

Mining Network Traffic Data

Ljiljana Trajkovic

ljilja@cs.sfu.ca

Communication Networks Laboratory

<http://www.ensc.sfu.ca/cnl>

Roadmap

- Introduction
- Traffic data and analysis tools:
 - data collection, statistical analysis, clustering tools, prediction analysis
- Case studies:
 - wireless network: *Telus Mobility*
 - public safety wireless network: *E-Comm*
 - satellite network: *ChinaSat*
 - packet data networks: *Internet*
- Conclusions and references

Roadmap

- Introduction
- Traffic data and analysis tools:
 - data collection, statistical analysis, clustering tools, prediction analysis
- Case studies:
 - wireless network: Telus Mobility
 - public safety wireless network: E-Comm
 - satellite network: ChinaSat
 - packet data network: Internet
- Conclusions and references

Network traffic measurements

- Traffic measurements in operational networks help:
 - understand traffic characteristics in deployed networks
 - develop traffic models
 - evaluate performance of protocols and applications
- Traffic analysis:
 - provides information about the user behavior patterns
 - enables network operators to understand the behavior of network users

User cluster analysis

- Clustering analysis groups or segments a collection of objects into subsets or **clusters** based on similarity
- An object can be described by a set of measurements or by its relations to other objects
- Clustering algorithms can be employed to analyze network user behaviors
- Network users are classified into clusters, according to the similarity of their behavior patterns
- With user clusters, traffic prediction is reduced to predicting and aggregating users' traffic from few clusters

Traffic prediction

- Traffic prediction: important to assess future network capacity requirements and to plan future network developments
- Auto-Regressive Integrated Moving Average (ARIMA) model:
 - general model for forecasting time series
 - past values: AutoRegressive (AR) structure
 - past random fluctuant effect: Moving Average (MA) process
- Seasonal ARIMA (SARIMA): a variation of the ARIMA model that captures seasonal pattern

Roadmap

- Introduction
- Traffic data and analysis tools:
 - data collection, statistical analysis, clustering tools, prediction analysis
- Case studies:
 - wireless network: Telus Mobility
 - public safety wireless network: E-Comm
 - satellite network: ChinaSat
 - packet data networks: Internet
- Conclusions and references

Telus Mobility CDPD network

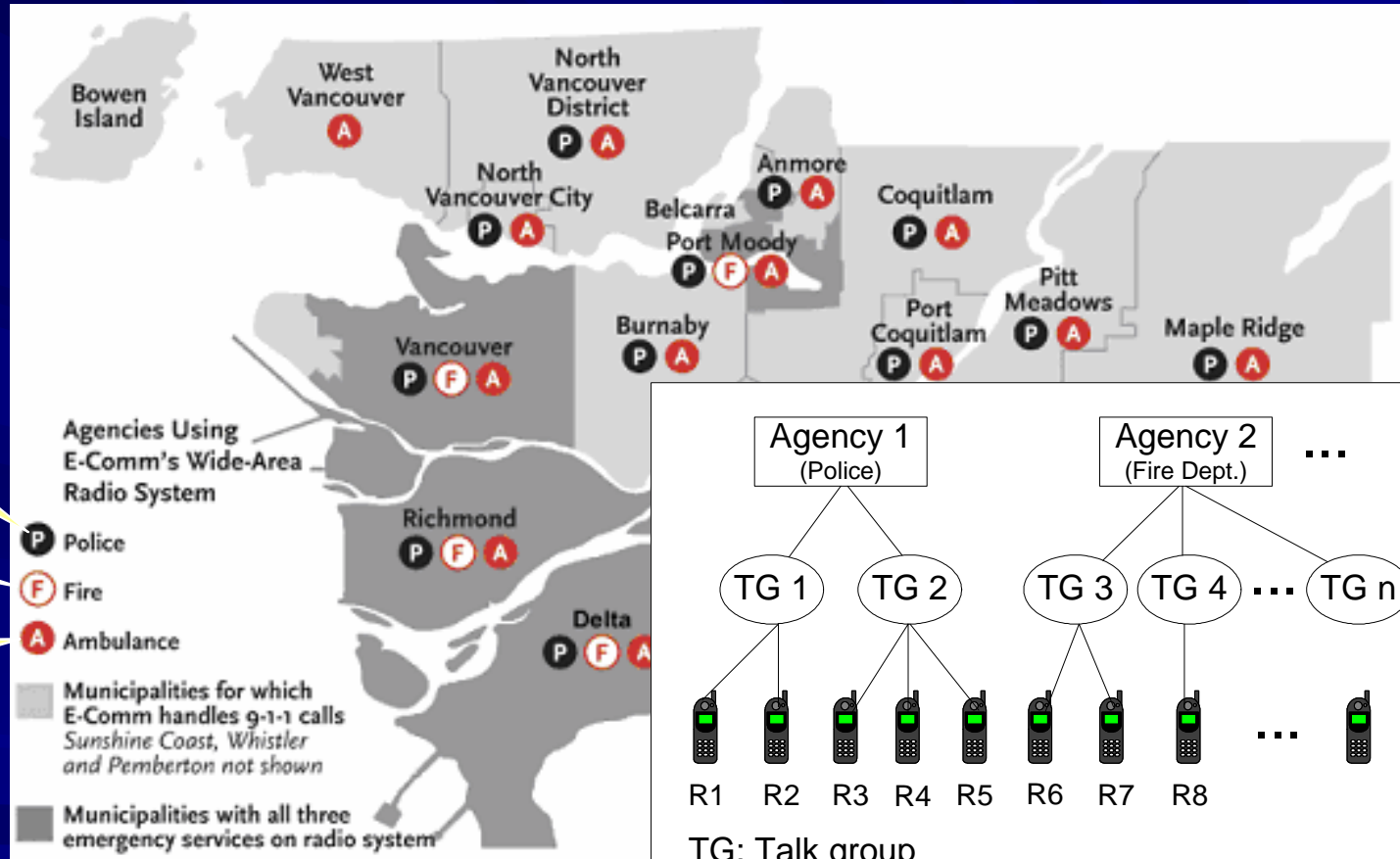
- Downtown Vancouver area:
June 12, 1998: 14:56:37.56 to 15:24:46.88
- The trace consisted of Mobile Data Link Layer Protocol (MDLP) frames
- CDPD network subscribers used both TCP and UDP over IP protocols
- Channel capacity: 19.2 kbps
- 10 mobile end systems appeared in the cell

