MARKETING SCIENCE
Vol. 27, No. 5, September–October 2008, pp. 764–778
ISSN 0732-2399 | EISSN 1526-548X | 08 | 2705 | 0764

DOI 10.1287/mksc.1070.0349

Network Formation and the Structure of the Commercial World Wide Web

Zsolt Katona
Haas School of Business, University of California at Berkeley, Berkeley, California 94720-1900,
zskatona@haas.berkeley.edu

Miklos Sarvary
INSEAD, 77305 Fontainebleau, France, miklos.sarvary@insead.edu

We model the commercial World Wide Web as a directed graph that emerges as the equilibrium of a game in which utility maximizing websites purchase (advertising) in-links from each other while also setting the price of these links. In equilibrium, higher content sites tend to purchase more advertising links (mirroring the Dorfman-Steiner rule) while selling less advertising links themselves. As such, there seems to be specialization across sites in revenue models: high content sites tend to earn revenue from the sales of content, whereas low content ones earn revenue from the sales of traffic (advertising). In an extension, we also allow sites to establish (reference) out-links to each other and find that there is a general tendency to establish reference links to sites with higher content. Finally, we explore network formation in the presence of search engines and find that the higher the proportion of people using them, 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 online advertising, and the choice of Internet business models. They also shed light on why Google can use the web's link structure to rank sites by content.

Key words: Internet advertising; game theory; network formation
History: This paper was received November 2, 2005, and was with the authors 1 year and 4 months for 2 revisions; processed by J. Miguel Villas-Boas.

1. Introduction

The Internet and its most broadly known application, the World Wide Web (WWW), are gaining tremendous importance in our society. They represent 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, as a decentralized network, it evolves on its own based on its members' incentives and activities. In this paper we develop a model that helps us to understand what structure emerges from this decentralized network formation process.

The WWW includes an extremely broad community of websites 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 "subnetwork." 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' content is related to their connectedness on the Web. In turn, websites 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 reaching 125% in 2005.

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 in which each site is a potential advertiser and a potential seller of advertising, what determines the tradeoff

[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 *The Economist* 2006).

[2] In 2006, Internet advertising has reached $10 billion with a yearly growth rate of over 25% (see eMarketer Inc. 2006).

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," which sites establish to other sites 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) that 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 beyond buying and selling advertising links, in which 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 not, so far, 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-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, whereas 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 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 in which 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 the 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 noncooperative network formation in which 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 that 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 important, the above papers consider that individuals in the network are identical. For example, in Bala and Goyal (2000), linking to

[3] See also Jackson and Wolinsky (1996) for an early paper concerned with the relationship between social network stability and efficiency and Jackson (2004) for a recent summary of this literature.

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' content 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," and their respective generalizations.

Our work also relates to the vast literature on advertising (see Bagwell 2007 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. Advertising expenditures have also been shown to be affected by product quality in a variety of contexts. Nelson (1974) and Schmalensee (1978) develop 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. 2008). 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, to signal quality, or to affect brand loyalty. More important, in our model, advertisers and the media are not separate entities. Each site is a buyer *as well as* a seller of advertising. The central questions are: 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 to improve their own value as perceived by consumers. With these elements, we develop a model that is more consistent with the reality of the WWW than what is described by the existing network formation literature. This model is presented next.

3. The Model

We describe websites 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 website that may possibly include multiple pages. Technically, on the WWW, the nodes correspond to the Web pages. However, most of the time, a website 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* website. All the links going out of and coming in to a site's subpages should be assigned to this one node.[5] 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

[4] 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.

[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 subdirectories 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.

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 another. 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 effects 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 (PR), 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 site i, with probability δ, s/he randomly follows a link going out from that site or stays there, choosing each of these $d_i^{\text{out}}+1$ options with equal probability.[6] 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 $1/(1-\delta)$. 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 websites. This limit distribution is called PR.[7] It can be thought of as the number of visitors at a website per unit of time. By definition, PR has to satisfy the following equation:

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

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

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

Notice that the ith 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 website 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 nonzero 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 $[U]_{ij} = 1/n$, $rU = (1/n, 1/n, \ldots, 1/n)$. Hence (3) can be written as

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

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

[7] Although PR usually refers to the score that websites receive from Google, we use PR to describe the scores that are calculated of this simple version of the algorithm.

