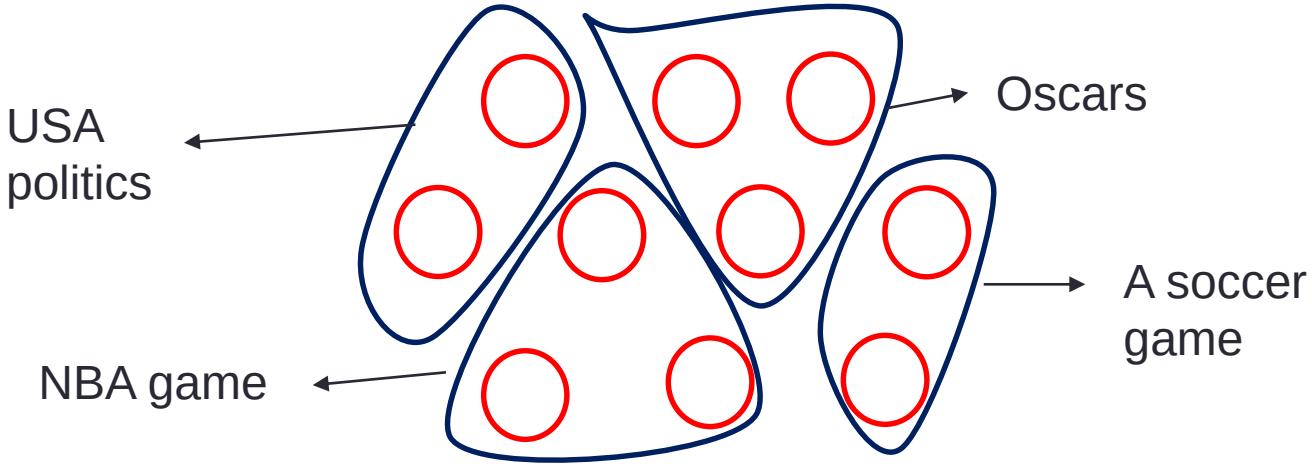
PARTITIONING AND RANKING TAGGED DATA SOURCES

Milad Eftekhar, Nick Koudas

University of Toronto

Introduction

- Explosion of data through social activities
  - More than a billion pieces of content on facebook per day
  - Hundreds of millions of blog post daily
  - Hundreds of millions of tweets every day



- Management
  - Organizing data
  - Group related content
  - Understanding trending topics and discussions

Introduction (Cont'd)

"The film receiving the most nominations was Lincoln with twelve #politics #usa #oscars2013"

"Argo won the #bestpicture in #Oscars2013"

"Oscars #redcarpet The Best Moments You Didn't See …"

"Who brought home the award for best dressed at the #Oscars? #redcarpet"

Introduction (Cont'd)

"The film receiving the most nominations was Lincoln with twelve #politics #usa #oscars2013"

"Argo won the #bestpicture in #Oscars2013"

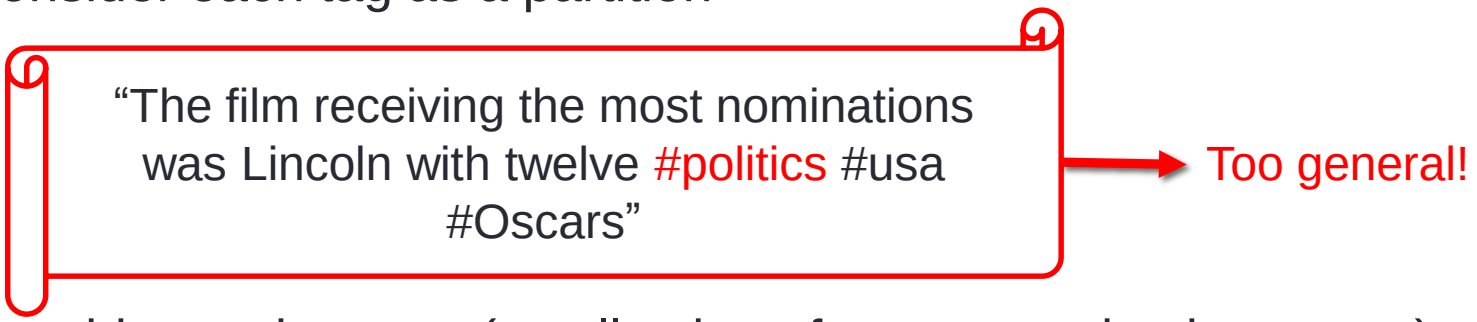
"Oscars #redcarpet The Best Moments You Didn't See …"

