The Journal of Symbolic Logic
Volume 86, Number 2, June 2021

TWO-VARIABLE LOGIC HAS WEAK, BUT NOT STRONG, BETH DEFINABILITY

HAJNAL ANDRÉKA AND ISTVÁN NÉMETI

Abstract. We prove that the two-variable fragment of first-order logic has the weak Beth definability property. This makes the two-variable fragment a natural logic separating the weak and the strong Beth properties since it does not have the strong Beth definability property.

§1. Introduction. One of the many expressibility properties of first-order logic with equality FO is the Beth definability property BDP. It states that if a relation can be specified by some extra means then it can be specified explicitly without using the extra means. In more detail, if Th is a first-order logic theory in a language $\mathcal{L}$ and Σ is another first-order logic theory in the language $\mathcal{L}$ expanded with an extra relation symbol R such that in each model of Th there is at most one relation R satisfying Σ, then this unique relation can be defined in the original language $\mathcal{L}$ without using the extra relation symbol, i.e., there is a formula φ in $\mathcal{L}$ such that $\mathsf{Th} \cup \Sigma \models \forall \overline{x}(\mathsf{R}(\overline{x}) \leftrightarrow \varphi)$. In this context, Σ is called the implicit definition and φ is called the explicit definition of R.

Investigating BDP for fragments of FO means showing that if all the formulas of the theory and the implicit definition belong to the fragment, then the explicit definition, too, belongs to it. Thus, having BDP or not shows a kind of "integrity" of the fragment, and a kind of "complexity-property" of FO itself. For example, the guarded fragment GF of FO has BDP [18]. Thus, if the theory Th and the implicit definition Σ consist of guarded formulas, then the explicit definition can be chosen to be guarded, too. We note that in the strict sense, GF does not have Craig interpolation property, yet it is worth deciding when the interpolant belonging to guarded Th and Σ can be chosen to be guarded itself [19]. For work of the similar kind, see e.g., [5].

It is known that n-variable fragments FOn of FO do not have BDP, for all finite $n \geq 2$, see [1]. This means that if the theory Th and the implicit definition Σ all use only n variables, the explicit definition φ may need more than n variables. That the BDP fails for FO2 is kind of surprising, because FO2 usually behaves "better" than FOn for $n \geq 3$. For example, FO2 is decidable while FOn, for $n \geq 3$ is not.

Received October 2, 2020.
2020 *Mathematics Subject Classification*. 03C40, 03B20, 03C95.
Key words and phrases. weak Beth definability, abstract model theory, two-variable logic, homogeneous model.

0022-4812/21/8602-0016
DOI:10.1017/jsl.2021.7

The weak Beth definability property wBDP was introduced by Harvey Friedman [12]. The definition of wBDP is the same as that of BDP except that only those implicit definitions have to be made explicit which also have the existence property, not only the uniqueness property. Let us call these strong implicit definitions. Thus we require that R has an explicit definition only when in each model of Th there is exactly one relation R satisfying Σ, as opposed to having at most one such relation. (For formal definition see Definition 2 at the beginning of Section 3.) Thus, BDP implies wBDP, since wBDP requires fewer definitions to be equivalent to an explicit one. In mathematical practice, one almost always requires both existence and uniqueness for an implicitly defined object. For this reason, wBDP sometimes is considered to be a more natural definability property than BDP itself (see, e.g., [26, p. 129]). We note that wBDP was intensely investigated in abstract model theory in connection with logics stronger than FO, see, e.g., [6, 24].

It is known that FOn does not have wBDP either, whenever $n \geq 3$ ([30] for $n = 3$, [16] for $n \geq 5$, and [3] for $n \geq 3$). Proving failure of wBDP amounted to finding also strong implicit definitions in FOn that could not be made explicit. It remained open whether FO2 had wBDP or not.

In this paper we prove that FO2 does have wBDP, that is, in FO2 all strong implicit definitions can be made explicit. This restores two-variable logic's image that it behaves better than n-variable logics for $n \geq 3$. This theorem may also point to wBDP being a more natural property than BDP.

As far as we know, the present paper contains the first proof for a logic to have the wBDP not via showing that it has the stronger BDP. So, the difference between BDP and wBDP was not tangible so far in the sense that there was no example for a logic that distinguished the two properties. FO augmented with the quantifier "there exists uncountable many" $L(Q)$ was a good candidate for such a distinguishing logic, since it does not have BDP [12] and it is consistent with set theory that it has wBDP [26]. However, it is still an open problem whether wBDP can be proved for $L(Q)$ in set theory or not.

It is also satisfying that the distinguishing logic FO2 is a well-investigated, natural logic. Luckily, we did not have to construct a logic to show that wBDP and BDP are distinct properties, a well-known logic turned out to do the job for us. Two-variable logic and its extensions are quite popular in computer science and in modal logic.

Our proof hinges on the fact that FO2 has a special property that FOn with $n \geq 3$ do not have. Namely, each model of FO2 is FO2-equivalent with a model in which elements of the same FO2-type also have the same automorphism-type (Theorems 1 and 2). FO3 does not have this property (Theorem 3).

The structure of the paper is as follows. In Section 2 we define the above kind of models that we will call transitive and we prove the key property of FO2 about the abundance of transitive models. In Section 3 we prove that FO2 has wBDP by relying on these transitive models (Theorem 4). We close the paper with a remark about some connections with algebra and the literature.

§2. Transitive models of FO2. We use the notation of [9], if not stated otherwise. By FO we mean first-order logic with equality, but we do not allow function or constant symbols.

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

Let n be a finite number. By FOn we mean the fragment of FO that uses only the first n variables. Strictly speaking, the fragment FOn of FO is defined by taking for all languages $\mathcal{L}$ all the models of $\mathcal{L}$ but restricting the set of formulas of FO to those that contain the first n variables only. Thus, relation symbols of arbitrarily high rank can be allowed in FOn. For simplicity, in this paper in FOn we will allow only languages with relation symbols of rank at most n. We go further, we allow relation symbols of rank n only. These are not important restrictions. Usually, we do not indicate the language, but we will always work with similar models, i.e., models having the same language, if not stated otherwise.

We say that $\mathfrak{M}$ and $\mathfrak{N}$ are *n-equivalent*, in symbols $\mathfrak{M} \overset{n}{\equiv} \mathfrak{N}$, when the same n-variable formulas are true in $\mathfrak{M}$ and in $\mathfrak{N}$. In this paper, we concentrate on $n = 2$. By the *type* of an element in a model we understand the set of FO2-formulas true of it. We say that $\mathfrak{M}$ is *transitive* if whenever a and b are elements having the same type in it, there is an automorphism of $\mathfrak{M}$ taking a to b. In concise form:

$$(\mathfrak{M},a) \overset{2}{\equiv} (\mathfrak{M},b) \text{ implies } (\mathfrak{M},a) \cong (\mathfrak{M},b).$$

A typical transitive model is the set of integers with the successor relation. A typical non-transitive model is a connected graph on a set with more than three elements in which there are nodes of distinct degree.

Let us call a model *binary* if all its basic relations are of rank 2. We prove in this section that each binary model is 2-equivalent to a transitive model (Theorem 2). The idea of the proof is that we construct a 2-equivalent version of any binary model via replacing each binary relation in it with a suitable set of different successor relations.

A stronger version of the above will be proved in this section. We call a model $\mathfrak{M}$ 2-*homogeneous* if whenever a and b have the same type in it, to any element $c \in M$ there is $d \in M$ such that $(\mathfrak{M},a,c) \overset{2}{\equiv} (\mathfrak{M},b,d)$. This is a straightforward analogue of the definition of α-homogeneity in [9] where α is any ordinal. We note that transitivity implies 2-homogeneity, but not the other way round. Further, a finite model is always 2-homogeneous (see the proof of Theorem 2).

Finally, we need the notion of 2-partial isomorphism. The notion of n-partial isomorphism was defined in [7, p. 259] as a natural restriction of the usual notion of partial isomorphisms between models of FO (see [9]). We recall the definition of 2-partial isomorphism in detail because we will rely on it.

DEFINITION 1 (2-partial isomorphism). The set I is a 2-*partial isomorphism* between models $\mathfrak{M},\mathfrak{N}$ if (i)– (iv) below hold:

(i) I relates elements as well as pairs of M and N, i.e., it is a subset of $(M \times N) \cup (M^2 \times N^2)$,

(ii) local isomorphism property:
related pairs of I are isomorphisms between $\mathfrak{M}$ and $\mathfrak{N}$ restricted to the first and second parts of the pair, respectively,

(iii) restriction property:
if $\langle (a,a'),(b,b') \rangle \in I$ then $\langle a,b \rangle \in I$ and $\langle a',b' \rangle \in I$, and

(iv) back-and-forth property:
$\forall a \in M \exists b \in N \langle a,b \rangle \in I$ and vice versa, $\forall b \in N \exists a \in M \langle a,b \rangle \in I$,
$\forall \langle a,b \rangle \in I \forall a' \in M \exists b' \in N \langle (a,a'),(b,b') \rangle \in I$, and vice versa
$\forall \langle a,b \rangle \in I \forall b' \in N \exists a' \in M \langle (a,a'),(b,b') \rangle \in I$.

