Estimation of Stochastic Preferences: An Empirical Analysis of Demand for Internet Services

Walter Beckert[*], University of Florida

November 30,2000

Abstract

The rapid increase in demand for Internet services and new, bandwidth- and time-intensive applications require high quality access to the Internet. Service quality may be assured through efficient allocation of Internet access capacity. Efficient capacity allocation can be achieved through nonlinear prices. Nonlinear pricing is motivated through preference heterogeneity. The objective of this paper is to develop an econometric model of Internet users' preferences over service attributes that calibrates unobserved preference heterogeneity. To this end, a stochastic preference model is proposed and estimated on data from the U.C. Berkeley Internet Demand Experiment (INDEX).

A structural econometric framework is developed within which a consumer learns in the process of service consumption and randomness in choices, conditional on prices and expenditure, arises from unobserved heterogeneity in the consumer's preferences. Heterogeneity in preferences is estimated using a simulation-assisted estimation methodology. The empirical analysis of Internet user data shows that considerable heterogeneity in preferences exists, among different users and for each user over the observation horizon. Moreover, users appear to learn in the consumption process, deviating in their on-line valuations from their ex ante consumption plans. A user's variation in on-line valuations also typically exceeds the variation in ex ante valuations. For the purpose of demand management, the estimated model appears to quite accurately predict the distribution of the continuous choice variables, conditional on discrete service choice, and can be used to explore different pricing scenarios.

KEYWORDS: Internet Demand, Preference Heterogeneity, Bayesian Learning, Discrete-Continuous Choices, Method of Simulated Moments

CORRESPONDENCE should be directed to:
Walter Beckert, Department of Economics, 224 Matherly Hall, PO Box 117140, University of Florida, Gainesville FL 32611 - 7140; Tel: 352 - 392 0113, Fax: 352 - 392 7860, e-mail: beckert@ufl.edu

[*]I thank Chunrong Ai, David Brillinger, Daniel McFadden, Paul Ruud, David Sappington, Kenneth Train, Pravin Varaiya, Hal Varian and the INDEX team, and participants at the MIT/Tufts ISQE Conference for helpful comments and discussions. All errors are mine. This work was funded in part by the National Science Foundation, grant ANI-9714559, and in part by the Cal@Silicon-Valley Fellowship of U.C. Berkeley.

1 Introduction

The rapid increase in demand for Internet services and the emergence of new, bandwidth- and time-intensive applications require high quality access to the Internet. Service quality may be assured through efficient allocation of Internet access capacity. In theory, efficient capacity allocation can be achieved through nonlinear pricing. Specifically, quality-differentiated and usage-based prices may be superior to the flat-rate pricing prevalent in today's marketplace. The fundamental economic motivation for nonlinear pricing is preference heterogeneity (see, for example, Wilson (1993)). Diversity in consumers' preferences renders different service and associated pricing options advantageous because offering such options enhances allocative efficiency by allowing consumers to self-select the service they desire. The objective of this paper is to develop a structural econometric model of Internet users' preferences over Internet services that can calibrate unobserved preference heterogeneity. To this end, a stochastic preference model is proposed and estimated on data from the U.C. Berkeley Internet Demand Experiment (INDEX). INDEX provides a prototype implementation of the technology necessary to implement nonlinear Internet service pricing. This technology also makes it possible to collect disaggregate demand data that allow calibration of Internet users' preference heterogeneity and make implementation of optimal nonlinear pricing feasible.

Assessing how heterogeneous consumers value quality-differentiated services and utilize them, more generally, is important in its own right. It is of particular interest in the context of services provided by capacity-constrained resources. In such cases aggregate utilization of a service typically determines service quality delivered to the individual consumer. Access capacity to the Internet, measured as bandwidth or transmission speed in kilobits per second (*kbps*), is typically shared among users. Aggregate utilization of a given nominal transmission speed determines the service quality in terms of effective transmission speed that is actually delivered to the individual user. The stochastic preference model proposed and estimated in this paper can be used to assess how Internet users differ in their valuation of service quality and utilization behavior and how the consumption experience per se alters their valuations. While the model is motivated in the context of demand for Internet services, it is cast in sufficient generality to use it as a template, easily adaptable to other contexts in which customers select a service capacity or nominal service quality and subsequently choose how to utilize it. The case of demand for electricity, almost prototypical for nonlinear pricing, is a good example: Customers select a service quality (availability, priority and assurance of dispatch) and subsequently choose power level, duration and thereby total energy. Another topical example is cellular phone service. Customers compete for capacity resources, in particular at peak hours; they initially choose coverage area and a nonlinear tariff and subsequently utilization. In Europe, wireless networks are also already being used for quality-differentiated data services.

The choice model proposed in this paper mimics the choice situation that INDEX users face and, thus, addresses two choice questions. The user is offered a menu of different bandwidth choices. Each bandwidth in the choice set comes with a different price for byte volume transmitted and subscription time spent in this bandwidth. The user chooses a bandwidth

and subsequent utilization in terms of transmitted volume and time. The choice model has two distinct features. First, it models the user's discrete bandwidth capacity choice as jointly endogenous with the user's continuous choices of byte volume and subscription time, given the chosen bandwidth. Second, unlike in previous analyses of discrete-continuous choice problems (Dubin and McFadden (1984), Dubin (1985)), the *sequence* of discrete choice followed by continuous choices is modeled, distinguishing ex ante preferences that give rise to the initial capacity choice from ex post or on-line preferences that rationalize subsequent continuous utilization choices. This distinction arises because the on-line consumption experience itself may alter the user's valuations. The modelling framework stipulates that (Bayesian) consumers have preferences over service products that can be produced through different service technologies; the quality or productivity of technology inputs is ex ante unknown, but consumers can learn about it in the process of consumption; "reduced-form" preferences, as the composition of utility and service production technology, thus are ex ante stochastic. One interpretation of Internet users' ex ante uncertainty about their on-line service valuations is uncertainty about the quality of information embodied in transfered data. Data are transmitted on-line; the quality of embodied information is revealed in the process of data transmission and allows users to update beliefs about latent quality parameters. The user's expectation about the quality of information determines ex ante valuations and thereby the initial bandwidth choice. The quality of information revealed in the transmission process then determines ex post valuations and thereby the total byte volume being transmitted.

The distinction of ex ante and on-line (reduced-form) valuations has a number of important implications. Since the quality of information is unknown ex ante and beliefs about it are re-assessed online, the user may demand connection time in excess of the minimal time necessary for transmission. This excess time may be used to intellectually process and evaluate information. The price paid for intellectual processing time reflects the value that the user places on the convenience to keep the option to transmit data while processing information, rather than disconnecting. Intellectual processing time can therefore synonymously be thought of as convenience time. Thus, uncertainty about information quality or productivity gives rise to the distinction between ex ante and on-line valuations and introduces a new good, convenience time, into the analysis. To the extent that users demand convenience time, capacity is claimed, but not fully utilized, and users are willing to pay for the mere option to utilize it. Connection time for convenience can be clearly identified from cumulative utilization curves as in figure 1 that graph cumulative utilization of a bandwidth against time connected to it; the fraction of time for which utilization is zero represents claims on capacity for convenience.[1] The figure also suggests that demand for convenience time is price sensitive. Capacity that is claimed but not utilized is generally allocated inefficiently, but optimal usage-based prices can enhance the allocative efficiency. The structural modeling approach proposed in this study permits to estimate the impact of the consumption experience on valuations, as a prerequisite

[1]Reading the graph from the vertical axis, for any given utilization u, it displays the fraction of time of a user's connection to 128 *kbps* for inbound traffic during which actual utilization of 128*kbps* was at least as high as u. A connection is defined as the time between log-on and log-off to a bandwidth. A more detailed discussion of this and related figures can be found in section 4.1.1.

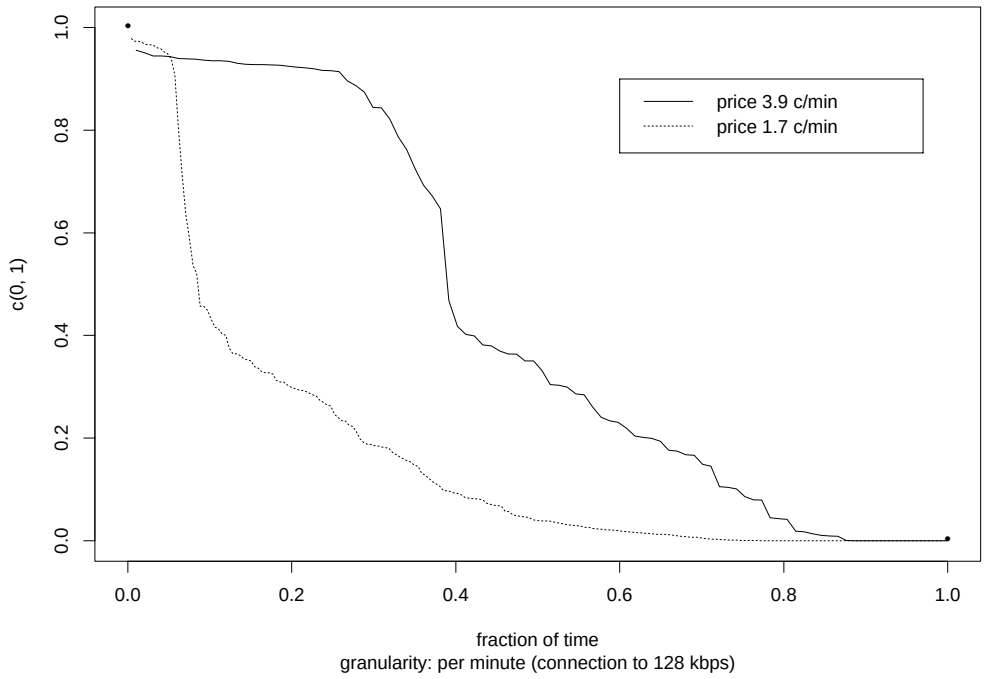

Figure 1: Utilization of 128 *kbps*, at 3.9 and 1.7 c/min.

for the design of optimal prices.

Total transmitted volume and convenience time are chosen once the user is committed to a bandwidth and once the quality of information is revealed in the process of transmission. The observed capacity choice may then appear suboptimal ex post: Another bandwidth choice may have afforded the same continuous choices at lower cost and possibly higher speed. But this bandwidth was not perceived optimal ex ante, for otherwise it would have been chosen. The distinction between ex ante and ex post valuations thus has the second implication that it rationalizes choice behavior which is seemingly suboptimal ex post. Econometrically, ex post suboptimal choice behavior allows to identify the degree of ex ante uncertainty about the actual on-line service valuation.

The contribution of this analysis is threefold. First, it develops a structural econometric framework within which consumers learn in the process of consumption. Consumers' discrete-continuous choices are modeled as jointly endogenous, emerging from ex ante and ex post service valuations respectively. In this framework randomness in discrete and continuous demands, conditional on prices and expenditure, arises from preference heterogeneity. This distinguishes this work from the usual motivation of randomness in reduced-form continuous demand systems as arising from measurement error. As seemingly suboptimal choices, violating revealed preference postulates, are utilized to identify intertemporal preference heterogeneity, this approach

complements analyses of the impact of heterogeneity on average demand systems (Beckert (2000), Blundell et al. (1998), Brown and Matzkin (1995, 1995A), Brown and Walker (1989), McElroy (1987), Lewbel (1996)).[2] Second, it outlines and implements a simulation-assisted estimation methodology to estimate heterogeneity in preferences of Internet users on data for a subset of the INDEX subject pool. Third, the empirical analysis displays some of the potential usefulness of this approach for demand management. The results demonstrate that considerable heterogeneity in preferences exists, both among different users and for each user across the user's connections to the service provider. Moreover, a user's variation in ex ante valuations is typically surpassed by variation in ex post valuations. And users appear to learn in the comsumption process, deviating in their on-line service valuations from their ex ante valuations. The results also show that the estimated heterogeneity in preferences is consistent with nonparametrically detectable features the data. Finally, from the perspective of capacity management, the estimated model appears to quite accurately predict the distribution of the continuous choice variables, conditional on a discrete choice. This recommends the model for demand management. It is shown how the model can be used to explore the impact of different pricing scenarios.

The paper proceeds as follows. The second section gives an exposition of the proposed discrete-continuous choice model, with particular emphasis on the distinction between ex ante and ex post valuations. The third section investigates the statistical properties of the model, parameter identification and feasible estimation methodologies. The fourth section gives an exploratory account of the INDEX data and summarizes the econometric data analysis in the context of the proposed model. The final section concludes. A further discussion of modeling issues and limitations can be found in an appendix.

2 Econometric Model Specification

2.1 Overview

The development of the econometric model is guided by the INDEX data available at the time of this study. INDEX is a market trial for quality-differentiated Internet service. For details about technology and experimental design, see Rupp et al. (1998). INDEX provides home Internet access over ISDN lines to a group of users affiliated with U.C. Berkeley (students, faculty and staff). Users connect to a control gateway, selecting a service in terms of a bandwidth (nominal transmission rate, measured in kilobits per second, *kbps*), and pay for their usage according to prices that are differentiated by quality of service. The data are observed for each of

[2]This literature, in part, is concerned with conditions on the specification of heterogeneity parameters that induce reduced form demand systems which satisfy the well known implications of utility maximization for demands, and with biases introduced into estimation and non-parametric revealed preference tests by averaging across heterogeneous consumers.

a user's connections. A connection is defined as the time between log-on to a chosen bandwidth and log-off. A user is not committed to a bandwidth, but can switch to another bandwidth or disconnect instantaneously, by clicking on a button on a control window. In each connection, the user then chooses a bandwidth capacity (in *kbps*) out of a menu of discrete bandwidths, and, conditional on this capacity choice, through web-based applications subsequently chooses transmitted volume v (in bytes) and convenience time t (in seconds) consumed in this connection.[3] Next to these usage data, prices for unit volume, $p_v(b_i)$, and unit connection time, $p_t(b_i)$, are observed for all of m available bandwidths $\{b_i\}_{i=1}^m$. Covariates characterizing the connection, like time of day and a working day dummy, are recorded as a vector $\mathbf{z}'$. The data are best thought of as non-equispaced, unbalanced panel data.

It should be admitted at the outset that the informational content of these data available at the time of this study, while allowing to calibrate variation in preferences both across users and distinguishing ex ante and ex post valuations, do not permit to attribute such differences to covariates that characterize applications, types of user activity or give an indication of the higher-level consumption activities that Internet services feed into. This is a data question, not a modeling question. Once such covariate information is made available, this additional information can serve to explain some of the variation that this analysis finds in users' preferences. Such information is also a precondition for studying switching between bandwidths. This analysis therefore does not attempt to model switching behavior.

An Internet user's preferences are presumably defined over some Internet service product which is generated by some technology or web-based application for which transmitted volume and convenience time act as inputs, much in the style of either a Lancaster-type theory of consumption or two-stage budgeting models[4]. Both this product and its production technology are unobservable to the analyst and therefore empirically unidentifiable; only volume and convenience time are measurable, conditional on capacity choice. This motivates the specification of a "reduced form utility" model, with a user's preferences defined over observable inputs such as byte volume and time, conditional on a chosen bandwidth, rather than higher-level characteristics or goods which are unobservable in the data available for analysis. The following subsections lay out such a reduced form model for stochastic preferences. It distinguishes the user's ex ante from his or her ex post valuation of these inputs and motivates this distinction as a consequence of learning about the ex ante unknown productivity of inputs to the Internet product technology, such as byte volume. And it derives the user's discrete bandwidth capacity choice on the basis of the ex ante valuation of anticipated services, while continuous demands result from ex post, on-line valuation or actual service experience.

The model described in the following subsections is cast in the context of demand for Internet

[3] While bandwidth and byte volume are directly observable, convenience time is not and needs to be estimated. Section 4.1.1 provides details on how this can be done.

[4] For the former, see Lancaster (1966, 1971, 1979). For the latter, see e.g. Hausman, Kinnucan and McFadden's (1979) model for household electricity demand under time-of-day pricing; households choose aggregate electricity consumption per day on the first stage, and relative consumption for each time interval in a partition of the day on the second stage.

services. The notions of capacity or service quality and utilization in terms of time and volume, however, are sufficiently general to make it more widely applicable. The examples of demand for electricity generation or cellular phone services given in the introductory section may serve as an example. In the latter example, for instance, the analyst might observe time spent in the coverage area, the time when the phone is used and mere stand-by time, the nonlinear tariff of the service provider and, next to bills paid to the provider, total expenditure on all communication services. The range of potential applications seems worth the extra cost of a general modeling framework. One might think that, e.g, a budget constraint for each discrete-continuous choice instance may not matter in the Internet context, but does matter in the electricity and cellular phone service examples, where prices and expenditures are much higher. Both the conceptual modeling framework and the estimation methodology are developed at this level of generality. For the empirical analysis of INDEX data, the model is estimated conditional on expenditure for Internet services, thus replacing the budget constraint by a constraint on joint expenditure. The presentation of the more general model framework also allows to assess potential estimation biases that arise if joint expenditures are correlated with the residuals in the stochastic demand equations. Section 4.2 discusses how such biases can be tested for.

2.2 Related Work

The econometric specification starts from a structural stochastic preference or random utility model as a representation of heterogeneous Internet user preferences. Preference heterogeneity is the motivation for and primitive in the theoretical development of optimal nonlinear prices. In order to implement the prescriptions of the theory of nonlinear prices for optimal demand management, it is necessary to empirically assess its primitives, i.e. to calibrate preference heterogeneity. This logic calls for a structural, rather than reduced-from econometric approach. Compared to reduced form approaches, as in Varian (2000), the structural approach, jointly modeling the sequential discrete and continuous choices, allows to estimate the change in a user's service valuations induced by learning about the service (inputs) in the process of consumption, in short: by the consumption experience. Such changes in tastes are identifiable through bandwidth capacity choices that ex post appear suboptimal.

