MIS Quarterly

RESEARCH ARTICLE

RECOMMENDATION NETWORKS AND THE LONG TAIL OF ELECTRONIC COMMERCE[1]

Gal Oestreicher-Singer
Recanati Graduate School of Business, Tel Aviv University, Tel Aviv 69978 ISRAEL {galos@post.tau.ac.il}

Arun Sundararajan
Stern School of Business, New York University, 44 West 4th Street,
New York, NY 10012 U.S.A. {asundara@stern.nyu.edu}

It has been conjectured that the peer-based recommendations associated with electronic commerce lead to a redistribution of demand from popular products or "blockbusters" to less popular or "niche" products, and that electronic markets will therefore be characterized by a "long tail" of demand and revenue. We test this conjecture using the revenue distributions of books in over 200 distinct categories on Amazon.com and detailed daily snapshots of co-purchase recommendation networks in which the products of these categories are situated. We measure how much a product is influenced by its position in this hyperlinked network of recommendations using a variant of Google's PageRank measure of centrality. We then associate the average influence of the network on each category with the inequality in the distribution of its demand and revenue, quantifying this inequality using the Gini coefficient derived from the category's Lorenz curve. We establish that categories whose products are influenced more by the recommendation network have significantly flatter demand and revenue distributions, even after controlling for variation in average category demand, category size, and price differentials. Our empirical findings indicate that doubling the average network influence on a category is associated with an average increase of about 50 percent in the relative revenue for the least popular 20 percent of products, and with an average reduction of about 15 percent in the relative revenue for the most popular 20 percent of products. We also show that this effect is enhanced by higher assortative mixing and lower clustering in the network, and is greater in categories whose products are more evenly influenced by recommendations. The direction of these results persists over time, across both demand and revenue distributions, and across both daily and weekly demand aggregations. Our work illustrates how the microscopic economic data revealed by online networks can be used to define and answer new kinds of research questions, offers a fresh perspective on the influence of networked IT artifacts on business outcomes, and provides novel empirical evidence about the impact of visible recommendations on the long tail of electronic commerce.

Keywords: Networks, social networks, electronic commerce, recommender systems, Gini coefficient, long tail, influence, social media, Web 2.0

[1]Vijay Gurbaxani was the accepting senior editor for this paper. Ravi Bapna served as the associate editor.

The appendices for this paper are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

Introduction and Related Work

An important by-product of the sustained recent increase in electronic commerce and interaction is the emergence of a number of *visible* hyperlinked electronic *networks* that connect products and their consumers. These networked IT artifacts include social networking sites like Facebook which link friends to each other, business networking sites like LinkedIn which connect professionals, and *product networks* like those created by Amazon and YouTube which link the landing pages of products or online content. If one imagines the process of browsing an e-commerce site as being analogous to walking the aisles of a physical store, then the aisle structure of an online retail space is determined not by a built physical system of shelves, but instead by the *IT artifact* consisting of the electronic network of interconnected products whose landing pages link to each other. The location of a product in this network is thus analogous to its virtual shelf placement. Much like shelf position in traditional commerce, it is natural to expect that the *position* occupied by products in these electronic hyperlinked networks will be economically influential.

Perhaps the oldest example of an electronic and visible product network of peer products is the "co-purchase" network of Amazon.com,[2] which, for many years now, has presented Amazon.com's consumers with links to complementary products made visible under the label "Consumers who bought this item also bought…." This is illustrated in Figure 1. Will making the product complementarity relationships embodied in co-purchase recommendation links *explicitly visible* redistribute the attention that each product receives from its potential consumers and, as a consequence, the relative fraction of demand and revenue across products? The conjectured answer to this question seems to be that the peer-based recommendations associated with e-commerce (along with wider product selection, costless search, and product unbundling) will increase consumer awareness of relatively obscure products and cause e-commerce demand and revenue distributions to have a *long tail,* whereby less popular products constitute a larger fraction of total sales (Anderson 2006, Brynjolfsson et al. 2006).

While it is natural to anticipate that attention, demand, and revenue will be redistributed on account of recommendation networks, predicting the ensuing shift in the demand distribution is actually not straightforward. As predicted by Anderson (2006), recommendation networks could increase the demand for niche products by making consumers aware of items they might not otherwise have noticed. In contrast, however, since popular products are frequently purchased, they are also on average more likely to be co-purchased, and are thus more likely to receive consumer attention via a co-purchase recommendation link. This could lead to demand being shifted toward blockbusters and away from niche products that might otherwise have been discovered via undirected search. In fact, studies looking for this anticipated distribution of demand from the small but growing long tail literature (Elberse 2008; Fleder and Hosanagar 2008; Tucker and Zhang 2008) have uncovered mixed evidence, which suggests a need for further investigation.

In this paper, we provide an approach for linking recommendation networks to the long tail by connecting the *position* of products in these networks to the relative *demand* and *revenue* within their respective categories. Our empirical analysis relates the influence of Amazon's recommendation network to the demand and revenue distributions for over 200 categories of books, comprising over 250,000 titles sold on Amazon.com, over the course of 28 days in 2007. We model the influence of the network on each book by computing each book's PageRank, which measures the centrality of its network position. We then quantify the "evenness" of each category's demand and revenue distributions by constructing their Lorenz curves and computing their associated Gini coefficients, a measure of inequality that is normalized for size and average magnitude.[3]

Our results present significant and persistent evidence that categories whose books are more *highly* and *evenly* influenced by visible networks have consistently flatter demand and revenue distributions, even after controlling for the average demand and price in the category as well as the number of products in the category. We estimate that doubling the average network influence on a category is associated with an *increase* of about 50 percent in the relative revenue for the least popular 20 percent of products, and with a *reduction* of up to 15 percent in the relative revenue for the most popular 20 percent of products. We also show that this effect is enhanced when there is higher assortative mixing in the network, or when a large fraction of recommendations terminating within a category also originate from the category. Further, categories whose products are more evenly influenced by the network have flatter demand distributions. The

[2]Such co-consumption networks are not unique to Amazon.com. Barnes and Noble has a similar feature; more recently, YouTube introduced a similar graphical network of co-viewed videos.

[3]Our paper is not about documenting changes in individual product outcomes, but rather about assessing the relationship between the influence of a visible recommendation network and the *distributions* of demand and revenue across products. An individual-level analysis is available in Oestreicher-Singer and Sundararajan (2012).

Figure 1. The Outgoing Co-Purchase Links for a Sample Book

direction of these results persists across all 28 days, across both demand and revenue distributions, and across both daily and weekly demand aggregations.

Put simply, our findings imply that the influence of visible recommendation networks is positively associated with the widely documented phenomenon of the long tail of demand. The observable emergence of these recommendation networks, new IT artifacts that are as fundamental to electronic commerce as the physical artifacts of retail shelves are to traditional commerce, is a basic way in which e-commerce differs from traditional face-to-face commerce. Our results provide evidence that this distinction might explain some of the documented contrast in demand patterns between online and offline commerce.

We add to a small but growing literature in information systems and marketing documenting the drivers and extent of the long tail. Early work (Anderson 2006; Brynjolfsson et al. 2003; Clemons et al. 2006) posited that wider product variety would drive sales away from popular products and predicted the emergence of a long tail of demand. A subsequent paper by Brynjolfsson et al. (2006) provides evidence supporting the theory that reduced search costs foster such diversity in demand. Choi and Bell (2011) find that online sales are higher for niche categories, especially in locations where customers suffer from preference isolation. Bailey et al. (2008) find that the demand for niche products may be systematically underestimated due to a bias in research that focuses on larger retailers, or in other words, after accounting for smaller retailers, the long tail may in fact be longer than we think. Using a novel combination of research methods, Tucker and Zhang (2008) show that the marginal benefits of visible popularity information (the number of consumers who have previously visited) are higher for niche products, and this in turn may contribute to a more prominent long tail.

However, more recent evidence has suggested that this long tail effect may not be as simple as originally conjectured. For example, Elberse (2008) discusses distributions of DVD sales between 2000 and 2005. She reports evidence of a lengthening of the tail of demand in 2005, documenting a doubling of the number of niche products which regularly sell a small number of copies. In parallel, she provides evidence of an amplification of the "superstar" effect: there are fewer products in the highest selling *quantiles*, each of which has a *higher* individual demand level. Similarly, Zhao et al. (2008) study the influence of word-of-mouth on hit and non-hit products, documenting how positive word-of-mouth has a higher impact on hit products than on non-hit products, while negative word-of-mouth has an opposing effect. Their paper takes an interesting micro approach to the analysis, although it does not control for the amount of attention due to word-of-mouth relative to the product's absolute demand. Goh and Bockstedt (2008) examine how the unbundling of music online impacts the relative demand for popular and niche products, showing that the disaggregation of digital goods may actually shift demand toward more popular products and away from niche products that previously received much of their demand on account of being bundled with hits.

In contrast with Elberse, we document cross-sectional rather than longitudinal variations in demand and revenue distributions and also explicitly associate them with the influence of a recommendation network, a specific conjectured driver of this shift. This focus also differentiates our work from

Brynjolfsson et al. (2006), whose emphasis is on reduced search costs and an online–offline comparison. Tucker and Zhang study the influence of a different kind of electronically visible information on demand patterns; our work differs in its focus on recommendation networks, detailed inter-category comparisons, as well as a more nuanced measurement of influence. Broadly, our work is further distinguished from the prior empirical literature on the long tail in a couple of salient ways. First, our study focuses entirely on products sold online, contrasting the demand distributions of categories based on the extent to which they are influenced by a recommendation network. Second, we use a measure of centrality (PageRank) to quantify this influence while controlling for total demand levels; accounting for intrinsic popularity in this way allows us to focus more carefully on variations in the *distribution* of demand and revenue across categories.

