Massachusetts Institute of Technology
Engineering Systems Division

Working Paper Series

ESD-WP-2008-04

MEASURING DATA BELIEVABILITY: A PROVENANCE APPROACH

Nicolas Pratt[1] and Stuart E. Madnick[2]

[1]ESSEC Business School
prat@essec.fr

[2]Sloan School of Management
Massachusetts Institute of Technology
smadnick@mit.edu

January 2008

Measuring Data Believability: a Provenance Approach

Nicolas PRAT
ESSEC Business School
Avenue Bernard Hirsch
BP 50105 - 95021 Cergy Cedex - FRANCE
prat@essec.fr

Stuart MADNICK
MIT Sloan School of Management
30 Wadsworth Street – Room E53-321
Cambridge MA 02142 - USA
smadnick@mit.edu

Abstract

Data quality is crucial for operational efficiency and sound decision making. This paper focuses on believability, a major aspect of quality, measured along three dimensions: trustworthiness, reasonableness, and temporality. We ground our approach on provenance, i.e. the origin and subsequent processing history of data. We present our provenance model and our approach for computing believability based on provenance metadata. The approach is structured into three increasingly complex building blocks: (1) definition of metrics for assessing the believability of data sources, (2) definition of metrics for assessing the believability of data resulting from one process run and (3) assessment of believability based on all the sources and processing history of data. We illustrate our approach with a scenario based on Internet data. To our knowledge, this is the first work to develop a precise approach to measuring data believability and making explicit use of provenance-based measurements.

1. Introduction

Data quality is crucial for operational efficiency and sound decision making. Moreover, this issue is becoming increasingly important as organizations strive to integrate an increasing quantity of external and internal data. This paper addresses the measurement of data believability. Wang and Strong [1] define this concept as "the extent to which data are accepted or regarded as true, real and credible". Their survey shows that data consumers consider believability as an especially important aspect of data quality. Besides, the authors characterize believability as an intrinsic[1] (as opposed to context- i.e. task-dependant) data quality dimension.

From the definition of believability, it is clear that the believability of a data value depends on its origin (sources) and subsequent processing history. In other words, it depends on the data provenance (aka lineage), defined in [2] as "information that helps determine the derivation history of a data product, starting from its original sources". There exists a substantial body of literature on data provenance. Several types of data provenance have been identified, e.g. "why-provenance" versus "where-provenance" [3] [4], and schema-level versus instance-level provenance [5]. Major application areas include e-science (e.g. bioinformatics) [2] [6] [7] [8], data warehousing and business intelligence [9], threat assessment and homeland security [10] [11] [12]. Among the several possible uses of provenance information, data quality assessment is widely mentioned [2] [6] [13] [14]. Ceruti et al. [12] even argue that a computational model of quality (enabling quality computation at various aggregation levels) should be an integral part of a provenance framework. However, in spite of the relationship between data provenance and quality, no computational model of provenance-based data quality (and more specifically believability) can be found in extant data-provenance literature. It should be noted that some papers, including [10], [15] and [16] address knowledge (as opposed to data) provenance. More specifically, [10] and [15] deal with the issue of trust-based belief evaluation. However, those papers deal with the believability of knowledge (represented as logical assertions). In contrast, we focus on data believability.

In the literature of data quality, believability has been defined in [1]. Guidelines for measuring this quality dimension may be found in [17] (pp. 57-58). However, these guidelines remain quite general and no formal metrics are proposed. An earlier data quality paper [18] addresses the issue of lineage-based data

[1] Although the distinction between intrinsic and contextual data quality is not always clear-cut and often more a matter of degree, this makes believability more easily amenable to automatic computation than other contextual dimensions like relevancy or timeliness.

quality assessment (even if the concept of lineage/provenance is not explicitly mentioned). However, the authors address data quality (defined as the absence of error) in a general and syntactic way. We argue that the different dimensions of quality (and, more particularly, of believability) have different semantics, which should be explicitly considered for quality computation.

Summing up the contribution of extant literature, (1) the literature on provenance acknowledges data quality as a key application of provenance, but does not provide an operational, computational model for assessing provenance-based data believability and (2) the literature on data quality has defined believability as an essential dimension of quality, but has provided no specific metrics to assess this dimension. Consequently, the goal of our work is to develop a precise approach to measuring data believability and making explicit use of provenance-based measurements.

The rest of the paper is structured as follows. Section 2 presents the dimensions of believability. Section 3 presents our provenance model. This model aims at representing and structuring the data which will then be used for believability computation. The approach for believability measurement is presented in section 4. It is structured into three increasingly complex building blocks: (1) definition of metrics for assessing the believability of data sources, (2) definition of metrics for assessing the believability of data resulting from one process run and (3) global assessment of data believability. Section 5 applies our approach to an example scenario based on Internet data, and section 6 concludes with a discussion and points to further research.

