2007 UNIVERSITY OF PENNSYLVANIA INVITATIONAL CHOICE CONFERENCE

TRACK ON HIGH-DIMENSIONAL DATA

Prasad Naik and Michel Wedel

JUNE 13-17, 2007

The Wharton School
The University of Pennsylvania
Philadelphia, PA

THE INN AT PENN
3600 SANSOM STREET

JON M. HUNTSMAN HALL
3730 WALNUT STREET

Table of Contents

Participants' Information for the Session Track on High-Dimensional Data

Nos	*Name and Email*	*Affiliation*	*Domain Expertise*
1	Dr. Lynd Bacon lbacon@lba.com	LBA Associates, Woodside CA	Database design and management; Data & text mining; Direct response modelling; Database scoring; Bayesian modeling and simulation, AI. For more information see: http://www.lba.com/
2	Prof. Anand Bodapati Assistant Professor of Marketing anand.bodapati @anderson.ucla.edu	University of California, Los Angeles	Customer Relationship Management and Direct Marketing; Consumer Response to Advertising; Marketing Research. For more information, see http://www.anderson.ucla.edu/ x2021.xml
3	Prof. Wagner Kamakura Ford Motor Company Professor of Global Marketing kamakura@duke.edu	Duke University	Market Segmentation and Structure, Database Marketing, CRM, Consumer Choice Behavior. For more information, see http://www.fuqua.duke.edu/ faculty/alpha/kamakura.htm
4	Dr. Jeffrey Kreulen Manager, operations technologies kreulen@almaden.ibm.com	IBM Research	Web-based service delivery; Integrated text and data analysis; eClassifier and BIKM technology for customer scenarios. For more information, see http://domino.watson.ibm.com/com m/pr.nsf/pages/bio.kreulen.html
5	Prof. Peter Lenk Professor of Statistics plenk@umich.edu	University of Michigan	Bayesian Data Analysis, Nonparametric Analysis, Multivariate Data Analysis, Choice Modeling For more information, see http://www.bus.umich.edu/ FacultyBios/ FacultyBio.asp?id=000119800

6	Prof. David Madigan Professor of Statistics Dean, Physical and Mathematical Sciences dmadigan@rutgers.edu	Rutgers University	Bayesian Data Analysis, Data Mining, Text Categorization For more information, see http://www.stat.rutgers.edu/%7Emadigan/
7	Prof. Alan Montgomery Associate Professor of Marketing alm3@andrew.cmu.edu	Carnegie Mellon University	Intelligent computer agents; Browsing behavior; Click-stream Data; Data mining; Time Series Analysis; Bayesian Statistics For more information, see http://business.tepper.cmu.edu/display_faculty.aspx?id=100
8	Prof. Prasad Naik Chancellor's Fellow and Professor of Marketing (Co-chair) panaik@ucdavis.edu	University of California Davis	High-Dimensional Data Analysis, Inverse Regression, Mixture Regression, Variable Selection, Kalman Filter, Hidden Markov Models, Integrated Marketing Communications For more information, see http://www.gsm.ucdavis.edu/Faculty/index.aspx?id=582
9	Prof. Michel Wedel Pepsico Professor of Consumer Science (Co-chair) mwedel@rhsmith.umd.edu	University of Maryland	Bayesian Data Analysis, Hidden Markov Models, Mixture Models, Recommendation systems. http://www.rhsmith.umd.edu/marketing/faculty/wedel.html

Schedule of Sessions for High-Dimensional Data Track

Wednesday, June 13--

6:00-9:00 p.m. **Welcome Reception and Conference Registration**

The Inn at Penn Hotel--Regent/St. Mark's Ballroom

Thursday, June 14--

7:00-8:30 a.m. **Continental Breakfast**
The Inn at Penn

8:30-10:00 a.m. **Introduction to Sessions**

10:00-10:30 a.m. **Coffee Break**
The Inn at Penn AND 8th Floor, Huntsman Hall

10:30-12:00 noon **Alan Montgomery,**
Computational Challenges for Real-Time Marketing with Large Datasets

12:00 -1:30 p.m. **Lunch**
8TH Floor, Huntsman Hall

1:30-3:00 p.m. **Lynd Bacon**
Understanding Choices and Preferences with Massive, Complex On-line Data

3:00-3:30 p.m. **Coffee Break**
The Inn at Penn AND 8th Floor, Huntsman Hall

3:30-5:00 p.m. **Wagner A. Kamakura**
Some rambling comments on "High Dimensional Data Analysis"

6:00-9:00 p.m. **A TASTE OF PHILLY**
Drinks and Casual Dinner
The Annenberg Center for Performing Arts
Main Lobby
37th and Walnut Streets, Philadelphia

9:00-11:00 p.m. **Nightcap at the Inn at Penn**

Friday, June 15, 2007—

7:00-8:30 a.m. **Continental Breakfast**
The Inn at Penn

8:30-10:00 a.m. **Jeffrey T. Kreulen**
Leveraging Structured and Unstructured Information Analytics to Create Business Value

10:00-10:30 a.m. **Coffee Break**
The Inn at Penn AND 8th Floor, Huntsman Hall

10:30-12:00 noon **David Madigan**
Statistical Modeling: Bigger and Bigger

12:00 -1:30 p.m. **Lunch**
8th Floor, Huntsman Hall

2:30-5:00 p.m. **HISTORIC TROLLEY TOUR OF PHILADELPHIA**
Depart from The Inn at Penn. Tour will Terminate at the National Constitution Center/Independence Hall Area

5:00-9:00 p.m. **EVENING AT THE NATIONAL CONSTITUTION CENTER**
Cocktails, Dinner, and Show
Mingle with our founding fathers as the story of the US Constitution comes to life!
Trolleys will transport you back to The Inn at Penn at the close of the evening

Saturday, June 16, 2007—

7:00-8:30 a.m. **Continental Breakfast**
The Inn at Penn

8:30-10:00 a.m. **Anand Bodapati**
Issues in the Modeling of Behavior in Online Social Networks

10:00-10:30 a.m. **Coffee Break**
The Inn at Penn AND 8TH Fl, Huntsman Hall

10:30-12:00 noon **Michel Wedel**
State of the Art Recommendation Systems: the Good, the Bad and the Ugly

12:00 -1:30 p.m. **Lunch**
MBA Lounge and Koo Plaza
2nd Floor Huntsman Hall

1:30-3:00 p.m. **Peter Lenk**
Approximate Bayes Methods for Massive Data in Conditionally Conjugate Hierarchical Bayes Models

3:00-3:30 p.m. **Coffee Break**
The Inn at Penn AND 8th Floor, Huntsman Hall

3:30-5:00 p.m. **Prasad Naik**
Inverse Regression Methods: A Review, Applications and Prospects

6:00-9:00 p.m. **Dinner**
The Inn at Penn Hotel--Woodlands Ballroom

Sunday, June 17, 2007--

7:00-8:30 a.m.	**Continental Breakfast** **The Inn at Penn**
8:30-10:00 a.m.	**Plenary Session 1** **G-06 Auditorium, Huntsman Hall**
10:00-10:30 a.m.	**Coffee Break** **Forum Level Lounge, Huntsman Hall**
10:30-12:00 noon	**Plenary Session 2** **G-06 Auditorium, Huntsman Hall**
12:00 NOON	**ADJOURN**

EXTENDED ABSTRACTS
ON HIGH-DIMENSIONAL DATA TRACK

Understanding Choices and Preferences with Massive, Complex On line Data

Lynd Bacon

LBA Associates

The accelerating rate of technological innovation is producing more and more data that might be (and should be!) used to better expain and predict choices and preferences("EPCP"). With this increasing volume of potentially useful data comes increasing complexity, which presents additional opportunities and also some new challenges.

