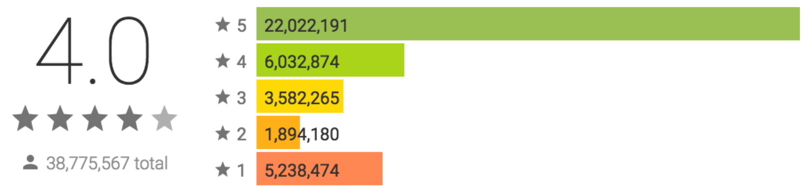
Voting with Their Feet: Inferring User Preferences from App Management Activities

Huoran Li[1], Wei Ai[2], Xuanzhe Liu[1], Jian Tang[3],
Gang Huang[1], Feng Feng[4], Qiaozhu Mei[2]

[1] Peking University [2] University of Michigan
[3] Microsoft Research [4] Wandoujia Lab

How to evaluate an app's quality?

Numerical ratings



Like/Dislike

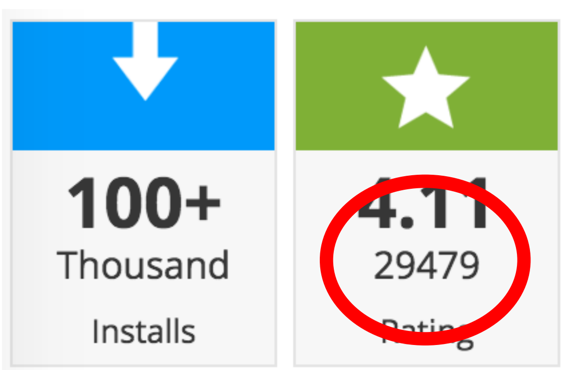
1958 万 人安装 | 1136 人喜欢 | 2838 人评论

Free-text comments

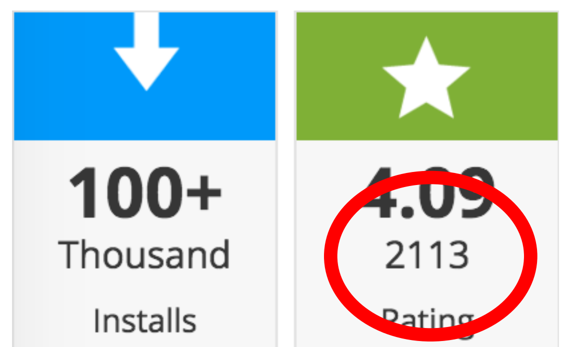
The most useful app in the world !
Best . Useful . Quick . Only drawback is the popping notifications which cannot be avoided even in the settings

However

- For many apps, although they have been downloaded lots of times.
- Only a small proportion of users have left reviews.
  - Some apps even have no reviews at all!
  - May causes biases



Lots of ratings



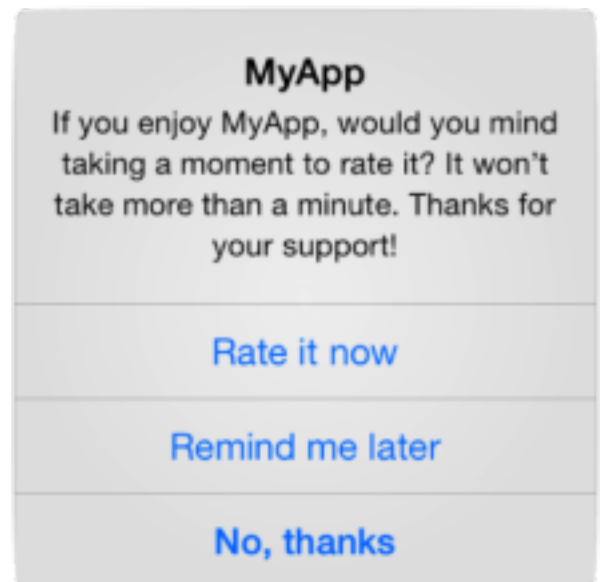
Not very much ratings

Think about it: have you ever…

- Downloaded an app
- Tried it
- Then uninstalled it…

- Uninstall an addicting game
- Get it back within one day

Have you left a rating for the app?



Users might be lazy

Has she left any ratings/comments? NO :(

Has she shown her attitudes? YES! :)

- Downloaded an app
- Tried it
- Then uninstalled it…

No, I don't like it

- Uninstall an addicting game
- Get it back within one day

Yes! I like it! I need it!