This formula helps interpret the meaning of PR 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.[8] Thus, PR 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 there are n nodes (sites) with given constants $c_1 \leq \cdots \leq c_n$, representing their contents. These content parameters can be thought of as some measure of the website's 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 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 in which 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 $\delta r_i/(d_i^{\text{out}}+1)$. Then, the total price of an advertising link from site i will be $p_i = \delta r_i q_i/(d_i^{\text{out}}+1)$.

If another node purchases a link, then this link will be created and point from the seller to the buyer. Given prices, nodes make simultaneous decisions about their incoming links, that is, which other nodes they buy 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 §4.2. 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 (16) in the appendix). In §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. Price-per-click increasing in content allows us to capture the basic tradeoff between keeping a consumer or handing him/her 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.[9]

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,

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

3.3. 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.*

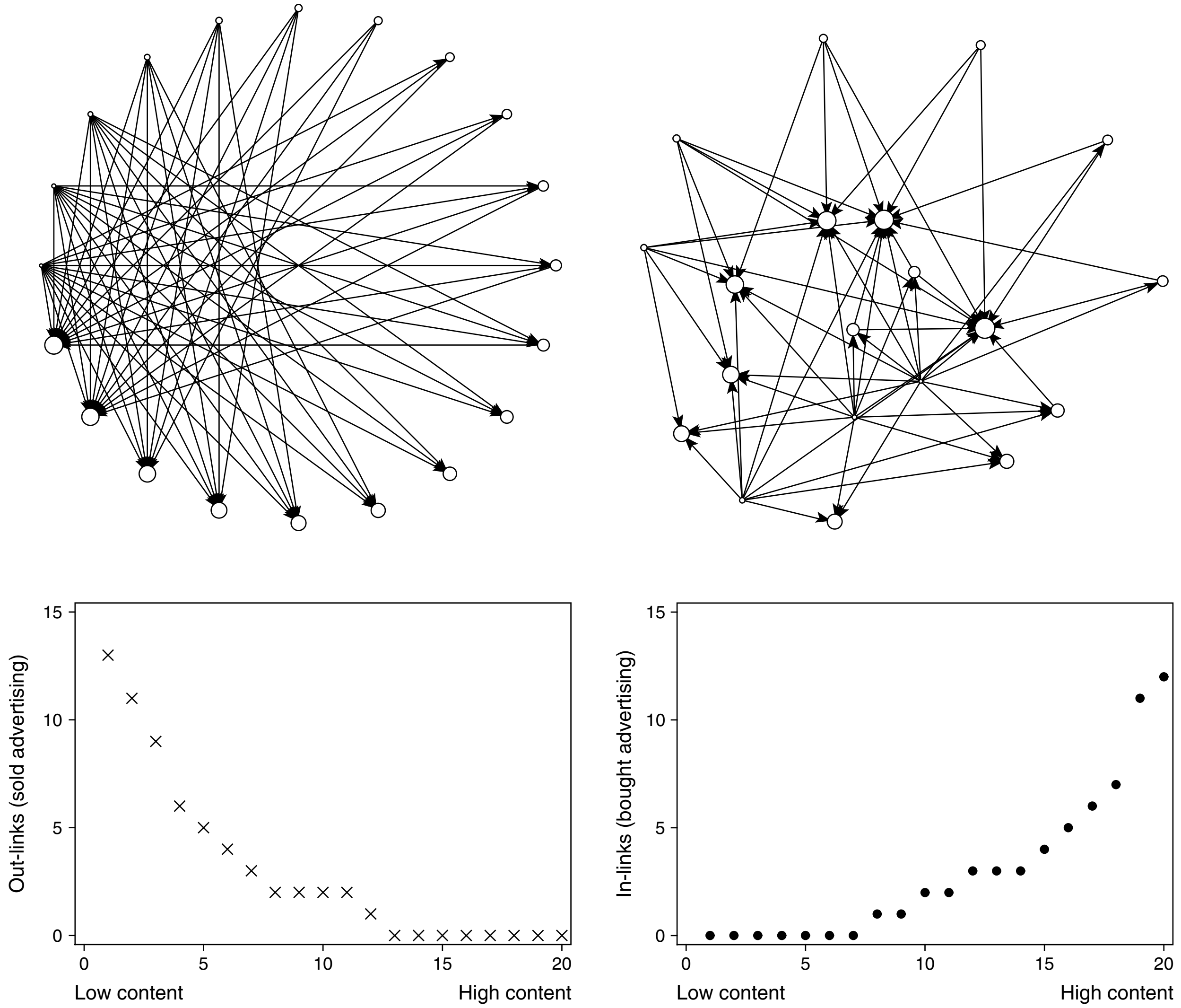
(i) *The out-degree is a weakly decreasing function of content in the following sense. If, for a given pair of nodes $c_k < c_l$, then $d_k^{\text{out}} \geq d_l^{\text{out}}$.*

