

# **A Customer Purchase Incidence Model Applied to Recommender Services**

WEBKDD 2001

August 26, 2001, San Francisco, CA

---

Andreas Geyer-Schulz,  
Informationsdienste und Elektronische Märkte, Universität Karlsruhe (TH), D-76128  
Karlsruhe.

Michael Hahsler,  
Informationswirtschaft, Wirtschaftsuniversität Wien, A-1090 Wien.

Maximilian Jahn,  
Informationswirtschaft, Wirtschaftsuniversität Wien, A-1090 Wien.

# **Table of Contents**

---

- Introduction and Motivation
- Recommendations for a Broker System
- Ehrenberg's Repeat Buying Theory for Bundles of Information Products
- A Small Example
- First Empirical Results
- Further Research

# **Introduction**

---

1. We transfer a consumer purchase incidence model to Web-based information products, and
2. we build an anonymous recommender service based on repeated cross-selling.

We use:

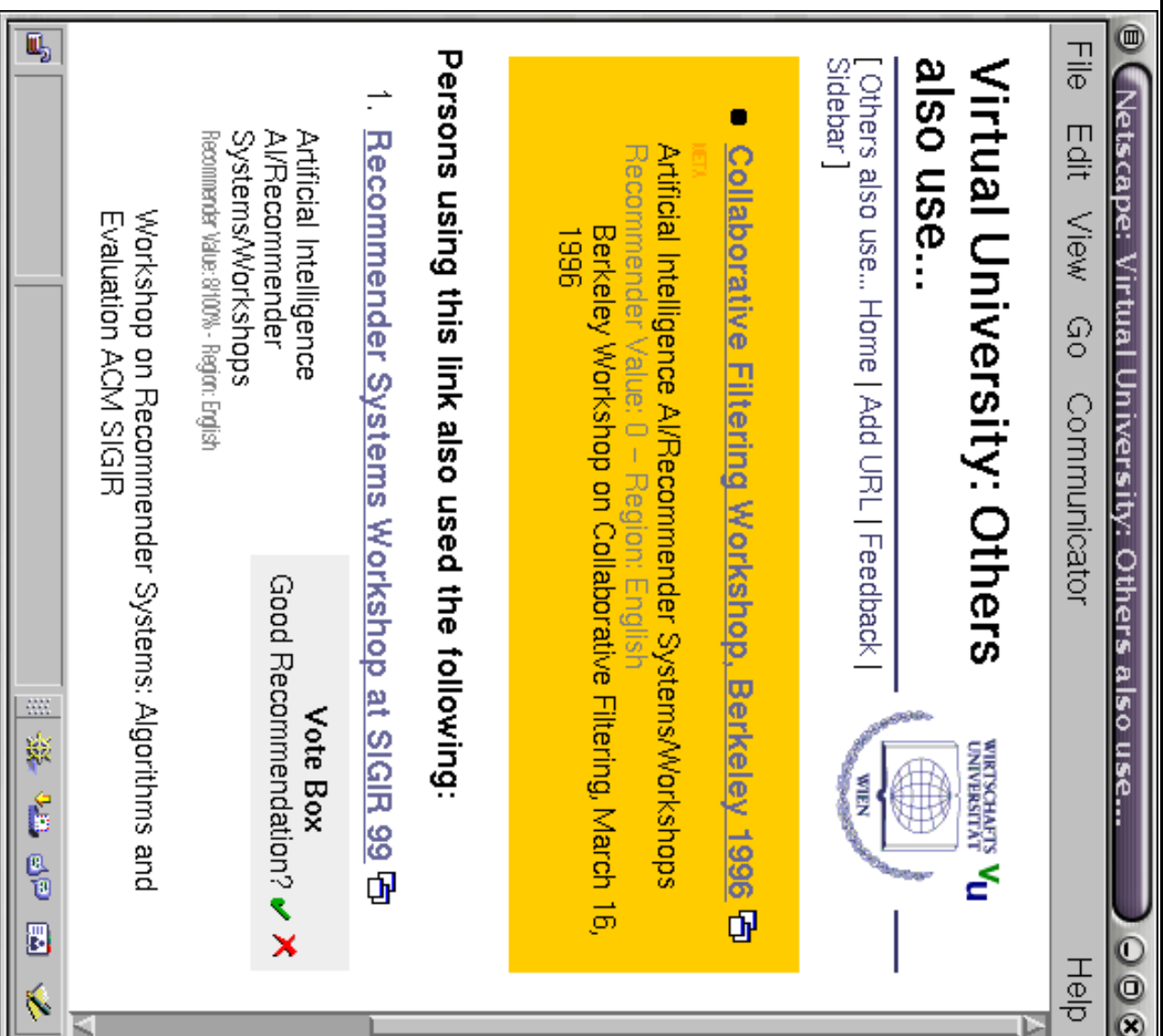
- Information market (broker)
- Information products (Web-sites)
- Observed consumer behavior (purchases of information products)
- Market baskets (Web browser sessions)
- Item-to-item similarity, correlation
- No purchase history (anonymous users)

