Detecting User Interests on Twitter via Seed Set Expansion

Amit Goyal, Praveen Bommannavar, Stuart Anderson, Alek Kolcz, Kurt Smith

Social Networks

f
Insta
foursquare
LinkedIn
YouTube
tumblr.
B
flickr
P
S

User Interests Modeling

- *Question*: Which users are interested in what topics?

- *Several use-cases:*
  - Recommendations
  - Search
  - Consumer insights
    - What kind of users are interested in which topics?
    - How many users are interested in each topic?
    - Which topics are popular in a specific country?
    - What are the growth trends among users interested in various topics?
    - Which topics are growing/shrinking, in terms of active user counts?
    - How do various events impact growth trends in various topic populations?

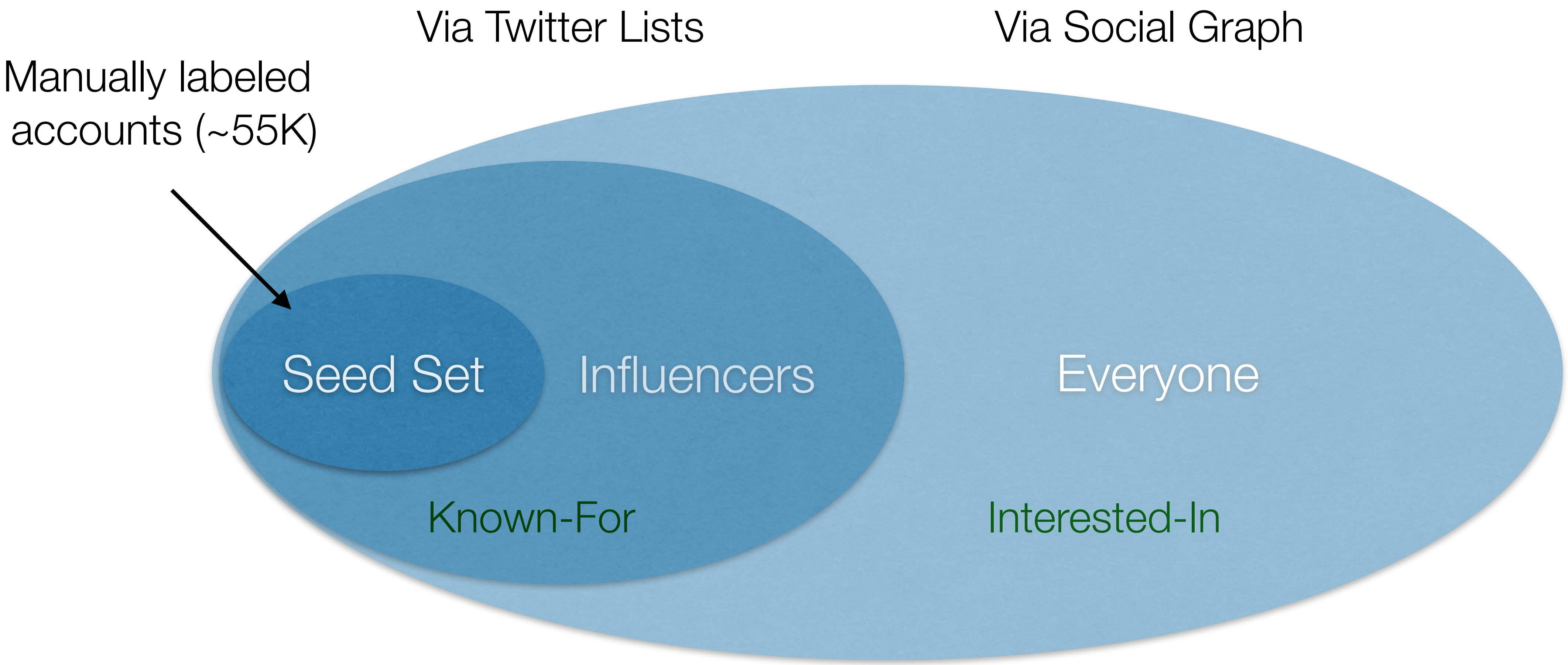
Challenges

- Text processing?
  - Difficult to scale to international markets.
  - Tweets are short (140 characters).
  - Sparsity in data — several users tweet rarely.

