On Economic Heavy Hitters: Shapley value analysis of 95th-percentile pricing

Rade Stanojevic
Telefonica Research

Nikolaos Laoutaris
Telefonica Research

Pablo Rodriguez
Telefonica Research

ABSTRACT

Cost control for the Internet *access providers* (AP) influences not only the nominal speeds offered to the customers, but also other, more controversial, policies related to traffic shaping and discrimination. Given that the cost for the AP is determined by the peak-hour traffic (e.g. through the 95th-percentile), the individual user contribution towards the aggregate cost is not a linear function of its byte usage. In this paper we propose a metric for evaluating the contribution each individual user has on the peak demand, that is based on Shapley value, a well known game-theoretic concept. Given the computational complexity of calculating the Shapley value, we use a Monte Carlo method for approximating it with reasonable accuracy. We employ our methodology to study a dataset that logs per-subscriber temporal usage patterns over one month period for 10K broadband subscribers of a European AP and report observed results.

Categories and Subject Descriptors

C.2.3 [**Computer-Communications networks**]: Network Operations; Network monitoring; C.4 [**Performance of systems**]: Measurement techniques

General Terms

Measurement, Economics

Keywords

Heavy-hitters, Network economics, Net-neutrality, Shapley value, Monte-Carlo method

1. INTRODUCTION

A large number of Internet Access Provider (AP) adopted flat-rate pricing as a de-facto standard for charging of broadband services as such pricing appears to be preferred by the customers [18]. This creates many difficulties for the APs as it does not allow APs to transparently control the uplinking (transit + infrastructure) costs and forced many APs to create nontransparent rules for traffic shaping and violate net-neutrality as a means for control of their costs [9]. Using the terminology of [22], uplinking costs are the single most expensive component of the costs for broadband connectivity for a majority of currently used technologies, including DSL, cable and WiFi (mash and access point). An important property of the uplinking costs (influenced by the transit costs and the cost of infrastructure) is that they are determined by the peak demand (e.g. through the 95th-percentile) rather than average demand, which makes it hard to assess per-customer contribution towards these costs.

Flight ticket prices typically depend on the time of travel and hotel rooms in tourist resorts are less expensive during off-season. Similarly, a byte downloaded in peak-hour costs more (for the provider) than a byte of traffic in off-peak hours. In this paper we study per-user contribution in the AP peak hour demand. More precisely, we measure per-customer contribution towards the 95th-percentile of the aggregate demand series[1]. For the purpose of quantifying per-user contribution to the 95th-percentile, we use Shapley value, an intuitive concept from coalitional game theory that characterizes fair cost sharing among involved players (customers). Shapley value framework allows us to: (1) accurately quantify the contribution of each customer towards the peak-hour traffic, (2) analyze the relationship between the aggregate usage (in bytes) and the peak-hour contribution and (3) formally measure how cost of bandwidth is related to the demand pattern. We validate our methodology over a dataset that logs temporal usage of 10K broadband customers of a European AP.

Note that we talk about costs customers generate for the AP rather than the price they pay; retail prices are often strongly impacted by other market, competition and social factors [22]. For various mechanisms for pricing the communication services in the context of revenue (or social welfare) maximization, see [6].

1.1 Toy example

For measuring the peak demand we use the 95th-percentile of the aggregate demand, the most standard measure for billing of the transit traffic and an indicator of the network utilization, used for the dimensioning of the infrastructure; see Appendix A for a brief description. To understand the concept of Shapley value and how it applies to the 95th-

[1]We stress, however, that the framework is general enough to accommodate any other metric that measures the peak demand.

IMC'10, November 1–3, 2010, Melbourne, Australia.

percentile billing let us consider a synthetic example of an AP ISP providing service to only two users that have demand patterns that are depicted in Figure 1. The user 1 generates a demand of $1Mbps$ during the whole day except for the four-hour period [15-19h]. The user 2 is idle for 22 hours and generates $3Mbps$ traffic during two hours: [16-18h]. The 95-th percentile of the aggregate user demand is the peak-hour traffic $v_{95th} = 3Mbps$ and the price the ISP would need to pay to its transit provider is $v_{95th} \cdot A_0$ (where A_0 is the price in USD per $Mbps$). The following question arises: What is the fair cost sharing among the two involved users? The Shapley value concept gives an answer to this question and the intuition behind it is described bellow.

