Workload Characterization – Practical Examples

1. Characterizing Scholar Popularity: A Case Study in the Computer Science Research Community, G. Gonçalves et al. Proc. ACM/IEEE Digital Libraries 2014.

2. On the Dynamics of Social Media Popularity: A YouTube Case Study, F. Figueiredo et al.,ACM TOIT 2015

3. Understanding Video-Ad Consumption on Youtube: A Measurement Study on User Behavior, Popularity and Content Properties, M.Arantes et al., Proc. ACM Web Science 2016

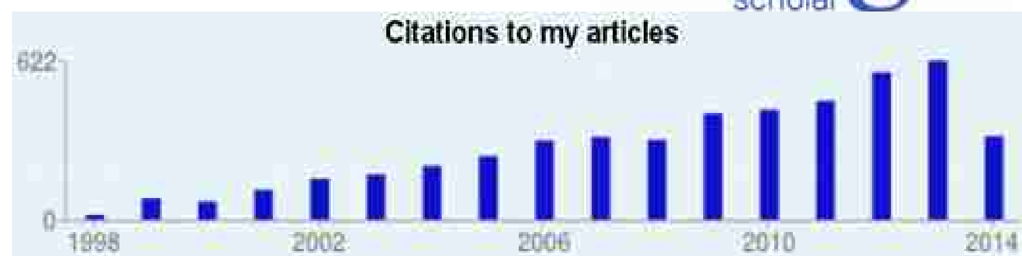
4. Tips, Dones and To-Dos: Uncovering User Profiles in Foursquare, M. Vasconcelos et al., Proc. WSDM 2012

Characterizing Scholar Popularity: A Case Study in the Computer Science Research Community

Scholar Popularity

- Popularity is one measure of scholarly success
  - Not the only one; but very sought-after
  - Scholar popularity = number of citations [DiCr2011]
  - Which factors impact a scholar's popularity?
  - Some old issues persist:
    - Quantity vs. quality of publication venues
    - Role of co-authorship network

Popularity Temporal Dynamics



- Are there common patterns of popularity evolution?
  - That is: are there common scholar popularity profiles?
  - How do they correlate with various academic features?
- How to extract such profiles?

Our Goals

1. Investigate and quantify the importance of factors that impact scholar popularity

2. Study scholar popularity dynamics:
   a. identify common profiles
   b. characterize their academic features

- Draw insights into the design of popularity prediction models

Methodology

Data Collection

- ArnetMiner:
  - 2,244,018 publications
  - 831,763 authors
  - 8,274 venues
  - 38,770,182 citations

- Microsoft Academic Search:
  - 624,784 author time series

    number of publications

    number of citations

Data Collection

- ArnetMiner:
  - 2,244,018 publications
  - 831,763 authors
  - 8,274 venues
  - 38,770,182 citations

characterization of features
that impact popularity

- Microsoft Academic Search:
  - 624,784 author time series

    number of publications

    number of citations

study
popularity dynamics

Both datasets: COMPUTER SCIENCE

Scholar's Experience Level

- Split each dataset into 5 experience groups
  - Experience level = # years since scholar's 1st publication
  - Reduce natural heterogeneity across scholars

Experience Level

	[0;5]	(5;10]	(10;15]	(15;20]	$(20,\infty)$
# authors	53,642	148,567	91,932	50,435	58,144
# popularity time series	76,980	133,501	99,234	60,521	67,210

Academic Features

Characterize each scholar by 10 academic features

Notation	Description
$nPubs$	total number of publications
$yPubRate$	yearly publication rate
$nVenues$	number of distinct venues
$CitVen_{max}$	maximum number of citations of any venue
$CitVen_{avg}$	average number of citations per venue
$CitPubVen_{max}$	maximum number of citations per publication of any venue
$CitPubVen_{avg}$	average number of citations per publication per venue
$nCoauthors$	number of co-authors
$closeness$	closeness in the co-authorship network
$PageRank$	PageRank in the co-authorship network

Academic Features

Characterize each scholar by 10 academic features

Notation	Description
$nPubs$	total number of publications
$yPubRate$	yearly publication rate
$nVenues$	number of distinct venues
$CitVen_{max}$	maximum number of citations of any venue
$CitVen_{avg}$	average number of citations per venue
$CitPubVen_{max}$	maximum number of citations per publication of any venue
$CitPubVen_{avg}$	average number of citations per publication per venue
$nCoauthors$	number of co-authors
$closeness$	closeness in the co-authorship network
$PageRank$	PageRank in the co-authorship network

1

Group 1: scholar productivity

Academic Features

Characterize each scholar by 10 academic features

	Notation	Description
1	$nPubs$	total number of publications
	$yPubRate$	yearly publication rate
	$nVenues$	number of distinct venues
2	$CitVen_{max}$	maximum number of citations of any venue
	$CitVen_{avg}$	average number of citations per venue
	$CitPubVen_{max}$	maximum number of citations per publication of any venue
	$CitPubVen_{avg}$	average number of citations per publication per venue
	$nCoauthors$	number of co-authors
	$closeness$	closeness in the co-authorship network
	$PageRank$	PageRank in the co-authorship network

Group 1: scholar productivity

Group 2: venue quality

Academic Features

Characterize each scholar by 10 academic features

	Notation	Description
1	$nPubs$	total number of publications
	$yPubRate$	yearly publication rate
	$nVenues$	number of distinct venues
2	$CitVen_{max}$	maximum number of citations of any venue
	$CitVen_{avg}$	average number of citations per venue
	$CitPubVen_{max}$	maximum number of citations per publication of any venue
	$CitPubVen_{avg}$	average number of citations per publication per venue
3	$nCoauthors$	number of co-authors
	$closeness$	closeness in the co-authorship network
	$PageRank$	PageRank in the co-authorship network

Group 1: scholar productivity
Group 2: venue quality
Group 3: role in co-authorship network

Characterization Results:

Impact of Academic Features on Scholar Popularity

Impact of Academic Features

- Quantify the importance of each academic feature to scholar popularity
  - Correlation analysis
  - Regression analysis

- Data source:
- ArnetMiner dataset

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

Correlations tend to strengthen with scholar experience for most features

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	$(20,\infty)$
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

Correlations tend to strengthen with scholar experience for most features

exception: venue quality features

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

Less experienced scholars: quality of venues is more important than role in coauthorship network

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

More experienced scholars: coauthorship network is more important

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

Number of publications: most important feature for all experience groups

Correlation Between Academic Feature and Popularity

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

Number of publications: most important feature for all experience groups

Other features are also strongly correlated with popularity

Is there a subset of the features that explain most of the popularity variations?

Which of the considered features are redundant?

- Key questions for designing prediction models

Regression Analysis

Build a model to **describe** scholar popularity

(1) Build model with all academic features (*k* = 10)

$$log(\mathcal{R}) = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + \cdots \beta_k log(x_k)$$

(2) Quantify importance of each feature

(3) Disregard redundant and unnecessary features

Quantify Feature Importance

Regression Model	Model Quality (R^2)				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
All Features	0.450	0.621	0.696	0.737	0.785
$nPubs$ (-)	0.337	0.566	0.656	0.699	0.741
$yPubRate$ (-)	0.449	0.619	0.694	0.736	0.785
$nVenues$ (-)	0.449	0.621	0.696	0.736	0.785
$CitVen_{max}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitVen_{avg}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitPubVen_{max}$ (-)	0.445	0.617	0.694	0.735	0.784
$CitPubVen_{avg}$ (-)	0.430	0.591	0.666	0.709	0.763
$nCoauthors$ (-)	0.444	0.617	0.693	0.733	0.781
$closeness$ (-)	0.450	0.621	0.696	0.735	0.784
$PageRank$ (-)	0.450	0.620	0.695	0.734	0.782

Quantify Feature Importance

Regression Model	Model Quality (R^2)				
	[0;5]	(5;10]	(10;15]	(15;20]	$(20,\infty)$
All Features	0.450	0.621	0.696	0.737	0.785
$nPubs$ (-)	0.337	0.566	0.656	0.699	0.741
$yPubRate$ (-)	0.449	0.619	0.694	0.736	0.785
$nVenues$ (-)	0.449	0.621	0.696	0.736	0.785
$CitVen_{max}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitVen_{avg}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitPubVen_{max}$ (-)	0.445	0.617	0.694	0.735	0.784
$CitPubVen_{avg}$ (-)	0.430	0.591	0.666	0.709	0.763
$nCoauthors$ (-)	0.444	0.617	0.693	0.733	0.781
$closeness$ (-)	0.450	0.621	0.696	0.735	0.784
$PageRank$ (-)	0.450	0.620	0.695	0.734	0.782

Importance of feature: reduction in model quality (R^2) when feature is removed

Quantify Feature Importance

Regression Model	Model Quality (R^2)				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
All Features	0.450	0.621	0.696	0.737	0.785
$nPubs$ (-)	0.337	0.566	0.656	0.699	0.741
$yPubRate$ (-)	0.449	0.619	0.694	0.736	0.785
$nVenues$ (-)	0.449	0.621	0.696	0.736	0.785
$CitVen_{max}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitVen_{avg}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitPubVen_{max}$ (-)	0.445	0.617	0.694	0.735	0.784
$CitPubVen_{avg}$ (-)	0.430	0.591	0.666	0.709	0.763
$nCoauthors$ (-)	0.444	0.617	0.693	0.733	0.781
$closeness$ (-)	0.450	0.621	0.696	0.735	0.784
$PageRank$ (-)	0.450	0.620	0.695	0.734	0.782

The model explains well the popularity of scholars ≥ 5 yrs.

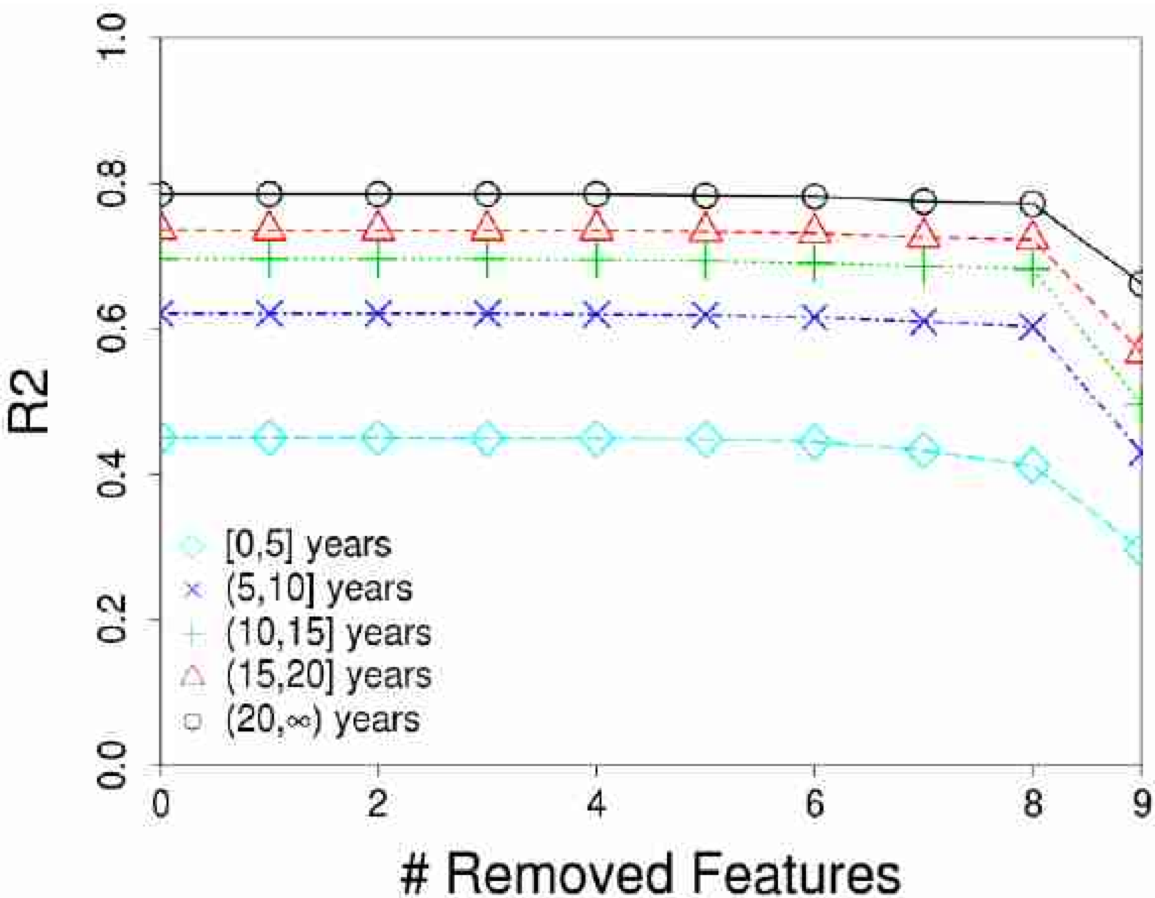
Quantify Feature Importance

Regression Model	Model Quality (R^2)				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
All Features	0.450	0.621	0.696	0.737	0.785
$nPubs$ (-)	0.337	0.566	0.656	0.699	0.741
$yPubRate$ (-)	0.449	0.619	0.694	0.736	0.785
$nVenues$ (-)	0.449	0.621	0.696	0.736	0.785
$CitVen_{max}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitVen_{avg}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitPubVen_{max}$ (-)	0.445	0.617	0.694	0.735	0.784
$CitPubVen_{avg}$ (-)	0.430	0.591	0.666	0.709	0.763
$nCoauthors$ (-)	0.444	0.617	0.693	0.733	0.781
$closeness$ (-)	0.450	0.621	0.696	0.735	0.784
$PageRank$ (-)	0.450	0.620	0.695	0.734	0.782

nPubs or CitPubVenue_{avg}: largest reductions on R^2

Are the other 8 features redundant ?

Disregard Redundant Features



Quality of regression as features are removed in decreasing order of importance

Only two of the features are needed to explain scholar popularity:

nPubs is the most important one

Importance of *nPubs* increases with experience

Characterization Results:

Temporal Dynamics of Scholar Popularity

Popularity Temporal Dynamics

- Identify common profiles of popularity dynamics
- Characterize scholars in each profile
- Data source:
  - Microsoft Academic Search dataset
    (time series of citations per year)
  - Focus on 2 most experienced groups of scholars
    (15,20] , (20,∞) years

Identifying Popularity Profiles

K-Spectral Clustering Algorithm (KSC) [YaLe2011]

- Group times series based on the *shape*
  - Popularity scale and time shift invariants
- K-means with distance metric:

$$dist(s_a, s_b) = \min_{\alpha, q} \frac{||s_a - \alpha s_{b(q)}||}{||s_a||}$$

Vectors representing popularity time series of scholars A and B

Identifying Popularity Profiles

K-Spectral Clustering Algorithm (KSC) [YaLe2011]

