NET Institute*

www.NETinst.org

Working Paper #10-18

September 2010

Is Oprah Contagious? Identifying Demand Spillovers in Product Networks

Eyal Carmi
Tel-Aviv University

Gal Oestreicher-Singer
Tel-Aviv University

Arun Sundararajan
Stern School of Business, NYU

Is Oprah Contagious? Identifying Demand Spillovers in Product Networks

Eyal Carmi *(Tel-Aviv University),*

Gal Oestreicher-Singer *(Tel-Aviv University)* and

Arun Sundararajan *(NYU Stern School of Business)*

Abstract

We study the online contagion of exogenous demand shocks generated by book reviews featured on the Oprah Winfrey TV show and published in the New York Times, through the co-purchase recommendation network on Amazon.com. These exogenous events may ripple through and affect the demand for a "network" of related books that were not explicitly mentioned in a review but were located "close" to reviewed books in this network. Using a difference-in-differences matched-sample approach, we identify the extent of the variations caused by the visibility of the online network and distinguish this effect from variation caused by hidden product complementarities. Our results show that the demand shock diffuses to books that are upto five links away from the reviewed book, and that this diffused shock persists for a substantial number of days, although the depth and the magnitude of diffusion varies widely across books at the same network distance from the focal product. We then analyze how product characteristics, assortative mixing and local network structure, play a role in explaining this variation in the depth and persistence of the contagion. Specifically, more clustered local networks "trap" the diffused demand shocks and cause it to be more intense and of a greater duration but restrict the distance of its spread, while less clustered networks lead to wider contagion of a lower magnitude and duration. Our results provide new evidence of the interplay between a firm's online and offline media strategies and we contribute methods for modeling and analyzing contagion in networks.

1. Introduction and Research Questions

Online commercial interactions have increased dramatically over the last decade. An important by-product of this process is the emergence of visible *product networks*. For example, most electronic commerce sites are organized as a collection of webpages, each featuring a single product (e.g. a book, video, or other content item). These product pages are linked by hyperlinks to other product pages, thus creating a network where the products are the nodes. Perhaps the oldest example of a visible electronic product network is the "co-purchase" network of Amazon.com.[1]

The presence of the hyperlinked network structure is one fundamental way in which electronic commerce differs from traditional commerce. One can imagine the process of browsing an electronic store as being analogous to walking the aisles of a physical store, where the product network of interconnected webpages forms the electronic "aisle structure", and the position of a product in the network is its virtual "shelf placement." Thus, it is natural to assume that in contrast with what models of costless electronic search might suggest (Bakos 1997), the set of products to which cognitively bounded consumers actually pay attention is altered by the hyperlinks between these pages.

In this paper we study the online contagion of exogenous demand shocks created by media events on such product networks. Specifically, we focus on reviews and their impact on demand. While previous research has focus on the effect of such events on a single product, our goal is to show that the visibility of product networks affects consumers' demand patterns by causing exogenous demand shocks (resulting from marketing campaigns) to spill over to other products (we refer to this as a "ripple effect"). One of the challenges in studying this ripple effect is separating the effect of contagion through the product network from other effects, that is, disentangling this influence from correlation due to hidden product similarities (affinity). Using data collected from a large-scale real-world product network, we are able to measure and describe the structure of the network and gain important insights regarding the

[1] Amazon.com provides hyperlinks that connect products, under the heading "Consumers who bought this item also bought …". While Amazon was one of the first to introduce a recommendation network, today almost every major e-commerce website (Barnes & Noble, YouTube, Yelp, iTunes, etc.) implements a recommendation system that can be modeled as a product network.

connection between local network structure and the patterns of contagion across the network.

Our study is based on data derived from the co-purchase network on the Amazon.com website as well as exogenous shocks created by book reviews featured on the *Oprah Winfrey Show* and in the *New York Times* [2]. Product reviews that appear on television or in newspapers are known to have a high impact on the sales of the reviewed products. Specifically, prior research shows that a review on the *Oprah Winfrey Show* can transform a reviewed book into a bestseller literally overnight (Balogh 2008; Illouz 2003; Rooney 2005). Similarly, Deschatres and Sornette (2005) and Sorensen and Rasmussen (2004) show that book reviews published in the *New York Times* newspaper also significantly increase the sales of reviewed books.

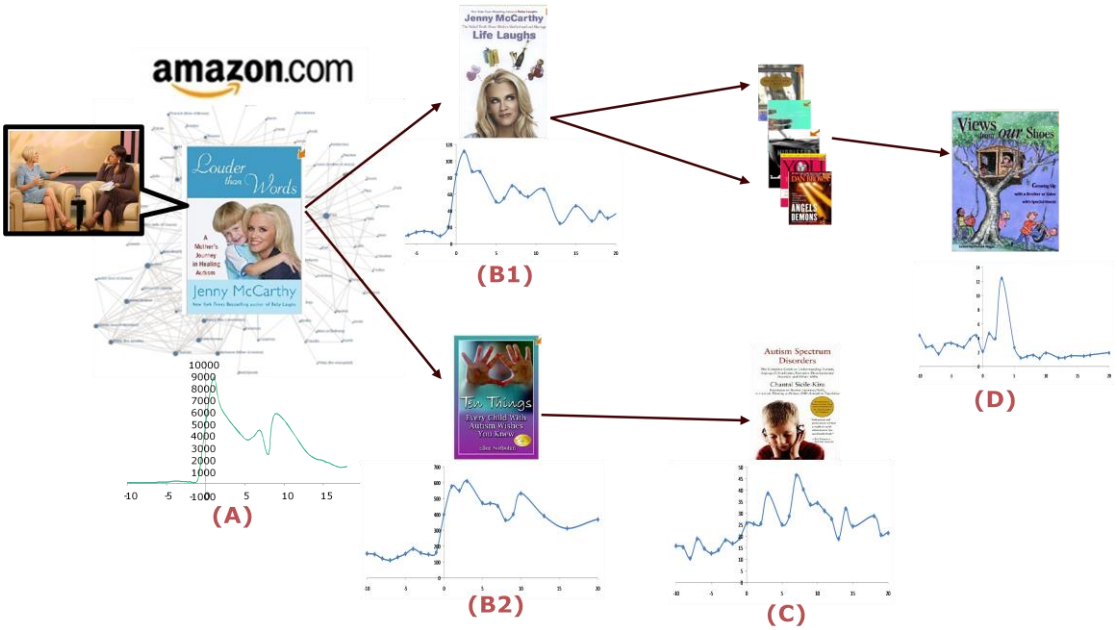

Figure 1-1: Example of the ripple of exogenous demand shocks across product networks. The book "Louder than Words" by Jenny McCarthy (A) was featured on the Oprah Winfrey Show in September of 2007 and immediately experienced an increase in demand of close to 9000%. We also witness an increase in demand in books that are one click away from the reviewed book (B1 & B2); two clicks away from the reviewed book (C); and even books that are four clicks away from the reviewed book (D). Note that in each of these graphs, the x-axis is time in days, where 0 represents the day of the review; the y-axis represents a measure of demand.

[2] We use book reviews as an example for exogenous demand shocks; other examples of exogenous shocks to a product network include attention drawn to a blog in a network of blogs (similar to the blog network described by Mayzlin and Yoganarasimhan (2008)) due to a scandal following a blogger's political activity. It is natural to assume that the increased attention to one blog will spill over and create shocks to the demand for neighboring blogs. Similarly, in the citations network, academic community attention captured by an award-winning paper may spill over to other papers it cites or in which it is cited.

It is not surprising that such reviews result in a strong exogenous shock to the demand of the reviewed book. The impact of this shock,, however, limited to the reviewed product. The featured review creates interest in the product, which spills over to other products as well (see Figure 1-1 for an example).

Specifically, our paper addresses the following three research questions: First, do exogenous demand shocks diffuse across product networks? Second, what is the magnitude, the depth and the persistence of this ripple effect across products in co-purchase recommendation networks? And third, which network structure characteristics influence the rate and persistence of the ripple?

We first provide statistical evidence of the existence of cross-product spillover of demand shocks in product networks. The main challenge we address in the paper is the identification of the cross-product spillover effect, separating it from hidden product complementarities (what social network theorists might call "homophily"). We used a quasi-natural experiment of exogenous demand shocks created by reviews to focal products in the network to study the influence of the network on other products that were not mentioned in the reviews. Our identification strategy is based on a difference-in-differences extension of propensity-score-based matching.

Our empirical estimations show a significant influence of the visible network on neighbors up to three links away from the reviewed book. Though the effect of the network beyond third-degree neighbors is not significant on average, it can be seen as far as the fourth and fifth neighbors, and it decays both with distance from the source of the shock and with time. These results provide compelling evidence that exogenous demand shocks cause statistically and economically significant changes to the demand for neighboring books, and that these changes travel quite deep in the network.

We next analyze the variance in the resistance of network neighbors to an exogenous shock; we find evidence of a strong influence of both assortative mixing and local network structure. Cross-product similarities such as sharing an author with a reviewed book or having the same binding type highly influence the probability of being affected by the shock. Both network proximity and local clustering around a book were also found to play an important role in increasing neighbors' probability of being affected by the shock (even when controlling for the global network structure). These results suggest that cross-product spillover processes across product networks

are consistent with the idea of "complex contagion" (Centola and Macy 2007) and are highly moderated by assortative mixing.

The third part of our research provides evidence that while these observed diffused demand shocks are at times remarkably persistent, there is considerable variation in the persistence of these aftershocks across books located at similar distances from the source of a shock. Building on duration model theory, we estimate an exponential hazard rate model that captured the influence of network structure and proximity on the persistence of these diffused aftershocks. We show that shock persistence differs fundamentally between close neighbors (one or two clicks away from the reviewed book) and distant neighbors (three clicks or more). The persistence of a shock to close neighbors is highly affected by their geodesic distance from the reviewed book (due to the significantly greater exposure of first neighbors), whereas ripple to distant network neighbors depends on the presence of multiple paths linking to them from the source of the shock (which are necessary to direct sufficient consumer attention).

While the local clustering around a neighbor (consistent with the previous analysis) positively increases the persistence of a shock, we find that local clustering around the source of the shock creates *a* "*fishing net*" effect, trapping consumer attention in the network neighborhood close tothe reviewed book. This structure increases the persistence of the shock among close neighbors and decreases the persistence of shocks to distant neighbors.

This paper is organized as follows: section 2 reviews the related literature; section 3 describes the data used for the empirical part of the paper and the operationalization of variables; the identification issues are discussed and analyzed in section 4; section 5 explores the product-level resistance to shocks; and in section 6 we conclude and provide avenues for future research.

2. Related Work

Our paper contributes to three major streams of research: product networks, exogenous shocks in networks, and reviews' impact on demand.

Most importantly, our work advances the understanding of product networks, a new and relatively unstudied field. Recently, social networks have received much attention from researchers in a variety of fields, such as business, economics,

epidemiology and computer science[3]. In view of the extensive study of social networks, the limited attention given to research of product networks is perhaps surprising. Work on product networks includes a study of the network of videos on YouTube by Oh et al. (2008), a study of the network of blogs by Mayzlin and Yoganarasimhan (2008)[4], and a study of the network of news reports by Dellarocas et al. (2009) studied the strategic interaction between content sites, which can also be thought of as a product network. However, product networks were not explicitly mentioned in those studies. Goldenberg et al. (2010) studied the interaction between product networks and social networks in the context of YouTube. Oestreicher-Singer and Sundararajan (2008) studied the network of books on Amazon.com and quantified the incremental correlation in book sales attributable to the product networks' visibility. Our work contributes to this stream of research by analyzing the ripple process across products following exogenous shocks.

Somewhat related to this topic is the literature on multi-product ripple in marketing (e,g., Chintagunta and Haldar 1998; Libai et al. 2008; Niraj et al. 2008). These studies measure correlations in sales among products or product categories; however, focus has traditionally been on a small set of similar products. For example, Niraj et al. (2008) studied the cross-category spillover between two product categories (bacon and eggs) and estimated the cross-category profit impact of promotions (also see Edwards and Allenby 2003; Manchanda et al. 1999). To the best of our knowledge, our work is novel in examining how product networks affect multi-product ripple on a large scale.

We also add to the literature on demand shifts following expert reviews or celebrity endorsement as well as marketing campaigns. Particularly, the impact of reviews on demand has been extensively studied in marketing literature in the context of traditional commerce (Boatwright et al. 2007; Reinstein and Snyder 2005) and e-commerce (Deschatres and Sornette 2005; Forman et al. 2008; Sorensen and Rasmussen 2004; Sornette et al. 2004). Specifically, Oprah Winfrey's endorsement was shown to have a powerful economic (and political) impact (Balogh 2008; Illouz

[3] A complete review of this literature is beyond the scope of this paper; for an extensive review of the study of social networks in economics the reader is referred to: Jackson (2009); Kempe (2010); Mayer (2009) and Newman et al. (2006).
[4] The network of blogs can also be thought of as a social network.

