NetFlow-based bandwidth estimation in IP networks

Rodrigo Alarcón-Reyes
Department of Electrical and Computer Engineering
McGill University

March 13, 2012

Outline

1. Introduction to NetFlow
2. Definition of the bandwitdh estimation problem
3. Proposed Solution
   - Notation
   - Assumptions
   - Solution
4. Study of the uplink/downlink ratio from real traces
5. Future work

Introduction

- NetFlow is an application protocol developed by Cisco for collecting and reporting IP traffic information in networks

- Information is exported on a per-flow basis

- A flow is defined by the following 7-tuple:
  - Source IP address
  - Destination IP address
  - Source L4 port
  - Destination L4 port
  - IP protocol
  - Input interface
  - IP ToS

NetFlow operation

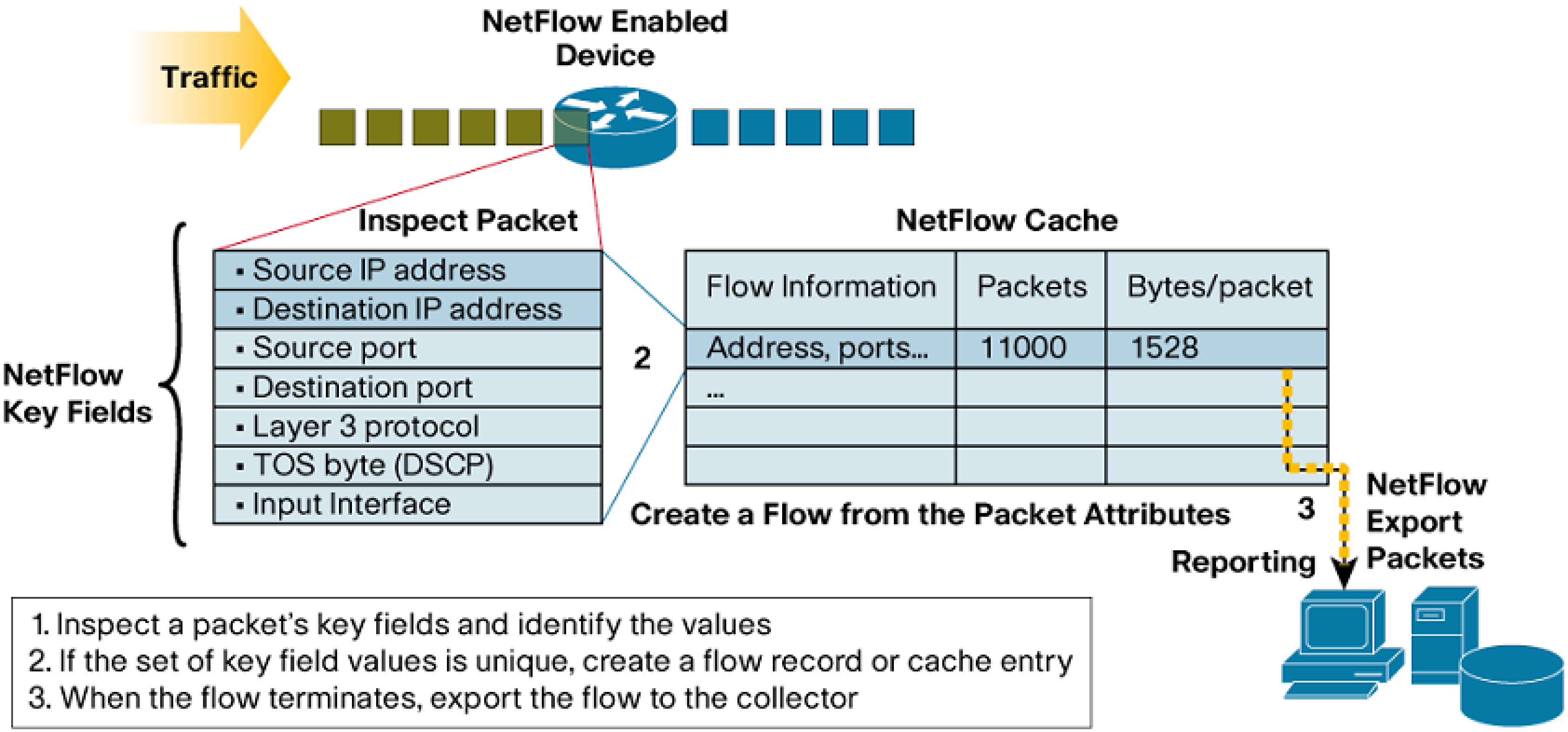
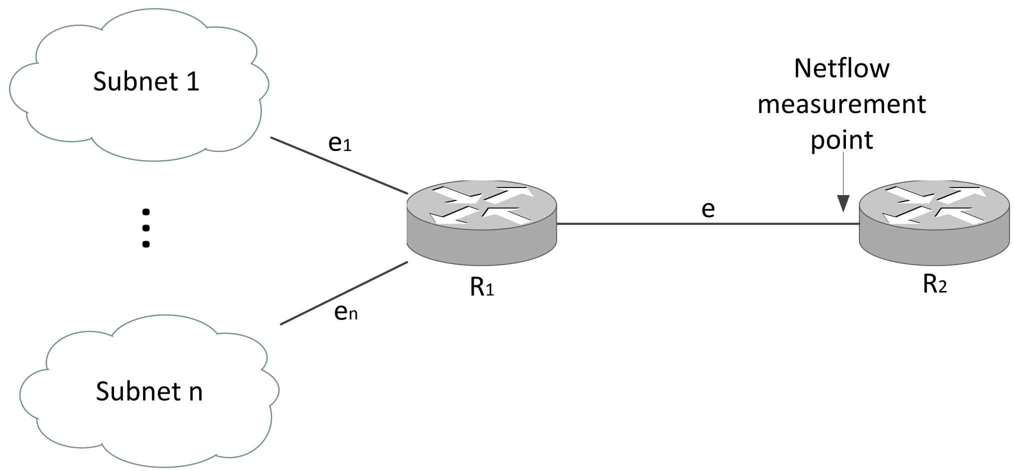