In what might be called the domain of "traditional" data mining, "large" has usually meant having so much data that computational cost is metered in disk activity, rather than in CPU cycles. The definition is in terms of hardware. It's also the case that the data themselves have usually be "frozen" in data marts built from production loads so that they are not changing. At least that was the preference of data miners, for obvious reasons. But for the most part, size mattered.

"Size" is relative, of course. If size is considered from the perspective of processing effort or time required, one thousand customer verbatims would be considered a "large" data set if they were to be content coded by one person. So would several hundred multiple-page answers to an essay exam that needed to be "labeled" with letter grades.

Complexity matters, too. As innovation has been driving data availability, it has also been driving data *complexity*. This complexity is multidimensional, and it may be more significant in terms of explaining and predicting choices and preferences than sheer amount of data. I posit that the relevant dimensions of this complexity include the following:

- degree and nature of structure
- availability
- velocity
- perishability

Nowhere are data complexity and volume growing more rapidly than on the Internet and the World Wide Web. The evolution of Web 2.0 applications, easier-to-use content creation tools, and broadband penetration are significant factors underlying this growth. The "democratization" of the Web is definitely playing a role, and it is beginning to facilitate the kind of co-creation of value and collaboration indicated by some as being of strategic importance to many businesses(see for example Prahalad & Ramaswamy,2004; Miles et al. 2005). My subsequent comments pertain to the burgeoning on line environment.

The Web: an exploding domain of complex, twinking data that bears the footprints of preference and choice

Since the initial development and implementation of software for web servers in 1990, and the release of this software into the public domain 1993, the world wide web (WWW) has been growing exponentially in terms of content and utilization. The content of the web is complex, but not unstructured, of course. And, more recent considerations for web standards and design are focussed on making the web, the "Semantic Web," an environment with smarter data that supports reasoning (World Wide Web Consortium, 2005).

At the same time that efforts are being pursued to make the Web a more valuable knowledge network, new applications have been coming available that enable non-technical users to create and publish various kinds of content. These applications include content management systems and blogging software. Several such applications are available on a "hosted" basis, further lowering the cost of using them. As a result of the effects of these technologies and other factors, web sites are becoming more participatory: more open, and empowered by social networking effects. These are so-called "Web 2.0" sites and applications (O'Reilly, 2005). The participatory nature of these applications moves them from a publishing (i.e. pushing of content) paradigm to a participatory, aggregative sort of functionality. And as a result, a large number of consumers (> 1B) are on line, and are increasingly posting product evaluations and comments (Trendwatching.com, 2007). Companies like Umbria and Wize are attempting to extract insights from the growing volume of diverse and heterogeneous consumer-generated data. The challenges include locating the data, and getting it into analyzable form, of course.

On line search, retrieval, and recommendation are preference prediction problems, and a difficult ones, at that. From the perspective of a search engine company, the problem is to produce a list of preferred items sorted in decreasing order of preference. For an on line retailer, the problem is to produce the best approximation to the item sought by a customer. Recommendation is another example. Current practice relies mainly on notions of "similarity" for retrieval and recommendation. Similarity may be defined in terms of object characteristics, e.g. features of digital cameras, or in terms of associations between behaviors directed at objects, such as in *collaborative filtering*.

Similarity-based approaches to search and recommendation of the type that are currently deployed are never likely to lead to highly accurate results for several reasons. These include sparse data at the individual level, lack of contextual data, and in the case of retrieval based on typed queries, the usual problems of polysemy and semantic ambiguity. This is in addition to the problems caused by variations in data structure, coding and so on.

Spam and "anti-preference." Spam filtering can be thought of as a problem in predicting disutility, or "anti-preference." One feature that makes spam filtering a difficult machine learning problem is that the "labels" (indicators of spam-ness) are not well defined, or are essentially missing. One person's spam may be another's sonnet. Also, (anti)preference for spam may vary over time and context for any individual. Companies like Cloudmark and Mcafee are aggregating data from their users to develop collaborative filtering procedures by leveraging the collective judgments of email recipients.

Structuring collective on line activity in the interest of EPCP

Some of the difficulties in learning about choices and preferences from consumer-generated content on the Web is consumers don't always provide the data that would be most useful for EPCP, and what they do provide isn't usually in a form that's convenient to use, either because they have no reason to, or because the sites they are using aren't designed to do so. When will the *Semantic Web for EPCP* be here? Anyway, one way of improving on this situation is to get consumers to participate in activities that structure what they do so as to produce more usable data and metadata, and that reward them appropriately for their efforts. Such rewards may be monetary or psychosocial. An example of the latter type can be found in the "usefulness" ratings for product reviews on Amazon.com.

CMU has offered a simple example of organizing collective efforts in their ESP Game, an effort to label images available on the Web. This application is a game in the sense that there are rules and desired outcomes from activity. Predictive markets, sometimes also called virtual stock markets (e.g. the Iowa Electronic Markets, and HSX)

Other game-like methods have been used for ideation and product development. Prelec (2001) describes a guessing game played over a network that can be used to develop customer descriptions of products and product concepts. Toubia (2006) describes a system for generating ideas in which incentives are aligned with performance. Bacon & Sridhar(2006) summarized different, on line "innovation tools" that can be used for identifying preferences and solving hard problems, including a game-oriented discussion system based on Kunz and Rittel's (1970)concept of an issues-based planning system(see also Conklin & Begman, 1988), that aligns rewards with incentives. An example implementation of this system is available for review by "massive data" session participants. Information about accessing it is provided at:

https://hound2.lba.com/community/?q=opengames_example

These and other kinds of purposed, structured activities have the potential to produce rich data about preferences that can be modeled in scalable and effective ways, given that participant contributions can be appropriately constrained, and that metadata are defined adequately. They can be used to leverage the complementary competencies of machines and humans working together over networks to EPCP.

