Network Formation and the Structure of the Commercial World Wide Web

Zsolt Katona and Miklos Sarvary*

June 23, 2006

*Zsolt Katona is a Ph.D. student and Miklos Sarvary is Associate Professor of Marketing at INSEAD, Bd. de Constance, 77305, Fontainebleau, France. E-mail: zsolt.katona@insead.edu, miklos.sarvary@insead.edu. Tel: +33 1 60 71 26 05 Fax: +33 1 60 74 55 00.
Abstract

We model the World Wide Web (WWW) as a graph emerging as the equilibrium of a game in which utility maximizing Web sites (nodes) choose to establish reference links to and/or purchase advertising links from other sites. A key feature of our model is that we consider nodes to be heterogeneous in terms of their content, i.e. their general appeal to the public. We find that the equilibrium structure of such a graph matches the empirical structure of the WWW better than that proposed by other models. There is a general incentive to establish reference link to sites with higher content. Also, higher content sites tend to purchase more advertising links mirroring the Dorfman-Steiner rule. Overall, there is a strong positive correlation between a site’s content and the number of its in-links. Sites with higher content sell less advertising links and offer such links at higher prices, i.e. there seems to be specialization across sites in terms of revenue models. High content sites tend to earn revenue from the sales of content while low content ones from the sales of traffic (advertising links). These results have interesting practical implications with respect to ‘search-engine optimization’, the pricing of Internet advertising and the trade-off between selling traffic versus content. 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: graph formation, game theory, network with heterogenous agents.
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.\(^1\) The goal of this paper is to develop a model that helps understand what structure emerges from this decentralized network formation process and what incentives drive the behavior of its constituencies.

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 commercial WWW, 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 determines 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. Indeed, “search-engine optimization” has grown into a $1.25 billion business

\(^1\)The WWW is not the only such network and our findings relate to many other networks with similar characteristics, including citation networks or even the human nervous system.
with a growth rate of 125% in 2005.\(^2\) Similarly, sites also need to address the problem of driving traffic to themselves through the purchase of advertising. In fact, the Web provides multiple opportunities for each commercial site to leverage the acquired traffic. Beyond selling its content to the public, a site can decide to sell the traffic itself by selling advertising links to other sites. In this case, appropriate pricing of advertising is needed. These practical issues all require the understanding of the “forces” that drive the evolution of the network’s structure and the resulting competitive dynamics.

The purpose of this paper is to address these questions by building an analytic model that combines graph theory and game theory. Specifically, we propose a model in which the nodes represent rational economic agents (sites) who make simultaneous and deliberate decisions on the (reference) out-links they establish to and the (advertising) in-links they purchase from each other. Agents are strategic and heterogeneous with respect to their endowed “content”, which may be thought of as their inherent value in the eyes of the public/market. Consumer traffic to a site is related to the site’s connectedness in the network. Sites are assumed to be able to generate revenue from two sources: (i) selling their content to the public and (ii) selling links to other sites. First, we explore a model where each node establishes/purchases links simultaneously given the prices but later we also analyze the situation when sites also optimize the price they set for advertising links.

We find that our model provides an equilibrium network structure that matches significantly better the empirical reality of the WWW, than structures proposed by the existing literature. Specifically, 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. Furthermore, we also show how the number of established links depends on a node’s content and explain why famous search engines (such as Google, for instance) have so much success using ranking statistics (e.g. PageRank) for

ordering pages in their search function. Specifically, sites with higher content generally establish more reference links and also purchase more advertising links as is consistent with the Dorfman-Steiner rule. As a result, there is a strong positive correlation between a site’s content and the number of its in-links. This confirms the general search-engine heuristic that the number of in-links predicts content. We also find that sites with higher content sell less advertising links and offer such links at higher prices. This suggests that there is specialization across sites in terms of revenue models. High content sites tend to earn revenue from the sales of content while low content sites mostly benefit from the sales of traffic, i.e. the sales of advertising links.

The paper is organized as follows. Next, we review the relevant literature. Section 3 presents a model of network formation treating prices as parameters and describing the properties of the equilibrium graph. Next, in section 4, we explore a game where network formation is preceded by the price setting decision for advertising links. The paper ends with concluding remarks. To improve readability, longer 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 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. These models have several features, which do
not really apply to the WWW. First, they concentrate on the cost of link formation, which is shown to be critical for the outcome. More importantly, the above papers consider that individuals in the network are identical. For example, in Bala and Goyal (2000), linking to a well-connected person costs the same as connecting to an idle one. This is clearly not the case on the WWW, where large differences exist between the sites’ contents and their connectedness. Also, on the WWW the cost of establishing a link largely depends on the type of this link (e.g. an out link versus an in-link) as well as 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.

Outside the social sciences, Barabási and Albert (1999) were the first to introduce random network models to describe the structure of the Internet. They point out that many complex real-world networks cannot be adequately described by the classical Erdős-Rényi random graph model (Erdős 1947), where the possible links between nodes are included independently, with the same probability. Instead, Barabási and Albert suggest that starting with a small number of nodes, at every step we add a new node with fixed number of links ($m$) that connect this new node to $m$ different nodes already present in the system. To incorporate preferential attachment, they suggest that the probability of linking a new node to an old one be proportional to the old node’s degree, that is, the number of nodes that the old node is connected to.

