Network Structure and the Long Tail of Electronic Commerce[1]

Gal Oestreicher-Singer and Arun Sundararajan

Leonard N. Stern School of Business, New York University
44 West 4th Street, New York, NY 10012
goestrei@stern.nyu.edu, asundara@stern.nyu.edu

September 2006

Summary: We report on a research project that studies how network structures affect demand in electronic commerce, using daily data about the graph structure of Amazon.com's co-purchase network for over 250,000 products. We describe how the presence of such network structures alters demand patterns by changing the distribution of traffic between ecommerce web pages. When this traffic distribution generated by the presence of the network is less skewed than the intrinsic or "real world" traffic distribution, such network structures will even out demand across products, leading to a demand distribution with a longer tail. We estimate an econometric model to validate this theory, and report on preliminary confirmation by contrasting the demand distributions of products within over 200 distinct categories on Amazon.com. We measure the overall extent to which a product influences the network by adapting Google's PageRank algorithm, applying it to a weighted composite of graphs over four distinct 7-day periods, and we characterize the demand distribution of each category using its Gini coefficient. Our results establish that categories whose products are influenced more by the network structure have significantly flatter demand distributions, which provides an additional explanation for the widely documented phenomenon of the long tail of ecommerce demand.

1. Introduction

There are numerous networks associated with electronic business. Some of these can describe the relationships between consumers who communicate product information and influence each others purchasing, others can describe how the demand for different products are related based on shared purchasing patterns, and yet others may describe the patterns of trade between

[1] We thank Vasant Dhar, Nicholas Economides, William Greene, Panos Ipeirotis, Roy Radner and seminar participants at New York University, the Second Annual Statistical Challenges in Electronic Commerce Research Symposium and the 34th Annual Telecommunications Policy Research Conference for their feedback.

firms. Many recent studies in information systems have recognized the role that such network structure play in IS research as diverse as measuring the productivity of information workers, characterizing the effectiveness of information systems for knowledge, studying the internal dynamics of communities, and modeling the adoption of products that display local network effects.

In this paper, we report on a project that aims to model and measure the effect that such network structures have on outcomes in electronic commerce. A good example of such a structure is the network of product pages on an ecommerce site. Each product on an ecommerce site has a network position, which is determined by the products it links to, and those that link to it. If one imagines the process of browsing an ecommerce site as being analogous to walking the aisles of a physical store, then the ecommerce aisle structure is the this graph of interconnected products, and the network position of a product in this graph is analogous to its virtual shelf placement. A product that is linked to by an intrinsically popular one is likely to enjoy an increase in sales on account of this aspect of its network position. A product linked to by hundreds of others is likely to get more “network traffic” more than one linked to by just a few. Thus, both the structure of the networks and the nodes that comprise them seem to matter.

We measure the extent to which the position of a product in such a network structure will affect its demand, based on the idea that the network structure redirects the flow of consumer attention, which results in a redistribution of traffic and demand. The manner in which attention is redistributed can be measured using certain properties of the network structure, and these properties can be associated with observed variations in both individual and aggregate product demand. One specific prediction of our theory is that network structures with common degree distributions will even out traffic between products (nodes), thereby reducing demand inequity between products.

Our preliminary evidence is base on econometric estimates of how the intensity of the network structure affects the demand distributions of over 250,000 products within over 200 distinct categories on Amazon.com. Briefly, we compute an adapted version of the PageRank coefficient for each node of four composites of seven daily instances of Amazon.com's co-purchase network. We characterize the demand distribution of each category by constructing its Lorenz curve and measuring its Gini coefficient. We show that when network structure has a greater influence (when the average Weighted PageRank is higher) on a category, its demand distribution displays significantly lower inequity (its Gini coefficient is significantly lower). In other words, the presence of the network structure flattens ecommerce demand distribution. This provides a new explanation for the widely documented long tail of ecommerce demand.

