AusNOG 2016

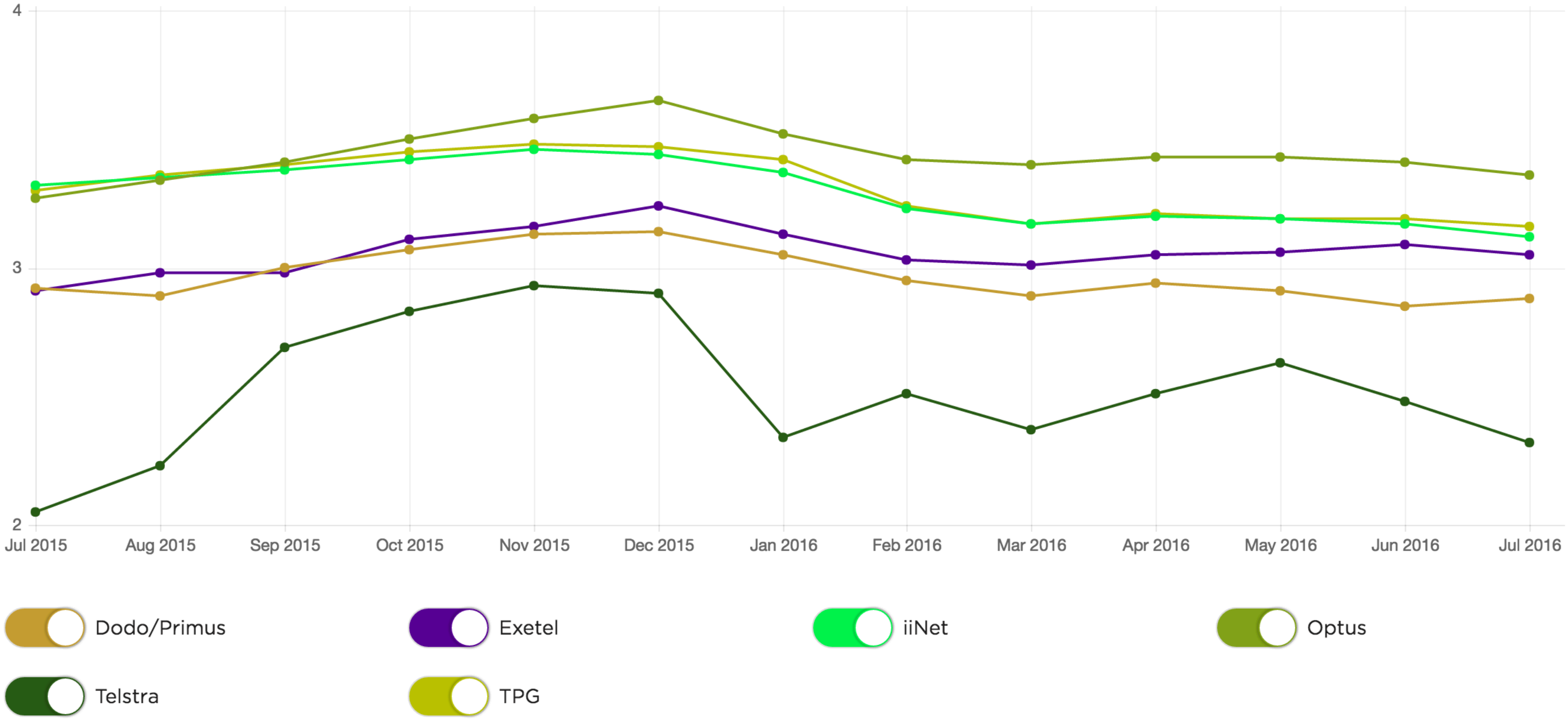
1 & 2 September 2016, Swissotel, Sydney

Rethinking Broadband Performance using Big Data from M-Lab

Xiaohong DENG
xiaohong.deng@xhdeng.com
Thanchanok sutjarittham,
Vijay Sivaraman
UNSW
Blanca Gallego
Macquarie University

1 September 2016

How do you think of your ISP?



This Talk

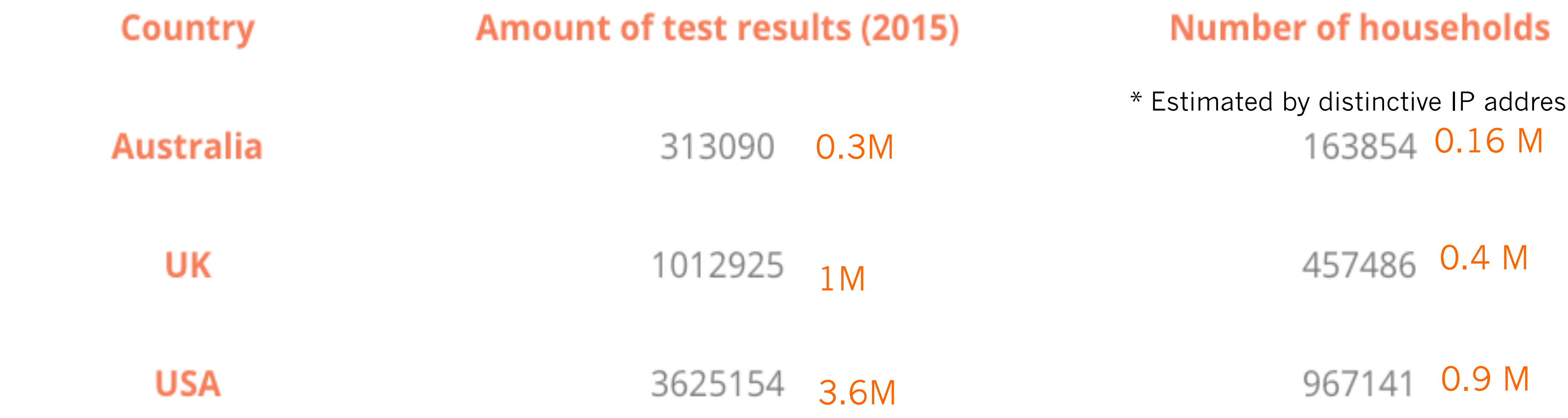
- Overview: Data, Speed, user test pattern
- Problems:
  - Various Network Variables effect on Speeds; different distribution for different ISP
  - Sampling bias
- Solution:
  - Consider all affecting variables:
    - Casual inference model
  - De-Biasing
- Conclusion

- Data at a Glance

*

NDT does:

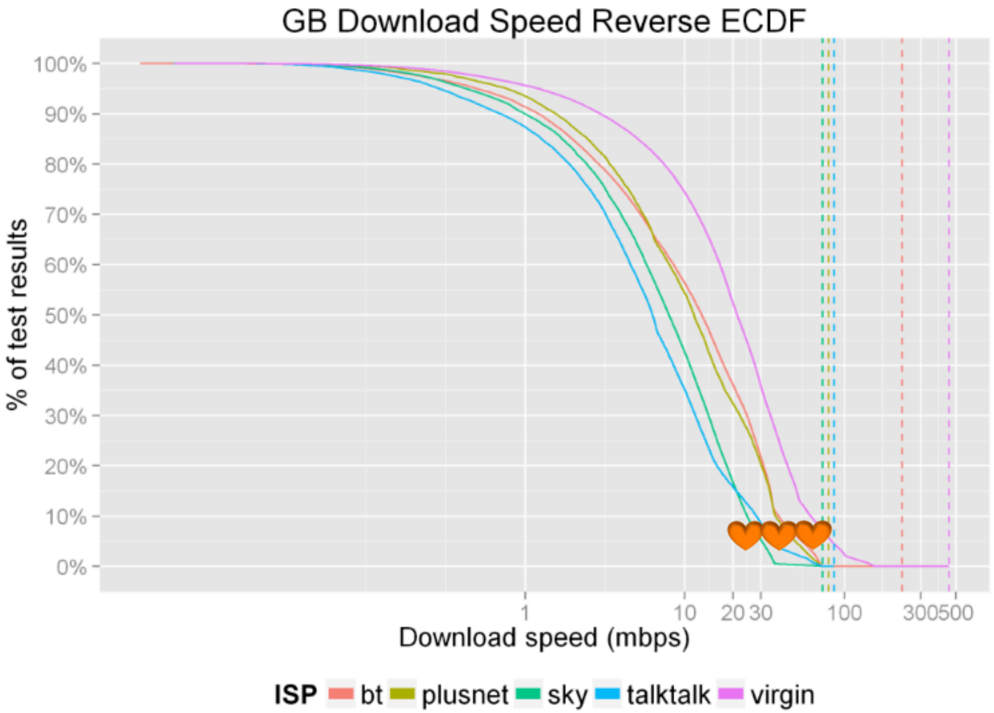
- TCP performance / speed test:
- Record TCP web100 variables during one test session: RTT, loss, MSS, Congestion signal counts, ECN..



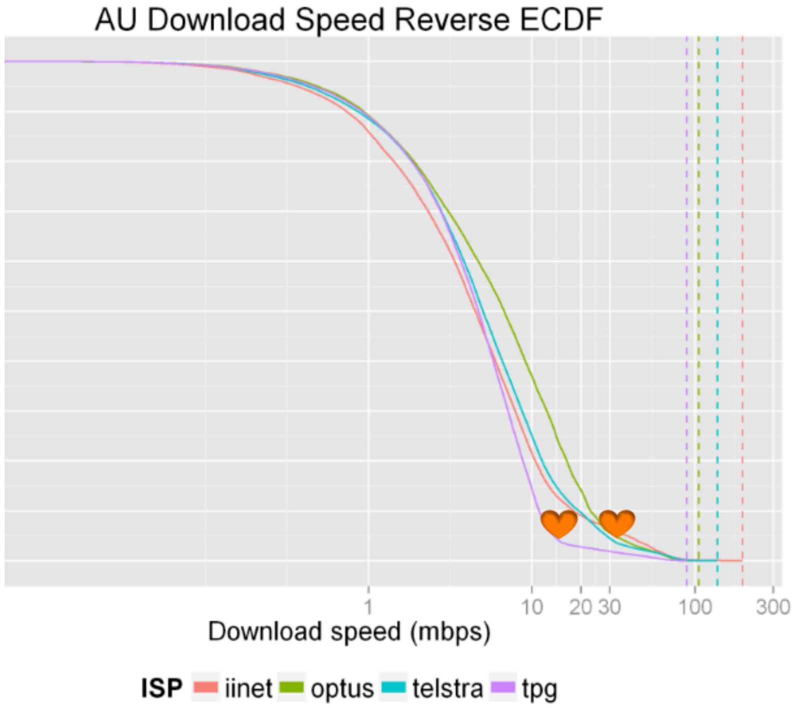
Country	Amount of test results (2015)	Number of households*
Australia	313090 0.3M	163854 0.16 M
UK	1012925 1M	457486 0.4 M
USA	3625154 3.6M	967141 0.9 M

