Pricing The Last Mile: Data Capping For Residential Broadband

Konstantinos Poularakis
ECE Dept.
University of Thessaly
Volos, Greece

Ioannis Pefkianakis
Technicolor Labs
Paris, France

Jaideep Chandrashekar
Technicolor Labs
Paris, France

Leandros Tassiulas
Electrical Eng. Dept. &
Institute of Network Science
Yale University
New Haven CT 06520

ABSTRACT

In recent years, ISPs in mature residential broadband markets have been moving away from flat-rate pricing schemes to ones that involve usage limits (or "data caps"). These changes have led to a lively debate about the merits of data capping. While previous work has studied the theoretical benefits of data capped pricing models, there is relatively little work that provides insights into exactly how to design data capped tariffs, and their impact on the parties involved, i.e., ISPs and their broadband subscribers. In this paper, we formulate the problem of selecting optimal data capped tariffs based on a simple, extensible model of ISP costs and revenues, and a novel model that captures how users evaluate a set of offered tariffs. We show that the problem is NP-Hard and describe two heuristics that provide reasonable solutions. We also describe an approximation algorithm that yields bounded solutions. We apply this framework on a novel broadband usage dataset collected from real-users and examine the cost-benefit impact of data capping for various cost regimes that reflect different regional markets around the world. We find that in highly competitive markets with relatively high traffic costs, data capping can indeed help the ISP to manage costs while resulting in a more equitable per unit data cost across the users. However, in the absence of meaningful competition or when traffic costs are low, data capping largely benefits the ISP and is to the detriment of subscribers.

Categories and Subject Descriptors

C.2.3 [**Computer Communication Networks**]: Network Operations; C.4 [**Performance of Systems**]: Measurement Techniques, Modeling Techniques

CoNEXT'14, December 2–5, 2014, Sydney, Australia.
Copyright 2014 ACM 978-1-4503-3279-8/14/12 ...$15.00.
http://dx.doi.org/10.1145/2674005.2675008.

Keywords

Broadband Access Networks; Usage-based Pricing; Data Caps

1. INTRODUCTION

Flat rate models, where subscribers pay a fixed monthly price independent of actual usage, have traditionally been the dominant pricing model in many residential broadband markets and this by virtue of appealing to both providers and subscribers. Subscribers find the predictability of a fixed monthly bill compelling and easy to understand. At the same time, service providers benefit by doing away with complicated accounting systems. However, in recent years, a significant number of ISPs are moving away from such models towards those that involve usage limits or "data caps". Here, a subscription plan is associated with a data usage (traffic volume) cap over a time interval (typically the billing period). When the data cap is exceeded, the provider applies an overage policy to the subscriber and this can range from lowering or degrading the service levels, charging block price increments, or even interrupting the service. While usage limits have long existed in certain regions with poor network connectivity (e.g., S.E. Asia, Africa), their adoption in highly connected, mature markets is a recent trend. Some notable examples include Verizon and Comcast in the US, Belgacom and Deutsche Telekom in Europe. The rationalization offered by providers while introducing such pricing plans largely centers around managing increases in bandwidth costs, caused by increasing traffic demands on their networks from online video streaming [1] or BitTorrent traffic. However, these pricing changes have led to a consumer outcry against the practice of data capping [2, 3], arguing that the practice is unfair [1] and hinders new innovative services.

While both sides present valid arguments, the actual merits of data capping, or the lack thereof, depend on a host of factors that vary by region – broadband competition, subscriber behavior, transit and capacity prices, and so on. Despite the fact that there has been considerable work applying economic theory to the issue of data caps, the investigations have either focused on examining the theoretical benefits of data capping [4–7], or quantitatively evaluated the benefits for a pre-specified set of data capping tariffs [8, 9]. There has been relatively less work in understanding exactly how

[1] Some service providers often exclude their own bandwidth intensive Video-on-Demand service from the data caps.

a provider should design a data capped pricing structure, and how the provider (and subscribers) will fare as these behaviors and costs change over time (or from region to region). Such an investigation is critical to cutting through the rhetoric that exists today and informing the current debate on exactly who benefits from the practice of data capping. To carry this out, we need a valid mechanism to identify a set of feasible tariffs, which takes into account the ISP's cost structure the motivation and behavior of subscribers; however, identifying such a mechanism is challenging for the following reasons. First, ISP cost structures are typically confidential and not (easily) available to the research community. Second, subscribers may be unpredictable (or even irrational), which makes modeling behavior difficult at a high level. Lastly, the tariffs must be evaluated on a realistic and sufficiently fine-grained dataset of broadband usage in the context of various regional cost scenarios.

In this paper, we present an initial attempt to tackle these challenges. To address the first, we design the *Tariff Formulation Problem* that marries a simple, extensible model of ISP costs and revenues to a model that captures user utility in a framework that can select a set of optimal tariffs. We show this problem to be NP-Hard. Subsequently, we identify two heuristics and an approximation algorithm that yields bounded solutions. The TFP framework enables a quantitative analysis of the impact of introducing data capping on both providers and subscribers. To overcome the last challenge, we leverage a dataset of detailed traffic usage data obtained from 260 subscribers of a large European ISP. Collating traffic pricing information from several sources, we synthesize a variety of regional markets and investigate the impact of data capping in each. Our work in this paper combines strong theoretical contributions with a careful empirical data analysis from real ISP subscribers. The results we present provide valuable insights in the current data capping debate and clearly differentiate situations in which data capping can indeed help users, and those in which the imposition of data caps largely benefits the ISPs bottom line at the cost of subscribers. The main contributions of this paper are summarized as follows.

1) We analyze a dataset of 260 ISP subscribers and investigate the merits of data capping (which limits data volume) as a mechanism to reduce traffic rates (which affect network traffic costs). We find that aggregate data usage of subscribers is proportional to their contribution to the ISP's peak rates (and costs). This establishes that under certain conditions, moving to a data capped regime can lead to transit and capacity cost reductions for the ISP (§4).

2) We study the problem of determining optimal tariffs (price, speed, data cap) for an ISP. Using simple and intuitive models of user behavior and ISP costs & revenue, we formulate the *Tariff Formation Problem* (TFP) and show that it is NP-Hard. We present a set of three algorithms, one of which is an approximation algorithm that yields solutions with bounded performance (§5).

3) We examine various cost scenarios reflecting various regional markets around the world and, evaluate the impact of moving from flat-rate pricing to a data capped model. We find that ISPs are able to use data capping to indirectly reduce network costs only in markets with relatively high transit and capacity prices. In the best case scenario – with traffic prices near $1€/\mathrm{Mbps}$, and with 8 tariffs – an ISP can improve their bottom line by 23% by instituting caps, and with a small net change in subscription revenues (less than 5%). In competitive markets where traffic prices are low, ISPs gain relatively little from instituting data caps (they have more to lose from customers departing). On the other hand, a lack of competition reduces the probability of subscribers departing, which allows the ISP more freedom in instituting caps – even when traffic prices are low (§6).

2. RELATED WORK

There is a large body of existing work that has investigated usage-based pricing – data-capped models are a specific instance – from the perspective of economic theory [4–7, 10–13]. The general consensus across these works is that, in theory, data caps are overall beneficial to both service providers and subscribers: they can raise the ISP's revenue, alleviate cross-subsidization across services, and potentially reduce network congestion. In our work, we take these benefits as theoretical possibilities and seek to assess, quantitatively, how the ISP is affected by adopting data caps and by exploring a number of cost scenarios that reflect markets worldwide. In order to do so, we first construct a framework to actually determine the appropriate data caps for a set of cost structures (and a subscriber population). Our work is also novel in that we perform an empirical evaluation using a fine-grained broadband usage dataset from real ISP subscribers.

There have been a few efforts related to ours that provide an empirical analysis of data capped pricing plans in residential broadband markets [8, 9, 14]. However, our own work is different in the several respects. The work in [8, 9] evaluates the effect of a given set of data caps that are taken as inputs; in contrast, one of the main contributions of our work is a mechanism to design *optimal* data capped tariffs, which we then evaluate on a number of different market conditions. In contrast to the datasets presented in [8, 9], the dataset used in this paper spans different access technologies and speed tiers, allowing us to contrast behavior across these populations. Finally in [14], the authors propose a scheme where access-*rates* are capped at certain times, rather than aggregate data usage. While such an approach is potentially interesting, it has not yet been adopted by any ISPs to date.

While data capped pricing plans have recently become popular in mature markets, they are the most common offerings of almost all mobile cellular providers today [6], possibly due to spectrum scarcity, and very high costs of deploying new infrastructure. In [15], it is shown that usage limits on SMS, voice and data can inflate the difference between revenues and costs by a factor of 2 without a significant erosion of the subscriber base. A more recent work has also shown that a wireless provider obtains a higher revenue with usage-based (or flat-rate) time-dependent scheme if the capacity is insufficient (or sufficient) [16]. Finally, previous work has looked at the issue of pricing in Internet transit markets [17]. In our own work, apart from the focus on a different market (residential broadband), we go beyond setting tariff prices and extract tariff sets which requires an optimal distribution of data caps across speed tiers in conjunction with fixing the price.

3. BACKGROUND

In this section, we first provide a very brief overview of the residential broadband market and the popular pricing models seen today. Subsequently, we describe the broadband usage dataset of real users that we leverage to understand the impact of data caps.

