The Complex Dynamics of Sponsored Search Markets[1]

Valentin Robu[ab] Han La Poutré [ac] Sander Bohte [a]

[a]*CWI, Science Park 123, 1098XG Amsterdam, NL*
[b]*University of Southampton, SO17 1BJ Southampton, UK*
[c]*TU/e, De Lismortel 2, Postbus 513, 5600 MB Eindhoven, NL*

Sponsored search, the payment by advertisers for clicks on text-only ads displayed alongside search engine results, has become an important part of the Web. It represents the main source of revenue for large search engines, such as Google; and Microsoft's Live.com, and sponsored search is receiving a rapidly increasing share of advertising budgets worldwide. At the same time, sponsored search also present exciting research opportunities, for fields as diverse as economics, artificial intelligence and multi-agent systems.

Here, based on large-scale Microsoft sponsored search data, we provide a detailed empirical analysis of such data. We carry out this empirical analysis as most existing work on the dynamics of electronic markets (e.g. in agent-based computational economics (ACE)) has been based on simulations, as there are few sources of large-scale, empirical data from real-world automated markets. In this context, empirical data made available from sponsored search provides an excellent opportunity to test the assumptions made in such models in a real market. To do this, we deploy several techniques derived from computational economics, and especially complex systems theory. Complex systems analysis has been shown to be an excellent tool for analyzing large social, technological and economic systems, including web systems [4, 3, 1].

The study provided in this paper is based on a large dataset of sponsored search queries, obtained from the website Live.com through a Microsoft Beyond Search grant. The search data provided consists of two distinct data sets: a set of sponsored search dataset (URLs returned are allocated to advertisers, through an auction mechanism) and an organic search dataset (standard, unbiased web search). The sponsored search data consists of 101,171,081 distinct impressions (i.e. single displays of advertiser links, corresponding to one web query), which in total received 7,822,292 clicks. This sponsored dataset was collected for a roughly 3-month period in the autumn of 2007. The organic search data set consists of 12,251,068 queries, and was collected in a different 3-month interval in 2006 (therefore the two data sets are chronologically disjoint).

It is important to stress that in the results reported in this paper are based mostly on the sponsored search data set. Furthermore, the sponsored search data we had available only provides partial information, in order to protect the privacy of Microsoft Live.com customers and business partners. For example, we have no information about financial issues, such the prices of different keywords, how much different advertisers bid for these keywords, the budgets they allocate etc. Furthermore, while the database provides an anonymized identifier for each user performing a query, we cannot trace individual users for any length of time. Nevertheless, one can extract a great deal of useful information from the data. For example, the identities of the bidders; for which keyword combinations their ads were shown (i.e. the impressions); for which of these combinations they received a click; the position their sponsored link was in when clicked etc...

In our analysis, we first study how the display rank of a URL link influences its click frequency, for both sponsored search and organic search. Second, we study the market structure that emerges from these queries, especially the market share distribution of different advertisers. We show that the sponsored search market is highly concentrated, with less than 5% of all advertisers receiving over 2/3 of the clicks in the market. Furthermore, we show that both the number of ad impressions and the number of clicks follow power law distributions of approximately the same coefficient. However, we find this result does not hold when studying the same distribution of clicks per rank position, which shows considerable variance, most

[1]This work been presented in the 2009 AAMAS Workshop on Agents and Data Mining Interaction (ADMI'09), and will appear in LNCS/LNAI post-proceedings. A pre-print is available at http://homepages.cwi.nl/ sbohte/publication/admi09.pdf. This work was performed based on a Microft Research "Beyond Search" award. The authors wish to thank Microsoft Research for their support.

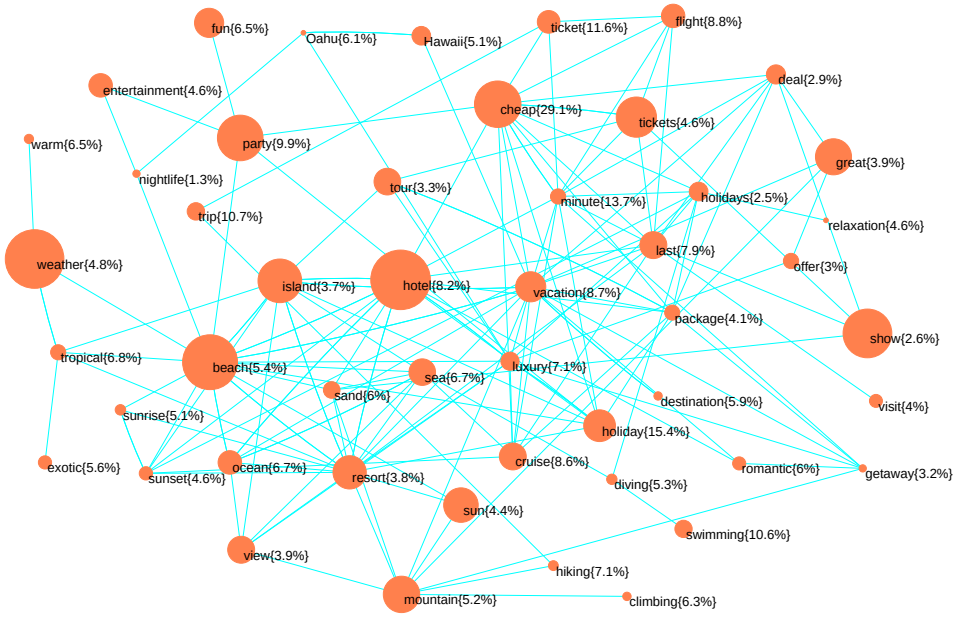

Figure 1: Visualization of a search term correlation graph, for a set of search terms related to the tourism industry. Each search term is assigned one colored dot. The size of the dots gives its relative weight (in total number of clicks received), while the distances between the dots are obtained through a spring-embedder type algorithm and are proportional to the co-occurrence of the two search terms in a query. Each dot is marked with its success rate (percentage of the total number of impressions associated with that query word that received a click).

likely due to the way advertisers divide their budget on different keywords. Finally, we turn our attention to how such sponsored search data could be used to provide decision support tools for bidding for combinations of keywords. We provide a method to visualize keywords of interest in graphical form, as well as a method to partition these graphs to obtain desirable subsets of search terms based on modularity [4].

There exists previous work that has applied similar co-occurence-based techniques to organic search logs or tagging systems [2, 3]. However, our focus in this paper is different: we do not aim to to merely deduce what is the semantic distance between keywords in the general sense, but what kind of combinations of keywords are financially interesting for a sponsored search advertiser to bid on. This is the reason why the size of the nodes and distances computed in Fig. 1 are built using only queries which lead to an actual click on a sponsored ad. Basically, this is equivalent to filtering only the “opinion” (expressed through queries) of the subset of users that are likely to buy something online, rather than all search engine users. To our knowledge, this is the first paper to use sponsored search click data in this way.

References

[1] T. Carter. A short trip through entropy to power laws, 2007. Complex Systems Summer School, Santa Fe Institute, NM.

[2] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, 2007.

[3] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proc. 16th Int. World Wide Web Conf. (WWW’07)*, pages 211–220. ACM, 2007.

[4] M. Newman. Power laws, pareto distributions and zipf’s law. *Contemporary Physics*, 46:323–351, 2005.