

An Experimental Evaluation of Bidders' Behavior in Ad Auctions

Gali Noti^{*}
School of Computer Science
and Engineering
Hebrew University of
Jerusalem, Israel
galinoti@gmail.com

Noam Nisan
Microsoft Research and
School of Computer Science
and Engineering
Hebrew University of
Jerusalem, Israel
noam@cs.huji.ac.il

Ilan Yaniv
Department of Psychology
Hebrew University of
Jerusalem, Israel
ilan.yaniv@huji.ac.il

ABSTRACT

We performed controlled experiments of human participants in a continuous sequence of ad auctions, similar to those used by Internet companies. The goal of the research was to understand users' strategies in making bids. We studied the behavior under two auction types: (1) the Generalized Second-Price (GSP) auction and (2) the Vickrey–Clarke–Groves (VCG) payment rule, and manipulated also the participants' knowledge conditions: (1) explicitly given valuations and (2) payoff information from which valuations could be deduced. We found several interesting behaviors, among them are:

- No convergence to equilibrium was detected; moreover the frequency with which participants modified their bids increased with time.
- We can detect explicit “better-response” behavior rather than just mixed bidding.
- While bidders in GSP auctions do strategically shade their bids, they tend to bid higher than theoretically predicted by the standard VCG-like equilibrium of GSP.
- Bidders who are not explicitly given their valuations but can only deduce them from their gains behave a little less “precisely” than those with such explicit knowledge, but mostly during an initial learning phase.
- VCG and GSP yield approximately the same (high) social welfare, but GSP tends to give higher revenue.

Categories and Subject Descriptors

K.4.4 [Computers and Society]: Electronic Commerce

^{*}Partially supported by a grant from the Israeli Science Foundation (ISF) and by the Google Inter-university Center for Electronic Markets and Auctions.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
<http://dx.doi.org/10.1145/2566486.2568004>.

Keywords

Advertising Auctions; Experimental Economics

1. INTRODUCTION

1.1 Background

Ad auctions generate billions of dollars of revenue for Internet giants such as Google, Microsoft, or Yahoo and indeed are famously the “killer application” of algorithmic mechanism design. The most common format of these ad auctions (to date) sells a sequence of “ad positions” on a web page viewed by some user, where the different positions have different Click-Through Rates (CTR), i.e., likelihoods of being followed by the user. This format is usually used for advertising on search-results web pages, and is also referred to as “sponsored-search auctions” or “keyword auctions.” Advertisers place bids (for combinations of keywords and user population) “per click” and the highest bidder – after adjustment for ad quality – gets the best position, the second highest gets the second-best position, and so on. Such auctions were first studied in [11, 3] and have since received much attention.

Simplifying slightly (see fuller descriptions, e.g., in [8]), the most often used pricing rule, the Generalized Second-Price (GSP) auction, charges each advertiser the price (per click) offered by the bidder following him in the ranking of bidders (again, after adjusting for ad quality), i.e., the minimum price needed to keep the position that he won. This is different both from first-price auctions that charge the bid itself and from the more complex Vickrey–Clarke–Groves (VCG) payment rule that charges according to the marginal gains from getting better positions and theoretically ensures “incentive compatibility”, i.e., that bidders never gain from strategic manipulation. The reason that the first-price auction is not used (anymore) for sponsored-search auctions is that it requires each bidder to continuously monitor his competition and modify his bids so as to overbid it just slightly. The reason that the theoretically appealing VCG auction is not commonly used in sponsored-search auctions is less clear, with various explanations offered or refuted at various places (e.g., cognitive complexity, lower revenue, dependence on CTRs), as well as suggestions to indeed switch to VCG auctions at other places.

While there has been considerable work on the theoretical aspects of ad auctions as well as much work analyzing data from actual ad auctions, there has so far been very little

experimental evaluation of ad auctions; we are only aware of [4, 2]. This is in contrast to the very large body of work in experimental game theory; see, e.g., the surveys in [7, 6]. Such experimental evaluation may be useful despite the large amount of data from real-life auctions, as it allows us to ask “what if” questions and to isolate different aspects of user behavior that cannot be answered based just on real-world data. Such questions include possible effects of changing the auction format to VCG, questions related to how well the advertisers “know” their valuations, as well as more detailed modeling of bidder behavior.

1.2 Our Experiment

In this experiment we recruited participants in groups of five, and each group played a game involving a continuous sequence of ad auctions. Specifically, in each instance of the experiment, five participants simulated the roles of advertisers and competed in a stream of ad auctions that lasted 25 minutes. We used a flexible auction experimentation software platform we developed that enabled us to control the auction details as well as players’ knowledge and values. The auctions were conducted continuously, one auction in each second; thus there were 1500 auctions within the 25-minute period. The participants could modify their bids at any time, and each auction was performed with the current set of five bids. Each player was assigned a “valuation,” i.e., a monetary value that he obtains from each user who clicks on his ad (we used 21, 27, 33, 39, and 45 “coins”). Each ad auction sold five ad positions in decreasing order of CTRs from top (first on “page”) to bottom (we used 38%, 29%, 20%, 11% and 2%), and every time an advertiser whose valuation was v won a position with CTR p , he got an income of $p \cdot v$ from that auction. This income was added to his balance and the appropriate payment according to the auction rule was deducted from his balance. The players were given a graphical user interface in which they could modify their bids as often as they wished, and follow the results of the auctions until then. Figure 1 shows a screenshot of the user interface.

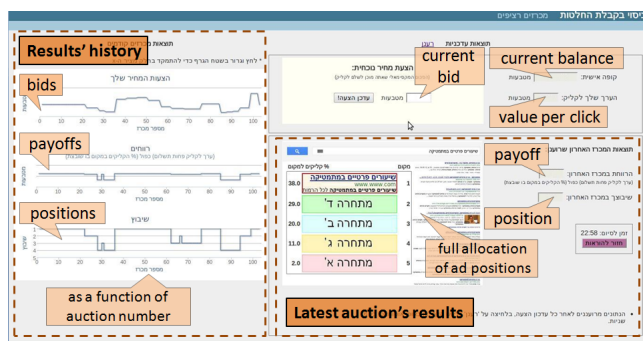


Figure 1: Game interface

We paid particular attention to the repeated auction dynamics. Most analyses of ad auctions focus on a single auction. However, in reality, as well as in our experiment, a long sequence of auctions is played in which players can learn, can signal to each other, etc., and we wanted to study how user behavior changes with time. The experiment had a two-way

(2x2) design; thus there were four experimental conditions. The two factors were:

1. **Payment Rule:** We compared the (theoretically appealing) VCG payment rule with the (commonly used) GSP payment rule.
2. **Valuation Knowledge:** While the starting point of analyzing behavior in auctions is the “valuation” of the bidder, it is questionable to what extent users fully know their valuations. We compared the case where bidders were directly provided with their valuations (Given Value (GV)) and were explained its significance, and a case where bidders were not directly given their valuations, but rather could only see their payoffs – information from which the valuation may be deduced, but could alternatively be directly used to guide the bidding (Deduced Value (DV)).

