Constructing Realizers of Semantic Entity Relationship Schemes

Konrad Engel and Sven Hartmann

Universität Rostock, Fachbereich Mathematik, 18051 Rostock, Germany

Abstract

We use the Ford-Bellman algorithm to construct feasible class cardinality functions for a given semantic entity relationship scheme (SER scheme). Further we show that these functions form a distributive lattice under a natural ordering. Finally we present an algorithm to construct the minimum element of this lattice which can be interpreted as the class cardinality function of a realizer of minimum size for the given SER scheme.

1 Introduction

Within the recent years, ER modeling has become one of the most popular methods in high-level data design. It provides a convenient, powerful and implementation independent description of objects to be stored in a database. The ER approach and its variations are frequently used in the conceptual design of database applications, and lots of data design tools employ its concepts.

In the ER model, data are described as entities and relationships. Each entity consists of a number of instances representing real-world objects which can be characterized by the same attributes. On the other hand, a relationship among several entities describes a set of associations between real-world objects. A relationship consists of a number of instances each of whom represents an association among instances of the participating entities. In all what follows, we use the term class to refer to an entity or a relationship.

Often, a relationship comes along with certain constraints that limit the possible combinations of instances of the participating entities. These constraints are frequently used to model real-world dependencies between objects. Among the most popular kinds of constraints there are the cardinality ratio constraints. They specify restrictions on the number of instances of a fixed relationship that an instance of a fixed entity may be involved in.

Every collection of entities and relationships representing a real-world situation shall be called an ER scheme. In addition, ER schemes with cardinality ratio constraints are said to be semantic (SER schemes). These schemes are used in modern data design to model databases. This has given rise to the main problem: Given an SER scheme, obtain a database that satisfies the cardinality ratio constraints.

2 The graph-theoretical model

From the graph-theoretical point of view, an *SER scheme* $\mathcal{S}$ is a bipartite digraph $G = (V, A)$, where $V = E \cup R$ and $A \subseteq E \times R$, together with a pair of functions $\alpha_S : A \to I\!N \cup \{0\}$ and $\beta_S : A \to I\!N \cup \{\infty\}$ defined on the edge set of G such that $\alpha_S(a) \leq \beta_S(a)$ for all $a \in A$. (We use $I\!N = \{1, 2, \ldots\}$.)

The vertices in E and R are just the entities, resp. the relationships of the SER scheme. Two vertices $e \in E$ and $r \in R$ are connected by an edge in G iff the entity e participates in the relationship r. Henceforth, we shall denote the neighborhood of a vertex v by $N(v)$.

Let us now define the satisfiability problem for SER schemes. Consider a bipartite digraph $\mathcal{R} = (V^R, A^R)$ with vertex set V^R and edge set A^R. It is called a *realizer* of a given SER scheme $\mathcal{S}$ iff there exist two mappings $\phi : V^R \to V$ and $\psi : A^R \to A$ satisfying the following conditions:

(R1) $\psi(a^R) = (\phi(e^R), \phi(r^R))$ for all $a^R = (e^R, r^R) \in A^R$,

(R2) $\phi^{-1}(v) \neq \emptyset$ for all $v \in V$,

(R3) $|\phi^{-1}(e) \cap N(r^R)| = 1$ for all $a = (e, r) \in A$ and all $r^R \in \phi^{-1}(r)$,

(R4) $\alpha_S(a) \leq |\phi^{-1}(r) \cap N(e^R)| \leq \beta_S(a)$ for all $a = (e, r) \in A$ and all $e^R \in \phi^{-1}(e)$.

An SER scheme is said to be *(strongly) satisfiable* iff it has at least one realizer. According to condition (R1), a satisfiable SER scheme is the homomorphic image of its realizers under the mappings ϕ and ψ. The pre-images of a vertex v are just the instances of the class v of the SER scheme.

According to condition (R2), a realizer contains at least one instance of every class of the SER scheme. The number of pre-images of a vertex $v \in V$ shall be called the *class cardinality* of the class v and denoted by $x_R(v)$. The function $x_R : V \to \mathbb{N}$ is called the *class cardinality function* of $\mathcal{R}$.

The inequalities in condition (R4) are the *cardinality ratio constraints* of the SER scheme. They give us a lower and an upper bound on the number of instances of the relationship r that instances of the participating entity e are involved in.

