ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-49

Conservative Social Laws

Thomas Ågotnes[1] and **Wiebe van der Hoek**[2] and **Michael Wooldridge**[3]

Abstract. Social laws – sets of constraints imposed on the behaviour of agents within a multi-agent system with the goal of some desirable overall behaviour resulting – are an important mechanism for coordinating multi-agent behaviour. When considering social laws in human environments, the inspiration for social laws in multi-agent systems, we argue that a key design principle is *least change*. That is, social laws are more likely to be accepted and adopted, and hence successful, if they are *conservative*, in the sense that they represent the smallest change possible from the pre-existing status quo that is required to effect the desired objective. Our aim in the present paper is to introduce, formalise, and investigate the notion of a *conservative social law* for multi-agent systems. To make the idea of a conservative social law precise, we formalise the notion of a *distance metric* for social laws, and discuss a range of possible properties for such metrics. We then formulate the conservative social law problem, (i.e., the problem of constructing an effective social law that requires the least change according to this metric), discuss some possible interpretations of distance in this context, and discuss some issues surrounding conservative social laws.

1 Introduction

Social laws, or normative systems, are a widely studied approach to coordination in multi-agent systems [8, 9, 1, 4]. The basic idea is to coordinate a social system by placing restrictions on the activities of the agents within the system; the purpose of these restrictions is typically to prevent some destructive interaction from taking place, or to facilitate some positive interaction. In the original framework of Shoham and Tennenholtz [8], the aim of a social law was to restrict the activities of agents so as to ensure that all individual agents are able to accomplish their personal goals. In [9], this idea was generalised to allow for the objective of a social law (i.e, what the designer intends to accomplish with the social law) to be specified as a logical formula. Variations on the same theme have subsequently been explored in a number of papers.

We believe that a key design principle for social laws in human society, the inspiration for social laws in multi-agent systems, is the principle of *least change*. That is, a social law is easiest to implement, e.g., because it more likely to be accepted and adopted or because it is less costly to implement, and is hence more likely to be successful, if it is *conservative*, in the sense that it represents the smallest change possible from the pre-existing status quo that is required to realise the desired objective. Our aim in the present paper is to introduce, formalise, and investigate the notion of a *conservative social law* for multi-agent systems. To do this, we use the CTL-based social law/normative system framework of Ågotnes *et al.* [1], which derives from the work of Shoham and Tennenholtz [8]. We emphasise that this framework is just one possible expression of the notion of a social law, but we find it a natural one in which to express the ideas of the present paper. The study of conservative social laws is motivated by very similar considerations as in *belief revision* where minimal change is taken as a paramount first principle.

To be able to make the idea of a conservative social law precise, we introduce the notion of a *distance metric* for social laws. A distance metric in our setting is used to measure the degree of change that a social law induces from the pre-existing status quo. We begin with a high-level definition of what we mean by a distance metric for social laws, and then introduce and discuss a range of possible axiomatic properties that such metrics might satisfy. For example, one of the axioms we consider says that if two systems are *bisimilar*, then we must regard the distance between them as being 0. The rationale is that, if they are bisimilar, then we cannot distinguish between them as logical structures using temporal logics such as CTL and CTL* [5], and so we should regard the distance between them as 0.

Having discussed the possible properties that a distance metric can or should satisfy, we move on to consider some actual metrics. For example, the simplest distance metric we consider (the *Kripke distance*) simply counts the number of transitions that are deleted in the implementation of a social law, i.e., the number of actions it forbids. We then move on to evaluate these distance metrics against the axioms given earlier: we systematically consider which concrete distance metrics satisfy which axioms. We then formulate the conservative social law problem, (i.e., the problem of constructing an effective social law that requires the least change according to some given metric), and discuss some issues surrounding conservative social laws.

2 The Formal Framework

The model of social laws we use here is that of [9, 1]; we give a complete but terse summary of the model, referring to the above cited papers for more details.

Kripke Structures: We use conventional *Kripke structures* as our semantic model for multi-agent systems (see, e.g., [6]). A Kripke structure (or *model*) K over a set of Boolean variables Φ is given by a tuple $K = \langle S, s_0, R, \pi \rangle$, where S is a finite set of states, $s_0 \in S$ is the initial state, $R \subseteq S \times S$ is a binary transition relation on S, and $\pi : S \rightarrow \mathbf{2}^{\Phi}$ is a labelling function, associating with each state in S the set of Boolean variables that are true in that state. We require R to be *total*, by which we mean that every state has a successor. We let $\mathcal{K}$ denote the set of Kripke structures (over some Φ).

When R is a transition relation in a Kripke structure and s is a state, let $next(s, R) = \{s' : (s, s') \in R\}$. Let $rch(s, R)$ denote the set of states reachable from state s in transition relation R, i.e., $rch(s, R) = next(s, R^*)$ where R^* is the reflexive transitive closure of R. When R is clear from context, we simply write $next(s)$ and $rch(s)$. Thus,

[1] University of Bergen, Norway, e-mail: thomas.agotnes@infomedia.uib.no
[2] University of Liverpool, UK, e-mail: wiebe@csc.liv.ac.uk
[3] University of Oxford, UK, e-mail: mjw@cs.ox.ac.uk

the semi-total requirement on transition relations that we mentioned above may be formalised as: $\forall s \in rch(s_0), \exists t \in S, (s,t) \in R$.

