Provenance-based Belief

Adriane Chapman, Barbara Blaustein, Chris Elsaesser
The MITRE Corporation
{achapman, bblaustein, chris}@mitre.org

Abstract

Provenance has been touted as a basis to establish trust in data. Intuitively, belief in a hypothesis should depend on how much one trusts the relevant data. However, current proposals to assess trust based solely on provenance are insufficient for rigourous decision making. We describe a model of provenance and belief that is necessary and sufficient to incorporate "trust in the data" in a way that supports normative inference. The model is based on the observation that provenance can be viewed as a causal structure which can be used to compute belief from assessments of the accuracy of sources and transformations that produced relevant data. In our model, data sources are like sensors with associated conditional probability tables. Provenance identifies dependencies among sensors. Together, this information allows construction of causal networks that can be used to compute the belief in a state of the world based on observation of data. This model formalizes the role of source accuracy, and provides a method for formally assessing belief that uses only information in the provenance store, not the contents of the data.

1. Introduction

The Open Provenance Model (OPM) [12] says, "We assume that provenance of objects (whether digital or not) is represented by an annotated causality graph". While it is possible to argue about whether every provenance graph reflects true causality, this paper focuses on those domains in which reports about the world are collected and fused. In applications such as biosurveillance or global warming, we want to determine how much to believe derived data.

Some in the provenance community assume that knowing the source of data and how it was manipulated, i.e., its provenance, is sufficient to allow a user of the data to make decisions based on how much they trust the data. Researchers are developing methods to use trust metrics on the assumption that they will exist. For example, Dai, et al. [4] assume that a measure of trust of a data item's source(s) exist and propose to use it to return the most trustworthy results for a query.

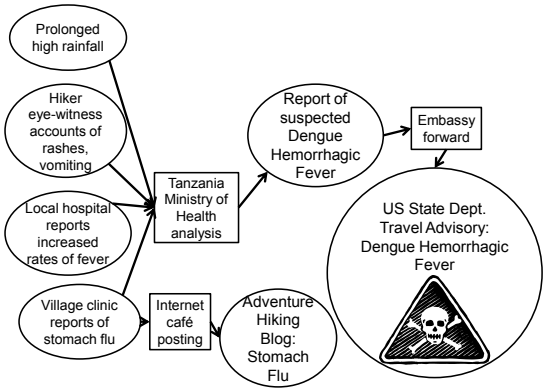

Figure 1: Provenance Information for Dengue Hemoragic Fever Symptoms Report

However, most provenance systems simply record the origin of data and the processes used to transform original data [1, 11, 15]. In these systems, a user reviews the provenance of a data item and arrives at her personal belief in the veracity of data based on subjective assessment of its provenance. In other words, data veracity or trust is based on a "gut feeling" that is derived externally of provenance management systems. Even probabilistic databases [3, 13] merely store the assigned probabilities and manipulate them appropriately during query execution.

Previous works in computing trust [14] or data quality based on provenance [5, 7-9] require information that might exist in the data but is not in a typical provenance store. For instance, [14] require information such as the "reasonableness of data." Moreover, many data quality metrics are based on content [7], uncertainty of the data [5], timeliness as calculated with data expiry age [9], or accuracy of the data [2], all of which are based on information in the data, not the provenance. Our proposed model is based solely upon information that uses and augments a provenance graph.

Consider the following example.

Example: While planning for a trip to Kilimanjaro, you notice a State Department advisory cautioning about a report of an outbreak of Dengue Hemorrhagic Fever

(DHF). On the other hand, your favorite adventure hiking blog merely reports stomach flu.

The provenance for these reports is in Figure 1. In light of the blog posting, might the State Department be overstating the situation? Should you trust the report and alter your travel plans?

With current models of provenance, our intrepid hiker has two options. The first is to view the provenance of the reports, assign a "gut feeling" about each based on the sources that contributed to the reports, and then fuse these to arrive at a belief in the final report. The second option is to obtain the original reports, assess the accuracy, data quality, timeliness, etc. that went into creating those reports and use those assessments to determine if the reports correctly indicate existence of DHF symptoms. Unfortunately for our hiker, Option 1 incorrectly equates the (fused) accuracy of the report's source(s) with the probability that DHF symptoms are present at Kilimanjaro and, therefore, exhibits the Base Rate Fallacy [16]. Option 2 requires more information than exists in the provenance store. Further, such information may not exist by the time a decision must be made. (It is impossible to measure the accuracy of a report on predicted corn consumption in 2010 until 2010 is over and the results have been tallied.)

