Fake View Analytics in Online Video Services

Liang Chen, **Yipeng Zhou**, Dah Ming Chiu

Shenzhen University

The Chinese University of Hong Kong

What is “Fake View”

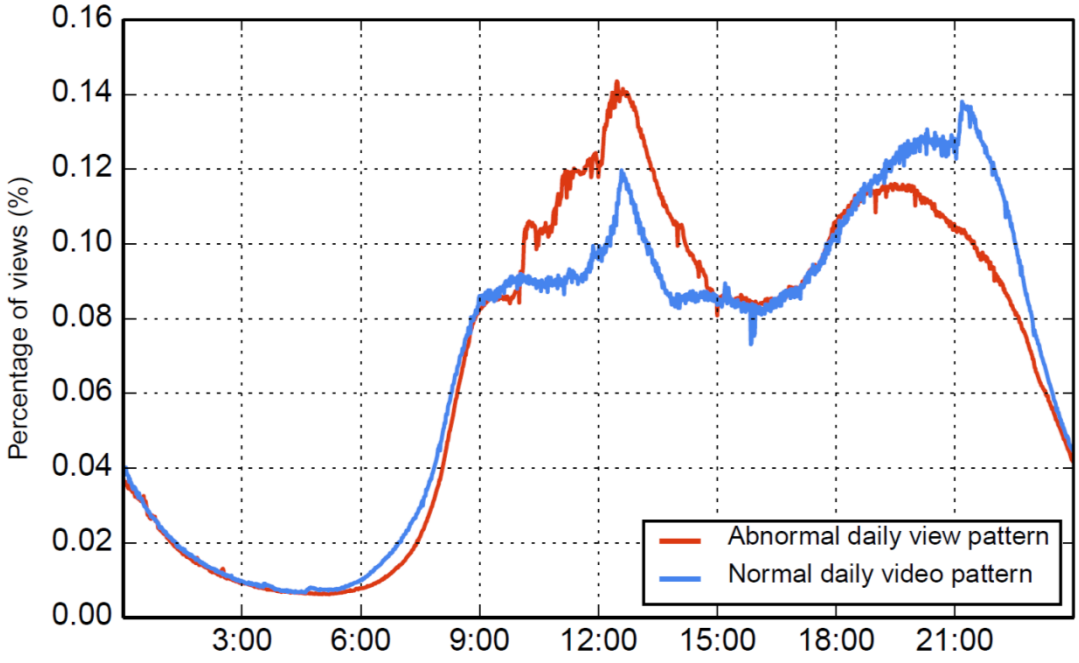
- View count effect
  - Viewer: recommendation reference
  - Content owner: measure video popularity
  - Advertiser: currency
- **Fake view – view count created by non-human**
- YouTube kills billions of video views faked by music industry (Dec. 27, 2012)
  - Universal lost more than 1 billion views.
  - Sony BMG: 850 million → 2.3 million.
  - RCA: 159 million → 120 million.
- We studied how to detect fake views automatically in Tencent Video, one of the larges online video provider in China.

Outline

- Background
  - -Platform of Tencent Video
  - -The Motivation to Make Fake View
  - -The Method to Generate Fake View
- Fake View Detection
  - -User Dimension
  - -IP Dimension
  - -Video Dimension
- Conclusion

Fake Views in Tencent Video

- Abnormal pattern of daily view count in Tencent Video



走!(快乐大本营 12/11/1...
歌手：周笔畅
发布：2012-11-11
播放：640.6万

- **From multiple machines**
- **Target video: a music video**

Why “Fake View”

- Attract eye balls
  - **Attack the ranking based on view count**
  - **Make the target video popular**

