ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

A. Avenali, C. Batini, P. Bertolazzi, P. Missier

BROKERING INFRASTRUCTURE FOR MINIMUM COST DATA PROCUREMENT BASED ON INFORMATION QUALITY - QUANTITY MODELS

R. 614 Settembre 2004

Alessandro Avenali – Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza", Roma, Italy. Email: avenali@dis.uniroma1.it.

Carlo Batini – Dipartimento di Informatica e Sistemistica, Università di Milano Bicocca, Milano, Italy. Email: batini@bicocca.mi.it.

Paola Bertolazzi – Istituto di Analisi dei Sistemi ed Informatica, Consiglio Nazionale delle Ricerche, Roma, Italy. Email: bertola@iasi.rm.cnr.it.

Paolo Missier – School of Computer Science, The University of Manchester, Manchester, UK. Email: Missier@cs.man.ac.uk.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

Abstract

Inter-organization business processes involve the exchange of structured data across information systems. We assume that data are exchanged under given condition of quality (offered or required) and prices. Data offer may include bundling schemes, whereby different types of data are offered together with a single associated price and quality. We describe a brokering algorithm for obtaining data from peers, by minimizing the overall cost under quality requirements constraints. The algorithm extends query processing techniques over multiple database schemas to automatically derive an integer linear programming problem that returns an optimal matching of data providers to data consumers under realistic economic cost models.

1. Introduction

For large businesses and public sector agencies, good management of information assets has long been a key to their effectiveness in delivering quality services to users, and many organizations have processes to manage the quality of their data.

Recently, advances in the technology for large-scale deployment of information services, for example over service-oriented software infrastructures, have enabled cost-effective data exchange across organizations. In business terms, this means that it is becoming increasingly feasible for organizations to (i) purchase or otherwise acquire data from other peers, and (ii) exploit their own information assets for marketing purposes. These capabilities may be used to offer advanced services to users.

Example: A local council that relies on families' addresses to deliver information on child benefits, may obtain accurate and up-to-date street information from a third-party source such as the postal service, as well as occupational information from employment agencies.

Thus, a general common trend is for organizations to acquire the information needed to support user services from third-parties. Several studies have analyzed the economic relevance of the potential information market. Public agencies have been found to be the greatest producers of information by far, and the information they create and disseminate is often relevant for both the private and public processes, products, and services. In [3] an analysis of the commercial exploitation of *public sector information* is presented both for the USA and the European Union (EU). The study shows that the economic value of the information market in the EU for year 2000 amounted approximately to 10% of that of the US, where it was 750 billion dollars, and it recommended regulating the information market, to provide further incentives for the public sector information trading across and within member states. With the final goal to improve this kind of market in the EU, rules for managing the reuse of information owned by public sector bodies of the member states have since been issued [4].

To understand the implications of this trend, the size of the information market must be compounded with the issue of its *quality*, as a factor that will presumably affect the cost of data and hence the overall information market. Quality of data has been an issue since the nineties. General frameworks are available from the literature for describing data quality properties, or *dimensions* [38, 5, 54, 7]. For instance, *accuracy* characterizes how well data represents its corresponding real-world entities. Another main issue concerned with information market is represented by offering *bundles* of data, which are indivisible units of data, each one with a single associated price and quality level. In fact, both the cost structure behind the production and the selling of digital information goods, and the necessity of implementing anti-competitive strategies can induce more and more data providers to offer indivisible units of different types of data (for example [20] and [24, 25]).

In order to introduce a motivating example, we focus our analysis on the public sector. It is well known that public agencies, in order to provide services to citizens and businesses, manage large registries with overlapping and heterogeneous data, and exchange large amounts of data flows.

Example: In countries where agencies are organized in several administrative tiers, different registries are managed at the different tiers:

- at the *central level* usually dozens and even hundreds of registries exist on individuals, usually covering part of the population, such as tax payers, workers, retired personnel, etc; several overlapping registries exist also for businesses, social insurance, chambers of commerce, etc.

- at the *regional level* there are local income tax payers registries, health care registries, etc.
- at the *province level* (present in some countries as an intermediate level between regions/districts and municipalities) we may assume that scholarship registries are held.
- in every *municipality* a personal data registry for resident people and a separate registry for the civil status of resident people may be managed.

Such a huge number of registries, from one side is characterized by a high overlap, from the other side they are usually managed and updated with different policies, resulting in different levels of accuracy and other quality dimensions. In many data intensive processes sources are combined, and it is important for agencies and private users to be able to choose and compose data on the basis of the desired target quality. In other terms, the availability of such overlapping sources of data may be seen as an opportunity for the data demand, that may use a *quality driven query processing* strategy [45] that builds the global data set on the basis of the differentiated offer of data characterized by different qualities. Furthermore, the quality of data has a cost, and, at the same time, heavily influences the quality, the cost, and the revenues of the processes that use the data. While considering the relationship between the quality and cost of quality issues, some authors start their analysis from a parallel between the emerging information market and established markets for other goods [8, 9], with the final purpose of defining criteria for data quality control and improvement. These activities, like for other types of goods, have a cost which is a component of the selling price. Furthermore, in order to conceive rational methodologies for improving the quality of data, several authors have proposed data quality cost classifications [15] and [16] and *cost/quality optimization procedures* [14] that investigate the various different types of cost of non quality of data, and of data quality improvement activities. As a consequence, costs can be quantified and optimized with reasonable approximation. Issues of quality driven query processing and cost/quality optimization have been addressed only recently so far.
In this paper we propose a *brokering algorithm* that provides a *cost quality broker service* for facilitating the procurement of data from third parties, based on the assumptions that consumer interest for data is based both on its cost and on its quality, and that distinct data can be sold together in a bundle with a single associated quality and price. More precisely, the algorithm (see Figure 1) starting from: (i) the *offer of data* with possible bundling schemes from a set of providers, its quality and cost, (ii) the global, integrated knowledge on the information content offered by providers, and (iii) a query, that expresses the *data demand*, namely data requested by consumers and their quality, provides the optimal choice in terms of selected data, their quality and cost.

We note that the broker service can be used as a decision support system for managers who have the responsibility of information acquisition activities. A straightforward extension of the service could be used in a coordinated spot market to establish prices of data, where suppliers, subordinated and coordinated by a single mediator, simultaneously announce the offered bundles to the mediator with relative associated prices (one for every combination quantity-quality which the bundle can be sold with). Many consumers also simultaneously submit one or more global queries to the mediator by also specifying a reserve price for every submitted global query.

The rest of the paper is organized as follows. In Section 2 the information procurement scenario underlying our approach is presented, together with a motivating example. Furthermore, a first overview of the algorithm is presented, and basic definitions are provided. The two phases of the algorithm, decomposition and optimization, are detailed in Sections 3 and 4, respectively. A discussion on related work is presented in Section 5; Section 6 concludes the paper.

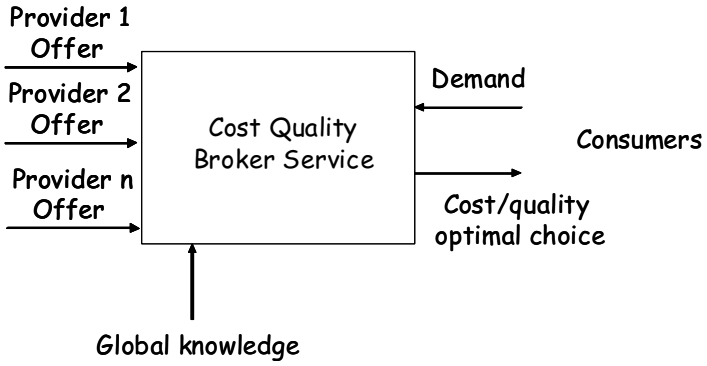

Figure 1: Black box view of the broker service

2. Overview of the approach and basic definitions

2.1. Information Procurement Scenario and Cost Model

The input/output description of the cost quality broker service in Figure 1 is now detailed in terms of an underlying *information procurement scenario* and a *cost model* that we describe in detail. Concerning the information procurement scenario, we assume the following:

- the information market is made of a potentially wide number of organizations; each organization may have a role both as data provider and consumer relative to other organizations;
- providers offer a description of the data they can procure, along with a cost model and the quality of the data offered;
- the data offer that providers can procure is made available by them in terms of bundles of data (see Section 1);
- consumers express their data demand as queries, along with constraints on the minimal acceptable quality level;
- for each type of data of interest to a consumer, multiple providers that are capable of fulfilling at least part of the demand may exist.

A distinguishing feature of our information procurement scenario is the concept of bundle and the *cost model* adopted for it. Several economic reasons motivate providers to offer bundles of data. Digital information goods are typically characterized by high costs in the production and promotion of the first copy (high fixed cost of development), while additional copies are cheap to reproduce (low marginal cost). These goods are termed *non-rival*, because one's consumption does not limit the consumption of others. Since data have negligible marginal costs, the pricing strategy of bundling is profitable [17, 18, 19]. In fact, offering bundles can result in cost savings due to economies of scale/scope [20]. Bundles can also be used to provide discounts to consumers who acquire two or more (complementary) information goods. Moreover, bundling is a selling strategy that allows producers to reduce the uncertainty associated in the consumer's willingness to pay for individual goods. It has been shown [21] that the willingness to pay for bundles exhibits lower variance than for separate individual items. Thus, the producer is able to extract more

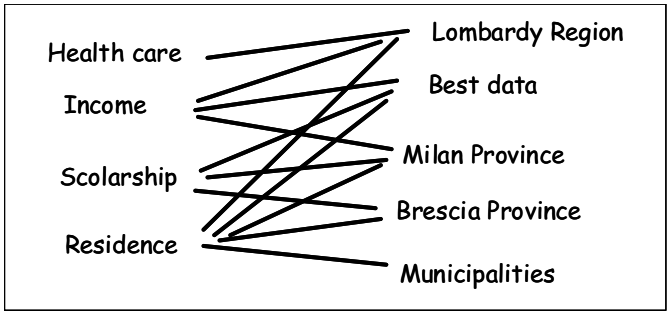

Figure 2: Providers and procured areas

surplus from the consumers. By adding new items to a bundle, a provider can also reduce the risk involved in offering consumers new types of goods; therefore, the provider can stimulate the sale of new complementary goods.