Notice the difference to other analyses of discrete-continuous choice problems, like Dubin and McFadden (1984) and Dubin (1985). These study a single discrete-continuous choice pair of an economic decision maker. Their model of the unit-of-electricity-consumption's conditional indirect utility allows for unobserved characteristics. They maintain, however, the hypothesis that discrete and continuous choices are made contemporaneously. For these reasons, they lack the ability to distinguish ex ante from ex post valuations of the decision maker and cannot assess how preferences change as a result of the consumption experience per se. The focus on the joint endogeneity of discrete–continuous choices adopted here, at the expense of a more elaborate serial dependence in a dynamic programming framework, distinguishes this work from Rust (1987,1994) on controlled stochastic processes. Rust models a sequence of binary

indicators of consecutive discrete investment decisions as the optimal policy of an intertemporal cost minimization problem, leaving aside the joint endogeneity of investment and equipment utilization.[5]

2.3 Theoretical Framework: Bayesian Learning by Consuming

Consumers derive utility from consumption goods. Broadly definded, these can be distinguished as physical goods and service goods. The latter category comprises, for example, communication, entertainment, research, purchasing or budget and financial management. Suppose that utiliy is defined over a service level A and an unspecified outside good x, say as $\tilde{U}(A, x)$, for a continuous, monotonically increasing, concave and strictly quasiconcave utility function $\tilde{U} : \mathbb{R}^2_+ \to \mathbb{R}$. The service level A, in turn, can be achieved using different technologies. In the case of communication, for instance, available technologies may be traditional phone service, voice over IP, eletronic data transfer as in e-mail, surface mail etc.; in the case of research, such services can be websearches, telephone surveys, or library searches. Each of these technologies may itself be offered in various quality gradations.

Each service production technology $i = 1, \ldots, m$ out of a menu of m available technologies uses certain production inputs $\mathbf{w}_i$. Ex ante, the actual productivity of each technology may be unknown to the consumer. This uncertainty can be interpreted as uncertainty about the quality of the inputs the technology uses. In the research example, for instance, the level of research success depends on the quality of information that each technology is capable to retrieve; the quality of information typically is ex ante unknown. Similarly, the effectiveness of communication depends on the quality of transmitted information. Let α denote the true, latent and unknown parameter that captures this ex ante uncertainty. Then, each production technology may be represented by a production function $a_i(\mathbf{w}_i, \alpha)$, monotonically increasing in $\mathbf{w}_i$ for each α, where the productivity of inputs $\mathbf{w}_i$ depends on α. Assuming that the consumer chooses one technology to obtain a service, the achieved service level is $A = \prod_{i=1}^{m} \left[a_i(\mathbf{w}_i, \alpha)\right]^{1_{\{i\}}}$, where $1_{\{i\}}$ denotes an indicator taking value 1 if technology i is chosen. This leads to $\tilde{U}(A, x) = \tilde{U}\left(\prod_{i=1}^{m} \left[a_i(\mathbf{w}_i, \alpha)\right]^{1_{\{i\}}}, x\right) = \sum_{i=1}^{m} 1_{\{i\}} \tilde{U}\left(a_i(\mathbf{w}_i, \alpha), x\right)$. Assume for the remainder that the composition of the functions $\tilde{U}$ and a_i is concave in α.

Suppose that the consumer has prior beliefs $\pi(\alpha)$ about the distribution of α. Even though α is unknown, the consumer may learn about it in the process of service utilization and revise his or her ex ante beliefs accordingly. Suppose the consumer, after having chosen a service production technology, receives a signal θ, conditional on α. Once a technology is chosen, in the process of consumption itself the consumer can revise his or her beliefs about the latent α via the signal θ. Learning can be thought of as processing inputs and the signal θ in order to form posterior beliefs $\pi(\alpha|\theta, i) = \frac{\pi(\theta|\alpha,i)1_{\{i\}}\pi(\alpha)}{E[\pi(\theta|\alpha,i)]}$ about α. Note that, as learning about α occurs conditional on a technology choice, this framework allows different technologies to offer different

[5] Cp. Rust (1987), p. 1004-1005.

potentials for learning. Consumers, then, conditional on choosing technology $i^\star$, make continuous consumption (input) choices $\mathbf{w}_{i^\star}$ on the basis of $E\left[\tilde{U}\left(a_{i^\star}\left(\mathbf{w}_{i^\star},\alpha\right),x\right)|\theta,i^\star\right]$. And they make initial technology choices on the basis of $E\left[E\left[\tilde{U}\left(a_{i^\star}\left(\mathbf{w}_{i^\star},\alpha\right),x\right)|\theta,i^\star\right]\right]$, $i=1,\ldots,m$, where the outer expectation is taken with respect to the marginal distribution of θ. Notice that consumers' preferences are ex ante stochastic through the dependence on the random signal θ.

Two specific features of expected utility, conditional on θ and service i, emerge. First, since the marginal productivity of inputs $\mathbf{w}_i$ depends on α, the marginal expected utility of $\mathbf{w}_i$, conditional on θ, depends on θ through the moments of the posterior $\pi(\alpha|\theta,i)$. Second, it follows from the assumed concavity of $\tilde{U}(a_i(\mathbf{w}_i,\alpha))$ with respect to α and a second-order Taylor's series expansion of $\tilde{U}$ about the conditional mean of α that $\tilde{U}$ is inversely related to the posterior variance of α. So learning will be valuable when the posterior distribution is more precise than the prior. The posterior will have lower variance if the precision of the prior on α is higher and if the precision of the signal itself is higher, i.e. if the conditional variance of θ, given α, is lower. Determining the precision of the signal requires some effort on the part of the consumer who receives the signal. This effort is time-consuming. Let t_i denote the time allocated to (intellectually) process the signal, and suppose that the conditional variance of the signal θ, given α, is decreasing in t_i. This means that processing effort and signal precision are positively related. Therefore, $\tilde{U}$ is increasing in t_i. Since the posterior variance of α, given θ, is higher the higher the prior variance of α and the higher the conditional variance of the signal θ, given α, the marginal utility of t_i depends positively on these variance parameters. Specifically, the more precise the prior on α, the lower the marginal valuation of time, ceteris paribus. The class of utility models exhibiting these features is henceforth referred to as "reduced-form utility" models, since the higher-level service consumption is only indirectly modeled, via the choice of technology i and augmented production inputs $(\mathbf{w}_i',t_i)'$.

2.4 A "Reduced-Form Utility" Model Specification for Discrete-Continuous Choices

For the analysis of INDEX data, identify a service i with its nominal bandwidth b. Let $\mathbf{w}_b=v$, where subscripts on v and t are henceforth omitted for notational convenience. Suppose the relationship between latent α and signal θ is linear; this is implied, for instance, if α and θ are jointly normal. And assume that the prior variance as well as the conditional variance of the signal are both proportional to a variance parameter σ^2. Then the marginal variance of the signal θ is proportional to σ^2 and the marginal utility of t is increasing in σ^2. In the appendix, it is shown how specific functional forms for utility $\tilde{U}$ and service production technologies a_b, together with assumed joint normality of α and θ, yield the "reduced-form" preference

specification

$$\begin{aligned} E\left[\tilde{U}\left(a_b\left(v,\alpha\right),x\right)|\theta,b\right] &= U(v,t,x,b;\theta,\epsilon,\zeta_b) \\ &= e^{\epsilon_1+\theta}\ln(v)+\sigma_\theta^2\ln(t)+e^{\epsilon_2}\ln(x)+\zeta_b, \end{aligned}$$

where θ is normally distributed with variance $\sigma_\theta^2=\sigma^2$, $\epsilon=(\epsilon_1,\epsilon_2)'$ are parameters in $\tilde{U}$ and ζ_b is a paremeter in a_b, which are known to the consumer. This model resides within the theoretical framework of the preceding subsection. The dependence of the marginal utiltiy of input v on the signal θ is a consequence of the dependence of the marginal productivity of v on α in the specification for the technology $a_b(v,\alpha)$.[6] The ex ante random parameter θ gives rise to preferences which are stochastic from the user's ex ante perspective. Its variation determines the degree to which the user's on-line service valuation, conditional on a bandwidth choice, differ from the user's ex-ante anticipated service valuation. And the dependence of the marginal utility of t on σ^2 arises from assumptions about the joint distribution of (α,θ) and the concavity of $\tilde{U}(a_b(v,\alpha))$ in α. Thus, the ex ante uncertainty that the consumer faces determines the marginal valuation of convenience or intellectual processing time t.[7] In this specification, the ex ante marginal rate of substitution between volume and convenience time depends on the degree of ex ante uncertainty σ_θ^2. This implies that the user ex ante anticipates to consume some amount of convenience time. The ex post marginal rate of substitution between volume and convenience time depends in addition on θ. This implies that the revealed quality of information also determines the amount of convenience time that is actually consumed.

The utility function depends on a bandwidth-specific shift parameter ζ_b through the Internet service production technology. It can rationalize situations in which a user chooses a bandwidth and thereby associated prices for which the opportunity set, at each expenditure level, is dominated by the opportunity set of another bandwidth. The parameters ζ_b can, thus, be viewed as explaining what ex post appears as optimization error.

Suppose the time-line of the choice problem is as follows. The user first chooses a bandwidth, in ignorance about θ, and is subsequently committed to it. Then, the user starts data transmission. In the process of data transmission, θ is revealed to the user. Given θ, the user finally determines the values of v and t. The user solves this choice problem recursively. The solution algorithm is assumed to consist of the following steps. The user forms expectations about θ and computes the expected utility function for each bandwidth. For each bandwidth b with associated prices $\mathbf{p}(b)'=(p_v(b),p_t(b))$, the user maximizes expected utility over the associated budget constraint. Thus the user computes the indirect utility function of the expected utility maximization problem for each capacity choice alternative. The user then makes a capacity choice, choosing the capacity with maximal indirect expected utility. Subsequently, being committed to the chosen bandwidth, θ is revealed in the process of data transmission, and the user chooses total volume and convenience time for this connection.

[6] The significance of the parameters $\epsilon=(\epsilon_1,\epsilon_2)'$ will be clarified below.

[7] It is straightforward to check, on the basis of the formulae for ex ante anticipate and ex post realized demand functions, that these are continuous in σ_θ^2 at $\sigma_\theta^2=0$. Thus, the model does not exhibit any irregularities at the boundary of the parameter space.

Note that the assumed solution algorithm is a heuristic procedure for the user. It allows the user to circumvent the analytical intractability of the solution to the full dynamic program. Computing the expected value function of the random utility maximization problem is analytically intractable due to the nonlinearity of the model in θ. Instead, in the heuristic procedure, the user computes the value function of the expected utility maximization problem. In the former case, in the backward recursive solution, maximization precedes integration at the first stage of the solution algorithm, while in the heuristic procedure integration precedes maximization. Analytically intractable solutions for the value functions in the former case are computationally expensive and as such not plausible as decision criteria for the user. Therefore, the heuristic approach is adopted. A condition under which the heuristic yields the same choices as the full dynamic program is given below. It will also be seen how the heuristic approach can explain observed choices that ex post seem suboptimal.

Formally, the solution algorithm of the user's heuristic proceeds as follows. Suppose that θ is normally distributed with mean μ and variance σ_θ^2. First, for any bandwidth b, anticipated continuous choices are derived as

$$(\hat{v}, \hat{t}, \hat{x}) = \arg \max_{(v,t,x)\in\mathbb{R}^3_{++}} \left\{ E_\theta \left[U(v, t, x, b; \theta, \epsilon, \zeta_b) \right] : p_x x + p_v(b)v + p_t(b)t = M \right\},$$

where M denotes exogenous total outlays. The solutions to this problem are the anticipated continuous choices

$$\begin{aligned}
\hat{v} &= \hat{\nu}(b, \epsilon) = \frac{M}{p_v(b) + p_t(b)/b} \left(\frac{e_1^\epsilon + \mu \frac{1}{2}\sigma_\theta^2}{\sigma_\theta^2 + e^{\epsilon_2} + e^{\epsilon_1 + \mu + \frac{1}{2}\sigma_\theta^2}} \right) \\
\hat{t} &= \hat{\tau}(b, \epsilon) = \frac{M}{p_t(b)} \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + e^{\epsilon_2} + e^{\epsilon_1 + \mu + \frac{1}{2}\sigma_\theta^2}} \right) \\
\hat{x} &= \hat{x}(\epsilon) = \frac{M}{p_x} \left(\frac{e^{\epsilon_2}}{\sigma_\theta^2 + e^{\epsilon_2} + e^{\epsilon_1 + \mu + \frac{1}{2}\sigma_\theta^2}} \right),
\end{aligned}$$

where $\epsilon' = (\epsilon_1, \epsilon_2)$. Note that $\hat{x}$ does not depend on b, as a consequence of the assumption that utility is separable in x. Evaluating the expected utility function at $\hat{\nu}(b, \epsilon)$, $\hat{\tau}(b, \epsilon)$ and $\hat{x}(\epsilon)$ yields the indirect utility function of the expected utility maximization problem for b,

$$V(b, \epsilon) = E_\theta \left[U(\hat{\nu}(b, \epsilon), \hat{\tau}(b, \epsilon), \hat{x}(\epsilon), b; \theta, \epsilon, \zeta_b) \right].$$

The observed bandwidth capacity choice then is

$$\begin{aligned}
b &= \arg \max_{\{b_i\}_{i=1}^m} \{V(b_i, \epsilon)\} \\
&= \arg \max_{\{b_i\}_{i=1}^m} \left\{ e^{\epsilon_1 + \mu + \frac{1}{2}\sigma_\theta^2} \ln(\hat{\nu}(b_i, \epsilon)) + \sigma_\theta^2 \ln(\hat{\tau}(b_i, \epsilon)) + \zeta_{b_i} \right\}
\end{aligned}$$

Once the user is committed to b and starts data transmission, θ is revealed. Continuous volume

and convenience time choices, given b, are then obtained from the realized demand functions

$$\begin{aligned} v &= \nu(b,\theta,\epsilon) = \frac{M}{p_v(b)+p_t(b)/b}\left(\frac{e^{\epsilon_1+\theta}}{\sigma_\theta^2+e^{\epsilon_2}+e^{\epsilon_1+\theta}}\right) \\ t &= \tau(b,\theta,\epsilon) = \frac{M}{p_t(b)}\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+e^{\epsilon_2}+e^{\epsilon_1+\theta}}\right). \end{aligned}$$

At this point, it may be worth comparing the user's heuristic decision procedure to the solution of the full dynamic program. The heuristic employs the functions $\hat{\nu}(b,\epsilon)$ and $\hat{\tau}(b,\epsilon)$ which are referred to as anticipated demand functions, rather than expected demand functions. The latter term is reserved for the expectation of the actually ex post realized demand functions $\nu(b,\theta,\epsilon)$ and $\tau(b,\theta,\epsilon)$ with respect to θ. Due to the nonlinearity of the realized demand functions in the stochastic component θ, expected and anticipated demand functions do not coincide in general. In fact, one has the following

Result:

$$\begin{aligned} & E_\theta[\nu(b,\theta,\epsilon)] \geq \hat{\nu}(b,\epsilon) \\ \textit{and} \quad & E_\theta[\tau(b,\theta,\epsilon)] \leq \hat{\tau}(b,\epsilon) \\ \textit{if and only if} \quad & \mu \leq \ln(\sigma_\theta^2+e^{\epsilon_2})-\epsilon_1. \end{aligned} \tag{2-1}$$

A proof of this result is given in an appendix. If the last weak inequality holds with equality for a given ϵ, the heuristic decision procedure and full dynamic programming solution yield the same discrete choices for the user. Otherwise, the two approaches differ.

Their difference can be interpreted in the following manner. Anticipated convenience time $\hat{\tau}$ can be viewed as a measure for the size of the underlying of an option[8] on convenience time which the user ex ante desires to hold. The size of the underlying of this option depends on the expectation of the valuation parameter θ, μ, and on its variance σ_θ^2. For given expectations μ about the valuation parameter, ex ante uncertainty σ_θ^2 exceeding the benchmark $e^{\mu+\epsilon_1}-e^{\epsilon_2}$ leads the user to ex ante desire a convenience time option that provides excess insurance, since the size of the option is larger than what the user ends up consuming on average. In this case, ceteris paribus the user is likely to choose a low bandwidth in which idle time is cheap. Since the user on average ends up consuming less idle time than anticipated, ex post it might appear that a higher bandwidth would have been superior, both in terms of time and pecuniary cost. Similarly, if σ_θ^2 is relatively low, ceteris paribus the user ex ante desires an option for volume transmission that provides excess insurance when compared to the average byte volume actually transmitted. In this case, a lower bandwidth choice is likely to be less expensive ex post. Furthermore, it is easy to show that the probability of ex post seemingly suboptimal bandwidth choices is higher under the heuristic than under the full dynamic programming algorithm. Also, under the full dynamic programming approach, ex post seemingly suboptimal bandwidth choices should on average occur symmetrically. Given any bandwidth choice, the risk of it ex post being too low should balance the risk of it being too high. Under the heuristic,

[8]The underlying of an option is the asset that the holder of the (call) option is entitled to if the option is exercised.

these risks may differ, depending on the type of option the user desires ex ante. The exploratory analysis of INDEX user data summarized in section 4.1 below provides evidence of asymmetric risk. It thus lends support to the assumption that users make bandwidth choices on the basis of the type of option they desire. The assumption of users employing the heuristic approach thus gives some suggestion how ex post seemingly suboptimal choices arise from the decision process.[9]

This completes the outline of the model structure. Before proceeding to the econometric version of the model, the vector of covariates $\mathbf{z}'$ is included in the model. These covariates are assumed exogenous to the user's choices. They enter the model through a parametric function $f(\mathbf{z}) = f(\mathbf{z};\xi)$, for ξ a vector of parameters. This function is assumed additive to ϵ_1 and therefore impacts both discrete and continuous choices.

2.5 The Econometric Model

The econometrician does not possess the same information about the user's preferences as the user herself. Specifically, the utility function is only known to the analyst up to a vector of parameters. Some of these are assumed fixed across connections, while others are allowed to vary. Specifically, the following assumptions are maintained. The analyst does not observe θ, the vector $\epsilon' = (\epsilon_1, \epsilon_2)$ and the bandwidth-specific shift parameters $\zeta' = [\zeta_{b_i}]_{i=1,\ldots,m}$. Moreover, the analyst knows that the marginal distribution of θ is normal with mean $\mu = 0$, but its marginal variance σ_θ^2, known to the user, is an unknown constant. The function $f(\mathbf{z};\xi)$ is known up to the fixed parameter vector ξ. The unknown parameters ϵ, θ and ζ are allowed to differ across connections for the user. The econometric model therefore allows these model parameters to be stochastic. These assumptions give rise to the econometric version of the stochastic preference model for the user.