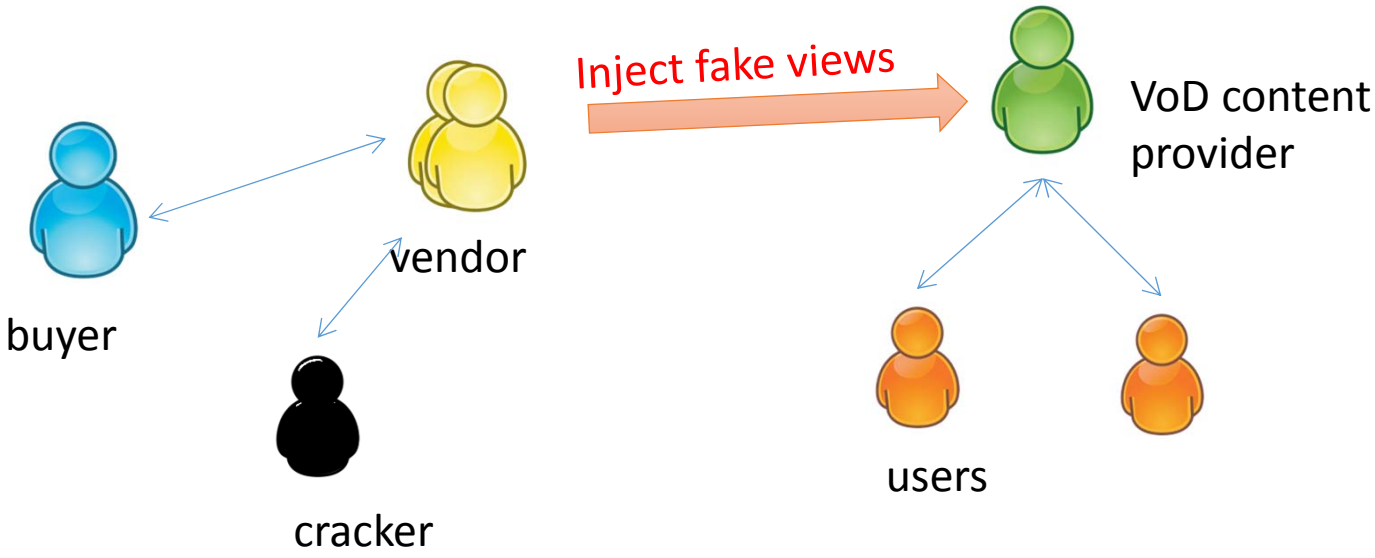
- Make an impact
  - **High view count can be referenced publicly**
  - **Content creator can benefit from a large number of views**

The Impact of “Fake View”

- **Network resource allocation**
  - CDN resource
  - Schedule workload
- **Recommendation system**
  - User experience on recommended videos
- **Business intelligence**
  - The most popular videos in reality
  - Advertising
  - Product analysis
- **End users**
  - Be tricked, waste time, lost trust in recommendation

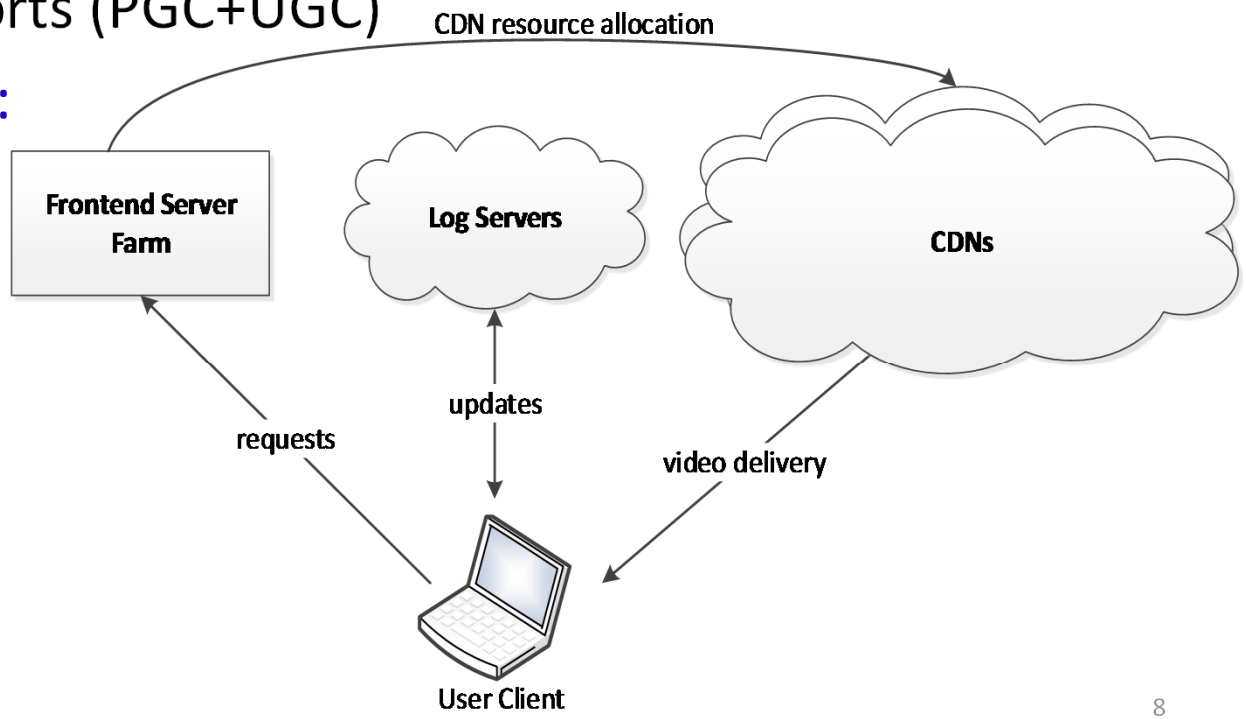
Who is Making the “Fake View”