2003; Rooney 2005). Similarly, book reviews published in the *New York Times* newspaper significantly increase the sales of the reviewed books (Sorensen and Rasmussen (2004), Deschatres and Sornette (2005)). Our research focuses on the ripple process across products, rather than the diffusion of demand for a single product across a network of individual consumers. We offer a novel analysis of the connection between product network structure and demand.

Broadly, we add to the network analysis literature by providing an extensive analysis of the characteristics of product networks, a type of large, real-world network. Within the study of networks, one stream of literature that is particularly relevant to our work is the effect of exogenous shocks in networks. Exogenous shocks have been studied in biology (Kakimura et al. 2002), marketing (Groot 2006), finance (Bae et al. 2003; McDonald et al. 2008; Sornette et al. 2002) and other fields (Sornette 2002, 2006). The majority of studies, especially in the context of epidemiology (Mike J. Jeger (2007)), the spread of computer viruses (Lloyd and May 2001) and word-of-mouth and information diffusion (Aral et al. 2009; Cointet and Roth 2007; Libai et al. 2010; Watts and Dodds 2007), treat diffusion as an unbounded process (stochastic or deterministic). They focus on conditions (typically based on the base-rate of contagion or the global network characteristics) that may cause an event (disease outbreak, virus infection, technology innovation, product adoption, etc.) to spread across the network until the entire network is affected. However, there is also evidence from the literature (Fowler and Christakis 2010; Karrer and Newman 2010) that the influence of an actor in real-world networks is limited to a small area in the network. Though these studies were done in different domains, our current findings lend additional support to the latter approach.

Finally, from a methodological point of view, this research adds additional support to the body of literature that shows causality in complex networks. The general identification challenge, one that most empirical research on networks deals with, is: what is the true process that drives the results we observe and how could one separate the effect of the presence of the network from other confounding effects? The approach used in this paper adds to a recent stream of literature that tries to identify causality using "quasi-natural experiments". We demonstrate how to identify causality and estimate treatment effects in the context of large-scale quasi-natural

experiments and provide additional support for the validity and importance of this stream of research.

3. Data

The following section provides an overview of our data set and of the operationalization of variables we use. We combine data sets from three main sources: (1) information about network structure and demand for books on the Amazon.com website, (2) information about book reviews that appeared on the Oprah Winfrey show on television, and (3) information about book reviews that appeared in the "Sunday Book Review" section of the online edition of the *New York Times*. Using two different sources for exogenous demand shocks (The *Oprah Winfrey Show* and the *New York Times*) contributes to the robustness of the results of this research.

Network structure and demand data from Amazon.com

The data set we use includes daily product, pricing, demand and "network" information for over 700,000 books sold on Amazon.com. Each product on Amazon.com has an associated webpage, displaying a set of "co-purchase links," which are hyperlinks to products that were co-purchased most frequently with that product on Amazon.com. The co-purchase set for each webpage is limited to five[5] items and is listed under the heading "Customers who bought this item also bought …" (See Figure 3-1 for an illustration). Conceptually, the co-purchase network is a directed graph in which nodes correspond to products, and edges to directed co-purchase links. (A sample part of a graph is illustrated in Figure 3-1.)

Data on this graph are collected using a Java-based crawler that 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 on the title book, 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: The size of the daily collected

[5] Currently Amazon.com provides a list of more than five items in each co-purchase network. Although users are initially exposed to the top five due to screen-size limitations, users can click to view the next five products. We began collecting data before 2007, when this was not the case, and we assume that only five links are available per product page.

connected component varies and is 260,000 books on average. The algorithm used for data collection is provided in Appendix A.

Figure 3-1: Illustration of co-purchase links on a product webpage on Amazon.com (Left), and illustration of a subset of paths in the co-purchase graph for *The Da Vinci Code* (Right).

We use the following data, collected between January 2006 and June 2008 for this study, for each book: ASIN (a unique serial number given to each book by Amazon.com), List Price, Sales Price (the price on the Amazon.com website that day), Co-purchases (ASINs of the five books that appear on the co-purchases list), Sales Rank (a number associated with each product on Amazon.com, which measures its demand relative to other products), Author, Category, User Reviews and Average Star Rating.

Exogenous shocks from the *Oprah Winfrey* TV Show

We collected information about book reviews that appeared on the *Oprah Winfrey Show*. Each book review on the *Oprah Winfrey Show* has a dedicated webpage on the Oprah.com website (See Figure 3-2). We collected review-related data from January 2006 to April 2008. The data set contains 400 book reviews. For each review, the book's title, author and review date were collected using a PHP-based crawler and then manually verified.

Exogenous shocks from the *New York Times*

We collected data about book reviews that appeared in the "Sunday Book Review" section of the online edition of the *New York Times* between January 2006 and June 2008; the dataset contains over 2,000 book reviews. Every week, the

NYTimes.com publishes a section ("Sunday Book Review") containing 10–15 book reviews. Each book review on the "Sunday Book Review" has a dedicated webpage on the NYTimes.com website (See Figure 3-2). The collection method and data are similar to those described above for the Oprah Winfrey reviews.

Figure 3-2: Illustration of a sample of book reviews. The figure shows a sample of book reviews taken from the "Sunday Book Review" section of the online edition of the New York Times (right-hand image) and "Books Seen on the Show" page on Oprah.com (left-hand image). Since mid-2008, "Books Seen on the Show" data are no longer publicly available on Oprah.com.

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 (See Appendix B for details).

An important observation that guides this research is that even though the global structure of the network seems to be stable over time, the local structure of the network can vary significantly across different areas of the network (a summary of network statistics for the event sub-networks is provided in Appendix B). We therefore explore the connection between the local area of the network and the ripple of exogenous shocks across the network.

Operationalization of Variables

We developed several measures to represent the magnitude of the shock, local network structure, and link properties. An extended description of the constructed variables as well as summary statistics are provided in Appendix C.

Shock Parameters

Sales Rank (*SR*) is a number associated with each product on Amazon.com, which measures the product's demand relative to other products. The best-selling product is therefore ranked 1, followed by 2, 3, and so on.

SalesRankRatio (*SRR*) measures the magnitude of the product's Sales Rank at time *t* following an event, in comparison to the pre-event average Sales Rank of the product.

SalesRankShock (*SRS*) measures the maximal short-term change in the Sales Rank of a book following the exogenous shock, and represents the peak of the sales increase relative to the pre-event average.

Affected is a binary variable, splitting our sample into books that showed a significant reaction (greater than one standard deviation from the pre-event average level) to the exogenous shock and those that did not.

Persistence of the Shock (*PSR*) measures how long it takes (in days) before the effect diminishes and the demand returns to within one standard deviation of its pre-event average level.

Book/Network Parameters

Distance of a book is defined as the number of links on the minimal path extending across the network to the reviewed book.

Network Proximity extends the simple *distance* variable by taking into consideration a weighted average of all possible paths between A and B.

Local Clustering is a measure of how close a node and its neighbors are to being a clique (Watts 2003; Watts and Strogatz 1998).

Assortative Mixing and Link (relation) Parameters

We define several dyad (book-to-book) characteristics for links in the Amazon.com co-purchase network to reflect cross-product similarity. These characteristics include: category similarity, author, price, binding type (hardcover, soft cover, spiral) and vintage (difference in years between year of review and release year).

4. Identification of Cross Product Spillovers (Ripple Effect)

To study the depth of the spillover effect across the network we can compare the effect of the exogenous shock (i.e. the book review) on the demand for products at different distances from the source of the shock. Perhaps surprisingly, the data show that, on average, the spillover effect is limited to a close neighborhood around the source of the shock (the reviewed book). First neighbors see on average a 660% increase in *SalesRank* after the shock, whereas second and third neighbors see on average 50% and 13% increases in *SalesRank*, respectively. The results suggest that fourth and fifth neighbors may also be affected; however, these effects are not statistically significant (see Table 5-1 for summary statistics).

Nevertheless, the above analysis may be misleading; an observed correlated change in demand across products (which might be interpreted as ripple) can also arise from reasons other than the presence of a visible product network.

First, one should control for global changes in demand (for example, due to seasonality). To do that, a control group should be constructed, so that a difference-in-difference model can be used. One possibility is randomly selecting untreated products (i.e., not neighbors) as a control group. However, while this will control for global changes over time, it will not control for hidden product similarity (homophily).

The difficulty in analyzing real-world natural experimental settings is due to the lack of random assignment to treatment and control groups, creating selection bias. In our context, selection bias is introduced by two sources—selection of the product to be reviewed and selection of network neighbors to be presented.

It is natural to assume that books are not randomly selected to be reviewed, but rather, that there is some underlying process of selection (for example based on compatibility with taste of existing fans, popularity, the agenda *Oprah* wishes to promote, as well as various marketing efforts exerted by publishers). It is therefore possible that the types of books *Oprah* selects all have an unobserved set of shared characteristics, and those should be controlled for. We partially control for this source of bias by using two very different independent sources of exogenous shocks (i.e., the *New York Times* and *Oprah*). We also verified that the category distributions of the

books reviewed by *Oprah* and by the *New York Times* are very different. Acknowledging this limitation, we note that the extent of such an identification effect in the experimental setup we propose is expected to be somewhat limited compared with classical natural experiments. The reason is that the focus here is on the effect of the treatment (e.g., the demand shock to the reviewed book) on the network neighbors and not the effect of the review on the reviewed book; there is no evidence that *Oprah* or the *New York Times* has any influence in selecting the network neighbors of reviewed books[6].

The second source of selection bias is introduced by the selection of network neighbors. We would expect that a reviewed book's network neighbors share observed and unobserved characteristics with that book, thus making them potentially more susceptible to being affected by the review (due to group affiliation). For example, it may be the case that other books written by the same author experience an increase in sales regardless of the presence of a visible hyperlink. From the analysis of the co-purchase network we know that on average, one of five links on a given book's page (see Table 10-7) leads to a book with the same author. If this neighboring book also experiences an increase in demand, we may mistakenly attribute all of the change to the presence of the visible link, not taking into account the propensity of the neighboring book to be influenced due to the similarity between the two products. Therefore, the main endogeneity challenge in estimating the depth of the ripple involves the selection of the reference group in a way that controls for those similarity effects.

These issues are addressed in the following sections using a difference-in-differences model where the second difference is based on a matched sample that accounts for group affiliation.

Selection Model

More formally, the identification issue arises since the products in the network were not randomly assigned to treatment groups, denoted *T* (i.e. members and non-members of the reviewed book's sub-network). Therefore, we are unable to control

[6] We also know from Amazon's public statements and from conversations with senior managers at Amazon that Amazon does not interfere with the structure of the network.

for observed and unobserved characteristics that drive the selection into treatment groups.

Possible choices for reference groups include books of the same category, or books written by the same author. However, neither group captures all possible unobserved characteristics. One way the literature proposes to create a more reliable reference group is to use a matched sample based on propensity scores (Heckman et al. 1998b; Rosenbaum and Rubin 1983); for a recent use of propensity scores see Aral et al. (2009); Hill et al. (2006); Oestreicher-Singer and Zalmanson (2010) [7].

In a nutshell, instead of grouping products on the basis of observed covariates (such as category or author), we compute the propensity of each book to be treated, and group the books according to propensity score. Implementing this in our context, we created a matched sample based on a propensity score with nearest neighbor matching (Leuven and Sianesi 2003). Propensity score was computed using observed book characteristics[8] (Author, Category, average *SalesRank*, Price, Binding and Rating). For each network neighbor (b_{kj}) of a reviewed book (b_k^o) given event *k*, we assigned a matched book ($b_{kJ}^{'}$) to the matched sample such that the probability of $b_{kJ}^{'}$ to be a network neighbor was equal (or close enough) to that of b_{kj}, on the basis of a specific propensity score. Ideally, one would want the matched book $b_{kJ}^{'}$ to be as similar as possible to the network neighbor b_{kj}, and in general for the distribution of all observed properties of $\{b_{kj}\}$ and $\{b_{kJ}^{'}\}$ to be identical so that the only difference is

[7] Econometric literature (Greene 2008) suggests two possible solutions to the sample selection problem: Regression and Matching. In many cases, choosing whether to use a regression approach or a matching approach does not affect the results. The difference is that matching focuses on modeling the selection process, while regression assumes that one can model the outcome generation process. When the researcher understands the selection mechanism better than the outcome mechanism, a matching approach is likely to be more convincing.

[8] This is estimated using a logit model. However, the number of positive examples (network neighbors) is smaller in orders of magnitude from the number of negative examples (remaining candidate books from the co-purchase network). Under these conditions, the logit estimator is known to be biased (Ben-Akiva and Lerman 1985). We therefore followed the choice-based sampling suggested by the literature (Ben-Akiva and Lerman 1985), and under-sampled the negative observations in order to get a more balanced sample. Note that choice-based sampling is known to lead to inconsistent intercept estimation when using MLE (which can be corrected, by subtracting a constant term from the estimated intercept (Manski and Lerman 1977). In our case, however, the computed propensity score is only used as part of the matching procedure, and thus, a subtraction of a constant is superfluous.

the treatment. Once matching was complete, we could compare the effect of the exogenous shock on the treatment (neighbors) and control groups[9].

