PREDICTION IN ECONOMIC NETWORKS: USING THE IMPLICIT GESTALT IN PRODUCT GRAPHS

Completed Research Paper

Vasant Dhar
Stern School of Business,
New York University
44 West Fourth Street
New York, NY 10012
vdhar@stern.nyu.edu

Tomer Geva
Google Israel
23 Menachem Begin Road
Tel-Aviv, 66183
Israel
tomergev@post.tau.ac.il

Gal Oestreicher-Singer
Recanati Business School
Tel Aviv University
Tel Aviv 69978
Israel
galos@post.tau.ac.il

Arun Sundararajan
Stern School of Business
New York University
44 West Fourth Street
New York, NY 10012
asundara@stern.nyu.edu

Abstract

We define an economic network as a linked set of products, where links are created by realizations of shared outcomes between entities. We analyze the predictive information contained in an increasingly prevalent type of economic network, a "product network" that links the landing pages of goods frequently co-purchased on e-commerce websites. Our data include one million books in 400 categories spanning two years, with over 70 million observations. Using autoregressive and neural-network models, we demonstrate that combining historical demand of a product with that of its neighbors improves demand predictions even as the network changes over time. Furthermore, network properties such as clustering and centrality contribute significantly to predictive accuracy. To our knowledge, this is the first large-scale study showing that a non-static product network contains useful distributed information for demand prediction, and that this information is more effectively exploited by integrating composite structural network properties into one's predictive models.

Keywords: Product networks, prediction, co-purchase network, predictive modeling, neural networks, autoregressive models, network-based prediction, PageRank.

Introduction

The growth in commercial and social interaction online has made electronic networks of different kinds increasingly prevalent. These networks can contain useful embedded information for predicting future states of the network entities, information that is not available if these elements are considered in isolation.

It is recognized that social networks that describe relationships between individuals such as Facebook and LinkedIn contain useful relational information that can be used for a variety of purposes. A different kind of network, one that has received less attention in the literature and popular media thus far, but which we believe is central to graph-based data mining and predictive modeling, is an *economic network*, which we define as a network in which the links between the nodes are created by actual realizations of shared economic outcomes between entities which are represented by the nodes of the network.

A common economic network might be the result of linking consumers based shared prior choice, purchasing or browsing behaviors; such graphs have also been referred to as "pseudo-social networks" (Martens and Provost, 2011). Another economic network might be the result of linking the product landing pages on an electronic commerce website based on shared demand or purchasing outcomes. Of course, this is a subset of all networks of products which link different items sold online; after all, on 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 is linked by a variety of hyperlinks to other product pages. If one imagines the process of browsing an ecommerce site as being analogous to walking the aisles of a physical store, then the "aisle structure" of ecommerce is defined by this graph of interconnected products, and the network position of a product in this graph is its virtual "shelf position."

One of the longest standing examples of such a visible network is the "co-purchase" network of Amazon.com, which, for many years now, has presented its consumers with links to complementary products made visible under the label "Consumers who bought this item also bought…" Furthermore, this network of products is indeed an economic network, since the links are created by shared dyadic purchasing overlaps. We focus specifically on networks of this kind for the remainder of the paper, simply referring to them as "*product networks*" for brevity.

This research is motivated by the belief that such product networks may contain useful embedded information for predicting future states of products. The product network offers a unique prediction task – each node is a product and hence, in a way, its demand represents the aggregation of different diffusion or adoption processes. Therefore, in the context of the product network, demand prediction means that parallel diffusion or adoption processes provide predictive information about each other.

The structure of the specific network we consider can provide two sources of predictive information. First, neighboring products (products that directly link to one another) may be found to be a good reference group with high demand correlation. After all, this network is based on frequent co-purchase behavior, and when items tend to be purchased concurrently, chances are their demand is highly correlated. While demand correlation across products has always existed, the explicit visibility of these relationships embodied in hyperlinks makes them much more readily available for prediction purposes. Moreover, this information can be very useful, especially if what is driving the correlation is distinct from observable characteristics of the products such as author, subject category, or other attributes that describe the product. Note that for the purpose of predicting future states of an entity, the drivers of such correlation and the direction of causation are irrelevant (more on this in what follows).

Second, the structural properties of a product's network position may incorporate valuable information. Unlike offline stores, which have limited dimensionality, the online store has a complex structure, and the position of a node in that network may include predictive information. For example, it is possible that products in highly clustered areas of the network will exhibit different (higher or lower) demand patterns from products in less clustered areas. In such cases, the clustering coefficient of a product may be predictive of its demand. Similarly, global structural characteristics, such as product centrality, may prove to have high predictive power.

This research asks whether product networks contain any predictive information. In other words, if we want to predict a particular state of an entity, e.g., its future demand, is there information embedded in the network that can be used to make more accurate predictions than could be made using information

about that entity alone? Further, we break down this question into two parts: (a.) Does current and past information regarding neighboring entities contain predictive information? (b.) Do the network's structural properties (as measured by network measures such as PageRank and clustering coefficient) contain additional predictive information? To the best of our knowledge the effects of network-based measures on the accuracy of predicting the future state of network entities have not been previously reported in related literature. We therefore contribute to the existing literature on prediction using networked data by highlighting the unique case of product networks as well as by showing the importance of network structure characteristics, on top of entity-specific information.

There is reason to expect that information on neighboring entities together with network-based measures would contribute to predictive accuracy in product networks. First, it is highly likely that, as in social networks, changes in the state of one entity at the current time are predictive of subsequent changes in the states of entities linked to it, because the propagation of information is not instantaneous. Second, the structural properties of parts of the network, such as the density of connections of entities and their centrality, can be expected to moderate the flow of information in various parts of the network. In other words, we posit that there is a "gestalt" associated with the distributed information implicit in product networks that can be exploited to build better predictive models.

Indeed, the literature has reported on various applications in which "the network matters", that is, applications in which an entity's connections embody information about the entity that is not "contained" in the entity itself. Much of the previous research in this area has focused on methods for network classification (Chakrabarti et al. 1998; Macskassy and Provost 2007; Sen et al. 2008) where the objective is to determine the class probability distribution of a node, and the number of classes is generally small (i.e. fraudulent or non-fraudulent, influenced or not-influenced, etc). Our work builds on these ideas, of exploiting local and global network data, as well as collectively using the network data for inference, but is unique in two important respects. First, we are dealing with a regression forecasting problem, not classification, in that we wish to predict future demand of an entity based on the current state of the network. Second, we treat time as a distinguished variable by ensuring that the network structure at a certain point in time cannot possibly be used to make inferences about the network previous to this point in time, and by allowing the network to change at every point in time. The last point is subtle but important. Previous approaches treat the network as "fixed" at a certain point in time and then make inferences. In contrast, our network structure is not static; we do not know what the future network structure will be. Rather, our concern is only with using the current and past network structure and its entities' states to make a prediction about each entity's future state.

Our paper also carries clear managerial implications. Improved demand prediction is a key factor in inventory planning, inventory management, and supply chain optimization. Accurate demand prediction could also have far-reaching secondary effects on many important aspects such as sales, pricing, and customer satisfaction.

