technical report

TR-IMDEA-Networks-2012-3

On weather and Internet traffic demand

Juan Camilo Cardona

Rade Stanojević

Rubén Cuevas

September 2012

On weather and Internet traffic demand - Technical Report

Juan Camilo Cardona Restrepo
Institute IMDEA Networks
Madrid, Spain

Rade Stanojevic
Telefonica Research
Barcelona, Spain

Ruben Cuevas
UC3M
Madrid, Spain

ABSTRACT

The weather is known to have a major impact on demand of utilities such as electricity or gas. Given that the Internet usage is strongly tied with human activity, one could guess the existence of similar correlation between its traffic demand and the weather conditions. In this paper, empirical in nature, we demonstrate and quantify such correlation between weather conditions and the Internet traffic demand on different time-scales (from hourly to yearly). For that purpose we collect and use the data from 8 Internet eXchange Points (IXP), geographically spread on 5 different continents, as indicators of the Internet demand in those particular areas. We observe that the seasonal traffic demand variability exists in the locations with large yearly variations in temperature, while the traffic demand in locations close to the equator (with low variability of temperature) is season independent. Using a fine-grain dataset, from three European IXPs, we show that precipitation increases the traffic demand for up to 6%, and somewhat surprisingly that in regards to the impact of precipitation on the demand all major types of ISPs (mobile, residential, content, etc.) observe very similar behavior. One of the implications of the observed time-of-the-day dependent impact of the precipitation is that precipitation has a mild impact on the IP transit costs. Finally, we hint on the possible benefits of the seasonal variations on the energy-proportional computing and scheduling large-scale software releases.

1. INTRODUCTION

Analysis, modeling and forecasting of the Internet traffic is a well studied topic with a number of applications in domains from traffic engineering [24, 29] and network planning [29] to energy efficient routing [8, 31] and server provisioning [11]. Such studies of the Internet traffic dynamics have used traditional statistical tools to capture the dominant characteristics of the dynamics, without explicit modeling of the dependence with the external factors such as weather conditions or social events that are typically accounted as noise [11, 28, 29]. While it has been known that these external factors have significant impact on the demand intensity of some utilities such as gas or electricity, their relationship with the Internet traffic demand is less obvious and not well understood.

In this paper we study the relationship between the Internet traffic demand[1] and one of the factors that plays a significant role in traffic variability: weather. Our approach, essentially empirical, aims to understand the relationship between the traffic demand and the weather both qualitatively and quantitatively. For that purpose we collect and use the traffic data from 8 vantage points, Internet eXchange Points (IXPs) located in 5 different continents, in diverse climate zones, aggregating the traffic of hundreds of ISPs operating in the corresponding regions. To collect the weather data we queried the public database `wunderground.com` for the same locations.

The interaction between the weather conditions and the traffic demand happens on several timescales. Short term weather events, like thunderstorm, snow or heavy wind, impact the accessibility/importance of the Internet access, and have direct, short-term, effect on the traffic demand. Longer term effects reflected through seasonal changes in temperature and daylight duration, have slower and not so direct and immediate influence on the way the Internet traffic is generated. Here we study both the short-term and long-term effects.

An example of a short-time effect is rainfall or other form of precipitation, typically lasting less than a few hours. We find that the precipitation periods have a tendency of increased traffic demand. However, this tendency is dependent on the time of the day and is most noticeable in the late afternoon, when precipitation tends to increase the traffic demand for around 6%.

The periodicity due to end-user temporal cycles is a widely known property of the Internet traffic demand. While the daily and weekly periodicity have been studied extensively [11, 29, 28], and applied in various domains from bulk-traffic transfer scheduling [24] to energy management [11], the seasonal variability over 12-month periods is not well understood, partly because of the fact that the Internet traffic demand has been

[1]In the rest of the paper we will use terms *traffic demand*,*demand* or *load* interchangeably.

dominated by the exponential growth on top of which seasonal changes may be hard to observe and quantify. The six vantage points, geographically spread across the globe, allow us to study the season-dependent traffic variability in various climates. Our data suggests that seasonal traffic variability is strongly tied with temperature variability over the year: the regions far from the equator exhibiting strong seasonal traffic variation while the regions close to the equator show no such seasonal traffic changes.

1.1 Background

Our work is partially motivated by the decades-long research on the demand forecasting in electricity distribution/production systems [16]. Both the short- and long-term forecasting is important for the electricity market: the short-term fluctuations are directly responsible for price variations (in de-regulated markets) while long-term evolution drives the investments in production plants and long-distance transmission links. While the Internet is in many ways a different kind of utility compared to the electricity, the two bear many similarities too. For example: the pricing of the wholesale Internet (IP transit) and the electricity depends on the peak usage, though pricing methods are different: implicit 95th-percentile pricing for the IP transit and explicit variable unit (MWh) pricing of the electricity. Creating new capacity in the Internet (by lying sub-ocean cables) and the electricity distribution systems (building long-distance transmission lines and power plants) are large investments that require accurate understanding of the load dynamics.

The goal of this paper is to empirically analyze the relationship between the weather and what appears to be noise, i.e. how variations in the Internet traffic demand are affected by the weather conditions. Accurate estimation and forecast of the Internet traffic demand has crucial impact on a number of recently designed systems [11, 24]. Even though our work does not directly deal with the traffic forecasting, it suggests that weather conditions should be taken into account when designing networked systems whose performance depend on the traffic demand.

1.2 Contributions

The main contributions of this paper are:

- We collect a number of traffic volume datasets (fine-grain and coarse), spanning 6 different regions as well as the corresponding weather data.

- Studying three IXPs with fine-grain traffic information, we demonstrate that the periods with precipitation in average receive higher traffic demand, and we quantify this relationship in appropriate metrics.

- We show that long-term seasonality of the traffic is not an universal property of the Internet. While regions in Europe and North America do exhibit significant seasonal traffic variations, regions close to the equator show no sign of such seasonal variations.

- We reflect on possible implications of the observed phenomena on IP transit costs, energy-proportional computing and large remote backups/software releases.

While the data we use is public, collection and transformation to a usable format was not straightforward. The traffic data is typically reported via visual images, and we use a custom-made optical function/character recognition tool for converting the visual images into the numeric traffic data.

