Analysis of billing data from a hybrid satellite network

Savio Lau

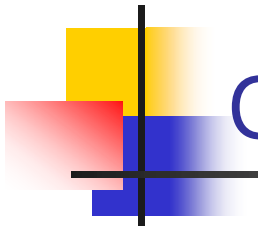
saviol@cs.sfu.ca

Communication Networks Laboratory

http://www.ensc.sfu.ca/research/cnl

School of Engineering Science

Simon Fraser University



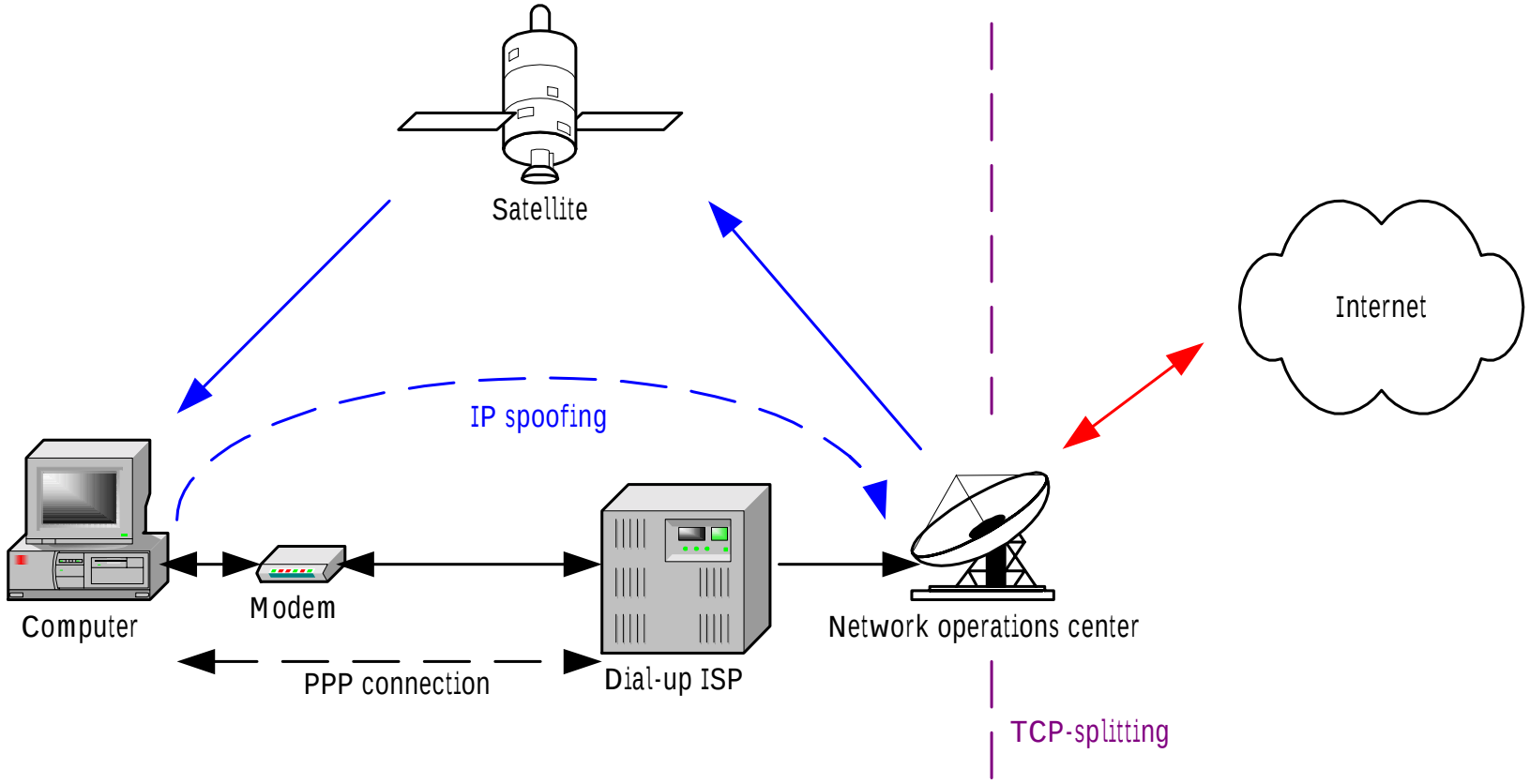
Outline

- Introduction
- Previous work
- Billing data:
  - processing
  - analysis
- Current work
- References

ChinaSat hybrid satellite network

- Manufactured by Hughes Network Systems Inc.
- Employs geosynchronous satellites
- Provides data and television services:
  - DirecTV: satellite television service
  - DirecPC: unidirectional satellite data service
  - DirecWay: bi-directional satellite data service replacing DirecPC

DirecPC system diagram



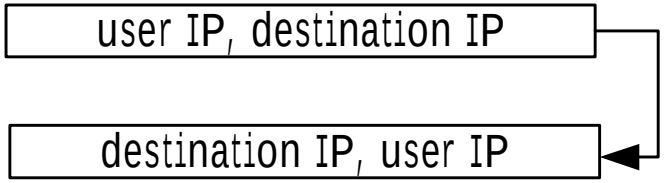
PPP: Point-to-point protocol

IP spoofing and TCP splitting

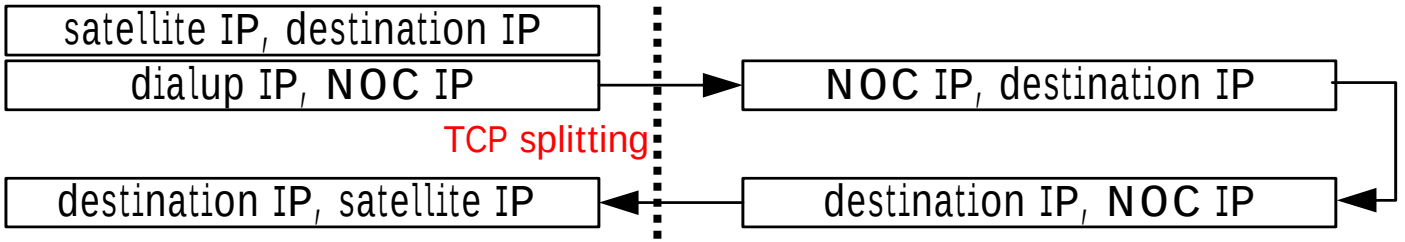
- Each user has 2 IP addresses: dialup IP and satellite IP
- Data transfer pair for dialup, DSL, or cable modem is:
  - source (user) IP, destination IP
- DirecPC connection:
  - dialup IP is used to transmit to network operations center (NOC)
  - embedded in the packet is the satellite IP and destination address (IP tunneling and spoofing)
  - NOC transmits to the destination on behalf of the user (TCP splitting) and replies back to the user via satellite using satellite IP (IP spoofing)

IP spoofing and TCP splitting

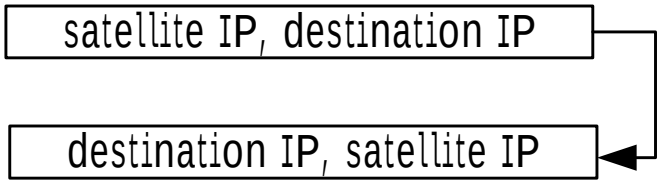
Dialup, cable, or DSL transmissions:



Hybrid satellite based Internet transmissions:



Hybrid satellite based Internet transmissions viewed by application/user:



IP packet address field:

source IP, destination IP

Collected ChinaSat data

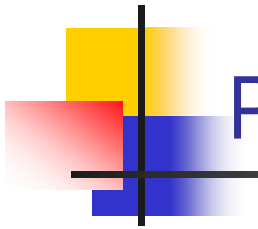
- Collected ChinaSat DirecPC data:
  - continuous billing data from October 31st 2002 to January 10th, 2003
  - 35 GBytes of data from tcpdump traces from December 14th, 2002 to January 10th, 2003

Outline

- Introduction
- Previous work
- Billing data:
  - processing
  - analysis
- Current work
- References

Previous work

- Previous work (Q. Shao) on the tcpdump traces:
  - tcp protocol analysis: >90% of connections are http traffic
  - ranking of visited websites with the discrete Gaussian exponential (DGX) distribution
  - self-similarity parameter estimation:
    - wavelet transform, periodogram, and variance time plot
  - modeling of interarrival time and received bytes per connection using common distributions:
    - exponential, lognormal, and Pareto



Previous work

- Previous work (Q. Shao) on the tcpdump traces:
  - short term prediction using AR+atrous redundant wavelet
- Previous work (Q. Shao) on ChinaSat billing data:
  - long term prediction on billing data using Seasonal Autoregressive Integrative Moving Average (SARIMA) method

Outline

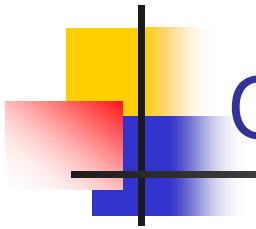
- Introduction
- Previous work
- Billing data:
  - processing
  - analysis
- Current work
- References

Billing data processing

- ChinaSat billing data is collected in ~1,700 text files from two systems turbo1 and turbo2:
  - each file contains summary data from one system for one hour
- Billing data has the following format:
  - RecLen, RecTyp, SiteID (HEX), Start Time, Stop Time, Cmin, Bill, CRxByte, CTxByte, CRxPkt, and CTxPkt
  - important columns are: siteID, Start and Stop times, Cmin, CTxByte, CRxByte, CTxPkt, and CRxPkt
  - Cmin contains errors (hours with > 60 min usage)

Billing data processing

- Data is processed for Matlab use:
  - a linux shell script concatenates separate data files into one large text file and replaces the hexadecimal siteID identified to base 10
  - two hours are discarded manually: January 6th 2002, at 21:00 and January 7th, 2002, at 15:00
  - cleaned data is imported into Matlab for combining overlapped turbo1 and turbo2 entries
  - hours are normalized from 0 to 1,690



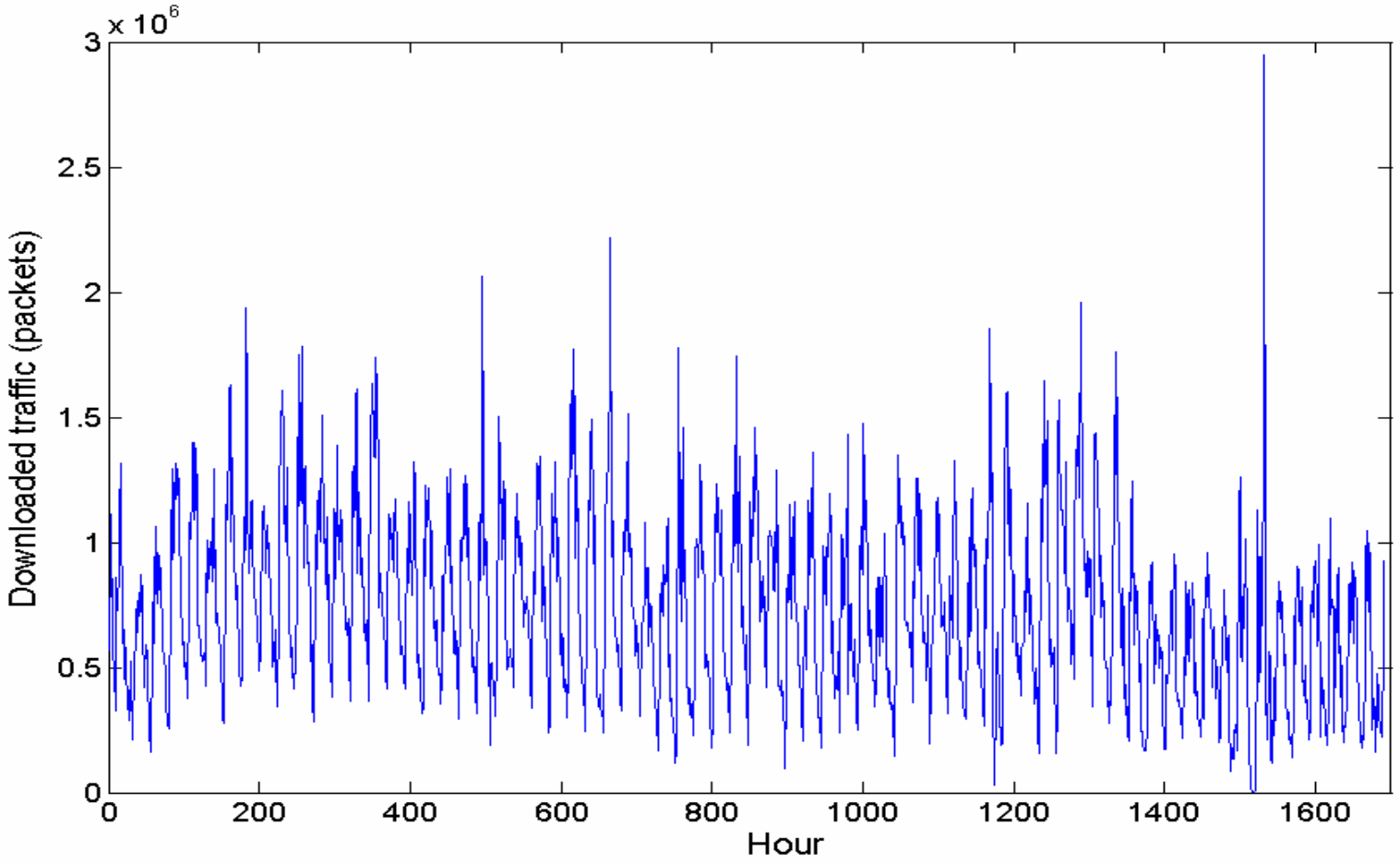
Outline

- Introduction
- Previous work
- Billing data:
  - processing
  - analysis
- Current work
- References