"Who brought home the award for best dressed at the #Oscars? #redcarpet"

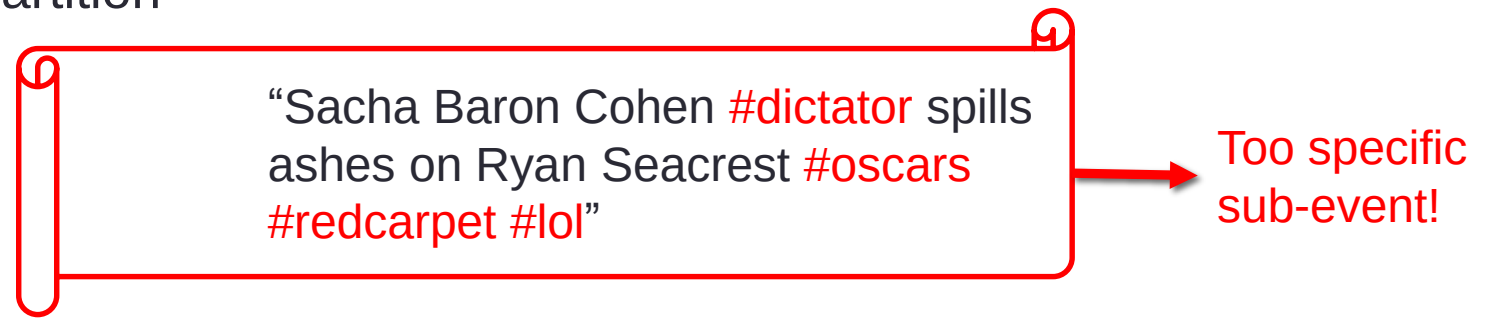
- Human-generated tags
  - content discovery and description
- Goal: Utilize tags to automatically organize and partition data

Simple approaches fail

- Naïve approaches:
  - Consider each tag as a partition



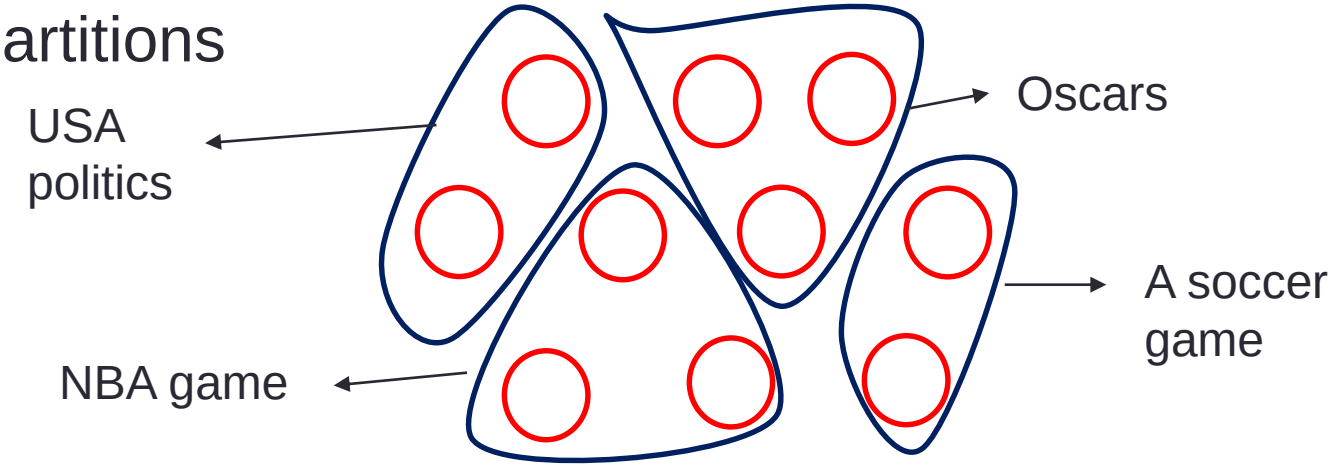
  - Consider each tagset (a collection of tags occurring in a tweet) as a partition



