On Computing the Measures of First-Order Definable Sets of Trees

Marcin Przybyłko*
University of New Caledonia
University of Warsaw
M.Przybylko@mimuw.edu.pl

We consider the problem of computing the measure of a regular language of infinite binary trees. While the general case remains unsolved, we show that the measure of a language defined by a first-order formula with no descendant relation or by a Boolean combination of conjunctive queries (with descendant relation) is rational and computable. Additionally, we provide an example of a first-order formula that uses descendant relation and defines a language of infinite trees having an irrational measure.

1 Introduction

The problem of computing a measure of a set can be seen as one of the fundamental problems considered in the study of probabilistic systems. This problem has been studied mostly implicitly, as it is often one of the intermediary steps in solving stochastic games, cf. [4], answering queries in probabilistic databases, cf. [16], or model checking for stochastic branching processes, cf. [5].

To us, this problem naturally arises in the study of stochastic games. Many of the games considered in literature can be seen as instances of the stochastic version of Gale-Stewart games [6]. Such games have winning conditions expressed as sets of winning plays, i.e. a set of (in)finite words representing winning plays. Hence, computing the value of a stochastic game involves computing the measure of the winning set with respect to the probabilistic space generated by the stochastic elements of the game.

Mio [11] introduced branching games, i.e. stochastic games for which the plays are represented as (in)finite trees, rather than words. Then, the questions concerning whether a set of trees has a measure [7], and whether that measure can be computed [10] have been raised and partially answered.

In this work, we focus our attention on the problem of computing the uniform measure of a set of labelled infinite trees defined by a first-order formula using child and descendant relations.

Related work The problem of computing the measure of an arbitrary regular language of trees has been already, explicitly or implicitly, studied. Gogacz et al. [7] prove that regular languages of trees are universally measurable. In the case of infinite trees, Chen et al. [5] show that the measure of a language accepted by deterministic automaton is computable; Michalewski and Mio [10] extend the class of languages with computable measure to the class of languages defined by game automata. In the case of finite trees, Amarilli et al. [1] show that with measures defined by fragments of the probabilistic XML, i.e. where the support of the measure consists of the trees of bounded depth, the measure is computable for arbitrary regular sets of trees. In the case of regular languages of infinite words, Staiger [15] shows that

*The author has been supported by Poland's National Science Centre grant no. 2016/21/D/ST6/00491.

A. Orlandini, M. Zimmermann (Eds.): 9th Symposium on
Games, Automata, Logics and Formal Verification (GandALF'18)
EPTCS 277, 2018, pp. 206–219, doi:10.4204/EPTCS.277.15

the measure of an arbitrary regular language of words is computable. Note that, in all the above results the inherent deterministic nature of the involved automata plays an important role.

The problem of computing the measure of tree a language has been, implicitly, considered in probability games. The problem is a special case of computing the value of a probability game when the strategies of players are already chosen. In the case of infinite trees, Przybyłko and Skrzypczak [14] consider branching games with regular wining sets. In the case of words, for the survey of probabilistic ω-regular games on graphs see e.g. Chatterjee and Henzinger [4].

The problem of computing the measure can be also seen as the problem of query evaluation on probabilistic databases, e.g. Amarilli et al. [2] enquire into the evaluation problem of conjunctive queries over probabilistic graphs. For an introduction to probabilistic databases see e.g. [16].

Our contribution We provide algorithms to compute the measures of tree languages definable by some restricted classes of first-order formulae. We show that, in the case of first-order formulae using unary predicates and child relation, the standard measure can be computed in three-fold exponential time. Moreover, in the case of Boolean combinations of conjunctive queries using unary predicates, child relation, and descendant relation, the measure can be computed in exponential space.

We also provide an example of a first-order formula over a two letter alphabet for which the defined language has an irrational standard measure. An example of a regular language with an irrational standard measure was already presented in [10], however that language is not first-order definable.

Organization of the paper In Section 2 we define basic notions used in this article. In Section 3 we showcase the properties of the standard measure on selected examples. The computability of the measure of the regular languages defined by first-order formulae is discussed in Section 4. The computability of the measure of the regular languages defined by conjunctive queries is discussed in Section 5. In the last section, we discuss obtained results and propose some directions of future research.

2 Preliminaries

In this section we present crucial definitions used throughout this work. We assume basic knowledge of logic, automata, and measure. For introduction to logic and automata see [17], for introduction to topology and measure see [9].

Words and trees An alphabet is any non-empty finite set. A *word* is a partial function $w\colon \mathbb{N} \rightharpoonup \Gamma$, such that $\mathbb{N}$ is the set of natural numbers, Γ is an alphabet, and the domain $Dom(w)$ of w is $\leq$-closed. By $|w|$ we denote the length of the word w, i.e. the size of its domain. By ε we denote the empty word, i.e. the unique word of length 0. If the domain of a word w is finite, then the word is called finite; otherwise, it is called infinite. The set of all finite words over an alphabet Γ is denoted Γ^*, the set of all infinite words over an alphabet Γ is denoted Γ^ω. Let $n \in \mathbb{N}$ and $\bowtie \in \{<, \leq, =, \geq, >\}$, then the set of all words over an alphabet Γ of length l such that $l \bowtie n$ is denoted $\Gamma^{\bowtie n}$.

A word w is called a prefix of a word v, denoted $w \sqsubseteq v$, if $Dom(w) \subseteq Dom(v)$ and for every $i \in Dom(w)$ we have that $w(i) = v(i)$. By $w \cdot v$, or simply wv, we denote the concatenation of the words w and v.

A tree is any partial function $t\colon \{\mathtt{L}, \mathtt{R}\}^* \rightharpoonup \Gamma$, where the domain $Dom(t)$ is prefix-closed and Γ is a finite alphabet. The elements of the set $\{\mathtt{L}, \mathtt{R}\}$ are called directions (left and right, respectively) and the elements of the set $\{\mathtt{L}, \mathtt{R}\}^*$ are called positions. For a given tree t, the elements of the set $Dom(t)$ are called nodes of the tree t, or nodes for short. A tree t is either finite, if its domain $Dom(t)$ is finite, or infinite. The

tree t is called a full binary tree of height k if $Dom(t) = \{\mathtt{L},\mathtt{R}\}^{\leq k}$. A tree t is called a full binary tree if $Dom(t) = \{\mathtt{L},\mathtt{R}\}^*$. Let Γ be an alphabet. The set of all trees over an alphabet Γ is denoted by $\mathcal{T}_\Gamma$; the set of all finite trees by $\mathcal{T}_\Gamma^F$; the set of all full binary trees of height k by $\mathcal{T}_\Gamma^{=k}$; the set of all full binary trees by $\mathcal{T}_\Gamma^\omega$. A tree t_1 is called a prefix of a tree t_2, denoted $t_1 \sqsubseteq t_2$, if $Dom(t_1) \subseteq Dom(t_2)$ and for every $u \in Dom(t_1)$ we have that $t_1(u) = t_2(u)$. We say that a tree t_2 is a sub-tree of t_1 in node $u \in Dom(t_1)$ if $u \cdot Dom(t_2) \subseteq Dom(t_1)$ and for every $v \in Dom(t_2)$ we have that $t_1(uv) \subseteq t_2(v)$. Let t_1 be a tree, by $t_1.u$ we denote the $\sqsubseteq$-biggest sub-tree of t_1 in the node $u \in Dom(t_1)$, i.e. biggest with the respect to the containment of the domains. For a tree t and a position u, by $\mathbb{B}_{t,u}$ we denote the set of full binary trees in which t is a sub-tree in node u, i.e. the set

$$\mathbb{B}_{t,u} \stackrel{\text{def}}{=} \{t' \in \mathcal{T}_\Gamma^\omega \mid t \sqsubseteq t'.u\}, \tag{1}$$

with $\mathbb{B}_t \stackrel{\text{def}}{=} \mathbb{B}_{t,\varepsilon}$.

