Coalition Formation through Motivation and Trust

Nathan Griffiths
Dept of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
nathan@dcs.warwick.ac.uk

Michael Luck
Dept of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
mml@ecs.soton.ac.uk

ABSTRACT

Cooperation is the fundamental underpinning of multi-agent systems, allowing agents to interact to achieve their goals. Where agents are self-interested, or potentially unreliable, there must be appropriate mechanisms to cope with the uncertainty that arises. In particular, agents must manage the risk associated with interacting with others who have different objectives, or who may fail to fulfil their commitments. Previous work has utilised the notions of motivation and trust in engendering successful cooperation between self-interested agents. Motivations provide a means for representing and reasoning about agents' overall objectives, and trust offers a mechanism for modelling and reasoning about reliability, honesty, veracity and so forth. This paper extends that work to address some of its limitations. In particular, we introduce the concept of a *clan*: a group of agents who trust each other and have similar objectives. Clan members treat each other favourably when making private decisions about cooperation, in order to gain mutual benefit. We describe mechanisms for agents to form, maintain, and dissolve clans in accordance with their *self-interested* nature, along with giving details of how clan membership influences individual decision making. Finally, through some simulation experiments we illustrate the effectiveness of clan formation in addressing some of the inherent problems with cooperation among self-interested agents.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence

General Terms

Algorithms, Experimentation

Keywords

Cooperation, coalitions, clans, trust, motivation

1. INTRODUCTION

Autonomous agents situated in dynamic multi-agent environments may need to cooperate to achieve their goals. Differences in capabilities, knowledge and resources mean that an individual may not be able to perform a particular task alone. In this paper we consider cooperation among autonomous agents, focusing in particular on *motivated* agents, i.e. agents whose autonomy results from motivations. Motivations can be thought of as an agent's high level desires, guiding *all* aspects of its behaviour. Ascribing motivations to an agent gives considerable flexibility and robustness, and several researchers suggest that some form of motivation is necessary for achieving true autonomy [6, 15, 18].

Existing models of cooperation can be broadly divided into two strands: teamwork and coalition formation. Typically, teamwork is task-based and concerned with attaining cooperation in the short-term to achieve a specific task. Approaches to teamwork, generally focus on assigning tasks to agents and coordinating actions. By contrast, coalition formation is concerned with establishing a group of agents in pursuit of a common aim or goal, either to achieve goals that cannot be achieved alone or to maximise net group utility. The cooperative groups formed through coalition formation tend to be longer term than those formed by teamwork approaches, and the members of a coalition are generally expected to be inclined to act on behalf of others in the coalition, even when there may be more (direct) utility from acting alone. By virtue of their task-based nature, teamwork approaches tend to require all members of the team to remain involved, while coalitions often allow agents to join or leave the coalition at any time (depending on their other commitments).

Task-based teamwork approaches to cooperation are particularly suited to situations in which agents have common goals, since teams are formed for the achievement of specific tasks [20]. However, this gives rise to the main limitation of a task-based approach, that unless agents have common goals *at the time of forming the team*, they will not cooperate, regardless of whether their goals are similar *in the long-term*. In the long-term, if their goals are similar (but out of step in terms of time), they may benefit *overall* from cooperating even when there is no direct *immediate* benefit. Standard teamwork approaches do not consider such missed opportunities for cooperation. Furthermore, task-based approaches typically result in cooperation for a single goal, requiring a team to be created for subsequent goals even if the team members are the same. Taking a long-term view can significantly reduce the computation required in achieving cooperation for a goal, especially in environments containing a significant number of agents. Thus, we need an approach to cooperation that takes a long-term view to avoid missed opportunities and reduce the computation involved in forming teams.

Coalition formation typically takes a long-term view, although often still directed towards a particular goal. Here, the benefit to an individual of joining a coalition tends to be assessed according to the utility gained by the group, with respect to achieving a

AAMAS'03, July 14–18, 2003, Melbourne, Australia.

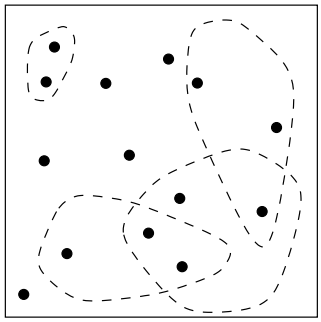

Figure 1: The partitioning of the space of agents

goal, if a coalition is formed [3, 13, 19]. The utility afforded to a group by achieving a goal is often taken to be the sum of the utilities afforded to the individuals involved. This group utility can be fairly divided between the coalition members, according to the proportion of each agent's contribution to the group achievement of the goal. Calculation of this utility must either be performed at an external system level, or the agents must (at least partially) reveal their utilities. Although this allows for self-interest, since agents only join a coalition if they gain sufficient utility, this does not account for any motivational reasons an agent might have. Furthermore, the motivational value[1] afforded to a group cannot be determined by combining individual motivations. Motivations represent an agent's high level desires and, since different agents have different desires, these desires differ, so motivational value cannot be compared across agents. Consequently, utility-based coalition formation, which considers group utility, cannot be directly applied to motivated agents; a motivation-based approach is needed.

