UNIVERSITY OF PADOVA - DEPARTMENT OF INFORMATION ENGINEERING

COMPUTER ENGINEERING

THESIS

A WEB SEARCH MARKET MODEL

Supervisor: **Ch.mo Prof. Enoch Peserico Stecchini Negri De Salvi**

Student: **Andrea Mambrini**

Academic Year 2010-2011

24-10-2011

Abstract

In this work we analyze a web search market where there is real competition between search engines (which also are advertising agencies). We will show that in this kind of market the revenue of the search engines depends on the precision of the predictions of users' preferences and that advertisers and search engines' users can gain from the competition between search engines.

Contents

Introduction

In recent years a new discipline named *computational advertising* emerged [12, 13]. The aim of this discipline is to merge the knowledge of different topics like microeconomics, information retrieval, auction theory and machine learning to find the best match between a given user in a given context and a suitable advertisement.

This discipline is very important over the Internet to find optimal strategies for *contextual advertising*. This kind of advertising tries to infer the user preferences in order to tailor the advertising to the user. The idea is that if the user is interested in the product advertised it is more likely that he will pay attention to the advertising.

We can divide contextual advertising into two categories: the ones that appear on normal websites and the ones that are shown on the result pages of search engines. The latter are usually called *sponsored search*. The main search engines (i.e. Google, Yahoo, Bing) give two sets of results to a user query: organic results and sponsored results. Organic results are given by the search engines according to some ranking algorithms (e.g. PageRank [10]) that ranks websites according to their relevance to the keyword searched by the user. Sponsored results take also into consideration how much the website is paying the search engine in order to be shown among the sponsored results.

In this thesis we will analyze a market where there are many search engines offering sponsored search and where the user and the websites can chose which one to use in order to maximize their revenue. We will see that the revenue of the search engine will be less easy to obtain and it will be related to its ability to predict user preferences. The model also allow users and websites to earn money from sponsored search.

In Chapter 1 we will give an overview of the search engine market. In Chapter 2 we will define the model describing the entities involved, how they interact and how they measure their revenue. In Chapter 3 we will see how different level of accuracy in user preference prediction leads to different revenue for the search engines. Finally in Chapter 4 we will summarize the results and we will analyze limit and strengths of the model.

Chapter 1

The sponsored search Market

In this chapter we describe the situation of the web search market, the entities involved and the economical relevance of this market. We will see that few forces dominate the market and that the competition is not enough to avoid abuse of market power.

1.1 Organic search

The main function of a web search engine is to provide links as result to a query based on what it is available on the web. Most of the results a search engine provides are *organic search results*. Those results should be distinguished from paid search results (Fig. 1.1) that we will discuss in the following sections.

In order to provide organic results search engines regularly explore the web and rank its content. The ways the exploration and the ranking are exactly done are confidential business secrets, anyway some basic principles are usually followed:

- **Crawling and indexing**. Search engines crawl the web and create an index of web site content [3]. When a query is submitted, the search engine try to match the keyword used in the query to the index in order to provide relevant results.

- **Reputation**. A measure of the reputation of a website is really important in order to rank websites providing similar content. In this way the search engine can provide to the user the best results available. Usually reputation is commonly measured by the number of links on other web sites that point to it. In particular, links from popular web sites improve the ranking of a specific web site [10].

- **Past user behaviour**. Information about previous search queries is used to improve search results [2].
- **Personalized search**. At each query search engines receive additional information about the user (IP, browser type, language, cookies). This information is used to tailor search results to the user and improve the quality of results [7]. For example the IP address can be used to narrow down the results according to the location of the user.

1.2 Web search and advertising

The aim of search engines is to make the information available on the Internet easily available to the users. That is usually provided free of charge for the users so, in order to turn this business profitable the search engine had to start to sell users' attention to advertisers. In fact by submitting a query to the search engine, a user provides information about his current interests to the search engine. Those information are really useful for advertisers that can adapt their advertises to user's interests. From providing those information to advertisers search engines started to be profitable.

Among with normal search results, search engines provided something named "sponsored links" (see Figure 1.1). The difference between normal results and sponsored links is that the website has to pay the search engine for appearing among sponsored links when the users look for a specific keyword. The mechanism to sell those slots is called slot auction. The benefit for advertisers is that they have the possibility to advertise websites that can be interesting for the users. In this way the advertising is more effective. In Figure 1.1 we see that the user is looking for an used car and that the proposed advertising is about used cars vendors.

