Evaluation without Experimentation

Measuring the impact of relational organizing with causal inference

Emily Riederer
Bluebonnet Data Fellow for Two Million Texans

Two Million Texans wanted to understand whether their **all-volunteer, largest-ever relational organizing network** drove midterm turnout

>5,000 Volunteers **added** **Friend & Family to Network**

and encouraged network to GOTV, supported by the Reach app and personalized emails

Treatment Applied

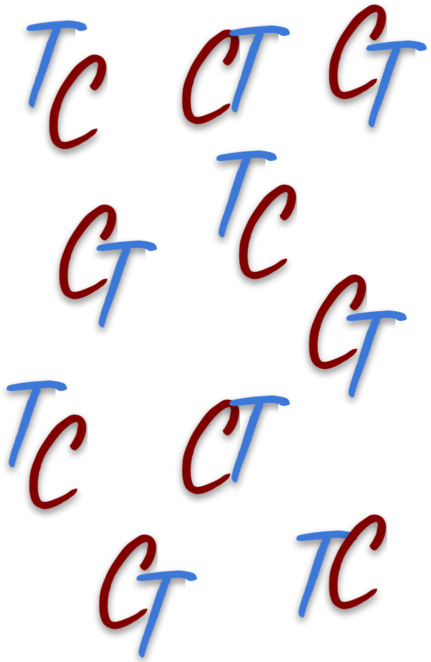
Voter presence in volunteer network *(contact assumed)*

Outcome of Interest

Increased voter turnout in 2022 midterm election

Did it work?!

Why not experiment?



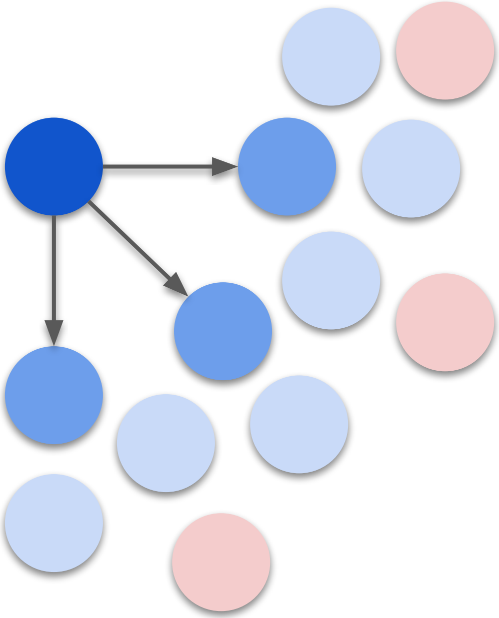
In industry, strategies are measured with random experiments

- Randomly assign people to ‘treatment’ and ‘control’
- Only intervene (e.g. encourage turnout) for treatment
- Compare results between groups

Field experimentation is not ideal in organizing

- Every vote matters! *Especially* for state and local races
- Unintuitive to request that volunteers *not* contact network

Why not *not* experiment?



Cannot just compare 2018 to 2022 for same voter set

Many causes of cycle-to-cycle change besides our campaign:

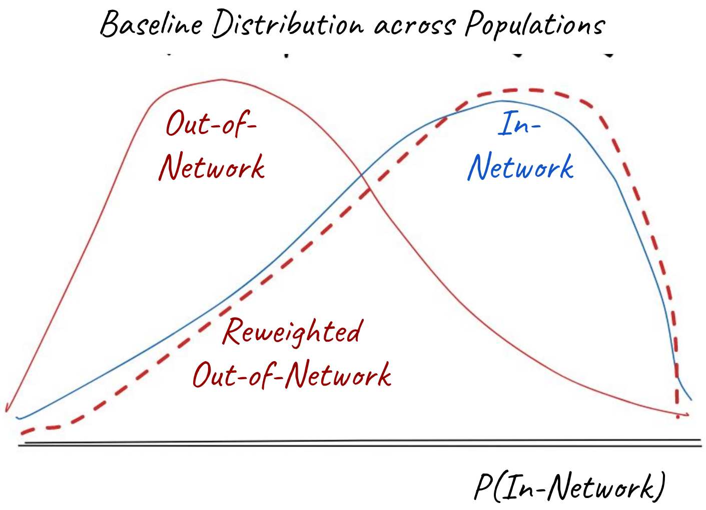
- Fundamental differences in coverage between cycles
- Presence of high-profile local races chance by-cycle behavior
- Redistricting

