Bisimulation for Neighbourhood Structures

Helle Hvid Hansen[1,2,*] Clemens Kupke[1,†] Eric Pacuit[3,‡]

[1] Centrum voor Wiskunde en Informatica (CWI)

[2] Vrije Universiteit Amsterdam (VUA).

[3] University of Amsterdam (UvA).

August 23, 2007

Abstract

Neighbourhood structures are the standard semantic tool used to reason about non-normal modal logics. In coalgebraic terms, a neighbourhood frame is a coalgebra for the contravariant powerset functor composed with itself, denoted by 2^2. In our paper, we investigate the coalgebraic equivalence notions of 2^2-bisimulation, behavioural equivalence and neighbourhood bisimulation (a notion based on pushouts), with the aim of finding the logically correct notion of equivalence on neighbourhood structures. Our results include relational characterisations for 2^2-bisimulation and neighbourhood bisimulation, and an analogue of Van Benthem's characterisation theorem for all three equivalence notions. We also show that behavioural equivalence gives rise to a Hennessy-Milner theorem, and that this is not the case for the other two equivalence notions.

Keywords: Neighbourhood semantics, non-normal modal logic, bisimulation, behavioural equivalence, invariance.

1 Introduction

Neighbourhood semantics (cf. [7]) forms a generalisation of Kripke semantics, and it has become the standard tool for reasoning about non-normal modal logics in which (Kripke valid) principles such as $\Box p \wedge \Box q \; \rightarrow \; \Box(p \wedge q)$ and $\Box p \rightarrow \Box(p \vee q)$ (mon) are considered not to hold. In a neighbourhood model, each state has associated with it a collection of subsets of the universe (called neighbourhoods), and a modal formula $\Box\phi$ is true at states which have the truth set of ϕ as a neighbourhood. The modal logic of all neighbourhood models is called classical modal logic.

During the past 15-20 years, non-normal modal logics have emerged in the areas of computer science and social choice theory, where system (or agent)

[*] Supported by NWO grant 612.000.316.

[†] Supported by NWO under FOCUS/BRICKS grant 642.000.502.

[‡] Supported by NSF grant OISE 0502312.

properties are formalised in terms of various notions of ability in strategic games (e.g. [2, 21]). These logics have in common that they are monotonic, meaning they contain the above formula (mon). The corresponding property of neighbourhood models is that neighbourhood collections are closed under supersets. Non-monotonic modal logics occur in deontic logic (see e.g. [9]) where monotonicity can lead to paradoxical obligations, and in the modelling of knowledge and related epistemic notions (cf. [25, 18]). Furthermore, the topological semantics of modal logic can be seen as neighbourhood semantics (see [24] and references).

In the present paper we try to find the "logically correct" notion of semantic equivalence in neighbourhood structures. For monotonic neighbourhood structures, this question has already been addressed (cf. [20, 13]), but as mentioned in [20], it is not immediate how to generalise monotonic bisimulation to arbitrary neighbourhood structures. This is where coalgebra comes in. Neighbourhood frames are easily seen to be coalgebras for the contravariant powerset functor composed with itself, denoted 2^2. Based on this observation the general theory of coalgebra (cf. [23, 15]) provides us with a number of candidates: behavioural equivalence, 2^2-bisimulation and a third notion (based on pushouts), which we refer to as neighbourhood bisimulation. From the logic point of view, a good equivalence notion E should have the following properties: (rel) E is characterised by relational (back-and-forth) conditions which can be effectively checked for finite models; (hm) the class of finite neighbourhood models is a Hennessy-Milner class with respect to E; and (chr) classical modal logic is the E-invariant fragment of first-order logic interpreted over neighbourhood models, i.e., we would like an analogue of Van Benthem's characterisation theorem ([3]) to hold. These logic-driven criteria form the main points of our investigation.

In section 2 we define basic notions and notation. In section 3 we investigate the three equivalence notions, first for arbitrary set functors, and then for 2^2-coalgebras. We provide relational characterisations for 2^2-bisimulation and neighbourhood bisimulation, and we show, by means of examples, that in general neighbourhood bisimilarity is stronger than behavioural equivalence, and weaker than 2^2-bisimilarity. However, when considered on a single model, the three notions coincide. The above-mentioned examples also demonstrate that 2^2-bisimilarity and neighbourhood bisimilarity fail to satisfy (hm). Furthermore, in much work on coalgebra (cf. [23]) it is often assumed that the functor preserves weak pullbacks, however, it is not always clear whether this requirement is really needed. In [11], weaker functor requirements for congruences are studied, and 2^2 provides an example of a functor which does not preserve weak pullbacks in general, but only the special ones consisting of kernel pairs. Finally, in section 4 we prove the analogue of the Van Benthem characterisation theorem, for all three equivalences. To this end, we introduce a notion of modal saturation for neighbourhood models, and since we can show that in a class of modally saturated models, modal equivalence implies behavioural equivalence, it follows that behavioural equivalence has the property (hm).

So although behavioural equivalence fails at the property (rel), we still consider it the mathematically optimal equivalence notion. Taking computational aspects into consideration, we find that neighbourhood bisimulations provide a good approximation of behavioural equivalence, while still allowing a fairly simple relational characterisation. 2^2-bisimulations, however, must be considered too strict a notion.

2 Preliminaries and Notation

In this section, we settle on notation, define the necessary coalgebraic notions, and introduce neighbourhood semantics for modal logic. For further reading on coalgebra we refer to [23, 26]. Extended discussions on neighbourhood semantics can be found in [7, 12].

Let X and Y be sets, and $R \subseteq X \times Y$ a relation. For $U \subseteq X$ and $V \subseteq Y$, we denote the R-image of U by $R[U] = \{y \in Y \mid \exists x \in U : xRy\}$, and the R-preimage of V by $R^{-1}[V] = \{x \in X \mid \exists y \in V : xRy\}$. The domain of R is $\mathsf{dom}(R) = R^{-1}[Y]$, and the range of R is $\mathsf{rng}(R) = R[X]$. Note that in the special case that R is (the graph of) a function, then image, preimage, domain and range amount to the usual definitions. Given a set X, we denote by $\mathcal{P}(X)$ the powerset of X, and for a subset $Y \subseteq X$, we write Y^c for the complement $X \setminus Y$ of Y in X.

Let $\mathsf{At} = \{p_j \mid j \in \omega\}$ be a fixed, countable set of atomic sentences. The basic modal language $\mathcal{L}(\mathsf{At})$ is defined by the grammar: $\phi ::= p_j \mid \neg\phi \mid \phi\wedge\phi \mid \Box\phi$, where $j \in \omega$. To ease notation, we write $\mathcal{L}$ instead of $\mathcal{L}(\mathsf{At})$. Formulas of $\mathcal{L}$ are interpreted in neighbourhood models.

Definition 2.1 A *neighbourhood frame* is a tuple $\langle S, \nu\rangle$ where S is a nonempty set and $\nu : S \to \mathcal{P}(\mathcal{P}(S))$ is a neighbourhood function which assigns to each state $s \in S$ a collection of neighbourhoods. A *neighbourhood model* based on a neighbourhood frame $\langle S, \nu\rangle$ is a tuple $\langle S, \nu, V\rangle$ where $V : \mathsf{At} \to \mathcal{P}(S)$ is a valuation function. ◁

Let $\mathcal{M} = \langle S, \nu, V\rangle$ be a neighbourhood model and $s \in S$. Truth of the atomic propositions is defined via the valuation: $\mathcal{M}, s \models p_i$ iff $s \in V(p_i)$, and inductively over the boolean connectives as usual. For the modal operator, we write $\mathcal{M}, s \models \Box\phi$ iff $(\phi)^{\mathcal{M}} \in \nu(s)$, where $(\phi)^{\mathcal{M}} = \{t \in S \mid \mathcal{M}, t \models \phi\}$ denotes the *truth set of ϕ in $\mathcal{M}$*. Let also $\mathcal{N}$ be a neighbourhood model. Two states, s in $\mathcal{M}$ and t in $\mathcal{N}$, are *modally equivalent (notation: $s \equiv t$)* if they satisfy the same modal formulas, i.e., $s \equiv t$ if and only if for all $\phi \in \mathcal{L}$: $\mathcal{M}, s \models \phi$ iff $\mathcal{N}, t \models \phi$. A subset $X \subseteq S$ is *modally coherent* if for all $s, t \in S$: $s \equiv t$ implies $s \in X$ iff $t \in X$. Another way of stating that $X \subseteq S$ is modally coherent would be to require that X is a union of modal equivalence classes. Note that X is modally coherent if and only if its complement $X^c = S \setminus X$ is modally coherent.

The maps between neighbourhood structures which preserve the modal structure will be referred to as bounded morphisms. These have previously been studied in the context of algebraic duality ([8]), and monotonic neighbourhood structures ([12]), in which neighbourhood collections are closed under supersets.

Definition 2.2 If $\mathcal{M}_1 = \langle S_1, \nu_1, V_1\rangle$ and $\mathcal{M}_2 = \langle S_2, \nu_2, V_2\rangle$ are neighbourhood models, and $f : S_1 \to S_2$ is a function, then f is a *(frame) bounded morphism from $\langle S_1, \nu_1\rangle$ to $\langle S_2, \nu_2\rangle$* (notation: $f : \langle S_1, \nu_1\rangle \to \langle S_2, \nu_2\rangle$), if for all $X \subseteq S_2$, we have $f^{-1}[X] \in \nu_1(s)$ iff $X \in \nu_2(f(s))$; and f is a *bounded morphism from $\mathcal{M}_1$ to $\mathcal{M}_2$* (notation: $f : \mathcal{M}_1 \to \mathcal{M}_2$) if $f : \langle S_1, \nu_1\rangle \to \langle S_2, \nu_2\rangle$ and for all $p_j \in \mathsf{At}$, and all $s \in S_1$: $s \in V_1(p_j)$ iff $f(s) \in V_2(p_j)$. $\lhd$

As usual, bounded morphisms preserve truth of modal formulas.

Lemma 2.3 *Let $\mathcal{M}_1 = \langle S_1, \nu_1, V_1\rangle$ and $\mathcal{M}_2 = \langle S_2, \nu_2, V_2\rangle$ be two neighbourhood models and $f : \mathcal{M}_1 \to \mathcal{M}_2$ a bounded morphism. For each modal formula $\phi \in \mathcal{L}$ and state $s \in S_1$, $\mathcal{M}_1, s \models \phi$ iff $\mathcal{M}_2, f(s) \models \phi$.*

Proof. By a straightforward induction on the formula structure. Details left to the reader. QED

We will work in the category Set of sets and functions. Let $\mathrm{F} : \mathsf{Set} \to \mathsf{Set}$ be a functor. Recall that an F-*coalgebra* is a pair $\langle S, \sigma\rangle$ where S is a set, and $\sigma : S \to \mathrm{F}(S)$ is a function, sometimes called the *coalgebra map*. Given two F-coalgebras, $\langle S_1, \sigma_1\rangle$ and $\langle S_2, \sigma_2\rangle$, a function $f : S_1 \to S_2$ is a *coalgebra morphism* if $\mathrm{F}(f) \circ \sigma_1 = \sigma_2 \circ f$.

