Weaker Forms of Monotonicity for Declarative Networking: a more fine-grained answer to the CALM-conjecture Link
Peer-reviewed author version

Made available by Hasselt University Library in Document Server@UHasselt

Reference (Published version):

DOI: 10.1145/2594538.2594541
Handle: http://hdl.handle.net/1942/16393
Weaker Forms of Monotonicity for Declarative Networking: 
a More Fine-grained Answer to the CALM-conjecture

Tom J. Ameloot* 
Hasselt University &
transnational University of Limburg 
tom.ameloot@uhasselt.be

Frank Neven 
Hasselt University &
transnational University of Limburg 
frank.neven@uhasselt.be

Bas Ketsman 
Hasselt University &
transnational University of Limburg 
bas.ketsman@uhasselt.be

Daniel Zinn 
LogicBlox, Inc 
daniel.zinn@logicblox.com

ABSTRACT

The CALM-conjecture, first stated by Hellerstein [23] and proved in its revised form by Ameloot et al. [13] within the framework of relational transducer networks, asserts that a query has a coordination-free execution strategy if and only if the query is monotone. Zinn et al. [32] extended the framework of relational transducer networks to allow for specific data distribution strategies and showed that the non-monotone win-move query is coordination-free for domain-guided data distributions. In this paper, we complete the story by equating increasingly larger classes of coordination-free computations with increasingly weaker forms of monotonicity and make Datalog variants explicit that capture each of these classes. One such fragment is based on stratified Datalog where rules are required to be connected with the exception of the last stratum. In addition, we characterize coordination-freeness as those computations that do not require knowledge about all other nodes in the network, and therefore, can not globally coordinate. The results in this paper can be interpreted as a more fine-grained answer to the CALM-conjecture.

Categories and Subject Descriptors

H.2 [Database Management]: Languages; H.2 [Database Management]: Systems—Distributed databases; F.1 [Computation by Abstract Devices]: Models of Computation

Keywords

Distributed database, relational transducer, consistency, coordination, expressive power, cloud programming

*PhD Fellow of the Fund for Scientific Research, Flanders (FWO).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PODS’14, June 22–27, 2014, Snowbird, UT, USA.
Copyright 2014 ACM 978-1-4503-2375-8/14/06 ...$15.00.
http://dx.doi.org/10.1145/2594538.2594541.

1. INTRODUCTION

Declarative networking is an approach where distributed computations are modeled and programmed using declarative formalisms based on extensions of Datalog. On a logical level, programs (queries) are specified over a global schema and are computed by multiple computing nodes over which the input database is distributed. These nodes can perform local computations and communicate asynchronously with each other via messages. The model operates under the assumption that messages can never be lost but can be arbitrarily delayed. An inherent source of inefficiency in such systems are the global barriers raised by the need for synchronization in computing the result of queries.

This source of inefficiency inspired Hellerstein [11] to formulate the CALM-principle which suggests a link between logical monotonicity on the one hand and distributed consistency without the need for coordination on the other hand. A crucial property of monotone programs is that derived facts must never be retracted when new data arrives. The latter implies a simple coordination-free execution strategy: every node sends all relevant data to every other node in the network and outputs new facts from the moment they can be derived. No coordination is needed and the output of all computing nodes is consistent. This observation motivated Hellerstein [23] to formulate the CALM-conjecture which, in its revised form1 states

“A query has a coordination-free execution strategy iff the query is monotone.”

Ameloot, Neven, and Van den Buseche [13] formalized the conjecture in terms of relational transducer networks and provided a proof. Zinn, Green, and Ludäscher [32] subsequently showed that there is more to this story. In particular, they obtained that when computing nodes are increasingly more knowledgeable on how facts are distributed, increasingly more queries can be computed in a coordination-free manner. Zinn et al. [32] considered two extensions of the original transducer model introduced in [13]. In the first extension, here referred to as the policy-aware model, every computing node is aware of the facts that should be assigned to it and can consequently evaluate negation over schema relations. In the second extension, referred to as the

1CALM stands for Consistency And Logical Monotonicity.
2The original conjecture replaced monotone by Datalog [13].
domain-guided model, data distribution is restricted as follows: each possible domain value $d$ is assigned to at least one node; and, when an input fact contains value $d$, this fact is given to all nodes that $d$ is assigned to. It was shown in \cite{22} that the coordination-free computations within the original, policy-aware, and domain-guided models form a strict hierarchy and that the non-monotone win-move query can be computed by a coordination-free domain-guided transducer network. The central objective of this paper is to characterize these increasingly larger classes of coordination-free computations in terms of increasingly weaker forms of monotonicity thereby obtaining a more fine-grained answer to the CALM-conjecture.

Towards this goal, we introduce the set of domain-distinct-monotone and the set of domain-disjoint-monotone queries, which we denote by $M_{\text{distinct}}$ and $M_{\text{disjoint}}$, respectively. Recall that a query is monotone if the output does not decrease (w.r.t. set inclusion) when new facts are added. The classes of domain-distinct-monotone and domain-disjoint-monotone queries then correspond to queries with non-decreasing output (again w.r.t. set inclusion) when only facts are added that contain at least one and only new domain elements, respectively. While $M_{\text{distinct}}$ is a reformulation of the class of queries preserved under extensions (c.f., Section 3.2), $M_{\text{disjoint}}$ appears to be a new class. We semantically characterize the coordination-free computations within the policy-aware and domain-guided model in terms of $M_{\text{distinct}}$ and $M_{\text{disjoint}}$, respectively, to obtain the following answer to the CALM-conjecture: for a query $Q$,

(i) $Q$ can be computed by a coordination-free (original) transducer network iff $Q$ is monotone; \cite{13}

(ii) $Q$ can be computed by a coordination-free policy-aware transducer network iff $Q \in M_{\text{distinct}}$; and,

(iii) $Q$ can be computed by a coordination-free domain-guided transducer network iff $Q \in M_{\text{disjoint}}$.

It is tricky to formally define coordination-freeness because ideally it should forbid communication for coordination purposes but it should allow communication to exchange data values, for instance, to compute joins. We employ the formalization of coordination-freeness as introduced in \cite{13}. While we do not claim this notion of coordination-freeness to be the only possible one, the results in this paper imply that it is a sensible one. In particular, we show that coordination-free computations can not globally coordinate across all computing nodes. This is made precise by proving that every coordination-free transducer is equivalent to one that has no knowledge of all other nodes in the network. We refer to Section 4.1.5 and Section 4.3 for a more thorough discussion.

In its original formulation \cite{22}, the CALM-conjecture did not refer to the general class of monotone queries, but rather to the monotone queries definable in Datalog. Therefore, it is interesting to investigate subclasses of Datalog with negation that remain within $M_{\text{distinct}}$ and $M_{\text{disjoint}}$, respectively. As mentioned before, $M_{\text{distinct}}$ corresponds to the well-known class of queries which are preserved under extensions, denoted by $\mathcal{E}$. Afrati et al. \cite{6} obtained that semi-positive datalog, denoted SP-Datalog, is included in $\mathcal{E}$, while Cabibbo \cite{18} showed that SP-Datalog extended with value invention captures $\mathcal{E}$ and therefore $M_{\text{distinct}}$. We show that semi-connected stratified Datalog, denoted semicon-Datalog$, ^\gamma$,

a fragment of stratified Datalog where only ‘connected’ rules are allowed (except for the last stratum) and which contains SP-Datalog, is included in $M_{\text{disjoint}}$ and that this fragment extended with value invention captures precisely $M_{\text{disjoint}}$. Furthermore, all queries definable in SP-Datalog and semicon-Datalog$^\gamma$ are coordination-free within the policy-aware and domain-guided transducer network model, respectively.

The results of this paper are summarized in Figure 2.

Outline. In Section 2 we introduce the necessary definitions. In Section 3 we investigate the classes $M_{\text{distinct}}$ and $M_{\text{disjoint}}$. In Section 4 we semantically characterize coordination-free transducer networks in the policy-aware and domain-guided models. In Section 5 we consider Datalog fragments for $M_{\text{distinct}}$ and $M_{\text{disjoint}}$. We discuss related work in Section 6 and conclude in Section 7.

2. DEFINITIONS

Queries and instances. We assume an infinite set $\text{dom}$ of data values. A database schema $\sigma$ is a collection of relation names $R$ where every $R$ has arity $ar(R)$. We call $R(d)$ a fact when $R$ is a relation name and $d$ is a tuple in $\text{dom}$. We say that a fact $(d_1,\ldots,d_k)$ is over a database schema $\sigma$ if $R \in \sigma$ and $ar(R) = k$. A (database) instance $I$ over $\sigma$ is simply a finite set of facts over $\sigma$. We denote by $\text{dom}(I)$ the set of all values that occur in facts of $I$. When $I = \{I\}$, we simply write $\text{dom}(I)$ rather than $\text{dom}(\{I\})$. By $|I|$ we denote the number of facts in $I$. A query over a schema $\sigma$ to a schema $\sigma'$ is a generic mapping $Q$ from instances over $\sigma$ to instances over $\sigma'$. Genericity means that for every permutation $\pi$ of $\text{dom}$ and every instance $I$, $Q(\pi(I)) = \pi(Q(I))$. For a set of facts $I$ and a schema $\sigma$, we write $I_{\sigma}$ to denote the maximal subset of $I$ that is over $\sigma$.