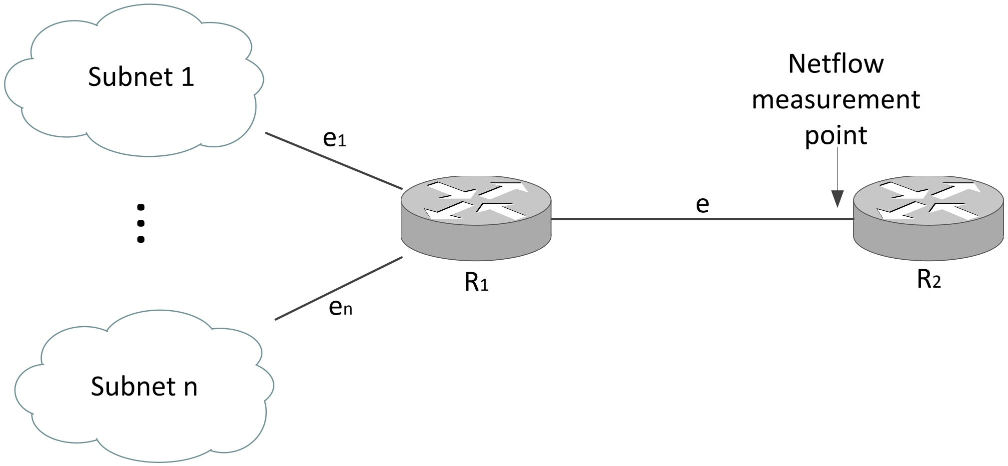
Illustration from «Introduction to Cisco IOS® Flexible NetFlow», Technology White Paper. Cisco Systems Inc., 2008.