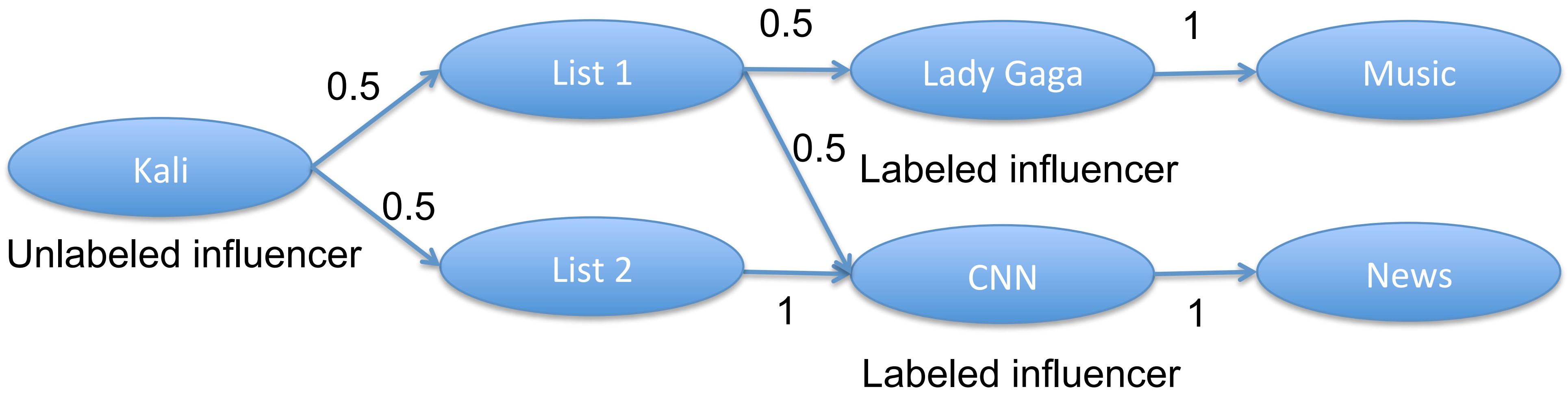
In this work, we propose a text independent graph-based approach to user interest modeling.

Our Approach

- We distinguish b/w influencers and others.
  - Known-For topics for influencers
    - E.g. Justin Bieber is Known-For Pop Music
    - An influencer is someone who has >= 10K followers
  - Interested-In topics for everyone



