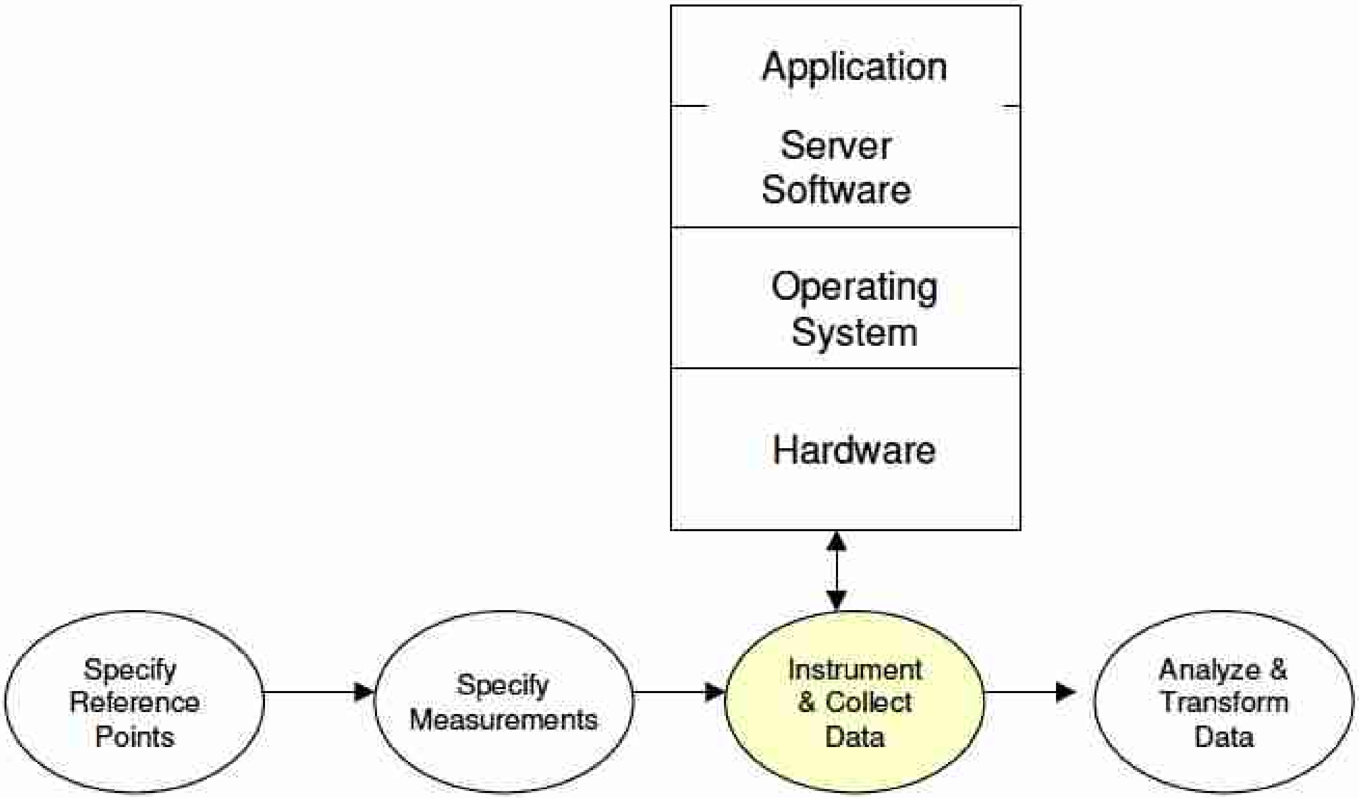
Métodos Quantitativos para Ciência da Computação Experimental

Caracterização de Cargas

Jussara Almeida
DCC-UFMG
2017

Framework



Application
Server
Software
Operating
System
Hardware
Specify
Reference
Points
Specify
Measurements
Instrument
& Collect
Data
Analyze &
Transform
Data

Data Collection Tools

- Hardware monitors
- Software monitors
- Accounting systems
- Program analyzers
- Logs: typical format of the access log of a Web server includes:

```
hostname -- [dd/mm/yyy:hh:mm:ss:zz tz] request
status bytes
```

Web Access Log

```
perf.xyz.com - - [24/Jan/19xx:13:41:41 -0400] "GET i.html HTTP/1.0" 200 3185
perf.xyz.com - - [24/Jan/19xx:13:41:41 -0400] "GET 1.gif HTTP/1.0" 200 1210
h0.south.com - - [24/Jan/19xx:13:43:13 -0400] "GET i.html HTTP/1.0" 200 3185
h0.south.com - - [24/Jan/19xx:13:43:14 -0400] "GET 2.gif HTTP/1.0" 200 2555
h0.south.com - - [24/Jan/19xx:13:43:15 -0400] "GET 3.gif HTTP/1.0" 200 36403
h0.south.com - - [24/Jan/19xx:13:43:17 -0400] "GET 4.gif HTTP/1.0" 200 441
cs.uni.edu - - [24/Jan/19xx:13:46:45 -0400] "GET i.html HTTP/1.0" 200 3185
cs.uni.edu - - [24/Jan/19xx:13:46:45 -0400] "GET 2.gif HTTP/1.0" 200 2555
cs.uni.edu - - [24/Jan/19xx:13:46:47 -0400] "GET 3.gif HTTP/1.0" 200 36403
cs.uni.edu - - [24/Jan/19xx:13:46:50 -0400] "GET 4.gif HTTP/1.0" 200 98995
sys1.world.com - - [24/Jan/19xx:13:48:29 -0400] "HEAD index.html" 400 -
```

Web Access Log: an example

Workload characterization

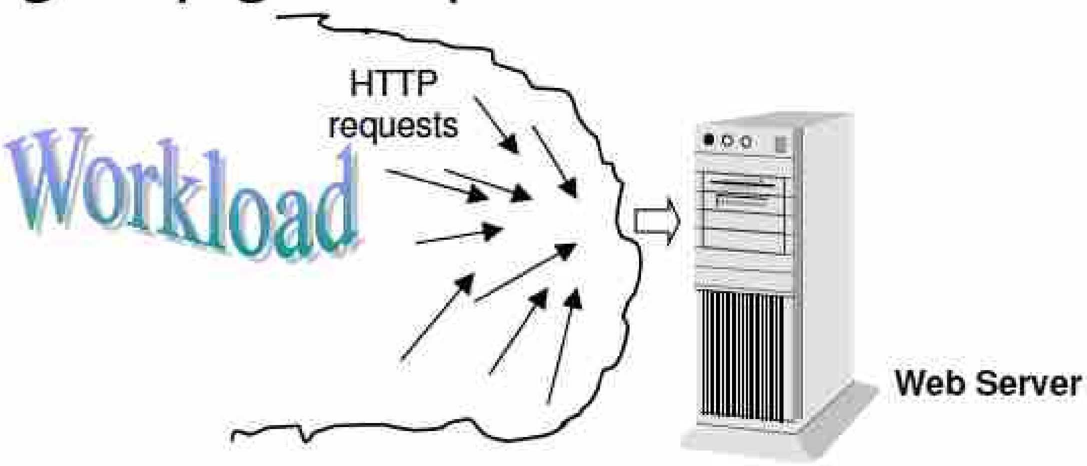
- measuring interval T= (13:48:29 - 13:41:41) = 408 sec
- number of requests = 11
- avg. arrival rate λ = 11/408 req./sec
- avg. size of transferred files = 188,117/10 = 18,811.7 bytes
- minimum file size = 441 bytes
- maximum file size = 98,995 bytes
- Logs do not provide all of the information needed by performance models

What is Workload Characterization?

