actors involved in a particular workflow instance.

4.1. Process Trust

Process trust represents the confidence in the sequence of steps carried out to perform a task, for example a workflow. By evaluating the steps in a workflow, a reviewer can judge its soundness and decide on whether or not to trust the workflow. As mentioned in use case 1 in Section 3 – this may be a subjective measure. Hence, even if the workflow is sound, the reviewer still might not trust the result. This evaluation can be based on the provenance information about the algorithm or services used in the workflow.

4.2. Service Trust

The notion of service trust refers to the trust placed on the individual actors involved in the workflow. Based on existing work, generally the trustworthiness of actors is influenced by users' perceptions and recommendations from others; but here we focus on the behavior history of an actor for an objective indication of trustworthiness. In BDW scenario, the execution time of services is relevant for the bioinformatition, so that the services can be relied upon for delivering messages within some maximum time bound. Thus, using the previous execution times (actor provenance), one can estimate "reliability" (likelihood of completion) of a service with reference to execution time.

We propose such estimation to be generated through the evidence of the actor's past behaviors (actor provenance) by using a method of probability theory. We use beta probability distribution [2] that is useful for modelling random probabilities and proportions, particularly in the context of Bayesian analysis. In particular, the Bayesian theory uses standard beta distributions to model posterior probability estimates of observed binary events with two possible outcomes. We choose the beta function that takes the integer number of past observations with two possible outcomes: reliable (below execution time threshold) or unreliable with reference to past execution times of an actor to estimate the probability of "reliability". Using beta function allows us to predict the maximum uncertain probability (peak of the distribution, see figure 3) with which the actor can be relied upon by using the actor provenance (execution time).

The Beta Density Function: The Beta Distribution is a continuous probability distribution with the probability density function defined on the interval [0, 1]. Beta distribution is defined in terms of parameters α and β. A continuous random variable has a beta distribution with parameters α and β where,($\alpha, \beta >$0), its density function $f(x|\alpha, \beta)$ can be expressed as;

$$f(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B[\alpha,\beta]} \tag{1}$$

Where, (0$\leq x \leq$1) and $B[\alpha,\beta]$ is the beta function with parameter α and β which is given as;

$$B[\alpha,\beta] = \int_0^1 x^{\alpha-1}(x-1)^{\beta-1}dx \tag{2}$$

We use the implementation of Eq.(1) for calculating the probability distribution for an actor with evidence observed over time. For example, availability of an actor would have a set of boolean values gathered over a period (such actor provenance metric is recorded by a client about a service for that interaction). The boolean values are represented as the two parameters α and β respectively in Eq.(1) for calculating probability distribution of the actor being available in future. Out of 14 interactions, if the actor was available in 10 instances and unavailable in 4 instances, then, $f(availability|10,4) = 0.75$. This distribution is plotted in figure 3 which expresses the probability that the actor will be available in the future. Here, we use value 0.75 which is the uncertain maximum probability and denoted as $Umax(p)_{availability}$. It could be supposed that the regularity of the actor's availability in future is rather uncertain and that the most likely value is 0.75.

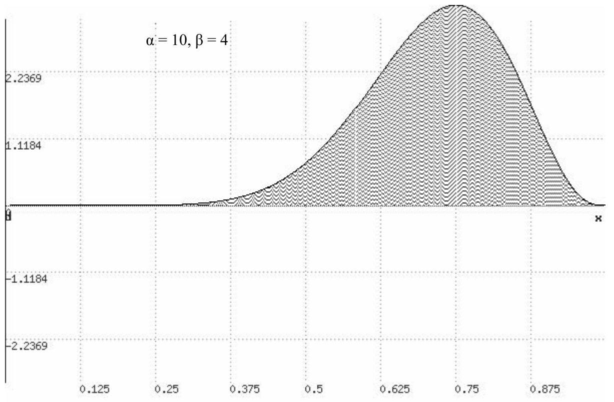

Figure 3. Probability distribution of an actor with observations of α = 10 and β = 4

