Working Paper #09-03

January 2009

Recommendation Networks and the Long Tail of Electronic Commerce

Gal Oestreicher-Singer
Tel-Aviv University and New York University

Arun Sundararajan
New York University

Recommendation Networks and the Long Tail of Electronic Commerce[1]

Gal Oestreicher-Singer[2] and Arun Sundararajan[3]

January 2009

Summary: 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. In this paper, we develop a novel method to test this conjecture and we report on results contrasting the demand distributions of books in over 200 distinct categories on Amazon.com. Viewing each product as having a unique position in a hyperlinked network of recommendations between products that is analogous to shelf position in traditional commerce, we quantify the extent to which a product is influenced by its recommendation network position by using a variant of Google's PageRank measure of centrality. We then associate the average level of network influence 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 recommendations have significantly flatter demand distributions, even after controlling for variations in average category demand, the category's size and measures of price dispersion. Our empirical findings indicate that doubling the average influence of recommendations on a category is associated with an average increase in the relative demand for the least popular 20% of products by about 50%, and a average reduction in the relative demand for the most popular 20% by about 12%. We also show that this effect is enhanced when there is assortative mixing in the recommendation network, and in categories whose products are more evenly influenced by recommendations. The direction of these results persist across time, across both demand and revenue distributions, and across both daily and weekly demand aggregations. Our work offers new ideas for assessing the influence of networks on demand and revenue patterns in electronic commerce, and provides new empirical evidence supporting the impact of visible recommendations on the long tail of electronic commerce.

[1] We thank Vasant Dhar, Nicholas Economides, William Greene, Panos Ipeirotis, Roy Radner and seminar participants at the 2007 International Workshop and Conference on Network Science, the Wharton School, New York University, Tel-Aviv University, the Zentrum für Europäische Wirtschaftsforschung, the Second Annual Statistical Challenges in Electronic Commerce Research Symposium, the 34th Annual Telecommunications Policy Research Conference, the INFORMS Conference on Information Systems and Technology and the 2006 International Conference on Information Systems for their feedback. Financial support from the NET Institute (http://www.netinst.org/) is gratefully acknowledged.

[2] Tel-Aviv University and New York University. *galos@post.tau.ac.il*

[3] New York University. *asundara@stern.nyu.edu*

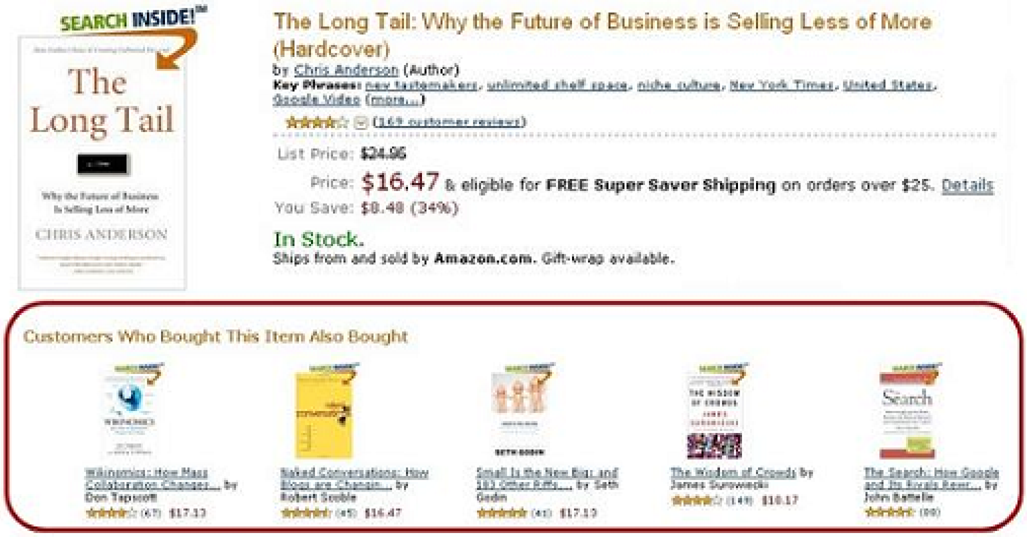
1. Introduction and Related Work

An important by-product of the sustained recent increase in electronic commerce and interaction is the *visible* emergence of a number of hyperlinked *networks* that connect products and their consumers. These include social networks like Facebook which link consumers, business networks like LinkedIn which link professionals, and co-consumption networks like those created by Amazon.com and YouTube which link products or content. Much like shelf position does in traditional commerce, it seems likely that *position* in the latter networks of interconnected products will influence a product's demand. After all, if one imagines the process of browsing an ecommerce site as being analogous to walking the aisles of a physical store, then the aisle structure of an online retailer is defined by a network of interconnected products whose landing pages link to each other, and the "position" of a product in this graph is thus analogous to its virtual shelf placement.

Perhaps the oldest example of a electronic and visible network of peer products is the "copurchase" network of Amazon.com[4], which, for many years now, has presented its consumers with links to complementary products made visible under the label “Consumers who bought this item also bought…”. This is illustrated in Figure 1.1. While consumers have always co-purchased complementary products, and these complementary products by definition influence each others' demand levels, the central conjecture of this paper is that the explicit *visibility* of these relationships is likely to *redistribute* the attention that each product receives from its potential consumers. This conjecture is in line with recent ideas that the wide product selection, costless search, unbundling and peer-based recommendations associated with ecommerce increase consumer awareness of relatively obscure products and cause ecommerce demand distributions to have a *long tail,* whereby less popular products constitute a larger fraction of total sales (Anderson, 2006, Brynjolfsson el at., 2006).

Anticipating a redistribution of attention and thus of demand on account of recommendation networks seems quite natural. However, the direction of this redistribution is not immediately intuitive. As predicted by the trade press, recommendation networks could increase the demand for niche products by making items that consumers might otherwise have not been aware of visible to them. In contrast,

[4] 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.


The Long Tail: Why the Future of Business is Selling Less of More (Hardcover)
by Chris Anderson (Author)
Key Phrases: new tastemakers, unlimited shelf space, niche culture, New York Times, United States, Google Video (more...)
(169 customer reviews)

List Price: ~~$24.95~~
Price: $16.47 & eligible for **FREE Super Saver Shipping** on orders over $25. Details
You Save: $8.48 (34%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Customers Who Bought This Item Also Bought

Wikinomics: How Mass Collaboration Changes... by Don Tapscott (67) $17.13

Naked Conversations: How Blogs are Changin... by Robert Scoble (45) $16.47

Small Is the New Big: and 183 Other Riffs,... by Seth Godin (41) $17.13

The Wisdom of Crowds by James Surowiecki (149) $10.17

The Search: How Google and Its Rivals Rewr... by John Battelle (89)

Figure 1.1: Illustrates the outgoing copurchases links for a sample book

however, visible recommendations based on copurchasing patterns might also increase the level of attention paid to popular products. Since they are frequently purchased, these products are also more likely to be co-purchased, and thus more likely to receive consumer attention via a recommendation link. Evidence about the anticipated and realized distribution of demand documented in the small but growing "long tail" literature thus far (Elberse and Oberholzer-Gee, 2006, Tucker and Zhang, 2008, Fleder and Hosanagar, 2008; more on these later) is actually mixed, suggesting a need for further investigation.