Apresentacao derivada dos slides originais de Virgilio Almeida

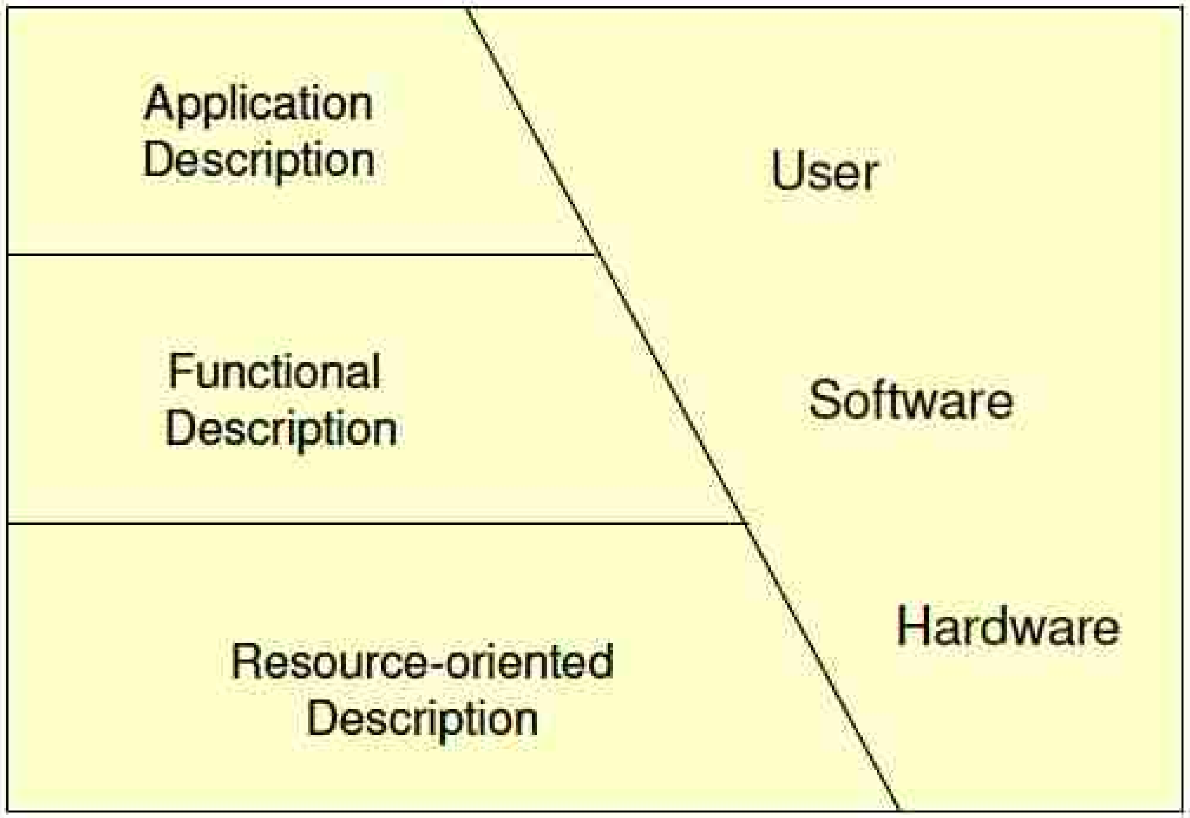
Workload

- The workload of a system can be defined as the set of all inputs that the system receives from its environment during any given period of time.



Apresentacao derivada dos slides originais de Virgilio Almeida

Workload Description



Execution of HTTP Requests (sec)

Request No.	CPU time	I/O time	Elapsed time
1 small	0.0095	0.04	0.071
2 medium	0.0130	0.11	0.145
3 medium	0.0155	0.12	0.156
4 small	0.0088	0.04	0.065
5 medium	0.0111	0.09	0.114
6 medium	0.0171	0.14	0.163
7 large	0.2170	1.20	4.380
8 medium	0.0129	0.12	0.151
9 small	0.0091	0.05	0.063
10 medium	0.0170	0.14	0.189

Workload Characterization

- Depends on the purpose of the study
  - cost x benefit of a proxy caching server
  - impact of a faster CPU on the response time
- Common steps
  - specification of a point of view from which the workload will be analyzed;
  - choice of set of relevant parameters;
  - monitoring the system;
  - analysis and reduction of performance data
  - construction of a workload model.

Workload Characterization: concepts and ideas

- Basic component of a workload refers to a generic unit of work that arrives at the system from external sources.
  - Transaction,
  - interactive command,
  - process,
  - HTTP request, and
  - depends on the nature of service provided

user
group of users
query
file

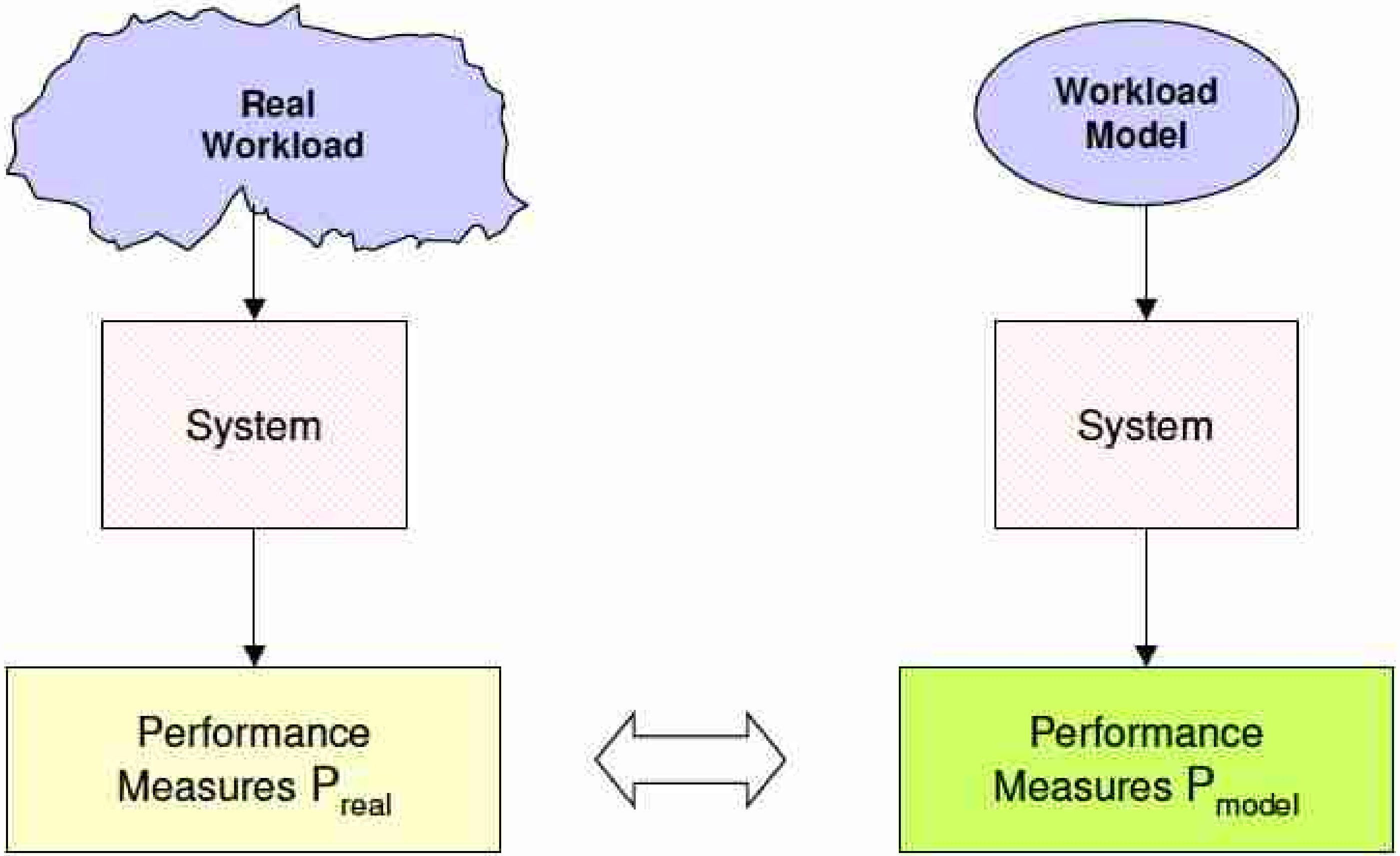
Workload Characterization: concepts and ideas

- Workload characterization
  - **workload model is a representation that mimics the workload under study.**

- Workload models can be used:
  - selection of systems
  - performance tuning
  - capacity planning

Uncovering relevant patterns

Representativeness of a Workload Model



Apresentacao derivada dos slides originais de Virgilio Almeida

A Web Server Example

- The pair **(CPU time, I/O time)** characterizes the execution of a request at the server.