A *path* over a transition relation R is an infinite sequence of states $\tau = s_0, s_1, \ldots$ which must satisfy the property that $\forall u \in \mathbb{N}$: $s_{u+1} \in next(s_u)$. If $u \in \mathbb{N}$, then we denote by $\tau[u]$ the component indexed by u in τ (thus $\tau[0]$ denotes the first element, $\tau[1]$ the second, and so on). A path τ such that $\tau[0] = s$ is an *s-path*. Let $paths_R(s)$ denote the set of s-paths over R; we often omit reference to R, and simply write $paths(s)$. We will refer to and think of an s-path as a possible computation, or system evolution, from s.

For two Kripke structures $K_1 = \langle S, s^0, R_1, \pi\rangle$ and $K_2 = \langle S, s^0, R_2, \pi\rangle$ we will say that K_1 is a *subsystem* of K_2, denoted $K_1 \sqsubseteq K_2$ or $K_2 \sqsupseteq K_1$, iff $R_1 \subseteq R_2$.

Computation Tree Logic (CTL): CTL is a branching time temporal logic intended for representing the properties of Kripke structures [6]; since CTL is widely documented in the literature, our presentation will be brief. The syntax of CTL is defined by the following BNF grammar, where $p \in \Phi$:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathsf{E}\bigcirc\varphi \mid \mathsf{E}(\varphi\,\mathcal{U}\,\varphi) \mid \mathsf{A}\bigcirc\varphi \mid \mathsf{A}(\varphi\,\mathcal{U}\,\varphi)$$

The semantics of CTL are given with respect to the satisfaction relation "$\models$", which holds between pairs of the form K, s, (where $K \in \mathcal{K}$ is a Kripke structure and s is a state in K, such pairs are also called *pointed structures*), and formulae (where $p \in \Phi$):

$K, s \models \top$; $K, s \models p$ iff $p \in \pi(s)$;
$K, s \models \neg\varphi$ iff not $K, s \models \varphi$;
$K, s \models \varphi \vee \psi$ iff $K, s \models \varphi$ or $K, s \models \psi$;
$K, s \models \mathsf{A}\bigcirc\varphi$ iff $\forall\tau \in paths(s) : K, \tau[1] \models \varphi$;
$K, s \models \mathsf{E}\bigcirc\varphi$ iff $\exists\tau \in paths(s) : K, \tau[1] \models \varphi$;
$K, s \models \mathsf{A}(\varphi\,\mathcal{U}\,\psi)$ iff $\forall\tau \in paths(s), \exists u \in \mathbb{N}$, s.t. $K, \tau[u] \models \psi$ and $\forall v, (0 \leq v < u) : K, \tau[v] \models \varphi$
$K, s \models \mathsf{E}(\varphi\,\mathcal{U}\,\psi)$ iff $\exists\tau \in paths(s), \exists u \in \mathbb{N}$, s.t. $K, \tau[u] \models \psi$ and $\forall v, (0 \leq v < u) : K, \tau[v] \models \varphi$

The remaining classical logic connectives are defined as usual. CTL temporal operators are defined: $\mathsf{A}\Diamond\varphi \equiv \mathsf{A}(\top\,\mathcal{U}\,\varphi)$; $\mathsf{E}\Diamond\varphi \equiv \mathsf{E}(\top\,\mathcal{U}\,\varphi)$; $\mathsf{A}\Box\varphi \equiv \neg\mathsf{E}\Diamond\neg\varphi$; $\mathsf{E}\Box\varphi \equiv \neg\mathsf{A}\Diamond\neg\varphi$.

We say φ is *satisfiable* if $K, s \models \varphi$ for some Kripke structure $K \in \mathcal{K}$ and state s in K; φ is *valid* if $K, s \models \varphi$ for all Kripke structures K and states s in K. The problem of checking whether $K, s \models \varphi$ for given K, s, φ (*model checking*) can be done in deterministic polynomial time, while checking whether a given φ is satisfiable or whether φ is valid is EXPTIME-complete [6]. We write $K \models \varphi$ if $K, s_0 \models \varphi$, and $\models \varphi$ if $K \models \varphi$ for all K. For a set of formulas F, we write $K \models F$ if for all $\varphi \in F$, we have $K \models \varphi$.

Bisimulation: The expressiveness of CTL over Kripke structures is characterised by the notion of *bisimulation equivalence*. Formally, a *bisimulation relation* between two Kripke structures $K = \langle S, s_0, R, \pi\rangle$ and $K' = \langle S', s'_0, R', \pi'\rangle$ is a binary relation $\mathcal{Z} \subseteq S \times S'$ such that for all s and s' such that $s\mathcal{Z}s'$:

1. $\pi(s) = \pi'(s')$,
2. for any s_1 such that sRs_1 there is a s'_1 such that $s'R's'_1$ and $s_1\mathcal{Z}s'_1$
3. for any s'_1 such that $s'R's'_1$ there is a s_1 such that sRs_1 and $s_1\mathcal{Z}s'_1$.

Two pointed structures K, s and K', s' are *bisimulation equivalent* or *bisimilar*, which we denote by $K, s \underline{\leftrightarrow} K', s'$, if there exists a bisimulation relation $\mathcal{Z}$ between K and K' such that $s\mathcal{Z}s'$. If $K, s_0 \underline{\leftrightarrow} K', s'_0$, we also write $K \underline{\leftrightarrow} K'$. We have:

Proposition 1 (See, e.g., [5].) *For any pair of Kripke structures $K, K' \in \mathcal{K}$ and states s in K and s' in K', we have that $K, s \underline{\leftrightarrow} K', s'$ iff for all CTL formulae φ: $K, s \models \varphi$ iff $K', s' \models \varphi$.*

Social Laws: For our purposes, a *social law*, or a *normative system*, is simply *a set of constraints on the behaviour of agents in a system* [1]. More precisely, a social law defines, for every possible system transition, whether or not that transition is considered to be legal. Formally, a social law η (w.r.t. a Kripke structure $K = \langle S, s_0, R, \pi\rangle$) is a subset of R, such that $R \setminus \eta$ is a semi-total relation. The latter is a *reasonableness* constraint: it prevents social laws that lead to states with no successor. Let $N(R) = \{\eta : (\eta \subseteq R)$ and $(R \setminus \eta$ is semi-total$)\}$ be the set of social laws over R. The intended interpretation of a social law η is that $(s, s') \in \eta$ means transition (s, s') is forbidden in the context of η; hence $R \setminus \eta$ denotes the *legal* transitions of η.

