

# A Customer Purchase Incidence Model Applied to Recommender Services

Andreas Geyer-Schulz, Michael Hahsler, and Maximillian Jahn

*Abstract*—In this contribution we transfer a customer purchase incidence model for consumer products which is based on Ehrenberg’s repeat-buying theory to Web-based information products. Ehrenberg’s repeat-buying theory successfully describes regularities on a large number of consumer product markets. We show that these regularities exist in electronic markets for information goods too, and that purchase incidence models provide a well founded theoretical base for recommender and alert services.

The article consists of two parts. In the first part Ehrenberg’s repeat-buying theory and its assumptions are reviewed and adapted for Web-based information markets. Second, we present the empirical validation of the model based on data collected from the information market of the Virtual University of the Vienna University of Economics and Business Administration at <http://vu.wu-wien.ac.at> from September 1999 to May 2001.

## I. INTRODUCTION

In this article we concentrate on an anonymous recommender service of the correlation-type made famous by Amazon.com applied to an information broker. It is based on consumption patterns for information goods (web sites) from market baskets (web browser sessions) which we treat as consumer purchase histories with unobserved consumer identity. In Resnick and Varian’s design space [13] this recommender service is characterized as:

1. The contents of a recommendation consists of links to web sites.
2. It is an implicit service based on user behavior.
3. The service is anonymous.
4. The aggregation of recommendations is based on identifying outliers with the help of a stochastic purchase incidence model.
5. A sorted list of recommended related web sites is offered to the user of a web site (see figure 1).

This recommender service is part of the first educational and scientific recommender system integrated into the Virtual University of the Vienna University of Economics and Business Administration (<http://vu.wu-wien.ac.at>) since September 1999. A full description of all recommender services of this educational and scientific recommender system can be found in [8].

For example, figure 1 shows the recommended list of web-sites for users interested in web-sites re-

lated to the Collaborative Filtering Workshop 1996 in Berkeley. The web-site (in figure 1 the Collaborative Filtering Workshop 1996 in Berkeley) in the yellow box (gray in print) is the site for which related web-sites have been requested.

We have presented the architecture of an information market and its instrumentation for collecting data on consumer behavior in [7]. We consider an information broker with a clearly defined system boundary. Clicking on an external link which leaves the system is equated as “purchasing an information product”. In marketing, we assume that a consumer will only repeatedly purchase a product or a product combination, if he is sufficiently content with it. The rationale that this analogy holds even for *free* information products stems from an analysis of the transaction costs of a user of an information broker. Even *free* information products burden the consumer with search, selection and evaluation costs. Therefore, in this article we derive recommendations from products which have been repeatedly used (= purchased) together in the same sessions (= buying occasions) [4]. Such recommendations are attractive for information brokers for the following reasons:

- Observed consumer purchase behavior is the most important information for predicting consumer behavior online [2] and offline [6].
- In traditional retail chains, basket analysis shows up to 70 percent cross-selling potential [3]. Such recommendations facilitate “repeat-buying”, which should be one of the main goals of e-commerce sites [2].
- Most important in a university environment is that such recommendations are not subject to several incentive problems found in systems based on explicit recommendations (as e.g. free-riding, bias, ...) which are analyzed in [1]. The transaction cost of faking such recommendations is high, because only one co-occurrence of products is counted per user-session as usual in consumer panel analysis [6]. Free-riding is impossible, because by using the information broker each user contributes usage data for the recommendations. The user’s privacy is preserved.
- And, last but not least, the transaction costs for the broker are low, since high-quality recommendations can be generated without human effort. No editor, no author, no web-scout is needed.

Andreas Geyer-Schulz is with Informationsdienste und Elektronische Märkte, Geb. 20.20 Rechenzentrum, Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany. Michael Hahsler and Maximillian Jahn are with Informationswirtschaft, WU-Wien, A-1090 Wien, Austria. E-mails: [andreas.geyer-schulz@em.uni-karlsruhe.de](mailto:andreas.geyer-schulz@em.uni-karlsruhe.de), [Michael.Hahsler@wu-wien.ac.at](mailto:Michael.Hahsler@wu-wien.ac.at), [Maximillian.Jahn@wu-wien.ac.at](mailto:Maximillian.Jahn@wu-wien.ac.at)