- Our basic workload: 10 HTTP requests

- First case: only one document size (15KB)
- 10 executions ---> (0.013 sec, 0.09 sec)

- More realistic workload: documents have different sizes.

Execution of HTTP Requests (sec)

Request No.	CPU time	I/O time	Elapsed time
1	0.0095	0.04	0.071
2	0.0130	0.11	0.145
3	0.0155	0.12	0.156
4	0.0088	0.04	0.065
5	0.0111	0.09	0.114
6	0.0171	0.14	0.163
7	0.2170	1.20	4.380
8	0.0129	0.12	0.151
9	0.0091	0.05	0.063
10	0.0170	0.14	0.189
Average	0.0331	0.205	0.550

A Refinement in the Workload Model

- The average response time of 0.55 sec does not reflect the behavior of the actual server.
- Due to the heterogeneity of its components, it is difficult to view the workload as a single collection of requests.
- Three classes
  - small documents
  - medium documents
  - large documents

Three-Class Characterization

Type	CPU time (sec)	I/O time (sec)	No of omponents
Small Docs.	0.0091	0.04	3
Medium Docs.	0.0144	0.12	6
Large Docs.	0.2170	1.20	1
Total	0.331	2.05	10

Workload Models

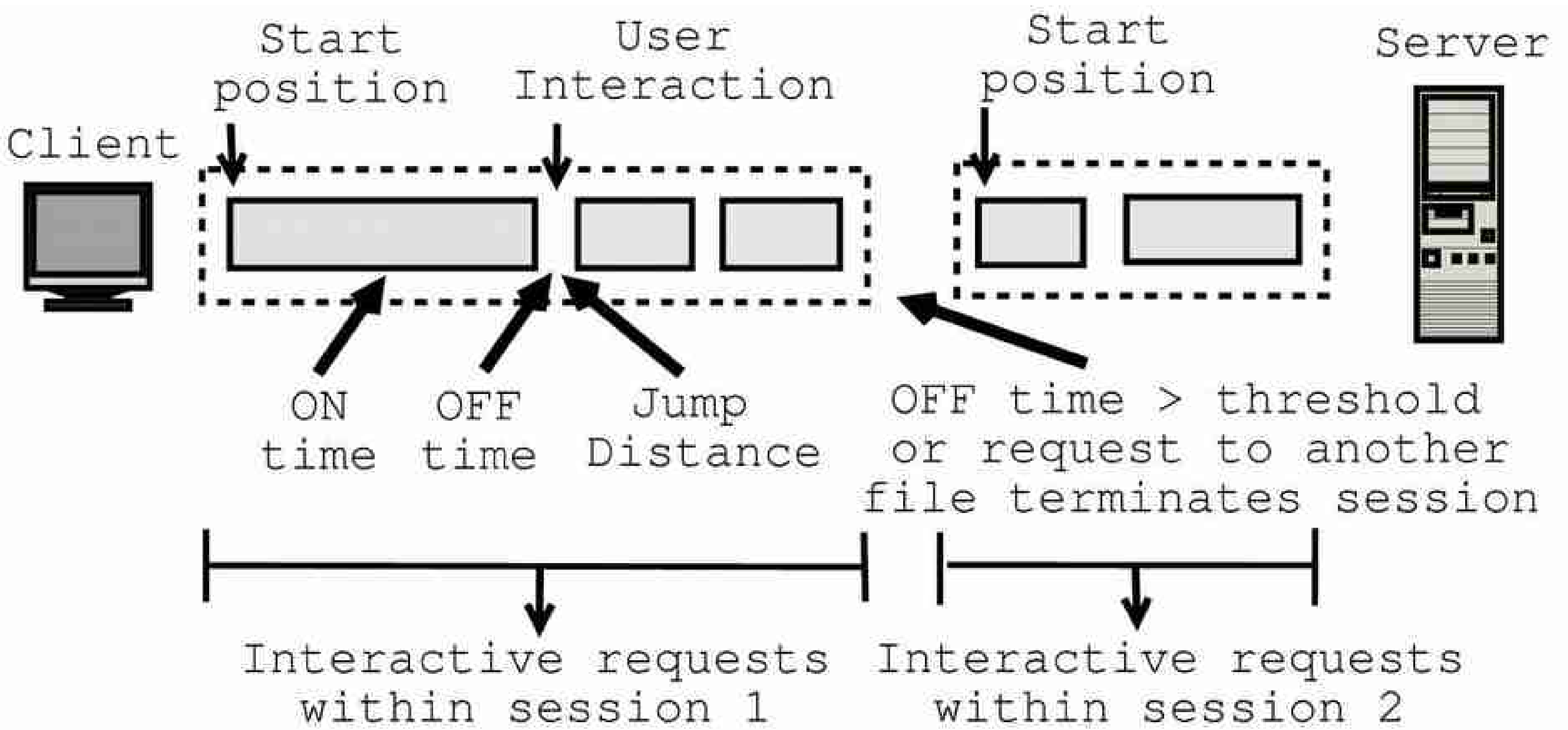
- A model should be representative and compact.
- Natural models are constructed either using basic components of the real workload or using traces of the execution of real workload.
- Artificial models do not use any basic component of the real workload.
  - Executable models (e.g.: synthetic programs, artificial benchmarks, etc)
  - Non-executable models, that are described by a set of parameter values that reproduce the same resource usage of the real workload.

A Workload Characterization Methodology

- Choice of an analysis standpoint
- Identification of the basic component
- Choice of the characterizing parameters
- Data collection
- Partitioning the workload
- Calculating the class parameters

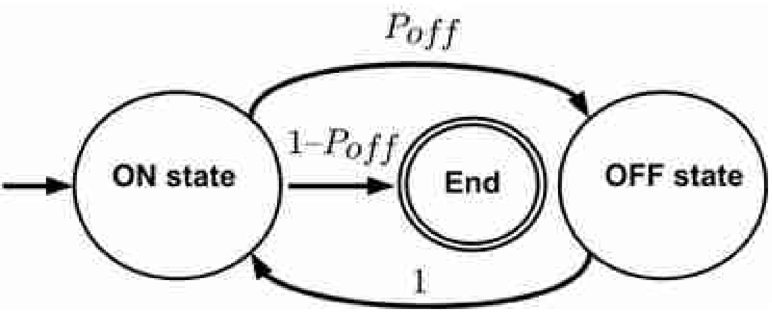
Qual o seu objetivo?

