

Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems

Zan Huang

Department of Supply Chain and Information Systems, Pennsylvania State University, 419 Business Building,
University Park, Pennsylvania 16802, zanhuang@psu.edu

Daniel D. Zeng, Hsinchun Chen

Department of Management Information Systems, The University of Arizona, McClelland Hall 430,
1130 East Helen Street, Tucson, Arizona 85721 {zeng@eller.arizona.edu, hchen@eller.arizona.edu}

We apply random graph modeling methodology to analyze bipartite consumer-product graphs that represent sales transactions to better understand consumer purchase behavior in e-commerce settings. Based on two real-world e-commerce data sets, we found that such graphs demonstrate topological features that deviate significantly from theoretical predictions based on standard random graph models. In particular, we observed consistently larger-than-expected average path lengths and a greater-than-expected tendency to cluster. Such deviations suggest that the consumers' product choices are not random even with the consumer and product attributes hidden. Our findings provide justification for a large family of collaborative filtering-based recommendation algorithms that make product recommendations based only on previous sales transactions. By analyzing the simulated consumer-product graphs generated by models that embed two representative recommendation algorithms, we found that these recommendation algorithm-induced graphs generally provided a better match with the real-world consumer-product graphs than purely random graphs. However, consistent deviations in topological features remained. These findings motivated the development of a new recommendation algorithm based on graph partitioning, which aims to achieve high clustering coefficients similar to those observed in the real-world e-commerce data sets. We show empirically that this algorithm significantly outperforms representative collaborative filtering algorithms in situations where the observed clustering coefficients of the consumer-product graphs are sufficiently larger than can be accounted for by these standard algorithms.

Key words: random graph theory; consumer-purchase behavior; topological features; recommender systems; collaborative filtering

History: Accepted by Brian Uzzi and Luis Amaral, special issue editors; received September 8, 2004. This paper was with the authors 5 months for 2 revisions.

1. Introduction

The past few years we have witnessed dramatic growth in research using random graph theory to study complex systems in a wide variety of scientific, engineering, and social domains. Examples include telecommunication networks and the World Wide Web (Faloutsos et al. 1999); biological systems such as genetic and metabolic regulation, protein folding, and neural networks (Amaral et al. 2000, Jeong et al. 2001); scientific literature coauthorship and citation networks (Barabási et al. 2002, Newman et al. 2002); and social networks of collaborators and acquaintances (Amaral et al. 2000, Watts 1999). Various graph-generation mechanisms have been explored to explain the topological structure of these real-world networks and to predict global and local network features—e.g., the robustness of the Internet (Albert et al. 2000) and the creativity and performance of teams in collaboration networks (Guimerà et al. 2005, Uzzi and Spiro 2005).

Several recent studies have adopted the methodology of random graph modeling to study phenomena in the business world, including interlocking boards of directors (Conyon and Muldoon 2004, Davis et al. 2003, Robins and Alexander 2004), corporate governance and ownership (Baum et al. 2003, Kogut and Walker 2001), and electronic bidding behavior (Yang et al. 2003). These studies have focused on characterizing the statistical topological properties of the business networks and linking these properties to underlying individual and organizational behavioral explanations. Many of these studies have shown that random graph modeling methodology can be fruitfully applied to business problems and can bring about useful insights. Two examples of such findings are the resilience of the structure of the corporate elite to macro and micro changes affecting corporate governance (Davis et al. 2003), and the domination of online auctions by an unusually active minority (Yang et al. 2003).

This paper applies random graph modeling to study consumer purchase behavior in e-commerce settings. In particular, we are interested in studying the evolution of interactions among consumers and products reflected in online-sales transactions. It has been recognized in marketing research that observed consumer-purchase behavior is the key to predicting consumer behavior offline (Ehrenberg 1988) and online (Bellmann et al. 1999). By representing consumers and products as vertices, and sales transactions as edges linking consumer and product vertices, we view the entire transaction history as a growing consumer-product graph. We examine the structure of this consumer-product graph with an attempt to gain insights regarding the underlying mechanism that governs consumer-purchase behavior and improve predictive analysis of the consumers' future purchases.

Sales transactions, as a major form of interaction between consumers and products, have also been extensively studied computationally in recommender system research (Resnick and Varian 1997). As an important component of the marketing and customer relationship management process, a recommender system suggests products and services to potential customers based on the observed customer behavior and the customer and product attributes. Sales transaction data are a major input to many algorithmic engines for commercial recommender systems and personalization systems (Huang et al. 2004, Schafer et al. 2001). In fact, the collaborative filtering approach (Hill et al. 1995, Resnick et al. 1994), arguably the most successful recommendation approach, relies on transaction data alone to make recommendations. The success of such a recommendation approach relies on the existence of consistent consumer-purchase patterns reflected in the sales transactions alone (as opposed to the consumer demographic and product attribute information) and on how well the algorithms can capture such patterns. Our research is aimed at exploring the use of random graph modeling of sales transactions to reveal the characteristics of consumer-behavior patterns and to improve the performance of recommender systems.

The intended contributions of this paper are two-fold. First, to the best of our knowledge, we present the first empirical study of consumer-product graphs using the complex systems/random graph analysis methodology. Using two online retail data sets, we compare measures of topological characteristics of consumer graphs and product graphs (projected from consumer-product graphs representing the sales transactions) with the theoretical predictions given by a random graph model. We find that these consumer and product graphs show a consistently higher tendency to cluster over time and have

path lengths or distances that are relatively short, but still longer than those predicted by the random graph model. These findings strongly suggest that the product choices of a community of consumers are not random, and in turn empirically confirm the fundamental assumption of recommender system research. Further simulation-based investigation embedding two popular recommendation algorithms as the graph-generation mechanism shows that such algorithms can lead to graphs that provide better matches with real-world consumer-product graphs than those produced by the purely random model. However, significant deviations remain. Our second contribution is a new graph partitioning-based recommendation algorithm motivated by the empirical and simulation-based findings summarized above. Our experimental study using the real-world e-commerce data sets shows empirically that this algorithm significantly outperforms representative collaborative filtering algorithms in situations where the clustering coefficients of the consumer-product graphs are sufficiently larger than can be accounted for by these standard algorithms.

The remainder of this paper is structured as follows. In §2, we provide a brief review of the two streams of literature relevant to this paper. Section 2.1 reviews the literature on random graphs and introduces topological measures and random graph models that are relevant to our research. Section 2.2 briefly surveys the recommender system literature and discusses representative recommendation algorithms relevant to our research. Section 3 describes the consumer-product graph representation of sales transactions and illustrates how to project this consumer-product graph into consumer and product graphs. We also summarize theoretical predictions of the topological measures of the consumer and product graphs under the assumption of a randomly generated consumer-product graph. In §4, we present the empirical and simulation-based findings on consumer-product graphs using two online retail data sets. In §5, we discuss conceptual implications of our findings in the context of recommender systems. These findings motivated the development of a new graph-based recommendation algorithm. Section 5 contains a detailed discussion of this algorithm and related computational evaluation. We conclude the paper in §6 by summarizing our findings and pointing out future research directions.

2. Research Background and Related Work

2.1. Related Random Graph Modeling Research

Graphs have been used to represent various types of relationships in a wide range of complex systems.

As is common in the random graph modeling literature, we use graph and network interchangeably in this paper. Random graph modeling research exploits a graph representation of complex systems and is aimed at capturing the mechanisms that determine the network topology of such systems. The key assumption is that the fundamental mechanism that governs the generation of relationships among components of a system leaves certain identifiable traits in the resulting network topology. Thus, a simple graph-generation model that can reproduce similar topological features of the real network may bring important insights to understanding the actual mechanism that governs the real system. The domain-specific interpretations of such abstract graph-generation models could potentially lead to development of theories regarding the interactions among the system constituents.

Many recent studies show that real-world networks demonstrate surprisingly consistent topological characteristics across different domains (Albert and Barabási 2002). Three major concepts related to such topological features are *small world*, *clustering*, and *scale-free phenomena*.

Small World. The small world concept describes the fact that despite their often large size, most networks exhibit a relatively short path between any two vertices. The distance between two vertices is defined as the number of edges along the shortest path connecting them. The *average path length* (or typical/characteristic distance) measure L , defined as the average of the path lengths of all connected vertex pairs, quantifies this property.

Clustering. Many real-world networks show an inherent tendency to cluster. A typical example is social networks, in which cliques form, representing circles of friends or acquaintances in which every member knows every other member. Such a tendency is quantified by the clustering-coefficient measure (Newman et al. 2001, Watts and Strogatz 1998). We adopt the Newman definition:

$$C = \frac{3 \times (\text{number of triangles in the graph})}{\text{number of connected triples}}, \quad (1)$$

where a triangle is a set of three vertices, each of which is connected to both of the others, and a connected triple is three vertices x - y - z , with both vertices x and z connected with y (note that x - y - z and x - z - y are considered the same connected triple). The factor 3 in the numerator accounts for the fact that each triangle contributes to three connected triples of vertices. The clustering coefficient C is strictly bounded between 0 and 1 and measures the extent to which being a neighbor is a transitive property. In our context, for example, a consumer graph represents relationships between consumers who purchase the same

products. In a consumer graph with a high clustering coefficient (close to 1), such a copurchase relationship tends to be transitive under most cases, i.e., if consumers a and b purchase the same products and consumers b and c purchase the same products, then consumers a and c are highly likely to do so as well.

Scale-Free Phenomena. The scale-free property is linked to the degree distribution of a graph. The *degree* of a vertex in a graph is the number of edges incident on that vertex. We define p_k , known as the degree distribution of the graph, to be the probability that a vertex chosen uniformly at random has degree k (i.e., the fraction of vertices that have degree k). Scale-free graphs refer to graphs with power-law degree distributions as described by (2):

$$p_k \sim k^{-\alpha}, \quad (2)$$

where α is a positive constant. Power-law degree distributions have been observed in a wide range of networks, including many of the real networks mentioned previously.

Various graph-generation models have been explored to explain empirically observed topological characteristics of real-world networks. The classic random graph model developed by Paul Erdős and Alfred Rényi (Erdős and Rényi 1959), called the *ER model*, studies a graph as N labeled nodes connected by n edges that are chosen randomly from $N(N-1)/2$ possible edges. In the ER model, the generation of the graphs is conditional only on the size of the graph and the vertex connection probability (or number of edges). Despite its simplicity, the ER model serves as the foundation for the analysis of statistical properties of graphs generated by a wide range of probabilistic graph-generation models. Although these models assume certain nonrandom principles governing the graph-generation process, they still contain certain random elements that differentiate them from deterministic graph models. In the literature, the term *random graph* is sometimes used to refer to the purely random ER model. For simplicity, in this paper we use this term in a broader sense to refer to both the purely random ER model and newer models that incorporate various types of nonrandom graph-generation mechanisms.

