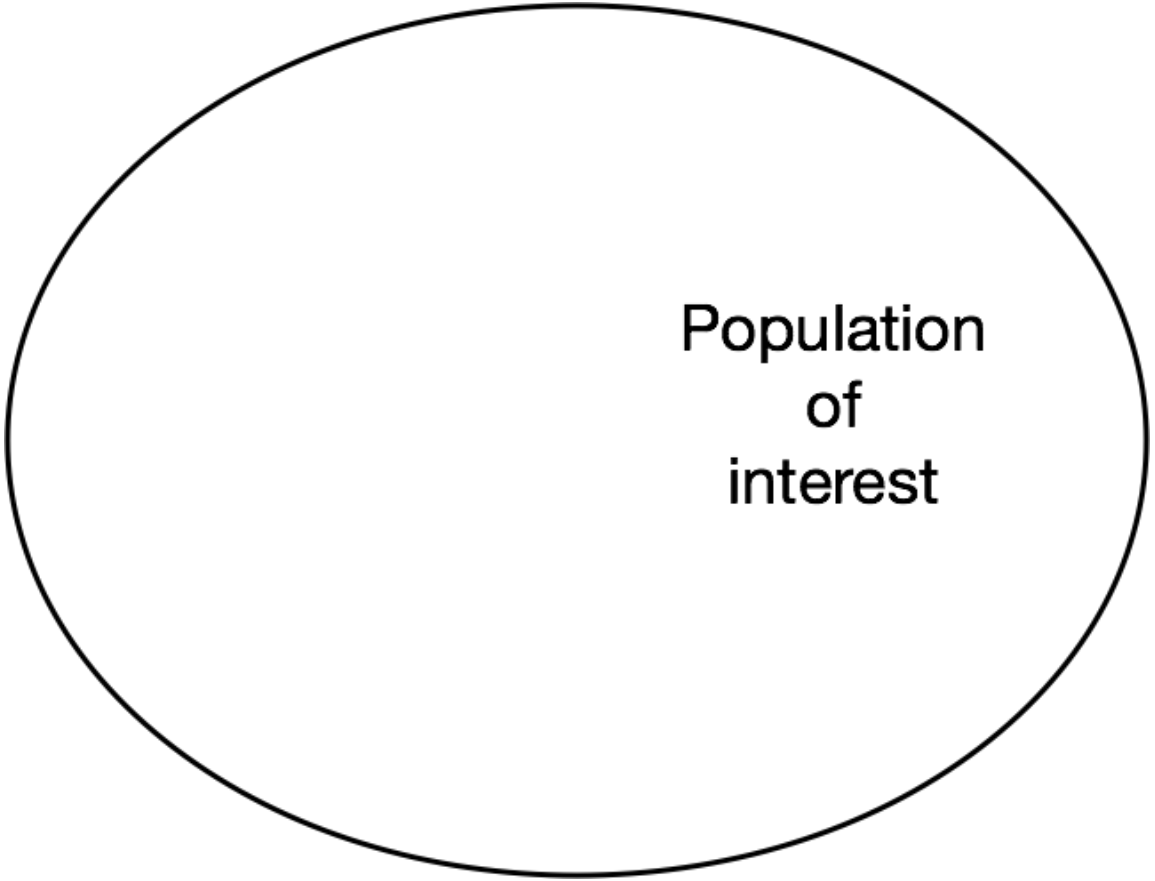
Estimating internet adoption around the world using a sample of Facebook users

PAA 2018

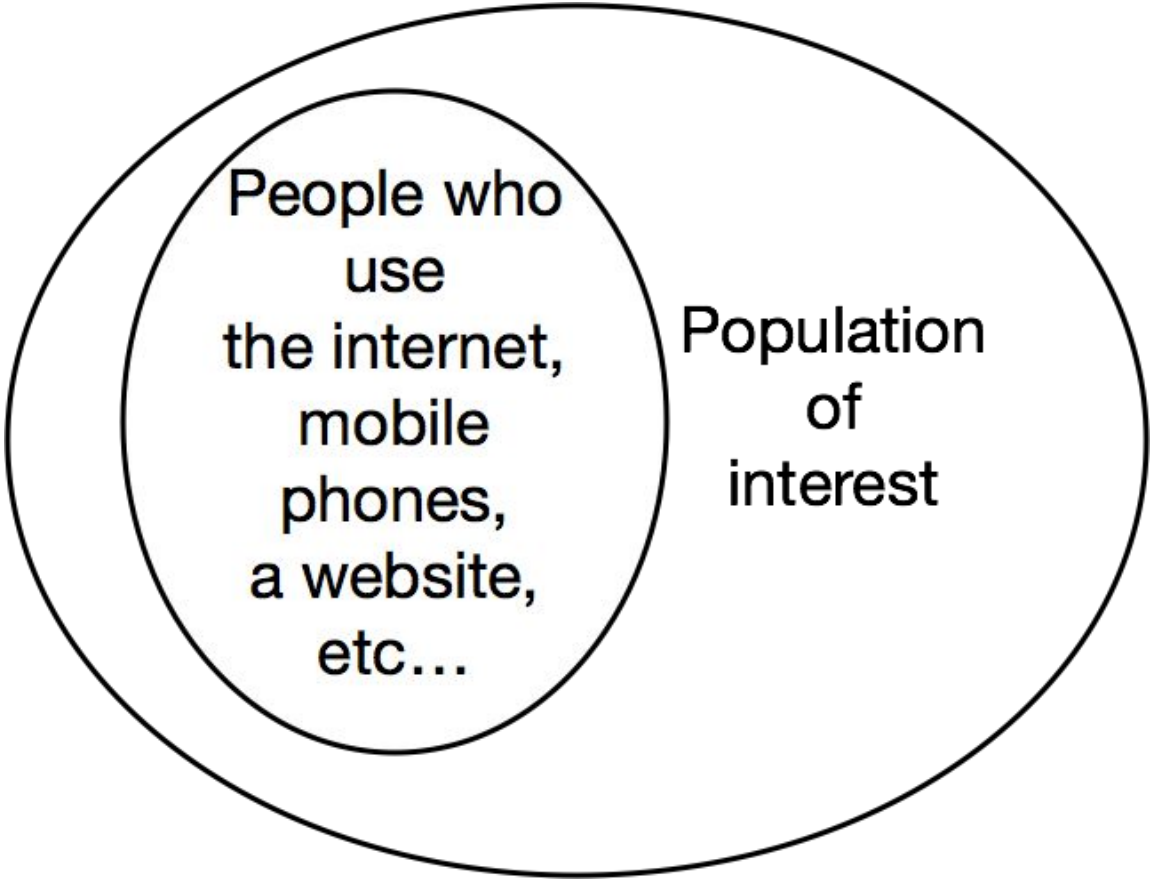
Dennis M. Feehan
UC Berkeley

Curtiss Cobb
Facebook

The problem



The problem


People who use the internet, mobile phones, a website, etc…
Population of interest

The problem

Digital inequalities and why they matter

Laura Robinson[a], Shelia R. Cotten[b], Hiroshi Ono[c], Anabel
Wenhong Chen[f], Jeremy Schulz[g*], Timothy M. Hale[h] and

*[a]Department of Sociology, Santa Clara University, Santa Clara, C
Information, Michigan State University, East Lansing, MI, USA; [c]H
School of International Corporate Strategy, Hitotsubashi Universit
and Media Studies/Sociology, The University of Western Ontario,
nthropology, University
, Austin, TX, USA; [g]Insti
Connected Health, Parti
er for Excellence in Surv*