* Estimated by distinctive IP addresses

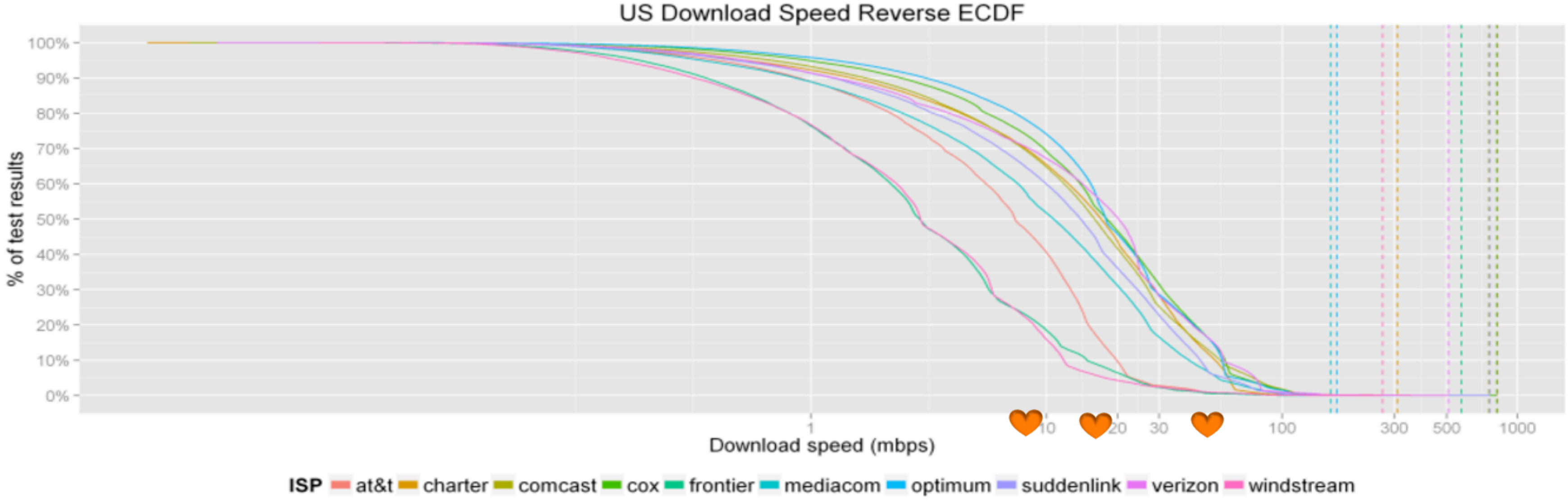
Amount of NDT Data in 2015



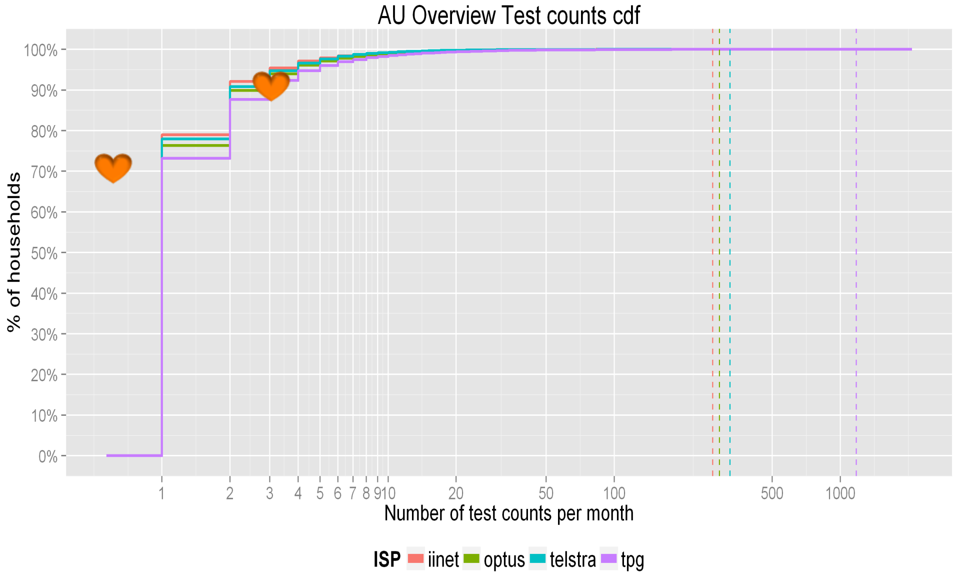
GB Download Speed Reverse ECDF

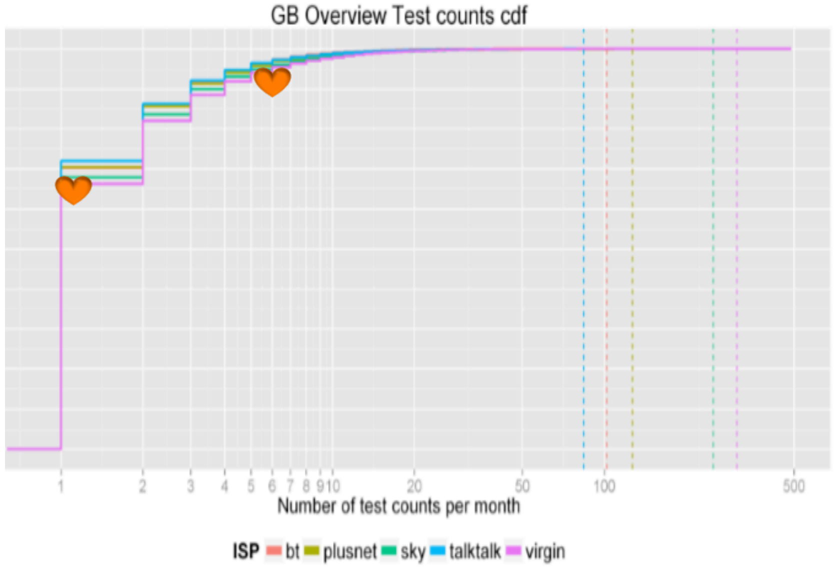


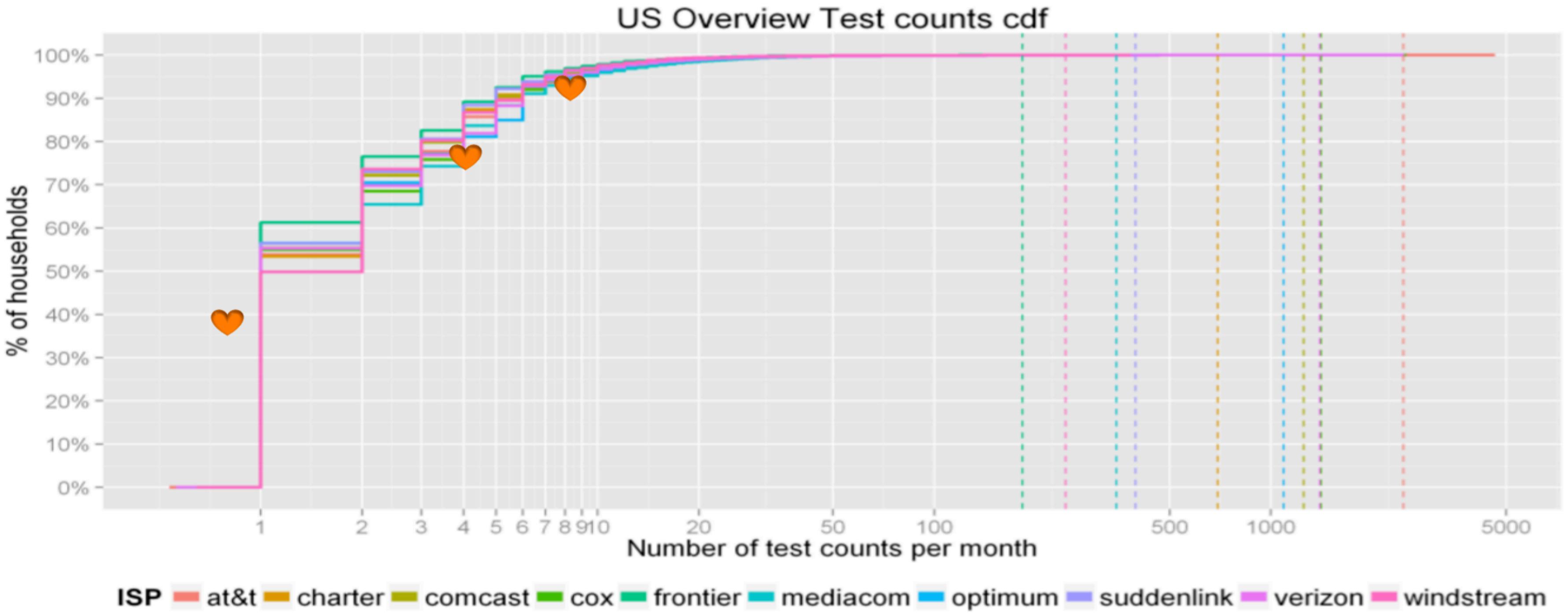
AU Download Speed Reverse ECDF



US Download Speed Reverse ECDF


AU Overview Test counts cdf
% of households
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
20
50
100
500
1000
Number of test counts per month
ISP
iinet
optus
telstra
tpg


GB Overview Test counts cdf
1
2
3
4
5
6
7
8
9
10
20
50
100
500
Number of test counts per month
ISP
bt
plusnet
sky
talktalk
virgin


US Overview Test counts cdf
% of households
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
20
50
100
500
1000
5000
Number of test counts per month
ISP
at&t
charter
comcast
cox
frontier
mediacom
optimum
suddenlink
verizon
windstream