The effect of *implementing* a social law on a Kripke structure is to eliminate from it all transitions that are forbidden according to this social law (see [9, 1]). If K is a Kripke structure, and η is a social law over K, then $K \dagger \eta$ denotes the Kripke structure obtained from K by deleting transitions forbidden in η. Formally, if $K = \langle S, s_0, R, \pi\rangle$, and $\eta \in N(R)$, then $K \dagger \eta = K'$ is the Kripke structure $K' = \langle S, s_0, R', \pi\rangle$ such that $R' = R \setminus \eta$ and all other components are as in K. Social laws are implemented for a reason. The designer of a social law typically has some *objective* in mind: the goal of designing a social law is that by restricting the behaviour of agents within the system appropriately, the objective is satisfied. Following Ågotnes *et al.* [9, 1], we will most of the time represent the designer's objective as a CTL formula φ. Given a Kripke structure K, a CTL objective φ, and a social law η, we say that η is *effective* if $K \dagger \eta \models \varphi$. Notice that checking effectiveness can trivially be done in polynomial time. The *feasibility* problem for social laws is the problem of determining whether, given a Kripke structure K and a CTL objective φ, there exists a social law η such that $K \dagger \eta \models \varphi$. The feasibility problem is NP-complete [8, 9].

Example 1 *Consider the models in Figure 1. This could model a situation where a system administrator can hand out a resource (say, a laptop) to either the director, or to one of the IT teachers. The director will give the resource back or keep it, and each of the three teachers can either keep the resource, give it to a colleague, or hand it back. Let us suppose we have three atoms: an atom b (the laptop is at the base) which is only true in s_0, an atom d (the laptop is with the director) which is only true in d, an atom t (the laptop is with one of the teachers), which is only true in t_1, t_2 and t_3. Since in our language we do not care about which teacher is owning the laptop if it is with the teachers, we have $K \underline{\leftrightarrow} K'$ and $K' \underline{\leftrightarrow} K''$ and $K \underline{\leftrightarrow} K''$. So all three structures agree on all properties, whether they are expressed in* LTL, CTL *or* CTL*. *Some examples of* CTL *formulas that are true in state s_0 are $\mathsf{A}\Box\mathsf{E}\bigcirc b$ (on all paths, it is always the case that the laptop can be returned to the administrator in the next step), and $\mathsf{E}\bigcirc\mathsf{E}\Box t$ (there is an evolution of the system s.t. in the next state, there is a computation where the teachers keep the laptop forever).*

3 Distance Metrics

As we indicated above, the designer of a social law will typically have some overall objective in mind when designing it, which in our framework is represented as a CTL formula φ. The *primary* criterion by which a social law will be judged will be whether the social law is

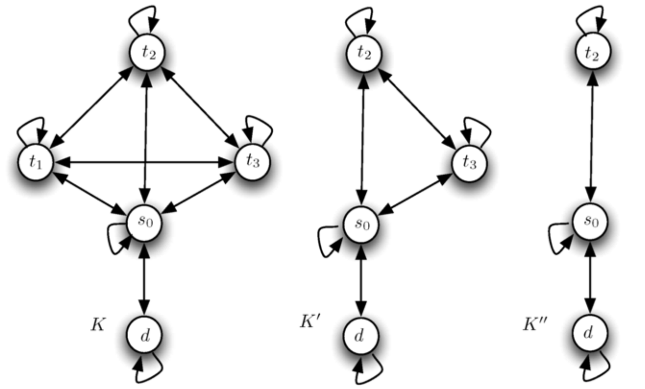

Figure 1. Three bisimilar models. We have not drawn non-reachable worlds (an isolated state t_1 in K' and two isolated states t_1 and t_3 in K'').

effective, i.e., whether, after implementing it, the objective φ holds. However, this will not in general be the *only* criterion. For example, Ågotnes and Wooldridge argued that in some cases it makes sense to weigh the costs and benefits of social laws, and to aim for social laws that strike an optimal balance between these [2]. In this paper, we consider a related issue: when considering two social laws that achieve some objective, we argue that the social law which brings the *least change* is likely to be more readily accepted by a society. However, in order to make this idea precise, we need to formalise and quantify in some way exactly what we mean by "least change". To do this, we now introduce *distance metrics*.

To understand the notion of a distance metric for our setting, we first recall some mathematical definitions. Let X be some set of objects with an *indistinguishability* relation $\sim \subseteq X \times X$ defined on this set. For example, it could be that $X = \mathbb{N}$, the set of natural numbers, with $\sim$ being the usual mathematical equality relation, "=". Now, a function $d : X \times X \to \mathbb{R}_+$ is said to be a *distance metric* if it satisfies the following three axioms:

1. *Indistinguishability*: $d(x,y) = 0$ iff $x \sim y$.
2. *Symmetry*: $d(x,y) = d(y,x)$.
3. *Subadditivity*: $d(x,z) \leq d(x,y) + d(y,z)$.

