A Correctness Criterion for Schema Dominance Centred on the Notion of 'Information Carrying' between Random Events

JUNKANG FENG and KAIBO XU
E-Business Research Institute
Business College, Beijing Union University
No. A3 Yanjing Li East, Chaoyang District, Beijing
CHINA

Database Research Group
School of Computing, University of the West of Scotland
High Street, Paisley
UNITED KINGDOM
junkang.feng@uws.ac.uk
kaibo.xu@bcbuu.edu.cn
(*The contributions of the authors are equal.*)

Abstract: - In systems development and integration, whether the instances of a data schema may be recovered from those of another is a question that may be seen profound. This is because if this is the case, one system is dominated and therefore can be replaced by another without losing the capacity of the systems in providing information, which constitutes a correctness criterion for schema dominance. And yet, this problem does not seem to have been well investigated. In this paper we shed some light on it. In the literature, works that are closest to this problem are based upon the notion of 'relevant information capacity', which is concerned with whether one schema may replace another without losing the capacity of the system in storing the same data instances. We observe that the rational of such an approach is over intuitive (even though the techniques involved are sophisticated) and we reveal that it is the phenomenon that one or more instances of a schema can tell us truly what an instance of another schema is that underpins a convincing answer to this question. This is a matter of one thing carrying information about another. Conventional information theoretic approaches are based upon the notion of *entropy* and the preservation of it. We observe that schema instance recovery requires looking at much more detailed levels of informational relationships than that, namely *random events* and *particulars* of random events.

Key-Words: - Database design, Schema dominance, Schema transformation, System integration, Information content, Information capacity

1 Introduction

We observe that whether the instances of a data schema may be recovered from those of another is a question that may be seen profound for systems design and integration as this underpins the validity of a design and the superiority of one design over another. This is because if this is the case, one system is dominated and therefore can be replaced by another without losing the capacity of the systems in providing information. This, we are convinced, would constitute a probably more insightful and therefore better correctness criterion than those presented in the literature for *schema dominance* as defined in the literature. This question does not seem thus far to have drawn sufficient attention and been made prominent and explicit. The notion of 'schema dominance' has been investigated, which is concerned with how a conceptual data schema may have at least the same capacity in terms of its instances as that of another, for example, references [6], [9] and [10]. In some of such investigations, Shannon's information theory [13] is used. For example, Lee in [7] and [8] puts forward an entropy preserving approach to measuring whether the entropy is lost when a schema changes. Arenas and Libkin [1] look at normal forms of relational and XML data by means of conditional entropy. In [9] and [10], a notion called 'information capacity preserving' is used to verify schema transformation. These alone, we maintain, cannot answer our question adequately. This is because the notion of 'information content' in these approaches is based upon the notion of 'types', and yet '*only particulars can carry information*' [2, p.26], that is, it is individual things in the world that carry information. The instances of a schema are at the level of particulars of random events. We will elaborate these ideas through the sections that follow.

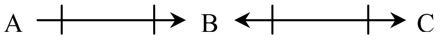
We motivate the discussion with a simple example of normalization of relational data schemata in section 2. We define the foundation and basic notions for our approach in sections 3 and 4 before we describe our approach *per se* in section 5. Then we apply our approach to normalization by revisiting the motivating example in section 6 and to the schema structural transformations of [10] in section 7, which shows the validity and usefulness of our ideas. We make concluding remarks in section 8.

2 A Motivating Example

Before introducing our approach in details, we present a small example first concerning normalization of relational databases. Two schemata S_1 and S_2 with one of their respective instances are shown below. S_2 is a *good* decomposition of S_1 [12]. We also draw their respective SIG (this stands for Schema Intension Graph proposed by Miller et al [10], which represents a schema in terms of nodes and edges) diagrams as follows.

S_1

A	B	C
1	2	3
2	2	3

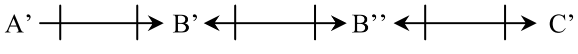


(a)

S_2

A’	B’
1	2
2	2

B”	C’
2	3



(b)

Fig. 1. An Example of Normalization

Let us take a look at how the instances of path P_{AC} of S_1 may be recovered from that of S_2. From the normalization decomposition algorithm that was used to create S_2 from S_1, we know that there is a bijection (i.e., an ‘one to one’ relationship, and it is represented by an arrow and a vertical bar at the both ends of an edge) between node A and node A’, and also another bijection between node C and node C’. We propose to call such things ‘inter-schemata constraints’ as they are logical limitations on the relationship between two schemata. Inter-schemata constraints capture underlying regularities that govern the relationship between two schemata. Moreover, we find that given an element of path $P_{A'C'}$, there is only one element of path $P_{A'C'}^{\nabla}$ = (A, A’, B’, B’’, C’, C) corresponding to it, and each element of $P_{A'C'}^{\nabla}$ is uniquely determined by at least one element of path $P_{A'C'}$. For example, $P_{A'C'}^{\nabla}$ = (1, 1, 2, 2, 3, 3) is uniquely determined by $P_{A'C'}$ = (1, 2, 2, 3). $P_{A'C'}^{\nabla}$ = (2, 2, 2, 2, 3, 3) is uniquely determined by $P_{A'C'}$ = (2, 2, 2, 3).