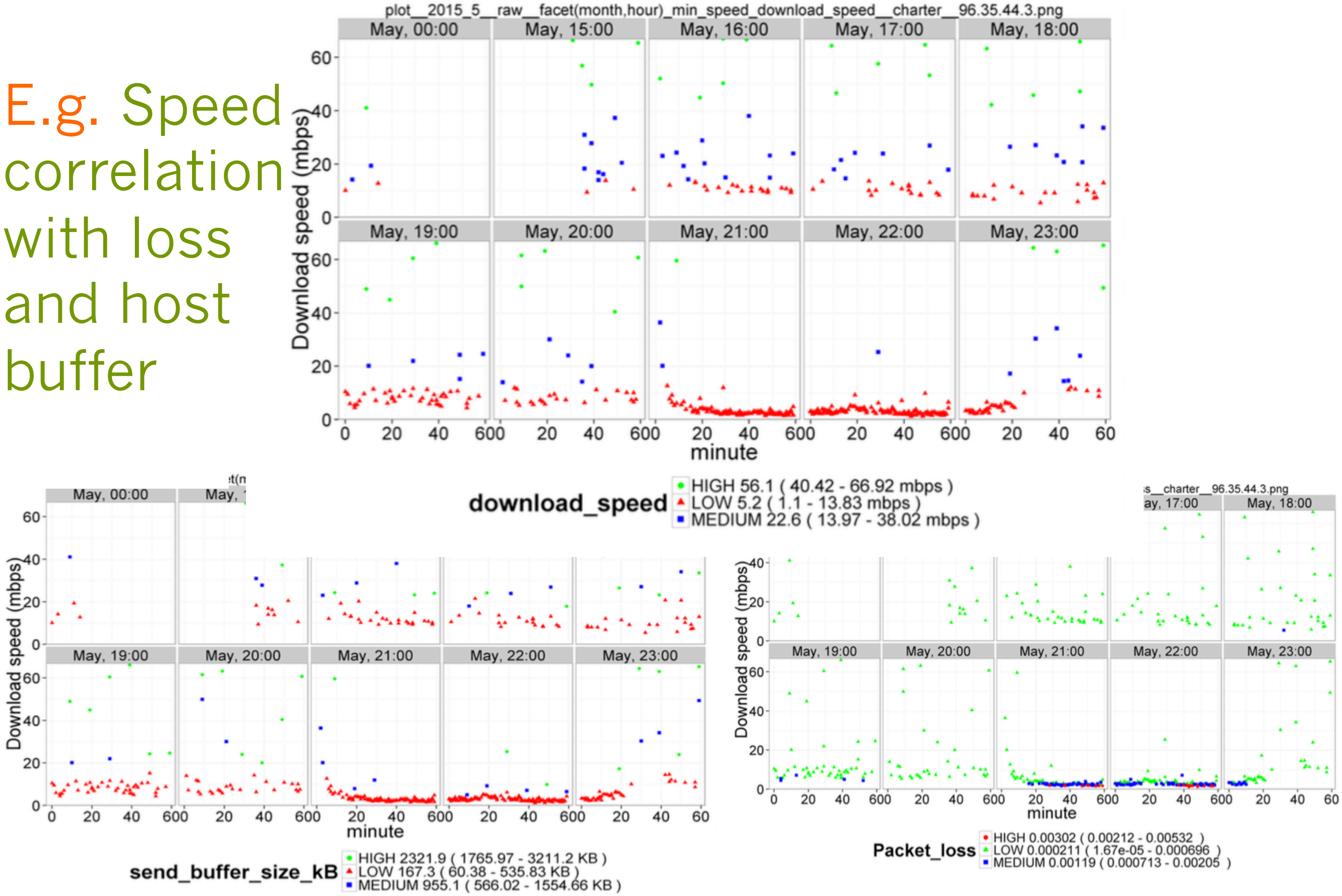
Frequent households' influence

- Various Network Variables effect on Speeds

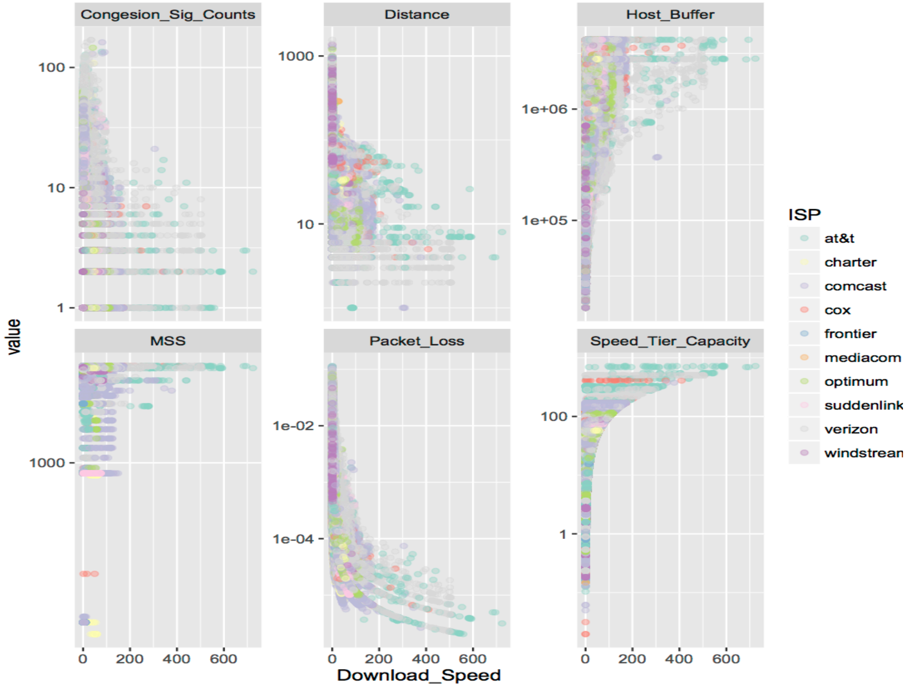
Network Variables(configurations) affects Speed

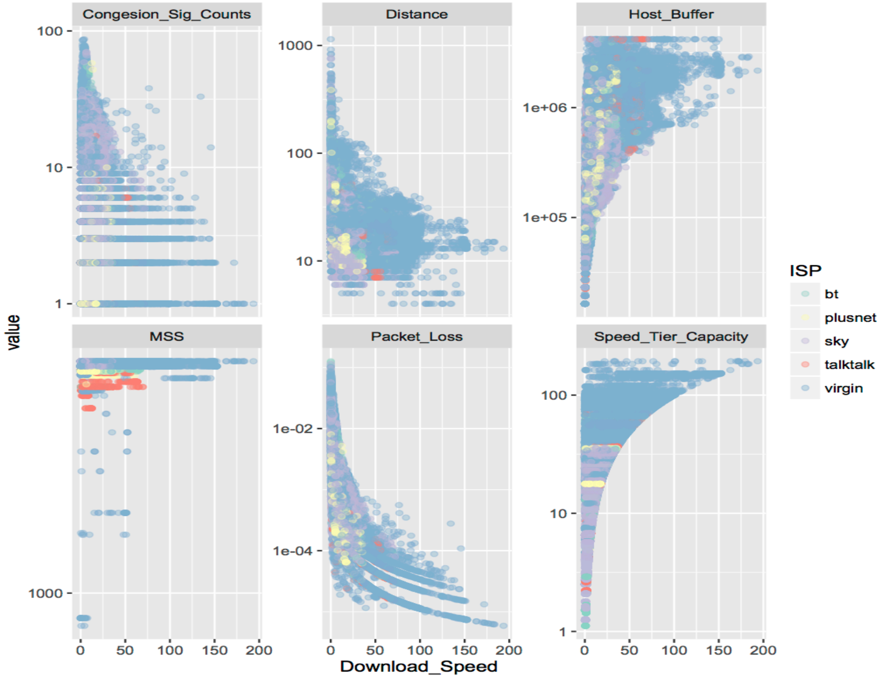
- User subscribed speed tier *estimated
- User host's configuration : TCP send buffer
- TCP MSS
- Client – Server Distance
- IP address Family (IPv4/IPv6):
  - Data not currently available with NDT
- Time of a day

E.g. Speed correlation with loss and host buffer

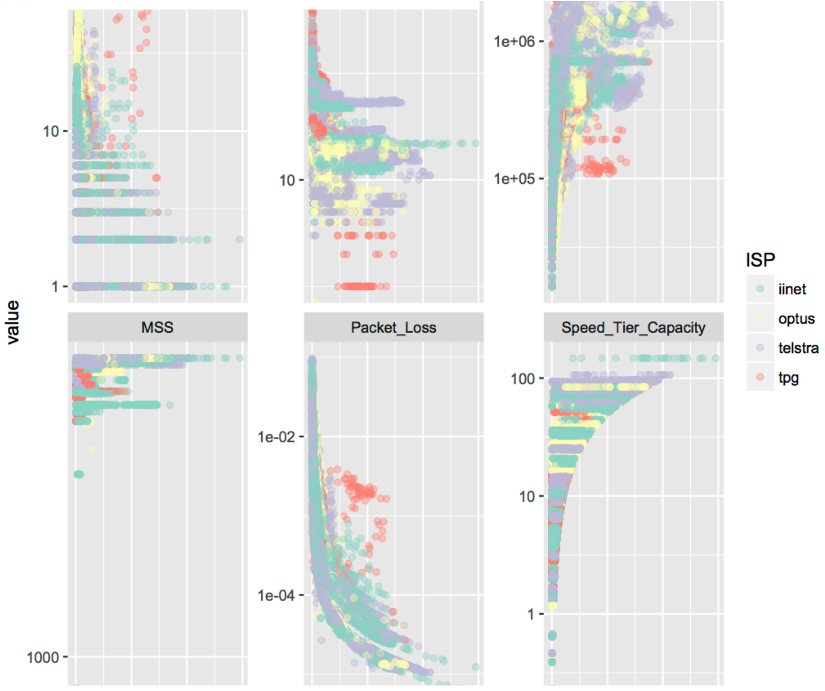


http://104.154.87.31/static/





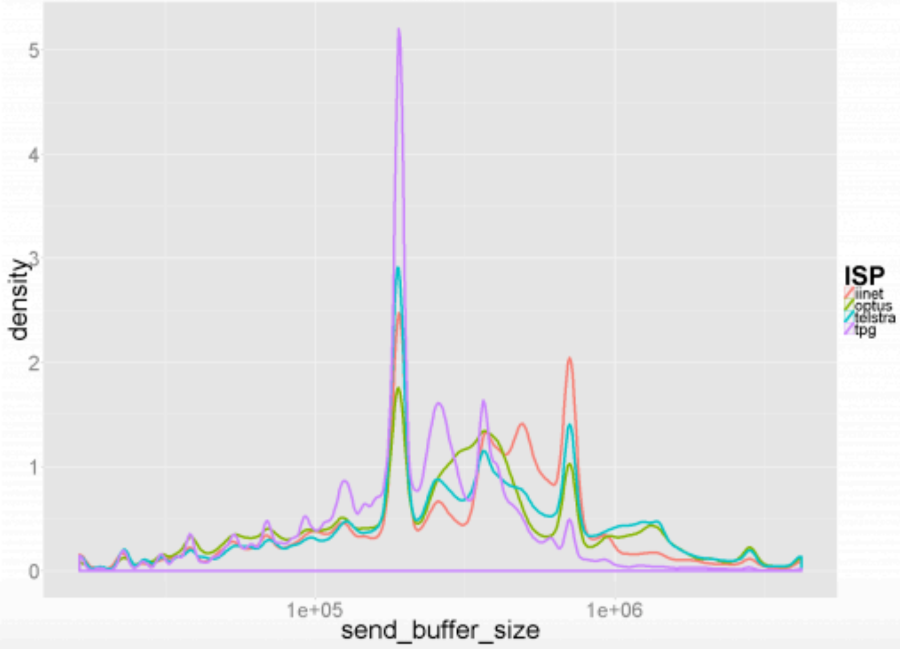
Three counties' speed response to Network Variables

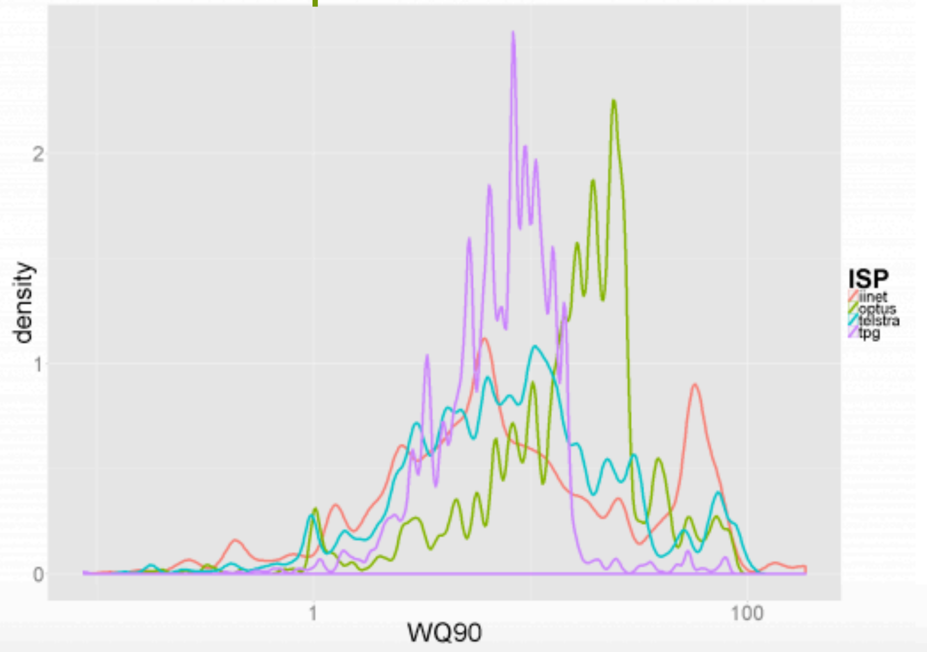


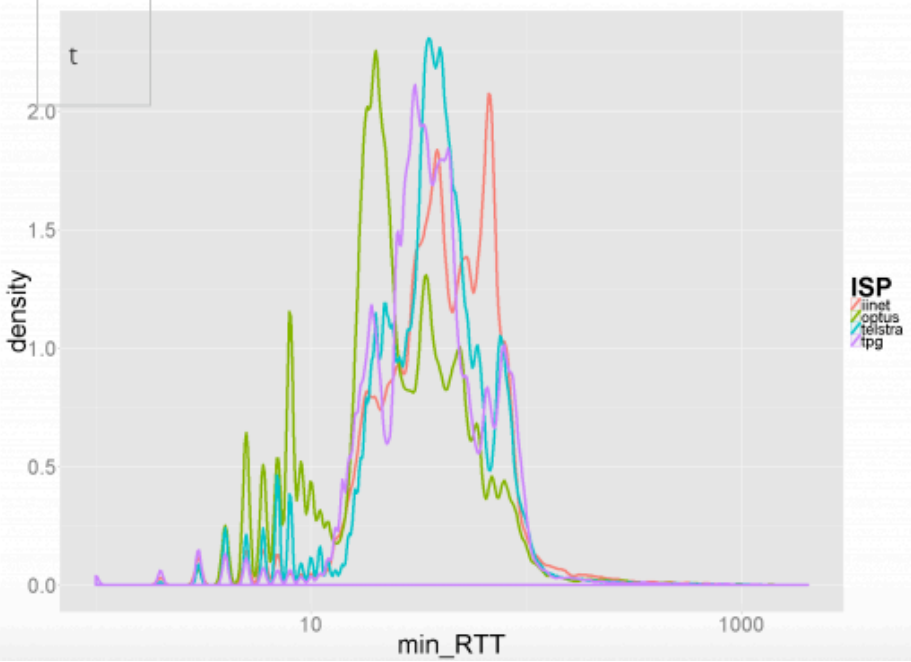
1) Mis-matched Network Variable distribution