Our focus on prediction is motivated by a number of factors. In his famous treatise titled "Conjectures and Refutations," Popper (1968) argues that the primary criterion of the scientific status of a model is its falsifiability, or refutability, or testability. Prediction, he argues, especially if it is non-trivial, is an essential validation of a model. In a recent paper, Shmueli (2010) provides an extensive review of the literature on explanation versus prediction, showing how the former has dominated the research in information systems largely because of its emphasis on testing hypothesized causal relationships. In the context of product networks, previous work has focused on identifying the causal effect of visible links in a product network on demand (Oestreicher-Singer and Sundararajan, 2012). In contrast, this research focuses on showing that there is predictive information embedded in a product network that can be used to estimate future demand of nodes in the network.

There are several reasons why building a predictive model as opposed to an explanatory one for prediction in online product networks is a worthwhile complementary goal and an important contribution to the literature.

First, explanation and prediction often serve different goals which can lead to significantly different models. Hastie, Tibshirani and Friedman (2009) point out that there are three sources of error in prediction: bias due to miss-specifying the statistical model; estimation error due to the use of a sample to estimate the model; and randomness that will exist even if the model is correctly specified. While

prediction models aim to minimize the total error, explanatory models focus on minimizing bias. Because of this, a model that minimizes the total error need not be the "best" model in terms of describing the "true" phenomenon (Shmueli 2010). Shmueli and Koppius (2011) make the case for the importance of predictive models, specifically, their ability to stimulate theory development, to contribute to the development of new measures, to improve/assess existing theories or to compare competing theories, and to establish benchmarks for the inherent predictability of measurable phenomena. From a managerial perspective, explanatory models can be important in understanding the underlying process and possibly manipulating and controlling those processes to impact outcomes (for example, using marketing tools). Predictive modeling, on the other hand, by minimizing the total error, provides valuable tools for estimating future outcomes based on available data at a point in time.

A second point, related to the point above, is that online networks provide network-centric measures, such as centrality and PageRank, which are increasingly being used to describe these networks. We make use of these and a novel network-centric variable for demand prediction with the expectation that our results will further motivate the use of such variables for prediction as well as for explanatory purposes in other types of problems. In other words, we expect our results to stimulate further theory development in product networks.

Finally, predictive models such as ours provide a benchmark in that they quantify the inherent predictability in product networks based on available information. The explosion in prediction competitions, exemplified by Netflix, and now available on the Kaggle platform, provides researchers with useful performance benchmarks to improve upon, thereby stimulating research on methods, measures, and even on better understanding the underlying phenomena. Our model quantifies how well we do, and should motivate researchers to try to do better, either by building better models, better network metrics or by using more available information than we did. Such quantification also provides a practical benefit to online merchants by enabling them to forecast demand more accurately, thereby optimizing inventory and supply chain management.

We investigate our hypotheses using a massive co-purchase data set gathered from Amazon.com. The data cover nearly 1 million books over 2 years, resulting in a total of over 70 million daily observations. The “naïve” forecast for the next period’s demand would be today's demand. A richer input set would include historical demand for an entity. Richer still would be input from the product’s related entities (e.g., its network neighbors). Finally, the richest input set would include local properties of the network, such as an entity’s local clustering coefficient (Watts and Strogatz 1998) or its PageRank (Brin and Page 1998). We would expect such local and global properties to provide the highest predictive accuracy.

In addition to comparing predictive models based on different sets of inputs, we also compare two classes of modeling choices: The first model is a linear autoregressive (AR) model which is widely known in the time series prediction and economic literature. The second is a neural network (NN), which is a powerful method for modeling complex nonlinear models in situations where plenty of data are available, as is the case in this study.

Evaluating performance while using these diverse algorithms has two main purposes: First, it is interesting to discern whether network-based models perform better regardless of the type of algorithm or whether a certain type of algorithm is better suited to exploit the network “gestalt”. Second, often it is not a priori known which model will provide better predictions, especially for out-of-sample data. While non-linear algorithms such as NN are recognized for their ability to capture complex relations, these algorithms were also found in various cases to over-fit the data. On the other hand, a simple AR model may be inferior in capturing complex relations, but it is also much less likely to over-fit the data. Thus, employing the simpler AR model could serve as a benchmark. While it is possible to use many other models, our choice reflects two standard model types that are commonly used in econometric and machine learning modeling.

Prior Work

Prior work in the information systems, computer science, and marketing literature has recognized the importance of social and economic networks in describing patterns associated with a variety of business outcomes.

There has been considerable interest recently in data mining that is based on social network data (Kleinberg 2007; Hill et al. 2006), and parallel interest in mining web-based data sources for developing micro-data-based decision support systems for making more data-enabled business decisions (Marsden 2008). Some of the popular applications for networked data sets include prediction of new product diffusion and churn, fraud detection, and counter-terrorism. An early paper by Domingos and Richardson (2001) uses knowledge sharing sites to mine implied social relationships between consumers towards improving targeted marketing and optimizing the marketing funds allocated to each consumer. More recent work has focused largely on how electronic instantiations of "social networks" (for example, like those implied by corporate email communication patterns, online conversations or by an instant messenger network) serve as conduits for information flow (Godes and Mayzlin 2004), how network structures affect such flows (Bampo et al. 2008; Kiss and Bichler 2008), what fraction of these flows can be ascribed to true influence (Aral et al. 2009; Ma et al. 2009; Sun and Zeng 2008), and more broadly, how to improve the mining of information by leveraging information about social networks (Song and van der Aalst 2008).

Prior work on product networks includes a study of the network of videos on YouTube by Susarla et al. (2011), a study of the network of blogs by Mayzlin and Yoganarasimhan (2012), and a study of the network of news reports by Dellarocas et al. (2009). Goldenberg et al. (2012) studied the interaction between product networks and social networks in the context of YouTube. Oestreicher-Singer and Sundararajan (2012) studied the network of books on Amazon.com and quantified the incremental correlation in book sales attributable to the product network's visibility. In contrast with our predictive focus, these papers are concerned with the causal impact of the product network on demand for individual products. Rhue and Sundararajan (2010) examine how effectively one might explain the future donation choices of consumers who are linked based on shared prior charitable giving behaviors; Martens and Provost (2011) provide a method for target marketing based on a 'pseudo-social network' (PSN): consumers are linked if they have transferred money to the same entities in the past. Our work contributes to this stream of research by analyzing the predictive potential of a different kind of economic network, a product network.

Our current question steps away from deconstructing information flows or distinguishing influence from other drivers of correlation, and instead examines whether outcomes can be predicted in the context of network links that are created by a high fraction of copurchases. We focus on using state changes of an entity in predicting state changes of other entities in the network. Specifically, a change in an entity in one period results in a cascade of changes across the entire network, where the changes are predictions for the next time period. We identify the local neighbors of a given product and analyze the structural characteristics (global and local) of the product within the network as a basis for predicting future demand of related products.

It is important to emphasize that the interconnection between economic objects (agents, products) is not necessarily on account of their explicitly sharing one or more observable features or characteristics such as author, topic or genre. As mentioned earlier, such features are part of what we term a product's "intrinsic features" and are typically used as a basis for prediction in data mining. While we may leverage such information, our primary interest is in demonstrating that an entity's future demand is more accurately predicted by combining its historical demand with that of its neighbors and its network "positioning" than by considering its demand alone.

Data

We use a large time-series data set of recommendation networks for nearly 1 million books sold on Amazon.com. Each product on Amazon.com has an associated webpage. Each page has a set of co-purchase links, which are hyperlinks to the set of products that were copurchased most frequently with this product on Amazon.com. This set is listed under the title "Customers who bought this also bought:" An example of co-purchase links is illustrated in Figure 1.