In parallel, two recent theoretical papers provide new arguments linking recommendation networks to the long tail of demand. Fleder and Hosanagar (2008) simulate the effects of recommendation systems on the distribution of demand and predict that recommender systems that base recommendations on sales and ratings reinforce the popularity of already popular products. The researchers do so with the caveat that their results depend heavily on assumptions about how the recommender system works. In contrast, Hervas-Drane (2007) shows that while recommendations largely benefit mainstream consumers, when recommender systems based on social filtering are introduced alongside traditional word-of-mouth recommendations, there is a positive impact on consumers interested in niche products, since such recommenders are more likely to draw attention to niche products. The competing theoretical predictions of these papers further motivates our empirical work.

More broadly, our work highlights an ongoing transformation engendered by the emergence of visible IT-based networked artifacts. Our social, economic, and cultural connections to one another are made more persistently visible by virtue of their being encapsulated and displayed as digital artifacts, and we believe that the visibility of these networks by itself will alter their socioeconomic impact. Our focus in this paper is on *product networks*, a new and relatively understudied class of socioeconomic networks. While the last few years have witnessed the emergence of a vibrant research stream about social networks, less attention has been given to the analysis of online product networks. This is especially surprising given their ubiquity and potential influence on business. Some notable exceptions include Mayzlin and Yoganarasimhan (2008) who analyze the network of hyperlinked blogs, Dellarocas et al. (2010) who study links between news websites, and Katona and Sarvary (2008) who examine the strategic interaction between content sites. Clearly, our work departs from these studies in its context, its quantifying actual demand and revenue, and its methods for associating demand inequity with network structure and position.

While the notion of an online product network is fairly new, extensive attention has been given to inter-product associations in the context of traditional brick-and-mortar retailing, although generally focused on correlations between dyads or a small number of entities. Such inter-product correlations are of interest to marketers since they can affect optimal pricing decisions (Niraj et al. 2008), new product sales forecasts (Sriram et al. 2010), assessments of cross-selling opportunities (Li et al. 2005), or competitive dynamics (Wedel and Zhang 2004). The idea that position affects demand is also fairly well-established in the context of traditional brick-and-mortar retailing, a point made repeatedly in the literature on shelf positioning and placement (see, for example, Desai 2001; Lariviere and Padmanabhan 1997; a more detailed survey of this literature is available in Oestreicher-Singer and Sundararajan 2012). We treat network position as given, focusing on assessing the demand influence garnered from how central this position is, rather than addressing programmatic or strategic allocation to positions. This distinction actually highlights an interesting differentiating feature of position that is defined by co-purchases: the virtual aisle location of a product is determined, in part, *endogenously and collectively* by consumers rather than being chosen based on fees paid by manufacturers, or explicit strategic considerations by the retailer.

A final stream of related research has associated network properties with a variety of adoption and diffusion outcomes in organizations and markets including the influence of social networks on the spread of products and information (for example, Aral et al. 2009; Goldenberg et al. 2001; Muller et al. 2009; Van den Bulte and Wuyts 2007), the relationship between diffusion processes and associated social activity levels (Oh et al. 2008), identifying opinion leaders (Keller and Berry 2003; Watts and Dodds 2007) and innovators (Valente 1996), the role of spatial proximity in the process of product and service adoption (Barrot et al. 2008), and the extent to which network position and information diffusion affect the productivity and performance of employees in organizations (Aral et al. 2007). Related research streams at the interface of networks and IS include the use of networks and collective inference for predictive modeling (Dhar et al. 2009; Hill et al. 2006) and the use of graphs for modeling underlying social structures in network games (Bramoulle and Kranton 2007; Galeotti et al. 2010; Sundararajan 2007).

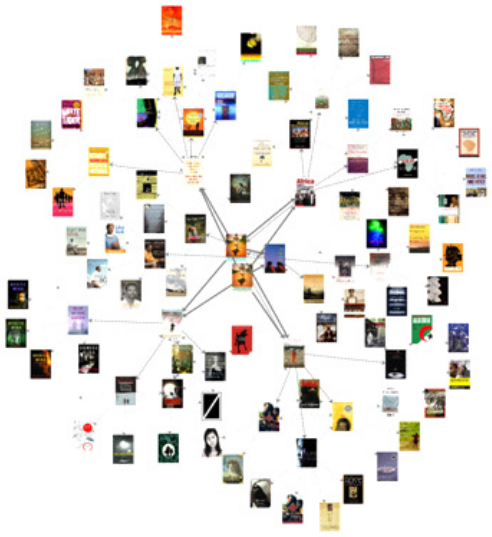


Figure 2. Subset of the Recommendation Network (Highlighting the Network Structure)

Overview of Data and How It Is Collected

Our empirical work is based on a set of observations of recommendation networks for over 250,000 books sold on Amazon.com. Each product on Amazon.com has an associated webpage. Each such product page has a set of co-purchase links, which are hyperlinks to the set of products that were co-purchased most frequently with this product on Amazon.com. This set is listed under the title "Customers who bought this also bought." This is illustrated in Figure 1.

Conceptually, the co-purchase network is a directed graph in which nodes correspond to products, and edges correspond to directed co-purchase links. We collect data about this graph using a Java-based crawler, which starts from a popular book and follows the co-purchase links using a depth-first search algorithm. At each page, the crawler gathers and records information for the book whose webpage it is on, as well as the co-purchase links on that page, and terminates when the entire connected component of the graph is collected. This is repeated daily. An illustrative subset of the graph is presented in Figure 2. The algorithm used for data gathering is provided in Appendix A.

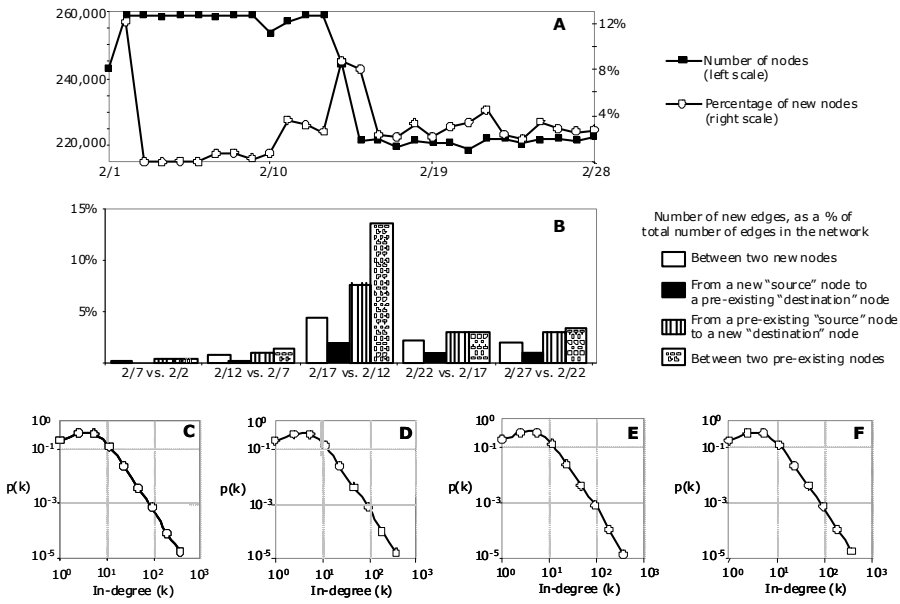
We have chosen to focus on books because they have the largest number of individual titles, the product set is relatively stable (compared to electronics, for instance), and because the influence of recommendations based on shared purchasing patterns (that reveal underlying product similarities not easily observable in expressed product characteristics) is likely to be significant for books.

The data collection began in August 2005 and is currently ongoing. The graph is traversed every day. Apart from the co-purchases, each book's ISBN, list price, sale price, category affiliation, secondary market activity, author, publisher, publication date, and consumer ratings are gathered. An additional script collects the demand information for all books on the graph every 3 hours for the 24-hour period following the collection of the graph.[4]

The following data is available for each book on the co-purchase graph, for each day:

- **ASIN**: A unique serial number given to each book by Amazon.com. Different editions and different versions have different ASIN numbers.
- **List Price**: The publisher's suggested price.
- **Sale Price**: The price on the Amazon.com website that day.

[4]The demand for books is computed using the SalesRank information provided by Amazon.com. More details are available in Appendix B.



(A) plots the variation in the size of the graph and the differences between the identities of the nodes from day to day. The changes to the edges on successive graphs are categorized in (B). The degree distribution (in-degree) is plotted after logarithmic binning of the data for February 1 (C), 10 (D), 19 (E), and 28 (F).

Figure 3. Summary of the Evolution of the Co-purchase Network

- **Co-purchases**: ASINs of the books that appear as its co-purchases.[5]

- **SalesRank**: The SalesRank is a number associated with each product on Amazon.com, which measures its demand relative to that of other products. The lower the number is, the higher the sales of that particular product.

- **Category Affiliation**: Amazon.com uses a hierarchy of categories to classify its books. Thus, each book is associated with one or more hierarchical lists of categories, starting with the most general category affiliation, and ending with the most specific one. For example:

  Subjects > Business & Investing > Biographies & Primers > Company Profiles

  (for *The Search* by John Batelle).

- **Author**: The name of the author or authors of the book.

- **Publisher**: The name of the publisher of the book.

- **Publication Date**: The date of publication of the book (by that publisher).

As illustrated in Figure 3, for a sample month, the component of the co-purchase network we study changes substantially over time. It contains new nodes every day (over 6,500 per day, on average) and there are frequent daily changes to the edges between existing nodes. The occasional large shifts in the component's size are due to one or more clusters of nodes detaching from the large connected component; this was often accompanied by a different set of clusters of nodes attaching to this component. There was also a significant redistribution of edges in the graph in the middle of the month, probably because of the seasonal demand spike associated with Valentine's Day. Despite the variation in the graph's composition, its in-degree[6] distribution remained quite stable through the month. Between 18 percent and 20

[5]Our work is based on data from 2007, when Amazon.com provided just five co-purchase links per product. Currently Amazon.com provides a list of up to 100 such links for each book. Users are initially exposed to the top six due to screen size limitations, and they can then click on a link to view the next six products.

[6]In-degree refers to the number of incoming links to a node in the network. In our context, it is the number of hyperlinks that terminate at that book's page.

percent of the books have one incoming link, a little over 30 percent have two or three incoming links, roughly the same fraction have between four and seven incoming links, and the in-degree distribution of the remaining 15 percent or so follows a power-law distribution.

