

Rationalizing YouTube Commenting behavior

William Hoiles, Vikram Krishnamurthy, Kunal Pattanayak

YouTube > 1 billion users. 300 hours of content created every minute.

- **How to model, predict user engagement?**
How do thumbnail, title, description of video affect popularity?
How does interaction of YouTube channels with users affect popularity? User → Video Content → User
- **Why?** Content Caching in 5G. Recommender systems, Strategic Design for content producers (e.g. BBTv).
- **Main Result:** Based on massive YouTube dataset, Bayesian Revealed Preferences with Deep Embedded Clustering yields accurate model that relates viewcount to commenting behavior.

YouTube Commenting as Bayesian Utility Maximization

State: $X \in \{\text{low viewcount, high viewcount}\}$,
 $X \sim \pi_0$ (prior)

Observation: $Y \sim \alpha(y|x)$. Visual cues from
thumbnail, video description

Attention function of viewer: α

Action: a . Comment count (high or low),
sentiment (positive, negative or neutral)

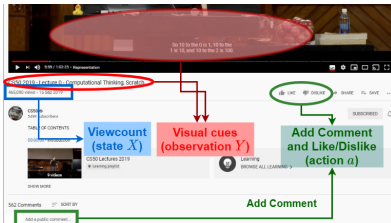
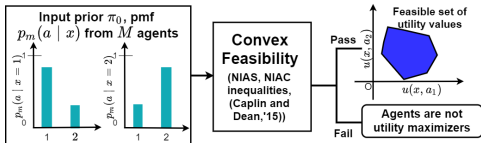
Commenter's reputation: $u(x, a)$

Rational Inattention Cost: $C(\alpha, \pi_0)$

Optimize commenting behavior:

$$(\alpha^*, a^*) \in \arg \max_{\alpha, a} \mathbb{E}_{\pi_0, \alpha} \{u(X, a(Y))\} - C(\pi_0, \alpha)$$

Analyst aim: Given $\pi_0, \{p_m(a|x), m \in \mathcal{M}\}$
from M agents, detect utility maximization.



Massive dataset: 140k videos, 25k channels.

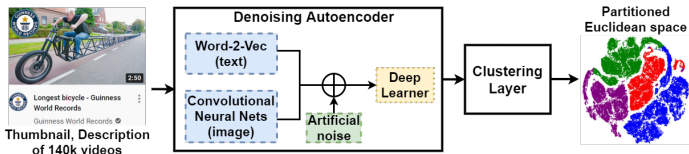
Dimension Reduction. How to group
videos of specific topic with similar
commenting behavior?

(i) *User-centric:* Deep Clustering using
thumbnail & description.

(ii) *Content-centric:* Video category.

Main Result: YouTube commenting is
consistent with utility maximization.
Estimated utility can predict commenting
behavior. (83% accuracy).

Deep Embedded Clustering and Dataset Analysis



Autoencoder partitions YouTube dataset into 8 distinct clusters (agents).

How well does Bayesian Utility Maximization explain dataset?

General Rational Inattention cost: All 8 clusters satisfy test.

Renyi/Shannon mutual information cost: 2/8 clusters satisfy test.

Finer Granularity. 18 categories using topic (Gaming, Politics, Education, etc.)

Result: 10 categories satisfy general cost, 2 categories satisfy Renyi/Shannon.

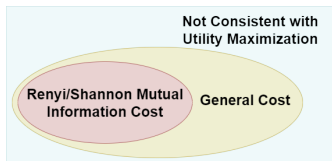
Key Insights:

- Clusters fail Renyi/Shannon by small margin \implies model is robust.
- Utility (reputation) is substantially higher for popular videos.
- *Predictive Accuracy.* Given a video in a specific category, predicts comment count with 83% accuracy; sentiment with 80% accuracy.

Robustness of utility maximization test

Quantifying robustness:

- For categories that satisfy utility maximization, how far are they from failing.
- For categories that don't satisfy, how close are they to passing.



1. For categories that fail general cost, find min. perturbation to pass.

Result: Average $\epsilon_1 = 1.2 \times 10^{-3}$.

2. For categories that satisfy general cost, find max. perturbation to fail.

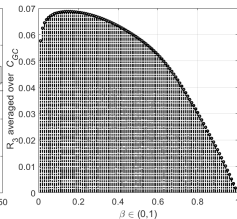
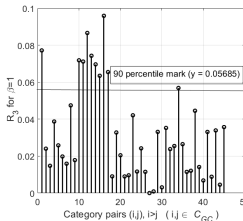
Result: Average $\epsilon_2 = 7.01 \times 10^{-3}$.

Conclusion: $\epsilon_1/\epsilon_2 \approx 6$, hence categories are much closer to satisfying general cost than failing.

3. For categories that satisfy general cost, find min. perturbation to satisfy Renyi or Shannon cost.

Renyi Entropy: $H_\beta(p) = \sum_{i=1}^n \log(p_i^\beta)/(1 - \beta)$.

Shannon cost: Renyi cost with $\beta \rightarrow 1$.



Extensions and Limitations

- *Sequential Information sampling*: How to incorporate the visual cues of the viewer while watching the video into the attention function?
- *Temporal dependence of commenting behavior*: How to incorporate the effect of existing comments on a video on future comments and commenter reputation? Herding and information cascades.
- *Analyzing effect of changing meta-features*: How to detect a change in commenter reputation if the thumbnail/description is modified after the video is posted?
- *Feedback Control*: How to adapt video title & keywords to maximize view count given knowledge of commenter reputation?
- *Richer models of commenting reputation*: The five dominant meta-level features that affect the popularity of a video are: first day view count, number of subscribers, contrast of the video thumbnail, Google hits, and number of keywords. How do all these features affect commenting behavior?