The UNIVERSITY of NORTH CAROLINA
at CHAPEL HILL

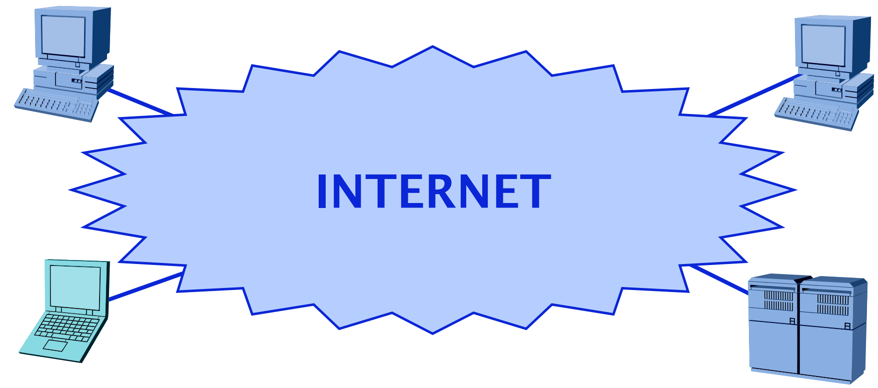
Understanding Patterns of TCP Connection Usage with Statistical Clustering

Félix Hernández-Campos
Kevin Jeffay
Don Smith
Department of Computer Science

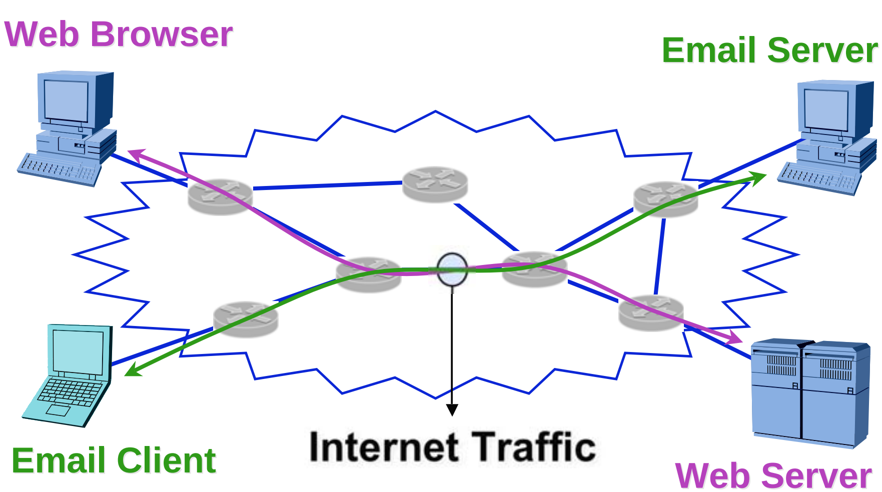
Andrew Nobel
Department of Statistics

http://www.cs.unc.edu/Research/dirt

Motivation
Modeling Internet Traffic



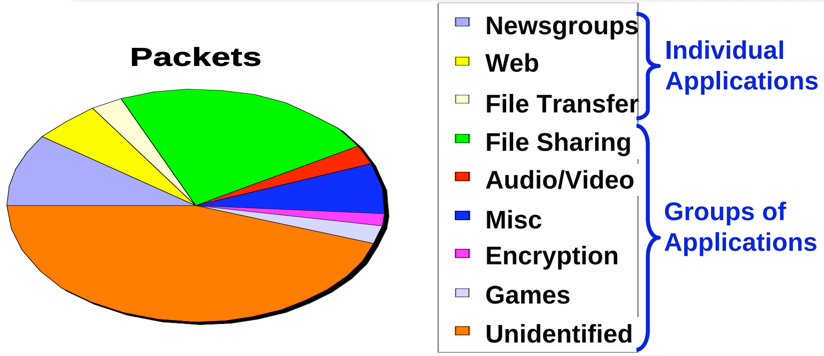
Motivation
Modeling Internet Traffic



Motivation
Experimental Networking Research

- Evaluating network technologies requires *realistic experiments* in a controlled laboratory environment
- A key component of these experiments is the *traffic workload*
  - Traffic is created by distributed applications running at the end hosts
- A natural approach for traffic generation is to simulate these applications using models of their behavior
  - This is known as *source-level modeling*