3.1 Pricing in Residential Broadband

There are three dominant access technologies for residential broadband available today – ADSL2, Cable and Fiber. ADSL and its variants reuse legacy twisted pair copper wire meant for fixed line telephony. The latest ADSL2+ deployments offer capacity up to 24/1Mbps (downstream/upstream), but the speeds degrade quickly with increasing distance from the and performance is adversely impacted by distance from the local point of presence. Newer variants, e.g., VDSL offer significantly higher capacity ($\sim$100 Mbps) by shortening the local loop (and using alternative technologies for back haul). Cable broadband networks deliver data over the coax cable TV infrastructure (the access network is shared by subscribers behind a cable *head end*), offering download speeds of up to 42.88 Mbps, with emerging standards promising rates upwards of 100 Mbps. Finally, residential Fiber Internet service covers a range of technologies that transmit data over fiber optic cables at speeds that exceed the other technologies.

Independent of the access technology, there are two dominant pricing models in existence. With **flat-rate** pricing, the ISP charges a fixed, typically monthly price, allowing for *unlimited* use by the subscriber. Such flat-rate models have so far been predominant in residential markets (particularly in the western hemisphere). In contrast, **data capped** pricing models associate a subscription price with a usage limit (the data-cap), and apply an overage policy when this limit is exceeded. To illustrate one such policy, AT&T, Comcast and MediaCom charge increments of $10 (for each 50GB of data used above the cap). Data caps are typically specified monthly; however, some ISPs (e.g., CableOne) are even experimenting with daily usage limits. In both models, it is common for providers to offer several *speed tier* offerings (downstream/upstream *speed* limits) at different prices. Often, ISPs offer several data cap choices for a given speed tier (with suitable price differentiation). Finally, additional services – VoIP, Television, Home Automation, etc. – are often *bundled* together with the basic broadband connectivity service. Subscribers enjoy a small cost benefit from the bundling and in return, increase their affinity to the ISP (which benefits the latter).

3.2 Dataset Description

The broadband usage dataset used in this paper was collected from 260 residential gateways over six months of normal operation (Oct. 2013 - Mar. 2014). Each subscriber gateway[2] corresponds to a single billing address, and thus a single home network. The 260 subscribers in our dataset were all enrolled in triple-play service offerings (i.e., Broadband, IPTV + VoIP), but vary in the access technology and speed tier. The subscribers in our dataset were part of a voluntary trial deployment wherein an additional software module was added to the ISP provided home gateway. This module extracted a set of low level traffic usage counters that were timestamped and sent to a back end server approximately every minute where it was recorded into a database. Among other things, most relevant to this paper, the counters recorded the number of incoming (and outgoing) bytes and packets seen at the home gateway to WAN interface.

While the subscribers in our dataset were enrolled in triple-play packages, our empirical analysis only focuses on the Internet broadband service, so the prices we consider are based on what the unbundled service would have cost a subscriber. Table 1 presents a summary of the subscriber population in our dataset, across the different service tiers; the adjusted monthly monthly subscription price is indicated with each tier. Our dataset has the following salient properties: (i) it is temporally fine-grained (1 minute samples), which allows us to understand traffic demand – *individually and in the aggregate* – over short time scales, (ii) by extending to six-months, it is unlikely to be biased by any short term variations in subscriber behavior (vacations, seasonal effects, etc.), (iii) it includes subscribers from different tariffs and access technologies, allowing us to contrast behavior across these populations. While our dataset captures behavior from a relatively small population (the actual ISP has in excess of a million users), we do see a great deal of variability in data usage across subscribers (even within the same tier). Furthermore, we compare our dataset against a much

[2]We use "subscriber", "home", "gateway" and "user" interchangeably.

Access	Speed Tier(r_t) Mbps	Monthly Price(p_t) €	% Subscribers
ADSL	24/1	21.94	34.8
Fiber	30/3	25.20	5.3
Fiber	100/10	33.33	59.9

Table 1: Tariff plans for all the subscribers in our dataset. The ISP currently only offers flat-rate tariffs.

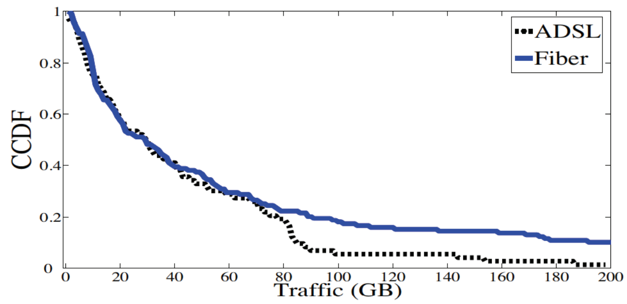


Figure 1: Tail Distribution of Monthly Data Usage for Dec. 2013.

larger dataset, which has been made available by FCC Measuring Broadband America program [18]. In Figure 1, we plot the tail distribution of monthly data usage (in gigabytes) seen across the 260 subscribers for one particular month (we observed similar statistics for the remaining months). The shape of the distribution closely resembles the one presented in the FCC study (cf. chart 24 in [18]). Overall, the FCC statistics indicate that the data usage is slightly higher than observed in our dataset; we believe this can be explained by the penetration of online video streaming services in the US (Netflix, Hulu, etc.), which did not exist in Europe at the time of this study.

4. DO DATA CAPS WORK?

Putting aside some of the arguments about why data capping is finding favor with ISPs, does it even work as a mechanism? That is, can an ISP reduce traffic costs by imposing data caps on users. At first glance, it does not appear to be the case. On the one hand, traffic costs, particularly at transit and peering links are generally tied to *rates*. For example, transit (traffic) payments are based on the 95%-ile of rates, computed over 5 min. intervals, seen over a month (we denote this P_{95}). Capacity upgrades are often triggered by sustained, upward shifts in the peak traffic, i.e. the 100%-ile of the traffic rates in the ISP's network. On the other hand, data caps have the effect of limiting the aggregate traffic volume of the ISP over a *long period* (usually a month). Thus, instituting data caps may not necessarily lower the transit and capacity cost. Moreover, it may not target the right users. Consider two subscribers with the same monthly data usage, but very different behaviors. One of them may create traffic *only* in the peak times and thus bears a higher traffic cost than the other. Data caps penalize them both!

To understand the prevalence of such cases, we examined the correlation between a subscriber's monthly data usage and its impact on the traffic costs of the ISP. The latter metric is computed as follows: we first aggregate traffic across all the users and compute the P_{95} statistic for each month. Then, we identify all the 5 min. intervals where the aggregate traffic rate exceeds P_{95}. We call these the *peak traffic* windows and the peak traffic volume is the volume of traffic across these windows. Each subscriber's traffic cost contribution is simply the ratio of the traffic volume that she generates (in the peak traffic windows) to the peak traffic volume of the ISP.

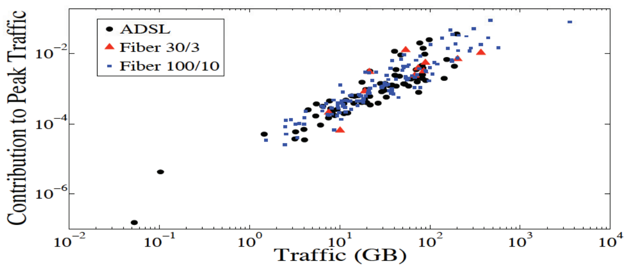


Figure 2: Relationship between subscriber's (monthly) data usage and contribution to ISP peak traffic (and costs).

Figure 2 is a scatter plot across the subscribers in our dataset which shows the subscriber monthly data usage (on x-axis) against their traffic cost contribution for the month of Dec. 2013 (on y-axis). Surprisingly, we find a strong linear relationship that extends several orders of magnitude, and across *all* the different service offerings. We obtained similar results in all the other months and we omit these due to a lack of space. As captured in Fig. 2, the more data a subscriber consumes over the month, the higher his contribution to the traffic cost becomes. This relationship may be explained by the fact that most subscribers (typically) tend to use their devices at particular times of the day (evenings, nights and weekends) and these periods coincide with the peak traffic windows. While this relationship might not hold for some kinds of application usage (for e.g., heavy bit-torrent users), it is likely to apply for typical application usage. Importantly, video streaming – which drives much of the traffic demand today – requires user engagement with the device, and this can happen only at given periods in a day. The takeaway from this analysis is that imposing data caps would on average limit the monthly data usage of a subscriber, and consequently, his contribution to the traffic cost. Thus, under certain circumstances, implementing data caps could possibly reduce the traffic cost burden of an ISP (this can be a significant cost).

Figure 2 also comments on the *unfairness* inherent in flat-rate pricing. The monthly data usage spans two orders of magnitude across subscribers. Importantly, this holds inside each service class, particularly in Fiber/100/10 and ADSL/24/1, which are the dominant populations. This wide disparity in usage, even as all users in a service class pay the same monthly price (with flat-rate), implies that subscribers bear different "costs" for the ISP; typically, heavy users cost the ISP more than light users. In order not to suffer from having to service the heavy users, the ISP has two choices: (i) increase the costs uniformly so that no subscriber leads to a negative gain for the ISP, or (ii) spread the higher cost from heavy users across the lightweight users. The former approach risks losing customers while the latter is inherently unfair. Clearly, flat rate pricing cannot lead to a scheme that is fair and satisfactory to all parties. In contrast, a pricing model with differentiated data caps, i.e., where subscribers with a type of service are offered multiple data caps (at different prices), reduces the inequity across subscribers.