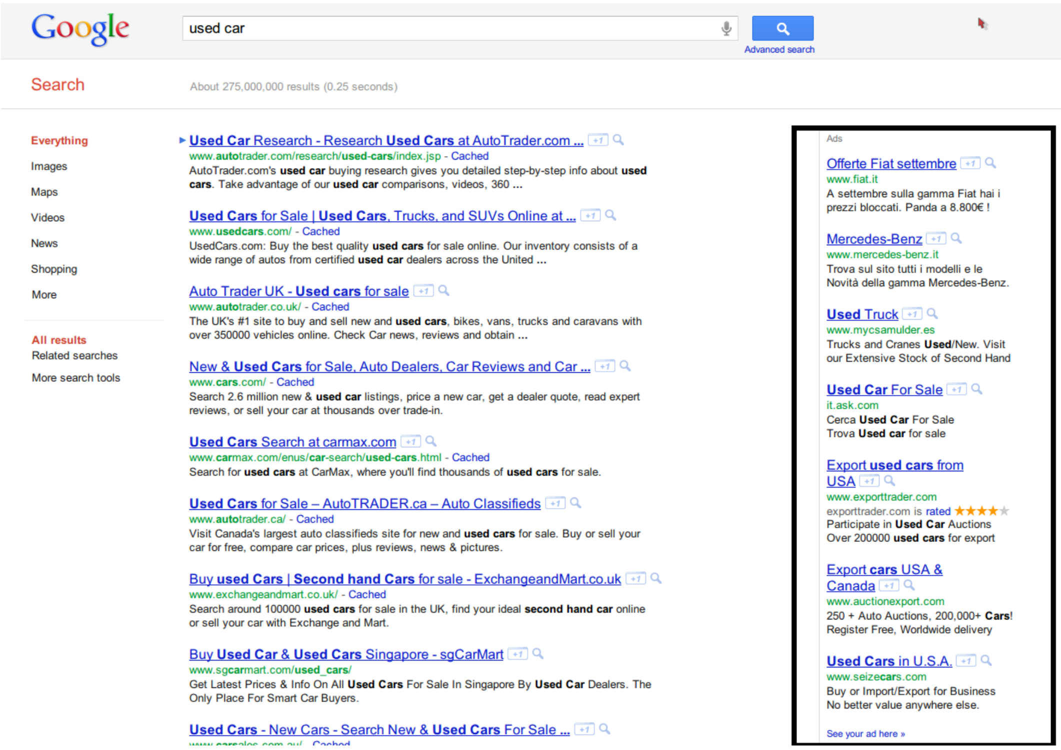

Figure 1.1: Sponsored links on the right, organic results on the left

1.2.1 Slot auction

Advertising slots are sold in auctions where advertisers submit bids for keyword combinations. Each such bid indicates the willingness of an advertiser to pay for every time that users click on an ad shown as a result of a query for a specific keyword or a combination of keywords. Advertisers pay for each click on the sponsored link and this gain goes to the search engine.

Many mechanisms for auction have been developed. The scheme most of the search engines uses is Generalized Second-Price (GSP)[4]. In this scheme advertiser who submits the highest bid wins the best slot but pays only the second-highest bid. In the same way, the second highest bidder wins the second-best slot and pays the third-highest bid price, etc.

Another auction scheme is Vickrey–Clarke–Grove (VCG)[9]. This scheme is more complicated but has an important property GSP doesn't have: it is truth-telling [5]. This means that bidding the real value is a dominant strategy for advertisers. This benefits the advertisers since it no longer encourages them to invest into bidding robots to game the system. The con for search engines is that if advertisers bid the same amounts under GSP mechanism and under VGC mechanism than the search engine revenue will be higher in the former case rather than in the latter [4].

1.3 Competition

1.3.1 Market share

The web search market can be separated into different countries . Table 1.2 shows market shares of the biggest providers in web search in selected countries. All markets are highly concentrated. With the exception of Russia, the largest provider in each market has secured a share of between 60 and 90 percent. Google is the market leader in most countries including the US, the UK, France and Germany.

USA	UK	France	Germany	China	Russia
Google (65.0%)	Google (91.3%)	Google (89.8%)	Google (86.6%)	Baidu (60.9%)	Yandex (47.4%)
Yahoo! (20.1%)	Yahoo! (2.8%)	MSN+Live (2.9%)	MSN+Live (6.0%)	Google (27.0%)	Google (31.2%)
MSN+Live (8.0%)	Ask (1.7%)	Yahoo! (2.5%)	Yahoo! (2.7%)	Sogou (3.1%)	Rambler (9.7%)
Ask (3.9%)	Live(0.9%)	AOL (1.7%)	T-Online (1.8%)	Yahoo (2.4%)	Mail.ru (7.0%)

Figure 1.2: Market share by search queries. Sources: USA (ComScore, 2009), UK (Hitwise UK, 2009), France (At Internet Institute, 2009), Germany (Webhits, 2009), China (China IntelliConsulting, 2008), and Russia (ComScore, 2008b). [11]

1.3.2 Reasons of market concentration

Market concentration is mainly driven by three features of the web search market:

Cost structure of a search engine business

The cost structure in the search engine market is characterized by high fixed costs for Research & Development as well as hardware and software infrastructure and almost zero variable cost for providing an additional query or placing an additional advertisement. Industry experts estimate the minimum value to start a search engine business and paying off the fixed costs as 2 billion USD [11]. Even with such an amount of resources the competition will be hard with the market leaders. In 2010 Google spent 3.7 bilion USD for R&D (7.8 % of its revenue) [6], Yahoo did 1.08 bilion USD (17 of its revenue)[6].

Prevalence of network effects

In economic theory, a positive network effect describes a situation where the value of a good or service for individual consumers increase with the total number of consumers. In this case we can say that the quality of the service provided by a search engine increase with the number of users. As we have described in Section 1.1 search engines use past queries to increase the quality of results. So users see an increasing of quality of the service with the increasing of the number of users. The same happens for advertisers: increasing the number of users increases the probability of finding a good match with a possible costumer so it increases the quality of the service the search engine provides to advertisers. The presence of positive network effects for both users and advertisers creates a positive feedback loop in which a search engine becomes more valuable to users and advertisers as more users join.

Inertia of users to switch to another search engine

Typically, users cannot fully assess the quality of search results and use search engines without knowing how its search algorithm works. Instead, they mainly trust a search engine's choice and believe in its quality. Web search can therefore be considered a *credence good*[1], since users need to develop trust based on the reputation of a search engine as well as on their previous experiences. For users, switching to another search engine means losing the brand reputation of their previous provider. Since Google and Yahoo! are listed among the most valuable global brand names[8] it can be assumed that the costs of switching away from one of the major web search engines are very high.

1.3.3 Bargaining Power

The market of sponsored search is economically really relevant. Google's revenue from sponsored search is 2010 was 19.5 bilion USD[6], in the same Yahoo's revenue from sponsored search has been of 3.2 bilion USD[14]. This revenue come from GSP mechanism. As we stated in Section 1.2.1, this auction mechanism maximize search engine revenues but it is harder to manage for advertisers. The fact that the main search engines (i.e. Google and Yahoo) use GSP instead of VCG shows that the strong concentration of forces in very few number of search engines leads to an asymmetry in the bargaining power of search engine, users and advertisers that put the search engines in a very strong position. This can also leads to a potential risk of abuse of market power [11].