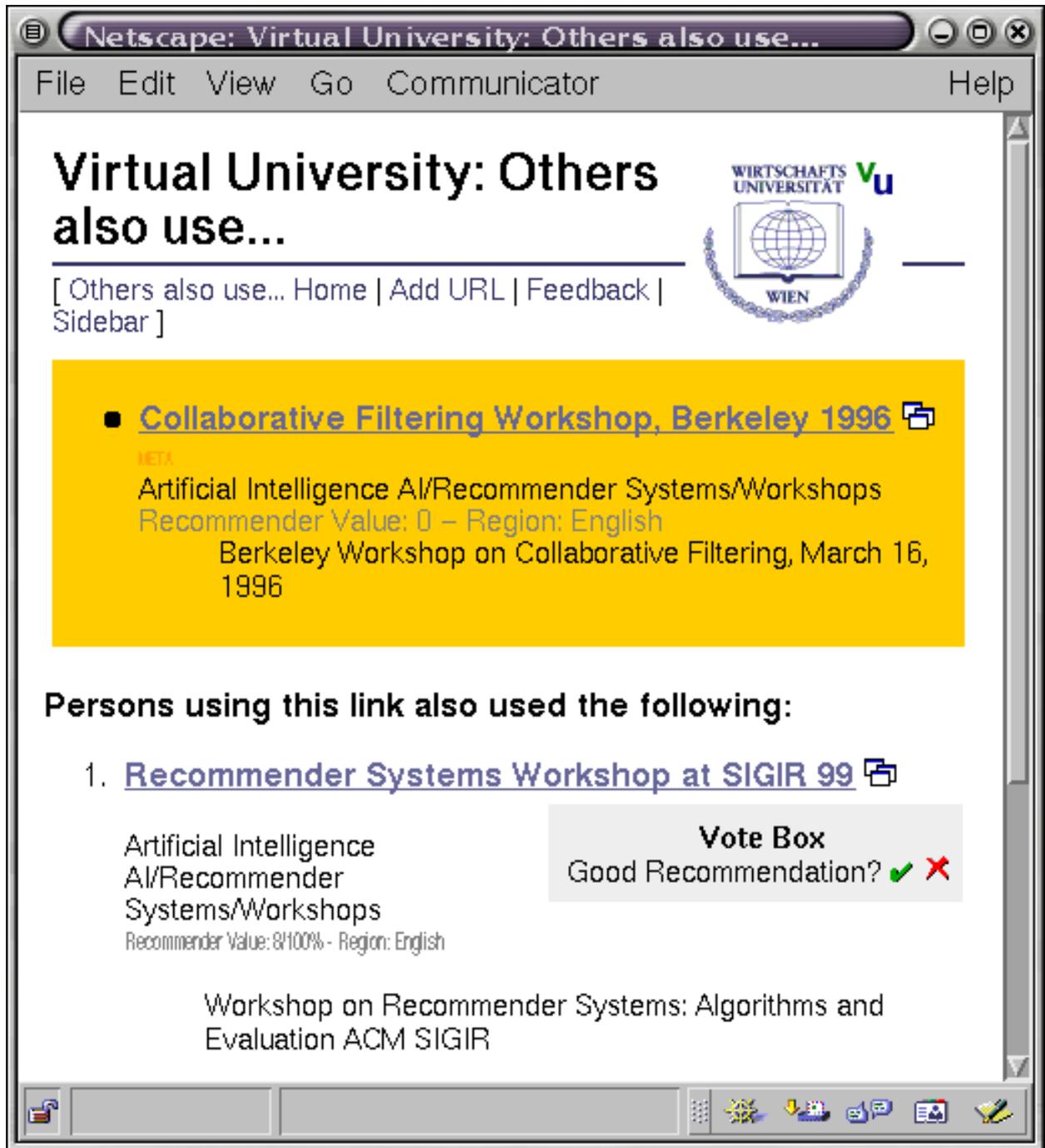


Fig. 1. Example: An Anonymous Recommender Based on "Observed Purchase Behavior"

However, anonymous recommendations based on consumption patterns nevertheless have the following two problems which we address in this article with the help of Ehrenberg's repeat-buying theory (see [6]):

- Which co-occurrences of products qualify as non-random?
- And how many products should be recommended?

Ehrenberg's repeat-buying theory provides us with a reference model for testing for non-random outliers, because of the strong stationarity and independence assumptions in the theory discussed in section II. What makes this theory a good candidate for describing the consumption behavior for information products is that it has been supported by strong empirical evidence in several hundred consumer product markets since the late 1950's. Ehrenberg's repeat-buying theory is a descriptive theory based on consumer panel data. It captures how consumers behave, but not why. Several very sophisticated and general models of the theory (e.g. the Dirichlet model ([10]) exist and have a long tradition in marketing research. However, for our purposes, namely identifying non-random purchases of two information products, the simplest model – the logarithmic series distribution (LSD) model – will prove quite adequate. For a survey on stochastic consumer behavior models, see e.g. [14].

One of the main (conceptual) innovations of this paper is that we explain, how we can apply a theory for analyzing purchase histories from consumer panels to mere market baskets.

## II. EHRENBURG'S REPEAT-BUYING THEORY AND BUNDLES OF INFORMATION PRODUCTS

*Of the thousand and one variables which might affect buyer behavior, it is found that nine hundred and ninety-nine usually do not matter. Many aspects of buyer behavior can be predicted simply from the penetration and the average purchase frequency of an item, and even these two variables are interrelated.* A.S.C. Ehrenberg (1988).

In purchasing a product a consumer basically makes two decisions: when does he buy a product of a certain product class (purchase incidence) and which brand does he buy (brand choice). Ehrenberg claims that almost all aspects of repeat-buying behavior can be adequately described by formalizing the purchase incidence process for a single brand and to integrate these results later (see figure 2).

In a classical marketing context Ehrenberg's repeat-buying theory is based on purchase histories from consumer panels. The purchase history of a consumer is the sequence of the purchases in all his market baskets over an extensive periods of time (a year or more). For information products, the purchase history of a consumer corresponds to the sequence of sessions of a user in a per-

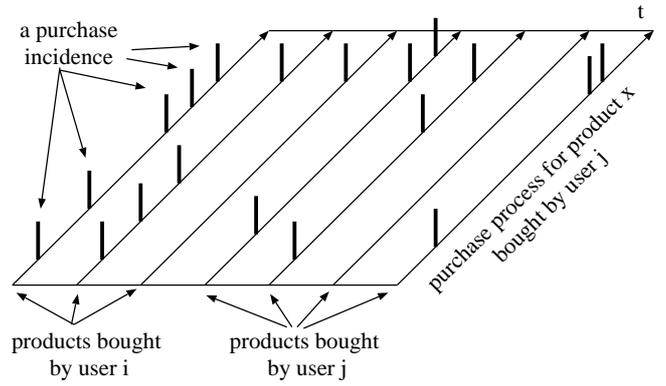


Fig. 2. Purchase Incidences as Independent Stochastic Processes

sonalized environment.