For convenience, we restrict our attention to schemas for which all relations have arity greater than zero. In particular, this means that queries, and therefore also Datalog programs, can not define nullary relations. We address how the addition of nullary facts changes our results in Section 7.

Datalog with negation. Let $\text{var}$ be the universe of variables, disjoint from $\text{dom}$. An atom is of the form $R(u_1,\ldots,u_k)$ where $R$ is a relation name and each $u_i \in \text{var}$. We call $R$ the predicate. A literal is an atom or a negated atom and is called positive in the former case and negative in the latter case.

We recall Datalog with negation \cite{3}, abbreviated Datalog$^\gamma$. Formally, a Datalog$^\gamma$ rule $\varphi$ is a quadruple $(\text{head}_\varphi, \text{pos}_\varphi, \text{neg}_\varphi, \text{ineq}_\varphi)$ where $\text{head}_\varphi$ is an atom; $\text{pos}_\varphi$ and $\text{neg}_\varphi$ are sets of atoms; $\text{neg}_\varphi$ is a set of inequalities $(u \neq v)$ with $u, v \in \text{var}$; and, the variables of $\varphi$ all occur in $\text{pos}_\varphi$. The components $\text{head}_\varphi$, $\text{pos}_\varphi$, and $\text{neg}_\varphi$ are called respectively the head, the positive body atoms and the negative body atoms. We refer to $\text{pos}_\varphi \cup \text{neg}_\varphi$ as the body atoms. Note, $\text{neg}_\varphi$ contains just atoms, not negative literals. Every Datalog$^\gamma$ rule $\varphi$ must have a head, $\text{pos}_\varphi$ must be non-empty and $\text{neg}_\varphi$ may be empty. If $\text{neg}_\varphi = \emptyset$ then $\varphi$ is called positive. Of course, a rule $\varphi$ may be written in the conventional syntax. For instance, if $\text{head}_\varphi = T(u,v)$, $\text{pos}_\varphi = \{R(u,v)\}$, $\text{neg}_\varphi = \{S(v)\}$, and $\text{ineq}_\varphi = \{u \neq v\}$, with $u, v \in \text{var}$, then we can write $\varphi$ as $T(u,v) \leftarrow R(u,v), \neg S(v), u \neq v$. The set of variables of $\varphi$ is denoted $\text{vars}(\varphi)$. A rule $\varphi$ is said to be over schema $\sigma$ if for each atom $R(u_1,\ldots,u_k) \in$
We define \( \text{sch}(P) \) to denote the (minimal) database schema that \( P \) is over. We define \( \text{idb}(P) \subseteq \text{sch}(P) \) to be the database schema consisting of all relations in rule-heads of \( P \). We abbreviate \( \text{edb}(P) = \text{sch}(P) \setminus \text{idb}(P) \). As usual, the abbreviation “\( \text{idb} \)” stands for “intensional database schema” and “\( \text{edb} \)” stands for “extensional database schema”. A valuation for a rule \( \varphi \) in \( P \) w.r.t. an instance \( I \) over \( \text{edb}(P) \), is a total function \( V : \text{vars}(\varphi) \to \text{dom} \). The application of \( V \) to an atom \( R(u_1, \ldots, u_k) \) of \( \varphi \), denoted \( V(R(u_1, \ldots, u_k)) \), results in the fact \( R(a_1, \ldots, a_k) \) where \( a_i = V(u_i) \) for each \( i \in \{1, \ldots, k\} \).

This is naturally extended to a set of atoms, which results in a set of facts, in a set of atoms, which results in a set of facts. The valuation \( V \) is said to be satisfying for \( \varphi \) on \( I \) if \( V(\text{pos}_\varphi) \subseteq I \), \( V(\text{neg}_\varphi) \cap I = \emptyset \), and \( V(u) \neq V(v) \) for each \( (u \neq v) \in \text{ineq}_\varphi \). If so, \( \varphi \) is said to derive the fact \( V(\text{head}_\varphi) \).

Positive and semi-positive Datalog. A Datalog\(^-\) program \( P \) is positive if all rules of \( P \) are positive. We say that \( P \) is semi-positive if for each rule \( \varphi \in P \), the atoms of \( \text{neg}_\varphi \) are over \( \text{edb}(P) \). We now give the semantics of a semi-positive Datalog\(^-\) program \( P \). Let \( T_P \) be the immediate consequence operator that maps each instance \( J \) over \( \text{sch}(P) \) to the instance \( J' = J \cup A \) where \( A \) is the set of facts derived by all possible satisfying valuations for the rules of \( P \) on \( J \). Let \( I \) be an instance over \( \text{edb}(P) \). Consider the infinite sequence \( I_0, I_1, I_2, \ldots \) inductively defined as follows: \( I_0 = I \) and \( I_i = I_{i-1} \cup T_P(I_{i-1}) \) for each \( i \geq 1 \). The output of \( P \) on \( I \) denotes \( \{\text{pos}_\varphi \setminus \text{idb}(P) \mid \varphi \in \text{sch}(P) \} \) as \( |P| = \cup_i I_i \); this is the minimal fixpoint of the \( T_P \) operator.

We denote by \( \text{Datalog} \), \( \text{Datalog}(\neq) \), and \( \text{SP-Datalog} \) the class of positive Datalog\(^-\) programs without inequalities, the positive Datalog\(^-\) programs, and the class of semi-positive Datalog\(^-\) programs, respectively. Note that the last two classes may use inequalities.

Stratified semantics. We say that \( P \) is syntactically stratifiable if there is a function \( \rho : \text{sch}(P) \to \{\ldots, \text{idb}(P)\} \) such that for each rule \( \varphi \in P \), having some head predicate \( T \), the following conditions are satisfied: (1) \( \rho(R) \leq \rho(T) \) for each \( R(u) \in \text{pos}_\varphi \cap \text{idb}(P) \); and, (2) \( \rho(R) < \rho(T) \) for each \( R(u) \in \text{neg}_\varphi \cap \text{idb}(P) \). For \( R \in \text{idb}(P) \), we call \( \rho(R) \) the stratum number of \( R \). Intuitively, \( \rho \) partitions \( P \) into a sequence of semi-positive Datalog\(^-\) programs \( P_1, \ldots, P_k \) with \( k = |\text{idb}(P)| \) such that for each \( i = 1, \ldots, k \), the program \( P_i \) contains the rules of \( P \) whose head predicate has stratum number \( i \). This sequence is called a syntactic stratification of \( P \). We can now apply the stratified semantics to \( P \): for an input \( I \) over \( \text{sch}(P) \), we first compute the fixpoint \( P_1(I) \), then the fixpoint \( P_2(P_1(I)) \), etc. The output of \( P \) on \( I \) denotes \( P(I) \), defined as \( P_k(P_{k-1}(\ldots(P_1(I))\ldots)) \).

It is well known that the output of \( P \) does not depend on the chosen syntactic stratification (if more than one exists). Not all Datalog\(^-\) programs are syntactically stratifiable. By stratified Datalog we refer to all Datalog\(^-\) programs which are syntactically stratifiable.

Since we only consider syntactically stratifiable programs in this paper, we denote from now on the class of stratified Datalog\(^-\) programs simply by Datalog\(^-\).

Computing Queries. For a query \( Q \) with input schema \( \sigma \) and output schema \( \sigma' \), and a stratifiable Datalog\(^-\) program \( P \), we say that \( P \) computes \( Q \) if \( Q(I) = P(I)|_{\sigma'} \) for all instances \( I \) over \( \sigma \).

Figure 1: Monotonicity hierarchy

We assume that for each Datalog\(^-\) program some idb-relations are marked as the intended output. In our example Datalog\(^-\) programs, we use the convention that relation ‘\( O \)’ denotes that output. The input relations are recognizable as the edb-relations.

In the sequel, we overload notation and denote both the fragment of Datalog\(^-\) programs as well as the queries expressed by programs in that class with the same notation. For instance, we use SP-Datalog to denote both the class of semi-positive Datalog\(^-\) programs as well as the queries which are expressible by a positive Datalog\(^-\) program.

In examples, we use a unary idb-relation \( \text{Adom} \) that contains the active domain of the input. This predicate is computed as the union of the projections of all positions of all edb-relations. We omit the rules to compute \( \text{Adom} \).

3. Weaker forms of monotonicity

We introduce in Section 3.1 two weaker forms of monotonicity that will be used in Section 4 to characterize classes of coordination-free transducers. In Section 3.2, we relate these notions with the well-known classes of queries preserved under homomorphisms and extensions.

3.1 Domain distinct & disjoint monotonicity

We say that a fact \( f \) is domain distinct from instance \( I \) when \( \text{dom}(f) \setminus \text{dom}(I) \neq \emptyset \) (i.e., \( f \) should contain a new domain element not occurring in \( I \)); \( f \) is domain disjoint when \( \text{dom}(f) \cap \text{dom}(I) = \emptyset \). Furthermore, an instance \( J \) is domain distinct (resp., domain disjoint) from \( I \), when every fact \( f \in J \) is domain distinct (resp., domain disjoint) from \( I \).