There were a total of 24 experimental sessions, 6 sessions for each of the 4 experimental conditions (thus we get 12 sessions for each factor). The groups (of five players each) were randomly assigned to the four experimental conditions, giving a total of $n = 120$ participants. For further details regarding the experimental setup see Section 2.

Before going into our results, let us reflect shortly on what may be theoretically expected from rational players engaged in such auctions. For VCG, the prediction is simple: truthful bidding is a dominant strategy so a strong prediction is that players will simply bid their true valuation. Even if the valuation is not explicitly given but can only be deduced, we can expect the players to quickly learn their valuation and then settle on the (deduced) truthful bid. For GSP, a pure equilibrium exists, and in particular the well-studied VCG-like equilibrium (the “lowest-LEF equilibrium” [11, 3]), is naturally reached by a sequence of best replies [1, 9]. Therefore, we would expect convergence to equilibrium. This equilibrium will certainly have participants “shading their bids” (i.e., always bidding less than their true valuation) and will, possibly, be the VCG-like equilibrium. We examine whether, and to what extent, people follow the theoretical expectations or deviate from them either due to psychological biases or due to the repeated setting which provides rational players with wide-ranging opportunities for signaling and cooperation.¹

1.3 Results

We start by analyzing bidder activity (how often bids are modified) over time. As discussed above, for VCG, theory suggests that bidders bid the dominant strategy and never modify their bid (at least after a short learning phase, if valuations are not given but rather must be learned). For GSP, theory would also suggest convergence to equilibrium. Surprisingly, this was not the case in the study: not only do we not observe any convergence to fixed bids for VCG or GSP, but we actually observe an *increase* in frequency of bid modifications as time progresses, and this is observed across all four experimental conditions. As may be expected, we see more frequent bid modifications when players are not given

¹Our experiment, technically, has a finite horizon and so the folk theorem that ensures this wide range of equilibria does not formally apply. However, this seems to be too strict an interpretation and one would naturally assume that humans treat our experiment as an infinite-horizon setup.

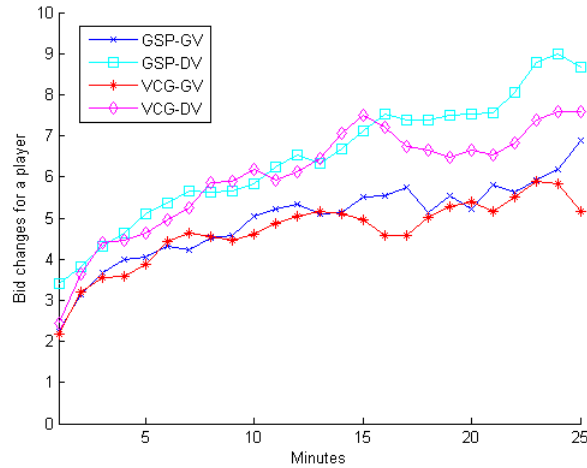


Figure 2: Bid modifications frequency²: The average number of bid changes per minute made by a player, in each of the experimental conditions over time.

their valuations but rather must deduce them, although unexpectedly the gap does not diminish after an initial learning phase. Figure 2 shows the average activity level (i.e., frequency of bid change) as a function of time across all auctions. As theoretically expected, we could not find any evidence that this high level of activity was beneficial in any way to the bidders. We did find that bidders with lower types (values) tended to modify their bid more often, perhaps indicating a psychological discomfort of being in lower positions (see Section 3).

Since we see that bidders keep modifying their bids, we try to identify regularities in these modifications: how do they change their bids? We have found good evidence of a better-reply behavior – as opposed to just randomly mixing – across all auction settings. In particular, as a response to the last-seen bids of the others, we find that changes in bids were 50% more likely to be better responses than worse responses. Moreover, it seems that this better response to the last-seen actions of the others also provides a significant gain in utility when played against the actual (updated) bids of the others (see Section 4 for some delicate issues, though). As expected, these gains are largest in the early learning phase, but they are maintained throughout the auctions. Figure 3 shows the effects of bid modifications as a function of time both according to the last-seen bids of the others and according to the actual updated bids of the others.

Our next analysis compares the actual levels of the bids made by participants to those expected by theory. In the VCG condition, theory makes a firm prediction that participants should bid their true valuations as this is a dominant strategy. In the GSP condition, theory does not make as firm a prediction, since the equilibrium is not unique, but would point to the VCG-like equilibrium³ identified in [9, 3]; in any case, bidders should theoretically bid below their

²All line graphs presenting progress over time in this paper were smoothed using a 3-point moving average.

³In our settings, the bids in this equilibrium are calculated to be (17.18, 21.6, 25.14, 28.42, > 28.42).

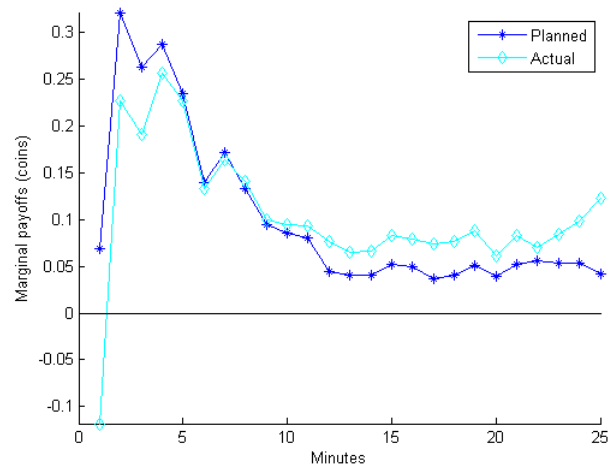


Figure 3: Impact of bid modifications on bidders' payoffs: The average marginal payoffs resulted from bid changes over time, according to both planned gains (i.e., against last-seen bids of others) and actual gains (i.e., against actual updated bids of others).

true values. As expected, players acted strategically, making lower bids, on average, in the GSP auctions than in the VCG auctions. In the VCG auctions, the average bids were not far from the predicted ones (i.e., the true values), while in the GSP auctions the average bids tended to be consistently higher than those predicted by the VCG-like equilibrium. Figure 4 shows the average bids according to player type and experimental condition.

