

Voting Mechanism and Consensus Involving

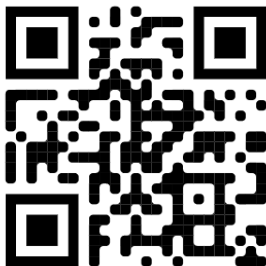
Shirui Zhou, Sharon Chuang, etc

PPOL564: Final Class Status Update. 11.30.2022

Lets Do a Experimentation Before We Start

Polis

- ▶ Link: <https://pol.is/2hkcy8chj>
- ▶ Topic: **China Zero Covid Protests**
- ▶
 - ▶ User can agree / disagree on different comments
 - ▶ Its **anonymous!!**.
 - ▶ Based in the US
 - ▶ → *youwon' tblacklist!!*



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

Prior Research

Polis

- ▶ a real-time system for gathering, analyzing and understanding what large groups of people think in their own words, enabled by advanced statistics and machine learning.
- ▶ method:
 1. Read about the polis opinion matrix
 2. Dimensionality reduction
 3. Clustering



China Zero Covid Protests

"In memory of the people who died in the fire in Urumuqi in November 24th, and the ones who are silenced, under arrest, and have their freedom taken away by the Chinese government with their brutal measure against covid-19 pandemic.

Understand the Protests in China

- The Toll of 'Zero Covid': The protests against China's strict pandemic policy come after President Xi Jinping's unbending approach hurt businesses and strangled growth.
- At a 'Tipping Point': For the protesters, public dissent was unimaginable until days ago. Our columnist asked young people what led them to take the risk.
- The Economic Fallout: The growing unrest in the world's biggest manufacturing nation is injecting a new element of uncertainty and instability into the global economy.
- Reasserting Control: The Communist Party is drawing on its decades-old policy of repression and surveillance — along with some new tactics — to quash the protests. (NewYork Times, 2022)

Welcome to a new kind of conversation — vote on other people's statements.

Anonymous wrote: 5 remaining

In my opinion, it's brave to stand out and voice the concern of people, but i don't to be the one who died for bravery

Agree
 Disagree
 Pass / Unsure

Are your perspectives or experiences missing from the conversation? If so, **add them** in the box below.

What makes a good statement?

- Stand alone idea
- Raise new perspectives, experiences or issues
- Clear & concise (limited to 140 characters)

Please remember, statements are displayed randomly and you are not replying directly to other participants' statements.

Share your perspective...

[Submit](#)

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

- ▶ Vtaiwan 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

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: Process flow

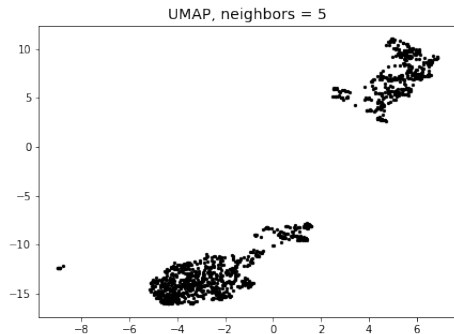
1. Data Clean
2. Dimensionality reduction:
 - ▶ PCA
 - ▶ UMAP
3. Cluster:
 - ▶ Kmean

Method: Data Cleaning

```
1 def count_finite(row):
2     finite = np.isfinite(row[val_fields])
3     return sum(finite) # count number of True values in `
4     finite `
5
6 def select_rows(df, threshold):
7     number_of_votes = df.apply(count_finite, axis=1)
8     valid = number_of_votes >= threshold
9     return df[valid]
10
11 df_votes = select_rows(df, 7)
12
13 ## remove statements (columns) which were moderated out
14 statements_all_in = sorted(list(df_comments.loc[
15     df_comments["moderated"] > 0].index.array), key = int
16 )
```

Methods: Algorithms

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



Methods: Dimension reduction

- ▶ 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.

