



IS OPRAH CONTAGIOUS? THE DEPTH OF DIFFUSION OF DEMAND SHOCKS IN A PRODUCT NETWORK

Eyal Carmi

Google, 76 Ninth Avenue, New York, NY 10011 U.S.A. {eyal.carmi@gmail.com}

Gal Oestreicher-Singer and Uriel Stettner

Coller School of Management, Tel Aviv University, Tel Aviv 69978 ISRAEL {galos@tau.ac.il} {urielste@tau.ac.il}

Arun Sundararajan

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

Appendix A

Descriptive Statistics

Table A1 depicts comparison statistics on the matched sample versus the treatment groups for each distance. Similarly, Table A2 depicts comparison statistics on the network-based group used as a control versus the treatment groups. The variable *Total25* is used to control for Amazon's \$25 shipping policy, capturing whether the sum of the price of the reviewed book and any of the books in the local network passes the \$25 free shipping threshold; *indegree* captures the number of books directed to a focal book in the product network; *local clustering* measures the degree to which books in the product network tend to cluster together to create groups characterized by a their density of ties; *same binding* indicates whether a purchased book is of the same binding type (e.g., hardcover, paperback) as the reviewed book; *sale price* controls for the sales price of a purchased book; *average rating* captures the subjective evaluation of a book as reported by consumers; and *sales rank* measuring a book's demand relative to other products.

Table A1. Summary Statistics for the Group Used as a Control Based on a Matched Sample									
		Cor	ntrol			Trea	ated		
	Distance 1	Distance 2	Distance 3	Distance 4	Distance 1	Distance 2	Distance 3	Distance 4	
Total 25	0.827	0.849	0.861	0.884	0.819	0.836	0.852	0.856	
TOTAI25	(0.378)	(0.358)	(0.346)	(0.320)	(0.385)	(0.371)	(0.355)	(0.351)	
Indegree	7.599	7.214	6.855	7.229	7.481	6.950	6.792	6.900	
Indegree	(15.758)	(15.572)	(15.406)	(17.184)	(14.944)	(14.794)	(15.291)	(16.383)	
	0.450	0.388	0.374	0.365	0.443	0.381	0.368	0.364	
Local clustering	(0.167)	(0.132)	(0.128)	(0.124)	(0.168)	(0.132)	(0.130)	(0.129)	
Samo catogory	0.582	0.467	0.458	0.408	0.764	0.605	0.525	0.431	
Same calegory	(0.493)	(0.499)	(0.498)	(0.491)	(0.425)	(0.489)	(0.499)	(0.495)	
Samo binding	0.401	0.379	0.355	0.333	0.684	0.626	0.556	0.490	
Same binding	(0.490)	(0.485)	(0.479)	(0.471)	(0.465)	(0.484)	(0.497)	(0.500)	
Solo prico	2069.605	1756.587	1847.640	1946.760	1571.420	1568.508	1587.784	1624.669	
Sale price	(2515.388)	(1691.524)	(1808.255)	(1918.071)	(613.206)	(625.593)	(735.234)	(936.690)	
	4.235	4.243	4.226	4.240	4.195	4.171	4.162	4.167	
Average rating	(0.579)	(0.561)	(0.590)	(0.625)	(0.534)	(0.540)	(0.566)	(0.558)	
Salas Bank	116417.400	129004.200	158465.200	181031.200	76259.260	86075.950	96597.280	113820.500	
Sales Rank	(185952.000)	(196460.1)	(215149.400)	(223633.400)	(143612.800)	(153471.900)	(159435.100)	(173966.000)	