In this paper, we provide a new approach and new evidence that connects the *position* of products in recommendation networks to *aggregate outcome patterns* in ecommerce. Specifically, we study the extent to which the position of products in this kind of network will affect their relative *demand and revenue*. The idea is that the visibility of the network redirects the flow of consumer attention, which results in a redistribution of traffic, demand and eventually revenue. We analyze this empirically by relating the influence of Amazon's recommendation network to the demand distribution for over 200 categories of books, comprising over 250,000 titles sold on Amazon.com, over 25 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[5] and computing their associated Gini coefficient, a measure of inequality that is normalized for size and average magnitude.

Our empirical 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 in the category as well as the number of products in the category. We estimate that doubling of the average network influence on a category is associated with an *increase* in the relative demand for the least popular 20% of products by about 50%, and a *reduction* the relative demand for the most popular 20% by upto 12%. We also show that this effect is enhanced when there is more 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 direction of these results persist across all 25 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 associated with the widely documented phenomenon of the "long tail" of demand. The visibility of recommendation networks is one fundamental way in which ecommerce differs from traditional face-to-face commerce, and our findings suggest that their presence might explain some of the documented differences in demand patterns between the two settings.

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) has emphasized the role of wider product variety in driving sales away from popular products and towards the "tail" of the demand distribution. More recently, Brynjolfsson el at. (2006) compare sales of identical products online and via catalogs, and provide evidence supporting the theory that reduced search costs foster diversity in demand. Bailey, Gao and Lucas (2008) suggest 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

[5] The Lorenz curve is a widely used depiction of distributional equality, most commonly used to compare income distributions across regions and time.

than we think.

Current evidence, however, has suggested that the long tail effect may not be as simple as originally conjectured, and other effects may play a role in expanding demand for both hit and niche products. For example, Elberse and Oberholzer-Gee (2006) contrast the distribution of DVD sales between 2000 and 2005. They find evidence of a "lengthening" of the tail of demand in 2005, documenting a doubling of the number of products in the tail which regularly sold a small number of copies. However, in parallel, they provide 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 non-hit products, while negative word-of-mouth affects non-hit products more. Their paper takes a novel "micro" approach to the analysis, although not controlling 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 disagrregation of digital goods may actually shift demand towards more popular products and away from niche products that previously received a lot of their demand on account of being bundled with hits. Tucker and Zhang (2008) use a combination of research methods to show that the marginal benefits of visible popularity information (the number of consumers who have previously visited) is higher for niche products, and this in turn may contribute to a more prominent long tail.

Further, two recent papers provide theoretical arguements 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 recommendation on sales and ratings reinforce the popularity of already popular products. They do so with the caveat that their results do depend heavily on assumptions about how the recommender system works. In contrast, Hervas-Drane (2008) 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 mixed predictions of these theory papers further motivate our empirical assessment.

In contrast with Elberse and Oberholzer-Gee (2006), we document variations in demand distributions as well as associating these with a conjectured driver of this shift, the influence of a recommendation network. This driver of the long tail we focus on also differentiates our work from Brynjolfsson et al. (2007), whose focus is on reduced search costs and an online-offline comparison. Tucker and Zhang (2008) study the influence of a different kind electronic visibility on demand patterns; our work is different in its focus on recommendation networks, detailed inter-category comparisons, as well as a more nuanced measurement of influence. Our work is further distinguished from the prior 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 the intrinsic popularity allows us to focus more carefully on variations in the *distribution* of demand across categories.

Our work also draws from and adds to other related streams of prior work in information systems and marketing. First, while our results provide a new association between *online network position* and variations in observed demand/revenue, the idea that "position" affects demand is fairly well-established in the context of traditional bricks-and-mortar retailing, a point made repeatedly in the literature on shelf positioning and placement (this literature started with Cox, 1964 and Curhan, 1972; a more detailed survey of this literature is available in Oestreicher-Singer and Sundararajan, 2008). In contrast with this literature, we treat network position as given, focusing instead 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. This contrasts our work with the extensive literature on *slotting allowances/fees*[6]. Proponents of these fees emphasize its power as a

[6] Payments by manufacturers to persuade retailers to stock and display new products more prominently.

signaling tool when manufacturers have better demand information than retailers[7]. However, there is a tradeoff when delegating this decision because slotting fees can increase the channel power of retailers relative to smaller manufacturers (Shaffer, 1991) and bias demand distributions in favor of large manufacturers (Bloom, et al., 2000). Our data suggests that this tradeoff can be addressed by indirectly delegating slotting to the *consumers* by basing it on their observed shared purchasing patterns. Simply put, delegating slotting decisions "collectively" to the consumers (through the use of copurchase links) seems far more effective in mitigating issues of information asymmetry than delegating it to manufacturers; our evidence suggests that such a move also might mitigate the demand bias in favor of large manufacturers, because the resulting recommendation network redistributes attention towards niche products.

We also add to an extensive literature in the social sciences that provides evidence of the benefits garnered from an advantageous network position and other structural properties of one's local network, which might include the resources of one's direct network (Lin, 2001), the number and strength of ties (Bell et. al., 2007; Granovetter, 1973), closure or local clustering (Coleman 1990; Lin, 2001), along with other more subtle structural properties like the extent to which actors in a network span structural holes (Burt, 1992). A survey of this literature is beyond the scope of this paper; for more information, see Wasserman and Faust (1994).

A more recent literature has associated network properties with a variety of adoption and diffusion outcomes in organizations and markets. Within this literature, some work studies the influence of social networks on the diffusion of word-of-mouth (for example, Goldenberg et al, 2001, Van den Bulte and Wuyts, 2007), while other papers have paid closer attention to identifying agents that hold specific roles. For example, network positioning and properties of the network have been used to identify opinion leaders (Watts and Dodds, 2007; Keller and Berry, 2003) and innovators (Valente, 1996). Other papers[8] have studied the role of spatial proximity in the process of products and service adoption (Barrot et

[7] Chu (1992) and Desai (2001) compare slotting favorably to advertising efforts in this regard. Lariviere and Padmanabhan (1997) relate slotting fees to wholesale prices and retailer fixed costs, suggesting that signaling and cost compensation are the primary motivation for slotting allowances. Other advantages documented are of risk shifting, efficient shelf space allocation and enabling retail price reductions (Bloom, et al., 2000).