A market basket is simply the list of items (quantity and price) bought in a specific trip to the store. In a consumer panel the identity of each user is known.

For information products the corresponding concept is a session which contains records of all information products visited (used) by a user. In anonymous systems (e.g. most public web-sites) the identity of the user is not known.

Very early in the work with consumer panel data it turned out that the most useful unit of analysis is in terms of *purchase occasions*, not in terms of quantity or money paid. A purchase occasion is coded as yes, if a consumer has purchased one or more items of a product in a specific trip to a store. We **ignore** the number of items bought or package sizes and concentrate our attention on the frequency of purchase.

For information products we define a purchase occasion as follows: a purchase occasion occurs if a consumer visits a specific information product at least once in a specific session. We **ignore** the number of pages browsed, repeat visits in a session, amount of time spent at a specific information product, ... Note, that this definition of counting purchases or information product usage is basic for this article and crucial for repeat-buying theory to hold.

Analysis is carried out in distinct time-periods (such as 1-week, 4-week, quarterly periods) which ties in nicely with other standard marketing reporting practices. A particular simplification from this time-period orientation is that most repeat-buying results for any given item can be expressed in terms of penetration and purchase frequency.

The *penetration*  $b$  is the proportion of people who buy an item at all in a given period. For this article, penetration is of no concern to us, because in anonymous public Internet systems we simply cannot determine the proportion of users who use a specific web-site at all. (That, of course, changes in personalized recommender systems.)

The *purchase frequency*  $w$  is the average number of

times these buyers buy at least one item in the period. The mean purchase frequency  $w$  is itself the most basic measure of repeat-buying in the theory [6] and in this article.

In the following we consider anonymous market baskets as consumer panels with **unobserved consumer identity** – and as long as we work only at the aggregate level, everything works out fine, as long as Ehrenberg’s assumptions on consumer purchase behavior hold.

Figure 2 shows the main idea of purchase incidence models: a consumer buys a product according to a stationary Poisson process which is independent of the other buying processes. Aggregation of these buying processes over the population under the (quite general) assumption that the parameters  $\mu$  of the Poisson distributions (the long-run average purchase rates) follow a truncated  $\Gamma$ -distribution results in a logarithmic series distribution as Chatfield et al. have shown [5]. We present Chatfield’s proof in detail, because the original proof is marred by a typesetting error:

1. The probability  $p_r$  that a buyer makes  $r$  purchases is Poisson distributed:

$$\frac{e^{-\mu} \mu^r}{r!}$$

2. We integrate over all buyers in the truncated  $\Gamma$ -distribution:

$$\begin{aligned} p_r &= c \int_{\delta}^{\infty} \left( \frac{e^{-\mu} \mu^r}{r!} \right) \left( \frac{e^{-\mu/a}}{\mu} \right) d\mu \\ &= \frac{c}{r!} \int_{\delta}^{\infty} e^{-(\mu+\mu/a)} \mu^{r-1} d\mu \\ &= \frac{c}{r!} \int_{\delta}^{\infty} e^{-(1+1/a)\mu} \frac{(1+1/a)^{r-1}}{(1+1/a)^{r-1}} \mu^{r-1} d\mu \\ &= \frac{c}{r!(1+1/a)^{r-1}} \int_{\delta}^{\infty} e^{-(1+1/a)\mu} ((1+1/a)\mu)^{r-1} d\mu \\ &= \frac{c}{r!(1+1/a)^{r-1}} \int_{\delta}^{\infty} e^{-(1+1/a)\mu} ((1+1/a)\mu)^{r-1} \frac{1}{(1+1/a)} d(1+1/a)\mu \\ &= \frac{c}{r!(1+1/a)^r} \int_{\delta}^{\infty} e^{-(1+1/a)\mu} ((1+1/a)\mu)^{r-1} d(1+1/a)\mu \end{aligned}$$

Since  $\delta$  is very small, for  $r \geq 1$  and setting  $t = (1+1/a)\mu$  this is approximately

$$p_r \approx \left( \frac{c}{r!(1+\frac{1}{a})^r} \right) \Gamma(r)$$

$$\begin{aligned} &= \frac{c}{(1+\frac{1}{a})^r r} \\ &= c \frac{q^r}{r} \\ &= qp_{r-1}(r-1)/r \end{aligned}$$

with  $q = \frac{a}{1+a}$ .

3. If  $\sum p_r = 1$  for  $r \geq 1$ , we get  $p_1 = \frac{-q}{\ln(1-q)}$  and  $p_r = \frac{-q^r}{r \ln(1-q)}$ . (However, this is the LSD. q.e.d.)

The logarithmic series distribution (LSD) describes the following frequency distribution of purchases ([6]), namely how many buyers buy a specific product 1, 2, 3, ... times (without taking into account the number of non-buyers)?

$$P(r \text{ purchases}) = \frac{-q^r}{r \ln(1-q)}, \quad r \geq 1 \quad (1)$$

$$\text{Mean purchase frequency } w = \frac{-q}{(1-q) \ln(1-q)} \quad (2)$$

The variance is:

$$\sigma^2 = \frac{-q \frac{1+q}{\ln(1-q)}}{(1-q)^2 \ln(1-q)} \quad (3)$$

For more details on the logarithmic series distribution, we refer the reader to [11]. The logarithmic series distribution results from the following assumptions about the consumers’ purchase incidence distributions:

1. The share of never-buyers in the population is not specified. In our setting of an Internet information broker with anonymous users this definitely holds.

2. The purchases of a consumer in successive periods follow a Poisson distribution with a certain long-run average  $\mu$ . The purchases of a consumer follow a Poisson distribution in subsequent periods, if a purchase tends to be independent of previous purchases (as is often observed) and a purchase occurs in such an irregular manner that it can be regarded as if random (see [14]).