Cannot just compare in-network versus out-of-network

Many systemic differences between in- and out-of-network:

- Volunteers are more engaged than general population
- People tend to know people more like them
- Volunteers are steered to contact their ‘top targets’

We can ‘find’ comparable control individuals among out-of-network voters with Inverse Propensity of Treatment Weighting (IPTW)



Recipe:

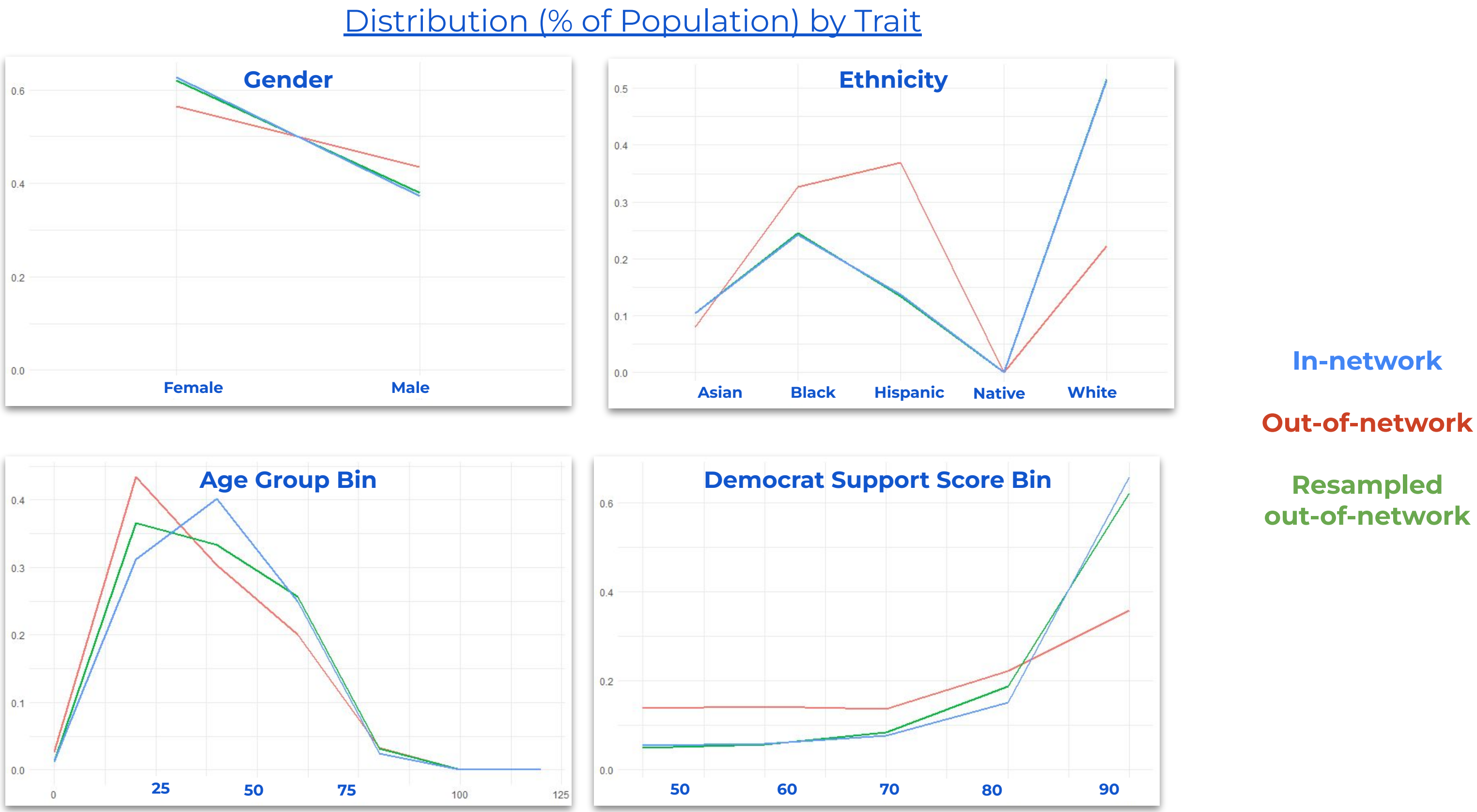
1. Model Probability(Treatment), **p**, based on voter traits
2. Compute IPTW weights*, **p / (1-p)**, for out-of-network voters
3. Weights represent similarity of each voter to our network
4. Calculate turnout for in/out-of-network using weights
5. Compare results