2. Related work

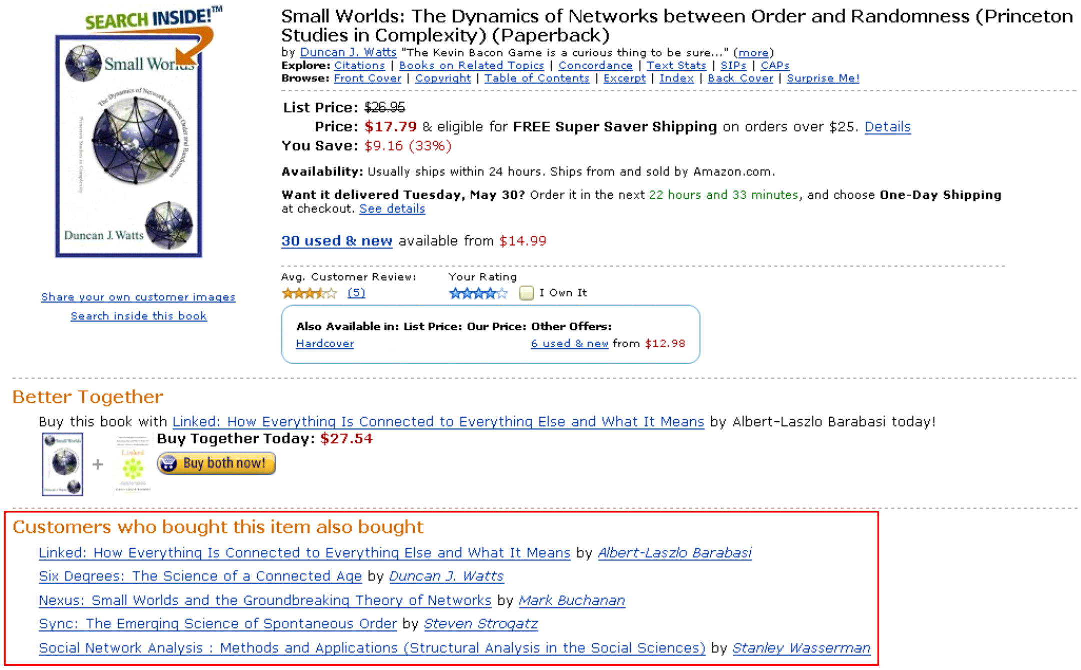
Many recent studies in the information systems field have recognized the role that such network structures play in IS. Those studies have begun to introduce concepts from social network and complex network research to different streams of IS research. For example, researchers have studied the evolution and development of network structures in online communities over time (Schoberth, Preece and Heinzl, 2003) and the internal dynamics of communities (Butler, 2001). Models of network structure have been recently applied to studying knowledge management by Alavi and Kane (2006), whose multimodal knowledge networks include both people and information systems as nodes within a single network, and facilitate exploring more complex interactions (like IS centrality) and their effects on knowledge sharing. Aral, Brynjolfsson and Van Alstyne (2006) study the influence of network structures on IT worker productivity, relating technology use and social network characteristics to economic measures, and providing evidence that the structure and size of workers' communication networks, including such social network metrics as betweenness and structural holes are highly correlated with performance. 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).

There has been some prior research in marketing which aimed to assess the structure of preferences for products based on purchase data. These studies have used scanner panel data, and are based on a similar notion: that such data contain important information based on revealed preference about the structure of brand preferences both within and across product categories. For example, the time series of purchasing data has been used to compute segment-product distances (Ramaswamy and DeSarbo, 1991), segment consumers with respect to brand preferences (Russell and Kamakura, 1997, Matthias, Bauer and Hammerschmidt, 2002) and build probabilistic models that provide spatial representations of product structure (Erdem, Imai, and Kean,1999). These are based on co-purchase bundles, and none of these papers has used a co-purchase network, or exploited any structural properties of these networks in making inferences about the nature of demand.

Beyond the scope of traditional IS research, network structures have received a significant amount of attention from sociologists studying relationships between people, from physicists and computer scientists studying the Web. Our work is also related to a growing literature on using network structures to create sophisticated ranking algorithms, one well-known contribution being the PageRank algorithm of Brin and Page, 1998 (for further information, see Langville and Meyer, 2005).

3. Data

We collect daily product, pricing, demand and “network” information for over 250,000 books sold on Amazon.com. Each product on Amazon.com has an associated webpage. Those pages each have 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" and is limited to 5 items (see Figure 3.1).



Small Worlds: The Dynamics of Networks between Order and Randomness (Princeton Studies in Complexity) (Paperback)
by Duncan J. Watts "The Kevin Bacon Game is a curious thing to be sure..." (more)
Explore: Citations | Books on Related Topics | Concordance | Text Stats | SIPs | CAPs
Browse: Front Cover | Copyright | Table of Contents | Excerpt | Index | Back Cover | Surprise Me!

List Price: ~~$26.95~~
Price: $17.79 & eligible for **FREE Super Saver Shipping** on orders over $25. Details
You Save: $9.16 (33%)

Availability: Usually ships within 24 hours. Ships from and sold by Amazon.com.