The contravariant powerset functor $\mathsf{2} : \mathsf{Set} \to \mathsf{Set}$ maps a set X to $\mathcal{P}(X)$, and a function $f : X \to Y$ to the inverse image function $f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X)$. The functor $\mathsf{2}^{\mathsf{2}}$ is defined as the composition of $\mathsf{2}$ with itself. It should be clear that neighbourhood frames are $\mathsf{2}^{\mathsf{2}}$-coalgebras and vice versa, although we follow standard logic practice and exclude the empty coalgebra from being a neighbourhood structure. Similarly, a neighbourhood model $\langle S, \nu, V\rangle$ corresponds with a coalgebra map $\langle \nu, V'\rangle : S \to \mathsf{2}^{\mathsf{2}}(S) \times \mathcal{P}(\mathsf{At})$ for the functor $\mathsf{2}^{\mathsf{2}}(-) \times \mathcal{P}(\mathsf{At})$ by viewing the valuation $V : \mathsf{At} \to \mathcal{P}(S)$ as a map $V' : S \to \mathcal{P}(\mathsf{At})$ where $p_i \in V'(s)$ iff $s \in V(p_i)$. Moreover, it is straightforward to show a function $f : S_1 \to S_2$ is a bounded morphism between the neighbourhood frames $\mathcal{S}_1 = \langle S_1, \nu_1\rangle$ and $\mathcal{S}_2 = \langle S_2, \nu_2\rangle$ iff f is a coalgebra morphism from $\mathcal{S}_1$ to $\mathcal{S}_2$. Similarly, $\mathsf{2}^{\mathsf{2}}(-) \times \mathcal{P}(\mathsf{At})$-coalgebra morphisms are simply the same as bounded morphisms between neighbourhood models. In what follows we will switch freely between the coalgebraic setting and the neighbourhood setting.

Finally, we will need a number of technical constructions. The disjoint union of two sets S_1 and S_2 is denoted by $S_1 + S_2$. Disjoint unions of neighbourhood frame/models are instances of the category theoretical notion of *coproducts*, and they lift disjoint unions of sets to neighbourhood frames/models such that the inclusion maps are bounded morphisms. This amounts to the following definition for neighbourhood models; the definition for neighbourhood frames is obtained by leaving out the part about the valuations.

Definition 2.4 Let $\mathcal{M}_1 = \langle S_1, \nu_1, V_1\rangle$ and $\mathcal{M}_2 = \langle S_2, \nu_2, V_2\rangle$ be two neighbourhood models. The *disjoint union of $\mathcal{M}_1$ and $\mathcal{M}_2$* is the neighbourhood model $\mathcal{M}_1 + \mathcal{M}_2 = \langle S_1 + S_2, \nu, V\rangle$ where for all $p_j \in \mathsf{At}$, $V(p_j) = V_1(p_j) \cup V_2(p_j)$; and for $i = 1, 2$, for all $X \subseteq S_1 + S_2$, and $s \in S_i$, $X \in \nu(s)$ iff $X \cap S_i \in \nu_i(s)$. $\lhd$

In the sequel we will also use pullbacks and pushouts. We now remind the reader of how these can be constructed in Set. More information about pullbacks in Set can be found in [10]. For the general definition we refer to any standard textbook on category theory (e.g. [1]).

First, given a relation $R \subseteq S_1 \times S_2$, we can view R as a relation on $S_1 + S_2$. We denote by $\hat{R}$ the smallest equivalence relation on $S_1 + S_2$ that contains R, and $(S_1 + S_2)/\hat{R}$ is the set of $\hat{R}$-equivalence classes.

Definition 2.5 Let $f_1 : S_1 \to Z$ and $f_2 : S_2 \to Z$ be functions. The *canonical pullback of* f_1 *and* f_2 (in Set) is the triple $(\mathrm{pb}(f_1, f_2), \pi_1, \pi_2)$, where $\mathrm{pb}(f_1, f_2) := \{(s_1, s_2) \in S_1 \times S_2 \mid f_1(s_1) = f_2(s_2)\}$; and $\pi_1 : \mathrm{pb}(f_1, f_2) \to S_1$ and $\pi_2 : \mathrm{pb}(f_1, f_2) \to S_2$ are the projections.

Let $R \subseteq S_1 \times S_2$ be a relation with projections $\pi_1 : R \to S_1$ and $\pi_2 : R \to S_2$. The *canonical pushout of* R (in Set) is the triple $(\mathrm{po}(\pi_1, \pi_2), p_1, p_2)$, where $\mathrm{po}(\pi_1, \pi_2) := \big(S_1 + S_2\big)/\hat{R}$, and $p_1 : S_1 \to \mathrm{po}(\pi_1, \pi_2)$ and $p_2 : S_2 \to \mathrm{po}(\pi_1, \pi_2)$ are the obvious quotient maps. $\triangleleft$

The fact that both the canonical pullback and pushout are a pullback and pushout respectively is well-known (cf. e.g.[1]).

3 Equivalence Notions

In this section we will study various notions of "observational equivalence" for neighbourhood frames in detail. In the first part we list the three coalgebraic equivalence notions that we are going to consider. In the second part we spell out in detail what these three equivalence notions mean on neighbourhood frames.

3.1 Three coalgebraic notions of equivalence

Remark 3.1 In this subsection we introduce behavioural and relational equivalences. We want to stress that we use the word "equivalence" to indicate that a relation relates only equivalent points. We do not require these equivalences to be equivalence relations.

The main observation for defining equivalences between coalgebras is that coalgebra morphisms preserve the behaviour of coalgebra states. This basic idea motivates the well-known coalgebraic definitions of bisimilarity and behavioural equivalence. In the following F denotes an arbitrary Set functor.

Definition 3.2 Let $\mathcal{S}_1 = \langle S_1, \nu_1 \rangle$, $\mathcal{S}_2 = \langle S_2, \nu_2 \rangle$ be F-coalgebras. A relation $R \subseteq S_1 \times S_2$ is an (F-)bisimulation between $\mathcal{S}_1$ and $\mathcal{S}_2$ if there exists a function $\mu : R \to \mathrm{F}(R)$ such that for both $i = 1, 2$ the projection map $\pi_i : R \to S_i$ is a coalgebra morphism from $\langle R, \mu \rangle$ to $\mathcal{S}_i$. Two states s_1 and s_2 are *(*F*-)bisimilar* if they are linked by some bisimulation (notation: $s_1 \leftrightarrows s_2$). We call $R \subseteq S_1 \times S_2$ a *behavioural equivalence* between $\mathcal{S}_1$ and $\mathcal{S}_2$ if there are a F-coalgebra $\langle Z, \lambda \rangle$ and F-coalgebra morphisms $f_i : \langle S_i, \nu_i \rangle \to \langle Z, \lambda \rangle$ for $i = 1, 2$ such that

$R = \mathrm{pb}(f_1, f_2)$. Two states s_1 and s_2 that are related by some behavioural equivalence are called *behaviourally equivalent* (notation: $s_1 \mathrel{\underline{\leftrightarrow}}^b s_2$). $\lhd$

It has been proven in [23] that two states are F-bisimilar iff they are behavioural equivalent under the assumption that the functor F is weak pullback preserving. The same article, however, tells us that the functor 2^2 that we want to study lacks this property. Therefore it makes sense to look at both 2^2-bisimulations and behavioural equivalences on our quest for the right notion of equivalence. In fact, we will also look at a third notion that, to the best of our knowledge, has not been considered before, namely the notion of a *relational equivalence*. The motivation for introducing relational equivalences is to remedy one obvious shortcoming of behavioural equivalences: in general it is difficult to provide some criterion for a relation R to be a behavioural equivalence. Bisimulations, in contrast, can be nicely characterized using relation lifting (cf. e.g. [22]). For example when considering Kripke frames ($\mathcal{P}$-coalgebras) this characterization yields the well-known forth and back conditions for Kripke bisimulations. We want to have a similar characterization of behavioural equivalence - even if the functor does not preserve weak pullbacks.

Definition 3.3 Let $\mathcal{S}_1 = \langle S_1, \nu_1 \rangle$ and $\mathcal{S}_2 = \langle S_2, \nu_2 \rangle$ be F-coalgebras. Furthermore let $R \subseteq S_1 \times S_2$ be a relation and let $\langle Z, p_1, p_2 \rangle$ be the canonical pushout of R (cf. Def. 2.5). Then R is called a *relational equivalence* between $\mathcal{S}_1$ and $\mathcal{S}_2$ if there exists a coalgebra map $\lambda : Z \to \mathrm{F}(Z)$ such that the functions p_1 and p_2 become coalgebra morphisms from $\mathcal{S}_1$ and $\mathcal{S}_2$ to $\langle Z, \lambda \rangle$ (see the diagram below). If two states s_1 and s_2 are related by some relational equivalence we write $s_1 \mathrel{\underline{\leftrightarrow}}^r s_2$.

$$\begin{tikzcd}
 & R \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_2"] & \\
S_1 \arrow[r, "p_1"] \arrow[d, "\nu_1"'] & Z \arrow[d, dashed, "\exists\lambda"'] & S_2 \arrow[l, "p_2"'] \arrow[d, "\nu_2"] \\
\mathrm{F}(S_1) \arrow[r, "\mathrm{F}(p_1)"] & \mathrm{F}(Z) & \mathrm{F}(S_2) \arrow[l, "\mathrm{F}(p_2)"']
\end{tikzcd}$$

$\lhd$

We note that the definition of a relational equivalence is independent of the concrete representation of the pushout. This follows easily from the fact that pushouts are unique up-to isomorphism.

Remark 3.4 The main advantage of relational equivalences is that they can be characterized by some form of relation lifting [1]: Let $\langle S_1, \nu_1 \rangle$ and $\langle S_2, \nu_2 \rangle$ be F-coalgebras, let $R \subseteq S_1 \times S_2$ with projections π_1, π_2 and let $(\mathrm{po}(\pi_1, \pi_2), p_1, p_2)$ the canonical pushout of R. We define the F-lifting $\hat{\mathrm{F}}$ of R, by $\hat{\mathrm{F}}(R) := \mathrm{pb}(\mathrm{F}p_1, \mathrm{F}p_2) \subseteq \mathrm{F}(S_1) \times \mathrm{F}(S_2)$. It is not difficult to see that R is a relational equivalence iff for all $(s_1, s_2) \in R$ we have $(\nu_1(s_1), \nu_2(s_2)) \in \hat{\mathrm{F}}(R)$.