An important extension to the ER model, known as the *configuration model* (Newman 2003), studies random graphs generated according to various mechanisms conditional on the degree distribution (Newman et al. 2001). The main ideas behind this model are described as follows. We first specify a degree distribution p_k and then choose a *degree sequence*, which is a set of N values of the degrees k_i of vertices from this distribution, where $i = 1, \dots, N$, and N is the number of vertices in the graph. The configuration model is

thus a random graph model consisting of the ensemble of all graphs with this given degree sequence. The difference between the predictions of the ER model and the configuration model can be interpreted as the effect on the graph topology determined by the given degree distribution as compared to the expected Poisson degree distribution following the purely random vertex connection under the ER model. Furthermore, theoretical predictions of various topological characteristics of the configuration model are known (Newman et al. 2001).

In addition to incorporating additional information about the actual graph, the random graph model has also been extended in other dimensions. For instance, one of the key findings of complex systems research is that many real networks demonstrate the small-world property: a small average path length coexisting with a large clustering coefficient. This small-world property deviates from the theoretical predictions made by the ER model, i.e., *small* average path lengths and *small* cluster coefficients, and motivated the development of the small-world model (Watts and Strogatz 1998). Based on the small-world model, graphs are generated by randomly rewiring a regular graph (a graph is said to be regular of degree r if each vertex has the same degree r). This model is essentially a hybrid model that combines a completely random graph and a completely regular graph to reproduce the observed small-world property. A large number of follow-up studies have investigated similar hybrid graph-generation models. An important family of such models, called *scale-free networks*, focuses on reproducing the empirically observed power-law degree distribution (Barabási and Albert 1999, Dorogovtsev and Mendes 2001).

2.2. Recommender Systems Research

Recommender systems automate the process of suggesting products, services, and information items to potential consumers and are increasingly being adopted by major e-tailers such as Amazon, Half.com, CDNOW, Netflix, and Yahoo!. Such systems have been acknowledged to help increase online and catalog sales and improve customer loyalty (Schafer et al. 2001). The overall marketing power of these systems is also well recognized (Gladwell 1999, Vrooman et al. 2002). Many software companies provide generic recommendation technologies, with the top five providers (NetPerceptions, Epiphany, Art Technology Group, BroadVision, and Blue Martini Software) reaching a combined market capital of over \$600 million as of December 2004 (based on information from finance.yahoo.com).

Significant academic interest has been devoted to recommender system-related research issues. Our recent search on the Science Citation Index resulted

in around 315 journal publications on recommender systems from 1997 through 2004. A similar search on the ACM Digital Library revealed that close to 400 conference articles were published on recommender systems during the same time period. At the heart of recommender systems are the algorithms for making recommendations. Researchers and practitioners have investigated and experimented with a variety of recommendation approaches, taking three types of data as input: product attributes, consumer attributes, and previous interactions between consumers and products (including purchase, rating, and other types of interaction).

One of the most commonly used successful recommendation approaches is the *collaborative filtering approach* (Hill et al. 1995, Resnick et al. 1994, Shardanand and Maes 1995), which utilizes only consumer-product interaction data in the form of historical sales-transaction data and ignores consumer and product attributes. One basic collaborative filtering algorithm is the *user-based neighborhood algorithm*. To predict the potential interests of a given consumer, this approach first identifies a set of similar consumers based on past transactions and product feedback information and then makes a prediction based on the observed behavior of these similar consumers.

Because collaborative filtering algorithms utilize only consumer-product interactions to generate recommendations, they can be viewed as a predictive application based on the consumer-product graph representation of the past sales, taking as input a consumer-product graph at a certain time point and trying to recommend candidate product vertices for individual consumers to form future edges. In other words, the collaborative filtering problem can be recast as the one that predicts the future state(s) of the consumer-product graph conditional on the current graph (and possibly past ones).

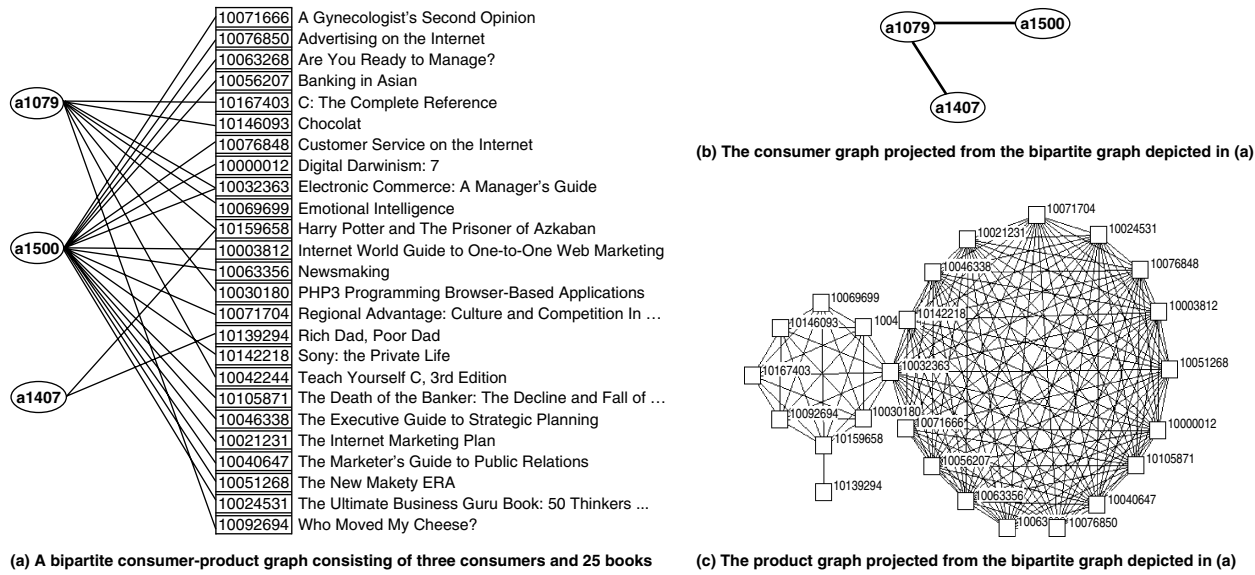
3. Modeling the Consumer-Product Graph

In this section, we first present the consumer-product graph representation of the sales-transaction data and its projections to a *consumer copurchase graph* and a *product copurchased-by graph*. We then summarize the known analytical results concerning these projected graphs' topological measures.

3.1. Graph Representation of Sales Transactions

A sales-transaction data set can be naturally represented as a graph by treating consumers and products as vertices and transactions involving consumer-product pairs as edges. We refer to this graph as the *consumer-product graph*, which is a bipartite graph

Figure 1 A Book Sales Example: Consumer-Product Graph, Consumer Graph, and Product Graph



consisting of two disjoint sets of vertices and no edges connecting vertices in the same set. In the social network literature, this type of graph is also called an *affiliation network*, whose edges represent affiliation relations (e.g., Davis et al. 2003, Robins and Alexander 2004). Figure 1(a) shows a small portion of a consumer-product graph constructed from a real-world sales-transaction data set from an online bookstore.

Because most of the existing theoretical results concerning graph-topological characterization are on unipartite graphs, most previous empirical studies have projected a bipartite graph into two unipartite graphs, each with only one type of vertex, and have analyzed the topological properties of these projected unipartite graphs. We adopt a similar approach in our study. In our context, a bipartite consumer-product graph is projected into a consumer graph and a product graph. In the consumer graph, an edge between two consumer vertices represents that the two consumers had purchased at least one common product previously (copurchase relationship). Similarly, in the product graph, an edge between two product vertices represents that the two products had previously been purchased by at least one common consumer (copurchased-by relationship). Figures 1(b) and 1(c) present the unipartite projections of the bipartite graph shown in Figure 1(a), one for the consumers and the other for the products.

3.2. Random Bipartite Graphs and the Generating-Function Method

Previous studies have investigated graphs projected from bipartite graphs that are similar to the consumer and product graphs in our context, such as the movie

actor collaboration network (Watts and Strogatz 1998), scientific collaboration network (Barabási et al. 2002), word co-occurrence network (Cancho and Sole 2001), and board of directors network (Davis et al. 2003). Most of these networks exhibit drastically larger clustering coefficients than the expected values predicted by the random unipartite graph model, typically over 1,000 times larger. However, as we can see in the example consumer and product graphs in Figures 1(b) and 1(c), the clustering tendency of a projected graph can to a large extent be attributed to the projection process itself. Each consumer who has purchased multiple products will result in a fully connected complete subgraph in the projected product graph; therefore, the product graph is actually comprised of (potentially overlapping) complete subgraphs of this kind. In general, a unipartite graph projected from a bipartite graph is guaranteed to have larger clustering coefficients than a random unipartite graph of the same size and number of edges. For this reason, the predictions given by the unipartite random graph model are not appropriate benchmarks for such projected unipartite networks (Conyon and Muldoon 2004). Newman et al. (2001) introduced the generating-function approach, which gives theoretical predictions of the topological measures of graphs projected from a random bipartite graph following the given degree distributions for the two types of vertices. We adopt this bipartite configuration model-based approach in our study to compare empirically observed consumer-product graphs with such random graphs conditional on consumer and product degree distributions. Note that ideally the topological features of the bipartite consumer-product graphs should be directly studied, as opposed to those of the

projected unipartite consumer and product graphs, because the projection process results in information loss. However, to the best of our knowledge, no such random bipartite graph theory exists in the literature. In this paper, we use the projection-based approach exclusively.

We summarize the theoretical predictions by the generating-function method below. Given the degree distribution of a graph v_k , which describes the probability of a randomly chosen vertex to have k neighbors, one can construct a *probability-generating function*, $G(x)$, of the graph:

$$G(x) = \sum_{k=0}^{\infty} v_k x^k. \quad (3)$$

Newman et al. (2001) derived the theoretical predictions of the topological measures of a random graph conditioned on a given degree distribution using the generating function defined in (3). These topological measures include average degree, average path length, and clustering coefficient. They have also shown that the generating functions corresponding to the theoretical degree distribution of the graphs projected from a random bipartite graph can be derived from the two generating functions associated with the degree distributions of the two types of vertices. The theoretical predictions of the topological measures of the unipartite graphs projected from a bipartite graph can then be derived from such a generating function.

In our context, a bipartite consumer-product graph has two empirical degree distributions, from which two generating functions can be constructed. One, denoted as $f_0(x)$, generates the degree distribution for consumers:

$$f_0(x) = \sum_{j=0}^{\infty} p_j x^j, \quad (4)$$

where p_j denotes the frequency with which one finds that a consumer purchased j distinct products. The other function for products, denoted as $g_0(x)$, is defined similarly:

$$g_0(x) = \sum_{k=0}^{\infty} q_k x^k, \quad (5)$$

where q_k is the frequency with which one finds a product being purchased by k distinct customers.