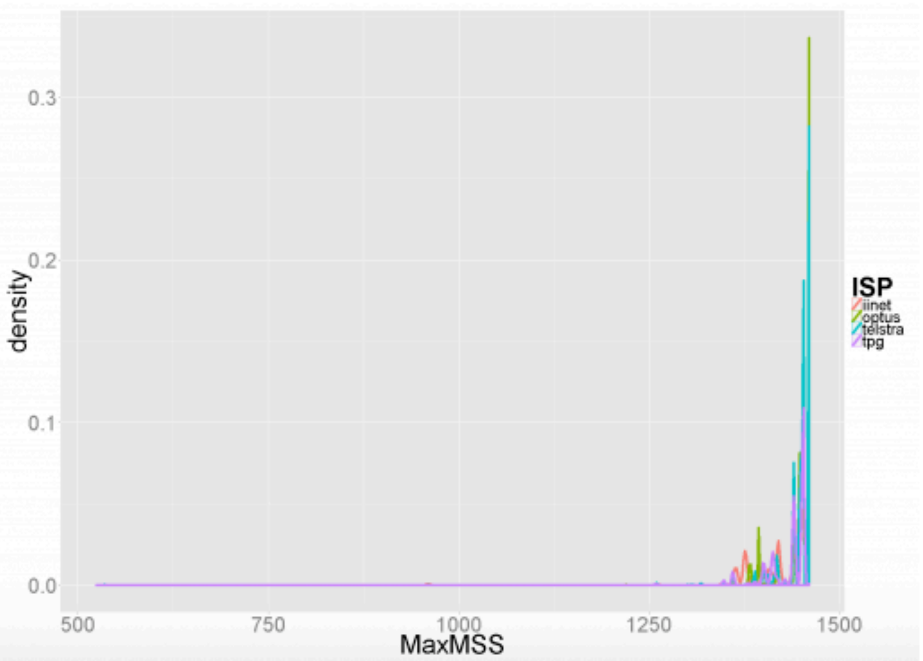
2) Sampling Bias

AU ISPs’ Network Variables Comparison

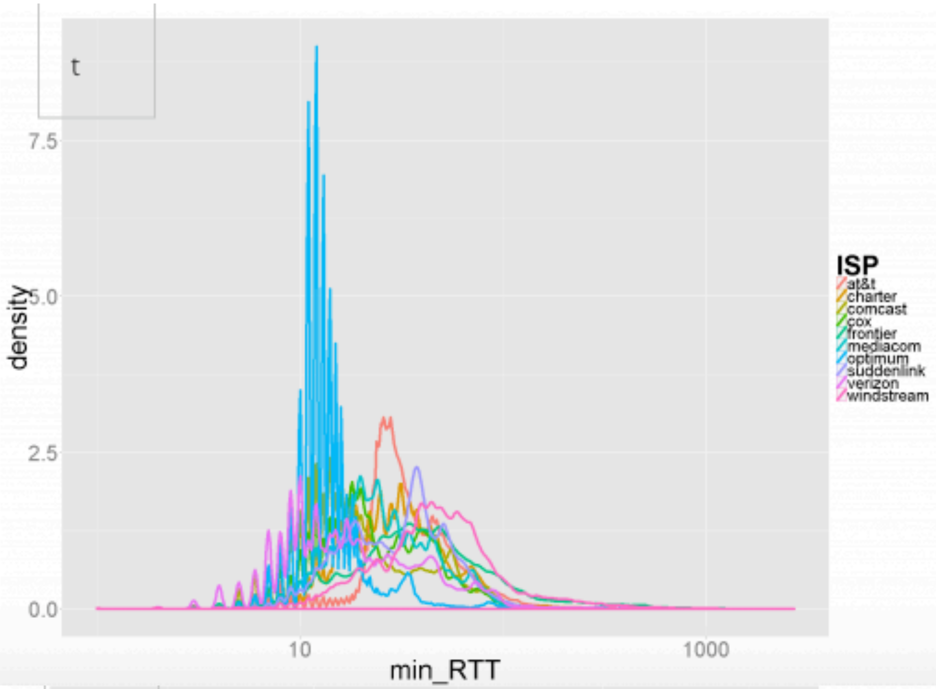


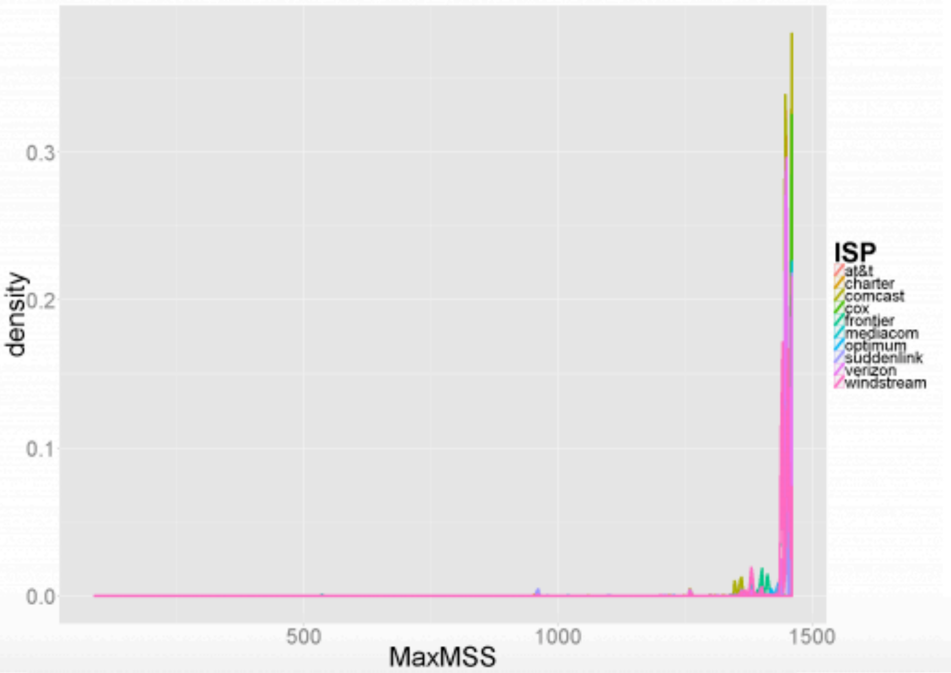


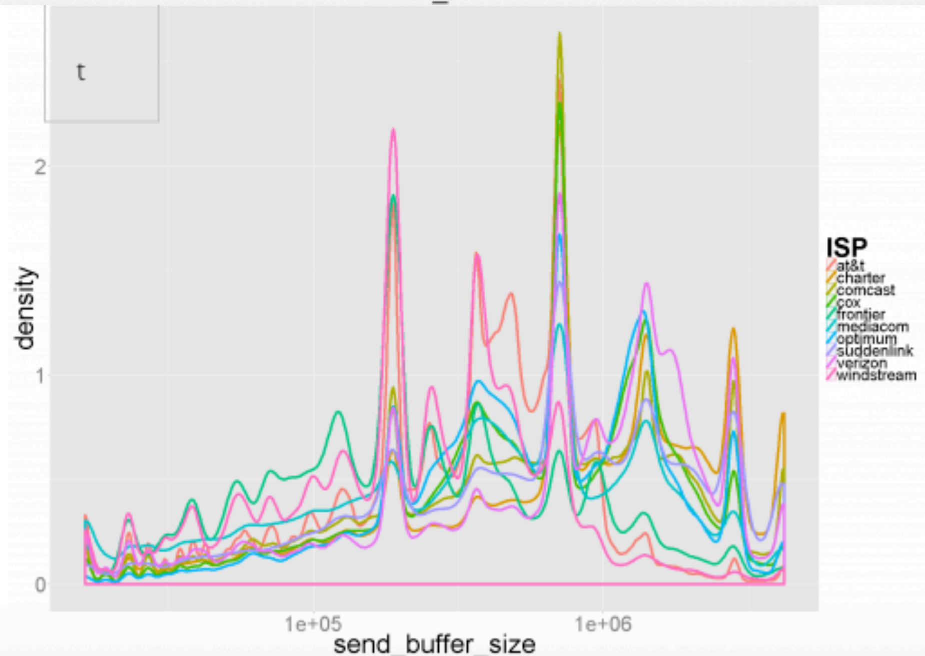


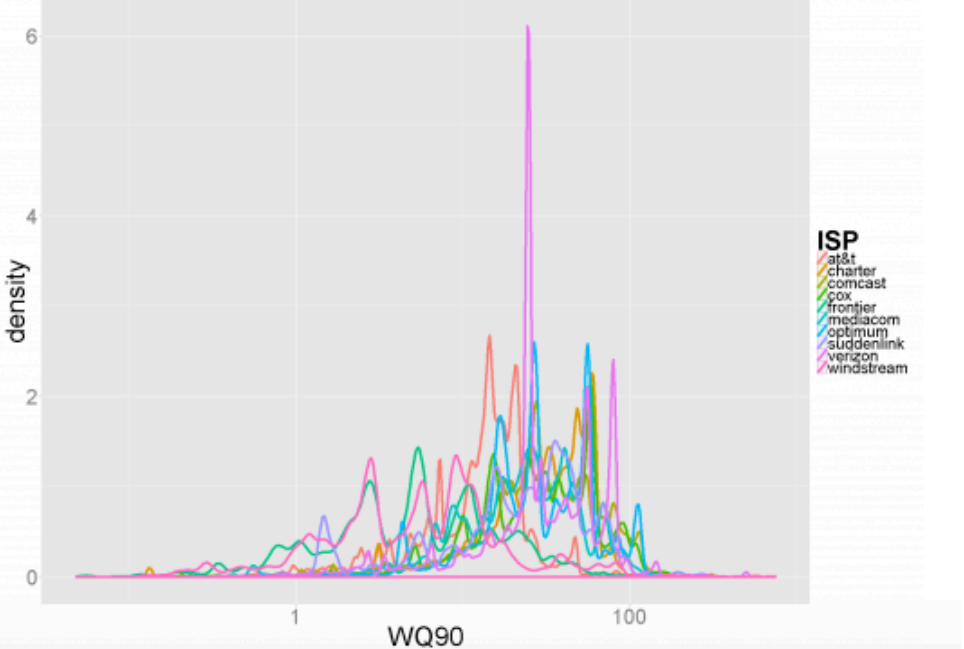


US ISPs' Network Variables Comparison