[8] See Bradlow et al. (2005) for a survey of spatial models in marketing, and Nair et al. (2006) for a detailed survey of the literature about peer effects in marketing.

al., 2008, Garber et al., 2004), and the extent to which network position and information diffusion affects the productivity and performance of employees in organizations (Aral et al., 2007) Features based on network structure have been shown to improve the predictions of data mining models used for targeted marketing (Hill, Provost and Volinsky, 2006). A recent literature on "network games" (for example, Bramoulle and Kranton, 2005, Galeotti et. al., 2008, Sundararajan, 2007) has begun studying theoretically how the properties of the equilibria of specific classes of IT-related games played on a graph depend on network structure. A more detailed survey is available in Sundararajan (2007).None of these papers view network position in the way that we do, or study its effects on aggregate demand patterns, across as extensive a range of titles, or over time. Further, rather than studying social influence between individuals in the context of adoption of one product or service, we consider the unique case of a network of products that is constructed by aggregating decisions made by individuals[9].

2. Overview of data and how it is collected

We use a large time series of recommendation networks for over 250,000 books sold on Amazon.com. Each product on Amazon.com has an associated webpage. These pages each have a set of “copurchase links” which are hyperlinks to the set of products that were copurchased most frequently with this product on Amazon.com. This set is listed under the title “Customers who bought this also bought:“. This was illustrated in Figure 1.1.

Conceptually, the copurchase network is a directed graph in which nodes correspond to products and edges correspond to directed copurchase links. We collect data about this graph using a Java-based crawler, which starts from a popular book and follows the copurchase 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 copurchase links on that page, and terminates when the entire connected component of the graph is collected. This is repeated daily. A sample part of the graph is illustrated in Figure 2.1. The algorithm used for data gathering is provided in Appendix A.

We have chosen to focus on books because they have, by far, the largest number of individual titles,

[9] A couple of papers do study influence in "product" networks although with fairly different questions. Mayzlin and Yonagarasimhan (2008) study hyperlinked competing blogs, and Oh et al (2008) study a network of videos on YouTube.

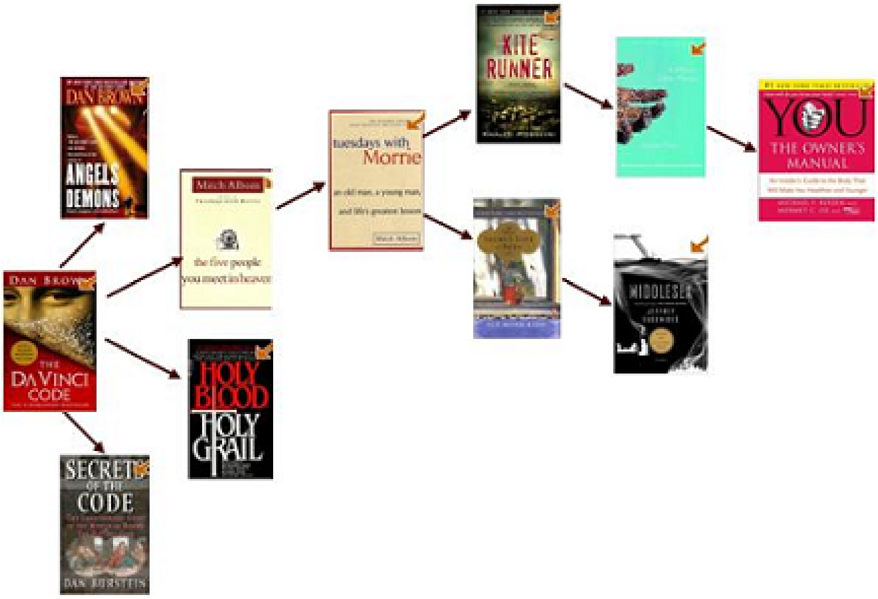


Figure 2.1: Illustrates a subset of paths in the graph

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 this category.

The data collection began in August 2005 and is currently ongoing. The graph is traversed every day. Apart from the copurchases, 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[10].

The following data is available for each book on the copurchase 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.

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

Copurchases: ASINs of the books that appear as its copurchases.

SalesRank: The sales rank is a number associated with each product on Amazon.com, which measures its demand of relative to 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 2.2 for a sample month, the component of the copurchase network we study changes substantially over time. It contains new nodes every day (over 6500 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 distribution remained quite stable through the month. Between 18% and 20% of the books have one incoming link, a little over 30% have two or three incoming links, roughly the same fraction have between 4 and 7 incoming links, and the in-degree distribution of the remaining 15% or so follows a power-law distribution.

3. Network position and demand distributions

This section describes how we construct our variables relating to network position/influence and the distribution of demand/revenue. Recall that Amazon.com uses a hierarchy of categories to classify its

