Measuring the Internet: challenges and applications

Telecommunication Group presentation

Speaker:
Marco Mellia

Politecnico di Torino – 7/12/2011

The Internet today

A very complex scenario

- many eterogeneous services
- many content providers

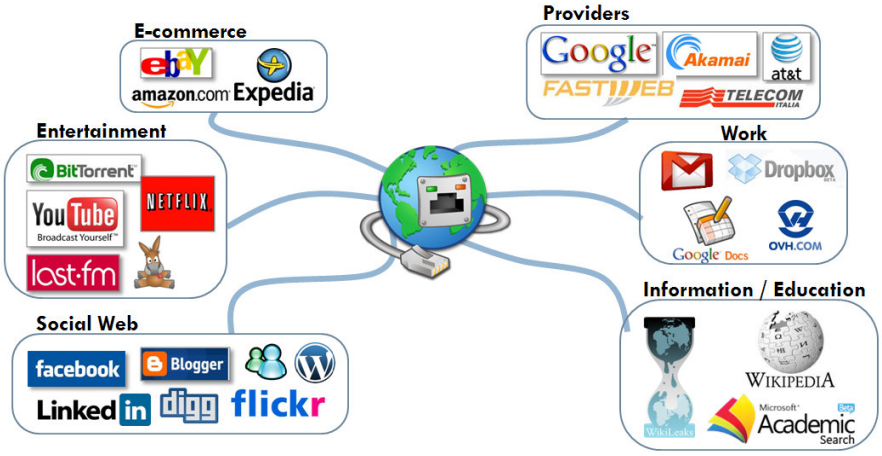
*"The Internet is the first thing that humanity has built that humanity doesn't **understand**, the largest experiment in **anarchy** that we have ever had."*

Eric Schmidt – Google Exec. Chairman

Why we want to study "the anarchy"?

3

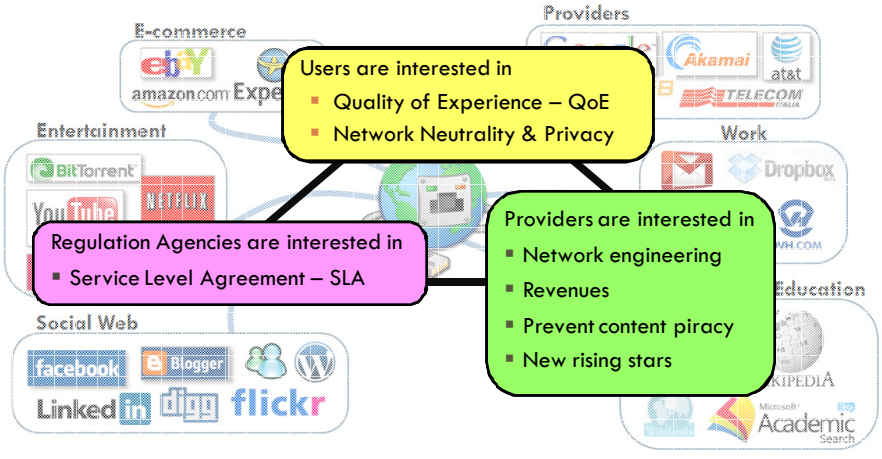
Behind "the anarchy" there are many structured markets where several players participate having different interests



Why we want to study "the anarchy"?

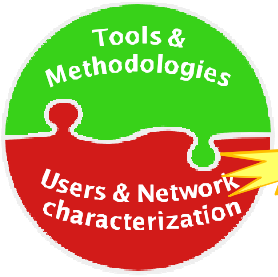
4

Behind "the anarchy" there are many structured markets where several players participate having different interests



How to study Internet?

5



Research

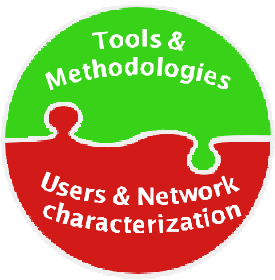
- We need Tools and Methodologies as to
  - Inspect and characterize the traffic both as an aggregate and per flow basis
  - Classify i.e. identifying the traffic related to a specific application

Research

- We can characterize Users and Networks
  - Network engineering & monitoring
    - Capacity provisioning, billing
  - User Applications popularity and volumes
  - Anomaly Detection & Security
    - DDoS, malware

Methodologies

6

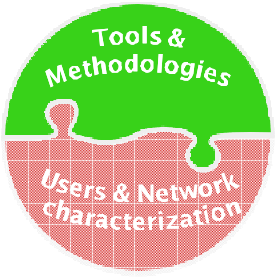


Machine learning techniques: e.g.