The strength of this model is that it is consistent with data collected from several real-world networks (WWW, citation network, nervous system or directory trees) in one specific aspect, the so-called degree distribution. In particular, it was found that the degree distribution of these networks is a power-law degree distributions, that is, the proportion of nodes with a fixed degree $k$ is asymptotically $k^{-\xi}$, with an exponent of around $-3$ for
most of the undirected networks. However, in the case of the WWW the exponent is around \(-2\) and only slightly differ for the in- and out-degrees (Barabási and Albert 1999, Broder et al. 2000, Faloutsos et al. 1999, Katona 2005). Figure 1 illustrates such a degree distribution for the Hungarian Web (Benczúr et al. 2003). Roughly speaking, this means that the proportion of nodes reached from a large number of other nodes tends to be quite small, i.e. the outcome of the network formation process is a “mild” winner-take-all situation. Theoretical work, extending the Barabási and Albert model (Bollobás et al. 2001, Cooper and Frieze 2003, Katona and Móri 2006) provide evidence that this model is robust and its generalizations result in the same degree distribution.

Although the Barabási and Albert model is consistent with data in one respect, it has important shortcomings. First, as Bala and Goyal (2000),
it treats the network members as identical, which is a rather unrealistic assumption in the context of Web-sites. A model with homogeneous network members and thus a fixed number of out-links cannot say much about the distribution of out-degrees in the graph. Empirical evidence shows that this distribution is similar to the one for in-degrees but the Barabási and Albert model cannot predict this. Also, a fixed number of out-links means that the total number of links in the network (i.e. the density of the graph) is exogenous. This is clearly an unreasonable assumption. Finally, and most importantly, this model is ‘ad hoc’ in the sense that it does not consider the network members as economic agents acting deliberately in their own interest. Why would nodes (or agents in the nodes) select to establish links one after the other without any interaction? Why would they consider solely the degrees to make their decisions? From a marketing perspective the interesting question is: what incentives drive agents’ choices of the nodes and how these choices depend on their inherent characteristics (e.g. their content)? Beyond describing the structure of the Web, our goal is to understand these incentives, i.e. the mechanism of the network formation process. A model to achieve these goals is presented next.

3 The model of network formation

Assume $n$ nodes (Web-sites) with given constants $0 < c_1 < \ldots, < c_n$, representing the content of the nodes. These content parameters can be thought of as some measure of the sites’ value for the public. As such, we assume that the content parameters refer to a broad content domain (e.g. travel, books or entertainment), i.e. they are comparable between each other. It is also important to precise 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. Beyond structural reasons, another reason to consider sites as the unit of analysis is that they represent a single decision maker.

Let $N$ denote the set of players, who manage the sites and decide simultaneously which links to establish with other sites. We consider two types of (directed) links. Reference links, denoted by $\rightarrow_R$, are used to increase the referring sites’ content with the help of the referred pages. The number of reference links going out from a site is denoted by $d_{out}^R$. Every node is allowed to establish one reference link from itself to every other node at maintenance cost $\kappa$. The second type of links are advertising links, denoted by $\rightarrow_A$, that are meant to increase the buying sites’s popularity. Each site can buy one advertising link from every other site if it pays the advertising fee, denoted $p$, of the seller. The number of incoming advertising links is denoted by $d_{in}^A$. In this section, we assume that these prices are fixed albeit different across sites. Later, we will consider a two-stage game with price setting first, followed by network formation.

Thus, for this stage, the strategy of player $i$ can be described by two vectors, each consisting of 0’s and 1’s. The first vector $x_i^R$ 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 $x_i^A$ 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

---

3 Most of the time when Google calculates the rank of a page in its search function, 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.
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 nodes together with all the links (both types) form a directed graph $G$, with the vertices of the graph corresponding to the Web-sites and the directed edges corresponding to the links between them. Let $i \rightarrow_R j$ denote if there is a reference link (edge) from $i$ to $j$ and $i \rightarrow_A j$ if there is an advertising link (edge) between them. We can refer separately to the graph formed by the reference links by $G^R$ and to the graph formed by the advertising links by $G^A$.

The total utility (payoff) of node $i$ is defined as follows.

$$u_i = f(d_{inA}^i, d_{inR}^i) \left( c_i + \beta \sum_{i \rightarrow R j} c_j - \gamma d_{outA}^i \right) - \kappa d_{outR}^i + p_i \cdot d_{outA}^i - \sum_{j \rightarrow A i} p_j. \quad (1)$$

In (1), $f(d_{inR}^i, d_{inA}^i)$ is site $i$’s popularity, i.e. 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 ($d_{inA}^i$) and reference links ($d_{inR}^i$). 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”.\(^4\) We also make the natural assumption that $f$ has increasing differences in $d_{inR}^i$ and $d_{inA}^i$. 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.