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

Instead of 2-partial isomorphism we will simply say 2-*isomorphism*. It is known that if two pairs are related by a 2-isomorphism, then the same FO2-formulas are true of them, this is straightforward to show by induction. Thus, if there is a 2-isomorphism between two models, then they are 2-equivalent.

Being 2-isomorphic is stronger than being 2-equivalent. We are going to prove that a binary model is 2-isomorphic to a transitive model if and only if it is 2-homogeneous. This stronger theorem will be used in the proof of weak Beth definability property for 2-variable logic FO2 (Theorem 4).

THEOREM 1. *A binary model* $\mathfrak{M}$ *is* 2*-isomorphic to a transitive model* $\mathfrak{N}$ *if and only if* $\mathfrak{M}$ *is* 2*-homogeneous.*

PROOF. We prove sufficiency of 2-homogeneity first. Let $\mathfrak{M}$ be any 2-homogeneous binary model. We are going to define another model $\mathfrak{N}$ and a 2-isomorphism I between them. $\mathfrak{N}$ will be finite when $\mathfrak{M}$ is so. Then we show that $\mathfrak{N}$ is transitive.

Some notation and terminology. The variables of FO2 will be denoted by x, y. By the 2-*type* of $a,b \in \mathfrak{M}$ we understand the set of FO2-formulas $\rho(x,y)$ that are true for a,b in $\mathfrak{M}$. Formally,

$$\mathsf{Type}(a,b,\mathfrak{M}) = \{\rho(x,y) \in FO2 : \mathfrak{M} \models \rho[a,b]\}.$$

Now we define

$$[a] = \{b \in M : \mathsf{Type}(b,b,\mathfrak{M}) = \mathsf{Type}(a,a,\mathfrak{M})\},$$
$$[a,b] = \mathsf{Type}(a,b,\mathfrak{M}),$$
$$\mathsf{Types}([a],[b]) = \{[p,q] : p \in [a],\ q \in [b]\}.$$

We call the elements of $\{[a] : a \in M\}$ 1-*Types*. Note that $[a]$ is a subset of the model, while $[p,q]$ is a set of FO2-formulas.

The type $[p,q]$ determines $[q,p]$, and we call the latter the *converse type* $[p,q]^{\smile}$ of $[p,q]$. Taking the converse is a bijection between $\mathsf{Types}([a],[b])$ and $\mathsf{Types}([b],[a])$. We call a type $[p,q]$ *symmetric* when $[p,q] = [q,p]$, otherwise we call it *asymmetric*. There is a unique element of $\mathsf{Types}([a],[a])$ which contains $x = y$, we call this the *identity type on* $[a]$, it is denoted by $\mathsf{id}(a)$, and it is symmetric. A *non-identity* type is a type which is not an identity type.

We begin the definition of $\mathfrak{N}$. A group $\mathfrak{G} = \langle G, + \rangle$ will be used in the definition of $\mathfrak{N}$. That is, G is the universe of the group, and $+$ is its group-operation. The identity element (or zero-element) of $+$ is denoted by 0, and the inverse of an element a is denoted by $-a$. We call a group *asymmetric* when its zero-element is its only element of order 2, i.e., when $x + x = 0$ implies $x = 0$ in it. Let $\mathfrak{G}$ be any commutative asymmetric group of size at least twice that of $\mathsf{Types}([a],[b])$ for any $a,b \in M$, i.e.,

$$|G| \geq 2 \cdot |\mathsf{Types}([a],[b])| \quad \text{for all } a,b \in M. \tag{g1}$$

There is such a $\mathfrak{G}$ because for each odd number n the group $\mathbb{Z}_n$ of integers smaller than n and with addition modulo n is asymmetric, and then one can construct an asymmetric group of any infinite size by taking an elementary submodel of a sufficiently big ultraproduct of the $\mathbb{Z}_n$ with odd n. Equivalently, use the upward Löwenheim–Skolem–Tarski Theorem [9, Corollaries 2.1.5 and 2.1.6].

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

The universe N of $\mathfrak{N}$ is defined as

$$N = \{([a],g) : a \in M, g \in G\}.$$

Clearly, if M is finite, then G can be chosen to be finite, and then N is finite.

For the definition of the relations of $\mathfrak{N}$, let $\lambda = \langle \lambda_{[a],[b]} : a,b \in M \rangle$ be a system of functions mapping G to the types of $\mathfrak{M}$ that satisfies the following conditions for all $a,b \in M$. We will write $\lambda_{a,b}$ in place of $\lambda_{[a],[b]}$, for easier readability.

$$\lambda_{a,b} : G \to \mathsf{Types}([a],[b]) \text{ is surjective,}$$
$$\lambda_{a,b}(g) = \lambda_{b,a}(-g)^{\smile} \quad \text{and} \quad \lambda_{a,a}(0) = \mathsf{id}(a).$$

There is such a system λ of functions, because of the following. When $[a] \neq [b]$, take any surjective $\lambda_{b,a} : G \to \mathsf{Types}([b],[a])$, there is such since $|G| \geq |\mathsf{Types}([b],[a])|$ by (g1), then define $\lambda_{a,b}(g) = \lambda_{b,a}(-g)^{\smile}$, and this is also surjective because $\mathsf{Types}([a],[b]) = \{-t^{\smile} : t \in \mathsf{Types}([b],[a])\}$.

Assume now $[a] = [b]$. Since G is asymmetric, there is a set $P \subseteq G$ such that P, $-P = \{-g : g \in P\}$ together with $\{0\}$ form a partition of G. Let S denote the set of all non-identity symmetric types in $T = \mathsf{Types}([a],[a])$, then there is a set $A \subseteq T$ such that A, $A^{\smile} = \{t^{\smile} : t \in A\}$, and S together with $\{\mathsf{id}(a)\}$ form a partition of T. Now, $|P| \geq |S \cup A|$ by (g1), so there is a surjective function $L : P \to (S \cup A)$. Define now $\lambda_{a,a}(g) = L(g)$ for $g \in P$, $\lambda_{a,a}(g) = L(-g)^{\smile}$ for $g \in -P$, and $\lambda_{a,a}(0) = \mathsf{id}(a)$. This function satisfies the required conditions (because both functions of taking inverse in the group and taking converse in the types are their own inverses).

Now we define, for all $a,b \in M$ and $g,h \in G$

$$\mathsf{ty}(([a],g),([b],h)) = \lambda_{a,b}(h-g),$$

where $h-g$ denotes $h+-g$ as usual in group theory. Then ty maps $N \times N$ to the types of $\mathfrak{M}$. Let R be an arbitrary binary relation symbol in the language of $\mathfrak{M}$. The binary relation $R^{\mathfrak{N}}$ belonging to R in $\mathfrak{N}$ is defined as

$$R^{\mathfrak{N}} = \{(p,q) \in N \times N : \mathsf{R}(x,y) \in \mathsf{ty}(p,q)\}.$$

By this, the model $\mathfrak{N}$ has been defined.

Next, we exhibit a 2-isomorphism between $\mathfrak{M}$ and $\mathfrak{N}$. We define $I \subseteq (M \times N) \cup (M^2 \times N^2)$ by requiring for all $a,b \in M$ and $p,q \in N$ that

$$\langle a,p \rangle \in I \quad \text{iff} \quad p = ([a],g) \text{ for some } g,$$
$$\langle (a,b),(p,q) \rangle \in I \quad \text{iff} \quad [a,b] = \mathsf{ty}(p,q).$$

We now show that I is a 2-isomorphism between $\mathfrak{M}$ and $\mathfrak{N}$. From the conditions defining a 2-isomorphism, I clearly satisfies (i) by its very definition.

Next we show that the restriction property (iii) holds for I. Assume that $\langle (a,b),(p,q) \rangle \in I$. This means that $[a,b] = \mathsf{ty}(p,q)$. Assume that $p = ([c],g)$ and $q = ([d],h)$. From the definition of $\mathsf{ty}(p,q)$ it is clear that $\mathsf{ty}(p,q) \in \mathsf{Types}([c],[d])$. Thus, $[a,b] = [r,s]$ for some $r \in [c]$ and $s \in [d]$. But this implies that $[a] = [r] = [c]$ and $[b] = [s] = [d]$, hence $\langle a,p \rangle \in I$ and $\langle b,q \rangle \in I$.

