ITMgen - A First-principles Approach to Generating Synthetic Interdomain Traffic Matrices

Jakub Mikians
UPC BarcelonaTech
jmikians@ac.upc.edu

Nikolaos Laoutaris
Telefonica Research
nikos@tid.es

Amogh Dhamdhere
CAIDA
amogh@caida.org

Pere Barlet-Ros
UPC BarcelonaTech
pbarlet@ac.upc.edu

***Abstract*—We present the design and evaluation of `ITMgen`, a tool for generating synthetic but representative Interdomain Traffic Matrices (ITMs). `ITMgen` is motivated by the observation that gravity-based models do not reflect application level or regional characteristics of Internet traffic. `ITMgen` works at the level of connections, taking into account the relative sizes of ASes, their popularity with respect to various applications, and the relation between forward and reverse traffic for different application types. The necessary parameters for integrating application types and the distribution of content popularity can be realistically estimated by combining public sources like Alexa that capture traffic trends at a macro level with local traffic sampling (NetFlow, DPI) for providing an additional enhancement layer at the micro level. Using the above philosophy we demonstrate that we can synthesize ITMs that match real-world measurements closer than the current state of the art. In addition, the modular design philosophy of `ITMgen` makes it easy to integrate additional enhancement layers that improve the accuracy of our existing implementation.**

I. Introduction

The knowledge of interdomain traffic characteristics is important for a number of reasons, particularly related to economics and policy, as the flow of money on the Internet depends on the flow of traffic. A comprehensive understanding of interdomain traffic characteristics has consistently remained elusive, primarily due to the difficulty of obtaining representative traffic data which is often viewed as sensitive information. However, we need realistic interdomain traffic matrices in order to model and simulate new interdomain interconnection policies, pricing schemes, or routing protocols. Moreover, simulations of the interdomain Internet often need to be at different scales than the real Internet (which consists of more than 40,000 networks), either "shrinking" the actual traffic matrix for scalable modeling and simulation, or to investigate "what-if" scenarios in the evolution of the Internet. Researchers have mostly had to rely on synthetic interdomain traffic matrices generated using ad-hoc methods, reproducing some high-level characteristics of the interdomain traffic matrices such as heavy-tailed traffic volume distributions, or the presence of large traffic sources and sinks [10], [12], [14]. However, the research community lacks a configurable tool for producing synthetic traffic matrices of arbitrary size that match basic real interdomain traffic characteristics in more detail.

To fill the gap, we present in this paper the design and evaluation of `ITMgen`, a new tool to generate representative synthetic interdomain traffic matrices. `ITMgen` is based on first-principles, and incorporates several features that result in more representative traffic matrices than the current state of the art [9]. First, we model interdomain traffic at the level of *connections*, taking into account the relative sizes of ASes measured by the number of users they serve. Second, we model multiple content (or application) types, and their effect on interdomain traffic in terms of the ratio of forward to reverse traffic that each application type produces. Third, `ITMgen` captures the fact that the popularity of content objects shows regional effects - certain websites, for instance, may be more popular in specific countries or geographical regions. Finally, `ITMgen` is designed to be parameterized with high-level input data that is available publicly, and we provide such a *canonical* parameterization that represent present-day interdomain traffic characteristics. `ITMgen` is designed to be highly configurable and extensible; when new content types emerge and data about them becomes available, `ITMgen` can be easily extended to incorporate the new data. *We are making the* `ITMgen` *tool, and the data required to parameterize it available to the research community [1].*

The remainder of this paper is organized as follows. Sec. II discusses related work. Sec. III describes the design of `ITMgen`. Sec. IV describes the datasets used. Sec. V demonstrates how `ITMgen` can be parametrized and how to synthesize a matrix. The validation is presented in Sec. VI. Sec. VII concludes the work.

II. Related work

Most prior work on traffic matrix estimation and generation focused on *intradomain traffic* (see [7], [10], [17], [18], [20], [21] and references therein). Although those solutions give useful hints about synthesizing interdomain traffic matrices, they cannot be applied directly to the interdomain context. A prior paper on modeling intradomain traffic that inspired our work was by Erramilli et al. [11], which modeled intradomain traffic at the level of individual connections.