Figure 5 shows the frequency of overbidding behavior as a function of time in the different auction conditions. While it may seem puzzling that overbidding behavior persisted so often, as it is strategically harmful (see Section 5), this finding is in line with previous observations of overbidding even in simple second-price auctions (see [5]). Particularly noteworthy is the low frequency of truthful bidding under the VCG conditions (less than 20% even when valuations were explicitly given). This is despite the fact that it was explained to the participants that bidding the true value is a dominant strategy. See Section 5 for more details.

Finally we consider the bottom lines, namely, the social welfare and revenue achieved in each condition. Figure 6 shows the average social welfare and average revenue achieved as a function of time. The GSP and VCG mechanisms led to the same high level of social welfare, about 82% of the optimum,⁴ with no clear advantage to VCG, in contrast to the report in [4]. Social welfare increased over time, attaining over 85% of the optimum level towards the end of the session. Participants were at a clear disadvantage when values had to be deduced (compared with the condition where values were given explicitly), but even then 80% of the op-

⁴This is in a scale giving the correct sorting of bidders 100, and giving the exact opposite order 0; since even the opposite order achieves about 2/3 of the social welfare in our setting, the actual fraction of social welfare obtained is about 95%.

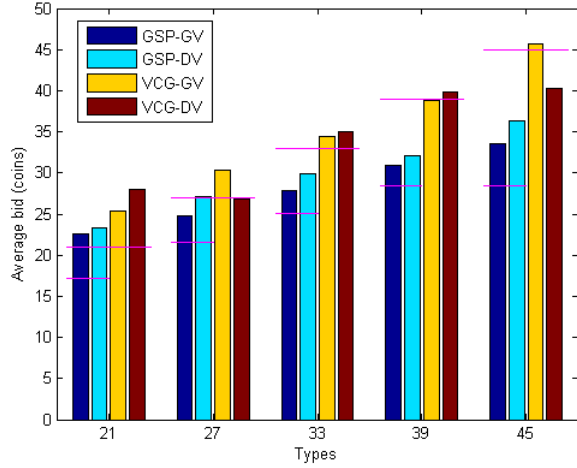


Figure 4: Average bid level: The average bid for each of the experimental conditions and player types. For each type, the horizontal higher and lower lines mark the truthful and VCG-like equilibrium bids, respectively.

timium was achieved, with most of the loss occurring in the initial phase, while participants were still learning. The GSP auctions had a clear advantage in terms of revenue, corroborating the findings of [4] and supporting the common use of GSP in practice. We note though that the gap was largest at the beginning of the session and decreased with time. In particular the revenue of GSP decreased in the initial learning phase as bidders learned to shade their bids while the revenue of VCG increased. In both auction mechanisms revenues were above the predicted equilibrium revenue: the auctioneer captured 76% of the social welfare using GSP and 68% of it using VCG, but the theoretical prediction for the revenue share is only 63% (obtained in a VCG auction with our parameters and truthful players). See Section 6 for more analysis of the revenue.

2. EXPERIMENTAL SETUP AND PROCEDURE

We ran laboratory simulations of ad auction games, in which five human players compete on five ad positions in a sequence of ad auctions. Each game lasted 25 minutes during which 1500 auctions were performed at a rate of one auction per second. Players could continuously change their bids, and each auction was performed using their most updated bids.

The experiment involved a 2x2 between-participants design and thus had four conditions. The first factor was mechanism: we compared behavior under two mechanism settings, GSP and VCG. The second factor was information structure: players were either directly given their value per click (i.e., Given Value (GV)), or could only indirectly deduce this value (i.e., Deduced Value (DV)).

We conducted the experiment using a web application we developed for the experimentation of ad auctions. This new platform allows us to control the various aspects of the ad

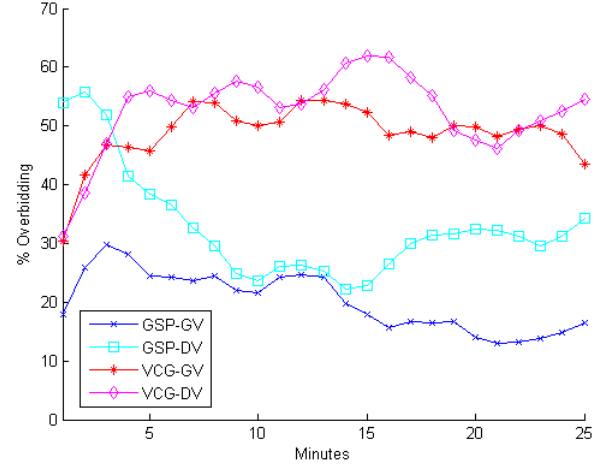


Figure 5: Overbidding frequency: The percentage of bids above personal values in each of the experimental conditions over time.

auctions’ unique environment, as well as provides a realistic user-interface simulating a search-results web page (see screenshot in Figure 1). All phases of the experiment – the instructions phase, the ad auctions game, and a short questionnaire run in the end – were conducted using this software.

In the beginning of each experiment session, each of the five players privately read the instructions from his computer screen.⁵ The players were presented with a scenario involving a competition among five players for placing ads on an internet page. Specifically, they were instructed to imagine that they were each advertising private lessons on a search-results web page, and that they were competing with the other four participants on ad positions. The players were given information on the Click-Through Rates (CTR) associated with each ad position, from top to bottom: {38%, 29%, 20%, 11%, 2%}. Thus, the positions were displayed in a decreasing order of CTRs, so that the position on the top of the page received the highest CTR.

Each player was randomly assigned a different monetary value from the set: {21, 27, 33, 39, 45}, and was told that: “... each click on your ad is worth to you an expected income of V game coins.” In the given-value conditions the “V” was replaced with the exact value, while in the deduced-value conditions it remained a placeholder. In both value-information conditions players were told that their value remained constant throughout the game, and that it could differ from one player to another. Players further learned that in each auction, positions were allocated according to the decreasing order of their most recent bids (ties were broken randomly), and that after each auction their resulting payoff (gain minus payment to the search engine) would be added to their total balance. They were further told they would be paid for their participation according to their final balance (in addition to a fixed participation payment). This provided genuine incentives for playing the auctions game.

⁵Players were asked not to speak to each other during the session, and their computers were separated by partitions.

