How many people do you know?: Efficiently estimating personal network size

Tian Zheng
Department of Statistics
Columbia University

April 22nd, 2009

Acknowledgements

- Collaborators
  - Tyler McCormick (Statistics, Columbia University)
  - Matt Salganik (Sociology, Princeton University)
- NSF for research support.

Social networks

- *Social Network Analysis* is the study of the structure of relationships between individuals.
  - Studied widely in the social sciences with gain in popularity in the physical sciences and in government.
- We concentrate on one aspect of social network analysis —estimating personal network sizes.

Why estimate personal network size?

- Personal network size, or *degree*, is the number of people known by a particular individual.
- Degree is a topic of interest in its own right to social scientists and it can also be useful in helping to explain other social phenomenon. For example:
  - Conley (2004) wanted to know whether siblings who knew more people tended to be more successful.
  - Study the dynamics of social processes such as spread of diseases and the evolution of group behavior.
- Few cases where it is practical to survey *all* actors in a network

Killworth, McCarty et al. 's "how many X's do you know" surveys

- McCarty et al (2001) Comparing two methods for estimating network size. Human Organization 60, 28-39.
- Telephone surveys of 1370 Americans.
- Responses from 32 questions of the form "how many X's do you know".
- Define "knowing someone": *you know them and they know you* **by sight or by name**, *that you could* **contact them**, *that they* **live within the United States**, *and that there has been* **some contact** *(either in person, by telephone or mail) in the past two years.* (Symmetric "knowing".)
- Example: how many people do you know who are currently homeless?

Killworth, McCarty et al. 's "how many X's do you know" surveys

- (Female names) Stephanie, Jacqueline, Kimberly, Nicole, Christina, Jennifer
- (Male names) Christopher, David, Anthony, Robert, James, Michael
- (Occupations) Commercial pilot, gun dealer, postal worker
- (Ethnicity) American Indian
- (Experiences) Twin, woman adopted kid in past year, gave birth in past year, widow(er) under 65, opened business in past year, homicide victim, suicide in past year, died in auto accident, diabetic, kidney dialysis, AIDS, HIV-positive, rape victim, male in prison, homeless, member of Jaycees

Previous attempts to estimate degree with "how many x's" data

The scale-up method (Killworth et al. 1998)

- Suppose you know 1 person named Nicole
- At the time of the survey, 358,000 out of 280 millions Americans are named Nicole
- Assume the 1 Nicole represents $0.13\% = \frac{358,000}{280,000,000}$ of your acquaintances
- Estimate: you know $1/0.0013 = 770$ people.

Assumptions held by the scale-up methods

- Killworth's method assumes that all people in the population are *equally likely* to know members of a given subpopulation. (i.e., *Random mixing in social space.*)
- It also assumes that respondents will be able to *accurately* recall their acquaintances during a survey.

Previous attempts to estimate degree with "how many x's" data

The Zheng et al (2006) Overdispersed model

- y_{ik} = number of persons in group k known by person i.
- $y_{ik} \sim \text{Poisson}(a_i b_k g_{ik})$
- For k, we let g_{ik} follow a gamma distribution with mean 1 and $1/(\omega_k - 1)$ as the shape parameter. Thus, $y_{ik} \sim \text{Negative Binomial}(\text{mean} = a_i b_k, \text{overdispersion} = \omega_k)$.
- $a_i = e^{\alpha_i}$, "gregariousness" of person i—degree.
- $b_k = e^{\beta_k}$, size of group k in the social network
- ω_k is overdispersion of group k
  - $\omega_k = 1$ is no overdispersion (Poisson model)
  - Higher values of ω_k show overdispersion
- Overdispersion represents social structure

Zheng et al (2006) overdispersed model

- Model the overall extent of non-random mixing (social structure) but didn't correct the degree estimation.
- Didn't address the issue of imperfect recall.

Three sources of errors for degree estimation

- Barrier effect: Non-random mixing is common.
- Transmission errors: the possibility that a respondent knows someone in a category without realizing it.
  - Solution: we use first names as our X's to eliminate transmission errors.
- Imperfect recall.