Even though, qualitatively, one may have guessed that the precipitation may have an immediate impact on the Internet traffic, the quantitative effects of that relationship are less obvious. We show that the relative precipitation-related increment of traffic is time-of-the-day dependent, peaking in the mid-afternoon and not in the late evening when the aggregate traffic peaks. Our per-ISP analysis shows that the effects of the precipitation on the individual ISPs are quantitatively similar.

Choosing six representative vantage points, we show a strong evidence that regions with large yearly temperature variation do exhibit seasonal variation, while those with low yearly temperature variability do not. We use the Pearson correlation coefficient between the traffic and the temperature to quantify the correlation.

The IP transit is typically priced based on the peak-hour traffic demand, using the 95th-percentile method [14, 36]. While some periods of the day do see relatively large precipitation-related increment on traffic, surprisingly the peak-hour periods are only mildly affected by the precipitation, implying relatively low impact of precipitation on the ISP transit costs. The other two aspects mentioned above: (1) energy-proportional computing and (2) large remote backups/software releases, can benefit from the seasonal variations in that they can exploit: (1) cheap cooling when the demand is high and (2) yearly summer valleys, respectively. See Section 5 for more details.

2. DATASETS DESCRIPTION

To understand the interaction between the traffic demand and weather conditions we collect and use a number of datasets described below.

2.1 Traffic data

As indicator of the Internet traffic demand in particular areas we use the traffic information publicly available at several Internet eXchange Points (IXP). IXPs

IXP name	city	duration (months)	granularity	peak traffic	# of ISPs
AMS-IX	Amsterdam	16	$24h$	$1.4Tbps$	463
TORIX	Toronto	13	$24h$	$70Gbps$	144
PTT.br	Sao Paolo	13	$24h$	$60Gbps$	243
WAIX	Perth	13	$24h$	$2Gbps$	64
NIXI	Mumbai	13	$24h$	$8Gbps$	31
SIX	Bratislava	24	$24h$	$40Gbps$	52
SIX	Bratislava	18	$5min$	$54Gbps$	52
FICIX	Helsinki	18	$5min$	$35Gbps$	28
INEX	Dublin	8	$5min$	$31Gbps$	56

Table 1: The traffic datasets and details of corresponding IXPs.

facilitate convenient and inexpensive peering between ISPs operating in a region or country [2]. It is shown in [1], that for several ISPs the IXP traffic is a good proxy of their non-IXP traffic and thus the traffic trends observed in the IXP traffic can be seen in the aggregate (IXP and non-IXP) traffic. Therefore, in this paper we use IXP traffic to analyze the traffic trends (driven mainly by human activity) in a particular area.

Traffic demand information of a particular ISP is often considered confidential, and is rarely shared externally. However, many IXPs for marketing purposes reserve rights to report publicly the information regarding the traffic passing the IXP infrastructure. Most of the IXPs publish only the traffic data aggregated across all ISPs and all ports [15], while some also report per-ISP (or per-port) traffic demand time series.

The traffic data is rarely provided in an easily usable numeric format[2]. The de facto standard in reporting such traffic data is through png images: output of mrtg/rrdtool [27] for graphic representation of raw traffic data. Examples of such images are shown in Figure 1. Extracting the numeric values from such mrtg images is a rather time-consuming task. In order to automatize the process we used a piece of code that reads the png image pixel-by-pixel and outputs the relevant traffic data in a numeric format. While such optical character/function recognition was not intellectually challenging, it required serious efforts to capture many small details and resolve various issues due to different png output formats.

In order to properly understand the short-term weather impact on the Internet traffic demand we have to focus on a limited geographical area in which we can guarantee constant weather conditions for the whole area. Furthermore, we need fine-grain data for both traffic demand and weather reports. The data from three IXPs (Slovak-IX, FICIX and INEX) dataset fulfills these requirements as we will see in Section 2.1.2.

2.1.1 Long-term traffic trends

The traffic demand trends over long time intervals (12 months or more) may reveal dynamics in traffic demand time series, that are correlated with seasonal changes. The mrtg output typically outputs the traffic demand data on different granularity (5-minute to 24-hours slots) covering the dynamics of the previous day (reporting 5-min averages), week (30-min), month (2-hour) and year (24-hour). The example of daily and yearly graphs are shown in Figure 1.

In this paper we use the traffic data from several IXPs: the largest European (AMS-IX), Northern American (TORIX), Southern American (PTT.br), Australian (WAIX), and Indian (NIXI) IXP that report their traffic information publicly. As a remark, AMS-IX, PTT.br and NIXI are indeed the largest (in terms of traffic and membership size) European, South-American and Indian IXPs respectively. While we suspect that there are some IXPs in North America and Australia that carry more traffic than TORIX and WAIX, they do not publish their traffic data and we go with the largest available IXPs. In addition to those 5 IXPs we also utilize one medium sized IXP from central Europe, Slovak-IX (SIX).

The data reports daily average of the traffic exchanged in the IXP for the period of 13 to 18 months, depending on the local configuration of the mrtg/rrdtool. Table 1 contains the details about the dataset (duration, granularity) and the IXPs: the peak traffic, the number of ISPs participating at the IXP and the annual growth rate (AGR)[3].

2.1.2 Short-term traffic trends

To investigate the immediate effects of precipitation on the traffic usage, the traffic averaged over a day is too crude to provide valuable information. As we mentioned above, mrtg/rrdtool reports the traffic data of finer granularity, but such data spans small time periods, e.g. 5-min averages are reported for the previous one day or so. However, since March 2011 we have been collecting and storing the whole snapshot of the Slovak-IX and the Finish IX (FICIX) website, once per day. Thus, we have the coverage of 18 months, with the 5-min granularity. Additionally, we have also stored data for INEX, the largest Irish IX, since January of 2012.

[2]Of 8 IXPs analyzed here, only AMS-IX published raw numeric data, albeit only on monthly granularity.

[3]AGR is defined as the ratio between the traffic in the same month(s) in 2011 and 2010; see Eq (1).