In the next chapter we will build a model of the web search market where many entities offer sponsored search and we will see how the distribution of bargaining power would be more equilibrated.

[1] products whose quality is difficult to assess even after they are consumed. Other examples of credence goods are services provided by doctors, car repair, or legal consulting.

Chapter 2

Model

In this section we will introduce a model for the web search market. We can divide the set of actors in the market in 3 subsets: users, search engines and websites. Each entity has its own utility and acts in the market to maximize it. We will analyze how this market behaves and what are the equilibria.

2.1 Entities interaction

There are 3 sets of actors: a set U of users, a set E of search engines, and a set W of websites. We denote by $u \in U$, $e \in E$ and $w \in W$ respectively a user, a search engine and a website.

A user u can perform a search query $q_u = (e_u, k_u)$ to a search engine e_u using a keyword k_u . The search engine e answers the user with a set V_u of websites. Each search engine e has its own algorithm to decide which websites include in V_u . We assume that $|V_u| = T$ is fixed for the system, which means that each search engine visualizes the same number of websites as an answer to a query.

A website can ask a search engine to be indexed (i.e. to have the possibility to be visualized by that search engine as a result to a search query).

Each user u assigns a value ν_{uw}^{k} to a website w (i.e., ν_{uw}^{k} is high if the the user u is happy to visualize the website w as a result for a query when looking for keyword k). In the same way each website w assigns a value ω_{wu}^{k} to a visualization received by the user u (i.e. ω_{wu}^{k} is high if the website w is happy to be visualized as a result from a query from the user u when it looks for the keyword k).

Each website w decides whether it wants to have the possibility to be shown as result of a query for a keyword k by the engine e . If it does, it makes a bid to the search engine b_{ew}^{k} for the keyword k. This will affect the chances to get visualized as we will show in Section 2.3. Each time the website w is visualized to a user u by the engine e for the keyword k it pays (or get payed if the value is negative) ψ_{euw}^{k}. This values is decided by the search engine as we will show in section 2.3

The user decides which search engine to use. The user u pays (or get payed if the value is negative) μ_{euw}^{k} for each websites he/she see as result for a query to the search engine e for the keyword k. So it actually pays $\sum_{w\in V_u}\mu_{euw}^{k}$ for each query. The value of μ_{euw}^{k} is decided by the search engine as we will show in section 2.3.

We call a round an interval of time in which each user select a search engine, does a query on it and get an answer from it. Given the set U of users we get, for each round the following sets

- $Q = \left\{q_1, \cdots, q_{|U|}\right\}$ the set of all the queries $q_u = (e_u, k_u)$
- $V = \left\{V_1, \cdots, V_{|U|}\right\}$ the set of all the answers to each query $V_u(q_u) = \{w_1, \cdots, w_T\}$

We have that $|U| = |Q| = |V|$

We also introduce two combined parameters:

- $\beta_{uew}^{k} = \nu_{uw}^{k} + b_{ew}^{k}$ is the *bid-value*
- $\gamma_{uw}^{k} = \nu_{uw}^{k} + \omega_{wu}^{k}$ is the *match-value*

2.1.1 A note about the users

The model is fine-grained enough to represents all the users of the search engines as independent entities. On the other hand search engine will hardly consider each user as different entities. What search engines usually do is consider a set of user as one "kind of user". That means that the search engine can, for example set the group of users that study computer science as the "computer science student" kind of user. So it will deal with all the users in this set as they are the same $u^* \in U$.

So, in the following sections we will indifferently refer to a $u \in U$ either as a user or a kind of user.

2.2 Gain

We assume that at each round all the users $u \in U$ make a query $q_u = (e_u, k_u)$ to a search engine e_u for the keyword k_u . For each entity (users, websites and search engines) we define a gain on the round. The gain on the round is high if the entity is happy after that round. We define respectively g_u, g_w and g_e the gain for users, websites and search engines for each round.

- $g_u = \sum_{w \in V_u} (\nu^k_{uw} - \mu^k_{euw})$ where w are the websites visualized by the engine e as an answer for the query for the keyword k done by the user u
- $g_w = \sum_{u: w \in V_u} (\omega^k_{wu} - \psi^k_{euw})$ where u is in the set of users that have visualized the website w as result for a query for the keyword k , and e is the engine that have been used
- $g_e = \sum_{u: q_u = (e, k_u)} \sum_{w \in V_u} (\mu^k_{euw} + \psi^k_{euw})$ where u is in the set of users that have visualized the website w as result for a query for the keyword k using the engine e

2.2.1 Surplus value

We define a value for the system that represents the sum of the gains of all the entities in a round. We call it *system surplus value* and we indicate it with S. Let's calculate it

$$S = \sum_{u \in U} g_u + \sum_{w \in W} g_w + \sum_{e \in W} g_e$$

$$\sum_{u \in U} g_u = \sum_{u \in U} \sum_{w \in V_u} (\nu^k_{uw} - \mu^k_{euw})$$

$$\sum_{w \in W} g_w = \sum_{w \in W} \sum_{u : w \in V_u} (\omega^k_{wu} - \psi^k_{ew}) = \sum_{u \in U} \sum_{w \in V_u} (\omega^k_{wu} - \psi^k_{ew})$$

$$\sum_{e \in E} g_e = \sum_{e \in E} \sum_{u : q_u = (e, k_u)} \sum_{w \in V_u} (\mu^k_{eu} + \psi^k_{ew}) = \sum_{u \in U} \sum_{w \in V_u} (\mu^k_{eu} + \psi^k_{ew})$$

So

$$S = \sum_{u \in U} g_u + \sum_{w \in W} g_w + \sum_{e \in W} g_e = \sum_{u \in U} \sum_{w \in V_u} \nu^k_{uw} + \omega^k_{wu} = \sum_{u \in U} \sum_{w \in V_u} \gamma^k_{uw}$$

That means that the game is not zero-sum. The surplus value for each round is the sum of the matching-value $\gamma^k_{uw} = \nu^k_{uw} + \omega^k_{wu}$ for each matching (u, w) between users and websites that is created by the system (all the search engines).

