

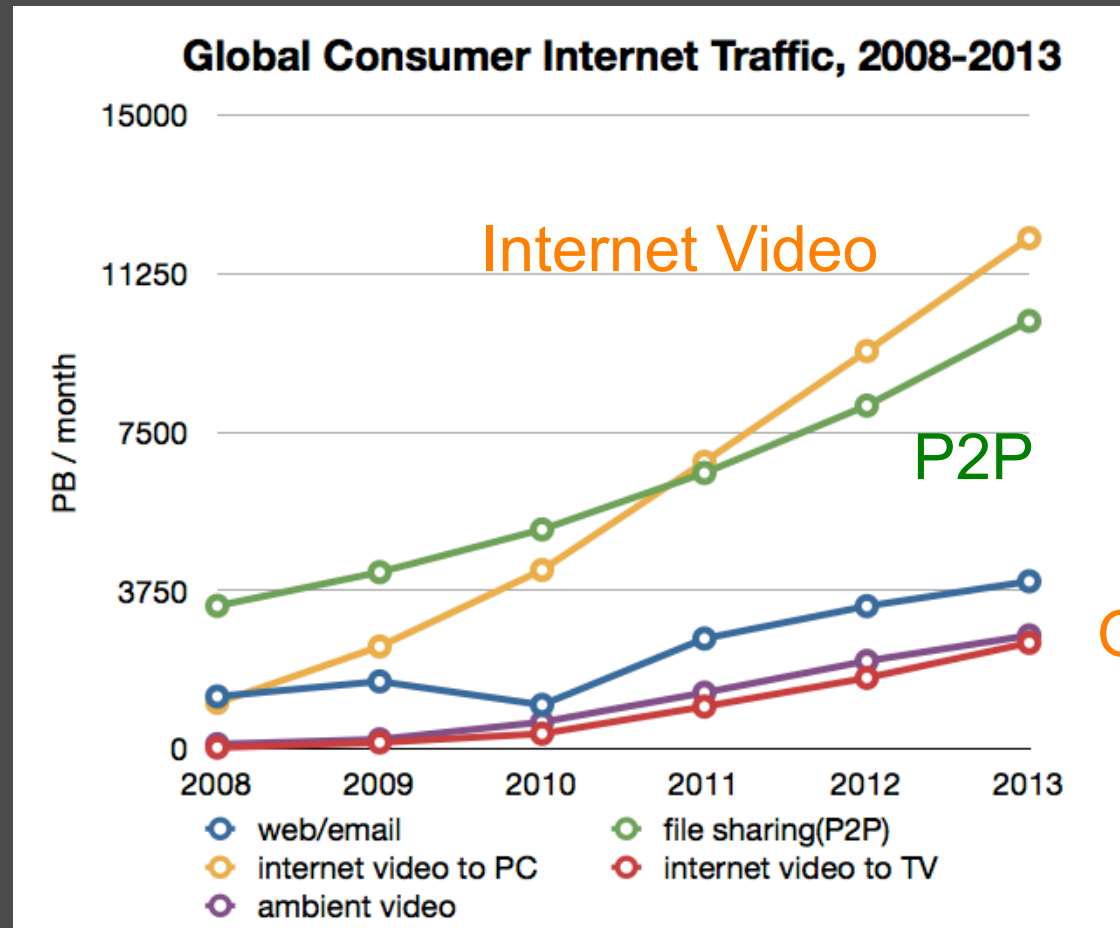
CHARACTERIZING TRAFFIC FLOWS ORIGINATING FROM LARGE-SCALE VIDEO SHARING SERVICES

TMA workshop 2010
April 7 2010

T. Mori, R. Kawahara,
H. Hasegawa, and S. Shimogawa

NTT Service Integration Laboratories

Background



Other Videos

Statistics cited from:
Cisco Systems, Visual Networking Index
Forecast and Methodology, 2008-2013





Goal of this work

- ④ Understanding the new class of traffic – video sharing services (Internet video)
 - For creating realistic traffic workload model
 - Addressing implications for network management