Internet Traffic Mixes
Internet2 Applications (Nov 4 2002)



- Dozens of different applications are commonly used
- There is a large percentage of unidentified traffic

Difficulties in Source-Level Modeling

- *Real* Internet traffic is the result of aggregating many individual applications into a *traffic mix*
- Requires protocol specifications
  - Closed applications have to be reverse engineered
- Applications change quickly
- Privacy considerations complicate data acquisition

➢ It is simply infeasible to develop models for each application and maintain them up to date

Modeling of Internet Traffic Mixes
Goals

- Develop source-level models of traffic mixes
  - Easy to populate and update
  - Derived from very large data sets

➢ Model communication patterns in an abstract manner
  - ***Application-independent source-level modeling***

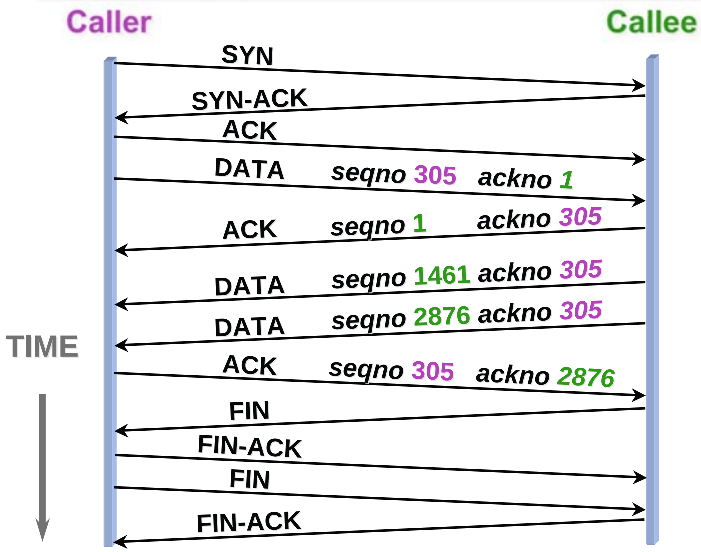
- Construct flexible traffic generators
  - Reproduce a wide range of traffic mixes

➢ Find the fundamental patterns of communication
  - ***Cluster-based traffic generation***

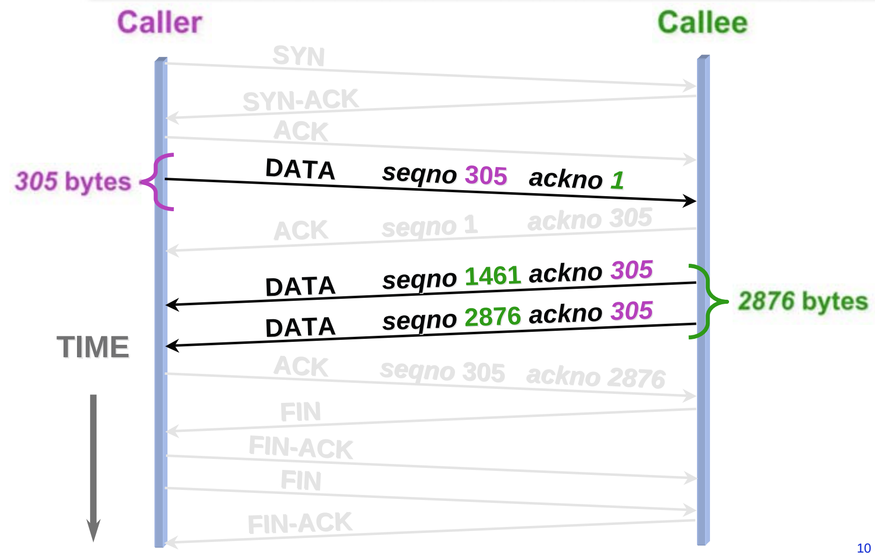
Our Approach
Finding Patterns in TCP Connections