# **Advantages of Recommender based on the Customer Purchase Incidence Model**

---

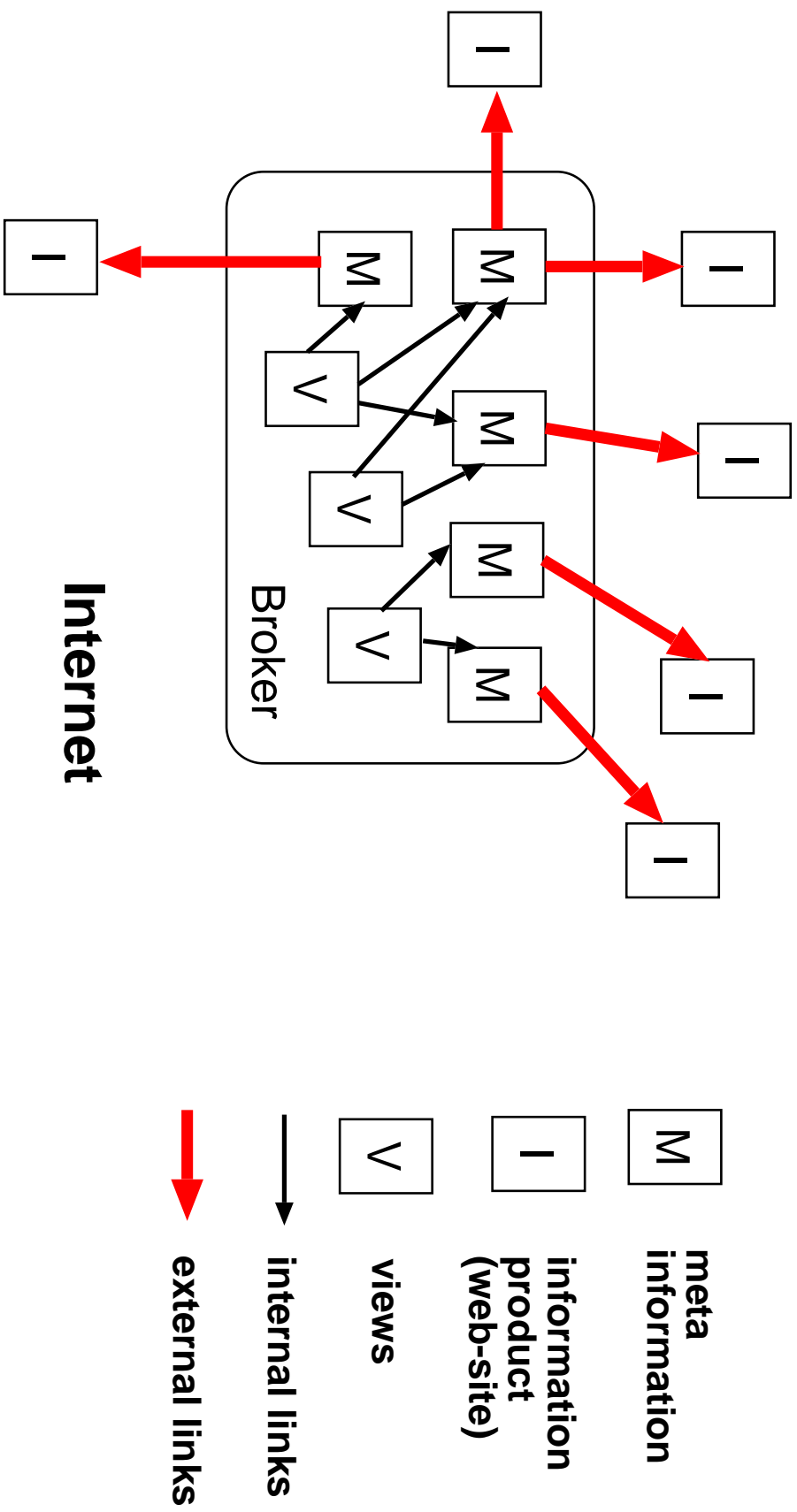
- Observed consumer behavior is the most important information for predicting consumer behavior online and offline.
- Market basket analysis shows up to 70 percent cross-selling potential.
- Facilitates “repeat-buying”.
- Not subject to incentive problems of explicit recommendations. (Transaction cost of fakes high, no free-riding, ...)
- Privacy preserved (anonymous service).
- Low transaction cost for broker, fully automatic. (No editor, author, web-scout, ...)