Summarizing, in circumstances that cover most kinds of subscriber application usage, data caps can be a feasible mechanism to reduce the traffic volume, and also the traffic rates needed to be supported by the ISP. However, the overall *cost* benefit is likely to depend on a number of factors – subscriber behavior, bandwidth unit costs, ISP competition, and so on. In the next section, we introduce a model and a tariff computation framework that allows us to understand the cost impact of imposing data caps over a range of traffic pricing scenarios.

5. DETERMINING DATA CAPS

We formulate the *Tariff Formation Problem* (TFP), which takes as input (i) a set of subscribers with their current tariff prices, (ii) a set of cost parameters and (iii) a *target* number of tariffs and which returns a *set* of tariffs, each defined as a triple (speed tier, data cap, price). We show that TFP is NP-Hard and describe two heuristics and an approximation algorithm. The solutions obtained reflect viable choices for an ISP trying to balance subscription revenue with network costs. However, in real life, an ISP may have to balance a number of other considerations – some very long term – when selecting tariff offerings and so we consider the tariff solutions we obtain to establish a lower bound.

5.1 The ISP Model

We model an ISP that has a set of subscribers, $\mathcal{N}$. Each subscriber $n \in \mathcal{N}$ selects a particular tariff t from a set of tariffs, $\mathcal{T}$, where $T = |\mathcal{T}|$. Each tariff t represents a triple (p_t, r_t, x_t), where p_t is the monthly subscription price (Table 1, column 3), r_t is the speed tier (Table 1, column 2), and x_t is the associated data cap. The speed tier r_t is selected from a *finite* set of tiers, $\mathcal{L}$. Finally, x_t is selected from a finite, possibly large set $\mathcal{X}$. We make two simplifying assumptions: (i) subscribers do not exceed the data caps associated with the tariff selected (this allows us to ignore the effect of ISP overage policies); (ii) a subscriber can select any one of the offered tariffs, even though in practice all the offerings may not be available to them (e.g., fiber access may not be available in rural areas).

The (network) cost of serving traffic in a residential broadband network has two main components: (i) *capacity* and infrastructure costs, which includes all costs related to networking equipment, last mile access, power, etc., and (ii) *transit* costs, paid for the fraction of traffic going through upstream transit providers. In general, there are a number of other considerations that may come into play – sunk costs relating to last mile infrastructure upgrades, building and personnel costs, and so on. However, we expect these costs to be less dynamic and managed on far slow(er) timescales (over years). On the other hand, the transit and capacity costs are affected each month by the behavior of subscribers. Our model mainly focuses on these, noting that the sunk costs can be captured by an additional constant factor (when amortized over a long enough period of time).

We then define the residual (or balance) of an ISP, denoted J, as the net difference between incoming subscriber revenue and the outgoing costs associated with serving the subscribers.

$$J = \sum_{t \in \mathcal{T}} (p_t \cdot D_t) - C(\mathbf{D}) - G(\mathbf{D}) \qquad (1)$$

D_t is the number of subscribers that choose $t \in \mathcal{T}$ at a price p_t. Thus, $p_t \cdot D_t$ represents revenue from the single tariff t. We denote the demand vector across tariffs as $\mathbf{D} = (D_t)$. The functions $C(\mathbf{D})$ and $G(\mathbf{D})$ capture the capacity and transit costs, respectively. In the following, we first discuss a way to model the demand vector $\mathbf{D}$ and subsequently describe the cost functions for transit and capacity.

5.1.1 Subscriber Demand

Previous efforts at modeling subscriber utility have considered speed tier and data caps *individually*. For example, the work in [19, 20] considers subscriber utility to be logarithmically related to the speed tier and this is based on the principle of diminishing returns. Users that stream movies regularly are able to watch higher quality videos with higher access speeds, but only up to a point (the video bit rate is bounded); after this point, the utility increases very

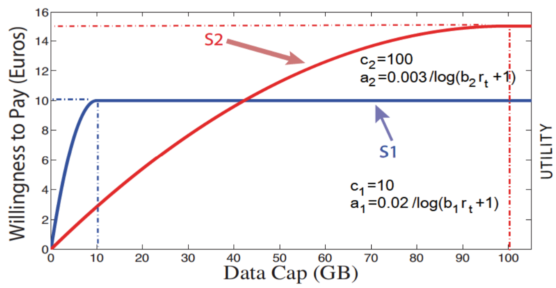


Figure 3: Utility function for two users (S1, S2) realized for a single speed tier r_t.

slowly or not at all. Furthermore, users that carry out very different activities (e.g., video streaming and web browsing) are like to value high speed tiers very differently (web browsing quality is not improved with higher access speeds). Considering data caps, the work in [9, 21] considers utility to be quadratically related to them until a certain saturation point. Intuitively, each subscriber typically has a finite (possibly large) set of activities that she may wish to carry out over a period of time – watch movies or videos, stream music, video chat with friends and family, and so on. These activities, which are limited by the time available to the subscriber, and can be associated with a certain data need. Having data caps that are significantly larger than this *need*, is not likely to be more valuable to the user. In our work, we combine these two factors, speed tier and data caps, and construct a utility function governed by three user parameters: (i) a scaling parameter a_n that captures how different people evaluate a fixed amount of resource, (ii) a parameter b_n that specifies the slope of the logarithmic function for the speed, and (iii) a parameter c_n that determines the saturation point for data usage.

Based on these, we define the utility of user n for tariff t as:

$$U_n(t) = \begin{cases} a_n \cdot \log(b_n \cdot r_t + 1) \cdot (c_n \cdot x_t - \frac{{x_t}^2}{2}) & : 0 \leq x_t \leq c_n \\ a_n \cdot \log(b_n \cdot r_t + 1) \cdot (\frac{{c_n}^2}{2}) & : x_t > c_n \end{cases} \tag{2}$$

For a fixed speed tier r_t, utility increases with the data cap x_t, until $x_t = c_n$, and flattens out beyond this value. Moreover, $U_n(t) = 0$ if $r_t = 0$, since all users need *some* bandwidth (even if small). Similarly, for a given cap x_t, utility increases as r_t increases and where the shape of this increase is controlled by the parameter b_n. To illustrate this, we depict the utility functions for two subscribers S1 and S2 (for a single speed tier, r_t) in Figure 3. Here, S1 and S2 have identical values for b_n and only differ in the parameters a_n and c_n. Here, based on their respective utility functions, S1 is willing to pay at most 10€ for a 10GB cap while S2 is willing to pay at most 15€ for a 100GB cap. If the ISP were to impose a data cap less than 10GB (for S1) and 100GB (for S2), their individual utilities would decrease. In order to assess demand for a particular tariff, users are assigned to particular tariffs, and we consider two distinct families of demand functions, *deterministic* and stochastic *logit*, that allow D_t to be reconstructed for every $t \in \mathcal{T}$.

Deterministic demand model: In this model, a subscriber $n \in \mathcal{N}$ is interested purely in maximizing his *surplus*, defined as $U_n(t) - p_t$. The surplus is the gap between how much a certain tariff is worth to the user and how much the ISP charges for it. Thus, subscriber n will pick tariff t as $\arg\max_{t \in \mathcal{T}}(U_n(t) - p_t)$.

If the surplus is negative over all offered tariffs, i.e., the price of every offered tariff is higher than the subscriber's willingness to pay, then the subscriber leaves the ISP. To simplify the notation when dealing with a large set of subscribers $\mathcal{N}$, we define the binary decision $q_{n,t}$, such that $q_{n,t} = 1$ if subscriber n selects tariff t, and 0 otherwise. Then, $\mathbf{q} = (q_{n,t} : n \in \mathcal{N}, t \in \mathcal{T})$ is the tariff selection vector.

There are two shortcomings with the deterministic demand model. First, it assumes that subscribers always are rational and driven by self-interest which may not be the case in real life. For e.g., a subscriber may desire to stay with a higher priced tariff even if equally suitable, cheaper tariffs are available [4], or else may choose to stay with the ISP even if there is no tariff with positive surplus because the cost of switching ISPs is high [22]. Second, the deterministic model does not capture ISP competition. In non-monopoly markets, a subscriber may depart for a competitor if the price or convenience is better than the status quo. The *logit* model, widely used in econometric demand estimation [23] and also applied previously towards modeling Internet transit pricing [17], can model such uncertainties.

Logit demand model: This stochastic model defines a probability distribution over the set of offered tariffs, where the probabilities depend on the individual utility functions. Thus, the probability that subscriber n selects the tariff t is given as:

$$s_{n,t} = \frac{e^{u_n(t) - p_t}}{\sum_{t' \in \mathcal{T}} (e^{u_n(t') - p_{t'}}) + c} \tag{3}$$

Note that unlike in the deterministic model, a subscriber could have a non zero probability associated with more than one tariff. In Eq. 3, c is non-negative constant that incorporates the notion of competition in the ISP market. In a monopoly market, $c = 0$ and $\sum_{t \in \mathcal{T}} s_{n,t}{=}1, \forall n$, and each subscriber will pick one of the offered tariffs. However, when ISPs compete, $c > 0$ and $\sum_{t \in \mathcal{T}} s_{n,t} < 1$, the probability that the subscriber n will leave the ISP is given by $1 - \sum_{t \in \mathcal{T}} s_{n,t}$. It should be pointed out that c does not count the number of competing entities, but simply captures the level of competition. In general, higher c is associated with a larger risk of a subscriber departing to a competitor. Finally, with the *logit* model, we describe demand for tariff t simply as the sum of probabilities over subscribers: $D_t = \sum_{n \in \mathcal{N}} s_{n,t}$.