Network Position and the Distribution of Demand

This section describes how we construct our variables relating to network position/influence and the distribution of demand/revenue. To quantify the distribution and to compare it across groups of books, we first have to partition the books we analyze into groups. The most natural partition is by category affiliation. Recall that Amazon.com uses a hierarchy of categories to classify its books, and each book is associated with one or more hierarchical lists of categories (see example above). We use this exogenous categorization as a grouping for comparing demand distribution across books. Using the second level of the hierarchy, there are 1,472 such categories across all books sold, of which between 203 and 225 (depending on the day) have 100 or more nodes (books) represented in our co-purchase network.

In order to relate the network position of a set of products to their relative revenue fraction, we follow the following sequence of steps:

1. Quantify the distributions of demand and revenue. We characterize the demand and revenue distributions of each category by constructing each distribution's Lorenz curve and measuring its Gini coefficient (more on this later).

2. Characterize the extent to which the position of a book in the co-purchase network is related to the influence of the network on the book's demand by using PageRank, a measure of centrality. We then compute the average PageRank for each category as a measure of the category's centrality.

3. Associate variation in (2) with variation in (1) at a group-specific level of analysis. This is repeated for 28 different instances of the co-purchase network. We have also repeated the same analysis for four distinct composites of seven daily graph instances, and 22 overlapping composites of seven daily graph instances, with a remarkable level of stability across our empirical findings.

Quantifying the Distribution of Revenue: The Gini Coefficient

We quantify the shape of the revenue distribution within categories in a way that is comparable across categories by calculating the Gini coefficient of each category of books (Gini 1921). The Gini coefficient is a measure of distributional inequality, a number between 0 and 1, where 0 corresponds to perfect equality (in our case: where all the books in that category have the same revenue) and 1 corresponds to perfect inequality (where one book in the category has all the revenue, and all other books in the category have zero revenue).

The Gini coefficient is based on the Lorenz curve (Lorenz 1905), a widely used summary of distributional equality most commonly seen in comparisons of income distributions across regions and time. In our analysis, the Lorenz curve of a category's revenue (demand) ranks the products in increasing order of revenue (sales), then plots the cumulative fraction $L(\rho)$ of revenue (sales) associated with each ascending rank percentile ρ, where $0 < \rho \leq 1$. More precisely, define $N = \{1, 2, 3, ..., n\}$ as the set of all books in a category of size n, and define $q(i)$ as the revenue for book i. To compute the Lorenz curve, we define, for each book i, $R(i)$ as the size of the set $\{x : x \in N, q(x) \leq q(i)\}$, which is the set of all products with revenue less than or equal to that of i. $R(i)$ is thus simply the (inverse) rank of the product within its category, with the product with the lowest revenue having the lowest rank. Next, we define

$$S(r) = \{y \in N, R(y) \leq r\} \tag{1}$$

as the set of product indices whose rank is less than or equal to r. Then, for each percentile ρ (which corresponds to the books ranked ρn or lower), the Lorenz curve is defined by

$$L(\rho) = \frac{\sum_{y \in S(n\rho)} q(y)}{\sum_{y \in N} q(y)} \tag{2}$$

Notice that the Lorenz curve is increasing and piecewise (weakly) convex.

The Gini coefficient is computed as twice the area between the Lorenz curve $L(\rho)$ and the 45-degree line between the origin and (1, 1). We calculate it for each category by first computing the entire area above the Lorenz curve, the Lorenz upper area

$$LU = \sum_{y=1}^{n} \left[1 - L\left(\frac{y}{n}\right) \right] \quad (3)$$

and then using the identity

$$Gini = 2(LU) - 1 \quad (4)$$

Figure 4 illustrates this computation for two categories in our data set.

The Gini coefficient is especially suitable for this study for a variety of reasons. Most importantly, it measures inequality in the revenue distribution regardless of the category's size or average demand (popularity), which facilitates comparing different categories despite their intrinsic differences and independent of their scale. An important point to note here is that we compute the Lorenz curve and Gini coefficients *for each category*. While the SalesRank-to-demand conversion computations that follow use a standard assumption, that across Amazon.com, overall demand is Pareto-distributed by rank, different categories, being varying subsets of the entire sample, have demand distributions of varying forms.[7]

Measuring Network Influence: Weighted PageRank

Our measure of the influence the recommendation network has on a product is called *WeightedPageRank*. This is a computed measure of the *global* influence of the recommendation network on outcomes. It is based on (and essentially identical to) *PageRank* as computed by Google's original algorithm (Brin and Page 1998; Brin et al. 1999). It iteratively computes the influence of the entire network on each product over time. It can operate either on an individual daily graph, or on an average graph, constructed as a weighted composite of a few co-purchase networks. The original PageRank algorithm provides a ranking of the *importance* of web pages based on the link structure of the web created by the hyperlinks between the pages by using the following model:

$$PageRank(i) = \frac{(1-\alpha)}{n} + \alpha \sum_{j \in G(i)} \frac{PageRank(j)}{OutDegree(j)} \quad (5)$$

where $j \in G(i)$ if there is a link originating from product j to product i (meaning that product j is a network neighbor of product i) and $OutDegree(j)$ is the total number of links originating from product j.

PageRank is based on a simple model of behavior—one of consumers who "surf" the recommendation network randomly. A surfer follows any one of the links on a page with equal probability or jumps to a randomly chosen page with probability $(1 - \alpha)$ (this probability is also referred to as the *damping factor*, and is what differentiates PageRank from a commonly used notion of centrality in social network theory). The algorithm divides a page's PageRank evenly among its successors in the network. The ranking of a page ends up being the long-run steady-state probability that a random surfer who starts at a random page will visit that specific page. Thus, a page can achieve a high rank by either having many pages pointing to it or having a few highly ranked pages pointing to it. The PageRank of all pages in the network is computed iteratively, until some convergence estimator is met. For more information about the PageRank algorithm see Appendix C.

Since the influence of a network may diffuse out over some period of time during which the network itself changes, we adapt the PageRank algorithm to account for the fact that one might wish to measure the average influence the network has on a product over a weighted composite of networks. In this adapted model

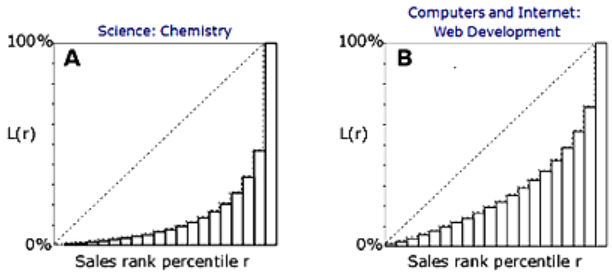
$$WeightedPageRank(i) = \frac{(1-\alpha)}{n} + \alpha \sum_{j \in G(i)} Weight(j,i) \frac{WeightedPageRank(j)}{\sum_{k \in G(j)} Weight(j,k)} \quad (6)$$

where $Weight(j,i)$ is the fraction of the days that the link was present in the co-purchase network.[8]

It is important to note that while this kind of measure of centrality is widely used as a measure of *importance* in ranking algorithms (such as Google's), we reinterpret PageRank as also being a measure of *network-induced attention*. That is, we are exploiting the fact that, fundamentally, measures like PageRank simply represent the probability that a random surfer will arrive at a hyperlinked page if he or she traverses just the hyperlinks of the network. Thus, a product with a higher PageRank is *more likely to get traffic* from the network than one with a lower PageRank. Consequently, network centrality or PageRank of a product assesses the consumer attention a product would get if all of

[7]Note that while the Gini coefficient is linked to the Pareto distribution, as is our demand estimation method (see Appendix B for details), our measures still pick up on variations in the demand distribution across different categories. The absolute magnitude of the Gini coefficients may be slightly different if we choose a different Pareto distribution for the entire population, but the relative magnitudes will still be similar.

[8]When computed on a single co-purchase network, $Weight(j,i)=1$.



Lorenz curves for the "Science: Chemistry" and "Computers and Internet: Web Development" categories respectively. L(r) plots the fraction of the category's total revenue from the books whose SalesRanks are in the category's lowest r^2 percentile. (The data has been binned for illustrative purposes in the figure.) The size of the dotted area is proportionate to (and is one half of) the category's Gini coefficient. The categories' Gini coefficients are 0.76 (A) and 0.51 (B) respectively. Notice that a category whose revenue is more highly concentrated on the higher-ranked products has a higher Gini coefficient.

Figure 4. Illustration of Lorenz Curves

its demand came exclusively from consumers traversing the recommendation network. In other words, it isolates the *extent* to which the network we are interested in—the co-purchase network—*influences* the product in question.[9] As we are interested in the aggregate influence of the network on different categories and the co-variation of this influence with aspects of the demand and revenue distribution, in what follows, we will use the average PageRank, the variance in PageRank, and the kurtosis of the PageRank computed for each category.

Analysis and Results: Recommendation Networks and the Long Tail

Having defined our two main variables—PageRank and Gini—we now turn to motivating our empirical analysis. We do so by presenting an illustrative model of how the presence of a recommendation network might change the distribution of demand, and by examining how an increase in its influence might enhance or diminish the long tail of e-commerce demand.