Bundling results in a profitable pricing strategy even when the willingness to pay for goods is independent of the possible consumption of other goods in the bundle, and when there is no costs saving [22, 23]. In fact, incumbents often make use of this selling practice [24, 25], since, for instance, bundling allows operators with significant market power to introduce new entry barriers without lowering the prices of the individual goods in a form of anti-competitive behavior; [26]. For instance, depending on the reserve prices of two information goods, a provider could gain much more in cross-subsidizing these goods by selling them in bundle, rather than separately (pure bundling).

2.2. Running Example

We assume that in the public sector scenario proposed in the introduction, several types of data related to citizens are procured by public or private providers as a result of elaborations on registries owned by public agencies or else specific inquiries. We focus in the following on a specific region, Lombardy, of a specific country, Italy. We assume for the sake of simplicity that the Lombardy region has two provinces, Milan and Brescia. Data sets concern different areas of interest, such as income, health care, scholarship, and province of residence location. The data sets are offered by different providers. Among them three are public, namely the *Lombardy region*, the *provinces of Milan and Brescia*, and, potentially, the roughly 800 *municipalities* present in Lombardia. Besides the public providers, a hypothetical company *Bestdata* sells data, on the basis of a processing performed on data acquired from public providers. The correspondence between areas of interest and providers is represented in Figure 2.

Due to the many to many correspondence between providers and areas, the different providers own different sets of properties (usually called *attributes* in the relational database terminology [36]) related to citizens, the correspondence is shown in Figure 3, where attributes are also numbered with identifiers that will be used in the following for brevity.

Data sets are provided in bundles, whose shape is decided by providers on the basis of economic issues discussed previously. Figure 4 shows possible examples of bundles provided, in terms of the provider offering it, and the scope of the bundle within the data set.

Furthermore, the quality of properties in data sets and, consequently, in bundles, depends on

Property/Provider	1. PersonId	2. Province	3. Age	4. Income	5. Pathology	6. LastClassAttended
Lombardy region	X	X	X	X	X	
Best data (only for age > 18)	X	X	X	X		X
Milan province	X	X	X	X		X
Brescia province	X	X	X			X
Municipalities	X	X	X			

Figure 3: Providers and procured attributes

Provider	Scope
Lombardy region	Single province
Lombardy region	PersonId, Age, Pathology
Bestdata	The whole data set
Bestdata	Single province
Bestdata	PersonId, Age
Bestdata	PersonId, Age, LastClassAttended
Milan Province	The whole data set
Brescia Province	The whole data set
Municipalities	The whole data set

Figure 4: Providers and bundles

Property/Provider	1. PersonId	2. Province	3. Age	4. Income	5. Pathology	6. Last Class attended
Lombardy region	100%	93%	85%	90%	99%	
Best data (only for age > 18)	100%	98%	98%	98%		85%
Milan province	100%	99%	95%	96%		96%
Brescia province	100%	99%	99%			99%
Municipalities	100%	98%	99%			

Figure 5: Providers, attributes and qualities (level of completeness) offered

the provider and on the attribute, as shown in Figure 5. In the figure we show realistic values of accuracy that measures, in this case, the percentage of correct data in the data sets.

The assumption that the quality varies with the provider and the attribute is reasonable, since the quality of a set of data depends heavily on the process followed for acquiring the data through surveys or elaborations on external sources, and productions of them. Finally, providers sell the bundles at a given price, fixed according to their internal production cost, market size and expected revenues.

We assume that several private companies are interested in getting data sets with quality greater than a given threshold for their own purposes. For instance, franchising companies could be interested in planning products and services to sell to individuals in the whole Lombardy, on the basis of `Age` and `Income`, but only in case the overall accuracy of such attributes in the result is grater than 95 per cent. A further goal is to achieve these data and threshold qualities at the minimum price. As another example, a company selling e-learning courses for permanent education could be interested in purchasing data related to the `LastClassAttended` at the best price, only for citizens with a secondary school degree, provided that the accuracy is greater than 90 %.

2.3. Algorithm Overview

A structured view of the brokering algorithm is shown in Figure 6, where the two main phases of the algorithm are highlighted, *decomposition* and *optimisation*. We now provide a high level description of: (i) the inputs to the algorithm, (ii) the two phases, and (iii) the intermediate products exchanged by the two phases.

The underlying data architecture is grounded in the framework of federated database systems with a mediated query processing architecture [28]. We assume that data is described using the relational model, that each provider manages a local relational schema, and the local schemas are defined as mappings over a common *global schema*, that represents the whole available information content in an integrated view.

In this setting, the global schema is used by consumers to express their demand. The global query processor includes a *mediator*, an architectural component with functionality to recognize which of the local sources may contribute to the result of the global query and to translate the global query into a collection of local queries to be issued to the local sources. Each local query result represents a partial contribution to the global query result. The mediator formulates a query plan that includes the execution of the local queries, and then starting from these results, it produces a single consistent result to be delivered to the user. Quality constraints from

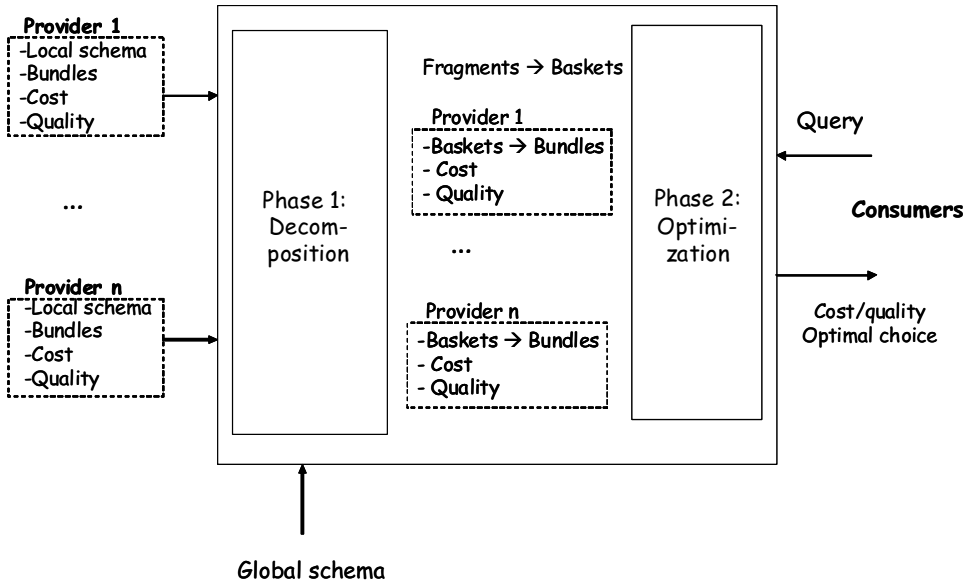

Figure 6: input/output two phase view of the brokering algorithm

consumers are expressed in the global schema, while the quality and cost of the available data are associated to the local relations. Bundles, whose economic relevance have been discussed in Section 1, will be formally defined in the next section.

The *decomposition phase* accepts and interprets a consumer query and related quality constraints. It decomposes each local query result into a set of *fragments*, in such a way that fragments from different local sources can be compared. This homogeneity allows classes of equivalent fragments, called *baskets*, to be defined across all participating providers. Fragments and baskets will be formally defined in Section 2.5. At this point, the issue of provider-defined bundles as indivisible sets of data is considered; given a configuration of bundles available from multiple providers. The bundles are mapped to sets of fragments within each basket, along with their associated cost and quality.

The *optimization phase* formulates an *integer linear programming (ILP) model* to compute a cost-optimal combination of those partial results that is as complete as possible and satisfies all the constraints. The information on fragments, baskets, and bundles is translated into variables for an ILP model, in which:

- variables are introduced to represent membership of fragments to baskets;
- quality constraints are mapped to constraints in the space of the program variables;
- completeness constraints are introduced using the baskets, to model coverage of the global result in terms of fragments;
- bundle constraints are introduced to ensure that only allowed combinations of fragments are selected;
- the objective is to minimize a cost function.

The optimization phase of the algorithm solves an ILP problem on this solution space, by minimizing the cost, taking the constraints on quality and bundling into account. As we will see in the rest of the paper, the decomposition phase includes algorithms that are time-polynomial under suitable and realistic assumptions on the query conditions, while the optimization phase solves an ILP problem, which is NP-hard in general, relying on both efficient exact and heuristic procedures.

The results of this work extend and generalize those presented in [27]. In particular, our algorithm features the following innovative aspects:

- a query decomposition algorithm that is grounded in well-known research on distributed database systems [2], [1] and federated database systems with a mediated query processing architecture [28];
- an algorithm that, starting from the partial results, automatically synthesizes the constraints and the objective function of a ILP;
- the new cost model considers a more general offer scenario.