Telus Mobility CDPD network

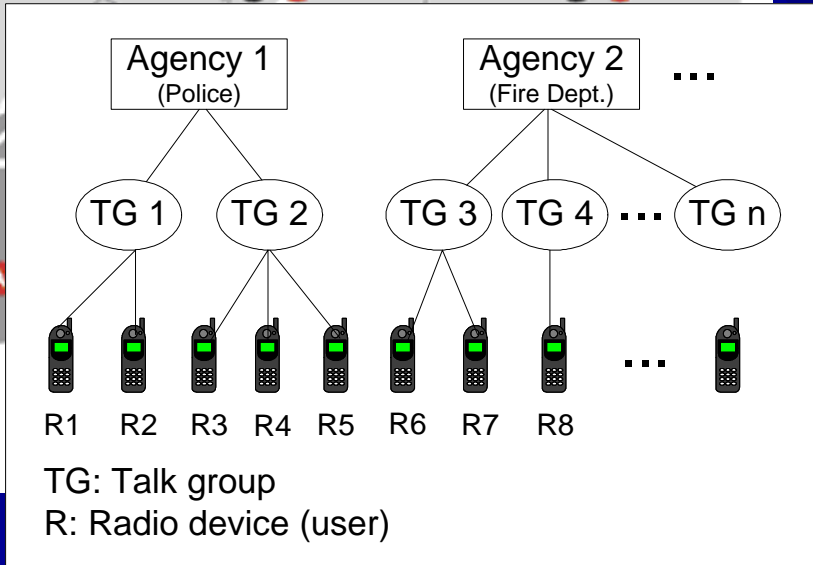
- Aggregated traffic: the total input traffic to the mobile data base station
- Only a 20 minute interval available
- Telus Mobility CDPD network traffic trace:

Duration	Number of link layer frames	Number of bytes	Average traffic load	Network utilization
20 min	1,281	152,439	1,016 bps	5.29 %

