

The University of North Carolina at Chapel Hill

Talk at Interface 2003

March 13, 2003

**Statistical Clustering of Internet
Communication Patterns**

Félix Hernández-Campos

(UNC-Chapel Hill Computer Science)

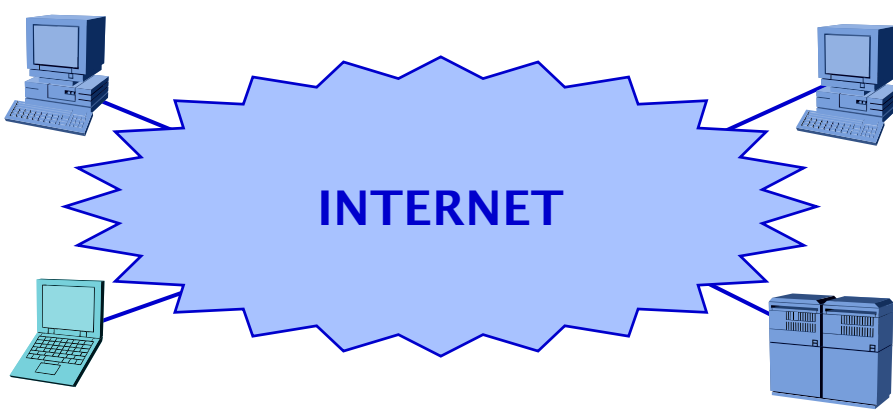
Joint work work with:

Andrew Nobel
(UNC-CH Statistics)

Don Smith *Kevin Jeffay*
(UNC-CH Computer Science)

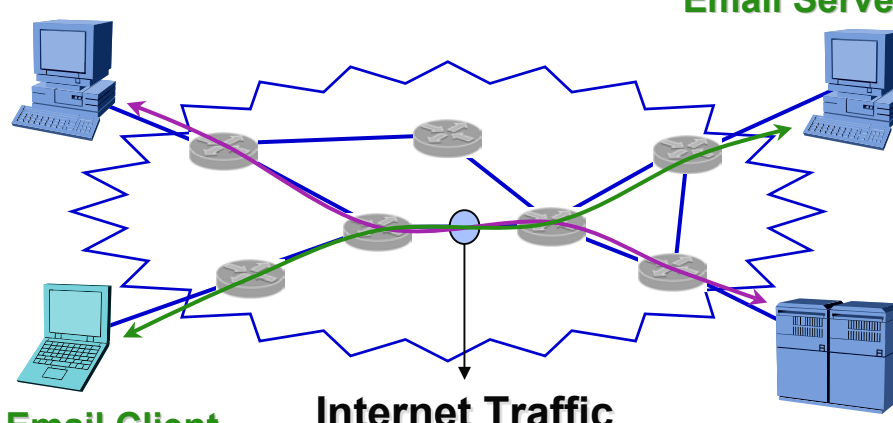


Motivation
Modeling Internet Traffic



Motivation

Modeling Internet Traffic



The diagram illustrates a network topology for modeling Internet traffic. It features a central hub of several interconnected routers. Four end hosts are connected to this network: a Web Browser (top left), an Email Server (top right), an Email Client (bottom left), and a Web Server (bottom right). Colored arrows represent traffic flows: a purple arrow from the Web Browser to the Web Server, a green arrow from the Email Server to the Email Client, and a red arrow from the Web Server to the Email Client. A blue arrow also shows traffic from the Web Browser to the Email Client. The central routers are connected by a mesh of blue lines, and a central node is labeled 'Internet Traffic' with a downward-pointing arrow.