A technique related to coalition formation is *congregating*, which aims to reduce the cost of locating others to work with, such that rather than searching the whole population, agents congregate into groups and search within the congregation [4, 5]. Congregations avoid some of the limitations of task-based approaches since a single goal connecting members of a congregation is not required. Existing literature on congregations does not consider motivational value, but is instead focused on enabling agents to form groups with particular *similarities*. Agents are divided into *labellers* and *congregators* such that the former label their congregations so as to attract similar congregator agents. Our work has a broader aim than reducing the search and advertising cost, and although we can utilise the concept of congregations, we cannot use them directly in enabling cooperation between motivated agents. In particular we are concerned with constructing a model of cooperation that avoids missed opportunities for cooperation and redundant computation in re-constructing similar groups, and accounts for the motivational aspects of cooperation.

In this paper we take a *medium-term* view of cooperation and describe a mechanism for the formation of medium-term coalitions, which we call *clans* to distinguish them from existing approaches. We propose a mechanism through which autonomous, self-interested, agents form clans to enhance their individual goal achievement, with increased group performance occurring as a useful consequence. We focus, in particular, on the roles played by the notions of *motivation* and *trust* in the formation of clans. Trust is used as a mechanism for modelling and reasoning about others' honesty and reliability, etc.; prior to cooperating, an agent considers the trustworthiness of the others involved. By forming clans, the space of agents is partitioned according to the clans an agent belongs to. The form of the resulting structure is illustrated in Figure 1, in which the dots represent agents, and agents contained within a dashed line are considered to belong to a clan. An agent can to belong to several clans, with each clan being given equal importance when reasoning about them. Thus, the resulting organisations have a flat structure.

[1]Motivational value can be thought of as analogous to utility. An action, goal, or situation may be of benefit to one or more of the agent's motivations, and we call this benefit the motivational value. More details can be found in [10].

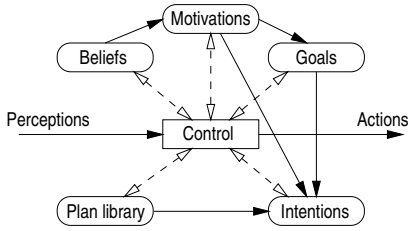

Figure 2: mBDI agent architecture

2. MOTIVATION AND TRUST

2.1 Motivated Agents

It is useful, before continuing, for us to be precise about the kind of agent that we are concerned with. Specifically, we adopt a BDI-based approach of taking an agent to comprise: *beliefs* about itself, others and the environment; a set of *desires* representing the states it wants to achieve; and *intentions* corresponding to the plans adopted in pursuit of these desires [2]. In addition to the traditional BDI model however, we concur with the views of some that motivation is an extra component required to achieve true autonomy in such agents [15].

Motivations are high-level desires that characterise an agent; they guide behaviour and, at a fundamental level, control reasoning. In addition to causing the generation and subsequent adoption of goals, motivations direct an agent's reasoning and action at both an individual and a cooperative level. An agent has a fixed set of motivations, each with a particular intensity which varies according to the current situation. We represent a motivation by a tuple (m, i, t, f_i, f_g, f_m), where m is the name of the motivation, i is its current intensity, t is a threshold, f_i is the intensity update function, f_g is the goal generation function, and f_m is the mitigation function.

The intensity of an agent's motivations changes in accordance with its beliefs (as determined by f_i), which in turn are determined by perceptions. When the intensity of a motivation exceeds its threshold, the motivation uses the function f_g to generate a set of goals. Thus, an agent responds to changes in its beliefs, resulting from perception, by generating goals according to its motivations and beliefs. These goals are then evaluated according to their motivational value (i.e. the amount by which their achievement would reduce the motivational intensity, as determined by f_m), and those that are considered sufficiently important are adopted as intentions by selecting an appropriate plan, and committing to its execution. Finally, an agent selects a particular intention to pursue and acts toward its achievement, again using motivational value as the guiding measure. The resulting architecture is illustrated in Figure 2, in which solid arrows represent the flow of information, and dotted arrows the control structure. The corresponding reasoning cycle is as follows.

1. Perceive the environment and update beliefs accordingly.
2. For each motivation apply f_i to update its intensity.
3. For each motivation apply f_g to generate a set of new goals.
4. Select an appropriate plan for the most motivated[2] of these newly generated goals, and adopt it as an intention.
5. Select the most motivationally valuable intention and act towards it by perform the next step in the plan
6. If the intention is completed then, for each motivation, apply the mitigation function f_m to reduce the intensity according to the motivational importance of achieving the goal.
7. Finally, return to the beginning of the cycle.

2.2 Trust

Cooperation involves a degree of risk arising from the uncertainties of interacting with autonomous self-interested agents. The notion of *trust* is recognised by several researchers as a means of assessing the perceived risk in interactions [7, 16]; trust represents an agent's estimate of how likely another agent is to fulfill its cooperative commitments. The risk of whether to cooperate, and with whom, may be determined by, among other things, the degree of trust. As agents interact they can infer trust values based on their experience and, over time, improve their models of trustworthiness. We base our model of trust upon Marsh's formalism [16] and the work of Gambetta [9], and define the trust in an agent α, to be a value from the interval between 0 and 1: $T_\alpha \in [0, 1]$. The numbers merely represent comparative values, and are not meaningful in themselves. Values approaching 0 represent complete distrust, and those approaching 1 represent complete, blind trust.