E-Comm network: coverage and user agencies



- RCMP and Police
- Fire
- Ambulance
- Other



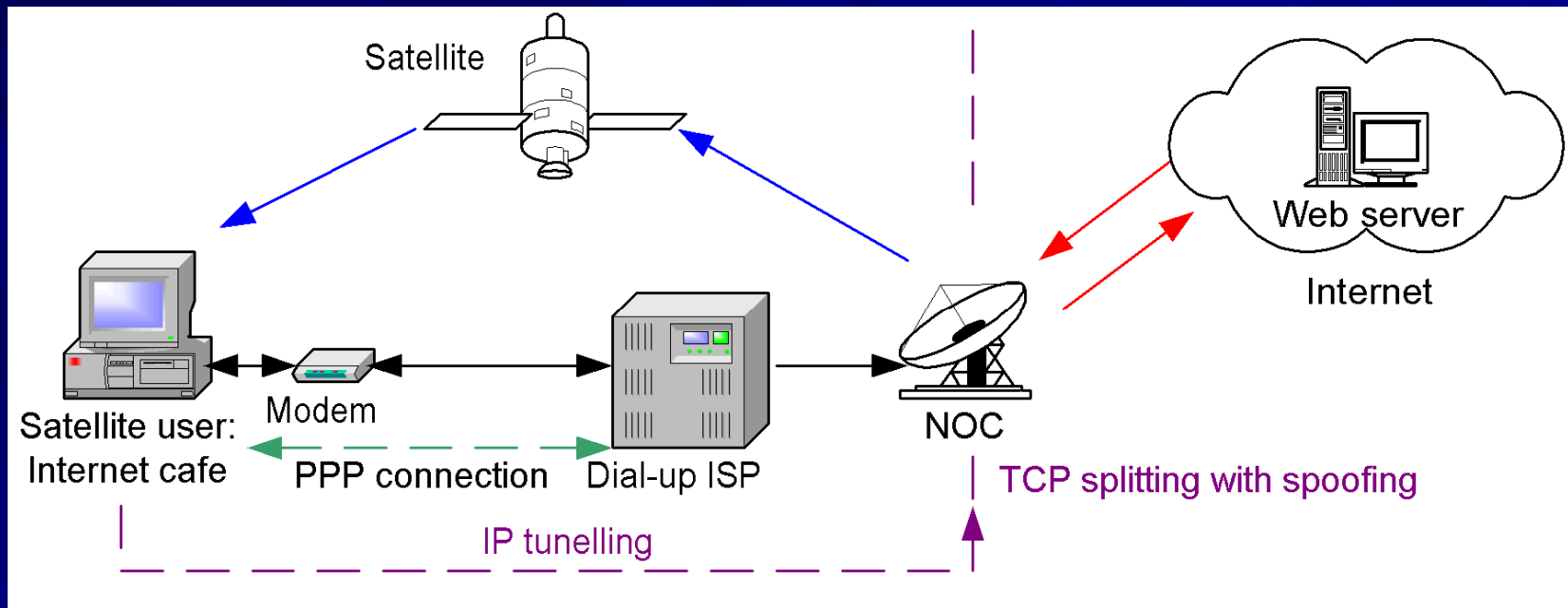
Traffic data

- 2001 data set:
 - 2 days of traffic data
 - 2001-11-1 to 2001-11-02 (110,348 calls)
- 2002 data set:
 - 28 days of continuous traffic data
 - 2002-02-10 to 2002-03-09 (1,916,943 calls)
- 2003 data set:
 - 92 days of continuous traffic data
 - 2003-03-01 to 2003-05-31 (8,756,930 calls)

ChinaSat hybrid satellite network

- Employs geosynchronous satellites deployed by Hughes Network Systems Inc.
- Provides data and television services:
 - DirecPC (Classic): unidirectional satellite data service
 - DirecTV: satellite television service
 - DirecWay (Hughnet): new bi-directional satellite data service that replaces DirecPC
- DirecPC transmission rates:
 - 400 kb/s from satellite to user
 - 33.6 kb/s from user to network operations center (NOC) using dial-up
- Improves performance using TCP splitting with spoofing

DirecPC system diagram



NOC: Network operations center
PPP: Point-to-point protocol

Analysis of collected data

- Analysis of patterns and statistical properties of two sets of data from the ChinaSat DirecPC network:
 - billing records
 - tcpdump traces
- Billing records:
 - daily and weekly traffic patterns
 - user classification:
 - single and multi-variable k-means clustering based on average traffic
 - hierarchical clustering based on user activity

Analysis of collected data

- Analysis of tcpdump trace
 - tcpdump trace:
 - protocols and applications
 - TCP options
 - operating system fingerprinting
 - network anomalies
 - Developed C program pcapread:
 - processes tcpdump files
 - produces custom output
 - eliminates the need for packet capture library libpcap

Detecting network anomalies

- Scans and worms
- Denial of service
- Flash crowd
- Traffic shift
- Alpha traffic
- Traffic volume anomalies

Billing records

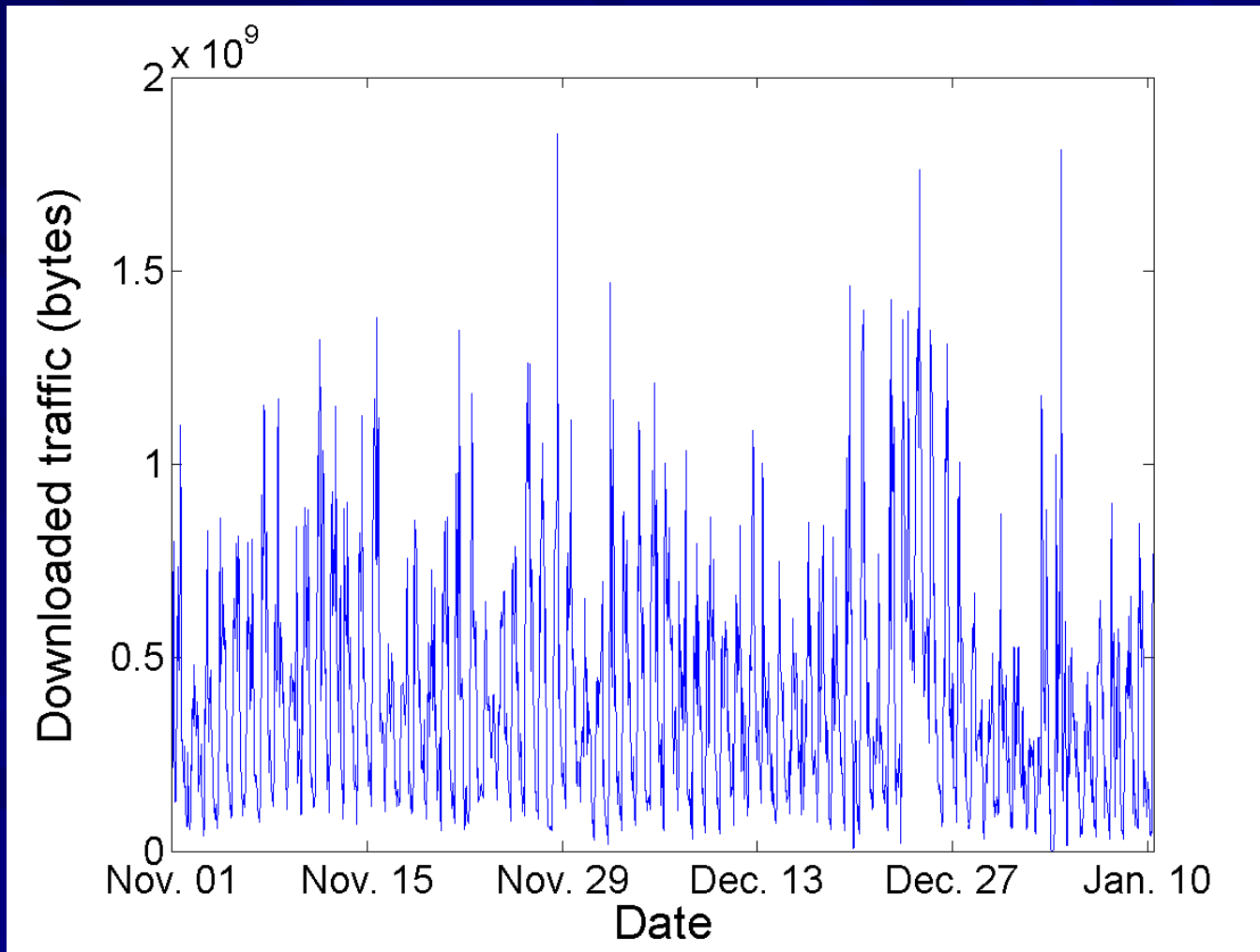
- Records collected during the continuous period from 23:00 on Oct. 31, 2002 to 11:00 on Jan. 10, 2003
- Each file contains the hourly traffic summary for each user
- Fields of interest:
 - SiteID (user identification)
 - Start (record start time)
 - CTxByt (number of bytes downloaded by a user)
 - CRxByt (number of bytes uploaded by a user)
 - CTxPkt (number of packets downloaded by a user)
 - CRxPkt (number of packets uploaded by a user)

