Analysing YouTube Traffic

Irina Antonova
Seminar: Internet Measurement
Berlin, den 23.02.2009

Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik
Intelligente Netze und Management verteilter Systeme (INET)
Research Group Prof. Anja Feldmann, Ph.D.
An-Institut Deutsche Telekom Laboratories

Einleitung:

- YouTube als Beispiel für Web2.0

  > Web2.0
    > Dynamische Websites
      = jeder kann Inhalte publizieren
    > Interaktion
      = z.B. Rate, Comment, Reply

  > Kontrast zu Web1.0
    > Statische Websites
    > begrenzte Anzahl Content Provider
    > begrenzte Menge Content

Einleitung: Traffic Analyse

- Traffic Analyse
  - Beobachtung des Netzwerkes
    (z.B. Packet Capturing / Web Crawler)
  - Identifizieren von Charakteristiken
    (z.B. Auslastung, Verlauf)
  - Abschätzen zukünftigen Netzwerkerhaltens
- Ziel der Web2.0 Traffic Analyse
  - Management des steigenden Datenvolumens
  - Zur Entwicklung von Netzwerk-Management- und Kapazitätsplanungsstrategien

Einleitung: YouTube =Web2.0 ?

- größte Video-Sharing Website
- Ca. 60% aller Web-Videos
- Zuwachsrate von 65 000 Videos/Tag
  - > YouTube als Repräsentant für Web2.0

Überblick

- Einleitung
- Methodologie
- Auswertung
  - Datenbestand
  - Videocharakteristiken
  - Popularität der Videos
  - Lokalitätscharakteristiken
- Zusammenfassung
- Literatur

Methodologie: A View from the Edge

- 3-Monatiges Beobachten von YouTube-Traffic
- Lokale Perspektive:
  - Uni-Campus
  - Packet Capturing: Protokollierung aller YouTube relevanten Packete
- Globale Perspektive:
  - Untersuchung der Most-Popular-Listen auf www.youtube.com

Methodologie: I Tube, You Tube, Everybody Tubes

- Crawl YouTube
  - Sammeln der Meta-Informationen zu allen Videos in einer Kategorie
  - 6 Tage in Folge werden u.a. Alter, Views und Rating zu Videos protokolliert
- Beschränkt auf die Kategorie „Science & Technology“