In our approach, trust values are associated with a measure of *confidence*, and as an agent gains experience this confidence increases. Trust values are initially infered according to an agent's *disposition*: optimistic agents infer high values, while pessimists infer low values. This disposition also determines how trust is updated after interactions [17]. After a successful interaction, optimists increase their trust more than pessimists, and conversely, after an unsuccessful interaction pessimists decrease their trust more than optimists. The magnitude of change in trust is a function of several factors depending on the agent concerned, including the current trust and the extent of the agent's optimistic or pessimistic disposition. The use of initial values combined with updates according to disposition means that each agent has an estimate of the trustworthiness of other agents, even if these values are simply the default.

We adopt Marsh's approach of representing both *general* and *situational* trust for a given agent [16]. General trust gives an overall view, based on all previous interactions. Situational trust is finer grained, and based on previous interactions in similar situations. In our model, *similar* situations are defined in terms of similar motivational value rather than specific capabilities. Although the latter is more powerful, it requires knowledge of why a particular cooperative interaction failed. For an agent to maintain models of the trustworthiness of others at a capability level it is necessary to know which capability caused cooperation to fail and why.

Existing models of teamwork and coalition formation do not generally consider trust except for a small number that consider trust for individual tasks, or for very specific constrained situations, such as electronic marketplaces [3]. In our view consideration of the trust in others in forming a clan is beneficial and, moreover, trust is a useful notion for binding a group together and providing additional justification for an agent deciding to perform an action that is not of direct immediate benefit on behalf of another.

We have described elsewhere [10, 11] a mechanism for agents to obtain assistance from other autonomous agents, through consideration of trust. Assistance is only obtained where it is motivationally valuable to each agent involved in the cooperative interaction. However, the approach we have developed is task-based, and suffers from the same limitations of other approaches described above. In particular, agents miss opportunities for cooperation that would be motivationally beneficial to *each individual agent*. This is the key point. We are not concerned with imposing some external utility function on the system as a whole, since that would detract from the benefits of autonomy through motivation [7, 8]. Instead, we wish to achieve cooperation with respect to the motivations of individuals.

2.3 Motivated Cooperation

Earlier work on motivation has been rather short-term in its view of cooperation [10, 15]. An agent acts, and cooperates if appropriate, according to the motivation that is currently of the highest importance. Missed opportunities for cooperation can arise when agents' motivations are out of step in time, leading to reduced benefit in the long-term. (This is analogous to goals being out of step in task-based teamwork approaches.) Cooperation arises when an agent's goal cannot be achieved alone (or is better achieved through cooperation). In such situations an agent uses its knowledge of others to determine who to ask for assistance. On receiving a request for assistance, agents inspect their own motivations and commitments (or *intentions*) to decide whether or not to agree, and send appropriate responses to the requesting agent; motivations determine whether agents *want* to cooperate, and intentions determine whether they *can* cooperate. An agent will agree to cooperate if there is no conflict of intentions[3] and the goal for which cooperation is requested is of motivational value.

There are two key problems with this approach. Firstly, the importance of a motivation fluctuates, and cooperation requests that would be valuable in the long-term, may be rejected. Secondly, the short-term view of cooperation leads to missed opportunities for cooperation as described above, since agents' goals may be out of step. In dynamic environments, fluctuations in the importance of motivations can lead to failures to establish cooperation that would actually benefit the individuals. We are concerned with maximising motivational value for individuals, with the aim of generating a positive consequence for the system as a whole, since typically (as system builders) we are concerned with the success of collections of agents rather than an individual. However, we are also concerned with maintaining the robustness and flexibility benefits that motivations afford.

3. COOPERATION FRAMEWORK

Previous work has described a framework for cooperation founded upon the notions of trust and motivation [10]. Cooperation is more than simultaneous actions and individual intentions; agents need some form of *commitment* to the activity of cooperation itself [1, 14] along with an appropriate set of *conventions* [21] specifying when and how a commitment can be abandoned. Where a group forms appropriate commitments to cooperate and adopts suitable conventions we say that they have formed a *cooperative intention*.

[2]The most motivated goal is the one that has the most motivational value, as determined by the mitgation function f_m.

[3]More specifically, cooperation occurs only if any conflicts are resolved in favour of cooperation.

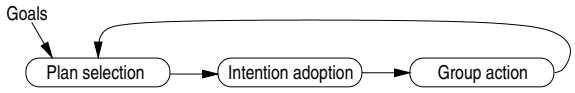

Figure 3: Stages of cooperation

There are several distinct tasks surrounding the formation and execution of a cooperative intention. If an agent is faced with a plan that requires cooperation, it must seek assistance and form an appropriate cooperative intention before the plan can be fully executed. The process of setting up a cooperative intention can be broken down into the stages of plan selection, intention adoption, and group action as illustrated in Figure 3. We give a brief overview of these stages below, but due to space constraints we do not discuss the details.

