# Dependency Discovery in Data Quality

Daniele Barone[1], Fabio Stella[2], and Carlo Batini[2]

[1] Department of Computer Science, University of Toronto, Toronto, ON, Canada
barone@cs.toronto.edu
[2] Department of Informatics, Systems and Communication,
University of Milano-Bicocca, Milano, Italy
{stella,batini}@disco.unimib.it

**Abstract.** A conceptual framework for the automatic discovery of dependencies between data quality dimensions is described. Dependency discovery consists in recovering the dependency structure for a set of data quality dimensions measured on attributes of a database. This task is accomplished through the data mining methodology, by learning a Bayesian Network from a database. The Bayesian Network is used to analyze dependency between data quality dimensions associated with different attributes. The proposed framework is instantiated on a real world database. The task of dependency discovery is presented in the case when the following data quality dimensions are considered; accuracy, completeness, and consistency. The Bayesian Network model shows how data quality can be improved while satisfying budget constraints.

**Keywords:** Data quality, Bayesian networks, Data mining.

## 1 Introduction

In the last two decades, research and practical efforts to improve data quality did not recognize, with the exception of a few works (e.g., *logical interdependence analysis* [1], *tradeoff analysis* [2,3,4,5,6] and *data dependency analysis* [7]), the relevance of studying and analyzing potential dependencies among data quality dimensions, namely, correlations and reciprocal influences among them. To give an example, data that are up-to-date (thus having a low currency) have a high chance to be incorrect too.

Nowadays, the issue of data quality and of dependencies among quality dimensions is gaining more and more importance in the life of organizations. For example, concerning those Information Systems used for decision-making and problem solving in business, i.e., *Decision Support Systems* (DSSs) and *Management Information Systems* (MISs), it is well known [8] that their effectiveness is strictly related to the quality of information resources involved in the decision making process. Making correct decisions is clearly dependent on the quality of data used [9]; however, complete knowledge regarding the quality of data cannot be gained without knowing what the existing relationships are among data quality dimensions. In fact, as shown in the following examples, dimensions

B. Pernici (Ed.): CAiSE 2010, LNCS 6051, pp. 53–67, 2010.

can be strictly related to each other and dependencies among them can play an important role in a decision making process:

- *Accuracy and Timeliness*: often the information becomes better over time, i.e., more *accurate*. However, it is also possible that for a given context, the information becomes less relevant and critical over time [2]. Consider an air traffic control center which receives data from several controller stations. To regulate air traffic, the traffic control center has to cope with uncertain data. Thus, the decision process must balance the delay in receiving more accurate data of airplane positions and the critical period of time in which an "effective" decision must be made to regulate traffic;
- *Consistency vs Completeness*: often the information is based on incomplete but consistent data or on complete but less consistent data [4]. A classic situation concerns human resource data in which different evaluators (i.e., data sources), provide information regarding a particular employee. Those evaluators could have used different attributes to characterize an employee and could have evaluated him over a potentially extended time period. In a scenario where the goal is to promote one out of several employees to a position in senior management, the decision process can use different strategies; it can use all available data even though some items are inconsistent, or only use recent data obtained by a common evaluator.

Therefore, taking into account data quality dependencies is a basic ingredient for rationale decision making and activity planning. Moreover, the knowledge of data quality dependencies can also be extremely important for improvement activities; in fact, it contributes to: i) *diagnose* which is the most probable cause of the bad quality for the considered data quality dimensions and thus helps to identify error sources in an Information System; ii) *select* the most effective data quality improvement strategy, i.e., the one which maximizes data quality when we are subject to budget constraints. Finally, since the quality of data quickly degenerates over time[1], a complete "knowledge" of quality issues represents a fundamental set of *quality requirements* in those (Evolution) Information Systems [11] in which the capability to actively react to organization changes must also take into account data quality problems[2].

This paper presents the Dependency Discovery in Data Quality (D$^3$Q) framework which extends, from bivariate to multivariate, the data-driven analysis in [7]. As shown in Fig. 1, the D$^3$Q framework is a flexible component that can be added as an extra layer to whatever data quality assessment solution is available. It provides a "comprehensive" data quality knowledge for supporting improvement activities. The results provided by the assessment activities (*Assessment Knowledge* (AK)), are exported into the *Dependency Discovery Model* (DDM)

[1] Experts say that 2% of the records in a customer file become obsolete in a month because customers die, divorce, marry and move [10].

[2] Data entry errors, systems migrations, and changes to source systems, generate bucket loads of errors. Thus, the knowledge of the cause-effect structure helps to diagnose what the error sources are.