5.1.2 *ISP Network Costs*

We now describe how we model and estimate the actual cost functions for transit and capacity. Note that in each case, we take the unit price(s) as parameters; we elaborate on this further in Section 6.

Transit Cost: Of the total traffic generated by the subscribers of a residential ISP, a fraction $0 \leq f \leq 1$, goes from the ISP network towards one of its transit providers; these essentially sell connectivity to the rest of the Internet. The particular value of f depends on several factors such as ISP size, geographic area, user demographics, and so on. For example, f is around 0.3 in the case of some Japanese ISPs, while it can be as large as 0.8 in parts of Africa [24]. Transit providers charge their customers on the P_{95} rate and with a unit price of t_p €/Mbps that is negotiated between the parties. Stanojevic et al. observed an increase in P_{95} with measured total data usage up to a saturation point [24]. This may be explained by noting that increased cumulative traffic usage is likely to be distributed in time, and contribute marginally less to P_{95} which depends on volumes during peak periods. With this insight, we can model the transit cost as a concave function of the total data usage of subscribers. Given that the total traffic volume of the ISP is bounded by the sum over all subscriber caps, we consider that transit traffic is a concave function of the data caps. One such function is the following:

$$P_{95} = f \cdot d \cdot log(\sum_{t \in \mathcal{T}} D_t \cdot x_t) \quad (4)$$

Here, d is a positive (scaling) constant that relates volume to rates. We estimate this term by extracting the summation term in Eq. 4 for each month of data in our dataset, and fitting this to the values of P_{95} which is also extracted from the data. Finally, the network transit cost function is:

$$G(\mathbf{D}) = t_p \cdot P_{95} \quad (5)$$

Network Capacity Costs: ISPs typically provision core network capacity to satisfy peak demand (with a small over-provision factor) to ensure that utilization stays under a threshold. Based on anecdotal evidence and informal discussions with some providers, we believe this threshold to be between $50-70\%$. As traffic demand shifts upwards over time, the utilization starts to approach the threshold which forces the ISP to carry out capacity upgrades and network expansion in the core. Note that this is quite different from last mile upgrades, which tend to be expensive and are planned a long time in advance. We denote the average utilization as v, and the unit cost to carry traffic as c_p. Note that this is pegged to the peak traffic level [13, 25]. Since the total network traffic is bounded by the cumulative data caps, the capacity cost is likely to scale proportionally with the latter:

$$C(\mathbf{D}) = \frac{c_p}{v} \cdot g \cdot \sum_{t \in \mathcal{T}} D_t \cdot x_t \quad (6)$$

where g is a positive (scaling) constant and is again estimated by curve fitting the relevant variables using data from several months.

5.1.3 Tariff Formation Problem

We now formulate the exact problem of designing data capped tariffs, using the model expressed in Eq. 1 and the utility function described in Eq. 2. Our goal is to identify a finite set of tariffs which maximizes the balance in Eq. 1, for a given set of per-subscriber coefficients a_n, b_n and c_n. First, we describe the formulation assuming the *deterministic demand* model. Subsequently, we show how this can be extended to the stochastic *logit* model. The *TFP* is expressed as:

$$\max_{\mathbf{p},\mathbf{r},\mathbf{x},\mathbf{q}} \sum_{t \in \mathcal{T}} (p_t \cdot D_t(\mathbf{q})) - C(\mathbf{D}(\mathbf{q})) - G(\mathbf{D}(\mathbf{q})) \quad (7)$$

$$\text{s.t. } q_{n,t} \cdot (U_n(t) - p_t) \geq q_{n,t} \cdot (U_n(t') - p_{t'}), \forall t' \neq t, n \in \mathcal{N} \quad (8)$$

$$q_{n,t} \cdot (U_n(t) - p_t) \geq 0, \forall n \in \mathcal{N} \quad (9)$$

$$\sum_{t \in \mathcal{T}} q_{n,t} \leq 1, \forall n \in \mathcal{N} \quad (10)$$

$$p_t \geq 0, r_t \in \mathcal{L}, x_t \in \mathcal{X}, q_{n,t} \in \{0,1\}, \forall t \in \mathcal{T}, n \in \mathcal{N} \quad (11)$$

Here, $\mathbf{q}$ is the tariff selection vector. Eq.8 ensures that subscribers pick tariffs that maximize their surplus (as prescribed by the deterministic model) and Eq.9 guards against the possibility of selecting a tariff that yields negative surplus (which implies the subscriber leaves the ISP). Finally, Eq.10 forces subscribers to pick at most one tariff. The remaining constraints defined in Eq.11 constrain the feasible ranges for the optimization variables. For convenience, we denote the solution to TFP as OPT(TFP).

The optimization setup for the *logit* model is largely similar. The tariff selection vector, with the associated constraints defined on it, are omitted and D_t is redefined based on the distribution realized by $s_{n,t}$. Given the similarity, we omit a complete description.

Note that TFP contains both continuous (e.g., p_t) as well as integer (e.g., r_t,x_t) variables and this leads to a non-linear optimization problem which usually lack efficient solutions. By mapping it onto the well known Maximum Independent Set Problem (MISP) problem, we formally prove that it is NP-Hard; the proof is provided in the Appendix. Besides, since it is NP-Hard to approximate MISP within an exponential to the number of vertices *multiplicative factor* away from the optimal solution, the mapping indicates the inapproximability of the TFP problem as well. Thus, it is impractical to try to obtain optimal solutions when the subscriber populations are large. A common approach for solving integer programming optimization problems is to develop the *linear relaxation* of the original problem which is more tractable. In the next section, we derive this form and then describe two heuristics based on it. Since the linear relaxation & rounding procedure may deteriorate the quality of the solution of the above heuristic algorithms in an uncontrollable manner, we also present a third algorithm of different philosophy that produces solutions which are provably within an *additive gap* from the optimal one. We refer to the latter as an approximation algorithm.

5.2 Heuristic and Approximation Solutions

To obtain the linear relaxation (LR) form of TFP, we consider $\mathcal{L}$ as an ordered set and convert the integer variable r_t to a binary vector $(r_{t,l})_{l \in \mathcal{L}}$: here, $r_{t,l} = 1$ if the speed-tier of the tariff t is the l-th element in $\mathcal{L}$, and $r_{t,l} = 0$ otherwise. Additionally, we restrict each tariff to be associated with a single speed tier by imposing an additional constraint, $\forall t, \sum_{l=1}^{|\mathcal{L}|} r_{t,l} = 1$. The LR form of TFP, which we denote LTFP, is obtained by relaxing the constraints on x_t and $r_{t,l}$ and allowing them to take on continuous values. LTFP can be solved with computational efficient algorithms (e.g., branch and bound methods), and we denote the optimal solution to this (relaxed) formulation as OPT(*L*TFP). Note that, by ignoring the integer constraints, OPT(LTFP) can often produce infeasible solutions that overestimate OPT(TFP) which is hard to compute. Thus, it offers an useful comparison point for algorithms that yield feasible solutions, and we describe two such approaches in the following.

Naive Algorithm: This heuristic starts by obtaining a solution to OPT(LTFP) and the greedily rounds all the speed tier variables in one step. It operates as follows:

1. Solve OPT(LTFP)
2. For each t, picks the speed tier with the highest (fractional) solution r_{tl}, $\forall l$
3. Solve OPT(LTFP) a second time, but restrict the r variables previously picked to be equal to 1, and zero elsewhere
4. Rounds each of the computed x values to the nearest value in $\mathcal{X}$, and assign each subscriber n to the tariff t', which satisfies

$$t' = \arg\max_t (U_n(t) - p_t), \text{if } U_n(t') - p(t') \geq 0$$

Sequential Rounding algorithm (SR): This heuristic operates by assigning tariffs to the speed tiers one by one, at each step taking into account the subset of tariffs already assigned to a speed tier in previous iterations. Thus, it carries out T iterations, successively building up each tariff, and each iteration proceeds as follows:

1. Solve for OPT(LTFP).

2. Pick the pair (tariff t',speed tier l') with the highest (fractional) solution among all the r_{tl} $\forall t, l$
3. Set $r_{t'l'} = 1$ and $r_{t'l} = 0$ for all other levels $l \neq l'$
4. Restrict the solution to the LTFP at the next iterations to preserve the above rounded (binary) values of r.

After T iterations are completed, all the r_{tl} variables will be restricted to a binary value. Then, we again solve the restricted $LTFP$ and round the computed x values to the nearest feasible value in $\mathcal{X}$ and assign subscriber n to the tariff t', where

$$t' = \arg\max_t (U_n(t) - p_t), \text{if } U_n(t') - p(t') \geq 0$$

In practice, both *Naive* and *SR* lead to feasible solutions that *can* be close to the optimal, but this is not guaranteed. Such a property might be unsatisfactory for an ISP diligent about reducing costs (via data caps). To this end, we describe a hierarchical algorithm which provides solutions with tunable approximation ratios. The following description assumes the *deterministic demand* model.

