

# Modeling Purchase and Browsing Behavior Using Clickstream Data

Alan Montgomery  
*Associate Professor*  
**Carnegie Mellon University**  
Graduate School of Industrial Administration

e-mail: [alan.montgomery@cmu.edu](mailto:alan.montgomery@cmu.edu)  
web: <http://www.andrew.cmu.edu/user/alm3>

*Choice Symposim, 2 June 2001*



## Outline

- What is clickstream data?
- User Profiling
- Modeling Browsing/Shopping Behavior
- Shopbot Design
- Future Research Directions

@

# My/Our World View





## Defining Clickstream Data

## What is clickstream data?

- A record of an individual's movement through time at a web site
- Contains information about:
  - Time
  - URL content
  - User's machine
  - Previous URL viewed
  - Browser type

7

## A clickstream example

Household ID: Female born 12Jul42

Demographics: Philadelphia Area, Male and Female Married (husband born: 27Sep46), 3 members in household, income: \$75,000-\$99,999, Graduated College, employed 35 or more hours, 1 child age 13 to 17 (daughter born: 5Jul80), own single family home, white collar, own car & truck, microwave, three dogs, five cats

|                  |     |  |
|------------------|-----|--|
| 18JUL97:18:55:57 | 47  | <a href="http://www.voicenet.com/">www.voicenet.com/</a>   |
| 18JUL97:18:56:44 | 37  | <a href="http://www.weather.com/weather/us/cities/HI_Lahaina.html">www.weather.com/weather/us/cities/HI_Lahaina.html</a>             |
| 18JUL97:18:57:25 | 105 | <a href="http://www.weather.com/weather/us/cities/MI_Traverse_City.html">www.weather.com/weather/us/cities/MI_Traverse_City.html</a> |
| 18JUL97:19:03:00 | 7   | <a href="http://www.weather.com/weather/us/cities/MI_Traverse_City.html">www.weather.com/weather/us/cities/MI_Traverse_City.html</a> |
| 18JUL97:19:03:56 | 2   | <a href="http://www.weather.com/weather/us/cities/MI_Traverse_City.html">www.weather.com/weather/us/cities/MI_Traverse_City.html</a> |
| 18JUL97:19:03:58 | 6   | <a href="http://www.weather.com/weather/us/cities/HI_Lahaina.html">www.weather.com/weather/us/cities/HI_Lahaina.html</a>             |
| 18JUL97:19:04:58 | 2   | <a href="http://www.weather.com/weather/us/cities/HI_Lahaina.html">www.weather.com/weather/us/cities/HI_Lahaina.html</a>             |
| 18JUL97:19:05:00 | 1   | <a href="http://www.weather.com/weather/us/cities/MI_Traverse_City.html">www.weather.com/weather/us/cities/MI_Traverse_City.html</a> |
| 18JUL97:19:15:24 | 39  | <a href="http://www.weather.com/weather/us/cities/MI_Traverse_City.html">www.weather.com/weather/us/cities/MI_Traverse_City.html</a> |
| 18JUL97:19:17:00 | 7   | <a href="http://www.weather.com/weather/us/cities/MI_Traverse_City.html">www.weather.com/weather/us/cities/MI_Traverse_City.html</a> |
| 18JUL97:19:17:07 | 13  | <a href="http://www.realastrology.com/">www.realastrology.com/</a>   |
| 18JUL97:19:17:20 | 44  | <a href="http://www.realastrology.com/libra.html">www.realastrology.com/libra.html</a>   |

8

## What does the provider see?

*www.voicenet.com would be able to see:*

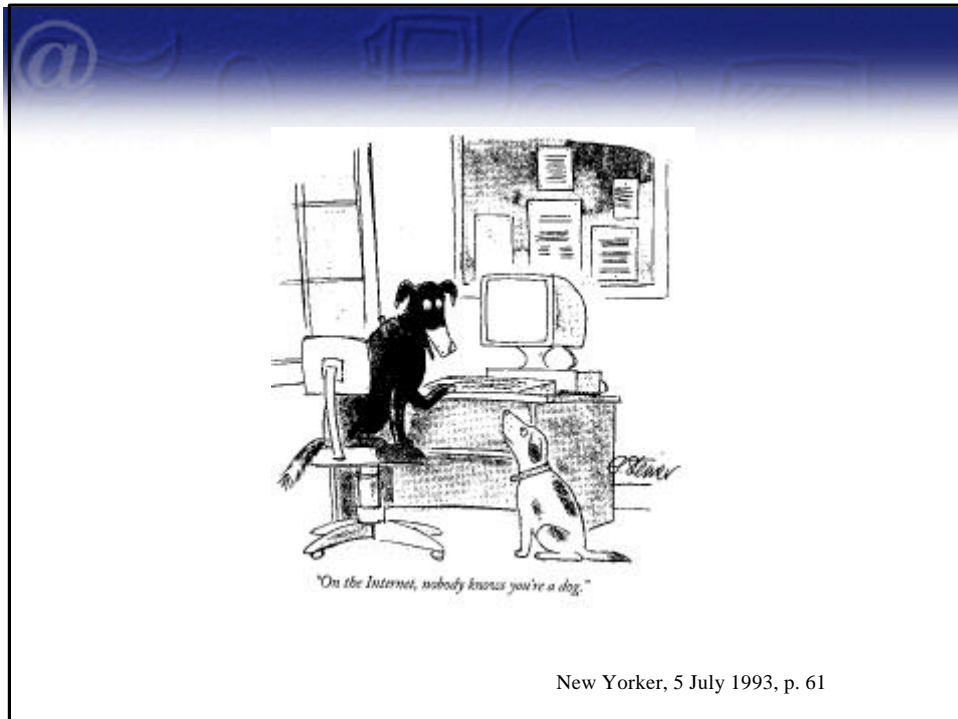
|                  |  |
|------------------|--|
| 18JUL97:18:56:44 | www.voicenet.com/<br>www.weather.com/weather/us/cities/HI_Lahaina.html |
| 18JUL97:18:57:25 | www.weather.com/weather/us/cities/MI_Traverse_City.html                |