Logic A tree t over an alphabet Γ can be seen as a relational structure $t = \langle Dom(t), s_\mathtt{L}, s_\mathtt{R}, s, \lessdot, (a^t)_{a\in\Gamma}\rangle$, where

- $Dom(t)$ is the domain of t;
- $s_\mathtt{L}, s_\mathtt{R} \subseteq Dom(t) \times Dom(t)$ are the left child relation ($u\ s_\mathtt{L}\ u\cdot\mathtt{L}$) and right child relation ($u\ s_\mathtt{R}\ u\cdot\mathtt{R}$), respectively;
- s is the child relation $s_\mathtt{L} \cup s_\mathtt{R}$;
- $\lessdot$ is the ancestor relation, i.e. the transitive closure of the relation s;
- $a^t \subseteq Dom(t)$ is a subset of $Dom(t)$, for $a \in \Gamma$, and the family of sets $(a^t)_{a\in\Gamma}$ is a partition of $Dom(t)$.

The partition $(a^t)_{a\in\Gamma}$ induces a labelling function $\lambda_t : Dom(t) \to \Gamma$ in the natural way, i.e. $\lambda_t(u) = a$ if and only if $u \in a^t$.

Regular languages Formulae of Monadic Second-Order logic (MSO) can quantify over positions in trees $\exists x, \forall x$ and over sets of positions $\exists X, \forall X$. A First-Order (FO) formula is an MSO formula that does not quantify over the sets of positions. A sentence is a formula with no free variables.

We say that an MSO formula φ is over a signature Σ if φ is a well-formed formula built from the symbols in Σ together with the quantifiers and logical connectives. Let Γ be an alphabet, in this paper, we consider only the formulae over the signatures Σ such that $\Sigma \subseteq \{s_\mathtt{L}, s_\mathtt{R}, s, \lessdot, \varepsilon\} \cup \Gamma$.

Let φ be a first-order formula. We write $t, v \vDash \varphi(x_1, \ldots, x_k)$, if the tree t, as a logical structure, with the valuation $v \in Dom(t)^k$ satisfies the formula $\varphi(x_1, \ldots, x_k)$. If φ is a sentence, we simply write $t \vDash \varphi$. We say that a formula $\varphi(x_1, \ldots, x_k)$ is satisfiable if there is a tree t and a tuple $v \in Dom(t)^k$ such that $t, v \vDash \varphi(x)$.

Let Γ be an alphabet, the set defined by an MSO sentence φ, denoted $\mathrm{L}(\varphi)$, is the set of all full binary trees over the alphabet Γ that satisfy φ, i.e. $\mathrm{L}(\varphi) \stackrel{\text{def}}{=} \{t \in \mathcal{T}_\Gamma^\omega \mid t \vDash \varphi\}$. A language defined by an MSO formula is called regular. This definition of regular languages of trees is equivalent to the automata based definition, cf. e.g. [17].

Measure The set of all full binary trees over an alphabet Γ, denoted $\mathcal{T}_\Gamma^\omega$, is the set of all functions $t\colon\{\mathtt{L},\mathtt{R}\}^* \to \Gamma$. This set can naturally be enhanced with a topology in such a way that it becomes a homeomorphic copy of the Cantor set, see Gogacz et al.[7] for more detailed definitions. The *uniform* (or *standard*) measure μ^* defined on the set of trees $\mathcal{T}_\Gamma^\omega$ is the probability measure such that for every finite tree $t \in \mathcal{T}_\Gamma^\omega$ we have that $\mu^*(\mathbb{B}_t) = |\Gamma|^{-|Dom(t)|}$. In other words, this measure is such that for every node $u \in \{\mathtt{L},\mathtt{R}\}^*$ and label $a \in \Gamma$ the probability that in a random full binary tree t the node u is labelled with the letter a is $\frac{1}{|\Gamma|}$, i.e. $\mu^*(\{t \in \mathcal{T}_\Gamma^\omega \mid t(u) = a\}) = \frac{1}{|\Gamma|}$.

By the following theorem we conclude that every regular language of trees L has well defined standard measure $\mu^*(L)$.

Theorem 2.1 ([7])**.** *Every regular language L of infinite trees is universally measurable, i.e. for every Borel measure μ on the set of trees, we know that L is μ-measurable.*

Hence, the following problem is well-defined.

Problem 2.2 (The $\mu^*(\mathrm{MSO})$ problem)**.** *Is there an algorithm that given an MSO formula φ computes $\mu^*(\varphi)$?*

With the $\mu^*(\mathrm{MSO})$ problem we associate the following decision problem.

Problem 2.3 (The positive $\mu^*(\mathrm{MSO})$ problem)**.** *Given an MSO formula φ, decide whether $\mu^*(\varphi) > 0$.*

If $\mathcal{C}$ is a class of MSO formulae, then by *the (positive) $\mu^*(\mathcal{C})$ problem*, we understand the above where possible input formulae are restricted to the class $\mathcal{C}$. If we restrict the class $\mathcal{C}$ to formulae over the signature Σ we denote it by $\mathcal{C}(\Sigma)$.

The problem, in this form, was stated by Michalewski and Mio [10]. It is open in the general case, but some partial results have been obtained, see paragraph *Related work* for details.

3 Measures of simple languages

To better understand the properties of the standard measure, we start with some examples of sets of infinite trees and their measures. The examples will be used in the proofs in the following sections.

Lemma 3.1. *Let t be a binary tree over the alphabet Γ, $u \in \{\mathtt{L},\mathtt{R}\}^*$ a position.*

1. *If t is finite and $L_{fp} = \mathbb{B}_{t,u} = \{t' \in \mathcal{T}_\Gamma^\omega \mid t \sqsubseteq t'.u\}$, then $\mu^*(L_{fp}) = \Gamma^{-|Dom(t)|}$.*
2. *If t is finite and $L_{fs} \overset{def}{=} \{t' \in \mathcal{T}_\Gamma^\omega \mid \exists v.(u \sqsubseteq v) \wedge (t \sqsubseteq t'.v)\}$, then $\mu^*(L_{fs}) = 1$.*
3. *If t is infinite and $L_{ip} = \mathbb{B}_{t,u} = \{t' \in \mathcal{T}_\Gamma^\omega \mid t \sqsubseteq t'.u\}$, then $\mu^*(L_{ip}) = 0$.*

Proof. The proof of Item 1 is straightforward. To prove Item 2, for $i \geq 0$ let L_i be the language $L_i \overset{\text{def}}{=} \mathbb{B}_{t,v\mathtt{L}^i\mathtt{R}} = \{t' \in \mathcal{T}_\Gamma^\omega \mid t \sqsubseteq t'.(v\mathtt{L}^i\mathtt{R})\}$. Then, for every $j \geq 0$ we have that $L_j \subseteq L_{\mathrm{fs}}$ and, in consequence, $\overline{L_{\mathrm{fs}}} \subseteq \bigcap_{j \geq i \geq 0} \overline{L_i}$. Hence, for every $j \geq 0$ we have that

$$1 - \mu^*(L_{\mathrm{fs}}) = \mu^*(\overline{L_{\mathrm{fs}}}) \leq \mu^*\Big(\bigcap_{j>i\geq 0} \overline{L_i}\Big) \leq \big(1 - |\Gamma|^{|Dom(t)|}\big)^j,$$

where the last inequality follows from the fact that the nodes $v\mathtt{L}^l\mathtt{R}$ and $v\mathtt{L}^k\mathtt{R}$ are incomparable for $k \neq l$, thus L_i are independent sets and

$$\mu^*\Big(\bigcap_{j>i\geq 0} \overline{L_i}\Big) = \prod_{0\leq i<j} \mu^*(\overline{L_i}) = \prod_{0\leq i<j} \big(1 - |\Gamma|^{|Dom(t)|}\big) = \big(1 - |\Gamma|^{|Dom(t)|}\big)^j$$

Taking the limit, we conclude Item 2.