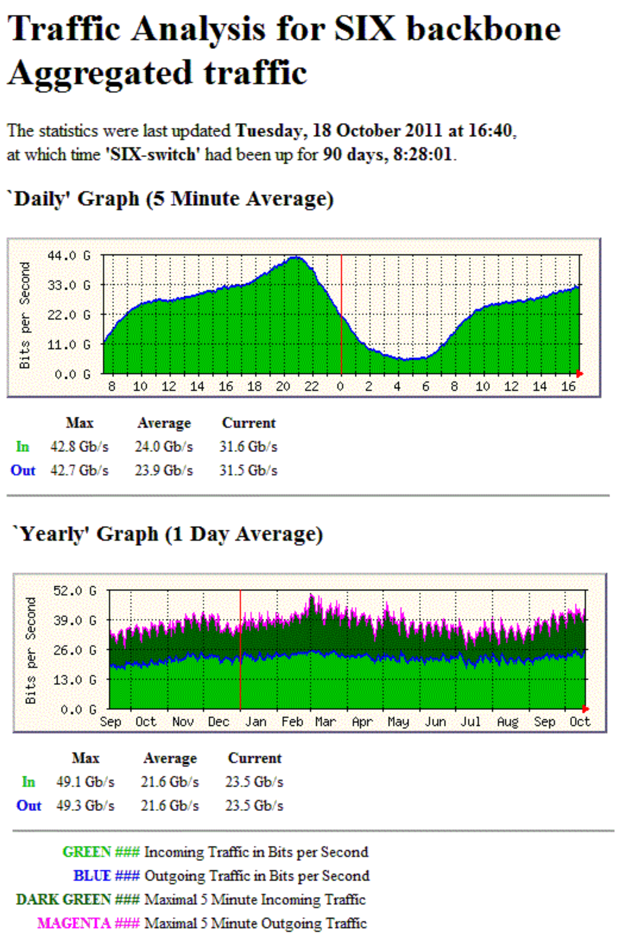

Figure 1: mrtg output for aggregated SIX traffic.

In addition to the mrtg data on the aggregate traffic (the sum across all ports), SIX and FICIX also reports per member mrtg statistics, that we store too every day for the same period. Since mrtg produces the time-stamp at the time images are created synchronization of the data was straightforward except for the several days just after the daylight saving time change which we synchronized manually.

2.1.3 Limitations of the data

We conclude this section with a brief discussion of the limitation of the traffic data. Namely, the traffic data used in this paper is local in nature: it is the traffic that is exchanged between ISPs operating in the region of the IXP and that peer between each other at the IXP. There are several factors that may impact the qualitative and quantitative results including the following.

For a given ISP, both the traffic exchanged locally via the IXP and the rest of the traffic is predominately generated by the human activity, and intuitively these two quantities are expected to have proportional dynamics. This proportionality of the IXP and non-IXP traffic dynamics is shown empirically in [1] using detailed traffic data from 4 ISPs [7, 18, 20, 33]. In both SIX and FICIX, more than 3/4 of the traffic is generated by the local/national ISPs, implying very high locality of the traffic. In the case of INEX, although we do not possess the traffic for each of its members, more than 67% of its members are local (including the 5 largest Internet Providers of the country)[21, 13], which hints a similar case.

Weather conditions, such as rain or clear sky, are location dependent and vary across the country. It is necessary, before our analysis, to assign the Internet traffic of each IXP to specific geographical points. However, it is unfeasible to estimate the distribution of traffic of each exchange point among the cities of each region, since IXP's members might peer in different locations outside the IXP. We choose then to analyze how each IXP's traffic is impacted by the weather over the city where each IXP is located: Helsinki[4], Bratislava and Dublin. These cities host a large percentage of the broadband connections of each country[26, 17, 6]. Additionally, the presence of the IXP in the city makes it ideal for the companies peering on it to exchange the local sourced and destined content through it. The above reasons justify that our selected points are responsible for enough of the traffic of the IXP to give us insights on the correlations between the traffic and the weather of the city. We remark though that with a fine-grain data on the traffic from narrower regions, one would get a more accurate picture on the relationship between the weather and traffic demand.

[4]FICIX's two main locations are situated inside Helsinki Metropolitan Area.

The other factor that may impact the representativeness of the data is the peering dynamics of the ISPs that participate at the IXP. Namely, if a pair of ISPs stop/start exchanging traffic via the IXP, that creates a disturbance in the aggregate IXP traffic dynamics. However, the IXPs analyzed here are mature organizations, and the membership and the peering agreements are rather stable. Consequently we do not observe any large sudden changes in the traffic levels that could be expected in case of significant peering removal/creation.

Duration of the data monitoring affects the confidence in observed quantitative results. We currently collect and store the daily traffic data from dozens of IXPs spread across the globe which will be used to build more accurate quantitative models of the weather-traffic relationship. Additionally we are planning to release the collected data for public use.

2.2 Weather data

There are several online sources of free historical weather datasets. Since they report quantities that are easily measurable they are very similar between each other. Some inconsistencies may exist between different datasets, but they are typically very small (e.g. the measurement point at the city center and the airport are likely to have slightly different weather conditions.). In this paper we use the weather data provided by the Weather Underground, easily accessible database available at `http://www.wunderground.com/`.

We queried the database for the cities that host the IXPs listed in Table 1, for the period that covers our traffic data. The arguments for a single query are the `location` and `day`. The output is a table that specifies the weather conditions in the given `location` and on the given `day` with the granularity of 30 minutes; i.e. every 30 minutes weather parameters are reported. The `wunderground.com` publishes a number of weather parameters including temperature, dew point, humidity, pressure, precipitation. Granularity of 30 minutes allows fine analysis of the relationship between the weather and traffic, though some very short events such as hail or short storms, may be missed in the 30-minute sampling.

For the purpose of this paper we mainly use two weather parameters: temperature and precipitation. While the other parameters may be correlated with the Internet traffic demand, these two affect the Internet traffic in the most direct way. Understanding the (second order) relationships between the traffic demand and the other weather parameters is out of scope of the present paper.

3. TRAFFIC VS. PRECIPITATION

We start the analysis by examining the dependance between the traffic demand and the precipitation. For that purpose we use the datasets dataset described in Section 2.1.2 that contains the fine-grain traffic statistics (both inbound and outbound) exchanged at three IXP. Such data allows us to notice changes that happen on short-time scales (say hourly) and compare them against the weather conditions. As we explain above, we focus on precipitation as the parameter that is most likely to affect the instantaneous human behavior towards the Internet usage.

