Analysing Norms with Transition Systems

Trevor BENCH-CAPON
Department of Computer Science, The University of Liverpool, UK

Abstract.
The design and analysis of norms is a somewhat neglected topic in AI and Law. In recent years powerful techniques to model and analyse norms have been developed in the Multi-Agent Systems community. In this paper I consider these techniques from and AI and Law perspective, and suggest a framework for the exploration of these issues.

Keywords. Norms, Models, Legislation Design

1. Introduction

Sometimes an interesting idea appears but remains largely unexplored until the technology to handle it is developed and it reappears in a new context, One such idea can be found in Winkels and den Haan's 1995 ICAIL paper on automatic legislative drafting [12], in which they explore the roles that "deep structures" (formal descriptions of the environment in which the norms will operate) might play in drafting legislation. The paper demonstrates, among other things, how the same objectives can be achieved by different norms, according to the position taken on four different dimensions: the default status (desirable or undesirable), the viewpoint (legislator, norm subjects), the level of abstraction (high or low) and the favoured deontic operators (permission, prohibition or obligation). This paper has attracted too little attention (at the time of writing it has only eight citations on Google Scholar), but this is perhaps because the time was not then right. The development of Multi-Agent Systems (MAS), and within MAS the emergence of Electronic Institutions (e.g. [3]) and the general idea that open agent systems can be controlled by norms (which can be traced back to [9]), has led to a great revival of interest in these topics. Thus the design and specification of norms for multi-agent systems is now a much studied topic. For one important and representative strand see e.g. [11], [2], [1]). The deep model in the MAS systems is provided by a transition diagram which indicates the effects of an agent's action in the context of the system, and norms are specified as constraints on such behaviour. Specification in [11] and subsequent papers is in terms of Kripke structures and Computation Tree Logic (CTL) [6], but we will not use any particular formal structures in this paper.

The term "norm" will be used in this paper to refer to any statement intended to influence behaviour, including e.g. laws, social conventions, and rules for electronic institutions. Typically there will be an authority which issues the

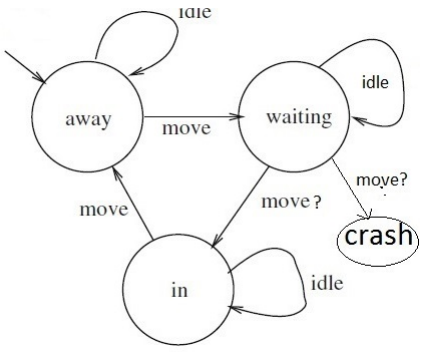

Figure 1. Individual perspective of train with crash

norm (e.g. the legislature, systems designers) and a group of subjects intended to comply with it (e.g. citizens, software agents). The MAS work is directed towards software systems and its concerns are mainly software engineering concerns, such as effectiveness, efficiency and liveness. Our discussion will return to the theme of [12]: we will be thinking mainly in terms of human agents and their legal systems.

2. Models

In this paper we will mainly discuss a model taken from [11]. We will also use an important motivating example from [10], Ullmann-Margalit's seminal work on norms, in which she uses a variety of public goods games to discuss social norms and their origins. Our main example is of two trains travelling in different directions around a circular track. Mostly the tracks are separate, but at one point both tracks pass through a narrow tunnel and a crash will occur if both trains are in the tunnel at the same time. The trains can either move forwards or remain where they are. With respect to the tunnel, a train may be in the tunnel, waiting to enter, or away from the tunnel. Note that this description is very abstract: it may be that there are several points away from the tunnel any of which can be considered *away*, and this further level of detail is something we might need to consider in some circumstances.

Two viewpoints are relevant: that of the trains (the norm subjects) and that of the system as a whole. We can represent these viewpoints using transition diagrams. Figure 1 shows the individual perspective (ignore for the moment the "crash node" and the question marks), while Figure 2 shows the system perspective. The system perspective is more complicated because each train can be in any of three states, giving rise to nine states instead of the original three, and because the transitions must now consider the *joint actions* of *both* agents, so that we have nine actions instead of three. The system perspective is also more informative: we can see that the state where both trains are *in* has no exit: so that idleness is enforced in that state. That moving in the waiting state may or may not result in a crash is not captured by Figure 1 until we add the "crash" node and the question marks.

