INSEAD
The Business School
for the World

Faculty & Research Working Paper

Network Formation and the Structure of the Commercial World Wide Web

Zsolt KATONA
Miklos SARVARY
2006/60/MKT (revised version of 2006/04/MKT)

Network Formation and the Structure of the Commercial World Wide Web

Zsolt Katona*

and

Miklos Sarvary**

November 3, 2006

* PhD student of Marketing at INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France, zsolt.katona@insead.edu

** Associate Professor of Marketing, Director, The International Centre for Learning Innovation (ICLI) and Coordinator, Marketing Area at INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France, miklos.sarvary@insead.edu, Tel: +33 1 60 71 26 05 Fax: +33 1 60 74 55 00.

Network Formation and the Structure of the Commercial World Wide Web

Abstract

We model the commercial World Wide Web (WWW) as a directed graph emerging as the equilibrium of a game in which utility maximizing Web sites purchase (advertising) in-links from each other, while also setting the price of these links. A key feature of our model is that we consider sites to be heterogeneous in terms of their "content", i.e. their inherent value to consumers. In a world where consumers 'surf' on the WWW, sites' revenues/profits originate from two sources: (i) the sales of content (products) to consumers, and (ii) the sales of links (traffic) to other sites. We find that in equilibrium, higher content sites tend to purchase more advertising links mirroring the Dorfman-Steiner rule. Sites with higher content sell less advertising links and offer such links at higher prices. As such, there seems to be specialization across sites in terms of revenue models: high content sites tend to earns revenue from the sales of content while low content ones from the sales of traffic (advertising). In an extension, we also allow sites to establish (reference) out-links to each other beyond the sales of advertising links and find that there is a general tendency to establish reference link to sites with higher content. Overall, there is a strong positive correlation between a site's content and the number of its in-links. We also explore network formation in the presence of search engines and find that the higher the proportion of people using these, the more sites have an incentive to specialize in certain "content areas". Our results have interesting practical implications for 'search-engine optimization', the pricing of Internet advertising as well as the choice of Internet business models. They also shed light on why successful search engines (e.g. Google) can use simple heuristics based on in-links to rank sites with respect to their content.

Keywords: Internet advertising, game theory, network formation.

1 Introduction

The Internet and its most broadly known application, the World Wide Web (WWW) is gaining tremendous importance in our society. It represents a new medium for doing business that transcends national borders and attracts an ever larger share of social and economic transactions. 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 this paper is to develop a model that helps understand what structure emerges from this decentralized network formation process.

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. 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.

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] 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).

$1.25 billion business with a growth rate in 2005 reaching 125%.

Similarly, the primary way through which sites can drive traffic to themselves is the purchase of advertising links.[2] 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.

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

[2]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).

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 from low content sites 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-view (or 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. 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.

The paper is organized as follows. The next section reviews the relevant literature. Section 3 presents the basic model, which considers advertising

links and exogenous prices. Section 4 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 5 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 paper ends with a general discussion and concluding remarks. To improve readability, most proofs have been delegated to the Appendix.

2 Relevant Literature

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 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.[3] 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 de-

[3]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.

pends 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).[4] 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. More recent papers in marketing (see, e.g. 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 two 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. 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

[4]See also 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.

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 consistent with the reality of the WWW than those of the existing network formation literature. This model is presented next.

3 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 $i \not\to j$ 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.[5] Beyond

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

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.

3.1 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 (even approximately) is not a trivial task because the weight (or "vote") 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 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 a site, with probability δ, s/he randomly follows a link going out from that site, choosing each out-link with equal probability. 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.

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). 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:

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

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 (i.e. the j-th site linking to site i).

