The Economics of Internet Traffic Management: An Empirical Prospective

Aviv Nevo

Based on work with Jon Williams, Jacob Malone and John Turner

CRESSE, July 2016

Motivation

- The telecommunications sector is undergoing major changes
- A driving force: growing importance of data services
  - the use of the Internet is growing at astonishing rates
  - residential broadband (BB) usage has grown 30-40% annually over last two decades
  - file-sharing drove early growth, over-the-top video (OTTV) recently
- BB capacity has largely kept pace with demand
  - investment averaged $75 billion per year since 1996 (FCC 2015), or $1.5 trillion over past two decades
  - new private (Google) and public (Muni BB) networks
- Congestion continues to periodically impact networks
  - congestion can happen in several places
  - my focus will be on the "last mile"

Policy Debates

- The rapid change has sparked several policy debates in the US and Europe
  - *Pricing:* Data caps, zero rating, peak-load pricing
  - *Entry:* Google Fiber, municipal broadband
  - *Evolving Choice Set:* OTTV, cord cutting, satellite
  - *Mergers:* Charter-TWC, Comcast-TWC, ATT-DirecTV (horizontal and vertical elements)
  - *Regulation:* net neutrality, opening set-top boxes, expanding residential networks
- Up to now much of the economic debate has been theoretical
- Our goal in this project (spread over several papers) is to bring some facts and empirical analysis to the policy questions

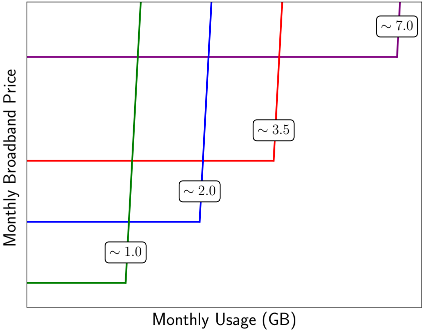
Outline of Talk

- Present descriptive analysis using high-frequency usage data
- Outline model + estimation of demand for residential BB
- Results and implications for traffic management
  - focus on pricing and investment incentives

Data

- Detailed usage data from North American provider
  - Hourly observations at subscriber level
  - Data include bytes down/up and packet drops and delays
  - Covers roughly 45,000 subscribers from Feb–Dec 2015 (334 days)
- Hourly average network (*node*) utilization
- Daily subscriber billing records
  - Includes speed, usage allowance, and price information
- In our data the plan prices are usage-based
  - not common in the US
- Key for the estimation of demand
  - will use shadow-price variation to estimate demand

Internet Plan Features



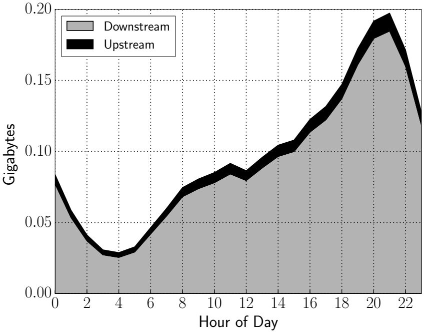
- Plans are 3PTs: access fee, usage allowance, and overage fee
- More expensive plans come with faster speeds and larger usage allowances: complicates identification if only plan choice observed
- Average user pays \$58.89 for 22 Mbps down and a 287 GB allowance (representative for US plans)

Heterogeneity in Daily Usage

	Tier 1	*Tier 2*	*Tier 3*	*Tier 4*	*All*
Mean	1.4 GB	3.4 GB	5.4 GB	8.2 GB	2.3 GB
Std. Dev.	2.9	5.0	7.3	10.4	4.5
25^{th} %tile	0.0	0.3	0.6	1.3	0.1
Median	0.4	1.5	3.1	5.3	0.6
75^{th} %tile	1.5	4.7	7.6	11.4	2.7
90^{th} %tile	4.1	9.0	13.6	19.4	6.7
95^{th} %tile	6.3	12.5	18.5	26.1	10.2
99^{th} %tile	12.8	22.3	32.0	46.2	20.3
N	8,539,830	2,910,234	1,117,680	320,085	12,887,829

