Mining Network Traffic Data

Ljiljana Trajković
ljilja@cs.sfu.ca

Communication Networks Laboratory
http://www.ensc.sfu.ca/cnl
School of Engineering Science
Simon Fraser University, Vancouver, British Columbia
Canada

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 networks: 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
- Traffic prediction: important to assess future network capacity requirements and to plan future network developments

Self-similarity

- Self-similarity implies a "fractal-like" behavior: data on various time scales have similar patterns
- A wide-sense stationary process $X(n)$ is called (exactly second order) self-similar if its autocorrelation function satisfies:
  - $r^{(m)}(k) = r(k)$, $k \geqslant 0$, $m = 1, 2, \ldots, n$,

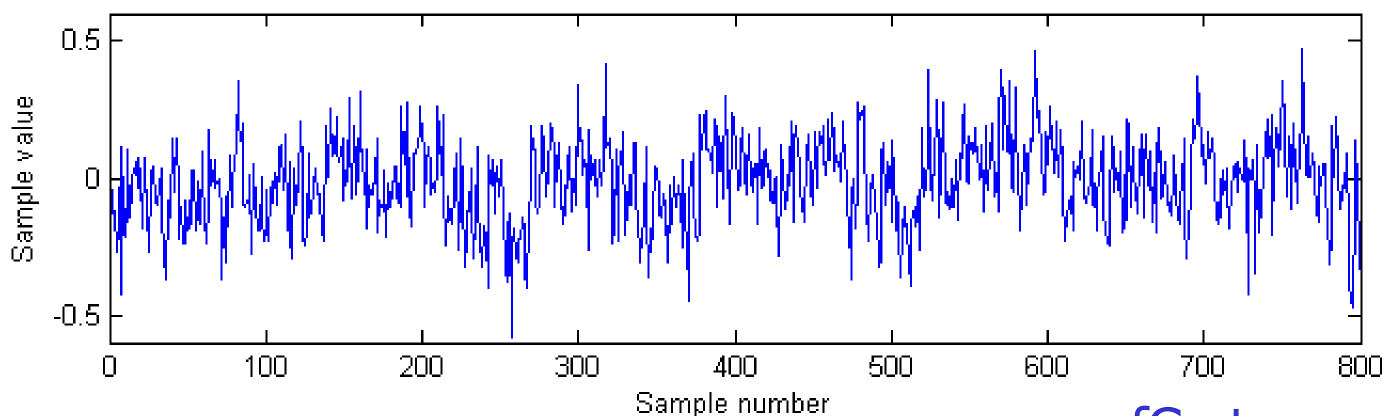
    where m is the level of aggregation
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become "smoother" when aggregated (unlike Poisson traffic)

Self-similar processes

- Properties:
  - slowly decaying variance
  - long-range dependence
  - Hurst parameter (H)
- Processes with only short-range dependence (Poisson): H = 0.5
- Self-similar processes: 0.5 < H < 1.0
- As the traffic volume increases, the traffic becomes more bursty, more self-similar, and the Hurst parameter increases

Long-range dependence: properties

- High variability:
  - when the sample size increases, variance of the sample mean decays more slowly than expected
- Burstiness over a range of timescales:
  - long runs of large values followed by long runs of small values, repeated in aperiodic patterns

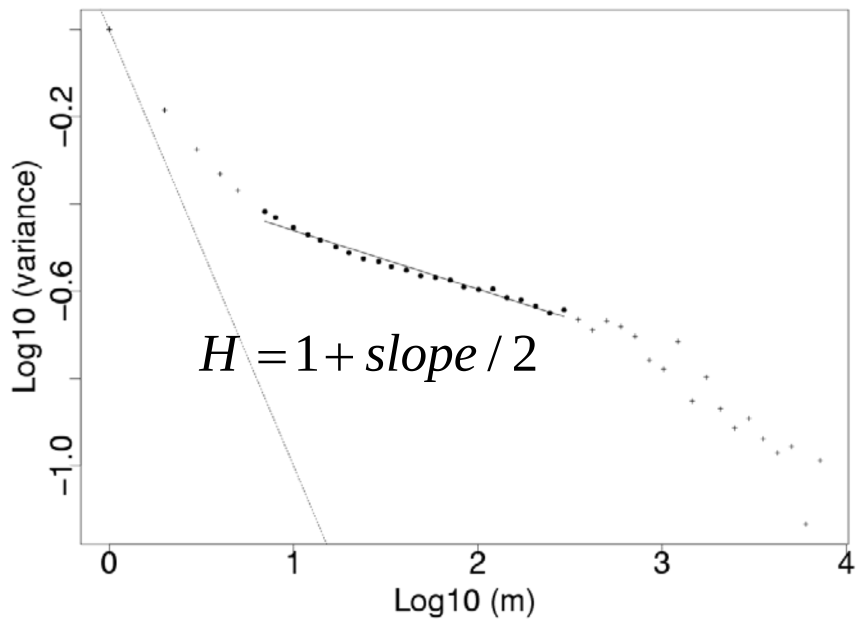


fGn trace

Estimation of H

Various estimators:

- variance-time plots
- R/S plots
- periodograms
- wavelets



Their performance often depends on the characteristics of the data trace under analysis

Clustering 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

Clustering analysis

- Groups collection of objects into subsets (clusters):
  - resulting intra-cluster similarity is high while inter-cluster similarity is low
- The inter-cluster distance reflects dissimilarity between clusters:
  - Euclidean distance between two cluster centroids (mean value of objects in a cluster, viewed as cluster's center of gravity)
- The intra-cluster distance expresses coherent similarity of data in the same cluster:
  - average distance of objects from their cluster centroids
- Better clustering:
  - large inter-cluster and small intra-cluster distances

Clustering quality

- Overall clustering quality: defined as difference between minimum inter-cluster and maximum intra-cluster distances
  - larger indicator implies better overall clustering quality
- Silhouette coefficient (x):

  $$(b(x) - a(x)) / \max \{a(x), b(x)\}$$

  a(x) and b(x) are average distances between data point x and other data points in clusters A and B, respectively
  - independent of number of clusters K

Clustering algorithms

- Two approaches:
  - partitioning clustering (k-means)
  - hierarchical clustering
- Clustering tools:
  - AutoClass tool
  - k-means algorithm

P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., AAAI Press/MIT Press, 1996.

L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

Clustering algorithms: k-means

- The k-means algorithm is commonly used for data clustering
- The algorithm is well-known for its simplicity and efficiency
- Based on the input parameter k, it partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high and the inter-cluster similarity is low
- Similarity of clusters is measured with respect to the mean value of the objects in a cluster (viewed as the cluster's center of gravity)

k-means: partitioning clustering

- Constructs k partitions of the data from n objects, where $k \leq n$
- Two constraints:
  - each cluster must contain at least one object
  - each object must belong to exactly one group
- Requires exhaustive enumeration of all possible combinations to find the optimal cluster solution

k-means clustering

- Generates k clusters from n objects
- Requires two inputs:
  - k: number of desired partitions
  - n objects
- Uses random placement of initial clusters
- Determines clustering results through an iteration technique to relocate objects to the most similar cluster:
  - similarity is defined as the distance between objects
  - objects that are closer to each other are more similar
- Computational complexity of $O(nkt)$, where t is the maximum number of iterations

Finding number of clusters

- The number of clusters k is not known a priori
- k-means algorithm is repeated for different k values
- Number of clusters is found by comparing average SC value for various values of k:
  - average SC is calculated for all objects
  - the natural number of clusters k is found at the local maxima

SC: silhouette coefficient

Traffic prediction: ARIMA model

- 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
- ARIMA model explicitly includes differencing
- ARIMA (p, d, q):
  - autoregressive parameter: p
  - number of differencing passes: d
  - moving average parameter: q