A Different Signal

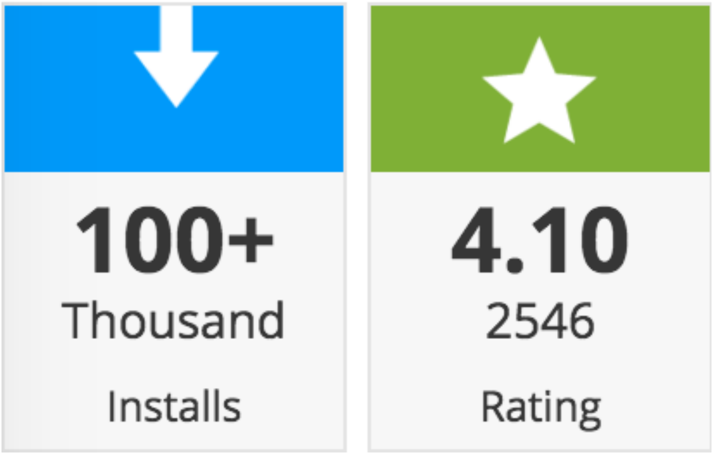
- Users may not rate an app
- But they vote with their feet!

User's management activities

can be used to evaluate apps' quality!

But wait a minute…

Number of Downloads as App Quality



- Recall that users may just download an app, have a try, and then abandon it
- One single number is not reliable.

More than Download

Download Uninstallation Update

⇩

Download – Uninstall => Bad

Uninstall – Download => Good

Download – Update – Update – Update => Love it!

⇩

Intuition: app management activities → better indicator of the user preference and the quality of an app ***even if it is not rated by any user!***

Research Questions

- Can we construct a large and representative dataset for this study?
- Are existing measures (rating & popularity) good enough?
- Can we find better indicators of app quality?
- Can they be combined to predict app quality?

Research Questions

- Can we construct a large and representative dataset for this study?
- Are existing measures (rating & popularity) good enough?
- Can we find better indicators of app quality?
- Can they be combined to predict app quality?

The Wandoujia Dataset

- A leading app marketplace in China
  - 200 million users till 2016
  - Similar to Apple App Store and Google Play
- Wandoujia provides a native management app, by which people can manage their apps
  - Search
  - Download
  - Update
  - Uninstall
- Wandoujia will record users' management actions

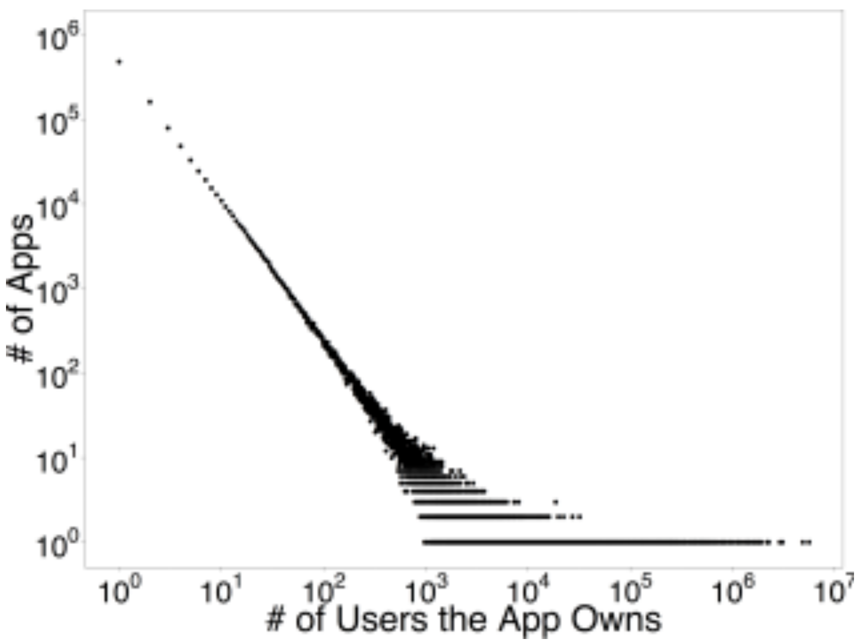
The Wandoujia Dataset

Time Span	5 months (May~September,2014)
# of Users (Devices)	17,303,122
# of Apps	1,054,969
# of Activities	240,108,930
# of Online Ratings	4,225,153

