4 - Crowd-sourced online data

Diego Alburez-Gutierrez
MPIDR
European Doctoral School of Demography 2021-22

24 February 2022

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH
MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

Agenda

1. Q&A
2. Crowd-sourced data
3. **Break**
4. User-generated family trees
5. Limitations and bias

Q&A

- Questions about the assignment
- Questions about the study last week
- Other?

Crowd-sourced data

What is crowd-sourced data?

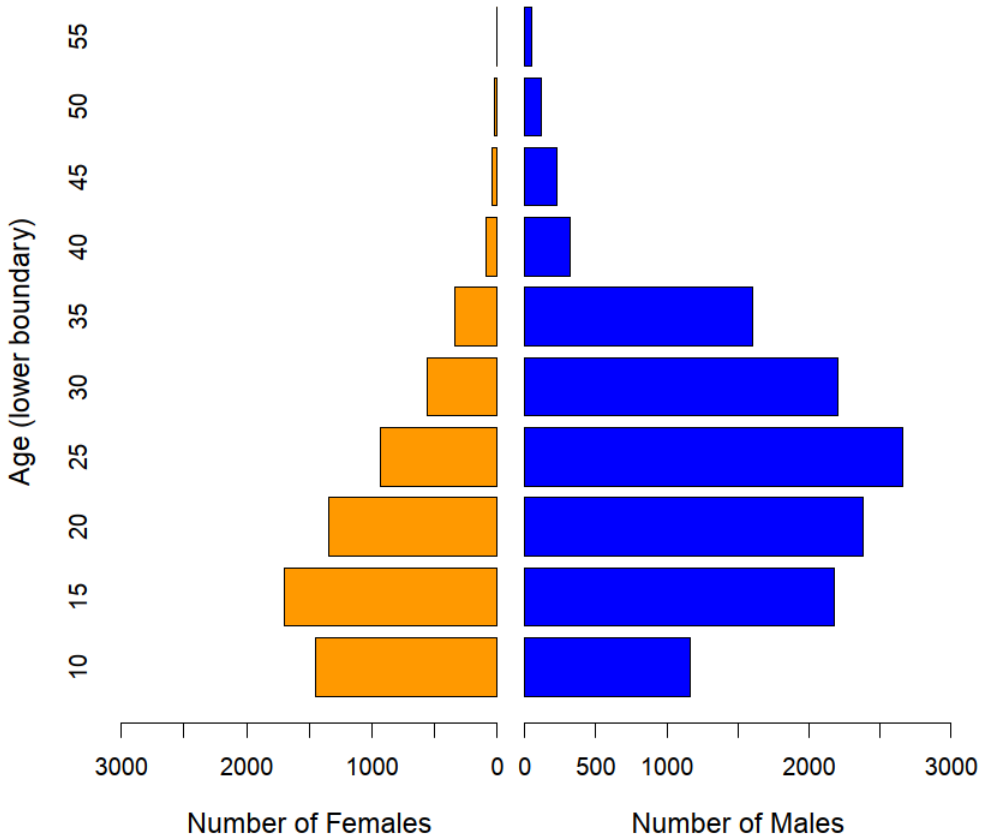
- 'Bottom-up' user-generated content
- Usually large
- Available online - may have been produced offline
- By-product of decentralized activity

Examples in this session

1. Recruitment platforms
2. Family history research

Group discussion

Consider this image of Twitter user demographics:



1. Do you see anything odd?
2. What might be causing the issue?

Online recruitment + post-stratification

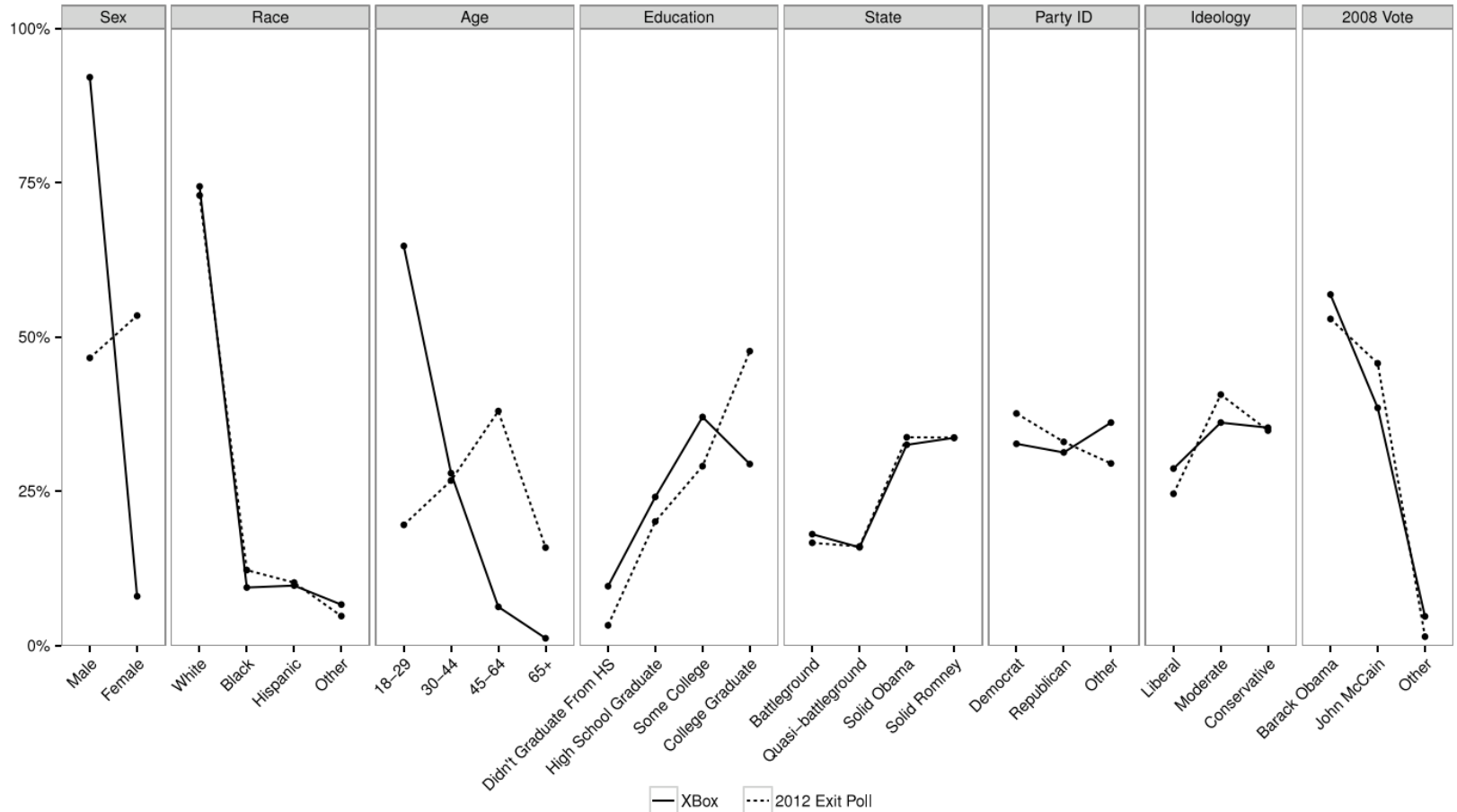


Figure 1: Different demographics

Wang, Rothschild, Goel, and Gelman 2015. Forecasting elections with non-representative polls. International Journal of Forecasting, 31 (3).