The same method may be applied to other metrics for example, performance and reliability of an actor to produce its uncertain maximum probabilities. Therefore, we represent an overall probability value to trust an actor as:

$$TP_{actor} = \sum_{i=1}^{n} w_i(Umax(p)_i) \tag{3}$$

where n represents the number of metrics being measured, and $Umax(p)_i$ is the maximum probability for the i^{th} metric. w_i is the weight determining the *relative importance* attached to the i^{th} metric and $\sum_{i=1}^{n} w_i = 1$. Extending $Umax(p)_{availability}$ = 0.75, suppose the uncertain maximum probability measures for two other metrics are: $Umax(p)_{reliability}$ = 0.66 and $Umax(p)_{performance}$ = 0.75. And the weights are allocated as following; $w_{availability}$ = 0.0, $w_{reliability}$ = 0.5 and $w_{performance}$ = 0.5. This indicates that the probability of the actor being available is of no importance for measuring the overall probability value for trusting the actor. Whereas reliability and performance have equal importance in the overall calculation for the user to make the decision on whether or not to trust the actor. Applying these values in Eq.(3), the overall probability to trust the actor would be;

$$TP_{actor} = (0.75)(0.0) + (0.66)(0.5) + (0.75)(0.5)$$

$$= 0.70$$

It is concluded that trust in the actor is 70%. It can be speculated that a low reliability in this case suggests that the actor's execution time fluctuated in the past, making the actor less reliable. If "reliability" is of major concern, then a higher weight allocated to this attribute would significantly lower the overall trust. Such trust calculations are important if the aim is to re-execute the experiment later for comparison.

4.3. Data Trust

This refers to trust placed on the data consumed and generated by intermediate services that form part of a workflow. Trust on a source and intermediate data is important since errors initiated may accumulate through workflow stages. In light of the use cases in Section 3, data trust can be categorized into two main perspectives: (i) user's subjective view of data requirements, and (ii) objective assessment of data. It should be noted that trust based on the availability or reliability (Section 4.2) of an actor does not influence trust in data – for example accuracy of the data may depend on the algorithm used by the service.

1. **User dependent:** This is a subjective view of trust placed by a user in data, and includes:

   Type of input data: This represents a users perception about output data suitable for a particular workflow stage. This parameter is from use case 3 in Section 3.

   Source confidence: This parameter provides a user's trust placed on the source of the data. For example, provenance information about who provided the data that was used in the process (research institute or public archives) determines the extent to which the data can be trusted.

2. **User independent:** This is intended to be an objective assessment of trust – and independent of any one user. This view is based on assessing for example: a) generated data and comparison of its type with respect to an existing schema; b) conflicts between data produced by a service, and data consumed by the next service in a workflow; c) authentication of services.

5. Decision Tree Model for Result Trust

The three different types of trust measurements may be described as a decision tree. The goal is to identify a question sequence that will help us assess the trust that can be associated with data produced from a process. Using a decision tree consists of first picking the root node – representing the main question being asked. This process is repeated for every node until every path leads to a leaf node.

In our model, the response to a question (represented by a node in the tree) may be either positive (something that adds to trustworthiness of the result) or negative (something that limits or reduces trustworthiness of the result). Although many of the questions in the decision tree are far more complicated, a positive or negative assessment will still give some idea of where the decision maker stands. The decision tree is designed as a tool to help consider different factors objectively and provide an ability to generate an automated approximate trust measure.

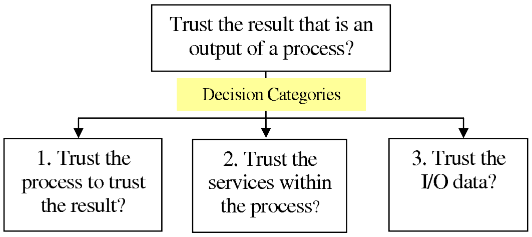
Trust the result that is an output of a process?
Decision Categories
1. Trust the process to trust the result?
2. Trust the services within the process?
3. Trust the I/O data?

Figure 4. Decision Tree Categories

To answer the root question of whether to trust the result that is an output of a process, in figure 4 the trust decision question at the root node is split into three decision categories as a starting point. Depending on the significance of the questions in figure 4, the decision maker may choose to follow either only one path or all the three paths consecutively. The ideal scenario for our model would be to follow all the three paths one after another, thereby producing a complete assessment.

