Buying and Selling Traffic: The Internet as an Advertising Medium

A dissertation presented by
Zsolt Katona
to INSEAD faculty
in partial fulfilment of the requirements for the degree of
PhD in Management

February 2008
Dissertation Committee:
Miklos Sarvary (chairman)
Elie Ofek
Paddy Padmanabhan
Timothy Van Zandt

Abstract

The Internet is rapidly growing as a marketing medium. This year online advertising expenditures will reach approximately $20 billion in the US alone. Two formats dominate online advertising: (i) Web sites buying advertising links from each other and (ii) search engines selling sponsored links on their results pages. The first part of the dissertation studies the former advertising model and investigates the network structure that emerges from advertising links. In a world in which consumers 'surf' the WWW, Web sites' revenues originate from two sources: the sales of content (products and services) to consumers, and the sales of links (traffic) to other sites. In equilibrium, higher content sites tend to purchase more advertising links, mirroring the Dorfman-Steiner rule. Sites with higher content sell fewer advertising links and offer these links at higher prices. Thus, sites seem to specialize in terms of revenue models: high content sites tend to earn revenue from sales of content, whereas low content sites tend to earn revenue from sales of traffic (advertising). I test these findings in a variety of empirical studies. The second part of the dissertation explores the other dominant form of online advertising: paid placement. Here, a search engine auctions sponsored links next to the search results. Advertisers submit bids for the price that they are willing to pay for a click. The model focuses on two key characteristics of this problem: (i) the interaction between the search list and the list of sponsored links and (ii) the dynamic forces that influence bidding behavior when sites compete for the sponsored links over time. The findings explain the seemingly random order of sites on the sponsored links list and their variation over time. The results have important managerial implications for both sellers and buyers of online advertising.

Contents

1 Introduction

The Internet and its most broadly known application, the World Wide Web (WWW) are gaining tremendous importance in our society. The Web represents a new medium for doing business that transcends national borders and attracts a significant share of social and economic transactions. A large part of these transactions involves advertising. The most basic form of advertising on the Web is when a Web site sells an advertising link by displaying an ad on one or more of its pages for which the advertiser pays a fee based on the page impressions or the clicks on the ad. A site can be an advertiser and a publisher of advertising at the same time. In this way, Web sites buy and sell the traffic of potential consumers who visit them.

A key feature of the WWW is that it is a decentralized network that evolves on its own, based on its members' incentives and activities. The goal of the dissertation's first part is to develop a model that helps understand what structure emerges from this decentralized network formation process. Understanding this network structure is important for all firms participating in e-commerce. The network structure has a crucial role in determining the flow of potential consumers to each site, which is key for demand generation. A primary interest of search engines, for instance, is to understand how sites' contents are related to their connectedness on the Web. In turn, Web-sites need to be strategic about connecting themselves in the Web to ensure that search engines correctly reflect or even boost their rank under a given search word.[1] Indeed, "search-engine optimization" has grown into a $1.25 billion business with a growth rate in 2005 reaching 125%.

The second part will examine a new but rather popular form of advertising:

---

[1] In response to Google's regular updates of its search algorithm, different sites shuffle up and down wildly in its search rankings. This phenomenon, which happens two or three times a year is called "Google Dance" by search professionals who give names to these events as they do for hurricanes (see "Dancing with Google's spiders", *The Economist*, March 9, 2006).

search advertising. Potential advertisers bid for a place on the list of sponsored links that appears on a search engine's "results" page for a specific search word. In 2006, the revenues from such paid placements have doubled compared to 2005, reaching almost \$16 billion[2]. This fast growing market is increasingly dominated by Google, which today, controls some 56 % of Internet searches. How such advertising is priced and what purchase behavior will advertisers follow for this new form of advertising is investigated in this section.

I develop a model, that takes into account different aspects of paid search advertising. In doing so, my goal is to shed light on the advertising patterns observed on Google search pages. Specifically, search pages can be characterized by a variety of patterns in terms of the identity and position of sponsored links. In particular, there is no clear relationship between the "results list" of search and the list of sponsored links. Sometimes, a site may appear in both or in only one (either one) of the lists. For example, for the search word "travel", the two lists are different. However, for the search word "airlines", United Airlines appears as the first search result and the second sponsored link. One can also observe significant fluctuations in the sites' order in the sponsored links list. Besides generating normative guidelines to both advertisers and the search engine on how to buy and sell sponsored links, my model generates testable hypotheses that account for the variations described above.

It is important to confront the analytical results with empirical data. The third part of the dissertation contains several empirical studies. In the first study, I compare the results to previous empirical work (Broder et al. 2000, Faloutsos et al. 1999) that examined the degree distribution of the WWW. A broad result found across these studies is that links follow a scale-free power-law distribution with an exponent of around 2. It is an empirical puzzle however, that this degree distribution

---

[2]See "Where is Microsoft Search?", *Business Week*, April 2, 2007, p. 30. Total revenues from paid placements is expected to reach \$45 billion by 2011.

is the same for both in- as well as out-links. In this study I show, how the model can explain this pattern. In the second study, I collect data from a search engine. For a variety of search words, I record how much advertising Web sites in different positions sell and relate this to their content. This study confirms the hypothesis that Web sites with lower content sell more advertising. Finally, in a third study, I examine sites that buy advertising on Google search pages in the form of sponsored links. On these sites, I estimate the amount of sold advertising and confirm that this quantity is in an inverse relationship with the site's profitability.

The rest of this dissertation is organized as follows. In Section 2, I summarize the model and the results on the structure of the Web. Then, in Section 3, I present the search advertising model. In Section 4, I describe the empirical analysis. Finally, I conclude with a discussion of the results.

2 Network Formation and the Structure of the Commercial World Wide Web

The WWW includes an extremely broad community of Web sites with a vast array of motivations and objectives. We cannot pretend to be able to capture all relevant behaviors on such a diverse network. Rather, we restrict our attention to the *commercial* WWW, by which we mean the collection of interlinked sites' whose objective is to profit from economic exchange with the public and/or each other. In the following, by WWW, we will always refer to this "sub-network". As such, our goal is to explain the network formation process and the resulting network structure of the commercial WWW.

The primary way through which sites can drive traffic to themselves is the purchase of advertising links.[3] At the same time, each site also has the option to sell the traffic reaching it by selling such advertising links to other sites. In a network where each site is a potential advertiser and a potential seller of advertising, what determines the tradeoff between selling content or advertising? In particular, how does this tradeoff depend on the site's popularity or attractiveness to the browsing public? A closely related question is how should sites price their advertising links as a function of their content. Finally, even on the commercial WWW, many of the links are so-called "reference links", that sites establish to other sites in order to boost their own content or credibility (Mayzlin and Yoganarasimhan 2006). Sites need to understand, how such links complement or interact with advertising links to determine the ultimate network structure. Addressing these practical problems requires the understanding of the "forces" that drive the evolution of the network's structure and the resulting competitive dynamics.

---

[3]In 2006, Internet advertising has reached $10 billion with a yearly growth rate of over 25% (see "Marketing Budgets Are Up 46% for Q2", www.emarketer.com, July 5, 2006).

Specifically, we propose a network model in which the nodes represent rational economic agents (sites) who make simultaneous and deliberate decisions on the advertising in-links they purchase from each other. Agents are heterogeneous with respect to their endowed "content", which may be thought of as their inherent value in the eyes of the public/market. Consumers are assumed to 'surf' on the web of nodes according to a random process, which is nevertheless closely linked to the network structure. Sites generate revenue from two sources: (i) by selling their content to consumers and (ii) by selling links to other sites. We start by assuming that the price per traffic of each link is an increasing function of the originating site's content. Next, we show that this is indeed the case in an equilibrium where sites first set their prices for advertising links and then purchase links at these prices in a second stage. We also extend the model to the case where beyond buying and selling advertising links, sites can also establish reference out-links to each other at a small cost. Finally, we explore the situation when a substantial part of the public uses search engines. In this context, we ask what happens when nodes represent multiple content "areas".

We find that in equilibrium, higher content sites tend to buy more advertising links, mirroring the Dorfman-Steiner rule well-known for traditional media but, so far, not explored for a network medium. Similarly, reference links tend to point to high content ones. As such, in equilibrium, the number of *all* in-links is closely correlated with the site's content. This explains why search engines have so much success using algorithms based primarily on in-links (e.g. Google's Page Rank) for ordering pages in terms of content in the context of a search word. The model also has a number of practical implications for the pricing of Internet advertising. We find for instance, that sites with higher content should set a higher price-per-click for their advertising links. This, combined with our result on the purchase of advertising links indicates that there is a tendency for specialization of commercial sites' business models. Higher content sites emphasize product sales driving traffic to the site, while

lower content ones emphasize the sales of traffic by mainly selling advertising links. Therefore, high content sites tend to sell fewer advertising links than low content sites. Figure 1 shows the example of "aa.com" and "kayak.com". Both sites sell airline tickets and related products. American Airlines supposedly makes a higher margin on its visitors since it sells its own tickets, whereas Kayak does not get any revenue from selling the tickets, therefore the former is a high content site, whereas the latter is a low content site. As the figure shows, the sold advertising quantities support our results that the low content site sells more advertising (on the right under "sponsored") than the high content site.

The two sites in Figure 1 constitute the two extreme types. However, according to the results sites with a medium content also sell advertising but not huge amounts. The example in Figure 2 show the site "travelocity.com". This site also sells plane tickets and charges a fixed amount for each ticket, therefore its profit margin is higher than Kayak's but lower than an airline's. As the snapshot of the site shows, it sells one advertising link in the bottom of the page, which fits into the pattern that our results suggest.

A tendency for specialization also exists in content areas. Specifically, if we allow sites to cover multiple content areas, we can show that, the more consumers use search engines, the more sites have an incentive to specialize in terms of content areas. Finally, we can show that the above equilibrium patterns are generally consistent with the empirical reality of the commercial WWW. In particular, we find that in-links follow a similar degree distribution as out-links as it is empirically observed on the WWW, but not predicted by existing models of network formation.

While the marketing literature related to the Internet has grown considerably in recent years, there is virtually no research exploring the link-structure of this new medium or the likely forces that drive its evolution. This is not to say that social sciences and economics in particular have not examined the endogenous formation

Figure 1: High and low content sites: aa.com and kayak.com

Figure 2: A "medium" content site: travelocity.com

of networks. In an influential paper, Bala and Goyal (2000), for instance, develop a model of non-cooperative network formation where individuals incur a cost of forming and maintaining links with other agents in return for access to benefits available to these agents. Recent extensions of the model (Bramouille et al. 2004) also consider the choice of behavior in an (anti-)coordination game with network partners beyond the choice of these partners.[4] These models have several features, which do not really apply to the WWW. First, they concentrate on the cost of link formation, which is shown to be critical for the outcome. More importantly, the above papers consider that individuals in the network are identical. For example, in Bala and Goyal (2000), linking to a well-connected person costs the same as connecting to an idle one. This is clearly not the case on the WWW, where large differences exist between the sites' contents and their connectedness. Also, on the WWW the cost of establishing a link largely depends on where this link originates from. Finally, the equilibrium networks emerging from the above models clearly do not comply with the structure of the WWW. Bala and Goyal (2000), for instance, find two possible equilibrium network architectures, the "wheel" and the "star" or their respective generalizations.

Our work also relates to the vast literature on advertising (see Bagwell (2005) for a good recent review).[5] Of particular interest for us are studies dealing with advertising firms' choices of advertising quantities and the pricing of advertising by media firms. Advertising quantities have been known to be determined by the advertisers' product margins (Dorfman and Steiner 1954) and, of course, by the effectiveness of advertising. Advertising expenditures have also been shown to be affected by product quality in a variety of context. Nelson (1974) and Schmalensee (1978) develop

---

[4]See also Jackson and Wolinsky (1996) for an early paper concerned with the relationship between social network stability and efficiency and Jackson (2003) for a recent summary of this literature.

[5]See Zeff and Aronson (1999) for an early summary of advertising on the Internet and Hoffman and Novak (2000) for a qualitative description of online advertising pricing models. See also Iyer and Padmanabhan (2006) on Internet referral services.

a theory of advertising as a signal of quality. Villas-Boas (2004) studies advertising effort in the context of discrimination between high and low quality products and Agrawal (1996) computes equilibrium advertising levels in the presence of differential brand loyalty. Our model does not map into these situations but our results linking advertising quantities to sites' content relate to the variety of outcomes identified in these papers.

On the supply side, recent papers in marketing (see Dukes and Gal-Or 2003) have shown that advertiser- and media-competition also have a significant effect on advertising quantities. Advertising prices have also been shown to be influenced by the above market features but recently, two additional factors have been revealed to be of further interest: (i) the disutility of advertising (Masson et al. 1990) and (ii) the competitive pricing of media *content* (Godes et al. 2006). Our paper builds on this literature but is markedly different from it in many respects. First, our model studies advertising via links of a network, i.e. advertising effectiveness is endogenous as it depends on the network's structure. Also, advertising is used to increase traffic, not to inform, nor to signal quality or affect brand loyalty. More importantly, in our model, advertisers and the media are not separate entities. Each site is a buyer *as well as* a seller of advertising. A central question is: which one of these activities dominates and how does this decision depend on the site's content.

Finally, our work is also related to recent papers modeling consumers' browsing process on the WWW. Our demand structure is based on the classic model by Brin and Page (1998) to provide a consistent description of how consumers flow on a complex network of sites. We use some of the recent mathematical results related to this framework, in particular Langville and Meyer (2004). We extend our model using the concept of a reference-link, as in Mayzlin and Yoganarasimhan (2006), to designate out-links that sites establish to other sites in order to improve their own perceived value by consumers. With these elements, we develop a model that is more

15

consistent with the reality of the WWW than those of the existing network formation literature.

The next section presents this basic model, which considers advertising links and exogenous prices. Section 2.2 extends this model to a two-stage game where sites price advertising links in the first stage and then, purchase in-links from each other. Section 2.3 explores two further extensions: (i) the introduction of reference out-links and (ii) the existence of search engines in a context where content is multi-dimensional. The section ends with a general discussion and concluding remarks. To improve readability, most proofs have been delegated to the Appendix.

2.1   The Model

We describe Web sites and the links between them as a directed graph, G. The nodes of the graph correspond to the sites and the directed edges to the links between the sites. Let $i \to j$ denote if there is a link from node i to node j and $i \nrightarrow j$ if there is no link between them. The number of links going out from a site is the out-degree of the site, denoted by d_i^{out}, and the in-degree is the number of its incoming links, denoted by d_i^{in}.

It is important to note that we consider as the unit of analysis a single Web *site*, which may possibly include multiple pages. Technically, on the WWW, the nodes correspond to the Web pages. However, most of the time, a Web site offering a single product consists of several pages having almost all links established between them. The incoming links of the site usually go to one of the main pages and the outgoing links can go from any page. We argue that in a model of network formation, these pages should be considered as one single node representing *the* Web site. All the links going out and coming into a site's sub-pages should be assigned to this one node.[6]

---

[6]This perspective is shared by search professionals. When Google calculates the rank of a page

Beyond structural reasons, considering sites as the unit of analysis also makes sense because they represent a single decision maker.

In what follows, we will describe consumers' browsing behavior on such a graph, followed by the description of the network formation game played by the sites. In doing so, we need to stay at a relatively high level of abstraction. In particular, we will consider a homogeneous group of consumers and a reduced form profit function for sites.

2.1.1 Consumer browsing process

The primary task in modeling the WWW is to describe the process through which users browse the Web, i.e. how they move from one site to the other. We will consider these users as potential consumers, who may buy the content (product) sold at a particular site. We normalize their total number to 1. Furthermore, we will neglect consumer heterogeneity and simply assume that a consumer reaching a site may consume the content of that site or "purchase" it with probability ρ, that we can assume to be 1, without loss of generality. Our goal is to establish the number of visitors at a site (in a given unit of time). To do this consistently is not a trivial task because the weight (incoming traffic) of incoming links depends on how much traffic reaches *their* originating sites, i.e. how many in-links the incoming links themselves have. Obviously, two incoming links have very different effect on a site's traffic if they originate from different locations. In other words, we need to describe the flow of consumers consistently across *all* nodes of the network.

We will use the simple but very powerful solution proposed to this problem by Brin and Page (1998), which became one of the basic principles for Page Rank, the

---

in its search function for instance, it calculates it for the whole site and not for single pages within a site. A possible way to do this is to consider all the pages that are in the sub-directories under the same domain name of a site. For example any page with an address "www.amazon.com/..." is considered as part of the "Amazon" site.

algorithm that Google's search engine uses to order Web pages. Assume n sites and imagine that the total mass of consumers (1 unit) is initially distributed equally between these n sites. A consumer follows a random browsing behavior in every step. Starting from site i, with probability δ, s/he randomly follows a link going out from that site or stays there, choosing each of these $d_i^{out} + 1$ options with equal probability.[7] With probability $1 - \delta$, s/he jumps to a random site on the Web, again choosing each site with equal probability. The number of steps while the user follows the links without jumping then follows a geometric distribution, with expectation $\frac{1}{1-\delta}$. δ is called the "damping factor" and in practice it is often set to $\delta = 0.85$, which corresponds to an expected "surfing distance" of around 6.67, that is, almost seven links. Figure 3 illustrates the flow of consumers following the links.

It can be shown that the iteration of the above process results in a limit distribution of consumers between Web sites. This limit distribution is called Page Rank (PR).[8] It can be thought of as the number of visitors at a Web site per unit time. By definition, PR has to satisfy the following equation:

$$r_i = \frac{1 - \delta}{n} + \delta \left(\frac{r_i}{d_i^{out} + 1} + \frac{r_{i1}}{d_{i1}^{out} + 1} + \frac{r_{i2}}{d_{i2}^{out} + 1} + \ldots + \frac{r_{ik}}{d_{ik}^{out} + 1} \right), \qquad (1)$$

where r_i is the Page Rank of site i (i.e. the proportion of visitors reaching it), $i1, i2, \ldots, ik$ are the sites linking to site i and d_{ij}^{out} denotes the number of links going out from site ij, that is the j-th site linking to site i (without counting the loops).

Describing the process over time for all sites, let $r^{(t)}$ denote the row vector resulting from the iteration after step t. With this notation $r^{(0)}$ denotes the initial vector

---

[7]The event when a consumer stays at the website can be formally represented by drawing a loop around the node.

[8]Although Page Rank usually refers to the score that Web sites receive from Google, we use Page Rank to describe the scores that are calculated of this simple version of the algorithm.
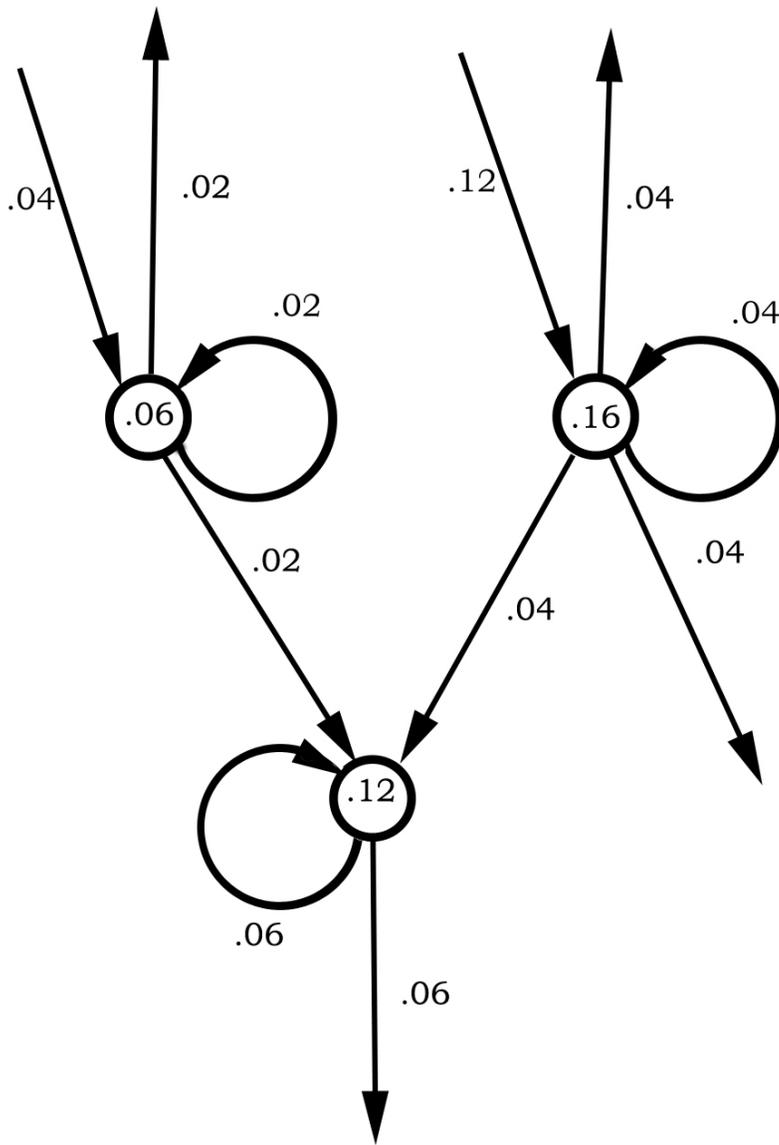
Figure 3: Flow of visitors not showing those who jump to random pages.

of the iteration which, we set without loss of generality to $r^{(0)} = (\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$, i.e. we distribute browsers uniformly across all nodes. The iteration is defined through the M transition probability matrix, whose cells are:

$$[M]_{ij} = \begin{cases} \frac{1}{d_i^{out}+1}, & \text{if } (i \rightarrow j), \\ 0 & \text{otherwise.} \end{cases}$$

Notice, that the i-th row of the matrix represents node i and the number in cell ij represents the probability of moving to node j from node i. Using M, the iteration described above reads:

$$r^{(t+1)} = \delta \cdot r^{(t)} M + (1 - \delta) r^{(0)}. \tag{2}$$

If the series $r^{(t)}$ is convergent as $t \rightarrow \infty$ and it converges to r, then r provides the PR values of the nodes in the network. These can be thought of as the steady number of visitors at a Web site per unit time. It can be shown using Markov-chain theory that the iteration is indeed convergent if the graph satisfies some properties (see Langville and Meyer (2004) for details). We only use the following lemma.