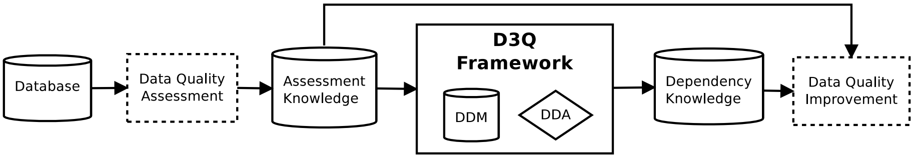

**Fig. 1.** The D$^3$Q framework

which feeds the *Dependency Discovery Algorithm* (DDA) implemented by a data mining algorithm to provide the *Dependency Knowledge* (DK). The assessment activity plays a critical role for an "effective" discovery of dependencies, since the benefit of the proposed approach is directly influenced by the results obtained during the assessment phase. This paper focuses on the data quality dependency discovery, while assessment that received much attention from the data quality community is not discussed. The interested reader can refer to [12] for a rich and comprehensive survey on data quality assessment methodologies and techniques. Bayesian Networks have been used to implement the DDA component. The D$^3$Q framework has been tested on a real world database, namely the Italian Social Security Contributors' Anagraph database, which is assessed along the three most important data quality dimensions, i.e., syntactic accuracy, completeness, consistency, and the most adopted metrics[3]. However, it is worthwhile to mention that it can be applied to any given number and type of data quality dimensions.

The rest of the paper is organized as follows. Section 2 describes the D$^3$Q framework together with its main components, i.e., the DDM and the DDA. Section 3 reports on the instantiation of D$^3$Q framework to the selected database, namely the Italian Contributors' Anagraph database. This section describes the main features of the Italian contributors' anagraph database, the related assessment activities and the results of a rich set of numerical experiments concerning the task of D$^3$Q. The last section is devoted to conclusions and directions for further research.

## 2 The D$^3$Q Framework

The proposed framework aims to discover the dependency structure of the assessed data quality dimensions for a set of attributes belonging to a target database. In this paper the authors focus their attention on the case where the data quality dimensions are described by means of binary variables, i.e., (true,

[3] In [12], about 70% of the reviewed methodologies use syntactic accuracy, and about 60% adopt the NULL value as a completeness measure. Moreover, these dimensions and the associated metrics are also part of the ISO/IEC 25012:2008 - SQuaRE - Data Quality Model [13], in which they appear in the first three positions of the list of dimensions belonging to the model.

false). However, it is possible to extend the proposed approach to include the case where the data quality dimensions are represented through discrete multi-value and/or continuous variables.

### 2.1 The Dependency Discovery Model

The Dependency Discovery Model receives the results of the data quality assessment component in input in order to build the corresponding learning dataset $LD^{DDM}$, i.e., the dataset used to discover the dependency structure of the assessed data quality dimensions. To clarify how the DDM is used to solve the D$^3$Q problem and how the $LD^{DDM}$ is obtained, we need to introduce several quantities. Let $\mathbf{X} = \{x_i, i = 1, ..., N\}$ be the set of instances of a database consisting of $M$ attributes $A = \{A_1, ..., A_M\}$ on which $K$ data quality dimensions $D = \{d_1, ..., d_K\}$ have been assessed. Furthermore, let the data quality dimension $d_j$ for the attribute $A_l$ be described by means of a binary variable $Y_{(l,j)}$. Then, for the instance $x_i$ we have that $Y_{(l,j)}^{(i)} = true$ in the case where the value of the attribute $A_l$ for the instance $x_i$, $A_l(x_i)$, is correct w.r.t the data quality dimension $d_j$. The $DDM(\mathbf{X}) = \{Y_{(l,j)}^{i}(\mathbf{X}) \mid i = 1, ..., N, l = 1, ..., M, j = 1, ..., K\}$ model consists of $N$ realizations of $M \cdot K$ binary variables $Y_{(l,j)}$. The learning dataset $LD^{DDM}(\mathbf{X})$ (hereafter $LD(\mathbf{X})$) is obtained exploiting the results provided by the assessment component.

### 2.2 The Dependency Discovery Algorithm

The Dependency Discovery Algorithm is implemented through a Bayesian Network (BN) which exploits the information stored using the DDM, i.e., the learning dataset $LD^{DDM}$, to discover the dependency structure. Bayesian Networks (BNs) implement an optimal trade-off between complexity and interpretability when we have to cope with highly dimensional domains. BNs are an alternative to rule based models. BNs differ from rule based models in three ways: i) instead of modeling the knowledge of the domain expert, they model the domain; ii) instead of using a non-coherent uncertainty calculus tailored for rules, they use classical probability calculus and decision theory; iii) instead of replacing the expert, they support her/him. The main problem of rule based models is coherence. Indeed, rule based models do not deal properly with uncertainty which naturally arises for the following reasons; observations may be uncertain, the information may be incomplete, the relations in the domain may be of a non-deterministic type. A way to incorporate uncertainty in rule based models is to extend the production rule to include the concept of certainty for both the left and the right part of the rule. The rule based model must be extended with new inference rules, which shall ensure a coherent reasoning under uncertainty. However, it is not possible to capture reasoning under uncertainty with inference rules. Inference rules are context free; while coherent reasoning under uncertainty is sensitive to the context in which the certainties have been established [14].