Motivations give rise to goals that must be adopted as intentions by selecting a plan and committing to its execution. The applicable plans for a particular goal may include plans that require cooperation. If an agent selects such a plan, which we call a *cooperative plan*, it is electing to cooperate for the achievement of the goal; the choice to cooperate is incorporated into plan selection. In order to select between cooperative plans, an agent must consider the nature of those it may cooperate with; it should consider the likelihood of finding agents to assist and the likelihood that they will execute the plan successfully, i.e. their trustworthiness. We have described in [11] a mechanism for assessing the contributions contained in a plan, in terms of the risk associated with the agents who are believed capable of executing them. This assessment is combined with more traditional standard planning heuristics (such as cost and plan length) to obtain a measure for selecting between plans that balances these, often contradictory, aims to minimise both cost and risk. Using this approach an agent's choice about whether to cooperate or not is embodied in its choice of plan.

After selecting a plan, an agent commits to its execution by forming an intention. If the plan does not require assistance from others it can simply be adopted, and action towards it can begin, otherwise the agent must solicit assistance towards its execution. In order to gain assistance, the initiating agent must determine who to request assistance from. This is achieved by iterating through the steps of the plan, annotating each contribution with the identifier of the agent that the initiator considers best able to perform it, based on knowledge of their capabilities, and their believed trustworthiness. This annotation process is described in detail in [12] The assistance of these agents can then be requested. On receiving a request for assistance, these agents inspect their own motivations and intentions to decide whether or not to agree, and send an appropriate response. If sufficient agents agree then a cooperative intention can be established among them. However, if insufficient agents agree then either the plan can be reannotated, or failure is conceded.

Once a group of agents has formed a cooperative intention it can be executed. On successful completion of the cooperative intention, the agents dissolve their commitment and cooperation is finished. Alternatively, if execution fails, the agent that first comes to believe this informs the others in accordance with the specified conventions, and again their commitments are dissolved. In both cases, agents update the information they store about others to aid future decisions about cooperation. In particular, the trust values infered for these agents are updated.

4. COALITION FORMATION

In this paper, we address the limitations of existing cooperation approaches identified in Section 1, in particular with respect to motivation and trust, by enabling agents to form medium-term clans. Short-term clans are not appropriate since motivations are too dynamic when considered over a short period of time, i.e. over a small number of tasks; in the short-term agents' motivations may be out of step leading to the problems described above in Section 2. Similarly, long-term clans are unsuitable since although a given agent has a fixed set of motivations, the general trend of which motivations are active may change. For example, if two agents collaborate to obtain and maintain information, then although their active motivations might change in the short-term, they are similar in the medium-term, while they may diverge long-term (once the information is finished with). In other words, in the short-term motivations fluctuate, in the long-term they slowly change, but in the medium-term they are relativity stable.

There are a number of key issues that must be addressed in order for agents to form clans for medium-term durations based around their motivations, as follows.

1. Why find others to cooperate with — what motivational benefit is there to forming a clan?
2. When should an agent initiate the process of forming a clan?
3. How can other suitable agents be identified, and how should a clan be formed?
4. Under what circumstances must an agent reconsider its membership of a clan, and how should an agent leave a clan?
5. When should an agent act in favour of the clan rather than for short-term motivational value? More specifically, since a self-interested agent acts solely according to its own individual motivations, what is the motivational justification for acting in favour of the clan?

In the remainder of this paper we address these questions, and provide a broad mechanism for clan formation based on the notions of trust and motivation. The limitations identified above can be seen as a set of reasons for forming a clan. In particular an agent may suffer as a result of: missed opportunities for cooperation, problems of scalability, lack of information in decision making, and lack of robustness of commitments in highly dynamic environments. These factors can be combined to determine whether to attempt to form a clan. Figure 4 gives a skeletal algorithm outlining the decision process. In the remainder of this section we describe the component steps of this algorithm.

4.1 Missed opportunities

Recall from Section 2 that the dynamic nature of motivations can lead to missed opportunities for cooperation. Therefore, an agent needs some means of assessing the extent to which it is missing such opportunities. Unfortunately, this cannot be quantified directly, even from a system-level perspective. At the system-level we can perform a limited analysis based on knowledge of agents' motivations and how they evolve. However, we cannot predict how the environment will change, nor consequently how important a motivation will be at some future time. We can undertake a retrospective analysis, but this is complex and of little benefit in providing a mechanism to guide behaviour at run-time.

In our model, an agent selects a plan for its goal, and attempts to annotate cooperative actions in that plan with the identifiers of agents to seek assistance from. If an agent is missing opportunities for cooperation, then its attempts to cooperate are likely to fail at annotation time. Therefore, each missed opportunity leads to a

```
function ATTEMPT-TO-FORM-CLAN returns boolean
  local: missed-opportunities ←false
    scalability ←false
    lack-of-information ←false
    high-failure-rate ←false
  if (plan-annotation-failure-rate > annotation-threshold)
    and (MOTIVATIONAL-VALUE( FILTER(previous-rejected-requests)) > rejection-threshold) then missed-opportunities ←true
  if PROPORTION-COOPERATIVE(recently-applicable-plans) > scalability-threshold then scalability ←true;
  if (AVERAGE-TRUST(agent-models) < trusted-threshold) or (AVERAGE-CONFIDENCE(agent-models) < confidence-threshold)
    and (exists agents such-that (AVERAGE-TRUSTWORTHINESS(agents) > trusted-threshold)
    and (AVERAGE-CONFIDENCE(agents) > confidence-threshold))
    then lack-of-information ←true
  if (failure-rate > failure-threshold) then high-failure-rate ←true
  if (missed-opportunities = true) or (scalability = true) or (lack-of-information = true) or (high-failure-rate = true)
    then return true else return false
end
```