The co-purchase network is a directed graph in which nodes correspond to products, and edges to directed co-purchase links. We collect data about this graph using a Java-based crawler, which starts from a popular book and follows the co-purchase links using a depth-first algorithm. At each page, the crawler gathers and records information for the book whose webpage it is on, as well as the co-purchase links on that page, and terminates when the entire connected component of the graph is collected. This is repeated

daily. A sample part of the graph is illustrated in Figure 2a, and the corresponding larger segment of a co-purchase network is shown in Figure 2b.

We have chosen to focus on books because they are the product category with by far the largest number of individual titles, their product set is relatively stable (compared to electronics, for instance), and their network data are observable.

We use data collected from August 2005 till September 2007. The graph is traversed every day. We utilize the following data fields, which are available for each book on the co-purchase graph for each day:

- ASIN: a unique serial number given to each book by Amazon.com. Different editions and different versions have different ASIN numbers.
- Copurchases: ASINs of the books that appear as the focal book's copurchases.
- SalesRank: The SalesRank is a number associated with each product on Amazon.com, which measures its demand relative to that of other products. The lower the number is, the higher the sales of that particular product.
- Category Affiliation: Amazon.com uses a hierarchy of categories to classify its books. Thus, we extract for each book its association with two categories: Cat_h: a high-level category affiliation (e.g., Business & Investing); Cat_l: a specific, "low-level" category affiliation (e.g., Accounting).

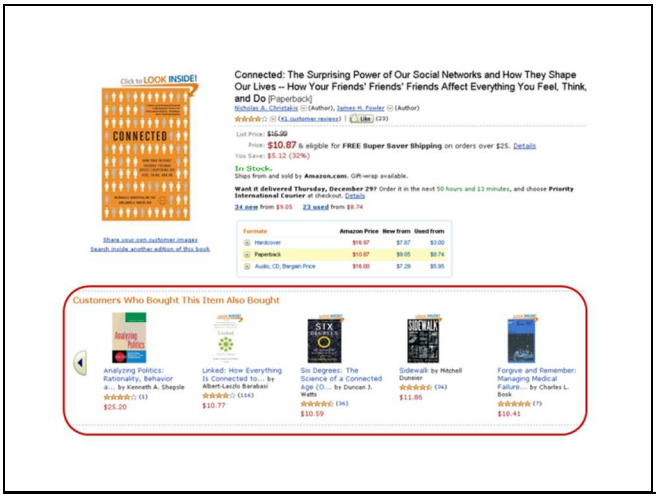

Figure 1. An Example of Co-purchase Links

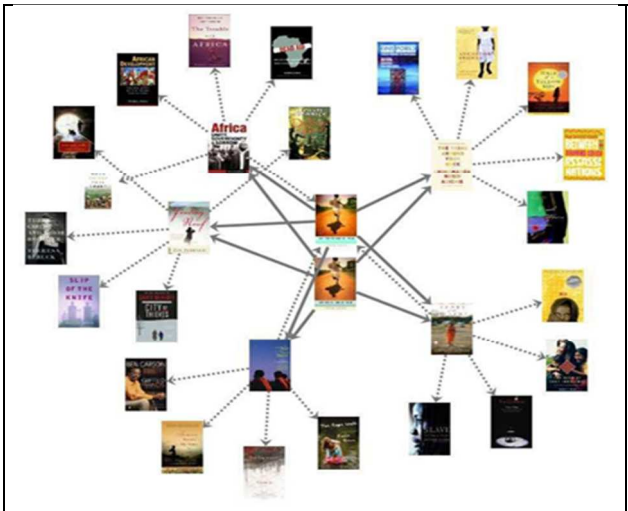

Figure 2a. Small Segment of a Co-Purchase Network

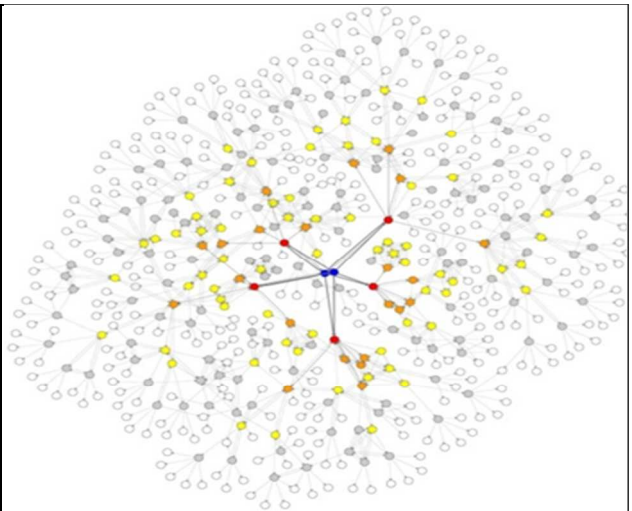

Figure 2b. Larger Subset of the Co-Purchase Network

An additional script collects the SalesRank for each book on the graph every 3 hours for the 24-hour period following the collection of the graph. We convert SalesRank data into demand since we are dealing with the prediction of demand, not SalesRank, and since demand is subject to arithmetic operations (like addition), whereas SalesRank is more nebulous in this respect as well as conceptually. The demand computed is based on the SalesRank data generated by Amazon and following a log-linear conversion model suggested by Chevalier and Goolsbee (2003) and by Brynjolfsson et al. (2003). Based on the collected data, we construct the following variables for each node:

- **Neighboring Nodes' Demand Data (InDemand):** We calculate the cumulative demand for each node's (incoming) neighbors.[1]
- **Network Structure-Based Variables:** For each book we calculate two variables that are based on the network structure: PageRank and Local Clustering Coefficient.[2]

We calculate each book's PageRank score according to Google's original algorithm (Brin and Page 1998). The original PageRank algorithm provides a ranking of the "importance" of webpages based on the link structure of the "web" created by the hyperlinks between the pages. In our case, we use it to compute the "importance" of each book (i) according to the network's daily graph structure, using the following mode:

$$PageRank\ (i) = \sum_{j \in G(i)} \frac{PageRank\ (j)}{OutDegree\ (j)} \qquad (1)$$

where $G(i)$ is the set of books which have outgoing links to book (i), and OutDegree(j) is the total number of links originating from book j.

For each book we also calculate the local clustering coefficient (Watts and Strogatz 1998). The local clustering coefficient $C(i)$ for a given node i is a measure of how close the node and its neighbors are to being a clique and is computed as:

$$C(i) = \frac{\text{Number of edges between the neighbors of } i}{k(i)(k(i)-1)} \qquad (2)$$

where $k(i)$ is the number of direct incoming neighbors of node i, and is the maximal potential number of connections between node i's direct incoming neighbors.[3]

Sampling and Data Cleaning

The entire book data set collected as part of this study is immensely large. In order to construct and evaluate prediction models, we utilized a more manageable and "cleaner" set of books involving three million instances (30,000 books per day, over 100 days). First, we filtered out time periods with unusual purchasing behavior. Specifically we removed seven days before and after the following holidays: Christmas, New Year's Day, Thanksgiving and Valentine's Day. Second, we removed from the sample books whose data were missing or partial. Third, we removed unusual or extreme observations, filtering out the top and bottom deciles for daily book demand as well as books whose daily PageRank scores (during the preceding 7 days) displayed unusual behavior—going up or down by a factor of 20 (assuming that those books must be subject to unusual exogenous factors, like marketing campaigns). Last, we randomly selected 100 days from our data set.[4] From each of these days we selected a random sample of 30,000 books. Random selection of the days forced us to sample across many periods, thereby eliminating any bias that may exist within a specific time period, and hence increasing the generalizability of the predictive model.