The second term in brackets is the site’s “accumulated” or “effective” content. It consists of three 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 \( 0 \leq \beta < 1 \) and (iii) the disutility of advertising links, i.e., the advertising out-degree multiplied by a constant \( \gamma \geq 0 \). The constant \( \beta \) describes the efficiency of reference links while \( \gamma \) measures the negative effect of advertisement on the site. For simplicity, we assume that \( \gamma \) is small enough that the accumulated content of a site is always at least zero.\(^5\)

\( \kappa \) is the maintenance cost of reference links that we assume to be the same for all reference links, and sufficiently small as is typically the case in practice.\(^6\) We assume that every node offers links for a fixed price, \( p_i \), which may be different for different nodes. Obviously, a lot depends on how the price of advertising of a given node is defined. In this section, we treat these prices as parameters, hence most of the results we obtain are valid for any price setting. In Section 4.2 we examine a simplified problem, where the sites make decisions both about their links and prices.

In sum, this formulation assumes that a node’s source of income is twofold. First, it generates income by selling its content to consumers. This income is defined as its popularity multiplied by the node’s accumulated content - see first term in (1). Second, a node sells advertising links to other pages for which it receives a fee - third term in (1). A node’s expenses are also twofold. It pays a maintenance cost for its reference links (second term in (1)) and an advertising fee for the advertising links (last term in (1)). At first glance one can see that every site wants to increase its in-degree and content but advertising has a price and maintaining reference links has a cost. The strategic interaction between sites occurs through two factors. First, when

\(^5\)If we want to allow a larger \( \gamma \), we can assume that \( f(.,.) \) is zero whenever the accumulated content goes below zero.
\(^6\)See the proof of Proposition 3 for a more specific condition on \( \kappa \).
a site receives reference links from other sites, this has a positive effect on popularity. Second, selling advertising can have a negative effect on the site’s accumulated content.

We are looking for the pure strategy Nash-equlibria of this game. Notice that if a player makes the decision about the number of reference links s/he wants to establish and the number of advertising links s/he wants to buy, it is straightforward to select the highest content nodes for reference links and the lowest priced nodes for advertising links. Therefore, there is a one to one mapping between the set of equilibria of this game and the restricted game where the players are only allowed to select the number of reference links and the number of advertising links they want to establish. We will analyze this restricted game from this point. Now, the strategies of player $i$ can be described by the vector $(d_i^{\text{outR}}, d_i^{\text{inA}})$. Let $(N, S, \{u_i|i \in N\})$ denote this modified game, where $S = \times_{i \in N} S_i$ and $S_i$ denotes the strategies of player $i$ consisting of the 2-vectors described above. For the sake of simplicity we allow the degrees to be non-integers, that is $S = [0, N]^{2N}$. A partial link results in benefits and costs proportional to the weight of the link. Thus, the utility function will be:

$$u_i = f(d_i^{\text{inA}}, d_i^{\text{outR}}) \left( c(i) + \beta C(d_i^{\text{outR}}) - \gamma d_i^{\text{outR}} \right) - \kappa d_i^{\text{outR}} + p_i \cdot d_i^{\text{outA}} - P(d_i^{\text{inA}}),$$

where $c(x) = c_i$ if $i - 1 < x \leq i$ for any $0 < i \leq N$ integer. The function $C(x)$ denotes the total content of the $x$ highest content nodes, that is, $C(x) = \int_x^n c(y)dy$. It is strictly increasing and concave. The function $P(x)$ denotes the total price of the $x$ cheapest nodes.\(^7\) If all the prices are different, then $P(x)$ is strictly increasing and convex. Although the functions $C()$ and $P()$ are not twice continuously differentiable they can be approximated by such functions. For the sake of simplicity we will assume that they are twice continously differentiable, strictly increasing and strictly concave (convex) functions.

\(^7\)First we define it for integers, then we extend it linearly.
The equilibria of this modified game also represent a graph and our main interest is understanding the structure of this graph. First, we will show that any equilibrium structure has certain properties regarding its degree distribution.

3.1 Degree distribution

It is important to confront our model to empirical evidence about the degree distribution of the WWW. An important empirical pattern about the Web, is that it has a scale-free power-law degree distribution, that is, the ratio of nodes with degree \( k \) is \( P(k) \sim k^{-\xi} \), where \( \xi \) is around 2. More importantly, this empirical observation stands for both in- and out-degrees. Previous models, could claim this finding for in-degrees only but not for out-degrees.

In our model it is more appropriate to study the degree distribution using the ratio of nodes with degree "at least \( k \)". The connection between these two quantities can be easily established: a ratio of \( k^{-\xi} \) in the former case yields a ratio of \( k^{-\xi+1} \) in the latter case (i.e. with degrees at least \( k \)). Let us denote the number of nodes with in-degree at least \( k \) by \( N(k) \) and the number of nodes with out-degree at least \( l \) by \( M(l) \). First, we define a class of graphs, which have the special property that reference links go to the highest content nodes and advertising links go to the lowest priced nodes.

**Definition 1** A directed graph \((G,V,E)\) is said to be oriented if vertices can be ordered such that for a given node \( v_i \) and for a given pair \( k < j \) the existence of a directed edge \((v_i, v_k)\) implies the existence of the directed edge \((v_i, v_j)\).