References

*Bacon, L. & Sridhar, A. "Interactive innovation tools and Methods." Annual Convention of the Marketing Research Association, Washington D.C., June 2006

*Conklin, J. & Begman, M. (1988), "gIBIS: A hypertext tool for exploratory policy discussion." ACM Transactions on Office Information Systems, 6(4), 303-331.

Kunz, W. & Rittel, H. (1970), "Issues as elements of information systems." Berkeley: Institute of Urban and Regional Development, Working Paper 131.

Miles, R., Miles, G. & Snow, C. Collaborative Entrepreneurship. Stanford CA: Stanford Business School Press, 2005.

O'Reilly, T. (2005), "What is Web 2.0: Design patterns and business models for the next generation of software. http://www.oreillynet.com/lpt/a/6228

Prahalad, C. & Ramaswamy, V. The Future of Competition: Co-Creating Unique Value with Customers. Boston: Harvard Business School Press, 2004.

*Prelec, D. (2001), "Readings Packet on the Information Pump," MIT Sloan School of Management.

*Toubia, O. (2006), "Idea generation, creativity, and incentives." Marketing Science, 25(5), 411-425

Trendwatching.com (2007),"Transparency," May 2007 Trendwatching briefing,http://trendwatching.com/briefing/

World Wide Web Consortium (2005), "The Semantic Web." www.w3.org/2005/Talks/1111-Delhi-IH/

Note: materials marked with a "" are available to Symposium participants at:*

http://flyingdog.LBA.com/choice_symposium

Some rambling comments on "High Dimensional Data Analysis"

Wagner A. Kamakura
Duke University

To start, I must say I had great difficulty writing this extended abstract because of my limited experience with large-scale or truly "high dimensional" models." Like many of my marketing colleagues, I prefer to play with "toy" problems with a limited number of variables on a small sample size, and to pretend that my models are scalable to real sizes, or that Intel will ultimately produce the super-desktop (with an optical/biological processor?) needed to make the model feasible to real-sized problems.

The scalability of marketing models was not a major issue in the past because most firms traded with anonymous consumers, and therefore most of their marketing-research data was obtained from small not-so-random samples, through surveys that produced a limited number of variables. As more and more firms become customer-focused, they are amassing massive customer databases that integrate all aspects of their customer relationships. A typical bank in the US, Europe, Asia or Latin America collects information on millions of customers on no less than hundreds of variables including deposits, withdraws, balances, contacts, costs, fees, etc. One of the largest banks in the US has over 100 million accounts and maintains a standard "minimal" database with 150 variables on each of them.

This shift in focus from anonymous consumers in traditional mass marketing (e.g., packaged goods) towards individual identifiable customers (e.g., services) calls for a new focus in our modeling efforts. In traditional mass marketing it was acceptable to pretend that the data being analyzed came from a representative random sample of the population, and to draw inferences from that small sample to the population at large. Customer-focused enterprises on the other hand, want more specific customer insights that lead to policies that can be implemented at the customer or identifiable-segment levels, requiring that the model be scalable up to the entire customer base.

I see the "high dimensional analysis" problem in two dimensions (pardon the pun): 1) number of variables and 2) number of cases.

The "number of cases" problem

A common solution to the "number of cases" problem is to draw a random sample from the population, calibrate the model on the sample, and then apply the calibrated model to the entire population. It is rather ironic that while traditional mass marketers do not hesitate in making decisions on the basis of small non-random and hardly-representative samples, customer-focused marketers tend to resist doing so, because the data are at their disposal for the entire population of customers. The irony lies in the fact that because data are available for the entire population of customers, one can draw perfect textbook-case random samples from customer databases, something that rarely happens in traditional marketing research. There are, however, some situations where simple random sampling might not be most appropriate, for example when the purpose is to predict rare behaviors such as churn in well-managed customer relationships. In such cases where the event being studied is rare, *endogenous stratified sampling* (King and

Zeng 2001) or *case-control* design (Breslow 1996) can be used. Otherwise, the usual strategy is *exogenous stratified sampling*, which ensures that the population is well represented in the sample in terms of the predictors used in the modeling effort. One direction for future research would be in devising sampling strategies that ensure sample representativeness across all variables contained in the customer database. In other words, instead of stratifying the sample solely on the dependent or predictor variables, create a feasible stratification scheme covering all variables in the database.

The "number of variables" problem

This problem is more easily solvable through modeling, and has been the focus of some of my own work. The problem here is to reduce a large multivariate problem into a more manageable one. Suppose a retailer has sales (Y) and pricing (X) data on each of k=1,…,30 brands for each of its i=1,…,300 stores for t=1,…,36 months, and wants to understand the (cross) effects of price on brand sales, while allowing the price elasticities to vary across stores and over time. An aggregate model would require the estimation of 900 price coefficients, which is feasible given the available data. A random-coefficients model, only accounting for unobserved heterogeneity would also require the estimation of 901x450 = 405,450 covariance terms! If the modeler also wants to account for shifts in the parameters over time, the number of parameters would increase even further. In order to transform this random-coefficients model into a feasible one while allowing for the possibility that the (cross)-elasticities might be correlated across stores and over time, Kamakura and Kahn (2007) propose a factor-regression model where the covariance of the random regression coefficients is decomposed into its principal components across stores and over time. With a two-factor solution, the model would require 1,800 estimates for the covariance of the price coefficients, still large but more manageable than a full-covariance model, given the data available.

A more traditional approach for reducing a high dimensional problem to a more manageable one is by factor-analyzing the dependent variables (Wedel and Kamakura 2001). Du and Kamakura (2007), for example, propose a stochastic frontier factor model that takes advantage of the covariance structure among the dependent variables to define efficiency frontiers for each of multiple product categories based on the observed sales in all categories, as well as exogenous variables. This factor-analytic approach makes it possible to consider, in a parsimonious way, the information that all other categories might contain in defining the relative market potential for one focal category.

Even if the "number of variables problem" is solved with some form of data-reduction model, it is still critical that the calibrated model can be implemented in the entire database. Because some customer databases contain billions of records, simulation-based methods requiring many iterations per case might not be as scalable as simpler data-mining models based on singular-value decomposition. I hope someone in our workshop will focus on these data-mining models, which are grossly overlooked in the marketing literature, probably because we like to play with "toy" problems…

References

Breslow, Norman E. 1996. “Statistics in Epidemiology: The Case-Control Study.” *Journal of the American Statistical Association* 91:14–28

Du, Rex and Wagner A. Kamakura (2007) “How efficient is your category management? A stochastic-frontier factor model for internal benchmarking”, *Working Paper*

Kamakura, Wagner A. and Wooseong Kang (2007) Chain-wide and store-level analysis for cross-category management, *Journal of Retailing*, 83(2) : 159-70.

King, Gary and Langche Zeng (2001) “Logistic regression in rare events data,” *Political Analysis*, (February) 137-163.