3. The distribution of  $\mu$  in the population follows a truncated  $\Gamma$ -distribution, so that the frequency of any particular value of  $\mu$  is given by  $(ce^{-\mu/a}/\mu)d\mu$ , for  $\delta \leq \mu \leq \infty$ , where  $\delta$  is a very small number,  $a$  a parameter of the distribution, and  $c$  a constant, so that  $\int_{\delta}^{\infty} (ce^{-\mu/a}/\mu)d\mu = 1$ . A  $\Gamma$ -distribution of the  $\mu$  in the population may have the following reason (see [6]): If for different products  $P, Q, R, S, \dots$  the average purchase rate of  $P$  is independent of the purchase rates of the other products, and  $\frac{P}{(P+Q+R+S+\dots)}$  is independent of a consumer’s total purchase rate of buying all the products. These independence conditions are likely to hold approximately in practice.

4. The market is in equilibrium (stationary).

Next, consider for some fixed information product  $x$  in the set  $X$  of information products in the broker, the purchase frequency of pairs of  $(x, i)$  with  $i \in X \setminus x$ . The

probability  $p_r(x \wedge i)$  that a buyer makes  $r$  purchases of products  $x$  and  $i$  at the same buying occasion which follow independent Poisson processes with means  $\mu_x$  and  $\mu_i$  is [12]:  $p_r(x \wedge i) = \frac{e^{-\mu_x} \mu_x^r}{r!} \frac{e^{-\mu_i} \mu_i^r}{r!}$ . For our recommender services for product  $x$  we need the conditional probability that product  $i$  has been used under the condition that product  $x$  has been used in the same session. It is easy to see that the conditional probability  $p_r(i | x)$  is again Poisson distributed by

$$\begin{aligned} p_r(i | x) &= \frac{p_r(x \wedge i)}{p_r(x)} \\ &= \frac{\frac{e^{-\mu_x} \mu_x^r}{r!} \frac{e^{-\mu_i} \mu_i^r}{r!}}{\frac{e^{-\mu_x} \mu_x^r}{r!}} \\ &= \frac{e^{-\mu_i} \mu_i^r}{r!} \end{aligned}$$

Because of the independence assumptions outlined above, the frequency distribution that such pairs occur 1, 2, 3, ..., -times, follows a logarithmic series distribution by the same line of reasoning as above.

We expect that non-random occurrences of such pairs occur more often than predicted by the logarithmic series distribution and that we can identify non-random occurrences of such pairs and use them as recommendations.

We can estimate this logarithmic series distribution for the whole market (over all consumers) from market baskets, that is from anonymous web-sessions. The limitation is that we can not analyze the behavior of different types of consumers (e.g. light and heavy buyers).

What kind of behavior is captured by the LSD-model? Because of the independence assumptions, the LSD-model estimates the probability that a product pair has been used by chance  $r$ -times together in a session. For example, consider that a user reads – as his time allows – some Internet newspaper and that he uses an Internet-based train-schedule for his travel-plans. Clearly, the use of both information products follows independent stochastic processes. And because of this, we would hesitate to recommend to other users who read the same Internet newspaper the train schedule. The frequency of observing this pair of information products in one session is as expected from the prediction of the LSD-model.

Next, consider complementarities between information products: Internet users usually tend to need several information products for a task. E.g. to write a paper in a foreign language the author might repeatedly need an on-line dictionary as well as some help with  $\LaTeX$ , his favorite type-setting software. In this case, however, we would not hesitate to recommend a  $\LaTeX$ -online documentation to the user of the on-line dictionary. And the frequency of observing these two information products in the same session is (far) higher than predicted by the LSD-model.

---

Algorithm for computing recommendations:

---

1. Compute for all information products  $x$  in the market baskets the frequency distributions for repeat-purchases of the co-occurrences of  $x$  with other information products in a session, that is of the pair  $(x, i)$  with  $i \in X \setminus x$ . Several co-occurrences of a pair  $(x, i)$  in a single session are counted only once.
  2. Discard all frequency distributions with less than  $l$  observations.
  3. For each frequency distribution:
    - (a) Compute the **robust** mean purchase frequency  $w$  by trimming the  $x$  percentil of the high repeat-buy pairs.
    - (b) Estimate the parameter  $q$  for the LSD-model from  $w = \frac{-q}{(1-q)(\ln(1-q))}$  with either a bisection or Newton method.
    - (c) Apply a  $\chi^2$ -goodness-of-fit test with a suitable  $\alpha$  (e.g. 0.01 or 0.05) between the observed and the expected LSD distribution with a suitable partitioning.
    - (d) Determine the outliers in the tail. (We suggest to be quite conservative here: Outliers at  $r$  are above  $\sum_r^\infty p_r$ .)
    - (e) Finally, we prepare the list of recommendations for information product  $x$ , if we have a significant LSD-model with outliers.
- 

TABLE I  
ALGORITHM FOR COMPUTING RECOMMENDATIONS

A *recommendation* in this setting simply implies that co-occurrences occur more often than expected from independent random choice acts and that a recommendation reveals a complementarity between information products.

The main purpose of the LSD-model in this setting is to separate non-random co-occurrences of information products (outliers) from random co-occurrences (as expected from the LSD-model). We use the LSD-model as a benchmark for discovering regularities.

Finally, we show a short overview of the algorithm for computing recommendations in table I.

Note, that in step 1 of the algorithm repeated usage of two information products in a single session is counted once as required in repeat-buying theory.

In addition, ignoring high-repeat buy outliers by trimming the sample (step 3a) considerably improves the chances of finding a significant LSD-model. This is supported by the data in column V of table III.

