Data Plan Throttling: A Simple Consumer Choice Mechanism

Barbara M. Anthony
Math and Computer Science Department
Southwestern University
Georgetown, Texas 78626
Email: anthonyb@southwestern.edu

Christine Chung
Computer Science Department
Connecticut College
New London, Connecticut 06230
Email: cchung@conncoll.edu

***Abstract*—Despite only a small portion of unlimited data plan users experiencing throttling each month, it is a prominent source of complaints from users and a significant concern for mobile network operators. We propose a simple mechanism that allows users to choose when they want their data transmission "fast," and when they want it "slow." Users still have the same cap on total high-speed transfer before being throttled, and hence may still be subject to throttling, but now they are given some control. We propose a basic model of payoffs, and demonstrate that the proposed mechanism would be preferable to the user over the throttling policies currently in place. We then consider the impacts that extend beyond a single user, and provide a framework for determining the aggregate effects of such a mechanism.**

I. Introduction

A. Background and Motivation

While a number of mobile network operators offer 'unlimited' data plans, those plans do not guarantee truly unlimited data at maximum speed at all times. Currently, many operators *throttle* some of their unlimited data plan users: after exceeding their allotment of data transfer capacity (the "cap"), users are no longer able to transfer data over high-speed networks (like 4G) and are demoted to transferring all remaining data for the month at slower speeds. While typically only a small percentage of users are affected (see, for example, [1] and [10]), many users are personally concerned about throttling and similar practices, sparking lawsuits and other actions [3].

We propose a simple mechanism that allows users to specify which data they want transferred immediately and quickly, and which they want transferred, but not as urgently. The motivation for our mechanism is that users currently are forced to use their fast data transfer capacity, even for tasks that are rather data intensive, but are often not time-sensitive (downloading apps, syncing photos, etc). Knowing they may be throttled after reaching their cap causes users to police their own activity on their mobile devices, needlessly reducing their freedom as consumers. In turn, their utility (as end-users) to content-providers and other online services is diminished. (As one anecdotal example, Facebook, Google, Dropbox, and others provide a convenient feature where photos from your phone are automatically uploaded; this is a feature that end-users concerned about their data usage likely opt out of using.)

We demonstrate via a simple payoff model that under our proposed mechanism, the individual user benefits, and some metrics related to the average experience of the users collectively improve, which in turn can be better for the operators. It requires minimal effort on the users' part, is simple to implement, and does not require additional infrastructure from the operators. In addition, given the consumer research studies that suggest that users respond more favorably when they perceive that they have more control [5], implementing this simple mechanism may result in users being happier with their wireless providers, even if (just as in the existing mechanism) not all data can be sent quickly all the time.

B. Proposed Mechanism

We propose a simple mechanism in which, at any time, users may set the "mode" of the phone to one of two choices: *fast* or *slow*. As long as the user is not being throttled and network conditions permit, in *fast* mode, data is transferred at high speeds, while in *slow* mode, data is transferred at low speeds. This is in contrast to the existing mechanism where high-speed transfer is used whenever the user has not exceeded her cap. Under the proposed mechanism, the cap naturally (and critically) applies only to data transferred in *fast* mode. After exceeding the cap, data is transferred at low speeds, regardless of the user's choice; this is identical to the existing mechanism. While users still may be throttled after exceeding the cap on *fast* data, now they are in control, rather than being forced to use *fast* data transfer up to the cap. It is up to an individual user to decide how frequently to switch between the two modes, if at all.

Note that our proposed mechanism is a theoretical abstraction and simplification of the actual underlying details of the implementation required for such a mechanism. This work should be viewed as a high-level proposal and a theoretical study of the simple, abstract model that would overlay the real-world actualization of such a solution. Therefore, we do not seek to address lower-level issues in the present work.