download: satellite to user
upload: user to NOC

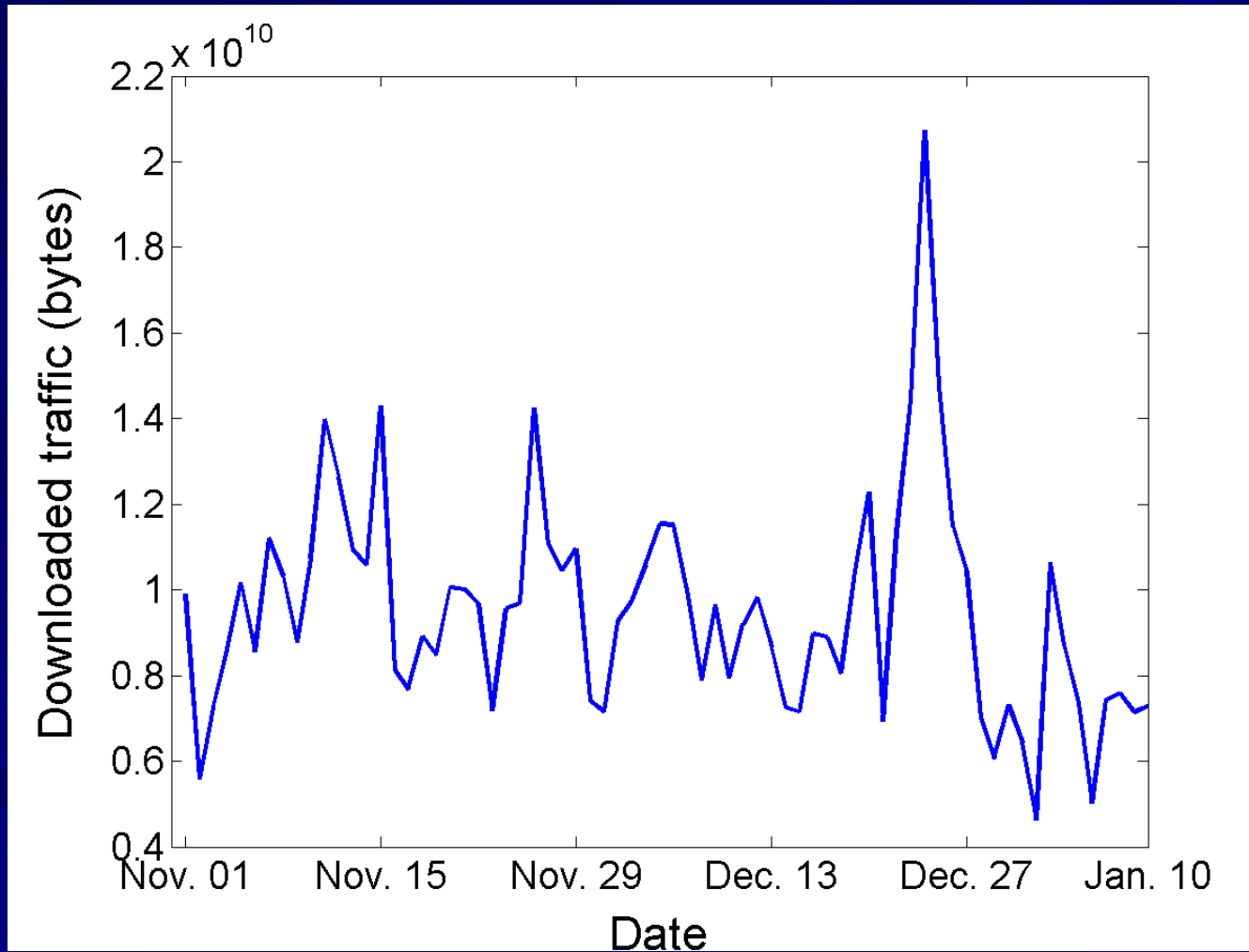
Billing records: characteristics

- 186 unique SiteIDs
- Daily and weekly cycles:
 - lower traffic volume on weekends
 - daily cycle starts at 7 AM, rises to three daily maxima at 11 AM, 3 PM, and 7 PM, then decrease monotonically until 7 AM
- Highest daily traffic recorded on Dec. 24, 2002
- Outage occurred on Jan. 3, 2003