The problem of constructing realizers of SER schemes has been studied by several authors. Lenzerini and Nobili [5] proposed a two-phase approach to solve it. In a first phase, one looks for legal class cardinalities for the entities and relationships in the scheme. This provides us the vertex set of a realizer. The edges are constructed afterwards in a second phase.

For the first phase, Lenzerini and Nobili proposed to solve a system of linear inequalities which yields then the class cardinalities. For the second phase, they gave an efficient algorithm using congruences of numbers.

The objective of this paper is to formulate algorithms for the first phase which efficiently construct the cardinalities of the classes. The paper is organized as follows. In the next section, we formulate the class cardinality problem in terms of combinatorial optimization. Using shortest-path methods we provide a polynomial-time algorithm to solve this problem. In Section 4 we study the structure of the set of class cardinality functions of the realizers of an SER scheme. We show that it forms a distributive lattice (under an appropriate definition of join and meet operations for the lattice). In Section 5 we consider the problem of constructing a realizer of minimum size and formulate an algorithm for determining the cardinalities of the classes under this additional condition.

3 The class cardinality problem

We begin by defining the class cardinality problem for a given SER scheme $\mathcal{S}$. Assuming $\mathcal{S}$ is satisfiable, consider a realizer $\mathcal{R}$ with class cardinality function x. By summing up over all pre-images of e, conditions (R3) and (R4) imply

$$\alpha_S(a)x(e) \leq x(r) \leq \beta_S(a)x(e) \text{ for all } a = (e, r) \in A. \tag{1}$$

Hence, to obtain a realizer of an SER scheme $\mathcal{S}$, we are required to find class cardinalities that are *feasible* in that inequality (1) is satisfied.

In the sequel, we use the term class cardinality function to refer to an integral function defined on the vertex set of G. The *class cardinality problem* is that of finding a feasible class cardinality function for a given SER scheme $\mathcal{S}$, i.e. a function $x : V \to \mathbb{N}$ satisfying inequality (1) (such a function may not even exist).

Lenzerini and Nobili proved that, for every feasible class cardinality function x, there is a realizer $\mathcal{R}$ such that x_R equals x. We will henceforth restrict our attention to the class cardinality problem.

For a given SER scheme $\mathcal{S}$ consisting of a digraph $G = (V, A)$ and functions α_S and β_S, we consider a new digraph $G' = (V, A \cup A^{-1})$ which can be obtained from G by adding to each edge $a = (e, r)$ its reverse edge $a^{-1} = (r, e)$.

On the edge set of G' we define a length function $l_S : A \cup A^{-1} \to \mathbb{R} \cup \{\infty\}$ by

$$l_S(a) := \begin{cases} \infty & \text{if } \beta_S(a) = \infty, \\ \log \beta_S(a) & \text{otherwise} \end{cases}$$

and

$$l_S(a^{-1}) := \begin{cases} \infty & \text{if } \alpha_S(a) = 0, \\ -\log \alpha_S(a) & \text{otherwise} \end{cases}$$

where $a \in A$.

A subset $P = \{a_1, a_2, \ldots, a_t\}$ of edges from $A \cup A^{-1}$ with $a_i = (v_{i-1}, v_i)$, $i = 1, \ldots, t-1$, for some vertices $v_i \in V$, $i = 1, \ldots, t$, is called a *directed path* and, if additionally $v_0 = v_t$, a *directed cycle* in G'. We define the length of a directed path P by

$$l_S(P) := \sum_{a \in P} l_S(a).$$

A directed cycle C is said to be *critical for a scheme $\mathcal{S}$* if it has negative length with respect to l_S, i.e. iff $l_S(C) < 0$.

Let x be an arbitrary class cardinality function and put $p(v) := \log x(v)$ for all $v \in V$. Then (1) is equivalent to

$$-l_S(a^{-1}) \leq p(r) - p(e) \leq l_S(a) \text{ for all } a = (e, r) \in A,$$

i.e. to

$$p(v) - p(u) \leq l_S(a) \text{ for all } a = (u, v) \in A \cup A^{-1}. \tag{2}$$

So we are lead to the problem of finding a function $p : V \to I\!N \cup \{0\}$ satisfying (2). Such a function is called a *potential function in G' with respect to l_S*. The existence of a potential function is known to be ensured for a digraph with strongly connected components and finite length function iff it contains no directed cycles of negative length. (The necessity is clear and the sufficiency follows from the correctness of the Ford-Bellman algorithm described below.)