The Arrival of Fast Internet and Skilled Job Creation in Africa*

Jonas Hjort
Columbia University
& BREAD & NBER

Jonas Poulsen
Harvard University

September 10, 2016

January 2015)

continues to expand in
ther forms of inequality
in and outside the field
ostantive problem and as
on multiple aspects
age, skills, and self-perc
s makes the case that di
of inequality in the t
ould not be only the pr
cial scientists concerned
e trajectories. As we a
ad range of individual-
e, and class, as well as

nmunication; digital divi

ng its current frontier in devel-
rican coastal cities of submarine
etwork that connects those cities
covering 14 countries show large
fset by a bigger increase in jobs
etween more and less educated
s to investigate how higher aver-
ob creation. We find an increase
l the productivity of workers in
ry in South Africa in the sectors
comes available, and (iii) work-
ren countries. Finally, we show
mes. Our findings shed light on
b creation, structural change, job

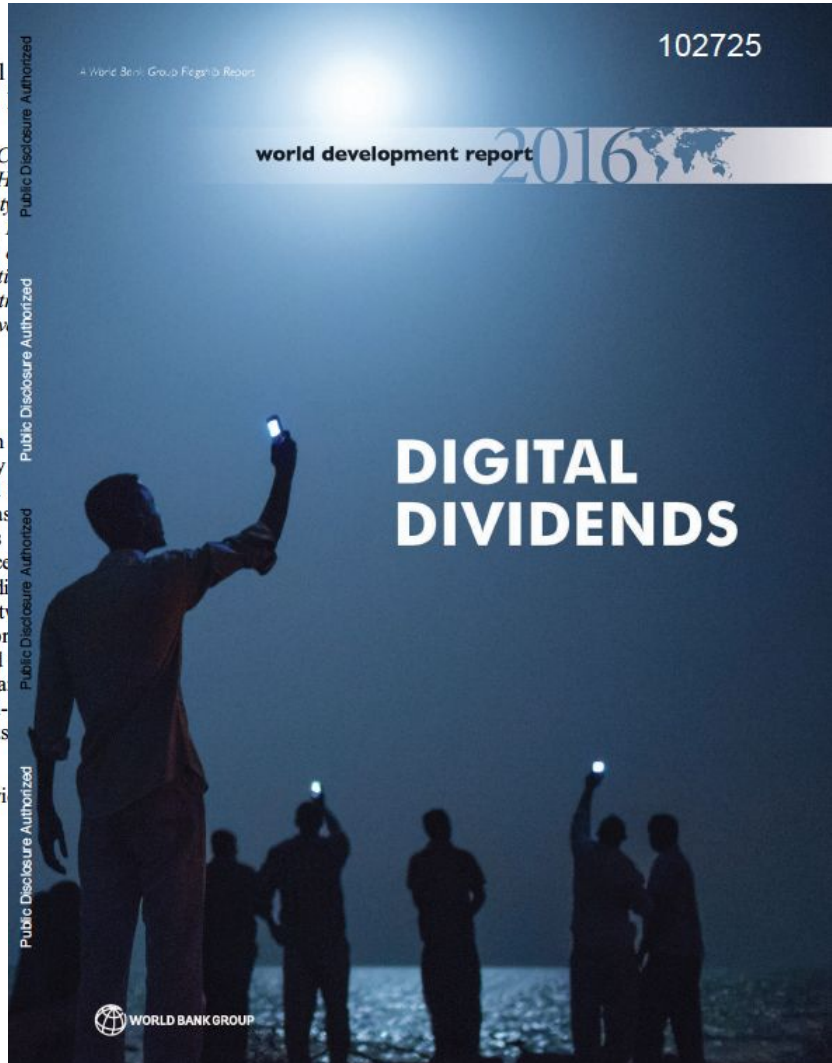
Chinn, Menzie D.; Fairlie, Robert W.

Working Paper

The Determinants of the Global Digital Divide : A Cross-Country Analysis of Computer and Internet Penetration

IZA Discussion paper series, No. 1305

Provided in Cooperation with:
Institute for the Study of Labor (IZA)



The plan

- Methods: network reporting with an online sample
- Study design: estimating internet adoption in 5 countries
- Results: estimates and sensitivity
- Next steps

Methods: network reporting

Methods: network reporting

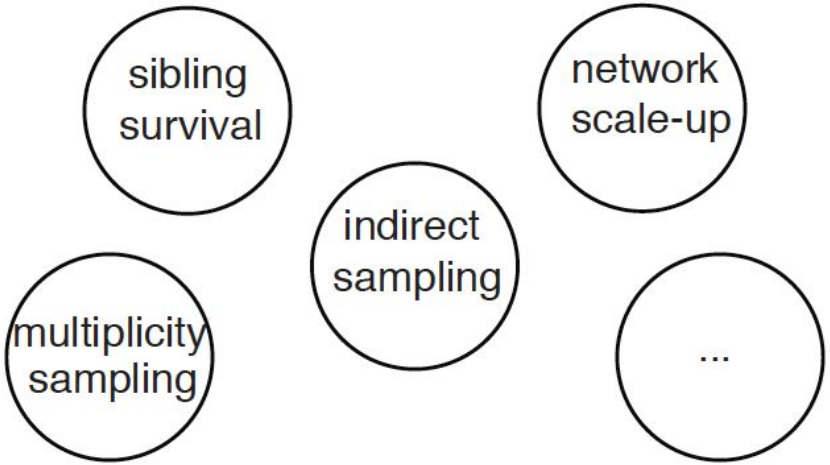
The idea: survey respondents are connected to other people through many different kinds of personal networks

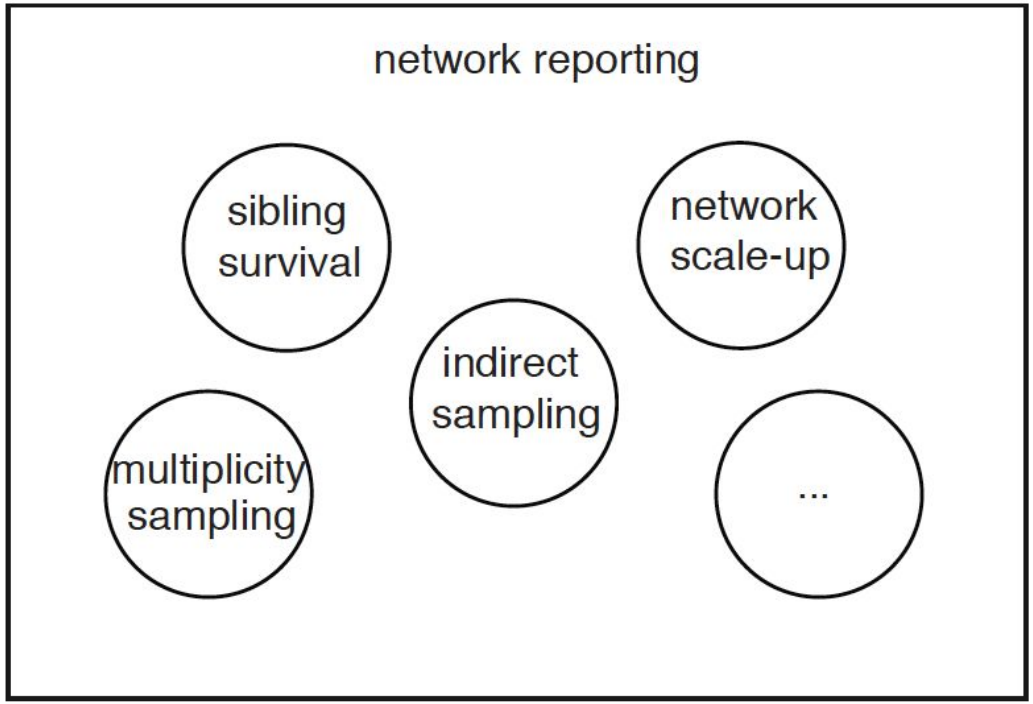
We can ask respondents questions about their personal network and learn about more than just the respondent.