For simplicity the remaining part of the decision tree is split into two parts in Figures 5 and 6. Each node in the tree is numbered to help in explanation, and does not influence in any way the setup of the decision tree algorithm itself. The horizontally shaded nodes indicate the question of whether to stop the assessment at that point or to continue to the next level. In order to explain the working of the decision tree, assumption is made that the path in the decision

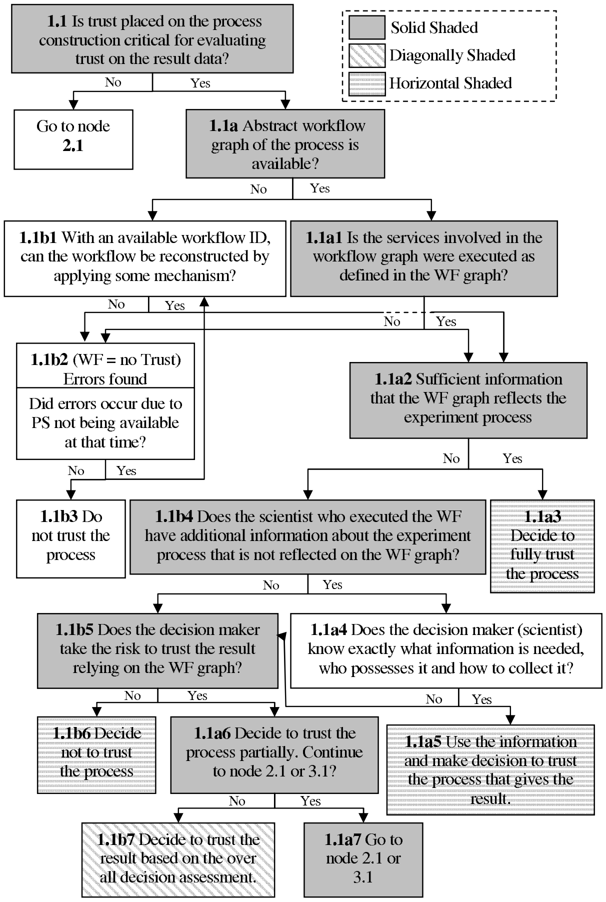
1.1 Is trust placed on the process construction critical for evaluating trust on the result data?
Solid Shaded
Diagonally Shaded
Horizontal Shaded
No
Yes
Go to node 2.1
1.1a Abstract workflow graph of the process is available?
No
Yes
1.1b1 With an available workflow ID, can the workflow be reconstructed by applying some mechanism?
1.1a1 Is the services involved in the workflow graph were executed as defined in the WF graph?
No
Yes
No
Yes
1.1b2 (WF = no Trust) Errors found
Did errors occur due to PS not being available at that time?
1.1a2 Sufficient information that the WF graph reflects the experiment process
No
Yes
No
Yes
1.1b3 Do not trust the process
1.1b4 Does the scientist who executed the WF have additional information about the experiment process that is not reflected on the WF graph?
1.1a3 Decide to fully trust the process
No
Yes
1.1b5 Does the decision maker take the risk to trust the result relying on the WF graph?
1.1a4 Does the decision maker (scientist) know exactly what information is needed, who possesses it and how to collect it?
No
Yes
No
Yes
1.1b6 Decide not to trust the process
1.1a6 Decide to trust the process partially. Continue to node 2.1 or 3.1?
1.1a5 Use the information and make decision to trust the process that gives the result.
No
Yes
1.1b7 Decide to trust the result based on the over all decision assessment.
1.1a7 Go to node 2.1 or 3.1