Unfortunately, the length function l_S defined above may not even be finite. Therefore, let us consider the scheme $\mathcal{N}$ consisting of the digraph G and the functions $\alpha_N : A \to I\!N \cup \{1/M\}$ and $\beta_N : A \to I\!N$ defined by

$$\begin{aligned} \alpha_N(a) &:= \begin{cases} 1/M & \text{if } \alpha_S(a) = 0, \\ \alpha_S(a) & \text{otherwise,} \end{cases} \\ \beta_N(a) &:= \begin{cases} M & \text{if } \beta_S(a) = \infty, \\ \beta_S(a) & \text{otherwise,} \end{cases} \end{aligned} \tag{3}$$

where $a \in A$ and

$$M := \prod_{\substack{a \in A, \\ \alpha_S(a) \neq 0}} \alpha_S(a) \tag{4}$$

or $M := 1$ if $\alpha_S(a) = 0$ for all $a \in A$. (In the latter case the problem becomes trivial: we may put $x(v) := 1$ for all $v \in V$.)

We call $\mathcal{N}$ the *normalized scheme* of $\mathcal{S}$. Note, that $\mathcal{N}$ is not always an SER scheme since α_N is not necessarily integral. As above, we define a length function $l_N : A \cup A^{-1} \to I\!R$ on the edge set of G' by $l_N(a) := \log \beta_N(a)$ for all $a \in A$ and $l_N(a^{-1}) := -\log \alpha_N(a^{-1})$ for all $a^{-1} \in A^{-1}$.

Lemma 1 *An SER scheme $\mathcal{S}$ contains a critical directed cycle iff its normalized scheme $\mathcal{N}$ contains a critical directed cycle.*

Proof. Consider the digraph G' and let C be a directed cycle in G'. Assume C has negative length with respect to l_S. We show, that C is critical for $\mathcal{N}$, too. By

$$l_S(C) = \sum_{a \in C} l_S(a) < 0$$

we obtain $l_S(a) \neq \infty$ for all $a \in C$. This implies $l_N(a) = l_S(a)$ for all $a \in C$ and

$$l_N(C) = \sum_{a \in C} l_N(a) = \sum_{a \in C} l_S(a) < 0,$$

i.e. C is critical for $\mathcal{N}$.

On the other hand, assume C is critical for $\mathcal{N}$. Then

$$l_N(C) = \sum_{a \in C} l_N(a) < 0$$

implies

$$\prod_{a \in C \cap A} \beta_N(a) < \prod_{a' \in C \cap A^{-1}} \alpha_N(a').$$

By

$$\prod_{a' \in C \cap A^{-1}} \alpha_N(a') \leq M$$

we have $\beta_N(a) < M$ for all $a \in C \cap A$, and by

$$1 \leq \prod_{a \in C \cap A} \beta_N(a)$$

we obtain

$$1 < \prod_{a' \in C \cap A^{-1}} \alpha_N(a') = \alpha_N(a) \prod_{\substack{a' \in C \cap A^{-1} \\ a' \neq a}} \alpha_N(a') \leq \alpha_N(a)\ M$$

and $1/M < \alpha_N(a)$ for all $a \in C \cap A^{-1}$. Hence, $l_S(a) = l_N(a)$ for all $a \in C$, which implies

$$l_S(C) = \sum_{a \in C} l_S(a) = \sum_{a \in C} l_N(a) < 0,$$

i.e. C is critical for $\mathcal{S}$. ■

Lemma 2 *Let $\mathcal{S}$ be an SER scheme and $\mathcal{N}$ its normalized scheme. Then a class cardinality function x is feasible for $\mathcal{S}$ if it is feasible for $\mathcal{N}$.*

Proof. This is an easy consequence of the inequalities $\alpha_S(a) \leq \alpha_N(a)$ and $\beta_N(a) \leq \beta_S(a)$ for all $a \in A$. ■

The usual way for computing a potential function in a digraph with strongly connected components and finite length function would be the following: For every component choose an arbitrary vertex u and put $p(u) := 0$. The potential of a vertex v would be the length of a shortest directed path from the fixed vertex u in the same component to v. It follows from properties of shortest directed paths that the resulting potential function p satisfies (2).