Lemma 1 (Langville and Meyer 2004) *If $r^{(t)}$ is a probability distribution for every t, then the series is convergent as $t \rightarrow \infty$.*

Obviously, in the initial step, $r^{(0)}$ is a probability distribution, but $r^{(t+1)}$ does not satisfy this unless each row of the matrix M contains at least one non-zero element, that is, every node in the graph has at least one out-link. The loops added to the nodes ensure that this holds.

Using the matrix form of definition (1), if iteration (2) is convergent and it converges to r, then it has to satisfy:

$$r = \delta \cdot rM + (1 - \delta) r^{(0)}. \tag{3}$$

Notice that if r is a probability distribution, then for any matrix $[U]_{ij} = \frac{1}{n}$, $rU = (\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$. Hence (3) can be written as

$$r = \delta \cdot rM + (1 - \delta)rU = r(\delta M + (1 - \delta)U). \tag{4}$$

This formula helps interpret the meaning of Page Rank by describing it as the weighted average of two matrices (M and U) each representing a different random process. M contains the transition probabilities across linked sites, i.e. it moves browsers along the links of the network. Thus, it encapsulates the structure of the Web. In contrast U represents a process that scatters browsers randomly around to any of the sites. The weights given to these two processes are defined by δ, the damping factor.[9] Thus, Page Rank and the underlying process is a consistent description of how traffic is distributed across sites for any given link structure of the network.

2.1.2 Network formation

Assume that there are n nodes (sites) with given constants $c_1 \leq, \ldots, \leq c_n$, representing their contents. These content parameters can be thought of as some measure of the Web sites' value for the public in a particular content domain. For instance, the site may sell a product and c may represent consumers' willingness to pay for this product. Then, the variation in c may be thought of as heterogeneity across sites in terms of product quality. In this spirit, we assume that the site's net revenue from a consumer is proportional to this parameter: the higher the public values the site, the higher the income from a consumer visiting it. The site's net revenue will also be proportional to the total number of consumers being at the site, as measured by

---

[9]It is also interesting to note that r is the eigenvector of the matrix $\delta M + (1 - \delta)U$ with its principal eigenvalue, 1.

r_i, i.e. site i's total income from its consumers is:

$$r_i c_i.$$

The cost of each site has a fixed and a variable component. The fixed component can be set to 0 without loss of generality. We assume that the variable component (e.g. a shipping cost) that is proportional to the number of visitors is identical across sites. Let C denote this per-visitor cost. Then, the total cost of a site is:

$$r_i C.$$

We assume that there is a market for links between sites. Every node, i offers links for a fixed price-per-click, q_i, which varies across nodes as will be clarified below. This is consistent with general media (or Internet) practice where ad rates are typically quoted as "rates per click-through".

The number of clicks on a particular link can be calculated from the consumer flow model. If site i has traffic r_i and d_i^{out} out-links, then the number of visitors clicking on a particular out-link will be $\delta r_i / (d_i^{out} + 1)$. Then, the total price of an advertising link from site i will be $p_i = \delta r_i q_i / (d_i^{out} + 1)$.

If another node purchases a link then this link will be created and pointing from the seller to the buyer. Given prices, nodes makes simultaneous decisions about their incoming links, that is, which other nodes they buys links from. Each node is allowed to buy one link from every other node. Essentially, this market can be thought of as the advertising market. If a node buys a link, it pays for an advertisement to be placed on the seller's page.

In our baseline model, the per-click prices for links are exogenous but we will relax this assumption in Section 2.2.2. Specifically, in this section we will assume that $q_i = q(c_i)$ is an increasing function of content c_i and that prices are not too high (see (24) in the Appendix). In Section 2.2.2, we show that in a two-stage game

where prices are set first, followed by the purchase of links, equilibrium prices are indeed set this way. Nevertheless, even this exogenous pricing structure as reflected by the choice of $q(c)$ is quite intuitive. Price-per-click increasing in content allows us to capture the basic tradeoff between keeping a consumer or handing it over to another site. The higher the gain from a consumer (i.e. the higher c), the higher the site wants to charge for potentially letting him/her to surf to another site. In other words, this price function captures the tradeoff between sites' two revenue streams.[10]

With these elements, a site's profit, for a given network structure consists of its income from its consumers plus the advertising income (from sold links) minus the advertising costs (of bought links). Formally:

$$u_i = r_i(c_i - C) + p_i \cdot d_i^{out} - \sum_{j \to i} p_j. \tag{5}$$

2.1.3 Equilibrium analysis

Our objective is to determine the Nash-equilibria of a game where players' objective function is given by (5) and their strategies consist of buying links from one another in a simultaneous decision. These equilibria represent a network or a graph (a set of links between the nodes) and our main interest is in understanding the structure of this graph. The following proposition describes the general structure of these equilibria.

Proposition 1 *At least one Nash-equilibrium always exists and all the equilibria have the following properties.*

---

[10]Notice, that in our model, sites control their sold advertising links only through their pricing. This may not entirely capture the strategic interaction between sites. For example, a site may not allow advertising by a strong rival even at a high price. We will discuss this issue in detail at the end of the paper and would like to thank the review team for pointing it out.

(i) *The out-degree is a weakly decreasing function of content in the following sense. If, for a given pair of nodes $c_k < c_l$, then $d_k^{out} \geq d_l^{out}$.*

(ii) *If all the content parameters are different, then in-degree and Page Rank are increasing functions of content.*

Proof (Sketch): Here we give the main logic of the proof while the detailed proof is provided in the Appendix. In the first step, we show that in equilibrium all the nodes buy links from the nodes with the lowest q's. This does not mean that they will buy from the nodes charging the lowest price for links, but rather from those, which sell their traffic at the lowest "per-click price". Based on the increasing price structure, these must be the sites with lowest content parameters, hence out-degree is a decreasing function of the content parameter. Then, we show that nodes with higher content can buy more links, hence in-degree is an increasing function of the content. Due to the special structure of the network this yields that the Page Rank is also an increasing function of content. $\square$

Figure 4 shows a possible equilibrium network structure. Once the nodes are arranged according to their content (top left graph), the network structure reveals the simple tendency whereby most links originate from low content pages (small dots) and are directed towards high ones (large dots). The lower part of the figure shows how in- and out-links depend on content, where nodes are arranged in increasing order of content. Of course, if we suppose that all the content parameters are different, then (i) is equivalent to saying that the out-degree is a decreasing function of the content parameter. If there are identical content values, the nodes can still be ordered (as is done on the figure) such that both the contents are increasing and the out-degrees are decreasing.

This general equilibrium structure of the model, that advertising links tend to go from lower content sites to higher content ones, is quite interesting. Essentially,

24
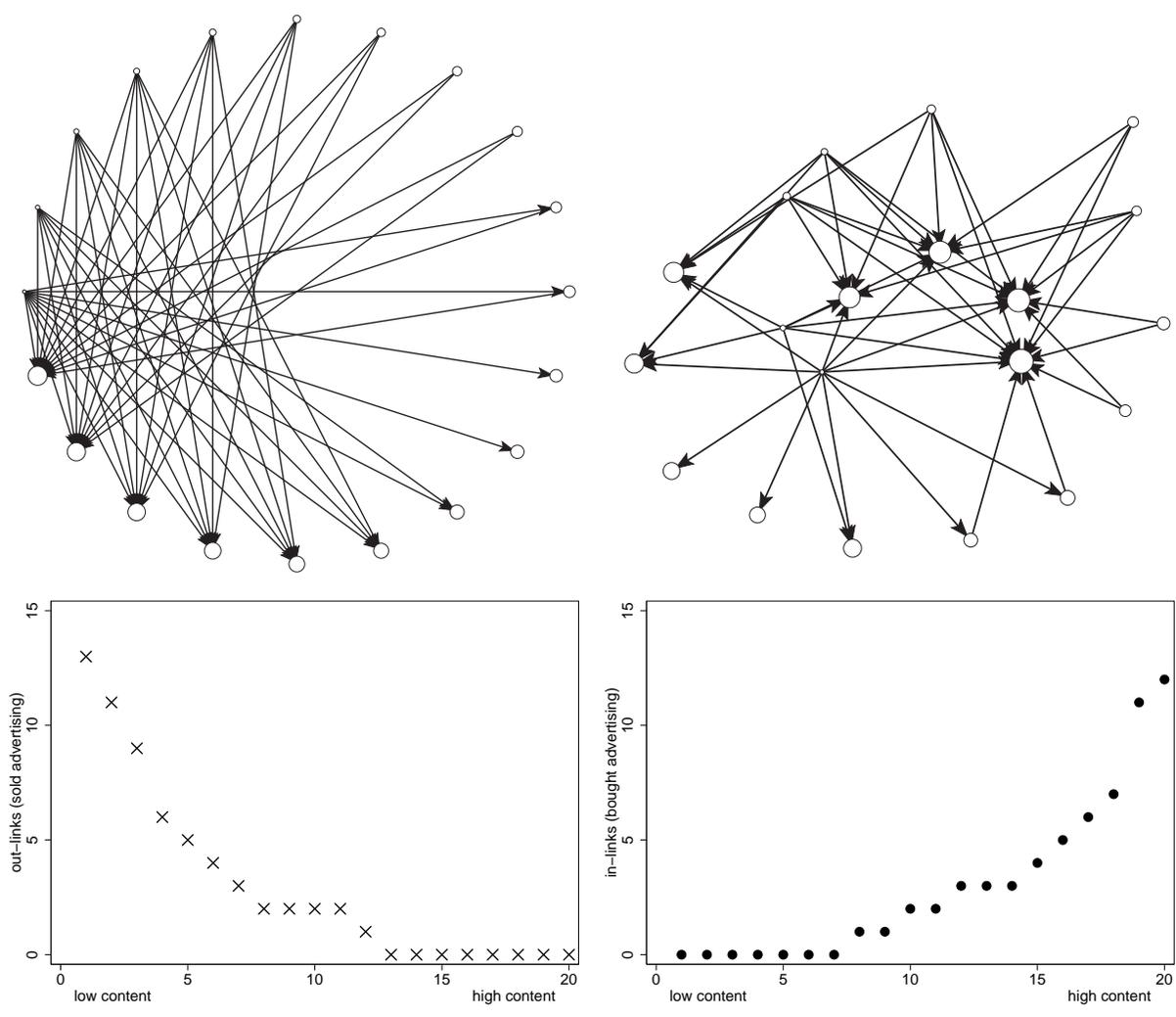
Figure 4: The top two figures depict the same network, a possible equilibrium network, where larger nodes denote higher content. The bottom graphs represent the number of out- and in-links for each node, where nodes are arranged in increasing order of content.

it means that high content sites are the most important buyers of advertising. This result is similar to the Dorfman-Steiner advertising rule well-known in traditional media.[11] It is particularly interesting that this result continues to hold even in a network context where sellers of advertising are competing for traffic to sell their own content. The result also seems to have face validity as the biggest advertising sites tend to be large well-known brands. Surveying the last decade in online advertising, DoubleClick, for example, documents that by 2005, Fortune 500 companies' share of all online advertising reached 30% and has steadily increased over time. Similar, trends emerge for Europe as well.[12]

The result is also interesting, because it suggests that sites have a tendency to specialize in their business model. Certain sites, the ones with low content specialize in selling links (i.e. traffic), while sites with high content tend to buy links (advertise) in order to benefit from content (product) sales. However, there are also sites that do both, which is specific to the Web.

To summarize, the network's formation is characterized by two features: (i) pages tend to buy links from other sites with lower contents and (ii) the higher the content of a site the more links it will buy from other sites. This results in a network where the number of in-links correlates with the value of the corresponding site.

2.2 Endogenous prices and infinitely many sites

After analyzing network formation with per-click prices as parameters, we now study a game where prices and links are both decision variables. In particular, a key driver of our results so far was the assumption that q_i is increasing in content. Our goal is to show that this is true even with endogenous prices and that the network formation

---

[11]We would like to thank the Area Editor for pointing out this similarity.

[12]See, "The Decade in Online Advertising" and "The Online Advertising Landscape in Europe", DoubleClick, April/September 2005 as well as Zeff and Aronson (1999) p.7.

results hold. Specifically, we analyze a two-stage game where in the first stage, sites set per-click prices for advertising links and in the second stage, they establish links between each other, given prices. The second stage game, as it was described in Section 2.1.2, would be too complex to solve for any fixed set of q_i parameters. However, the size of the Web suggests that we should consider the case when the number of players is large enough so that a single site's decision does not have a significant effect on the other sites. To capture this idea, we suppose that there are infinitely many sites or a continuum of sites. We describe such a model next.

2.2.1 Network formation

In the infinite version of the original network formation game, suppose that the set of players is the interval $I = [0, 1]$ and each player corresponds to a node of the infinite directed graph.

Definition 1 *A directed graph on the set I is defined as a subset $G \subseteq I \times I$, where an element $(x, y) \in G$ corresponds to a directed link from $x \in I$ to $y \in I$.*

The definition of the degrees of the graph requires measure theory. We will call the subsets of I measurable if they are measurable with respect to the Lebesgue-measure on the interval I, denoted Λ.

Definition 2 *The out-degree of $x \in I$ in the graph G, is the measure of those nodes to which links from x exist, that is $d^{out}(x) = \Lambda\{y \in I | (x, y) \in G\}$ if the set is measurable, otherwise the out-degree does not exist. Similarly, the in-degree of $y \in I$ is defined as $d^{in}(y) = \Lambda\{x \in I | (x, y) \in G\}$ if the set is measurable.*

We will restrict ourselves to graphs where all the degrees exist, that is, the corresponding sets are measurable. We will show that any equilibrium graph has to be

27
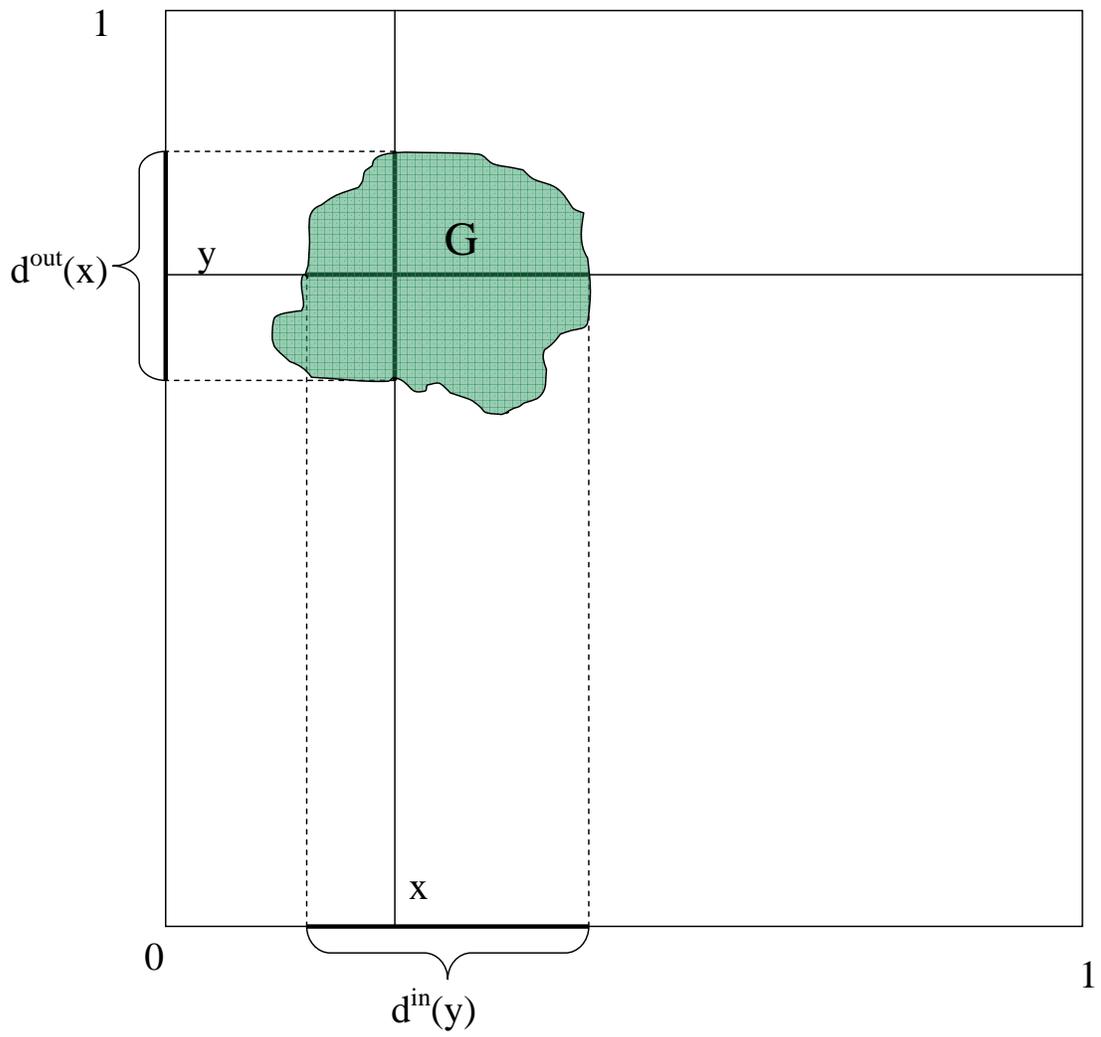
Figure 5: The representation of an infinite network and its degrees.

such. Directly generalizing the game, we assume that the measurable function $c(i)$ provides the content of site $i \in I$ and the measurable function $q(i)$ represents the per-click prices. We can assume without loss of generality that $c(i)$ is increasing, i.e. sites are ordered by content on I. The Page Rank 'function' is also directly generalizable. However, in the infinite case, we have to deal with the problem of zero out-degrees. If the set of nodes that buy links from node i, is a zero measure set, then $d^{out}(i) = 0$. In the finite case, the solution is to establish a loop around node i, but that would also be a zero-measure set in the infinite case. Hence, we introduce the variable $s > 0$, accounting for the visitors who stay at site i. Then, the proportion of visitors who stay at the site is $\frac{s}{s+d^{out}(i)}$. Therefore, the equation defining Page Rank will be

$$r(i) = (1 - \delta) + \delta \frac{s}{d^{out}(i) + s} r(i) + \delta \int_{x \to i} \frac{r(x)}{d^{out}(x) + s} dx. \tag{6}$$

It can be interpreted as a density function describing the marginal probability of visitors being at different sites. A $(1-\delta)$ proportion of visitors is jumping to random pages and the rest of them are following the links. Note that, in the $s = 0$ case, we can derive (6) by multiplying (1) by n and changing the notation to $r(i) := nr_i$. Then, as $n \to \infty$ we obtain (6). To make sure that players are not indifferent between different choices, we assume that $\Lambda(q^{-1}(x)) = 0$ for every x, that is, not many sites have the exact same price. The total price for a link at site i is $p(i) = \delta r(i) q(i) / (d^{out}(i) + s)$. Then, site i has the following utility function.

$$u_i = r(i)(c(i) - C) - p(i) \cdot d^{out}(i) - \int_{j \to i} p(j) dj. \tag{7}$$

For this infinite game, the main results that were valid for the discrete case still hold. If $q(.)$ is an increasing function of content and satisfies (24), there always exists an equilibrium and in this equilibrium, in-degree is increasing and out-degree is decreasing in content (and in i). Proposition 2 formally states this result.

29

Proposition 2 *If $q(i)$ is increasing satisfying (24), and the functions c and q are continuous, at least one pure-strategy Nash-equilibrium exists and in any equilibrium $d^{in}(i)$ is increasing and $d^{out}(i)$ is decreasing.*

Proof: See the Appendix.

Since the number of players is infinite, a single player does not have a significant impact on the game. Let us capture this by the following definition.

Definition 3 *Two measurable functions q and q' : $[0, 1] \rightarrow \mathbf{R}$ are equal almost everywhere ($q = q'$ a.e.) if $\Lambda\{x | q(x) \neq q'(x)\} = 0$, that is, if they only differ in a small set.*

Lemma 2 *If $q = q'$ a.e., then the set of equilibria of the games corresponding to the two functions are equal a.e., that is, for any equilibrium function $d^{in}()$ for q, there exists an equilibrium for q' with a $d^{in'}() = d^{in}()$ a.e.*

Proof: Let X denote the set $\{i | q(i) \neq q'(i)\}$. The payoffs and the optimal decisions do not change for the sites that are not in X. For those, who are in X, the optimal decisions may be different, but these players are in a null set. $\square$

Now that we have characterized the equilibria in the second stage (network formation) game, we will show that $q(i)$ is increasing in any equilibrium of the two-stage game.