Profit Participation Algorithm (PPA): This algorithm is inspired by the model described in [26] and works by solving OPT(TFP) for a carefully chosen subset of the subscribers. To begin with, PPA defines $\mathcal{Y}$ as the cartesian product of $\mathcal{L}$ and $\mathcal{X}$ i.e., all the speed tier, data cap combinations and denotes p_y as the price associated with $y \in \mathcal{Y}$. Next, it defines a distance metric between subscribers which is given as:

$$d_{n,m} = \max_{y \in \mathcal{Y}} |U_n(y) - U_m(y)|, \quad \forall n, m \in \mathcal{N}$$

PPA takes the set of subscribers, their respective data usage(s), and two parameters $\epsilon > 0$ and $0 \leq \tau \leq 1$, and operates as follows:

1. The set of subscribers $\mathcal{N}$ is partitioned such that for n, m in the same cluster, $d_{n,m} \leq \epsilon$. In other words, the clustering assigns subscribers with similar utility to the same cluster.
2. A single "stereotype" subscriber, denoted n_i, is selected from each cluster i. Thus, the set $\{n_i\}$ represents the set of all stereotypes. Note that $\{n_i\}$ is typically much smaller than $\mathcal{N}$.
3. OPT(TFP) is solved for the set $\{n_i\}$. This is feasible even with exhaustive search methods given that $\{n_i\}$ is typically small. Suppose the solution assigns subscriber n_i the tariff (p_t, r_t, x_t). Based on the definition of $\mathcal{Y}$, we can denote as $(r_t, x_t) = y \in \mathcal{Y}$ and p_t to be the price of y (p_y).
4. For each $y \in \mathcal{Y}$ where a price was assigned, PPA discounts the price by a factor that is controlled by τ parameter:

$$p'_y = p_y - \tau \cdot (p_y - cost_y)$$

where $cost_y$ is the cost to the ISP (based on transit & capacity cost) of assigning a subscriber to y.
5. For each (non-stereotype) subscriber n, PPA assigns her to the tariff y', where

$$y' = \arg\max_{y \in \mathcal{Y}} (U_n(y) - p_y), \text{if } U_n(y') - p_{y'} \geq 0$$

The quality of the solution obtained from PPA is controlled by ϵ: large values lead to fewer clusters and a faster, but less optimal solution; but as $\epsilon \to 0$, PPA generates a larger number of stereotypes and the solution approaches the optimal case (at higher computational cost). Note that ϵ effectively bounds the distance between elements in the same cluster. We implement PPA with agglomerative hierarchical clustering [27], which explicitly incorporates a distance threshold. PPA returns solutions within $2 \cdot (\tau + \frac{2 \cdot \epsilon}{\tau}) \cdot N$ of the optimal.

Theorem 1 *If P_{OPT} is the optimal solution of the TFP problem, N the number of subscribers, ϵ is the largest difference in subscriber utility in any cluster, and τ is the discounted factor, then the ISP balance of the PPA algorithm is at least $P_{OPT} - 2 \cdot (\tau + \frac{2 \cdot \epsilon}{\tau}) \cdot N$*

The proof is based on a reduction to the *Screening* problem that is described in [26]; the validity of this reduction is presented in the Appendix □

Notice that for small values of ϵ and N, the term $2 \cdot (\tau + \frac{2 \cdot \epsilon}{\tau}) \cdot N$ is also small, and hence the worst case performance of PPA is very close to the optimal one (P_{OPT}). However, as the number of subscribers (N) or the distance between the elements in the same cluster (ϵ) increase, the term $2 \cdot (\tau + \frac{2 \cdot \epsilon}{\tau}) \cdot N$ increases as well, which deteriorates the theoretical worst-case performance.

Importantly, PPA requires hard assignments of subscribers to tariffs and thus can only be used in conjunction with the *deterministic demand* model. In contrast, Naive and SR are agnostic of the underlying demand model and can be applied to both. To distinguish the two models, we use the Naive-det and Naive-logit (similarly with SR) to denote the combination of heuristic algorithm and demand model. In the next section, we apply the three algorithms described above against our dataset to obtain a set of tariffs which are compared against the bound established by OPT(*L*TFP).

6. IMPACT OF DATA CAPPING

In this section, we evaluate the impact of moving from a flat rate price structure to one with data capped tariffs. The evaluations cover a number of regional cost scenarios, which lead to an understanding of how the impact varies in different regions and markets. Overall, we find that ISPs can extract more value by implementing data caps, but this heavily depends on transit and capacity prices. In the best scenario, an ISP can improve its bottom line by 23% by offering 8 different data capped tariffs with a very small increase (<5%) in subscription revenues. Importantly, we show that setting data caps *naively* can actually lead to a net loss for the ISP. In the rest of this section, we discuss these results in detail; we begin by describing the parameters used in the later evaluations.

6.1 Evaluation Setup

To guide the set of tariffs generated by the algorithms, we constrain the parameters related to speed tiers and data caps as follows. The set of speed tiers, $\mathcal{L}$, is identical to the current offerings seen in our dataset i.e., {24/1Mbps, 30/3Mbps, 100/10Mbps}. We vary the available data caps from 10GB to 200GB, with 5GB increments.

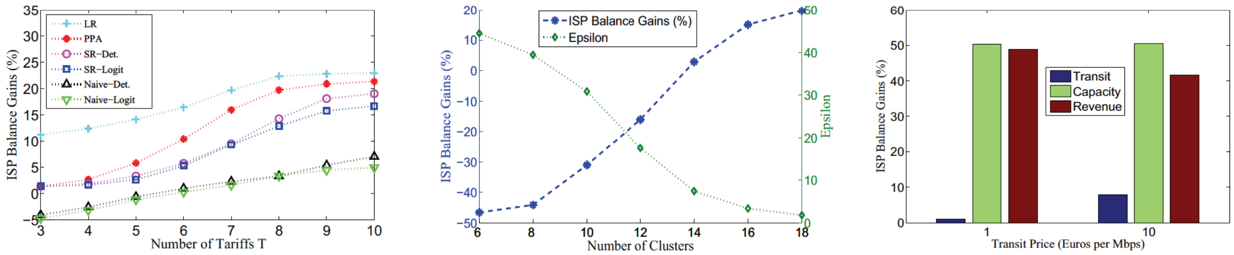
Next, we describe how to extract the utility coefficient parameters, a_n, b_n and c_n which are introduced in Eq. 2. To start, we make the assumption, motivated by the work in [17], that each subscriber, in the current flat-rate offering, evaluates the tariff exactly equal to the price p_t they currently pay for it (Table 1). This is reasonable, since the user would probably have switched providers if this were not the case. Then, replacing the left side of Eq.2 with p_t, yields:

$$a_n = \frac{p_t}{log(b_n \cdot r_t + 1) \cdot (c_n \cdot x_t - \frac{x_t{}^2}{2})}$$

The coefficients in this equation are determined as follows:

$\mathbf{r_t}$**:** This is taken as the speed tier currently being subscribed by the user.

$\mathbf{x_t}$**:** Since the flat-rate plan t allows for unlimited use, i.e. $x_t = +\infty$, we use the expected data usage of the subscriber instead, and we derive this as the *average* traffic volume for the subscriber, over all the months in the dataset.



(a) Balance gains for 3 to 10 tariffs. (b) Impact of clustering on PPA. (c) Distribution of balance gains.

Figure 4: Data cap pricing balance gains over flat-rate.

$\mathbf{b_n}$**:** This captures how a subscriber evaluates speed tiers. Large values reflect an indifference to speed tier variations, while low values reflect that the subscriber is affected by changes in speed (e.g., someone that frequently watches video). In our dataset, we suspect users with high traffic volume to have low b_n values and lightweight users to associate with high b_n values. To reflect a range of behavior, we partition the subscribers into four bands based on their average traffic volume, i.e., $<$10, 10$-$30, 30$-$100, $>$100 (GBs) and these groups are assigned b_n values of 1000, 100, 10 and 1 (respectively).

$\mathbf{c_n}$**:** This parameter captures how much data volume a subscriber may need. We assume that this is bounded by the maximum data usage over the dataset for a subscriber and set it to the largest value (monthly usage) seen over the dataset for the subscriber.

6.2 ISP Impact

We use the notation J^f (J^c) to represent the solution to Eq.1 in a flat rate (data capped) setting. Our main evaluation metric in this section is *gain*, defined as the difference $J^c - J^f$ which is the improvement in the ISPs bottom line when data capped pricing is adopted.

Throughout, we adopt the transit and capacity cost functions defined in Eq. 5 and 6 respectively. Unless otherwise specified, we set the parameters: $f = 0.66$ and $t_p = 1$ as in [24], $c_p = 10$ and $v = 0.7$ as in [13, 25] and $c = 1$ as in [17].

We first investigate the question of how many tariffs are needed. Figure 4(a) plots *gain* as a function of the number of tariffs (T). We find that across all the algorithms (save for Naive), there is an improvement in gain as the number of tariffs are increased until $T = 8$. This can be explained by the fact that adding an extra tariff increases the number of price/byte offerings; this slightly increases the likelihood of some subscribers getting higher utility with the extra offering. However, once T "matches" the heterogeneity of data usage, additional tariffs are unlikely to improve subscriber surplus. While in general, the results obtained by using the deterministic model cannot be quantitatively compared with those obtained using the logit demand model, we consolidate these in Fig. 4(a) to preserve space. We do note however, that the two families yield similar results for SR and $Naive$ (there is no logit equivalent for PPA). In the following discussion, we focus mainly on the results from the deterministic model.

