

Finding the Transitive Closure of Functional Dependencies using Strategic Port Graph Rewriting

János Varga

Department of Informatics

King's College London

janos.varga@kcl.ac.uk

We present a new approach to the logical design of relational databases, based on strategic port graph rewriting. We show how to model relational schemata as attributed port graphs and provide port graph rewriting rules to perform computations on functional dependencies. Using these rules we present a strategic graph program to find the transitive closure of a set of functional dependencies. This program is sound, complete and terminating, assuming that there are no cyclical dependencies in the schema.

1 Introduction

Traditionally, steps of relational database design include Conceptual, Logical and Physical modelling. The theory behind these steps is well-understood and is part of the syllabus of many databases courses. Yet, database professionals often consider Logical Design (normalisation) too cumbersome and do not apply normalisation theory, despite the clear advantages of normalised database designs. Badia and Lemire [4] also highlight that conceptual and logical models do not always carry enough information about database semantics thereby leading the architect to a sub-optimal design.

We use *attributed port graphs* to represent a relational schema and its semantics. Port graphs are graphs where edges are connected to nodes at specific points, called ports. In port graphs, nodes, edges and ports can have attributes, which are used to represent properties of the system modelled. In this paper we focus on using port graphs in the logical phase of database design. We show that port graphs are a good choice of data structure to store relational metadata and can be transformed without the loss of metadata.

To specify the transformations applied to relational schemata, we use *port graph rewriting systems*, a general class of graph rewriting systems [10]. The implementation framework we use is PORGY [12] – a visual, interactive tool for the specification, simulation and analysis of systems based on port graph rewriting. PORGY provides a graphical interface, where users can define a system and specify its dynamics by means of port graph rewrite rules and strategies. Port graphs have node, port and edge *attributes*, whose values are taken into account in port graph morphisms (used to define rewriting steps) and in strategy expressions (to control the application of rules). *Strategic graph programs* [12], consisting of an initial port graph and a set of rewrite rules controlled by a strategy, are the essence of PORGY. The strategy language offers separate primitives to select subgraphs of the model as focusing positions for rewriting and to select the rewrite rules to be applied, following the separation of concerns principle which makes programs easier to maintain and adapt. The strategy language also allows users to define strategies using not only operators to combine graph rewriting rules but also operators to deal with graph traversal and management of rewriting positions in a graph. PORGY provides a visual representation of

the set of rewrite derivations (a *derivation tree*) and includes features such as cycle detection, to facilitate debugging.

We extend the rule language by adding the possibility to specify application conditions for a rule. That is, as part of a rewrite step, in the rule editor, users can define a set of conditions which are evaluated after a morphism has been found. The rewrite step is applied only if the rule condition evaluates to true. As a use-case we provide a set of port graph rewriting rules and a strategy to calculate the transitive closure of a set of functional dependencies. Although there are a number of tools already available to do the same, a distinctive advantage of our implementation is that it is visual and backtrackable (thanks to the derivation tree feature of PORGY). Our strategy is sound, complete and terminating, given the restriction that there are no cyclical dependencies in the schema.

Summarising, our contributions are:

1. a new visual language, based on port graphs, for logical design of relational schemata,
2. generic application conditions for rules (a port graph rewriting language extension),
3. a strategic graph program to find the transitive closure of a set of functional dependencies.

This last contribution is a key step towards building strategic graph programs to find minimal covers, candidate keys and Third Normal Form (3NF) relation schemata.

Related Work. Graph theory and graph rewriting is by no means a new addition to the set of tools that have been used for relational database design. In [11] hypergraphs are used and their well-formed property (called a canonical hypergraph) determines the quality of design they represent. The authors of [16] used directed graphs to find all candidate keys of a relation in polynomial time. A special family of labelled graphs, FD-graphs, were introduced in [2] to obtain meaningful closures of a set of functional dependencies. In terms of graph transformations and rewriting we highlight two works. Hypergraph rewriting was used in [5] for the manipulation of functional dependencies and Triple Graph Grammars were used in [13] to optimize an already existing schema. Section 3 translates the work presented in [2] to port graphs and extends it.

Organization. This paper is organised as follows. We briefly review relational database theory and port graph rewriting background in Section 2. We present our port graph visual language for logical design of relation schemata in Section 3. In Section 4 we define the syntax of the language for generic rule application conditions. Section 5 illustrates how the visual language and the generic rule application conditions can be used to find the transitive closure of a set of FDs. We then conclude in Section 6 by highlighting how these results can be used in future work.

2 Background

In this section we will review the definitions and background in relational databases and port graph rewriting that we are going to use throughout this paper. Due to space constraints, for formal definitions and proofs, this section will refer the reader to the relevant works rather than recalling them. We will also briefly review related work that used graphs to represent or transform relational schemata.

2.1 Relational Database Design

We assume that the reader is familiar with the theory of logical design of relational databases [8, 9]. In particular, the definitions of: *relation schema*, *attribute*, *candidate key* and *functional dependency* (FD) [8, 9]. We refer to a single attribute with letters from the beginning of the alphabet A, B, \dots and to

attribute sets with letters from the end of the alphabet X, Y, Z . This paper will denote the set of all FDs of a relation schema Σ_R or just Σ , where appropriate. We also assume familiarity with the inference rules of functional dependencies, also known as Armstrong’s Axioms [6]: Transitivity, Trivial Dependency, Augmentation, Union, Decomposition, Pseudotransitivity. We call the set of all FDs that can be inferred from Σ , using Armstrong’s Axioms, the *syntactic closure* Σ^+ . It was shown that Armstrong’s Axioms are sound and complete, which means that they find only and all (respectively) semantically correct dependencies. This work assumes that a) FDs are in canonical form ie. only single attributes appear on the right-hand sides of the FDs and b) there are no cyclical dependencies.