Forecasting with non-representative polls

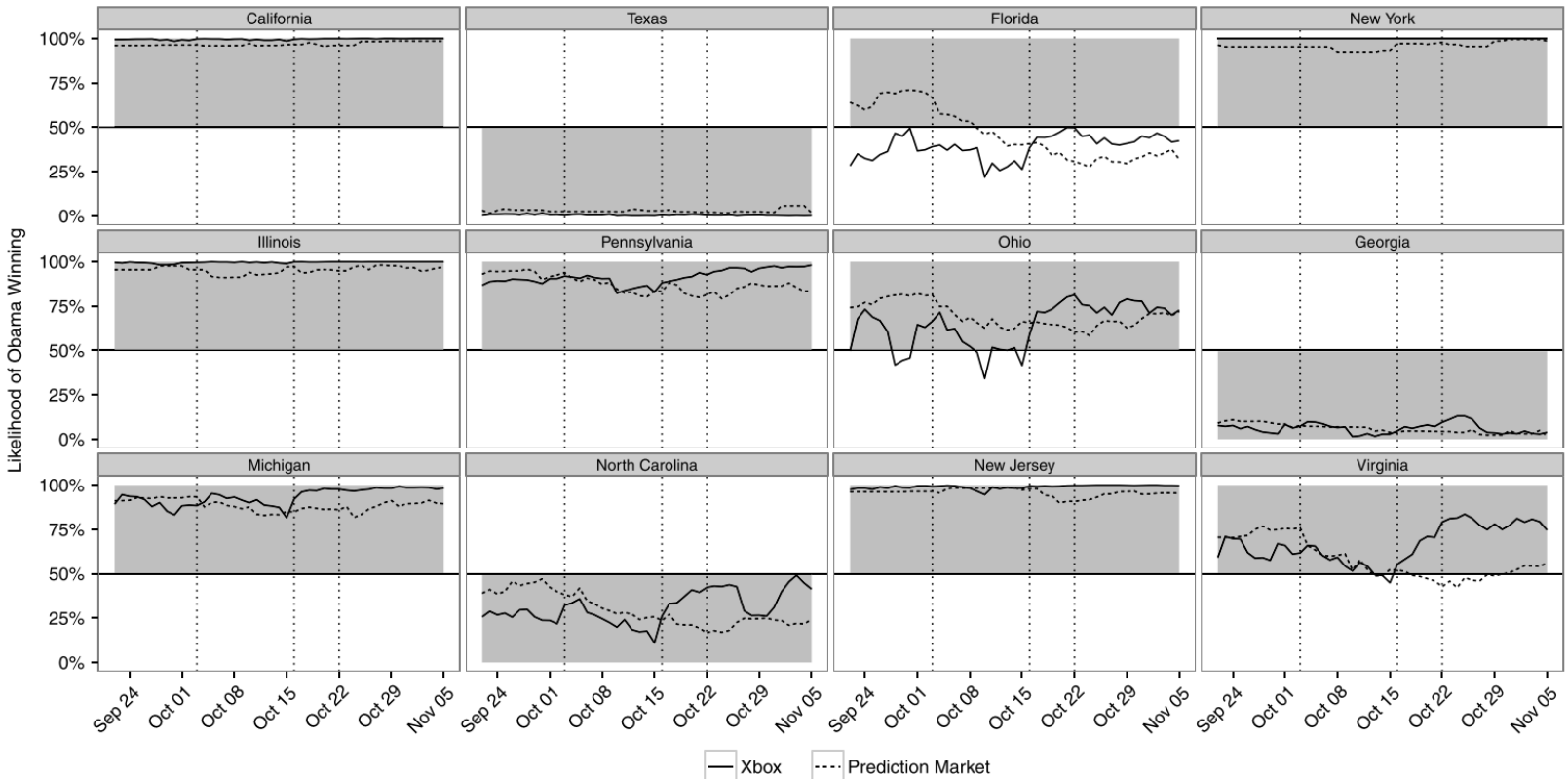


Figure 2: Likelihood of Obama victory

Wang, Rothschild, Goel, and Gelman 2015. Forecasting elections with non-representative polls. International Journal of Forecasting, 31 (3).

An example closer to home

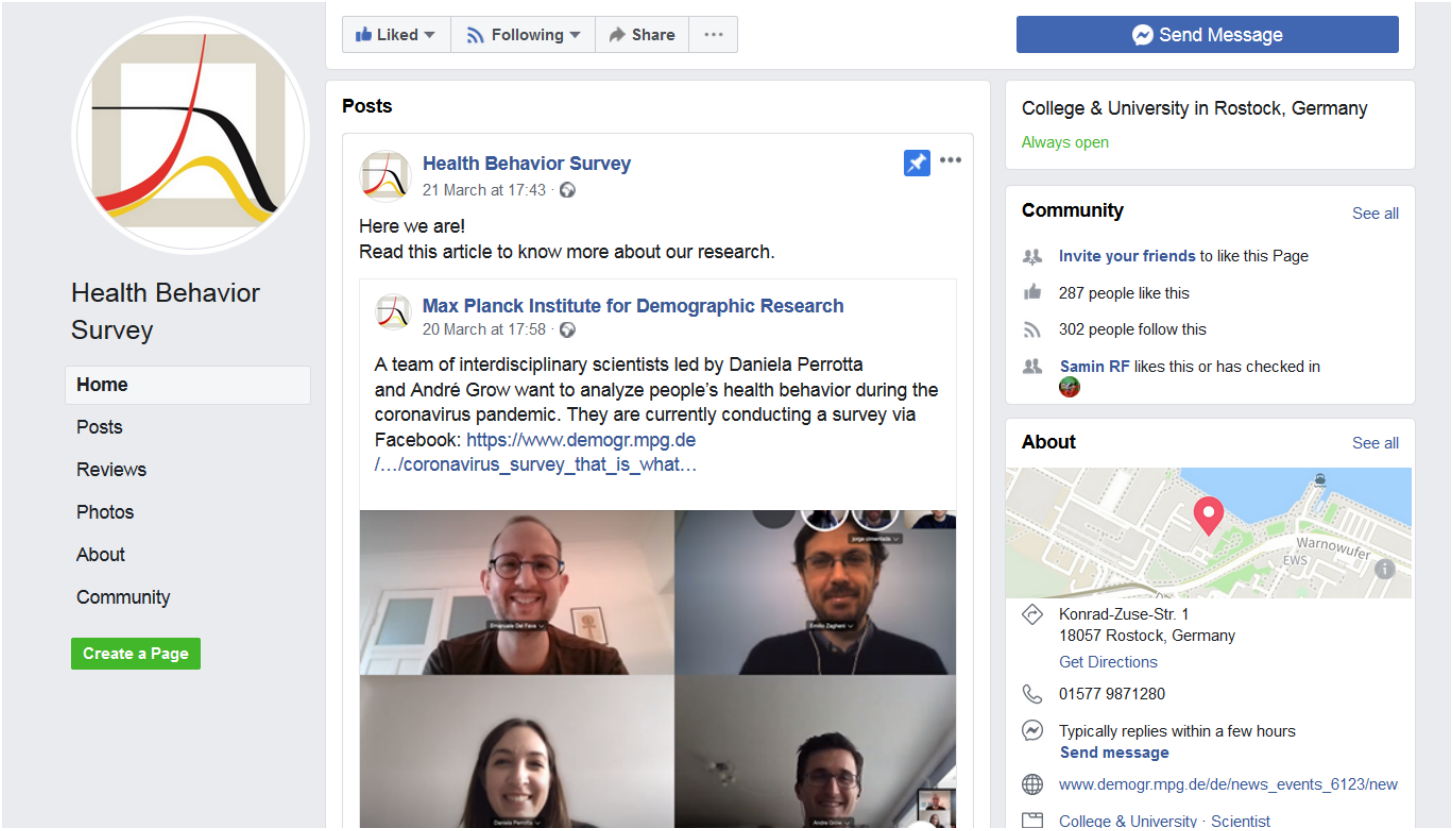


Figure 3: Learning more about the coronavirus

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., and Zagheni, E. (2020). Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations. Journal of Medical Internet Research 22(12):e20653