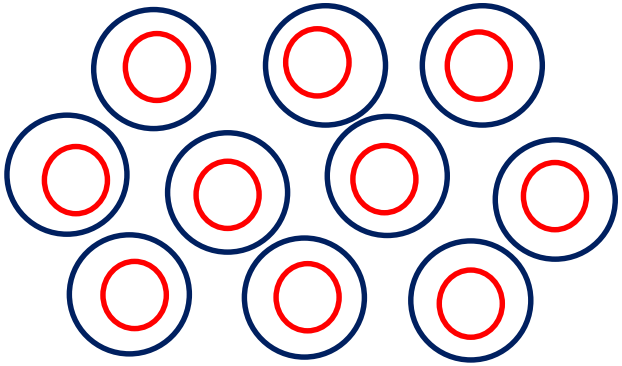
- Put all in a partition describing Oscars

Partitioning

- Segment the entire collection of tweets into disjoint partitions

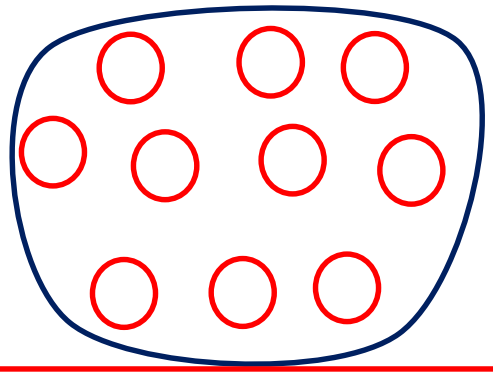


Extreme case 1 (one in one)



Goal: large size for each partition

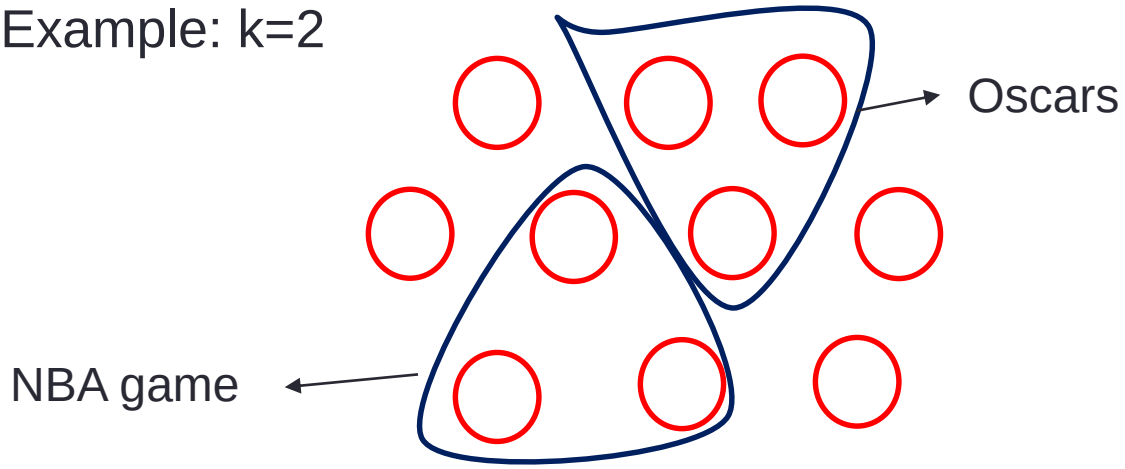
Extreme case 2 (all in one)



Goal: small number of distinct tagsets in a partition

Ranking

- Identify “top”-k partitions fast
  - Example: k=2



- A weight function that depends on
  - Size of the partition
  - Number of distinct tagsets

Problems Studied in this paper

Partitioning Problem 1 (MSP):

- Constraint: $size \geq c$
  - Do not create small partitions
- Objective: Maximize number of partitions
  - Prevent merging unrelated tweets: few tagsets

Ranking Problem 1 (MST):

- Objective: k partitions with the max aggregate weight

Partitioning Problem 2 (MMP):

- Constraint: number of partitions = k
- Objective: Min size is maximized
  - Balanced partitions in size

Ranking Problem 2 (MWT):

- Objective: k partitions such that the min weight is maximized

What you will learn if you read the paper!