C. Related Work

Ultimately, most people would find it desirable if users experienced less throttling without operators having to invest in more infrastructure. Various pricing schemes have been extensively studied, including both flat-rate plans and different pricing strategies that would be advantageous to particular constituents: see [9] for a comprehensive survey. Throttling could potentially be avoided by charging per use, yet in some cases it is argued that such ‘fine-scale charging’ is undesirable, both in terms of tracking the transaction costs and from the users’ perspectives [7]. Furthermore, users often prefer flat-rate pricing to pricing based on usage [8], evidenced by the fact that most major carriers still offer flat-rate plans. These compelling factors led us to consider what mechanisms can be implemented while maintaining plans with caps on the amount of data.

By allowing users control of when they use their fast capacity, rather than forcing them to use it first, it may be possible to provide better utilization of the existing infrastructure. Along the lines of efficient utilization, concern about the looming possibility of spectrum exhaust [2] has led to research into cognitive radio networks as a means of improving spectrum utilization [4]. Other work has studied what happens when heavy users are ‘deprioritized’ in wireless networks [11] and how that compares to throttling.

D. System Model and Terminology

Let C be the total amount of fast data that a user can transmit before being throttled, that is, the *cap* or capacity on fast usage per billing cycle (month). This cap is, naturally, the same for all users. We typically think of C as being a few GB. Verizon Wireless, for example, indicates that “using more than 2 GB of data in a month, you’re in the top 5% of data users and will be impacted by Network Optimization when you’re connected to congested 3G cell sites” [10] though on the same page they indicate that Network Optimization is not the same as throttling. Other operators’ caps are similar: AT&T indicates that 3GB of usage in a billing cycle can result in lowered throughput speed [1].

Let d_i denote the total amount of data transfer demanded by user i (per month). We make the simplifying assumption that the amount of data the user wants fast is independent of when the user hits the cap, letting $0 \leq f_i \leq 1$ be the portion of the demanded data d_i that the user wants fast.

In Section II we consider an individual user’s perspective, detail a payoff matrix, and discuss the ramifications of the proposed mechanism for an individual in isolation. In Section III we consider all users in aggregate, both from their perspective and that of the operator, and discuss how the collective usage changes under various scenarios. In Section IV we conclude with some recommendations regarding circumstances in which the proposed mechanism should be used.

II. An Individual User’s Perspective

We restrict our discussion in this section primarily to *heavy* users, that is, those whose data demands exceed the cap, i.e. users i such that $d_i > C$. Non-heavy, or *light*, users can request all data be sent fast and not be impacted by throttling, per published operator policies. Yet light users may also benefit from improved network congestion. Within this section we consider only one user, dropping the subscripts on f_i and d_i.

A. The Payoff Matrix

We conceptualize the existing and proposed mechanisms using a payoff matrix based on the choices made by the user and the operator. The user chooses her data transfer to be *fast* or *slow*, and can indicate such (knowing that there is a cap on the amount of fast data per month), and said data can then either be transferred *fast* or *slow* by the operator. Thus, four different outcomes can arise, yielding different payoffs for the user, as shown in Table I. Let $r > 0$ be the positive utility (a “reward”) that a user derives when the data she wants fast is transferred fast. Let $p < 0$ be the negative utility (a “penalty”) the user incurs when data she wants fast is transferred slowly. Let b be a non-negative utility (a small “bonus”) with $r > b \geq 0 > p$ that the user receives when data she wants slow is transferred fast. Let 0 be the payoff for the user when data she wants sent slow is transferred slow; this may be assumed without loss of generality, by additive scaling as needed.

TABLE I. Payoffs for the user, based on the speed the user requests, and the speed at which the operator transfers the data, with $r > b \geq 0 > p$.

		Mobile Operator	
		Fast	Slow
User	Fast	r, a reward	p, a penalty
	Slow	b, a bonus	0, neutral

TABLE II. Amount of data transferred at each speed by the operator based on speed requested by any heavy user ($d > C$) in the existing mechanism.

		Mobile Operator	
		Fast	Slow
User	Fast	fC	$f(d-C)$
	Slow	$(1-f)C$	$(1-f)(d-C)$

B. Results

We first analyze the payoffs for heavy users, considering separately the cases where $fd > C$ (the user wants more fast data than the cap), and $fd \leq C$ (the amount of fast data the user wants is less than the cap, but the user is still a heavy user, where total desired data exceeds the cap). We then calculate the payoff for a light user in each mechanism. Let π and π^* denote, respectively, the payoff in the existing and proposed mechanisms.