Notice that this result does not depend on the way the engines decide the payments. That is the surplus value does not depend on the values of μ^k_{euw} and ψ^k_{euw}. We can than state that the Surplus value is a function of the matching between users and websites and that it actually measures the quality of this matching.

It is also interesting to define the *engine surplus value* that we indicate as S_e as

$$S_e = g_e + \sum_{u : q_u = (e, k_u)} g_u + \sum_{w : (w \in V_u), q_u = (e, k_u)} g_w$$

That is the sum of the gain of the search engine e and of the other entities from using the search engine e. We also have that S_e is the sum of the matching-values of the websites visualized to the users making queries to that search engine e, that is

$$S_e = \sum_{u:q_u=(e,k_u),w\in V_u} \gamma^k_{uw}$$

and that the sum of all the engine surplus values is the system surplus value

$$S = \sum_{e\in E} S_e$$

Of course we have that if there is just one search engine the engine surplus value and the system surplus value are the same. In this case we indicate it as just *surplus value*.

2.3 Slot auction

2.3.1 Motivation

We said that each search engine shows T results as an answer to a query for a specific keyword. It usually happens that more than T websites want to be shown as a result for a specific keyword. So the search engine needs a way to select T of them to be shown. We call a place in the result page a *slot* (so there are T slots for each query) and the mechanism to assign the slots a *slot auction*.

At each round we want to maximize the system surplus value. In this way there is more gain to distribute among the entities. In order to do that each search engines have to select the T websites that maximize the engine surplus value. We have shown in Section 2.2.1 that the engine surplus value is equal to the sum of the match values of each shown website. So the search engines have to show the best T websites according to $\gamma^k_{uw} = \nu^k_{uw} + \omega^k_{wu}$. The problem is that the search engine does not know this value. It can have an estimation of ν^k_{uw} but not about ω^k_{wu}. The former is in fact what actually search engines use in the real world to rank the websites (i.e. Google's pagerank) while the latter depends on the business structure of the website and it is related to the return on investment of the visualization of the website in a result page. So the search engine need a strategy to know ω_{wu}^k.

2.3.2 Slot auction mechanism

We will now show an auction mechanism that maximize the engine surplus value and make the search engine to know the exact value of ω_{wu}^k. When a user u makes a query for the keyword k to the search engine e the slot auction happens as following:

1. The search engine knows each website bid b_{ew}^k and an estimation of the user value ν_{uw}^k for the websites.
2. The search engine order all the websites in descending order according to the bid-vales $\beta_{uew}^k = \nu_{uw}^k + b_{ew}^k$.
3. It selects the first T websites according to this rank and put them in the results set V_u. We call $\overline{\beta}_{ue}^k$ the bid-value of the first websites not selected (i.e. the websites in position $T+1$ in the rank)
4. The search engine asks each website that is in the result set V_u to pay ψ_{euw}^k
5. The search engine asks the user to pay μ_{euw}^k for each visualized website.

2.3.3 The website payment

We have to define the website payment (ψ_{euw}^k). We set it to

$$\psi_{euw}^k = b_{ew}^k - (\beta_{uew}^k - \overline{\beta}_{ue}^k)$$

and we show that in this way we get what we were looking for in Section 2.3.1.

Lemma 2.3.1. *This auction mechanism is truth-telling. That is bidding* $b_{ew}^k = \omega_{wu}^k$ *is a dominant strategy for a website.*

Proof. We want to prove that a dominant strategy for a website is to bid its real value for the slot. That is bidding $b_{ew}^k = \omega_{wu}^k$ is a dominant strategy. Let's assume that all the websites made their bid for a keyword k to the search engine e, the search engine made the ranking and that $\overline{\beta}^k_{ue}$ is the bid-value in position $T+1$ in the ranking. For a website w we can have the following situations:

- $\nu^k_{uw} + \omega^k_{wu} < \overline{\beta}^k_{ue}$: in this case bidding $b^k_{ew} \leq \omega^k_{wu}$ the website is not in the best T bidders so it is not visualized and its gain is 0. If it bids $b^k_{ew} > \omega^k_{wu}$ the gain is either 0 (if it is not in the best T bidders) or

$$\omega^k_{wu} - \psi^k_{ew} = \omega^k_{wu} - (b^k_{ew} - (\beta^k_{uew} - \overline{\beta}^k_{ue})) = \nu^k_{uw} + \omega^k_{wu} - \overline{\beta}^k_{ue} < 0$$

  So the best move in this situation is to bid $b^k_{ew} \leq \omega^k_{wu}$.
- $\nu^k_{uw} + \omega^k_{wu} \geq \overline{\beta}^k_{ue}$: in this case bidding $b^k_{ew} \geq \omega^k_{wu}$ the website gets in the best T websites and gains $\nu^k_{uw} + \omega^k_{wu} - \overline{\beta}^k_{ue} > 0$. Bidding $b^k_{ew} < \omega^k_{wu}$ the websites can either get in the first T websites and gains, as before, $\nu^k_{uw} + \omega^k_{wu} - \overline{\beta}^k_{ue} > 0$ or not making to be in the first T websites and therefore gains 0. So the best move in this situation is to bid $b^k_{ew} \geq \omega^k_{wu}$.

So the dominant strategy in both the situations is bidding $b^k_{ew} = \omega^k_{wu}$

□

Corollary 2.3.1.1. *With this auction mechanism it is maximize the system surplus value.*

Proof. The consequence of $b^k_{ew} = \omega^k_{wu}$. is that the bid-value is equal to the match-value ($\beta^k_{uew} = \gamma^k_{uw}$). So when the search engine is selecting the best T website according to β^k_{uew} it is actually doing it according to γ^k_{uw} so the engine surplus value is maximize. Since all the search engines are following this strategy the system surplus value (that is equal to the sum of all the engine surplus values) is maximized. □

2.3.4 The user payment

We still have to define what users pays (μ^k_{euw}). To do that we observe that, if the estimations of ν^k_{uw} are the same for all the search engines the search engine gain must be 0. In fact all the search engines are offering the same service and have the same information. If there are two search engines and $g_{e1} > g_{e2}$ it means that users, websites or both of them gain less using the first search engine rather than using the second. In fact the sum of all the gains (the surplus value) is the same for all the search engine (it is the maximum surplus value). That means that users and websites will move to the second search engine to gain more and the first search engine will die. So competition makes the search engines to low down their gains until they reach the minimum which is 0.