Similarly, each element of P_{AC} is uniquely determined by at least one element of path $P_{A'C'}^{\nabla}$. Through transitivity, each element of P_{AC} is uniquely determined by at least one element of path $P_{A'C'}$. Note that P_{AC} is a path in S_1, and $P_{A'C'}$ in S_2, thus the instance of the former shown in (a) of Fig. 1 above is fully recoverable from that of the latter shown in (b) of Fig. 1. As the instance shown in Fig. 1 is arbitrarily chosen, this example shows that any instance of S_1 is recoverable from instances of S_2, and this is one of the main reasons why S_1 can be replaced by S_2 without losing data that would otherwise be stored in S_1.

This example, even though simple, may show something profound. That is, the uniqueness of the instance of S_1 shown in Fig. 1 given the instance of S_2 shown in Fig. 1 is a result of the latter carrying all the information about the former in that the latter *can tell us truly* [4, P.64] all the details of the former. This is what we mean by ‘information carrying’ between states of affairs, and this is our foundation to approach the problem of schema instance recoverability. We know define the notion of ‘information carrying’ relation between systems.

3 Information-carrying Relation

To answer the question whether a data schema, moreover a system may be recovered from another, we propose a concept of ‘information carrying’ between systems, which means that ‘what information a signal carries is what it is capable of “telling” us, telling us *truly*, about another state of affairs’ [4, p.64].

This idea is established upon Dretske’s *semantic theory of information* [4], Devlin’s notion of ‘*infon*’ and *situation semantics* [3] and Floridi’s *information philosophy* [5]. To address how much (i.e., the amount of) information is generated and transmitted associated with a given set of state of affairs, Shannon [13] uses the notion of *entropy*, which is based upon probability theory to measure the amount of information. His approach calculates the quantity of information during a process of information transmission. However, Dretske [4, p.40] points out that apart from the quantity of information, the *content* of information should be considered, which is more relevant to the ordinary notion of 'information' than the quantity of it. For example, any toss involved in tossing a fair coin creates one bit of information (i.e., log2 = 1), which is the quantity of information. Moreover, we also need the content of information that whether it is the 'tail' or the 'head' that is facing up. If this piece of information is carried by a message, then the message not only carries one bit of information, but also tells us truly that the 'tail' or the 'head' is facing up. That is, the message carries both the quantity of information and the content of information. In this section, we extend Dretske's idea to define the notion of 'information carrying', which reveals and formulates the phenomenon that 'what information a signal carries is what it is capable of "telling" us, telling us truly, about another state of affairs' [4, p.64].

Here is an example of 'information carrying'.

Information Source	Information Carrier
Grade A	PASS
Grade B	
Grade C	
Grade D	
Grade E	FAIL
Grade F	

Table 1. A Grade Evaluation System

The input of this grade evaluation system is taken as an information source. The system showing the evaluation result is an information carrier for the existing information source.

3.1 States of Affairs of an Information Source and an Information Carrier

To describe the notion of 'information carrying', we look at the information source and the information carrier as two separate systems first, and then explore how they are related whereby one can tell us truly about the other. The whole information transmission is represented by the fact that a state of affairs of the information carrier is capable of telling us (i.e., carries the piece of information) that a particular state of affairs of the information source exists.

Following Shannon [13] and Dretske [4] we model both the information source and the information carrier as a selection process under a certain set of conditions with a certain set of possible outcomes. Let *s* be a set of state of affairs (described by a random event) among others at a selection process *S*. Similarly, let *r* be a set of state of affairs among others at a selection process *R*. Let *P*(*s*) denote the probability of *s* and *P*(*r*) denote the probability of *r*. Let *I*(*s*) denote the *surprisal* for *s* [4], which is taken as the information quantity created by *s* and *I*(*r*) denote the *surprisal* for *r*. Then we have:

$$I(s) = -\log P(s)$$
$$I(r) = -\log P(r)$$

Let *I*(*S*) denote the *entropy* of the selection process *S*, namely the weighted mean of surprisals of all random events of *S*. Then

$$I(S) = -\Sigma P(s_i)\log P(s_i),\ i = 1,\dots, m.$$

For the selection process *R*, we have:

$$I(R) = -\Sigma P(r_j)\log P(r_j), j = 1,\dots, n.$$

For our grade evaluation system, the input, which is the information source, can be seen as a random variable having six different possible values, namely those listed in the left column in Table 1. The random variable having a particular value, i.e., one of the six grades being inputted, is a random event. And also, such random events, which reflect the results of the selection process, show that all possible 'run' of the selection process results in the realization of all possible state of affairs. Therefore and hereafter we shall take the term 'random events' and the term 'state of affairs' as interchangeable.