We introduce two weaker forms of monotonicity by restricting the set of instances for which the monotonicity condition should hold:

Definition 1. Let \( Q \) be a query. Then,

- \( Q \) is monotone if \( Q(I) \subseteq Q(I \cup J) \) for all database instances \( I \) and \( J \);
- \( Q \) is domain-distinct-monotone if \( Q(I) \subseteq Q(I \cup J) \) for all instances \( I \) and \( J \) for which \( J \) is domain distinct from \( I \); and,
- \( Q \) is domain-disjoint-monotone if \( Q(I) \subseteq Q(I \cup J) \) for all instances \( I \) and \( J \) for which \( J \) is domain disjoint from \( I \).

We denote the class of monotone, domain-distinct-monotone, and domain-disjoint–monotone queries by \( \mathcal{M} \), \( \mathcal{M}_{\text{distinct}} \), and
Theorem 3.1. For every $i, j \geq 1$ with $i < j$,
1. $M \subseteq M_{\text{distinct}} \subseteq M_{\text{disjoint}} \subseteq C$;
2. $M = M^i$;
3. $M_{\text{distinct}} \subseteq M_{\text{distinct}}^{i+1} \subseteq M_i$;
4. $M_{\text{disjoint}} \subseteq M_{\text{disjoint}}^{i+1} \subseteq M_i$;
5. $M_{\text{distinct}}^i \subseteq M_{\text{distinct}}^{i+1}$;
6. $M_i^{\text{disjoint}} \subseteq M_i$;
7. $M_i^{\text{disjoint}} \subseteq M_i$.

Proof. We focus on the inequalities and sketch the separating examples. All queries are over directed graphs that are defined over the binary edge relation $E$.

For (1), first, $M \subseteq M_{\text{distinct}}$ follows from SP-Datalog $\subseteq M_{\text{distinct}}$. Second, we show $Q_{\text{path}} \in M_{\text{disjoint}} \setminus M_{\text{distinct}}$; where $Q_{\text{path}}$ is the query that computes the complement of the transitive closure of the edge relation. To see $Q_{\text{path}} \in M_{\text{disjoint}}$, for all instances $I$ with a path missing from vertex $a$ to vertex $b$, i.e., $(a, b) \in Q_{\text{path}}$, the addition of domain-disjoint subgraphs will not create this path. To see $Q_{\text{path}} \notin M_{\text{distinct}}$, the addition of domain-distinct subgraphs can create a path $E(a, c), E(c, b)$, where $c$ is a new vertex. Third, for $M_{\text{disjoint}} \subseteq C$, take the query that outputs all triangles on condition that no two disjoint triangles exist.

For (2), by definition, $M_{\text{distinct}}^{i+1} \subseteq M_i$. We show that $M_i \subseteq M_{\text{distinct}}^{i+1}$. Let $Q \in M^i$ and let $I$ be arbitrary instances such that $|J| \leq i + 1$. Pick an arbitrary fact $f \in J$ and set $J' = J \setminus \{f\}$. By assumption, $Q(I) \subseteq Q(I \cup J')$ and $Q(I \cup J') \subseteq Q((I \cup J') \cup \{f\})$. Hence, $Q(I) \subseteq Q(I \cup J)$.

For (3), it suffices to show that $M_{\text{distinct}}^{i+1} \setminus M_{\text{distinct}}^{i+1} \neq \emptyset$. Ignoring the direction of edges, let $Q_{\text{clique}}^i$ be the query that outputs the edge relation when no clique of $k$ vertices exists and the empty relation otherwise. We show $Q_{\text{clique}}^{i+2} \in M_{\text{distinct}}^{i+1}$. Let $I$ be an input not containing $(i + 2)$-size cliques, upon which the output is thus nonempty. To expose the non-monotone behavior of $Q_{\text{clique}}^{i+2}$, we can try to extend any $(i + 1)$-size cliques in $I$ to $(i + 2)$-size cliques by adding a domain-distinct instance $J$. For this to work, $J$ needs to contain a star: one new value is the center and it points at old clique vertices of $I$, requiring $|J| \geq i + 1$. Such instances $J$ are not considered in the definition of $M_{\text{distinct}}^{i+1}$; but they are considered in the definition of $M_{\text{distinct}}^{i+1}$.

For (4), again, we only show that $M_{\text{disjoint}}^{i+1} \subseteq M_{\text{disjoint}}^{i+1}$.

Lemma 3.2. $H \subseteq H_{\text{adj}} = M \subseteq C = M_{\text{distinct}}$.

Proof. We only argue that $E = M_{\text{distinct}}$ as the rest is folklore (see, e.g., [6, 26, 29]). The equality follows immediately as $J$ is an induced subinstance of $I$ iff $I \setminus J$ is domain distinct from $J$.\qed
4. COORDINATION-FREENESS

A relational transducer is essentially a collection of queries that transforms a sequence of input facts to a sequence of output facts while maintaining a relational state $[5, 20, 21]$. In the distributed context, the functionality at each node of a network can be described via a relational transducer, giving rise to so-called relational transducer networks $[13]$. Subsequently, relational transducer networks have been extended in two ways to give nodes restricted access to distribution policies $[22]$. Such policies model deterministic input data distributions based on hashing. In this section, we review these two extensions, here referred to as policy-aware and domain-guided transducer networks, and characterize the corresponding classes of coordination-free queries. In particular, we show the following: the queries that are coordination-free in policy-aware transducer networks and domain-guided transducer networks correspond to the query classes $M_{\text{distinct}}$ and $M_{\text{disjoint}}$, respectively. It was shown in $[13]$ that coordination-free queries in the original transducer networks correspond precisely to the set $M$ of all monotone queries. The results in this section therefore provide a refinement of the CALM-conjecture in terms of weaker forms of monotonicity.

We formalize the transducer networks in Section 4.1 and present the results in Section 4.2. We provide some additional discussion in Section 4.3.

4.1 Transducer models

We review the two extensions of Zinn et al. $[32]$ to the original transducer network model $[13]$. Here, we refer to these extensions as policy-aware and domain-guided transducer networks, respectively.

4.1.1 Networks, data distribution, and policies

A network $N$ is a nonempty finite set of values from dom, which we call nodes. Let $\sigma$ be a database schema. A distributed database instance $H$ over $\sigma$ and $N$ is a function that maps each $x \in N$ to an instance over $\sigma$. A distributed database instance $H$ over $\sigma$ models a distribution with potential replication of data over the schema $\sigma$.

For a set $X$, let $P^+(X) = P(X) \setminus \{\emptyset\}$ denote the set of all non-empty subsets of $X$. We write facts$(\sigma)$ to denote the set of all possible facts over $\sigma$, i.e., using all possible values from dom. A distribution policy $P$ for $\sigma$ and $N$ is a total function from facts$(\sigma)$ to $P^+(N)$. Intuitively, $P$ says how to distribute any instance $I$ over $\sigma$ to the nodes of $N$, possibly with replication. Concretely, we define dist$\sigma(I)$ to be the distributed database instance $H$ over $\sigma$ and $N$ that satisfies $H(x) = \{f \in I \mid x \in P(f)\}$ for each $x \in N$.

A domain assignment $\alpha$ for $N$ is a total function from dom to $P^+(N)$. A distribution policy $P$ for $\sigma$ and $N$ is called domain-guided if there exists a domain assignment $\alpha$ for $N$ such that for each $R(a_1, \ldots, a_k) \in$ facts$(\sigma)$ we have $P(R(a_1, \ldots, a_k)) = \bigcup_{i=1}^k \alpha(a_i)$ $[3]$ Intuitively, for each value $a \in$ dom, function $\alpha$ says which nodes get input facts containing $a$.

Example 4.1. Suppose dom is the set $\mathbb{N}$ of natural numbers. Let $N = \{1, 2\}$ be a two-node network and let the schema $\sigma$ contain the single relation symbol $E$ of arity 2.

Consider the following distribution policy $P_1$ for $\sigma$ and $N$:

$$P_1(\{E(a, b)\}) = \begin{cases} \{1\} & \text{if } a \text{ is odd} \\ \{2\} & \text{otherwise} \end{cases}$$

for each $a, b \in \mathbb{N}$. Note that $P_1$ partitions any input $I$ over $\sigma$ based on its first attribute. If $I = \{E(1, 3), E(3, 4), E(4, 6)\}$, the distributed database instance dist$_1(I)$ is

$$\{1 \mapsto \{E(1, 3), E(3, 4)\}, 2 \mapsto \{E(4, 6)\}\}.$$

This input $I$ demonstrates that $P_1$ is not a domain-guided policy: neither node is assigned all facts containing domain value 4.