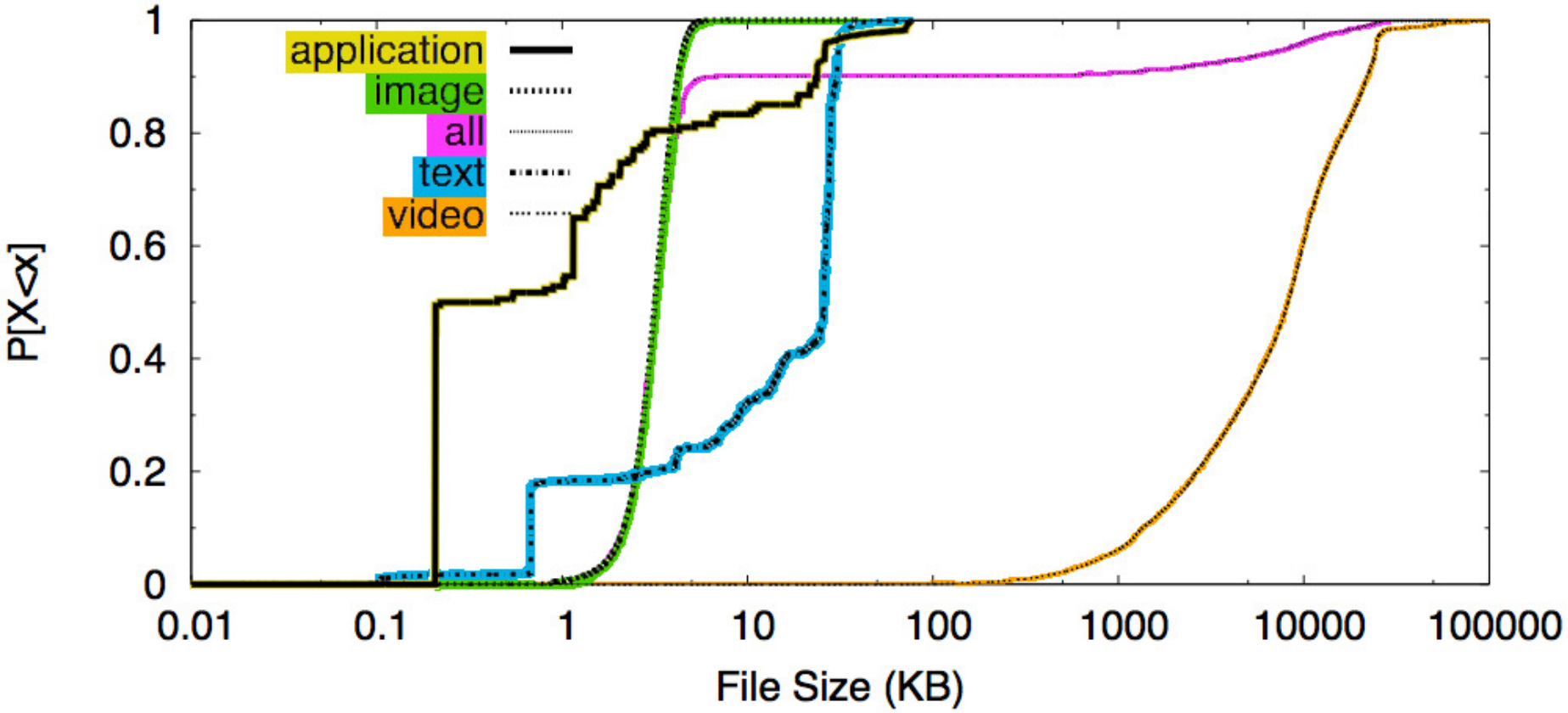
Auswertung: Datenbestand(Edge)

- Datenaufnahme über 85 Tage
- Lokaler YouTube-Datenbestand

Item	Information
Start Date	Jan. 14, 2007
End Date	Apr. 8, 2007
Total Valid Transactions	23,250,438
Total Bytes	6.54 TB
Total Video Requests	625,593
Total Video Bytes	6.45 TB
Unique Video Requests	323,677
Unique Video Bytes	3.26 TB

  - > Videoanfragen machen >98% des Datenvolumens aus
  - > nur ~50% der Videos sind unterschiedlich

