View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2006 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2006

The Impact of Network Structures on Electronic Commerce

Gal Oestreicher
New York University

Follow this and additional works at: http://aisel.aisnet.org/amcis2006

Recommended Citation

Oestreicher, Gal, "The Impact of Network Structures on Electronic Commerce" (2006). *AMCIS 2006 Proceedings*. 100.
http://aisel.aisnet.org/amcis2006/100

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

The Impact of Network Structures on Electronic Commerce

Gal Oestreicher-Singer
New York University
44 West 4th Street,
New York, NY 10012
goestrei@stern.nyu.edu

INTRODUCTION

There are numerous networks associated with electronic business[1]. Information systems have been used for many years to gather and analyze detailed data about the structure of some of these networks[2]. More recently, electronic commerce sites have made publicly available data about "co-purchase networks": retailers like Amazon.com determine the most frequent co-purchases of many of their products, and publish this information on the web pages on which these products are listed for sale.

The set of products that consumers actually pay attention to is affected by the hyperlinks between these pages, in contrast with what a model of costless electronic search might suggest. In other words, products on an ecommerce site have a "network position", which is determined by the products and other pages it links to, and those that link to them. 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 ecommerce site is the graph of interconnected products, and the location of a product in this graph (or the network position of the product) is analogous to its virtual shelf placement. For example, a product with an incoming link from another product that happens to be intrinsically popular is likely to enjoy an increase in sales on account of this aspect of its network position. A web page with 200 incoming links is likely to benefit more than one with only a few incoming links. Thus, both the structure of the network and the identity of the nodes that comprise these networks matter.

RESEARCH QUESTIONS

My dissertation proposes to study such network structures in electronic commerce. Broadly, it will address the following questions:

(A) Can hyperlinked network structures explain observed variations in e-commerce outcomes?

(B) Can hyperlinked network structures be used to predict outcomes in e-commerce?

My first essay will address question (A). It will develop methods for measuring the influence a network structure has on demand and associate variations in these measures of influence with variations in demand.

My second essay will address question (B). Hyperlinked networks, such as the co-purchase network, are generated by the decentralized economic actions of consumers (rather than being centralized strategic seller choices). Thus, they could contain information that is better suited for certain predictions.

Some key contributions that answering these questions will make:

- The presence of hyperlinked network structures is one fundamental way in which electronic commerce differs from traditional face-to-face commerce. This presence may explain some of the documented differences in demand between these two settings (my preliminary results provide the first empirical evidence that network structures do indeed explain the "long tail", or the equalizing of demand across popular and niche products in electronic commerce).

[1] Some of these can describe the relationships between consumers, 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 firms.

[2] For example, a number of retailers give their customers frequent shopper cards in order to understand what groups of products are purchased by the same consumer.

- The hyperlinked network structures may be viewed as an IT-based design variable for electronic retailers; so identifying properties of structures that are associated with desirable outcomes seems important.
- An analysis of the structure of co-purchase networks can deepen our understanding of relationships between products in electronic commerce, which can lead to more effective business intelligence systems.
- Add to a growing set of studies in information systems that recognize the influence that network structure can have on a variety of outcomes.

(A summary of the related literature is available on request)

OVERVIEW OF THE DATA AND HOW I COLLECT IT

I 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[3]. Conceptually, the co-purchase network is a directed graph in which nodes correspond to products, and edges to directed co-purchase links[4]. I 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 is repeated daily. The co-purchase links of a sample book are illustrated in Figure 1.

