Fair, Flexible and Feasible ISP Billing*

Vamseedhar Reddyvari Raja[1], Srinivas Shakkottai[1], Amogh Dhamdhere[2], and Kc Claffy[2]
[1]Texas A&M University, College Station, TX
[2]CAIDA/UCSD, La Jolla CA

ABSTRACT

The 95^{th} percentile method for calculating a customer's billable transit volume has been the industry standard used by transit providers for over a decade due to its simplicity. We recently showed [1] that 95^{th} percentile billing can be *unfair*, in that it does not reflect a customer's contribution to the provider's peak load. The 95^{th} percentile method is also inflexible, as it does not allow a provider to offer incentives to customers that contribute minimally to the provider's peak load. In this paper we propose a new transit billing optimization framework that is fair, flexible and computationally inexpensive. Our approach is based on the *Provision Ratio*, a metric that estimates the contribution of a customer to the provider's peak traffic. The proposed mechanism has fairness properties similar to the optimal (in terms of fairness) Shapley value allocation, with a much smaller computational complexity.

1. INTRODUCTION

Transit providers are an important piece of the Internet ecosystem, providing customers with access to the rest of the Internet. But the future role of transit providers is uncertain, given continuously falling transit prices and increased propensity for networks to interconnect directly (peering) [2, 3], essentially routing around traditional transit providers. These business risks increase the pressure on transit providers to optimize their transit billing schemes to remain competitive. This work offers a new metric and associated framework to support such optimization.

There are two components to today's Internet transit billing scheme: the volume of traffic for which a customer network is billed (the *billing volume*), and a function that computes price based on this volume. The industry standard for determining the billing volume is the 95^{th} *percentile* method [4, 5]: a transit provider measures the utilization of a customer link in 5-minute bins throughout a month, and then computes the 95^{th} percentile of these values as the billing volume. The 95^{th} percentile method has three attractive properties: it is simple to implement; it uses data that the provider typically already collects; and it approximates the load a customer imposes on the provider's network while forgiving a few anomalous traffic bursts. An important aspect of the second billing component (the pricing function) is that providers generally offer volume discounts, such that the per-bit price decreases as billing volume increases [2].

In our recent examination of the first aspect of transit billing – the 95th percentile method – we showed that this mechanism can be *unfair*, as it charges all customers at the same percentile, and does not account for the load that a customer actually imposes on the provider. While solutions such as the Shapley value method exist to assign billing volumes to customers in a fair manner, they are computationally too expensive to implement at scale without approximations. Further, those methods are not flexible enough to accommodate all the constraints of transit providers, e.g., restricting billing percentiles to a certain range or offering incentives to certain classes of customers.

In this work we present a framework for determining the billing volume for each customer in a manner that is fair, computationally inexpensive, and flexibly allows the provider to provide incentives (discounts) to certain customers. Our billing framework is based on a new metric called the *Provision Ratio*, which reflects a customer's contribution to the provider's peak traffic load. By assigning billing volumes per-customer, providers can exercise fine-grained control over their billing and provide discounts to customers that contribute minimally to the provider's peak traffic. The transit provider can use such incentives as a means for attracting new customers.

2. MEASURING BILLING VOLUMES

We present a framework for percentile-based measurement of billing volumes. Consider a transit provider with N customers indexed by $i,\ i \in \{1, 2, \ldots, N\}$. Each month, the transit provider must determine the billing volumes of each customer.

The relationship between billing volume and billing percentile can be expressed using the cumulative distribution function (CDF) of the customer network's traffic. First, both the inbound and outbound traffic volumes are measured in 5-minute intervals, and are used to calculate the average transmission rates during each interval. Denote the empirical CDFs of customer network i's transmission rates by $\mathcal{F}_{i(in)}(.)$ and $\mathcal{F}_{i(out))}(.)$, for inbound and outbound directions, respectively. Also, denote the inverse cumulative distribution functions by $\mathcal{F}^{-1}_{i(in)}(.)$ and $\mathcal{F}^{-1}_{i(out)}(.)$. If the CDF function is not one-to-one, the inverse will be an interval (due to monotonicity of the CDF function). If this is the case, we take the supremum of the interval to be the value of the inverse, *i.e.,*

$$\mathcal{F}^{-1}_{i(in)}(y) = \sup \left\{ x | \mathcal{F}_{i(in)}(x) = y \right\}, \qquad (1)$$

and similarly for outbound traffic. We then decide on whether the traffic is inbound or outbound dominated by comparing the 95^{th} percentile volumes of the two, *i.e.,* we compare $V_{i(in)}(0.95) = \mathcal{F}^{-1}_{i(in)}(0.95)$ with $V_{i(out)}(0.95) = \mathcal{F}^{-1}_{i(out)}(0.95)$. We choose the overall CDF of the customer i's traffic to be the one with the larger