2.2 Port Graph Rewriting

An attributed port graph is a labelled attributed graph where nodes have specific points-of-connection called ports, and edges that are attached to ports. In this subsection we recall the most important port graph rewriting constructs from Sections 2 and 3 of [12], where the full formal definitions can also be found.

A *port graph rewrite rule* is a port graph $L \Rightarrow_C R$ consisting of two subgraphs L and R together with an *arrow* node that links them. Each rule is characterised by its arrow node, which has a unique name (the rule’s label), a condition *Where* restricting the rule’s matching, and ports to control the rewiring operations when rewriting steps are computed. Edges that run between ports of L , R and the arrow node are coloured red by PORGY to distinguish them from normal edges. We recall the definition of the *matching morphism* that states that a *match* $g(L)$ of the left-hand side is found in G if there is a total port graph morphism g from L to G such that if the arrow node has an attribute *Where* with value C , then $g(C)$ is true in G . C is of the form $\text{saturated}(p_1) \wedge \dots \wedge \text{saturated}(p_n) \wedge B$. The predicate $\text{saturated}(g(p_i))$ is true if there are no edges between $g(p_i)$ and ports outside $g(L)$ in G – this ensures that no edges will be left dangling in rewriting steps. B is a Boolean expression such that all its variables occur in L . To aid visual design of rewrite rules, PORGY allows us to name nodes in L and R but these are treated by the system as *node name variables*. This means that node name variables identify the nodes on both sides of the rule but are instantiated when $g(L)$ is found using actual values from the matching nodes.

Our contribution to the rewrite rule language is to provide the functionality of generic application conditions. This task was two-fold: we created a grammar for B and implemented a PORGY Rule Editor plug-in (called Rule Conditions).

We also recall here that a *strategic graph program* consists of a *located graph* (a port graph with two distinguished subgraphs that specify the locations where rewriting should take place or not), a set of *located rewrite rules*, and a *strategy expression*. In a located graph G_P^Q , P represents the *position* subgraph of G where rewriting steps may take place and Q represents the *banned* subgraph of G where rewriting steps are forbidden. A located rewrite rule $L \Rightarrow_C R_M^N$ can update P and Q in a rewrite step such that $P' = (P \setminus g(L)) \cup g(M)$ and $Q' = (Q \setminus g(L)) \cup g(N)$.

Our work to find the transitive closure is implemented in the form of a strategic graph program.

2.3 Abstract Reduction Systems

We use the theoretical framework of Abstract Reduction Systems (ARS) [3, Chapter 2] to prove termination. Various techniques to construct termination proofs have been published and we shall recall one here. This technique requires the embedding of the ARS (A, \longrightarrow) into another ARS $(B, >)$ of which we know that it terminates. Our choice, $(\mathbb{N}, >)$ terminates because every descending chain $a_0 > a_1 > \dots$ is finite.

Definition 1 (Monotone mapping). *The mapping $\varphi : A \rightarrow B$ is monotone if $x \rightarrow x' \Rightarrow \varphi(x) > \varphi(x')$. φ is also known as the measure function.*

It does not automatically follow from the above that to prove termination of an abstract reduction system it suffices to find a measure function φ to embed the system into, for example, $(\mathbb{N}, >)$. But if we can prove that the ARS is finitely branching then we can make use of the following lemma:

Lemma 2. *A finitely branching reduction terminates iff there is a monotone embedding into $(\mathbb{N}, >)$.*

3 A Visual Language for Relational Schema Design

We now show how relational schemata (using functional dependencies only) can be modelled as attributed port graphs. We use the '.' (dot, member-of) operator to refer to a particular port of a node.

We define relation schema attributes and FDs as nodes. The fact that an attribute belongs to the right- or left-hand side of a FD is represented by edges. However, when adding FDs to the visual language, we face a challenge. Because of the semantics of a FD (i.e. LHS determines RHS), strategic graph programs executed on this visual language have to be able to distinguish between LHS and RHS attributes of FDs. A non-trivial FD must have at least one attribute on both sides where $(\text{RHS} \not\subseteq \text{LHS})$. Also, as per the separation of concerns principle, a FD has to be aware of the list of attributes on its sides, not the other way around. Formally, we say that:

Definition 3 (Functional Dependency Port Graph, FDPG). *Let R be a relation schema and Σ its set of functional dependencies. A Functional Dependency Port Graph representing Σ is an attributed port graph [12] $G_\Sigma = (V, P, E, D)$ and is defined as:*

- $V = V_A \cup V_{FD}$ is a union of two disjoint sets of nodes:
 - V_A : set of Attribute nodes, one node for every attribute in R ;
 - V_{FD} : set of Functional Dependency nodes, one node for every functional dependency in Σ ;
- $P = P_A \cup P_{FD}$ is a union of two defined sets of ports:
 - $P_A = \{pFD\}$ and
 - $P_{FD} = \{pFDLHS, pFDRHS\}$;
- E is a finite set of edges between ports; two ports may be connected by only one edge;
- D a set of records [12];

and a set \mathcal{F} of functions Connect, Attach and \mathcal{L} such that:

- **Connect:** for each edge $e \in E$, $\text{Connect}(e)$ is the pair (p_1, p_2) of ports connected by e where the only allowed pairs are $(pFD, pFDLHS)$ and $(pFDRHS, pFD)$. For every dependency $\varphi \in \Sigma : X \rightarrow A$ the pFD port of every attribute node corresponding to X will be connected to the $pFDLHS$ port of the dependency node corresponding to φ and the $pFDRHS$ port of the FD node φ will be connected to the pFD port of the attribute node representing A .
- **Attach:**
 - for each port $p \in P_A$, $\text{Attach}(p)$ is the node $n \in V_A$ to which the port belongs;
 - for each port $p \in P_{FD}$, $\text{Attach}(p)$ is the node $n \in V_{FD}$ to which the port belongs;
- \mathcal{L} a labelling function [12].

