Measurement of Youtube traffic from Orange networks

Fabrice Guillemin, Thierry Houdoin, Stéphanie Moteau

Orange Labs, 22300 Lannion

Workshop "Optimization of Network Ressources for Content Access and Delivery"

September 6, 2012

Outline

Introduction

- Objective: to show results for measurements of Youtube traffic in Orange networks
- Special focus on Otarie probes located in Paris, Lyon, and Bordeaux. The measurements in Paris are composed of data collected by two Otarie probes (higher aggregation level than in Bordeaux and Lyon)
- Additional measurements from all Otarie probes located in France
- Data are analyzed
  1. to compute popularity curves
  2. to determine which content is popular
  3. to characterize the request arrival pattern (validity of the IRM assumption)

Approximation of popularity curves

In all the measurement results reported below, we approximate the popularity curve of Youtube video files by a truncated Pareto function of the form

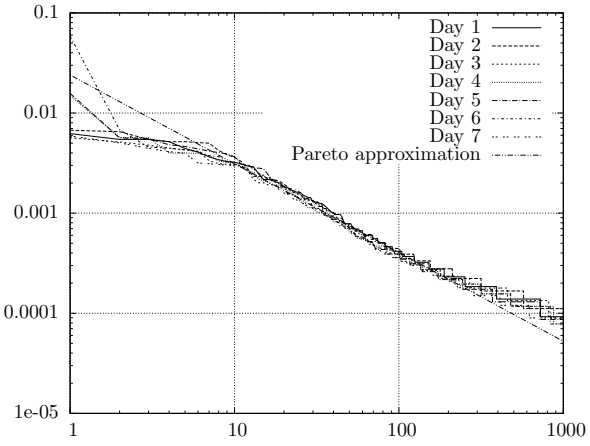
$$f(x) = \mathbb{1}_{\{x_{\min} \leq x \leq x_{\max}\}} \frac{c}{x^{\alpha}}. \tag{1}$$

The parameters c and α as well as the range $[x_{\min}, x_{\max}]$ may change, depending on the duration of the measurements and the location where measurements are performed.
When $\alpha < 1$, this the range $[x_{\min}, x_{\max}]$ is necessarily finite.

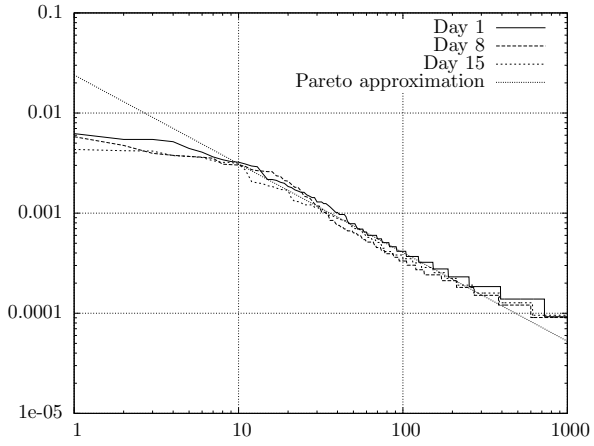
Some issues regarding measurements

- Files are rarely transmitted over a single TCP connection (even in the case of progressive download)
- It is necessary to "reconstruct the session": Aggregate the TCP connections between the same IP addresses (client and server) within a given time frame (e.g., 30 seconds).
- This point is critical for estimating the volume of a video file.
- Issue for CDN: It is necessary that the CDN server acts as a proxy, even if the complete file is not totally viewed by the end user (reneging, zapping, bad quality). Only "relevant" files should be cached!

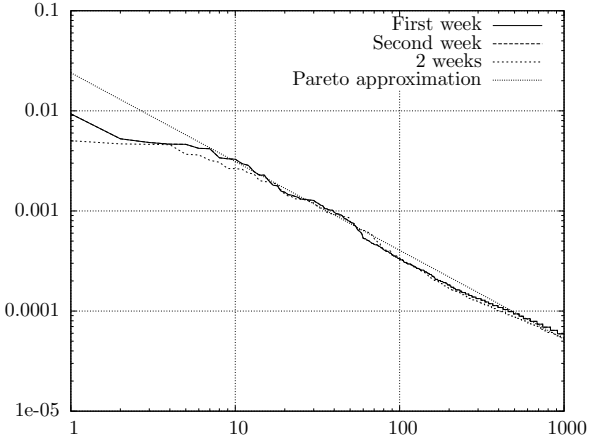
Popularity curves in Bordeaux



(a) One week.