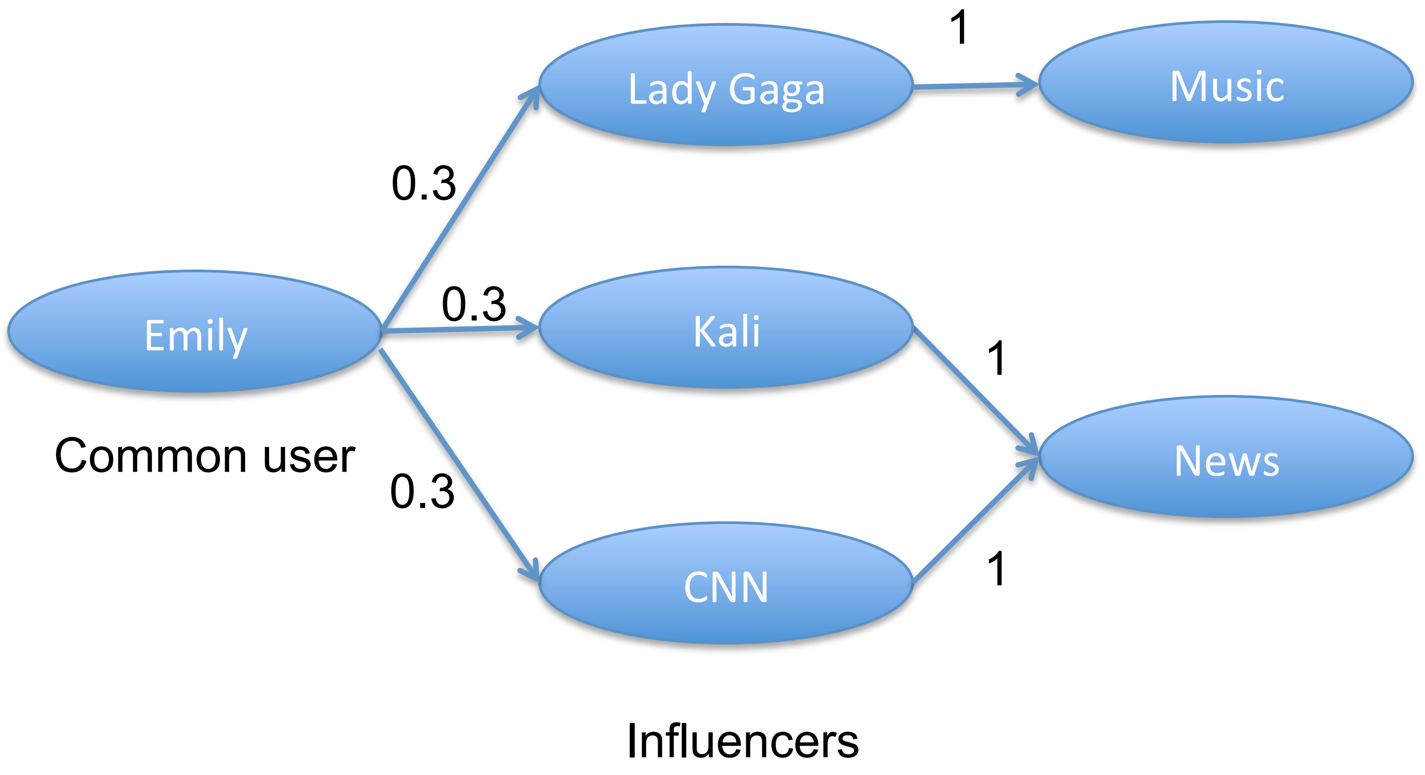
Learning Known-For Labels



Kali is Known-for for News with prob. 0.75

- Twitter Lists
  - On Twitter, a user can create her own lists, or follow the lists created by other users.
  - E.g. A user would put Lady Gaga and Justin Bieber in a list to have a filtered timeline for Pop Music.
  - Barack Obama, Bill Clinton, George Bush may be in another “Government & Politics” list.
- From Seed Set of 55K labeled accounts to influencers.
  - 336K influencers — 6x.
- Only for influencers (users with >= 10K followers)
- One Known-For label for an influencer.