A domain-guided policy assigns domain values (rather than facts) to nodes in the network. Consider the domain assignment $\alpha$ for $N$ that maps odd numbers to $\{1\}$ and even numbers to $\{2\}$. The corresponding domain-guided distribution policy $P_2$ for $\sigma$ and $N$ allocates a fact $I$ to node 1 if any of its two values is odd; to node 2 if any of its two values is even. For the specific input $I$ from above, the instance dist$_2(I)$ is

$$\{1 \mapsto \{E(1, 3), E(3, 4)\}, 2 \mapsto \{E(4, 3), E(4, 6)\}\}.$$

Note that node identifiers can occur as data in relations. □

4.1.2 Policy-aware relational transducers

In the following, we write $R^{(k)}$ to denote a relation symbol $R$ of arity $k$. A (policy-aware) transducer schema $T$ is a tuple $(\text{ins}_N, \text{out}_N, \text{sys}_N, \text{msg}_N)$ of database schemas with disjoint relation names, with the additional restriction that

$$\text{sys}_N = \{\text{Id}^{(1)}, \text{All}^{(1)}, \text{MyAdom}^{(1)}\} \cup \{\text{policy}^{(k)} \mid R^{(k)} \in \text{ins}_N\}.$$

These schemas are called respectively “input”, “output”, “message”, “memory”, and “system”.

A (policy-aware relational) transducer $\Pi$ over $T$ is a quadruple $(Q_{\text{ins}}, Q_{\text{out}}, Q_{\text{del}}, Q_{\text{sys}})$ of queries having the input schema $\text{ins}_N \cup \text{out}_N \cup \text{sys}_N \cup \text{msg}_N$ and such that

- queries $Q_{\text{ins}}$ and $Q_{\text{del}}$ both have target schema $\text{sys}_N$;
- query $Q_{\text{sys}}$ has target schema $\text{sys}_N$.

These queries form the mechanism by which a node on a network produces output, updates its memory (through insertions and deletions), and sends messages. We note that the transducers in $[13]$ do not have the relations $\text{MyAdom}$ and $\text{policy}_P$ with $R$ in $\text{ins}_N$.

4.1.3 Policy-aware transducer networks

A (policy-aware) transducer network $\Pi$ is a quadruple $(N, T, \Pi, P)$ where $N$ is a network, $T$ is a transducer schema, $\Pi$ is a transducer over $T$, and $P$ is a distribution policy for $\text{ins}_N$ and $N$. We say that $\Pi$ is domain-guided if the distribution policy $P$ is domain-guided. Note that domain-guided transducer networks are a special kind of policy-aware transducer networks.

Intuition. We will next give the semantics of $\Pi$ on an input $I$ over $\text{ins}_N$. We start with the underlying intuition. Intuitively, we put a copy of transducer $\Pi$ on each node of $N$, and the policy $P$ tells us how to initialize each node with a fragment of $I$. Then we choose an arbitrary node $x \in N$ and make it “active”: we execute the queries of $\Pi$ on the
union of the local input facts at \( x \), the output and memory
facts stored at \( x \), any message facts (over \( \Upsilon_{\text{msg}} \)) received by \( x \), and also some facts over \( \Upsilon_{\text{sys}} \). The facts over \( \Upsilon_{\text{sys}} \) consist of the following: relation \( \text{Id} \) provides the identifier of \( x \) (i.e., just \( 'x' \)); relation \( \text{all} \) provides the identifiers of all nodes in the network; relation \( \text{MyAdom} \) provides for convenience the active domain that \( x \) knows about (either from its local facts or from received messages); and, the relations \( \text{policy}_R \) provide the set of facts over this active domain that \( x \) is responsible for (i.e., that are assigned to \( x \)) in policy \( P(i) \).

Intuitively, by considering only \( \text{policy}_R \)-facts over the active domain at \( x \), we provide “safe” access to the distribution policy, i.e., we prevent \( x \) from using values outside \( \text{dom}(I) \cup \mathcal{N} \). Next, the queries of \( \Pi \) update the output and memory at \( x \), and generate new messages that are sent to the other nodes. Such an active moment of a node is called a transition of \( \Pi \). The semantics of \( \Pi \) is described by so-called runs which are infinite sequences of transitions.

**Example 4.2.** Recall \( \text{dom} \), \( \mathcal{N} \), \( \sigma \), and \( P_1 \) from Example 4.1. Consider a policy-aware transducer network \( \Pi = (\mathcal{N}, \Upsilon, \Pi, P_1) \) with \( \Upsilon_{\text{in}} = \sigma \). We leave \( \Upsilon_{\text{out}}, \Upsilon_{\text{msg}}, \) and \( \Upsilon_{\text{mem}} \) unspecified. Let us focus on the node \( 1 \). On input \( I = \{E(1,3), E(3,4), E(4,6)\} \), at least the following facts will be exposed to node \( 1 \) during each transition: the local input facts \( E(1,3) \) and \( E(3,4) \); the system facts \( \text{Id}(1), \text{all}(1), \text{all}(2) \), \( \text{MyAdom}(a) \) for each \( a \in \{1,2,3,4\} \), and \( \text{policy}_k(a,b) \) with \( a \in \{1,3\} \) and \( b \in \{1,2,3,4\} \). If node \( 1 \) would later receive (and store) the value 6, then also \( \text{MyAdom}(6) \) will be exposed, and the \( \text{policy}_k(a,b) \)-facts with \( a \in \{1,3\} \) and \( b = 6 \). Also, note that node \( 1 \) can in principle deduce that \( E(3,2) \) is not part of \( I \) since \( \text{policy}_k(3,2) \) is present at node \( 1 \) but not in \( E(3,2) \).

**Formal semantics.** Let \( \Pi = (\mathcal{N}, \Upsilon, \Pi, P) \) be a policy-aware transducer network. A configuration of \( \Pi \) is a pair \( \rho = (s, b) \) of functions \( s \) and \( b \) such that:

- \( s \) maps each \( x \in \mathcal{N} \) to a set of facts over \( \Upsilon_{\text{out}} \cup \Upsilon_{\text{mem}} \);
- \( b \) maps each \( x \in \mathcal{N} \) to a multiset of facts over \( \Upsilon_{\text{msg}} \).

We call \( s \) and \( b \) respectively the state and (message) buffer. Intuitively, \( s \) specifies for each node what output and memory facts it has locally available, and \( b \) specifies for each node what messages have been sent to it and that are not yet delivered. The reason for having multisets for the buffers, is that the same message can be sent multiple times to the same recipient and thus multiple copies can be floating around in the network simultaneously. The start configuration of \( \Pi \) is the unique configuration \( \rho = (s, b) \) that satisfies \( s(x) = \emptyset \) and \( b(x) = \emptyset \) for each \( x \in \mathcal{N} \).

A transition of \( \Pi \) on an input \( I \) over \( \Upsilon_{\text{in}} \) is a quadruple \((\rho_1, x, m, \rho_2)\) where \( \rho_1 = (s_1, b_1) \) and \( \rho_2 = (s_2, b_2) \) are configurations of \( \Pi \), \( x \in \mathcal{N} \), \( m \) is a submultiset of \( b_1(x) \), and, letting

\[
H = \text{dist}_P(I),
\]
\[
M = m \text{ collapsed to a set,}
\]
\[
J = H(x) \cup s_1(x) \cup M,
\]
\[
A = \mathcal{N} \cup \bigcup_{r \in J} \text{dom}(f),
\]
\[
S = (\text{Id}(x)) \cup \{\text{all}(y) \mid y \in \mathcal{N}\} \cup \{\text{MyAdom}(a) \mid a \in A\} \cup \{\text{policy}_R(a_1,\ldots,a_k) \mid R(a) \in \Upsilon_{\text{in}}, \{a_1,\ldots,a_k\} \subseteq A, x \in P(R(a_1,\ldots,a_k))\},
\]
\[
D = J \cup S,
\]
for the state \( s_2 \) we have

\[
s_2(x)_{\Upsilon_{\text{out}}} = s_1(x)_{\Upsilon_{\text{out}}} \cup Q_{\text{out}}(D),
\]
\[
s_2(x)_{\Upsilon_{\text{mem}}} = [s_1(x)_{\Upsilon_{\text{mem}}} \cup (Q_{\text{ins}}(D) \setminus Q_{\text{del}}(D))] \backslash (Q_{\text{del}}(D) \setminus Q_{\text{ins}}(D)),
\]
\[
s_2(y) = s_1(y) \text{ for each } y \in \mathcal{N} \setminus \{x\}.
\]

We call \( \rho_1 \) and \( \rho_2 \) respectively the source and target configuration of the transition and we refer to \( x \) as the active node. If \( m = \emptyset \), then we call the transition a heartbeat.

For an input \( I \) over \( \Upsilon_{\text{in}} \), a run \( R \) of \( \Pi \) on \( I \) is an infinite sequence of transitions of \( \Pi \) on \( I \), such that the start configuration of \( \Pi \) is used as the source configuration of the first transition, and the target configuration of each transition is the source configuration of the next transition. Note that runs represent nondeterminism: each transition can choose what node becomes active and what messages to deliver from the buffer of that node. We consider only fair runs. These are the runs that satisfy the following additional conditions: (i) each node is the active node in an infinite number of transitions; and, (ii) if a fact occurs infinitely often in the message buffer of a node then this fact is infinitely often delivered to that node. Intuitively, the last condition demands that no sent messages are infinitely delayed.