(b) The first days of three consecutive weeks.



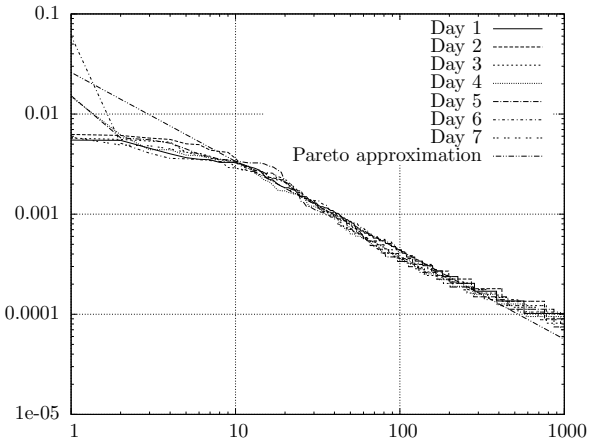
(c) The first and second weeks and the two weeks.

$$c = 0.024004, \quad \alpha = 0.886537$$

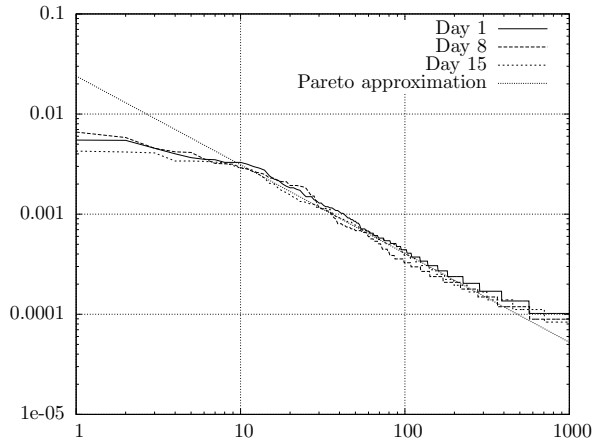
Statistics of downloads in Bordeaux

	number	number of downloads	volume
Day 7			
Files	23,071	33,351	2,427.6 GB
More than twice	1,193	9,315 (28 %)	701.6 GB
Only once		19,720 (59.1 %)	1247.9 GB
First week			
Files	110,106	202,717	6239 GB
More than twice	10,510	90,203 (44.5 %)	2736.6 GB
Only once		86,678 (42.7 %)	2512 GB
Two weeks			
Files	208,289	425,266	9,354.8 GB
More than twice	24,291	215,158 (50.6 %)	4,462.7 GB
Only once		157,888 (37 %)	3,486.2 GB

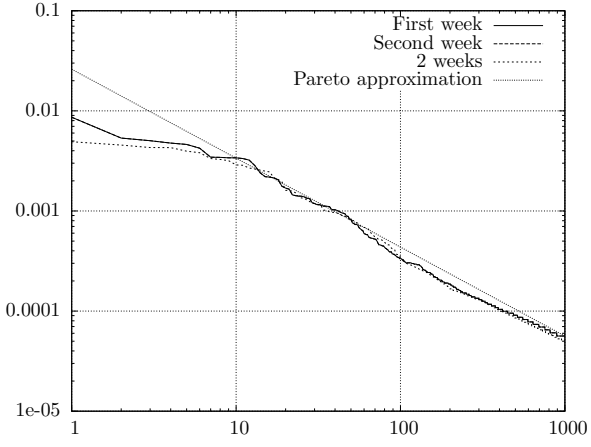
Popularity curves in Lyon



(d) One week.



(e) The first days of three consecutive weeks.



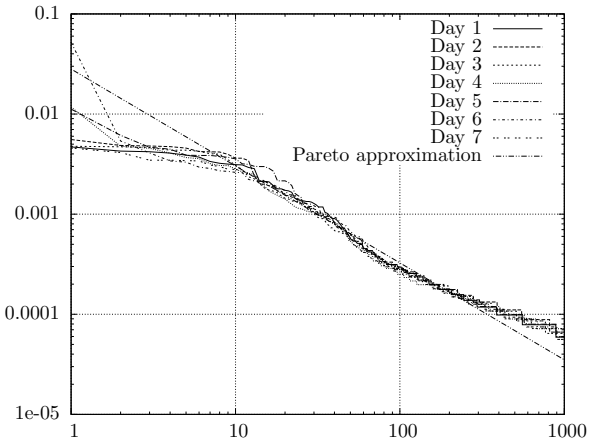
(f) Aggregation over the first and second weeks and over the two weeks.