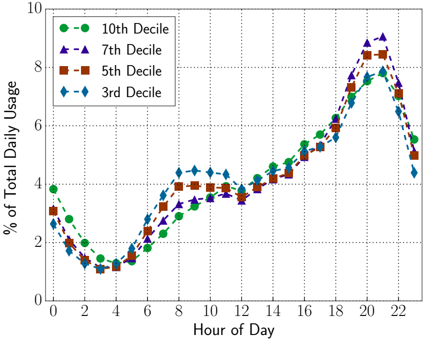
- About 90% of sample are from Tiers 1 and 2
- Strong selection effect into plans, much higher usage on more expensive plans

Average Usage by Hour and Direction



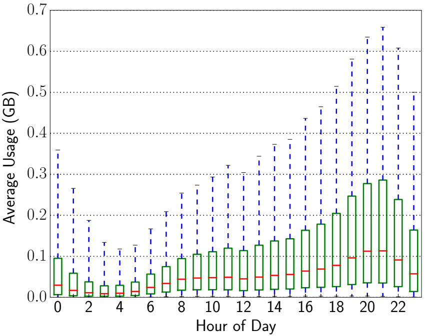
- Construct temporal profile (level) for each user
- Peak average usage 5 times greater than trough, 90% downstream
- Potentially ripe for peak-use pricing or technological advances to more efficiently use network

Hourly Usage by Monthly-Usage Decile



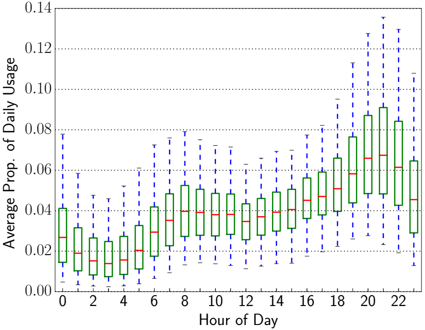
- Construct temporal profile (proportion) for each user
- Average proportion at each hour by monthly-usage decile
- Almost no correlation between level and temporal profile

Heterogeneity in Hourly Usage Levels



- Construct temporal profile (levels) for each user
- Calculate different quantiles at each our hour
- Substantial heterogeneity, median peak usage is only 15% of 95^{th}% user's peak usage, service cost proportional to monthly usage

Heterogeneity in Temporal Usage Patterns



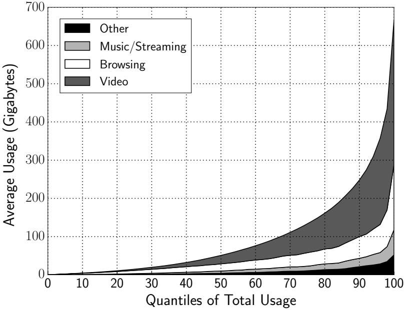
- Construct temporal profile (proportions) for each user
- Substantial heterogeneity in hourly pattern (nearly flat to very peaked), despite no relationship with level

DPI Application Groupings

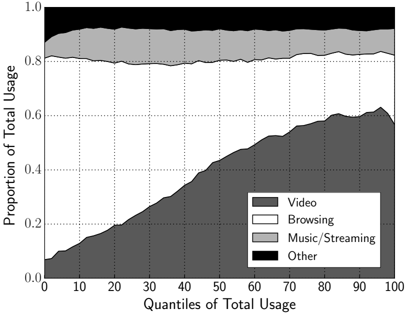
Groups	*Description (Examples)*	*% of All Usage*
Administration	System administrative tasks (STUN, ICMP)	1.19
Backup	Online storage (Dropbox, SkyDrive)	0.58
Browsing	General web browsing (HTTP, Facebook)	26.70
CDN	Content delivery networks (Akamai, Level3)	2.95
Gaming	Online gaming (Xbox Live, Clash of Clans)	3.06
Music	Streaming music services (Spotify, Pandora)	3.40
Sharing	File sharing protocols (BitTorrent, FTP)	0.20
Streaming	Generic media streams (RTMP, Plex)	6.26
Tunneling	Security and remote access (SSH, ESP)	0.07
Video	Video streaming services (Netflix, YouTube)	55.47
Other	Anything not included in above groups	0.13

- Deep-Packet Inspection (DPI) data from nationwide sample of users, different ISP and not used in estimation
- Video and browsing are dominant sources of traffic