Several studies have measured interdomain traffic characteristics. An early study by Fang et al. [12], confirmed by [10], [17], showed that interdomain traffic distributions are highly non-uniform. Labovitz et al. [14] reported that interdomain traffic has been consolidating. Maier et al. [15] characterized residential broadband traffic. Bharti et al. [7] report on the sparseness of the ITM, and propose methods to infer the invisible elements of the ITM. Mikians et al. [16]

confirmed the sparseness of ITM, heavy-tailed distribution of sent and received traffic volumes, and measured the global and regional popularity associated with content sources. Feldmann et al. [13] present a methodology to estimate web traffic demands by analyzing CDN logs. While these studies do not directly measure ITM, the research community has mostly relied on measurements reported in these studies to synthesize ITM for modeling and simulation purposes.

The only work presenting a full approach to model interdomain traffic matrices by Chang et al. [9], which uses the *gravity model* to estimate the traffic between the ASes. The authors model ASes with a mix of "utilities" (business, residential, web hosting) and attribute the traffic accordingly to the interacting AS types. In contrast, we do not attribute types or "utilities", but rather distribute users and content, and model their interactions. A further difference is that our model is *topology agnostic*, and does not require knowledge of the interdomain topology in order to synthesize an ITM.

III. ITMgen design

The design of `ITMgen` is based on first-principles, modeling traffic at the level of connections and taking into account traffic asymmetries based on application type and the effects of regional/global content popularity. We emphasize that we focus on generating *static snapshots* of the interdomain traffic matrix. Although such a static model might be sufficient for applications such as Internet economics or network formation, other areas may requite a model that captures temporal effects. We strive to expand the model along the temporal dimension. Next, we summarize the key decisions underlying the design of `ITMgen`.

Connection-based

The interdomain traffic matrix, by definition, is concerned with the terminating ends of the Internet, i.e., it measures the traffic that originates at an AS X and terminates at AS Y. We recognize that such traffic is from *connections that originate from and/or terminate at individual users*. We thus make the design decision to model interdomain traffic at the level of connections, and the traffic exchanged by an AS will depend on the number of users in that AS.

Content types

The Internet caters to a variety of different applications, such as web, peer-to-peer file sharing, streaming video, conferencing, etc. Given a connection, the ratio of traffic flowing in the two directions (traffic asymmetry) over that connection depends on the nature of the application. For example, in the case of client-server applications, the traffic asymmetry will be determined by the ratio of the size of data packets to the size of acknowledgements. In the case of P2P traffic, we expect more symmetric traffic. We explicitly model different application types in `ITMgen`. Note, however, that we are considering different traffic types, and not necessarily network types; the same network can thus host different applications.

Regional and global popularity

We recognize that content popularity on the Internet shows both global and regional effects. With respect to web content, for example, websites such as Google and Facebook are popular worldwide; on the other hand, some websites cater to specific countries or regions. Such regional websites may be highly popular traffic sources for ASes in the same region, but they are not popular for ASes in a different region. `ITMgen` takes into account the global and regional popularity associated with content objects.

Parameterizable using commonly available data

`ITMgen` can be parameterized using commonly available data sources, which measure interdomain traffic characteristics at a high level. Further, we have designed `ITMgen` to be *extensible* to accommodate new application types that may emerge in the future. A user can extend the tool whenever data about new application types - the traffic parameters for the new application type, the global and regional popularities of ASes w.r.t. that application type - are available.