$c = 0.02606,\ \alpha = 0.8885$

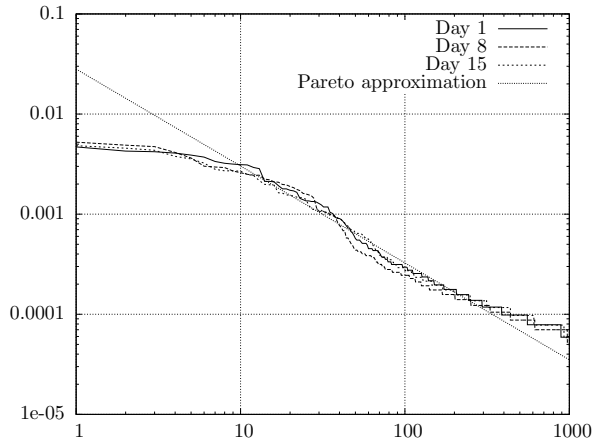
Statistics of downloads in Lyon

	number	number of downloads	volume
Day 7			
Files	22,768	34,025	2,493.1 GB
More than twice	1,311	10,414 (30.6 %)	755.2 GB
Only once		19,303 (56.7%)	1,295.6 GB
First week			
Files	122,461	230,766	6,552.4GB
More than twice	12,182	106,127 (46 %)	2,980.7 GB
Only once		95,919 (41.6 %)	2,529.9 GB
Two weeks			
Files	232,579	492,008	9,956.4 GB
More than twice	27,899	257,736 (52.4 %)	4,834.1 GB
Only once		175,088 (35.6 %)	3,653.6 GB

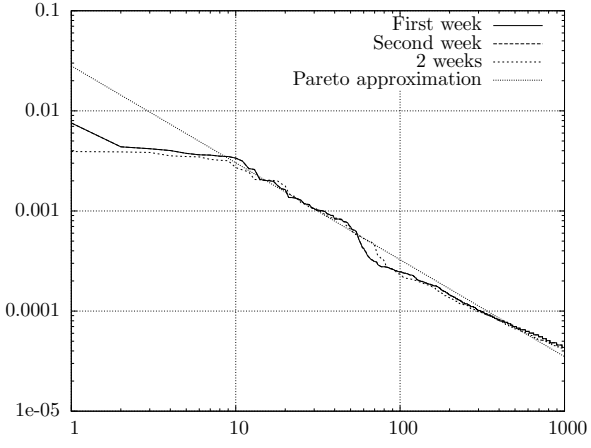
Popularity curves in Paris



(g) One week



(h) The first days of three consecutive weeks.



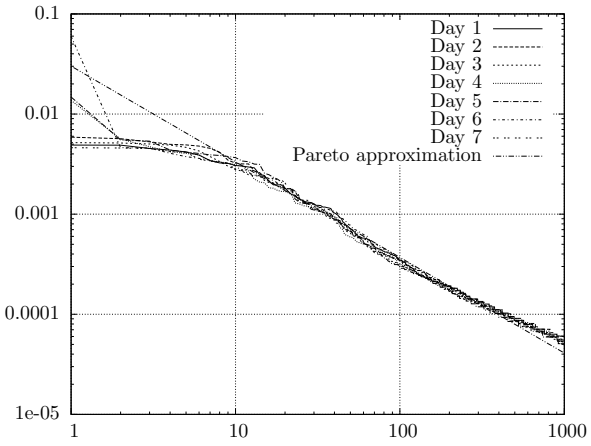
(i) Aggregation over the first and second weeks and over the two weeks.