A BN $\mathbf{B}$ consists of $n$ discrete random variables $X_1$, ..., $X_n$ and an underlying Directed Acyclic Graph (DAG) $G = (V, E)$, where $V$ is the set of vertices while

$E$ is the set of directed links. Each random variable is uniquely associated with a vertex of the DAG. The BN model $\mathbf{B}$ is fully specified by means of the DAG $G$, together with a set of conditional probability tables $P(X_i|pa[X_i])$, $i = 1$, ..., $n$, where $pa[X_i]$ denotes the parents of node $X_i$, i.e., the set of variables which directly influence the random variable $X_i$. The main characteristic of the BN model $\mathbf{B}$ is that the joint probability distribution for the random vector $(X_1, ..., X_n)$ can be represented through the following factorization:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa[X_i]). \tag{1}$$

In the case where the random variable $X_i$ has no parents (no directed links oriented towards the node associated with $X_i$), $P(X_i|pa[X_i])$ is simply its marginal probability $P(X_i)$. One of the interesting features of BNs is that one can infer conditional variable dependencies by visually inspecting the DAG and exploiting the concept of conditional independence [15]. Another interesting aspect is that many algorithms are available for learning their structure from data [16]. The main structural learning algorithms, namely PC and NPC [17], are based on making dependence tests that calculate a test statistic which is asymptotically chi-squared distributed when assuming (conditional) independence. If the test statistic is large for a given independence hypothesis, the hypothesis is rejected; otherwise, it is accepted. The probability of rejecting a true independence hypothesis is given by the *level of significance*. Heckerman [17] provides a full discussion of how the Bayesian estimation approach can be exploited to construct BNs from data. It is worthwhile to mention that inference and learning algorithms are available for multi-valued and continuous random variables. However, for multi-valued random variables the complexity grows exponentially with the number of parents for each node. While for continuous random variables, inference and learning are restricted to gaussian and conditionally gaussian distributions.

## 3 Dependency Structure Discovery

This section is devoted to present the Italian Social Security Contributors' List database together with the corresponding BN model, learnt from the available learning dataset. The BN model is queried to show how it can be used for inference on data quality dimensions dependency. A simple example illustrating how the BN model could be exploited for data quality improvement is given.

### 3.1 Italian Social Security Contributors' List

The *Italian Social Security Contributors'List* (ISSCL) database, maintained by the Italian social security administration, contains data related to Italian contributors on social security for retirement funds. The ISSCL database consists of the following six attributes; *Social Security Number/Value Added Tax Number*, *Juridical/Physical Person Name*, *Street*, *ZipCode*, *Country*, and *Region*. A portion of the ISSCL database, containing 1,483,712 records, is used; while for privacy reasons, only the following three attributes ($M = 3$): *ZipCode*, *Country*, and *Region*, have been included in the dataset $\mathbf{X}$. Three data quality dimensions ($K = 3$) have been considered: namely *accuracy*, *completeness* and *consistency.* The *(syntactic) accuracy* has been assessed by using a *comparison function*, which computes the distance between the value $v$ of a given attribute and the true values belonging to a domain look-up table. A relevant example of comparison function is offered by the Edit distance [18]. Through the DDM we addressed only exact matching by means of binary variables. However, it is worthwhile to mention that it is possible to use a threshold value to deal with approximate matching. The assessment of the *completeness* dimension is based on i) the close world assumption [19], which states that only the values actually present in a relation, and no other values, represent true facts of the real world, and ii) the relational data model with $NULL$ values. The *consistency*, which is defined as the logical coherence of different information [20], has been assessed by finding the violations of semantic rules. A semantic rule is defined over data items (a set of), where the items can be either tuples (e.g., $x_i$) of relational tables or records belonging to a given file. In regard to $\mathbf{X}$, the following business rule is applied:

$$\textbf{IF } \textit{ZipCode}(x_i) = z \textbf{ THEN } (\textit{Country}(x_i) = u \; AND \; \textit{Region}(x_i) = v), \quad (2)$$

which means that if, for a given tuple $x_i$, the attribute *Zip* equals $z$, then the values of *Country* ($u$) and *Region* ($v$) are univocally determined by means of a domain look-up table.