The following properties directly follow from Definition 3.

Property 4 (Cardinality of set V). In a FDPG $G_\Sigma = (V, P, E, D)_{\mathcal{F}}$, $|V| = |V_A| + |V_{FD}| = |R| + |\Sigma|$.

Proof. The mappings $R \rightarrow V_A$ and $\Sigma \rightarrow V_{FD}$ are bijections. \square

Property 5 (Cardinality of set E). Given the set of functional dependencies $\Sigma = \{\phi_1, \dots, \phi_k\}$ and a FDPG $G_\Sigma = (V, P, E, D)_{\mathcal{F}}$, $|E| = |\Sigma| + \sum_{i=1}^k |LHS(\phi_i)|$.

Proof. Number of edges = one right-hand side edge per dependency + sum of the sizes of the left-hand side of each functional dependency. \square

We implement FDPGs in PORGY. Firstly, using the set D of records, we introduce an attribute called *RelDbType* which denotes the role of the node in the relational context. Every new node and port created in a FDPG-based logical model have to have a constant *RelDbType* value, placing it in the appropriate set of V_A, V_{FD}, P_A or P_{FD} . Attribute nodes have *RelDbType* = ATTR, FD nodes have *RelDbType* = FD. The port of an attribute that handles the connection to either side of a FD has *RelDbType* = pFD. The LHS and RHS connection ports of a FD node have *RelDbType* = FDLHS and FDRHS, respectively. Both FDLHS and FDRHS ports have an integer attribute *FunctionalArity* defined that allows the system to store the number of attributes on each side. This is required because the matching algorithm does not enforce exact arity since a particular port can be connected to other ports outside the match found, however, when matching on FDs and their LHSs, every single LHS attribute has to be in the matching subgraph.

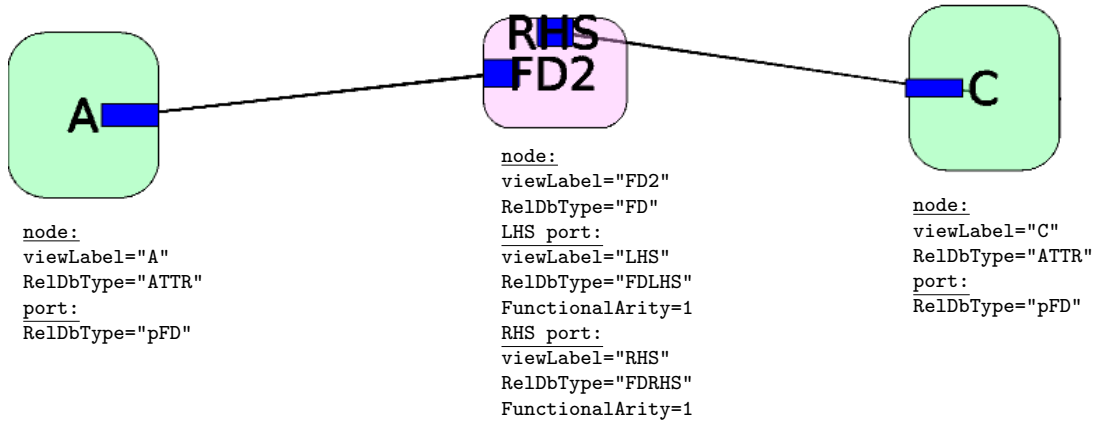


Figure 1: The functional dependency $A \rightarrow C$.

Both FD and ATTR nodes have an integer UID attribute that allows the rules to assign a numeric identity value to them. This is useful when a rule adds a new FD and we want control over the value of the unique identifier. We make use of the built-in *viewLabel* attribute to assign meaningful node name constants (in models) and variables (in rules) to nodes. Figure 1 shows a functional dependency $A \rightarrow C$ as a port graph, with the relevant attribute values.

Based on the FD-path definition of Ausiello et al. [2] we define an FDPG-path as follows:

Definition 6 (FDPG-Path). Given an FDPG $G_\Sigma = (V, P, E, D)_{\mathcal{F}}$, an attribute node set $X \subseteq V_A$ and an attribute node $j \in V_A$, a (directed) FDPG-Path $\langle X, j \rangle$ from X to j is a minimal subgraph $G'_\Sigma = (V', P', E', D')_{\mathcal{F}'}$ of G_Σ such that $X \cup \{j\} \subseteq V'_A$ and one of the following conditions holds:

1. there exists $v \in V'_{FD}$ such that for all $x_i \in X$, $(x_i.pFD, v.pFDLHS) \in E'$ and $(v.pFDRHS, j.pFD) \in E'$, i.e. there exists $v \in V'_{FD}$ such that X is the set of all left-hand side attributes and j is the right-hand side attribute of the functional dependency represented by v ;

2. there exist $v \in V_{FD}'$ and $k \in V_A'$ such that $(k.pFD, v.pFDLHS) \in E'$ and $(v.pFDRHS, j.pFD) \in E'$ and there is an FDPG-Path $\langle X, k \rangle$ included in G_Σ' ;
3. there exist $v \in V_{FD}'$ and $K \subseteq V_A'$ such that for all $k_i \in K$, $(k_i.pFD, v.pFDLHS) \in E'$ and $(v.pFDRHS, j.pFD) \in E'$ and n FDPG-Paths $\langle X, k_1 \rangle, \dots, \langle X, k_n \rangle$ are included in G_Σ' .