Break

Online genealogies

A genealogy is the history of a population

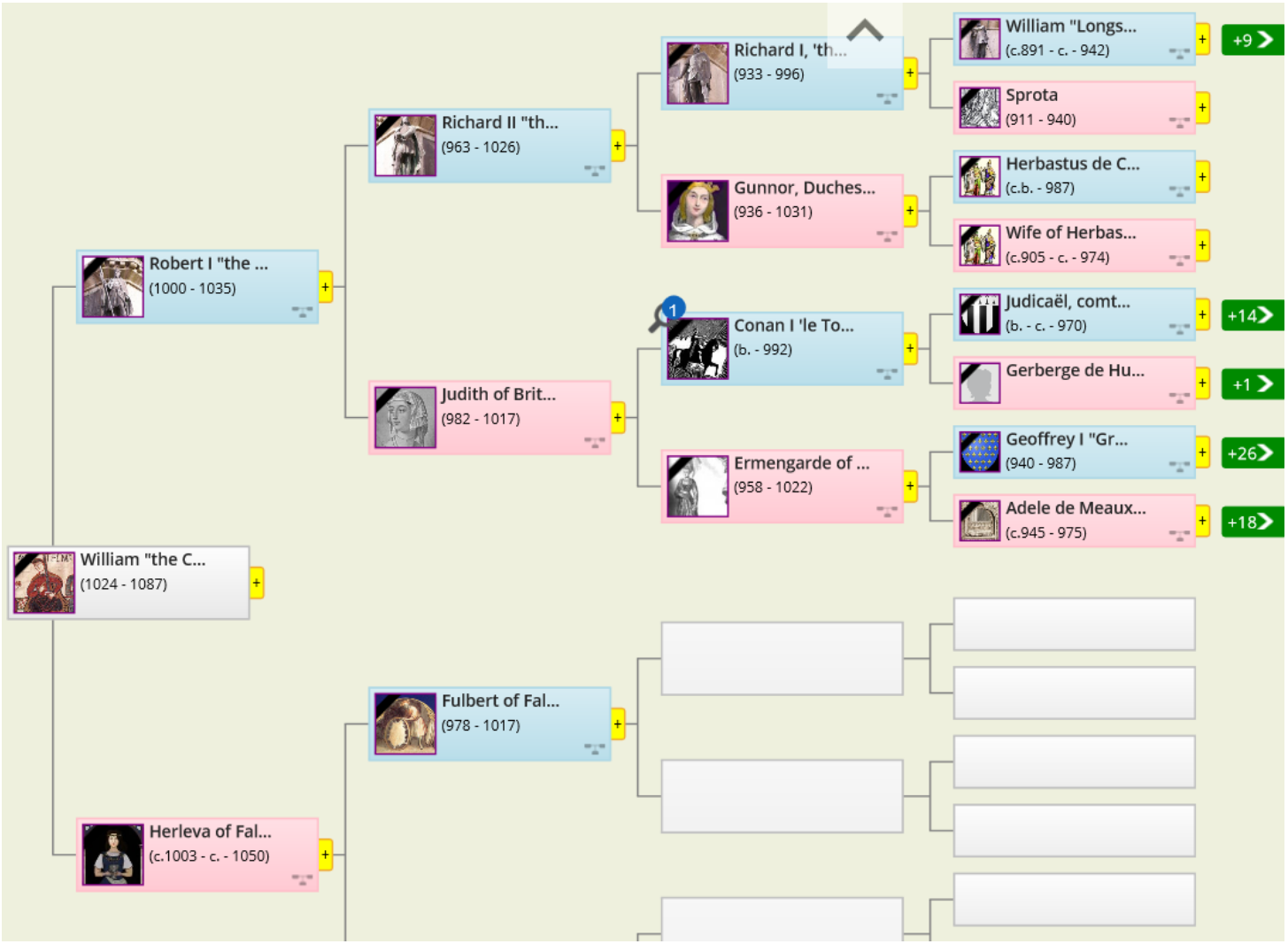


Figure 4: A Geni.com family tree

WikiTree

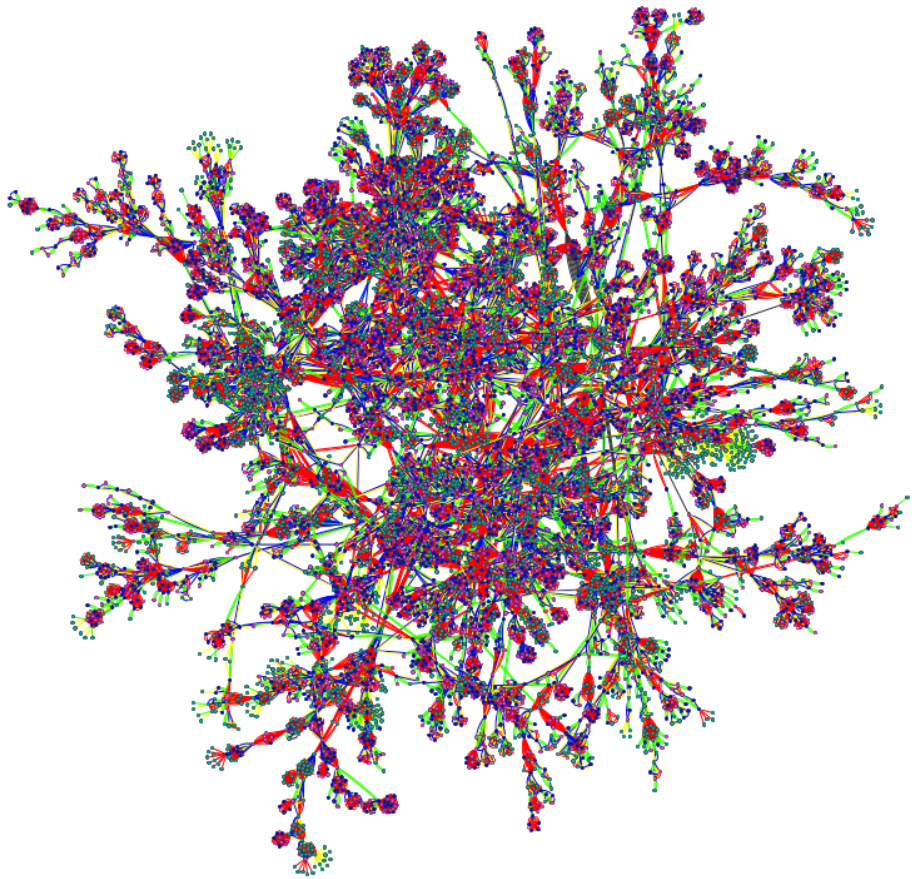


Figure 5: Cool dataviz, but what do we learn from the data?

Fire, M. and Elovici, Y. (2015). Data mining of online genealogy datasets for revealing lifespan patterns in human population. ACM Trans. Intell. Syst. Technol. 6(2):28:1–28:22.