The following algorithm uses this idea in order to construct a feasible class cardinality function of a given SER scheme $\mathcal{S}$. We assume that the digraph $G = (V, A)$ with $V = E \cup R$ and the functions α_S and β_S are given as input to the algorithm. In addition, let $G_1, \ldots, G_s$ be the components of G' and $W_1, \ldots, W_s$ their vertex sets which can be determined with the help of breath first search or the union-find algorithm (cf. Jungnickel [4] or Sedgewick [6]).

Algorithm 1.

1. $M \leftarrow 1$;
2. for $a \in A$ do
3. if $\alpha_S(a) \neq 0$ then $M \leftarrow M\alpha_S(a)$ fi;
4. od;
5. for $a \in A$ do
6. if $\alpha_S(a) = 0$ then $\alpha_N(a) \leftarrow 1/M$ else $\alpha_N(a) \leftarrow \alpha_S(a)$ fi;
7. if $\beta_S(a) = \infty$ then $\beta_N(a) \leftarrow M$ else $\beta_N(a) \leftarrow \beta_S(a)$ fi;
8. od;
9. for $v \in V$ do $x^0(v) \leftarrow \infty$ od;

10. for $i = 1$ to s do
11. choose $u \in W_i$;
12. $x^0(u) \leftarrow M$;
13. $k \leftarrow 0$;
14. repeat
15. $k \leftarrow k + 1$;
16. for $e \in E$ do
17. $x^k(e) \leftarrow \min(x^{k-1}(e), x^{k-1}(r)/\alpha_N(a) \,:\, a = (e, r) \in A)$;
18. od;
19. for $r \in R$ do
20. $x^k(r) \leftarrow \min(x^{k-1}(r), \beta_N(a)x^{k-1}(e) \,:\, a = (e, r) \in A)$;
21. od;
22. until $x^k(v) = x^{k-1}(v)$ for all $v \in W_i$;
23. $t_i \leftarrow k$;
24. for $v \in W_i$ do $x(v) \leftarrow x^{t_i}(v)$ od;
25. od.

Theorem 3 *Let $\mathcal{S}$ be an SER scheme without critical directed cycles. Then Algorithm 1 terminates and finds a feasible class cardinality function x.*

Proof. The digraph G forms together with the functions α_N and β_N, which we compute in Steps 6 and 7, the normalized scheme $\mathcal{N}$ of $\mathcal{S}$. According to Lemma 1 the scheme $\mathcal{N}$ contains no critical directed cycles.

It is easy to see, that the length function l_N defined on the edge set of G' is finite. Hence, we can use the Ford-Bellman algorithm [2] for computing a potential function in G'. In steps 9 to 25 we formulate this algorithm directly for the original problem.

Consider the component G'_i $(i = 1, \ldots, s)$ of G'. By induction on k one can show, that $\log x^k(v) - \log M$ equals the length of a shortest directed path from the chosen vertex u in W_i to v using at most k edges (cf. [4, 3.6.8]). Since $\mathcal{N}$ has no critical cycles this shortest path contains no directed cycles, and this implies that $x^k(v)$ is an integer for $k = 1, \ldots, t_i$ (the numbers $\alpha_N(a)$ appear at most once in the denominator). Since the shortest directed path from u to v cannot contain more than $|W_i| - 1$ edges we have $t_i \leq |W_i|$, i.e. the algorithm terminates for every component after a finite number of steps.