For the purposes of this paper, we are interested in metrics that can be used to measure the size of change induced by a social law. That is, we are given a Kripke structure K and a social law η, and we want to know what the distance is between the Kripke structure K and the Kripke structure $K \dagger \eta$. In this case, the objects we are measuring the distance between are not arbitrary members of $\mathcal{K}$, the set of all Kripke structures. We know that K is a supersystem of $K \dagger \eta$, i.e., $K \sqsupseteq K \dagger \eta$. Thus the distance metrics we will consider only need to be defined for pairs (K, K') when $K \sqsupseteq K'$. For this reason, we will not be concerned with considering the symmetry axiom in this paper, although we will see versions of indistinguishability and subadditivity.

Formally, we will model distance metrics as partial functions

$$\delta : \mathcal{K} \times \mathcal{K} \to \mathbb{R}_+$$

where the value $\delta(K, K')$ need only be defined when $K' \sqsubseteq K$.

Now, relating to distance metrics, two important questions suggest themselves; we address them in the subsections that follow:

- First is the issue of *what counts as a reasonable distance metric* – that is, what *criteria* we would expect a "reasonable" distance metric to have. As we will see, there are several possible interpretations of the axioms listed above in our setting.
- Second is the issue of *what practical measures we can use to measure distance* – that is, how can we actually measure distance between Kripke structures, in such a way as to be reasonable according to the criteria we set out above.

3.1 Axioms for Distance Metrics

In this section, we explore the question of what possible properties we would expect a distance metric to satisfy. We state these properties as a set of axioms. We will start with the simplest, most obvious, and arguably weakest axiom that makes sense.

The equality axiom, which we will denote by (*Equal*), says that if two Kripke structures are equal, then the distance between them should be 0. Formally, a distance metric δ satisfies this axiom if:

$$(K = K') \to \delta(K, K') = 0. \qquad (\textit{Equal})$$

Now, it seems hard to argue against this axiom in terms of whether it makes sense for distance metrics. If two Kripke structures are equal in the sense that they agree on every component, then surely we should accept that the distance between them is 0. However, as we will now argue, this requirement (of equality), while surely reasonable, may in some cases be stronger than necessary.

We will refer to the next axiom as the *bisimulation axiom*. A distance metric δ satisfies this axiom if:

$$(K \leftrightarrows K') \to \delta(K, K') = 0. \qquad (\textit{Bisim}_1)$$

Now, the rationale for this axiom is the following. If the properties of a system that we care about are expressed in a language like CTL, then the fact that two Kripke structures are bisimilar means that we *cannot tell them apart*, and hence we should regard the distance between them as being 0. Notice that there is a fairly strong condition on this statement: this axiom makes sense if the properties we are interested in can be captured in CTL, but not necessarily otherwise. In fact, we can strengthen this statement: it is known that if two Kripke structures are bisimilar then in fact they must agree on the truth status of all formulae expressable in a much richer logic, namely CTL* [5]. Moreover, since CTL* subsumes linear time temporal logic (LTL), this implies that if the properties of Kripke models that we are interested in are expressed in LTL, then axiom (*Bisim₁*) seems reasonable.

Next, we will consider a related axiom, also concerned with bisimulation, this time between *end systems*. This axiom says that if we have two systems K' and K'', both of which are subsystems of K, such that K' and K'' are bisimilar, then the distance between K and K' must be the same as the distance between K and K''. Formally:

$$(K \sqsupseteq K') \wedge (K \sqsupseteq K'') \wedge (K' \leftrightarrows K'') \to \delta(K, K') = \delta(K, K''). \qquad (\textit{Bisim}_2)$$

The motivation for this axiom is, arguably, less compelling than that for (*Bisim₁*), but seems nevertheless reasonable. If we cannot tell two Kripke structures apart, then *how we got to them* is arguably not important. That is, this axiom says that it is not the mechanism by which a Kripke structure is reached that is significant, but the properties the structure satisfies; and if two structures satisfy the same properties, then we should regard the distance to them as being the same

Along the lines of the previous metric, it is also natural to consider bisimulation between *source systems*.

$$(K \sqsupseteq K'') \wedge (K' \sqsupseteq K'') \wedge (K \leftrightarrows K') \to \delta(K, K'') = \delta(K', K''). \qquad (\textit{Bisim}_3)$$

This requirement says that if two bisimilar Kripke structures K and K' can both lead to the same subsystem K'', then the effort this reduction takes from K to K'', should be the same as from K' to K''.

The next axiom, which we refer to as *monotonicity*, says that if we impose a social law on a system, and then impose a second social law on the resulting system, then the total distance between the original system and the final system is at least as large as either of the two individual distances. Formally:

$$\begin{array}{l}(K \sqsupseteq K' \sqsupseteq K'') \to \\ \quad (\delta(K,K'') \geq \delta(K,K')) \wedge (\delta(K,K'') \geq \delta(K',K''))\end{array} \qquad (Mon)$$

The next axiom, *subadditivity*, states that, if we impose two successive social laws, then the distance from the start to the end point is no greater, and may be smaller, than he distance in the two successive social laws. Formally:

$$(K \sqsupseteq K' \sqsupseteq K'') \to \delta(K,K'') \leq (\delta(K,K') + \delta(K',K'')) \qquad (Sub)$$

Related to subadditivity is the *superadditivity* axiom, which states that, if we impose two successive social laws, then the distance from the start to the end point is at least as great, and possibly greater, than the distance in the two successive social laws. Formally:

$$(K \sqsupseteq K' \sqsupseteq K'') \to \delta(K,K'') \geq (\delta(K,K') + \delta(K',K'')) \qquad (Sup)$$

Notice that superadditivity implies monotonicity, although of course the converse does not hold.