Network reporting

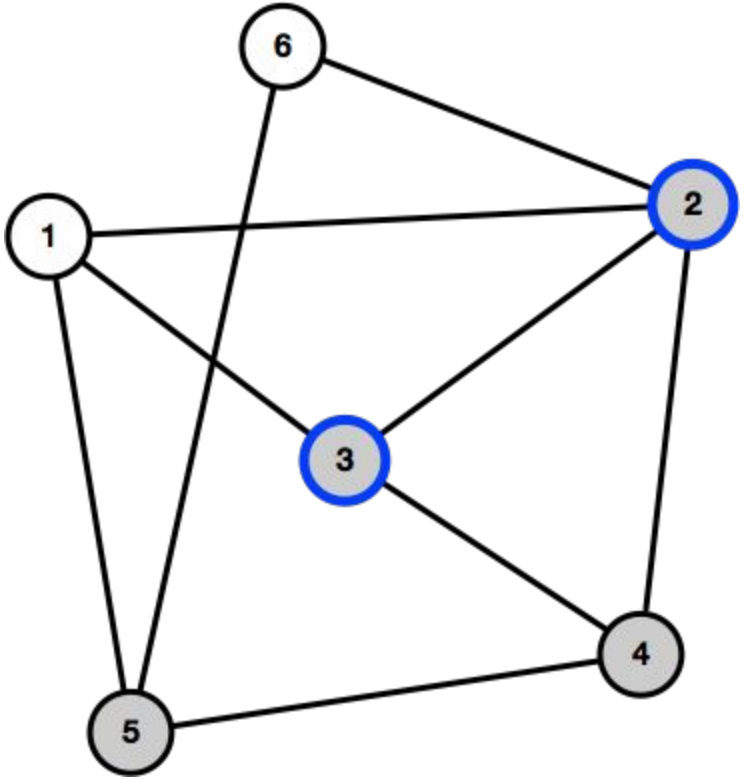
Approaches like this have been used in lots of different situations

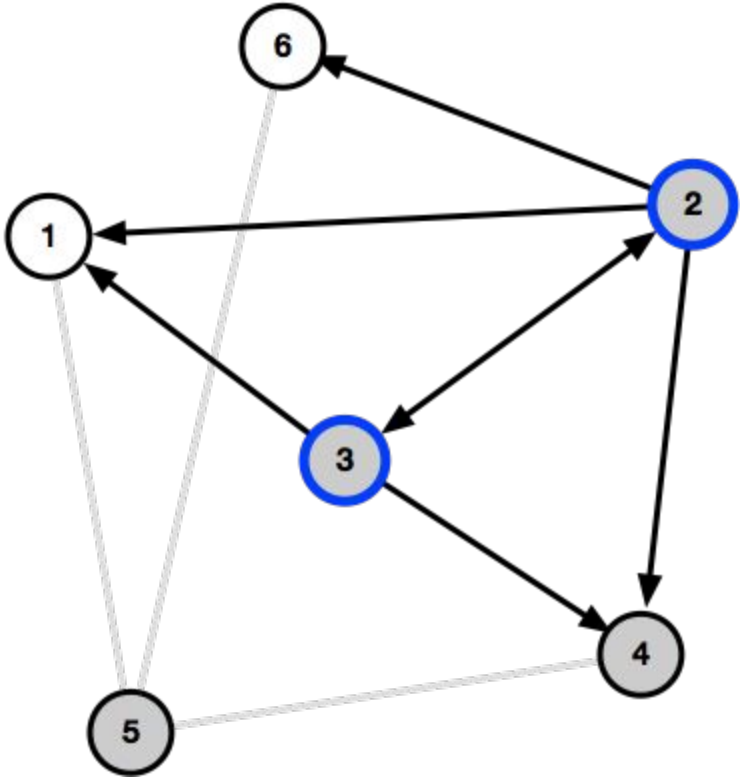
- Deaths
- Epidemiologically important groups (drug injectors, sex workers)
- Migrants
- … and many others

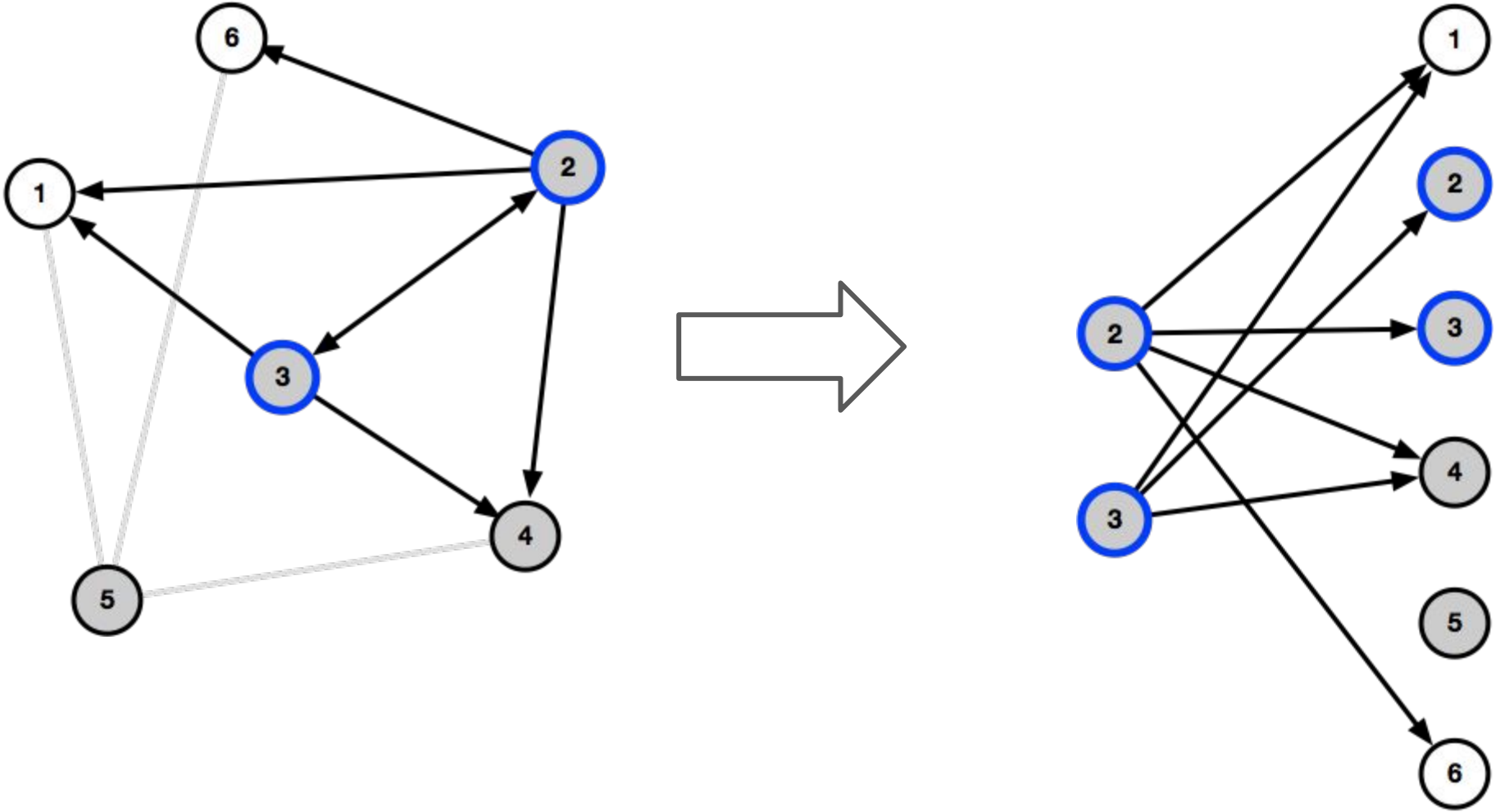

sibling survival
network scale-up
indirect sampling
multiplicity sampling
...


