Why Does Collaborative Filtering Work?
— *Recommendation Model Validation and Selection by Analyzing Bipartite Random Graphs*

Zan Huang
Supply Chain and Information Systems
Smeal College of Business
Pennsylvania State University
Email: zanhuang@psu.edu

Daniel D. Zeng
Management Information Systems
Eller College of Management
University of Arizona
Email: zeng@eller.arizona.edu

Abstract
A large number of collaborative filtering (CF) algorithms have been proposed in the literature as the core of automated recommender systems. However, the underlying justification for these algorithms is lacking and their relative performances are typically domain- and data-dependent. In this paper, we aim to develop initial understanding of the validation and model/algorithm selection issues based on the graph topological modeling methodology. By representing the input data in the form of consumer-product interactions such as purchases and ratings as a bipartite graph, we develop bipartite graph topological measures to capture patterns that exist in the input data relevant to recommendation. Using a simulation approach, we observe the deviations of these topological measures for given recommendation datasets from the expected values for simulated random datasets. These deviations help explain why certain CF algorithms work for the given datasets. They can also serve as the basis for a comprehensive model selection framework that chooses appropriate CF algorithms given the characteristics of the dataset under study. We validate our approach using two real-world e-commerce datasets.

1. Introduction
Recommender systems such as those employed by *Amazon* and *Netflix* automate the process of recommending products and services to consumers based on various types of data concerning consumers, products, and previous interactions between consumers and products. Consumer-product interactions can take different forms, such as product/service purchases, ratings, paper citations, and catalog and website browsing activities. Recommender systems are being increasingly adopted in a wide range of applications, especially in e-commerce applications. They have become a standard e-commerce technology that helps increase online and catalog sales and improve customer loyalty.

The *collaborative filtering* (CF) approach [3] has been acknowledged to be the most commonly-used recommendation approach. This approach utilizes only consumer-product interaction data and largely ignores consumer and product attributes. The majority of recommendation algorithms proposed in the literature falls into this category [1]. Example CF approaches include the standard user-based and item-based neighborhood algorithm, dimensionality reduction algorithms [10], generative models [5], clustering approaches [3], and graph-based algorithms [6]. In addition, a large number of CF algorithm evaluation studies have been reported, revealing that relative performances of different algorithms are largely domain- and data-dependent. These previous evaluation studies are mostly based on computational testing of the performance of various algorithms using a limited set of recommendation datasets. Such studies, albeit informative and important, do not directly answer the fundamental question behind automated recommendation: Are the future interactions between consumers and products *inherently predictable* given the past consumer-product interactions?

We see a critical need to investigate this fundamental "predictability" question and in turn develop a model selection and meta-level analysis framework that is able to extract relevant features from recommendation datasets and "recommend" appropriate recommendation algorithms in a principled manner. This paper reports our initial attempts towards this direction. Our study is based on a bipartite graph-based consumer-product interaction representation, in which the two types of nodes represent consumers and products, respectively, and links between consumer and product nodes represent the interactions between them. We focus on binary interactions (e.g., the presence of a link represents an observed sales transaction) in this paper and leave the weighted graph representation for the multi-scaled

data such as rating data for future research. Under this graphical representation, the CF problem can be viewed as a problem of predicting future links in a growing graph/network based on previously observed links. We adopt the graph topological modeling methodology to analyze the global topological properties of the consumer-product graph and link these properties to the predictability of future transactions and the relative performance of different collaborative filtering algorithms.

The remainder of the paper is organized as follows. Section 2 presents in detail the bipartite graph representation of consumer-product interactions and the underlying graph topological modeling methodology. Section 3 summarizes several representative collaborative filtering algorithms and proposes several bipartite topological measures relevant to link prediction and recommendation. We present an empirical study in Section 4 using two real-world e-commerce datasets to demonstrate the usefulness of the proposed topological measures and summarize findings connecting these measures to the effectiveness of various CF approaches. We conclude the paper in Section 5 by summarizing contributions and pointing out the future directions.

2. Consumer-product Graphs and the Graph Topological Modeling Methodology

2.1 Consumer-product Graphs

Consumer-product interactions captured in a sales transaction dataset can be naturally represented as a graph by treating consumers and products as nodes, and transactions involving consumer-product pairs as links between these nodes. We refer to this graph as the *consumer-product graph* in this paper. This type of graph is a *bipartite graph* whose nodes can be divided into two distinct sets and whose links only make connections between the two sets. The input data for collaborative filtering has been traditionally represented by an interaction matrix. The example shown in Figure 1, which involves 3 consumers, 4 products, and 7 past transactions, shows a direct mapping between the two representations.