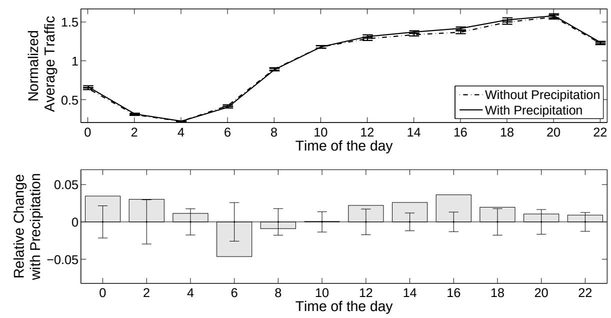

Figure 4: Impact of weather of Kosice in SIX's traffic.

For that purpose we split the time into 2-hour timeslots. For the ISP A, $u_A(t)$ and $d_A(t)$ denote the average upstream and downstream traffic (in $Mbps$) at time slot t. With

$$u(t) = d(t) = \sum_{\text{for all members } A \text{ of SIX}} u_A(t)$$

we will denote the aggregate traffic exchanged at the IXP[5]. The time-series $u(t)$ is depicted in Figure 2. Finally, in order to smooth-out the seasonal effects (e.g. one can notice from Figure 2 that the March traffic is higher than the July traffic; see Section 4 for more details) we normalize $u(t)$ with the average traffic over a two week period centered at t:

$$\bar{u}(t) = \frac{u(t)}{average(u(t-84), \ldots, u(t+84))}.$$

For each time-slot t there are 4 weather reports (sometimes more than 4, depending on reporting system configuration) in our weather dataset, and we set a binary variable $wet_\alpha(t)$ to be 1 if the fraction of the weather reports from time slot t that report precipitation (snow, shower, rain, storm...) is not smaller than α, otherwise $wet_\alpha(t) = 0$, where $\alpha \in (0, 1]$ is a parameter. In other words if the fraction of the time it precipitates during the time slot is greater than or equal α, we declare that the time slot to be wet, otherwise we declare it non-wet. We use $\alpha = 0.15$ unless we specify a different value, thus

[5]Since all the traffic that is sent by one member-ISP is received by another member-ISP, the aggregate upstream is identical to the aggregate downstream traffic volume at any IXP.

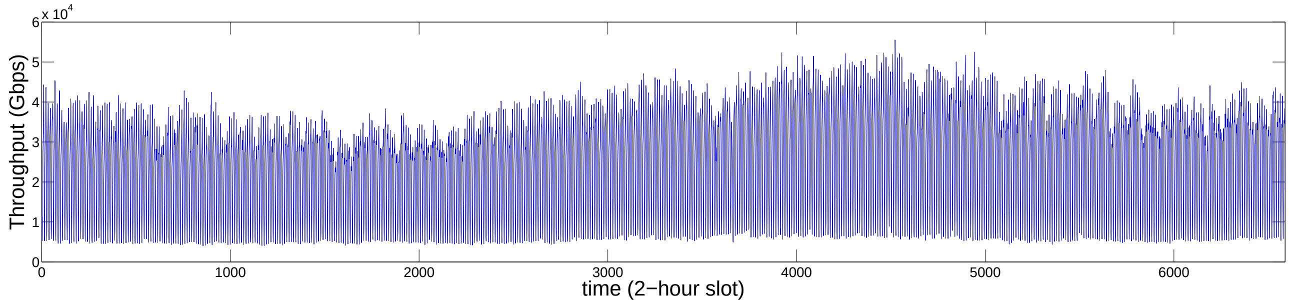

Figure 2: SIX aggregate traffic over a 18-month period (Mar/2011-Oct/2012).

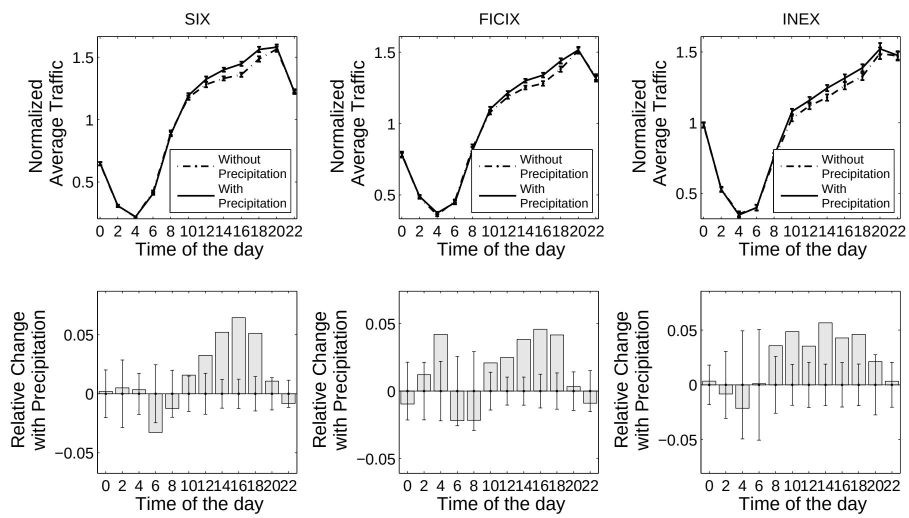

Figure 3: Normalized daily SIX demand with and without precipitation.

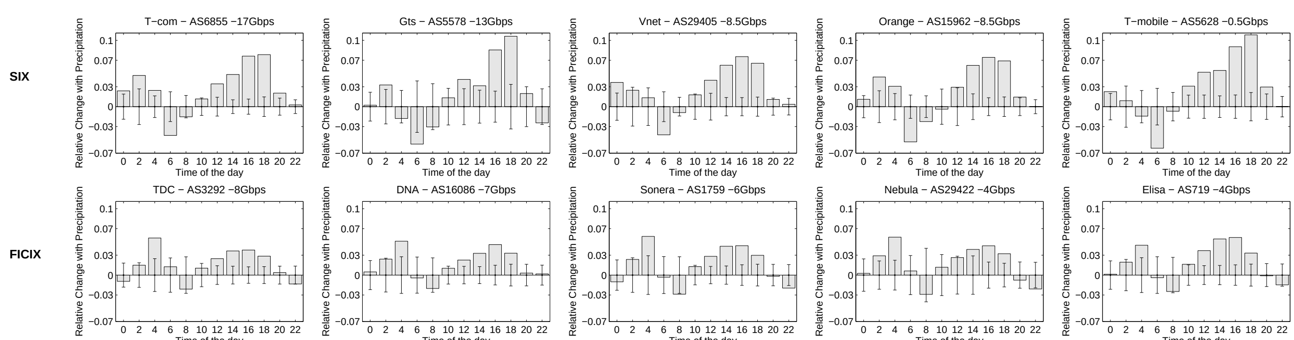

Figure 5: Relative change of demand with precipitation.

announcing a slot 'wet' if it receives precipitation for at least 15% of the time. Furthermore, we also declare a slot to be wet if any of the two preceding time slots presented an alpha of at least 15%.