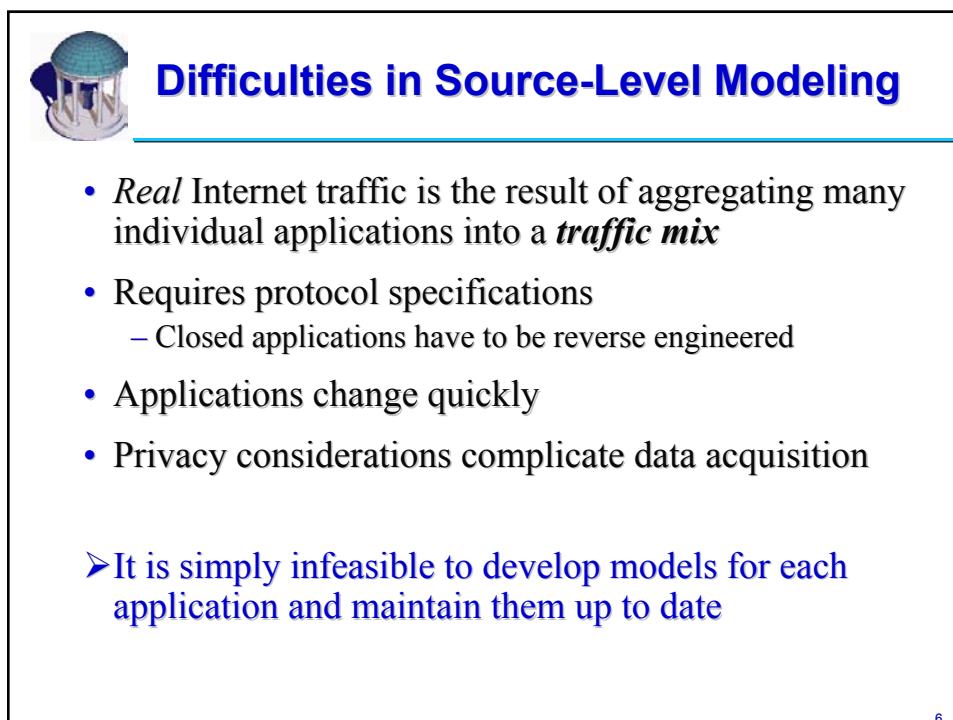
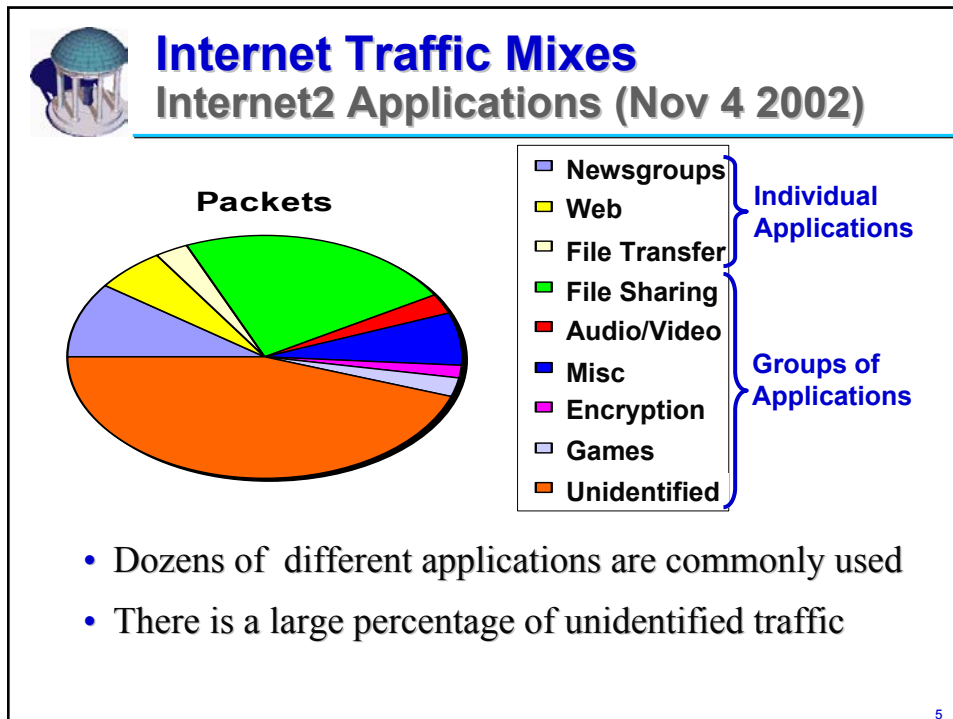
3