Looking at Fig. 4(a), we find that SR consistently produces more beneficial solutions compared to $Naive$. This is likely due to the differences in how they assign speed tiers to tariffs. $Naive$ performs a "one-shot" assignment, while SR iteratively improves the solution obtained at each stage. Importantly, note that $Naive$, with $T{<}5$, has *negative gain*, i.e., the ISP does worse *after* instituting data caps. Examining this carefully, we uncover that with $T{<}5$, most of the plans are (myopically) allocated to the higher (more expensive) speed tier(s) since user utility increases with speed, ignoring all else. However, the higher prices make it less likely for subscribers to find a suitable tariff, and they leave. The ISP can only counter this by lowering prices across tariffs (reducing the revenue). These conflicting factors result in the ISP being worse off with data caps. The key takeaway here is that tariff selection is non trivial: simplistic solutions may not extract value from subscribers effectively. Finally, we see in Fig. 4(a) that PPA and $SR{-}Det$ return solutions that are qualitatively similar and are also approach the OPT(LTFP) solution (labeled as LR), indicating that these are quite close to the optimal solution of the TFP problem.

To understand the intuition about why increasing tariffs improves gain, we examine two tariff assignments in detail. Table 2 enumerates the tariffs selected by PPA for $T{=}3$ and $T = 10$. When $T = 3$, each speed tier is allocated one data cap. In fact, this scenario reflects the system where operators – Comcast, for example – apply a single (uniform) data cap for each speed tier. However, when $T = 10$, a higher fraction of data caps are assigned to the 100/10 Mbps tier. In our dataset, the subscribers in this speed tier exhibited the most variation in monthly data usage (over the other 2). By allocating more tariffs to the tier with the most variation, PPA increases the likelihood of subscribers finding maximal surplus tariffs. The overall gains are dictated by the number of subscribers that are able to identify such tariffs amongst the offerings (and staying with the ISP). Given that PPA is the best performing scheme, we focus the rest of the discussion on the results for PPA with $T = 8$.

Recall that with PPA, the solution quality (and computational complexity) depends on the required number of clusters, which is a parameter to the algorithm. Fig. 4(b) plots this (on the x-axis) against two different terms on the y-axis – *gain* on the left y-axis, and ϵ on the right y-axis (this parameter controls the approximation bound of the obtained tariff solution). When the number of clusters is less than 12, PPA actually returns solutions with negative gain. We believe this is due to the number of clusters being unable to capture the heterogeneity of the subscriber population. However, with larger number of clusters, the extracted tariff solutions have positive gain. The relationship between the number of clusters and ϵ is consistent with this observation. Lower values of ϵ, which determines the tightness of the clusters, force it to find solutions that are closer to the optimal, and this necessitates a higher number of clusters (and hence complexity).

PPA requires only 8 tariffs for the ISP to improve its gain by 23%. This is significant, and comes without requiring any capital expenditure from the ISP. However, in practice, ISPs may not be willing to use more than a few tariffs in order to make their offerings simple to understand. Subscribers may face significant cogni-

Speed tier	24/1 Mbps	30/3 Mbps	100/10 Mbps
			T=3
Cap(GB)	45	55	70
Demand(%)	32.2	7.8	60
Price(€)	21.8	25.32	33.8
			T=10
Cap(GB)	15,40	20, 60	20,25,35,50,80,100
Demand(%)	9.55,17.73	1.82,5	16.36,15,11.82,10,8.18,9.09
Price(€)	21.8,22.65	24.58, 26.55	32.51,33.41,34.1,34.75,35.22,35.75

Table 2: Data cap pricing tariffs when T is 3 and 10 (PPA algorithm).

Region	transit price t_p (€ per Mbps)	transit fraction f (%)	balance gains(%) c_p=0.1 (€ per Mbps)	c_p=10 (€ per Mbps)
NY	0.5	66	10.2	19.6
UK	2	66	11.6	20.8
Tokyo	3	33	10.9	20.2
Sao Paulo	6	66	14.8	24.1
Rural Texas	50	66	26.5	33.4
Sub-Saharan	100	80	36.1	38.2

Table 3: Traffic cost parameters and ISP gains in various markets.

tive dissonance is choosing between a large number of tariffs [28], which works against the ISP. Thus in practice, ISPs may choose to use fewer tariffs (than are optimal) and thus not realize the full possible benefit in order to attract (or retain) customers. A detailed analysis of these interactions is outside of the scope of this paper.

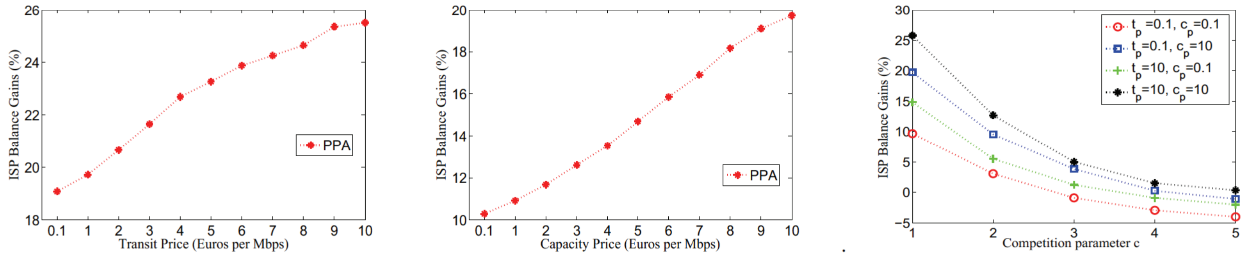
Dissecting the gains: The ISP can improve its gain by adopting data caps due to two factors: (i) by reducing traffic costs (transit and capacity), or (ii) by pushing users to more expensive tariffs (higher speeds and/or caps). While much of the recent arguments offered by ISPs in moving to data capped schemes center on the former, *both* of these factors are important and we find evidence of both even as the relative effects vary by region. We start by examining two extreme cases for transit prices in Figure 4(c), which presents the breakdown of ISP gain in two different markets. The parameters used are as enumerated in the beginning of the section. In the first market (on left) with t_p=1€ and the second market (on right) with t_p=10€. On the left, transit is plentiful and cheap and hence effective transit *costs* are low, 50% of the gain from moving to data caps is realized by lowered capacity costs, 48% is contributed by increased revenue from users going to more expensive tariffs, and only about 2% of the gain is contributed by the lowered transit costs. We contrast this with the scenario on the right, where transit is ten times as expensive. Here, 10% of the gain is contributed by lowered transit costs, and this is for the exact same subscriber and traffic composition. Thus, we can generalize this to say that when traffic prices (transit or capacity) are low, the ISP gains relatively little by reducing traffic volume (which is what caps accomplish), making it more likely that the ISP will look elsewhere to balance the increased costs (e.g., increasing tariff prices significantly for heavy users). On the other hand, when traffic prices are high, the ISP is able to reduce its cost overhead significantly by lowering the traffic. A related observation is that transit (and capacity) costs affects the undesirability of heavyweight users to the ISP. With low transit prices, ISPs are less sensitive to heavyweight users: with the transit cost at 1€, we find 1% of subscribers leave the ISP when data-capped, while this grows to 5.2% when the transit price is 10€.

The transit cost settings used in Figure 4(c) may seem contrived, but in fact, they are very much in line with prices reported today. Table 3 is a compendium of network cost parameters reflecting various regional markets around the world, and were assembled from several sources [24, 29–31]. The second column in the table indicates the unit price for transit traffic, and column 3 represents the transit fraction of total traffic for a sample ISP in the region. The remaining two columns indicate the overall gain with using $T = 8$ data capped tariffs (under two capacity cost regimes); we return to this shortly. Examining column 2, we find tremendous variability across ISPs depending on the size, geography, and the number of available transit providers. For e.g., an ISP in rural Texas recently reported transit prices over 50$ due to a lack of competition in the local transit market [29]. We also uncovered examples of extremely high transit prices in sub Saharan Africa and small island nations [30]. Thus, the cost examples used in Fig. 4(c) reflect prices well inside what can be found today. In the following, we examine transit and capacity prices independently and analyze how gains vary in each case. We do this independently because these prices are not necessarily coupled: some regions may have very good intra-region connectivity, but sparsely connected to other regions, and vice versa. The last two columns of Table 3 summarize the overall ISP gain for two extreme capacity price points.

In Figure 5(a), we look at the impact of varying transit prices on the ISP gains. We fix T=8, c_p=10€, and f=0.66. As expected, when transit prices are high, data caps benefit the ISP by reducing the aggregate traffic, and hence $G(.)$ – the transit cost. When the unit transit price is greater than 1€, the ISP stands to improve its bottom line by at least 20%. The gains are less striking when transit prices drop below this price point. In fact, most markets in the US and Europe today have transit prices that are sub 1€ and in such markets, data capping is not likely to noticeably impact the ISPs bandwidth costs (even if traffic rates decrease). On the other hand, in markets with high transit prices – parts of South America and S.E. Asia, Africa, etc. – moving to a data capped model significantly lowers the traffic costs for the ISP and benefits subscribers as well. This observation is consistent with the fact that data capped pricing plans have existed for a long time in such areas.