Difference-in-Differences Model

The difference-in-differences model is the most common statistical method designed to handle experimental designs involving data from several time periods (before and after a treatment is given) both for a group that received the treatment and for a control group that did not receive the treatment (Meyer (1995)).

With the increasing use of natural experiments as a basis for econometric studies, difference-in-differences methods have grown in popularity for the identification of average treatment effects (a few recent examples include: Chen et al. 2006; Chevalier and Mayzlin 2006; Danaher et al. 2010). Difference-in-differences extensions of matching have also been suggested (see, for example, Heckman et al. 1997; Heckman et al. 1998a) in which the assumption that the assignment to treatment group is not confounded is relaxed by only requiring unconfoundedness conditional on observables (Chandra and Collard-Wexler 2009).

We therefore define the following difference-in-differences model:

$$SRR_{it} = \beta_0 + \beta_{1t}\delta_t + \beta_{2t}T_i + \beta_{3t}\delta_t T_i + \beta_4 x_{it} + \varphi_i + \varepsilon_{it}$$

Where $t = \{0..3\}$ corresponds to $(t'..t'+2) = \{(-3..-1), (0..2), (3..5), (6..8)\}$, δ_t are time fixed effects, T_i is the assignment to treatment groups, x_{it} are observed covariates while ε_{it} are unobserved, and φ_i is a vector of either review source or review-level fixed effects. The coefficient β_{3t} on the interaction term between the time fixed effects and the group assignment is the difference-in-differences estimator.

Model estimation

We estimated the difference-in-differences regression model for all network neighbors up to five links away from a reviewed book and for their corresponding matched samples from the control group. The dependent variable in the model is the *SalesRankRatio* of the books, which measures the change in demand relative to the

[9] We note that an alternative to using a propensity score is performing "hard" matching based on all observed characteristics, the caveat being that typically it is hard to find matching candidates over the set of all observed characteristics. Nevertheless, using a propensity score (logit or probit models are most commonly used) may result in non-intuitive specific matching while preserving the global distributions over the treatment and control groups.

pre-event average level. Since the responses of products in the sub-network related to a single review event may be correlated, we clustered the standard errors at the review event level. Therefore, within each of the 83 review events, all neighboring books are allowed to be correlated. The values of the difference-in-differences estimator (β_{3t}) and its standard errors are shown in Table 4-1; column (B) adds review fixed effects, which produce similar results.

The coefficients for all difference-in-differences estimators are all positive and significant, suggesting that belonging to the treatment group (i.e. belonging to the network neighborhood of a reviewed book) has a positive influence on the *SalesRankRatio* following the review event; the change to *SalesRankRatio* is far beyond that observed in the control group, thus it is attributed to the visibility of the network.

Difference in Differences estimates using OLS Regression

	(A)	(B)
$\boldsymbol{diff - in - diff_{t=1}}$	0.223***	0.223***
	(0.0450)	(0.0651)
$\boldsymbol{diff - in - diff_{t=2}}$	0.238***	0.238***
	(0.0286)	(0.0399)
$\boldsymbol{diff - in - diff_{t=3}}$	0.221***	0.221***
	(0.0379)	(0.0410)
Constant	1.093***	1.094***
	(0.00352)	(0.0210)
Time Fixed Effects	Yes	Yes
Review Fixed Effects	No	Yes
Observations	131,533	131,533
F	130.3	58.45

Standard errors between parentheses, clustered at the review event level. Asterisks represent significance at the 10% (*), 5% (**) and 1% (***) levels.

Table 4-1: Results of Difference in Differences model for *SalesRankRatio* using OLS Regression.

To test whether the diffused shocks are limited to a local area around the source of the shock – we broke down the results according to distance from the reviewed book and ran the difference-in-differences model for all separate distance groups (see Table 4-2). The results of the estimation were consistent with previous findings; the difference-in-differences estimator was significant with a relatively large coefficient for the reviewed books and for their first, second and third network neighbors. These results suggest that the shock is limited to a small environment of

neighbors that are up to three links away (shortest distance) from the reviewed book. The number three is also surprisingly consistent with prior and recent studies (though conducted in other domains) on real-world networks with a high level of clustering (Fowler and Christakis 2010; Friedkin 1983).

For third neighbors, we see a positive and significant coefficient only for the two later time periods and an insignificant coefficient for the first time period after the review event. Similar positive coefficients are shown for fourth and fifth neighbors, though they are not significant.

One possible interpretation of the results for the distant (fourth and fifth) network neighbors is that distant network neighbors, on average, are prone to receive less attention from the network due to local network structures.

Difference in Differences estimates using OLS Regression of by distance

Distance	**0**	**1**	**2**	**3**	**4**	**5**
$\boldsymbol{diff-in-diff_{t=1}}$	39.87***	3.623***	0.421**	-0.0248	-0.0661	-0.0930
	(7.275)	(1.166)	(0.168)	(0.0443)	(0.0781)	(0.0640)
$\boldsymbol{diff-in-diff_{t=2}}$	19.35***	2.044***	0.415***	0.174***	0.0587*	0.0773***
	(3.895)	(0.619)	(0.137)	(0.0626)	(0.0322)	(0.0242)
$\boldsymbol{diff-in-diff_{t=3}}$	8.542***	1.165***	0.323**	0.226***	0.188	0.112***
	(1.584)	(0.366)	(0.125)	(0.0737)	(0.117)	(0.0284)
Constant	1.045	1.095***	1.060***	1.097***	1.095***	1.096***
	(1.545)	(0.265)	(0.0520)	(0.0222)	(0.0198)	(0.0157)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Review Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	624	2880	7504	16269	34124	70132
Adj. R^2	0.58	0.11	0.04	0.02	0.01	0.01
F	7.798	6.246	7.125	14.83	22.67	31.28

* Standard errors between parentheses, clustered at the review event level.
* Asterisks represent significance at the 10% (*), 5% (**) and 1% (***) levels.

Table 4-2: Results of difference-in-differences model for SalesRankRatio using OLS Regression.

For example, we can hypothesize that a high clustering coefficient around a reviewed book will direct a consumer's attention back to the close area around that book. This explanation is also consistent with the high variance observed in the response of the far network neighbors.

5. Which books are affected?

A considerable portion of the books in the network of a reviewed book showed a statistically significant change in demand following a shock (See Table 5-1). However, not all books were affected. For books reviewed on the *Oprah Winfrey Show*, 62% of first neighbors were significantly affected, as were 38% of second neighbors and 33% of third, fourth and fifth neighbors. For books reviewed in the *New York Times*, 47% of first neighbors were significantly affected, as were 36% of second neighbors and 33% of third, fourth and fifth neighbors. In comparison, over the previously mentioned random samples, we found that, on average, 17–23% (consistent across samples, review source and review date) of the books showed an increase in sales.

In this section we present an econometric model that explains the variance in the effect of the shock on different books in the network (i.e. why the effect on some books is significantly stronger than others) by directly modeling the different components, including local and global network structure and hidden product complementarities (assortative mixing).

Persistence and Shock Based on Sales Rank											
Source	Distance	(a) All Books			(b) Affected Books (Persistence>0)				(c) Books Not Affected (Persistence=0)		
		#	Average PSR	Average SRS	#	%	Average PSR	Average SRS	#	%	Average SRS
NYT	0	43	19.23	55.22	43	***100%	***19.23	55.22	0	0%	0.00
	1	214	2.75	4.34	101	***47%	***5.82	5.98	113	53%	2.87
	2	625	1.68	4.18	227	*36%	*4.63	5.47	398	64%	3.45
	3	1539	1.37	2.10	501	33%	4.19	3.19	1038	67%	1.57
	4	3435	1.28	1.99	1154	34%	3.80	3.30	2281	66%	1.32
	5	7370	1.38	2.00	2468	33%	4.12	3.22	4902	67%	1.39
Oprah	0	40	20.48	146.77	39	***98%	***21.00	146.89	1	3%	142.33
	1	191	6.37	29.57	118	***62%	***10.31	47.11	73	38%	1.23
	2	419	1.88	3.65	161	**38%	**4.89	6.79	258	62%	1.69
	3	879	1.37	2.07	288	33%	4.19	3.06	591	67%	1.58
	4	1734	1.06	1.92	552	32%	3.33	2.88	1182	68%	1.47
	5	3180	1.27	1.90	1035	33%	3.91	2.98	2145	67%	1.38

* Asterisks represent significance at the 10% (*), 5% (**) and 1% (***) levels for one sample paired t-test compared to the matched sample.

Table 5-1: Persistence and Shock statistics based on Sales Rank. Divided according to: (A) All books; (B) Books that were affected by the shock; and (C) Books that were not affected by the shock.

To explore the variations in resistance to exogenous demand shocks and analyze the factors that determine which neighboring book is affected by a shock, we use the following binary logistic model for the probability of a book being affected by the exogenous demand shock (see Table 5-2 for a description of the model variables) [10].

$$\ln\left(\frac{P(affected_i)}{1-P(affected_i)}\right) = \alpha_0 + \sum_{j=1}^{J} \alpha_j LOCAL_{ij} + \sum_{k=1}^{K} \beta_k GLOBAL_{ik} + \sum_{l=1}^{L} \gamma_l MIXING_{il} + \sum_{m=1}^{M} \delta_m CONTROL_{im}$$

Group	Measure	Description
LOCAL	Distance	Minimal distance from the reviewed book across the network.
	NetworkProximity	Normalized assessment of how "close" the neighboring book is to the reviewed book, taking into account all possible paths between them.
	LocalClustering	Measures how close a book and its neighbors are to being a clique.
GLOBAL	InDegree	Indegree of the book.
MIXING	SameAuthor	Books share the same author.
	SameCategory	Books belong to the same second-level category (based on Amazon's categories tree.
	SameVintage	Books have the same age (release date minus review date).
	SameBinding	Books have the same binding (Hardcover, Paperback, Spiral-bound).
	SamePrice	Price difference is up to $10.
CONTROL	AverageSalesRank [11]	Average *SalesRank* of the book in the two weeks prior to the event.
	DiscountRate	The discount rate of the book on the day of the event.
	Re-Run[12]	Dummy variable indicating the review was featured on a re-run show.
	Day of the Week	The day of the week when the review was published.
	Customer Reviews	Average rating and number of reviews.
	Review Source	Oprah ("1") or New York Times ("0").
* Fixed effects by day of week and review event		

Table 5-2: Description of variables in the binary logistic model for the probability of being affected by the exogenous shock.

[10] We note that after establishing the effect of treatment though matching, the logistic model we presented here is a regression-based approach for analysis. We believe that the combination of both approached adds robustness to the results of this work.

[11] The average Sales Rank was divided by 100,000 when entered into the logistic regression for readability reasons of the coefficient.

[12] Defined only for Oprah Winfrey's reviews.

Analysis of the co-purchase global network structure and of the event sub-network structures shows that there are large variations in local network structure, whereas the global network structure remains stable. This observation may suggest that both local and global network structure play a role in any process that takes place over the network. We therefore include the two types of structural network properties: global and local. Following the literature, we use the indegree centrality as a global measure of centrality[13].

Our local structure measures build on literature from social network analysis, which suggests that the more relations an actor is involved in, the higher the actor's visibility; this notion has been extended to construct a large family of centrality and prestige measures. Centola and Macy (2007) studied the process of complex contagion. They suggest that multiple sources of activation are required in order to spread complex contagions[14]. On the basis of these concepts, we defined three local network structure variables that draw from the notion of centrality—distance, network proximity and local clustering—and we adapted them for the context of product networks and the local influence limitation.

Our analysis aims to disentangle the relevant drivers of shock susceptibility (network structure and assortative mixing), controlling for other known drivers that, according to prior literature, influence demand (such as the day of week effect, the influence in changes in price through the discount rate of each product), as well as controlling for quality through consumer reviews and average rating).

Model estimation

We fully estimate the fixed effects models (day of week and review event) by maximizing the log likelihood. The incidental parameters estimation issue (which may lead to an inconsistent estimator due to small individual sample size) is less significant in our case due to the relatively large sizes of the individual samples; the average unbalanced sample includes 236 observations, which is clearly greater than

[13] This specification uses the degree centrality of a node. We also experimented with other types of centralities such as PageRank and eigenvector centrality, and results were robust.

[14] They also note that further work is required in order to understand the effects of heterogeneity of thresholds in the dynamics of complex diffusion, and they stress the importance of identifying the influence rather than homophily.

the range of 8 to 20 observations which was suggested to be sufficient by prior research (Greene 2001; Heckman 1981).