# An Motivational Example



# Recommendation I - Information Market with Broker

---



1. We treat web-sites as information products.
2. Following an external link = purchase of an information product

# **Recommendation II - Problems**

- Co-occurrence of combinations of Web-sites used in sessions (cross-selling)
- Repeat cross-selling
- Anonymous user sessions

## **Questions**

- **How to measure repeat usage for anonymous user sessions?**
- **Which co-occurrences qualify as non-random?**
- How many products should be recommended?
- How should the recommendations be presented?

# Ehrenberg's Repeat Buying Theory I

- Descriptive theory of consumption behavior
- Generalization of regularities
- Strong empirical evidence for several hundred consumer product markets since the 1950's

*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

# Ehrenberg's Repeat Buying Theory II

## **Consumer's decision:**

1. Whether/When does a consumer buy a certain product-class?
2. If so, what brand does he buy? (brand choice)

## **-> Formalizing the purchase process:**

- Concept of the “purchase occasion”
- Analysis variables
  1. analysis period
  2. penetration - proportion of customers who bought the product
  3. average purchase frequency - average number of times the product is bought
- Market is in equilibrium (stationary condition)

# Ehrenberg's Repeat Buying Theory III

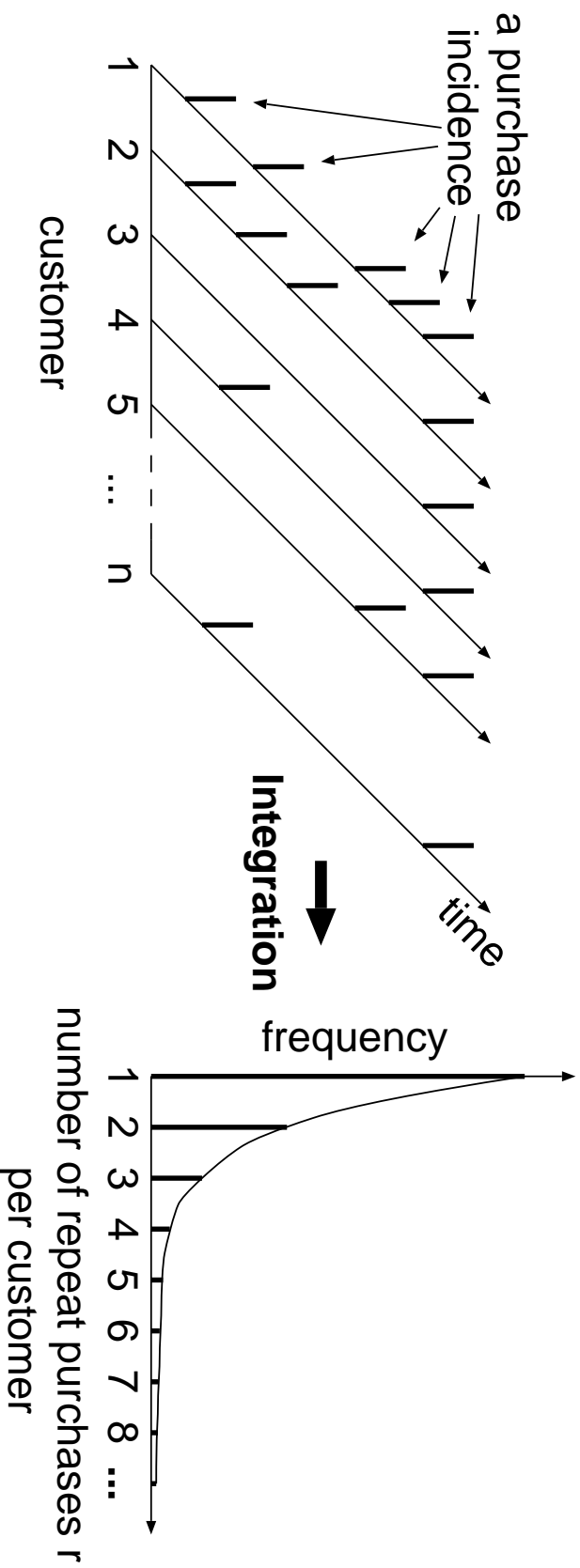
## Models and parameters:

- Negative binomial distribution (NBD) model
  - $m$  - Average number of purchases per informant
  - $k$  - Negative binomial exponent (estimated with penetration  $b$ )
- **Logarithmic series distribution (LSD) model:** A simpler model where the penetration is unknown but below 20%
  - $q$  - estimated from  $W$  (average number of purchases by buyer = mean purchase frequency)
- Dirichlet model

# The LSD Model

Independent Stochastic Purchase Processes  
for Product  $x$  or Product-Class  $x$

Frequency distribution of  
Repeat Purchases  
for Analysis period  $t$



1. Purchase processes of customers follow independent, stationary Poisson processes
2. The parameters  $\mu$  follow a truncated  $\Gamma$ -distribution

-> The frequency distr. follows a logarithmic series distribution

# LSD Model: Proof

---

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:

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

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

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

$$\begin{aligned} p_r &\approx \left( \frac{c}{r!(1+\frac{1}{a})^r} \right) \Gamma(r) \\ &= \frac{c}{(1+\frac{1}{a})^r r} \\ &= c \frac{q^r}{r} \\ &= qp_{r-1}(r-1)/r \end{aligned}$$

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.)

## LSD Model: Usage

The logarithmic series distribution (LSD) describes how many buyers buy a specific product 1, 2, 3, ... times (without taking into account the number of non-buyers):

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

Mean purchase frequency:

$$\mu = \frac{-q}{(1-q) \ln(1-q)} \quad (2)$$

Variance:

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

# **LSD Model: Assumptions**

---

- **Share of never-buyers in the population is not specified.**
  - > True for most services on the Web
- **Purchases of a consumer in successive periods follow a Poisson distribution with a certain long-run average  $\mu$ .**
  - > Purchases tend to be independent of previous purchases and they occur in such an irregular manner that they can be regarded as if random
- **The distribution of  $\mu$  in the population follows a truncated  $\Gamma$ -distribution.**
  - > A quite general shape and also follows independence assumptions
- **The market is in equilibrium (stationary).**
  - > Empirical analysis shows that most markets for consumer products are most of the time in a near equilibrium state.

# Combination of products

- Two products  $x, i$
- Two independent purchase processes (Poisson processes with means  $\mu_x$  and  $\mu_i$ )

$$P_r(x \wedge i) = \frac{e^{-\mu_x} \mu_x^r}{r!} \frac{e^{-\mu_i} \mu_i^r}{r!} \quad (4)$$

$$P_r(i | x) = \frac{P_r(x \wedge i)}{P_r(x)} = \frac{\frac{e^{-\mu_x} \mu_x^r}{r!} \frac{e^{-\mu_i} \mu_i^r}{r!}}{\frac{e^{-\mu_x} \mu_x^r}{r!}} = \frac{e^{-\mu_i} \mu_i^r}{r!} \quad (5)$$

-> The frequency distribution follows again LSD

# Consumer Panel vs. Browser Sessions

The repeat buying theory uses information from consumer panels

<b>Consumer Panel</b>	<b>Browser Sessions</b>
consumer products	web-sites
purchase incidence	selection of a web-site
number of items bought ignored	repeat visits per session ignored
package size ignored	number of pages browsed ignored
identity of customer known	anonymous user sessions
purchase history known	history unknown
non-buyers known	non-buyers unknown

- Browser sessions can be used as consumer panels with **unobserved consumer identity**. The identity is not needed to use the model at the aggregate level!
- Non-buyers are not needed for the LSD model

# **Back to the problems**

## **1. How to measure repeat usage for anonymous user sessions?**

The model already implies repeat-buying in aggregated data.