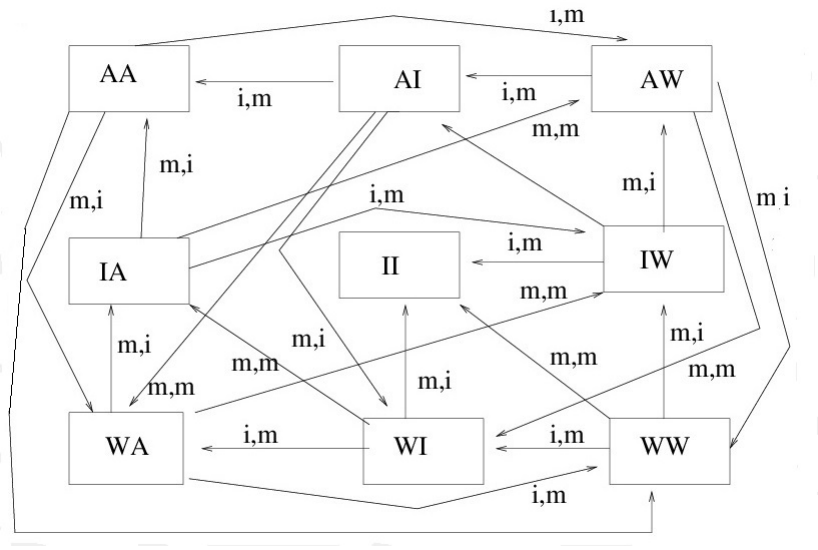

Figure 2. System perspective on trains

The technique of [11] is to identify an objective (e.g. avoiding the state where both trains are in the tunnel), and express a norm as a behavioural constraint which will ensure that the objective is achieved. The behavioural constraint used in [11] is to prohibit both trains from waiting in the tunnel and to prohibit the eastbound train from entering the tunnel, except where it is waiting and the westbound train is away. This norm gives priority to the westbound train. The norm can be represented by removing the transitions which include waiting in the tunnel and the eastbound train moving from states WI and WW. Model checking tools can now be used to verify the effectiveness of the norm.

The example from [10] is the *Machine Gunner's Dilemma*, a variant on the Prisoner's Dilemma. I will extend the example slightly here. The idea is that there are two machine gunners who need to delay the enemy if reinforcements are to arrive, and victory is to be won. If both run away, the enemy will win, and both will be taken as prisoners of war. If one stays and one runs, the deserter will escape unharmed while the other will delay the enemy sufficiently even though in the end he will die (remember the Alamo). If both stay there is a possibility that both will survive (as at Rorke's Drift) or that both will die, but only after having succeeded in their mission (as with the three hundred of Thermopoylae), according to the skill and determination of the enemy.

The objective of the army will be that at least one, or preferably both, the machine gunners remain at their post. The machine gunners most want to avoid dying, but also wish to avoid capture. The dominant game-theoretic strategy is to run, but this is not best for either the gunners themselves or the army. Ullmann-Margalit discusses three possibilities. One is simply to make desertion impossible, by chaining the gunners to their guns: this is very similar to the usual MAS approach - the undesired actions are made unavailable. The second is to impose sanctions: if deserters are shot, then they will both choose to stay as this is the

only possibility of survival. This works not by removing actions but changing the states: capture and escape are no longer possible. The third possibility is the use of norms to change the preferences of the agents: two military sayings are *Death or Glory*[1] and *Death Before Dishonour*[2]. The idea is to instill sufficient trust in or loyalty to their comrades or sufficient confidence in the fighting qualities of their regiment that they will choose to remain at their posts. Military training typically attempts all three through teaching military history, team building etc. The sanctionless military norms are unwritten, but also have implications for legal norm formulation.

The following three sections will discuss enforcement, sanctions and norms which rely neither on enforcement nor sanctions.