We show that I satisfies local isomorphism property (ii). Let R be an arbitrary relation symbol in the language of $\mathfrak{M}$. Assume that $\langle a,p \rangle \in I$. Then $p = ([a],g)$ for some g by the definition of I, and we have to show that $R(p,p)$ holds in $\mathfrak{N}$ iff $R(a,a)$

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

holds in $\mathfrak{M}$. By definition, $R(p,p)$ holds in $\mathfrak{N}$ iff $\mathsf{R}(x,y) \in \mathsf{ty}(p,p) = \lambda_{a,a}(g-g) = \lambda_{a,a}(0) = \mathsf{id}(a)$, and $\mathsf{R}(x,y) \in \mathsf{id}(a)$ iff $R(a,a)$ holds in $\mathfrak{M}$, by the definition of $\mathsf{id}(a)$.

Assume that $\langle (a,b),(p,q)\rangle \in I$. Then $[a,b] = \mathsf{ty}(p,q)$ by the definition of I. We have to show that each of $a = b, R(a,a)$, $R(a,b)$, $R(b,a)$, and $R(b,b)$ holds in $\mathfrak{M}$ iff the same holds in $\mathfrak{N}$ for p,q in place of a,b. Assume that $p = ([c],g)$ and $q = ([d],h)$, then $\mathsf{ty}(p,q) = \lambda_{c,d}(h-g)$ by the definition of ty. Now, $a = b$ iff $x = y \in [a,b] = \mathsf{ty}(p,q) = \lambda_{c,d}(h-g)$, and $x = y \in \lambda_{c,d}(h-g)$ iff ($[c] = [d]$ and $h - g = 0$) iff $p = q$. Similarly, $R(a,b)$ iff $\mathsf{R}(x,y) \in [a,b] = \mathsf{ty}(p,q)$ iff $R(p,q)$, by the definition of $\mathfrak{N}$. For the next case, we want to show that

$$\mathsf{ty}(p,q) = \mathsf{ty}(q,p)^{\smile}. \tag{t1}$$

Indeed, $\mathsf{ty}(p,q) = \lambda_{c,d}(h-g) = \lambda_{d,c}(g-h)^{\smile} = \mathsf{ty}(q,p)^{\smile}$, by the second condition that λ has to satisfy. By this, (t1) is proved.

Now, $R(b,a)$ iff $\mathsf{R}(x,y) \in [b,a] = [a,b]^{\smile} = \mathsf{ty}(p,q)^{\smile} = \mathsf{ty}(q,p)$ iff $R(q,p)$. Finally, $R(a,a)$ iff $R(p,p)$ and $R(b,b)$ iff $R(q,q)$ hold by the first case of (ii), since we have already shown the restriction property (iii). Thus, $\langle (a,b),(p,q)\rangle$ indeed specifies a partial isomorphism between $\mathfrak{M}$ restricted to $\{a,b\}$ and $\mathfrak{N}$ restricted to $\{p,q\}$.

We check the back-and-forth property (iv) for I. To check the first part, notice that to any $a \in M$ there is at least one $p = ([a],g)$ in N, because G is nonempty. The second part is clear, since no 1-Type $[a]$ is empty. To check the third and fourth parts of the back-and-forth property, assume that $\langle a,p\rangle \in I$ with $p = ([a],g)$. Let $b \in M$ be arbitrary. We have to find a $q \in N$ such that $[a,b] = \mathsf{ty}(p,q)$. By surjectivity of $\lambda_{a,b}$, there is $f \in G$ with $[a,b] = \lambda_{a,b}(f)$. Let $q = (b, f+g)$. Then $\mathsf{ty}(p,q) = \lambda_{a,b}(f+g-g) = \lambda_{a,b}(f) = [a,b]$ and we are done with the third part. Let now $q = ([b],h) \in N$ be arbitrary. We have to find $c \in M$ such that $[a,c] = \mathsf{ty}(p,q)$. Now, $\mathsf{ty}(p,q) \in \mathsf{Types}([a],[b])$ which means that there are $a' \in [a]$ and $b' \in [b]$ such that $\mathsf{ty}(p,q) = \mathsf{Type}(a',b',\mathfrak{M})$. By $a' \in [a]$ we have $\mathsf{Type}(a',a',\mathfrak{M}) = \mathsf{Type}(a,a,\mathfrak{M})$, so by 2-homogeneity of $\mathfrak{M}$ there is $c \in M$ such that $\mathsf{Type}(a,c,\mathfrak{M}) = \mathsf{Type}(a',b',\mathfrak{M})$ and this shows that $[a,c] = \mathsf{ty}(p,q)$. So, (iv) holds for I. We have seen that I is a 2-isomorphism between $\mathfrak{M}$ and $\mathfrak{N}$.

We show that $\mathfrak{N}$ is transitive. We have to show that if $p,q \in N$ are of the same type in $\mathfrak{N}$, then there is an automorphism of $\mathfrak{N}$ that takes p to q. Assume that $p = ([a],g)$ and $q = ([b],h)$. Then $\langle a,p\rangle \in I$ and $\langle b,q\rangle \in I$, by the definition of I. Since I is a 2-isomorphism between $\mathfrak{M}$ and $\mathfrak{N}$, we get that the same formulas hold in $\mathfrak{N}$ for p as in $\mathfrak{M}$ for a, and the same for q and b. Hence, a and b are of the same type in $\mathfrak{M}$ (since p and q are of the same type in $\mathfrak{N}$). It is easy to check that a and b are of the same type in $\mathfrak{N}$ iff $[a] = [b]$. Let $k = -g + h$ and define $\alpha : N \to N$ by

$$\alpha(([c],f)) = ([c], f+k) \quad \text{for all } c \in M \text{ and } f \in G.$$

Now, $\alpha(p) = q$ by $[a] = [b]$ and $h = g + k$. We show that α is an automorphism of $\mathfrak{N}$. First, α is a permutation of N because $\mathfrak{G}$ is a group. Let R be a binary relation symbol in the language of $\mathfrak{N}$, and let $r = ([c],i) \in N$, $s = ([d],j) \in N$. Now, $R(r,s)$ holds in $\mathfrak{N}$ iff $\mathsf{R}(x,y) \in \mathsf{ty}(r,s) = \lambda_{c,d}(j-i)$. Similarly, $R(\alpha(r),\alpha(s))$ holds in $\mathfrak{N}$ iff $\mathsf{R}(x,y) \in \mathsf{ty}(\alpha(r),\alpha(s)) = \lambda_{c,d}(j+k-(i+k))$. However, $j-i = j+k-(i+k)$ because $\mathfrak{G}$ is commutative. Thus, α is indeed an automorphism, and we are done with showing that $\mathfrak{N}$ is transitive.

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

We have seen that $\mathfrak{N}$ is transitive, and this finishes the proof of one direction of Theorem 1.

The other direction of Theorem 1, necessity of 2-homogeneity, follows from the facts that a transitive model is always 2-homogeneous, and 2-isomorphisms preserve being 2-homogeneous. In more detail, assume that I is a 2-isomorphism between $\mathfrak{M}$ and the transitive $\mathfrak{N}$. We have to show that $\mathfrak{M}$ is 2-homogeneous. Let $a,b,c \in M$ be such that $[a] = [b]$. By the back-and-forth property in the definition of a 2-isomorphism, there are $a',b',c' \in N$ such that $\langle a,a'\rangle$, $\langle b,b'\rangle$, $\langle (a,c),(a',c')\rangle$ are all in I. Then $\mathsf{Type}(a',a',\mathfrak{N}) = \mathsf{Type}(b',b',\mathfrak{N})$ since I is a 2-isomorphism and $[a] = [b]$. Since $\mathfrak{N}$ is 2-transitive, there is an automorphism of $\mathfrak{N}$ that takes a' to b'. Let d' be the image of c' under this automorphism. Then $\mathsf{Type}(a',c',\mathfrak{N}) = \mathsf{Type}(b',d',\mathfrak{N})$ since automorphisms preserve 2-types of elements. By the back-and-forth property of I again, there is $d \in M$ such that $\langle (b,d),(b',d')\rangle \in I$. Then $\mathsf{Type}(a,c,\mathfrak{M}) = \mathsf{Type}(a',c',\mathfrak{N}) = \mathsf{Type}(b',d',\mathfrak{N}) = \mathsf{Type}(b,d,\mathfrak{M})$ and we are done. ⊣

Next we state a corollary of Theorem 1.