*Standard errors in parentheses

Table A2. Summary Statistics for the Group Used as a Control Based on Network									
		Con	trol			Trea	ated		
	distance 1	distance 2	distance 3	distance 4	distance 1	distance 2	distance 3	distance 4	
Totol 25	0.836	0.852	0.856	0.865545	0.819	0.836	0.852	0.856	
TOTAIZS	(0.371)	(0.355)	(0.351)	(0.341145)	(0.385)	(0.371)	(0.355)	(0.351)	
Indograa	6.950	6.792	6.900	6.697625	7.481	6.950	6.792	6.900	
indegree	(14.794)	(15.291)	(16.383)	(15.75062)	(14.944)	(14.794)	(15.291)	(16.383)	
Local eluctoring	0.381	0.368	0.364	0.35946	0.443	0.381	0.368	0.364	
Local clustering	(0.132)	(0.130)	(0.129)	(0.12478)	(0.168)	(0.132)	(0.130)	(0.129)	
Somo ostorony	0.605	0.525	0.431	0.382425	0.764	0.605	0.525	0.431	
Same category	(0.489)	(0.499)	(0.495)	(0.485986)	(0.425)	(0.489)	(0.499)	(0.495)	
Some binding	0.626	0.556	0.490	0.427813	0.684	0.626	0.556	0.490	
Same binding	(0.484)	(0.497)	(0.500)	(0.494768)	(0.465)	(0.484)	(0.497)	(0.500)	
Colo prizo	1568.508	1587.784	1624.669	1686.151	1571.420	1568.508	1587.784	1624.669	
Sale price	(625.593)	(735.234)	(936.690)	(1203.321)	(613.206)	(625.593)	(735.234)	(936.690)	
Average rating	4.171	4.162	4.167	4.180571	4.195	4.171	4.162	4.167	
Average rating	(0.540)	(0.566)	(0.558)	(0.582775)	(0.534)	(0.540)	(0.566)	(0.558)	
Soloo Book	86075.950	96597.280	113820.500	134070.7	76259.260	86075.950	96597.280	113820.500	
Jaies Ralik	(153471.900)	(159435.100)	(173966.000)	(205551.2)	(143612.800)	(153471.900)	(159435.100)	(173966.000)	

*Standard errors in parentheses

Appendix B

Algorithm for Data Collection from Amazon.com

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

Graph Collection: The program that 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, one popular book was chosen as a seed) 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 last-in-first-out (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 for which exploration was not exhausted. The program terminates when the entire connected component of the graph is collected.

For example, in the graph in Figure B1, the nodes are numbered in the order in which the crawler traverses the graph. In this case, collection starts at node 1. Its co-purchase links are nodes 2, 6, and 7. Therefore, these numbers are added to a LIFO stack. The script will then proceed to node 2, whose co-purchases are nodes 3, 4, and 5, and thus, those numbers will be added to the LIFO stack, which will now include 3, 4, 5, 6, and 7. The script will continue to node 3. Since there are no co-purchase links to that node, it will move on to node 4. In the same way, the script will collect data on node 5, node 6 and node 7.

Since node 7 has co-purchase links to nodes 8 and 9 they will be added to the stack. After visiting nodes 8, 9, and 10, data collection will terminate. As can be seen, the script stops only after information about the entire connected component has been collected.

The collection of the entire connected component on Amazon.com takes between 4 and 5 hours. The script is run each day at midnight.

Sales Rank Collection: A second program collects the demand information for all books on the graph at 3-hour intervals for the 24-hour period following the collection of the graph. This script collects the Sales Ranks of all the books that ever appeared in the graph. Therefore, it also tracks the sales of books that are no longer in the graph.



Appendix C

Network Statistics

Co-Purchase Networks

Table C1 presents basic network statistics on each of the daily co-purchase graphs that were collected in the period of 2006–2008. Each daily product network consists of a daily average of 270K books and over 1.2M edges. The average density is very low (\sim 1.45*10⁻⁵) due to the truncation to 5 outgoing links per node¹; however, the fraction of reciprocal links in the network is very high (55% on average) and the average clustering coefficient is 0.39. These data are reasonable since the network represents co-purchased products.

The global structure of the network is relatively stable over time; we observe a relatively low standard deviation in network properties such as the average clustering coefficient, the average indegree and the fraction of reciprocal links. The degree distribution is stable across days and exhibits a power law shape (see Figure C1 for degree distribution and distribution of betweenness centrality on a sample daily network).

Table C1. Amazon Co-purchase Networks Statistics								
Variable	# Nodes	# Edges	Average In Degree	Fraction of Reciprocal Links	Average Clustering Coefficient			
Mean	274,179	1,246,986	4.7	55%	0.39			
Median	273,255	1,230,800	4.7	56%	0.39			
Maximum	368,760	1,657,400	4.8	56%	0.40			
Minimum	120,620	362,580	3.5	43%	0.27			
Std. Dev.	40,547	182,999	0.1	2%	0.01			
Skewness	-0.37	-0.71	-5.3	-4.56	-6.46			
Kurtosis	2.58	4.43	42.4	26.95	55.09			
Jarque-Bera	9.80	55.61	22,822	8976	39355			
Probability	0.01	0.00	0.0	0.00	0.00			
Observations	328	328	328	328	328			

$$\frac{5n}{n(n-1)} = \frac{5}{n-1} \cong 1.8 \times 10^5$$

¹Since each node has up to five outgoing edges, the maximal theoretic network density (a proxy for the average level of activity in the network) is



Event Networks

Each review event was cross-referenced with the corresponding network and sales data from Amazon.com and went through a series of manual and automatic cleaning procedures. Details on these procedures are available upon request.