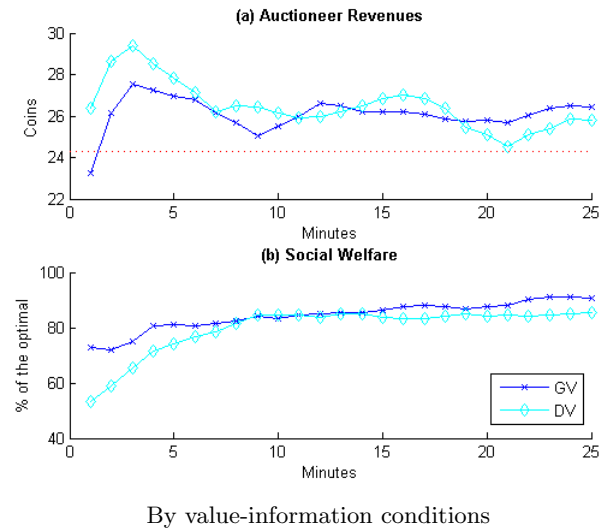
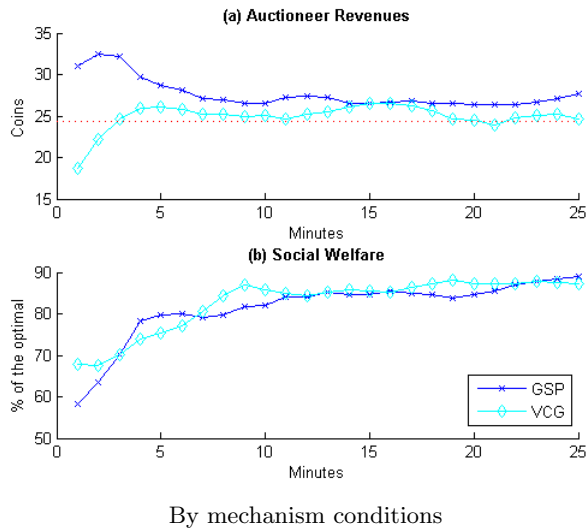


Figure 6: Social welfare and auctioneer revenues: Results by mechanism conditions (on the left) and by value-information conditions (on the right). For each factor: (a) The average revenues for the auctioneer achieved over time. The horizontal dotted line marks the VCG-equilibrium revenue outcome. (b) The average social welfare achieved over time (see Section 6 for the scaling method that we used).

The instructions given in the two mechanism conditions differed in two respects: first, the description of the auction pricing rule was written according to either the GSP or the VCG mechanism; second, the players were advised about their best strategy in the following way (based on theory): in the VCG condition the players were informed that in each auction, bidding their true value would always be their most beneficial strategy, independent of the other players' bids; the GSP players were informed that the bid that would give them the highest payoff might depend on other players' bids. Such advice was given since we were interested in testing people's behavior when they were aware of the real properties of the auction mechanism.

After all players finished reading and correctly answered several comprehension questions, a summary of the instructions was read publicly (skipping the private values), to let the players know it was common knowledge. Then, an initial bidding page was displayed, in which players were asked to state how much they were willing to pay per click for their ad. Once all initial bids were given, the repeated auctions game started, as described above.

During the game, players were given a graphical user interface in which they could change their bid as often as they wished, and follow the results of the last (i.e., most recent) auction and the history of the results until then (see screenshot in Figure 1). In particular, each player was provided with an input field for updating his bid (restricted to 2 digits before and after the decimal point), with his value per click (only in GV condition), and with the following feedback: his current balance; the payoff and position he had achieved in the last auction instance; the full allocation of ad positions in the last auction (the competitors were represented by colored labels); and three dynamic graphs of his results until then, describing the history of bids, payoffs, and posi-

tions.⁶ This information was automatically updated either every 6 seconds or after the player changed his bid or after he pressed the “refresh” button. In addition, a timer at the bottom of the screen indicated the remaining time in the game. Finally, after 25 minutes the game was over and each player was shown a summary of his profits. Thereafter players filled out a short questionnaire, after which they were paid privately and had thus completed their participation.

The participants in the experiment were undergraduate students at the Hebrew University of Jerusalem, who did not come from any specific discipline. They were run in groups of five in separate sessions that lasted up to 50 minutes each. Each session involved one continuous ad auctions game. There were 6 such sessions in each of the 4 experimental conditions. Thus, all together there were 24 experimental sessions with a total of 120 participants. The average payment for a participant was 65 NIS (18 US dollars).

3. BIDDING ACTIVITY

The bidders are put in a continuous stream of auctions, and are allowed to change their bids as often as they want. Once a bid is put in, it keeps being used by all coming auctions, until the bid is changed. Theoretical analysis expects players to bid truthfully in VCG auctions as this is a dominant strategy and this was explicitly explained to the participants before the auctions started. Also in GSP, we would expect convergence to the VCG-like equilibrium. Therefore, bidding activity is generally expected to disappear after a short learning phase, or at least to decrease with time.

Surprisingly, we observed a high level of bidding activity in each of the four conditions of our experiment, which significantly increased in frequency throughout the game (see Figure 2). Specifically, the average number of bid modi-

⁶As in real life, our players were not provided with either the other players' valuations or their bids, and could only infer them from allocation results.

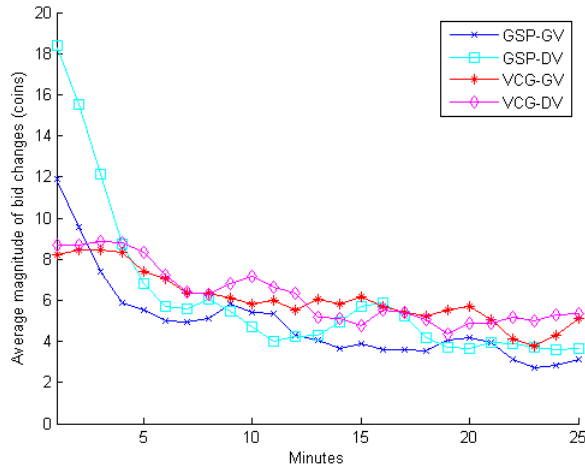


Figure 7: Bid modifications magnitude: The average magnitude of bid changes in each of the experimental conditions over time.

fications made by a player in the last third of the game was significantly higher than it was in the first third of the game (Wilcoxon paired two-sided signed rank test, $p < 0.05$ for each condition except for VCG-DV for which $p = 0.06$). In the last third of the game the high level of frequency reached an average of 6.56 bid modifications per minute for a single player. Further examination shows that the magnitude of the bid modifications, although decreasing in the beginning, did not become trivial (see Figure 7). Even in the last third of the game the average magnitude of each bid modification was 4.29 coins – rather considerable compared to the difference between the valuations of adjacent player types which was 6 coins in our settings. Therefore we see that contrary to theoretical expectations, bids did not converge to a fixed level.

We tried to test whether being more active improved the players’ long-term payoffs, but, as theoretically expected, this was found not to be the case, and no significant correlation was found. We also tested whether the more active bidders also bid higher values, perhaps signifying an escalation in their commitment (see [10]). However, we did not find a consistent connection between the bid levels and the activity level of a player either.