- Modeling of data exchange patterns in TCP connections
  - Application-independent, network-independent
- Statistical clustering of TCP connection patterns
  - Find the fundamental subpopulations
  - Construct empirical or parametric models of subpopulations
- Development of new, flexible traffic generators
  - Cluster-based synthetic traffic
- Validation
  - Compare synthetic traffic with some *gold standard*

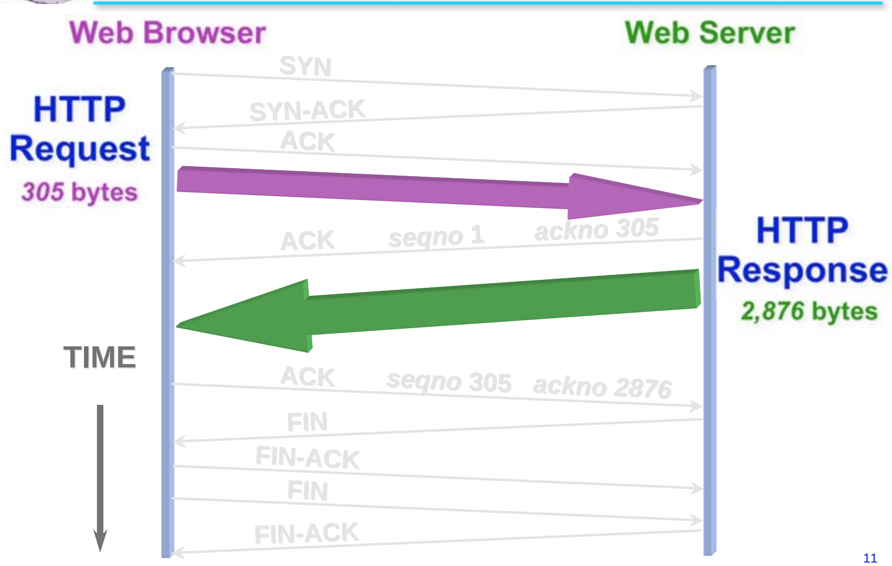
Modeling of Data Exchange Patterns
ADU Inference from TCP Packet Headers



Modeling of Data Exchange Patterns
ADU Inference from TCP Packet Headers

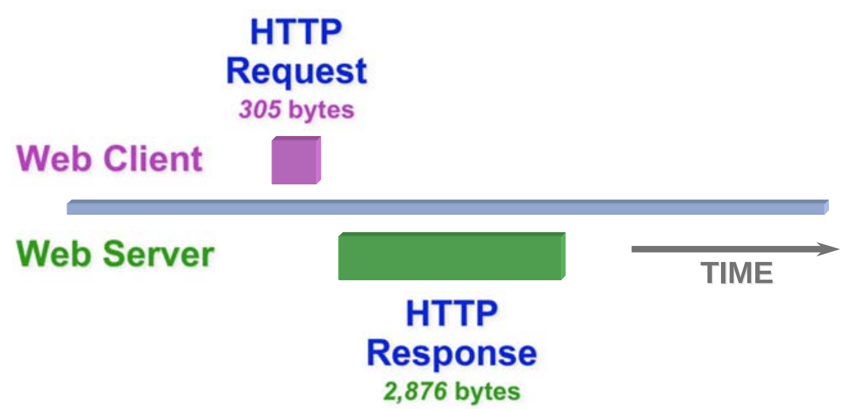


Modeling of Data Exchange Patterns
ADU Inference from TCP Packet Headers



Modeling of Data Exchange Patterns
HTTP Connection (Web Traffic)