To prove Item 3, let t_i be a sequence of finite trees such that for every $i \geq 0$ we have that $t_i \sqsubseteq t_{i+1} \sqsubseteq t$ and $|Dom(t_i)| < |Dom(t_{i+1})|$. Since the sequence of sets $\mathbb{B}_{t_i,v}$ is decreasing and its limit contains the set $\mathbb{B}_{t,v}$, i.e. $\mathbb{B}_{t_i,v} \supseteq \mathbb{B}_{t_{i+1},v} \supseteq \mathbb{B}_{t,v}$, we have that $\mu^*(\mathbb{B}_{t,v}) \leq \lim_{i\to+\infty} \mu^*(\mathbb{B}_{t_i,v}) = \lim_{i\to+\infty} |\Gamma|^{-|Dom(t_i)|} = 0$. □

Example 3.2. *Let* $\Gamma = \{a, b, c\}$.

1. *If* L_a *is the language of trees over the alphabet* Γ *with arbitrarily long sequences of* a*-labelled nodes, i.e.,* $L_a = \{t \in \mathcal{T}_\Gamma^\omega \mid \forall k \geq 0.\ \exists w, v \in \{\mathtt{L},\mathtt{R}\}^*.\big(\,(|v| \geq k) \wedge \forall\ u \sqsubseteq v.\ t(wu) = a\big)\}$*, then* $\mu^*(L_a) = 1$.
2. *If* L_{a3} *is the language of trees over the alphabet* Γ *with an infinite* $\{a\}$*-labelled path, i.e.* $L_{a3} = \{t \in \mathcal{T}_\Gamma^\omega \mid \exists w \in \{\mathtt{L},\mathtt{R}\}^\omega.\forall u \sqsubseteq w.\ t(u) = a\}$*, then* $\mu^*(L_{a3}) = 0$.
3. *If* L_{a2} *is the language of trees over the alphabet* $\{a, b\}$ *with an infinite* $\{a\}$*-labelled path, i.e.* $L_3 = \{t \in \mathcal{T}_{\{a,b\}}^\omega \mid \exists w \in \{\mathtt{L},\mathtt{R}\}^\omega.\forall u \sqsubseteq w.\ t(u) = a\}$*, then* $\mu^*(L_{a2}) = 0$.
4. *If* L_{ab} *is the language of trees over the alphabet* Γ *with an infinite* $\{a, b\}$*-labelled path, i.e.* $L_{ab} = \{t \in \mathcal{T}_\Gamma^\omega \mid \exists w \in \{\mathtt{L},\mathtt{R}\}^\omega.\forall u \sqsubseteq w.\ t(u) \in \{a, b\}\}$*, then* $\mu^*(L_{ab}) = \frac{1}{2}$.

Calculating the measures. To show Item 1, let t^i be a complete tree of height i with every node in $Dom(t^i)$ labelled a and let L^i be the language of trees having t^i as a sub-tree. Then by Lemma 3.1 part 2 we have that $\mu^*(L^i) = 1$. Moreover, $\bigcap_{i\geq 1} L^i \subseteq L_a$ and $L^{i+1} \subseteq L^i$. Since the measure is monotonically continuous, we have that

$$\mu^*(L_a) \geq \mu^*\Big(\bigcap_{i\geq 1} L^i\Big) = \lim_{n\to+\infty} \mu^*\Big(\bigcap_{n\geq i\geq 1} L^i\Big) = 1. \tag{2}$$

Let $\phi(t)$ stay for "there is an infinite a-labelled path in the tree t" then Item 2 follows from the fact that the language in question is regular, thus measurable, and its measure satisfies the following equation.

$$\begin{aligned}\mu^*(L_{a3}) = {} & \mu^*(t(\varepsilon)\neq a)\cdot 0 + \\ & \mu^*(t(\varepsilon)=a)\cdot\big(\mu^*(\phi(t.\mathtt{L})) + \mu^*(\phi(t.\mathtt{R})) - \mu^*(\phi(t.\mathtt{L}) \wedge \phi(t.\mathtt{R}))\big)\end{aligned}$$

Thus, we get the equation

$$\mu^*(L_{a3}) = \frac{1}{3}\cdot\big(2\mu^*(L_{a3}) - \mu^*(L_{a3})^2\big) = \frac{2}{3}\mu^*(L_{a3}) - \frac{1}{3}\cdot\mu^*(L_{a3})^2 \tag{3}$$

implying that $\mu^*(L_{a3}) = 0$, since the measure cannot be negative.

Similarly, in Item 3 we get the equation

$$\mu^*(L_{a2}) = \frac{1}{2}\cdot\big(2\mu^*(L_{a2}) - \mu^*(L_{a2})^2\big) = \mu^*(L_{a2}) - \frac{1}{2}\cdot\mu^*(L_{a2})^2 \tag{4}$$

implying that $\mu^*(L_{a2}) = 0$.

In Item 4, we get the equation

$$\mu^*(L_{ab}) = \frac{2}{3}\cdot\big(2\mu^*(L_{ab}) - \mu^*(L_{ab})^2\big) = \frac{4}{3}\mu^*(L_{ab}) - \frac{2}{3}\cdot\mu^*(L_{ab})^2 \tag{5}$$

implying that either $\mu^*(L_{ab}) = \frac{1}{2}$ or $\mu^*(L_{ab}) = 0$. Thus we need to look at this example a bit more carefully. Consider a sequence of languages A^i, where $A^0 = \mathcal{T}_\Gamma$ and A^i is the language such that there is a

$\{a,b\}$-labelled path of length i beginning at the root. Then, by König's lemma we have that $\bigcap_{i\geq 1} A^i = L_{ab}$. Moreover, for every $i > 0$ we have that $A^{i+1} \subseteq A^i$ and

$$\mu^*(A^{i+1}) = \frac{2}{3}\cdot\left(2\mu^*(A^i) - \mu^*(A^i)^2\right) = \frac{4}{3}\mu^*(A^i) - \frac{2}{3}\cdot\mu^*(A^i)^2. \tag{6}$$

Now, note that if $\mu^*(A^i) \geq \frac{1}{2}$, then $\mu^*(A^{i+1}) \geq \frac{1}{2}$. Indeed, the quadratic function $f(x) = \frac{2}{3}(2x - x^2)$ is monotonically increasing on the interval $[-\infty, 1]$ and we have that $f(1) = \frac{2}{3}$ and $f(\frac{1}{2}) = \frac{1}{2}$. Since $\mu^*(A^0) = 1 \geq \frac{1}{2}$, we conclude that $\mu^*(L_{ab}) = \frac{1}{2}$. □

4 First-order definable languages and their standard measures

The ideas presented in both Lemma 3.1 and Example 3.2 allow us to compute the measures of tree languages defined by some FO formulae.