Want it delivered Tuesday, May 30? Order it in the next 22 hours and 33 minutes, and choose **One-Day Shipping** at checkout. See details

30 used & new available from $14.99

Share your own customer images
Search inside this book

Avg. Customer Review: ★★★½☆ (5)
Your Rating ★★★★☆ □ I Own It

Also Available in: List Price: Our Price: Other Offers:
Hardcover 6 used & new from $12.98

Better Together

Buy this book with Linked: How Everything Is Connected to Everything Else and What It Means by Albert-Laszlo Barabasi today!
Buy Together Today: $27.54

Customers who bought this item also bought

Linked: How Everything Is Connected to Everything Else and What It Means by *Albert-Laszlo Barabasi*
Six Degrees: The Science of a Connected Age by *Duncan J. Watts*
Nexus: Small Worlds and the Groundbreaking Theory of Networks by *Mark Buchanan*
Sync: The Emerging Science of Spontaneous Order by *Steven Strogatz*
Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences) by *Stanley Wasserman*

Figure 3.1: Illustrates sample copurchase links on Amazon.com

Conceptually, the co-purchase network is a directed graph in which nodes correspond to products, and edges 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 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 process is repeated daily.

We have chosen to focus on books because they are in the product category with by far the largest number of individual titles, whose product set is relatively stable (compared to electronics, for instance), and it seems to be a class of products for which the network we study would actually matter.

The data collection began in August 2005 and is currently ongoing. The graph is traversed every day, and we thus have over 300 co-purchase graphs collected so far. Apart from the co-

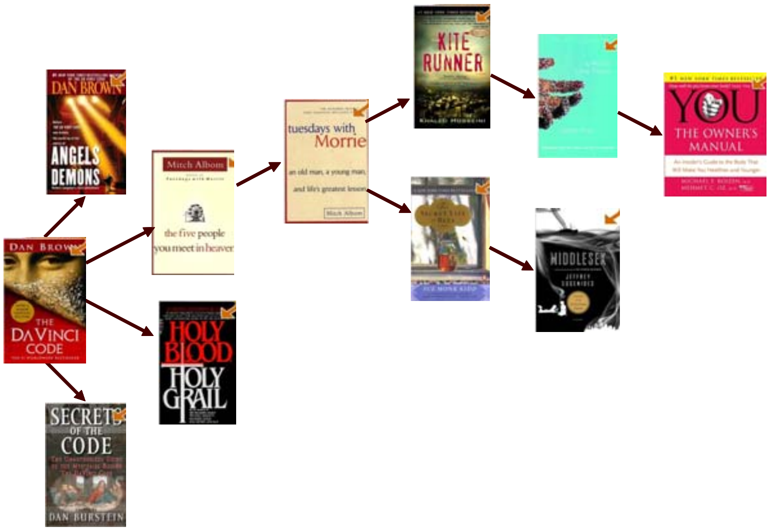


Figure 3.2: Illustrates a subset of paths in the graph

purchases, each book's ISBN, list price, sale price, category affiliation, secondary market activity, author, publisher, publication date, and consumer ratings are gathered. A sample part of the graph is illustrated in Figure 3.2

The following data that we gather are 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.

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).

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, the minimum category size we analyze.

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).

4. Characterizing ecommerce demand and its distribution

In order to relate the network position of a product to variation in its demand, we do the following:

1. Infer demand levels from the SalesRank data reported by Amazon.com, thereby associating a periodic demand level with each product.

2. Characterize the extent to which the network structure influences a product based on its network position.

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.

4.1. Estimating demand from Amazon.com salesranks

To estimate the actual level of demand, $Demand(j)$, of a book from its sales rank, $SalesRank(j)$, we use a conversion model suggested by Goolsbee and Chevalier (2003) and by Brynjolfsson, Hu and Smith (2003).

$$Log[Demand(j)] = a + bLog[SalesRank(j)] \tag{4.1}$$

This 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 a later study, Brynjolfsson, Hu and Smith (2003) use 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 $a = 10.526, b = -0.871$.

We have used the latter estimate in our study. In future work, we propose to conduct an independent purchasing and demand estimation experiment in order to update these estimates. Note, however, that since our results are all based on comparisons between categories, the fact that these parameters are dated are unlikely to affect our results directionally.

4.2. Quantifying the distribution of demand: the Gini coefficient

Next, we compute the Gini coefficient of each category of books. The Gini coefficient is a measure of distributional inequality, a number between 0 and 1, where 0 corresponds with perfect equality (in our case: where all the books in that category have the same demand) and 1 corresponds with perfect inequality (where one book has all the demand, and all other books have zero demand).