The results of the estimation are presented in Table 5-3 and strongly demonstrate the importance of both assortative mixing and network structure in the spillover patterns across the network. The coefficients of the majority of operationalized assortative mixing variables (such as: author, vintage and binding) are statistically significant. For example, having the same author as the reviewed book more than doubles the odds ratio of being affected by the shock.

Both local (local clustering) and global (indegree) network properties were found to have statistically significant effects. Interpreting the estimates for the clustering coefficient, each additional edge between the first neighbors of a book (which increases the local clustering coefficient by 1/30) results in a 1.2% increase in the odds ratio of being affected by the shock. The more clustered the network around a book, the greater the odds of the book being affected by the shock. The higher the indegree of a book, the lower the odds of it being affected by the shock, which is consistent with a positive coefficient (odds ratio > 1) on the average *SalesRank* of the book. This means that books in the tail are more likely to be affected by the shock.

Network proximity is also statistically significant and positively contributes to the probability of being affected by the shock. An additional 2-link path from the reviewed book to a specific book (which increases network proximity by 1/25) increases the odds ratio by 12.9%. Similarly, an additional 3-link path from the reviewed book results in a 2.5% increase to the odds ratio.

A more intuitive interpretation of the odds ratio is given by calculating the changes in predicted probability, by setting all other parameters to their mean values and fixing the test variable to the desired value. For example, **w**e see that first neighbors have a substantially higher probability of being affected compared with other neighbors (e.g. 11.3% more than second neighbors), which is expected due to their direct visibility from the same page as the reviewed book that is the source of the shock. Being a second neighbor (i.e. having a 2-link path from the reviewed book), in addition to being a first neighbor (which defines a triad), adds 3% to the target book's predicted probability of being affected by the shock. Similarly, being a third neighbor in addition to being a first neighbor (which creates a tetrad) adds 0.6% to the predicted probability of being affected by the shock.

Odds Ratio Estimates: Logit Model

	Distance Only	Network	Network and Mixing	Full Model	Full Model with DOW FE	Full Model with DOW & Event FE
Distance	1.03994*	1.02474	1.0205	1.01422	1.013	1.01505
	(0.0214)	(0.0212)	(0.0226)	(0.0228)	(0.0228)	(0.0243)
Network Proximity	35.92456***	35.62395***	21.17335***	21.55474***	20.69582***	24.58680***
	(15.4064)	(15.2546)	(10.1648)	(10.3870)	(10.0126)	(15.3290)
Local Clustering		1.26571**	1.21746*	1.43159***	1.44412***	1.48112***
		(0.1164)	(0.1244)	(0.1518)	(0.1535)	(0.2155)
In Degree		0.99266***	0.99321***	0.99603***	0.99598***	0.99623***
		(0.0008)	(0.0009)	(0.0008)	(0.0008)	(0.0009)
Same Author			2.11562***	2.05425***	2.08298***	2.12771**
			(0.4057)	(0.4187)	(0.4261)	(0.6610)
Same Category			1.10439*	1.03582	1.04138	1.06527
			(0.0650)	(0.0639)	(0.0643)	(0.0788)
Same Vintage			1.00998***	1.00727**	1.00731**	1.00599*
			(0.0028)	(0.0029)	(0.0029)	(0.0034)
Same Price			0.89582**	1.00313	0.99714	1.0066
			(0.0398)	(0.0512)	(0.0511)	(0.0703)
Same Binding			0.93393*	0.93605*	0.94082*	0.92957
			(0.0328)	(0.0343)	(0.0346)	(0.0479)
Average Sales Rank				1.16144***	1.16212***	1.15892***
				(0.0160)	(0.0161)	(0.0016)
Discount Rate				1.35237**	1.35576**	1.38654**
				(0.1918)	(0.1921)	(0.2218)
Re-Run				0.83289***	0.83810***	
				(0.0485)	(0.0502)	
Total Reviews				1.022	1.020	1.028*
				(0.0133)	(0.0134)	(0.0163)
Average Rating				1.231*	1.226*	1.170
				(0.136)	(0.135)	(0.156)
Review Source	1.02585	1.05226	1.03465	0.90505**	0.83867**	
	(0.0332)	(0.0343)	(0.0363)	(0.0422)	(0.0707)	
Observations	19,586	19,574	17,547	16,969	16,969	16,963
Log Likelihood	-12469.38	-12405.141	-11077.381	-10607.289	-10602.64	-10251.111
LR Chi Square	97.335	185.177	223.978	404.367	411.145	261.214
Event Fixed Effects	-	-	-	-	-	+
Day of Week Fixed Effects	-	-	-	-	+	+

* Standard errors between parentheses.

* Asterisks represent significance at the 10% (*), 5% (**) and 1% (***) levels.

Table 5-3: Estimation of a binary logistic model for the probability of being affected by the exogenous shock.

Surprisingly, when network proximity is controlled for, distance from the reviewed book does not significantly affect the odds of being affected by the shock. This indicates that each path between the reviewed book and the focal book matters. These results further support our conjecture that the visible hyperlinks of the product network influence the ripple process.

6. Post-shock Persistence

In this section we study the persistence of the multiple sequential aftershocks created by the ripple effect. We witness large variance in persistence, even among neighbors of the same distance from the source of a shock, and we study the factors that affect the observed persistence. Understanding the persistence of demand shocks is of great importance since the shape of the decay and its duration are central to understanding the economic value of these shocks.

In light of our results with regard to nodes' resistance to shocks, we expect that both product complementarities (assortative mixing) and network structure (local and global) will determine the persistence of these sequential aftershocks, and we model them based on duration model theory.

Duration model of shock persistence

To model the persistence of the diffused exogenous shocks we follow duration model theory and use a hazard-rate model (Greene 2008) where the hazard rate *h(t)* is the probability of the extinction of the diffused shock. In the context of this work, a "failure" occurs when the demand of a book returns to within one standard deviation of its pre-event average (we assume a "failure" occurs only once for each target book).

We first carried out a non-parametric Kaplan-Meier maximum likelihood estimation of the survival function. The results of the estimation (presented in Figure 6-1) provided an important insight: close (first and second) neighbors seemed to behave differently from distant (third and fourth) neighbors, and distant neighbors seemed to behave according to a similar survival function; this observation was also validated by the log rank test (p-value < 1%). Following this, we extended the analysis of the hazard rate model to allow separate hazard rate functions for neighbors based on their distance.

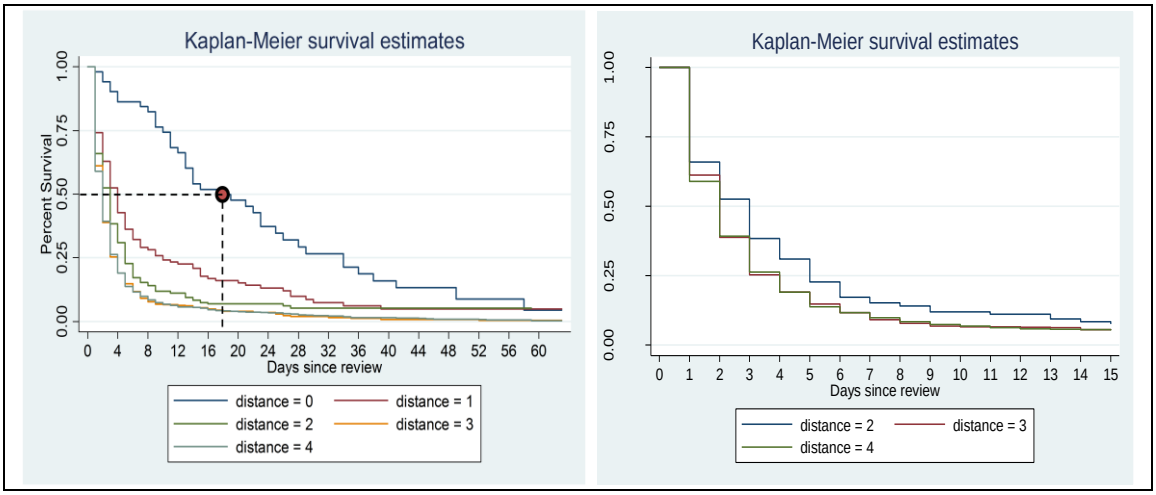

Figure 6-1: Kaplan-Meier estimations of the survival functions. On the left, the estimations for the reviewed books (distance=0) and their network neighbors (distance=1..4); On the left, a zoom into the Kaplan-Meier estimation of the survival functions for second, third and fourth neighbors (distance=2..4). The interpretation of the survival function is demonstrated, for example, by the marker point on the left chart, which shows that in 50% of the reviewed books the exogenous shock persisted over 18 days.

We therefore estimate the following exponential hazard rate model (model variables are similar to those used for the estimation of the binary logistic model; see Table 5-2 for a complete list):

$$h(t) = exp\left\{\alpha_0 + \sum_{j=1}^{J} \alpha_j LOCAL_{ij} + \sum_{k=1}^{K} \beta_k GLOBAL_{ik} + \sum_{l=1}^{L} \gamma_l MIXING_{il} + \sum_{m=1}^{M} \delta_m CONTROL_{im}\right\}$$

The results of the estimation of the exponential model are presented in Table 6-1. The estimation strategy is designed to provide robustness to the specification of variables by running several nested variations of the model (adding each of the parts one at a time: LOCAL, GLOBAL, MIXING and CONTROL). In addition, to evaluate robustness to different functional forms for the parametric distribution of the survival function, we consider a Weibull distribution (see column e) for the basic survival function, and a semi-parametric Cox-proportional hazard rate model (see column f). Following the Kaplan-Meier estimations, we also repeat the estimation of the model for different neighbors, grouped by their minimal distance from the reviewed book (see Table 6-2). The results of all parametric models (see Table 6-1) show that the main coefficients under interest are statistically significant and stable across all specifications. Naturally, only books that were affected were included in this

estimation, leaving us with 6,605data points. Note that survival analysis models allow censored data to be incorporated into the model. In this case, censoring was incorporated in two cases: (1) books for which the shock persisted for over 60 days were labeled as censored after 60 days (4 books in total) and (2) books with missing data for which the shock persisted until the last day of available data (50 books in total).

Distance and Network Proximity

Network proximity has a statistically significant and positive effect on shock persistence even when we control for the minimal distance from the reviewed book. Coefficients suggest a decrease of 10% in hazard rate for each additional 2-link path from the reviewed book and a decrease of 2% for each additional 3-link path. These results suggest that the persistence of the shock is highly influenced by the total number of paths connecting the node to the reviewed book (i.e., its *proximity*).

Interestingly, when we break down the analysis based on distance groups (Table 6-2), we find that when a book (node) is close to the reviewed book (at a distance of one or two clicks), the persistence of the shock is highly influenced by the node's *distance* from the reviewed book. However, we find that when a book is distant from the reviewed book (at a distance of three clicks or more) the persistence of the shock is highly influenced by the node's *proximity* to the reviewed book.

Local Clustering and the Fishing Net Effect

The analysis of the local structure of the sub-networks shows a large variation in clustering coefficients around the reviewed books. These differences drastically affect the number of neighbors and the structure of the sub-networks (for more details see Appendix B). A large clustering coefficient of a sub-network (represented by the "Shock Local Clustering" coefficient) suggests that customers traversing the links are highly likely to encounter the same set of books (which reside inside the cluster) over and over. This repetitive feedback may play a role in the duration of the shock to this

set of products (network neighbors) inside the cluster (close neighbors) and outside the cluster (distant neighbors)[15].

Our results show that when all network neighbors are pooled together (Table 6-1), the degree of local clustering around the reviewed book has a statistically significant effect on persistence, with a coefficient greater than one. Generally, this means that when the local area of the network around the shock is more clustered, the shock is likely to end sooner (persist less). Interestingly, the signs of the coefficients for close (first and second) and distant (third, fourth and fifth) neighbors are in opposite directions (Table 6-2). For example, an additional edge between first neighbors of the reviewed book (which increases the degree of local clustering by 1/30) results in a 5.4% reduction in the hazard rate for first neighbors, yet a 1.1% increase in the hazard rate for third, fourth and fifth neighbors.

Such results imply the existence of a *fishing net effect*: As the clustering coefficient of the sub-network composed of the reviewed book and its immediate (first) neighbors increases, the probability of triad and tetrad formation also increases. This process "traps" a greater proportion of the diffused influence closer to the reviewed book (books inside the "*fishing net*" enjoy a positive increase in persistence) rather than allowing it to spread further (so that books outside the "*fishing net*" suffer from a decrease in persistence). Inside this "*fishing net*" environment, a user entering the network at the reviewed book node or at one of it first neighbors has a greater chance of being re-directed to one of the books inside the net (i.e., the reviewed book and its close neighbors).

[15] Note that we also control for the focal book's local clustering coefficient (*Book Local Clustering*), which does not yield significant results.