Case 1: Heavy user, $fd > C$

In the existing mechanism, the first C units of data for a user is sent fast, and then the user is throttled so that the remaining $d - C$ data is sent slow. Since the user wanted f of the data fast, both before and after reaching the cap C, we can easily determine how much data is sent at the speeds the user requested, and how much is not, as illustrated in Table II. Thus the payoff to the user is calculated to be

$$\begin{aligned}\pi &= r \cdot fC + b \cdot (1-f)C + p \cdot f(d-C) + 0 \cdot (1-f)(d-C) \\ &= rfC + b(1-f)C + pf(d-C).\end{aligned}$$

In the proposed mechanism, the user specifies which data is sent fast or slow, and is allowed a maximum of C fast data. Again we can determine how much is sent in accordance with the user’s wishes in Table III. The payoff is thus

$$\pi^* = rC + p(fd - C).$$

Lemma 2.1: In the case $fd > C$, the payoff π^* for a heavy user in the proposed mechanism is at least as large as the payoff π in the existing mechanism.

TABLE III. AMOUNT OF DATA TRANSFERRED AT EACH SPEED BY THE OPERATOR BASED ON SPEED REQUESTED BY A HEAVY USER IN THE PROPOSED MECHANISM, CASE 1 ($fd > C$).

		Mobile Operator	
		Fast	Slow
User	Fast	C	$fd - C$
	Slow	0	$(1-f)d$

TABLE IV. AMOUNT OF DATA TRANSFERRED AT EACH SPEED BY THE OPERATOR BASED ON SPEED REQUESTED BY A HEAVY USER IN THE PROPOSED MECHANISM, CASE 2 ($fd \leq C$).

		Mobile Operator	
		Fast	Slow
User	Fast	fd	0
	Slow	$C - fd$	$d - C$

Proof: The payoff in the existing mechanism is $\pi = rfC + b(1-f)C + pf(d-C)$, while the payoff in the proposed mechanism is $\pi^* = rC + p(fd - C)$. Since $b < r$,

$$\begin{aligned} & rfC + b(1-f)C + pf(d-C) \\ & \leq rfC + r(1-f)C + pf(d-C) \\ & \leq rC + pf(d-C) \\ & \leq rC + pfd - pfC \\ & \leq rC + pfd - pC \end{aligned}$$

where the last inequality is due to the fact that $fC \leq C$ and p is negative. ■

Corollary 2.2: When $fd > C$ and $b = 0$, $\pi^*/\pi \geq 1/f$.

As reflected in the tables, this category of users exceeds the cap in the amount of data they want fast, so they would be throttled in both the existing and proposed mechanisms. Observe that the payoff for the user in the proposed mechanism is identical to the payoff in the existing mechanism precisely when $f = 1$, i.e., the user wants all data fast. See the full version of the paper for the lower bound on π^*/π when $b \neq 0$.

Case 2: Heavy user, $fd \leq C$

Since we are concerned only with the payoff improving with the proposed mechanism, we need only to show that it increases from the existing mechanism to the proposed mechanism. In the existing mechanism, the first C units of data are again all sent fast, regardless of the user's preference, and thus the amounts are unchanged from the previous case, so Table II still applies.

In the proposed mechanism, the user specifies which data is sent fast or slow, and is allowed a maximum of C fast data. Thus all $fd \leq C$ data that the user wants to be sent fast is in fact sent fast, an additional $C - fd$ that the user would have been happy to have sent slow can be sent fast, and the rest is sent slow, as shown in Table IV. The payoff is thus

$$\pi^* = rfd + b(C - fd).$$

We can easily observe from the tables that while this type of user is throttled in the existing mechanism, she is not in the proposed mechanism as there is no data she wants fast that is sent slow since $fd \leq C$.

Lemma 2.3: In the case $fd \leq C$, the payoff π^* for a heavy user in the proposed mechanism is at least as large as the payoff π in the existing mechanism.