### 3.2 Bayesian Network Learning and Inference

The meanings of the binary random variables: *CO-acc, CO-com, RE-acc, RE-com, ZIP-acc, ZIP-com and* Cons, belonging to the learning dataset $LD(\mathbf{X})$, are described in Table 1. It is worthwhile to mention that the variable *Cons* refers to the triplet of attributes (*Country*, *Region*, *ZipCode*) and measures the consistency of the entire tuple by using the business rule (2). The BN modeling tasks, structural learning and inference, have been accomplished by using Hugin Ver. 6.3. The structural learning tasks, performed by using PC and NPC with the value of the *Level of Significance* parameter set to 0.01, resulted in the same BN model (Fig. 2). The BN model represents the dependency structure of

**Table 1.** Variables for the ISSCL dataset

| Name | Meaning |
|---:|:---|
| *CO-acc* | Country accuracy |
| *CO-com* | Country completeness |
| *RE-acc* | Region accuracy |
| *RE-com* | Region completeness |
| *ZIP-acc* | ZIP code accuracy |
| *ZIP-com* | ZIP code completeness |
| *Cons* | tuple consistency |

the assessed data quality dimensions for the considered subset of attributes in a compact and efficient way. The model summarizes **qualitative** and **quantitative** knowledge extracted from the considered database. These two kinds of knowledge are the basic elements of the DK component in Fig. 1. The qualitative knowledge is associated with the graph component of the BN model, i.e., the DAG; while the quantitative knowledge is obtained through the inferential process over the nodes of the BN model.

**QuaLitative Knowledge (QLK)** is read from the DAG by exploiting the property of the *Markov blanket* of a node, i.e., the set of its parents, children and nodes with which it shares children. Indeed, the *Markov blanket property* states that a node is independent from the rest of the BN's nodes when the state of its *Markov blanket* is known.

*Example 1. An example of qualitative knowledge extracted from the analyzed database is as follows: the variable Cons is conditionally independent from the variables* $ZIP - com$*,* $CO - com$ *and* $RE - com$*, given the variables* $ZIP - acc$*,* $CO - acc$ *and* $RE - acc$*. This means that whenever the states of the variables ZIP-acc, CO-acc, and RE-acc are known, the knowledge about the state of one or more of the following variables ZIP-com, CO-com and RE-com brings no information about the variable Cons. Conditional independence is symmetric. Thus, the variables ZIP-com, CO-com and RE-com are conditionally independent from the variable Cons, given the variables ZIP-acc, CO-acc and RE-acc.*

*Comments.* The above example makes explicit what is known from theory, i.e., the completeness property is a necessary condition for the accuracy property. In fact, we can evaluate the syntactic accuracy of a value only in the case where we

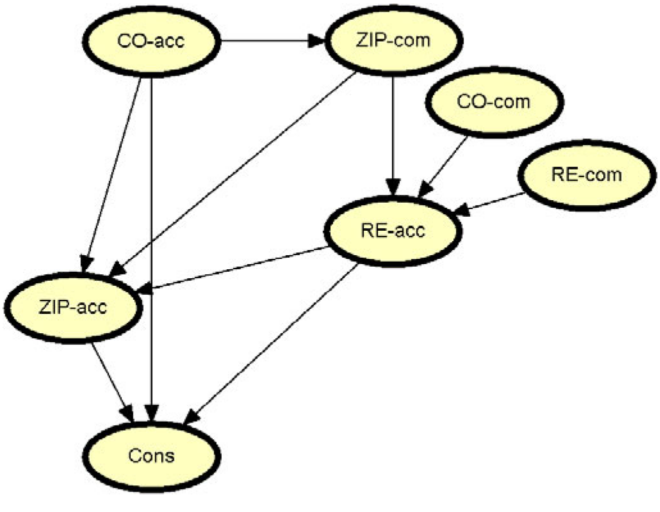

**Fig. 2.** BN model of the ISSCL dataset

have some information to assess. Therefore, it is also true that a value which is not complete is also not accurate. Notice that, a missing value does not always represent an error of completeness. For instance, in a table of employees that has a column for the name of the employeers's manager, that value will be missing for the president of the company. In this case, no value is applicable for the president and therefore, no information can be assessed along the accuracy dimension. However, in this scenario, the NULL value can be seen as complete information (in referring to the semantic expressed by the employees table) but also accurate, though we have no information to assess.

*Example 2. A second example of qualitative knowledge, which refines what presented through Example 1, is as follows. Cons is conditionally independent from* $CO - com$ *and* $RE - com$ *given the variables* $ZIP - com$ *and* $RE - acc$*, while it is not conditionally independent from* $CO - acc$ *and* $ZIP - acc$*.*

*Comments.* The above example enforces the statement we provided about the completeness property as a necessary condition for the accuracy property. In fact, $Cons$ becomes conditionally independent from $ZIP - com$, $CO - com$ and $RE - com$, if we already have some accuracy values for $ZIP - acc$, $CO - acc$ and $RE - acc$, i.e., we had some values (complete information) on which to perform activity assessment along the accuracy dimension.