What we did in this work:

- ① Developed a simple and effective technique that identifies flows originating from video sharing services
- ② Revealed the basic characteristics of network traffic flows of video sharing services from a network service provider view

Video sharing services we studied

Service	Location of Provider	Capacity (free)	Capacity (premium)	Content type
YouTube 	US	2GB and 10 mins	20GB	flv, 3gp, mp4
Smile video 	Japan	40MB	100MB	flv
MEGAVIDEO 	China(HK)	100MB	5GB	flv
Dailymotion 	France	20mins or 150MB	--	flv, mp4

Identifying video flows

⦿ Constraint:

- We take an approach that does not use DPI (Deep Packet Inspection)
- DPI does not scale (think of 100+Gbps world)
- User privacy issues.

⦿ Idea:

- Use source IP addresses as a hint
- YouTube video traffic may be originating from AS 36561 (YouTube)
- But, we need more fine-grained information
 - YouTube is also originating from Google AS
 - Some other objects are also originating from YouTube

Web objects originating from YouTube

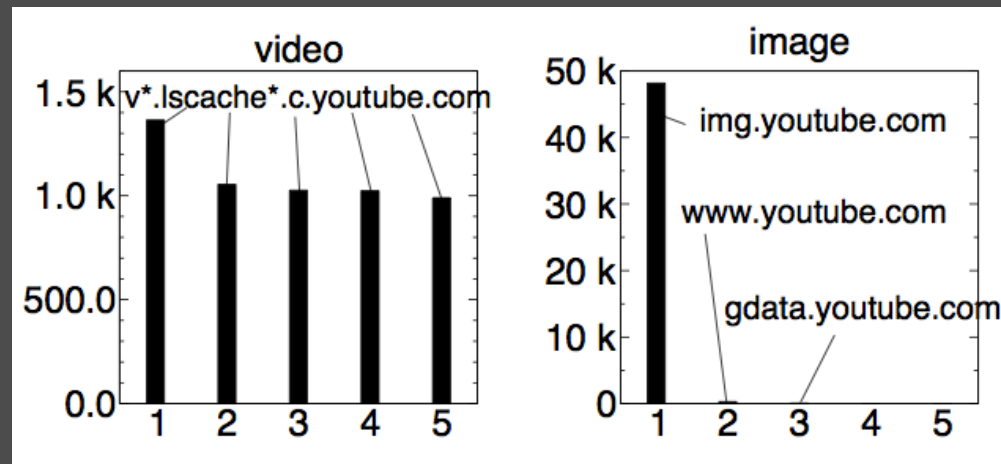
◎ Various http/content-types

- Application:
 - atom+xml, javascript, octet-stream, x-shockwave-flash, etc.
- Image:
 - gif, jpeg, png, x-icon
- Text:
 - html, css, javascript, plain, x-cross-domain-policy, xml
- Video:
 - mp4, flv, x-flv

◎ Observation:

- each content type is served by different server groups in large web farms

Separation of server roles in large-scale web farms



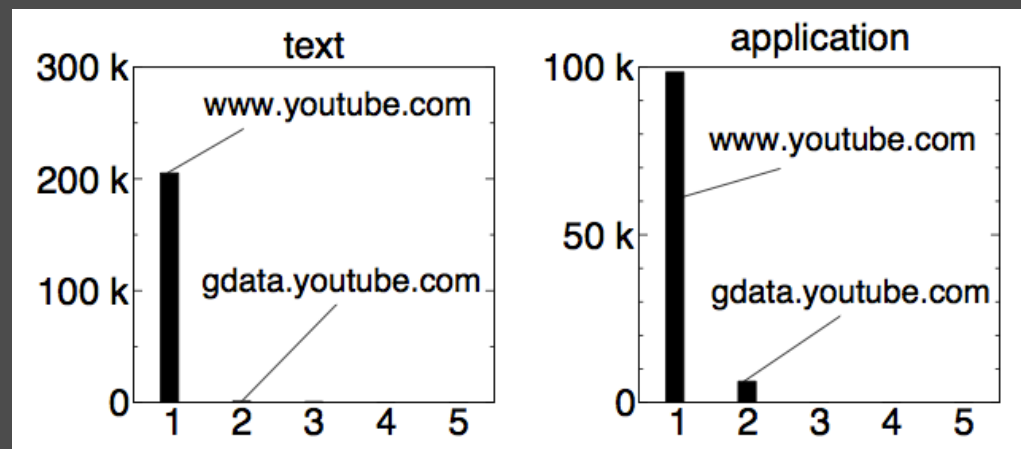
Video: `v*.lscache*.c`

Image: `img`

Text: `www`

Apps: `www, gdata`

These hostnames are assigned different set of IP addresses



Extracting YouTube hosting servers (1)

1. Analysis of massive web URLs

- web proxy server logs (100M records / month)
- Extract hostname if it is serving content-type of “video” and its domain part matches /youtube\.com\$/
- Reverse engineering the naming rules of hostname: v*.cache*.c.youtube.com

2. Compile a hostname list

- Complement unobserved hostnames.
If we see v12 and v14 → we may also have v13
- If a domain has proper A record, add it in the list.

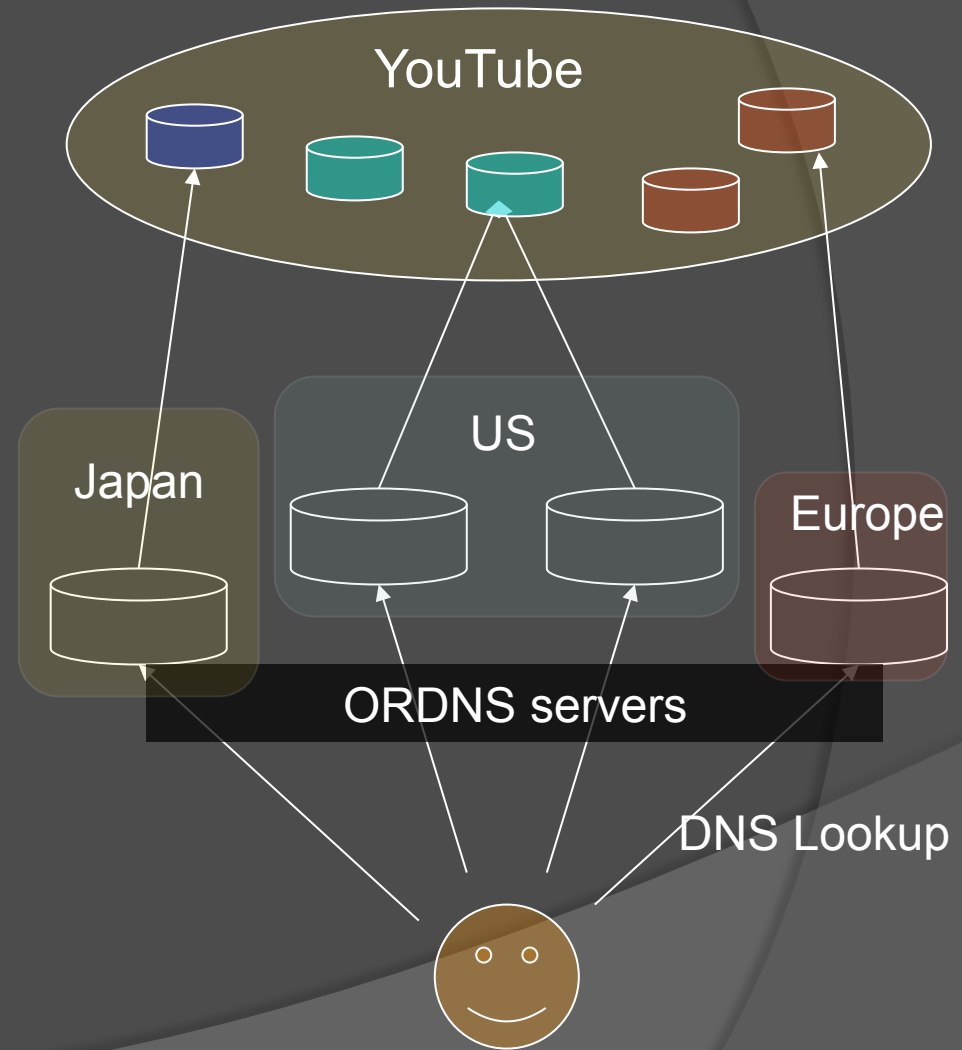
Extracting YouTube hosting servers (2)