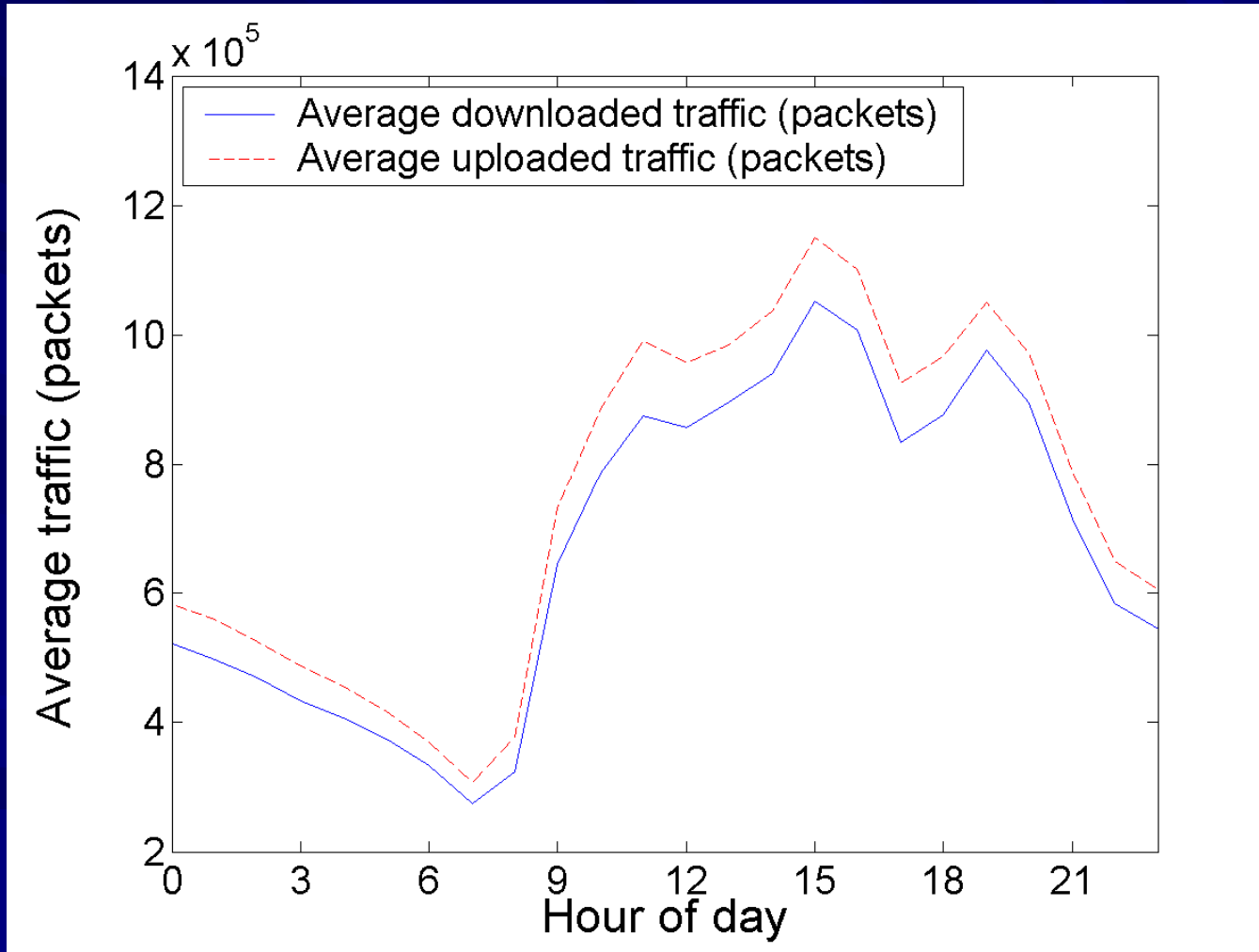
Aggregated hourly traffic



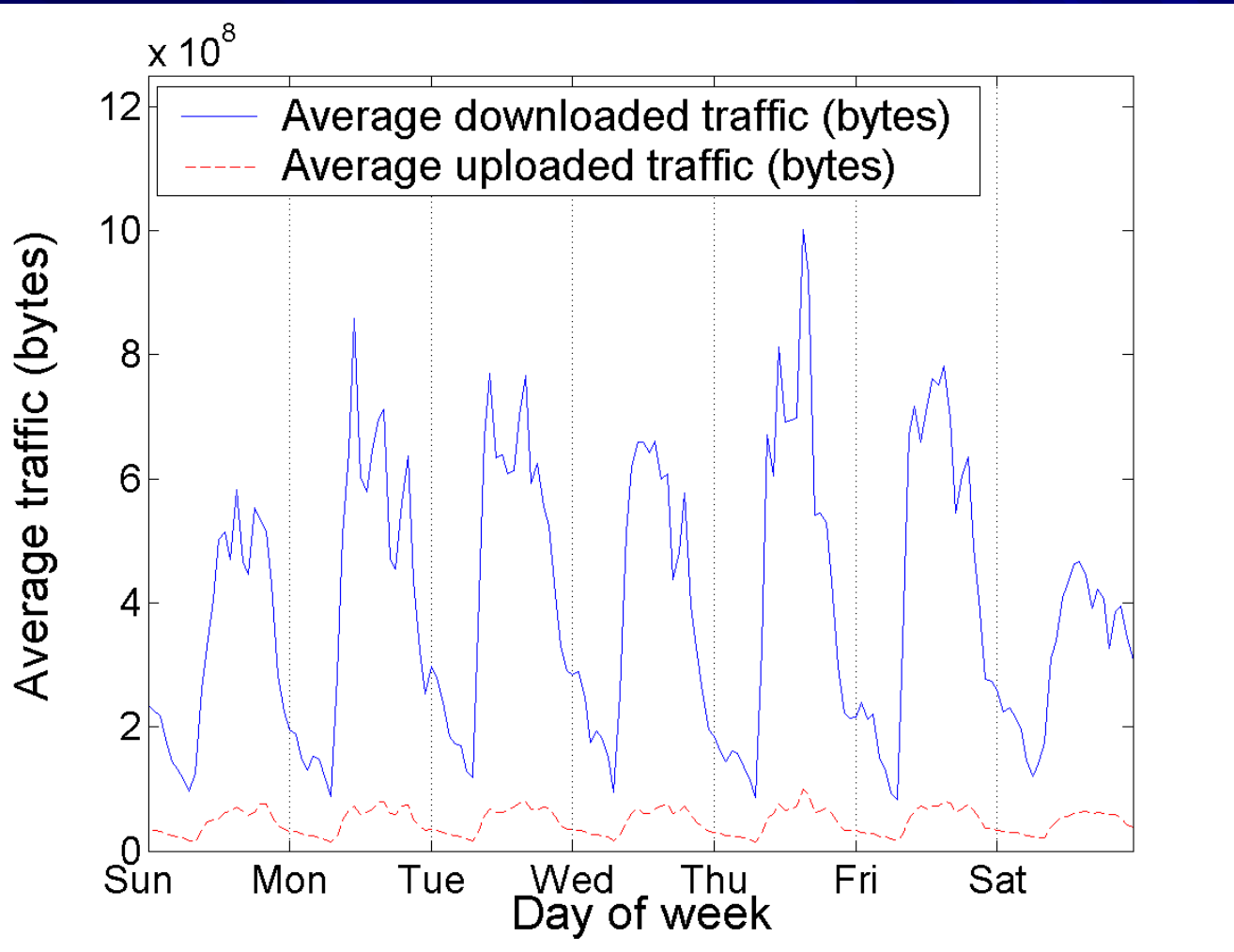
Aggregated daily traffic



Daily diurnal traffic: average downloaded bytes



Weekly traffic: average downloaded bytes



Ranking of user traffic

- Users are ranked according to the traffic volume
- The top user downloaded 78.8 GB, uploaded 11.9 GB, and downloaded/uploaded ~205 million packets
- Most users download/uploaded little traffic
- Cumulative distribution functions (CDFs) are constructed from the ranks:
 - top user accounts for 11% of downloaded bytes
 - top 25 users contributed 93.3% of downloaded bytes
 - top 37 users contributed 99% of total traffic (packets and bytes)

tcpdump traces

- Traces continuously collected from 11:30 on Dec. 14, 2002 to 11:00 on Jan. 10, 2003 at the NOC
- The first 68 bytes of a each TCP/IP packet were captured
- ~63 GB of data contained in 127 files
- User IP address is not constant due to the use of the private IP address range and dynamic IP
- Majority of traffic is TCP:
 - 94% of total bytes and 84% of total packets
 - HTTP (port 80) accounts for 90% of TCP connections and 76% of TCP bytes
 - FTP (port 21) accounts for 0.2% of TCP connections and 11% of TCP bytes

tcpdump output example

```
12/15/2002 04:27:05.328455 192.168.1.83.63260 > 211.167.92.197.6732: . ack 489 win 8192
12/15/2002 04:27:05.331020 211.100.18.48.80 > 192.168.1.164.41842: S