Let $G_0(x)$ denote the generating function for the theoretical degree distribution of the projection onto the consumer graph. Newman et al. (2001) show that it is given by

$$G_0(x) = f_0\left(\frac{g_0'(x)}{g_0'(1)}\right). \quad (6)$$

The corresponding theoretical predictions of average degree z , average path length L , and triangle clustering

coefficient C are given by

$$z = G_0'(1), \quad L = 1 + \frac{\ln(N/G_0'(1))}{\ln\left(\frac{f_0''(1)}{f_0'(1)}\right)\left(\frac{g_0''(1)}{g_0'(1)}\right)}, \quad (7)$$

$$C = \frac{M g_0''(1)}{N G_0''(1)},$$

where M is the total number of products and N is the total number of customers. The predictions for the product graph can be derived similarly by interchanging f and g , and then M and N in (7).

4. An Empirical Study

4.1. Data Sets

We used two e-commerce data sets in our study: a book data set provided by one of Taiwan's largest online bookstores and a retail data set provided by the online division of a U.S.-based company in the apparel industry. The sales transactions in both data sets contained time stamps, which enabled us to reconstruct the evolution of the consumer-product graph by examining the "snapshots" of the graph at a series of time points (t_1, \dots, t_K) . Consistent with typical marketing studies on sales prediction and recommendation research, we selected a set of consumers who made their initial purchases in a predetermined observation time period (from t_1 to t_T , $T < K$) to form the cohort for our analysis (Fader and Hardie 2001, Jain and Vilcassim 1991). The transaction data involving this set of consumers in the observation time period typically are used as the training set for estimation or learning purposes, whereas the transaction data for these consumers in the remaining time periods are used as the testing set to evaluate the predictive performance of the model. For analysis purposes, we constructed a consumer-product graph at each time point t_k ($k = 1, \dots, K$) by including transactions involving these consumers before t_k . The characteristics of the samples we used from the two data sets for our analysis are presented in Table 1.

4.2. Comparing Actual and Random Graphs Given Consumer/Product Degree Distributions

In our analysis, we used Equation (7) to calculate expected average degrees, average path lengths, and clustering coefficients of the two projections of a random bipartite graph with given consumer/product degree distributions at different time periods. These consumer and product degree distributions were computed directly from the sales transactions. We also calculated the actual topological measures of the projected consumer and product graphs. The deviations of the actual values from the expected values of the three topological measures would indicate the

Table 1 Data Samples from the Book and Retail Sales-Transaction Data

Data set	Number of consumers	Total number of products	Total number of transactions	Time span	Training time period (T)	Testing time period ($K - T$)
Book	1,000	7,667	12,314	04/1998–08/2001	36 months	4 months
Retail	1,000	7,175	9,047	10/2002–12/2002	9 weeks	4 weeks

presence of certain nonrandom principles that govern the consumers' choices. More importantly, the direction and magnitude of these deviations may shed light on the nature of such governing principles and could potentially be exploited for predicting consumers' future purchases.

Figures 2 and 3 show selected topological measures of the consumer and product graphs projected from the actual and random bipartite consumer-product graphs: average degree, average path length, and clustering coefficient. The graphs corresponding with the early time points were so sparse that certain measures were not well defined. For example, the clustering coefficient in Equation (2) is meaningless for a graph that contains no connected triples. Thus, in Figures 2 and 3 we included only graphs of the last several time periods (last 20 months for the book data set and last six weeks for the retail data set) that were sufficiently dense to assure well-defined measures and robust theoretical predictions. We projected

the consumer-product graphs at each time point t_k to the corresponding consumer and product graphs, as illustrated in the example shown in Figure 1. The actual topological measures were computed based on these projected unipartite graphs. In Figures 2 and 3, we also show the theoretical predictions of the topological measures obtained from Equation (7), taking as input the consumer and product degree distributions of the actual consumer-product graphs.

In Figures 2 and 3, we observe that the topological measures of the actual and random consumer-product graphs vary widely across the data sets and time periods due to the changing consumer and product degree distributions. For instance, the drastic increases in the average degree and clustering coefficient of the consumer graph for the book data set at the 12th and 18th months were due to surges in sales for two popular *Harry Potter* books. However, consistent patterns regarding the differences between the topological measures of the random and actual

Figure 2 Topological Measures of the Projections of the Actual and Random Consumer-Product Graphs: Book Data Set, Last 20 Months

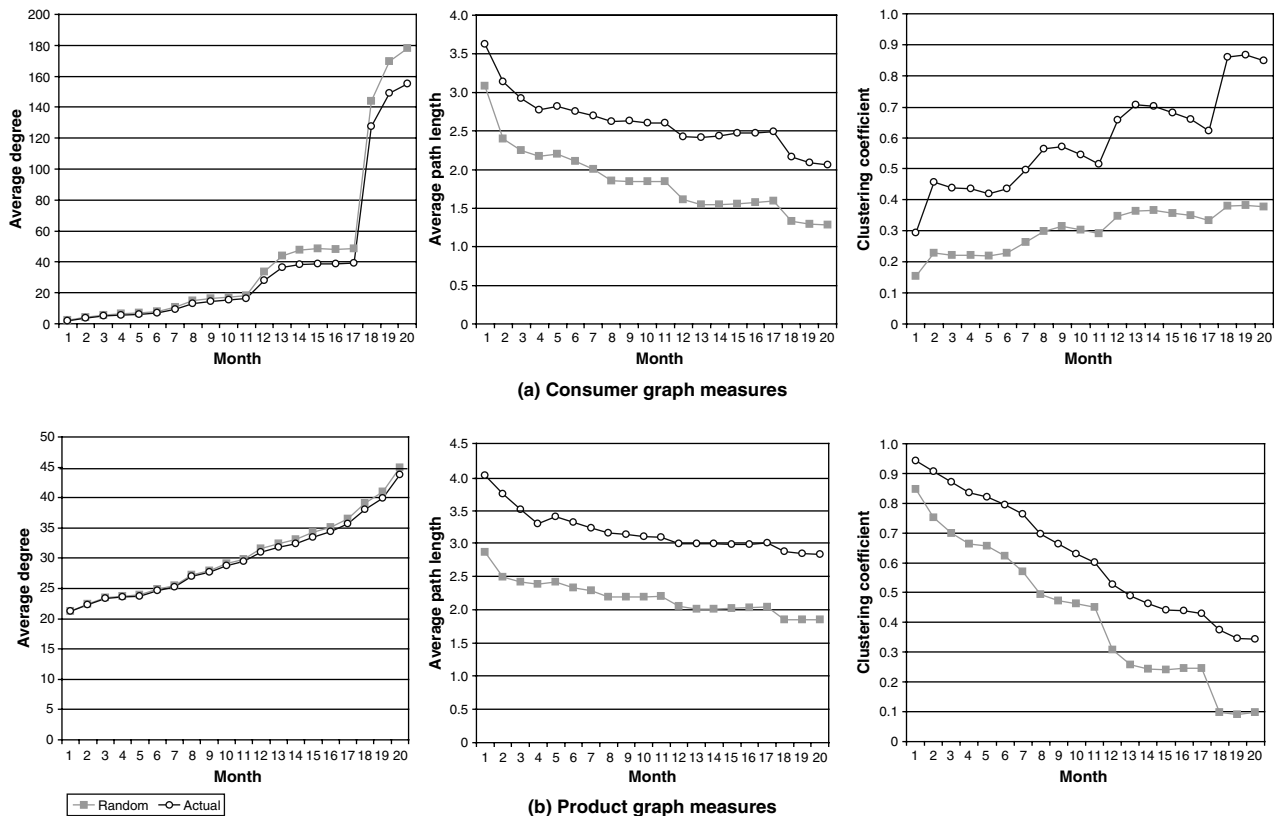
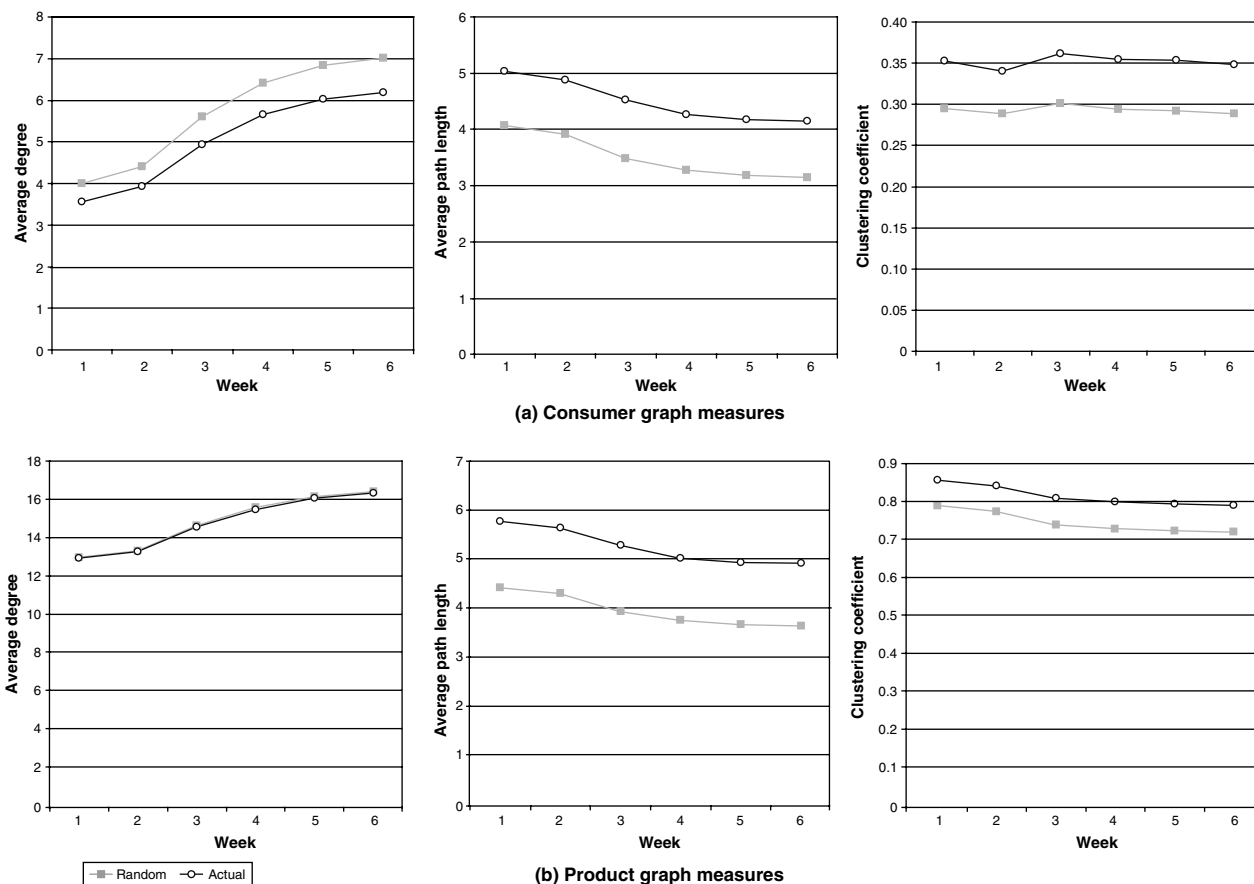


Figure 3 Topological Measures of the Projections of the Actual and Random Consumer-Product Graphs: Retail Data Set, Last Six Weeks



graphs were observed.¹ The consumer graphs and product graphs of both data sets exhibited a substantially larger average path length and clustering coefficient than those of the random graphs, while the average degree measures were more consistent with the random graph predictions. Such patterns were consistent over all the time periods included in the analysis. For both data sets, the topological measures for the consumer graphs showed larger deviation from the random graph predictions than those for the product graphs. Overall, we observed that the topological measures of the book-sales graphs exhibited much larger deviation from the expected values than those of the retail graphs. In particular, for the book data set at the last time period, the clustering coefficients of the actual consumer and product graphs were about 2.4 times larger than those predicted by the random graph model; the average path

lengths of the actual consumer and product graphs were about 1.5 times larger than the predictions. The consumer graph also showed significant deviation for the average degree measure, with the actual graph measuring about 30% smaller than that of its random counterpart.