Our goal is to observe whether time series $wet_\alpha(t)$ and $\bar{u}(t)$ are correlated. To that end, we split the day in twelve 2-hour interval, and calculate average normalized traffic with and without precipitation for each of the twelve intervals:

$$A(i) = \frac{\sum_{\bmod (s,12)=i} \bar{u}(s) wet_\alpha(s)}{\sum_{\bmod (s,12)=i} wet_\alpha(s)} \quad i = 0..11$$

$$B(i) = \frac{\sum_{\bmod (s,12)=i} \bar{u}(s)(1 - wet_\alpha(s))}{\sum_{\bmod (s,12)=i} (1 - wet_\alpha(s))} \quad i = 0..11$$

thus for the twelve time intervals $0h-2h, 2h-4h, \ldots, 22h-24h$, $A(i)$ and $B(i)$ represent the average normalized load in the interval $[2ih, (2i+2)h]$ with and without precipitation, respectively.

In Figure 3(top) we depict the values of $A(i)$ and $B(i)$ together with the relative difference between $B(i)$ and $A(i)$ for the three IXPs. To determine if the difference between $A(i)$ and $B(i)$ is statistically significant to claim that the means of the samples with and without precipitation are different, we use Welch's t-test [37], which is well-suited for this case as the number of samples for each random variable is different and relatively large[6]. Figure 3(bottom) also include the interval outside of which Welch's t-test rejects the null-hypothesis for a significance level of 0.05. Thus from early afternoon to early evening, with 95% of confidence we can affirm for all IXPs that the mean normalized traffic is larger in timeslots with precipitation than in timeslots without precipitation. For the other periods of the day (except for a couple of time period in the early morning), the difference between the means is not statistically significant to support that precipitation impacts the traffic. **Remark**: While the above properties may be explained by some intuitive characteristics of the human daily cycles, we do not attempt to model human behavior in this paper, and leave it for future work.

As explained in Section 2.1.3, our analysis is based in the impact of precipitation of the city where each IXP is located. Although traffic from other cities of the same region can also be exchanged in the same IXP, it will be extremely hard to estimate the portion of traffic that these cities represent. In Figure 4 we depict the effect of precipitation of the city of Kosice (second largest city of Slovakia) in the traffic of the Slovak-IX. The figure shows that precipitation has a considerable lower impact in traffic than the ones already described. Nonetheless, this result is not discouraging. It only means that we require a better understanding of the footprint of a specific IXP to perform a better analysis of the impact of traffic in the other cities of the region.

[6] In all cases the number of samples obtained is larger than 40.

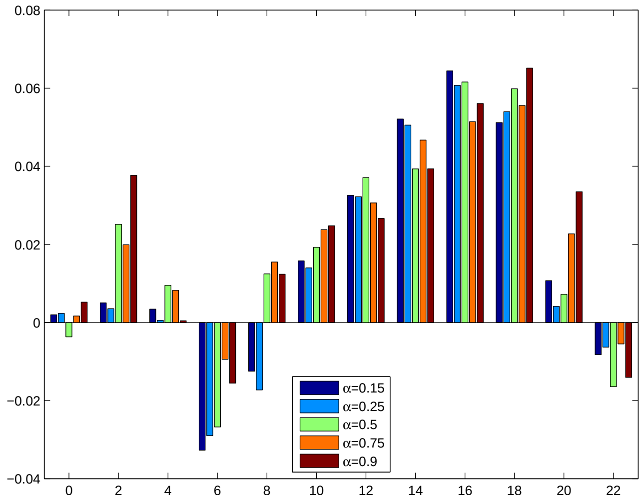

Figure 6: The impact of the parameter α.

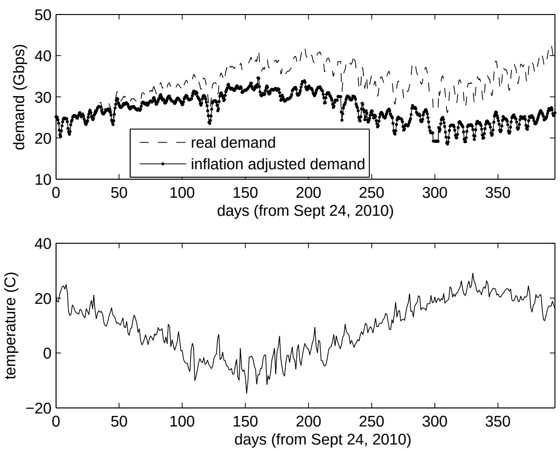

Figure 7: The real and inflation adjusted demand at TORIX (top). Average daily temperature at Toronto (bottom).

3.1 Precipitation effects at different ISPs

In order to understand which ISPs contribute to the observed effect of precipitation on the demand, we consider the 4 largest ISPs (in terms of traffic exchanged), plus the largest mobile (3G) ISP in SIX and FICIX. In Figure 5 we report the relative difference (per 2-hour) between periods with and without precipitation, for these 12 ISPs, and also the Welch's t-test confidence intervals. Even though the daily patterns among the 12 ISPs differ, it is interesting to see that the increasing effect of weather on the traffic between noon and early evening is common for all ISPs and tops around 10%. With the exclusion of the 4am-6am slot in FICIX's ISPs and the 6am-8am in SIX's ISPs, the night and early morning periods suffer no clear effect from precipitation.

3.2 The effect of parameter α

In the previous analysis we used the value of $\alpha = 0.15$, the slot was declared 'wet' if it precipitates for at least 15% of the time. However, the choice of this parameter influences the extent to which the precipitation affects the load. In order to examine the effect of this parameter, we evaluate the relative change with precipitation (as in Figure 3, bottom) for 5 different values of α: $0.15, 0.25, 0.5, 0.75$ and 0.9. The observed values are reported in Figure 6. We can observe that larger α implies more consistent rain which is reflected in the larger change with precipitation, in most of the time slots.