- Group times series based on the *shape*
  - Popularity scale and time shift invariants
- K-means with distance metric:

$$dist(s_a, s_b) = \min_{\alpha, q} \frac{||s_a - \alpha s_{b(q)}||}{||s_a||}$$

Shift s_b by q units

Identifying Popularity Profiles

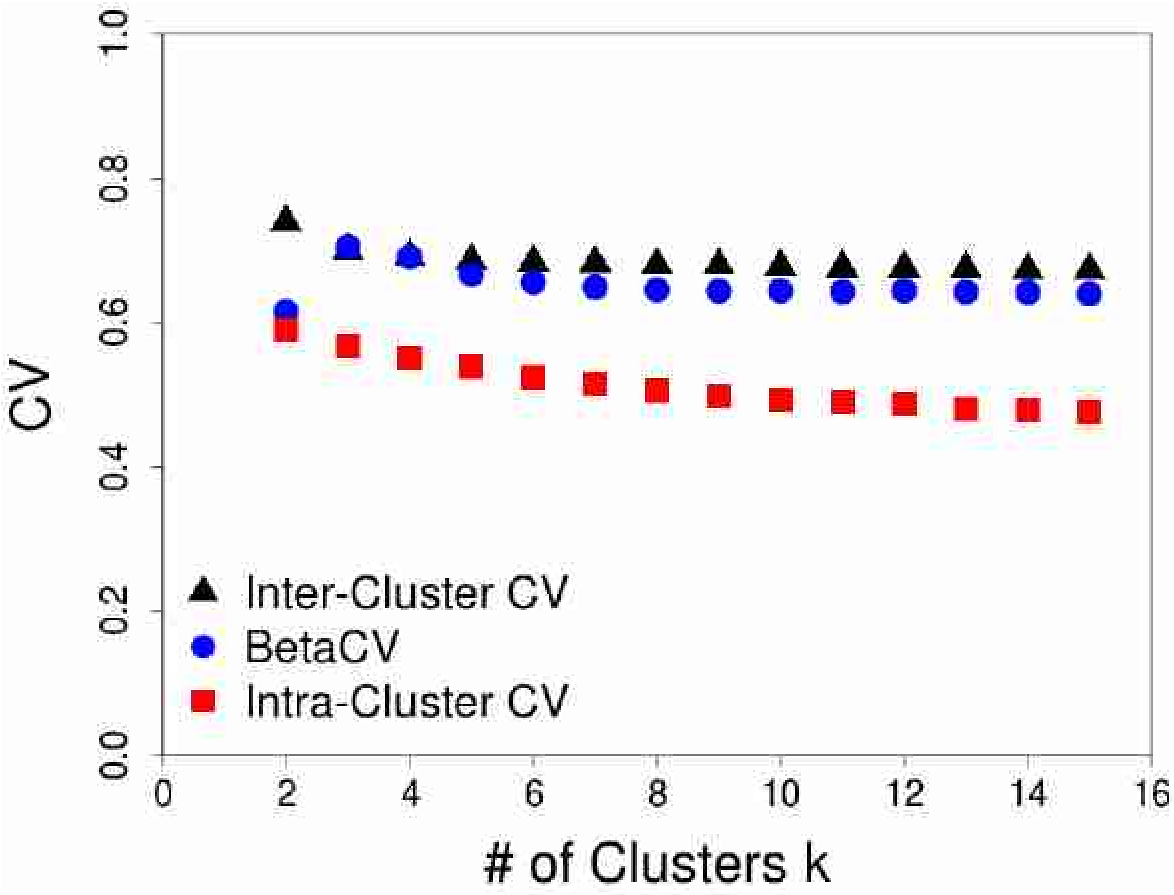
K-Spectral Clustering Algorithm (KSC) [YaLe2011]

- Group times series based on the *shape*
  - Popularity scale and time shift invariants
- K-means with distance metric:

$$dist(s_a, s_b) = \min_{\alpha, q} \frac{||s_a - \alpha s_{b(q)}||}{||s_a||}$$

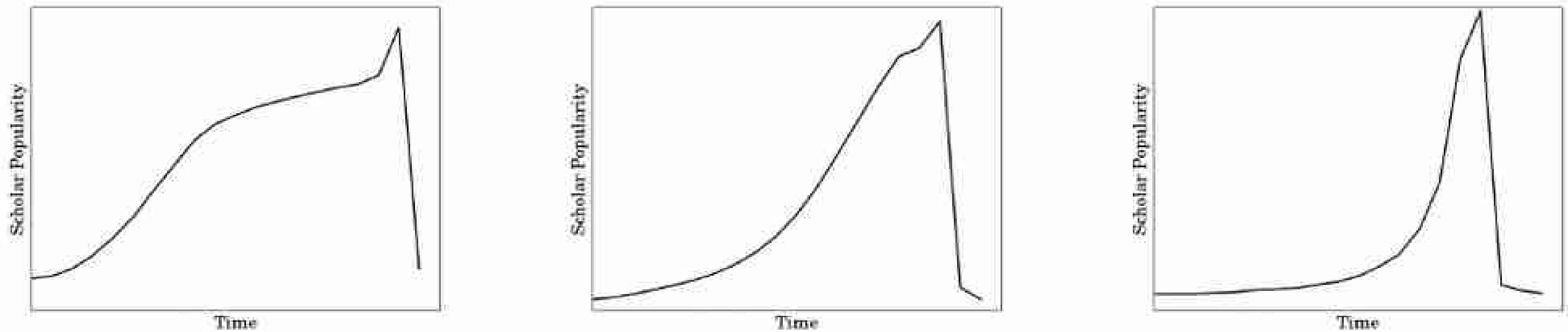
Scale s_b by α units

Determining Number of Profiles



- # of clusters based on β_{CV} metric
- β_{CV} stabilizes for $k \geq 5$: more clusters should be of little help

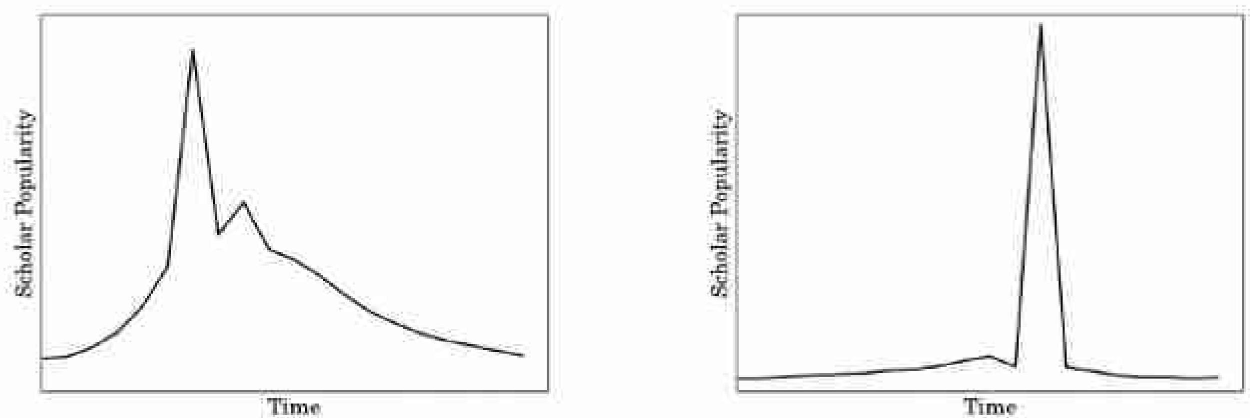
Profiles of Popularity Dynamics



(a) C0 (23% of the scholars)

(b) C1 (27% of the scholars)

(c) C2 (16% of the scholars)

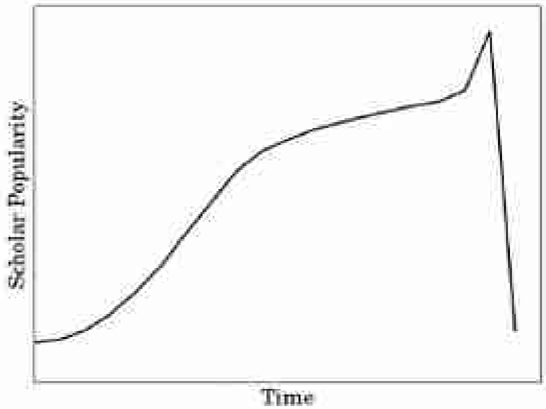


(d) C3 (19% of the scholars)

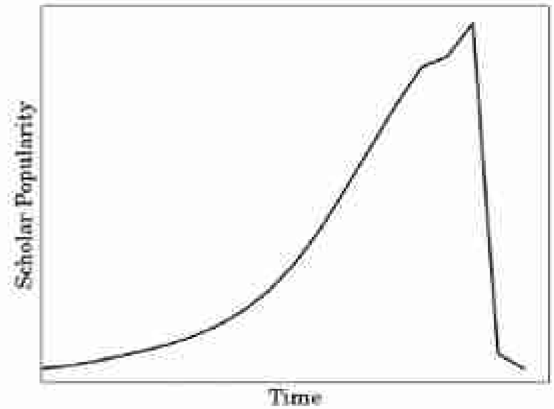
(e) C4 (15% of the scholars)

"Average" popularity curve for scholars in the cluster (centroids)

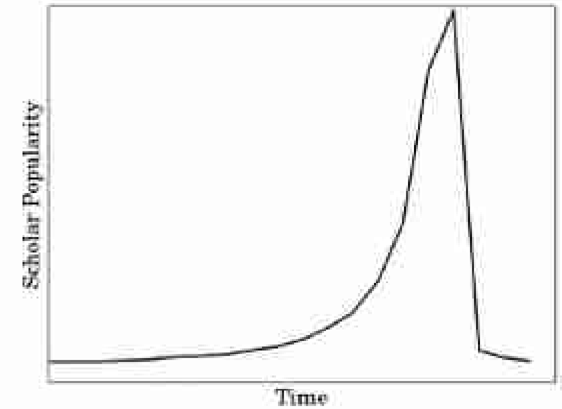
Profiles of Popularity Dynamics



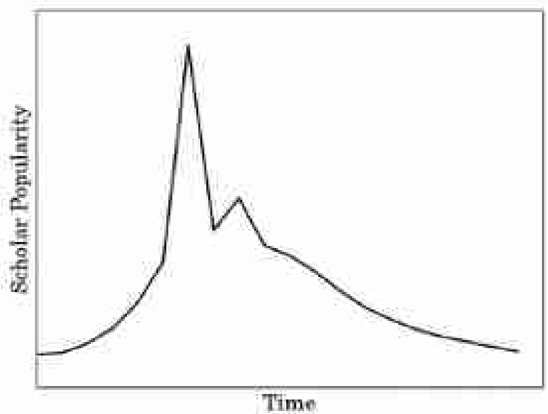
(a) C0 (23% of the scholars)



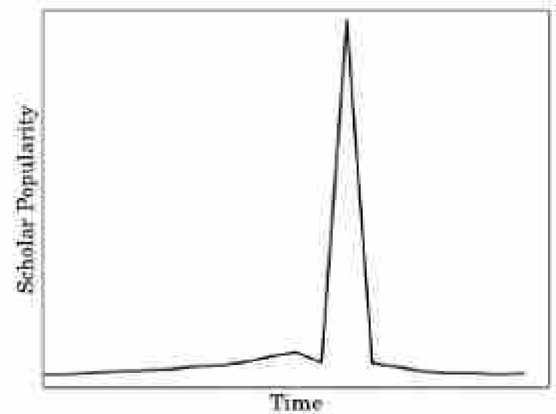
(b) C1 (27% of the scholars)



(c) C2 (16% of the scholars)



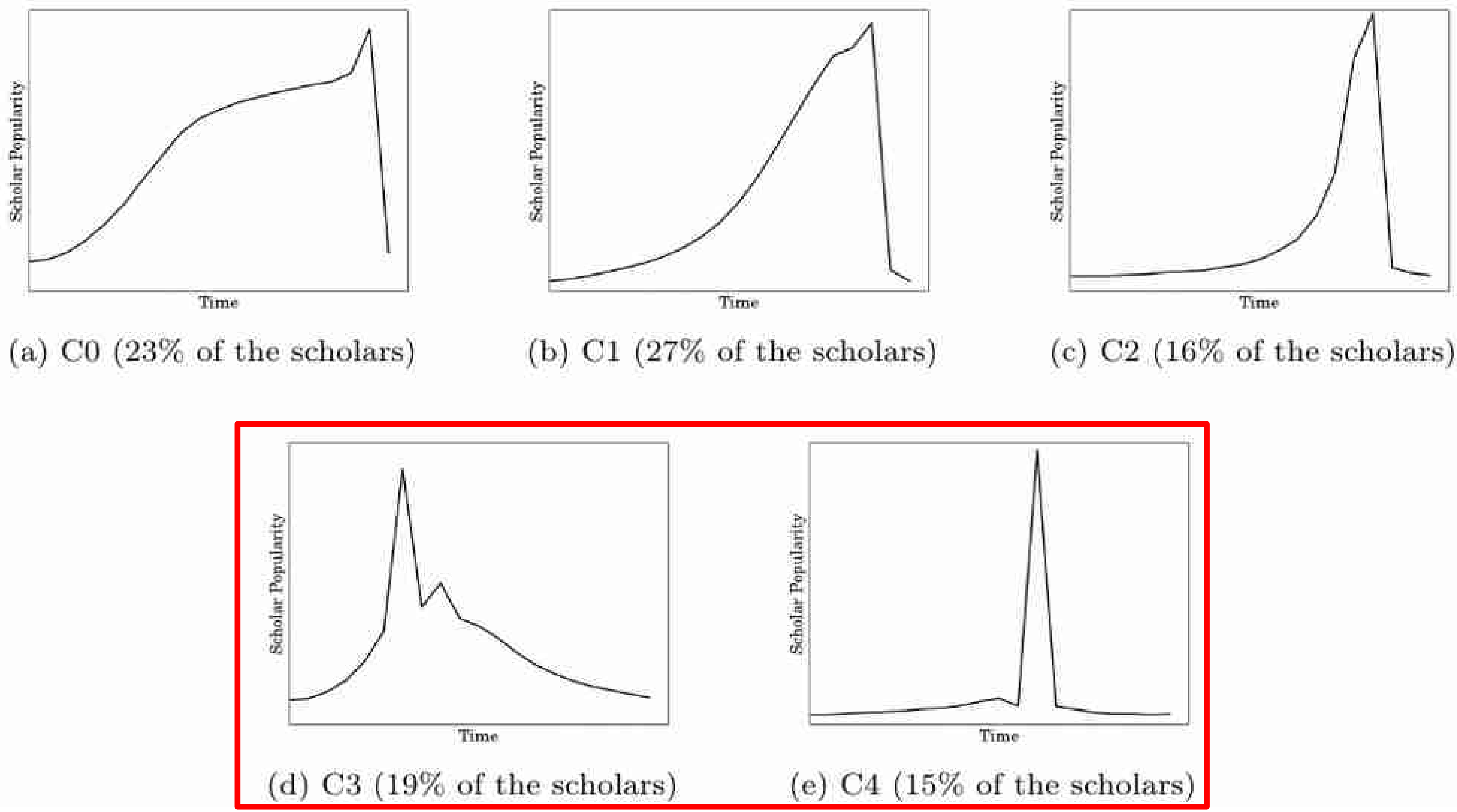
(d) C3 (19% of the scholars)



(e) C4 (15% of the scholars)

C0, C1, C2: scholars succeed in becoming increasingly popular

Profiles of Popularity Dynamics



(a) C0 (23% of the scholars)

(b) C1 (27% of the scholars)

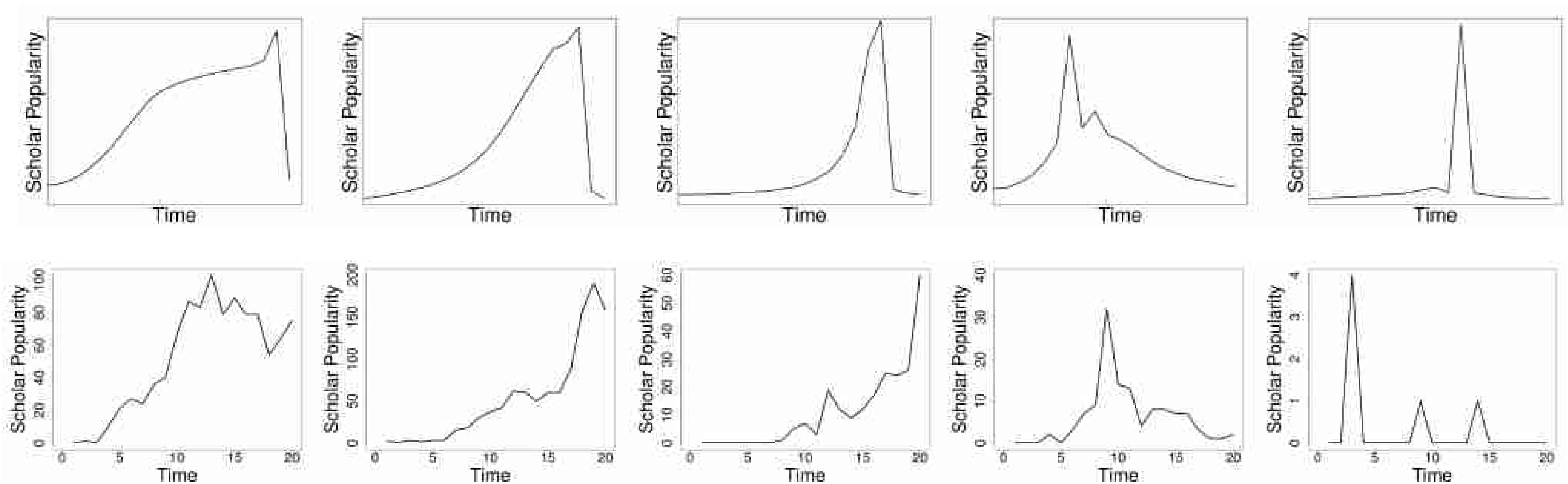
(c) C2 (16% of the scholars)

(d) C3 (19% of the scholars)