$$c = 0.028128, \ \alpha = 0.96802.$$

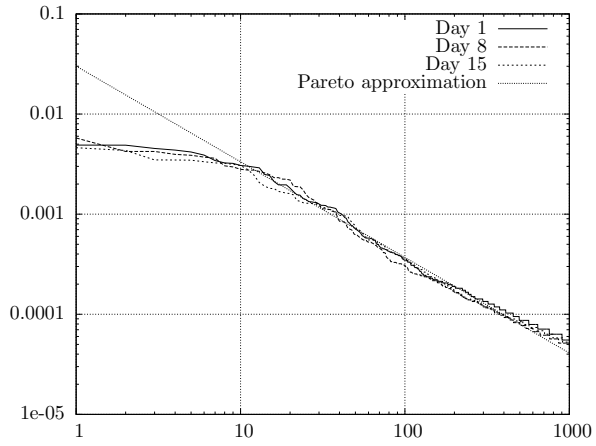
Statistics of downloads in Paris

	number	number of downloads	volume
Day 7			
Files	39,172	55,520	2,919.7 GB
More than twice	1,908	14,580(26.2 %)	836.7 GB
Only once		33,598 (60.5 %)	1,551.4 GB
First week			
Files	226,943	414,807	8,609.9 GB
More than twice	21,566	182,093 (43.9 %)	3,548.6 GB
Only once		178,040 (42.3 %)	3,712.4 GB
Two weeks			
Files	415,915	841,916	12,961 GB
More than twice	48,013	420,205 (49.9 %)	5,732.7 GB
Only once		314,093 (37.3 %)	5,321.8 GB

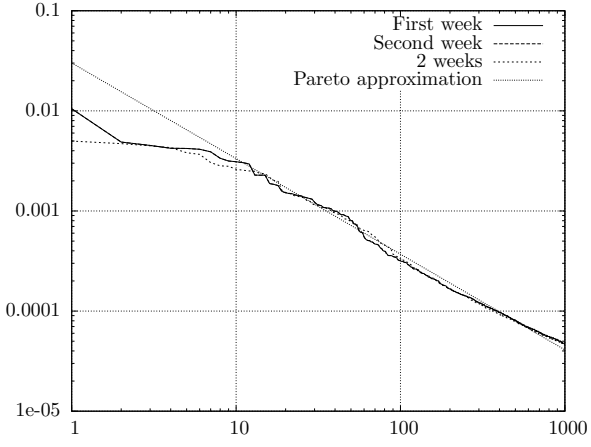
All Otarie probes



(j) One week.



(k) The first days of three consecutive weeks.



(l) Aggregation over the first and second weeks and over the two weeks.

$c = 0.03010,\ \alpha = 0.9553$

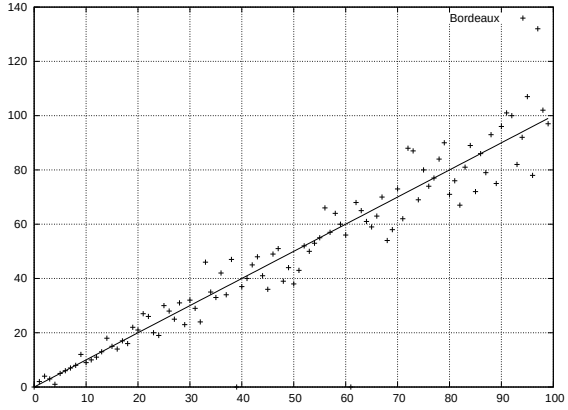
Statistics of all downloads

	number	number of downloads	volume
Day 7			
Files	107,381	1114,065	6,484 GB
More than twice	8,611	76,319 (40 %)	2,593.8 GB
Only once		87,154 (46.7 %)	2,816 GB
First week			
Files	473,131	1,112,936	15,000.7 GB
More than twice	61,145	640,122 (57.5 %)	7,252.7 GB
Only once		351,158 (31.6 %)	5565 GB
Two weeks			
Files	891,781	2,472,450	22,503.9 GB
More than twice	138,269	1,597,572 (64.6 %)	11,391.4 GB
Only once		632,145 (25.5 %)	7,973.6 GB

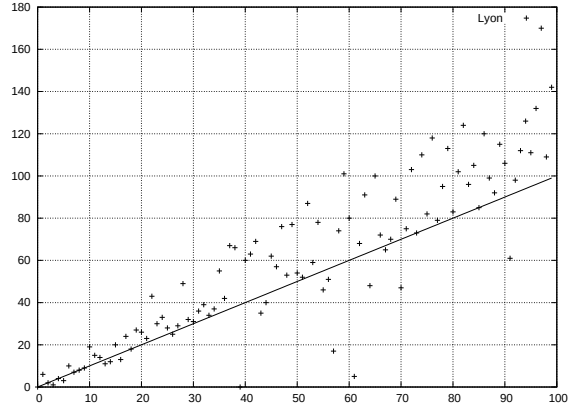
Lessons from statistics

- The popularity curve of video files estimated over one day is quite stable in time and can be well approximated by a Pareto curve with a shape parameter less than one.
- The same Pareto approximation holds when aggregating measures over longer time periods (over one week or two weeks);
- The Pareto approximation is however valid only for those video files which are seen a significant number of times.
- A huge number of video files are seen only once or twice, indicating that the tail of the popularity curve is flat.
- The number of cacheable flows increases in time. Caching capacities may overflow (need for replacement policies).
- The mean volume of Youtube files is much larger than those observed in previous studies (Youtube has recently changed and more voluminous content is available).