- Supervised machine learning
  - Bayesian classifiers, kernel based approaches, SVM, decision trees, neural networks, rule based approaches,…
- Unsupervised machine learning
  - Clustering algorithms (k-means, db-scan, EM,)
  - Augmented **by domain knowledge**

Contributions (1/3)

7

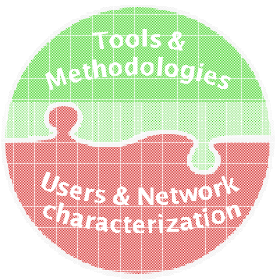


Traffic classification: identify the application that has generated a certain flow

- **DPI – Deep Packet Inspection**: if the packet payload starts with `'Bittorrent'` then is a Bittorrent connection
- **Behavioral**: if the distrib. of the packet inter-arrival and packet size are X and Y then is a VoIP call using a certain voice encoder
- **Machine learning:** map flow features (e.g. length of the first N packets) to an hyper-space and verify how points clusterize

Contributions (2/3)

8

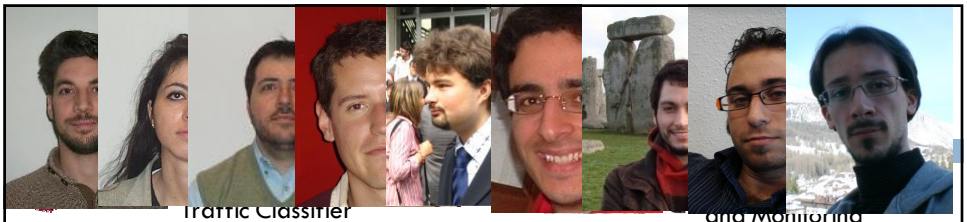


User Characterization: e.g.

- Application popularity
  - Rise of video streaming and drop of P2P
- Access line usage
  - Very few application (Filehosting) can saturate the available download bandwidth

Network and Service Characterization:

- YouTube
- Content Delivery Network - CDN

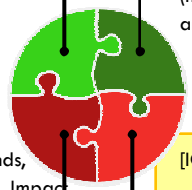


Traffic Classifier

[ToN10] KISS: Stochastic Packet Inspection Classifier for UDP Traffic
[ICC10] Stochastic Packet Inspection for TCP Traffic
[ICC10] Comparing P2PTV Traffic Classifiers
…

and Monitoring

[WWIC10] Live Traffic Monitoring with Tstat: Capabilities and Experiences
Deploy of monitoring probes in Europe (Italy, France, Poland, Hungary, Austria) and US (Purdue University)



[TNSM] Characterization of ISP Traffic: Trends, User Habits and Access Technology Impact (submitted)

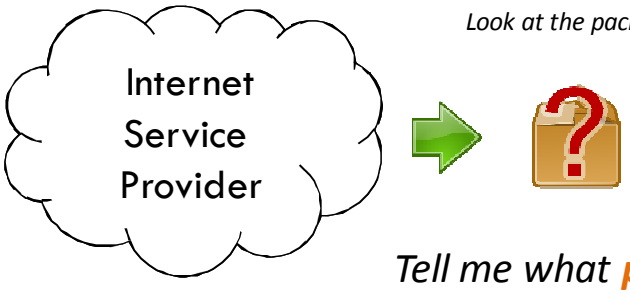
[ICDCS11] Dissecting Video Server Selection Strategies in the YouTube CDN
[IMC11] YouTube Everywhere: Impact of