Learning Interested-In Labels



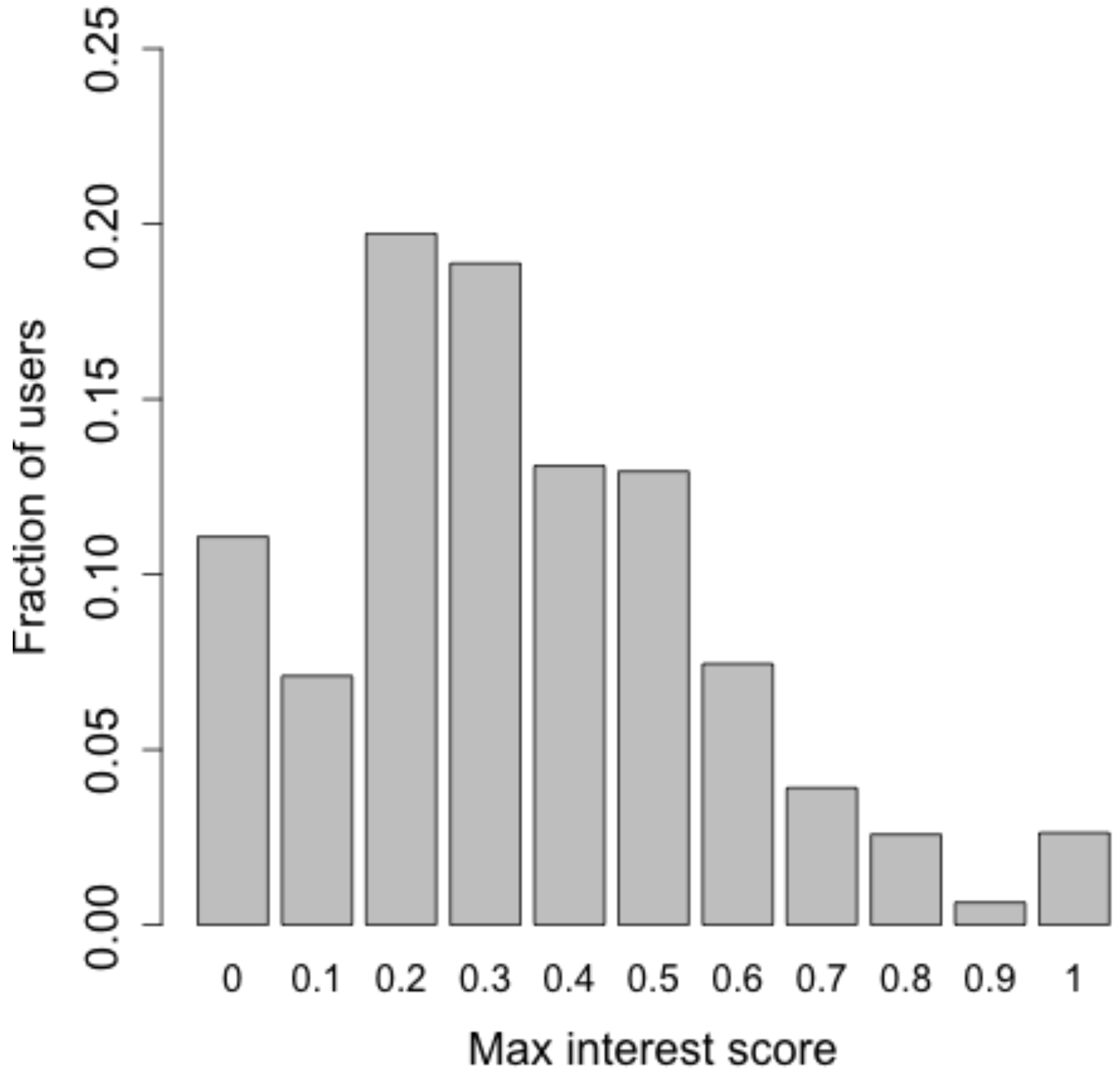
Interest_score(Emily, Music) : 0.33
Interest_score(Emily, News) : 0.67

- Via social graph — from influencers to everyone.
- Can be several Interested-In labels for a user.
- Sum of interest scores for a user = 1.

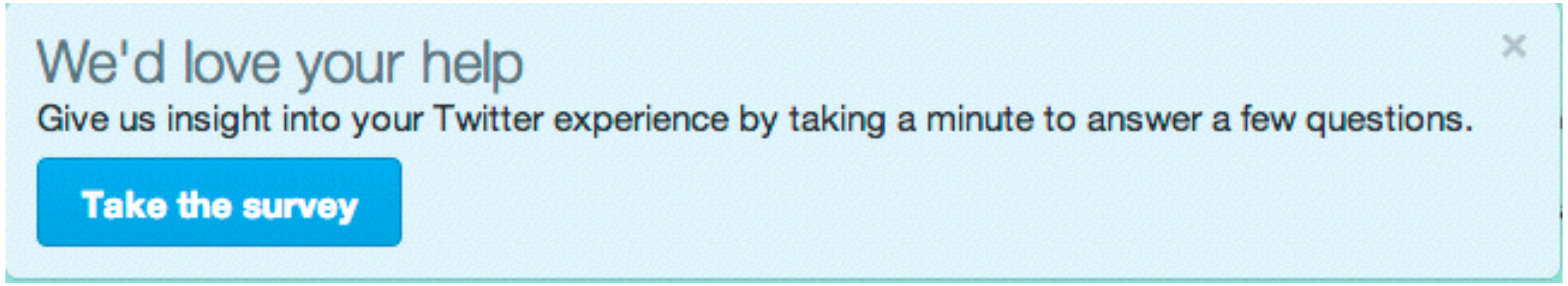
Further Improvements

- Issue of overfitting:
  - In cases when a user follows only one influencer.
- Issue of limited coverage:
  - This method provides us the coverage of 78%.
- Solution:
  - 2-hop random walk instead of 1-hop.
  - Performed for users who are
    - not covered in the 1-hop random walk.
    - may lie in overfitting case.
- Coverage increases to 88%.