Having this information we can state that the user payment is the one that make the search engine gains zero. We see that this happens setting

$$\mu^k_{euw} = \nu^k_{uw} - \overline{\beta}^k_{ue}$$

Lets in fact calculating the gain for the users, websites, and search engines.

$$
\begin{aligned}
g_u &= \sum_{w \in V_u} \nu^k_{uw} - \mu^k_{euw} = \sum_{win_u} \nu^k_{uw} - (\nu^k_{uw} - \overline{\beta}^k_{ue}) = \\
&= T * \overline{\beta}^k_e = T * \overline{\gamma}^k_e \\
g_w &= \sum_{u:w \in V_u} \omega^k_{wu} - \psi^k_{euw} = \sum_{u:w \in V_u} \gamma^k_{uw} - \overline{\beta}^k_e = \\
&= \sum_{u:w \in V_u} \gamma^k_{uw} - \overline{\gamma}^k_e \\
g_e &= \sum_{u:q_u=(e,k_u)} \sum_{w \in V_u} \mu^k_{euw} + \psi^k_{euw} = \\
&= \sum_{u:q_u=(e,k_u)} \sum_{w \in V_u} (\nu^k_{uw} - \overline{\beta}^k_{ue}) + (b^k_{ew} - (\beta^k_{uew} - \overline{\beta}^k_{ue}) = \\
&= \sum_{u:q_u=(e,k_u)} \sum_{w \in V_u} \nu^k_{uw} + b^k_{ew} - \beta^k_{uew} = 0
\end{aligned}
$$

Notice that the users gains as he is being shown the first excluded website. The websites gain the difference between their match-value and the one from the first excluded website. Finally the search engine gains 0 as we wanted.

2.3.5 Example

Let's show an example with some data (See Figure 2.1). The user search for the keyword $k =$ "pizza" to a search engine $e = "Iauu"$. There are 4 websites that made a bid for the keyword "pizza" to the search engine "Iauu".

The first website $W_1 =$ "ThePizzaEncyclopedia" has high quality information about pizza but want to get paid to have access to this information. So it has an high positive value for ν and a negative value for ω.

The second website $W_2 =$ "WeDeliverYourPizza" makes good pizza and gain from selling it. The user is happy to visit that website and the website is happy to sell pizza to the users. So both ν and ω are positive.

The third website $W_3 =$ "BoostYourselfWithPizza" is a spam website trying to sell some weird product the user don't need. So it has a low value for the user ($\nu < 0$) and an high value for the website that gains from selling something useless and overpriced.

The last website $W_4 =$ "WikiOpenPizza" is a no profit website with not that bad content. So ν is positive but ω is 0 because the website does not get anything from being visualized.

Website	ν	ω	b	β	$\beta - \overline{\beta}$
$W_1 =$ "ThePizzaEncyclopedia"	15	-4	-4	11	5
$W_2 =$ "WeDeliverYourPizza"	4	5	5	9	3
$W_3 =$ "BoostYourselfWithPizza"	-3	9	9	6	0
$W_4 =$ "WikiOpenPizza"	5	0	0	5	-1

Figure 2.1: Websites parameters

Let's set $T = 2$ so the first 2 websites are selected and $\overline{\beta} = \beta_2 = 6$.

We have that the **user**:

- pays $\mu = \sum_i \nu_i - \overline{\beta} = (15-6)+(4-6) = 7$
- gets $\sum \nu_i = 15+4 = 19$
- gains $g_u = 19 - 7 = 12$.

The **website** W_1

- pays $\psi_1 = b_1 - (\beta_1 - \overline{\beta}) = -4 - (11-6) = -9$
- gets $\omega_1 = -4$
- gains $g_{w1} = \omega_1 - \psi_1 = -4 - (-9) = +5$

The **websites** W_2

- pays $\psi_2 = b_2 - (\beta_2 - \overline{\beta}) = +5 - (9-6) = +2$
- gets $\omega_2 = +5$
- gains $g_{w2} = \omega_2 - \psi_2 = +5 - 2 = +3$

All the other **websites** (W_3 and W_4):

- pays 0
- gets 0
- gains 0

The surplus value is

$$S = g_u + g_{w1} + g_{w2} = 12 + 5 + 3 = 20$$

We see that this is equal to $S = \gamma_1 + \gamma_2 = (\nu_1 + \omega_1) + (\nu_2 + \omega_2)$ as predicted in Section 2.2.1. Notice also that the gain for users and websites are the same calculated at the end of Section 2.3.

Chapter 3

About noise

In the previous section we have assumed that search engines know exactly the value of a website from the user point of view (ν^k_{uw}). That is not always true since inferring users preferences is an hard task. We can model this lack of information as noise in the value of ν^k_{uw}. We will see that good estimation for the user preferences (so low noise on ν^k_{uw}) ensures a search engine some gain over its competitors.

3.1 Definition

Each search engine e has estimate values for ν^k_{uw} that we call $\tilde{\nu}^k_{uw}$. We assume there is additive Gaussian, zero average, noise on the real value of each parameter. So for each engine e we have that:

$$\tilde{\nu}^k_{uw} = \nu^k_{uw} + N(\sigma, 0)$$

So each search engine e has a noise represented by σ

3.2 Matching problem with noise

Given the set of users U and the set of websites W we call optimal matching for a search engine e the matching between users and websites that maximize the engine surplus value. We call the optimal engine surplus value S_{opt}. The optimal matching is obtained selecting the T best websites for each user according to the matching-value γ_{uw}^k. It means that for each user u, V_u contains the websites w_i such that γ_{ui}^k is in the top T values.