Geni.com: a social network for genealogists

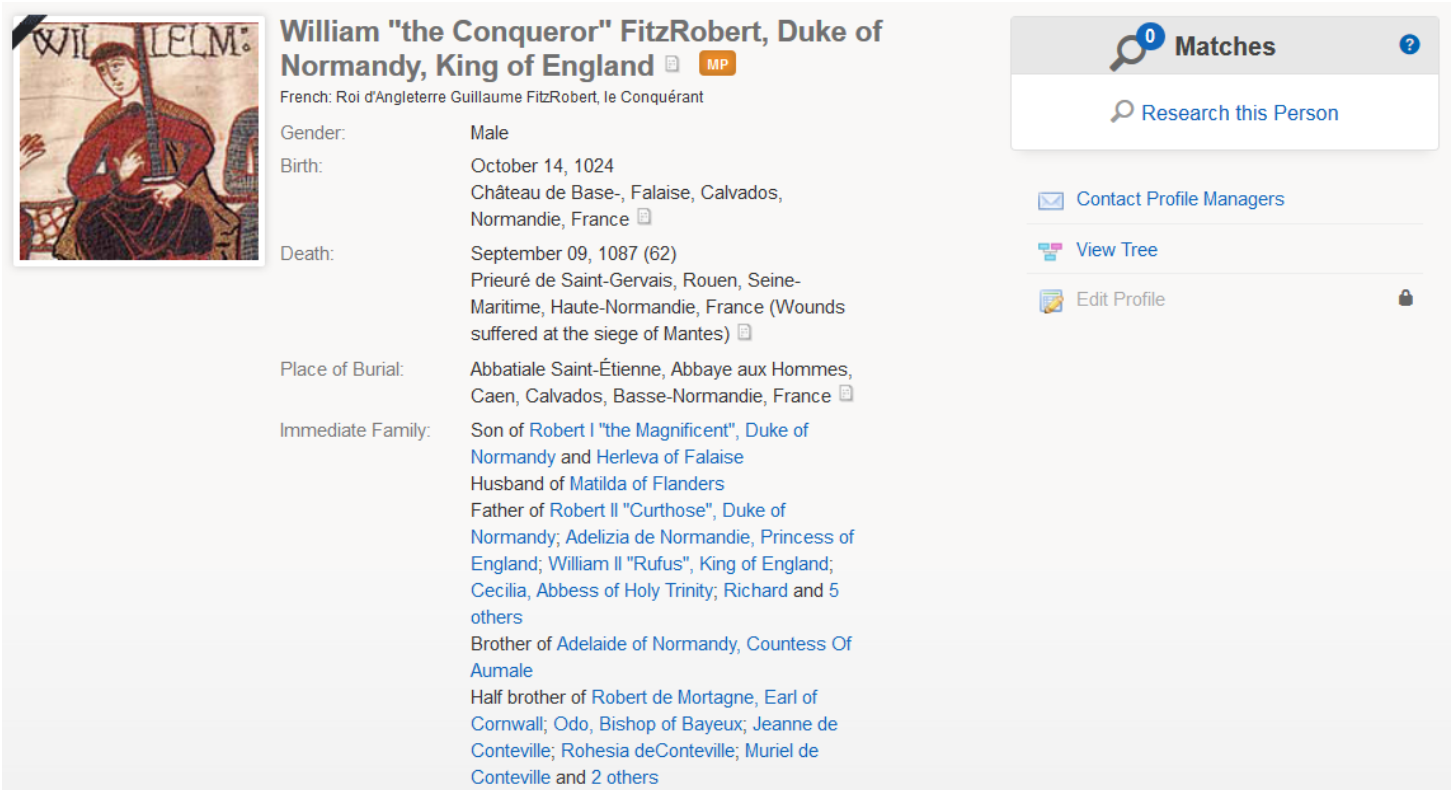


Figure 6: Everyone's relative

Built on top of (private) genomic data

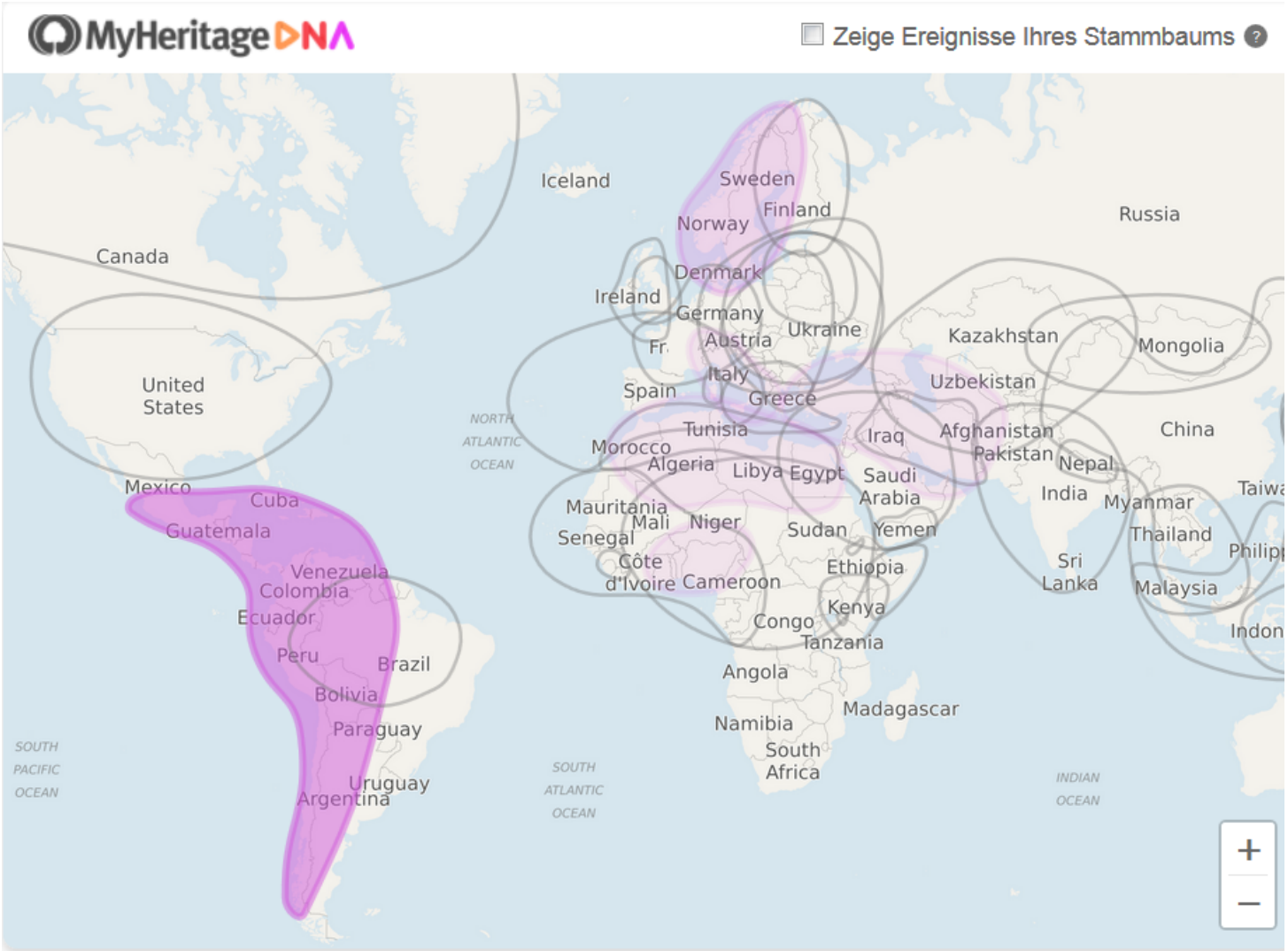


Figure 7: 'Ethnicity' estimates

Our example: Familinx data

1. Genealogy-driven social media data
2. Goal: register entire population of the world
3. 86M unique profiles over last 400 years
4. Curated, with quality checks
5. Geo-coded events - 55% Europe; 30% North America

Kaplanis, J., et al. (2018). Quantitative analysis of population-scale family trees with millions of relatives. Science 360(6385):171–175.

Geographic distribution in Familinx

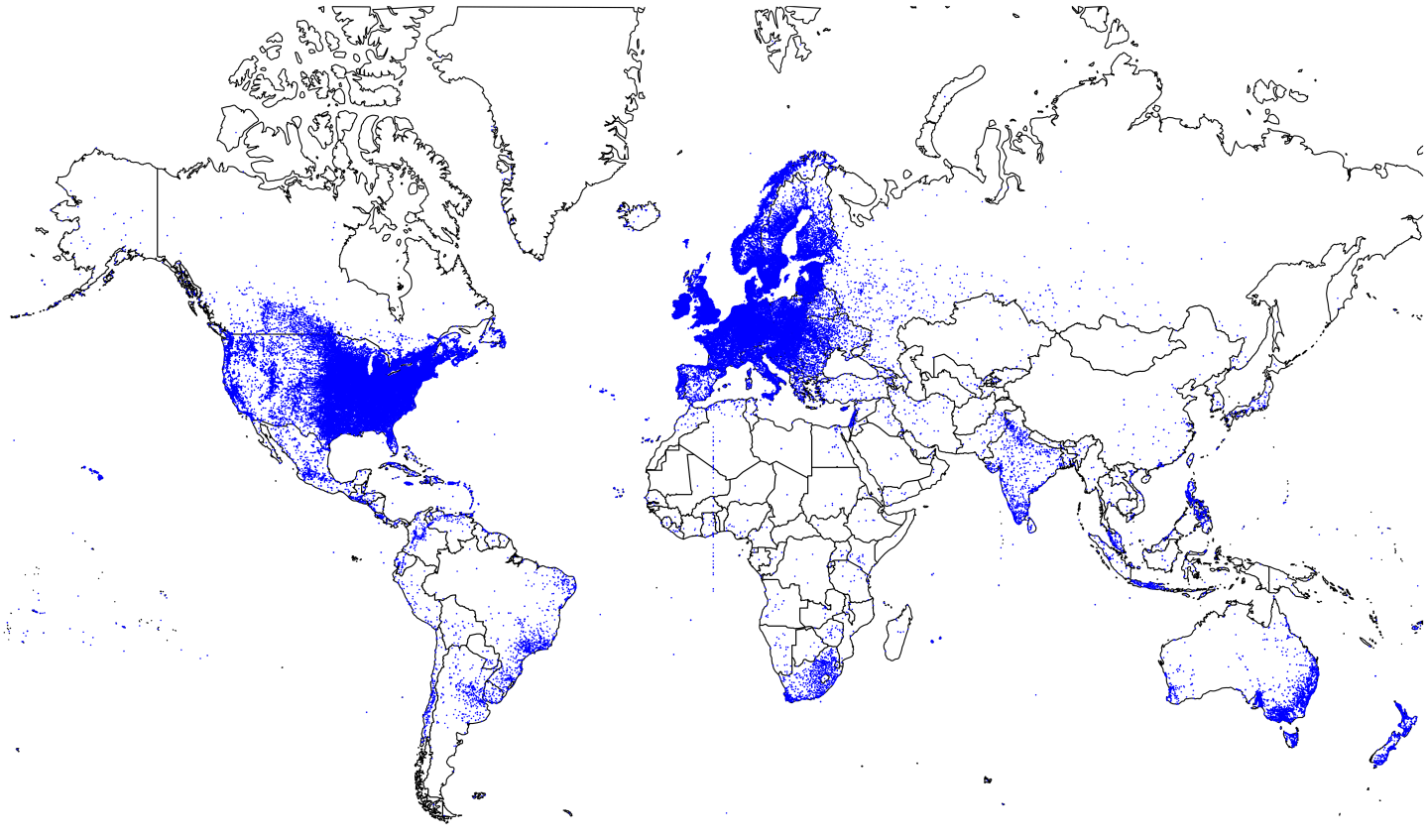


Figure 8: Birth events worldwide

Population alive by age group (all countries)

