Towards Fixing Internet Measurement Infrastructure Biases

Emile Aben, Romain Fontugne

December 2020

1 Problem Statement

A common misinterpretation about large scale Internet measurement infrastructure (LSIMI) is that effects measured by using all data coming off of these is actually a fair reflection of the state of the Internet. LSIMI like RIPE Atlas for data plane data, and RIS, Routeviews, and Isolario on the capture of control plane data are large, but do have their biases, that make them quite poor if one wants to see a representative picture of the Internet. There is an inherit bias in the way new data collection points are added and no structural fix to address this bias. Many studies either explicitly or implicitly assume that effects seen by LSIMI are either qualitatively or quantitatively representative for the wider Internet, and we have little or no methodology to even know to what extent this is true.

It is all to easy to say 40% of RIS full feeds saw route X, where that gets (implicitly or explicitly) interpreted as 40% of the Internet saw route X. Whoever reads studies based on LSIMI data has to continuously stay vigilant of their own interpretation of such simple facts.

2 Research Question

This leads to the following research questions: **Can we measure the bias of large scale Internet Measurement infrastructures**? And a follow up question: **Can we counter the bias of large scale Intermet Measurement infrastructures**? Can we work on deployment strategies and/or data analysis methods that work towards this unbiasing. To make future data collection more sustainable, issues related to bias should be considered a priori to data collection, as these put more focus on making data collection fit for purpose.