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*

4



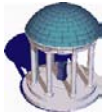


Goals

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

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

7

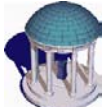


Our Approach

- 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*

8

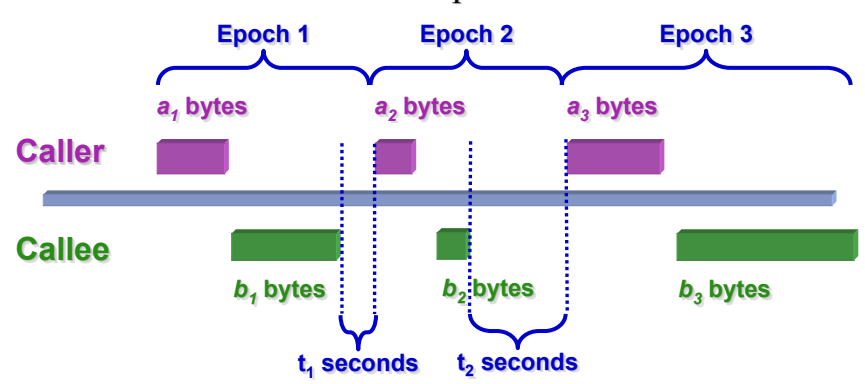


Abstract Communication Model


The *a-b-t* Model

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

$$((a_1, b_1, t_1), (a_2, b_2, t_2), \dots, (a_e, b_e, \perp))$$
 where e is the number of epochs