- Fake view ecosystem:
  - Online video service provider
  - Video viewers
  - Fake view vendor
  - Fake view buyer
  - Cracker
- Value chain



Tencent Video Platform

- 50 million active daily users
- More than 2 million users online during busy hours
- Movie, TV episodes, music/entertainment video, short clips of news and sports (PGC+UGC)
- Viewing reports:



How to Create Fake Views?

- On the market
  - Google “buying view”
- General approaches
  - With tools:
    - Artificial views: Open multiple web browser tabs successively to view video
    - Forged reports: Send forged viewing reports (by cracking the ICP’s protocol)
  - With distributed network:
    - Like DDoS
    - Schedule the requests sending time and frequency

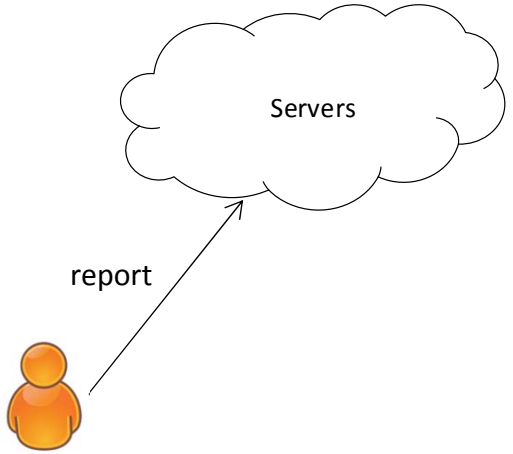
Fake View Attack Methods

	Artificial Views	Forged Reports
Single IP	< 10k/day	~ 10m/day
Multiple IPs	100k ~ 10m/day	> 10m/day

- I. **Artificial view**: open video in multiple browser tabs continuously and periodically
- II. **Forged report**: send lots of viewing reports to server by cracking service provider's protocol
- We focus on the daily offline detection from single or multiple IPs by forged report

Fake View Features

- Feature candidates:
  - # of views by a user?
  - Request frequency?
  - IP address based feature
  - Video based feature
  - Release time



How frequently?

- 1000 /min
- 900 /min
- …

Data Used for Study

- Daily view records are collected by Tencent Video's log servers

IP	TCP	HTTP message [uid, vid, timestamp, ...]

- Users are not required to login in advance, most view records have no user ID.

Our Idea:

1) **User entropy based observation**
2) **IP entropy based detection**
3) **Video entropy based detection**

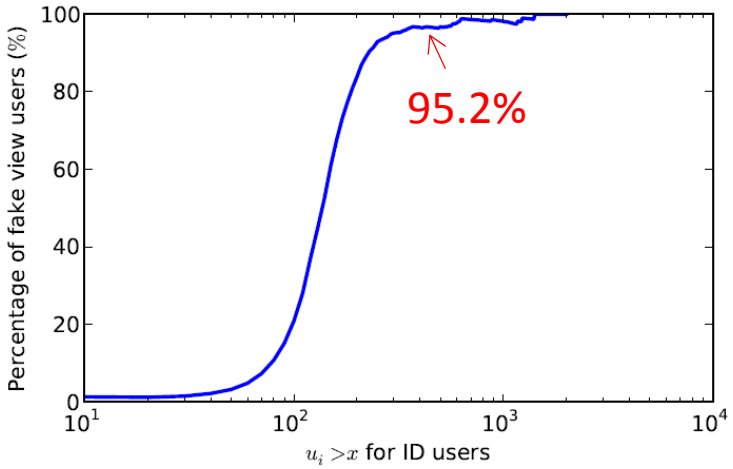
User Dimension Observation

- User's video access matrix

$$A_{m\times n} = \begin{array}{c} \\ \text{user}_1 \\ \text{user}_2 \\ \vdots \\ \text{user}_m \end{array} \begin{array}{c} \begin{array}{cccc} \text{vid}_1 & \text{vid}_2 & \dots & \text{vid}_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Detection: $a_{ij} > K$