As final remark, our weather dataset does report the daily amount of precipitation (in *mm*) but unfortunately does not report such amount per time-slot, and therefore the problem of quantifying the correlation between the amount of precipitation and the traffic variations remains open.

4. SEASONAL TRENDS

Yearly seasonality of the Internet traffic has been poorly understood phenomenon for a number of reasons. Firstly, the data on the traffic over long-time periods is rarely available. Secondly, for decades since the beginning of the commercial Internet, the traffic dynamics has been dominated by the exponential growth which makes the characterization of the seasonal effects very challenging. And finally, seasonal effects may not be present at all in some geographical regions.

In this section we examine the (yearly) seasonality trends of the Internet traffic demand, across a geographically diverse set of IXPs. The data used here covers 6 regions, from 5 continents and is described in detail in Section 2.1.1. The traffic demand datasets we collected cover slightly over 1 year time span, which is a minimal duration for inferring basic characteristics of the yearly seasonality.

To extract the seasonal effects from the exponential growth of the traffic demand, we utilize the following procedure. For the traffic time series covering a time period of 365+Δ days[7], we calculate the *annual growth rate* (AGR) as the ratio between the traffic in the last Δ and the first Δ days:

$$AGR = \frac{\text{total traffic in days } 366 \text{ to } 365+\Delta}{\text{total traffic in days } 1 \text{ to } \Delta}. \qquad (1)$$

[7]In different datasets, the number of days on top of a full year, Δ, varies from 30 to 180 days, see Table 1.

. Then, we calculate the inflation adjusted demand (IAD) at day t as:

$$IAD(t) = \frac{D(t)}{AGR^{\frac{t}{365}}},$$

where $D(t)$ is the average demand (in $Mbps$) at day t. Intuitively, the inflation adjusted demand, filters out the exponential growth of the Internet demand and allows us to focus on seasonal effects.

To illustrate the inflation adjustment operation we depict both the real and the inflation adjusted demand for the TORIX in Figure 7. We can see that the traffic demand, after inflation adjustment, appears to follow a sin-wave pattern that is out-of-phase with the average daily temperature observed in Toronto.

4.1 Demand vs. temperature

To visualize the correlation between the time series of average daily inflation adjusted demand $IAD(t)$ and the average daily temperature $Y(t)$, we use the scatter graphs in Figure 8. There we depict the scatter graphs for six IXPs listed in Table 1. One can observe that the IXPs located close to the equator (WAIX, PTT.br, NIXI) show no visually obvious dependence between the IAD and the temperature, while the dependence between the IAD and the temperature is visually observable from the Figure 8 for those IXPs that lie in the regions with large temperature variation. In order to quantify the correlation between $IAD(\cdot)$ and $Y(\cdot)$ we also calculate the Pearson correlation coefficient:

$$corr = \frac{\sum_{t=1}^{N}(Y(t)-\bar{Y})(IAD(t)-\overline{IAD})}{\sqrt{\sum_{t=1}^{N}(Y(t)-\bar{Y})^2}\sqrt{\sum_{t=1}^{N}(IAD(t)-\overline{IAD})^2}},$$

where we used the notation $\bar{X}$ for the sample mean of the time series X. The Pearson coefficient is a metric that ranges between [-1,1]. The closer the value is to 0 the less correlated are the variables, whereas positive and negative values closer to 1 and -1 indicate strong positive and negative correlation, respectively.

The Pearson correlation coefficient is reported in the same figure, for all of the six studied IXPs. For the three regions close to the equator, the *corr* value is very close to zero, indicating nonexistence of any significant correlation. For the other three IXPs, the *corr* value is relatively high indicating strong negative correlation between the temperature and the demand.

From Figure 8 we can also observe relatively high variability of the daily traffic averages. While the correlation with the temperature may explain some of this variability, there are still many external factors (eg. social events) with unknown influence on the Internet traffic load.

5. IMPLICATIONS

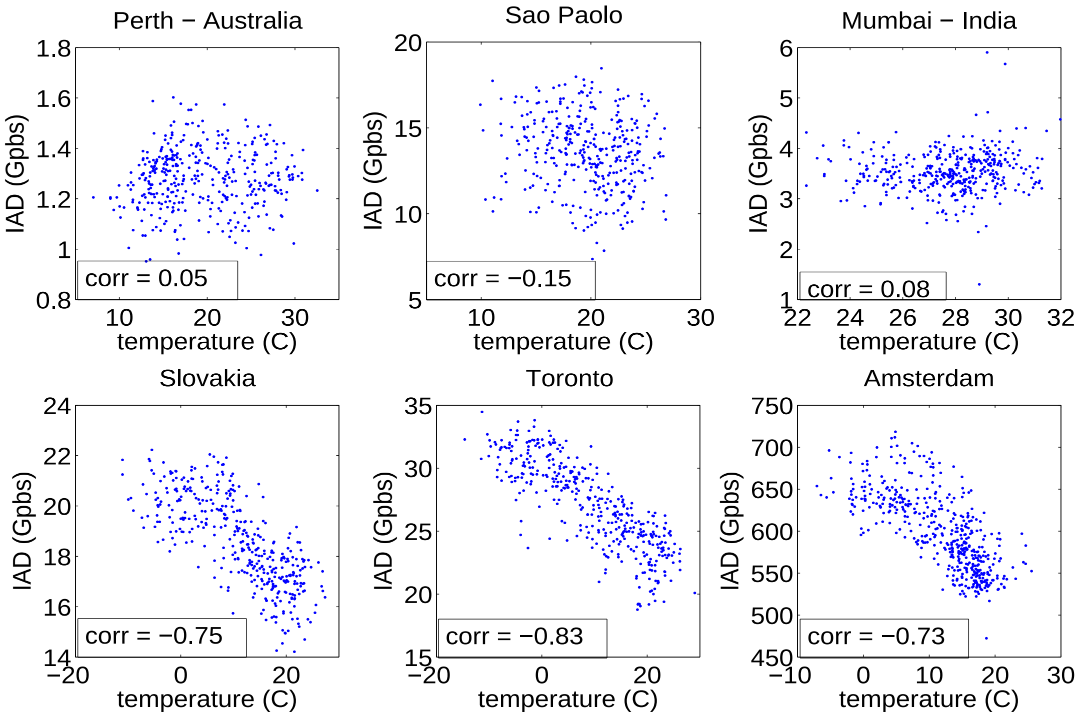

Figure 8: Correlation between the average daily temperature and the inflation adjusted demand (IAD).