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 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:

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

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 $t \to \infty$ 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 $t \to \infty$.*

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. To overcome this, Langville and Meyer (2004) suggest to add links from nodes with no regular out-links to every other node. In the browsing process this corresponds to the idea that if a browser arrives to a node with no out-links s/he is forced to jump. However, this is not consistent with the original definition of the process since equation (1) does not contain the "votes" from sites with no out-links. Therefore, we apply a second method, which is consistent with the original definition of Brin and Page (1998). We add a virtual node to the graph, representing inactive users. We link every other node to this virtual node and finally link it to itself. In this model, if users get to nodes with no regular out-links they switch to an inactive state and stay there until they decide to stop browsing or jump to a random site. If we calculate the Page Rank of the nodes in this graph, for the regular nodes we get the same values as we would from the original definition.

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.[6] 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.

3.2 Network formation

Assume that to the n nodes (sites) correspond 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 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

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

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-visitor, 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 thousand impressions" (CPM) or "rates per click-through".[7] If another node purchases a link then this link will be created and point from the seller to the buyer. Given prices, each node makes a simultaneous decision about its incoming links, that is, which other nodes it 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-visitor prices for links are exogenous but we will relax this assumption in Section 4.2. Specifically, in this section we will assume that $q_i = q(c_i)$ is an increasing function of content c_i. Then, the total price of an advertising link from site i is $p_i = r_i q(c_i)$. In Section 4.2, 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. In our model, we do not explicitly describe consumers' behavior at a site, this would result in a model that is prohibitively complex. We simply say that the gross profit of the site is proportional to the number of

[7]Clearly, in our model that uses a representative consumer, any measure of click-through would be proportional to the number of visitors. Also, the boundaries between click-through and CPM are increasingly disappearing. As online ads adopt similar formats to those of traditional media (e.g. 'Video Clip Module' for real time video sequencing) sites tend to use the same metrics as TV commercials - see Eyeblaster press release, June 7, 2004.

consumers flowing through the site as well as the site's content. This does not allow us to capture the basic tradeoff between keeping a consumer or handing it over to another site. The price of a link increasing in the site's content does exactly that however. The higher the gain from a consumer (i.e. the higher c), the higher the site wants to charge for handing it over to another site. In other words, this price function captures the tradeoff between sites' two revenue streams.

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:

$$\begin{aligned} u_i &= r_i(c_i - C) + p_i \cdot d_i^{out} - \sum_{j \to i} p_j \\ u_i &= r_i(c_i - C + q_i \cdot d_i^{out}) - \sum_{j \to i} r_j q_j. \end{aligned} \tag{5}$$

3.3 Equilibrium analysis

We are looking for the pure strategy 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 pure strategy Nash-equilibrium always exists and all the equilibria have the following properties.*

(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 $d_k^{out} \geq d_l^{out}$. If $c_k = c_l$ then $d_k^{out} + 1 \geq d_l^{out}$.

(ii) If we suppose that 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 $q(c)$'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-traffic price". In the next step, we show that 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. □

Figure 1 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 small content pages (small dots) and are directed towards large 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. However, the meaning of this more general statement is that if two sites have different contents then the one with lower content has the more out-degrees (not necessarily strictly) and if two sites have the same content parameter then the difference in out-degrees can be at most one. If we plot out-degrees as a function of content, this means that out-degree is mainly decreasing but at equal content nodes it can jump up one. Of course the nodes can be ordered (as is done on the figure) such that both the contents are increasing and the out-degrees are decreasing.