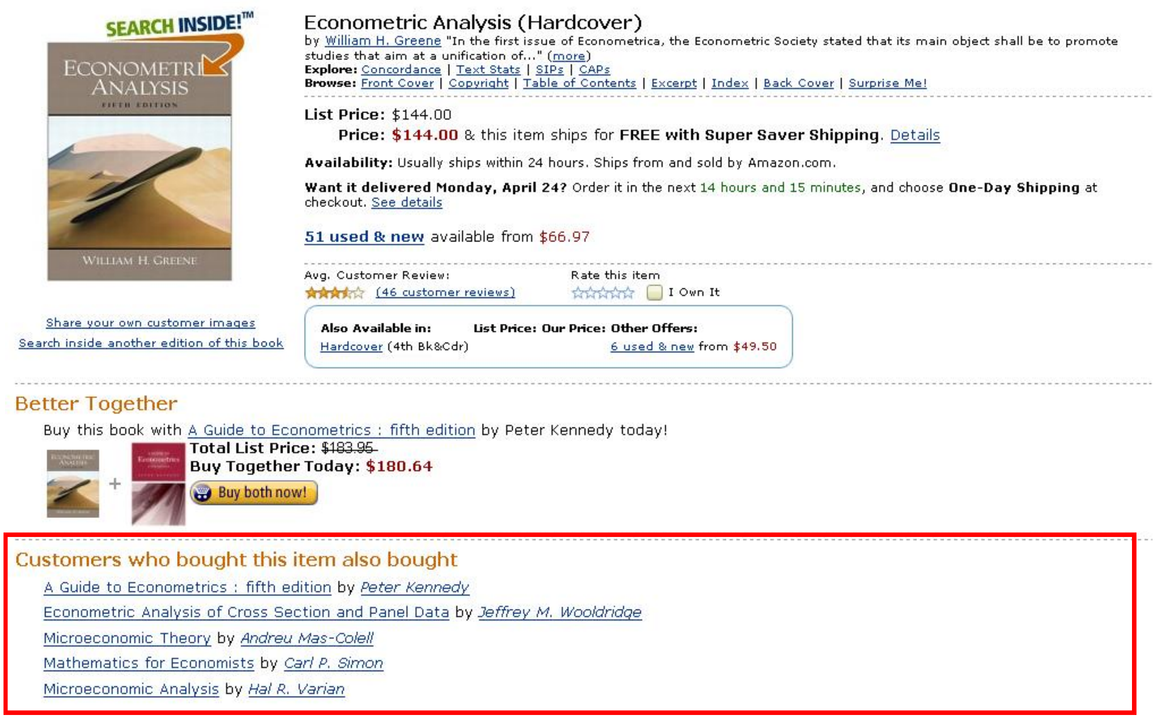

Figure 1: Illustrates the co-purchase links for a sample book

[3] This set is listed under the title "Customers who bought this also bought:" and is limited to 5 items.

[4] The full co-purchase graph is inherently symmetric. However, Amazon only displays the top 5 most frequently co-purchases of each book, and those are not necessarily symmetric. Therefore, the resulting co-purchase graph is directed.

The data collection began in August 2005 and is currently ongoing. Apart from the co-purchases, each book's ISBN, list price, sale price, category affiliation, sales rank, secondary market activity, author, publisher, publication date, binding information, and consumer ratings are gathered.

ESSAY 1: HOW NETWORK STRUCTURE AFFECTS ECOMMERCE DEMAND

My first essay studies the extent to which the position of products in the network structure can explain variations in their demand. The presence of the hyperlinks associated with network structures alter the distribution of traffic across the web pages of an electronic commerce site, and thus its influence can be measured by comparing groups of products which differ in the extent to which the network influences their demand. In order to relate the network position of a product to variation in its demand, one needs to

(a) Infer demand levels from the SalesRank data reported by Amazon.com

(b) Characterize the extent to which the network structure influences a product based on its network position.

(c) Associate variation in (b) with variation in (a)

Measuring demand and its distribution

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

$$Log[Demand(j)] = a + bLog[SalesRank(j)]$$

I use the values $a = 10.526$, $b = -0.871$ computed for books on Amazon.com by Brynjolfsson, Hu and Smith. I then compute the Gini coefficient of each of 182 categories of books. The Gini coefficient is a measure of distributional inequality, a number between 0 and 1[5]. To compute the Gini coefficient of each category one first ranks the books in the category in increasing order of their demand, thereby constructing the Lorenz curve. The Gini coefficient is calculated as a ratio of the area under the Lorenz curve to the area under the "perfect equality" line [figure 3].

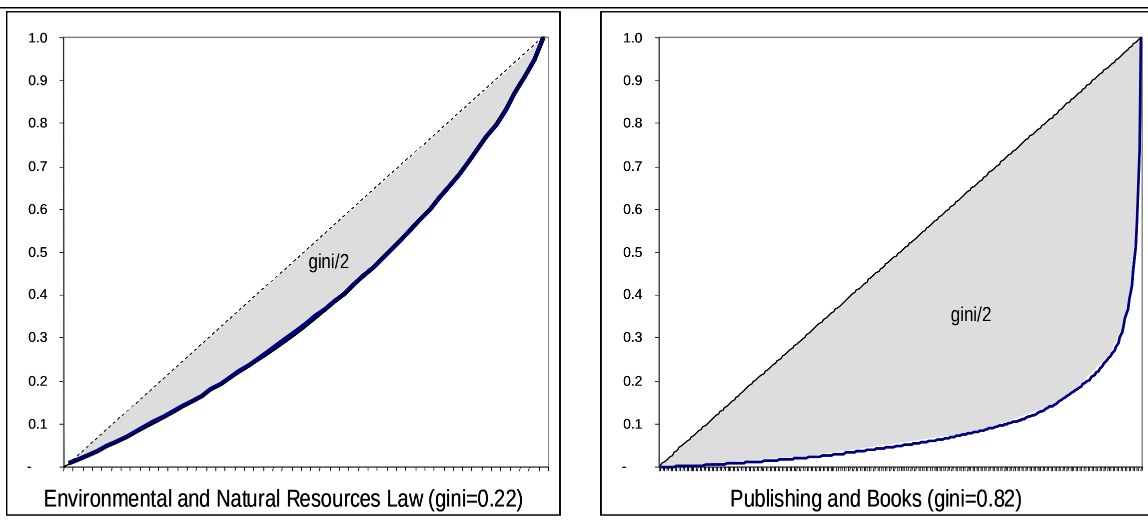

Figure 3: Illustrates the Gini coefficients for two categories.