- Observation on user dimension



Theoretical Analysis Results I

$$A_{m\times n} = \begin{array}{c} \\ \text{user}_1 \\ \text{user}_2 \\ \vdots \\ \text{user}_m \end{array} \begin{array}{c} \begin{array}{cccc} \text{vid}_1 & \text{vid}_2 & \dots & \text{vid}_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

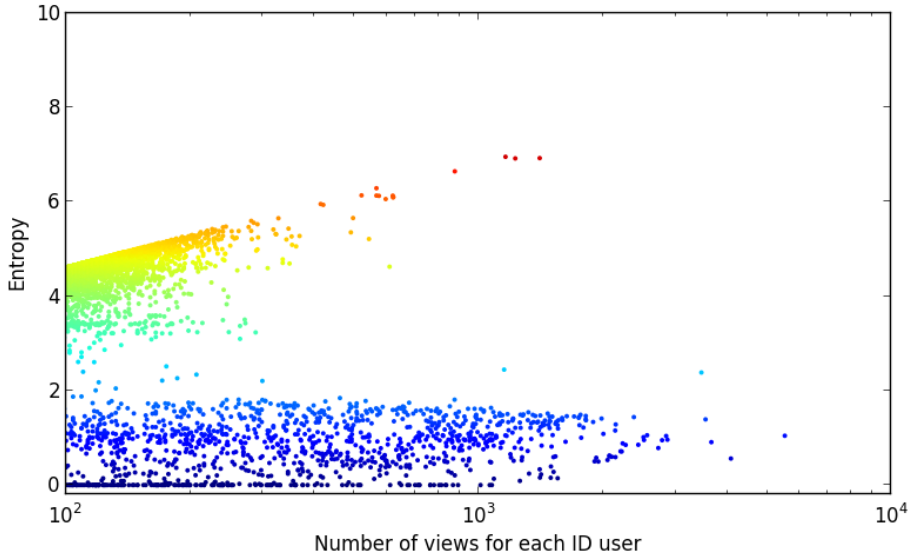
$$H_u(i) = -\sum_{j=1}^{n} \frac{a_{ij}}{u_i} \ln \frac{a_{ij}}{u_i} \cdot I(a_{ij} > 0),$$

$$u_i = \sum_j a_{ij}$$

- Observation: the probability to replay a video is very low for most users.
- Most users' entropy should increase logarithmically with view counts.

Entropy of Viewing Distribution

Entropy for each user identified by user ID



	$H_u = 0$	$H_u < 1$	$H_u < 3$
# of users	166	649	1171
percent of anomaly	100%	100%	98.2%

Theoretical Analysis Results II

- IP entropy increases at most logarithmically with views since multiple users may share a common IP and the replay probability is larger than single user.

$$H_w(i) \leq \ln w_i$$

- Video entropy increases logarithmically if views are from different IP addresses.