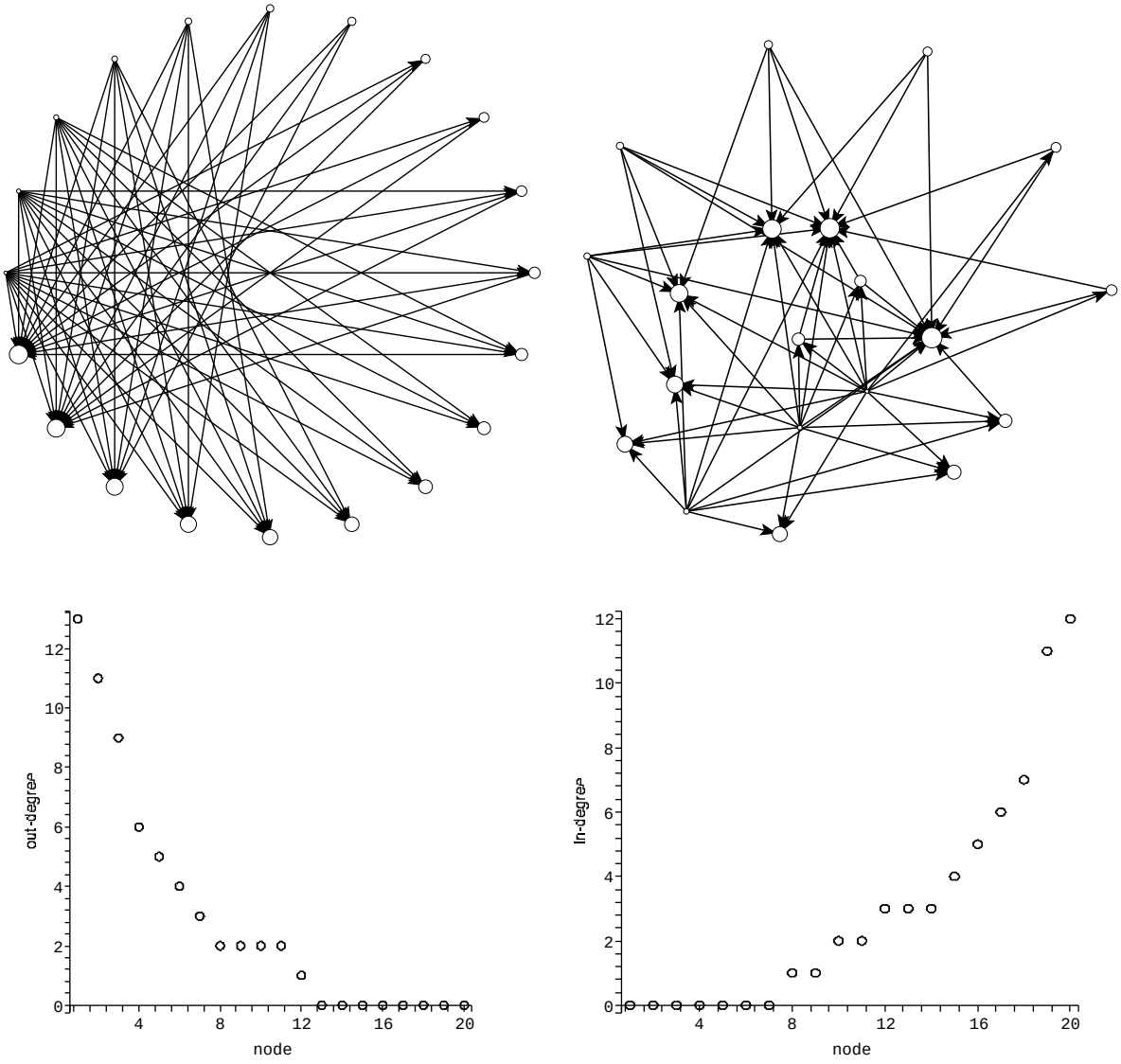

Figure 1: The top two figures depict the same network, a possible equilibrium network, where larger nodes denote higher content. The bottom graphs represent the number of out- and in-links for each node, where nodes are arranged in increasing order of content.

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, 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.[8] 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.[9] 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.

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 pages. This results in a network where the number of in-links correlates with the value of the corresponding site.

4 Endogenous prices and infinitely many sites

After analyzing network formation with per-visitor 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

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

[9] See, "The Decade in Online Advertising" and "The Online Advertising Landscape in Europe", DoubleClick, April/September 2005 as well as the list of top 50 advertisers online in Zeff and Aronson (1999), p.7.

is increasing in content. Our goal is to show that this is true even with endogenous prices. Specifically, we analyze a two-stage game where in the first stage, sites set per-visitor 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 3.2, 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.

