Is Oprah Contagious?

Identifying Demand Spillovers in Online Networks

Eyal Carmi (Tel-Aviv University),
Gal Oestreicher-Singer (Tel-Aviv University) and
Arun Sundararajan (NYU Stern School of Business)

Abstract

We study the spread of exogenous demand shocks generated by book reviews featured on the Oprah Winfrey TV show and published in the New York Times through the online co-purchase recommendation network on Amazon.com. We analyze the co-purchase recommendation network on Amazon.com to determine how such exogenous events might affect the demand for books that were not explicitly mentioned in a review but are located “close” to reviewed books in the network. Using a difference-in-differences matched-sample approach, we identify the extent of variation caused by membership in this network. Our results show that the demand shock diffuses to books that are up to five links away from the reviewed book, and that this diffused shock persists for a substantial number of days. However, the depth and the magnitude of diffusion vary widely across books at the same network distance from reviewed products. We also describe how product characteristics, assortative mixing and local network properties can explain this variation in the depth and persistence of contagion. Specifically, highly clustered local networks “trap” the diffused demand shocks, causing them to last longer and be more pronounced while restricting the distance of the shocks’ spread. Conversely, less clustered networks lead to wider contagions of lower magnitude and duration. We discuss the significance of these results and their implications for the design of networks of products as well as optimizing digital marketing spillovers.
1. Introduction and Research Questions

In September 2007, the book *Louder than Words* by Jenny McCarthy was featured on the *Oprah Winfrey Show*. Sales for the book spiked dramatically after this television appearance, and demand for copies increased over 200-fold on Amazon.com\(^1\) soon after the show aired. More interestingly, the economic impact of McCarthy’s appearance on Winfrey’s show was not restricted just to the specific book discussed during the show. Our data from Amazon.com also show that demand increased for books that were "recommended" on the reviewed book’s page, and even for books located several recommendation clicks away. The book *"Life Laughs: The Naked Truth about Motherhood, Marriage, and Moving On"* by Jenny McCarthy and the book *"Ten Things Every Child with Autism Wishes You Knew"* by Ellen Notbohm, located one click away on the co-purchase network, experienced a 16-fold and 6-fold increase in their demand, respectively\(^2\). The book *“The Kid-Friendly ADHD and Autism Cookbook”* by Pamela Compart, located two clicks away, experienced a 42-fold increase in its demand\(^3\). We documented statistically significant demand increases in response to this event for a number of other books located three and even four clicks away from the reviewed book (as illustrated in Figure 1).

While this is not an isolated example, evaluating the impact of such exogenous shocks on an interconnected set of products has been difficult in the past. However, the transition to online commerce has led to the emergence of new website structures that enable us to study potential cross-product spillover effects. One such structure is the visible product network, in which related products sold online are explicitly linked to one another. In most electronic commerce sites, each product (e.g. a book, video, or other content item) is featured on its own designated webpage, and each product page

\(^1\) Estimated demand increased from average sales of 110 copies per week to over 22,000 copies per week. Demand estimates computed using Amazon's reported SalesRank we collected from their XML feed and the demand estimation methods suggested by Goolsbee and Chevalier 2003 and Brynjolfsson et al. (2003).

\(^2\) Estimated demand increased from an average of 134 books per week to 798 books per week and an average of 12 books per week to 195 books per week, respectively.

\(^3\) Estimated demand increased from an average of 6 books per week to 242 books per week.
We also witness an increase in demand for books that are one click away from the reviewed book (B1 & B2); two clicks away from the reviewed book (C); and even 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 is linked by hyperlinks to other product pages, thus creating a visible network in which the products are the nodes. Perhaps the oldest example of this type of visible electronic product network is the co-purchase network of Amazon.com. Much like social networks enable the flow of “word-of-mouth” between individuals, product networks may represent conduits for the collective flow of attention between different products. This role of the network is of particular interest in cases where one product is subject to an exogenous demand shock, since it is now possible to assess how the impact of an event can spill over into a network of products that are related and visibly connected to the focal product.

Our goal in this paper is to investigate whether this kind of “contagion”, which is widely believed to occur between consumers in social networks, exists in networks of products, and to contribute towards understanding its characteristics and economic impact. We do this by analyzing the spillover of exogenous demand shocks through

---

4 Amazon.com provides hyperlinks that connect products, under the heading “Consumers who bought this item also bought …” (see also section 3). 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.
one product network, the Amazon.com co-purchase network for books. The specific demand shocks we study are caused by mass-media book reviews (on the *Oprah Winfrey* television show and in the *New York Times* “Sunday Book Review”). Of course, it is well known that such media appearances have a positive impact on the sales of the reviewed products. For example, a review on the *Oprah Winfrey Show* can transform a reviewed book into a bestseller literally overnight (as shown by Balogh (2008); Illouz (2003); Rooney (2005)); similarly, Deschatres and Sornette (2005) and Sorensen and Rasmussen (2004) have shown that book reviews published in the *New York Times* have a significant positive impact on the sales of the featured books. What is less understood is whether the impact of these reviews spreads.

We pose three research questions. First, do exogenous demand shocks diffuse through product networks? Second, what are the magnitude, the depth and the persistence of this “spillover effect” across products, and what kinds of products are most susceptible to “catching” the demand contagion? Finally, do the structural properties of the network influence the magnitude, depth and persistence of this demand spillover?

Our research design uses quasi-natural experiments, in which we investigate exogenous demand shocks created by reviews of focal products in the network in order to study the network’s influence 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 results show a significant average influence of the visible network on neighbors up to three links away from the reviewed book. Additionally, increases in demand can be observed as far away as the fourth and fifth neighbors. This effect 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.

