## Weighted sample networks

* **How will the networks properties change depending on an introduced bias?**

**We will generate a network with all the properties and use it as our representation of the population. We then take a sample from the population and see how the properties will change if we do not control for the different groups. We will use predefined groups to show the effects on polarisation.**

**What to do?**

* **Generate a Population (network, integrate groups, a bunch of different attributes) [check]**
* **At random select a subset of as “survey”**
* **Allow resampling and get the same size as before but better!**

->> **Results:** We can generate a synthetic data set with a certain division of groups. We also can randomly add attributes to participants, like age, etc. We get weighted subsample through an iterative proportional fitting.

- In the analysis, I suggest to construct a population in form of a survey, or data. From that, we can first apply our network method to generate a network and obtain network statistics. In a second step, draw a survey from the “population” a weighted and a random one. We calculate network statistics for these two samples as well. In a third step, we resample from the samples X times and can get a sub-samples [weighted (with/without replace) and unweighted versions]. We calculate the network statistics and compare them to the sample and to the ground truth.

**Intention:**

* Show how the network statistics change from weighted to unweighted
* Weaknesses especially in weighted data networks
* What is the effect on our analysis?
* What bias in the data is still getting the same results?
* Polarisation measure for the two groups?
* Generalisations
* How do you it generalising it?
* Parametric? Bootstrapping – coverage probability. Compare it to the population? How often to high?
* 4 groups 🡪 4 groups 🡪 get a lot of questions 🡪
* How does Bias influence our survey data?

**Problems:**

* Random sample is highly dependent on the first sample. Due to the randomness, it is still getting about the same proportions as the original data.
* What are the relevant network statistics? How can I compare the networks with each other? What does it mean? Are there network statistics that should be the same?
* Could we compare the opinion space?
* How can I generate an artificial bias?
* Must I generate different population networks?

What did I miss?

Sample at random and maybe grouped

Clustering coefficient

Get the proportion on ANES data

* Issues for network statistics (polarisation & average path length (is size dependent)
* No community detection,
* Seeded – use a big size network as a population

Read an iterational proportional fitting 🡪

ANES 🡪 data set, where are different? Results for network

How big is the issue of the weighted different sample?

* Figure out the sample high amount of variance in the questions; weights?
* ANES (parameter)
* Wellcome

Weights used in ANES:

1. Base weights were calculated as the reciprocal of selection probability for each address. Within each of the three sample groups (web-only, mixed web, or mixed video) the selection probabilities were the same.

2. An adjustment was made for unknown address eligibility.

3. A screener nonresponse adjustment was made, weighting by the inverse of base-weighted response rates within groups defined by Census region, Census division, dwelling type, and whether or not the address had a telephone number associated with it on the sampling frame.

4. An adjustment was made for within-household sampling based on the selected person’s probability of selection within their household.

5. A pre-election nonresponse adjustment was made using two variables from the screener: gender and the number of eligible adults in the household.

6. Pre-election **raking** adjustments were made to make survey estimates match known population characteristics (for which the main source was the March 2020 Current Population Survey): age by gender, marital status by gender, race/ethnicity by educational

attainment, race/ethnicity by Census region, nation of birth, home tenure by metropolitan status, population density, household income, and early voter status. Large weight values were trimmed.

7. Post-election raking adjustments were made using the same characteristics as the pre-election adjustment.

8. Composite adjustments were made for combinations of sample groups. These composites were raked using the dimensions listed above and also adjust for different design effects in different sample groups by weighting each sample group according to its effective (design-adjusted) sample size.

Meeting (08/11/2021) DAVID

Build a Questionnaire start with simulated data

* Bias: different amount of people
* Chapter: bootstrap,
* 2,3 groups
* 3,4 questions
* Category: Dem, Rep or something else as a group
* Average connections: Compare to “entire” population”/ see how to do the bootstrapping
* How might we actually do this for the ANES data?
  + Survey data: CHECK the number of the individuals who are the same, mark them,
* Mark the individuals who are the same on our network (sample vs. overall sample); how many edges are there which represent the full agreement?
* Where does it break down?
  + READ chapter “bootstrap