[1] The definition of $\hat{F}$ goes back to an idea by Kurz ([14]) for defining a relation lifting of non weak pullback preserving functors.

Definition 3.3 ensures that relational equivalences only relate behavioural equivalent points. The following proposition provides a first comparison between the three equivalence notions.

Proposition 3.5 *Let $\mathcal{S}_1 = \langle S_1, \nu_1 \rangle$ and $\mathcal{S}_2 = \langle S_2, \nu_2 \rangle$ be* F*-coalgebras. We have for all $s_1 \in S_1$ and $s_2 \in S_2$: $s_1 \underline{\leftrightarrow} s_2$ implies $s_1 \underline{\leftrightarrow}^r s_2$ implies $s_1 \underline{\leftrightarrow}^b s_2$.*

Proof. Suppose that $s_1 \underline{\leftrightarrow} s_2$ and let R be a bisimulation with $(s_1, s_2) \in R$. Then it is straightforward to check that R meets the requirement of Remark 3.4 and thus R is a relational equivalence, i.e., $s_1 \underline{\leftrightarrow}^r s_2$. Now suppose that R is a relational equivalence and let (Z, p_1, p_2) be the pushout of (R, π_1, π_2) where the π_i's denote the projection functions. Then by the definition of a relational equivalence it is clear that $\mathrm{pb}(p_1, p_2)$ is a behavioural equivalence. Because $R \subseteq \mathrm{pb}(p_1, p_2)$ we get $(s_1, s_2) \in \mathrm{pb}(p_1, p_2)$ and hence $s_1 \underline{\leftrightarrow}^b s_2$. QED

This proposition is clearly not enough to justify the introduction of relational equivalences: our motivation was to give a characterization of behavioural equivalence using a relation lifting. We will demonstrate, however, that behavioural equivalences give us in general a strictly weaker notion of equivalence between coalgebras than relational equivalences. Luckily both notions coincide if we restrict our attention to "full" relations. In particular, we obtain the result that behavioural equivalence and relational equivalence amount to the same thing when studied on a single coalgebra.

Lemma 3.6 *If $\mathcal{S}_1 = \langle S_1, \nu_1 \rangle$ and $\mathcal{S}_2 = \langle S_2, \nu_2 \rangle$ are* F*-coalgebras and $R \subseteq S_1 \times S_2$ is a behavioural equivalence between $\mathcal{S}_1$ and $\mathcal{S}_2$ that is* full*, i.e. $\mathsf{dom}(R) = S_1$ and $\mathsf{rng}(R) = S_2$, then R is a relational equivalence.*

Proof. Let R be a behavioural equivalence with projection maps $\pi_1 : R \to S_1$ and $\pi_2 : R \to S_2$. Then there are some F-coalgebra $\langle Z, \lambda \rangle$ and coalgebra morphisms $f_i : S_i \to Z$ for $i = 1, 2$ such that $R = \mathrm{pb}(f_1, f_2)$. Let $\langle Z', p_1, p_2 \rangle$ be the canonical pushout of R. We are going to define a function $\lambda' : Z' \to \mathrm{F}(Z')$ such that p_i is a coalgebra morphism from $\mathcal{S}_i$ to $\langle Z', \lambda' \rangle$ for $i = 1, 2$.

By the universal property of the pushout there has to be a function $j : Z' \to Z$ such that $j \circ p_i = f_i$ for $i = 1, 2$. We claim that this function is injective. First it follows from the definition of the canonical pushout that both p_1 and p_2 are surjective, because R is a full relation. Let now $z_1, z_2 \in Z'$ and suppose that $j(z_1) = j(z_2)$. The surjectivity of the p_i's implies that there are $s_1 \in S_1$ and $s_2 \in S_2$ such that $p_1(s_1) = z_1$ and $p_2(s_2) = z_2$. Hence $j(p_1(s_1)) = j(p_2(s_2))$ which in turn yields $f_1(s_1) = f_2(s_2)$. This implies $(s_1, s_2) \in R$ and as a consequence we get $p_1(s_1) = p_2(s_2)$, i.e., $z_1 = z_2$. This demonstrates that j is injective and thus there is some surjective map $e : Z \to Z'$ with $e \circ j = \mathrm{id}_{Z'}$. Now put $\lambda' := \mathrm{F}(e) \circ \lambda \circ j$. It is straightforward to check that for $i = 1, 2$ the function p_i is a coalgebra morphism from $\mathcal{S}_i$ to $\langle Z', \lambda' \rangle$ as required. QED

Theorem 3.7 *Let $\mathcal{S} = \langle S, \nu \rangle$ be an* F*-coalgebra. Every behavioural equivalence $R \subseteq S \times S$ on $\mathcal{S}$ is contained in a relational equivalence. Hence $s \underline{\leftrightarrow}^b s'$ iff $s \underline{\leftrightarrow}^r s'$ for all $s, s' \in S$.*

Proof. The theorem is a consequence of Lemma 3.6 and the fact that every behavioural equivalence R on a coalgebra $\langle S, \nu \rangle$ is contained in a full one: If $R = \mathrm{pb}(f_1, f_2)$ for two coalgebra morphisms f_1 and f_2 we construct the coequalizer h of f_1 and f_2 in the category of F-coalgebras (cf. e.g. [23, Sec. 4.2]). If we put $f := h \circ f_1$ we obtain $R \subseteq R' := \mathrm{pb}(f, f)$, and R' is obviously full. QED

3.2 Equivalences between neighbourhood frames

In this subsection we instantiate the three coalgebraic equivalence notions for 2^2-coalgebras, i.e., for neighbourhood frames.

We first consider 2^2-bisimulations. Recall from Def. 3.2 that a relation $R \subseteq S_1 \times S_2$ is a 2^2-bisimulation between two 2^2-coalgebras $\mathcal{S}_1 = \langle S_1, \nu_1 \rangle$ and $\mathcal{S}_2 = \langle S_2, \nu_2 \rangle$ if the projection maps π_1 and π_2 are bounded morphisms (2^2-coalgebra morphisms) from some 2^2-coalgebra (R, μ) to $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. By Definition 2.2 of a bounded morphism this means that for $(s_1, s_2) \in R$ and $i = 1, 2$:

$$U \in \nu_i(s_i) \quad \text{iff} \quad \pi_i^{-1}[U] \in \mu(s_1, s_2) \qquad \text{for } U \subseteq S_i.$$

This leads to two "minimal requirements" on the neighbourhood functions ν_1 and ν_2 for pairs (s_1, s_2) related by a 2^2-bisimulation. For all $U_i, U_i' \subseteq S_i$, $i = 1, 2$:

1. $\pi_i^{-1}[U_i] = \pi_i^{-1}[U_i']$ implies $U_i \in \nu_i(s_i)$ iff $U_i' \in \nu_i(s_i)$,
2. $\pi_1^{-1}[U_1] = \pi_2^{-1}[U_2]$ implies $U_1 \in \nu_1(s_1)$ iff $U_1' \in \nu_2(s_2)$.

The following definition will help us to state these requirements in a concise way.

Definition 3.8 Let $R \subseteq S_1 \times S_2$ be a relation with projection maps $\pi_i : R \to S_i$ for $i = 1, 2$. A set $U \subseteq S_1$ is called *R-unrelated* if $U \cap \mathsf{dom}(R) = \emptyset$. Similarly we call $V \subseteq S_2$ *R-unrelated* if $V \cap \mathsf{rng}(R) = \emptyset$. Furthermore we say two sets $U \subseteq S_1$ and $V \subseteq S_2$ are *R-coherent* if $\pi_1^{-1}[U] = \pi_2^{-1}[V]$. $\lhd$

It is easy to check that for sets $U, U' \subseteq S_i$ we have $\pi_i^{-1}[U] = \pi_i^{-1}[U']$ iff the symmetric difference $U \Delta U'$ of U and U' is R-unrelated, i.e., iff U and U' only differ in points that do not occur in the relation R. The notion of R-coherency can also be formulated in terms of the relation R: Let $R \subseteq S_1 \times S_2$ be a relation and let $U \subseteq S_1$, $V \subseteq S_2$. Then U and V are R-coherent iff $R[U] \subseteq V$ and $R^{-1}[V] \subseteq U$.

Using the notions of R-coherency and R-unrelatedness we can reformulate the previous requirements and prove that they in fact characterize 2^2-bisimulations.

Proposition 3.9 *Let $\mathcal{S}_1 = \langle S_1, \nu_1 \rangle$ and $\mathcal{S}_2 = \langle S_2, \nu_2 \rangle$ be neighbourhood frames. A relation $R \subseteq S_1 \times S_2$ is a* 2^2-bisimulation *between $\mathcal{S}_1$ and $\mathcal{S}_2$ iff for all $(s_1, s_2) \in R$, for all $U_1, U_1' \subseteq S_1$ and for all $U_2, U_2' \subseteq S_2$ the following two conditions are satisfied:*

1. $U_i \Delta U_i'$ is R-unrelated implies $U_i \in \nu_i(s_i)$ iff $U_i' \in \nu_i(s_i)$, for $i = 1, 2$.

2. U_1 and U_2 are R-coherent implies $U_1 \in \nu_1(s_1)$ iff $U_2 \in \nu_2(s_2)$.

Proof. It is a matter of routine checking that every 2^2-bisimulation R fulfills conditions 1 and 2. Let now $R \subseteq S_1 \times S_2$ be a relation that fulfills the conditions 1 and 2 for all $(s_1, s_2) \in R$. We define the neighbourhood function $\mu : R \to 2^2(R)$ by $\mu(s_1, s_2) := \{\pi_1^{-1}[U] \mid U \in \nu_1(s_1)\} \cup \{\pi_2^{-1}[V] \mid V \in \nu_2(s_2)\}$. In order to show that R is a 2^2-bisimulation it suffices to prove that for $i = 1, 2$ the projection functions $\pi_i : \langle R, \mu \rangle \to \mathcal{S}_i$ are bounded morphisms. We only provide the details for the proof that π_1 is a bounded morphism. We have to demonstrate that for all $(s_1, s_2) \in R$ and all $U \subseteq S_1$ we have

$$U \in \nu_1(s_1) \quad \text{iff} \quad \pi_1^{-1}[U] \in \mu(s_1, s_2). \tag{1}$$

Let $(s_1, s_2) \in R$ and $U \subseteq S_1$. By definition of $\mu(s_1, s_2)$ the direction from left to right in (1) is immediate. In order to prove the other implication in (1) suppose that $\pi_1^{-1}[U] \in \mu(s_1, s_2)$ for some $U \subseteq S_1$. According to the definition of $\mu(s_1, s_2)$ the following cases can occur:

Case $\pi_1^{-1}[U] = \pi_1^{-1}[U']$ for some $U' \in \nu_1(s_1)$. Then $U \Delta U'$ is R-unrelated and hence U must be also in $\nu_1(s_1)$ by condition 1 of the proposition.