Let s_a, s_b, s_c, s_d, s_e and s_f denote the six random events, namely one of the six grades being inputted to the system. Suppose that the six random events are equally likely, then the probability of s_a, s_b, s_c, s_d, s_e and s_f are all 1/6. The surprisals of them can be listed as:

$$I(s_a) = I(s_b) = I(s_c) = I(s_d) = I(s_e) = I(s_f) = -\log P(s_a) = \log 6 \text{ (bits)}$$

The entropy would be:

$$I(S) = -\Sigma P(s_i)\log P(s_i) = 6 \times \frac{1}{6} \times \log 6 = \log 6 \text{ (bits)}$$

Similarly, the information carrier can also be taken as a selection process. Let r_a, r_b denote two random events 'PASS' and 'FAIL' respectively. The probabilities for the random events of the information carrier are: $\frac{2}{3}$ and $\frac{1}{3}$ respectively. We would then have the surprisals for the information carrier R

$$I(r_a) = \text{-log}P(r_a) = \log\frac{3}{2} = \text{log}3 - 1 \text{ bits}$$

$$I(r_b) = \text{-log}P(r_b) = \log 3 \text{ bits}$$

The entropy of R is thus

$$I(R) = \text{-}\Sigma P(r_j)\log P(r_j) = \frac{2}{3} \times (\text{ log}3 - 1) + \frac{1}{3} \times \text{ log } 3 = \log 3 - \frac{2}{3} \text{ bits}$$

The states of affairs of the information carrier namely the outputs of the grade evaluation system are not independent of those of the information source namely the inputs of the system. There is some regularity between them as shown in Table 1. For example, whenever the input is a 'Grade A' then the output would be a 'Pass'. This is to say, seeing the 'Pass', we would know that the grade would definitely be one of A, B, C and D and be neither E nor F. But an information carrier may not carry all the information created at the information source. Moreover, it is not always the case that all the information created at an information carrier is acounted for by that created at the information source. Such situations are captured with the notions of 'equivocation' and 'noise'.

3.2 Equivocation and Noise

An information carrier can tell us truly something about the information source. When the 'something' is not 'everything', information is not fully carried. That is, there must be some information created at the information source and not carried by the information carrier and therefore lost in the process of information transmission. Such information is termed *equivocation*. This is on the one hand. On the other hand, the information created at the information carrier does not necessarily come from the source. This may be caused by some reason of the carrier itself or its being affected by something else other than the source [11]. Such information is termed *noise*.

Fig. 2 shows the notions of equivocation and noise in relation to the information source and the information carrier in an information carrying relationship.

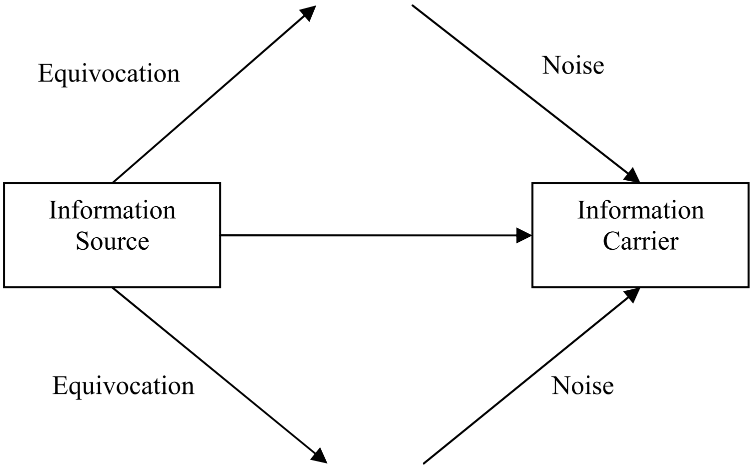


Fig. 2. Equivocation and Noise

How can we calculate equivocation and noise? Just like the measure of surprisal above, these two terms can be measured as long as the probabilities of random events at the source and the carrier are available. We now show how this can be done.

Recall that equivocation is the *lost information* that is created at the source but not carried by the carrier [4]. Let $P(s_i|r_j)$ denote the probability of the source event s_i under the condition that the carrier event r_j occurs. Let E_{s_i} (r_j) denote the equivocation in relation to s_i and r_j. We would have

$$E_{s_i}(r_j) = \text{-log}P(s_i|r_j)$$

This is because -log$P(s_i|r_j)$ is the amount of the part of the *uncertainty* reduced due to the occurrence of s_i that is *not* carried by the occurrence of r_j. If the latter does carry all the information created due to the occurrence of the former, which can be formulated as 'whenever the latter happens, the former happens as well', that is $P(s_i|r_j) = 1$, then -log$P(s_i|r_j)$ would be 0 bits. That is, the equivocation in relation to s_i and r_j would be none.

Similarly, noise can be seen as the information that is created at the carrier but is not accounted for by the source. Let $P(r_j|s_i)$ denote the probability of the carrier event r_j under the condition that the source event s_i occurs. Let N_{r_i} (s_j) denote the noise between r_j and s_i. We would have

$$N_{r_i}(\text{s}_\text{j}) = \text{-log}P(r_j|s_i)$$

Here is an example to summarize our analysis of equivocation and noise.

Information Source ***S***	**Information Carrier** ***R***
S_1	R_1
S_2	R_1
S_3	R_2
S_3	R_3

Table 2 An Example of Equivocation and Noise

Assume that some regularity that controls the situation illustrated in Table 2 is as such that the occurrence of Rj j = 1, 2, 3 are fully determined by that of Si, I = 1, 2, 3; S_1, S_2 and S_3 are equally likely to happen; and R2 and R3 are equally likely to happen in responding to S3. Then in relation to $S = S_1$ and $R = R_1$, we would have

$$I(S_1) = \log 3 \text{ (bits)};$$
$$I(R_1) = \log 3/2 \text{ (bits)};$$
$$E_{S_1}(R_1) = -\log P(S_1|R_1) = \log 2 \text{ (bits)};$$
$$N_{R_1}(S_1) = 0 \text{ (bits)}.$$

The above results show that equivocation exists and noise does not. This means that $R = R_1$ does not carry all the information that $S = S_1$.