Non-random mixing

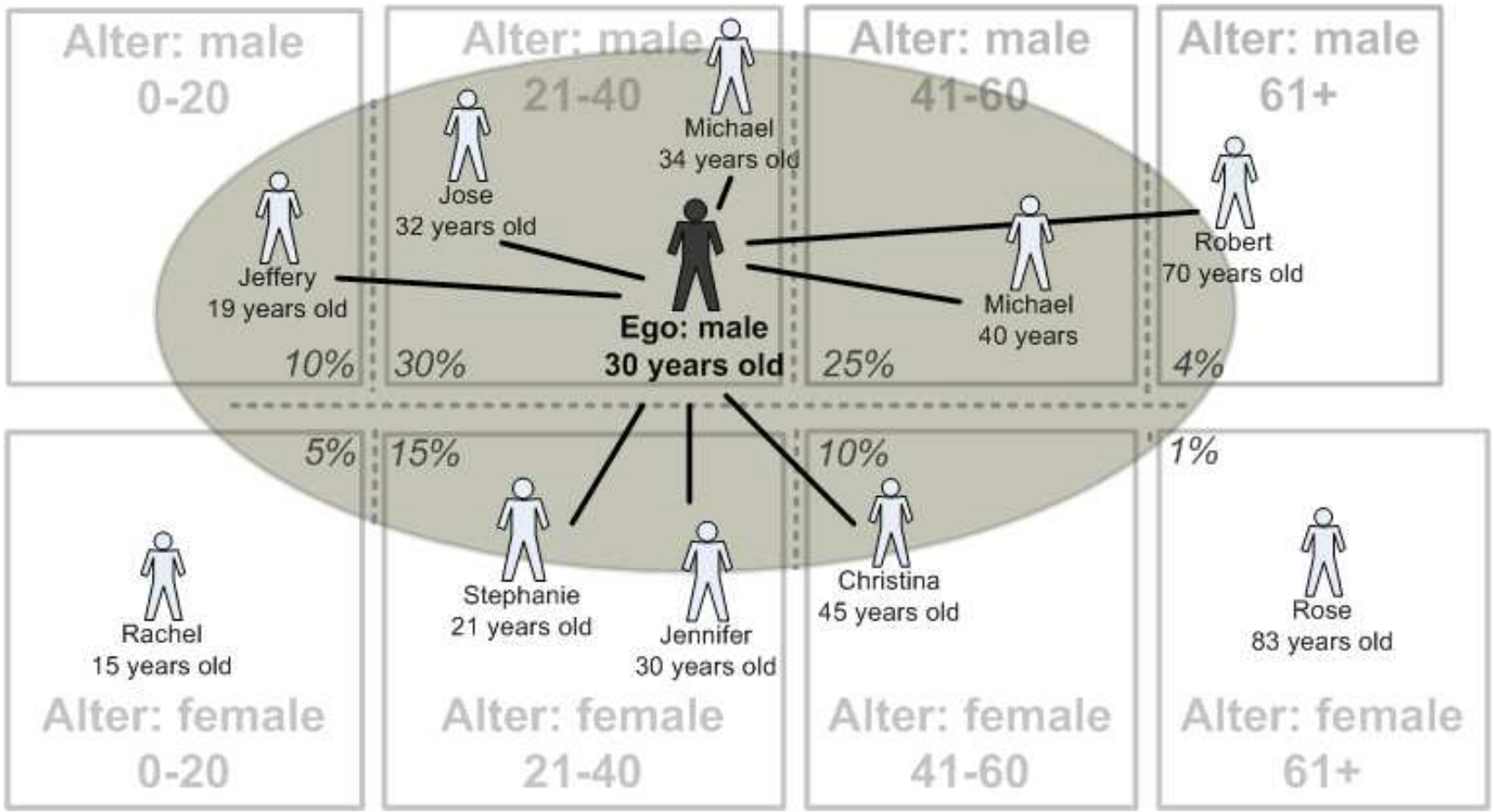


Figure: An acquaintanceship network.