5.1 Mild impact on the 95th-percentile

The motivation for a vast amount of literature on accurate estimation of the electricity demand, is mainly economic. The electricity demand (especially at peak hours), has a direct effect on the MWh prices, and consequently it is of huge importance for both the buyers and the sellers to accurately estimate and predict the demand and hence optimize the costs, particularly at peak hours.

In the IP transit market, the pricing is also peak-based, albeit indirect through the 95th-percentile [14, 36]. In contrast with unit-based pricing employed in the electricity markets in which the charge for a consumed demand is calculated on the spot, the 95th-percentile pricing is implicit in the sense that the impact of an instantaneous change in the demand is not possible to estimate until the end of the charging period, which is typically one month long. In this section we report the impact of precipitation on the 95th percentile of the traffic in the short SIX dataset.

For each month in our 7-month SIX dataset, we calculate the 95th-percentile of the traffic time series depicted in Figure 2. Additionally, we approximate the traffic demand time series discounted for the precipitation increment: in each time slot, we subtract an appropriate percentage of the traffic subject to time of the day and the relative increment depicted in Figure 3. For the discounted traffic time series, the reduction in the monthly 95th-percentile is in the range between 0% (in March) to 2.4% (in June). Such mild effect on the 95th-percentile is the consequence of the two factors. Firstly, the duration of the rain periods is relatively low. And secondly, the 95th percentile is mainly impacted by the peak-hour period ($20h-22h$), that surprisingly does not experience large precipitation increment.

5.2 Positive impact on energy

With the trend of increasing energy prices and concerns related to the negative effects of the carbon emissions a large fraction of the research community is focused on the topics related to energy-proportional computing and networking. The goal of the energy-proportionality is designing systems in which the energy consumption is proportional to the load, which is rarely the case in the systems that are in operation today.

We want to stress here, that the seasonal demand variability observed in regions in Europe and North America is beneficial for the energy-proportional systems. Namely, a large fraction of the energy consumed in a data-center is devoted for cooling. As authors of [4] show, appropriately designed cooling systems take advantage of low external temperature and are significantly more efficient in cold weather than in the hot periods. Intuitively, in the energy-proportional systems it is advantageous to have 'cheap' cooling when the demand is high. Having no access to relevant data to quantify the effect of the Internet usage demand on the energy consumption, we do not proceed in assessing the impact of this phenomenon and leave it for future work.

5.3 Large (yearly) backups/software release

Daily backups and software updates are examples of applications that require nontrivial amount of bandwidth with relatively soft performance constraints. In order to minimize their effect on the interactive traffic, such data transfers can be scheduled to happen over night [9]. Similarly, large data transfers that can tolerate delays in the order of magnitude of several months, can be scheduled for the months with lower traffic load, which can have beneficial effects for performance of the interactive traffic and/or the cost of IP transit, depending on the physical capacity of the link from the user to its ISP as well as the pricing contract between them.

One possible example of such application is 'large' backup of a global enterprize. Normal daily/weekly backups may be executed and stored locally. However to secure the data in long term, the system administrator may choose to store accumulated data in a remote data-center, occasionally. Such infrequent, data-intensive, backups can be scheduled over the weeks with low traffic usage, and the seasonal variability offers opportunities for cheap/efficient large transfers.

Another example of such application can be large software release such as Mac OS X Lion, that attracted 1 million downloads in the first 24 hours after the release [25]. With the installation download size of around 4Gbytes, it is equivalent to $370Gbps$ of traffic averaged[8] over $24h$. Very few ISPs are capable of accepting such enormous amount of traffic over a short period of time. Therefore, picking the seasons with low traffic load for such releases, may be beneficial both for the transport ISPs as well as for the end users through reduced likelihood of traffic throttling due to overload. Not surprisingly, the release date for Lion was in the middle of the summer, July 20th, 2011.

6. RELATED WORK

The Internet traffic has been extensively studied by both the practitioners and the theorists over the previous two decades.

Understanding the evolution and statistical properties of the Internet traffic has been subject of several studies. Papagiannaki et al [29], study IP traffic time series from multiple vantage points in a tier 1 ISP, and conclude that such traffic exhibits strong daily and weekly trends and derive a methodology for implicit extraction of long term trends that could be used for predicting time and place of link upgrades/additions in the core of an ISP. Cho et al [10] analyze the traffic data from a set of residential Japanese ISPs to quantify several properties including: per-customer traffic distribution, the impact of the heavy-hitters, correlation between the inbound and outbound traffic usage, etc. Roughan et al. [32], introduces a metric for measuring the variability in the periodic daily traffic patterns. In understanding the periodicity of daily/weekly traffic patterns, the variability has been considered as white noise by previous studies. Here we show that some of this noise can be accounted to weather conditions, and we believe that a large fraction of the remaining 'noise' can be correlated with social or other events identifiable by public information such as Twitter.

Several systems are recently proposed that critically rely on accurate traffic prediction. Energy-aware server provisioning have been suggested in [11] for connection-intensive Internet services such as Windows Live Messenger. The system uses fine-grained load estimator to decide when and which servers to turn off, and it has been shown that the performance of the system strongly depends on the accuracy of the traffic estimator. NetSticher [22] uses off-peak periods to push delay tolerant bulk transfers without increasing IP transit bill and in order to do so must have an accurate forecast of the 95th-percentile and the traffic load.

As we mentioned in Section 1.1, the motivation for our work comes from the electricity markets where load affects daily operation and costs. Chen et al. [12] show that the electricity load can be represented as a sum of four components: periodic trend, the weather and special event related components and the random noise. In this paper we show that the weather indeed does have an immediate impact on the Internet traffic, however impact of the special events and 'pure' noise is an open problem.

Most recently Shulman and Spring [34] observe a strong correlation between the IP network failures and the weather conditions. To the best of our knowledge it is the first study on understanding the impact of weather on the operation of IP networks. Such failures are likely to be the consequence of direct impact of the weather on the ISP infrastructure (such as equipment damage by lightning strikes or degradation of satellite link quality due to high humidity) and are independent of the human behavior. In contrast the phenomenon of correlation between the weather and traffic usage is heavily impacted by the humans.