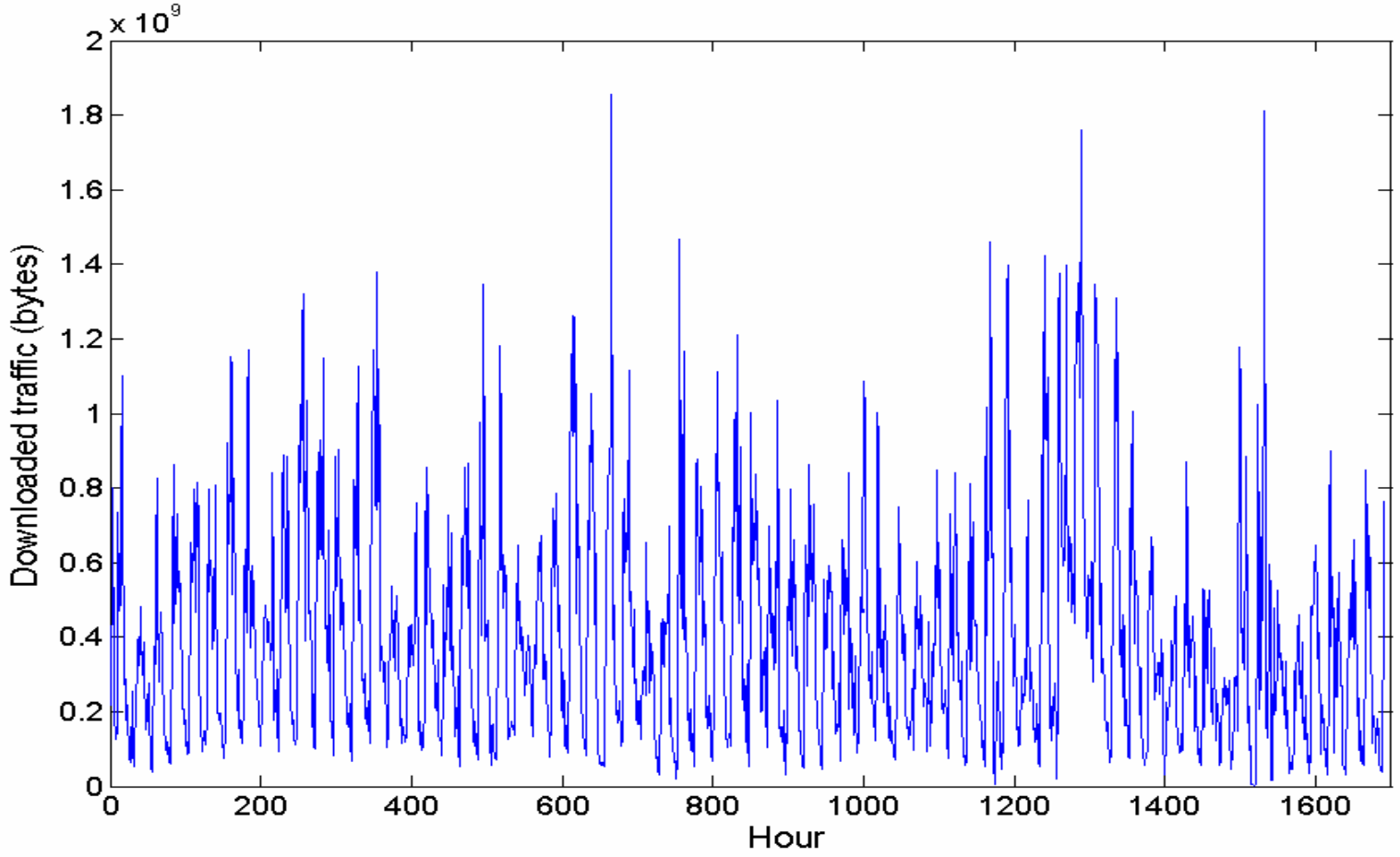
Billing data analysis

- Matlab scripts are written to generate statistics:
  - aggregated hourly traffic
  - aggregated daily traffic
  - hourly traffic for each individual siteID
  - daily traffic for each individual siteID
  - siteIDs traffic ranked by traffic volume
  - k-means clustering based on siteIDs' traffic volume

Aggregated hourly traffic



Aggregated hourly traffic



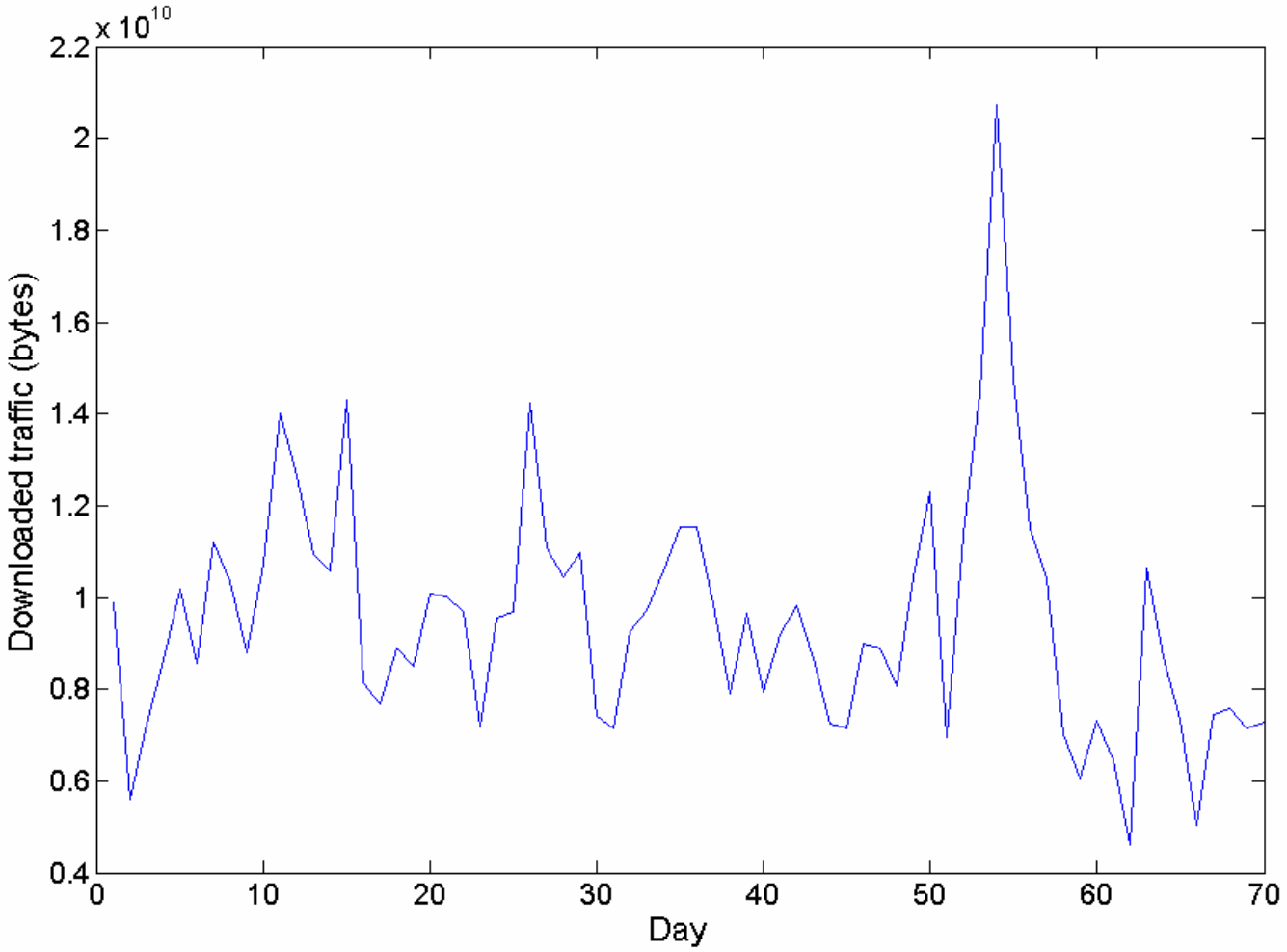
Aggregated hourly traffic: observations

- 1,691 hours are aggregated
- Daily maximum occurs most often between 14:00 and 18:00
- Maximum traffic volume (packets) occurs at hour 1,533:
  - corresponds to Friday, January 3rd, 2003, at 19:00
  - possible outage at 17:00:
    - lowest traffic volume was recorded at hour 1,531
  - increased traffic volume could have been caused by the system outage