    2928120965:2928120965(0) ack 3324468 win 64240 <mss 1460,nop,nop,sackOK> (DF)
12/15/2002 04:27:05.331612 61.135.137.66.9013 > 192.168.1.164.41806: P
    3091059901:3091060177(276) ack 11834706 win 5840 (DF)
12/15/2002 04:27:05.343507 192.168.1.164.41806 > 61.135.137.66.9013: . ack 276 win 8192
12/15/2002 04:27:05.343748 192.168.1.242.45045 > 210.51.17.96.9065: P 25309490:25309522(32)
    ack 1436759200 win 8192 (DF)
12/15/2002 04:27:05.359048 192.168.1.242.44991 > 211.167.92.226.6732: P 17:25(8) ack 16 win
    8192 (DF)
12/15/2002 04:27:05.359218 192.168.1.83.64228 > 61.242.153.168.11745: udp 92
12/15/2002 04:27:05.359383 192.168.1.164.9668 > 211.150.186.218.4000: udp 60
12/15/2002 04:27:05.359537 192.168.1.83.64228 > 61.242.153.168.11745: udp 92
12/15/2002 04:27:05.359693 192.168.1.83.64228 > 61.242.153.168.11745: udp 92
12/15/2002 04:27:05.359694 61.152.252.11.55901 > 192.168.1.242.45311: P 48:56(8) ack 1 win
    62851 (DF)
12/15/2002 04:27:05.362315 210.51.17.96.9065 > 192.168.1.242.45045: . ack 32 win 32120 (DF)
12/15/2002 04:27:05.366415 61.135.137.26.9013 > 192.168.1.242.45533: P 112:138(26) ack 1 win
    6432 (DF)
```

Network anomalies

- Ethereal/Wireshark, tcptrace, and pcapread
- Four types of network anomalies were detected:
 - invalid TCP flag combinations
 - large number of TCP resets
 - UDP and TCP port scans
 - traffic volume anomalies