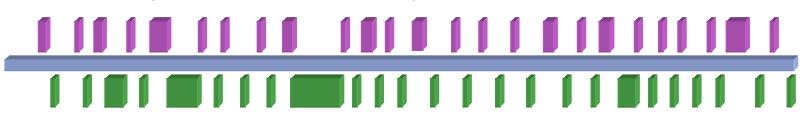



13




The *a-b-t* Model

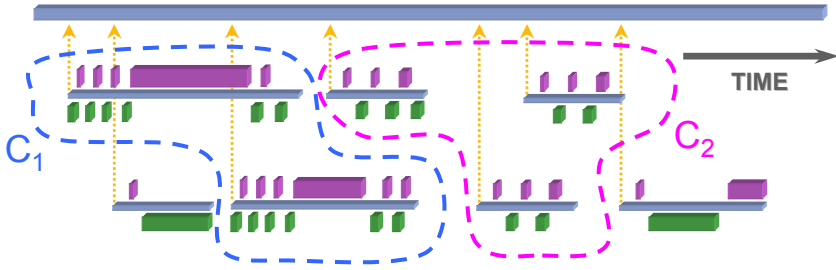
Typical Communication Patterns

- SMTP (send email)**

- Telnet (remote terminal)**

- FTP-DATA (file download)**


14




Clustering Communication Patterns



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

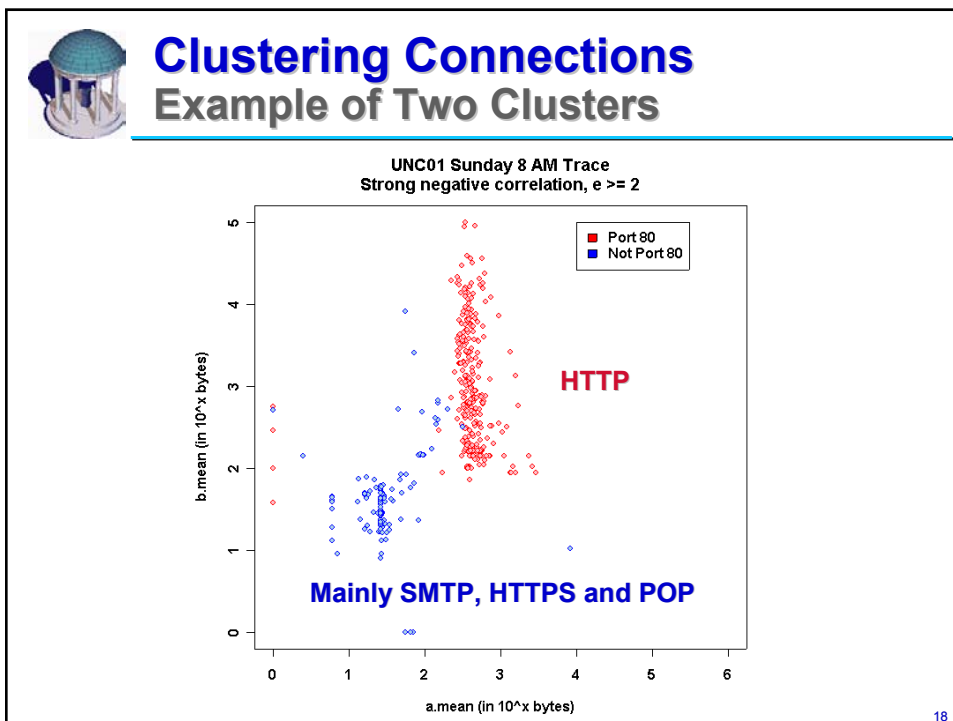
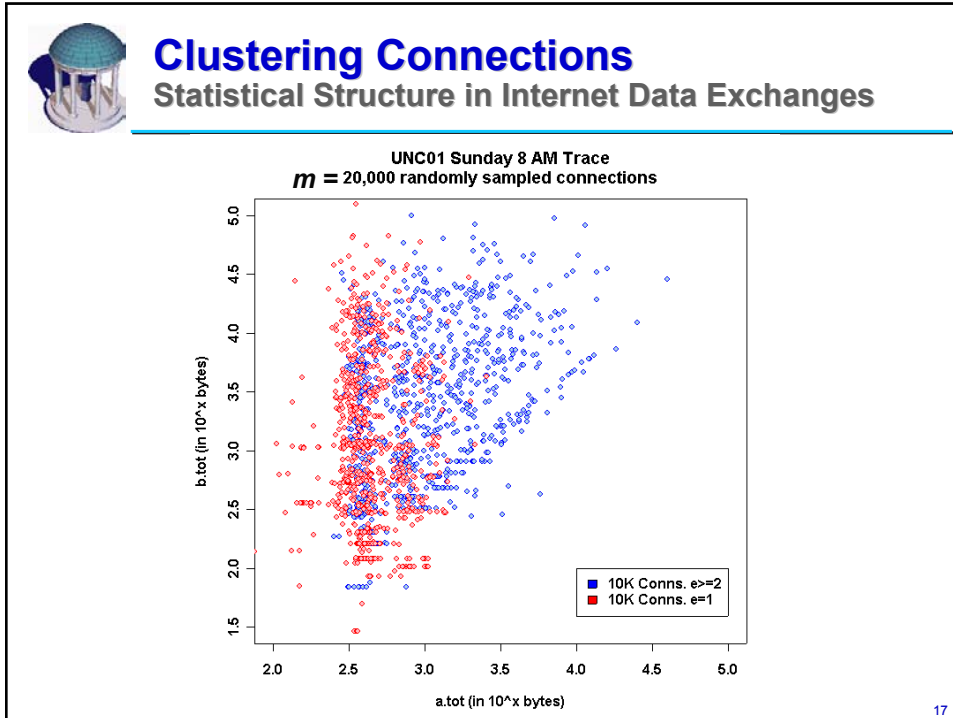
15

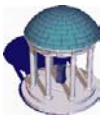


Statistical Features of an A-b-t Vector

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}	1 st 2 nd 3 rd 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	