If a distance metric satisfies both subadditivity and superadditivity, then it satisfies the following *additivity* axiom:

$$(K \sqsupseteq K' \sqsupseteq K'') \to \delta(K,K'') = \delta(K,K') + \delta(K',K'') \qquad (Add)$$

Since a distance metric satisfies additivity if, and only if, it satisfies both subadditivity and superadditivity, it is not an independent axiom. For this reason, we will not consider it any further.

3.2 Concrete Distance Measures

In this section, we turn our attention to concrete measures of distance, and consider the extent to which these concrete measures do or do not satisfy the axioms we discussed in the preceding section

Kripke Distance: Given that in our model, social laws are sets of transitions to be deleted from a Kripke structure, a very natural measure of distance would seem to be counting how many transitions we are deleting. We call this the *Kripke distance*. Formally, where $K = \langle S, s_0, R, \pi\rangle$ and $K' = \langle S, s_0, R', \pi\rangle$ are Kripke structures such that $K \sqsupseteq K'$, we denote the Kripke distance between K and K' by $\delta_{\mathcal{K}}(K,K')$, and define this value by:

$$\delta_{\mathcal{K}}(K,K') = |R \setminus R'|.$$

Example 2 *Take the system K from Figure 1. Let K_1 be the system that only differs from K by leaving out (s_0,t_1) from K. It is easy to see that $K, s_0 \leftrightarroweq K_1, s_0$, so both systems verify the same formulas, but still, their Kripke distance would be 1.*

Kripke Distance on Minimal Models: We saw that the Kripke Distance may not always make sense, since we do not distinguish between eliminating 'useful' transitions from transitions that 'do not matter'. If we want to give an account of the intuition that every change in properties should be accounted for in the distance metrics, we should look at models that are *contraction minimal*.

This concept can be defined quite generally, for a modal logic with language $\mathcal{L}$, as follows. Let $K = \langle S, s_0, R, \pi\rangle$ be a Kripke model. Define, for any $s, s' \in S$: $s \equiv_{\mathcal{L}} s'$ iff $\forall\varphi \in \mathcal{L}$, $K, s \models \varphi \Leftrightarrow K, s' \models \varphi$. We henceforth take $\mathcal{L}$ to be the language of CTL. Since we know that in this case logical equivalence ($K, s \models \varphi \Leftrightarrow K, s' \models \varphi$) coincides with bisimilarity (Prop. 1), an equivalent definition is: $\equiv_{\mathcal{L}} = \leftrightarroweq$.

It is obvious that $\equiv_{\mathcal{L}}$ is an equivalence relation. So, with $[s]$ we denote $\{s' \in S \mid s \equiv_{\mathcal{L}} s'\}$. The *minimal contraction* $MC(K)$ of K is defined to be the model $L = \langle T, t_0, U, \rho\rangle$ where $T = \{[s] \mid s \in S\}$, $t_0 = [s_0]$, $U[s][t]$ iff $\exists s', t' : s' \in [s]$ and $t' \in [t]$ and $Rs't'$, and $p \in \rho([s])$ iff $p \in \pi(s)$. Let $K = \langle S, s_0, R, \pi\rangle$ be a Kripke model, and $L = \langle T, t_0, U, \rho\rangle$ be its minimal contraction. Then:

1. K, s_0 and L, t_0 are bisimilar;
2. There is no strict submodel $L' \sqsubset L$ that is bisimilar to K.

So we can think of the minimal contraction $MC(K)$ of a pointed structure K as the smallest model L that is bisimilar to K, s_0, a model where no world can be eliminated without losing an expressible property. As an example, the minimal contraction of all structures in Figure 1 is K'' (all with initial state s_0).

We say that a structure K is *contraction-minimal* (or the *bisimulation contraction*) if it is its own minimal contraction.

We can now define $\delta_{min}(K,K')$ for any two models for which $MC(K') \sqsubseteq MC(K)$:

$$\delta_{min}(K,K') = \delta_{\mathcal{K}}(MC(K), MC(K'))$$

For future reference, let $MC(\mathcal{K})$ denote the set of all Kripke structures that are contraction-minimal. This is in some sense not a restriction, since for every structure K, there is a structure $C \in MC(\mathcal{K})$ such that $K \models \varphi$ iff $C \models \varphi$, for all properties φ.

Example 3 *Given the structures of Figure 1, we have that $MC(K) = MC(K') = MC(K'')$ and hence the distance between all of them is 0 according to δ_{min}, even $\delta_{min}(K'',K) = 0$. Note that for any of the other metrics δ in this paper, $\delta(K'',K)$ is not defined (because K'' is a proper subsystem of K). We also have, for any $X, X' \in \{K, K', K''\}$ in Figure 1 and $Y \in \{L_0, L_1, L_2, L_3\}$ from Figure 2 that $\delta_{min}(X,Y) = \delta_{min}(X',Y)$. In fact, we have $\delta_{min}(X,L_i) = i$, for $0 \leq i \leq 3$.*

Feature Sets: Using the measure δ_{min}, we at least know that any change in the model is accounted for in the distance. However, this metric does not discriminate between possible 'un-important' changes, or crucial ones, when going from K to $K' \sqsubseteq K$. The next idea we discuss is to have a set $\mathcal{F}$ of *features*, which represent properties of the system that we are reluctant to lose in implementing a social law. We measure the distance between Kripke structures K and K' as being the number of features of K that are lost in the move to K'. Now, since we have a language specifically intended to capture the properties of Kripke structures, i.e., CTL, it seems very natural to represent features as CTL formulae. Formally, then, a *feature set* $\mathcal{F}$ is a set of CTL formula: $\mathcal{F} = \{\varphi_1, \ldots, \varphi_k\}$. The distance metric $\delta_{\mathcal{F}}$ induced by a feature set $\mathcal{F}$ is defined as follows:

$$\delta_{\mathcal{F}}(K,K') = |\{\varphi \in \mathcal{F} : K \models \varphi\}| - |\{\varphi \in \mathcal{F} : K' \models \varphi\}|.$$

Of course, this definition does not rule out the possibility that some features are false in K but true in K', and hence that the distance between K and K' is in fact negative. For this reason we typically assume feature sets are *normal*, in the sense that all features in $\mathcal{F}$ are satisfied in the initial Kripke structure K, i.e., $\forall\varphi \in \mathcal{F}, K \models \varphi$.