*What would www.weather.com see?*


9

## User Profiling with Clickstream Data

*Does 'Where you go' define 'Who you are'?*



## Clickstream Example: Who is this web person?



Example

| Session Number | Start Time           | Seconds Actively Viewing | Seconds Elapsed between start and stop | URL   |
|----------------|----------------------|--------------------------|--|---|
| 1              | 06.Jul.1995:20:49:31 | 4                        | 100                                    | http://www.mserrysbe.anies.com/mserrysbe.anies/pokacksch.htm        |
| 2              | 06.Jul.1995:20:51:11 | 52                       | 132                                    | http://www.mserrysbe.anies.com/mserrysbe.anies/pokacksch.htm        |
| 3              | 06.Jul.1995:20:53:23 | 51                       | 85                                     | http://www.aol.com/inet/ind/kidshome.html                           |
| 4              | 06.Jul.1995:20:54:49 | 51                       | 213                                    | http://www.mserrysbe.anies.com/mserrysbe.anies/pokacksch.htm        |
| 5              | 06.Jul.1995:20:58:20 | 50                       | 11765                                  | http://www.blae.planet.com/fun/packman/                             |
| 6              | 07.Jul.1995:08:14:28 | 51                       | 74                                     | http://find.web.aol.com/chain/ffind/malechamseffind/?query=eric.tmc |
| 7              | 07.Jul.1995:08:15:42 | 50                       | 162635                                 | http://www.lycos.com/network/                                       |
| 8              | 08.Jul.1995:21:26:17 | 4                        | 113                                    | http://www.zvooorder.com/sst/ogon.asp                               |
| 9              | 08.Jul.1995:21:28:10 | 19                       | 8                                      | http://www.zvooorder.com/sst/verificatlon.asp                       |
| 10             | 08.Jul.1995:21:28:29 | 4                        | 43                                     | http://find.web.aol.com/chain/ffind/malechamseffind/?query=stargate |
| 11             | 08.Jul.1995:21:28:12 | 1                        | 601                                    | http://members.aol.com/maest  |
| 12             | 08.Jul.1995:21:39:13 | 2                        | 65                                     | http://members.aol.com/maest  |
| 13             | 08.Jul.1995:21:40:10 | 36                       | 8                                      | http://index.simplenet.com/simplenet/links.htm                      |
| 14             | 08.Jul.1995:21:40:54 | 132                      | 172                                    | http://www.e-net.or.jp/uses/oc/1701/dcf                             |
| 15             | 08.Jul.1995:21:43:46 | 1                        | 29                                     | http://www.e-net.or.jp/uses/oc/1701/dcb06/99sg1.html                |
| 16             | 08.Jul.1995:21:44:06 | 1                        | 5                                      | http://www.e-net.or.jp/uses/oc/1701/dcf                             |
| 17             | 08.Jul.1995:21:44:11 | 1                        | 15                                     | http://www.e-net.or.jp/uses/oc/1701/dcb06/99sg1.html                |
| 18             | 08.Jul.1995:21:44:26 | 6                        | 6                                      | http://www.e-net.or.jp/uses/oc/1701/dcb06/99sg1.html                |
| 19             | 08.Jul.1995:21:44:32 | 1                        | 15                                     | http://www.e-net.or.jp/uses/oc/1701/dcf                             |
| 20             | 08.Jul.1995:21:44:47 | 1                        | 16                                     | http://www.e-net.or.jp/uses/oc/1701/awaso/smaie.html                |
| 21             | 08.Jul.1995:21:45:00 | 1                        | 75                                     | http://www.e-net.or.jp/uses/oc/1701/awaso/smaie.html                |

12

## How much have you learned about this person?

- Gender
  - Age
  - Race
  - Marital Status
  - Geographic Location
  - City Size
  - Household Size
  - Household Composition
  - Household Income
  - Rent or Own
  - Education
  - Age and presence of children
  - Car or truck ownership
  - Dog or cat ownership
- Female
  - 34 years old
  - White
  - Single
  - East South Central
  - 250,000-499,999
  - 2 household members
  - Female head living with others related
  - \$25,000-\$29,999
  - Own
  - Graduated High School
  - No Children Under 18
  - Two cars, no trucks
  - No dogs or cats

13

## Is this user male or female?

User visits the following five sites in the Doubleclick network



95% probability that user is female

14

## Bayesian updating formula

Test the hypothesis that a user is female by updating the current guess using new information

$$\bar{p} = \frac{p \cdot p}{p \cdot p + (1-p)(1-p)}$$

The formula is annotated with blue arrows:
 

- An arrow labeled "New probability" points to the updated probability  $\bar{p}$ .
- An arrow labeled "New information" points to the first  $p$  in the numerator.
- An arrow labeled "Old probability" points to the second  $p$  in the numerator.
- Brackets under the denominator label the terms: "Female" under  $p \cdot p$  and "Male" under  $(1-p)(1-p)$ .