Biases in the degree estimates based on first names due to non-Random mixing

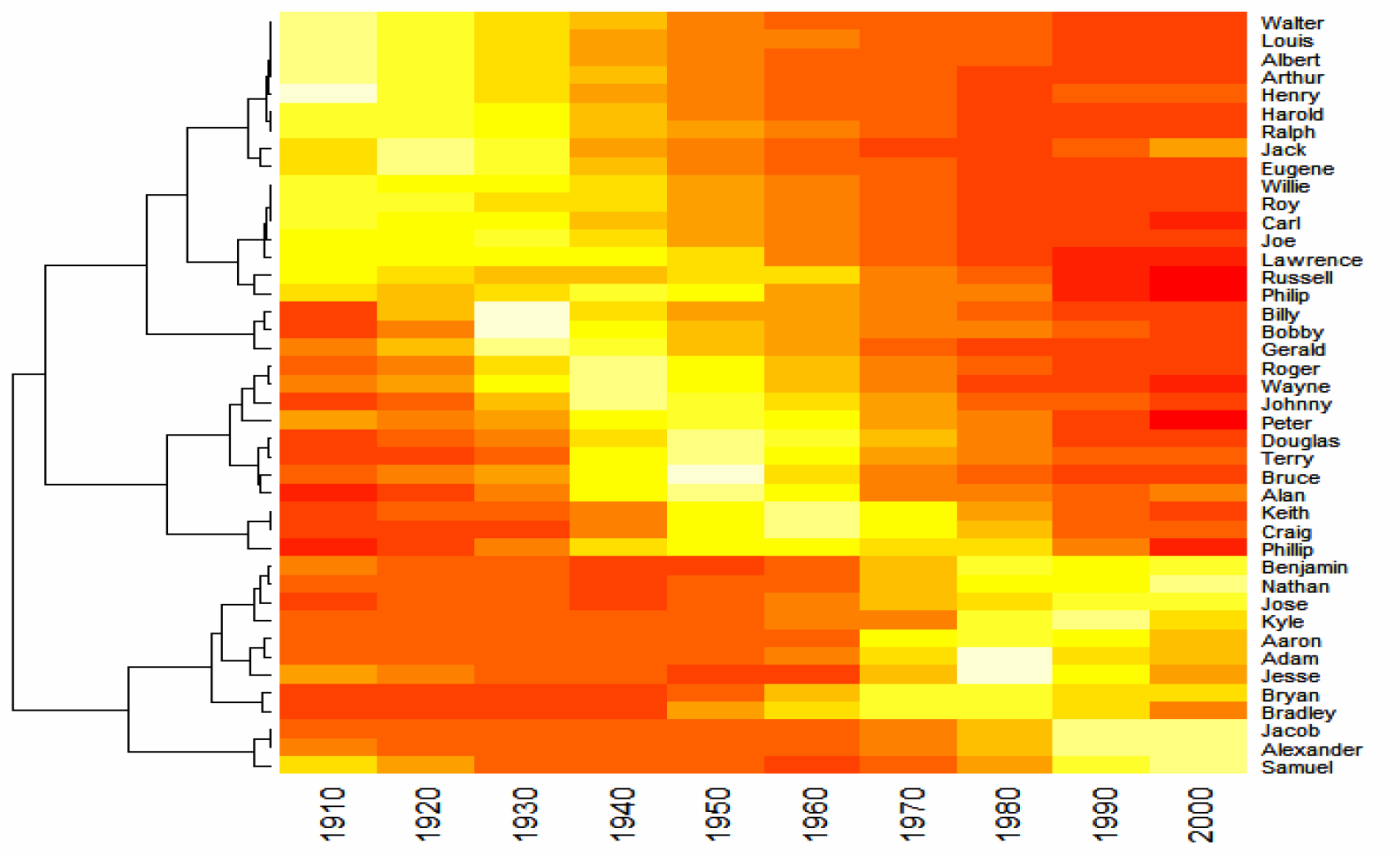


Figure: Age distributions of male names.

Effects of the under-recall of large subpopulations

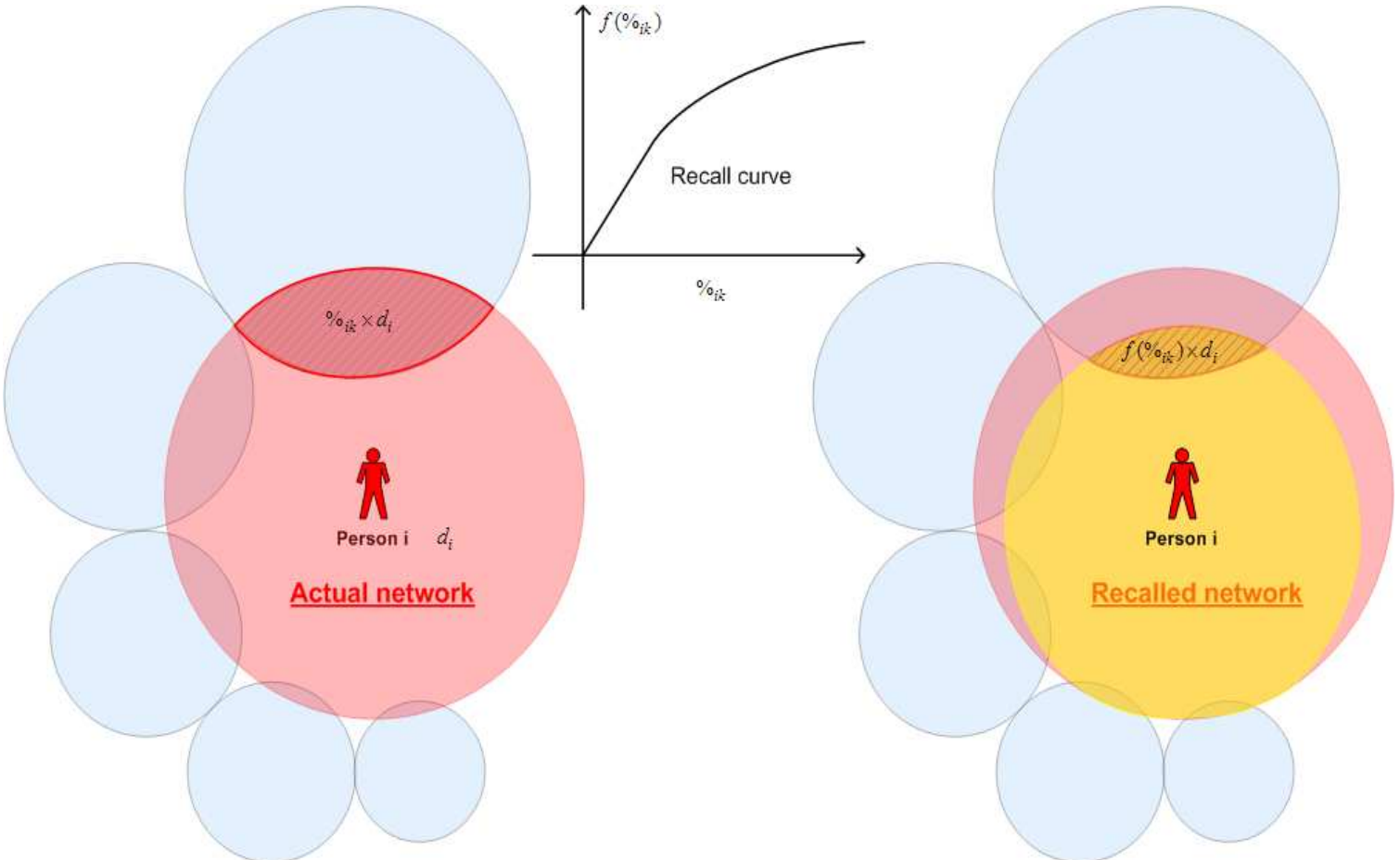


Figure: The calibration curve.

