Coalition Detection and Identification (Extended Abstract)

Reid Kerr
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
rckerr@cs.uwaterloo.ca

ABSTRACT

This document outlines doctoral research being undertaken by Reid Kerr, under the supervision of Robin Cohen. In this research, we are investigating the problem of detecting and identifying groups of agents that are collaborating, within a larger population. In the scenarios with which we are primarily concerned, agents may cooperate to further their own interests, despite the fact that this may be unwelcome or forbidden. Colluding agents, then, are unlikely to advertise their membership in a team, and may actively seek to conceal their cooperation. We address the problem of identifying such coalitions and their activities.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Systems—*Multiagent Systems*

General Terms

Measurement, Security

Keywords

Coalitions, collusion, multiagent systems, marketplaces

1. MOTIVATION

Trust and reputation researchers are concerned with multiagent systems where an agent's success depends to a large degree on the reliability or trustworthiness of the agents with whom it chooses to interact. A prominent example of such a scenario (and the one with which we concerned ourselves) is that of large electronic marketplaces. In such marketplaces, agents buy and sell with one another; success depends to a large degree on trading with reliable agents. Trust and reputation systems aim to aid agents by helping them to find trustworthy partners and/or avoid untrustworthy ones.

A multitude of trust and reputation systems have been proposed; typically, these systems provide some degree of protection against agents that are untrustworthy, and act individually. Most researchers, however, readily acknowledge that their systems provide little protection against coalitions of agents. In fact, trust and reputation researchers have made little progress on the problem of collusion (i.e., coordinated actions by a coalition to undermine a system).

Cite as: Coalition Detection and Identification (Extended Abstract), Reid Kerr, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

2. COALITIONS AND TEAMWORK

The importance and persistence of this problem inspired us to investigate the issue of collusion, and of teamwork in general, for my doctoral research. While there has been substantial work on the problems of coalition formation and stability, recognition of coalition behaviour has received less attention. Some researchers have investigated the problem of multiagent plan recognition, but it has been noted (e.g., [4]) that there has been a limited quantity of such work.

Existing multiagent plan recognition work, while important, focuses on different scenarios than those with which we are concerned. For example, it may be assumed that the membership of the team (or teams) is known, and we seek only to identify the plans used by the team (e.g., [1]), or how the team has organized itself internally to execute a plan. Moreover, existing work (following related work on single agent plan recognition) often presumes the existence of a model of the team or known library from which plans are drawn. (e.g., [4, 5]). By matching the observed behaviours of the agent(s) to plans in the library, the likelihood that a specific plan is in use can be evaluated, or the best match can be determined. Beyond simply identifying plans, this approach can yield other information: for example, it may aid in identifying the members of subteams that have split to perform subtasks, or detecting when subteams have merged.

2.1 Motivating Scenarios and Issues

Our work focuses on different scenarios than the multiagent plan recognition work with which we are aware. For example, in the marketplace scenario noted above, there may be a very large number of participating agents. It is extremely unlikely that all of the agents belong to one coalition, or are partitioned into two competing teams. Instead, there may be zero, one, or many coalitions at work; each may represent a small or large subset of the total set of agents. Coalitions may or may not be disjoint. In such a scenario, simply identifying whether a coalition exists may be important—it may signal a need to take corrective action, to investigate further, to halt marketplace activity, etc. Identification of the members of each coalition is also likely to be an important outcome; to what extent false positives or false negatives are important is likely to depend on the intended application. In contrast to the focus of existing work, it may or may not be important to identify the intended action of a coalition, or their plan for achieving it.

Many important scenarios can be identified which share characteristics with that described above. Examples include: detecting teamwork in online games, combating cheating in gambling activities such as poker; identifying criminal, insurgent or terrorist activity within a larger population, and detecting insider trading.

A key issue in such situations is that, unlike scenarios studied in many works, we cannot presume possession of a known plan library. We may have knowledge of some of the strategies coalitions may employ, but we cannot assume that such a list is comprehensive. This issue was highlighted in our investigation of security in trust and reputation systems [2]. In this area, we developed an important catalog of known vulnerabilities/attacks. During our research, we uncovered new attacks, highlighting the difficulty in developing a comprehensive library, and similarly, the danger in depending on the fact that a library is complete. Indeed, passing familiarity with news reports on scams and fraud, on computer crime and security breaches, etc., reveals an astonishing ability for adversaries to find new strategies to prey on the unwary, because the strategy space is so large that it makes potential attacks difficult to foresee. Under such circumstances, the assumption of a known library seems to seriously limit the value of a work.

