Voting, Algorithms, and Complexity

Palash Dey
Supervisors: **Prof. Y. Narahari** and **Prof. Arnab Bhattacharyya**

Department of Computer Science and Automation
Indian Institute of Science, Bangalore

April 23, 2016

Talk Overview

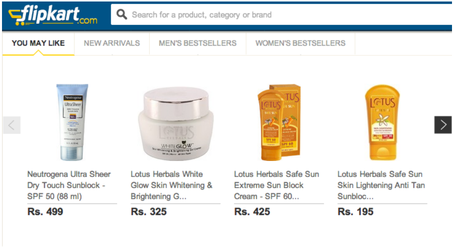
Motivation

Framework

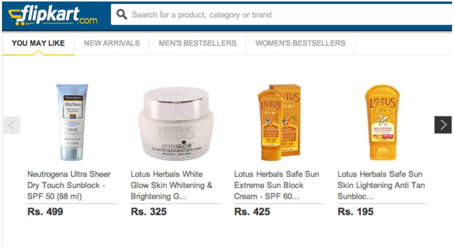
Our Contributions

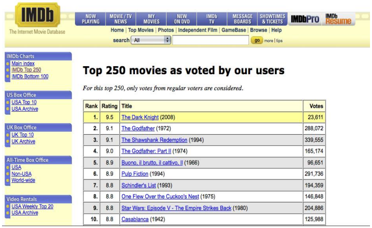
Voting Applications

Voting Applications



Voting Applications




IMDb
The Internet Movie Database

NOW PLAYING | MOVIE / TV NEWS | MY MOVIES | NEW ON DVD | IMDb TV | MESSAGE BOARDS | SHOWTIMES & TICKETS | IMDbPro | IMDb Resume

Home | Top Movies | Photos | Independent Film | GameBase | Browse | Help

search All go more | tips

IMDb Charts
- Main index
- IMDb Top 250
- IMDb Bottom 100

US Box Office
- USA Top 10
- USA Archive

UK Box Office
- UK Top 10
- UK Archive

All-Time Box Office
- USA
- Non-USA
- World-wide

Video Rentals
- USA Weekly Top 20
- USA Archive

Top 250 movies as voted by our users

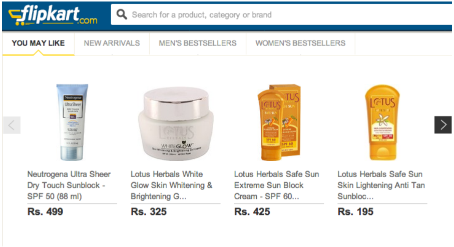
For this top 250, only votes from regular voters are considered.

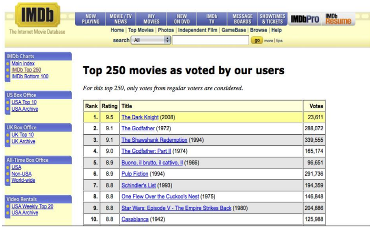
Rank	Rating	Title	Votes
1.	9.5	The Dark Knight (2008)	23,611
2.	9.1	The Godfather (1972)	288,072
3.	9.1	The Shawshank Redemption (1994)	339,555
4.	9.0	The Godfather: Part II (1974)	165,174
5.	8.9	Buono, il brutto, il cattivo, Il (1966)	96,651
6.	8.9	Pulp Fiction (1994)	291,736
7.	8.8	Schindler's List (1993)	194,359
8.	8.8	One Flew Over the Cuckoo's Nest (1975)	146,848
9.	8.8	Star Wars: Episode V - The Empire Strikes Back (1980)	204,886
10.	8.8	Casablanca (1942)	125,988

Voting Applications

Google

PageRank





Top 250 movies as voted by our users

For this top 250, only votes from regular voters are considered.

Rank	Rating	Title	Votes
1.	9.5	The Dark Knight (2008)	23,611
2.	9.1	The Godfather (1972)	288,072
3.	9.1	The Shawshank Redemption (1994)	339,555
4.	9.0	The Godfather: Part II (1974)	165,174
5.	8.9	Buono, il brutto, il cattivo, Il (1966)	96,651
6.	8.9	Pulp Fiction (1994)	291,736
7.	8.8	Schindler's List (1993)	194,359
8.	8.8	One Flew Over the Cuckoo's Nest (1975)	146,848
9.	8.8	Star Wars: Episode V - The Empire Strikes Back (1980)	204,886
10.	8.8	Casablanca (1942)	125,988

WIKIPEDIA

The Free Encyclopedia

Talk Overview

Motivation

Framework

Our Contributions

Voting Setting

- A set C of m candidates
- A set V of n votes
- Vote - a complete order over C
- Voting rule - $r : \mathcal{L}(C)^n \longrightarrow C$