If there was only user 1 or only user 2 in the system, the 95th percentile would have been:

$$v_{95th}(\{1\}) = 1Mbps \quad \text{and} \quad v_{95th}(\{2\}) = 3Mbps$$

respectively. As we already observed, the 95th percentile of the union of these two users is

$$v_{95th}(\{1,2\}) = 3Mbps.$$

The Shapley value of user i, ϕ_i is now the average marginal contribution that user i imposes to the coalition cost. In other words:

$$\phi_i = \frac{1}{2}\left(v_{95th}(\{i\}) + (v_{95th}(\{i,i'\}) - v_{95th}(\{i'\}))\right),$$

where $\{i'\} = \{1,2\} \setminus \{i\}$. In our example the per user Shapley values are:

$$\phi_1 = 0.5Mbps \quad \text{and} \quad \phi_2 = 2.5Mbps.$$

Thus by entering the coalition, the fair cost sharing of the 95th-percentile $v_{95th}(\{1,2\}) = 3Mbps$ would be the one in which the user 1 is accounted for $\phi_1 = 0.5Mbps$ and the user 2 for $\phi_2 = 2.5Mbps$. The nature of the 95th-percentile pricing is such that even though the user 1 generates in total 3.3 times more traffic than user 2, its contribution to the 95th-percentile is 5 times lower.

COMMENT 1. *We can learn two lessons from the above example: firstly, the user that sends/receives more data does not necessarily have higher impact on the 95th-percentile; and secondly, even if a user does not generate any traffic in the peak hours that does not imply that its impact towards the 95th-percentile is zero. Shapley value balances between these two extremes (aggregate usage and peak-only usage) by evaluating the average marginal contribution of each user (eq. 1).*

1.2 Summary of contributions

Briefly, the main contributions of this paper are the following:

- We develop a new methodology for studying heavy users in an operational ISP. We use the standard concept from cooperational game theory, known as Shapley value, to quantify per-user cost contribution in the context of 95th-percentile pricing.
- Using the Shapley value methodology, we study a month-long dataset that tracks temporal usage patterns from 10K broadband users of a European ISP. We quantify several relevant metrics over this dataset. In particular we find that for approximately 10% of users, the relative cost contribution (Shapley value) is less than half of the relative byte usage (off-peak users), and that for additional 10% of users the relative cost contribution is more than twice of their relative byte usage (peak users). Finally, we use the Shapley value framework to formalize the intuitive wisdom "a byte in the peak-hour has a higher value/cost than an off-peak byte" by quantifying the hourly per-byte bandwidth prices that approximate best the measured Shapley value.

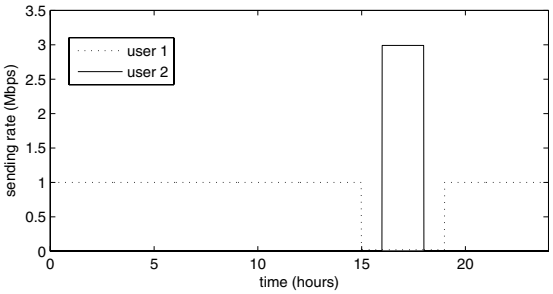

Figure 1: Toy example. Two users with different demand pattern.

2. APPROXIMATING SHAPLEY VALUE

In this section we will briefly introduce the Shapley value concept for general cooperative games, relate it to our framework in which the cost of a user coalition is determined by the 95th-percentile of the traffic they generate and propose a randomized method for efficiently computing SV for large number of players.

2.1 Shapley value: definition

Consider a set $\mathcal{N}$ of N players[2]. For each subset (coalition) $S \subset \mathcal{N}$ let $v(S)$ be the cost of coalition S. In other words if S is a coalition of players which agree to cooperate, then $v(S)$ determines the total cost from this cooperation.

For given cost function v, the Shapley value is a (uniquely determined) vector $(\phi_1(v), \ldots, \phi_N(v))$ defined bellow that is "fair" in that it satisfies four intuitive properties (see [20, 23]) for sharing the cost $v(\mathcal{N})$ that exhibits the coalition of all players. It can be shown that Shapley value of player i is determined by

$$\phi_i(v) = \frac{1}{N!} \sum_{\pi \in S_N} (v(S(\pi, i)) - v(S(\pi, i) \setminus i)) \qquad (1)$$

where π is a permutation or arrival order of set $\mathcal{N}$ and $S(\pi, i)$ is the set of players arrived in the system not later than i. In other words, player i is responsible for its marginal contribution $v(S(\pi, i)) - v(S(\pi, i) \setminus i)$ averaged across all $N!$ arrival orders π. Note that the Shapley value defined by (1) satisfies (so called efficiency) property:

$$\sum_{i \in \mathcal{N}} \phi_i(v) = v(\mathcal{N}).$$

2.2 The 95th-percentile cost

The 95th-percentile billing is a method of measuring bandwidth usage based on peak utilization, defined in Appendix

[2]We interchangeably use terms player, user, customer and subscriber.

A. Informally it measures close-to-peak demand but it also allows usage to exceed a specified threshold for brief periods of time without the financial penalty.