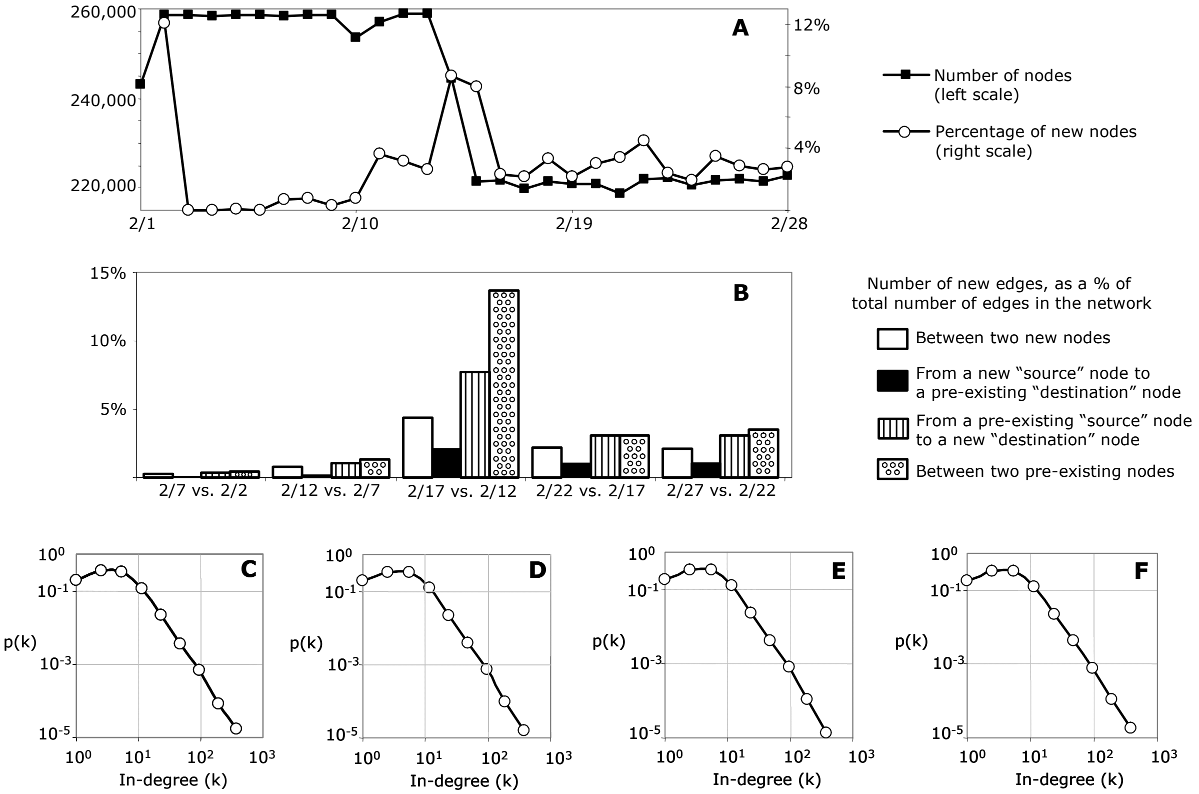


Figure 2.2: Summarizes the evolution of the copurchase network – the connected component of the graph I study. (A) plots the variation in the size of the graph and the differences between the identities of the nodes day-over-day. The changes to the edges on successive graphs are categorized in (B). The degree distribution (in-degree) is plotted after logarithmic binning of the data for February 1st (C), 10th (D), 19th (E) and 28th (F).

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).

Using the second level of the hierarchy, there are 1472 such categories across all books sold, of which between 203 and 225 have 100 or more nodes represented in our copurchase network.

In order to relate the network position of a product to variation in its demand, we follow the following sequence of steps

1. Quantify the distribution of demand. We characterize the demand distribution of each category by constructing its *Lorenz curve* and measuring its *Gini Coefficient* (more on this later).

2. Characterize the extent to which the position of a book in the copurchase network is related to the influence of the network on the book's demand by using *PageRank*, a measure of centrality.

3. Associate variation in (2) with variation in (1) at both a product-specific level of analysis and at a group-specific level of analysis. This is repeated for 25 different instances of the copurchase 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.

3.1. Quantifying the distribution of demand: the Gini coefficient

To quantify the demand distribution and comparing it across groups of books, we first have to partition the set of books. As mentioned earlier, Amazon.com places its books into a hierarchy of categories. We use this exogenous categorization as a natural grouping for comparing demand distribution across books. We have chosen the second level of the categorization hierarchy. There are over 1400 such categories across all books in our data set, of which between 203 and 225 have 100 or more books represented in our copurchase network, the minimum category size we analyze.

We quantify the shape of the demand 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 demand) and 1 corresponds to perfect inequality (where one book in the category has all the demand, and all other books in the category have zero demand).

The Gini coefficient is based on the Lorenz curve (Lorenz, 1905), a widely used depiction of distributional equality, most commonly used to compare income distributions across regions and time. In our analysis, the Lorenz curve of a category's demand (revenue) ranks the products in increasing order of sales (revenue), then plots the cumulative fraction $L(\rho)$ of sales (revenue) 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 $q(i)$ is the demand 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 demand 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 demand having the lowest rank. Next, define

$$S(r) = \{y \in N, R(y) \leq r\}, \tag{3.1}$$

which is 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:

$$L(\rho) = \frac{\sum\limits_{y \in S(n\rho)} q(y)}{\sum\limits_{y \in N} q(y)}. \tag{3.2}$$

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:

$$LU = \sum_{y=1}^{n} \left[1 - L(y/n)\right], \tag{3.3}$$

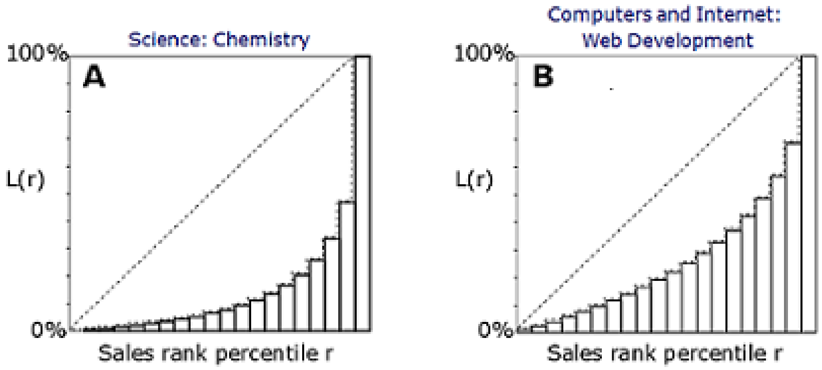


Figure 3.1: Illustrate the Lorenz curves for the "Computers and Internet: Web Development" and "Science: Chemistry" categories respectively. L(r) plots the fraction of the category's total demand from the books whose sales ranks are in the category's lowest r^2 percentile. (The data has been binned for 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 category's Gini coefficients are 0.751 (A) and 0.502 (B) respectively. Notice that a category whose demand is more highly concentrated on the higher-ranked products has a higher Gini coefficient.

and then using the identity

$$Gini = 2(LU) - 1. \tag{3.4}$$

Figure 3.1 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 demand 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.

3.2. Measuring network influence: Weighted PageRank

Our measure of the influence the recommendation network has on a product is called *WeightedPageRank*. This is a 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 on either an individual daily graph, or on an "average graph", constructed as a weighted composite of a time series of copurchase 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, based on the following model:

$$PageRank(i) = \frac{(1-\alpha)}{n} + \alpha \sum_{j \in G(i)} \frac{PageRank(j)}{OutDegree(j,k)}, \tag{3.5}$$

PageRank is based on a simple model of behavior – consumer who "surfs" the recommendation network randomly. This surfer follows any one of the links on a page with equal probability or jumps to a random page with probability $(1-\alpha)$ (this probability is also referred to as the “dumping 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-stage probability that a random surfer who starts at a random page will visit the specific page. Thus, a page can gain a high ranking by either having many pages pointing to it or having 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 mare information about the PageRank algorithm see Appendix C.