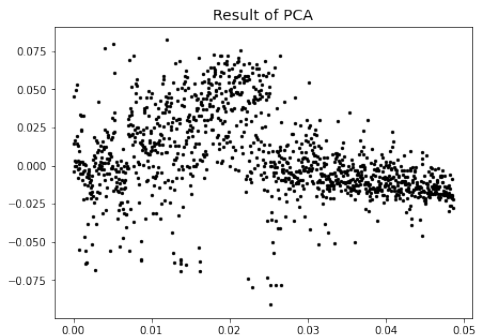


Abbildung 1: Visualize w/ PCA

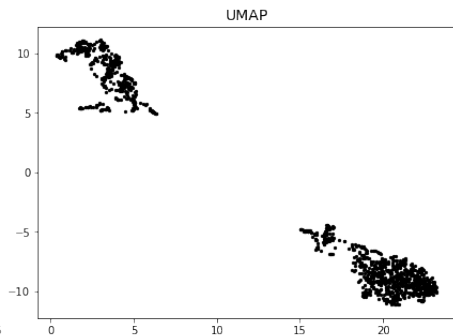


Abbildung 2: Visualize w/ UMAP

Methods: Dimension reduction

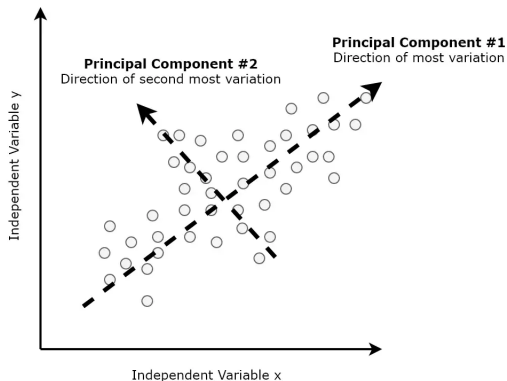
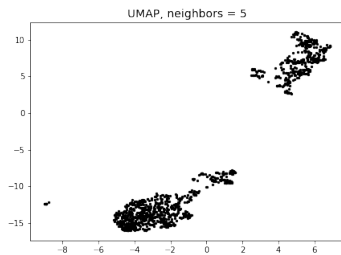
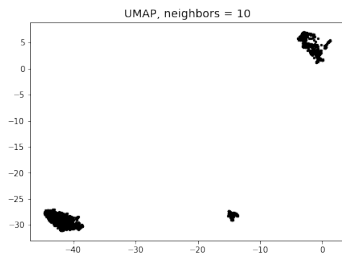


Abbildung 3: Source: Packt_Pub, via Hackernoon

► PCA

- effective for visualizing clusters or groups of data points and their relative proximities.
- Identifying the hyperplane which lies closest to the data and then

Methods: Dimension reduction



- ▶ UMAP (Uniform Manifold Approximation and Projection)
 - ▶ effective for visualizing clusters of data points.
 - ▶ nonlinear dimensionality reduction method
 - ▶ Scalability: can be applied directly to sparse matrices

Methods: K-means

► Algorithm

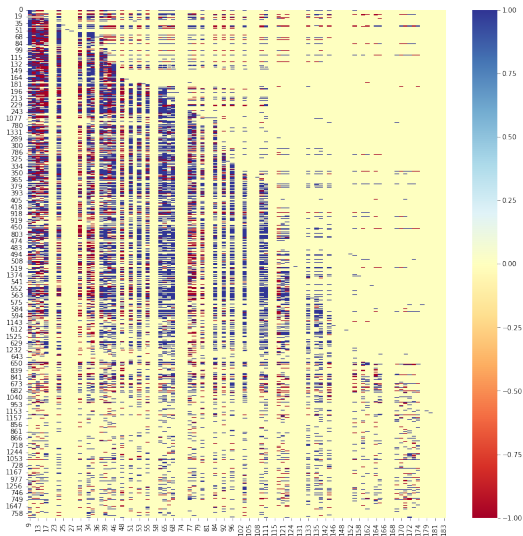
1. Randomly initialize k cluster means (here: $k = 2$)
2. Iterate:
3. Assign each object to the nearest cluster mean
4. Recompute cluster means
5. Stop when clustering converges

```
def kmean_get_grouped(embeds, k):  
    grouped_embed = KMeans(n_clusters=2, r  
    grouped_embed = grouped_embed.fit(embeds)  
    labs = grouped_embed.labels_  
  
    A_key = np.where(labs == 0 )[0]  
    B_key = np.where(labs == 1 )[0]  
    return A_key, B_key
```

Describing the analytic sample

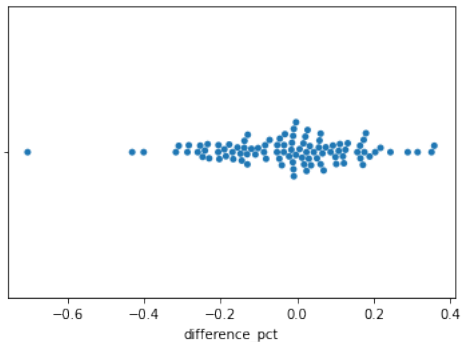
Name	Estimand
Dimensions of pre-processed matrix	(1921, 203)
Dimensions of post-processed matrix	(1269, 198)
Total number of possible votes	251262
Total number of agrees	30237
Total number of disagrees	11661
Total without vote	208097
Percent sparse	0.8282072100039003

Full participants * comments matrix



Results: Divisiveness of Comments

- ▶ How divisive was the conversation?
- ▶ Statement close to zero are voted on the same way - either both opinion groups agree or disagree
- ▶ Statement far from zero are divisive - participant were split between agreement and disagreement



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，以及以後計程車可權的就...
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或其新創模式再度優秀，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 目前無法駕駛客保意外險，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...