2. Dimensions of believability

Believability is itself decomposed into sub-dimensions. Lee et al. [17] propose three sub-dimensions, namely believability: (1) of source, (2) compared to internal common-sense standard, and (3) based on temporality of data. Table 1 refines this typology (the notations introduced in the table will be used in section 4).

Table 1. Dimensions of believability

DIMENSION (NOTATION)	DEFINITION
1. Trustworthiness of source (S_i)	The extent to which a data value originates from trustworthy sources.
2. Reasonableness of data (R_i)	The extent to which a data value is reasonable (likely).
2.1 Possibility ($R1_i$)	The extent to which a data value is possible.
2.2 Consistency ($R2_i$)	The extent to which a data value is consistent with other values of the same data.
2.2.1. Consistency over sources ($R21_i$)	The extent to which different sources agree on the data value.
2.2.2. Consistency over time ($R22_i$)	The extent to which the data value is consistent with past data values.
3. Temporality of data (T_i)	The extent to which a data value is credible based on transaction and valid times.
3.1. Transaction and valid times closeness ($T1_i$)	The extent to which a data value is credible based on proximity of transaction time to valid times.
3.2. Valid times overlap ($T2_i$)	The extent to which a data value is derived from data values with overlapping valid times.

3. Provenance model

Several “generic” provenance models have been proposed in the literature. These models are generic in that they may be used for a wide variety of applications. The W7 model is proposed by Ram and Liu [11]. This model represents the semantics of provenance along 7 complementary perspectives: “what” (the events that happen to data), “when” (time), “where” (space), “how” (actions), “who” (actors), “which” (devices) and “why” (reason for events, including goals). The W7 model is expressed with the ER formalism [19]. [8] presents ZOOM, a generic model to capture provenance for scientific workflows. Finally, [20] presents initial ideas concerning the data model of the Trio system. One of the characteristics of Trio is the integration of lineage with accuracy/uncertainty.

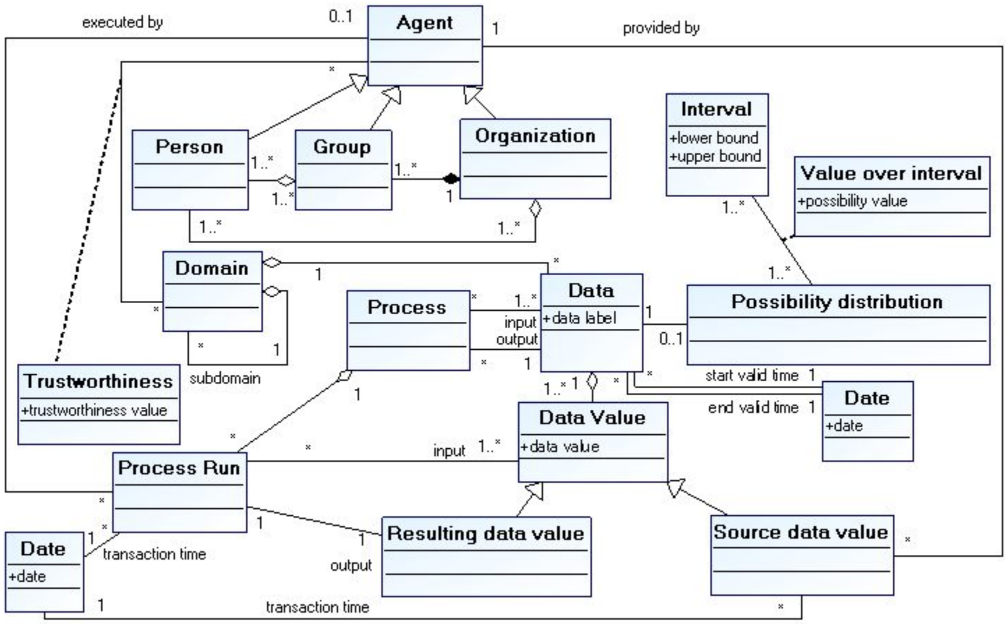

Figure 1. UML representation of the provenance model

Contrary to the above-presented provenance models, which are generic, we have a specific objective in mind, namely the computation of the different dimensions of believability. Therefore, some semantic perspectives need to be developed more thoroughly, while others are of secondary interest. For example, using the terminology of the W7 model, the reasons for events ("why") are of little interest for computing believability. On the contrary, the "when" is crucial in our case, especially for assessing the third dimension of believability (temporality of data).

Figure 1 represents our provenance model in UML notation [21]. Section 4 will illustrate how the elements of the provenance model are used for computing the different dimensions of data believability (introduced in section 2).

Since our goal is to assess the believability of data values, they are the central concept of the model. A data value may be atomic or complex (e.g. relational records or tables, XML files…). Our current research is focused on atomic, numeric data values. Other types of values will be explored in further research, and the provenance model will be refined accordingly.