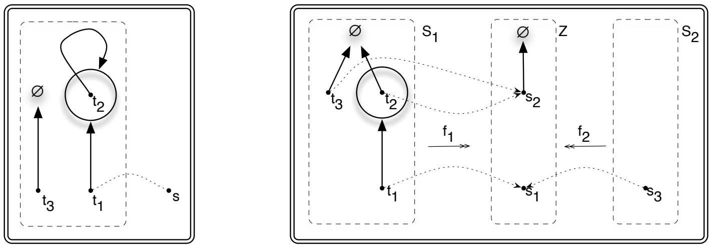
Case $\pi_1^{-1}[U] = \pi_2^{-1}[V]$ for some $V \in \nu_2(s_1)$, i.e., the sets U and V are R-coherent. Condition 2 yields therefore $U \in \nu_1(s_2)$ as required. QED

We will now demonstrate with an example that 2^2-bisimulations are too restrictive, i.e., we give an example of two states that *should be* regarded as equivalent but which are not 2^2-bisimilar.

Example 3.10 Let $T := \{t_1, t_2, t_3\}$ and $S := \{s\}$. Furthermore put $\nu_1(t_1) = \nu_1(t_2) := \{\{t_2\}\}$, $\nu_1(t_3) := \{\emptyset\}$ and $\nu_2(s) := \emptyset$ (cf. Fig. 1). We claim that there is no 2^2-bisimulation between $\langle T, \nu_1 \rangle$ and $\langle S, \nu_2 \rangle$ which relates t_1 and s.

We first note that t_3 and s cannot be related by a 2^2-bisimulation. This follows easily from the fact that $\emptyset \subseteq T$ and $\emptyset \subseteq S$ are R-coherent, and $\emptyset \in \nu_1(t_3)$ and $\emptyset \notin \nu_2(s)$. Suppose now R is a 2^2-bisimulation such that $(t_1, s) \in R$. It must then be the case that $\{t_3\} = \{t_3, t_2\} \Delta \{t_2\}$ is R-unrelated as we saw above. Therefore it follows by condition 1 of Proposition 3.9 that $\{t_3, t_2\} \in \nu_1(t_1)$ - a contradiction.

But what justifies our claim that t_1 and s *should be* bisimilar? The reason is that t_1 and s are modally equivalent: in order to see this one has first to observe that the states t_1 and t_2 are obviously modally equivalent since they have the same neighbourhoods. Therefore $\{t_2\}$, the only neighbourhood set of t_1, is *undefinable*, i.e. every formula that is true at t_2 will be also true at t_1. The semantics of the $\Box$-operator, however, only takes *definable* neighbourhoods into account, i.e. those neighbourhoods which consist exactly of those states that



Example 1

Example 2

Figure 1: Examples

make a certain modal formula true. Hence it is possible to prove the modal equivalence of t_1 and s using an easy induction on the structure of a formula.

So let us have a look at our second candidate for an equivalence notion between 2^2-coalgebras, namely what we called *relational equivalence*. In the sequel the relational equivalences between neighbourhood frames will be referred to as *neighbourhood bisimulations*. The following proposition gives a characterization of neighbourhood bisimulations in set-theoretic terms.

Proposition 3.11 *Let $\mathcal{S}_i = \langle S_i, \nu_i \rangle$, $i = 1, 2$, be neighbourhood frames. A relation $R \subseteq S_1 \times S_2$ is a* neighbourhood bisimulation *iff for all $(s_1, s_2) \in R$ and for all R-coherent sets $U_1 \subseteq S_1$ and $U_2 \subseteq S_2$: $U_1 \in \nu_1(s_1)$ iff $U_2 \in \nu_2(s_2)$.*

The following technical lemma is needed for proving Proposition 3.11.

Lemma 3.12 *Let $R \subseteq S_1 \times S_2$ be a relation with projections π_1 and π_2 and let $(Z, p_1 : S_1 \to Z, p_2 : S_2 \to Z)$ be the canonical pushout. Furthermore let $P := \mathrm{pb}(p_1, p_2)$. If two sets $U_1 \subseteq S_1$ and $U_2 \subseteq S_2$ are R-coherent then they are also P-coherent.*

Proof. Let U_1 and U_2 be R-coherent sets, i.e., $R[U_1] \subseteq U_2$ and $R^{-1}[U_2] \subseteq U_1$ and let $P := \mathrm{pb}(p_1, p_2)$. We claim that U_1 and U_2 are P-coherent. We only show that $P[U_1] \subseteq U_2$ - the other half of our claim, namely that $P^{-1}[U_2] \subseteq U_1$ can be proven analogously. The first observation to be made is that $(s, t) \in P$ implies $p_1(s) = p_2(t)$. Therefore $(s, t) \in P$ implies that $(s, t) \in \hat{R}$ where $\hat{R}$ denotes the smallest equivalence relation on $S_1 + S_2$ that contains R. It is well known that $\hat{R}$ can be computed as $(R \cup R^{-1})^*$ where the $*$-operator denotes the reflexive, transitive closure of a relation. In particular this means that $(s, t) \in \hat{R}$ iff there is some sequence of pairs in $R \cup R^{-1}$ of the form

$$w_{(s,t)} = (s_0, t_0)(t_0, s_1)(s_1, t_1) \ldots (s_m, t_m),$$

of length $2m + 1$ and with $s_0 = s$ and $t_m = t$. We call the sequence $w_{(s,t)}$ a witness of $(s, t) \in \hat{R}$.

We now prove that $s \in U_1$ and $(s, t) \in \hat{R}$ implies $t \in U_2$. The proof goes by induction on the length $2m + 1$ of the shortest witness of $(s, t) \in \hat{R}$: for $m = 0$

there is nothing to show, because in this case $(s,t) \in R$ and therefore $t \in U_2$ by the assumption that U_1 and U_2 are R-coherent. Suppose now that $(s,t) \in \hat{R}$ and the shortest witness

$$w_{(s,t)} = (s,t_0)\dots(s_m,t_m)(t_m,s_{m+1})(s_{m+1},t_{m+1})$$

has length $2(m+1)+1$. Then $(s,t_m) \in \hat{R}$ with a witness of length $2m+1$. Hence by I.H. we get $t_m \in U_2$. But then R-coherency implies that from $(t_m,s_{m+1}) \in R^{-1}$ and $(s_{m+1},t_{m+1}) \in R$ we get $s_{m+1} \in U_1$ and finally $t_{m+1} \in U_2$ as required. This finishes the proof of the lemma, because for all $s \in U_1$ and $(s,t) \in P$ we have $(s,t) \in \hat{R}$ and thus by our claim $t \in U_2$, i.e., $P[U_1] \subseteq U_2$. QED

We now turn to the proof of Proposition 3.11.

Proof of Prop. 3.11. Let $R \subseteq S_1 \times S_2$ be a relation with projections $\pi_i : R \to S_i$. First suppose that R fulfills the condition of the proposition. We have to show that R is a relational equivalence between the 2^2-coalgebras $\mathcal{S}_1$ and $\mathcal{S}_2$. Let $\langle Z, p_1, p_2 \rangle$ be the canonical pushout of R (cf. Def. 2.5). It is straightforward to check that $p_1^{-1}[U]$ and $p_2^{-1}[U]$ are R-coherent for all $U \subseteq Z$. By our assumption on R this implies $p_1^{-1}[U] \in \nu_1(s_1)$ iff $p_2^{-1}[U] \in \nu_2(s_2)$ for all $U \subseteq Z$ and all $(s_1,s_2) \in R$. Hence we have $(2^2 p_1)(\nu_1(s_1)) = (2^2 p_2)(\nu_2(s_2))$ for all $(s_1,s_2) \in R$ which is by Remark 3.4 sufficient to prove that R is a neighbourhood bisimulation.

Suppose now that R is a neighbourhood bisimulation, i.e., there is a function $\lambda : Z \to 2^2(Z)$ such that the p_i's are bounded morphisms. Furthermore let (s_1,s_2) be some element of Z and let $U_1 \subseteq S_1$ and $U_2 \subseteq S_2$ be R-coherent. We have to show that $U_1 \in \nu_1(s_1)$ iff $U_2 \in \nu_2(s_2)$ $(*)$.

First we prove that $p_i^{-1}[p_i[U_i]] = U_i$ for $i = 1,2$. Let us prove this for U_1: The inclusion from right to left is obvious. For the other direction let $s \in p_1^{-1}[p_1[U_1]]$, i.e., $p_1(s) \in p_1[U_1]$. Then there is some $s' \in U_1$ such that $p_1(s) = p_1(s')$. If $s \notin \mathsf{dom}(R)$ it follows from the definition of the pushout and $p_1(s') = p_1(s)$ that $s = s'$ and thus $s \in U_1$. If $s \in \mathsf{dom}(R)$ there is some $t \in S_2$ with $(s,t) \in R$. Then $p_1(s) = p_1(s') = p_2(t)$. This implies $(s',t) \in P = \mathrm{pb}(p_1,p_2)$ by the definition of the pullback P. By Lemma 3.12 we know that U_1 and U_2 are P-coherent and therefore $s' \in U_1$ implies $t \in U_2$. Now R-coherency of U_1 and U_2, $(s,t) \in R$ and $t \in U_2$ entails $s \in U_1$. This finishes the proof of $p_i^{-1}[p_i[U_i]] = U_i$ for $i = 1,2$. Furthermore P-coherency of U_1 and U_2 implies $p_1^{-1}[p_2[U_2]] \subseteq U_1$ and $p_2^{-1}[p_1[U_1]] \subseteq U_2$.

We still have to prove that $(*)$ holds. Define $V := p_1[U_1] \cup p_2[U_2]$. Then the results from the previous paragraph yield $p_i^{-1}[V] = U_i$ for $i = 1,2$. Therefore

$$\begin{aligned} U_1 = p_1^{-1}[V] \in \nu_1(s_1) &\overset{\text{morphism}}{\Leftrightarrow} V \in \lambda(p_1(s_1)) = \lambda(p_2(s_2)) \\ &\overset{\text{morphism}}{\Leftrightarrow} U_2 = p_2^{-1}[V] \in \nu_2(s_2) \end{aligned}$$

which shows that $(*)$ indeed holds. QED

The good news about neighbourhood bisimuations is that they capture the fact that the states t_1 and s in Example 3.10 are equivalent: The reader is invited to check that in this case $R := \{(t_1, s), (t_2, s)\}$ is a neighbourhood bisimulation. The next question is: how are neighbourhood bisimulations and behavioural equivalences related? Unfortunately the following example shows that neighbourhood bisimilarity is strictly stronger than behavioural equivalence.