### 4.1.4 Computing queries

We are interested in transducers that produce the same facts over \( \Upsilon_{\text{out}} \) regardless of the network, the distribution policy, and the order of transitions. These transducers are said to compute a query.

Formally, let \( Q \) be a query with input schema \( \sigma_1 \) and output schema \( \sigma_2 \). Further, let \( \Pi = (\mathcal{N}, \Upsilon, \Pi, P) \) be a policy-aware transducer network. We define the output of a run \( R \) of \( \Pi \), denoted \( \text{out}(R) \), to be the union of all output facts jointly produced by the nodes of \( \mathcal{N} \) during \( R \), i.e., all facts over \( \Upsilon_{\text{out}} \). Note that once a fact is added to \( \Upsilon_{\text{out}} \), it can never be retracted. We say that \( \Pi \) computes \( Q \) if \( i \) \( \Upsilon_{\text{in}} = \sigma_1 \) and \( \Upsilon_{\text{out}} = \sigma_2 \); and, (ii) for each input \( I \) over \( \sigma_1 \), every (fair) run \( R \) of \( \Pi \) on \( I \) satisfies \( \text{out}(R) = Q(I) \).

Now, we say that a policy-aware transducer \( \Pi \) over transducer schema \( \Upsilon \) (distributively) computes \( Q \) (for all policies) if for all networks \( \mathcal{N} \) and all distribution policies \( P \) for \( \Upsilon_{\text{in}} \) and \( \mathcal{N} \), the policy-aware transducer network \( (\mathcal{N}, \Upsilon, \Pi, P) \) computes \( Q \). We say that \( \Pi \) (distributively) computes \( Q \) un-
der domain-guidance if for all networks \(\mathcal{N}\) and all domain-guided policies \(P\) for \(\Upsilon\), the transducer network \((\mathcal{N}, \Upsilon, \Pi, P)\) computes \(Q\).

4.1.5 Coordination-freeness

We define coordination-freeness for policy-aware transducers similarly as for the original transducer model \[13\].

Definition 3. Let \(\Pi\) be a policy-aware transducer over a schema \(\Upsilon\). We say that \(\Pi\) is coordination-free if (1) \(\Pi\) distributedly computes a query \(Q\), and (2) if for all networks \(\mathcal{N}\), for all inputs \(I\) for \(Q\), there is a distribution policy \(P\) for \(\Upsilon\), such that the policy-aware transducer network \((\mathcal{N}, \Upsilon, \Pi, P)\) has a run on input \(I\) in which \(Q(I)\) is already computed in a prefix consisting of only heartbeat transitions.\[4\]

Similarly, we say that \(\Pi\) is coordination-free under domain-guidance if (1) \(\Pi\) distributedly computes a query \(Q\) under domain-guidance, and (2) if for all networks \(\mathcal{N}\), for all inputs \(I\) for \(Q\), there is a domain-guided policy \(P\) for \(\mathcal{N}\), such that the transducer network \((\mathcal{N}, \Upsilon, \Pi, P)\) has a run on input \(I\) in which \(Q(I)\) is already computed in a prefix consisting of only heartbeat transitions.

We write \(F_1\) to denote the set of queries distributively computed by coordination-free policy-aware transducers. We write \(F_2\) to denote the set of queries distributively computed by policy-aware transducers that are coordination-free under domain-guidance.

It is useful to reflect on what it means to be coordination-free. Of course, one could prohibit any form of communication but that would be too drastic and unworkable when data is distributed and communication is already needed for the simple purpose of exchanging data (for instance, to compute joins). Therefore, the present formalization tries to separate the ‘data’-communication (that can never be eliminated in a distributed setting) from the ‘coordination’-communication by requiring that there is some ‘ideal’ distribution on which the query can be computed without any communication.\[3\] The intuition is as follows: because on the ideal distribution there is no coordination (as communication is prohibited), and the transducer network has to correctly compute the query on all distributions, communication is only used to transfer data on non-ideal distributions and is not used to coordinate. While we do not claim our notion of coordination-freeness to be the only possible one, the results in Section 4.3 confirm that the just described intuition is not too far off. Indeed, it follows that coordination-freeness corresponds precisely to those computations that do not require the knowledge about all other nodes in the network, and hence, can not globally coordinate. Specifically, we show that when a node has no complete overview of all the nodes in the network, which can be achieved by removing the relation \(\text{A11}\), then every transducer is coordination-free. More importantly, the converse holds true as well. That is, we show that every coordination-free transducer is equivalent to one that does not use the relation \(\text{A11}\).

Theorem 4.3. \(F_1 = \mathcal{M}_{\text{distinct}}\).

Proof. We sketch the proof.

\(\mathcal{M}_{\text{distinct}}\) Let \(Q\) be a query from \(\sigma\) to \(\sigma'\) distributively computed by a coordination-free policy-aware transducer \(\Pi\). Let \(I\) and \(J\) be two input instances for \(Q\), such that \(J\) is domain-distinct from \(I\). Let \(f \in Q(I)\). We show \(f \in Q(I \cup J)\). Because \(\Pi\) distributedly computes \(Q\), transducer \(\Pi\) also computes \(Q\) on a network \(\mathcal{N}\) with at least two nodes. By coordination-freeness, there is a distribution policy \(P\) for \(\sigma\) and \(\mathcal{N}\) such that the transducer network \((\mathcal{N}, \Upsilon, \Pi, P)\) when given input \(I\), has a run \(R_1\) in which \(Q(I)\) is already computed in a prefix consisting of only heartbeat transitions. Let \(x \in \mathcal{N}\) be a node that outputs \(f\) in this prefix. We can make \(x\) output \(f\) on input \(I \cup J\), so that \(f \in Q(I \cup J)\).

To do this, fix an arbitrary node \(y \in \mathcal{N}\setminus\{x\}\). Consider the following distribution policy \(P\) for \(\sigma\) and \(\mathcal{N}\): \(P_1(g) = \{y\}\) for all \(g \in J\), and \(P_2(g) = P_1(g)\) for all \(g \in \text{facts}(\sigma)\setminus J\). Denote \(\Pi_1 = (\mathcal{N}, \Upsilon, \Pi, P_1)\). It can be verified that \(\Pi_1\) on input \(I \cup J\) gives to \(x\) the same local input facts as \(\Pi\) on input \(I\). So, when running \(\Pi_2\) on input \(I \cup J\), if we initially do only heartbeat with active node \(x\), the node \(x\) goes through the same state changes as in the heartbeat-prefix of \(R_1\). So after a while, after \(k\) heartbeats, node \(x\) outputs \(f\). This finite prefix can be extended to a full fair run \(R_2\) of \(\Pi_2\) on input \(I \cup J\), for which \(\text{out}(R_2) = Q(I \cup J)\) holds by assumption on \(\Pi\). Hence, \(f \in Q(I \cup J)\).

\(\mathcal{M}_{\text{distinct}} \subseteq F_1\) Let \(Q\) be an input for \(Q\) that is distributed over a network \(\mathcal{N}\). We say that a subset \(C \subseteq \text{adom}(I) \cup \mathcal{N}\) is complete at a node \(x \in \mathcal{N}\), when \(x\) knows for every fact \(f \in \text{facts}(\sigma)\) with \(\text{adom}(f) \subseteq C\) whether \(f \in I\) or \(f \notin I\). If \(C\) is indeed complete at \(x\), node \(x\) will output \(Q'(I')\) where \(I' = \{f \in I : \text{adom}(f) \subseteq C\}\). Note that \(Q'(I') \subseteq Q(I)\) by domain-distinct-monotonicity of \(Q\). We will construct a policy-aware transducer that postpones executing \(Q\) at a node \(x\) until the system relation \(\text{MyAdom}\) is complete at \(x\).

To implement this idea, the nodes broadcast their locally given input facts. Each node \(x\) stores every received input fact. This way, the system relation \(\text{MyAdom}\) grows at \(x\). During each transition, \(x\) checks for each input relation \(R^{(k)}\) and each \(k\)-tuple \((a_1, ..., a_k)\) over \(\text{MyAdom}\) whether the fact \(\text{policy}_{R}(a_1, ..., a_k)\) is shown to \(x\). If so, then \(x\) is responsible for the fact \(R(a_1, ..., a_k)\) under the distribution policy. In that case, if \(R(a_1, ..., a_k)\) is absent in the local input at \(x\), node \(x\) can conclude that \(R(a_1, ..., a_k)\) is actually globally absent from the entire input. Then \(x\) broadcasts the absence of this fact. These absences are also accumulated at all nodes. Now, consider a transition of node \(x\), and let \(I' \subseteq I\) denote the set of collected input facts at \(x\) so far.

For each potential input fact \(f\) over \(\text{MyAdom}\), node \(x\) checks whether \(f \in I'\) or that node \(x\) knows \(f \notin I\) (by means of the explicit absences). If this is so then \(\text{MyAdom}\) is complete at \(x\), and \(x\) subsequently computes \(Q\) on \(I'\).