15

## Probability user is female

|                         | <i>Probability a<br/>Female Visits<br/>the site</i> | <i>Probability<br/>visitor is<br/>Female given<br/>visits to</i> |
|-------------------------|---|--|
| <i>Overall Internet</i> | 45%   | 45.0%  |
| cbs.com                 | 54%   | 49.0%  |
| ivillage.com            | 66%   | 65.1%  |
| libertynet.org          | 63%   | 76.0%  |
| nick.com                | 57%   | 80.8%  |
| onlinepsych.com         | 83%   | 95.4%  |

16



## What is this user's gender?

### *Web sites visited during one month:*

|     |                    |     |                   |
|-----|--------------------|-----|-------------------|
| 48% | aol.com            | 63% | libertynet.org    |
| 64% | astronet.com       | 39% | lycos.com         |
| 75% | avon.com           | 27% | netradio.net      |
| 52% | blue-planet.com    | 57% | nick.com          |
| 56% | cartoonnetwork.com | 59% | onhealth.com      |
| 54% | cbs.com            | 83% | onlinepsych.com   |
| 76% | country-lane.com   | 44% | simplenet.com     |
| 47% | eplay.com          | 76% | thriveonline.com  |
| 41% | halcyon.com        | 59% | valupage.com      |
| 70% | homearts.com       | 71% | virtualgarden.com |
| 66% | ivillage.com       | 66% | womenswire.com    |

*Percentage of female viewers using PC Meter data*

## Results

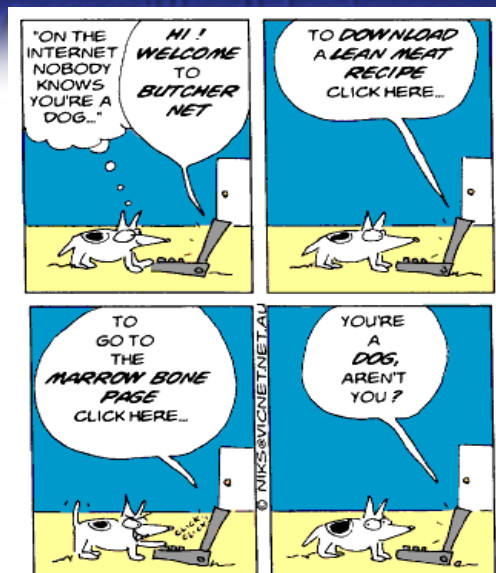
- Analysis shows that there is a >99% probability this user is female.
  - Using only DoubleClick sites the probability is 95%.
- Using all user data for one month:
  - 90% of men are predicted with >80% confidence (81% accuracy)
  - 25% of women are predicted with >80% confidence (96% accuracy)

## Key Points of User Profiling

*We can identify 'who you are' from 'where you go'*

- What the user views on the web reveals their interests and preferences
  - We can personalize the web experience without explicitly requiring customers to login and identify themselves
- Browsing and product choices can reveal key information about interest and price sensitivity
- Requires marketers to be smarter in designing their websites and analyzing their information. Big profitability gains if this is done correctly.

19



[Http://www.moreinfo.com/au.cranlerma/fo12.htm](http://www.moreinfo.com/au.cranlerma/fo12.htm)

## Using the clickstream to predict purchases

### Problem



- How do we predict (or influence) a purchase decision?
- How much information is there in the path that an individual takes about purchase? Price sensitivity?
- What is an appropriate form of the model?

## Motivation

- Assume that our independent variable is a latent measure of interest (e.g., utility) instead of time web used
- Again relate current interest to past values of interest and unpredictable changes
- Also incorporate measures of information and content about a web page
- Finally, assume the coefficients follow a switching model for 'surfing' behavior and 'goal' oriented behavior

23

## Model

Latent utility condition/ordered probit model:

$$y_{it} = \begin{cases} 1, & \text{if } u_{it}^* > 0 \\ 0, & \text{if } 0 \leq u_{it}^* \end{cases}$$

Purchase  
Continue to browse  
Leave web site

ARIMAX model for latent utility:

State condition (browse/search) - Markov Model/Mixture Model

$$(1 - \mathbf{f}_{i1}^s)u_{it}^* = \mathbf{b}_i^s x_{it} + (1 - \mathbf{q}_{i1}^s)\mathbf{e}_{it}, \quad \mathbf{e}_{it} \sim N(0,1)$$

