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, future work, 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, future work, 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:
  - $r^{(m)}(k) = r(k), \quad k \geq 0, \quad m = 1, 2, \ldots, n,$
  - 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** \((x)\):
  
  \[
  \frac{(b(x) - a(x))}{\max \{a(x), b(x)\}}
  \]
  
  \(a(x)\) 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


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 study:
  - satellite network: *ChinaSat*
- related studies:
  - wireless network: Telus Mobility
  - packet data networks: Internet
  - public safety wireless network: E-Comm
- Conclusions, future work, 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 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**
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
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

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

![Graph showing aggregated daily traffic with peaks on December 27.](image)
Daily diurnal traffic:
average downloaded bytes

![Graph showing daily diurnal traffic with average downloaded and uploaded traffic in packets over the hour of day. The graph indicates peaks in traffic during the day, with a significant drop in the early morning hours.]
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

<table>
<thead>
<tr>
<th>Traffic pattern</th>
<th>Number of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle</td>
<td>162</td>
</tr>
<tr>
<td>Active</td>
<td>16</td>
</tr>
<tr>
<td>Semi-active</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total number of users</strong></td>
<td><strong>186</strong></td>
</tr>
</tbody>
</table>
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 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

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

- 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

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
UDP port scans direct to the ChinaSat network

- 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 ±3σ 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 network: Internet
- Conclusions, future work, and references
Conclusions

- Traffic data from deployed networks (Telus Mobility, E-Comm, ChinaSat, the Internet) were 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
Current and future projects

- Measuring traffic from BC.NET: http://www.bc.net/
  BCNET builds high-performance networks for British Columbia's research and education institutes. A not-for-profit society, BCNET is collectively funded by BC's universities, federal and provincial governments.

- Collecting user traffic and BGP data from routing tables

- Measuring equipment:
  - Endace Ninjabox 5000 (10 Gbps): 16 GB RAM, 16 TB RAID storage with write-to-disk performance of 5 Gbps
  - Endace Ninjabox 504 (1 Gbps): 8 GB RAM, 8 TB RAID storage with write-to-disk performance of 2 Gbps

BGP: border gateway protocol
BC.NET traffic measurements

Current transit loads:
500 Mbps In
500 Mbps Out

10000 Capture system
10/100/1000/10000 Ethernet switch 1
10/100/1000 Capture system
10/100 Ethernet switch 2

Supplier 1
Supplier 2
Supplier 3

Proposed additions
10 GigE link
1 GigE link

Red
Green
Black
References: downloads

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

References: self-similarity


References: self-similarity


References: time series


References: cluster analysis


References: data mining

References: protocols

References: fingerprinting

References: anomalies

References: spectral analysis


References: traffic analysis