Measuring network influence

[5] 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).

I have developed two different measures of network influence:

ImmediateInfluence: 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 demand, is divided equally and flows to those products it has direct co-purchase links to. It captures the influence of a product's immediate neighbors. I construct the *ImmediateInfluence* variable in the following way:

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

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

WeightedPageRank is based on Google's PageRank algorithm (Page and Brin, 1998), and iteratively computes the influence of the entire network on each product over time, ignoring variations in intrinsic traffic across pages. It operates on an "average graph", constructed as a weighted composite of a time series of networks.

$$WeightedPageRank(i) = (1-\alpha)+\alpha \sum_{j \in G(i)} Weight(j,i)\left(\frac{WeightedPageRank(j)}{OutDegree(j)}\right)$$

This is based on the assumption that a random surfer follows any of the links on a page with equal probability or jumps to a random page with probability [1 – α] (the "dumping factor"). PageRank algorithm divides a page's PageRank evenly among its successors in the network. The ranking of a page ends up being the long run steady-state probability that a random surfer, will visit the specific page. A page can gain a high ranking by either having many pages pointing to it or having few highly ranked pages pointing to it.

How network structure influences individual demand

To test the hypothesis that local co-purchase network links explain demand variation, I estimate the following model:

$$\boldsymbol{Log[Demand] = a + \beta_1 Log[ImmediateInfluence]}$$

I control for the unobserved heterogeneity by using the fixed effects (within) transformation by book. I estimate this equation using data collected during two distinct two-week periods, spaced six months apart: August 10th through August 23rd 2005, and February 10th through February 23rd, 2006. The results of this estimation are available on request, and establish that Immediate Influence explains a significant amount of variation in demand.

Examining the distribution of demand and influence (Figure 4): both have the same mean (influence is demand being redistributed), but the range and standard deviation of influence are larger than that of demand. This leads to suspect that the network redistributes demand in a more "equal' manner, a hypothesis I test in what follows.

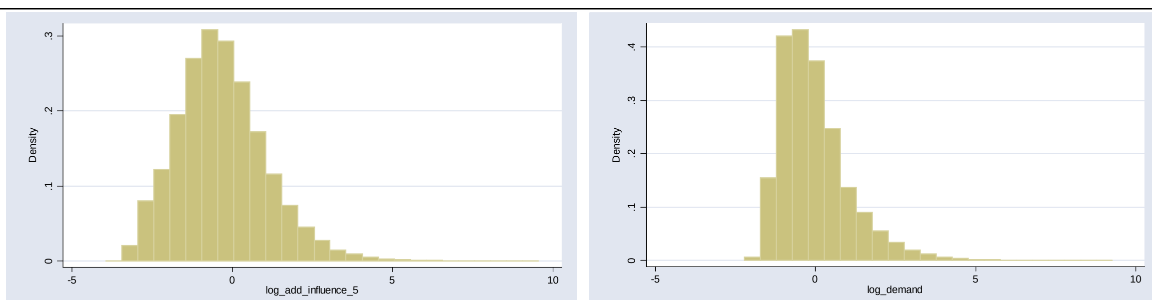

Figure 4: Distribution of log(ImmediateInfluence) and log(Demand)

My eventual dissertation will

- Ground my research in one or more structural models of consumer behavior.

- Identify product characteristics as well as link characteristics that are associated with variation in network demand across products.
- Study network properties (such as centrality) that can explain demand variation.

How network structure influences the distribution of demand

I next group the books according to category affiliation and test the hypothesis that a higher average *Weighted PageRank* for a category will be associated with a lower Gini coefficient for that category, by estimating the following equation:

$$\begin{aligned} Log[Gini] = a + b_1 Log[NumberOfBooks] + b_2 Log[AverageDemand] + b_3 Log[AveragePageRank] \\ + b_4 Log[VariancePageRank] + b_5 Log[VariancePrice] + b_6 Log[Average\ Price] \end{aligned}$$

Based on a comparative analysis across 182 categories of products, I conclude that an increase in the extent to which the network structure is influential leads to a ***flattening of demand***. I find that the average *Weighted PageRank* of the books in the category is negatively associated with the Gini coefficient of the category. This provides the first evidence that ecommerce network structures can explain outcomes such as the widely documented "long tail" of ecommerce demand (Anderson, 2004) that are uniquely characteristic of conducting business via electronic commerce. The results of the estimations are presented in table 2. Further discussion of those results is available on request.