As far as the first and the second items are concerned, it is well known that the formulation of an optimization model is not an easy task, due to the lack of competence in most private and public organizations. In this paper we show that this task can be automated. Concerning the cost model, we extend the model proposed in [27], where only one price is associated with each offered bundle, by allowing suppliers to apply discounts in the case that multiple copies of the same bundle are demanded and/or lower quality levels for the data in the bundle are required.

2.4. Schema Mappings and Query Rewriting

In this section we introduce the concepts of global schema, local schema, and the type of mapping we define among them, according to the local as view (LAV) model. We make use, as in the rest of the paper, of a slightly simplified version of the running example of Section 2.2.

In our setting the *global schema* is a relational schema R in the relational model [36], defined over a set of attributes $\mathcal{A} = A_1, \ldots, A_n$. We will assume, without loss of generality, that the *primary key* of the global schema is the first attribute A_1.

In the running example, the global schema includes all attributes of section 2.2:

```
AllRes(PersonId, Province, Age, Income, Pathology, LastClassAttended)
```

where the underlined attribute `PersonId` is the primary key.

The basic idea of the local as view (LAV) model [29, 28, 30] is that each local schema is defined as a *view mapping*, that is, a relational expression on the global schema: $\pi_{\mathcal{A}_L}(\sigma_p(R))$, where symbol $\pi_{\mathcal{A}_L}$ denotes a *projection* relational operator defined on the set of attributes $\mathcal{A}_L$ of local schema L and symbol $\sigma_p(R))$ denotes a *selection* operator defined on the predicate p. The pair $[P, A_L]$ will be called in the following *selection projection condition* or, simply *condition*. Furthermore, we make the assumption that, given pk, the primary key for R, for every pair of local schemas L_1 and L_2, the condition

$$pk \subseteq \mathcal{A}_{L_1} \wedge \mathcal{A}_{L_2}$$

must hold. This means that, in order to merge data referring to the same tuples in the global schema, it is always possible to join tuples of different views through the same key.

Provider Id	Provider	Content	Local Schema	Mapping	
				Predicate p	Attribute set A_L
MP	Milan Province	Residents of Milan Province	Res-Mi(PersonId, Province, Age, Income, LastClass Attended)	Pro = 'MI'	12346
BD	Best Data	Adults (Age >= 18)	Adults(PersonId, Province, Age, Income, LastClass Attended)	Age >= 18	12346
LR	Lombardy Region	All residents in Lombardy	AllRes(PersonId, Province, Age, Income, Pathology)	true	12345
BP	Brescia Province	Residents of Brescia Province	Res-Bs(PersonId, Province, Age, LastClass Attended)	Pro = 'BS'	1236

Figure 7: The four local schemas of the running example

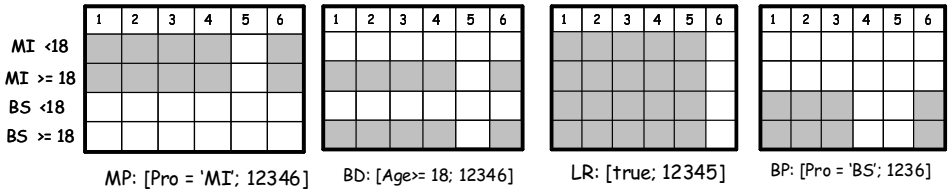

Figure 8: Set oriented representation of the four schemas

To understand the idea underlying the LAV model, imagine that the global schema has been materialized, and that the following query is issued:

$$V = \pi_{Personid,Province,Age,Income,LastclassAttended}(\sigma_{Prov='MI'}(\text{AllRes}))$$

By saying that the query V is the LAV mapping from the global schema *AllRes* to a local schema *Milan province*, we say that the extension of *Milan province* can be obtained by computing V on the global schema *AllRes*. Of course, since in reality the materialized schema is that of *Milan province*, we may interpret the mapping as a specification of the contribution of a local schema to a possible materialization of the global schema.

In our running example, the four local schemas, corresponding to Milan province, Bestdata, the Lombardy region, and the Brescia province rows of Figure 3, are described in Figure 7 in terms of a content description, the local schema, and the view mapping, expressed in terms of the condition $[p, \mathcal{A}_L]$.

In Figure 8 we show the local schemas by means of a set oriented representation, where the part of the global schema present in the local schema is in gray; for example the local schema *MP* is defined for cells where the predicate Province = 'MI' holds, and only for attributes 12346 (on the left hand side of the figure the predicates corresponding to relevant groups of tuples appear with a shorthand notation).

The literature on the LAV approach, cited above, states that, given a set of mappings, the answer to a query Q requires a process of rewriting so that Q is expressed solely in terms of the mapping-defining views. The main theoretical results indicate that (i) query processing is NP-complete in the worst case [32], and (ii) because the mappings can be incomplete, the goal of rewriting is to compute a maximal subset of the complete result, rather than a provably complete one (for instance [29, 31] for a thorough description of the problem). In brief, there are

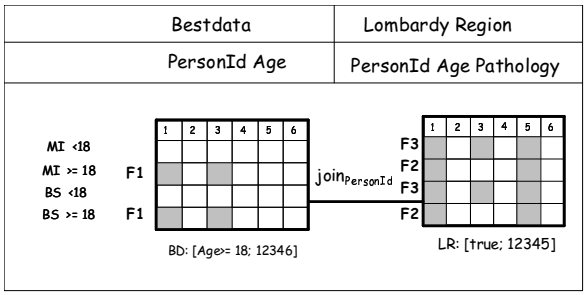

Figure 9: Fragment combinations selected in the second query

two main sources of complexity for this problem: (i) there are an exponential number of query rewritings, and (ii) testing containment for one such rewriting is itself NP-complete with respect to the length of the query. In our formulation, however, we make the simplifying but realistic assumption that none of the queries contain repeated predicates, making the problem linear. We further restrict queries to *conjunctive predicates* that are either (i) relational operators on ordered domains, of the form x *relop* c, of the types $=, \neq, <, \leq, >, \geq$ with c constant, or (ii) disjunctive predicates over set membership expressions, that is, $x =' c'$.

2.5. Fragments and Baskets

We investigate the issues related to query construction, and express the demand of data, defining the concepts of fragment and basket. Consider the following global query:

$$Q \equiv \pi_{PersonId,Age,Pathology}(AllRes) \tag{1}$$

and assume that the LAV mappings have been defined as in Figure 7. Several options exist for composing a global result from the local schemas. First, one can simply issue a single query to provider LR alone, as follows:

$$\pi_{PersonId,Age,Pathology}(LR) \tag{2}$$

because LR provides all the required attributes and tuples.

A logically equivalent result can also be obtained combining *fragments* of local schemas, that is groups of values belonging to a set of attributes and a set of tuples of some local schema. For example, the following fragments can be calculated that may be combined to obtain the complete result:

$$F_1 = \pi_{PersonId,Age}(BD)$$

$$F_2 = \pi_{PersonId,Pathology}(\sigma_{Age \geq 18}(LR))$$

$$F_3 = \pi_{PersonId,Age,Pathology}(\sigma_{Age<18}(LR))$$

The result is expressed (see Figure 9 with the intermediate results labeled as F_1, F_2, and F_3) as:

$$(F_1 \bowtie_{PersonId} F_2) \cup F_3$$

For this composition to be feasible, we make use of the assumption made formerly, that local schemas include the common key `PersonID` as part of its attributes. Quality and cost are the criteria that in our approach drive the choice. In the example, we could prefer the query (2) if

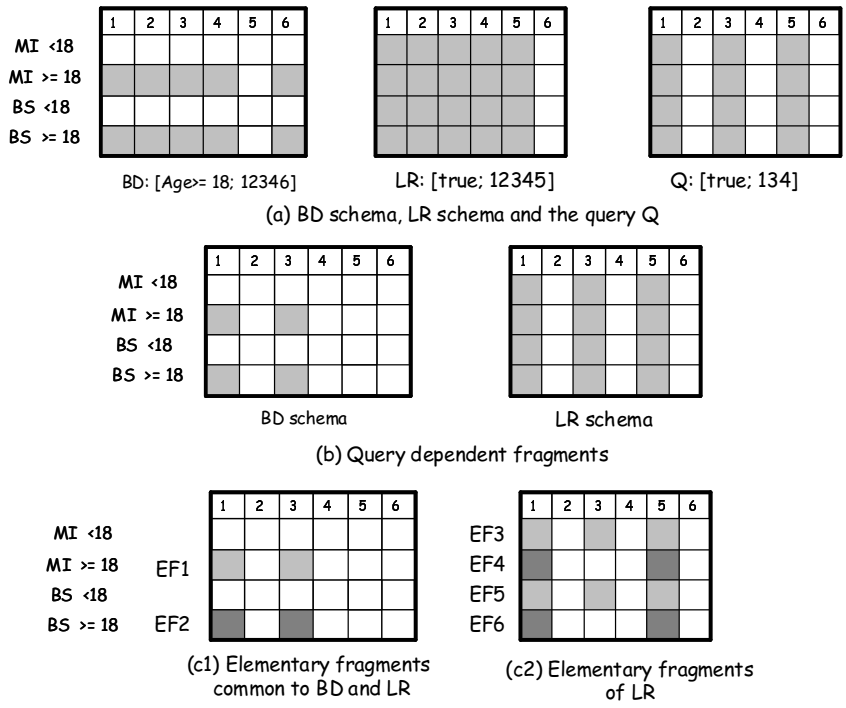

Figure 10: Fragments resulting from comparison of BD and LR fragments in Step2

we know that provider LR owns good quality data for attributes `Age` and `Pathology`, and the cost of these data is not prohibitive. The second choice is preferred if we know that provider BD, for resident person with `Age` ≥ 18 bears ages of residents with better data qualities than LR and reasonable cost.