---

[1] An outgoing link from book X to book Y is recorded in cases that book X's web page reports that book Y is co-purchased with book X.

[2] These measures represent both global and local properties of the network and were utilized due to the fact that they have rapid computational estimation which makes them suitable for applicative purposes.

[3] In later stages we utilize a log transformation of the different variables. To avoid log(0), in cases where C(i)=0, we replace it with a value of 10^{-5}.

[4] Note: for each date included in our sample we collected relevant information during the preceding 7 days.

Models and Data Sets

We use several sets of data inputs to predict future demand. We begin with the data set we expect to be the least predictive and successively augment the data with additional explanatory variables. The different sets are specified below:

(a) **"Naïve" Baseline:** We use *log(Demandi,1)* as an estimate for *log(Demandi,0)*, where *Demandi,0* is node *i*'s unobserved demand today, and *Demandi,1* is node *i*'s observed demand yesterday.

(b) **Historical Demand:** We use *log(Demandi,1)*, …, *log(Demandi,N)* to predict *log(Demandi,0)*, where *Demandi,1* is node *i*'s observed demand yesterday, and *Demandi,N* is node *i*'s observed demand *N* days ago.

(c) **Historical Demand and (Incoming) Neighboring Nodes' Demand:** We use *log(Demandi,1)*, …, *log(Demandi,N)* as well as *log(InDemandi,1)*, …, *log(InDemandi,N)* in order to predict *log(Demandi,0)*, where *InDemandi,1* is yesterday's total observed demand for node *i*'s direct (incoming) neighbors (nodes with directional links pointing at node *i*), and *InDemandi,N* is the total observed demand of node *i*'s direct neighbors *N* days ago.

(d) **Historical Demand, Neighboring Nodes' Demand and PageRank:** We use *log(Demandi,1)*, …, *log(Demandi,N)*, *log(InDemandi,1)*, …, *log(InDemandi,N)*, and *log(PageRanki,1)*,…, *log(PageRank i,N)* in order to predict *log(Demandi,0)*, where *PageRanki,1* is yesterday's observed PageRank of node *i*, and *PageRank i,N* is the PageRank of node *i N* days ago.

(e) **Historical Demand, Neighboring Nodes' Demand and Local Clustering**: We use *log(Demandi,1)*, …, *log(Demandi,N)*, *log(InDemandi,1)*, …, *log(InDemandi,N)*, and *log(LocalClusti,1)*,…, *log(LocalClusti,N)* in order to predict *log(Demandi,0),* where *LocalClusti,1* is yesterday's observed Local Clustering Coefficient of node *i*, and *LocalClusti,N* is the Local Clustering Coefficient of node *i N* days ago.

Prediction Models

To capture the effect on prediction accuracy of successively adding information, we separately used two well-known models to generate predictions based on each of the data sets: (b), (c), (d), and (e) which were detailed above.[5]

(a.) An Auto-Regression (AR) Model with a least-squares estimator. This is a simple linear model often used to model time-series data. We implement the algorithm using the R software.

(b.) A Back-Propagation-Based Neural Network (NN) algorithm. This is a non-linear model that is estimated by the backprop algorithm (Werbos 1974). Neural networks attempt to "learn" patterns from data directly, by sifting the data repeatedly, searching for relationships, automatically building models, and correcting over and over again the model's own mistakes. The technique can derive good models even when the data are incomplete or noisy as long as they are plentiful and representative of the variety of situations in which the models will make predictions (Dhar and Stein 1997). For a rich historical perspective on neural networks, see Hinton (2002), and Rumelhart and McClelland (1984). We use the R software package "nnet" to implement the backprop algorithm. This implementation involves one layer of hidden neurons, and the minimization of a Sum of Square Errors criterion.
Last, each prediction model was recreated several times time using a different number of historical lags (*N*), from one day up to seven days. The use of multiple lags of the network past states provides the forecasting model with information in the form of multiple "snapshots" of the network past states.

[5] We note that a plethora of data mining algorithm is suitable for handling this problem. However, our motivation is not to test competing data mining algorithms. Rather, we focus on two well-known algorithms to test the predictive contribution of the different data sets, and to discern whether imposing/ not imposing linearity on the model affects the results. For this purpose, we use two popular algorithms. A linear AR model and a flexible NN model.

Results

The literature on model validation is extensive. We use a validation framework that unifies several streams of literature and is described in Sobehart et al. (2001). The framework views model validation along two dimensions. The first is across time: How well does a model built on data over a certain period of time perform on future data? The second is across universe: If a model is built on a certain universe (such as books, financial instruments, etc.), how well does it perform on elements of the universe that were not used in building the model? The validation framework is presented in Figure 3. The lower rightmost quadrant represents the most stringent validation test in that the data used for testing a model represent a future point in time and entities in the universe that were not used to build the model.

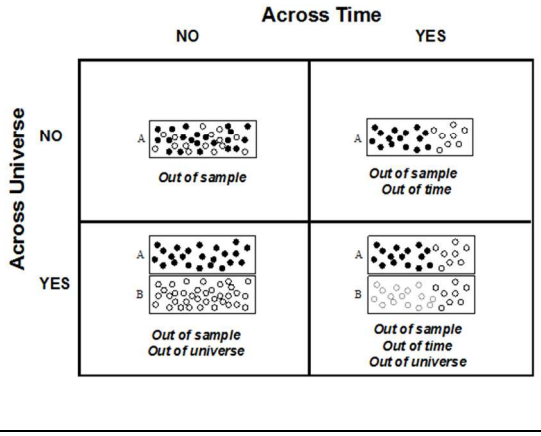

Figure 3. Validation Framework

We partitioned the data sample into two sets: a training set, which includes the first 66 days (1,980,000 data instances) and is used to construct the prediction models, and a test set, which includes the remaining 34 days (1,020,000 data instances) and is used to independently evaluate the forecasting models. This ensures validation across time in that a model is tested only on future data. We further ensure that the test set contains books that never appeared in the training set. These two guidelines provide a robust basis for model validation.

Since NN performance is known to be sensitive to the number of hidden neurons (HNs), we check the robustness of our NN models using three different architectures. Each architecture uses one layer of HNs, but the architectures differ in the proportion of HNs to the number of inputs. We determine the number of HNs as 0.5, 1, or 2 times the number of inputs.[6] While there were some differences in performance, overall, the three architectures yielded similar findings. For the sake of brevity, we report only on the simplest NN (with the number of HNs set as half the number of inputs). This network was the most conservative in terms of improving overall performance.