Hierarchical Feature Sets: With feature sets as we have just introduced them, all features are considered equally important: in developing a social law, we will simply be aiming to develop one that minimises the total number of features that we lose. However,

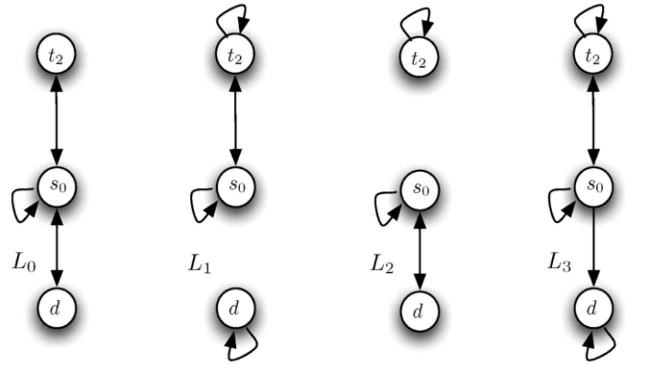

Figure 2. All models L_i are substructures of K, K' and K'' from Figure 1.

in many settings some features will be more important than others. This motivates us to consider the notion of *hierarchical* feature sets. A hierarchical feature set $\mathcal{H}$ is an ordered list of feature sets, i.e., $\mathcal{H} = (\mathcal{F}_1, \ldots, \mathcal{F}_k)$, where $\mathcal{F}_i$ for $1 \leq i \leq k$ is a feature set. Intuitively, the features in $\mathcal{F}_k$ are more important than the features in $\mathcal{F}_{k-1}$, while the features in $\mathcal{F}_{k-1}$ are more important than the features in $\mathcal{F}_{k-2}$, and so on. Given a hierarchical feature set $\mathcal{H} = (\mathcal{F}_1, \ldots, \mathcal{F}_k)$, we define a distance metric $\delta_{\mathcal{H}}$ as follows.

$$\delta_{\mathcal{H}}(K, K') = \begin{cases} \max\{i : \exists \varphi \in \mathcal{F}_i \; K \models \varphi \; \& \; K' \not\models \varphi_i\} & \text{if exists} \\ 0 & \text{else} \end{cases}$$

Thus, according to this measure, a social law will be considered preferable to another social law if it loses features that are lower down the feature set hierarchy $\mathcal{H} = (\mathcal{F}_1, \ldots, \mathcal{F}_k)$. Notice that this metric does not consider *how many* features are changed; it only looks at how far up the hierarchy those changes propagate. For example, suppose that K satisfies already all properties in all $\mathcal{F}_i$'s, then it could be that one social law falsifies all properties in $\mathcal{F}_i$, for all $i < k$ but falsifies no properties in $\mathcal{F}_k$, while another social law falsifies a single property in tier $\mathcal{F}_k$; then the first would still be considered preferable to the second, because it is regarded as causing changes of less significance than the second.

Example 4 *Consider the following objective for a social law:* $\varphi = \mathsf{A}\square(d \to \neg(\mathsf{E}\bigcirc d \wedge \mathsf{E}\bigcirc \neg d))$*, i.e., when the director has the laptop, it should be clear for him where it should go next, there should be no choice. Also assume we have three feature sets:* $\mathcal{F}_i = \{\mathsf{A}\Diamond\mathsf{E}\Diamond p_i\}$*, with* $p_1 = d, p_2 = t$ *and* $p_3 = b$*. So,* $\mathcal{F}_3$*, the most important feature, requires that it should always be possible to return the laptop to the administrator,* $\mathcal{F}_2$ *demands the same for the teachers, and* $\mathcal{F}_1$ *for the director. Let our starting system be* K'' *of Figure 1. Note that* K'' *does not satisfy the objective* φ*, but it does satisfy all three features. Now consider the 4 structures* L_i *from Figure 2. We invite the reader to check that they all satisfy the objective, and also for all of them, we have* $L_i \sqsubset K''$*. So is any of them 'closest' to* K''*?*

We have $L_3 \models \neg\varphi_3$*: it falsifies the most important feature (in* L_3*, the laptop may never return to the base station). So* $\delta_{\mathcal{H}}(K'', L_3) = 3$*. The structures* L_0, L_1 *and* L_2 *satisfy* φ_3*, so their distance to* K *is less than 3. In fact, it is not hard to see that* $\delta_{\mathcal{H}}(K'', L_i) = i$*. In particular,* L_0 *an example of a norm that implements the objective, and is closest to* K''*, in the sense that, like* K'' *itself, it makes all features true.*

