Detecting and Identifying Coalitions

(Extended Abstract)

Reid Kerr
University of Waterloo
Waterloo, Ontario, Canada
rckerr@cs.uwaterloo.ca

Robin Cohen
University of Waterloo
Waterloo, Ontario, Canada
rcohen@cs.uwaterloo.ca

ABSTRACT

In multiagent scenarios, subsets of a population (*coalitions*) may attempt to cooperate, for mutual benefit. We present a technique for detecting the presence of coalitions (malicious or otherwise) and identifying their members, and demonstrate its effectiveness.

Categories and Subject Descriptors

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

General Terms

Experimentation, Security

Keywords

Coalitions, Collusion, Trust and Reputation, Multiagent Systems

1. INTRODUCTION/RELATED WORK

In multiagent systems, groups of agents (*coalitions*) may seek to coordinate their activities in some way, to further their goals; where such activity is unwelcome, it may be called *collusion*. Coalitions represent a persistent and pervasive problem for many multiagent systems. Despite this, there has been little progress towards a solution. Here, we present a technique for detecting coalitions in an environment, and for identifying coalition members. Detection might, e.g., allow remediation, or might serve as a deterrent.

Because our approach is based on the concept of benefit rather than on domain-specific features, and because it requires no knowledge of the plans in use, we believe it to be applicable to a wide variety of domains: e.g., cheating in games, 'shilling' or 'astroturfing', or insurgent activity. Here, we apply our technique to trust and reputation systems for marketplaces, where two forms of collusion are well-known problems: *ballot-stuffing* (false positive reviews, to inflate teammates' reputations), and *bad-mouthing* (false negative reviews, to damage competitors' reputations). Both attacks seek to improve team members' chances of being selected by other agents. We demonstrate strong detection performance, with excellent resistance to false positives. As such, this work represents an important step towards addressing the challenges posed by coalitions.

Key characteristics of the scenarios of interest should be noted. First, we have no knowledge of communication or sharing of resources by coalition members outside the system. Second, and importantly, we assume no knowledge of the plans in use.

While several areas of research share some relation to our problem, the work in each targets fundamentally different scenarios than our own. Coalition formation and stability (e.g., [4]) assumes, for example, that the capabilities of agents, and payouts, are known. Similarly, work in multiagent plan/behavior recognition (e.g., [5]) assumes known plan libraries. Community finding, in social networks (e.g., [3]), typically uses metrics (e.g., connectivity, frequency of interaction) that are of limited value for our problem.

2. METHOD

Because we have no access to a plan library, our method must rely on fundamental properties of the observable actions themselves. In particular, self-interested agents belong to coalitions because they expect to improve their benefit (or reduce harm done to them). We might expect that coalition members are more likely to help one another than to help outsiders, and/or more likely to harm outsiders than to harm one another. The important insight is that because coalition members favor the same set of agents (each other), there is likely similarity in terms of the agents they benefit, and harm.

Our technique is a two step process. First, we identify 'candidate' sets of agents; second, we characterize each candidate group as either a coalition, or not.

We define the *benefit space* as a high-dimensional space reflecting the degree of benefit (and harm) rendered to each agent in the system. This is a key insight—the benefit space formulation allows possible coalitions to be detected using existing tools such as clustering. Specifically, given N total entities in the system, the benefit space $\mathcal{B}$ is a space $\mathbb{R}^N$, where the value in each dimension β_i represents an amount of net benefit (i.e., total benefit minus total harm) to entity i. Each entity maps to a point in the benefit space, reflecting the amount of (observable) net benefit it has rendered to each entity in the system. Because members of a coalition are likely to be similar in terms of the sets of agents that they favor, we would expect them to be close in this benefit space. Using Euclidian distance as our dissimilarity measure, we have used k-means clustering to partition the population P into a set of clusters $\{C_1, C_2, ..., C_n\}$, each of which is a *candidate coalition*.

Similarity does not *necessarily* imply that a set of agents is a coalition; for example, agents may simply have similar preferences, so they select the same sellers. Thus, we must characterize each candidate cluster to determine if it is, in fact, a coalition. We might expect a true coalition T to be more 'self-serving' (i.e., benefiting each other more than outsiders) than a 'non-coalition' group G. In this case, we would expect the benefit flowing *from* members of T *to* members of T to be greater than the benefit flowing from members of G to members of G. (Similarly, we might expect a coalition

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

to damage outsiders more than a 'normal' group would. The discussion of this is omitted, for brevity.) Consider any given set of agents S, where $m = |S|$. There are $m(m-1)$ (directed) relationships between agents in S. The average benefit (per relationship) flowing *from* agents in S, *to* agents in S, then, is:

$$\bar{\beta}_S = \frac{\sum_{i \in S} \sum_{j \in S, j \neq i} \beta_j(i)}{m(m-1)} \tag{1}$$

Using Formula 1, we can find $\bar{\beta}_C$, the average benefit within C. To know whether the computed value is abnormally high, we need a benchmark to which to compare it. For this, we take random samples of m agents (drawn from the entire population P, including agents in C). For each sample G, we compute $\bar{\beta}_G$, using Formula 1. Doing so over a large number of samples, we estimate the mean and standard deviation over $\bar{\beta}_G$. With this, we can estimate the probability of obtaining a measure as high as $\bar{\beta}_C$ by chance, using the normal distribution. If this probability is too low (i.e., below α, a parameter), we conclude that members of C abnormally benefit one another; we label all agents in C as coalition members.