4.3. Recommendation Algorithm-Induced Graphs

In addition to the actual and random consumer-product graphs, we also studied graph models that directly *embed* recommendation algorithms as the underlying graph-generation mechanism. We refer to them as *recommendation algorithm-induced graphs* or simply *induced graphs*. Comparing the topological features of these graphs with the actual and the random consumer-product graphs can reveal whether and to what extent the deviations of the actual consumer-product graphs from the purely random graphs can be explained by the assumptions driving the design of various recommendation models. In this section, we first introduce two representative algorithms to be incorporated in the graph-generation process: the standard user-based neighborhood algorithm (or simply the user-based algorithm) and the generative model algorithm that is closely related to

¹ The theoretical prediction concerning the variances of the topological measures of random graphs is yet to be developed. We observed from our simulation results that the realized sample variances are relatively small. Furthermore, nonparametric statistical tests confirm that all differences in average path length and clustering-coefficient measures discussed in this paper are statistically significant.

the modeling of unobserved consumer/product heterogeneity in the marketing literature (Allenby and Ginter 1995, Ansari et al. 2000, Gershoff and West 1998, Rossi et al. 1996). We then describe a generic, simulation-based graph-generation model that can embed any given recommendation algorithm and report the findings on the topological features of the graphs induced by these two recommendation algorithms relative to those of the random and actual consumer-product graphs.

4.3.1. Recommendation Algorithms. Collaborative filtering algorithms take a consumer-product graph as input. Such a graph can be equivalently represented by a consumer-product *interaction matrix* $A = (a_{ij})$, with M rows representing the consumers $C = \{c_1, c_2, \dots, c_M\}$ and N columns representing the products $P = \{p_1, p_2, \dots, p_N\}$. Element a_{ij} takes the value of one if there is an edge (transaction) between consumer i and product j , and zero otherwise. The output of a collaborative filtering algorithm is an $M \times N$ matrix $Z = (z_{ij})$ storing potential scores of products for individual consumers, representing the likelihood of future transactions. A wide range of collaborative filtering algorithms have been investigated in the literature, including the standard user-based and item-based neighborhood algorithms (Herlocker et al. 2004, Hill et al. 1995, Resnick et al. 1994, Shardanand and Maes 1995), cluster and generative models (Hofmann 2004, Kumar et al. 1998, Ungar and Foster 1998), rule-based approaches (Adomavicius and Tuzhilin 2001, Lin et al. 2002), and advanced matrix analysis approaches (Azar et al. 2001, Goldberg et al. 2001, Sarwar et al. 2000). For our study, we focus on two representative algorithms: user-based and generative model-based.

User-Based Algorithm. The user-based algorithm predicts a target consumer's future transactions by aggregating the observed transactions of *similar* consumers. The algorithm first computes a consumer similarity matrix $WC = (wc_{st})$, $s, t = 1, 2, \dots, M$. The similarity score wc_{st} can be calculated using a vector similarity function based on the corresponding row vectors of A (Breese et al. 1998). A high similarity score wc_{st} indicates that consumers s and t have similar preferences or tastes because they have previously purchased many common products. The matrix product $WC \cdot A$ then gives potential scores of the products for each consumer based on which recommendation can be generated. In essence, the element at the c th row and p th column of $WC \cdot A$ aggregates the similarity scores between consumer c and other consumers who have purchased product p previously. The underlying hypothesis of the user-based algorithm is that consumers naturally form homogeneous segments with consistently correlated preferences and that consumer preferences are adequately expressed

in the sales-transaction data. The user-based algorithm provides an automatic mechanism to identify preference correlations and utilize the segment homogeneity to make recommendations.

Generative Model-Based Algorithm. Under this approach, latent class variables are introduced to explain the patterns of interactions between consumers and products (Hofmann 2004, Ungar and Foster 1998). Typically, one can use one latent class variable to represent the unknown cause that governs the interactions between consumers and products. The interaction matrix A is considered to be generated from the following probabilistic process: (1) select a consumer c with probability $P(c)$; (2) choose a latent class with probability $P(z | c)$; and (3) generate an interaction between consumer c and product p (i.e., setting a_{cp} to 1) with probability $P(p | z)$. Thus, the probability of observing an interaction between c and p is given by $P(c, p) = \sum_z P(c)P(z | c)P(p | z)$. Using A as training data, all relevant probabilities and conditional probabilities are estimated using a maximum-likelihood procedure called the *Expectation Maximization (EM) Algorithm* (Dempster et al. 1977). Based on the estimated probabilities, $P(c, p)$ can be computed to give the potential score of product p for consumer c .

The generative model algorithm is closely related to the marketing literature on modeling of unobserved consumer heterogeneity using scanner panel data (Allenby and Ginter 1995, Rossi et al. 1996). Latent class variable approaches were also applied by several marketing studies in this context (Kamakura and Russell 1989, Kamakura et al. 1994). Most studies in the marketing literature have focused on analyzing sales transaction histories of panel households on a relatively small number of products or product brands that are well described by observed attributes. However, in e-commerce contexts, recommender systems are typically applied to a large number of products (e.g., movies, books, and music products) that are difficult to describe adequately using a few observable attributes. Two recent marketing studies have explicitly analyzed unobserved product heterogeneity (Gershoff and West 1998) and the combination of unobserved consumer and product heterogeneities (Ansari et al. 2000) using recommendation data sets. The preference and appeal structure heterogeneities are typically modeled using both observed and unobserved consumer/product attributes. Note that, in our context, generative model-based recommendations are solely based on the sales-transaction data and the consumer and product heterogeneities are captured entirely by their associations with the latent class variable, $P(z | c)$ and $P(p | z)$.

4.3.2. Graph Generation Based on Recommendation Algorithms. In this section, we present a generic approach to embed any given recommendation algo-

rithm into the consumer-product graph-generation process. This induced graph approach is inspired by the configuration model introduced in §2.1. All of the input items assumed by the configuration model, i.e., the number of edges (transactions) and the consumer/product degree sequences of the actual consumer-product graph at each time period t , are kept in this approach. The key difference between the original configuration model and our model is as follows. Instead of randomly generating the graphs at each time period t according to the given number of edges and degree sequences, we start with a randomly generated graph G_0 following the configuration model for the time period t_0 just before the start of the analysis period. The graph G_k at each time period t_k within the analysis period is then generated by populating additional edges to “grow” G_{k-1} into G_k , while restricting G_k to have the given consumer/product degree sequences at time t_k . When selecting edges to add to G_{k-1} , the potential scores of consumer-product pairs computed by the given recommendation algorithm using G_{k-1} as training data are used to determine the probability of each possible edge to be selected. Note that the potential scores given by the recommendation algorithm are only available for pairs involving existing consumers and products appearing in G_{k-1} . For each existing consumer, we reset the potential scores of existing products with zero-valued potential scores, and those of new products to a small nonzero positive number. We then normalize all potential scores to determine the probability of selecting a potential product to form new edges. For a new consumer, we use a uniform probability distribution to select products randomly to form new edges. The ensemble of all the graphs generated in this manner is the universe of the graphs with exactly the same degree sequences as the actual graph, but whose growth are governed by the given recommendation model. The graph-generation process is still random except as to the number of edges, the consumer/product degree distributions, and a recommendation model. Following the philosophy of random graph modeling methodology, if the actual consumer-product graph-generation process follows the recommendation model, the actual graph is expected to exhibit topological features similar to these induced graphs. This recommendation algorithm-induced graph-generation model is summarized as follows.

Recommendation Algorithm-Induced Graph Generation Model

Input: $G_{t-1} = \langle C_{t-1}, P_{t-1}, E_{t-1} \rangle$, consumer and product vertex degree sequences of G_t : $D_t^C = \{k_1^t, \dots, k_{M_t}^t\}$ and $D_t^P = \{g_1^t, \dots, g_{N_t}^t\}$, and a recommendation algorithm $r: G \rightarrow Z(M \times N)$.

Output: G_t^* such that G_{t-1} is a subgraph of G_t^* and G_t^* has the same degree sequences as G_t .

Step 1. Initialize G_t^* to be $\langle C_t, P_t, E_t^* \rangle$, where $C_t = \{c_1, \dots, c_{M_t}\}$, $P_t = \{p_1, \dots, p_{N_t}\}$, and $E_t^* = E_{t-1}$.

Step 2. Compute the potential score matrix: $Z^{t-1} = r(G_{t-1})$.

Step 3. Repeat for each consumer $c_i \in C_t$.

3.1. If $c_i \in C_{t-1}$

3.1.1. $z_{ij}^{t-1} = \varepsilon$ if $z_{ij}^{t-1} = 0$ or $j > M_{t-1}$, where ε is a small positive value.

3.1.2. Compute the connecting probability $p_{ij} = z_{ij}^{t-1} / \sum_j z_{ij}^{t-1}$.

3.2. Else

3.2.1. $p_{ij} = 1/N_t$.

Step 4. Obtain the differential degree sequences: $D_\Delta^C = \{k_1^\Delta, \dots, k_{M_t}^\Delta\} = \{k_1^t - k_1^{t-1}, \dots, k_{M_{t-1}}^t - k_{M_{t-1}}^{t-1}, k_{M_{t-1}+1}^t, \dots, k_{M_t}^t\}$, $D_\Delta^P = \{g_1^\Delta, \dots, g_{N_t}^\Delta\} = \{g_1^t - g_1^{t-1}, \dots, g_{N_{t-1}}^t - g_{N_{t-1}}^{t-1}, g_{N_{t-1}+1}^t, \dots, g_{N_t}^t\}$.

Step 5. Repeat for each consumer $c_i \in C_t$.