Example 3.13 We are going to describe the situation that is depicted on the right in Figure 1. Let $S_1 := \{t_1, t_2, t_3\}$, $S_2 := \{s_3\}$ and define the neighbourhood functions ν_1 and ν_2 as follows: $\nu_1(t_1) := \{\{t_2\}\}$, $\nu_1(t_2) = \nu_1(t_3) := \{\emptyset\}$ and $\nu_2(s_3) := \emptyset$. We claim that the relation $R := \{(t_1, s_3)\}$ is a behavioural equivalence. Let $Z := \{s_1, s_2\}$, $\lambda(s_1) := \emptyset$ and $\lambda(s_2) := \{\emptyset\}$. Furthermore for $i \in \{1, 2\}$ we define functions $f_i : S_i \to Z$ by putting $f_1(t_1) := s_1$, $f_1(t_2) = f_1(t_3) := s_2$ and $f_2(s_3) := s_1$. Then it is straightforward to check that f_1 and f_2 are in fact bounded morphisms and that $R = \mathrm{pb}(f_1, f_2)$ as required.

At first this might look a bit surprising, because the neighbourhood frames (S_1, ν_1) and (S_2, ν_2) look rather different. But again it is not difficult to see that the states t_1 and s_3 should be considered equivalent because they are modally equivalent. Like in Example 3.10 the modal equivalence of t_1 and s_3 follows from the fact that the set $\{t_2\}$, the only neighbourhood of t_1, is not definable: all formulas that are true at t_2 are also true at t_3.

However t_1 and s_3 are not neighbourhood bisimilar: suppose for a contradiction that $(t_1, s_3) \in R'$ for some relational equivalence R'. Then it is easy to see that also $(t_2, s_3) \in R'$ (otherwise we obtain a contradicition from the fact that $\{t_2\}$ and $\emptyset$ are R-coherent). But $\emptyset \in \nu_1(t_2)$ now would imply $\emptyset \in \nu_2(s_3)$ because $\emptyset$ and $\emptyset$ are R'-coherent - a contradiction.

To sum it up: Example 3.10 showed that neighbourhood bisimulations are a clear improvement when compared to 2^2-bisimulations. Example 3.13, however, demonstrates that neighbourhood bisimulations are in general not able to capture behavioural equivalence of neighbourhood frames. If we consider behavioural equivalences on one neighbourhood frame all equivalence notions coincide.

Proposition 3.14 *Let $\mathcal{S} = (S, \nu)$ be a neighbourhood frame and $s_1, s_2 \in S$. Then $s_1 \leftrightarrows s_2$ iff $s_1 \leftrightarrows^r s_2$ iff $s_1 \leftrightarrows^b s_2$.*

Proof. The first equivalence is a consequence of Prop. 3.9 and Prop. 3.11. The second equivalence is an instance of the more general result in Theorem 3.7. Alternatively, the proposition can be proven using the result in [11] that congruence relations are bisimulations in case the functor weakly preserves kernel pairs - a property that the functor 2^2 has. QED

Remark 3.15 The results of this section can be easily extended to neighbourhood *models*: a relation R is a (neighbourhood) bisimulation/behavioural equivalence between neighbourhood models, if R is a (neighbourhood) bisimulation/behavioural equivalence between the underlying neighbourhood frames which relates only points that satisfy the same propositions.

4 The Classical Modal Fragment of First-Order Logic

We will now prove that the three equivalence notions described in section 3 all characterise the modal fragment of first-order logic over the class of neighbourhood models (Theorem 4.5). This result is an analogue of Van Benthem's characterisation theorem for normal modal logic (cf. [3]): *On the class of Kripke models, modal logic is the (Kripke) bisimulation-invariant fragment of first-order logic.* The content of Van Benthem's theorem is that the basic modal language (with $\Box$) can be seen as a fragment of a first-order language which has a binary predicate $\mathsf{R}_\Box$, and a unary predicate P for each atomic proposition p in the modal language. Formulas of this first-order language can be interpreted in Kripke models in the obvious way. Van Benthem's theorem tells us that a first-order formula $\alpha(x)$ is invariant under Kripke bisimulation if and only if $\alpha(x)$ is equivalent to a modal formula.

4.1 Translation into first-order logic

The first step towards a Van Benthem-style characterisation theorem for classical modal logic is to show that $\mathcal{L}$ can be viewed as a fragment of first-order logic. It will be convenient to work with a *two-sorted* first-order language. Formally, there are two sorts $\{\mathsf{s},\mathsf{n}\}$. Terms of the first sort (s) are intended to represent states, whereas terms of the second sort (n) are intended to represent neighbourhoods. We assume there are countable sets of variables of each sort. To simplify notation we use the following conventions: $x, y, x', y', x_1, y_2, \ldots$ denote variables of sort s (*state variables*) and $u, v, u', v', u_1, v_1, \ldots$ denote variables of sort n (*neighbourhood variables*). The language is built from a signature containing a unary predicate P_i (of sort s) for each $i \in \omega$, a binary relation symbol N relating elements of sort s to elements of sort n, and a binary relation symbol E relating elements of sort n to elements of sort s. The intended interpretation of $x\mathsf{N}u$ is "u is a neighbourhood of x", and the intended interpretation of $u\mathsf{E}x$ is "x is an element of u". The language $\mathcal{L}_1$ is built from the following grammar:

$$\phi \quad ::= \quad x = y \mid u = v \mid \mathsf{P}_i x \mid x\mathsf{N}u \mid u\mathsf{E}x \mid \neg\phi \mid \phi \wedge \psi \mid \exists x\phi \mid \exists u\phi$$

where $i \in \omega$; x and y are state variables; and u and v are neighbourhood variables. The usual abbreviations (eg. $\forall$ for $\neg\exists\neg$) apply.

Formulas of $\mathcal{L}_1$ are interpreted in two-sorted first-order structures $\mathfrak{M} = \langle D, \{P_i \mid i \in \omega\}, N, E\rangle$ where $D = D^{\mathsf{s}} \cup D^{\mathsf{n}}$ (and $D^{\mathsf{s}} \cap D^{\mathsf{n}} = \emptyset$), each $P_i \subseteq D^{\mathsf{s}}$, $N \subseteq D^{\mathsf{s}} \times D^{\mathsf{n}}$ and $E \subseteq D^{\mathsf{n}} \times D^{\mathsf{s}}$. The usual definitions of free and bound variables apply. Truth of sentences (formulas with no free variables) $\phi \in \mathcal{L}_1$ in a structure $\mathfrak{M}$ (denoted $\mathfrak{M} \models \phi$) is defined as expected. If x is a free state variable in ϕ (denoted $\phi(x)$), then we write $\mathfrak{M} \models \phi[s]$ to mean that ϕ is true in $\mathfrak{M}$ when $s \in D^{\mathsf{s}}$ is assigned to x. Note that $\mathfrak{M} \models \exists x\phi$ iff there is an element $s \in D^{\mathsf{s}}$ such that $\mathfrak{M} \models \phi[s]$. If Ψ is a set of $\mathcal{L}_1$-formulas, and $\mathfrak{M}$ is an $\mathcal{L}_1$-model, then $\mathfrak{M} \models \Psi$ means that for all $\psi \in \Psi$, $\mathfrak{M} \models \psi$. Given a class $\mathbf{K}$ of $\mathcal{L}_1$-models, we denote the *semantic consequence relation over* $\mathbf{K}$ by $\models_{\mathbf{K}}$. That is, for a set of $\mathcal{L}_1$-formulas $\Psi \cup \{\phi\}$, we have $\Psi \models_{\mathbf{K}} \phi$, if for all $\mathfrak{M} \in \mathbf{K}$, $\mathfrak{M} \models \Psi$ implies $\mathfrak{M} \models \phi$.

We can translate modal formulas of $\mathcal{L}$ and neighbourhood models to the first-order setting in a natural way:

Definition 4.1 Let $\mathcal{M} = \langle S, \nu, V \rangle$ be a neighbourhood model. The *first-order translation* of $\mathcal{M}$ is the structure $\mathcal{M}^\circ = \langle D, \{P_i \mid i \in \omega\}, R_\nu, R_\ni \rangle$ where

- $D^{\mathsf{s}} = S$, $D^{\mathsf{n}} = \nu[S] = \bigcup_{s \in S} \nu(s)$
- $P_i = V(p_i)$ for each $i \in \omega$,
- $R_\nu = \{(s, U) \mid s \in D^{\mathsf{s}}, U \in \nu(s)\}$,
- $R_\ni = \{(U, s) \mid s \in D^{\mathsf{s}}, s \in U\}$.

$\lhd$

Definition 4.2 The *standard translation* of the basic modal language is a family of functions $st_x : \mathcal{L} \to \mathcal{L}_1$ defined as follows: $st_x(p_i) = \mathsf{P}_i x$, $st_x(\neg\phi) = \neg st_x(\phi)$, $st_x(\phi \wedge \psi) = st_x(\phi) \wedge st_x(\psi)$, and

$$st_x(\Box\phi) = \exists u(x\mathsf{N}u \wedge (\forall y(u\mathsf{E}y \leftrightarrow st_y(\phi))).$$

$\lhd$

Standard translations preserve truth; the easy proof is left to the reader.

Lemma 4.3 *Let $\mathcal{M}$ be a neighbourhood model and $\phi \in \mathcal{L}$. For each $s \in S$, $\mathcal{M}, s \models \phi$ iff $\mathcal{M}^\circ \models st_x(\phi)[s]$.*

In the Kripke case, every first-order model for the language with $\mathsf{R}_\Box$ can be seen as Kripke model. However, it is not the case that every $\mathcal{L}_1$-structure is the translation of a neighbourhood model. Luckily, we can axiomatize the subclass of neighbourhood models up to isomorphism. Let $\mathbf{N} = \{\mathfrak{M} \mid \mathfrak{M} \cong \mathcal{M}^\circ \text{ for some neighbourhood model } \mathcal{M}\}$, and let NAX be the following axioms

(A1) $\exists x(x = x)$

(A2) $\forall u \exists x(x\mathsf{N}u)$

(A3) $\forall u, v(\neg(u = v) \to \exists x((u\mathsf{E}x \wedge \neg v\mathsf{E}x) \vee (\neg u\mathsf{E}x \wedge v\mathsf{E}x)))$

It is not hard to see that if $\mathcal{M}$ is a neighbourhood model, then $\mathcal{M}^\circ \models \mathsf{NAX}$. The next result states that, in fact, NAX completely characterizes the class $\mathbf{N}$.