Scenario



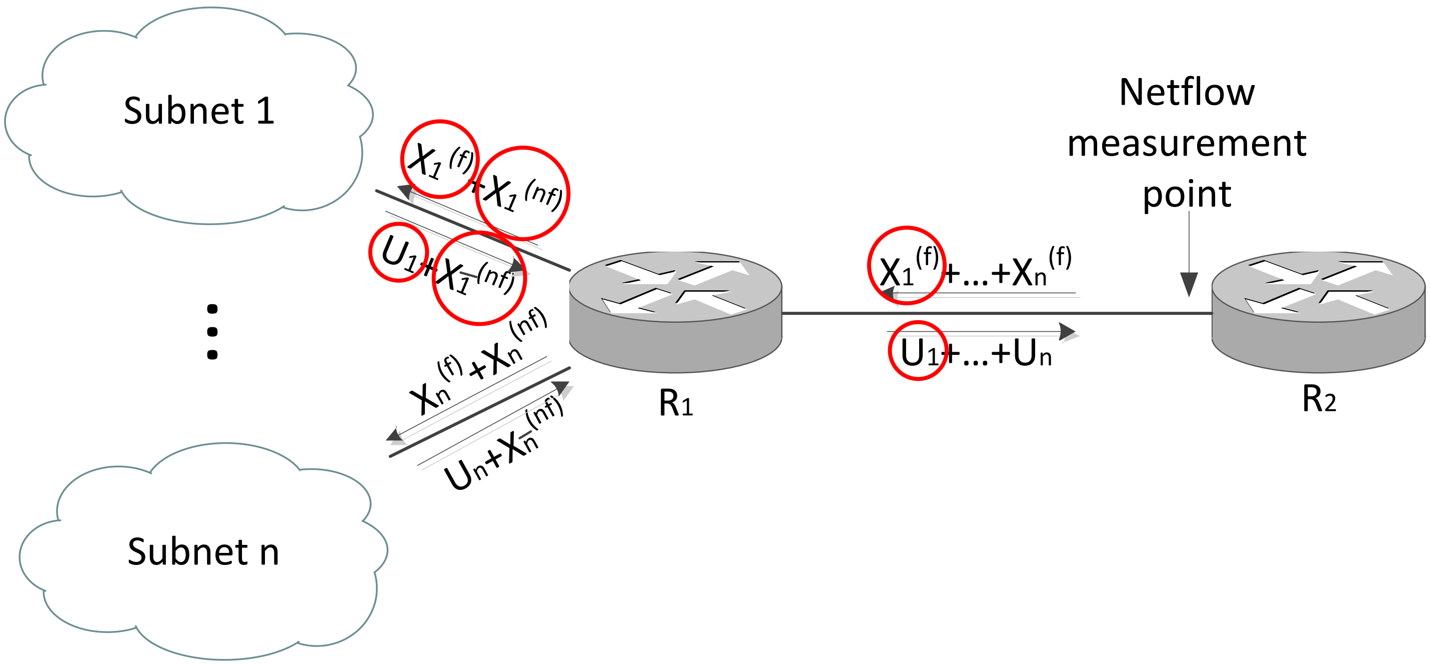
- IP network with tree topology
- Router R_1 connected to n subnets through links $e_1, \ldots, e_n$
- R_1 also connected to router R_2
- NetFlow agent configured in R_2 for reporting incoming and outgoing traffic from R_1

Bandwidth Estimation Problem



- During a specific time interval (t_1, t_2) we want to estimate the average downlink bandwidth of traffic over links e_1 to e_n
  - Based on the NetFlow measurements in R_2
  - Without any packet injection

Proposed Solution: Notation



- $X_i^{(f)}$: Portion of downlink traffic over link e_i that is forwarded from R_2
- $X_i^{(nf)}$: Portion of downlink traffic over link e_i not forwarded from R_2
- U_i : Portion of uplink traffic over link e_i that is forwarded to R_2
- $X_i^{(nf)}$: Portion of uplink traffic over link e_i not forwarded to R_2

Proposed Solution: Assumptions

1. For each link e_i the uplink traffic is directly proportional to the downlink traffic, with proportionality constant λ_i

   e.g. $U_1 + X_{\bar{1}}^{(nf)} = \lambda_1 \, (X_1^{(f)} + X_1^{(nf)})$

2. The uplink traffic $X_{\bar{i}}^{(nf)}$ is distributed as downlink traffic in the remaining subnets, in a per-link proportion $w_{k,i} \in (0,1)$

   e.g. $X_{\bar{1}}^{(nf)} = w_{2,1}X_2^{(nf)} + \dots + w_{n,1}X_n^{(nf)}$

Proposed Solution

- With the previous asumptions, we can derive the vectorial relation

$$\underline{X}^{(nf)} = (\boldsymbol{W} - \boldsymbol{\Lambda})^{-1}\, \boldsymbol{\Lambda}\, \underline{X}^{(f)} - (\boldsymbol{W} - \boldsymbol{\Lambda})^{-1}\, \underline{U} \qquad \boldsymbol{W}_{ij} = w_{j,i}\,,\ \boldsymbol{W}_{ii} = 0$$

$$\boxed{\underline{X}^{(nf)} = \boldsymbol{A}\, \underline{X}^{(f)} - \boldsymbol{B}\, \underline{U}} \qquad \boldsymbol{\Lambda}_{ii} = \lambda_i$$

- Using linear regression it is possible to obtain estimates of the matrices A and B
  - A combination of SNMP polling and NetFlow records can be using for the training period
  - After the training period, only NetFlow records are required

- By knowing the traffic $\underline{X}^{(f)}$ and $\underline{X}^{(nf)}$, the downlink bandwidth can be finally estimated as the average over a desired interval