In Figure 5(b), we look at the impact of changing capacity prices on ISP gain. While we expect this to qualitatively follow the same behavior as seen with transit prices, the actual costs have minor differences (a log factor, cf. Eq.5 and Eq.6). Again, ISP gains from data capping steadily increase as the unit capacity prices increase, and thus, data capping becomes a more viable tool to manage costs as the capacity costs increase. Unlike with transit prices, where regional prices are sometimes available, capacity costs are much harder to discover (ISPs consider them confidential). Typically, we expect capacity prices to be slightly lower than transit; the latter is likely to have a small margin since it is offered as a service. In any case, our evaluation covers two orders of magnitude in capacity prices, and we believe most markets fall somewhere into this range.

Deterministic vs. logit: Turning now to the *logit* demand model, we first note that the algorithms $Naive{-}logit$ and $SR{-}logit$ perform very similarly to the deterministic case as seen in Fig.4(a). This is because the logit model assigns users probabilistically to tariffs, based on their utility. When computing tariffs for each user, they eventually get assigned to the tariff associated with the highest probability. Hence the users are mapped to tariffs in a manner that largely mimics the deterministic model. One of the salient features of the *logit* model is that it allows us to investigate the impact of competition, captured by the parameter c in Eq.3. In monopoly or duopoly, as is the situation most US markets, an ISP may be able to set data caps (alternatively, set prices) aggressively – without fear of losing subscribers. On the other hand, in markets with a large number of competing providers, such as in most of Western Europe, small pricing tweaks may result in subscribers departing the



(a) Balance gains for various transit prices (b) Balance gains for various capacity prices (c) Impact of market competition on ISP gains

Figure 5: Data cap pricing balance gains over flat-rate with varying transit price, capacity price, and competition.

ISP and this is likely to make the ISP more risk averse. In Figure 5(c), we examine various levels of market competition for four cost settings. Recall that the numerical value of c does not reflect the number of competitors; it simply characterizes the intensity of competition. As c increases, we find that more subscribers leave the ISP when data caps are imposed and the net impact varies with the cost scenarios. When the traffic costs are high, and as observed previously, the ISP can afford to "shed" customers up to a point. Even with high competition ($c = 5$), the ISP still finds a positive gain with $t_p, c_p = 10$€ and the gain decreases as the traffic costs drop. Thus, overall, the impact of data capping to an ISP's bottom line is also affected by the level of competition in the market. In many cases, ISPs often use incentives such as service-bundling or promotional pricing to retain customers to counteract the increased competition.

While putting these results into perspective, we point out an important assumption in our methodology – the monthly price paid by subscribers (between $21-33$€) goes entirely toward the capacity and transit costs borne by the ISP. In reality, only a small fraction of this is likely to used toward traffic costs, with a much larger fraction is likely to be toward cost factors not easy to uncover – leasing costs, salaries, etc. If the evaluations were to be carried out with more detailed and fine grained cost structures, we expect the economic gains from data capping to be much more striking.

6.3 Subscriber Impact

We now examine data capping from the viewpoint of individual subscribers. Figure 6(a) plots the distribution of the difference in monthly subscription prices (p_t) between a flat rate pricing scheme and a data capped scheme, across all the subscribers. We observe that roughly 25% of the subscribers see a net (small) reduction in their monthly bills. These generally correspond to the lightweight users that pay a disproportionately high cost per unit of data usage in the flat-rate setting. Note that cheaper tariffs associated with the lower end of the data capped plans are likely to attract new customers to the service and increase revenue – a side effect that we cannot model easily from our dataset. Thus overall, data capping can benefit the (potential) subscriber with modest needs by lowering costs. In our dataset, 18% of the users have a significant increase (>2€ per month) in the new scheme. These particular users correspond to the most heavyweight users; and who are willing to pay more for increased speed or data cap (as captured by the utility function). The remaining users are charged almost the same or a slightly higher price with a data capped tariff structure. We do note that some users may choose to modulate their behavior (download less rather than pay more) when faced with steeper charges. However, such behavior is very challenging to model, and we leave that as an open question.

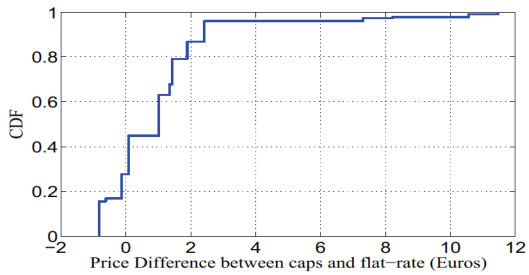
Next, we compare the *fairness* between flat rate and data capped pricing. In a *completely fair* scheme, subscribers would pay proportional to their actual data usage (all subscribers pay same unit cost). Comparing two pricing schemes, the fairer one would exhibit less variance in this unit cost across subscribers. Figure 6(b), plots the distribution of the ratio of price paid/usage over all the subscribers for the two pricing models. As can be expected, we observe smaller price/byte variations among the users under caps. The most visual case is for the heavy users (area in CDF near zero) who shift to the right, since they are forced to pay more for maintaining their data usage or they are capped. The standard deviation of price/byte among subscribers drops by 1€ under caps. Thus, our empirical results validate some of of the existing theoretical benefits of data capping for subscribers [4–7, 10–13].

7. DISCUSSION

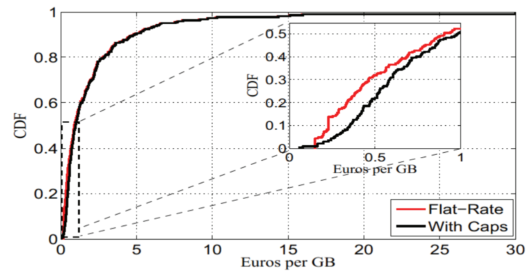
In this section, we revisit some assumptions made in our model and methodology and also describe some extensions to our work so far.

Cost model: The profit and cost model we use is relatively simple. Making it more realistic would involve obtaining accurate and fine grained cost structures from an ISP. However, even this might not be quite enough given that there is a great deal of subjectivity in how costs are applied (and over what time frame). To use an example: consider the real estate cost of housing the networking equipment. Some buildings are owned outright, some are rented, and some are obtained using public funds or support (as is the case for national carriers in some countries). Another example is the cost of manning the help desks that users call into – staffed by a combination of salaried full time and temporary personnel. This is often a large expense to ISPs and the outlay depends on the expected volume of service calls. In such cases, the month-to-month amortization is unclear and open to debate. Our simple model avoids such issues and importantly, it captures how the nominal profits/costs depend on subscriber behavior.

User utility: The utility function we use leverages previous work to describe (user) satisfaction entirely in terms of the speed tier and usage allowance. While we believe it to be reasonable, it may not be universal (some users who are extremely price sensitive might not tolerate *any* price hikes). In addition, we consider that each subscriber can pick any speed tier that is offered; this is not the case in reality: rural subscribers may not have access to a faster fiber service and cannot switch to it. Yet another shortcoming of our model is that it assumes users can upgrade/downgrade speed tiers easily; in reality, the switching cost might be non-trivial, or there could be other dissuading factors [4, 9]. Finally, our work ignores the impact of particular overage policies when data-caps



(a) Tariff changes across subscribers



(b) CDF of €$/GB$ in flat rate & data capped schemes.

Figure 6: Impact of Data Caps on Subscribers

are exceeded. We plan to investigate other utility functions and extend our models to account for overage policies in future work.

Alternate pricing models: Our empirical findings hint at two extensions to current pricing models. Recall that subscriber traffic volume correlates with their impact on transit and capacity costs. Given this, it is interesting to imagine time-based pricing models (such as those being discussed in the context of "smart grid") to incentivize users to shift their elastic consumption (file downloads, etc.) to off-peak times. This could reduce the ISP costs (by driving down the peak traffic) and these savings could be passed on to subscribers. This model has been previously studied in a theoretical framework [32], so fine-grained usage data such as in our dataset could provide empirical validation.

8. CONCLUSION

In this paper we investigate the issues surrounding data capping in residential broadband networks. To this end, we develop a methodology that yields a set of optimal tariffs, taking a set of cost components and a population of subscribers. This framework builds upon a simple, extensible model of ISP costs and a model that relates subscribers' utility to bandwidth speeds (speed tiers) and data caps. The developed framework enables to examine the relative benefits, for providers and their subscribers, of moving from flat rate pricing (the predominant model in most western broadband markets) to data capped pricing models, and to understand these benefits (or the lack) as traffic costs change.

We then apply the framework to a novel dataset that captures fine grained, broadband data usage from a set of 260 ISP subscribers. By referencing traffic prices that reflect different broadband markets around the world, we are able to identify the particular conditions when moving to a data capped pricing model is beneficial to an ISP, in the sense of limiting bandwidth costs. Overall, we find that data capping can be an effective mechanism for ISPs to reduce their costs only when transit and capacity costs are relatively high. On the other hand, when these costs are relatively low, data capping is not a very useful mechanism to balance bandwidth costs. In addition, the level of competition directly influences the outcome of data capping; in competitive markets, data capping may not be effective in managing costs as subscribers may leave the service and reduce the incoming subscription revenues. Our work in this paper is a first attempt to enable ISPs or policy makers to obtain quantitative insights of the impact of data capping.

9. ACKNOWLEDGMENTS

We would like to thank Augustin Soule and Pascal Le Guyadec for facilitating access to the data. We also thank our shepherd, Kenjiro Cho, and the anonymous reviewers for their constructive feedback.