3. Compile IP address list

- A host name is assigned multiple IP addresses (for the purpose of efficient content delivery)
- Lookup the hostnames from multiple vantage points.
- We used open recursive DNS servers (ORDNS) that are located over the world (5000+ servers in 68 countries).



We now have list of IP addresses for YouTube video hosting servers



Extracted hostnames and IP addresses

hostnames

hostnames	complemented range	#observed transactions
v \odot .lscache \otimes .c	$1 \leq \odot \leq 24, 1 \leq \otimes \leq 8$	130,286
v \odot .cache \otimes .c	$1 \leq \odot \leq 8, 1 \leq \otimes \leq 8$	27,485
tc.v \odot .cache \otimes .c	$1 \leq \odot \leq 24, 1 \leq \otimes \leq 8$	1626
v \odot .nonxt \otimes .c	$1 \leq \odot \leq 24, 1 \leq \otimes \leq 8$	25
lax-v \odot .lax	$1 \leq \odot \leq 308$	19
sjl-v \odot .sjl	$1 \leq \odot \leq 50$ (with exceptions)	19

service	hostnames	complemented range
Smiley videos	smile-com $\odot\otimes$.nicovideo.jp	$0 \leq \odot \leq 6, 0 \leq \otimes \leq 3$
Smiley videos	smile-clb $\odot\otimes$.nicovideo.jp	$0 \leq \odot \leq 6, 0 \leq \otimes \leq 3$
Megavideo	www \odot .megavideo.com	\odot can be any positive integer.
Dailymotion	proxy- $\odot\otimes$.dailymotoin.com	$0 \leq \odot \leq 9, 0 \leq \otimes \leq 9$