Example

- $C = \{x, y, z\}$
- Votes
  - ✓ Vote 1: $x > y > z$
  - ✓ Vote 2: $z > y > x$
  - ✓ Vote 3: $x > z > y$

Plurality rule: winner is candidate with most top positions

Plurality winner: x

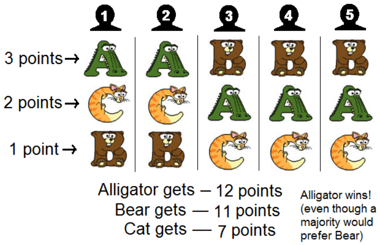
Example: Scoring Rules

Scoring Rule

- Score vector: $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$
- A vote $x_1 > x_2 > \cdots > x_m \Rightarrow x_i$ gets score α_i
- Winner: candidate with the highest score

Important Special Cases

- Plurality: $(1, 0, \cdots, 0)$
- Veto: $(0, \cdots, 0, -1)$
- Borda: $(m-1, m-2, \cdots, 0)$

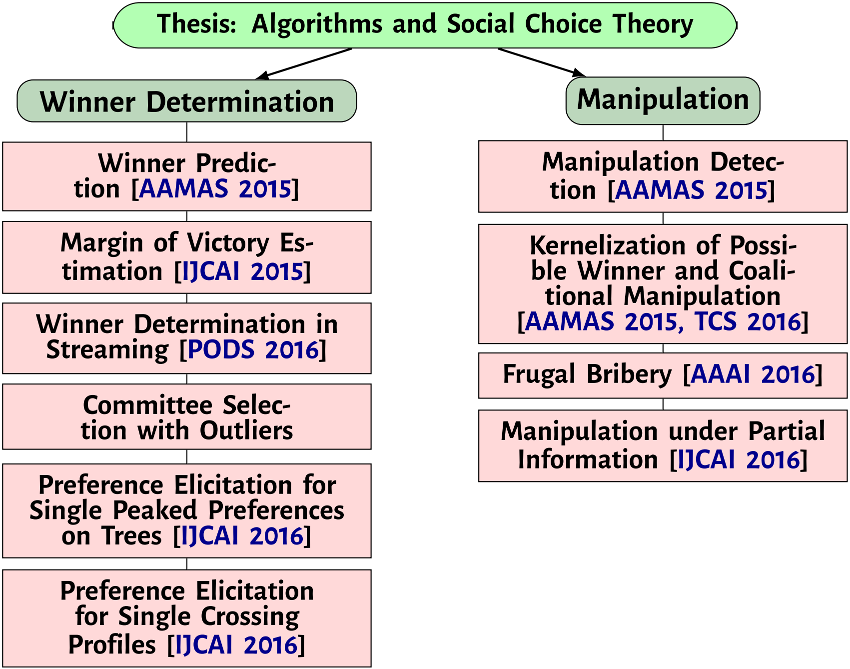


Talk Overview

Motivation

Framework

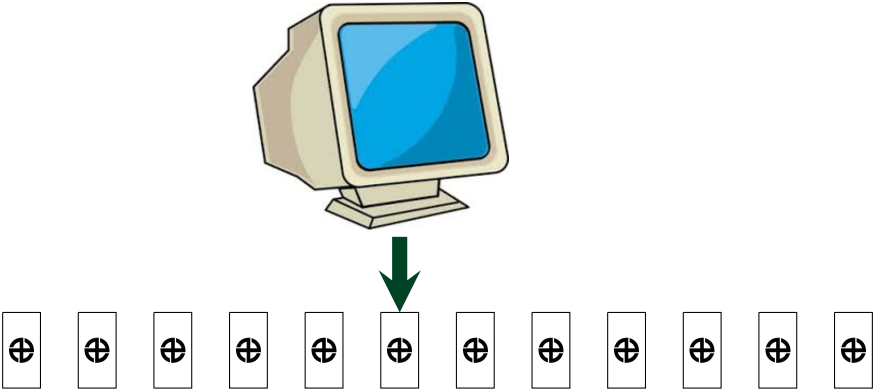
Our Contributions


Thesis: Algorithms and Social Choice Theory
Winner Determination
Winner Prediction [AAMAS 2015]
Margin of Victory Estimation [IJCAI 2015]
Winner Determination in Streaming [PODS 2016]
Committee Selection with Outliers
Preference Elicitation for Single Peaked Preferences on Trees [IJCAI 2016]
Preference Elicitation for Single Crossing Profiles [IJCAI 2016]
Manipulation
Manipulation Detection [AAMAS 2015]
Kernelization of Possible Winner and Coalitional Manipulation [AAMAS 2015, TCS 2016]
Frugal Bribery [AAAI 2016]
Manipulation under Partial Information [IJCAI 2016]