The space of all possible combinations of fragments can be visualized by overlapping the local schemas, based on their LAV mappings and the global query expression. We are looking here for *elementary fragments*, that is fragments that may make a contribution to the query and that it is not worthwhile to further decompose. In Figure 10 we see the elementary fragments result of the comparison of: (i) the LR local schema, (ii) the BD local schema, and (iii) the query Q. Figure 10(a) shows the two schemas and the query Q; Figure 10(b) represents the fragments of the two schemas that independently may make a contribution to the query. Finally, Figure 10(c) shows the elementary fragments in two sets: on the left hand side we see the fragments which are common to the two schemas, on the right hand side we see the fragments which belong to a single schema, namely the schema LR.

Notice that fragments EF_1 and EF_2 in Figure 10(c) are offered by both providers, while exactly one fragment from either of the two schemas should be used in a possible query to compose the result. Considering for example EF_1, we refer to the corresponding condition $[Pro = MI \wedge Age \geq 18; 13]$, generalized as $[p; A]$ as a set of (two) fragments called *basket*, whose members are the two logically equivalent fragments

$$\pi_{PersonId,Age}(\sigma_{Pro=MI}(BD))$$

and

$$\pi_{PersonId,Age}(\sigma_{Pro=MI \wedge Age \geq 18}(LR))$$

Thus, the definition of a basket is intensional (a condition) and applies to several local schemas, while a fragment is a collection of (parts of) tuples obtained from the application of the basket expression to a specific schema.

2.6. Representing Quality and Bundles

We extend our formalism to deal with data qualities and the demand and offer of data. Data quality deals with a wide number of different *dimensions*, that express properties of data. The most investigated dimensions are *accuracy* and *completeness*. Accuracy, as we have seen in Section 1 measures the closeness of the actual data values to their exact values. Completeness measures the extension of data in representing the real world. Other dimensions are currency, timeliness, consistency. Definitions of such dimensions can be found in [5]. Measures or *metrics* can be assigned to dimensions: in relational tables, dimensions, and corresponding metrics, they can refer to tuples (e.g. a null value in a tuple results in low completeness) or else to the entire relation (e.g. only 80 % of the tuples are correct, resulting in an accuracy equal to 0.8), or else to attributes defined in the relation; for example if we know that values of the attribute `Age` are specified only in 90 % of cases, *null* otherwise, the value of completeness is 0.9. In this paper, for generality purposes, dimensions refer to attributes. Furthermore, we assume that for a relation with attributes $A_1, \ldots, A_m$, the data quality of the set of attributes is represented as a m-tuple q of vectors $q_i = (q_{i1}, \ldots, q_{in})$, one for each attribute A_i. We assume, as for example in [52], that the primary key has, for all qualities, the topmost quality value. In other words, values of the primary key are always correct and different from *null*. Furthermore, we assume that it is the responsibility of the providers to assess the quality vectors; also, some of the values may be missing from any of the q_i.

We now have to define how data quality is expressed in the demand of data by consumers, and in the offer of data by providers. Consumers express quality constraints on the global schema, alongside the global query. Constraints are of the form q_{ij} *relop* c_{ij}, where q_{ij} refers to the value of quality dimension j for the attribute i and *relop* is a relational operator. For example, considering query (1) above, the expression $q_{Age,completeness} > 0.6$ indicates the threshold value for dimension *completeness* relative to the `Age` global attribute.

From the provider's perspective, data sets and their quality are associated to bundles. A *bundle* is a triple $bu = (c, p, q)$ that specifies a relation the provider is committed to sell as: (i) a condition c on the local schema, expressed, as usual, as a pair $[p, A]$, (ii) its price c, and (iii) its quality vector q. A provider may declare several bundles, possibly overlapping in content. The following are valid bundles for provider *Lombardy region* (for conciseness, we denote here attributes with their identifiers):

i $bu_{LR1} = ([Prov =' MI' \vee Prov =' BS'; 12345], p_1, q_1)$,

ii $bu_{LR2} = ([Prov =' MI' \wedge Age < 18; 12345], p_2, q_2)$,

iii $bu_{LR3} = ([Prov =' MI' \wedge Age \geq 18; 12345], p_3, q_3)$,

iv $bu_{LR4} = ([Age \geq 18; 12345], p_4, q_4)$.

Thus, provider *Lombardy region* offers a bundle for the the whole set of citizens resident in Lombardy, two different bundles for residents of Milano, based on age distinction, and a bundle for adults; in all bundles the whole owned set of attributes is offered. This means that for example a request for Milano residents may be fulfilled entirely by bundle 1, and partially by each of the other bundles.

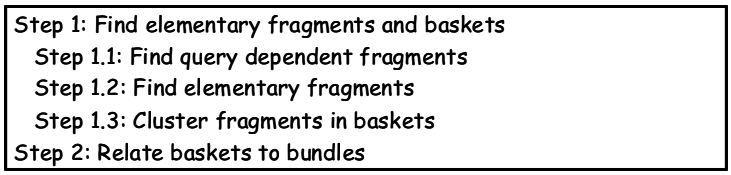

Figure 11: Steps of the decomposition phase

We assume that the quality of attributes in local schemas is homogeneous among the different parts of the local schema. As a consequence, bundles inherit by definition the quality vector associated to the local schema. Finally, quality also influences the composition of fragments in the process of query construction. When composing different relations with given quality dimensions and metrics with a union or join operation, we may wonder if functions exist that allow to automatically compute the values of metrics for the composed relation. Such composition functions have been investigated in the literature (see Section 5). Given two relations r_1 and r_2 with sizes respectively $|r_1|$ and $|r_2|$, in case of the union operator we assume as composition functions for quality dimensions *accuracy* and *completeness* (in short *qd*) the following expression:

$$qd(r_1 \bigcup r_2) = (|r_1| * qd(r_1) + |r_2| * qd(r_2))/(|r_1| + |r_2|)$$

The above formula provides an approximation of the correct value, since in general the two relations may overlap; see Section 5 for a more detailed discussion on comparison functions. In case of join, we simply have to juxtapose the quality vectors of the joined attributes.

3. Decomposition

In this section, we describe the first phase of the brokering algorithm in detail. It consists of the two steps shown in Figure 11. The goal of Step 1 is to find all fragments and corresponding baskets in local schemas that may potentially contribute to satisfy the demand of data, expressed by a query Q.

The goal of Step 2 is to relate the demand of data, expressed by baskets, and the offer of data, expressed by bundles procured by providers. For each basket we have to find all the bundles that contain it, which can be used to satisfy the demand. The relationship between baskets and bundles is the input to the second phase of the algorithm, the optimization phase. Now we describe the two steps in detail, using the global query as an example:

$$Q \equiv \pi_{PersonId,Province,Age,Income,LastClassAttended}(\sigma_{Age \geq 10}(AllRes)) \tag{3}$$

represented, using our shorthand notation for attributes as:

$$Q \equiv \pi_{12346}(\sigma_{Age \geq 10}(AllRes)) \tag{4}$$

3.1. Finding Elementary Fragments and Baskets

We have to remark here that in a traditional optimal plan algorithm for a distributed query [1], where one is interested in minimizing the cost of transmission, we are interested, given a selection projection query Q whose extension is the fragment F, in finding all conditions expressing fragments that *contain* F. Here we have to solve a different task, that is, to find all

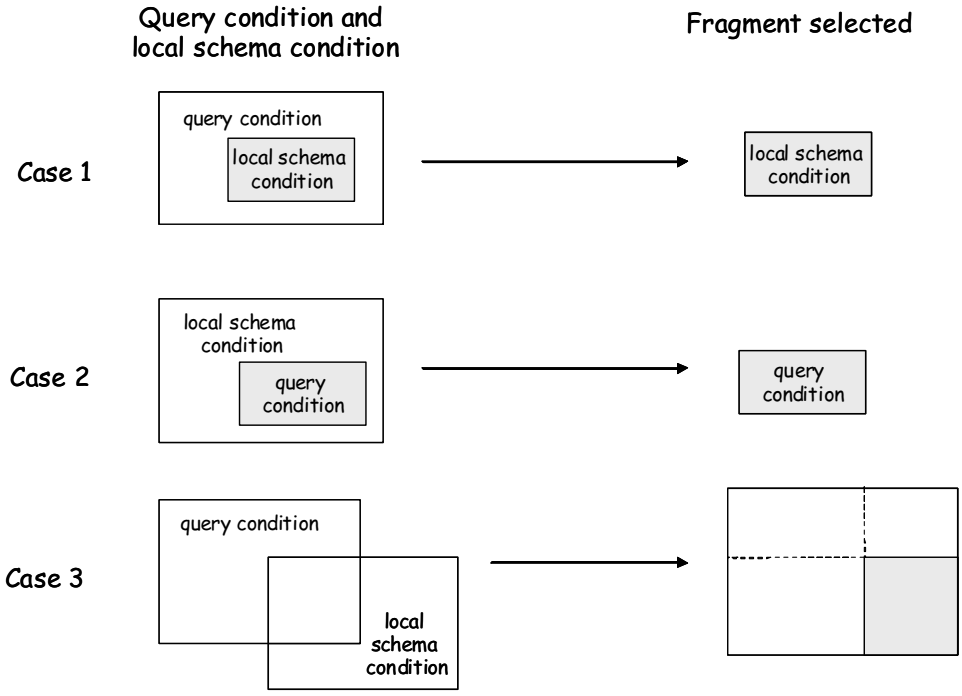

Figure 12: The three cases resulting from the comparison of fragments

fragments that may give a contribution to the final result and consequently *are contained in* F. They can also be optimally *composed*, knowing their cost and quality. This different target is due to the fact that only within this setting we are free to choose the most suitable mix of fragments to fit the quality/cost problem.