We are now in a position to elaborate our approach in details. But first let us give a few basic notions.

4 Basic Notions

Definition 1: Paths

Let $G = (N, E)$ be a SIG and A an annotation (i.e., constraints on edges) on G, where N is a finite set for nodes, and E a finite set for edges. A *path*, P: $N_1 – N_k$, in G is a (possibly empty) sequence of edges e_1: $N_1 – N_2$, e_2: $N_2 – N_3$, ..., e_{k-1}: $N_{k-1} – N_k$ and is denoted $e_{k-1} \circ e_{k-2} \circ \ldots \circ e_1$. A path is *functional* (respectively *injective*, *surjective* or *total*) if every edge in the path is functional (respectively injective, surjective or total). The trivial path is a path from a node to itself containing no edges.

Definition 2: Instances of a Schema in SIG

An *instance* of G is a function whose domain is the sets N of nodes and E of edges.

Let $I_Y(S_1)$, $I_Y(S_2)$ denote the set of instances of S_1 and S_2 respectively. Let $\Im_1(S_1) \ldots \Im_n(S_1)$ denote instances of S_1. Then $I_Y(S_1) = \{\Im_1(S_1) \ldots \Im_n(S_1)\}$. Let A be part of G, $\Im(S_1)[A]$ denotes the part of $\Im(S_1)$ that is concerned with A, and it is called the projection of $\Im(S_1)$ on A.

Definition 3: Connections

A *connection* of a path P is an instance of P made up of instances of nodes that are linked by edges of P. That is, a connection of a path P is a link that associates individuals each of which belongs to one node of P and all nodes of P contribute at least one individual to the link. Let P = (node$_1$, node$_2$, …, node$_n$), individual nodes node$_{11}$, node$_{12}$, …node$_{1m}$ belong to node$_1$. For example, in a path ‘a student consults with a teacher on different occasions’ that connects students and teachers, the instances of node student are students appearing at different occasions for consulting a teacher. Any set of instances of nodes such as (node$_{11}$, node$_{23}$, …, node$_{nm}$) that are linked with one another is a connection of P.

As our approach is based upon ‘information carrying’ between schemata and their instances, we formalise a SIG by means of a set of mathematical notions centred on the concept of ‘random event’. As a result, a schema is looked at on a number of different levels. The following are a few definitions for this purpose.

Definition 4: A connection of a path say P may be of one of many possible types, which cannot be pre-determined. Thus what a connection of P could be is a *random variable*.

Definition 5: That a connection of a path happens to be of a particular type of those possible ones is a *random event*.

Definition 6: A specific connection of a path P that happens to be of type σ is a *particular* (i.e., an individual occurrence) of the random event that a connection happens to be of σ.

5 An Approach Centered on the Notion of ‘Information Carrying’

With the basic notions in place, now we present our approach with propositions and a further definition.

Proposition 1.

All instances of S_1 can be *recovered* from instances of S_2, if for any arbitrarily chosen instance $\Im_i(S_1)$ of S_1, there is at least one instance $\Im_j(S_2)$ of S_2 such that by looking at it (i.e., $\Im_j(S_2)$), we can know exactly how $\Im_i(S_1)$ would have been.

We now use SIG [10] as a tool to explore how this might be possible.

Definition 7.
Let SIG schema S_2 (meaning that S_2 is expressed in SIG format) be $G = (N, E)$. Let N_1 be the end nodes of a path P_{S1} in another SIG schema S_1. Let E_1 be an edge between N_1 and N due to constraints between N_1 and N, which can be formulated as annotations. $G' = (N', E')$, where $N' = N \cup N_1$, $E' = E \cup E_1$, is *an extended SIG schema* S_2 by taking into consideration inter-schemata constraints between S_1 and S_2.

In some cases, between schemata S_1 and S_2, there are sets of constraints $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_n\}$ and $\beta = \{\beta_1, \beta_2, ..., \beta_m\}$ such that α links one end node, say N_1 of a path P_{S1} in schema S_1 and one end node of each path in a set of paths $P_{S2}^* = \{P_1, P_2, ..., P_n\}$ in schema S_2, and β links the other end node, say N_2 of P_{S1} and the other end node of each path of P_{S2}^*. That is, $P_{S2}^* = \{P_1, P_2, ..., P_n\}$ is linked with P_{S1} through α and β. In such a case we would have $P_{S2}^{\nabla} = (N_1, P_1, N_2, P_2, N_1, \ldots)$ which walks through N_1, N_2 and P_{S2}^* by α and β.

Individual connections of a path could be of a same type in the sense that individual nodes that are connected are the same, for example, Dr Jones and Student Jane, and the meanings of the connections are also the same, for example, Dr Jones teaches Student Jane. Such a connection (a connection of a particular kind) may occur more than once for one reason or another.

For the structure just described, it is a random event that an individual connection of P_{S2}^* happens to be of type a, denoted $P_{S2}^* = a$, and it is also a random event that an individual connection of P_{S2}^{∇} happens to be of type b, denoted $P_{S2}^{\nabla} = b$. If under the condition of $P_{S2}^* = a$, it is always the case that $P_{S2}^{\nabla} = b$, then we have $p(\ P_{S2}^{\nabla}=b \mid P_{S2}^*=a) = 1$, which denotes the probability of $P_{S2}^{\nabla}=b$ under the condition of $P_{S2}^*=a$ is 1, i.e., a certainty.