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:

$$WeightedPageRank(i) = \frac{(1-\alpha)}{n} + \alpha \sum_{j \in G(i)} Weight(j,i) \frac{WeightedPageRank(j)}{\sum\limits_{k \in F(j)} Weight(j,k)}, \tag{3.6}$$

where $Weight(j,i)$ is the fraction of the days that the link was present on the copurchase graph[11].

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 are exploiting the fact that fundamentally, Weighted PageRank measures the probability that a “random surfer” will arrive at a hyperlinked page if he were to traverse just the hyperlinks of the network. In other words, a product with a higher Weighted PageRank is *more likely to get traffic* from the network than one with a lower Weighted PageRank, and

[11] When computed on a single copurchase networks, $Weight(j,i) = 1$.

this therefore measures the *extent* to which the network we are interested in – the copurchase network – *influences* the product in question.[12]

4. 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 a very simple 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 ecommerce 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 pageviews) that product i gets is $\alpha_I(i)$, and the conversion rate associated with this attention is $c_I < 1$. The demand for product 1 and 2 are, respectively:

$$\begin{aligned} q_I(1) &= c_I \alpha_I(1); \quad (4.1) \\ q_I(2) &= c_I \alpha_I(2). \end{aligned}$$

Without any loss in generality, assume that $\alpha_I(2) > \alpha_I(1)$. It follows from (3.1) that $S(1) = \{1\}, S(2) = \{1, 2\}$, and after using (3.2) and (3.3) to compute the Lorenz upper area, one can show that the Gini coefficient for the category in the absence of the recommendation network is:

$$Gini_I = \frac{q_I(2) - q_I(1)}{q_I(2) + q_I(1)}, \quad (4.2)$$

which can be rewritten as:

$$Gini_I = \frac{1 - \frac{q_I(1)}{q_I(2)}}{1 + \frac{q_I(1)}{q_I(2)}}, \quad (4.3)$$

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

[12] For a survey on the use of PageRank in the literature, see Langville and Mayer, 2005.

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:

$$\begin{aligned} q_N(1) &= c_I'\alpha_I(1) + c_N\alpha_N(1); \qquad (4.4) \\ q_N(2) &= c_I'\alpha_I(2) + c_N\alpha_N(2), \end{aligned}$$

and correspondingly (following (3.1-3.4) and a sequence of analytical steps similar to those described above) the new Gini coefficient of the category is[13]:

$$Gini_N = \frac{1 - \frac{q_N(1)}{q_N(2)}}{1 + \frac{q_N(1)}{q_N(2)}}. \qquad (4.5)$$

It follows from (4.3) and (4.5) that $Gini_N < Gini_I$ if and only if

$$\frac{q_N(1)}{q_N(2)} > \frac{q_I(1)}{q_I(2)}, \qquad (4.6)$$

or if

$$\frac{c_I'\alpha_I(1) + c_N\alpha_N(1)}{c_I'\alpha_I(2) + c_N\alpha_N(2)} > \frac{c_I\alpha_I(1)}{c_I\alpha_I(2)}. \qquad (4.7)$$

Equation (4.7) can be rearranged as

$$[c_I\alpha_I(2)][c_I'\alpha_I(1) + c_N\alpha_N(1)] > [c_I\alpha_I(1)][c_I'\alpha_I(2) + c_N\alpha_N(2)], \qquad (4.8)$$

which, upon multiplying out and rearranging, reduces to:

$$\frac{\alpha_N(1)}{\alpha_N(2)} > \frac{\alpha_I(1)}{\alpha_I(2)}. \qquad (4.9)$$

One can use (4.7) to show that the condition in () holds if and only if the following condition holds:

$$\frac{\alpha_N(1)}{\alpha_N(2)} > \frac{c_I'\alpha_I(1) + c_N\alpha_N(1)}{c_I'\alpha_I(2) + c_N\alpha_N(2)}, \qquad (4.10)$$

[13] 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.

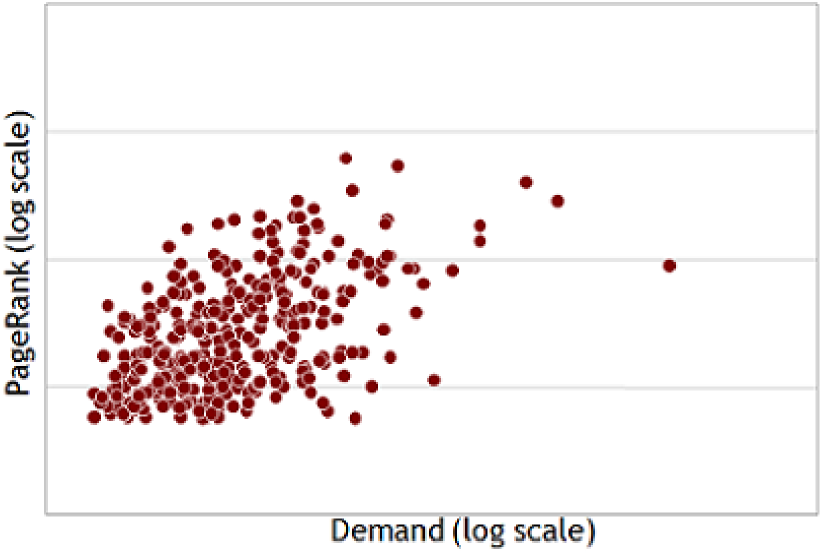


Figure 4.1: Plots SalesRank versus PageRank for a sample of the data. Illustrates the fact that while they are weakly (negatively) correlated, there are factors beyond network position that affect a product's demand.

or equivalently, if

$$\frac{\alpha_N(1)}{\alpha_N(2)} > \frac{q_N(1)}{q_N(2)}, \tag{4.11}$$

The above condition 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 4.1 and 4.2 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 (4.11) 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.

Addtionally, 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 recieves, on average, more attention from the network. Returning to our illustrative model, suppose the level of attention flowing from the network to

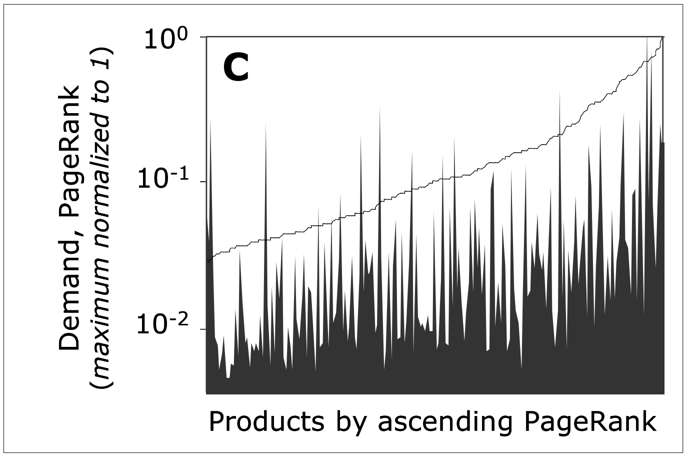


Figure 4.2: Contrasts the PageRanks (ascending line) of a random sample of books across all categories with their corresponding demand levels (dark spikes), with the maximum PageRank and demand levels in the sample normalized to 1. The correlation coefficient between demand and PageRank across the entire data set averaged 0.03 across the 28 days.

