Efficient Statistical Methods for Evaluating Trading Agent Performance

Eric Sodomka, John Collins, and Maria Gini*
Dept. of Computer Science and Engineering, University of Minnesota
{sodomka,jcollins,gini}@cs.umn.edu

Abstract

Market simulations, like their real-world counterparts, are typically domains of high complexity, high variability, and incomplete information. The performance of autonomous agents in these markets depends both upon the strategies of their opponents and on various market conditions, such as supply and demand. Because the space for possible strategies and market conditions is very large, empirical analysis in these domains becomes exceedingly difficult. Researchers who wish to evaluate their agents must run many test games across multiple opponent sets and market conditions to verify that agent performance has actually improved. Our approach is to improve the statistical power of market simulation experiments by controlling their complexity, thereby creating an environment more conducive to structured agent testing and analysis. We develop a tool that controls variability across games in one such market environment, the Trading Agent Competition for Supply Chain Management (TAC SCM), and demonstrate how it provides an efficient, systematic method for TAC SCM researchers to analyze agent performance.

Introduction

The Trading Agent Competition for Supply Chain Management (TAC SCM) (Collins *et al.* 2005) is a market simulation in which autonomous agents act as manufacturers in a two-tier supply chain marketplace. Agents are responsible for purchasing components from suppliers, manufacturing finished products, and selling these products to customers via reverse auction. TAC SCM is interesting to many researchers because it provides a competitive environment in which dynamic, agent-based supply chain methods can be evaluated without the costs and risks associated with a real-world supply chain.

The primary measure of agent performance is total profit over a simulated year of activity. Availability and prices of parts in the procurement market, and unmet demand and prices in the customer market, are influenced by both the mix of agents, known as the *profile space* (Wellman *et al.* 2006), and by random variations in supply, demand, and other market parameters, which we refer to as the *market space*.

Like many market simulations, the high dimensionality of both the profile and market space, and the time required to complete a TAC SCM simulation (nearly an hour), cause systematic analysis to be intractable. The complexity is particularly burdensome during the testing phase of agent design— whenever changes are made to the agent, a reliable evaluation of performance requires a large number of simulations to marginalize over the profile and market space.

If the complexity of these market simulations could be reduced or controlled, researchers would better be able to analyze the performance of their agent and the decisions of their competitors. While the size of the profile space is extremely large, the researcher does have control in selecting the set of competitors. The market space, however, is not controlled by the researcher, but instead by the TAC SCM server. Our objective is to put control over the market space back into the researcher's hand.

In this paper, our contributions are twofold:

1. Propose a *new method for agent evaluation* in TAC SCM and other competitive simulation environments. We introduce a tool that enables us to evaluate agents with this method, and show through statistical power analysis that this method requires fewer simulations for significance testing than other methods currently in use.
2. Demonstrate how using this tool has *led to new insights* about agent interactions that are otherwise difficult to obtain, such as the effects of stochastic agents and specific market factors on various game metrics.

Related Work

Currently, most TAC SCM researchers test their agents in a market simulation environment provided by the Swedish Institute of Computer Science (SICS).[1] The server provided by SICS for agent testing is the same server used in the annual TAC SCM tournament. Such a framework, in conjunction with a user-submitted agent repository, allows for testing under conditions that are identical to the competition environment, to the extent that the actual competition agents are available. However, within this environment, there still

*Partially supported from the National Science Foundation under award NSF/IIS-0414466.

[1] http://www.sics.se/tac/page.php?id=1

exists a large amount of variability within the profile space and market space.

Managing Profile Variability

The first difficulty that arises in trading agent performance testing is choosing a set of agents to compete against, as it is intractable to test against every possible set of opponents available. Some approaches taken by different research teams are to run against different variations of their own agent (He *et al.* 2006), the set of dummy agents included in the TAC SCM server (He *et al.* 2006), and a fixed set of agents from the agent repository (Pardoe & Stone 2004; Pardoe, Stone, & VanMiddlesworth 2006). The University of Michigan team has attempted to find the set of agents closest to game-theoretic equilibrium, which could then be used as background opponents in agent testing (Wellman *et al.* 2006).

Managing Market Variability

In addition to the complexity introduced by the profile space, the TAC SCM server is specifically designed to generate highly variable market conditions in order to challenge the adaptability of the competing agents. Dimensions of variability include availability of components across component types and suppliers, customer demand across multiple market segments, randomly-assigned interest rates and warehousing charges, and other random processes, such as the customers' decisions when two agents offer identical prices.