Assumptions:

- Non-treated population contains some individuals that are ‘similar to’ each treated individual
- Common causes of treatment and turnout are observable

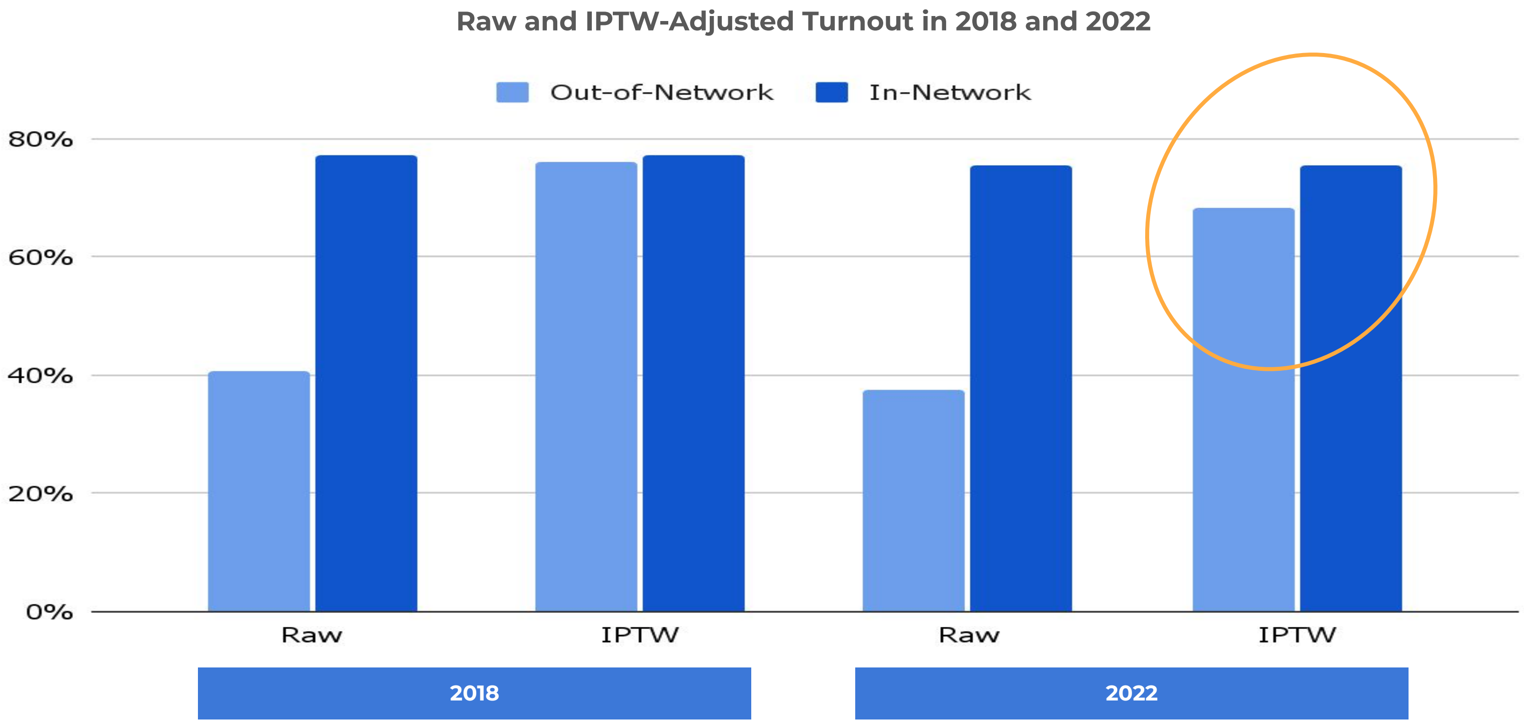
Note: See appendix for formulas and justification for different mappings of probabilities to weights

Reweighting adjustment in action on baseline voter characteristics
(example: Harris County)



Note: Dimensions shown for example purposes only. More features were used in reweighting

Reweighting adjustment in action on pre/post-treatment outcomes
(example: Harris County)



IPTW closes the gap in the baseline for 2018

The remaining **gap** in 2022 suggests an effect

Note: See appendix for fully worked example

We increased turnout by +4-6 percentage points in our core counties

All-Election Turnout by Treatment of ‘In-Network’ of Highly Engaged User

		Effect on Turnout*	
County	**N**	**Percentage Point Increase within Treatment**	**Number of Voters (N * PP Increase)**
Harris	31,712	**+5.9**	**1,871**
Fort Bend	13,015	**+4.2**	**547**
Travis	45,361	**+4.8**	**2,177**

Results suggest impact exceeded win margin in key local judicial races!