Partitioning Problem 1 (MSP):
- NP-complete
- AMSP: a 2-factor approximation algorithm
  - Time: $\theta(l)$
    - l: number of distinct tagsets

Ranking Problem 1 (MST):
- Problem in P
- MST: An optimal algorithm
  - Time: $\theta(l)$

Partitioning Problem 2 (MMP):
- NP-complete
- MMP: a heuristic algorithm
  - Time: $\theta(kl)$
    - k: number of partitions

Ranking Problem 2 (MWT):
- NP-complete
- AMWT: approximation algorithm
  - Time: $\theta(k^2 l)$
- HMWT: heuristic algorithm
  - Time: $\theta(kl)$

Experiments

- 10 days of Twitter fire hose
- 250-300 million tweets daily
- 13% of tweets are tagged

Time

Partitioning Problem 1 (MSP):
- 9.3 seconds

Ranking Problem 1 (MST):
- 9.2 seconds

Partitioning Problem 2 (MMP):
- 18.7 minutes

Ranking Problem 2 (MWT):
- AMWT
  - 2.9 minutes
- HMWT
  - 32.9 seconds

100 partitions

Qualitative results (Feb 26, 2012)

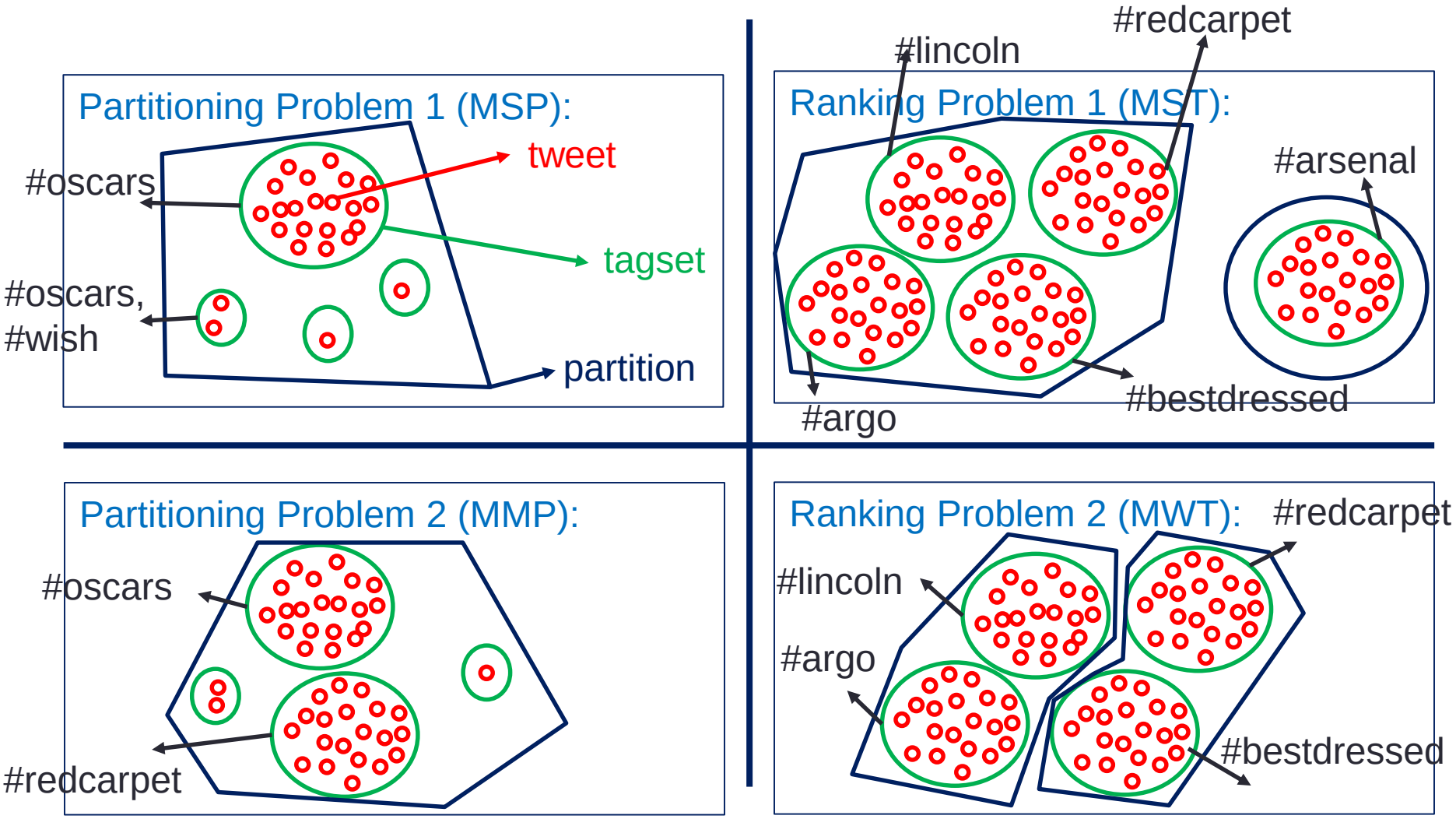
Partitioning Problem 1 (MSP):
- #OSCARS
- #ALLSTARGAME
- #YNWA
- #TEAMEAST
- #ARSENAL
- #SONGFESTIVAL
- …