network reporting
sibling survival
network scale-up
indirect sampling
multiplicity sampling
...

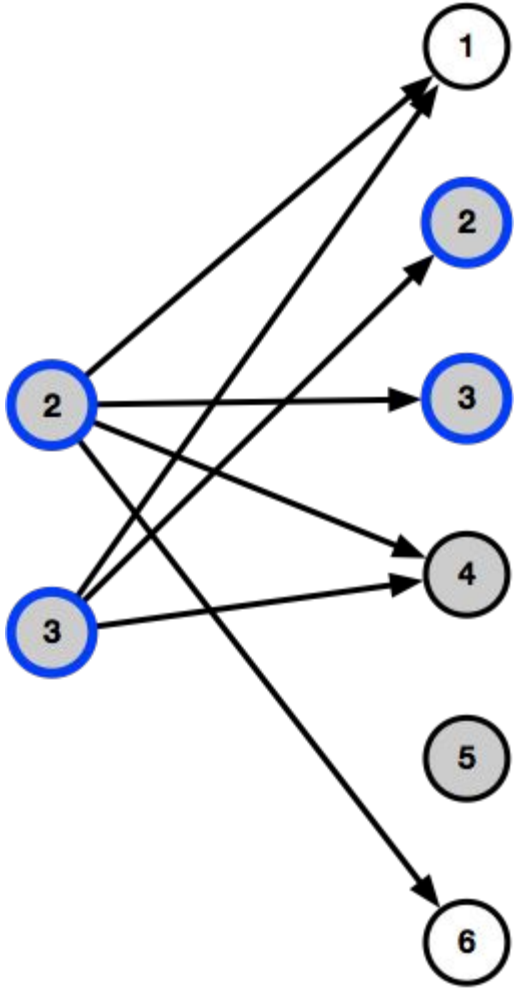
How it works


6
2
1
3
4
5


6
2
1
3
4
5


6
1
2
3
4
5
1
2
3
4
5
6

$$\begin{aligned}
&\text{total out-reports} = \text{total in-reports} \\
\iff &\text{total out-reports} = (\text{number of internet users} \times \\
&\qquad \text{average in-reports per internet user}) \\
\iff &\text{number of internet users} = \frac{\text{total out-reports}}{\text{average in-reports per internet user}}
\end{aligned}$$



$$\#\text{ of internet users} = \frac{\text{total reported connections to internet users}}{\text{average in-reports per internet user}}$$

Study design

$$\#\text{ of internet users} = \frac{\text{total reported connections to internet users}}{\text{average in-reports per internet user}}$$

$$\#\text{ of internet users} = \frac{\text{total reported connections to internet users}}{\text{average in-reports per internet user}}$$

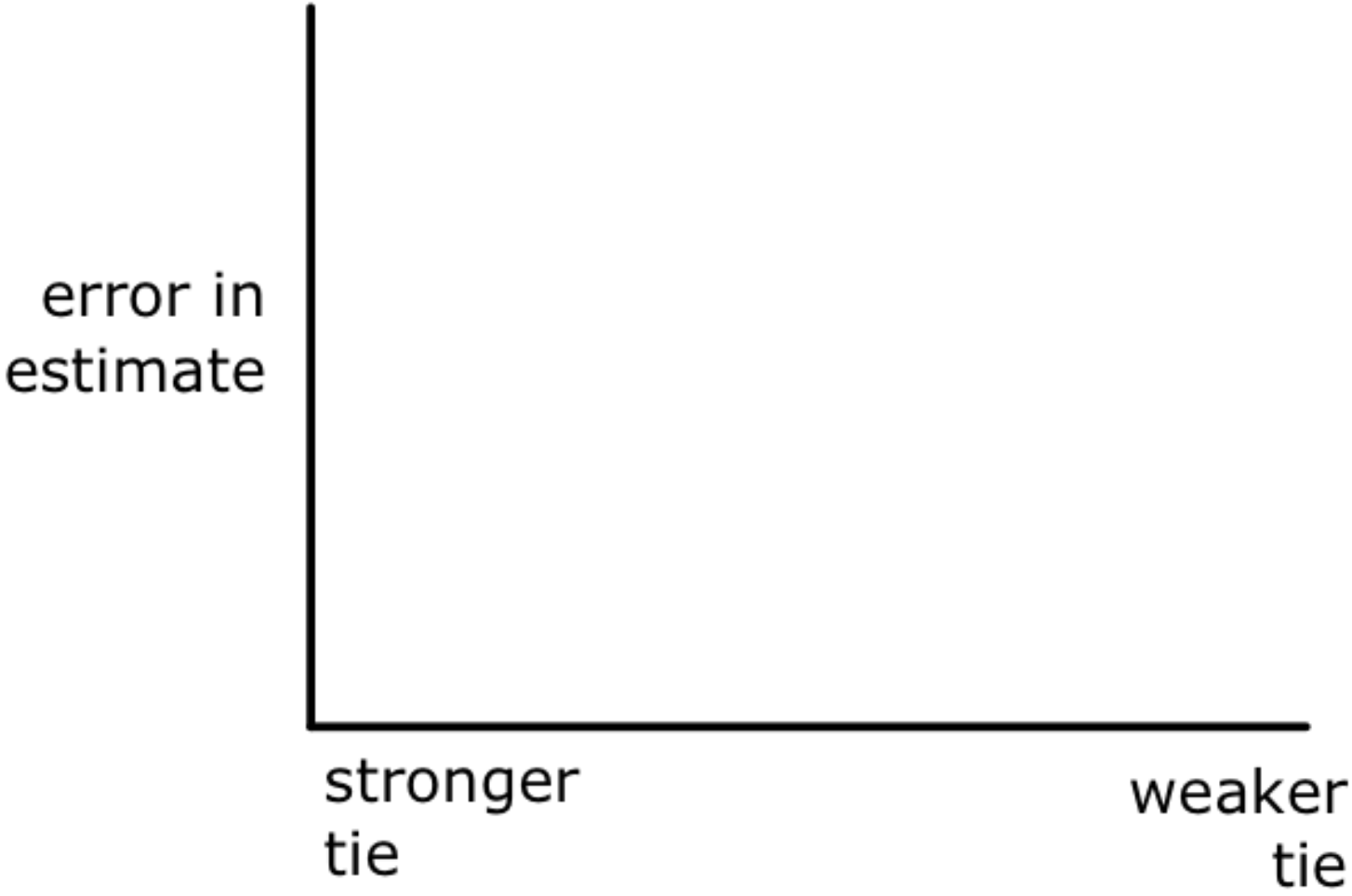
We can ask respondents questions like “how many people are in your network?”