24

## Model Elements

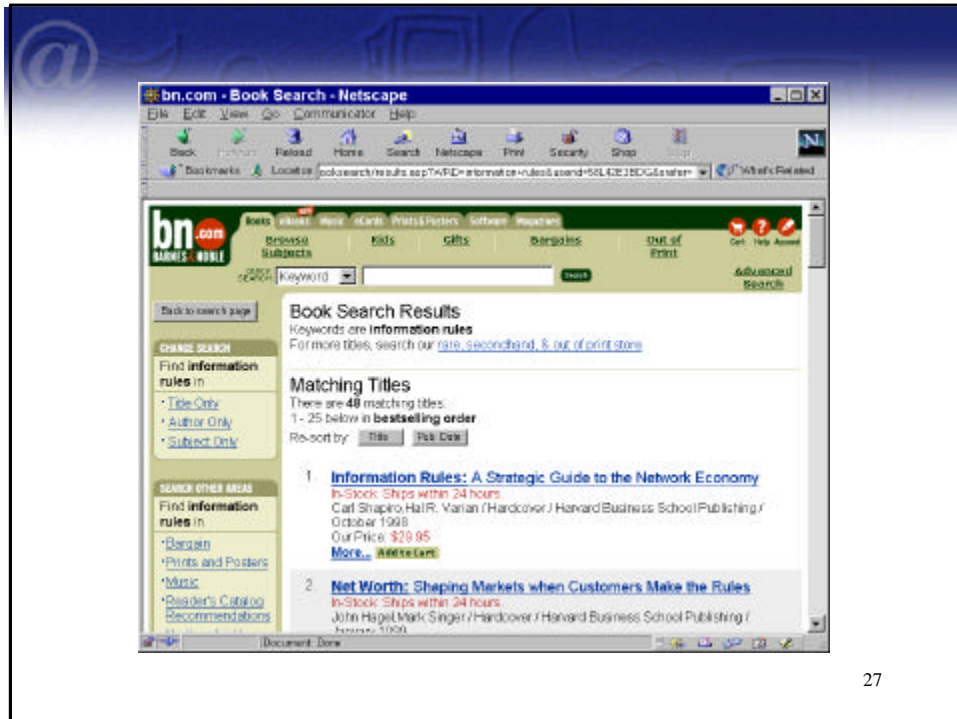
- Apply a multinomial probit model to the choice of links on the page (implicitly models purchase when 'checkout' is chosen)
- Incorporate
  - Attributes of the product/page
  - Interaction with items in the basket
  - Position on the screen
  - Time Effects
  - Effects of Last Viewing

25

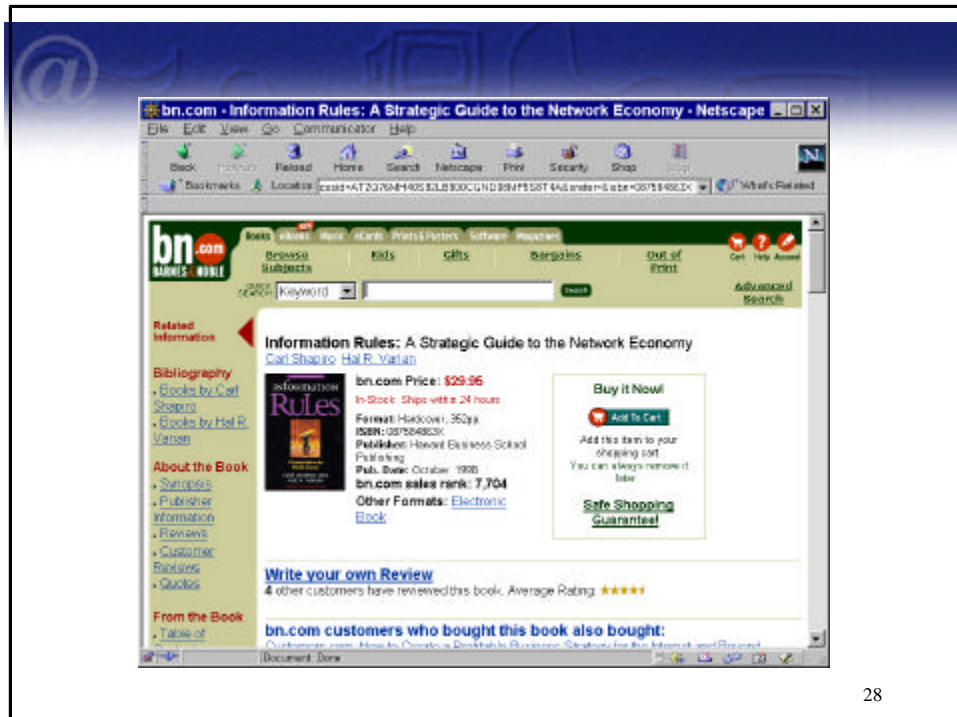


The screenshot shows a Netscape browser window displaying the BarnesandNoble.com website. The browser's address bar shows "barnesandnoble.com (www.bn.com) - Netscape". The website's navigation bar includes links for "Books", "Subjects", "Gifts", "Bargains", and "Out of Print". A search bar is visible with the text "Keyword" and a "Search" button. The main content area features several promotional banners and articles. On the left, there is a banner for "May 2, 2000" with a "HAVE \$10" offer. The central article is titled "Brown at Her Best" and discusses Sandra Brown's new book, "THE STANDOFF". Below this, there is a section for "In Music" featuring Billy Joel's "2K" album. On the right side, there are several smaller promotional boxes, including "Safe Shopping Guarantee", "Order Status", "On the Go" (wireless shopping), "Tom Robbins", and "New from Oprah". The browser's status bar at the bottom shows "Document Done".

