Voting Mechanism and Consensus Involving

Shirui Zhou, Sharon Chuang, etc

PPOL564: Final Class Status Update. 11.30.2022

Outline

- Motivation
- Research Questions
- Data
- Methods
- Results thus far
- Limitations/Next Steps

Motivation

- Problems
  - 1.‘Horse race’ problem and misinterpretation in polls and probabilistic modeling
  - 2. Echo chamber and polarization
  - 2. No alternative voice in some authoritarian countries
- Solution
  - Voting mechanism served as decentralized opinion aggregation platform
  - Vote on the ”comments”by participants
  - Review and understand the most trivial concern of the other side

Research Questions

- Observation
  - Static: Opinion landscape of the participants and representativeness of comments
  - Dynamic: Is a consensus forming in the voting?
- Activism
  - Would prioritizing the comments from the opposite 'echo chamber' make a difference?

Data Sources

- Polis open source data
  - Experimentation of voting mechanism (45s test)
  - Uber Issue: Should Uber be regulated in Taiwan
  - Main data set: participants vote csv, comment csv

Experimentation

Method: Data Acquisition

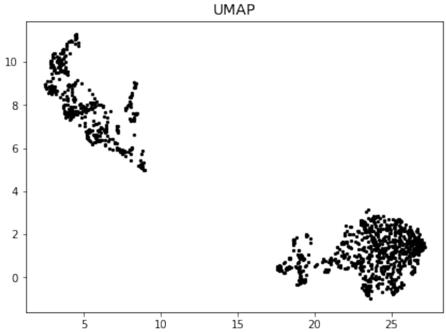
- participants-votes.csv
  - meta-data: participant, group-id, n-comments, n-notes, n-disagree, n-agree
  - sparse matrix: participants (x-axis) vote on each comments (y-axis)
- comments.csv
  - variables: comment-id, author-id, moderated, comment-body, timestamp

Method: Data Cleaning

```
def count_finite(row):
    finite = np.isfinite(row[val_fields])
    return sum(finite) # count number of True values in `finite`

def select_rows(df, threshold):
    number_of_votes = df.apply(count_finite, axis=1)
    valid = number_of_votes >= threshold
    return df[valid]
df_votes = select_rows(df, 7)

## remove statements (columns) which were moderated out
statements_all_in = sorted(list(df_comments.loc[df_comments["moderated"] > 0].index.array), key = int)
```

Methods: Algorithms

- process flow
- Dimensionality reduction
  - PCA
  - UMAP
- Clustering
  - Kmeans



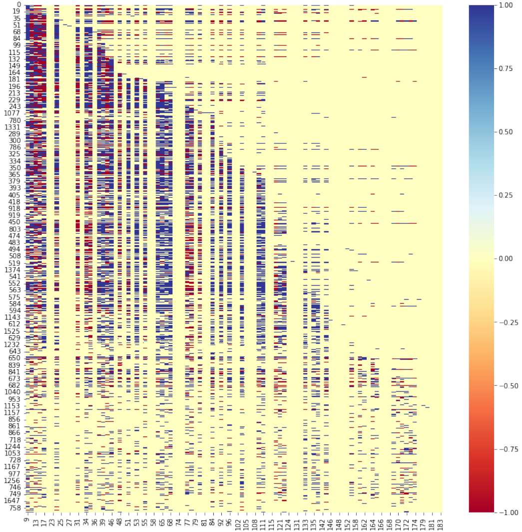
Methods: Dimensionality reduction

- PCA
  - enables us to visualize participants in relation to each other within the opinion landscape
  - Participants are closer together in this landscape if they tend to agree. And further apart if they tend to disagree.
- UMAP
  - Participants are closer together in this landscape if they tend to agree. And further apart if they tend to disagree.