(e) C4 (15% of the scholars)

C3, C4: scholars have a clear popularity peak but fail to remain popular

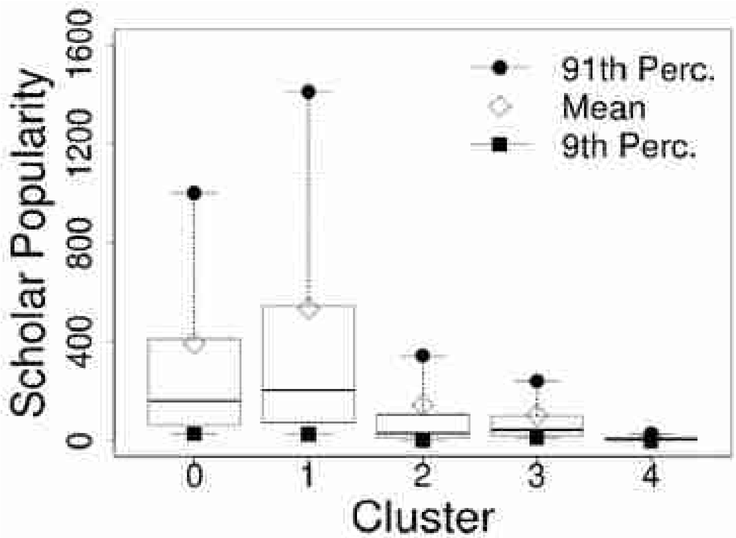
Examples of Scholars in Each Profile



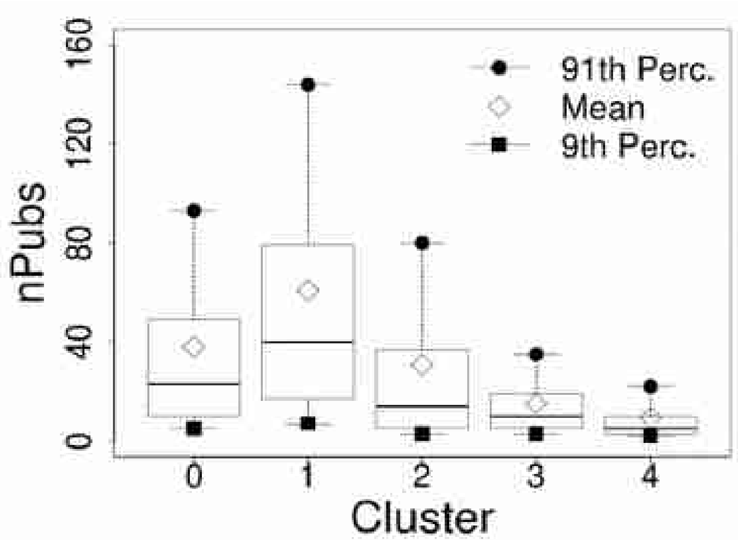
(a) Example in C0 (b) Example in C1 (c) Example in C2 (d) Example in C3 (e) Example in C4

The centroids approximate well the general trends of the individual examples

Characterization of Profiles

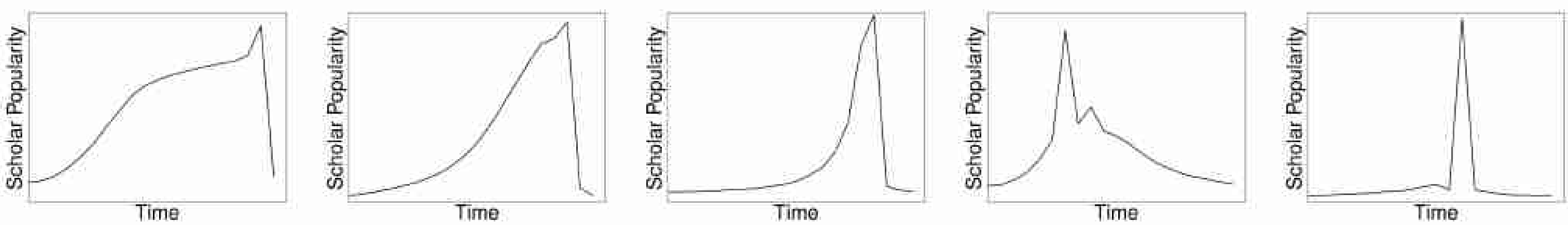


(a) Popularity

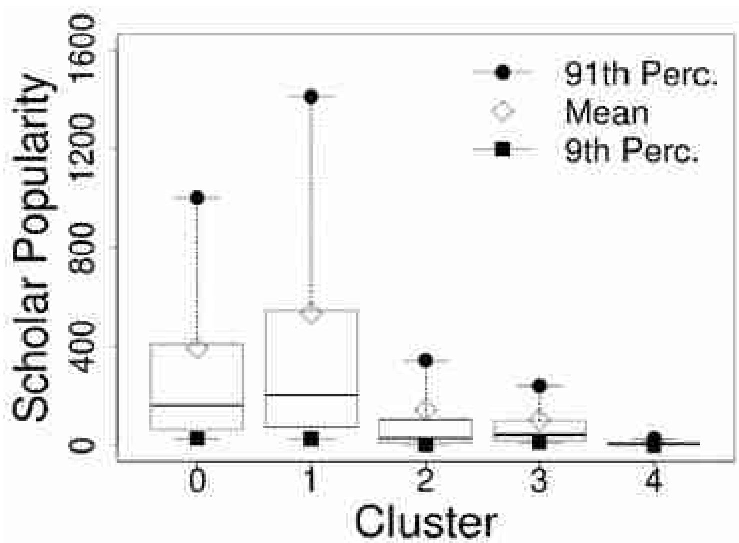


(b) Number of publications

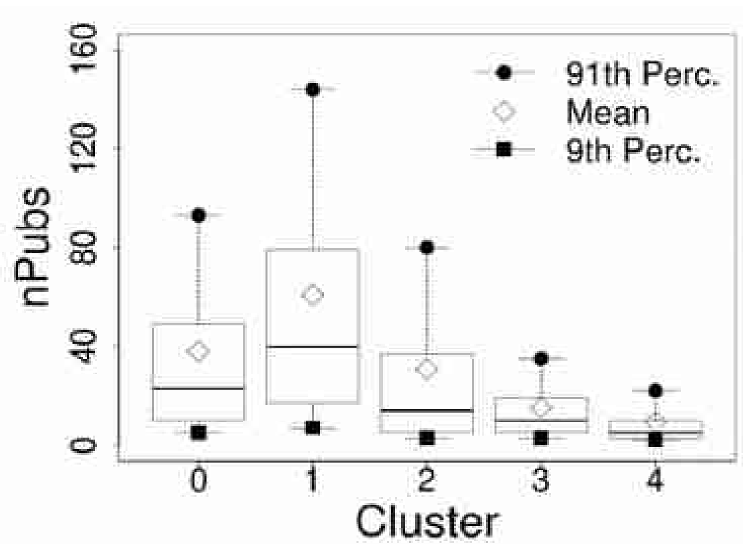
Characterization of Profiles



(a) C0 (23%) (b) C1 (27%) (c) C2 (16%) (d) C3 (19%) (e) C4 (15%)

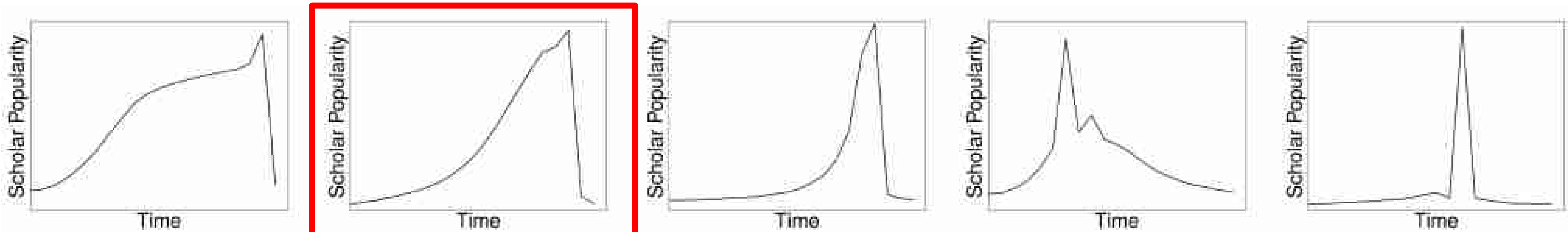


(a) Popularity

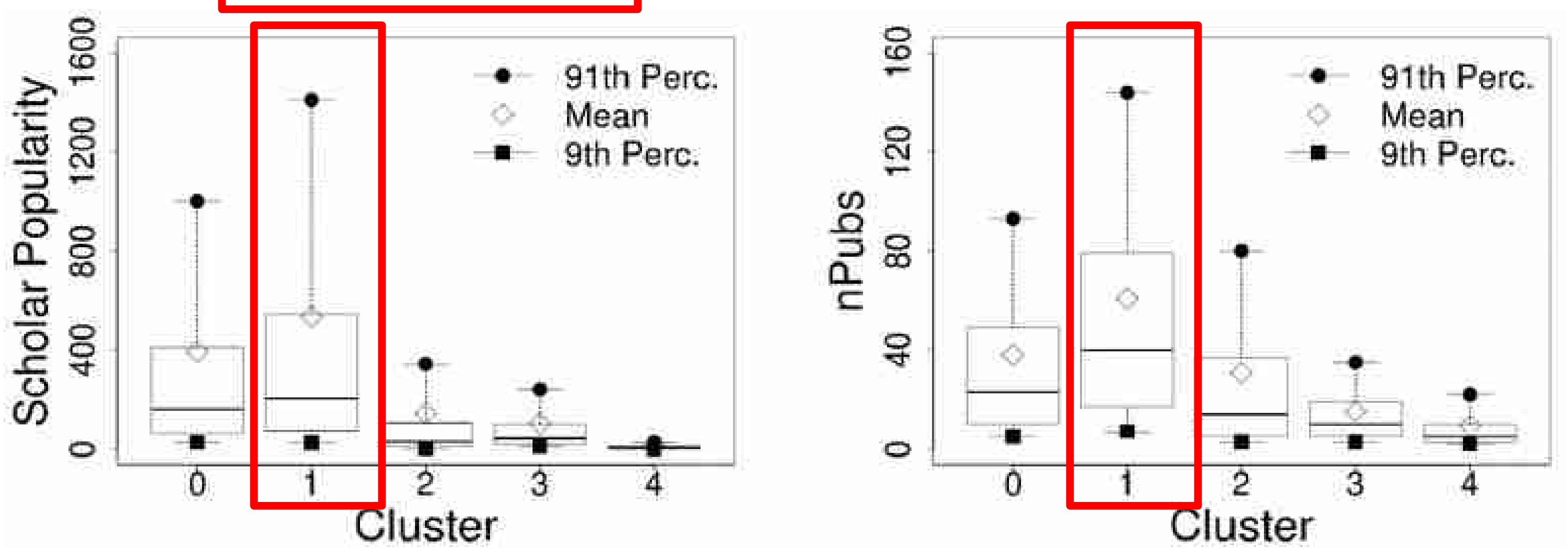


(b) Number of publications

Characterization of Profiles



(a) C0 (23%) (b) C1 (27%) (c) C2 (16%) (d) C3 (19%) (e) C4 (15%)



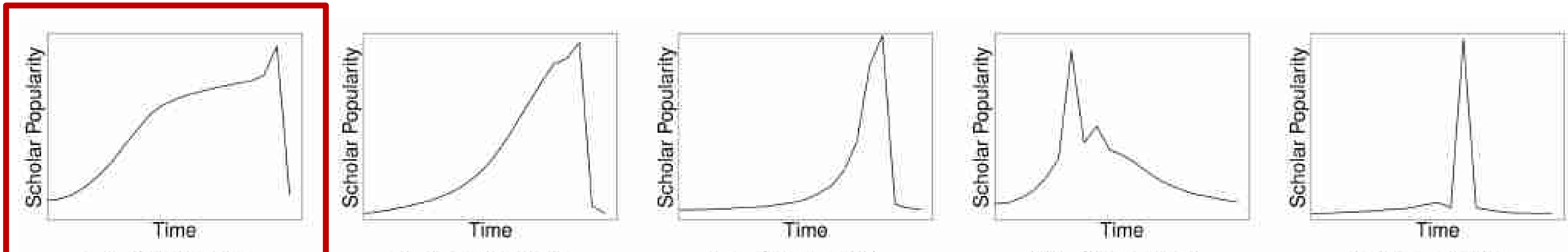
(a) Popularity (b) Number of publications

C1: most popular scholars, with largest # of publications

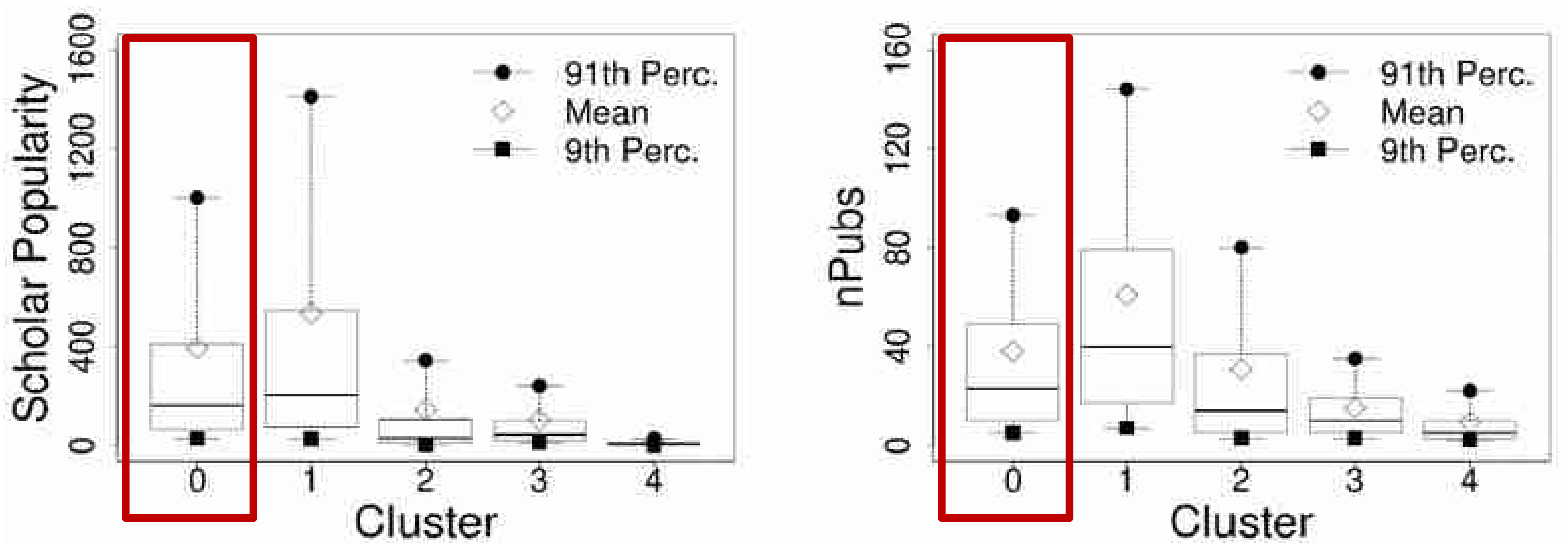
Popularity: 535 (mean) and 1410 (91th percentile)

Number of publications: 61(mean) and 144 (91th percentile)

Characterization of Profiles



(a) C0 (23%) (b) C1 (27%) (c) C2 (16%) (d) C3 (19%) (e) C4 (15%)