$$H_v(j) \leq \ln v_i$$

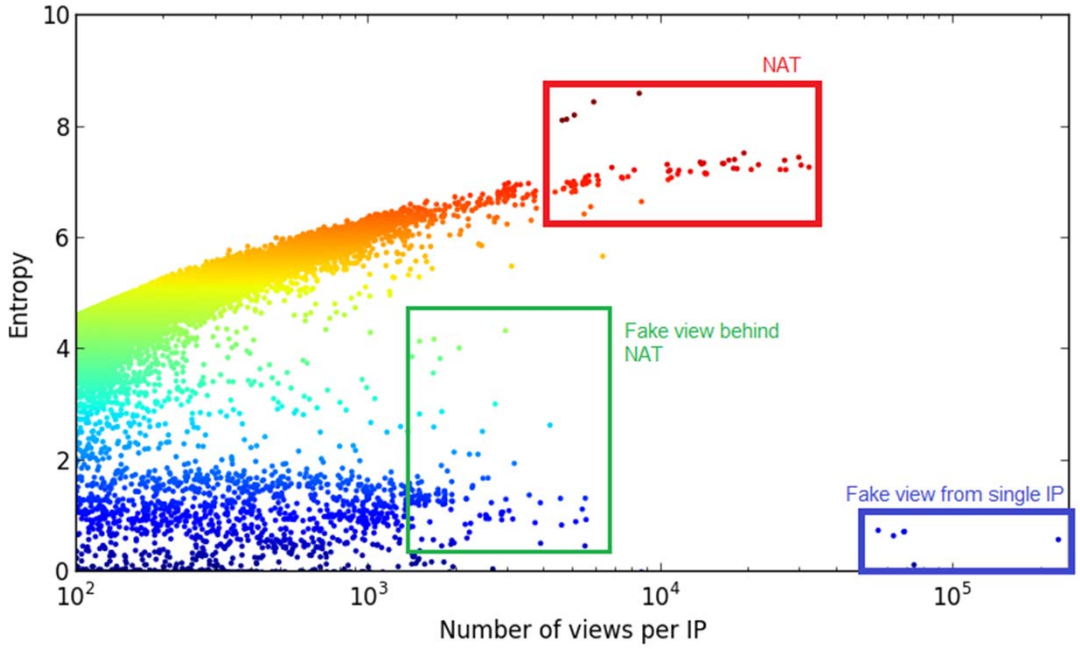
IP Dimension on Video Access

- IP access matrix

$$A_{m\times n} = \begin{array}{c} \\ \mathrm{IP}_1 \\ \mathrm{IP}_2 \\ \vdots \\ \mathrm{IP}_m \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{vid}_1 & \mathrm{vid}_2 & \dots & \mathrm{vid}_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

$$H_w(i) = -\sum_{j=1}^{n} \frac{a_{ij}}{w_i} \ln \frac{a_{ij}}{w_i} \cdot I$$

$$w_i = \sum_{j}^{n} a_{ij}$$



Entropy for Videos

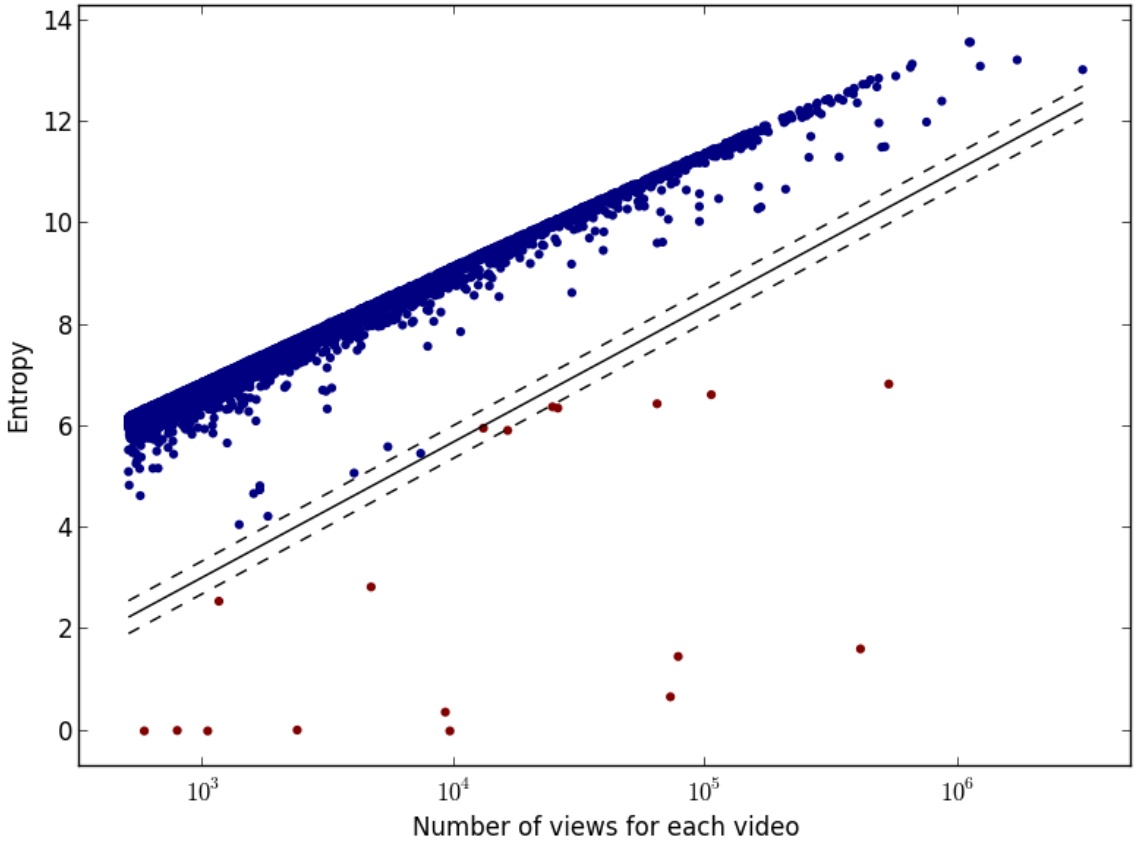
$$H_v(j) = -\sum_{i=1}^{m} \frac{a_{ij}}{v_j} \ln \frac{a_{ij}}{v_j} \cdot I(a_{ij} > 0),$$