We have already ensured that no wrong outputs are produced. To show that at least \(Q(I)\) is produced, we note that at some point, each node \(x\) has received all available input
facts and all absences of facts over \( \text{adom}(I) \cup \mathcal{N} \). At that moment, \( x \) computes \( Q \) on \( I \), causing at least \( Q(I) \) to be output in each run.

Transducer \( \Pi \) is indeed coordination-free according to the formal definition: for all networks \( \mathcal{N} \), and for all inputs \( I \), the full output will be computed at some node \( x \) with only heartbeats when \( x \) is made responsible (under the distribution policy) for all facts over \( \mathcal{F}_x \) made with the values \( \text{adom}(I) \cup \mathcal{N} \); then \( x \) will immediately detect that relation MyAdom is complete. \( \square \)

**Theorem 4.4.** \( \mathcal{F}_2 = \mathcal{M}_{\text{disjoint}} \).

**Proof.** We sketch the proof. The proof of \( \mathcal{F}_2 \subseteq \mathcal{M}_{\text{disjoint}} \) is similar to the proof of \( \mathcal{F}_1 \subseteq \mathcal{M}_{\text{distinct}} \).

\[
\mathcal{M}_{\text{disjoint}} \subseteq \mathcal{F}_2 \quad \text{Let } Q \in \mathcal{M}_{\text{disjoint}}. \quad \text{Let } I \text{ be an input for } Q \text{ that is distributed over a network } \mathcal{N}, \text{ by means of a domain-guided distribution policy. We call a subset } C \subseteq \text{adom}(I) \cup \mathcal{N} \text{ complete at a node } x \text{ when } x \text{ knows it has collected all facts } f \in I \text{ for which } \text{adom}(f) \cap C \neq \emptyset. \quad \text{If } C \text{ is indeed complete at } x, \text{ node } x \text{ will compute } Q(I') \text{ where } I' = \{ f \in I | \text{adom}(f) \cap C \neq \emptyset \}. \quad \text{Note that } Q(I') \subseteq Q(I) \text{ by domain-disjoint-monotonicity of } Q. \quad \text{We will construct a policy-aware transducer that postpones executing } Q \text{ at a node } x \text{ until the system relation MyAdom is complete at } x.

To implement this idea, the nodes broadcast the active domain of their local input fragment. These values are accumulated at each node. Note that when a node \( x \) has some value \( a \) in relation MyAdom, node \( x \) is responsible for \( a \) under the domain assignment if and only if \( \text{policy}_q(a, \ldots, a) \) is shown to \( x \) for at least one input relation \( R \). If \( x \) is indeed responsible for \( a \) then \( x \) is already locally given all facts of \( I \) containing \( a \) (because the distribution policy is domain-guided). Now, when \( x \) is not responsible for \( a \), node \( x \) will send out the request \((x, a)\). Any node \( y \) responsible for \( a \) under the domain assignment will then send all input facts containing \( a \) to \( x \). When \( x \) has acknowledged all these facts to \( y \), node \( y \) will send “OK(\( x, a \))”. Now, consider a transition of \( x \), and let \( I' \subseteq I \) denote the set of collected input facts at \( x \) so far. Node \( x \) checks for each value \( a \) in MyAdom whether \( x \) is responsible for \( a \) or that \( x \) has “OK” for \( a \). If this is so, the protocol above implies that \( x \) has obtained all input facts containing values from MyAdom, i.e., that MyAdom is complete at \( x \). Subsequently, \( x \) computes \( Q \) on \( I' \).

We have already argued that no wrong outputs are produced. To see that at least \( Q(I) \) is computed, we note that each node \( x \) eventually knows of the entire active domain (because the broadcasted domain will eventually arrive), and will thus at some point compute \( Q \) on the entire set of collected input facts.

The transducer is coordination-free according to the formal definition: for all networks \( \mathcal{N} \), and all inputs \( I \), the output will be computed with only heartbeats at some node \( x \) when \( x \) is made responsible (under the domain assignment) for all values \( \text{adom}(I) \cup \mathcal{N} \); then \( x \) will immediately detect that relation MyAdom is complete. \( \square \)

### 4.3 Discussion

We contrast the evaluation algorithms in the above proofs for \( \mathcal{M}_{\text{distinct}} \) and \( \mathcal{M}_{\text{disjoint}} \), with that for \( \mathcal{M} \). It is important to realize that the formulated algorithms are naive in the sense that the whole database is sent to all nodes and every node computes the result of the query. It is the type of monotonicity that determines when a node can start producing output:

- \( \mathcal{M} \): every node broadcasts all local input facts; output is generated for every newly received fact;
- \( \mathcal{M}_{\text{distinct}} \): every node broadcasts all local input facts as well as non-facts (missing input facts that the node is responsible for); output is generated for every complete (cf. proof of Theorem 4.3) subset of facts; and,
- \( \mathcal{M}_{\text{disjoint}} \): every node broadcasts the active domain of the local input facts; whenever a new domain element is received that the node is not responsible for, a coordination protocol is initiated with nodes responsible for this value; output is generated for every complete (cf. proof of Theorem 4.4) subset of facts.

While it might seem contradictory that the coordination-free evaluation of queries in \( \mathcal{M}_{\text{disjoint}} \) requires the use of a coordination protocol, it is important to realize that this coordination is only determined by the way data is distributed and does not require global coordination between all nodes. Indeed, we show below that transducers that do not access the system relation \( \mathcal{A} \), containing the names of all the nodes in the network, are automatically coordination-free. Moreover, the classes of queries that are distributively computed by policy-aware transducers without relation \( \mathcal{A} \) still capture \( \mathcal{M}_{\text{distinct}} \) and \( \mathcal{M}_{\text{disjoint}} \). That is, in absence of relation \( \mathcal{A} \), we do not need the notion of coordination-freeness to characterize \( \mathcal{M}_{\text{distinct}} \) and \( \mathcal{M}_{\text{disjoint}} \).

Formally, to prevent usage of relation \( \mathcal{A} \), at a node \( x \) we modify the semantics of transitions from Section 4.1.3 as follows: we define the set \( A \) now as \( A = \{ x \} \cup \bigcup_{f \notin x} \text{adom}(f) \); and, the set \( S \) is defined as before, but now without the \( \mathcal{A} \)-facts. For this resulting model, let \( A_1 \) denote the set of queries distributively computed by policy-aware transducers, and let \( A_2 \) denote the set of queries distributively computed by policy-aware transducers under domain-guidance.

**Theorem 4.5.** \( A_1 = \mathcal{M}_{\text{distinct}} \) and \( A_2 = \mathcal{M}_{\text{disjoint}} \)

**Proof.** First, the transducers constructed for the proofs of directions \( \mathcal{M}_{\text{distinct}} \subseteq \mathcal{F}_1 \) and \( \mathcal{M}_{\text{disjoint}} \subseteq \mathcal{F}_2 \) can be used unmodified to respectively show that \( \mathcal{M}_{\text{distinct}} \subseteq A_1 \) and \( \mathcal{M}_{\text{disjoint}} \subseteq A_2 \), because they do not use \( \mathcal{A} \).

We sketch the proof of \( A_1 \subseteq \mathcal{M}_{\text{distinct}} \). Let \( Q \) be a query that is distributedly computed by a policy-aware transducer \( \Pi \) without \( \mathcal{A} \). By assumption, transducer \( \Pi \) also computes \( Q \) on a single-node network \( \{ x \} \). So, on input \( I \), the single-node transducer network will produce \( Q(I) \) with just heartbeats (no messages can be sent). Let \( J \) be an input instance for \( Q \) that is domain-distinct from \( I \). To show \( Q(I) \subseteq Q(I \cup J) \), we consider a two-node network \( \{ x, y \} \) on which we run \( \Pi \), and consider the distribution policy that assigns \( J \) to just \( y \) and the other facts to \( x \). Then \( x \) is still given precisely \( I \) when the two-node transducer network is given \( I \cup J \) as input. Now, if we do only heartbeats at \( x \), then \( x \) will behave the same as on the single-node network because it can not detect the difference with the two-node network in absence of relation \( \mathcal{A} \). So, \( x \) will produce \( Q(I) \) after a finite number of heartbeats. This prefix can be extended to a full fair run of the two-node transducer network on input \( I \cup J \).

The inclusion \( A_2 \subseteq \mathcal{M}_{\text{disjoint}} \) can be shown similarly, except that now we assign all values of \( \text{adom}(J) \) to \( y \) and all other domain values to \( x \). \( \square \)
It is interesting to note what happens when transducers are not aware of the distribution policies, i.e., when we do not provide the policy_relations. In the resulting model, the set \( F_0 \) of queries distributedly computed by coordination-free transducers is precisely the set \( M \) of monotone queries, and relating to the above, the set \( A_0 \) of queries distributedly computed by transducers without relation \( A_1 \) is also \( M \).

For completeness, we also mention the existence of so-called oblivious transducers in the original transducer model of Ameloot et al. [13]: these transducers may use neither relation \( 1d \) nor relation \( A_1 \). The set of queries distributedly computed by oblivious transducers is again the set \( M \).