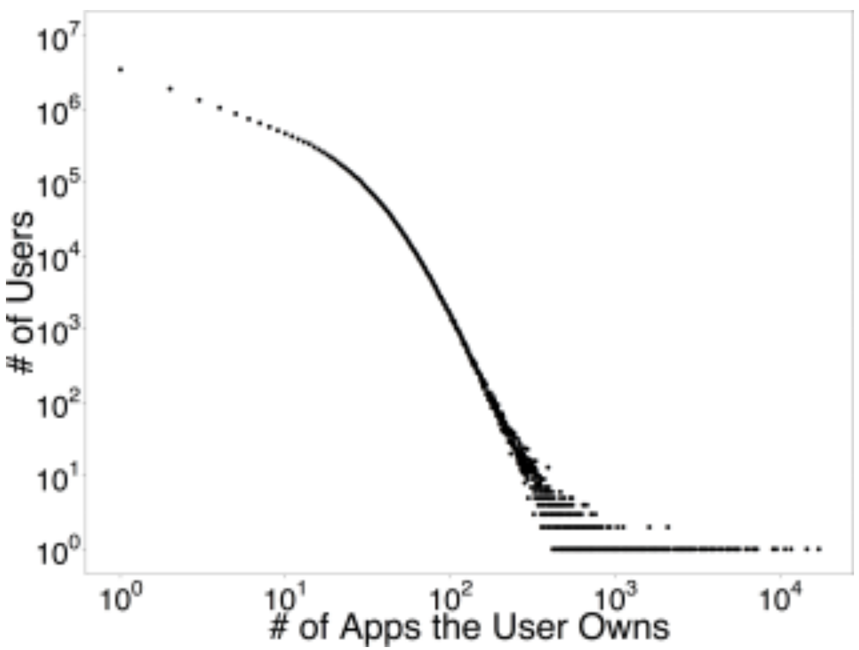
- We took a series of steps to preserve the privacy of involved users in our dataset.
  - All raw data was kept within the secure warehouse servers.
  - User identifiers are anonymized.

A Descriptive Analysis

users per app follows power-law.

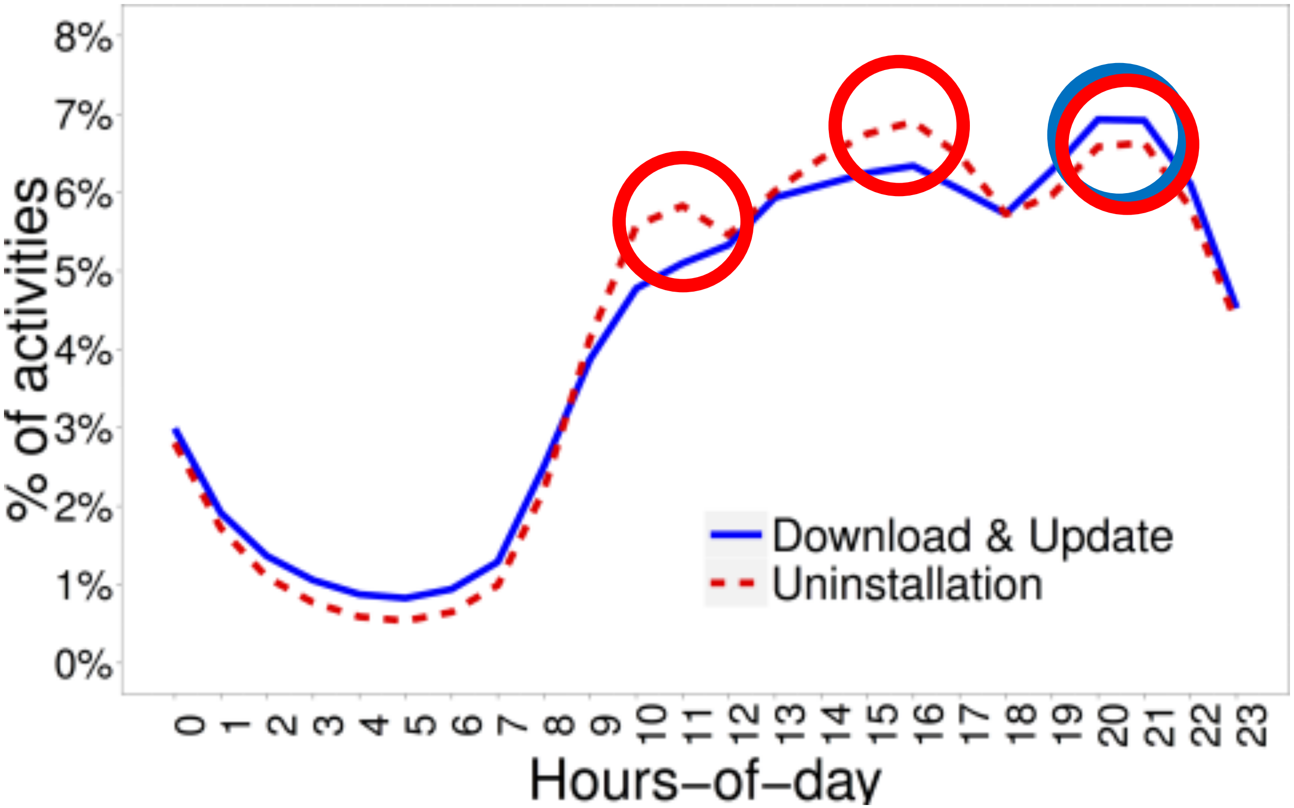


apps per user follows power law in the tail.