Traffic prediction: SARIMA model

- Seasonal ARIMA is a variation of the ARIMA model
- Seasonal ARIMA (SARIMA) model:

$$(p,d,q)\times(P,D,Q)_S$$

  - captures seasonal pattern
- SARIMA additional model parameters:
  - seasonal period parameter: S
  - seasonal autoregressive parameter: P
  - number of seasonal differencing passes: D
  - seasonal moving average parameter: Q

SARIMA models: selection criteria

- Order (p,d,q) selected based on:
  - time series plot of traffic data
  - autocorrelation and partial autocorrelation functions
- Validity of parameter selection:
  - Akaike's information criterion:
    - *AIC*
    - corrected *AICc*
  - Bayesian information criterion *BIC*

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

ChinaSat data: analysis

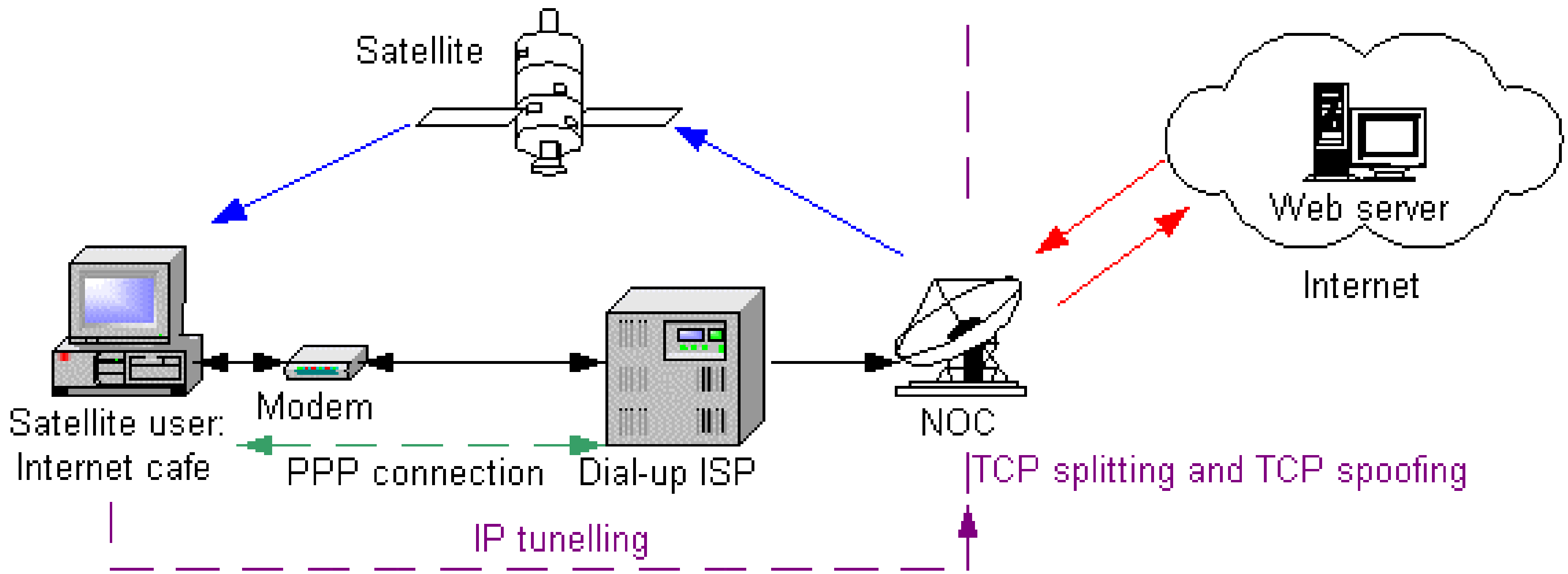
- Analysis of network traffic:
  - characteristics of TCP connections
  - network traffic patterns
  - statistical and cluster analysis of traffic
  - anomaly detection:
    - statistical methods
    - wavelets
    - principle component analysis

TCP: transport control protocol

Network and traffic data

- ChinaSat: network architecture and TCP
- Analysis of billing records:
  - aggregated traffic
  - user behavior
- Analysis of tcpdump traces:
  - general characteristics
  - TCP options and operating system (OS) fingerprinting
  - network anomalies

DirecPC system diagram



Characteristics of satellite links

- Large coverage area
- High bandwidth
- Long propagation delay
- Large bandwidth-delay product
- High bit error rates:
  - 10^{-6} without error correction
  - 10^{-3} or 10^{-2} due to extreme weather and interference
- Path asymmetry

Characteristics of satellite links

- ChinaSat hybrid satellite network
  - Employs geosynchrous 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

ChinaSat data: analysis

- ChinaSat traffic is self-similar and non-stationary
- Hurst parameter differs depending on traffic load
- Modeling of TCP connections:
  - inter-arrival time is best modeled by the Weibull distribution
  - number of downloaded bytes is best modeled by the lognormal distribution
- The distribution of visited websites is best modeled by the discrete Gaussian exponential (DGX) distribution

ChinaSat data: analysis

- Traffic prediction:
  - autoregressive integrative moving average (ARIMA) was successfully used to predict uploaded traffic (but not downloaded traffic)
  - wavelet + autoregressive model outperforms the ARIMA model

Q. Shao and Lj. Trajkovic, "Measurement and analysis of traffic in a hybrid satellite-terrestrial network," in *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 329–336.

S. Lau and Lj. Trajkovic, "Analysis of traffic data from a hybrid satellite Q. terrestrial network," in *Proc. QShine 2007*, Vancouver, BC, Canada, Aug. 2007.

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

Network anomalies

- Scans and worms:
  - packets are sent to probe network hosts
  - used to discover and exploit resources
- Denial of service:
  - large number of packets is directed to a single destination
  - makes a host incapable of handling incoming connections or exhausts available bandwidth along paths to the destination

Network anomalies

- Flash crowd:
  - high volume of traffic is destined to a single destination
  - caused by breaking news or availability of new software
- Traffic shift:
  - redirection of traffic from one set of paths to another
  - caused by route changes, link unavailability, or network congestion

Network anomalies

- Alpha traffic:
  - unusually high volume of traffic between two endpoints
  - caused by file transfers or bandwidth measurements
- Traffic volume anomalies:
  - significant deviation of traffic volume from usual daily or weekly patterns
  - classified as:
    - outages: caused by unavailable links, crashed servers, or routing problems
    - short term increases in demand: caused by short term events such as holiday traffic
  - involve multiple sources and destinations