We did find that the lower-type players (i.e., those with lower values) updated their bids more frequently than higher-type players. Specifically, by testing players in all experimental conditions, we found a significant negative correlation between the player’s type and the number of times he modified his bid during the game ($N = 120$, $p < 0.02$). This pattern was consistent during the whole game, and is presented in Figure 8. The fact that lower types were more active may suggest a psychological difficulty with being in lower positions. This is in spite of the fact that being in the lower positions was more beneficial for them, and they knew their value per click was constant throughout the game and it was not in their ability to change it.

As may be expected, we found that under the deduced-value (DV) settings bidders were more active than under the

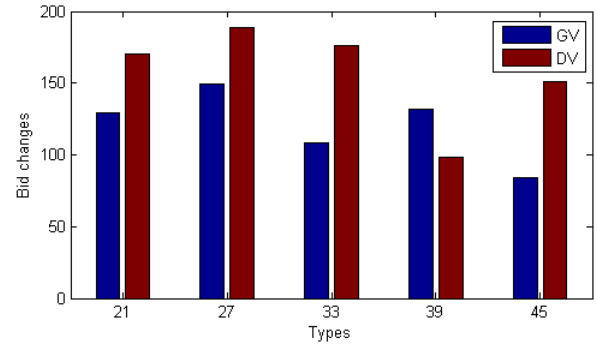


Figure 8: Bid modifications: The average number of bid changes (throughout the game) as a function of player type and value-information condition.

given-value (GV) settings (see Figure 8). However, while this result may seem reasonable, assuming that DV bidders had to compensate for their lack of information by learning and exploring the environment more intensively, we also found that this gap did not diminish after an initial learning phase. Specifically, in the DV sessions bidders performed significantly more bid modifications than in the GV sessions (t-test, $N = 12$, $p < 0.05$), and the same significant results were obtained when we tested the last third of the game separately.

Finally, in contrast to our expectations, the difference in activity level between the two auction formats (GSP and VCG) was insignificant, with an only marginally higher level of activity on the side of the GSP bidders.

4. BID MODIFICATIONS

After observing this intensive bidding activity, and finding that, in general, active bidders did not do better or worse than less active bidders, it is interesting to take a closer look and to check how the bidders changed their bids.

We found good evidence that bidders did better reply to the others. Specifically, according to the state of the game as it appeared on their screen at the time they updated their bid (i.e., *planned* results), bidders responded to the others in a way that often improved their payoffs: averaged over all four conditions of the experiment, 46% of the bid modifications improved bidders’ payoffs, 31% reduced their payoffs, and the remaining 23% did not affect their payoffs. We did not find any indication that bidders replied in terms of fictitious play, which attempts to reply to the empirical distribution of others’ bids through the history of the game. In addition, we also tested the bid change impact according to the actual results the bidder received (i.e., with the *actual* updated bids of the others). Interestingly, here the picture was different and bidders were not as successful: only 35% of the bid modifications improved their payoffs compared to 36% that reduced them. That is, although it seems that players did plan to better-reply to the others, each time they modified their bid they had almost an equal chance to either increase or decrease their actual profits. This pattern was similar in all conditions (see Figure 9), and the proportions of beneficial bid modifications did not improve with time.

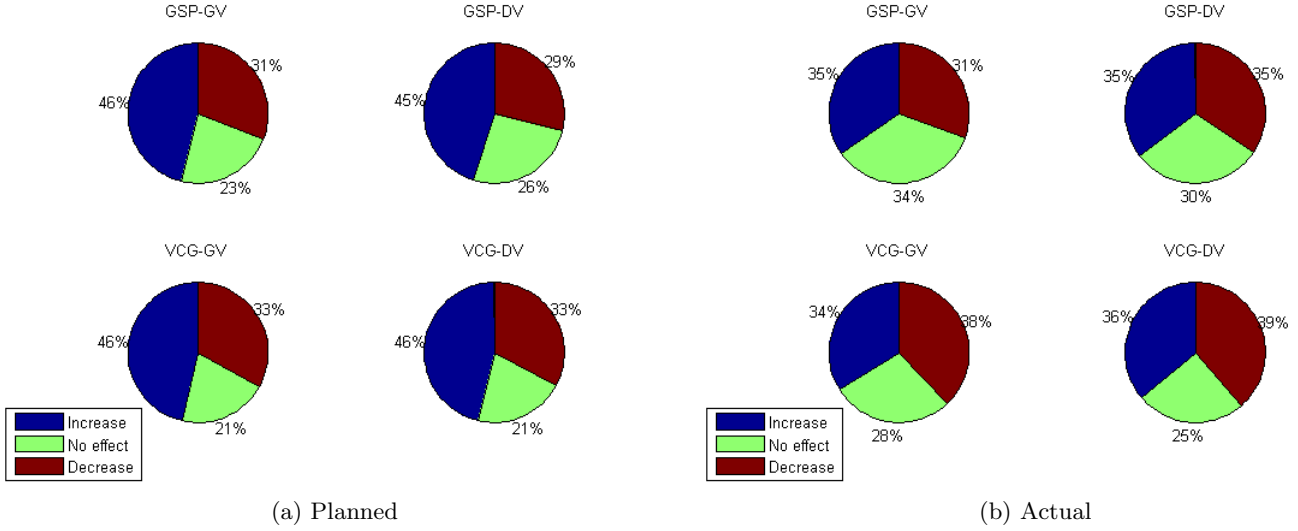


Figure 9: Impact of bid modifications on bidders' payoffs: The proportions of bid changes that increased, decreased, or did not affect the bidder's payoff, in each of the four conditions, and according to (a) planned gains (i.e., against last-seen bids of others) and (b) actual gains (i.e., against actual updated bids of others).

Is it possible that players did so poorly in predicting the immediate progress of the game as it is implied by the equal proportions of beneficial and harmful modifications? We further examined the average surplus bidders remained with as a result of their bid modifications. That is, this time we tested not only whether bid modifications improved or harmed bidders' payoffs, but how much they gained from these modifications. Figure 3 shows the marginal payoffs from bid modifications over time for planned and actual results, averaged over all conditions and player types. For both lines, it can be seen that the average marginal payoff increases rapidly and reaches its peak at the beginning, and thereafter decreases and seems to converge around the middle of the game. Except for a higher peak observed in DV conditions relative to GV conditions (that may be expected due to the initial uncertainty), the picture was similar across the four conditions and between the different player types.