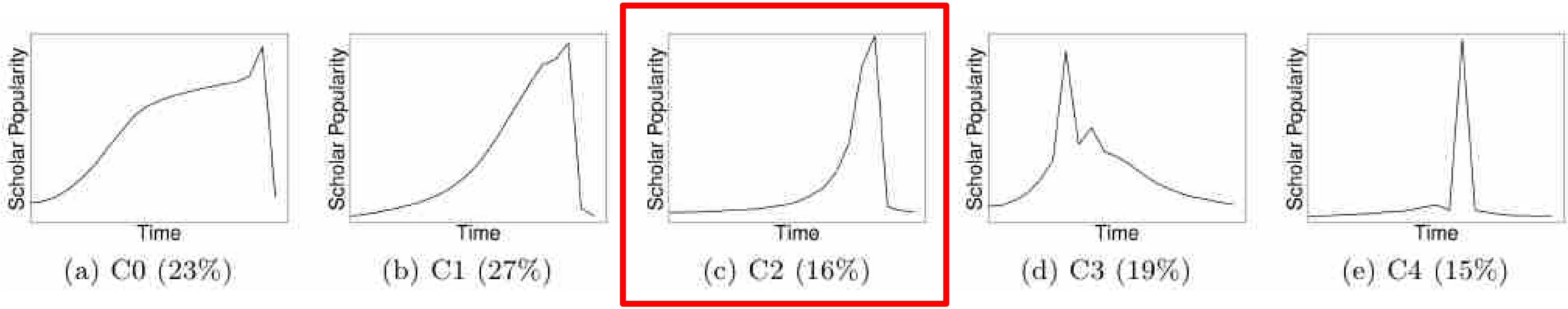


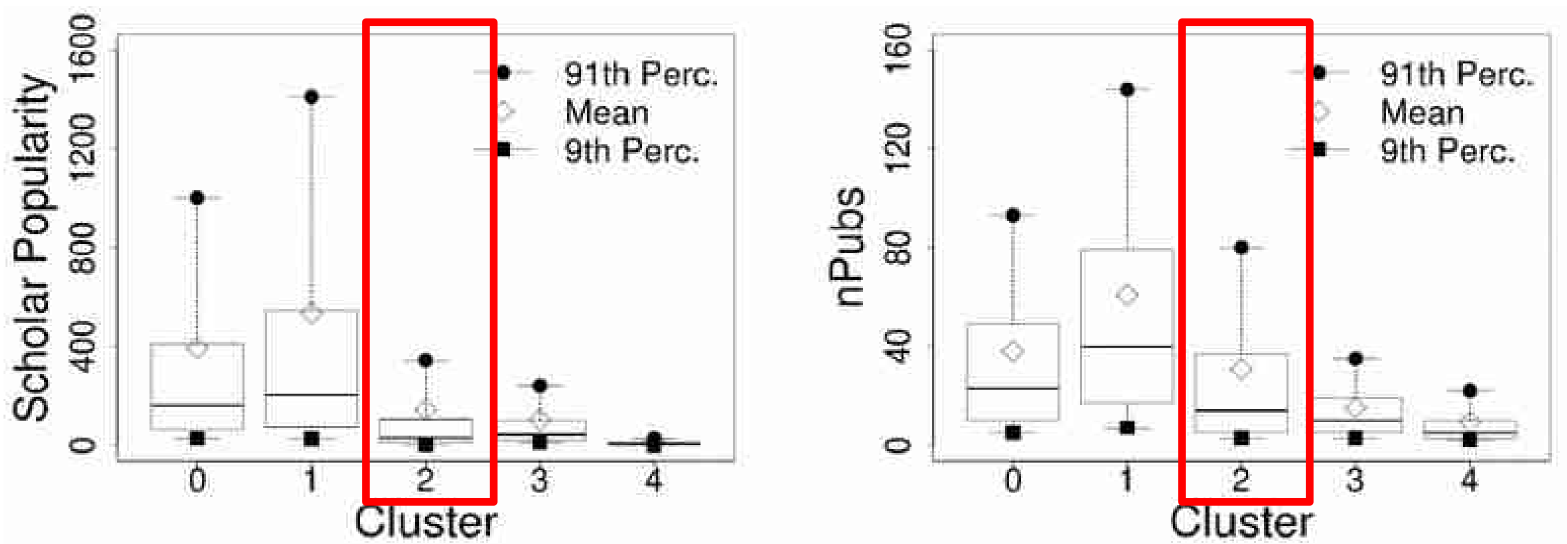
(a) Popularity (b) Number of publications

C0: also very popular/productive scholars (but less than C1)

(91th percentile = 1001 citations)

Characterization of Profiles

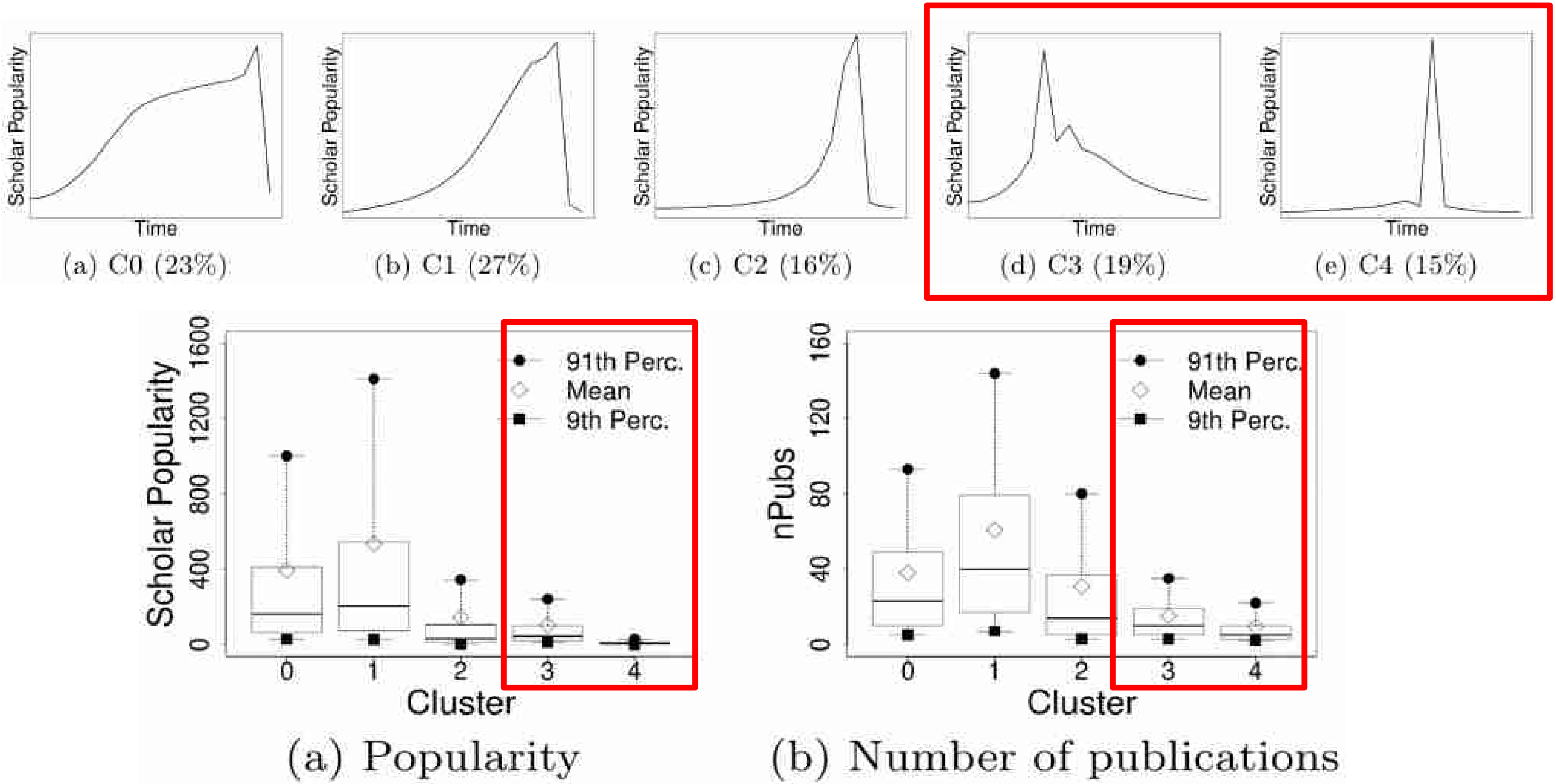




(a) Popularity (b) Number of publications

C2: distributions skewed towards fewer citations/publications

Characterization of Profiles



C3/C4: scholars do not publish often and are not popular

Average popularity, # publications : 101 and 15 (C3); 13 and 9 (C4)

Main Insights

- Venue quality features are less correlated with popularity for more experienced scholars
- *nPubs* and $CitPubVenue_{avg}$ explain most variations in popularity
  - Relative importance of $CitPubVenue_{avg}$ decreases for the most experienced groups
- 5 profiles of scholar popularity dynamics
- "Publish or perish" culture

Conclusions and Future Work

- Conclusions
  - Importance of academic features to scholar popularity
  - Study of scholar popularity dynamics (profiles)
  - Quantitative study based on two large datasets
- Future Work
  - Investigate popularity dynamics in other fields
  - Develop scholar popularity prediction methods
  - Expand investigation for other indices and metrics of scholarly success

On the Dynamics of Social Media Popularity: A YouTube Case Study

Is it possible to **understand** and to provide **good enough predictions** on how the popularity of UGC evolves over time?

Research Goals

RG1: Feature and Content Importance

What "causes" popularity growth?

RG2: Prediction of Popularity Growth

Is it possible to predict how the popularity of individual videos evolves over time?

RG3: Applications for Popularity Prediction

What can we do with this knowledge?

How?

RG1: Feature and Content Importance

- Characterization
- User Study

RG2: Prediction of Popularity Growth

- Learning Methods
- Time Series Data Mining

RG3: Applications for Popularity Prediction

- Revenue Models

RG1: Feature and Content Importance

Datasets

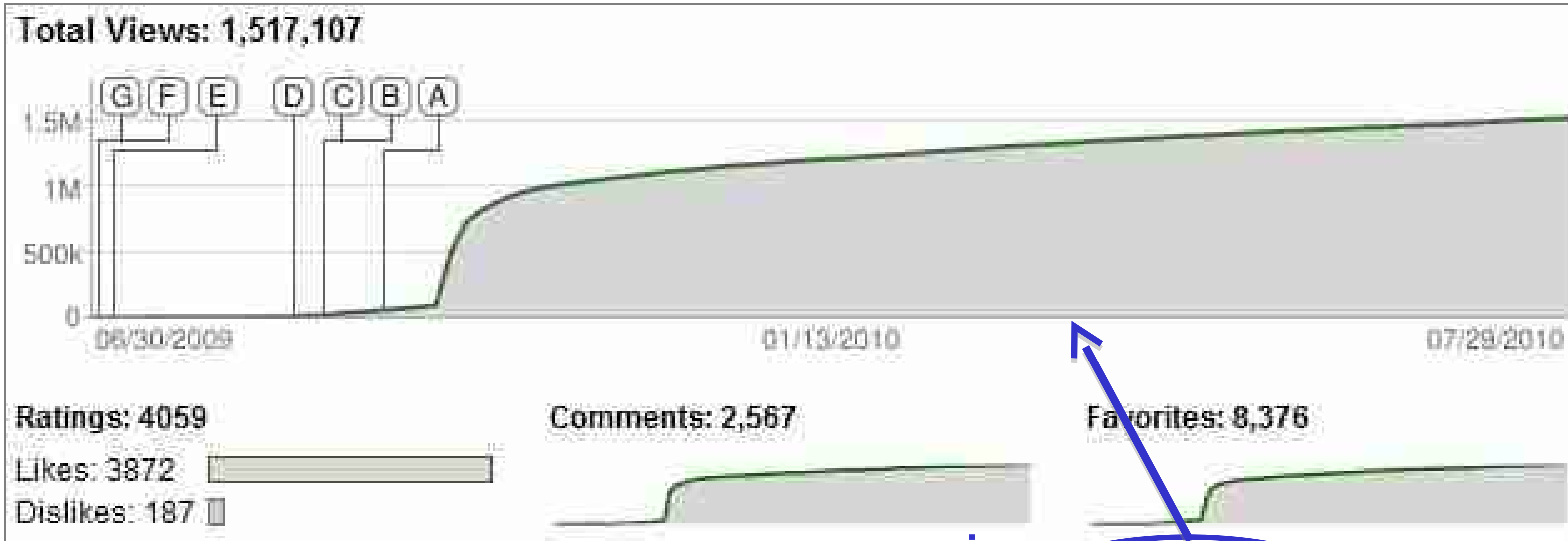
YouTube Case Study

Top: Videos that appeared on top lists

YouTomb: Copyright violated videos

Random: Videos selected based on random queries

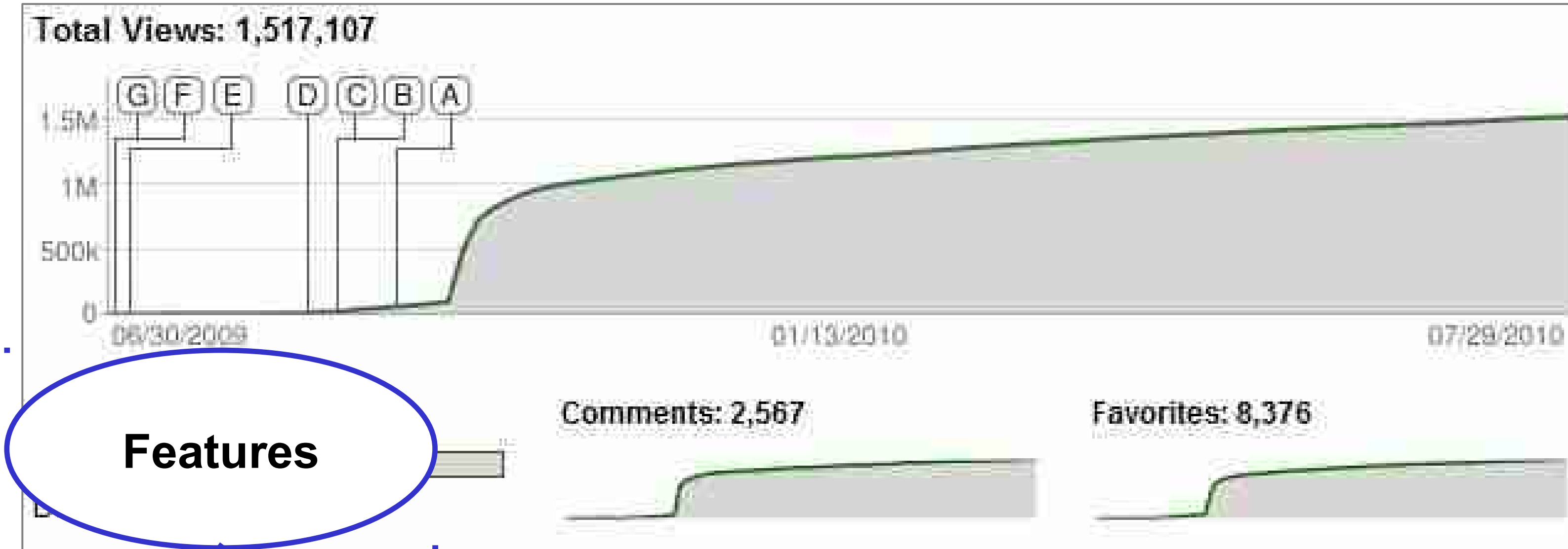
Datasets



Links

	Date	Link	Views
A	Sep 17, 2009	First featured video view	120,021
B	Sep 02, 2009	First view from a mobile device	83,883
C	Aug 30, 2009	First referral from YouTube search - rio de janeiro	25,807
D	Aug 23, 2009	First embedded on - www.facebook.com	130,515
E	Jul 07, 2009	First referral from YouTube - Homepage	99,886
F	Jul 02, 2009	First embedded on - www.orkut.com.br	42,353
G	Jul 01, 2009	First referral from YouTube search - rio 2016	145,217

Datasets



Features

Links

	Date	Link	Views
A	Sep 17, 2009	First featured video view	120,021
B	Sep 02, 2009	First view from a mobile device	83,883
C	Aug 30, 2009	First referral from YouTube search - rio de janeiro	25,807
D	Aug 23, 2009	First embedded on - www.facebook.com	130,515
E	Jul 07, 2009	First referral from YouTube - Homepage	99,886
F	Jul 02, 2009	First embedded on - www.orkut.com.br	42,353
G	Jul 01, 2009	First referral from YouTube search - rio 2016	145,217

Feature and content importance

How fast does a video become popular?

How concentrated is popularity?