Figure 4: Assessing when to form a clan.

failure in plan annotation[4], and consequently many missed opportunities for cooperation result in a high failure rate at annotation time. Unfortunately, annotation may fail for other reasons and a high annotation failure rate may not imply a high number of missed opportunities. A partial solution is for an agent to maintain a record of the requests for assistance that it has itself received and declined due to lack of motivational value. Where a high annotation failure rate is a result of missed opportunities then previous requests for assistance will contain the requests corresponding to those opportunities. If an agent experiences high plan annotation failure rates it can inspect these previous requests. If these requests contain several that are currently motivationally valuable, and are similar in nature to the plan being annotated, this indicates that cooperation opportunities are being missed. We can construct a heuristic based on this information for an agent to use in deciding whether to attempt to form a clan. This heuristic tests whether the annotation failure rate exceeds the $annotation - threshold$, and whether the current motivational value of the (similar) previously rejected requests exceeds the $rejection - threshold$. If both tests are positive, then we take the agent to be at risk of missed opportunities.

4.2 Scalability

Existing approaches to cooperation can have problems scaling to large numbers of agents. In our model, when annotating a plan an agent must consider each agent of whom it has a model to determine whether they are trusted and have suitable capabilities. Moreover, once a set of agents has been identified, communicating with them and processing their responses is costly. Clearly, as the number of agents increases, the search space and communication cost also increases. Our proposed notion of clans can be seen as taking the essence of Brooks and Durfee's congregations solution to this problem [4, 5], and applying it in the context of a trust and motivation based cooperation.

At the simplest level, the number of agents modelled gives an indication of the scalability problem, since an agent must search through all its models. However, the scalability problem also depends on the frequency of searching — if cooperation is rare, then the impact is much less than if each plan requires cooperation. The percentage of an agent's cooperative plans influences the frequency with which it cooperates. However, an agent does not necessarily utilise all of its plans, so we can filter out those that are less likely to be significant. In particular, we can measure the percentage of plans that are cooperative in the last n reasoning cycles, where n is a direct function of the agent's memory length, i.e. we consider all *applicable plans* in last n iterations. An agent can inspect the set of recently applicable plans and, if the proportion of these that are cooperative exceeds the $scalability - threshold$, then it should attempt to form a clan in order to address the scalability problem.

[4]Actually, this is a simplification since the agent may cooperate with a less trusted agent. However, as estimates of trustworthiness become more accurate over time, the likelihood of cooperating with less trusted agents reduces.

4.3 Lack of information

There are two areas where lack of information affects cooperation. Firstly, an agent may have little information about others' trustworthiness (having had few previous interactions with them). Secondly, an agent may not have sufficient information about others' capabilities to make decisions about cooperation. On joining a multi-agent system, both of these represent problems for an agent, since it will not have a history of interactions to provide information. However, although membership of a clan may be beneficial at this point, it is not possible to join a clan since there is no information with which to evaluate the benefits and risk of so doing. Over time, agents accumulate this information, which they can use in reasoning about cooperation in general, but also in reasoning about clan formation.

Recall that in modelling the trustworthiness of others an agent maintains a measure of confidence in its trust assessments. Trust models based on limited experience are given low confidence in comparison with models based on extensive experience. In considering clan formation an agent must consider both the confidence placed in its models, and their extent. There is a lower bound below which clan formation is not practical due to lack of information (or confidence) about its potential members. In particular, it is only sensible to join (or form) a clan with agents who are trusted to a reasonable degree of confidence. Therefore, an agent should inspect its models of others; if there are many untrusted agents or agents whose models have low confidence, it can attempt to form a clan (provided there is a subset of confidently trusted agents with whom to form a clan).

4.4 High execution failure rate

Clan membership may help reduce a high failure rate of cooperative intentions at execution time. Although cooperative intention requires a commitment to all agents involved, in our model the duration of this commitment is determined solely by an agent's motivations. Highly dynamic environments give rise to fluctuations in motivation intensity, and can lead to excessive failures. Membership of a clan provides an additional degree of commitment and, importantly, a mechanism for an agent to obtain motivational value through acting in what may appear to be a semi-benevolent man-

ner. In the following section, we describe the motivational aspects of clan membership.

5. CLAN FORMATION

In common with other aspects of cooperation, clan formation is guided by trust and motivation. For an initiating agent, the influence of motivation is determined in relation to the motivational value that may be gained by clan membership (in terms of a higher success rate and quality of future interactions). This is indirectly accounted for in the decision to form a clan in the first instance. The influence of trust, however, is more direct. At a fundamental level, trust determines whether it is practical to form a clan. As discussed above, if an agent has a low trust in others or has low confidence in its models of those it considers trustworthy, it should not form a clan. However, if it does have adequate trust in others, it can attempt to form a clan with the most trusted agents possible. In our current model, for reasons of simplicity and computational cost, an agent simply attempts to form a clan with the set of most trusted agents. Their capabilities are not considered, since the agent cannot predict which plans will require cooperation in the future.