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

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

[9] 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.

Figure 1 A Typical Equilibrium Network Structure



Notes. 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.

Proof (Sketch). Here we give the main logic of the proof and provide the detailed proof in the appendix. In the first step, we show that in equilibrium all the nodes buy links from the nodes with the lowest qs. This does not mean that they will buy from the nodes charging the lowest price for links but rather from those that sell their traffic at the lowest "per-click price." Based on the increasing price structure, these must be the sites with the 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 PR 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 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) so 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, 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.[10] 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.[11]

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), whereas 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) sites 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.

4. 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 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 §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 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, $d^{\text{out}}(x) = \Lambda\{y \in I \mid (x, y) \in G\}$ if the set is measurable; otherwise, the out-degree does not exist. Similarly, the in-degree of $y \in I$ is defined as $d^{\text{in}}(y) = \Lambda\{x \in I \mid (x, y) \in G\}$ 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-click prices. We can assume without loss of generality that $c(i)$ is increasing; i.e., sites are ordered by content on I. The PR 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 $d^{\text{out}}(i) = 0$. 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 $s/(s + d^{\text{out}}(i))$. Therefore, the equation defining PR will be

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

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 $p(i) = \delta r(i) q(i)/(d^{\text{out}}(i) + s)$. Then, site i has the following utility function:

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

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

[11] See *DoubleClick* (2005a, b), as well as Zeff and Aronson (1999, p. 7).

For this infinite game, the main results that were valid for the discrete case still hold. If $q(\cdot)$ is an increasing function of content and satisfies (16), 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, satisfying* (16), *and the functions c and q are continuous, at least one pure-strategy Nash equilibrium exists and in any equilibrium $d^{\text{in}}(i)$ is increasing and $d^{\text{out}}(i)$ is decreasing.*

Proof. See the appendix.

Because 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 $q'\colon [0,1] \to \mathbf{R}$ are equal almost everywhere ($q = q'$ a.e.) if $\Lambda\{x \mid q(x) \neq q'(x)\} = 0$, 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 $d^{\text{in}}()$ for q, there exists an equilibrium for q' with a $d^{\text{in}'}() = d^{\text{in}}()$ a.e.*

Proof. Let X denote the set $\{i \mid q(i) \neq q'(i)\}$. 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. Because the two-stage game may have several subgame perfect Nash equilibria (SPNE), even unreasonable ones, we will rule out some of them based on Lemma 2.

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

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

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

This definition makes sure that to any refined SPNE corresponds an SPNE and that 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 $q(\cdot)$ 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. 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.[12] The intuition behind the result is that sites with 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 becomes their customers, resulting in a higher average margin per visitor. In the second stage these sites purchase more advertising, because they can more effectively leverage the traffic they buy.