We found that on average, bidders gained positive profits from their bid modifications throughout the game. In particular, according to either planned or actual results, the average marginal payoffs bidders gained from their bid modifications over all experiment sessions was found to be significantly positive in each third of the game (Wilcoxon two-sided signed rank test, $N=24$, $p<0.01$ in the first and second thirds, and $p<0.05$ in the last third of the game).⁷ That is, although the frequencies of improving and harmful bid modifications were similar in respect to the actual results, the better response to last-seen bids of the others also gave bidders positive gains when played against the actual bids of the others. Moreover, in terms of the average marginal

⁷We found both that the activity level of players did not affect their total payoffs, and that they still gain (on average) positive profits from these actions. These findings may be settled in various ways. For example, bidders may constantly gain profits from their own changes but lose them from the others' changes, or it may be that the effectiveness of bid changes is different between bidders of different activity levels. Which of the possible explanations is the most suitable here still has to be verified.

payoffs for the bidders, the planned and actual results were not found to be significantly different in any of the three thirds of the game. That is, the somewhat higher gains the actual results seem to give relative to the planned results in the second half of the game, as it appears in Figure 3, were not found to be significantly higher, although it may suggest the possibility that bidders learned and predicted the game dynamics better with time.

5. BID LEVELS

In this section we compare the patterns of the bids observed in our experiment to the theoretical predictions. Figure 4 shows the average bids per player type and condition throughout the game. Figure 10 shows the bid distributions per each player type and mechanism and their corresponding average payoffs. Next we will discuss patterns that arise from these plots.

As mentioned before, bidders in VCG are expected to bid truthfully since this is their dominant strategy. In GSP bidders are predicted to strategically bid lower, and in particular get closer to the lowest-LEF equilibrium bids. In line with the theoretical predictions, we found that bidders in GSP bid significantly lower than in VCG; specifically, the average bid in GSP sessions was 28.80 coins relative to 34.46 coins in VCG sessions (t-test, $N=12$, $p<0.001$). Also, when testing the average bid separately for each player type, GSP bidders bid significantly lower than VCG bidders (t-test, $N=12$, $p<0.05$ for all types except 27).

In VCG, for each of the player types, except for the lowest (21), the average bid was not found to be significantly different from their values; i.e., as theoretically predicted bidders stayed close to their values. However, the lowest type was found to bid significantly above his value (t-test, $N=12$, $p<0.01$). The same results were obtained also when we tested the first and last thirds of the game separately.

In GSP, each of the player types, except the lowest, shaded their bids (on average) toward the predicted equilibrium, but not enough to reach it (see Figure 4). The lowest type over-

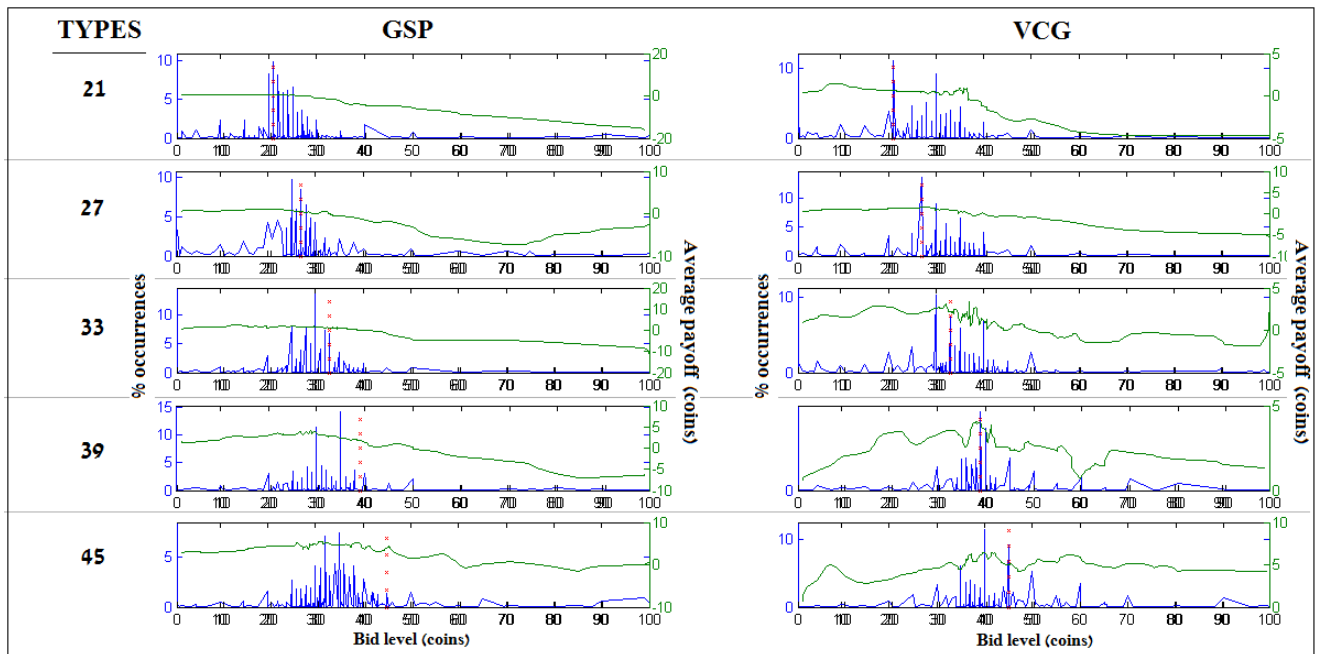


Figure 10: Bid distributions and payoffs: The relative frequency of bid levels as a function of auction mechanism and player type (the left y-axis, in blue), alongside their corresponding average payoffs for the bidder (the right y-axis, in green). For each player type, the truthful bid is marked by vertical red dots.

bids and seems to “block” the others from below. First, a comparison of the GSP bids with the lowest-LEF equilibrium bids shows that for each player type the average bid throughout the game was significantly higher than its corresponding equilibrium bid (t-test, $N=12$, $p<0.05$). Also, when testing the auctions in the last third of the game, bids were still significantly higher than the equilibrium bids (t-test, $N=12$, $p<0.05$ for all types except 39), indicating that bids did not converge to this equilibrium by the end of the game. Second, a comparison of the GSP bids with the bidders’ values shows that the three highest types significantly shaded their bids (t-test, $N=12$, $p<0.01$ for each of the types: 33, 39 and 45), but the average bids for the two lowest types (21 and 27) were not found to be significantly distant from their values.

Overbidding.