The main requirement is that the provenance system be extended to capture accuracy (sensitivity and specificity) of sources. When this is so, the computations needed to support decision making are straightforward and efficiently performed by off-the-shelf Bayesian network algorithms. Our model relies on information in the provenance store about how information is propagated through the graph, and how accurate each source is. This information is used to compute belief in derived data items.

Section 2 describes the models that underlie our approach. The model developed to compute belief based upon provenance information is presented in Section 3. Section 4 describes our planned future work and conclusions.

2. Provenance Foundations

The choice of a model for lineage information is completely independent from the base systems' data models and except for a linking identifier, has no communication with the actual base data. We follow the OPM convention and represent artifact and process entities as nodes [12]. A *lineage graph,* then, is a triple, consisting of a graph identifier *G*, a set of nodes, *N*, and a set of edges, *E*. Provenance information forms a Directed Acyclic Graph (DAG). This paper is not concerned with the implementation of the graphs, which could be relational, RDF, XML, etc.

3. Belief, Evidence, and Causality

We are interested in using provenance to support decision making. For instance, should you go to Kilimanjaro, not go, or acquire additional data that might clarify your travel decision? If we are to use provenance to make such decisions in a rigorous manner, we must augment it with probabilities [10].

In order to formalize our derivations, we employ propositional semantics. A proposition is a sentence expressing something true or false. Belief in a proposition is one's subjective probability that the proposition is true. Notationally, belief in proposition C that there are symptoms of DHF at Kilimanjaro is written p(C).

3.1. The influence of evidence on belief

Belief often is not static; rather, it is influenced by evidence. In the example, it stands to reason that one's prior belief in the presence of DHF symptoms at Kilimanjaro, p(C), might increase in light of one or both of the reports. Belief in proposition C in light of proposition E is written p(C|E) and called the conditional probability of C given E. The definition of conditional probability is:

$$p(C|E) = \frac{p(C \wedge E)}{p(E)} \qquad (1)$$

In words, the probability C is true (e.g., DHF symptoms are present at Kilimanjaro) given E (the State Department or blog report) is the proportion of the times one expects C and E to co-occur when E occurs. By division, p(C∧E) = p(E|C) * p(C). Substituting this identity in (1) yields Bayes' rule:

$$p(C|E) = \frac{p(E|C)*p(C)}{p(E)} \qquad (2)$$

Where p(E) = p(E|C) * p(C) + p(E| ¬C) * p(¬C).

3.2. Source accuracy and weight of evidence

Accuracy is the proportion of true results – both positive and negative – in all the results produced by a source. Thus, both components of accuracy, – a source's true positive rate p(E|C) and a source's true negative rate p(¬ E| ¬C) [equal to 1-p(E| ¬C)] are required to calculate belief in proposition C given evidence E. Note that p(C|E) is *not* equal to the accuracy of the source or the sources true positive rate p(E|C) or p(E).

3.3. Causal chains

In this exposition we denote evidence by E and the state of the world by C in part to evoke the idea that evidence (i.e., data) is an effect caused by a state of the world. In our example, E is either the State Department advisory or the blog report, and C is the disease symptom that engendered that report.

The occurrence of a symptom DHF is likely not the primary concern of our hiker, but rather it is the possibility of the presence of DHF that is the ultimate concern. An advantage of causal models is that they can be extended to represent a chain of causes and effects that allows us to address sequences where the effects represented by data can be traced back to the original source.

Let us denote the presence of DHF at Kilimanjaro by S. The causal network that captures the knowledge that S may have caused the symptom C that engendered report E is: S → C → E.