16



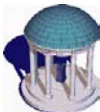


Clustering Communication Patterns Data Set

- 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 <i>m</i>	0.11	0	0	...	1

19

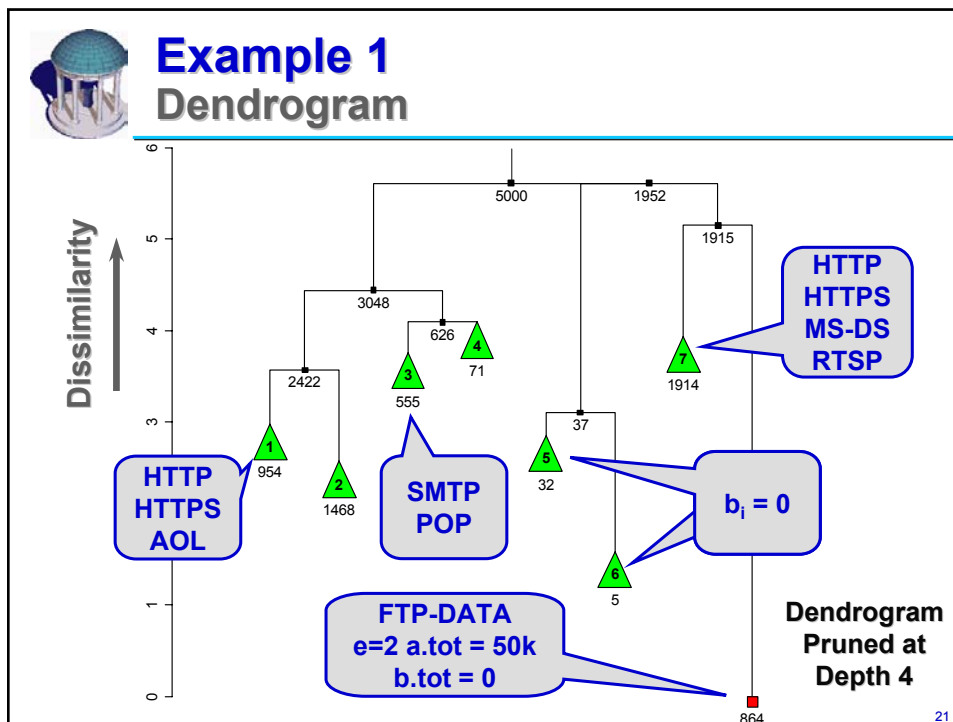


Example 1 Divisive Hierarchical Clustering

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

20

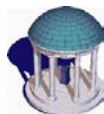
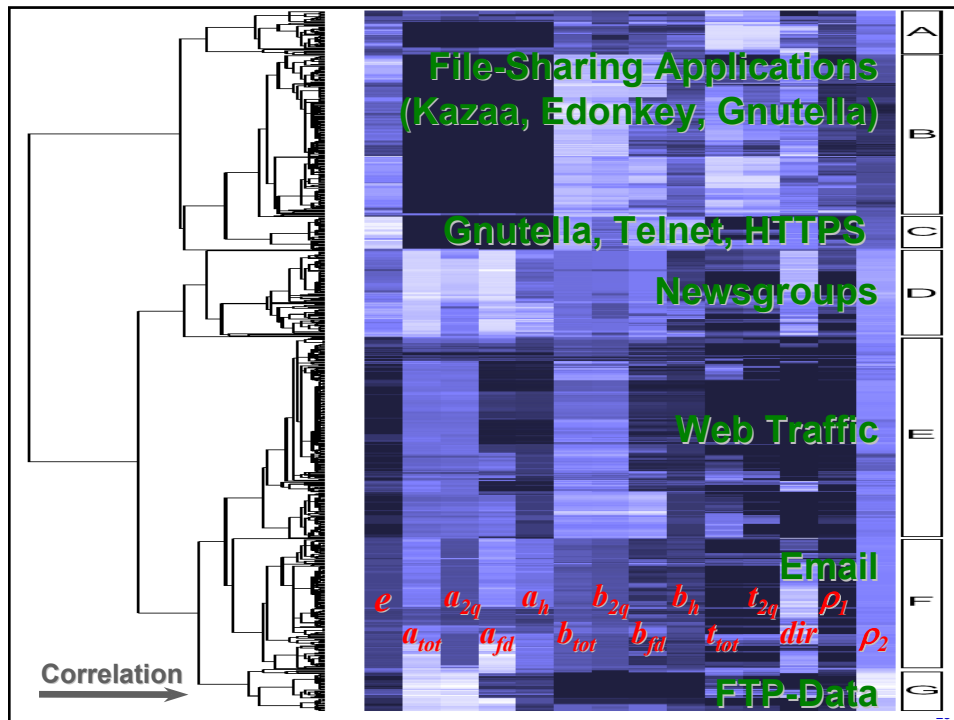


Example 2 Agglomerative Hierarchical Clustering

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

22



Summary and Current Work

- Developed an application-independent model of Internet communication patterns
 - 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