2.2.2 Price setting

In the first stage, every site selects its $q(i)$ simultaneously, only knowing the content function. In the second stage, sites establish links. Since the two-stage game may have several sub-game perfect Nash-equilibria, even unreasonable ones, we will rule out some of them based on Lemma 2.

Definition 4 *A sub-game perfect equilibrium* $(q, E(q))$ *of the two-stage game is a refined sub-game perfect Nash-equilibrium, if*

(i) $E(q)$ is a pure-strategy Nash-equilibrium of the second stage and

(ii) If $q = q'$ a.e., then $E(p) = E(p')$ a.e.

This definition makes sure, that to any refined SPNE corresponds an SPNE, and any SPNE with the property that an infinitesimal perturbation in prices $(q \sim q')$ leads to a qualitatively different network in the second stage is not a refined SPNE. Therefore, sites have an expectation about the second stage's network structure in the first stage, and this expectation does not change if only a few sites change their prices. This approach ignores certain direct strategic effects of the pricing decision. Specifically, we assume that sites react to the distribution of prices across all other sites. With infinitely many sites, this distribution does not change if a single site alone changes its price. This assumption is realistic in the context of the WWW where there are over 10 billion pages and no site dominates the traffic on the entire network. Using this equilibrium concept, our main result is the following.

Proposition 3 *For any refined SPNE of the two-stage game, the first stage's $q(.)$ function has to be increasing.*

Proof: See the Appendix.

The significance of Proposition 3 is that it supports our assumption that in the network formation stage of the game, the per-click prices of advertising links increase with respect to the sites' content. Among other findings, this reinforces our previous result that sites tend to be specialized in terms of their revenue models. Sites with low content tend to sell traffic to higher content sites by selling advertising links for relatively low prices. Figure 6 shows a possible infinite equilibrium network. High-

31
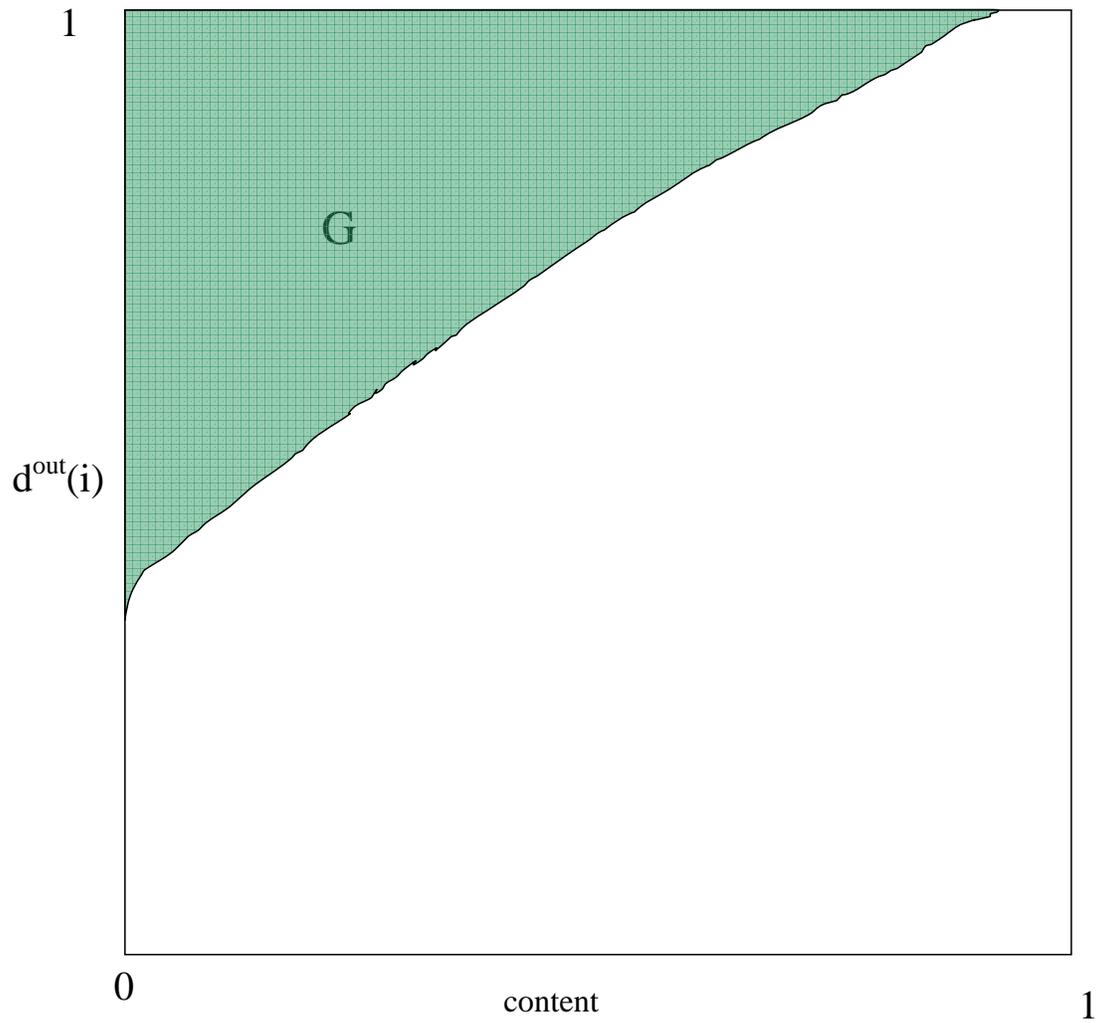
Figure 6: The representation of an equilibrium network. The horizontal axis represents content and the outdegrees can be read from the vertical axis.

content sites on the other hand benefit more from the sales of their content to the public. They price their advertising links high and, as a result, sell few advertising links.[13] The intuition behind the result is that sites with a higher content have a higher potential of making profits on their visitors. Hence they set higher prices to be able to sell fewer links. This way a higher proportion of their visitors become their customers, resulting in a higher average margin per visitor. In the second stage these sites purchase more advertising, since they can more effectively leverage the traffic they buy.

2.3 Extensions

In what follows, we explore three extensions to the model. First, we allow sites to create reference links. These are out-links that sites may establish to boost their effective content. Second, we incorporate advertising disutility in model, by assuming that potential consumers tend to spend less if there are too many ads on a site. Finally, we explore the impact of search engines allowing sites to have multiple content areas.

2.3.1 Reference links

So far, we have focused on a specific type of links: advertising links. These links are established for a fee to direct consumers to the Web site of the advertiser. Here, we introduce another type of link that is commonly used in the non-commercial Web: reference links.[14] These links also have an important role in forming the structure of the commercial Web. Reference links are used to increase the referring sites' content with the help of the referred pages (Mayzlin and Yoganarasimhan 2006). The number

---

[13] "Hot, well-targeted content sites have [..] been able to command very high prices." Zeff and Aronson (1999), Chapter 7, p.176.

[14] We are indebted to one of the reviewers for suggesting this extension.

of reference links going out from (coming in) a site is denoted by d^{out_R} (d^{in_R}). Every node is allowed to establish one reference link from itself to every other node at maintenance cost κ. Each site is allowed to establish an (outgoing) reference link to every other site. The advertising links are still included in the model, as they were in the original version, that is, each site is allowed to buy one (incoming) advertising link from every other site. Let $i \rightarrow_R j$ denote if there is a reference link from i to j and $i \rightarrow_A j$ if there is an advertising link between them, whereas the number of incoming (outgoing) advertising links is denoted by d^{in_A} (d^{out_A}).

Thus, the strategy of player i can be described by two vectors, each consisting of 0's and 1's. The first vector $\mathbf{x_i^R}$ determines to which nodes player i establishes reference links to ($x_i^{R(j)} = 1$ if s/he forms a reference link to node j and 0 if not). The second vector $\mathbf{x_i^A}$ describes which nodes s/he buys advertising links from ($x_i^{A(j)} = 1$ if s/he buys a link from node j and 0 if not). In the case when i decides to refer to j and j decides to buy an advertising link from i, we assume that both links are established and this is the only case when two links pointing in the same direction are allowed between two nodes. Also, in order to get around the problem that players might be indifferent between two or more possible choices of links, we will assume that if a player is indifferent s/he establishes as many links as possible.

The incentive to create reference links is to increase a site's content by referring to other sites. Therefore, we generalize the payoff function by using the "accumulated" or "effective" content term, which consists of two elements: (i) the site's resident content, c_i, (ii) the sum of the content of sites linked to through reference links multiplied by a scaling constant $0 \leq \beta < 1$. Therefore, the total payoff of node i is defined as follows:

$$u_i = r_i \left(c_i + \beta \sum_{i \rightarrow_R j} c_j - C \right) - \kappa d_i^{out_R} + p_i \cdot d_i^{out_A} - \sum_{j \rightarrow_A i} p_j. \tag{8}$$

Introducing the reference links makes the problem much more complex, since a site cannot control its traffic by buying the appropriate number of advertising links, the traffic is also affected by the incoming reference links. In order to solve the game we use the following simplification. Instead of using the stochastic model, to describe the flow of consumers, we use a traffic function with the following properties. Let $r_i = f(d_i^{in_R}, d_i^{in_A})$ be the traffic or demand that reaches the site. f is a function of the site's in-degrees and we assume that it is increasing and strictly concave in both advertising links $(d_i^{in_A})$ and reference links $(d_i^{in_R})$. This assumption is strongly supported by practice and is one of the basic principles behind search engine design. Describing Google's search engine, *The Economist* claims for example, that "[t]he most powerful determinant of a Web page's importance is the number of incoming referral links, which is regarded as a gauge of a site's popularity".[15] We also make the natural assumption that f has increasing differences in $d_i^{in_R}$ and $d_i^{in_A}$. That is, $f(x+h_1, y+h_2) - f(x, y+h_2) \geq f(x+h_1, y) - f(x, y)$ for any $x, y \geq 0$ and $h_1, h_2 \geq 0$, i.e. the two kinds of in-degrees are weakly complements. Then, the utility function becomes:

$$u_i = f(d_i^{in_A}, d_i^{in_R}) \left(c_i + \beta \sum_{i \to_R j} c_j - C \right) - \kappa d_i^{out_R} + p_i \cdot d_i^{out_A} - \sum_{j \to_A i} p_j. \qquad (9)$$

With this generalization we can show the following.

Proposition 4 *If $p_i = p(c_i)$ is increasing, then the game has an equilibrium, and in any equilibrium, if $c_i > c_j$ then $d_i^{in_R} \geq d_j^{in_R}$, $d_i^{out_A} \leq d_j^{out_A}$, $d_i^{in_A} \geq d_j^{in_A}$ and $d_i^{out_R} \geq d_j^{out_R}$.*

Proof: See the Appendix.

[15]Ibid. See also "How Google works", *The Economists Technology Quarterly*, September 18, 2004.
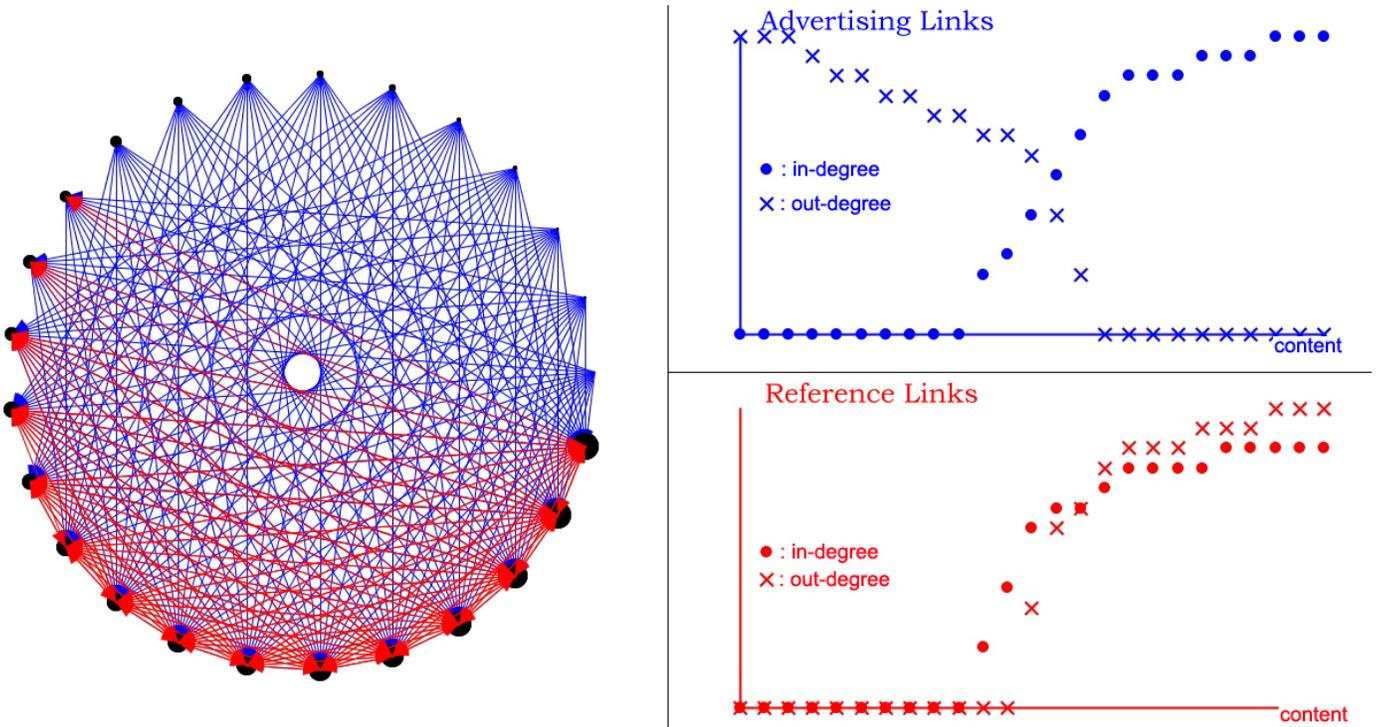
Figure 7: A possible equilibrium network obtained in a simulation, n=25

Keeping the assumption that prices are increasing in content, we can show that the structure of the network formed by the advertising links is qualitatively the same as without reference links. The network formed by the reference links has a similar structure but with the opposite order of out-degrees. For both networks, the in-degrees are increasing in content, whereas the out-degrees are decreasing in content for advertising links and increasing for reference links. Figure 7 shows a possible equilibrium network.

The intuition for the distribution of reference links is quite simple. Clearly, each site will try to establish reference links to the highest content sites, which benefit

36

more from these in-links as they have a higher margin on the additional traffic generated by these in-links. Therefore, high content sites can afford to establish more reference out-links increasing their margin even more. The presence of advertising links intensifies this effect since outgoing reference links and incoming advertising links are complements. The more reference links a site establishes the more advertising links it has an incentive to buy. Thus, the increased traffic from these advertising links results (indirectly) in extra profit from outgoing reference links.

The general feature of the equilibrium network, that higher content results in more reference in-links is very interesting. It provides, for instance, an explanation for why the famous search engine, Google had so much success introducing the quantity Page Rank for search. Google's objective is not only to find all the pages containing the search expression, but also to rank them according to their content. Since measuring content directly is difficult, it can use Page Rank as an indirect measure because, according to our model, in equilibrium, high Page Rank should be correlated with high content.

2.3.2 Advertising disutility

The obvious downside of selling advertising links is that visitors leave the site before making a purchase. However, consumers may also be annoyed by ads leading to a decreased willingness to pay. Here, we extend the model by assuming that consumers' utility decreases if the site that they visit contains many advertisements. This will decrease their willingness to spend money on that Web site. We will capture this phenomenon by introducing a negative element in the content term that linearly increases with the (advertising) out-links. Thus, the total payoff of site i is defined as follows.

$$u_i = r_i(c_i - \gamma d_i^{out} - C) + p_i \cdot d_i^{out} - \sum_{j \to i} p_j,$$

37

in the case with only advertising links and

$$u_i = r_i \left(c_i + \beta \sum_{i \to_R j} c_j - \gamma d_i^{out_A} - C \right) - \kappa d_i^{out_R} + p_i \cdot d_i^{out_A} - \sum_{j \to_A i} p_j$$

in the general case with reference links. $\gamma \geq 0$ measures the disutility for advertising. A closer examination shows that the introduction of advertising disutility does not change the complexity of the problem; the outcome of the game and the proofs are essentially the same. The reason is that in all the results the out-degrees are decreasing in content. Subtracting this decreasing term from the increasing content makes it even more increasing. This makes the results more accentuated with a higher γ parameter, that is, with consumers more sensitive to the negative effects of advertising.

2.3.3 Search engines and multiple content areas

Search engines (SE) play an important role in the formation of the network. If some consumers use SEs, then the number of visitors at a Web site does not only depend on the structure of the network but also on how search engines display the site in the result of a given search. Today's SEs use a twofold method to determine which pages and in what order to display the result of a search. On the one hand, they measure content directly, on the other hand, they measure content indirectly through the structure of the network, using methods such as Page Rank. To examine the effect of SEs we will assume a single SE that filters the s highest content sites for its users, where s is a fixed integer. We also assume that traffic is distributed across these s sites proportional to each site's Page Rank. Note that we do not consider the SE as a strategic player.

As will become clear later, when considering SEs, we need to generalize our model in another respect, letting content have multiple dimensions. Specifically, we assume

that content is a D-dimensional vector $\mathbf{c_i} = (c_i^1, c_i^2, \ldots, c_i^D)$. These dimensions can be seen as content areas (e.g. entertainment or e-commerce in various domains, etc.). We assume that a particular consumer visiting the site is only interested in one dimension of the site.[16] The proportion of consumers interested in the different dimensions is represented by the weight vector $\mathbf{w}$. This vector can also be interpreted as the probability distribution on content dimensions describing the interest of a randomly selected consumer. Thus, the expected consumer-specific content at site i is the scalar product $\mathbf{w} \cdot \mathbf{c_i}$, which can also be called the (weighted) average content of a page.

Then, in the generalization of our model (5), the income of a Web site from selling its content changes from $r_i c_i$ to $r_i \cdot \mathbf{w} \cdot \mathbf{c_i}$. Thus, still without the presence of SEs, the total utility of node i is

$$u_i = r_i(\mathbf{w} \cdot \mathbf{c_i} - C) + p_i d_i^{out} - \sum_{j \to i} p_j, \tag{10}$$

where we assume that $p_i = \delta q_i r_i/(d_i^{out}+1)$ and $q_i = q(\mathbf{w} \cdot \mathbf{c_i})$ is an increasing function of average content.

It is easy to see that this generalized model results in the same equilibrium as the one described in Proposition 1. The only difference is that we need to replace content with the weighted average content in the Proposition. This shows that without introducing the SEs in the model, multi-dimensional content does not make much difference. In particular, if sites had the possibility to change the allocation (distribution) of their total content across specific content areas, they would not have an incentive to do so, since only (weighted) average content matters.[17]

What happens if we incorporate SEs in the model? Let us assume that only a

---

[16]This assumption can be relaxed. If a consumer is interested in several dimensions we assign a probability distribution to his/her interest.

[17]Notice that the "cost of content" associated with a certain area is proportional to the consumer interest in that dimension.

b proportion of consumers is browsing according to the process described in Section 2.1.1. The remaining $(1 - b)$ consumers use a SE in every step of browsing, which directs them to a Web site in the following way. As we mentioned before, a consumer is only interested in one dimension of content, hence s/he runs a search in that dimension. Through the result of the search, the SE directs the consumer randomly to one of the top content sites in that dimension. More precisely, the SE selects the pages with the s highest content parameters in every dimension and directs consumers to one of these with probability proportional to their Page Rank.[18] Let S_d denote the set of the s highest content pages in dimension d and I_i^d denote the indicator of the event $(i \in S_d)$, that is, whether the content of site i in dimension d is among the top s contents. Then, the probability that a consumer from a SE gets to a given page in dimension d is either 0, if it is not one of the top content sites in the search dimension, or r_i/R_d, where $R_d = \sum_{l \in S_d} r_l$ is a normalizing constant in dimension d. Thus, the income from consumers in dimension d at site i is:

$$br_i c_i^d + (1 - b) r_i c_i^d \frac{I_i^d}{R_d} = r_i c_i^d (b + (1 - b) I_i^d / R_d).$$

Using notation $\mathbf{C_i} = (C_i^1, C_i^2, \ldots, C_i^D)$, where $C_i^d = c_i^d I_i^d / R_d$, the expected income from selling content at page i is: $r_i(b\mathbf{w} \cdot \mathbf{c_i} + (1 - b)\mathbf{w} \cdot \mathbf{C_i})$. It is important to see the difference between $\mathbf{c_i}$ and $\mathbf{C_i}$, the latter being the content vector truncated by the search engine by eliminating (setting to 0) the dimensions that do not make it in the top s ranks. The term $(1 - b)\mathbf{w} \cdot \mathbf{C_i}$ can then be interpreted as the expected reward from the search engine for being a top site in one of the content dimensions i.e. a sort of "specialization reward". Let E_i denote the modified average content $b\mathbf{w} \cdot \mathbf{c_i} + (1 - b)\mathbf{w} \cdot \mathbf{C_i}$. Then, the total utility of site i is

$$u_i = r_i(E_i - C) + p_i d_i^{out} - \sum_{j \to i} p_j, \tag{11}$$

---

[18]This is consistent with practice. For example, there are very few consumers who go beyond the second page of Google's search results.

where $p_i = \delta q_i r_i / (d_i^{out} + 1)$ and $q_i = q(.)$ is an increasing function of the modified average content, E_i, as defined before.