Hierarchical Transition Relations: The next metric we introduce can be understood as a semantic counterpart to hierarchical feature sets. Instead of having a hierarchy of feature sets, we separate the transition relation R into a hierarchy, with the idea that being that we consider edges further up the hierarchy to be more significant than edges lower down the hierarchy. Formally, if $R \subseteq S \times S$ is the transition relation of a Kripke structure, then a *hierarchical transition relation*, $\mathcal{R}$, for R is an ordered, indexed list of relations over S (typically sub-relations of R), i.e., $\mathcal{R} = (R_1, \ldots, R_k)$ such that $R_i \cap R_j = \emptyset$ for $i \neq j$ and $R \subseteq R_1 \cup \cdots \cup R_k$. Given a hierarchical transition relation $\mathcal{R} = (R_1, \ldots, R_k)$ for a Kripke structure $K = \langle S, s_0, R, \pi \rangle$, and second Kripke structure $K' = \langle S, s_0, R', \pi \rangle$ such that $K \sqsupseteq K'$, we define the corresponding distance metric $\delta_{\mathcal{R}}(K, K')$ by (the condition C is short for 'if this maximum exists'):

$$\delta_{\mathcal{R}}(K, K') = \begin{cases} \max\{i : \exists (s, s') \in R_i \cap R \; \& \; (s, s') \notin R'\} & \text{if } C \\ 0 & \text{else} \end{cases}$$

Example 5 *In the structures of Figure 1, it might be that transitions* (t, t) *and* (d, d) *have a low priority, since when it comes to fairness, it seems reasonable the users of the laptop don't hang on for it for too long. Also note that this assumption might make more sense in* K *than in* K''*: if we remove* (t_i, t_i) *transitions from* K*, it only means that an individual teacher can not keep the laptop for two time units, but the teachers as a collective would still be able to pass it around.*

Syntactic and Semantic-based Metrics: We have seen semantic-based metrics ($\delta_{\mathcal{K}}, \delta_{min}$ and $\delta_{\mathcal{R}}$) and syntactic-based metrics ($\delta_{\mathcal{F}}$ and $\delta_{\mathcal{H}}$). Both may have their virtues, but on the class $MC(\mathcal{K})$, it appears that the syntactic-based measures are more general than the semantic-based ones. We now show that that is in fact the case. First note that on $MC(\mathcal{K})$, the two measures $\delta_{\mathcal{K}}$ and δ_{min} coincide, since for every $K \in MC(\mathcal{K})$, we have $MC(K) = K$.

Proposition 2 *Consider the two metrics* $\delta_{\mathcal{K}}$ *and* $\delta_{\mathcal{R}}$*, and suppose we only consider Kripke structures that are contraction minimal, i.e., take models from* $MC(\mathcal{K})$*. Then:*

1. *There is a procedure, that, given* $\delta_{\mathcal{K}}$ *and a minimal Kripke structure* K*, generates a set of features* $\mathcal{F}$ *such that for all* $K' \sqsubseteq K$*,* $\delta_{\mathcal{K}}(K, K') = \delta_{\mathcal{F}}(K, K')$*.*
2. *There is a procedure, that, given* $\delta_{\mathcal{R}}$ *and a minimal Kripke structure* K*, generates a hierarchical set of features* $\mathcal{H}$ *such that for all* $K' \sqsubseteq K$*,* $\delta_{\mathcal{R}}(K, K') = \delta_{\mathcal{H}}(K, K')$*.*

Note that the converse of Proposition 2 does not hold. Suppose $\mathcal{F} = \{\mathsf{E}\bigcirc(p \wedge q), \mathsf{E}\bigcirc p\}$. It is well possible that $\mathcal{F}$ is true in K, while there are two substructures K' and K'', both obtained from K by deleting one transition (i.e., $\delta_{\mathcal{R}}(K, K') = \delta_{\mathcal{R}}(K, K'') = 1$), while one substructure loses two features from $\mathcal{F}$, the other only one.

3.3 Properties of Distance Metrics

Now that we have a set of axioms and a set of concrete distance metrics, it is natural to evaluate the metrics against the axioms. Table 1 summarises these results.

Proposition 3 *The characterisations of distance metrics and the axioms they satisfy given in Table 1 are sound.*

So δ_F and δ_{min} are two metrics that satisfy all axioms. Note the rather different behaviour between $\delta_{\mathcal{R}}$ and $\delta_{\mathcal{H}}$: despite their seemingly similar definitions, they have very different axiomatic properties.

4 Conservative Social Laws

We can now formulate some computational problems, which we collectively refer to as the CONSERVATIVE SOCIAL LAW problems. When considering these problems, it should be understood that the distance metric δ is one of the distance metrics discussed above.

	Distance Metrics				
	Semantic			Syntactic	
Axiom	$\delta_{\mathcal{K}}$	δ_{min}	$\delta_{\mathcal{R}}$	$\delta_{\mathcal{F}}$	$\delta_{\mathcal{H}}$
(*Equal*)	yes	yes	yes	yes	yes
(*Bisim₁*)	no	yes	no	yes	yes
(*Bisim₂*)	no	yes	no	yes	yes
(*Bisim₃*)	no	yes	no	yes	yes
(*Mon*)	yes	yes	yes	yes	no
(*Sub*)	yes	yes	yes	yes	no
(*Sup*)	yes	yes	no	yes	no

Table 1. Some distance metrics and the axioms they satisfy.

CONSERVATIVE SOCIAL LAW (DECISION):
Instance: Kripke structure $K = \langle S, s_0, R, \pi \rangle$, CTL formula φ, distance metric δ, and bound $b \in \mathbb{R}_+$.
Question: Does there exist a normative system $\eta \in N(R)$ such that $K \dagger \eta \models \varphi$ and $\delta(K, K \dagger \eta) \leq b$?