Property 7 (Transitivity). Condition 2 of Definition 6 represents Armstrong's Transitivity axiom.

Proof. An FDPG-Path $\langle X, k \rangle$ represents FD $X \rightarrow k$.

Edges $(k.pFD, v.pFDLHS) \in E$ and $(v.pFDRHS, j.pFD) \in E$ represent FD $k \rightarrow j$. The fact that there is an FDPG-Path $\langle X, j \rangle$ means the FD $X \rightarrow j$ exists. \square

Property 8 (Union). Condition 3 of Definition 6 represents Armstrong's Union axiom.

Proof. The n FDPG-Paths $\langle X, k_1 \rangle, \dots, \langle X, k_n \rangle$ represent FDs $X \rightarrow k_1, \dots, X \rightarrow k_n$.

Edges $k_i \in K$, $(k_i.pFD, v.pFDLHS) \in E$ and $(v.pFDRHS, j.pFD) \in E$ represent FDs $k_1 \rightarrow j, \dots, k_n \rightarrow j$. The fact that n FDPG-Paths $\langle X, j \rangle$ exist means that $X \rightarrow k_1, \dots, X \rightarrow k_n$ can be unified into $X \rightarrow K$. \square

We now turn our attention to extending the port graph rewriting language so that strategic graph programs can be created to find the transitive closure of a Functional Dependency Port Graph.

4 Generic Rule Application Conditions

We recall the structure of the arrow node *Where* attribute defined in Section 2.2. In this section we extend the rewrite rule language to provide the functionality of generic application conditions. Firstly, we define the context-free grammar for B , secondly, we present the PORGY Rule Editor plug-in (called Rule Conditions).

The EBNF grammar for B is defined in Figure 2. The structure of the grammar was inspired by C++ and follows the operator precedence of C++, too.

We point out that when referring to a node or edge the user has to use its internally assigned id. Also, due to the implementation of PORGY, there is no `port` construct in the grammar – they have to be referred to as `node`. This is because the underlying graph engine (TULIP) processes ports as nodes. Terminal number can be any integer or floating-point number and `quoted_string` is an arbitrary-length string made up of letters, digits and symbols in double quotes. Similarly, `quoted_attribute_name` is a valid name of an attribute of `node`, `edge` or `port`.

We highlight the `NotNode()` operator: it iterates all nodes of G and checks if there exists a node with an attribute `quoted_attribute_name` and if the comparison on them evaluates to true. Intuitively, if at least one such node is found in G , `NotNode()` returns false. It is very important to note here that this check is performed on the entire graph G , not just in $g(L)$. This is fundamentally different from the rest of the rule application condition grammar, which only applies to $g(L)$. This is a consequence of the definition of the port graph rewrite rule which states that all variables in the Boolean expression of the *Where* attribute have to occur in L , so that the matching algorithm can work with them. When a match $g(L)$ is found, all variables of B are mapped so that their actual values can be found. However, when checking the absence of a node, we are not constrained by this, because we are not specifying a node in `NotNode()` – we are only specifying an attribute comparison that *must* evaluate to false on all nodes of G .

PORGY offers a modular plug-in system allowing developers to create Python/C++ plug-ins. We implemented an LL-parser for the above detailed context-free grammar in C++ using the Boost Spirit Parser

$\langle node \rangle$::= 'n(' valid LHS node id ')'
$\langle edge \rangle$::= 'e(' valid LHS edge id ')'
$\langle element\ attribute \rangle$::= ($\langle node \rangle$ $\langle edge \rangle$) '.' quoted_attribute_name
$\langle factor \rangle$::= number quoted_string '(' $\langle expression \rangle$ ')' '!' $\langle factor \rangle$ $\langle element\ attribute \rangle$ 'max(' $\langle expression \rangle$ ',' $\langle expression \rangle$ ')' 'min(' $\langle expression \rangle$ ',' $\langle expression \rangle$ ')' 'random(' $\langle factor \rangle$ ')'
$\langle term \rangle$::= $\langle factor \rangle$ { ('*' '/' '%') $\langle factor \rangle$ }
$\langle expression \rangle$::= $\langle term \rangle$ { ('+' '-') $\langle term \rangle$ }
$\langle comp\ operator \rangle$::= '==' '!=' '>' '<' '>=' '<='
$\langle comparison \rangle$::= $\langle expression \rangle$ $\langle comp\ operator \rangle$ $\langle expression \rangle$ 'NotNode('quoted_attribute_name $\langle comp\ operator \rangle$ $\langle expression \rangle$)'
$\langle logical\ expression \rangle$::= $\langle logical\ term \rangle$ { ' ' $\langle logical\ term \rangle$ }
$\langle logical\ term \rangle$::= $\langle logical\ factor \rangle$ { '&&' $\langle logical\ factor \rangle$ }
$\langle logical\ factor \rangle$::= $\langle comparison \rangle$ '!' $\langle logical\ factor \rangle$ '(' $\langle logical\ expression \rangle$ ')'
$\langle rule\ condition \rangle$::= { $\langle logical\ expression \rangle$ };

Figure 2: The rule application condition grammar.