Proposition 4.4 *Suppose $\mathfrak{M}$ is an $\mathcal{L}_1$-model and $\mathfrak{M} \models \mathsf{NAX}$. Then there is a neighbourhood model $\mathfrak{M}_\circ$ such that $\mathfrak{M} \cong (\mathfrak{M}_\circ)^\circ$.*

Proof. Let $\mathfrak{M} = \langle D^{\mathsf{s}} + D^{\mathsf{n}}, \{P_i \mid i \in \omega\}, N, E\rangle$ be an $\mathcal{L}_1$-model such that $\mathfrak{M} \models \mathsf{NAX}$. We will construct from $\mathfrak{M}$ a neighbourhood model $\mathfrak{M}_\circ = \langle S, \nu, V\rangle$ such that $\mathfrak{M} \cong (\mathfrak{M}_\circ)^\circ$. First, define the map $\eta : D^{\mathsf{n}} \to \mathcal{P}(D^{\mathsf{s}})$ by $\eta(u) = \{s \in D^{\mathsf{s}} \mid uEs\}$. We take $S = D^{\mathsf{s}}$. Note that since $\mathfrak{M} \models \mathsf{A1}$, $S \neq \emptyset$. Now define for each $s \in S$ and each $X \subseteq S$: $X \in \nu(s)$ iff there is a $u \in D^{\mathsf{n}}$ such that sNu and $X = \eta(u)$, and define for all $i \in \omega$, $V(p_i) = \{s \in S \mid \mathfrak{M} \models \mathsf{P}_i[s]\}$. Then $\mathfrak{M}_\circ$ is clearly a well-defined neighbourhood model, and we claim that the map $\mathrm{id} + \eta : D^{\mathsf{s}} + D^{\mathsf{n}} \to D^{\mathsf{s}} + \bigcup_{s\in D^{\mathsf{s}}} \nu(s)$ is an isomorphism from $\mathfrak{M}$ to $(\mathfrak{M}_\circ)^\circ = \langle S \cup \nu[S], \{P'_i \mid i \in \omega\}, R_\nu, R_\ni\rangle$ (cf. Definition 4.1).

Firstly, it follows directly from $\mathfrak{M} \models \mathsf{A3}$, that η is injective. Secondly, by the definition of η and the $()^\circ$-construction, the range of η, $\mathsf{rng}(\eta)$, contains $\bigcup_{s\in D^{\mathsf{s}}} \nu(s)$. The inclusion $\mathsf{rng}(\eta) \subseteq \bigcup_{s\in D^{\mathsf{s}}} \nu(s)$ follows from the assumption that $\mathfrak{M} \models \mathsf{A2}$, since this implies that for every $u \in D^{\mathsf{n}}$ there is an $s \in D^{\mathsf{s}}$ such that $\eta(u) \in \nu(s)$. Finally, we check the structural conditions: We have for all $i \in \omega$, $s \in P_i$ iff $s \in V(p_i)$ iff $s \in P'_i$. Similarly, for all $s \in D^{\mathsf{s}}$, and all $u \in D^{\mathsf{n}}$: sNu iff $\eta(u) \in \nu(s)$ iff $sR_\nu\eta(u)$, and uEs iff $s \in \eta(u)$ iff $\eta(u)R_\ni s$. QED

Thus, in a precise way, we can think of models in $\mathbf{N}$ as neighbourhood models. In particular, if $\mathfrak{M}$ and $\mathfrak{N}$ are in $\mathbf{N}$ we will write $\mathfrak{M} + \mathfrak{N}$ by which we (strictly speaking) mean the $\mathcal{L}_1$-model $(\mathfrak{M}_\circ + \mathfrak{N}_\circ)^\circ$ (which is also in $\mathbf{N}$).

Furthermore, Proposition 4.4 implies that we can work relative to $\mathbf{N}$ while still preserving nice first-order properties such as compactness and the existence of countably saturated models. These properties are essential in the proof of Theorem 4.5. Recall (cf. [6]) the definition of countable saturation. Let $\mathfrak{M}$ be a first-order $\mathcal{L}_1$-model with domain M. For a subset $C \subseteq M$, the *C-expansion $\mathcal{L}_1[C]$ of $\mathcal{L}_1$* is the two-sorted first-order language obtained from $\mathcal{L}_1$ by adding a constant $\underline{c}$ for each $c \in C$. Now $\mathcal{L}_1[C]$-formulas are interpreted in $\mathfrak{M}$ by requiring that a new constant $\underline{c}$ is interpreted as the singleton $\{c\}$. The $\mathcal{L}_1$-model $\mathfrak{M}$ is *countably saturated*, if for every finite $C \subseteq M$, and every collection $\Gamma(x)$ of $\mathcal{L}_1[C]$-formulas with one free variable x the following holds: If $\Gamma(x)$ is finitely satisfiable in $\mathfrak{M}$, then $\Gamma(x)$ is satisfiable in $\mathfrak{M}$.

4.2 Characterisation theorem

We are now able to formulate our characterisation theorem. Let $\sim$ be a relation on model-state pairs. Over the class $\mathbf{N}$, an $\mathcal{L}_1$-formula $\alpha(x)$ is *invariant under* $\sim$, if for all models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ in $\mathbf{N}$ and all sort s-domain elements s_1 and s_2 of $\mathfrak{M}_1$ and $\mathfrak{M}_2$, respectively, we have $\mathfrak{M}_1, s_1 \sim \mathfrak{M}_2, s_2$ implies $\mathfrak{M}_1 \models \alpha[s_1]$ iff $\mathfrak{M}_2 \models \alpha[s_2]$. Over the class $\mathbf{N}$, an $\mathcal{L}_1$-formula $\alpha(x)$ *is equivalent to the translation of a modal formula* if there is a modal formula $\phi \in \mathcal{L}$ such that for all models $\mathfrak{M}$ in $\mathbf{N}$, and all s-domain elements s in $\mathfrak{M}$, $\mathfrak{M} \models \alpha[s]$ iff $\mathfrak{M} \models st_x(\phi)[s]$.

Theorem 4.5 *Let $\alpha(x)$ be an $\mathcal{L}_1$-formula. Over the class* $\mathbf{N}$ *(of neighbourhood models) the following are equivalent:*

1. *$\alpha(x)$ is equivalent to the translation of a modal formula,*
2. *$\alpha(x)$ is invariant under behavioural equivalence,*
3. *$\alpha(x)$ is invariant under neighbourhood bisimilarity,*
4. *$\alpha(x)$ is invariant under $\mathsf{2}^{\mathsf{2}}$-bisimilarity.*

Our proof of Theorem 4.5 uses essentially the same ingredients as the proof of Van Benthem's theorem (see e.g. [5]). In particular, we define a notion of modal saturation which ensures that modal equivalence implies behavioural equivalence. To this end, we need the following notion of satisfiability. Let Ψ be a set of $\mathcal{L}$-formulas, and let $\mathcal{M} = \langle S, \nu, V \rangle$ be a neighbourhood model. We say that Ψ *is satisfiable in a subset* $X \subseteq S$ of $\mathcal{M}$, if there is an $s \in X$ such that for all $\psi \in \Psi$, $\mathcal{M}, s \models \psi$. The set Ψ *is finitely satisfiable in* $X \subseteq S$, if any finite subset $\Psi_0 \subseteq \Psi$ is satisfiable in X. Recall (from pg. 3) that $X \subseteq S$ is modally coherent if for all $s, t \in S$: $s \equiv t$ implies $s \in X$ iff $t \in X$.

Definition 4.6 [Modal saturation] A neighbourhood model $\mathcal{M} = \langle S, \nu, V \rangle$ is *modally saturated*, if for all modally coherent neighbourhoods $X \in \nu[S]$, and all sets Ψ of modal $\mathcal{L}$-formulas the following holds:

1. If Ψ is finitely satisfiable in X, then Ψ is satisfiable in X, and
2. If Ψ is finitely satisfiable in X^c, then Ψ is satisfiable in X^c.

$\lhd$

The reason we need modally saturated models is that they allow quotienting with the modal equivalence relation. The property which ensures this *modal quotient* is well-defined is that in a modally saturated model, a modally coherent neighbourhood is definable by a modal formula. The consequence is that modally equivalent states are identified in the modal quotient, and hence behaviourally equivalent via the quotient map, and we have the following proposition.

Proposition 4.7 *Let $\mathcal{M} = \langle S, \nu, V \rangle$ be a modally saturated neighbourhood model. We have for all $s, t \in S$: $s \equiv t$ iff $s \leftrightarroweq^b t$.*

Proof. Behaviourally equivalent states are modally equivalent, since modal formulas are invariant under bounded morphisms. In order to show that in a modally saturated neighbourhood model, modal equivalence implies behavioural equivalence, we build the quotient $\mathcal{M}_\equiv$ of $\mathcal{M}$ with the modal equivalence relation. We denote the modal equivalence class of a state s by $s_\equiv$, and the natural map which sends a state to its modal equivalence class is denoted by ε, i.e. $\varepsilon(s) = s_\equiv$. Let $\mathcal{M}_\equiv := \langle S_\equiv, \nu_\equiv, V_\equiv \rangle$ be defined by taking

- $S_{\equiv} = \{s_{\equiv} \mid s \in S\}$,
- for all $s_{\equiv} \in S_{\equiv}$: $s_{\equiv} \in V_{\equiv}(p)$ iff $s \in V(p)$,
- for all $s_{\equiv} \in S_{\equiv}$, and all $Y \subseteq S_{\equiv}$: $Y \in \nu_{\equiv}(s_{\equiv})$ iff $\varepsilon^{-1}[Y] \in \nu(s)$.