- Communication pattern was (a_1, b_1)
  - *E.g.*, (305 bytes, 2,876 bytes)



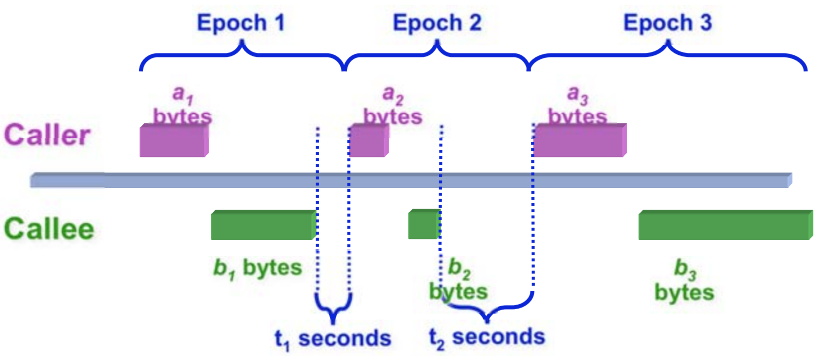
Abstract Communication Model

The *a-b-t* connection vector model

- General model (*a-b-t* vector):

$$((a_1, b_1, t_1), (a_2, b_2, t_2), \ldots, (a_e, b_e, \perp))$$

where e is the number of epochs

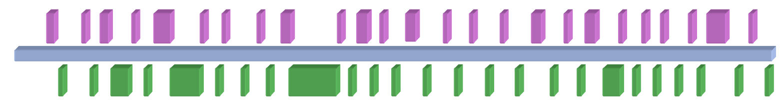


a-b-t Connection Vectors

Typical Communication Patterns

- SMTP (send email)

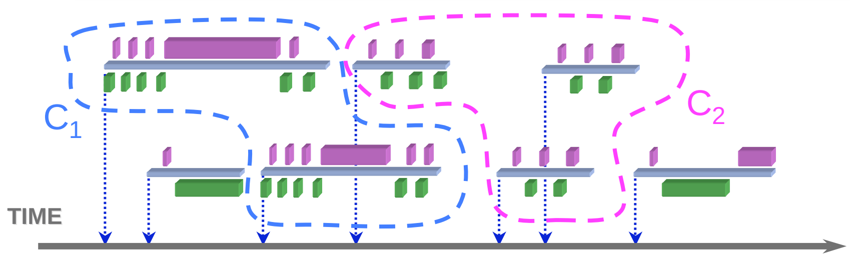
- Telnet (remote terminal)



- FTP-DATA (file download)

a-b-t Connection Vectors

Clustering communication patterns



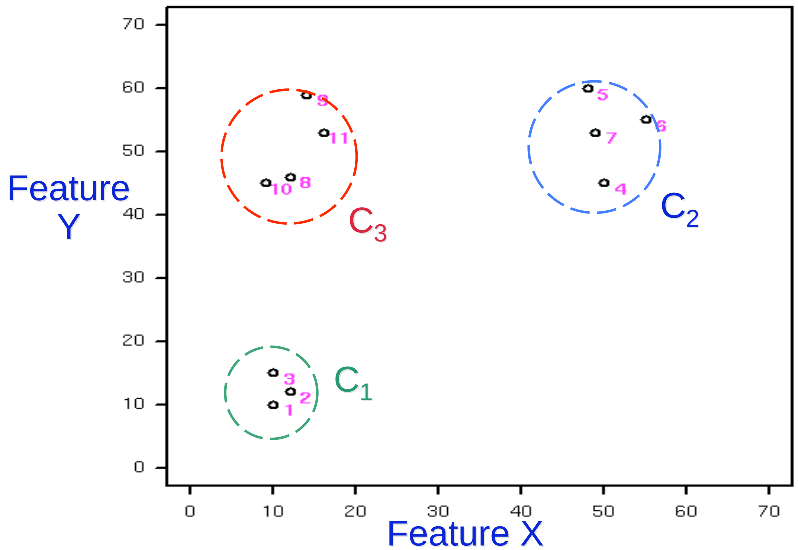
- Find statistically homogeneous communication patterns
  - Study this *mixture of populations*
- Address scalability using *statistical clustering*