Study of the uplink/downlink ratio from real traces

- Anonymized dataset from real traces
  - Nearly 10 MM NetFlow records collected during a 3-hours window at an aggregation point
  - No information was provided about the topology of the network
  - Only the top 15 sender and top 15 receiver subnetworks were considered for the analysis

- For each considered subnetwork:
  - Records were grouped in 10-minutes intervals
  - *Partial* Uplink/Downlink ratio $U_i / X_i^{(f)}$ was computed for each interval, based on total bytes count per interval

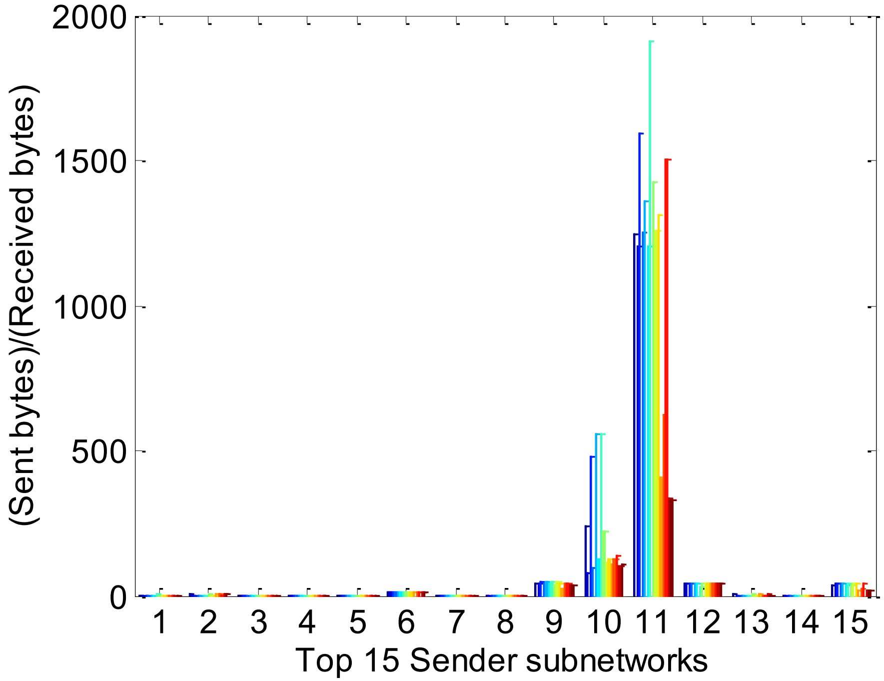
Top 15 Sender Subnetworks

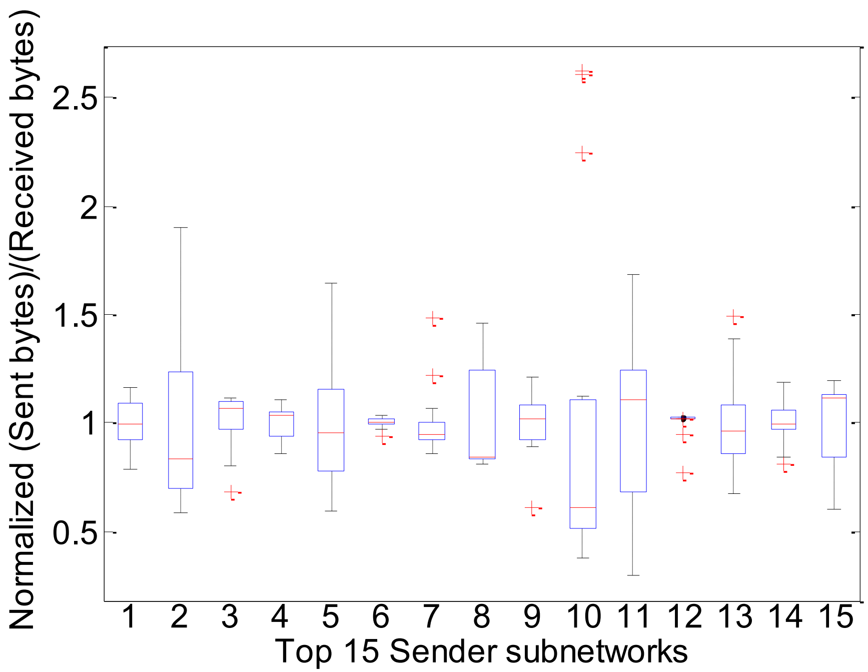
#	Anonymized Subnet (/16)	Total Sent (MB)	Total Received (MB)
1	132.207.0.0	3298.0	647.1
2	231.166.0.0	1591.4	291.8
3	230.194.0.0	1445.0	3350.2
4	140.164.0.0	1045.7	2390.4
5	137.32.0.0	1032.7	864.6
6	241.124.0.0	961.6	72.7
7	230.205.0.0	756.7	377.1
8	162.227.0.0	658.0	1506.8
9	122.171.0.0	569.0	12.6
10	243.92.0.0	491.7	3.9
11	230.68.0.0	373.7	0.4
12	245.242.0.0	300.8	6.6
13	165.195.0.0	298.4	55.5
14	123.33.0.0	256.2	253.6
15	6.180.0.0	242.0	5.9