We display the different models' results (over the independent validation set) in Figures 4–8. We present the results using the well-known Mean Square Error (MSE) criterion. This measure penalizes in a similar manner excess demand as well as shortage and penalizes more severely larger errors. Since both AR and NN implementations share the goal of minimizing the sum of square errors, this criterion is also suitable for making a proper evaluation of how the two algorithms utilize the various sets of data. For completeness, we repeated the analysis using another well-known measure-- mean absolute percentage error (MAPE), and the results were very similar. (MAPE results are not presented here for the sake of

[6] There are no hard and fast rules about what is the optimal complexity of the network in terms of hidden nodes. Nevertheless, some guidelines have been proposed in the literature (Zhang et al. 1998). One such guideline is that the number of hidden units should be proportional to the number of inputs. Specific architectures have been proposed, e.g., the number of inputs times 0.5, 1, and 2. In our experiments, we use these three levels of complexity, choosing architectures where the number of hidden neurons is half, one, and twice the number of inputs. (For the NN with half the number of HNs—in cases in which the number of inputs is an odd number we round up the number of HNs.)

brevity). Confidence intervals (95%) for the difference in MSE values, between different data sets, are reported in Appendix A.

Several patterns are apparent in the results. Figure 4 compares a Naïve model that uses the previous day's demand as the next period's demand versus an AR model with Historical Demand, as well as versus an AR model with Historical Demand and Neighboring Nodes' Demand (InDemand). The simple AR model with Historical Demand information does much better than the Naïve model, which is in line with expectations. The more past periods used, the better the performance. Adding Neighboring Nodes' Demand to the AR model provides further improvement to the prediction accuracy.

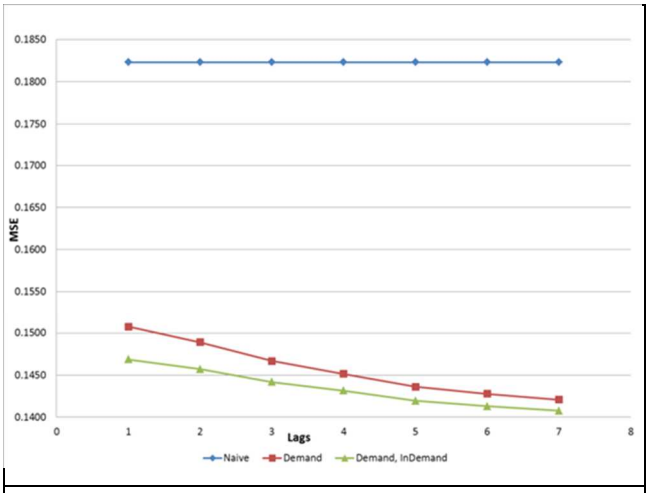

Figure 4. AR Model Inputs: Naive Vs Demand Vs Demand & InDemand

Figure 5 shows the results for the AR model when network-based variables are added, namely, PageRank and local clustering. Interestingly, the additional information does not yield significantly better predictions compared with an AR model that uses Historical Demand and InDemand information. See also Appendix A (Table A.1, two rightmost columns), which shows that in most cases the 95% confidence intervals for MSE difference include zero.

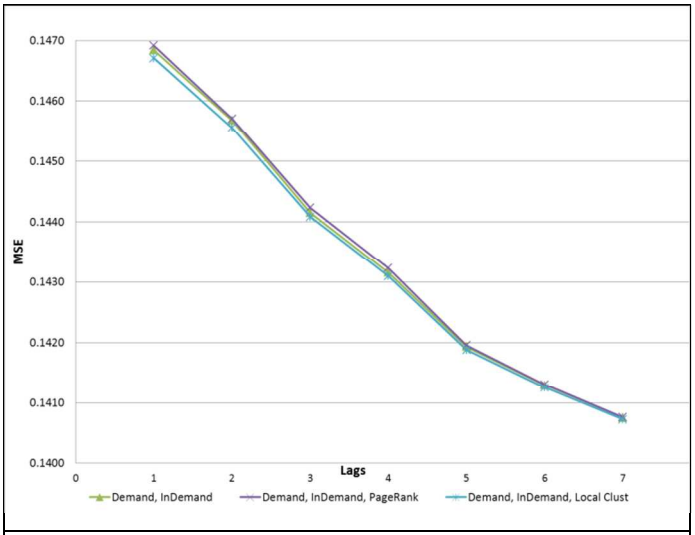

Figure 5. AR Model: Adding PageRank or Local Clustering to Demand & InDemand

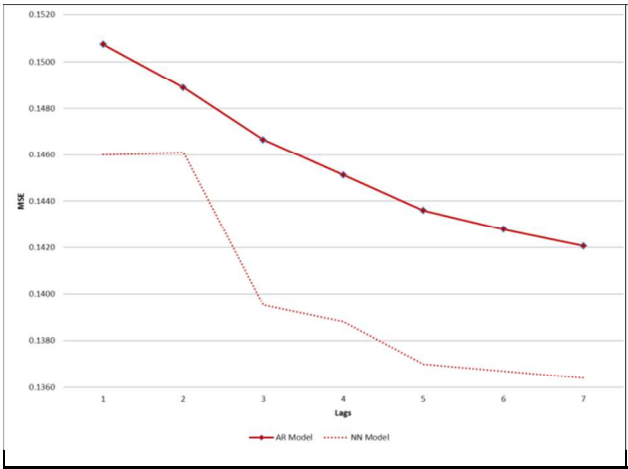

Figure 6a. Demand Data: AR vs NN

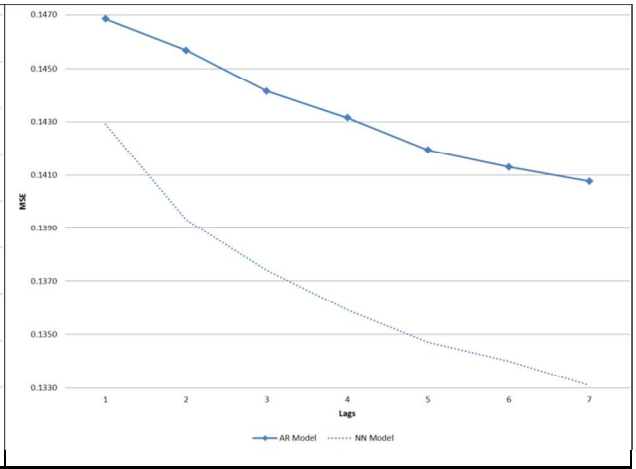

Figure 6b. Demand & InDemand Data: AR vs NN

Figure 6a compares the NN model to the AR model based on Historical Demand data of the entities. It is notable that the neural network model considerably outperforms the AR model when using demand data.

Figure 6b compares the NN model to the AR model with the addition of InDemand. Again, the NN model considerably outperforms the AR model when using both Historical Demand data and InDemand data.

Figure 7 compares the performance of the NN model given different sets of inputs. In general, the NN model not only performs better than the AR model but is also able to "extract" additional predictive power out of the network-based variables—something that the AR model fails to achieve. The general conclusion here is that the network-based variables provide predictive power as long as the model is capable of extracting it.