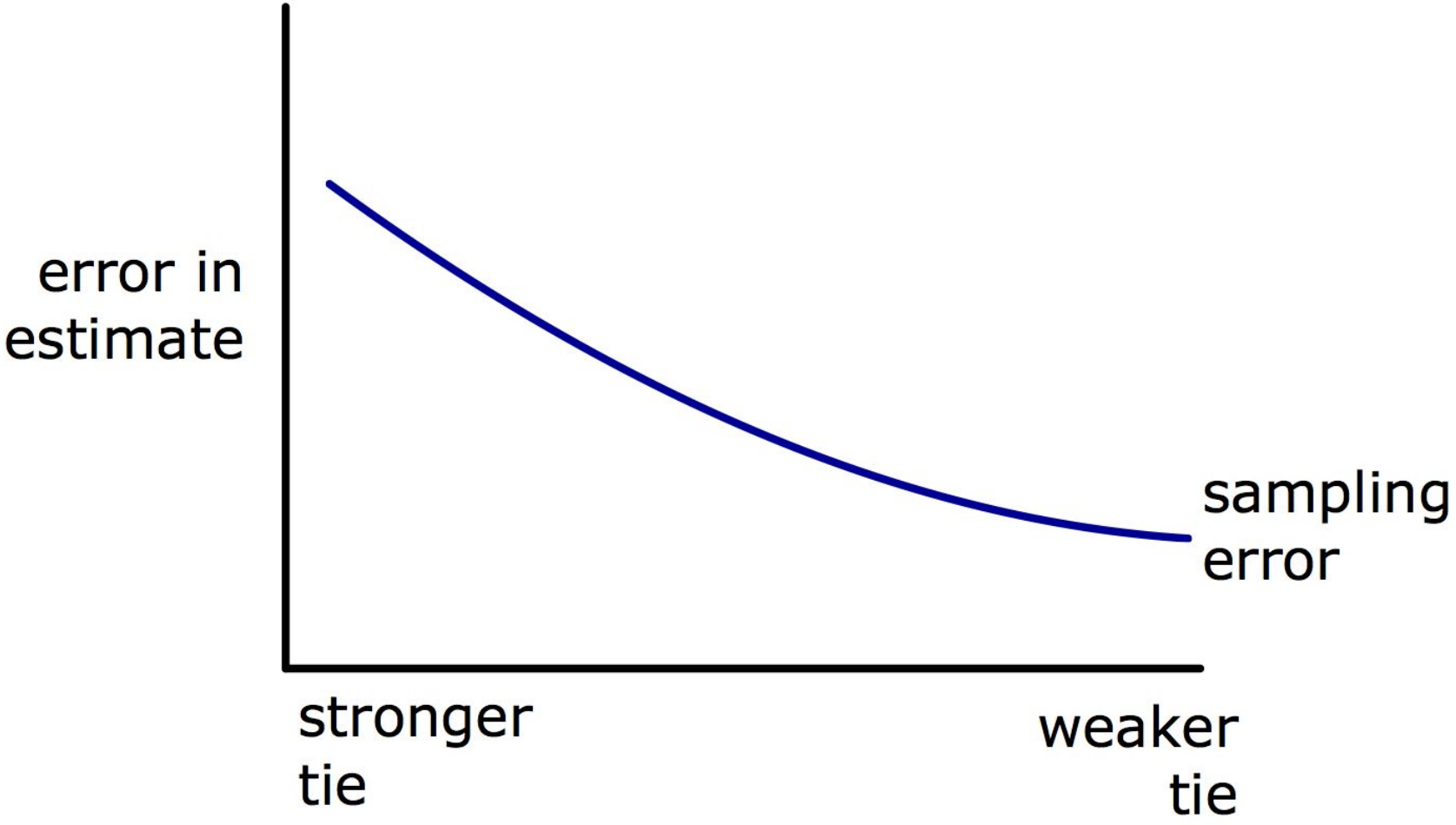
And then, “which of these people uses the internet?”

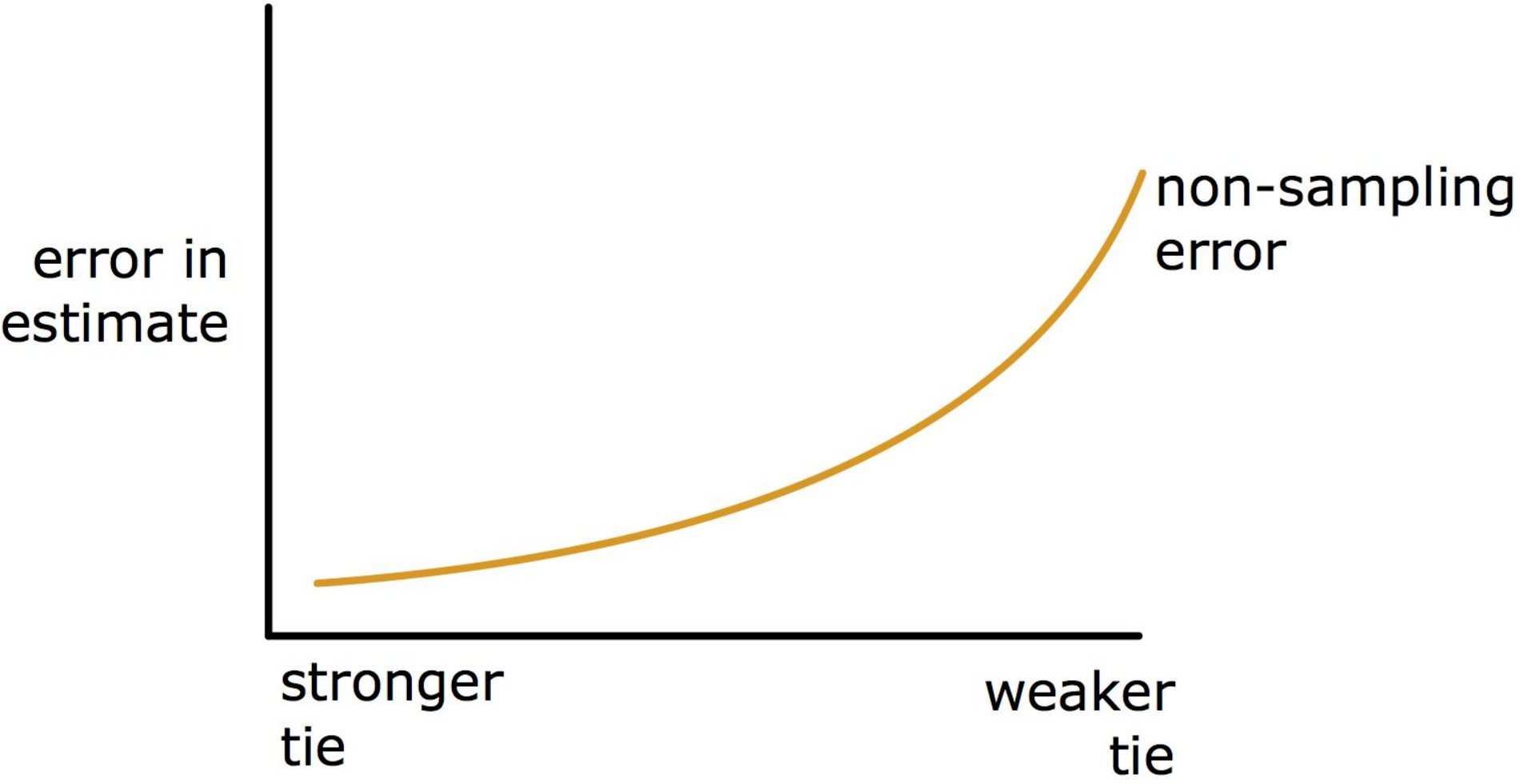
We can ask respondents questions like “how many people are in your network?”

