

## Context

***On Thursday, Dec. 3rd, Herman Wagter and Benoît Felten published on their respective blogs a joint article expressing doubt at the considerable impact on network performance ascribed to so-called "bandwidth hogs" by some network operators and ISPs.***

***In order to prove or disprove this doubt, they issued a request to ISPs willing to participate: by sharing near real-time customer usage data, the exact role of heavy users in network disruption could be assessed. The aim of this document is to describe a dataset structure that would allow analysis of heavy usage patterns amongst residential internet end-users dependant on a shared backhaul link where congestion is likely to occur.***

***For more information on this topic, see [www.fiberevolution.com](http://www.fiberevolution.com) and [www.dadamotive.com](http://www.dadamotive.com).***

## Why participate ?

In our experience, most ISP operations lack the internal resources and time to do fine-grained analysis of customer usage patterns, and yet assumptions about usage patterns orient crucial strategic decisions.

The situation for small ISPs is particularly difficult since they are often squeezed between high backhaul bandwidth costs and customers who don't understand why the broadband service they subscribe to is not as "unlimited" as they thought it was. The introduction of bandwidth caps has been a response to the perceived problem of heavy user disruption, but if our hypothesis is correct, bandwidth caps will not contribute much to solving congestion issues. We hope this will give us all a better understanding of the real problems, and possible ways ahead. Also it will provide a much needed reference and support to small ISPs in their customer communication. Finally, it might help attract attention to the middle mile issues that make so many rural operations complex to sustain.

What we're offering is to perform a usage analysis on the basis of the dataset described herein. The results of this analysis will be - of course - presented to the participating ISP, but will also be published in aggregated form as a series of case studies and comparisons. The conditions of this collaboration are as follows:

- Our offer is to perform this work on our spare time. While we will try to be as diligent as possible, instantaneous results cannot be expected of us.
- While we aim to publish the results of our analysis, especially as it proves or disproves our hypothesis about "bandwidth hogs", publication will only identify ISPs who wish to be identified. Anonymity will be scrupulously respected for those who don't.
- Datasets shared with us should also anonymize end-users (see below)
- We reserve the right to refuse a particular collaboration if the conditions are not aligned with the fact this is voluntary work on our part. Our intention is neither to do huge amounts of work pro-bono nor is it to compete with regular data analysis businesses.

## Definitions:

- **Heavy Users:** Heavy Users are defined as customers who rank amongst the top downloaders / uploaders each month. They are the ones commonly called "bandwidth hogs" in media and telco parlance.
- **Disruptive Users:** Disruptive Users are defined as customers who create or amplify congestion by using up a significant proportion of bandwidth thus causing other users to experience a lower

performance. They would be identified as users that retain usage of a significant portion of bandwidth allocation even when congestion occurs.

- **Congestion:** Congestion is defined as a state of the shared bottleneck (usually the backhaul link) when it is saturated beyond a certain % of overall capacity (to be determined together with ISP, most likely around 70-80%)
- **Aggregation Cluster:** An aggregation cluster is a group of customers all connected to the same aggregation or backhaul link. As such, they are potentially interdependent.

### Targeted output:

- Assess whether Disruptive Users exist and if yes in what proportion.
- Examine bandwidth allocated to Disruptive Users during congestion
- Examine whether disruption is caused by the same users over multiple instances of Congestion
- Examine the two-way correlations between Disruptive Users and Heavy Users
- Examine correlations between number of open Virtual Circuits and Disruptive Usage

### Dataset definition:

In order to understand the correlations between usage patterns and congestion, it's necessary to examine end-user bandwidth usage across time and in an area where all users affect each other. In order to obtain a compromise between a detailed dataset tracking all peaks and one that remains of manageable size, I propose that the analysis be over a full month, with measurements every two hours.

Therefore the scope of the dataset needs to be **all users dependant on a single aggregation or backhaul link** where congestion is likely to occur. For the sake of expedient analysis, it would be ideal if said dataset didn't regroup more than 10.000 users. If business and residential users share the same backhaul link they should both be included, but business customers should be flagged. In the rest of this document this group of interdependent customers will be called an aggregation cluster.

The dataset should essentially be comprised of three tables:

**Table 1: Spot Measure of Bandwidth Usage in an aggregation cluster.**

Table 1 will be constituted of time stamped records of upload, download and VC usage by every customer in the cluster every two hours. The base will look something like this:

- **Customer ID:** anonymized unique customer identifier
- **Timestamp**
- **Download usage in Mb/s**
- **Upload usage in Mb/s**
- **Number of open VCs**

Note that this table will have around 360 lines per customer ID, which is why it seems sensible to limit the analysis to a single cluster of customers to begin with (if said cluster is roughly 5.000 customers, that's already 1 800 000 lines.)

**Table 2: Spot Measure of Aggregation Link Congestion**

- **Timestamp**
- **Aggregated Download in Mb/s**
- **Aggregate Upload in Mb/s**
- **Total Number of open VCs**
- **Total Link Download Capacity**
- **Total Link Upload Capacity**

Note that the timestamps here need to be exactly identical to the timestamps in table 1 for the analysis to have any validity.

**Table 3: Customer Information**

- **Customer ID:** anonymized unique customer identifier
- **Total Monthly Download in GB**
- **Total Monthly Upload in GB**
- **Nominal Download Rate (from customer plan)**
- **Nominal Upload Rate (from customer plan)**

Let me stress again that all user information should be anonymized before being submitted!

## **Data Format:**

Data Format should be CSV for easy import into SPSS, the software we will use for analysis. If this is not convenient, please contact us to discuss alternatives.