Global gender estimation from distribution of first names

Manolis Antonoyiannakis[1,2], Hugues Chaté[3,4], Serena Dalena[1], Jessica Thomas[1], and Alessandro S. Villar[1]

[1]*American Physical Society, 1 Physics Ellipse, College Park, 20740 Maryland, USA*
[2]*Department of Applied Physics & Applied Mathematics, Columbia University, New York, USA*
[3]*Service de Physique de l'Etat Condensé, CEA, CNRS, Université Paris-Saclay, CEA-Saclay, 91191 Gif-sur-Yvette, France and*
[4]*Computational Science Research Center, Beijing 100193, China*

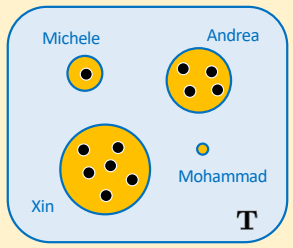
https://ggem.app

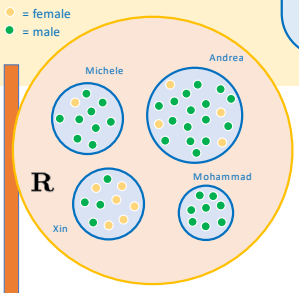
Problem: Find gender makeup given a list of names

β = female fraction

Andrea: 4
Michelle: 1
Mohammad: 0
Xin: 6

(Example)

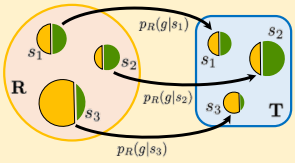




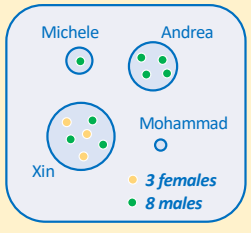
"Cheat sheet": Population at large

$p(\text{male}|\text{name} = \text{"Andrea"}) = 86.5\%$
$p(\text{male}|\text{name} = \text{"Michele"}) = 95.4\%$
$p(\text{male}|\text{name} = \text{"Mohammad"}) = 99.99\%$
$p(\text{male}|\text{name} = \text{"Xin"}) = 52.5\%$

Solution from the literature: Distribute according to the observed proportions

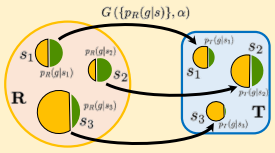


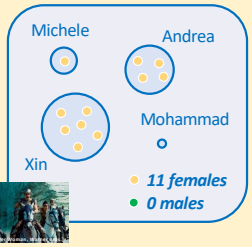
Implicit assumption of *fair sampling*:
... gender independence?!?



(effectively, a gender-balance approximation)

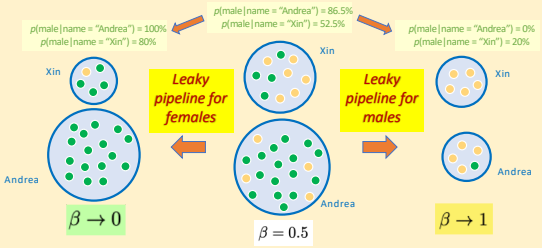
global Gender Estimation Method (gGEM): Estimate how a social dynamic shapes the name list





$p(\text{male}|\text{name} = \text{"Andrea"}) = 0\%$
$p(\text{male}|\text{name} = \text{"Michele"}) = 0\%$
$p(\text{male}|\text{name} = \text{"Mohammad"}) = 0\%$
$p(\text{male}|\text{name} = \text{"Xin"}) = 0\%$

How do conditional probabilities transform? A "leaky pipeline" model of social dynamic



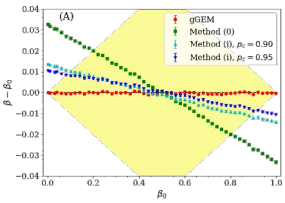
$$p(g|s) = p(g|s;\beta) = G(\{p_R(g|s)\}, \beta)$$

Results: Performance comparison

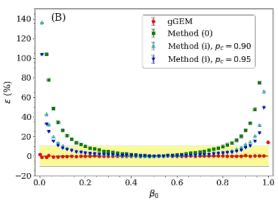
- SSN + census data for: US, Brazil, France
- 1000 test populations
- 1000 individuals per test population

- Better than 1 percentage point in gGEM estimate

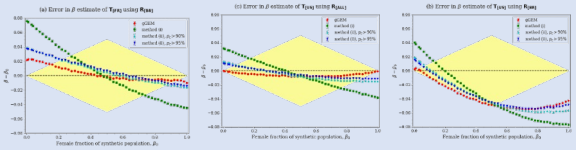
Absolute error



Relative error



Mismatched populations



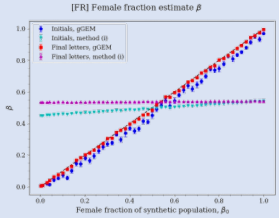
Conclusion: Performance comparison

Methods in the literature:
Accurate if target population is gender-balanced
[but: assumption not true in several cases of interest]
More accurate if gender-ambiguous names are removed
[price: smaller sample]
=> Otherwise, ***systematic overestimation***

Global gender estimation method:
Works for any gender mix
Uses the whole set of names
Robust against mismatches in gender-name correlations
Picks up very weak correlations (e.g. first-names initials)
Interpretation in terms of social dynamic shaping gender distribution
Accuracy: No intrinsic methodological systematic effects

Incomplete information

- Using only initials or final letters



References

[1] Ross CO, Gupta A, Mehrabi N, Muric G, Lerman K. The Leaky Pipeline in Physics Publishing. arXiv:2010.08912 [physics.soc-ph].
[2] Squazzoni F, Bravo G, Farjam M, et al. Peer review and gender bias: A study on 145 scholarly journals. Sci. Adv. 2021;7:eabd0299.
[3] Dworkin JD. The extent and drivers of gender imbalance in neuroscience reference lists. Nat. Neurosci. 2020;23:918.
[4] Hu Y, Hu C, Tran T, Kasturi T, Joseph E, Gillingham M. *What's in a Name? – Gender Classification of Names with Character-Based Machine Learning Models*. arXiv:2102.03692 [cs.LG].