While a considerable fraction of the books in the network of a reviewed book experience a statistically significant change in demand following a shock, not all books are affected. For example, among books reviewed on the *Oprah Winfrey Show*, we observed significant effects for 62% of the books’ first neighbors and 38% of their second neighbors. What causes these books to benefit from the spillover while not
others? Towards understanding this difference better, we analyze the variance in the reception of network neighbors to the spillovers. We find evidence of a strong influence of both assortative mixing (similarity) and local network characteristics. Both network proximity and local clustering around a book as well as cross-product similarities (such as sharing an author with a reviewed book) strongly influence the probability of being affected by the shock. 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 theory, we estimate an exponential hazard rate model that captures 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 away). The persistence of a shock to close neighbors is highly affected by the neighbors’ geodesic distance from the reviewed book (due to the significantly greater exposure of first neighbors), whereas spillover 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 positively increases the persistence of a shock, we find that local clustering around the source of the shock creates something analogous to a “fishing net”, trapping consumer attention in the network neighborhood close to the reviewed book. This structure increases the persistence of the shock among close neighbors and decreases the persistence of shocks to distant neighbors.

The rest of 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 depth of the spillover effect and the associated identification issues are discussed and analyzed in section 4. Following this, section 5 explores the product-level resistance to shocks, section 6 explores the product-level persistence of the shocks (how long they last) and section 7 discusses robustness tests.
We conclude, summarize key managerial insights and provide avenues for future research in section 8.

2. Related Work

While social networks have received much attention from researchers in a variety of fields such as business, economics, epidemiology and computer science, the limited attention given to research of networks of products is perhaps surprising. Prior 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), and a study of the network of news reports by Dellarocas et al. (2009). 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 network’s visibility. Our work contributes to this stream of research by analyzing cross-product spillover processes following exogenous shocks.

Somewhat related to this topic is the literature on multi-product diffusion in marketing (for example, 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 spillover on a large scale.

We also add to the literature on demand shifts following expert reviews or celebrity endorsement as well as marketing campaigns. 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-

\footnote{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).}

\footnote{The network of blogs can also be thought of as a social network.
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 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 spillover process across products, rather than on the increase 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.

We also draw selectively from the network analysis literature, in particular from studies on the effect of exogenous shocks in networks, and apply their insights to the product network, a type of large, real-world network. The majority of studies, especially in the context of epidemiology (Leger et al. (2007)), the spread of computer viruses (Lloyd and May 2001) and word-of-mouth/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 aim to distinguish actual spillovers from other causes of correlated outcomes between neighboring nodes 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 can 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 collected 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. Our use of 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-related 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 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.)

Currently Amazon.com provides a list of more than five items in each co-purchase network. Users are initially exposed to the top six due to screen-size limitations and can click to view the next six products. Our data set predates this change.
We collected information about book reviews that appeared on the Oprah Winfrey TV 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 by research assistants. ** how many books???
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, NYTimes.com publishes a section (“Sunday Book Review”) containing between 10 and 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.

Event networks

Each review event was linked to the corresponding temporal network and sales data from Amazon.com and went through a series of manual and automatic data 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 spillover of exogenous shocks across the network.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shock parameters</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Sales Rank (SR)</strong></td>
<td>A number associated with each product on Amazon.com that measures the product’s demand relative to the demand for other products. The best-selling product is therefore ranked 1, followed by 2, 3, and so on.</td>
</tr>
<tr>
<td><strong>SalesRankRatio (SRR)</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>SalesRankShock (SRS)</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Affected</strong></td>
<td>Binary variable that splits our sample into books that showed a significant reaction to the exogenous shock (post-event demand that is greater than one standard deviation from the pre-event average level) and those that did not.</td>
</tr>
<tr>
<td><strong>Persistence of the Shock (PSR)</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Product/Network Parameters</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Distance</strong></td>
<td>The number of links on the minimal path between the two books.</td>
</tr>
<tr>
<td><strong>Network Proximity</strong></td>
<td>Extends the simple distance variable by taking into consideration a weighted average of the lengths of all possible paths between a given book and the reviewed book.</td>
</tr>
<tr>
<td><strong>Local Clustering</strong></td>
<td>A measure of how close a node and its neighbors are to being a clique (Watts 2003; Watts and Strogatz 1998).</td>
</tr>
</tbody>
</table>

Table 3-1: Summary of variables

**Operationalization of variables**

We develop several measures to represent the magnitude of the shock, local network structure, and link properties. The key variables are summarized in Table 3.1. In addition, 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 the following: category similarity, author, price, binding type (hardcover, soft cover, spiral) and vintage (difference in years between year of review and release year). An extended description of these variables as well as summary statistics are provided in Appendix C.
4. Identification of Cross-Product Spillovers

One initial measure of the depth of exogenous shock diffusion in the online product network is obtained by evaluating the average Sales Rank of the reviewed book and of its neighbors in the three-day period preceding the review and comparing them to the average Sales Rank of the same products in the three-day period immediately following the review. We find that on average, neighbors up to three clicks away from a reviewed book experience a demand shock immediately following the event, and this shock is both economically and statistically significant. Following a review by Oprah Winfrey, we observe, on average, a 29.57-fold increase in the Sales Rank of first neighbors, a 3.65-fold increase in Sales Rank for second neighbors and a 2.07-fold increase in Sales Rank for third neighbors (see Table 5-1 for more details). Though the effect of the network beyond third-degree neighbors is not significant on average, increases in demand can be observed as far as the fourth and fifth neighbors.

These results provide initial evidence of the role of the product network in the spillover of exogenous shocks. In order to verify that the effects truly stem from membership in the visible network, however, it is necessary to control for potential self selection biases as well as for temporal changes in the demand patterns on the website. In what follows, we estimate a difference-in-differences model using a matched-sample approach, to control for potential sources of bias.