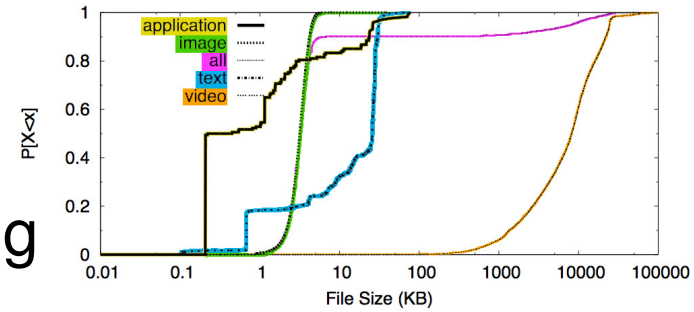
Auswertung: Dateigrößen



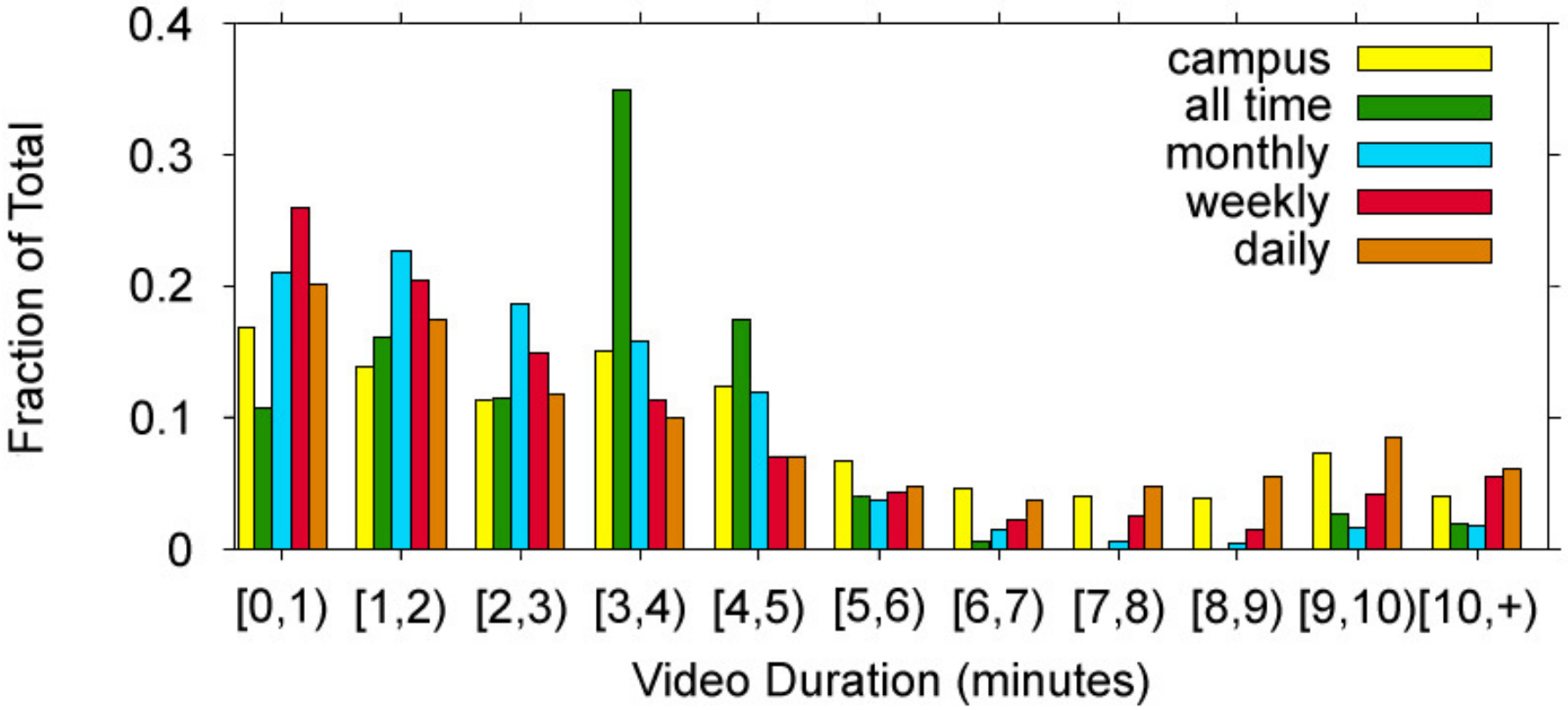
Auswertung: Dateigrößen

- aus Content-Length-Feld in Http-Header
- Haupsächlich werden Videos und Bilder übertragen
- Videodateigröße zw. 500 und 100 000KB
  - > größer als bei anderen Dateitypen