Theorem 4.1. *Let φ be a* FO *sentence over the signature $\Gamma \cup \{\varepsilon, s_{\mathrm{L}}, s_{\mathrm{R}}, s\}$. Then, the measure $\mu^*(\mathrm{L}(\varphi))$ is rational and computable in three-fold exponential time.*

The proof utilises the *Gaifman locality* to partition the formula into two separate sub-formulae. Intuitively, one sub-formulae describes the neighbourhood of the root while the other describes the tree "far away from the root".

Before we prove the above theorem, we define the idea of a root formula, i.e. a formula that necessarily describes the neighbourhood of the root. Let $\mathcal{A}$ be a logical structure. The *Gaifman graph* of $\mathcal{A}$ is the undirected graph $G^{\mathcal{A}}$ where the set of vertices is the universe of $\mathcal{A}$ and there is an edge between two vertices in $G^{\mathcal{A}}$ if there is a relation R in $\mathcal{A}$ and a tuple $x \in R$ that contains u and v. The distance $d(u,v)$ between two elements u, v of the universe of $\mathcal{A}$ is the distance between u, v in the Gaifman graph.

Before we proceed, let us note that in this section we disallow the use of $\lneq$ relation in formulae. Hence, the Gaifman graf of a tree t is induced by the child relations only, and so is the notion distance.

We say that a first-order formula $\varphi(x)$ is a r-local formula around x if the quantifiers are restricted to r-neighbourhood of x, i.e. if $\varphi(x)$ uses the quantifiers $\forall^{\leq r}$ and $\exists^{\leq r}$ defined as follows: $\exists^{\leq r} y.\psi(y) \stackrel{\text{def}}{=} \exists y.\psi(y) \wedge d(x,y) \leq r$ and $\forall^{\leq r} y.\psi(y) \stackrel{\text{def}}{=} \forall y.(d(x,y) \leq r) \to \psi(y)$, where $d(x,y) \leq r$ is a first-order formula stating that the distance between x and y is at most r.

We say that a first-order sentence φ is a *basic r-local sentence* if it is of form

$$\exists x_1, \ldots, x_n \Big(\bigwedge_{i=1}^{n} \varphi_i(x_i) \wedge \bigwedge_{0\leq i<j\leq s} d(x_i, x_j) > 2r, \Big) \tag{7}$$

where $\varphi_i(x)$ are r-local formulae around x and $d(x,y) > 2r$ is a first-order formula stating that the distance between x and y is strictly greater than $2r$.

Let $t \in \mathcal{T}_\Gamma$ be a tree and let φ be a basic r-local sentence, then $t \vDash \varphi$ if and only if there is a function $\tau : x_1, \ldots, x_n \to Dom(t)$ mapping variables $x_1, \ldots, x_n$ to the nodes of t so that for every $i \in \{1, \ldots, n\}$ we have that $t, \tau(x_i) \vDash \varphi_i(x_i)$.

Theorem 4.2 (Gaifman)**.** *Every first-order sentence is equivalent to a Boolean combination of basic r-local sentences, where r is a number depending on the size of the formula. Furthermore, r can be chosen so that $r \leq 7^{qr(\varphi)}$, where $qr(\varphi)$ is the quantifier rank of φ.*

As proved by Heimberg et al., cf. [8], the translation to Gaifman normal form can be costly.

Theorem 4.3 ([8])**.** *There is a three-fold exponential algorithm on structures of degree 3 that transforms a first-order formula into its Gaifman normal form. Moreover, there are first-order formulae for which the three-fold exponential blow-up is unavoidable.*

Let $\psi(x)$ be a r-local formula around x. We say that $\psi(x)$ is a *root* formula if for every tree $t \in \mathcal{T}_\Gamma$ and every node $u \in \{\mathtt{L},\mathtt{R}\}^*$ if $t, u \vDash \psi(x)$ then $\mathrm{d}(u,\varepsilon) < r$. Notice that every unsatisfiable formula is, by the definition, a root formula. Let φ be a basic r-local sentence. We say that $\varphi_i(x)$ for $i \in \{1,\dots,n\}$ *is a root formula of* φ if $\varphi_i(x)$ is a root formula.

Fact 4.4. *For every satisfiable basic r-local sentence there is at most one root formula.*

With the above definitions, we can describe the connection between the basic local sentences and the standard measure.

Lemma 4.5. *Let φ be a basic r-local sentence, i.e. as in Equation* (7)*. If φ is*

- *not satisfiable, then $\mu^*(\mathrm{L}(\varphi)) = 0$,*
- *satisfiable and has no root formula, then $\mu^*(\mathrm{L}(\varphi)) = 1$,*
- *satisfiable and has a root formula φ^*, then for every t^r that is a complete tree of height $2r+1$*

$$\mu^*(\mathrm{L}(\varphi) \cap \mathbb{B}_{t^r}) = \begin{cases} \mu^*(\mathbb{B}_{t^r}) & \text{if there is } u \in \{\mathtt{L},\mathtt{R}\}^{\leq r} \text{ such that } \quad t^r, u \vDash \varphi^*(x); \\ 0 & \text{otherwise.} \end{cases}$$

Proof. If φ is not satisfiable then $\mathrm{L}(\varphi) = \varnothing$ and $\mu^*(\mathrm{L}(\varphi)) = 0$. Therefore, let us assume that φ is satisfiable. By Fact 4.4 we know that there is at most one root formula in φ. Let I be the set of indices of not root formulae, i.e. for $i \in I$ we have that φ_i is not a root formula. Since φ is satisfiable then for every $i \in I$ there is a finite tree $t_i \in \mathcal{T}_\Gamma$ and a node $u_i \in \{\mathtt{L},\mathtt{R}\}^*$ of length $|u_i| > r$, such that $t_i, u_i \vDash \varphi_i(x)$, and the set $Dom(t_i)$ contains the r-neighbourhood of u_i.

Let $W = \{v_i\}_{i=1}^n$ be a set of n $\sqsubseteq$-incomparable nodes such that for $i \in I$ we have that $|v_i| > 2r$. Let $F = \bigcap_{i\in I} L_i$ where L_i is the set of trees having t_i as a sub-tree rooted below the node v_i, i.e. $L_i \stackrel{\text{def}}{=} \{t' \in \mathcal{T}_\Gamma \mid \exists u.(v_i \sqsubsetneq u) \wedge (t_i \sqsubsetneq t'.u)\}$. Since, by Lemma 3.1, every L_i has measure 1, we have that $\mu^*(F) = 1$. Moreover, for every tree $t \in F$ and index $i \in I$ there is a node $v_i' \in \{\mathtt{L},\mathtt{R}\}^*$ such that $\mathrm{d}(v_i, v_i') > r$, $v_i \sqsubseteq v_i'$ and $t, v_i' \vDash \varphi_i(x)$.

Now, if there is no root formula in φ, i.e. $I = \{1\dots,n\}$, then $F \subseteq \mathrm{L}(\varphi)$. Indeed, let $t \in F$ then for $i \neq j$ we have that $\mathrm{d}(v_i', v_j') > 2r$ and we can infer that $t \vDash \varphi$. Hence, the sequence of inequalities

$$1 = \mu^*(F) \leq \mu^*(\mathrm{L}(\varphi)) \leq 1$$

is sound and proves the second bullet.