Parameter Estimates: Exponential Duration Model (Hazard Rates)

	(a)	(b)	(c)	(d)	(e)	(f)
	Distance Only	**Network**	**Network and Mixing**	**Full Model (Exponential)**	**Full Model (Weibull)**	**Full Model (Cox)**
Distance	0.997	0.993	0.998	0.994	0.994	0.996
	(0.00780)	(0.0112)	(0.0215)	(0.0112)	(0.0117)	(0.0114)
Network Proximity	0.0777**	0.0616**	0.0669*	0.0688*	0.0678*	0.169*
	(0.0982)	(0.0797)	(0.0979)	(0.0988)	(0.103)	(0.154)
Shock Local Clustering		1.322**	1.347***	1.333***	1.338***	1.072***
		(0.175)	(0.00307)	(0.127)	(0.110)	(0.0103)
Book Local Clustering		1.144	1.184	1.176	1.177	1.140*
		(0.254)	(0.208)	(0.189)	(0.196)	(0.0904)
In Degree		1.001	1.001	1.003***	1.003***	1.002***
		(0.00161)	(0.00141)	(0.000518)	(0.000412)	(0.000408)
Same Author			0.844***	0.792***	0.792***	0.810***
			(0.0327)	(0.00408)	(0.00682)	(0.0374)
Same Category			1.135***	1.116***	1.117***	1.049*
			(0.0100)	(0.0156)	(0.0109)	(0.0259)
Same Vintage			0.997*	0.994**	0.994**	0.996***
			(0.00189)	(0.00303)	(0.00284)	(0.00119)
Same Price			1.155*	1.176***	1.177***	1.157***
			(0.0865)	(0.0430)	(0.0390)	(0.0474)
Same Binding			0.908**	0.912***	0.911***	0.966*
			(0.0419)	(0.0276)	(0.0313)	(0.0196)
Total Reviews				0.930***	0.929***	0.972**
				(0.0115)	(0.0143)	(0.0138)
Average Rating				0.888	0.887	0.947
				(0.107)	(0.112)	(0.0957)
Re-Run				1.046	1.046	1.040
				(0.0571)	(0.0583)	(0.0466)
Number of observations	6605	6605	5909	5711	5711	5711
AIC	19328.6	19313.6	17267.2	16572.5	16572.2	86184.8
Log Pseudolikelihood	-9663.3	-9655.8	-8632.6	-8285.3	-8285.1	-43091.4

*** Exponentiated coefficients (hazard rates). A value greater than 1 means that the parameter increases the hazard rate; standard errors between parentheses, adjusted for correlation among books belonging to the same review source.**

*** Asterisks represent significance at the 10% (*), 5% (**) and 1% (***) levels.**

Table 6-1: Table presents the estimation results of the exponential hazard rate model (a)-(d); and for the full model, the results from estimating a parametric Weibull hazard rate model (e) and the semi-parametric Cox-proportional hazard rate model (f).

Assortative Mixing

All assortative mixing variables are statistically significant (see Table 6-1), suggesting strong influence of product similarities on the persistence of the diffused shocks. The signs of the coefficients suggest that similarity has a complex influence. While consistent across model specifications, some dimensions of similarity (author, vintage and binding) reinforce the demand shock, whereas others (category and price) seem to have the opposite effect.

Having the same author clearly reduces the hazard rate and increases the persistence of the shock; this can be explained by the exposure the author receives from the review itself, which is translated into a persistent increase in sales of other books from the same author. Belonging to the same category, however, doesn't seem to increase the persistence of the shock; on the contrary, distant books that belong to the same category experience a reduction in persistence, suggesting that when consumers take the time to traverse the network and search for more books they are likely to diversify and purchase books from a different category.

Moreover, we find that for close neighbors of a reviewed book, the effect of similarity is not statistically significant. This may be related to the increase in the number of alternatives the consumer is exposed to as they explore more of the product network.

Consistent with prior literature highlighting the importance of consumer reviews (Chevalier and Mayzlin 2006; Duan et al. 2008; Forman et al. 2008; Ghose and Ipeirotis 2010), consumer reviews and ratings were found to be statistically significant and reduce the hazard rate, i.e. increase the persistence of the shock.

Parameter Estimates: Exponential Duration Model (Hazard Rates) by distance groups

	(a)	(b)	(c)
	First and Second Neighbors	**First Neighbors**	**Third, Fourth and Fifth Neighbors**
Distance	1.545**		0.954
	(0.275)		(0.0651)
Network Proximity	0.635	0.374*	0.00473*
	(0.796)	(0.223)	(0.0479)
Shock Local Clustering	0.478	0.187***	1.383**
	(0.224)	(0.0664)	(0.194)
Book Local Clustering	1.799	6.123	1.150
	(1.363)	(11.29)	(0.270)
In Degree	1.010***	1.013***	1.003***
	(0.00142)	(0.000299)	(0.000233)
Same Author	0.935	1.138	0.698***
	(0.0896)	(0.117)	(0.0265)
Same Category	1.254	0.962	1.110***
	(0.309)	(0.291)	(0.0229)
Same Vintage	0.982	0.978***	0.994***
	(0.0230)	(0.00392)	(0.00144)
Same Price	0.798	0.783	1.207***
	(0.142)	(0.296)	(0.0282)
Same Binding	1.223	1.094***	0.889*
	(0.211)	(0.0150)	(0.0596)
Total Reviews	0.933***	0.948***	0.933***
	(0.0111)	(0.00260)	(0.0131)
Average Rating	2.493	1.036	0.850*
	(2.174)	(0.669)	(0.0788)
Re-Run	0.945	0.616	1.074***
	(0.420)	(0.361)	(0.0167)
Number of observations	538	194	5173
AIC	1613.9	588.5	14901.9
Log Pseudo likelihood	-806.0	-293.3	-7449.9

*** Exponentiated coefficients (hazard rates), A value greater than 1 means that the parameter increases the hazard rate; Standard errors between parentheses, adjusted for correlation among books belonging to the same review source.**

*** Asterisks represent significance at the 10% (*), 5% (**) and 1% (***) levels.**

Table 6-2: Table presents the estimation results of the exponential hazard rate model for several test groups based on (minimal) distance.

7. Robustness

Prior literature has suggested equations to convert Sales Rank data into demand estimations (Goolsbee and Chevalier 2003; Brynjolfsson et al. (2003); Brynjolfsson et al. 2009). For robustness, all the analysis presented in this paper was repeated two more times - once using the demand estimations suggested in Brynjolfsson et al. (2003) rather than Sales Rank, and once using the demand estimations suggested in Brynjolfsson et al. (2009). We did not find changes in the magnitude or signs of coefficients; all results are available upon request (see Appendix D for details).

We also studied the sensitivity of our results to our definition of Affected. In section 3, we defined *Affected* (and corollary *Persistence*) as a maximal change in *SalesRank* that is greater than one standard deviation from the pre-event average level. Following prior literature on extreme events (Chollete 2009), we can generalize the definition of *Affected* to $\omega - affected$, which represents a maximal change in *SalesRank* that is greater than ω standard deviations from the pre-event average level, i.e.:

$$\omega - affected_i = \begin{cases} 1, & SR_{Peak,i} < \overline{SR}_i - \omega\sigma_{SR_i} \\ 0, & o/w \end{cases}$$

In this framework, the variable *Affected* presented above can be viewed as $1-affected$.

For robustness, we repeated the all the analysis presented in this paper replacing "*Affected*" with "$2 - affected$". The results of those estimates are very similar to the results presented here and are available upon request.

8. Conclusions

Our world is undergoing one of the largest technological revolutions of all time. Information is becoming available to all, and due to the rapid increase in the quantity of information, search engines and information technology tools such as recommender systems are important in assisting cognitively bounded consumers to find products that meet their needs (whether an interesting article to read or a product to purchase). As a result, the development of new information technologies has implications spanning far beyond mere technological advancement. Such technologies

often influence managerial, organizational and consumer behavior, and they transform business and society (Dhar and Sundararajan 2006).

This paper studies an increasingly important type of information technology in e-commerce that is relatively under-researched. The presence of hyperlinked product recommendation networks is one of the principal differences between the online and traditional channels of commerce. Hyperlinked product recommendation networks facilitate consumers' foraging among products, for example by directing them through pre-defined paths along virtual store aisles. A better understanding of the properties of these product networks allows us to gain insight into consumers' purchase behaviors, understand changes in patterns of demand, and influence future design and implementation of e-commerce information systems.

In this paper, we focus on the online contagion of exogenous demand shocks created by media events. The media events we consider are book reviews featured on the *Oprah Winfrey* television show and in the Sunday edition of the *New York Times*. We study the impact and ripple effect of these exogenous events on the demand for a "network" of related books that were not explicitly mentioned in a review but were located "close" to a reviewed book in the online co-purchase product network of Amazon.

Using a difference-in-differences matched sample approach, we identified the extent of the variations caused by the visibility of the online network (i.e., by consumers clicking on visible hyperlinks) and distinguished this effect from variation caused by hidden product complementarities. We found a strikingly high level of ripple of exogenous shocks through such networks. Neighboring books experienced a dramatic increase in their demand levels, even though they were not actually featured in a review; this effect is indicative of the depth of contagion in online recommendation networks following exogenous shocks. However, in comparison to prior research on ripple in networks and the potential extent of ripple (given the size of the network), this effect is limited to a relatively small area (up to three clicks away) around the source of the shock, mainly due to the local structure of these networks.

We find that product characteristics, assortative mixing and local network structure play an important role in explaining which books will be affected by the shock, as well as the relative persistence of the multiple sequential aftershocks. The

local network structure and specifically the number of directed paths (which direct consumers' attention from the source of the shock) were found to affect the persistence of a shock to distant neighbors. Most interestingly, we found that clustered networks "trap" a higher fraction of the contagion closer to the reviewed book. This structure increases the persistence of the shock among close neighbors and decreases the persistence of shocks to distant neighbors.

This research provides an important documentation of the magnitude and persistence of ripple of demand shocks across product networks, as well as evidence of the important role and influence of product networks in electronic commerce (specifically in the presence of exogenous shocks). These findings have significant managerial implications, for design as well as for marketing and strategy.

Product recommendation networks are growing and becoming standard in modern e-commerce. (Examples of sites that integrate product networks are Amazon, Barnes & Noble, YouTube, iTunes, and even Yelp, which provides a network of co-viewed restaurants.) This research demonstrates the potential and importance of studying product networks, which allow us to gain insights into consumers' behavior and analyze changes in demand patterns. The use of a network as a research framework allows us to model and study these additional features within the same construct (in the example given here by adding "forward edges" or "shortcuts" to the network) and should be studied more by researchers.

Acknowledgments

The authors thank Eitan Muller, Barak Libai and David Godes for many helpful discussions, and seminar participants at New York University, Carnegie Mellon University, the 2009 SCECR Conference, the 2009 Marketing Science Institute Conference, the 2009 INFORMS-CIST Conference and the 2009 ICIS Conference for helpful comments. Financial support from the NET Institute and the Google and WPP Marketing Research award is gratefully acknowledged.

9. References

Aral, S., Muchnik, L., and Sundararajan, A. "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences of the United States of America* (106:51) 2009.

Bae, K.-H., Karolyi, G.A., and Stulz, R.M. "A New Approach to Measuring Financial Contagion," *Rev. Financ. Stud.* (16:3), July 1, 2003 2003, pp 717-763.

Bakos, Y. "Reducing buyer search costs: implications for electronic marketplaces," *Management Science* (43:12) 1997, pp 1676-1692.

Bala, V., and Goyal, S. "Learning from Neighbours," *The Review of Economic Studies* (65:3) 1998, pp 595-621.

Balogh, S. "Oprah Winfrey rallies voters for Barack Obama," in: *Heraldsun.com.au*, 2008.

Ben-Akiva, M., and Lerman, S. *Discrete choice analysis: theory and application to travel demand* The MIT Press, 1985.

Boatwright, P., Basuroy, S., and Kamakura, W. "Reviewing the reviewers: The impact of individual film critics on box office performance," *Quantitative Marketing and Economics* (5:4) 2007, pp 401-425.

Brynjolfsson, E., Hu, Y., and Smith, M. "A Longer Tail?: Estimating The Shape of Amazon's Sales Distribution Curve in 2008," in: *Workshop on Information Systems and Economics (WISE)*, 2009.

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

Centola, D. "The New Petri-Dish: Social Science Experiments on the Internet," The Workshop on Information in Networks (WIN), 2009.

Centola, D., and Macy, M. "Complex Contagions and the Weakness of Long Ties 1," *American Journal of Sociology* (113:3) 2007, pp 702-734.

Chandra, A., and Collard-Wexler, A. "Mergers in Two-Sided Markets: An Application to the Canadian Newspaper Industry," *Journal of Economics & Management Strategy* (18:4) 2009, pp 1045-1070.