Proof: Since $d > C$ for a heavy user and $r > b$, the payoff for the user in the proposed mechanism, $rfd + b(C - fd)$, is guaranteed to be at least as large as $rfC + b(C - fC)$. The additional term involving the negative value p in the payoff in the existing mechanism, $rfC + b(1-f)C + pf(d-C)$, can only lessen the sum, giving the desired guarantee. ■

Corollary 2.4: When $fd \leq C$ and $b = 0$, $\pi^*/\pi \geq d/C$.

Case 3: Light user

Lemma 2.5: The payoff for a light user is unchanged between the proposed mechanism and the existing mechanism if the user chooses to send all data in fast mode.

Proof: A light user, that is, one for whom $d \leq C$, would have all data sent fast in the existing mechanism, thus receiving a payoff of $\pi = rd$. Likewise, in the proposed mechanism, she may choose to send all data fast, again receiving a payoff of $\pi^* = rd$. ■

Lemmas 2.1, 2.3 and 2.5 immediately imply Theorem 2.6.

Theorem 2.6: The payoff for any user, heavy or light, is at least as large in the proposed mechanism as it is in the existing mechanism; for some users the payoff is strictly better.

C. Discussion

The results clearly indicate that from an individual user's standpoint, the proposed mechanism produces a more desirable payoff than the existing mechanism for heavy users. Light users (those who would not hit the cap) would see no difference if operators only throttle when a user exceeds the cap. Naturally, there are other factors to consider as well regarding the implementation and the broader implications, both for this user, fellow users, and the operator.

While the idea behind the mechanism is that users would request data to be sent fast when they truly do want it immediately and request data slow when they do not, potentially improving the performance of the overall network, it is natural to question whether it is advantageous to users to be more strategic about their requests. In fact, a user who would not exceed the cap has no incentive to not request all data fast. However, a user who will exceed the cap will be well-served by being honest about her preferences; requesting fast when she wants slow limits the amount she is able to request fast later on in the cycle (and as a heavy user she will exceed C if she requests all of her data fast), while requesting slow when she wants fast means she often will not get the speed desired.

Though each user has the same amount of fast capacity C potentially available in the proposed mechanism, concerns about fairness may understandably arise. Since there are many possible interpretations of fairness (see [6] for a survey), different users may perceive the mechanism differently. Yet, it is quite likely that a user would find the proposed mechanism fair in many regards, as each user has the same amount of fast capacity available in a month and exerts individual control over when to use it.

Successful mechanisms are often easy to implement, both in terms of an operator providing the means to do so, and the user easily knowing how to behave. The proposed mechanism is straightforward in both regards. As discussed, users can set the mode of their phone. Note that non-heavy users, or those who wish not to take advantage of the proposed mechanism, could thus leave their device always in fast mode.

III. The Collective Perspective

We now consider the collective perspective, looking at overall utilization by all users at each timestep, noting how this impacts both users in aggregate, and mobile network operators. Recall that in the existing mechanism, users run out of fast capacity based on their total usage, but in the proposed mechanism users could save their fast capacity. Thus, a natural question is: what are the right objectives to use to compare the two mechanisms from the operator's perspective?

While one metric could be the maximum total demand at any point in time, such a measure poses some concerns. Suppose a global event (whether anticipated like a sporting event or unpredictable like a natural disaster) results in an unprecedented number of users wanting to use fast capacity during a short interval of time. In the proposed mechanism, more people may have not yet used C fast data this month, and thus there may be a larger maximum demand in the proposed mechanism than the existing mechanism. At the same time, the occasional spike may be tolerated by both the operator and users as long as the frequency and/or duration are limited. Perhaps operators would perceive it as an improvement if demanded data better utilizes the existing network capacity, and is smoother overall.

We aim to provide a framework that will allow a carrier to see how this proposed mechanism would affect their network, based on their (often proprietary) actual data about user behavior, both indicating scenarios under which it is an improvement and acknowledging its limitations.

A. Additional Definitions and Assumptions

We assume that each user's data plan is on a one month cycle (30 days), and that start dates of users' cycles are uniformly distributed. Thus we may restrict our attention to the usage on an arbitrary, given day. (While there may be variations due to weekends versus weekdays, or other such patterns, an operator can apply this framework to a day of their choosing, perhaps their worst such day.) We then discretize a day into m timesteps, $t \in [0, \ldots, m-1]$.