*Research was supported in part by NSF grants CNS-1149458, CNS-1017064 and a Cisco URP grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or Cisco.

Contact email: vamseedhar.reddyvaru@neo.tamu.edu.

95^{th} percentile . Thus, if $V_{i(in)}(0.95) > V_{i(out)}(0.95)$, then we set $\mathcal{F}_i(x) = \mathcal{F}_{i(in)}(x)$. Correspondingly, the volume billed by the 95^{th} percentile scheme is $V_i(0.95) = \mathcal{F}_i^{-1}(0.95)$, and the sum total volume of billed traffic is $V_{95} = \sum_{i=1}^{N} V_i(0.95)$. As described in Section 1, the 95^{th} percentile method is unfair because it does not account for the fact that the temporal traffic profile of customers might impose very different loads on the transit provider. For instance, a customer whose traffic is concentrated in the peak periods of overall traffic would require the transit provider to provision more capacity than one whose traffic is in the off-peak periods. A fair scheme should ensure that the amount of resources used by a customer should be reflected in its corresponding billing volume.

2.1 Shapley Value

The Shapley value is a means of representing the contribution of each group member to the overall value of a group. It is generally considered *fair* as it is equal to the average marginal increase in the value of the group due to the presence of each member. It is also *efficient* in that the sum of Shapley values is the total value of the group. In our case, the members are the customer ISPs, and the value is actually the cost of the network capacity needed to support their traffic. In earlier work, Stanojevic *et al.* [6] showed that the Shapley value can be used to assign costs to different customer ISPs based on their traffic profiles. We describe a similar scheme, modified to convert Shapley values into billing percentiles.

In order to calculate the Shapley value, we first need to define a value function. This value function maps each possible subset of customers to a real number. We define the value function of a group as the 95^{th} percentile of the total traffic obtained by adding the traffic of all members in the group. The ISP needs to provision for this quantity of traffic (it is also the volume for which the ISP would be billed by its own transit provider). The Shapley value (ϕ_i) of customer i is obtained by the equation $\phi_i = \frac{1}{N!}\sum_{\pi \in \Pi} \left(\mathcal{V}(S(\pi, i) - \mathcal{V}(S(\pi, i)\backslash i)\right)$, where $\mathcal{V}$ is the value function, Π is the set of all possible permutations of players $\mathcal{N}$ and $S(\pi, i)$ is the set of all customer ISPs in ordering π before i including i. Essentially, to calculate the Shapley value of a customer i, we calculate the difference in the value of a group with i and without i, and average over all possible groups.

Once we determine the Shapley values, we normalize them according to $\sigma_i = \phi_i / \sum_{j=1}^{N} \phi_j$ Then, if we wish to ensure that the same volume is billed as with the 95^{th} percentile method, but assign these volumes according to the Shapley value, we should choose billing volumes as $S_i = \sigma_i * V_{95}$. Finally, for purposes of comparison with the 95^{th} percentile scheme, we can translate these values to a Shapley value percentile (SVP) using $p_i^S = \mathcal{F}_i(S_i)$.

We determined the SVPs for customers over a month using two different data traces. We provide more details on the data in Section 4. We found that the SVP is as low as 0.58 for some customers, and as high as 0.98 for some others. Thus, some customers' billing volume should be lower (than the 95th percentile) since their resource usage corresponds to off-peak periods. The 95^{th} percentile billing mechanism ignores this difference between customers, and calculates all their billing volumes using the same percentile. From a resource usage perspective, this means that some customers are billed for too much, while others are billed for too little.

Though the SVP gives a fair way of calculating billing volumes, the calculation of Shapley value is computationally intensive as the complexity is $\mathcal{O}(N!)$. For a network with 50 customers, this is of the order of 10^{64}. In our earlier work [1], we developed a proxy for Shapley value called the Provision Ratio. This method of calculating contribution addresses complexity issue of the Shapley value.