4.1 Network formation

In the infinite version of the original network formation game, suppose that the set of players is the continuous 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 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-traffic prices. We can assume without loss of generality that $c(i)$ is increasing, i.e. sites are ordered by content on I. Then, the Page Rank 'function' is also directly generalizable for each i:

$$r(i) = (1-\delta) + \delta \int_{\{x \in I | (x,i) \in G\}} \frac{r(\lambda)}{d^{out}(\lambda)}. \tag{6}$$

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. Then, site i has the following utility function.

$$u_i = r_i(c(i) - C + q(i) \cdot d^{out}(i)) - \int_{\{j | j \to i\}} r(j)q(j). \tag{7}$$

For this continuous game, the main results that were valid for the discrete case still hold. If $q(.)$ is an increasing function of content, 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.

Proposition 2 *if $q(i)$ is increasing, and the functions c and p are differentiable, 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 $q' : [0,1] \to \mathbf{R}$ are equivalent ($q \sim q'$) if $\Lambda\{x | p(x) \neq p'(x)\} = 0$, that is, if they only differ in a small set.*

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

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. □

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.

4.2 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 $q \sim q'$, then $E(p) \sim E(p')$.*

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. Using this definition, 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-traffic 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. High-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.[10]

5 Extensions

In what follows, we explore two 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 explore the impact of search engines allowing sites to have multiple content areas.

5.1 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.[11] 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 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

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

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

(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:

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

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 us-

ing 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".[12] 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:

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

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.

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

[12] Ibid. See also "How Google works", *The Economists Technology Quarterly*, September 18, 2004.

the out-degrees are decreasing in content for advertising links and increasing for reference links.

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. Also, as can be readily seen from (9), the marginal benefit from a reference link with any given content level is higher for a high content site than for a low content site because the former has more traffic reaching it (f is higher). Furthermore, high content sites also have more resources. Because of these two reasons, high content sites will establish more 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.

5.2 Search engines and multiple content areas

Search engines (SE) play an important role in the formation of the network. 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 try to measure content directly, on the other hand, they measure content indirectly through the structure of the network, using methods such as Page Rank.

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.[13] 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 $p_i = q_i r_i$, where $q_i = q(\mathbf{w} \cdot \mathbf{c_i})$ is an increasing function of average content. The utility can be then rewritten as:

$$u_i = (\mathbf{w} \cdot \mathbf{c_i} - C + q_i d_i^{out}) r_i - \sum_{j \to i} q_j r_j. \tag{11}$$

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.[14]

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

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

What happens if we incorporate SEs in the model? Let us assume that only a proportion b of consumers is browsing according to the process described in Section 3.1. The remaining $(1-b)$ consumers use a SE in every step of browsing, which directs them to a Web site in the following way.[15] 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.[16] 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 $b\mathbf{w} \cdot \mathbf{c_i} + (1-b)\mathbf{w} \cdot \mathbf{C_i}$.

[15] We assume that all the SEs are fundamentally the same.

[16] This is consistent with practice as there are very few consumers who go beyond the 2nd page of Google's search results.

Then, the total utility of site i is

$$u_i = r_i(E_i - C) + p_i d_i^{out} - \sum_{j \to i} p_j, \tag{12}$$

where we assume that $p_i = q_i r_i$, where $q_i = q(.)$ is an increasing function of the modified average content, E_i. The utility can then be rewritten as:

$$u_i = (E_i - C + q_i d_i^{out}) r_i - \sum_{j \to i} q_j r_j. \tag{13}$$

Clearly, with a single content area, the existence of search engines 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 weighted average content in the following sense. If, for a given pair of nodes $E_k < E_l$, then $d_k^{out} \geq d_l^{out}$. If $E_k = E_l$ then $d_k^{out} + 1 \geq d_l^{out}$.

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

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

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_i is the linear combination of the average content of site i and the expected reward from SEs for offering leading content in particular dimensions, the proposition implies that in the presence of search engines 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 content. Thus, we would expect sites with low total content to specialize, while those with high general content to diversify.

6 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-traffic 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.