Chellappa, R.K., and Chen, C. "On the Temporal Nature of Sales-Rank Relationships of Albums and Digital Tracks in the Music Industry: The Relevance of Billboard Charts Post-Digitization," The proceedings of the INFORMS annual meeting, Washington, 2008.

Chen, P., Dhanasobhon, S., and Smith, M. "All reviews are not created equal," in: *SSRN Working Paper*, 2006.

Chevalier, J., and Mayzlin, D. "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research* (43:3) 2006, pp 345-354.

Chintagunta, P.K., and Haldar, S. "Investigating purchase timing behavior in two related product categories," *Journal of Marketing Research*) 1998, pp 43-53.

Chollete, L. "The propagation of financial extremes," *Discussion Papers*) 2009.

Cointet, J.-P., and Roth, C. "Information diffusion on realistic networks," Proceedings of AlgoTEL 9th "francophone summit on algorithms for telecommunications, The Centre for Research in Social Simulation, 2007.

Danaher, B., Dhanasobhon, S., Smith, M.D., and Telang, R. "Converting Pirates Without Cannibalizing Purchasers: The Impact of Digital Distribution on Physical Sales and Internet Piracy," *SSRN eLibrary*), March 3 2010.

Dellarocas, C., Katona, Z., and Rand, W. "Media, aggregators and the link economy, Working paper," 2009.
Deschatres, F., and Sornette, D. "Dynamics of book sales: Endogenous versus exogenous shocks in complex networks," *Physical Review E* (72:1) 2005, p 016112.
Dhar, V., and Sundararajan, A. "Does IT matter in business education? Interviews with business school deans," 2006.
Domingos, P., Nath, A., and Richardson, M. "Modeling and Optimizing Word of Mouth with Markov Logic," The Workshop on Information in Networks (WIN), 2009.
Duan, W., Gu, B., and Whinston, A.B. "Do online reviews matter? - An empirical investigation of panel data," *Decis. Support Syst.* (45:4) 2008, pp 1007-1016.
Edwards, Y.D., and Allenby, G.M. "Multivariate analysis of multiple response data," *Journal of Marketing Research* (40:3) 2003, pp 321–334.
Eguíluz, V.M., and Klemm, K. "Epidemic Threshold in Structured Scale-Free Networks," *Physical Review Letters* (89:10) 2002, p 108701.
Eugene, A., Eric, B., Susan, D., and Robert, R. "Learning user interaction models for predicting web search result preferences," in: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, Seattle, Washington, USA, 2006.
Forman, C., Ghose, A., and Wiesenfeld, B. "Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research* (19:3) 2008, pp 291-313.
Fowler, J., and Christakis, N. "Cooperative Behaviour Cascades in Human Social Networks," *Arxiv preprint arXiv:0908.3497*) 2009.
Fowler, J., and Christakis, N. "Cooperative behavior cascades in human social networks," *Proceedings of the National Academy of Sciences*) 2010.
Friedkin, N. "Horizons of observability and limits of informal control in organizations," *Social Forces* (62:1) 1983, pp 54-77.
Ghose, A., and Gu, B. "Search Costs, Demand Structure and Long Tail in Electronic Markets: Theory and Evidence," in: *NET Institute Working Paper No. 06-19*, 2006.
Ghose, A., and Ipeirotis, P.G. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering (TKDE), forthcoming.*) 2010.
Ghose, A., Smith, M., and Telang, R. "Internet exchanges for used books: An empirical analysis for product cannibalization and social welfare," *Information Systems Research* (17:1) 2006, pp 3–19.
Goldenberg, J., Oestreicher-Singer, G., and Reichman, S. "The quest for content: The integration of product networks and social networks in online content exploration, Available at SSRN: http://ssrn.com/abstract=1538283," 2010.
Goolsbee, A., and Chevalier, J. "Measuring Prices and Price Competition Online: Amazon and Barnes and Noble," *Quantitative Marketing and Economics* (1) 2003, pp 203-222.
Granovetter, M. "The Strength of Weak Ties: A Network Theory Revisited," *Sociological Theory* (1) 1983, pp 201-233.
Greene, W. "Estimating Econometric Models with Fixed Effects," 2001.
Greene, W. *Econometric Analysis*, (6th Edition ed.) Prentice Hall, 2008.

Groot, R.D. "Consumers don't play dice, influence of social networks and advertisements," *Physica A: Statistical Mechanics and its Applications* (363:2) 2006, pp 446-458.

Heckman, J. "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," *Structural analysis of discrete data with econometric applications*) 1981, pp 179-195.

Heckman, J., Ichimura, H., and Smith, J. "ET P. TODD (1997),"Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme"," *Review of Economic studies* (64) 1997, pp 605-654.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. "Characterizing selection bias using experimental data," *Econometrica* (66:5) 1998a, pp 1017-1098.

Heckman, J., Ichimura, H., and Todd, P. "Matching as an econometric evaluation estimator," *Review of Economic studies* (65:2) 1998b, pp 261-294.

Hill, S., Provost, F., and Volinsky, C. "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science* (21:2) 2006, p 256.

Illouz, E. *Oprah Winfrey and the Glamour of Misery: An Essay on Popular Culture* New York: Columbia University Press, New York, 2003.

Jackson, M.O. "Networks and Economic Behavior," *Annual Review of Economics* (1:1) 2009.

Kakimura, J.-I., Kitamura, Y., Takata, K., Umeki, M., Suzuki, S., Shibagaki, K., Taniguchi, T., Nomura, Y., Gebicke-Haerter, P.J., Smith, M.A., Perry, G., and Shimohama, S. "Microglial activation and amyloid-ß clearance induced by exogenous heat-shock proteins," *FASEB J.* (16:6), April 1, 2002 2002, pp 601-603.

Karrer, B., and Newman, M. "A message passing approach for general epidemic models," *Arxiv preprint arXiv:1003.5673*) 2010.

Kempe, D. "Structure and Dynamics of Information in Networks," Department of Computer Science, University of Southern California, 2010.

Laura, A.G., Thorsten, J., and Geri, G. "Eye-tracking analysis of user behavior in WWW search," in: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, Sheffield, United Kingdom, 2004.

Leuven, E., and Sianesi, B. "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing," 2003.

Libai, B., Muller, E., and Peres, R. "The Role of Within-Brand and Cross-Brand Communications in Competitive Growth," *Journal of Marketing*) 2008.

Libai, B., Muller, E., and Peres, R. "Source of Social Value in Word of Mouth Programs," in: *Marketing Science Institute working paper*, 2010, pp. 10-103.

Lloyd, A.L., and May, R.M. "EPIDEMIOLOGY: How Viruses Spread Among Computers and People," *Science* (292:5520), May 18, 2001 2001, pp 1316-1317.

Manchanda, P., Ansari, A., and Gupta, S. "The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions," *Marketing Science* (18:2) 1999, pp 95-114.

Manski, C., and Lerman, S. "The estimation of choice probabilities from choice based samples," *Econometrica: Journal of the Econometric Society* (45:8) 1977, pp 1977-1988.

Mayer, A. "Online social networks in economics," *Decision Support Systems* (47:3) 2009, pp 169-184.

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

McDonald, M., Suleman, O., Williams, S., Howison, S., and Johnson, N.F. "Impact of unexpected events, shocking news, and rumors on foreign exchange market dynamics," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* (77:4) 2008, pp 046110-046112.

Meyer, B.D. "Natural and Quasi-Experiments in Economics," *Journal of Business & Economic Statistics* (13:2) 1995, pp 151-161.

Mike J. Jeger, M.P.O.H.M.W.S. "Modelling disease spread and control in networks: implications for plant sciences," *New Phytologist* (174:2) 2007, pp 279-297.

Morris, M. "Sexual networks and HIV," *AIDS (London, England)* (11) 1997, pp 209-216.

Morris, S. "Contagion," *The Review of Economic Studies* (67:1) 2000, pp 57-78.

Newman, M.E.J. "Ego-centered networks and the ripple effect," *Social Networks* (25:1) 2003a, pp 83-95.

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

Newman, M.E.J., Barabási, A.-L., and Watts, D.J. *The structure and dynamics of networks* Princeton Univ Pr, 2006.

Newman, M.E.J., and Park, J. "Why social networks are different from other types of networks," *Physical Review E* (68:3) 2003, p 036122.

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

Oestreicher-Singer, G., and Sundararajan, A. "The Visible Hand of Social Networks in Electronic Markets," Working Paper, New York University, 2008.

Oestreicher-Singer, G., and Zalmanson, L. "Paying for Content or Paying for Community? The Effect of Consumer Involvement on Willingness to Pay on Media Web Sites," *SSRN eLibrary*) 2010.

Oh, J., Susarla, A., and Tan, Y. "Examining the Diffusion of User-Generated Content in Online Social Networks," SSRN, 2008.

Reinstein, D.A., and Snyder, C.M. "The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics," *Journal of Industrial Economics* (53:1), March 2005 2005, pp 27-51.

Rooney, K. *Reading with Oprah: The Book Club That Changed America* University of Arkansas Press, Fayetteville, 2005.

Rosenbaum, P., and Rubin, D. "The central role of the propensity score in observational studies for causal effects," *Biometrika* (70:1) 1983, p 41.

Rosenthal, M. "Amazon Sales Rank For Books," 2010.

Sorensen, A., and Rasmussen, S. "Is Any Publicity Good Publicity? A Note on the Impact of Book Reviews," in: *Working Paper*, 2004.

Sornette, D. "Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth," *Proceedings of the National Academy of Sciences of the United States of America* (99:Suppl 1), February 19, 2002 2002, pp 2522-2529.

Sornette, D. "Endogenous versus Exogenous Origins of Crises," in: *Extreme Events in Nature and Society*, 2006, pp. 95-119.

Sornette, D., Deschatres, F., Gilbert, T., and Ageon, Y. "Endogenous Versus Exogenous Shocks in Complex Networks: An Empirical Test Using Book Sale Rankings," *Physical Review Letters* (93:22) 2004, p 228701.

Sornette, D., Malevergne, Y., and Muzy, J.F. "Volatility fingerprints of large shocks: Endogeneous versus exogeneous," *RISK* (16:2) 2002, pp 67-71.

Watts, D.J. *Small worlds: the dynamics of networks between order and randomness* Princeton University Press, Princeton, N.J., 2003.

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

Watts, D.J., and Strogatz, S.H. "Collective dynamics of /`small-world/' networks," *Nature* (393:6684) 1998, pp 440-442.

Yook, S.-H., Jeong, H., and Barabási, A.-L. "Modeling the Internet's large-scale topology," *Proceedings of the National Academy of Sciences of the United States of America* (99:21), October 15, 2002 2002, pp 13382-13386.

10. Appendices

10.1. Appendix A – 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 give 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 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 10-1, 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 four and five hours. The script is run each day at midnight.

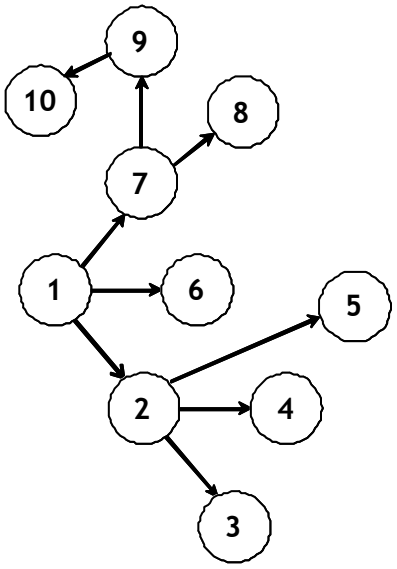

Figure 10-1: Illustrates depth-first search used for graph traversal.

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.

10.2. Appendix B – Network Statistics

10.2.1. Co-Purchase networks

Table 10-1 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 (~$1.45*10^{-5}$) due to the truncation to 5 outgoing links per node[16]; 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 10-2 for degree distribution and distribution of betweenness centrality on a sample daily network).

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

Table 10-1: Network statistics for the large connected component of the Amazon co-purchase networks.