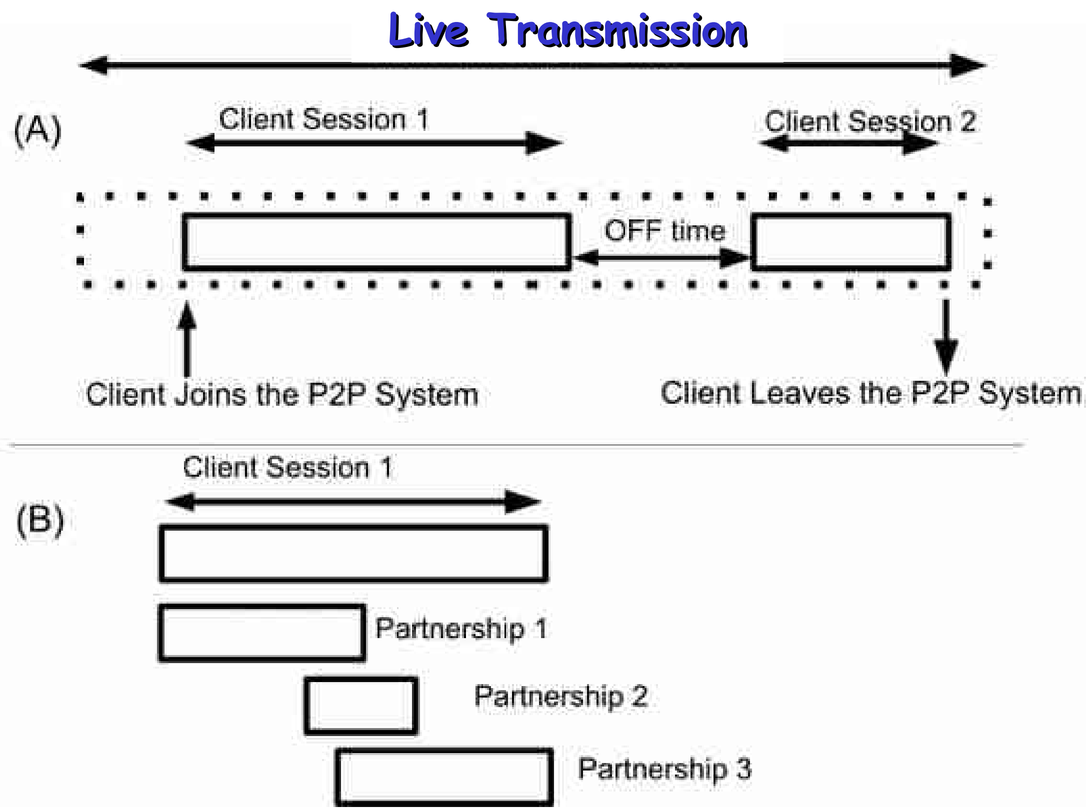
Caracterizar comportamento de usuário de VoD interativo



Qual o seu objetivo?

Caracterizar comportamento de usuário de live P2P TV





Qual o seu objetivo?

Caracterizar qualidade dos atributos textuais em aplicações Web 2.0 como potencial fonte de dados para serviços de informação

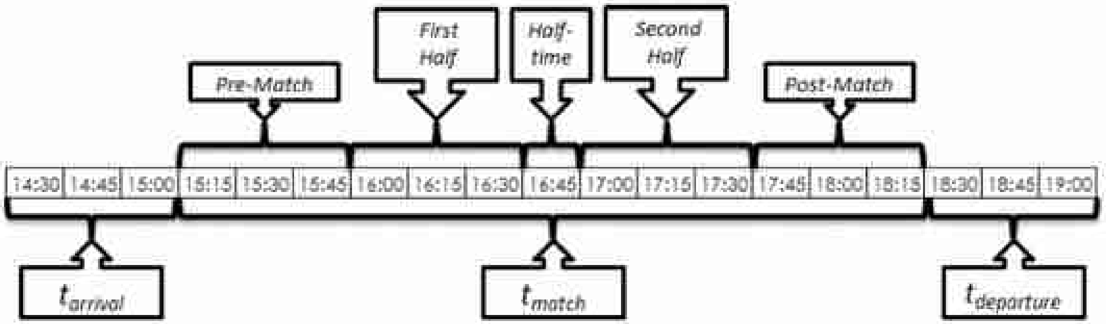
- **Quais atributos?**
  - **Título, tags, descrição, comentários**
- **Caracterizar qualidade sob que perspectiva? Quais métricas usar?**

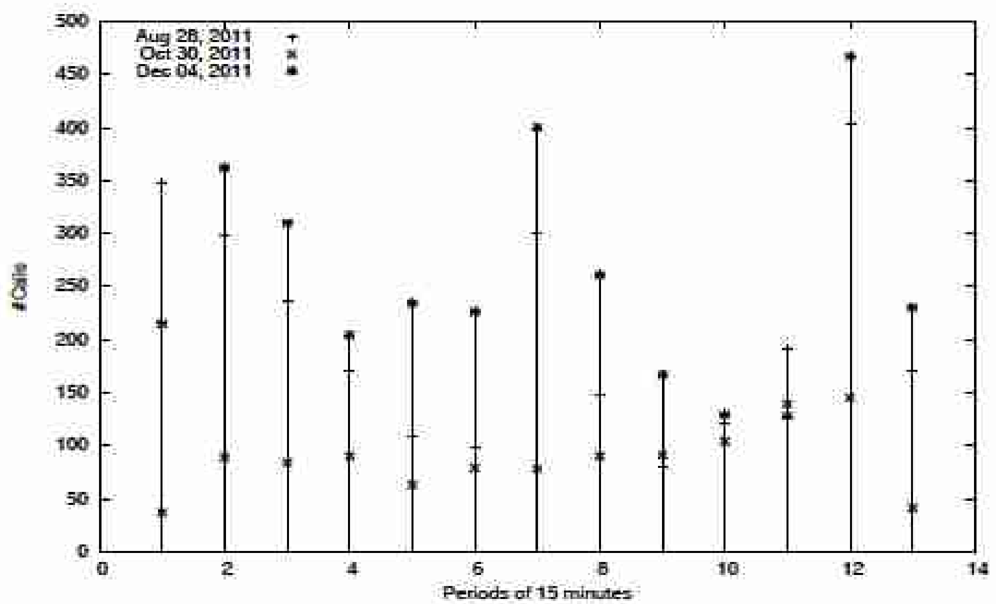
Uma proposta: qualidade depende:

- **Quantidade de conteúdo: # de termos únicos**
- **Poder descritivo: heurística derivada do TF**
- **Poder discriminativo: heurística derivada do IDF**

Qual o seu objetivo?

Caracterizar a sobrecarga na infra-estrutura de redes de comunicação de um evento de larga escala?

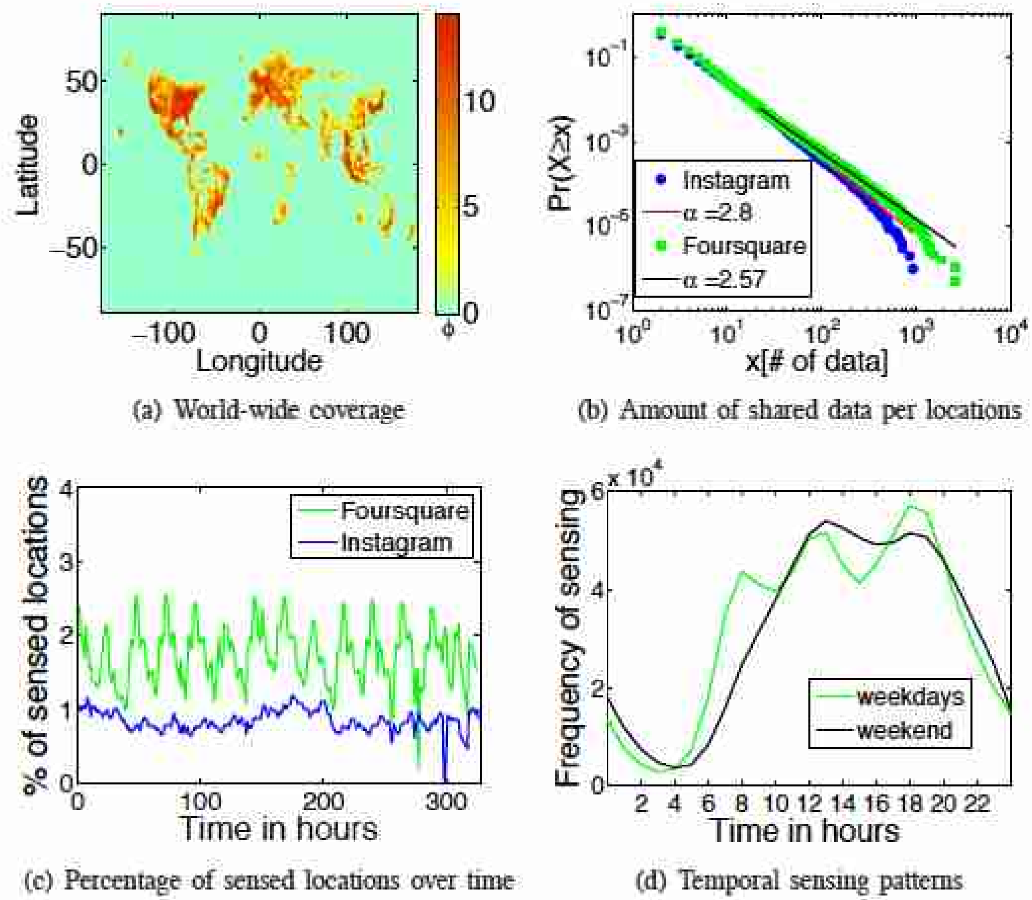




Qual o seu objetivo?