Clearly, with a single content area, the existence of a search engine does not matter qualitatively. It simply makes the "divide" between low and high content pages more pronounced. Assuming multiple content areas, the equilibria can be described by the following proposition.

Proposition 5 *At least one pure strategy Nash-equilibrium always exists and all the equilibria have the following properties.*

(i) *The out-degree is a weakly decreasing function of the modified average content in the following sense. If, for a given pair of nodes $E_k < E_l$, then $d_k^{out} \geq d_l^{out}$.*

(ii) *If we suppose that all the modified average contents are different, then the in-degree and the Page Rank are increasing functions of the modified average content.*

Proof: The proof follows from that of Proposition 1, replacing c_i with E_i. $\square$

The above properties of the equilibrium graph show that the sites with the highest E_i will have the highest in-degree and Page Rank. Since E_k is the linear combination of (i) the average content of site k and, (ii) the expected reward from the SE for offering leading content in particular dimensions, the proposition implies that in the presence of a search engine the allocation of content between dimension really matters. Specifically, there is an incentive to specialize in a certain content area in order to be one of the top sites of a particular dimension and, in this way maximize the "specialization reward". On the other hand, this incentive to specialize decreases as the average content of a site is higher, since a high average content site does not have to allocate all its resources to one dimension, it can afford to diversify its

41

content. Thus, we would expect sites with low total content to specialize, while those with high general content to diversify. However, as more and more people use search engines the advantage from high *average* content disappears and ultimately all sites compete for higher content in a specific area.

2.4   Discussion and conclusion

We proposed to model the commercial WWW based on the idea that profit maximizing Web sites purchase (advertising) in-links from each other to direct traffic to themselves in order to sell their content. A key feature of the model is that sites are heterogeneous in terms of their content. Homogeneous consumers are assumed to browse the Web in a random process directed by the network's link structure. First, we supposed exogenous per-click prices for in-links that increase in content. Later, we showed that with endogenous prices this pattern is confirmed in equilibrium. In two extensions, we introduced the presence of search engines and the possibility for sites to establish reference out-links to each other. In each case, we were interested in the equilibrium network structure as well as sites' differing incentives as a function of their content.

Overall, we found that in all equilibria, both advertising and reference links point to higher content sites. This result strongly supports the broadly accepted search heuristic, which heavily relies on the number of in-links to rank sites with respect to content. This can explain, for instance, why Google's Page Rank algorithm works so well in practice, by showing that in equilibrium, the number of in-links is positively related to a site's content. In contrast to in-links, the pattern of out-links is markedly different for advertising and reference links. Sites tend to purchase advertising links from lower content sites, i.e. the number of advertising out-links is negatively related to the content of a given site. In the case of reference links however, it is higher

content sites that tend to establish more out-links. We also show that, in the presence of search engines, this structure becomes more pronounced.

These results provide useful guidelines for marketing managers on how to manage their firms' site(s) in terms of their connectedness in the Web. First, competition seems to provide strong incentives for sites to specialize in terms of their business models. Low content sites benefit more from the sales of traffic (advertising) even though they can only price such traffic at modest rates. High content sites on the other hand, benefit more from revenues earned from content sales to consumers. These sites should charge high prices for advertising links and, as a result, sell few of these. Instead, they are better off attracting traffic by purchasing advertising links. Because of this increased traffic, high content sites also benefit more from reference links and should therefore, establish more such links. Finally, if we consider multiple content areas, then we can show that low content sites have an incentive to specialize by area while high content ones benefit more from diversification. Translating to practice, this may mean that in the context of e-commerce for instance, a strong online retail brand, like Amazon.com can afford to have a broad product assortment, while a small retail brand may have to specialize in one category to be successful.[19]

Limitations and future research

Our stylized model is limited in several ways. Probably the most severe limitation comes from our assumptions on consumer behavior. We have assumed away explicit consumer search and reduced it to a random browsing process. More importantly, we ignored consumer heterogeneity in preferences for content. Such heterogeneity could be of two kinds: vertical and/or horizontal. With respect to the first, while we assume

---

[19]In our context, Amazon is a 'high content' site in the sense that consumers' willingness to pay for items (books, CDs, etc.) is higher than their willingness to pay for the *same* items at another online retailer.

sites to be different in terms of content that could be broadly identified with 'quality', we do not model heterogeneity in terms of consumers' willingness to pay for content. Such considerations would need to take explicitly into account sites' pricing of content that would make the model prohibitively complex. Similarly, in one extension, we consider heterogeneity in consumers' interest for certain 'content areas' but we do not allow firms to influence this interest. Again, this would require the explicit consideration of pricing and maybe even the modeling of the advertising message (i.e. positioning). Clearly, neglecting these important aspects of consumer behavior limits the practical applicability of the paper. Rather than providing very specific recommendations for firms, our results should be interpreted as broad structural patterns/tendencies spanning the WWW. A more detailed modeling of consumers (including search and heterogeneity in preferences) is an obvious direction for future research.

Our model has important limitations on the firms' side as well. For example, we assumed a generic profit function across sites that only differed in terms of sites' content. Another limitation is that sites were not allowed to strategically choose their out-links. Rather, the creation of out-links is only influenced by each site's pricing strategy, which in turn only depends on the distribution of prices. This aspect of the model may not fully represent the competitive dynamics between sites. For example, two sites competing head on for consumers may not accept advertising from one another even if they would do so for other sites at a given price. Again, such idiosyncratic relationships would change the micro-structure of links around certain key sites. One could only speculate that, in these cases, rather than the regular patterns of our equilibrium structures, one would expect the emergence of clusters around a few large sites.

One way to account for a site's strategic decisions about out-links would be allowing sites to price discriminate. In a possible generalization of the model, sites

could sell their out-links charging different (per-click) prices to different sites. We do not solve this general model but we conjecture that the equilibrium structure would be similar to that in our simple model. High content sites would generally charge higher prices and a particular site's price would be increasing in the content of the potential buyer. The intuition is that high content sites still want to sell fewer links, thus charge higher prices, but they also want to make the highest possible profit on sold links. Therefore, a site would ask for a higher price if the buyer is willing to pay more (if it has higher content). Other ways to consider the strategic formation of out-links and the resulting link structures is certainly a valuable direction for future research.

Given the above limitations, one should naturally ask: are the presented equilibrium network patterns consistent with empirical evidence? In Section 4.1, we compare our results to previous empirical work (Broder et al. 2000, Faloutsos et al. 1999) that examined the degree distribution of the graph (i.e. the histogram of links) formed by the WWW. A broad result found across these studies is that links follow a scale-free power-law distribution with an exponent of around 2. It is an empirical puzzle however, that this degree distribution is the same for both in- as well as out-links. Our model can explain this pattern. Specifically, in Section 4.1, we establish the relationship between the degree distributions of in- and out-links. In particular, we show that, if either of these is a scale-free power-law distribution with an exponent of around 2, then in- *and* out-links follow the *same* degree distribution as is the case in reality. As such, our equilibrium network structure is more consistent with the empirical features of the WWW than those of previous theoretical models' that do not consider heterogeneity across sites and/or do not treat sites as utility maximizing agents. In this respect, a key contribution of our model is that it *explains* what drives Web sites' choices of links.

The WWW is a fascinating new medium with an important effect on our economy

and society. This paper is just a small step towards understanding its structure. As discussed above, there are many opportunities for both theoretical and empirical work to further explore the drivers of its evolution.

3 The Race for Sponsored Links: A Model of Competition for Paid Placement on a Search Engine

Previous research studying search advertising has focused on the problem of multi-item (or position) auctions to study the optimal bidding behavior of advertisers (Varian 2007, Edelman et al. 2007). However, a key characteristic of paid placement is that the consumer is facing two "competing" lists of sites that are both relevant in the context of the particular search: (i) the "results list" of the search and (ii) the list of sponsored links (see Figure 8). Furthermore, membership and position on the results list is exogenous and typically represents the site's popularity or inherent value. The search engine cannot use this list strategically without losing credibility from users. Thus, the existence of this search list cannot be ignored when one evaluates sites' bidding behavior for sponsored links appearing on the same page.

Another key characteristic of the problem is that the search engine can take into account advertisers' inherent traffic when awarding paid links. As the bids correspond to payments per-click, this information is important in determining the search engine's total revenue from a given sponsored link. Google's superior technology, Ad-Works consists in taking sites' click-through rates into account in addition to their per-click bids when awarding paid placements. Finally, the search engine can also determine how many sponsored links it auctions away on its site. Again, advertisers' incentives for bidding and, in turn, the search engine's revenue will depend on this decision.

We develop a model, that takes into account these aspects of paid search advertising. In doing so, our goal is to shed light on the advertising patterns observed on Google search pages. Specifically, search pages can be characterized by a variety of patterns in terms of the identity and position of sponsored links. In particular,

47

Google

Web   Images   Groupes   Actualités   **plus »**

travel   [Rechercher]   Recherche avancée
                                                           Préférences

Rechercher dans : ● Web ○ Pages francophones ○ Pages : France

Web   Résultats **1 - 10** sur un total d'environ **768 000 000** pour **travel** (0,04 secondes)

Les séjours Carlson Wagonlit **Travel** : tous les séjours dans le **...**
Travelonweb, le site de Carlson Wagonlit **Travel**, vous propose des séjours, week-end, et
croisières aux meilleurs prix. Retrouvez sur notre site les plus **...**
www.**travel**onweb.com/ - 14k - En cache - Pages similaires

travel price
www.**travel**price.com/ - 4k - En cache - Pages similaires

Hotel, séjours, voyage, location de voiture avec Roots **Travel**
Réservez votre hotel, votre voyage ou votre séjour en ligne. Roots **Travel**, votre agence de
voyage spécialisée dans les voyages et séjours chez l'habitant.
www.roots**travel**.com/ - 69k - En cache - Pages similaires

Sante, Vaccination
Pour tout savoir sur la vaccination, le paludisme dans le monde.
www.**travel**sante.com/ - 12k - En cache - Pages similaires

Tahiti Nui **Travel** - Voyage à Tahiti, Bora Bora et les atolls de **...**
Informations visiteurs | Cartes | Webcam Tahiti et Bora Bora| Réservez vos vols, hôtels et les
resorts de Tahiti, croisières, package plongée ou votre lune **...**
www.tahitinui**travel**.fr/ - 93k - En cache - Pages similaires

Travel: Airfares, hotels, vacations, cruises, car rentals and more **...** - [Traduire cette page]
Offers a variety of **travel** services, including fare finder, hotel and car rental information,
reservations, and destination information.
www.**travel**ocity.com/ - 98k - En cache - Pages similaires

Yahoo! **Travel** - [Traduire cette page]
Yahoo! **Travel** is a comprehensive online **travel** destination, where you can reserve cheap
airline tickets, hotels, cars, cruises and vacation packages, **...**
travel.yahoo.com/ - 183k - 22 mai 2007 - En cache - Pages similaires

Voyages: Absolute **Travel** Mag, participez.com / magazine du voyage **...**
Voyages: Absolute **Travel** Mag, magazine mensuel du voyage et de la découverte.
www.participez.com/ - 19k - En cache - Pages similaires

Liens commerciaux

Billets Avion : comparez
10 millions de Vols à comparer
Trouvez le billet le moins cher !
www.easyvols.fr

Cheap **Travel**
Book cheap flights, Car rentals,
Hotels... Great offers !
www.ebookers.fr

Cheap **Travel**
Trouvez votre billet d'avion à prix
tout compris en un clin d'œil !
www.liligo.fr

Petchey **Travel** Shop
Cheap flights to the Algarve
beach front accommodation
www.petchey**travel**shop.com

Billets d'Avion
Comparez et trouvez en moins de
30 secondes le vol le moins cher !
illico**travel**.com/comparer_vols_secs

Figure 8: Search for "travel" at google.fr. The list on the left side is the results list of the search and the list on the right side represents the list of sponsored links or paid placements. In this case, the two lists do not overlap.

there is no clear relationship between the "results list" of search and the list of sponsored links. Sometimes a site may appear in both or in only one (either one) of the lists. For example, on Figure 8 the two lists are different. However, on Figure 9, representing the results page for the search word "airlines", United Airlines appears as the first search result and second on the sponsored links list. One can also observe significant fluctuations in the sites' order in the sponsored links list. Finally, the number of items listed in the sponsored list is also changing over time. Beside generating normative guidelines to both advertisers and the search engine on how to buy and sell sponsored links, our model generates testable hypotheses that account for the variations described above.

Since search advertising is mostly responsible for the growth of the online advertising business, it has attracted significant interest in the economics literature[20]. Edelman et al. (2007) analyze the generalized second price auction that is used by most search engines to allocate sponsored links on search pages. The paper focuses on equilibrium properties and compares these to other auction mechanisms. Varian (2007) studies a similar problem but assumes away uncertainty and shows that the equilibrium behavior matches empirical pricing patterns for sponsored links. In a related paper, Chen and He (2006) study bidding for paid placements but assume differentiated advertisers and consumers who are initially uncertain about their valuations. They show how the auction mechanism improves the efficiency of consumer search and results in possible price dispersions for advertising.

Our work is different from this literature in that our focus is on the interaction between the search engine's basic role to find relevant sites in a given search context and its private objective to sell sponsored links on search pages. We model the

---

[20]The other dominant advertising model - sites buying ads on each other's pages - is analyzed in Katona and Sarvary (2006). That paper studies equilibrium advertising prices and the endogenous network structure determined by the advertising links.

Figure 9: Search for "airlines" at google.fr. United Airlines is a top member on both search and sponsored lists.

inherent competition between the output of these two processes and evaluate its effect on advertisers' behavior. In terms of modeling the allocation of sponsored links, our paper is closest to Varian (2007) but our focus is elsewhere. In addition, a key difference is that we assume a concave response function to advertising that is well documented in marketing. Finally, as opposed to the above papers, we also explore the endogenous choice of the number of sponsored links offered by the search engine.

The dynamic advertising model we study in the second part of the paper is related to previous work on the dynamic setting of marketing variables in a competitive context using a Markovian game. For an application on advertising see Villas-Boas (1993), while an application for dynamic R&D competition can be found in Ofek and Sarvary (2003). Our work uses a similar framework and relates to the results of both papers. The possibility of an alternating advertising pattern is similar to Villas-Boas (1993) and is largely driven by decreasing returns on advertising. However, in our model, as in Ofek and Sarvary (2003), we have a contest as advertisers' bid for each position on the list with only one winner. Finally, this dynamic model is also somewhat related to the dynamic auction model of Zeithammer (2006). However, in our case this is a repeated auction for a per-period prize while his paper considers dynamic bidding for a single item.

Finally, our paper is also related to recent empirical work on search advertising (Rutz and Bucklin 2007a, b). This research broadly studies the effectiveness of paid placements with particular attention devoted to spillover and lagged effects as well as contexts when multiple search words are used. Our model extensions are largely motivated by these papers (see our dynamic model and the discussion on multiple search words) although the present paper focuses on the pricing of search advertising and advertisers' bidding behaviors.

The rest of the section is organized as follows. Next, we describe the basic model

51

description in Section 3.1 and equilibrium analysis in Section 3.2. Section 3.3 studies dynamic advertising purchases by firms, whereas Section 3.4 generalizes the model to incorporate multiple content areas.

3.1 The Model

We assume n websites that are indexed with respect to their exogenously given, inherent click-through rates (CTRs), $\gamma_1 > \gamma_2 > ... > \gamma_n$. These rates represent the value of the sites in the eyes of the consumers or can be thought of as their popularity *in the context of a search word*. The $(n+1)$th player is a search engine (SE), a special website[21]. The SE ranks the sites according to their popularity in a given search context, that is, the click-trough rates determine the ranking. Thus, its basic service lies in finding sites, that consumers are most interested in (click-through-rates are known by sites and the SE but not by consumers). The search engine returns the r highest ranked sites as the search result. Next to the regular results, the SE also displays an s number of "sponsored links". The order of these links can be chosen by the SE and this choice is based on the bids submitted by websites. Let $l_1, l_2, ..., l_s$ denote the sites winning the sponsored links, in order of appearance. Thus, the output of the SE is modeled as a page with two lists: a search list and a sponsored ad list. Google's search page is exactly like this (see Figures 8 and 9) and other search engines have a similar format.

3.1.1 Consumers' behavior on the search page

We assume, that the SE attracts a unit traffic of consumers which is distributed in the following way. When a consumer sees the SE's page generated by the search, s/he either clicks on one of the regular results, one of the sponsored links or leaves

---

[21]We assume that the SE is a monopolist. While this is not entirely true in practice, Google dominates the search industry with over 56% of all searches, a proportion that is growing.

the page without clicking. We assume that consumers' clicking behavior is affected by the following four factors.

1. The order in which the sites are listed on both types of lists.

2. Differences in click probabilities between the sponsored list and the search result list.

3. Individual differences between sites in inherent click-though rates or popularity.

4. Whether the site appears in both search and sponsored lists or only one of the lists.

For the first factor, assume that $\alpha_1, \alpha_2, ... > 0$ denote the psychological order constants that determine how the possible clicks are distributed through an ordered list of items. That is, whenever someone sees an ordered list of *equally interesting* items s/he chooses the ith item with probability proportional to α_i. Similarly, for the second factor, let $\beta < 1$ denote the ratio of consumers who click on a sponsored link rather than an equally interesting link on the organic search list. Combining the two factors, the distribution of consumers among the links, not taking into account individual differences, is determined by the parameters: $\alpha_1, \alpha_2, ..., \alpha_r$ and $\beta\alpha_1, \beta\alpha_2, ..., \beta\alpha_s$. Since the search engine has a unit traffic, we have to normalize

$$\sum_{i=1}^{r} \alpha_i + \beta \sum_{i=1}^{s} \alpha_i$$

to 1.

For the third factor, that takes individual differences into account, we can multiply these parameters with the inherent click-through rates of the sites. In any particular position, a site with a higher CTR is more likely to attract a click than another site in the same position having a lower CTR. Finally, for the fourth factor, we assume that if a site is listed both among the regular search results and the

53

Figure 10: Probabilities of a user clicking on the links.

sponsored links, the latter will have a lower click-probability than if the site were listed *only* on the sponsored links list. Specifically, let δ denote the strength of this effect, that is, the proportion of people who do not click on a sponsored link if it is also displayed among the regular results but click on the regular link instead. Note that the parameter δ does not have an effect on the total traffic that a site gets from the search engine because it simply changes the origin of this traffic. However, it determines the traffic coming from a sponsored link, which will be important in the sites' bidding process and will also affect the SE's revenue. Figure 10 illustrates the probabilities that a user clicks on a specific link.

Given these factors, we now determine how the traffic of the search engine is distributed through the websites. Let $A(i)$ denote the function that takes a value of 0 if Site i does not win a sponsored link, that is, $i \notin \{l_1, l_2, ..., l_s\}$ and α_j if Site i wins the jth sponsored link. With these, the total traffic that Site $i \leq r$ gets from

the search engine is proportional to:

$$t_i = t_i^R + t_i^S = \gamma_i \alpha_i + \gamma_i \beta A(i).$$

For $i > r$, the traffic is

$$t_i = t_i^S = \gamma_i \beta A(i).$$

3.1.2 Websites

Websites make profits from the traffic that arrives to their sites from the search engine[22]. Let us assume, that there is a common $R(t)$ function for all sites that determines the revenue associated with t amount of traffic. We naturally assume, that $R(t)$ is increasing and concave[23]. In order to obtain sponsored links, sites have to submit bids to the search engine. The bid that Site i submits, b_i is the amount that it is willing to pay for unit traffic (per-click). If the search engine decides to include Site i among the sponsored links, Site i has to pay an advertising fee of $p_i t_i^S$, where $p_i \leq b_i$ is set by the search engine. Therefore, Site i's utility is

$$u_i = R(t_i^R + t_i^S) - p_i t_i^S$$

if it wins a sponsored link and $u_i = R(t_i^R)$ otherwise, where t_i^S depends on which sites win the sponsored links.

At this point, the SE is completely free to determine the order of winners and advertising fee it charges for a click, $p_i \leq b_i$. First, we will show that in a one-period game the SE sets $p_i = b_i$ corresponding to a first price auction, then we will discuss the different types of auctions that search engines use in practice. Based on this

---

[22]Thus, we ignore the fact, that sites could already have different amounts of incoming traffic from other sources. If we naturally assume that sites with a higher click-through rate also have higher outside traffic, then the results still hold.

[23]See Rutz and Bucklin (2007b) for a detailed analysis on how $R(t)$ could be estimated in practice.

discussion, in Section 3.1.3, we will restrict the SE's strategies and define the types of equilibria we use in subsequent analysis.

The timing of the game is the following. First, websites simultaneously submit the b_i bids, knowing all the click-through rates and $R()$. Then, the search engine decides which sites it will include among the sponsored links and in what order. Finally, sites pay the advertising fee to the search engine and realize profits from the traffic they receive.