**QuanTitative Knowledge (QTK)** is accessed through the inferential process applied to the BN model. The inferential process allows to quantify multivariate relationships between data quality dimensions, i.e., to quantify how data quality dimensions relate to each other and impact on each other. The inferential process includes the computation of the most probable configuration of the BN model, the computation of the posterior distribution for a set of nodes when some evidence is available, as well as the computation of the most probable cause for the available evidence. This last type of inferential instance is known as BN diagnosis and is particularly useful for selecting the correct inspection policy when dealing with decision making under uncertainty. By inspection policy we mean the process through which we acquire information on the BN variables to discover what the cause of the available evidence is.

*Example 3. The first information extracted from the BN model depicted in Fig. 2 concerns the most probable joint assignment of the states for the variables, i.e., the most probable configuration of a database record. This joint assignment is (CO-acc=true, CO-com=true, RE-acc=true, RE-com=true, ZIP-acc=true, ZIP-com=true, Cons=true), and it occurs with probability* 0.743995*. Thus, the probability that a tuple belonging to the considered database is affected by at least one data quality problem equals* 0.256005*.*

*Comments.* The above example shows how the QTK easily answers simple but important questions, such as: “What is the probability that a tuple belonging to the database is wrong?”. From a business point of view, if we consider the following example, the relevance of the answer becomes clear. In fact, referring

**Table 2.** Posterior probability $P(RE - acc, RE - com|CO - acc, CO - com)$

| (CO-acc, CO-com) | (RE-acc, RE-com) | $P(RE - acc, RE - com\|CO - acc, CO - com)$ |
|---|---|---|
| (true,true) | (true,true) | 0.964357 |
| | (true,false) | 0.000000 |
| | (false,true) | 0.035640 |
| | (false,false) | 0.000003 |
| (false,true) | (true,true) | 0.963785 |
| | (true,false) | 0.000000 |
| | (false,true) | 0.036213 |
| | (false,false) | 0.000002 |
| (false,false) | (true,true) | 0.399998 |
| | (true,false) | 0.000000 |
| | (false,true) | 0.600000 |
| | (false,false) | 0.000002 |

**Table 3.** Posterior probability $P(CO - acc, CO - com|RE - acc, RE - com)$

| (RE-acc, RE-com) | (CO-acc, CO-com) | $P(CO - acc, CO - com\|RE - acc, RE - com)$ |
|---|---|---|
| (true,true) | (true,true) | 0.962532 |
| | (true,false) | 0.000000 |
| | (false,true) | 0.037464 |
| | (false,false) | 0.000004 |
| (false,true) | (true,true) | 0.961879 |
| | (true,false) | 0.000000 |
| | (false,true) | 0.038064 |
| | (false,false) | 0.000057 |
| (false,false) | (true,true) | 0.962522 |
| | (true,false) | 0.000000 |
| | (false,true) | 0.037476 |
| | (false,false) | 0.000002 |

to a generic selling process which uses customer information to perform transactions, it is possible to estimate the probability of process failure exploiting the probability of a (customer) tuple being wrong.

*Example 4. Table 2 and Table 3 report two examples of the computation of the posterior distribution for a set of nodes when some evidence is available. Indeed, Table 2 (Table 3) reports the posterior probability distribution of accuracy and completeness for the attribute* Region *(*Country*) depending on the accuracy and completeness for the attribute* Country *(*Region*). Therefore, from Table 2 we discover that the probability for* Region *to be jointly accurate and complete (RE-acc=true, RE-com=true) decreases from* 0.964357*, when* Country *is jointly accurate and complete (CO-acc=true, CO-com=true), to* 0.399998*, when* Country *is neither accurate nor complete (CO-acc=false, CO-com=false). Furthermore, from Table 3 we discover that the probability for* Country *to be complete but not* *accurate (CO-acc=false, CO-com=true) increases from* 0.037464*, when* Region *is jointly accurate and complete (RE-acc=true, RE-com=true), to* 0.038064*, when* Region *is complete but not accurate (RE-acc=false, RE-com=true).*

*Comments.* The QTK allows the following considerations: i) if *Country* is complete and accurate then *Region* tends to be accurate; ii) if *Country* is not complete and therefore not accurate, the quality of *Region* significantly decreases; and iii) the probability for *Country* being complete but not accurate is slightly affected by the accuracy of *Country*, given that *Country* is complete. In fact, a possible case for i) is to suppose that a sales employee is registering information about a new customer who lives in "Castelfiorentino" in the province of "Florence" and that he/she is able to correctly type "Castelfiorentino" since he/she knows this place; then there is a high probability that he/she correctly inputs the associated *Region* "Florence". Analogous cases can be found for ii) and iii).