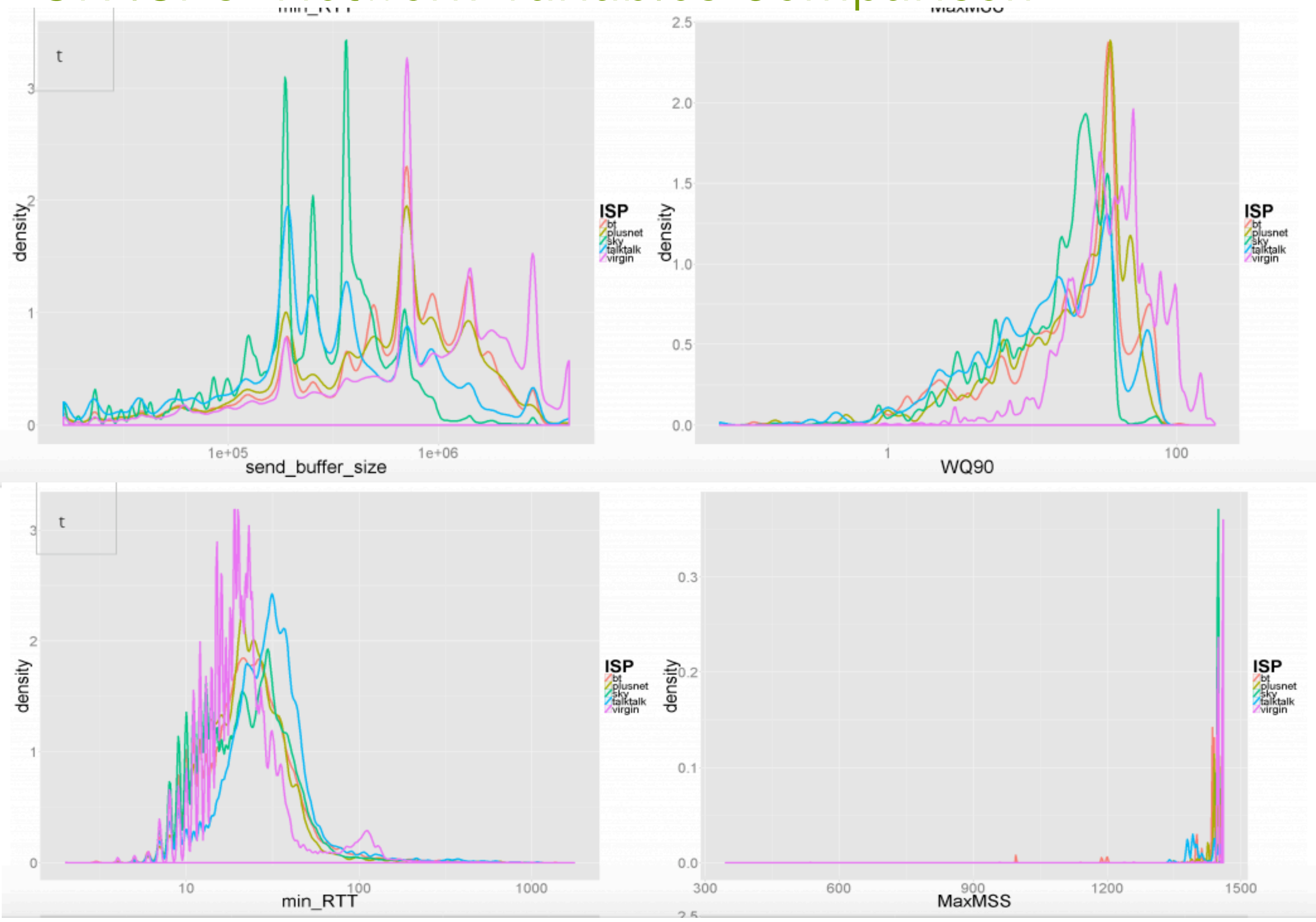






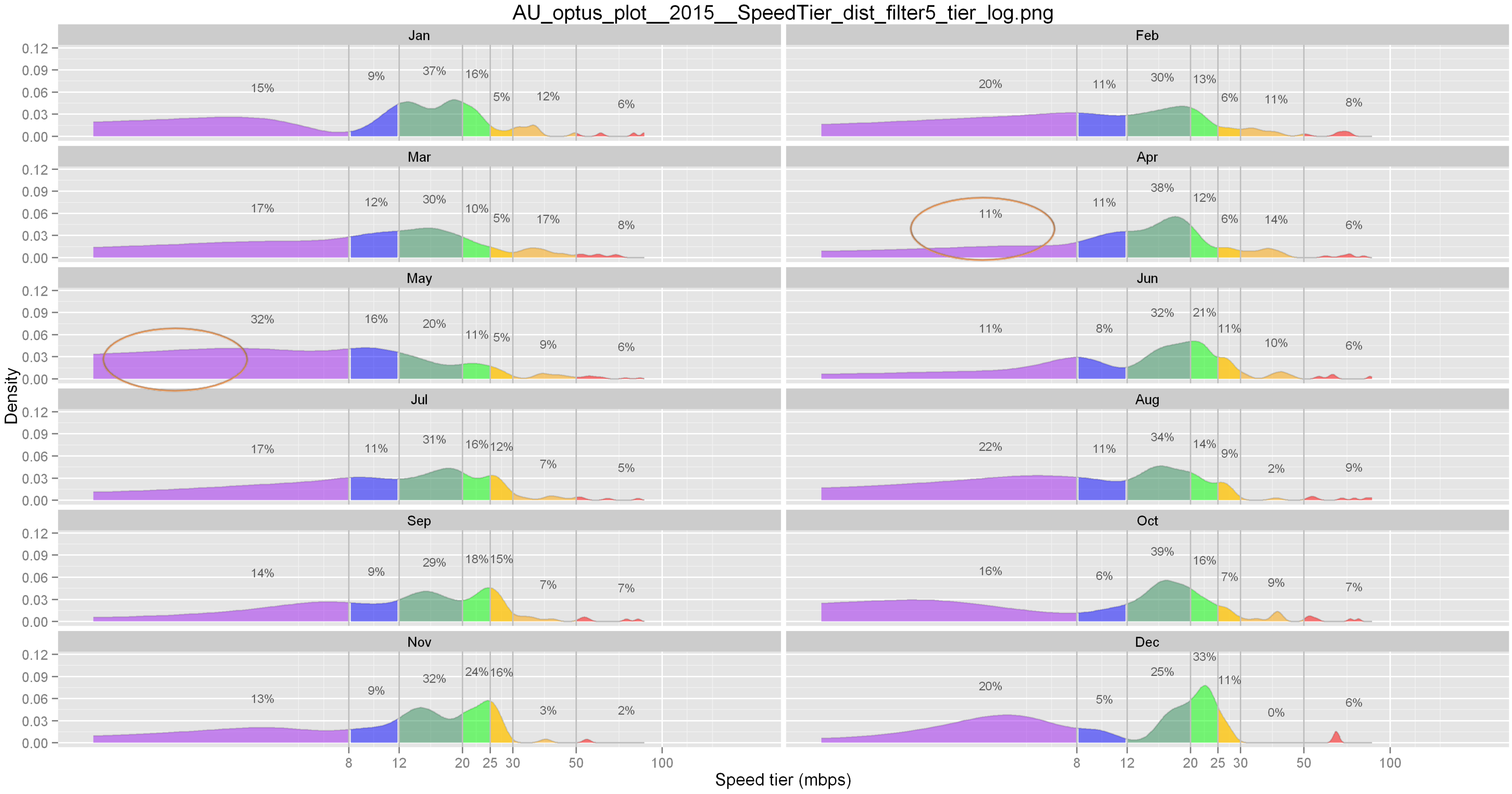


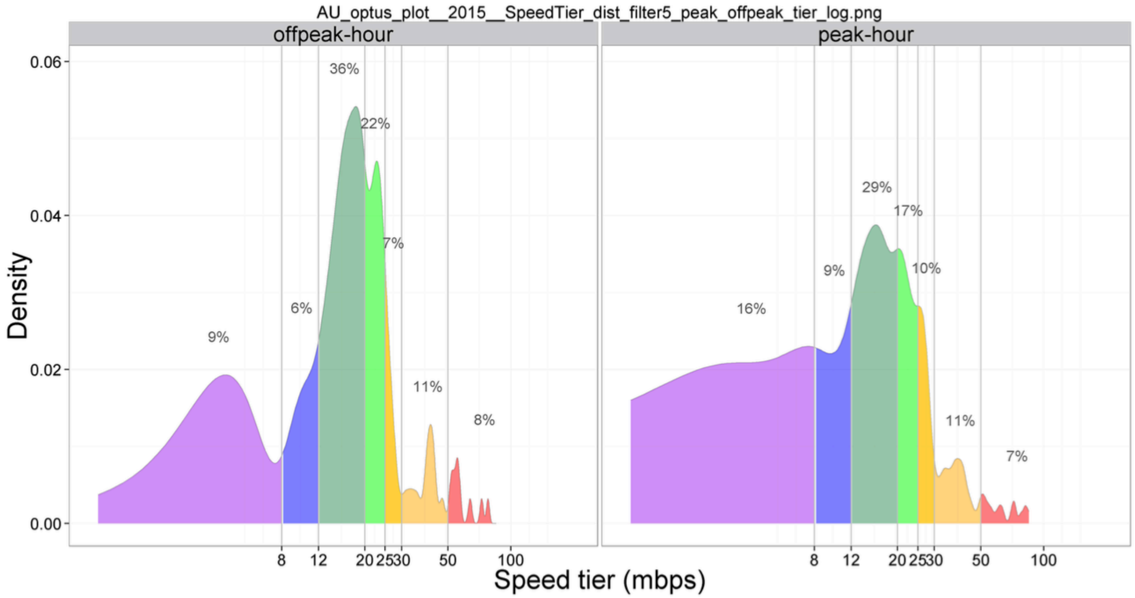
UK ISPs’ Network Variables Comparison



- Sampling Bias

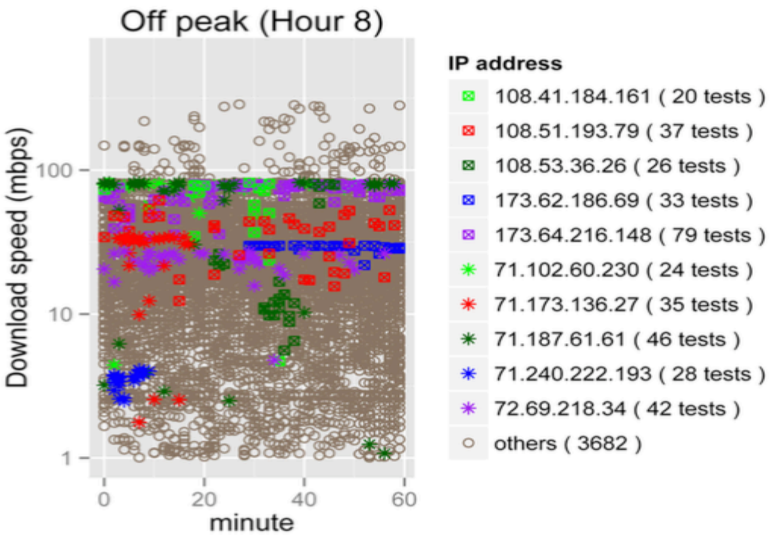
Monthly fluctuation caused by sampling bias