THEOREM 2. *Each binary model is* 2*-equivalent to a transitive model. Each finite binary model is* 2*-isomorphic to a finite transitive model.*

PROOF. For the definition of an ω-saturated model see, e.g., [9]. First we prove

$$\mathfrak{M} \text{ is } \omega\text{-saturated implies that } \mathfrak{M} \text{ is 2-homogeneous.} \tag{S}$$

Assume that $\mathfrak{M}$ is ω-saturated. Let $a,b,c \in M$ be such that $\mathsf{Type}(a,a,\mathfrak{M}) = \mathsf{Type}(b,b,\mathfrak{M})$. Let $Y = \{b\} \subseteq M$ and let $\Gamma(x) = \{\rho(x,b) \in FO2 : \rho(c,a) \text{ in } \mathfrak{M}\}$. Then $\Gamma(x)$ is a set of formulas in the language of $\langle \mathfrak{M},b\rangle$. We show that it is consistent with the theory of $\langle \mathfrak{M},b\rangle$. Let Δ be a finite subset of $\Gamma(x)$, let $\delta(y)$ denote the formula $\exists x \bigwedge \Delta[b/y]$ that we get from $\exists x \bigwedge \Delta$ by replacing b everywhere with y. Then $\mathfrak{M} \models \delta(a)$ by the definition of $\Gamma(x)$, and so $\mathfrak{M} \models \delta(b)$ since a,b have the same 1-Type in $\mathfrak{M}$. But $\mathfrak{M} \models \delta(b)$ means that $\langle \mathfrak{M},b\rangle \models \exists x \bigwedge \Delta$ that shows that Δ is consistent with the theory of $\langle \mathfrak{M},b\rangle$. Since Δ is an arbitrary finite subset of $\Gamma(x)$, this means that $\Gamma(x)$ is consistent with the theory of $\langle \mathfrak{M},b\rangle$. Since $\mathfrak{M}$ is ω-saturated, then there is $d \in M$ for which $\Gamma(d)$ holds. This means that $\mathsf{Type}(a,c,\mathfrak{M}) = \mathsf{Type}(b,d,\mathfrak{M})$ and we are done with showing (S).

Now, Theorem 2 follows from (S) and Theorem 1 by using that each infinite model is elementarily equivalent—hence 2-equivalent—with an ω-saturated one (see [9, Lemma 5.1.4]), that each finite model is ω-saturated (see [9, Proposition 5.1.2]), and that the model $\mathfrak{N}$ in the proof of Theorem 1 can be constructed to be finite whenever $\mathfrak{M}$ is finite. ⊣

Theorem 3 below serves as a contrast to Theorem 2.

THEOREM 3. *There is a finite binary model that is not* 3*-equivalent to any transitive model.*

PROOF. The binary model $\mathfrak{M}$ has 45 elements and 4 basic relations S,G,R,B. Let $\langle \mathbb{Z}_5,+\rangle$ denote the group of non-negative numbers smaller than 5 with addition modulo 5, and let $9 = \{0,1,\ldots,8\}$ denote the set of non-negative integers smaller than 9. We define

$$M = 5 \times 9.$$

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

Let s,g be permutations of 9 defined, in cycle form, as
$s=(012)(345)(678)$ and $g=(136)(147)(258)$,
and let $r,b\subseteq 9\times 9$ be defined as
$r=\{0,3,6\}\times\{0,1,2\}\cup\{1,4,7\}\times\{3,4,5\}\cup\{2,5,8\}\times\{6,7,8\}$ and
$b=\{0,4,8\}\times\{0,5,7\}\cup\{1,5,6\}\times\{1,3,8\}\cup\{2,3,7\}\times\{2,4,6\}$.

Now, the basic relations of $\mathfrak{M}$ are defined as

$$S=\{\langle(i,j),(i,s(j))\rangle : i\in 5, j\in 9\},$$
$$G=\{\langle(i,j),(i,g(j))\rangle : i\in 5, j\in 9\},$$
$$R=\{\langle(i,j),(i+1,k)\rangle : i\in 5,(j,k)\in r\},$$
$$B=\{\langle(i,j),(i+2,k)\rangle : i\in 5,(j,k)\in b\}.$$

We show that $\mathfrak{M}$ is not 3-equivalent to any transitive model. Let us call a model 3,1-*transitive* when any two elements of the same 3-type can be taken to each other by an automorphism. A transitive model $\mathfrak{N}$ is 3,1-transitive, because let $a,b\in N$ have the same 3-types, then they have the same types in $\mathfrak{N}$, therefore there is an automorphism taking a to b, by transitivity of $\mathfrak{N}$. Thus, it is enough to show that if $\mathfrak{M}$ is 3-equivalent to $\mathfrak{N}$ then $\mathfrak{N}$ is not 3,1-transitive.

Assume that $\mathfrak{M}$ is 3-equivalent to $\mathfrak{N}$. For a first-order formula $\rho(x,y)$ with free variables among x,y let $\rho(\mathfrak{M})$ denote the relation that ρ defines in $\mathfrak{M}$, i.e., $\rho(\mathfrak{M})=\{(a,b):\mathfrak{M}\models\rho[a,b]\}$. Let $\mathfrak{Ra}(\mathfrak{M})$ be the relation algebra of FO3-definable binary relations of $\mathfrak{M}$, i.e., the universe of $\mathfrak{Ra}(\mathfrak{M})$ is $\{\rho(\mathfrak{M}):\rho(x,y)\in FO3\}$, and the operations of $\mathfrak{Ra}(\mathfrak{M})$ are the operations of taking union, converse, and relation composition of binary relations together with the (base-sensitive) operations of taking complement in $M\times M$ and the identity constant $\{(u,u):u\in M\}$ on M. Let $\mathfrak{Ra}(\mathfrak{N})$ denote the similar algebra of FO3-definable binary relations of $\mathfrak{N}$. We show the following:

$$\mathfrak{M}\text{ is 3-equivalent to }\mathfrak{N}\text{ implies that }\mathfrak{Ra}(\mathfrak{M})\text{ is isomorphic to }\mathfrak{Ra}(\mathfrak{N}). \tag{1}$$

Indeed, it is easy to check that the relation $\{(\rho(\mathfrak{M}),\rho(\mathfrak{N})):\rho(x,y)\in FO3\}$ is an isomorphism between $\mathfrak{Ra}(\mathfrak{M})$ and $\mathfrak{Ra}(\mathfrak{N})$ when $\mathfrak{M}$ is 3-equivalent to $\mathfrak{N}$.

A *base-automorphism* of $\mathfrak{Ra}(\mathfrak{N})$ is a permutation α of N that leaves all elements of $\mathfrak{Ra}(\mathfrak{N})$ fixed when taking Z to $\{(\alpha(u),\alpha(v)):(u,v)\in Z\}$. Now, $\mathfrak{Ra}(\mathfrak{N})$ is called *c-permutational* iff any element of N can be taken to any other by a base-automorphism.

$$\mathfrak{N}\text{ is 3,1-transitive implies that }\mathfrak{Ra}(\mathfrak{N})\text{ is c-permutational.} \tag{2}$$

To check (2), notice first that all elements in $\mathfrak{N}$ have the same 3-type. This is so because $\mathfrak{M}$ is such and this property can be expressed with FO3 formulas $\{\exists x\rho(x,x)\to\forall x\rho(x,x):\rho(x,y)\in FO3\}$. Therefore, $\mathfrak{N}$ is 3,1-transitive means that each element of N can be taken to any other element of N by an automorphism of $\mathfrak{N}$. Finally, an automorphism α of $\mathfrak{N}$ is a base-automorphism of $\mathfrak{Ra}(\mathfrak{N})$ and we are done.

From now on, we will use [2].

$$\mathfrak{Ra}(\mathfrak{M})\text{ is the algebra }\mathfrak{A}\text{ defined in [2, section 2].} \tag{3}$$

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

Indeed, it can be checked that the basic relations S, G, R, B of $\mathfrak{M}$ coincide with the relations s, g, r_0, b_0 in [2, Section 2]. It is stated in [2, p. 375, line 16] that $\mathfrak{A}$ is generated by these four elements, so each element of A is FO3-definable in $\mathfrak{M}$. In the other direction, it is a theorem of relation algebra theory that all FO3-definable elements of $\mathfrak{M}$ can be generated from the basic relations of $\mathfrak{M}$ with the operations of $\mathfrak{Ra}(\mathfrak{M})$, see e.g., [32, Section 3.9] or [15, Theorem 3.32]. Thus, the elements of A are exactly the FO3-definable binary relations of $\mathfrak{M}$ and we are done.

In the proof of [2, Theorem 1] it is proved that $\mathfrak{A}$ is not isomorphic to any c-permutational algebra, and so $\mathfrak{N}$ cannot be 3,1-permutational by (1) and (2). The proof of Theorem 3 is complete. ⊣

§3. Two-variable fragment of FO has the weak Beth definability property. We recall the definition of when the two-variable fragment FO2 has the weak Beth definability property (wBDP).