Framework. This framework generates and executes the parser design-time and builds and evaluates an abstract syntax tree run-time. When evaluated, the Boolean result is ANDed to the rest of the arrow node *Where* attribute by the matching algorithm. The parser ensures that all nodes, edges, ports and attributes referred to in the conditions exist on the LHS of the rule.

We also added a UI extension to PORGY that allows users to specify and parse/check the rule conditions. A screenshot of PORGY with the Conditions editor is presented in Figure 3. Examples of rule conditions can be found in Section 5.

5 Transitive Closure Strategy

In this section we present the strategy to find the transitive closure of a set of FDs using the previously introduced visual language and rule conditions.

In PORGY, the starting point of port graph rewriting, the original model, is referred to as G_0 . In our case, G_0 is a Functional Dependency Port Graph, as defined in Section 3. The task is to find a relational transitive closure of G_0 , i.e. to generate all the transitive dependencies.

Inspired by the Chase Algorithm [1, 15] we *iterate* every FD in G_0 (and also those added by the rules), apply the rules that detect a transitive dependency pattern and create the new FD. Once we have every

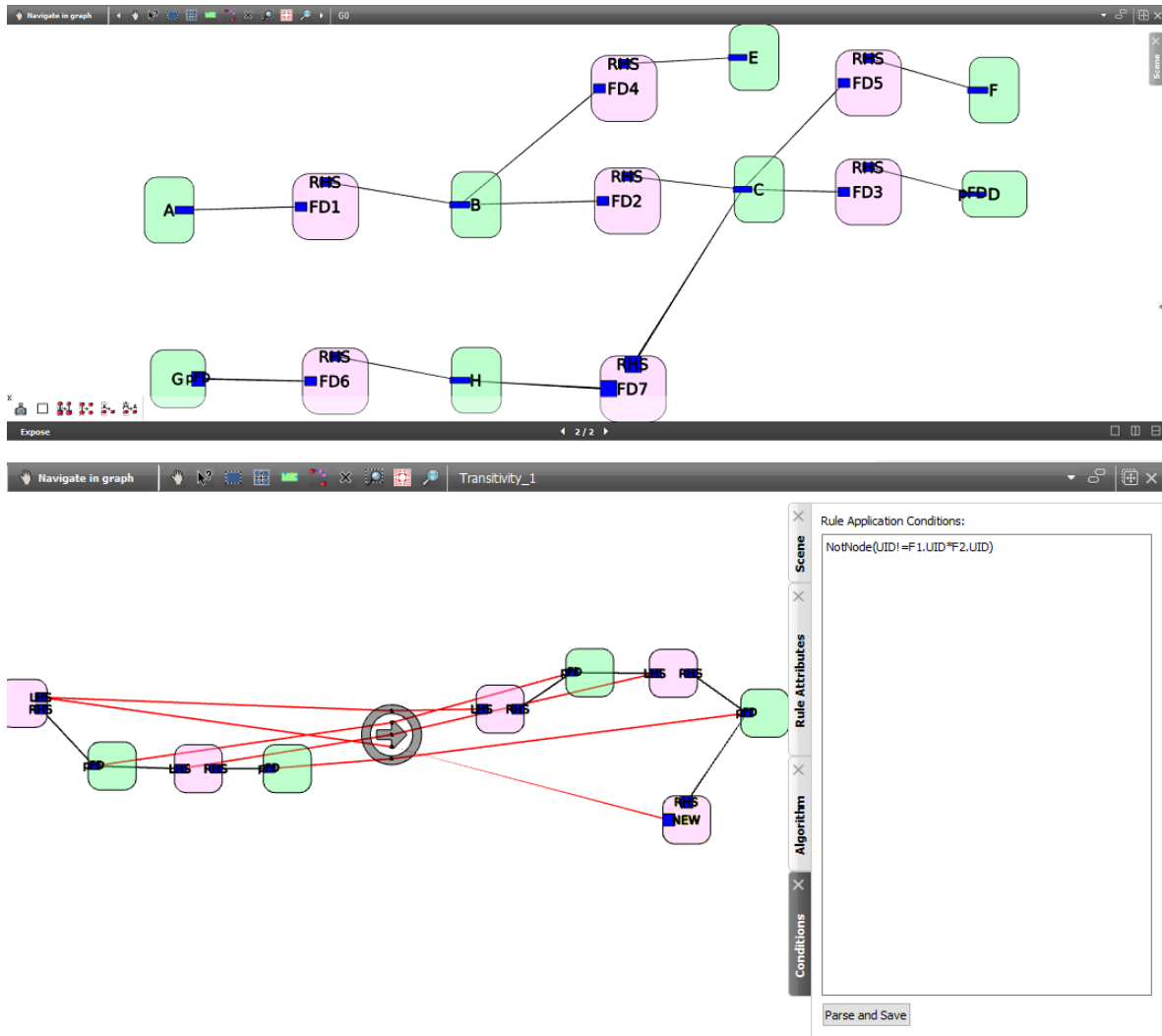
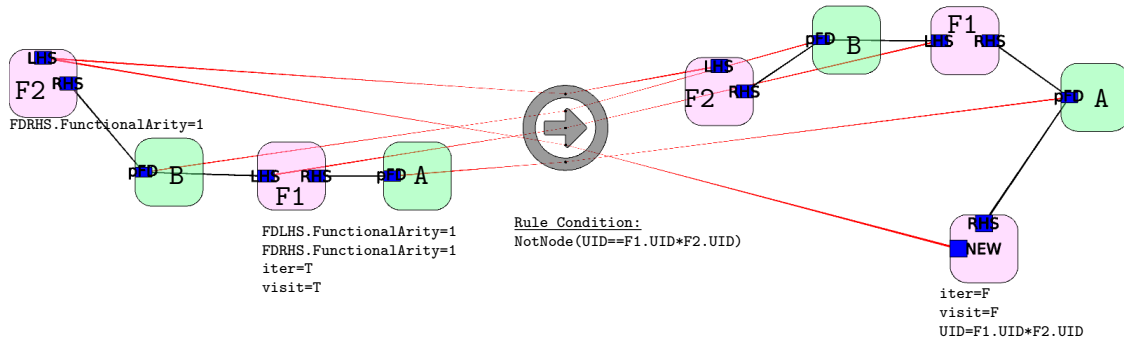
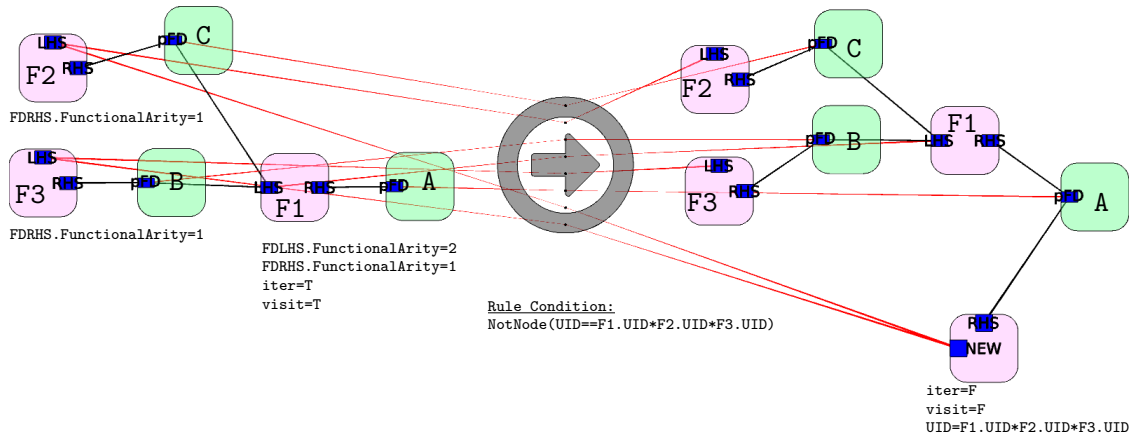