Clustering Communication Patterns

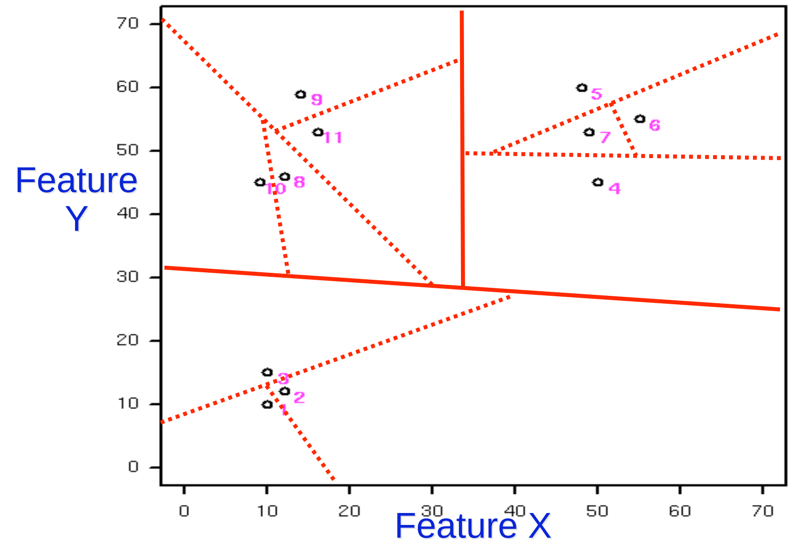
Clustering 101

- Procedure that divides a given set of feature vectors into disjoint groups, or clusters, $C_1, C_2, \ldots, C_m$
- The goals of clustering schemes:
  - Clusters are small and mutually far apart
  - Clustering is done automatically
    - » Clustering is a form of unsupervised learning
- Statistical clustering is a well founded technique
  - Successfully applied to Gene Micro-array classification, Data Mining,…