3. Enforcement

Enforcement is the method normally adopted in MAS specification, and is quite natural, and relatively easy to implement, in a software system. For example in an Electronic Institution users may be offered a menu of actions, and the forbidden actions may be greyed out, or otherwise made non-executable in certain contexts. But in the real world it is harder to enforce norms in this way: it is difficult to make it impossible for the eastbound train to move and the idea of chaining soldiers to their guns is surely contrary to the Geneva Convention. Moreover, it is usually not even desirable to make the prohibited action *impossible*: sometimes there is an overriding reason to violate the norm. The use of deontic modalities (as opposed to alethic modalities which result from making prohibited actions actually impossible) is important when violations need to be considered [7]. In general the option of rendering actions impossible is not available to legislatures, although it remains a reasonable way to model a fully complied with norm, and so remains a convenient way of testing the effects of a norm, assuming universal compliance.

We can consider the solution of [11] and other MAS systems on the dimensions of [12]. First the viewpoint is that of the system: we need to consider both trains and their interaction. Second the default is that all actions are permitted. Third the level of abstraction is high: it is useful to be able to describe a whole set of possible positions as "away". Lastly the preferred modality is prohibition: it is easier to remove actions. An alternative (for a software system) would be to start with no actions available to the agents, and then add permitted actions and their transitions. This is unrealistic in an open system where the agents participating are designed by people other than those running the institution, and so their capacities are beyond the control of the institution. Even more is this true when the norm subjects are real people. Normally prohibition (removing an action) will be simpler than obligation, which may involve removing a large number of actions (e.g. if there are five actions available to an agent, we would need to remove four to represent the obligation).

[1] The Regimental Motto of the 17th/21st Lancers, now part of the Queen's Royal Lancers
[2] A saying apparently widely used in the US Marines Corps

We can also see from this example that the rule and exception structure widely found in law is a very natural and concise way of specifying norms. In [12] the norm is specified on a state by state basis. But this is not feasible, except for very limited problems where the number of states is small. Much more concise is to say that *a train must move except if it is eastbound and waiting when it must idle except when the westbound train is away when it must move*. Note, however, that the preferred modality has now become obligation.

One should also note the role of abstraction. In the train example all locations of the train other than *in* the tunnel and *waiting* to enter are represented as the single state *away*. If *away* is a single location the eastbound train will not have very long to wait for the westbound train to reach the away state, and can see that the westbound train is away, and so will know when it is permitted to enter the tunnel. But at a more detailed level, we might find that the circuit was much longer so that there were many locations contained within the "away" abstraction. In this case the westbound train might be invisible to the eastbound train for long periods, making it uncertain whether it can enter the tunnel, and making the wait until it can be certain rather long. Two points are important: one is to frame the norms so that the subjects can know whether they are complying with them or not, and the other is that an asymmetry may appear to be acceptable at one level of abstraction but not at an another. We must therefore, when framing norms, be careful to use the appropriate level of abstraction, both to facilitate compliance and to frame acceptable norms.

Before moving on to sanctions, observe that in MAS treatment of norms the objective is to prevent a *state*, but the norm forbids an *action*. This is also true of norms in general (consider the ten commandments). Why is this so? I believe the main reason is the uncertainty of the effect of actions, because agents do not know in which joint action they will participate. Because we do not know for certain whether an action will lead to the unwanted state it is necessary to prohibit any action which *could* result in the unwanted state. In the train example, moving into the tunnel will often not result in a crash: waiting is only necessary if the other train enters the tunnel at the same time. But to *ensure* that this does not occur, that action must be prohibited unless the other train is *away*.

4. Sanctions

The approach of enforcement left the transition diagrams unchanged, except for removing certain transitions so as to make certain states unreachable. Sanctions operate very differently since they add *additional* states, actors and transitions to the diagram. Suppose we wish to supplement the norm proposed for the trains example discussed above by fining the eastbound train if it violates the norm. We need an additional agent to impose the sanctions, and we need to represent the action of imposing the sanction and additional states to record its effect. Thus the system perspective transition diagram ceases to be that of Figure 2 and becomes the diagram shown in Figure 3 (transitions following fines are omitted). If we can assume that the sanctions are enforced, we can remove any states reachable only by unpunished violation, which should, of course, include all the undesirable states.