Several less conservative options for determining the outliers in the tail of the distribution (step 3d) are discussed in the next section.

## Java Code Engineering & Reverse Engineering

Persons using the above entry used the following entries too:

1. Free Programming Source Code
2. Softwareentwicklung: Java
3. Developer.com
4. Java-Einfuehrung
5. The Java Tutorial
6. JAR Files
7. The Java Boutique
8. Code Conventions for the Java(TM) Programming Language
9. Working with XML: The Java(TM)/XML Tutorial
10. Java Home Page
11. Java Commerce
- === Cut =====
12. Collection of Java Applets
13. Experts Exchange
- === Cut =====
14. The GNU-Win32 Project
15. Microsoft Education: Tutorials
16. HotScripts.com
- ...

Fig. 3. List of entries with cuts

### III. A SMALL EXAMPLE: JAVA CODE ENGINEERING & REVERSE ENGINEERING

In figure 3 we show the first 16 candidates for recommendations of the list of 117 web-sites generated for the site `Java Code Engineering & Reverse Engineering` by the method described in table I. 101 other information products have been found in market baskets together with this research site. The mean purchase frequency is 1.564. After trimming the highest 2.5 percentil (ignoring two observations with 7 and 8 repeat buys, respectively), the (robust) mean purchase frequency is 1.460, the parameter  $q$  of the LSD-model is 0.511. A  $\chi^2$  goodness-of-fit test is highly significant ( $\chi^2 = 1.099$  which is considerably below 3.841, the critical value at  $\alpha = 0.05$ ). This indicates that ignoring high repeat-buy outliers improves the fit of the LSD-model. Visual inspection of figures 4 and 5 shows that the theoretical LSD-model properly describes the empirical data. This is impression is supported by comparing the columns  $f(x)_{ob}$  and  $f(x)_{theo}$  in the second part of table II as well as looking at the  $\chi^2$ -values in table II. For more details, see table II.

All outliers whose observed repeat-purchase frequency is above the theoretically expected frequency are candidates to be selected as recommendations. In figure 5 (with a logarithmic y-axis) we explore three options of determining the cut-off point for such outliers:

*Option 1.* Without doubt, as long as the observed repeat-purchase frequency is above the cumulated theoretically expected frequency, we have detected outliers. In our example, this holds for all observations of more than 11 co-purchases which correspond to the top 11 sites shown as recommendations in figure 3. (This is the most conservative choice. Inspecting these recommendations shows that all of them are more or less directly related to Java programming, which is probably the task in which students use the example site.)

*Option 2.* Discounting any model errors, as long as the observed repeat-purchase frequency is above the theoretically expected frequency is a less conservative option. For the example, we select all co-purchases with more than 3 occurrences as recommendations. For the example, this coincides with the option described above. See the top 11 sites in figure 3.

*Option 3.* If we take the cut, where both cumulative purchase frequency distributions cross, we get 13 recommendations regarding all co-purchases occurring more than twice as nonrandom – see the top 13 sites in figure 3. However, it seems, that entries 12 and 13, namely `Collection pf Java Applets` and `Experts Exchange` seem to be not quite so related to Java programming.

Note, that the last three entries shown in figure 3 seem to be of little or no relevance for Java programming.

We have implemented the most conservative approach,

```

# File: wu01_74 (Mon May 7 16:48:37 2001)
# Heuristic was: distr=NBD - Case 4: NBD heuristic var>mean
# Heuristic was: mean= 1.56410256410256 Var=1.64456233421751
# Total number of observations: 117
#
# Robustify trimmed begin: 0 / end: 0.025 (2 observations)
# Robustify left mean: 1.46086956521739
# W/Robustify estimated q: 0.511090921020508

#Plot:
#Rep r f(x)ob f(x)theo 1-F(x)ob 1-F(x)theo
1 87 83.565419 117 117
2 17 21.354763 30 33.434580
3 2 7.276150 13 12.079816
4 5 2.789080 11 4.803665
5 3 1.140379 6 2.014585
6 1 0.485697 3 0.874205
7 1 0.212773 2 0.388508
8 1 0.095153 1 0.175734

# Getting fat tails:
# Method 1-F(x) intersection at: 2 (leaves 13 nonrandom outliers)
# Method f(x) intersection at: 3 (leaves 11 nonrandom outliers)
# Method mixed intersection (f(x) obs w/ 1-F(x) theo) at:
# 3 (leaves 11 nonrandom outliers)

#Chi Square Test:
#class obs theoretic chi2 trimmed chi2 trimmed
#1 87 83.565 0.141 87 0.141
#2 17 21.355 0.888 17 0.888
#3 13 12.080 0.070 2 0.097

# Sum of chisquare value: 1.09930205129959
# Sum of chisquare value trimmed: 1.12573175901045
# Test border (at 99% w/1 d.f.): 10.828 (95% would be 3.841)
# *** Significant at 95% ***

```

TABLE II  
STATISTICS FOR ENTRY WU01.74

namely option 1, in the recommender service based on a check of the face validity of the recommendations for a small sample of information products (25 products). We think that, at least in cases with a considerable number of candidates for the recommendation list, this is a suitable approach.

Next, let us be a little bit more precise about what constitutes an outlier. Consider, for example, the number of product combinations which have been bought 8 times together in table II. Theoretically, we would expect that this is a chance event roughly in one out of ten cases. Now, we have observed 5 product combinations with 4 repeat-buys. Unfortunately, theoretically 2.789 product combi-

nations can be expected from pure chance. In this class we observe now a mixture of random product combinations and non-random product combinations, but we are not able to distinguish them. However, we can specify a threshold for the chance of falsely presenting a random co-occurrence, e.g. below 0.40. In the example, we would then present the entries in classes 5, 7, and 8, but not the entry in class 6. That is, we would present entries 1, 2, 4, 5, and 6 in figure 3, but not entry 3, Developer.com and, indeed, this site definitely is not exclusively devoted to Java programming. In the analysis of outliers there is still room for improvement as e.g. by developing statistical test for identifying outliers.