CISCO FASTWEB vodafone BOEING NARUS

10

THE PROBLEM
THE TOOL

Traffic classification



Look at the packets…

*Tell me what **protocol** and/or **application** generated them*

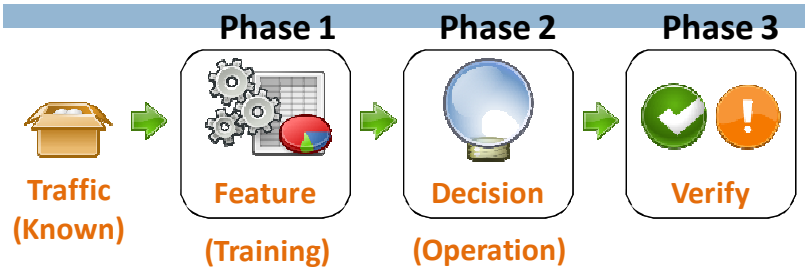
Typical approach:
Deep Packet Inspection (D

It fails more and more:
P2P
Encryption
Proprietary solution
Many different flavours

Port: ?
Payload: *RTP protocol*

Port: …/4672 ?
Payload: *E4/E5*

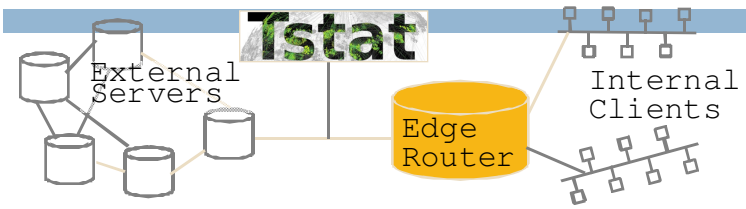
Possible Solution: Behavioral Classifier



1. **Statistical** characterization of traffic (given source)
2. Look for the **behaviour** of unknown traffic and assign the class that better fits it
3. Check for possible classification mistakes

METHODOLOGY: SUPERVISED MACHINE LEARNING

Tstat

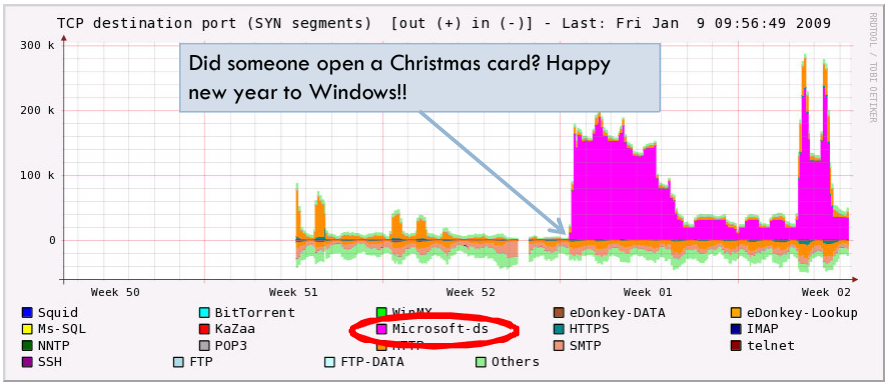


- Traffic classifier
  - Deep packet inspection
  - Statistical methods
- Persistent and scalable monitoring platform
  - Round Robin Database (RRD)
  - Histograms
- Open Source

TOOL & DATA: software & packets

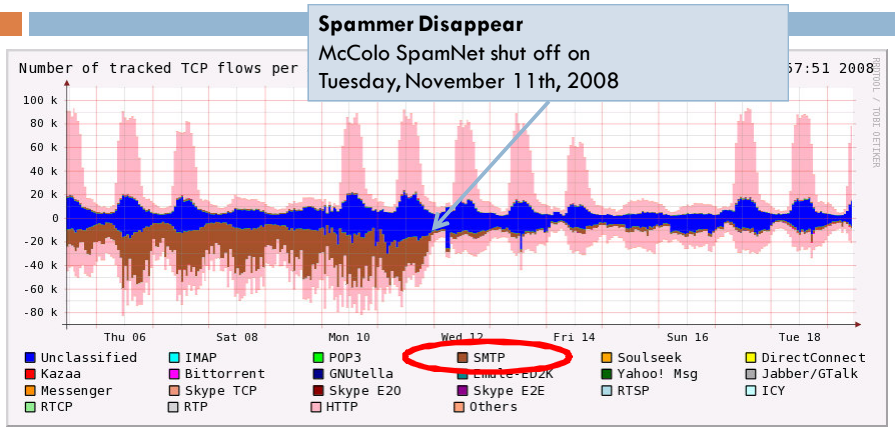
http://tstat.tlc.polito.it

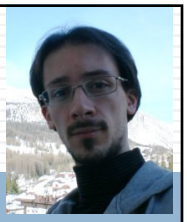
Worm and Viruses?