3. APPROACH

Based on the above, our goal is to use techniques that require no plan library. To accomplish this, we must use observable actions, yet without matching them against known patterns—we must rely on more fundamental properties.

Typically, a self-interested agent will be part of a coalition because it sees a benefit in doing so. (An agent might participate because, for example, it has been coerced to do so. It can still be argued, however, that there is benefit in such a case: the agent avoids the damage it might incur if it did not cooperate.) We presume that each agent is individually rational, and will only be part of a coalition if it sees a net benefit from doing so.

Benefits to an agent may or may not be observable; for example, an agent may receive payments from another agent via a 'back channel'. Note, however, that the agent making such payments is also rational—it must be accruing benefits at least as great as the payments it is making. Note, too, that the agent receiving the payment is likely to be benefiting others in the coalition—if it simply received payments, and didn't make any contributions, then it would not be desired as a teammate by others in the group. Thus, we would expect net positive (observable) benefits from the mutual actions of coalition members.

Similarly, we would not expect coalition members to do substantial harm to one another. Certainly, an agent might harm a teammate accidentally, or to avoid the appearance that they are in the same coalition. Note, however, that the former case should not occur frequently, and in the later case, the net positive benefits should outweigh the harm.

At present, we are investigating the use of these properties to identify groups with shared interests. One can envision a number of ways in which one might attempt to distinguish coalition membership based on these properties. Using mutual net benefit or mutual harm as the basis of similarity measures, we are exploring the use of clustering to identify these agents. There is likely to be much noise in a sample (for example, as agents do things that benefit those that are not part of its coalition, such as making honest sales to outsiders). It is anticipated, however, that the relatively stronger benefit/weaker harm to coalition members will provide detectable signal. The use of graphical models is another avenue of investigation.

These principles can be applied directly to domains such as marketplaces (where, e.g., paying a user equates to benefitting them, giving a negative review harms them, etc.) and games like poker (where, e.g., betting against another player harms him). It is also possible that the same basic principle can be applied in less structured domains. In a battlefield scenario, shooting someone is an obvious case of harm. But correlated movement, for example, can also provide signals: moving in parallel with another agent may indicate supporting behaviour, while converging movement, attempts to conceal oneself, etc. may indicate antagonism.

Detecting coalitions in this way may provide means to characterize the activity, as well. For example, when a candidate set of agents has been identified, the related actions underlying the identification might be analyzed using a technique to extract plans (e.g., [1]).

Our goals require us to detect coalitions when the activity does not match known plans. They do not preclude using a library-based approach for those plans which *are* known. We intend to explore hybrid approaches, were a plan-based approach is used to identify well-known or previously detected behaviours, while a 'planless' approach is used to detect teamwork where the plan-based approach cannot.

It is also possible that our work might be useful as a preprocessing step or input into a plan-based approach, identifying teams to reduce the search space for library searches.

4. EVALUATION

As noted, we have a particular interest in the marketplace scenario, and it provides a well-structured environment in which to evaluate our techniques. To that end, we have developed an experimental testbed modelling a marketplace environment [3]. This testbed was specifically formulated to allow investigation of collusion by agents. The marketplace can be populated with agents exhibiting a wide variety of behaviours, both honest and dishonest. Within this population, sets of colluding agents (using a variety of techniques) can be added; the accuracy of the identification techniques can be evaluated under a wide range of conditions.

5. REFERENCES

[1] G. Kaminka, M. Fidanboylu, A. Chang, and M. Veloso. Learning the sequential coordinated behavior of teams from observations. *Lecture notes in computer science*, pages 111–125, 2003.

[2] R. Kerr and R. Cohen. Smart cheaters do prosper: Defeating trust and reputation systems. In *Proceedings of AAMAS'09*, Budapest, Hungary, 2009.

[3] R. Kerr and R. Cohen. TREET: The Trust and Reputation Experimentation and Evaluation Testbed. *Electronic Commerce Research*, To appear.

[4] G. R. Sukthankar. *Activity recognition for agent teams.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2007. Adviser-Katia Sycara.

[5] M. Tambe. Tracking dynamic team activity. In *Proceedings of the National Conference on Artificial Intelligence*, pages 80–87, 1996.