10. REFERENCES

[1] Will Netix Kill The Internet?, Dec 2010.

[2] `www.stopthecap.com`.

[3] Public Knowledge Petition Against Comcast. http://tinyurl.com/ofp2azs, 2012.

[4] Andrew Odlyzko, Bill St Arnaud, Erik Stallman, and Michael Weinberg. Know your limits: Considering the role of data caps and usage based billing in internet access service. *Public Knowledge*, 2012.

[5] Daniel A Lyons. Impact Of Data Caps And Other Forms Of Usage-based Pricing For Broadband Access. *Mercatus Center*, 2012.

[6] Soumya Sen, Carlee Joe-Wong, Sangtae Ha, and Mung Chiang. Pricing data: A look at past proposals, current plans, and future trends. *arXiv preprint arXiv:1201.4197*, 2012.

[7] Richard Edell and Pravin Varaiya. Providing internet access: What we learn from index. *IEEE Network*, 13(5):18–25, 1999.

[8] J. Williams. A. Nevo, J. Turner. Usage-based Pricing And Demand For Residential Broadband. *Social Science Research Network*, 2013.

[9] Anja Lambrecht, Katja Seim, and Bernd Skiera. Does uncertainty matter? consumer behavior under three-part tariffs. *Marketing Science*, 26(5):698–710, 2007.

[10] FCC Open Internet Advisory Committee. Policy issues in data caps and usage-based pricing, 2013. Annual Report.

[11] Gregory L Rosston, Scott J Savage, and Donald M Waldman. Household Demand For Broadband Internet. *The BE Journal of Economic Analysis & Policy*, 10(1), 2010.

[12] Andrew Odlyzko, Papak Nabipay, and Zhi-Li Zhang. Flat vs. Metered Rates, Bundling, and "Bandwidth Hogs". In *NetEcon*, 2011.

[13] Wei Dai and Scott Jordan. Design and impact of data caps. In *GLOBECOM*, December 2013.

[14] Parminder Chhabra, Nikolaos Laoutaris, Pablo Rodriguez, and Ravi Sundaram. Home is Where the (Fast) Internet is: Flat-rate Compatible Incentives for Reducing Peak Load. In *HomeNets*. ACM, 2010.

[15] Jeremy Blackburn, Rade Stanojevic, Vijay Erramilli, Adriana Iamnitchi, and Konstantina Papagiannaki. Last Call for the Buffet: Economics of Cellular Networks. MobiCom '13, 2013.

[16] Liang Zhang, Weijie Wu, and Dan Wang. Time-dependent Pricing In Wireless Data Networks: Flat-rate Vs. Usage-based Schemes. In *INFOCOM*. IEEE, 2014.
[17] Vytautas Valancius, Cristian Lumezanu, Nick Feamster, Ramesh Johari, and Vijay V. Vazirani. How many tiers?: Pricing in the internet transit market. SIGCOMM '11. ACM, 2011.
[18] Federal Communication Commission, A Report on Consumer Wireline Broadband Performance in the U.S., 2014. FCC report.
[19] Costas Courcoubetis and Richard Weber. *Pricing Communication Networks: Economics, Technology and Modeling*. John Wiley & Sons, 2003.
[20] Asuman Ozdaglar and R. Srikant. Incentives and pricing in communication networks. In *In Algorithmic Game Theory*, pages 571–591. Cambridge Press, 2007.
[21] J. Bauer et. al. Economics Of Usage-based Pricing In Local Broadband Markets. *Michigan State U.*, 2012.
[22] Ana Amante and João Vareda. Switching Costs In The Portuguese Telecommunications Sector: Results From A Customer Survey. *Concorrência e Regulação*, pages 115–135, 2011.
[23] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1974.
[24] Rade Stanojevic, Ignacio Castro, and Sergey Gorinsky. CIPT: Using Tuangou to Reduce IP Transit Costs. CoNEXT '11. ACM, 2011.
[25] Wei Dai and Scott Jordan. Modeling ISP Tier Design. In *International Teletraffic Congress*, pages 1–9. IEEE, 2013.
[26] Kristóf Madarász and Andrea Prat. Screening with an Approximate Type Space. *STICERD - Theoretical Economics Paper Series*, 2010.
[27] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The Elements Of Statistical Learning*. Springer, 2009.
[28] Barry Schwartz. *The Paradox Of Choice*. Ecco, New York, 2004.
[29] 2014 Transit Prices and Peering Projections. `http://drpeering.net`, Oct 2013.
[30] Global Internet Geography: Executive Summary. `http://www.telegeography.com`, 2013.
[31] IP Transit Pricing Service. `www.telegeography.com`.
[32] Prashanth Hande, Mung Chiang, Robert Calderbank, and Junshan Zhang. Pricing under constraints in access networks: Revenue maximization and congestion management. In *INFOCOM*, pages 1–9. IEEE, 2010.

APPENDIX

Lemma 1 *The TFP problem is NP-Hard.*

PROOF. We prove the NP-Hardness of the TFP problem by reduction from the Maximum Independent Set Problem (MISP), which is NP-Hard. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the sets $\mathcal{V}$ and $\mathcal{E}$ denote the vertices and the edges, a maximum independent set is the maximum-sized set of non-adjacent vertices. For any instance of MISP, we construct the following equivalent instance of TFP:

For each vertex $v \in \mathcal{V}$, we create a tariff described by the pair of speed and cap (r_v, x_v), with cost that is independent of the demand and equal to the degree of v (d_v). The price of the tariff is equal to $1 + \frac{\gamma}{d_v}$, where γ is an arbitrarily small positive constant. Thus, the ISP benefits by offering the tariff (r_v, x_v) iff it is picked by at least d_v subscribers. Finally, we create one subscriber per edge $e \in \mathcal{E}$, who obtains arbitrarily high utility from each tariff (r_v, x_v) such that v is an adjacent vertex to e; otherwise zero. The TFP problem asks for determining the subset of tariffs to offer to the users.

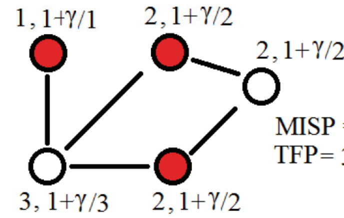


Figure 7: An example of the reduction from the MISP with 5 vertices (tariffs) and 5 edges (users). Vertex labels denote the costs and the prices of the respective tariffs.

Given a solution of value $k \cdot \gamma$ to TFP, one can find a solution of value k to the MISP instance. This is because, this TFP solution specifies that exactly k tariffs of the form (r_v, x_v) are offered by the ISP, each one of them picked by *all* the d_v subscribers corresponding to an adjacent edge to vertex v, who pay in total $d_v + \gamma$. Since, the cost for offering a tariff (r_v, x_v) is equal to d_v, the ISP benefits by γ for each one of the k tariffs that it offers, and the solution value is $k \cdot \gamma$. But, recall that each offered tariff (r_v, x_v) corresponds to a separate vertex v, and these vertices must be non-adjacent, since each subscriber can pick *at most one* tariff. Thus, there are k non-adjacent vertices in $\mathcal{G}$.

Figure 7 describes a simple example. There exist 5 possible pairs of speed and cap (tariffs) with costs $\{1, 2, 2, 2, 3\}$. The ISP balance-maximizing policy is to offer 3 out of 5 tariffs to the subscribers (those that correspond to the marked vertices). The prices per tariff are $\{1 + \frac{\gamma}{1}, 1 + \frac{\gamma}{2}, 1 + \frac{\gamma}{2}, 1 + \frac{\gamma}{2}, 1 + \frac{\gamma}{3}\}$. □

Screening Problem: The "Screening problem" [26] commonly refers to the case that a seller offers a menu of item-price specifications to a buyer or a continuum of buyers. The buyer's item preferences depend on her private type $b \in \mathcal{B}$, while the seller only knows the probability density function $f(.)$ of it. The evaluation of the buyer of type b for an item $i \in \mathcal{I}$ is equal to $V(b, i)$. The seller wishes to find the menu $\mathcal{M} \subseteq \mathcal{I}$ of items to offer and the prices $p(i)$, $\forall i \in \mathcal{M}$ aimed at maximizing her profit: $\sum_{b \in \mathcal{B}} f(b)(p(y(b)) - c(y(b))$, where $y(b)$ denotes the item picked by the buyer of type b and $cost_i$ is the production cost of item i. The buyer of type b will purchase the item $y(b)$ in $\mathcal{M}$ such that: $V(b, y(b)) - p(y(b)) \geq V(b, j) - p(j), \forall$ pairs $(j, p(j)) \in \mathcal{M}$.

Theorem 1 *The TFP problem is polynomial-time reducible to the Screening Problem.*

PROOF. Consider an instance of the TFP problem defined by the set of subscribers $\mathcal{N}$, the possible tariffs $\mathcal{L} \times \mathcal{X}$, the utility functions U and the costs C. Then, the above is equivalent to the special instance of the Screening problem, where we set: $\mathcal{I} = \mathcal{L} \times \mathcal{X}$, $\mathcal{B} = \mathcal{N}$, $f(b) = \frac{1}{|\mathcal{N}|}, \forall b \in \mathcal{B}$, $V(b, i) = U_n(t)$ and $cost_i = C(t)$, where type b corresponds to a subscriber n and item i corresponds to a tariff t. □