$$v_j = \sum_{i}^{m} a_{ij}$$

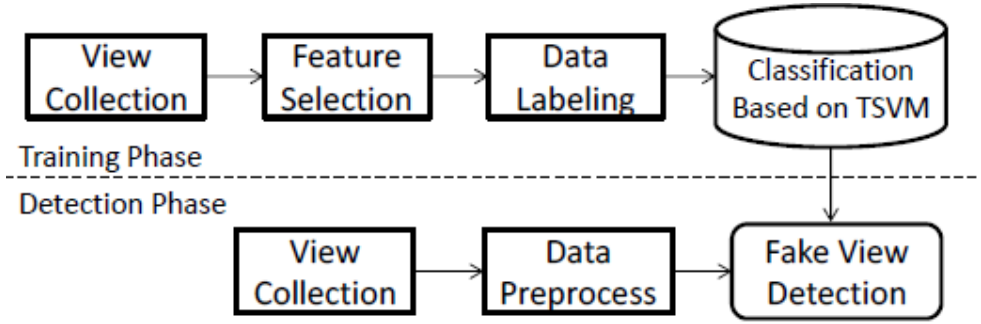
Around 800 million views per day

Manually checking 10 thousand video at most

Machine learning approach: TSVM

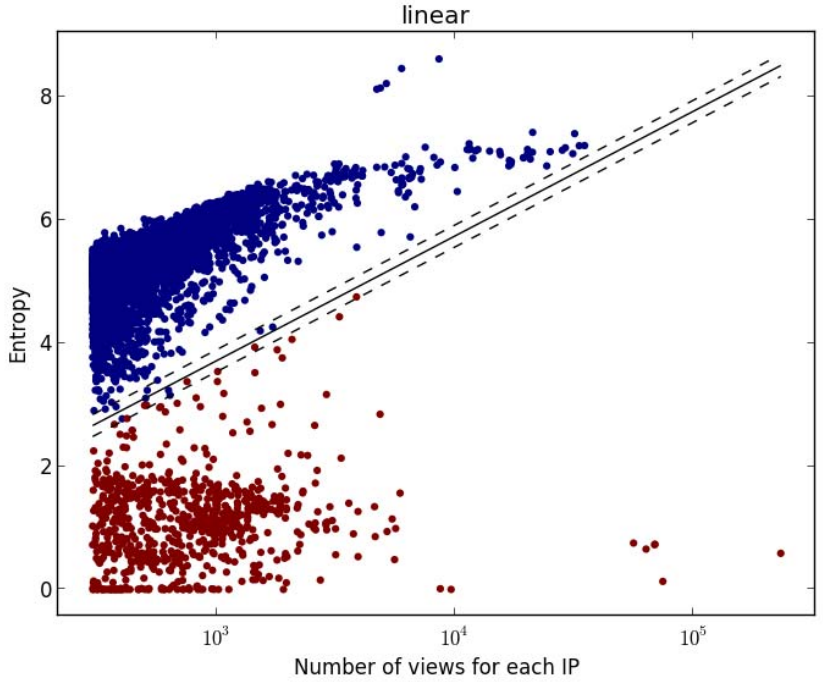


Detecting Fake View IPs

- 

• Classification based on multiple features.

Accuracy is about 99%



Observation

- Most fake view videos are UGC and MV (account for more than 90% fake view videos), but also some popular TV series (usually the first episode).

- For UGC, many video creators have incentive to promote their videos.
- For MV, it may involve public relation companies and fans to introduce the fake views.

TABLE III
CASE STUDY FOR FAKE VIEW VIDEOS.

	Type	# of Views	# of IP	Entropy
video1	UGC	10552	1	0
video2	MV	409409	162	1.62
video3	TV	1388461	219584	5.02

Video1:10552 views are from single IP.
Video2:99.95% views are from 6 IPs.
Video3:63.5% views are from 10 IPs.

Conclusion

- Introduce the fake view problem in online video services
- **Offline Detection** of fake views caused by forged reports
- Based on the **IP entropy** and **video entropy**
- Challenges:
  - Distributed network attack (DDoS)
  - Online algorithms for real-time detection

Thank you!

- *Liang Chen* is looking for post-doc positions right now
- leoncuhk@gmail.com

- Q&A