Figure 5. Process Trust Decision Tree

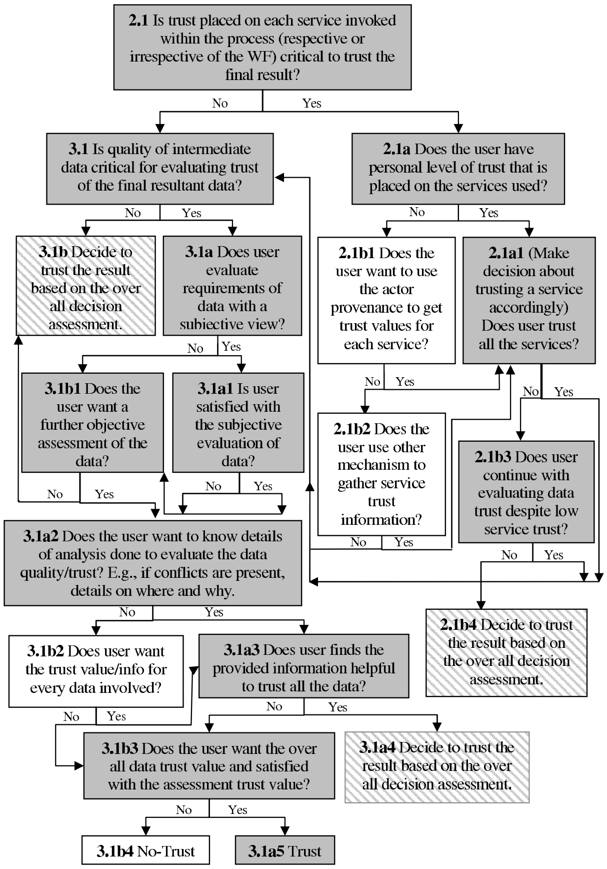
2.1 Is trust placed on each service invoked within the process (respective or irrespective of the WF) critical to trust the final result?
No
Yes
3.1 Is quality of intermediate data critical for evaluating trust of the final resultant data?
2.1a Does the user have personal level of trust that is placed on the services used?
No
Yes
No
Yes
3.1b Decide to trust the result based on the over all decision assessment.
3.1a Does user evaluate requirements of data with a subiective view?
2.1b1 Does the user want to use the actor provenance to get trust values for each service?
2.1a1 (Make decision about trusting a service accordingly) Does user trust all the services?
No
Yes
3.1b1 Does the user want a further objective assessment of the data?
3.1a1 Is user satisfied with the subjective evaluation of data?
No
Yes
No
Yes
2.1b2 Does the user use other mechanism to gather service trust information?
2.1b3 Does user continue with evaluating data trust despite low service trust?
No
Yes
No
Yes
3.1a2 Does the user want to know details of analysis done to evaluate the data quality/trust? E.g., if conflicts are present, details on where and why.
No
Yes
No
Yes
No
Yes
3.1b2 Does user want the trust value/info for every data involved?
3.1a3 Does user finds the provided information helpful to trust all the data?
2.1b4 Decide to trust the result based on the over all decision assessment.
No
Yes
No
Yes
3.1b3 Does the user want the over all data trust value and satisfied with the assessment trust value?
3.1a4 Decide to trust the result based on the over all decision assessment.
No
Yes
3.1b4 No-Trust
3.1a5 Trust

Figure 6. Service and Data Trust Decision Tree

tree is triggered at node 1 in figure 4 followed by the linked node 1.1 in figure 5.

5.1. Path Scenario: A complete Trust Assessment

An example of this decision path is denoted with the solid shaded nodes in figure 5 and figure 6. In the process assessment decision tree (figure 5), the question assess the extent to which the provenance about the process construction is available and accurate. For instance, question 1.1a2 recognizes that despite provenance data about workflow process being available, there can still be doubts about whether to trust the process, as provenance could be limited to some constructs (e.g,. services linked together in sequence) with less or no further explanations on the experiment process itself. Question 1.1b5 recognizes that the decision maker may not rely on the constructs or workflow graph and perceives that the risk of trusting it may exceed a threshold. The example path follows toward question 1.1a6, trusting the process partially and continues to question 2.1. Question 1.1a6 contemplates the uncertainty about the algorithm used within the process, for example the accuracy of the result produced by an actor compared to another.