Wedel, Michel and Wagner A. Kamakura (2001), “Factor Analysis with Mixed Observed and Latent Variables in the Exponential Family,” *Psychometrika*, 66 (4), pp. 515-30.

Leveraging Structured and Unstructured Information Analytics to Create Business Value

Jeffrey T. Kreulen, Ph.D.
Senior Manager, Senior Technical Staff Member
IBM Almaden Research Center

Enterprises understand that timely accurate knowledge can mean improved business performance. Two technologies are central in improving the quantitative and qualitative value of the knowledge available to decision makers. They are structured and unstructured information. We believe that the analysis of structured and unstructured data will converge over time, which has the potential to redefine the industry's traditional approach to BI, search and information management and analysis. Until now, text analysis has been significantly underutilized. Incorporating it into an information warehouse can enhance decision making processes and provide customers with a deeper insight into historical and extrapolated business conditions.

Business Intelligence has leveraged the functionality, scalability and reliability of modern database management systems to build constantly larger data warehouses and to utilize data mining techniques to extract business advantage from the vast available enterprise data. BI has traditionally focused on analysis of transactions and their associated attributes. Unstructured information technologies, while less mature than BI, are now capable of combining today's content management systems and the Web with vastly improved searching and text mining capabilities to derive more value from the explosion of textual information. Search has been widely espoused as the solution for unstructured information. We believe search is one tool in a workbench of capabilities that will be required to create business value from unstructured information. Unstructured information analytics will also require taxonomy generation and editing, classification, information extraction, statistical analysis, visualization and data structures to support these capabilities at the scale and dimensionality of unstructured information. We believe that structured and unstructured analytical systems will blend over time, leveraging techniques from each other and inspiring new approaches that can analyze data and text together seamlessly.

Many domains can benefit from the integrated analysis of both text and data. We have explored the application of these technologies in such domains as customer relationship management (CRM), health care and life sciences, intellectual property analysis, risk and compliance solutions, enterprise content analysis, and world wide web analytics.

References

"Mining Patents Using Molecular Similarity Search," with James Rhodes, Stephen Boyer, Ying Chen, and Patricia Ordonez, Pacific Symposium on Biocomputing 2007, pp. 304-315.

"*Interactive Methods for Taxonomy Editing and Validation,"* with W. Spangler, Next Generation of Data-Mining Applications, ISBN 0-471-65605-4, Chapter 20, pp. 495-522.

"*Generating and Browsing Multiple Taxonomies over a Document Collection*,", with W. Spangler and J. Lessler, Journal of Management Information Systems, Vol. 19, No. 4, Spring 2003. pp. 191-212.
"The Integration of Business Intelligence and Knowledge Management,", with W. Cody, W. Spangler and V. Krishna, IBM Systems Journal, Vol. 41, No. 4, 2002.

Approximate Bayes Methods for Massive Data in Conditionally Conjugate Hierarchical Bayes Models

Peter Lenk
University of Michigan

Massive datasets present challenges for Bayesian inference, which require numerically intensive computational procedures. Markov Chain Monte Carlo (MCMC) algorithms recursively generate the model's parameters from the full conditional distributions given the data. Straight forward implementations of MCMC require holding all of the data in memory or repeatedly reading from the data file at each iteration of the MCMC. Neither of these approaches is feasible with massive datasets.

DuMouchel (1999) used empirical Bayes (EB) for the analysis of very large frequency tables. EB can be viewed as a large sample approximation to hierarchical Bayes (HB) models where the parameters for the population-level distributions are estimated by non-Bayes methods, such as maximum likelihood, often after marginalizing over individual-level parameters. Ridgeway and Madigan (2002) proposed a fully Bayesian method that first performs traditional MCMC on a subset of the data, then applies importance sampling and resampling to the remainder of the data. Balakrishnan and Madigan (2006) modified this method by applying particle filters (Gordon, Salmond, and Smith 1993 and Liu and Chen 1998) to obtain a one-pass method.

Our proposed method relies on a property of Bayesian methods for conditionally independent observations. Suppose that the data set $Y = (y_1, \ldots, y_N)$ of N observations is too large for traditional MCMC methods. Divide the data into m manageable blocks $Y = (Y_1, .., Y_m)$. The algorithm recursively analyzes each block Y_j and uses the posterior distribution from the previous analysis as the "updated prior" for the next analysis. To fix ideas, $[y_i \mid \Theta]$ is the distribution of the i^{th} observation given the parameter Θ, and $[\Theta]$ is the prior distribution. Assume that the observations are independent given Θ. After observing the first block of n_1 observations $Y_1 = (y_1, \ldots, y_{n1})$, the posterior distribution $[\Theta|Y_1]$ expresses all of the information in Y_1 about Θ. Once one has this posterior distribution, Y_1 is no longer needed and can be erased from memory. Before observing the next block of n_2 observations $Y_2 = (y_{n1+1}, \ldots, y_{n1+n2})$, $[\Theta|Y_1]$ becomes the "updated prior" distribution for Θ. After observing Y_2, the updated posterior distribution is $[\Theta|Y_1,Y_2]$, which is proportional to $[Y_2|\Theta][\Theta|Y_1]$. Once again, after obtaining $[\Theta|Y_1,Y_2]$, the block of observations Y_2 is no longer needed and can be erased from memory. The recursion continues until of the data has been processed. The sequential updating results in the same posterior distribution as if the data were analyzed together.

If the model is conjugate, then the posterior distributions belong to the same family D of distribution as the prior distribution but with different parameters. In this case, the prior to posterior analysis for Y_1 is the same as Y_j given $Y_1, \ldots, Y_{j-1}$. When the model is not conjugate, then the posterior distributions belong to a different family of distributions, often of unknown form. In this case, it becomes numerically intensive to specify the "updated prior" $[\Theta| Y_1, \ldots, Y_{j-1}]$ for processing the next block Y_j.

The goal of the paper is to explore the feasibility of sequential processing of the observations in hierarchical Bayes (HB) models were the "population-level" model is conditionally conjugate. The HB model we will consider has two levels. The subject

level model, $[y_i|\beta_i, \sigma]$ is the distribution if the i^{th} observation and depends on subject-level parameters β_i, and parameters σ that are common to all observations. The population level model $[\beta_i|\Theta]$ expresses the heterogeneity in the subject level parameters across the population. The observations are conditionally independent and the joint distribution is given by:

$$\prod_{i=1}^{N} [Y_i \mid \beta_i, \sigma][\beta_i \mid \Theta][\Theta, \sigma].$$

We focus on models where the full conditional of Θ and σ given $\beta_1, \ldots, \beta_N$ belongs to the same family of distributions D as the prior for Θ and σ. The distribution of the subject-level parameters need not be conjugate. In the sequential updating process, $[\Theta,\sigma|Y_1, \ldots, Y_j]$ will be a mixture of distribution from D.