The definition captures the phenomenon, that every player has the same preference in ordering the other sites when deciding where to establish a reference link. Although the definition considers out-links (from \( v_i \)), the reverse would result in an equivalent definition. That is, the nodes can
be numbered such that the existence of a directed edge \((v_i, v_l)\) implies the existence of the directed edge \((v_j, v_l)\), if \(j < i\). The following proposition shows that these graphs have special degree distributions. Let \(\pi\) denote a permutation such that \(d_{out}(v_{\pi(l)}) \geq d_{out}(v_{\pi(j)})\) if \(l < j\).

**Proposition 1** For any oriented graph, \(M(N(k)) = k\) for any \(k\) such that \(d_{out}(v_{\pi(k)}) > d_{out}(v_{\pi(k+1)})\), that is, \(M(.\, ) = N^{-1}(.,\, )\), where the inverse is well defined.

**Proof:** See the Appendix.

If we consider any equilibrium graph of our model, both the graph formed by the advertising links and the graph formed by the reference links are oriented and the corresponding ordering is the same as ordering by content for reference links and the same as ordering by price for advertising links. As a consequence Proposition 1 holds for both of the graphs. Specifically, it follows from the proposition that if one of the in- and out-degree distributions is power-law scale free with exponent \(\xi\) then the other one is also power-law scale-free with exponent \(1 + \frac{1}{\xi-1}\). In particular, the in- and out-degree distributions are identical if one of them is a power-law distribution with exponent \(\xi = 2\). This is an important result as it is consistent with patterns found in the real WWW: the exponent of the in- and out-degree distributions are around two and only slightly differ (Broder et al. 2000, Faloutsos et al. 1999). To our knowledge, our model is the first to exhibit this characteristic, providing external validity for it.

### 3.2 No advertising disutility

Beyond the distribution of degrees, we also want to describe the nature and relationship of links to the sites' contents. For the sake of simplicity we first assume that advertising does not decrease the accumulated content, i.e. there
is no advertising disutility. To do so, we set $\gamma = 0$ in (1). In this case, the
game is supermodular.

**Lemma 1** If $\gamma = 0$, then the game $(N, S, \{u_i | i \in N\})$ is supermodular.

**Proof:** One can see that the payoff function has increasing differences in
the players’ own decisions ($d_i^{inA}$, $d_i^{outR}$) and in the pairs composed of an own
decision variable and another player’s decisions variable.

In this case, 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 special structural properties.

Since every node refers to the highest content nodes and buys links from
the cheapest nodes, obviously we have $d_i^{inR} \geq d_j^{inR}$ if $c_i > c_j$ and $d_i^{outA} \leq d_j^{outA}$
if $p_i > p_j$. Thus, we can show, that in equilibrium, the actions of players are
ordered with respect to their content.

**Proposition 2** If $\gamma = 0$, then in any equilibrium of the game $(N, S, \{u_i | i \in N\})$, if $c_i > c_j$ then $d_i^{inA} \geq d_j^{inA}$ and $d_i^{outR} \geq d_j^{outR}$.

**Proof:** See the Appendix.

The result in Proposition 2 is quite interesting. First, it says that if a
site has large content then it is likely to buy more advertising links. This
is intuitive as in our model content is a proxy for “margin”. A site with
higher content benefits more from a marginal increase in traffic than a site
with lower content. This effect is even more pronounced by the fact that -
as also stated by the Proposition - a high-content site’s effective content is
further increased relative to low content sites’ because it tends to establish
more reference links as well. In other words, this first result is equivalent to
the Dorfman-Steiner advertising rule well-known for traditional media and
generalized here for a network medium as well.\footnote{We would like to thank
the Area Editor for this insight.}

Notice also, that there is a strong positive correlation between a site’s
popularity and content. This is natural but what is interesting is that pop-
ularity can be gauged from the equilibrium network structure. Specifically,
sites that in equilibrium have a large number of in-links (both reference and
advertising) also have higher content, i.e. represent more value to the pub-
lic. This is important because it explains why popular search engines like
Google are so successful with their search heuristic that is primarily based
on measuring a site’s in-links to gauge its content.

Finally, the second result in Proposition 2, namely that high-content sites
tend to establish more reference links is also intuitive. High content sites
generate more traffic (have more in-links) and, as a result, benefit more from
a unit increase in their content. As such they have more incentive to establish
reference links.

Our analysis so far reveals the relationships between a site’s content and
its in-links (reference and advertising) as well as its reference out-links. We
also know that sites that sell the most advertising (out-)links are the ones
that sell advertising at the lowest price. However, in order to identify which
are these sites in terms of content we need to have a further assumption on
prices.

\subsection{Price is increasing in content}

If advertising links hurt the accumulated content of the site selling them, i.e.
if $\gamma > 0$, then the game is not supermodular anymore. However, assuming
that the prices at which sites offer advertising links increases with content,
every equilibrium graph still possesses the properties shown previously. More
precisely, let us assume that $p_i = p(c_i)$, where $p(.)$ is a strictly increasing positive function. In Section 4, we will show that this relationship holds even if prices are endogenous. However, even as an assumption, a price function increasing in content makes sense. In our model, we do not explicitly describe consumers’ behaviour 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 popularity of the site and its effective 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, a price function increasing in content captures the tradeoff between the sites’ two revenue streams.