Effects of the under-recall of large subpopulations

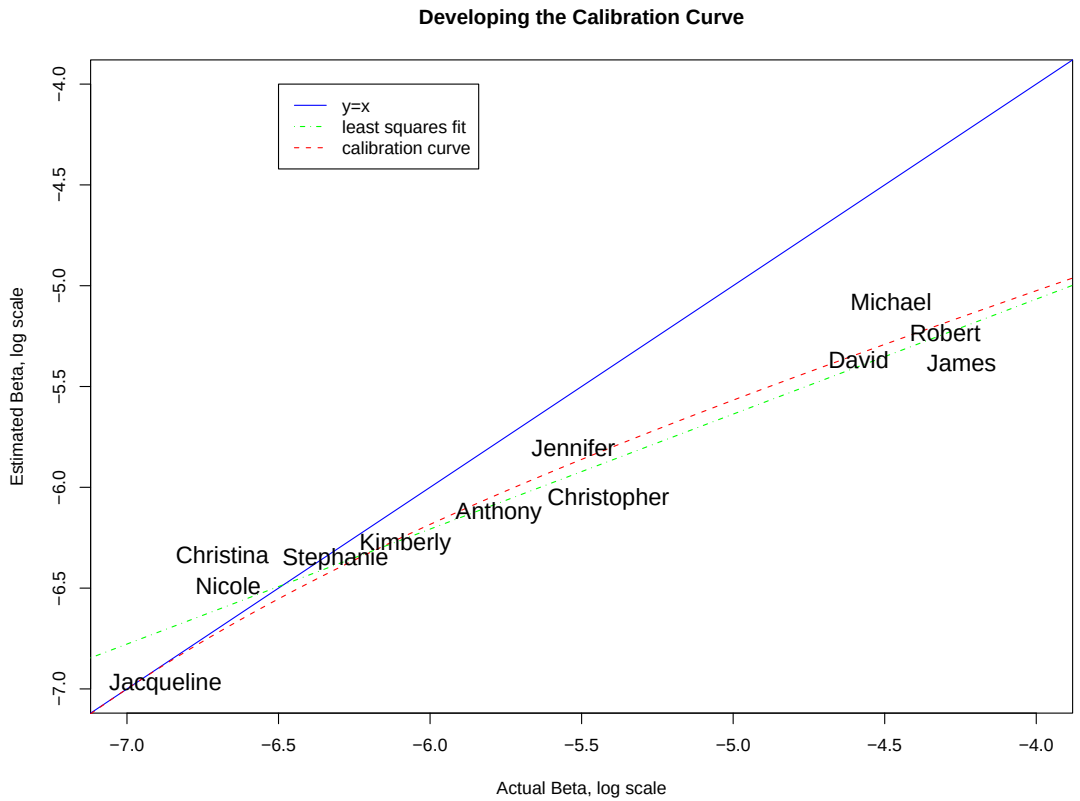


Figure: Twelve names without the calibration curve.

Three sources of errors for degree estimation

- Barrier effect: Non-random mixing is common.
- Transmission errors: the possibility that a respondent knows someone in a category without realizing it.
  - Solution: we use first names as our X's to eliminate transmission errors.
- Imperfect recall.
  - Solution 1: Use calibration curve in the model to adjust for under-recall of large groups.
  - Solution 2: Use first names that are about 0.1-0.2 percent of the population, where the calibration curves is close to the $y = x$ line.