In the service assessment decision tree figure 6, the questions assess the extent to which actor (service) provenance is available and accurate. Question 2.1a determines the preference to judge the services involved. Here, the user's trust on services can be based on aspects such as the identity of the provider. The example path followed assumes that trust on the services is based on judgement and understanding of the accuracy of the algorithm(s) used by the service. Alternatively, when trust cannot be determined by a decision makers own judgement, question 2.1b1 provides the option to use actor provenance to evaluate trust on the actors (Section 4.2).

From here on, assessment of the data quality is performed to trust the data that is used or produced within the process. Question 3.1a1 is asked to determine the subjective assessment of data described in Section 4.3. Objective analysis of the data using the parameters in Section 4.3 would identify any conflicts or differences present in the

Table 1. Trust Analysis at different decision path

Decision Path	Number of Yes	Number of No	Trust Probability
Complete path	13	4	0.77
Partial path	7	4	0.66
Incomplete path	4	3	0.59
no-trust path	1	3	0.0

data. Question 3.1a2 signifies that the decision maker may choose to view these details to conclude his decision. The path from 3.1a2 in figure 6 follows the course, but does not lead to any conclusion. Question 3.1b3 is relevant for providing estimates over all trust values produced from the objective assessment. Based on this information the user decides to trust (or not) the data ending the path at node 3.1a5.

5.2. Trust Measure Analysis

Our primary aim is to provide an approximate probable measure of trust that depends on the path that is followed in the decision tree. Table 1 presents trust analysis of three different decision paths, with the numbers of *yes* and *no* gathered for each path and the probability values for trusting the result based on the information gathered via the followed path.

The trust probability is calculated by applying the collected boolean values to Eq.(1). Using the beta distribution in Eq.(1) keeps the analysis consistent in terms of implementation, and provides a statistical trust measure that is mathematically sound. Each decision path is discussed below:

1. Complete Path: The complete path in Table 1 is the path described in Section 5.1 with the Figures 5 and 6. This path provides a complete assessment, as it includes all the three stages of trust described in Section 4. The trust probability value of 0.77 reflects mainly three aspects of the decision process:

   (a) There is no additional information about the experiment process available at question 1.1b4 in figure 5.

   (b) It is assumed that the accuracy of a result for some of the algorithms cannot be trusted by the user at node 2.1a1 in figure 6 – as described in Section 5.1. The decision maker finding that some algorithms used in the process give results which are inaccurate, but not incorrect, may influence the overall trust value.

   (c) At node 3.1a1, it is assumed that the user is not satisfied with the subjective assessment of some data which leads to the reduced overall trust value.

2. Partial and Incomplete Paths: Partial assessment indicates ending the trust assessment in the diagonally shaded leaf nodes in the decision tree so that the approximate trust value for the result is calculated at these points. For all the paths in Table 1, the data is gathered and an assessment is made following the same example path explained in Section 5.1. But this path is arranged to give two directions ending at node 2.1b4 in figure 6 or at node 1.1b7 in figure 5, giving the partial and incomplete paths respectively. For instance at node 2.1b4, the decision is made to trust the result based on the rate of process correctness and service trustworthiness. At node 2.1b3, the accuracy of outcome of some algorithms may be used, resulting in low trust in the services. Also the assessment ending at node 2.1b4 shows that data quality assessment is not important for the decision maker to trust the result. These factors are reflected in the comparatively lower overall trust probability value of 0.66. Whereas in case of the incomplete path, the trust probability value of 0.59 reflects a low service and data trust.

3. No-Trust Path: When the path is followed from node 1.1 ending at the leaf node 1.1b3, the trust probability value is 0.0. This indicates that the particular experimental result cannot be trusted, ending the assessment. This indicates that there is limited information available about the workflow run that produced the result in the Provenance Store.

6. Trust Framework Evaluation

Our framework focuses on overcoming the difficulty of answering the question of how much trust can be placed on the result generated by workflow executed given its provenance. The goal of presenting the decision tree is to provide structured support in answering this question and in relation to the motivating use cases of BDW application scenario. We present a short evaluation on application of our trust framework using the decision tree in context of the BDW workflow scenario.