A number of methods are currently in place by the TAC SCM community to manage the variability caused by the market space. The University of Texas and Southampton teams have run tests with different variations of their agent in the same simulation, causing each agent variation to see the same set of market conditions (Pardoe, Stone, & VanMiddlesworth 2006; He *et al.* 2006). While analyzing self-play performance may be worthwhile at times, there are situations where running an agent against itself (or a slightly modified clone) is not representative of how the agent would perform in a real-world situation, particularly if one usually competes against a set of opponents with less similar strategies. The simulation becomes particularly misrepresentative when more than two agent variations are being tested at once. Experimental results have shown that the set of agents competing in self-play is typically the least strategically stable of all possible profiles (Wellman *et al.* 2006). Performance tests also provide evidence that as more of the same type of agent are added to a game, that agent's performance decreases (He *et al.* 2006). Additionally, we would like to test over a variety of different competitor profiles, but when multiple agent variations are being tested concurrently, this profile space is constrained.

Instead of comparing profit values directly, another option is to use control variates to calculate *demand-adjusted profit* (DAP) (Wellman *et al.* 2006). This metric factors out profit variations caused by differences in demand, and allows agent profit levels to be compared between any two games. While this can greatly increase significance if a highly correlated variate is used, we will show that there are other factors influencing profit that are not considered by this demand-adjusted metric. Additionally, if new strategies are introduced to the profile space, the coefficients used for this control variate may be inaccurate, and will have to be recalculated.

The University of Minnesota team (Borghetti & Sodomka 2006) has created a controlled server which allows for researchers to run the same market conditions across multiple games. This paper expands on that work by allowing individual aspects of the market, such as supply or customer demand, to be repeated in multiple games, and shows how this control of market conditions benefits agent testing and evaluation.

Reusable Trajectories

Outside of the TAC SCM domain, Kearns et al have introduced *reusable trajectories* to reliably evaluate a restricted class of strategies in a partially observable Markov decision process (Kearns, Mansour, & Ng 2000).

The authors describe two methods of generating reusable trajectories. The first method uses a strong generative model to create a number of *trajectory trees*, and then evaluates the average return of each strategy in the given trees. The generative model required for the trajectory tree method is not available in TAC SCM.

The second method does not require a strong generative model. Instead, a number of *random trajectories* are generated, and the value of a strategy is evaluated as the average return of all random trajectories with the same observable history as the given strategy.

This method is difficult to apply to TAC SCM because there is an incredibly large number of actions an agent can choose each day. For example, agent offer prices can vary by as little as $1, and the quantity of parts requested from each supplier can range from 0 to many thousands. Thus, it is highly unlikely that an agent would ever choose the same actions as the random agent.

Approach

We now describe our method for evaluating agents in highly variable environments, and we compare it with other methods. We also provide a method for analyzing the effects of individual market factors on agent profit and order prices.

Testing in Paired Markets

Consider a situation where a researcher has an agent A, and has made some modifications to this agent, resulting in agent A'. It must now be determined whether or not these changes have actually improved or hurt agent performance, or resulted in no significant effect. In this situation, we propose the method of *paired market testing* to efficiently analyze agent performance.

First, we randomly choose N different sets of market conditions. We then run N simulations with agent A and N simulations with agent A'. *The N different sets of market conditions seen by one agent variation are the same as the N different sets of market conditions seen by the other.* We can thus compare the profit difference between the two agent variations for each corresponding set of market conditions

using a paired-means t-test, as performed by other TAC researchers when both agent variations compete in the same simulation (Pardoe, Stone, & VanMiddlesworth 2006). The paired market testing method removes the possibility of the agent variations interacting in the same game, which could potentially distort the results. We expect to see that using our method will result in a smaller standard deviation in profit than other methods that also test agent variations in seperate games.

Our method is only possible if market conditions can be repeated across games. We have developed a framework that allows for such market repeatability, and we describe it later. First, we present a way in which the performance of these different testing methods can be quantified.

Analysis of Testing Methods

A *statistical power analysis* (Cohen 1988) is used to determine the probability of obtaining statistically significant results from a hypothesis test that compares profit levels of two agent variations. The relationship can be described by four parameters:

1. Statistical power, the probability of correctly recognizing a difference between the two agent variations.
2. Sample size N, the number of simulations for each agent.
3. Significance level α, the maximum acceptable probability of incorrectly detecting a significant difference in profit between the two agents (type I error).
4. Effect size ES, the minimum profit difference required for us to consider there to be an important difference between the two agents.