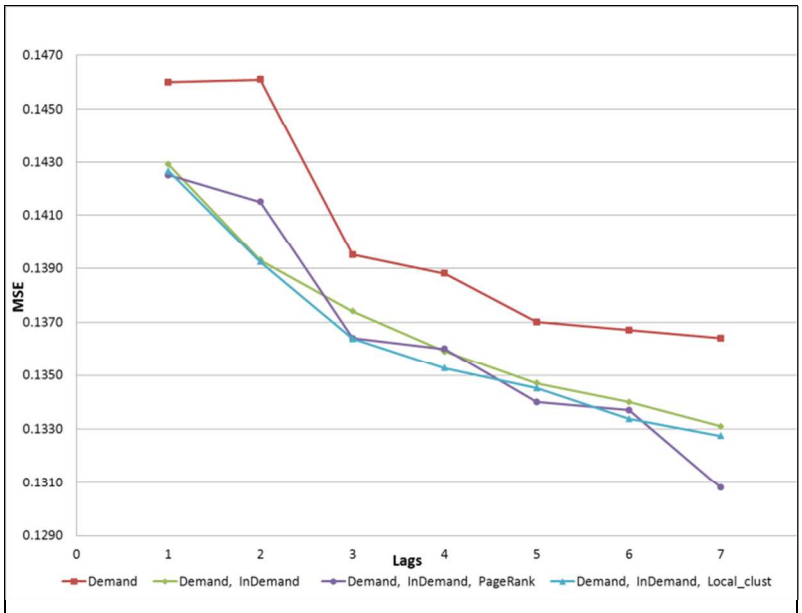

Figure 7. NN Model

It is important to note that even small visual differences in Figure 7 have a substantial business impact. First, the 95% confidence intervals (Appendix A, Table A.2) for the difference in MSE between pairs of data sets are generally very narrow and do not include zero. For instance, in Figure 7 there is a difference in MSE of ~0.00035 between a model using one lag of Historical Demand and InDemand Information and a model utilizing one lag of Historical Demand, InDemand as well as PageRank Information. Nevertheless, the 95% confidence interval for this difference is between 0.0003 and 0.0004.

Second, for a large network such as this one, which involves the maintenance of a large inventory of books over time, small improvements in prediction have a considerable economic impact. For instance, the 0.00035 improvement in MSE in the example above could be translated into an aggregated reduction in prediction errors of ~500,000 books per year.[7]

Sensitivity Analysis

We have performed a number of sensitivity analyses to obtain additional insights into our results. Our first type of analysis seeks to discern whether predictive performance depends on network structure or product characteristics. For this analysis, we focus on the performance of NN models using seven lags of data since this configuration provides the best results across all types of data inputs and model choices.

Figure 8 presents the performance of the different models across five quintiles of *Demand*$_1$ data (Q1–Q5). In this figure, dots represent MSE values; and lines, above and below the dots, represent two standard deviations above and below the MSE value, respectively. From this figure it is evident that all models perform much better when Demand is high (Q5). This is interesting and may occur because demand spikes are probably less pronounced for products with already high demand. Interestingly, for high demand, models using network-based parameters perform significantly better than models without the network-based parameters. It is also noteworthy that with the exception of Q3, models using PageRank perform consistently well. It is also notable that across all quintiles, adding local clustering to Demand and InDemand data never makes the results worse, and in some cases it improves them.

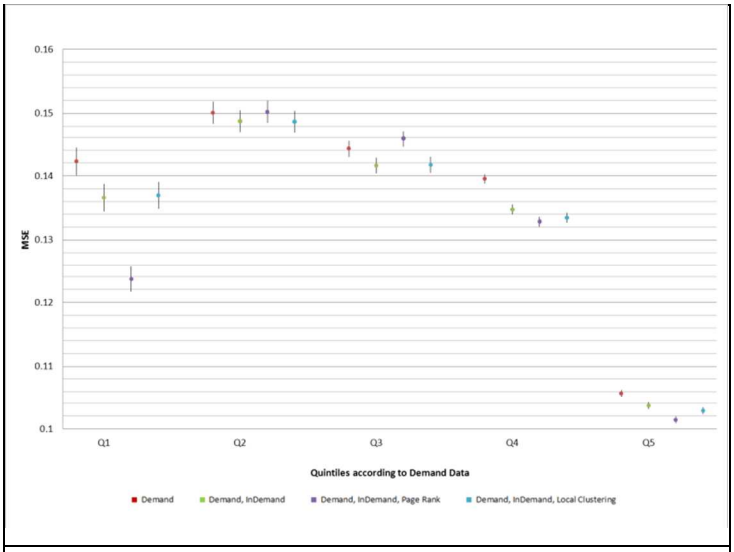

Figure 8. NN Model Performance: Different Inputs by Demand Quintiles

Figure 9 presents the performance of the different models across five quintiles of *InDemand*$_1$ data. Network information provides value across the quintiles. The contribution of adding PageRank is more noticeable at the lower range (Q1, Q2). We have noticed that the volatility of *InDemand* is the lowest in the lower quintiles. This could account for the improved performance in these quintiles.

Figure 10 presents the performance of the different models across five quintiles of *PageRank*$_1$ data. Again,

[7] This improvement amounts to approximately 42 books per day due to reduction in prediction errors for a daily sample of 30,000 books. Assuming that Amazon has to offer at least one million physical books with similar characteristics, this would translate into ~1,400 books per day. Over a period of one year this would sum up to ~500,000 books. In reality, this calculation may be overly simplistic as it does not take into account many factors such as inventory ordering schedules, purchase costs and others. However, it does indicate the considerable impact of small improvements in prediction accuracy.

adding PageRank data to the prediction models provides the most significant improvement at the lower ranges of PageRank (Q1, Q2), while adding local clustering data provides a small improvement to the baseline (e.g., using demand data) across the entire range.

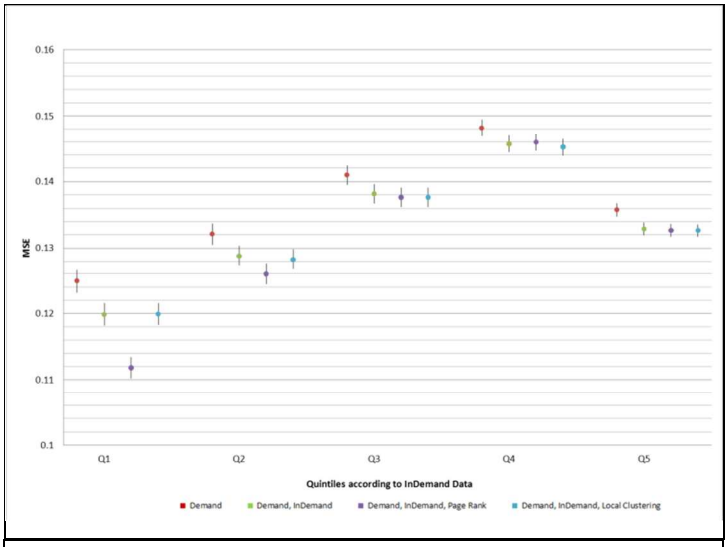

Figure 9. NN Model Performance: Different Inputs by InDemand Quintiles

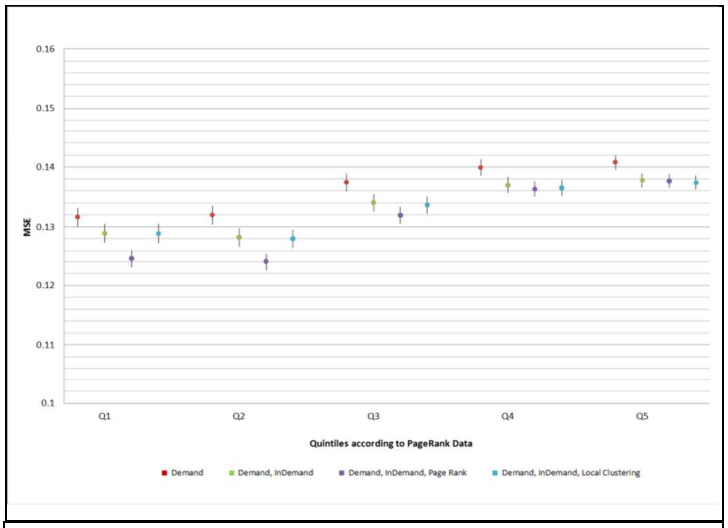

Figure 10. NN Model Performance: Different Inputs by PageRank Quintiles

Finally, a similar picture appears in Figure 11, which presents the performance across five quintiles of $LocalClust_1$ data. In this case, PageRank provides a significant improvement at the bottom and top quintiles while local clustering provides a small improvement in predictive performance across the entire LocalClust range.