IP addresses

Service	#addresses
YouTube	2,138
Smiley videos	74
Megavideo	670
Dailymotion	100

Traffic dataset

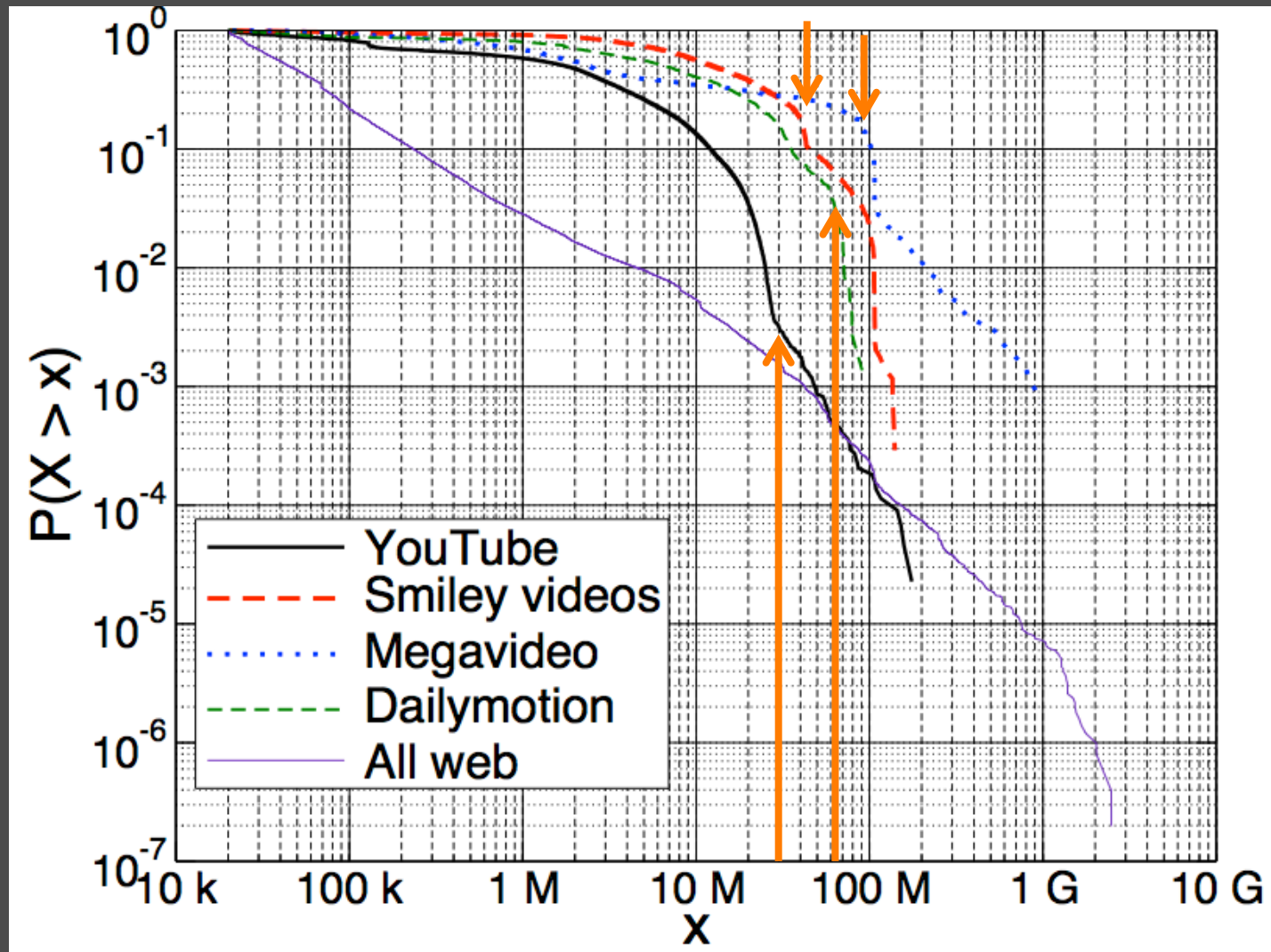
- ⦿ Collected on GbE edge link of an Internet edge site
- ⦿ Incoming traffic collected over 9 hours
- ⦿ Non-sampled flow records:
 - Start time, end time, server IP, client ID, source port, destination port, #pkts, size
- ⦿ # of total flows: 100 M

Statistics of collected flows

Service	#flows	mean size	mean rate	mean duration
YouTube	43,960	4.1 MB	1.3 Mbps	41.8 sec
Smiley videos	3,438	21.3 MB	2.6 Mbps	139.8 sec
Megavideo	1,354	30.0 MB	1.3 Mbps	232.6 sec
Dailymotion	730	13.7 MB	1.5 Mbps	96.0 sec
All web	5,043,927	0.33 MB	0.9 Mbps	16.5 sec