## **2. Which co-occurrences qualify as non-random?**

The model is based on strict independence assumptions. It estimates the probability that a product combination is used together  $r$ -times by chance.

We expect non-random choices (complementarity between two products) to occur more often than the model would predict.

# Algorithm

---

1. Compute for all information products  $x$  in the browser sessions the frequency distributions for repeat-purchases of the co-occurrences of  $x$  with other information products in a session.
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  $x$  percentile of the high repeat-buy pairs.
  - (b) Estimate the parameter  $q$  for the LSD-model from
$$w = \frac{-q}{(1-q)(\ln(1-q))}$$
  - (c) Apply a  $\chi^2$ -goodness-of-fit test with a suitable  $\alpha$  between the observed and the expected LSD distribution with a suitable partitioning.
  - (d) Determine the outliers in the tail.
  - (e) Finally, prepare the list of recommendations for information product  $x$ , if the LSD-model is significant and outliers exist.

# A Small Example I

Java Code Engineering & Reverse Engineering

1. Free Programming Source Code
2. Softwareentwicklung: Java
3. Developer.com
4. Java-Finfuehrung
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
- .. .
117. GNU's Not Unix! - the GNU Project and the FSF

# A Small Example II

---

```
# File: wu01_74 (Mon May 7 16:48:37 2001)
# Robustify left mean: 1.46086956521739, estimated q: 0.51109092
#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

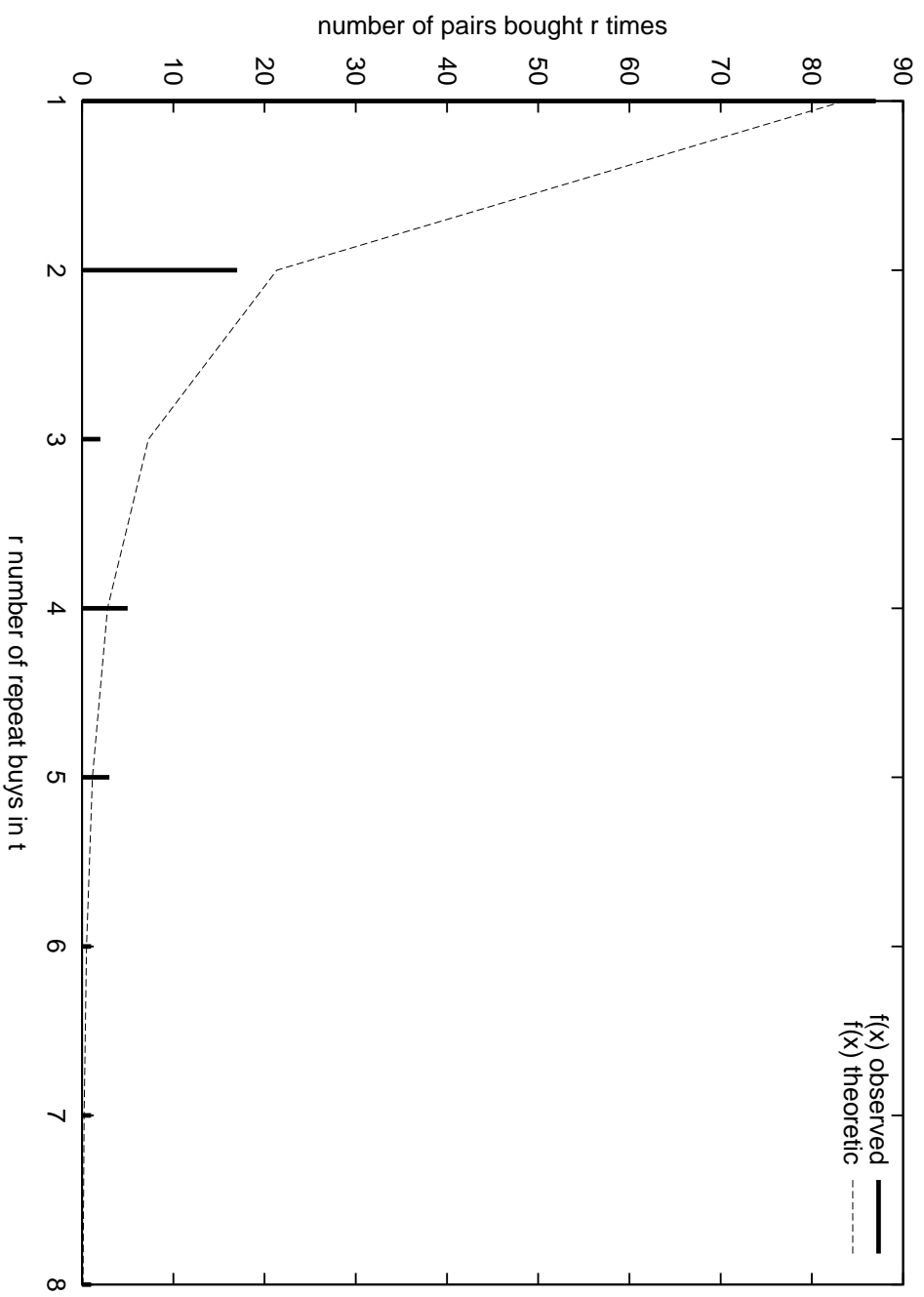