5.1. Set $l = 0$.

5.2. Repeat while $k_i^\Delta > 0$.

5.2.1. Randomly choose a product $p_j \in P_t$ according to p_{ij} .

5.2.2. If $(c_i, p_j) \notin E_t^*$ and $g_j^\Delta > 0$

5.2.2.1. Add (c_i, p_j) into E_t^* , $k_i^\Delta = k_i^\Delta - 1$, $g_j^\Delta = g_j^\Delta - 1$.

5.2.3. Else

5.2.3.1. Set $l = l + 1$.

5.2.4. If $l < T$ (a predetermined control parameter)

5.2.4.1. Go to Step 5.1.1.

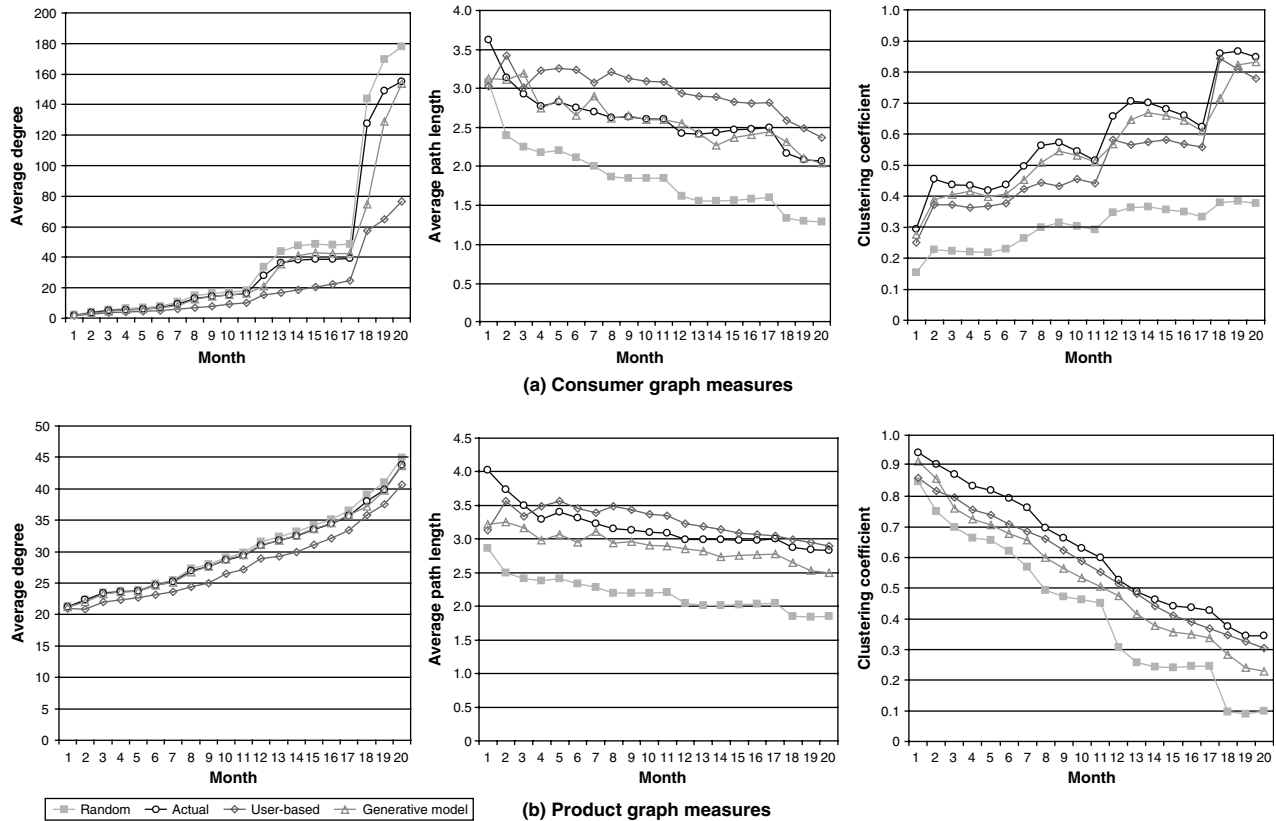
5.2.5. Else

5.2.5.1. Exit the loop and go to Step 1.

4.3.3. Empirical Findings. In Figures 4 and 5, we report the topological measures of the consumer and product graphs projected from the consumer-product graphs induced by the user-based and generative model recommendation algorithms, in addition to those projected from the random and actual graphs reported in §4.2.

From Figure 4, we observe that for the book data set, both consumer and product graph projections from the consumer-product graph induced by the generative model were generally the closest match with the projections from the actual consumer-product graph for all three topological measures under study. The user-based model-induced graph did not provide a good match with the actual graph, with substantially smaller average degree for the consumer graphs in the later time periods, slightly higher average path length, and slightly lower clustering coefficients for both the consumer and product graphs. For the product graph clustering coefficient, the user-based model-induced graph provided a slightly better match than the generative model-induced graph.

Figure 4 Topological Measures of Projections of the Actual, Random, and Recommendation Algorithm-Induced Consumer-Product Graphs: Book Data Set, Last 20 Months



In Figure 5, we observe that the user-based model-induced graph had the best overall match with the actual consumer-product graph, with slight deviations for the average degree and average path length measures, but still substantially smaller clustering coefficients for both the consumer and product graphs. The generative model-induced graph generally matched well with the random consumer-product graph.

Our empirical findings support the following general observations. (1) The graphs induced by recommendation algorithms typically match better with the actual graph than a random one, although the degree of matching is recommendation algorithm and data set specific. (2) The recommendation algorithm-induced graphs still consistently deviate from the actual graph. One of the consistent patterns across the recommendation models and data sets is that the actual consumer-product graph has a larger clustering coefficient than random and induced graphs. This pattern will be exploited in §5 to improve recommendation algorithm design.

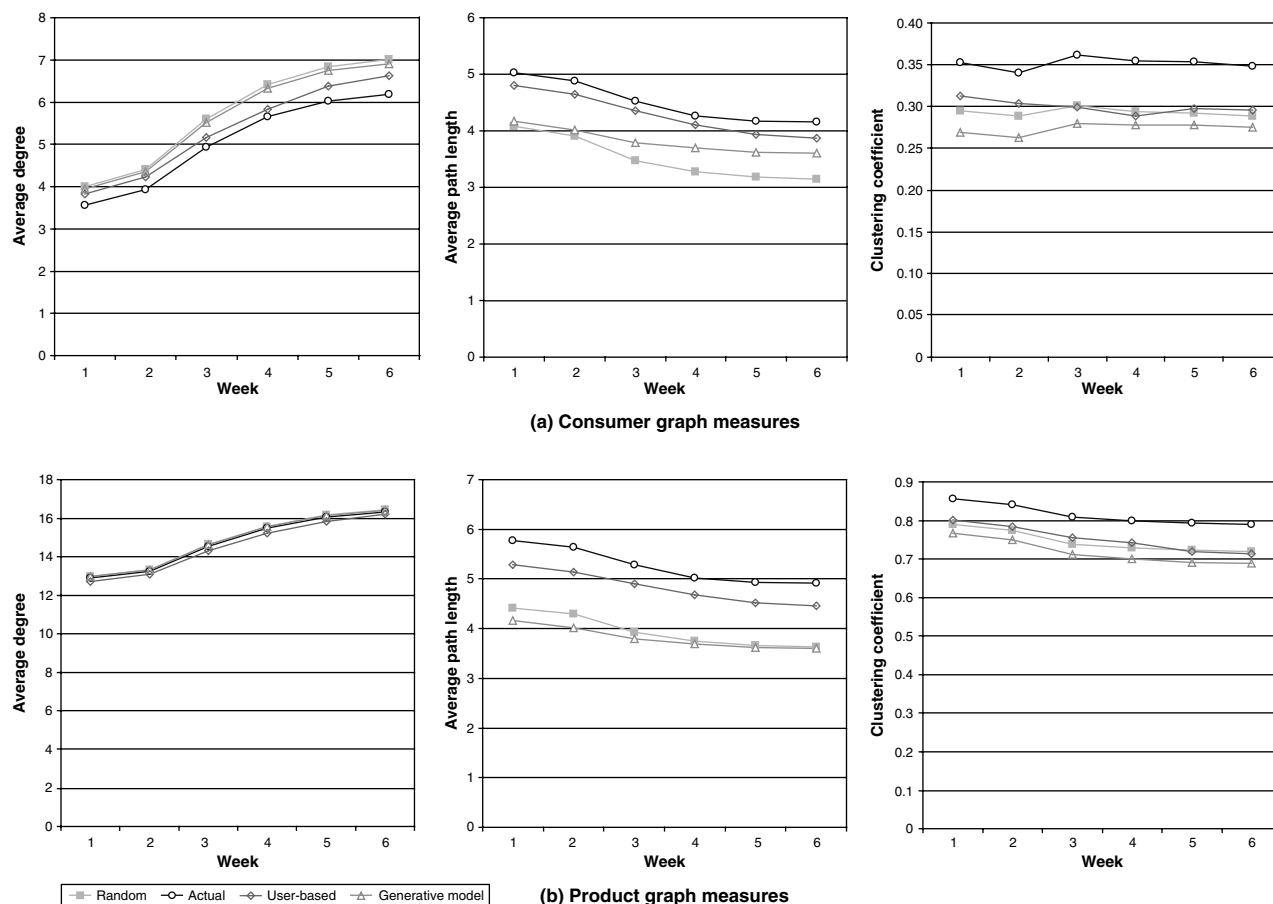
4.4. Discussion

Comparison of the actual topological measures and their expected values showed that the underlying

transaction-generation process governing the evolution of the consumer-product graph deviated significantly from a random process. Because the theoretical predictions based on the generating function method inherently related to the graph projection process and the actual consumer and product degree distributions, the observed deviations point to the existence of additional nonrandom underlying graph-generation principles. Note that graph-generation models relying on concepts such as preferential attachment (Barabási and Albert 1999, Dorogovtsev and Mendes 2001) are not relevant to our observations because the vertex degree distributions are specified exogenously in our case. The nonrandom phenomena in our findings are largely attributable to the characteristics of purchase choices and preference structures of repeat-buying consumers rather than the consumer and product sales distributions.

There exists a large body of marketing literature on consumer purchase behavior. However, most research has focused on modeling timing of repeat purchases (Cox 1972, Ehrenberg 1988, Jain and Vilcassim 1991), predicting aggregate sales for particular products (Bradlow and Fader 2001) or the entire store (Fader and Hardie 2001), or studying consumers' choices

Figure 5 Topological Measures of the Projections of the Actual, Random, and Recommendation Algorithm-Induced Consumer-Product Graphs: Retail Data Set, Last Six Weeks



among a small number of alternative brands or products (Jain et al. 1994, McFadden 1974). Our findings contribute to this literature by analyzing the sales transactions of a large number of consumers who repeatedly choose from a vast number of alternative products, which is typical in e-commerce settings. Although the exact underlying mechanisms that govern the growth of the consumer-product graph are yet to be discovered, such purchase choice and preference structure patterns may provide useful and actionable insights for applications relying on sales transaction data as input, including recommender systems. In the next section, we propose and evaluate a new graph-based algorithm motivated by our empirical findings on the topological characteristics of consumer-product graphs reported in this section.