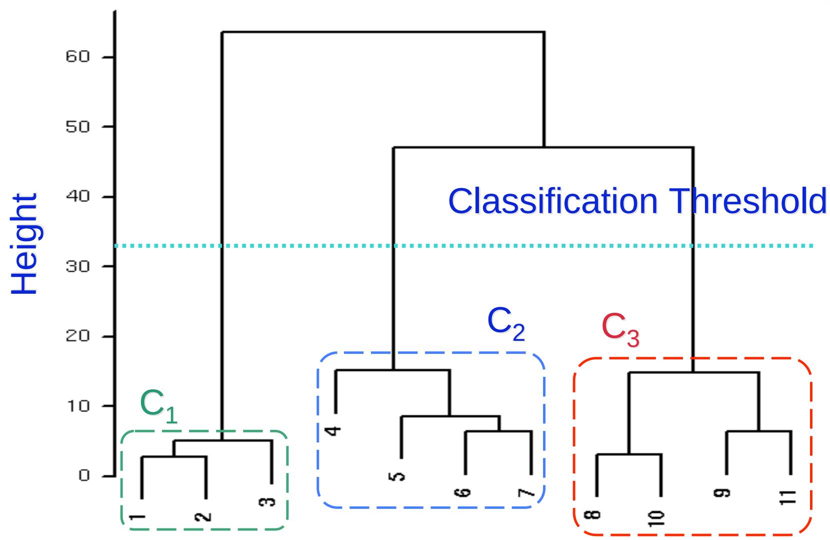
Example
Clusters in a 2D Data Set



Example
Divisive Hierarchical Clustering



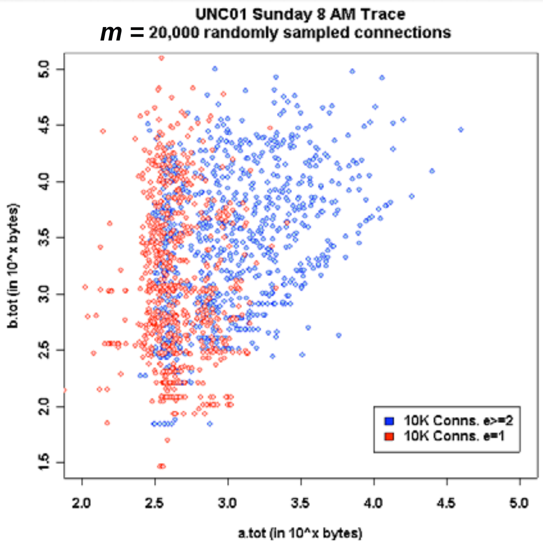
Divisive Hierarchical Clustering
Dendrogram



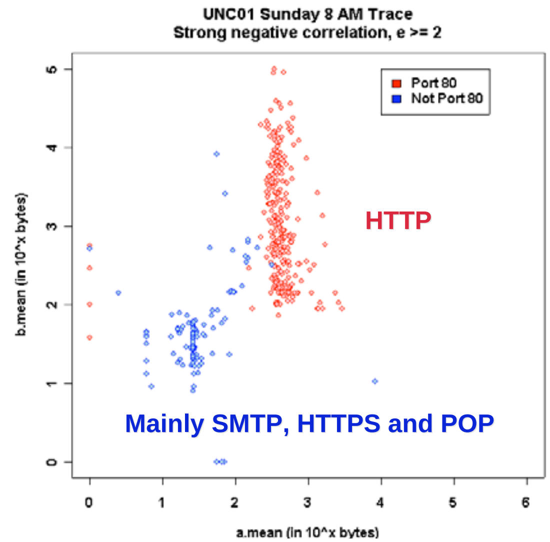
Statistical Features of *a-b-t* Connection Vectors

UNIVARIATE				MULTIVARIATE		
a_{tot}	b_{tot}	t_{tot}	Total bytes/time	*cor.a.b*	*cor.a.t*	*cor.b.t*
a_{max}	b_{max}	t_{max}	Max bytes/time	Correlations		
a_{min}	b_{min}	t_{min}	Min bytes/time	*cor.a.b.x*	*cor.a.t.x*	*cor.b.t.x*
a_{mean}	b_{mean}	t_{mean}	Mean bytes/time	Lagged Correlations		
a_{xq}	b_{xq}	t_{xq}	1st 2nd 3rd Quartiles	*crc.a.b*	*crc.a.t*	*crc.b.t*
a_{stdev}	b_{stdev}	t_{stdev}	Standard Deviation	Cross-correlations		
$a_{cor.x}$	$b_{cor.x}$	$t_{cor.x}$	Autocorrelations	*dir1.a.b*	*dir2.a.b*	
a_{hx}	b_{hx}	t_{hx}	Homogeneity	Directionality		
a_{vs}	b_{vs}	t_{vs}	Total Variation	UNIVARIATE		
a_{vm}	b_{vm}	t_{vm}	Max First Diff.	*e*	No. of Epochs	

Clustering Connections
Statistical structure in data exchanges



Clustering Connections
Example of two clusters



Clustering Communication Patterns
Data Set

- Each feature is approximately normalized to [0,1]
  - Many features have heavy-tailed distributions