26



27



28



29

## Model

Latent utility condition/ordered probit model:

$$y_{it} = \begin{cases} 2, & \text{if } u_{it}^* > \mathbf{g}_i & \text{Purchase} \\ 1, & \text{if } \mathbf{g}_i \geq u_{it}^* > 0 & \text{Continue to browse} \\ 0, & \text{if } 0 \leq u_{it}^* & \text{Leave web site} \end{cases}$$

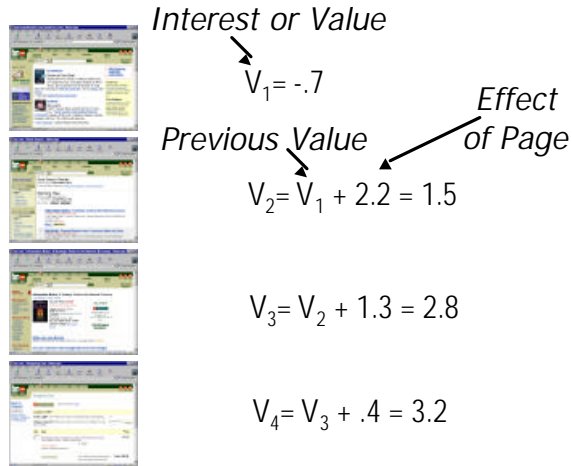
ARIMAX model for latent utility:

State condition (browse/search) - Markov Model/Mixture Model

$$(1 - \mathbf{f}_{i1}^s)u_{it}^* = \mathbf{b}_i^s x_{it} + (1 - \mathbf{q}_{i1}^s)\mathbf{e}_{it}, \quad \mathbf{e}_{it} \sim N(0,1)$$

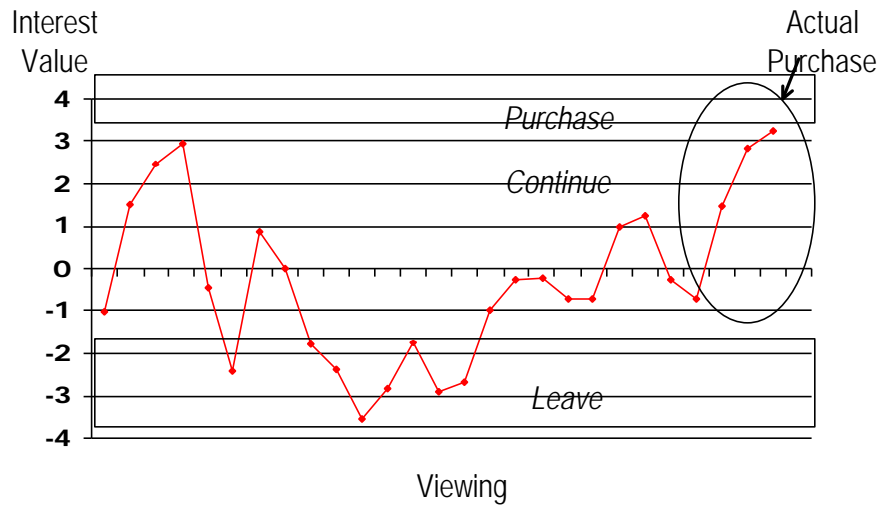
30

## Illustrating the Model




31

## How the model works



32





## Implementation

- Statistical models can describe modeling very well
- The biggest challenge is extracting textual information and navigation information from the web page

33



## Improving Shopbot Design

## Which would you choose?

**A** Amazon  
\$43.90  
5-10 days

**B** Kingsbooks.com  
\$32.06  
16 days

**C** 1Bookstreet.com  
\$35.96  
6-21 days

**D** Barnesandnoble.com  
\$37.19  
5-9 days

35

## Which store would you shop?

The left screenshot displays a price comparison tool interface. It features a search bar at the top and a table of results. The table has columns for 'Retailer', 'Price', 'Shipping', 'Total', 'Days', 'Status', and 'Link'. Below the table, there is a note: 'Click on one offer (1st column) to proceed to the respective shop!'. The right screenshot shows the Amazon.com homepage, featuring a navigation bar, a search bar, and several promotional banners, including one for 'Summer Reading Sale' and another for 'Hot Summer Deals'.

36

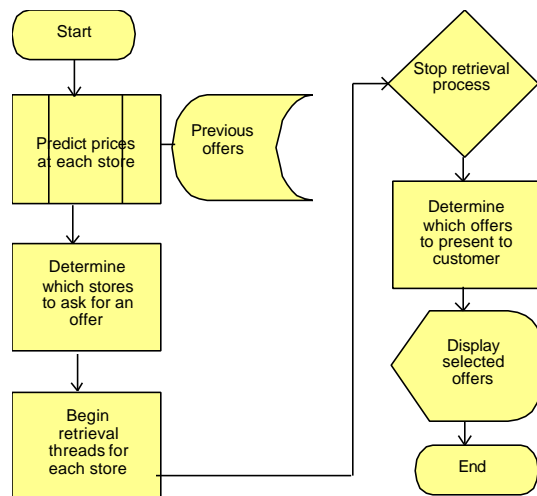
# Current ShopBot Design



- Enter a book and evenbetter.com will search 30+ bookstores and return all searches ordered by price
- Yahoo currently lists over 100 bookstores
- Makes search quick and simple
- Search is valuable
  - Average range is \$12
  - Amazon lowest only 5% of time (and dropping quickly)

37

# Shopbot Operational Flow



38

## Operational Decisions

- Which stores to search?  
Shopbots can form prior expectations about prices that can help eliminate searching at high price stores
- How long to wait?  
About 5% of store requests time out, also it may be better to interrupt searches at a certain point
- Which offers to present?  
It is very cognitively taxing for consumers to have to search through scores of offers. Consumer research tells us they will use less efficient comparison rules.