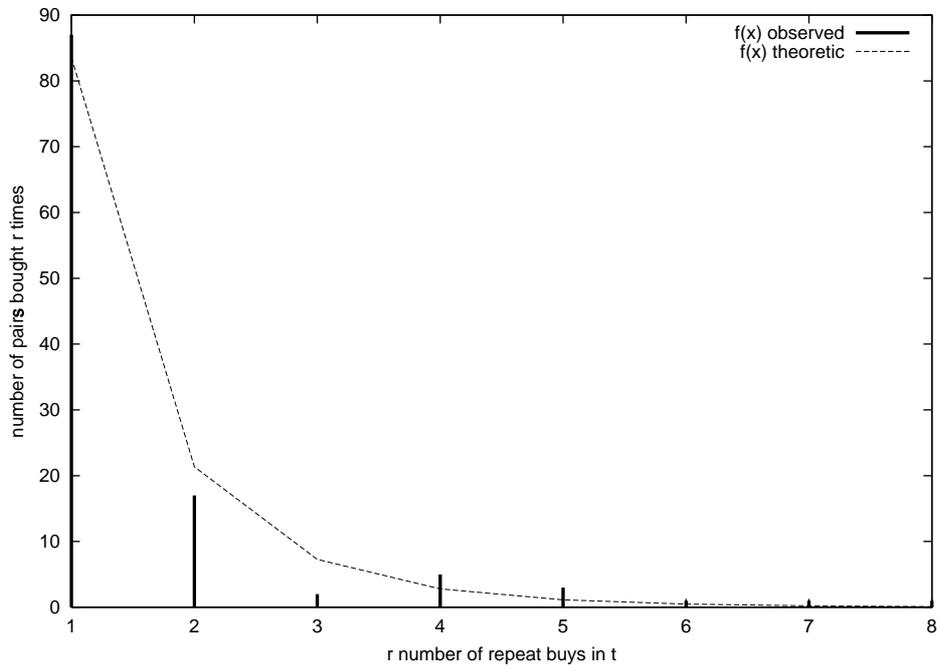


Fig. 4. Plot of frequency distribution

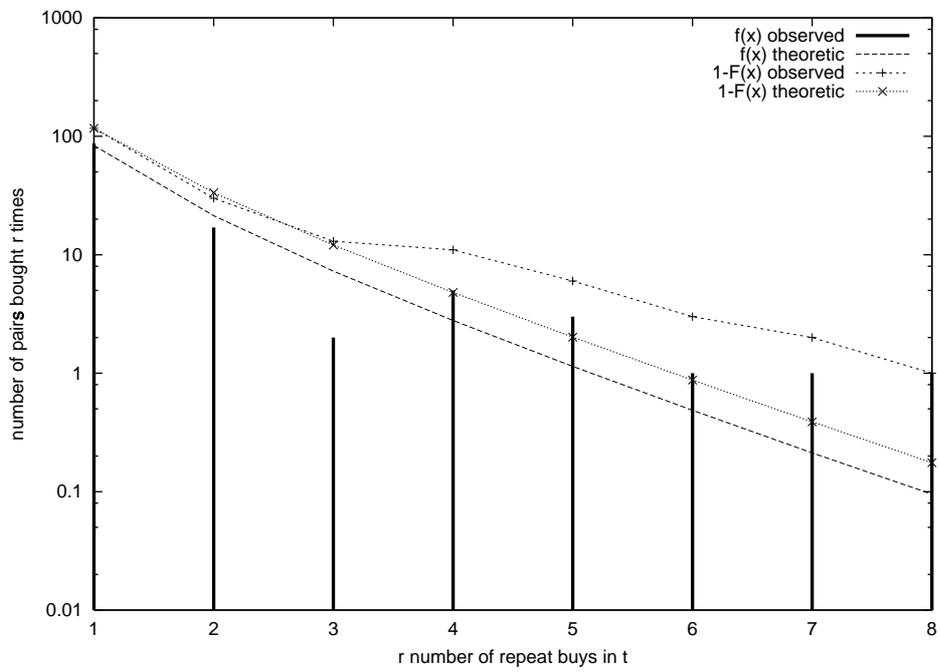


Fig. 5. Plot of log frequency distribution

	I $q$ undef.	II no $\chi^2$ ( $< 3$ classes)	III Sign. $\alpha = 0.05$	IV Sign. $\alpha = 0.01$	V Sign. (trim)	VI Not sign.	$\Sigma$
A							
Obs. $< 10$	1128 (0)	66 (63)	0 (0)	0 (0)	0 (0)	0 (0)	1194 (63)
B							
$\bar{x} = 1$	1374 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1374 (0)
C							
$\bar{x} > \sigma^2$ $r \leq 3$	2375 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2375 (0)
D							
$\bar{x} > \sigma^2$ $r > 3$	201 (0)	617 (605)	105 (105)	46 (46)	15 (15)	372 (145)	1356 (916)
E							
$\sigma^2 > \bar{x}$	3 (0)	86 (86)	222 (222)	194 (194)	93 (93)	253 (253)	851 (848)
$\Sigma$	5081 (0)	769 (754)	327 (327)	240 (240)	108 (108)	625 (398)	7150 (1827)

TABLE III

DETAILED RESULTS. (OBSERVATION PERIOD: 1999-09-01 – 2001-05-07). NUMBERS IN PARENTHESIS INDICATE LISTS WITH AT LEAST 1 OUTLIER

#### IV. FIRST EMPIRICAL RESULTS

To establish that a recommender service based on Ehrenberg’s repeat buying-theory is supported by empirical evidence, we proceed as follows:

1. In section IV-A we investigate, how well the LSD-model explains the actual data for 7150 information products.
2. However, that the LSD-model fits the data well, does not yet mean that the outliers we have identified are suitable recommendations for a user. In section IV-B we present the results of a first small face evaluation experiment whose result suggests that these outliers are indeed valuable recommendations.

