STATISTICAL METHODS FOR COMPUTATIONAL ADVERTISING

David Banks (Duke University),
Nancy Heckman (University of British Columbia),
Nancy Reid (University of Toronto)

October 3, 2021–October 8, 2021

1 Overview of Computational Advertising

Computational advertising is a young, fast-moving field. The online advertising market is projected to be valued at $982 billion by 2025. It is the dominant revenue stream for many major IT companies.

The ecology of ad buy and auctions is complex, and companies with better information and better statistics will make more profit. At a high level, when a user types in a keyword, such as ”pizza”, to the browser, it triggers an auction among companies that think that keyword is an opportunity to encourage a purchase. The highest ”qualified” bidder gets the most prominent ad position on the browser, the second highest bidder gets the next best placement, and so forth. Bidder qualification is a secret weighting factor that favors prominent, established companies over shady or unreliable businesses.

The auction is complicated, but the basic steps are as follows:

1. A person’s browser contacts the publisher’s website (e.g., CNN.com).

2. The publisher’s website sends back content including placements that will need to be fulfilled though an AdServer.

3. The browser contacts the publisher’s AdServer to fulfill placements that will not go up for auction. If the user has an in-app or a browser ad blocker this interchange will not happen.

4. The publisher’s AdServer sends back predefined ad content.

5. For other placements the browser will contact an Exchange with placement information and an indication on whether the placement should go out for bid, or if another private (guaranteed) deal has been setup for the placement.

6. If the placement is marked for real-time bidding (RTB), an auction is set up by the Exchange.

7. Once the auction is initiated the demand side platforms (DSPs) are simultaneously contacted to participate in the auction. This all runs in parallel.

8. If a DSP decides to bid its bid offer and a director tag (for tracking) are returned to the auction.

9. The winning bid from the auction and information on the bidder, the DSP, is sent to the Exchange

10. The Exchange returns the DSP wrapped ad tag that contains everything needed to track the bid, impression, and director log joins. It's essentially a link to the director so one can obtain user information and pass back the actual creative ad tag.

11. The browser contacts the DSP's director to obtain the AdServer tag, which is a link to the advertising agency for (one of) the company's that won the bid.

12. The browser receives the AdServer tag.

13. The browser then uses the AdServer tag to contact the advertiser's AdServer requesting the impression.

14. The advertiser's AdServer returns all of the impression assets and creative content back to the browser.

15. If the campaign/line item is using a third party for tracking it is contacted with the impression and user information. This service is now almost always used.

16. The third party tracker sends back a 1×1 pixel as a verification or handshake token.

17. The ad tag (sent in #10) contains the javascript that the browser/app will load when the impression renders. These signals are passed back ONLY AFTER the creative content is loaded which is why these actions/behaviors are so far down the chain.

18. The DSP sends back a 1×1 pixel as a verification/handshake.

19. If the DSPs are tracking pixel fires this information is sent from the browser.

The auction must be completed within ten milleseconds. But this is still a simplification. The field is changing rapidly—it used to be that nearly all auctions were second-price auctions, but now they are nearly all first-price auctions.

Computational advertising touches on many aspects of statistics. When a company develops a new ad, process control methodology monitors its clickthrough rate, to discover when an ad has become stale (which corresponds to drift) or when a competitor has developed a superior ad (which corresponds to shift). But unlike classical process control literature, these processes are adversarial—producing an effective ad incentivizes the competition to create a better one.

Experimental design is also important. Google runs several thousand designed experiments a year, and famously compared 47 shades of blue for their hyperlinks. Netflix recently announced that it wanted to hire an expert in response surface methodology. But novel problems arise: one can be exposed to one treatment level on one's laptop, but a different level on one's cellphone. Such contamination has not been previously considered in the classical experimental design literature.

A third area of statistical research is recommender systems. These are the engine of computational advertising, since they determine which ads to show and which products to suggest. There are two strategies, collaborative filtering and content-based filtering, and both require sophisticated statistical models that approximately factor a sparse ratings matrix into two low-rank matrices.

An emerging area is active recommender systems. In these, the system interacts with the user, asking questions about the user's taste in, say, books. To learn efficiently, the system would ask questions for which it has prior probability 0.5 that the user will answer yes or no. But such questions require a complexity penalty: "Do you prefer this list of 100 books to this other list of 120 books?" is not a question that humans can cognitively process. There are additional difficulties—the search tree of questions that are asked must contain many branches for which the prior probability is near 0.5 and few that are near 0.1, which is a novel problem in optimization.

There are many statisticians working on computational advertising in industry, but few in academics. The industry statisticians do not have time to develop principled solutions and general theory—they put out fires. And their employers do not encourage them to publish, since it can reduce the competitive edge. So this is a ripe field for academic research, and one that is important to the economy of the future.

2 Outcome of the Meeting

The meeting had 27 presentations and 52 virtual participants. One result of that meeting was an Introductory Overview session given at the 2022 Joint Statistical Meetings in Washington DC by David Banks and Nathaniel Stevens.

A second output is a paper on recommender systems that is under review by the *Journal of the American Statistical Association*; it is an invited revision, which means publication is likely. The authors are Patrick LeBlanc, David Banks, Linhui Fu, Ryan Tang, and Qiuyi Wu, three of whom were part of the BIRS meeting.

A third result is that computational advertising figured prominently in the 2023 Deming Lecture. This named lecture recognizes the pioneering work of Ed Deming in quality management, and his ideas for manufacturing often carry forward to information technology industries. The lecture was given by David Banks.

Additionally, David Banks and Nathaniel Stevens have given lectures on aspects of both computational advertising and agent-based models, a related tool, at a number of conferences and university seminars.