The goal of Step 1.1 is to find the set of fragments in the local schema for each local schema that may contribute to the final query. When comparing the query condition and the local schema condition of a local schema V_1, three cases are given, shown in Figure 12:

Case 1 the local schema fragment is contained in the query fragment; in this case the local fragment is selected;

Case 2 the query fragment is contained in the local schema fragment; here the query fragment is selected;

Case 3 there is a non null intersection between the two fragments. This case is expressed more formally in Figure 13, where compared fragments are marked with bold lines and all possible combinations of predicates p_Q and p_1 and set relationships among sets of attributes A_1 and A_Q are shown. In this case only the sub fragment indicated in Figure 13 with *yes* has to be selected, since this is the unique contribution of the local schema to the query.

In Figure 14 we see the fragment corresponding to the query and the four fragments resulting from application of Step 1.1 to the four local schemas. The goal of Step 1.2 is to relate all fragments obtained after Step 1.1 in order to find the *elementary fragments*, namely fragments that (i) contribute to the query, (ii) it is not fruitful to further decompose, and (iii) do not contain other fragments. Also in this step the three cases of Figure 12 are given. Case 1 and Case 2,

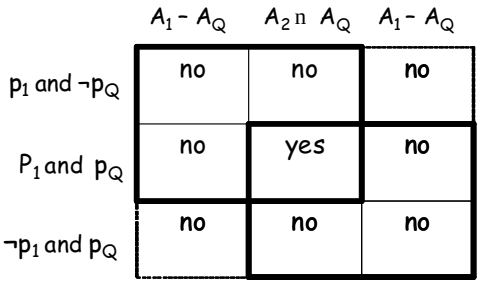

Figure 13: The case of partial overlapping between conditions

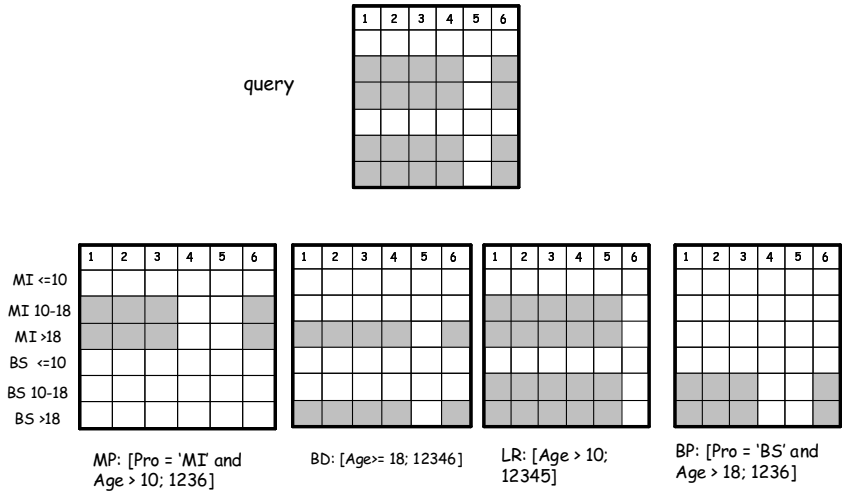

Figure 14: Fragments resulting from Step 1.1

corresponding to one fragment containing the other one, do not result in new fragments. Case 3, corresponding to partial overlapping among fragments has the possible subcases as shown in Figure 15. In this case we have to include as fragment each fragment denoted with *yes* in the figure, since all these fragments are part of the two original fragments, and may contribute to the query result. The two cases denoted as *no* are to be excluded, since they do not belong to either of the original fragments.

In Figure 16 we see the elementary fragments resulting from the comparison of *LR* tp *BD* local schemas, subdivided in common and uncommon fragments.

In Step 1.3 we collect all elementary fragments having the same condition expression $[p; A]$ and consequently, correspond to the same basket. In Figure 17 we show baskets and corresponding providers for the fragments considered in Figure 16. The algorithm requires to test predicate containment, an NP-complete operation in the most general case, which we simplify to linear operations by considering only the type of logical operators already mentioned.

3.2. Relating Elementary Baskets to Bundles

We recall from Section 2.1 that a provider offers its local data in bundles of the form $bu = (c, p, q)$, which specify a data set as a local condition c, its price p and its quality vector q. In *Step 2* of the

	$A_1 - A_2$	$A_2 \cap A_1$	$A_2 - A_1$
p_1 and $\neg p_2$	yes	yes	no
p_1 and p_2	yes	yes	yes
$\neg p_1$ and p_2	no	yes	yes

Figure 15: The case of partial overlapping between conditions of demand fragments

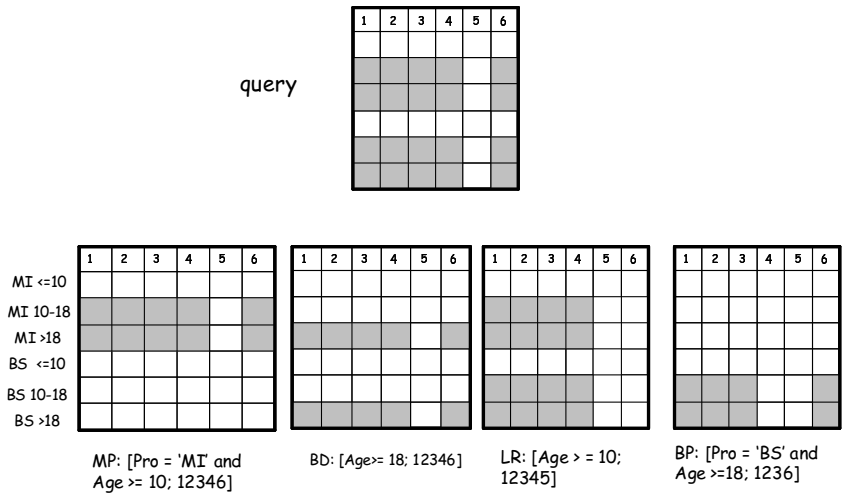

Figure 16: Fragments resulting from comparison of BD and LR fragments in Step 2

Basket	Basket condition	Providers
bk_1	Pro = 'MI' and Age >= 18; 1234	BD, LR
bk_2	'Pro = 'BS' and Age >= 18; 1234	BD, LR
bk_3	Pro = 'MI' and 10 < = Age < 18; 1234	LR
bk_4	Pro = 'MI' and Age >= 18; 16	BD
bk_5	Pro = 'BS' and 10 < = Age < 18; 1234	LR
bk_6	Pro = 'BS' and Age >= 18; 16	BD

Figure 17: Baskets and corresponding providers for the example of Figure 16

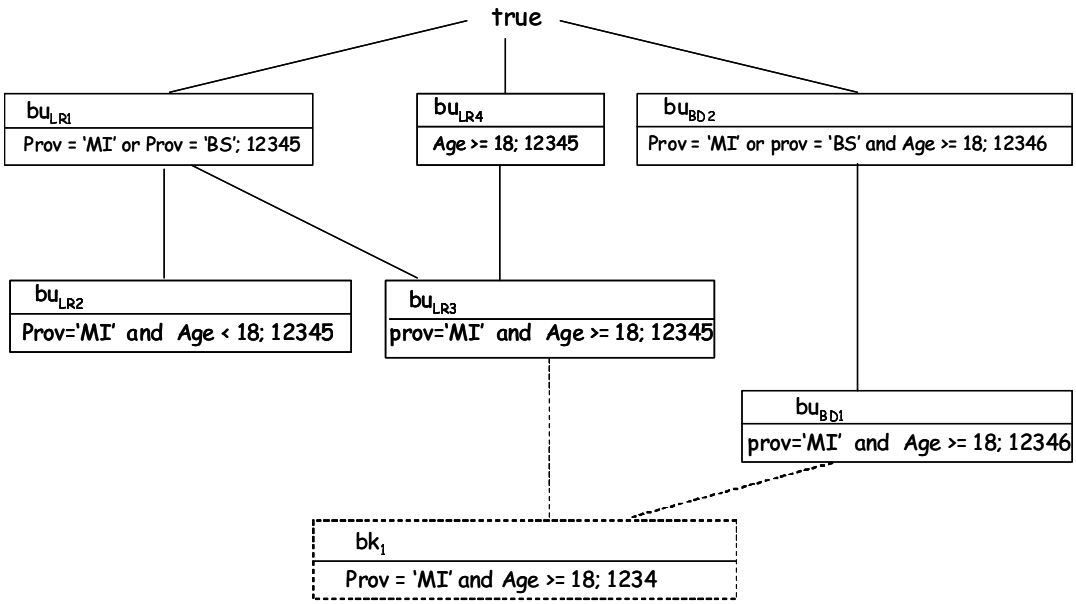

Figure 18: Baskets and their relation with bundles

decomposition procedure we now need to describe their relationship with the basket formulation described so far.

Bundles are atomic units of data that can be purchased, and thus they represent a "business view" of a provider's offering, which is independent of any query. On the other hand, the query processor provides a description of the local data in terms of atomic baskets, determined by the query, from which individual fragments can be selected. Selecting one fragment from each basket provides a coverage of the query result.