Consider a category with two products labeled 1 and 2. In the absence of the recommendation network, suppose the level of attention (for example, number of page views) that product i gets is $\alpha_I(i)$, and the conversion rate associated with this attention is $c_I < 1$. The demand levels for products 1 and 2 are, respectively,

$$\begin{aligned} q_I(1) &= c_I \alpha_I(1) \\ q_I(2) &= c_I \alpha_I(2) \end{aligned} \tag{7}$$

Without any loss in generality, assume that $\alpha_I(2) > \alpha_I(1)$. It follows from (1) that $S(1) = \{1\}$, $S(2) = \{1,2\}$, and after using (2) and (3) to compute the Lorenz upper area, one can show that the Gini coefficient for the category in the absence of the recommendation network is

$$Gini_I = \frac{q_I(2) - q_I(1)}{q_I(2) + q_I(1)} \tag{8}$$

which can be rewritten as

$$Gini_I = \frac{1 - \frac{q_I(1)}{q_I(2)}}{1 + \frac{q_I(1)}{q_I(2)}} \tag{9}$$

Now, suppose the presence of the recommendation network has two effects. First, it introduces a new source of network attention $\alpha_N(1)$ and $\alpha_N(2)$ for the two products. Since this is a different attention source, we assume it has a different associated conversion rate c_N. Further, suppose the presence of the network also changes the conversion rate from intrinsic attention from c_I to c_I'. It follows that the demand for the two products when they receive both intrinsic and network attention will be

$$\begin{aligned} q_N(1) &= c_I' \alpha_I(1) + c_N \alpha_N(1) \\ q_n(2) &= c_I' \alpha_I(2) + c_N \alpha_N(2) \end{aligned} \tag{10}$$

and correspondingly (following equations (1–4) and a sequence of analytical steps similar to those described above)

[9]For a survey on the use of PageRank in the literature, see Langville and Meyer (2005).

the new Gini coefficient of the category is[10]

$$Gini_N = \frac{1 - \frac{q_N(1)}{q_N(2)}}{1 + \frac{q_N(1)}{q_N(2)}} \quad (11)$$

It follows from (9) and (11) that $Gini_N < Gini_I$ if and only if

$$\frac{q_N(1)}{q_N(2)} > \frac{q_I(1)}{q_I(2)} \quad (12)$$

or if

$$\frac{c_I'\alpha_I(1) + c_N\alpha_N(1)}{c_I'\alpha_I(2) + c_N\alpha_N(2)} > \frac{c_I\alpha_I(1)}{c_I\alpha_I(2)} \quad (13)$$

Equation (13) can be rearranged as

$$\left[c_I\alpha_I(2)\right]\left[c_I'\alpha_I(1) + c_N\alpha_N(1)\right] > \left[c_I\alpha_I(1)\right]\left[c_I'\alpha_I(2) + c_N\alpha_N(2)\right] \quad (14)$$

which upon multiplying out and rearranging, reduces to

$$\frac{\alpha_N(1)}{\alpha_N(2)} > \frac{\alpha_I(1)}{\alpha_I(2)} \quad (15)$$

One can use (13) to show that the condition in (15) holds if and only if the following condition holds:

$$\frac{\alpha_N(1)}{\alpha_N(2)} > \frac{c_I'\alpha_I(1) + c_N\alpha_N(1)}{c_I'\alpha_I(2) + c_N\alpha_N(2)} \quad (16)$$

or equivalently, if

$$\frac{\alpha_N(1)}{\alpha_N(2)} > \frac{q_N(1)}{q_N(2)} \quad (17)$$

The condition in (17) intuitively implies that the presence of the network will flatten the demand distribution of a category if the distribution of attention from the recommendation network is more "even" than the distribution of observed demand in the presence of the network. For a random sample of books across categories, Figures 5 and 6 contrast the PageRank distribution with the distribution of demand. Both comparisons illustrate that rather than being proportionate to demand, PageRank is more evenly and randomly spread among books. Since we have argued that PageRank is a measure of the *network attention* received by products, the condition in (17) from our illustrative model leads us to hypothesize that the presence of the recommendation network will lower the Gini coefficient, or reduce the inequality in demand across products.

Additionally, different categories are influenced differentially by the presence of the recommendation network. We quantify this difference by assessing the average PageRank of books in a category, based on the idea that a category with a higher average PageRank receives, on average, more attention from the network. Returning to our illustrative model, suppose the level of attention flowing from the network to a category's products increases by a factor of $\beta > 1$. The analysis above indicates that this increase will lower the category's Gini coefficient if and only if it leads to an increase in the ratio $\left[\frac{q_N(1)}{q_N(2)}\right]$. Rewriting (10) to reflect the introduction of β

$$\frac{q_N(1)}{q_N(2)} = \frac{c_I'\alpha_I(1) + \beta c_N\alpha_N(1)}{c_I'\alpha_I(2) + \beta c_N\alpha_N(2)} \quad (18)$$

this in turn suggests that if the derivative of the RHS of (18) with respect to β is positive, an overall increase in the level of attention from the network (an increase in average PageRank) will reduce the category's Gini coefficient and increase demand for the "tail." We examine this by differentiating both sides of (18) with respect to β

$$\frac{d}{d\beta}\left[\frac{q_N(1)}{q_N(2)}\right] = \frac{c_N\alpha_N(1)}{c_I'\alpha_I(2) + \beta c_N\alpha_N(2)} - \frac{\left[c_I'\alpha_I(1) + \beta c_N\alpha_N(1)\right]c_N\alpha_N(2)}{\left[c_I'\alpha_I(2) + \beta c_N\alpha_N(2)\right]^2} \quad (19)$$

which simplifies to

$$\frac{d}{d\beta}\left[\frac{q_N(1)}{q_N(2)}\right] = \frac{c_N c_I'\left[\alpha_N(1)\alpha_I(2) - \alpha_N(2)\alpha_I(1)\right]}{c_I'\alpha_I(2) + \beta c_N\alpha_N(2)} \quad (20)$$

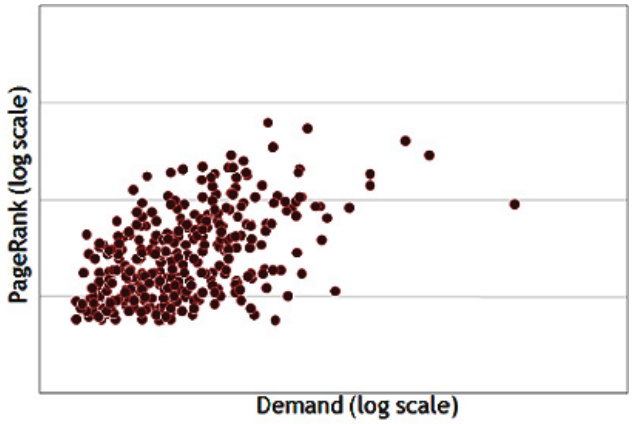
The RHS of (20) is positive if its numerator is positive, or if

$$\alpha_N(1)\,\alpha_I(2) > \alpha_N(2)\,a_I(1) \quad (21)$$

which is precisely the condition of equation (15). In our illustrative model, we have, therefore, shown that if the distribution of attention generated by the network is more even than the intrinsic distribution of attention to products (the condition of equation (15)), then an *increase* in the influence of the network on a category (an increase in β in our model, or an increase in average PageRank) will *reduce* the Gini coefficient of the category, or shift demand from the popular products to the tail. This is our main testable conjecture.[11]

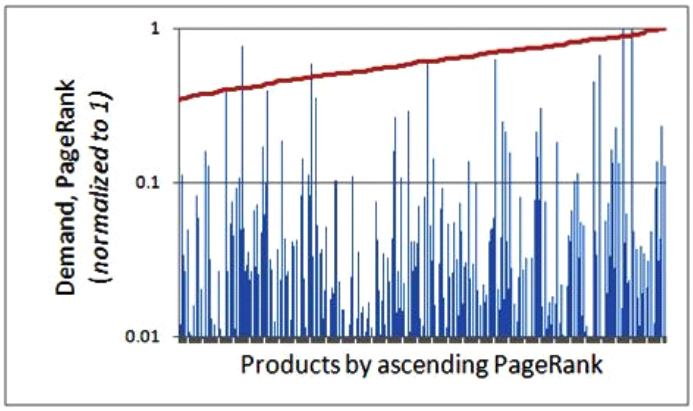
[10]This assumes that the presence of the recommendation network does not reverse the ordering of popularity of the two products. We return to this later.

[11]We can extend this model to accommodate (one at a time) differing c_I across the two categories, or differing c_N across the categories, with the same conclusion.



Illustrates the fact that while the two variables are weakly (negatively) correlated, there are factors beyond network position that affect a product's demand.

Figure 5. Plots of Revenue Versus PageRank for a Sample of the Data



Contrasts the PageRanks (ascending line) of a random sample of books across all categories with their corresponding demand levels (dark spikes), with the maximum PageRank and demand levels in the sample normalized to 1. The correlation coefficient between demand and PageRank across the entire data set averaged 0.03 across the 28 days.

Figure 6. Contrast of PageRank of Random Sample of Books

To test this main conjecture, we estimate the relationship between a category's Gini coefficient (*RevenueGini*) and the average PageRank of its books (*AvgPageRank*) using ordinary least-squares regression.[12] We use logarithmic transformations of all our variables to facilitate interpretation of their coefficients as percentage changes, and because the empirical distributions of the transformed variables are more suitable for OLS. We use the variance in PageRank across the category's books (*PageRankVar*), the kurtosis of the PageRank distribution across the category's books (*PageRankKurtosis*), the category's average demand (*AvgDemand*), the number of books in the category (*Size*), the fraction of co-purchase links to the category's books that are from other books within it (*Mixing*),[13] the average clustering coefficient of books in the

[12] *RevenueGini* refers to the Gini coefficient computed based on the category's revenue distribution. Results using the demand Gini are discussed later.

[13] For more information on measuring assortative mixing, see http://en.wikipedia.org/wiki/Assortative_mixing.

category (*Clustering*),[14] the average list price (*ListPrice*), and the average sale price (*SalePrice*) as control variables. We thus report on our estimation of the following reduced-form equation:

$$\begin{aligned} Log[RevenueGini] = {} & a + b_1 Log[AvgDemand] + \\ & b_2 Log[AvgPageRank] + \\ & b_3 Log[PageRankVar] + \\ & b_4 Log[PageRankKurtosis] + \\ & b_5 Log[Size] + b_6 Log[Mixing] + \\ & b_7 Log[Clustering] + b_8 Log[ListPrice] + \\ & b_9 Log[SalePrice]. \end{aligned}$$

Given the computational complexity in handling this large data set, we limit our analysis to 28 randomly chosen days and estimate this equation using data about the network on those dates. Summary statistics for our data across the 28 days are provided in Table 1.