A. Traffic model

`ITMgen` models the traffic between two ASes as an interaction between users and content within the ASes, facilitated by a set of distinct applications. Consider an example where users in ASes U_1 and U_2 are accessing objects stored on machines in ASes M_1 and M_2, using an application A_1. The volume of this user-to-machine (U2M) traffic depends on the number of users in U_i, the popularity of content in M_j, and the nature of traffic produced by application A_1. Moreover, the popularity of M_1 can be different for U_1 and U_2, for example due to a regional bias. Users in U_1 and U_2 also interact using application A_2, generating user-to-user traffic (U2U). The obvious examples of applications that produce U2M and U2U traffic are browser-based services and P2P, respectively. In our study we omit machine-to-machine (M2M) traffic. The reason is twofold: traffic reports like [5] do not indicate that M2M traffic volume is significant in access networks. Also, access to the packet level data at the level of non-access ASes (*i.e.*, business ASes) is highly restricted. Therefore we only acknowledge that M2M traffic estimation will require further effort.

The traffic represented in the ITM is thus an aggregate of all the individual interactions between users and content in different ASes. There are two levels at which these interactions need to be characterized in order to generate an ITM. At the *macro level*, the traffic between two ASes depends on the number of users and the popularity of the content hosted within those ASes. The common gravity models [9], [19] operate at this level. This level of description is insufficient to capture more elusive aspects of the traffic, namely what happens at the application level. Therefore, we enhance the macro-information with the *micro-level* view which describes the actual interaction between users and content objects.

Combining the macro and micro-level views, traffic from AS i to AS j can be expressed as

$$T_{i,j} = \sum_{\kappa} m_{\kappa} \left(S_i p_i^{\kappa}(j) + d_{\kappa} S_j p_j^{\kappa}(i) \right) \quad (1)$$

S_i denotes the number of users in AS i. $p_i^{\kappa}(j)$ denotes the relative popularity of j *subjective to* i *and with respect to*

application κ. The two terms in the summation represent the traffic from a user to an object due to application κ, and the traffic produced by that application in the reverse direction. The (a)symmetry in the two directions of traffic due to application κ is denoted by d_κ, and this parameter is application-dependent. The parameter m_κ represents the contribution of each application to the overall traffic mix.

In the rest of this paper, we describe how to parameterize `ITMgen` with respect to these two applications. Although both groups contain more applications (Skype, mail, etc.), web and P2P in particular contribute to the bulk of interdomain traffic [14], [5]. `ITMgen` can easily be extended to add more application types, as long as the relevant information to parameterize them (m_κ, d_κ and popularity) is available.

A curious reader could notice that we do not consider network topology. Our goal is to model traffic resulting from an interaction between users and content; the user's decision to access a specific content does not depend on the topology.

IV. Dataset description

We give a brief overview of the datasets we have used to parameterize (Sec. V) and validate (Sec. VI) `ITMgen`.

We use the **Alexa** [2] list of global top 1 million websites to measure the popularity of ASes with respect to web content. Alexa also provides per-country statistics[1], which we use to determine the *regional* popularity of ASes. To estimate the popularity of ASes with respect to peer-to-peer traffic, we rely on data obtained by crawling the BitTorrent (BT) tracker (openbittorrent.com). To obtain the number of users per AS, we relied on open **marketing reports** from ISPs [1].

To obtain the micro-level information regarding application characteristics (ratio of forward to reverse traffic) and the fraction of traffic accounted for by various application types, we rely on a two-week long **packet level trace from CESCA** [3]. Although CESCA is a fully fledged AS and access to the packet level data at that level is difficult, we strive to confirm our results with other data sources. We deliver `ITMgen` with preconfigured parameters in case a researcher does not have access to the relevant low level data.

For validating `ITMgen`, we use traffic statistics for 3 ISP ASes from **Telefonica**, a world-wide Internet connectivity provider. For those ISPs we analyze traffic statistics for the top 1000 ASes; as the traffic distribution was heavy tailed, those top entries contribute to more than 95% of the total traffic. The statistics come from international access links, therefore some of the regional (country) traffic can be undervalued and we use this data only where this shortcoming is insignificant.

V. Parameterization and synthesis

In this section we describe how each of the parameters in (1) can be estimated from real-world measurements. We provide this measurement data and the associated parameterization as the *canonical parmeterization* of `ITMgen`.