Evaluation: Coverage



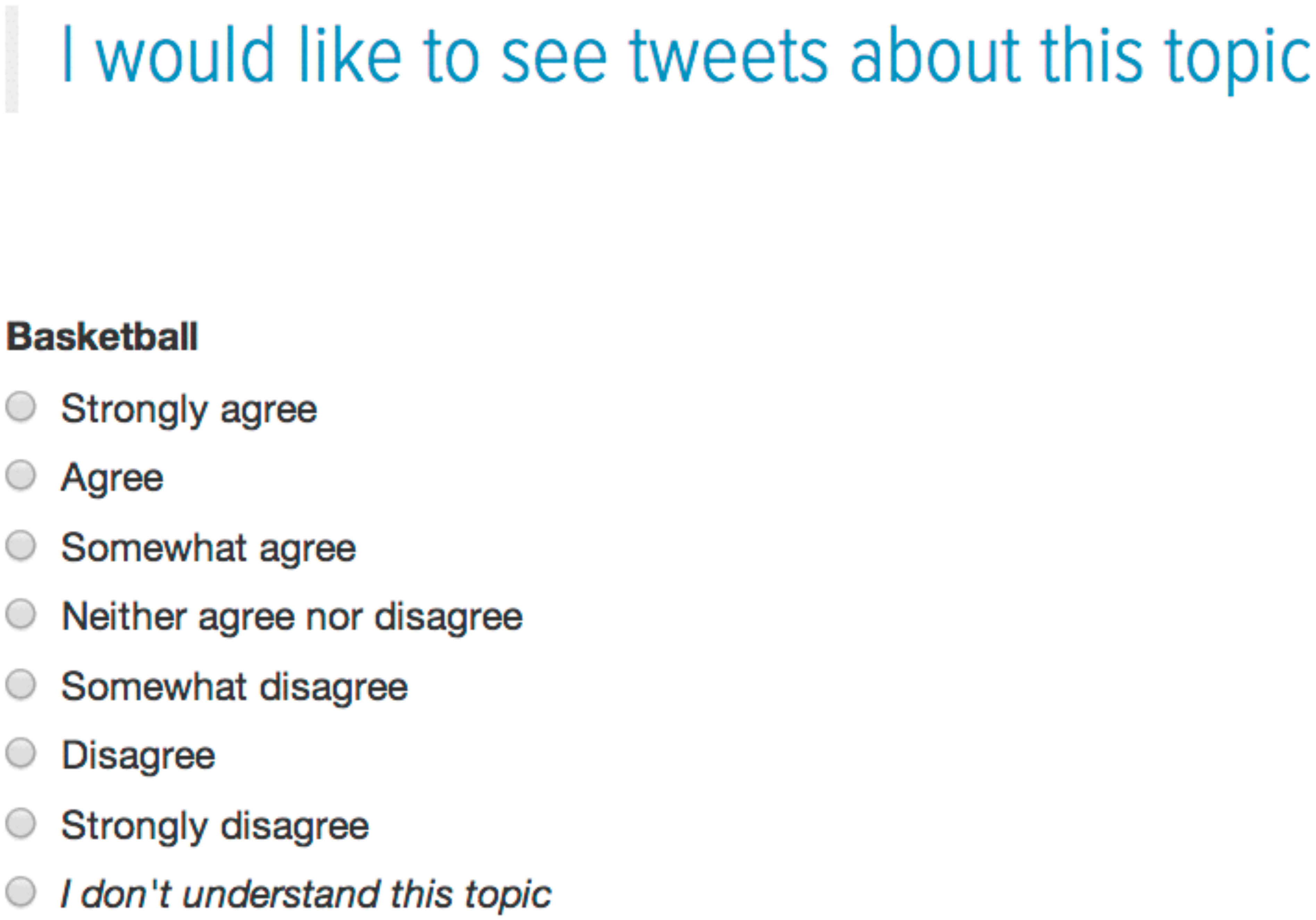
Evaluation: User interest surveys



We'd love your help

Give us insight into your Twitter experience by taking a minute to answer a few questions.