Correcting for Under-Recall: calibration curve $f(\cdot)$

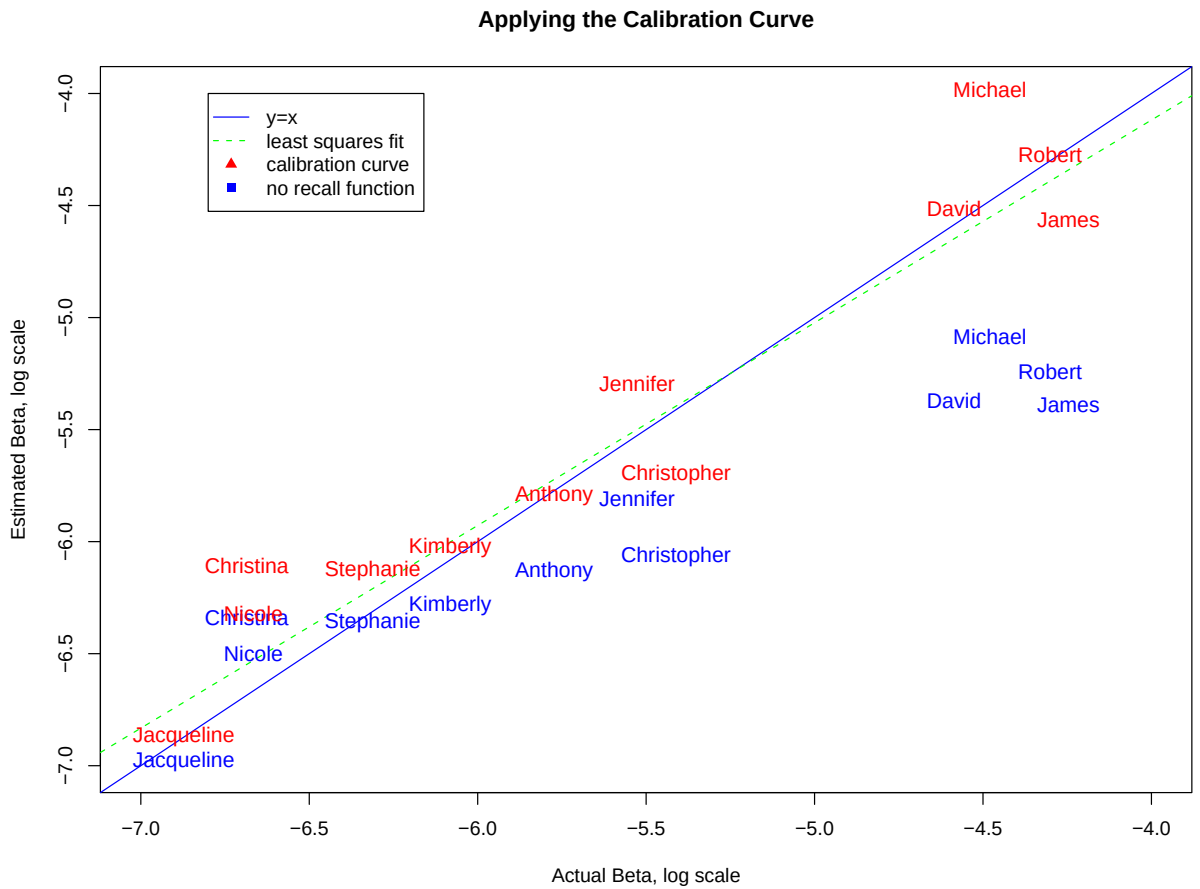


Figure: The Calibration Curve.

The latent non-random mixing model

Modeling goals and data

- account for non-random mixing due to gender/age
- We use data from McCarty et al. (2001). There are 1375 respondents and twelve names.

Modeling "How many X do you know" with non-random mixing and imperfect recall

$$y_{ik} \sim \text{Neg-Binom}(\mu_{ike}, \omega_k),$$

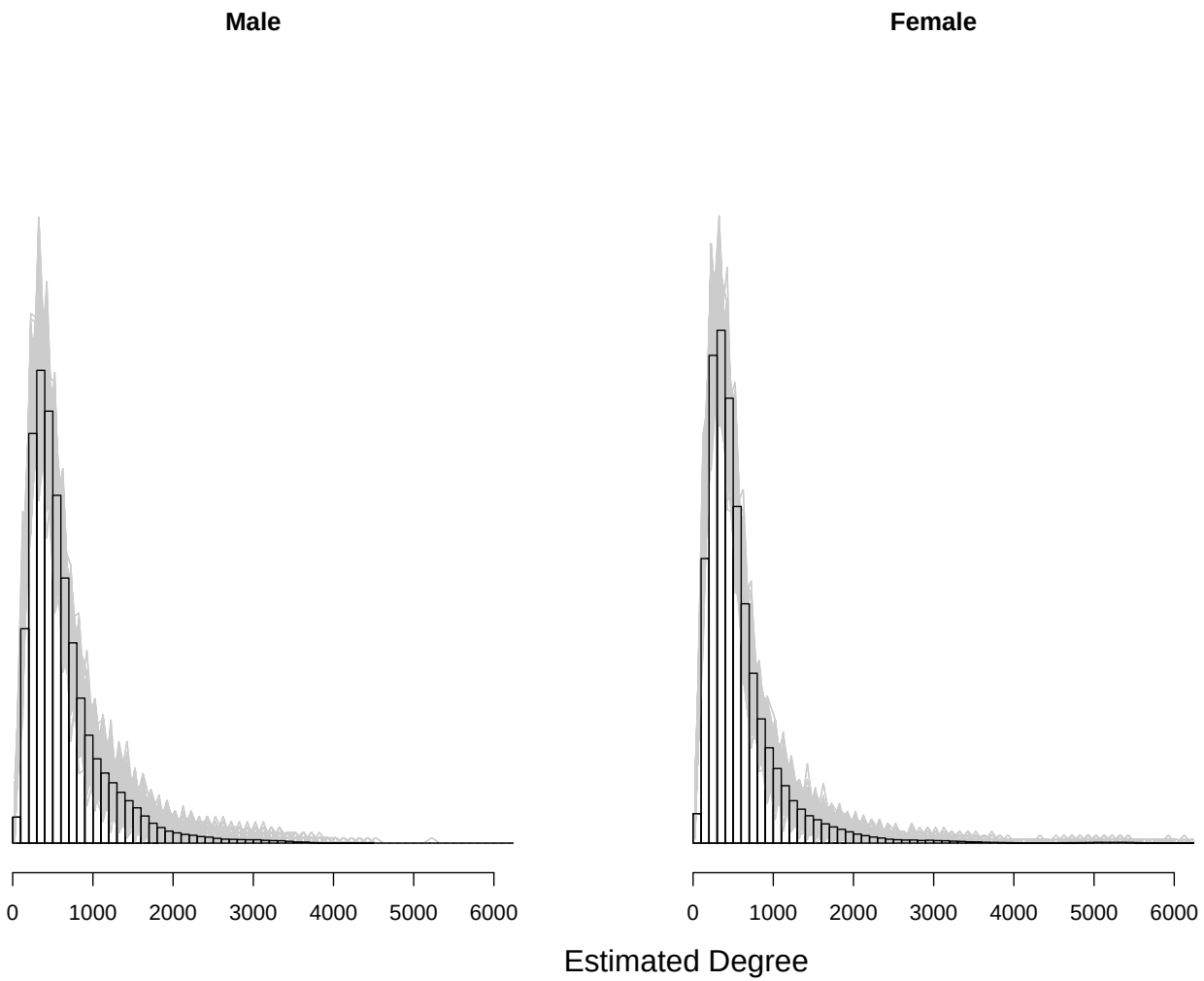
where $\mu_{ike} = d_i f\left(\sum_{a=1}^{A} m(e,a)\,\frac{N_{ak}}{N_a}\right)$.