Monthly Usage by Quantile and Traffic Type



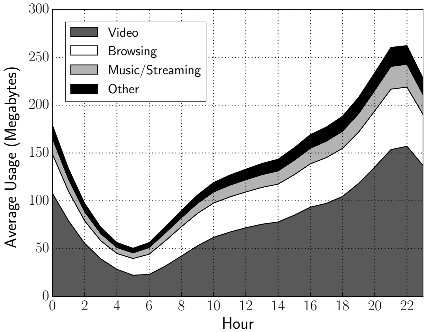
(a) Average Monthly Usage



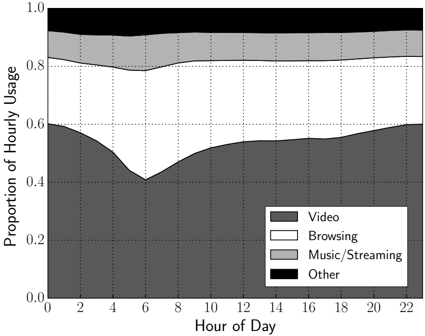
(b) Proportions of Monthly Usage

- Video is a larger proportion of usage for heavier users
- Encouraging for economic and technological solutions to congestion since video is passive activity (download any time to watch later unlike browsing)

Hourly Usage by Group



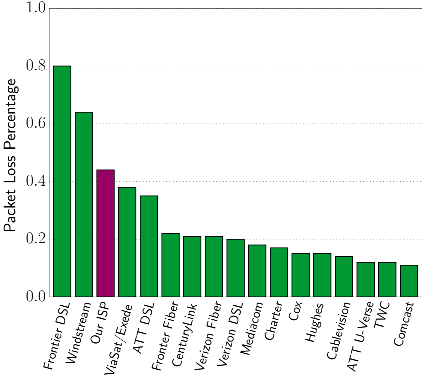
(a) Level: Decomposed Hourly Usage



(b) %: Decomposed Hourly Usage

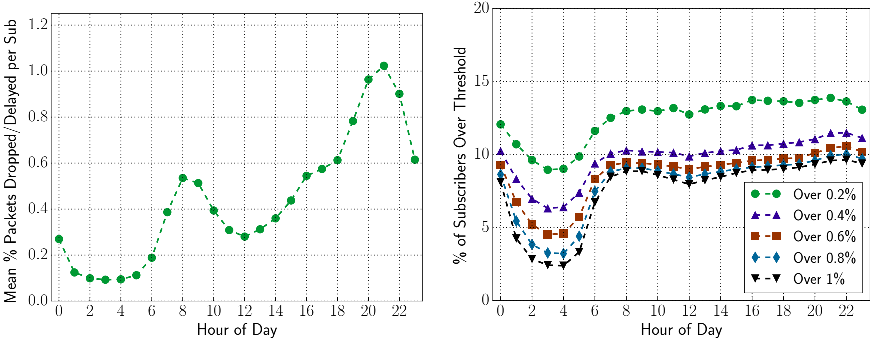
- Pattern across the day is observed in every data set collected
- Video is a more peak-intensive activity, encouraging again for local-cache technology

Congestion: Comparison Across ISPs



- Reproduced from FCC (2015)
- For comparison, we report the packet loss at the average *subscriber-hour* level (third worst ISP)
- Great opportunity with sample to study congestion

Congestion: Packet Loss by Hour

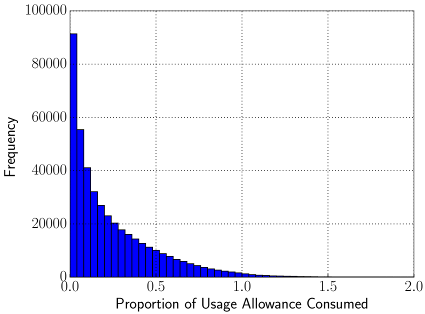


(a) Average Packet Loss

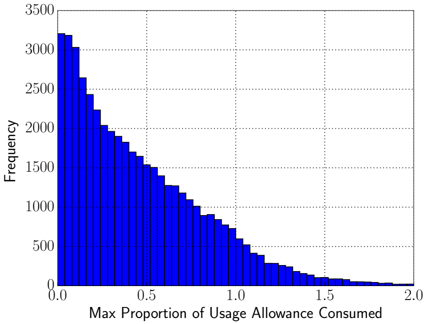
(b) Variation in Packet Loss

- 1% of packet loss is considered very high, average at 9PM exceeds 1%
- Packet loss is highly right-skewed across users
- About 10% of users experience over 1% packet loss during peak hours