The research will investigate approximations to the $[\Theta,\sigma|Y_1, \ldots, Y_j]$ that facilitate the sequential updating. A mixture of distribution from the family D can often be well approximated by a single member of D. The approximation can be constructed to preserve the location and scale of the mixture. With massive data sets, the amount of information on Θ and σ is very large, and the posteriors become asymptotically normal. If the prior family D allows for normal distribution as a limiting case, then approximation of the mixture by a single distribution from D will asymptotically converge to the true posterior.

An application of particular interest is mixtures of Dirichet processes (MDP Escobar and West 1996), which provides a flexible description for the distribution of heterogeneity. MDPs can be used in data-mining to uncover important clusters in the data.

References

Balakrishnan, S. and Madigan, D. "A One-Pass Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets," Bayesian Analysis, 2006, 1, Number 2, pp. 345-362

DuMouchel, W. "Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System," The American Statistician, August 1999, Vol. 53, No. 3, pp 177-190.

Escobar, M. D. and West, M. "Bayesian density estimation and inference using mixtures," Journal of the American Statistical Association, 1996, 90, pp. 577-588.

Gordon, N., Salmond, D., and Smith, A. F. M. "Novel approach to nonlinear/non-gaussian Bayesian state estimation," IEE Proc. -F Radar, Sonar Navig., 1993, vol. 140, pp. 107–113.

Liu, J. S. and R. Chen, R. "Sequential Monte Carlo methods for dynamical systems," J. Amer. Statist. Assoc., 1998, vol. 93, pp. 1032–1044.

Ridgeway, G. and Madigan, D. "A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets." Journal of Knowledge Discovery and Data Mining, 2002, 7 pp 301-319.

Statistical Modeling: Bigger and Bigger

David Madigan
Rutgers University

Statistics as a discipline exists to develop tools for analyzing data. As such, Statistics is an engineering discipline and methodology is its core. In methodology research, mathematics used to reign supreme. Historically, data existed primarily on paper, and useful inferences owed their existence to deft, computation-avoiding mathematics. Two interrelated events have utterly changed this landscape: ubiquitous computing and ubiquitous databases. Statistics as a discipline has been slow to catch up and our graduate programs in particular are frequently antediluvian. Nonetheless, extraordinarily exciting times await us. In my talk I will describe some challenging statistical applications I have encountered in the last few years that have altered my perspective on the computational scale of things to come.

In the mid-1960's, diagonalizing a 7 x 7 matrix represented a significant computational challenge. Stunning progress has occurred since then leading some, for example, Benzécri, to claim that in data analysis "there is no longer any problem of computation" (Benzécri, 2005). I strongly believe this is not the case and argue that Statistics has an insatiable appetite for computing power.

Application 1: Localization in Wireless Networks
Consider wireless internet service in a public space with multiple access points. The task is to use the vector of signal strengths from the different access points to provide a probabilistic estimate of device location. Considerable commercial and research interest has focused on this problem in recent years (e.g., www.ekahau.com). Statistical approaches dominate: gather signal strength measurement vectors at known locations and build a predictive model that facilitates location estimation. The data-gathering phase ("profiling") is expensive, and Bayesian approaches with carefully chosen prior distributions can drastically reduce the data requirements. However, the computational requirements for the requisite Markov chain Monte Carlo algorithms, even for modest sized applications, prove formidable. The application demands real-time location estimation and tracking and compute-power several orders of magnitude greater than currently available seems to be required.

Application 2: Safety Surveillance for Licensed Pharmaceutical Products
Government-run spontaneous report databases represent the primary data source for monitoring the safety of licensed drugs. Individuals, physicians, and drug companies submit reports containing drug(s), adverse event(s), and basic demographic information. The statistical challenge is to detect novel drug-adverse event associations as they emerge. Since most reports contain multiple drugs and multiple adverse events, the potential also exists for detecting complex multi-drug interactions. About 15,000 drugs are available in the US market. By coincidence, the adverse event dictionary employed by the FDA contains about 15,000 unique adverse events. Thus, from one perspective, the challenge reduces to performing a multivariate regression with a 15,000-dimensional response variable and 15,000 input variables. Including just two-way interactions leads to

over 100,000,000 input variables, well beyond the capabilities of current algorithms. Emerging approaches to drug safety focus on longitudinal databases amassed by HMOs and insurers where the computational challenges are even greater.

Application 3: Bayesian Modeling with Structured Inputs
Many predictive modeling scenarios involve structured input variables. For example, one application I have been involved with attempts to predict the survival of laboratory animals based on twenty assays, each measured at upwards of ten time points over two years. Modified versions of recent ideas in the regularized regression literature (e.g., group lasso and fused lasso) have utility in this context and can deal sensibly with the temporal structure. However, even in modest sized applications, a fully Bayesian approach that properly accounts for uncertainty, quickly overwhelms current computing power. An ideal approach would go even further and simultaneously model latent structure within the data, identifying for example, groups of pathogen-resistant animals. I suspect that we will look back in a relatively few number of years and consider the scale of these applications rather quaint.

Relevant References

Application 1:
Madigan, D., Ju, W., Krishnan, P., and Krishnakumar, A.S. (2006). Location estimation in wireless networks: A Bayesian approach. Statistica Sinica, 16, 495-522.
http://www.stat.rutgers.edu/~madigan/PAPERS/wireless.pdf

Application 2:
Hauben, M., Madigan, D., Gerrits, C., and Meyboom, R. (2005). The role of data mining in pharmacovigilance. Expert Opinion in Drug Safety, 4(5), 929-948.

Application 3:
Balakrishnan, S. and Madigan, D. (2007). LAPS: Lasso with Partition Search. Manuscript.
http://www.stat.rutgers.edu/~madigan/PAPERS/laps.pdf

This paper:
Airoldi, E.M., Blei, D.M., Fienberg, S.E., & Xing, E.P. (2006). Stochastic blockmodels of mixed-membership: General formulation and nested variational inference. ICML Workshop on Statistical Network Analysis, Pittsburgh PA.
http://www.cs.cmu.edu/~eairoldi/nets/icml_sna/paper1_final.pdf
describes a novel model for relationship data of direct relevance to Application 2.

This paper:
Genkin, A., Lewis, D.D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. Technometrics, to appear.
http://www.stat.rutgers.edu/~madigan/PAPERS/techno-06-09-18.pdf
describes some high-dimensional logistic regression algorithms and software that have found widespread use.

Computational Challenges for Real-Time Marketing with Large Datasets

Alan L. Montgomery (alanmontgomery@cmu.edu)
Carnegie Mellon University

Research into the use of data intensive marketing strategies has been intense over the past decade. For example, researchers have proposed methods for analyzing store transaction data to set optimal pricing strategies (Montgomery 1997), the analysis of clickstream data for adaptive web design (Montgomery et al. 2004), and analyzing consumer choice for shopbot design (Brynjolfson, Smith, and Montgomery 2007). This research is just a sampling of a broad stream of applications of Bayesian statistical techniques to marketing problems (Rossi and Allenby 2003). Much of this progress has been due to new statistical estimation procedures that rely upon simulation (Chib 2003).