7. SUMMARY

In this paper we demonstrated a dependance between the Internet traffic demand and the weather, both on short and long time scales, and we hinted on possible implications of this dependance on the cost and opera-

[8] The peak traffic is likely to be in the range twice greater than the average due to the human diurnal cycle.

tion of Internet service providers.

We believe that phenomena observed here opens an avenue for understanding the impact of external factors on the Internet traffic demand. We conclude the paper with a list of exciting questions that remain open.

Open Problem 1. Our initial analysis indicates that other weather parameters (clouds/wind/humidity/daylight etc.) have a relatively small impact on the Internet traffic demand. Exact quantification of these effects is an open problem.

Open Problem 2. Is the impact of precipitation similar in other regions/IXPs?

Open Problem 3. Can observed phenomena be explained by a simple (first principle) model of human behavior?

Open Problem 4. Intuitively, special events, such as sport games or elections, impact the Internet usage pattern. Can we demonstrate and quantify such correlation by analyzing the public news data such as Twitter?

8. REFERENCES

[1] Stanojevic, R. and Castro, I. and Gorinsky, S. CIPT: using tuangou to reduce IP transit costs. CoNext 2011.

[2] B. Augustin, B. Krishnamurthy, W. Willinger. "IXPs: Mapped?". In Proc. of ACM IMC 2009.

[3] Ager, B. and Chatzis, N. and Feldmann, A. and Sarrar, N. and Uhlig, S. and Willinger, W. "Anatomy of a Large European IXP". In Proc. of ACM SIGCOMM 2012.

[4] E. Bonetto, M. Mellia, M. Meo. "Energy Profiling of ISP Points of Presence", under submission 2011.

[5] J. C. Cardona Restrepo, R. Stanojevic. "A history of an internet exchange point". ACM CCR, vol. 42(2) 2012.

[6] Central Statistics Office (Ireland). `http://cso.ie/en/index.html`

[7] CESNET, Traffic Stats. `http://www.ces.net/netreport/`.

[8] J. Chabarek et al. "Power awareness in network design and routing". In Proc. of IEEE Infocom 2009.

[9] A. Chervenak, V. Vellanki, Z. Kurmas. "Protecting file systems: A survey of backup techniques". In Proc. of the Sixth Goddard Conference on Mass Storage Systems and Technologies, March, 1998.

[10] K. Cho, K. Fukuda, H. Esaki, A. Kato. "The Impact and Implications of the Growth in Residential User-to-User Traffic". In Proc. of ACM SIGCOMM 2006.

[11] G. Chen, W. He., J. Liu, S. Nath, L. Rigas, L. Xiao, F. Zhao. "Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services". In Proc. of NSDI 2008.

[12] H. Chen, C. Canizares, A. Singh. "ANN-based short-term load forecasting in electricity markets". IEEE Power Engineering Society Transmission and Distribution Conference, 2001.

[13] Commision for Communications Regulation (Ireland). `http://www.comreg.ie/_fileupload/publications/ComReg1263a.pdf`

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

[15] European Internet Exchange Association, `www.euro-ix.net`.

[16] E. Feinberg, D. Genethliou. "Load forecasting". Applied mathematics for power systems, 2005.

[17] Finnish Communications Regulatory Authorit. `http://www.viestintavirasto.fi/en/index/asiointi-info/tutkimusjatilastot/yleinenmarkkinatieto.html`

[18] GRNET, Traffic Stats. `http://mon.grnet.gr/rg/`.

[19] P. C. Gupta, K. Yamada. "Adaptive short-term load forecasting of hourly loads using weather information", IEEE Transactions on Power Apparatus and Systems, 1972.

[20] HEANET, Traffic Stats. `http://www.hea.net/`.

[21] INEX, Member list and Traffic Stats. `https://www.inex.ie/`.

[22] P. Kennedy. A Guide to Econometrics. MIT Press, Cambridge, MA. USA, 1998.

[23] Roe K. and Vandebosch H. (1996). "Weather to view or not: That is the question". European Journal of Communication, 11(2), 201-216.

[24] N. Laoutaris, M. Sirivianos, X. Yang, P. Rodriguez. "Inter-datacenter bulk transfers with NetStitcher". In Proc. of ACM SIGCOMM, 2011.

[25] Mac OS® X Lion, the first day report. `http://www.apple.com/pr/library/2011/07/21Lion-Downloads-Top-One-Million-in-First-Day.html`.

[26] Ministry of Transport, Construction and Regional Development of the Slovak Republic. Electronic Communications - Statistic Data. `http://www.telecom.gov.sk/files/statistika_vus/telekom_ang/index.htm/`.

[27] Multi Router Traffic Grapher. `http://oss.oetiker.ch/mrtg/`.

[28] D. Niu, Z. Liu, B. Li, S. Zhao. "Demand forecast and performance prediction in peer-assisted on-demand streaming systems". In Proc. of IEEE INFOCOM 2011.

[29] K. Papagiannaki, N. Taft, Z.L. Zhang, C. Diot. "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models". In Proc. of IEEE INFOCOM 2003

[30] S. Rahman, R. Bhatnagar. "An expert system based algorithm for short term load forecast". IEEE Transactions on Power Systems vol. 3(2) 1988.

[31] J.C.C. Restrepo et al. "Energy profile aware routing". In Proc. of IEEE ICC workshops, 2009.

[32] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, Y. Zhang. "Experience in measuring backbone traffic variability: models, metrics, measurements and meaning". In Proc. of Internet Measurement Workshop 2002.

[33] SANET, Traffic Stats. `http://samon.cvt.stuba.sk/`.

[34] A. Schulman, N. Spring. "Pingin' in the Rain". In Proc. of ACM IMC 2011.

[35] R. Stanojevic, I. Castro, S. Gorinsky. "CIPT: using tuangou to reduce IP transit costs". In Proc. of ACM CoNEXT 2011.

[36] R. Stanojevic, N. Laoutaris, P. Rodriguez."On Economic Heavy Hitters: Shapley Value Analysis of the 95th-Percentile Pricing". In Proc. of ACM IMC 2010.

[37] Welch, B. L. "The generalization of Student's problem when several different population variances are involved". Biometrika 34 (1-2): 28-35. doi:10.1093/biomet/34.1-2.28. MR19277. 1947