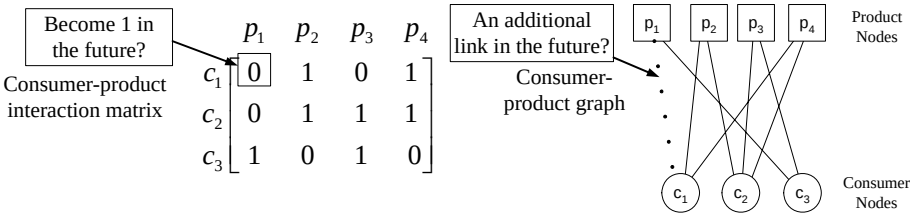

Figure 1. Example consumer-product interaction matrix and consumer-product graph

Under this graphical representation, the CF problem can be reformulated as a link prediction problem: How to predict the future links between the existing consumer and product nodes based on the observed links in the consumer-product graph? A natural question that is fundamental to CF arises: Is the growth of the graph *inherently predictable* based on the observed existing links? One approach that may shed light on this question is to investigate whether a given consumer-product graph is random [2, 7]. Only when such a graph is not random in certain aspect there is room for *any* CF algorithm to make successful and meaningful recommendations.

2.2 Random Graph and Topological Modeling Methodology

The main methodology for investigating the randomness of graphs is random graph theory and related graph topological modeling methods [2]. As common in this literature, we use graph and network interchangeably in this paper. The key assumption of random graph analysis 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.

To determine whether a graph generation model of certain degree of randomness can reproduce the topological features of the real network, we analytically or numerically derive the distributions of the topological measures of the ensemble of graphs generated by this model conditional on certain constraints derived from the real network (e.g., the number of nodes, number of links, or the observed degree distribution). The topological measures of the real network are then compared with these expected distributions to determine the fit between the model and the real network, in a similar manner as statistical

hypothesis testing of single variables. In recent years, many studies have analyzed topological characteristics of large-scale real-world networks in various application domains, including World-Wide Web, Internet, movie actor collaboration network, science collaboration network, and cellular network, among others [2]. These real-world networks have demonstrated surprisingly consistent non-random topological characteristics across different domains. Three major concepts related to such topological features are: "small world," "clustering," and "scale-free" phenomena [2, 8].

The existing literature on graph topological modeling has largely focused on analyzing the unipartite graphs where links are allowed to form between any pair of nodes. In our context, consumer-product graphs are inherently bipartite. The topological measures developed for the unipartite graphs are often not meaningful, or not directly relevant to, bipartite graphs. The dominant approach in the current literature to analysis of bipartite graphs is to project it into two unipartite graphs. The consumer-product graph has been previously analyzed in this manner being projected into a consumer co-purchase graph and a product co-purchased-by graph [7]. However, valuable structural information critical to the analysis of the recommendation problem is lost during the projection process. In this paper, we develop new bipartite graph topological measures particularly motivated for the purpose of explaining CF algorithm performance and selecting appropriate algorithms for individual datasets.

3. Collaborative Filtering Algorithms and Bipartite Graph Topological Measures

3.1 Collaborative Filtering Algorithms

A naïve recommendation algorithm makes recommendation simply based on popularity of the products, i.e., recommending to each consumer the most popular products that are not purchased previously by this consumer. We refer to this algorithm as the *top-N most popular* algorithm.

One basic CF algorithm is a well-tested *user-based* neighborhood algorithm using statistical correlation [3]. To predict the potential interests of a given consumer, this algorithm first identifies a set of similar consumers based on correlation coefficients or similarity measures using the past transactions, and then makes a prediction based on the behavior of these similar consumers. The fundamental assumption is that consumers who have previously bought a large set of the same products will continue to buy the same set of new products in the future. The *item-based* algorithm [4] is different from the user-based algorithm only in that product similarities are computed instead of consumer similarities. The assumption here is that products that have been bought by the same set of consumers will continue to be co-purchased by other consumers. The user-based and item-based algorithms are the mostly commonly used CF algorithms. Under the graphical representation, both algorithms rely on the paths of length 3 (involving 4 nodes, which we refer to as 4-node paths) to make recommendations: "target consumer – purchased product – similar consumer – unpurchased product" or "target consumer – purchased product – other consumer – similar product as the purchased ones."