It remains to show, that the resulting function x is a feasible class cardinality function for $\mathcal{S}$. By the definition of α_N and β_N, we know that the function x is finite and integral, i.e. a class cardinality function. Moreover, we have

$$\begin{aligned} x(e) &= x^{t_i}(e) \leq x^{t_i-1}(r)/\alpha_N(a) \\ &= x^{t_i}(r)/\alpha_N(a) = x(r)/\alpha_N(a) \text{ for all } e \in E \cap W_i \text{ and all } a=(e,r) \in A, \\ x(r) &= x^{t_i}(r) \leq \beta_N(a)x^{t_i-1}(e) \\ &= \beta_N(a)x^{t_i}(e) = \beta_N(a)x(e) \text{ for all } r \in E \cap W_i \text{ and all } a=(e,r) \in A, \end{aligned}$$

where $i = 1, \ldots, s$. Hence, x is feasible for $\mathcal{N}$ and, by Lemma 2, also for $\mathcal{S}$. ■

Remark. The scheme $\mathcal{N}$ has a critical directed cycle iff the condition in Step 22 is not satisfied after $|W_i|$ iterations for at least one $i = 1, \ldots, s$. So the algorithm either detects a critical directed cycle or provides a solution to the problem. By Lemma 1, this immediately implies the following corollary which is equivalent to Theorem 3 of Lenzerini and Nobili [5].

Corollary 4 *An SER scheme has a feasible class cardinality function iff it contains no critical directed cycle.*

4 The lattice structure

Our aim in this section is to investigate the structure of the set of feasible class cardinality functions for a given SER scheme $\mathcal{S}$. We shall show that this set forms a distributive lattice. The necessary definitions can be found e.g. in [1].

Given two class cardinality functions x and y, we say that y *dominates* x and write $x \leq y$ iff, for all classes v, $x(v) \leq y(v)$. Henceforth, we denote by X the set of all feasible class cardinality functions. It is easy to see, that X forms a poset together with the dominance relation defined above. We shall denote this poset also by X.

Theorem 5 *The poset X is a distributive lattice.*

Proof. Let C be the chain $(1 < 2 < 3 < \ldots)$. Obviously, X is an induced subposet of the direct product $P = \prod_{v \in V} C_v$ where C_v is a copy of C for every

$v \in V$. It remains to show that X is a sublattice of P (cf. [1, Theorem 2.8]). So we must verify that, for any two elements of X, also their meet and join (in P) belong to X. For feasible class cardinality functions x and y their join $x \vee y$ (in P) is given by $(x \vee y)(v) = \max(x(v), y(v))$ for all $v \in V$.

Now $x \vee y$ is also a feasible class cardinality function since for every edge $a = (e,r)$ we have $\alpha(a)x(e) \leq x(r) \leq \beta(a)x(e)$ and $\alpha(a)y(e) \leq y(r) \leq \beta(a)y(e)$. This implies

$$\begin{aligned} \alpha(a)(x \vee y)(e) &= \max(\alpha(a)x(e), \alpha(a)y(e)) \\ &\leq (x \vee y)(r) = \max(x(r), y(r)) \\ &\leq \max(\beta(a)x(e), \beta(a)y(e)) = \beta(a)(x \vee y)(e). \end{aligned}$$

In the same way we may argue for the meet $x \wedge y$ which is given by $(x \wedge y)(v) = \min(x(v), y(v))$, $v \in V$. ■

5 The minimum class cardinality problem

For satisfiable SER schemes there exist an infinite number of different feasible class cardinality functions. It is easy to see, that there is always a minimum feasible class cardinality function (with respect to the dominance relation defined in Section 4). Algorithm 1 in Section 3 usually does not find this one. In order to determine the minimum feasible class cardinality function, we present the following algorithm. We assume, that the SER scheme $\mathcal{S}$ is given by the digraph $G = (V, A)$ with $V = E \cup R$ and the functions α_S and β_S as input to the algorithm. As usually, $\lceil x \rceil$ denotes the smallest integer not smaller then x.

Algorithm 2.

1. $k \leftarrow 0$;
2. for $v \in V$ do $x^0(v) \leftarrow 1$ od;
3. repeat
4. $\quad k \leftarrow k+1$;
5. $\quad$ for $e \in E$ do
6. $\quad\quad x^k(e) \leftarrow \max(x^{k-1}(e), \lceil x^{k-1}(r)/\beta_S(a) \rceil \,:\, a = (e,r) \in A)$;
7. $\quad$ od;
8. $\quad$ for $r \in R$ do

$\quad\quad x^k(r) \leftarrow \max(x^{k-1}(r), \alpha_S(a)x^{k-1}(e) \,:\, a=(e,r) \in A)$;
$\quad$ od;
until $x^k(v) = x^{k-1}(v)$ for all $v \in V$;
$t \leftarrow k$;
for $v \in V$ do $x(v) \leftarrow x^t(v)$ od.