On the other hand, let there be a root formula in φ. Without loss of generality φ_1 is the root formula and $I = \{2,\dots,n\}$. Let F and v_i's be as before and let $t \in F \cap \mathbb{B}_{t^r}$. If there is $u_1 \in \{\mathtt{L},\mathtt{R}\}^{\leq r}$ such that $t^r, u_1 \vDash \varphi^*(x)$ then we take $v_1' \stackrel{\text{def}}{=} u_1$. Now, again, for $i \neq j$ we have that $\mathrm{d}(v_i', v_j') > 2r$ and for all $i \in I$ we have that $t, v_i' \vDash \varphi_i(x_i)$. In other words, if there is $u_1 \in \{\mathtt{L},\mathtt{R}\}^{\leq r}$ such that $t^r, u_1 \vDash \varphi^*(x)$ then $F \cap \mathbb{B}_{t^r} \subseteq \mathrm{L}(\varphi) \cap \mathbb{B}_{t^r}$. Moreover, since F is of measure 1, the following sequence of inequalities is sound

$$\mu^*(\mathbb{B}_{t^r}) = \mu^*(F \cap \mathbb{B}_{t^r}) \leq \mu^*(\mathrm{L}(\varphi) \cap \mathbb{B}_{t^r}) \leq \mu^*(\mathbb{B}_{t^r}).$$

Furthermore, if there is no such u_1 then φ_1 is not satisfiable in t^r. Since φ_1 is a root formula, we have that $\mu^*(\mathbb{B}_{t^r} \cap \mathrm{L}(\varphi)) = 0$, which concludes the proof. □

Intuitively, the above lemma states that when we consider the uniform measure and a basic r-local sentence, the behaviour of the sentence is almost surely defined by the neighbourhood of the root. This intuition can be formalised as follows.

Lemma 4.6. *Let φ be a basic r-local sentence. Then, there is a sentence φ^* such that for every complete tree t^r of height $2r+1$*

$$\mu^*(\mathrm{L}(\varphi) \cap \mathbb{B}_{t^r}) = \mu^*(\mathrm{L}(\varphi^*) \cap \mathbb{B}_{t^r}).$$

Moreover, for every tree $t \in \mathbb{B}_{t^r}$ we have that $t \vDash \varphi$ if and only if $t^r \vDash \varphi^*$. We call the formula φ^* the reduction of φ.

Proof. If φ has a root formula φ_i then we take $\varphi^* \stackrel{\text{def}}{=} \exists x.\varphi_i(x) \wedge \mathrm{d}(x,\varepsilon) < r$. If φ has no root formulae but is satisfiable then we take $\varphi^* \stackrel{\text{def}}{=} \exists x.\varepsilon(x)$. Otherwise we take $\varphi^* \stackrel{\text{def}}{=} \bot$. □

The above result can be extended to Boolean combinations of r-local basic formulae by the following property of measurable sets.

Lemma 4.7. *Let M be a measurable space with measure μ, S be a measurable set and $\{S_i\}_{i\in I}$ be a family of measurable sets such that for every $i \in I$ either $\mu(S \cap S_i) = 0$ or $\mu(S \cap S_i) = \mu(S)$. Then for every set W in the Boolean algebra of sets generated by $\{S_i\}_{i\in I}$ we have that either $\mu(S \cap W) = 0$ or $\mu(S \cap W) = \mu(S)$.*

Proof. The proof goes by a standard inductive argument. □

Hence, by Lemma 4.5 and the above lemma, we obtain the following.

Lemma 4.8. *Let ϕ be a boolean combination of basic r-local formulae and t be a complete tree of height $2r+1$. Then, $\mu^*(\mathrm{L}(\phi) \cap \mathbb{B}_t) = \mu^*(\mathrm{L}(\phi^*) \cap \mathbb{B}_t)$, where ϕ^* is the reduction of ϕ, i.e. the Boolean combination ϕ with its every basic r-local sentence φ replaced by its reduction φ^*, as defined in Lemma 4.6.*

Moreover, $\mu^*(\mathrm{L}(\phi^*) \cap \mathbb{B}_t) = \begin{cases} \mu^*(\mathbb{B}_t) & \text{if } t \vDash \phi^*; \\ 0 & \text{otherwise.} \end{cases}$

With above lemmas we can finally prove Theorem 4.1.

Proof of Theorem 4.1. Let φ be a first order sentence as in the theorem. We utilise the Gaifman locality theorem (see Theorem 4.2 on page 211) to translate the sentence φ into a Boolean combination ϕ of basic r-local sentences. Now, let ϕ^* be the reduction of ϕ, as in Corollary 4.8, and let $S \subseteq \mathcal{T}_\Gamma$ be the set of all complete trees of height $2r+1$. Then

$$\begin{array}{lcllcl} \mu^*(\mathrm{L}(\phi)) & \stackrel{1}{=} & \mu^*(\mathrm{L}(\phi) \cap (\bigcup_{t\in S} \mathbb{B}_t)) & \stackrel{2}{=} & \mu^*(\bigcup_{t\in S} (\mathrm{L}(\phi) \cap \mathbb{B}_t)) \\ & \stackrel{3}{=} & \sum_{t\in S} \mu^*(\mathrm{L}(\phi) \cap \mathbb{B}_t) & \stackrel{4}{=} & \sum_{t\in S} \mu^*(\mathrm{L}(\phi^*) \cap \mathbb{B}_t) \\ & \stackrel{5}{=} & \sum_{t\in S \wedge t\vDash\phi^*} \mu^*(\mathbb{B}_t) & \stackrel{6}{=} & |\{t \in S \mid t \vDash \phi^*\}| \cdot \frac{1}{2^{2^{r+1}-1}}. \end{array}$$

The first equation follows from the fact that $\{\mathbb{B}_t \mid t \in S\}$ is a partition of the space. The second from operations on sets and the third from a simple property of measures. The fourth from the first part of Lemma 4.8, while the fifth follows from the second part of this lemma. The last equation is a consequence of the fact that $\mu^*(\mathbb{B}_t) = 2^{-|Dom(t)|}$.

Since $\mu^*(\mathrm{L}(\phi)) = \frac{|\{t\in S | t\vDash\psi\}|}{2^{2^{r+1}-1}}$, it is enough to count how many complete trees of height $2r+1$ satisfy the reduced combination. Thus, the complexity upper bound comes from the fact that translating a

first-order formula into a Gaifman normal form can produce a three-fold exponential formula in result, see Theorem 4.3, which then can be checked against two-fold exponential number of trees of size that is two-fold exponential in the size of the original formula. □

Algorithm 1 FO measure

Require: a FO formula φ
$S \stackrel{\text{def}}{=}$ the set of all of all complete trees of height $2r+1$
$\phi \stackrel{\text{def}}{=} \text{Gaifman}(\varphi)$
$\phi \stackrel{\text{def}}{=} \text{extractRootFormulae}(\phi)$
$S \stackrel{\text{def}}{=} \{t \in S \mid t \vDash \phi\}$
return $|S| \cdot 2^{-2^{r+1}+1}$

The technique used to prove Theorem 4.1 cannot be extended to formulae utilising the descendant relation because, as presented in Example 4.9, languages defined by such formulae can have irrational measures.