Results: examples of nice things you may see

Anomalies (Good!)





Dissecting YouTube

[ICDCS11] Dissecting Video Server Selection Strategies in the YouTube CDN

[IMC11] YouTube Everywhere: Impact of Device and Infrastructure Synergies *on User Experience*

Collaboration with:

Prof. Sanjay Rao

Dr. Ruben Torres

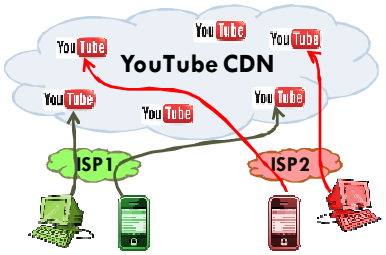
PURDUE UNIVERSITY

Motivations

18

YouTube is the most popular video download system on the Internet (*)

- 13 million hours of video uploaded during 2010
- It is a big share of the mobile traffic
  - more than 500 tweets per minute contain a YouTube link



YouTube Data Center

Questions:

- What is the system design?
- How the system handle PC or Mobile requests?
- What about the performance?

(*) *www.youtube.com/t/press_statistics*

Which devices?

19

We separate the devices in two categories:

PC-player:

- Regular PC / Laptop / nettop having a web browser with the Adobe Flash plugin or that is HTML5 compliant

Mobile-player:

- A smarthphone, an Internet Tablet or a set-top-box using a custom application to access to YouTube
- No distinction/difference among the different operating systems

[CoNEXT11] Network Characteristics of Video Streaming Traffic (INRIA)

Collection Tool

20

- Traffic classification using Tstat (*)
  - L4 (TCP) statistics
    - Per-connection statistics (#bytes, #pkts, ...)
  - L7 DPI to inspect the HTTP messages
    - Classify the type of content and device
    - Identify the "control" messages
    - Per-video statistics (video duration, resolution, codec, ...)

(*) http://tstat.polito.it

Data sets

21

Name	Type	Flows	Vol.[GB]	SrcIP	Videos
US-Campus	Campus	2,172,250	10,898	20,455	446,870
EU1-Campus	Campus	173,024	714	1,203	50,205
EU1-ADSL	Home	740,330	2,615	8,154	189,788
EU1-FTTH	Home	135,907	480	1,136	33,762
EU2-ADSL	Home	830,476	3,688	5,826	205,802

Week-long collections on Sep. 2010

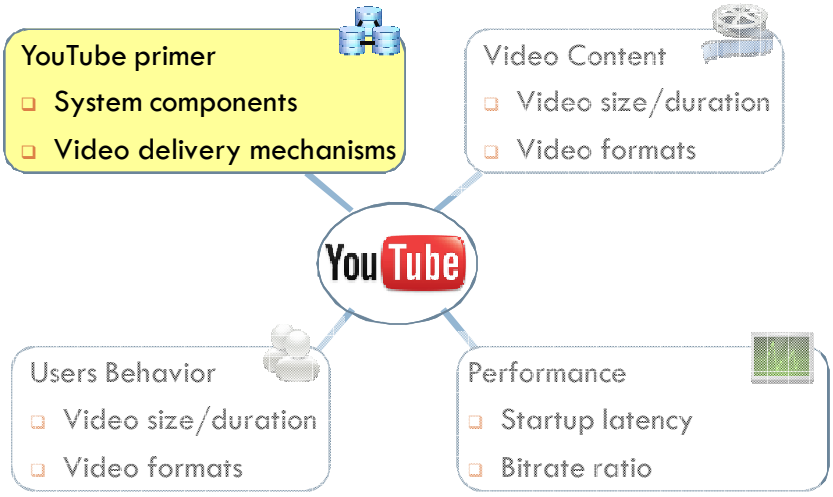
- 5 vantage points in Europe and US
- 4 access technologies - ADSL, Fiber-To-The-Home, Ethernet, WiFi
- Both Residential ISPs and Campus networks
- Mobile-player access YouTube via WiFi