Proposition 2.
For a pair of schemata, say S_1 and S_2, there could be a kind of relationship between them, namely, for every possible instance of schema S_1, i.e., $\Im_i(S_1) \in I_Y(S_1)$, there is at least one instance of S_2, i.e., $\Im_j(S_2) \in I_Y(S_2)$, such that the following condition holds:

If for every possible type a of connections of P_{S2}^{∇}, there is at least one type b of P_{S2}^* such that $p(\ P_{S2}^{\nabla}=a \mid P_{S2}^*=b) = 1$, and a similar relation holds between P_{S1} and P_{S2}^{∇}. Moreover, each individual connection of P_{S2}^{∇} can be ascertained by the existence of at least one individual connection of P_{S2}^*, and the same applies to individual connections of P_{S1} and individual connections of P_{S2}^{∇}, then $\Im_i[P_{S2}^*]$ carries all the information of $\Im_i[P_{S1}]$.

If this applies to every path of S_1, then $\Im_j(S_2)$ carries all the information of $\Im_i(S_1)$. If this in turn applies to every possible $\Im_i(S_1) \in I_Y(S_1)$, then every instance of schema S_1 can be recovered from those of schema S_2. Following the definitions of annotations of SIG [10], the above condition entails a total injective binary relation f: $P_{S2}^{\nabla} \rightarrow P_{S2}^*$ and a total injective binary relation f': $P_{S1} \rightarrow P_{S2}^{\nabla}$.

Lemma 1.
Let $\Im_i(S_2)^{\nabla}$ denote an instance of $S_2 \cup \alpha \cup \beta$, where α and β are inter-schemata constraints between S_1 and S_2. For each possible type a of connections of $\Im_j(S_1)$, if the following criteria are satisfied then $\Im_j(S_1)$ can be recovered from $\Im_i(S_2)$:

- For every type a of $\Im_j(S_1)$, there is at least one type b of $\Im_i(S_2)$ and one type of c of $\Im_i(S_2)^{\nabla}$ such that $p(\Im_i(S_2)^{\nabla} = c \mid \Im_i(S_2) = b) = 1$ and $p(\Im_j(S_1) = a \mid \Im_i(S_2)^{\nabla} = c) = 1$.
- Each individual connection of $\Im_i(S_2)^{\nabla}$ can be ascertained by the existence of at least one individual connection of $\Im_i(S_2)$ and the same applies to individual connections of $\Im_j(S_1)$ and $\Im_i(S_2)^{\nabla}$.

Proof: The proof uses the transitivity of conditional probability.

6 The Example of Normalization Revisited

Now we explore, with our approach centered on the notion of ‘information-carrying’, whether normalization with non-additive join (also called ‘lossless join’) satisfies the above condition, i.e., whether all instances of the original schema can be recovered from those of the resultant schema of normalization, and therefore the former can be replaced by the latter without losing the capacity of storing data. We still use the same example in Fig. 1, which shows that S_1 has three paths, which are denoted as P_{AB}, P_{BC}, P_{AC} respectively, and S_2 has paths $P_{A'B'}$, $P_{A'B''}$, $P_{A'C'}$, $P_{B'C'}$ and $P_{B''C'}$ etc.

From the perspective of random events and particulars of random events, Proposition 2 states that for each possible type a of connections of path P_{S1} in $\Im_i(S_1)$, if the following criteria are satisfied then $\Im_i(S_1)$ can be recovered from $\Im_j(S_2)$:

- There is at least one type b of P_{S2}^* and one type of c of P_{S2}^{∇} in $\Im_j(S_2)$ such that $p(\ P_{S2}^{\nabla}=c \mid P_{S2}^*=b) = 1$ and $p(\ P_{S1} = a \mid P_{S2}^{\nabla}=c) = 1$.
- Each individual connection of P_{S2}^{∇} can be

ascertained by the existence of at least one individual connection of P_{S2}^{*}, and the same applies to individual connections of P_{S1} and P_{S2}^{∇}.

For the instance of S_1 shown in Fig. 1 (a), we find an instance of S_2 shown in Fig. 1 (b), which is a result of restructuring the former according to the structure of S_2. Now we justify that the condition identified above is met, i.e., to justify that the former can be recovered by looking at the latter through our approach outlined above, rather than the conventional 'joining the two relations in S_2'.

Here is the SIG diagram, which shows S_2 and its links with attributes (represented by nodes) A and C of S_1.

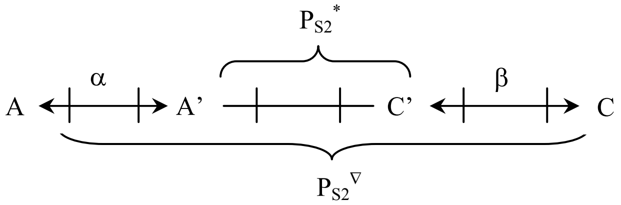


Fig. 3. SIG Diagram for analysing the normalisation example

The above diagram explains from 'binary relation' or topological characteristics we can be sure that the connections between A and C are uniquely determined by the rest of the diagram. This is one of the results from the normalisation having been conducted following a standard lossless join decomposition algorithm. Now we check P_{AB} in S_1 as an example to show our approach.