3.1.3   The Search Engine's Best Response

First, we determine the SE's best response to given bids $b_1, b_2, ..., b_n$ in the second stage of the game. Although it would seem so, the best strategy is not to simply assign the sponsored links to websites in the order of their bids. The SE has to consider the sites' click-through rates, since the total traffic it sells to them and thus its revenue depends on these rates. Therefore, a site with a high click-through rate may pay a higher total fee even if its bid is low. An opposite effect is that sites with the highest inherent click-through rates will also likely appear on the regular search list. As a result, they will attain fewer clicks on the sponsored link because a δ proportion of the consumers will click on the regular search results link instead. Formally, the SE maximizes its profit,

$$\Pi_{SE} = \sum_{i=1}^{s} t_i^S p_i.$$

The following claim summarizes the SE's best response to the bids. Let $I(i)$ denote the function that takes the value 1 if $i \leq r$ and 0 otherwise. The SE's decision can be described by the series $w_1, w_2, ..., w_n$, where Site w_i will get sponsored link i. Sites $w_{s+1}, w_{s+2}, ..., w_n$ will not get a sponsored link.

Claim 1 *In equilibrium,*

$$\gamma_{w_i} b_{w_i}(1 - \delta I(w_i)) \geq \gamma_{w_j} b_{w_j}(1 - \delta I(w_j))$$

holds for $i < j$, where $i \leq s$ and the SE sets $p_i = b_i$.

In other words, the search engine ranks the sites according to their $\gamma_i b_i (1 - \delta I(i))$ and charges each site's bid. That is, for sites that are not in the top r among the search results, their position among the sponsored links is determined by their inherent CTR multiplied by their bid. For top sites this value is multiplied by $(1 - \delta)$, accounting for consumers who choose to click on the results link instead of the sponsored link.

As a result of Claim 1, in a non-repeated game, the search engine's best strategy is to charge the highest bid (corrected with the CTR). This corresponds to a first price auction. However, in reality search engines use second price auctions (some of them correcting for differences in CTRs, some of them not) to avoid the problem that when multiple items with different values are auctioned away then the first price auction typically does not have an equilibrium. This is because bids in a first price auction always converge towards each other, which makes it impossible to reflect the differences in valuations for the different items[24]. Thus, it is important to discuss the different types of auctions and equilibria that can be used in our models.

In our analysis, we assume that websites have full information about each others' bids, valuations and click-through rates. This is consistent with reality: quite well known valuations across sites are typical characteristics of auctions of sponsored links. When competitors' valuations are known, a first price auction for a single item typically has an infinity of equilibria. For example, let $v_1 > v_2 > ... > v_n$ be the valuations of n bidders for a single item. If a first price auction is applied then the

---

[24] Another reason to use a second price auction is that, if valuations are uncertain, then the second price auction is a mechanism that leads to truth-telling in a single-item auction.

winner pays its bid. In equilibrium, the winner is always player 1 and the winning bid, b_1 can take any value in the $(v_2, v_1]$ interval. Thus, the auctioneer's revenue is between v_2 and v_1. We denote these equilibria by FNE (first price Nash equilibrium).

In the case of a second price single-item auction, anyone can win the auction in equilibrium (SNE). If every player bids zero except player i, who bids $v_0 > v_1$, then the winner is player i, who has to pay nothing. In general, the second highest bid is always below v_1, so the auctioneer's revenue is somewhere between 0 and v_1. To restrict the possible outcomes of a second price auction, Varian (2007) introduced the notion of symmetric equilibria for multi-item second price auctions, (SSNE), which is a subset of the pure-strategy Nash-equilibria. In such an equilibrium, the player in position k is better off paying the bid of the player in position $k + 1$, then would be in position l paying the bid of player $l + 1$. This is a stronger restriction than in an SNE for moving up in the ranking because in an SNE a player is only supposed to be better off paying bid $k + 1$ for position k than paying bid l for position l. Since bid l is higher than bid $l + 1$, an SSNE is always an SNE but the opposite is not true. According to Varian (2007), in an SSNE, the order of winners is always $1, 2, 3, ...$, that is, in case of a single item the winner is always player 1. Furthermore, the auctioneer's maximum SSNE revenue is the same as the maximum SNE revenue and is equal to v_1 in case of a single item. Since the equilibria in a first price single-item auction (FNE) and symmetric equilibria in a second price single-item auction (SSNE) give the same results for the bid orders and maximum revenues of the seller, we can use the two concepts interchangeably for our analysis, if there is only one sponsored link. For multiple links, the FNE usually does not exist, so in this case, we will always use the SSNE as the equilibrium concept.

We always correct for click-through rates as it is established in Claim 1. Player i's bid is multiplied by $\gamma_i(1 - \delta I(i))$ and the search engine ranks the

$$F_i = \gamma_i b_i (1 - \delta I(i)),$$

values when determining the order of sites and the prices. In a first price auction, Site i has to pay $p_i = b_i$ for a click, corresponding to a total fee of $\beta A(i) F_i$, where $A(i)$ reflects its position. In a second price auction, if Site i is followed by Site j in the order then Site i has to pay

$$p_i = \frac{F_j}{\gamma_i(1 - \delta I(i))} = b_j \frac{\gamma_j(1 - \delta I(j))}{\gamma_i(1 - \delta I(i))}$$

for a click, totaling to a fee of $\beta A(i) F_j$. The next section determines the equilibrium bids.

3.2 Equilibrium analysis

3.2.1 Bidding strategies for one sponsored link

To illustrate the primary forces that work in the game, we first consider the case in which, there is only one sponsored link offered, that is, $s = 1$. Let

$$G(i) = R(I(i)\gamma_i \alpha_i + \gamma_i \beta \alpha_1) - R(I(i)\gamma_i \alpha_i)$$

denote the revenue gain for Site i of winning the sponsored link. Clearly, the total fee Site i will pay for the sponsored link cannot exceed $G(i)$. Let $w_1, w_2, ..., w_n$ be a permutation of sites such that $G(w_1) > G(w_2) \geq ... \geq G(w_n)$ holds[25]. Furthermore, let P_1 denote the total fee that the winner pays for the sponsored link,[26] which is equal to the seller's revenue.

---

[25] The assumption that there is a single highest value eases the presentation of results, but does not change them qualitatively.

[26] In case of a first price auction, this is calculated from its own bid. In case of a second price auction, it is calculated from the second highest bid, corrected for CTRs.

Proposition 6 *In any FNE and SSNE, the winner of the sponsored link is Site w_1 and the total fee it pays is $G(w_1) \geq P_1 \geq G(w_2)$.*

Given the assumption that $R()$ is increasing and concave, the winner can be any site from 1 to $r + 1$, depending on the parameters. For example, if $R()$ were linear then the site with the highest $\gamma_i \beta \alpha_1$, that is, Site 1 would be the winner. However, if the γ_i's are not too far from each other, that is $\gamma_1 - \gamma_{r+1} \to 0$, then the winner is Site $r + 1$. These two cases illustrate the two forces that work against each other in determining the outcome. On one hand, since $R()$ is concave, sites who already receive traffic from the search engine through regular results have a lower benefit from winning the link[27]. On the other hand, sites with a higher γ_i obtain more traffic from a sponsored link, therefore, they are willing to pay more for such a link. If the latter effect is stronger, then a top site wins, otherwise a regularly lower ranked site wins the sponsored link. In reality, these two cases translate to the distinct, observed scenarios illustrated by Figures 8 and 9. On the first figure, the sponsored links and search result are distinct. In contrast, on the second, a site appearing among the top search results also obtains a (top) sponsored link.

The following corollary describes the equilibrium bids.

Corollary 1 *The winning bid in an FNE is*

$$\frac{G(w_1)}{\beta \alpha_1 \gamma_{w_1}(1 - \delta I(w_1))} \geq b_1 > \frac{G(w_2)}{\beta \alpha_1 \gamma_{w_1}(1 - \delta I(w_1))}.$$

In an SSNE, the winning bid can be arbitrarily high, but the second highest bid is

$$\frac{G(w_1)}{\beta \alpha_2 \gamma_{w_2}(1 - \delta I(w_2))} \geq b_2 > \frac{G(w_2)}{\beta \alpha_2 \gamma_{w_2}(1 - \delta I(w_2))}.$$

---

[27]This force is even stronger if we assume that sites with a high CTR have a larger traffic independent from the SE.

Note that the bids largely depend on the parameters. Sites with similar valuations might submit significantly different bids based on their CTR's or their position among the regular search results.

3.2.2 Bidding strategies for multiple sponsored links

We will now discuss the general case, with multiple sponsored links ($s > 1$). As mentioned before, the first price auction does not work in this case, thus we analyze the SSNE only. Let

$$G_j(i) = R(I(i)\gamma_i\alpha_i + \gamma_i\beta\alpha_j)) - R(I(i)\gamma_i\alpha_i)$$

denote the revenue gain for Site i of winning the sponsored link j ($j = 1, ..., s$). Let $w_1, w_2, ..., w_n$ denote the sites in the order of their CTR-corrected bids (F_i's). Furthermore, let P_i denote the total fee that Site i pays for the advertising:

$$P_i = b_{w_{i+1}}\alpha_{w_i}\beta\gamma_{w_i}(1 - \delta I(w_i)).$$

The search engine ranks the sites according to their CTR-corrected bids, that is, if the order is $w_1, w_2, ...$, then the following have to hold for every $2 \geq i \geq s$:

$$\frac{P_{i-1}}{\alpha_{i-1}} > \frac{P_i}{\alpha_i}. \tag{12}$$

In any equilibrium, Site w_k does not have an incentive to bid less and get to a lower position. Therefore,

$$G_k(w_k) - P_k \geq G_l(w_k) - P_l. \tag{13}$$

Furthermore, according to the definition of a symmetric equilibrium, Site w_l does not want to get into position k even if it has to pay P_k (and not P_{k-1}). That is,

$$G_l(w_l) - P_l \geq G_k(w_l) - P_k. \tag{14}$$

Combining (12), (13) and (14), we get the following inequalities, describing the equilibria of the auction:

$$G_k(w_k) - G_l(w_k) \geq P_k - P_l \geq G_k(w_l) - G_l(w_l). \tag{15}$$

The complexity of the problem does not allow us to characterize all the SSNEs. Multiple equilibria may exist, where the order of winners is different. The following example illustrates the complexity of the problem even in a simple case.

Example 1 Assume $s = 2$ and $n = 3$, with the following valuations:

$$G_1(1) = 10, \ \ G_2(1) = 8, \ \ G_1(2) = 9, \ \ G_2(6) = 6, \ \ G_1(3) = 8, \ \ G_2(3) = 7.$$

These gains can be derived from a suitable $R()$ function, γ-s and α-s. Note that with prices $P_1 = 9$ and $P_2 = 7$, the equilibrium order of sites can be either $(w_1 = 1, w_2 = 3, w_3 = 2)$ or $(w_1 = 2, w_2 = 1, w_3 = 3)$.

To solve for the maximum and minimum revenue equilibria in the general problem, we would have to solve the linear program defined by (12) and (15) for every i, k and l. While this problem is still very complex, with a minor restriction, we can easily solve it.

Definition 5 *We say that the preferences of sites i and j are aligned, if $G_1(i) > G_1(j)$ implies $G_k(i) - G_l(i) > G_k(j) - G_l(j)$ for every $1 \leq k, l \leq s + 1$.*

The assumption of aligned preferences is rather natural. It says that there is a consensus between players about the value of different positions. With this, we can determine the equilibrium ranking of sites.

Lemma 3 *In any SSNE, $G_k(w_1) \geq G_k(w_2) \geq ... \geq G_k(w_{s+1})$ for any $1 \geq k \geq s + 1$.*

In order to fully describe the equilibria we also have to assume that sites' valuation for the position they are in is high enough relative to the next site' valuation of the next position. Specifically, we assume that

$$G_j(w_j) - G_{j+1}(w_j) > \frac{\alpha_j - \alpha_{j+1}}{\alpha_{j+1} - \alpha_{j+2}}(G_{j+1}(w_{j+1}) - G_{j+2}(w_{j+1})) \qquad (16)$$

holds for every $1 \geq j \geq s-1$ (see the Appendix for more details on this assumption). With these assumptions, we can describe the SSNE, following the path proposed by Varian (2007).

Proposition 7 *If all the sites' preferences are aligned and (16) holds, then an SSNE exists. Furthermore,*

1. *The maximum SSNE income of the seller is*

$$M(s) = \sum_{j=1}^{s-1} [j(G_j(w_j) - G_{j+1}(w_j))] + sG_s(w_s).$$

2. *The maximum SSNE income is equal to the maximum SNE income.*

The results are similar to the case in which there is only one sponsored link to bid for. The set and order of winners is determined by two factors. Sites with higher traffic from other sources, such as regular search results, have a lower marginal valuation for traffic, however sites with higher CTRs value sponsored links higher. It is clear that the order among those sites that do not receive regular search results will be decreasing in the CTR, that is, $r + 1, r + 2, ..., n$. However, the top r sites may end up in any position depending on their parameters. Also, the misalignment of preferences is only possible among the top r sites, that is, in some cases multiple equilibria may exist and they may switch positions.

Figure 11 shows the possible valuations of twenty sites for five sponsored links. The parameters are such that sites 11 and 12 have the highest valuations for the

sponsored links because they are the sites with the highest click-through rates that are *not* listed among the regular search results. Since the advertising response function is concave, these sites have a higher marginal valuation for a click. As a result, the winner of the first sponsored link is Site 11, followed by sites 12, 3, 4, and 2. Figure 12 shows the equilibrium prices the sites pay and the bids they submit. Here, the sites are listed in their order of appearance. It is not surprising, that the total fee they pay is decreasing with the position they are in. However, it is interesting to see that higher per-click bids do not automatically lead to a better position. Generally, sites with higher inherent CTRs do not need to bid too high, however, top sites (such as 3, 4 and 2) still have to bid higher than others for the same position because their higher CTRs guarantees them a position on the SE's search results list, which in turn directs traffic away from the sponsored link.

3.2.3 The number of sponsored links

So far we have considered the number of sponsored links displayed by the search engine given. In this section, we compare the search engine's revenue in case of offering different numbers of links. For the sake of simplicity we assume a linear revenue function, that is, $R(t) = at$. Then $G_k(i) = \beta\gamma_i\alpha_k$. We assume that the search engine makes a decision about the number of sponsored links and announces it prior to the auction. When it makes the decision it has to take into consideration two forces. First, if it offers more links for sale, it will receive fees from more sites. However, when the number of links is increased, the traffic flowing through one goes down. Let us compare the cases when the search engine offers s sponsored links and when it offers $t < s$ instead. If $\beta\alpha_j$ is the traffic going to sponsored link j in the first case, then it increases to

$$\beta\alpha_j' = \beta\alpha_j(1 + \beta\alpha_{t+1} + ... + \beta\alpha_s).$$
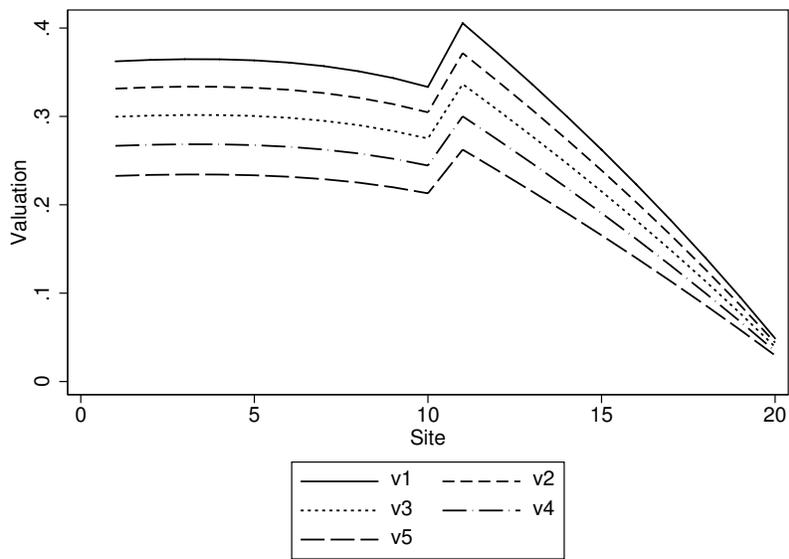
Figure 11: Sites' valuation of the five sponsored links. The parameters are: $n = 20$, $r = 10$, $s = 5$, $\gamma_i = 0.5 - 0.025(i - 1)$, $\alpha_i = (20 - (i - 1))/232.5$, $\beta = 0.5$, $\delta = 0.6$, and $R(x) = \log(1 + 30x)$.

65
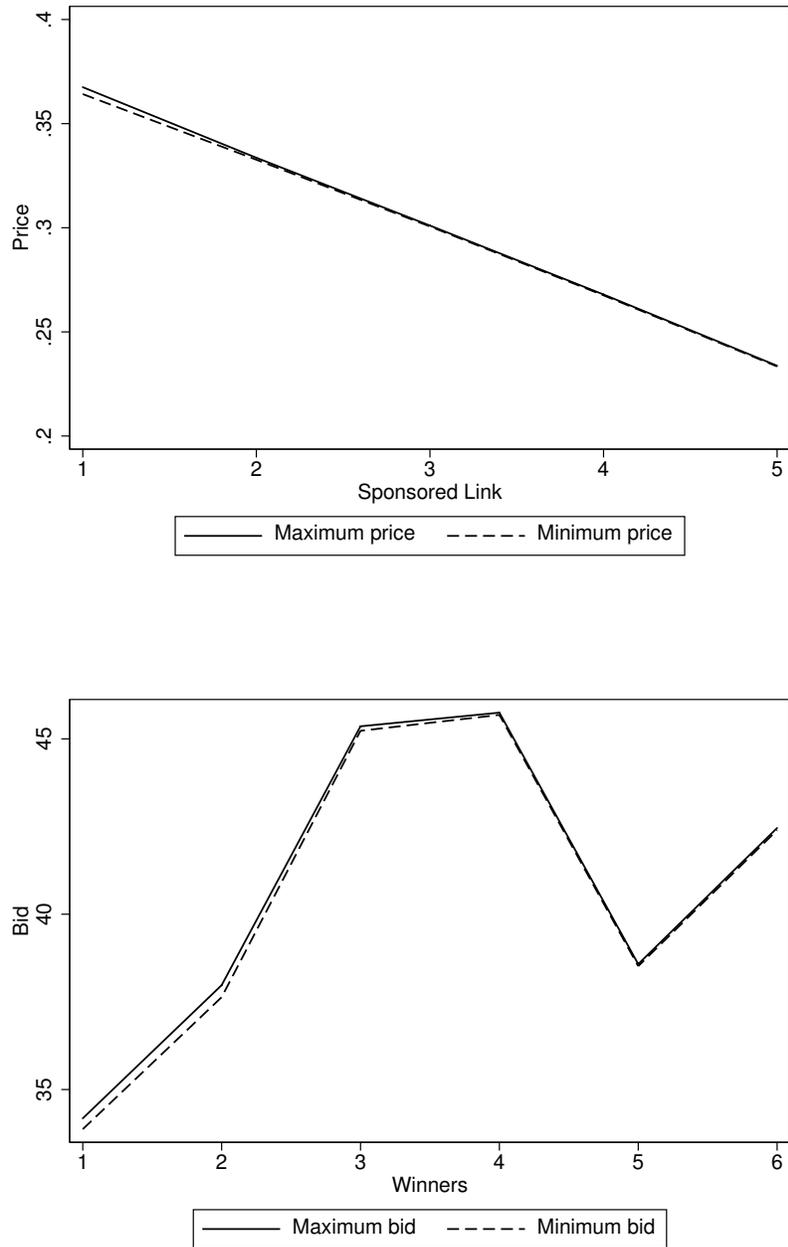
Figure 12: Prices and bids of the five winners in their order of appearance: (11,12,3,4,2). The parameters are: $n = 20$, $r = 10$, $s = 5$, $\gamma_i = 0.5 - 0.025(i-1)$, $\alpha_i = (20 - (i-1))/232.5$, $\beta = 0.5$, $\delta = 0.6$, and $R(x) = \log(1 + 30x)$.

As we saw in the previous section, there are usually many equilibria and the revenue of the SE cannot be determined. Here, we will only compare the maximum revenues the SE can attain by selling different number of sponsored links.

Proposition 8 *The SE can attain a higher maximum revenue by offering $t < s$ sponsored links instead of s, if and only if,*

$$\beta(\alpha_{t+1} + ... + \alpha_s)\left(\sum_{j=1}^{t} j\gamma_j\alpha_j - \sum_{j=1}^{t-1} j\gamma_j\alpha_{j+1}\right) > \sum_{j=t+1}^{s} j\gamma_j\alpha_j - \sum_{j=t}^{s-1} j\gamma_j\alpha_{j+1}.$$

Decreasing the number of sponsored links increases the traffic on the remaining ones. Thus, the sites are willing to pay more for them. The LHS of the inequality is equal to this benefit. However, by forgoing sponsored links $t + 1$ to s, the SE loses $s - t$ advertisers. The resulting loss is the RHS of the inequality. Note that the RHS is sometimes negative, that is, even without the increased traffic on the remaining links the SE may have an incentive to decrease the number of links.