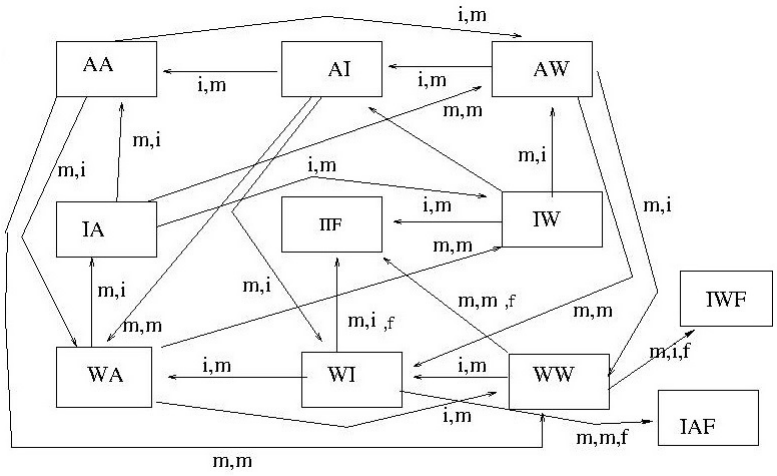

Figure 3. Trains with Sanctions

We can now classify the states which result from sanctions to reflect the different ways in which sanctions can operate. It may be that the receipt of the fine makes the situation actually desirable. An example of this may be library fines: the library may well prefer to receive the income for the book than have it unborrowed on their shelves, and the library user may well be willing to pay to retain the book, or to return it at a more convenient later date. In this case the sanction makes everyone happy: it is less a fine than a charge for an extra service. The sanction works by making all states desirable: a win-win situation. If there is no level of fine which can make everyone happy, the fine should be sufficient that at least the authority is made happy. Certain motoring offences, such as parking (and, some motorists claim, speeding) may fit this. The fines provide revenue, which the authority welcomes, but which the users resent. But this does not matter: users can choose to obey the norm, and the states so reached will also be acceptable to the authority. In some cases norm subjects may disagree as to whether the sanction serves as a fee or as a discouragement. Some people see parking fines as worth paying and will not worry about accumulating parking tickets, whereas others will avoid parking illegally because of the fines.

The final possibility is that the consequences are so bad that no level of sanction is able to compensate. Here the purpose of the sanction is not to *compensate* for the violation and so make it acceptable, but to *deter*. The crash situation in the train example is one such. The objective is to ensure that there are no crashes: any revenue that might result from violations that do not lead to crashes is neither here nor there, since the important thing is to eliminate violations and hence the possibility of a disaster. Here the sanctions have to be such as to make the situation clearly undesirable to all norm subjects, so that they will choose to comply with the norm, unless there is some abnormal situation which makes violation absolutely necessary. Whereas fines are natural sanctions for some offences, they do not work so well as a deterrent: some people may always consider the price worth paying. Thus such violations tend to be associated with custodial

sanctions, which are a very different type of sanction, since imposing them represents a cost rather than compensation. For this reason they should be reserved for avoiding situations which are so undesirable that they cannot be compensated for.

The above assumes that the sanctions will be always enforced. This is often the case, where detection is not a problem: library fines can always be exacted when the book is returned. In the train example a single CCTV camera would ensure that violations were detected. But speeding offences often go undetected, and speed limits are frequently exceeded as a result. Thus the norm subject can be uncertain whether violating the norm will reach the state where the penalty is imposed or not. If the subject is to be deterred by the penalty, it will need to assess the likelihood of detection: if it is certain that the sanction will be imposed, the norm will be obeyed and if it is certain not to be imposed the norm will be violated, but if the sanction is a more or less likely consequence, the norm subject must weigh the benefits of violation against the risk of the sanction. In this case the authority must make the sanction larger than it would be if detection was thought certain, so that the *expected value* will be undesirable to the norm subject, even with the smaller risk of detection.