The data set used for the example given in section III and in this section is from the anonymous recommender services of the Virtual University of the Vienna University of Economics and Business Administration (<http://vu.wu-wien.ac.at>) for the observation period from 1999-09-01 to 2001-03-05. Co-occurrences have been observed for 8596 information products, After elimination of web-sites which ceased to exist in the observation period, co-occurrences for 7150 information products remain available for analysis.

##### A. Fit of Data to LSD-Models

Table III summarizes the results of applying the algorithm for computing recommendations presented in table I. If the sample variance is larger than the sample

mean, this may indicate that an NBD-model (and thus its LSD-approximation) is appropriate (see [11, p.138]). This heuristic suggests 851 candidates for an LSD-model (see table III, row E).

The rows of the table represent the following cases:

A The number of observations is less than 10. In this row we find co-occurrence lists either for very young or for very rarely used web-sites. These are not included into the further analysis. Cell (A/II) in table III contains lists which have repeated co-occurrences despite the low number of observations. In this cell good recommendation lists may be present (4 out of 5).

B No repeat buys, just one co-occurrence. These are discarded from further analyses.

C Less than 4 repeat-buys and trimmed sample mean larger than variance. Trimming outliers may lead to the case that only the observations of class 1 (no repeat buys) remain in the sample. These are discarded from further analyses.

D More than 3 repeat-buys and trimmed sample mean larger than variance. In cell (D/I) after trimming only class 1 entries remain in the trimmed sample (no repeat buys). As a future improvement, the analysis should be repeated without trimming. In cell (D/II) the  $\chi^2$ -test is not applicable, because less than 3 classes remain.

E (Trimmed) sample variance larger than sample mean. For cell (E/I) we recommend the same as for cell (D/I). For cell (E/II) we observe the same as for cell (D/II).

The columns I – VI of table III have the following meaning:

*I* The parameter  $q$  of the LSD model could not be estimated. For example, only a single co-occurrence has been observed for some product pairs.

*II* The  $\chi^2$  goodness-of-fit test could not be computed, because of lack of observations.

*III, IV* The  $\chi^2$  goodness-of-fit test is significant at  $\alpha = 0.05$  or  $\alpha = 0.1$ , respectively. The LSD-model was estimated without trimming outliers.

*V* The  $\chi^2$  goodness-of-fit test is significant at  $\alpha = 0.1$  for LSD-models estimated with trimmed data. All high-repeat buy pairs in the 2.5 percentil have been excluded from the model estimation.

*VI* The  $\chi^2$ -test is not significant.

	n	%
Information products	9498	100.00
Products bought together with other products	7150	75.28
Parameter $q$ defined	2069	21.78
Enough classes for $\chi^2$ -test	1300	13.69
LSD with $\alpha = 0.01$ (robust)	675	7.11
LSD not significant	625	6.58
LSD fitted, no $\chi^2$ -test	703	7.40
$n < 10$ and no $\chi^2$ -test	66	0.69

TABLE IV

SUMMARY OF RESULTS. (OBSERVATION PERIOD: 1999-09-01 – 2001-05-07)

As summarized in table IV we fitted a LSD-model for the frequency distributions of co-occurencies for 1300 information products. For 675 information products, that is more than 50 percent, the estimated LSD-models pass a  $\chi^2$  goodness-of-fit test at  $\alpha = 0.01$ .

### B. Face Validation of Recommendations

In order to establish the plausibility of the recommendations identified by the recommender service previously described, we performed a small scale face validation experiment. The numbers in parenthesis in table III indicate the number of lists for which outliers were detected. From these lists 100 lists of recommendations were randomly selected. Each of the 1259 recommendations in these lists was inspected for plausibility. Plausible recommendations were counted as good recommendations by pressing the affirmative symbol (a hook) in the Vote Box shown in figure 1.

This small scale face validation experiment of inspecting recommendations for plausibility led to a quite satisfactory result:

- For the 31 lists for which a significant LSD-model could be fitted, 87.71 % of the recommendations were

judged as good recommendations.

- 25 lists for which a LSD model was not significant contained 89, 45 % good recommendations.

- Only 75.74 % good recommendations were found in the 44 lists for those LSD-models where no  $\chi^2$  test could be computed which is a significantly lower percentage.

Surprisingly, the class of models where the LSD model was not significant contains a slightly higher number of recommendations evaluated as good. However, a number of (different) reasons may explain this:

- First, we might argue that even if the LSD-model is insignificant, it still serves its purpose, namely to identify non-random outliers as recommendations.

- A close inspection of frequency distributions for these lists revealed the quite unexpected fact that several of these frequency distributions were for information products which belong to the oldest in the data set and which account for many observations. The reasons for this may be explained e.g. by a shift in user behavior (non-stationarity) or too regular behavior (as e.g. for cigarettes in consumer markets [6]). If too regular behavior is the reason that the LSD-model is insignificant, again, we still identified the non-random outliers.