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 website of the advertiser. Here, we introduce another type of link that is commonly used in the noncommercial Web: reference links.[13] These links also have an important role in

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

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

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 to) 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, 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, and (ii) the sum of the content of sites linked to through reference links multiplied by a scaling constant $0 \le \beta < 1$. Therefore, the total payoff of node i is defined as follows:

$$u_i = r_i\left(c_i + \beta \sum_{i \rightarrow_R j} c_j - C\right) - \kappa d_i^{\text{out}_R} + p_i \cdot d_i^{\text{out}_A} - \sum_{j \rightarrow_A i} p_j. \quad (8)$$

Introducing the reference links makes the problem much more complex; because a site cannot control its traffic by buying the appropriate number of advertising links, the traffic is also affected by the incoming reference links. 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 $r_i = f(d_i^{\text{in}_R}, d_i^{\text{in}_A})$ 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 ($d_i^{\text{in}_A}$) and reference links ($d_i^{\text{in}_R}$). 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* (2006) 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 (p. 11)."[14] We also make the natural assumption that f has increasing differences in $d_i^{\text{in}_R}$ and $d_i^{\text{in}_A}$. That is, $f(x + h_1, y + h_2) - f(x, y + h_2) \ge f(x + h_1, y) - f(x, y)$ for any $x, y \ge 0$ and $h_1, h_2 \ge 0$; i.e., the two kinds of in-degrees are weakly complements. Then, the utility function becomes

$$u_i = f(d_i^{\text{in}_A}, d_i^{\text{in}_R})\left(c_i + \beta \sum_{i \rightarrow_R j} c_j - C\right) - \kappa d_i^{\text{out}_R} + p_i \cdot d_i^{\text{out}_A} - \sum_{j \rightarrow_A i} p_j. \quad (9)$$

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 $d_i^{\text{in}_R} \ge d_j^{\text{in}_R}$, $d_i^{\text{out}_A} \le d_j^{\text{out}_A}$, $d_i^{\text{in}_A} \ge d_j^{\text{in}_A}$, and $d_i^{\text{out}_R} \ge d_j^{\text{out}_R}$.*

Proof. See the appendix.

Keeping the assumption that prices are increasing with 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.

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 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 because 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

[14] See also *Economist Technology Quarterly* (2004).

had so much success introducing the quantity PR for searches. Google's objective is not only to find all the pages containing the search expression but also to rank them according to their content. As measuring content directly is difficult, it can use PR as an indirect measure because, according to our model, in equilibrium, high PR 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. If some consumers use SEs, then the number of visitors at a website depends not only 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 to display as the result of a search and in what order. 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 PR. 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 PR. 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.[15] 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 website from selling its content changes from r_ic_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

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

where we assume that $p_i = \delta q_i r_i/(d_i^{\text{out}}+1)$ 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, because only (weighted) average content matters.[16]

What happens if we incorporate SEs in the model? Let us assume that only a b proportion of consumers is browsing according to the process described in §3.1. The remaining $(1-b)$ consumers use an SE in every step of browsing, which directs them to a website 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 PR.[17] 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 an 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

$$br_ic_i^d + (1-b)r_ic_i^d\frac{I_i^d}{R_d} = r_ic_i^d(b + (1-b)I_i^d/R_d).$$

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

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

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

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

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

where $p_i = \delta q_i r_i/(d_i^{\text{out}} + 1)$ and $q_i = q(\cdot)$ 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 $d_k^{\text{out}} \geq d_l^{\text{out}}$.*

(ii) *If we suppose that all the modified average contents are different, then the in-degree and the PR are increasing functions of the modified 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 PR. Because 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 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; because 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. 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.