Difference-in-differences model

The difference-in-differences model is a common statistical method used to analyze data that is obtained from experimental designs that collects observations across 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)). The presence of a control group controls for global trends and seasonal factors that might alter overall demand patterns across books in general, as well as possibly for factors other than network membership that might account for correlations in the demand of the reviewed book and other books (more on this later). In our case, "treated" products are those that are part of the local network around a reviewed book (up to 5 links away).
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). Our specification includes four time windows each of which is three days long – one before the review (\( t = (-3, -1) \)), and three following the review (\( t = \{0, 2, 3, 5, 6, 8\} \)). We therefore estimate the following model:

\[
SRR_{it} = \beta_0 + \beta_{1i} \cdot \text{Period}_{it} + \beta_{2i} \cdot \text{Treatment}_{it} \\
+ \beta_{3i} \cdot \text{Period}_{it} \cdot \text{Treatment}_{it} + \beta_{4} \cdot x_{it} + \varphi_{i} + \varepsilon_{it},
\]

where \( t = \{0, 3\} \) corresponds to the four time windows (\( \{0, 2, 3, 5, 6, 8\} \)), \( \text{Period}_{it} \) are time-period fixed effects, \( \text{Treatment}_{it} \) represents the assignment to treatment groups, \( x_{it} \) are observed covariates (author, category, average Sales Rank, price, binding and rating), \( \varphi_{i} \) are unobserved covariates, and \( \varepsilon_{it} \) is a vector of either review source or review-level fixed effects. The coefficients whose significance is important are those of the interaction terms (that is, the \( \beta_{3i} \) coefficients).

**Choosing the appropriate control group**

We next turn to our construction of a control group. A natural first candidate might include randomly selected products that are not part of the reviewed book’s local network (and thus not “treated”). However, while this accounts for global changes over time, it will not control for factors other than network membership that cause linked products to have correlated demand patterns.

In our context, selection bias is introduced by two sources—selection of the product to be reviewed and selection of network neighbors to be featured on the product’s page. 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 Winfrey wishes to promote, as well as various marketing efforts exerted by publishers). It is therefore possible that the types of books Oprah Winfrey 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 the *Oprah Winfrey Show*).
We also verified that the category distributions of the books reviewed on *Oprah* and in the *New York Times* are very different. Acknowledging this limitation, we note that the extent of such an identification effect our setting is quite likely to be minimal. 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; put differently it seems quite unlikely that the producers of *Oprah* or the publishers of the *New York Times* have any influence in selecting the network neighbors of reviewed books\(^8\), which are instead chosen by an algorithm on Amazon.com.

The second source of selection bias is introduced by the recommendation algorithm’s 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), regardless of the existence of a visible hyperlink. For example, books written by the same author are more likely to experience an increase in sales following the review and are also highly likely to be network neighbors. Thus, the increase in demand for such a neighbor may be mistakenly attributed to the presence of the visible link. Before concluding that the spillover was in fact on account of the network, we need to address the endogeneity caused by this source of selection bias.

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). Possible choices for appropriate control groups that might mitigate this 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. 1998; Rosenbaum and Rubin 1983) which have been used in the past to control for selection bias in networks. See, for example, Aral et al. (2009); Hill et al. (2006); Oestreicher-Singer and Zalmanson (2010)\(^9\). Thus, instead of grouping products on the basis of

\(^8\) 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 determined by the recommendation algorithm.

\(^9\) 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
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). We computed the propensity score using observed book characteristics (author, category, average Sales Rank, price, binding and rating). For each network neighbor \( b_{i,k} \) of a reviewed book \( b_{i,k} \), given event \( k \), we assigned a matched book (to the matched sample such that the probability of \( b_{i,k} \) to be a network neighbor was equal (or close enough) to that of \( b_{i,k} \), on the basis of a specific propensity score. Ideally, one would want the matched book \( b_{i,k} \) to be as similar as possible to the network neighbor \( b_{i,k} \), and in general for the distribution of all observed properties of \( \{b_{i,k}\} \) and \( \{b_{i,k}\} \) to be identical so that the only difference is the treatment.

**Model estimation**

We estimated a 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 which measures the change in demand relative to the 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 \( (\bar{D}_{GR}) \) and its standard errors are shown in Table 1.

The coefficients for the difference-in-differences estimators are 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 results of prior 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.
and recent studies (though conducted in other domains) on real-world networks with a high level of clustering (Fowler and Christakis 2010; Friedkin 1983).

### Difference in Differences estimates using OLS Regression of by distance

<table>
<thead>
<tr>
<th>Distance</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{31}$</td>
<td>39.87***</td>
<td>3.62***</td>
<td>0.42**</td>
<td>-0.02</td>
<td>-0.06</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>(7.27)</td>
<td>(1.16)</td>
<td>(0.17)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>$\beta_{32}$</td>
<td>19.35***</td>
<td>2.04***</td>
<td>0.41***</td>
<td>0.17***</td>
<td>0.06*</td>
<td>0.07***</td>
</tr>
<tr>
<td></td>
<td>(3.89)</td>
<td>(0.62)</td>
<td>(0.13)</td>
<td>(0.06)</td>
<td>(0.03)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\beta_{33}$</td>
<td>8.54***</td>
<td>1.16***</td>
<td>0.32**</td>
<td>0.23***</td>
<td>0.19</td>
<td>0.11***</td>
</tr>
<tr>
<td></td>
<td>(1.58)</td>
<td>(0.37)</td>
<td>(0.12)</td>
<td>(0.07)</td>
<td>(0.12)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.04</td>
<td>1.09***</td>
<td>1.06***</td>
<td>1.10***</td>
<td>1.09***</td>
<td>1.09***</td>
</tr>
<tr>
<td></td>
<td>(1.54)</td>
<td>(0.26)</td>
<td>(0.05)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Observations</td>
<td>624</td>
<td>2880</td>
<td>7504</td>
<td>16269</td>
<td>34124</td>
<td>70132</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.58</td>
<td>0.11</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>F</td>
<td>7.79</td>
<td>6.24</td>
<td>7.12</td>
<td>14.83</td>
<td>22.67</td>
<td>31.28</td>
</tr>
</tbody>
</table>

* 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.