Figure 3: PORGY, an initial graph (G_0), $Transitivity_1$ rule and the conditions editor.

possible new transitive dependency that goes through the iterated FD, we mark it *visited*. We define two new Boolean node attributes: *iter* to flag the node currently being iterated and *visit* to permanently flag the node as visited. The two rules controlling the iteration are *IterOn* and *IterOff* (omitted). *IterOn* randomly selects a node with attribute values $RelDbType=FD$, $iter=false$, $visit=false$ and sets the two flags: $iter=true$, $visit=true$. *IterOff* rule selects the currently iterated FD node $RelDbType=FD$, $iter=true$, $visit=true$ and turns the iteration flag off: $iter=false$, $visit=true$.

The first two rules we created can be seen on Figures 4 and 5. These rules detect FDPG-Paths representing transitive functional dependency chains e.g. $X \rightarrow Y \rightarrow A$ which, per Armstrong’s Transitivity Axiom, means that $X \rightarrow A$ holds. Red edges in the rules go through the *bridge* arrow node ports. These edges mean that all edges from outside the matching subgraph into the port on the LHS end of the red edge are to be copied to connect to the port on the RHS end of the red edge. For example, the red edges running between the pFDLHS ports of nodes F2 in both rules and of node F3 in $Transitivity_2$ rule ensure that

Figure 4: Transitivity₁ rule.Figure 5: Transitivity₂ rule.

- the pFDLHS port of the newly created dependency will be connected to all attributes that were on the LHS of F2 (and F3);
- the pFDLHS port of the new instance of F2 (and F3) will be connected to all attributes that were on the LHS of F2 (and F3) (the original dependency is preserved).

To offer extra, context-specific backtracking functionality, we assign prime number values to the UID attribute of every FD. When a new FD is created by any transitivity rule, the UID of the new FD is the product of the UIDs of the FDs that lead to the new FD. This is calculated and assigned by the Rule Algorithm feature of PORGY. We use the Rule Condition functionality specified in Section 4 to control the applicability of the Transitivity rules. For example, the condition `NotNode(UID==F1.UID*F2.UID)` means that if there is a node in the entire graph being rewritten (not only the matching subgraph) with a value in UID equal to the product of the UIDs of F1 and F2 (meaning the transitive dependency has already been found) then the rule shouldn't apply.

Finally, the strategy that uses the above defined rules is Strategy 1. The *ResetVisitedFlags* rule sets the flags `visit=F`, to allow for a whole new loop to run on all FD nodes.

Note that the node being iterated (node name variable: F1) plays a pivotal role in the matching and its LHS subgraph increases in size as the *FunctionalArity* of its FDLHS port increases.