Billing records

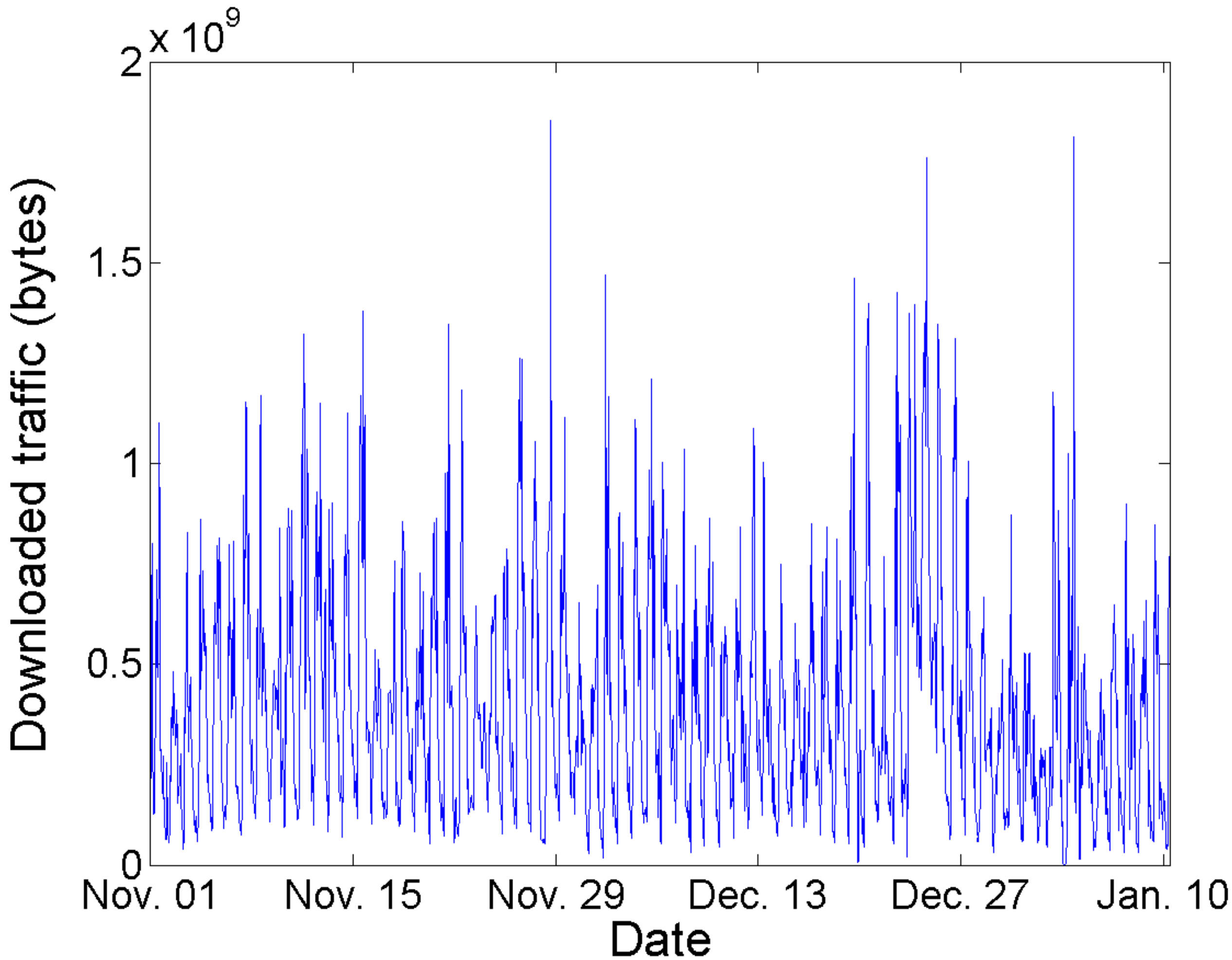
- Records were collected during the continuous period from 23:00 on Oct. 31, 2002 to 11:00 on Jan. 10, 2003
- Each file contains the hourly traffic summary for each user
- Fields of interests:
  - SiteID (user identification)
  - Start (record start time)
  - CTxByt (number of bytes downloaded by a user)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