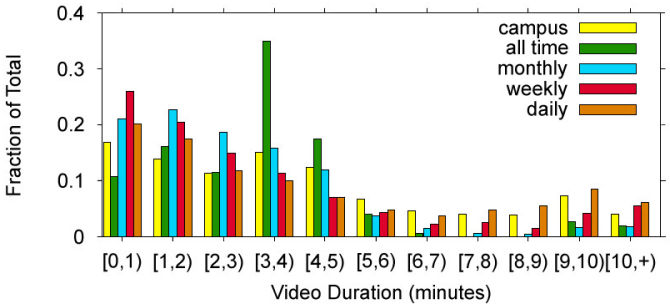
⇒ Netzwerkentlastung durch Caching



Auswertung: Videodauer



Auswertung: Videodauer



- mit YouTube-API erfragt
- Durchschnittliche Videodauer: 4,15min
- „Most-Popular“-Videos haben kürzere Dauer:
  - > 0-1min größter Anteil in Daily-Listen
  - > 3-4 min größter Anteil in All-Time-Listen

Auswertung: Popularität von Videos

- Untersuchung des Zusammenhangs zwischen der Aufrufhäufigkeit von Videos und deren Popularität
  - Edge: Aufrufhäufigkeit lokal gemessen (Campus Netz)
  - I Tube: Aufrufhäufigkeit anhand Viewanzahl (Global innerhalb YouTube Kategorie „Science & Technology“)
- Methoden:
  - Zipf-Regel
  - Konzentrationsanalyse / Pareto