Evaluation: User interest surveys



> I would like to see tweets about this topic

Basketball

○ Strongly agree

○ Agree

○ Somewhat agree

○ Neither agree nor disagree

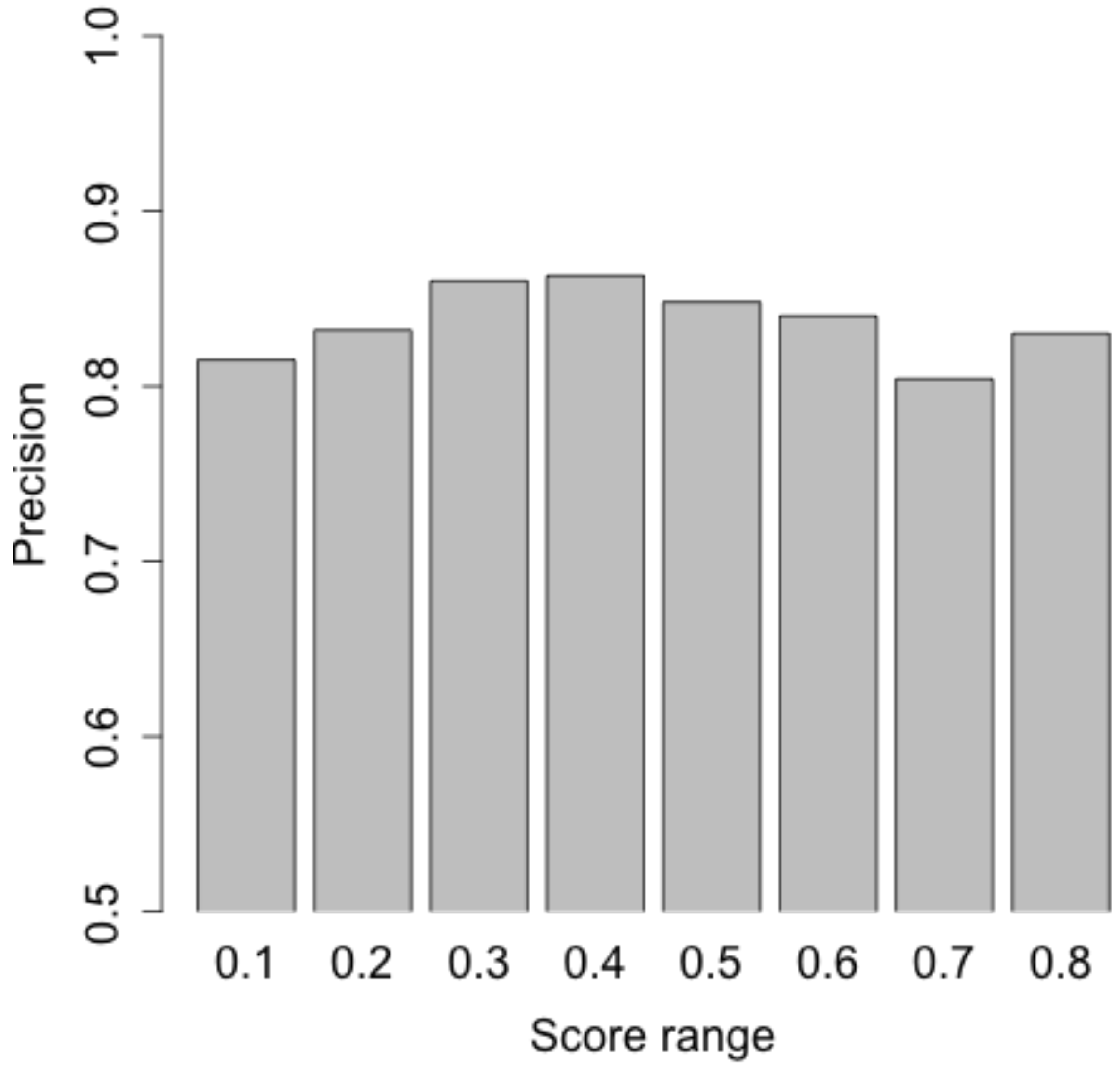
○ Somewhat disagree

○ Disagree

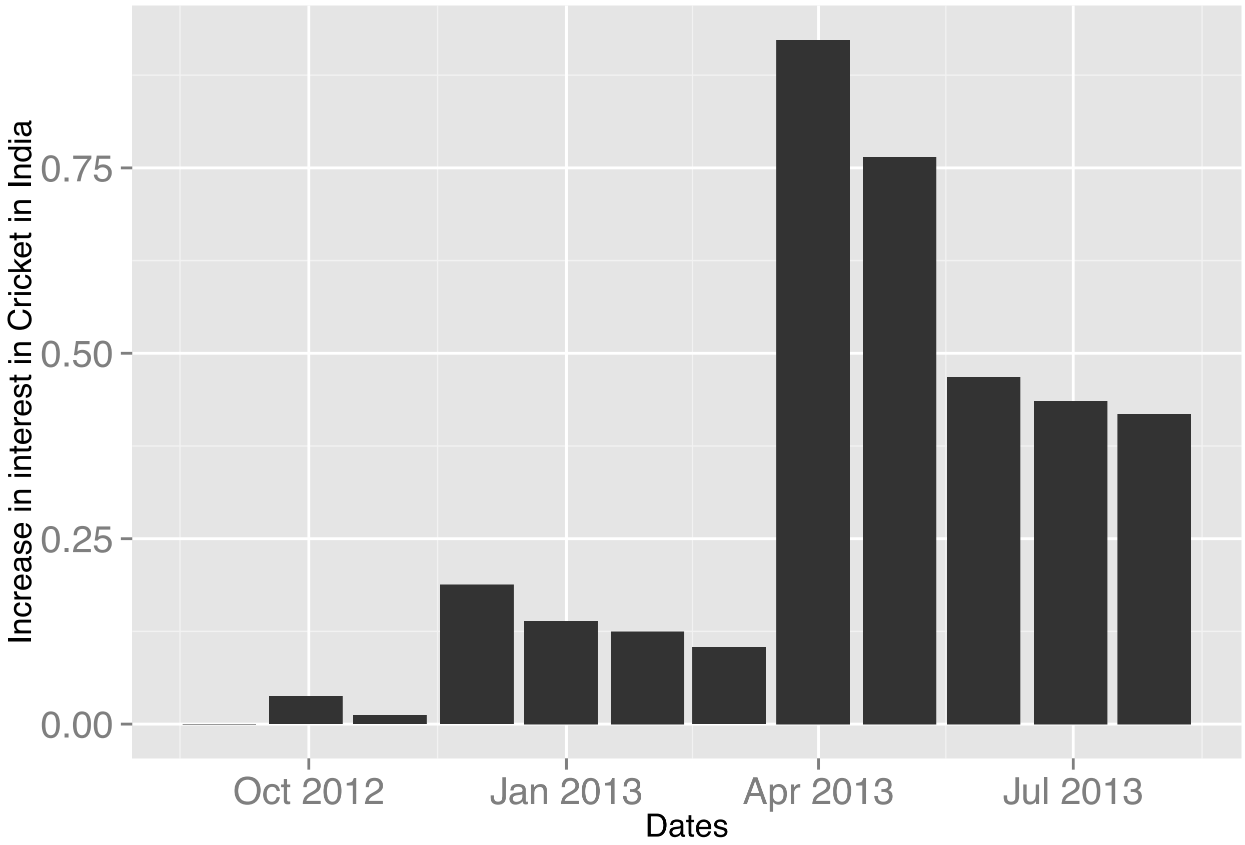
○ Strongly disagree

○ *I don't understand this topic*

Evaluation: Precision



Case Study: IPL Cricket Season



Summary

- Mine user interests via seed set expansion
  - First use lists to expand known for labels
  - Then use follow graph to infer interests

- Several wins over purely text based methods
  - Avoid inherent difficulties in language specific methods and internationalization
  - Learn about users even if they don’t engage with tweets much

- Evaluation & case study
  - High coverage (88% worldwide) at a reasonable precision (> 80%)
    - Can be used with several other signals to achieve higher precision, if desired
  - Consumer insights - IPL Cricket
    - How much do events affect expression of interests on Twitter?