For third neighbors, we see a positive and significant coefficient only for the later time periods and have no evidence that there is a significant shock on average in the first three days after the review event. This might suggest a lag in the impact of the shock for more distant neighbors. Similar positive coefficients are shown for fourth and fifth neighbors, though they are not significant.

### 5. Patterns of Demand Spillover

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.
### Persistence and Shock Based on Sales Rank

| Source | Distance | (a) All Books | | | (b) Affected Books | | | (c) Books Not Affected | | |
|--------|----------|---------------|---|---|---------------|---|---|---------------|---|
|        |          | Average # | PSR | SRS | Average # | % | PSR | 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.

We estimate a binary Logit model to 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 the effect on others) by directly accounting for the different components, including local and global network structure and product similarities ( assortative mixing). Specifically, we estimate 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).

\[
\ln \left( \frac{P(affected)}{1-P(affected)} \right) = \alpha_0 + \sum_{j=1}^{I} \gamma_j \text{LOCAL}_{ij} + \sum_{k=1}^{K} \beta_k \text{GLOBAL}_{ik} + \sum_{l=1}^{L} \delta_l \text{MIXING}_{il} + \sum_{m=1}^{M} \zeta_m \text{CONTROL}_{im}
\]

Analysis of the co-purchase global network structure and of the event local sub-network structures shows that there are large variations in local network structure, whereas the global network structure remains stable. We therefore include the two types of structural network properties—global and local—while focusing on the local...
network structures. We use the indegree centrality as a global measure of centrality\textsuperscript{10}. Our local structure measures build on the social network analysis literature, 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, suggesting that multiple sources of activation are required in order to spread complex contagions. Hence, we focus on three local network structure variables—distance, network proximity and local clustering.

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

* 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.

Our measures for product similarities include: sharing the same author, category, vintage, binding and price range. We also control for other drivers of shock

\textsuperscript{10} We also experimented with other types of centralities such as PageRank and eigenvector centrality, and results were robust.

\textsuperscript{11} The average Sales Rank was divided by 100,000 when entered into the logistic regression for readability reasons of the coefficient.

\textsuperscript{12} Defined only for Oprah Winfrey's reviews.
susceptibility, including the day of week, changes in price, the existence of a discount, and quality (as measured by the number of consumers' reviews and average ratings).

Model estimation

We estimate the fixed effects models (day of week and review event) by maximizing the log likelihood. 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. The more clustered the network around a book, the greater the odds of the book being affected by the shock. For example, each additional edge between the first neighbors of a book results in a 1.2% increase in the odds ratio of being affected by the shock. We also find that a higher indegree is associated with lower odds of being affected by the shock, which is consistent with a positive coefficient (odds ratio > 1) on the average Sales Rank of the book. This means that books at the tail are more likely to be affected by the shock.

Network proximity to a reviewed book 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 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, we 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, as first neighbors appear on the same page as the reviewed book (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.

<table>
<thead>
<tr>
<th>Odds Ratio Estimates: Logit Model</th>
<th>Distance Only</th>
<th>Network Only</th>
<th>Network and Mixing</th>
<th>Full Model with day-of-week FE</th>
<th>Full Model with day-of-week &amp; Event FE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>1.04*</td>
<td>1.02</td>
<td>1.02</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Network Proximity</td>
<td>35.92***</td>
<td>35.62***</td>
<td>21.17***</td>
<td>21.55***</td>
<td>20.69***</td>
</tr>
<tr>
<td></td>
<td>(15.40)</td>
<td>(15.25)</td>
<td>(10.16)</td>
<td>(10.38)</td>
<td>(10.01)</td>
</tr>
<tr>
<td>Local Clustering</td>
<td>1.26**</td>
<td>1.21*</td>
<td>1.43***</td>
<td>1.44***</td>
<td>1.48***</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.12)</td>
<td>(0.15)</td>
<td>(0.15)</td>
<td>(0.21)</td>
</tr>
<tr>
<td>In Degree</td>
<td>0.99***</td>
<td>0.99***</td>
<td>0.99***</td>
<td>0.99***</td>
<td>0.99***</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Same Author</td>
<td>2.11***</td>
<td>2.05***</td>
<td>2.08***</td>
<td>2.12**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.40)</td>
<td>(0.41)</td>
<td>(0.42)</td>
<td>(0.66)</td>
<td></td>
</tr>
<tr>
<td>Same Category</td>
<td>1.10*</td>
<td>1.03</td>
<td>1.04</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td>Same Vintage</td>
<td>1.01***</td>
<td>1.01**</td>
<td>1.01**</td>
<td>1.01*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td></td>
</tr>
<tr>
<td>Same Price</td>
<td>0.89**</td>
<td>1.01</td>
<td>0.99</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td>Same Binding</td>
<td>0.93*</td>
<td>0.93*</td>
<td>0.94*</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td></td>
</tr>
<tr>
<td>Average Sales Rank</td>
<td>1.16***</td>
<td>1.16***</td>
<td>1.15***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discount Rate</td>
<td>1.35**</td>
<td>1.35**</td>
<td>1.38**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.19)</td>
<td>(0.19)</td>
<td>(0.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Re-Run</td>
<td>0.83***</td>
<td>0.83***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.05)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Reviews</td>
<td>1.02</td>
<td>1.02</td>
<td>1.03*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rating</td>
<td>1.23*</td>
<td>1.22*</td>
<td>1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.13)</td>
<td>(0.15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review Source</td>
<td>1.02</td>
<td>1.05</td>
<td>1.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>19,586</td>
<td>19,574</td>
<td>17,547</td>
<td>16,969</td>
<td>16,969</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-12469.38</td>
<td>-12405.14</td>
<td>-11077.38</td>
<td>-10607.28</td>
<td>-10602.64</td>
</tr>
<tr>
<td>LR Chi Square</td>
<td>97.33</td>
<td>185.17</td>
<td>223.97</td>
<td>404.36</td>
<td>411.145</td>
</tr>
<tr>
<td>Event Fixed Effects</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Day of Week Fixed Effects</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* 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 spillover process.