After determining the most trusted agents, they must be sent a request to form a clan. In the ideal case, no further information would be required, since the agents might also consider it beneficial to join a clan based on their own assessments of missed opportunities, trust, and previous failures. However, due to differences in individuals' experiences this is generally not the case. In particular, although an agent may benefit from clan membership, its own assessment of whether it should attempt to form a plan (using the algorithm in 4) may not indicate this. An agent must, therefore, give some incentive for joining the clan. Since we do not assume that agents have negotiation or persuasion capabilities abilities, we take a simple mechanistic approach. Specifically, the request to join must include the set of most frequently generated goals from the most active motivations as a means for others to assess the usefulness of clan membership. The disadvantage to this approach lies in revealing what is essentially private information. However, since the agent should (hopefully) gain motivational benefit from forming the clan, we argue that this is justified. The number of agents requested is domain dependent, and must be determined empirically. Figure 5 outlines this process in the function INITIATE-CLAN-FORMATION.

On receiving a request to join a clan an agent must consider the motivational value of so doing. Firstly, the criteria described above are considered, namely, the perceived extent of missed opportunities and lack of information etc. This gives an overall indication of how beneficial the agent would find clan membership *in general*. Secondly, the goals contained in the request are used to estimate how useful it would be to join the clan *in particular*. The motivational value of each goal is considered in a situation independent manner. If both the general and specific exceed a threshold then the agent agrees to form a clan. If sufficient agents respond positively then, on receiving the responses, the initiator sends acknowledgements and a clan is formed (with those who acceded). Alternatively, if insufficient agents accede then those agents that did accede are informed and clan formation abandoned. Figure 5 outlines this process in the function PROCESS-FORMATION-REQUEST.

```
function INITIATE-CLAN-FORMATION
  target-agents ←SELECT-MOST-TRUSTED(agent-models,
      confidence-threshold)
  goals-to-communicate ←EXTRACT-GOALS(active-motivations)
  for agent in target-agents do
    REQUEST-FORM-CLAN(goals-to-communicate)
  end
end

function PROCESS-FORMATION-REQUEST returns response
  input: request-goals
  local: motivational-value ←0
  if ATTEMPT-TO-FORM-CLAN = true then return accept
  for goal in request-goals do
    motivational-value ←motivational-value
        + MOTIVATIONAL-VALUE(goal)
  end
  if motivational-value > threshold return accept
    else return decline
end
```

Figure 5: Clan formation.

5.1 Reasoning in a Clan

There are two aspects to the influence of clan membership on behaviour: sharing of information, and increased likelihood of cooperating and fulfilling cooperative intentions. The first of these is relatively simple, since if a clan member requires assistance but does not know of any trusted agents having the required capabilities, then it can request information from other clan members. If another member is aware of such an agent then it informs the requester of the identity of the capable agent. It does not communicate trust information, since although in our model of trust each agent represents trustworthiness on a scale of 0 to 1, the values held by different agents are *not comparable*. Highly trustworthy may equate to different numerical values for different agents. Unfortunately, there is little by way of practical solution to this. Marsh suggests that a simple approach would be to use a stratification of trust for inter-agent communication, dividing the numerical range into equal subranges [16], but goes on to describe a number of associated problems with this approach. Instead, we take the approach that if agents have joined a clan together there is a degree of commonality, and this is sufficient. Investigating mechanisms for agents to share trust information and reason about reputation is an area of ongoing work. Trust values are internal to an agent and depend on its disposition and experience; they are not directly comparable across agents. Therefore, if A trusts B and B trusts C it is not necessarily true to say that A trusts C.

The second aspect through which clan membership influences behaviour is in terms of motivational value. Clan members are more likely to cooperate and to fulfil their commitments. Both of these aspects result from the motivational value ascribed to the cooperative interaction. In order to ascribe motivational value to clan membership, and to ensure that agents remain self-interested, we introduce an additional *kinship motivation* to all agents. This motivation is mitigated by offering assistance to other clan members. In practise it functions like any other motivation — its influence is taken into account when deciding whether to cooperate, and in determining when to rescind commitments. Kinship intensity is determined by such factors as the proportion of goals that require cooperation, and the extent and quality of the information an agent has about others, namely the very criteria that led to clan membership in the first place.

At a philosophical level, introducing a kinship motivation can perhaps be seen to undermine the fundamentally self-interested nature of agents. Recall, however, that agents choose to join a clan for specific reasons that are undeniably self-interested. Furthermore, the kinship motivation is just one of a set of motivations, and does not override the others; if it did then the agent would certainly cease to be self-interested. Given sufficient information and reasoning resources the kinship motivation could be avoided, since an agent would be able to reason explicitly about the benefit

it may in the future receive from agreeing to cooperate, or sharing information. However, in practise such information and reasoning resources are unrealistic, and we take this simple approach of introducing an additional motivation.

Finally, if an agent joined a clan to address scalability problems, i.e. to reduce the search cost of finding cooperative partners, then it can simply search through the members of the clan. This is a special case arising from the reason for forming a clan. Due to space constraints we do not give details here but, in broad terms an agent goes through the standard process of attempting to form a cooperative intention but restricted to agent models corresponding to clan members. If this fails, then the standard cooperative intention formation procedure is undertaken.