Example 2 Assume that $s = 2$ and $t = 1$. The SE is better off offering one link, iff,

$$\beta\alpha_1 > \frac{2\gamma_2 - \gamma_1}{\gamma_1}.$$

In essence, the SE will offer only one sponsored link, if the second highest CTR is relatively low. In particular, if $\gamma_2 < \gamma_1/2$, then the SE is better off selling one link even if the second link still drains the traffic.

3.3   Repeated bidding for sponsored links

In the previous models, we assumed that the process through which the sponsored links are assigned is a one-shot game. However, the auctions for the links take place repeatedly. We cannot always ignore the effects that previous bids and results

have on the current auction. An important effect is, for example, that when a site wins a sponsored link, the traffic that it receives through the link may have a lagged effect. Such lagged effects have been documented in Rutz and Bucklin (2007a). Some consumers who get to a website through advertising may become regular customers of the site. If they want to return to the site they do not need the sponsored link again, they may remember or "bookmark" the site's address. This effect however, decreases with time. For the sake of simplicity we assume that it lasts only for one time period and that there is only one sponsored link. Precisely, if a consumer arrives from the SE to the site in a given time period, then with probability q s/he will return in the next period without the help of the search engine. Then, if Site i receives traffic t_i from the search engine in a given period, then the lagged effect of this traffic is qt_i in the next period.

Now let us examine how this effect changes site's valuations of the sponsored link. If a site did not win the sponsored link in the previous period then the gain associated with winning it is

$$G_l(i) = R((1+q)I(i)\gamma_i\alpha_i + \gamma_i\beta\alpha_1)) - R((1+q)I(i)\gamma_i\alpha_i),$$

where we also deal with the lagged effect of regular search results. On the other hand, if the site did win the sponsored link in the previous period, then its gain is

$$G_w(i) = R((1+q)I(i)\gamma_i\alpha_i + (1+q)\gamma_i\beta\alpha_1)) - R((1+q)I(i)\gamma_i\alpha_i + q\gamma_i\beta\alpha_1).$$

Therefore, if $R()$ is strictly concave, then $G_w(i) < G_l(i)$. The intuition is that due to decreasing marginal returns, the site values the sponsored link less, if it has already won the link in the previous period. To solve the repeated game we use the concept of Markov-perfect equilibrium, where players' actions only depend on the states of the world. In this case, the states represent the possible winners of the auction and when a site wins the auction, the world moves to that state. In such an equilibrium,

forward looking players choose their strategies to maximize their profits over time using the discount factor δ. Let $V_i^{(j)}$ denote Site i's discounted equilibrium profits counted from a period, when the previous winner is Site j. Sites' payoffs in the current period will be determined by their bids. If Site i does not win the auction, it does not make any profit in the current period, that is, its overall discounted profit will be

$$\delta V_i^{(w)},$$

where w is the winner of the current auction. On the other hand, if Site i wins the auction then it will make a profit of $v_i = G_w(i) - P$ if $i = j$ and $v_i = G_l(i) - P'$ if $i \neq j$, where P and P' are the prices the winner has to pay (these depend on the bids). Therefore, its overall discounted profit will be

$$v_i + \delta V_i^{(i)}.$$

In equilibrium, player i chooses its bid to maximize this quantity, that is,

$$V_i^{(j)} = \max_{b_i}(\delta V_i^{(w)}, v_i + \delta V_i^{(i)}),$$

where w and v_i both depend on b_i.

Since there is only one sponsored link, we can use the first price auction's equilibrium and the second price auction's symmetric equilibrium concepts interchangeably. We will determine the Markov-perfect first price Nash-equilibria (MFNE) and Markov-perfect second price symmetric Nash-equilibria (MSSNE) of the game. Regarding the valuations, let us assume that only the first two sites have a high enough valuation to win the auction, that is $G_l(j) < \min(G_w(1), G_w(2))$ for $j \geq 3$. Then, we only have the examine the auction where Sites 1 and 2 bid for the link. Without loss of generality we can assume that $G_l(1) > G_l(2)$, that is, site 1 has a higher valuation for the sponsored link, not taking into account the lagged traffic.

Proposition 9

1. *If $G_l(2) < G_w(1)$, then Site 1 is the winner in every period and*

$$G_w(1) \geq P_1 \geq G_l(2),$$

that is, the seller's maximum discounted income is

$$M_1 = \frac{G_w(1)}{1 - \delta}.$$

2. *If $G_l(2) > G_w(1)$, then the two sites alternate winning, and the seller's maximum discounted income is*

$$M_2 = \frac{G_l(1) + \delta G_l(2)}{1 - \delta^2}.$$

In essence, if Site 1 values winning the link for a second time higher than Site 2 does for the first time, then Site 1 is the winner always. Otherwise, the two sites alternately win and lose the auction. The intuition is that when Site 1 wins the link in one period, then its valuation goes down in the next period and Site 2 is willing to pay more for the link. Now that Site 2 wins the auction, the valuations will again cross each other leading to the alternation. Next, we examine how the value of q affects the type of equilibrium.

Corollary 2 *There is a $q^* > 0$, such that*

- *if $0 < q < q^*$ then the winner is always Site 1,*

- *if $q^* < q$ then the two sites alternate as winners.*

In other words, as the ratio of returning customers increases, at one point the type of equilibrium changes and the two sites start winning alternately. This critical value is smaller if the marginal returns on traffic decrease quickly.

Next, let us compare the search engine's income in the two cases. It is worth noting, that M_1 and M_2 not only represent the SE's maximum income in the two cases, but also the total surplus of all players (SE and sites) in all the equilibria of the given type[28]. We compare these values around the boundary of the two regions, which separates the alternating and non-alternating equilibria, that is, where $G_w(1) = G_l(2)$.

Corollary 3

$$\lim_{G_w(1)-G_l(2)\to 0^+} M_1 < \lim_{G_w(1)-G_l(2)\to 0^-} M_2,$$

and the difference increases in q and δ.

We find a discontinuity in the total income at the boundary of the two regions, because the SE and the sites are strictly better off in the case of an alternating equilibrium. The intuition is that the alternating assignment of the SE's traffic is a more efficient allocation than when one site is the winner in every period. This extra revenue is higher if the ratio of returning consumers is higher and if the discount rate is higher. Whether the SE or the sites appropriate this extra revenue depends on the actual bids.

3.4 Multiple keywords

So far we assumed that every consumer is interested in the same topic and the results include the same pages for every query. Obviously, this is rather unrealistic, thus, in this section, we will relax this assumption. Our objective is to explore how the

---

[28]Individual incomes depend on how this surplus is divided in a given equilibrium.

allocation of sponsored links in relation to a given search word changes when multiple *interacting* search words are considered. As reported in Rutz and Bucklin (2007b), most advertisers manage/bid for a bundle of key words.

Let us assume that there are d different topics that consumers are interested in and $(p_1, p_2, ..., p_d)$ denotes the distribution of these interests, i.e. p_k is the proportion of consumers who search in topic k or the proportion of searches that consumers initiate in topic k.

Web sites may offer content in every topic, although their relevance may vary from topic to topic. In other words, the inherent click-trough rates may be different for the same site in different topics. For example travelocity.com may have a high click-through rate in the context of travel but most likely has a lower one when consumers are searching for home appliances. Accordingly, let γ_i^k denote the inherent click-through rate of Site i in topic k. Note that the order of sites with respect to their click-trough rates may be different for different topics. We index sites with respect to their click-through rates in topic 1, that is $\gamma_1^1 > \gamma_2^1 > ... > \gamma_n^1$. For convenience, we also assume that two sites cannot have identical click-trough rates in the same topic. We introduce the permutations $\sigma_k()$, that ranks sites with respect to their click-trough rates in topic k, that is, $\gamma_{\sigma_k(1)}^k > \gamma_{\sigma_k(2)}^k > ... > \gamma_{\sigma_k(n)}^k$, where obviously, $\sigma_1(i) = i$. Websites' revenues are now a function of the traffic they receive in the d topics, that is, Site i benefits $R(t_i^1, t_i^2, ..., t_i^d)$ from the traffic it receives from the SE.

In every topic the search engine lists the top r sites, that is, the ones with the highest click-through rates. Next to the regular search results, it lists the sponsored links as a result of the same bidding process described before. The sponsored links for every search word are allocated after a separate bidding process. First, websites simultaneously submit their bids in every topic, then the SE allocates the sponsored

links for every word. Then the traffic Site i receives in topic k will be

$$t_i^k = p_k \gamma_i^k (I^k(i)\alpha_i + \beta A^k(i)),$$

where $I^k(i)$ and $A^k(i)$ are the analogues of $I(i)$ and $A(i)$.

First, we examine the case when $R()$ is additive over topics, that is,

$$R(t^1, t^2, ..., t^d) = R^1(t^1) + R^2(t^2) + ... + R^d(t^d). \tag{17}$$

Under this assumption, the results reduce to the single topic equilibria studied before:

Proposition 10 *If $R()$ satisfies (17), then equilibrium strategies in topic k are identical to those in Proposition 7, with $R() = R^k()$.*

The result states that the bids in different topics are independent and are the same as in the single-topic model. This shows that Proposition 7 can be applied to determine the equilibria in this general model if the return on traffic is independent across search words.

However, the focus of our investigation is the interaction between the bids in different topics. Therefore, from this point, we assume a specific form of revenue function,

$$R(t^1, t^2, ..., t^d) = R(t^1 + t^2 + ... + t^d),$$

that is, traffic received in different topics is equivalent and the revenue of a website only depends on the total traffic it receives. To exclude the additive case, we assume that $R()$ is increasing and strictly concave. The complexity of the problem does not allow us to describe all possible equilibria. We focus on the simple case of $d = 2$, to capture the interaction between the auctions in different content areas. Let $G_S(i)$ denote Site i's gain form winning the sponsored links for the keywords in set S, where

73

$S = \{1\}, \{2\}, \{1, 2\}$. For example,

$$G_{\{1,2\}}(i) = R(I^1(i)p_1\gamma_i^1\alpha_i + I^2(i)p_2\gamma_i^2\alpha_{\sigma^{-1}(i)} + (p_1\gamma_i^1 + p_2\gamma_i^2)\beta\alpha_1)) \qquad (18)$$
$$-R(I^1(i)p_1\gamma_i^1\alpha_i + I^2(i)p_2\gamma_i^2\alpha_{\sigma^{-1}(i)}).$$

Since $R()$ is concave, G is subadditive on the sets, that is, $G_{\{1,2\}} < G_{\{1\}} + G_{\{2\}}$. In order to determine the winners in equilibrium, we have to know which site has the highest valuation for a given keyword. Let $x_1, x_2, ...$ be a permutation of the sites according to their valuation of the first keyword and $y_1, y_2, ...$ be one according to their valuation of the second keyword. That is, $G_{\{1\}}(x_1) > G_{\{1\}}(x_2) > ...$ and $G_{\{2\}}(y_1) > G_{\{2\}}(y_2) > ...$ Without loss of generality, we will assume that

$$G_{\{2\}}(y_1) + G_{\{1\}}(x_2) > G_{\{1\}}(x_1) + G_{\{2\}}(y_2). \qquad (19)$$

Let P^k be the total fee that the winner of the auction pays for keyword k (denoted by w^k). Then, the SE's income will be $P^1 + P^2$. The following proposition describes the outcome under this scenario.

Proposition 11 *Both an SSNE and an FNE exist and they all satisfy the following conditions in which the inequalities are binding.*

1. *If $x_1 \neq y_1$, then $w^1 = x_1$ and $w^2 = y_1$. The seller's revenue is*

   $$G_{\{1\}}(x_1) + G_{\{2\}}(y_1) \geq P^1 + P^2 \geq \max(G_{\{1\}}(x_2) + G_{\{2\}}(y_2), G_{\{1,2\}}(x_1), G_{\{1,2\}}(y_1)).$$

2. *If $x_1 = y_1$ and $G_{\{1,2\}}(x_1) \geq G_{\{2\}}(y_1) + G_{\{1\}}(x_2)$, then $w^1 = w^2 = x_1$. The seller's revenue is*

   $$G_{\{1,2\}}(x_1) \geq P^1 + P^2 \geq G_{\{1\}}(x_2) + G_{\{2\}}(y_2).$$

74

3. If $x_1 = y_1$, $x_2 = y_2$, and $G_{\{1,2\}}(x_1) < G_{\{2\}}(y_1) + G_{\{1\}}(x_2)$, then $w^1 = x_2$ and $w^2 = y_1$. The seller's revenue is

$$G_{\{2\}}(y_1) + 2G_{\{1\}}(x_2) - G_{\{1\}}(x_1) \geq P^1 + P^2.$$

4. If $x_1 = y_1$, $x_2 \neq y_2$, and $G_{\{1,2\}}(x_1) < G_{\{2\}}(y_1) + G_{\{1\}}(x_2)$, then two types of outcomes may exist.

   (a) As in the previous case, $w^1 = x_2$ and $w^2 = y_1$ and the seller's revenue is

   $$G_{\{2\}}(y_1) + 2G_{\{1\}}(x_2) - G_{\{1\}}(x_1) \geq P^1 + P^2.$$

   (b) If $G_{\{2\}}(x_1) + G_{\{1\}}(y_2) \leq G_{\{1\}}(x_1) + G_{\{2\}}(y_2)$, and $G_{\{1,2\}}(x_1) < G_{\{1\}}(x_1) + G_{\{2\}}(y_2)$ then in another type of equilibrium, $w^1 = x_1$ and $w^2 = y_2$ and the seller's revenue is

   $$G_{\{1\}}(x_1) + 2G_{\{2\}}(y_2) - G_{\{2\}}(y_1) \geq P^1 + P^2.$$

In all the cases, except (4b), the allocation is efficient in the sense that the winners of the auction are those sites that have the maximum total valuation for the words that they win the auction for (in the second case this is a single site). In (4b), the winner of word 1 is the site that has the highest valuation for it, and the winner of word 2 is the site with the second highest valuation for it, in spite of the fact that the first site for word 2 and the second site for word 1 have a higher total valuation.

Examining the relationship between the two words, we can see that if the words are unrelated, that is, if different sites have the highest valuation for the two words, then these sites with highest valuation win the auction (case 1). On the other hand, if the words are related and presumably the same site has the highest valuation for both words, then either that site wins both auctions (if its total valuation is high enough - case 2) or it wins only one of them (cases 3 and 4). In these latter cases, the

intuition is that winning one auction boosts the winner's traffic, therefore, it does not value the traffic in the other auction that high, leaving the opportunity to the site with the second highest valuation to win there. In the extreme case, if there are nearly as many related words as bidders, then even the site with the lowest valuation can end up winning a sponsored link.

3.5 Conclusion

In this paper, we have modeled the race for sponsored advertising links on a SE's page between websites endowed with different click-through rates. We argue that the SE's problem can not simply be described as a multi-item auction. The existence of the search results list on the SE's page represents an important externality for both types of players. In addition to exploring the effect of this externality on the allocation outcomes we also study a variety of other issues: the endogenous choice of the number of sponsored links, bidding for links in multiple search words and the dynamics of the bidding behavior.

Our key result is that we explain the machanism that may lead to wildly different patterns observed in the behavior of sponsored links. In particular, top sites who rank high on the SE's search results list are likely to benefit less from advertising links. Furthermore, from the SE's perspective, even if they bid high for a sponsored link, consumers may actually not click on this link but rather click on the search result link instead. These two effects may cause secondary sites to end up winning the auction on the sponsored list. On the other hand, if the popularity of a site is large enough compared to secondary sites then the above effects are not enough to compensate for the inherent advantage of a site in directing traffic to itself and top sites may still end-up high on the list of sponsored links.

We also explore a number of extensions. Endogenizing the number of sponsored

76

links allocated by the SE, we show that the SE can increase click-though rates by decreasing the number of these links. A decrease in this number increases the value of the links and may result in compensating the loss associated with a smaller number of links. We also explore the case when multiple search words are considered for sponsored links. We identify conditions under which bidding across words should be independent. A key result here is that the relationship between the search words is the main driver of the bidding outcomes. Finally, we examine a dynamic model, where online advertising has a lagged effect on the site that wins the sponsored link. We identify dynamic bidding patterns that lead to alternating or constant allocations of the sponsored links, depending on the strength of the lagged effect.

All along, the paper proposes testable hypotheses to be confronted with online advertising data. Furthermore, the results also provide normative insights to managers of both sellers and buyers of sponsored links.

4 Empirical Analyses

4.1 Degree Distribution

It is important to confront our model to empirical evidence about the degree distribution of the WWW. An important empirical pattern about the Web, is that it has a scale-free power-law degree distribution (Barabási and Albert 1999, Broder et al. 2000, Faloutsos et al. 1999, Katona 2005). In particular, the ratio of nodes with degree k is $P(k) \sim k^{-\gamma}$, where γ is around 2. More importantly, this empirical observation stands for both in- and out-degrees. Figure 13 illustrates such a degree distribution for the Hungarian Web (Benczúr et al. 2003). Roughly speaking, this means that the proportion of nodes reached from a large number of other nodes tends to be quite high. Previous models from mathematics (e.g. Barabási and Albert (1999), Bollobás et al. (2001), Cooper and Frieze (2003), Katona and Móri (2006)) could claim this finding for in-degrees only but not for out-degrees.

Without assuming a specific distribution of c across sites and a specific functional form for $q(c)$, we cannot predict the actual degree distribution observed. However, we can explain why do we observe an almost identical degree distribution for in- and out-degrees. To our knowledge, our model is the first one to exhibit this characteristic.

In our model it is more commode to study the degree distribution using the ratio of nodes with degree "at least k". The connection between these two quantities can be easily established: a ratio of $k^{-\gamma}$ in the former case yields a ratio of $k^{-\gamma+1}$ in the latter case (i.e. with degrees at least k). Let us denote the number of nodes with in-degree at least k by $N(k)$ and the number of nodes with out-degree at least l by $M(l)$.

Proposition 12 *If we suppose that all the content parameters are different then, in an equilibrium network, $M(N(k))=k$ for any k such that $d_k^{out} > d_{k+1}^{out}$, that is,*
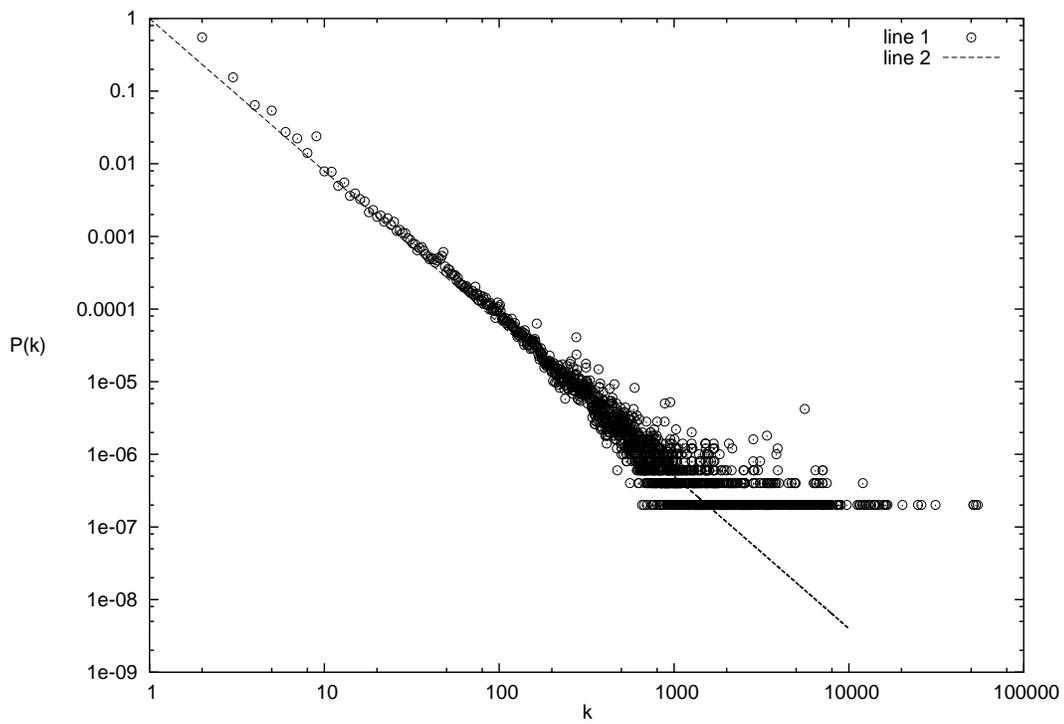
Figure 13: The in-degree distribution of the Hungarian Web. The horizontal axis represents in-degree (k), while the vertical axis measures the number of nodes with a given in-degree ($P(k)$), both on logarithmic scales. The empirical distribution is $P(k) \sim k^{-2.29}$.

$M() = N^{-1}()$, *where the inverse is well defined. Specifically, the in- and out-degree distribution are identical if one of them is a power-law distribution with exponent* $\gamma = 2$.

It follows from the proposition that if one of the in- and out-degree distributions is power-law scale free with exponent γ then the other one is also power-law scale-free with exponent $1 + \frac{1}{\gamma - 1}$. Thus, if γ is about 2 for one of the degree distributions, then it has to be 2 for the other one as well.

4.2 Sold Advertising as a function of content

I defined a Web site's content as its value in the eye of the public and argued that a site's margin is proportional to its content. Thus, people surfing on the Web obtain a higher utility from visiting a high content site than a low content site. However, due to the complexity of the web and the high number of web sites it is hard to find these pages. Search engines provide a service which makes this easier. They find the pages (for every search word) that surfers value the most. Therefore, I may assume that a search engine's goal is to find the highest content Web pages in every area.