6. Post-shock Persistence

We witness large variance in the persistence of the increase in demand (see Table 5-1), even among neighbors of the same distance from the source of a shock. Therefore, in this section we study what explains variation in persistence of the spillover effect. 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.

Duration model of shock persistence

To model the persistence of the spillover effects we follow duration model theory and use a hazard-rate model 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).

Figure 6—1 presents the non-parametric Kaplan-Meier maximum likelihood estimation of the survival function and provides an important insight: close (first and second) neighbors seemed to have a decay pattern that is distinct from that of the more distant (third and fourth) neighbors (this observation was also validated by the log rank test with p-value < 1%). Following this, we extend 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. Left: The estimations for the reviewed books (distance=0) and their network neighbors (distance=1..4); Right: 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):

\[ h(t) = \exp \left\{ \alpha_0 + \sum_{i=1}^{I} \beta_{i,\text{LOCAL}} t + \sum_{k=1}^{K} \gamma_{k,\text{GLOBAL}} t + \sum_{l=1}^{L} \psi_{l,\text{MIXING}} t + \sum_{m=1}^{M} \sigma_{m,\text{CONTROL}} t \right\} \]

The results of the model estimations are presented in Table 6-1. The estimation strategy was 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 considered a Weibull distribution (see column (e) of Table 6-1) for the

<table>
<thead>
<tr>
<th>Parameter Estimates: Exponential Duration Model (Hazard Rates)</th>
<th>(a) Distance Only</th>
<th>(b) Network</th>
<th>(c) Network and Mixing</th>
<th>(d) Full Model (Exponential)</th>
<th>(e) Full Model (Weibull)</th>
<th>(f) Full Model (Cox)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distance</strong></td>
<td>0.99 (0.01)</td>
<td>0.99 (0.01)</td>
<td>0.99 (0.02)</td>
<td>0.99 (0.01)</td>
<td>0.99 (0.01)</td>
<td>0.99 (0.01)</td>
</tr>
<tr>
<td><strong>Network Proximity</strong></td>
<td>0.07** (0.09)</td>
<td>0.06** (0.07)</td>
<td>0.06* (0.09)</td>
<td>0.06* (0.09)</td>
<td>0.0678* (0.11)</td>
<td>0.16* (0.15)</td>
</tr>
<tr>
<td><strong>Shock Local Clustering</strong></td>
<td>1.32**</td>
<td>1.34***</td>
<td>1.33***</td>
<td>1.34***</td>
<td>1.07***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.00)</td>
<td>(0.12)</td>
<td>(0.11)</td>
<td>(0.01)</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>In Degree</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01***</td>
<td>1.01***</td>
<td>1.01***</td>
<td></td>
</tr>
<tr>
<td>Same Author</td>
<td>0.84***</td>
<td>0.79***</td>
<td>0.79***</td>
<td>0.81***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same Category</td>
<td>1.13***</td>
<td>1.11***</td>
<td>1.11***</td>
<td>1.04*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same Vintage</td>
<td>0.99*</td>
<td>0.99**</td>
<td>0.99**</td>
<td>0.99***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same Price</td>
<td>1.15*</td>
<td>1.17***</td>
<td>1.17***</td>
<td>1.15***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same Binding</td>
<td>0.91**</td>
<td>0.91***</td>
<td>0.91***</td>
<td>0.96*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Reviews</td>
<td>0.93***</td>
<td>0.92***</td>
<td>0.97**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rating</td>
<td>0.88</td>
<td>0.88</td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Re-Run</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of observations</td>
<td>6605</td>
<td>6605</td>
<td>5909</td>
<td>5711</td>
<td>5711</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>19328.6</td>
<td>19313.6</td>
<td>17267.2</td>
<td>16572.5</td>
<td>16572.2</td>
<td></td>
</tr>
<tr>
<td>Log Pseudolikelihood</td>
<td>-9663.3</td>
<td>-9655.8</td>
<td>-8632.6</td>
<td>-8285.3</td>
<td>-8285.1</td>
<td></td>
</tr>
</tbody>
</table>

* 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).

basic survival function, and a semi-parametric Cox-proportional hazard rate model (see column (f)). Following the Kaplan-Meier estimations, we also repeated the estimation of the model for different neighbors, grouped by their minimal distance from the reviewed book (see Table 6-2). Naturally, only books that were affected were included in this estimation, leaving us with 6,605 data 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 (4 books in total; these books were labeled as censored after 60 days); 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 strongly influenced by the total number of paths connecting the node to the reviewed book.

Interestingly, when we break down the analysis according to 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 network 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 significantly 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 determining the duration of the shock to the set of products (network neighbors) inside the cluster (close neighbors) and to the products outside the cluster (distant neighbors).13

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. This means that when the local area of the network around the shock is more clustered, the shock is

13 Note that we also control for the focal book's local clustering coefficient (Book Local Clustering), which does not yield significant results.
likely to end sooner (persistence less). Interestingly, the signs of the coefficients for close (first and second) neighbors and for distant (third, fourth and fifth) neighbors are in opposite directions (Table 6-2). For example, an additional edge between first neighbors of a 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 in 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 its 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).

**Assortative mixing**

All assortative mixing variables are statistically significant, 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.
Table 6-2: Table presents the estimation results of the exponential hazard rate model for several test groups based on (minimal) distance.

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, does not 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), number of consumer reviews and ratings were found to be statistically significant and reduce the hazard rate, i.e., increase the persistence of the shock.

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; see Appendix D for details). For robustness, we repeated the analysis presented in this paper twice - 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.