A data value (e.g. 25 580 000) is the instance of a data (e.g. "the total population of Malaysia in 2004"). A data value may be a source or resulting data value, where a resulting data value is the output of a process run. We introduce this distinction between source and resulting data values because different believability metrics (presented in section 4) are used for these two types of values. The notion of source data value is relative to the information system under consideration: very often, a "source" data value is itself the result of process runs, but these processes are outside the scope of the information system.

A process run is the instantiation (i.e. execution) of a process. This distinction between process runs and processes parallels the distinction between data values and data, respectively. The distinction is similar to the one between steps and step-classes proposed in ZOOM [8]. In our approach, processes may have several inputs but only have one output. This restricted notion of process aims at simplifying believability computation. However, this notion of process is quite general. For example, similarly to the process of data storage [18], the paste operation can be represented as a process whose input is the source and output the target data value.

A data value has a transaction time. For a resulting data value, the transaction time is the execution time of the process run that generated the data value. For a source data value, the transaction time is attached directly to the data value. For example, if a source data value comes from a Web page, the transaction time can be defined as the date when the Web page was last updated. In addition to transaction time, we use the notion of valid time, defined as follows in [22] (p. 53): "The valid time of a fact is the time when the fact is true in the modeled reality. A fact may have associated any number of instants and time intervals, with single instants and intervals being important special cases." Contrary to transaction time which depends on process execution, valid time depends on the semantics of data. For example, for the data "the total population of

Malaysia in 2004”, the start valid time is January 1 and the end valid time is December 31, 2004. The distinction between valid time and transaction time is crucial in our approach. These concepts are used explicitly in the assessment of the two sub-dimensions of temporality. Although transaction time and valid time are standard concepts in temporal databases, we haven’t encountered this distinction in extant provenance models.

When computing data believability (more precisely, when assessing the first sub-dimension of the dimension “reasonableness of data”), we will use the concept of possibility defined in possibility theory [23]. Accordingly, a possibility distribution is associated with data. Possibility distributions may be acquired from experts. They take their values between 0 (impossible) and 1 (totally possible) and may be defined on intervals [24]. For example, if one considers that the total population of Malaysia in 2004 is somewhere between 10 000 000 and 40 000 000, this can be expressed by a possibility distribution with a value of 1 in the [10 000 000 ; 40 000 000] interval, and 0 outside. In this case, the possibility distribution is equivalent to an integrity constraint stating that the total population of Malaysia in 2004 should be in the [10 000 000 ; 40 000 000] range. However, possibility distributions allow for a fine-tuned representation of uncertainty, by using possibility values between 0 and 1. The possibility distribution then approaches a bell-shaped curve, with a value of 1 around the center of the interval (e.g. between 20 000 000 and 30 000 000 in our example), and decreasing values as one gets closer to the extremities of the interval. Like our provenance model, Trio combines provenance with uncertainty. However, contrary to Trio, we use the possibility theory instead of probabilities to represent uncertainty. We believe that possibilities provide a more pragmatic approach. In particular, possibility distributions are easier to acquire from experts than probability distributions.

Processes are executed by agents (organizations, groups or persons). This concept also represents the providers of the source data values. For example, if a data value comes from the Web site of the Economist magazine, the agent is the Economist (an organization).

When computing believability, we are not interested in agents per se, but in the trustworthiness of these agents. The concept of trustworthiness is essential for assessing the dimension “trustworthiness of source”. We use the term “trustworthiness” in a similar way as [25]. Trustworthiness is evaluated for an agent, for a specific knowledge domain [25] [26]. Examples of knowledge domains are “management”, “engineering”… Trustworthiness is closely related to trust and reputation. Reputation is similar to our concept of trustworthiness, but we consider this term as too general i.e. reputation does not depend on a specific domain. Trust, contrary to reputation, is subjective i.e. depends on a particular evaluator, the “trustor” [26]. We avoid introducing this subjectivity in our approach. This is consistent with the finding that data consumers consider believability and reputation as an intrinsic part of data quality [1]. However, trust is a function of reputation [27], and a natural extension of our work would be a more subjective, user-centered assessment of believability.

Trustworthiness in an agent for a domain is measured by a trustworthiness value, normalized between 0 and 1. The computation of these values is outside the scope of our work. We assume that these values are obtained from outside sources, e.g. reputation systems [28]. Thus, the trustworthiness of the magazine “The Economist” is available from Epinions (www.epinions.com). Heuristics may also be used to propagate trustworthiness. For example, [29] shows that an individual belonging to a group inherits a priori reputation based on that group’s reputation.

Summing up, our provenance model is specific to believability assessment. Consequently, it integrates all the concepts that we will need for provenance-based believability assessment. The model was elaborated by integrating concepts from existing models, by specifying these concepts and adding new concepts (e.g. possibility). Our model is represented with an object-oriented formalism (UML), thus enabling a more precise representation of semantics than with the standard ER formalism. Finally, our provenance model is also guided by pragmatic considerations: several provenance metadata used in our approach (e.g. process execution data like transaction time, input or output values, actors…) are relatively easy to trace and/or readily available in existing tools (e.g. log files in workflow tools, “history” tab in Wikepedia – www.wikipedia.org –, …).