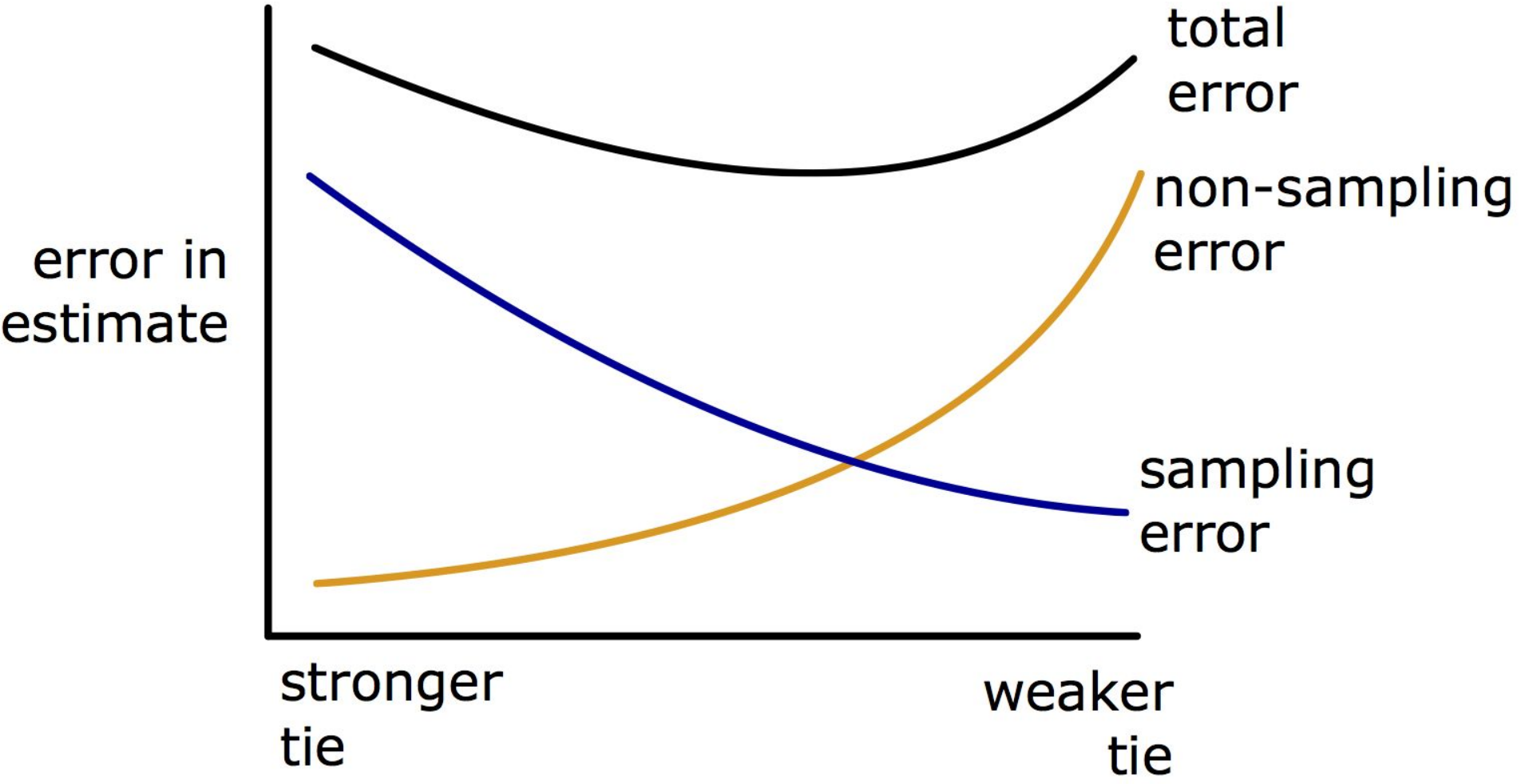
… but what does it mean to ‘know’ someone?

=> we need to choose a **tie definition**


error in
estimate
stronger
tie
weaker
tie


error in estimate
sampling error
stronger tie
weaker tie


error in estimate
non-sampling error
stronger tie
weaker tie


total
error
non-sampling
error
error in
estimate
sampling
error
stronger
tie
weaker
tie

Tie definition: survey experiment

- Previous research has found some evidence of a tie strength / accuracy tradeoff
- We designed an experiment to further test this question in our setting

Tie definition: survey experiment

- Previous research has found some evidence of a tie strength / accuracy tradeoff
- We designed an experiment to further test this question in our setting

Conversational Contact Network

- How many people did you have conversational contact with yesterday? By conversational contact, we mean anyone you spoke with face to face for at least three words.

Tie definition: survey experiment

- Previous research has found some evidence of a tie strength / accuracy tradeoff
- We designed an experiment to further test this question in our setting

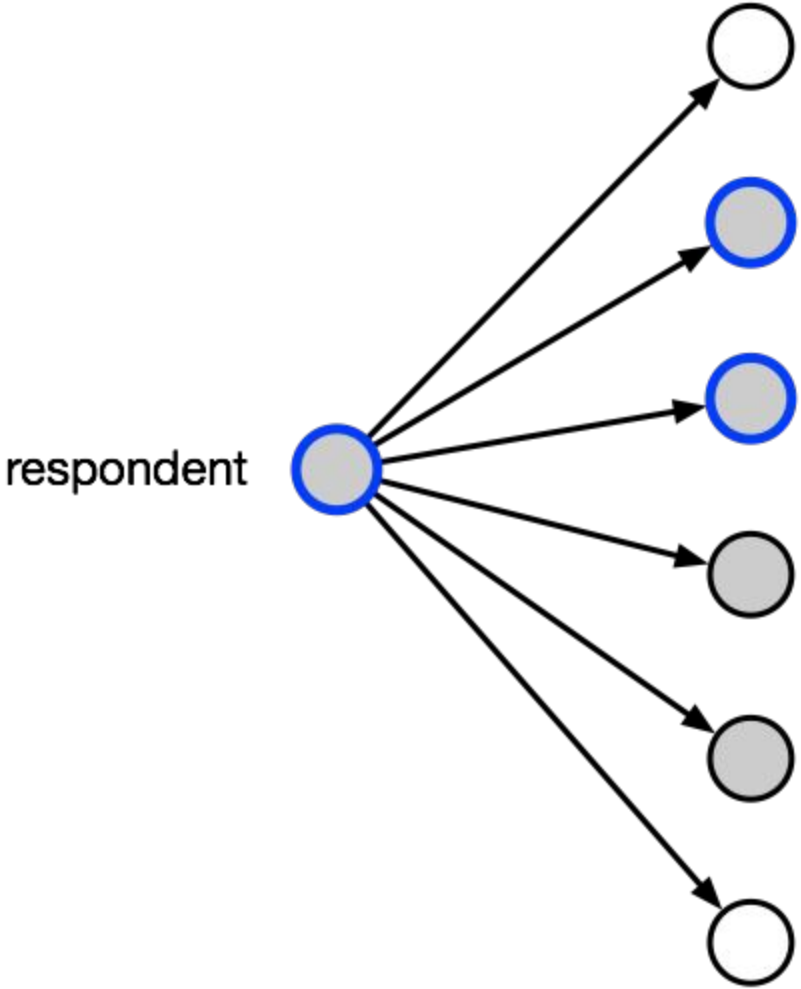
Conversational Contact Network

- How many people did you have conversational contact with yesterday? By conversational contact, we mean anyone you spoke with face to face for at least three words.