*Example 5. The last instance of the inferential process, i.e., BN diagnosis, is probably the most interesting one. Suppose we are concerned with the most probable cause of the inaccuracy for the attribute* Region*, i.e., we want to discover which data quality dimension/s for which attribute/s is associated with the inaccuracy of attribute* Region*. While the prior marginal probability for each variable, is reported in Table 4, the BN model is queried with the following evidence RE-acc=false to compute the posterior probability for the remaining variables CO-acc, CO-com, RE-com, ZIP-acc, ZIP-com and Cons (Table 5). The posterior probability of the variable Cons equals* 1*. This is obvious when recalling the definition of tuple consistency implemented through the business rule (2). The most probable single cause for the inaccuracy of the attribute* Region *is the inaccuracy for the attribute* Country *(*$P(CO - acc = false|RE - acc = false) = 0.038065$*), the second most probable single cause is the inaccuracy for the attribute* ZipCode *(*$P(ZIP - acc = false|RE - acc = false) = 0.028000$*) while the third most probable single cause is the incompleteness for the attribute* ZipCode *(*$P(ZIP - com = false|RE - acc = false) = 0.000624$*) and so on. However, according to posterior probability values listed in Table 5, CO-acc=false is about one and a half times more probable than ZIP-acc=false, which in turn is about fifty times more probable than ZIP-com=false. Therefore, it is very likely that by inspecting the attribute* Country*, we find it to be accurate, i.e., CO-acc=true. Then, we can think that the next two data quality dimensions to inspect are the accuracy and the completeness for the attribute* ZipCode*. However, this is not the correct database inspection policy. Indeed, we must introduce the new evidence into the BN model to compute the new posterior probability for the non evidenced variables Cons, ZIP-acc, ZIP-com, RE-com and CO-com (Table 6). Under the new evidence (RE-acc=false,CO-acc=true), we deduce that the attribute* ZipCode *is complete, i.e.,* $P(ZIP - com = false|RE - acc = false, CO - acc = true) = 0$.

*Comments.* The BN diagnosis process suggests what the right inspection policy is to discover the cause of the observed evidence, lack of data quality. From a data quality analyst point of view, this can be a useful instrument for discovering cause-effect patterns in order to identify the most relevant sources of errors.

**Table 4.** Prior probability

| VARIABLE | $P(Variable = false)$ |
|---|---|
| CO-acc | 0.037487 |
| CO-com | 0.000003 |
| RE-acc | 0.035663 |
| RE-com | 0.000004 |
| ZIP-acc | 0.168496 |
| ZIP-com | 0.000022 |
| Cons | 0.256005 |

**Table 5.** Posterior probability for the evidence $[RE - acc = false]$

| VARIABLE | $P(Variable = false \| RE - acc = false)$ |
|---|---|
| Cons | 1.000000 |
| CO-acc | 0.038065 |
| ZIP-acc | 0.028000 |
| ZIP-com | 0.000624 |
| RE-com | 0.000113 |
| CO-com | 0.000057 |

**Table 6.** Posterior probability for the evidence $[RE - acc = false, CO - acc = true]$

| VARIABLE | $P(Variable = false \| RE - acc = false, CO - acc = true)$ |
|---|---|
| Cons | 1.000000 |
| ZIP-acc | 0.013515 |
| RE-com | 0.000113 |
| CO-com | 0.000057 |
| ZIP-com | 0.000000 |

### 3.3 Data Quality Improvement

The DK component

- assists the data quality expert in discovering which the most probable *sources of non quality* are. An example is described in [7], in which a positive association between the *timeliness* of the Standard & Poor's (S&P's) and Moody's data sources has been discovered. This association was caused by an inappropriate loading process for the considered data sources; the loading process, responsible for feeding the internal database by using the S&P's and Moody's data sources, was not updated with the same frequency as that of the external bond rating data provider;
- allows the selection of the most effective improvement activity. If the syntactic accuracy of $Y$ depends on the syntactic accuracy of $X$, then it could be the case that improving the quality of $Y$ will result in a quality improvement for $X$, while improving the quality of $X$ does not necessarily bring an improvement of the quality for $Y$;
- implements the correct process for decision making; it avoids redundant activities to be performed and thus minimizes the improvement costs.