The Gini coefficient is based on the Lorenz curve, 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 ranks the products in increasing order of sales, then plots the cumulative fraction $L(\rho)$ of 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 recall that $q(i)$ 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

$$S(r) = \{y \in N, R(y) \leq r\}, \tag{4.2}$$

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:

$$L(\rho) = \frac{\sum\limits_{y \in S(n\rho)} q(y)}{\sum\limits_{y \in N} q(y)}. \tag{4.3}$$

Notice that the Lorenz curve is increasing and 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:

$$LU = \sum_{y=1}^{n} \left[1 - L(y/n)\right], \tag{4.4}$$

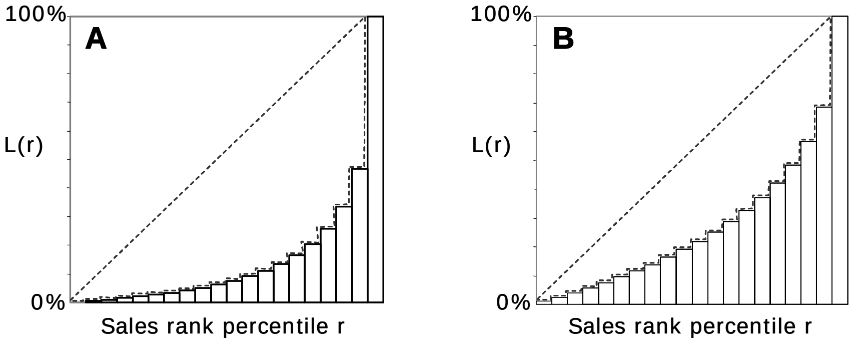


Figure 4.1: Illustrates the Lorenz curves and the Gini coefficients for two categories in our data set: Computers and Internet: Web Development (A) and Science: Chemistry (B) respectively. Their Gini coefficients are about 0.75 and 0.5 respectively.

and then using the identity

$$Gini = 2(LU) - 1. \tag{4.5}$$

Figure 4.1 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 average demand (or popularity), which facilitates comparing different categories despite their intrinsic differences and independent of their scale.

4.3. Two measures of network influence: Immediate Influence and Weighted PageRank

We have developed two different measures of network influence:

ImmediateInfluence: This is a measure of the traffic which flows into a product's webpage from its neighbors in the network. It is based on the assumption that the influence exerted by each product is proportionate to its total incoming traffic, is divided equally and flows to those products it has direct co-purchase links to (note, that this model does not allow back clicking). It therefore captures the influence of a product's immediate neighbors. Therefore, the influence that the co-purchase network has on demand depends on two factors: the local structure of the network and the amount of traffic associated with each link in the network. To combine the demand information with the structure of the network, we construct the ImmediateInfluence variable in the following way:

$$ImmediateInfluence(i) = \sum_{j \in G(i)} \frac{Demand(i)}{OutDegree(i)}, \tag{4.6}$$

where $G(i)$ is the set of books that link to book i.

WeightedPageRank: This is based on Google's PageRank algorithm, and iteratively computes the influence of the entire network on each product over time, although ignoring variations in intrinsic traffic across pages. It operates on an "average graph", constructed as a weighted composite of a time series of 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. This ranking forms the basis for Google's search engine. The PageRank algorithm is based on a simple model of behavior – a random surfer. 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.

We adapt the PageRank algorithm to account for the fact that we wish to measure the

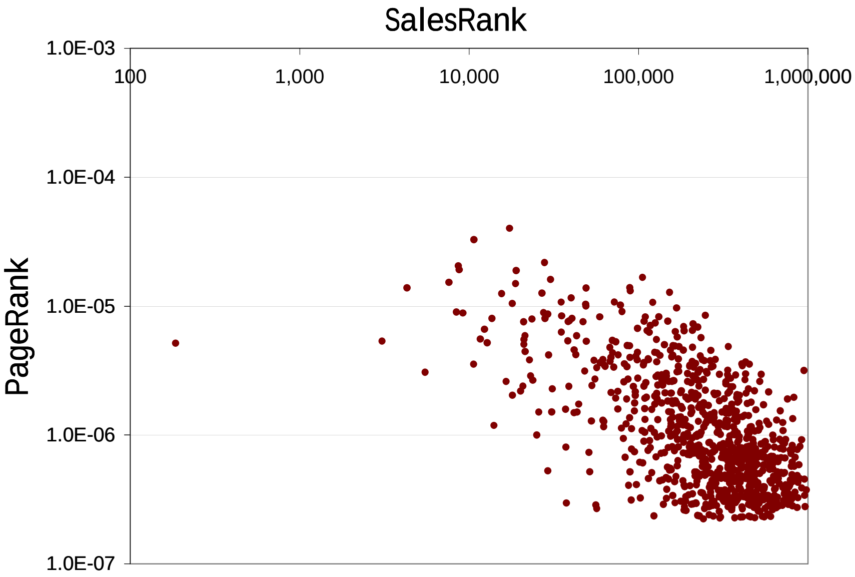


Figure 4.2: Plots SalesRank versus PageRank for a sample of the data. Illustrates the fact that while they are weakly (negatively) correlated, there are factors beyond network position that affect a product's demand.