Caracterizar o potencial de redes sociais geo-referenciadas para criar redes de sensoriamento participativo:

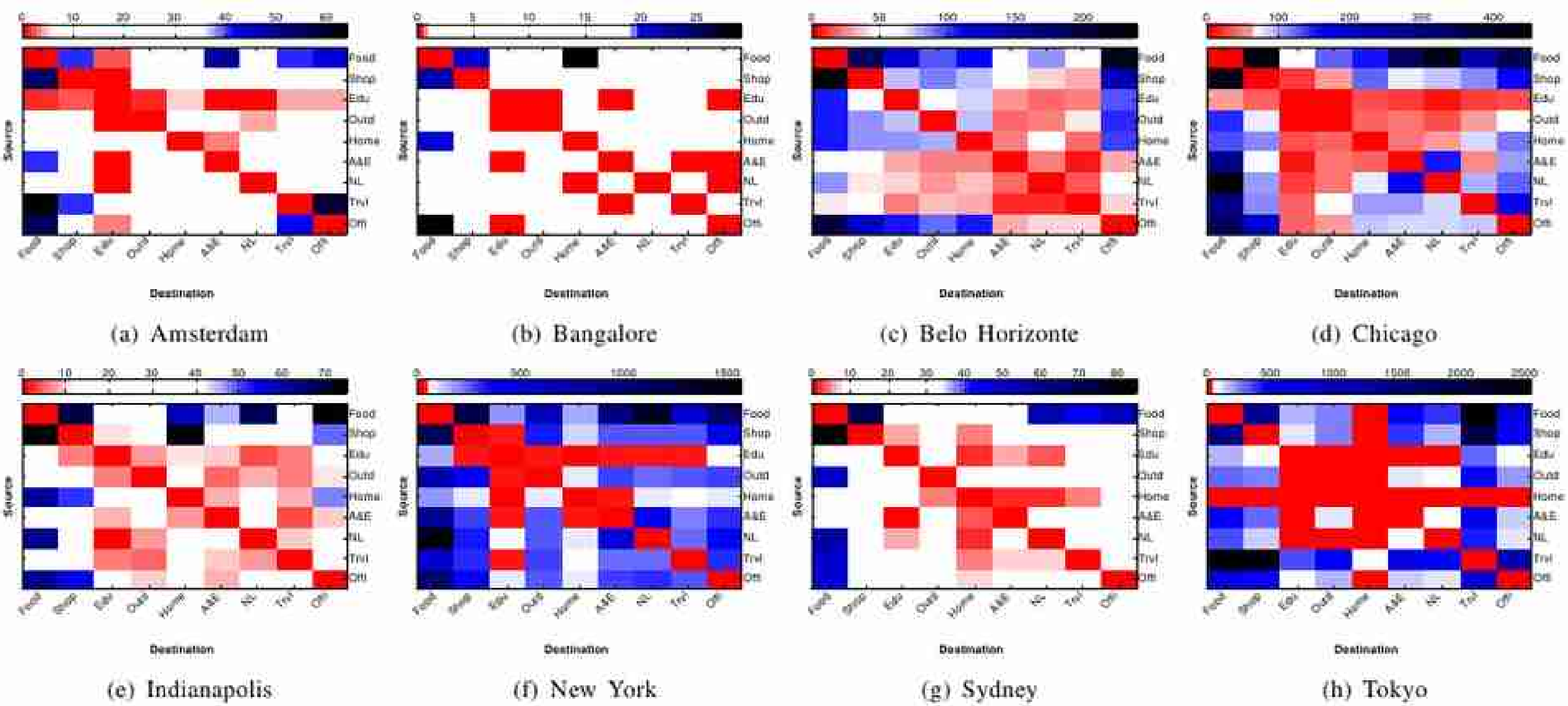
Aspectos chaves: cobertura espacial e temporal



(a) World-wide coverage
(b) Amount of shared data per locations
(c) Percentage of sensed locations over time
(d) Temporal sensing patterns

Qual o seu objetivo?

Caracterizar os padrões de deslocamentos das pessoas em diferentes cidades do mundo, utilizando check-ins do Foursquare



(a) Amsterdam (b) Bangalore (c) Belo Horizonte (d) Chicago

(e) Indianapolis (f) New York (g) Sydney (h) Tokyo

Data Collection

- This step assigns values to each component of the model.
  - Identify the time windows that define the measurement sessions.
  - Monitor and measure the system activities during the defined time windows.
  - From the collected data, assign values to each characterizing parameters of every component of the workload.

Challenge: collected data must be representative
When? How? For how long?

Partitioning the workload

- Motivation: real workloads can be viewed as a collection of heterogeneous components.
- Partitioning techniques divide the workload into a series of classes such that their populations are composed of quite homogeneous components.
- What attributes can be used for partitioning a workload into classes of similar components?

Partitioning the Workload

- Resource usage
- Applications
- Objects
- Geographical orientation
- Functional
- Organizational units
- Mode

Workload Partitioning: Resource Usage

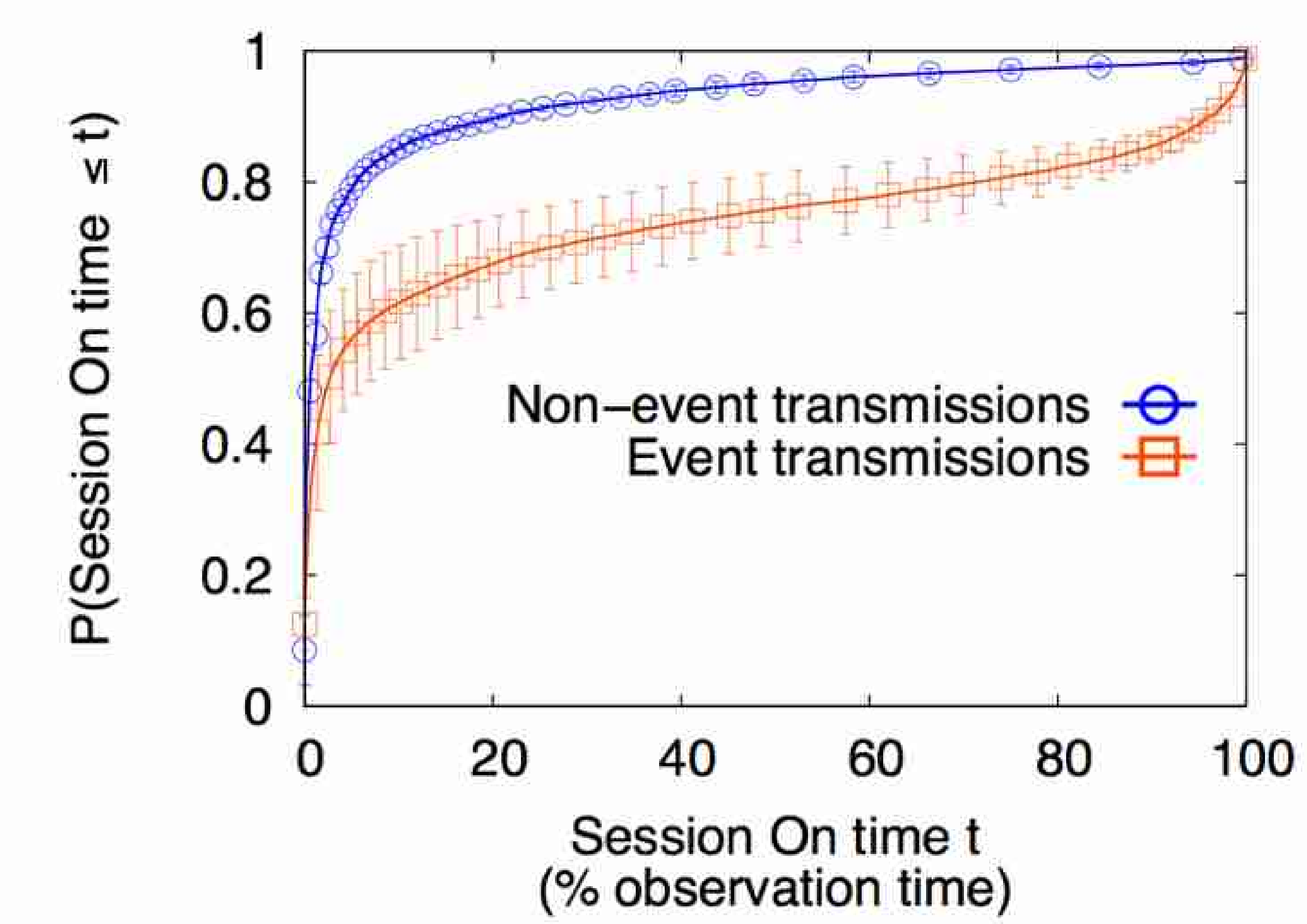
Transaction Classes	Frequency	Maximum CPU time (msec)	Maximum I/O time (msec)
Trivial	40%	8	120
Light	30%	20	300
Medium	20%	100	700
Heavy	10%	900	1200