To clarify how the DK can be used to improve data quality lets us present the following simple example. Assume we want to maximize the probability of *Consistency* for the ISSCL database. Suppose, for any attribute belonging to the database, we can force the accuracy to hold true. However, to force accuracy to be true an amount of money has to be paid, where the amount of money depends on the attribute to act on. Furthermore, it is usual to impose an upper bound on the budget to maximize the probability of *Consistency.* The amount of money required to force accuracy on the three attributes of the ISSCL database is reported in Table 7, while the budget for each tuple of the database is set to 2.20€. If the accuracies for the considered attributes are assumed to be independent, then the probability of *Consistency* could be maximized by means of a simple ranking procedure. Indeed, we can compute the conditional probability of *Cons* to be *true* when the accuracy for the considered attribute is set to *true.* These conditional probability values are ranked in Table 8. Information in Table 7 and in Table 8 allow to conclude that the solution which maximizes the probability of *Consistency* consists of forcing *ZIP-acc=true* and *CO-acc=true.* This solution costs 1.50€+ 0.60€= 2.10€< 2.20€, which satisfies the budget constraint, while enforcing also *RE-acc=true* further maximizes the probability of *Consistency* but violates the budget constraint (1.50€+ 0.60€+ 0.40€= 2.50€> 2.20€). However, this solution (*ZIP-acc=true*,*CO-acc=true*) results in a probability of *Consistency* equal to 0.926090 ($P(Cons = true|ZIP - acc = true, CO - acc = true) = 0.926090$), as computed through inference on the BN model. Thus, it is not the optimal solution which is obtained from the BN model to be (*ZIP-acc=true*,*RE-acc=true*). This solution is both feasible (1.50€+ 0.40€= 1.90€< 2.20€) and optimal $P(Cons = true|ZIP - acc = true, RE - acc = true) = 0.933681$).

The above example emphasizes the importance of knowing the structure of dependency between data quality dimensions when concerned with decision making under uncertainty. The knowledge of the dependency structure between the *ZIP-acc, ZIP-com, RE-acc, RE-com, CO-acc, CO-com* and *Cons* data quality dimensions allows to efficiently and effectively improve the data quality of the considered database.

**Table 7.** Costs to enforce accuracy for the ISSCL database

| Attribute | Cost |
|---:|:---|
| ZipCode | 1.50€ |
| Country | 0.60€ |
| Region | 0.40€ |

**Table 8.** Sorted conditional probabilities $P(Cons = true|Variable = true)$

| Variable | $P(Cons = true\|Variable = true)$ |
|---:|:---:|
| ZIP-acc | 0.894757 |
| CO-acc | 0.772971 |
| RE-acc | 0.771509 |

# 4 Related Work

The analysis of dependencies between data quality dimensions has been mainly investigated in terms of tradeoff. An example of such an approach is presented in [2], where the authors investigate how the improvement of the timeliness dimension negatively affects the accuracy dimension. The paper presents a theoretical framework to select the most appropriate changes to data managed in information systems. An accuracy-timeliness utility function figures prominently in the analysis, while in the case when it is not possible to determine the utility function, the authors describe how to identify a suitable approximation. A similar framework, that allows the systematic exploration of the tradeoff between completeness and consistency is presented in [4]. The relative weight (importance) of completeness and consistency to the decision maker is an input to the analysis. In order to examine the tradeoff the authors explore various facets of the two dimensions that produce analytical expressions for the measurement activity. The utility of various combinations of completeness and consistency, for fixed and variable budgets, provides guidance to evaluate the appropriate tradeoff of these factors for specific decision contexts. The main difference between the frameworks presented in [2,4] and our approach, is that $D^3Q$ does not limit to a tradeoff between two dimensions but addresses multi-variate dependencies. Moreover, in [2,4], the user must provide: i) a *weight*, that represents the importance of a dimension versus another dimension, and ii) a *functional relationship*, that defines the binding between the dimensions involved[4]. Instead, the $D^3Q$ is able to learn such information (probabilities) directly from the data. In [5], a practical case of tradeoff among timeliness and other generic data quality dimensions in the context of the Bureau of Labor Statistics Covered Employment and Wages is described. The goal is to determine if data quality decreases as a result of receiving data earlier than the current due date. No significant quality deterioration is detected. Although this work describes a real case where dependencies play an important role, a general framework is not provided. In [21] a rigorous and pragmatic methodology for information quality assessment, called AIMQ, is presented. The AIMQ methodology consists of three components: i) a model representing what information quality means to information consumers and managers; ii) a questionnaire for measuring information quality along the dimensions; iii) analysis techniques for interpreting the assessments gathered by the questionnaire. The results of the proposed methodology highlight the fact that information quality is a single phenomenon where dimensions are not inherently independent. The table of correlations among dimensions, calculated on the basis of answers, is reported in the paper. The AIMQ methodology shows how knowledge on dependencies can be exploited to offer a comprehensive quality solution; however, the approach we presented here is more powerful than the one based on *correlation* [21]. In [1], within the context of business scenarios,

[4] When the user cannot provide such functional relationship, the framework suggests the use of generic families of functions to approximate the required true functions; but still, functional parameters must be provided by the user.

a set of logical interdependencies among dimensions is presented. The authors define a taxonomy for data quality dimensions composed of *direct* and *indirect* attributes. Direct attributes represent the main dimensions which directly influence the results of business operations when a change in their values occurs. The indirect attributes determine and also contribute to the direct attributes; hence indirectly influence the results. The $D^3Q$ could be used in a complementary way with [1] to evaluate if such logical interdependencies are satisfied in a real scenario and, therefore, allowing to validate such taxonomy. Finally, a data-driven tool for data quality management is described in [22] which suggests rules and identifies conformant and non-conformant records. The authors focused on the discovery of context-dependent rules, namely conditional functional dependencies (CFDs), i.e., that hold only over a data portion. The tool outputs a set of functional dependencies together with the context in which they hold. To avoid returning an unnecessarily large number of CFDs a set of interest metrics are evaluated, and comparative results using real datasets are reported.