For P_{AB} in S_1, we find a path $P_{A'C'}$ in S_2 that satisfies the condition. Fig. 1 (b) is a restructured Fig. 1 (a) and thus no new data value is involved. More importantly, because it is a lossless-join decomposition, which guarantees that a natural join of the relations after the decomposition results in exactly the same relation as that before the normalisation decomposition, each tuple in S_1 can be recovered from one tuple of the result of the natural join performed on relations of S_2. That is to say, $\Im_i(S_1)$ is fully recoverable from $\Im_j(S_2)$. Therefore for individual instances of A of P_{AC} and A' of $P_{A'C'}$, we have P_{AC}.A = $P_{A'C'}$.A'. Here we denote α as a constraint between P_{AC} and $P_{A'C'}$, which serves as an edge that links P_{AC}.A and $P_{A'C'}$.A'.

That is, α = (P_{AC}.A, $P_{A'C'}$.A') with individual connections (i.e., instances) (1, 1) and (2, 2):

P_{AC}.A	$P_{A'C'}$.A'
1	1
2	2

Fig. 4. Inter-schemata constraint α for path P_{AC} and path $P_{A'C'}$

Similarly we have β = (P_{AC}.C, $P_{A'C'}$.C') with two individual connections (2, 2):

P_{AC}.C	$P_{A'C'}$.C'
3	3
3	3

Fig. 5. Inter-schemata constraint β for path P_{AC} and path $P_{A'C'}$

Notice that the two individual connections of (3, 3) should not been seen as one. The first individual connection links C in P_{AC} = (1, 3) and C' in $P_{A'C'}$ = (1, 3), and the second links C in P_{AC} = (2, 3) and C' in $P_{A'C'}$ = (2, 3).

For the type ((1, 1) (1, 3), (3, 3)) of the connections of path P_{S2}^{∇} = (A, $P_{A'C'}$, C), we find that the type (1, 3) of connections of path P_{S2}^{*} = $P_{A'C'}$ such that the probability of the latter is 1 given the former. Similarly, type ((2, 2) (2, 3), (3, 3)) of the connections of path P_{S2}^{∇} = (A, $P_{A'C'}$, C) is ascertained by the type (2, 3) of connections of path P_{S2}^{*} = $P_{A'C'}$.

As for individual connections of path P_{S2}^{∇}, we want to check whether 'each individual connection of P_{S2}^{∇} can be ascertained by the existence of at least one individual connection of P_{S2}^{*}'. We find that it is the case due to the inter-schema constrains α and β both being a bijection.

Now we examine the relationship between P_{S1} and P_{S2}^{∇} along a similar line. Because it is a lossless-join decomposition, $\Im_i(S_1)$ is fully recoverable from $\Im_j(S_2)$. As a result, the two aforementioned bijective inter-scheme constraints α and β between node A in S_1 and node A' in S_2 and between node A in S_1 and node A' in S_2 are in existence between the two schemata. Also, because $\Im_i(S_1)$ is fully recoverable from $\Im_j(S_2)$, each type of instances of any part of S_1 can be recovered from some types (could be just one type) of instances of one or more parts of S_2. Therefore, in the case of path P_{AB}, we find that type (1, 3) of the connections of path P_{AC} is ascertained by the type ((1, 1) (1, 3), (3, 3)) of the connections of path P_{AC}^{∇}. As path P_{AB} is a constituting part of P_{AC}, type (1, 2) of the connections of path P_{AB} is ascertained by type (1, 3) of the connections of path P_{AC}. As for individual connections, again due to $\Im_i(S_1)$ being fully recoverable from $\Im_j(S_2)$ and path P_{AB} is a constituting part of P_{AC}, each individual connection of any type of the connections of P_{AB} is ascertained by a connection of P_{AC} of which the former is a constituting part. Thus each individual connection of $P_{S1} = P_{AB}$ can be ascertained by another individual connection of $P_{S2}^* = P_{A'C'}$ via the individual connection P_{S2}^∇. It means that path P_{AB} can be recovered from path $P_{A'C'}$.

7 Miller's αος-dominance

We continue with the validly check of our ideas by examining an influential work on schema transformation by Miller et al [10], which is based upon Hull's [6] notion of 'relative information capacity'. The basic notion of their work is called '*αος*-dominance relation' between two data schemata, which are set of schema structural transformations. We look at the elementary transformations (dominance or equivalence) that make up such dominance, namely *α*-dominance, *ο*-dominance and *ς*-dominance.

To examine these transformations, we require a couple of definitions.

Definition 8. Corresponding instances of S_1 and S_2
For a given instances of S_1 denoted $\Im_s(S_1)$, if there is an instance of S_2 denoted $\Im_t(S_2)$ such that from the latter the former can be fully recovered, then $\Im_s(S_1)$ and $\Im_t(S_2)$ are said to be a pair *corresponding instances*.