Let's analyze a scenario in which the entity that matches users and websites has a noisy information about ν_{uw}^k (and so about γ_{uw}^k) as defined in Section 3.1. It happens that the optimal matching (so the sets V_u) is decided according to the noisy values $\tilde{\gamma}_{uw}^k$. This matching has a surplus value of $\tilde{S}_{opt} = \sum_{u \in U} \sum_{w \in V_u} \gamma_{uw}^k$. Notice that this is calculated using the real values for γ_{uw}^k. That's because even if the matching is decided according to the noisy values the surplus value is still calculated using the real values.

The auction mechanism change as following:

1. The search engine knows each website bid b_{ew}^k and an estimation of the user value $\tilde{\nu}_{uw}^k$ for the websites.
2. The search engine order all the websites in descending order according to the noisy bid-vales $\tilde{\beta}_{uew}^k = \tilde{\nu}_{uw}^k + b_{ew}^k$.
3. It selects the first T websites according to this rank and put them in the results set V_u. We call $\overline{\beta}_{ue}^k$ the (noisy) bid-value of the first websites not selected (i.e. the value of $\tilde{\beta}$ for websites in position $T+1$ in the rank)
4. The search engine ask each website that is in the result set V_u to pay
$$\psi_{euw}^k = b_{ew}^k - (\tilde{\beta}_{uew}^k - \overline{\beta}_{ue}^k)$$
5. The search engine ask the user to pay
$$\mu_{euw}^k = \tilde{\nu}_{uw}^k - \overline{\beta}_{ue}^k$$
For each visualized website.

It is easy to prove that the noise low down the expected value for the engine system surplus value.

Lemma 3.2.1. *Let e_1 and e_2 be two search engines such that the first one know exactly the values of ν_{uw}^k while the second one knows them with noise we get that for each query $S_1 \geq S_2$*

Proof. That's obvious. In fact the engine surplus value is the sum of all the matching-value for each matching done by the engine. The search engine e_1 will select the optimal matching. The search engine e_2 will try to do the same but since the values of γ are affected by noise it will include in the result set some websites that should not be in that and will exclude some valuable website. So the engine surplus value for e_2 will be lower or equal than the one of e_1 □

3.2.1 Example

We now provide an example with the same data from Section 2.3.5 (see Figure 3.1) . In this situation there is noise so the final ranking is different.

Website	ν	noise	ω	b	β	$\tilde{\beta}$	$\tilde{\beta} - \overline{\beta}$
W_1 = "ThePizzaEncyclopedia"	15	0	-4	-4	11	11	6
W_3 = "BoostYourselfWithPizza"	-3	+2	9	9	6	8	3
W_2 = "WeDeliverYourPizza"	4	-4	5	5	9	5	0
W_4 = "WikiOpenPizza"	5	-2	0	0	5	3	-2

Figure 3.1: Websites parameters

Let's set $T = 2$ so the first 2 websites are selected and $\overline{\beta} = \beta_3 = 5$. Notice that the values of β that are considered for the ranking are the noisy ones. In fact the search engine knows the noisy values ($\tilde{\beta}_i$) but not the real ones (β_i).

We have that the **user**:

- pays $\mu = \sum_i \tilde{\nu}_i - \overline{\beta} = (15 - 5) + (-1 - 5) = 4$
- gets $\sum \nu_i = 15 - 3 = 12$
- gains $g_u = 12 - 4 = 8$.

The **website** W_1

- pays $\psi_1 = b_1 - (\tilde{\beta}_1 - \overline{\beta}) = -4 - (11 - 5) = -10$
- gets $\omega_1 = -4$
- gains $g_{w1} = \omega_1 - \psi_1 = -4 - (-10) = +6$

The **websites** W_3

- pays $\psi_3 = b_3 - (\tilde{\beta}_2 - \overline{\beta}) = +9 - (8 - 5) = 6$
- gets $\omega_3 = +9$
- gains $g_{w3} = \omega_3 - \psi_3 = +9 - 6 = +3$

All the other **websites** (W_2 and W_4):

- pays 0
- gets 0
- gains 0

The surplus value is

$$S = g_u + g_{w1} + g_{w2} = 8 + 6 + 3 = 17$$

We see that this is still equal to $S = \gamma_1 + \gamma_3$, so it is the sum of the matching values of the selected websites. That's because the surplus values does not depend on the payments (μ and ψ) as we proved in Section 2.2.1. In this situation the only difference with the situation without noise are the assignment of the payments. So the surplus value is still equal to the sum of the matching values of the visualized websites.

Notice also that the surplus value is lower than the one obtained when there is no noise, as proved in Lemma 3.2.1.

3.2.2 Numeric simulation

We provide a numeric simulation to show the implications of the lemmas above. In this system there are 10 users, 10 websites and $T = 1$. That means that we are matching each user with one website.

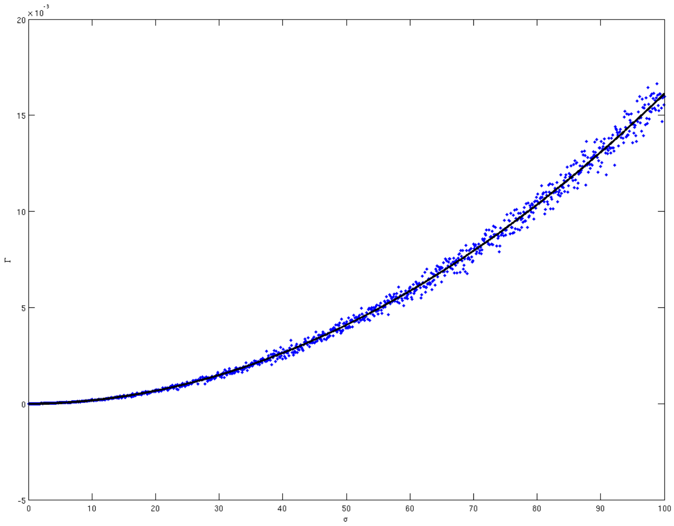

Figure 3.2: Numeric simulation for $\Gamma = \frac{S_{opt}-S}{S_{opt}}$ function of the noise variance σ and quadratic interpolation ($y = 10^{-4}(0.0159\sigma^2 + 0.0250\sigma - 0.1944)$)