4. Provenance-based believability assessment

Based on the information contained in the provenance model, our approach computes and aggregates the believability of a data value across the different dimensions and sub-dimensions of believability (as presented in Table 1). The approach is structured into three building blocks.

4.1. Believability of data sources

This section presents the metrics and parts of the associated algorithms for computing the sub-dimensions of the believability of data sources. The metrics are real values ranging from 0 (total absence of quality) to 1 (perfect quality). The algorithms use an object-like notation (for example, for a data value v, v.data is the object of class Data corresponding to the data value v).

The **trustworthiness** of a source data value v (noted $S_1(v)$) is defined as the trustworthiness of the agent which provided the data value (the knowledge domain for which the trustworthiness of the agent is evaluated has to match with the knowledge domain of the data).

In order to compute the **reasonableness** of a source data value v (noted $R_1(v)$), we need to define metrics for possibility ($R1_1(v)$), consistency over sources ($R21_1(v)$), consistency over time ($R22_1(v)$), and aggregate these metrics.

The possibility $R1_1(v)$ of a data value v is retrieved directly from the provenance model, using the possibility distribution of the corresponding data.

To compute consistency over sources ($R21_1(v)$), the intuition is as follows: we consider the other values of the same data, provided by other sources. For each such value, we determine the distance between this value and the value v (to compute this distance, we use a formula widely used in case-based reasoning [30]). We transform distances into similarities by taking the complement to 1, and compute the average of all similarities. Our approach for computing consistency over sources is similar to the approach described by Tversky [31] for computing the prototypicality of an object with respect to a class (this prototypicality is defined as the average similarity of the object to all members of the class). More formally, based on the UML provenance model represented in Figure 1, the metric $R21_1(v)$ is defined as follows:

Let *d:Data* such that *v.data =d*

Call *Min* and *Max* the smallest (respectively largest) values for which the possibility distribution of d is >0 ([Min ; Max] is thus the range of possible values for data d)

Let *Set1={v':Data value* such that *v'.data=v.data* AND *v'.provided by≠v.provided by}*

$$R21_1(v) = \left(\sum_{v' \in Set1} \left(1 - \frac{|\mathrm{v - v'}|}{\mathrm{Max - Min}}\right)\right) \Big/ Card(Set1)$$

For consistency over time ($R22_1(v)$), the intuition is that values of the same data should not vary too much over time, otherwise they are less believable. The basic principle for computing this metric is similar to the previous metric. However, the specific semantics of time has to be taken into account. Also, this metric assumes that effects of seasonality are absent or may be neglected.

The reasonableness $R_1(v)$ of a source data value v is computed by aggregating the values of the above-presented metrics. In order to compute the value of a dimension based on the values of its sub-dimensions, the most common aggregation functions are Min, Max, and (weighted) Average [17]. The choice of the appropriate aggregation function depends on the semantics of the dimensions and sub-dimensions, and on the available information. Here, consistency may be defined as the weighted average of the values of its two sub-dimensions (by default, the weights are equal). However, to compute reasonableness from possibility and consistency, the Min operator is more appropriate. Possibility depends solely on the experts' evaluation, while consistency is strongly correlated with the different data values considered for comparison. Therefore, we make the most cautious choice for aggregating possibility and consistency, namely the Min operator. Alternatively, if the criterion of consistency is considered too much dependant on context or the computation cost too high, the measurement of reasonableness may be based on possibility only. Formally, we have:

Let *$r21_1$ and $r22_1$ be the respective weights of consistency over sources and consistency over time ($r21_1 + r22_1 = 1$)*

$R_1(v)=MIN(R1_1(v), R2_1(v))$

$=MIN(R1_1(v), (r21_1 * R21_1(v) + r22_1 * R22_1(v)))$

To compute the **temporal** believability of a data value v ($T_1(v)$), we consider two aspects: believability based on transaction and valid times closeness, and believability based on valid times overlap.

For believability based on transaction and valid times closeness, the intuition is that a data value computed in advance (estimation) is all the more reliable as the valid time (especially the end valid time) of the data value approaches. To capture this idea, various metrics may be used (e.g. linear, exponential). Here, drawing from the metrics proposed for data currency in [32], we propose an exponential function. The function grows exponentially for transaction times before the end valid time. When transaction time is equal or superior to the end valid time, the value of the metric is 1. A decline coefficient [32] may be used to control the shape of the exponential function. Alternatively, we could use other metrics, e.g. metrics using a different function before and after the start valid time.

Believability based on valid times overlap measures the extent to which a data value resulting from a process is derived from data values with "consistent" i.e. overlapping valid times. Thus, this metric is defined for resulting data values and shall be developed in section 4.2. For source data values, the value of this metric may be defaulted to one (or, alternatively, the weight of the sub-dimension “believability based on valid times overlap” may be set to zero).