If any three of these parameters are known, the fourth can be determined. In our case, we want to estimate how many simulations must be run in order to achieve significant results with a reasonably high probability. We use an α value of 0.05, and follow the convention established by Cohen that considers a "reasonably high" value for statistical power to be 0.8 (Cohen 1988). Effect size in this case is defined as the mean profit difference between the two groups divided by the standard deviation.We calculate the root mean square of the two standard deviations when independent groups are used (Cohen 1988), and the standard deviation of the profit differences when testing in paired markets (Gibbons, Hedeker, & Davis 1993).

We perform our analysis on three different agent testing methods: a baseline case, in which each agent is tested independently and average profit levels are compared; the demand-adjusted case, which is similar to the baseline, except that profit levels are first adjusted based on the level of demand in the simulation; and the paired market case, which is similar to the baseline, except that for each simulation, the market conditions faced by one agent are identical to the market conditions faced by the other.

Analyzing Effects of Individual Market Factors

While it is clear that market variability, as a whole, has a significant effect on agent performance, we have not yet discussed *which* specific market factors are responsible for performance variations. Such knowledge would be particularly valuable during the agent design process—variable factors that cause little to no variation in the final result can be ignored in future research, thus reducing dimensionality and simplifying learning models. We may also find market factors that look promising to focus on in future research, because they are shown to be significant in the model output. Once we find the important variability factors, future research can examine exactly how they affect agent performance.

One possible approach is to reduce or remove variability of all market factors except for one, and examine how varying that single factor affects prices and profit levels. However, such an aggressive manipulation of the market space runs the risk of producing results that are misrepresentative of an actual simulation.

Instead, the method we use is a *factors fixing sensitivity analysis* (Saltelli 2002). First, consider all market factors (such as individual supplier and customer demand walks) that could potentially affect some simulation output (such as daily order prices or profit levels). The interactions of the market inputs with the simulation output can be described with a high dimensional model representation:

$$\begin{aligned} Y = \quad & f(\vec{X}) = \textstyle\sum_{i=1} f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) \\ & + ... + f(X_1, X_2, ..., X_n) \end{aligned}$$

where the output Y, which could be order price, is a function of the various market factors X_i, each of which can have an effect individually or through some joint interaction with other factors.

If each market factor is controlled by its own pseudo-random sequence, then we can repeat the pseudo-random sequences across simulations for all market factors but one, and allow the remaining factor to vary across multiple simulations. The "uncontrolled" factor, the one for which we do not force a repeated pseudo-random sequence, affects every term in the above equation which contains this factor. We can subsequently observe how the variability of that single market factor X_i affects the variance of the output. The variance of the output Y, given repeated pseudo-random sequences for every market factor except for X_i, is represented as $V(Y|X_{-i})$.

Note that the variance in this case is likely dependent upon the pseudo-random sequences we have chosen to repeat. Thus, it is important to run across many different repeated pseudo-random sequences, or market conditions. If we run simulations with a number of different repeated sequences for factors besides X_i, we find the expected, or average, output variance $E(V(Y|X_{-i}))$. If this is not statistically distinguishable from the inherent variance caused by stochastic agents alone, then the market factor X_i must not have any significant influence on the specified output.

We can thus define the total sensitivity index for market factor X_i to be the ratio of the expected value of the variance that X_i is contributing, either individually or through interactions with other market factors, to the total variance of the output.

$$S_i^T = \frac{E(V(Y|X_{-i}))}{V(Y)} \tag{1}$$

Ideally, we would like to get total sensitivity indices for each of the individual factors in the model. However, because each TAC SCM game takes almost an hour, running enough games to find sensitivity values for every factor is impractical. We settle for a method that first considers groups of factors, such as supply and demand, and treats them as single factors in the factors fixing method. While there may be some interactions that are overlooked, our abstraction allows us to greatly reduce the complexity of the space to show which are at least the most promising sets of factors to decompose in the future.

Controlling the Market Space

In order to support our testing approach, we extend the TAC SCM server to allow for repeatable pseudo-random sequences of any individual market factor or combination of factors. The actual values of these factors are not explicitly selected; instead, the research may decide *which* of these factors will vary across games, by selectively locking the starting conditions of the various pseudo-random processes. We refer to our extension of the TAC SCM server as the *controlled server*. In total, there are thirty-seven different pseudo-random processes that can be controlled. Any process that is uncontrolled will vary in the same way that it does in the original server.