If $\mathcal{M}_{\equiv}$ is well-defined, it is immediate that the natural map $\varepsilon : S \to S_{\equiv}$ is a bounded morphism, and the behavioural equivalence we need is obtained as the pullback of $\mathcal{M} \xrightarrow{\varepsilon} \mathcal{M}_{\equiv} \xleftarrow{\varepsilon} \mathcal{M}$. It is easy to see that $V_{\equiv}$ is well-defined, since modally equivalent states satisfy the same atomic propositions. To see that also $\nu_{\equiv}$ is well-defined, let $Y \subseteq S_{\equiv}$, and $s, t \in S$ be such that $s \equiv t$. We need to show that $\varepsilon^{-1}[Y] \in \nu(s)$ iff $\varepsilon^{-1}[Y] \in \nu(t)$. First note that $\varepsilon^{-1}[Y]$ is modally coherent: if $x \equiv y$ then $x \in \varepsilon^{-1}[Y]$ iff $y \in \varepsilon^{-1}[Y]$. If we can show that modally coherent neighbourhoods in $\mathcal{M}$ are definable, then we have: $\varepsilon^{-1}[Y] \in \nu(s)$ implies there is a formula $\delta \in \mathcal{L}$ such that $\varepsilon^{-1}[Y] = (\delta)^{\mathcal{M}}$, hence $\mathcal{M}, s \models \Box\delta$, and so by the modal equivalence of s and t, $\mathcal{M}, t \models \Box\delta$ which implies $\varepsilon^{-1}[Y] \in \nu(t)$. The other direction can be shown similarly. In the remainder of the proof we will show that modally coherent neighbourhoods in $\mathcal{M}$ are indeed definable.

Assume $X \in \nu[S]$ is modally coherent, i.e., for all $s, t \in S$, if $s \equiv t$, then $s \in X$ iff $t \in X$. If $X = S$ or $X = \emptyset$, then $X = (\top)^{\mathcal{M}}$ or $X = (\bot)^{\mathcal{M}}$, respectively. So suppose now that $X \neq S$ and $X \neq \emptyset$. For each $x \in X$ we have for every $y \notin X$ a modal formula δ_x^y such that $\mathcal{M}, x \models \delta_x^y$ and $\mathcal{M}, y \models \neg\delta_x^y$. Let $\Delta_x := \{\delta_x^y \mid y \notin X\}$. By construction, Δ_x is not satisfiable in X^c, hence by the modal saturation of $\mathcal{M}$, there is a finite subset $\Delta_x^0 \subseteq \Delta_x$ which is also not satisfiable in X^c. Let

$$\delta_x := \bigwedge_{\delta_x^y \in \Delta_x^0} \delta_x^y, \qquad \text{and} \qquad \Delta := \{\neg\delta_x \mid x \in X\}.$$

Then Δ is not satisfiable in X, and hence by modal saturation there is a finite subset $\Delta_0 \subseteq \Delta$ which is not satisfiable in X, hence for all $x' \in X$, $\mathcal{M}, x' \models \bigvee_{\neg\delta_x \in \Delta_0} \delta_x$, that is,

$$X \subseteq \bigcup_{\neg\delta_x \in \Delta_0} (\delta_x)^{\mathcal{M}}. \tag{2}$$

On the other hand, since each δ_x is not satisfiable in X^c, we have for each $x \in X$, $X^c \subseteq (\neg\delta_x)^{\mathcal{M}}$, and therefore also $X^c \subseteq \bigcap_{\neg\delta_x \in \Delta_0} (\neg\delta_x)^{\mathcal{M}}$, hence

$$X \supseteq \Big(\bigcap_{\neg\delta_x \in \Delta_0} (\neg\delta_x)^{\mathcal{M}} \Big)^c = \bigcup_{\neg\delta_x \in \Delta_0} (\delta_x)^{\mathcal{M}}. \tag{3}$$

¿From (2) and (3) it follows that we have $X = (\delta)^{\mathcal{M}}$ by taking

$$\delta := \bigvee_{\neg\delta_x \in \Delta_0} \delta_x.$$

QED

Since all three equivalence notions coincide on a single model (Proposition 3.14), we obtain the following corollary.

Corollary 4.8 *Let $\mathcal{M} = \langle S, \nu, V\rangle$ be a modally saturated neighbourhood model. We have for all $s, t \in S$: $s \equiv t$ iff $s \mathrel{\underline{\leftrightarrow}}^b t$ iff $s \mathrel{\underline{\leftrightarrow}}^r t$ iff $s \mathrel{\underline{\leftrightarrow}} t$.*

Furthermore, it can easily be shown that finite neighbourhood models are modally saturated, hence the modal quotient of the disjoint union of two finite neighbourhood models is well-defined. This means that over the class of finite neighbourhood models, we can always construct a behavioural equivalence containing any given pair of modally equivalent states. In other words, finite neighbourhood models form a Hennessy-Milner class with respect to behavioural equivalence. This, however, is not the case with respect to 2^2-bisimulation or neighbourhood bisimulation, as Examples 3.10 and 3.13 in section 3 show. We sum up in the next proposition.

Proposition 4.9 *Over the class of finite neighbourhood models, modal equivalence implies behavioural equivalence, but not 2^2-bisimilarity nor neighbourhood bisimilarity.*

Proof. Example 3.10 of section 3 exhibits a pair of finite neighbourhood models containing states that are modally equivalent, but not 2^2-bisimilar. Similarly, Example 3.13 shows the existence of modally equivalent states that are not neighbourhood bisimilar. Hence, on the class of finite neighbourhood models, modal equivalence implies neither 2^2-bisimilarity nor neighbourhood bisimilarity. To prove that this does hold with respect to behavioural equivalence, it suffices to show that finite neighbourhood models are modally saturated, since then the disjoint union of finite neighbourhood models is modally saturated. Hence if $\mathcal{M}_1$ and $\mathcal{M}_2$ are finite neighbourhood models containing states s_1 and s_2, respectively, such that $\mathcal{M}_1, s_1 \equiv \mathcal{M}_2, s_2$, then also $\mathcal{M}_1 + \mathcal{M}_2, \mathrm{inc}_1(s_1) \equiv \mathcal{M}_1 + \mathcal{M}_2, \mathrm{inc}_2(s_2)$. Now from Proposition 4.7 there are a neighbourhood model $\mathcal{M}$ and bounded morphisms $f_i : \mathcal{M}_1 + \mathcal{M}_2 \to \mathcal{M}$, $i = 1, 2$, such that $f_1(\mathrm{inc}_1(s_1)) = f_2(\mathrm{inc}_2(s_2))$, and so $s_1 \mathrel{\underline{\leftrightarrow}}^b s_2$ via $f_i \circ \mathrm{inc}_i : \mathcal{M}_i \to \mathcal{M}$, $i = 1, 2$.

$$\begin{tikzcd}
\mathcal{M}_1 \arrow[r, "\mathrm{inc}_1"] & \mathcal{M}_1 + \mathcal{M}_2 \arrow[d, "\varepsilon"] & \mathcal{M}_2 \arrow[l, "\mathrm{inc}_2"'] \\
 & (\mathcal{M}_1 + \mathcal{M}_2)_{\equiv} &
\end{tikzcd}$$

To prove that finite neighbourhood models are modally saturated, let $\mathcal{M} = \langle S, \nu, V\rangle$ be a finite neighbourhood model, and let Ψ be a set of modal $\mathcal{L}$-formulas. Suppose now that $X = \{x_0 \ldots, x_n\}$ is a neighbourhood of some state s, and Ψ is not satisfiable in X. This means that for each $i = 0, \ldots, n$, there is an $\mathcal{L}$-formula $\psi_i \in \Psi$ such that $\mathcal{M}, x_i \not\models \psi_i$. But then $\{\psi_0, \ldots, \psi_n\}$ is a finite subset of Ψ which is not satisfiable in X. The other saturation condition is shown similarly. QED

In the proof of the characterisation theorem, we will need to construct modally saturated models from arbitrary neighbourhood models. The first step towards this is to obtain ω-saturated $\mathcal{L}_1$-models. This can be done in the form of ultrapowers using standard first-order logic techniques: Every $\mathcal{L}_1$-model has an ω-saturated, elementary extension (see e.g. [6]). The second step is to show that any ω-saturated neighbourhood model (viewed as a $\mathcal{L}_1$-model) is modally saturated. Before we state and prove this lemma, we recall (cf. [6]) the definition of ω-saturation. Let $\mathfrak{M}$ be a first-order $\mathcal{L}_1$-model with domain M. For a subset $C \subseteq M$, the *C-expansion $\mathcal{L}_1[C]$ of $\mathcal{L}_1$* is the two-sorted first-order language obtained from $\mathcal{L}_1$ by adding a constant $\underline{c}$ for each $c \in C$. Now $\mathcal{L}_1[C]$-formulas are interpreted in $\mathfrak{M}$ by requiring that a new constant $\underline{c}$ is interpreted as the singleton $\{c\}$. The $\mathcal{L}_1$-model $\mathfrak{M}$ is *ω-saturated*, if for every finite $C \subseteq M$, and every collection $\Gamma(x)$ of $\mathcal{L}_1[C]$-formulas with one free variable x the following holds: If $\Gamma(x)$ is finitely satisfiable in $\mathfrak{M}$, then $\Gamma(x)$ is satisfiable in $\mathfrak{M}$.

Lemma 4.10 *Let $\mathfrak{M}$ be a model in $\mathbf{N}$, and let $\mathfrak{M}_\circ$ be its corresponding neighbourhood model. If $\mathfrak{M}$ is ω-saturated, then $\mathfrak{M}_\circ$ is modally saturated.*

Proof. Let $\mathfrak{M}$ be an $\mathcal{L}_1$-model in $\mathbf{N}$, $\mathfrak{M}_\circ = \langle S, \nu, V\rangle$ its corresponding neighbourhood model (cf. Proposition 4.4), and assume that $\mathfrak{M}$ is ω-saturated. Let Ψ be a set of modal $\mathcal{L}$-formulas, and let $U \subseteq S$ be a neighbourhood of some state s. Then U corresponds to a domain element $u \in D^{\mathsf{n}}$ of $\mathfrak{M}$ via the isomorphism $\mathfrak{M} \cong (\mathfrak{M}_\circ)^\circ$. If Ψ is finitely satisfiable in U in $\mathfrak{M}_\circ$, then the set of $\mathcal{L}_1[\{u\}]$-formulas $\{R_\ni \underline{u}x\} \cup \{st_x(\psi) \mid \psi \in \Psi\}$ is finitely satisfiable in $\mathfrak{M}$, and hence satisfiable, which implies that Ψ is satisfiable in U. Similarly, if Ψ is finitely satisfiable in U^c, then the set of $\mathcal{L}_1[\{u\}]$-formulas $\{\neg R_\ni \underline{u}x\} \cup \{st_x(\psi) \mid \psi \in \Psi\}$ is finitely satisfiable in $\mathfrak{M}$, and hence satisfiable, which implies that Ψ is satisfiable in U^c. QED

We are now ready to prove Theorem 4.5. The proof proceeds along the same lines as the proof of Van Benthem's theorem with one modification. One of the elements used in Van Benthem's proof is that over a class of modally saturated Kripke models, modal equivalence implies Kripke bisimilarity. Note however, that Proposition 4.7 does not provide us with such an analgoue, since it requires modal equivalence in a single modally saturated model. The solution is to first take the sum to obtain a single neighbourhood model, and then take a modally saturated, elementary extension in which modal equivalence does imply behavioural equivalence, and hence also neighbourhood bisimilarity and 2^2-bisimilarity (Corollary 4.8).