Additionally, we repeat the estimation including date dummies to control for unobserved heterogeneity due to time or seasonality (those results are presented in the third column of Table 2).[15] To further control for the intrinsic quality of the books, we add controls for the average rating of books in the category and the variance of their ratings (those results are presented in the rightmost column of Table 2). We also repeat this estimation for each of the 28 days separately.

The results of the latter estimations are summarized in Figure 7 and are strikingly consistent. We summarize some key observations below.

Recommendation Networks and the Evenness of Revenue

The coefficients of *AvgPageRank* in each of our models show that categories with a higher average PageRank are consistently associated with a significantly *lower* Gini coefficient. In other words, in categories that are, on average, *more influenced* by the recommendation network, revenue is *more evenly distributed*. In our base model estimates, the coefficient value of the *AvgPageRank* variable is -0.28, with the following interpretation: a doubling of the average PageRank of a category's books is associated with an 18 percent *decrease* in the Gini coefficient of the category.

Our results, therefore, establish that, based on a comparative analysis across more than 200 categories of books, more influential recommendation networks are associated with flatter revenue distribution or an increase in the relative revenue from niche (rather than blockbuster) products. Figure 7 shows the shift in the fractions of revenue obtained by the most and least popular books for an illustrative doubling of influence of the recommendation network on a category. As the figure illustrates, these revenue shifts can lead to pretty substantial changes in the relative revenue contributions of popular and of niche products. For example, the revenue fraction of the least popular 20 percent of products is about 2 percent for a category with a Gini of 0.8, or just half the corresponding revenue fraction of a category with a Gini of 0.6.

These results are further reinforced when controlling for unobserved heterogeneity of product category and time. The estimates with date dummies (Table 2) indicate that the *AvgPageRank* coefficient is essentially unchanged by controlling for unobserved variations across different dates (which might have been caused, for example, by media events that shifted demand toward more popular products in specific categories). We also find that, even after controlling for unobserved variation across categories, the coefficient of the *AvgPageRank* variable is significant and negative, although its magnitude is smaller. This reduction in the coefficient value may be due to accounting for category-specific factors that are unrelated to the influence of the recommendation network (for example, the category may be one in which there are fewer best-sellers or one in which there are multiple focused subject areas with distinct customer segments). Even so, an 11 percent decrease in Gini coefficient with a doubling of average PageRank is striking, and represents a nontrivial change in the relative revenue fractions. Most importantly, it demonstrates that the relationship between the long tail and the influence of recommendation networks is real and not merely on account of some unobserved category covariate.

The coefficients of many of our control variables are consistently significant and are worth mentioning since they each strengthen our central finding.

Category Size, Average Demand, and Rating

We find that categories with more products (measured by the variable *Size*) are more likely to contain very popular products. The categories in our data set have 100 to over 10,000 books in them. It is natural to expect that when all else is equal, a category with over 10,000 books is more likely to have higher variance in the revenue for its books than a category with about 100 books.

[14]For more information about measuring clustering coefficient, see http://en.wikipedia.org/wiki/Clustering_coefficient.

[15]The coefficients for the dummy variables are not presented.

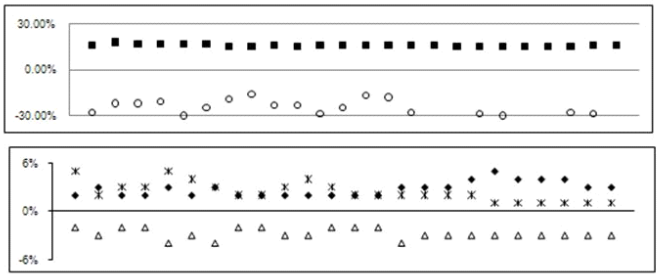
Table 1. Summary Statistics

Variable	Range	Mean	Standard Deviation
RevenueGini	[0.43, 0.98]	0.73	0.09
AvgDemand	[0.47, 81.28]	3.02	3.6
AvgPageRank	$[1.37 \times 10^{-6}, 5.07 \times 10^{-6}]$	2.78×10^{-6}	3.60×10^{-7}
PageRankVar	$[1.33 \times 10^{-12}, 4.10 \times 10^{-10}]$	1.69×10^{-11}	1.89×10^{-11}
PageRankKurtosis	$[1.22 \times 10^{21}, 8.76 \times 10^{24}]$	3.81×10^{23}	4.24×10^{21}
Size	[100, 14293]	1, 344	2, 041
Mixing	[0.01, 0.78]	0.34	0.18
Clustering	[0.22, 0.59]	0.38	0.06
ListPrice	[8.42, 99.75]	31.13	16.69
SalePrice	[7.78, 97.43]	27.56	16.40
AvgRating	[3.75, 4.6]	4.25	0.15
VarRating	[0.26, 1.41]	0.55	0.16
Log[*RevenueGini*]	[-0.84, -0.01]	-0.31	0.13
Log[*AvgDemand*]	[-0.73, 4.3]	-0.82	0.67
Log[*AvgPageRank*]	[-13.50, -12.19]	-12.80	0.13
Log[*PageRankVar*]	[-27.34, -21.61]	-25.09	0.75
Log[*PageRankKurtosis*]	[48.55, 57.43]	53.81	1.05
Log[*Size*]	[4.60, 9.56]	6.47	1.17
Log[*Mixing*]	[-4.21, 0.23]	-1.25	0.65
Log[*Clustering*]	[-1.48, 0.49]	-0.96	0.16
Log[*ListPrice*]	[2.12, 4.60]	3.31	0.47
Log[*SalePrice*]	[2.05, 4.57]	3.17	0.51
Log[*AvgRating*]	[1.32, 1.52]	1.44	0.04
Log[*VarRating*]	[-1.33, 0.34]	-0.62	0.28

Table 2. Coefficient Estimates

	Estimated Value (SE)			
Variable	Base Model	Controlling for Category	Controlling for Date	Controlling for Rating
AvgDemand	0.15*** (0.01)	0.17*** (0.01)	0.15*** (0.01)	0.16*** (0.01)
AvgPageRank	-0.28*** (0.02)	-0.11*** (0.03)	-0.25*** (0.02)	-0.26*** (0.02)
PageRankVar	0.03*** (0.01)	0.007* (0.00)	0.02*** (0.00)	0.03*** (0.00)
PageRankKurtosis	-0.005** (0.00)	-0.003(0.00)	-0.004* (0.00)	-0.004** (0.00)
Size	0.03*** (0.01)	0.09*** (0.01)	0.03*** (0.00)	0.03*** (0.01)
Mixing	-0.03*** (0.01)	-0.03(0.01)	-0.03*** (0.01)	-0.03*** (0.01)
Clustering	0.05*** (0.00)	0.08*** (0.02)	0.04*** (0.01)	0.05*** (0.01)
ListPrice	0.14*** (0.01)	0.36*** (0.02)	0.06*** (0.01)	0.14*** (0.01)
SalePrice	-0.11*** (0.01)	-0.53*** (0.02)	-0.04*** (0.01)	-0.10*** (0.01)
Log[*AvgRating*]				0.19*** (0.03)
Log[*VarRating*]				0.06*** (0.01)
Constant	-3.32*** (0.12)	-1.74*** (0.25)	-3.05*** (0.12)	-3.39*** (0.13)
Observations	7070	7070	7070	7070
R^2	71.8%	89.1%	73.5%	72.53%
Number of omitted dummies		216	27	

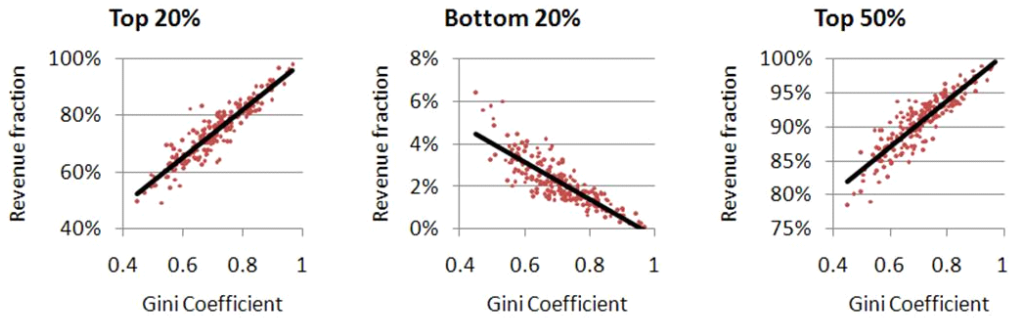
*Significant with $p \leq 0.05$
**Significant with $p \leq 0.01$
***Significant with $p \leq 0.001$



Regression coefficient

- AVGDEMAND: Average demand for books in the category
- AVG PAGERANK: Average PageRank for books in the category
- PAGERANKVAR: Variance in PageRank across books in the category
- SIZE: Number of books in the category
- AMIXING: Fraction of co-purchase links to books in the category that originate from a book within the same category

$$\text{Log[REVENUEGINI]}=a+b_1\text{Log[AVGDEMAND]}+b_2\text{Log[AVGPAGERANK]}+b_3\text{Log[PAGERANKVAR]}+b_4\text{Log[PAGERANKKURTOSIS]}+b_5\text{Log[SIZE]}+b_6\text{Log[MIXING]}+b_7\text{Log[CLUSTERING]}+b_8\text{Log[LISTPRICE]}+b_9\text{Log[SALEPRICE]}$$



The top two figures depict the estimated coefficients of the regression equation, on two separate graphs with different scales for clarity. Only coefficients that are significant at least at the 5% level are plotted. The bottom three panels further illustrate how the Gini coefficient measures the distribution of revenue across more and less popular books in a category. Consider a category with a Gini coefficient of 0.75. A doubling of the average PageRank of its books will, on average, be associated with a decrease of about 18% in its Gini coefficient, to about 0.61. Contrasting the corresponding revenue fractions associated with these two Gini values, this suggests a marked decrease in the fraction of revenue realized by the 20% of titles that are most popular, from about 75% of total revenue to about 60%. Similarly, it corresponds to an increase of about 50% (from about 2% to 3%) in the fraction of revenue realized by the 20% of titles that are least popular, and again, of about 50% (from 8% to 14%) in the fraction of revenue realized by the titles in the top half. While this example is for illustrative purposes, it is based on our empirical data, and indicates that the differences in revenue fractions from more and less popular products across categories with different average PageRanks is economically quite substantial.