# Getting fat tails:
# Method 1-F(x) intersection at: 2 (leaves 13 nonrandom outliers)
# Method f(x) intersection at: 3 (leaves 11 nonrandom outliers)
# Method mixed intersection (f(x) obs w/ 1-F(x) theo) at:
#      3 (leaves 11 nonrandom outliers)

#Chi Square Test:
#class  obs      theoretic  chi2      trimmed      chi2  trimmed
#1      87      83.5665      0.141      87       0.141
#2      17      21.3555      0.888      17       0.888
#3      13      12.0800      0.070      2        0.097

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

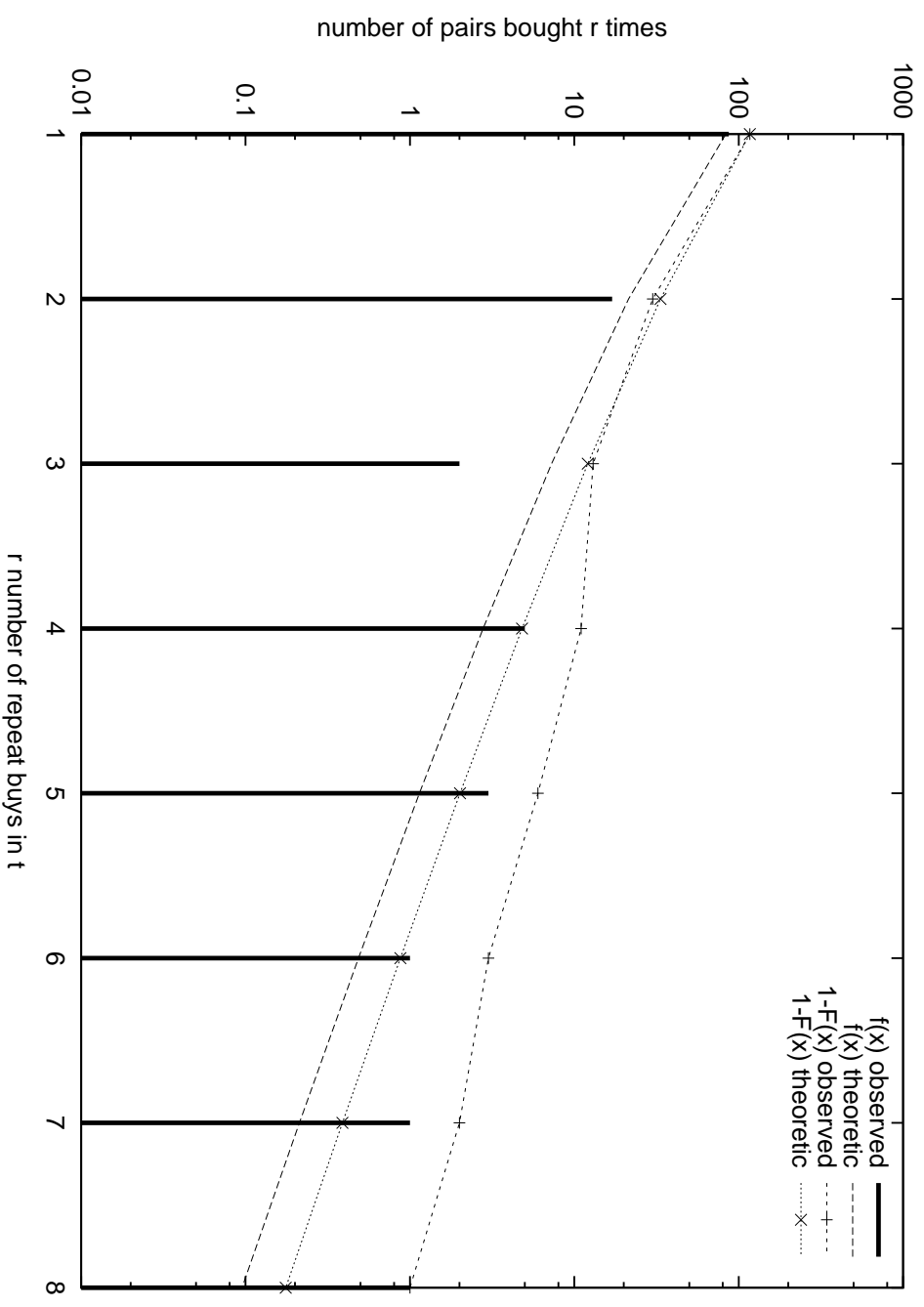
# A Small Example III

---



# A Small Example IV

---



# First Empirical Results

---

## Questions:

1. How well does the LSD-model explain our actual data?
2. Are the outliers valuable recommendations?

## Data set:

Observed co-occurrences for 8596 information products in the information broker of the Virtual University of the Vienna University of Economics and Business Administration (<http://vu.wu-wien.ac.at>) from January 1999 to May 2001.

# Fit of LSD-Models to the Data Set

We used the  $\chi^2$  goodness-of-fit test with  $\alpha = 0.01$  and  $\alpha = 0.05$

	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	100.00
LSD with $\alpha = 0.05$	327	3.44	25.15
<b>LSD with <math>\alpha = 0.01</math> (robust)</b>	<b>675</b>	<b>7.11</b>	<b>51.92</b>
LSD not significant	625	6.58	48.08
LSD fitted, no $\chi^2$ -test	703	7.40	
$n < 10$ and no $\chi^2$ -test	66	0.69	

# Face Validation of Recommendations

## Sample:

- Recommendation lists for 100 information products were randomly selected (from the generated 1827 lists)
- For these products 59750 co-occurrences where found