[1]We use "page views" metric provided by Alexa, together with per-country breakdown.

A. Number of users: S_i

Our model requires an estimate of the number of users in each AS, which we characterize as follows. Using publicly available marketing data and annual reports, we obtained the market shares of ISPs for the top-10 countries in the world according to the number of Internet subscribers [6]. This gives us insight into ISP market shares, but not per-AS estimates. For each ISP, we then obtained the set of ASes belonging to that ISP using *whois* data. For these ASes we measured the number of IP addresses in our BT logs, and split the subscribers of the ISP among different ASes in proportion to the number of IP addresses seen from each of those ASes in BT. The assumption is that approximately the same fraction of users in each AS belonging to an ISP participate in BT file sharing. This gives us an empirical distribution of the number of Internet users per AS, for about 400 ASes. Although this represents only 1% of the total number of ASes, these contribute to about 60% of the total number of Internet subscribers in the world.

In addition to the number of users per AS, we need to determine the fraction of ASes in the world that *do not serve any users*. Such ASes could host content (pure content providers), or provide transit service (pure transit providers). Pure transit providers do not appear in the interdomain traffic matrix, as they do not source or sink any traffic. To find pure content providers, we obtain the set of ASes that host websites represented in the Alexa list. Of these ASes, we separate the ones that do not show any BT activity in our BT logs. We thus find that at least 42% of ASes do not serve end users. We emphasize that these are rough estimates and can be easily improved as more precise data becomes available.

B. Content Popularity: $p_i^\kappa(j)$

Another macro-level parameter is the popularity of an AS with respect to various content types. Vector $p_i^\kappa(j)$ describes the fraction of traffic generated by an average user in AS i that is sent to AS j. Recall from Sec. III-A that the traffic between two ASes is proportional to the popularity of the content objects hosted by that AS. Moreover, the popularity can be *subjective*, i.e., some ASes are likely to be popular only in their own region, while others are globally popular. The bias can be easily observed in the rankings of ASes calculated from Alexa: for top-10 most popular ASes in 20 examined regions, by average 53% of them were from that region. As a result, we assign to each AS i a popularity vector $p_i^\kappa(j) : \sum_j p_i^\kappa(j) = 1$ describing the *subjective* popularities of the other ASes, as visible from i.

Web popularity

To gain an insight into the popularity of the WEB traffic, we used *Alexa page views* statistics. Although this metric does not reflect literally the actual traffic volume for the AS, the number of page accesses per AS will impact the generated traffic, and we believe that it can serve as the basis for comparison between ASes. Figure 2 shows the distribution of AS popularity for different regions. Strikingly, the underlying distribution appears to be similar for all 20 examined regions (not shown in the figure) and the corresponding Zipf slope falls typically

into range $(1.13, 1.28)$. Some ASes, e.g., Google, Facebook, etc. are expected to be popular in many regions. Moreover, an AS that is among the most popular ASes in region A can also be among the most popular ASes in region B. To confirm the intuition, we computed the pairwise Spearman correlation of the rankings in all the considered regions. We found a relatively high Spearman correlation (0.62), indicating that there does exist correlation between the top ranking ASes in different regions.

To synthesize the ITM, we need to define a procedure to create $p_i^\kappa(j)$ that (1) has a certain statistical distribution (resembling measurements), (2) keeps the notion of local and global popularity of ASes (to distinguish between, for example, a global content provider and large regional hosting provider), and (3) preserves ordering (e.g., for two globally popular ASes X and Y, X will be always more popular than Y). The following procedure builds $p_i^\kappa(j)$ for an AS i that captures those three properties. First, we split all ASes into three ordered groups: globally popular, locally popular and the remaining ones. Next, from a Zipf distribution we generate a random vector q (sorted in descending order) of length n, where n is a number of ASes. This vector contains the popularities of remote ASes from the perspective of AS i. Then, n times we pick an AS j from a random group (globally popular, locally popular, or other) and assign the next value from q to $p_i^\kappa(j)$. This way we build $p_i^\kappa(j)$ for a specific AS i.