Proportion of Monthly Allowance Utilized



(a) Proportion - Subscriber-Month



(b) Max Proportion - Subscriber

- About 2.5% of people exceed allowance (average 40% of allowance used), less shadow-price variation than Nevo et al (2016) so less shadow-price variation
- Over 15% of people exceed allowance at some point during sample, so panel is crucial source of within-user variation

Data: Key Take Aways

- Substantial heterogeneity across users in level and temporal pattern of usage
  - economics: policies that discriminate between consumers might be welfare improving
  - modeling: need a model with rich heterogeneity
- Within day usage pattern suggests peak load pricing
- Dynamics
  - variation in shadow price of usage
  - consumers respond to the variation in shadow price (not shown today)
- Congestion
  - it exists
  - consumers respond to reduction (evidence from node splits, not shown)

Outline of Model

- The goal is to use the variation in shadow price over the month to estimate demand
- Model
  - daily (or even peak/off-peak) usage
  - plan choice
- Need to allow for rich heterogeneity
- Key modeling elements
  - utility from content
  - cost of usage related to value of time (creates satiation)
  - connection speed relates to value of time
  - random shocks
  - plan choice: assumed optimal
  - significant heterogeneity

Sample Model: Utility Function

- Utility from content of consumer type h on plan k (for each day t):

$$u_{hk}(c_p, c_{op}, \psi, \upsilon) = \upsilon_1 \left(\frac{(c_{op} + c_p)^{1-\alpha_h}}{1-\alpha_h} \right) - c_{op}^2 \left(\frac{\upsilon_2 \kappa_h}{\ln(s_k)} \right) - c_p^2 \left(\frac{\kappa_h}{\ln(\psi s_k)} \right)$$

  where: c – GB of content; s_k – connection speed; υ_1 and υ_2 – shocks to preferences for content; ψ – network state
- ψ follows first-order Markov process G_ψ
- $\upsilon_1 \sim EXP(\lambda_1)$ and $\upsilon_2 \sim EXP(\lambda_2)$
- Consumer type $(\alpha_h, \kappa_h, \lambda_{1h}, \lambda_{2h})$: α_h – curvature to utility; κ_h – content wait time/preference for speed; λ_{1h} and λ_{2h} – time-varying shock parameters capture level and temporal variation in usage
- Satiation in usage for finite speeds, proportional rationing of speed, and additive usage determines benefit

Solving the Model

- Usage (conditional on plan choice)
  - consumers solve finite-horizon (T) dynamic program, choosing daily (or peak and off-peak) usage
  - solve for optimal usage on each of the offered plans
- Plan choice
  - identify optimal plan (at $t = 0$) and assume that consumers choose this plan
  - optimal plan assumption assigns each type to a plan

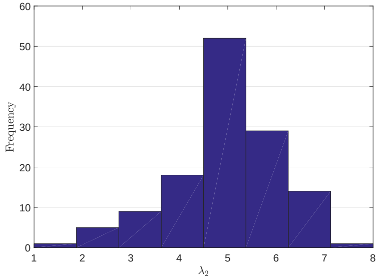
Estimation: Overview

- Estimation procedure follows ideas in Fox et al. (2015)
- Two separable steps:
  1. Computational:
     - type defined by vector, $(\alpha_h, \kappa_h, \lambda_{1h}, \lambda_{2h})$
     - solve dynamic program for large number of types
     - store optimal plan, policy function, and value function for each type
  2. Estimation
     - estimate the distribution of types that best matches the data and the model predictions
     - NTW (2016) use a "random effects" approach: match weighted predicted moments to observed moments
     - MNW (in progress) propose a "fixed effects" approach
     - key: estimation is quick and flexible (the time consuming part is in the first stage)

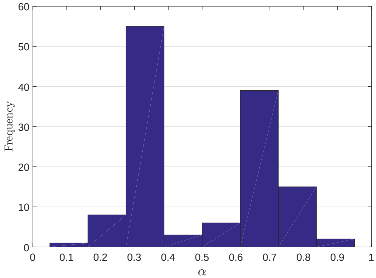
Sample Results: Type Distribution

- Difficult to succinctly summarize results: distribution of types
- 129 types with positive weights (of 4,096 considered)
- Slightly more uniform weights across types than Nevo et al. (2016)
- Model fits the data quite well (over 97% correlation between data and fitted optimal-type's expected behavior)
- Most of the estimated positive types (85) come from top tier
  - Top tier only around 2.5% of sample
  - Wide variety of behavior to explain within tier differences in usage (choose plan for either speed or allowance)