DEFINITION 2 (wBDP for FO2). Let $\mathcal{L}$ be a language with relation symbols of rank 2, and let Th be any set of formulas of $\mathrm{FO2}(\mathcal{L})$, the set of formulas of language $\mathcal{L}$ that contain, bound or free, only the variables x, y. Assume that R is a binary relation symbol not occurring in $\mathcal{L}$, let $\mathcal{L}^+$ denote $\mathcal{L}$ expanded with R. Let $\Sigma(\mathsf{R})$ be a set of formulas of $\mathrm{FO2}(\mathcal{L}^+)$.

(i) We say that $\Sigma(\mathsf{R})$ is a *strong implicit definition of* R *w.r.t.* Th when in each model $\mathfrak{M}$ of Th there is exactly one relation $R \subseteq M \times M$ such that $\langle \mathfrak{M}, R\rangle \models \Sigma(\mathsf{R})$. We say that $\Sigma(\mathsf{R})$ is just a *weak implicit definition of* R *w.r.t.* Th when in each model $\mathfrak{M}$ of Th there is at most one relation R such that $\langle \mathfrak{M}, R\rangle \models \Sigma(\mathsf{R})$. That is, with weak definitions it is allowed that in some models there is no relation at all satisfying the implicit definition.

(ii) We say that $\Sigma(\mathsf{R})$ *can be made explicit w.r.t.* Th, or that R *has an explicit definition over* Th when there is a formula $\varphi \in FO2(\mathcal{L})$ such that $\mathsf{Th} \cup \Sigma(\mathsf{R}) \models \forall x, y(\mathsf{R}(x, y) \leftrightarrow \varphi)$. In this case, we say that φ is an explicit definition of R in Th.

(iii) Now, *FO2 has the weak Beth definability property* means that each strong implicit definition of FO2 can be made explicit.

It is proved in [1] that there is a weak implicit definition in FO2 that cannot be made explicit. The question is whether there is also a strong implicit definition that cannot be made explicit. The following theorem gives a negative answer to this.

THEOREM 4. *FO2 has the weak Beth definability property.*

The proof of Theorem 4 uses the following two lemmas. We say that $R \subseteq M \times M$ *does not cut* 2*-types in* $\mathfrak{M}$ if for all $a, b, c, d \in M$ we have $(R(a, b)$ iff $R(c, d))$ whenever the 2-type of (a, b) is the same as that of (c, d) in $\mathfrak{M}$.

LEMMA 1 (No-cut lemma). *Assume that* Σ *is a weak implicit definition of* R *in* Th *and* $\mathsf{Th} \cup \Sigma$ *consists of FO2-formulas. Assume that* $\mathfrak{M}$ *is transitive and* $\langle \mathfrak{M}, R\rangle \models \mathsf{Th} \cup \Sigma$. *Then* R *does not cut* 2*-types in* $\mathfrak{M}$.

PROOF OF LEMMA 1. Let $\mathsf{Th}, \Sigma, \mathfrak{M}, R$ be as in the first two sentences of the statement of Lemma 1. Assume that R cuts a 2-type in $\mathfrak{M}$. By using this, we are going to define

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

a relation S distinct from R which also satisfies Σ in $\mathfrak{M}$, contradicting that Σ is an implicit definition.

Assume that

$$R(a,b) \text{ and not } R(c,d) \text{ in } \mathfrak{M} \tag{a}$$

for some a,b,c,d such that $\mathsf{Type}(a,b,\mathfrak{M}) = \mathsf{Type}(c,d,\mathfrak{M})$. Let $T = \mathsf{Type}(a,b,\mathfrak{M})$ and

$$t = \{(m,n) \in M \times M : \mathsf{Type}(m,n,\mathfrak{M}) = T\}.$$

We note that

$$T \text{ is not an identity type} \tag{d}$$

by (a) and $\mathfrak{M}$ being transitive: assume $(m,m) \in t$ for some $m \in M$, then $(a,a),(c,c) \in t$ and there is an automorphism α of $\mathfrak{M}$ taking a to c. Hence $(a,a) \in R$ iff $(c,c) \in R$ because automorphisms preserve meanings of formulas and hence they leave solutions of implicit definitions fixed. However, $a = b$ and $c = d$ by t being an identity type, contradicting (a).

For a binary relation Z, let $Z^{-1} = \{(v,u) : (u,v) \in Z\}$ denote its inverse, Z is symmetric means that $Z = Z^{-1}$. We define $S \subseteq M \times M$ as follows. If $T \neq T^{\smile}$ or $t \cap R$ is symmetric, then we define S by "interchanging" $t \cap R$ with $t \setminus R$, i.e.,

$$S = (R \setminus t) \cup (t \setminus R).$$

If $T = T^{\smile}$ and $t \cap R$ is not symmetric then S is defined by "interchanging" $(t \cap R \setminus R^{-1})$ with its inverse $(t \cap R^{-1} \setminus R)$, i.e.,

$$S = (R \setminus [t \cap R \setminus R^{-1}]) \cup (t \cap R^{-1} \setminus R).$$

Then S is distinct from R by $t \cap R$ being nonempty in the first case, and by $(t \cap R \setminus R^{-1}) \cup (t \cap R^{-1} \setminus R)$ being nonempty in the second case. However,

$$S \text{ and } R \text{ differ only inside } t, \tag{s}$$

that is, $[R(m,n)$ iff $S(m,n)]$ for all $(m,n) \in M \times M \setminus t$. We are going to show that $\langle \mathfrak{M}, S\rangle \models \Sigma$. We define $J \subseteq (M \times M) \cup (M^2 \times M^2)$ by requiring for all $m,n,p,q \in M$ that

$$\begin{aligned}
&\langle m,p\rangle \in J \quad \text{iff} \quad \mathsf{Type}(m,m,\mathfrak{M}) = \mathsf{Type}(p,p,\mathfrak{M}),\\
&\langle (m,n),(p,q)\rangle \in J \quad \text{iff} \quad \mathsf{Type}(m,n,\mathfrak{M}) = \mathsf{Type}(p,q,\mathfrak{M}) \text{ and}\\
&\quad [(m,n) \in R \leftrightarrow (p,q) \in S)] \text{ and } [(n,m) \in R \leftrightarrow (q,p) \in S)].
\end{aligned}$$

We now show that J is a 2-isomorphism between $\langle \mathfrak{M}, R\rangle$ and $\langle \mathfrak{M}, S\rangle$. We check properties (i)–(iv) in the definition of a 2-isomorphism. (i) is satisfied by the definition of J. Restriction property (iii) is satisfied, because if $\mathsf{Type}(a,b,\mathfrak{M}) = \mathsf{Type}(p,q,\mathfrak{M})$ then $\mathsf{Type}(a,a,\mathfrak{M}) = \mathsf{Type}(b,b,\mathfrak{M})$ and $\mathsf{Type}(p,p,\mathfrak{M}) = \mathsf{Type}(q,q,\mathfrak{M})$.

Checking local isomorphism property (ii): Assume $\langle (m,n),(p,q)\rangle \in J$. Then J is a local isomorphism with respect to the language of $\mathfrak{M}$ by $\mathsf{Type}(m,n,\mathfrak{M}) = \mathsf{Type}(p,q,\mathfrak{M})$. We now check local isomorphism with respect to the new relation symbol R. Indeed, $R(m,m)$ iff $S(p,p)$ holds because $R(m,m)$ iff $R(p,p)$ by the restriction property, and $R(p,p)$ iff $S(p,p)$ by (s) and (d). We have $R(m,n)$ iff $S(p,q)$ and $R(n,m)$ iff $S(q,p)$ by the definition of J. Thus property (ii) holds.

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

Checking back-and-forth property (iv): The first line is satisfied by $\langle a,a\rangle \in J$ for all $a \in M$. For checking the second line, let $m,n,m' \in M$, $\langle m,n\rangle \in J$. Let α be an automorphism of $\mathfrak{M}$ that takes m to n. There is such an α because $\langle m,n\rangle \in J$ and $\mathfrak{M}$ is transitive.

Assume that $\mathsf{Type}(m,m',\mathfrak{M}) \notin \{T,T^\smile\}$. Let $n' = \alpha(m')$. Then $\mathsf{Type}(n,n',\mathfrak{M}) = \mathsf{Type}(m,m',\mathfrak{M}) \notin \{T,T^\smile\}$, because automorphisms do not change 2-types. By (s) then $R(e,f)$ iff $S(e,f)$ for all (e,f) in $\{(m,m'),(m',m),(n,n'),(n',n)\}$, thus $\langle(m,m'),(n,n')\rangle \in J$ by the definition of J.