Describing the analytic sample

Name	Estimand
Dimensions of pre-processed matrix	(1921, 203)
Dimensions of post-processed matrix	(1269, 198)
Total number of possible votes	251262
Total number of agrees	30237
Total number of disagrees	11661
Total without vote	208097
Percent sparse	0.8282072100039003

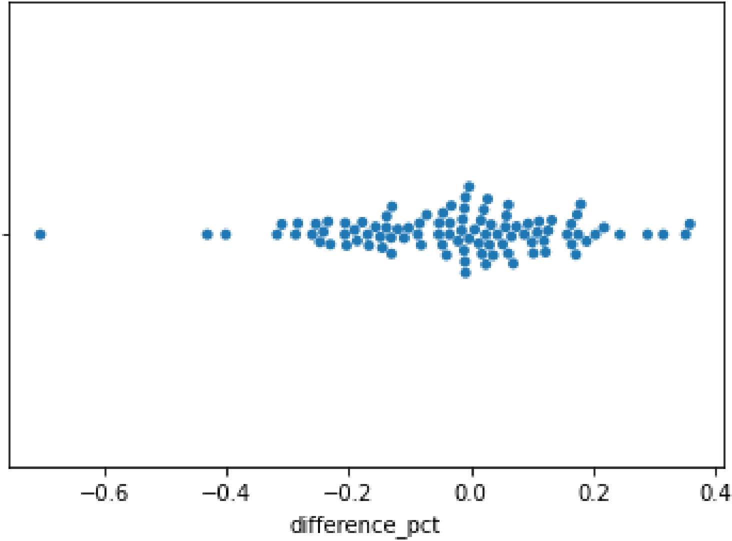
Full participants * comments matrix



Method: in words

- Umap and K-means
  - detect two communities
- construct two opinion groups
  - the divisiveness, consensus comments and dispute comments

Results: Divisiveness of Comments



- statement on the left vote on the same

Results: Consensus Comments

	comment-id	author-id	a_agree	a_disagree	a_support_share	b_agree	b_disagree	b_support_share	comment-body
107	94	0	122	46	0.452381	163	61	0.455357	我覺得要提前考慮電腦自動駕駛的情況，包括現代大眾系統是否還有興建必要，以及以後計程車司機的就...
165	141	258	91	11	0.784314	101	12	0.787611	我覺得繳稅是一個企業在台灣經營的義務，Uber或其新創模式再麼優秀，在台灣的所得都應該要盡到...
95	150	2938	73	11	0.738095	71	11	0.731707	應修改讓自用車的乘客保險也能保障乘客的權益或由UBER統一投保
49	32	0	194	106	0.293333	222	119	0.302053	我覺得 UberX 目前無法幫乘客保意外險，讓我感到沒有保障。
65	50	0	118	126	−0.032787	142	149	−0.024055	我覺得大眾運輸工具普及後，營業車的空車率已不斷攀升，開放自用車載客不會擴大需求，只會讓計程車...