Without going into details, note that Bayes' rule is the normative way to compute p(S|C) and that using a chain of conditional probabilities it is straightforward to compute the belief p(S|E).[1]

This means that provenance graphs such as depicted in Figure 1 can be translated into causal Bayesian networks that support inference about evidence provided by data.[2]

3.4. Integrating causal reasoning with provenance

One issue of significance is that causal networks begin someplace and that someone must provide the *a priori* probability of each node that has no parent. In our example, there are four such nodes. But notice that these nodes represent data and not what caused the data to be observed (possibly incorrectly). In reality, the state of interest to our hiker is DHF which may have caused the reports at the head of the provenance graph but is not represented in the graph. Without knowing the probability of DHF at Kilimanjaro before any evidence was acquired, it is impossible to compute p(DHF|reports). This illustrates that a provenance graph is not sufficient for inference; an external domain model must augment it.

From the exposition above, it should be clear that for a single source of a single piece of data (about a single external cause) we require the probability the source will report the data when the causing state is true *and* the probability the source might report the data when the causing state is *not* true (i.e., the probability the source issues an incorrect report). But our exposition was only about a single source and a single data item. There are only two other cases:

Figure 2(a) depicts single cause C, or multiple *independent* sources of data. This graph structure means that p(C | E1 & … & En) = p(C|E1) * … * p(C|En). Therefore, to support inference we need only the individual conditional probabilities associated with each source (E nodes).

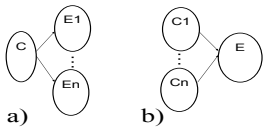

Figure 2: Sample Causal Graphs (a) single cause, (b) multiple sources

Figure 2(b) shows data derived from multiple sources. In this case, the conditional probabilities for E must specify the probability of each state of E for every possible combination of states of C1-Cn.

3.4.1. Generating conditional probability tables

There are three possible ways to enumerate such a conditional probability table. The least desirable is to ask an expert on the sources. This can be an onerous task prone to cognitive biases.

A second way is to use a learning algorithm. For each combination of values of the C nodes, one would initially assign equal probability to each possible value of E. These probabilities would be updated as one obtained verification of the accuracy of the sources. The problem with this approach is that it requires knowledge of results which may be in short supply for rare events (e.g., how many times have you assessed the accuracy of health alerts by the US embassy in Tanzania?).

A third way to produce such tables, which we are investigating in our research, is to create a set of models that are parameterized according to what is immediately "upstream" in the causal graph from C1-Cn. For example, if all predecessors of E have a common ancestor in the provenance graph, that means they are not conditionally independent from node E's point of view. In such a case p(E | C1), …, p(E | Cn), and p(E | C1 & … & C2) might be assumed to be approximately the same when C1-Cn are of the same class having members with approximately the same accuracy in repeating what the common source says. If C1-Cn draw from independent sources we can use what is called a "noisy or" in Bayesian network terminology.

3.4.2. Independence and the Single Source Problem

Utilizing the causal reasoning described above, a large problem in the provenance world is solved automatically: the single source problem. For example, knowing that an assertion that Iraq was developing weapons of mass destruction was based on a single source code-named "Curveball", as opposed to four independent sources, might have influenced belief in the WMD assertion. Meanwhile, four independent sources should create a higher belief in the resulting report [6]. Because provenance is a DAG, this can be accounted for with conditional probabilities. Moreover, partial dependencies in the graph, such as the one that exists in Figure 1, are

[1] In practice these computations are performed using algorithms that implement Bayesian belief networks and are available in a number of off-the-shelf systems.

[2] The translation from a provenance graph to a causal Bayesian network is direct since provenance forms a directed acyclic graph.

also automatically comprehended by the use of causal reasoning.

3.4.3. Impact of Processes

Processes have a large impact on the belief of their derived data. Consider in our example, the process "Embassy Forward". Suppose this was done via a disenchanted intern, whose selection of what material to copy and forward was done haphazardly. The final Travel Advisory could look very different from one created by the bright, excited intern who fully read and understood the Tanzanian Ministry of Health's report. Initially, all processes can use a default conditional probability table. However, these could be altered if some information is known about a specific process; e.g. the good intern's conditional probability table would be the identity matrix while the bad intern's table would give less credence to the information produced. Figure 3 contains samples of all three conditional probability tables for the "Embassy Forward" process.