5. An Application in Recommender Systems

Past and current random graph research has been largely exploratory or descriptive in nature. One of our intended contributions is to explore ways through which the findings based on random graph-modeling

methodology can be utilized to enable or improve business decision making in concrete settings. Recommender systems, in particular collaborative filtering systems, provide an ideal context for such an application. Section 5.1 discusses implications of our empirical findings reported in §4 to recommender systems. In §5.2, we present a preliminary design for a graph partitioning-based recommendation algorithm motivated by our empirical findings. Section 5.3 summarizes an experimental study that demonstrates the effectiveness of this new algorithm using the two real-world e-commerce data sets examined in §4.

5.1. Implications of Our Empirical Findings to Recommender Systems

Current research in collaborative filtering has been mainly focused on algorithm design (Sarwar et al. 2000, Hofmann 2004, Ungar and Foster 1998, Breese et al. 1998, Heckerman et al. 2000). Many previous studies have demonstrated the predictive power of these algorithms empirically, using various kinds of data sets. However, to the best of our knowledge, no well-grounded metalevel analysis has been conducted

to answer the fundamental question regarding the predictability of future interactions between consumers and products based on their previous interactions.

A key finding of our empirical analysis is that the underlying mechanism governing the generation of the consumer-product graph over time is not a purely random process. This general finding provides support for use of collaborative filtering-based recommendation algorithms. In fact, various collaborative filtering algorithms, the generative model algorithms being a prominent example (Hofmann 2004, Kumar et al. 1998, Ungar and Foster 1998), implicitly make certain assumptions without validation regarding the underlying transaction-generation process and exploit such assumptions or patterns to make predictions.

Random graph research offers a useful framework to validate such assumptions. Given a particular assumption, one can derive either analytically or numerically the expected topological measures for the growing consumer-product graph following the random graph modeling framework. In this derivation process, the other aspects of graph generation can be accounted for by the random process (see §2.1). Then, these expected values can be compared against the observed ones that are calculated using the real transactions to determine whether or to what extent the assumption holds true and the collaborative filtering algorithm based on such an assumption is suited for the application context under study.

As an example, generative model algorithms assume that consumers and products belong to certain latent types and that their interactions are governed by their types (Hofmann 2004, Kumar et al. 1998, Ungar and Foster 1998). Using the proposed random graph modeling framework, we can derive or simulate the evolution of the consumer-product graph under this assumption, taking as input observed consumer and product degree distributions at different time points, and obtain the expected values of the topological measures. The deviation of the actual topological measures from these expected ones would provide a useful indication as to the “fitness” between the assumption and the actual data-generation process. Such a framework can also explain in a principled manner why certain collaborative filtering approaches work better than others for specific data sets.

The proposed application of random graph modeling methodology to recommender system research as a metalevel model verification and algorithm selection mechanism is still preliminary. Nonetheless, it represents a potentially important initial step toward an underlying theory for recommender systems and provides new insights that could lead to effective new algorithms. In the next section, we present a new graph partitioning-based algorithm motivated by our

empirical findings as a concrete example of such an application.

5.2. A New Graph Partitioning-Based Recommendation Algorithm

Our algorithm is motivated by the larger-than-expected clustering coefficients of the consumer and product graphs observed in our empirical study. Large clustering coefficients indicate that consumers and products tend to form cliques in the graphs projected from the bipartite consumer-product graph. The work of Watts and Strogatz (1998) on small-world models has shown that the short average path lengths and large clustering coefficients are consistent with a hybrid random graph model consisting of both random and regular components (Watts and Strogatz 1998). Our empirical findings strongly suggest that the consumer-product graph is also a hybrid one. The tendency to form regular graphs of periodic patterns (e.g., the ordered lattices) may account for the large average path length and clustering coefficients we have observed.

Although a definitive theory to explain these empirical findings has yet to be developed, in this paper we attempt to directly exploit these findings algorithmically. Because neither adding edges randomly (§4.2) nor adding edges following (representative) existing recommendation algorithms (§4.3) can result in the observed clustering coefficients, we aim to develop a new algorithm that recommends new edges between consumer and product vertices (purchases), leading to the larger cluster coefficients of the projected graphs.

One interesting observation is that new edges given by the user-based algorithm do not change the clustering coefficients of the projected graphs. In fact, such an algorithm does not even change the projected graphs. Consider a simple example: Consumers A and B both purchased Product a , and Consumer B also purchased Product b . In the projected consumer graph, A is connected with B because of the common purchase of a . In the projected product graph, a is connected with b because B purchased both of them. Product b is likely to be recommended to Consumer A based on the user-based algorithm. However, the newly added edge b — A does not change the projected graphs. (The variation in clustering coefficients of the user-based algorithm-induced graphs in Figure 4 was due to the random component in the induced graph-generation model to assure that the degrees of newly added consumer and product vertices match the given configuration.) This observation suggests the need to go beyond the immediate neighbors of vertices to increase the clustering coefficient. The graph partitioning-based approach described below is one such method that explores neighborhoods of a more global nature.

We start with partitioning the consumer-product graph into H subgraphs. Graph partitioning is a well-studied problem that is concerned with dividing a graph into H disjoint partitions by cutting the minimal number of edges while maintaining a similar number of vertices in each partition (Bondy and Murty 1976). Such partitions with a similar number of vertices are referred to in the literature as *balanced partitions*. Immediate neighbors of consumer and product vertices are likely to fall into the same or nearby subgraphs or partitions. Based on these partitions, we devise a recommendation algorithm that encourages the generation of edges that are within strongly connected graph partitions to achieve high clustering coefficients.

We use a hypothetical example shown in Figure 6 to illustrate how this graph partitioning-based algorithm can lead to the high clustering tendency. Figure 6(a) shows a bipartite consumer-product graph, in which circles represent consumer vertices and blocks represent product vertices. This bipartite graph contains three balanced partitions, G_1 , G_2 , and G_3 . Suppose now that we need to make recommendations for Consumer A. Potential products to recommend include the ones that are not currently linked to Consumer A, i.e., products a , b , c , d , and e . Figure 6(b) shows the consumer graph projected from Figure 6(a). Figures 6(c)–6(g) show the consumer graphs projected from the consumer-product graph after connecting Consumer A with each of the recommendable products, respectively.

A critical observation is that the change in the clustering coefficients of the projected graphs before and after adding an edge depends on the *within-partition degrees* of the related vertices, and the *distance* and *connection strength* between the partitions to which the consumer and product vertices belong. The within-partition degree of a vertex is defined as the number of its neighbors that belong to the same partition. The notions of distance and connection strength between

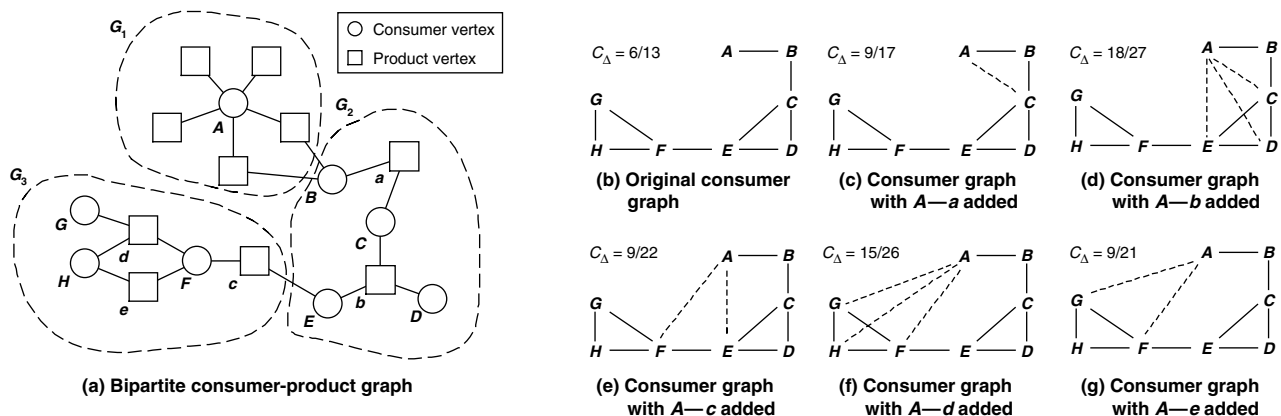
partitions are determined by the length and number of the partition-level paths connecting the two partitions. The distance measure is defined as the length of the shortest path connecting the two partitions. The connection strength measure is defined as the number of distinctive paths connecting the two partitions. In the example shown in Figure 6, G_2 and G_3 are considered zero and one partition away from G_1 , respectively. There are two edges connecting G_1 and G_2 and one edge connecting G_3 and G_2 and thus two partition-level paths from G_1 and G_3 . A criterion to rank unpurchased products p_j for consumer c_i can then be given as

$$r(i, j) = d_i u^{-D(G_i, G_j)/S(G_i, G_j)}, \quad (8)$$

where G_i and G_j denote the partitions to which c_i and p_j belong, respectively; d_i denotes the within-partition degree of p_j ; and $D(G_i, G_j)$ and $S(G_i, G_j)$ denote the partition-level distance and connection strength between G_i and G_j , respectively. Parameter $u > 1$ is a balancing parameter that trades off between distance and connection strength considerations. In the example shown in Figures 6(c)–6(g), using this criterion, we can reach the following ranked order with decreasing resulted clustering coefficients: adding A – b , adding A – d , adding A – a , adding A – e , and adding A – c . Setting the balancing parameter u to be any number between 1 and 2 will result in the same order as given above. In contrast, the standard user-based neighborhood algorithm will only recommend product a to Consumer A because only Consumer B is considered to be A’s neighbor due to the two past overlapped purchases, and a is the only other purchase made by B. Similar results can be obtained with the product graph.

The ranking function defined in Equation (8) plays a critical role in our proposed algorithm, called the *graph partitioning-based recommendation algorithm*. The main computational steps of this algorithm are presented below. The input to this algorithm is the

Figure 6 An Example Illustrating the Graph Partitioning-Based Algorithm



$M \times N$ interaction matrix $A = (a_{ij})$ associated with M consumers $C = \{c_1, c_2, \dots, c_M\}$ and N products $P = \{p_1, p_2, \dots, p_N\}$. We focus on recommendation that is based on transactional data; thus, a_{ij} has two possible values of one and zero, with one representing a past sale of p_j to c_i and zero otherwise. This interaction matrix can be represented as a bipartite graph $G = \langle V, E \rangle$, where

$$V = \{v_1, v_2, \dots, v_{M+N}\},$$

$$v_i = \begin{cases} c_i, & i \leq M, \\ p_{i-M}, & i > M \end{cases}, \quad i = 1, \dots, M+N; \quad E = \{(v_i, v_j)\},$$

where $a_{ij-M} = 1$ or $a_{ji-M} = 1$.