Assume now that $\mathsf{Type}(m,m',\mathfrak{M}) = T$ and $T \neq T^\smile$ or $t \cap R$ is symmetric. If $R(m,m')$, then let β be an automorphism that takes c to n and let $n' = \beta(d)$. Then $(n,n') \in (t \setminus R)$ by (a), so $(n,n') \in S$ by the definition of S. Assume $T \neq T^\smile$, then $(m',m) \in R$ iff $(n',n) \in S$ by (s). Assume that $T = T^\smile$ and $t \cap R$ is symmetric. Then $t \cap R^{-1}$ is also symmetric, hence $(m',m) \in R$ and $(n',n) \notin R$, so $(n',n) \in S$. Thus $\langle(m,m'),(n,n')\rangle \in J$. If not $R(m,m')$, then let β be an automorphism that takes a to n and let $n' = \beta(b)$. From here on, the argument showing $\langle(m,m'),(n,n')\rangle \in J$ is completely analogous to the previous case.

Assume now that $\mathsf{Type}(m,m',\mathfrak{M}) = T = T^\smile$ and $t \cap R$ is not symmetric, say, $(e,f) \in t \cap R \setminus R^{-1}$. If $(m,m') \in (R \cap R^{-1})$ then let $n' = \alpha(m')$. Then $(n,n') \in (R \cap R^{-1})$ by $\alpha(R) = R$ and so $\alpha(R^{-1}) = R^{-1}$. Also, $(n,n') \in S \cap S^{-1}$ in this case, by the definition of S. Hence $\langle(m,m'),(n,n')\rangle \in J$. The case $(m,m') \notin (R \cup R^{-1})$ is completely analogous. Assume $(m,m') \in R \setminus R^{-1}$. Let β be an automorphism taking f to n, and let $n' = \beta(e)$. Then $(n,n') \in R^{-1} \setminus R$, so $(n,n') \in S \setminus S^{-1}$ by the definition of S, and so $\langle(m,m'),(n,n')\rangle \in J$ by the definition of J. The case $(m,m') \in R^{-1} \setminus R$ is completely analogous.

The case when $\mathsf{Type}(m,m',\mathfrak{M}) = T^\smile$ can be proved in a completely analogous way. By this, checking the second line is finished. The third line of (iv) follows in our case from the second one by noticing that J is symmetric both in $M \times M$ and in $M^2 \times M^2$.

Thus J is a 2-isomorphism between $\langle\mathfrak{M},R\rangle$ and $\langle\mathfrak{M},S\rangle$, so $\langle\mathfrak{M},S\rangle \models \Sigma$. Thus both R and the distinct S satisfy Σ, this contradicts Σ being an implicit definition. The proof of Lemma 1 is complete. ⊣

The next lemma is a kind of characterization of those implicit definitions that cannot be made explicit.

LEMMA 2 (Cut lemma). *Assume that* Σ *is a weak implicit definition of* R *in* Th, *and* $\mathsf{Th} \cup \Sigma$ *consists of FO2 formulas. Statements* (*i*)–(*iii*) *below are equivalent*.

(i) Σ *can be made explicit in* Th *by an FO2-formula*.
(ii) R *does not cut 2-types in* $\mathfrak{M}$ *whenever* $\langle\mathfrak{M},R\rangle \models \mathsf{Th} \cup \Sigma$.
(iii) R *does not cut 2-types in* $\mathfrak{M}$ *whenever* $\langle\mathfrak{M},R\rangle \models \mathsf{Th} \cup \Sigma$ *and* $\mathfrak{M}$ *is 2-homogeneous*.

PROOF OF LEMMA 2. Clearly, (i) implies (ii) and (ii) implies (iii). To show that (iii) implies (i), assume that (i) does not hold, we will infer the negation of (iii).

We will refer to the negation of (i) just as "R is not 2-definable". Thus we want to show that there is a model $\langle\mathfrak{M},R\rangle \models \mathsf{Th} \cup \Sigma$ such that $\mathfrak{M}$ is 2-homogeneous and R cuts a 2-type in $\mathfrak{M}$.

By a 2-*partition* we understand a system $\langle\pi_i : i \leq n\rangle$ of FO2-formulas in the language of Th such that $\mathsf{Th} \cup \Sigma \models (\bigvee\{\pi_i : i \leq n\} \wedge \bigwedge\{\neg(\pi_i \wedge \pi_j) : i < j \leq n\}$. We

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

say that R *cannot cut* π when $\mathsf{Th} \cup \Sigma \models \forall x,y(\pi(x,y) \to \mathsf{R}(x,y)) \vee (\forall x,y(\pi(x,y) \to \neg\mathsf{R}(x,y))$. We say that R *cannot cut into the* 2*-partition* $\langle \pi_i : i \leq n \rangle$ when R cannot cut any π_i for $i \leq n$. We will use the following statement

$$\mathsf{R} \text{ is 2-definable} \quad \text{iff} \quad \text{there is a 2-partition R cannot cut into.} \tag{L1}$$

Indeed, if R is definable by the FO2-formula $\rho(x,y)$, then R cannot cut into $\langle \rho, \neg\rho \rangle$. In the other direction, assume that R cannot cut into $\langle \pi_i : i \leq n \rangle$. Let us treat natural numbers in von Neumann's sense, i.e., each natural number n is the set of smaller natural numbers: $n = \{i \in \omega : i < n\}$. For $J \subseteq n$ let $\pi(J) = \bigwedge\{\pi_j : j \in J\} \wedge \bigwedge\{\neg\pi_j : j \in n \setminus J\}$. By the assumption that R cannot cut into $\langle \pi_i : i < n \rangle$ we have that R is a union of some π_is in each model of $\mathsf{Th} \cup \Sigma$, i.e., $\mathsf{Th} \cup \Sigma \models \bigvee\{\mathsf{R}(x,y) \leftrightarrow \pi(J) : J \subseteq n\}$. Since Σ is an implicit definition, we have $\mathsf{Th} \cup \Sigma(\mathsf{R}) \cup \Sigma(\mathsf{R}') \models \mathsf{R}(x,y) \leftrightarrow \mathsf{R}'(x,y)$ where R' is a brand new binary relation symbol, so by compactness

$$\mathsf{Th} \cup \Sigma_0(\mathsf{R}) \cup \Sigma_0(\mathsf{R}') \models \mathsf{R}(x,y) \leftrightarrow \mathsf{R}'(x,y) \quad \text{for some finite } \Sigma_0 \subseteq \Sigma. \tag{s}$$

Let $\sigma(\mathsf{R}) = \bigwedge \Sigma_0(\mathsf{R})$ for a Σ_0 satisfying (s). Assume that $\langle \mathfrak{M}, R \rangle \models \mathsf{Th} \cup \Sigma$. Then there is a unique J such that $\mathfrak{M} \models \mathsf{R}(x,y) \leftrightarrow \pi(J)$, since $\pi(J) \wedge \pi(K)$ is inconsistent for distinct J and K. By $\langle \mathfrak{M}, R \rangle \models \Sigma$ and $\Sigma_0 \subseteq \Sigma$ we have that $\mathfrak{M} \models \sigma(\pi(J))$. Also, $\mathfrak{M} \models \sigma(\pi(K))$ is not true for $K \neq J$ by (s) since $\pi(J)$ and $\pi(K)$ define distinct relations in $\mathfrak{M}$. This shows that $\mathsf{Th} \cup \Sigma \models \mathsf{R}(x,y) \leftrightarrow \bigwedge\{\sigma(\pi(J)) \to \pi(J) : J \subseteq n\}$, and this is a 2-definition for R. Statement (L1) has been proved.

To prove (iii), we construct a 2-type T in the language of Th such that the set

$$\mathsf{Th} \cup \Sigma \cup \{\mathsf{R}(x,y), \neg\mathsf{R}(z,v)\} \cup T(x,y) \cup T(z,v) \tag{L2}$$

of FO-formulas is consistent. That T is a 2-type means that either $\rho \in T$ or $\neg\rho \in T$ for all FO2-formulas ρ in the language of Th. We say that a set of open FO-formulas is consistent when there is a model and an evaluation that make the set true. Equivalently, one can consider the free variables in the set to be constants, we will use this second option.

Let τ be an FO2-formula in the language of Th. A 2-*partition of* τ is a system $\langle \pi_i : i \leq n \rangle$ such that $\mathsf{Th} \cup \Sigma \models (\tau \leftrightarrow \bigvee\{\pi_i : i < n\}) \wedge \bigwedge\{\neg(\pi_i \wedge \pi_j) : i < j < n\}$. Let T be a set of FO2 formulas in the language of Th. We say that T *is good* iff R can cut into any 2-partition of $\bigwedge T_0$, for all finite subsets T_0 of T. Clearly, a directed union of good sets is a good set again, so there is a maximal one among the good sets by Zorn's lemma. We will show that any maximal good set is a 2-type and that the set in (L2) with any good T is consistent. We begin with this second statement.