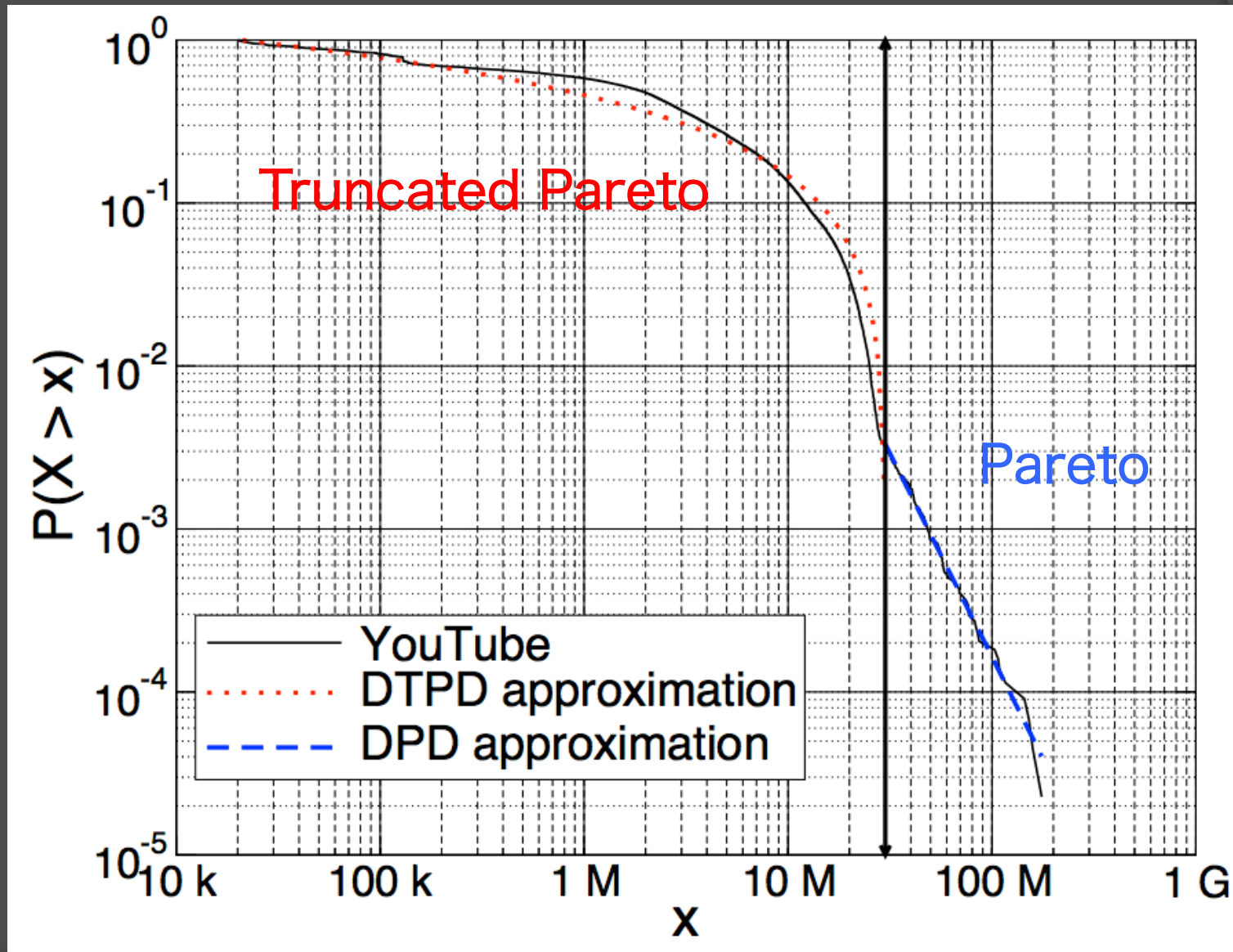
Flows whose sizes > 20KB

Flow size distribution