The output is the $M \times N$ matrix $Z = (z_{ij})$ storing potential scores of products for individual consumers.

Graph Partitioning-Based Recommendation Algorithm

Input: Consumer-product graph $G = \langle V, E \rangle$, a heuristically set-partition parameter H , a partition-level distance-connection strength trade-off parameter $u > 1$.

Output: The $M \times N$ potential score matrix Z .

Step 1. Use a graph-partitioning algorithm to obtain a balanced exclusive partition $\pi: G = \bigcup_{l=1}^H G_l$, and a membership function $f: V \rightarrow \{1, \dots, H\}$ gives the partition of each vertex.

Step 2. Repeat, for each product vertex v_j ($j > M$).

2.1. Obtain within-partition degree $d_j = |\{v_i; v_i \in G_{f(v_j)} \text{ and } (v_i, v_j) \in E\}|$.

Step 3. Repeat for each partition G_m .

3.1. Identify the set of partitions that are reachable from G_m , denoted by $R_m = \{G_l; \exists v_i, v_j \text{ such that } v_i \in G_m, v_j \in G_l, \text{ there exists a path } \sigma_{ij} \text{ connecting } v_i \text{ and } v_j\}$.

3.2. Repeat for each G_l in R_m .

3.2.1. Identify all paths connecting G_m and G_l .

3.2.2. The length of the shortest path gives the partition-level distance $D(G_m, G_l)$.

3.2.3. The number of such connecting paths gives the partition-level connection-strength $S(G_m, G_l)$.

Step 4. Repeat for each consumer vertex v_i ($i < M$).

4.1. Repeat for each product vertex v_j ($j > M$).

4.1.1. Compute potential score: $z_{ij-M} = d_j u^{-D(G_{f(v_i)}, G_{f(v_j)})/S(G_{f(v_i)}, G_{f(v_j)})}$.

5.3. Experimental Evaluation

We conducted an experimental study using the two e-commerce transaction data sets analyzed in §4 to evaluate the proposed graph partitioning-based recommendation algorithm. Performances of the user-based neighborhood algorithm and the generative model algorithm were also analyzed for comparison purposes.

For both data sets, we used the consumer-product graphs corresponding to the training time period as

input and treated the purchase transactions in the testing time period (last four months for the book data set and last four weeks for the retail data set; see Table 1) as unknown “future” edges to evaluate the recommendation algorithms. The algorithms were set to generate a ranked list of recommendations of Y products. For each consumer, the recommendation quality was measured based on the number of *hits* (recommendations that match the products actually purchased as recorded in the testing set) and their positions in the ranked list. We adopted the following recommendation-quality metrics regarding the relevance, coverage, relevance and coverage combined, and ranking quality of the ranked list recommendation from the literature (e.g., Breese et al. 1998):

- (a) precision: $P_c = \frac{\text{number of hits}}{Y}$,
- (b) recall: $R_c = \frac{\text{number of hits}}{\text{number of products in the testing set}}$,
- (c) F measure: $F_c = \frac{2 \times P_c \times R_c}{P_c + R_c}$, and
- (d) rank score: $RS_c = \sum_j \frac{q_{cj}}{2^{(j-1)/(h-1)}}$,

where j is the index for the ranked list; h is the viewing *halflife* (the rank of the product on the list such that there is a 50% chance the user will purchase that product); and

$$q_{cj} = \begin{cases} 1 & \text{if } j \text{ is in } c\text{'s testing set,} \\ 0 & \text{otherwise.} \end{cases}$$

For precision, recall, and F measure, an average value over all consumers tested was adopted as the overall metric for the algorithm. For the rank score, an aggregated rank score for all consumers tested was derived:

$$RS = 100 \frac{\sum_c RS_c}{\sum_c RS_c^{\max}},$$

where RS_c^{\max} was the maximum achievable rank score for consumer c if all future purchases had been at the top of a ranked list. The precision, recall, and F measure are standard performance measures to gauge the relevance and coverage of the recommended items relative to the consumers' potential purchases. For instance, 10% precision indicates that one out of 10 recommendations would actually be purchased by the target consumer; 10% recall indicates that if 10 products were to be purchased in the future by the target consumer, one of them would appear in the recommendation list. The F measure is a harmonic mean of the two. Because precision and recall are potentially competing measures, the F measure provides a single-index performance measure balancing these

two. The rank score measure was proposed in Breese et al. (1998) and adopted in many follow-up studies (e.g., Deshpande and Karypis 2004, Huang et al. 2004) to evaluate the ranking quality of the recommendation list. For instance, suppose that the two recommendation lists, one given by Algorithm A and the other by Algorithm B, had exactly the same set of correct recommendations (matched with future purchases) for a target consumer. If the index of each of the correct recommendations in the list by Algorithm A is exactly one less than (closer to the top of the list by one position) that of the matching recommendation by Algorithm B, both algorithms would achieve the same precision, recall, and F measures, but Algorithm A would achieve a rank score doubling that of Algorithm B.

Our implementation of the graph partitioning-based recommendation algorithm uses a software package called ParMETIS, which implements the parallel multilevel k -way graph-partitioning algorithm described in Karypis and Kumar (1998). In our experiments, we set the number of partitions to 500 for both data sets ($H = 500$). The number of recommendations Y was set to 5, 10, and 20, respectively, for different experimental configurations. The experimental results are presented in Table 2. The performance measures in the bold font indicate the best performance across the three algorithms under study for the corresponding measure, data set, and number of recommendations. Differences between these best and second-best performances are statistically significant at the 5% level.

We observe that the newly proposed graph partitioning-based recommendation algorithm out-

performed the standard user-based neighborhood algorithm in almost all four performance measures for both data sets, with one exception for the retail data set when the recommendation lists consist of five recommendations. For the retail data set, this new algorithm also significantly outperformed the generative model. On the other hand, the generative model always delivered the best performance for the book data set. These findings are generally consistent with the patterns observed regarding the topological characteristics of the random, actual, and induced graphs (see Figures 4 and 5). For the book data set, the generative model-induced graph matched closely with the actual graph (see Figure 4). In Figure 5, we observe that the actual consumer-product graph for the retail data set had a substantially larger clustering coefficient than other graphs. This is consistent with the experimental results showing that the new graph partitioning-based algorithm is the best-performing approach for the retail data set. These findings demonstrate that our proposed algorithm is potentially capable of exploiting the principles governing the generation of consumer-product graphs, which to some extent are not fully captured by existing recommendation algorithms. As a side test, we analyzed the graphs induced by the graph partitioning-based algorithm, which were generated in the same manner as those reported in §4.3. It is confirmed that these graphs induced by our algorithm indeed exhibit clustering coefficients larger than the graphs induced by the user-based and generative model algorithms, resulting in a closer match with the actual graphs. We also note that in general the retail data set had much worse recommendation

Table 2 Recommendation Performance Measures

Number of recommendations	Data set	Algorithm	Precision	Recall	F	Rank score
$Y = 5$	Book	User-based	0.0217	0.0463	0.0275	3.1446
		Generative model	0.0493	0.0983	0.0582	18.4739
		Partitioning-based	0.0197	0.0406	0.0234	3.9818
	Retail	User-based	0.0047	0.0111	0.0064	0.6768
		Generative model	0.0058	0.0136	0.0077	0.6768
		Partitioning-based	0.0113	0.0346	0.0162	3.6131
$Y = 10$	Book	User-based	0.0133	0.0556	0.0202	4.5292
		Generative model	0.0336	0.1263	0.0485	9.6162
		Partitioning-based	0.0255	0.1065	0.0377	9.0444
	Retail	User-based	0.0027	0.0128	0.0044	0.9729
		Generative model	0.0036	0.0201	0.0059	0.6486
		Partitioning-based	0.0086	0.0523	0.0142	4.7786
$Y = 20$	Book	User-based	0.0064	0.0729	0.0148	5.5068
		Generative model	0.0193	0.1382	0.0317	10.5059
		Partitioning-based	0.0153	0.1291	0.0259	9.1581
	Retail	User-based	0.0019	0.0204	0.0034	1.1985
		Generative model	0.0022	0.0234	0.0040	1.3536
		Partitioning-based	0.0056	0.0686	0.0101	4.7786

quality than the book data set. This finding is consistent with the observation that the consumer-product graph from the retail data set in general deviated less significantly from a purely random graph than the one from the book data set, as measured by the topological measures (shown in Figures 2 and 3). In other words, there was little room for nonrandom data patterns for recommendation algorithms to exploit in the retail data set.

Note that the proposed graph partitioning-based recommendation algorithm is meant to be an illustration of how the empirical observations of the consumer-product graphs guided by the random graph modeling framework can lead to actionable algorithm design ideas. We acknowledge that the performance of such algorithms may still be dependent on the characteristics of the underlying data set. For example, in our experiments on the book data set, the generative model algorithm outperforms the graph partitioning-based algorithm because the underlying graph does not show much unaccounted-for clustering tendency. It is not our intention in this paper to develop a universally best-performing recommendation algorithm. Rather, we focus on developing a well-grounded general framework to validate and evaluate recommendation algorithms, and generating new ideas for future recommendation algorithm development.

6. Conclusions and Future Directions

In this paper, we apply random graph modeling methodology to study consumer purchase behavior in e-commerce settings. We represent sales transactions as a bipartite consumer-product graph, and then study the topological characteristics of the consumer and product graphs projected from this consumer-product graph. We show that the topological characteristics of several real-world consumer and product graphs deviate significantly from the theoretical predictions based on a random bipartite graph. In particular, these graphs exhibit larger-than-expected average path lengths and a stronger-than-expected tendency to cluster. These findings confirm the fundamental assumption of collaborative filtering-based recommender systems research regarding the nonrandomness of consumers' choices even with no consumer or product information beyond past sales transactions. We also studied consumer-product graphs induced by two popular recommendation algorithms and observed that in general they are capable of explaining some of the nonrandom elements of the consumer-product graph-generation process. However, significant differences between actual and these induced graphs remain. Such differences shed light on the nature of the principles that govern the consumer-product interaction, but are not fully

captured by the existing recommendation methods. These observed differences also motivated the design of a new graph partitioning-based recommendation algorithm whose effectiveness was shown through an experimental study with two well-known collaborative filtering algorithms as the benchmark.

We summarize the contributions and potential implications of our work before discussing future research. Our work represents a novel and practical application of random graph modeling in recommendation systems. We have provided general justification for collaborative filtering-based recommendation in e-commerce applications. We have also demonstrated the potential to develop a practical metalevel recommendation algorithm validation, analysis, and selection framework based on the concepts from random graph theory. From a practical standpoint, if completed, such a framework can guide the design of recommender systems by (potentially automatically) "recommending" appropriate recommendation algorithms based on the topological similarity between the consumer-product graph observed in the application context of interest and various algorithm-induced graphs. Our graph partitioning-based recommendation algorithm not only demonstrates the practical relevance of random graph modeling from a decision-making perspective, but also is a highly effective recommendation approach itself, with potentially wide applications. Furthermore, although our work focuses on consumer purchase behavior, the underlying research framework and the proposed graph partitioning-based algorithm have the potential to be applied in any of the problem domains in which bipartite affiliation networks and link/edge prediction play a central role (e.g., authorship networks and the World Wide Web).