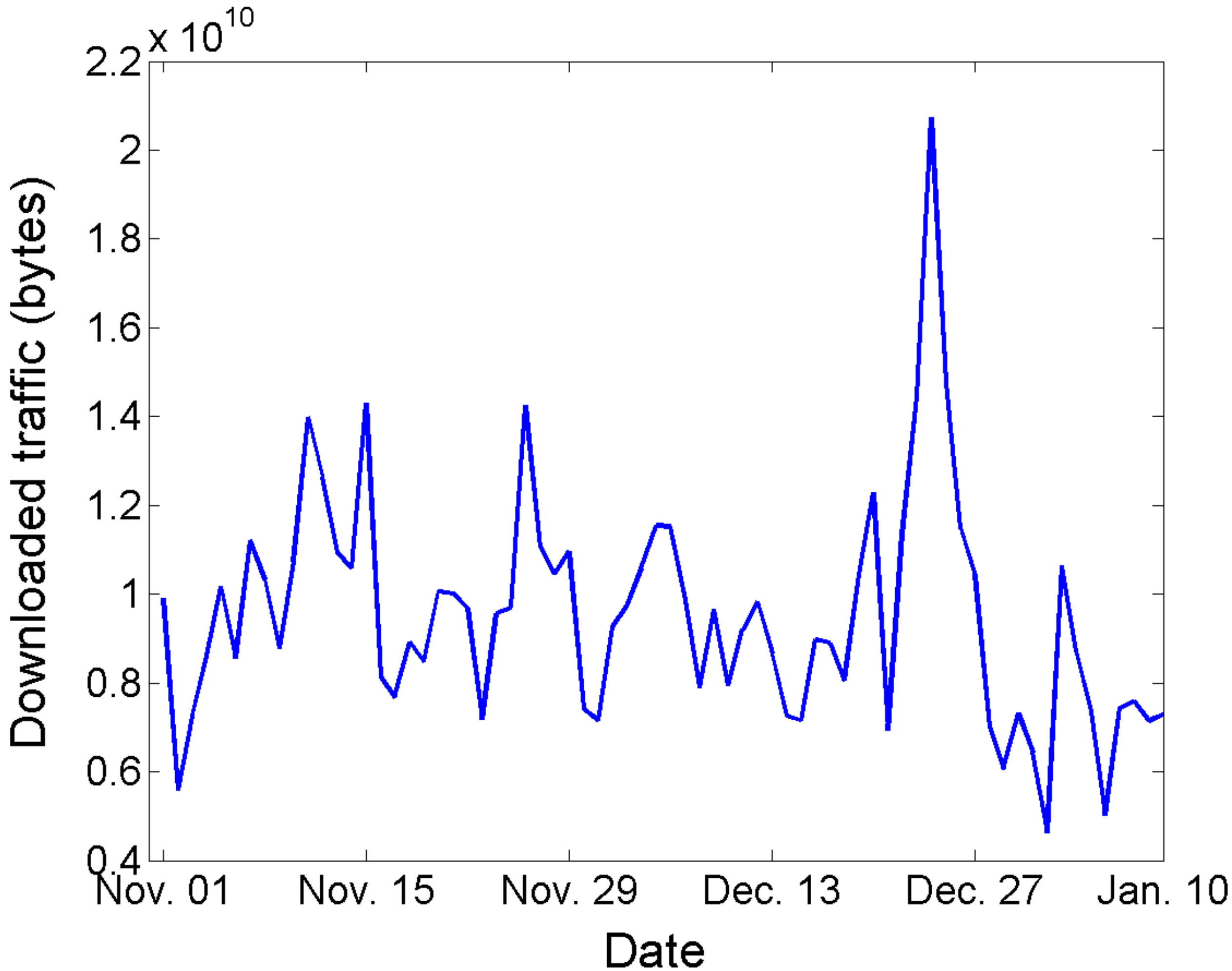
Billing records: characteristics

- 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

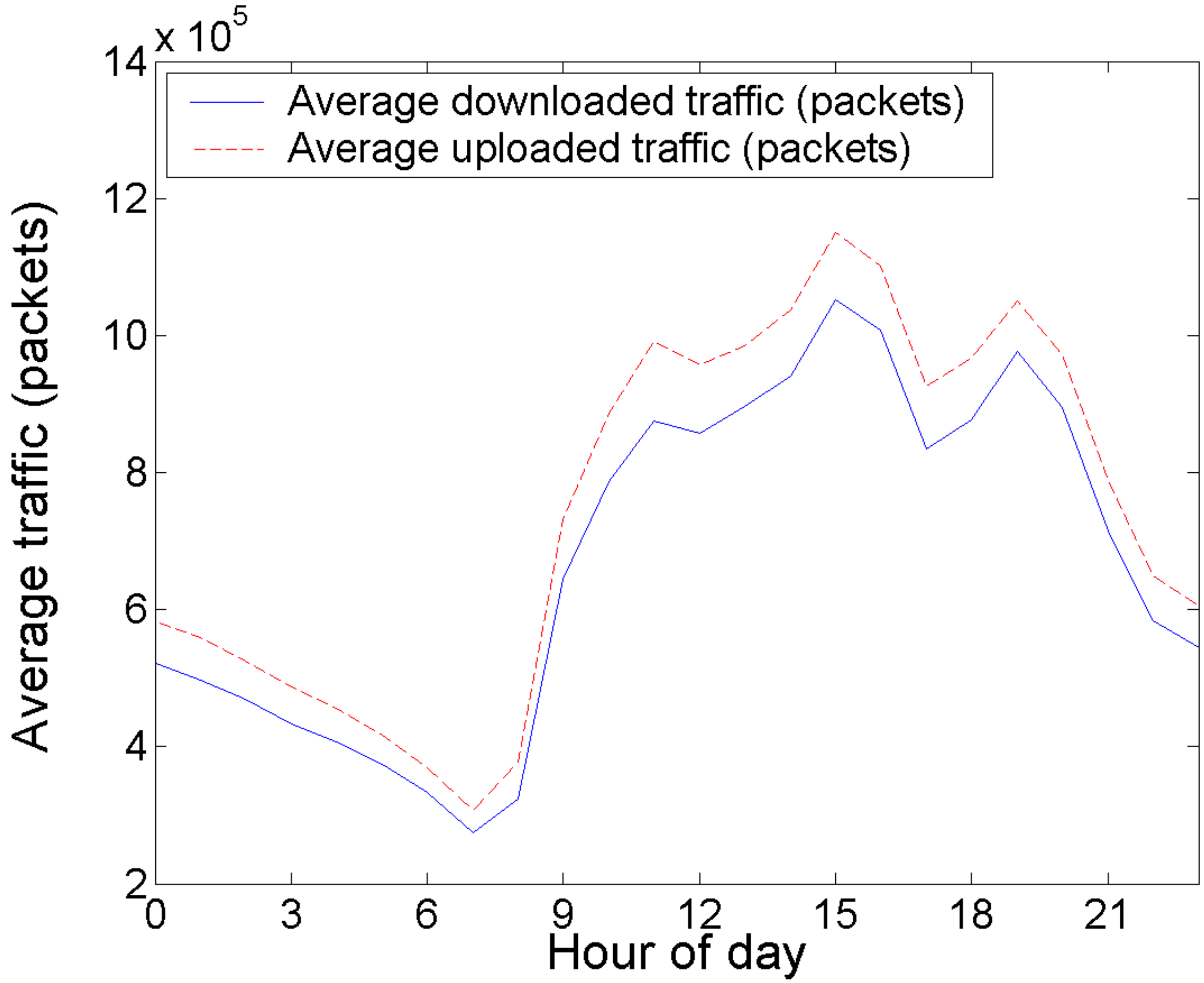
Aggregated hourly traffic



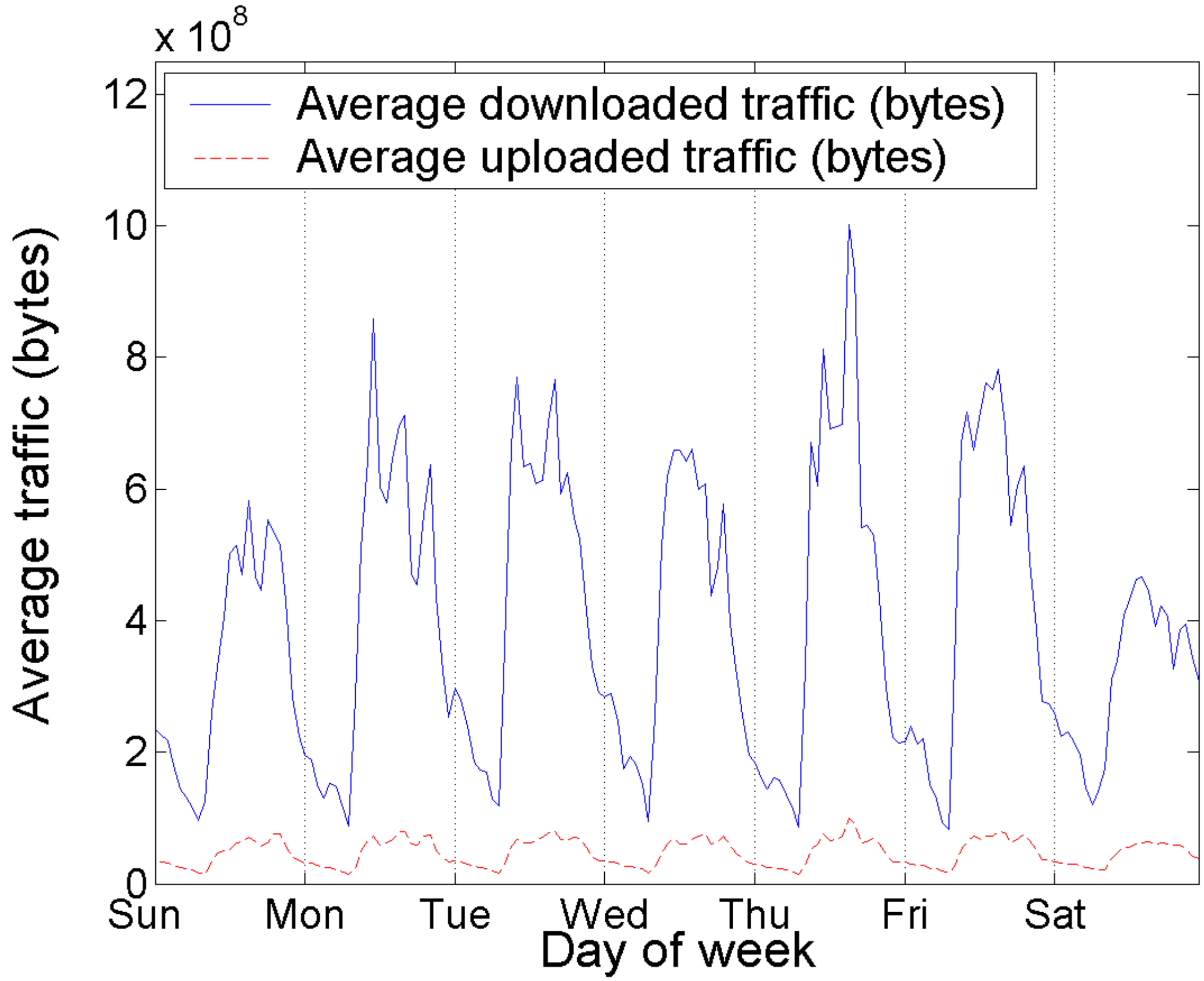
Aggregated daily traffic



Daily diurnal traffic: average downloaded bytes



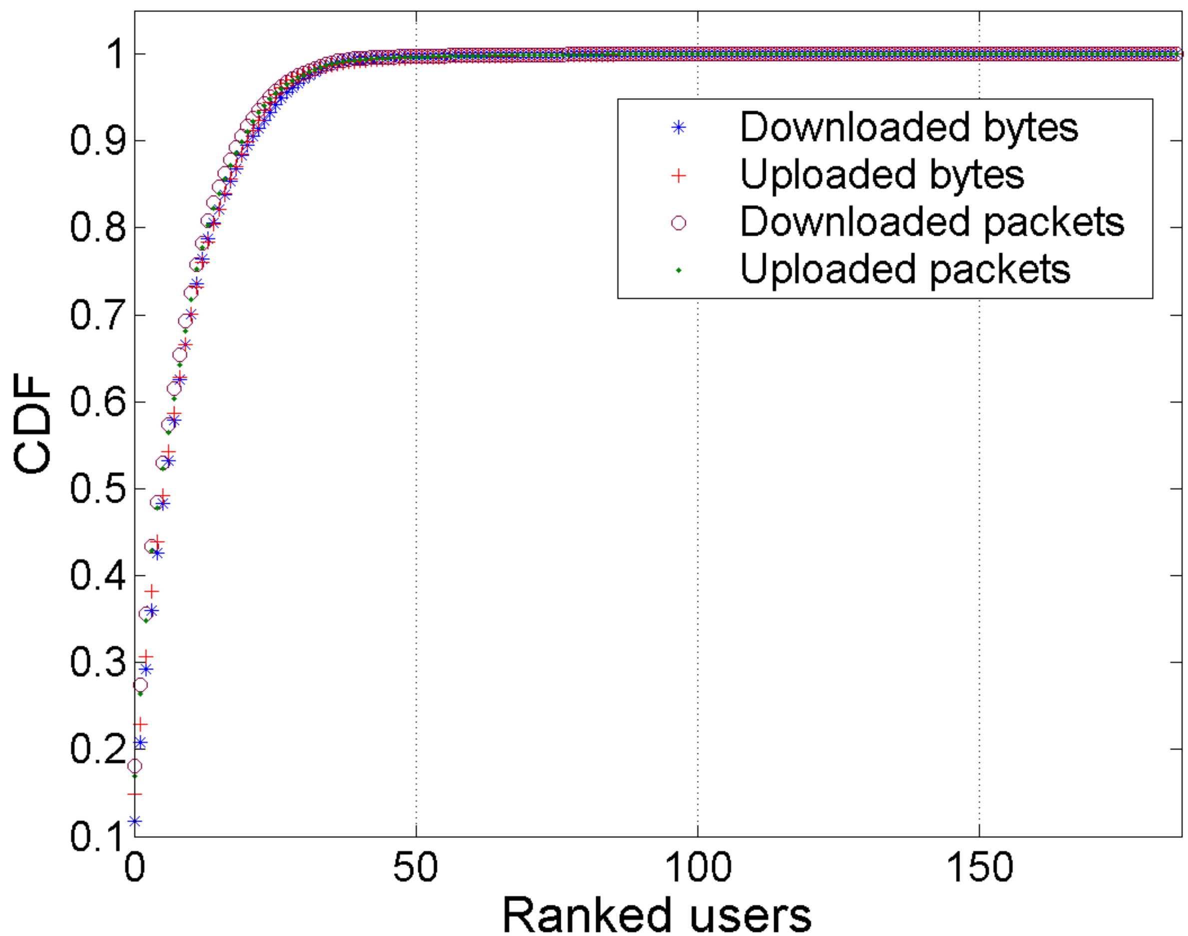
Weekly traffic: average downloaded bytes



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)

Cumulative distribution functions



k-means: clustering results

- Natural number of clusters is k=3 for downloaded and uploaded bytes
- Most users belong to the group with small traffic volume
- For k=3:
  - 159 users in group 1 (average 0.0–16.8 MB downloaded per hour)
  - 24 users in group 2 (average 16.8–70.6 MB downloaded per hour)
  - 3 users in group 3 (average 70.6–110.7 MB downloaded per hour)

Refinement: three most common traffic patterns

- Idle users:
  - rarely download/upload traffic
  - represented by zero traffic
- Active users:
  - download/upload traffic for more than 18 hours a day
  - represented by traffic over 24 hours each day
- Semi-active users:
  - download/upload traffic for 8–12 hours a day
  - represented by a cycle of 10 hours ACTIVE/14 hours IDLE cycle for each day

Refinement: clustering results

Traffic pattern	Number of users