Auswertung: Popularität von Videos

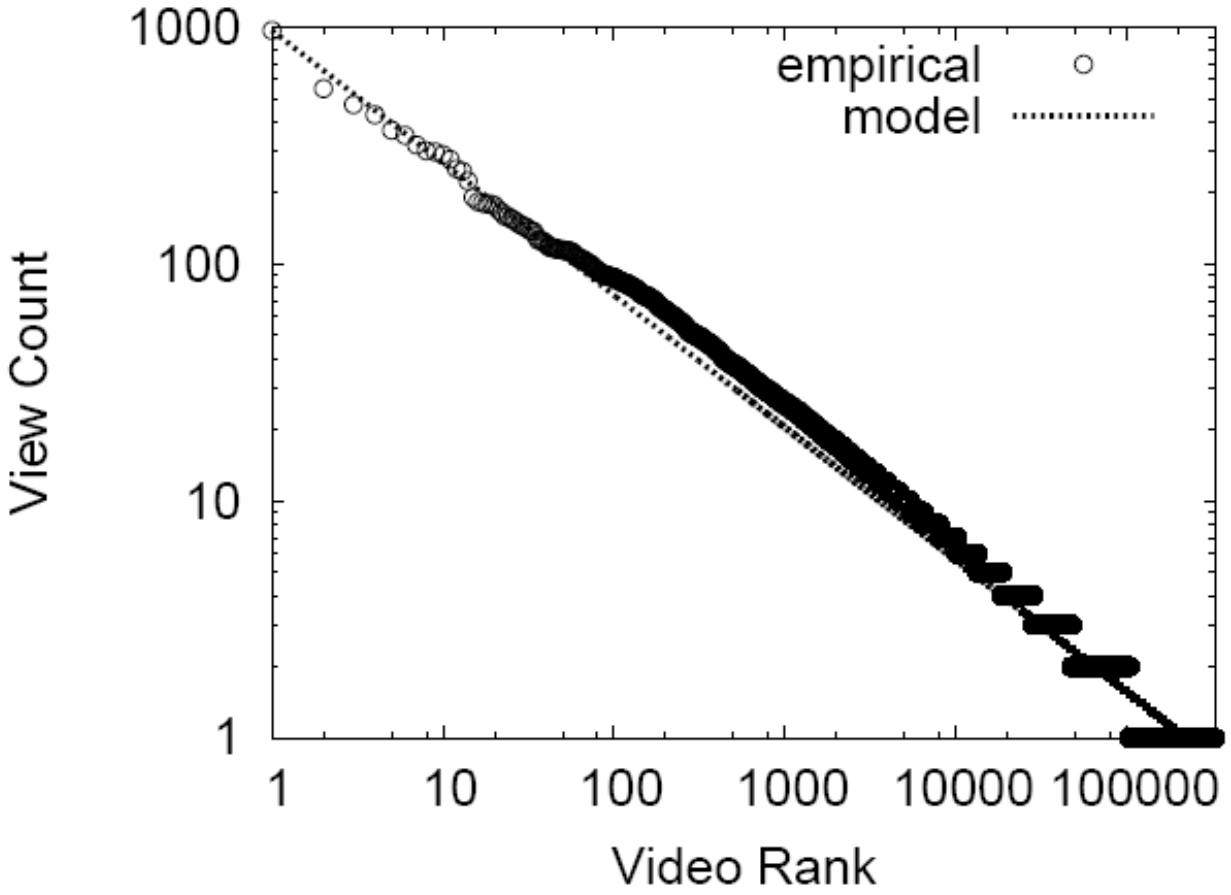
Zipf-Regel

- Anwendbarkeit der Zipf-Regel:
  - Sortieren aller Videos anhand ihrer Aufrufhäufigkeit (Beliebtestes Video bekommt Rang 1, unbeliebtestes Rang n)
  - Plotten der Aufrufhäufigkeit und des Rangs in ein logarithmisches Koordinaten-System