We take values for ν_{uw}^k uniformly distributed in the range $(-400, 400)$. We than plot the factor $\Gamma = \frac{S_{opt}-S}{S_{opt}}$ as a function of the variance (σ) of the Gaussian noise. For each value of σ the noise and the values for ν_{uw}^k have been generated 1000 times, 1000 values for S have been calculated and we took the average. The plot has been generated with 1000 different values of σ in the range $(0, 100)$. The matlab code for the numeric simulation can be found in Section 5.1.

As we see from the simulation the value of Γ has a quadratic relation with the variance of the noise.

3.3 Consequences of the noise on the market

We realized that the noise on the values of ν affects the expected value of the engine surplus value. We will now see how that also affect the gain of search engines.

The engine surplus value is the additional value the search engine create from the matching. This additional value has to be distributed between the three entities of the market: websites, the search engine and users. That means that with an higher surplus value there is more value to distribute and so it is easier to make the entities happy. A search engine that is able to provide an higher surplus value will have more chances of gain over the other search engines as we see in the following lemma.

Lemma 3.3.1. *If there are two search engine: e_1 and e_2 such that for each query $S_1 \geq S_2$ we have that e_1 can set $\epsilon > 0$ and for each query it gains $g_{e1} = S_1 - S_2 - \epsilon$ while $g_{e2} = 0$*

Proof. The idea is that the first search engine has more surplus value to distribute so it can keep some of it for itself. In fact the first search engine can make the users and the websites gain the same they would get using the engine e_2 plus a little bias. In this way all the users and the websites using e_2 would move to e_1 because their gains would be higher with the latter. In fact for each user $u \in U$ and for each website $w \in W$ we have that the gains of users and websites using the first search engine are:

$$g'_u = g_u + \delta$$

$$g'_w = g_w + \delta$$

where g_u and g_w are respectively the gains of the users and of the websites using the second search engine. Setting δ such that $\epsilon = (|W|+|U|)\delta$ we get that

$$\sum_{u \in U} g'_u + \sum_{u \in W} g'_w = S_2 + \epsilon$$

So we have that

$$S_1 = g_{e1} + \sum_{u \in U} g'_u + \sum_{u \in W} g'_w = g_{e1} + S_2 + \epsilon$$

and we get

$$g_{e1} = S_1 - S_2 - \epsilon$$

□

Chapter 4

Conclusion

4.1 Results

We have introduced a web search market model for sponsored search dividing the entities involved in 3 sets: Users, Websites, Search Engines. We have defined the action each entity can take and we have introduced the gain as a measure each entity is trying to maximize.

We have defined the System Surplus Value as the sum of all the gains and we have shown how that is a measure of the optimality of the matching between users and websites (Section 2.2.1). We have further shown that a search engine that can provide an higher surplus value can improve its gain (Lemma 3.3.1). From that we have stated that a search engine will always try to maximize the surplus value trying to provide optimal matching between users and websites. To do that it needs two information about each possible matching: ω_{wu}^k and ν_{wu}^k. The former can be obtain from websites' bids using a truth-telling auction mechanism (Lemma 2.3.1), the latter must be inferred by the search engine in some way. We have modeled the different quality of predictions of ν_{wu}^k that different search engines can provide, using a noise model (Section 3.1).

With Lemma 3.2.1 we have shown that the precision of the predictions of ν_{wu}^k affects the surplus value provided by the search engine's matching and therefore the search engine's gain (Lemma 3.3.1). In Section 3.2.2

we have shown with a numeric simulation that there is a quadratic relation between the variance of the noise on the prediction of ν_{wu}^{k} and the expected surplus value of the matching and therefore the search engine's gain.

This pushes the search engines to invest money in mechanisms trying to predict users' preferences and leads to a better quality of search results. Also high quality websites would gain from that since their quality would be better recognized.

We can than conclude that according to our model a more open and competitive market for search engines would lead to better quality of search results and higher gains for both websites and users.

4.2 Limits of the model

We observe that the real web search market is different from what our model predicts. Search engines keep all the revenue from slot auctions and the dominant position of Google is difficult to attack. We should than analyze what are the limits of our model in order to explain why reality seems to be different from what we predict.

4.2.1 Rationality

We have assumed that all the entities take the best rational decision to maximize their gain. That is not usually true in reality. Each entity should develop a tool that can take the best decision. That requires knowledge and resources. Search engines can usually afford it but we can't say the same for websites and users.

Websites should develop a tool that analyze the available search engines and decide where to place a bid to maximize their gain.

Similarly users need to decide where to submit a query. In order to help them some *meta search engine* can be developed. A meta search engine receives a query from a user, forward it to one or more search engines and then process the results in some way to provide a result page to the user. The meta search engine has the resources to develop smart query strategies and can keep part of the gain of the queries for itself as its gain.

4.2.2 Monetization of information

When we define the gain of users $g_u = \sum_{w \in V_u} (\nu^k_{uw} - \mu^k_{euw})$ we assume that there is a way to convert the information received from the search results into money. That might be hard and very subjective. The point is that it doesn't have to be done explicitly. The user can directly "feel" his g_u without explicitly converting ν^k_{uw} into money and can than decide whether it worths to carry on using that search engine. On the other hand search engines, given some website value measure they obtain from ranking algorithms, can learn how to convert it into money checking how much the users are willing to pay (i.e. what is the ranking-money conversion that makes the user stop using the search engine)

4.2.3 Few search engines

The model assume that there are enough search engines such that the competition is real and it is difficult to reach a cartel agreement between competitors. But in reality there are few quality search engines. That is because search engines need a huge amount of computational power and storage so few companies can afford to start one.