Abbildung 1: the comment from

Results: Dispute Comments

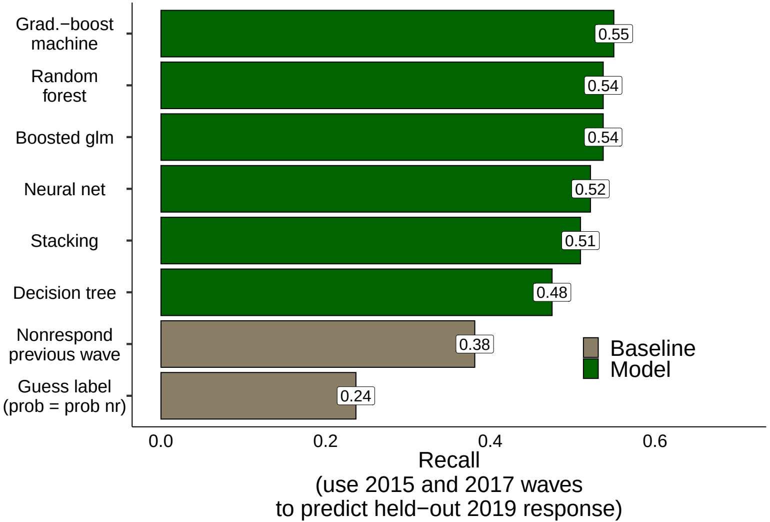
	comment-id	author-id	a_agree	a_disagree	a_support_share	b_agree	b_disagree	b_support_share	comment-body
179	13	0	85	245	−0.484848	195	125	0.218750	我覺得計程車身一定要塗裝成黃色的，和其他車輛顏色不同。
51	37	0	178	92	0.318519	286	41	0.749235	我覺得在徵集意見前，各級相關政府單位都應該先明確表示立場。
66	51	476	172	60	0.482759	269	30	0.799331	我覺得任何的創新服務以達營利目的的確是社會進步的重要過程，但必須完全在法律的規範下合法經營，...
108	96	0	128	26	0.662338	240	24	0.818182	我覺得所有營業車輛都應該一律採取乘客評分機制，而不是只靠政府核發營業許可。
47	30	0	48	237	−0.663158	125	214	−0.262537	我覺得 UberX 現行不法行為應盡一切努力使其停業，不需要國人表示意見。

Next step

- Dynamic analysis
  - time-series hypothesis testing
  - 4 sub-graphs, calculate the social distance of opinion group A and B
- Limitation
  - clustering method produce different result
  - testing leiden-graph method

Here are some slides that have examples of different syntax you may find helpful (not required to do things like a tikz diagram)

Example of loading a figure you've uploaded (can be pdf or png)



Example of a table

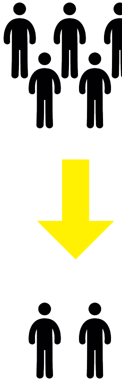
Name	Estimand	Estimator	Why?
Divergence between sample and population	$(\bar{X} - E[\hat{\hat{X}} \mid T = 1]) - (\bar{X} - E[\hat{\hat{X}} \mid T = 0])$	Regress distance between population mean ($\bar{X}$) and sample mean ($\hat{X}$) on treatment indicator	Measures whether treatment produces sample quantities closer to population mean
Response rate	$\frac{1}{n} \sum_{\{i:S_i=1\}} Y_i(T = 1) - Y_i(T = 0)$	Regress response on treatment	In combination with bias measure helps us understand whether we increase both response rate and decrease non-response bias, or only increase response rate with no reductions in bias

Example of inserting code snippet using fragile environment and listings

```
def clean_yelp_json(one_biz):

    ## restrict to str cols
    d_str = {key:value for key, value in one_biz.items()
   if type(value) == str}

    df_str = pd.DataFrame(d_str, index = [d_str['id']])
    return(df_str)

yelp_stronly = [clean_yelp_json(one_b)
        for one_b in yelp_genjson['businesses']]
yelp_stronly_df = pd.concat(yelp_stronly)
```

Example of splitting slide using minipage and tikz diagram

Random targeting:

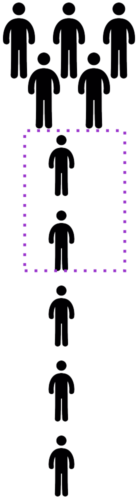


Potential respondents

Randomly sample

Provide additional incentives

Risk-based targeting



Potential respondents

1. Define estimand to predict (e.g., nonresponse risk; case importance)
2. Predict that label
3. Rank by $\hat{y}$
4. Provide additional incentives based on risk/importance

Another example tikz diagram

Use 80% sample ($N = 67,136$) to train model
with 162 features
to predict 2019 nonresponse;
select hyp. via CV

Select model that
optimizes recall (GBM)

Estimation data containing all sampled units' 2015 and 2017 features (aggregated so one prediction per unit)

Validate on 20% held out set ($N = 16,783$)

Fit top model to data containing training units' 2015, 2017, <u>and 2019</u> features (aggregated so one prediction per unit)

Use that model to predict 2021 nonresponse