Theorem 6 *Let $\mathcal{S}$ be an SER scheme without critical directed cycles. Then Algorithm 2 terminates and finds the minimum feasible class cardinality function x.*

Proof. To begin with, we will show, that the class cardinality functions x^k, $k = 1, \ldots, t$ are dominated by each feasible class cardinality function y of $\mathcal{S}$. We use induction on k. For $k = 0$ the statement follows immediately from condition (R2).

Let $0 < k \leq t$ and suppose y dominates x^{k-1}. This implies $y(v) \geq x^{k-1}(v)$ for all $v \in V$.

By (1) and the induction hypothesis we know

$$\begin{aligned} y(e) &\geq y(r)/\beta_S(a) \geq x^{k-1}(r)/\beta_S(a) \text{ for all } e \in E \text{ and all } a=(e,r) \in A, \\ y(r) &\geq \alpha_S(a)y(e) \geq \alpha_S(a)x^{k-1}(e) \text{ for all } r \in E \text{ and all } a=(e,r) \in A. \end{aligned}$$

Since $y(v) \in I\!N$ for all $v \in V$, we obtain

$$\begin{aligned} y(e) &\geq \max(x^{k-1}(e), \lceil x^{k-1}(r)/\beta_S(a) \rceil \,:\, a=(e,r) \in A) = x^k(e) \text{ for all } e \in E, \\ y(r) &\geq \max(x^{k-1}(r), \alpha_S(a)x^{k-1}(e) \,:\, a=(e,r) \in A) = x^k(r) \text{ for all } r \in R. \end{aligned}$$

This completes the induction step. Since $\mathcal{S}$ is satisfiable it has at least one realizer $\mathcal{R}$. Thus, $x(v)$ is bounded from above by $x_R(v)$ for all $v \in V$ and Algorithm 2 terminates. It remains to prove, that x is feasible. We have

$$\begin{aligned} x(e) &= x^t(e) \geq x^{t-1}(r)/\beta_S(a) \\ &= x^t(r)/\beta_S(a) = x(r)/\beta_S(a) \text{ for all } e \in E \text{ and all } a=(e,r) \in A, \\ x(r) &= x^t(r) \geq \alpha_S(a)x^{t-1}(e) \\ &= \alpha_S(a)x^t(e) = \alpha_S(a)x(e) \text{ for all } r \in R \text{ and all } a=(e,r) \in A. \end{aligned}$$

This implies $\alpha_S(a)x(e) \leq x(r) \leq \beta_S(a)x(e)$ for all $a = (e,r) \in A$, and by $x(v) \in I\!N$ for all $v \in V$, we obtain the feasibility of the class cardinality function x. Hence, x is indeed the minimum feasible class cardinality function of the given satisfiable SER scheme $\mathcal{S}$. ■

Remark. A realizer of a given SER scheme whose class cardinality function is just the minimum feasible class cardinality function has the minimum total number of instances as well as the minimum number of instances for each of the entities and relationships in the SER scheme.

Acknowledgement

We are thanking B. Thalheim for introducing us into this subject and for helpful discussions.

References

[1] M. Aigner, Combinatorial Theory (Springer, Berlin, 1979).

[2] R.E. Bellman, On a routing problem, Quart. Appl. Math. 16 (1958) 87-90.

[3] P. Chen and H.-D. Knoell, Der Entity-Relationship-Ansatz zum logischen Systementwurf (BI-Wissenschaftsverlag, Mannheim, 1991).

[4] D. Jungnickel, Graphen, Netzwerke und Algorithmen (BI-Wissenschaftsverlag, Mannheim, 1990).

[5] M. Lenzerini and P. Nobili, On the satisfiability of dependency constraints in Entity-Relationship schemata, Information Systems Vol.15, 4 (1990) 453-461.

[6] R. Sedgewick, Algorithms (Addison-Wesley, Reading, Mass., 1988).

[7] B. Thalheim, Foundations of Entity-Relationship Modeling, Annals of Mathematics and Artificial Intelligence 6 (1992).

[8] B. Thalheim, Fundamentals of Entity-Relationship Models (Springer, Berlin, 1993).