Default		
E1\E2	T	F
T	.9	.1
F	.1	.9

Good Intern		
E1\E2	T	F
T	1	0
F	0	1

Bad Intern		
E1/E2	T	F
T	.8	.2
F	.4	.6

Figure 3: Conditional Probability Tables for Default processes, and modifications when better background knowledge exists for the "Embassy Forward" process. E1 is the Tanzanian Report, E2 is the intern output.

4. Conclusions and Future Work

In this work, we highlight the need to formally model and compute trust utilizing provenance information. Unlike previous works, we rely purely upon the graphical structure contained in the provenance store to provide a base assessment of the belief in the final resulting data item. If the user has any extra knowledge about the quality of the processes utilized during transformations, this can be incorporated for a better calculation of belief, but is not required for a basic calculation. This approach has two major benefits. First, it decouples the assessment of belief from any information that is not directly stored in the provenance graph. Second, it gracefully accounts for independent, shared source and single-source reports.

We intend to explore further areas of research. First, we wish to build on this work to refine the model for computing belief based on initial assessments of source quality (separate from the data produced by that source) as well as better automatic computation of process effect on transformed evidence. Second, we will apply these belief models to inform users of the likelihood of different hypothesis. For instance, we could use the belief in the reports, as discussed herein, to propose a hypothesis that better explains the evidence. Finally, we will utilize the belief computed here, along with different disease models to direct an investigator's search for better evidence. For instance, based on the belief calculated for the US Warning of DHF, it would be better for our hiker to search out an additional, independent report, such as one from the World Health Organization.

5. Bibliography

[1] A. Baptista, B. Howe, J. Freire, D. Maier, and a. C. T. Silva, "Scientific Exploration in the Era of Ocean Observatories," *IEEE Computing in Science & Engineering*, vol. 10, pp. 53-58, 2008.

[2] D. Becker, W. Mcmullen, and K. Hetherington-young, "A Flexible and Generic Data Quality Metamodel," 2008.

[3] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom, "ULDBs: Databases with Uncertainty and Lineage," *VLDB Seoul, Korea*, pp. 953-964, 2006.

[4] D. L. Chenyun Dai, Murat Kantarcioglu, Elisa Bertino, Ebru Celikel, Bhavani Thuraisingham, "Query Processing Techniques for Compliance with Data Confidence Policies," in *SDM*, 2009, pp. 49-67.

[5] A. deKeijzer and M. vanKeulen, "Quality Measures in Uncertain Data Management," *Scalable Uncertainty Management*, vol. 4772, pp. 104-115, 2007.

[6] L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: the role of source dependence," in *VLDB*, 2009, pp. 550-561.

[7] Y. Gil and D. Artz, "Towards content trust of web resources," *Web Semant.*, vol. 5, pp. 227-239, 2007.

[8] J. Golbeck, "Trust on the World Wide Web: A Survey," *Found. Trends Web Sci.*, vol. 1, 2006.

[9] O. Hartig and J. Zhao, "Using Web Data Provenance for Quality Assessment," in *Proceedings of the 1st International Workshop on the Role of Semantic Web in Provenance Management (SWPM) at the International Semantic Web Conference*, 2009.

[10] D. V. Lindley, "Scoring Rules and the Inevitability of Probability," *International Statistical Review*, vol. 50, pp. 1-26, 1982.

[11] P. Missier, K. Belhajjame, J. Zhao, and C. Goble, "Data lineage model for Taverna workflows with lightweight anotation requirements," Data lineage model for Taverna workflows with lightweight anotation requirements, 2008.

[12] L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, and P. Paulson, "The Open Provenance Model," University of Southampton 2007.

[13] A. Nierman and H. V. Jagadish, "ProTDB: Probabilistic Data in XML," *28th VLDB Confernece, Hong Kong, China*, pp. 1-10, 2002.

[14] N. Prat and S. Madnick, "Measuring Data Believability: A Provenance Approach," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*: IEEE Computer Society, 2008.

[15] T. Stef-Praun, B. Clifford, I. Foster, U. Hasson, M. Hategan, S. Small, M. Wilde, and Y. Zhao, "Accelerating Medical Research using the Swift Workflow System," *Health Grid*, 2007.

[16] A. Tversky and D. Kahneman, *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press, 1982.