Meal Network

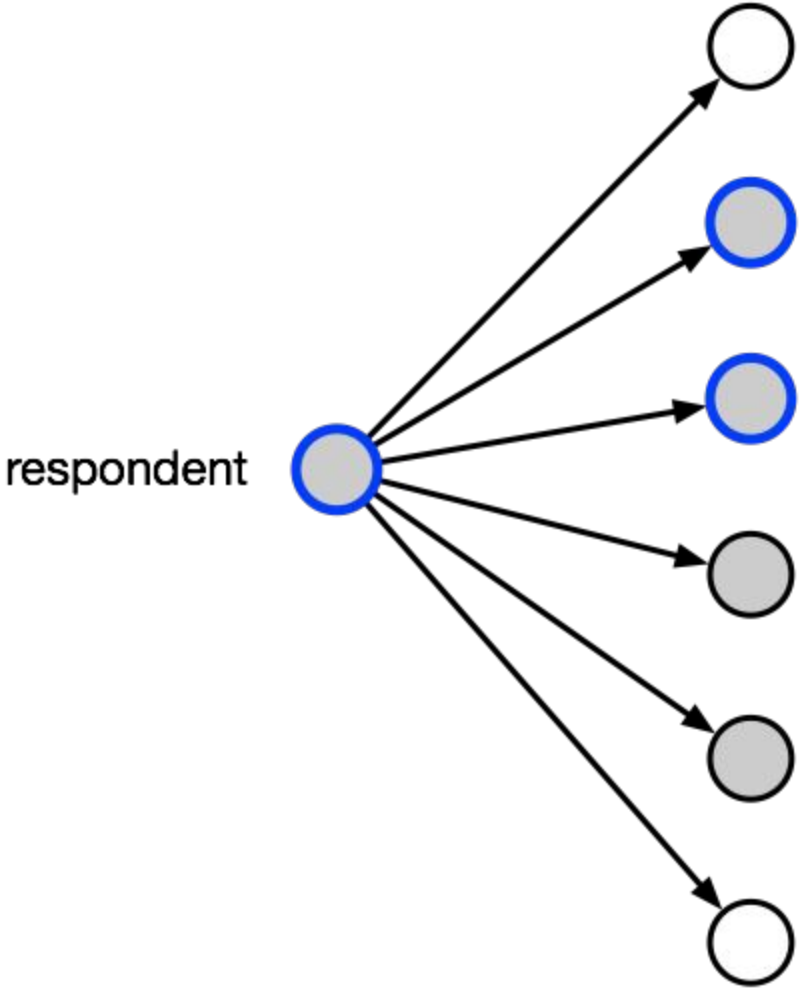
- How many people did you share food or drink with yesterday? These people could be family members, friends, co-workers, neighbors, or other people. Please include all food and drink taken at any location, including at home, at work, at a cafe, or in a restaurant.

How many people did you share food or drink with yesterday?



How many people did you share food or drink with yesterday?

=> response tells us about network size

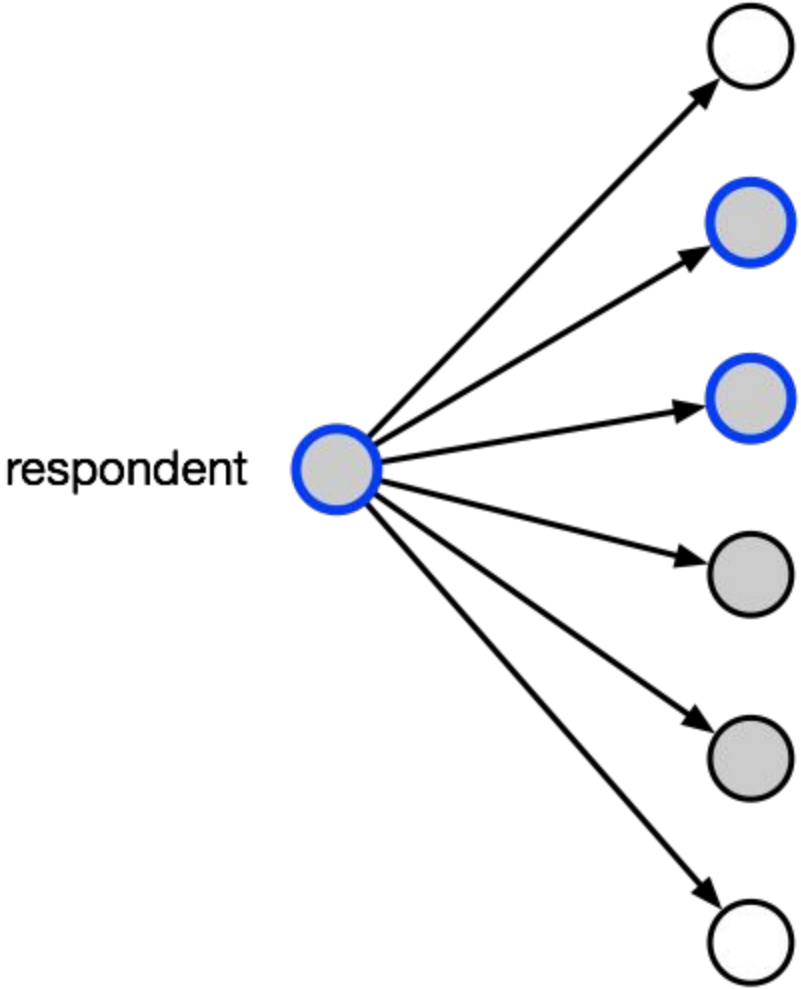


How many people did you share food or drink with yesterday?

=> response tells us about network size

Next, we want to know what proportion of respondent's network uses the internet.