average influence the network has on a product over four successive one-week periods. In our adapted model:

$$WeightedPageRank(i) = \frac{(1-\alpha)}{n} + \alpha \sum_{j \in G(i)} Weight(j,i) \frac{WeightedPageRank(j)}{\sum\limits_{k \in F(j)} Weight(j,k)}, \tag{4.7}$$

where $Weight(j,i)$ is the fraction of the 7 days that the link was present on the copurchase graph. The contrast between SalesRank and PageRank is illustrated in Figure 4.4.

It is important to note that while this measure is widely used in ranking algorithms (such as Google's), we use 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 this therefore measures the *extent* to which the network structure we are interested in – the co-purchase network –

influences the product in question.

The Weighted PageRank and the Immediate Influence measures described above are two different measures, which differ in the following key ways:

1. Weighted PageRank does not take the demand or intrinsic traffic variation across books into account. It is based only on the structure of the network. In contrast, Immediate Influence is based on both the structure of the network and the demand associated with each page.

2. The two measures use information about the structure of the network differently. Immediate Influence only includes the information about the immediate neighbors of the page, while Weighted PageRank measures the influence of the entire network.

5. How network structure influences demand and its distribution

The results we discuss in what follows were obtained using data for four distinct one-week periods between February 1st and February 28th . There is a seasonal demand pattern associated with sales around Valentine's Day in our data set, and we observed substantial changes in the edges of the co-purchase graph during this period (close to 20% of the edges changed). As will be seen, despite such changes in the identities of the books linking to each other, our results remained relatively stable. Some summary statistics of the daily graphs are in Figure 5.1.

We first study the variation between demand for each individual product, and the corresponding ImmediateInfluence of that product. Our results indicate that the immediate influence of a product explains a significant amount of the variation in the demand for the product. Our final section reports on refinements we are working on that account for endogeneity in these estimates.

What is more pertinent to our main results is the contrast between the distribution of