- d_i is the degree of person i
- $m(e,a)$ is a matrix of mixing coefficients, the proportion of respondent group e's network made up of alter group a
  - We use 8 alter groups (4 age categories and gender) and 6 ego groups (3 age categories and gender).
- $\frac{N_{ak}}{N_a}$ is the proportion of alter group a made up of people with name k and is assumed known (available from SSA).
- ω_k is the overdispersion
- $f()$ is the calibration curve.

Priors and Hyperparameters

- Parameters are estimated using a multilevel model and Bayesian inference.
- The network size and mixing coefficients are modeled with log-normal distributions.
- This parameterizations is based on previous research about the degree distribution of of the acquaintanceship network (see McCarty *et al.* 2001).
- Hyperparameters are assigned noninformative uniform priors.

Results-Estimated Degree



Results-Estimated Degree

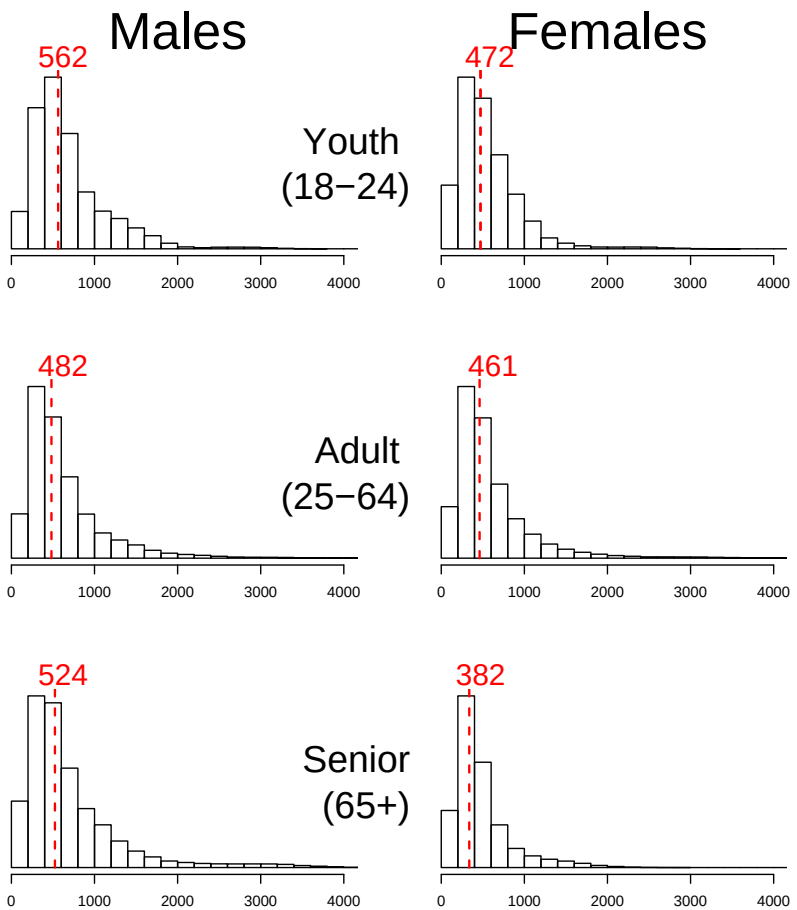


Figure: Histograms of the estimated degree distribution. Red lines show the estimated median.