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

$\omega_{\text{affected}} = \begin{cases} \omega_{\text{peak}} - \omega_{\text{SR}} & \text{if } \omega_{\text{peak}} > \omega_{\text{SR}} \\ 0 & \text{if } \omega_{\text{peak}} < \omega_{\text{SR}} \end{cases}$

In this framework, the variable Affected presented above can be viewed as $1 - \text{affected}$. 

27
For robustness, we repeated the analysis presented in this paper replacing “Affected” with “2–affected”.

In the logistic regression model, use of the variable 2–affected rather than 1–affected resulted in an increase in the effect of assortative mixing (having the same author) and pricing policy (discount rate) on the shock, while local clustering was no longer significant. Nevertheless, network proximity was statistically significant, and the magnitude of its coefficient was high relative to the coefficients of the other variables. These results suggest the importance of the joint analysis of network structure and assortative mixing.

In the hazard rate model, estimations using 2–affected showed an increase in the levels of significance of both distance and network proximity. Also, the fishing net effect and the effect of local clustering seemed to become more dominant (showing both greater significance and higher magnitude of coefficients).

The full sets of results of those estimates are very similar to the results presented here and are available upon request.

8. Concluding Remarks

The presence of hyperlinked product recommendation networks is one of the principal differences between the online and traditional channels of commerce, and is yet surprisingly understudied. 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 Book Review” section of the New York Times. We study the impact and spillover 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 observed a strikingly high level of spillover 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 spillover in networks and the potential extent of spillover (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 found 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 important documentation of the magnitude and persistence of spillover 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). Put together, our results suggest that, on average, the increase in sales of a reviewed book constitutes only 67% of the total effect of the review event. These findings have significant managerial implications, for design as well as for marketing and strategy. From the retailer's perspective, they may carry implications as to inventory and pricing decisions preceding media events. Publishers, who sell multiple books, may use the product network structure in order to better allocate advertising budgets to maximize their total revenue. More broadly, our work may also carry implications to marketers placing ads on those product pages. After all, if a product's page is about to
experience a shock in amount of attention directed to it, ad placement and pricing should be planned accordingly.

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.
Acknowledgments

The authors thank Michael Ward for an insightful early discussion, Barak Libai and Eitan Muller for many helpful discussions, and seminar participants at Cambridge University, Carnegie Mellon University, Microsoft Research, New York University, the Statistical Challenges in E-Commerce Research Symposium, the Marketing Science Institute Conference, the INFORMS Conference on Information Systems and Technology, the Workshop on Information in Networks, and the International Conference on Information Systems. Financial support from Google, the WPP Group, the Israeli Science Foundation, NEC Corporation and the NET Institute is gratefully acknowledged.
9. References


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


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 last-in-first-out (LIFO) stack. If the algorithm finds it is on the page of a product that it has visited already, it "backtracks" and returns to the most recent product for which exploration was not exhausted. The program terminates when the entire connected component of the graph is collected.

For example, in the graph in Figure A-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 4 and 5 hours. The script is run each day at midnight.
**Figure A-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.

**Appendix B – Network Statistics**

**Co-purchase networks**

Table B-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\(^{14}\); 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

\[ \frac{5n}{2(n-1)} = \frac{5}{n-1} \approx 1.8 \times 10^8 \]

\(^{14}\) 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 \( \frac{5n}{2(n-1)} \).
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 B-1 for degree distribution and distribution of betweenness centrality on a sample daily network).

<table>
<thead>
<tr>
<th>Variable</th>
<th># Nodes</th>
<th># Edges</th>
<th>Average In Degree</th>
<th>Fraction of reciprocal links</th>
<th>Average Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>274,179</td>
<td>1,246,986</td>
<td>4.7</td>
<td>55%</td>
<td>0.39</td>
</tr>
<tr>
<td>Median</td>
<td>273,255</td>
<td>1,230,800</td>
<td>4.7</td>
<td>56%</td>
<td>0.39</td>
</tr>
<tr>
<td>Maximum</td>
<td>368,760</td>
<td>1,657,400</td>
<td>4.8</td>
<td>56%</td>
<td>0.40</td>
</tr>
<tr>
<td>Minimum</td>
<td>120,620</td>
<td>362,580</td>
<td>3.5</td>
<td>43%</td>
<td>0.27</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>40,547</td>
<td>182,999</td>
<td>0.1</td>
<td>2%</td>
<td>0.01</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.37</td>
<td>-0.71</td>
<td>-5.3</td>
<td>-4.56</td>
<td>-6.46</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.58</td>
<td>4.43</td>
<td>42.4</td>
<td>26.95</td>
<td>55.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Jarque-Bera</th>
<th>Probability</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td># Nodes</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td>328</td>
<td></td>
</tr>
</tbody>
</table>

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

Figure B-1: 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.

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 5\textsuperscript{th} 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 B-2). 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 B-2 presents basic network statistics on the sub-networks up to a distance of 5 links away (the 5\textsuperscript{th} network neighbor of the reviewed book). The relatively high variance in the average clustering coefficient of these networks (as illustrated in Figure B-3) shows that they are significantly different from each other, which may be reflected in the way exogenous shocks diffuse through the network.

![Figure B-2: Reviewed books time series data, classified into two categories: Exogenous Shocks (top); Endogenous & Multiple Shocks (bottom).](image)
Table B-2: Network statistics across the sub-networks up to the 5th network neighbor for each of the reviewed books’ events.