The optimisation variant of the problem is:

CONSERVATIVE SOCIAL LAW (OPTIMISATION):
Instance: Kripke structure $K = \langle S, s_0, R, \pi \rangle$, CTL formula φ, and distance metric δ.
Question: Compute some η^* satisfying:

$$\eta^* \in \arg \min_{\eta \in N(R), K \dagger \eta \models \varphi} \delta(K, K \dagger \eta).$$

Thus, the aim of the optimisation problem is to actually find an effective social law for the objective that minimises the δ-distance.

It is not our aim in the present paper to discuss these problems in detail. However, it is not hard to see that all the problems inherit the NP-hardness of their "parent" problem (i.e., the problem of checking whether, given a Kripke structure K and CTL objective φ, there exists a social law η such that $K \dagger \eta \models \varphi$). It is also similarly easy to see that, using similar arguments to those presented in [2], the CONSERVATIVE SOCIAL LAW (OPTIMISATION) problem for feature set based distance metrics can be solved with a "small" (logarithmic) number of queries to an NP-oracle (in technical terms, it is $\mathrm{FP}^{\mathrm{NP}[\log_2 |\mathcal{F}|]}$-complete). However, we will leave a detailed study of the computational complexity of these problems for future work.

5 Discussion

Our starting point was to represent the behaviour of a multi-agent system by a Kripke structure K, where the accessibility relation models possible transitions in the system. We then interpreted a social law as a restriction on the possible transitions, leading us to talk about possible subsystems $K' \sqsubseteq K$, where the idea is that K' could be the implementation of a possible social law, applied to K. The main question we address then in this setting is to shed some light on the question whether we can say that one subsystem K' may be 'better' than another subsystem K'' of K. And the intuition we tried to capture in the possible answers to this question is that one norm η' might be favoured over another norm η'' because the changes it brings about in K, are smaller than those that η'' brings about. In other words, the distance from K to $K' = K \dagger \eta'$ is smaller than the distance from K to $K'' = K \dagger \eta''$. This, in turn, requires a notion of distance over Kripke structures, which we here formalised as a metric.

We formulated some general principles such a metric could satisfy. Since the notion of bisimulation on finite Kripke models captures 'when two models are the same', it should come at no surprise that this notion plays a prominent role in our axioms. Subsequently, we formulated a number of concrete metrics to measure the distance between Kripke structures: some of them focused on syntactic properties (formulas) that they satisfy or not and others on the change in structure of the underlying graph of the models.

Similar ideas are pursued in [3], which proposes to use distance metrics to measure the difference between possible *protocol modifications* in order to avoid a modifications that are "far" from some "desired" specification. These metrics are defined on a space of "specification points" of protocols, while the metrics we discuss in the current paper are defined on a very general model class, namely Kripke models. Furthermore, [3] does not define or discuss concrete metrics or abstract properties or axioms of metrics; metrics are assumed to be "application specific" and it is assumed that there exists a "logic programming implementation of a given metric". [7] considers *minimal* social laws, which are social laws that constrain the behaviour of agents as little as possible. This is very similar to our Kripke distance metric $\delta_{\mathcal{K}}$. However, our axiomatic treatment, and the other metrics we consider, are different. Finally, the main problems considered in this paper are somewhat reminiscent of some concepts in *belief revision*. In belief revision, the effect of operations like revision, contraction or expansion is also governed by minimal change, and the notion of *entrenchement* in belief revision has a similar flavour as our notion of metric. The notion of expansion in belief revision ('add φ to the belief set') is related to our notion: 'find a social law η for K such that $K \dagger \eta \models \varphi$'. Given this analogy, it is interesting to not only look at norms that *restrict* the behaviour of agents, but also look at modifications that *add* transitions, (or indeed states), to a structure. The interpretation of such a modification would be: although the current system might not cater for it, it should become true that φ (where φ is a CTL formula). In our example model K'' for instance, a modification of the system might require that it should always be possible that a teacher hands the data projector directly to the director. Our notion of metrics might again be employed to reason about 'minimal modifications' in this sense. Obviously, this would become rather more complex if the change would involve the addition of new states, but it is not difficult to imagine how K in Figure 1 might evolve from K'' by an objective that says 'there should be several teachers and they should be able to pass the laptop around'.

REFERENCES

[1] T. Ågotnes, W. van der Hoek, J. A. Rodriguez-Aguilar, C. Sierra, and M. Wooldridge, 'On the logic of normative systems', in *Proceedings of IJCAI*, (2007).

[2] T. Ågotnes and M. Wooldridge, 'Optimal social laws', in *Proceedings AAMAS*, (2010).

[3] Alexander Artikis, 'Dynamic protocols for open agent systems', in *Proceedings of AAMAS*, (2009).

[4] G. Boella and L. van der Torre, 'Delegation of power in normative multiagent systems', in *Proceedings of DEON 2006*, (2006).

[5] M. C. Browne, E. M. Clarke, and O. Grümberg, 'Characterizing finite kripke structures in propositional temporal logic', *Theoretical Computer Science*, **59**, (1988).

[6] E. A. Emerson, 'Temporal and modal logic', in *Handbook of Theoretical Computer Science Volume B: Formal Models and Semantics*, 996–1072, Elsevier, (1990).

[7] D. Fitoussi and M. Tennenholtz, 'Choosing social laws for multi-agent systems: Minimality and simplicity', *Artificial Intelligence*, **119**(1-2), 61–101, (2000).

[8] Y. Shoham and M. Tennenholtz, 'On the synthesis of useful social laws for artificial agent societies', in *Proceedings of AAAI*, (1992).

[9] W. van der Hoek, M. Roberts, and M. Wooldridge, 'Social laws in alternating time: Effectiveness, feasibility, and synthesis', *Synthese*, **156**(1), 1–19, (2007).