Winner Determination in a Stream of Votes

To appear: ACM SIGMOD conference on Principles of DB Systems (PODS-16)

Data Streams



Data Stream

- Suitable model for many large sources of data
  - ✓ Stream of network packets
  - ✓ Sensor networks
- Impractical and undesirable to store and process the entire data exactly
  - ✓ Instead design algorithms to find approximate solutions
  - ✓ Quickly build summary with one pass over data
- Active area of research for last 15 years, history goes back 35 years

(ε, φ)-Plurality

Let $0 < \varepsilon < \varphi < 1$ and f_i be the plurality score of candidate i

Problem Definition

Find a set S of candidates with the following property:

- S contains every candidate i with $f_i > \varphi n$
- S contains no candidates j with $f_j < (\varphi - \varepsilon)n$

Moreover, for every candidate $i \in S$, output an estimate $\tilde{f}_i$ such that $|f_i - \tilde{f}_i| \leqslant \varepsilon n$

(ε, φ)-Plurality

Let $0 < \varepsilon < \varphi < 1$ and f_i be the plurality score of candidate i

Output all popular candidates

Problem Definition

Find a set S of candidates with the following property:

- S contains every candidate i with $f_i > \varphi n$
- S contains no candidates j with $f_j < (\varphi - \varepsilon)n$

Moreover, for every candidate $i \in S$, output an estimate $\tilde{f}_i$ such that $|f_i - \tilde{f}_i| \leqslant \varepsilon n$

Don't output any unpopular candidate

Estimate plurality score of popular candidates

(ε, φ)-Plurality

Let $0 < \varepsilon < \varphi < 1$ and f_i be the plurality score of candidate i

Output all popular candidates

Problem Definition

Find a set S of candidates with the following property:

- S contains every candidate i with $f_i > \varphi n$
- S contains no candidates j with $f_j < (\varphi - \varepsilon)n$

Moreover, for every candidate $i \in S$, output an estimate $\tilde{f}_i$ such that $|f_i - \tilde{f}_i| \leqslant \varepsilon n$

Don't output any unpopular candidate

Estimate plurality score of popular candidates

This problem is popularly known as **(ε, φ)-Heavy hitters** in the streaming literature

Main Theorem

We show that space complexity of (ε, φ)-Plurality is* :

$$\boxed{\Theta\left(\frac{1}{\varepsilon}\log\frac{1}{\varphi} + \frac{1}{\varphi}\log n + \log\log m\right)}$$

with $O(1)$ worst case update and query response times. Our algorithm is randomized

*If $n \geqslant \left(\frac{1}{\varepsilon}\right)^{1.0001}$

Main Theorem

We show that space complexity of (ε, φ)-Plurality is* :

$$\boxed{\Theta\left(\frac{1}{\varepsilon}\log\frac{1}{\varphi} + \frac{1}{\varphi}\log n + \log\log m\right)}$$

with $O(1)$ worst case update and query response times. Our algorithm is randomized

- Resolves a 30 years old open question in data streaming literature
- Resolving this was mentioned as a key research challenge in IITK Workshop on Data Streams (2006)

*If $n \geqslant \left(\frac{1}{\varepsilon}\right)^{1.0001}$

Other Results

ε-veto

$$O\left(\frac{1}{\varepsilon}\log\log\frac{1}{\varepsilon}+\log\log m\right), \Omega\left(\frac{1}{\varepsilon}+\log\log m\right)$$

ε-Borda

$$\Theta\left(n(\log\frac{1}{\varepsilon}+\log n)+\log\log m\right)$$

ε-maximin

$$O\left(\frac{n}{\varepsilon^2}\log^2 n+\log\log m\right), \Omega\left(\frac{n}{\varepsilon^2}+\log\log m\right)$$

Other Results

Optimal upto $O(\log\log\varepsilon^{-1})$

ε-veto

$$O\left(\frac{1}{\varepsilon}\log\log\frac{1}{\varepsilon}+\log\log m\right),\Omega\left(\frac{1}{\varepsilon}+\log\log m\right)$$

Optimal upto $O(1)$

ε-Borda

$$\Theta\left(n(\log\frac{1}{\varepsilon}+\log n)+\log\log m\right)$$

Optimal upto $O(\log^2 n)$

ε-maximin

$$O\left(\frac{n}{\varepsilon^2}\log^2 n+\log\log m\right),\Omega\left(\frac{n}{\varepsilon^2}+\log\log m\right)$$

- I acknowledge all the useful discussions and continuous support from **Prof. Y. Narahari**, **Prof. Arnab Bhattacharyya**, **Prof. Neeldhara Misra** and all my co-authors.
- I gratefully acknowledge financial support from MHRD and Google India.