Definition 9. Corresponding nodes of S_1 and S_2
Let A_i (i = 1, …, *p*) be nodes of S_1, and B_j (j = 1, …, *q*) be nodes of S_2. If for any A_i there is at least one set of B_j such that $\psi(A_i) = B_j$ for any pair of corresponding instances of S_1 and S_2, say $\Im_s(S_1)$ and $\Im_t(S_2)$, that is, for $\Im_s(S_1)[A_i]$, there is at least $\Im_t(S_2)[B_1 + B_2 + \ldots + B_m]$ such that the former is uniquely determined by therefore fully recoverable from the latter, then A_i and $(B_1 + B2 + \ldots + B_m)$ are said to be a pair *corresponding nodes*.

With these definitions in place, now we examine each elementary transformation in turn.

- *α*-dominance

An example of *α*-transformation is shown in Fig. 6.

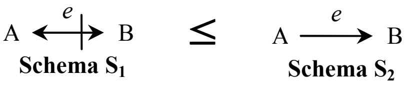


Fig. 6. An *α*-transformation

Let α denote a set of constraints, which are inter-schema constraints formulated in terms of annotations defined by Miller et al [10], such that α links node A of schema S_1 and node A of schema S_2, and β links B of S_1 and B of S_2.

α-dominance requires that nodes are the same. That is, for any node of S_1, there must a node of S_2 such that the mapping from the former to the latter is total, injective and internal. Therefore, in the above example, α and β should be total, injective and internal. Now we use our approach to justify whether an α-dominance relation satisfies the condition of 'information-carrying' that we defined earlier.

Let P_{S1} denote the path (A, B) of S_1 and P_{S2}^* denote the path (A, B) of S_2. Let $P_{S2}^\nabla = \alpha \circ P_{S2}^* \circ \beta$ (i.e., Let P_{S2}^∇ denote the path $(\alpha, P_{S2}^*, \beta)$). First of all, we check whether there is at least one type *b* of P_{S2}^* and one type of *c* of P_{S2}^∇ in $\Im_j(S_2)$ such that $p(P_{S2}^\nabla = c \mid P_{S2}^* = b) = 1$ and $p(P_{S1} = a \mid P_{S2}^\nabla = c) = 1$, where *a* is an arbitrary type of P_{S1} that we want ascertained. Secondly, we check whether each individual connection of P_{S2}^∇ can be ascertained by the existence of at least one individual connection of P_{S2}^*, and the same applies to individual connections of P_{S1} and P_{S2}^∇, the purpose of which is to ascertain all individual connections of P_{S1} so that $\Im[P_{S1}]$ can be fully determined (recovered).

As α and β extend P_{S2}^* to P_{S2}^∇, and both α and β are total, injective and internal, for any type of P_{S2}, there must be at least one type of P_{S2}^∇ such that P_{S2}^∇ can be ascertained by P_{S2}^*. At the individual connections level, also because of both α and β being total, injective and internal in a *α*-transformation, we have that each individual connection of P_{S2}^∇ can be ascertained by at least one individual connection of P_{S2}^*.

The same approach can be applied to the examination of the relationship between P_{S1} and P_{S2}^∇. We know that for α-transformation, every instance of S_1 is also an instance of S_2, that is, they are exactly the same. That is, the instances of the end nodes of P_{S1} are ascertained by P_{S2} as they are part of P_{S2} and the connections of the instances of the end nodes are ascertained by P_{S2} . Therefore given a path P_{S2}^∇ (which is determined by P_{S2} in S_2) we would have a unique path P_{S1} in S_1.

Therefore, an *α*-dominance (transformation) satisfies the condition of 'information-carrying'.

- *ο*-dominance

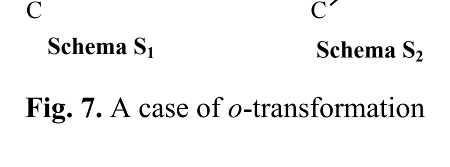


Fig. 7. A case of *ο*-transformation

According to Miller et al [10], the above example of *o*-transformation is an 'information capacity preserving transformation', which shows that under the condition of $\Im[g] = \Im[e] \circ \Im[p]$, every instance of edge *e* determines a unique instance of *g* so. In terms of our approach, they uniquely determine each other, and therefore any instance of *e* is recoverable from an instance of *g* that corresponds to it. We now verify *o*-transformation is of information carrying. To this end,

1. We first check whether edge $e \circ p$ of schema S_1 and its individual connections are ascertained by edge *g* and edge *p* of schema S_2 and its individual connections respectively, namely whether $\Im[p] \circ \Im[e]$ is ascertained by $\Im[g]$ and $\Im[p]$.
2. If it is so, we would then know that edge *e* of S_1 and its individual connections are ascertained by edge $e \circ p$ of S_1 and its individual connections respectively.

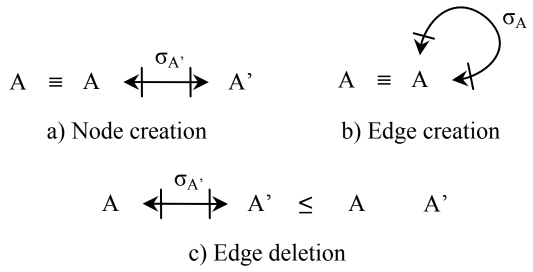
We first check the relationship between edge $e \circ p$ of S_1 and edge *g* and edge *p* of S_2. Let P_{S1} denote path (A, B) of S_1 and P_{S1}' denote path (C, A, B) of S_1. Let P_{S2} denote the whole structure, namely the set of path (C, B) and path (C, A)) of S_2. Let α denote the inter-schema constraints that link node C of S_1 and node C of S_2. Let β links B of S_1 and B of S_2.