Consequently, $T_1(v)$ is defined as follows:

Let *tt:Date* such that *v.transaction time = tt*
Let *vt:Date* such that *v.data.end valid time = vt*
Let *t1 be a decline factor (t1>0)*

$$T1_1(v) = MIN(e^{(-t1*(vt-tt))}, 1)$$

Let $t1_1$ *and* $t2_1$ *be the weights of the two sub-dimensions of temporality* ($t1_1 + t2_1 = 1$)

$$T_1(v) = t1_1 * T1_1(v) + t2_1 * T2_1(v) = t1_1 * T1_1(v) + t2_1$$

4.2. Believability of process results

The quality of any data value depends on the quality of the source data and on the processes. By combining data, processes may amplify data errors (i.e. quality defects), reduce them, or leave them unchanged, depending on the processes; moreover, processes themselves may be error-prone [18].

Following the line of [18], we present metrics for assessing the believability of data resulting from one process run, as the next building block of our approach for global believability assessment. More precisely, we consider a process P whose input data values are denoted by v_i (i=1...n, where n is the number of input parameters of the process). We want to determine the believability of the data value (noted v) resulting from P, along the different dimensions of believability. Departing from [18], which treats all types of quality errors uniformly, we claim that as data are transformed through processes, the evolution of the different dimensions of believability (and, more generally, quality), depends not only on the data and processes, but also on the dimensions considered and on their semantics. Therefore, as in section 4.1, we distinguish between the different dimensions of believability.

For simplicity, this paper assumes that processes are error-free (e.g. a process specified as dividing one number by another makes the division correctly).

To compute the source **trustworthiness** $S_2(v)$ of an output data value v based on the source trustworthiness of the input data values, we use partial derivatives, adapting the general algorithm proposed in [18] for error propagation (in this paper, we consider the particular case of processes for which these partial derivatives are defined). An error caused on v by a lack of trustworthiness of an input value v_i has an incidence on v which depends not only on the value of v_i itself, but also on the “weight” (influence) of v_i in process P, as measured by the derivative. Consequently, to measure the lack of trustworthiness of v, we compute the weighted average of the lack of trustworthiness for the v_i (i=1...n). The weight of v_i is the value of v_i multiplied by the value of the derivative dP/dx_i. We normalize the weights such that their sum equals one, and take absolute values to avoid negative weights.

$\forall$ *v_i (i=1...n), call $S(v_i)$ the source trustworthiness of v_i ($S(v_i)$ may have been determined with the metric S_1 presented in section 4.1 if v_i is a source data value, or with the metric S_2 if v_i is itself a value resulting from a previously executed process).*

$$S_2(v) = 1 - \frac{1}{\sum_{i=1}^{n} \left| \frac{dP}{dx_i}(v_i) * v_i \right|} * \left(\sum_{i=1}^{n} \left| \frac{dP}{dx_i}(v_i) * v_i \right| * (1 - S(v_i)) \right)$$

As an illustration of this metric, consider a process P defined by: $P(y) = 3*x_1 + 2*x_2$, and suppose that the value of x_1 is 2 with trustworthiness 0.8 while the value of x_2 is 3 with trustworthiness 0.6. In the present case, the derivatives (3 and 2 respectively) are constant. Applying the metric, the trustworthiness of the resulting data value (12) is 0.7, i.e. the average of the trustworthiness of the two input data values. In this case, the input data values equally contribute in the assessment of the trustworthiness of the result. Assuming now that the value of x_2 is 30, the trustworthiness of the resulting data value (66) is 0.62, reflecting a much more significant role of x_2 in the output data value.

To compute the **reasonableness** $R_2(v)$ of an output data value, we need to consider the sub-dimensions of reasonableness. Concerning consistency, since it depends on the data values considered for comparison, it may not easily be derived from the consistency computed for the input values v_i. Therefore, if consistency is used to assess reasonableness, it has to be computed again for the data value v, based on the metric presented in section 4.1.

In order to compute the possibility of v based on the input values v_i, we follow similar lines of reasoning as for combining trustworthiness (i.e. combination based on derivatives, assuming again that all partial derivatives of process P are defined).

To compute **temporal** believability, we consider its two sub-dimensions.

Concerning believability based on transaction and valid times closeness, the principle is the same as in section 4.1. (The transaction time is the transaction time of the process run).

Believability based on valid times overlap measures the extent to which the valid times of the input values v_i of process P are consistent with each other, i.e. their degree of overlap. In order to define the corresponding metric, we assume here, for the sake of simplicity, that there are only two input values v_1 and v_2; we also assume that the objective of process P is not to compute an evolution (in the later case, it is normal that the input data – e.g. the total sales in fiscal year 2005 and the total sales in fiscal year 2006 – do not have overlapping valid times). Formally, the metric for believability based on valid times overlap is defined as follows:

Call *VTv_1 the valid time interval of v_1 (interval delimited by the start and end valid times of v_1).*
Call *VTv_2 the valid time interval of v_2.*

If *($VTv_1 \cap VTv_2 = ø$)*
Then
$T2_2(v)=0$
Else

$$T2_2(v)=\frac{length(VTv_1 \cap VTv_2)}{MAX(length(VTv_1), length(VTv_2))}$$

Endif

4.3. Global believability

At this point, it is clear that to compute the believability of a data value v, we need to consider the provenance/lineage of this data value, i.e. its origin and processing history.