Activity Distribution over 24 Hours

- Downloading activities peak at television time.
- Uninstallations present three peaks.

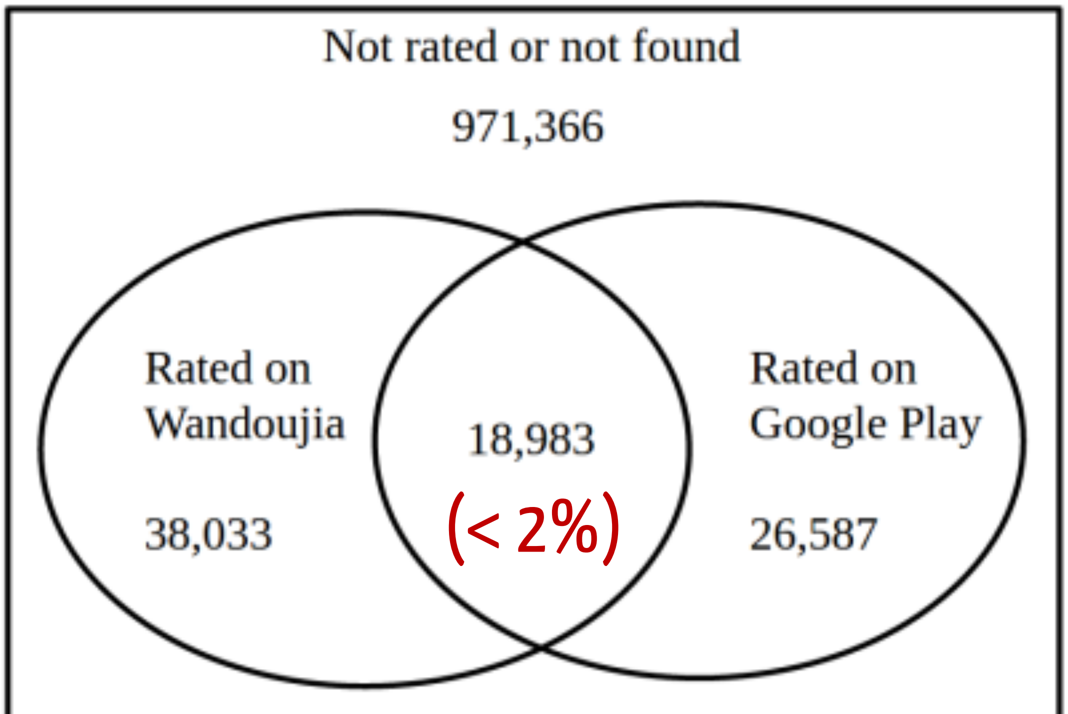


Research Questions

- Can we construct a large and representative dataset for this study?
- Are existing measures (rating & popularity) good enough?
- Can we find better indicators of app quality?
- Can they be combined to predict app quality?