With this assumption every node still refers to the highest content nodes and buys links from the cheapest ones. However, we know that the cheapest sites are the ones with the lowest content, hence we now have $d_{i^{in}} \geq d_{j^{in}}$ and $d_{i^{out}} \leq d_{j^{out}}$ if $c_i > c_j$. Moreover, we can still show the following.

**Proposition 3** If $p_i = p(c_i)$, then the game $(N, S, \{u_i| i \in N\})$ has an equilibrium, and in any equilibrium, if $c_i > c_j$ then $d_{i^{in}} \geq d_{j^{in}}$ and $d_{i^{out}} \leq d_{j^{out}}$.

**Proof:** See the Appendix.

In summary, assuming exogenous prices for advertising links that are increasing in the sites’ content we found a relatively simple equilibrium network structure. This structure becomes apparent if we order sites according to their content as is done in Figure 2. In equilibrium, advertising links tend to go from low content sites to high content sites, suggesting that sites tend to specialize in their revenue model. Low content sites tend to generate revenue from the sales of their traffic while high content sites generate returns from their content. With respect to reference links, we find that most reference
Figure 2: The two figures on the top depict the same sub-network formed by the advertising links in a possible equilibrium, whereas the bottom figures show the reference links. Larger nodes denote higher content.
links are between high content sites only. Indeed, for reference links both in- and out-links increase in the sites’ content. To the extent that reference links themselves increase content, it seems that content is fairly concentrated in the equilibrium network.

4 Endogenous prices and infinitely many sites

After analyzing network formation with 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 prices are 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 prices for advertising links and in the second stage they establish links between each other given prices. For the sake of simplicity we ignore reference links and concentrate on advertising links, where prices have a direct effect. Thus, we set $\beta = 0$ assuming that reference links have no benefit. Given the positive maintenance cost, $\kappa$, in equilibrium, no site will establish any reference links. Hence we consider the simplified game $(N, S', \{u'_i | i \in N\})$, where players only make decisions about how many advertising links they want to buy and from which sites. The payoff function of site $i$ is now:

$$u'_i = f(d'^{in}_i)(c_i - \gamma d'^{out}_i) + p_i \cdot d'^{out}_i - \sum_{j \to A_i} p_j,$$

with $f(d'^{in}_i) = f(d'^{in}_i, 0)$ increasing in $d'^{in}_i$.

Since the second stage of the game may have several equilibria and it is not supermodular, we cannot select a “greatest” (e.g. a Pareto-optimal) equilibrium. As such, it is not appropriate to consider this as a second stage of the game. 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 structure next.

4.1 Network formation

In the infinite version of the \((N, S', \{u'_i | i \in N\})\) 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 2** 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 \(\Lambda\).

**Definition 3** 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 \(d^{\text{out}}(x) = \Lambda\{y \in I | (x, y) \in G\}\) if the set is measurable, otherwise the out-degree does not exist. Similarly, the in-degree of \(y \in I\) is defined as \(d^{\text{in}}(y) = \Lambda\{x \in I | (x, y) \in G\}\) if the set is measurable.

Directly generalizing \((N, S', \{u'_i | i \in N\})\), we assume that the measurable function \(c(i)\) provides the content of site \(i \in I\) and the measurable function \(p(i)\) represents prices. We can assume without loss of generality that \(c(i)\) is increasing and site \(i\) has the following utility function. To make sure that players are not indifferent between different choices, we assume that \(\Lambda(p^{-1}(x)) = 0\) for every \(x\), that is, not many sites have the same price.

\[
u_i^p = f(d^{\text{in}}_i)(c(i) - \gamma d^{\text{out}}_i) + p(i) \cdot d^{\text{out}}_i - \int_{\{j: j \to i\}} p(\lambda).
\] (4)
Every site selects its own in-degree and the graph is formed by establishing links from the lowest price nodes. That is, a link is established from every point of the set $F(i) := p^{-1}([0, q(d^{in}(i))])$ to node $i$, where $q(x)$ is defined such that $\Lambda(p^{-1}([0, q(x)])) = x$. Thus, the graph $G = \{(x, y)| x \in F(y)\}$ is formed by this procedure. The game is denoted by $(I, S^I, \{u^i_p| i \in I\})$ if the price function is given by $p$.

**Lemma 2** If $d^{in}()$ is a measurable function then, in the graph $G$, $d^{out}(i)$ exists for every $i$ and it is also a measurable function.

**Proof:** By definition $d^{out}(i) = \Lambda(j|d^{in}(j) \geq r(i))$, where $r(i) := \Lambda\{x|p(x) \leq p(i)\}$, hence $d^{in}(i)$ exists. Furthermore, for any $j$, the set $\{x|d^{out}(x) < j\} = \{y|r(y) > \Lambda\{x|d^{in}(x) \geq d^{in}(j)\}\}$ is measurable, hence the function $d^{out}()$ is also measurable. 