Top 15 Receiver Subnetworks

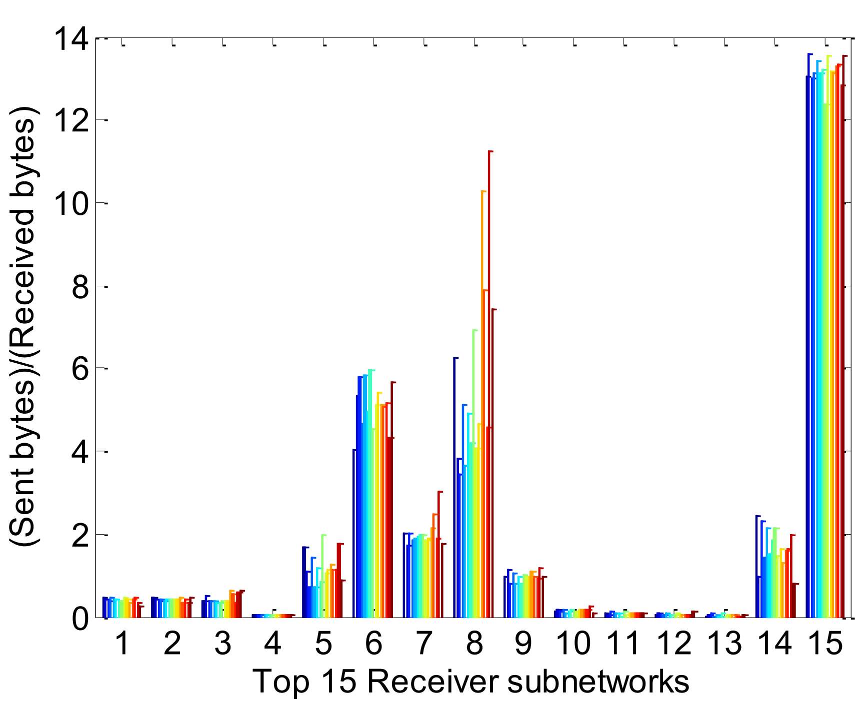
#	Anonymized Subnet (/16)	Total Sent (MB)	Total Received (MB)
1	230.194.0.0	1445.0	3350.2
2	140.164.0.0	1045.7	2390.4
3	162.227.0.0	658.0	1506.8
4	137.51.0.0	92.1	1151.7
5	137.32.0.0	1032.7	864.6
6	137.207.0.0	3298.0	647.1
7	230.205.0.0	756.7	377.1
8	231.166.0.0	1591.4	291.8
9	123.33.0.0	256.2	253.6
10	224.92.0.0	28.6	171.9
11	125.1.0.0	13.1	137.7
12	116.171.0.0	8.5	119.1
13	241.35.0.0	4.7	113.2
14	137.253.0.0	135.5	87.1
15	241.124.0.0	961.6	72.7