P2P Popularity

As we mentioned in Sec. III-A a prevailing U2U application is P2P file sharing, which is responsible for most of U2U traffic between ISPs. In this section we describe the parametrization of P2P popularity vector $p_i^{P2P}(\cdot)$.

We estimate the relative popularity of different ASes for P2P content using BT measurements. To this end, we measured the number of IP addresses from each AS seen in our BT crawls. Figure 4 shows the distribution of the active P2P peers per AS. The flat section of the plot suggests an underlying power-law distribution, which is more evident after binning the data. The bent tail could be the effect of an information bottleneck, e.g., insufficient measurement time [8]. To build the popularity vector w.r.t. P2P traffic $p_i^{P2P}(\cdot)$, we draw a vector of random variables from the fitting Zipf distribution with slope 1.63 and assign the generated values, from the highest to the lowest, to ASes in the order of descending number of users. We refrain from modelling regional popularity in P2P traffic, as those effects are difficult to estimate precisely from BT data. As P2P contributes a relatively small fraction of overall Internet traffic [5], [14], we accept the error introduced by not considering the locality of P2P. Nevertheless, it is possible to use measurement-based insights on the regional distribution for P2P [19], together with a procedure similar to the one used for WEB to assign regional P2P popularities.

C. Application mix: m_κ, d_κ

As mentioned in Sec. III, `ITMgen` recognizes the fact that traffic at its micro level is a mix of different applications, which is expressed in (1) by κ. There are two crucial parameters we must estimate at the micro-level. The parameter d_κ, which describes the ratio between the two directions of traffic generated by application κ, and m_κ, the fraction of the traffic generated by an average user, due to application κ. These application-specific characteristics cannot be obtained from the macro level data; to this end, we monitored the CESCA access link for 14 days. To classify the applications we used the commercial *PACE* [4] tool for deep packet inspection, which in our case yielded only 13% of unclassified traffic[2].

Our measurements indicate that in the case of WEB traffic the ratio per flow $log_{10}(d_\kappa)$ typically falls into the range $(0.4, 1.5)$ and for P2P traffic into the range $(-0.87, 1.25)$. It is unsurprising that d_{WEB} is skewed, since for WEB one direction of the traffic is predominant. Also the upper bound of d_{WEB} is determined by MTU as $log_{10}\frac{MTU}{TCP\ Ack} \approx 1.56$. Interestingly, the ratio for P2P traffic is both positive and negative, suggesting that some P2P clients use the same connections, once established, to both upload and download the exchanged content. In the latter examples we use the statistical distributions that best fit the measurements, i.e., normal d_{WEB} and uniform and d_{P2P}

To explain the exact role of m_κ consider the following example: a user downloads a file from a server and $d^{WEB} = \frac{MTU}{Ack}$. Also, the same user exchanges P2P traffic, and $d^{P2P} = 1$. If both applications use the same total bandwidth, it does not mean that the upstream flows (from the user to the object) are the same size: the upstream flow of WEB is smaller than that of P2P. The parameter m_κ reflects this difference in the traffic mix originally generated by an average user. Based on our measurements we choose $m_{P2P} = 0.65$ and $m_{WEB} = 0.35$. We strive to compare those results with the data from other vantage points.

VI. Validating ITMgen

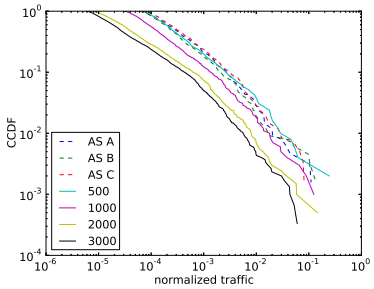
In this section we validate `ITMgen`. First, we perform some sanity checks to show that a synthetic ITM generated by `ITMgen` reproduces well-known characteristics of the real ITM. Later, we discuss the advantages of `ITMgen` over a common gravity model (GM) [9].