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 4** Two measurable functions $p$ and $p': [0, 1] \rightarrow \mathbb{R}$ are equivalent ($p \sim p'$) if $\Lambda\{x|p(x) \neq p'(x)\} = 0$, that is, if they only differ in a small set.

**Lemma 3** If $p \sim p'$ then the set of equilibria of the games $(I, S^{I}, \{u^i_p| i \in I\})$ and $(I, S^{I}, \{u^i_{p'}| i \in I\})$ are equivalent in the two cases, that is, for any equilibrium function $d^{in}()$ in one case there exists an equilibrium in the other case with a $d^{out}() \sim d^{in}()$.

**Proof:** A node’s decision only depends on others’ prices but the set of sites with a price difference between the two cases is a null set. Therefore the payoffs do not change for those sites, who have the same price. For those

---

9 Notice, that if $p$ is increasing, then the set $F(i)$ is simply $[0, d^{in}(i)]$. 

---

19
who have different prices, the payoff might change but the optimal decision does not since the maximization does not depend on their own price.

For this continuous game, the main result that were valid for the discrete case still hold. If price is an increasing function of content, there always exists an equilibrium and in this equilibrium, in-degree is increasing in content (and in \( i \)). The next proposition formally states this, together with the existence of the equilibrium.

**Proposition 4** If the functions \( f, c, p \) are differentiable, at least one pure-strategy Nash-equilibrium exists. Furthermore, if \( p(i) \) is increasing, then in any equilibrium, \( d^m(i) \) is also increasing.

**Proof:** For the proof of the existence of the equilibrium, see the Appendix. The proof of the second part is the same as in the discrete case, since the utility function has increasing differences in \( i \) and \( d_i^m \).

Next, we need to show that prices increase in content.

### 4.2 Price setting

In the first stage, every site selects its price simultaneously, only knowing the content function. Since the second stage might have several equilibria, we have to select an equilibrium for any price function. We assume that selecting such an equilibrium is possible although we do not precise the selection mechanism. Let \( E(p) \) denote this “focal” equilibrium. \( E \) can be an arbitrary operator assigning an equilibrium to every measurable price function with the only restriction that \( E(p) \sim E(p') \) if \( p \sim p' \). \( E \) should be thought of as a “consensus” between players, who expect that given the price function \( p \), the outcome of the network formation game will be \( E(p) \). Let \( d_p^m() \) denote the in-degree and \( d_p^{out}() \) the out-degree function in that equilibrium. Notice
that the equivalency class of the focal equilibrium of the second stage does not depend on a single site’s price decision if the others’ prices are fixed.

**Proposition 5** For any potential equilibrium of the two-stage game, the first stage’s price function has to be increasing.

**Proof:** Consider site \( i \in I \) and for the moment let \( \pi \) denote \( p(i) \) and \( \delta \) denote \( d^\text{init}(i) \). Let \( D(\pi) \) be the aggregate demand for out-links, that is \( D(\pi) = d^\text{out}_p(p^{-1}(\pi)) \) and \( K(\delta) \) the cost of \( \delta \) links, that is, \( K(\delta) = \int_{\{j|j \to i\}} p(\lambda) \). Obviously \( K() \) is increasing and \( D() \) is decreasing. Every site maximizes

\[
    u_i(\delta, \pi) = f(\delta)(c(i) - \gamma D(\pi)) + \pi \cdot D(\pi) - K(\delta),
\]

in \( \delta \) for any \( \pi \) and then in \( \pi \). This is equivalent to maximizing in \( \delta \) and \( \pi \) simultaneously. The solution of the maximization problem (5) is increasing in \( i \), because the function

\[
    u(i, \delta, \pi) = f(\delta)(c(i) - D(\pi)) + \pi D(\pi) - K(\delta)
\]

has increasing differences in \((i, \delta), (i, \pi), \) and \((\delta, \pi)\) hence both coordinates of the solution are increasing in \( i \). \( \square \)

The significance of Proposition 5 is that it supports our assumption that in the network formation stage of the game, the 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.
5 Conclusions

In this paper, we have analyzed a model where sites act as economic agents and establish links between each other to maximize their benefit. Such benefit was assumed to originate from two sources: (i) the sales of the site’s content to the traffic generated at the site and (ii) revenues from the sales of advertising to other sites. We considered two types of links between sites: (i) low cost reference links that were meant to increase the initiating site’s effective content and (ii) advertising links that could be purchased from other sites to increase a site’s popularity or traffic. We were interested in the equilibrium network structure of such a game. In our first analysis, we considered exogenous prices for advertising links assuming that such prices are increasing in content. Subsequently, we showed that in equilibrium, prices will indeed have this characteristic.