In the current study, I use the results from a search engine to examine how the content of a site (estimated from its position among the search results) affects the number of advertising links and size of the advertising space sold on that site. I collected fifty popular search words and ran a Google search for each of them. I selected Web sites from different positions (1st, 201st, 401st, 601st, and 801st)[29] from the results of a given search. On each result page, I counted the number of sold advertising links (num) and estimated the advertising surface (size) on each of these sites. Table 1 shows the results for a number of words.

---

[29]Although Google's estimate for the number of results containing a word is over a million, it only lists around a thousand results. I selected the above numbers to span the entire range of results.

Word	1st num	1st size	201st num	201st size	401st num	401st size	601st num	601st size	801st num	801st size
airlines	0	0	0	0	5	2	2	3	10	6
finance	0	0	0	0	0	0	7	4	10	8
maps	1	2	0	0	2	4	0	0	7	6
pictures	0	0	0	0	0	0	4	7	3	3
sports	0	0	0	0	1	1	5	4	1	1
video	0	0	3	1	0	0	7	6	3	3

Table 1: Results for selected words. Number of sold advertising links and occupied advertising surface on pages in different positions in a Google search for a given word.

Examining the results for the fifty words suggests that both quantities increase with the position in the Google search. Figure 14 shows the averages over the fifty words.

Applying a fixed effects model also confirms that these quantities are significantly increasing ($t = 4.75, p < 0.001$ for *num* and $t = 4.28, p < 0.001$ for size). The coefficient reveals that there is an about 0.23 increase in the number of sold advertising links as the site's position in the search list increases by 100. The results support the hypothesis that sites with lower content sell more advertising, since these sites are in a lower position among search results.

Although this study accounts for sold advertising links, it would also be important to examine bought advertising links. However, this method does not allow us to do that. Sold advertising links are easily identifiable by just looking at a Website, whereas it is much more complicated to gather the advertising links that one site buys. In order to do this, one would have to visit all the sites that a particular site could possibly buy advertising from.
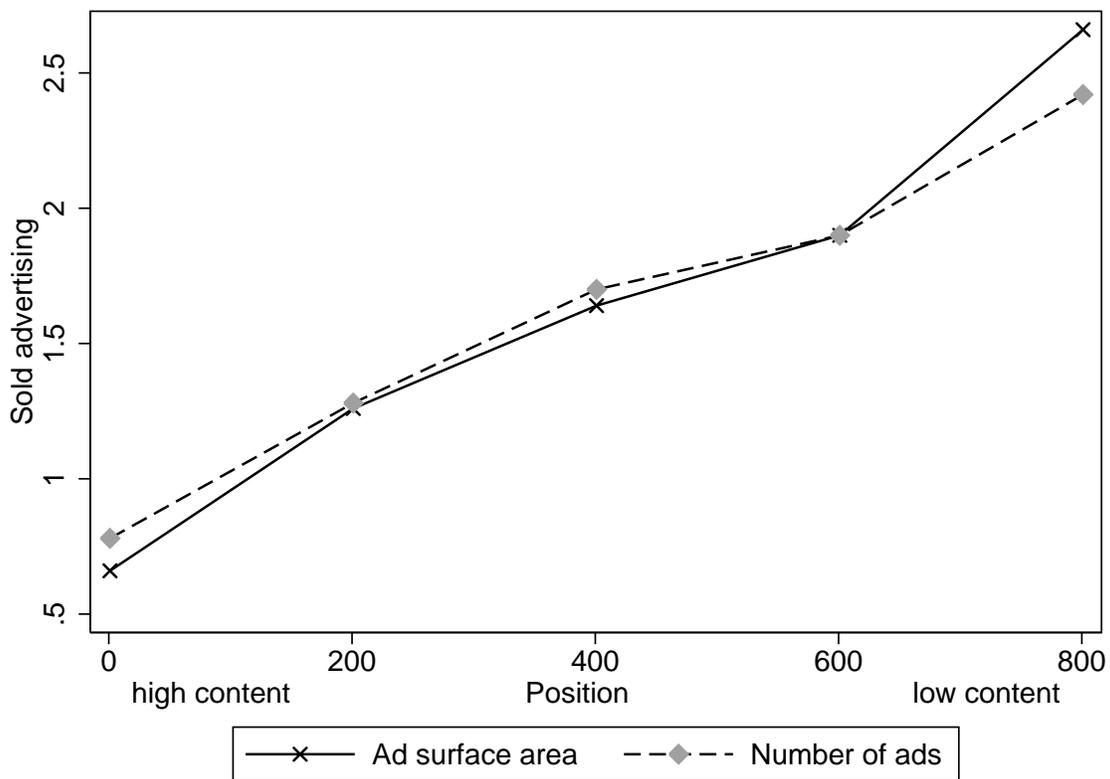
Figure 14: Average number of sold advertising links and advertising space as a function of Google ranking.

4.3 Sponsored links and sold advertising

The previous study estimated the relationship between an indirect measure of content and sold advertising. However, sponsored links allow us to find out more about a site's content. As described in Section 3, sponsored links on a search engine are sold through an auction where, in general, sites which are willing to pay more for a click get to a better position. Thus, there is a clear relationship between profit margin of a site and its position among the sponsored links. This makes it possible to find high content sites and low content sites in a given content area is. The difficulty is that there are only a limited number of sponsored links for a word and all of these point to sites that are relatively high content, since they are willing to buy traffic.

For each of the fifty words that I used in the previous study, I ran a Google search and clicked on the first and last sponsored links. Then, following the same method as in Section 4.2, I estimated the advertising surface occupied by advertisements on the target page. As expected, I only found out-going advertising links on these pages for a low percentage of words (32%), since these are all relatively high content sites. However, the overall size of advertising on the pages that were in the last position among the sponsored links is significantly higher than on those which were in the first position ($t = 3.52, p < 0.001$). Table 2 shows the results for a number of words

These results also support the hypothesis that having a higher content decreases the willingness to sell advertising links. Together with the other two studies, the results provide empirical evidence that the patterns obtained in the model are similar to those that exists in reality.

	size	
Word	first	last
airlines	0	2
cars	0	3
finance	0	2
horoscopes	2	4
songs	0	3
sports	0	4
weather	0	2

Table 2: Results for selected words. Number of sold advertising space on the target page of the first and last sponsored link

5   Conclusion

The dissertation studies two dominant forms of online advertising. The first part analyzes how Web sites buy and sell advertising links directly from/to each other. Overall, I found that both advertising and reference links point to higher content sites. This result strongly supports the broadly accepted search heuristic, which heavily relies on the number of in-links to rank sites with respect to content. This can explain, for instance, why Google's Page Rank algorithm works so well in practice, by showing that in equilibrium, the number of in-links is positively related to a site's content. In contrast to in-links, the pattern of out-links is markedly different for advertising and reference links. Sites tend to purchase advertising links from lower content sites, i.e. the number of advertising out-links is negatively related to the content of a given site. In the case of reference links however, it is higher content sites that tend to establish more out-links. I also show that, in the presence of search engines, this structure becomes more pronounced. Furthermore, search engines provide an incentive to sites to specialize in content areas.

In the second part, I model another important form of online advertising: paid placement. Web sites endowed with different click-through rates race for sponsored advertising links on a search engine's results page. I argue that the SE's problem cannot simply be described as a multi-item auction. The existence of the search results list on the SE's page represents an important externality for both types of players. My key contribution is that I explain the process that may lead to wildly different patterns observed in the behavior of sponsored links. In particular, top sites who rank high on the SE's search results list are likely to benefit less from advertising links. Furthermore, from the SE's perspective, even if they bid high for a sponsored link, consumers may actually not click on this link but rather click on the search result link instead. These two effects may cause secondary sites to end up winning

the auction on the sponsored list. On the other hand, if the popularity of a site is large enough compared to secondary sites then the above effects are not enough to compensate for the inherent advantage of a site in directing traffic to itself and top sites may still end up high on the list of sponsored links.

The empirical section provides evidence that patterns found in the models are consistent with reality. The first study explains the connection between the degree distribution of in- and out-links in the Web. Previous research has shown that these distributions follow a power-law scale-free form, but have not established a connection between the in- and out-degrees.

The second study examines the amount of advertising that different Web sites sell. It is not possible to directly measure a site's content, but search engines rank sites according to how valuable they are in the eye of the public. This ranking gives makes it possible to compare how much space advertising occupies on pages in different positions. The results confirm the hypothesis derived from the model that sites with a higher content sell fewer advertising links.

Finally, the third study connects the two different types of on-line advertising. It examines the amount of sold advertising on sites that pay different prices for sponsored links. Sites that want to get the first sponsored links have to submit a higher bid than sites that just want an arbitrary sponsored link. That is, sites of the former type value traffic more, presumably because they can better leverage traffic. According to the model these site should sell fewer advertising links. The results of the third study confirm this hypothesis.

Overall, the results provide useful guidelines for marketing managers on how to manage their firms' site(s) in terms of their connectedness in the Web. First, competition seems to provide strong incentives for sites to specialize in terms of their business models. Low content sites benefit more from the sales of traffic (advertising)

even though they can only price such traffic at modest rates. High-content sites on the other hand, benefit more from revenues earned from content sales to consumers. These sites should charge high prices for advertising links and, as a result, sell few of these. Instead, they are better off attracting traffic by purchasing advertising links. Because of this increased traffic, high content sites also benefit more from reference links and should therefore, establish more such links. The second part also has several important managerial implications. First, managers of a site should be aware that when they are bidding for sponsored links their site's inherent popularity can significantly change how high they have to bid to get a sponsored link in a good position. Furthermore, there are important interactions between the position occupied in the organic search results list and the sponsored links list. Sites that are among the top search results may not need to win a sponsored link under certain conditions. On the other hand, if they do need a sponsored link they might have to bid higher than sites who are not on the organic results list. Finally, sites have to consider the lagged effect of sponsored links which may lead to a pulsing advertising strategy.

Appendix: Proofs

PROOF OF PROPOSITION 1:

First, we prove that if an equilibrium exists then it has to satisfy (i) and (ii). Although we do not know the Page Rank values, we know how a node's rank is related to its in-neighbors ranks. In particular

$$r_i = \frac{d_i^{out} + 1}{d_i^{out} + 1 - \delta} \left(\frac{1 - \delta}{n} + \delta \sum_{j \to i} \frac{r_j}{d_j^{out} + 1} \right). \tag{20}$$

Therefore, we can transform (5) to

$$u_i = \frac{d_i^{out} + 1}{d_i^{out} + 1 - \delta} (1 - \delta) \frac{\left(c_i - C + \delta q_i \frac{d_i^{out}}{d_i^{out} + 1} \right)}{n} +$$

$$+ \delta \sum_{j \to i} r_j \frac{\frac{d_i^{out} + 1}{d_i^{out} + 1 - \delta} \left(c_i - C + \delta q_i \frac{d_i^{out}}{d_i^{out} + 1} \right) - q_j}{d_j^{out} + 1}. \tag{21}$$

The first term does not depend on player i's decision, therefore it is enough to maximize the sum in the second term if the other agents' decisions are fixed. Player i makes a decision about which in-links to buy, hence s/he only decides which terms to include in the sum. Thus, the sum is maximal if only those terms are included which are non-negative. Hence player i buys a link from player j if and only if

$$\frac{d_i^{out} + 1}{d_i^{out} + 1 - \delta} \left(c_i - C + \delta q_i \frac{d_i^{out}}{d_i^{out} + 1} \right) - q_j \geq 0. \tag{22}$$

This inequality shows that a node buys links from those nodes for which q_j is the lowest. Therefore, in an equilibrium, if $q_k < q_l$ for a given pair of nodes (k, l), then the nodes who buy from node l must form a subset of those who buy from node k, implying that $d_k^{out} \geq d_l^{out}$. Since $q_k = q(c_k) \geq q_l = q(c_l)$ and q is an increasing function, $c_k < c_l$ implies $d_k^{out} \geq d_l^{out}$, completing the proof of part (i) of the proposition.

91

In order to prove part (ii), we have to continue the above argument. Rearranging inequality (22), we get

$$T(i) := \frac{d_i^{out} + 1}{d_i^{out} + 1 - \delta}(c_i - C) + \delta q_i \frac{d_i^{out}}{d_i^{out} + 1 - \delta} \geq q_j. \tag{23}$$

Node i buys a link from node j if and only if this holds. If prices are such that $T(i)$ is increasing, then the number of bought links is increasing in content. We can ensure this by assuming

$$q_i \leq c_i \frac{\delta}{1 - \delta}. \tag{24}$$

However, in Section 2.2.2, we will show that if sites are allowed to set prices, $T(i)$ will be increasing. Therefore, if $c_k < c_l$, that is $k < l$, then $T(k) < T(l)$, hence site l buys more links than site k. The threshold increases as the content increases, therefore the in-degree is an increasing function of the content. As a consequence of the special structure of the graph, if a node has higher content than another, it not only buys more links, but the set of nodes s/he buys links from contains that of the lower content nodes. Since Page Rank is a linear combination of those pages a node buys links from, this ensures that Page Rank is also increasing in content, proving part (ii).

Finally, we will prove that at least one equilibrium exists. We will use the result that any game with convex and compact strategy space and continuous payoff function, which is quasi-concave in the players' own strategies has a pure-strategy Nash-equilibrium. Although, the strategy space in our case is discrete, we will extend it. We will allow the sites to establish partial links. If a site establishes a link partially with weight $0 < w \leq 1$, it only pays w fraction of the price and gets w proportion of the traffic. Fixing the other player's actions, let

$$U_{j \to i}(w) = w r_j \frac{\frac{d_i^{out} + 1}{d_i^{out} + 1 - \delta}\left(c_i - C + \delta q_i \frac{d_i^{out}}{d_i^{out} + 1}\right) - q_j}{d_j^{out} + 1} \tag{25}$$

92

denote the payoff of establishing link $j \to i$ with weight w for node i. These $U_{j\to i}(w)$ functions are linear, therefore the payoff function is quasi-concave, since it is the sum of these functions. Since we extended the strategy space, it is compact and convex. Also, the payoffs are continuous and concave in the players' own actions, hence an equilibrium exists. Furthermore, in this equilibrium, a player will only establish a partial link if s/he is totally indifferent about the link. If a site has a profit increase from establishing a link partially, it has an even higher increase from establishing it fully. In equilibrium, however, a player can only be indifferent about one link. Therefore, in this equilibrium, every player will establish at most one partial link, the rest of the links will be either fully or not established. Notice also, that we only show the existence of an equilibrium, but this may not be unique. $\qquad\square$

PROOF OF PROPOSITION 2:

We begin by proving that if $q(i)$ is increasing and $q(i) \leq \frac{\delta}{1-\delta}c(i)$, then in any equilibrium, $d^{in}(i)$ is also increasing and $d^{out}(i)$ is decreasing. Similarly to the discrete case, player i buys a link from player j if and only if

$$\frac{d^{out}(i) + s}{d^{out}(i) + s(1-\delta)}(c(i) - C) + \delta q(i)\frac{d^{out}(i)}{d^{out}(i) + s(1-\delta)} \geq q(j). \qquad (26)$$

This shows that a node buys links from those nodes for which $q(j)$ is the lowest. Therefore, in an equilibrium, if $q(k) < q(l)$ for a given pair of nodes (k, l), then the nodes who buy from node l must form a subset of those who buy from node k, implying that $d^{out}(k) \geq d^{out}(l)$, therefore $d^{out}(i)$ is decreasing.

In order to prove that $d^{in}(i)$ is increasing, we have to continue the above argument. We repeat inequality (26),

$$T(i) := \frac{d^{out}(i) + s}{d^{out}(i) + s(1-\delta)}(c(i) - C) + \delta q(i)\frac{d^{out}(i)}{d^{out}(i) + s(1-\delta)} \geq q(j). \qquad (27)$$

to recall the decision rule of a node. The left hand side defines a $T(i)$ threshold for node i, deciding from which nodes to buy links. The number of links bought buy

node i depends on this quantity. The higher $T(i)$ is, the more links it buys. If, for example, $q(i) \leq \frac{\delta}{1-\delta}c(i)$, then $T(i)$ is increasing. Furthermore, we will show in Proposition 3, that $T(i)$ is increasing if players set their prices. Finally, if $T(i)$ is increasing, $d^{in}(i)$ will also be increasing.

In order to prove the existence of an equilibrium we will use Tikhonov's fixed point theorem (Istratescu 1981). It states that if X is a compact convex subset of a a locally convex topological vector space (X) and $f : X \to X$ is continuous, then f has a fixed point. Recall equation (26), describing the decision rule of player i. Player j sells links to the nodes that satisfy $T(i) > q(j)$. Therefore,

$$d^{out}(j) = \Lambda\left(i|T(i) > q(j)\right). \tag{28}$$

Let $L(j)$ denote the right hand side of equation (28), which is a measurable function if $d^{out}()$ is measurable. A function $d^{out}(j)$ satisfying $d^{out}(j) = L(j)$ must represent an equilibrium. We will show that the operator mapping $L()$ to the function d^{out} is continuous. Since q is a continuous function, $\int_0^\infty |T^{-1}(j) - T'^{-1}(j)|dj \leq c_1 \int_I |d(i) - d'(i)|di$ with a suitable c_1 constant, where $T()$ and $T'()$ are the threshold functions corresponding to $d()$ and $d'()$, respectively. Also, let $L()$ and $L()'$ denote the functions that the operator assigns to $d()$ and $d'()$. Then, $\int_0^\infty |L(j) - L'(j)|dj = \int_0^\infty |T^{-1}(j) - T'^{-1}(j)|dq(j)$. Since $q(j)$ is continuous on a compact set, it has to bounded therefore, $\int_0^\infty |T^{-1}(j) - T'^{-1}(j)|dq(j) \leq c_2 \int_I |d(i) - d'(i)|di$, hence the operator is continuous. We will apply Tikhonov's theorem to this operator on the normed space of L_1 functions on $[0, 1]$. The fixed point of this operator must satisfy (28), thus it represents an equilibrium of the game. However, the equilibrium may not be unique. □

PROOF OF PROPOSITION 3:

Let us consider a refined SPNE $(q, E(q))$ and look at the optimization problem that a site faces in stage one. Let ζ denote $q(i)$, that is, the decision variable of site i

in stage one. We have seen in the proof of Proposition 2, that in the second stage a site essentially only decides how many links to buy and establishes them from the cheapest sites. Let ψ denote $d^{in}(i)$, that is, the decision variable in the second stage. Let $D(\zeta)$ be the aggregate demand for out-links in the second stage (in the equilibrium $E(q)$), that is, the measure of the set of sites who want to buy a link from site i (or any site). Let $K(\psi)$ denote the cost of ψ links, that is, $K(\psi) = \int_{j \to i} p(j)dj$. Obviously, $K(\psi)$ is increasing and $D(\zeta)$ is decreasing. Also, rewriting (6) Page Rank is

$$r(i) = \frac{d_i^{out} + s}{d_i^{out} + s(1 - \delta)} \left((1 - \delta) + \delta \int_{x \to i} \frac{r(x)}{d^{out}(x) + s} dx \right).$$

Decomposing this into two factors, let

$$r_1(\zeta) = \frac{D(\zeta) + s}{D(\zeta) + s(1 - \delta)}$$

denote the first factor and

$$r_2(\psi) = (1 - \delta) + \delta \int_{x \to i} \frac{r(x)}{d^{out}(x) + s} dx$$

the second. Then, rewriting the utility function, we have

$$u_i(\psi, \zeta) = r_2(\psi)r_1(\zeta) \left(c(i) - C + \delta\zeta \frac{D(\zeta)}{D(\zeta) + s} \right) - K(\psi). \qquad (29)$$

Since $(q, E(q))$ is a refined SPNE, ζ and ψ has to maximize this function, as if the price and in-link decisions were simultaneously made. If we fix i, the solution of the maximization problem in ζ is the same for all ψ's. This optimal $\zeta^*(i)$ is increasing in i, because the function

$$T(i, \zeta) = r_1(\zeta) \left(c(i) - C + \delta\zeta \frac{D(\zeta)}{D(\zeta) + s} \right) =$$
$$= \frac{D(\zeta) + s}{D(\zeta) + s(1 - \delta)}(c(i) - C) + \delta\zeta \frac{D(\zeta)}{D(\zeta) + s(1 - \delta)} \qquad (30)$$

95

has increasing differences in (i, ζ), since the term that contains both variables is a product of two increasing functions (of i and ζ, respectively). Furthermore, the optimal T, that is, $T^*(i) = T(i, \zeta^*(i))$ is also increasing, because if $l > k$ then

$$T^*(l) = T(l, \zeta^*(l) \geq T(l, \zeta^*(k) > T(k, \zeta^*(k)) = T^*(k).$$

Therefore, in equilibrium both $q(i)$ and $T(i)$ are strictly increasing (if $c(i)$ is strictly increasing), hence the second stage results hold. $\square$

PROOF OF PROPOSITION 4:

We will show that the payoff function has increasing differences in the players' own decisions ($d_i^{in_A}$, $d_i^{out_R}$) and in the pairs composed of an own decision variable and another player's decisions variable. Although (9) is not written as a direct function of other players' decisions, these are captured by $d_i^{in_R}$ and $d_i^{out_A}$. If another player buys more advertising links $d_i^{out_A}$ either increases or does not change. If another player establishes an extra reference link $d_i^{in_R}$ does not change or increases. Then it is straightforward to check that the payoff function has increasing differences in the above mentioned variable pairs, because with the exception of $f(.,.)$, which has increasing differences in its variables by definition, the relevant terms are always products of functions which are increasing in the variables in question.