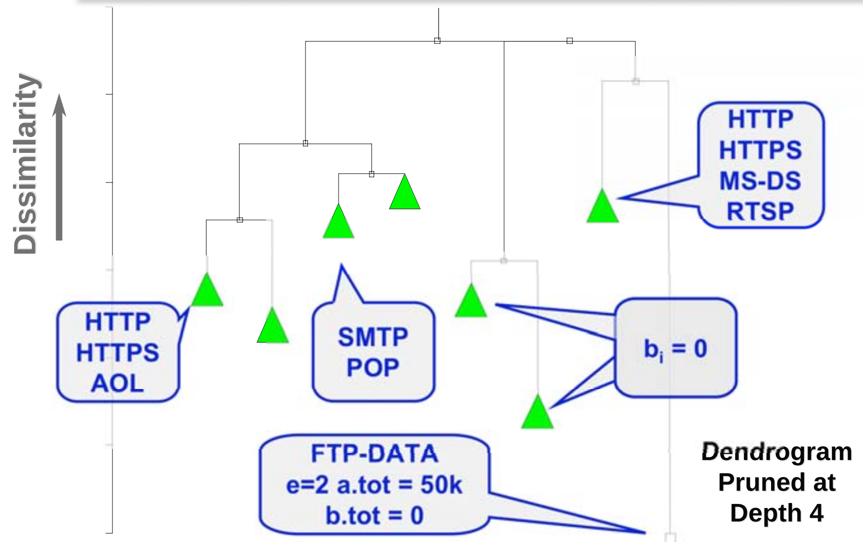
Features / Observations	e	a.max	a.min	…	dir2.a.b
Connection 1	0.66	0.23	0.12	…	0.61
Connection 2	0.24	1.03	0.45	…	0.23
…	…				
Connection *m*	0.11	0	0	…	1

Example 1
Divisive Hierarchical Clustering

- Packet header trace collected from UNC main Internet access link
  - April 2002
- Random sample of 5,000 connections
  - $e \geq 2$
- Analysis performed using R's implementation
  - Using the `diana` algorithm
- Euclidean distance

26 Features			
e			No. of Epochs
a_{tot}	b_{tot}		Total bytes/time
a_{max}	b_{max}	t_{max}	Max bytes/time
a_{min}	b_{min}		Min bytes/time
$a_{\mu,\sigma}$	$b_{\mu,\sigma}$		1st 2nd Moments
a_{xq}	b_{xq}		1st 2nd 3rd Quartiles
a_{vs}	b_{vs}		Total Variation
a_h	b_h		Max/Min Ratio
r_a	r_b		Lag-1 Autocorr.
$\rho_1(a\text{'s}, b\text{'s})$			Spearman's Correl.
$\rho_2(b\text{'s}, a\text{'s})$			Lag-1 Cross Corr.