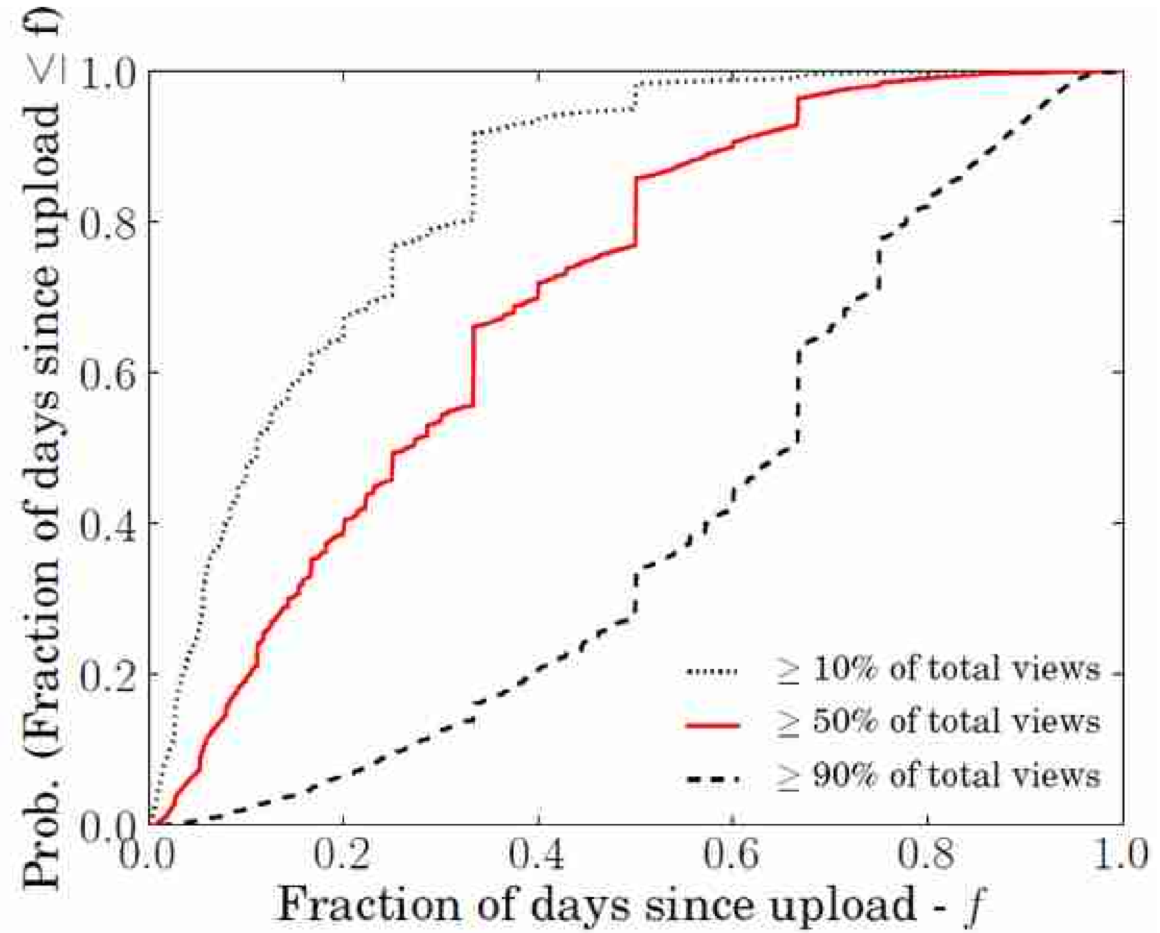
Why does a video become popular?

- Content / Referrers

Learning of similar trends amongst videos

- Clustering [Yang2011]

How fast?

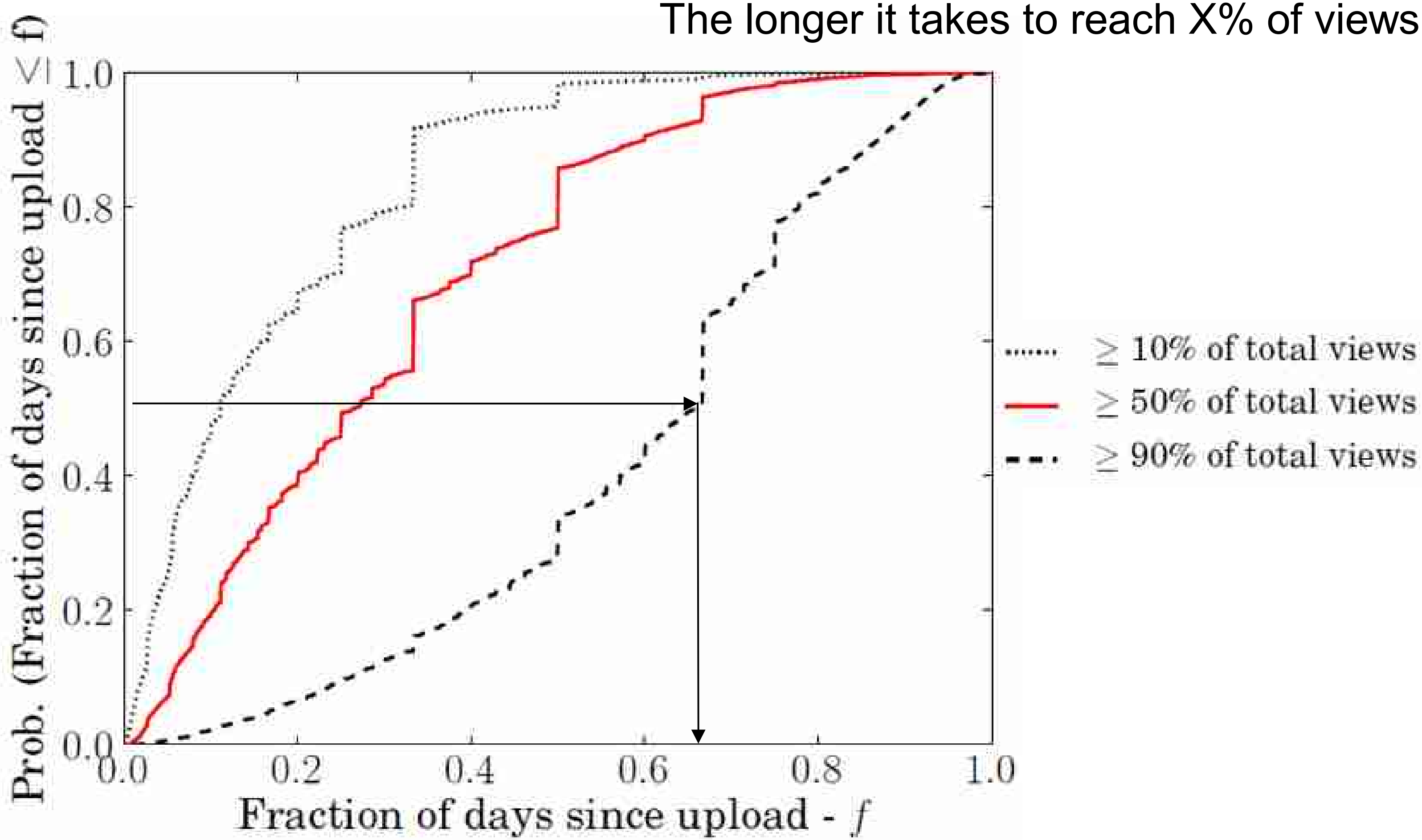


(a) Top

CDF of the fraction of time until X% of popularity reached

How fast?

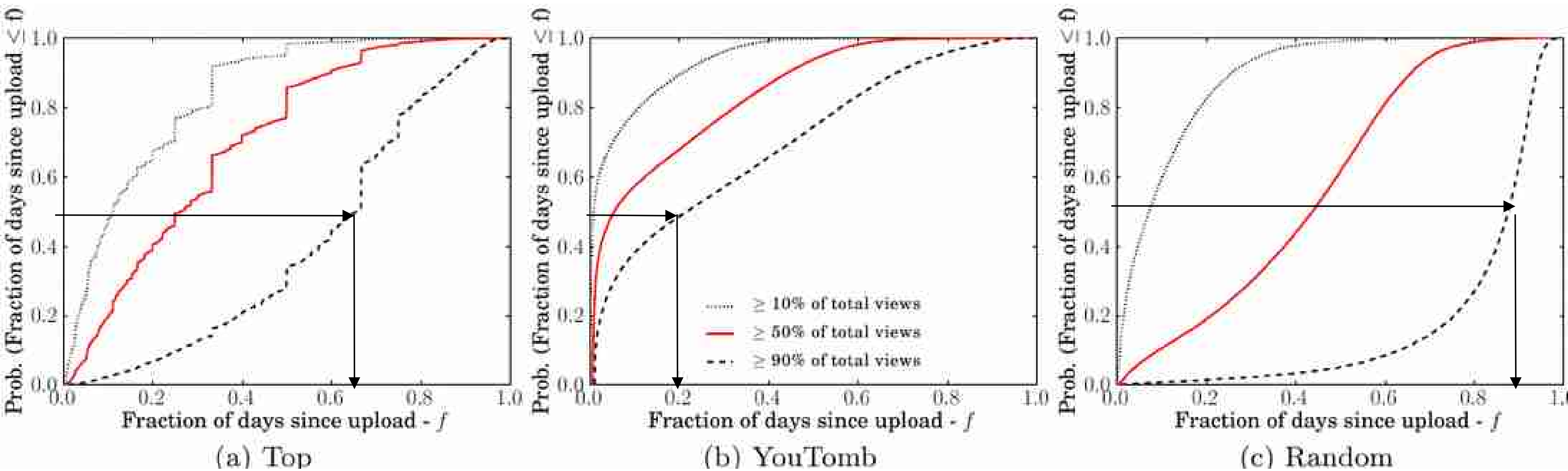
The more to the right the curve is,
The longer it takes to reach X% of views



(a) Top

50% of videos take 65% of lifetime to achieve 90% of views

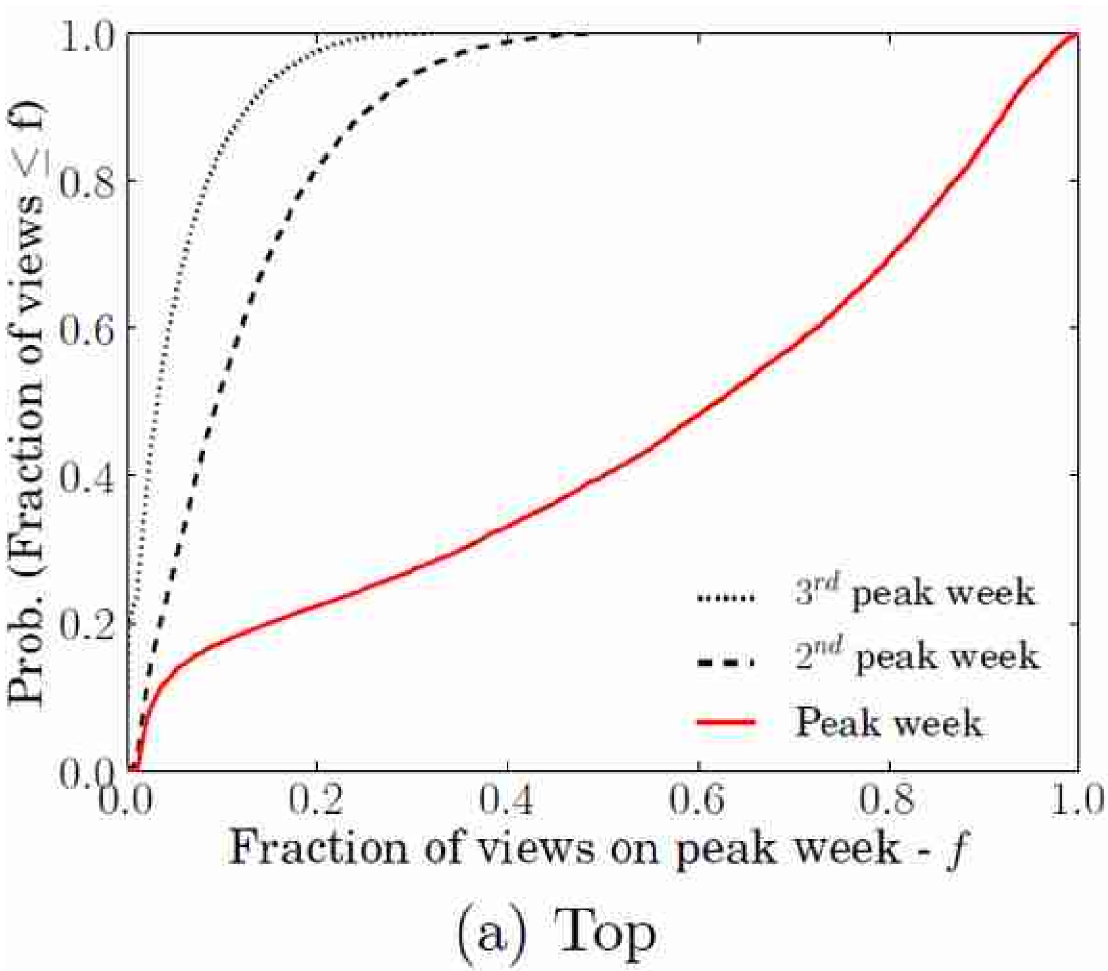
How fast?



For 50% of the videos:

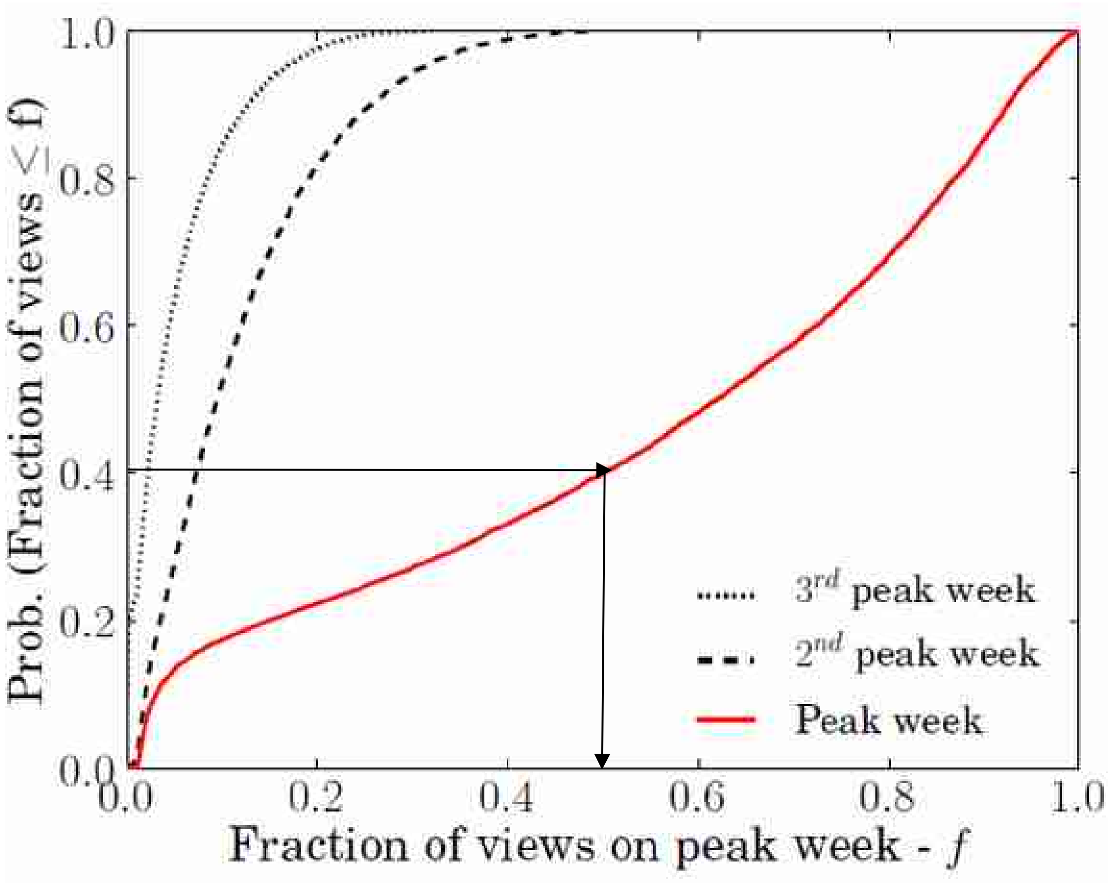
- YouTomb: 21% of lifetime to reach 90% of final popularity
- Top: 65% of lifetime for same 90%
- Random: 87% for same 90%

How concentrated?