Idle	162
Active	16
Semi-active	8
Total number of users	186

tcpdump traces

- Traces were 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

OS fingerprinting results

- Analyzed 9 hours of tcpdump trace on Dec. 14, 2002 using the open-source tool p0f v2
- Assumed constant IP addresses
- Detected 171 users:
  - 137 users did not initiate any connections and cannot be identified (no SYN packets)
  - 14 users employ Microsoft Windows
  - 2 users employ Linux
  - 1 user employs an unknown OS (identified as an MSS-modifying proxy)

OS: operating system

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

M. Arlitt and C. Williamson, "An analysis of TCP reset behaviour on the Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 1, pp. 37–44, Jan. 2005.

UDP and TCP port scans

- UDP port scans are found on UDP port 137 (NETBEUI)
- TCP port scans are found on these 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
- No HTTP(S) servers were active in the ChinaSat network
- MSSQL vulnerability was discovered on Oct. 2002, which may be the cause of scans on TCP port 1433
- The Subseven trojan is a backdoor program used with malicious intents

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

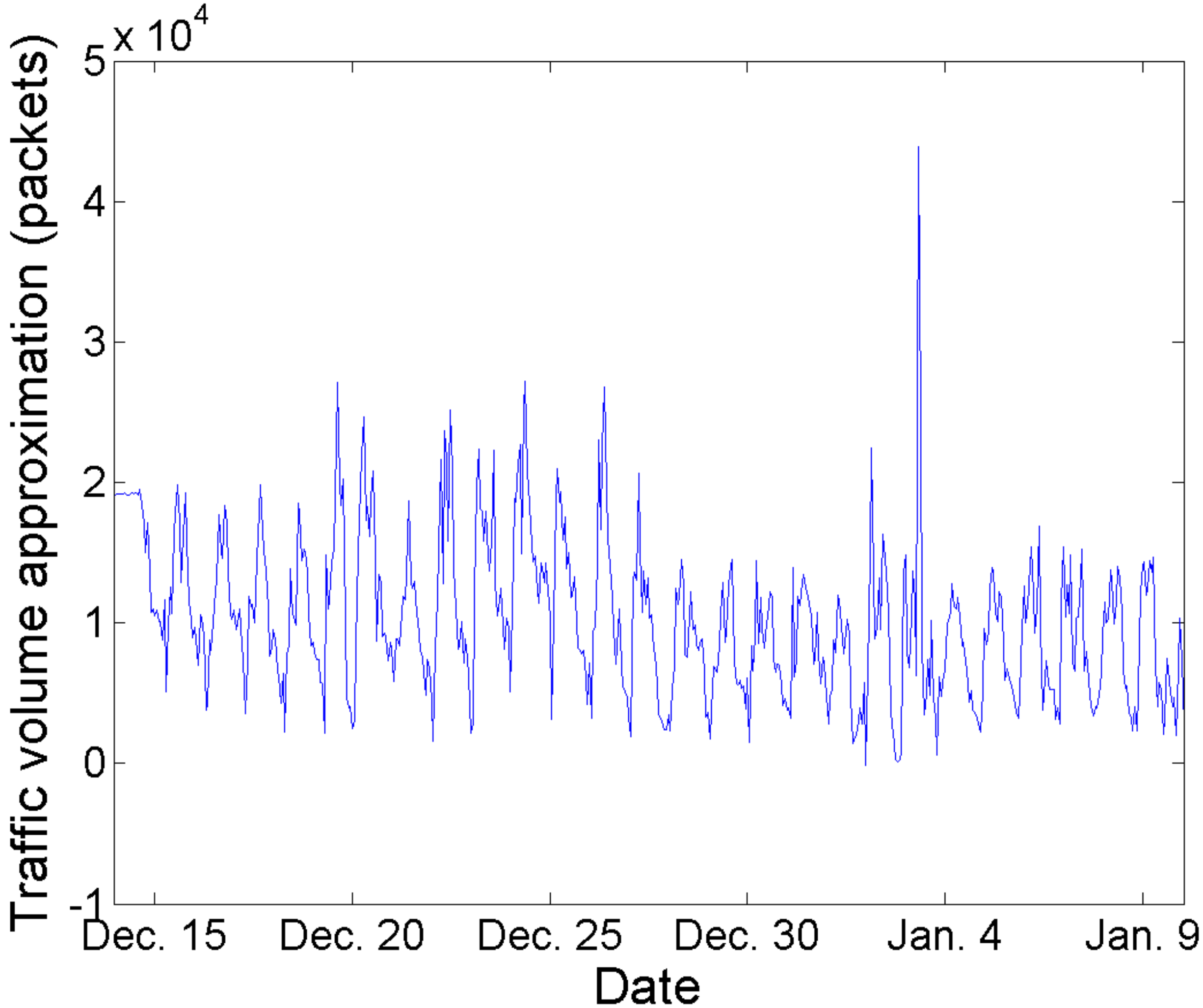
Detection of traffic volume anomalies using wavelets