The econometric model maintains a number of distributional assumptions. The econometric error terms ϵ and ζ represent unobserved heterogeneity in ex ante valuations. Its component ϵ_1 may arise for instance from applications the user has planned and that are unobserved in the data for analysis. It is interpreted as unobserved variation in anticipated valuations of byte volume. As such, it may be correlated with the unanticipated component in valuations, θ. The remaining components of the econometric error terms, ϵ_2 and ζ, are assumed to be

[9] To test the validity of the assumption that users employ the described heuristic approach, one might proceed more formally as follows. The values of σ_θ^2 and the mean of θ, given the econometric error, can be estimated, as will be shown in the section on estimation. Suppose ϵ was known to the analyst. Then, the relationships in $(2-1)$ could be checked by determining the direction of the third inequality and evaluating the functions $\hat{\nu}$ and $\hat{\tau}$ and simulating the expected choices, given the estimate of σ_θ^2 and ϵ. Checking the relationships for all observations, if the number of violations of the relationships relative to the number of cases for which they are satisfied is large - in some statistical sense -, then the validity of the assumption would have to be called into question. The econometric model, outlined in the following section, assumes that ϵ is unknown to the analyst. In the section on the likelihood function, it is demonstrated, however, how the model can essentially be inverted so as to uncover or estimate ϵ.

independently distributed. The econometric model then postulates that the vector (ϵ', θ) has a multivariate normal distribution, with means zero, variances $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}$ and σ^2_θ and covariance $\sigma_{\theta\epsilon_1}$ between ϵ_1 and θ; all other covariances are zero. The idiosyncratic errors ζ_{b_i}, $i = 1, \ldots, m$, are assumed to be identically extreme value distributed, mutually independent and independent of ϵ and θ. All stochastic components are assumed independent across connections. Under these distributional assumptions, all expectations with respect to θ in the previous section are to be re-interpreted in the econometric model as being conditional on ϵ_1.

Issues concerning the justification and various limitations of the econometric model are discussed in an appendix. At this point, three model assumptions should perhaps be commented on here. Since in the application of this model to INDEX data total expenditure M is not observable, the assumption of separability of x and independence of ϵ_2 are essential to estimate the model parameters. Estimation proceeds conditional on observed expenditure on byte volume and duration. This renders $\sigma^2_{\epsilon_2}$ unidentifiable, given the data. A second comment concerns the normality assumption. Normality of θ is convenient since, conditional on ϵ, the expectation of U with respect to θ exists. Normality is not essential, however. The random component θ can have any distribution, as long as its conditional moment generating function $M_{\theta|\epsilon}(t) = E_{\theta|\epsilon}[e^{t\theta}]$ exists at $t = 1$. In the case of normality, this requirement is fulfilled since the conditional moment generating function of θ exists for all real and finite t. There exist cases, however, when the conditional expected utility function fails to exist because this requirement is not met.[10] Finally, the assumption of the idiosyncratic error ζ_b being independently distributed may be questionable. Loosely speaking, ζ_b represent the ex ante utility of nominal bandwidth b. Since the congestion of the network determines the actually delivered transmission speed, the utility of b may well have to be re-assessed on-line. The delivered transmission speed depends on applications such as ftp or web traffic. These are unobserved by the analyst and captured by the econometric error ϵ_1. This suggests that a more elaborate analysis might allow for correlations between idiosyncratic errors ζ_b and ϵ_1 as well as possibly θ. This, however, increases the number of parameters by $2m$. In terms of the estimation methodology outlined below, it is clear how to estimate such a larger model, but this computationally more expensive task is left for future work.

The next section describes how measurements of byte volume v, convenience time t and bandwidth b can be used to identify and estimate the proposed structural model from the reduced form.

[10] Consider, for instance, the case in which $\theta \sim \exp(\lambda)$, $\lambda > 0$, independent of ϵ and ζ_b. Then, $M_\theta(t) = \lambda/(\lambda - t)$, provided $t < \lambda$. Therefore, the expected utility function does not exist if $\lambda \leq 1$.

3 Estimation

3.1 The Likelihood Function

Denote the induced distribution of the observations (v,t,b) by $f(v,t,b)$. Then, suppressing prices $\mathbf{p}$, covariates $\mathbf{z}$, ξ and the vector of distributional parameters $\Sigma' = (\sigma^2_\theta, \sigma^2_{\epsilon_1}, \sigma_{\theta\epsilon_1}, \sigma^2_{\epsilon_2})'$ for notational simplicity,

$$\begin{aligned} f(v,t,b) &= f(v,t;b)\Pr(b) \\ &= \int_{\epsilon_1} f(v,t;b,\epsilon_1)\Pr(b|\epsilon_1)\frac{1}{\sigma_{\epsilon_1}}\phi\left(\frac{\epsilon_1}{\sigma_{\epsilon_1}}\right)d\epsilon_1, \end{aligned}$$

where the conditional choice probabilities are of the conditional logit form

$$\Pr(b|\epsilon_1) = \frac{e^{-\left(e^{\epsilon_1+\mu_{\theta|\epsilon_1}+\frac{1}{2}\sigma^2_{\theta|\epsilon_1}}\ln(p_v(b)+p_t(b)/b)+\sigma^2_\theta\ln(p_t(b))\right)/\lambda}}{\sum_{i=1}^m e^{-\left(e^{\epsilon_1+\mu_{\theta|\epsilon_1}+\frac{1}{2}\sigma^2_{\theta|\epsilon_1}}\ln(p_v(b_i)+p_t(b_i)/b_i)+\sigma^2_\theta\ln(p_t(b_i))\right)/\lambda}}, \tag{3-2}$$

where $\lambda > 0$ is the scale parameter of the extreme value distribution.

The conditional density of the continuous choice variables, given the discrete choice and ϵ_1, can only be expressed symbolically, due to the nonlinearity of the stochastic demand functions in the stochastic components. Define

$$\begin{aligned} h(b,\theta,\epsilon) &:= \begin{bmatrix} \nu(b,\theta,\epsilon) \\ \tau(b,\theta,\epsilon) \end{bmatrix} \qquad \text{and} \\ g(v,t;b,\epsilon_1) &:= h^{-1}_{(\theta,\epsilon_2)'}(b,\theta,\epsilon) \\ &= \begin{bmatrix} \theta \\ \epsilon_2 \end{bmatrix}. \end{aligned}$$

It follows from Beckert (1999), Lemma 1, that the Jacobian of h with respect to θ and ϵ_2 exists and has full rank with probability one. Therefore, the inverse function $g(v,t;b,\epsilon_1)$ exists. Moreover, the lemma implies that the distribution of v and t, conditional on b and ϵ_1 is non-degenerate on the hyper-rectangle $[0, M/(p_v(b)+p_t(b)/b)] \times [0, M/p_t(b)] \subset \mathbb{R}^2_{++}$. Thus, the econometric model is such that it does not implicitly a priori restrict choice behavior.

Now the conditional density $f(\nu,\tau;b,\epsilon_1)$ can be expressed as

$$\begin{aligned} f(v,t;b,\epsilon_1) &= \left|\begin{bmatrix} \sigma^2_{\theta|\epsilon_1} & 0 \\ 0 & \sigma^2_{\epsilon_2} \end{bmatrix}^{-\frac{1}{2}}\right| \phi\left(\begin{bmatrix} \sigma^2_{\theta|\epsilon_1} & 0 \\ 0 & \sigma^2_{\epsilon_2} \end{bmatrix}^{-\frac{1}{2}} g(v,t;b,\epsilon_1)\right) \left|\left[\nabla_{(\nu,\tau)'}g(v,t;b,\epsilon_1)\right]^{-1}\right| \\ &= \frac{1}{\sigma_{\theta|\epsilon_1}\sigma_{\epsilon_1}}\phi\left(\frac{1}{\sigma_{\theta|\epsilon_1}\sigma_{\epsilon_1}}g(v,t;b,\epsilon_1)\right)\left|\left[\nabla_{(\nu,\tau)'}g(v,t;b,\epsilon_1)\right]^{-1}\right|. \end{aligned}$$

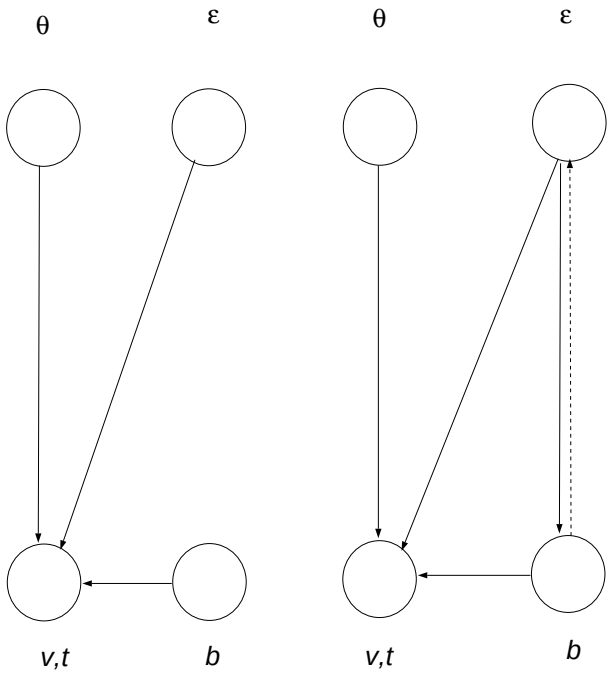

Figure 2: Stochastic Demand Models — Continuous choice model and discrete/continuous choice model

Bayes' Theorem can be used to symbolically express the likelihood function of the parameters, given the observable continuous choice variables v and t and given the discrete choice b, in two useful ways. These are

$$\begin{aligned} f(v,t,b) &= f(v,t;b)\Pr(b) \\ &= E_{\epsilon_1}[f(v,t;b,\epsilon_1)\Pr(b|\epsilon_1)] \\ &= E_{\epsilon_1|b}[f(v,t;b,\epsilon_1)]\Pr(b). \end{aligned} \tag{3-3}$$

The decomposition ($3-3$) of the likelihood of the data suggests the following three insights: With a model for the discrete choice of b, given ϵ_1, using Bayes' Theorem, it is possible to separate the variation in ϵ_1 from the variation in θ and therefore to identify σ_θ^2 separately from the variation in the econometric errors. Secondly, in general, the distribution of $\epsilon_1|b$ has $\text{supp}(\epsilon_1|b) \subset \text{supp}(\epsilon_1)$. Therefore, it also follows that $\text{supp}(\theta|b,\epsilon_1) \subset \text{supp}(\theta)$. The information embedded in the discrete choice of b reduces the uncertainty in the econometric error ϵ_1, and this informational gain spills over to reduce the uncertainty about θ. It thereby allows more efficient estimation of the parameters of interest. Figure 2 graphically displays the model for the continuous choices, with and without a model for the discrete choice; the dashed line in the discrete-continuous model corresponds to the application of Bayes' Theorem. Finally, the alternative decompositions of the likelihood as $E_{\epsilon_1|b}[f(v,t;b,\epsilon_1)]\Pr(b)$ and $E_{\epsilon_1}[f(v,t;,,\epsilon_1)P(b|\epsilon_1)]$ suggest two ways of simulating the likelihood in the unmodified model - the first based on sampling from a conditional distribution and the second sampling from the marginal distribution -

that could be used to simulate the likelihood if the conditional density $f(v,t;b\epsilon_1)$ were analytically tractable. Because of its intractability, however, the model in general must be estimated from its conditional moments. That these can be simulated using the same principles will be shown after a brief excursion on identification.

3.2 Identification

The goal of this section is to argue that the parameters of the random utility model, in particular, the covariance matrix of $(\epsilon', \theta)'$, can be identified from the observed discrete and continuous choices. Intuitively, two features of the experiment are instrumental for identification. First, INDEX is a randomized experiment, so that there is no concern that covariates like prices are endogenous. Second, every user makes a series of sequential discrete and continuous choices. This allows to compare the initial discrete bandwidth choice with the ex post optimal bandwidth choice, given the observed continuous choices. Discrepancies, i.e. ex post apparent optimization errors, identify changes in tastes as a consequence of the consumption experience. This permits calibration of σ_θ^2. The discussion of estimation results in section 4.2 uses this logic to form heuristics in order to validate the estimates.

In the remainder of this subsection, the question of identification is addressed formally. An impediment to a completely rigorous analysis is the fact that, due to the nonlinearity of the model, the density of ν and τ can only be expressed symbolically. It may therefore be helpful to look at the data from a number of different perspectives. Also, to make the discussion more transparent, the following, equivalent re-parameterization of the model is considered. Let $\zeta \sim N(0, \mathbf{I}_3)$. If $(\epsilon_1, \theta, \epsilon_2)' \sim N(0, \Sigma)$ and σ_{ij} denotes the ij element of Σ, then

$$\begin{aligned}
\begin{bmatrix} \epsilon_1 \\ \theta \\ \epsilon_2 \end{bmatrix} &= \Gamma \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{11}^{1/2} & 0 & 0 \\ \frac{\sigma_{12}}{\sigma_{11}^{1/2}} & \left(\sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)^{1/2} & 0 \\ \frac{\sigma_{13}}{\sigma_{11}^{1/2}} & \frac{\sigma_{23} - \frac{\sigma_{12}\sigma_{13}}{\sigma_{11}}}{\left(\sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)^{1/2}} & \left(\sigma_{33} - \frac{\sigma_{13}}{\sigma_{11}^{1/2}} - \frac{\sigma_{23} - \frac{\sigma_{12}\sigma_{13}}{\sigma_{11}}}{\left(\sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)^{1/2}}\right)^{1/2} \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix}
\end{aligned} \tag{3-4}$$

produces a model with equivalent stochastic features, since $\Sigma = \Gamma\Gamma'$. The question of identification centers around the σ_{ij}'s and the suppressed parameters in $f(\mathbf{z})$. Assume, as above, that ϵ_2 is uncorrelated - and under normality independent - of ϵ_1 and θ, which amounts to assuming that $\sigma_{13} = \sigma_{23} = 0$, so that $\Gamma_{31} = \Gamma_{33} = 0$.

First consider the discrete choice problem. Separability of the outside good x in the random

utility function implies that

$$b = \arg \max_{\{b_i\}_{i=1}^m} E_{\zeta_2} \left[\sigma_{22}^2 e^{\Gamma_{1.}\zeta + \Gamma_{2.}\zeta + f(\mathbf{z})} \ln\left(\nu\tau(b_i)\right) + \ln(\hat{\tau}(b_i)) + \zeta_\beta \right],$$

for the abbreviated notation $\hat{\nu}(b_i) = \hat{\nu}(b_i, \epsilon)$ and similarly for $\hat{\tau}(b_i)$. Under the re-parameterization in this section, ($3-2$) is given by

$$\Pr(b|\zeta_1) = \frac{e^{-\left(e^{(\Gamma_{11}+\Gamma_{21})\zeta_1 + \frac{1}{2}\Gamma_{22}} \ln(p_v(b)+p_t(b)/b) + \sigma_\theta^2 \ln(p_t(b))\right)}}{\sum_{i=1}^m e^{-\left(e^{(\Gamma_{11}+\Gamma_{21})\zeta_1 + \frac{1}{2}\Gamma_{22}} \ln(p_v(b_i)+p_t(b_i)/b_i) + \sigma_\theta^2 \ln(p_t(b_i))\right)}}$$

Notice the normalization $\lambda = 1$. Since different shift and scale parameters of the i.i.d. extreme value terms ζ_b yield affine transformations of the corresponding expected utility functions, i.e. different representatives of an equivalence class, these parameters are normalized to zero and one. From these conditional logit choice probabilities, $\sigma_\theta^2 = \sigma_{22}$ is identified, and $\Gamma_{11} + \Gamma_{21}$ and Γ_{22} are identified up to scale. This allows identification of $\frac{\sigma_{12}^2}{\sigma_{11}}$ up to scale.

Now turn to the expressions for the stochastic demand functions ($2-1$), where ϵ_1, θ and ϵ_2 are expressed according to ($3-4$). That is, denoting the ith row of Γ by $\Gamma_{i.}$,

$$\begin{aligned} v &= \nu(b,\zeta) \\ &= \frac{M}{p_v(b) + \frac{p_t(b)}{b}} \left(\frac{1}{1 + e^{\Gamma_{3.}\zeta + \sigma_{22}^2 e^{\Gamma_{1.}\zeta + \Gamma_{2.}\zeta + f(\mathbf{z})}}} \right) \\ &= \frac{M}{p_v(b) + \frac{p_t(b)}{b}} \left(\frac{1}{1 + e^{\Gamma_{3.}\zeta + \sigma_{22}^2 e^{(\Gamma_{11}+\Gamma_{21})\zeta_1 + \Gamma_{22}\zeta_2 + f(\mathbf{z})}}} \right), \end{aligned}$$

and an analogous expression can be obtained for $\tau(b,\zeta)$. The scale of $\Gamma_{11}+\Gamma_{21}$ and Γ_{22} is identified from the conditional first moments of the continuous choice variables. Given the functions of the covariance parameters identified from the discrete choice analysis, ζ_1 can be sampled and viewed as an observation, imputed by sampling from its conditional distribution, given b. Given the imputed observation ζ_1, the bivariate system $(\nu, \tau)'$ has a nonsingular distribution on $\mathbb{R}_+^2$ which is induced by the distribution of $(\Gamma_{22}\zeta_2, \Gamma_{32}\zeta_2 + \Gamma_{33}\zeta_3)'$, which has variance parameters Γ_{22}^2 and $\Gamma_{32}^2 + \Gamma_{33}^2$, respectively, and covariance parameter $\Gamma_{22}\Gamma_{32} = \sigma_{23} - \frac{\sigma_{12}\sigma_{13}}{\sigma_{11}}$; under the assumption of $(\epsilon_1, \theta)'$ and ϵ_2 being independent, this covariance is zero. This bivariate system obeys all the requirements of known results on identification in stochastic demand systems that arise from stochastic preferences (Beckert (1999), Proposition 1), so that these three functions of parameters are identified. Then, given Γ_{22}, σ_{22}, σ_{11} and σ_{12} are identifiable from the identified parameter functions obtained through the discrete choice analysis. The covariance parameter of the bivariate system then identifies σ_{23}, and $\Gamma_{32}^2 + \Gamma_{33}^2$ finally permits identification of σ_{33}. Thus, all the covariance parameters are identifiable in this model.

3.3 Estimation by Simulation

There are in principle two possibilities to estimate the proposed model. The most efficient estimation methodology would proceed by standard maximum likelihood estimation. Due to the nonlinearity of the relationship between endogenous variables and the stochastic model components, this approach, while conceptually appealing, is impractical, for essentially two reasons. First, as shown above, the likelihood function is non-standard since it involves the determinant of the inverse of the Jacobian of the nonlinear transformation, which in most cases is analytically intractable. If the determinant of the inverse Jacobian could be dealt with easily, then maximum simulated likelihood estimation would be feasible. This approach is described first. An alternative route is to estimate the model parameters from conditional and unconditional moments. In this case, again due to model nonlinearity, simulated analogues of the analytically intractable theoretical moments must be used to estimate the model by the method of simulated moments, initially proposed by McFadden (1989) and Pakes and Pollard (1989). This approach is described in the second place and carried out in the empirical analysis.