A. Sanity checks

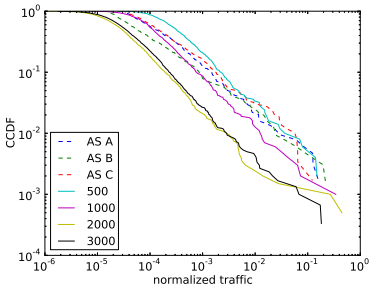
One of the properties of the ITM observed in [7], [16] is it's low rank, meaning that the matrix can be approximated by a small number of independent vectors. The reason of the low rank is that a small number of the most popular ASes (rows/columns) capture the bulk of the traffic. To verify that `ITMgen` produces ITMs with this property, we computed the eigenvalues of a synthetic ITM with 1,000 ASes. Less than 30 out of 1,000 values were significant, confirming that the low rank property holds for `ITMgen` generated matrices.

Next, we compare the statistical distributions of the traffic exchanged by ASes in the synthetic ITM and that seen in the Telefonica dataset. Figure 1 shows results for 3 ISPs from Telefonica, and selected ASes from the generated ITMs that

[2]The overall accuracy was affected by the packet capturing process (e.g., packet drops and truncated flows), which are not related with PACE.



(a) Tr. produced



(b) Tr. consumed

Fig. 1. Statistical distribution of the traffic produced and consumed by the observed ASes, for the Telefonica data (dashed line) and the model (solid) for the synthetic ITMs of different sizes.

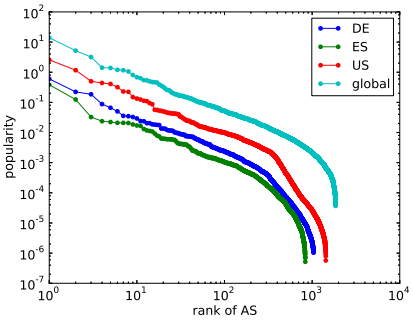
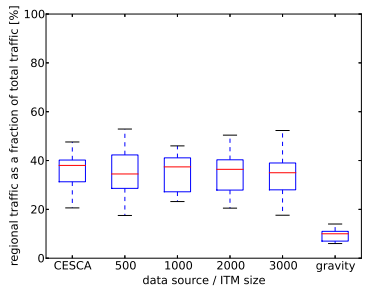
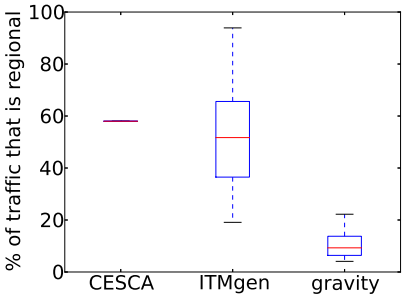


Fig. 2. WEB popularity distribution of ASes, globally and for three example regions.



(a) Traffic exchanged with ASes within same region; matrices of 4 different sizes are shown.



(b) Regional traffic of CPs.

Fig. 3. Regional traffic exchange.

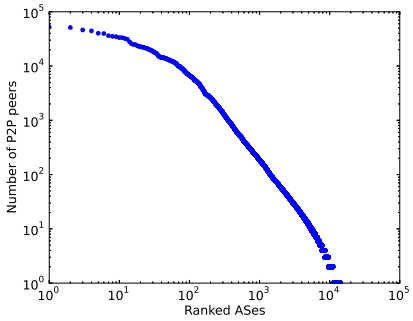


Fig. 4. P2P activity distribution.

have a similar number of users, relatively to the total number of users in all ASes. We perform this analysis for synthetic ITMs of different sizes. We observe that the distribution of traffic produced by synthetic ASes is qualitatively similar to that in the measurements. On the other hand, the traffic consumed by synthetic ASes appears to be more skewed than in the measurement data (the slope of 0.89 for the measurements and 0.69 for the model). Although the mismatch is visible, we do not aim to match exactly the special case visible in the plot. Simulation parameters, in particular content popularity distribution, can be adjusted to match desired special cases.