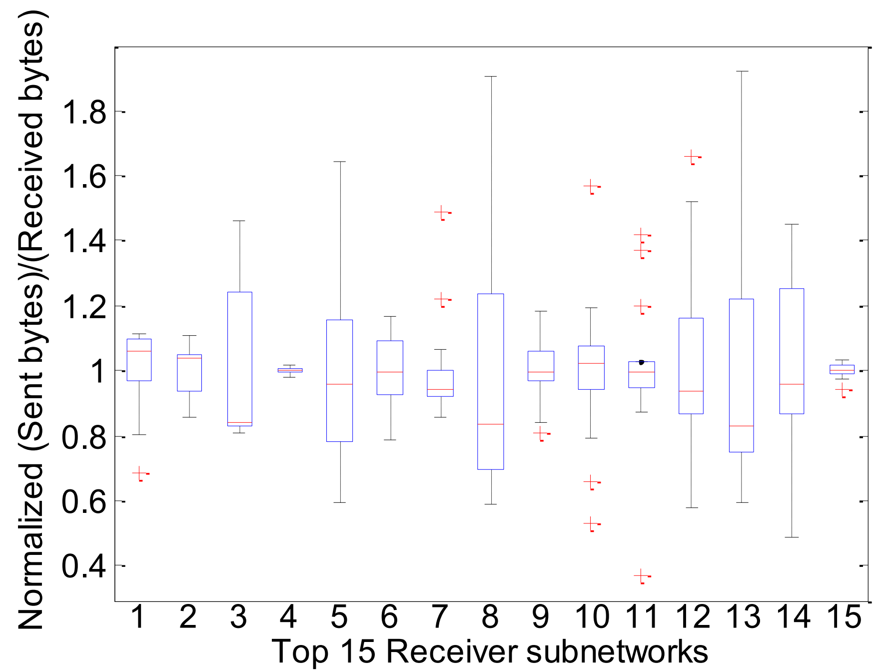
Partial ratio for top 15 senders (using time intervals of 10 minutes)





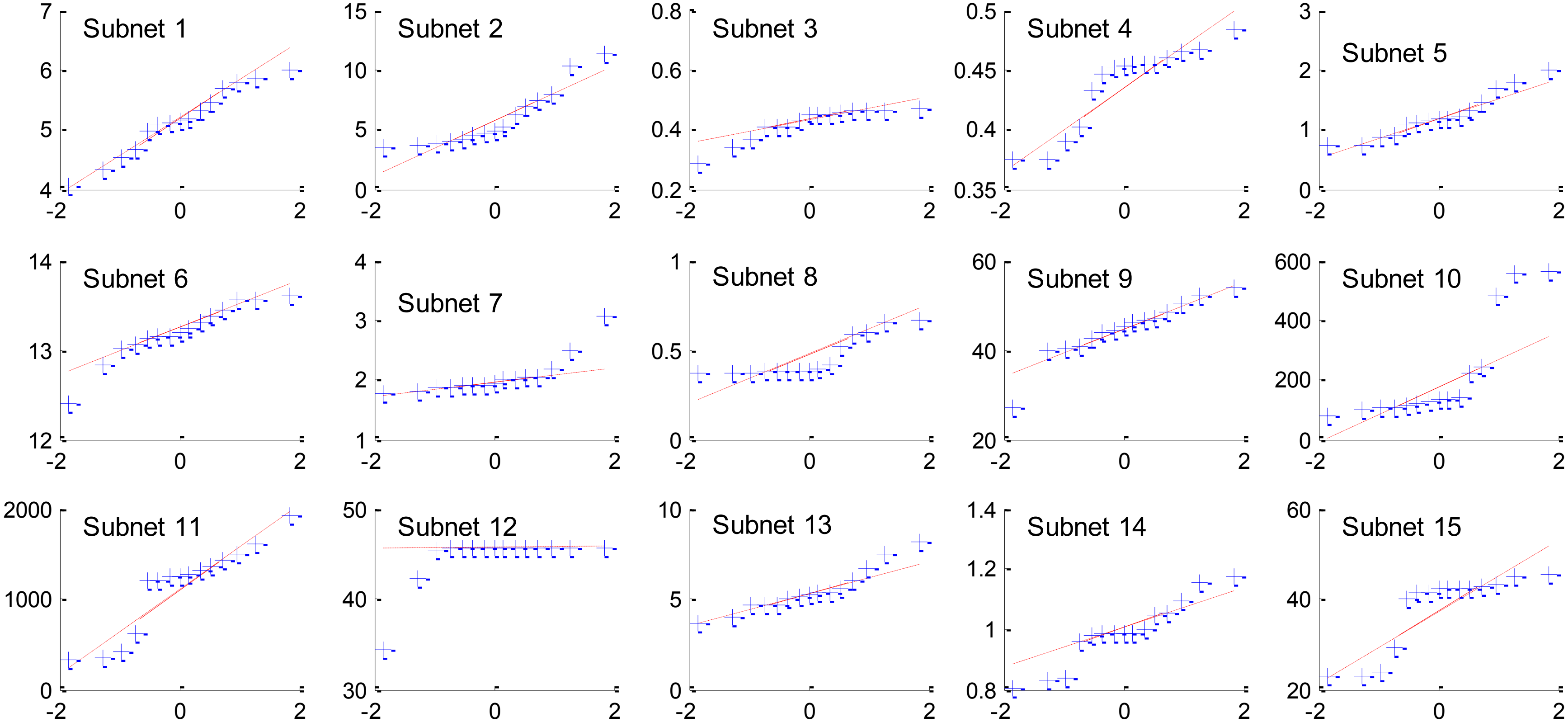
Partial ratio for top 15 receivers (using time intervals of 10 minutes)