3.3.1 Maximum Simulated Likelihood

Recall from ($3-3$) that

$$f(v,t,b) = E_{\epsilon_1}[f(v,t;b,\epsilon_1)\Pr(b|\epsilon_1)].$$

This suggests the possibility of estimating the model by maximizing the simulated likelihood function, if this function were analytically tractable. It typically is not, and so the procedure described in this subsection is given for completeness only. Specifically, consider simulating the unobservable endogenous variables $\hat{v}$ and $\hat{t}$, as well as the observable discrete choice variable b. Given values for the parameters Σ, draw $\epsilon_1^\star$ from its marginal distribution. For the observed capacity choice b in the set of possible capacity choices, compute simulated conditional logit choice probabilities, conditional on $\epsilon_1^\star$, and approximate $E_{\epsilon_1}[f(v,t;b,\epsilon_1)\Pr(b|\epsilon_1)]$ by its simulated analogue,

$$E_{\epsilon_1^\star}^T\left[f(v,t;b,\epsilon_1^\star)\Pr(b|\epsilon_1^\star)\right] = \frac{1}{T}\sum_{t=1}^{T}\left(f(v,t;b,\epsilon_{1,t}^\star)P(b|\epsilon_{1,t}^\star)\right),$$

for T draws $\{\epsilon_t^\star\}_{t=1}^T$ from $f(\epsilon)$. To benefit from the added information about ϵ_1 embedded in the observed discrete choices, one could draw more than T values from the marginal distribution of ϵ_1 and just keep those T draws that maximize the discrete choice probability of the observed choice b. For a sequence of observations $(\mathbf{v},\mathbf{t},\mathbf{b})' = \{v_s,t_s,b_s\}_{s=1}^S$, assuming, for now, serial independence of $\theta_s,\epsilon_{1s},\epsilon_{2s}$, the simulated likelihood is

$$f^\star(\mathbf{v},\mathbf{t},\mathbf{b};\xi,\Sigma) = \prod_{s=1}^{S} E_{\epsilon_{1,s}^\star}^T\left[f(v_s,t_s;b_s,\epsilon_{1s_t}^\star)\Pr(b|\epsilon_{1s_t}^\star)\right],$$

where $\epsilon^\star$ are $S \times T$ draws from $f(\epsilon)$. Then, the maximum simulated likelihood estimates $(\hat{\xi}_{MSLE}, \text{vec}(\hat{\Sigma}_{MSLE}))'$ are obtained as

$$(\hat{\xi}_{MSLE}, \text{vec}(\hat{\Sigma}_{MSLE}))' = \arg\max_{(\xi,\Sigma)} f^\star(\mathbf{v}, \mathbf{t}, \mathbf{b}; \xi, \Sigma).$$

3.3.2 Method of Simulated Moments

An alternative simulation assisted estimation methodology is based on comparing the moments of the distribution of simulated choices with the moments of the distribution of observed choices. To motivate this approach, observe that, suppressing covariates and parameters,

$$\begin{aligned}
E[\nu|b] &= \int_{\theta,\epsilon} \nu(b,\theta,\epsilon) f(\theta,\epsilon|b) d\theta d\epsilon \\
&= \int \nu(b,\theta,\epsilon) f(\theta|\epsilon,b) f(\epsilon|b) d\theta d\epsilon \\
&= \int \nu(b,\theta,\epsilon) f(\theta|\epsilon,b) \frac{\Pr(b,\epsilon)}{\Pr(b)} d\theta d\epsilon \\
&= \int \nu(b,\theta,\epsilon) f(\theta|\epsilon,b) \frac{\Pr(b|\epsilon_1)}{\Pr(b)} f(\epsilon) d\epsilon d\theta \\
&= \frac{1}{\Pr(b)} E_\epsilon \left[E_{\theta|\epsilon,b} \left[\nu(b,\theta,\epsilon) \right] \Pr(b|\epsilon_1) \right]. \qquad (3\text{-}5)
\end{aligned}$$

Moreover, from the discrete choices,

$$\Pr(b) = E_{\epsilon_1}[\Pr(b|\epsilon_1)]. \qquad (3\text{-}6)$$

Analogous expressions hold for $E[\tau|b]$. These again suggest to replace the expectations by their simulated counterparts. In both cases, all the moments of $(\epsilon', \theta)'$ are available as well.

The conditional discrete choice probabilities are directly simulated according to ($3-2$), with ϵ_1 sampled from its marginal distribution. Either all draws can be retained or, in a modified rejection step to benefit from the possibility to increase efficiency, simply those draws are retained that either produce $\Pr(b|\epsilon_1^\star) \geq \Pr(\beta|\epsilon_1^\star)$ for all β or give the highest simulated choice probability $\Pr(b|\epsilon_1^\star)$. Draw $\theta^\star$ from the distribution of θ, given a (retained) draw $\epsilon_1^\star$; in the case of joint normality and the assumption that θ and ϵ_2 are independent, this amounts to drawing from a normal distribution with mean $\frac{\sigma_{\theta\epsilon_1}}{\sigma^2_{\epsilon_1}}\epsilon_1^\star$ and variance $\sigma^2_\theta - \frac{\sigma^2_{\theta\epsilon_1}}{\sigma^4_{\epsilon_1}}$. From ($2-1$), for a sequence of $S \times T$ draws $\{\epsilon^\star_{s,t}, \theta^\star_{s,t}\}_{s=1,t=1}^{S,T}$, this yields conditional moments for the continuous choice variables, given $\mathbf{p}$ and M,

$$d^\star_{1s}(\xi,\Sigma) = \begin{bmatrix} v_s - \frac{1}{E^T_{\epsilon^\star_{1,s}}[P(b_s|\epsilon^\star_{1,s})]} E^T_{\epsilon^\star_s} \left[E^T_{\theta^\star_s|\epsilon^\star_{1,s},b} \left[\nu(b_s, \theta^\star_{s,t,k}, \epsilon^\star_{1s,t}, \epsilon^\star_{2s,t}) \right] P(b_s|\epsilon^\star_{1s,t}) \right] \\ t_s - \frac{1}{E^T_{\epsilon^\star_{1,s}}[P(b_s|\epsilon^\star_{1,s})]} E^T_{\epsilon^\star_s} \left[E^T_{\theta^\star_s|\epsilon^\star_{1,s},b} \left[\tau(b_s, \theta^\star_{s,t,k}, \epsilon^\star_{1s,t}, \epsilon^\star_{2s,t}) \right] P(b_s|\epsilon^\star_{1s,t}) \right] \end{bmatrix},$$

where the parameters ξ and Σ are suppressed on the right-hand side. Conditional moments of the discrete choice problem can be simulated for each capacity b_i in the choice set as

$$d^{\star}_{2s}(\xi, \Sigma) = 1_{\{b_s = b_i\}} - E^{T}_{\epsilon^{\star}_{1,s}}[P(b_i|\epsilon^{\star}_{1,s})].$$

Since ($3-5$) and ($3-6$) hold conditional on $\mathbf{p}$ and M, any function of $\mathbf{p}$ and M and possibly Σ and ξ can serve as additional instruments and be used to form further unconditional moment conditions, by taking iterated expectations. Denote the vector of all unconditional simulated moments by $D^{\star}_{s}(\xi, \Sigma)$. For a symmetric, positive definite weight matrix $\mathbf{Q}_S$, the length with respect to the metric $\mathbf{Q}_S$ of the average deviation of the observations and their simulated unconditional moments is $E_S\left[D^{\star}_{s}(\xi, \Sigma)\right]' \mathbf{Q}_S E_S\left[D^{\star}_{s}(\xi, \Sigma)\right]$. Convenient choices for $\mathbf{Q}_S$, in a first step of a two step feasible MSM estimation, are $\mathbf{Q}_S = \mathbf{I}_r$ or $\mathbf{Q}_S = E_S[\mathbf{W}_s \mathbf{W}_s']$, where r is the number of available moment conditions and $\mathbf{W}_s$ is an array of instruments used to form unconditional moments. Let

$$Q(\mathbf{v}, \mathbf{t}, \mathbf{b}; \epsilon^{\star}, \theta^{\star}, \xi, \Sigma, \mathbf{Q}_S) = E_S\left[D^{\star}_{s}(\xi, \Sigma)\right]' \mathbf{Q}_S E_S\left[D^{\star}_{s}(\xi, \Sigma)\right];$$

for identification purposes, the standard order condition for Generalized Method of Moments estimation requires, of course, that the vector $D^{\star}(\xi, \Sigma)$ have at least as many components as there are distinct elements in ξ and Σ. Then, the method of simulated moments estimates $(\hat{\xi}_{MSME}, \mathrm{vec}(\hat{\Sigma}_{MSME}))'$ are obtained as

$$(\hat{\xi}_{MSME}, \mathrm{vec}(\hat{\Sigma}_{MSME}))' = \arg\min_{\xi, \Sigma} Q(\mathbf{v}, \mathbf{t}, \mathbf{b}; \epsilon^{\star}, \theta^{\star}, \xi, \Sigma, \mathbf{Q}_S).$$

The following section describes the data and their analysis, both exploratory and in the context of the model, using the outlined estimation methodology to estimate the model parameters.

4 Data Analysis

4.1 A Brief Description and Exploratory Analysis of the Data

The first part of this section characterizes the data of a "representative" user out of the INDEX subject pool, while the second part presents estimation results for the model outlined above.

The data are collected when a connection between an INDEX host and the INDEX control gateway is established. Strictly speaking, it cannot be precluded that more than one person uses a host. More generally, the notion of user then refers to the person(s) using a host. Hosts are located in INDEX subjects' residences. While this still leaves the possibility of multiple users, it suggests the econometric framework laid out here as a reasonable approximation. For demand

management, of course, the number of users behind a host is immaterial if billing is based on traffic and capacity usage generated by a host. For the purpose of most of this anlaysis, the user data, available in minute–by–minute detail, are aggregated to individual connections to a bandwidth, as defined above. Metered usage corresponds to TCP connection-oriented protocol, which amounts to almost all traffic on the Internet. Discrete capacity choices are measured in terms of bandwidth (in $kbps$). With each such choice, both transmitted volume (inbound and outbound, in bytes) and duration (in seconds), the start second (relative to GMT) of this connection and the prevailing prices in the symmetric, variable bandwidth experiment are recorded. In this experiment, inbound and outbound volume are transmitted subject to the same capacity choice out of the menu $\{8kbps, 16kbps, 32kbps, 64kbps, 96kbps, 128kbps\}$, where $8kbps$ is free throughout the experiment, while the remaining capacities are priced in increasing order. Prices $p_t(b_i)$, in cents per minute, are drawn randomly from the interval $[0.1, 20.0]$. In this experiment, therefore, $p_v(b_i) = 0$ for all i; in subsequent experiments, for which data were not yet available when this anaylsis was conducted, volume based prices were positive. The symmetric, variable bandwidth experiment ran over six weeks. The first week was not priced, weeks 2 to 5 exhibited weekly price changes, while prices changed daily in the sixth week. Note that the pricing scheme is nonlinear and usage sensitive, where usage is measured in terms of subscription time and indexed by capacity.

4.1.1 Capacity Utilization and Convenience

This and the following subsections describe features of the data obtained from a particular subject. Most of these features reported here are also found for other subjects.

Define a session as a sequence of capacity choices by the user which is subject to payment and between two free states, i.e. either $0kbps$ or $8kbps$. The subject under investigation generated a total of 185 such session, 148 of which involved only a single capacity choice with positive price. So in about 80 percent of all sessions the subject decided to stick to a single, priced capacity, while in the remaining sessions the subject alternated between various priced capacities. This percentage, as most other findings reported here, are also characteristic for other subjects. Table 1 shows that, on average, higher volume and therefore more time actively used for transmission is associated with higher capacity choices. This suggests that discrete bandwidth capacity choices and subsequent continuous volume choices are indeed related choices, as the model stipulates.

capacity	$16kbps$	$32kbps$	$64kbps$	$96kbps$	$128kbs$
ave. volume(Mb)	0.37	0.73	0.35	0.99	1.60

Table 1: Capacity choices and average usage (priced)

Typically, users maintain a connection to a capacity for some time, even when they do

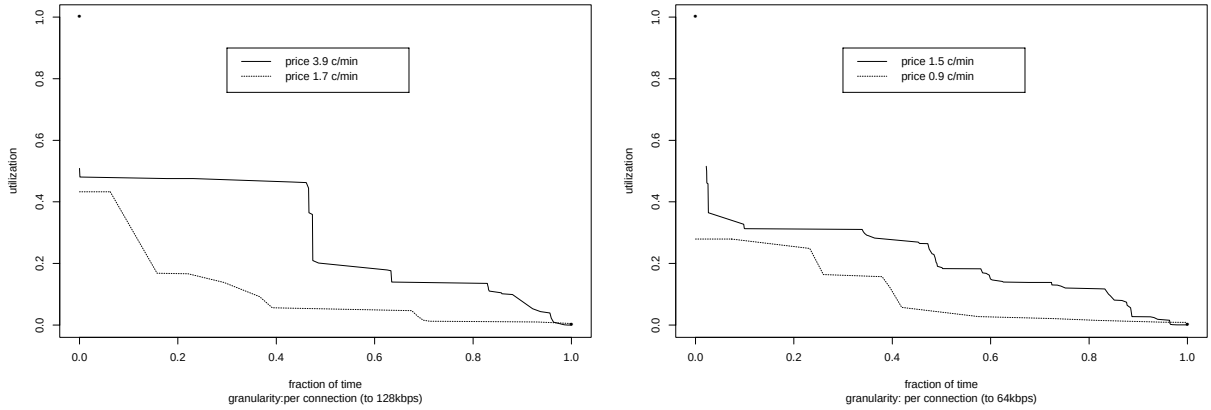

Figure 3: Utilization of 128*kbps* and 64*kbps*

not actively use it and even though this is costly under the time-based pricing scheme of this experiment. The reasons may be that, while processing the transmitted volume, users value having the option to continue traffic generation without disruption or that they experience some cost or disutility from disconnecting or switching to a less expensive bandwidth. This results in under-utilization of chosen capacity and reflects users' demand for convenience. Figure 3 presents load–duration curves for the representative user, measuring average utilization (in percent, per connection to 128*kbps*) against fraction of time, distinguishing utilization of a given capacity between different prices. When read from the vertical axis, the graphs, for any given utilization u, display the fraction of time during which utilization of 128*kbps* was at least as high as u. When read this way, the graphs can essentially be interpreted as 1 minus the empirical cumulative distribution function of the "random variable" utilization, having the unit interval as its support. The load–duration patterns are seen to be price sensitive; higher prices per unit time induce more conservative utilization, increasing the fraction of time during which a given utilization is achieved and thus reducing demand for convenience time and ensuing capacity under-utilization. This finding is robust in the sense that it holds for high capacities with relatively high prices, as well as for lower capacities with relatively low prices.

Figure 4 presents another look at essentially the same phenomenon: Here, rather than considering aggregate values per connection, the data is represented with per–minute resolution, displaying inbound capacity utilization for minutes connected to 128*kbps* and to 64*kbps*. Note that the overwhelming percentage of all transmission, 95.2 percent for this user, is inbound. This suggests the interpretation, that time in excess of the minimal time necessary for data transmission is intellectual processing time, used to assess the quality of information. Both graphs clearly show the effect of higher prices on convenience or intellectual processing time, i.e. time with no or very thinly distributed volume transmission. Convenience is reflected in transmission inactivity for up to forty percent of subscription time. It also emerges from the graphs that the user's decision whether or not to keep the option of such convenience – and if so, to what extent – is subject to prevailing prices. Higher prices for minutes of usage induce

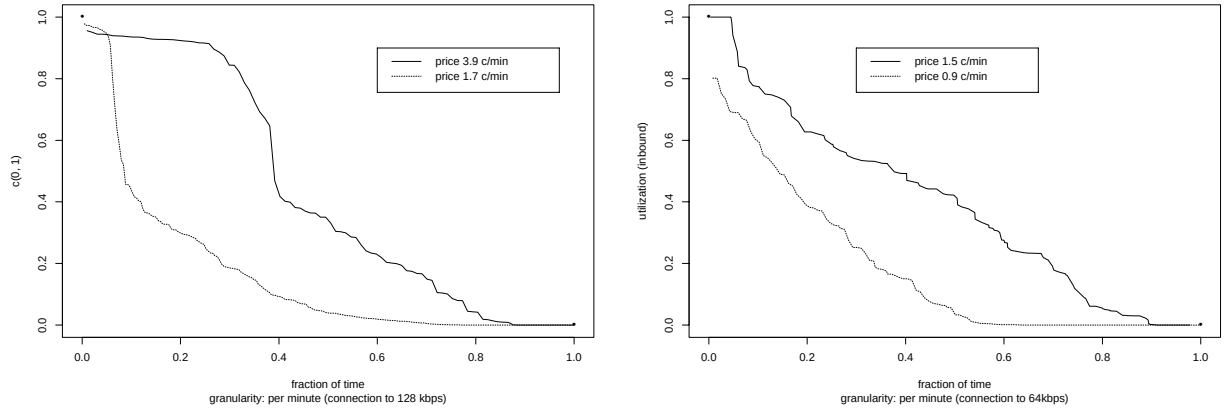

Figure 4: Utilization of 128*kbps* and 64*kbps*

more conservative capacity utilization, and this holds both for high as well as for low capacity levels and for prices in the higher as in the lower ranges.

Measuring convenience or intellectual processing time is not unproblematic, for essentially two reasons. The finest granularity of the data for analysis is per minute detail, which is a coarse resolution relative to some session durations. The second reason is that effective capacity utilization hinges on overall system conditions. While the INDEX setup can guarantee 100 percent of chosen capacity on the link to the Internet backbone, using a shared system implies for the user that realized utilization, speed and traffic flow depend on the overall contemporaneous competition among all users for capacity on every segment of the route that is assigned to the user's traffic. Therefore, even though INDEX can assure full capacity on the leg it provides to its users, a choice of any capacity may result in less effective capacity being delivered. This complicates the measurement of convenience time, since it implies that time actively used for user initiated traffic can substantially exceed v/b, when capacity b is chosen and volume v is requested for transmission.