The Transitive Closure Strategy is explained as follows:

Strategy 1: Transitive Closure Strategy

```

1. while(match(IterOn))do(
a)   one(IterOn);
b)   repeat(one(Transitivity1));
c)   one(IterOff)
    );
2. repeat(one(ResetVisitedFlags));
3. while(match(IterOn))do(
a)   one(IterOn);
b)   repeat(one(Transitivity2));
c)   one(IterOff)
    )

```

1. For as long as there is at least one FD node the strategy hasn't *visited* and *iterated*, do
 - (a) Pick one such FD at random (with equal probabilities) and mark it *visited* and *iterated*,
 - (b) Find and apply all possible applications of the rule *Transitivity₁* with the FD node picked in the previous step as the pivotal node F1,
 - (c) Mark F1 *visited* but *not iterated*;
2. Mark all FD nodes *not visited*;
3. For as long as there is at least one FD node the strategy hasn't *visited* and *iterated*, do
 - (a) Pick one such FD at random (with equal probabilities) and mark it *visited* and *iterated*,
 - (b) Find and apply all possible applications of the rule *Transitivity₂* with the FD node picked in the previous step as the pivotal node F1,
 - (c) Mark F1 *visited* but *not iterated*;

Property 9 (Derivation Tree of Strategy 1). Every non-leaf node in the Derivation Tree of Strategy 1 has only one child node. Conversely, there is only one leaf node.

Proof. From the semantics of strategic port graph programs [12] we know that strategy constructs we use (*match()*, *while()*, *repeat(one())* and *one()*) will not branch the Derivation Tree at all. The loop constructs we use execute their arguments sequentially as many times as they apply. \square

Rules *IterOn*, *IterOff* and *ResetVisitedFlags* are omitted due to space constraints. We discuss the semantic and rewriting characteristics of Strategy 1 in Theorems 10, 11 and 12.

Theorem 10. *The Transitive Closure Strategy is sound. That is, it **only** finds functional dependencies that can be inferred from the original set of FDs using Armstrong's Axioms but ignoring the meaningless dependencies that would be generated by the Reflexivity axiom.*

Proof. Let Σ be a set of functional dependencies and $G_\Sigma = (V, P, E, D)_{\mathcal{F}}$ the FDPG representing Σ . Let $K, L \subset V_A$ be attribute node sets representing attribute sets X, Y (resp.) and $n \in V_A$ an attribute node representing attribute A . We want to show that, after executing the Transitive Closure Strategy on G_Σ obtaining G'_Σ , an FDPG-Path $\langle K, n \rangle$ exists in G'_Σ **only if** the functional dependency $X \rightarrow A$ can be inferred from Σ using Armstrong's axioms.

It follows from Definition 6 and Properties 7 and 8 that the existence of FDPG-Path $\langle X, Y, A \rangle$ in a leaf

node of the Derivation Tree means that either Σ contained the explicit dependency $X \rightarrow A$ or that the Transitive Closure Strategy added it. Our rules and strategies (in fact, PORGY itself) follow the rewriting principle that only rules can change the structure of the graph, strategy expressions can not. We ignore rules *IterOn* and *IterOff* because they only affect values of node attributes *iter* and *visit* and not alter FDPG-Paths in the graph. This means we have to show that *Transitivity_{1..k}* rules create only valid FDPG-Paths. These rules detect FDPG-Paths $\langle K, L, n \rangle$, add a new Functional Dependency node to represent the transitive dependency $X \rightarrow A$, and (using plain and red edges) ensure that all LHS and RHS attributes are connected to the ports of the new dependency node. This creates the direct path $\langle K, n \rangle$. According to Definitions 3 and 6 this new path is valid. \square

Theorem 11. *The Transitive Closure Strategy is complete. That is, it finds **all** functional dependencies that can be inferred from the original set of FDs using Armstrong’s Axioms but ignoring the meaningless dependencies that would be generated by the Reflexivity axiom.*

Proof. We know that Armstrong’s Reflexivity, Transitivity and Union rules, together, form a complete system of inference rules. We have to show that these three rules can be deduced from the Transitive Closure Strategy, or more precisely, from rules *Transitivity_{1..k}*. As noted earlier, a repeating pattern forms on the left-hand side of F1 in our Transitivity rules as k increases. A k -ary Armstrong’s Union operation is performed by *Transitivity_k*. For example, *Transitivity₂*: if the left-hand sides of F2 and F3 are the same attribute set, and the rule simply copies those edges (which it does), then the system behaves as if it performed the Union operation. It is obvious from the structure of the FDPG-Rules that they perform what Armstrong’s Transitivity rule does. \square

Theorem 12. *Let Σ be a set of functional dependencies and $G_\Sigma = (V, P, E, D)_{\mathcal{F}}$ the FDPG representing Σ . The Transitive Closure Strategy (TCS) is a terminating program that never fails. That is, assuming that there are no cyclical dependencies in Σ , there is no infinite descending chain in the Derivation Tree.*

Proof. Never fails. We know from the semantics set out in [12] that the expressions `while(C)do(S)` and `repeat(S)` never fail. `one(IterOn)` will also never fail because `while(match(IterOn))` will check if it is possible to execute it at all. `one(IterOff)` will always apply because every execution of it is preceded by an execution of *IterOn* and *Transitivity_k* rules do not change the *iter* and *visited* values of any pre-existing nodes.

Finitely branching. It follows from Property 9 that the Derivation Tree of the strategic graph program (G_Σ, TCS) is finitely branching.