B. Regional effects

`ITMgen` explicitly models regional biases in content popularity. We present those effects introduced by `ITMgen`, and compare them with real measurements, and with a synthetic ITM produced by GM. For this purpose, we model 10 regions with equal number of ASes. As GM does not introduce locality, it will result in a random assignment of ASes to the regions.

We analyze traffic locality for two types of ASes: ISP and CP. From the generated ITM, we select randomly 25 ASes with a similar relative number of users, and calculate the traffic that those ASes exchange with the ASes within the same region. We repeat the same procedure for the ITM generated by GM. We calculate the regional traffic of each institution in CESCA. Figure 3(a) shows the fraction of traffic that is exchanged with ASes in the same region for matrices of different sizes generated by `ITMgen`, a synthetic ITM generated by GM, and measurement data from CESCA. We observe that CESCA traffic is regionally biased - almost 40% of the traffic is exchanged with ASes within the same region. This bias is also reflected in the synthetic matrices produced by `ITMgen`, regardless of their size, as shown in Fig. 3(a). We also observe that GM produces an ITM with significantly less regional traffic.

Next, in Fig. 3(b) we compare local traffic from the point of view of CPs in the CESCA data. Although we did not have access to a true CP AS, we analyzed the traffic from/to the content servers inside the CESCA AS. The figure shows that in the measurement data from CESCA, about 60% of traffic is local; the synthetic ITM produced by `ITMgen` shows similar fractions, while GM clearly underestimates regional traffic.

C. Application mix

An important feature of `ITMgen` is that it offers the possibility of modeling the traffic mix in terms of applications. To this end, we show the application mix resulting from `ITMgen`, and later discuss a what-if scenario that considers a new application.

Various reports [14], [5] and our DPI measurements at CESCA suggest that P2P contributes between 9% and 21%

of the overall traffic. In the synthetic ITM we observe that P2P contributes an average of 27% of the traffic. The overestimation can stem from the fact that we model only two applications, whereas the mentioned measurements consider all possible applications.

D. Use case - cloud storage

Here we discuss how to introduce a new application to the ones already modeled by `ITMgen`. We consider "cloud storage" (ST), a service that allows a user to synchronize her data over a cloud that is managed by an external enterprise.

Recall that to model a new application, the user must specify both the macro and micro-level properties of that application. First, we consider the macro level characteristics of ST, expressed by the popularity vector $p_i^{ST}(j)$ (see Sec. V-B). Recall that p describes the popularity of AS j, as seen from AS i. We consider a hypothetical scenario where the storage is provided by three major global content providers, and we assign $p_i^{ST}(j)$ proportionally to $p_i^{WEB}(j)$ so that the more an popular AS already is, the more ST traffic it will attract. Next, we specify the micro level parameters. We simulate that the users generate an additional 5% of upstream traffic due to ST ($m_{ST} = 0.05$). We also simulate that the traffic generated by ST is skewed (upload files from one point and send to many points), and we model the traffic ratio $log_{10}(d_{ST})$ with a normal distribution $N(0.7, 0.2)$.

This information is sufficient to model the new application. We generated synthetic ITMs with `ITMgen`, considering the new application in addition to Web and P2P traffic. Analysis of the synthetic ITMs suggests that ASes providing cloud storage will increase their traffic from 16% to 20%, and the overall traffic generated by all ASes will increase by 9.1%. This example shows how `ITMgen` can be used to model various what-if scenarios related to new application types.

VII. Conclusions and future work

Modeling the interdomain traffic matrix is a challenging task, as it is impossible to obtain its full view. In this paper, we present `ITMgen`, a tool to build synthetic ITMs of arbitrary size. To the best of our knowledge, `ITMgen` is the only alternative to the current state of the art in interdomain traffic matrix estimation [9]. `ITMgen` takes a first-principles approach, and differs from that work in several significant ways - it models traffic at the level of connections, is topology-agnostic, and takes into account both regional and global popularity of content types. We are aware that `ITMgen` has both advantages and disadvantages compared to GM. `ITMgen` is extensible; it can be easily extended as the dominant application mix of interdomain traffic changes, and data about new application types becomes available. We show how to parameterize `ITMgen` using mostly data that is available publicly. On the other hand, it might be challenging to parameterize and it describes only relative traffic between ASes.. We are releasing `ITMgen` as a tool to enable researchers to generate synthetic, but representative traffic matrices for modeling and simulation purposes.