Many recent CF algorithms explore data patterns beyond 4-node paths. The graph-based algorithms, such as the *spreading activation* algorithm [6], explicitly explore longer paths to exploit the transitive consumer-product associations. The fundamental assumption is that the behavior of the transitive neighbors (neighbors of the neighbors) is also informative in predicting the behavior of the consumer. Other algorithms, such as the dimensionality reduction algorithms [10], generative models [5], and clustering approaches [3], essentially analyze the global properties of the interaction matrices and implicitly explore the transitive associations.

The relative performances of different CF algorithms are not consistent across the studies reported in the literature. Furthermore, no simple descriptive measures (e.g., relative size of the consumer and product sets and the density level of the interactions) seem to be able to serve as the indicator for choosing the best-performing algorithms. In this paper we develop bipartite graph topological measures motivated by the fundamental assumptions of various CF algorithms discussed above to serve as the basis for model/algorithm selection.

3.2 Bipartite Graph Topological Measures

We are solely interested in topological measures, in particular, clustering measures that are relevant to recommendation algorithms. From the topological perspective, various CF algorithms involve paths and cycles. Figures 2a-2d show four recommendation-relevant configurations of a consumer-product graph.

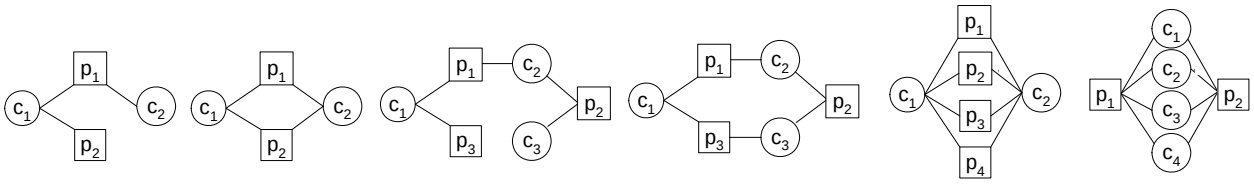

Figure 2. Local graph configurations: circles represent the consumers and squares represent the products

The 4-node path and 4-node cycle are the basic recommendation-relevant configurations. The user/item-based algorithms can be viewed as making recommendations such that many 4-node paths will be completed to form 4-node cycles. Thus, a measure that describes the tendency for the paths to become cycles in the observed data would be a good indicator regarding whether this assumption holds for the particular dataset and in turn whether the user/item-based algorithm would perform well. A natural extension of the unipartite graph clustering coefficient measure to the bipartite graphs (as adopted in previous studies in other domains [9]) may serve as such a measure: the *4-node bipartite clustering coefficient* defined as

$$C_4 = \frac{4 \times (\text{number of 4-node cycles in the bipartite graph})}{\text{number of 4-node paths}} \quad (1)$$

Following the same idea we define the *6-node bipartite clustering coefficient*:

$$C_6 = \frac{6 \times (\text{number of 6-node cycles in the bipartite graph})}{\text{number of 6-node paths}} \quad (2)$$

This measure is directly motivated by the fundamental assumption made by the algorithms that explore transitive consumer-product associations, such as the spreading activation algorithm. If, for instance, the number of 6-node cycles is much smaller than that of 6-node paths, it would be very difficult for algorithms exploring transitive associations to perform well since such approaches would be misguided by the target patterns that do not occur often enough. Similarly, if the C_4 is lower than the expectation, meaning consumers seldom co-purchase more than one product and products are seldom co-purchased by more than one consumer, one would not expect the CF algorithm to outperform the naïve top-N most popular recommendation. Clustering coefficients involving more nodes (8, 10, etc.) can be also defined for bipartite graphs. But the computational complexity grows exponentially with more nodes involved. In addition, the longer cycle patterns might be rare or decomposable to shorter cycle patterns thus do not provide much more information regarding the graph structure.

In addition to the two clustering coefficients, we also define the *consumer/product pair clustering distributions* to describe the prevalence of the two graph configurations shown in Figures 2e and 2f. Figure 2e shows a configuration with heavy consumer pair clustering participation, in which the two consumers participate in many 4-node cycles while each product pair only participates in one cycle. Figure 2f, on the other hand, shows a configuration with heavy product pair clustering participation. The two configurations are expected to correspond to the cases with which the user-based and item-based algorithms would outperform the other, respectively. Formally we define $P_c(k)$ ($P_p(k)$) as the probability for a pair of consumers (products) to participate in k 4-node cycles. The distribution of $P_c(k)$ ($P_p(k)$) is defined as the consumer (product) pair clustering distribution.