Assume that T is good and let $T_0 \subseteq T$ be finite. Then R can cut into any 2-partition of $\bigwedge T_0$, in particular R can cut into $\bigwedge T_0$. This means that there is a model of $\mathsf{Th} \cup \Sigma \cup \{\mathsf{R}(x,y), \neg\mathsf{R}(z,v)\} \cup T_0(x,y) \cup T_0(z,v)$. By compactness, the set in (L2) is consistent.

To show that a maximal good T is a 2-type, let T be any good set and let ρ be any FO2 formula in the language of Th. We show that either $T \cup \{\rho\}$ is good or $T \cup \{\neg\rho\}$ is good. Assume that neither of $T \cup \{\rho\}$ and $T \cup \{\neg\rho\}$ is good. Then there are finite subsets T_0, T_1 of T and 2-partitions $\pi = \langle \pi_i : i \leq n \rangle$ of $\rho \wedge \bigwedge T_0$ and $\delta = \langle \delta_j : j < m \rangle$ of $\neg\rho \wedge \bigwedge T_1$ such that R cannot cut into either of these two partitions. We can now combine π and δ to form a 2-partition σ of $\bigwedge(T_0 \cup T_1)$ by letting the members of the partition σ be $\pi_i \wedge \bigwedge T_1$ and $\delta_j \wedge \bigwedge T_0$ for $i < n, j < m$. Clearly, R cannot cut

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

into σ by our assumption that R cannot cut into either of π and δ; this contradicts to T being good. With this, we have proved that any maximal good T is a 2-type.

By the above, we now have a 2-type T such that the set $\Delta = \mathsf{Th} \cup \Sigma \cup \{\mathsf{R}(x,y), \neg\mathsf{R}(z,v)\} \cup T(x,y) \cup T(z,v)$ is consistent. Let then $\langle \mathfrak{M}, R\rangle$ be any ω-saturated model of Δ. Then $\mathfrak{M}$ is also ω-saturated, and so it is 2-homogeneous by statement (S) in the proof of Theorem 2. Also, R cuts the 2-type T in $\mathfrak{M}$ by $\langle \mathfrak{M}, R\rangle \models \Delta$. We derived the negation of (iii) from the negation of (i), and this finishes the proof of Lemma 2. ⊣

Proof of Theorem 4. Assume that Σ is a strong implicit definition of R w.r.t. Th. We are going to show that Σ can be made explicit, i.e., R has an explicit definition over Th that uses only two variables.

Take any 2-homogeneous model $\mathfrak{M}$ of Th, and let $\overline{\mathfrak{M}}$ be a transitive model with I a 2-isomorphism between $\mathfrak{M}$ and $\overline{\mathfrak{M}}$. There are such $\overline{\mathfrak{M}}$ and I by Theorem 1. $\overline{\mathfrak{M}}$ is a model of Th because 2-isomorphic models satisfy the same FO2-formulas. Since Σ is a *strong* implicit definition of R in Th, there is $\overline{R}$ which satisfies Σ in $\overline{\mathfrak{M}}$, i.e., $\langle \overline{\mathfrak{M}}, \overline{R}\rangle \models \Sigma$.

Since $\overline{\mathfrak{M}}$ is transitive, $\overline{R}$ does not cut 2-types in $\overline{\mathfrak{M}}$, by Lemma 1. We now "transfer" $\overline{R}$ to the model $\mathfrak{M}$ by the following definition: take any pair (a,b) in $\mathfrak{M}$. We define R so that this pair is related by R if and only if there is a pair of the same type in $\overline{\mathfrak{M}}$ which is related by $\overline{R}$. Formally:

$$R = \{(a,b) \in M \times M : \exists c,d[\overline{R}(c,d) \text{ and } \mathsf{Type}(a,b,\mathfrak{M}) = \mathsf{Type}(c,d,\overline{\mathfrak{M}})\}.$$

We now show that with this definition, the 2-isomorphism I between $\mathfrak{M}$ and $\overline{\mathfrak{M}}$ remains a 2-isomorphism between $\langle \mathfrak{M}, R\rangle$ and $\langle \overline{\mathfrak{M}}, \overline{R}\rangle$. Since I is a 2-isomorphism between $\mathfrak{M}$ and $\overline{\mathfrak{M}}$, it satisfies conditions (i), (iii), and (iv) in the definition of a 2-isomorphism, and it also satisfies condition (ii) for atomic formulas other than $R(v,z)$. Therefore, we only have to show that if $\langle (a,b),(a',b')\rangle \in I$, then $R(a,b)$ iff $\overline{R}(a',b')$.

Assume that $\langle (a,b),(a',b')\rangle \in I$. If $R(a,b)$, then there are $c,d \in \overline{\mathfrak{M}}$ such that $\overline{R}(c,d)$ and $\mathsf{Type}(c,d,\overline{\mathfrak{M}}) = \mathsf{Type}(a,b,\mathfrak{M})$, by the definition of R. The 2-type of (a',b') is also the same as that of (a,b), since they are I-related by assumption. Hence the 2-type of (c,d) is the same as that of (a',b') (since they both equal the 2-type of (a,b)). By $\overline{R}(c,d)$ we now get $\overline{R}(a',b')$, since we have seen that $\overline{R}$ does not distinguish elements of the same 2-type. In the other direction, assume that $\overline{R}(a',b')$. Since (a,b) is I-related to (a',b'), their 2-types equal, hence $R(a,b)$ by the definition of R. By this, we have seen that I is a 2-isomorphism between the expanded models $\langle \mathfrak{M}, R\rangle$ and $\langle \overline{\mathfrak{M}}, \overline{R}\rangle$. Since the latter is a model of Σ, we get that $\langle \mathfrak{M}, R\rangle$ is a model of Σ, too. By its definition, R does not cut 2-types in $\mathfrak{M}$.

Since Σ is a weak definition over Th and the 2-homogeneous $\mathfrak{M} \models \mathsf{Th}$ was chosen arbitrarily, we get that R does not cut 2-types in $\mathfrak{M}$ whenever $\langle \mathfrak{M}, R\rangle \models \mathsf{Th} \cup \Sigma$ and $\mathfrak{M}$ is 2-homogeneous. Thus, Σ can be made explicit in Th, by Lemma 2 and this finishes the proof of Theorem 4. ⊣

§4. On some connections with algebra and the literature.

(1) There is an underlying algebraic intuition behind the proofs in the paper. For example, one may wonder about the role of 2-homogeneity that shows up in

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

Theorem 1. For any model of any logic, the concepts (that is, explicitly defined notions) form a natural algebra with the connectives of the logic as operations, this is called the *concept algebra* of the model. The "types" of a model form a similar natural algebra only when some kind of homogeneity is satisfied. In the case of FO2, this condition is 2-homogeneity. The algebra of types is an atomic superalgebra of the concept algebra. A 2-partial isomorphism between binary models induces an isomorphism between the respective algebras of types, and vice versa, and any isomorphism between algebras of types induces a 2-isomorphism between the respective models. All this is part of a general approach to the "algebra behind logic", for more on this see [4, Part II].

(2) A bridge between logic and algebra is elaborated in, e.g., [8], [14, Section 4.3], or [4, Part II]. In this bridge, a class Alg(L) of algebras is associated with any (decent) logic L. Namely, Alg(L) is the infinitary quasi-equational hull of the class CA(L) of all concept algebras of models of logic. This correspondence is used for stating equivalence theorems of the kind: L has logical property LP if and only if Alg(L) has algebraic property AP. A satisfying fact is that, usually, to natural logical properties natural algebraic properties correspond this way (see the first sentence of [27]).

(3) To Beth definability BDP of a logic the corresponding algebraic property is surjectivity of epimorphisms in Alg(L) considered as a category (ES), see [14, Theorem 5.6] and [4, Theorem 6.11(i)]. Indeed, failure of the BDP for n-variable logics was proved first via showing that ES fails in the category of their concept algebras, see [1]. Failure of BDP for equality-free 2-variable logic is proved also in algebraic form in [29].

(4) Weak Beth definability wBDP possesses several natural algebraic equivalent properties, see [4, Theorems 6.11(iii) and 6.12] and [17]. Three of the corresponding algebraic properties are (i) "surjectivity of CA(L)-extendible epimorphisms", (ii) "Alg(L) is the smallest full reflective subcategory containing MA(L)", and (iii) "MA(L) generates Alg(L) by taking limits of diagrams of algebras". Here, MA(L) is the class of maximal (with respect to containment) members of CA(L).

(5) Craig Interpolation Property CIP and Beth definability property BDP are related, for example, BDP is often proved from CIP. The interpolation properties also have several variants that coincide in the case of FO while happen to be distinct in other logics, e.g., in some modal logic. In the spirit of the bridge theorems, to some interpolation property of a logic L some kind of amalgamation property of Alg(L) corresponds, see, e.g., [4, Theorem 6.15]. A landmark paper on amalgamation and related properties in algebraic versions of FO is [27].