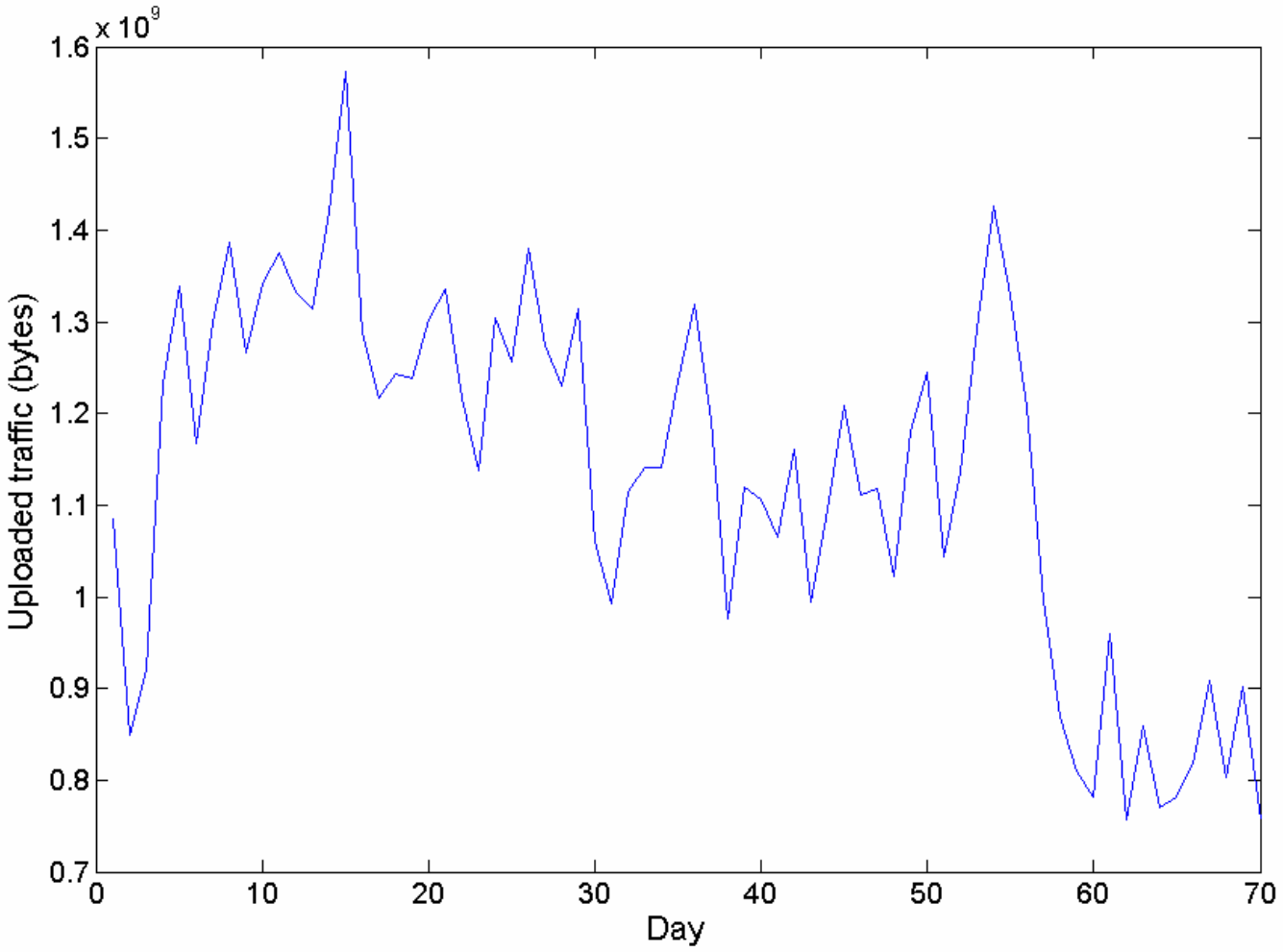
Aggregated hourly traffic: observations

- Traffic pattern changes about 1,300th hour:
  - the minimum traffic during the 1,300th hour is higher compared to other hours
  - the date: December 24th, 2002
  - this implies that holiday traffic patterns are different from regular (non-holiday) traffic
  - SARIMA traffic predictions may underestimate holiday traffic