5.2 Maintenance and Dissolution

In a dynamic environment the importance of an agent's motivations change, and addressing this provides one of the prime reasons for forming a clan. However, this addresses the short-term fluctuations in motivations, rather than a long-term change. Specifically, while motivations may change significantly in the short-term, the general set of active motivations is likely to be relatively static. In the long-term however, this set of active motivations is also likely to change. As an analogy, consider the role of a personal assistant agent whose hour-to-hour motivations are likely to fluctuate, but whose day to day motivations are likely to be broadly similar. Over time, however, user interests and priorities are likely to change and alter the set of day-to-day motivations. As an agent's motivations undergo long-term change it is likely that the benefit gained from membership of a specific clan will decrease (unless the other clan members have undergone similar changes). The consequence is that eventually one or more members of the clan no longer receive sufficient benefit to justify continued membership. Although above we have described the benefits of being a member of a clan, this has a cost, since a result of the kinship motivation may be that the agent acts to assist another clan member, rather than as it would otherwise. Provided the clan is operating effectively there will be sufficient reciprocal action for each member to receive net benefit overall. However, if the set of active motivations changes then it may no longer receive benefit from the clan and, given that agents are self-interested, should withdraw its membership by notifying the other members. In addition to the situation where an agent is no longer receiving benefit from clan membership, it should also withdraw its membership if it comes to distrust the other members.

6. RESULTS

In order to explore the effectiveness of our model we have developed a testbed simulation in which to explore the behaviour and effectiveness of various agent configurations. Specifically, we investigated the effect of incorporating our model of clans. For experimentation purposes we populated the testbed with agents whose capabilities, goals, plans, motivations, and reliability (in terms of the conditions in which their execution of actions will fail) are randomly selected. The actions contained in the plans are also randomly generated, as are their durations.

A number of simplifications were made in constructing the testbed. The most significant is that we assume a closed system, i.e. agents cannot join or leave during a simulation, so we do not investigate the effect of clans on scalability issues. However, given the parallel between clans and Brooks and Durfee's work on congregations, we hope to achieve similar benefits.

The testbed allows the monitoring of various measures of individual and system performance. Due to space constraints, we concentrate specifically on two: the total number of successful interactions, and the total number of interactions failing at execution time. These measures give an indication of system performance as a whole. However, our results show that an individual's performance loosely mirrors system performance. In particular, there is no subset of agents that benefits at the cost of others. This does occur if the number of clans is tightly restricted, but in general the majority of agents join a clan at one time or another and such perturbations are smoothed over time.

Figure 6 shows results averaged over several simulation runs, where the x axis represents time and the y axis the number of successful interactions. The left graph shows the number of failed interactions and the right graph the number of successful interactions. The graphs give the results of using clans with low, medium, and high kinship motivations, along with the control case of not using clans. As might be expected, clans are successful in reducing the number of failed interactions, and the extent of success is proportional to the importance of the *kinship* motivation. It can be seen that with low importance placed on kinship there is little effect, but larger values offer more significant gain.

The impact of clans on the number of successful interactions is more complex. It can be seen that there are fewer successful interactions using clans with a low kinship importance, than without using clans. High kinship importance, however, yields a significant increase in successful interactions. Other results, with different numbers of agents, thresholds, and simulation duration suggest that this results from the increased computational overhead placed on agents by the mechanisms required for clans. This overhead results in fewer actions performed due to time spent reasoning. When kinship is given a higher importance, a correspondingly higher number of requests are accepted and fewer commitments are broken, increasing the number of successful interactions. There is a transition point at which the increased number of successful interactions resulting from using clans is equal to the reduced number of actions resulting from the extra processing cost. This can be observed in the left graph, since although a reduced failure rate is observed for low kinship importance, this is actually due to there simply being fewer interactions. However, a higher kinship results in a more significant reduction, indicated by the lowest line on the left graph.

These results demonstrate that the model offers certain benefits to agents, but at a significant cost. Future work is needed in further exploring the implications of our approach to clans, to determine the bottlenecks, and to revise and optimise the model accordingly.

7. CONCLUSIONS

In this paper we have addressed some of the limitations of existing teamwork and coalition formation approaches, and of our previous trust and motivation based framework of cooperation. In particular, we have described how agents can utilise clans in addressing the problems of missed opportunities for cooperation, scalability, a lack of information, and high failure rates at execution time. We have outlined how an agent can form and maintain a clan, and demonstrated that it offers certain benefits to cooperation.