These two views are reconciled by mapping a basket onto one or more bundles, so that selecting a basket results in a choice of bundles that can be purchased. Here we make the assumption that for each basket, at least one bundle exists that contains it. This assumption looks reasonable, since providers tend to offer large data sets.

In this section again we use our running example. We consider two providers; let us assume that provider *Lombardy region* sells the four bundles described in Section 2.6 and provider *Bestdata* defines two bundles:

v $bu_{BD1} = ([Prov =' MI' \wedge Age \geq 18; 12346], p_5, q_5)$

vi $bu_{BD2} = ([Prov =' MI' \vee Prov =' BS' \wedge Age \geq 18; 12346], p_6, q_6)$

Consider the basket $bk_1 = [Prov =' MI' \wedge Age \geq 18; 1234]$ owned by providers *Bestdata* and *Lombardy region*. We now build the lattice relating the basket to the bundles containing it, leading to the lattice of Figure 18, where broken lines represent containment relations between the basket and the bundles and unbroken lines represent containment relations between bundles. The lattice represents an example of a *containing bundles* many to many relationship between baskets and bundles. In the example, the algorithm has the following choices for bundles:

- bu_{LR1} from provider *Lombardy region*;

- bu_{LR3} from provider *Lombardy region*;
- bu_{BD1} from provider *Bestdata*;
- bu_{BD2} from provider *Bestdata*;

Notice that bundles bu_{LR1} and bu_{BD2} lead to additional, unrequested data to be purchased with respect to bundles bu_{LR3} and bu_{BD1}, but they may still be convenient in terms of quality and price. Note that the query processor may then filter out the unrequested data, or it may decide to retain it and use it in combination with other bundles.

4. Optimization

As described in previous sections, given a global query, a set of baskets and a set of bundles are returned by the first phase of the algorithm. The set of baskets and corresponding containing bundles describe the relationship between the demand of data and the offer of data. Baskets contain different elementary data fragments, each one belonging exactly to one basket. Two or more fragments supplied by the same provider are in distinct baskets. The mediator allow us to provide a coverage of the global query by selecting exactly one fragment from each atomic basket. Bundles contain baskets and they are computed by the algorithm on the basis of the bundling conditions exposed by providers.

So far, only a quality vector and the related price are considered in the definition of a bundle. Actually, in real scenarios the supplier side can be characterized by discounts for packages of data both in terms of provided quality and number of sold copies. Therefore in the ILP formulation we propose in this section we also take into account quantity discounts. We assume that a bundle is a quadruple $bu = (c, p, q, qt)$, where qt copies of the data represented by the condition c characterized by a quality vector q are offered at price p. However in the running example, for simplicity of notation we will keep on considering only prices related to quality.

In this section we describe the second phase of the algorithm, which consists of first formulating and then solving the problem of selecting fragments across baskets in such a way that quality and quantity requirements are satisfied and the total cost to acquire bundles containing such fragments is minimized. We comment that also this second phase of the algorithm is automatically performed. In particular, as a first step a graph is constructed by taking as inputs the information returned by the first phase of the algorithm, then an ILP formulation is derived from this graph and finally an optimization solver is applied to find an optimal solution.

Selecting a bundle bu means that, for every basket bk in the *containing bundles* relationship with bu, the fragment supplied by the same provider of bu in bk must be chosen. We say that these fragments are *associated* to the bundle bu.

In our framework, when a consumer submits a global query he/she also specifies, for every demanded attribute A_k ($k = 1, \ldots, l$), the quality q_k and quantity qt_k required for A_k. Note that without loss of generality we assume that the consumer demands the first l attributes of the global schema. We denote by D_k the set of baskets whose fragments cover the demanded attribute A_k (i.e. the demanded attribute A_k is provided by selecting exactly one fragment from each basket in D_k). A complete description of the relevant offer for the submitted global query with quality and quantity requirements has to take into account information deriving from local schemas and exposed bundling conditions. We denote the offer scenario by a pair $(\mathcal{BK}, \mathcal{BU})$ where:

- $\mathcal{BK} = \{bk_1, bk_2, \ldots\}$ is the set of the generated baskets, and

- $\mathcal{BU} = \{bu_1, bu_2, \dots\}$ is the set of the offered bundles.

For instance, if we consider the running example, for the data providers *Lombardy region* and *Bestdata* the first phase of the algorithm returns:

- the set of baskets $\mathcal{BK} = \{bk_1, bk_2, bk_3, bk_4, bk_5, bk_6\}$, defined in Section 3.2;
- the set of bundles $\mathcal{BU} = \{bu_{LR1}, bu_{LR2}, bu_{LR3}, bu_{LR4}, bu_{BD1}, bu_{BD2}\}$, defined in Sections 2.6 and 3.2.

Given a scenario demand characterized by A_k, q_k, qt_k for $k = 1, \dots, l$ and the sets $(\mathcal{BK}, \mathcal{BU})$, we want to find a set of bundles satisfying all quality and quantity requirements with the minimum cost. We define this problem as the *minimum cost supplying* problem (MCS for short). We now provide a formulation of MCS. In order to represent the offer and demand scenarios we use a tripartite graph $G = (W, U, V, E)$, where:

- W is the set of vertices representing the *bundles* with the associated price matrices,
- U is the set of vertices representing the *fragments*,
- V is the set of vertices representing the *baskets*,
- $E = E_1 \cup E_2$ is the edge set such that E_1 (E_2) is the set of edges connecting W (V) with U.

The graph is constructed as follows. We associate exactly one vertex in W to every bundle $bu \in \mathcal{BU}$. Furthermore, given the vertex $w_i \in W$, we denote by qt_i, q_i and pr_i, respectively, the related quantity, the quality vector and the price characterizing the bundle bu associated to w_i. For every $w_i \in W$, each fragment associated to the bundle corresponding to w_i is represented by exactly one vertex $u_j \in U$. We recall from Section 2.6 that q_i is a m-tuple of quality vectors, each one associated to a specific attribute of the bundle. Given the fragment corresponding to u_j, we may associate to u_j an m-tuple of quality vectors denoted as $\tilde{q}_j$, obtained projecting q_i on the attributes defined in u_j. In particular, in $\tilde{q}_j$ every demanded attribute A_k in the fragment u_j is characterized by a quality vector $\tilde{q}_{j,k}$ (whose size depends on the number of quality dimensions). However, from now on, for the sake of simplicity and w.l.o.g. we limit the number of quality dimensions to one. Hence, the scalar $\tilde{q}_{j,k}$ is associated to every demanded attribute A_k in the fragment u_j. To keep a trace of the relation between the bundles and the associated fragments, an edge (i, j) between vertices w_i and u_j is added in E_1. Note that, on the supplier side, there is one vertex u for each fragment of every vertex w (this implies that multiple copies of a fragment can be represented in the graph). For each basket $bk_h \in \mathcal{BK}$ a vertex $v_h \in V$ is introduced. An edge (h, j) between vertices v_h and u_j belongs to E_2 iff the fragment represented by u_j is a member of the basket bk_h.

Coming back to the running example, the resulting tripartite graph $G = (W, U, V, E)$ is shown in Figure 19. Observe that, in order to improve the readability of the figure and w.l.o.g., we restrict our running example to two attributes, say A_4 and A_6.

Given the graph $G = (W, U, V, E)$, the MCS problem is formulated as an ILP problem. A binary variable x_i is associated to each vertex $w_i \in W$ and it assumes 1 if the corresponding bundle is selected to be sold, and 0 otherwise. Moreover, a binary variable y_j is associated to each vertex $u_j \in U$ and it is equal to 1 if the corresponding fragment is chosen to satisfy the consumer demand. We also denote the size of the fragment corresponding to u_j with r_j. The problem can thus be formulated as follows:

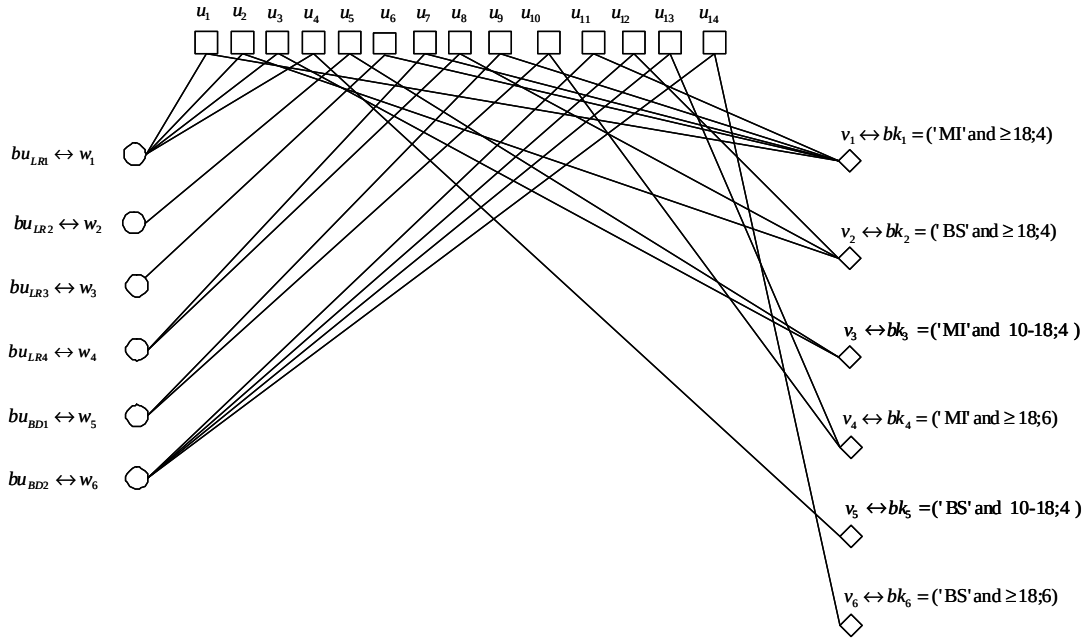

Figure 19: Tripartite graph

$$\min \sum_{i:w_i \in W} pr_i \cdot x_i$$

subject to the constraints:

$$y_j \leq x_i \qquad (i,j) \in E_1 \tag{5}$$

$$\sum_{j:(h,j)\in E_2, h:bk_h \in D_k} y_j = qt_k \qquad k \in \{1,\ldots,l\} \tag{6}$$

$$\sum_{j:(h,j)\in E_2, h:bk_h \in D_k} \tilde{q}_{j,k} \cdot r_j \cdot y_j \geq q_k \cdot \sum_{j:(h,j)\in E_2, h:bk_h \in D_k} r_j \cdot y_j \qquad k \in \{1,\ldots,l\} \tag{7}$$

$$x_i, y_j \in \{0,1\} \qquad i : w_i \in W, j : u_j \in U \tag{8}$$

Constraint (5) means that if a fragment from a bundle is chosen, then the entire bundle is chosen. Constraint (6) says that exactly qt_k fragments containing demanded attribute A_k must be chosen among baskets in order to satisfy the consumer's quantity requirements. Constraint (7) means that if a demanded attribute can be obtained as a merge of fragments (each in a different basket), then the overall quality of the merge is given by the weighted average of the elementary fragment qualities. Moreover, constraint (7) allow us to apply the composition functions discussed for accuracy and completeness in section 2.6. Obviously, in the case that quality dimension is larger than one, constraint (7) must be replicated for each quality dimension. The provided formulation takes inspiration from the mathematical model underlying the facility location problem ([34]) with no connection costs, where a given demand (computed baskets) has to be served by facilities (data fragments) which can be installed in several locations (the data

bundles offered, whose associated prices represent facilities costs); this problem, as known, is NP-hard. Note, however, that data differ from physical commodities in that the "overlapping" of their sources represents an opportunity to create a marketplace by diversifying the offer (e.g. by cost and quality), rather than a potential for wasting resources.

5. Related Work

The impact of information quality on actual decision quality using a theoretical and a simulation model is investigated in [37]. Relevant work has been done toward associating quality properties to relational databases, and using that information during query processing. A distinction should be made between two issues:

- *Quality composition* defines an algebra for composing data quality dimension values, for example given two relations whose completeness values are known. It provides formulae for computing the completeness of their union.

- *Quality-driven query answering* is the task of providing query results on the basis of a quality characterization of data at sources, and possibly cost of data.

As discussed in section 2.6, in this paper we adopt approximate formulae for composition functions, referring to the union and join operators. Quality composition has been addresses in several papers, Motro et al. [40], Wang et al. [54], Naumann et al. [52],Scannapieco et al. [53], Parssian et al. [55], and Florescu et al. [41]. Several characterizations exist for two dimensions, namely accuracy and completeness.

Concerning accuracy, [40] presents a model for estimating the quality of databases. [54] distinguishes between a *relation accuracy* and a *tuple accuracy*. In the hypothesis of the uniform distribution of errors that causes inaccuracy, the tuple accuracy is defined as *probabilistic tuple accuracy*, that coincides numerically with the overall relation accuracy. Several results are provided for selection and projection operators. Results provided in [55] are richer, and concern compositions in terms of cartesian product, projection, selection, and join operations.

Referring to completeness, major contributions are [52], and [53]. In the first paper the authors are interested to evaluate the quality of the process of composing sources, in order to put together data that is split in different sources. For this reason, they are interested in evaluating the behavior of several types of join operators, that extend the usual left/right outerjoin operators of relational algebra with merge features. Formulae for completeness composition functions are provided for several possible set containment relationships among sources. Similar results, extended to the union and difference operators are described in [53], within a different model, in which two different hypothesis are made on the (i) coincidence, or (ii) difference in the domain to which the two sources refer.

Regarding the second topic, that is, quality driven query answering, in the context of mediator-based data integration, Florescu, Koller, and Levy (1997) deal with the completeness problem by introducing various probability distributions regarding the content overlap across multiple database sources, and efficient ways to compute them. These distributions are then used by query planning algorithms, in order to select the most promising sources when answering a global query.

Naumann, Leser, and Freytag (1999) [42] propose extensions to their own local as view-based query processing algorithm by considering quality scores that are associated to the data sources. Quality scores are used during query planning, first to remove (prune) low-quality candidate

plans, thereby reducing the complexity of query optimization, and second to rank candidate plans according to their overall quality. With respect to our research, this work appears to only include a quality scoring model, but not a cost model, hence the selection of the best plan is essentially based on a quality ranking of the candidate plans.

The issue of how to resolve data inconsistencies is also addressed in [45], where data fusion functions are used to produce new versions of data for which different conflicting versions are available. Our own work does not address the issue of resolving inconsistencies among conflicting data, but rather, it exploits redundancy for selecting a cost-optimal combination of alternative and equivalent fragments, having different cost and quality.

The first setting of the quality cost optimization problem is due to Avenali, Batini, Bertolazzi, and Missier (2004). Successively the problem, to the best of our knowledge, has been addressed only in [44]. In this approach, in order to obtain the required data, customers must buy multiple data sets from different providers and then clean and merge them. In this case a broker architecture intermediates between users and syndicated data providers. On the basis of data quality and cost requirements, the broker builds the most suitable data set by integrating data fragments from different providers. In the selection phase, the broker uses optimization and negotiation mechanisms in order to satisfy requirements. The broker is modeled according to the Local-As-View perspective,where the data of a provider are represented as views of a global schema, called broker schema. The broker is in charge of managing the relationship with providers, but has no visibility on data values. The broker is also supposed to receive the average value of the quality of each data set from providers. Providers are responsible for the evaluation of data quality, that is that each provider has implemented data quality tools for the evaluation of data quality along the accuracy, completeness, and timeliness dimensions.

There are several differences between the approach presented in the above cited paper and the one presented in this paper. First, in our paper,in the decomposition phase, all candidate fragments are built, while in [44] fragments are given a priori, and the broker has no other choice than to manipulate them. Second, we have a concept of bundle, as a further constraint in the optimization process. Third, authors in [44] provide a non linear formulation of the problem of identifying a family of data sets to satisfy the query with a minimum overall price. However, the solutions found are not guaranteed to be optimal. In fact, a tabu search scheme is applied. Instead, we propose an integer linear problem which is solved to optimality.

For the optimization phase of the broker service two main areas must be cited. The first concerns the strategy of bundles, that has been discussed in the introduction. The second concerns the algorithms to solve ILPs similar to the one proposed in our paper.

The facility location problem underlying in the structure of the ILP presented in Section 4 (to minimize the cost of data procurement in a given offer scenario) is a widely studied topic in the operations research literature citeMirFra:. There are a number of papers that concern exact methods for this problem [47, 48, 49, 50]. Most of exact methods can be straightforwardly applied to our case.

6. Conclusions and extensions

We have presented a brokering algorithm that supports managers in the process of buying information from multiple data sources, that are characterized by different cost and quality. The algorithm accepts a query over a global schema, as well as the mappings from local to global schema (in a local-as-view setting), and computes the most complete answer to the global query with the best cost-quality ratio.

The algorithm consists of two phases. During the first phase, using the schema mappings, a set of local fragments for the query result are identified. In the second phase a variable is associated to each fragment, while their corresponding quality and cost are used to formulate constraints for an ILP problem. The problem solution contains a complete answer obtained under quality constraints, and at minimal cost.

The first phase includes a specific case of a condition subsumption problem. However the simplifying assumptions on the conditions make it polynomial. Although the second phase is known to be NP-complete, the size of the problem which is determined by the number of providers and of local schemas is expected to be small. The algorithm in practice is meant to be used by decision-makers with the responsibility of acquiring quality data from third parties. The algorithm can be usefully extended to support a *coordinated spot market*, where multiple consumers simultaneously require portions of data with specified quality levels, and multiple suppliers submit their offers and associated quantity-quality matrices to a central public supplier (CPS) mediator. For instance, the CPS might be in charge of selling data owned by multiple local public agencies to individuals, businesses and other public agencies. In such a case, in order to exploit the quantity/quality discounts as much as possible, the CPS could coordinate the purchasing process by collecting and then matching the overall demand and offer. In particular, the problem of allocating offered data among consumers can be formulated as a simple extension of the ILP presented in Section 4. We are interested in implementing the DSS presented in this paper and to develop the whole model underlying the coordinated spot market outlined above.

References

[1] T. Oszu, P. Valduriez, Principles of Distributed Database Systems (Second ed.) Prentice Hall, Englewood Cliffs, N.J., 1999.

[2] S. Ceri, G, Pelagatti, Distributed Databases - Principles and Systems, McGrawHill Inc. New York San Francisco, Washington D.C. (1984).

[3] Pira International, Commercial Exploitation of Europe's Public Sector Information, Final report for the European Commission, Directorate General for the Information Society (October 2000).

[4] Directive 2003/98/ec of the European Parliament and of the Council on the Re-use of Public Sector Information, November 2003.

[5] T. Redman, Data Quality for the Information Age, Artech House, 1996.

[6] R.Y. Wang, M. Ziad, Y.W. Lee, Data quality, Advances in Database Systems, Kluwer Academic Publishers, 2001.