Results-Mixing Matrix

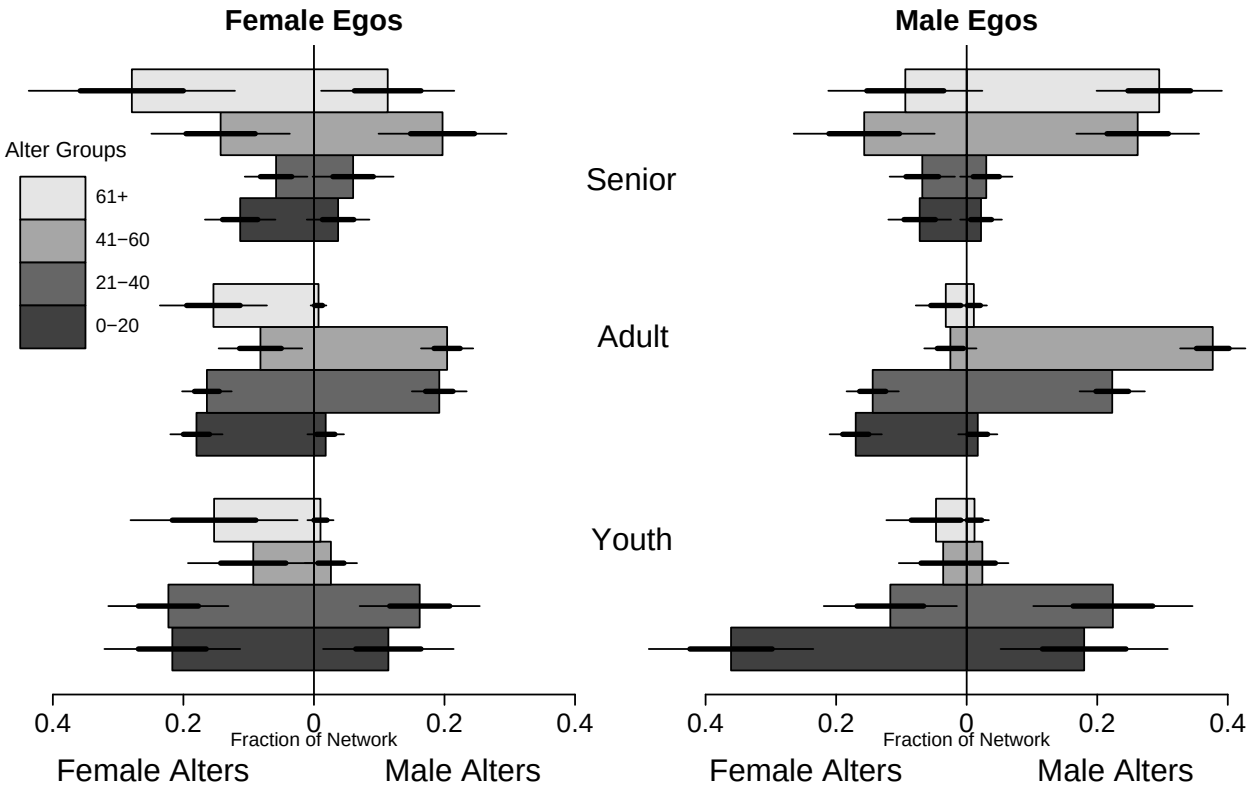


Figure: The mixing matrix estimates +/- 2 standard errors.

Three sources of errors for degree estimation

- Barrier effect: Non-random mixing is common.
  - Solution: the latent non-random mixing model.
- Transmission errors: the possibility that a respondent knows someone in a category without realizing it.
  - Solution: we use first names as our X's to eliminate transmission errors.
- Imperfect recall.
  - Solution 1: Use calibration curve in the model to adjust for under-recall of large groups.
  - Solution 2: Use first names that are about 0.1-0.2 percent of the population, where the calibration curves is close to the $y = x$ line.

Simple Estimates of Degree

- Our model addresses
  - non-random mixing
  - recall bias
- Our full model requires some expertise and time to fit.
- A simple design strategy would allow social scientists to obtain scale-up degree estimates that is equivalent to that of our method.

Deriving the Simple Estimate

Recall the expectation of our model:

$$\mu_{ike} = E(y_{ike}) = d_i \sum_{a=1}^{A} m(r,a) \frac{N_{ak}}{N_a}.$$

We the proceed by taking the sum over all K names.

$$\mu_{i\cdot e} = E\left(\sum_{k=1}^{K} y_{ike}\right) = \sum_{k=1}^{K}\left[d_i \sum_{a=1}^{A} m(r,a)\frac{N_{ak}}{N_a}\right]$$

$$= d_i \sum_{a=1}^{A} m(r,a)\left[\sum_{k=1}^{K}\frac{N_{ak}}{N_a}\right]$$

$$\left(\text{if } \sum_{k=1}^{K}\frac{N_{ak}}{N_a} = c, \text{ a constant}\right) = d_i c \sum_{a=1}^{A} m(r,a) = d_i c$$

The scale-down condition for unbiased scale-up estimates.

- The *scale-down* condition: the names are selected so that

$$\sum_{k=1}^{K} \frac{N_{ak}}{N_a} = constant.$$

- That is the names are balanced and have the same distribution on the demographic variables that define the alter groups.

Selecting Names

Suggestions for Selecting Subpopulations (X's):

1. Using first names eliminates *barrier* and *transmission* effects
2. Using rare names (suggest 0.1–0.2 percent of the population) minimizes recall bias
3. Select names with complimentary age profiles