Proof of Theorem 4.5. It is clear that *2* ⇒ *3* ⇒ *4* (cf. Proposition 3.5). To see that *4* ⇒ *2*, we only need to recall (cf. [23]) that graphs of bounded morphisms are 2^2-bisimulations. Furthermore, as truth of modal formulas is preserved by behavioural equivalence, *1* ⇒ *2* is clear. We complete the proof by showing that *2* ⇒ *1*.

Let $\mathrm{MOC}_{\mathbf{N}}(\alpha) = \{st_x(\phi) \mid \phi \in \mathcal{L},\ \alpha(x) \models_{\mathbf{N}} st_x(\phi)\}$ be the set of modal consequences of $\alpha(x)$ over the class $\mathbf{N}$. It suffices to show that $\mathrm{MOC}_{\mathbf{N}}(\alpha) \models_{\mathbf{N}} \alpha(x)$, since then by compactness there is a finite subset $\Gamma(x) \subseteq \mathrm{MOC}_{\mathbf{N}}(\alpha)$ such that $\Gamma(x) \models_{\mathbf{N}} \alpha(x)$ and $\alpha(x) \models_{\mathbf{N}} \bigwedge \Gamma(x)$. It follows that $\alpha(x)$ is $\mathbf{N}$-equivalent to $\bigwedge \Gamma(x)$, which is the translation of a modal formula. So suppose $\mathfrak{M}$ is a model in $\mathbf{N}$ and $\mathrm{MOC}_{\mathbf{N}}(\alpha)$ is satisfied at some element s in $\mathfrak{M}$. We must show that $\mathfrak{M} \models \alpha[s]$. Consider the set $T(x) = \{st_x(\phi) \ \mid \mathfrak{M}_\circ, s \models \phi\} \cup \{\alpha(x)\}$. $T(x)$ is $\mathbf{N}$-consistent, since suppose to the contrary that $T(x)$ is $\mathbf{N}$-inconsistent, then by compactness, there is a finite collection of modal formulas $\phi_1, \ldots, \phi_n$ such that $\mathfrak{M}_\circ, s \models \phi_i$ for all $i = 1, \ldots, n$ and $\alpha(x) \models_{\mathbf{N}} \neg \bigwedge_{i=1}^n st_x(\phi_i)$, which implies that $\neg \bigwedge_{i=1}^n st_x(\phi_i) \in \mathrm{MOC}_{\mathbf{N}}(\alpha)$. But this contradicts the assumption that $\mathfrak{M} \models \mathrm{MOC}_{\mathbf{N}}(\alpha)[s]$ and $\mathfrak{M} \models st_x(\phi_i)[s]$ for all $i = 1, \ldots, n$. Hence $T(x)$ is satisfied at an element t in some $\mathfrak{N} \in \mathbf{N}$, and by construction, s and t are modally equivalent: For all modal formulas $\phi \in \mathcal{L}$, $\mathfrak{M} \models st_x(\phi)[s]$ implies $st_x(\phi) \in T(x)$, and hence $\mathfrak{N} \models st_x(\phi)[t]$. Conversely, $\mathfrak{M} \not\models st_x(\phi)[s]$ iff $\mathfrak{M} \models \neg st_x(\phi)[s]$ which implies $\neg st_x(\phi) \in T(x)$, and hence $\mathfrak{N} \not\models st_x(\phi)[t]$.

Take now an ω-saturated, elementary extension $\mathfrak{U}$ of $\mathfrak{M} + \mathfrak{N}$. Note that $\mathfrak{U} \in \mathbf{N}$, since satisfiablity of NAX is preserved under elementary extensions. Moreover, the images s_U and t_U in $\mathfrak{U}$ of s and t, respectively, are also modally equivalent, since modal truth is transferred by elementary maps. Now since $\mathfrak{U}$ is modally saturated, it follows from Proposition 4.7 that s_U and t_U are behaviourally equivalent. The construction is illustrated in the following diagram; $\preceq$ indicates that the map is elementary.

$$\begin{tikzcd}
\mathrm{MOC}_{\mathbf{N}}(\alpha)[s] \mathrel{=\!\!|} \mathfrak{M} \arrow[r, "i"] & \mathfrak{M} + \mathfrak{N} \arrow[d, "\preceq"] & \mathfrak{N} \models \alpha[t] \arrow[l, "j"'] \\
& \mathfrak{U} &
\end{tikzcd}$$

Finally, we can transfer the truth of $\alpha(x)$ from $\mathfrak{N}, t$ to $\mathfrak{M}, s$ by using the invariance of modal formulas under bounded morphisms and standard translations *(bm+st)*; elementary maps *(elem)*; and the assumption that $\alpha(x)$ is invariant under behavioural equivalence *($\alpha(x)$-beh-inv)*.

$$\begin{array}{rcll}
\mathfrak{N} \models \alpha[t] & \iff & (\mathfrak{M}_\circ + \mathfrak{N}_\circ)^\circ \models \alpha[j(t)] & \textit{(bm+st)} \\
& \iff & \mathfrak{U} \models \alpha[t_U] & \textit{(elem)} \\
& \iff & \mathfrak{U} \models \alpha[s_U] & (s_U \mathrel{\underline{\leftrightarrow}}^b t_U \ \textit{and } \alpha(x)\textit{-beh-inv}) \\
& \iff & (\mathfrak{M}_\circ + \mathfrak{N}_\circ)^\circ \models \alpha[i(s)] & \textit{(elem)} \\
& \iff & \mathfrak{M} \models \alpha[s] & \textit{(bm+st)}
\end{array}$$

QED

5 Discussion and Related Work

The main result in our paper is the characterisation theorem (Theorem 4.5). Our proof builds on ideas from the original proof of the Van Benthem characterisation theorem ([3]). Closely related to our work are also the invariance results by Pauly ([20]) on monotonic modal logic and by Ten Cate et al. ([24]) on topological modal logic. Furthermore there seems to be a connection between our work and the results on Chu spaces in [4] where Van Benthem characterises the Chu transform invariant fragment of a two-sorted first-order logic. This and other model-theoretic results such as an interpolation theorem and a Goldblatt-Thomason Theorem (cf. [16]) will be discussed in the full version of our paper. We also want to explore the possibility of proving our result using game-theoretic techniques similar to the ones exploited by Otto ([17]).

We want to stress that the paper also contains observations that might be useful in universal coalgebra. We saw that relational equivalences capture behavioural equivalence on F-coalgebras for an *arbitrary* Set functor F (see Theorem 3.7). One advantage of these relational equivalences lies in the fact that they can be characterised by a kind of relation lifting (see Remark 3.4). Therefore we believe the notion of a relational equivalence might be interesting in situations where the functor under consideration does not preserve weak pullbacks. In particular, we want to explore the exact relationship of our results on relational equivalences and the work by Gumm & Schröder ([11]).

Finally our work might be relevant for coalgebraic modal logic (see e.g. [19]). Our idea can be sketched as follows: Given a collection of *predicate liftings* for a functor F we can turn any F-coalgebra into some kind of neighbourhood frame. We would like to combine this well-known connection with Theorem 4.5, in order to prove that, under certain assumptions, coalgebraic modal logic can be viewed as the bisimulation invariant fragment of some many-sorted first-order logic.

References

[1] J. Adámek. *Theory of Mathematical Structures.* Reidel Publications, 1983.

[2] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49(5):672–713, 2002.

[3] J. van Benthem. Correspondence theory. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume II, pages 167–247. Reidel, Doordrecht, 1984.

[4] J. van Benthem. Information transfer across Chu spaces. *Logic Journal of the IGPL*, 8(6):719–731, 2000.

[5] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic.* Cambridge University Press, 2001.

[6] C. Chang and H. Keisler. *Model Theory.* North-Holland, 1973.

[7] B. F. Chellas. *Modal Logic - An Introduction.* Cambridge University Press, 1980.

[8] K. Došen. Duality between modal algebras and neighbourhood frames. *Studia Logica*, 48:219–234, 1989.

[9] L. Goble. Murder most gentle: The paradox deepens. *Philosophical Studies*, 64(2):217–227, 1991.

[10] H.P. Gumm. Functors for Coalgebras. *Algebra Universalis*, 45:135–147, 2001.

[11] H.P. Gumm and T. Schröder. Types and coalgebraic structure. *Algebra universalis*, 53:229–252, 2005.

[12] H.H. Hansen. Monotonic modal logic (Master's thesis). Research Report PP-2003-24, ILLC, University of Amsterdam, 2003.

[13] H.H. Hansen and C. Kupke. A coalgebraic perspective on monotone modal logic. In *Proceedings of the 7th Workshop on Coalgebraic Methods in Computer Science (CMCS)*, volume 106 of *Electronic Notes in Computer Science*, pages 121–143. Elsevier, 2004.

[14] A. Kurz. personal communication.

[15] A. Kurz. *Logics for Coalgebras and Applications to Computer Science*. PhD thesis, Ludwig-Maximilians-Universität, 2000.

[16] A. Kurz and J. Rosický. The goldblatt-thomason-theorem for coalgebras. In *Proceedings of CALCO*, 2007.

[17] M. Otto. Bisimulation invariance and finite models. In *Logic Colloquium '02*, volume 27 of *Lecture Notes in Logic*. Association for Symbolic Logic, 2006.

[18] V. Padmanabhan, G. Governatori, and K. Su. Knowledge assesment: A modal logic approach. In *Proceedings of the 3rd Int. Workshop on Knowledge and Reasoning for Answering Questions (KRAQ)*, 2007.

[19] D. Pattinson. Coalgebraic modal logic: Soundness, completeness and decidability of local consequence. *Theoretical Computer Science*, 309:177–193, 2003.

[20] M. Pauly. Bisimulation for general non-normal modal logic. Manuscript, 1999.

[21] M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

[22] J.J.M.M. Rutten. Relators and metric bisimulations. In *Proccedings of CMCS'98*, volume 11 of *ENTCS*, 1998.

[23] J.J.M.M. Rutten. Universal coalgebra: a theory of systems. *Theoretical Computer Science*, 249:3–80, 2000.

[24] B. ten Cate, D. Gabelaia, and D. Sustretov. Modal languages for topology: Expressivity and definability. (Under submission).

[25] M. Vardi. On epistemic logic and logical omniscience. In J. Halpern, editor, *Proceedings TARK'86*, pages 293–305. Morgan Kaufmann, 1986.

[26] Y. Venema. Algebras and coalgebras. In *Handbook of Modal Logic*, volume 3, pages 331–426. Elsevier, 2006.