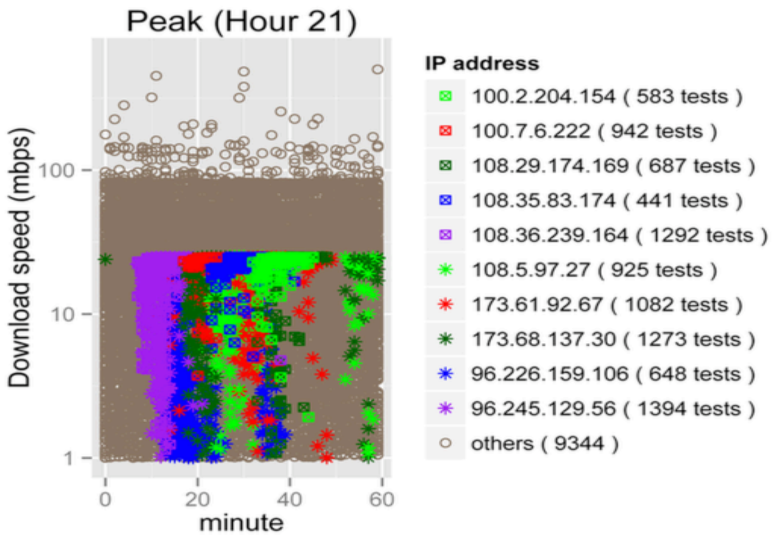




Peak hour v.s None peak hour sampling bias – User/ client behavior pattern

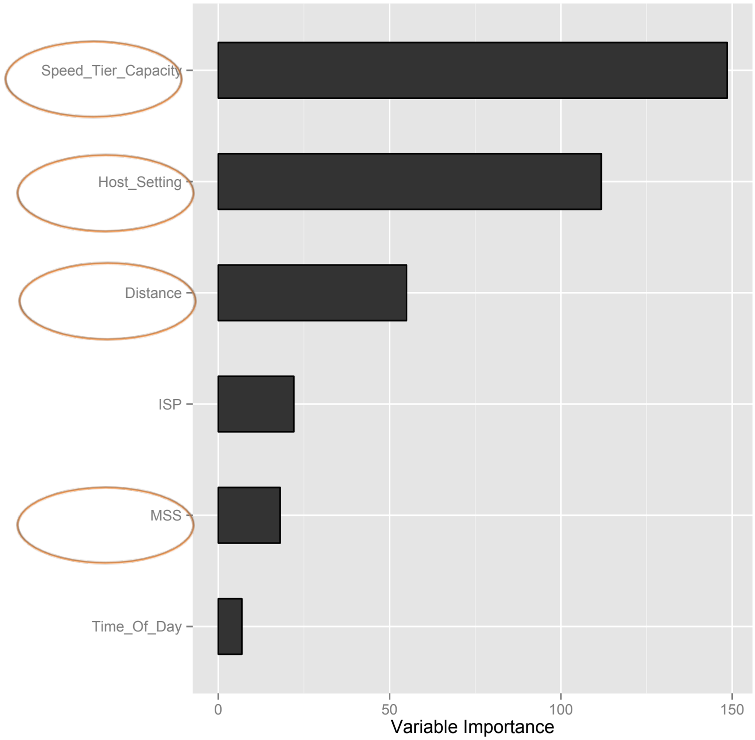
Observed low speed tier users from some ISPs performing More tests during peak hours





Verizon year 2015

ISPs’ Network Variables Importance*



Random Forests for Regression

1. % variance explained: 80.16,
A good modle fit

2. Time of Day has little affect or
Speeds

$$VI^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$$

$\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i) =$ predicted class before permuting

$\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j}) =$ predicted class after permuting X_j

$\mathbf{x}_{i,\pi_j} = \left(x_{i,1}, \ldots, x_{i,j-1}, \; x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p}\right)$

Note: $VI^{(t)}(\mathbf{x}_j) = 0$ by definition, if X_j is not in tree t

The permutation importance

*Variable importance (VIMP)

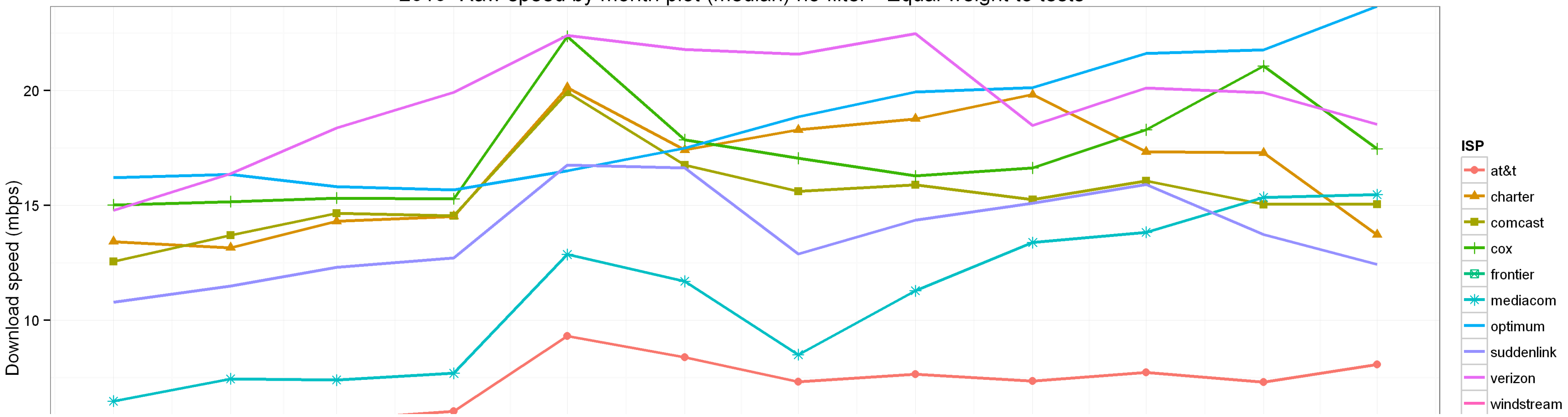
is the difference between OOB prediction error before and after permutation,
a large VIMP value indicates that misspecification detracts from the variable predictive accuracy.

1.De-Biasing:
Equal weight to household

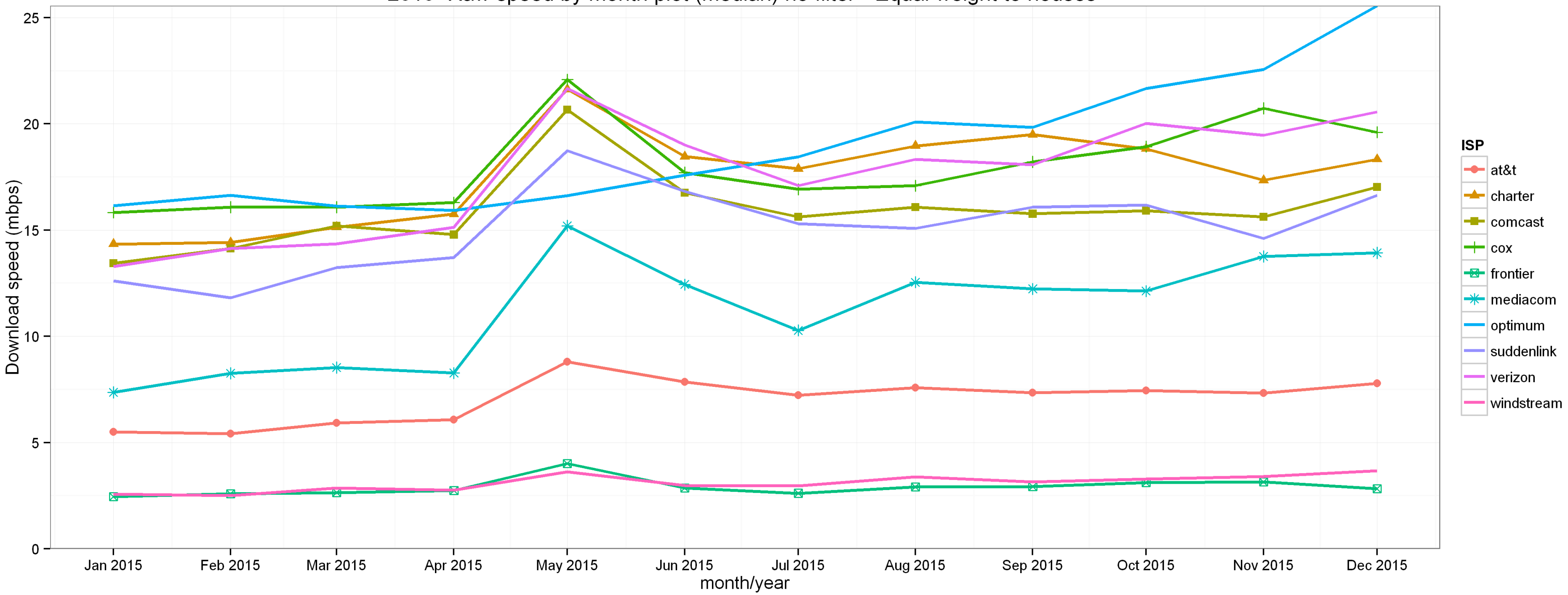
2. Casual Inference model:
Matching tests with similar network variables between two ISPs

Equal weight on household applied on US data

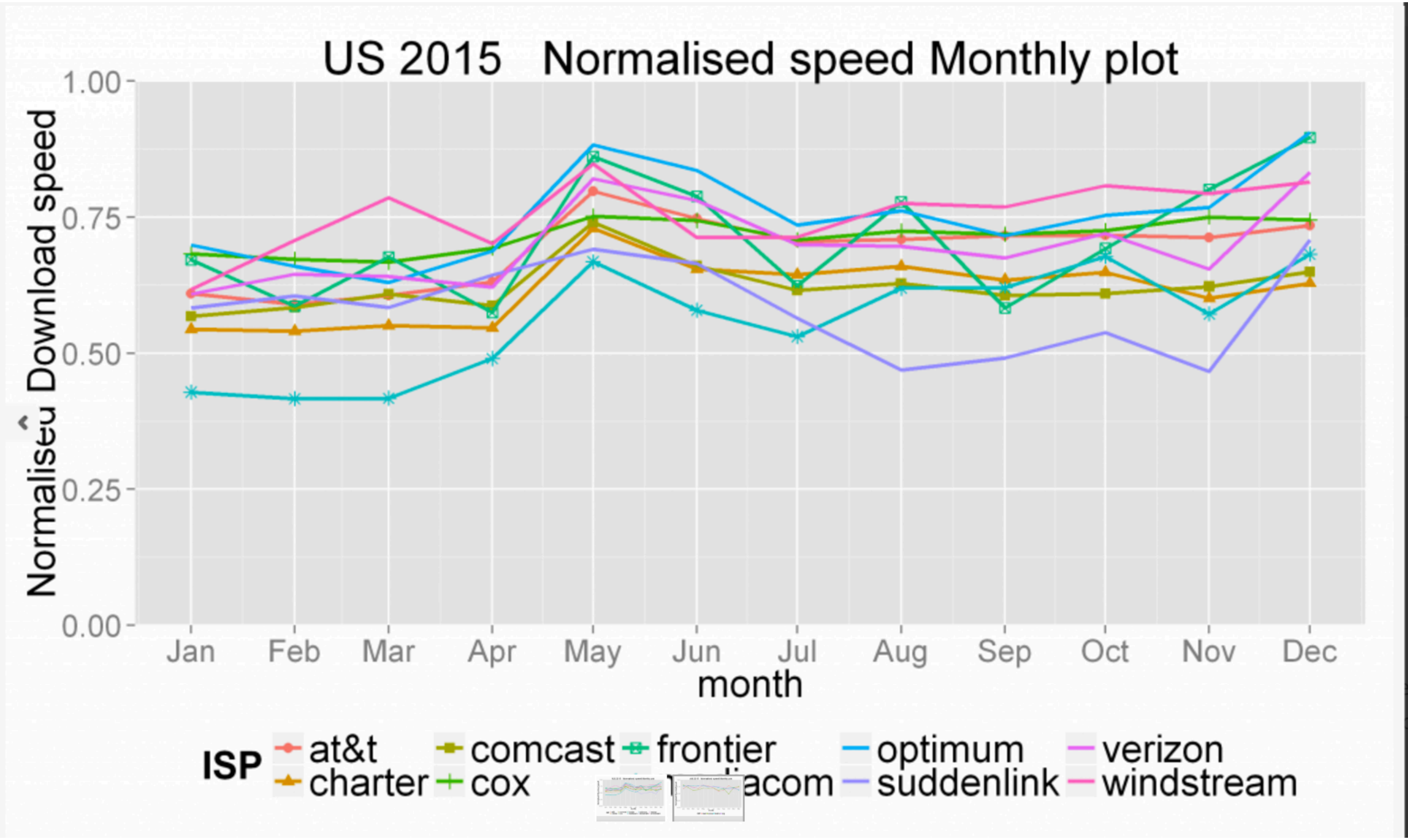
2015 Raw speed by month plot (median) no filter - Equal weight to tests