Our future research will be focused on developing a full-fledged random graph-based framework that can be used to evaluate the underlying assumptions of existing recommendation algorithms and to select the most appropriate recommendation algorithm for particular data sets. In particular, we are working toward a theory of bipartite random graphs to eliminate the need to project consumer-product graphs into unipartite graphs, which causes information loss (Robins and Alexander 2004). We are also working on improving our graph partitioning-based collaborative filtering algorithm with more comprehensive algorithm analysis/fine-tuning and large-scale comparative experiments. We plan to explore other computational techniques to fully exploit the information brought out by a random graph perspective for better recommendation performance.

Acknowledgments

This research was partly supported by an NSF Digital Library Initiative-II grant, "High-Performance Digital

Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999–March 2002; and by an NSF Information Technology Research grant, "Developing a Collaborative Information and Knowledge Management Infrastructure," IIS-0114011, September 2001–August 2005. The first author wishes to acknowledge support from a research grant from the Smeal College at Pennsylvania State University. The second author is also affiliated with the Key Lab of Complex Systems and Intelligence Science, the Institute of Automation, Chinese Academy of Sciences, Beijing, and wishes to acknowledge support from an open research grant (ORP-0303), an international collaboration grant (2F05NO1) from the Chinese Academy of Sciences, and a research grant (60573078), and a 973 program grant (2006CB705500) from the National Natural Science Foundation of China.

References

- Adomavicius, G., A. Tuzhilin. 2001. Using data mining methods to build customer profiles. *IEEE Comput.* **34**(2) 74–82.
- Albert, R., A.-L. Barabási. 2002. Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97.
- Albert, R., H. Jeong, A.-L. Barabási. 2000. Error and attack tolerance of complex networks. *Nature* **406** 378–382.
- Allenby, G. M., J. L. Ginter. 1995. Using extremes to design products and segment markets. *J. Marketing Res.* **32**(4) 392–403.
- Amaral, L. A. N., A. Scala, M. Bartheley, H. E. Stanley. 2000. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**(21) 11149–11152.
- Ansari, A., S. Essegai, R. Kohli. 2000. Internet recommendations systems. *J. Marketing Res.* **37**(3) 363–375.
- Azar, Y., A. Fiat, A. R. Karlin, F. McSherry, J. Saia. 2001. Spectral analysis of data. *Proc. 33rd ACM Sympos. Theory Comput.*, ACM Press, New York, 619–626.
- Barabási, A.-L., R. Albert. 1999. Emergence of scaling in random networks. *Science* **286** 509–512.
- Barabási, A.-L., H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, A. Schubert. 2002. Evolution of the social network of scientific collaborations. *Physica A* **311** 590–614.
- Baum, J. A. C., A. V. Shipilov, T. J. Rowley. 2003. Where do small worlds come from? *Indust. Corporate Change* **12** 697–725.
- Bellmann, S., G. L. Lohse, E. J. Johnson. 1999. Predictors of online buying behavior. *Comm. ACM* **42**(12) 32–38.
- Bondy, J. A., U. S. R. Murty. 1976. *Graph Theory with Applications*. American Elsevier Publishing, New York.
- Bradlow, E. T., P. S. Fader. 2001. Bayesian lifetime model for the "Hot 100" billboard songs. *J. Amer. Statist. Assoc.* **96** 368–381.
- Breese, J. S., D. Heckerman, C. Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proc. Fourteenth Conf. Uncertainty Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 43–52.
- Cancho, F. I., R. V. Sole. 2001. The small world of human language. *Proc. Roy. Soc. London Ser. B—Biol. Sci.* **268** 2261–2265.
- Canyon, M. J., M. R. Muldoon. 2004. The small world network structure of boards of directors. Social Science Research Network, <http://ssrn.com/abstract=546963>.
- Cox, D. R. 1972. Regression models and life-tables (with discussion). *J. Roy. Statist. Soc.* **B34** 187–220.
- Davis, G. F., M. Yoo, W. E. Baker. 2003. The small world of the corporate elite, 1982–2001. *Strategic Organ.* **1** 301–326.
- Dempster, A., N. Laird, D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**(1) 1–38.
- Deshpande, M., G. Karypis. 2004. Item-based top-*N* recommendation algorithms. *ACM Trans. Inform. Systems* **22**(1) 143–177.
- Dorogovtsev, S. N., J. F. F. Mendes. 2001. Scaling properties of scale-free evolving networks: Continuous approach. *Phys. Rev. E* **63**(5) 056125-1–056125-19.
- Ehrenberg, A. S. C. 1988. *Repeat-Buying: Facts, Theory and Applications*. Charles Griffin & Company Limited, London, UK.
- Erdős, P., A. Rényi. 1959. On random graphs. *Pub. Math.* **6** 290–297.
- Fader, P. S., B. G. S. Hardie. 2001. Forecasting repeat sales at CDNOW: A case study. *Interfaces* **31**(3) S94–S107.
- Faloutsos, C., P. Faloutsos, M. Faloutsos. 1999. On power law relationships of the Internet topology. *Proc. ACM SIGCOMM.*, ACM Press, New York, 251–262.
- Gershoff, A. D., P. M. West. 1998. Using a community of knowledge to build intelligent agents. *Marketing Lett.* **9**(1) 79–91.
- Gladwell, M. 1999. The science of the sleeper: How the information age could blow away the blockbuster. *The New Yorker* (October 4) 48–55.
- Goldberg, K., T. Roeder, D. Gupta, C. Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Inform. Retrieval* **4**(2) 133–151.
- Guimerà, R., B. Uzzi, J. Spiro, L. A. N. Amaral. 2005. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**(5722) 697–702.
- Heckerman, D., D. M. Chickering, C. Meek, R. Rounthwaite, C. Kadie. 2000. Dependency networks for inference, collaborative filtering, and data visualization. *J. Machine Learn. Res.* **1** 49–75.
- Herlocker, J. L., J. A. Konstan, L. G. Terveen, J. T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inform. Systems* **22**(1) 5–53.
- Hill, W., L. Stead, M. Rosenstein, G. Furnas. 1995. Recommending and evaluating choices in a virtual community of use. *Proc. ACM Conf. Human Factors in Comput. Systems CHI'95*, ACM Press, New York, 194–201.
- Hofmann, T. 2004. Latent semantic models for collaborative filtering. *ACM Trans. Inform. Systems* **22**(1) 89–115.
- Huang, Z., H. Chen, D. Zeng. 2004. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inform. Systems (TOIS)* **22**(1) 116–142.
- Jain, D. C., N. J. Vilcassim. 1991. Investigating household purchase timing decisions: A conditional hazard function approach. *Marketing Sci.* **10** 1–23.
- Jain, D. C., N. J. Vilcassim, P. K. Chintagunta. 1994. A random-coefficients logit brand-choice model applied to panel data. *J. Bus. Econom. Statist.* **12** 317–328.
- Jeong, H., S. P. Mason, A.-L. Barabási, Z. Oltvai. 2001. Lethality and centrality in protein networks. *Nature* **411** 41–42.
- Kamakura, W. A., G. J. Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Res.* **26** 379–390.
- Kamakura, W. A., M. Wedel, J. Agrawal. 1994. Concomitant variable latent class models for conjoint analysis. *Internat. J. Res. Marketing* **11** 451–464.
- Karypis, G., V. Kumar. 1998. Multilevel *k*-way partitioning scheme for irregular graphs. *J. Parallel Distributed Comput.* **48** 96–129.
- Kogut, B., G. Walker. 2001. The small world of Germany and the durability of national ownership networks. *Amer. Sociol. Rev.* **66**(3) 317–335.
- Kumar, S. R., P. Raghavan, S. Rajagopalan, A. Tomkins. 1998. Recommendation systems: A probabilistic analysis. *Proc. 39th Annual Sympos. Foundations Comput. Sci.*, IEEE Computer Society Press, Los Alamitos, CA, 664–673.
- Lin, W., S. A. Alvarez, C. Ruiz. 2002. Efficient adaptive-support association rule mining for recommender systems. *Data Mining Knowledge Discovery* **6** 83–105.

- McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. P. Zarembka, ed. *Frontiers of Econometrics*. Academic Press, New York, 105–142.
- Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Rev.* **45**(2) 167–256.
- Newman, M. E. J., S. H. Strogatz, D. J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**(2) 026118-1–026118-17.
- Newman, M. E. J., D. J. Watts, S. H. Strogatz. 2002. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **99** 2566–2572.
- Resnick, P., H. Varian. 1997. Recommender systems. *Comm. ACM* **40**(3) 56–58.
- Resnick, P., N. Iacovou, M. Suchak, P. Bergstorm, J. Riedl. 1994. GroupLens: An open architecture for collaborative filtering of netnews. *Proc. ACM Conf. Comput.-Supported Cooperative Work*, ACM Press, New York, 175–186.
- Robins, G., M. Alexander. 2004. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Comput. Math. Organ. Theory* **10** 69–94.
- Rossi, P. E., R. E. McCulloch, G. M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Sci.* **15**(4) 321–340.
- Sarwar, B., G. Karypis, J. Konstan, J. Riedl. 2000. Application of dimensionality reduction in recommender systems: A case study. *Proc. WebKDD Workshop ACM SIGKDD*, ACM Press, New York.
- Schafer, J., J. Konstan, J. Riedl. 2001. E-commerce recommendation applications. *Data Mining Knowledge Discovery* **5**(1–2) 115–153.
- Shardanand, U., P. Maes. 1995. Social information filtering: Algorithms for automating word of mouth. *Proc. ACM Conf. Human Factors Comput. Systems*, ACM Press, New York, 210–217.
- Ungar, L. H., D. P. Foster. 1998. A formal statistical approach to collaborative filtering. *Proc. Conf. Automated Learn. Discovery (CONALD'98)*, Pittsburgh, PA, <http://citeseer.ist.psu.edu/ungar98formal.html>.
- Uzzi, B., J. Spiro. 2005. Collaboration and creativity: The small world problem. *Amer. J. Sociol.* **111**(2) 447–504.
- Vrooman, E., J. Riedl, J. Konstan. 2002. *Word of Mouse: The Marketing Power of Collaborative Filtering*. Warner Business Books, New York.
- Watts, D. J. 1999. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, NJ.
- Watts, D. J., S. H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature* **393** 440–442.
- Yang, I., H. Jeong, B. Kahng, A.-L. Barabási. 2003. Emerging behavior in electronic bidding. *Phys. Rev. Lett. E* **68**(1) 016102-1–016102-5.