Therefore, the game is supermodular, hence we can use the machinery introduced by Topkis (1998) to describe the characteristics of the equilibria. It follows from supermodularity that the pure-strategy equilibria of the game form a non-empty complete lattice with a greatest and a least element, where the former is Pareto-optimal. Moreover, we can show that any equilibrium has the following special structural properties.

One can see that if a node select how many reference links to establish, it connects these to the highest content nodes. Also, every node buys advertising links from the

cheapest nodes, hence we obviously have $d_i^{in_R} \geq d_j^{in_R}$ if $c_i > c_j$ and $d_i^{out_A} \leq d_j^{out_A}$ if $p_i > p_j$, that is, if $c_i > c_j$. Now, we have to show, that in equilibrium, the actions of players are increasing with respect to their content.

Since every node buys advertising links from the lowest content nodes, and establishes reference links to the highest, the two decision variables of site i are only the number of links to establish: $d_i^{in_A}$ and $d_i^{out_R}$. It is easy to see that the payoff function has increasing differences in the pairs $(d_i^{in_A}, d_i^{out_R})$, $(d_i^{in_A}, i)$ and $(i, d_i^{out_R})$, checking the terms that contain two of the variables in question. Therefore, the optimal decisions $(d_i^{in_A*}, d_i^{out_R*})$ are increasing in i. That is, if $i > j$ (i.e. $c_i > c_j$) then $d_i^{out_R*} \geq d_j^{out_R*}$, and $d_i^{in_A*} \geq d_j^{in_A*}$. $\qquad \square$

PROOF OF CLAIM 1:

The search engine wishes to maximize the income from the s winners of the sponsored link. Given the order of sites it is obviously optimal to set the p_i's to the maximum, that is, $p_i = b_i$. Regarding the order of sites, if Site i acquires a sponsored link, the search engine will receive a total payment of $\beta A(i) F_i$ from that site, where $F_i = \gamma_i b_i (1 - \delta I(i))$. The F_i values are site specific and only depend on the site's parameters, whereas the $A(i)$ values are determined by the search engine, when it assigns the sponsored links. In order to maximize $\beta \sum_{i=1}^{n} A(i) F_i$, the SE has to assign the α's in a decreasing order of the F_i values. $\qquad \square$

PROOF OF PROPOSITION 6:

As we have discussed before, the winner both in an FNE and an SSNE is the site with highest valuation, The payment of the winner is between the first and second valuations. $\qquad \square$

PROOF OF LEMMA 3:

97

If sites' preferences are aligned, then (15) yields $G_1(w_l) \geq G_1(w_m)$ for every $l < m$, proving the lemma.

$\square$

PROOF OF PROPOSITION 7:

In order to prove the existence of an SSNE, we have to show that there exist $P_1 \geq P_2 \geq ... \geq P_s$, such that, they satisfy inequalities (13) and (14) for every $1 \leq k < l \leq s$. We will show that if the sites' preferences are aligned, then it is enough to check that $P_1 \geq P_2 \geq ... \geq P_s$ satisfy a subset of them, namely the following inequalities, for every j.

$$G_j(w_j) - G_{j+1}(w_j) \geq P_j - P_{j+1} \geq G_j(w_{j+1}) - G_{j+1}(w_{j+1}) \tag{31}$$

We have to show, that all the inequalities in (13) and (14) follow from these in (31). Let $1 \leq k < l \leq s$ be arbitrary indices. Summing (31) for $j = k$ to l, we get

$$\sum_{j=k}^{l-1}[G_j(w_j) - G_{j+1}(w_j) \geq P_k - P_l \geq \sum_{j=k}^{l-1}[G_j(w_{j+1}) - G_{j+1}(w_{j+1})].$$

Since the preferences are aligned, $G_j(w_k) - G_{j+1}(w_k) \geq G_j(w_j) - G_{j+1}(w_j)$ for $j > k$, therefore, we obtain

$$G_k(w_k) - G_l(w_k) \geq P_k - P_l,$$

and similarly

$$P_k - P_l \geq G_k(w_l) - G_l(w_l).$$

We have shown, that the system given by (13) and (14) is equivalent to that defined by (31). That is, it is always enough to check whether a site wants to get to a position which is one higher or lower. Therefore, given that (12) holds, the values of $P_j - P_{j+1}$ can be chosen arbitrarily from the intervals given in (31), fixing $P_{s+1} = 0$. In (16), we basically assume that selecting the maximum values does not violate (12). Thus,

we get the second part of proposition by summing the left hand sides of (31) in the following way.

$$\sum_{i=1}^{s} P_i = sP_s + \sum_{j=1}^{s-1} j(P_j - P_{j+1}).$$

For the fourth part, let us note that every SSNE is an SNE, therefore the maximum SNE income is at least as high as the maximum SSNE income. For the other direction, let P_i^N denote the expenditure of Site i in an SNE with maximum revenue and let P_i^S denote the same expenditure in a maximum revenue SSNE. From the previous part, we know that

$$P_j^S = P_{j+1}^S + G_j(w_j) - G_{j+1}(w_j),$$

However, according to the definition of an SNE,

$$P_j^N \leq P_{j+1}^N + G_j(w_j) - G_{j+1}(w_j).$$

Since $G_{s+1}(w_s) = 0$,

$$P_s^N \leq G_s(w_s) = P_s^S.$$

Then, it is easy to show recursively that $P_i^N \leq P_i^S$, completing the proof. $\qquad\square$

PROOF OF PROPOSITION 8:

According to Proposition 7, the maximum equilibrium revenue of the SE, in case of selling s links, is

$$M(s) = \beta \left(\sum_{j=1}^{s} j\gamma_j\alpha_j - \sum_{j=1}^{s-1} j\gamma_j\alpha_{j+1} \right).$$

If the SE decides to instead sell only t links, the traffic on the remaining links will increase by a factor of $(1 + \beta(\alpha_{t+1} + ... + \alpha_s))$. Therefore, the maximum equilibrium revenue will be

$$(1 + \beta(\alpha_{t+1} + ... + \alpha_s))\beta \left(\sum_{j=1}^{t} j\gamma_j\alpha_j - \sum_{j=1}^{t-1} j\gamma_j\alpha_{j+1} \right)$$

99

in this case. Comparing the two quantities, we get the expression in the proposition.
□

PROOF OF PROPOSITION 9:

First, we prove the second part of the proposition, that is, identify the conditions necessary for an alternating equilibrium. In such an equilibrium, bidding strategies are such, that if Site i has won the previous auction then Site $j = 3 - i$ is the current winner. Let $P^{(i)}$ denote the fee that Site $j = 3 - i$ has to pay in the auction when Site i is the previous winner. Let $V_i^{(j)}$ denote the discounted equilibrium profits of Site i from a given period when Site j is the previous winner. In an alternating equilibrium,

$$
\begin{aligned}
V_1^{(1)} &= \delta V_1^{(2)} \\
V_1^{(2)} &= G_l(1) - P^{(2)} + \delta V_1^{(1)} \\
V_2^{(1)} &= G_l(2) - P^{(1)} + \delta V_2^{(2)} \\
V_2^{(2)} &= \delta V_2^{(1)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
V_1^{(2)} &= \frac{G_l(1) - P^{(2)}}{1 - \delta^2} \\
V_2^{(1)} &= \frac{G_l(2) - P^{(1)}}{1 - \delta^2}
\end{aligned}
$$

The sufficient and necessary conditions these valuations and prices have to satisfy are that in a given auction, the winner has to have a higher valuation and fee payed by the winner must fall between the two players' valuations (both in an MFNE and MSSNE). For example, if the previous winner is Site 1, then the current winner must be Site 2, therefore,

$$
G_w(1) + \delta(V_1^{(1)} - V_1^{(2)}) \le P^{(1)} \le G_l(2) + \delta(V_1^{(1)} - V_1^{(2)})
$$

must hold. Plugging the corresponding formulas, we obtain

$$G_w(1) - \frac{1-\delta}{1-\delta^2}(G_l(1) - P^{(2)}) \leq P^{(1)} \leq G_l(2). \tag{32}$$

Comparing the valuations in a period when Site 2 is the previous winner, we get a similar inequality,

$$G_w(2) - \frac{1-\delta}{1-\delta^2}(G_l(2) - P^{(1)}) \leq P^{(2)} \leq G_l(1). \tag{33}$$

The set defined by (32) and (33) is a two-dimensional simplex. It is easy to see that it is non-empty iff $G_l(2) \geq G_w(1)$ (given the other restrictions on the parameters).

The maximum discounted income of the seller depends on the first period of the game. Let P_t denote its income in period t. If Site 1 is the first winner, then it would be

$$\sum_{t=1}^{\infty} \delta^{t-1} P_t = \frac{P^{(2)} + \delta P^{(1)}}{1-\delta^2}.$$

If Site 2 is the first winner, then it is

$$\frac{P^{(1)} + \delta P^{(2)}}{1-\delta^2}.$$

We determine the maximum for both and consider the higher value. Clearly, since Site 1 has higher valuations, the SE's income will be higher if Site 1 is the first winner. Maximizing $P^{(2)} + \delta P^{(1)}$ on the simplex defined by (32) and (33), we get

$$M_2 = \frac{G_l(1) + \delta G_l(2)}{1-\delta^2}.$$

The first part of the proposition can be proven by following the same steps. However, it is obvious, that since in both states site 1 has a higher valuation, it is always the winner. Then the price payed must be in the given range, yielding the stated maximum income. $\qquad\square$

PROOF OF COROLLARY 2:

The values of $G_l(1) > G_l(2)$ are independent of q. When $q = 0$, $G_l(i) = G_w(i)$ and as q increases $G_w(1)$ decreases. Let q^* be the unique solution of

$$R((1+q)I(1)\gamma_1\alpha_1 + (1+q)\gamma_1\beta\alpha_1)) - R((1+q)I(1)\gamma_1\alpha_1 + q\gamma_1\beta\alpha_1) =$$
$$= R((1+q)I(2)\gamma_2\alpha_2 + \gamma_2\beta\alpha_1)) - R((1+q)I(2)\gamma_2\alpha_2).$$

Then, for $0 < q < q^*$, we get the first case in Proposition 9 and for $q^* < q$, we get the second case. $\square$

PROOF OF COROLLARY 3:

Fixing $G_l(2)$ in Proposition 9, we can establish

$$\lim_{G_w(1)-G_l(2)\to 0^+} M_1 = \frac{G_l(2)}{1-\delta},$$

$$\lim_{G_w(1)-G_l(2)\to 0^-} M_2 = \frac{G_l(1) + \delta G_l(2)}{1-\delta^2} = \frac{G_l(2)}{1-\delta} + \frac{G_l(1) - G_l(2)}{1-\delta^2}.$$

Hence, the difference is

$$0 < \frac{G_l(1) - G_l(2)}{1-\delta^2} = \frac{G_l(1) - G_w(1)}{1-\delta^2},$$

which clearly increases in q and δ. $\square$

PROOF OF PROPOSITION 10:

The payoffs in one content area do not depend on the actions in the rest of the content areas. Therefore, the best responses can be determined separately in every content area and the equilibria will be the same that we described in Proposition 6. $\square$

PROOF OF PROPOSITION 11:

Before describing the equilibria in the different cases, we establish a general rule that drives the results. In any equilibrium (FNE or SSNE), the winner of the auction

102

for keyword i has to have the highest marginal valuation, given the result in the other auctions. Furthermore, the price payed by the winner has to between its marginal valuation and the second highest marginal valuation. Using this rule, we can now determine the equilibria in the different cases.

1. If two different sites have the highest valuation, then clearly they can be the only winners, otherwise there is a site with a higher marginal valuation.

2. Clearly, if Site x_1 wins one auction, then it still has the highest marginal valuation in the other one.

3. It is easy to see, that the two possible combination of winners is $(w^1 = x_2, w^2 = y_1)$ and $(w^1 = x_1, w^2 = y_2)$. Assume for a moment, that we have the latter case in equilibrium. Then, site x_1 does not have an incentive to give up keyword 1 in order to win keyword 2, that is,

$$G_{\{1\}}(x_1) - P^1 \geq G_{\{2\}}(x_1) - P^2.$$

Similarly, Site 2 does not have an incentive to give up keyword 2 in order to win keyword 1, that is,

$$G_{\{2\}}(x_2) - P^2 \geq G_{\{1\}}(x_2) - P^1.$$

Combining the two, we obtain

$$G_{\{1\}}(x_1) - G_{\{2\}}(x_1) \geq P^1 - P^2 \geq G_{\{1\}}(x_2) - G_{\{2\}}(x_2), \qquad (34)$$

yielding

$$G_{\{1\}}(x_1) + G_{\{2\}}(x_2) \geq G_{\{2\}}(x_1) + G_{\{1\}}(x_2),$$

contradicting our assumption in (19). This leaves us the case $(w^1 = x_2, w^2 = y_1)$.

To show that this type of equilibrium exists and to determine the maximum income, we determine the sufficient and necessary conditions on the prices. From the marginal valuation argument, we have

$$G_{\{1\}}(x_2) \geq \quad P^1 \quad \geq \max(G_{\{1,2\}}(x_1) - G_{\{2\}}(x_1), G_{\{1\}}(x_3)), \tag{35}$$

$$G_{\{2\}}(x_1) \geq \quad P^2 \quad \geq \max(G_{\{1,2\}}(x_2) - G_{\{1\}}(x_2), G_{\{2\}}(y_3)). \tag{36}$$

Furthermore, using the same argument as for (34), we obtain

$$G_{\{1\}}(x_2) - G_{\{2\}}(x_2) \geq P^1 - P^2 \geq G_{\{1\}}(x_1) - G_{\{2\}}(x_1), \tag{37}$$

One can check that the simplex defined by (37), (35), and (36) is non-empty and that the maximum of $P^1 + P^2$ is as stated.

4. As in the previous part we have to deal with the two possible types ($w^1 = x_2, w^2 = y_1$) and ($w^1 = x_1, w^2 = y_2$), but in this case both are possible. Determining the existence and the maximum income of the first goes as before.

For the second type, we have to examine the condition under which it exists. Similarly to the previous part, the following inequalities define the set of equilibria of this type.

$$G_{\{1\}}(x_1) \geq \quad P^1 \quad \geq G_{\{1\}}(x_2), \tag{38}$$

$$G_{\{2\}}(y_2) \geq \quad P^2 \quad \geq \max(G_{\{1,2\}}(x_1) - G_{\{1\}}(x_1), G_{\{2\}}(y_3)) \tag{39}$$

$$G_{\{2\}}(y_2) - G_{\{1\}}(y_2) \geq \quad P^2 - P^1 \quad \geq G_{\{2\}}(x_1) - G_{\{1\}}(x_1) \tag{40}$$

For (39), we need $G_{\{1,2\}}(x_1) < G_{\{1\}}(x_1) + G_{\{2\}}(y_2)$ and for (40), we need $G_{\{2\}}(x_1) + G_{\{1\}}(y_2) \leq G_{\{1\}}(x_1) + G_{\{2\}}(y_2)$. If these hold, the set is non-empty with the maximum income as stated.

□

104

PROOF OF PROPOSITION 12:   Recall that $N(k)$ denotes the number of nodes with in-degree at least k, while $M(k)$ denotes the number of nodes with out-degree at least k. The connection between these functions is the following. It follows from the proof of Proposition 1 that the kth largest out-degree is equal to the number of nodes with in-degree at least k because every node buys its in-links from the nodes with lowest content. Thus $d_k^{out} = N(k)$. Furthermore, since the out-degree is decreasing as k increases, the number of nodes which have out-degree higher than node k is at least k and is exactly k if $d_k^{out} > d_{k+1}^{out}$. Hence in the latter case $M(N(k)) = k$. To summarize the connection,

$$N() = d^{out}() = M^{-1}(),$$

where the inverse is well-defined. □

□

References

Agrawal, Deepak. 1996. Effect of brand loyalty on advertising and trade promotions: A game theoretic analysis with empirical evidence. *Marketing Science* **15**(1) 86–108.

Bagwell, Kyle. 2005. The economic analysis of advertising. Unpublished Manuscript.

Bala, V., S. Goyal. 2000. A noncooperative model of network formation,. *Econometrica* **68**(5) 1181–1229.

Barabási, A.-L., R. Albert. 1999. Emergence of scaling in random networks. *Science* **286** 509–512.

Benczúr, A., K. Csalogány, D. Fogaras, E. Friedman, T. Sarlós, M. Uher, E. Windhager. 2003. Searching a small national domain – preliminary report. *Proceedings of WWW*.

Bollobás, B., O. Riordan, J. Spencer, G. Tusnády. 2001. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms* **18** 279–290.

Bramouille, Y., D. Lopez-Pintado, S. Goyal, F. Vega-Redondo. 2004. Network formation and anti-coordination games. *International Journal of Game Theory* **33** 1–19.

Brin, S., L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* **30(1-7)** 107–117.

Broder, A.Z., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J.L. Wiener. 2000. Graph structure in the web. *Computer Networks* **33** 309–320.

Chen, Y., C. He. 2006. Paid placement: Advertising and search on the internet. *Working Paper, U. of Colorado Boulder* .

Cooper, C., A. Frieze. 2003. A general model of web graphs. *Random Structures and Algorithms* **22** 311–335.

Dorfman, R., P. O. Steiner. 1954. Optimal advertising and optimal quality. *American Economic Review* **44** 826–36.

Dukes, Anthony, Esther Gal-Or. 2003. Negotations and exclusivity contracts for advertising. *Marketing Science* **22**(2) 222–245.

Edelman, B., M. Ostrovsky, M. Schwarz. 2007. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review* **97**(1) 242–259.

Faloutsos, M., P. Faloutsos, C. Faloutsos. 1999. On power-law relationships of the internet topology. *Comp. Comm. Rev.* **29** 251–262.

Godes, David, Elie Ofek, Miklos Sarvary. 2006. Products vs. advertising:the impact of competition on media firm strategies. *Working paper, Harvard Business School* .

Hoffman, Donna L., Thomas Novak. 2000. Advertising pricing models for the world wide web. Deborah Hurley, Brian Kahin, Hal Varian, eds., *Internet Publishing and Be-*

yond: The Economics of Digital Information and Intellectual Property. MIT Press, Cambridge.

Istratescu, Vasile I. 1981. *Fixed Point Theory, An Introduction*. D.Reidel, Holland.

Iyer, Ganesh, V. Padmanabhan. 2006. Internet-based service institutions. *Marketing Science* **25**(6) 598–600.

Jackson, Mathew O. 2003. A survey of models of network formation: stability and efficiency. *mimeo* .

Jackson, Mathew O., Asher Wolinsky. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* **71** 44–74.

Katona, Zs. 2005. Width of a scale-free tree. *Journal of Applied Probability* **42**(3) 839–850.

Katona, Zs., T.F. Móri. 2006. A new class of scale free random graphs. *Statistics and Probability Letters* **76**(15) 1587–1593.

Katona, Zs., M. Sarvary. 2006. Network formation and the structure of the commercial world wide. *Insead working paper 2006/04/MKT* .

Langville, Amy N., Carl D. Meyer. 2004. Deeper inside pagerank. *Internet Mathematics* **1**(3) 335–400.

Masson, Robert T., Ram Mudambi, Robert J. Reynolds. 1990. Oligopoly in advertiser supported media. *Quarterly Review of Economics and Business* **30**(2) 3–16.

Mayzlin, Dina, Hema Yoganarasimhan. 2006. Link to success: How blogs build an audience by promoting rivals. *Working Paper,* Yale School of Management.

Nelson, Philip. 1974. Advertising and information. *Journal of Political Economy* **81** 729–54.

Ofek, E., M. Sarvary. 2003. R&d, marketing and the success of next generation products. *Marketing Science* **22**(3) 355–370.

Rutz, Oliver J., Randolph E. Bucklin. 2007a. From generic to branded: A model of spillover dynamics in paid search advertising. *Working paper, UCLA* .

Rutz, Oliver J., Randolph E. Bucklin. 2007b. A model of individual keyword performance in paid search advertising. *Working paper, UCLA* .

Schmalensee, Richard. 1978. A model of advertising and product quality. *Journal of Political Economy* **86**(31) 485–503.

Topkis, Donald M. 1998. *Supermodularity and Complementarity*, chap. Noncooperative Games. Princeton University Press, Princeton, New York, 175–206.

Varian, H. 2007. Position auctions. *Internation Journal of Industrial Organization* to appear.

Villas-Boas, J., Miguel. 2004. Communication strategies and product line design. *Marketing Science* **23**(3) 304–16.

Villas-Boas, M. 1993. Predicting advertising pulsing policies in an oligopoly: A model and empirical test. *Marketing Science* **12**(1) 88–102.

Zeff, Robbin, Brad Aronson. 1999. *Advertising on the Internet*. John Wiley and Sons.

Zeithammer, R. 2006. Forward-looking bidding in online auctions. *Journal of Marketing Research* **43**(3) 462–76.