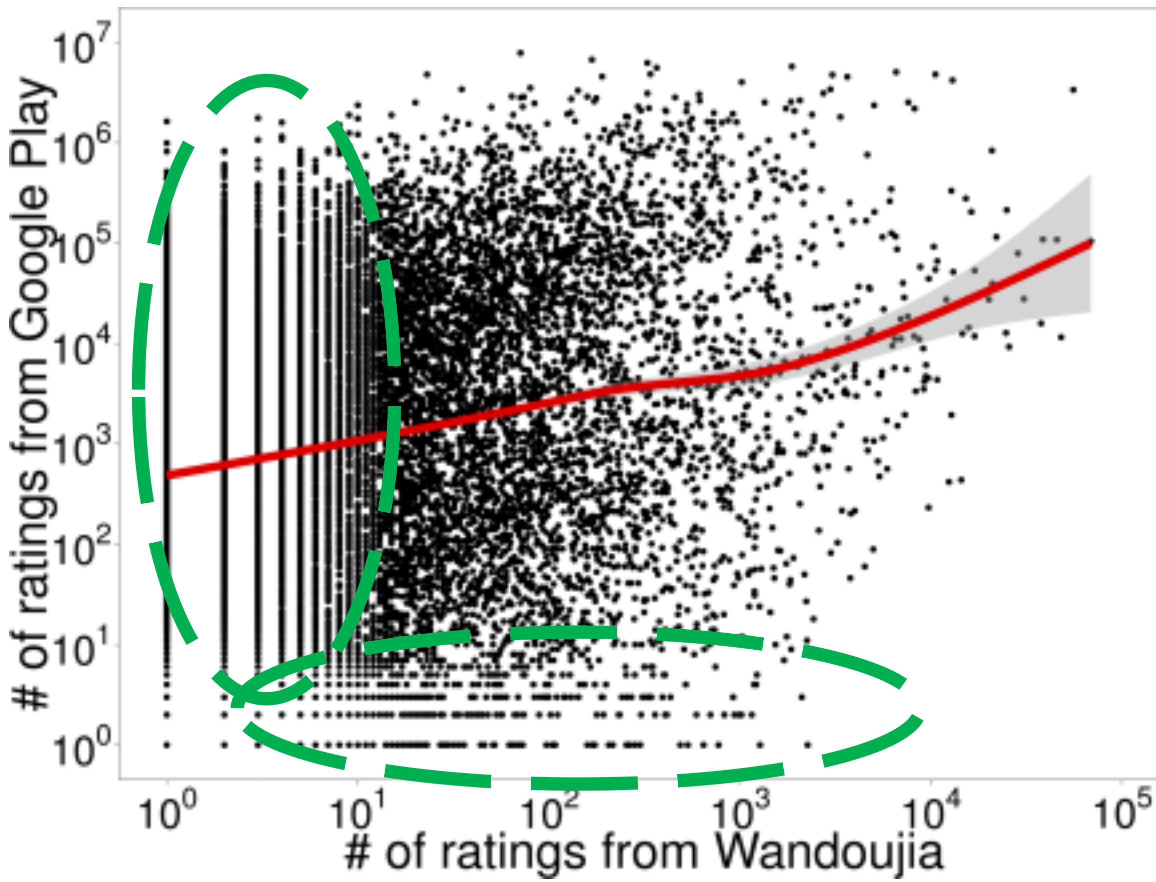
Limitations of User Ratings

Few apps get rated, even fewer get rated at both markets



Limitations of User Ratings

Number of ratings in two markets are correlated, but significant biases exist.

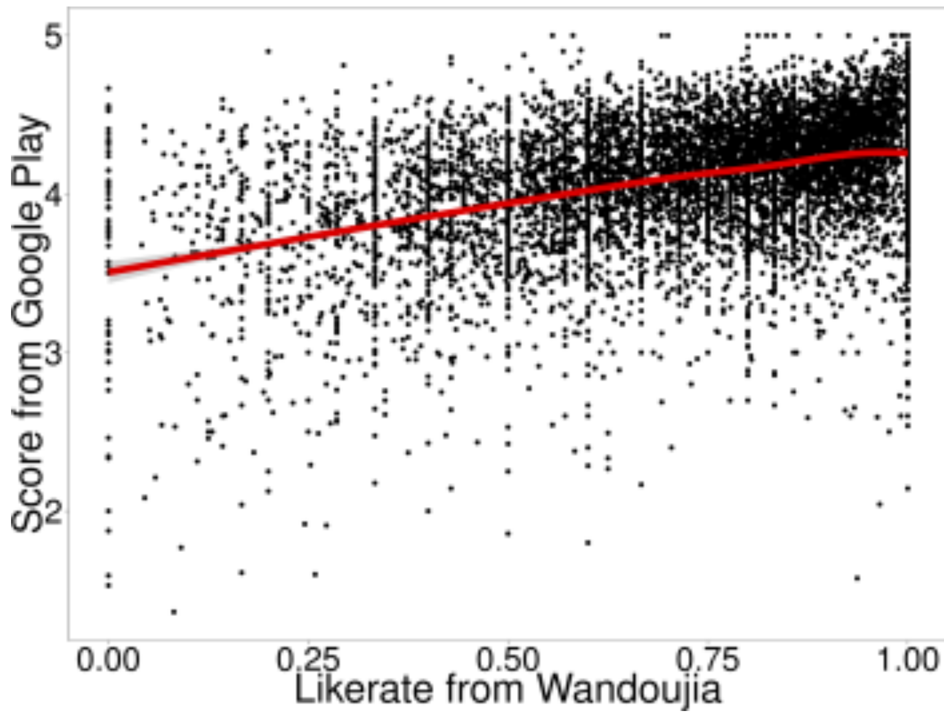


Limitations of User Ratings

Average ratings are correlated only if there are abundant of ratings.