**Corollary 4.6.** \( F_0 = A_0 = M \).

## 5. DATALOG FRAGMENTS

### 5.1 Semi-connected Datalog

As mentioned in the introduction, the original formulation of the CALM-conjecture links coordination-free computation to Datalog. It is therefore interesting to investigate subclasses of Datalog with negation that remain within \( \mathcal{M}_{\text{distinct}} \) and \( \mathcal{M}_{\text{disjoint}} \). While SP-Datalog \( \subseteq \mathcal{M}_{\text{distinct}} \) [6], it is not known whether SP-Datalog can be further extended while remaining within \( \mathcal{E} \).

We next identify a fragment of Datalog which is domain-disjoint-monotone. Let \( \varphi \) be a Datalog rule. We define \( \text{graph}^+(\varphi) \) as the graph where nodes are the variables in positive body atoms of \( \varphi \), and there is an edge between two variables if they occur together in a positive body atom of \( \varphi \). We say \( \varphi \) is connected if \( \text{graph}^+(\varphi) \) is connected. We say that an SP-Datalog program is connected when all rules are connected.

**Definition 4.** Let \( P \) be a program in Datalog. Then \( P \) is a connected stratified datalog program, when there is a stratification for \( P \) such that all strata are connected SP-Datalog programs. We say that \( P \) is semi-connected when there is a stratification such that all strata except possibly the last one are connected SP-Datalog programs.

In the following, we denote the class of connected and semi-connected stratified Datalog programs by con-Datalog and semicon-Datalog, respectively. Then (i) SP-Datalog \( \subseteq \) semicon-Datalog, (ii) SP-Datalog \( \subset \) con-Datalog, and (iii) con-Datalog \( \subset \) semicon-Datalog.

**Example 5.1.** Consider the following Datalog program \( P_1 \):

\[
T(x) \leftarrow E(x, y), E(y, z), E(z, x), y \neq x, y \neq z, z \neq x \\
O(x) \leftarrow \neg T(x), \text{Adom}(x)
\]

Then, \( P_1 \) is in con-Datalog, but \( P_1 \notin \mathcal{M}_{\text{distinct}} \). Indeed, \( P_1(\{E(a, b)\}) \neq \emptyset \), while \( P_1(\{E(a, b)\} \cup \{E(b, c), E(c, a)\}) = \emptyset \). Therefore, \( P_1 \notin \text{SP-Datalog} \).

Consider the program \( P_2 \) which is not a semicon-Datalog program:

\[
T(x, y, z) \leftarrow E(x, y), E(y, z), E(z, x), y \neq x, y \neq z, z \neq x, z \neq x \\
D(x_1) \leftarrow T(x_1, x_2, x_3), T(y_1, y_2, y_3), \bigwedge_{1 \leq i,j \leq 3} x_i \neq y_j \\
O(x) \leftarrow \neg D(x), \text{Adom}(x)
\]

**Note:** the specified query is not in \( \mathcal{M}_{\text{disjoint}} \). As we will see shortly, this implies that the query itself cannot be defined by any semicon-Datalog program.

We next relate connected programs to distributed evaluation over components. An instance \( I \) is a component of an instance \( J \) when \( J \subseteq I \), \( J \neq \emptyset \), \( \text{adom}(J) \cap \text{adom}(I \setminus J) = \emptyset \) and \( J \) is minimal with this property. That is, there is no strict nonempty subset \( J' \) of \( J \) for which \( \text{adom}(J') \cap \text{adom}(I \setminus J') = \emptyset \). Intuitively, a component is complete w.r.t. its active domain as it either contains all the facts in \( I \) in which a certain domain element occurs, or it contains none of these facts.

Denote by \( \text{co}(I) \) the components of \( I \).

**Definition 5.** A query \( Q \) distributes over components if for all instances \( I \): (1) \( Q(I) = \bigcup_{C \in \text{co}(I)} Q(C) \); and, (2) \( \text{adom}(Q(C)) \cap \text{adom}(Q(C')) = \emptyset \) for all \( C, C' \in \text{co}(I) \) with \( C \neq C' \).

The lemma below highlights an important property of con-Datalog and forms a crucial part of the proof of Theorem 5.3.

**Lemma 5.2.** Every query in con-Datalog distributes over components.

We are now armed to prove the following theorem:

**Theorem 5.3.** semicon-Datalog \( \subseteq \mathcal{M}_{\text{disjoint}} \)

**Proof.** Let \( P \) be a semicon-Datalog program with stratification \( P_1, \ldots, P_s \) where \( P_s \) is a connected SP-Datalog program for \( i < s \). Define \( P_{\leq i} \) as the composition \( P_i \circ \cdots \circ P_1 \) of the first \( i \) strata. Clearly, \( P = P_{\leq s} = P_s \circ P_{\leq s-1} \).

Let \( I \) and \( J \) be two disjoint instances. Then by Lemma 5.2, \( P_{\leq s-1}(I) \) is domain disjoint from \( P_{\leq s-1}(J) \). By domain-disjoint-monotonicity of SP-Datalog programs, it follows that

\[
P_s(I) \subseteq P_{\leq s-1}(I) \cup P_{\leq s-1}(J) \quad \text{(1)}
\]

We next show that

\[
P_{\leq s-1}(I) \cup P_{\leq s-1}(J) = P_{\leq s-1}(I \cup J) \quad \text{(2)}
\]

which together with (1) implies that \( P(I) \subseteq P(I \cup J) \) and therefore \( P \in \mathcal{M}_{\text{disjoint}} \).

It remains to prove (2). Lemma 5.2 implies that \( P_{\leq s-1}(I) = \bigcup_{C \in \text{co}(I)} P_{\leq s-1}(C) \), \( P_{\leq s-1}(J) = \bigcup_{D \in \text{co}(J)} P_{\leq s-1}(D) \) and \( P_{\leq s-1}(I \cup J) = \bigcup_{K \in \text{co}(I \cup J)} P_{\leq s-1}(K) \). Since \( I \) and \( J \) are domain disjoint, \( \text{co}(I) \cup \text{co}(J) = \text{co}(I \cup J) \). Therefore,

\[
P_{\leq s-1}(I) \cup P_{\leq s-1}(J) \\
= \bigcup_{C \in \text{co}(I)} P_{\leq s-1}(C) \cup \bigcup_{D \in \text{co}(J)} P_{\leq s-1}(D) \\
= \bigcup_{K \in \text{co}(I \cup J)} P_{\leq s-1}(K) \\
= P_{\leq s-1}(I \cup J).
\]

This completes the proof.

### 5.2 Datalog with value invention

The classes \( \mathcal{M} \) and \( \mathcal{E} \) (and therefore \( \mathcal{M}_{\text{distinct}} \)) are captured by fragments of ILOG, a declarative formalism in the style of stratified Datalog originally introduced in the context of object databases by Hull and Yoshikawa [21].
We follow Cabibbo [18] for the definition of ILOG\(^-\) and assume the existence of invention relations. These are relations with a distinguished first position, called the invention position or the invention attribute. An invention atom is an atom of the form \(R(\ast, u_1, \ldots, u_k)\), where \(R\) is an invention relation and each \(u_i \in \text{var}\). The symbol \(\ast\) is called the invention symbol. An ILOG\(^-\) program \(P\) over \(\sigma\) is a Datalog program with negation over \(\sigma\) where we allow head atoms to be either (ordinary) relation atoms or invention atoms.

Before defining the semantics of ILOG\(^-\), we associate to each invention relation \(R\) a distinct Skolem functor \(f_R\) of arity \(ar(R) - 1\). The Skolemization of \(P\), denoted by \(\text{Skol}(P)\), is the set of rules in \(P\) where every occurrence of the invention symbol is replaced by appropriate Skolem functor terms. For example, the Skolemization of the rule

\[ R(\ast, x_1, x_2) \leftarrow E(x_1, x_2) \]

is

\[ R(f_R(x_1, x_2), x_1, x_2) \leftarrow E(x_1, x_2). \]

The semantics of ILOG\(^-\) is defined similar to the semantics of Datalog with negation, but valuations are applied on the Skolemized rules instead of the rules itself and are w.r.t. the Herbrand universe \(\mathcal{H}_P\) for \(P\) which is the set of all ground terms built using elements from \(\text{dom}\) and Skolem functors for invention relations in \(P\). When the repeated application of the immediate consequence operator eventually gives rise to relations of infinite size, the output of the program is undefined, otherwise the output consists of the facts in the output relations.

An ILOG\(^-\) program \(P\) is safe when the output contains no invented values. Although safety is an undecidable property, there are syntactical restrictions that ensure safe programs. One of these is called weakly safe which we define next. We consider the set of unsafe positions in atoms of \(P\). This is the smallest set \(S\) containing pairs of the form \((R, i)\), where \(R\) is a relation name and \(i \leq ar(R)\), such that

- \((R, 1) \in S\) for every invention relation \(R\); and,
- if \((R, i) \in S\) and there is a rule \(\varphi\) in \(P\) such that \(R(x_1, \ldots, x_k) \in \text{pos}_{\varphi}\), \(T(y_1, \ldots, y_l) = \text{head}_{\varphi}\), and \(x_i\) and \(y_j\) refer to the same variable, then \((T, j) \in S\).