A summary conclusion from Figures 8-11 is that models using network-based parameters tend to make better predictions across all quintiles, and PageRank appears to be a consistent network metric in improving predictions. Local clustering tends to provide a small and consistent improvement across the entire data range relative to the baseline.

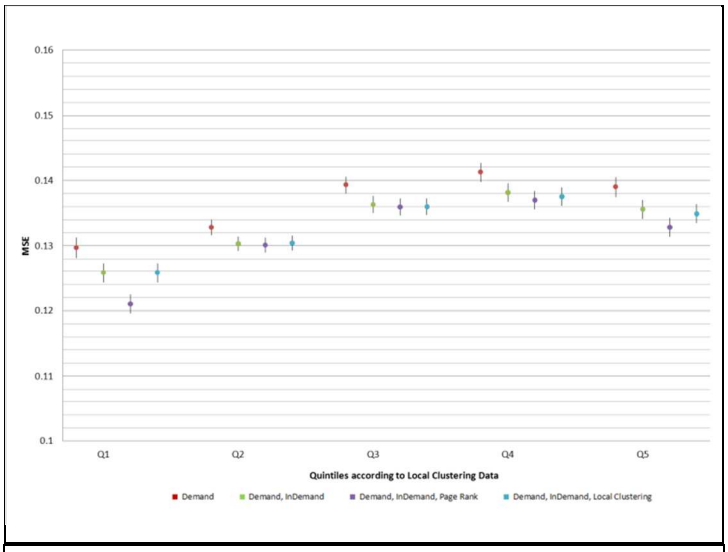

Figure 11. NN Model Performance: Different Inputs by Local Clustering Quintiles

In Figure 12 we present the effect of books' vintage (measured in number of years since the book was published) and category. Figure 12 shows that predictive performance is stronger for books aged between 2–20 years, whereas it is somewhat weaker for very new or very old books. As observed in previous analyses, models using network-based parameters generally outperform the baseline group across all vintage groups. Similarly, PageRank looks promising, yielding significant improvement in prediction performance for books aged between 2 and 20.

We also looked at predictive performance across different categories of books. (For brevity the results are not presented here and available upon request.) Although a few categories, e.g., Comics, Mysteries and Science Fiction, have higher error rates due to small sample size—consistent with previous analyses, the improvement in prediction using network information is apparent across several categories, as is the role of PageRank.

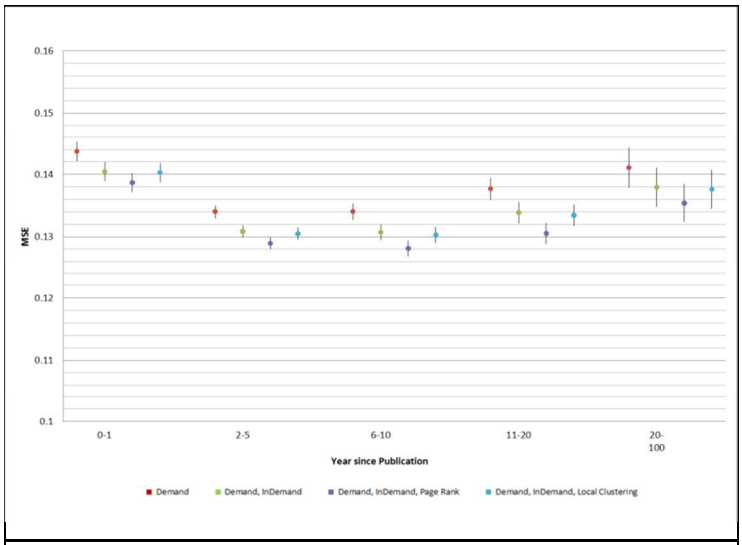

Figure 12. NN Model Performance: Different Inputs by Number of Years since Publication

Comparison with Average Category Demand

Last, we examine other possible explanations for our results besides the informative value of the network. One alternative explanation is that the network provides us with information about the identity of a group of products that are similar (or perhaps complements) to the focal product, and therefore additional information is almost always helpful, especially if it is of relevant products. That is, network-based information such as network neighbors' demand does not necessarily provide unique information; rather, any information about a relevant reference group will provide the same results.

To test this alternative explanation we use a comparable reference group—the group of products belonging to the same category. Given that in most brick and mortar book stores books are organized by category, this seemed like the most natural reference group to use. If, indeed, the explanation for the improvement in prediction accuracy is the mere use of reference group information, rather than the specific use of network-based data, then adding category information on top of historical demand information would provide an improvement in prediction results similar to that achieved by adding network neighbors' demand (InDemand). For brevity the results are not presented here; they are available upon request. The general conclusion is that adding average category demand to demand data provides at best a marginal performance improvement, and the models' performance is inferior to that obtained with both Historical Demand and InDemand data.

Discussion and Future Work

Oestreicher-Singer and Sundararajan (2012a) have previously shown that the impacts of visible network links can be econometrically identified and are significant. In this study, we set out to answer a different question, namely, whether changes in demand can be predicted, and specifically, whether using network information leads to higher predictive accuracy.

The results show that this is in fact the case, and that our hypothesis holds. To the best of our knowledge, we provide the first evidence using a large-scale study on the existence of information that is predictive of demand and is contained in the structure of product networks. We are able to show that, compared with naive predictions, our most elaborate model yields an aggregate improvement of sales predictions that translates into 32 million units per year. This clearly has major economic implications for online retailers.

There are many avenues for future work. In this first study using product networks for predictive modeling, we first restricted our attention to the simplest possible "neighborhood" of a product, its immediate in-neighbors, and then added other network-based information to the input set. Testing the predictive power of more distant neighbors remains a promising line of future inquiry.

While our intuition at the outset of the project was that the product network contains useful predictive information about sets of linked entities, it was not obvious that these would manifest as lagged effects. In contrast, if, for example, new information is reflected instantly in related entities, we would expect changes in the states of entities to be concurrent. Our results suggest that this is not the case, and that additional information is not instantaneously propagated through a product network. A natural next step in this direction would be to associate varying levels of lag influence with nodes that are different distances from the product in question. For example, one might expect the influence of a distant neighbor to take longer to measurably affect the economic outcomes we observe. The variation of this across product categories is another promising line of inquiry.

Having validated our initial central conjecture, namely that the product network contains predictive information, it seems natural for future research to consider the use of other machine-learning methods in addition to neural networks to build more accurate predictive models. Machine-learning methods can deal with problems with high levels of noise by discovering "local" models that can be applied to different partitioning of the data. These localized models make more aggressive forecasts for specific books, in contrast to the AR model, which is very conservative in its predictions, with small deviations from the mean. The results using the NN models validate this thinking.