Selecting Complimentary Name-Age Profiles

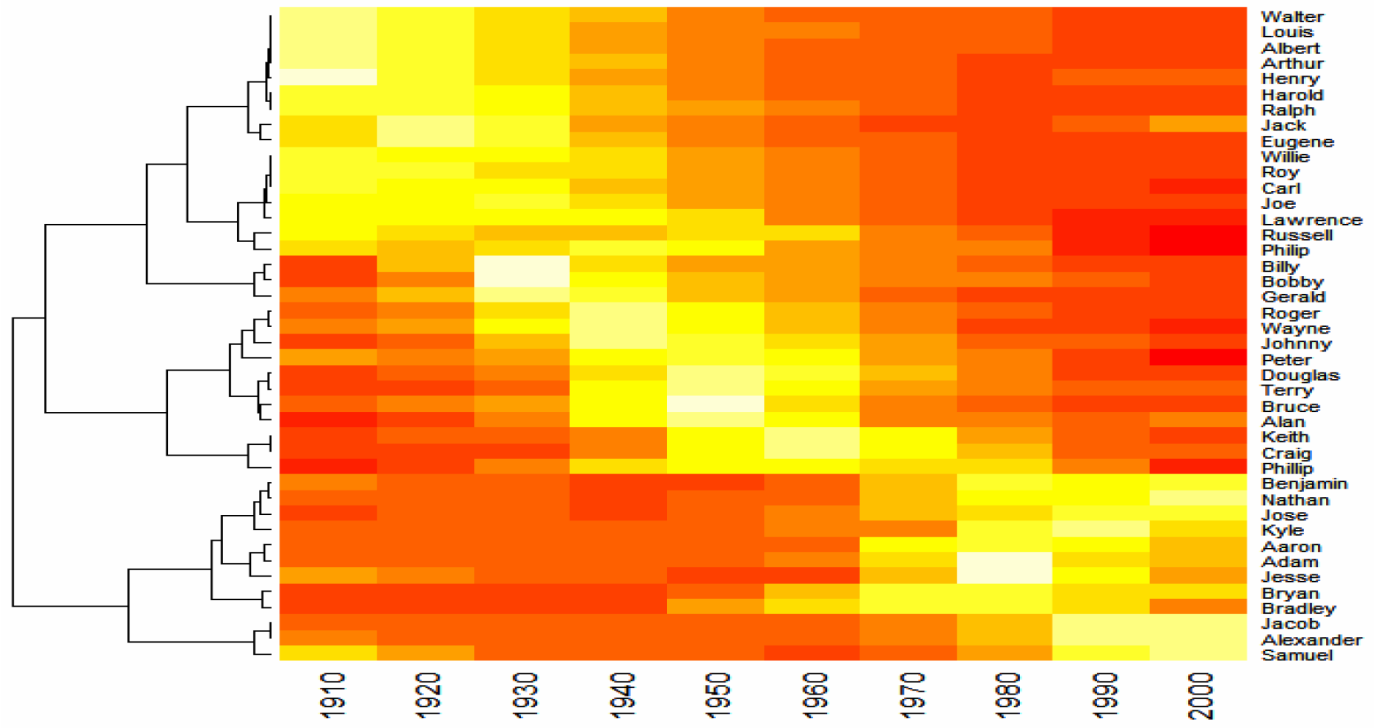


Figure: Selecting complementary name-age profiles.

Simulated Data Experiment

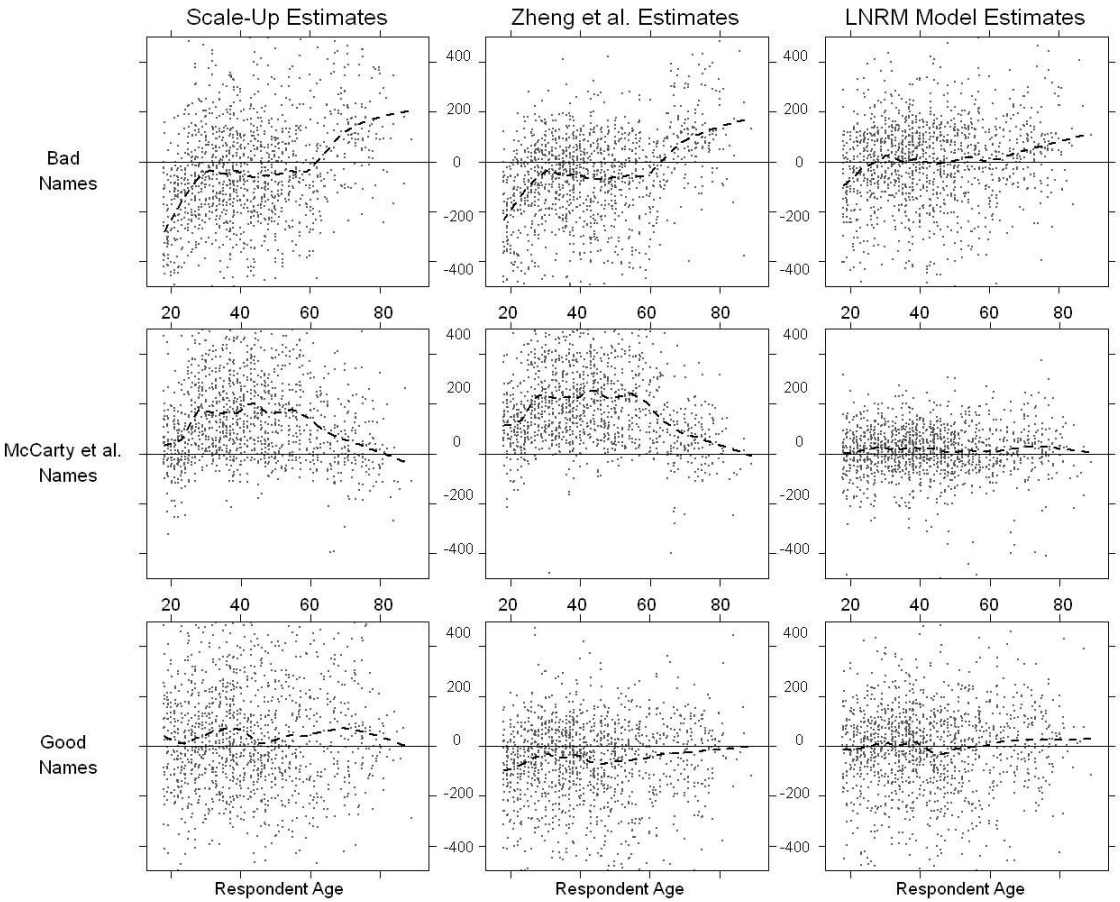


Figure: Comparison of full and simple model performance.

Three sources of errors for degree estimation

- Barrier effect: Non-random mixing is common.
  - Solution 1: the latent non-random mixing model.
  - Solution 2: select names that satisfy the scale-down conditions.
- Transmission errors: the possibility that a respondent knows someone in a category without realizing it.
  - Solution: we use first names as our X's to eliminate transmission errors.
- Imperfect recall.
  - Solution 1: Use calibration curve in the model to adjust for under-recall of large groups.
  - Solution 2: Use first names that are about 0.1-0.2 percent of the population, where the calibration curves is close to the $y = x$ line.

Limitations and Future Directions

- We account for non-random mixing from age and gender, but other factors could still be present. The census bureau collects other information on first names but does not release such data.
- Behavior of the calibration curve $f()$ on larger groups is not well established.

Conclusions

- We propose a model to estimate network size based on aggregated count data.
- This model accounts for non-random mixing based on gender and age, as well as corrects for under-recalling.
- We also show that the 'scale-up' approach is a reasonable estimate to our full model under given conditions.

THANK YOU!