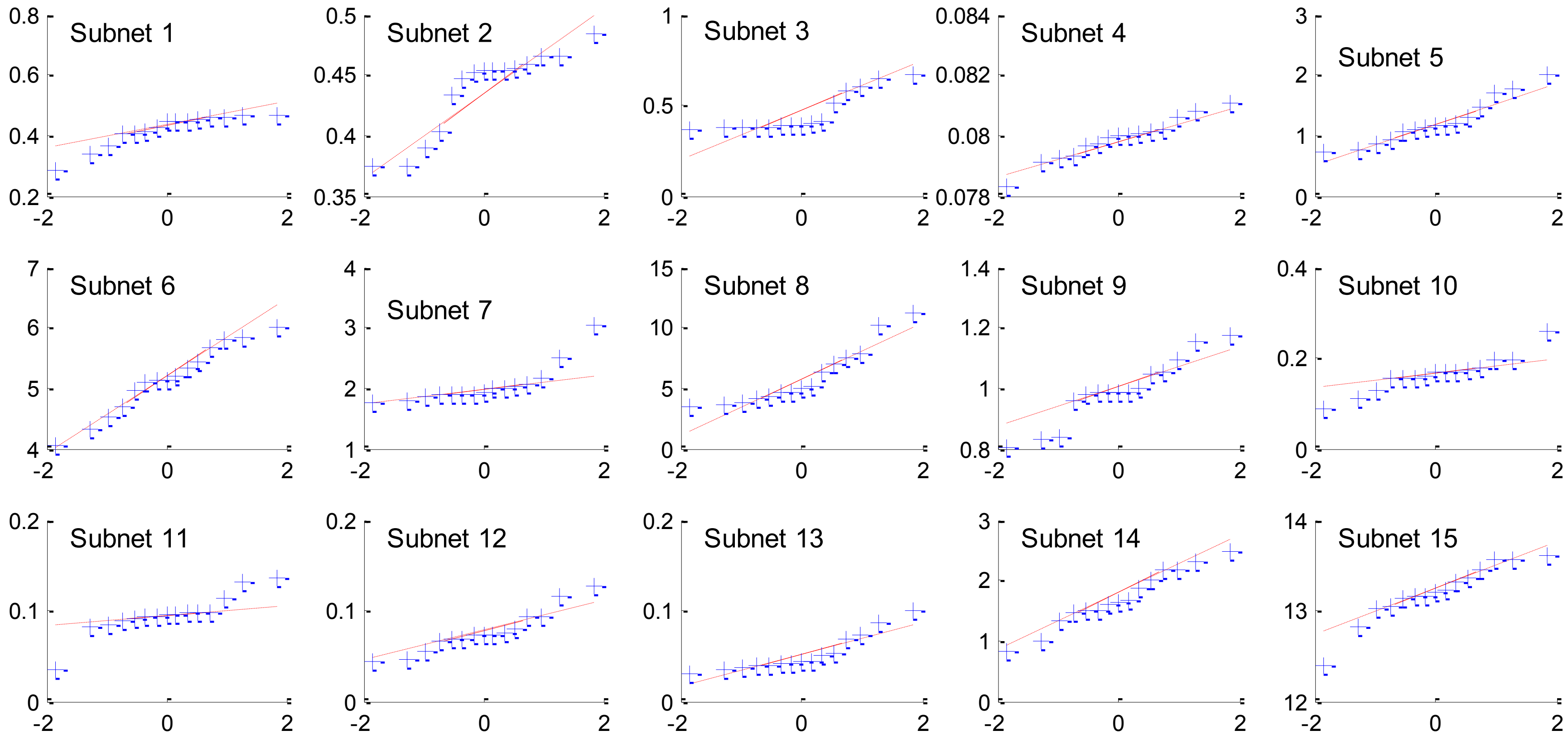
- Is it possible to bound the expected mean partial ratio?
  - Yes, at least if the values are normally distributed