From a research standpoint these methods have been highly successful, allowing the estimation of previously intractable models. Practically though these simulation techniques can be very time consuming, for example the estimation of a hierarchical choice model could take hours to complete (Allenby, Rossi, and McCulloch 1996). Amazon typically requires its system to respond in 2 seconds or less, which means that response needs to happen in seconds not hours. If the current stream of Bayesian research in marketing is to have an important impact in practice computational methods that can be implemented in real-time are necessary. The goal of this talk is to outline potential methods for developing feasible computational strategy for analyzing Bayesian choice models in a real-time environment.

A promising computational strategy that has been employed by Google, Amazon, Sun, and IBM (as well as a host of others) is to construct grid computing environments (Bryant 2007). This allows companies to harness the power of up to hundreds of thousands of low-cost personal computers by splitting up the computational tasks so that they can be handled in parallel. Grid computing provides a promising direction for being able to analyze the quantitative models employed in marketing problems. Unfortunately the simulation techniques like MCMC used currently to estimate marketing models are by their nature sequential. Brockwell (2005) proposed a method of parallelizing these algorithms. This approach is promising, as well as the ideas of pre-fetching and griddy approaches.

An alternative strategy is to use specialized electronic devices like ASICs and FPGAs (Brown and Rose 1996). Traditional microprocessors execute a set of instructions to perform a computation. By changing the software instructions, the functionality of the system is altered without changing the hardware. The downside of this flexibility is that the performance can suffer. The processor must read each instruction from memory, decode its meaning, and only then execute it. This results in a high execution overhead for each individual operation. The second method of algorithm execution is to use hardwired technology, like an Application Specific Integrated Circuit (ASIC). ASICs are designed specifically to perform a given computation, and thus they are very fast and efficient when executing the exact computation for which they were designed. However, the circuit cannot be altered after fabrication. FPGAs lie between these two extremes, being cheaper and more flexible than ASICs, while also being much faster than software programmable microprocessors. FPGAs contain an array of computational elements whose functionality is determined through multiple programmable configuration bits. For certain types of computations needing integer or fixed point arithmetic the benefits of FPGAs can be very significant, two orders of magnitude speed improvement compared to using a conventional cluster CPU. This performance increase comes without tuning to take advantage of the massive parallelism and pipelining.

The goal of this talk is to understand what computational problems must be addressed if currently proposed models in marketing are to be used in real-time environment. This means that not only do we need to make predictions with the model, but we must also be able to perform tasks such as estimation, inference and optimization in a real-time environment. These approaches must be designed to meet the unique attributes of the marketing problems which include large databases and decision theoretic problems.

References

Allenby, Greg M., Robert McCulloch, and Peter E. Rossi (1996), "The Value of Purchase History Data in Target Marketing", *Marketing Science*, 15, 321-340.

Brown, Stephen and Jonathan Rose (1996), "Architecture of FPGAs and CPLDs: A Tutorial", *IEEE Design and Test of Computers*, 13 (2), 42-57.

Bryant, Randal E. (2007), "Data-Intensive Supercomputing: The case for DISC", Carnegie Mellon University, School of Computer Science, Working Paper CMU-CS-07-128.

Brynjolfson, Erik, Michael Smith, and Alan Montgomery (2007), "The Great Equalizer: An Empirical Study of Choice in Shopbots", Carnegie Mellon University, Tepper School of Business, Working Paper.

Brockwell, Anthony E. and Joseph B. Kadane (2003), "A Gridding Method for Bayesian Sequential Decision Problems", *Journal of Computational and Graphical Statistics*, 12 (3), 566-584.

Chib, Siddhartha (2003), "Monte Carlo methods and Bayesian computation: Overview", S. E. Fienberg, J. B. Kadane, eds. *International Encyclopedia of the Social and Behavioral Sciences: Statistics*. Elsevier Science, Amsterdam, The Netherlands.

Montgomery, Alan L. (1997), "Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data", *Marketing Science*, 16 (4), 315-337.

Montgomery, Alan L, Shibo Li, Kannan Srinivasan, and John Liechty (2004), "Modeling Online Browsing and Path Analysis Using Clickstream Data", *Marketing Science*, 23 (4), 579-595.

Rossi, Peter E. and Greg M. Allenby (2003), "Bayesian Statistics and Marketing", *Marketing Science*, 22, 304-329.

Inverse Regression Methods: A Review, Applications and Prospects

Prasad A. Naik
University of California at Davis

In business and economics, parametric models such as logistic regression have become popular tools to predict customer choice, i.e., whether a customer would buy a firm's products or defect to competitor's services. However, parametric models suffer from the drawback that they have to pre-specify the shape of an underlying probability function. When the pre-specified function departs from the true unknown link, it cannot accurately predict rare events with low probability of occurrence. Nonparametric binary models overcome this drawback (Cosslett 1983). These models estimate a flexible probability function that does not depend on a particular functional form. But as the number of regressors increases, nonparametric models suffer from the curse of dimensionality, which induces the empty space phenomenon (see, e.g., Simonoff 1996, p. 101). Consequently, nonparametric approaches break down when datasets contain a large number of regressors (i.e., predictor variables), which are often encountered in many business applications, and so their applicability is restricted to small datasets with ten or fewer variables. Thus parametric models are not sufficiently flexible, whereas nonparametric models are not scalable to large databases encountered by businesses.

To overcome these problems, semi-parametric models have been proposed that relate an index — a linear combination of regressors — to the expected response via a nonparametric link function. Although semi-parametric models improve the computational properties at the expense of a somewhat restrictive formulation, they still have two limitations. First, they assume a single index to underlie consumer response; second, this index is limited to a small number of variables for computational reasons.

In this talk, I describe a model that addresses both these limitations. Specifically, Naik, Wedel and Kamakura (2007) introduce a Multi-Index Binary Response (MBR) model and develop a non-iterative two-step method to estimate it. They call the two steps *projection* and *calibration*. In the *projection* step, they combine information available in high-dimensional predictor space and project it into a low-dimensional index space. They achieve this dimension reduction by estimating an index structure (i.e., their composition in terms of original regressors and the number of indexes to retain) without specifying the link function (Cook and Weisberg 1991). In the *calibration* step, they estimate the unknown link function via local polynomial regression (or a parametric model).

One of the main advantages of the MBR model is that, by allowing for multiple indexes, it facilitates a more refined understanding of consumer behavior, as we will illustrate using real data from a telecom company. It is likely that consumers live in a two-dimensional plane, $z = (z_1, z_2) \in \Re^2$, some of them more responsive than others, if we find empirical evidence for the presence of two indexes. In contrast, consumers must lie on a line (i.e., $z \in \Re^1$) in the extant single-index probability models. Thus, the MBR model augments the scope of possibilities for profiling and targeting consumers.