Figures 5 and 6 display some of the representative user's sessions when 128*kbps* and 64*kbps* were chosen, both as cumulative utilization curves and as time profile, for inbound traffic. Again, convenience time clearly emerges, both at the beginning and end of a session as well as in the course of a session. Convenience time appears to be more prevalent when it is cheaper, as in the case of 64*kbps*, confirming the insights obtained from figure 4. Notice the variety of utilization patterns in the case of 128*kbps*, which suggests that different applications were being pursued in different sessions.[11] The following broad classification of applications seems plausible. High utilization, above 90 percent, say, suggests bulk traffic, e.g. ftp traffic. Medium range utilization suggests web traffic. In web traffic, it is possible to get high utilization for short time intervals, typically a couple of seconds; when averaged over a minute, yields medium

[11]Information about types of application is, in principle, available in the INDEX data, but not accessible to the analyst at the time when this analysis is conducted.

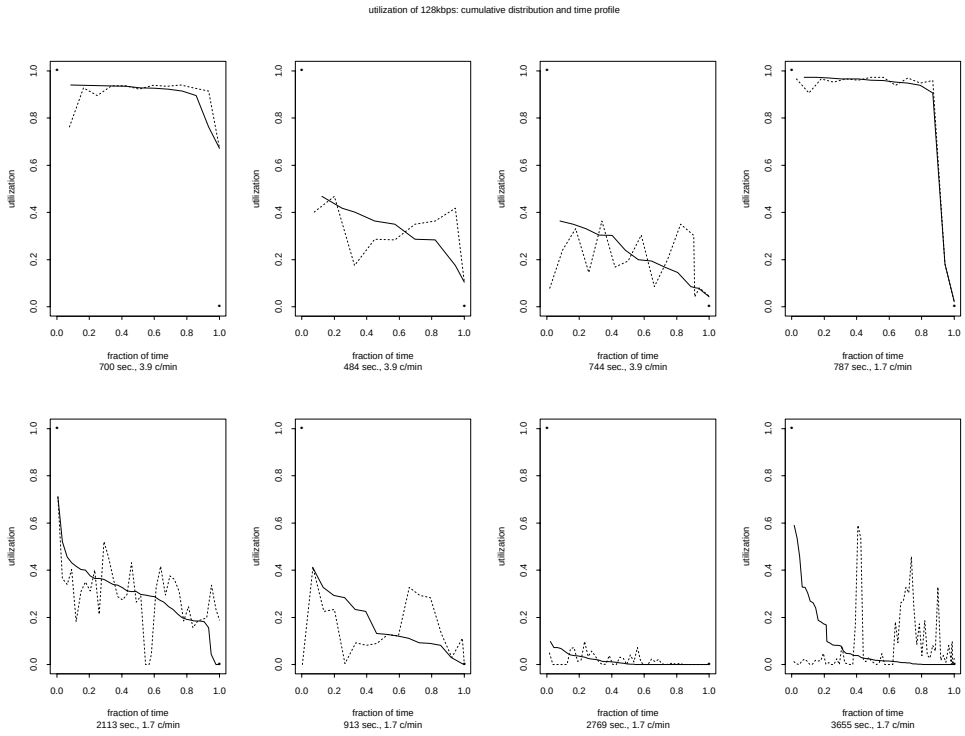

Figure 5: Selected connections to 128*kbps*

utilization. Low utilization suggests interactive traffic, e.g. games, and mere convenience. This is a crude classification, since within each minute, several TCP and UDP connections can be open simultaneously. This means that the actual distribution of applications during a connection to a bandwidth may be different from what this classification suggests.

One might consider the following approach to estimate convenience time. For any connection s, let t_s denote the entire duration, in seconds, of the session, with inbound and outbound volumes (in *kbits*) v_s^{in} and v_s^{out}, respectively, and b_s the chosen capacity (in *kbps*). Then, effective inbound and outbound utilizations for each (fraction of a) minute of connection s, indexed by subscripts m, are given by $u^i_{s_m} = (v^i_{s_m}/d_{s_m})/b_s$, where $i \in \{in, out\}$ and $t_{s_m} = \sum_{m=1}^{\lfloor t_s/60 \rfloor + 1} d_{s_m}$, for $d_{s_1} \geq 60$, $d_{s_m} = 60$ for $m = 2, \ldots, \lfloor t_s/60 \rfloor$, and $d_{s_{\lfloor t_s/60 \rfloor + 1}} \geq 60$. Then, overall network conditions in terms of minimum potential capacity can be approximated by the maximal achieved capacity utilization $\max_{m=1,\ldots,\lfloor t_s/60 \rfloor+1; i \in \{in,out\}} \{u^i_{s_m}\}$ in each minute of connection s and in each transmission direction. Using this approach to determine the condition of the network during a connection, an estimate for convenience time in connection s is $t_s - \frac{v_s}{b_s \max_{m,i}\{u^i_{s_m}\}}$, where $v_s = (v_s^{in} + v_s^{out})/2$. Figure 7 displays, as the solid curve, the empirical distribution of $\max_{m,i}\{u^i_{s_m}\}$ for the user's connections with 128*kbps* potential capacity, when the per minute prices were 3.9 c/min (left) and 1.7 c/min (right). The dotted curves below give average utilization for the corresponding sessions. The difference between dotted and solid curves, relative to the height of the solid curve, is a measure for the fraction of minimum

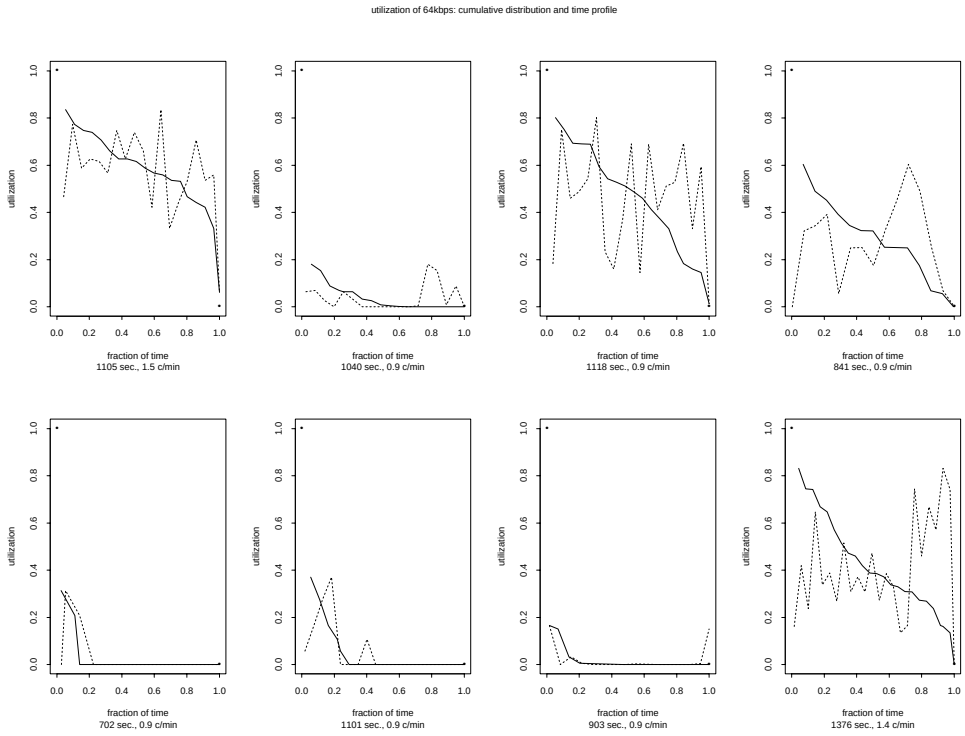

Figure 6: Selected connections to 64*kbps*

potential capacity that was kept for convenience. The two graphs again show that higher per minute prices for a capacity induce more conservative utilization of minimum potential capacity, as higher prices reduce the relative distance between the solid and broken curve. Table 2 condenses this in mean and median fraction of estimated convenience time relative to total duration.

capacity	price	mean	median
128*kbps*	3.9 c/min	0.509	0.499
	1.7 c/min	0.793	0.802
64*kbps*	1.5 c/min	0.532	0.478
	0.9c/min	0.741	0.715

Table 2: Mean and median fraction of convenience time

4.1.2 Volume Choice

Before proceeding to some heuristic characterization of preferences, it may be worth briefly characterizing the user's volume choices. As one might expect, due to the invertible relation-

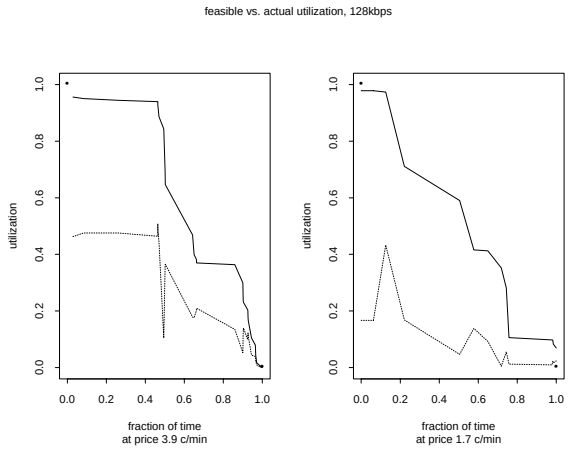

Figure 7: Utilization of minimum potential capacity (128*kbps*)

ship between volume and transmission-active time, given effective capacity, there exists a close relationship between per-minute capacity prices and demand for volume. Mean and median volume both are higher when per-minute capacity prices are lower: At 3.9 c/min for 128*kbps*, mean volume at this capacity was 1.026.202 bytes (median 35.013,0 bytes), while it was 2.731.610,0 bytes at 1.7 c/min (median 867.253,0 bytes). In the former case, prices were generally higher than in the latter, ranging form 1.1 c/min for 16*kbps* to 3.9 c/min for 128*kbps*, compared to 0.5 c/min for 16*kbps* to 1.7 c/min in the latter period. This price pattern induced generally lower volume transmission in the former period. Overall mean volume was then 498.850, 2 bytes (median 34.217,0 bytes), as opposed to 1.512.640 bytes (median 188.808 bytes) at the time when the lower prices prevailed. Notice the strong discrepancy between mean and median, which is a result of the apparent skewness of the volume distribution. Higher prices also seem to induce some substitution away from volume delivered under expensive capacity: The fraction of volume transmitted at 128*kbps* at 1.7 c/min, 75.7 percent, exceeded the corresponding fraction at 3.9 c/min for 128*kbps* (54.7 percent) by more than 20 percentage points. Table 3 shows the shift toward low capacities in the relative volume distribution.

capacity	price	percentage	price	percentage	price	percentage
8*kbps*	0	0	0	8.3	0	2.7
16*kbps*	0.5	0.4	1.1	0.05	0.2	7.3
32*kbps*	0.7	12.0	1.3	0.05	0.3	53.1
64*kbps*	1.4	10.8	1.5	26.5	5.7	0.3
96*kbps*	1.6	1.1	3.8	10.5	9.6	0
128*kbps*	1.7	75.7	3.9	54.7	9.8	36.6

Table 3: Fraction of volume transmitted in different bandwidths, in percent

4.1.3 "Optimization Errors" and Insurance

Observed discrete choices of capacity jointly with continuous choices of volume and convenience time allow for various heuristic assessments of revealed preferences. Consider a consumption bundle that consists of the observed transmitted volume and an estimate of convenience time during which the option to have the chosen capacity available was not forfeited, even though no volume was transmitted. For the following heuristic assessment, convenience time was (presumably over-) estimated as $t - v/b$. The consumption bundle can be priced according to the pricing schedule prevailing at the time it was observed. One may then ask the question whether there was a capacity choice and associated price other than the chosen one that made the same bundle affordable at lower pecuniary cost. For the user under consideration, for 17 percent of all sessions there did not exist an alternative capacity choice that would have been cheaper in money terms; if the estimate for convenience time described in the above discussion is used, this fraction rises to 45.9 percent. Corresponding values for other users also lie in the range between 17 and 25 percent and similarly increase when the more conservative measure for convenience time is used.

Comparing observed choices to all possible choices in pecuniary terms alone, however, may overlook that many cost superior choices involve lower capacities and, therefore, come at the expense of higher time costs, if time enters the user's utility through leisure. If the observed choices are compared in cost terms to the subset of potential capacity choices with higher bandwidths, then, typically, 93 to 96 percent of the observed choices are optimal, leaving about 5 percent of all choices for which an alternative existed that would have been superior in terms of both money and time. With the alternative estimate of convenience time, the degree of such ex post suboptimal bandwidth choices is higher, ranging from 15 to 26 percent.

Such seemingly suboptimal choices can occur if ex ante and ex post valuations differ, so that ex ante planned or anticipated continuous choices are optimal, given the capacity choice and its associated pricing scheme, while ex post realized choices differ from anticipated choices as a consequence of a stochastic preference or taste shock. Figure 8 illustrates the case in which the anticipated bundle is not dominated, given the bandwidth choice which determines the relative price between time and volume,[12] and where the realized continuous choice pair is dominated by another bundle in pecuniary terms only (left panel) and in terms of both time and cost (right panel). Under the assumption that users employ a fully dynamic programming approach to make decisions, both types of ex post seemingly suboptimal choices should be equally likely on average. Under the heuristic, on the other hand, the former should occur more often if users on average ex ante desire an option that provides excess insurance against large byte volume choices; the latter would be expected to occur more frequently if users ex ante desire an option that provides excess insurance against large convenience time choices. Among the users investigated in this study, for any bandwidth the likelihood of it ex post being too high exceeds the risk of it being too small. This suggests that users on average desire insurance against high

[12] Note that a higher capacity is associated with a higher relative price for volume.

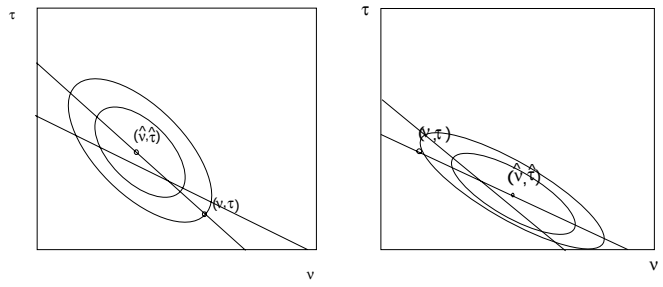

Figure 8: Anticipated and realized choices; in both cases, the budget lines are drawn for the same level of expenditure; ellipsoids represent stochastic variation of realized choices about anticipated choices.

byte volume transmission. It also casts doubt on the presumption that users employ a fully dynamic approach to solve their sequential choice problem.

4.1.4 Further Heuristics on Revealed Preferences

Some further heuristic inferences can be drawn from the comparison of actual and feasible choices. These concern the marginal valuation of time or cost of switching. If a certain bandwidth was chosen by a user and the same bundle of volume and inactive time that was obtained could have been realized at lower cost at a lower bandwidth, at the expense of longer total duration, then the cost difference between the two bundles relative to the time saving under the chosen capacity approximates the lower-bound of the marginal value of time for that user or the cost of switching to another bandwidth. Similarly, comparing costs and durations of the chosen bundle with the ones under higher bandwidth alternatives with higher money cost, the ratio of cost to time difference approximates an upper-bound on the user's marginal value of time or switching cost. It is important to emphasize that these are heuristics, since all minutes in an incremental time interval are treated symmetrically and since the comparisons assume perfect foresight, ignoring the distinction between ex ante and ex post valuations.[13]

A couple of a priori hypotheses about these marginal valuations may be formulated. One might expect that a user will choose a high bandwidth when his marginal value of time is high, and conversely a low and inexpensive bandwidth when his time costs are low. Moreover, if a user commits a larger amount of time to Internet activity, then, analogous to the standard

[13]In terms of the computations, this means that this analysis is exclusively based on observations satisfying the Weak Axiom of Revealed Preference.

assumption of diminishing marginal valuation in economic theory, one might expect that his marginal valuation is lower than otherwise.

The data of the representative user appear to lend support to these conjectures. Table 4 gives median lower and upper bounds on the user's marginal value of time, arranged by the user's capacity choice. Columns 1 & 2 provide median bounds, based on the the assumption that 100 percent of capacity was and would have been available, respectively; columns 3 and 4 give mean and median bounds, based on convenience time estimates derived as described in the previous subsection. High bandwidths are seen to be chosen when the user's marginal valuation of time is high, and vice versa, just as expected. The upper bounds are less informative The data for 128 *kbps* were less abundant than for the other capacity choices, which may explain that the reported lower bound does not obey the expected monotonicity. Correlations between the

capacity	med.l.b.	med.u.b	med.l.b.	med.u.b
32*kbps*	1.41	58.69	0.13	20.3
64*kbps*	1.87	42.25	0.84	5.5
96*kbps*	18.72	631.26	2.33	99.8
128*kbps*	11.54	n.a.	3.25	n.a.

Table 4: Median bounds on marginal valuation of time, in cents per minute

bounds in columns 1 & 2 and volume and duration, respectively, support the conjecture about diminishing marginal valuations; cp. table 5, which displays (with one exception) uniformly negative correlations between the bounds on marginal time valuation and the two measures of activity. Table 6 presents this observation with greater resolution, breaking up the relationship

capacity	duration		volume	
	l.b.	u.b.	l.b.	u.b.
32*kbps*	-0.23	-0.20	-0.30	-0.21
64*kbps*	-0.15	-0.10	-0.14	-0.11
96*kbps*	-0.13	0.06	-0.17	-0.32
128*kbps*	-0.21	n.a.	-0.22	n.a.