These cleaning procedures resulted in a sample of 123 review events; for each event we extracted a subnetwork from the co-purchase graph starting from the reviewed book and up to a distance of 5 links away (the fifth network neighbor of the reviewed book). Following Deschatres and Sornette (2005), we manually classified the review events into two categories: (1) exogenous shocks and (2) endogenous and multiple shocks (see Figure C2). All econometric models were applied to the final sample of 83 exogenous shocks (40 from the *Oprah Winfrey Show* and 43 from *The New York Times*) and to a total of 19,669 books in their subnetworks.

Table C2 presents basic network statistics on the subnetworks up to a distance of five links away (the fifth network neighbor of the reviewed book). The relatively high variance in the average clustering coefficient of these networks (as illustrated in Figure C3) shows that they are significantly different from each other, which may be reflected in the way exogenous shocks diffuse through the network.



Figure C2. Reviewed Books Time Series Data, Classified into Two Categories: Exogenous Shocks (top) and Endogenous and Multiple Shocks (bottom)

Table C2. Network Statistics Across the Su	bnetworks up to the Fiftl	h Network Neighbor fo	or Each of the
Reviewed Books' Events			

Amazon Co-purchase Networks Statistics										
Variable	# Nodes	# Edges	Average	Fraction of Reciprocal Links	Average Clustering					
Valiable	# NOUES	# Luges	III Degree	Recipiocal Links	Coefficient					
Mean	249	558	3.6	48%	0.33					
Median	231	534	3.6	47%	0.31					
Maximum	813	1524	5.0	80%	0.84					
Minimum	8	40	3.0	39%	0.17					
Std. Dev.	159	313	0.4	6%	0.10					
Skewness	0.72	0.46	1.1	1.62	1.98					
Kurtosis	3.33	2.73	4.8	7.77	9.85					
Jarque-Bera	11.22	4.74	39.7	170.23	320.89					
Probability	0.00	0.09	0.0	0.00	0.00					
Observations	123	123	123	123	123					



Appendix D

Summary Statistics

Summary statistics for a selection of shock constructed variables are given in Table D. We also see that on average, only 19% of the neighbors up to a distance of four clicks belong to the same category as the reviewed book, and only 2% were written by the same author.

To measure category mixing we utilize Amazon's multi-level category tree (see Table D2 for an example and Table D3 for summary statistics).

Further exploration of the distribution of persistence across different groups of neighbors based on minimal distance from the reviewed book (see Figure D1) shows a considerable amount of variation across books.

Table C1. Summary Statistics for a Selection of Constructed Variables								
Variable	Average Sales Rank	Persistence (Sales Rank)	SRS					
Mean	126,759	1.48	2.59					
Median	46,569	0.00	1.43					
Мах	4,340,296	64.00	477.62					
Min	10	0.00	0.08					
Std. Dev.	194,163	4.49	22.17					
Skewness	4	8.14	66.13					
Kurtosis	33	92.05	4,124.00					
Obs	19,669	19,669	19,669					



Figure D1. The distribution of persistence, the number of post-event days in which demand remained one standard deviation above the pre-event average demand for the reviewed books and first, second and third network neighbors. Graphs are based on the sub-networks of books reviewed by *Oprah* and the *New York Times* in 2007.

Defining category similarity is not a trivial task, since books belong to multiple categories at different levels of hierarchy. In the analysis that follows, two books are said to have the same category if they share at least one second-level category path. This definition is relatively liberal and will result in a high fraction of books sharing the same category. We also experimented with several alternative definitions: two books share at least one second-level categories; and (3) only the three top categories.