Workload Partitioning: Internet Applications

Application Classes	KB Transmitted
WWW	**4,216**
ftp	**378**
telnet	**97**
Mbone	**595**
Others	**63**

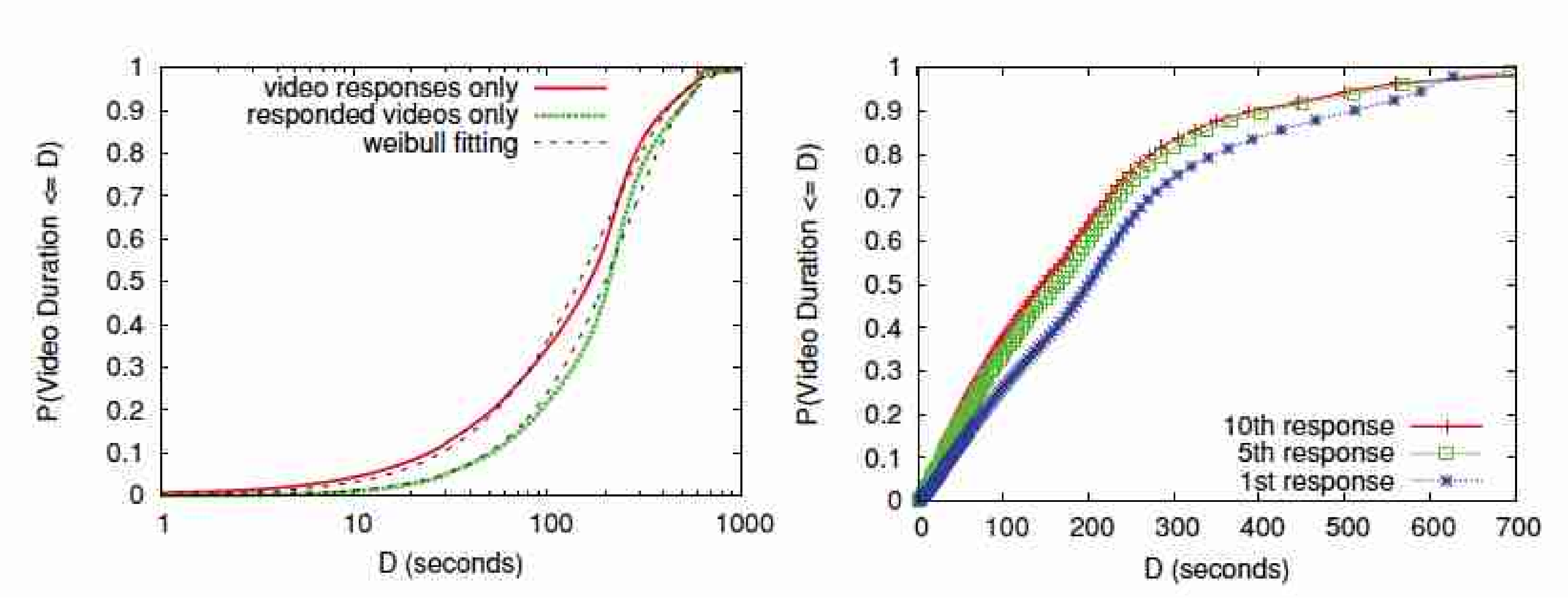
Workload Partitioning (Type of content)