Analysis of TCP flags

TCP flag	Packet count	% of total
SYN only	19,050,849	48.500
RST only	7,440,418	18.900
FIN only	12,679,619	32.300
*SYN+FIN	408	0.001
*RST+FIN (no PSH)	85,571	0.200
*RST+PSH (no FIN)	18,111	0.050
*RST+FIN+PSH	8,329	0.020
*Total number of packets with invalid TCP flag combinations	112,419	0.300
Total packet count	39,283,305	100.000

Large number of TCP resets

- Connections are terminated by either TCP FIN or TCP RST:
 - 12,679,619 connections were terminated by FIN (63%)
 - 7,440,418 connections were terminated by RST (37%)
- Large number of TCP RST indicates that connections are terminated in error conditions
- TCP RST is employed by Microsoft Internet Explorer to terminate connections instead of TCP FIN

UDP and TCP port scans

- UDP port scans are found on UDP port 137 (NETBEUI)
- TCP port scans are found on TCP ports:
 - 80 Hypertext transfer protocol (HTTP)
 - 139 NETBIOS extended user interface (NETBEUI)
 - 434 HTTP over secure socket layer (HTTPS)
 - 1433 Microsoft structured query language (MS SQL)
 - 27374 Subseven trojan

TCP: transport control protocol

UDP: user defined protocol

UDP port scans originating from the ChinaSat network

192.168.2.30:137 - 195.x.x.98:1025
192.168.2.30:137 - 202.x.x.153:1027
192.168.2.30:137 - 210.x.x.23:1035
192.168.2.30:137 - 195.x.x.42:1026
192.168.2.30:137 - 202.y.y.226:1026
192.168.2.30:137 - 218.x.x.238:1025
192.168.2.30:137 - 202.y.y.226:1025
192.168.2.30:137 - 202.y.y.226:1027
192.168.2.30:137 - 202.y.y.226:1028
192.168.2.30:137 - 202.y.y.226:1029
192.168.2.30:137 - 202.y.y.242:1026
192.168.2.30:137 - 61.x.x.5:1028
192.168.2.30:137 - 219.x.x.226:1025
192.168.2.30:137 - 213.x.x.189:1028
192.168.2.30:137 - 61.x.x.193:1025
192.168.2.30:137 - 202.y.y.207:1028
192.168.2.30:137 - 202.y.y.207:1025
192.168.2.30:137 - 202.y.y.207:1026
192.168.2.30:137 - 202.y.y.207:1027
192.168.2.30:137 - 64.x.x.148:1027

- Client (192.168.2.30) source port (137) scans external network addresses at destination ports (1025-1040):
 - > 100 are recorded within a three-hour period
 - targeted IP addresses are variable
 - multiple ports are scanned per IP
 - may correspond to Bugbear, OpaSoft, or other worms

UDP port scans direct to the ChinaSat network

210.x.x.23:1035 - 192.168.1.121:137
210.x.x.23:1035 - 192.168.1.63:137
210.x.x.23:1035 - 192.168.2.11:137
210.x.x.23:1035 - 192.168.1.250:137
210.x.x.23:1035 - 192.168.1.25:137
210.x.x.23:1035 - 192.168.2.79:137
210.x.x.23:1035 - 192.168.1.52:137
210.x.x.23:1035 - 192.168.6.191:137
210.x.x.23:1035 - 192.168.1.241:137
210.x.x.23:1035 - 192.168.2.91:137
210.x.x.23:1035 - 192.168.1.5:137
210.x.x.23:1035 - 192.168.1.210:137
210.x.x.23:1035 - 192.168.6.127:137
210.x.x.23:1035 - 192.168.1.201:137
210.x.x.23:1035 - 192.168.6.179:137
210.x.x.23:1035 - 192.168.2.82:137
210.x.x.23:1035 - 192.168.1.239:137
210.x.x.23:1035 - 192.168.1.87:137
210.x.x.23:1035 - 192.168.1.90:137
210.x.x.23:1035 - 192.168.1.177:137
210.x.x.23:1035 - 192.168.1.39:137