[7] R. Wang, D. Strong, Beyond Accuracy: what Data Quality Means to Data Consumers, Journal of Management Information Systems 12 (4).

[8] R. Y. Wang, M. Ziad, G. Shankaranarayanan, IP-MAP: Representing the Manufacture of an Information Product, in: Proceedings of the Eight International Conference on Information Quality (ICIQ-00), Cambridge, MA., 2000.

[9] D.Ballou, R.Wang, H.Pazer, G.K.Tayi, Modelling Information Manufacturing Systems to Determine Information Product Quality, Journal of Management Sciences, Vol. 44 (4).

[10] I. Fellegi, A. Sunter, A Theory for Record Linkage, Journal of the American Statistical Association, Vol. 64, (1969).

[11] W. E. Winkler, Methods for Record Linkage and Bayesian Networks, Technical Report, U.S. Census Bureau, Statistical Research Division (2002).

[12] A. Borthwick, M. Buechi, A. Goldberg, Key Concepts in the Choicemaker 2 Record Matching System, in: Proceedings of the First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, in conjunction with KDD 2003, Washington, DC, 2003.

[13] M. Eppler, M. Helfert, A Classification and Analysis of Data Quality Costs, in: Proceedings of the Eight International Conference on Information Quality (ICIQ04), MIT, Cambridge, MA, 2004.

[14] D. Ballou, H. Pazer, Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff, Information Systems Research, Vol. 6 (1).

[15] L. P. English, Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, 1st Edition, John Wiley & Sons, 1999.

[16] David Loshin Enterprise Knowledge Management - The Data Quality Approach - Chapter 4, Knowledge intelligence incorporated, 2004.

[17] H. Fang, P. Norman, To Bundle or not to Bundle, SSRI Working Paper 2003-18., Cowles Foundation, Yale University (2003).

[18] C. Brooks, R. Das, J. Kephart, J. MacKie-Mason, R. Gazzale, E. Durfee, Information Bundling in a Dynamic Environment, in: Proceedings of the IJCAI-01 Workshop on Economic Agents, Models, and Mechanisms, Seattle, Washington, 2001.

[19] S. Fay, J. MacKie-Mason, Competition Between Firms that Bundle Information Goods, in: 27th Annual Telecommunications Policy Research Conference, Alexandria, VA, 2001.

[20] G. Ulusoy, K. Karabulut, Determination of the Bundle Price for Digital Information Goods, working paper, University of Sabanci, Istanbul, (2003).

[21] Y. Bakos, E. Brynjolfsson, Bundling Information Goods: Pricing, Profits and Efficiency, Management Science 45 (12).

[22] W. Adams, J. Yellen, Commodity Bundling and the Burden of Monopoly, Quarterly Journal of Economics, Vol. 90 (3).

[23] R. McAfee, J. McMillan, M. Whinston, Multiproduct Monopoly, Commodity Bundling, Correlation of Values, Quarterly Journal of Economics 104 (1989) 371–84.

[24] I. Ayres, B. Nalebuff, Going Soft on Microsoft? the Eu's Antitrust Case and Remedy, The Economists' Voice 2 (2).

[25] B. Nalebuff, Competing Against Bundles, in: P. Hammond, G. Myles (Eds.), Incentives, Organization, and Public Economics, Oxford University Press, 2000.

[26] B. Nalebuff, Bundling as an Entry Barrier, Quarterly Journal of Economics 119 (1).

[27] A. Avenali, P. Bertolazzi, C. Batini, P. Missier, A Formulation of the Data Quality Optimization Problem in Cooperative Information Systems, in: Proceedings of the International Workshop on Data and Information Quality (DIQ 2004) in conjunction with CAiSE 04, Riga, Latvia, 2004.

[28] M.Lenzerini, Data integration: A Theoretical Perspective, in: Principles Of Database Systems, 2002, pp. 233–246.

[29] A. Y. Halevy, Answering Queries Using Views: a survey, VLDB Journal, vol. 10 (4) (2001), pp. 270–294.

[30] G. Grahne, A. O. Mendelzon, Tableau Techniques for Querying Information Sources Through Global Schemas, in: ICDT, 1999, pp. 332–347.

[31] A. Y. Halevy, Theory of Answering Queries Using Views, SIGMOD Record 29 (4) (2000) pp. 40–47.

[32] A. Y. Levy, A. O. Mendelzon, Y. Sagiv, D. Srivastava, Answering Queries Using Views, in: Principles Of Database Systems (PODS), 1995, pp. 95–104.

[33] A. Y. Levy, A. Rajaraman, J. J. Ordille, Querying Heterogeneous Information Sources Using Source Descriptions, in: Proceedings of the Internationl Conference on Very Large Data Bases (VLDB), 1996, pp. 251–262.

[34] R.L. Francis, L. McGinnis, J. White, Locational Analysis, European Journal of Operational Research, Vol. 12 (1983), pp. 220–252.

[35] H. Garcia Molina, J. Ullman, J. Widom, Database Systems: the Complete Book, Prentice Hall, N.J. 2002.

[36] R. Elmasri and S. Navathe, Fundamentals of Database Systems, Benjamin and Cummings, 2002.

[37] S. Raghunathan, Impact of Information Quality and Decision-maker Quality on Decision Quality: a Theoretical Model and Simulation Analysis, Decision Support Systems, Volume 26, Issue 4, October, 1999, pp. 275-286.

[38] R.Y. Wang, M.P. Reddy, H.B. Konource, Toward Quality Data: an Attribute-based Approach, Decision Support Systems Archive, Volume 13, Issue 3-4 (March 1995), pp. 349 - 372.

[39] A. Motro, Completeness of Information and its Application to Query Processing, in: W. W. Chu, G. Gardarin, S. Ohsuga, Y. Kambayashi (Eds.), Proceedings of the Twelfth International Conference on Very Large Data Bases, August 25-28, 1986, Kyoto, Japan, Morgan Kaufmann, 1986, pp. 170–178.

[40] A. Motro, I. Rakov, Estimating the Quality of Databases, in: T. Andreasen, H. Christiansen, H. L. Larsen (Eds.), FQAS, Vol. 1495 of Lecture Notes in Computer Science, Springer, 1998, pp. 298–307.

[41] D. Florescu, D. Koller, A. Y. Levy, Using Probabilistic Information in Data Integration, in: M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, M. A. Jeusfeld (Eds.), Proceedings of 23rd International Conference on Very Large Data Bases (VLDB), August 25-29, 1997, Athens, Greece, Morgan Kaufmann, 1997, pp. 216–225.

[42] F. Naumann, U. Leser, J.C. Freytag, Quality-driven Integration of Heterogenous Information Systems, in: VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases (VLDB), Morgan Kaufmann, Edinburgh, Scotland, UK, 1999, pp. 447–458.

[43] F.Naumann, J.C.Freytag, U.Leser, Completeness of Integrated Information Sources, Information Systems 29 (7) (2004) pp. 583–615.

[44] D. Ardagna, C. Cappiello, M. Comuzzi, C. Francalanci, B. Pernici, A Broker for Selecting and Provisioning High Quality Syndicated Data - International Conference on Information Quality, MIT, Boston, 2005.

[45] A.Motro, P.Anokhin, A. Acar, Utility-based Resolution of Data Inconsistencies, in: F. Naumann, M. Scannapieco (Eds.), International Workshop on Information Quality in Information Systems 2004 (IQIS'04), ACM, Paris, France, 2004.

[46] P. Mirchandani, R. Francis, Discrete Location Theory, John Wiley and Sons, New York, 1990.

[47] G. Nemhauser, L. Wolsey, The Uncapacitated Facility Location Problem, in: Mirchandani and Francis [46], pp. 119–171.

[48] D. Y. Holmberg K., M. Ronnqvist, An Exact Algorithm for the Capacitated Facility Location Problem with Single Sourcing, European Journal of Operational Research, Vol. 113 (1999), pp. 544–559.

[49] H. Pirkul, Efficient Algorithms for the Capacitated Concentrator Location Problem, Comput. Operations Research 14 (3) (1987) pp. 197–208.

[50] R. Galvao, The Use of Lagrangean Relaxation in the Solution of Uncapacitated Facility Location Problems, Location Science, Vol. 1 (1993), pp. 57–79.

[51] D. Erlenkotter, A Dual-based Procedure for Uncapacitated Facility Location, Operations Research, Vol. 26 (1978) 992–1009.

[52] F. Naumann, J.C. Freytag and U. Leser, Completeness of Integrated Information Sources, Information Systems, 29(7), 2004, pp.583-615.

[53] M. Scannapieco, C. Batini, Completeness in the Relational Model: a Comprehensive Approach - Proceedings of the International Conference on Data Quality, Boston Ma November 2004.

[54] R.Y. Wang, M. Ziad, and Y.W. Lee, Data Quality, Kluwer Academic Publisher, 2001, pp. 63-77.

[55] A. Parssian, S. Sarkar, and V.S. Jacob, Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product, Management Science, Vol. 50, No. 7, July 2004.

[56] C. Barahona, F.A. Chudak, Near-optimal Solutions to Large Scale Facility Location Problems, IBM Research Report RC21606.

[57] D.B. Shmoys, Approximation Algorithms for Facility Location Problems, in: K. Jansen, S. Khuller (Eds.), Approximation Algorithms for Combinatorial Optimization 2000, Vol. Lecture Notes in Computer Science 1913, 2000.

[58] M. Mahdian, Y. Ye, J. Zhang, Improved Approximation Algorithms for Metric Facility Location Problems, 2002, pp. 229–242.