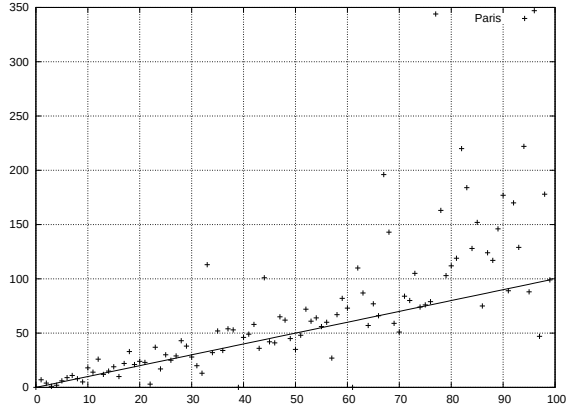
Cross popularity analysis



(m) Bordeaux.



(n) Lyon.



(o) Paris.

File are classified according to their global popularity (diagonal) - top 100 files. Points represent their popularity in ADSL areas.

Regional aspects of files

- Video files viewed during two weeks have been sorted according to their popularity on the various ADSL areas in the limit of 1 Terabytes for the cumulative volume.
- The limit of 1 TB for the most popular files corresponds to 1,238 files in global measurements, 1,582 files in Paris, 1,339 files in Lyon and 1,331 files in Bordeaux.

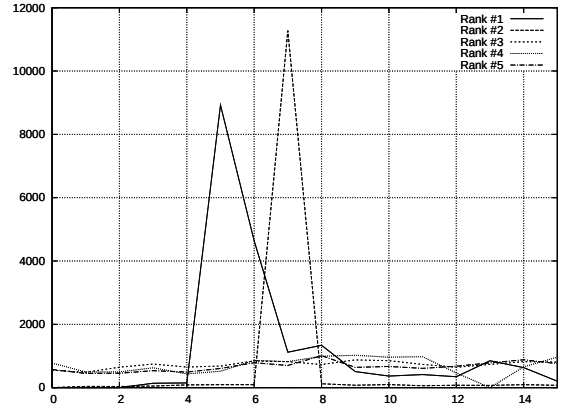
Location	Global	Bordeaux	Lyon	Paris
Global		831.5GB	847.3 GB	735.3 GB
Bordeaux	831.5 GB	166.9Gb	758.3 GB	686.1 GB
Lyon	847.3 GB	758.3 GB	174.5 GB	678.4 GB
Paris	735.3 GB	686.1 GB	678.4 GB	247GB

Analysis of cross popularity analysis

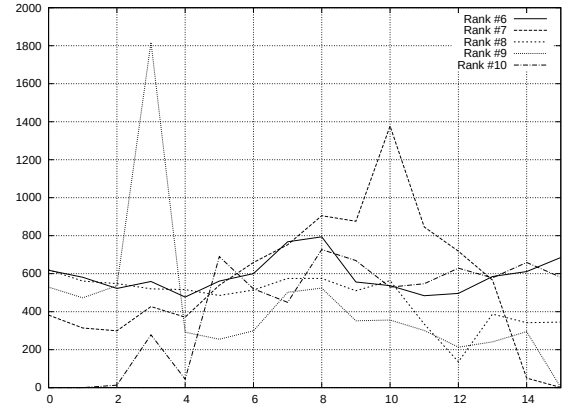
- Files which are popular in a given ADSL area are also popular in other ADSL areas.
- The fraction of files which are popular only in one ADSL area is rather small but far not negligible.
- It is worth distributing cache servers because of regional aspects of video files.
- A centralized cache can be used to cache files which are popular in all ADSL areas and medium cache servers to cache those files which are popular in a given ADSL area.

Caching popular files requires big storage capacities. It is necessary to understand request dynamics.

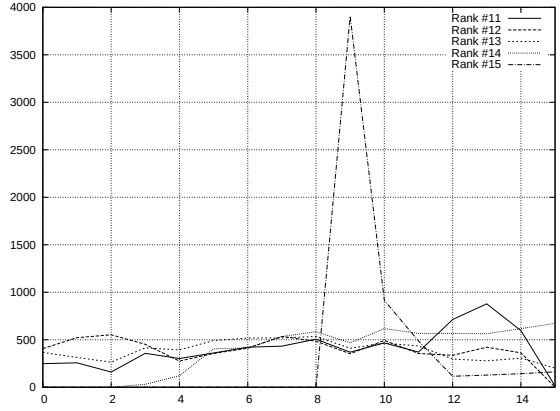
Request arrival processes of the top 15 files.