Proposition 4.9. *Let* Γ *be an alphabet. Then there is a language definable in by a* FO *formula over the signature* $\Gamma \cup \{s_{\mathtt{L}}, s_{\mathtt{R}}, \between\}$ *with an irrational standard measure.*

Proof. Let $\Gamma = \{a, b\}$, we define a language L in the following way $L \stackrel{\text{def}}{=} \{t \in \mathcal{T}_{\{a,b\}} \mid$ for every path the earliest node labelled b (if exists) is at an even depth$\}$. Now, the measure $\mu^*(L)$ is irrational, and there is a language L' definable by a first-order formula over the signature $\Gamma \cup \{s_{\mathtt{L}}, s_{\mathtt{R}}, \between\}$ such that $\mu^*(L') = \mu^*(L)$. We start with computing the measure of L, then we will define L'.

Observe that the measure $\mu^*(L)$ satisfies the following equation.

$$\mu^*(L) = \mu^*\big(\{t \in \mathcal{T}^{\omega}_{\{a,b\}} \mid t(\varepsilon){=}b\}\big) + \mu^*\big(t \in \mathcal{T}^{\omega}_{\{a,b\}} \mid t(\varepsilon){=}t(\mathtt{R}){=}t(\mathtt{L}){=}a\big) \cdot \mu^*(L)^4$$

After substituting the appropriate values, we obtain the equation

$$\mu^*(L) = \frac{1}{2} + \frac{1}{8}\mu^*(L)^4 \tag{8}$$

which by the *rational root theorem* has no rational solutions.

To end the proof, we will describe how to define the language L'. The crux of the construction comes from the beautiful example by Potthoff, see [13, Lemma 5.1.8]. We will use the following interpretation of the lemma: one can define in first-order logic over the signature $\{a, b, s_{\mathtt{L}}, s_{\mathtt{R}}, \between\}$ a language of finite trees over the alphabet $\{a, b\}$ where every a-labelled node has exactly two children end every b-labelled node is a leaf on an even depth.

To construct L' we simply utilise the formula defining the language in the Potthoff's example to define L' by substituting the leaf with the first occurrence of the label b. Note that the set L' agrees with L on every tree that has a label b on every infinite path from the root. On the other hand, the truth value of the modified formula on the trees that have an infinite path from the root with no b-labelled nodes, i.e. on the set L_{a2} from Example 3.2, is of no concern to us. Indeed, as previously shown, the standard measure of the set L_{a2} is 0.

To be precise, for every tree $t \in \mathcal{T}^{\omega}_{\{a,b\}} \setminus L_{a2}$ we have that $t \in L \iff t \in L'$, where L_{a2} is a language from the Example 3.2. Therefore, we have that $L \cup L_{a2} = L' \cup L_{a2}$. Since $\mu^*(L_{a2}) = 0$, we have that

$$\mu^*(L) = \mu^*(L \cup L_{a2}) = \mu^*(L' \cup L_{a2}) = \mu^*(L'),$$

which concludes the proof. □

5 Conjunctives queries and the standard measure

Introducing the ancestor/descendant relation to the tree structure causes that every two nodes in the Gaifman graph are in distance at most two from each other. Thus, for the purpose of having a relevant definition of the distance in the tree, we retain the child related notion of distance, i.e. in this section, as before, the notion of the distance is induced by the child relations only.

Conjunctive queries A *concjunctive query* (CQ) over an alphabet Γ is a formula of first-order logic, using only conjunction and existential quantification, over unary predicates $a(x)$, for $a \in \Gamma$, the root predicate $\varepsilon(x)$, and binary predicates $s_{\mathrm{L}}, s_{\mathrm{R}}, s, \dot{\leq}$.

An alternative way of looking at conjunctive queries is via graphs and graph homomorphisms. A pattern π over Γ is a relational structure $\pi = \langle V, V_\varepsilon, E_{\mathrm{L}}, E_{\mathrm{R}}, E_s, E_{\dot{\leq}}, \lambda_\pi \rangle$, where $\lambda_\pi \colon V \rightharpoonup \Gamma$ is a partial labelling, V_ε is the set of root vertices, and $G_\pi = \langle V, E_{\mathrm{L}} \cup E_{\mathrm{R}} \cup E_s \cup E_{\dot{\leq}} \rangle$ is a finite graph whose edges are split into left child edges E_{L}, right child edges E_{R}, child edges E_s, and ancestor edges $E_{\dot{\leq}}$. By $|\pi|$ we mean the size of the underlying graph.

We say that a tree $t = \langle Dom(t), s_{\mathrm{L}}, s_{\mathrm{R}}, \dot{\leq}, (a^t)_{a \in \Gamma} \rangle$ satisfies a pattern $\pi = \langle V, V_\varepsilon, E_{\mathrm{L}}, E_{\mathrm{R}}, E_s, E_{\dot{\leq}}, \lambda_\pi \rangle$, denoted $t \vDash \pi$, if there exists a homomorphism $h \colon \pi \to t$, that is a function $h \colon V \to Dom(t)$ such that

1. $h \colon \langle V, E_{\mathrm{L}}, E_{\mathrm{R}}, E_s, E_{\dot{\leq}} \rangle \to \langle Dom(t), s_{\mathrm{L}}, s_{\mathrm{R}}, s_{\mathrm{L}} \cup s_{\mathrm{R}}, \dot{\leq} \rangle$ is a homomorphism of relational structures,
2. for every $v \in V_\varepsilon$ we have that $h(v) = \varepsilon$,
3. and for every $v \in Dom(\lambda_\pi)$ we have that $\lambda_\pi(v) = \lambda_t(h(v))$.

Every pattern can be seen as a conjunctive query and vice versa. Hence, we will use those terms interchangeably. The class of conjunctive queries is denoted CQ, the class of formulae that are Boolean combination of conjunctive queries is denoted BCCQ.

Despite allowing the use of ancestor in conjunctive queries, the measure of the language defined by a conjunctive query is rational and computable.

Theorem 5.1. *Let q be a conjunctive query over the signature $\Gamma \cup \{\varepsilon, s_{\mathrm{L}}, s_{\mathrm{R}}, s, \dot{\leq}\}$. Then, the measure of the language $\mathrm{L}(q)$ is rational and computable in exponential space.*

To prove the theorem we will use the concept of *firm* sub-patterns, used e.g. in [12]. Intuitively, a *firm pattern* is a conjunctive query that has to be mapped in a small neighbourhood.

A sub-pattern π' is *firm* if it is a sub-pattern of a pattern π induced by vertices belonging to a maximal strongly connected component in graph $G_\pi = \langle V, E \rangle$ such that $\langle x, y \rangle \in E$ if either $xs_{\mathrm{L}}y$, $ys_{\mathrm{L}}x$, $xs_{\mathrm{R}}y$, $ys_{\mathrm{R}}x$, xsy, ysx, $x\dot{\leq}y$, or $\varepsilon(x)$. In particular, a pattern is firm if it has a single strongly connected component. We say that a sub-pattern is *rooted* if it contains predicate ε.

Proposition 5.2. *Let π be a firm pattern. Then for every tree t such that $t \vDash \pi$, for every two vertices x, y in V, and for every homomorphism $h \colon \pi \to t$ we have that $d(h(x), h(y)) < |\pi|$. Moreover, if π is rooted then for every vertex x we have that $d(h(x), \varepsilon) < |\pi|$.*

Proof. Let us assume otherwise, let $n = |\pi|$. Then, there is a tree t, a homomorphism h, and two vertices x, y such that $t \vDash \pi$ and $d(h(x), h(y)) \geq n$. We claim that x and y are not in the same strongly connected component.