2.2 Provision Ratio

We define the Provision Ratio of a network as the average fraction of its traffic that occurs during the times when the provider's total traffic is at its peak [1]. Peak slots are time slots during which the total traffic exceeds a threshold. We used a threshold of 95^{th} percentile of total traffic to define peak slots. However, the Provision Ratio is fairly robust to the threshold that we choose.

$$R_i = \frac{\text{Total traffic of } i \text{ during peak slots / \# of peak slots}}{95^{th} \text{ percentile of } i\text{'s traffic}}$$

The Provision Ratio is the average traffic during peak slots divided by the peak traffic (ignoring the top 5% bursts). This is an important parameter for billing because it captures the contribution of a customer network's traffic to the provider's peak. We showed in earlier work [1] that, in general, the order of two customers' Shapley values is also the order of their Provision Ratios, which enables us to use Provision Ratio as a low complexity alternative. As with the Shapley value, we normalize the Provision ratio using $\rho_i = R_i / \sum_{j=1}^{N} R_j$, and we can design a percentile-based billing mechanism using these normalized values.

While both the Shapley value and Provision Ratio can be translated into percentile-based volume measurement methods, they do not directly allow us to restrict the range of acceptable billing percentiles. As our objective is to incentivize customers to occupy off-peak periods, while not excessively dis-incentivizing those who do not, we desire a framework that incorporates both fairness as well as flexibility in choosing billing percentiles.

3. OPTIMIZATION FRAMEWORK

We seek a scheme whereby customers occupying off-peak periods are given rebates, while those that do not are charged extra. However, we also wish to ensure that the billing percentiles are not overly large or small. Finally, this must be done at no loss of net revenue to the transit provider. How can we achieve these goals?

Suppose that the transit provider uses a price function $\mathcal{B}(.)$ to translate traffic volumes into dollar charges. Often, this function is (approximately) concave and increasing [2] to ensure discounts for large volume customers. We do not propose to alter the billing function, but instead use $\mathcal{B}(.)$ as is. Let the revenue obtained through 95^{th} percentile based volume measurement be M_{95}. Then the solution to the following optimization problem attains our goals:

$$\max_{\{p_i\}} \sum_{i=1}^{N} (0.95 - p_i)\,\omega_i - \gamma(\sum_{i=1}^{N} (0.95 - p_i)^2 \qquad (2)$$

$$s.t. \quad L \leq p_i \leq H, \quad \forall i \in \{1, 2, \ldots, N\}. \qquad (3)$$

$$\sum_{i=1}^{N} \mathcal{B}\left(\widetilde{\mathcal{F}}_i^{-1}(p_i)\right) \geq M_{95} \qquad (4)$$

Here, the objective (2) is to ensure that the billing volume percentile is reduced below 0.95 as much as possible, i.e. provide the maximum possible incentives to customers. To provide incentives for off-peak customers, we set the weight $w_i = (1/\rho_i)^{\alpha}$, where $\alpha \geq 1$. Since the weight varies inversely with the normalized Provision Ratio, maximizing the objective would assign larger p_i values to customers with smaller weights i.e., high occupancy during peak times. The second term in the objective is to smooth it, as otherwise the solution would be to set p_i to extreme high or low values. Parameter γ is used to decide the desired smoothing.

We next have a (convex) constraint (3) that ensures that the percentiles output by the optimization lie in an acceptable interval between $[L, F]$. Constraint (4) ensures that the transit provider does not suffer any loss of revenue (as compared to 95^{th} percentile based volume measurements). As defined above, $\mathcal{B}(.)$ is a concave billing function. Now, since the inverse CDF of traffic, $\mathcal{F}_i^{-1}$, is empirical, it might not have any particular form. Hence, we approximate it using a concave function $\widetilde{\mathcal{F}}_i^{-1}$ in the range $[L, F]$. In practice, we employed an approximation of the form $\widetilde{\mathcal{F}}^{-1}(x) = a + bx + c\sqrt{x}$ with $c \geq 0$. Notice that the concave approximation immediately implies that the constraint becomes convex.