Our main results reveal that such a network formation model results in an equilibrium network structure that is consistent with the observed link structure of the World Wide Web. Specifically, our model is the first to show why in- and out-links follow the same degree distribution. Beyond consistency with the general empirical pattern of the Web, our findings also reveal a number of network characteristics that may be interesting for managers of commercial Web-sites. We find for example, that sites with higher content will have more in-links and that this feature is true for advertising as well as reference links. This feature of the network explains why search engines (e.g. Google’s PageRank system) have so much success using in-links as a proxy for ranking sites in terms of content relative to each other. The fact that sites with higher content purchase more advertising links is also interesting. This finding is parallel to the Drofman-Steiner advertising rule generalized for the case of a network, where advertising requires the purchase of a link.

We also found that higher content sites are more likely to establish reference links to other sites. As the primary purpose of such links is to increase
the content (relevance) of the initiating site, our model suggests that most content on the Web seems to be concentrated on a few interlinked sites, while the large majority of sites with low content tend to be isolated. Finally, we found that in equilibrium, sites seem to specialize in terms of their revenue models. Low content sites mainly sell advertising links, while high content sites tend to increase their traffic by the purchase of advertising to better leverage their content. These findings have important implications for managers of Web-sites who face strategic questions related to the choice of a revenue model, the setting of prices for advertising links or search-engine optimization.

Our model has many limitations. We could not explicitly model consumer behavior on the Web. We had to assume a generic response function assuming that consumers flow on the WWW according to its link structure. In another paper, we elaborate further on the flow of consumers on the network but we doubt that a model considering consumers as economic agents can be solved for an equilibrium network structure. We have also used a reduced form profit function for sites. Again, while this limits our results, it was necessary to provide a parsimonious description of sites’ objectives.

Our simple model can be extended in several ways. Among many possibilities, it would be interesting to explicitly include the presence of search engines. Also, we assumed that sites represent only a single content domain, which is clearly not true in many cases. This assumption could be relaxed to examine the tendency for sites to specialize in content. Finally, one could examine how the network of the commercial Web interacts with the traditional economy. We have left these and many other interesting questions for future research, including many opportunities for empirical work.
Appendix

Proof of Proposition 1

The number of nodes with out-degree at least \( k \) is equal to the \( k \)-th largest out-degree, that is, \( N(k) = d_{\text{out}}(v_{\pi(k)}) \). If \( d_{\text{out}}(v_{\pi(k)}) > d_{\text{out}}(v_{\pi(k+1)}) \), then obviously \( M(d_{\text{out}}(v_{\pi(k)})) = k \), that is, \( M(N(k)) = k \), otherwise \( M(N(k)) \geq k \).

Proof of Proposition 2

In equilibrium, each site maximizes its utility

\[
 u_i = f(d_i^{\text{inA}}, d_i^{\text{outR}}) \left( c_i + \beta C(d_i^{\text{outR}}) \right) - \kappa d_i^{\text{outR}} + p_i \cdot d_i^{\text{outA}} - P(d_i^{\text{inA}}) \quad (6)
\]

The two decision variables of site \( i \) are \( d_i^{\text{inA}} \) and \( d_i^{\text{outR}} \). Note that \( d_i^{\text{inR}} \) is increasing in \( i \), since higher content sites have more reference in-links. It is easy to check that the rest of the function has increasing differences in the pairs \((d_i^{\text{inA}}, d_i^{\text{outR}}), (d_i^{\text{inA}}, i) \) and \((i, d_i^{\text{outR}})\). Therefore, if the maximum is unique for every player, then the optimal decisions \((d_i^{\text{inA}*}, d_i^{\text{outR}*})\) are increasing in \( i \). That is, if \( i > j \) (that is, \( c_i > c_j \)) then \( d_i^{\text{outR}*} \geq d_j^{\text{outR}*} \), and \( d_i^{\text{inA}*} \geq d_j^{\text{inA}*} \).

If the solution of the maximization problem is not unique we have to use the assumption that if players are indifferent they establish as many links as possible - see the proof of Proposition 3 for details.

Proof of Proposition 3

First we prove that at least one pure-strategy Nash-equilibrium exists. We will use the result that any game with convex and compact strategy space and continuous and payoff function, which is quasi-concave in players’ own strategies has a pure-strategy Nash-equilibrium (Mas-Colell et al. 1995). Obviously, the action space \( S = [0, N]^{2N} \) is convex and compact. The payoff function is continuous, we only have to check that the payoff function is quasi-concave in a player’s own decision variables. Let \((x, y)\) denote the two
decision variables \((d_i^{inA}, d_i^{outR})\) for the moment. Then
\[
u(x, y) = f(x, \alpha_1)g(y) - \kappa y - P(x) + \alpha_2,
\]
where \(\alpha_1, \alpha_2\) denote the quantities \(d_i^{inR}, p_i \cdot d_i^{outA}\) respectively, which are fixed for the moment, and \(g(y) = c_i + \beta C(y) - \gamma d_i^{outA}\). Note, that \(g'(y) = \beta C''(y)\).