- The lists contain 1259 outliers (recommendations)
- Each of the 1259 recommendations was inspected for plausibility

## Percentage of plausible recommendations:

- 31 lists, significant LSD-model, 87.71 %
- 25 lists, LSD model not significant, 89.45 %
- 44 lists, no  $\chi^2$  test, 75.74 %

-> Even if the LSD-model is not significant, it still seems useful to identify non-random outliers

# **Contribution of this work**

- Anonymous Web browser sessions as consumer panels with unobserved identity.
- Evidence that information products also follow the repeat-buying patterns formalized by the repeat-buying theory.
- Evidence that the repeat-buying theory provides a good foundation to produce valuable recommendations.

# Further Research

---

- Establishing an empirical base for the validity of repeat-buying models for information products still requires additional evidence and careful investigation of additional data sets. E.g. Test the model's assumptions using a personalized a recommender system.
- Elimination of deficiencies of the current version of the anonymous recommender service used for this paper:
  - Avoid non-homogeneous age structure of products
  - Include time information to analyse fashions, emerging trends,...
  - Analyse the influence of linking from organizational units of the university directly to selected parts of the broker.
- Compare the system with other approaches (association rules using support/confidence, chi-squared test for correlation).
- Review the literature about customer incidence models and evaluate their utility for information product markets.