39

## Modeling Consumer Interaction with a Shopbot

- Use a compensatory utility model to determine consumer's tradeoff between price, delivery, tax, and waiting time
- Consider the cognitive costs that a consumer incurs in making comparisons
- Use past information from previous web retrievals to intelligently retrieve prices

Compare this approach with IR models that assume noncompensatory rules

40

## Utility Model

- Usual additive utility model for the  $i$ th product given  $P$  alternatives with  $A$  attributes in the set:

$$u_i = \overbrace{\sum_{j=1}^N b_{ij} a_{ij}}^{u_i^*} + e_i - \mathbf{x} \cdot T - \mathbf{w} \cdot Q - \mathbf{l} \cdot C$$

attributes: price, delivery time, etc.     
 waiting time =  $\max(t_1, \dots, t_M)$      
 system overhead to process requests     
 cognitive costs  $(A-1)(P-1)$

41

## Utility of the Choice Set

- Utility of basket with  $P$  choices from  $M$  alternatives:

$$\begin{aligned}
 U &= \max(u_{(1)}, \dots, u_{(P)}) \\
 &= \max(u_{(1)}^* - \mathbf{x} \cdot T - \mathbf{w} \cdot Q - \mathbf{l} \cdot C, \dots, u_{(P)}^* - \mathbf{x} \cdot T - \mathbf{w} \cdot Q - \mathbf{l} \cdot C) \\
 &= \max(u_{(1)}^*, \dots, u_{(P)}^*) - \mathbf{x} \cdot T - \mathbf{w} \cdot Q - \mathbf{l} \cdot (A-1)(P-1)
 \end{aligned}$$

present the best offers,  
ordered observations

42

## Formal Problem

- Sequential Optimization – solved backwards

$$\max_{q,p,t^*} E[\max(\mathbf{U}\langle p \rangle)] \quad s.t. \quad p \leq r \leq q$$

Variables:

- $q$  offers to query
- $r$  offers retrieved
- $p$  offers presented
- $t^*$  time to interrupt query

43

## Simulation Results

## Shopbot Choice Model

| <u>Parameter</u>        | <u>Estimate</u> |          |
|-------------------------|-----------------|----------|
| <i>Total Price</i>      |                 |          |
| Item Price              | -.19            | (\$1.00) |
| Shipping Price          | -.37            | (\$1.95) |
| U.S. Tax                | -.43            | (\$2.26) |
| <i>Delivery Average</i> | -.02            | (\$.10)  |
| <i>Delivery "n/a"</i>   | -.37            | (\$1.94) |
| <i>"Big 3"</i>          |                 |          |
| Amazon                  | .48             | (\$2.52) |
| BarnesandNoble          | .17             | (\$.89)  |
| Borders                 | .27             | (\$1.42) |

45

## Delivery Options for BarnesandNoble.com

| <u>Service</u>      | <u>Delivery</u> | <u>Cost</u> |
|---------------------|-----------------|-------------|
| U.S. Postal Service | 5-9 days        | \$3.95      |
| Standard Ground     | 4-7 days        | \$3.99      |
| FedEx Second Day    | 3-4 days        | \$7.95      |
| UPS 2nd Day Air     | 3-4 days        | \$7.99      |
| FedEx Overnight     | 2-3 days        | \$10.95     |