Our stylized model is limited in several ways. To concentrate on firms' network building strategies we had to reduce consumer search to a random browsing process. Also, we assumed a generic profit function across sites that only differed in terms of sites' content. In doing so, we also neglected an important aspect of advertising, the disutility that it represents for consumers. In a technical appendix, we tackle this problem and show that including advertising disutility does not change any of our result.

More importantly, one could ask: are our equilibrium network patterns consistent with empirical evidence? In the technical appendix, 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 the technical appendix, 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 and/or do not treat agents as utility maximizers. 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. There are many opportunities for both theoretical and empirical work to explore the drivers of its evolution.

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. Therefore using (1), we can transform (5) to

$$u_i = (1-\delta)\frac{c_i - C + q_i \cdot d_i^{out}}{n} + \sum_{j \to i} r_j \left(\frac{\delta(c_i - C + q_i \cdot d_i^{out})}{d_j^{out}} - q_j \right). \quad (14)$$

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 positive if the node decides to buy that link. Hence player i buys a link from player j if and only if

$$\frac{\delta(c_i - C + q_i \cdot d_i^{out})}{d_j^{out-i} + 1} - q_j > 0, \quad (15)$$

where d_j^{out-i} denotes the number of out-links from node j excluding the possible link to i. This equation can be transformed to

$$\delta(c_i - C + q_i \cdot d_i^{out}) > q_j(d_j^{out-i} + 1). \quad (16)$$

(16) shows that a node buys links from those nodes for which $q_j(d_j^{out-i} + 1)$ is the lowest. Therefore, in an equilibrium, if $q_k(d_k^{out} + 1) < q_l d_l^{out}$ 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 $d_k^{out} \geq d_l^{out}$. If $d_k^{out} + 1 \leq d_l^{out}$ then also $d_k^{out} < d_l^{out}$, which implies that $q_k(d_k^{out} + 1) \geq q_l d_l^{out}$, therefore $q_k \geq q_l$. Since $q_k = q(c_k) \geq q_l = q(c_l)$ and q is an increasing function, we have $c_k \geq c_l$. Thus $c_k < c_l$ implies $d_k^{out} + 1 > d_l^{out}$, that is, since out-degrees are integers,

$d_k^{out} \geq d_l^{out}$. Similarly, $c_k \leq c_l$ implies $d_k^{out} + 1 \geq d_l^{out}$, completing the proof of part (i) of the proposition.

In order to prove part (ii), we have to continue the above argument. Recall that $c_k < c_l$ implies that $d_k^{out} + 1 > d_l^{out}$, this must further imply that $q_k(d_k^{out} + 1) < q_l d_l^{out}$, that is, all the nodes who buy a link from node l must also buy from node k. We repeat inequality (16),

$$\delta(c_i - C + q_i \cdot d_i^{out}) > q_j(d_j^{out-i} + 1), \tag{17}$$

to recall the decision rule of a node. The left hand side defines a threshold for node i, deciding from which nodes to buy links. The number of links bought by node i depends on this quantity. The higher $\delta(c_i + q_i \cdot d_i^{out})$ is, the more links it buys. In the previous part we have proved that if $c_k < c_l$ then $d_k^{out} \geq d_l^{out}$, therefore $q_k d_k^{out} \leq q_l d_l^{out}$, hence it also implies that

$$\delta(c_k - C + q_k \cdot d_k^{out}) < \delta(c_l - C + q_l \cdot d_l^{out}).$$

Thus, 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

$$U_{j\to i}(w) = wr_j \left(\frac{\delta(c_i - C + q_i \cdot d_i^{out})}{d_j^{out} + w} - q_j \right) \tag{18}$$

denote the payoff of establishing link $j \to i$ with weight w for node i. One can check that the second derivative of $U_{j\to i}(w)$ is negative, hence $U_{j\to i}(w)$ is concave. Therefore the payoff function is also concave, since it is the sum of concave 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. □

Proof of Proposition 2:

First we prove that if $q(i)$ is increasing, then in any equilibrium, $d^{in}(i)$ is also increasing and $d^{out}(i)$ is decreasing. We proceed along the same lines as in the discrete case. As in the discrete case, a player i buys a link from player j if and only if

$$\delta(c(i) - C + q(i) \cdot d^{out}(i)) > q(j)d^{out}(j). \tag{19}$$

This shows that a node buys links from those nodes for which $q(j)(d^{out}(j)+1)$ is the lowest. Therefore, in an equilibrium, if $q(k)d^{out}(k) < q(l)d^{out}(l)$ 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 $q(k) \geq q(l)$. Since $q(i)$ is increasing, we have $k \geq l$. Thus $k < l$ implies $d^{out}(k) > d^{out}(l)$, therefore $d^{out}(i)$ is decreasing.

In order to prove that $d^{in}(i)$ is increasing, we have to continue the above argument. Since $k < l$ implies that $d^{out}(k) > d^{out}(l)$, this must further imply that $q_k d^{out}(k) < q_l d^{out}(l)$, that is, all the nodes who buy a link from node l must also buy from node k. We repeat inequality (19),

$$\delta(c(i) - C + q(i) \cdot d^{out}(i)) > q(j)d^{out}(j), \tag{20}$$

to recall the decision rule of a node. The left hand side defines a 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 $\delta(c(i) - C + q(i) \cdot d^{out}(i))$ is, the more links it buys. In the previous part we have proved that if $k < l$ then $d^{out}(k) \geq d^{out}(l)$, therefore $q_k d^{out}(k) \leq q_l d^{out}(l)$, hence it also implies that

$$\delta(c(k) - C + q(k) \cdot d^{out}(k)) < \delta(c(l) - C + q(l) \cdot d^{out}(l)).$$

Thus, the threshold increases as the content increases, therefore $d^{in}(i)$ is 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 (19), describing the decision rule of player i.

$$\delta(c(i) - C + q(i) \cdot d^{out}(i)) > q(j) d^{out}(j), \tag{21}$$

that is, player j sells links to the nodes that satisfy this equation. Therefore,

$$d^{out}(j) = \Lambda(i | \delta(c(i) - C + q(i) \cdot d^{out}(i)) > q(j) d^{out}(j)). \tag{22}$$

Let $L(j)$ denote the right hand side of equation (22), 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 apply Tikhonov's theorem to this continuous operator on the normed space of measurable functions on $[0, 1]$, that assigns $L()$ to the function d^{out}. The fixed point of this operator must satisfy (22), thus it represents an equilibrium of the game. □

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, $K(\psi) = \int_{\{j|j\to i\}} q(\lambda)$. Obviously, $K(\psi)$ is increasing and $D(\zeta)$ is decreasing. From the proof of Proposition 2 we can also see that $\zeta D(\zeta)$ is increasing. Then, rewriting the utility function, we have

$$u_i(\psi, \zeta) = r(\psi)(c(i) - C + \zeta D(\zeta)) - K(\psi). \tag{23}$$

Since $(q, E(q))$ is a refined SPNE, ζ and ψ has to maximize this function. However, the solution of the maximization problem (23) is increasing in i, because the function

$$u(i, \psi, \zeta) = r(\psi)(c(i) - C + \zeta D(\zeta)) - K(\psi)$$

has increasing differences in (i, ψ), (i, ζ), and (ψ, ζ). Therefore, the optimal ζ increases as i increases, yielding that $q(i)$ is increasing. □

Proof of Proposition 4:

One can see 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. 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 check that the payoff function has increasing differences in the pairs $(d_i^{in_A}, d_i^{out_R})$, $(d_i^{in_A}, i)$ and $(i, d_i^{out_R})$. 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 $d_i^{in_A*} \geq d_j^{in_A*}$. □

References

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.

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.

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.

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

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

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.

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.

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

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

Europe Campus
Boulevard de Constance,
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 6072 40 00
Fax: +33 (0)1 60 74 00/01

Asia Campus
1 Ayer Rajah Avenue, Singapore 138676
Tel: +65 67 99 53 88
Fax: +65 67 99 53 99

www.insead.edu

INSEAD

The Business School
for the World