<table>
<thead>
<tr>
<th>Variable</th>
<th># Nodes</th>
<th># Edges</th>
<th>Average In Degree</th>
<th>Fraction of reciprocal links</th>
<th>Average Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>249</td>
<td>558</td>
<td>3.6</td>
<td>48%</td>
<td>0.33</td>
</tr>
<tr>
<td>Median</td>
<td>231</td>
<td>534</td>
<td>3.6</td>
<td>47%</td>
<td>0.31</td>
</tr>
<tr>
<td>Maximum</td>
<td>813</td>
<td>1524</td>
<td>5.0</td>
<td>80%</td>
<td>0.84</td>
</tr>
<tr>
<td>Minimum</td>
<td>8</td>
<td>40</td>
<td>3.0</td>
<td>39%</td>
<td>0.17</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>159</td>
<td>313</td>
<td>0.4</td>
<td>6%</td>
<td>0.10</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.72</td>
<td>0.46</td>
<td>1.1</td>
<td>1.62</td>
<td>1.98</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.33</td>
<td>2.73</td>
<td>4.8</td>
<td>7.77</td>
<td>9.85</td>
</tr>
<tr>
<td>Jarque-Bera</td>
<td>11.22</td>
<td>4.74</td>
<td>39.7</td>
<td>170.23</td>
<td>320.89</td>
</tr>
<tr>
<td>Probability</td>
<td>0.00</td>
<td>0.09</td>
<td>0.0</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Observations</td>
<td>123</td>
<td>123</td>
<td>123</td>
<td>123</td>
<td>123</td>
</tr>
</tbody>
</table>

Figure B-3: Examples for sub-networks with increasing clustering coefficients.

Appendix C - Detailed Description of Constructed Variables

Shock parameters

Sales Rank (SR) is a number associated with each product on Amazon.com that measures its demand relative to other products\(^\text{15}\). The best-selling product is therefore ranked 1, followed by 2, 3, and so on. The Sales Rank of each product on

\(^\text{15}\) Amazon does not disclose the actual sales information.
Amazon is updated several times a day, and prior research has shown that there are intra-day fluctuations; therefore, we use a 24-hour 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** \(\overline{SR}_t\) - To assess the magnitude of response to the exogenous shock we follow a common procedure in extreme event studies (Chollette 2009) and compute the *pre-event average Sales Rank* \(\overline{SR}_t\) 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\(^ {16}\).

**SalesRankRatio** (SRR) measures the magnitude of the event at time \(t\) and is defined as: \(SRR_{i,t} = 1 - \frac{SR_{i,t}}{\overline{SR}_i}\), where \(SR_{i,t}\) is the average daily Sales Rank of book \(i\) on day \(t\). 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 increase relative to the pre-event average. Formally: \(SRS_i = 1 - \frac{SR_{peak,i}}{\overline{SR}_i}\), where \(SR_{peak,i}\) is the peak Sales Rank reached by the book in the 72-hour interval immediately following the review\(^ {18}\). SRS can therefore also be defined as \(\max\{SRR\}_{t=0}^{t=2}\).

---

\(^{16}\) 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.

\(^{17}\) 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.

\(^{18}\) 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.
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 $\sigma_{SR_i}$ of each book $i$ and compare it to the SalesRank peak of that book:

$$\text{Affected}_i = \begin{cases} 1, & \overline{SR}_{\text{peak}} < \overline{SR}_i - \sigma_{SR_i} \\ 0, & \end{cases}$$

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 ($\text{Affected}_i = 1$) 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.

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 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:

$$\text{Proximity}_{i} = \sum_{k=2}^{d} \frac{N_{ki}}{5^k};$$

where $N_{ki}$ is the number of times book $i$ is a $k$-neighbor
of the reviewed book\textsuperscript{19} and $d=5$\textsuperscript{20}. 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\textsuperscript{21}. (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:

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

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 that these types of networks are protected against the spread of viruses (Eguiluz and Klemm 2002). Nevertheless, Eguiluz 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.

\textsuperscript{19} Recall that each book in our network has five outgoing links, hence the choice of denominator.

\textsuperscript{20} 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.

\textsuperscript{21} We did not observe any significant change in results by changing this assumption.
In the context of product networks, it is interesting to study even a local spillover 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).

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).

Summary statistics

Summary statistics for a selection of shock constructed variables are given in Table C-1. 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 C-2 for an example and Table C-3 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 C-1: 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 C-1) shows a considerable amount of variation across books.

![Distribution of Persistence](image)

Figure C-1: 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.

<table>
<thead>
<tr>
<th>Level 1 Category</th>
<th>Level 2 Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children's Books</td>
<td>People &amp; Places</td>
</tr>
<tr>
<td>Children's Books</td>
<td>Science, Nature &amp; How It Works</td>
</tr>
<tr>
<td>Children's Books</td>
<td>Animals</td>
</tr>
<tr>
<td>Children's Books</td>
<td>Educational</td>
</tr>
<tr>
<td>Children's Books</td>
<td>Holidays &amp; Festivals</td>
</tr>
<tr>
<td>Literature &amp; Fiction</td>
<td>History &amp; Criticism</td>
</tr>
<tr>
<td>Literature &amp; Fiction</td>
<td>Poetry</td>
</tr>
<tr>
<td>Literature &amp; Fiction</td>
<td>Comic</td>
</tr>
<tr>
<td>Literature &amp; Fiction</td>
<td>Drama</td>
</tr>
<tr>
<td>Nonfiction</td>
<td>Education</td>
</tr>
<tr>
<td>Nonfiction</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>Nonfiction</td>
<td>Politics</td>
</tr>
</tbody>
</table>

Table C-2: Example of Amazon’s multi-level category tree, showing a subset from the two top-level categories.

<table>
<thead>
<tr>
<th>Number of categories (K)</th>
<th>Number of books with at least K categories</th>
<th>Number of categories (K)</th>
<th>Number of books with at least K categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>706,169</td>
<td>11</td>
<td>4,521</td>
</tr>
<tr>
<td>2</td>
<td>637,558</td>
<td>12</td>
<td>1,927</td>
</tr>
<tr>
<td>3</td>
<td>542,354</td>
<td>13</td>
<td>823</td>
</tr>
<tr>
<td>4</td>
<td>403,499</td>
<td>14</td>
<td>327</td>
</tr>
<tr>
<td>5</td>
<td>267,152</td>
<td>15</td>
<td>131</td>
</tr>
<tr>
<td>6</td>
<td>158,153</td>
<td>16</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>86,269</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>44,558</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>21,603</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>10,064</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Table C-3: Number of books with at least (K) second-level categories.