5. Norms Without Sanctions

In this section we will consider norms which are intended to be obeyed without enforcement and without sanctions. These are norms which secure compliance of the third type considered by Ullmann-Margalit when discussing the Machine Gunner's Dilemma in [10]. The first norms here are norms of *co-ordination*. Norms are needed when the norm subjects are confronted by a choice, and one choice will, or may, lead to an unacceptable situation. Often the norm subject will prefer the choice which may lead to the unacceptable situation: in the train example, the train wants to move, but there is a need to avoid crashes. In such cases norms serve to remove the choice (enforcement) or to alter the outcomes by imposing sanctions so that norm subject no longer wishes to make the undesirable choice.

But in some cases the norm subject is indifferent as to the choice, and the undesirable situation arises from two or more agents making an unfortunate choice. The classic example is norms relating to the side of the road that cars should drive on. Drivers do not particularly mind which side of the road they are required to drive on: some countries say left and others say right and no one argues for change. What matters is that everyone makes the same choice. Thus once the norm has been established, so that the behaviour of others is known, compliance becomes the only sensible choice: agents obey not because they cannot do otherwise, nor because they wish to avoid sanctions, but because it is in their interest to do so.

The train example can be seen as a coordination norm if we assume that, although both trains want to move, they are more concerned with safety. If this is the case, without the norm neither will dare to enter the tunnel and the rail system will grind to a halt. But if the norm is promulgated, and the trains assume that it will be obeyed, the westbound train can enter with confidence and the eastbound will be happy to suffer a short wait to ensure its safety. Similar solutions are used

for narrow bridges on roads: by giving priority to one direction, accidents can be avoided and the certainty the norm affords more than compensates for any delay for the unfavoured direction.

This leads to the notion of reasoning about the behaviour of others, since what the other chooses often determines whether the outcome will be good or bad. The train problem was discussed in [5], which used the machinery of [4] to describe the reasoning of agents considering the choices. There the transitions are labelled with the values promoted or demoted by following them, and the choice is made according to which set of values is preferred. In the train examples the relevant values are *Progress* and *Safety*. Progress will be promoted by a transition in which a train moves, and Safety will be demoted by a transition into the state where both are in the tunnel. If the trains prefer Safety, there will be standoff as they wait for one another, unless there is a co-ordination norm.

In this case the norm works by altering the labels on the transitions. Effectively we can see the norm as introducing one or more additional values, such as *Compliance*, promoted by complying with the norm and demoted by violating it. Such a value can itself be a reason for action, and a strong one: some people will obey the law simply and purely because it is the law. (Kant argued in [8] that this was the only moral reason to obey the law, and it is very often at least a contributory reason.) This additional value makes obeying the law more attractive (unless breaking the law is seen as a value in itself, as with the adolescent playing chicken). An alternative is to make the undesired choice less attractive, which can also be effected by a norm: if society attaches a stigma to breaking the law, then the norm violator's reputation will suffer. There is a similar contrast in the slogans mentioned in relation to the Machine Gunner's Dilemma: *death or glory* gives a value-based reason to comply, while *death before dishonour* gives a value-based reason not to violate. Different types of norm may relate to different values: obeying the law is a different value from conforming to a social convention: either may be rated more highly by an agent and they may even conflict (e.g in gang culture). The same norm can work differently for different agents: the norm will still work provided that one or other of the values, or a combination of the two, can serve to direct the agent into the desired choice.

6. Summary

Norms can be modelled using transition systems. From the above we can see that we normally need to adopt a system perspective, so that we can represent the effects of the actions of a group of agents acting simultaneously, since the outcome of an action will very often depend on what other agents choose to do. We therefore typically perform our analysis on a transition diagram in which the transitions relate to the joint actions of a group of agents, such as the Alternating Action-Based Transition Systems (AATS) of [11].

Norms are promulgated by an authority (legislature, social group, software designer, etc) when agents may act so as to realise a state undesirable to the authority (and perhaps also the agents themselves). This can be because of uncertainty, which may relate to uncertainty as to the current state (the train does

not know whether the other train is waiting or not), or uncertainty as to what the other will do, or because the effect of the actions is indeterminate and the agents consider the risk acceptable. The norm subjects may, however, wish to enter a state the authority finds undesirable because it is personally desirable to them. In summary a norm is required if an authority wishes to avoid a state and there is one or more norm subjects who can act so as to enter the state, and:

1. The norm subject desires to enter the state
2. The norm subject does not know whether its action will cause the state to be entered because
   (a) It is unsure of the current state;
   (b) It is unsure of the outcome of its action because of uncertainty as to what other agent will do;
   (c) It is unsure of the outcome of its action because the result of its action is indeterminate.