With the explosion of Web 2.0 technologies and the growing adoption of social media, we expect that numerous new product networks will become observable to firms and researchers over the next years. The tight association of the links in these networks with economic outcomes of interest makes these networks especially attractive as a basis for predictive modeling. These networks are the natural place to start when looking for ways to expand beyond product-centric features to improve predictive accuracy for network entities.

Appendix A.

Table A1 - AR Model, 95% Confidence Intervals for MSE Difference

MSE(Demand)-MSE(Demand, InDemand)		MSE(Demand, InDemand)-MSE(Demand, InDemand, PageRank)		MSE(Demand, InDemand)-MSE(Demand, InDemand, Local_Clust)	
0.0038	0.0040	-0.0001	-0.0001	0.0001	0.0001
0.0031	0.0033	0.0000	0.0000	0.0001	0.0001
0.0024	0.0026	-0.0001	-0.0001	0.0001	0.0001
0.0019	0.0020	-0.0001	-0.0001	0.0000	0.0001
0.0016	0.0017	0.0000	0.0000	0.0000	0.0001
0.0014	0.0015	0.0000	0.0000	0.0000	0.0001
0.0012	0.0014	0.0000	0.0000	0.0000	0.0000

Table A2 - NN Model, 95% Confidence Interval for MSE Difference

MSE(Demand)-MSE(Demand, InDemand)		MSE(Demand, InDemand)-MSE(Demand, InDemand, PageRank)		MSE(Demand, InDemand)-MSE(Demand, InDemand, Local_Clust)	
0.0030	0.0031	0.0003	0.0004	0.0002	0.0003
0.0066	0.0069	-0.0023	-0.0020	0.0001	0.0001
0.0020	0.0022	0.0009	0.0010	0.0009	0.0010
0.0029	0.0031	-0.0002	-0.0001	0.0005	0.0007
0.0023	0.0024	0.0005	0.0007	0.0001	0.0002
0.0026	0.0028	0.0002	0.0003	0.0005	0.0007
0.0032	0.0034	0.0021	0.0024	0.0003	0.0004

*Confidence intervals of 95% for the difference in MSE between various models were calculated by a bootstrapping procedure. For this purpose we used the "boot" library in R software (500 iterations, "basic" bootstrap category).

References

Aral, L., Muchnik, S., Sundararajan, A. 2009. "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences, U S A* (106:51), pp. 21544–21549.

Bampo, M., Ewing, M., Mather, D., Stewart, D., and Wallace, M. 2008. "The Effects of the Social Structure of Digital Networks on Viral Marketing Performance," *Information Systems Research* (19:3), pp. 273–290.

Brin, S., and Page, L. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems* (30:1-7), pp. 107–117,

Brynjolfsson, E., Hu, Y., and Smith, M. D. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* (49), pp. 1580–1596.

Chakrabarti, S., Dom, B., and Indyk, P. 1998. "Enhanced Hypertext Categorization Using Hyperlinks," In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 307–319.

Chevalier, J., and Goolsbee, A. 2003. "Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com," *Quantitative Marketing and Economics* (1:2), pp. 203–222.

Dhar, V., and Stein, R. 1997. *Seven Methods for Transforming Corporate Data into Business Intelligence*, Englewood Cliffs, NJ: Prentice-Hall.

Dellarocas, C., Katona, Z., and Rand. W.M., 2009. Media, Aggregators and the Link Economy: Strategic Hyperlink Formation in Content Networks. Working paper.

Domingos, P., and Richardson, M. 2001. "Mining the Network Value of Customers," Proceedings of the *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 57–66.

Godes, D., and Mayzlin, D. 2004. "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science* (23:4), pp. 545–560.

Goldenberg, J., Oestreicher-Singer, G., Reichman, S., 2012. "The Quest for Content: How User-Generated Links Can Facilitate Online Exploration". *Journal of Marketing Research* Vol. 49, No. 4, pp. 452-468

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hill, S., Provost, F., and Volinsky, C. 2006. "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science* (21:2), pp. 256–276.

Hinton, G. E. 2002. "How Neural Networks Learn from Experience," *Cognitive Modeling*, Polk, T. A., Seifert, C. (Eds), pp. 181–195.

Kiss, C., and Bichler, M. 2008. "Identification of Influencers: Measuring Influence in Customer Networks," *Decision Support Systems* (46:1), pp. 233–253.

Kleinberg, J. 2007. "Challenges in Mining Social Network Data: Processes, Privacy, and Paradoxes," *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, pp. 4–5.

Ma, A., Montgomery, L., and Krishnan, R. 2009. "Homophily or Influence? An Analysis of Purchase Decisions in a Social Network Context," Workshop on Information Systems and Economics.

Marsden, J. 2008. "The Internet and DSS: Massive, Real-Time Data Availability Is Changing the DSS landscape," *Information Systems and E-Business Management* (6:2), pp. 193–203.

Macskassy, S.A., and Provost, F. 2007. "Classification in Networked Data: A Toolkit and a Univariate Case Study*," Journal* of Machine Learning *Research* (8:May), pp. 935-983.

Mayzlin, D., and Yoganarasimhan, H. 2012 "Link to Success: How Blogs Build an Audience by Promoting Rivals" *Management Science,* forthcoming

Oestreicher-Singer, G., and Sundararajan, A. 2012. "The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets," *Management Science*, forthcoming.

Popper, K.. 1968. *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York: Basic Books.

Martens, D. and Provost, F, 2011. Pseudo-Social Network Targeting from Consumer Transaction Data. *http://ssrn.com/abstract=1934670*

Rhue, L. and Sundararajan, A., 2010. The Information Content of Economic Networks: Evidence from Online Charitable Giving. *Proceedings of the 31st International Conference on Information Systems. http://aisel.aisnet.org/icis2010_submissions/250/*

Rumelhart, D. E., and McClelland, J. L. 1984. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, MA: MIT Press.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. 2008. "Collective Classification in Network Data," *AI Magazine* (29:3), p. 93.

Shmueli, G. 2010. "To Explain or to Predict?" *Statistical Science* (25:3), pp. 289–310.

Shmueli, G., and Koppius, O. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553-572.

Sobehart, J., Keenan, S., and Stein, R. 2001. "Benchmarking Quantitative Default Risk Models: A Valuation Methodology," *Algo Research Quarterly* 4(1,2).

Song, M., and van der Aalst, W. 2008. "Towards Comprehensive Support for Organizational Mining," *Decision Support Systems* (46:1), pp. 300–317.

Sun, A., and Zeng, D. 2008. "Maximizing Influence through Online Social Networks," *Proceedings of the 18th Workshop on Information Technology and Systems*.

Susarla, A., Oh, J. H., and Tan, Y., 2011. Social networks and the diffusion of user-generated content: Evidence from YouTube. Information Systems Research

Watts, D. J., and Strogatz, S. 1998. "Collective Dynamics of 'Small-World' Networks," *Nature* (393:6684), pp. 440–442.

Werbos, P. J. 1974. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," PhD thesis, Harvard University.

Zhang, G., Patuwo, B. E., and Hu, M. Y. 1998. "Forecasting with Artificial Neural Networks: The State of the Art," *International Journal of Forecasting* (14), pp. 35–62.