Given the above application to choice modeling as a backdrop, I next review the broader class of methods known as "Inverse Regression Methods" for dimension reduction, such as *Sliced Inverse Regression* (SIR; Li 1991) or *Sliced Average Variance Estimation (*SAVE; Cook and Weisberg 1991), and *Constrained Inverse Regression* (CIR; Naik and Tsai 2005). In addition, I will describe recent approaches to determine how many reduced dimensions to retain in a given application, for example, by using analytical tests, permutation tests (Cook and Yin 2001), or information criteria (Zhu, Miao and Peng 2006). Finally, I will discuss recent advances in estimating models when the *number of regressors exceeds the number of observations* via a novel approach known as *Partial Inverse Regression* (PIR; Li, Cook and Tsai 2007).

References

Cook, R. D. and Weisberg, S. (1991), "Discussion of Li (1991)," *Journal of the American Statistical Association*, 86, 328-332.

Cook, R. D. and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis (with Discussions)," *Aust. N. Z. J. Stat*, 43 (2), 147-199.

Cosslett, S. R. (1987), "Efficiency Bounds for Distribution-Free Maximum Likelihood Estimator of the Binary Choice and Censored regression Models," *Econometrica*, 55, 559-585.

Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316-342.

Li, Lexin, R. D. Cook and C.-L. Tsai (2007), "Partial Inverse Regression," *Biometrika*, forthcoming.

Naik, Prasad A. and Chih-Ling Tsai (2005), "Constrained Inverse Regression for Incorporating Prior Information," *Journal of the American Statistical Association*, 2005, 100 (469), 204-211.

Naik, Prasad A., Michel Wedel, and Wagner Kamakura (2007), "Multi-Index Binary Response Analysis of Large Datasets," under review.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York, N. Y.: Springer.

Zhu, Lixing, Baiqi Miao, Heng Peng (2006), "On Sliced Inverse Regression With High-Dimensional Covariates," *Journal of the American Statistical Association*, 101 (474), 630-643.

State of the Art Recommendation Systems: the Good, the Bad and the Ugly

Michel Wedel
University of Maryland

Product Recommendation Systems have become the backbone of some of the Internet economy's largest and most successful firms. Companies such as Netflix, Eachmovie, Movielens and Blockbuster ask customers to rate movies they've viewed as a basis for recommendations; CDNow, iTunes, Audioscrobbler , MSN Music, RealPlayer MusicStore, Rhapsody, and Napster, use such systems for recommending music recordings; Amazon, Barnes & Noble and Storycode implement recommendation systems for books; Findory uses a collaborative filtering-based algorithm to recommend interesting news items; and TiVo does that for TV shows. User ratings are now collected on most e-tailers web-sites, but not always actively used to generate recommendations. The most recent trend, implemented by for example AudioScrobbler, StoryCode and MovieLens is to shift collaborative filtering from e-tailers to stand-alone services, moving to open source collaborative filtering systems, and from ratings-only data to a variety of actual listening/reading/viewing behavior, enabling users to update their own information.

The academic literature on recommendation systems has focused on the development of model-based methods that explicitly hypothesize a probabilistic data generation process. Model-based methods that have been used to generate product recommendations include Bayesian network models (Breese, Heckerman, and Kadie 1998), mixture models (Chien and George 1999), hierarchical Bayes' models (Ansari, Essegaier, and Kohli 2000), and hierarchical Bayes selection models (Ying, Feinberg and Wedel 2004). These models in most cases show substantial improvements in the quality of recommendations on test datasets, and are mostly estimated with MCMC algorithms.

An industry has emerged to offer solutions to businesses leveraging the product evaluations of prior customers and that industry has to address issues of the massive data being accumulated in the product recommendation databases. The academic literature has followed industry practice and has developed refined methods for making recommendations, but in most cases these methods are not applicable to the massive data sets that result from accumulating product ratings and/or transactions for the entire customer database of e-tailers. But in addition, there are several problems with recommendation systems based on user-provided product ratings.

1. *Massive numbers of choice alternatives.* Even after recommendations have been generated, the number of choice alternatives facing consumers is massive. In addition product attributes are hard to evaluate before consumption or not very helpful in anticipating the product's idiosyncratic consumption utility. Examples include hedonic products such as music, movies, books, restaurants, and other categories for

which there are massive numbers of alternatives and there is a substantial experiential component. And, typically these are the product categories for which recommendation systems have been implemented and/or have been found to be most useful. Recently statistical approaches have been developed to accommodate massive numbers of choice alternatives (Lans, Pieters, Wedel 2007)

2. *Missing data.* Rarely is the customer familiar with even a small fraction of the massive number of alternatives provided on the e-tailers' product catalogs. Customers can evaluate only products they have experienced, so they commonly rate only a very small subset of all available items, perhaps only those they like or dislike. Consequently, considered relative to the entire product catalog, the ratings history of any particular customer is extremely sparse, even within a single product category. In addition, much of the product rating data on which recommendation systems need rely is not only missing, but missing *non-randomly*. There is a statistical and marketing literature on how to address missing data that is applicable in this context (Ying, Feinberg and Wedel 2006).
3. *Scale usage heterogeneity.* Most recommendation systems rely on user-provided input, where product ratings are typically made on ordinal scales. For example, Amazon, Netflix, Eachmovie and Blockbuster each ask customers to award products 1-5 stars. Also, the academic work on recommendation systems has focused on ratings-based systems. But, the psychology and marketing literatures have revealed that people use scales differently, having various tendencies to score at the middle, or extremes. Recommendations based on such ratings may reflect scale usage behavior rather than product preference. Several approaches have been suggested in the literature to remove such scale usage behavior (Rossi, Gilula and Allenby 2001).
4. *Endogeneity.* Recommendation systems generate recommendations that are based on past usage or browsing behavior that reflects past interest. Subsequent choice behavior from customers is largely constrained by the recommendations received in the past. In making recommendations, recommendations generated themselves in the past should be accounted for as part of the data generating mechanism, which is not often the case. This creates problems for correlation, regression- and other model-based approaches that are based on the assumption that the design matrix is uncorrelated with the estimation error. Over time, biases will accumulate and the quality of the recommendation will decline. There is a substantial literature in economics and marketing on how to address endogeneity, even when "instrumental variables" are not available (Ebbes, Wedel, Bockenholt and Steerneman 2005).
5. *Shilling.* One way to increase recommendation frequency is to have a group of users (human or agent) provide specially crafted ratings to the recommendation system, that cause it to make the desired recommendation more often. Instances of these manipulations have been observed, for example, it has been shown that a number of book

reviews published on Amazon.com are actually written by the author of the book being reviewed. Shilling attacks have been shown to be effective in particular for infrequently recommended items (Lam and Riedl 2004).