Aggregated daily traffic



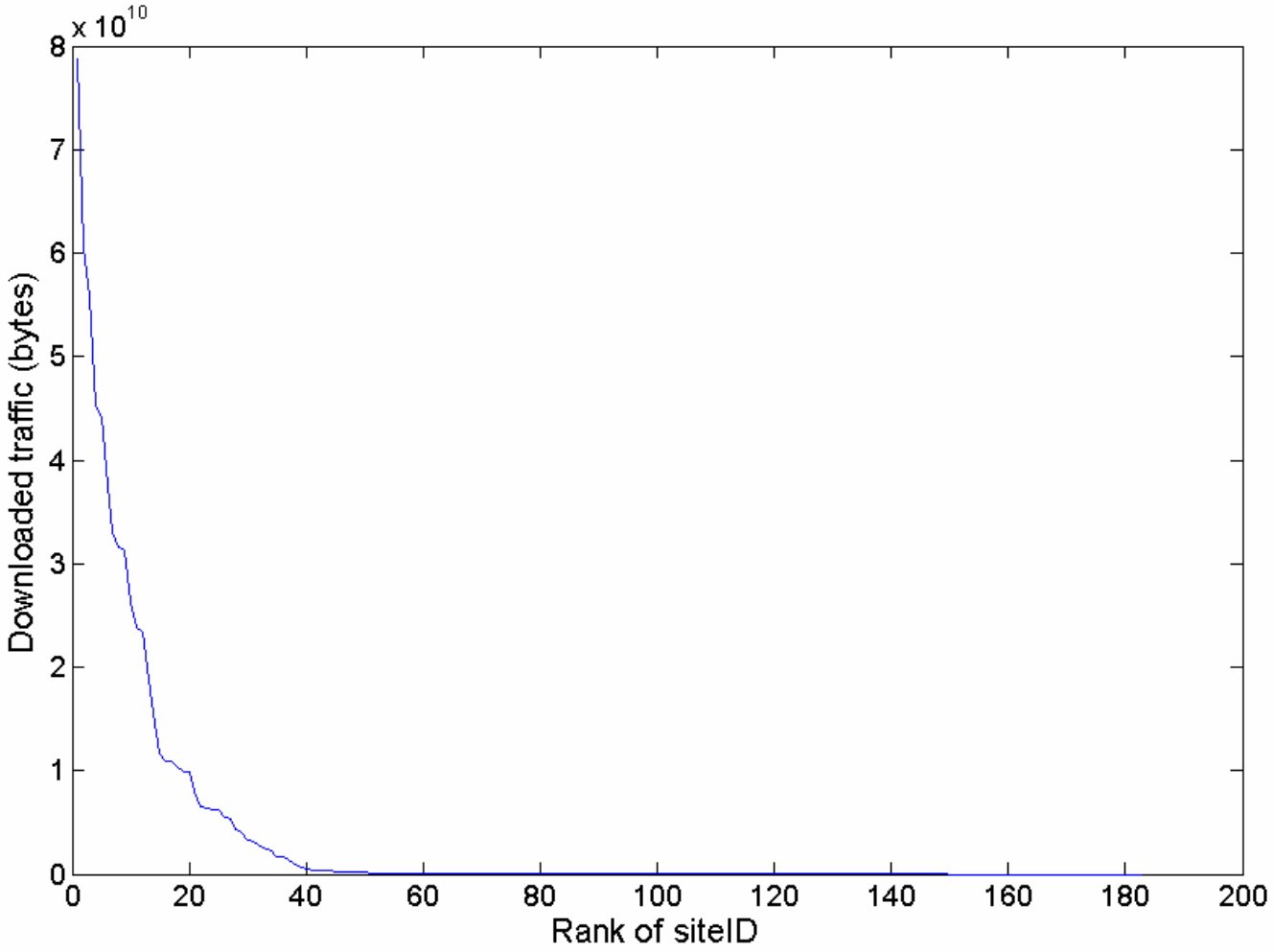
Aggregated daily traffic



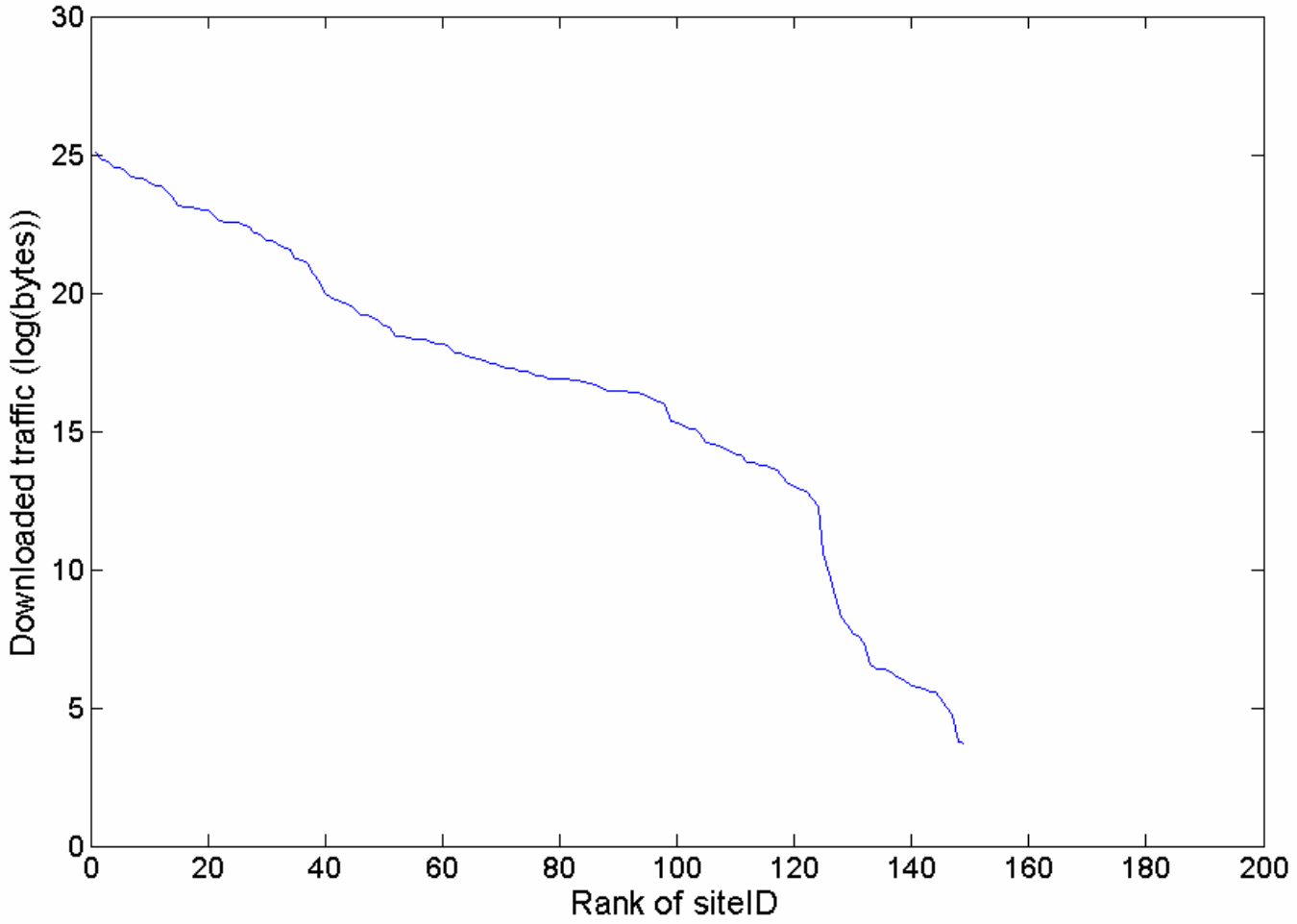
Aggregated daily traffic: observations

- 70 of 72 days are aggregated:
  - October 31, 2002 and January 10, 2003 are removed from the aggregation due to incomplete data (< 24 hours collected)
- Traffic volume is much lower from day 59 to day 70 compared to the first 58 days:
  - corresponds to January 1st, 2003 to January 10th, 2003
- A localized maximum can be detected on day 54:
  - corresponds to December 24th, 2002
  - agrees with hourly traffic observations