Figure 10 presents an intriguing pattern of data: players appear to overbid, namely, make bids above their personal values, quite frequently. Theoretically, overbidding is a weakly dominated strategy; that is, players can never gain better profits from overbidding, and hence it should never occur. Yet, overbidding was widespread: 38.8% of all bids were above the bidders’ values. Further, it was common in both value-information conditions, although as could be expected more in the DV than in the GV conditions (42.9% vs. 34.7%, respectively). In addition, overbidding in VCG was twice as frequent as in GSP (50.9% vs. 26.6%, respectively). We also found that lower-type players tended to overbid more than higher-type ones; specifically, there was a significant negative correlation between player types and the number of above-value bids, in both GSP conditions and in the VCG-DV condition ($N=30$, $p<0.001$). In the VCG-GV

condition this negative correlation was insignificant ($N=30$, $p=0.09$). Figure 5 shows the percentage of overbidding (of all bids) in each condition over time. It can be seen that overbidding remained a common behavior throughout the game, and had a different pattern of progress with time for each auction format: in the GSP conditions the frequency of overbidding decreased with time (consistent with the shading of bids described above) while in VCG it increased with time.

It seems clear that in spite of the theoretical non-profitability of overbidding, bidders do not hesitate to use this dominated strategy. Therefore we ask, why do they overbid? As can be seen in Figure 10, and consistent with the theoretical predictions, overbidding resulted on average in lower payoffs than the bidder could have obtained otherwise. Hence overbidding was not beneficial for them in the short term. We continued and tested whether it was somehow in the long term of the game beneficial for the bidders to overbid. However, as expected, we found the opposite picture: in GSP we found that the higher the player played (on median) the lower the payoff he got in total. Specifically, testing each type separately, we found a significant negative correlation between bid level (median) and the total payoff of a player ($N=12$, $p<0.05$ for each type except for type 39, for which the negative correlation was insignificant: $p=0.21$). Moreover, the negative correlation was found to be significant also when testing all types in GSP together on one scale, standardized by computing for each player the number of standard deviations from the mean achieved over players of his type and mechanism ($N=60$, $p<0.0001$). In VCG, there was an inconsistent, insignificant but still overall negative correlation between bidders’ long-term profits and their bidding levels ($N=60$, $p=0.12$). To conclude, it seems that the occurrence of overbidding could not be explained from

a rational perspective, and may be better understood using alternative approaches (e.g., bounded rationality, psychological motives, etc.; see, e.g., [6] and the references therein).

Truthfulness.

The mechanism design theory stresses the advantage of the VCG mechanism in which, as opposed to GSP, telling the truth is a dominant strategy. We already reported that we found a different distribution of bids for each mechanism, so that VCG bids were not significantly distant from their values, in contrast to the GSP in which bidders did shade their bids. Next we test whether the desired property of incentive compatibility of the VCG still holds with human players; that is, do players indeed play according to the theoretical firm prediction and bid *truthfully*?

We recall that the truthfulness property of the VCG mechanism was mentioned clearly in the instructions the players read before the game, while GSP players were told that their best bid may depend on other players' bids. By observing the initial bids in the GV condition of VCG and GSP, we can see that bidders did respond to the instructions: in VCG-GV 40% of the initial bids were truthful compared to 10% in GSP-GV. However, the total fraction of honest bidders decreased, so that only 19.5% of the bids in VCG-GV were the truthful dominant strategy, compared to 7.5% in GSP-GV. That is, although bidders seem to relatively "trust" the advice the instructions gave them, they end up preferring bids different than their value. Figure 10 shows that in VCG the truthful bid was among the bidders' most frequent bids (relative to bids in GSP, for which this was not the case), but it remained low in frequency for all types of players.

Again, as expected, we did not find any advantage in the long run of the game for players in VCG-GV who played the truthful bid less frequently. Specifically, the correlation between the frequency of truthful bids that a player played and his final payoff in the game among VCG-GV players was found to be positive. However, this result was not significant, probably due to the general low frequency of the use of truthful bids.

To conclude, it seems that although the VCG bids were closer to the bidders' values compared to GSP, the low fraction of truthful bids shows that human bidders still prefer to manipulate, and perhaps not to reveal, their true valuations. These results raise doubts regarding the truthfulness property of the VCG when played by human players.

6. BOTTOM LINE: REVENUE AND WELFARE

In this section we test how the behavioral bidding dynamics we have described so far are translated to the bottom line of the achieved social welfare and revenues for the auctioneer.

Social Welfare.

The Social Welfare achieved by an outcome of the auction is the sum of the incomes achieved by the bidders. Equivalently, it is the sum of the net utilities of the bidders and the revenue of the auctioneer. It is easy to see that the highest possible social welfare in an ad auction of the type studied here is achieved when the ad positions are allocated in the same order as the valuations, and the lowest possible social welfare is achieved by the exact opposite order. For our set-

tings of values and CTRs, the highest possible social welfare is 38.40 coins and the lowest possible is 27.60 coins. Every auction would achieve social welfare in this range. We measure the social welfare achieved by our auctions linearly on this scale, so that the lowest possible welfare is counted as 0 and the highest possible gets 100.

We found that both mechanisms resulted in similar high levels of social welfare. Specifically, GSP and VCG sessions achieved an average of 81.4% and 82.6% of the optimal social welfare on this scale, respectively ($N=12$, $\sigma = 5.1$ and $\sigma = 6.3$, respectively). The similarity of these distributions of the social welfare is in contrast to the advantage of VCG over GSP previously reported in [4]. Figure 6 shows how social welfare progresses with time. It shows that both mechanisms led to an increase in the social welfare throughout the game, with a faster increase at the beginning and slower thereafter.

Relatively high levels of social welfare were obtained in both value-information conditions, though a comparison reveals that the given-value condition achieved significantly higher social welfare than the deduced-value condition (84.3% vs. 79.8% of the optimal outcome in this scale, in the GV and DV conditions, respectively, t-test, $N=12$, $p<0.05$). The gap between value-information conditions was larger in the beginning of the game resulting from a particularly poor start achieved by the players who did not know their valuations (53% in the DV sessions, an almost random order, as expected, compared with 73% in the GV sessions). This gap decreased quickly with learning; however, the social welfare increased also for the GV players and stayed higher than for the DV players for most of the game (see Figure 6).

Revenues.

The revenue from each auction for the auctioneer (i.e., the search engine) is the sum of the bidders' payments. Theoretically, VCG players should implement the dominant strategies equilibrium and therefore produce its related revenue, which equals 24.3 coins per auction in our settings. As for the GSP, the VCG-like "lowest-LEF" equilibrium gives the auctioneer the same revenue as the VCG equilibrium. Therefore, it is predicted (firmly for VCG, possibly for GSP) that revenues from both mechanisms will converge to this outcome.

Our results were inconsistent with these predictions. First, both mechanisms produced higher revenues than expected throughout the game, but only the GSP sample was found to be significantly higher (t-test, $N=12$, $p<0.001$). The GSP revenues were significantly higher than the predicted outcome also when tested separately in each third of the game (t-test, $N=12$, $p<0.01$).