Figure 7. Results of Model Estimation

Similarly, categories whose books have a higher average demand (measured by the variable *AvgDemand*) are less likely to have evenly distributed revenue, perhaps because their higher average demand is on account of having a higher number of very popular products. An alternative interpretation of these results is that when intrinsic (non-recommendation network) demand is higher, the added demand due to network traffic has a lower relative effect on the distribution of revenue. To understand this result, consider two categories, both with the same average PageRank: Category A, with low average demand, and Category B, with high average demand. Since both categories have the same average PageRank, they receive the same traffic from the co-purchase network (the same number of consumers flowing in). This means they sell the same number of books to consumers who arrived at the books' pages via the co-purchase network. The network traffic has a flattening effect in both cases. In other words, the fraction of revenue that can be attributed to the best-selling books is lower. However, the impact that same number of additional copies sold will have on the fraction of revenue that comes from the best-selling books will be lower for category B. Thus, since the traffic from the network accounts for a smaller fraction of category B's sales, the flattening effect will be smaller in magnitude.

The positive coefficients for average and variance of rating of books in the category (measured by the variables *AvgRating* and *VarRating*) further support the notion that categories with higher intrinsic popularity are less likely to have evenly distributed revenue.

Assortative Mixing

The *Mixing* variable represents the number of co-purchase links that both originate from and terminate at books in the category (Newman 2003). It is measured as the fraction of the total number of outgoing co-purchase links from books in the category that terminate at books in that category, and is a simple measure of assortative mixing within categories. Thus, categories with a higher *Mixing* value have more homogeneous inter-product recommendations, while categories with a lower *Mixing* value send a larger fraction of their network traffic to books in other categories. We find that a higher level of assortative mixing is associated with a lower Gini coefficient. In other words, revenue within categories with higher assortative mixing is more evenly distributed. A possible explanation is that when a category's recommendations are largely to and from products from within the same category, the redistribution of traffic stays largely within the category and, therefore, has a higher impact on flattening revenue. On the other hand, recommendations across categories are, on average, likely to terminate at more popular products and, thus, a high level of disassortative mixing in the category is indicative of a substantial fraction of the flow of traffic from the category to more popular products outside it.

Clustering

The *Clustering* variable represents how close the books in the category and their neighbors are to being a clique (Watts and Strogatz 1998). It is measured as the proportion of links between the vertices originating at a book's immediate neighbors which also terminate at one of the book's immediate neighbors. We find that a higher level of clustering is associated with a higher Gini coefficient. In other words, revenue within categories with lower local clustering is more evenly distributed, or more highly clustered neighborhoods are associated with higher revenue fraction inequity. A possible explanation for this latter finding is that the influence of recommendations from more popular products stays largely within small clusters of books when there is high clustering, rather than being spread around the network, and these recommendations thus play a smaller role in flattening revenue. This finding and its possible explanation are especially interesting because, while the effect is relatively small, it highlights the fact that, theoretically, recommendation networks could just as easily increase demand and revenue inequality. This further emphasizes the importance of the direction of our main empirical findings relating to the flattening of the demand and revenue distributions.

Variation in Network Influence

An increase in the variance of PageRank within a category (measured by the *PageRankVar* variable, an inverse measure of how equally the network's influence on a category is distributed among its books) is associated with an increase in its Gini coefficient. That is, after controlling for differences in average PageRank, a higher variance in the ranking is associated with increased revenue inequality. Consider an illustrative example, again, of two categories with the same average PageRank: Category A, where all books have the same PageRank, and Category B, where there are a few books with a much higher than average PageRank, and correspondingly a number of books with a lower than average PageRank. One would expect that the revenue flattening effect will be stronger for category A than for category B. After all, most of the traffic that goes into category B goes to the same few books and is likely to enhance the inequality in revenue, thus increasing the Gini coefficient. In contrast, all books in category A get the same additional traffic from the network, so the relative differences in revenue decrease, thus flattening the revenue distribution.

This result is reinforced by our finding that an increase in the kurtosis of the PageRank distribution within a category (measured by the *PageRankKurtosis* variable, a measure of the "peakedness" of the PageRank distribution) is associated with a decrease in its Gini coefficient. That is, after controlling for differences in the mean and variance of PageRank, a higher kurtosis in the ranking distribution is associated with flatter revenue distribution. While the economic impact of this effect is relatively small, it highlights an interesting nuance of the connection between the distribution of influence and the distribution of revenue. Figure 8 illustrates six categories which have essentially the same variance in PageRank but which have substantially different kurtosis (for example the kurtosis of category "Health, Mind & Body: Death and Grief" is four times the kurtosis for category "Biographies and Memoirs: Specific Groups"). Notice that a higher kurtosis distribution has a more acute peak around the mean (that is, a higher mass near the mean than a comparable normally distributed variable of values) and fatter tails (that is, a higher mass of extreme values than a comparable normally distributed variable). Our results suggest that these "fat tails" of network-induced demand associated with higher kurtosis have

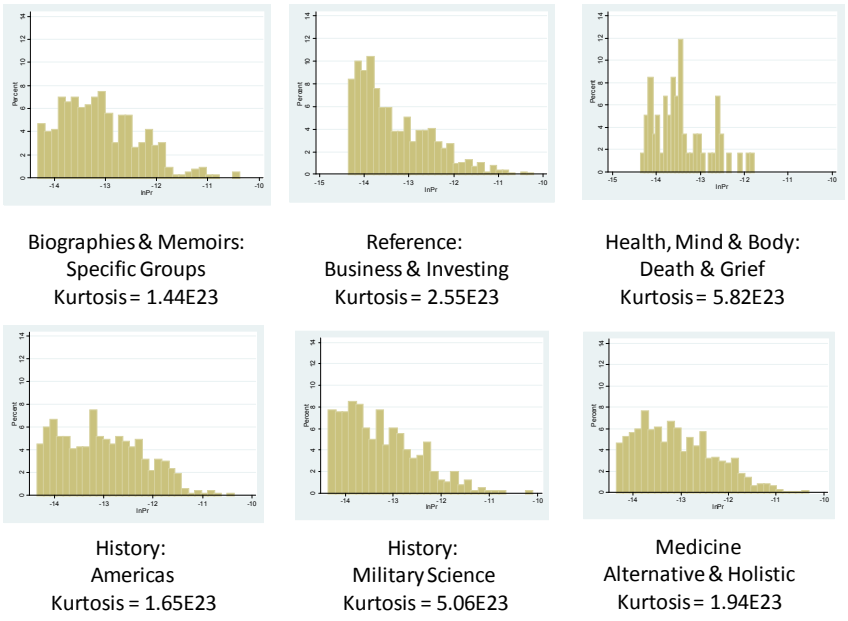


Figure 8. Plots of the Distribution of ln(PageRank) for Six Categories of Books with Variance Levels Within 10 Percent of Each Other

a dominant redistributive effect on demand (perhaps because of the low correlation between PageRank and SalesRank) which leads to an associated lower Gini coefficient. Put differently, in categories with high kurtosis, more network traffic is directed at the products which are at the tails of the influence distribution rather than in the middle. Consequently there is more dispersion in influence, and this leads to the network having a greater impact on flattening the distribution of revenue.

Revenue Versus Demand Distributions

We have replicated each of the results presented above for a model that studies the distribution of *demand* rather than revenue across categories. Strikingly, the results are directionally extremely similar. That is, an increase in the influence of the network *flattens the distribution of demand* across products as well. This is an important observation because it indicates that the revenue redistribution is not simply on account of niche products being inexpensive. These results are available on request.

Other Extensions

It is possible that the redirection of attention by a co-purchase link may cause revenue changes over a period of days rather than merely in the succeeding 24 hours. We explored this further by constructing composite weighted graphs for each of 22 overlapping seven-day intervals in February 2007, with weights on edges corresponding to the fraction of days they were present, implementing the WeightedPageRank measure on these networks, and estimating the relationship between the influence of the network and the revenue distribution measured over these overlapping week-long intervals. We did the same for four distinct seven-day composites. The results are strikingly similar to those summarized above, and are available on request.

Concluding Remarks

The long tail of e-commerce demand has been documented for a number of product categories sold in electronic markets. Many factors could be responsible for this demand redistribution, including an increase in product variety, lower search costs, and the redirection of attention due to outcome-based recommendations (Anderson 2006; Brynjolfsson et al. 2006). Our paper provides empirical evidence that relates the influence of one such recommendation network to the flattening of the demand and revenue distributions across more than 200 categories of books comprising a total of over 250,000 titles. We have used a global measure to quantify the

influence of such networked recommendations, and have computed measures of demand and revenue equality that control for variations in absolute demand levels and category sizes. To the best of our knowledge, this paper is the first study of its kind.

Our key findings are summarized below:

- We find that an increase in the influence of the recommendation network is consistently associated with a more even or flatter distribution of both revenue and demand. On average, a doubling of influence can increase the revenue from the bottom two deciles by up to 50 percent and reduce the revenue from the top two deciles from about 80 percent to about 70 percent of total revenue.

- Product categories with a higher number of titles and with a higher average demand display a shorter tail even with the same level of influence from the recommendation network. This is consistent with the conjecture that smaller categories with less popular products will have a more pronounced demand tail when influenced by recommendations.

- Holding average influence constant, the association between the influence of the network and flatter revenue distributions is enhanced when the influence is spread more evenly across the books in the category, rather than concentrated on a few books (popular or otherwise) within the category.

- The association between the influence of the network and flatter revenue distributions is enhanced when there is a high level of assortative mixing within the category's recommendations and a lower level of local clustering. Intuitively, when the recommendations originate and terminate within the category itself, the redistribution of attention they cause evens out attention more within the category, rather than redirecting high to a popular book in a different category. Similarly, when there is high clustering, the influence of recommendations is "trapped" within a small number of products rather than being spread around the network, and the long-tail effect of recommendations is thus diminished.