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 (4.4) to reflect the introduction of β:

$$\frac{q_N(1)}{q_N(2)} = \frac{c_I'\alpha_I(1) + \beta c_N \alpha_N(1)}{c_I'\alpha_I(2) + \beta c_N \alpha_N(2)}, \tag{4.12}$$

this in turn suggests that if the derivative of the RHS of (4.12) 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 (4.12) with respect to β:

$$\frac{d}{d\beta}\left[\frac{q_N(1)}{q_N(2)}\right] = \frac{c_N\alpha_N(1)}{c_I'\alpha_I(2) + \beta c_N \alpha_N(2)} - \frac{[c_I'\alpha_I(1) + \beta c_N \alpha_N(1)]c_N\alpha_N(2)}{[c_I'\alpha_I(2) + \beta c_N \alpha_N(2)]^2}, \tag{4.13}$$

which simplifies to

$$\frac{d}{d\beta}\left[\frac{q_N(1)}{q_N(2)}\right] = \frac{c_N c_I'[\alpha_N(1)\alpha_I(2) - \alpha_N(2)\alpha_I(1)]}{c_I'\alpha_I(2) + \beta c_N \alpha_N(2)}. \tag{4.14}$$

The RHS of (4.14) is positive if its numerator is positive, or if

$$\alpha_N(1)\alpha_I(2) > \alpha_N(2)\alpha_I(1), \tag{4.15}$$

which is precisely the condition of equation (4.9). In our illustrative model, we have therefore shown that if the distribution of attention to products generated by the network is more even than the intrinsic distribution of attention (the condition of equation (4.9)), 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.

To test this main conjecture, we estimate the relationship between a category's Gini coefficient (*GINI*) and the average PageRank of its books (*AVGPAGERANK*) using ordinary least-squares regression. We use a logarithmic transformations of all our variables to facilitate easily interpreting 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 category's average demand (*AVGDEMAND*), the number of books in the category (*SIZE*), and the fraction of copurchase links to the category's books that are from other books within it (*MIXING*) as control variables. We also used average price and the variance of price within a category as controls but these were never statistically significant. We thus report on our estimation the following reduced-form equation:

$$\begin{aligned} Log[GINI] &= a + b_1 Log[AVGDEMAND] + b_2 Log[AVGPAGERANK] \\ &\quad + b_3 Log[PAGERANKVAR] + b_4 Log[SIZE] + b_5 Log[MIXING] \end{aligned}$$

We estimate this equation independently for 25 randomly chosen days. Summary statistics for our data across the 25 days are provided in Table 4.1.

Variable	Range	Mean	StdDev
$GINI$	$0.36 - 0.97$	0.67	0.12
$AVGDEMAND^{*}$	$0.81 - 71.51$	3.48	4.50
$AVGPAGERANK^{**}$	$2.00 \times 10^{-6} - 7.32 \times 10^{-6}$	3.82×10^{-6}	7.69×10^{-7}
$PAGERANKVAR^{**}$	$1.72 \times 10^{-12} - 7.66 \times 10^{-10}$	6.92×10^{-11}	8.56×10^{-11}
$SIZE$	$100 - 10,657$	$1,002$	$1,543$
$MIXING$	$0.01 - 0.82$	0.32	0.18

Table 4.1: Summary statistics

The results of this estimation are summarized in Figure 4.3 and are strikingly consistent. We illustrate detailed results for one day in Table 4.2, and explain these results in some detail below.

Variable	Estimated Value (SE)
$Constant$	$-2.52^{***}(0.68)$
$AVGDEMAND$	$0.21^{***}(0.01)$
$AVGPAGERANK$	$-0.21^{***}(0.08)$
$PAGERANKVAR$	$0.04^{***}(0.01)$
$SIZE$	$0.03^{***}(0.01)$
$MIXING$	$-0.03^{***}(0.01)$
$R^2 = 0.83, n = 208$	* significant with $p \leq 0.05$ ** significant with $p \leq 0.01$ *** significant with $p \leq 0.001$

Table 4.2: Coefficient estimates for one sample day

Recommendation Networks and the Distribution of Demand: On each of the 25 days, categories with a *higher* average PageRank are associated with a significantly *lower* Gini coefficient. In other words, demand across categories with higher average PageRank is more evenly distributed. The coefficient value of the AVGPAGERANK variable ranged from -0.122 to -0.186, with the following interpretation: a doubling of the average PageRank of a category's books is associated with a 12.2%

to 18.6% *decrease* in the Gini coefficient of the category. (Across the 25 days, the highest average PageRank of a category was between 2.6 to 3.3 times higher than that of the lowest average PageRank, thus a doubling of average PageRank is a realistic notion.) Our results therefore establish that, based on a comparative analysis across over 200 categories of books, an increase in the extent to which the network structure is influential is associated with flatter demand, or an increase in the relative demand for niche (rather than blockbuster) products. Figure 4.3 further illustrates the shift in the fraction of demand obtained by the most and least popular books for a candidate doubling of influence of the recommendation network.

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

Category Size and Average Demand: We find that categories whose books have a higher average demand (measured by the variable AVGDEMAND) are less likely to have evenly distributed demand, perhaps because their higher average demand is on account of having a higher number of very popular products. Similarly, categories with more products (measured by the variable SIZE) are more likely to contain very popular products.