siteIDs ranked by traffic volume



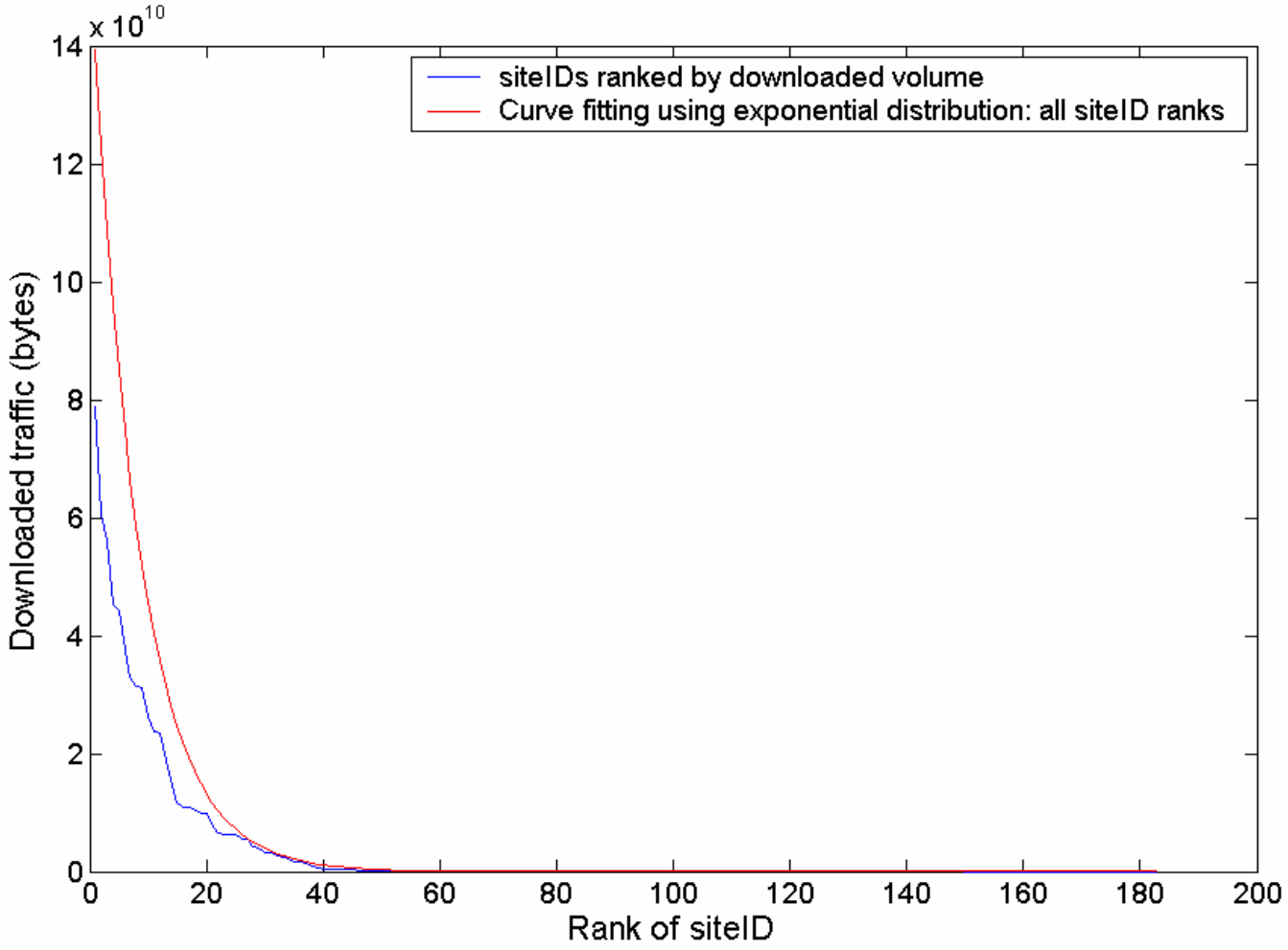
siteIDs ranked by traffic volume: log



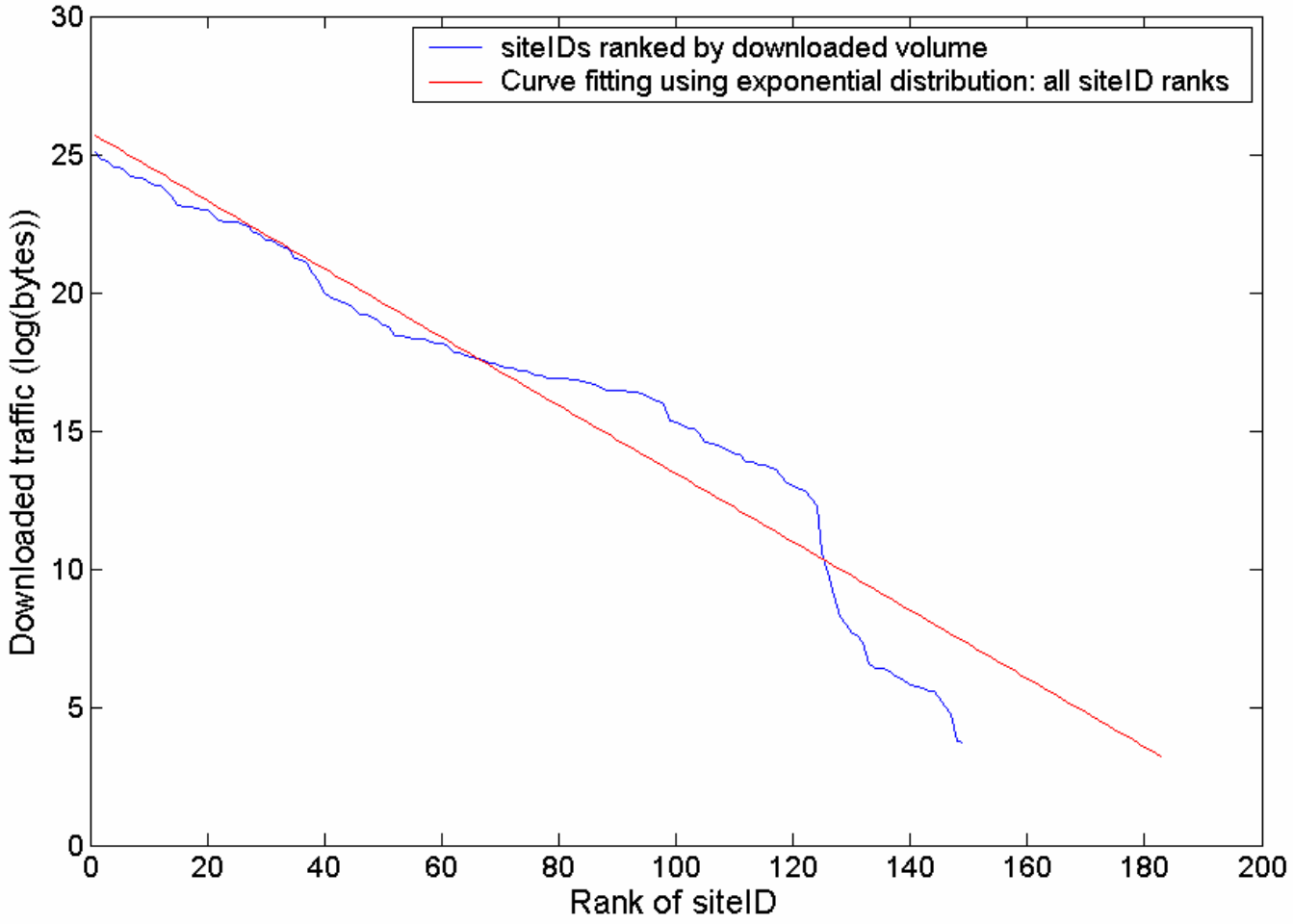
siteIDs ranked by traffic volume

- Only 151 of 183 siteIDs transmitted or received data during the recorded period:
  - 2 of 151 siteIDs transmitted data but did not receive data
  - 15 of 151 siteIDs received data but did not transmit data
- siteIDs ranked by number of downloaded bytes is best fitted using exponential distribution:
  - using all siteID ranks
  - using siteID ranks 1 to 149: all active receivers
  - using siteID ranks 1 to 123: all active receivers in the lower rank

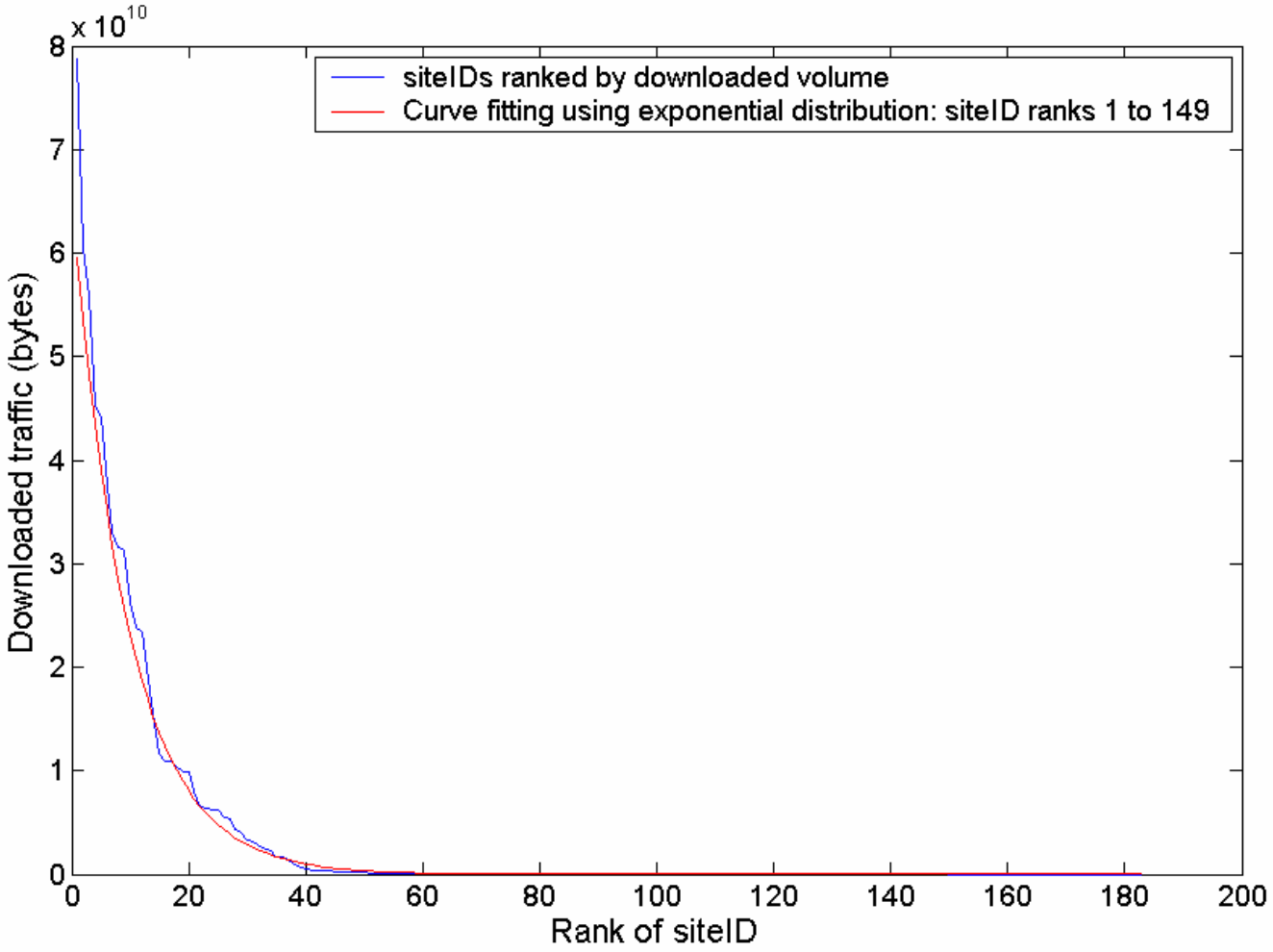
Fitting ranked siteIDs: all siteID ranks