- Traffic is decomposed into various frequencies using the wavelet transform
- Traffic volume anomalies are identified by the large variation in wavelet coefficient values
- The coarsest scale level where the anomalies are found indicates the time scale of an anomaly

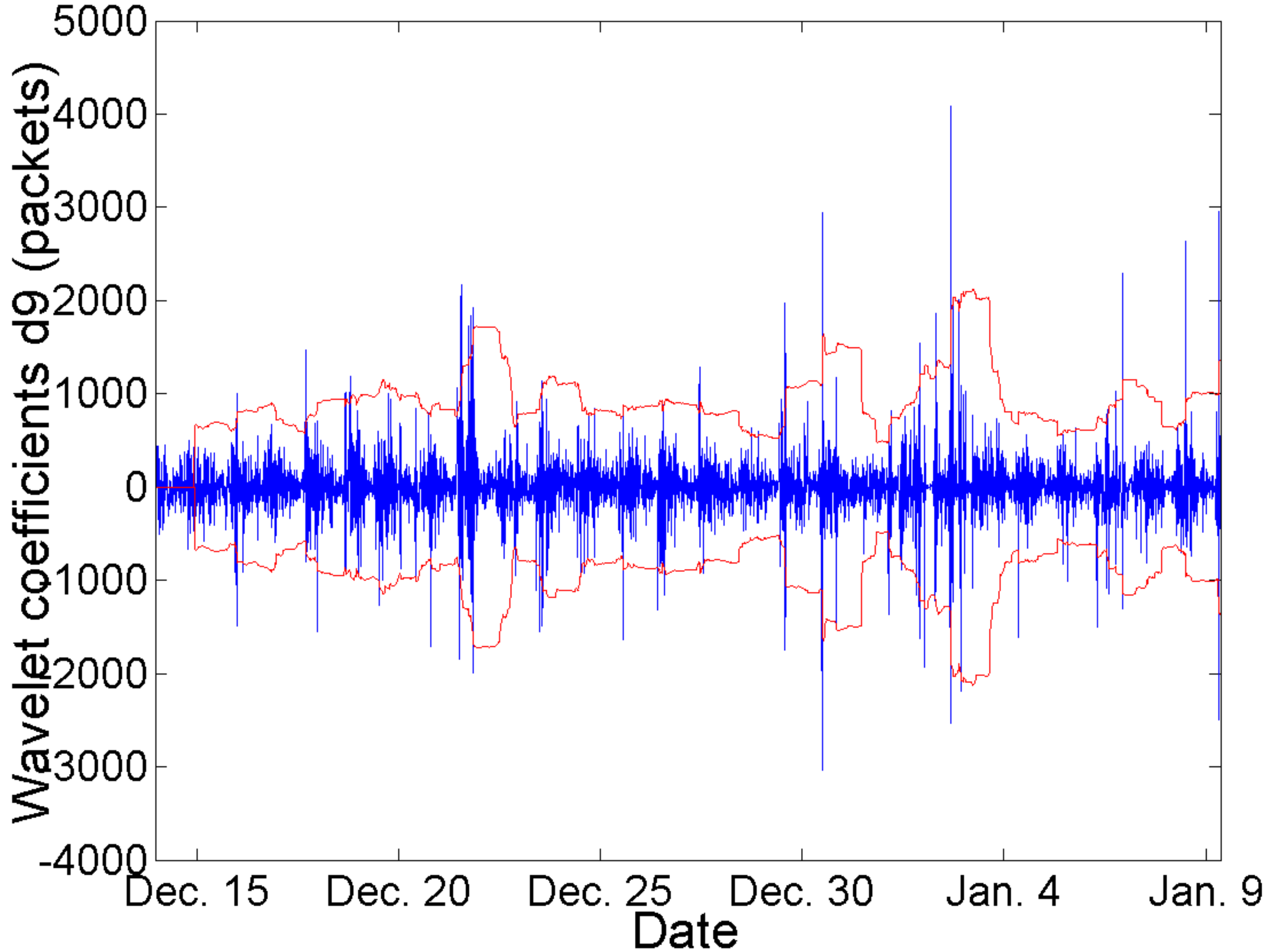
Detection of traffic volume anomalies using wavelets

- tcpdump traces are binned in terms of packets or bytes (each second)
- Wavelet transform of 12 levels is employed to decompose the traffic
- The coarsest level approximately represents the hourly traffic
- Anomalies are:
  - detected with a moving window of size 20 and by calculating the mean and standard deviation (σ) of the wavelet coefficients in each window
  - identified when wavelet coefficients lie outside the $\pm 3\sigma$ of the mean value