The model doesn't actually assume that there are *many* search engines. It just requires that there is real competition between them. To assure that it is sufficient that there is at least one quality search engine that keep little gain for itself (just to pay off the structure it needs). It might be an idea for a no-profit organization with enough resources to start a search engine with these characteristics in order to increase the competition in the web search market and get the results we have described in this thesis.

4.2.4 Transfer money at each query

In our model money is transfered between users and search engines and between websites and search engines at each query. Websites are used to do that since the bidding schemes that are actually implemented in the main search engines requires that[1]. On the contrary users are not used to exchange money at each query.

Users might don't like the idea of spending money for searching even if for some query they might actually receive money. That can again be solved with meta search engines that can provide the search service for free to users that are not interested in gaining from searching and that want just a easy way to search information. The meta search engine can keep a balance for each user and provide free search service until the sum of all the payments the users own to the search engine is positive. For providing this service it can keep this positive amount. Of course more skilled users can contact directly the search engines without using meta search engines.

4.2.5 Search engines could not accept to share revenue

Our model predict that users and websites can participate in search engine revenue. That should be enough to prove that this is possible under our assumptions. By the way it is useful to provide some evidences of this already happening in reality on similar situations.

Youtube provide a Partner program for revenue sharing[15]. Creator of original contents can earn revenue allowing relevant advertisements to be displayed with their videos and getting part of the revenue from advertising. HubPages has a similar program for user generated articles [1].

Revenue sharing in advertising market also happens in those websites that abuse of the pay per click mechanism to make money. Those websites usually create pages with many ads from some ad network (i.e.

[1]Payments usually happen on per-click base instead of per-view (and so at each query) but since there is a linear relation between views and clicks it doesn't make any difference.

AdSense) and then pay users to click on ads, sharing with them part of the revenue. This mechanism is known as *click fraud* and advertising network have tools to try to prevent it.

4.2.6 Are then the results realistic?

Our model is of course a simplification of reality and has some limits. But as we have shown above these limits can be easily overcome. Different and more complete models can be developed for example using a per-click payment scheme (we use a per-view), giving different values to the available T slots (in our model they all have the same value) and taking into consideration reputation and net effect, but our results will still be valid. In fact our model is founded on few simple assumptions that still hold in more complex models: rationality of entities, monetization of information, competition between entities, uncertainty on user-website matching values. Therefore the results that come from those assumption are still valid in more complex models.

Chapter 5

Appendix

5.1 Numeric simulation matlab code

```
clc; clear;

N = 10; % users/websistes
max_val = 400;  %values from -max_val to max_val
max_epsilon = max_val/4;  % values of noise from -max_epsilon
                                                % to max_epsilon
avg_steps = 1000;
epsilon_steps = 1000;
epsilon = zeros(epsilon_steps,1);
ratio = zeros(epsilon_steps,1);
S_noise_avg = zeros(epsilon_steps,1);
S_real_avg  = zeros(epsilon_steps,1);
for l=1:epsilon_steps
    epsilon(l) = max_epsilon * (l/epsilon_steps);
    S_noise = zeros(avg_steps,1);
    S_real = zeros(avg_steps,1);
    for k=1:avg_steps
        ni_real = 2*(max_val)*rand(N,N) - max_val;
        ni_real = tril(ni_real);
        ni_real = ni_real + tril(ni_real)';
        for i=1:N
                ni_real(i,i) = ni_real(i,i)/2;
        end
        omega_real = 2*(max_val)*rand(N,N) - max_val;
```

```
        omega_real = tril(omega_real);
        omega_real = omega_real + tril(omega_real)';
        for i=1:N
                omega_real(i,i) = omega_real(i,i)/2;
        end
        gamma_real = ni_real + omega_real;
        noise = normrnd(0,epsilon(l)/2,N,N); %gaussian noise
        noise = tril(noise);
        noise = noise + tril(noise)';
        ni_noise = ni_real + noise;
        gamma_noise = ni_noise + omega_real
        [values, indices] = max(gamma_noise,[],2);
        for i=1:N
            S_noise(k)= S_noise(k) + gamma_real(i,indices(i));
        end
        S_real(k) = sum(max(gamma_real,[],2));
    end
    S_noise_avg(l) = mean(S_noise);
    S_real_avg(l) = mean(S_real);
end
p = polyfit(epsilon,(S_real_avg - S_noise_avg) ./ S_real_avg,2)
plot(epsilon,(S_real_avg - S_noise_avg) ./ S_real_avg, '.b',
        0:0.05:max_epsilon,polyval(p,0:0.05:max_epsilon), '-k',
        'LineWidth', 3)
ylabel('\Gamma')
xlabel('\sigma')
```

Bibliography

[1] Hubpages. `http://hubpages.com`.

[2] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 19–26, New York, NY, USA, 2006. ACM.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[4] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. Working Paper 11765, National Bureau of Economic Research, November 2005.

[5] Ron Holzman, Noa Kfir-dahav, Dov Monderer, and Moshe Tennenholtz. Bundling equilibrium in combinatorial auctions, 2001.

[6] Google Inc. Google annual report 2010.

[7] Google Inc. Personalized search. `http://www.google.com/support/accounts/bin/answer.py?answer=54041`.

[8] Interbrand. Best global brand 2011. `http://www.interbrand.com/`.

[9] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

[11] Ferdinand Pavel. Competion in the web search market. A report for microsoft, DIW econ GmbH, July 2009.

[12] Yahoo Research. Computational advertising. `http://research.yahoo.com/Computational_Advertising`.

[13] Stanford University. Mse 239: Introduction to computational advertising. `http://www.stanford.edu/class/msande239/`.

[14] Yahoo. Yahoo annual report 2010.

[15] Youtube. Partner program. `http://www.youtube.com/partners`.