(p) For the 5 most popular.



(q) For the subsequent 5 most popular.



(r) For the last but 5 out of the 15 most popular.

Request process profiles

From the analysis of request arrival process of the top15, we can make the following points:

- Flat profile: files are continually requested and the intensity of requests is roughly constant. For this type of request pattern, the Independent Request Model (IRM) assumption is reasonable since the request process can be considered as stationary.
- Peaky profile: files are requested in bursts. There is a peak of requests that arrive and after the buzz, request vanish. In that case, the IRM assumption is clearly not satisfied.

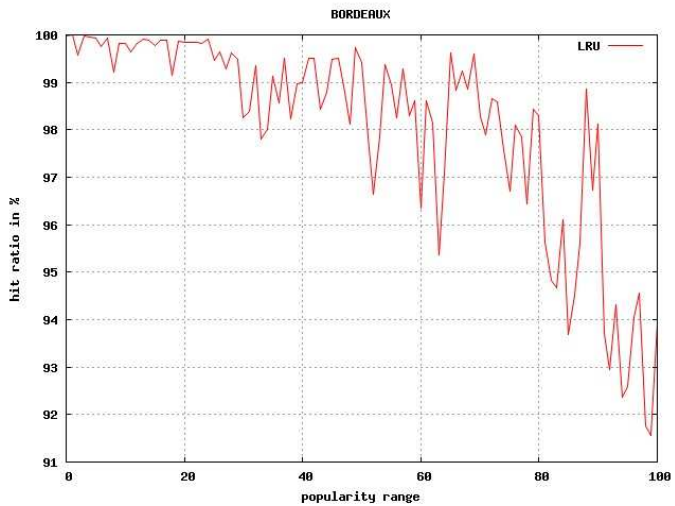
Even if the IRM assumption is not satisfied and files appears and disappear, the popularity curve is surprisingly constant over time periods with different lengths.

Impact of rare files on caching performance

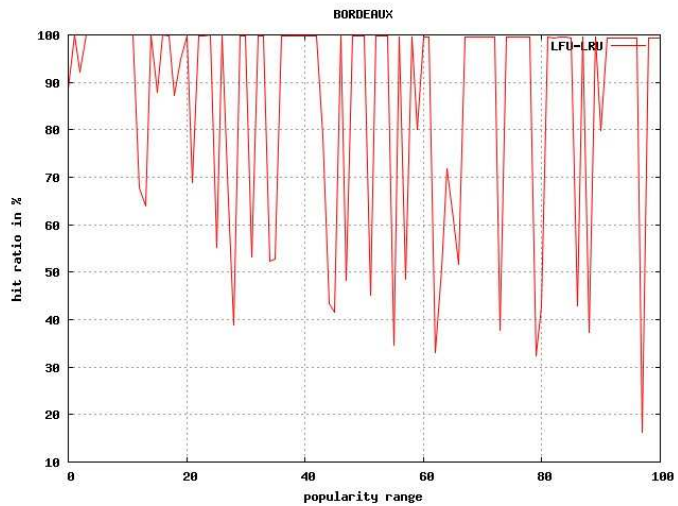
Hit ratios in simulated cache servers in Bordeaux and Lyon with the LRU and LFU+LRU disciplines (in percentage):

	LRU		LRU + LFU	
	global	files downloaded more than twice	global	files downloaded more than twice
Bordeaux	5.8	27.2	3.2	15
Lyon	5.9	26.8	4.2	19

Hit ratios for the top 100 files in Bordeaux:



(s) LRU in Bordeaux.



(t) LFU+LRU in Bordeaux.

Files requested only once or twice cause cache overflow

Future work

- A caching policy based on a big cache server for Youtube files popular in all ADSL areas and medium size cache servers for those files which are popular in one ADSL area makes sense. Investigate then gain in bandwidth consumption.
- Design a filter to eliminate files which are rarely requested (huge number) and to cache those files which are highly requested. This is equivalent to find heavy hitters (or elephants) in traffic.
- Analyze the consequences with regard to CCN. Is it useful to cache content everywhere in the network or is regional caching is sufficient?