6. *Scalability.* The academic literature on recommendation systems has focused on the development of model-based methods that explicitly hypothesize a probabilistic data generation process. While these models in most cases show substantial improvements in the quality of recommendations on toy datasets, they are mostly estimated with MCMC algorithms that are not scalable to datasets of the size encountered in practice. In addition, the models in question have been developed as forecasting models that predict the success of one product at-a-time, assuming the product with the highest predicted ranking to be recommended. Recently, the statistical literature has developed methods that address the scalability of MCMC methodology to real life problems (Ridgeway and Madigan 2002).

In the presentation, I will discuss the growth in number and content of recommendation systems based on a search of the Google patent database. And, I will present a number of (partial) solutions to each of the five problems mentioned above. I will advocate locally distributed model-based recommendation systems, estimated on usage data with MCMC, model averaging to leverage the information from different users, and batch recommendations based on sequential optimal experimental design.

References

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), "363-75.

Breese, Jack, David Heckerman, and Carl Kadie (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, WI: Morgan Kaufmann Publisher.

Chien, Yung-Hsin and Edward I. George (1999), "A Bayesian Model for Collaborative Filtering," *Working Paper*, University of Texas at Austin.

Ebbes, P., M. Wedel, T.Steerneman, U. Bockenholt (2005), "New Evidence for the Effect of Education on Income: Solving Endogeneity with Latent Instrumental Variables," *Quantitative Marketing and Economics*, **3,** 2005, 365-392.

Lam, S.K. and Riedl, J. (2004). "Shilling recommender systems for fun and profit," In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM Press (2004) 393–402.

Lans, R. Van der, F.G.M. Pieters, M.Wedel (200&). "Eye Movement Analysis of Search Effectiveness," *Journal of the American Statistical Association*, Forthcoming.

Ridgeway, G. and Madigan, D. (2002) "A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets." *Journal of Knowledge Discovery and Data Mining*, 7 pp 301-319.

Rossi, P., Z. Gilula and G. Allenby (2001), "Overcoming Scale Usage Heterogeneity: a Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, Vol. 96, 20-31

Ying, Y., F.Feinberg. M.Wedel (2006), "Improving Online Product Recommendations by Including Nonrated Items," *Journal of Marketing Research***,** 2006, 43 (August), 355-365.

Issues in the Modeling of Behavior in Online Social Networks

Anand V. Bodapati
UCLA Anderson School of Management

Social Networks are a centerpiece phenomenon in the so-called "Web 2.0" wave. Web 2.0 web sites are those where the content is largely user-generated rather than firm-generated. In most social network web sites, revenue is generated primarily through exposure of advertising to those who visit the site. For the firm to increase the number of opportunities for deploying advertising, it needs to (a) increase the number of users in its online community, (b) achieve "stickiness" so that retention of users once acquired is high, (c) the volume of page views per user per unit time is high. These three factors are of focal interest for pretty much any firm which operates on a revenue model. However, for a social networking site they are of special interest because the extent of factors (a), (b) and (c) for any user depends on the extents of these factors for other users. Users generate content and create online activity for others. If there is not enough of content generated by users, then there may not be a lot of consumers who consume content because there is not much content to consume. And if there are not enough content consumers, then there may not be enough of an audience for the content creators to feel incentivized enough to create content. Left on its own therefore, the audience base in social networks therefore may have a tendency to be self-propelling or self-deflating.

If the firm is fortunate, then the social network will be largely self-propelling, without much intervention on the part of the firm. This indeed is what happened with at least one dominant social network firm. But if the firm is to actively influence the dynamics of the social network, what is it to do? In the management of social networking sites, specific areas in which the firm can influence the course of the social networking site are those with respect to: (i) Creating sufficient initial mass of creators and consumers of content, (ii) Directing the information consumers to creators and information units that would be of interest, (iii) Providing appropriate incentives to the information creators and (iv) Identifying and nurturing those information creators who are of special importance to site usage volume.

In a user-generated content environment, the role of content creators users cannot be overestimated. To attract traffic, a social network site firm cannot go much beyond periodic updates of site features and design elements. The bulk of digital content —the driving force of the site's vitality and attractiveness — is produced by its users. Users, though, are not all created equal. There is a great deal of heterogeneity across community members in terms of the frequency, volume, type, and quality of digital content generated and consumed. In a recent article in The Wall Street Journal, Elizabeth Holmes provides anecdotal evidence that a user's interest in a site highly correlates with content updates produced by some key content contributors. Holmes reports that when a popular blogger left a particular site for a two-week vacation, the site's visitor tally fell. Clearly, not all user-generated content is demanded by other users. In Holmes's example, content produced by three invited substitute bloggers could not prevent site traffic from decreasing.

The situation with social sites is actually more complex than in the picture we have painted so far. We spoke as if some members create content and others consume content. However, individual users are often involved in both of these activities. Furthermore, the average user is likely to be attracted to the social network site by content produced not just by one, but by a number of other users. Accordingly, in contrast to Holmes's story, removing an individual content provider from the site may or may not affect the level of community participation. The community members may choose to maintain their usual activity levels; however, they would likely reallocate their content consumption to other sources available within the site. Finally, members of an online community might be attracted to the site not just by content produced by other members, but also could be motivated to contribute by others' consumption of their own content. Extending this analogy to the Holmes example, a popular blogger may stop producing new posts if it is evident that people have lost interest in reading them.

In my presentation, I will speak of the four areas I list above where the firm can manage the course of the social network. My particular emphasis will be on the the last of the four areas. From a managerial perspective, understanding who the users are and who keeps the social network site running is important. Who are the key players? What (or who) drives site visitation - and therefore advertising revenues? How should the social network site segment its customers? To manage the site, to make it a better place for consumers, and to obtain better information for advertisers, the site's management needs to know exactly what role is being played by each individual user within the network and whose actions have an impact on whom. An important objective ought to be to shed light on these questions and to develop an approach that will help managers identify those social network site users who are influential in affecting a firm's primary revenue driver: other members' site usage.

Some Key References

Breiger, Ronald, Carley, K., Pattison, P. (eds) (2003), *Dynamic Social Network Modeling and Analysis*, The National Academies Press.

Iacobucci, Dawn (ed.) (1996), *Networks in Marketing*, Sage Publications.

Trusov, Michael, Bodapati, A. V., Bucklin, R. E. (2007), "Your Members Are Also Your Customers: Marketing for Internet Social Networks," Working paper, UCLA Anderson School.

Van den Bulte, Christophe, Wuyts, S. (2007), *Social Networks and Marketing*, Marketing Science Institute".

Wasserman, Stanley, Faust, K. (1994), *Social Network Analysis*, Cambridge University Press.