Let N be the total number of users, indexed by i. Since all users can use fast data, this is the total number of users overall, not just those who will exceed the cap C. Let c_i denote the available fast capacity that user i has left at the beginning of the day in question (i.e. time $t = 0$). In the existing mechanism, this is thus C minus the total amount of data that user has used so far (or 0 if the value would be negative); under our proposed mechanism, it will be C minus the amount of fast data the user has requested thus far (again, limited to nonnegative values).

Let $0 < \alpha < 1$ be the portion of users who will exceed the cap C in a given month in the existing mechanism. Let H be the set of heavy users, i.e., those for whom $d_i > C$. (These were the users we focused on in Section II). Note that the number of heavy users, $|H|$, is αN. These users thus are throttled in the existing mechanism, and no longer have fast capacity, at some point in the cycle. Since each day of the cycle is assumed to have the same behavior, we thus have $\alpha N/30$ users who run out of fast capacity each day. Let $T = \sum_{i \in H} d_i$, which represents the total amount of data requested by the heavy users. (Note that there is no distinction in the existing mechanism between the total amount of data requested, and the total amount of *fast* data requested, as users cannot indicate a preference in the existing mechanism.)

Let $0 < \alpha' < 1$ be the portion of users who will exceed the cap C in a given month in the proposed mechanism. Let H' be the set of "fast-heavy" users, i.e., those for whom $f_i d_i > C$. (These were the users in Case 1 of Section II). Thus, $|H'| = \alpha' N$. Note that $H' \subseteq H$ and $\alpha' \leq \alpha$. Let $T' = \sum_{i \in H'} d_i$.

B. How Long Users Have Fast Capacity Available on Average

We now consider the average number of days it takes for specified groups of users to run out of fast capacity in the existing and proposed mechanisms. Denote by $\delta(H)$ the average number of days it takes for the heavy users H to run out of fast capacity in the existing mechanism, while $\delta^*(H)$ represents the quantity for the same users in the proposed mechanism. The quantities $\delta(H')$ and $\delta^*(H')$ are analogously defined. Next we explicitly calculate some of these values and determine some relationships between them. Since $T/(30\alpha N)$ is the amount of data demanded per heavy user per day,

$$\delta(H) = 30\alpha NC/T.$$

Since in the proposed mechanism users may choose to not always use their fast capacity,

$$\delta^*(H) \geq \delta(H).$$

Note that equality occurs when all users in H want all of their data to be transferred fast, that is, $f_i = 1$ for all $i \in H$. In fact, if all users want the same proportion of data fast with $f = f_i$ for all $i \in H$, then $\delta^*(H) = \min\{30, \delta(H)/f\}$. The upperbound of 30 is due to the length of the billing cycle, as heavy users exceed the cap C with total data, but not necessarily with fast data. It is easy to thus observe that if, for example, users want only half of their data fast (i.e. $f = .5$), the expected number of days until they run out of fast capacity doubles between the existing mechanism and the proposed mechanism. When the f_i are not all equal, but the values are known to the operator, $\delta^*(H)$ can be explicitly computed, analogous to the formula above for $\delta(H)$. Note that the f_i values need not be linked to particular users; knowing how many (or what fraction of) users have particular f_i values is sufficient to calculate $\delta^*(H)$.

The values for H' are calculated analogously, giving

$$\delta(H') = 30\alpha' NC/T', \text{ and } \delta^*(H') \geq \delta(H').$$

C. How Many Users Have Fast Capacity Left

Let $n_t \leq N$ and $n_t^* \leq N$ denote the number of users with fast capacity left at time t in the existing and proposed mechanisms, respectively. Thus n_0 (respectively, n_0^*) is the number of users with fast capacity left at the beginning of each day, that is, users with $c_i > 0$ in the existing (proposed) mechanism. We now compute the expected values of n_t and n_t^*.