Example 1
Dendrogram

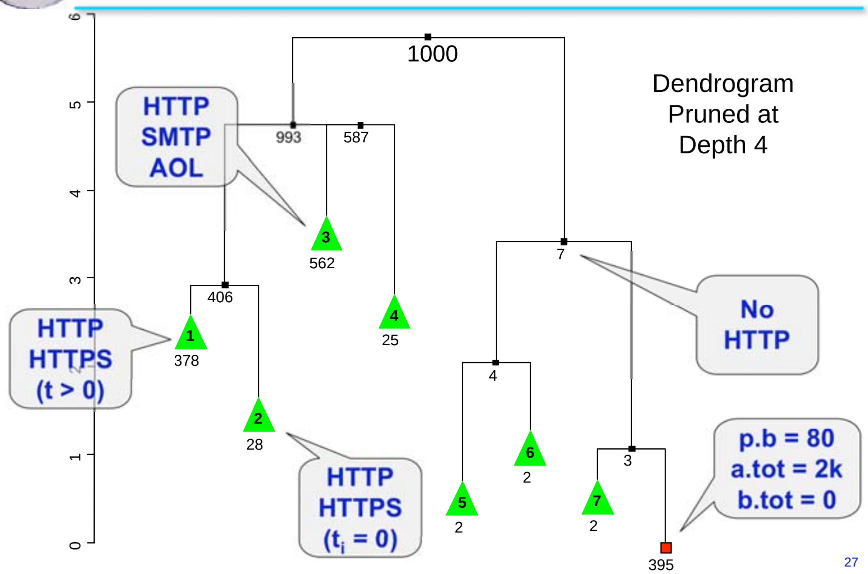


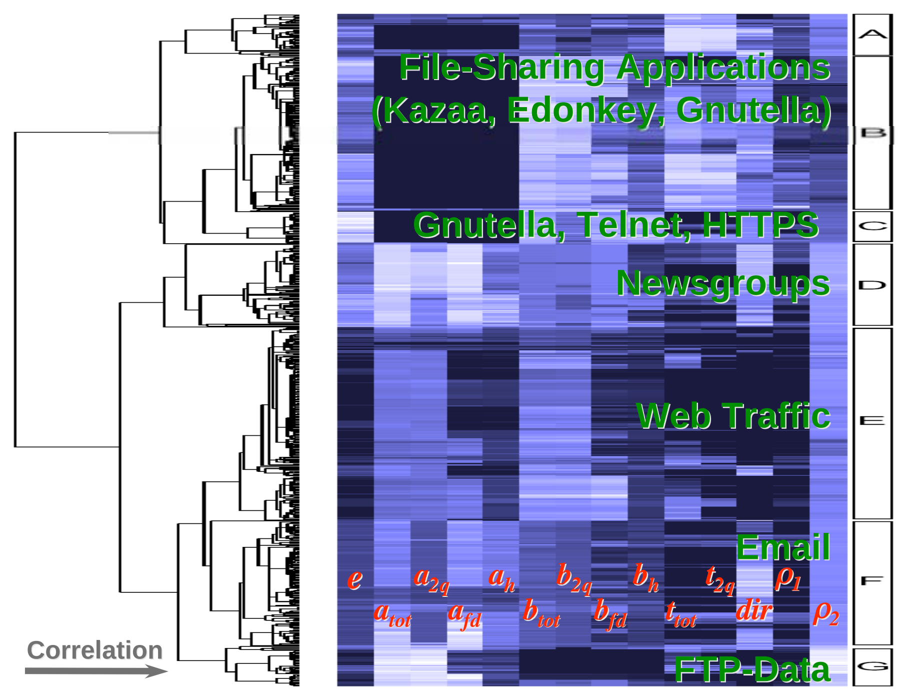
Example 2
Agglomerative Hierarchical Clustering

- Packet header trace collected from an Internet2 backbone link (Abilene-I data set)
  - August 2002
- Sample of 717 connections
  - $e \geq 2$
- Analysis performed using Eisen's software
  - Developed for ***microarrays***
- Pearson's correlation as distance metric

14 Features			
e			No. of Epochs
a_{tot}	b_{tot}	t_{tot}	Total bytes/time
a_{2q}	b_{2q}	t_{2q}	2nd Quartiles
a_{fd}	b_{fd}		Max First Diff.
a_{hx}	b_{hx}		Max/Min Ratio
dir			log (a_{tot} / b_{tot})
$\rho_1(a\text{'s}, b\text{'s})$			Spearman's Correl.
$\rho_2(a\text{'s}, b\text{'s})$			Lag-1 Sp. Corr.

Hierarchical Clustering
UNC01 1,000 Connection Sample





Summary and Current Work

- Developed an application-independent model of TCP communication patterns: the *a-b-t connection vector model*
  - Suitable for large scale data acquisition
- Applied statistical clustering to uncover fundamental subpopulations
  - Working on a *systematic approach* for feature selection and cluster identification (*i.e.* dendrogram pruning)
  - $O(n^2)$ is too slow, so we are also looking into data mining algorithms for clustering
- A synthetic traffic generator ("tmix") for reproducing TCP application workloads
  - Network specific workloads easily modeled given a packet header trace

The UNIVERSITY *of* NORTH CAROLINA
at CHAPEL HILL

Understanding Patterns of TCP Connection Usage with Statistical Clustering

Félix Hernández-Campos
Kevin Jeffay
Don Smith
Department of Computer Science

Andrew Nobel
Department of Statistics

http://www.cs.unc.edu/Research/dirt