Dependent Variable: Log[Gini]

Variable	Coeffient	Estimated Values (Standard Error)	
		Aug-05	Feb-06
Constant	*a*	-0.85 (0.2)***	-1.89 (0.57)**
Log[NumberOfBooks]	*b1*	0.011 (0.002)***	0.036 (0.007)***
Log[AverageDemand]	*b2*	0.11 (0.004)***	0.28 (0.01)***
Log[AveragePageRank]	*b3*	-0.078 (0.02)***	-0.16 (0.05)**
Log[VariancePageRank]	*b4*	0.02 (0.004)***	0.04 (0.01)***
Log[VariancePrice]	*b5*	-0.002 (0.002)	-0.007 (0.01)
Log[AveragePrice]	*b6*	-0.004 (0.007)	-0.003 (0.02)
	R^2	87%	83%

Table 2: How network structure affects the distribution of ecommerce demand

In my ongoing work, I intend to

- Repeat this estimation using the new measures of influence that I discuss in prior sections.
- Use a longer period of time, towards strengthening the results, as well as studying the evolution of the influence of network structure over time.

ESSAY 2: USING NETWORK STRUCTURES FOR BETTER DEMAND PREDICTION

In this essay, I propose to study whether economic network structures – networked data in which the links of the network are created by actual realizations of economic outcomes – can improve predictions of outcomes in ecommerce. I conjecture that such network structures inherently contain more information, — they isolate the relevant information about preferences of consumers without necessitating the explicit modeling of these preferences.

It is not realistic to build complete models of choices made by economic agents. Consider a simple example from electronic commerce. Many online retailers have tens of millions of consumers. Each of these consumers has unique preferences and willingness-to-pay for the (possibly) millions of products sold. In order to use data mining techniques for predicting optimal marketing choices, or for demand forecasting and planning, it is customary to create a coarse partition of these consumers, along a small set of readily observable dimensions, such as gender, zip code, and age, with the reasoning that consumers who share common characteristics along these dimensions are likely to make similar choices.

Now, consider an analogy. Equilibrium prices are "determined" for variety of products every day, and equilibrium demand levels are realized at posted prices, at a variety of retail stores. It is not possible to "reverse engineer" the actual preferences of decision makers or characteristics of products from these observed prices or demand levels, but such prices do contain aggregated summaries of these preferences. If one can relate these products, or these agents to one another based on similarities in such economic outcomes, the observed outcomes today for one set of products or agents can predict future outcomes for related sets of products or agents. A good example of the use of this notion is in collaborative filtering, which, in its simplest form, recommends products to a consumer based on the recent purchases of "similar" customers.

The kind of network structures I am interested in are outcome based – a link is created or altered between two economic objects based on them sharing an economic outcome. I operationalize the theory in the context of these links being created by a high fraction of copurchases. The interconnection between economic objects (agents, products) is not on account of their explicitly sharing one or more observable features or characteristics. Such features are part of what we term a products' "intrinsic features".

To provide a proof-of-concept, I set the dependent variable of the prediction task to be binary: will the average demand for the book go up in the forthcoming period (compared to the current period)?

Thus, one needs to:

(a) Determine a sufficiently comprehensive set of non-network features that can be used to benchmark predicting demand in the absence of the network.

Determine the set of "network features". I categorize network features as being of two kinds: **direct** (features that represent links corresponding to affiliation, like books of the same category, or by the same author) and **economic** (features based on links between products created by actual realizations of economic outcomes. In my context, this will refer to books that are immediate neighbors in the co-purchase network). Two types of such information will be used:

- Average information: such as number of incoming links; average price of the linked product;
- Non-network information about the most influential incoming links.

(b) Compare the performance of the following: (i) prediction using just non-network features. (ii) prediction using just network features and (iii) prediction using both network and non-network features

I have conducted pilot studies using a subset of non-network and network features, which provide preliminary evidence that economic network features improve demand predictions. Further details are available on request.

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 Information 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, Motwami, 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*, Vol. 49, No. 11.
6. Goolsbee, A. and Chevalier, J. A. (2003) Measuring Prices and Price Competition Online: Amazon and Barnes and Noble. *Qualitative Marketing and Economics* (1) P. 203-222.
7. Kane, J. and Alavi., M. (2006) Casting the Net: Towards a Theory of Multimodal Knowledge Networks. Mimeo, Emory University.
8. Langville, A. and Meyer, C. (2005) Deeper Inside PageRank. *Internet Mathematics*, 1(3):335--380.
9. Sundararajan, A. (2005) Local Network Effects and Complex Network Structure. Mimeo, NYU.