(6) The algebraic equivalent of CIP is strong amalgamation property SAP. The superamalgamation property SUPAP of Alg(L) is proved to correspond to a stronger version sCIP of Craig interpolation property in [23, 25]. The question whether they coincide for variants of FO is asked in [27]. This question is answered in [28], where a variant of n-variable logic, for any finite $n \geq 3$, is constructed that has CIP but not sCIP. This is an analogous result to the one in the present paper concerning Craig Interpolation Property in place of Beth

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press

definability property. We mention that all of the other questions in the two diagrams in [27] have been answered in the meantime, see [22].

(7) There is an extended literature connecting logic and algebra. A tiny sample is [10, 11, 13, 20, 21, 31].

In computer science, n-variable logic usually is used in a stronger version where it is endowed with infinite conjunctions and disjunctions, see [16]. Let $L^n_{\infty,\omega}$ denote this logic as in [16]. Failure of wBDP for $L^n_{\infty,\omega}$ in the case of $n \geq 3$ is proved in [16], but it also follows from [3] (since there the proof is based on finite counterexamples), and failure of BDP for $L^2_{\infty,\omega}$ follows from the proof in [1].

QUESTION 1. *Does $L^2_{\infty,\omega}$ have weak Beth definability property*?

Acknowledgments. We thank Zalán Gyenis and Gábor Sági for enjoyable (transitive) discussions on the subject. We also thank the referee for the many useful suggestions. We dedicate this paper to Harvey Friedman in respect for his work.

REFERENCES

[1] H. ANDRÉKA, S. D. COMER, J. X. MADARÁSZ, I. NÉMETI, and T. SAYED-AHMED, *Epimorphisms in cylindric algebras and definability in finite variable logic*. ***Algebra Universalis***, vol. 61 (2009), no. 3–4, pp. 261–282.

[2] H. ANDRÉKA, I. DÜNTSCH, and I. NÉMETI, *A nonpermutational integral relation algebra*. ***Michigan Mathematical Journal***, vol. 39 (1992), pp. 371–384.

[3] H. ANDRÉKA and I. NÉMETI, *Finite-variable logics do not have weak Beth definability property*, ***The Road to Universal Logic (Festschrift for the 50th Birthday of Jean-Yves Beziau, Vol II)*** (A. Koslow and A. Buchsbaum, editors), Studies in Universal Logic, Birkhäuser, Basel, 2015, pp. 125–133, Chapter 4.

[4] H. ANDRÉKA, I. NÉMETI, and I. SAIN, *Algebraic logic*, ***Handbook of Philosophical Logic, vol. 2***, second ed. (D. M. Gabbay and F. Guenthner, editors), Kluwer Academic, Dordrecht, 2001, pp. 133–247.

[5] V. BARANY, M. BENEDIKT, and B. T. CATE, *Some model theory of guarded negation*, this JOURNAL, vol. 83 (2018), no. 4, pp. 1307–1344.

[6] J. BARWISE and S. FEFERMAN (eds.), ***Model-Theoretic Logics***, Springer-Verlag, New York, 1985.

[7] J. F. A. K. BENTHEM, ***Language in Action: Categories, Lamdas and Dynamic Logic***, Elsevier, Amsterdam, 1991.

[8] W. J. BLOK and D. PIGOZZI, *Algebraizable logics*. ***Memoirs of the American Mathematical Society***, vol. 77 (1989), p. 396.

[9] C. C. CHANG and H. J. KEISLER, ***Model Theory***, third ed., North-Holland, Amsterdam, 1990, first published in 1973.

[10] S. D. COMER, *Galois-theory of cylindric algebras and its applications*. ***Transactions of the American Mathematical Society***, vol. 286 (1984), pp. 771–785.

[11] M. FERENCZI, *Probabilities defined on standard and non-standard cylindric set algebras*. ***Synthese***, vol. 192 (2015), no. 7, pp. 2025–2033.

[12] H. FRIEDMAN, *Beth's theorem in cardinality logics*. ***Israel Journal of Mathematics***, vol. 14 (1973), pp. 205–212.

[13] Z. GYENIS, *On atomicity of free algebras in certain cylindric-like varieties*. ***Logic Journal of the IGPL***, vol. 19 (2011), no. 1, pp. 44–52.

[14] L. HENKIN, J. D. MONK, and A. TARSKI, ***Cylindric Algebras. Parts I–II***, North-Holland, Amsterdam, 1971 and 1985.

[15] R. HIRSCH and I. HODKINSON, ***Relation Algebras by Games***, North-Holland, Amsterdam, 2002.

[16] I. HODKINSON, *Finite variable logics*. ***Bulletin of the European Association for Theoretical Computer Science***, vol. 51 (1993), pp. 111–140, updated version of paper appeared in this journal. With addendum in Vol. 52, http://www.doc.ic.ac.uk/~imh/papers/fvl_revised.pdf, p. 37.

[17] E. Hoogland, *Algebraic characterizations of various Beth definability properties*. ***Studia Logica***, vol. 65 (2000), pp. 91–112.

[18] E. Hoogland, M. Marx, and M. Otto, *Beth definability for the guarded fragment*, ***Logic for Programming and Automated Reasoning. LPAR 1999*** (H. Ganzinger, D. McAllester, and A. Voronkov, editors), Lecture Notes in Computer Science, 1705, Springer, Berlin, 1999, pp. 273–285.

[19] J. Jung and F. Wolter, *Interpolant existence in the guarded fragment*. Semantic Scholar Corpus ID:221018030, 2020.

[20] M. Khaled and T. Sayed Ahmed, *Vaught's theorem holds for L2 but fails for Ln when $n > 2$*. ***Bulletin of the Section of Logic***, vol. 39 (2010), no. 3–4, pp. 107–122.

[21] M. Khaled, G. Székely, K. Lefever, and M. Friend, *Distances between formal theories*. ***The Review of Symbolic Logic***, vol. 13 (2020), no. 3, pp. 633–654.

[22] J. Madarász and T. Sayed Ahmed, *Amalgamation, interpolation and epimorphisms*. ***Algebra Universalis***, vol. 56 (2007), no. 2, pp. 179–210.

[23] J. X. Madarász, *Interpolation and amalgamation; pushing the limits. Parts I and II*. ***Studia Logica***, vol. 61 (1998), no. 3, pp. 311–345 and vol. 62 (1999), no. 1, pp. 1–19.

[24] J. A. Makowsky, S. Shelah, and J. Stavi, *Delta-logics and generalized quantifiers*. ***Annals of Mathematical Logic***, vol. 10 (1976), no. 2, pp. 155–192.

[25] L. Maksimova, *Amalgamation and interpolation in normal modal logics*. ***Studia Logica***, vol. 50 (1991), pp. 457–471.

[26] A. H. Mekler and S. Shelah, *Stationary logic and its friends I*. ***Notre Dame Journal of Formal Logic***, vol. 26 (1985), no. 2, pp. 129–137.

[27] D. Pigozzi, *Amalgamation, congruence extension and interpolation properties in algebras*. ***Algebra Universalis***, vol. 1 (1972), no. 3, pp. 269–349.

[28] G. Sági and S. Shelah, *Weak and strong interpolation for algebraic logics*, this Journal, vol. 71 (2006), pp. 104–118.

[29] I. Sain, *Beth's and Craig's properties via epimorphisms and amalgamation in algebraic logic*, ***Algebraic Logic and Universal Algebra in Computer Science*** (D. H. Bergman, R. D. Maddux, and D. L. Pigozzi, editors), Lecture Notes in Computer Science, vol. 425, Springer Verlag, Berlin, 1990, pp. 209–226.

[30] I. Sain and A. Simon, *Beth properties of finite variable fragments of first order logic*, preprint, 1992, Mathematical Institute of the Hungarian Academy of Sciences, Budapest.

[31] T. Sayed Ahmed, *Varying interpolation and amalgamation for MV polyadic algebras*. ***Journal of Applied Non-Classical Logics***, vol. 25 (2015), no. 2, pp. 140–192.

[32] A. Tarski and S. R. Givant, ***A Formalization of Set Theory Without Variables***, AMS Colloquium Publications, 41, American Mathematical Society, Providence, RI, 1987.

SET THEORY, LOGIC AND TOPOLOGY DEPARTMENT
ALFRÉD RÉNYI INSTITUTE OF MATHEMATICS
BUDAPEST, REÁLTANODA ST. 13–15, H-1053, HUNGARY
E-mail: andreka.hajnal@renyi.hu
E-mail: nemeti.istvan@renyi.hu

https://doi.org/10.1017/jsl.2021.7 Published online by Cambridge University Press