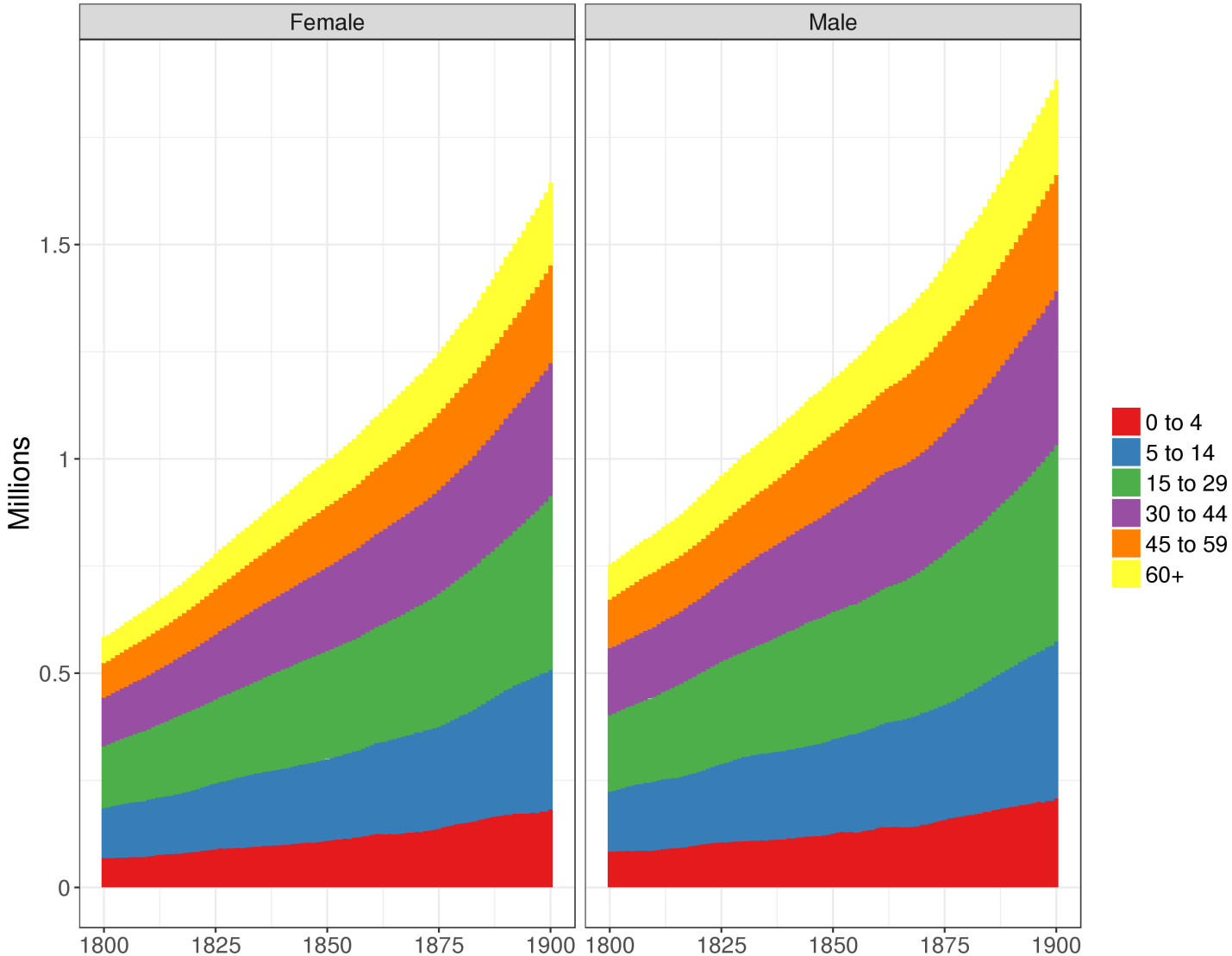


Figure 9: Yearly censuses from Familinx data

Male bias in online genealogies (Familinx)