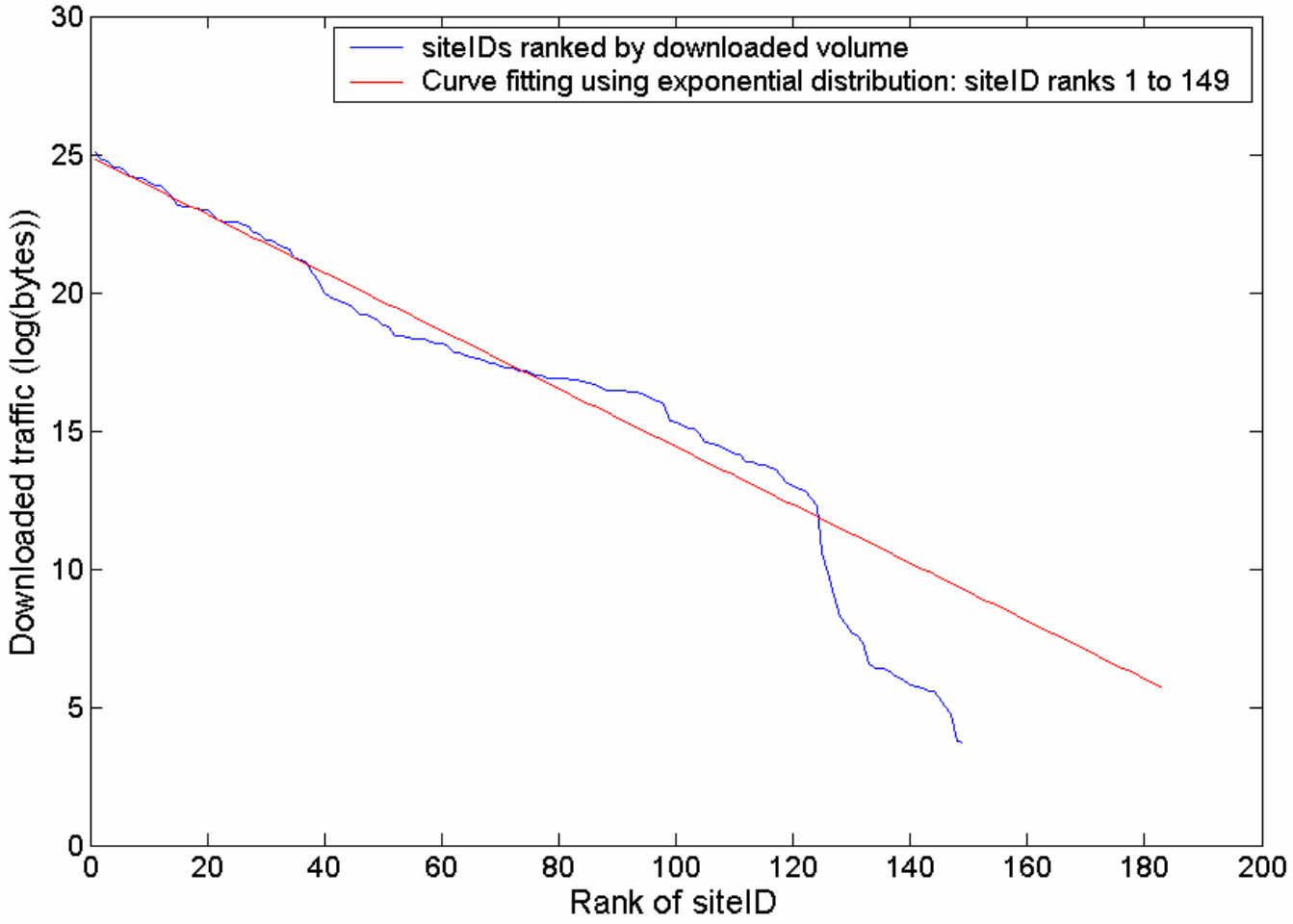
Fitting ranked siteIDs: all siteID ranks (log)



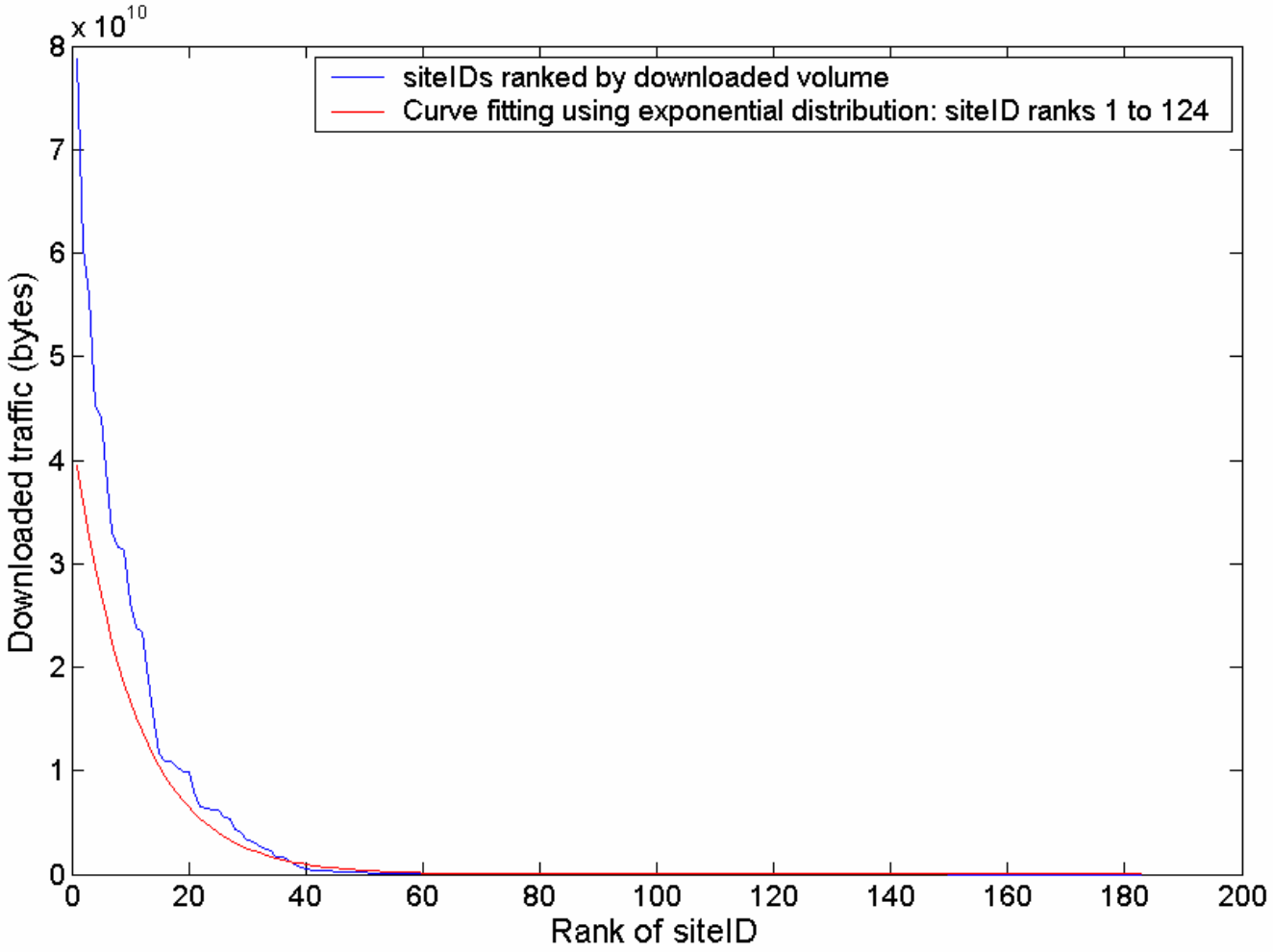
Fitting ranked siteIDs: siteID ranks 1 to 149