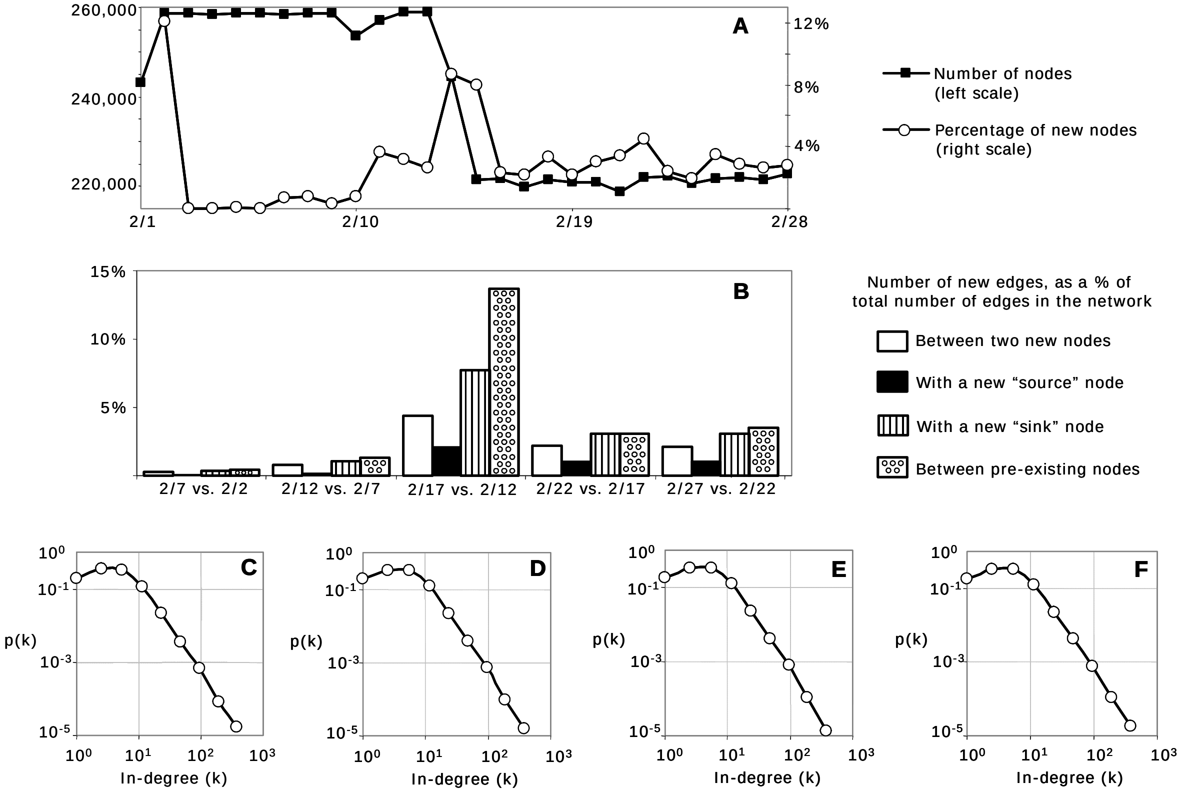


Figure 5.1: Graph statistics on the daily copurchase networks.

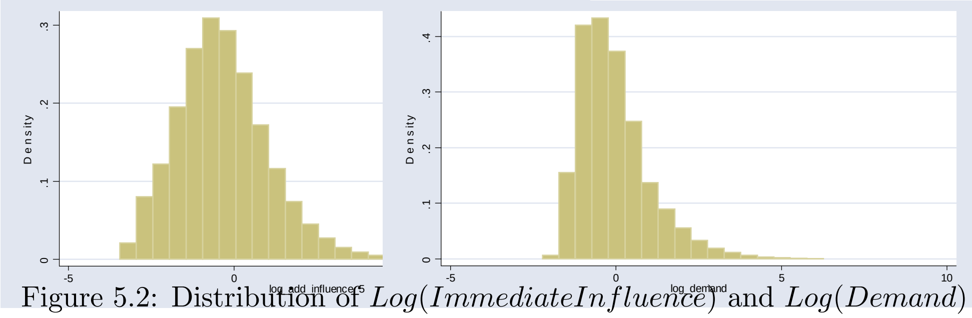
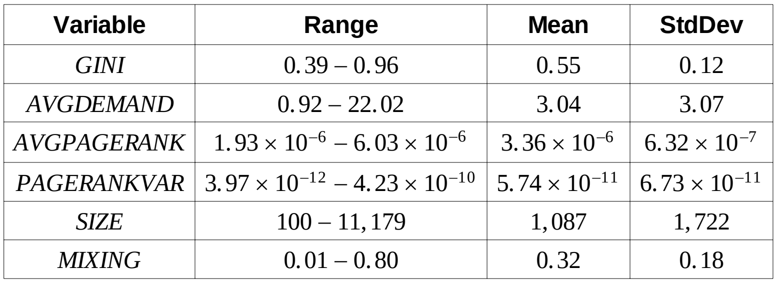


Figure 5.2: Distribution of $Log(ImmediateInfluence)$ and $Log(Demand)$

demand and Immediate Influence, which is illustrated in Figure 5.2. A casual examination of the distribution of demand and influence suggests the effect that the network structure might have on the demand distribution. As required by the definition of ImmediateInfluence, both have the same mean (influence is simply demand being redistributed), but the range and standard deviation of influence are larger than that of demand. This leads one to suspect that suspect that the network redistributes demand in a more “equal’ manner, a hypothesis we report on in the next section.

We now “zoom out” to an aggregated level (the category affiliation) and study how the network structure affects the market and the demand distribution. To study the effect of the network structure on the demand distribution, we group the books according to category affiliation. We test the hypothesis that a higher average weighted PageRank for a category will be associated with a lower Gini coefficient for that category. Following our interpretation of PageRank, a category whose products collectively have a higher average weighted PageRank, is, all else being equal, one whose products are influenced more by the network. The summary statistics for average weighted PageRank (and other variables we use as controls) are presented in Figure 5.3.