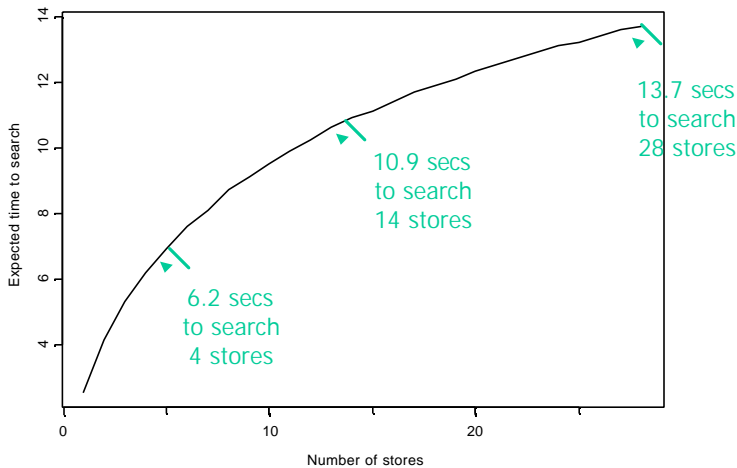
46

# Which stores to query?

| Store              | Mean | Std.Dev. | Store                | Mean | Std.Dev. |
|--------------------|------|----------|----------------------|------|----------|
| buy.com            | .52  | .10      | varsitybooks.com     | .75  | .05      |
| booksamillion.com  | .59  | .12      | 1bookstreet          | .76  | .13      |
| Bookbuyer's Outlet | .62  | .13      | bigwords.com         | .77  | .06      |
| Borders.com        | .62  | .13      | WordsWorth           | .83  | .10      |
| alldirect.com      | .63  | .05      | booksnow.com         | .88  | .06      |
| Amazon             | .63  | .13      | Cherryvalleybooks    | .89  | .02      |
| barnesandnoble.com | .63  | .13      | Rainy Day Books      | .89  | .05      |
| AlphaCraze.com     | .64  | .09      | Rutherfords          | .89  | .05      |
| Fatbrain           | .65  | .15      | Classbook.com        | .96  | .06      |
| Books.com          | .70  | .09      | Baker's Dozen online | .99  | .06      |
| HamiltonBook.com   | .70  | .07      | Book Nook Inc.       | .99  | .05      |
| BCY Book Loft      | .72  | .07      | Codys Books          | .99  | .06      |
| kingbooks.com      | .73  | .04      | computerlibrary.com  | .99  | .06      |
| A1 Books           | .75  | .06      | page1book.com        | .99  | .07      |

47

# How long to wait?



48



## Best set to offer

(if prices known)

| <u>Store</u>    | <u>Service</u>     | <u>Delivery</u> | <u>Price</u> | <u>Cost</u> | <u>Total</u> |
|-----------------|--------------------|-----------------|--------------|-------------|--------------|
| 1BookStreet.com | USPS Parcel Post   | 6-21 days       | \$15.19      | \$0         | \$15.19      |
| Amazon.com      | USPS Priority Mail | 5-10 days       | \$12.59      | \$3.95      | \$16.54      |
| Buy.com         | Standard Shipping  | n/a             | \$10.39      | \$3.95      | \$14.34      |
| Borders.com     | Standard           | 5-10 days       | \$12.39      | \$3.90      | \$16.29      |

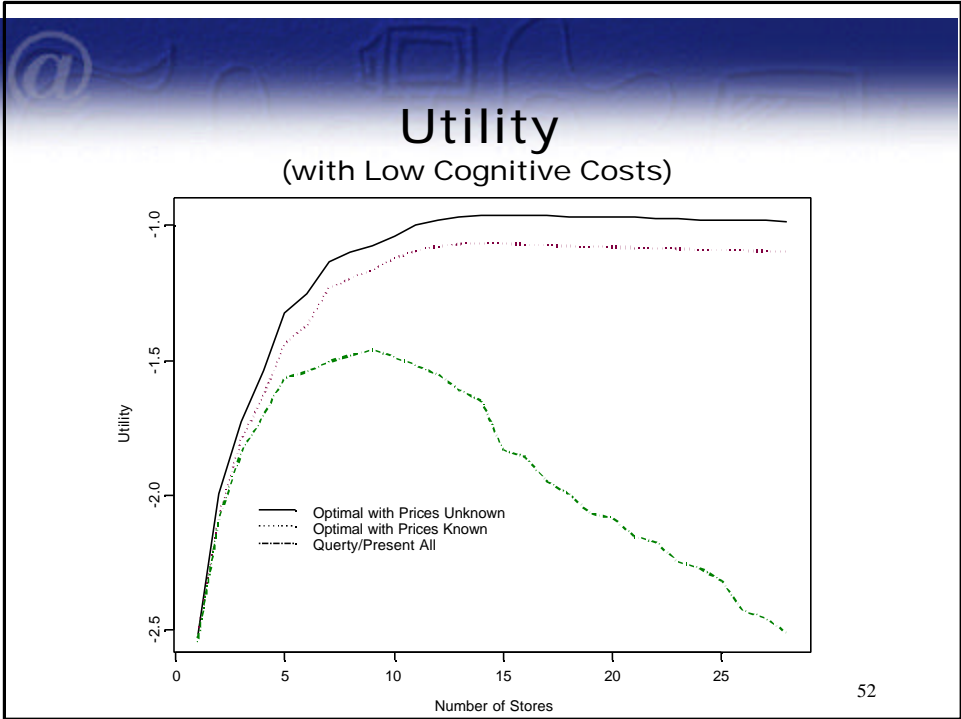
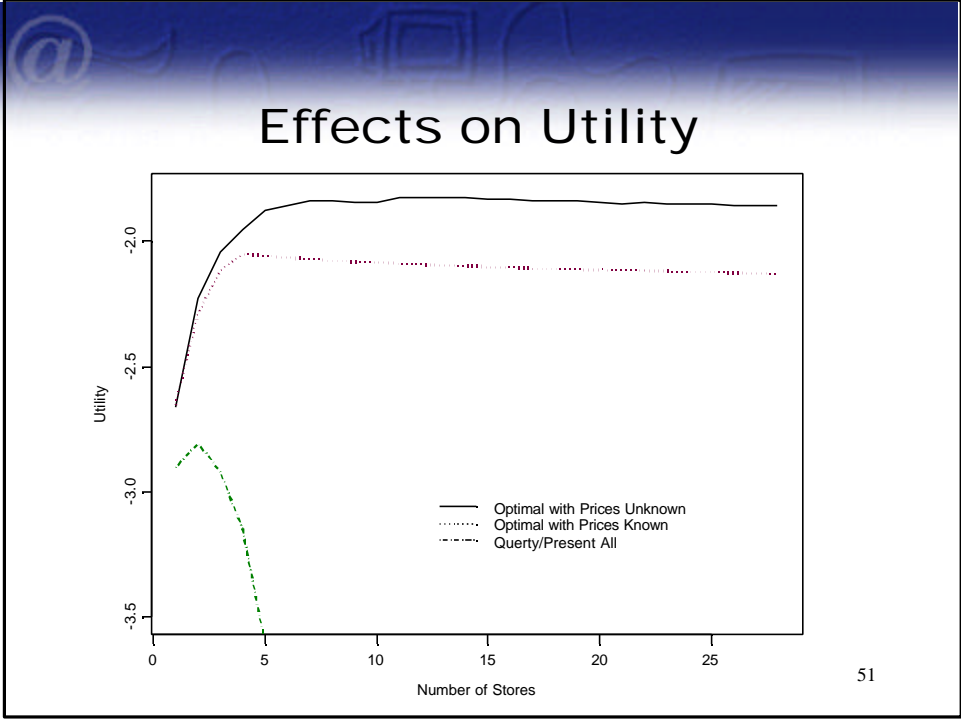
49