- ▶ min 'support share difference' on each comments
 - ▶ smaller difference in support share → *consensus*
- ▶ Consensus arguments
 - ▶ Main appeal of people (agreed by both groups)

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違章利目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	98	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。

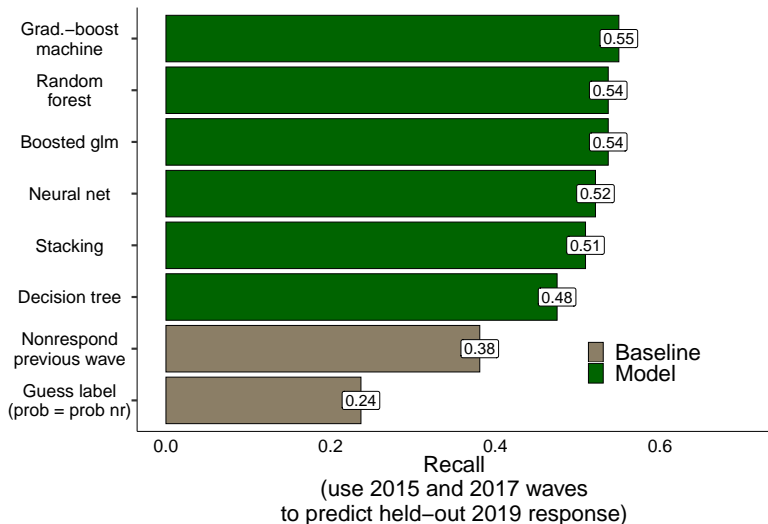
- ▶ max 'support share difference' on each comments
 - ▶ larger difference in support share dispute
- ▶ Dispute arguments
 - ▶ Need to be discussed and solved

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

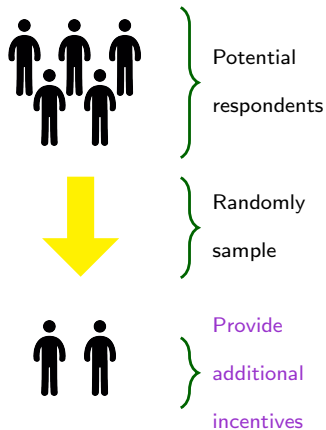
Name	Estimand	Estimator	Why?
Divergence between sample and population	$(\bar{X} - E[\hat{X} T = 1]) - (\bar{X} - E[\hat{X} T = 0])$	Regress distance between population mean (\bar{X}) and sample mean (\hat{X}) on treatment indicator	Measures whether treatment produces sample quantities closer to population 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

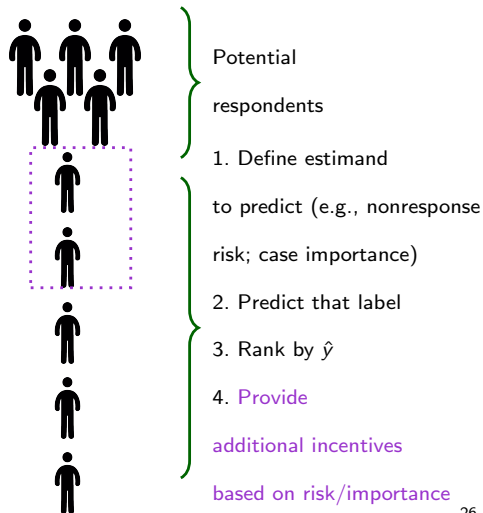
```
1 def clean_yelp_json(one_biz):
2
3     ## restrict to str cols
4     d_str = {key:value for key, value in one_biz.items()
5              if type(value) == str}
6
7     df_str = pd.DataFrame(d_str, index = [d_str['id']])
8     return(df_str)
9
10 yelp_stronly = [clean_yelp_json(one_b)
11                for one_b in yelp_genjson['businesses']]
12 yelp_stronly_df = pd.concat(yelp_stronly)
```

Example of splitting slide using minipage and tikz diagram

Random targeting:



Risk-based targeting



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 (aggrega-
ted so one prediction per unit)**

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

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

**Use that model
to predict 2021
nonresponse**