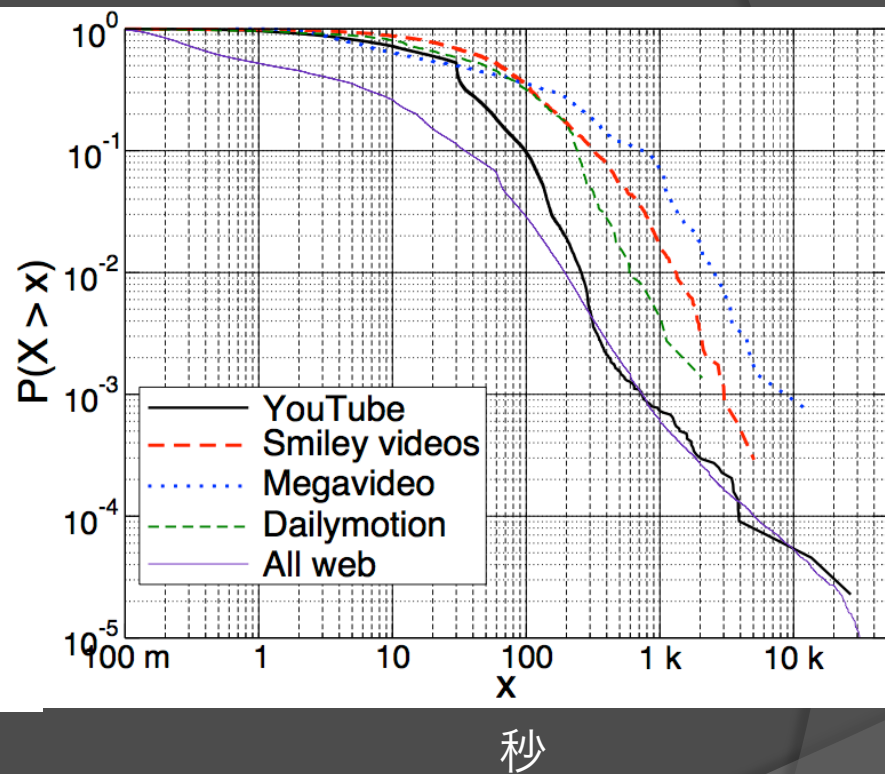
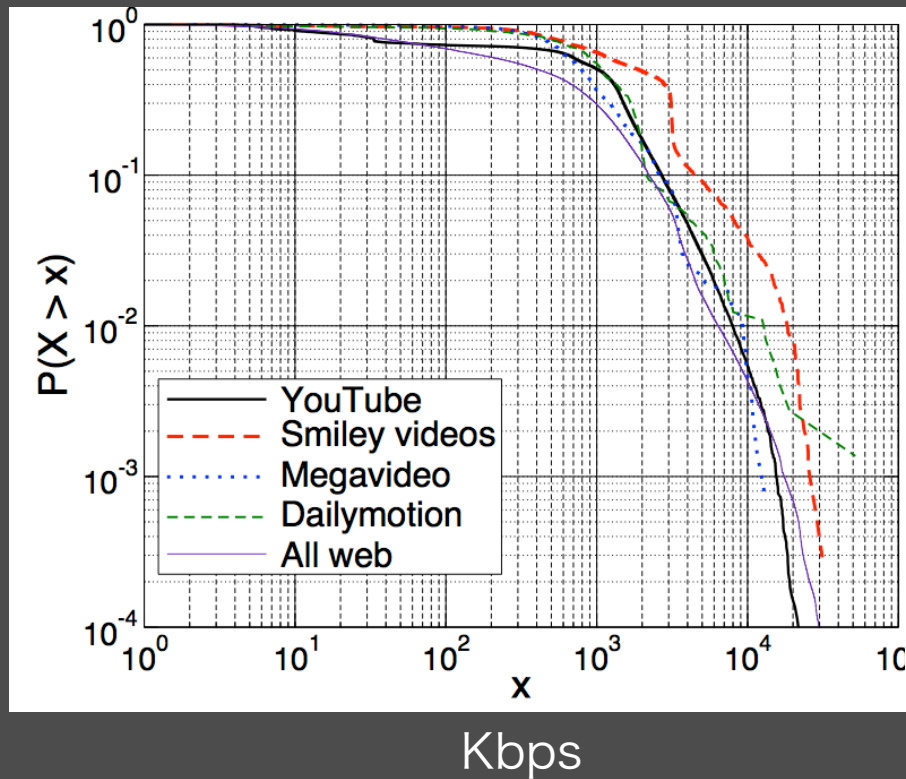
Capacity limitation affects the characteristics