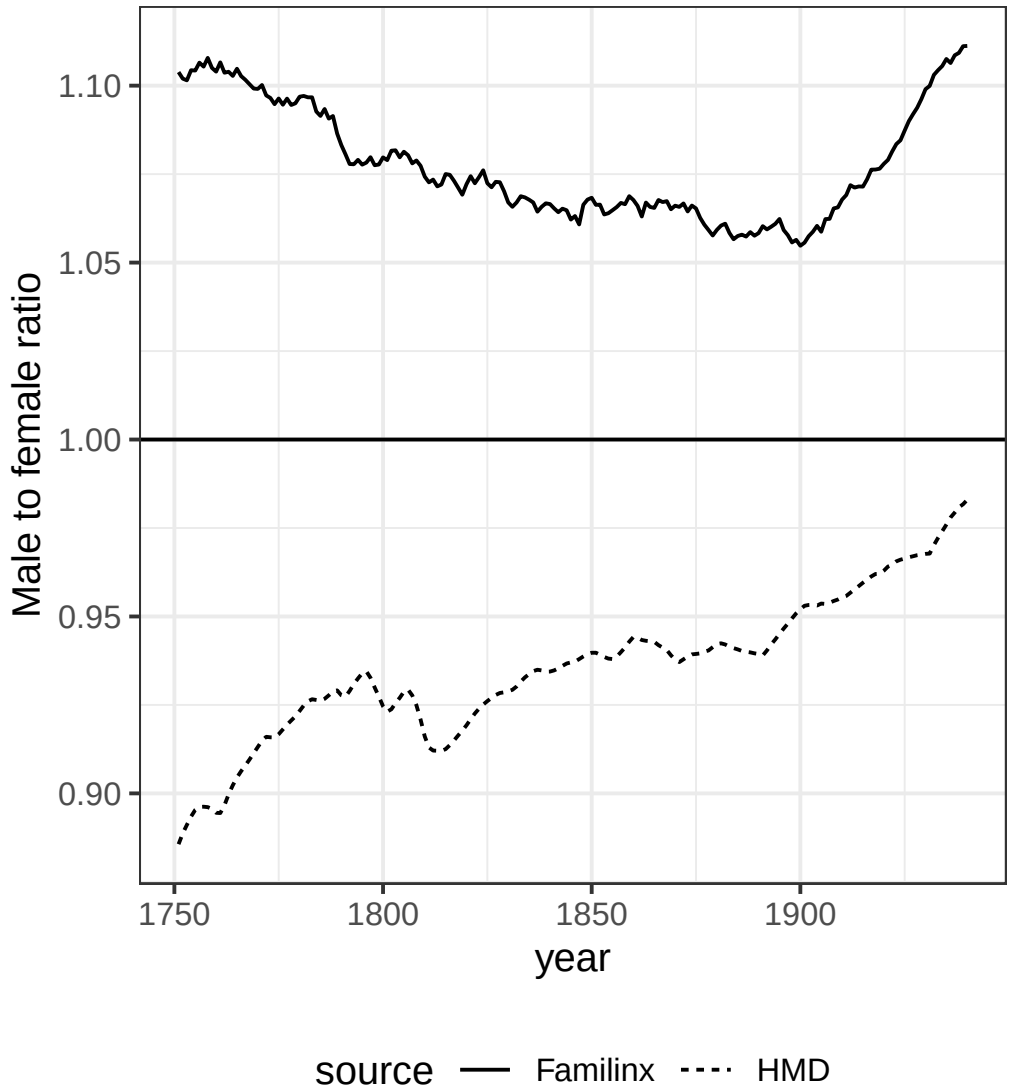


Figure 10: Sex ratio in Sweden

The data generating process

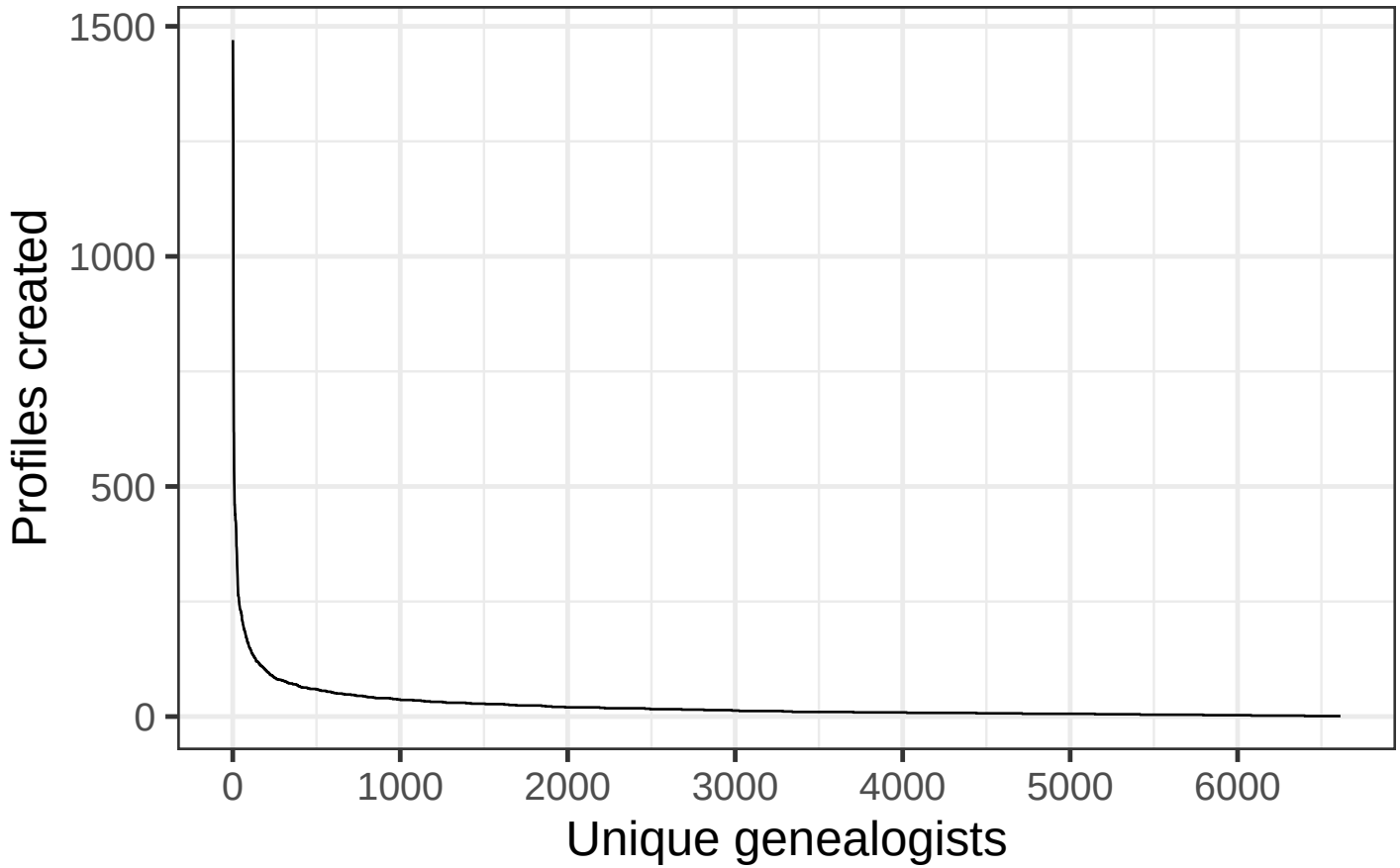


Figure 11: Power law function of genealogists

A closer look at the data generating process

Figure 12: The crowd-sourcers

https://blog.myheritage.com/2019/09/myheritage-live-2019-recap/

The Swedish sample

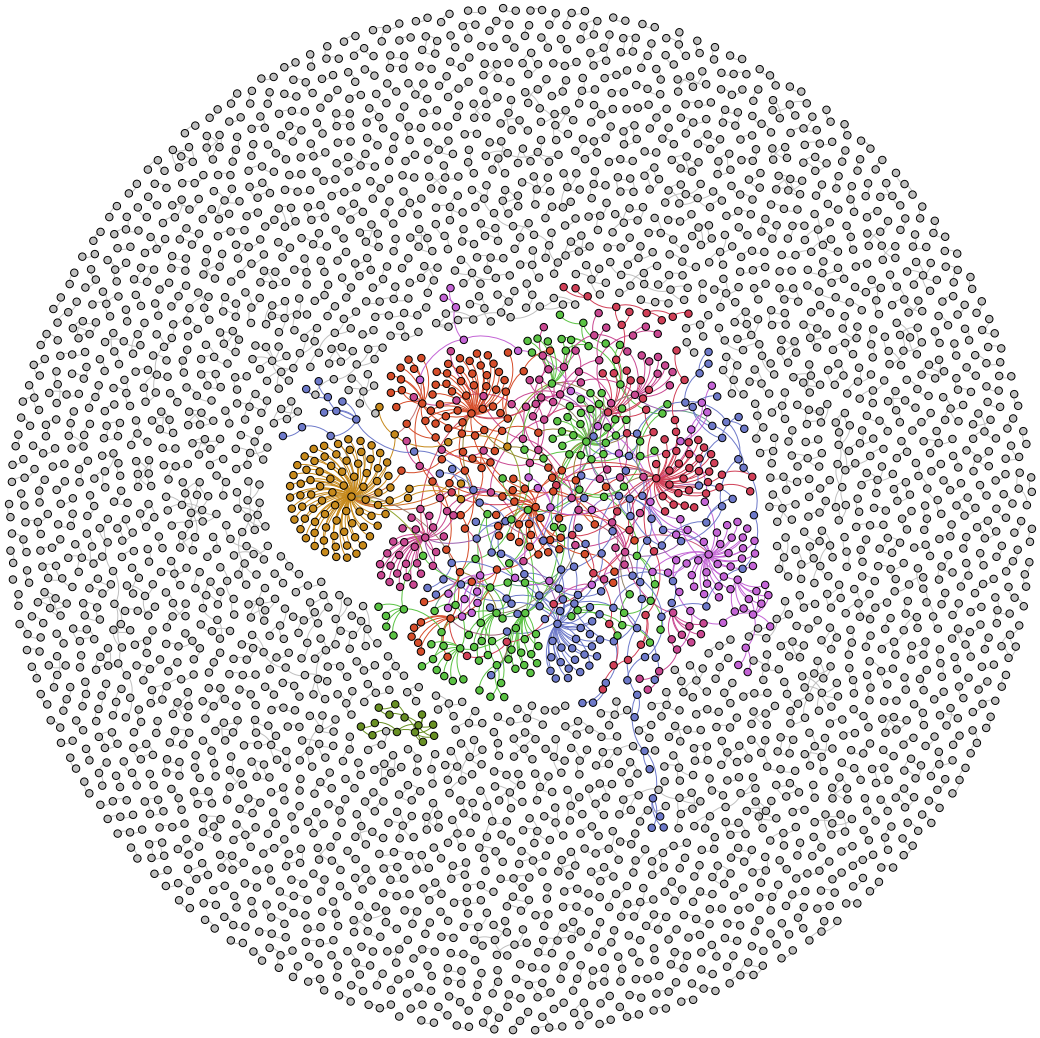


Figure 13: Subseting a genealogical network

Diego Alburez-Gutierrez and Samin Aref (2020)

Geographic distribution in Familinx data (sample)

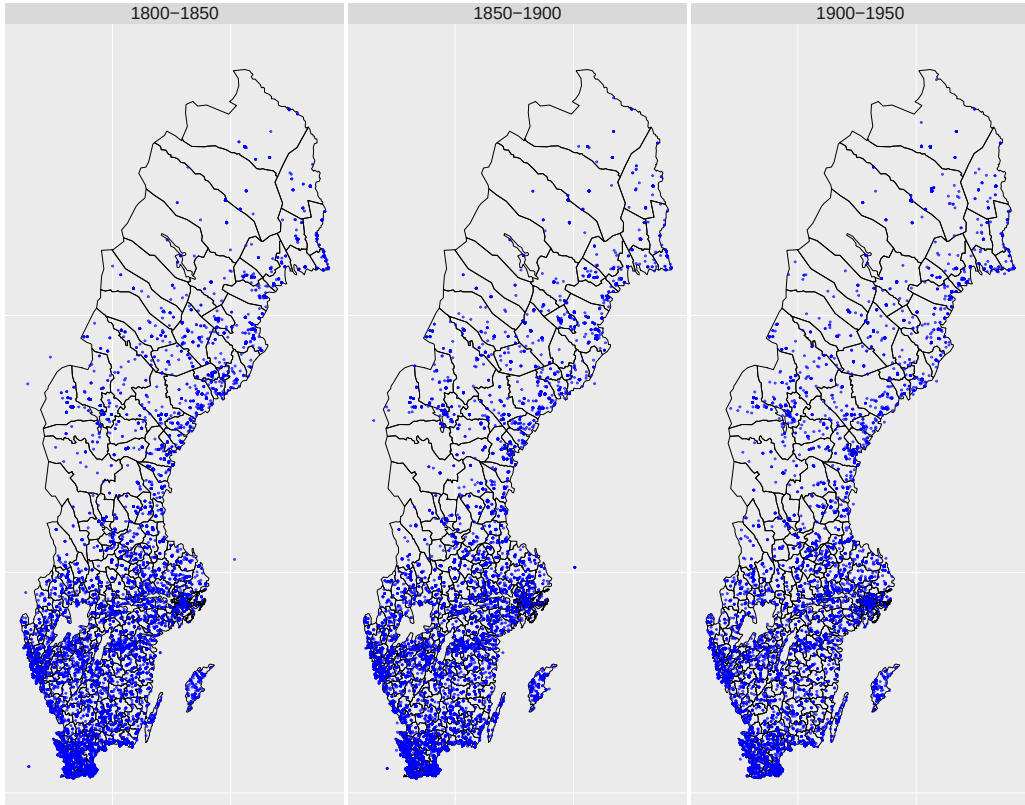


Figure 14: Birth events in Sweden over time

Diego Alburez-Gutierrez (2020)