Since for some m we have that $d(h(x), h(y)) = m - 1 \geq n$, there is a sequence of distinct nodes $u_1, u_2, \dots u_m$ such that $u_1 = h(x)$, $u_m = h(y)$ and for every i, u_i and u_{i+1} are in a child relation. Moreover, there is a node u such that $u = u_i$ for some $1 \leq i \leq m$, $u \notin h(\pi)$, and one of the nodes $h(x)$ or $h(y)$ is a descendant of u. Without loss of generality, let us say that $u \lneq h(y)$. Or, more precisely, that $u\mathtt{L} \sqsubseteq h(y)$.

If x and y were in a strongly connected component then there would be a path in the graph G_π that connects y to x, i.e. a sequence of vertices $y_1, y_2, \dots, y_k$, for some k, such that $y_1 = y$, $y_k = x$, and for every $i = 1, \dots, k-1$ there is an edge between y_i and y_{i+1} in G_π. In particular, this implies that for every i we have that $h(y_i)$ and $h(y_{i+1})$ are $\sqsubseteq$-comparable. Now, there would also exist an index $j \in \{1, \dots, k-1\}$ such that $h(y_{j+1}) \lneq u \lneq h(y_j)$. Indeed, if there would be no such index, then all the vertices y_i would satisfy $u\mathtt{L} \sqsubseteq h(y_i)$, as y_i and y_{i+1} are $\sqsubseteq$-comparable for every index i. But this impossible because if $u\mathtt{L} \sqsubseteq h(y_i)$ for all i, then we would have that $u\mathtt{L} \sqsubseteq h(y_k) = h(x)$. Now, since $u\mathtt{L} \sqsubseteq h(y)$ and $u\mathtt{L} \sqsubseteq h(x)$, then by the definition of the distance u would not belong to the sequence $u_1, \dots, u_m$. Which is a contradiction with our assumption.

Therefore, there is an index j such that $h(y_{j+1}) \lneq u \lneq h(y_j)$. Thus, by the definition of G_π we have that either $y_j s_\mathtt{L} y_{j+1}$, $y_{j+1} s_\mathtt{L} y_j$, $y_j s_\mathtt{R} y_{j+1}$, $y_{j+1} s_\mathtt{R} y_j$, $y_j s y_{j+1}$, $y_{j+1} s y_j$, $y_j \lneq y_{j+1}$, or $\varepsilon(y_j)$. Either of child relations is impossible because the distance between $h(y_j)$ and $h(y_{j+1})$ is at least two. Similarly, both $y_j \lneq y_{j+1}$ and $\varepsilon(y_j)$ are impossible because we have that $u \lneq h(y_j)$. Hence, there is no such sequence $y_1, \dots y_k$, thus x and y cannot belong to the same strongly connected component. This proves the first part of the lemma.

Now, if π is rooted then there is a vertex y such that for every homomorphism h we have that $h(y) = \varepsilon$. Hence, by the first part for all vertices $x \in \pi$ we have that $d(h(x), \varepsilon) = d(h(x), h(y)) < n$. □

Let π be a pattern. Consider a graph $\mathrm{G}^F_\pi = \langle V, E \rangle$ where V is the set of firm sub-patterns of π and there is and edge $\langle v_1, v_2 \rangle \in E \subseteq V \times V$ between two vertices $v_1, v_2 \in V$ if and only if there is an $\lneq$ labelled edge between some two vertices $w_1 \in v_1, w_2 \in v_2$. We call this graph the *graph of firm sub-patterns* of the pattern π.

Proposition 5.3. *The directed graph G^F_π of firm sub-patterns of a pattern π is acyclic and has at most one rooted firm sub-pattern. We call this sub-pattern* the root pattern.

Proof. By the definition of the firm sub-pattern, every vertex with predicate ε ends up in the same maximal strongly connected component. The acyclicity follows directly from the fact that firm sub-patterns are the maximal strongly connected components. □

As in the case of root formulae, the root pattern decides of the behaviour of a satisfiable conjunctive query.

Lemma 5.4. *Let Γ be an alphabet and q be a conjunctive query over the signature $\Gamma \cup \{\varepsilon, s_\mathtt{L}, s_\mathtt{R}, s, \lneq\}$. Then, either*

- *q is not satisfiable and $\mu^*(\mathrm{L}(q)) = 0$,*
- *q is satisfiable, has no root sub-pattern, and $\mu^*(\mathrm{L}(q)) = 1$,*
- *or q is satisfiable, has a root sub-pattern p, and $\mu^*(\mathrm{L}(q)) = \mu^*(\mathrm{L}(p))$.*

Proof. Let π be a pattern equivalent to q. If q is not satisfiable then $\mathrm{L}(q) = \varnothing$ and $\mu^*(\mathrm{L}(q)) = 0$. Let q be satisfiable, i.e. there is a tree t^q and a homomorphism $h \colon \pi \to t^q$. Let t^r be a finite tree such that

$h(\pi) \subseteq Dom(t^r)$ and let the set $S \subseteq \mathcal{T}_\Gamma$ be the set of all trees t such that for every node $u \in \{\mathtt{L},\mathtt{R}\}^{|q|+1}$ the tree t^r is a sub-tree of $t.u$. By Lemma 3.1 we have that $\mu^*(S) = 1$.

If π has no root firm sub-pattern then $S \subseteq \mathrm{L}(\pi)$ and we have that

$$\mu^*(\mathrm{L}(\pi)) \geq \mu^*(S) = 1.$$

On the other hand, if π has a root firm sub-pattern p then for every tree $t \in S$ we have that $t \vDash \pi$ if and only if $t \vDash p$. Thus, $\mathrm{L}(\pi) \cap S = \mathrm{L}(p) \cap S$ and since $\mu^*(S) = 1$ we have that

$$\mu^*(\mathrm{L}(\pi)) = \mu^*(\mathrm{L}(\pi) \cap S) = \mu^*(\mathrm{L}(p) \cap S) = \mu^*(\mathrm{L}(p)).$$

□

In other words, the problem of computing measure of a language defined by a conjunctive query reduces to the below problem of counting the models of fixed depth.

Problem 5.5 (The models counting problem)**.**

Input: *A conjunctive query q and a natural number n.*
Output: *Number of complete binary trees of height n that satisfy q.*

Proposition 5.6. *The models counting problem can be solved in exponential space.*

Indeed, all we need is to enumerate all binary trees of height linear in the size of the query. As an immediate consequence we infer Theorem 5.1.

For a lower bound of the problem of computing the standard measure of a conjunctive query, we observe that deciding whether the measure of a language defined by a conjunctive query is positive is intractable.

Proposition 5.7. *The positive $\mu^*(CQ)$ problem is NP-complete.*

Proof. Let q be a conjunctive query. Then, either q is not satisfiable and $\mu^*(q) = 0$, or q is satisfiable and has positive measure. That is, $\mu^*(q) > 0$ if and only if q is satisfiable. Deciding whether a conjunctive query is satisfiable is *NP*-complete, cf. e.g. [3]. □

As in the case of first-order formulae, we can lift Theorem 5.1 to Boolean combinations of conjunctive queries. Indeed, by Lemma 4.7, the measure of the language defined by a Boolean combination of conjunctive queries is computable.

Corollary 5.8. *Let φ be a Boolean combination of conjunctive queries. Then, the standard measure $\mu^*(\mathrm{L}(\varphi))$ can be computed in exponential space.*

For the lower bound, we observe.