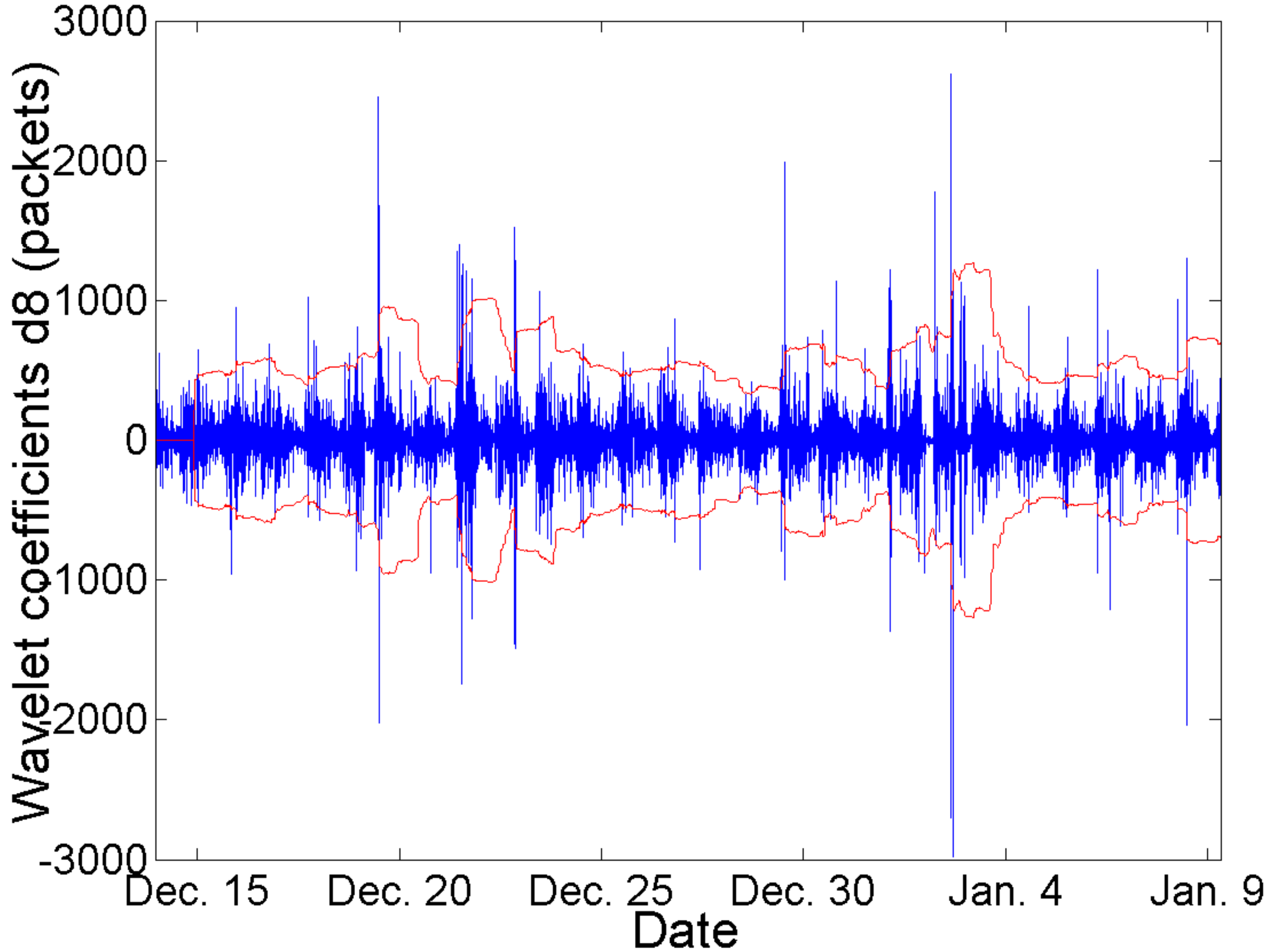
Wavelet approximation coefficients



Wavelet detail coefficients: d_9



Wavelet detail coefficients: d_8



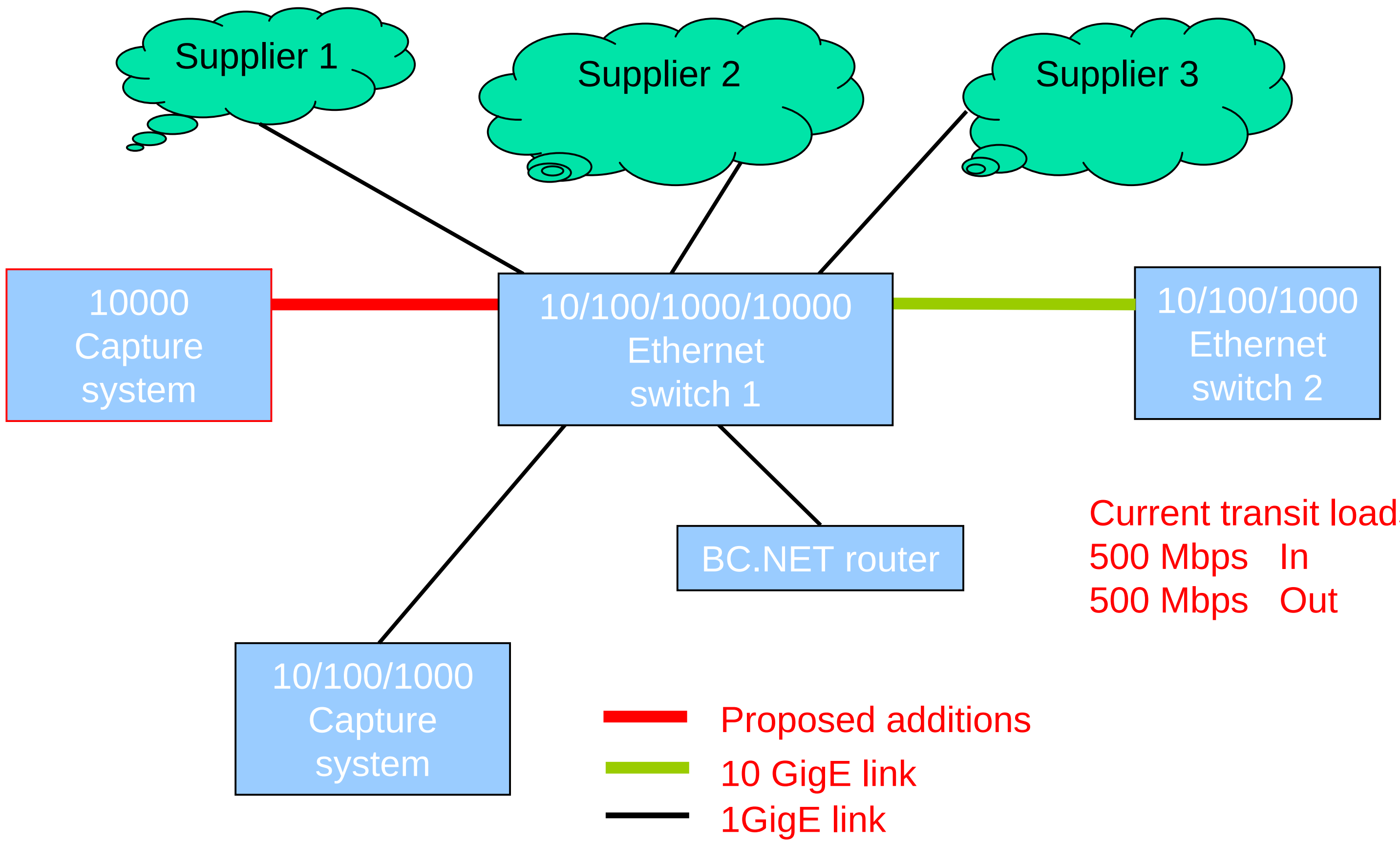
Roadmap

- 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:
- **Conclusions, future work, and references**

Conclusions

- Traffic data from deployed networks (Telus Mobility, E-Comm, ChinaSat, the Internet) were used to:
- evaluate network performance
- characterizr 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

BC.NET traffic measurements



Current transit loads:
500 Mbps In
500 Mbps Out

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.

References: self-similarity

- A. Feldmann, “Characteristics of TCP connection arrivals,” in *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., New York: Wiley, 2000, pp. 367–399.
- T. Karagiannis, M. Faloutsos, and R. H. Riedi, “Long-range dependence: now you see it, now you don't!,” in *Proc. GLOBECOM '02*, Taipei, Taiwan, Nov. 2002, pp. 2165–2169.
- W. Leland, M. Taqqu, W. Willinger, and D. Wilson, “On the self-similar nature of ethernet traffic (extended version),” *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
- M. S. Taqqu and V. Teverovsky, “On estimating the intensity of long-range dependence in finite and infinite variance time series,” in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications.* Boston, MA: Birkhauser, 1998, pp. 177–217.

References: self-similarity

- P. Abry and D. Veitch, "Wavelet analysis of long-range dependence traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 2–15, Jan. 1998.
- P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," in *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. New York: Wiley, 2000, pp. 39–88.
- P. Barford, A. Bestavros, A. Bradley, and M. Crovella, "Changes in Web client access patterns: characteristics and caching implications in world wide web," *World Wide Web*, Special Issue on Characterization and Performance Evaluation, vol. 2, pp. 15–28, 1999.
- Z. Bi, C. Faloutsos, and F. Korn, "The 'DGX' distribution for mining massive, skewed data," in *Proc. of ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, Aug. 2001, pp. 17–26.
- M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

References: time series

- G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, 2nd edition. San Francisco, CA: Holden-Day, 1976, pp. 208–329.
- P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd Edition. New York: Springer-Verlag, 2002.
- N. H. Chan, *Time Series: Applications to Finance*. New York: Wiley-Interscience, 2002.
- K. Burnham and D. Anderson, *Model Selection and Multimodel Inference*, 2nd ed. New York, NY: Springer-Verlag, 2002.
- G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

References: cluster analysis

- P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., AAAI Press/MIT Press, 1996.
- J. W. Han and M. Kamber, *Data Mining: Concepts And Techniques*. San Francisco: Morgan Kaufmann, 2001.
- T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

References: data mining

- J. Han and M. Kamber, *Data Mining: concept and techniques*. San Diego, CA:Academic Press, 2001.
- W. Wu, H. Xiong, and S. Shekhar, *Clustering and Information Retrieval*. Norwell,MA: Kluwer Academic Publishers, 2004.
- Z. Chen, *Data Mining and Uncertainty Reasoning: and integrated approach*. New York, NY: John Wiley & Sons, 2001.
- 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 Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, July. 2002.
- P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Reading,MA: Addison-Wesley, 2006, pp. 487–568.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an introduction to cluster analysis*. New York, NY: John Wiley & Sons, 1990.
- M. Last, A. Kandel, and H. Bunke, Eds., *Data Mining in Time Series Databases*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2004.
- W.-K. Ching and M. K.-P. Ng, Eds., *Advances in Data Mining and Modeling*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2003.