## 5 Conclusions

A framework to discover the dependency structure between data quality dimensions is described. The Dependency Discovery Algorithm uses the learning dataset, compiled by the Dependency Discovery Model, to learn the BN model for the data quality dimensions assessed on a real world database. The BN model implements the Dependency Knowledge component. Accuracy, completeness, and consistency are assessed on three attributes of the Italian Social Security Contributors's List database. The framework is general while the obtained BN model is context dependent, i.e., it depends on the specific database which has been analyzed. The obtained results allow to conclude that BNs

- provide the database analyst with an intuitive and efficient representation of the dependency structure between data quality dimensions;
- allow consistent evaluation of the data quality level associated with each tuple of the database to be certified;
- allow to compare alternative data quality improvement strategies, to perform costs/benefits analysis and thus to implement optimal decision making.

Directions for future work concern the evaluation of the proposed approach on a richer set of data quality dimensions and on a larger number of attributes. It is relevant to study how BNs can be exploited to develop effective data quality improvement strategies in an information system, by implementing a trade-off between the *cost of non quality* and the budget available for data quality improvement.

## Acknowledgments

The authors are grateful to the anonymous referees whose insightful comments enabled to make significant improvements to the paper. Daniele Barone would like to acknowledge the many valuable suggestions made by Fabrizio De Amicis, BSI Bank.

## References

1. Gackowski, Z.: Logical interdependence of some attributes of data/information quality. In: Proc. of the 9th Intl. Conference on Information Quality, Cambridge, MA, USA, pp. 126–140 (2004)
2. Ballou, D.P., Pazer, H.L.: Designing information systems to optimize the accuracy-timeliness tradeoff. Information Sys. Research 6(1), 51–72 (1995)
3. Han, Q., Venkatasubramanian, N.: Addressing timeliness/accuracy/cost tradeoffs in information collection for dynamic environments. In: Proc. of the 24th IEEE Intl. Real-Time Systems Symposium, Washington, DC, USA, p. 108 (2003)
4. Ballou, D.P., Pazer, H.L.: Modeling completeness versus consistency tradeoffs in information decision contexts. IEEE Trans. Knowl. Data Eng. 15(1), 240–243 (2003)
5. Sadeghi, A., Clayton, R.: The quality vs. timeliness tradeoffs in the BLS ES-202 administrative statistics. In: Federal Committee on Statistical Methodology (2002)
6. Fisher, C., Eitel, L., Chengalur-Smith, S., Wang, R.: Introduction to Information Quality, p. 126. The MIT Press, Poughkeepsie (2006)
7. DeAmicis, F., Barone, D., Batini, C.: An analytical framework to analyze dependencies among data quality dimensions. In: Proc. of the 11th Intl. Conference on Information Quality, pp. 369–383. MIT, Cambridge (2006)
8. Burstein, F. (ed.): Handbook on decision support systems. Intl. handbooks on information systems. Springer, Heidelberg (2008)
9. Berner, E., Kasiraman, R., Yu, F., Ray, M.N., Houston, T.: Data quality in the outpatient setting: impact on clinical decision support systems. In: AMIA Annu. Symp. Proc., vol. 41 (2005)
10. Eckerson, W.: Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. Technical report, The Data Warehousing Institute (2002)
11. Oei, J.L.H., Proper, H.A., Falkenberg, E.D.: Evolving information systems: meeting the ever-changing environment. Information Sys. Journal 4(3), 213–233 (1994)
12. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3), 1–52 (2009)
13. International Organization for Standardization: Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – data quality model. In: ISO/IEC 25012 (2008)
14. Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer, Heidelberg (2001)
15. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco (1988)
16. Baldi, P., Frasconi, P., Smyth, P.: Modeling the internet and the WEB: Probabilistic methods and algorithms. Wiley, Chichester (2003)
17. Heckerman, D.: A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research (1995)
18. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Trans. Knowl. Data Eng. 19(1), 1–16 (2007)
19. Reiter, R.: On closed world data bases. In: Logic and Data Bases, pp. 55–76 (1977)
20. Jarke, M., Jeusfeld, M., Quix, C., Vassiliadis, P.: Architecture and quality in data warehouses: an extended repository approach (1999)
21. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: AIMQ: a methodology for information quality assessment. Information Management 40(2), 133–146 (2002)
22. Chiang, F., Miller, R.J.: Discovering data quality rules. PVLDB 1(1), 1166–1177 (2008)