Using this data, we estimate the following reduced-form econometric model:



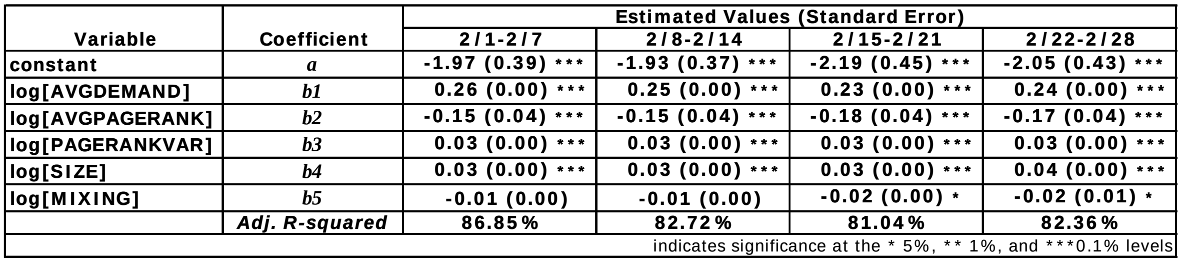
Variable	Range	Mean	StdDev
GINI	$0.39 - 0.96$	0.55	0.12
AVGDEMAND	$0.92 - 22.02$	3.04	3.07
AVGPAGERANK	$1.93 \times 10^{-6} - 6.03 \times 10^{-6}$	3.36×10^{-6}	6.32×10^{-7}
PAGERANKVAR	$3.97 \times 10^{-12} - 4.23 \times 10^{-10}$	5.74×10^{-11}	6.73×10^{-11}
SIZE	$100 - 11,179$	$1,087$	$1,722$
MIXING	$0.01 - 0.80$	0.32	0.18

Figure 5.3: Sample summary statistics

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

We chose a logarithmic specification because it facilitates ease of interpretation of the coefficients (in percentage terms), and because the empirical distributions of the transformed variables were more suitable for OLS regression.

The results of this estimation, presented in Figure 5.4, are striking. Based on a comparative analysis across over 200 categories of products, they establish that an increase in the extent to which the network structure is influential leads to a flattening of demand, or an increase in the relative demand for niche (rather than blockbuster) products. That is, we find that the average Weighted PageRank of the books in the category (the AVGPAGERANK variable) is negatively associated with the Gini coefficient of the category. This confirms that an increase in the extent to which network structure influences demand flattens the distribution of demand, or leads to a longer tail for demand, a phenomenon widely observed in electronic commerce (Anderson, 2004). Notice that this coefficient is not just statistically significant, but is economically significant as well. The highest average PageRank is generally about 3 times the lowest average PageRank. A doubling of average PageRank decreases the Gini coefficient by between 15% and 18% (since



Variable	Coefficient	Estimated Values (Standard Error)			
		2/1-2/7	2/8-2/14	2/15-2/21	2/22-2/28
constant	*a*	-1.97 (0.39) ***	-1.93 (0.37) ***	-2.19 (0.45) ***	-2.05 (0.43) ***
log[AVGDEMAND]	*b1*	0.26 (0.00) ***	0.25 (0.00) ***	0.23 (0.00) ***	0.24 (0.00) ***
log[AVGPAGERANK]	*b2*	-0.15 (0.04) ***	-0.15 (0.04) ***	-0.18 (0.04) ***	-0.17 (0.04) ***
log[PAGERANKVAR]	*b3*	0.03 (0.00) ***	0.03 (0.00) ***	0.03 (0.00) ***	0.03 (0.00) ***
log[SIZE]	*b4*	0.03 (0.00) ***	0.03 (0.00) ***	0.03 (0.00) ***	0.04 (0.00) ***
log[MIXING]	*b5*	-0.01 (0.00)	-0.01 (0.00)	-0.02 (0.00) *	-0.02 (0.01) *
	Adj. R-squared	86.85%	82.72%	81.04%	82.36%
indicates significance at the * 5%, ** 1%, and ***0.1% levels					

Figure 5.4: How network structure affects the distribution of ecommerce demand.

this is a log-log regression), which is pretty close to one standard deviation of Gini relative to its mean. Surprisingly, these results persist across four different weeks, one of which had substantial seasonal variation.

Moreover, the variance of the Weighted PageRank of different books within a category is positively correlated with the category's Gini coefficient. That is, after controlling for differences in average Weighted PageRank, a higher variance in the ranking (measured by PAGERANKVAR) is associated with increased inequality. To understand this result, consider two categories, both with the same average Weighted PageRank. Category A, where all books has the same Weighted PageRank and Category B, where there are a few books with a much higher than average Weighted PageRank, and correspondingly a number of books with a lower than average Weighted PageRank. It seems reasonable to expect that the 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.

The number of books in a category has a positive effect on the Gini coefficient. The categories in our data had 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 straight forward 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 Weighted PageRank. Category A, with low average demand and Category B, with high average demand. Since both categories have the same average Weighted PageRank, they receive the same traffic from the co-purchase network (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 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 for category A. Thus, since the traffic from the network accounts for a smaller fraction of category A's sales, the flattening effect will be smaller in magnitude.