The categories in our data have between 100 and over 10,000 books in them. It is natural to assume that when all else is equal, a category with over 10,000 books is more likely to have higher variance in the demand for its books than a category with about 100 books. Further, the average demand of the category has a positive effect on the Gini coefficient of the category. A straightforward interpretation of these results is that as the intrinsic demand increases, the added demand due to network traffic has a lower relative effect on the distribution of demand. 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 copurchase 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 copurchase network. The network traffic has a flattening effect in both cases. In other words, the fraction of demand which 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 demand that come from the best selling books will be lower

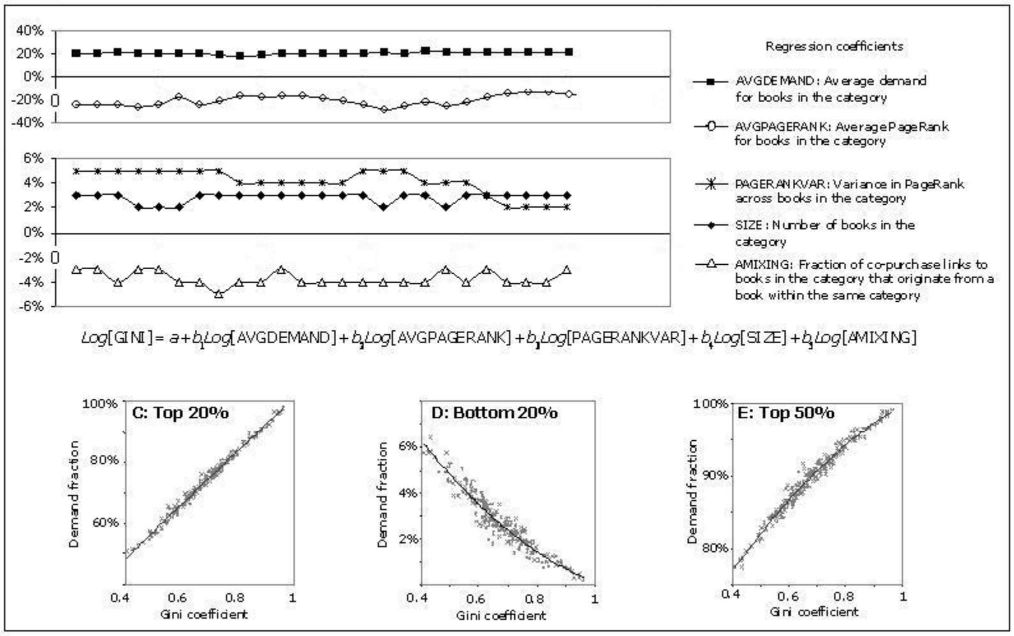


Figure 4.3: Results of model estimation. The top two figures depict the estimated coefficients of the regression equation, on two separate graphs with different scales for clarity. Only coefficients that are significant at least at the 5% level are plotted. Figures (C) through (E) further illustrate how the Gini coefficient measures the distribution of demand 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 16% in its Gini coefficient, to about 0.63. Contrasting the corresponding demand fractions associated with these two Gini values, this suggests a marked decrease in of the fraction of demand realized by the 20% of titles that are most popular, from about 80% of total demand to about 70%,. Similarly, it corresponds to an increase of about 50% (from about 2% to 3%) of the fraction of demand realized by the 20% of titles that are least popular, and again, of about 50% (from 8% to 12%) of the fraction of demand realized by the titles in the lower half. While this example is for illustrative purposes, it is based on our empirical data, and indicates that the differences in demand fractions from more and less popular products across categories with different average PageRanks is economically quite substantial.

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.

Assortative Mixing: The MIXING variable represents the number of copurchase links that both originate from and terminate at books in the category (Newman, 2003). It is measured as a fraction of the total number of outgoing copurchase links from books in the category that terminate at books in that category, and is a simple measure of assortative mixing within categories. We find that a higher level of assortative mixing is associated with lower Gini coefficient. In other words, demand 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 demand. 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 being to more popular products outside it.

Variance in Network Influence: Similarly, 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 demand inequality. To understand this result, consider two categories, both 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. It seems reasonable to conjecture that the demand 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 demand, 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 demand decrease, thus flattening the demand distribution.

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

Other extensions: We recognize that there may be sources of heterogeneity in the distribution of demand across books in a category that we do not observe (for example, the category may simply not produce bestsellers, or feature a diverse set of subtopics with distinct user bases). Towards accounting for these unobserved sources of variation across categories, we created a panel data set spanning on 28 consecutive days in February 2007, and estimated the relationship between Gini and average PageRank, using each of the control variables used earlier, and controlling for unobserved heterogeneity between categories using the fixed-effects transformation. The signs of the regression coefficients are identical to those reported in Figure 4.3, although the fixed effects transformation combined with our logarithmic transformation of the regression variables makes interpreting their magnitude difficult.

It is possible that the redirection of attention by a copurchase link may cause demand changes over a period of days rather than merely in the succeeding 24 hours. We explore 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 demand distribution measured over these overlapping week-long intervals. We did the same for four distinct 7-day composites. The results are strikingly similar to those summarized above, and are available on request.

5. Concluding Remarks

The long tail of ecommerce demand has been documented in a number of product categories sold online. It has been conjectured that 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 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 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 distribution of both revenue and demand across the books within a category. On average, a doubling of influence can increase the demand for the bottom two deciles by upto 50% and reduce the demand for the top two deciles from about 80% to about 70% of total demand.

- 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 a theory 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 demand distributions is enhanced when the influence is spread more evenly across the books in the category, rather than being concentrated on a few books (popular or otherwise) within the category.

- The association between the influence of the network and flatter demand distributions is enhanced when there is assortative mixing within the category's recommendations. Intuitively, when the recommendations originate and terminate from within the category itself, the redistribution of attention they cause evens out demand more within the category, rather than redirecting demand to a popular book in a different category.

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 demand distributions. In a related paper (Oestreicher-Singer and Sundararajan, 2008), we have provided a framework and a detailed set of estimates that allow us to make 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 ecommerce firm. We are exploring this possibility, and it remains an excellent direction for further research.

It is possible that the demand distributions of products in the last century – which often featured a high concentration of demand on a few popular products – were merely a historical aberration, and the flatter demand distributions that preceded the dominance of mass media are now returning. Redistribution of this kind seem important for progress in general, because it can increase creative and scientific efforts by enabling a subset of innovators whose creations are not "blockbusters" to benefit from their innovation more easily. Further, hyperlinked content networks such as Google's Scholar are becoming an increasingly accessed medium for aggregating and evaluating topic-specific research papers. The implicit acknowledgment of scientific influence and of having a shared topic that is embedded in scientific citations are converted into explicit hyperlinks by such networks; our findings suggest that an increase in the influence of these networks could lead to more equitable dissemination of the knowledge they aggregate.

6. References

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

2. Aral, S., Brynjolfsson, E. and Van Alstyne, M.,2007. Productivity Effects of Information Diffusion in Networks. Available at *http://ssrn.com/abstract=987499.*

3. 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." *Twentieth Workshop on Information Systems and Economics.*

4. Barrot, C., Rangaswamy, A., Albers, S., and Shaikh, N.I., 2008. "The Role of Spatial Proximity in the Adoption of a Digital Product," Mimeo.

5. Bell, D.R., and Song, S., 2007. "Social Contagion and Trial on the Internet: Evidence from Online Grocery Retailing," *Quantitative Marketing and Economics*, (5:4), pp. 361-400.