There are many areas for future work, the primary one being to explore how an agent should manage its clan membership when it is a member of more than one clan whose general objectives may differ. The model also needs to be the subject of more extensive experimentation in order to refine the heuristics and determine appropriate values for the thresholds used in decision making. We claim that the kinship motivation is a useful mechanism for an agent to balance the benefits of clan membership against the potential costs. In this sense the motivation can be thought of, in part, as a kind of utility function (although motivations are much more than simple utilities). Future work will investigate the effectiveness of the

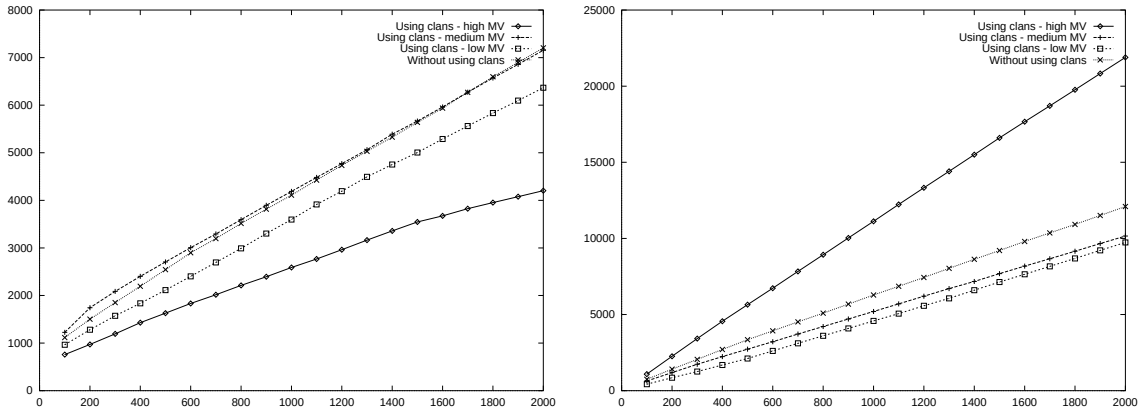

Figure 6: Failed cooperative interactions and successful interactions.

kinship motivation using a coarser external utility measure, to determine how well agents are balancing the cost of clan membership against the benefits. Our current model allows individual clan members to leave if they are no longer receiving benefit from clan membership. However, agents are currently unable to join existing clans, and this is an area of ongoing work. Finally, we intend to extend the simulation testbed to allow agents to join or leave at runtime, allowing us to explore scalability issues.

8. REFERENCES

[1] M. E. Bratman. Shared cooperative activity. *Philosophical Review*, 101(2):327–341, Apr. 1992.

[2] M. E. Bratman, D. Israel, and M. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[3] S. Breban and J. Vassileva. Long-term coalitions for the electronic marketplace. In B. Spencer, editor, *Proceedings of the E-Commerce Applications Workshop, Canadian AI Conference*, 2001.

[4] C. Brooks and E. Durfee. Congregating and market formation. In *Proceedings of the First International Joint Conference on Autonomous Agents in Multi-Agent Systems*, pages 96–103, Bologna, Italy, 2002. ACM Press.

[5] C. Brooks, E. Durfee, and A. Armstrong. An introduction to congregating in multiagent systems. In E. Durfee, editor, *Proceedings of the Fourth International Conference on Multi-Agent Systems (ICMAS-2000)*, pages 79–86, 2000.

[6] C. Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In M. J. Wooldridge and N. R. Jennings, editors, *Intelligent Agents: Proceedings of the First International Workshop on Agent Theories, Architectures and Languages (ATAL-94)*, pages 56–70. Springer-Verlag, 1995.

[7] C. Castelfranchi and R. Falcone. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS-98)*, pages 72–79, Paris, France, 1998.

[8] M. d'Inverno and M. Luck. *Understanding Agent Systems*. Springer-Verlag, 2001.

[9] D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Basil Blackwell, 1988.

[10] N. Griffiths. *Motivated Cooperation in Autonomous Agents*. PhD thesis, University of Warwick, 2000.

[11] N. Griffiths and M. Luck. Cooperative plan selection through trust. In F. J. Garijo and M. Boman, editors, *Multi-Agent System Engineering: Proceedings of the Ninth European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99)*. Springer-Verlag, 1999.

[12] N. Griffiths, M. Luck, and M. d'Inverno. Annotating cooperative plans with trusted agents. In R. Falcone and L. Korba, editors, *Proceedings of the Fifth International Workshop on Deception, Fraud and Trust in Agent Societies*, 2002.

[13] M. Klusch and O. Shehory. Coalition formation among rational information agents. In W. Van de Velde and J. W. Perram, editors, *Agents Breaking Away: Proceedings of the Seventh European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW-96)*, pages 204–217, 1996.

[14] H. J. Levesque, P. R. Cohen, and J. H. T. Nunes. On acting together. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, pages 94–99, Boston, MA, 1990.

[15] M. Luck and M. d'Inverno. A formal framework for agency and autonomy. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 254–260. AAAI Press/The MIT Press, 1995.

[16] S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, 1994.

[17] S. Marsh. Optimism and pessimism in trust. In *Proceedings of the Ibero-American Conference on Artificial Intelligence (IBERAMIA '94)*, 1994.

[18] T. J. Norman. *Motivation-based direction of planning attention in agents with goal autonomy*. PhD thesis, University of London, 1996.

[19] O. Shehory and S. Kraus. Task allocation via coalition formation among autonomous agents. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 655–661, Montréal, Québec, Canada, 1995.

[20] M. Tambe. Towards flexible teamwork. *Journal of Artificial Intelligence Research*, 7:83–124, 1997.

[21] M. Wooldridge and N. R. Jennings. Formalizing the cooperative problem solving process. In *Proceedings of the Thirteenth International Workshop on Distributed Artificial Intelligence (IWDAI-94)*, pages 403–417, Lake Quinhalt, WA, 1994.