Bias in lifespan dynamics from online genealogies

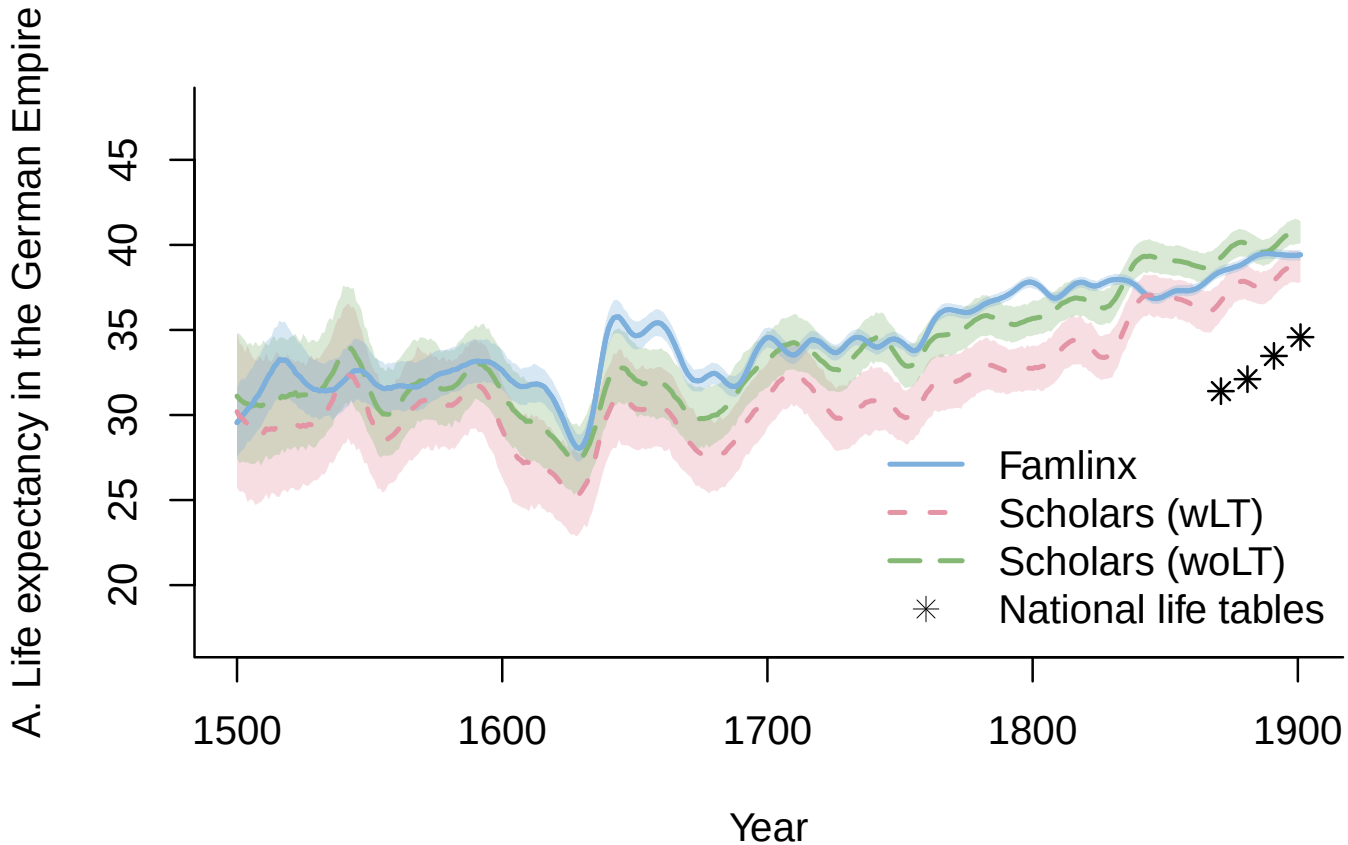


Figure 15: Life expectancies at age 30

Stelter, R. and Alburez-Gutierrez, D. (forthcoming). Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics. Proceedings of the National Academy of Sciences.

Correcting bias: an example using post-stratification weights

Starting point: Online genealogies are a non-representative sample of real-world genealogies.

- Online genealogies $\neq$ offline genealogies
- Unknown 'weights' - derive from comparison to trusted sources
- Understand data-generating process

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.

Numerator - death events

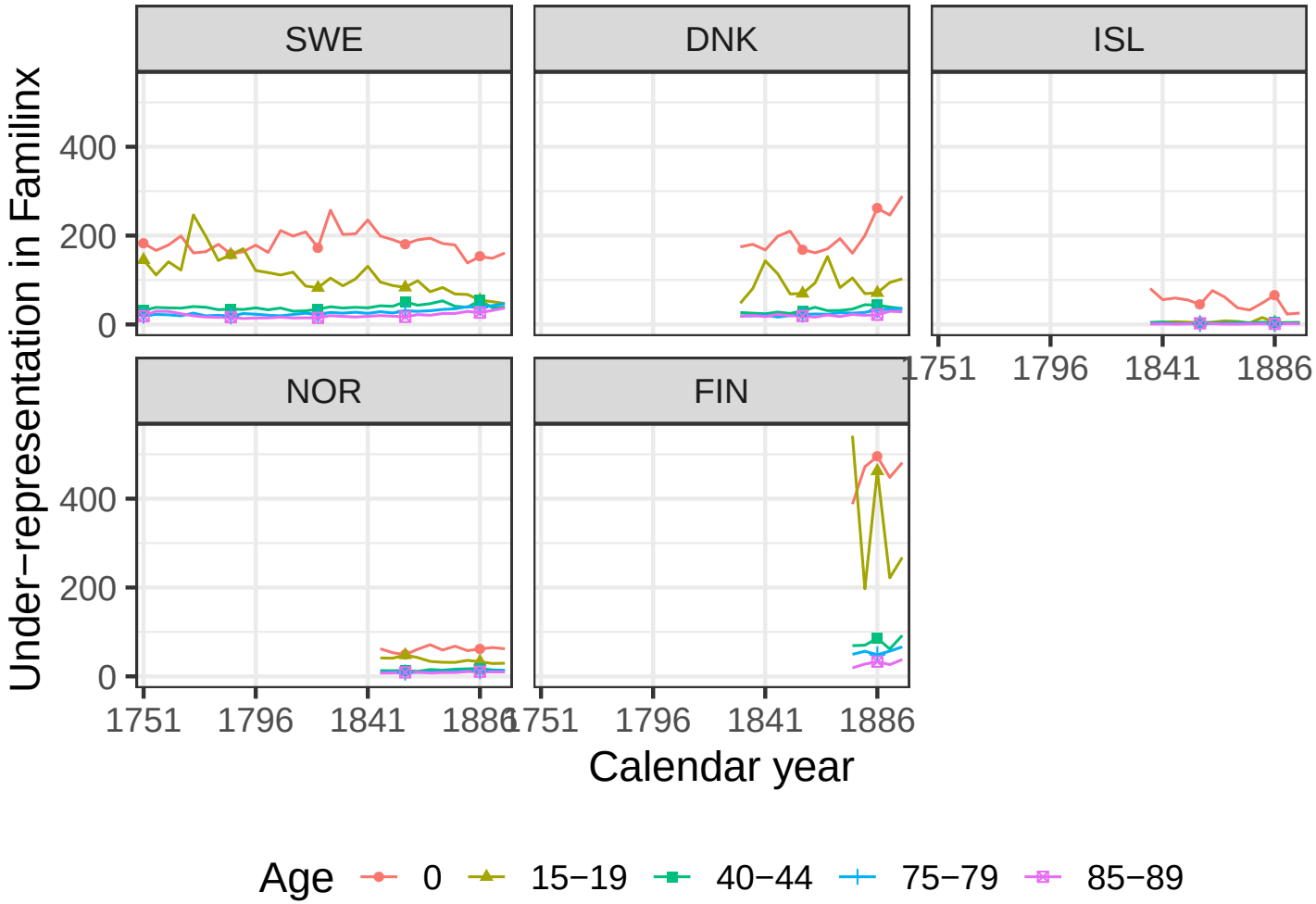


Figure 16: Observed deaths in four countries

Denominator or exposure

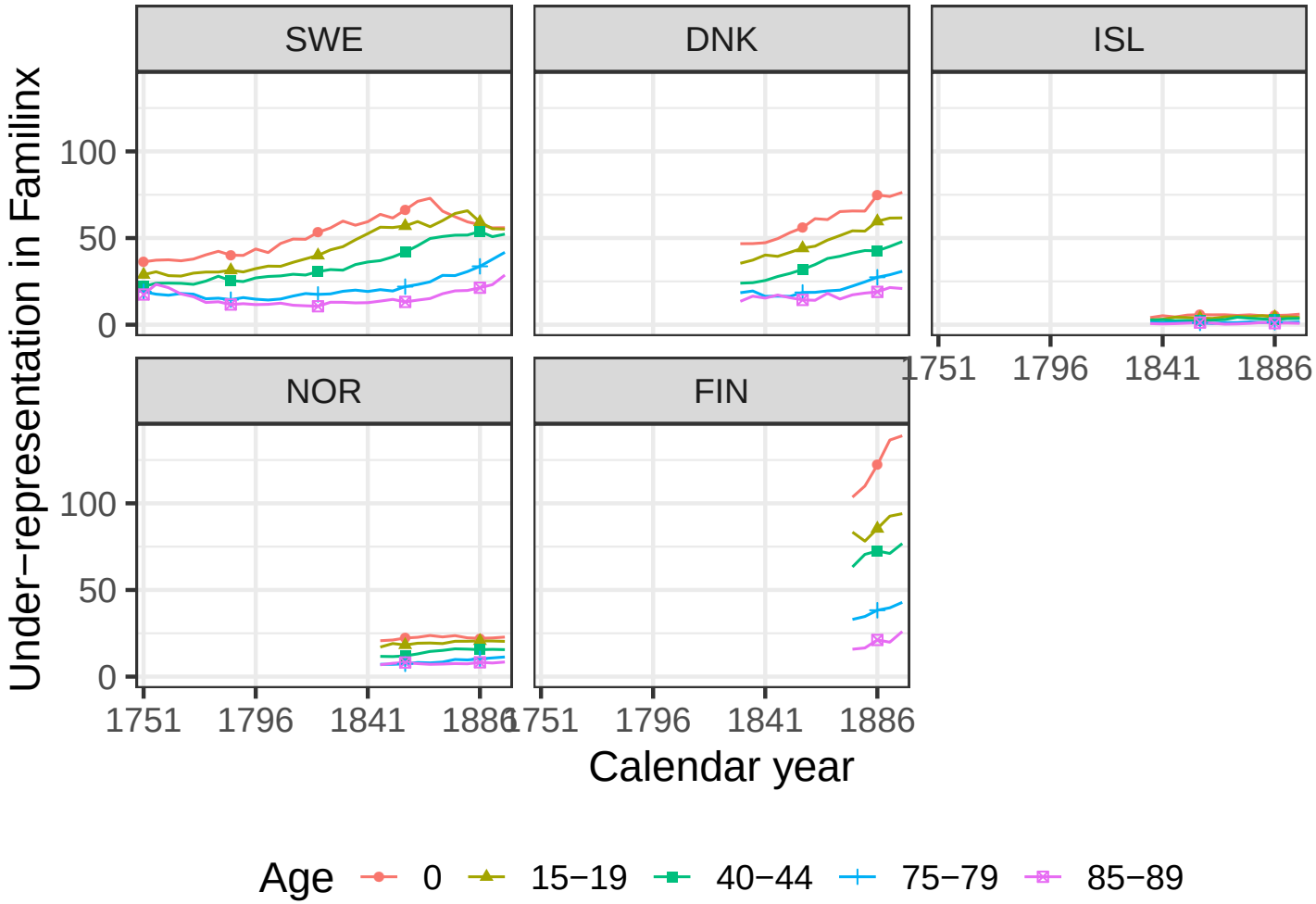


Figure 17: Population alive by age and sex in four countries

‘Corrected’ demographic rates (Finland)