- An other factor which might contribute to this problem is that several entries in this group belong to lists integrated in the web-sites of other organizational units. These lists, at least some of them, contain web-sites which have been carefully selected by the webmasters of these organizational units for their students. For example, the list of web-sites for student jobs is integrated within the main web-site of the university. The recommendations for such lists seem to reflect mainly the search behavior of the users. A similar effect is known in classic consumer panel analysis, if if the points of sale of different purchases are not cleanly separated. This implies that e.g. purchases in a supermarket are not distinguished from purchases from a salesman. In our analysis, the purchase occasions are in different web-sites, namely the broker system and the organizational web-site with the embedded list. Ehrenberg's recommendation is to analyse the data separately for each purchase occasion.

Also, the fact that the data set contains information products with different age may explain some of these difficulties. However, to settle this issue further investigations are required.

## V. FURTHER RESEARCH

The main contribution of this paper is that Ehrenberg's classical repeat-buying models can be applied to market baskets and describe – despite their strong independence and stationarity assumptions – the consumption patterns of information products – at least for the data set analyzed – surprisingly well. For e-commerce sites this implies, that a large part of the theory developed for con-

sumer panels may be applied to market basket data too, as long as the analysis remains on the aggregate level.

For anonymous recommender services they seem to do a remarkable job of identifying non-random repeated-choice acts of consumers of information products as we have demonstrated in section III. The use of the LSD-model for identifying non-random co-occurrences of information products constitutes a major improvement which is not yet present in other correlation-type recommender services.

However, establishing an empirical base for the validity of repeat-buying models in information markets as suggested in this article still requires a lot of additional evidence and a careful investigation of additional data sets. We expect that such an empirical research program would have a good chance to succeed, because to establish Ehrenberg's repeat-buying theory a similar research program has been conducted by Aske Research Ltd., London, in several consumer product markets (e.g. dentifrice, ready-to-eat cereals, detergents, refrigerated dough, cigarettes, petrol, tooth-pastes, biscuits, colour cosmetics, ...) from 1969 to 1981 [6].

The current version of the anonymous recommender services (and the analysis in this article) still suffers from several deficiencies. The first is that new information products are daily added to the information broker's data base, so that the stationarity assumptions for the market are violated and the information products in the data set are of non-homogenous age. The second drawback is that testing the behavioral assumptions of the model, e.g. by testing the behavioral assumptions with data from the personalized part of the VU, as well as validation either by studying user acceptance or by controlled experiments still has to be done. Third, for performance reasons the co-occurrence lists for each information product do not contain time-stamps. Therefore, the development of time-dependent e.g. alert systems has not been tried, although Ehrenberg's theory is in principle suitable for this task.

We expect Ehrenberg's repeat-buying models to be of considerable help to create anonymous recommender services for recognizing emerging shifts in consumer behavior patterns (fashion, emerging trends, moods, new sub-cultures, ...). Embedded in a personalized environment Ehrenberg's repeat-buying models may serve as the base of continuous marketing research services for managerial decision support which provide forecasts and classical consumer panel analysis in a cost efficient way.

## VI. ACKNOWLEDGEMENT

We acknowledge the financial support of the Jubiläumsfonds of the Austrian National Bank under Grant No. 7925 without which this project would not have been possible.

## REFERENCES

- [1] AVERY, C. and ZECKHAUSER, R. (1997): Recommender Systems for Evaluating Computer Messages. *Communications of the ACM*, 40(3), 88–89.
- [2] BELLMANN, S., LOHSE, G.L., and JOHNSON, E.J. (1999): Predictors of Online Buying Behavior. *Communications of the ACM*, 42(12), 32–38.
- [3] BLISCHOK, T.J. (1995): Every Transaction Tells a Story. *Chain Store Age Executive*, 71(3), 50–62.
- [4] BOEHM, W., GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2001): Repeat Buying Theory and its Application for Recommender Services. Submitted.
- [5] CHATFIELD, C., EHRENBURG, A. S. C., GOODHARDT, G. J. (1966): Progress on a Simplified Model of Stationary Purchasing Behavior. *Journal of the Royal Statistical Society A*, 129, 317–367.
- [6] EHRENBURG, A. S. C. (1988): *Repeat-Buying: Facts, Theory and Applications*. Charles Griffin & Company Limited, London.
- [7] GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2000): myVU: A Next Generation Recommender System Based on Observed Consumer Behavior and Interactive Evolutionary Algorithms. In: W. Gaul, O. Opitz, M. Schader (Eds.): *Data Analysis – Scientific Modeling and Practical Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, Vol. 18, Springer, Heidelberg, 447–457.
- [8] GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2001): Educational and Scientific Recommender Systems: Designing the Information Channels of the Virtual University. *International Journal of Engineering Education*, 17(2), 153–163.
- [9] GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2001): Recommendations for Virtual Universities from Observed User Behavior. In: W. Gaul, Ritter, M. Schader (Eds.): *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Heidelberg, to appear.
- [10] GOODHARDT, G.J., EHRENBURG, A.S.C., COLLINS, M.A. (1984): The Dirichlet: A Comprehensive Model of Buying Behaviour. *Journal of the Royal Statistical Society, A*, 147, 621–655.
- [11] JOHNSON, N.L., KOTZ, S. (1969): *Discrete Distributions*. Houghton Mifflin, Boston.
- [12] JOHNSON, N.L., KOTZ, S., BALAKRISHNAN N. (1997): *Discrete Multivariate Distributions*. John Wiley & Sons, New York.
- [13] RESNICK, P. and VARIAN, H.R. (1997): Recommender Systems. *Communications of the ACM*, 40(3), 56–58.
- [14] WAGNER, U. and TAUDES, A. (1987): Stochastic Models of Consumer Behaviour. *European Journal of Operations Research*, 29(1), 1–23.