6. Discussion and Conclusion

We proposed to model the commercial WWW based on the idea that profit maximizing websites purchase (advertising) in-links from each other to direct traffic to themselves 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 relies heavily on the number of in-links to rank sites with respect to content. This can explain, for instance, why Google's PR 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 of 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.[18]

6.1. 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 important, we ignored consumer heterogeneity in preferences for content. Such heterogeneity could be of two kinds: vertical

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

and/or horizontal. With respect to the first, while we assume 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. 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 results. Another limitation is that sites are 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 and charge 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 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 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 websites' 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 toward understanding its structure. As discussed above, there are many opportunities for both theoretical and empirical work to further explore the drivers of its evolution.

Acknowledgment

The paper is based on the first chapter of the first author's dissertation.

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 PR values, we know how a node's rank is related to its in-neighbors' ranks. In particular

$$r_i = \frac{d_i^{\text{out}}+1}{d_i^{\text{out}}+1-\delta}\left(\frac{1-\delta}{n}+\delta\sum_{j\to i}\frac{r_j}{d_j^{\text{out}}+1}\right). \qquad (12)$$

Therefore, we can transform (5) to

$$\begin{aligned} u_i = {} & \frac{d_i^{\text{out}}+1}{d_i^{\text{out}}+1-\delta}(1-\delta)\frac{1}{n}\left(c_i - C + \delta q_i \frac{d_i^{\text{out}}}{d_i^{\text{out}}+1}\right) \\ & + \delta\sum_{j\to i} r_j \frac{1}{d_j^{\text{out}}+1} \\ & \cdot\left[\frac{d_i^{\text{out}}+1}{d_i^{\text{out}}+1-\delta}\left(c_i - C + \delta q_i \frac{d_i^{\text{out}}}{d_i^{\text{out}}+1}\right) - q_j\right]. \qquad (13) \end{aligned}$$

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 nonnegative. Hence, player i buys a link from player j if and only if

$$\frac{d_i^{\text{out}}+1}{d_i^{\text{out}}+1-\delta}\left(c_i - C + \delta q_i \frac{d_i^{\text{out}}}{d_i^{\text{out}}+1}\right) - q_j \geq 0. \tag{14}$$

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 $d_k^{\text{out}} \geq d_l^{\text{out}}$. Because $q_k = q(c_k) \geq q_l = q(c_l)$ and q is an increasing function, $c_k < c_l$ implies $d_k^{\text{out}} \geq d_l^{\text{out}}$, completing the proof of part (i) of the proposition.

To prove part (ii), we have to continue the above argument. Rearranging inequality (14), we get

$$T(i) := \frac{d_i^{\text{out}}+1}{d_i^{\text{out}}+1-\delta}(c_i - C) + \delta q_i \frac{d_i^{\text{out}}}{d_i^{\text{out}}+1-\delta} \geq q_j. \tag{15}$$

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

$$q_i \leq c_i \frac{\delta}{1-\delta}. \tag{16}$$

However, in §4.2, 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 also the set of nodes s/he buys links from contains that of the lower content nodes. Because PR is a linear combination of those pages a node buys links from, this ensures that PR 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

$$\begin{aligned} U_{j\to i}(w) = {} & w r_j \cdot \frac{1}{d_j^{\text{out}}+1} \\ & \cdot \left[\frac{d_i^{\text{out}}+1}{d_i^{\text{out}}+1-\delta}\left(c_i - C + \delta q_i \frac{d_i^{\text{out}}}{d_i^{\text{out}}+1}\right) - q_j\right] \end{aligned} \tag{17}$$

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, as it is the sum of these functions. Because we extended the strategy space, it is compact and convex. Also, the payoffs are continuous and concave in the players' own actions, and 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. □

Proof of Proposition 2. We begin by proving that if $q(i)$ is increasing and $q(i) \leq (\delta/(1-\delta))c(i)$, then in any equilibrium $d^{\text{in}}(i)$ is also increasing and $d^{\text{out}}(i)$ is decreasing. Similarly to the discrete case, player i buys a link from player j if and only if

$$\frac{d^{\text{out}}(i)+s}{d^{\text{out}}(i)+s(1-\delta)}(c(i)-C)+\delta q(i)\frac{d^{\text{out}}(i)}{d^{\text{out}}(i)+s(1-\delta)} \geq q(j). \tag{18}$$

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 $d^{\text{out}}(k) \geq d^{\text{out}}(l)$; therefore, $d^{\text{out}}(i)$ is decreasing.