Table D2. Example of Amazon's Multilevel Category Tree, Showing a Subset from the Two Top-Level Categories

Level 1 Category	Level 2 Category
Children's Books	People & Places
Children's Books	Science, Nature & How It Works
Children's Books	Animals
Children's Books	Educational
Children's Books	Holidays & Festivals
Literature & Fiction	History & Criticism
Literature & Fiction	Poetry
Literature & Fiction	Comic
Literature & Fiction	Drama
Nonfiction	Education
Nonfiction	Social Sciences
Nonfiction	Politics

Table D3. Number of Books with at Least (K) Second-Level Categories								
Number of Categories (K)	Number of Books with at Least K Categories	Number of Categories (K)	Number of Books with at Least K Categories					
1	706,169	11	4,521					
2	637,558	12	1,927					
3	542,354	13	823					
4	403,499	14	327					
5	267,152	15	131					
6	158,153	16	50					
7	86,269	17	21					
8	44,558	18	7					
9	21,603	19	4					
10	10,064	20	1					

Summary statistics for a selection of network/mixing constructed variables are given in Table D4. We also see that, consistently with the findings of Oestreicher-Singer and Sundararajan (2008), on average about 44% of the neighbors up to a distance of five clicks from the reviewed book belong to the same category as the reviewed book, and only 1% were written by the same author.

Table D4. Summary Statistics for a Selection of Constructed Variables									
Variable	Network Proximity	CC_i	Same Author	Same Category	Same Price				
Mean	0.018	0.54	0.01	0.44	0.84				
Median	0.001	0.53	0.00	0.00	1.00				
Мах	1.00	1.00	1.00	1.00	1.00				
Min	0	0.023	0	0	0				
Std. Dev.	0.08	0.17	0.12	0.5	0.37				
Skewness	9.04	-0.02	8.46	0.24	-1.82				
Kurtosis	101.38	3.29	72.54	1.06	4.31				
Obs.	19669	19669	19669	19669	19669				

Breaking down category and author statistics (see Table D5), one can see that the percentage of books in the same category as the reviewed book drops as the distance from the reviewed book increases. An even sharper drop is seen (as expected) for books with the same author: The percentage of books with the same author among first neighbors is significantly higher.

Table C5. Category and Author Mixing Statistics by Distance from the Reviewed Book								
	Sa	ame Category	Statistics		Same Author Statistics			
Distance	All	Oprah Reviews	New York Times Reviews	All	Oprah Reviews	New York Times Reviews		
All neighbors	43.9%	44.4%	43.7%	1.3%	1.8%	1.1%		
(15)	(0.4%)	(0.6%)	(0.4%)	(0.1%)	(0.2%)	(0.1%)		
1	76.6%	80.4%	73.1%	20.7%	22.5%	19.3%		
I	(2.1%)	(2.9%)	(3.0%)	(2.0%)	(3.1%)	(2.7%)		
2	60.5%	63.6%	58.4%	4.6%	4.3%	4.8%		
2	(1.5%)	(2.3%)	(2.0%)	(0.6%)	(1.0%)	(0.9%)		
2	52.1%	54.6%	50.8%	0.9%	0.6%	1.1%		
5	(1.0%)	(1.7%)	(1.3%)	(0.2%)	(0.3%)	(0.3%)		
Λ	43.9%	42.3%	44.6%	0.2%	0.2%	0.2%		
4	(0.7%)	(1.2%)	(0.8%)	(0.1%)	(0.1%)	(0.1%)		
5	38.6%	37.0%	39.3%	0.1%	0.0%	0.1%		
0	(0.5%)	(0.9%)	(0.6%)	(0.0%)	(0.0%)	(0.0%)		

*Standard errors between parentheses.

Appendix E

Sales Rank Conversion to Demand I

To estimate the actual level of demand $Demand_{it}$ of a book *i* at time *t* on the basis of the book's *SalesRank* (*SR*_{it}), the following log-linear conversion model was suggested (Brynjolfsson et al. 2003; Goolsbee and Chevalier 2003):

 $Log[Demand_{it}] = a + bLog[SalesRank_{it}]$

This equation to convert Sales Rank data into demand estimations was first introduced by Goolsbee and Chevalier 2003. Their approach was based on making an assumption about the probability distribution of book sales, and then fitting some demand data to this distribution. They chose the standard distributional assumption for this type of rank data, which is the Pareto distribution (i.e., power law).

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 equation. They estimated the following parameters based on book sales data from 2000: a = 10.526, b = -0.871.