[16] Since each node has up to 5 outgoing edges, the maximal theoretic network density (a proxy for the average level of activity in the network) is $\frac{5n}{n(n-1)} = \frac{5}{n-1} \cong 1.8 \times 10^5$

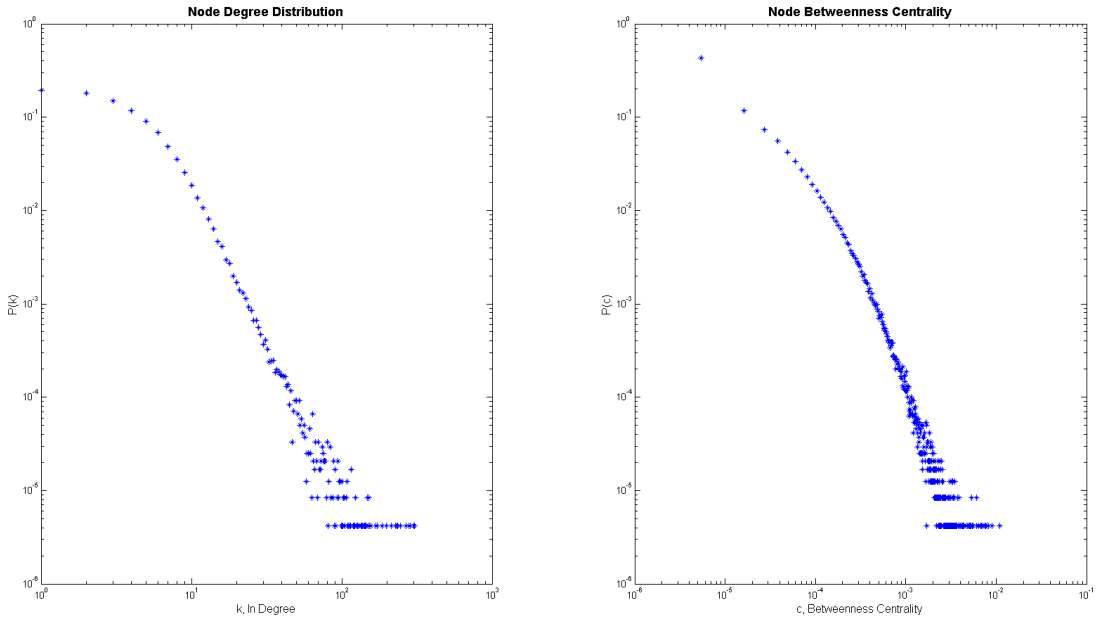

Figure 10-2: Node degree distribution of the large connected component of the Amazon co-purchase networks at 2007-09-16; the network has 319,340 nodes and 1,452,602 edges.

10.2.2. 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 sub-network from the co-purchase graph starting from the reviewed book and up to a distance of 5 links away (the 5th network neighbor of the reviewed book). Following Deschatres and Sornette (2005) we manually classified the review events into two categories: (1) Exogenous Shocks; (2) Endogenous & Multiple Shocks (See Figure 10-3). 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 sub-networks.

Table 10-2 presents basic network statistics on the sub-networks up to a distance of 5 links away (the 5th network neighbor of the reviewed book). The relatively high variance in the average clustering coefficient of these networks (as illustrated in Figure 10-4) shows that they are significantly different from each other, which may be reflected in the way exogenous shocks diffuse through the network.

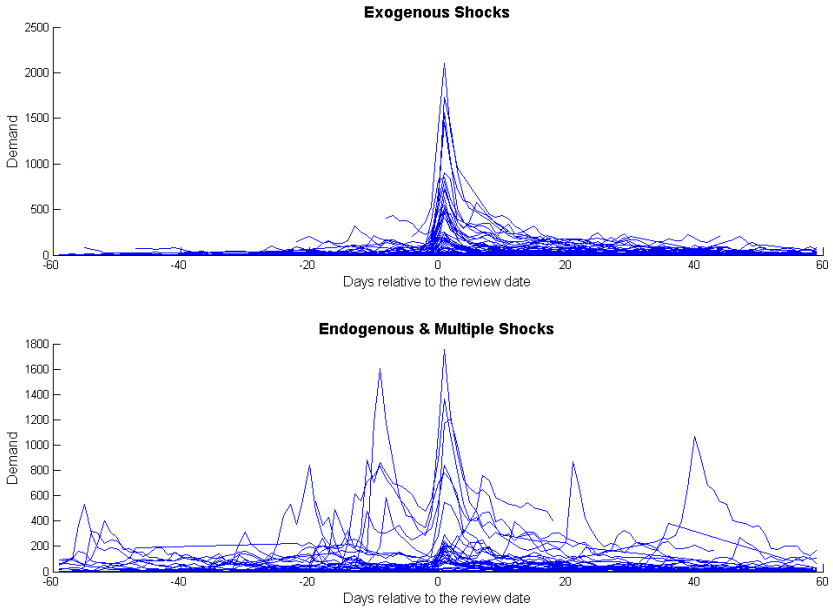

Figure 10-3: Reviewed books time series data classification into two categories: Exogenous Shocks (top); Endogenous & Multiple Shocks (bottom).

Variable	# Nodes	# Edges	Average In Degree	Fraction of reciprocal links	Average Clustering 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

Table 10-2: Network statistics across the sub-networks up to the 5th network neighbor for each of the reviewed books' events.

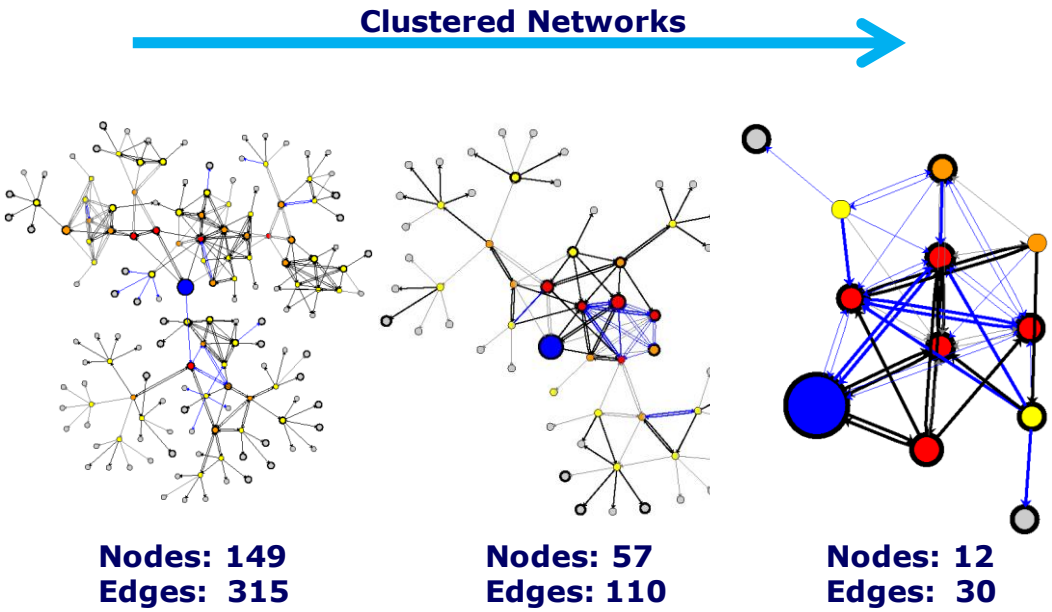

Figure 10-4: Examples for sub-networks with increasing clustering coefficient.

10.3. Appendix C - Detailed description of constructed variables

10.3.1. Shock Parameters

Sales Rank (*SR*) is a number associated with each product on Amazon.com, which measures its demand relative to other products[17]. The best-selling product is therefore ranked 1, followed by 2, 3, and so on. The Sales Rank of each product on Amazon is updated several times a day, and prior research has shown that there are intra-day fluctuations; therefore, we use a 24h average of the Sales Rank.

Prior literature has developed measures of estimated demand levels, based on Sales Rank data (Ghose and Gu 2006; Ghose et al. 2006; Oestreicher-Singer and Sundararajan 2008). However, those conversion measures are inappropriate when discussing high-selling products, such as some of the books in our sample. For an extended discussion on Sales Rank conversion to demand and evaluation of robustness see Appendix D.

Pre-Event Average Sales Rank ($\overline{SR_i}$) - To assess the magnitude of response to the exogenous shock we follow a common procedure in extreme event studies (Chollete 2009) and compute the *pre-event average Sales Rank* ($\overline{SR_i}$) of all products. This is based on the assumption that every book has a stable pre-event Sales Rank, which can be estimated using the average Sales Rank in the two weeks prior to the day of the review[18].

SalesRankRatio (*SRR*) measures the magnitude of the event at time *t* and is defined as: $SRR_{i,t} = 1 \Big/ \frac{SR_{i,t}}{\overline{SR_i}}$, where $SR_{i,t}$ is the average daily Sales Rank of book *i* on day *t*[19]. This measure is computed daily for each book in the sample (the reviewed books and their network neighbors) for the period ranging from two weeks prior to the date of the review until two months after the date of the review.

SalesRankShock (*SRS*) measures the maximal short-term change in the Sales Rank of a book following the exogenous shock, and represents the peak of the sales

[17] Amazon does not disclose the actual sales information.

[18] Choosing a large window is problematic since it increases the likelihood of interference from uncontrolled exogenous events. On the other hand, we would like to use the largest possible window in order to best characterize the pre-event patterns. We experimented with various window sizes; results were found to be robust with window sizes of 1 to 4 weeks.

[19] We use the reciprocal of the standard ratio since a lower Sales Rank corresponds to a higher level of sales; thus a decrease in the sales rank corresponds to an increase in sales.

increase relative to the pre-event average. Formally: $SRS_i = 1 \Big/ \frac{SR_{Peak,i}}{\overline{SR_i}}$, where $SR_{Peak,i}$ is the peak Sales Rank reached by the book in the 72-hour interval immediately following the review[20]. SRS can therefore also be defined as $\max\{SRR\}_{0 \le t \le 2}$.

Affected is a binary variable, splitting our sample into books that showed a significant reaction to the exogenous shock and those that did not. *Affected* is defined as "1" if the maximal change in Sales Rank is greater than one standard deviation from the pre-event average level. We therefore first compute the pre-event mean $\overline{SR_i}$ and standard deviation σ_{SR_i} of each book i and compare it to the *SalesRank* peak of that book: $Affected_i = \begin{cases} 1, & SR_{Peak,i} < \overline{SR_i} - \sigma_{SR_i} \\ 0, & o/w \end{cases}$;

All estimations were validated for robustness to the above specification of shock following prior literature on extreme events (Chollete 2009); see Appendix E for details.

Persistence of the Shock (*PSR*) measures how long it takes before the effect diminishes and the demand returns to its pre-event average level. Following event study methodology we estimate the *PSR* by computing the time required for the book to return to within one standard deviation of its pre-event average *SalesRank*. For each book which was affected by the shock ($Affected_i$ =1) we calculate the number of days until the *SalesRank* of the book first exceeds $\overline{SR_i} - \sigma_{SR_i}$. For computational reasons we truncate persistence to 64 days after the date of the review (truncation was necessary for 16 out of 20,024 books in our sample); however, the estimation method we use to study persistence (i.e. Duration Models) is able to incorporate truncated data such as these.

10.3.2. Book/Network Parameters

Distance of a book is defined as the number of links on the minimal path extending across the network to the reviewed book. By definition, the reviewed book has a distance of 0, its first neighbors have a distance of 1, its second neighbors will

[20] There is a tradeoff to consider when choosing the size of this window: extending the window size ensures we capture the full magnitude of the shock's peak, but it might also introduce noise. We experimented with window sizes of 24-72 hours following the initial response to the event, with no significant differences in the corresponding SRS values.

have a distance of 2, and so on. In graph theoretic terminology, distance is the geodesic distance between the reviewed book and the book in the network.

Network Proximity extends the simple *distance* variable (which provides a limited assessment of how "close" neighboring book A is to reviewed book B) by taking into consideration *all possible* paths between A and B. *Network Proximity* addresses this by providing a normalized assessment of how much attention potentially flows (assuming communication flows through all links in an identical manner) from one book to another based on a damped summation of all paths, given by: $Proximity_i = \sum_{k=0}^{d} \frac{N_{ik}}{5^k}$; where N_{ik} is the number of times book *i* is a *k-neighbor* of the reviewed book[21] and *d=5*[22]. There are two main assumptions we would like to note: (1) We ignore paths containing loops (backward edges), i.e. we assume that the conditional probability for a user to click on a link he or she already viewed (i.e. using a backward link) is 0. This assumption can be relaxed by assuming a similar probability to that of clicking a new link or some fraction of this probability[23]. (2) We assume all links are equal, while studies in the field of clicks on search engines have shown that the probability to click on a link drops sharply with rank (Eugene et al. 2006; Laura et al. 2004).

Local Clustering is a measure of how close a node and its neighbors are to being a clique (Watts 2003; Watts and Strogatz 1998) and is computed as:

$$CC_i = \frac{\left|\text{Edges between } v_i \text{ and its neighbors}\right|}{\left|\text{Outgoing edges from } v_i \text{ and its neighbors}\right|} = \frac{\left|\{e_{ij} \cup e_{ji}\}\right|}{k_i(k_i - 1)}; \quad e_{ij}, e_{ji} \in E, v_i \in V$$

The average of *local clustering* over all nodes in the network is called the *clustering coefficient* of the network. Empirical studies show that social networks exhibit a high average *clustering coefficient* (Newman 2003a; Newman and Park 2003) compared to random networks. The *clustering coefficient* has been shown to play an important role in the diffusion of information (Bala and Goyal 1998; Morris 2000). The finding that dense network clusters and overlapping neighbors may slow down the diffusion process (Bala and Goyal 1998; Granovetter 1983) led to claims

[21] Recall that each book in our network has five outgoing links, hence the choice of denominator.
[22] Our preliminary study shows that the diffusion of the shock is limited to a small radius around the reviewed book; this is also consistent with recent findings by Centola (2009); Domingos et al. (2009); Fowler and Christakis (2009). We also experimented with d=4 and results are robust.
[23] We did not observe any significant change in results by changing this assumption.

that these types of networks are protected against the spread of viruses (Eguíluz and Klemm 2002). Nevertheless, Eguíluz and Klemm (2002) also showed that for networks with scale-free distribution of degree, high clustering and a short average path length (which are typical of many real-world networks such as the Internet, as noted by Yook et al. (2002)), there is a threshold infection probability above which a virus can spread across the network. Cointet and Roth (2007) also argued that the clustering coefficient may have greater influence on diffusion than the commonly used degree distribution.