CDF of the fraction of views on peak week

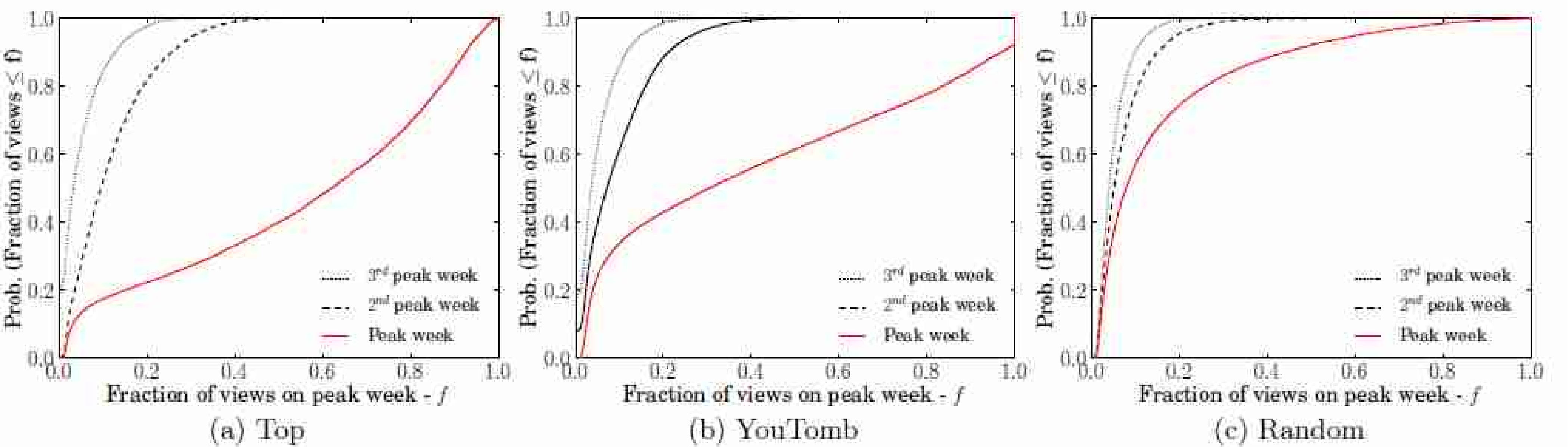
How concentrated?



(a) Top

At least 50% of views on peak week for 60% of videos

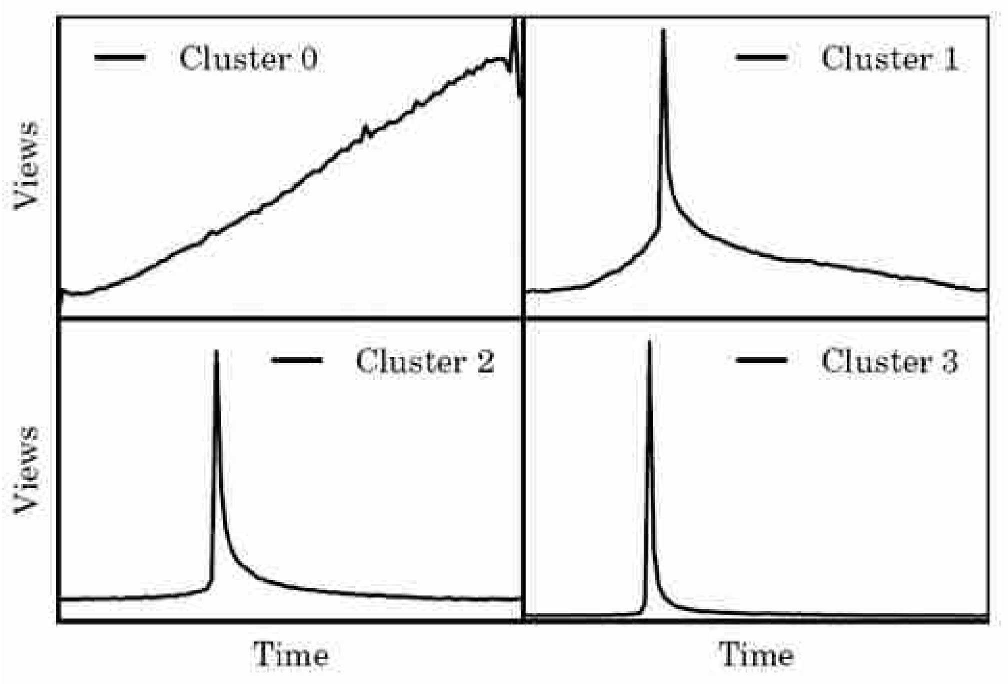
How concentrated?



For 60% of videos, most popular week corresponds to

- At least 50% of views for Top
- At least 40% of views for YouTomb
- At least 5% of views for Random

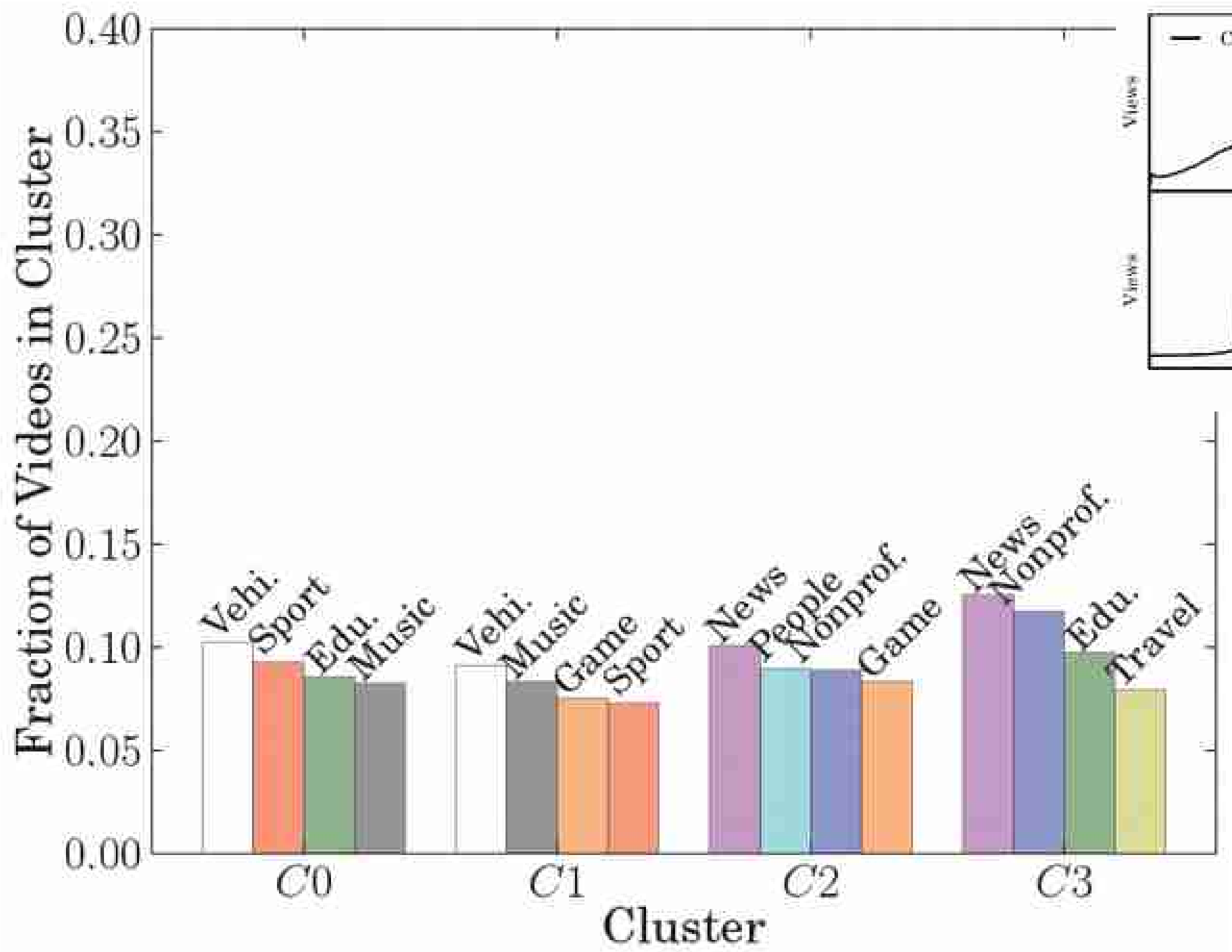
Patterns of Popularity Growth

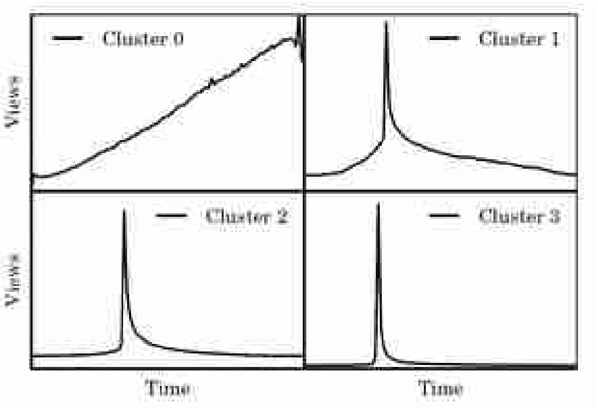


KSC-Algorithm

- 4 Clusters on all datasets
- Validation of [Crane2008,Figueiredo2011]

Types of Content per Cluster

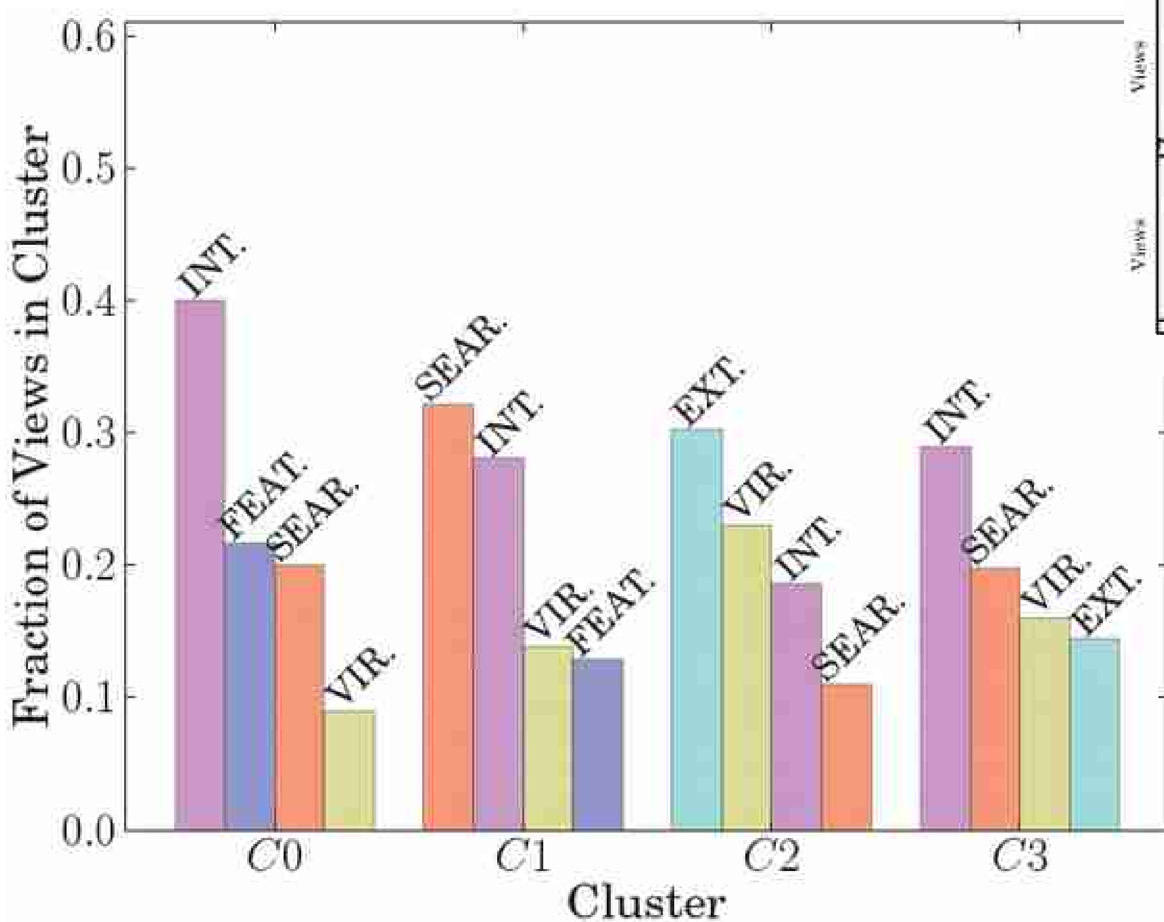


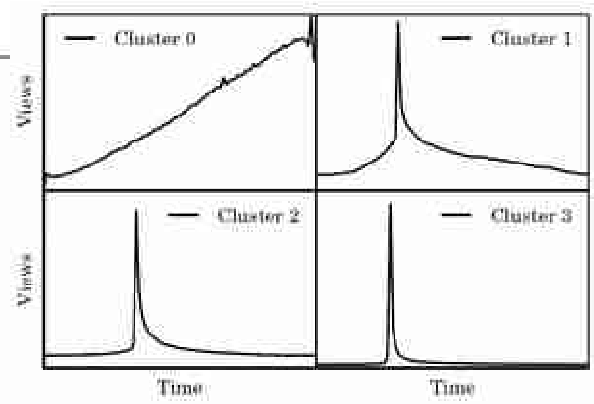


Fraction of videos per YouYube category

- Different types of content depending on the dataset
- Difference from whole dataset (chi-squared test)

How do users find this content?





Fraction of views per referrer type

- Search is very important, but also internal browsing
- Again, different concentration depending on cluster

Understanding Video-Ad Consumption on YouTube: A Measurement Study on User Behavior, Popularity and Content Properties,

RQ1: How do users consume video-ads?

RQ2: How does video-ad popularity evolve over time?

RQ3: What are the relationships (if any) between a video-ad and the video-contents with which it is associated?

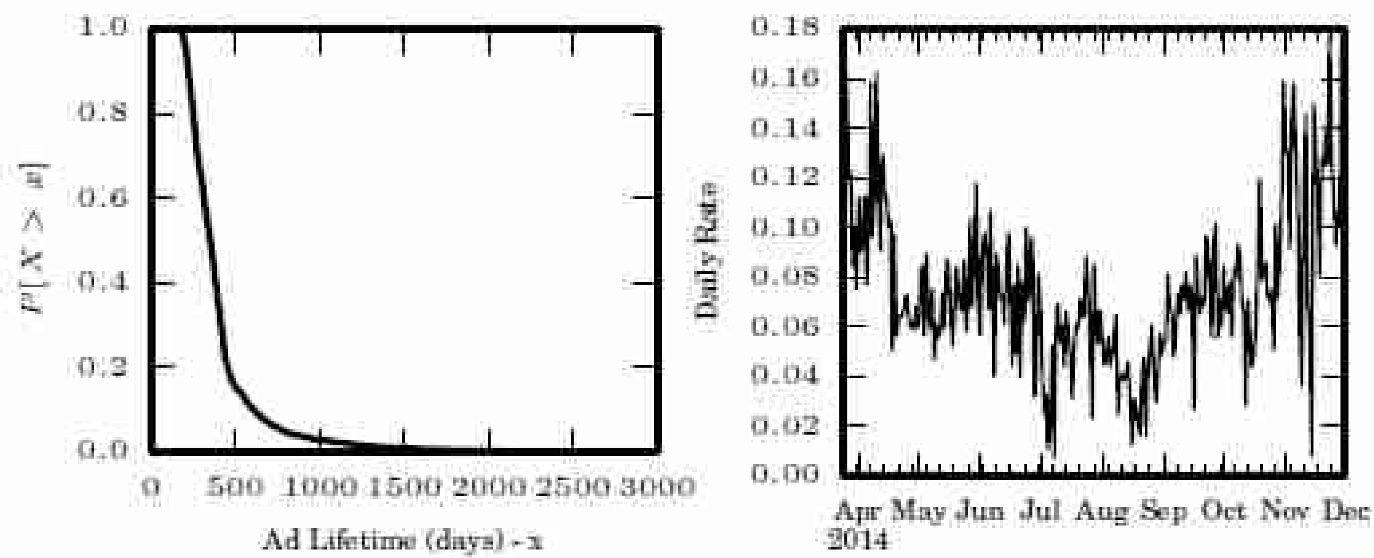
Datasets

Table 1: Summary of our datasets.