Miller's approach to schema transformation is to look at whether it is possible to restructure a given instance of S_1 to fit S_2 such that it becomes a unique instance of S_2. As all nodes remain the same through the transformation, α and β are both total, injective and internal. As α and β extend P_{S2} to P_{S2}^{∇}, for any type, say *c*, of P_{S2}^{∇}, there must be at least one type, say *b*, of P_{S2} such that $p(P_{S2}^{\nabla} = c \mid P_{S2} = b) = 1$. At the individual connection level, we also have that each individual connection of P_{S2}^{∇} can be ascertained by at least one individual connection of P_{S2}.

Now we examine the relationship between P_{S1}' (C, A, B) and P_{S2}^{∇} (i.e., edge *p* and edge *g* in S_2 extended with *α* and β). Under the condition $\Im[g] = \Im[e] \circ \Im[p]$, path P_{S1}' (which walks through node A, C, B) and its individual connections are ascertained by the whole structure of P_{S2} and its individual connections respectively. Notice that the composite path $e \circ p$ cannot be ascertained by edge *g* only.

Secondly, we check the relationship between edge *e* of S_1 and edge $e \circ p$ of S_1. It can be seen that whenever the whole $\Im[e] \circ \Im[p]$ is certain, part of it $\Im[e]$ is automatically certain.

- ς-dominance



a) Node creation

b) Edge creation

c) Edge deletion

Fig. 8. ς-transformations

Miller et al [10] defines *selection transformation* (ς-transformations) to look at whether information capacity is preserved through *node creation*, *node deletion* (if A = A'), *edge creation* and *edge deletion*, which are shown in Fig. 8. The idea is to check whether the instance of S_1 becomes a unique instance of S_2 after having been restructured to fit S_2. This implies that data values do not change.

Moreover, ς-transformation is only concerned with *selection* edges when talking about *edge creation* and *edge deletion*, which means that the end nodes of an edge are concerned with the same kind of objects that are involved in the 'is a' relationship. Note also that the selection edges that are involved in a ς-transformation are all bijections between the instances of their end nodes.

Therefore, any pair of corresponding instances of two schemata linked by a ς-transformation uniquely determines each other. That is, ς-transformation satisfies the condition of *information carrying* that we defined earlier. As a result, any instance of the original schema is ascertained by at least one instance of the transformed schema with both its types and individual connections.

8 Conclusions

Using a fundamental notion of 'information carrying' between states of affairs, we have presented in this paper an approach to the problem of under what conditions the instances of a schema can be recovered by those of another, which constitutes a correctness criterion for schema dominance. We presented a set of conditions in the form of propositions that are sufficient for this to take place. In exploration of this problem, we find that not only the normal 'random event' level but also the 'particulars of random events' level must be involved. We showed how our definitions, propositions and lemma would work on conventional normalization problems and the schema structural transformations that are ‘information capacity preserving’ put forward in [10].

References:

[1] Arenas M, Libkin L, An information-theoretic approach to normal forms for relational and XML data. *Journal of the ACM*, Vol. 52, 2005, pp. 246-283

[2] Barwise J, Seligman J, *Information Flow: the Logic of Distributed Systems*, Cambridge University Press, ISBN 0-521-58386-1, 1997

[3] Devlin, K. (1991). *Logic and information*, Cambridge University Press.

[4] Dretske F. I, *Knowledge and the flow of information*, Basil Blackwell, Oxford, 1981

[5] Floridi, L. (2005). Is Semantic Information Meaningful Data? *Philosophy and Phenomenological Research* 70(2): 351-370.

[6] Hull R, Relative information Capacity of Simple Relational Database Schemata, *SIAM Journal of Computing*, 1986, 15(3): pp. 856-886.

[7] Lee TT, An information-theoretic analysis of relational databases—part I: Data Dependencies and Information Metric. *IEEE Transactions on Software Engineering*, 1987, 13(10): 1049-1061

[8] Lee TT, An information-theoretic analysis of relational databases—part II: Information Structure of Database Schemas, *IEEE Transactions on Software Engineering*, 1987, 13(10): 1061-1072

[9] Miller RJ, Ioannidis YE, Ramakrishnan R, The Use of Information Capacity in Schema Integration and translation, In *Proceedings of the International Conference on Very Large Data Bases*, Dublin, Ireland, 1993, pp. 120-133.

[10] Miller RJ, Ioannidis YE, Ramakrishnan R, Schema Equivalence in Heterogeneous Systems, Bridging Theory and Practice, *Information Systems*, 1994, 19(1): pp. 3-31.

[11] Mingers 1995 Information and meaning: foundations for an intersubjective account, Info Systems J, 5

[12] Moody DL, Flitman AR, A decomposition method for entity relationship models: a systems theoretic approach, International Conference on Systems Thinking in Management 2000. In: Altmann, G., Lamp, J., Love, P.E.D., Mandal, P., Smith, R., Warren, M. (Eds.), *Proceedings of the International Conference on Systems Thinking in Management*, Technical University of Aachen, Geelong, Australia, 2000, pp. 462-469.

[13] Shannon C, A Mathematical Theory of Communication. *Bell.Syst.Tech.* J., 1948, 27, 379-423, 623-656.