Our problem formulation is in the form of convex optimization, and hence the solution can be easily computed using convex solvers. If we ensure that $0.95 \in [L, H]$, then 0.95 satisfies the constraints. Then setting $p_i = 0.95$ for all i would result in an objective value of zero. Maximization of the objective can only increase the value, which means that the optimal value should be non-negative. We denote the set of percentile values that solve the optimization problem (2)–(4) by $\{\hat{p}_i\}$, and refer to them as the *optimal weighted percentiles* (OWPs). In the next section we calculate the OWPs for customer networks using multiple data traces, and compare the values with the equivalent Shapley value percentiles (SVPs), in order to gauge the fairness achieved by this method.

4. DATA ANALYSIS

We compare the fairness achieved by SVP versus OWP, using data sets of traffic seen by real transit providers. Our first data set (the "SWITCH" data set) is from SWITCH, a European transit provider that serves educational institutions and some commercial organizations. The second data set (the "IXP" data set) is parsed from MRTG graphs published by three European Internet exchange points (IXPs): SIX, BIX and ILAN. From both data sets we extract traffic rates of each customer network at 5-minute intervals.

Our comparison of SVP and OWP proceeds as follows. For each customer i, we first calculate the 95^{th} percentile billing volume $V_i(0.95)$, and use a billing function $\mathcal{B}(x) = 50x^{0.7}$ to translate these volumes into dollar charges. This form of the billing function is based on real-world transit prices [2] and has been used in prior work [7, 3]. We refer to the sum total revenue obtained over all customers as R_{95}, and use it as the minimum target revenue that that both the SVP and OWP schemes should assure to the transit provider. We then compare which customers are targeted for higher/lower percentile billing in each method to check if both methods are aligned in their conception of fairness.

To find the Shapley value percentiles (SVPs) corresponding to the above revenue target, we use the same formulation as Section 2.1. Since calculating the Shapley value is computationally intensive, we used a Monte Carlo approximation [6] with 10000 iterations. Here, the idea is to pick random subsets of customers in Shapley value evaluation equation, and average the value over such subsets. Then, for each customer i, we set $S_i = \sigma_i R_{95}$, and determine the set of SVPs $\{p_i^S\}$ using $p_i^S = \mathcal{F}_i(S_i)$. Note, that for accurate results even this process is computationally expensive.

To find the optimal weighted percentiles (OWPs), we limit the allowable percentiles to 3 units above and below 95%, that is $L = 92\%$ and $H = 98\%$ in (3). We set $\alpha = 5$ when selecting the weights, and a smoothing parameter $\gamma = 200$. We used the Levenberg Marquardt algorithm [8] for approximating the inverse CDF function with a concave function, and found that the normalized least squares errors are less than 10^{-2}. The result of our optimization is a set of percentiles $\{\hat{p}_i\}$. Note that the complexity of these calculations is small as compared to determining the SVP.

We computed the SVPs and OWPs for four years of SWITCH data and 3 months of IXP data. When we plotted the distribution of these percentiles, the support of SVPs varied widely. For example the support of SVPs for February 2012 SWITCH data is $[0.83, 0.99]$. However, by design, the support of all OWP distributions is $[0.92, 0.98]$. Also, as desired, the OWP scheme reduced the billing percentiles of many customers, while increasing that of only a few. Since our conception of fairness is that of the Shapley value, we consider the OWP method fair if the *same* customers are targeted for high/low percentile billing as in the SVP method. We now show this kind of order preservation is largely maintained between SVP and OWP. We first visualize percentile information for a month in the SWITCH data set in Figure 1. We place the individual customers in increasing order of the reciprocal of their Provision Ratios on the x-axis, and their average traffic on the y-axis. For example, a customer with a large reciprocal of Provision Ratio (*i.e.,* it occupies off-peak periods) and small average traffic would appear in the bottom right of the plot. Each circle or square represents a customer network, while the size and intensity of fill color is proportional to the relative percentile used to bill them.