Second, we found that GSP produced significantly higher revenues than VCG did; specifically, the average revenue in GSP sessions was significantly higher than in VCG sessions (t-test, $N=12$, $p<0.02$, $\mu = 27.7$ and $\mu = 25.0$ coins, respectively). These findings are in line with the previous results reported in [4], and support the current implementation of ad auctions using the GSP mechanism. It seems that although the GSP bids were lower than VCG bids (see Section 5), this difference was not enough to balance out the difference in revenues resulting from the mechanism's payment rule. Figure 6 shows that the gap between revenue levels achieved under the two mechanisms was large in the

beginning and decreased with time. However, GSP revenues stayed higher than VCG revenues during most of the game.⁸

Finally, we report that value-information settings did not have a significant effect on revenue results; GV and DV sessions produced similar averages of 26.2 and 26.6 coins per auction for the auctioneer, respectively ($N=12$, $\sigma = 2.0$ and $\sigma = 3.6$, respectively).

Dividing the Pie.

To summarize the social welfare and revenues we have found, we shall check how the pie was divided between the advertisers and the search engine.

In our experiment the search engine captured a total of 76% of the social welfare in GSP sessions relative to 68% in VCG sessions. This is in contrast to 63% of the pie that the search engine would have captured if results were according to the theoretically predicted outcome. In a billion-dollar market this difference may be dramatic.

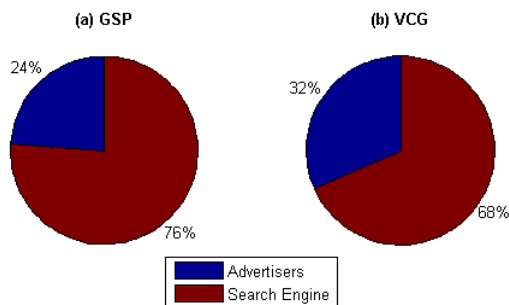


Figure 11: Share of resources in each mechanism.

7. CONCLUSION AND FUTURE RESEARCH

In this research we tested the behavior of participants in sponsored-search ad auction simulations and compared these results with the theoretical predictions.

Some of the results were in line with theory, most basically that bids in the VCG auction were, on average, close to the truthful values while in the GSP auction bidders did, as expected, shade their bids, and that both auction mechanisms achieved high levels of social welfare, indicating that, in general, allocation results were efficient.

Other results were compatible with theoretical reasoning to some degree, although they were not necessarily predicted by theory, e.g., the clear advantage that GSP had in terms of revenue, or the modest advantage bestowed by having explicit knowledge of the valuation.

On the other hand, bidders were also found to deviate from theoretical expectations in significant aspects. Notable deviations were the lack of convergence to equilibrium (with bid modification frequency actually increasing throughout the game), the low frequency of truthful bids in VCG auctions and fairly frequent overbidding in all auction settings.

⁸If we exclude the first 5 minutes as a learning phase, then the statistical significance of the revenue gap is a delicate issue. In a t-test with the $N=12$ auction averages of each format the gap is not significant ($p=0.11$). However, the gap is found to be significant if we are willing to assume independence over time (as was done in [4]); e.g., taking a t-test over $N=240$ minutes (last 20 minutes of each of the 12 sessions of each format) results in $p<0.001$.

In an attempt to explain these unexpected behaviors, we tested the possibility that they were beneficial to the players in the long run of the game, perhaps players using the sub-optimal strategies as signals. However, we did not find evidence for such long-term gains. For instance, players who overbid to a greater extent often ended up, on average, with lower total payoffs than those who made lower bids. Similarly, we did not find any long-term advantage for bidders who tended to modify their bids often. These results weaken attempts to explain the unexpected behaviors as some form of long-term “rational” signaling or collaboration strategies in the repeated game.

We suggest that human players’ behavior in an ad auctions environment should be analyzed with respect to other motives that they may have, according to (behavioral) heuristics they may be using, and considering cognitive constraints and biases. For instance, the players’ reports in the final questionnaire suggest that they strongly believed that their profits depended heavily on obtaining a particular position for their ad. Players also found the higher positions to be more attractive. While these may be generally useful heuristics, they could lead to sub-optimal results since obtaining higher ad positions need not be more profitable. The evidence for participants’ tendency to aim for higher positions arises also from the findings that lower player types were more active in changing their bids and tended to overbid more often than the higher player types.

We believe that much more about human bidding behavior in ad auctions can be explained and leave it as a direction for further research.

Acknowledgments

We thank Jennifer Klein for helping with running the experiments and Ely Levy for helping with the experimental infrastructure. We also thank Ilan Nehama and Effi Levi for comments on an early draft. Finally, special thanks to Yoav Kolumbus for his helpful advice and comments throughout this research.

8. REFERENCES

- [1] M. Cary, A. Das, B. Edelman, I. Giotis, K. Heimerl, A. R. Karlin, C. Mathieu, and M. Schwarz. Greedy bidding strategies for keyword auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, EC ’07, pages 262–271, New York, NY, USA, 2007. ACM.
- [2] Y.-K. Che, S. Choi, and J. Kim. An experimental study of sponsored-search auctions. *Working paper*, 2011.
- [3] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.
- [4] E. Fukuda, Y. Kamijo, A. Takeuchi, M. Masui, and Y. Funaki. Theoretical and experimental investigation of performance of keyword auction mechanisms. *Sixth Ad Auctions Workshop, 2010, Harvard University, Cambridge, Massachusetts*, 2010.
- [5] J. H. Kagel, R. M. Harstad, and D. Levin. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica*, 55(6):1275–1304, 1987.

- [6] J. H. Kagel and D. Levin. *The Handbook of Experimental Economics, Volume II*, chapter Auctions: A Survey of Experimental Research, 1995 - 2008. Princeton: Princeton University Press, forthcoming.
- [7] J. H. Kagel and A. E. Roth. *The Handbook of Experimental Economics*. Princeton University Press, Princeton, N.J., 1995.
- [8] S. Lahaie, D. M. Pennock, A. Saberi, and R. V. Vohra. Sponsored search auctions. *Algorithmic Game Theory*, pages 699–716, 2007.
- [9] N. Nisan, M. Schapira, G. Valiant, and A. Zohar. Best-response auctions. In *Proceedings of the 12th ACM Conference on Electronic Commerce, EC '11*, pages 351–360, New York, NY, USA, 2011. ACM.
- [10] B. M. Staw. The escalation of commitment to a course of action. *The Academy of Management Review*, 6(4):577–587, 1981.
- [11] H. R. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, 2007.