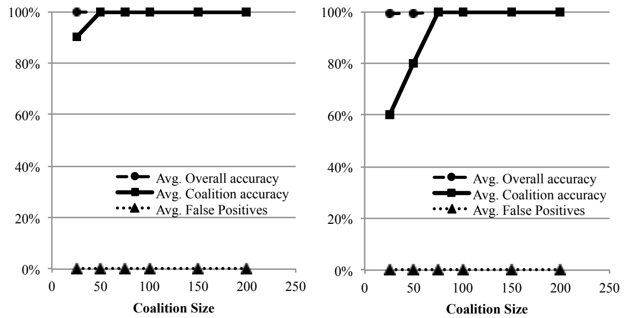
3. EXPERIMENTAL SCENARIO/RESULTS

Real-world colluders do not willingly reveal themselves as such, making it problematic to obtain real-world, labelled data that might be used for validation. Thus, the TREET marketplace testbed [2], populated by buying and selling agents, was used to validate our technique. Populations of 1000 agents made use of the Beta Reputation System [1]. Coalitions attempted to improve profits by bad-mouthing or ballot-stuffing. For each combination of parameter values, 10 trials were run (except where noted); the figures reported reflect the aggregate results across trials. The measure of benefit used to detect coalitions was the net sum of the review values given (counting a positive review as $+1$ and a negative review as -1), weighted by the dollar value of the transaction. After applying our technique, our classifications were compared to the true, hidden class of each agent to determine accuracy.

In the first set of tests, we evaluate the technique where exactly one coalition is present in the population. First, we consider coalition members engaged in bad-mouthing. These results are shown in Figure 1a, which contains three series. 'Avg. Overall accuracy', shows the percentage (across all trials) of agents that were accurately labelled as either coalition members or non-members. This metric can be misleadingly high, however, especially when the number of colluders is low. The second series, 'Avg. Coalition accuracy', depicts the fraction of coalition members that were accurately labelled as such. (This is equivalent to *recall*.) This shows that some colluders were missed for the smallest coalition size, but in general, performance is excellent. The third series, 'Avg. False Positives' shows the number of non-coalition members that were mistakenly identified as coalition members. (This value is equal to $1-$ *precision*.) This was zero, in all trials. Results for the ballot-stuffing case are depicted in Figure 1b; performance is slightly weaker, but very strong.

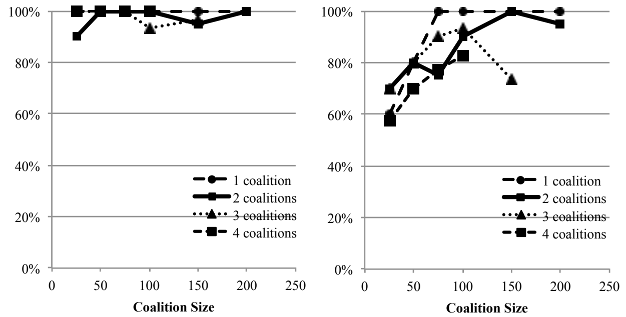
While performance is strong with exactly one coalition, it may be the case that there is no coalition present in a given population. Such situations provide a good test of the algorithm's resistance to false positives. We ran 120 trials with zero coalitions. In total, 3 agents were wrongly labelled as coalition members (a rate of 0.000025).

Just as a population might contain no coalitions, it might also contain multiple coalitions. We ran trials with up to 4 coalitions. The results for bad-mouthing are displayed in Figure 2a; those for ballot-stuffing are shown in Figure 2b. For clarity and brevity, false positive rates have been omitted from these charts. Again, they were zero in the vast majority of cases, and very low in the others.



(a) Bad-mouthing. (b) Ballot-stuffing.

Figure 1: Coalition detection accuracy, single coalition.



(a) Bad-mouthing. (b) Ballot-stuffing.

Figure 2: Coalition detection accuracy, multiple coalitions.

Overall performance is quite strong, in all cases. As in the single-coalition cases, performance is somewhat better for bad-mouthing than for ballot-stuffing; similarly, the general pattern of weaker performance on smaller coalitions is again evident in the ballot-stuffing data. Perhaps most importantly, note that there is no clear correlation between number of coalitions and performance: increasing the number of coalitions does not have the detrimental impact on performance that one might expect.

4. REFERENCES

[1] A. Jøsang and R. Ismail. The beta reputation system. 15th Bled Electronic Commerce Conference e-Reality: Constructing the e-Economy, June 2002.

[2] R. Kerr and R. Cohen. TREET: The Trust and Reputation Experimentation and Evaluation Testbed. *Electronic Commerce Research*, 10(3):271–290, 2010.

[3] B. Mehta and W. Nejdl. Unsupervised strategies for shilling detection and robust collaborative filtering. *User Modeling and User-Adapted Interaction*, 19(1-2):65–97, 2009.

[4] O. Shehory and S. Kraus. Feasible formation of coalitions among autonomous agents in nonsuperadditive environments. *Computational Intelligence*, 15(3):218–251, 1999.

[5] G. Sukthankar and K. Sycara. Robust recognition of physical team behaviors using spatio-temporal models. In *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 638–645, New York, NY, USA, 2006. ACM.