Ranking Problem 1 (MST):
- #OSCARS
- #YNWA
- #TEAMEAST
- #SONGFESTIVAL
- #ARSENAL
- #HALAMADRID
- …

Partitioning Problem 2 (MMP):
- #OSCARS
- #ALLSTARGAME
- #LFC
- #ARSENAL
- #YNWA
- #SONGFESTIVAL
- …

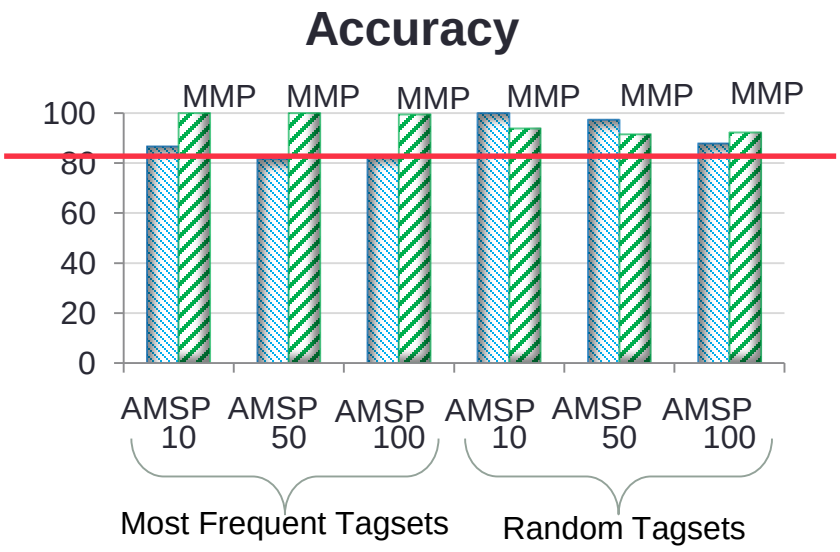
Ranking Problem 2 (MWT):
- #OSCARS
- #ALLSTARGAME
- #LFC
- #ARSENAL
- #YNWA
- #TEAMEAST
- …

Qualitative results' differences

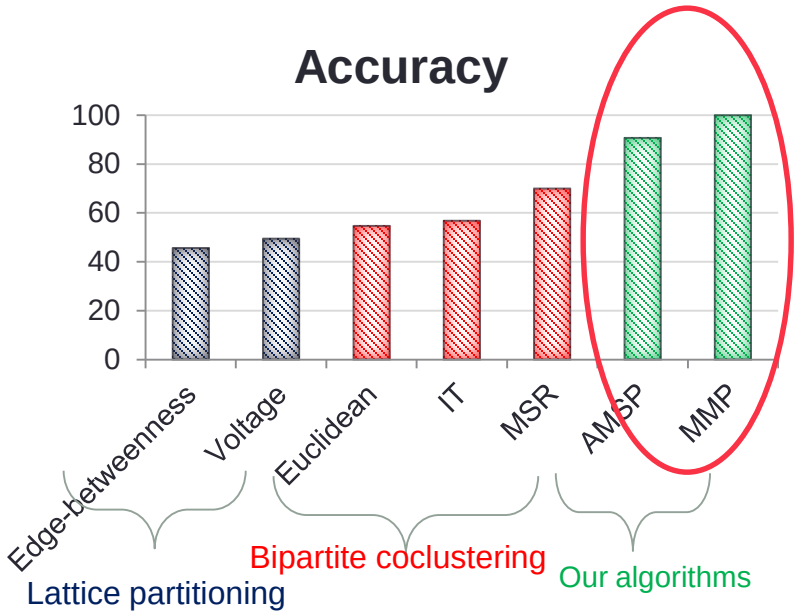


Comparison with baselines

- Baselines: clustering algorithms, co-clustering algorithms
- Datasets with known partitions



Partitioning quality



Baseline comparison

Conclusion and Future Works

- Partitioning and Ranking problems
- Top partitions represent Popular events
- Our algorithms result in up to 55% improvement in accuracy over baselines

- Better approximation bounds
- Taking additional aspects of the problem into account such as time, location, social ties

Thanks!

Question?