Methodology: data mining

[IMC'11] Over the Top Video: the Gorilla in Cellular Networks (AT&T)

Roadmap of the results

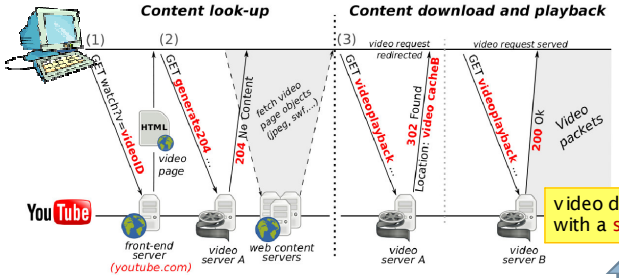
22



YouTube primer

23

PC-player Download



video downloaded with a single flow

memory constraints

video downloaded with multiple flow

Methodology: domain knowledge

Roadmap of the results

24

YouTube primer
- System components
- Video delivery mechanisms

Video Content
- Video size/duration
- Video formats

YouTube

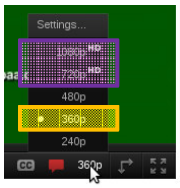
Users Behavior
- Video size/duration
- Video formats

Performance
- Startup latency
- Bitrate ratio

Video Content

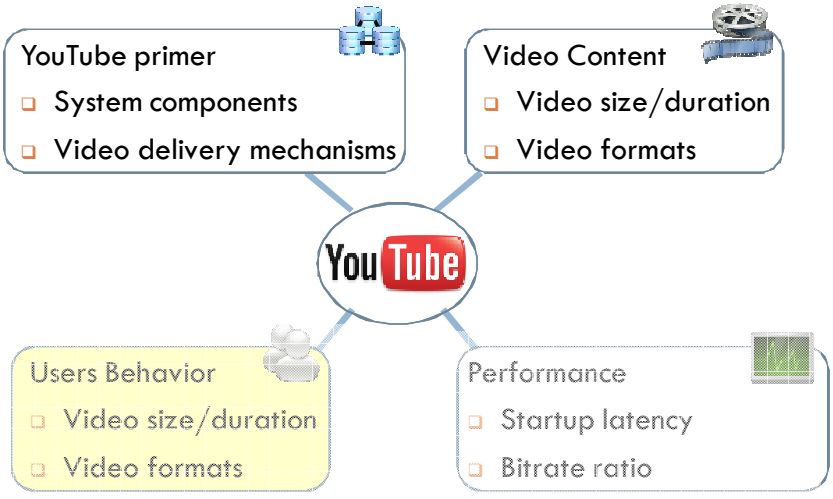
25

- 300-400M videos in the system
- Different video duration and size
  - Median duration: 3/5 min
  - Median size: 10MB
- More than 10 video format supported
  - The user just see a simple menu
  - 360p is the default resolution
    - PC use Flash as container while is Mpeg for Mobile
  - HD is negligible from PC but >2% from Mobile
    - Mobile devices download by default the best resolution available



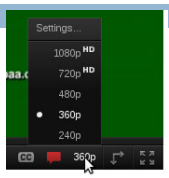
Roadmap of the results

26



Probability of resolution switch

27



Resolution switch:
The user starts to download in RES1 (e.g. 360p) and then jump to RES2 (e.g. 720p)

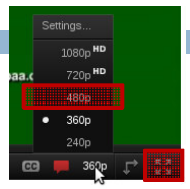
Let's focus on this

	PC-player			Mobile-player		
Data set	0	1	>1	0	1	>1
US-Campus	95.10	4.60	0.30	99.75	0.19	0.05
EU1-Campus	96.62	3.12	0.27	99.28	0.61	0.10
EU1-ADSL	95.27	4.45	0.28	99.63	0.28	0.09
EU1-FTTH	95.73	3.99	0.28	99.39	0.42	0.19
EU2-ADSL	95.14	4.40	0.46	98.07	1.36	0.57

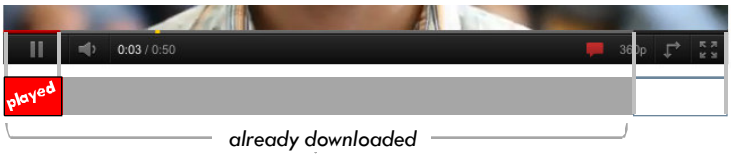
- Users stick to the default playback parameters!
- Why so? Intuition is that
  - users are not aware of this possibility
  - it is "difficult" to change resolution
  - inertia

Users behavior (1/3)

28



- Users do not change resolution
  - Only 5% perform 360p → 480p (full screen)
- Do the users watch the whole video?