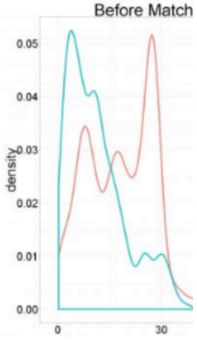
2015 Raw speed by month plot (median) no filter - Equal weight to houses

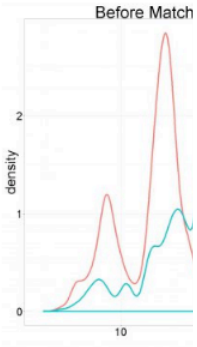


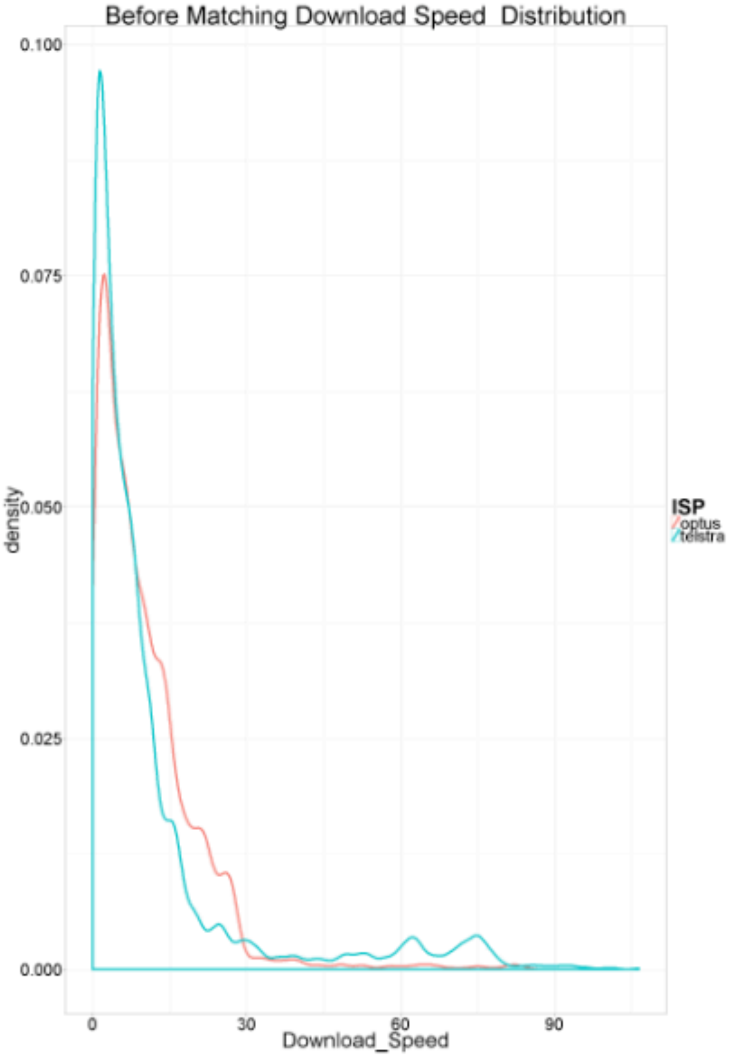
Normalisation: another aggregated performance indicator