Since no analytical results concerning the expected values of the above bipartite graph topological measures are readily available, we rely on simulation in this paper. We simulate a set of graphs that maintain the identical consumer and product node degree distributions to the observed graph. We then obtain the simulated distribution of the proposed clustering coefficients and node pair clustering distributions. The observed topological measures are then compared with these simulated distributions to determine the deviations. Note that we implicitly assume that the entire graph generation process is stationary in the sense that the underlying principle governing this process does not change over time.

4. An Empirical Study

4.1 Data and Recommendation Algorithm Performances

We used two real-world e-commerce recommendation datasets: a **retail** dataset from a major US online

apparel merchant and a **book** sales dataset from a major Chinese online bookstore. Table 1 presents the basic data statistics and the performance of the four recommendation algorithms measured by commonly used metrics of precision, recall, F measure, and rank score [3] that measure the accuracy, coverage, and ranking quality by matching the recommendation lists (10 products for each consumer) with withheld 20% later actual purchase records. Consistent with our earlier discussion, we observe that no one algorithm outperformed others for both datasets. The basic data metrics, such as the overall density level, average number of purchases per consumer, and average sales per product, are not quite informative and cannot explain the relative performances of different algorithms across the three datasets. These observations again exemplify the need for new topological measures to explain the algorithm performances.

Dataset	Retail	Book
# of Consumers	1,000	851
# of Products	7,328	8,566
# of Transactions	9,332	13,902
Density Level*	0.13%	0.19%
Avg. # of purchases per consumer	9.33	16.34
Avg. sales per product	1.27	1.62

Dataset	Retail				Book			
Algorithm	User-based	Item-based	Spreading Activation	Top-N Most Popular	User-based	Item-based	Spreading Activation	Top-N Most Popular
Precision	0.0042	0.0106	**0.0130**	0.0062	0.0122	0.0093	0.0231	**0.0258**
Recall	0.0305	0.0731	**0.0863**	0.0326	0.0753	0.0443	0.1155	**0.1316**
F	0.0073	0.0182	**0.0219**	0.0100	0.0202	0.0144	0.0362	**0.0405**
Rank Score	2.5770	4.9866	**5.2209**	1.3889	4.9332	3.2146	9.4955	**10.7814**
Algorithm Rank	3	2	1	4	3	4	2	1

Table 1. Left panel: basic data statistics; right panel: recommendation algorithm performance

4.2 Results and Discussions

The observed values and the simulated mean and standard deviation of C_4 and C_6 are reported in Table 2. We also report the z-score to indicate how much the observed values deviate from the simulated distribution. Our current simulation generates 20 graphs from the uniform graph distribution conditional on the actual degree distributions. The observed values of C_4 of the retail and book datasets both deviate significantly from the simulated distributions. The z-scores indicate that the observed C_4 of the retail dataset deviates much more significantly than that of the book dataset does (141.38 as compared to 23.35), which is consistent with the general superior performance of user/item-based algorithms compared to the top-N most popular recommendation for the retail dataset.

Dataset	Topological Measure	Observed	Simulated Mean	Simulated Stdev	Z-score
Retail	C_4	0.03787	0.00079	0.00026	141.38
	C_6	0.03649	0.00076	0.00015	231.71
Book	C_4	0.02519	0.01065	0.00062	23.35
	C_6	0.01589	0.00934	0.00047	14.04

Table 2. Observed and simulated values of the bipartite clustering coefficients

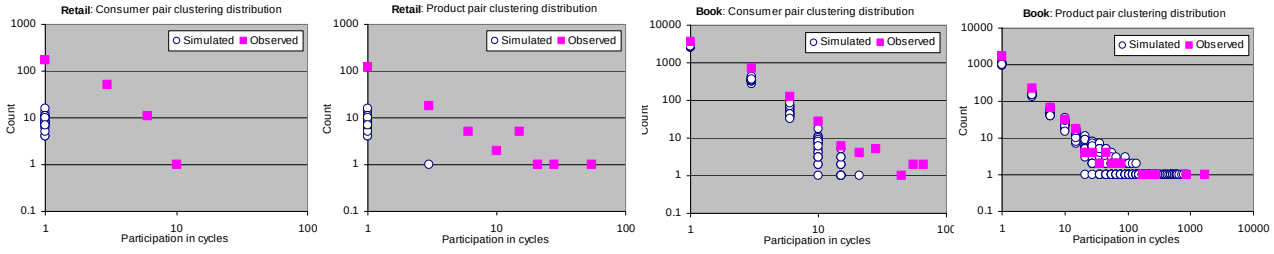

Figure 3. Observed and simulated consumer and product pair clustering distributions