In the context of product networks, it is interesting to study even a local ripple process (which does not spread across the entire network) since it may have substantial economic and marketing implications.

Following the above, we explore the effects of the network's level of clustering, focusing on the *local clustering* computed for the reviewed books and their network neighbors. We find that the average *local clustering coefficient* for books reviewed on the *Oprah Winfrey Show* is 0.5, while books reviewed by the *New York Times* had an average local clustering coefficient of 0.41; both are on average higher than the average *clustering coefficient* across the entire network (0.39).

10.3.3. Assortative Mixing and Link (Dyad) Parameters

Prior literature (Newman 2003b) draws a strong connection between network structure and the level of assortative mixing (link / relation characteristics). Extensive studies on social networks have also shown that assortative mixing and network structure affect the diffusion patterns across the network (Morris 1997). It was shown (Libai et al. 2008) that word-of-mouth generates both within-brand and cross-brand influence on sales, suggesting that an exogenous demand shock following a review for a specific book will result in an increase in demand in the entire category. Nevertheless, Oestreicher-Singer and Sundararajan (2008) used the Amazon.com network to demonstrate that the explicit presence of a recommendation link had a significant influence on demand even after controlling for category similarity.

We define the following book-to-book characteristics for links in the Amazon.com co-purchase network to reflect consumer taste: category similarity, author, price, binding type (hardcover, soft cover, spiral) and vintage (difference in years between year of review and release year).

10.3.4. Summary statistics

Summary statistics for a selection of shock constructed variables are given in Table 10-3. Consistently with the findings of Oestreicher-Singer and Sundararajan (2008), 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 10-4 for an example and Table 10-5 for summary statistics).

Variable	Average Sales Rank	Persistence (Sales Rank)	SRS
Mean	126,759	1.48	2.59
Median	46,569	0.00	1.43
Max	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

Table 10-3: Summary statistics for a selection of constructed variables.

Further exploration of the distribution of persistence across different groups of neighbors based on minimal distance from the reviewed book (see Figure 10-5) shows a considerable amount of variation across books.

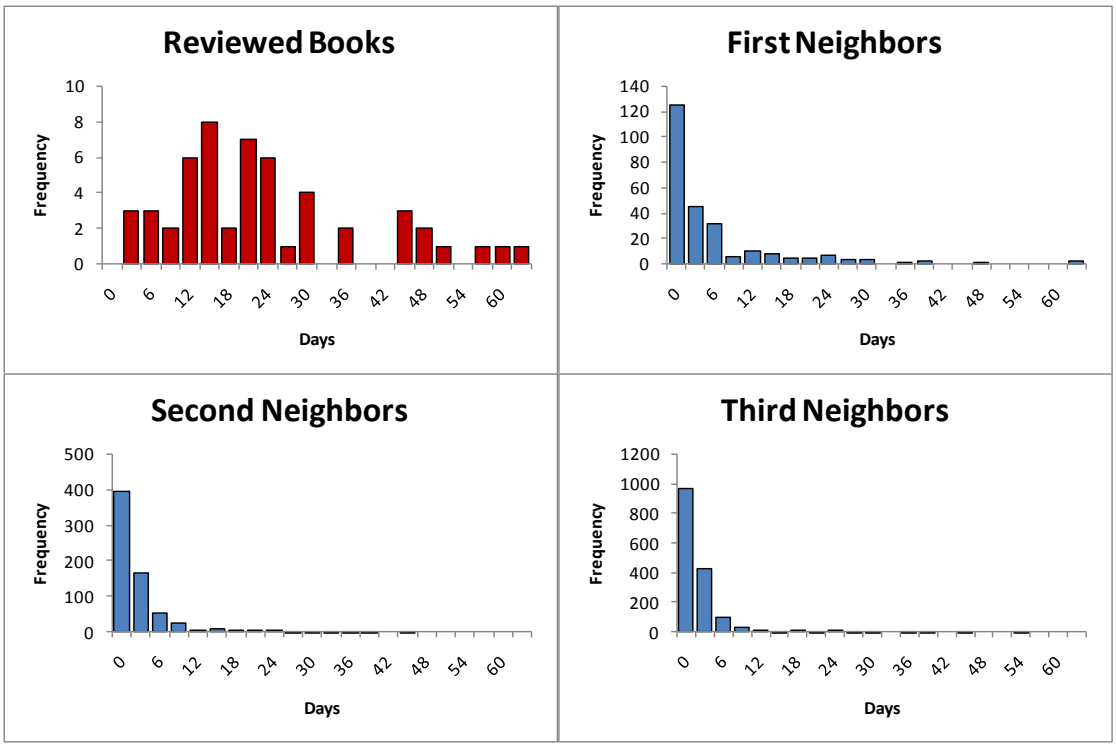

Figure 10-5: The distribution of persistence, the number of post-event days in which demand remained one standard deviation above the pre-event average demand for the reviewed books and first, second and third network neighbors. Graphs are based on the sub-networks of books reviewed by *Oprah* and the *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 category path comparing: (1) only the top category; (2) only the two top categories; (3) only the three top categories.

Level 1 Category	Level 2 Category
Children's Books	People & Places
Children's Books	Educational
Children's Books	Holidays & Festivals
Literature & Fiction	History & Criticism
Literature & Fiction	Poetry
Literature & Fiction	Drama
Nonfiction	Education
Nonfiction	Social Sciences
Nonfiction	Politics

Table 10-4: Example of Amazon's multi-level category tree, showing a subset from the two top-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

Table 10-5: Number of books with at least (K) second-level categories.

Summary statistics for a selection of network/mixing constructed variables are given in Table 10-6. Consistently with the findings of Oestreicher-Singer and Sundararajan (2008), we also see that, 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. The empirical results were robust to several definitions of clustering coefficient. Therefore, the results of all models are presented with CC_i as defined in section 10.3.2.

Variable	Network Proximity	CC_i^1	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
Max	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

Table 10-6: Summary statistics for a selection of constructed variables.

Breaking down category and author statistics (see Table 10-7), 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.

	Same Category Statistics			Same Author Statistics		
Distance	**All**	**Oprah Reviews**	**New York Times Reviews**	**All**	**Oprah Reviews**	**New York Times Reviews**
All neighbors (1..5)	43.9%	44.4%	43.7%	1.3%	1.8%	1.1%
	(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%
	(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%
	(1.5%)	(2.3%)	(2.0%)	(0.6%)	(1.0%)	(0.9%)
3	52.1%	54.6%	50.8%	0.9%	0.6%	1.1%
	(1.0%)	(1.7%)	(1.3%)	(0.2%)	(0.3%)	(0.3%)
4	43.9%	42.3%	44.6%	0.2%	0.2%	0.2%
	(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%
	(0.5%)	(0.9%)	(0.6%)	(0.0%)	(0.0%)	(0.0%)

* Standard errors between parentheses.

Table 10-7: Category & Author mixing statistics by distance from the reviewed book.

10.4. Appendix D – Sales Rank conversion to demand

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):

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

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 10-6 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.

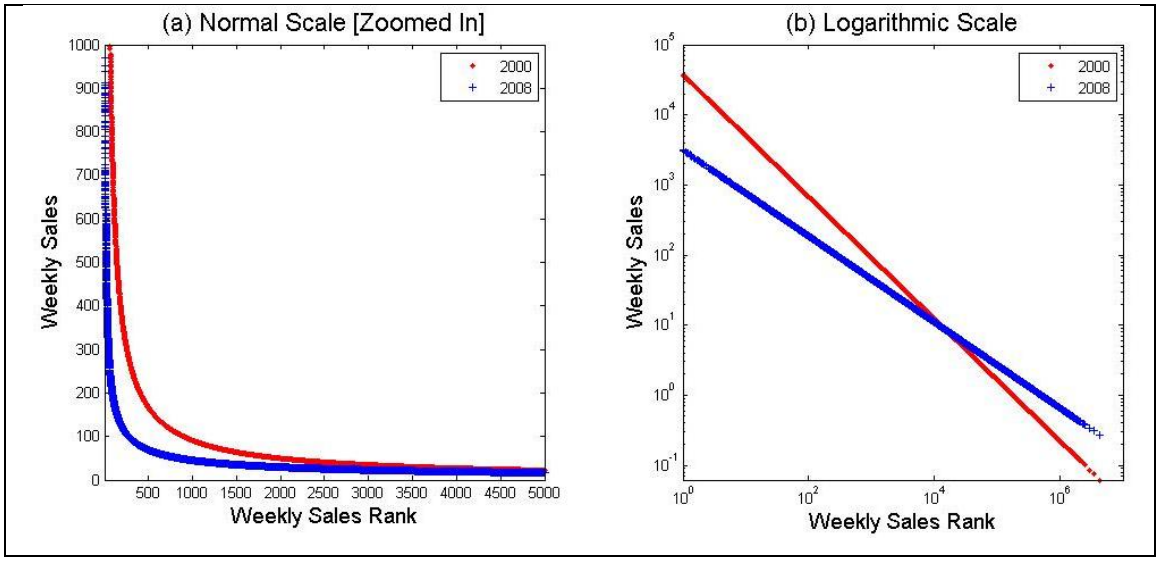

Figure 10-6: Sales Rank conversion to demand using 2008 estimation vs. 2000 estimations. The graphs present the conversion of the average Sales Rank of the books in our final sample to demand using the two estimations. The same data are presented in (a) normal scale (zoomed in to the range of 0 .. 5,000) and (b) logarithmic scale.

There are several other known issues regarding the use of converted demand estimations, especially for best-selling books (See 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 10-8 together with their demand-based counterparts (that is, demand estimated using the 2003 suggested estimates and the 2009 suggested estimates). 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 10-7).

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

Table 10-8: Summary statistics for a selection of constructed variables

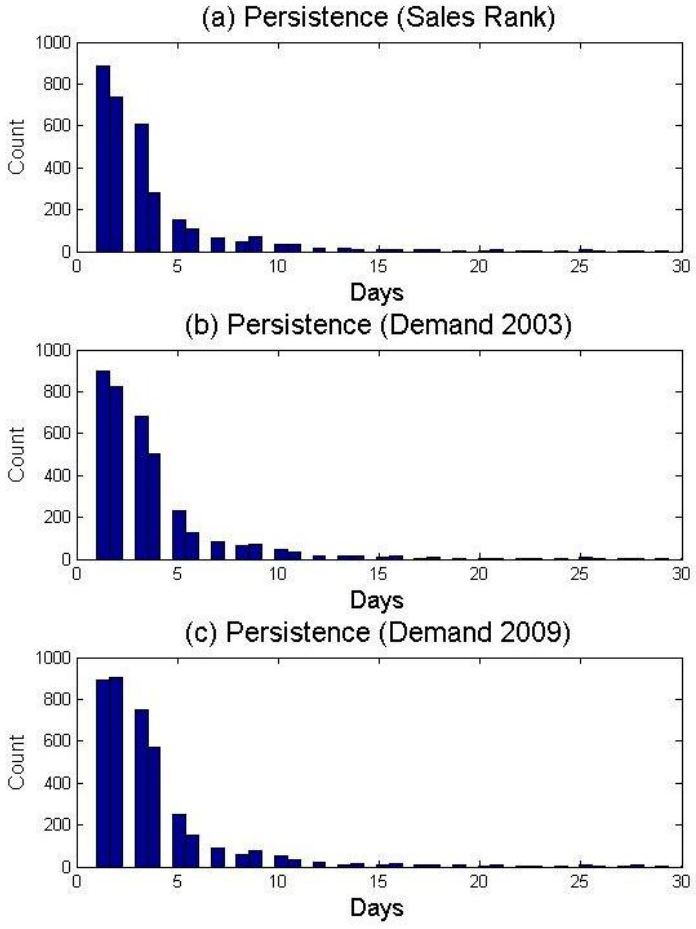

Figure 10-7: Distribution of persistence of the shock based on (a) Sales Rank, (b) Estimated demand using Brynjolfsson et al. (2003) and (c) Estimated demand using Brynjolfsson et al. (2009).