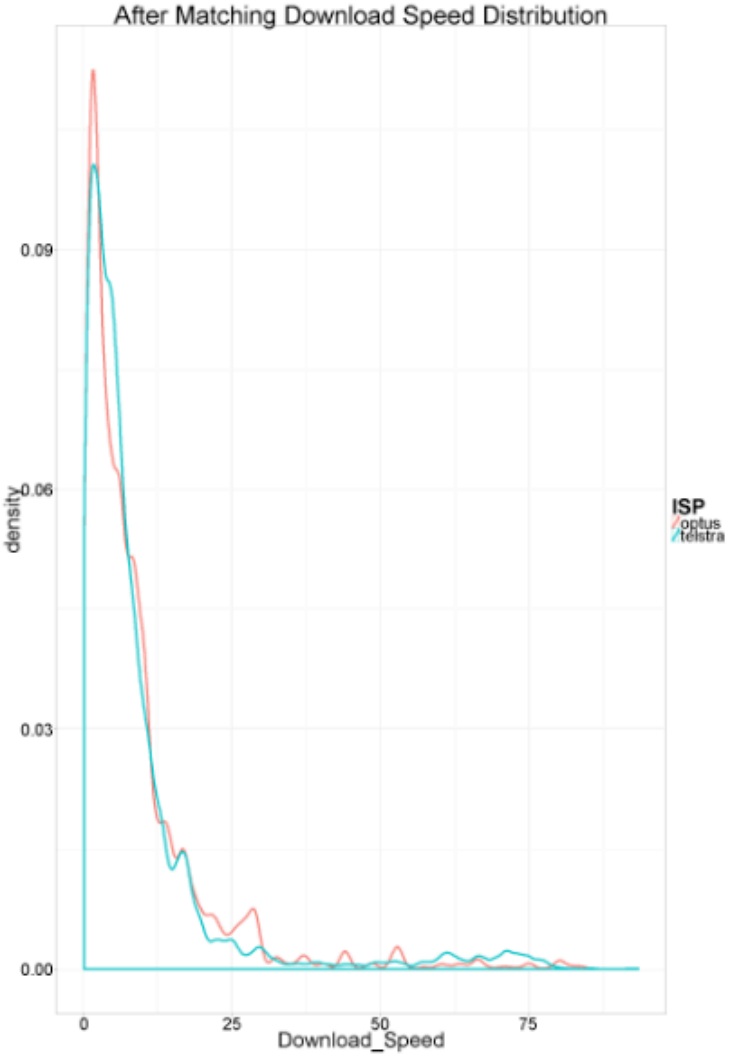


Before and After Matching: A pair of AU ISP

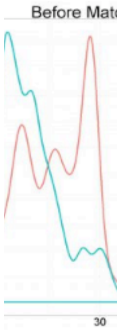


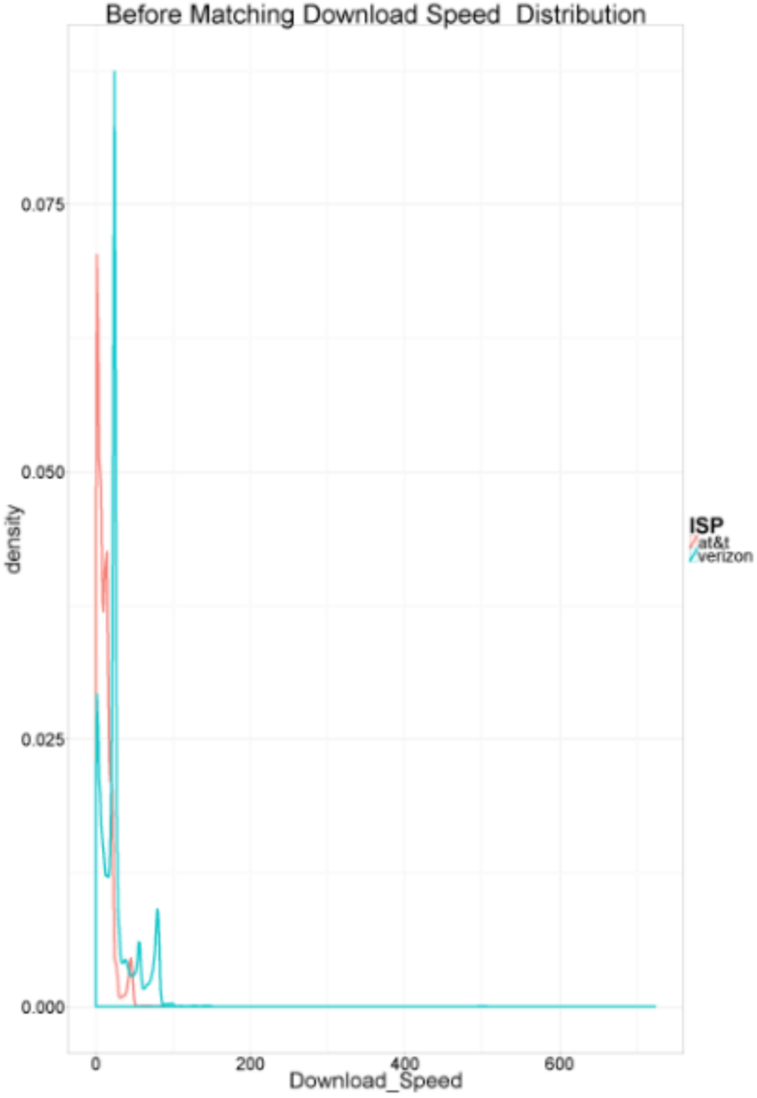


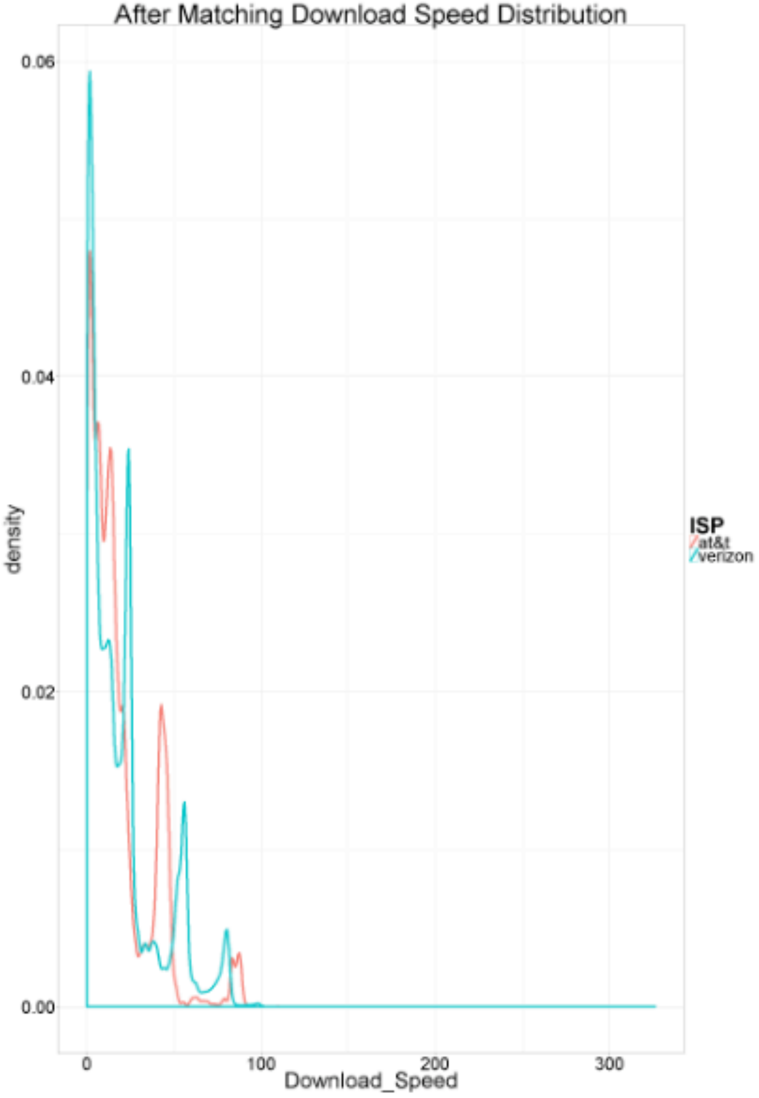


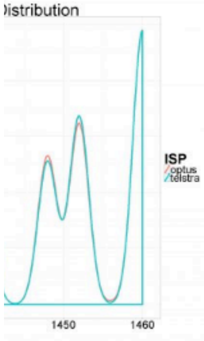


Before and After Matching: A pair of US ISP

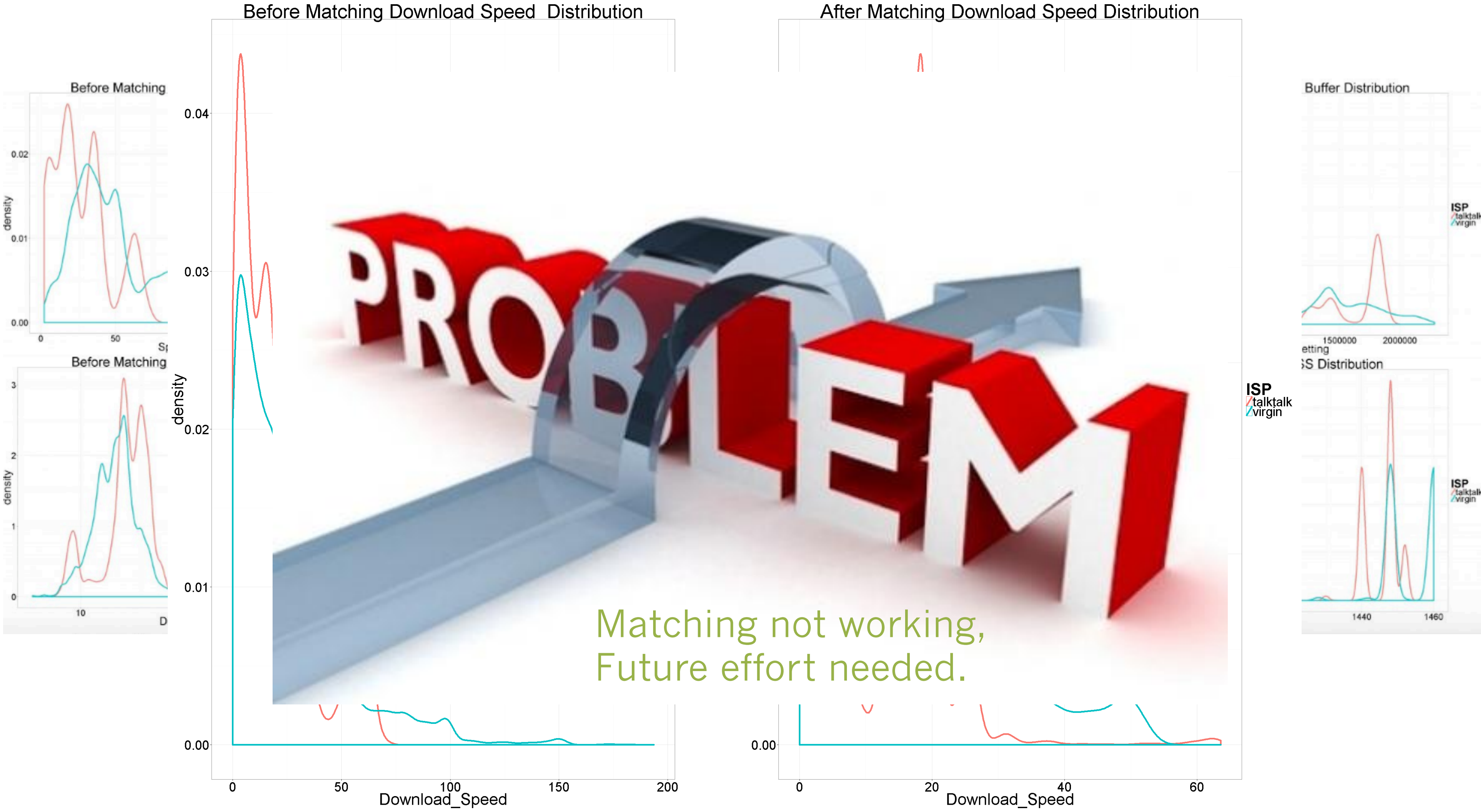




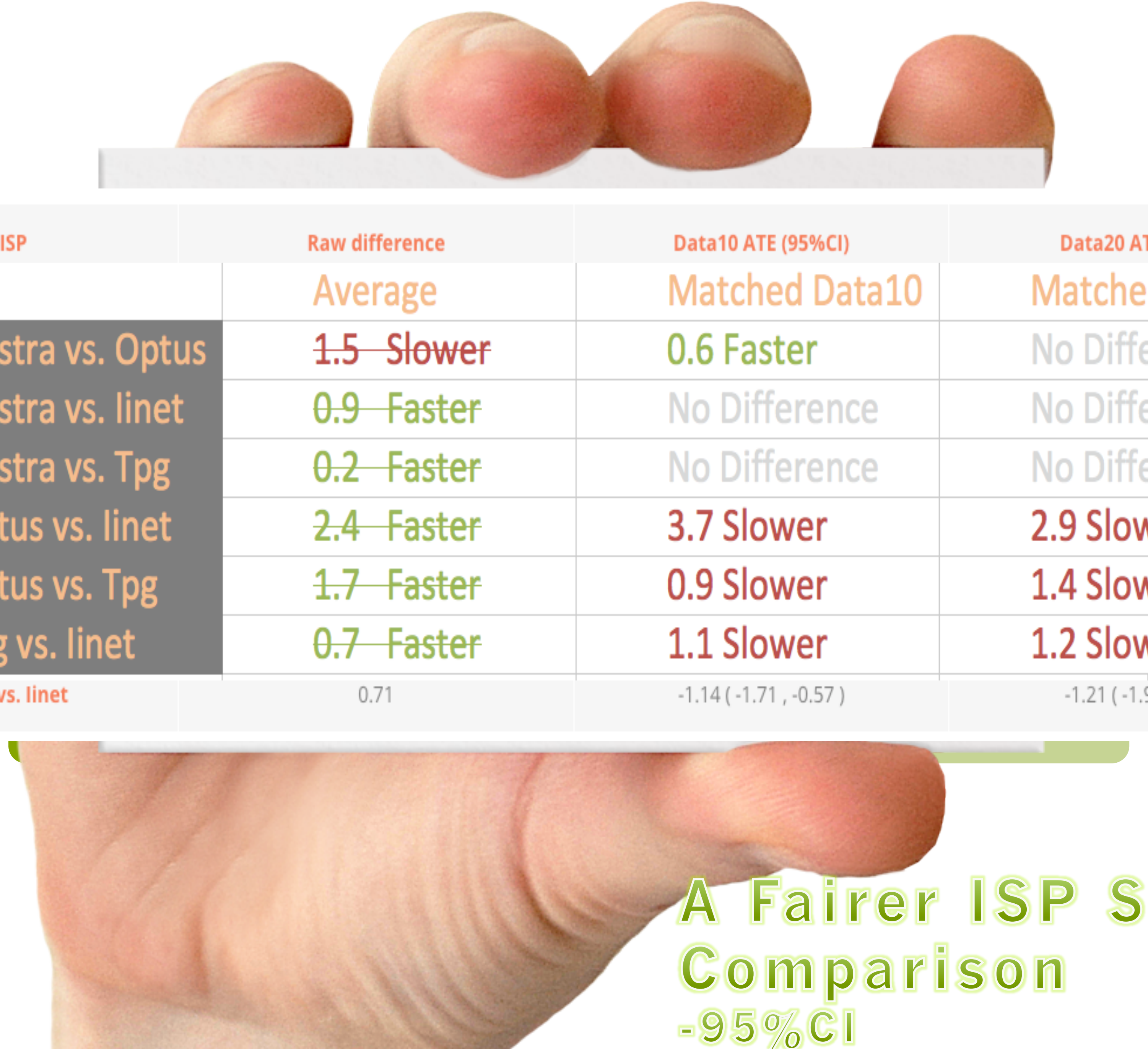




Before and After Matching: A pair of UK ISP



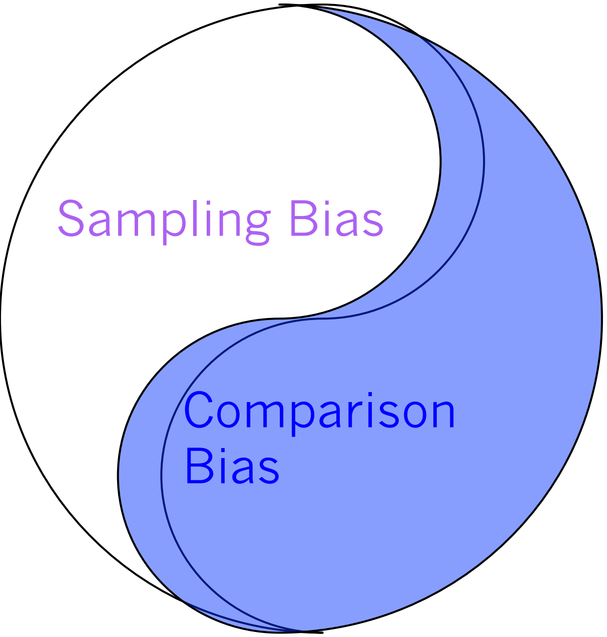
Matching not working,
Future effort needed.



ISP	Raw difference	Data10 ATE (95%CI)	Data20 ATE (95%CI)
ISP	Average	Matched Data10	Matched Data20
Telstra vs. Optus	~~1.5 Slower~~	0.6 Faster	No Difference
Telstra vs. Iinet	~~0.9 Faster~~	No Difference	No Difference
Telstra vs. Tpg	~~0.2 Faster~~	No Difference	No Difference
Optus vs. Iinet	~~2.4 Faster~~	3.7 Slower	2.9 Slower
Optus vs. Tpg	~~1.7 Faster~~	0.9 Slower	1.4 Slower
Tpg vs. Iinet	~~0.7 Faster~~	1.1 Slower	1.2 Slower
Tpg vs. Iinet	0.71	-1.14 (-1.71 , -0.57)	-1.21 (-1.99 , -0.43)

A Fairer ISP Speed Comparison

-95%CI



Insight gained:

1) Aggregated average number is impacted by sampling bias that caused by user behaviors.
2) Sampling Bias can cause ISP monthly fluctuation.
3) Sampling Bias opens doors to gaming the system. Warning

Insight gained:

Comparison Bias

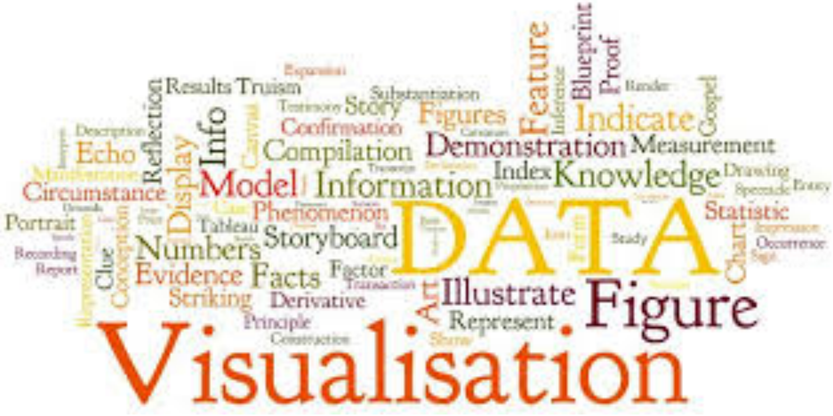
1) Simple Averaging Lead to misleading ISP rankings to Internet users

2) More sophiscated methods such as Statistical tools & Machine Learning shall be used, for a fairer ISP performance comparison.

RESULTS: A fairer comparison often shows a different ISP ranking – sometimes even opposite to ranking based on average

SOLUTIONS

a) Equal weight to household than test
b) Casual Inference Model, Bayes addictive Tree Model (future work)



http://104.154.87.31/static/

Questions
Answers

THANK YOU