The setup over which we apply the Shapley value framework is the following. We have the set $\mathcal{N}$ of N users that generate traffic over a charging period, say one month. The charging period is split into T sampling intervals, and at time $t \in [1, T]$, user i generates the traffic $Z_i(t)$ (measured in bytes). For a time series $D = (D(1), \ldots, D(T))$, the 95th-percentile $P_{95th}(D)$ is defined as the $\lceil \frac{T}{20} \rceil$-th largest number of the time series. For a coalition S of users the cost they generate is determined by the 95th-percentile of the aggregate demand pattern they generate:

$$v(S) = P_{95th}(\sum_{i \in S} Z_i(1), \ldots, \sum_{i \in S} Z_i(T)).$$

Given the cost function $v(\cdot)$, the contribution of each user to the 95th-percentile of the aggregate traffic $v(\mathcal{N})$ is defined by the Shapley value defined by (1). From the definition, one can notice that the 95th-percentile does not decrease by adding new users to the coalition, therefore implying that the cost function v is monotone:

$$(\forall S \subset \mathcal{N})(\forall i \in \mathcal{N}) \quad v(S \cup i) \geq v(S).$$

The monotonicity of the cost function v implies that the Shapley value of each user is indeed nonnegative.

2.3 Approximating Shapley value

Brute force application of formula (1) is computationally unfeasible once N becomes greater than 100. For APs with thousands (or millions) of subscribers such exact computation is not possible. In this Section we describe a simple randomized method for approximating the Shapley, that can scale with datasets of tens of thousands (if not millions) of subscribers.

The idea of the method is simple. The Shapley value of user i defined by (1) can be seen as the marginal cost increase by user i, averaged over all $N!$ arrival orders. In the example from Section 1.1, $N = 2$ and there are 2 arrival orders: $\pi_1 = (1, 2)$ and $\pi_2 = (2, 1)$ and the user 1 and user 2 Shapley values are

$$\phi_1 = \frac{1}{2}\left((v(\{1\}) - v(\emptyset)) + (v(\{1,2\}) - v(\{2\}))\right) =$$

$$= \frac{1}{2}\left((1 - 0) + (3 - 3)\right)) = 0.5.$$

$$\phi_2 = \frac{1}{2}\left((v(\{1,2\}) - v(\{1\})) + (v(\{2\}) - v(\emptyset))\right) =$$

$$= \frac{1}{2}\left((3 - 1) + (3 - 0)\right)) = 2.5.$$

While computing the exact Shapley value through the formula (1) is straightforward for small N, it becomes unfeasible for $N > 50$, as the number of different permutation orders grows with $N!$. However, the computational complexity can be significantly reduced by using the Monte Carlo method.

Instead of calculating the exact Shapley value as the average cost contribution across all $N!$ arrival orders, we estimate the Shapley value as the average cost contribution over a set Π_k of k randomly sampled arrival orders (permutations).

$$\hat{\phi}_i(v) = \frac{1}{k} \sum_{\pi \in \Pi_k} (v(S(\pi, i)) - v(S(\pi, i) \setminus i)) \qquad (2)$$

The parameter k determines the error between the real Shapley value and its estimate: the higher k the lower the error. So basically, one can control the accuracy of the estimators by increasing the number of sample permutation orders (see Section 3.2).

PROPOSITION 1. *The estimator $\hat{\phi}_i(v)$ is an unbiased estimator of the real Shapley value $\phi_i(v)$.*

PROOF. See [21]. □

Thus the Shapley value estimator (2) is unbiased. However the variance of the estimator is hard to model and in Section 3.2 we present empirical evidence that for reasonably small sample size (say, $k = 1000$) the estimator exhibits small variance, especially for the top users.

PROPOSITION 2. *The estimated Shapley values satisfy the efficiency property:*

$$\sum_{i \in \mathcal{N}} \hat{\phi}_i(v) = v(\mathcal{N}).$$

PROOF. See [21]. □

3. EMPIRICAL RESULTS

In this section we present the empirical results obtained by analyzing the dataset of around 10K broadband users of a major European ISP. In Section 3.1 we describe the dataset, then in Section 3.2 we analyze the accuracy of the randomized method for calculating Shapley value. We proceed by analyzing the correlation between per-user aggregate usage and its Shapley value in Section 3.3 and then in Section 3.4 we quantify the relative cost of bandwidth in time that would best approximate the Shapley value. Additional empirical results, related to the consistency of the Shapley value over time, as well as additional discussion on the relative cost of bandwidth, can be found in the Technical report [21].

3.1 Dataset description

The dataset consists of around 10K ADSL users of a major access provider in one European country. For each customer, its downstream and upstream consumption (in bytes) is captured during each hour for 30 days (thus spanning 720 hours). These users represent a random sample of ADSL users of the ISP and have diverse uplink/downlink capacities in the ranges of 256Kbps–10Mbps, and 1Mbps–20Mbps, respectively. The downstream traffic dominates the upstream traffic in the ratio 4 : 1, which is consistent with the recent findings from another European access provider ISP [15]. Virtually all ADSL users from the dataset pay flat-fee, without incentives to shift their traffic to the off-peak hours[11]. We stress that the empirical results derived from this dataset are mainly qualitative, used for the purpose of validating the Shapley value methodology and basic properties of Shapley value, and results derived here should not be generalized for other types of environments such as campus, backbone or enterprize networks.