## Best set to offer

(if prices are forecasted)

| <u>Probability</u>              | <u>Store</u>           | <u>Service</u>      | <u>Delivery</u> | <u>FIPrice</u> | <u>Cost</u> | <u>FITotal</u> |
|---------------------------------|------------------------|---------------------|-----------------|----------------|-------------|----------------|
| Still<br>include<br>these       | 82% 1BookStreet.com    | USPS Parcel Post    | 6-21 days       | \$ 15.19       | \$ -        | \$ 15.19       |
|                                 | 56% Amazon.com         | USPS Priority Mail  | 5-10 days       | \$ 12.59       | \$ 3.95     | \$ 16.54       |
|                                 | 51% Buy.com            | Standard Shipping   | n/a             | \$ 10.39       | \$ 3.95     | \$ 14.34       |
|                                 | 42% Borders.com        | Standard            | 5-10 days       | \$ 12.39       | \$ 3.90     | \$ 16.29       |
| Chance of<br>including<br>these | 34% bamesandnoble.com  | Standard Ground     | 4-7 days        | \$ 12.59       | \$ 3.99     | \$ 16.58       |
|                                 | 28% bamesandnoble.com  | U.S. Postal Service | 5-9 days        | \$ 12.59       | \$ 3.95     | \$ 16.54       |
|                                 | 26% booksamillion.com  | Standard Ground     | N/A             | \$ 11.79       | \$ 3.95     | \$ 15.74       |
|                                 | 21% Fatbrain.com       | UPS Ground          | 4-8 days        | \$ 12.99       | \$ 3.95     | \$ 16.94       |
|                                 | 17% AlphaCraze.com     | USPS Special Rate   | 5-15 days       | \$ 12.79       | \$ 3.50     | \$ 16.29       |
|                                 | 13% Bookbuyer's Outlet | Standarc            | n/a             | \$ 12.39       | \$ 4.50     | \$ 16.89       |

50



## Findings

- Present small number of the best alternatives (~4 items)
- Only need to search small number of stores if prices known with certainty
- If prices are unknown, it's best to continue searching at up to 14 stores. Since not all results presented (and little computational cost) go ahead and search more stores.
- Using traditional shopbot best to only search a couple of stores
- Probability consumer will prefer optimal shopbot over current shopbot >99.9% or \$2.00, or worth \$.20 compared with best result when all results presented from 2 stores

*What happens if customers less sensitive to cognitive costs?*

53

## Summary

- Intelligent design of shopbots can dramatically increase the utility that consumers garner from their use
- Instead of passively searching, can incorporate information about utility and price expectations to speed up search and satisfaction
- Incorporates cognitive effort, compensatory utility functions, and information retrieval

54

## Future Research Directions

## An improved shopbot?



- Ask users for filtering questions about preferences or use information from previous history
- Appropriately balance the cost of asking for the information with its benefits
- Allow further search
- Better understand how consumers perceive waiting time based on expectations, provide 'filler' tasks

56

## Future Shopbot Research

- Learning from past purchases and designing active shopbots (versus the passive design presented)
- Need better understanding of how to quantify cognitive costs and effects of waiting
- Train shopbots to be proactive in seeking out good deals. If bestseller status changed today & shopbot knows a store responds to status change in 2 days, it can make recommendations (“wait 2 days and price at amazon likely to be less by \$10”)
- Identify baskets of products or more complex products like travel (airline tickets, car rental, hotel, etc.)
- Applications to information goods (e.g., news stories, recommender systems, search engines)

57

## Reflections on E-Commerce Research

- Most of the existing marketing literature focuses on describing how consumers behave
- What is really needed is prescribing how agents should behave so they can work with and/or replace consumers

Can yield some new insights into old problems (e.g., how do consumers shop?)

58