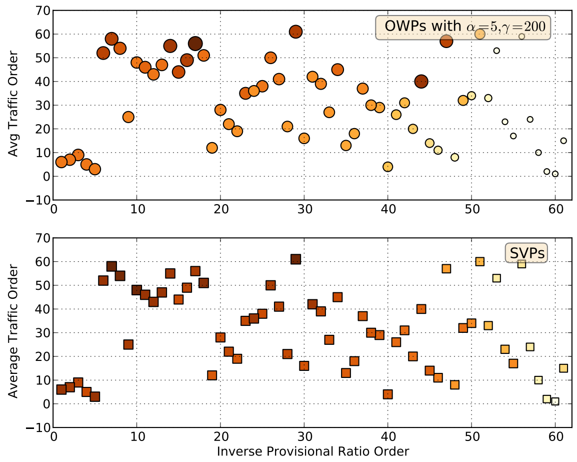

Figure 1: Inverse Provision Ratio order vs Average traffic order of SVPs and OWPs for February 2012 SWITCH data

We observe that in the SVP scheme, there is a gradual increase in billing percentiles from the bottom right to the top left (with a few exceptions). The same trend is observed in OWP. Although the actual percentiles used to bill are different, we see that by-and-large the same customers are targeted in both schemes.

While the visualization indicates the validity of the OWP scheme in preserving fairness, we would prefer to use numerical metrics. We define two such metrics, and show that SVP and OWP are well aligned on both metrics. Our first metric is that of order preservation. We say that order is preserved between two customers i and j if $p_i^s > p_j^s$ implies that $\hat{p}_i > \hat{p}_j$. We compute the percentage of orders preserved in each month of our data sets. For the SWITCH data set, this gives 48 samples over four years from 2009 to 2012. We plot the distribution of percentage of orders preserved in Figure 2. Here, the x-axis is the percentage of orders preserved in that sample, while the y-axis is the number of samples that had that value. We see that all values are above 70% and many values are around 80%, indicating strong order preservation between SVP and OWP. We observed similar results for IXP data sets: the percentage of orders preserved is above 78%. The second metric that we consider is the difference in ranks of billing percentiles. Consider the two sets of billing percentiles $\{p_i^s\}$ and $\{\hat{p}_i\}$, corresponding to SVP and OWP methods, respectively. We can arrange the percentile values in ascending order in each set. Let r_i^s and $\hat{r}_i$ refer

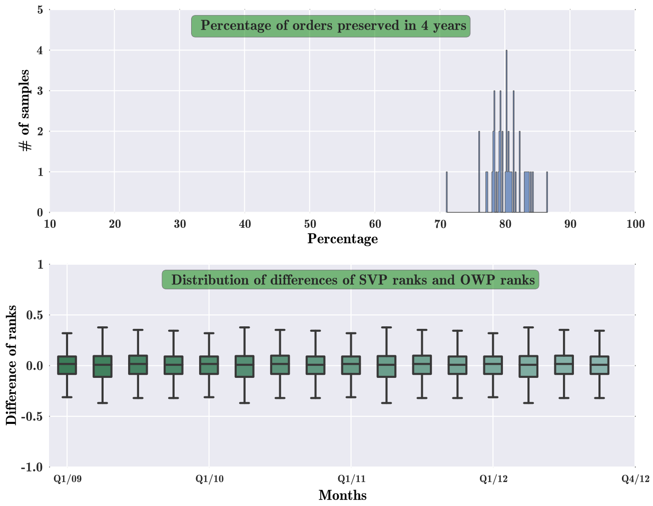

Figure 2: Distribution of percentage of orders preserved and Box and Whisker plot of difference in ranks for four years.

to the order in which $\{p_i^s\}$ and $\{\hat{p}_i\}$, respectively appear in the ordered sets. We call r_i^s and $\hat{r}_i$ as the *ranks* of customer i according to the two schemes, and consider the normalized rank difference $(r_i^s - \hat{r}_i)/N$. The difference must lie in $[-1, 1]$, and a large difference would mean that the ranks are very different, while a small one indicates that they are close to each other. We group the data into three-month intervals (quarters) and present a box-whisker plot of the distributions of the normalized absolute differences over each quarter. Here, the bottom and top of each "box" represents 1^{st} and 3^{rd} quartiles of the distribution for that quarter (*i.e.,* 50% of the samples are contained in both boxes together), while the bottom and top "whiskers" are equal to 1.5 times the 1^{st} and 3^{rd} quartiles of the distribution. We observe that the distributions tightly concentrate around 0, indicating strong preservation of ranks between the SVP and OWP schemes.