Some aspects of believability are transmitted along the transformation chain of data values. Such is the case with trustworthiness, which is transmitted along processes using the derivative-based metric presented in section 4.2. However, some other aspects of believability may not be transmitted as data move along the process chain. This may be the case, for instance, for possibility (a data value may appear completely possible even though it results from highly implausible data values). This can also happen with the sub-dimensions of temporality. For example, a data value v may be computed by a process P after the end valid time of this value (therefore performing well on the sub-dimension “believability based on transaction and valid times closeness”). However, the input values of P may themselves result from processes performing poorly on the sub-dimension “believability based on transaction and valid times closeness”.

Since some aspects of believability may not be transmitted as data move across processes, we need metrics accounting for this phenomenon, considering the complete lineage of a data value. For example, if a highly possible data value v results (directly or indirectly) from highly implausible values, this means that v is highly possible “by accident”. We want to reflect this in the believability computation of v.

The central idea of global believability assessment is to consider the complete lineage of a data value. Therefore, at this point, we need a more precise definition of data lineage. The lineage of a data value v is a labeled, directed acyclic graph representing the successive data values and processes leading to data value v. Figure 2 illustrates an example lineage, where data value v is computed by process P_2 from values v_{21} and v_{22}; v_{21} itself is computed by process P_1 from values v_{11} and v_{12}.

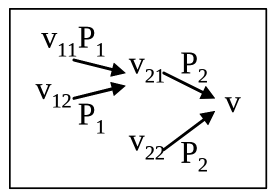

Figure 2. Example lineage

Based on lineage, the global believability of a data value v is computed as follows:

(1) For each of the three dimensions of believability, a global value for this dimension is computed, by considering the data lineage of v. For instance, if the dimension considered is temporal believability, the global temporal believability of v is noted $T_3(v)$. This global temporal believability is computed by averaging the temporal believability of all values in v’s lineage. For example, in the example above, $T_3(v)$ is computed by averaging $T_2(v)$ with $T_2(v_{21})$, $T_1(v_{22})$, $T_1(v_{11})$ and $T_1(v_{12})$. (According to the notation introduced in Table 1, T_1 and T_2 designate the temporal believability of data sources and of process results respectively). When computing a global value for any of the three believability dimensions, two types of weights are used (i.e. the average is a weighted average). The first weight is a “discount factor” [33], as often proposed in graph-based algorithms. This factor reflects the intuition that the influence of a vertex on another decreases with the length of the path separating the two vertices (the further away a value is in v’s lineage, the less it counts in the global believability of v). The discount factor may be different for the three dimension of believability, depending on the semantics of the dimension. In addition to discount factors, a second type of weight is used, based on derivatives, similarly to the approach

presented in section 4.2 for computing the source trustworthiness of output data values. These weights reflect the fact that for a given process, the input data values do not contribute equally to the process and, consequently, to its result.

(2) Once a global value has been defined for each of the three dimensions of believability, global believability is computed by multiplying these three values.

5. Application scenario

A communication group considers launching a new TV channel in Malaysia and Singapore, aimed more specifically at the Indian community. The group needs to know the total Indian population in Malaysia and Singapore. This figure is computed from source data found on Internet. We wish to asses the believability of this figure (the value v). The lineage of v is structured as in the graph of Figure 2. In this case, P_1 is the multiplication and P_2 the sum; the values in v's lineage and their characteristics are shown in Table 2. Start and end valid times are determined based on the semantics of the corresponding data, as expressed by the data labels. Transaction times are determined differently for source data values (v_{11}, v_{12} and v_{22}) and for resulting data values (v_{21} and v). For a source data value, transaction time is determined from temporal information found (when available) on the Web site providing the data value. For a processed data value, transaction time is the hypothetical date of computation of the value. The last column indicates the origin of the value. This origin is either the Web site of an organization (for a source data value), or the execution of a process.