Embedding. We want to find an embedding of (G_Σ, TCS) into $(\mathbb{N}, >)$. We do this by defining a measure for every loop in the strategy. For `repeat(one(Transitivityk))`, the measure is the number of possible matches of the LHS subgraph of F1. This is a good measure because a) the Rule Condition on UID prevents re-application of the rule and b) even though the NEW node is added, it is not part of the LHS subgraph of F1. For `while(match(IterOn))do(...)` the measure is $|V_{FD}|_{G_0} + |\Sigma^+| - |V_{FD}|_{G_i}$. That is, the initial number of FD nodes in G_Σ plus the size of the transitive closure of Σ less the number of FD nodes after the i th application of the loop. With one successful application of *Transitivity_k* the number of FD nodes increases therefore the measure decreases. Note that neither of these measures can be 0 or less. From these it follows that the above detailed measure provides a good monotone mapping into $(\mathbb{N}, >)$.

Then from Lemma 2 it follows that the strategy terminates. \square

6 Conclusion

Our results can be used to build a strategic graph program that takes the transitive closure as input and finds a minimal cover [2, 14]. From a minimal cover, all candidate keys of a relation schema can be found [16]. The minimal cover and the set of candidate keys can then be used as inputs to Bernstein's Synthesis Algorithm to synthesize Third Normal Form [7] schemata.

Future Work Definition 3 may be generalized such that the Attribute nodes have a list of dependency ports, each element of the list corresponding to a type of data dependency that can exist in a database (e.g. multi-valued dependency). Conversely, each dependency type requires the introduction of its own Dependency node type. Each new Dependency node type may represent bilateral dependencies (LHS and RHS) or multilateral ones. Each new dependency type will require its own Path definition.

References

- [1] Alfred V. Aho, Catriel Beeri & Jeffrey D. Ullman (1979): *The Theory of Joins in Relational Databases*. *ACM Trans. Database Syst.* 4(3), pp. 297–314. Available at <http://doi.acm.org/10.1145/320083.320091>.
- [2] Giorgio Ausiello, Alessandro D'Atri & Domenico Saccà (1983): *Graph Algorithms for Functional Dependency Manipulation*. *J. ACM* 30(4), pp. 752–766, doi:10.1145/2157.322404.
- [3] Franz Baader & Tobias Nipkow (1998): *Term rewriting and all that*. Cambridge University Press, doi:10.1017/CBO9781139172752.
- [4] Antonio Badia & Daniel Lemire (2011): *A Call to Arms: Revisiting Database Design*. *SIGMOD Rec.* 40(3), pp. 61–69, doi:10.1145/2070736.2070750.
- [5] Carlo Batini & Alessandro D'Atri (1978): *Rewriting Systems as a Tool for Relational Data Base Design*. In Volker Claus, Hartmut Ehrig & Grzegorz Rozenberg, editors: *Graph-Grammars and Their Application to Computer Science and Biology, International Workshop, Bad Honnef, October 30 - November 3, 1978, Lecture Notes in Computer Science 73*, Springer, pp. 139–154, doi:10.1007/BFb0025717.
- [6] Catriel Beeri, Ronald Fagin & John H. Howard (1977): *A Complete Axiomatization for Functional and Multivalued Dependencies in Database Relations*. In Diane C. P. Smith, editor: *SIGMOD Conference*, ACM, pp. 47–61. Available at <http://doi.acm.org/10.1145/509404.509414>.
- [7] Philip A. Bernstein (1976): *Synthesizing Third Normal Form Relations from Functional Dependencies*. *ACM Trans. Database Syst.* 1(4), pp. 277–298. Available at <http://doi.acm.org/10.1145/320493.320489>.
- [8] E. F. Codd (1970): *A Relational Model of Data for Large Shared Data Banks*. *Communications of the ACM* 13(6), pp. 377–387. Available at <http://doi.acm.org/10.1145/362384.362685>.
- [9] E. F. Codd (1971): *Normalized Data Structure: A Brief Tutorial*. In E. F. Codd & A. L. Dean, editors: *SIGFIDET Workshop*, ACM, pp. 1–17, doi:10.1145/1734714.1734716.
- [10] Bruno Courcelle (1990): *Graph Rewriting: An Algebraic and Logic Approach*. In: *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics (B)*, MIT Press, pp. 193–242. Available at <https://www.elsevier.com/books/formal-models-and-semantics/unknown/978-0-444-88074-1>.
- [11] David W. Embley & W. Y. Mok (2011): *Mapping Conceptual Models to Database Schemas*. XIX, Springer, pp. 123–164. Available at <http://www.springer.com/computer/swe/book/978-3-642-15864-3>.
- [12] Maribel Fernández, Hélène Kirchner & Bruno Pinaud (2018): *Strategic Port Graph Rewriting: an Interactive Modelling Framework*. *Mathematical Structures in Computer Science*, pp. 1–48, doi:10.1017/S0960129518000270. Available at <https://hal.inria.fr/hal-01251871>.
- [13] J. H. Jahnke & A. Zündorf (1999): *Applying Graph Transformations to Database re-engineering*. 2, World Scientific, pp. 267–286. Available at <http://www.worldscientific.com/worldscibooks/10.1142/4180>.

- [14] David Maier (1980): *Minimum Covers in Relational Database Model*. *J. ACM* 27(4), pp. 664–674, doi:10.1145/322217.322223.
- [15] David Maier, Alberto O. Mendelzon & Yehoshua Sagiv (1979): *Testing Implications of Data Dependencies*. *ACM Trans. Database Syst.* 4(4), pp. 455–469, doi:10.1145/320107.320115.
- [16] Hossein Saiedian & Thomas Spencer (1996): *An Efficient Algorithm to Compute the Candidate Keys of a Relational Database Schema*. *Comput. J.* 39(2), pp. 124–132, doi:10.1093/comjnl/39.2.124.