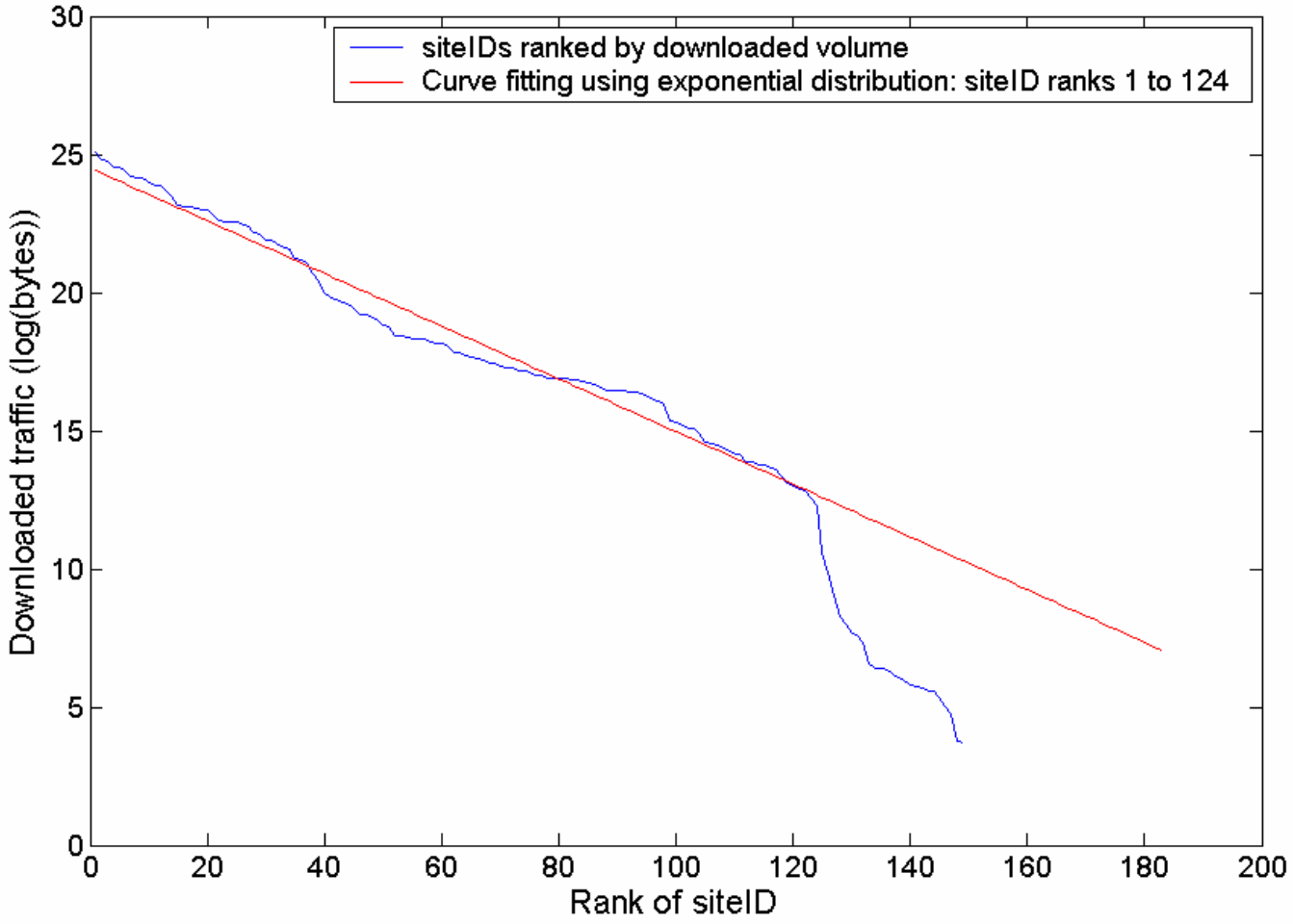
Fitting ranked siteIDs: siteID ranks 1 to 149 (log)



Fitting ranked siteIDs: siteID ranks 1 to 124



Fitting ranked siteIDs: siteID ranks 1 to 124 (log)



siteIDs ranked by traffic volume

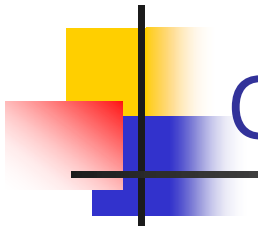
- Kolmogorov Smirnov goodness-of-fit test kstest2 is used to evaluate curve fitting:
  - returns a p-value < 1^{-40} for all 3 cases
  - hypothesis failed in all 3 cases

k-means clustering of siteIDs

- k-means algorithm is used to sort siteIDs based on downloaded bytes:
  - k=3 Cluster sizes: (5, 9, 169)
  - k=4 Cluster sizes: (3, 9, 15, 156)
  - k=5 Cluster sizes: (3, 9, 8, 12, 151)
  - k=6 Cluster sizes: (3, 9, 8, 7, 10, 146)
  - k=7 Cluster sizes: (3, 9, 8, 7, 5, 6, 134)
- For k=7, centroids are approximately 2 times larger compared to the next centroid:
  - centroid values are 10^{10} * (6.5088, 3.2945, 1.2287, 0.6317, 0.3512, 0.1743, 0.0040)

k-means clustering of siteIDs

- The cluster consisting of 3 siteIDs with the largest traffic volume corresponds to sites with 24 hours/7 days usage:
  - it corresponds to Internet cafes (customers with the highest network usage)

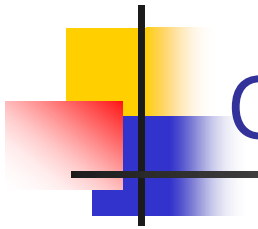


Outline

- Introduction
- Previous work
- Billing data:
  - processing
  - analysis
- Current work
- References

Current work

- Clustering of siteIDs is based on hours of activity:
  - peak vs. non-peak hours
  - total number of active hours for each siteID
- Ranking of siteIDs based on other criteria such as download/upload traffic ratio and activity hours:
  - curve fitting using other distributions for various types of ranking



Outline

- Introduction
- Previous work
- Billing data:
  - processing
  - analysis
- Current work
- References

References: billing data

- Q. Shao and Lj. Trajkovic, “Measurement and analysis of traffic in a hybrid satellite-terrestrial network,” *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 329–336.
- T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, July 2002.