Characterizing the distributions



It looks like a mixture distribution

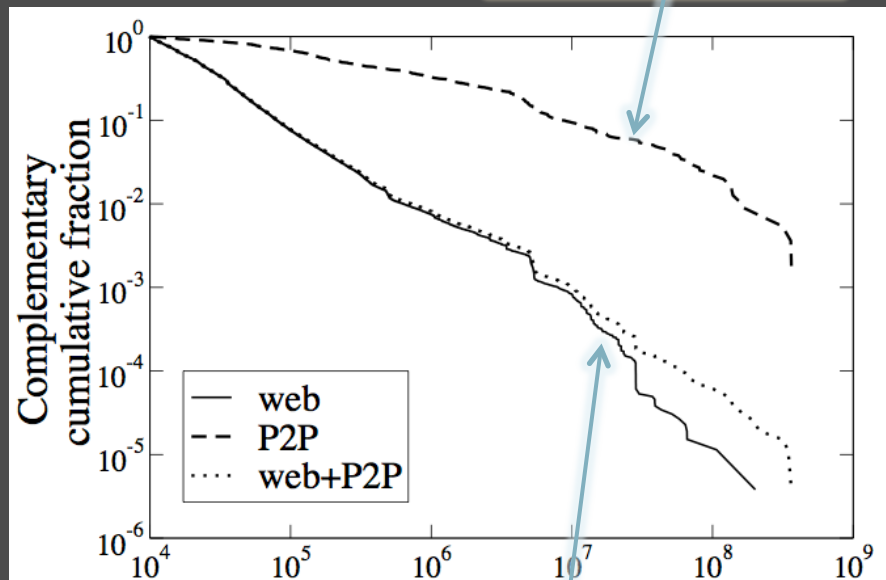
Flow rate / Flow duration



Not much differences among services.
Duration obeys power-law-like distribution.

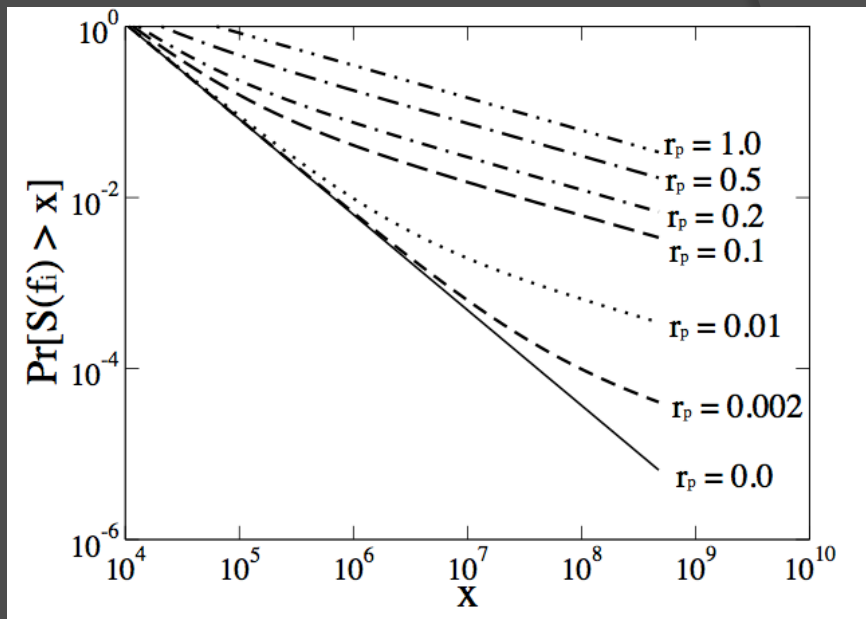
Comparison with conventional P2P flows

P2P: $\theta=0.38$



Web: $\theta=1.12$

Impact of P2P growth



Tatsuya Mori, Masato Uchida, Shigeki Goto,
Flow analysis of internet traffic: World Wide Web versus peer-to-peer.
Systems and Computers in Japan 36(11): 70-81 (2005)

Summary and Implications

- ⦿ Presented a simple and effective way of identifying traffic flows originating from video sharing services
Key idea: leverage naming convention of large-scale web farms + globally distributed measurement
- ⦿ Capacity limitation of video sharing services affects the characteristics of traffic flow sizes
Can be a control parameter both for content provider and ISPs
- ⦿ If video sharing services do not have capacity limitation, their traffic might be close to P2P traffic today (assumption)