Results: Marginal Type Distributions



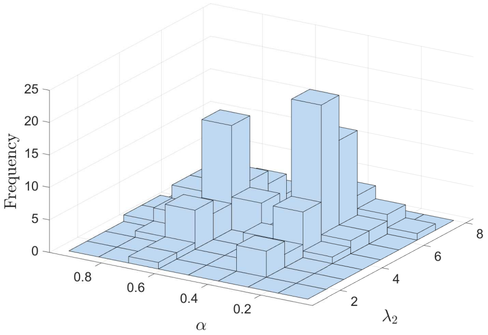
(a) Marginal of λ_2



(b) Marginal of α

- λ_2 distribution nearly normal, intuitive values since off-peak usage roughly 20% on average of peak usage
- α distribution irregular and bimodal, consistent with Nevo et al (2016)

Results: Joint Type Distributions



- Bimodal nature distribution of α clear visible, mixture of two normals might be reasonable
- Weak correlation, such that more price sensitive consumers have flatter temporal profiles (lower λ_2)

Key Findings

- Internet access is valuable
  - NTW: CS from free gigabyte service is 280$/month, 115$/month from typical cable
- Marginal usage is not very valuable
- Mean WTP for one additional Mb/s for one month:
  - NTW around \$2; MNW \$0.79
- Mean (median) WTP for one additional GB of allowance on first day of month is \$0.09 (\$0.04)
  - Implications for zero rating of content: price differential for types of content that count against allowance (e.g., Netflix) and those that don't (e.g. Comcast Stream)
  - Implications for peak load pricing

Counterfactual Analysis: Usage Based Pricing

NTW (2016) find that:

- UBP reduces consumer surplus (relative to unlimited plans), but that the loss is relatively small
  - mostly eliminates low value usage
  - key issue: what plans would be offered
- UBP results in higher revenues
  - mainly because consumers choose higher speed plans (which come with higher allowance)
- The sum of CS + revenue is largely unchanged, but cost is lower under UBP (due to less traffic)
- Data caps seem to be effective in sorting consumers and reducing low value usage

Counterfactual Analysis: Investment in Networks

NTW (2016) generally find that:

- Investment in next generation networks will be socially beneficial
- Large difference between private and social return
    - private firms have limited ability to get infra-marginal gains
- Allowing for more ways to price discriminate might yield efficient investment
- Justification for investment by Google (additional returns) or public sector (Muni BB)

MNW (in progress) find similar results for reducing congestion: consumer gains that are not reflected in ISP revenues

Counterfactual: Local-Caching Technology

- Technology is available, but not deployed, to permit many OTTV services to locally cache content like DVR technologies
- The model predicts large gains, larger than created by investment
- Hardware/development costs surely offset by surplus gains

Counterfactual: Peak-Use Pricing

- OTTV content providers currently have weak incentive to introduce local-caching technologies (quality improvement during peak hours)
- Peak-use pricing introduces right incentive
- Consider very simple form: reduce allowance by varying levels (30% and 50%) but only count peak usage against allowance, overage fee held constant
- Disadvantage is that it only weakly decreases off-peak cost (eliminates shadow price), so reduces overall cost of usage by less than local-cache technology

Counterfactual: Peak-Use Pricing

	Peak-Use Pricing		
Usage and Surplus	*Baseline*	*30% Reduction*	*50% Reduction*
Daily Usage (GB)	2.5	2.6	2.4
Peak Usage (GB)	1.8	1.7	1.5
Off-Peak Usage (GB)	0.7	0.9	0.9
Consumer Surplus ($)	70.22	72.33	69.01
Revenue ($)	57.42	57.21	58.54

- Not too successful in isolation (low intra-day elasticity)
- Small peak allowance leads to small consumer-firm transfer
- Substantially increases shadow price of peak-usage ($0.45 per GB on average), which provides strong incentive for introduction of local-cache technology

Concluding Comments

- Data and modeling quite useful in addressing policy relevant questions
- Still have many unanswered questions
- Consider net-neutrality
  - the economics boils down to who needs better incentives: edge provider or networks
  - on one hand, marginal content had low value; is this the content that would be lost?
  - on other hand, network are congested; but are other tools more effective in helping manage traffic?