Theorem 5.9. *The positive $\mu^*(BCCQ)$ problem is NEXP-complete.*

Sketch of the proof. First, we prove the upper bound. Let ϕ be a Boolean combination of conjunctive queries. Let m be the maximum over the sizes of conjunctive queries in ϕ. By Lemma 5.4 and Lemma 4.7, we can translate ϕ into a Boolean combination of firm, rooted conjunctive queries ϕ^* such that $\mu^*(\mathrm{L}(\phi)) = \mu^*(\mathrm{L}(\phi^*))$ and ϕ^* is of polynomial size with respect ot ϕ. This can be done in exponential time and requires verifying whether the patterns in ϕ are satisfiable.

Now, since every conjunctive query in ϕ^* is firm and rooted, either there is a finite tree t of depth $2n$ such that $t \vDash \phi^*$ or not. If there is no such tree, then $\mu^*(\mathrm{L}(\phi^*)) = 0$. If there is such a tree, then

$\mathbb{B}_t \subseteq \mathrm{L}(\phi^*)$ and by Lemma 3.1 we have that $\mu^*(\mathrm{L}(\phi^*)) > 0$. Hence, it is enough to guess the tree t and verify that $t \vDash \phi^*$. Since t is of exponential size in ϕ and model checking of a conjunctive query can be done in polynomial time, we infer the upper bound.

For the lower bound, we refer to the proof of Theorem 3 in Murlak et al. [12], the case of non-recursive schemas. The proof can be easily adapted to our needs. □

6 Conclusions and future work

We have shown that there exists an algorithm that, given a first-order sentence φ over the signature $\Gamma \cup \{\varepsilon, s_{\mathrm{R}}, s_{\mathrm{L}}\}$, computes $\mu^*(\mathrm{L}(\varphi))$ in three-fold exponential time. We also have shown that there exists an algorithm that, given a Boolean combination of conjunctive queries φ over the signature $\Gamma \cup \{\varepsilon, s_{\mathrm{R}}, s_{\mathrm{L}}, s, \lightning\}$, computes the standard measure $\mu^*(\mathrm{L}(\varphi))$ in exponential space. Establishing exact bounds of the problems is an interesting direction of future research. We provide some lower bounds for the conjunctive queries in the form of the positive measure problem. We claim, without a proof, that using the same techniques, we can give similar lower bounds for the first-order case.

Note, that the considered measure respects a form of a $0-1$-law. By Lemma 3.1, if t is a finite tree then with probability 1 it appears as a sub-tree in a random tree. It would be interesting to extend the enquiry to measures that do not posses such a property. Such measures can be expressed, for example, by graphs or by branching boards, cf. [14].

Obviously, the most interesting problem is to find an algorithm that can compute the standard measure of an arbitrary regular language of infinite trees. While we know that languages with irrational measures exist, we conjecture that for any regular language of trees L the standard measure $\mu^*(L)$ is algebraic.

Acknowledgements We thank Damian Niwiński and Michał Skrzypczak for inspiring discussions and careful reading of a preliminary version of this paper, and the anonymous referees for helpful comments motivating us to improve the presentation of the paper.

References

[1] Antoine Amarilli, Pierre Bourhis & Pierre Senellart (2015): *Provenance Circuits for Trees and Treelike Instances*. In: *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II*, pp. 56–68, doi:10.1007/978-3-662-47666-6_5.

[2] Antoine Amarilli, Mikaël Monet & Pierre Senellart (2017): *Conjunctive Queries on Probabilistic Graphs: Combined Complexity*. In: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pp. 217–232, doi:10.1145/3034786.3056121.

[3] Henrik Bjørklund, Wim Martens & Thomas Schwentick (2011): *Conjunctive query containment over trees*. *Journal of Computer and System Sciences* 77(3), pp. 450 – 472, doi:10.1016/j.jcss.2010.04.005. Database Theory.

[4] Krishnendu Chatterjee & Thomas A. Henzinger (2012): *A survey of stochastic ω-regular games*. *J. Comput. Syst. Sci.* 78(2), pp. 394–413, doi:10.1016/j.jcss.2011.05.002.

[5] Taolue Chen, Klaus Dräger & Stefan Kiefer (2012): *Model Checking Stochastic Branching Processes*. In: *Mathematical Foundations of Computer Science 2012 - 37th International Symposium, MFCS 2012, Bratislava, Slovakia, August 27-31, 2012. Proceedings*, pp. 271–282, doi:10.1007/978-3-642-32589-2_26.

[6] David Gale & Frank M. Stewart (1953): *Infinite games with perfect information*. In: *Contributions to the theory of games, Annals of Mathematics Studies, no. 28* 2, Princeton University Press, pp. 245–266.

[7] Tomasz Gogacz, Henryk Michalewski, Matteo Mio & Michał Skrzypczak (2014): *Measure Properties of Game Tree Languages*. In: *Mathematical Foundations of Computer Science 2014 - 39th International Symposium, MFCS 2014, Budapest, Hungary, August 25-29, 2014. Proceedings, Part I*, pp. 303–314, doi:10.1007/978-3-662-44522-8_26.

[8] Lucas Heimberg, Dietrich Kuske & Nicole Schweikardt (2013): *An Optimal Gaifman Normal Form Construction for Structures of Bounded Degree*. In: *Proceedings of the 2013 28th Annual ACM/IEEE Symposium on Logic in Computer Science*, LICS '13, IEEE Computer Society, Washington, DC, USA, pp. 63–72, doi:10.1109/LICS.2013.11.

[9] Alexander Kechris (1995): *Classical descriptive set theory*. Springer-Verlag, New York, doi:10.1007/978-1-4612-4190-4.

[10] Henryk Michalewski & Matteo Mio (2015): *On the Problem of Computing the Probability of Regular Sets of Trees*. In: *35th IARCS Annual Conference on Foundation of Software Technology and Theoretical Computer Science, FSTTCS 2015, December 16-18, 2015, Bangalore, India*, pp. 489–502, doi:10.4230/LIPIcs.FSTTCS.2015.489.

[11] Matteo Mio (2012): *Probabilistic modal μ-calculus with independent product*. *Logical Methods in Computer Science* Volume 8, Issue 4, doi:10.2168/LMCS-8(4:18)2012. Available at `https://lmcs.episciences.org/789`.

[12] Filip Murlak, Michal Oginski & Marcin Przybylko (2012): *Between Tree Patterns and Conjunctive Queries: Is There Tractability beyond Acyclicity?* In: *Mathematical Foundations of Computer Science 2012 - 37th International Symposium, MFCS 2012, Bratislava, Slovakia, August 27-31, 2012. Proceedings*, pp. 705–717, doi:10.1007/978-3-642-32589-2_61.

[13] Andreas Potthoff (1994): *Logische Klassifizierung regulärer Baumsprachen*. Ph.D. thesis, Universität Kiel.

[14] Marcin Przybylko & Michał Skrzypczak (2016): *On the Complexity of Branching Games with Regular Conditions*. In: *41st International Symposium on Mathematical Foundations of Computer Science, MFCS 2016, August 22-26, 2016 - Kraków, Poland*, pp. 78:1–78:14, doi:10.4230/LIPIcs.MFCS.2016.78.

[15] Ludwig Staiger (1998): *The Hausdorff Measure of Regular omega-languages is Computable*. *Bulletin of the EATCS* 66, pp. 178–182.

[16] Dan Suciu, Dan Olteanu, Christopher Ré & Christoph Koch (2011): *Probabilistic Databases*. Synthesis Lectures on Data Management, Morgan & Claypool Publishers, doi:10.2200/S00362ED1V01Y201105DTM016.

[17] Wolfgang Thomas (1996): *Languages, Automata, and Logic*. In: *Handbook of Formal Languages*, Springer, pp. 389–455.