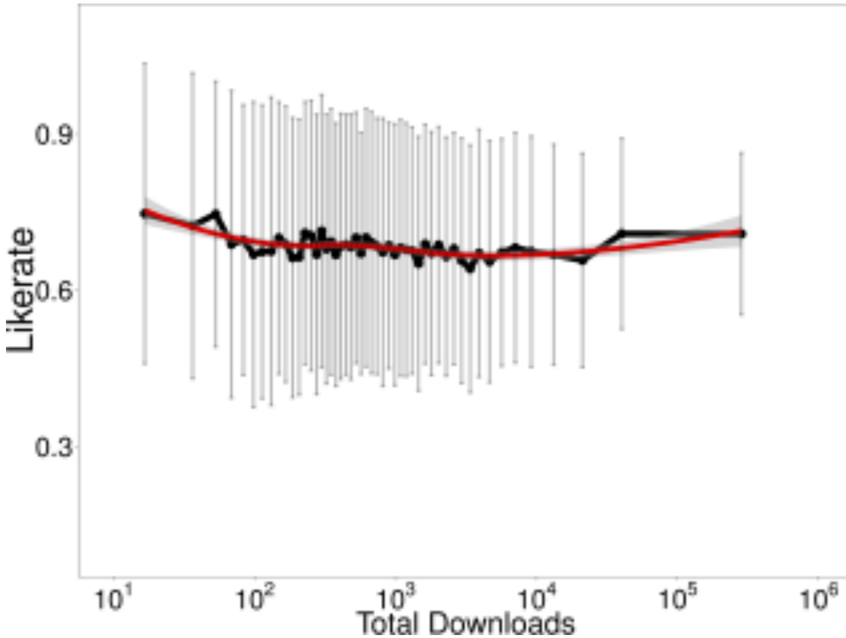
Score = Average rating

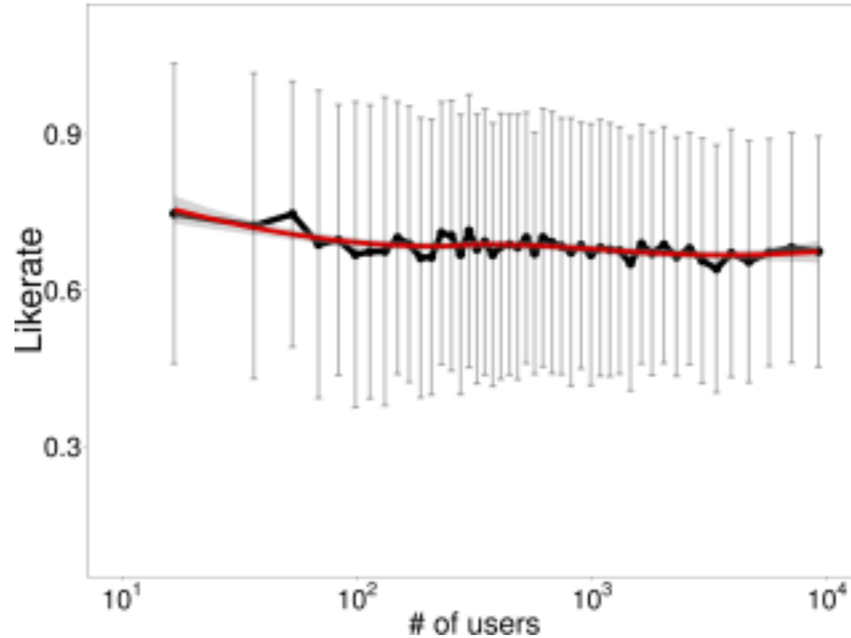
Likerate = like / (like + dislike)



(Apps receiving 5+ ratings from both marketplaces)

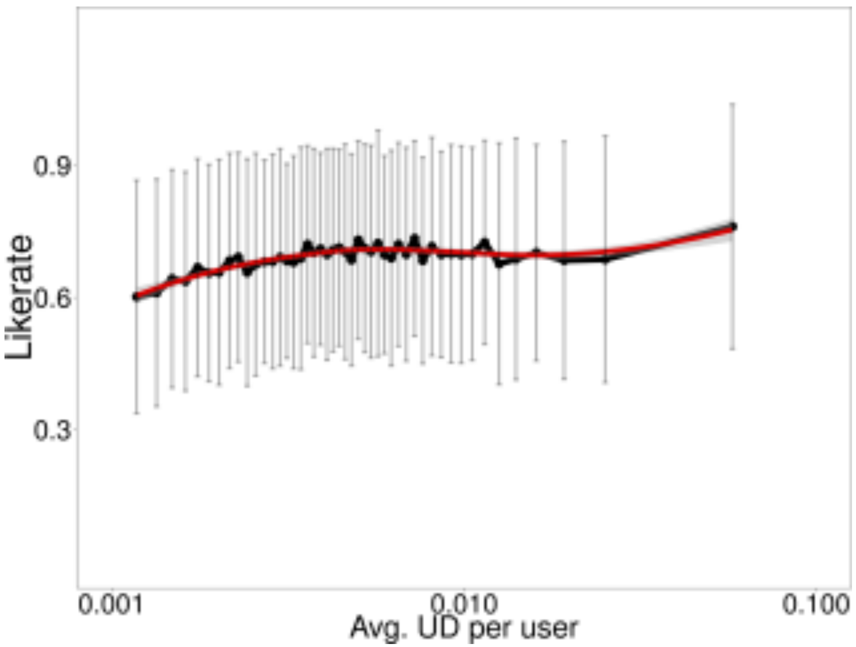
Popularity Indicators: not Reliable





A Sequential Feature can be More Promising

Uninstallation-Downloading



- Avg. #UD per user is in general ***positively*** correlated with the likerate of the app
- UD sequences are relatively rare
  - Only 4% sequences contain "UD."