Experimental Results

We now demonstrate how using the controlled server can make comparisons between agent variations more efficient. We also show how the controlled server can be used to learn information about the game that would otherwise be difficult to obtain, such as the effects of stochastic agents on the game, and which specific market factors have the greatest impact on game outputs.

The agents used for our experiments include: `DeepMaize` from the University of Michigan, `Maxon` from Xonar Inc, `MinneTAC` from the University of Minnesota, `PhantAgent` from Politechnica University of Bucharest, `RationalSCM` from the Australian National University, and `TacTex` from the University of Texas. Agents were obtained from the TAC SCM agent repository. We selected these agents because they were the most competitive agents that were readily available to us. Of the six agents, five were finalists in the 2006 competition, and the sixth, `RationalSCM`, was a finalist in 2005. `TacTex` was the winner of both the 2005 and 2006 competition.

Improved Significance Testing

We have run a number of performance tests to demonstrate the value of using the controlled server. We use two different versions of the `TacTex` agent—one that competed in 2005, and one that competed in 2006. We would like to know how much performance differs between these two agent variations.

To determine this, we run forty simulations with `TacTex05` and forty simulations with `TacTex06`. With the original TAC SCM server, market conditions are different in each of the 80 games. We compute the difference in average profit between `TacTex05` and `TacTex06`, as well as the standard deviation associated with that difference. We analyze our results using both simple profit and demand-adjusted profit.

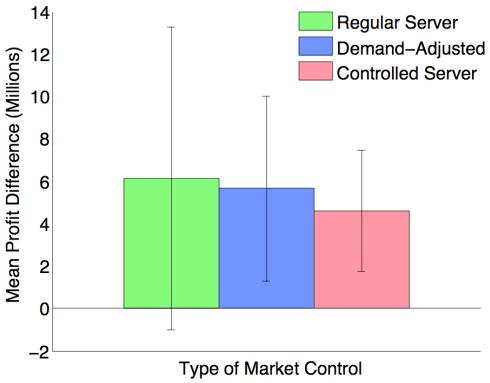

Figure 1: Standard deviation comparison of different performance evaluation methods. Results are shown for two versions of `TacTex`, for simple profit and demand-adjusted profit with the regular server, and for profit with the controlled server.

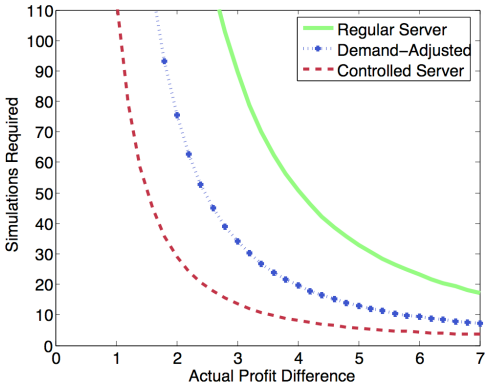

Figure 2: Number of samples required to detect, with significance, a given profit difference between two agents. These results are based upon a power analysis where $\alpha = .05$ and $power = .80$.

We then run tests with the controlled server. We randomly choose forty different sets of pseudo-random sequences, each set defining some market conditions. We again run forty simulations with `TacTex05` and forty with `TacTex06` using paired market testing; that is, the set of forty market conditions seen by `TacTex05` are the same as the forty seen by `TacTex06`.

Figure 1 shows the mean difference and the standard deviation of the difference for each of these methods. Clearly, the standard deviation of the profit difference is substantially smaller using the controlled server. These results also support the use of demand-adjusted profit, but because DAP is only adjusted for demand, it gives a slightly higher standard deviation than the controlled server method, which considers all market conditions.

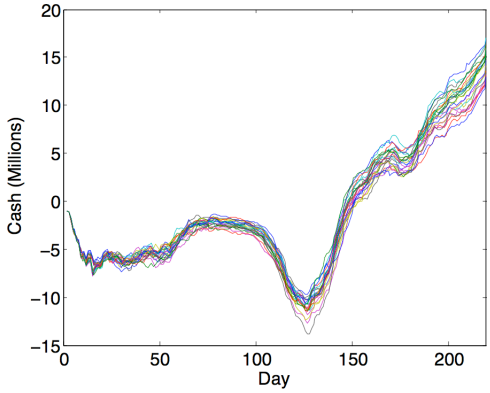

Figure 3: Daily profit values for TacTex06 using a set of repeated market conditions. Each line represents a different simulation. Other agents and fixed market conditions resulted in similarly low standard deviations in profit.