	Campus Network	API	HTML Stats
# of unique video-contents	58,082	47,007	-
# of unique video-ads	5,667	5,052	3,871
# video-ad exhibitions	99,658	-	-

Logs de trafego (local) + Dados do servidor (global)

Datasets

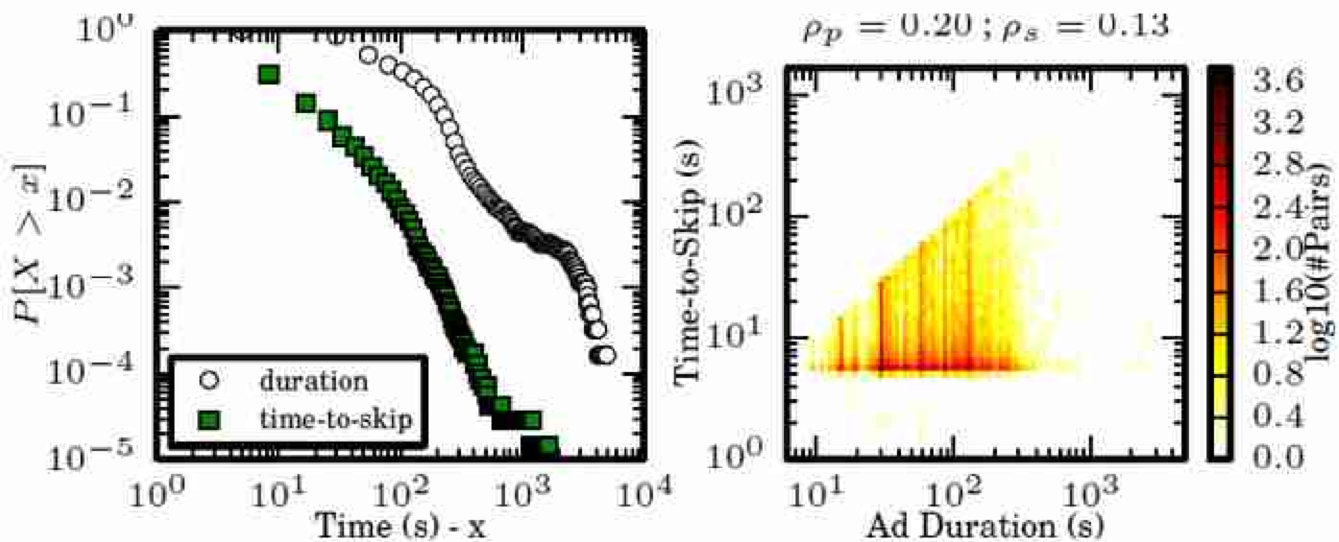


(a) Video-ad lifetimes (b) Daily fraction of pairings

Figure 3: Overview of video-ads in our datasets.

RQ1: How do users consume video-ads?

Only video-ads that were skipped by the user



(a) Duration and time-to-skip (b) Duration vs. time-to-skip

Figure 4: User behavior when exposed to video-ads: duration and time until user skips exhibition (time-to-skip).

29% video-ad exhibitions are in full

Out of those that were skipped:

- 35 %: skipping in less than 6 seconds
- 25 %:skipping after more than 10 seconds

RQ2: How does video-ad popularity evolve over time?

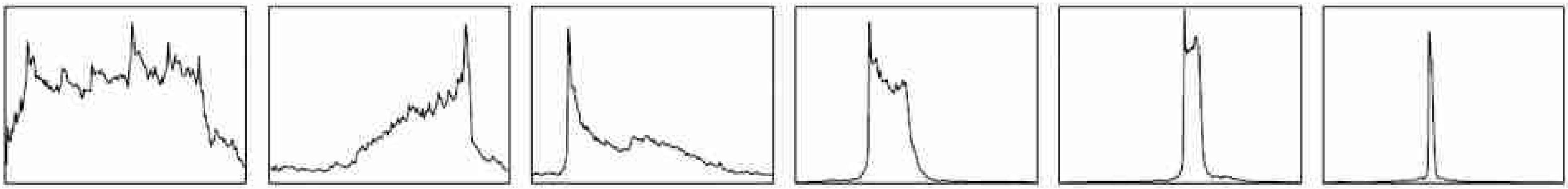
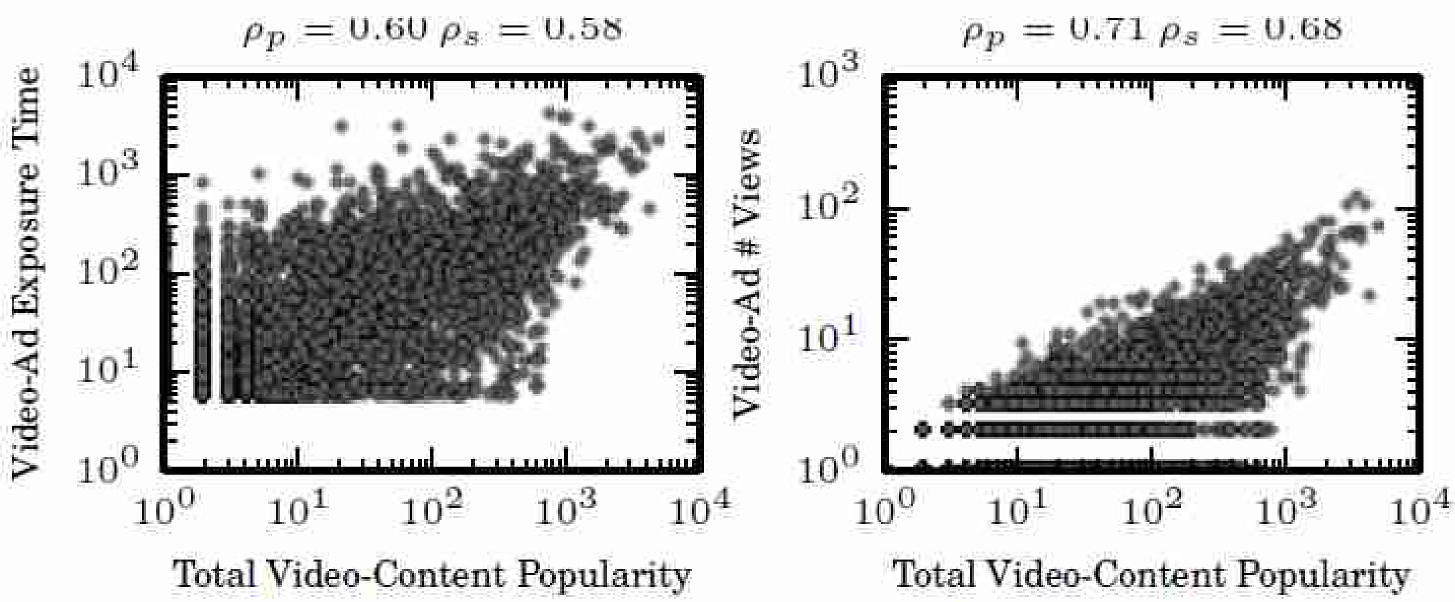


Figure 7: Trends (cluster centroids) of video-ad popularity evolution over time.

Table 2: Properties of each trend (cluster) of video-ad popularity evolution.

	C1	C2	C3	C4	C5	C6
# video-ads	69	108	109	293	467	569
Average Number of Views	1,486,175	1,869,906	4,882,094	1,789,798	1,451,894	984,175
Average Exposure Time	203,640,554	159,293,660	629,686,649	99,386,939	81,300,652	60,885,487
Average Exposure Time / Number of Views	137.02	85.19	128.98	55.53	56.0	61.86
Average Gini	0.24	0.61	0.58	0.82	0.9	0.92
Average Time to Peak	66	69	37	25	20	14

RQ3:What are the relationships (if any) between a video-ad and the video-contents with which it is associated?



(a) Exposure Time (b) Number of views

Figure 8: Popularity of video-ad versus total popularity (in # views) of all video-contents that were paired with the video-ad (measured in the campus network).

Motivation for popularity prediction models

Also observed: low content similarity between video-ad and video-contents (no contextual advertising?)

Tips, Dones and ToDos: Uncovering User Profiles in Foursquare

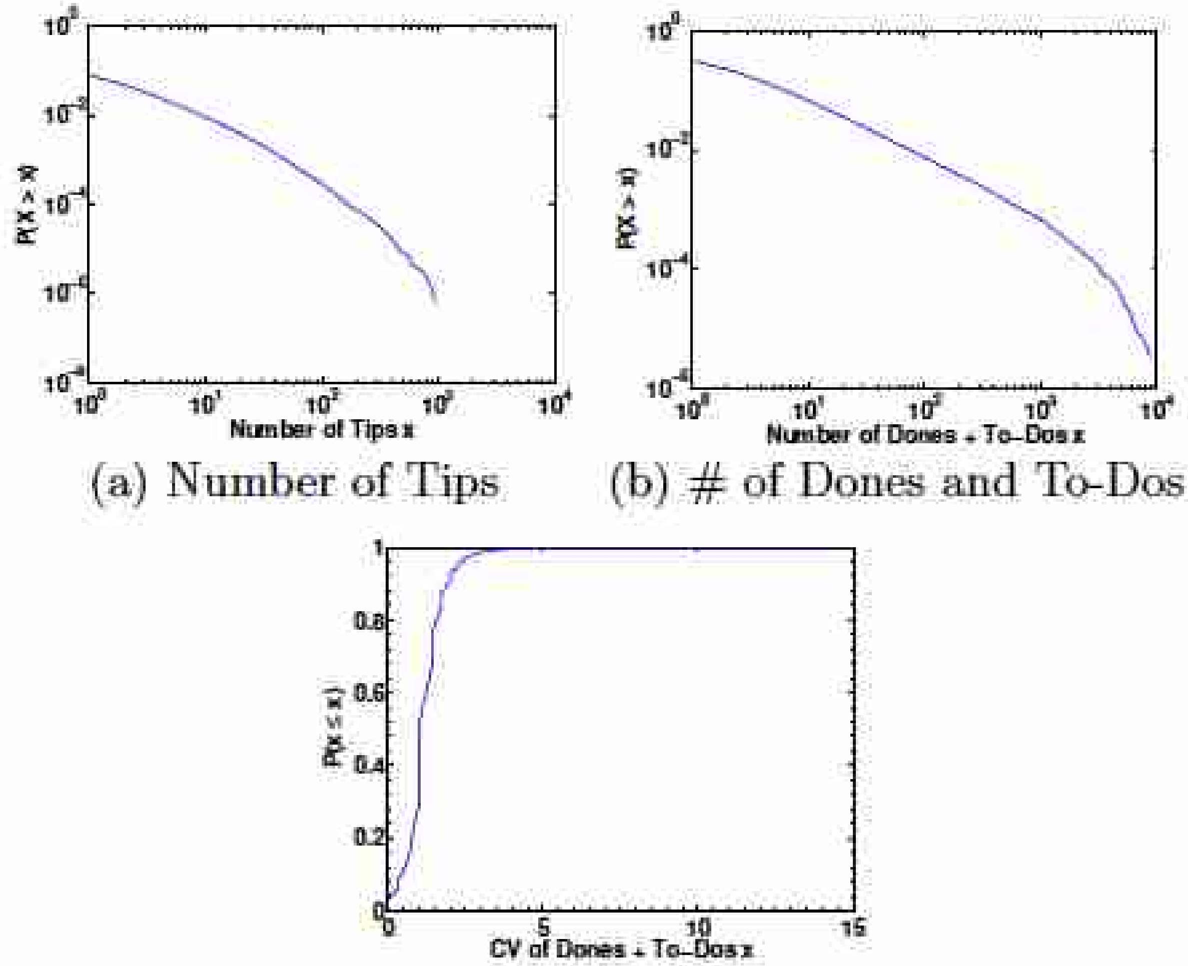
How do users use tips on Foursquare?

Dataset

Table 1: Summary of our Foursquare Dataset

Number of venues	1,601,412
Number of venues with at least one tip	296,217
Number of verified venues	61,378
Number of users	526,651
Number of brand users	1,248
Number of tips	984,251
Total number of dones for all tips	1,407,835
Total number of to-dos for all tips	393,574