Table 3 exhibits the values for the different dimensions of believability (computed with the algorithms of section 4.1. for source data values and 4.2 for resulting data values). The trustworthiness of v_{11}, v_{12} and v_{22} is the trustworthiness of the Malaysian Department of Statistics, the CIA and the Singapore Department of Statistics. We assume the values of trustworthiness to be 0.9, 0.8 and 0.9 respectively (The CIA is hypothesized to be less trustworthy in estimating demographic figures pertaining to Malaysia or Singapore, than the Department of Statistics of these countries). The trustworthiness of v_{21} is the average of the trustworthiness of the two input values of process P_1, reflecting an equal weight of the parameters for this type of process (multiplication). For the second process (sum), v_{21} has more weight than v_{22}. These two parameters play a symmetric role in process P_2, however the value of v_{21} is higher. When combining the trustworthiness of v_{21} and v_{22} using the derivative-based formula presented in section 4.2, we get the value

1– (1/2135280)*(1816180*0.15+319100*0.1)=0.857

Concerning reasonableness (second column of Table 3), we assume, due to space limitation, that only the sub-dimension "possibility" is considered and that all values are totally possible. The last three columns of Table 3 compute the two sub-dimensions of temporal believability, and the average of their values. The metrics reflect that value v_{22} is computed before the end of its valid time (we assume a value of 0.01 for the decline factor t1), and that v is computed based on incompatible valid times.

Table 2. Example scenario

Id	Data	Value	Transaction time	Start valid time	End valid time	Provided By/ Output of
v_{11}	Total population of Malaysia in 2004	25 580 000	31-Dec-05	1-Jan-04	31-Dec-04	Malaysian Dpt of Stats
v_{12}	% Indian population in Malaysia in 2004	7.1	31-Dec-04	1-Jan-04	31-Dec-04	CIA
v_{21}	Indian population in Malaysia in 2004	1 816 180	12-Feb-06	1-Jan-04	31-Dec-04	$P_1(v_{11},v_{12})=v_{11}*v_{12}$
v_{22}	Indian population in Singapore in 2006	319 100	30-Jun-06	1-Jan-06	31-Dec-06	Singapore Dpt of Stats
v	Indian population in Malaysia and Singapore in 2006	2 135 280	1-Jun-07	1-Jan-06	31-Dec-06	$P_2(v_{21},v_{22})=v_{21}+v_{22}$

Table 3. Metric values

Id	S	R	T1	T2	T
v_{11}	0.9	1	1	1	1
v_{12}	0.8	1	1	1	1
v_{21}	0.85	1	1	1	1
v_{22}	0.9	1	0.159	1	0.579
v	0.857	1	1	0	0.5

If the believability of data value v is computed by assigning a value of 0 to the discount factor for all dimensions (i.e. no adjustment of the score based on the lineage of v), the believability of v is 0.43 (S*R*T). This value should be compared to other values (e.g. by simulating the choice of other Internet sources and seeing if the quality is improved). The discount factor may also be assigned a non-zero value, e.g. for the dimension of temporal believability (in this case, the global believability score of data value v improves, which simply reflects the fact that v's lineage performs better than v itself on the dimension of temporal believability).

6. Discussion and conclusion

We have presented and illustrated metrics and a computational approach for measuring data believability. The believability of a data value is computed based on the provenance (lineage) of this value. We have presented the provenance model and associated computation approach, structured into three increasingly complex building blocks. Despite the importance of believability as a quality dimension and the relevance of provenance for its computation, the present work is – to the best of our knowledge – the first operationalizing provenance-based computation of data believability.

This work currently has some limitations. In particular, we only consider atomic data values, and processes for which a derivative is defined (thus excluding operators from relational algebra like selection for example). However, our approach may be applied in several domains, including data warehousing and business intelligence.

The next steps of this research will concentrate on the refinement of the proposed metrics in conjunction with further testing on real case studies, and the development of a tool to capture provenance metadata and use them for believability computation.

7. Acknowledgments

This work benefited from fruitful discussions within the CONTEXT research group at MIT. The authors address their special thanks to Frank Manola and Allen Moulton.

Nicolas Prat also thanks ESSEC Business School for supporting his stay at MIT.

8. References

[1] R. Wang and D. Strong, "Beyond Accuracy: what Data Quality Means to Data Consumers", Journal of Management Information Systems, vol. 12, no. 4, spring 1996, pp. 5-34.

[2] Y.L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science", SIGMOD Record, vol. 34, no. 3, September 2005, pp. 31-36.

[3] P. Buneman, S. Khanna, and W.-C. Tan, "Why and Where: a Characterization of Data Provenance", Proceedings of the 8th International Conference on Database Theory (ICDT 2001), London, UK, January 2001.

[4] R. Wang and S. Madnick, "A Polygen Model for Heterogeneous Database Systems: the Source Tagging Perspective", Proceedings of the 16th International Conference on Very Large Data Bases (VLDB'90), Brisbane, Australia, August 1990.

[5] Y. Velegrakis, R.J. Miller, and J. Mylopoulos, "Representing and Querying Data Transformations", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), Tokyo, Japan, April 2005.

[6] R. Bose and J. Frew, "Lineage Retrieval for Scientific Data Processing: a Survey", ACM Computing Surveys, vol. 37, no. 1, March 2005, pp. 1-28.

[7] P. Buneman, A.P. Chapman, and J. Cheney, "Provenance Management in Curated Databases", Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2006), Chicago, IL, USA, June 2006.