$\Rightarrow$ Ergibt sich eine gerade Linie, ist die Zipf-Regel anwendbar

Auswertung: Popularität von Videos

Zipf-Regel

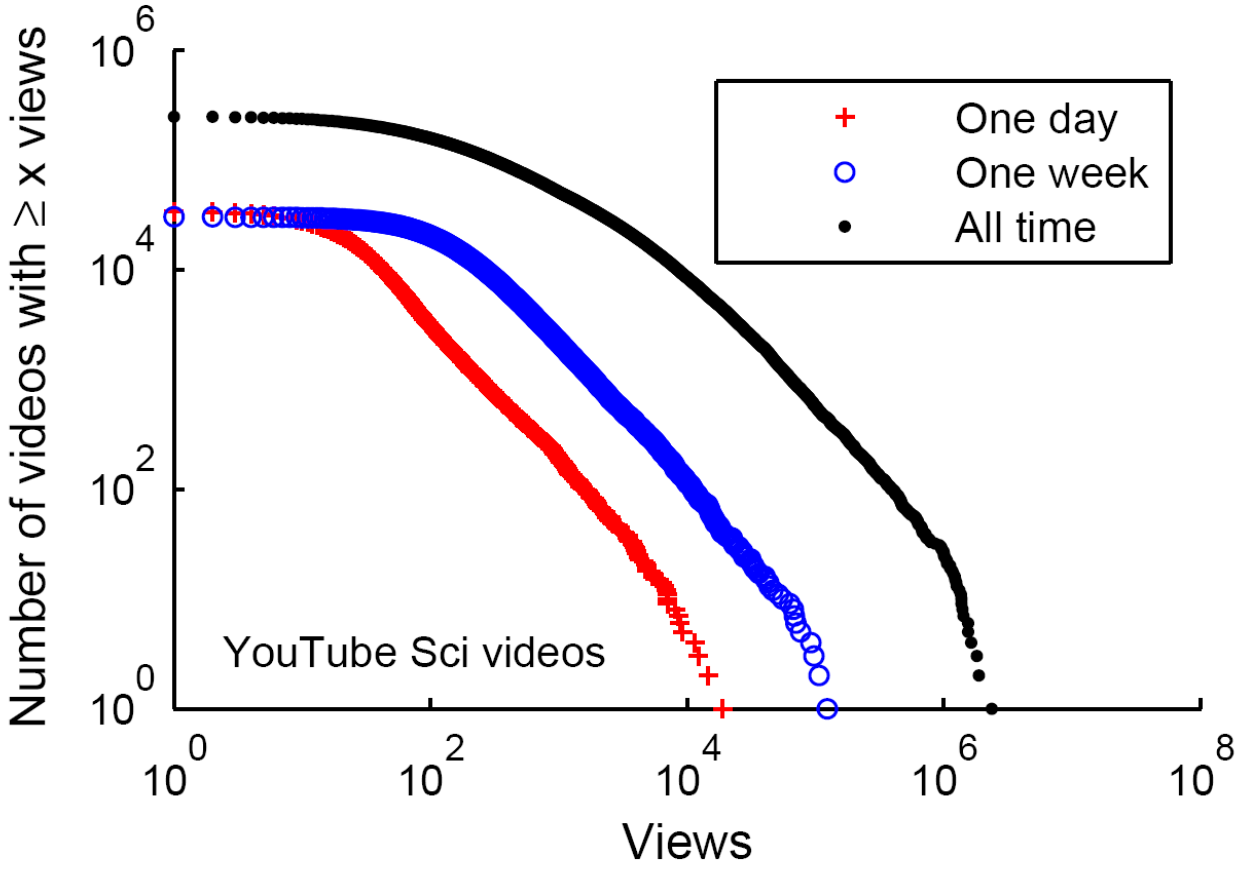


⇒ Zipf-Regel

(DatenBasis: Edge / Campus)

Auswertung: Popularität von Videos

Zipf-Regel



⇒ Zipf-Regel mit truncated Tail

(Datenbasis: YouTube Kategorie „Science & Technology“)