Table 5: Correlation of bounds with duration and volume

between the median lower bound of the user's marginal valuation of time according to observed usage, where usage is measured in terms of volume. Again, as volume increases, marginal valuations decrease, suggesting diminishing marginal valuation of time.

capacity	1st quartile	2nd quartile	3rd quartile	4th quartile
32*kbps*	5.35	2.28	0.61	0.34
64*kbps*	114.70	0.80	1.31	0.92
96*kbps*	44.02	21.31	14.03	2.12
128*kbps*	336.28	12.17	10.92	1.10

Table 6: Median lower bounds (in cents per minute) by volume quartiles

4.2 Estimation Results

This section reports the results from the estimation of the model. Since the sequence $\{M_s\}$ of total expenditures is not observed, the model is estimated conditional on observed expenditures on Internet services. This has the consequence that $\sigma^2_{\epsilon_2}$ is not identifiable. Also, it is clear from the functions $\nu(b, \theta, \epsilon)$, $\tau(b, \theta, \epsilon)$ and $x(\theta, \epsilon)$ that expenditures on volume v and convenience time t in the general model specification may very well be correlated with the residuals in the stochastic demand equations. This is the case if the outside good x is not redundant. In fact, it is easy to see that in this case the expectation of the continuous choice variables, given observed expenditure, exceeds the expectation, given exogenous total outlays M, by a positive bias term. The impact of this positive bias term in the corresponding moment conditions on the parameter estimates and how to test for it will be briefly discussed below. The remaining identifiable model parameters are $\sigma^2_{\epsilon_1}, \sigma^2_\theta, \sigma_{\theta\epsilon_1}$ and the parameters ξ in $f(\mathbf{z}; \xi)$. Since at the time of this study, no covariates other than the date of the observed connection and the start time are available, these are used to create proxies for whether the observed connection was work related or not. Whether or not a connection is work related may be reflected in a user's behavior. It may determine whether the user herself or her employer pays for the connection. It may also restrict the class of applications that make up the transmission activity of the connection. For lack of more accurate covariate data, the proxies used are two indicator variables, taking value one if the date of the connection corresponds to a regular working day (z_1) and a regular work hour, 7am-7pm, (z_2). These proxies are weak since many of the INDEX subjects are UC Berkeley students and faculty members whose schedules are likely to deviate from this notion of regularity. Once more accurate covariate data become available from the INDEX database, these are to be included instead. For simplicity, the linear specification $f(\mathbf{z}) = \mathbf{z}'\xi$ was chosen.

The model parameters are estimated by the method of simulated moments, as outlined above. The conditional moments used are the ones described in section 3.3.2, and unconditional moments are formed by choosing the vector of prices as instruments for conditional moments of the continuous choice variables. This leads to a total of $r = 20$ unconditional moments. Under regularity conditions, which essentially amount to uniform convergence in probability of the objective function, compactness of the parameter space and identification, this estimation procedure yields consistent estimates of the model parameters. For simulation, $T = 10$ simulation sample draws were used. The estimator is a two-stage feasible Method-of-Simulated Moments estimator. In a first step, initial consistent estimates $(\hat{\xi}', \mathrm{vec}(\hat{\Sigma})')'$ are

obtained by choosing the weighting matrix $Q_S = \mathbf{I}_r$. These are then used to form an estimate of the optimal weighting matrix $\mathbf{V}_0^{-1} = E[D(\xi_0, \Sigma_0)D(\xi_0, \Sigma_0)']^{-1}$, where ξ_0 and Σ_0 denote the unknown true parameters; a consistent estimate is given by $\hat{\mathbf{V}}_S = E_S[D_s^\star(\hat{\xi}, \hat{\Sigma})D_s^\star(\hat{\xi}, \hat{\Sigma})']$. In a second step, the unknown parameters are re-estimated with $\mathbf{Q}_S = \hat{\mathbf{V}}_S^{-1}$. Under some further regularity conditions, this two-step procedure yields asymptotically efficient, consistent estimates, given the set of instruments.[14] Note, however, that there exist further – in fact: an infinite number of – instruments that could be used to form unconditional moment conditions. Except for the estimates of the slope parameters ξ, the asymptotic distribution, strictly speaking, is not normal, due to the domain restrictions $\sigma_\theta^2 \geq 0$, $\sigma_{\epsilon_1}^2 \geq 0$ and $|\sigma_{\theta\epsilon_1}| \leq \sigma_\theta \sigma_{\epsilon_1}$.[15] Approximate asymptotic standard errors are computed on the basis of the usual normal approximation, using $E_S[\nabla_{(\xi', \mathrm{vec}(\Sigma)')} D^\star(\hat{\xi}, \hat{\Sigma})]$ as an estimate of $\mathbf{M}_0 = E[\nabla_{(\xi', \mathrm{vec}(\Sigma)')} D(\xi_0, \Sigma_0)]$ for the corresponding expression in the asymptotic variance covariance matrix $\frac{T+1}{T}\left(\mathbf{M}_0'\mathbf{V}_0^{-1}\mathbf{M}_0\right)^{-1}$.

Table 7 presents parameter estimates for a subset of users; in the table, $\rho_{\theta\epsilon_1} = \frac{\sigma_{\theta\epsilon_1}}{\sigma_{\epsilon_1}\sigma_\theta}$. The estimation results point to a number of observations. There appears to be considerable

userid	$\hat{\sigma}_{\epsilon_1}^2$	$\hat{\sigma}_\theta^2$	$\hat{\rho}_{\theta\epsilon_1}$	$\hat{\xi}_1$	$\hat{\xi}_2$
341645	3.94	4.42	-0.75	0.02	0.69
	(0.01)	(0.03)	(0.02)	(0.03)	(0.03)
595892	2.94	3.28	-0.12	2.51	-1.02
	(0.37)	(0.12)	(0.08)	(0.04)	(0.05)
335632	1.35	2.42	-0.91	-1.73	0.11
	(0.11)	(0.24)	(0.07)	(0.46)	(0.75)
588679	1.24	4.29	-0.89	0.40	-0.90
	(0.06)	(0.22)	(0.08)	(0.26)	(0.36)
648986	0.48	4.91	-0.44	0.52	-1.11
	(0.01)	(0.23)	(0.09)	(0.20)	(0.24)
883272	3.04	3.93	-0.65	-1.63	1.03
	(0.07)	(0.09)	(0.15)	(0.04)	(0.18)
935791	4.60	4.82	-0.01	-1.40	-1.47
	(0.03)	(0.02)	(0.01)	(0.10)	(1.05)

Table 7: MSM estimates; approximate standard errors in parenthesis

variation in ex ante valuations, measured by $\hat{\sigma}_{\epsilon_1}^2$, both between users and for each user when comparing the user's connections. A lot of this variation could presumably be explained by data on the types of applications that a user carries through in a particular connection. For ftp traffic within the UC Berkeley campus, INDEX can (almost) guarantee 100 percent of chosen capacity. For web surfing, on the other hand, the received transmission speed is determined

[14] The additional conditions needed are that a central limit theorem applies to the gradient of the vector of moments and that a uniform law of large numbers applies to the empirical analogue of the optimal weighting matrix.

[15] These parameters are estimated using transformations that map onto the real line.

by the level of congestion along the entire path between user and destination host. Therefore, planned applications are likely to determine the type of capacity that seems ex ante desirable. And the diversity in a user's types of activity then is reflected in the estimate of $\sigma^2_{\epsilon_1}$.

Users also exhibit similarly strong, if not larger variation in ex post valuations. This suggests that users differ in terms of their ex ante disposition toward Internet services, as reflected in the wide range of estimates $\hat{\sigma}^2_{\epsilon_1}$, and that the consumption experience itself induces a discrepancy between ex ante and ex post valuations, as portrayed by estimates $\hat{\sigma}^2_{\theta}$ dominating in size the estimates for the variation in ex ante valuations. The uniformly negative estimates of $\rho_{\theta\epsilon_1}$ suggest furthermore that users deviate in their on-line service valuations from their ex ante valuations. The estimates of the coefficients on the work proxies are to be interpreted with caution, for reasons already pointed out. There does not appear to be a regular pattern applicable to all users. For some users, on the premise of the validity of the work proxy, the result suggests a tendency for work-related activity to reduce convenience time, conditional on prices and expenditure.

How are the differences among users in the estimates for their idiosyncratic ex ante and ex post taste valuation to be interpreted? As already alluded to, high variation in ex ante valuations could possibly be ascribed to different applications that a user has planned. If a given user exhibits a higher estimate for $\sigma^2_{\epsilon_1}$ than another user, then this may reflect that the range of planned applications of the former user is wider than for the latter. If this were the case, then observed capacity choices should also reflect this wider range of applications, at least if the user chooses optimally. In other words, more dispersed capacity choices might be expected to be associated with higher values of $\sigma^2_{\epsilon_1}$, both intuitively and under the estimated model; cp. how ϵ_1 factors into the discrete choice probabilities ($3-2$). A measure of dispersion for the sequence of observed capacity choices, given a set of prices, is the conditional entropy,

$$E(\mathbf{p}) = -\sum_{b} \Pr(b;\mathbf{p}) \ln(\Pr(b;\mathbf{p})),$$

where $\Pr(b;\mathbf{p})$ is the conditional probability that b is chosen, given prices $\mathbf{p}$. An empirical analogue is formed by nonparametrically estimating the choice probabilities as the fraction of times that b was chosen while prices $\mathbf{p}$ prevailed out of all connections observed during that time. This measure, in analogy to its use in statistical physics, characterizes the energy in a sequence of discrete observations and intuitively is perhaps best thought of as summarizing the degree of surprises in a multinomial sequence or, more generally, as a measure of uncertainty in a random variable.[16] Since different users are confronted with different sets of prices, the average empirical entropy across all price sets in the experiment, $\bar{E} = \frac{1}{\#\mathbf{p}} \sum_{\mathbf{p}} E(\mathbf{p})$, can serve as a measure for dispersion of capacity choices. For table 8, two users with a similar number

[16]The entropy measure has interpretations in many fields. In Statistical Physics, its interpretation is the energy in a random variable. For a binary sequence, like in the Ising model for particles in a field, the energy is highest if the probability of the occurrence of one is 1/2 and the energy is zero if each particle takes on the same value with probability one. In Information Theory, defined in terms of the logarithm with base 2, it is interpreted as the number of bits that is on average required to describe the random variable; cp. Cover and Thomas (1901).

of paid connections S are chosen. Indeed, the user with the higher estimate of $\sigma^2_{\epsilon_1}$ also exhibits the higher average dispersion in his or her discrete choice behavior.

userid	$\hat{\sigma}^2_{\epsilon_1}$	$\hat{\sigma}^2_{\theta}$	E	ex post opt.error
341645	3.94	4.24	1.32	24.5%
595892	2.94	3.28	0.60	8.8%

Table 8: Differences in estimates between users

The variation in ex post valuations, in light of the model as well as intuitively, should render some capacity choices suboptimal ex post. A higher degree of variation might then be expected to lead to a higher percentage of ex post seemingly suboptimal bandwidth choices. This again can be found in the data for these subjects; here, apparent "optimization error" is measured as the percentage of all connections for which there would have been an alternative among the higher capacity alternatives that would have cost less, given the chosen bundle of observed transmitted volume and estimated convenience time.

As pointed out at the beginning of this subsection, joint expenditures on volume and convenience time may be correlated with the residuals from the difference between observed continuous choices and their expectation, given expenditure. This correlation amounts essentially to a selection bias in the moment conditions. Selection leads to an downward bias in the first moments of the continuous choice variables, conditional on observed expenditure; the true conditional moments exceed the ones which are assumed to hold for estimation. The impact of this bias on the estimates of $\sigma^2_{\epsilon_1}$ and σ^2_{θ} depends on the sign of the contribution due to observed preference heterogeneity $\mathbf{z}'\xi$. If this term is negative, then these variances tend to be overestimated; otherwise, they tend to be underestimated.[17] Exogeneity of expenditures can easily be tested. One includes a coefficient α on the component $\ln(x)$ in the utility function and tests for exogeneity by examining the null hypothesis $H_0 : \alpha = 0$. A score test is a convenient test procedure in this context since it obviates estimation of the alternative model. Under the null hypothesis, the score test statistic has a χ^2_1 distribution. The maximal score test statistic for the chosen group of users is 3.35, which still lies below the 95 percent critical level for rejection of 3.85. Therefore, the null hypothesis of exogenous expenditures cannot be rejected at the 95 percent significance level.

[17] This can intuitively be seen from the following comparison. Given $\mathbf{z}'\xi$ and ϵ_1, consider, on the one hand, the point about which the stochastic expenditure shares are symmetric in θ. If this point is larger than the mean of θ, then, in order to compensate for the downward bias in the conditional moments, the estimated variance of θ is larger than the true variance. The tendency of the bias in $\hat{\sigma}^2_{\epsilon_1}$ is more complicated, since the conditional logit choice probabilities, which do not depend on expenditure, are jointly used for estimation.

4.3 Prediction

One may finally wish to examine the estimated model's predictive ability. From a practical point of view, once a capacity choice is observed, one would hope to accurately predict for the imminent session the distribution of continuous choice variables from which a realization is drawn, conditional on the discrete choice; if this distribution is estimated, then, of course, all its moments can also be estimated. At this point, one does not observe total expenditure for this connection. But the distribution of past expenditures for connections in which the observed capacity was the chose one is known. So a natural simulation experiment would be to sample with replacement from the distribution of previous expenditures associated with the observed capacity and to use the sampled expenditures to simulate the continuous choice variables volume and convenience time to predict the distribution from which a draw is expected. To thoroughly check the predictive power of the model, one would split the sample in a training and a control group, use the former to estimate the model and simulate the responses conditional on the covariates of the latter to compare them with the associated observed responses. The number of observations (sessions) for a user and a given capacity, unfortunately, is not always large enough to allow for this method. Alternatively, in a jack-knife approach, the model could be estimated leaving out one data point at a time and subsequently be used to predict that data point. This is computationally expensive due to simulation effort in each estimation iteration.

Two different comparisons of model predictions and actual data are made. First, largely for convenience, the actual distribution of volume and estimated convenience time and the predicted one for re-sampled expenditures is compared. Figure 9 shows actual and predicted distributions for a user's $64kbps$ sessions; for this user, there were 81 such sessions, and the same number of choices were simulated on the basis of expenditures that were sampled with replacement from the empirical distribution of expenditures when $64kbps$ was observed. The predicted distribution of convenience time appears to place slightly more weight on lower values than the actual distribution, while the predicted distribution of byte volumes seems to track the actual distribution quite well. For the drawn sample of expenditures, the predicted maximal volume, however, is noticeably smaller than the maximum of observed volume choices.

Secondly, the sample was split by setting aside those data records that pertain to sessions in which $64kbps$ was chosen and the vector of per-minute prices was $\mathbf{p}' = (1.1, 1.3, 1.5, 3.8, 3.9)$; there are 66 such records for the chosen user. Then, the model parameters were estimated on the reduced sample, consisting of 164 observations which include 15 sessions in which $64kbps$ was chosen and a different price menu prevailed. The estimated model was then used to predict the left-out observations on the basis of re-sampled expenditures for $64kbps$. Note that this may be a challenge to the model since there are relatively few choices of $64kbps$ used for estimation. The predictive ability of the model nonetheless appears to be satisfactory. Figure 10 displays the distribution of actual and predicted choices. Again, the distributions deviate mainly in the tails, underestimating the extremes in the actual distribution; the quantiles in table 9 confirm this. Note that the predicted distributions adequately capture the skewness of the actual distributions.

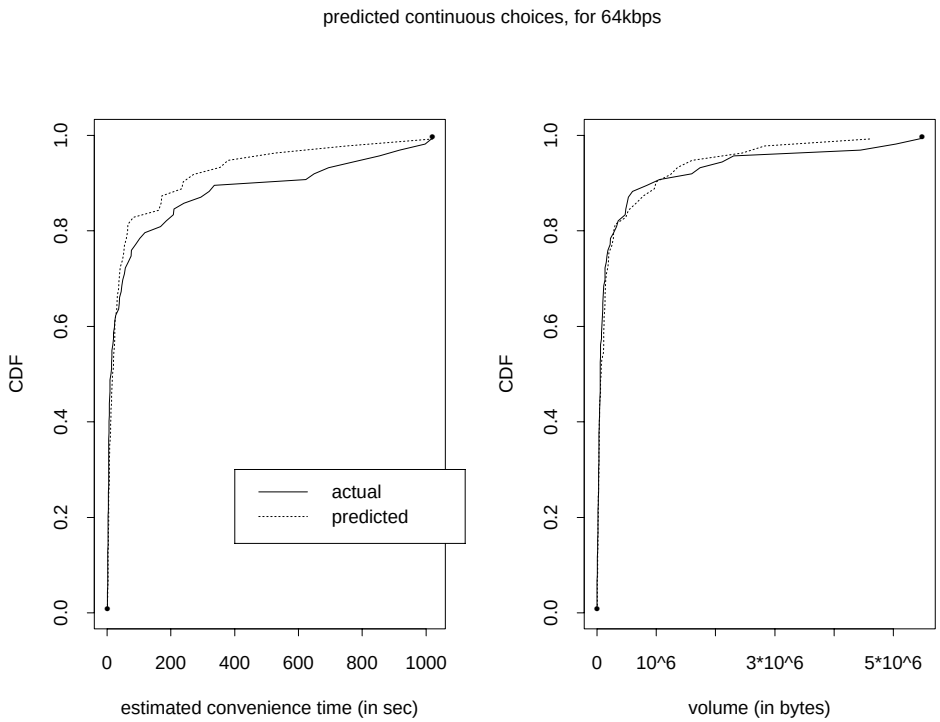

Figure 9: A user's actual and predicted continuous choices, for 64*kbps*

Finally, the model can be used to simulate the user's response to changes in prices. Suppose that, instead of charging 1.5 c/min in 64*kbps*, the charge is doubled. The model, estimated on the reduced sample, predicts the following changes. Since doubling the per-minute charge for 64*kbps* service reduces the indirect utility of this bandwidth choice relative to the alternative choices, the probability of 64*kbps* being chosen is diminished, while alternative choices become more likely. Table 10 displays the relative changes in the median estimated discrete choice probabilities. Furthermore, conditional on adhering to the choice of 64*kbps*, the amounts of volume and convenience time consumed substantially decrease. Figure 11 shows the predicted distributions for byte volume and convenience time, conditional on the choice of 64*kbps*, under the two price scenarios. Median volume is reduced by about 40 percent, while median convenience time decreases by 47 percent.

Since the distribution of the econometric errors ϵ, conditional on any bandwidth choice, depends on all prices, the contemplated change in prices affects the conditional distribution of volume and time choices, conditional on any bandwidth choice. This allows to simulate cross-price effects. After the price change, 32*kbps* is the bandwidth capacity associated with the highest discrete choice probability. Figure 12 displays the conditional distribution for the continuous choice variables, given that 32*kbps* service is chosen. The model predicts that the user indeed substitutes volume and time out of 64*kbps* into 32*kbps*.