Client Session Time in Live P2P TV



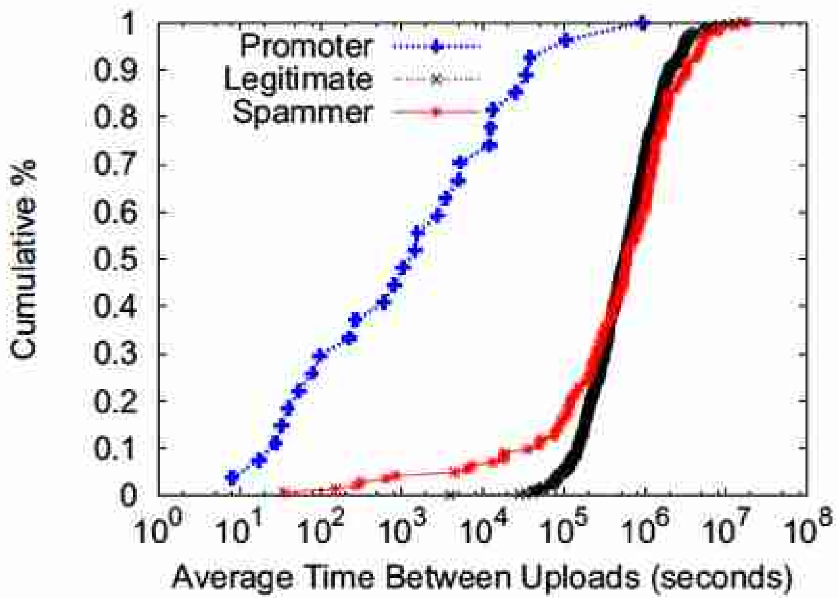
Workload Partitioning

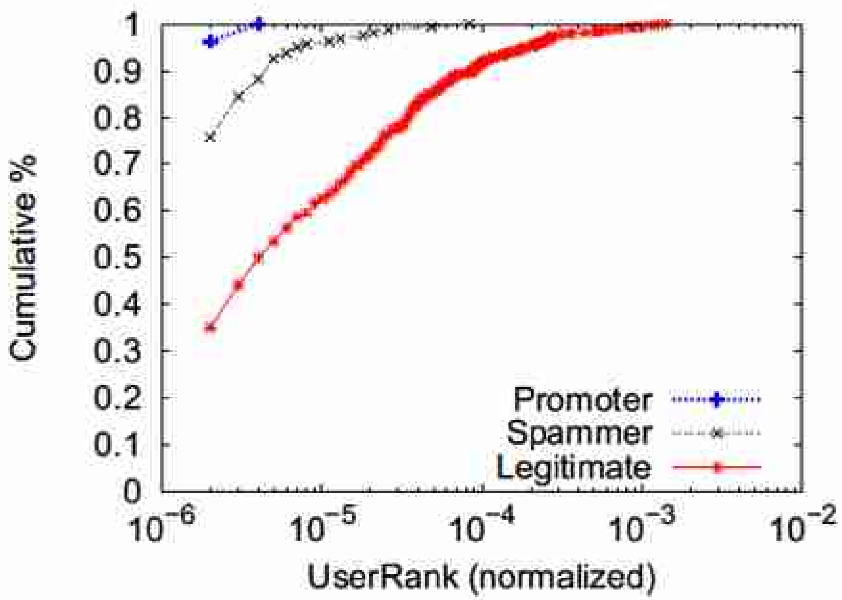
YouTube video durations



Workload Partitioning

User Behavior on YouTube





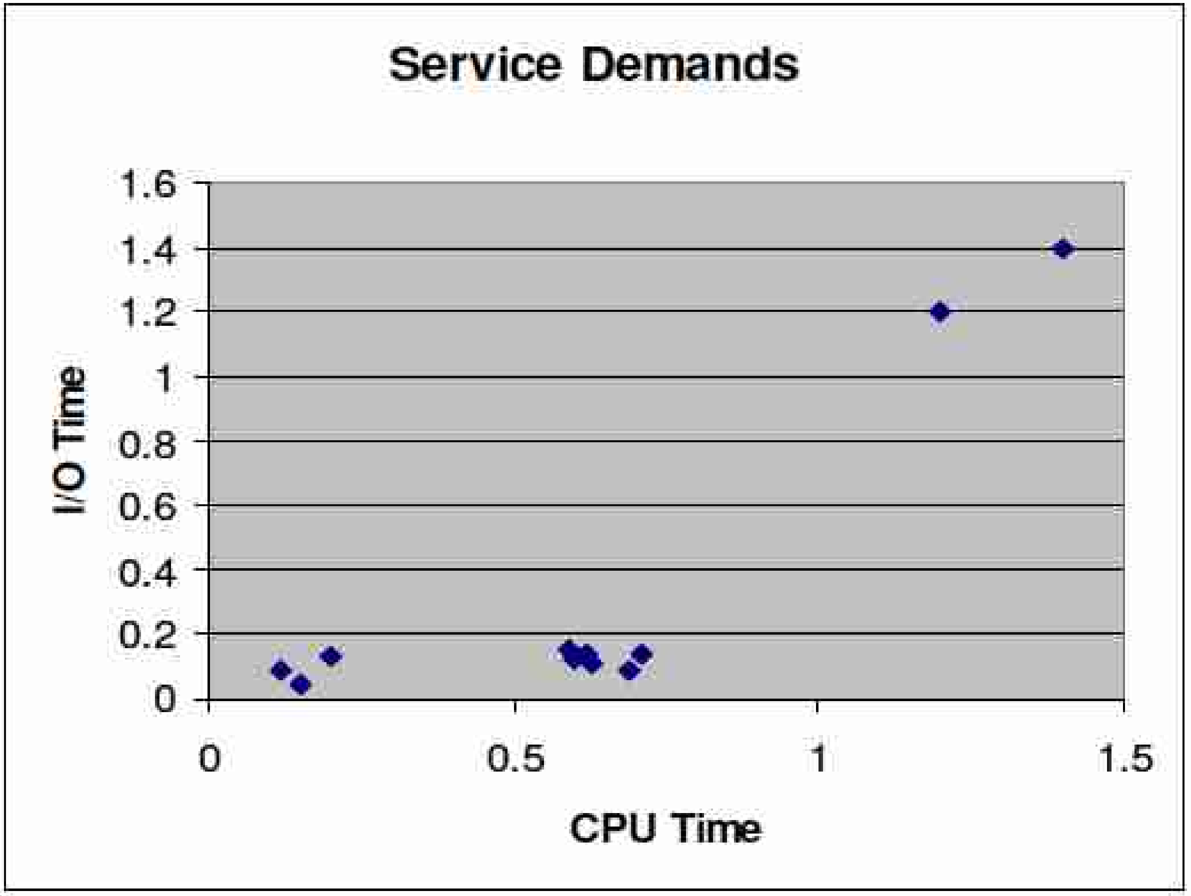
Workload Partitioning is important….

- Temporal variations: be careful!

- How to partition…
  - Based on prior knowledge/ intuition
    - E.g: live P2P TV users may behave differently if they watch a live TV news or a final match of an important championship

  - Clustering algorithms

Clustering Analysis



Clustering Analysis

- The centroid of a cluster is the point whose parameter values are the means of the parameter values of all points in the cluster.
- Clustering analysis can be automatically performed by some software packages
- The output of a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.

Clustering Algorithms

- The goal of a clustering algorithm is to identify natural groups of components, based on similar resource requirements.

- Minimal Spanning Tree (MST)

- k-means algorithm

Minimal Spanning Tree

- Set the initial number of clusters equal to the number of components of the workload ($j=p$)
- Repeat the following steps until the desired number of clusters is obtained:
  - determine the parameter values of centroid C_j of each of the j clusters
  - calculate the j x j intercluster distance matrix, where each element (m,n) represents the distance between the centroids of clusters m and n.
  - determine the minimum nonzero element (q,r) of the distance matrix. It indicates that clusters q and r are to be merged. Then decrease the number of clusters ($j \leftarrow j-1$)

Clustering Analysis: an example

Document	Size (KB)	Number of Accesses
1	12	281
2	150	28
3	5	293
4	25	123
5	7	259
6	4	241
7	35	75

Clustering Analysis:
Logarithmic transformation of parameters

Document	Size (KB)	Number of Accesses
1	1.08	2.45
2	2.18	1.45
3	0.7	2.47
4	1.40	2.09
5	0.85	2.41
6	0.60	2.38
7	1.54	1.88

Clustering Analysis:
Centroids of the initial clusters

Clusters	Size (KB)	Number of Accesses
C1	1.08	2.45
C2	2.18	1.45
C3	0.7	2.47
C4	1.40	2.09
C5	0.85	2.41
C6	0.60	2.38
C7	1.54	1.88

Clustering Analysis

- The distance between two points $w_1 = (90, 1300)$ and $w_2 = (500, 4000)$, using the Euclidean Metric, is given by:

- $d_{w1,w2} = ((90 - 500)^2 + (13000 - 4000)^2)^{1/2} = 9009.3$

Intercluster Distance Matrix (1)

Cluster	C1	C2	C3	C4	C5	C6	C7
C1	0	1.49	0.38	0.48	0.24	0.48	0.74
C2		0	1.79	1.01	1.64	1.83	0.76
C3			0	0.79	0.16	0.13	1.03
C4				0	0.64	0.85	0.26
C5					0	0.25	0.88
C6						0	1.07
C7							0

Intercluster Distance Matrix (2)

Cluster	C1	C2	C36	C4	C5	C7
C1	0	1.49	0.43	0.48	0.24	0.74
C2		0	1.81	1.01	1.64	0.76
C36			0	0.79	0.19	1.03
C4				0	0.64	0.26
C5					0	0.88
C7						0

Intercluster Distance Matrix (3)

Cluster	C1	C2	C365	C4	C7
C1	0	1.49	0.33	0.48	0.74
C2		0	1.73	1.01	0.76
C365			0	0.73	0.96
C4				0	0.26
C7					0

Intercluster Distance Matrix (4)

Cluster	C1	C2	C365	C47
C1	0	1.49	0.33	0.61
C2		0	1.73	0.89
C365			0	0.84
C47				0

Output of the Clustering Process

Type	Class	Doc. Size (KB)	No. of Acesses	No. of Components
Small	C1356	8.19	271.51	4
Medium	C47	29.58	96.05	2
Large	C2	150.00	28.00	1

How many clusters k to choose?