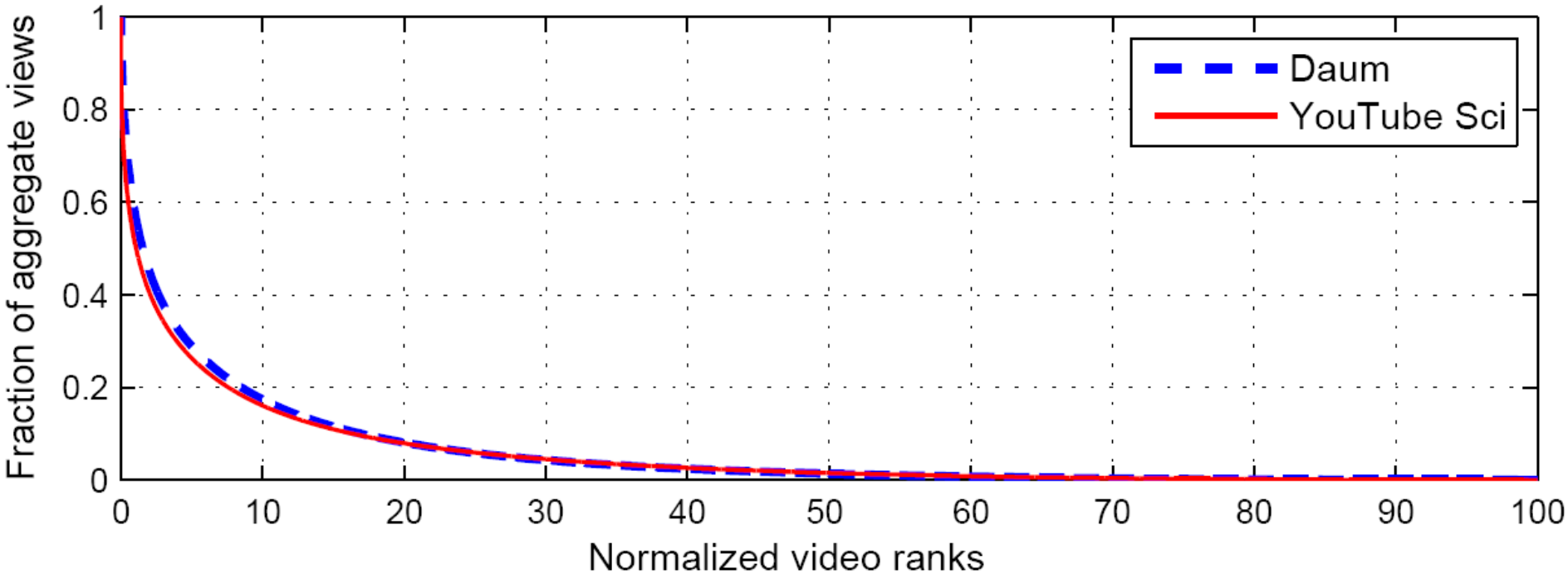
Auswertung: Popularität von Videos

Konzentrationsanalyse / Pareto

- Konzentrationsanalyse:
  - Überprüfung auf welchen Anteil aller vorhandenen Dateien sich der Großteil der Anfragen konzentriert
- Pareto-Regel:
  - Geringer Anteil an Gesamtmenge verursacht Großteil der Arbeit
    ~ Geringer Videoanteil verursacht Großteil der Anfragen

Auswertung: Popularität von Videos

Konzentrationsanalyse / Pareto



⇒ 10% der Videos machen 80% der Gesamtanzahl der Aufrufe aus
(Datenbasis: YouTube Kategorie „Science & Technology“)

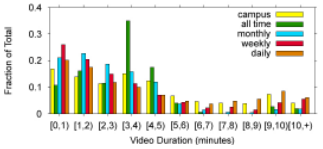
Auswertung: Lokalitätscharakteristiken

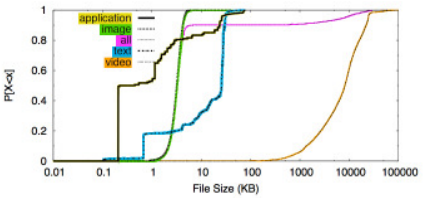
- Def.: Temporale Lokalität
  - Ableiten der Ereignisse der nahen Zukunft
    aus
    Ereignissen der jüngsten Vergangenheit
- Beobachtung am Campus:
  - 5-10% an Tag X aufgerufene Videos werden am Folgetag erneut aufgerufen

Zusammenfassung

Edge / Campusnetz:

- 98,8% des Datenvolumens im Campusnetz sind verursacht durch Videoanfragen
- Videodauer der relevanten Videos ist in 99% weniger als 10min
- 5-10% der Videos werden am Folgetag erneut aufgerufen

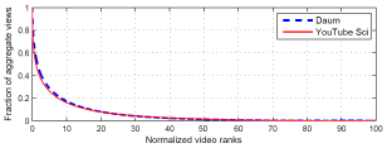


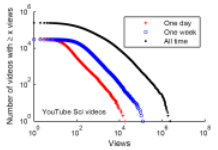


⇒ Potential zur Netzwerkentlastung durch Videocaching

YouTube / „Science & Technology“

- Aufrufhäufigkeit hängt von Videopopularität ab
  - Konzentration aus kleinen Anteil der gesamten Videos
  - unpopuläre Videos „gehen unter“ > Promoting-Potential





⇒ Konzentrationsverhältnisse gegeben wegen Informationsengpass

Literatur

- [1] Phillipa Gill, Martin Arlitt, Zongpeng Li, Anirban Mahanti. *YouTube Traffic Characterization: A View From the Edge*, IMC 2007
- [2] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. *I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System*, IMC 2007

- [3] Business Intelligence Lowdown. Top 10 Largest Databases in the World, Feb. 2007.
- [4] M. Gittens, Kim Yong, D. Godwin, *The vital few versus the trivial many: examining the Pareto principle for software,* COMPSAC 2005. 29th Annual International
- [5] D. N. Serpanos, G. Karakostas, W.H. Wolf, *Effective Caching of Web Objects using Zipf's Law,* ICME 2000. 2000 IEEE International Conference on Multimedia and Expo2000
- [6] USA Today. YouTube Serves up 100 million Videos a Day Online, July 2006