Using the general notion of exploring norms through models, we have seen that there are several ways of representing a norm in a transition system.

1. By removing transitions representing prohibited actions;
2. By removing all the transitions from a given state except those representing an obligatory action;
3. By imposing sanctions for violation. This involves modifying the transition diagram to include an additional agent and appropriate transitions to impose the sanction and additional states to represent that the sanction has been imposed. Sanctions may be designed to produce
   (a) A situation acceptable to both authority and norm subject. Such sanctions are akin to *fees*.
   (b) A situation acceptable to the authority. Such sanctions are intended to be *compensation*.
   (c) A situation unacceptable to the norm subject. Such sanctions are intended to be *deterrent*.
4. By labelling the transitions to represent promoted and demoted values (as in the AATS+V of [4]). The norm will then extend the set of possible labels (since the values will now include at least *Compliance* in addition to the existing values) and these new labels must be applied to the diagram. These values will then motivate the agents to avoid the undesirable situation. They can work
   (a) By removing uncertainty as to what the others will do (*co-ordination* norms).
   (b) By making the undesired state less attractive to the norm subject.
   (c) By making a choice other than the undesired state more attractive to the norm subject.
   (d) A combination of (4b) and (4c).

Of these, (1) and (2) represent *enforcement*. This is a simple and effective way of modelling a system in which the norm subjects can be forced to comply

(such as an Electronic Institution or other software system) or of determining the properties of the system under the assumption that all the norm subjects do, in fact, comply. It is, however, unsuitable if we want violation to be possible (either because there may be certain situations in which violation is acceptable, or even desirable), or where forced compliance is impossible (as in the normally the case in human societies). (3) represents sanctions and we can distinguish three sanction types (3a), (3b) and 3(c). (4) represents norms which require neither enforcement nor sanctions: it is intended that the agents will adopt the norms for their own reasons. This requires either the right sort of situation (so that what is needed is co-ordination: in particular the norm subjects themselves must wish to avoid the situation the authority finds undesirable) for (4a), or that the norm subjects have the appropriate value preferences (for (4b), (4c) and (4d)).

The design and analysis of norms is a potentially important topic for AI and Law, but has received very little attention in that field over the last two decades. In contrast the area has become increasingly active in the MAS field. In this paper I have considered the techniques used for MAS from an AI and Law standpoint, with the intention of providing a framework which will encourage further developments of the topic in AI and Law.

References

[1] T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge. Power in normative systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 145–152. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[2] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Robust normative systems. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 747–754. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[3] J. Arcos, M. Esteva, P. Noriega, J. Rodríguez-Aguilar, and C. Sierra. An integrated development environment for electronic institutions. In *Software agent-based applications, platforms and development kits*, pages 121–142. Birkhäuser Basel, 2005.

[4] K. Atkinson and T. Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artif. Intell.*, 171(10-15):855–874, 2007.

[5] K. Atkinson and T. Bench-Capon. States, goals and values: Revisiting practical reasoning. In *Proceedings of 11th Intl. Workshop on Argumentation in Multi-Agent Systems*, 2014.

[6] E.A. Emerson. Temporal and modal logic. *Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics (B)*, 995:1072, 1990.

[7] A. Jones and M. Sergot. Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1(1):45–64, 1992.

[8] I. Kant. *Groundwork of the Metaphysic of Morals*. Harper Perennial modern thought. Harper Collins, 2009.

[9] Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies (preliminary report). In *AAAI*, pages 276–281, 1992.

[10] E. Ullmann-Margalit. *The emergence of norms*. Clarendon Press Oxford, 1977.

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

[12] R. Winkels and N. Den Haan. Automated legislative drafting: Generating paraphrases of legislation. In *Proceedings of the 5th International Conference on Artificial Intelligence and Law*, pages 112–118. ACM, 1995.