- Must examine variation of two metrics:
  - Intra-cluster distance: Average distance between points of a cluster and its centroid
  - Inter-cluster distance: average distance between centroids

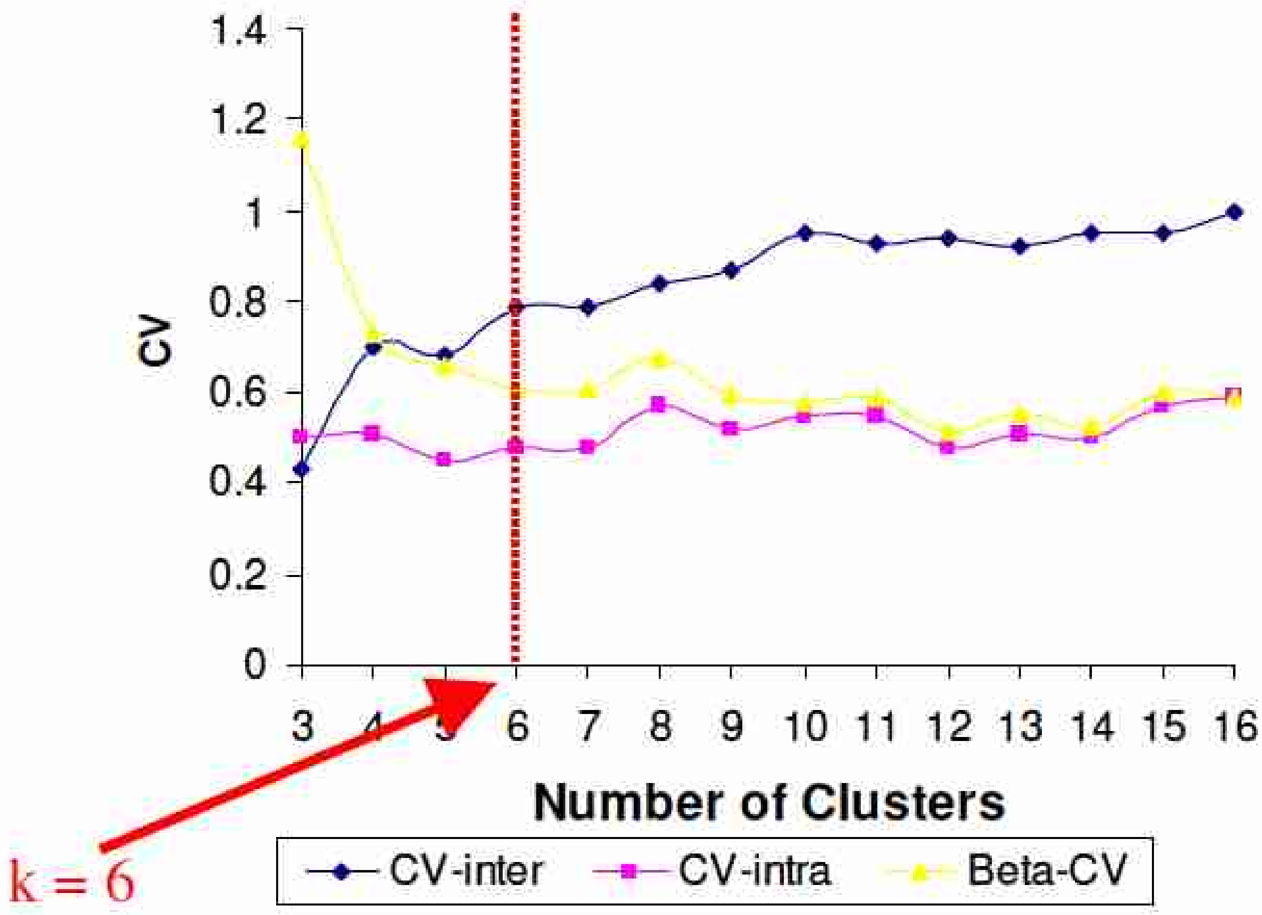
- Variation characterized by coefficient of variation(CV)
  - CV = standard deviation / average

- Must min intracluster CV while max'ing intercluster CV

How many clusters k to choose?

- Define β_{CV} = intracluster CV / intercluster CV
- Plot β_{CV} vs. number of clusters k
- Choose k after which that β_{CV} stabilizes

How many clusters k to choose?



Apresentacao derivada dos slides originais de Virgilio Almeida

How many clusters k to choose?

- Silhouette Index: avalia coesão e separação dos clusters

Explora a distância média dos pontos pertencentes ao cluster mais próximo para os pontos de um grupo

- Para cada ponto i, calcula-se o silhouette index

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

onde μ^{min}_{out} é a distância média de x_i para os pontos do cluster mais próximo e μ_{in} é a distância de x_i para os pontos do seu próprio cluster.

How many clusters k to choose?

$S_i = [-1, 1]$

$S_i \sim 1$: x_i está mais próximo dos pontos do seu cluster do que do cluster vizinho

$S_i \sim 0$: x_i está próximo da fronteira

$S_i \sim -1$: x_i está mais próximo dos pontos do outro cluster

A qualidade do clustering é estimada como a média do SI para todos os pontos

$$SilhouetteIndex = \frac{1}{n}\sum_{i=e}^{n} S_i$$

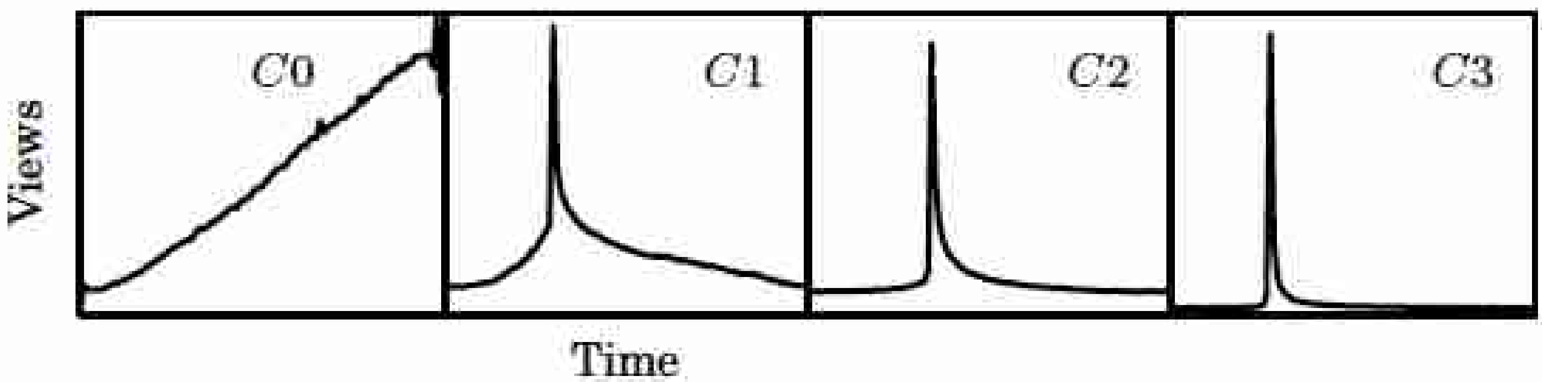
How many clusters k to choose?

Como interpretar o resultado de Silhouette Index (SI)?

- **Segundo Rousseeuw (1987):**
  - **$0.70 < SI \leq 1$: estrutura forte**
  - **$0.5 < SI \leq 0.7$: estrutura razoável**
  - **$0.25 < SI \leq 0.5$: estrutura fraca e pode ser artificial**
  - **$SI \leq 0.25$: nenhuma estrutura substancial**

Qual seu objetivo?

- **Entender como a popularidade de social media evolui com tempo**
  - **Identificar diferentes padrões**



Popularity Trends (Cluster Centroids) in Both Random and Top Datasets

Clustering de Time Series (KSC)

Benchmarks

- Benchmarking refers to running a set of representative programs on different systems and measuring the results
- What is a particular benchmarking really testing?
- How close does a benchmark resemble a user environment?
- What is it really measuring?

Benchmarks

- Common benchmarks:
  - www.tpc.org
  - www.spec.org
  - Httperf
  - Public datasets: LETOR, MovieLens