Q-Q analysis: Top 15 senders



x-axes: Normal quantiles
y-axes: Partial rate values

Q-Q analysis: Top 15 receivers



x-axes: Normal quantiles
y-axes: Partial rate values

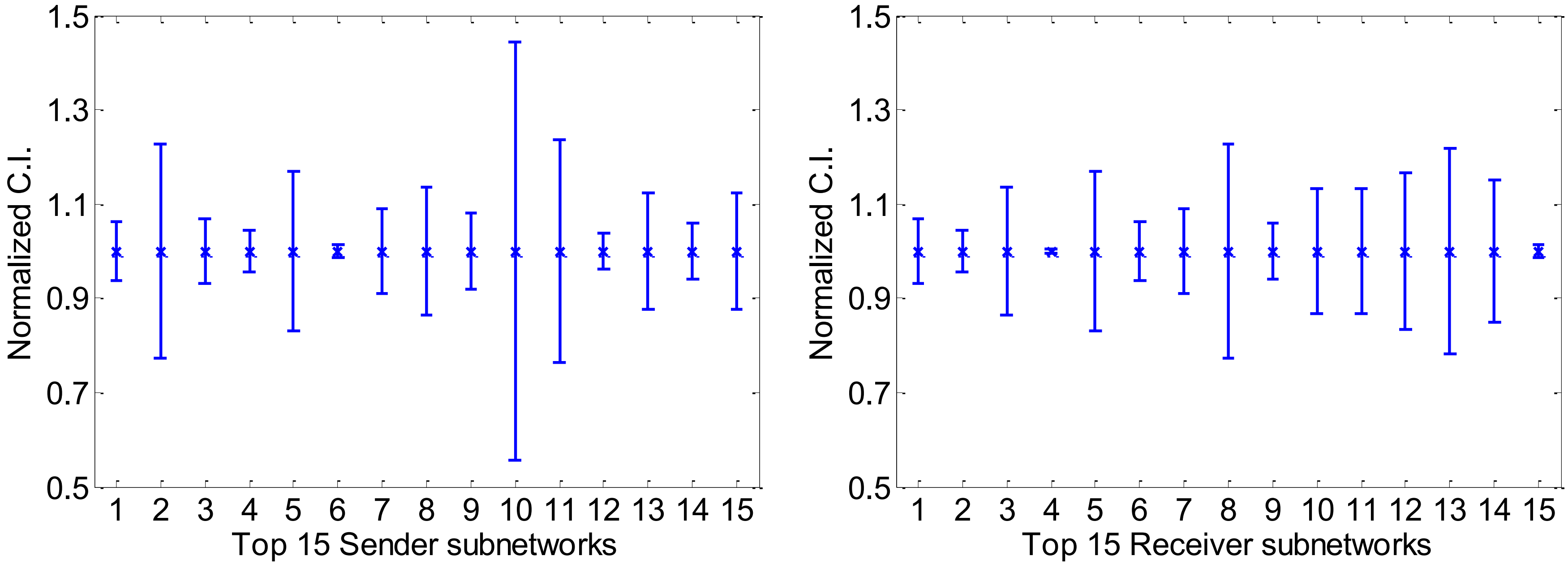
Hypothesis tests for normality: Top 15 sender subnetworks

Subnet	Lilliefors	Anderson-Darling	D'Agostino-Pearson
1	Yes	Yes	Yes
2	No	No	Yes
3	No	No	No
4	No	No	Yes
5	Yes	Yes	Yes
6	Yes	Yes	Yes
7	No	No	No
8	No	No	Yes
9	Yes	Yes	No
10	No	No	No
11	No	No	Yes
12	No	No	No
13	Yes	Yes	Yes
14	Yes	Yes	Yes
15	No	No	Yes

Hypothesis tests for normality: Top 15 receiver subnetworks

Subnet	Lilliefors	Anderson-Darling	D'Agostino-Pearson
1	No	No	No
2	No	No	Yes
3	No	No	Yes
4	Yes	Yes	Yes
5	Yes	Yes	Yes
6	Yes	Yes	Yes
7	No	No	No
8	No	No	Yes
9	Yes	Yes	Yes
10	Yes	Yes	Yes
11	No	No	No
12	Yes	Yes	Yes
13	No	No	Yes
14	Yes	Yes	Yes
15	Yes	Yes	Yes

Confidence interval for mean partial ratio (α =0.05)



- For most subnetworks, the expected mean *partial* ratio varies less than 20% around the sample average
  - Hence, the first assumption seems reasonable if the *total* ratio behaves like the *partial* ratio

Future (pending) Work

- Experimental setup to validate the proposed technique
  - Real topology v/s simulations

- Extension to Sampled NetFlow
  - Effect of packet sampling in the accuracy of the estimates

NetFlow-based bandwidth estimation in IP networks

Rodrigo Alarcón-Reyes
Department of Electrical and Computer Engineering
McGill University

March 13, 2012