To prove that $d^{\text{in}}(i)$ is increasing, we have to continue the above argument. We repeat inequality (18)

$$\begin{aligned} T(i) := {} & \frac{d^{\text{out}}(i)+s}{d^{\text{out}}(i)+s(1-\delta)}(c(i)-C) \\ & + \delta q(i)\frac{d^{\text{out}}(i)}{d^{\text{out}}(i)+s(1-\delta)} \geq q(j) \end{aligned} \tag{19}$$

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, $q(i) \leq (\delta/(1-\delta))c(i)$, 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, $d^{\text{in}}(i)$ will also be increasing.

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 locally convex topological vector space (X) and $f\colon X \to X$ is continuous, then f has a fixed point. Recall Equation (18), describing the decision rule of player i. Player j sells links to the nodes that satisfy $T(i) > q(j)$. Therefore,

$$d^{\text{out}}(j) = \Lambda(i \mid T(i) > q(j)). \tag{20}$$

Let $L(j)$ denote the right hand side of Equation (20), which is a measurable function if $d^{\text{out}}()$ is measurable. A function $d^{\text{out}}(j)$ satisfying $d^{\text{out}}(j) = L(j)$ 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, $\int_0^\infty |L(j)-L'(j)|\,dj = \int_0^\infty |T^{-1}(j)-T'^{-1}(j)|\,dq(j)$. Because $q(j)$ is continuous on a compact set, it has to bounded; therefore, $\int_0^\infty |T^{-1}(j)-T'^{-1}(j)|\,dq(j) \le c_2 \int_I |d(i)-d'(i)|\,di$—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 (20); 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 $d^{\text{in}}(i)$, 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 that 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) PR is

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

Decomposing this into two factors, let

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

denote the first factor and

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

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

$$u_i(\psi, \zeta) = r_2(\psi) r_1(\zeta) \left(c(i) - C + \delta\zeta \frac{D(\zeta)}{D(\zeta)+s} \right) - K(\psi). \quad (21)$$

Because $(q, E(q))$ is a refined SPNE, ζ and ψ have 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 ψs. This optimal $\zeta^*(i)$ is increasing in i, because the function

$$\begin{aligned} T(i,\zeta) &= r_1(\zeta) \left(c(i) - C + \delta\zeta \frac{D(\zeta)}{D(\zeta)+s} \right) \\ &= \frac{D(\zeta)+s}{D(\zeta)+s(1-\delta)} (c(i)-C) + \delta\zeta \frac{D(\zeta)}{D(\zeta)+s(1-\delta)} \end{aligned} \quad (22)$$

has increasing differences in (i, ζ), as 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

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

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

Proof of Proposition 4. We will show that the payoff function has increasing differences in the players' own decisions $(d_i^{\text{in}_A}, d_i^{\text{out}_R})$ and in the pairs composed of an own decision variable and another player's decision variable. Although (9) is not written as a direct function of other players' decisions, these are captured by $d_i^{\text{in}_R}$ and $d_i^{\text{out}_A}$. If another player buys more advertising links, $d_i^{\text{out}_A}$ either increases or does not change. If another player establishes an extra reference link, $d_i^{\text{in}_R}$ 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 $f(\cdot,\cdot)$, which has increasing differences in its variables by definition, the relevant terms are always products of functions that are increasing in the variables in question.

Therefore, the game is supermodular; hence, we can use the machinery introduced by Topkis (1998). In Chapter 4 he describes the characteristics of the equilibria. It follows from supermodularity that the pure-strategy equilibria of the game form a nonempty 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 selects 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 $d_i^{\text{in}_R} \ge d_j^{\text{in}_R}$ if $c_i > c_j$ and $d_i^{\text{out}_A} \le d_j^{\text{out}_A}$ 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.