6. Bloom, P., Gundlach, G., and Cannon J., 2000. "Slotting Allowances and Fees: Schools of Thought and the Views of Practicing Managers," *Journal of Marketing*, (64:2), pp. 92-108.

7. 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.

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

9. 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.

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

11. 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.

12. 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.

13. Burt, R. S., 1992. "Structural Holes: The Social Structure of Competition," Cambridge, Mass.: Harvard University Press.

14. Chu, W., 1992. "Demand Signaling and Screening in Channels of Distribution," *Marketing Science*, (11:4), pp. 327-347.

15. 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.

16. Cox, K., 1964. "The Responsiveness of Food Sales to Shelf Space Changes in Supermarkets," *Journal of Marketing Research*, (1:2), pp. 63-67.

17. Curhan, R.C., 1972. "The Relationship between Shelf Space and Unit Sales in Supermarkets," *Journal of Marketing Research*, (9:4), pp. 406-412.

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

19. Elberse, A., and Oberholzer-Gee, F., 2006. "Superstars and Underdogs: An Examination of the Long Tail Phenomenon in Video Sales," Harvard Business School Working Paper, No. 07-015.

20. Fleder, D. and Hosanagar, K., 2008. "Blockbuster Culture's Next Rise or Fall: the Impact of Recommender Systems on Sales Diversity," *Management Science* (forthcoming).

21. Galeotti, A., Goyal, S., Jackson, M. O., and Vega-Redondo, F., 2008. "Network Games," Mimeo, Stanford University.

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

23. Goh, K. H., and Bockstedt, J., 2008. "Unbundling and the Long Tail: New Evidence on the Consumption of Information Goods." *Twentieth Workshop on Information Systems and Economics.*

24. 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.

25. 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.

26. Granovetter, M., 1973. "The Strength of Weak Ties," *American Journal of Sociology* (87:6) pp. 1360-1380.

27. Hervas-Drane, A., 2008. "Word of Mouth and Recommender Systems: a Theory of the Long Tail," Working paper.

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

29. Keller, E., and Berry, J., 2003. "The Influences," The Free Press, New York.

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

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

32. Lin, N., 2001. "Social Capital: A Theory of Social Structure and Action," Cambridge, U.K.: Cambridge University Press.

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

34. Mayzlin, D., and Yoganarasimhan, H., 2008. "Link to Success: How Blogs Build an Audience by Promoting Rivals,".Mimeo, Yale University.

35. Nair, H., Manchanda, P., and Bhatia, T., 2006. "Asymmetric Peer Effects in Prescription Behavior: the Role of Opinion Leaders". Mimeo, Stanford University.

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

37. Oestreicher-Singer, G. and Sundararajan, A, 2008. "The Visible Hand of Social Networks in Electronic Markets," Available at *http://ssrn.com/abstract=1268516*

38. Oh. J. H., Susarla, A., and Tan Y., 2008. "Examining the Diffusion of User-Generated Content in Online Social Networks," Working Paper, University of Washington.

39. Shaffer, G., 1991. "Slotting Allowances and Resale Price Maintenance: A Comparison of Facilitating Practices," *The Rand Journal of Economics*, (22), pp. 120-135.

40. Sundararajan, A., 2007. "Local Network Effects and Complex Network Structure," *Contributions to Theoretical Economics,* (7:1).

41. Tucker, C., and Zhang, J., 2008. "Word of Mouth and Recommender Systems: A Theory of the Long Tail," Working Paper, MIT.

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

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

44. Watts, D. J., and Dodds, P. S., 2007. "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research* (340 pp. 441-458.

45. Wasserman, S., and Faust, K., 1994. "Social Network Analysis: Methods and Applications," Cambridge, U.K.: Cambridge University Press.

46. Zhao, X., Gu, B., and Whinston, A., 2008. "The Influence of Online Word-of-Mouth on Long Tail Formation: an Empirical Analysis". *INFORMS Conference on Information Systems and Technology.*

A. Algorithms for Data Collection

We use two computer programs for data collection. The first collects graph information and the second collects sales rank information. Both use the Amazon.com's XML data service. This service is part of the Amazon Web Services, which provides 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 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 hadn't finished exploring. The program terminates when the entire connected component of the graph is collected.

For example, in the graph on Figure B.1, 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, 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, 7. The script will continue to node 3. Since there are no co-purchase links to that node, it will move 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 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.

Sales Rank 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 Sale Rank of all the books which have appeared on the graph. Therefore, it follows

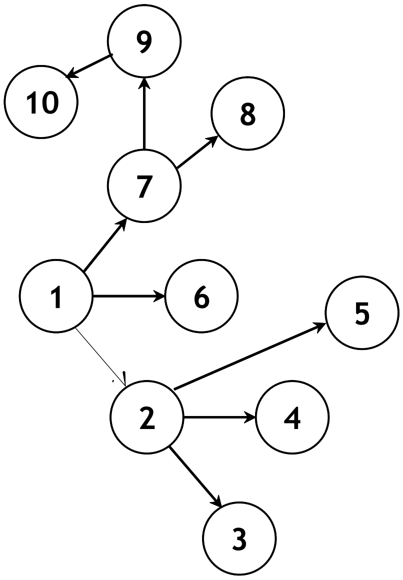


Figure A.1: Illustrates depth-first search used for graph traversal

the sales of some books that are no longer on the graph.

B. Converting Sales Ranks to Demand

SalesRank is a number associated with each product on Amazon.com and measures its 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 sales rank 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 Chevalier and Goolsbee (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 choose 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

$$\Pr(s > S) = \left(\frac{k}{S}\right)^{\theta}, \tag{B.1}$$

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

$$\frac{[SalesRank - 1]}{TotalNumberOfBooks} = \left(\frac{k}{Demand(j)}\right)^{\theta}. \tag{B.2}$$

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

$$Log[Demand(j)] = a + bLog[SalesRank(j)]. \tag{B.3}$$

The parameters a and b were estimated by Goolsbee and Chevalier using a couple of 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: $a = 10.526$, $b = -0.871$.

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 $B(u)$ 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:

$$R(u) = c \sum_{v \in B(u)} \frac{R(v)}{N(v)} \tag{C.1}$$

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 they point to. 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 column 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

$$R = cAR \tag{C.2}$$

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 web pages that point to each other but not to any other page. Suppose there is some web page which 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 “dumping factor” $(1 - \alpha)$ is introduced. The normalization factor c is then set to α. Thus, the full ranking formula is:

$$R'(u) = \alpha \sum_{v \in B(u)} \frac{R'(v)}{N(v)} + (1 - \alpha) \tag{C.3}$$

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