Users who have fast capacity left at a given time t can be grouped into three categories: those who never run out of fast capacity, those who run out today, and those who will run out 'soon' (based on δ values).

We focus first on the existing mechanism. Since only an α fraction of the users will run out of fast capacity in a given month, $(1-\alpha)\cdot N$ will not run out at all this month. The users who do run out do so uniformly on different days, so $\frac{1}{30}\alpha N$ will run out today. We assume the existence of some function $\mu(t)$ for $t \in \{0,\ldots,m-1\}$ that describes how quickly the people who will run out of fast capacity that day use it up. In particular, $\mu(t)$ is a decreasing function from 1 to 0 quantifying the proportion of people who will run out of fast capacity that day who have not yet done so. Note that a reasonable approximation for $\mu(t)$ may be determined based on an operator's historical usage data. Hence, the users who run out today at time t number $\mu(t)\frac{1}{30}\alpha N$. Finally, we account for users who will run out of capacity this month, but not today; since the average number of days it takes for a user to run out of fast capacity is $\delta(H)$, this adds a term of $\frac{\delta(H)-1}{30}\alpha N$. Thus,

$$E[n_t] = (1-\alpha)N + \mu(t)\frac{1}{30}\alpha N + \frac{\delta(H)-1}{30}\alpha N.$$

In our proposed mechanism, only users in $H' \subseteq H$ will run out of fast capacity during the billing cycle. Thus, when calculating n_t^* the primed values α' and H' are used. Additionally, we define $\mu^*(t)$ to be (like its unstarred counterpart) a decreasing function from 1 to 0 quantifying the proportion of people who will run out of fast capacity that day who have not yet done so. Note that we would typically expect $\mu^*(t)$ to drop off more slowly than $\mu(t)$ since in the proposed mechanism it is likely that some users save some portion of their fast capacity each day. We can still group the users who have fast capacity left at a given time t into three categories: those who never run out of fast capacity, those who run out today, and those who will run out 'shortly.' Thus, the expected number of users with available fast capacity is

$$E[n_t^*] = (1-\alpha')N + \mu^*(t)\frac{1}{30}\alpha' N + \frac{\delta^*(H')-1}{30}\alpha' N.$$

Note that these values are in expectation because $\delta(H)$ is an average value. Since only the second terms in the expressions for $E[n_t]$ and $E[n_t^*]$ depend on t, and $\mu(t)$ is a decreasing function of t, the following observation is immediate.

Observation 3.1: The expected values of $E[n_t]$ and $E[n_t^*]$ are maximized at $t = 0$, that is, the start of each day.

A Specific Example: Given the large number of parameters that are either specific to individual consumers or the operator, we consider some specific, reasonable values to aid in comparing n_t and n_t^*. Our choices along with some of the rationale for these decisions are detailed as follows, and Figure 1 illustrates how n_t and n_t^* compare during the day.

We first make the simplifying assumption that $f_i = f$ for all users, previously motivated by the ability to compute δ^* in terms of δ. In particular, we consider $f = .5, .7, .9$. We plot the results as a percentage of the users N, so its specific value is not important. We divide a day into minutes, so t ranges from 0 to 1440. We assume a cap of $C = 5$ GB, that $\alpha = .05$ (so five percent of users are throttled under the existing mechanism, consistent with industry statements), and the majority exceed the cap under the proposed mechanism, with $\alpha' = .035$. We assume that the average amount of data wanted by a user in H is 10GB, while for users in H' it is 12GB, recalling these are all heavy users who exceed the 5GB cap. Since μ is guaranteed only to be a decreasing function from 1 to 0, we let it be the quadrant of the ellipse that does so, with fewer users running out in early morning hours, and more later throughout the day. Since this mechanism has not been implemented, and thus data on μ^* is not available, we let $\mu^*(t) = \mu(t)$.