[8] S. Cohen-Boulakia, O. Biton, S. Cohen, and S. Davidson, "Addressing the Provenance Challenge Using ZOOM", Working Paper, Department of Computer & Information Science, University of Pennsylvania, May 2007, http://repository.upenn.edu/cis_papers/338/

[9] Y. Cui and J. Widom, "Lineage Tracing for General Data Warehouse Transformations", VLDB Journal, vol. 12, no. 1, May 2003, pp. 41-58.

[10] L. Ding et al, "On Homeland Security and the Semantic Web: a Provenance and Trust Aware Inference Framework", Proceedings of the 2005 AAAI Spring Symposium on AI Technologies for Homeland Security, Stanford University, CA, USA, March 2005.

[11] S. Ram and J. Liu, "Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling", Proceedings of the Active Conceptual Modeling of Learning Workshop (ACM-L 2006) in conjunction with the 25th International Conference on Conceptual Modeling (ER 2006), Tucson, AZ, USA, November 2006.

[12] M.G. Ceruti et al, "Pedigree Information for Enhanced Situation and Threat Assessment", Proceedings of the 9th International Conference on Information Fusion (ICIF 2006), Florence, Italy, July 2006.

[13] W.-C. Tan, "Research Problems in Data Provenance", IEEE Data Engineering Bulletin, vol. 27, no. 4, December 2004, pp. 45-52.

[14] R. Wang, H. Kon, and S. Madnick, "Data Quality Requirements Analysis and Modeling", Proceedings of the 9th International Conference on Data Engineering (ICDE 1993), Vienna, Austria, April 1993.

[15] M.S. Fox and J. Huang, "Knowledge Provenance in Enterprise Information", International Journal of Production Research, vol. 43, no. 20, October 2005, pp. 4471–4492.

[16] P. Pinheiro da Silva, D.L. McGuinness, R. McCool, "Knowledge Provenance Infrastructure", IEEE Data Engineering Bulletin, vol. 26, no. 4, December 2003, pp. 26-32.

[17] Y. Lee, L. Pipino, J. Funk, and R. Wang, Journey to Data Quality, MIT Press, Cambridge, MA, 2006.

[18] D.P. Ballou and H.L. Pazer, "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems", Management Science, vol. 31, no. 2, February 1985, pp. 150-162.

[19] P.P. Chen, "The Entity-Relationship Model – Toward a Unified View of Data", ACM TODS, vol. 1, no. 1, March 1976, pp. 9-36.

[20] J. Widom, "Trio: A System for Integrated Management of Data, Accuracy, and Lineage", Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 2005), Asilomar, CA, January 2005.

[21] Object Management Group, Unified Modeling Language (UML) specification, version 2.1.1, http://www.omg.org/technology/documents/formal/uml.htm

[22] C.S. Jensen et al, "A Consensus Glossary of Temporal Database Concepts", SIGMOD Record, vol. 23, no. 1, March 1994, pp. 52-64.

[23] D. Dubois and H. Prade, "An Introduction to Possibilistic and Fuzzy Logics", in P. Smets, A. Mamdani, D. Dubois, and H. Prade (eds), Non-Standard Logics for Automated Reasoning, Academic Press, London, 1988.

[24] D. Dubois, H. Prade, and R. Yager "Merging Fuzzy Information", in J.C. Bezdek, D. Dubois, and H. Prade (eds), Fuzzy Sets in Approximate Reasoning and Information Systems, Kluwer Academic Publishers, Boston, 1999.

[25] D. McGuinness et al, "Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study", Proceedings of the Workshop on Models of Trust for the Web (MTW'06) in conjunction with the 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland, May 2006.

[26] J. Huang and M. S. Fox, "An Ontology of Trust - Formal Semantics and Transitivity", Proceedings of the 8th International Conference on Electronic Commerce (ICEC 2006), Fredericton, New Brunswick, Canada, August 2006.

[27] L. Mui, M. Mohtashemi, and A. Halberstadt, "A Computational Model of Trust and Reputation", Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35 2002), Big Island, HI, USA, January 2002.

[28] A. Jøsang, R. Ismail, and C. Boyd, "A Survey of Trust and Reputation Systems for Online Service Provision", Decision Support Systems, vol. 43, no. 2, March 2007, pp. 618-644.

[29] S. Tadelis, "Firm Reputation with Hidden Information", Economic Theory, vol. 21, nos. 2-3, 2003, pp. 635-651.

[30] T. W. Liao, Z. Zhang, and C. Mount, "Similarity Measures for Retrieval in Case-Based Reasoning Systems", Applied Artificial Intelligence, vol. 12, no. 4, June 1998, pp. 267-288.

[31] A. Tversky, "Features of Similarity", Psychological Review, vol. 84, no. 4, July 1977, pp. 327-352.

[32] A. Even and G. Shankaranarayanan, "Utility-Driven Assessment of Data Quality", ACM SIGMIS Database, vol. 38, no. 2, May 2007, pp. 75-93.

[33] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust", Proceedings of the 13th International World Wide Web Conference (WWW 2004), New York, NY, May 2004.