5. DISCUSSION AND FUTURE WORK

In this work, we have presented a billing scheme in which an ISP can efficiently compute the billing volume on a per-customer basis. Our formulation is based on a new metric called the Provision Ratio which captures a customer's contribution to the provider's peak traffic load. By using the Provision Ratio as a weight factor in the optimization scheme, the ISP is able to assign lower percentiles to users that have a low contribution to the provider's peak periods. This scheme achieves a notion of fairness similar to the Shapley value, is efficient, and can provide rebates to customers that contribute less to the ISP's peak traffic.

A scheme that charges customers at different percentiles (possibly less than the standard 95^{th} percentile) raises the question: what is the incentive for a provider to charge a customer a smaller percentile? We believe that ISPs do have the incentive to offer lower billing percentiles to customers that contribute minimally to the ISP's peak. The ISP can identify *favorable traffic profiles*, and offer discounts to potential customers that have such traffic profiles. The ISP can use the fairness of our proposed billing mechanism to attract new customers who would be charged higher billing volumes by other providers.

A second question is about feasibility. Our proposed scheme can be expressed as a convex optimization problem, which can be solved efficiently. ISPs typically already collect traffic counts for each customer for each 5-minute period, so computing peak slots and the contribution of each customer is straightforward. A further issue with variable percentile billing is that a potential customer needs to know what billing percentile an ISP would charge it. This issue can be addressed as follows: A customer network shares its traffic profile with the provider. The provider can then add the potential customer's traffic profile to its existing set of customers, and re-run the optimization to determine the billing percentile for the new customer. While it is possible that a customer could infer the provider's peak slots and approximate traffic volume by repeatedly making this query, we believe that information about peak and off-peak periods is not extremely sensitive information. Several providers already make MRTG traffic graphs available online. By making some aspects of its traffic profile available publicly, the provider can also provide transparency into the percentile computation and demonstrate that it is not cheating.

A third question is whether such a scheme could cause oscillations as previously off-peak slots become peak slots in the future? First, since transit customers are generally not end-users, they have less elastic traffic profiles. Second, we conjecture that even in the presence of elastic customer traffic, our scheme will not lead to oscillations. As a customer is rewarded for moving traffic to off peak slots, a likely result is that the provider's overall traffic profile becomes smoother rather than peaking at a different time. We defer a study of the stability of this scheme to future work.

We have assumed that the total traffic load for a provider is the sum of the traffic from individual customers. In practice, however, traffic from all customers does not flow over the same infrastructure.Accounting for this requires knowledge of the provider's internal topology and a detailed cost model that determines total cost based on the traffic load on different portions of the topology. Incorporating topology information, a more realistic ISP cost model such as [9], and evaluating our scheme under more realistic settings are directions we plan to pursue in future work.

6. REFERENCES

[1] V. R. Raja, A. Dhamdhere, A. Scicchitano, S. Shakkottai, k. claffy, and S. Leinen, "Volume-Based Transit Pricing: Is 95 the Right Percentile?" in *Proceedings of PAM*, 2014.

[2] W. B. Norton, "Internet Transit Prices - Historical and Projected," *http://drpeering.net/AskDrPeering/blog*.

[3] A. Dhamdhere and C. Dovrolis, "The Internet is Flat: Modeling the Transition from a Transit Hierarchy to a Peering Mesh," in *Proceedings of CoNEXT*, 2010.

[4] A. Odlyzko, "Internet Pricing and the History of Communications," *COMPUTER NETWORKS*, vol. 36, pp. 493–517, 2001.

[5] X. Dimitropoulos, P. Hurley, A. Kind, and M. Stoecklin, "On the 95-Percentile Billing Method," in *Proc. of PAM*, 2009.

[6] R. Stanojevic, N. Laoutaris, and P. Rodriguez, "On Economic Heavy Hitters: Shapley Value Analysis of 95th-percentile Pricing," in *Proc. of ACM SIGCOMM IMC*, 2010.

[7] H. Chang, S. Jamin, and W. Willinger, "To peer or not to peer: Modeling the evolution of the internet's AS-level topology," in *Proc. IEEE INFOCOM*, 2006.

[8] K. Levenberg, "A method for the solution of certain problems in least squares," *Quarterly of applied mathematics*, vol. 2, pp. 164–168, 1944.

[9] M.Motiwala, A. Dhamdhere, N. Feamster, and A. Lakhina, "Towards a Cost Model for Network Traffic," *ACM SIGCOMM*, Jan. 2012.