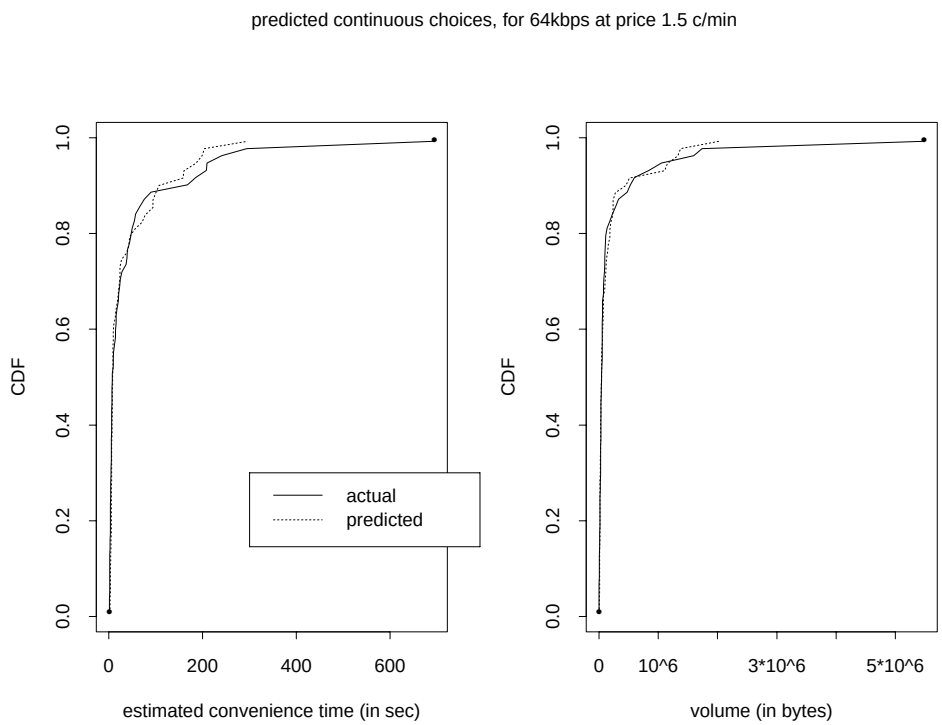

Figure 10: A user's actual and predicted continuous choices, for $64kbps$ at 1.5 c/min

5 Conclusions

This paper develops an econometric model of Internet user preferences over Internet service attributes that can assess users' valuations of Internet services. The general modeling framework allows consumers of service products to learn about latent, unknown productivity or quality of inputs to the service production technology. This allows intertemporal preference heterogeneity, distinguishing between a user's ex ante valuation, prior to service utilization, and ex post valuation, condition on his or her service technology choice. Discrete and continuous choices are modeled as jointly endogenous, and randomness in the observed choices emerges from preference heterogeneity, both intertemporal for a sequence of a user's Internet sessions and interpersonal for different users. In this regard, this work builds on the theoretical foundations of stochastic demand modeling laid out in Beckert (1999). Data from the U.C. Berkeley Internet Demand Experiment on discrete–continuous choice sequences allow to identify and estimate the model. It is found that considerable heterogeneity in preferences exists, both among different users and, perhaps more interestingly, for each user over time. Users' variation in ex ante valuations are typically accompanied by similar, if not larger, variation in ex post valuations. And users appear to deviate in their on-line service valuations from their ex ante valuations. The predictive power of the model appears satisfactory in out-of-sample predictions, recommending the model for demand management.

	time (sec)		volume (kbytes)	
quantile	actual	predicted	actual	predicted
0%	1.0	2.1	0.08	2.7
25%	3.6	5.1	15.7	14.4
50%	7.2	7.7	39.5	39.0
75%	37.8	27.0	96.6	120.2
100%	694.6	296.3	548.9	202.5

Table 9: Actual and predicted quantiles; $64kbps$ at 1.5 c/min

bandwidth b	$16kbps$	$32kbps$	$64kbps$	$96kbps$	$128kbps$
$\frac{\Delta \tilde{P}(b;\mathbf{p})}{\tilde{P}(b;\mathbf{p})}$	0.02	0.46	-0.61	0.65	0.67

Table 10: Relative changes in median discrete choice probabilities, in percent

Future work might attempt to address the difficult question how to connect the conceptual framework adopted here for the analysis of joint decisions on capacity and utilization with the work of Rust (1987, 1994) on controlled stochastic processes, aiming at a unified paradigm for the analysis of sequential discrete-continuous choices. A related conceptual extension might allow for the possibility that the timing of discrete choices impacts the degree of ex post uncertainty. This variant would be particularly interesting in capacity management problems in which customers have the option to postpone discrete choices in order to benefit from the ensuing reductions in ex post uncertainty. Yield management problems, like airline and car rental booking, naturally beg for this extension. The latter extension might be part of a broader attempt to empirically investigate issues of learning in the context of the INDEX data.

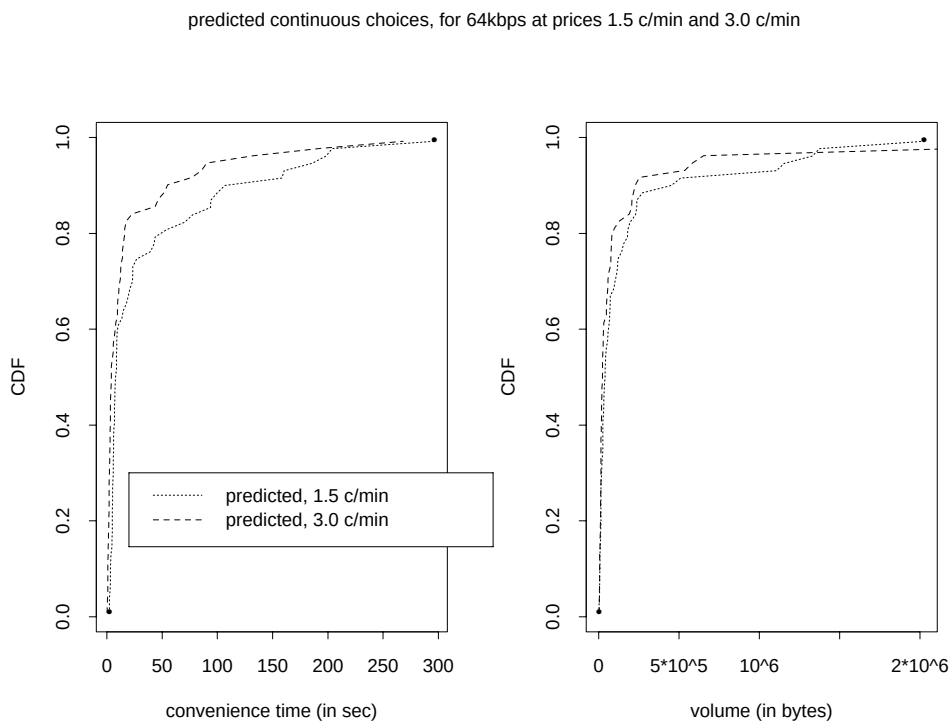

Figure 11: A user's predicted continuous choices, for $64kbps$ at prices 1.5 c/min and 3.0 c/min

Appendices

A Motivation of the "Reduced-Form Utility" Model Specification

This appendix develops the reduced-form utility model from specific functional forms for utility and service production technologies as well as assumptions on the joint distribution of the latent parameter α and the signal.

Assume the following specific functional forms. Let

$$\begin{aligned} \tilde{U}(A,x) &= \ln(A)+\ln(x) = \sum_{i=1}^{m} 1_i \tilde{U}_i \\ \tilde{U}_i &= \ln(a_i)+\ln(x), \qquad i=1,\ldots,m. \end{aligned}$$

For the service production technologies, suppose that there is a single input $\mathbf{w}_i = v$ for all i, and

$$a_i(v,\alpha) = \check{\zeta}_i v^{\exp(\alpha_1)} \exp(-\alpha_2^2),$$

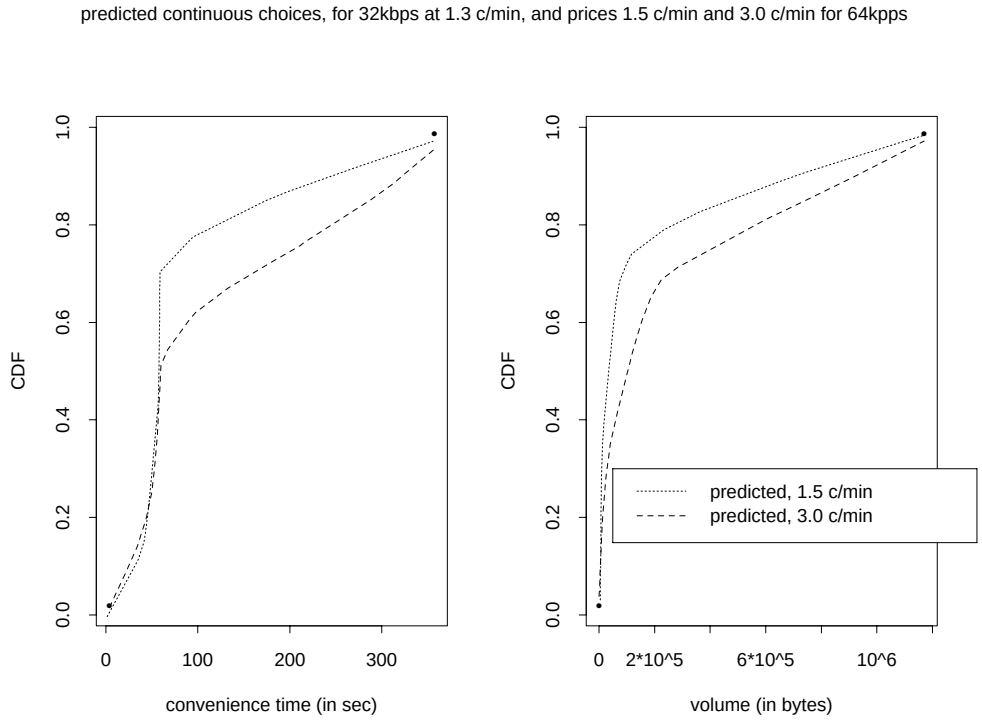

Figure 12: A user's predicted continuous choices, for 32*kbps* at prices for 64*kbps* of 1.5 c/min and 3.0 c/min

where $\alpha = (\alpha_1, \alpha_2)'$, α_1 governs the uncertainty about the productivity of v, α_2 governs the uncertainty about total productivity, and $\{\alpha_j, j = 1, 2\}$ are independently and identically distributed. Notice that a_i is log-concave, and so $\tilde{U}_i$ is concave in α.

Suppose that a signal $\check{\theta}$ is received, conditional on α, and that $\check{\theta}|\alpha$ is normal, with conditional mean $\iota'\alpha$ and conditional variance $\frac{1}{t_i - 1}\sigma^2$. Also, assume the priors $\pi(\alpha_j)$, $j = 1, 2$, are normal with mean 0 and variance σ^2. Then, the posteriors $\pi(\alpha_j|\check{\theta}, i)$, $j = 1, 2$, are conditionally normal with mean $\frac{t_i - 1}{t_i}\check{\theta}$ and variance $\frac{1}{t_i}\sigma^2$. The marginal distribution of $\check{\theta}$ is normal with mean 0 and variance $\frac{t_i}{t_i - 1}\sigma^2$.

Then,

$$\begin{aligned} U(v, t, x, i; \check{\theta}, \sigma^2, \zeta_i) &= E\left[\tilde{U}_i|\check{\theta}, i\right] \\ &= E\left[\exp(\alpha_1)|\check{\theta}, i\right]\ln(v) - E\left[\alpha_2^2\middle|\,\check{\theta}, i\right] + \ln(x) + \ln(\check{\zeta}_i) \\ &= \exp\left(\check{\theta} - \frac{1}{t}\left(\check{\theta} - \frac{1}{2}\sigma^2\right)\right)\ln(v) - \mathrm{var}(\alpha_2|\check{\theta}, i) - \left(E[\alpha_2|\check{\theta}, i]\right)^2 + \ln(x) + \ln(\check{\zeta}_i) \\ &= \exp\left(\check{\theta} - \frac{1}{t}\left(\check{\theta} - \frac{1}{2}\sigma^2\right)\right)\ln(v) - \frac{\sigma^2}{t_i} + \ln(x) + \ln(\check{\zeta}_i) + o_p(t_i^{-1}). \end{aligned}$$

Now $E\left[\exp(\check{\theta} - \frac{1}{t_i}(\check{\theta} - \frac{1}{2}\sigma^2))\right] = \exp(\frac{1}{2}\sigma^2)$. So, omitting terms of order $o_p(t_i^{-1})$, a representation of preferences that is *marginally* equivalent is

$$U((v, t, x, i; \theta, \sigma^2, \zeta_i) = exp(\theta)\ln(v) - \frac{\sigma^2}{t_i} + \ln(x) + \ln(\zeta_i),$$

where θ is normally distributed with mean zero and variance $\sigma_\theta^2 = \sigma^2$. This representation of preferences effectively imposes the restriction that U be separable in v and t_i conditional on θ. Finally, considering the marginal utility of t_i, a simple monotonic transformation of the units of t yields the equivalent representation

$$U(v, t, x, i; \theta, \sigma^2, \zeta_i) = \exp(\theta)\ln(v) + \sigma_\theta^2\ln(t_i) + \ln(x) + \ln(\check{\zeta}_i).$$

B Proof of the Result in Section 2.2

In this appendix, the result given in section 2.2 is established, as a straightforward consequence of the following

Lemma A1: *Let* $f(\theta; a, b) = \frac{e^{b+\theta}}{a+e^{b+\theta}}$, *for constants* $a > 0$ *and* $b \in \mathbb{R}$, *and* $\theta \sim N(E[\theta], \sigma^2)$, $\sigma^2 > 0$. *Then,* $E[f(\theta; a, b)] \leq f(E[\theta]; a, b)$ *if and only if* $E[\theta] \geq \ln(a) - b$.

Remark: Since f is strictly monotonically increasing in the parameter θ, Lemma A1 implies that $f(E[\theta]; a, b) < f(E[\theta] + \frac{1}{2}\sigma^2; a, b)$, producing the result in section 2.1 after substitution for

a and b. As an aside, notice that it is impossible to analytically express $E[f(\theta; a, b)]$, unless $E[\theta] = \ln(a) - b$ so that f is odd and the distribution of θ is symmetric about $E[\theta]$, in which case $E[f(\theta; a, b)] = \frac{1}{2}$.

Proof: Only the case $E[\theta] \geq \ln(a) - b$ is shown; the other case follows from a symmetric argument.

$E[f(\theta; a, b)]$ can be decomposed as

$$\begin{aligned} E[f(\theta; a, b)] &= E[f(\theta; a, b)|\theta \leq \ln(a) - b]P(\theta \leq \ln(a) - b) + \\ & \quad E[f(\theta; a, b)|\theta > \ln(a) - b]P(\theta > \ln(a) - b). \end{aligned}$$

Notice from the second derivative of f with respect to θ that $\frac{d^2}{d\theta^2} f(\theta_0; a, b) = 0$ if and only if $\theta_0 = \ln(a) - b$. For $\theta \leq \theta_0$, f is convex in θ, while it is concave for $\theta > \theta_0$. Therefore, by Jensen's inequality,

$$\begin{aligned} E[f(\theta; a, b)|\theta \leq \ln(a) - b] &\geq f(E[\theta|\theta \leq \ln(a) - b]; a, b) \\ E[f(\theta; a, b)|\theta > \ln(a) - b] &\leq f(E[\theta|\theta > \ln(a) - b]; a, b). \end{aligned}$$

If $E[\theta] \geq \ln(a) - b$, then $P(\theta \geq \ln(a) - b) > \frac{1}{2} > P(\theta < \ln(a) - b)$. Therefore,

$$E[f(\theta; a, b)] \leq 2f\left(E[\theta|\theta > \ln(a) - b]; a, b\right) P(\theta > \ln(a) - b).$$

Now $E[\theta] \leq E[\theta|\theta > \ln(a) - b]$ and $f(E[\theta]; a, b) > \frac{1}{2}$, because f is increasing in θ and $E[\theta] > \ln(a) - b$, and so $f(E[\theta|\theta > \ln(a) - b]; a, b) > f(E[\theta]; a, b) > \frac{1}{2}$. Also, $P(\theta > \ln(a) - b) > P(\theta > E[\theta]) = 1/2$. It then follows from the concavity of f for $\theta > \ln(a) - b$ that

$$\frac{P(\theta > \ln(a) - b)}{1/2} \leq \frac{f(E[\theta]; a, b)}{f(E[\theta|\theta > \ln(a) - b]; a, b)},$$

and so the conclusion follows, completing the proof. □

C Some Further Modeling Issues

This appendix attempts to motivate the proposed model and its features by appealing to more primitive and more general preference models that generate the features of the reduced form model under certain model assumptions and hence, under these assumptions, are observationally equivalent to this model. This section can therefore be viewed as a critical assessment of the model's limitations. It addresses essentially two model issues separately. The first part is devoted to the notion of time and its utilitarian value, spelling out in more detail the trade-off between time input to the production of an Internet product and leisure. In light of the empirically plausible conjecture that preferences for Internet services are subject to unanticipated taste and informational quality parameters evolving as a process over time during which the

service is consumed, the second part interprets the proposed one-shot decision problem in a dynamic framework. For a number of reasons, mentioned below, such interpretations are difficult to characterize analytically. Under some strong simplifying assumptions, conditions in an illustrative, yet restrictive dynamic model can be identified under which the optimal policy at each decision time amounts to solving the impending one-shot decision problem.

C.0.1 The Notion of Time in the Utility Model

The notion of time in the analysis of demand for Internet services is anything but unambiguous. On the one hand, there are reasons to believe that a trade-off exists between time spent on the Internet and leisure. This situation would arise if Internet services are used to accomplish work. On the other hand, one can think of numerous instances where time spent on the Internet is itself part of leisure time. In the latter case, even connections without continuous volume transmission can produce some form of recreational or aesthetic value that is part of the user's leisure time activities. Thinking in terms of a Lancaster-style theory of consumption[18], if leisure is associated with certain characteristics of social interaction and aesthetic pleasure, then volume transmission and processing time using the Internet can be as much input goods to a "consumption technology"[19] or intermediaries in the consumption process as going to a café, visiting a museum or buying a magazine. In the INDEX data available for analysis, such characteristics can at best be approximated by a broad site classification and different modes of transmission, like ftp, e-mail, audio, video etc. What is really measured are the amounts of time and volume consumed, and it is these commodities that an identifiable utility model should be defined over, even if it may obscure the complexity of the nature of consumption.

The remainder of this section formalizes the situation in which a trade-off between time using the Internet and leisure is envisaged, thereby disregarding the possibility that connection time may be associated itself with leisure time activity, and derives some condition on preferences that are sufficient to retain a positive marginal valuation of intellectual processing time, i.e. time beyond the technologically minimal amount necessary for volume transmission. These sufficient conditions, it will be shown, amount to the marginal utility of leisure being bounded at zero and to the relative marginal utility of intellectual processing time being uniformly large compared to the relative marginal utility of leisure.