Because 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: $d_i^{\text{in}_A}$ and $d_i^{\text{out}_R}$. It is easy to see that the payoff function has increasing differences in the pairs $(d_i^{\text{in}_A}, d_i^{\text{out}_R})$, $(d_i^{\text{in}_A}, i)$, and $(i, d_i^{\text{out}_R})$, checking the terms that contain two of the variables in question. Therefore, the optimal decisions $(d_i^{\text{in}_A *}, d_i^{\text{out}_R *})$ are increasing in i. That is, if $i > j$ (i.e., $c_i > c_j$), then $d_i^{\text{out}_R *} \ge d_j^{\text{out}_R *}$ and $d_i^{\text{in}_A *} \ge d_j^{\text{in}_A *}$. □

References

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

Bagwell, K. 2007. The economic analysis of advertising. M. Armstrong, R. Porter, eds. *Handbook of Industrial Organization*, Chapter 28. Elsevier, 1701–1844.

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. *Internat. J. Game Theory* **33** 1–19.

Brin, S., L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks 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. *Comput. Networks* **33** 309–320.

Dorfman, R., P. O. Steiner. 1954. Optimal advertising and optimal quality. *Amer. Econom. Rev.* **44** 826–836.

DoubleClick. 2005a. The decade in online advertising. (April).

DoubleClick. 2005b. The online advertising landscape in Europe. (September).

Dukes, A., E. Gal-Or. 2003. Negotations and exclusivity contracts for advertising. *Marketing Sci.* **22**(2) 222–245.

Economist. 2006. Dancing with Google's spiders. **378**(8468, March 11) 11–12.

Economist Technology Quarterly. 2004. How Google works. **372**(8393, September 18) 28–32.

eMarketer Inc. 2006. Marketing budgets are up 46% for Q2. (July 5), http://www.emarketer.com.

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

Godes, D., E. Ofek, M. Sarvary. 2008. Products vs. advertising: The impact of competition on media firm strategy. *Marketing Sci.* Forthcoming.

Hoffman, D. L., T. Novak. 2000. Advertising pricing models for the world wide web. D. Hurley, B. Kahin, H. Varian, eds. *Internet Publishing and Beyond*: *The Economics of Digital Information and Intellectual Property*. MIT Press, Cambridge, MA, 45–61.

Istratescu, V. I. 1981. *Fixed Point Theory, An Introduction*. D. Reidel Publishing Co., Dordrecht, The Netherlands.

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

Jackson, M. O. 2004. A survey of models of network formation: Stability and efficiency. G. Demange, M. Wooders, eds. *Group Formation in Economics: Networks, Clubs and Coalitions*, Chapter 1. Cambridge University Press, Cambridge, UK.

Jackson, M. O., A. Wolinsky. 1996. A strategic model of social and economic networks. *J. Econom. Theory* **71** 44–74.

Langville, A. N., C. D. Meyer. 2004. Deeper inside pagerank. *Internet Math.* **1**(3) 335–400.

Masson, R. T., R. Mudambi, R. J. Reynolds. 1990. Oligopoly in advertiser supported media. *Quart. Rev. Econom. Bus.* **30**(2) 3–16.

Mayzlin, D., H. Yoganarasimhan. 2006. Link to success: How blogs build an audience by promoting rivals. Working paper, Yale School of Management, New Haven, CT.

Nelson, P. 1974. Advertising and information. *J. Political Econom.* **81** 729–754.

Schmalensee, R. 1978. A model of advertising and product quality. *J. Political Econom.* **86**(31) 485–503.

Topkis, D. M. 1998. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ.

Villas-Boas, J. M. 2004. Communication strategies and product line design. *Marketing Sci.* **23**(3) 304–316.

Zeff, R., B. Aronson. 1999. *Advertising on the Internet*. John Wiley and Sons, New York.