Summary statistics for a selection of network/mixing constructed variables defined in Appendix C are given in Table C-4. Consistently with the findings of Oestreich-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 Appendix C.
### Table C-4: Summary statistics for a selection of constructed variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Network Proximity</th>
<th>Same Author</th>
<th>Same Category</th>
<th>Same Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.018</td>
<td>0.01</td>
<td>0.44</td>
<td>0.84</td>
</tr>
<tr>
<td>Median</td>
<td>0.001</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Max</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.08</td>
<td>0.17</td>
<td>0.5</td>
<td>0.37</td>
</tr>
<tr>
<td>Skewness</td>
<td>9.04</td>
<td>-0.02</td>
<td>0.24</td>
<td>-1.82</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>101.38</td>
<td>3.29</td>
<td>72.54</td>
<td>1.06</td>
</tr>
<tr>
<td>Obs.</td>
<td>19669</td>
<td>19669</td>
<td>19669</td>
<td>19669</td>
</tr>
</tbody>
</table>

Breaking down category and author statistics (see Table C-5), one can see that the percentage of books in the same category as the reviewed book drops as the distance from the reviewed book increases. An even sharper drop is seen (as expected) for books with the same author: The percentage of books with the same author among first neighbors is significantly higher.

### Table C-5: Category & Author mixing statistics by distance from the reviewed book.

<table>
<thead>
<tr>
<th>Distance</th>
<th>All</th>
<th>Oprah Reviews</th>
<th>New York Times Reviews</th>
<th>All</th>
<th>Oprah Reviews</th>
<th>New York Times Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>43.9%</td>
<td>44.4%</td>
<td>43.7%</td>
<td>1.3%</td>
<td>1.8%</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td>(0.4%)</td>
<td>(0.6%)</td>
<td>(0.4%)</td>
<td>(0.1%)</td>
<td>(0.2%)</td>
<td>(0.1%)</td>
</tr>
<tr>
<td></td>
<td>76.6%</td>
<td>80.4%</td>
<td>73.1%</td>
<td>20.7%</td>
<td>22.5%</td>
<td>19.3%</td>
</tr>
<tr>
<td></td>
<td>(2.1%)</td>
<td>(2.9%)</td>
<td>(3.0%)</td>
<td>(2.0%)</td>
<td>(3.1%)</td>
<td>(2.7%)</td>
</tr>
<tr>
<td></td>
<td>60.5%</td>
<td>63.6%</td>
<td>58.4%</td>
<td>4.6%</td>
<td>4.3%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>(1.5%)</td>
<td>(2.3%)</td>
<td>(2.0%)</td>
<td>(0.6%)</td>
<td>(1.0%)</td>
<td>(0.9%)</td>
</tr>
<tr>
<td></td>
<td>52.1%</td>
<td>54.6%</td>
<td>50.8%</td>
<td>0.9%</td>
<td>0.6%</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td>(1.0%)</td>
<td>(1.7%)</td>
<td>(1.3%)</td>
<td>(0.2%)</td>
<td>(0.3%)</td>
<td>(0.3%)</td>
</tr>
<tr>
<td></td>
<td>43.9%</td>
<td>42.3%</td>
<td>44.6%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td>(0.7%)</td>
<td>(1.2%)</td>
<td>(0.8%)</td>
<td>(0.1%)</td>
<td>(0.1%)</td>
<td>(0.1%)</td>
</tr>
<tr>
<td></td>
<td>38.6%</td>
<td>37.0%</td>
<td>39.3%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td>(0.5%)</td>
<td>(0.9%)</td>
<td>(0.6%)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
</tr>
</tbody>
</table>

* Standard errors between parentheses.

### Appendix D – Sales Rank conversion to demand

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

$$\log(\text{Demand}_{t,i}) = \alpha + \beta \log(\text{SalesRank}_{i,t})$$
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 D-1 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 D-1: 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 D-1 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 D-2).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Max</th>
<th>Min</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Sales Rank</td>
<td>126,759</td>
<td>46,569</td>
<td>4,340,296</td>
<td>10</td>
<td>194,163</td>
<td>3.67</td>
<td>32.83</td>
<td>19669</td>
</tr>
<tr>
<td>Average Demand (2003)</td>
<td>116.33</td>
<td>4.34</td>
<td>27404.55</td>
<td>0.06</td>
<td>572.38</td>
<td>18.66</td>
<td>669.32</td>
<td>19669</td>
</tr>
<tr>
<td>Average Demand (2009)</td>
<td>30.79</td>
<td>5.17</td>
<td>2360.51</td>
<td>0.27</td>
<td>86.83</td>
<td>7.12</td>
<td>95.34</td>
<td>19669</td>
</tr>
<tr>
<td>Persistence (Sales Rank)</td>
<td>1.476</td>
<td>0.000</td>
<td>64.000</td>
<td>0.000</td>
<td>4.486</td>
<td>8.14</td>
<td>92.05</td>
<td>19669</td>
</tr>
</tbody>
</table>

49
Table D-1: Summary statistics for a selection of constructed variables

<table>
<thead>
<tr>
<th>Persistence (Demand 2003)</th>
<th>1.332</th>
<th>0.000</th>
<th>64.000</th>
<th>0.000</th>
<th>4.045</th>
<th>8.84</th>
<th>111.57</th>
<th>19669</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence (Demand 2009)</td>
<td>1.365</td>
<td>0.000</td>
<td>64.000</td>
<td>0.000</td>
<td>4.093</td>
<td>8.68</td>
<td>107.93</td>
<td>19669</td>
</tr>
</tbody>
</table>

Figure D-2: 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).