Omitting for this discussion considerations about unobservable or stochastic taste parameters, suppose that a user's utility is defined over an Internet product[20] $\bar{v}$, leisure l and a composite outside good x and is given by $u(\bar{v}, l, x) = \psi(\bar{v})\Gamma(l) + \kappa(x)$. Furthermore, let $\bar{v}$ be generated by the technology $V(v, d)$ from two inputs, transmitted volume v and processing duration $d = t - v/b$, if bandwidth b is chosen. Notice that the time constraint $T = l + t$, for T

[18] Cp. Lancaster (1966, 1971, 1979);

[19] Lancaster (1966), p.137

[20] In Lancaster-style interpretation, this Internet product is presumably to be viewed as a means to achieve some "deeper consumption objectives"; cp. Lancaster (1979), p.7.

the exogenously given total amount of time available to the decision maker, links leisure time and processing duration and introduces a trade-off between time allocated to leisure and to the Internet product generation. To the analyst of data as the ones from INDEX, the desired product $\bar{v}$ is unobservable; only its inputs v and d are observed. It, therefore, is convenient to work with a reduced from utility function U, defined over v, d and x, the general properties of which can be justified from properties of the functions $\psi(\cdot)$, $\Gamma(\cdot)$ and $V(\cdot,\cdot)$. Assume that $\psi_{\bar{v}} = \frac{d}{d\bar{v}}\psi(\bar{v}) > 0$ and $\psi_{\bar{v}\bar{v}} < 0$ for all $\bar{v} \geq 0$, and similarly that $\Gamma_l > 0$ and $\Gamma_{ll} < 0$ for all $l \geq 0$ and that $V_v > 0$ and $V_d > 0$ for all $v > 0$ and $d \in (0, T)$. It, then, is clear that the reduced form utility function is also increasing in v and inherits the properties of u with respect to x. The remainder of this discussion focuses on the derivative with respect to d which is more complicated due to the time constraint. It will be seen that this derivative is positive for all $d \in (0, T - l - v/b), v \geq 0, l \geq 0$, given b, provided that the marginal utility of leisure at zero is bounded and the substitution elasticity between leisure l and the Internet product $\bar{v}$ is low, so that small sacrifices of leisure entail relatively large utilitarian gains in terms of $\bar{v}$.

In the above formulation, the marginal utility of processing duration is given by

$$\begin{aligned} u_d(V(v,d),l,x) &= V_d(v,d)\psi_{\bar{v}}(\bar{v})\Gamma(l)\left(1 - \frac{1}{V_d(v,d)}\frac{\psi(\bar{v})}{\psi_{\bar{v}}(\bar{v})}\frac{\Gamma_l(l)}{\Gamma(l)}\right) \\ &= V_d(v,d)\psi_{\bar{v}}(\bar{v})\Gamma(l)\left(1 - \frac{\bar{v}/l}{V_d(v,d)}\mathcal{E}_{l,\bar{v}}\right), \end{aligned} \tag{C-7}$$

where $\mathcal{E}_{l,\bar{v}} = \frac{\psi(\bar{v})/\bar{v}}{\psi_{\bar{v}}(\bar{v})}\frac{\Gamma_l(l)}{\Gamma(l)/l}$ denotes the substitution elasticity between l and $\bar{v}$. Under the assumption about the derivatives of ψ, Γ and V, the marginal utility of an incremental unit of processing duration is positive if the expression in parenthesis is positive for all v, d, l that satisfy the time constraint $T = l + d + v/b$, given b. For this to be the case, it is necessary for the marginal utility of leisure $\Gamma_l(l)$ to be bounded for all $l \geq 0$. Concavity of $\Gamma(\cdot)$ then implies that $\Gamma_l(l)|_{l=0} < \infty$ is sufficient for this necessary condition to hold. Also, uniformly low substitution elasticity $\mathcal{E}_{l,\bar{v}}$ between leisure and the Internet product render the expression in parenthesis positive if the marginal product of d, V_d, is not too low. As a specific example, consider the functional forms $\psi(\bar{v}) = \alpha(\bar{v})^\beta$ for $\alpha > 0$ and $\beta \in (0,1)$, $\Gamma(l) = \gamma(a+l)^\delta$ for $\gamma > 0$, $0 < a < \infty$ and $\delta \in (0,1)$, and $V(v,d) = vd$. Then, $u_d(V(v,d),l,x) \geq 0$ if $1 \geq \frac{\delta}{\beta}\frac{d}{a+l}$ for all $v, d, l : T = l + d + v/b$. Since $0 < a < \infty$, this condition is satisfied for all $l \in (0, T)$ satisfying the time constraint $T = l + v/b$; this again illustrates the requirement that that the marginal utility of leisure at zero be bounded. Moreover, the condition is satisfied for any $d \in (0, T)$ if $\frac{\delta}{\beta}T \leq a$, i.e. if it is satisfied for the maximal processing time, when $v = l = 0$; again, this latter condition in this example amounts to the previously stated additional sufficient condition that $\mathcal{E}_{l,\bar{v}} = \frac{\delta}{\beta}\frac{l}{a+l}$ be sufficiently low for all $l \in [0, T]$. The converse of these conclusions is that the reduced form model proposed above produces locally inaccurate predictions if leisure is valued highly relative to the Internet product.

C.0.2 The Model in a Dynamic Framework

Recall that one interpretation of the stochastic parameter θ appeals to ex ante randomness in the quality of information to be transmitted in an Internet session, so that actual continuous choices are determined by the ex post revelation of the quality of information. Since information accumulates bit by bit, θ could naturally be viewed as a continuous-time stochastic process $\{\theta_t, t \geq 0\}$, measurable with respect to a suitable filtration $\mathcal{F}_t = \bigcup_{s<t} \sigma(\theta_s)$, where $\sigma(\theta_s)$ denotes the σ-field generated by the random variable θ_s. This subsection sketches a dynamic model for discrete-continuous choice behavior in a continuous-time stochastic environment and, in light of the complexity of any such model, under rather restrictive assumptions attempts to identify some conditions on preferences that induce an optimal policy that is myopic and involves planning not to switch between capacities. Myopic optimal policies can be obtained in financial portfolio choice models with logarithmic and power utility specifications when the sole source of randomness is uncertainty about asset prices and decision times are exogenously given.[21] As will become clear shortly, the fact that choices are made sequentially and that subscription time to any discrete choice is itself a choice variable make the problem at hand less tractable.

In the interpretation of $\{\theta_t, t \geq 0\}$ as quality of information, this process may, but need not be exogenous to the agent's decisions; the latter case arises, for instance, through the choice of Internet sites to be visited in any given session. Since the objective here is to model the choice of capacity, volume and time, one of the simplifying assumptions maintained in this subsection is that this process is exogenous. It is also assumed that the information quality is revealed jointly with the transmission of bytes. As an example of a process $\{\theta, t \geq 0\}$, suppose that, for any $t > 0$, the process has independent increments with $d\theta_t|\mathcal{I}_{t-} \sim N(0, \sigma^2)$, $\sigma^2 > 0$, where $\mathcal{I}_{t-} = \bigcup_{s>0} \mathcal{F}_{t-s}$ is the information set up to time t; then, $E[\theta_{t+s}|\mathcal{I}_t] = \theta_t$ and $E[(\theta_{t+s} - \theta_t)^2|\mathcal{I}_t] = \sigma^2 s$, for all $s, t \geq 0$. Let $\{\hat{\nu}_s\}$, $\{\hat{\tau}_s\}$ and $\{\hat{x}_s\}$ denote the sequences of anticipated choices, conditional on the discrete choices $\{\beta_s\}$, where each subscript $s = 1, \ldots, S$, indexes a session defined by a discrete choice and subsequent continuous choices. As a convention, the realization of θ_t that pertains to session s is ascribed to the end of the session, i.e. to time $t_s = \sum_{s'=1}^{s} (\hat{\nu}_{s'}/\beta_{s'} + \tau_{s'})$. Two further strong assumptions that greatly facilitate the analysis are that there is no discounting and that utility is intertemporally separable. Also, the stochastic law governing $\{\theta_t, t \geq 0\}$ is assumed known to the decision maker; this entails the further assumption that no strategic experimentation occurs to make inference about this

[21] Cp. Ingersoll (1987); I thank Hal Varian for pointing me to this literature.

law. Then, omitting the econometric error, the dynamic decision problem is

$$\max_{\{\beta_s,\hat{\nu}_s,\hat{\tau}_s,\hat{x}_s\}} E_{\theta_{t_1}}\left[U(\hat{\nu}_1,\hat{\tau}_1,\hat{x}_1;\beta_1,\theta_{t_1})|\mathcal{I}_0\right] +$$

$$E_{\{\theta_{t_s}\}_{s=2}^S}\left[\sum_{s=2}^{S}\left(U\left(\hat{\nu}_s,\hat{\tau}_s,\hat{x}_s;\beta_s,C_s,\theta_{t_s}\right)-1_{\{\hat{\nu}_s>0,\hat{\tau}_s>0\}}K\right)|\mathcal{I}_0\right]$$

$$\text{subject to:} \qquad C_s=\sum_{s'=1}^{s-1}\hat{\nu}_{s'},\quad C_1=0,$$

$$t_s=\sum_{s'=1}^{s}(\hat{\nu}_{s'}/\beta_{s'}+\hat{\tau}_{s'}),$$

$$\lambda\left(\sum_{s=1}^{S}(p_v(\beta_s)\hat{\nu}_s+p_t(\beta_s)\hat{\tau}_s+p_x\hat{x}_s)-M\right)=0,$$

$$\lambda_{\nu_s}\hat{\nu}_s=0 \text{ for all } s,$$

$$\lambda_{\tau_s}\hat{\tau}_s=0 \text{ for all } s,$$

where λ,$\{\lambda_{\nu_s}\}$ and $\{\lambda_{\tau_s}\}$ are Lagrange multipliers, C_s is the accumulated information capital at the beginning of session s and $K>0$ is a utilitarian switching cost. Notice that the solution to this problem is complicated by the fact that the expectation is taken with respect to the conditional distribution of the θ sequence at the times $\{t_s\}_{s=1}^S$ which themselves are an outcome of the optimization.

The (approximate)[22] first order condition for $\hat{\nu}_1$ and $\hat{\nu}_2$ in the case $S=2$ is, in abbreviated notation,

$$E_{\theta_{t_1^\star},\theta_{t_2^\star}}\left[U_\nu(\hat{\nu}_1^\star)+U_C(\hat{\nu}_2^\star;C_2^\star)|\mathcal{I}_0\right]+\lambda_{\nu_1}^\star=\frac{p_v(b_1)}{p_v(b_2)}\left(E_{\theta_{t_2^\star}}\left[U_\nu(\hat{\nu}_2^\star;C_2^\star)|\mathcal{I}_0\right]-K+\lambda_{\nu_2}^\star\right);$$

where a star denotes candidate optimal values and $b_s=\beta_s^\star$ as above. Notice that $\hat{\nu}_1^\star>0$ implies $C_2^\star>0$ and $\lambda_{\nu_1}^\star=0$. Suppose that $U_\nu(\nu_2;C_2)=U_C(\nu_2;C_2)\geq 0$ almost everywhere[23] and bounded above a.e. by $Q<\infty$ for $C_2>0$; in this case, adding an incremental piece of information to C_2 is as valuable, at the intensive margin, as adding it to $\hat{\nu}_2$. Then, if $\hat{\nu}_2^\star>0$ and therefore $\lambda_{\nu_2}^\star=0$, $p_v(b_2)<p_v(b_1)$ is necessary for the last equality to hold, and so, in light of the monotonicity of the price schedule, the capacity chosen for session 2 must be lower than for session 1 in order to cover the utilitarian switching cost K. Suppose that $1<p_v(\beta^{\max})/p_v(\beta^{\min})=P<\infty$. Then, the condition (†) that, for $C_s>0$, any $t>0$ and any $s=2,\ldots,$ $U(\nu_s;C_s)$ be concave in ν_s and $(P-1)E_{\theta_t}[U_\nu(0;C_s)|\mathcal{I}_0]\leq PK$ is sufficient to force $\lambda_{\nu_2}^\star>0$ and therefore $\hat{\nu}_2^\star=0$. The latter ex ante optimal choice then also implies

[22]The displayed equation, in general, is an approximation, since changing $\hat{\nu}_1$ and $\hat{\nu}_2$ also changes t_1 and t_2 and thereby – through the effect on the variance of θ_{t_1} and θ_{t_2}, for example when $\{\theta_t,t\geq 0\}$ is Brownian motion – has an effect on the displayed expectations, as long as U is nonlinear in θ.

[23]That is to say that the equality holds for almost all values of τ_2, b_2 and all realizations of θ_{t_2}, except possibly on a set of measure zero.

$\hat{\tau}_2^\star = 0$. By induction, $\hat{\nu}_s = 0$ and $\hat{\tau}_s = 0$ for all $s = 2, \ldots S$. Since the condition (†) is preserved by scaling when considering more than merely two successive choices, it follows that it is always optimal to anticipate not to switch. Thus, switching costs that are sufficiently high relative to the expected marginal utility of transmitted volume after a switch prohibit planning to switch. The shadow prices of the expected consumption of the "commodities" $\{\hat{\nu}_s\}_{s=2}^S$, $\{\lambda_{\nu_s}\}_{s=2}^S$, are strictly positive since increasing the amounts consumed of theses commodities from zero to positive levels decreases utility by $(S-1)K$, without expected compensation through "consumption smoothing". Switching between bandwidths can still emerge once θ_{t_1} is revealed, since the first-order Markov property of $\{\theta_t, t \geq 0\}$ implies a revision of expectations of future realizations of θ_t, $t > t_1$, conditional on θ_{t_1}.

The proposed model can then be interpreted as arising from a time-separable continuous time utility model without discounting, built around an exogenous, first-order Markovian process $\{\theta_t, t \geq 0\}$, whose stochastic law is known to the decision maker, that involves switching costs satisfying the condition (†). Even with low switching costs, the decision maker may not plan to switch between capacities if there exists sufficient asymmetry in the way information generates utility – directly, in isolation from the accumulated information capital due to the interruption of the accumulation process as a consequence of the switch, as opposed to creating utility through uninterrupted information capital accumulation.

If any of the rather strong assumption maintained throughout this subsection is relaxed, the optimal policy may be quite different. Exploring such possibilities is left for future work. For now, the model can perhaps best be thought of as ignoring intertemporal strategic trade-offs, focusing on serially independent decision nodes for discrete–continuous choices. The focus on the joint endogeneity of the discrete–continuous choices adopted here, at the expense of a more elaborate serial dependence in a dynamic programming framework, distinguishes this work from Rust (1987,1994) on controlled stochastic processes as the optimal policy of an intertemporal cost minimization problem, the focus of which is the sequence of binary indicators of consecutive discrete investment decisions, abstracting from the joint endogeneity of investment and equipment utilization.[24]

References

[1] Beckert, W. (2000): "On Specification and Identification of Stochastic Demand Models", unpublished manuscript, Department of Economics, U.C. Berkeley

[2] Blundell, R., M. Browning and I. Crawford (1998): "Nonparametric Engel Curves, Revealed Preference and Welfare Bounds on Tax Reforms", unpublished manuscript, Department of Economics, UCL

[24] Cp. Rust (1987), p. 1004-1005.

[3] Brown, B.W. and M.B. Walker (1989): "The random utility hypothesis and inference in demand systems", *Econometrica*, vol.59, p.925-951

[4] Brown, D.J. and R.L. Matzkin (1995): "Estimation of a Random Utility Model from Data on Consumer Demand", mimeo, Yale University and Northwestern University

[5] Brown, D.J. and R.L. Matzkin (1995A): "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand", mimeo

[6] Cover, T.M. and J.A. Thomas (1991): *Elements of Information Theory*, New York: Wiley

[7] Dubin, J.A. (1985): *Consumer Durable Choice and the Demand for Electricity*, Amsterdam: North-Holland

[8] Dubin, J.A. and D.L. McFadden (1984): "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption", *Econometrica*, vol.52(2), p.345 -362

[9] Hausman, J.A., M. Kinnucan and D.L. McFadden (1979): "A Two-Level Electricity Demand Model; Evaluation of the Connecticut Time-of-Day Pricing Test", *Journal of Econometrics*, vol.10, p. 263-289

[10] Ingersoll, J.E. (1987): *The Theory of Financial Decision Making*, Totowa, N.J.: Rowman & Littlefield Studies in Financial Economics

[11] Lancaster, K.J. (1966): "A New Approach to Consumer Theory", *Journal of Political Economy*, vol.74(2), p. 132-157

[12] Lancaster, K.J. (1971): *Consumer Demand: A New Approach*, Columbia Studies in Economics, vol.5, NY: Columbia University Press

[13] Lancaster, K.J. (1979): *Variety, Equity, and Efficiency*, Columbia Studies in Economics, vol.10, NY: Columbia University Press

[14] Lewbel, A. (1996): "Demand systems with and without errors: reconciling econometric, random utility and GARP models", mimeo, Brandeis University

[15] McElroy, M.B. (1987): "Additive General Error Models for Production, Cost and Derived Demand or Share Equations", *Journal of Political Economy*, vol.95, p.737-757

[16] McFadden, D.L. (1974): "The Measurement of Urban Travel Demand", *Journal of Public Economics*, vol.3(4), p.303-328

[17] McFadden, D.L. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration", *Econometrica*, vol.57(5), p.995-1026

[18] Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators", *Econometrica*, vol.57(5), p.1027-1057

[19] Rupp, B, R. Edell, H. Chand and P. Varaiya (1998): "INDEX: A Platform for Determining how People Value the Quality of their Internet Access", in: *Proceedings of the 6th IEEE/IFIP International Workshop on Quality of Service*, p.85-90, Piscataway, NJ: IEEE Press

[20] Rust, J. (1987): "Optimal Replacement Of GMC Bus Engines: An Empirical Model of Harold Zurcher", *Econometrica*, p.999-1033

[21] Rust, J. (1994): "Structural Estimation of Markov Decision Processes", in: *Handbook of Econometrics*, vol.IV, R.F. Engle and D.L. McFadden, eds., Amsterdam, London and New York: Elsevier, North-Holland, p.3081-3143

[22] Varian, H.R. (2000): "Estimating the Demand for Bandwidth", unpublished manuscript, School for Information Management and Systems, U.C. Berkeley

[23] Wilson R.B. (1993): *Nonlinear Pricing*, Oxford: Oxford University Press