Our ongoing work is directed towards better parameterization of the model. This includes examining how allocation of the content on CDN servers affects popularity p of the ASes and considering M2M traffic. Also, we strive to capture the dynamics of the ITM, as we are aware that for many applications a static view of the traffic is insufficient. We are working on expanding the model with a temporal component. Moreover, an area of future work which we are exploring is to simultaneously generate a synthetic interdomain topology and the associated ITM. While we have taken a topology-agnostic approach in this work, we acknowledge that there may exist correlations between topological and traffic-related properties of an AS. Exploring such correlations, and incorporating them into a synthesis model will be the focus of our future work.

VIII. Acknowledgements

We acknowledge *ipoque* for kindly providing access to their PACE traffic classification engine for this research work. J. Mikians was funded by FI Grant 2010FI_B 00512 from Generalitat de Catalunya. The research was funded by the Spanish Ministry of Science and Innovation under contract TEC2011-27474 (NOMADS project). A. Dhamdhere was financially supported by the NSF (grant CNS-1017139). Authors would also like to thank CESCA for allowing them to collect the data used in this work.

References

[1] `ITMgen` tool and other resources (e.g., the marketing reports) can be found at http://monitoring.ccaba.upc.edu/itmgen
[2] http://www.alexa.com
[3] http://www.cesca.cat/en/communications/anella-cientifica, regional research network and AS, containing universities and research units.
[4] ipoque. PACE: Network Analysis with DPI. http://www.ipoque.com
[5] Sandvine Global Internet Phenomena Report: Fall 2011
[6] CIA - The World Factbook. https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html (2009), (accessed 2012)
[7] Bharti, V., Kankar, P., Setia, L., Gürsun, G., Lakhina, A., Crovella, M.: Inferring invisible traffic. In: CoNEXT (2010)
[8] Cha, M., et al.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: IMC (2007)
[9] Chang, H., Jamin, S., Mao, Z., Willinger, W.: An empirical approach to modeling inter-as traffic matrices. In: IMC (2005)
[10] Chang, H., et al.: The many facets of Internet topology and traffic. Networks and Heterogeneous Media (2006)
[11] Erramill, V., Crovella, M., Taft, N.: An independent-connection model for traffic matrices. In: IMC (2006)
[12] Fang, W., Peterson, L.: Inter-AS Traffic Patterns and Their Implications. In: GLOBECOM (1999)
[13] Feldmann, A., et al.: A methodology for estimating interdomain web traffic demand. In: IMC (2004)
[14] Labovitz, C., et al.: Internet inter-domain traffic. In: ACM SIGCOMM (2010)
[15] Maier, G., Feldmann, A., Paxson, V., Allman, M.: On dominant characteristics of residential broadband Internet traffic. In: IMC (2009)
[16] Mikians, J., et al.: Towards a statistical characterization of the interdomain traffic matrix. In: IFIP Networking (2012)
[17] Nucci, A., et al.: The problem of synthetically generating IP traffic matrices: initial recommendations. ACM SIGCOMM CCR (2005)
[18] Roughan, M.: Simplifying the synthesis of Internet traffic matrices. ACM SIGCOMM CCR (2005)
[19] Seibert, J., et al.: The Internet-wide Impact of P2P Traffic Localization on ISP Profitability. IEEE/ACM Transactions on Networking (2012)
[20] Zhang, Y., et al.: Fast accurate computation of large-scale IP traffic matrices from link loads. In: ACM SIGMETRICS (2003)
[21] Zhang, Y., et al.: Spatio-temporal compressive sensing and Internet traffic matrices. ACM SIGCOMM CCR (2009)