Take a Look Back

- User ratings:
  - Rareness ☹
  - Biases ☹
- Popularity indicators:
  - Not as reliable as expected ☹
- The sequential indicator UD:
  - Promising ☺
  - Rare ☹

What did we do?

Find all sequential indicators and study them systematically.

Research Questions

- Can we construct a large and representative dataset for this study?
- Are existing measures (rating & popularity) good enough?
- Can we find better indicators of app quality?
- Can they be combined to predict app quality?

Evaluation Metrics

- User ratings as ground-truth (if there are abundant)
- Compare the ranking of apps

Kendall's Tau

Takes the entire list into consideration, where all apps contribute equally.

Mean Average Precsion (MAP)

More weight on the top-ranked items.

Baseline: Ranking by Popularity

	Total Downloads	# Users
Kendall's Tau	-0.2372	-0.2436
MAP	0.8629	0.8574

- A reasonable MAP, a miserable Tau:
  - Good news for popular apps
  - Bad news for long tail or new apps

Experiment Setup: a Prediction Task

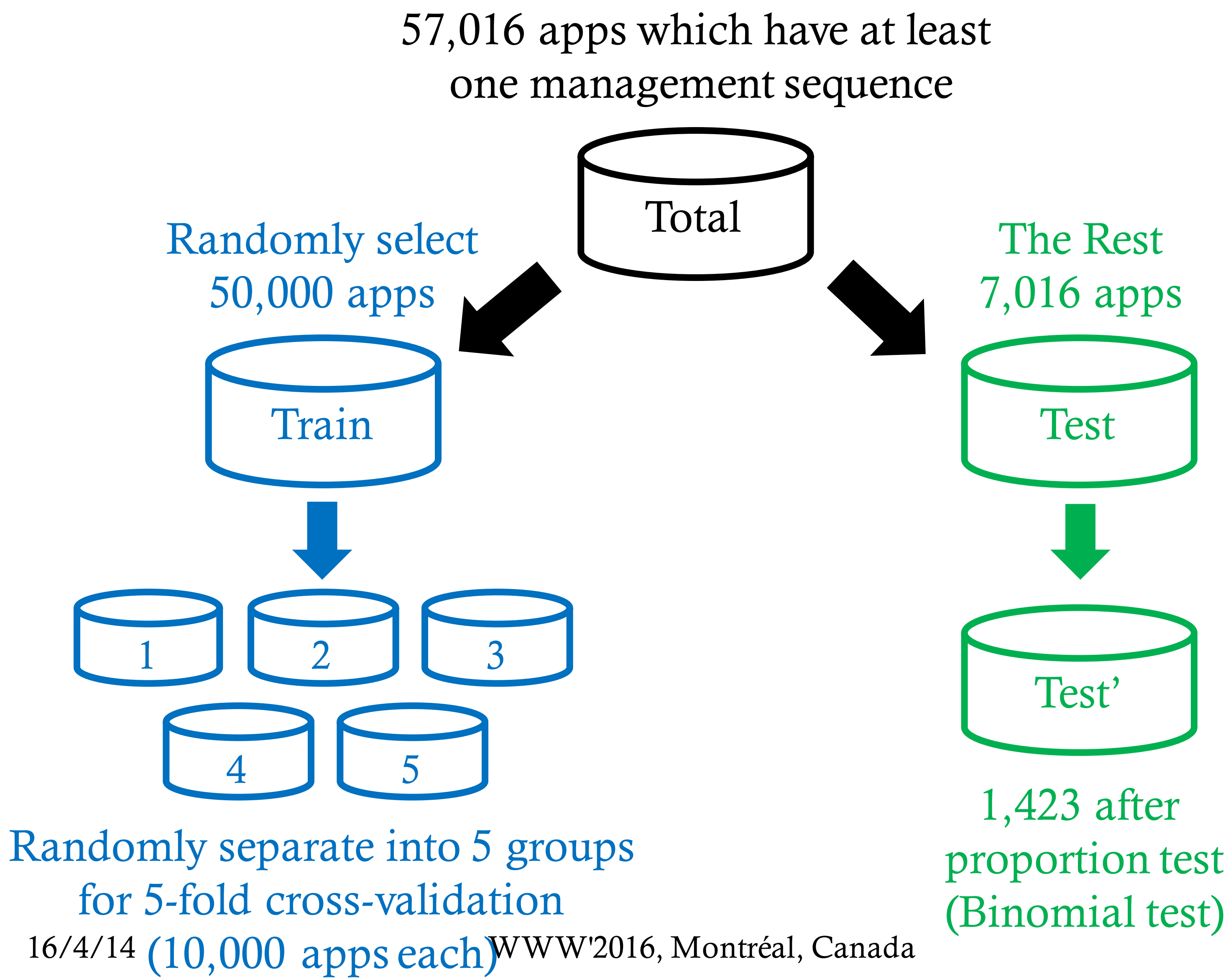
- Use sequential patterns as features

DU UD…

- Use ***likerate*** as outcome variable

- State-of-the-art machine learning algorithms for prediction
  - Ridge regression
  - Lasso regression
  - Random Forest (***RF*** for short)
  - Gradient Boosting Regression Tree (***GBRT*** for short)

Experiment Setup: Dataset Filtering



Experiment Results: Unit Features

- **#User**, Avg.Action, **Download**, Update, Uninstallation
- #User & Download are already used by marketplace

- MAP from 0.8574, 0.8629 to 0.9333