This conversion model has been used in many studies (see for example, Oestreicher-Singer and Sundararajan 2008; Sornette et al. 2004). However, estimating the actual level of demand is still not a trivial process, since demand patterns in electronic commerce tend to change over time, and the model may need to be updated. Brynjolfsson et al. (2009) recently carried out the estimation a second time, using the above log-linear model, and they found that the "long tail" of Internet book sales has gotten longer over the years. They estimated the coefficients based on book sales data from 2008 as: a = 8.046, b = -0.613.

The authors also suggested a new methodology to better fit the relationship between Sales Rank and sales: using a series of splines, each modeled as a negative binomial regression model (rather than a linear regression). Figure E1 shows the difference between the two estimations, computed over the average Sales Rank of each of the books in our final sample. We can see that our sample spans across a wide range of Sales Rank values and that the two curves cross each other when the Sales Rank equals 14,949.

There are several other known issues regarding the use of converted demand estimations, especially for best-selling books (see the discussion in Chellappa and Chen 2008; Rosenthal 2010; Sornette et al. 2004). These pose a more severe problem in our context, as several of the reviewed books attained best-seller status. We therefore directly use SalesRankRatios to compute the different variables.

Summary statistics for some of the constructed variables are given in Table E1 together with their demand-based counterparts (that is, demand estimated using the suggested estimates from Brynjolfsson et al. (2003) and the suggested estimates from Brynjolfsson et al. (2009)). We can see that the changes in estimation of the demand and Sales Rank actually translate to small changes in the computed persistence. This can also be seen when plotting the distribution of persistence based on each of the three estimation methods (see Figure E2).



Figure E1. Sales Rank Conversion to Demand Using 2008 Estimation Versus 2000 Estimations (The graphs present the conversion of the average Sales Rank of the books in our final sample to demand using the two estimations. The same data are presented in (a) normal scale (zoomed in to the range of $0 \dots 5,000$) and (b) logarithmic scale.)

Table E1. Summary Statistics for a Selection of Constructed Variables									
Variable	Mean	Median	Max	Min	Std. Dev.	Skewness	Kurtosis	Obs	
Average Sales Rank	126,759	46,569	4,340,296	10	194,163	3.67	32.83	19669	
Average Demand (2003)	116.33	4.34	27404.55	0.06	572.38	18.66	669.32	19669	
Average Demand (2009)	30.79	5.17	2360.51	0.27	86.83	7.12	95.34	19669	
Persistence (Sales Rank)	1.476	0.000	64.000	0.000	4.486	8.14	92.05	19669	
Persistence (Demand 2003)	1.332	0.000	64.000	0.000	4.045	8.84	111.57	19669	
Persistence (Demand 2009)	1.365	0.000	64.000	0.000	4.093	8.68	107.93	19669	



References

- Brynjolfsson, E., Hu, Y., and Smith, M. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* (49:11), pp. 1580-1596.
- Brynjolfsson, E., Hu, Y., and Smith, M. 2009. "A Longer Tail? Estimating the Shape of Amazon's Sales Distribution Curve in 2008," in *Proceedings of the Workshop on Information Systems and Economics*, Phoenix, AZ, December 14-16.
- Chellappa, R. K., and Chen, C. 2008. "On the Temporal Nature of Sales-Rank Relationships of Albums and Digital Tracks in the Music Industry: The Relevance of Billboard Charts Post-Digitization," in *Proceedings of the INFORMS Annual Meeting*, Washington, DC.
- Deschatres, F., and Sornette, D. 2005. "Dynamics of Book Sales: Endogenous Versus Exogenous Shocks in Complex Networks," *Physical Review E* 72, 016112.
- Goolsbee, A., and Chevalier, J. 2003. "Measuring Prices and Price Competition Online: Amazon and Barnes and Noble," *Quantitative Marketing and Economics* (1), pp. 203-222.
- Oestreicher-Singer, G., and Sundararajan, A. 2008. "The Visible Hand of Social Networks in Electronic Markets," Working Paper, New York University.
- Rosenthal, M. 2010. "Amazon Sales Rank for Books," http://www.fonerbooks.com/surfing.htm.
- Sornette, D., Deschatres, F., Gilbert, T., and Ageon, Y. 2004. "Endogenous Versus Exogenous Shocks in Complex Networks: An Empirical Test Using Book Sale Rankings. *Physical Review Letters* 93.228701.
- Watts, D. J. 2003. *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton, NJ: Princeton University Press. Watts, D. J., and Strogatz, S. H. 1998. "Collective Dynamics of 'Small-World' Networks," *Nature* (393), pp. 440-442.