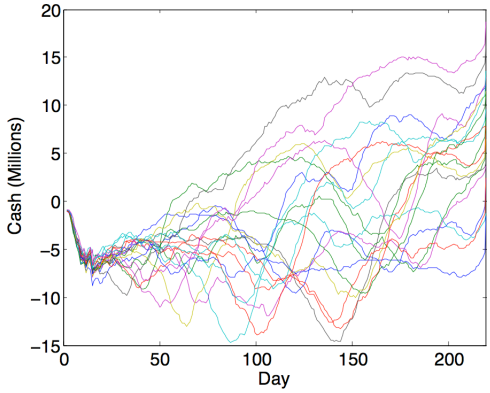

Figure 4: Daily profit values for TacTex06 using random market conditions. Other agents had similar profit variability.

Using the standard deviation present in each method to approximate effect size, we can estimate the number of samples required for each method to achieve statistical significance. Figure 2 provides sample size estimates required to detect a given profit difference between the agent variations.

The results suggest that our controlled server allows for researchers to perform significance tests with non-interacting agents in fewer games than has previously been possible. This added market control can be used not just for significance testing, but also for analyzing different interactions within the game that were previously obfuscated by market variability. We now present some additional ways we have performed analysis using the controlled server.

Measuring Effects of Stochastic Agents

The adaptive methods of some TAC SCM agents have created an incentive for other agents to behave stochastically.[2] The presence of these stochastic agents in TAC SCM is not disputed, but the effect they have on the simulation outcome is currently unknown. Even when the profile space and the market space are fixed, a single stochastic agent will cause output variability through both its own decisions and by affecting the decisions of other, potentially deterministic, agents.

Our tests attempt to measure the noise caused by these stochastic agents. To do so, we run a number of simulations with a controlled server and compare the results to those from a traditional random environment, where market conditions vary between simulations. In the controlled environment, the same profile space and market conditions are repeated across N simulations, and a measurement of standard deviation in profit is obtained. Because the market space and profile space are repeated across simulations, they cannot be contributing to profit variation from one simulation to the next. Thus, any observed variation in simulation output must be caused by random processes within the agents themselves.

Our results are presented in Table 1. We show each agent's average overall profit value (μ_R) and standard deviation (σ_R) in random market conditions, and the expected standard deviation ($E(\sigma_C)$) in controlled market conditions. To ensure the results we obtain are not a product of the market conditions we've randomly selected to repeat with the controlled server, we run the same tests with M different sets of market conditions, and average the results to get an expected value for the standard deviation (in these tests, $N = 20$ and $M = 5$). As it can be seen in the Table, there is a large difference in the standard deviations between the random and controlled cases.

Agent	μ_R	σ_R	$E(\sigma_C)$
TacTex06	8.0878	5.558	0.990
DeepMaize	6.303	5.186	1.067
PhantAgent	6.255	6.611	0.930
Maxon	1.99	4.101	0.782
MinneTAC	-1.331	3.470	0.867
RationalSCM	-1.623	5.301	1.024

Table 1: Average overall profit and standard deviation values for the agents playing with random vs controlled market conditions

See Figures 3 and 4 for a specific illustration of the data summarized in the table. Our results indicate that, while stochastic agents are indeed present in TAC SCM, their random behaviors do not have a significant effect on the profit levels of the agents. Similar results were observed when comparing the variability of the daily order prices instead of profit. From these results, it is clear that market conditions are far more influential in determining agent profit and order prices. This is promising because it shows that the source of variability we cannot control (stochastic behavior in opponent decision processes) does not have a large effect on simulation output, while the source of variability we can

[2]While agents are able to adapt *across* simulations during the TAC SCM competition, the agents we test with here are only able to adapt *within* the individual simulations.

control (market conditions) does have a large effect on agent performance and daily order price.

Measuring Effects of the Market Space

We applied the abstract factors fixing method with daily order price as its output across 170 different TAC simulations. For five different sets of market conditions, we ran ten different simulations, each time allowing either supply to vary, demand to vary, or having all factors fixed. We also used twenty random games from our tests in previous sections and the same profile space. From these tests, we were able to compute the total sensitivity indices for supply, demand, and the situation where everything is fixed but the stochastic behaviors of the agents.