Our findings should be viewed as a starting point for further discussion and inquiry, rather than a final causal statement about the influence of recommendation networks on the distribution of global demand patterns. Specifically, we acknowledge that our estimates do not provide scientific evidence of causation, and what we report are associations between the influence of the recommendation network and flatter revenue and demand distributions. In a related paper (Oestreicher-Singer and Sundararajan 2012), we have provided a framework and a detailed set of estimates that allow us to make more precise causal statements about the extent to which influence from one's immediate neighbors affects demand at the individual product level. An ideal research setting for extending this to making stronger causal claims about changes in demand distributions might involve studying the introduction of a recommendation network at a new e-commerce firm or a content web site. We are exploring this possibility, and it remains an excellent direction for further research.

Redistributions of the kind we have reported on seem important for progress in general, because they can increase creative and scientific efforts by enabling a subset of innovators whose creations are not blockbusters to benefit from their innovation more easily. Similar IT artifacts like Google's Scholar are becoming increasingly accessed media for aggregating and evaluating topic-specific research papers. Scientific citations represent an acknowledgment of scientific influence and of having a shared topic; electronic recommendation networks convert this implicit acknowledgment into explicit hyperlinks. Our findings suggest that an increase in the influence of such content networks could lead to more equitable dissemination of the knowledge they aggregate.

To conclude, we are at an important and exciting point in time for the information systems field, and the social sciences in general. The availability of massive networked electronic data sets that contain information about individual-level connections among people and between products provides researchers with an unprecedented microscopic view into the nature of commercial and social interaction. There is a wide array of emerging research into understanding the content and influence of information flows within these networks. While the research thus far has focused more actively on social networks rather than those connecting products, these product networks give us the ability to understand the nature of demand at a more detailed level than has been possible in the past, an agenda which simultaneously serves commercial and academic objectives. By providing a new perspective on how these data can be used to study the influence of networked IT artifacts on business outcomes, we hope this represents a starting point for attaining these objectives.

Acknowledgments

We thank Vijay Gurbaxani, Ravi Bapna, and the members of the review team for their careful and detailed suggestions during the review process. We also thank Vasant Dhar, Nicholas Economides, William Greene, Panos Ipeirotis, Roy Radner, and seminar participants at the International Workshop and Conference on Network Science, the Wharton School, New York University, Tel-Aviv

University, the Zentrum für Europäische Wirtschaftsforschung, the Statistical Challenges in Electronic Commerce Research Symposium, the Telecommunications Policy Research Conference, the INFORMS Conference on Information Systems and Technology, and the International Conference on Information Systems for their feedback. Financial support from the NET Institute (http://www.netinst.org/) is gratefully acknowledged.

References

Anderson, C. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*, New York: Hyperion Press.

Aral, S., Brynjolfsson, E., and Van Alstyne, M. 2007. "Information, Technology and Information Worker Productivity," *Center for Digital Business Research Brief* (9:2) (available at http://ebusiness.mit.edu/research/Briefs/Aral_IWP_Task-Level-Evidence.pdf).

Aral, S., Muchnik, L., and Sundararajan, A. 2009. "Distinguishing Influence Based Contagion from Homophily Driven Diffusion in Dynamic Networks," in *Proceedings of the National Academy of Science* (106:51), pp. 21544-21549.

Bailey, J., Gao, G., and Lucas, H. 2008. "The Long Tail is Longer than You Think: The Surprisingly Large Extent of Online Sales by Small Volume Sellers," paper presented at the 20th Workshop on Information Systems and Economics, Paris, France, December 13–14.

Barrot, C., Rangaswamy, A., Albers, S., and Shaikh, N. I. 2008. "The Role of Spatial Proximity in the Adoption of a Digital Product," working paper, Christian-Albrechts University at Kiel.

Bradlow, E., Bronnenberg, B., Russell, G., Arora, N., Bell, D., Duvvuri, S., Hofstede, F., Sismeiro, C., Thomadsen, R., and Yang, S. 2005. "Spatial Models in Marketing," *Marketing Letters* (16:3), pp. 267-278.

Bramoulle, Y., and Kranton, R. 2007. "Public Goods in Networks," *Journal of Economic Theory* (127:1), pp. 478-494.

Brin, S., and Page, L. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems* (33), pp. 107-117.

Brin, S., Page, L., Motwani, R., and Winograd, T. 1999. "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report 1999-0120, Computer Science Department, Stanford University.

Brynjolfsson, E., Hu, Y. J., and Smith, M. D. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* (49:11), pp. 1580-1596.

Brynjolfsson, E., Smith, M. D., and Hu, Y. J. 2006. "From Niches to Riches: Anatomy of the Long Tail," *Sloan Management Review* (47:4), pp. 67-71.

Choi, J., and Bell, D. R. 2011. "Preference Minorities and the Internet," *Journal of Marketing Research* (48:4), pp. 670-682.

Clemons, E. K., Gao, G. G., and Hitt, L. M. 2006. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), pp. 149-171.

Dhar, V., Oestreicher-Singer, G., Sundararajan, A., and Umyarov, A. 2009. "The Gestalt in Graphs: Prediction Using Economic Networks," Working Paper No. CEDER-09-06, Stern School of Business, New York University (available at SSRN: http://ssrn.com/abstract= 1500834).

Dellarocas, C., Katona, Z., and Rand, W. 2010. "Media, Aggregators and the Link Economy: Strategic Hyperlink Formation in Content Networks," working paper (available at http://idei.fr/doc/conf/sic/dellarocas.pdf).

Desai, P. S. 2001. "Multiple Messages to Retain Retailers: Signaling New Product Demand," *Marketing Science* (19:4), pp. 381-389.

Elberse, A. "Should You Invest in the Long Tail,?" *Harvard Business Review* (86:7/8), pp. 88-96.

Fleder, D., and Hosanagar, K. 2008. "Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity," *Management Science* (55:5), pp. 697-712.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., and Yariv, L. 2010. "Network Games," *Review of Economic Studies* (77:1), pp. 218-244.

Gini, C. 1921. "Measurement of Inequality and Incomes," *The Economic Journal* (31), pp. 124-126.

Goh, K. H., and Bockstedt, J. 2008. "Unbundling and the Long Tail: New Evidence on the Consumption of Information Goods," paper presented at the 20th Workshop on Information Systems and Economics, Paris, France, December 13–14.

Goldenberg, J., Muller, E., and Libai, B. 2001. "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letters* (12:3), pp. 211-223.

Hervas-Drane, A. 2007. "Word of Mouth and Recommender Systems: A Theory of the Long Tail," Working Paper No. 07-41, Stern School of Business, New York University.

Hill, S., Provost, F., and Volinsky, C. 2006. "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science* (21:2), pp. 256-276.

Katona, Z., and Sarvary, M. 2008. "Network Formation and the Structure of the Commercial World Wide Web," *Marketing Science* (27:5), pp. 764-778.

Keller, E., and Berry, J. 2003. *The Influences*, New York: The Free Press.

Langville, A., and Meyer, C. 2005. "Deeper Inside PageRank," *Internet Mathematics* (1:3), pp. 335-380.

Lariviere, M., and Padmanabhan, V. 1997. "Slotting Allowances and New Product Introductions," *Marketing Science* (16:2), pp. 112-128.

Li, S., Baohong, S., and Wilcox, R. T. 2005. "Cross-Selling Sequentially Ordered Products: An Application to Consumer Banking Services," *Journal of Marketing Research* (42:2), pp. 233-239.

Lorenz, M. O. 1905. "Methods of Measuring the Concentration of Wealth," *Publications of the American Statistical Association* (9), pp. 209-219.

Mayzlin, D., and Yoganarasimhan, H. 2008. "Link to Success: How Blogs Build an Audience by Promoting Rivals," working paper, School of Management, Yale University.

Muller, E., Peres, R., and Mahajan, V. 2009. *Innovation Diffusion and New Product Growth*, Cambridge, MA: Marketing Science Institute.

Newman, M. E. J. 2003. "The Structure and Function of Complex Networks," *SIAM Review* (45:2), pp. 167-256.

Niraj, R., Padmanabhan, V., and Seetharaman, P. 2008. "A Cross-Category Model of Households' Incidence and Quantity Decisions," *Marketing Science* (27:2), pp. 225-235.

Oestreicher-Singer, G., and Sundararajan, A. 2012. "The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets," *Management Science* (forthcoming).

Oh, J. H., Susarla, A., and Tan, Y. 2008. "Examining the Diffusion of User-Generated Content in Online Social Networks," working paper, Department of Information Systems and Operations Management, University of Washington, Seattle.

Sriram, S., Chintagunta, P. K., and Agarwal, M. K. 2010. "Investigating Consumer Purchase Behavior in Related Technology Product Categories," *Marketing Science* (29:2), pp. 291-314.

Sundararajan, A. 2007. "Local Network Effects and Complex Network Structure," *The B. E. Journal of Theoretical Economics* (7:1), Article 46.

Tucker, C., and Zhang, J. 2008. "Long Tail or Steep Tail? A Field Investigation into How Online Popularity Information Affects the Distribution of Customer Choices," working paper, Sloan School of Management, Massachusetts Institute of Technology.

Valente, T. W. 1996. "Social Networks Thresholds in the Diffusion of Innovations," *Social Networks* (18:1), pp. 69-89.

Van Den Bulte, C., and Wuyts, S. 2007. *Social Networks and Marketing*, Cambridge, MA: Marketing Science Institute.

Watts, D. J., and Dodds, P. S. 2007. "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research* (34:4), pp. 441-458.

Watts D. J., and Strogatz, S. 1998. "Collective Dynamics of 'Small-World' Networks," *Nature* (393:6684), pp. 440-442.

Wedel, M., and Zhang, J. 2004. "Analyzing Brand Competition Across Subcategories," *Journal of Marketing Research* (41:4), pp. 448-456.

Zhao, X., Gu, B., and Whinston, A. 2008. "The Influence of Online Word-of-Mouth on Long Tail Formation: An Empirical Analysis," paper presented at the INFORMS Conference on Information Systems and Technology, Washington, DC, October 12-15.