6. Conclusions and ongoing work

We have briefly outlined a new economic theory of how network structures in electronic commerce might affect demand and cause ecommerce demand to be different from what is observed in traditional bricks-and-mortar commerce. We have gathered a new and unique data set comprising hundreds of observations of a giant component of the co-purchase network of Amazon.com, along with the relevant economic variables for each of its constituent products. We have provided the first evidence that the presence of these network structures can cause changes in demand patterns that are consistent with the observed "long tail" of ecommerce demand. We do so by adapting the PageRank algorithm to measure the influence that the network structure has on each product, and then contrasting variations in the average such measure across categories that have different demand distributions. To the best of our knowledge, ours is the first study of its kind.

Our current work aims to extend these results in the following salient ways:

- Rather than being the "random" surfers used by the PageRank model, ecommerce consumers are strategic. They tend to visit more popular products more often, and their purchasing is affected by other economic variables like price, customer reviews and product age. Our first extension aims to develop a model of a "strategic surfer" that is grounded in more familiar economic theory, but with retains sufficient structure to allow the iterative estimation of the "importance" of the network. We have made substantial progress on this front, solving a first model. This gives us a basis for a structural model.

- The methods we have used so far do not explicitly separate demand that is caused by the presence of the hyperlinks associated with the network structure with the demand variation that complementary products might naturally experience together. Identifying these distinctly requires appropriate instrumental variables. We have begun experimenting with constructing suitable variables of this kind (using lagged demand, and contrasting demand for identical products over successive days over which the link appeared). Since we have a time series of over 300 days of data, with well over 200,000 products per day, we are optimistic that we can identify this suitably for a subset of our data. This will facilitate a clearer understanding of the value of such network structures as strategic IT design variables that are unique to ecommerce.

7. References

1. Anderson, C. The Long Tail. http://www.wired.com/wired/archive/12.10/tail_pr.html

2. Aral, S. Brynjolfsson, E., and Van Alstyne, M., 2006. Information, Technology and Infor-

mation Worker Productivity: Task Level Evidence. Mimeo, MIT.

3. Brin, S., and Page, L., 1998.The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 33:107–117.

4. 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.

5. Brynjolfsson, E., Smith, M. D., and Hu, Y. J., 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. *Management Science,* 49.

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

7. Butler, B., 2001. Membership Size, Communication Activity, and Sustainability: The Internal Dynamics of Networked Social Structures. *Information Systems Research* 12, 346–362.

8. Erdem, T., Imai, S., Kean, M. P., 1999. Econometric Modeling of Competition: A Multi-Category Choice-Based Mapping Approach,. *Journal of Econometrics* 89.

9. Ghose, A., M. Smith, and R. Telang. 2006. Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact. *Information Systems Research* 17, 3-19.

10. Goolsbee, A., and Chevalier, J. A., 2003. Measuring Prices and Price Competition Online: Amazon and Barnes and Noble. *Quantitative Marketing and Economics* (1) 203-222.

11. Hill, S. F. Provost and C. Volinsky, 2006. Network-based Marketing: Identifying Likely Adopters via Consumer Networks. *Statistical Science* 21, 256-276.

12. Kane, J. and Alavi., M., 2006. Casting the Net: Towards a Theory of Multimodal Knowledge Networks. Mimeo, Emory University.

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

14. Matthias, S., Bauer, H. H., and Hammerschmidt, M., 2002. Structuring Product Markets: An Approach Based on Customer Value. *Journal of the American Marketing Association.*

15. Newman, M. E. J., 2003, The structure and function of complex networks. *SIAM Review* 45

16. Ramaswamy, V., DeSarbo, W. S., 1990, SCULPTRE: A New Methodology for Deriving and Analyzing Hierarchical Product-Market Structures from Panel Data. *Journal of Marketing Research* 27(4)

17. Russell, G. J., Kamakura, W. A. 1997, Modeling Multiple Category Brand with Household Basket Data. *Journal of Retailing* 73(4)

18. Schoberth, T. Preece, J. and Heinzl, A., 2003. Online communities: a longitudinal analysis of communication activities. *Proceedings of the 36th Hawaii International Conference on Systems Sciences.*

19. Sundararajan, A., 2006. Local Network Effects and Complex Network Structure. *Contributions to Theoretical Economics* 6 (1), in press.

20. Watts, D. J. and Strogatz, S. H., 1998, Collective dynamics of ‘small-world’ networks. *Nature* 393