References: protocols

- D. E. Comer, *Internetworking with TCP/IP, Vol 1: Principles, Protocols, and Architecture*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2000.
- W. R. Stevens, *TCP/IP Illustrated (vol. 1): The Protocols*. Reading, MA: Addison-Wesley, 1994.
- J. Postel, Ed., "Transmission Control Protocol," RFC 793, Sep. 1981.
- J. Postel, "TCP and IP bake off," RFC 1025, Sep. 1987.
- J. Mogul and S. Deering, "Path MTU discovery," RFC 1191, Nov. 1990.
- V. Jacobson, R. Braden, and D. Borman, "TCP extensions for high performance," RFC 1323, May 1992.
- M. Allman, S. Floyd, and C. Partridge, "Increasing TCP's initial window," RFC 2414, Sep. 1998.
- M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, "TCP selective acknowledgment options," RFC 2018, Oct. 1996.
- M. Allman, D. Glover, and L. Sanchez, "Enhancing TCP over satellite channels using standard mechanisms," RFC 2488, Jan. 1999.
- M. Allman, S. Dawkins, D. Glover, J. Griner, D. Tran, T. Henderson, J. Heidemann, J. Touch, H. Kruse, S. Ostermann, K. Scott, and J. Semke, "Ongoing TCP research related to satellites," RFC 2760, Feb. 2000.
- J. Border, M. Kojo, J. Griner, G. Montenegro, and Z. Shelby, "Performance enhancing proxies intended to mitigate link-related degradations," RFC 3135, June 2001.
- S. Floyd, "Inappropriate TCP resets considered harmful," RFC 3360, Aug. 2002.

References: fingerprinting

- R. Beverly, "A Robust Classifier for Passive TCP/IP Fingerprinting," in *Proc. Passive and Active Meas. Workshop 2004*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 158–167.
- C. Smith and P. Grundl, "Know your enemy: passive fingerprinting," The Honeynet Project, Mar. 2002. [Online]. Available: http://www.honeynet.org/papers/finger/.
- Passive OS fingerprinting tool ver. 2 (p0f v2). [Online]. Available: http://lcamtuf.coredump.cx/p0f.shtml/.
- B. Petersen, "Intrusion detection FAQ: What is p0f and what does it do?" The SysAdmin, Audit, Network, Security (SANS) Institute. [Online]. Available: http://www.sans.org/resources/idfaq/p0f.php.
- T. Miller, "Passive OS fingerprinting: details and techniques," The SysAdmin, Audit, Network, Security (SANS) Institute. [Online]. Available: http://www.sans.org/reading room/special.php/.

References: anomalies

- P. Barford and D. Plonka, "Characteristics of network traffic flow anomalies," in *Proc. ACM SIGCOMM Internet Meas. Workshop 2001*, Nov. 2001, pp. 69–73.
- P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *Proc. ACM SIGCOMM Internet Meas. Workshop 2002*, Marseille, France, Nov. 2002, pp. 71–82.
- Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proc. ACM SIGCOMM Internet Meas. Conf. 2005*, Berkeley, CA, Oct. 2005, pp. 317–330.
- A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proc. ACM SIGCOMM Internet Meas. Conf. 2005*, Berkeley, CA, Oct. 2005, pp. 331–344.
- P. Huang, A. Feldmann, and W. Willinger, "A non-instrusive, wavelet-based approach to detecting network performance problems," in *Proc. ACM SIGCOMM Internet Meas. Workshop 2001*, San Francisco, CA, Nov. 2001, pp. 213–227.
- A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *Proc. ACM SIGCOMM Internet Meas. Conf. 2004*, Taormina, Italy, Oct. 2004, pp. 201–206.
- A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, pp. 219–230, Oct. 2004.
- M. Arlitt and C. Williamson, "An analysis of TCP reset behaviour on the Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 1, pp. 37–44, Jan. 2005.

References: spectral analysis

- M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the Internet topology,” *Proc. of ACM SIGCOMM ’99*, Cambridge, MA, Aug. 1999, pp. 251–262.
- H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, “Towards capturing representative AS-level Internet topologies,” *Proc. of ACM SIGMETRICS 2002*, New York, NY, June 2002, pp. 280–281.
- D. Vukadinovic, P. Huang, and T. Erlebach, “On the Spectrum and Structure of Internet Topology Graphs,” in H. Unger et al., editors, *Innovative Internet Computing Systems*, LNCS2346, pp. 83–96. Springer, Berlin, Germany, 2002.
- M. Mihail, C. Gkantsidis, and E. Zegura, “Spectral analysis of Internet topologies,” *Proc. of Infocom 2003,* San Francisco, CA, Mar. 2003, vol. 1, pp. 364-374.
- G. Huston, “Interconnection, peering and settlements-Part II,” *Internet Protocol Journal*, June 1999: http://www.cisco.com/warp/public/759/ipj_2-2/ipj_2-2_ps1.html.
- F. R. K. Chung, *Spectral Graph Theory*. Providence, Rhode Island: Conference Board of the Mathematical Sciences, 1997, pp. 2–6.
- M. Fiedler, “Algebraic connectivity of graphs,” *Czech. Math. J.*, vol. 23, no. 2, pp. 298–305, 1973.

References: traffic analysis

- Y. W. Chen, "Traffic behavior analysis and modeling sub-networks," *International Journal of Network Management*, John Wiley & Sons, vol. 12, pp. 323–330, 2002.
- Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE Trans. on Communications*, vol. 47, no. 7, pp. 1062–1072, July 1999.
- N. K. Groschwitz and G. C. Polyzos, "A time series model of long-term NSFNET backbone traffic," in *Proc. IEEE International Conference on Communications* (ICC'94), New Orleans, LA, May 1994, vol. 3, pp. 1400–1404.
- D. Papagiannaki, N. Taft, Z.-L. Zhang, and C. Diot, "Long-term forecasting of Internet backbone traffic: observations and initial models," in *Proc. IEEE INFOCOM 2003,* San Francisco, CA, April 2003, pp. 1178–1188.
- D. Tang and M. Baker, "Analysis of a metropolitan-area wireless network," *Wireless Networks*, vol. 8, no. 2/3, pp. 107–120, Mar.-May 2002.
- R. B. D'Agostino and M. A. Stephens, Eds., *Goodness-of-Fit Techniques*. New York: Marcel Dekker, 1986. pp. 63–93, pp. 97–145, pp. 421–457.
- F. Barceló and J. I. Sánchez, "Probability distribution of the inter-arrival time to cellular telephony channels," in *Proc. of the 49th Vehicular Technology Conference*, May 1999, vol. 1, pp. 762–766.
- F. Barcelo and J. Jordan, "Channel holding time distribution in public telephony systems (PAMR and PCS)," *IEEE Trans. Vehicular Technology*, vol. 49, no. 5, pp. 1615–1625, Sept. 2000.