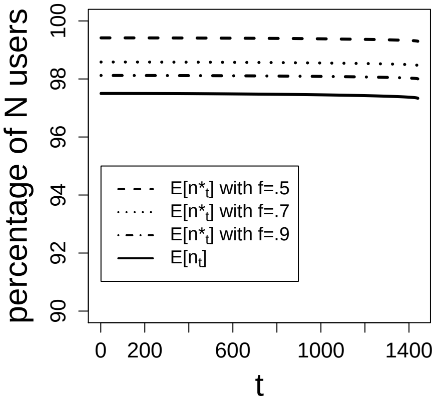

Fig. 1. Expected percentage of users with remaining fast capacity, both in the existing mechanism and the proposed mechanism with three different values of f, throughout the day, under a number of specified assumptions.

Observe that while Figure 1 shows that, not surprisingly, more users are expected to have fast capacity available in the proposed mechanism, the increase is small compared to the total number of users, for all three f values with these specified parameters. Operators ideally can use this framework along with detailed data about their users' behavior, individually or in aggregate, to determine the effect the proposed mechanism will have for them. Of course, much of this information is considered proprietary and not widely available.

D. Expected Data Transfer Demanded at time t

Since users having fast capacity left is not the same as requesting data at a particular time, we move on to providing a framework for calculating the expected amount of data transferred at any point in time. Naturally, this will be highly individualized, but operators can often make predictions based on historic usage as well as knowledge of current events.

Let Y_t (respectively Y_t^*) be the total amount of data transfer demanded at time t in the existing (proposed) mechanism. We provide expressions for $E[Y_t]$ and $E[Y_t^*]$. We assume that at any given time a user either wants to transfer NO data, SMALL data, or LARGE data. While an operator may decide on its own cutoff between SMALL and LARGE, one possible split is at 1MB, so that SMALL would include many emails with attachments, while most music and video would be in the LARGE category. When the user is able to specify the speed of data, LARGE requests that go 'slowly' may be treated as SMALL data, since they may effectively be parcelled out in small pieces for each of a number of (not necessarily successive) timesteps. We say that the probabilities of each of the three are p_0, p_{small}, and p_{large} with $0 \leq p_x \leq 1$ and $p_0 + p_{small} + p_{large} = 1$. Realizing that some people may alter their behavior in response to the new mechanism, we similarly define starred probabilities for the proposed mechanism, namely p_0^*, p_{small}^*, and p_{large}^* with $0 \leq p_x^* \leq 1$ and $p_0^* + p_{small}^* + p_{large}^* = 1$.

Without loss of generality, we may assume that SMALL data has size 1, scaling LARGE data as needed. Of the n_t people who have fast capacity left, $p_0 = 1 - p_{small} - p_{large}$ will transfer no data, p_{small} will transfer SMALL units, and p_{large} will transfer LARGE units. Thus, those n_t people contribute the following expected demand:

$$p_0 \cdot n_t \cdot 0 + p_{small} \cdot n_t \cdot 1 + p_{large} \cdot n_t \cdot \text{LARGE}$$

In addition, the $N - n_t$ people who have exceeded the cap C may still want to transfer data; however, if they wish to transfer LARGE data, they are forced to transfer it slowly (broken up into small data), thus having expected demand of

$$p_0 \cdot (N - n_t) \cdot 0 + p_{small} \cdot (N - n_t) \cdot 1 + p_{large} \cdot (N - n_t) \cdot 1$$

The total data demanded by all N users is thus, in expectation,

$$E[Y_t] = p_{small} N + p_{large}(N + (\text{LARGE} - 1)n_t).$$

The analogous calculations in the proposed mechanism give

$$E[Y_t^*] = p_{small}^* N + p_{large}^*(N + (\text{LARGE} - 1)n_t^*).$$

E. Discussion

Operators and consumers alike should be aware of one potential limitation of the proposed mechanism. Since users have the ability to save some of their fast capacity throughout the month, if users wanted to save enough fast capacity for a particular event (say a football game), and chose to do so, there could potentially be more requests for fast data during that particular interval than there would have been in the existing mechanism. Yet, this limitation is not necessarily fatal. In fact, if this event is that popular that heavy users will save their demand, it is worth considering that light users alone could request more than the system infrastructure allows, and thus would not achieve their requested speed due to network congestion alone, not throttling.