About the Authors

Gal Oestreicher-Singer is an assistant professor at Tel Aviv University's Recanati Graduate School of Business Administration. Her research studies the effects of visible networks on electronic markets and the economics of digital rights management. Her prior research has won the 2008 ACM SIGMIS Best Dissertation Award, European Union Marie Curie Early Career Award, an INFORMS CIST Best Paper Award, an ICIS Best Overall Paper award, and a MSI-WIMI User Generated Content Research Competition Award. She received her Ph.D. from New York University in 2008, and holds degrees in law and electrical engineering from the Hebrew University in Jerusalem and Tel Aviv University.

Arun Sundararajan is the NEC Faculty Fellow and associate professor of Information, Operations and Management Sciences at New York University's Leonard N. Stern School of Business, where he also serves as director of the IS Ph.D. program. His research studies the economics of information technology. His current interests include the use of information technologies to facilitate economic development, piracy and DRM, platforms, ecosystem evolutoin for networks and social media, reputation systems, and how networks affect economic outcomes. He has published in journals that include *Decision Support Systems, Economics Letters, Information Systems Research, Journal of Economic Literature, Journal of Management Information Systems, Management Science, Proceedings of the National Academy of Science*, and *Statistical Science*. Arun has degrees in electrical engineering, operations research and business administration from the Indian Institute of Technology, Madras, and the University of Rochester.

MIS Quarterly

RESEARCH ARTICLE

RECOMMENDATION NETWORKS AND THE LONG TAIL OF ELECTRONIC COMMERCE

Gal Oestreicher-Singer
Recanati Graduate School of Business, Tel Aviv University, Tel Aviv 69978 ISRAEL {galos@post.tau.ac.il}

Arun Sundararajan
Stern School of Business, New York University, 44 West 4th Street,
New York, NY 10012 U.S.A. {asundara@stern.nyu.edu}

Appendix A

Algorithm for Data Collection

We use two computer programs for data collection. The first collects graph information and the second collects SalesRank information. Both use Amazon.com's XML data service. This service is part of the Amazon Web Services, which provide developers with direct access to Amazon's platform and databases.

Graph Collection

The program (crawler) which collects the graph starts at a popular book. It then traverses the co-purchase network using a depth-first search. Intuitively, in a depth-first search, one starts at the root (in our case, the one popular book chosen) and traverses the graph as far as possible along each branch before backtracking. At each page, the crawler gathers and records information for the book whose webpage it is on, as well as the co-purchase links on that page. The ASINs of the co-purchase links are entered into a LIFO (Last-In-First-Out) stack. If the algorithm finds it is on the page of a product that it has visited already, it backtracks and returns to the most recent product it has not yet finished exploring. The program terminates when the entire connected component of the graph is collected.

For example, in the graph in Figure A1, the nodes are numbered in the order in which the crawler will traverse the graph. In this case, the collection starts at node 1. Its co-purchase links are nodes 2, 6, and 7. Therefore, those numbers are added to a LIFO stack. The script will then proceed to node 2, whose co-purchases are nodes 3, 4, 5, and thus, those numbers will be added to the LIFO stack, which will now include: 3, 4, 5, 6, and 7. The script will continue to node 3. Since there are no co-purchase links to that node, it will move on to node 4. In the same way, the script will collect data about node 5, node 6, and node 7.

Since node 7 has co-purchase links—nodes 8 and 9—they will be added to the stack. After visiting nodes 8, 9, and 10, the data collection will terminate. As can be seen, the script only stops once it has collected information about the entire connected component. The collection of the entire connected component on Amazon.com takes between four and five hours. The script is run each day at midnight.

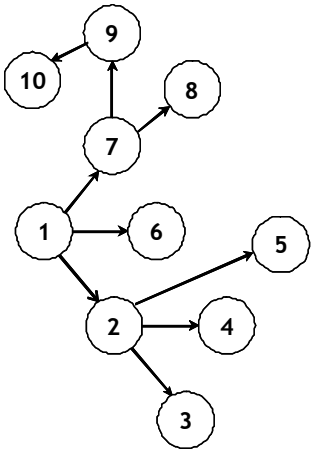


Figure A1. Depth-First Search Used for Graph Traversal

SalesRank Collection

A second computer program collects the demand information for all books on the graph every 3 hours for the 24-hour period following the collection of the graph. This script collects the SalesRank of each book that has ever appeared on the graph. Therefore, it follows the sales of some books that are no longer on the graph.

Appendix B

Converting SalesRank to Demand

SalesRank is a number associated with each product on Amazon.com that measures the product's demand relative to that of the other products sold on Amazon.com. The lower the number is, the higher the sales of that particular product. The SalesRank of a book is updated each hour to reflect recent and historical sales of every item sold on Amazon.com.

A formula to convert SalesRank information into demand information was first introduced by Goolsbee and Chevalier (2003). Their goal was to estimate demand elasticity. Their approach was based on making an assumption about the probability distribution of book sales, and then fitting some demand data to this distribution. They chose the standard distributional assumption for this type of rank data, which is the Pareto distribution (i.e., a power law). In the Pareto distribution, the probability that an observation's value exceeds some level S is an exponential function

$$\Pr(s > S) = \left(\frac{k}{s}\right)^{\theta} \tag{22}$$

where k and θ are the parameters of the distribution. The more important parameter is θ, the shape parameter that indicates the relative frequency of large observations. If θ is 2, for example, the probability of an observation decreases in the square of the size of the observation. With a value of 1, it decreases linearly.

For a given book, the number of books that have sales greater than those of that book is just one less than the book's rank. Therefore, the fraction of all books that have sales greater than those of a particular book is just [*SalesRank* – 1]/*TotalNumberOfBooks*. If there is a sufficient number of books to eliminate the approximation introduced by discreteness, then one can replace the equation above with:

$$\frac{[SalesRank - 1]}{TotalNumberOfBooks} = \left(\frac{k}{Demand(j)}\right)^{\theta} \tag{23}$$

Taking logs on both sides, and substituting θ with $-1/b$, this translates ranks into sales as follows:

$$Log[Demand(j)] = \alpha + bLog[SalesRank(j)] \tag{24}$$

The parameters a and b were estimated by Goolsbee and Chevalier using several parallel methods: by using data from the *Wall Street Journal* book sales index, which gives the actual quantity sold; by using sales information given by a publisher who sells on Amazon.com; and by conducting an experiment, buying copies of books with a steady SalesRank.

In a later study, Brynjolfsson et al. (2003), used data provided by a publisher selling on Amazon.com to conduct a more robust estimation of the parameters of the formula. They estimate the parameters as: $a = 10.526$, $b = -0.871$.

Appendix C

A More Detailed Description of PageRank

Let u be a web page. Let $F(u)$ be the set of pages u points to, and let $B(u)$ be the set of pages that point to u. Let $N(u)=|F(u)|$ be the number of links from u, and let c be a factor used for normalization (so that the sum of rank across all web pages is constant). A simple ranking, $R(u)$, is defined as

$$R(u) = c \sum_{v \in B(u)} \frac{R(v)}{N(v)} \tag{25}$$

This is a simplified version of PageRank. The rank of a page is divided among its forward links evenly to contribute to the ranks of the pages to which they point. Note that $c < 1$ because there are a number of pages with no forward links, and their weight is lost from the system. The equation is recursive, but it may be computed by starting with any set of ranks (commonly, equal rank for all pages) and iterating until convergence.

Stated another way, let A be a square matrix with the rows and columns corresponding to numbered web pages. Let $A(u,v) = \frac{1}{n(u)}$ if there is an edge from u to v, and $A(u, v) = 0$ otherwise. If we treat the rankings as a vector R over the linked pages, we have

$$R = cAR \tag{26}$$

So R is an eigenvector of A with eigenvalue c. In fact, the interesting one is the dominant eigenvector of A. It may be computed by repeatedly applying A to any non-degenerate start vector.

There is a small problem with this simplified ranking function. Consider two webpages that point to each other but not to any other page. Suppose there is some webpage that points to one of them. Then, during iteration, this loop will accumulate rank but never distribute any rank (since there are no outgoing edges). The loop forms a sort of trap which is called a *rank sink*. To overcome this problem of rank sinks, the damping factor $(1 - \alpha)$ is introduced. The normalization factor c is then set to α. Thus, the full ranking formula is

$$R'(u) = \alpha \sum_{v \in B(u)} \frac{R'(v)}{N(v)} + (1 - \alpha) \tag{27}$$

For further details and extensions, see Langville and Meyer (2005).

Appendix D

Correlation Matrix

The correlation matrix for our variables is presented in Table D1.

Table D1. Covariance Matrix

	RevenueGini	*AvgPagRank*	*PageRankVar*	*PagreRankKurtois*	*AvgDemand*	*Size*	*ListPrice*	*SalePrice*	*Mixing*	*Clustering*
RevenueGini	1.00									
AvgPageRank	0.10	1.00								
PageRankVar0.17	0.17	0.58	1.00							
PageRankKurtosis	0.09	-0.56	-0.25	1.00						
AvgDemand	0.55	0.05	0.06	0.10	1.00					
Size	0.19	0.19	0.15	0.18	0.03	1.00				
ListPrice	-0.13	0.22	0.16	-0.30	-0.12	-0.06	1.00			
SalePrice	-0.14	0.20	0.17	-0.28	-0.12	-0.06	0.94	1.00		
Mixing	-0.10	0.13	-0.09	0.10	-0.05	0.36	0.00	-0.02	1.00	
Clustering	-0.10	0.08	-0.23	0.17	0.14	-0.05	-0.41	-0.43	0.23	1.00

References

Brynjolfsson, E., Hu, Y. J., and Smith, M. D. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* (49:11), pp. 1580-1596.

Goolsbee, A., and Chevalier, J. A. 2003. "Measuring Prices and Price Competition Online: Amazon and Barnes and Noble," *Quantitative Marketing and Economics* (1), pp. 203-22.

Langville, A., and Meyer, C. 2005. "Deeper Inside PageRank," *Internet Mathematics* (1:3), pp. 335-380.