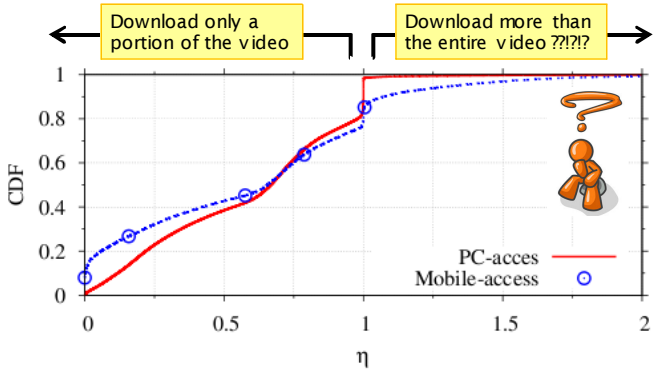


Fraction of video downloaded

29

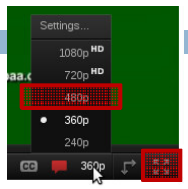
For each video session we compute:

$$\eta = \text{Fraction of video downloaded} = \frac{\text{Downloaded bytes}}{\text{Full video bytes}}$$

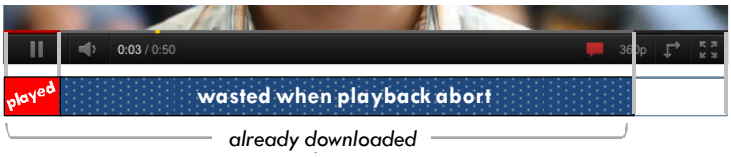


Users behavior (1/3)

30



- Users do not change resolution
  - Only 5% perform 360p → 480p (full screen)
- Do the users watch the whole video?

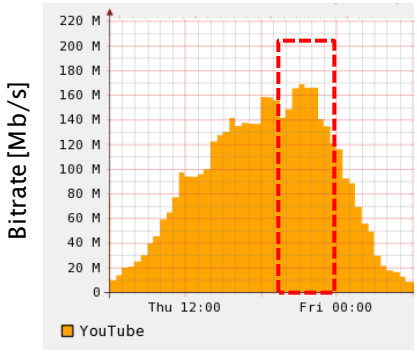


- Tstat monitor the progression of the download
  - If the TCP connection last for less than the video duration then the playback has been abruptly aborted
  - >80% of the playback are aborted during the download
  - Abruptly aborting the playback introduce waste of bandwidth

Users behavior (3/3)

31

Can we estimate the bandwidth wasted?



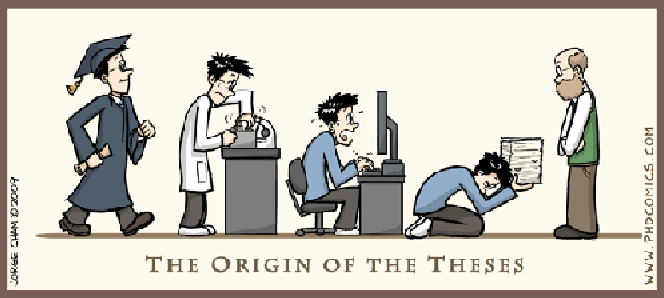
160Mb/s of YouTube @ peak hours

67Mb/s of traffic wasted!!!

Conclusions

32

- YouTube is an example of successful service but it is far from been optimized
  - Caching policies, bandwidth wasted, ecc…
- Generally speaking, we need measurement
  - Internet is in constant evolution
  - Operators are interested in acquiring and exploiting such knowledge
  - Methodologies need to be continuously updated
- It's fun to see what we do on the Internet
  - An to see how things works


THE ORIGIN OF THE THESES
WWW.PHDCOMICS.COM

?? || ##