We will also use the notation \(f(x) = f(x, \alpha_1)\). It is easy to check that \(u(x, y)\) is concave in \(x\) and \(y\) separately. However, it is not necessarily concave as a two-variable function. We will show that it is quasi-concave. The first order conditions are the following.
\[
u_x(x, y) = f'(x)g(y) - P'(x) = 0
\]
\[
u_y(x, y) = f(x)g'(y) - \kappa = 0
\]
The second derivatives have the following signs.
\[
u_{xx}(x, y) = f''(x)g(y) - P''(x) < 0
\]
\[
u_{yy}(x, y) = f(x)g''(y) < 0
\]
\[
u_{xy}(x, y) = f'(x)g'(y) > 0.
\]
In order to check for quasi-concavity we will analyze the bordered Hessian-matrix:
\[
\begin{vmatrix}
0 & u_x & u_y \\
u_x & u_{xx} & u_{xy} \\
u_y & u_{yx} & u_{yy}
\end{vmatrix}
\]
If its determinant is non-negative, then the function \(u(x, y)\) is quasi-concave for \(0 \leq x, y \leq N\). The determinant is
\[
2u_xu_yu_{xy} - u_x^2u_{yy} - u_y^2u_{xx},
\]
which is obviously non-negative, where \(u_x\) and \(u_y\) have the same signs, since \(u_{xy} \geq 0, u_{xx} \leq 0,\) and \(u_{yy} \leq 0\). If \(u_x\) and \(u_y\) have different signs, we have to check that
\[
2(f'(x)g(y) - P'(x))(f(x)g'(y) - \kappa) - (f'(x)g(y) - P'(x))^2f(x)g''(y) - (f(x)g'(y) - \kappa)^2(f''(x)g(y) - P''(x)) \geq 0
\]
If $\kappa$ is small enough, in the case $|u_x| \leq |u_y|$ we can show that

$$2((f'(x)g(y) - P'(x))(f(x)g'(y) - \kappa)) \leq -(f(x)g'(y) - \kappa)^2(f''(x)g(y) - P''(x)).$$

On the other hand, in the case $|u_x| \geq |u_y|$ we have

$$2((f'(x)g(y) - P'(x))(f(x)g'(y) - \kappa)) \leq -(f'(x)g(y) - P'(x))^2f(x)g''(y).$$

Therefore $u(x,y)$ is quasi-concave, that is, it follows that the game has at least one pure-strategy Nash-equilibrium.

In the second part of the proof we will show, that the actions of players are increasing in $i$ in any equilibrium. The payoff of player $i$ is again

$$u_i = f(d_i^{inA}, d_i^{out}) \left( c_i + \beta C(d_i^{out}) - \gamma d_i^{out} \right) - \kappa d_i^{out} + p_i \cdot d_i^{out} - P(d_i^{inA}) \quad (7)$$

We will show that this function has increasing differences in the pairs $(d_i^{inA}, d_i^{out})$, $(d_i^{inA}, i)$ and $(i, d_i^{out})$. The only difference compared to the proof of Proposition 2 is the term $-\gamma d_i^{out}$ in the payoff function. Since prices are ordered the same as contents in this case, $d_i^{out}$ is a decreasing function of $i$. Therefore $-\gamma d_i^{out}$ is increasing ensuring increasing differences. As such, if the maximum is unique for every player, the optimal decisions $(d_i^{inA}, d_i^{out})$ are increasing in $i$: i.e. if $i > j$ (that is, $c_i > c_j$) then $d_i^{out} \geq d_j^{out}$, and $d_i^{inA} \geq d_j^{inA}$. Again, if the solution of the maximization problem is not unique we have to use the assumption that if players are indifferent they establish as many links as possible. As it can be seen from the first part of the proof this assumption is valid because if there are two points where the function attains its maximum, both coordinates of the first point are either greater or smaller. That is, every player separately establishes as many reference links and buys as many advertising links as possible (still attaining the maximum payoff). In other words, $(d_i^{inA}, d_i^{out})$ are the upper bounds of the sets where the payoff function attains its maximum, therefore they are increasing in $i$.
Proof of Proposition 4

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 $f : X \rightarrow X$ is continuous, then $f$ has a fixed point. Recall equation (4), describing the payoff of player $i$. It can be written as

$$u^p_i = f(d^{in}_i)(c(i) - \gamma d^{out}_i) + p(i) \cdot d^{out}_i - P(d^{in}_i), \quad (8)$$

where $P()$ is similar as in the discrete case, it is the total price of the $x$ cheapest sites. That is,

$$P(x) = \int_{p^{-1}([0,q(x))]} p(\lambda),$$

which is a strictly convex function. Then the first order condition of a site $i$ is

$$f'(d)(c(i) - \gamma d^{out}_i) = P'(d),$$

where $d$ is the decision variable. Note that $f$ is concave, $P$ is strictly concave, thus at most one solution exists and that maximizes player $i$’s payoff.

A function $d^{in}(i)$ satisfying

$$f'(d^{in}(i))(c(i) - \gamma d^{out}_i) = P'(d^{in}(i))$$

must represent an equilibrium. We will apply Tikhonov’s theorem to the operator on the normed space of measurable functions on $[0,1]$, that assigns

$$d^{in} \circ f' \circ P^{-1} \cdot (c - d^{out})$$

to the function $d^{in}$, where $d^{out}$ is generated from $d^{in}$. The fixed point of this operator must satisfy the first order condition for every player, thus it represents an equilibrium of the game.
References