An ILOG\(^-\) program \(P\) is weakly safe when the output relations of \(P\) do not contain unsafe positions. Note that a weakly safe program is always safe. We denote weakly safe ILOG\(^-\) by wILOG\(^-\). Furthermore, wILOG\(^-\) where negation is restricted to predicates in \(\text{edb}(P)\) is denoted by SP-wILOG for semi-positive wILOG\(^-\). Stratification can be defined for wILOG\(^-\) in the same way as for Datalog\(^-\). As before, we only consider stratified negation in this paper.

Cabibbo investigated the expressiveness of ILOG\(^-\) over relational databases [18] and obtained the following results: stratified wILOG\(^-\) programs with two strata capture the class of all computable queries; wILOG\(^-\) programs where negation is restricted to inequalities and extensional database predicates captures \(M\) and \(E\), respectively.

We introduce the class of semi-connected wILOG\(^-\) programs and show that it captures precisely the domain-disjoint-monotone queries. Analogous to the definitions for Datalog\(^-\), we say that a SP-wILOG program is connected when all rules are connected. A wILOG\(^-\) program \(P\) is semi-connected when there is a stratification for \(P\) such that all strata, except possibly the last one, are connected SP-wILOG programs. Due to space restrictions, the proof of the next result is omitted.

**Theorem 5.4.** Semi-connected wILOG\(^-\) computes precisely all queries in \(M_{\text{disjoint}}\).

### 6. RELATED WORK

**Declarative networking.** Hellerstein [23] argues that the theory of declarative database query languages can provide a foundation for the next generation of parallel and distributed programming languages and formulates a number of related conjectures for the PODS community. Datalog has previously been considered for distributed systems, see e.g. [25], [18].

Ameloot et al. [13] have introduced the framework of relational transducer networks to formalize and prove the CALM-conjecture. The formalism was then parameterized by Zinn et al. [32] to allow for specific data distribution strategies. These authors showed that the classes of coordination-free queries in the original transducer network model \((\mathcal{F}_1)\), in the policy-aware transducer network model \((\mathcal{F}_2)\), and in the domain-guided transducer network model \((\mathcal{F}_2)\) form a strict hierarchy. In particular, they showed that the non-monotone win-move query is in \(\mathcal{F}_2\). Some further work on relational transducer networks was done by Ameloot and Van den Bussche who studied decidability of confluence [14] and consistency [12]. The CRON-conjecture, which relates the causal delivery of messages to the nature of the computations that those messages participate in (like monotone versus non-monotone) is treated by Ameloot and Van den Bussche [15].

Zinn introduced in [31] the idea of domain-distinct- and domain-disjoint-monotone queries (albeit under a different name). He defined these as \(M^1\) and \(M^1_{\text{disjoint}}\), respectively, and used them to show the corresponding version of Theorem 4.3 and Theorem 4.4. Although the proof and statement of the results are incorrect, the proposed approach did already contain the ideas on which the proofs presented in this paper are based. In fact, it was the enthusiasm of the first three authors over the approach in [31] that in a collaborative effort with the fourth author led to the present paper. In this way, Section 4.3 of the present paper can be seen as an extension and correction of [31]. The results in Section 5, Section 4.3, and Section 5 are not discussed in [31].

The networked relational transducer model is just one paradigm for studying distributed query evaluation. Another notable model (or language) is WebDanLog [3], which is a Datalog-variant for declarative networking. This language supports delegation, where a node can send rules to another node instead of just facts; transmitted rules can then be locally evaluated at the addressee. In general, the distributed computations expressed by WebDanLog do not assume a fixed set of nodes, but they allow previously unseen nodes to participate.

Another model, called the massively parallel (MP) model, is introduced by Koutris and Suciu [27]. Here, computation proceeds in a sequence of parallel steps, each followed by global synchronization of all servers. In this model, evaluation of conjunctive queries [27] as well as skyline queries [8] have been considered.
Datalog\(\neq\) \(\subset\) wILOG\(\neq\) \[\subseteq\]
SP-Datalog \(\subseteq\) SP-wILOG \[\subseteq\]
semicon-Datalog\(^-\) \(\subseteq\) semicon-wILOG\(^-\) \[\subseteq\]

\[\begin{array}{cccc}
\text{Fig. 2: Main results of this work: query classes introduced and relationships obtained in this paper are shown in bold-face. Non-bold-face relationships without annotation are part of database folklore.}
\end{array}\]

Related to the MP model, is the MapReduce framework where the evaluation of queries has been considered by a number of researchers. Afrati and Ullman \[9\] study the evaluation of join queries and take the amount of communication, calculated as the sum of the sizes of the input to reducers, as a complexity measure. Evaluation of transitive closure and datalog queries in MapReduce has been investigated in \[7, 10\].

As mentioned before, Afrati et al. \[6\] obtained that SP-Datalog \(\subseteq E\). Feder and Vardi \[22\] showed that all queries in SP-Datalog that are preserved under homomorphisms can already be expressed in Datalog. That is, SP-Datalog \(\cap H \subseteq \text{Datalog}\). Dawar and Kreutzer \[19\] showed that the latter result can not be extended to least fixed-point logic (LFP): LFP \(\cap H \not\subseteq \text{Datalog}\). The status of SP-Datalog w.r.t. \(E\) is less clear. It is, for instance, not known whether SP-Datalog = \(E \cap \text{Datalog}\)\(^-\). A related result here is the one by Rosen \[29\] who showed that FO[\(\exists^+ \forall^+\)] \(\cap \mathcal{E} \subseteq \text{SP-Datalog}\), where FO[\(\exists^+ \forall^+\)] denotes first-order logic formulas of the form \(\exists^+ \forall^+ \psi\) where \(\psi\) is quantifier-free. We note that Tait’s counterexample \[30\] separating FO[\(\exists\)] from \(E \cap \text{FO}\) is definable in SP-Datalog \[24\]. Atheros et al. \[16\] study \(E\) in relation to FO over restricted classes of structures.

7. CONCLUSION

Figure 2 summarizes the main findings of this paper. At the same time the figure formulates a more fine-grained answer to the CALM-conjecture which stipulates that a program has a coordination-free execution strategy if and only if the program is monotone. In particular, our results equate increasingly larger classes of coordination-free computations with increasingly weaker forms of monotonicity and make Datalog variants explicit for each of these classes. Furthermore, the last two columns, as already explained in Section 4.1, and Section 4.3 confirm that the notion of coordination-freeness as proposed in \[13\] is a sensible one. Indeed, the notion corresponds to the intended semantics in that coordination-freeness avoids global synchronization barriers through the absence of knowledge about all the nodes in the network. That said, we do not claim that our notion is the only possible one. Indeed, one could argue that, especially within \(F_1\) and \(F_2\), even though there is no global synchronization barrier, computing nodes are still prone to wait until complete subsets of the input data have been accumulated (cf. Section 1.2). Of course, the semantical characterizations in terms of weaker forms of monotonicity make precise that this waiting is determined by the way data is distributed. The query evaluation algorithms in the proofs in Section 4.2 are inefficient in that they require all data to be sent to all nodes, it remains to investigate how the insights obtained in this paper can lead to more practical algorithms.

Another contribution of this paper is the identification of (semi-)connected Datalog variants which to the best of our knowledge have not been considered before. We can prove that the connected variant of Datalog under the well-founded semantics, making use of the well-known “doubled program” approach, remains within \(M_{\text{disjoint}}\) which implies a simpler proof of the fact that win-move is in \(M_{\text{disjoint}}\) (one of the main results in \[32\]). It is open whether semi-connected Datalog under the well-founded semantics is in \(M_{\text{disjoint}}\). A direction for future work is to study the properties of connected Datalog.

As is to be expected, deciding whether a query class belongs to one of the monotonicity classes quickly turns undecidable. Still, it would be interesting to find decidable subclasses or identify sufficient conditions as this would provide insight on the way queries can be distributedly computed. In Section 3 we introduced the bounded classes \(M_{\text{distinct}}\) and \(M_{\text{disjoint}}\) mainly because of the mismatch with \[31\] as explained in Section 4. It remains to investigate their relationship with distributed evaluation of queries.

We avoided the use of nullary relations. This is not a fundamental restriction but a practical one. With general policies, all results in this paper carry over to programs with schemas that can derive nullary relations unaltered. For domain-guided policies, definitions need to be slightly adapted: a nullary fact is never domain disjoint from any relation. Furthermore, in a domain-guided distribution policy, all nullary facts always have to be assigned to all computing nodes, and the definition of a component has to be extended to include all nullary facts. We will provide all details in the journal version of this paper.

Acknowledgments.

We thank Georg Gottlob, Thomas Eiter, and Phokion Kolaitis for answering our questions on Datalog. We also thank Phokion Kolaitis for bringing \[20\] to our attention.
8. REFERENCES