- Tau from -0.2436, -0.2372 to 0.0923

Experiment Results: Sequential Features

- Construct a sequence for every user of every app
  Download(D), Update(P), Uninstallation(U), Start(S), Ending(E)
  - E.g., SDPPUDE
- Extract *N-gram* features from the sequences
  - $N = 2, 3, 4, 5$
- Tau increases from 0.0923 to 0.1180
- MAP increases from 0.9333 to 0.9423

Experiment Results: Sequential Features with Time Intervals

- Time interval between consecutive activities.

- Insert a “T” for every 24 hours into sequences.
- Replace 4+ consecutive Ts with “*”
- Insert “-” for two consecutive activities within 24 hours

- e.g. SDTTP*P-UTDE

- Tau increases from 0.1180 to 0.1716
- MAP increases from 0.9419 to 0.9512

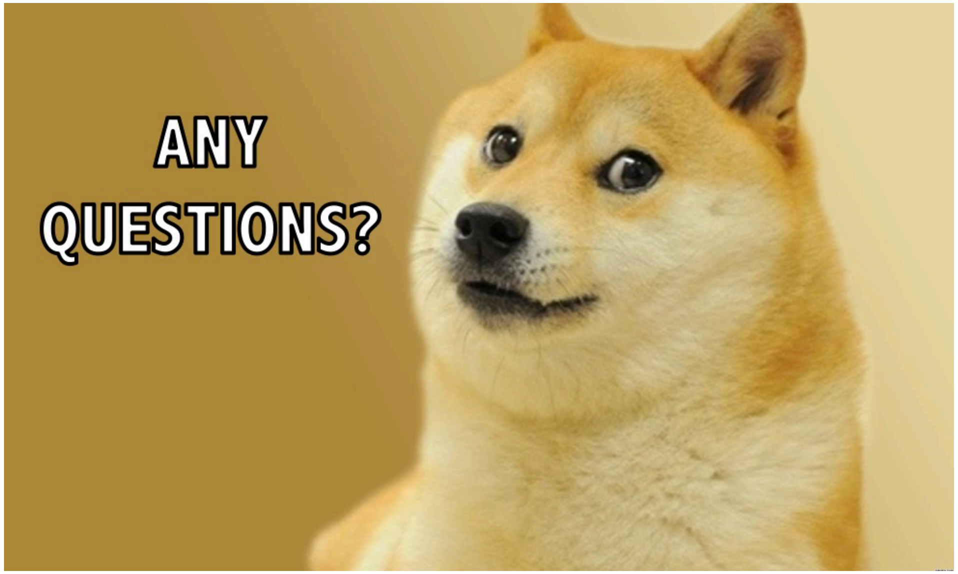
Interesting Findings

- Highest Tau (0.1716) achieved by Lasso.
  - Only one feature selected: "SD-U"

- Highest MAP achieved by GBRT.
  - Multiple features are utilized.
  - Many of them are variants of "DU": SD-U, D-UE, D-U, etc.
  - The "UD" indicator is also ranked high.

Take Away Points

1. User ratings are sparse.
2. Number of downloads is not a good indicator of app quality.
3. Users download, update and uninstall apps differently when they like or dislike the apps.
4. App management activities are good predictors, which alleviate the biases and sparsity of online ratings of apps.



Thank you for your attention!

Huoran Li[1], Wei Ai[2], Xuanzhe Liu[1], Jian Tang[3], Gang Huang[1], Feng Feng[4], Qiaozhu Mei[2]

[1] Peking University [2] University of Michigan
[3] Microsoft Research [4] Wandoujia Lab