The results of our analysis can be seen in Figure 5 for a single product. The analysis shows, in general, that demand plays a much larger role in determining order prices than does supply, which supports the use of demand control variates as a performance metric (Wellman *et al.* 2006). However, we can see that supply does have an influence on order price, as well. The shape of the curve is also interesting – for nearly all product indices, there are well-defined spikes at specific intervals, which suggest that these may be times when order prices are most affected by demand factors. However, such a statement cannot be made with certainty until we perform the factors fixing method with a greater number of different market conditions. As our testing continues, we expect this process to reveal a good deal of information about the interactions market conditions have with order prices and profit levels.

Conclusions and Future Work

We have proposed a method for more efficient agent testing and evaluation, and introduce a tool that makes such an evaluation possible. We show through a statistical power analysis that paired markets testing method requires fewer games to be run for significance testing. We have also used this tool to measure the amount of variability caused by stochastic processes in agents, and have demonstrated how researchers can determine which market factors most influence order prices and profit levels.

A more in-depth sensitivity analysis is a main priority for future work. Specifically, if we wish to continue using the abstract method of fixed factors, it will be important to show that the reduction in complexity is worth the loss of information such an abstraction causes.

While our work has been focused primarily on techniques that can be used to control the variability in the market space, we have only performed minimal tests in using these methods to better understand the interactions of the profile space. Just as the profile space was held fixed in our tests to simplify analysis, tests of the profile space could similarly use the controlled server to control variability from the market space.

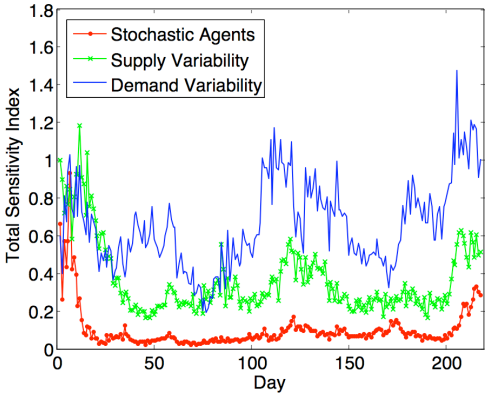

Figure 5: Total sensitivity indices for the daily order price of a particular product (Pintel CPU, 5.0 GHz), given variability of demand, supply, and stochastic agents. Results were similar with other products.

References

Borghetti, B., and Sodomka, E. 2006. Performance evaluation methods for the the trading agent competition. In *Proc. Nat'l Conf. on Artificial Intelligence*, 1857–1858.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum, 2nd edition.

Collins, J.; Arunachalam, R.; Sadeh, N.; Ericsson, J.; Finne, N.; and Janson, S. 2005. The supply chain management game for the 2006 trading agent competition. Technical Report CMU-ISRI-05-132, Carnegie Mellon University, Pittsburgh, PA.

Gibbons, R.; Hedeker, D.; and Davis, J. 1993. Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics* 18:271–279.

He, M.; Rogers, A.; Luo, X.; and Jennings, N. R. 2006. Designing a successful trading agent for supply chain management. In *Proc. of the 5th Int'l Conf. on Autonomous Agents and Multi-Agent Systems*.

Kearns, M.; Mansour, Y.; and Ng, A. 2000. Approximate planning in large pomdps via reusable trajectories. In *Advances in Neural Information Processing Systems 12*. MIT Press.

Pardoe, D., and Stone, P. 2004. Bidding for customer orders in tac scm: A learning approach. In *AAMAS04: Workshop on Trading Agent Design and Analysis*, 52–58.

Pardoe, D.; Stone, P.; and VanMiddlesworth, M. 2006. Tactex-05: An adaptive agent for TAC SCM. In *AAMAS06: Workshop on Trading Agent Design and Analysis (TADA/AMEC)*.

Saltelli, A. 2002. Sensitivity analysis for importance assessment. *Risk Analysis* 22(3).

Wellman, M. P.; Jordan, P. R.; Kiekintveld, C.; Miller, J.; and Reeves, D. M. 2006. Empirical game-theoretic analysis of the tac market games. In *AAMAS06: Workshop on Trading Agent Design and Analysis (TADA/AMEC)*.