As for the C_6 results, both the retail and book datasets show significant deviation from the simulated distribution, which is consistent with the better performance of the spreading activation algorithm than the user/item-based algorithms. The retail dataset exhibits much more significant deviation of the observed C_6 from the simulated distribution (z-score of 231.71), while the book dataset shows relatively minor deviation (z-score of 14.04). These observations are consistent with the superior performance of the spreading activation algorithm over the user/item-based algorithms for the retail dataset and the superior performance of the top-N recommendation over the spreading activation algorithm.

Figure 3 presents the observed and simulated consumer/product pair clustering distributions. These results are based on the same real and simulated graphs as used for Table 2. For example, the leftmost figure shows that for consumer pairs of the simulated retail datasets, about 10 pairs participate in one 4-

node cycles and no pairs participate in more than one cycle. For the observed data, a lot more pairs (172) participate in one cycle and 51, 11, and 1 consumer pairs participate in 3, 6, and 10 cycles. In general, the observed consumer/product pair clustering distributions for the book dataset deviate less from the simulated distributions than those of the retail dataset, which explains the overall better performance of the user/item-based algorithms for the retail dataset than for the book dataset. For the retail dataset, both consumer and product pair clustering distributions deviate significantly from the simulated distribution, with qualitatively a more significant deviation for the product pair clustering distribution. For the book dataset, the product pair clustering distribution is almost consistent with the simulated distribution while the consumer pair clustering distribution exhibits slight deviation, especially in the tail. These observations are consistent with the fact that item-based algorithm outperformed the user-based algorithm for the retail dataset and the reverse was observed for the book dataset.

In summary, a large deviation of C_4 from expected values under random assumption explains the superior performance of CF algorithms in general compared to the naïve recommendation, while whether or not C_6 deviates from expectation explains whether algorithms incorporating transitive associations would outperform user/item-based algorithms that only consider direct neighborhood relationships. The comparison between the consumer and product pair clustering distribution explains either user- or item-based algorithm outperforms the other.

5. Conclusions and Future Directions

This paper advocates the use of graph topological modeling methodology in developing a meta-level model validation and selection framework for CF algorithms. New bipartite topological measures are proposed to capture the recommendation-relevant data patterns. We demonstrated empirically the potential usefulness of our proposed measures in explaining the relative performances of several representative CF algorithms. Our ongoing and future work is focused on (a) extending the generalized random graph theory to analytically derive expected behavior (and its approximation if appropriate) of the measures used in the current study (as opposed to a simulation-based approach), (b) developing a more comprehensive set of topological measures including various distribution properties of cycles, (c) creating a taxonomy of both theory-driven and data-driven graph generation mechanisms relevant to recommender systems, and (d) conducting a large scale computational study using additional datasets and appropriate statistical tests to validate our approach and derive specific, verified lessons of practical relevance.

References

1. Adomavicius, G. and Tuzhilin, A. Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749, 2005.
2. Albert, R. and Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*. 47-97, 2002.
3. Breese, J.S., Heckerman, D. and Kadie, C., Empirical analysis of predictive algorithms for collaborative filtering. in *Fourteenth Conference on Uncertainty in Artificial Intelligence*, (Madison, WI, 1998), Morgan Kaufmann, 43-52, 1998.
4. Deshpande, M. and Karypis, G. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, *22*(1). 143-177, 2004.
5. Hofmann, T. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, *22*(1). 89-115, 2004.
6. Huang, Z., Chen, H. and Zeng, D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems*, *22*(1). 116-142, 2004.
7. Huang, Z., Zeng, D.D. and Chen, H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management Science, under review*, 2005.
8. Newman, M.E.J., Strogatz, S.H. and Watts, D.J. Random graphs with arbitrary degree distributions and their applications. *Physics Review, E 64*. 026118, 2001.
9. Robins, G. and Alexander, M. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, *10*. 69-94, 2004.
10. Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., Application of dimensionality reduction in recommender systems: a case study. in *WebKDD Workshop at the ACM SIGKKD*, 2000.