The accuracy of the workflow construction in use case 1 (Section 3) can simply be identified by either examining provenance of the available abstract workflow graph or directly querying for provenance data. Provenance record in both situations verifies if the experiment was being run to trust the workflow. Process trust depends on the trust the evaluator or decision maker may place on the way the workflow is constructed. This is supported in the decision tree indicating that the provenance which simply show the sequence of activities carried out may be insufficient to place complete trust on the process (figure 5). In case of BDW application, the accuracy of an algorithm is given as part of actor provenance. Although the logical way to use this information is by combining it in some way to determine an overall accuracy for the projected image, our current framework lacks such support but instead supports interpreting accuracy information for subjective examination of the algorithms/services during the service trust decision process. If any actor has less accuracy then expected, the decision maker may see it as being less trustworthy (figure 6).

Our decision tree based approach currently presents partial automation in context of consuming the information for the result trust evaluation. At this stage of our work, the framework addresses the issue a decision maker's personal actor trust assessment. Our approach seeks to apply both subjective and objective assessment of data as identified in the BDW scenario. The overall aim of our decision tree approach is to combine trust evaluation at different levels to automatically produce a trust value as a judgement made on a result of a workflow.

7. Conclusion

We have presented a trust framework that combines different aspects of trust with reference to a workflow application. In particular, we have considered the importance and use of process provenance and actor provenance during the trust assessment process, to generate the "trustworthiness" that may be associated with a final result that is generated. Indeed, the trust at different stages of assessment in the decision process may be evaluated, and the overall trust on the result is directly related to the trust the user has about all the three stages (process, actor/service and data) and what trust decision path is followed.

Our trust model for actors/services is based on a probabilistic beta distribution. This probability trust calculator is used for actor trust calculation using actor metrics, as well as the overall result trust calculation using the data gathered from the user decision path. Our future research direction focuses on the implementation of the remaining components of the trust framework. The focus of this paper is to provide a decision tree that determine how trust values can be estimated in an automated way using recorded provenance data.

8. Acknowledgments

Thanks to colleagues for discussions on provenance-based trust including Luc Moreau, Simon Miles and Paul Groth at the University of Southampton. This research is funded in part by EPSRC PASOA project (reference GR/S67623/01). Also, our thanks to the Bio-Diversity World Project members for help with the use cases.

References

[1] A. Abdul-Rahman and S. Hailes. Using recommendations for managing trust in distributed systems. In *Proceedings IEEE Malaysia International Conference on Communication*, 1997.

[2] J. M. Bernardo and A. F. Smith. *Bayes Theorem*, chapter 3, pages 116–117. John Wiley & Sons, West Sussex, England, May 2000.

[3] R. Falcone and C. Castelfranch. Social trust: A cognitive approach. *Trust and Deception in Virtual Societies Journal*, pages 55–90, 2001.

[4] gridprovenance. *http://gridprovenance.org*, 2005.

[5] A. Jones, R. White, N. Pittas, W. Gray, T. Sutton, X. Xu, O. Bromley, N. Caithness, F. Bisby, N. Fiddian, M. Scoble, A. Culham, and P.Williams. BiodiversityWorld: An architecture for an extensible virtual laboratory for analysing biodiversity patterns. In *UK e-Science All Hands Meeting, EPSRC*, pages 759–765, Nottingham, UK, 2003.

[6] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[7] myGrid. *http://www.mygrid.org.uk/*, 2004.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford Digital Library Technologies Project*, 1998.

[9] J. Sabater and C.Sierra. Regret: a reputation model for gregarious societies. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agents Systems*, 2002.

[10] A. Shaikh Ali, O. F. Rana, and R. J. Al-Ali. *High Performance Computing: Paradigms and Infrastructure*, chapter Evidence-aware Trust Model for Dynamic Services. 2005.

[11] M. Witkowski, A. Aritikis, and J. Pitt. Experiments in building experiential trust in a society of objective-trust based agents. *Trust in Cyper-societies*, pages 111–132, 2001.

[12] I. Wootten, S. Rajbhandari, O. Rana, and J. Pahwa. Actor provenance capture with ganglia. In *6th IEEE International Symposium Cluster Computing and the Grid (CCGrid2006)*, 2006.