Users are likely to be happier overall with many features of the proposed mechanism. Having choices and control may make users less discontented if they are throttled. In addition, having fast capacity available for a longer interval, on average, allows the user better access to what they want. When operators are able to provide these benefits without any additional infrastructure and via a simple mechanism, it is advantageous to them. While different operators may see different improvements, based on their users, the framework provided may help them to evaluate the mechanism for their unique needs.

Note that while many users (light users in particular) will not change their behavior in response to the proposed mechanism, it is conceivable that users could change their total data consumption, either increasing it or decreasing it as they consider how they want to allocate their fast data. Yet the demands placed on the system by fast data are still capped as they were before, and may in fact be lower if users actively participate in the choices provided by the proposed mechanism. Notably, users who do not exceed the cap but are concerned about doing so will likely exercise the ability to control their traffic provided by the proposed mechanism, potentially benefiting both users and operators.

IV. Conclusion

We proposed a simple mechanism that permits users to choose which data they want transferred fast, and which data they are willing to have transferred slow, that can mitigate some of the effects of data plan throttling. We showed that from an individual user's perspective, having this choice can only improve the user's payoff. While some unlimited data plan users may still be throttled each month, the ability to have an additional layer of control may reduce user dissatisfaction. At the same time, users are not required to adopt this new mechanism, and by default can have all of their data transferred fast until they would normally be throttled by their provider. We also showed that collectively users have fast capacity available for more days, on average, as well as an increase in the average number of users that have fast capacity available at various times throughout each day. Yet, we acknowledge there are limitations to this mechanism, and provide a framework for operators to use with their user data to determine the larger impacts. Future work includes performing simulations with various demand distributions noting typical diurnal fluctuation, as well as considering additional strategic interactions between the users and the operator.

References

[1] AT&T Data Usage Information & FAQs, http://www.att.com/esupport/datausage.jsp, Retrieved Feb. 7, 2013.

[2] T. Beard, G. Ford, L. Spiwak, and M. Stern, *Wireless Competition Under Spectrum Exhaust*, Phoenix Center Policy Paper No. 43, Feb. 1, 2012. Available at SSRN: http://ssrn.com/abstract=2131445

[3] B. Chen, "Victor in Throttling Case Publishes Guidelines on Taking AT&T to Court", *Bits*, New York Times, Feb. 28, 2012. Web. http://bits.blogs.nytimes.com/2012/02/28/att-small-claims/, Retrieved Feb. 11, 2013.

[4] S. Haykin, *Cognitive radio: brain-empowered wireless communications*, IEEE J.Sel. A. Commun. 23, 2 (Sept. 2006), pp. 201-220.

[5] M. Hui and J. Bateson, *Perceived Control and the Effects of Crowding and Consumer Choice on the Service Experience*, Journal of Consumer Research, Vol. 18, No. 2 (Sept. 1991), pp. 174-184.

[6] M.Z. Kwiatkowska, *Survey of fairness notions*, Information and Software Technology, Volume 31, Issue 7, Sept. 1989, pp. 371-386.

[7] D. Levinson and A. Odlyzko, *Too expensive to meter: The influence of transaction costs in transportation and communication*, In special issue on Networks: Modeling and control, Phil. Trans. Royal Soc. A, vol. 366, no. 1872, 2008, pp. 2033-2046.

[8] A. Odlyzko, B. St. Arnaud, E. Stallman, and M. Weinberg, *Know Your Limits: Considering the Role of Data Caps and Usage Based Billing in Internet Access Service*, Public Knowledge, May 2012.

[9] S. Sen, C. Joe-Wong, S. Ha and M. Chiang, *A Survey of Broadband Data Pricing: Past Proposals, Current Plans, and Future Trends*, accepted in ACM Computing Surveys, 2013.

[10] Verizon Wireless Network Optimization, http://support.verizonwireless.com/information/data_disclosure.html, Retrieved Feb. 7, 2013.

[11] H. Zhou, K. Sparks, N. Gopalakrishnan, P. Monogioudis, F. Dominique, P. Busschbach and J. Seymour, *Deprioritization of heavy users in wireless networks*, Communications Magazine, IEEE, vol.49, no.10, pp. 110-117, Oct. 2011.