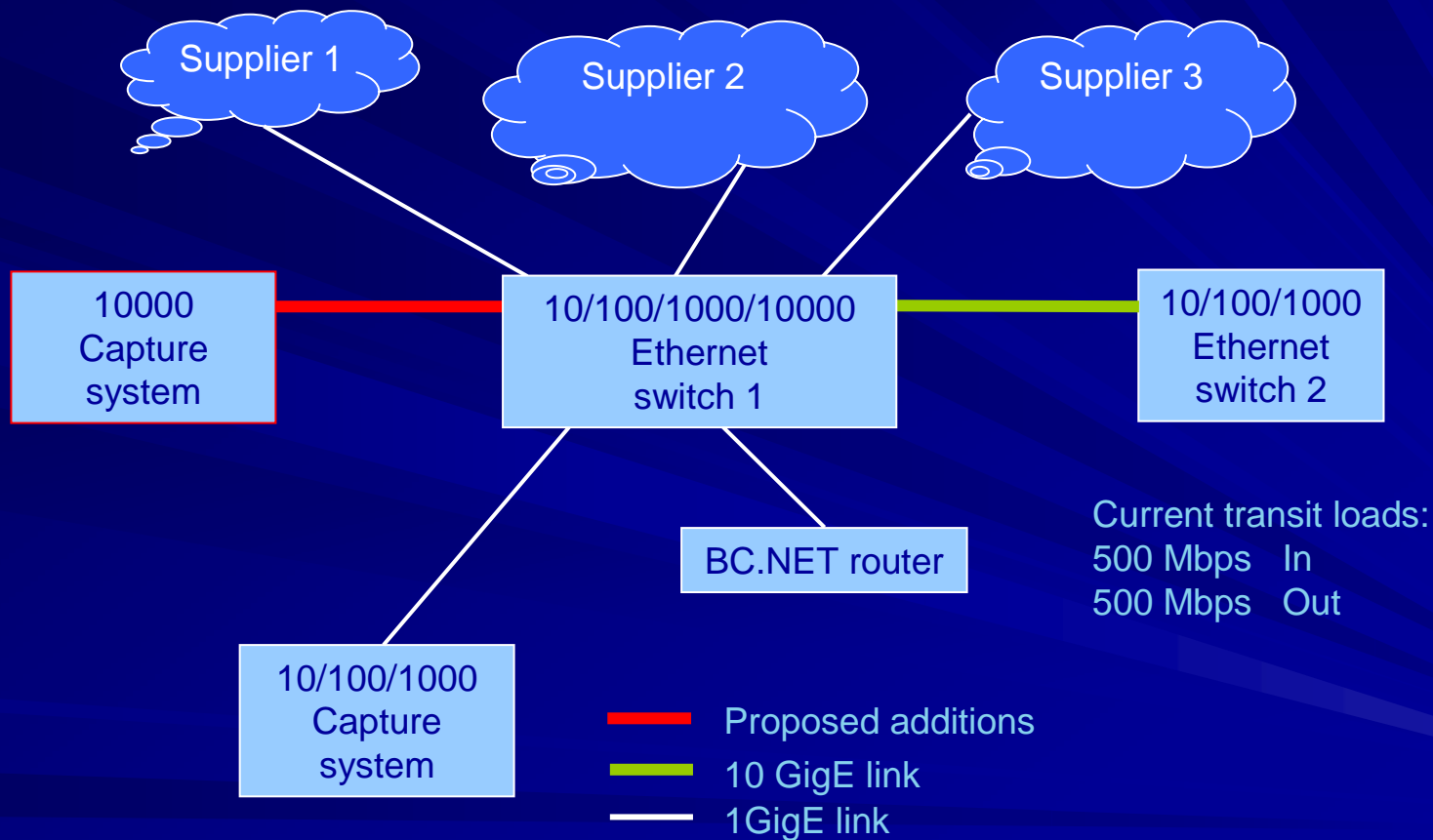
- External address (210.x.x.23) scans for port (137) (NETBEUI) response within the ChinaSat network from source port (1035):
 - > 200 are recorded within a three-hour period
 - targets IP addresses are not sequential
 - may correspond to Bugbear, OpaSoft, or other worms

Internet AS-level data

Source of data are routing tables:

- **Route Views:** <http://www.routeviews.org>
 - most participating ASs reside in North America
- **RIPE (Réseaux IP européens):** <http://www.ripe.net/ris>
 - most participating ASs reside in Europe

BC.NET traffic measurements



Conclusions

- Traffic data from deployed networks (Telus Mobility, E-Comm, ChinaSat, the Internet) was used to:
 - evaluate network performance
 - characterize and model traffic (inter-arrival and call holding times)
 - classify network users using clustering algorithms
 - predict network traffic by employing SARIMA models based on aggregate user traffic and user clusters
 - detect network anomalies using wavelet analysis

References: downloads

http://www.ensc.sfu.ca/~ljilja/publications_date.html

- S. Lau and Lj. Trajkovic, "Analysis of traffic data from a hybrid satellite-terrestrial network," in *Proc. QShine 2007*, Vancouver, BC, Canada, Aug. 2007.
- B. Vujičić, L. Chen, and Lj. Trajković, "Prediction of traffic in a public safety network," in *Proc. ISCAS 2006*, Kos, Greece, May 2006, pp. 2637-2640.
- N. Cackov, J. Song, B. Vujičić, S. Vujičić, and Lj. Trajković, "Simulation of a public safety wireless networks: a case study," *Simulation*, vol. 81, no. 8, pp. 571-585, Aug. 2005.
- B. Vujičić, N. Cackov, S. Vujičić, and Lj. Trajković, "Modeling and characterization of traffic in public safety wireless networks," in *Proc. SPECTS 2005*, Philadelphia, PA, July 2005, pp. 214-223.
- J. Song and Lj. Trajković, "Modeling and performance analysis of public safety wireless networks," in *Proc. IEEE IPCCC*, Phoenix, AZ, Apr. 2005, pp. 567-572.
- H. Chen and Lj. Trajković, "Trunked radio systems: traffic prediction based on user clusters," in *Proc. IEEE ISWCS 2004*, Mauritius, Sept. 2004, pp. 76-80.
- D. Sharp, N. Cackov, N. Lasković, Q. Shao, and Lj. Trajković, "Analysis of public safety traffic on trunked land mobile radio systems," *IEEE J. Select. Areas Commun.*, vol. 22, no. 7, pp. 1197-1205, Sept. 2004.
- Q. Shao and Lj. Trajković, "Measurement and analysis of traffic in a hybrid satellite-terrestrial network," in *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 329-336.
- N. Cackov, B. Vujičić, S. Vujičić, and Lj. Trajković, "Using network activity data to model the utilization of a trunked radio system," in *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 517-524.
- J. Chen and Lj. Trajkovic, "Analysis of Internet topology data," *Proc. IEEE Int. Symp. Circuits and Systems*, Vancouver, British Columbia, Canada, May 2004, vol. IV, pp. 629-632.

BCNET



Questions?

www.bc.net