Step-by-step implementation details are available in the appendix

Note: Estimates represent lower-bound of ‘true’ impact since treatment is ‘in-network’ and not observed contact

Questions?

↓ Get in touch ↓

@emilyriederer on Web | Twitter | GitHub | LinkedIn | Gmail

↓ Check out these resources ↓

Understanding propensity score weighting
Causal design patterns
Causal inference resource roundup

↓ Reference these (free!) books ↓

The Effect: an Introduction to Research Design and Causality
Causal Inference: the Mixtape
Causal Inference: What If?

↓ Find more math in the Appendix ↓

Appendix

Different mappings from propensity scores (P) to weights allow us to calculate different effects

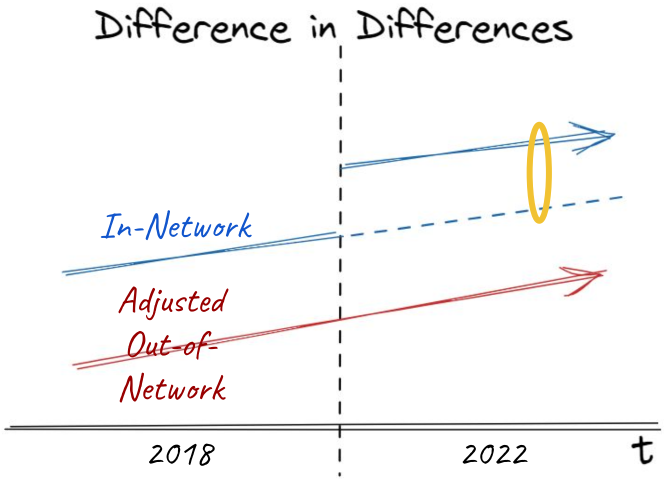
	Average Treatment Effect on the Treated (ATT)	**Average Treatment Effect (ATE)**	**Average Treatment Effect on the Control (ATC)**
Key Question	What effect did we accomplish where we were actually acting?	What effect could we accomplish if we could treat everyone?	What effect could we accomplish where we weren't acting?
Weight (Treated)	1	1/P	(1-P)/P
Weight (Control)	P / (1-P)	1/(1-P)	1

⭐ *Most often what we want to know for program evaluation!* ⭐

Intuition for ATT weights

Recall Unit Cancellation	Analogize to Weights
1 foot	1 control unit
X	X
(12 inches / 1 foot)	P treatment-like units / (1-P) control-like units
=	=
12 inches	P / (1-P) treatment-like units

Unexplained residual confounding in 2018 turnout was further reduced with a difference-in-differences strategy



When we have:

- Different baselines in comparison groups
- Variation across time (pre/post)

Recipe:

1. Compute difference in pre-treatment period (2018)
2. Compute difference in post-treatment period (2022)
3. Take the difference between (2) and (1) to find the effect

Assumptions:

- Decision to treat not influenced by anticipated outcome
- If not for the treatment, groups would have parallel trends
- Treatment of one group does not affect behavior of other

Reweighting adjustment in action (example: Harris County)

	Raw Turnout		Propensity-Score Weighted Turnout		Final Effect Estimate
Network	**2022**	**2018**	**2022**	**2018**	**Adjusted 2022 - 2018 PP+**
In	75.4%	77.1%	75.4%	77.1%	
Out	37.5%	40.7%	68.4%	76.0%	
Difference	**37.9%**	**36.4%**	**7.0%**	**1.1%**	**+5.9%**

Raw comparison between in/out of network suggests a massive effect

However, the existence of a similar gap in 2018 suggests systemic differences between groups

We 'close the gap' by reweighting the out-of-network observations to "look like" the in-network

Adjustment nearly eliminates the 2018 gap, so the "above and beyond" 2022 gap is driven by RO

Note: Difference-in-differences used to close the residual gap and control for unexplained confounding