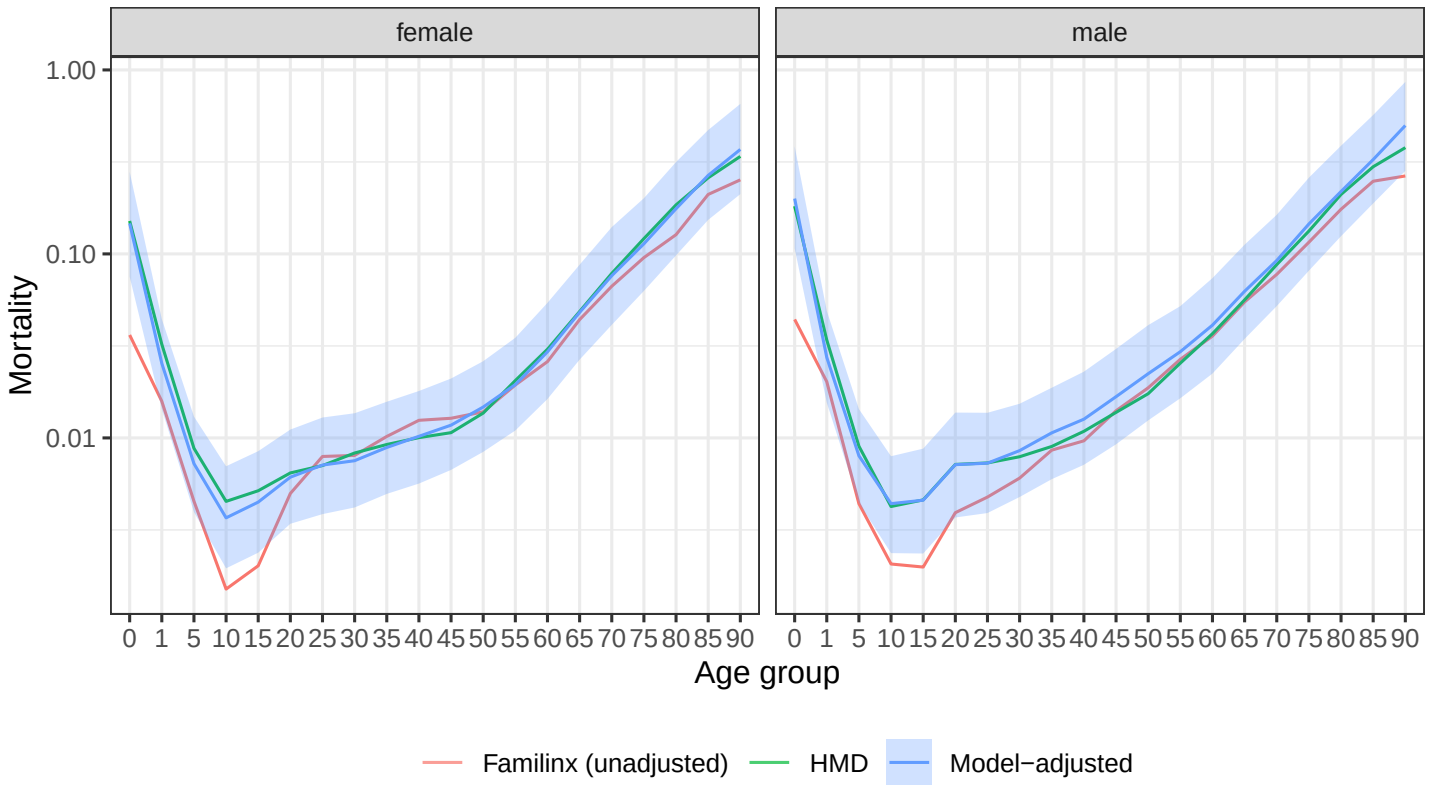


Figure 18: Weighted mortality rates

Chong, Alburez-Gutierrez, Del Fava, Alexander, Zagheni (2022). Identifying and correcting bias in big crowd-sourced online genealogies. MPIDR Working Paper. Rostock: Max Planck Institute for Demographic Research.DOI: 10.4054/MPIDR-WP-2022-005.

‘Corrected’ life expectancy (Finland)

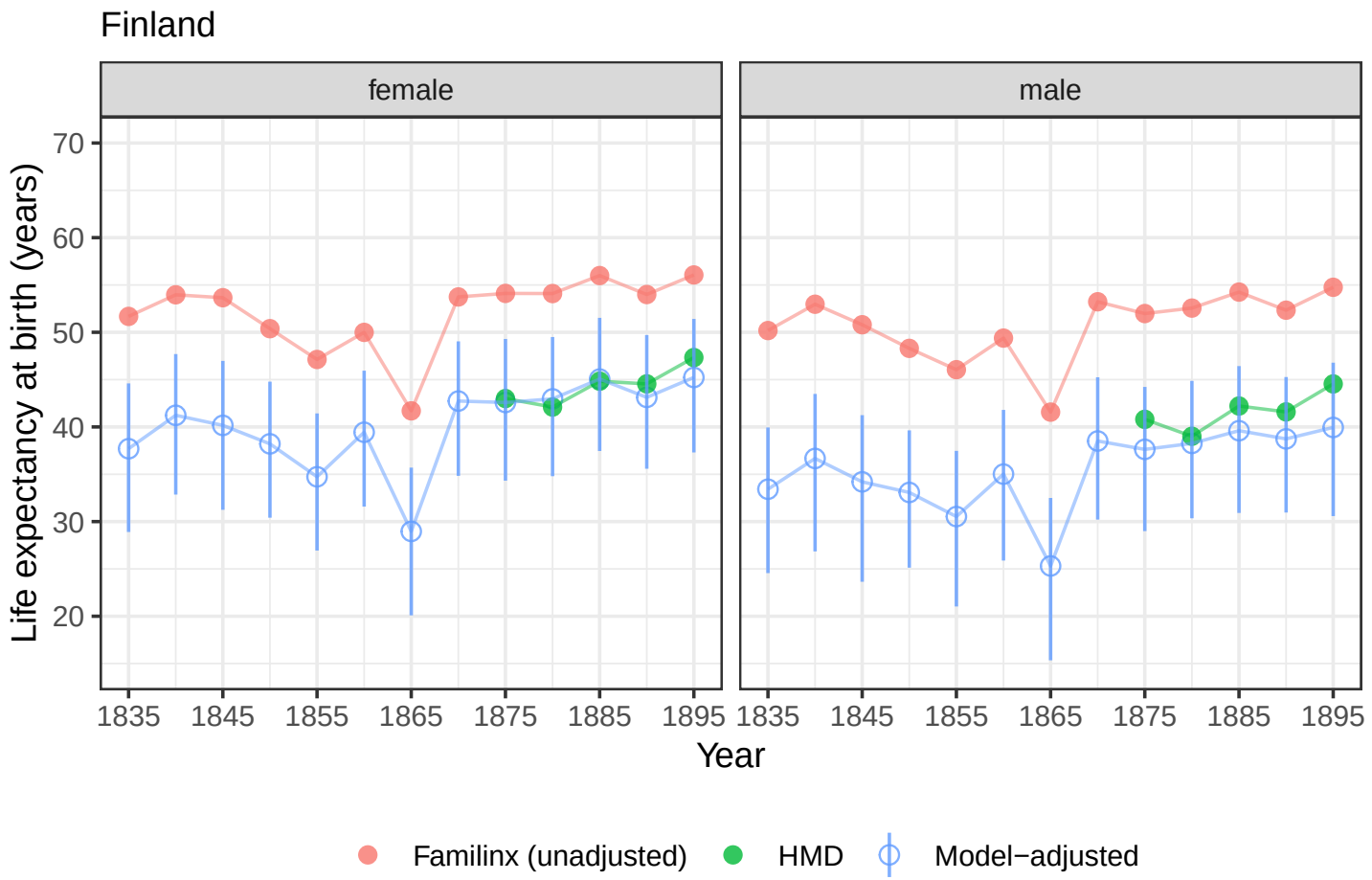


Figure 19: Weighted Life expectancy

Chong, Alburez-Gutierrez, Del Fava, Alexander, Zagheni (2022). Identifying and correcting bias in big crowd-sourced online genealogies. MPIDR Working Paper. Rostock: Max Planck Institute for Demographic Research.DOI: 10.4054/MPIDR-WP-2022-005.

Group discussion

Think about a way in which you could use crowd-sourcing in your own research.

1. Which platform would you use?
2. Which challenges do you foresee?
3. Special measures to protect privacy?