Venue Characteristics



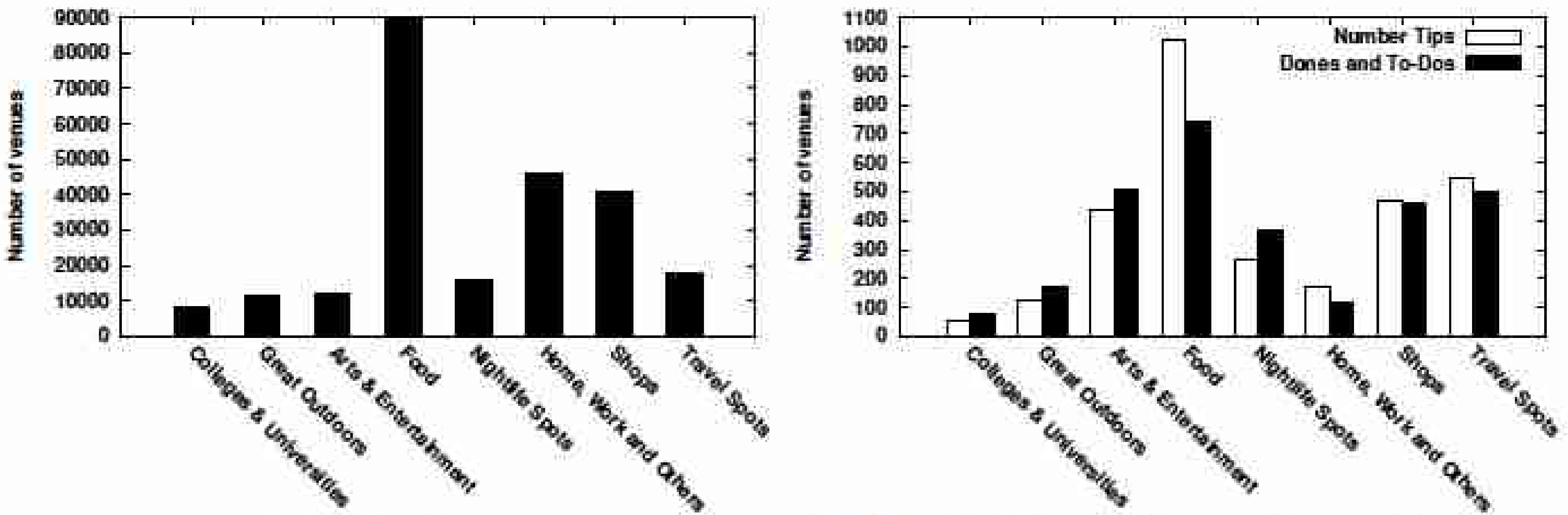
(a) Number of Tips

(b) # of Dones and To-Dos

(c) CV of Number of Dones and To-Dos per Tip

Figure 2: Distributions of Venue Attributes

Venue Characteristics

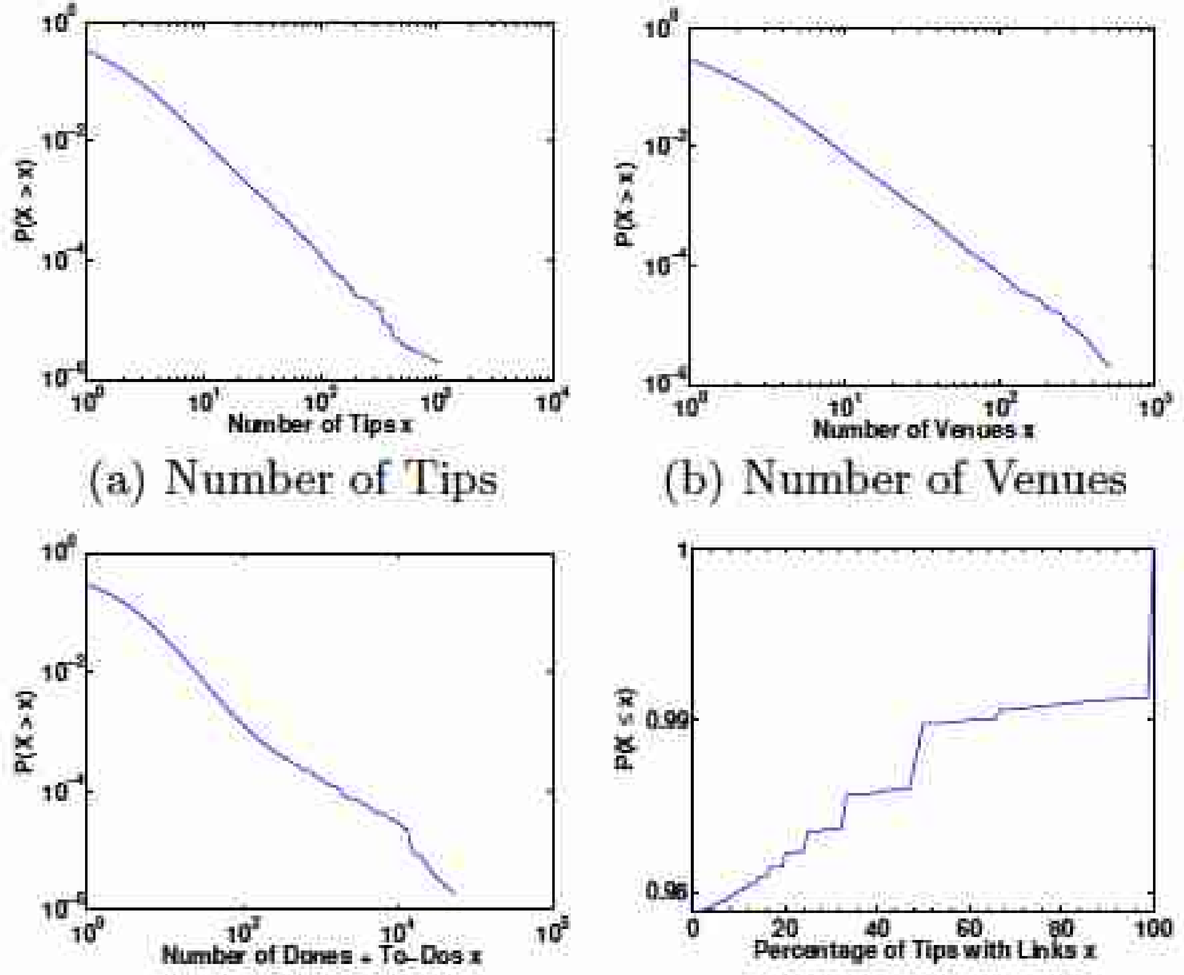


(a) All Venues

(b) Top 1% Venues According to Number of Tips and Number of Dones and To-Dos

Figure 3: Distribution of Venues Across Categories

User Characteristics

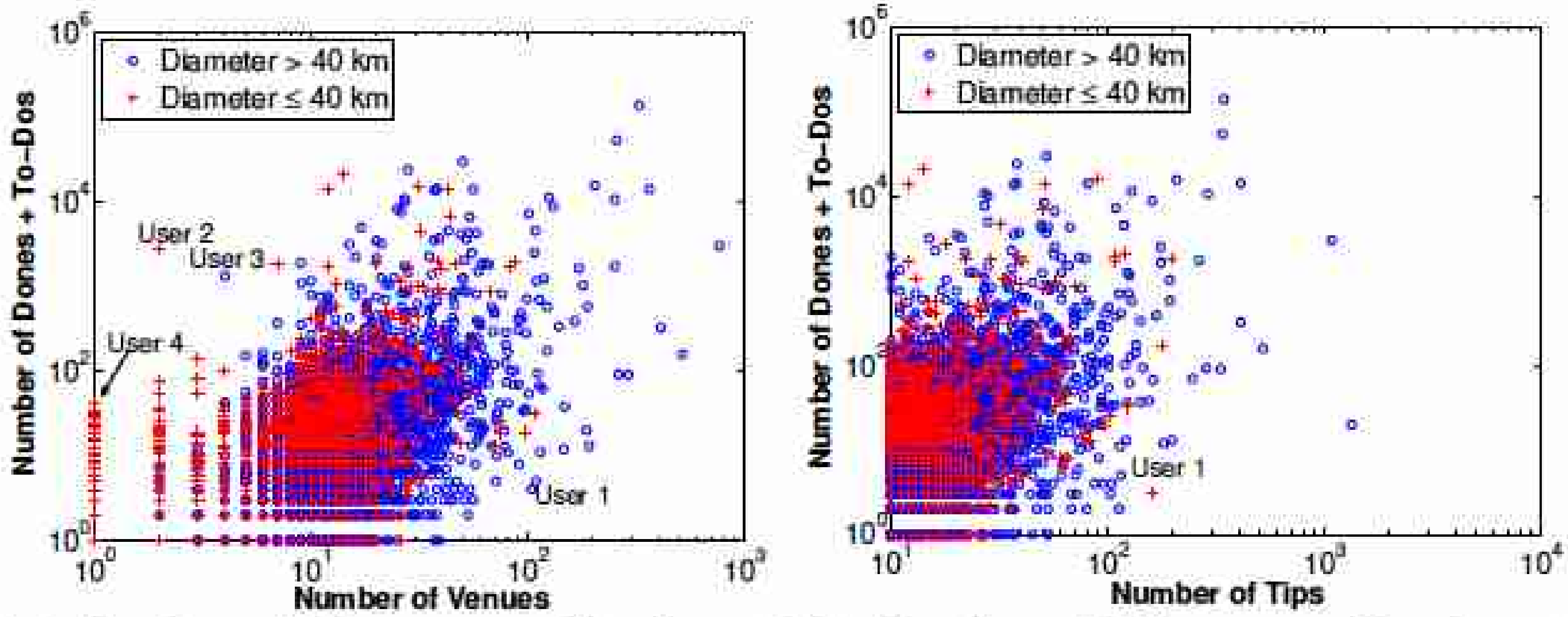


(a) Number of Tips (b) Number of Venues

(c) # of Dones and To-Dos (d) % of Tips with Links

Figure 4: Distributions of User Attributes

User Characteristics



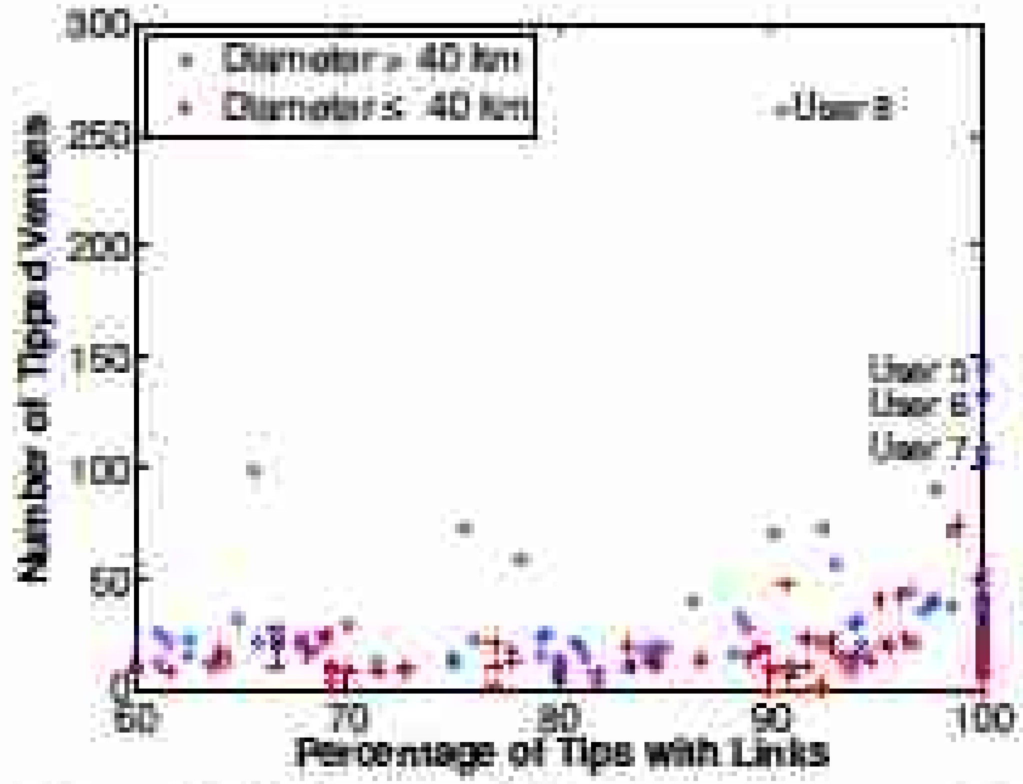
(a) Number of Venues vs. Number of Dones and To-Dos

(b) Number of Tips vs. Number of Dones and To-Dos

Figure 5: Correlation between User Attributes (only users with at least 10 tips)

Influence: both locally and globally (Users 2, 3 and 4 : brands)

Suspicious Behavior



(a) # Tipped Venues vs. % of Tips with Links

Spam: introduction of tips with unrelated links

User Profiles

Table 2: Summary of User Attributes Across Clusters

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	avg	cv	avg	cv	avg	cv	avg	cv
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Dones and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	

Clustering using K-Means

User Profiles

Table 2: Summary of User Attributes Across Clusters

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	avg	cv	avg	cv	avg	cv	avg	cv
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Dones and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	

Cluster 0 = spammers

Clustering using K-Means

User Profiles

Table 2: Summary of User Attributes Across Clusters

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	avg	cv	avg	cv	avg	cv	avg	cv
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Dones and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	

Cluster 1 = focused users who are not very active in the system

Clustering using K-Means

User Profiles

Table 2: Summary of User Attributes Across Clusters

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	avg	cv	avg	cv	avg	cv	avg	cv
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Dones and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	

Cluster 2 = more active users who receive many dones/to-dos

Clustering using K-Means

User Profiles

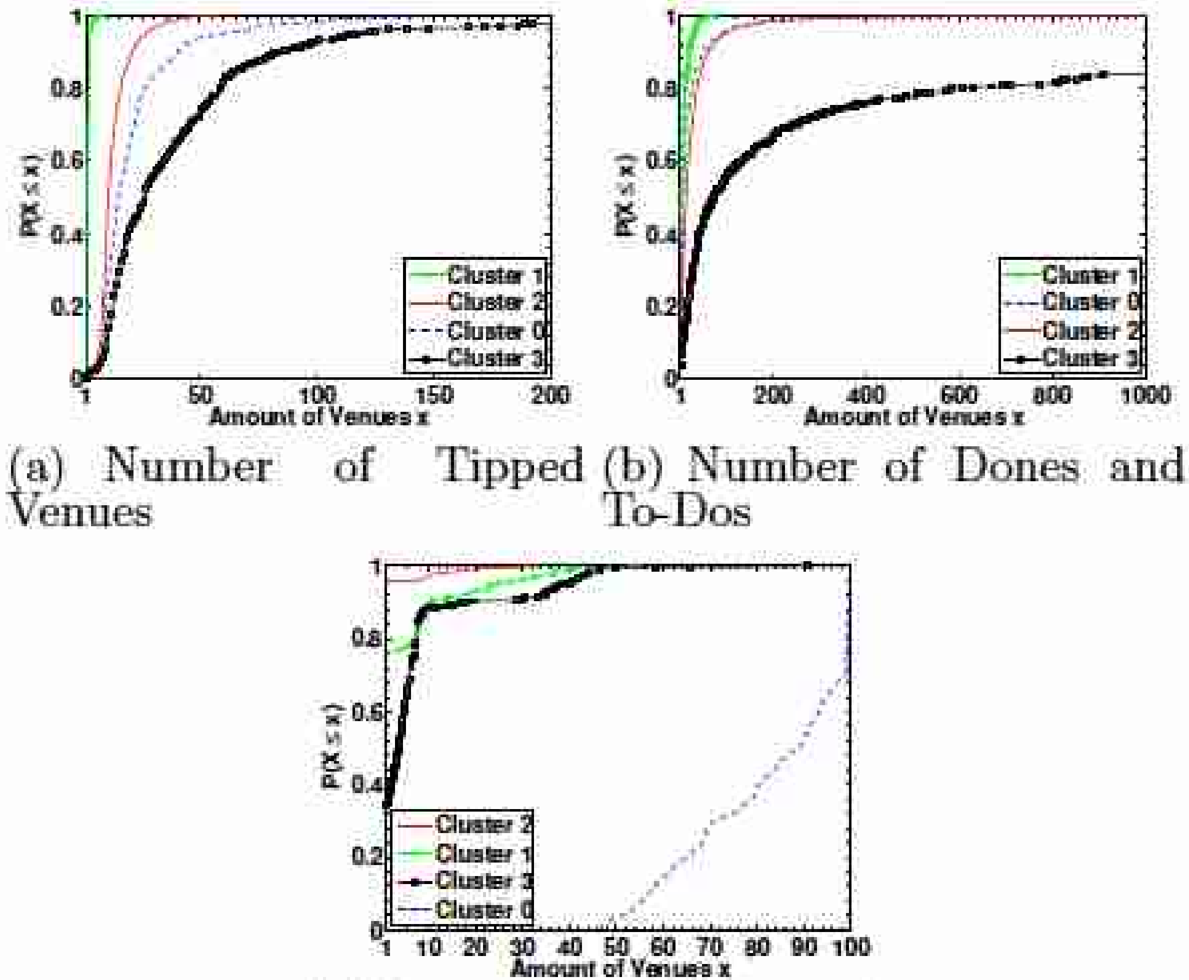
Table 2: Summary of User Attributes Across Clusters

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	avg	cv	avg	cv	avg	cv	avg	cv
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Dones and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	

Cluster 3 = very influential users who target many venues (often brands)

Clustering using K-Means

User Profiles



(a) Number of Tipped Venues (b) Number of Dones and To-Dos

(c) Percentage of Tips with Links

Figure 7: User Attribute Distributions

User Profiles

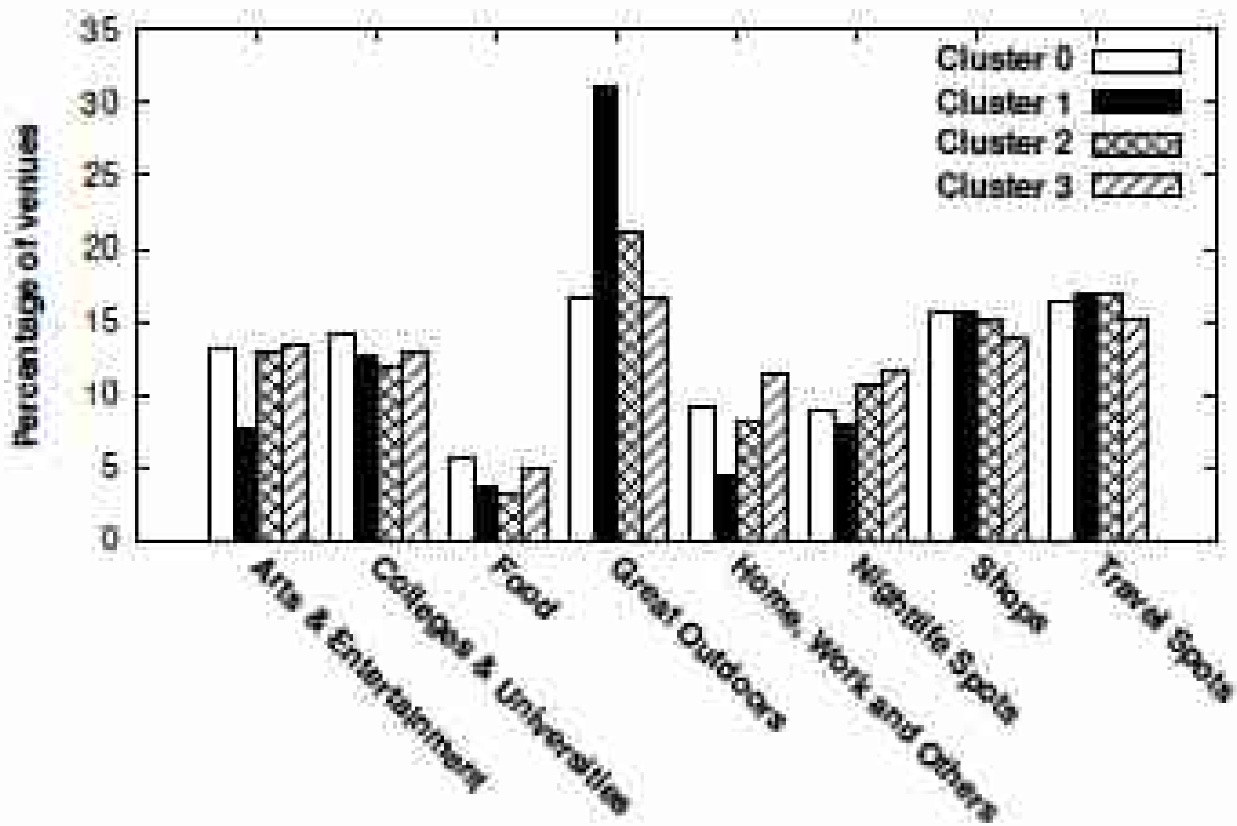


Figure 8: Venue Category Distributions