Ideally: ask respondent about each person in her network, one after another



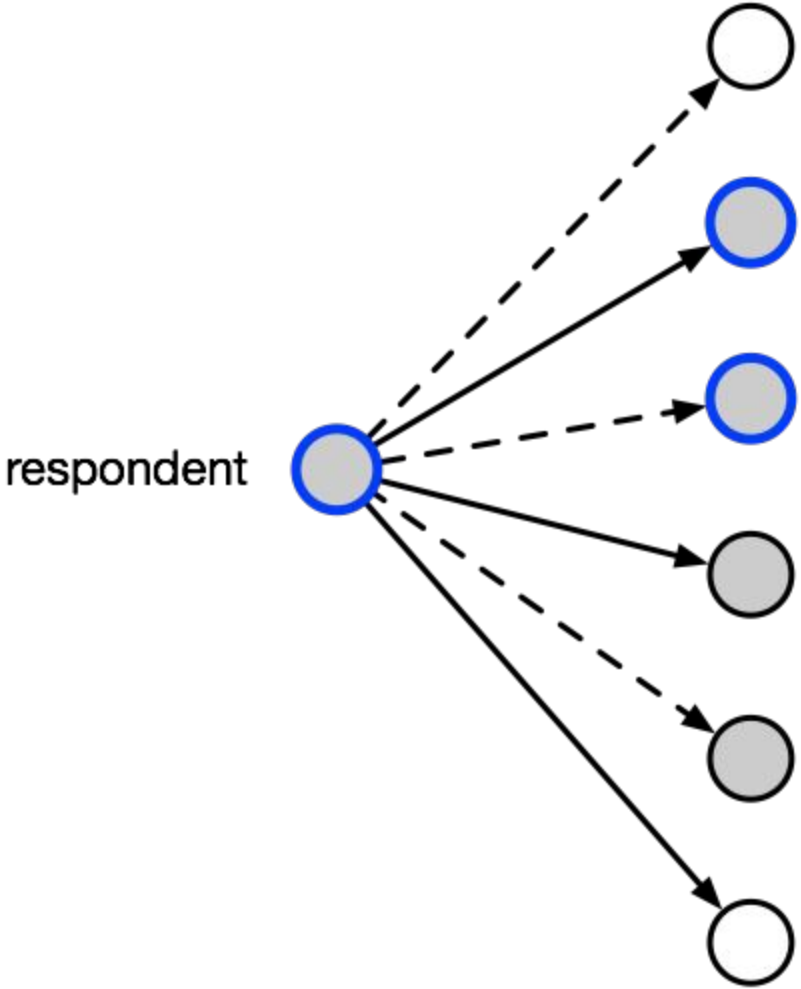
How many people did you share food or drink with yesterday?

=> response tells us about network size

Next, we want to know what proportion of respondent's network uses the internet.

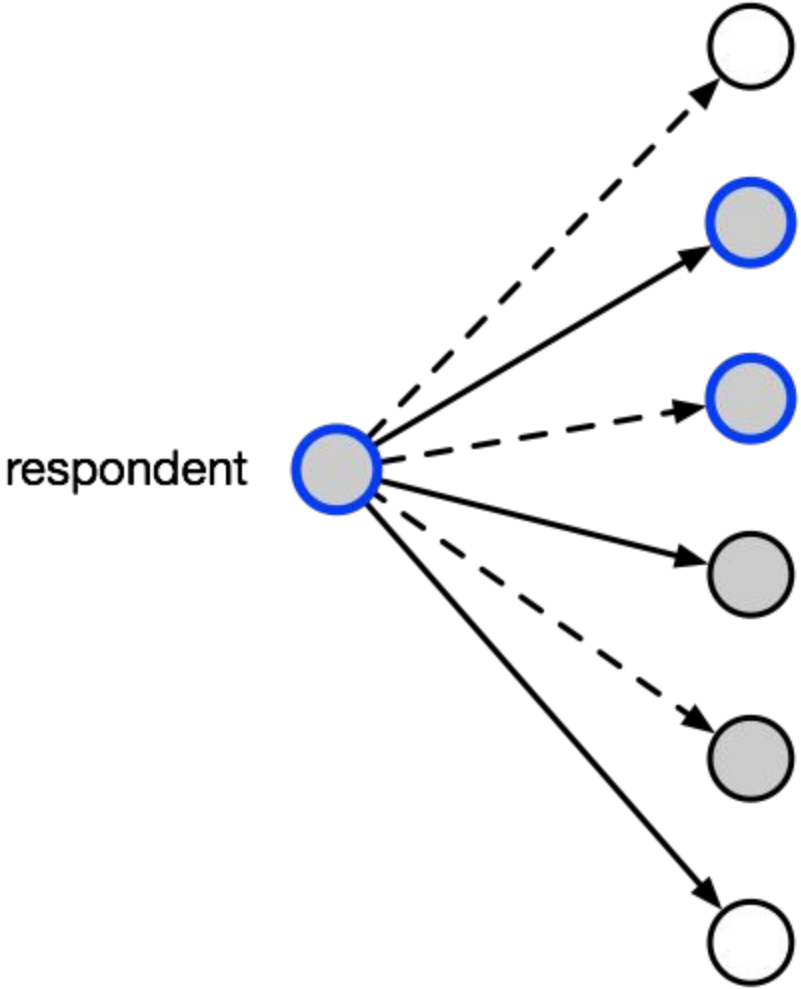
Ideally: ask respondent about each person in her network, one after another

Problem: this would likely cause unacceptable levels of respondent fatigue



Instead, we ask respondents about a subset of their network contacts; we call this subset the **detailed alters**

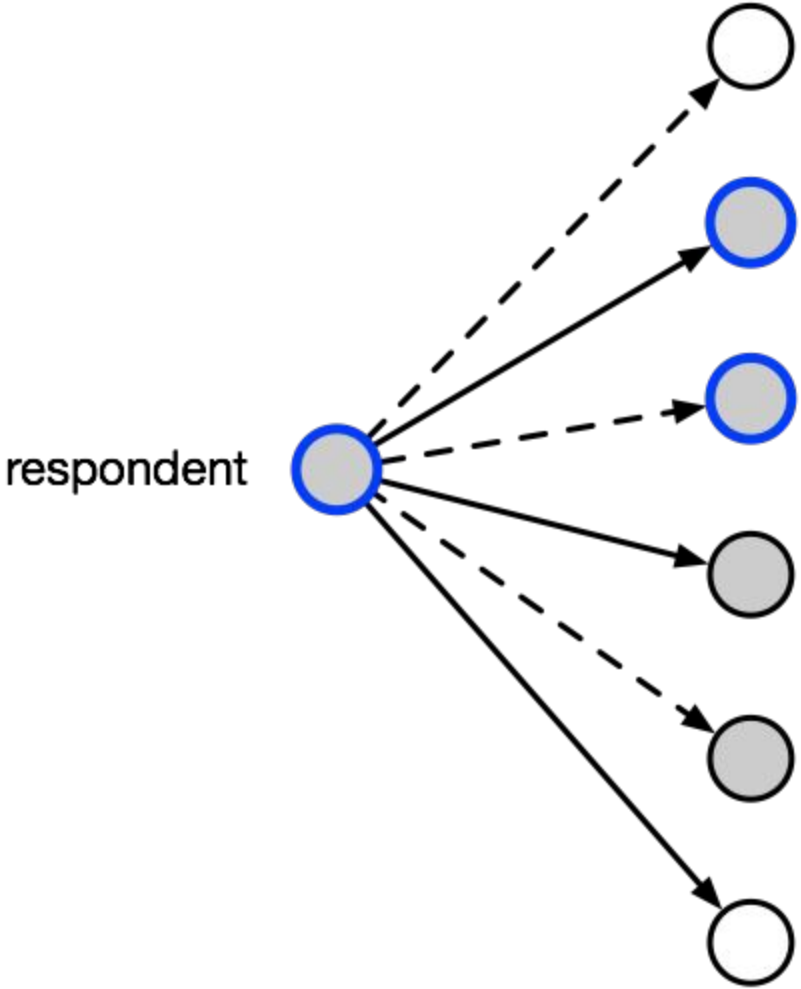
We ask for information about the three network members who ‘come to mind’ first



Instead, we ask respondents about a subset of their network contacts; we call this subset the **detailed alters**

We ask for information about the three network members who ‘come to mind’ first

We treat these three detailed alters as if they were a simple random sample of the respondent’s network members



Instead, we ask respondents about a subset of their network contacts; we call this subset the **detailed alters**

We ask for information about the three network members who ‘come to mind’ first

We treat these three detailed alters as if they were a simple random sample of the respondent’s network members

-> in reality, some alters are probably more likely to come to mind than others

-> paper mathematically describes how estimates are sensitive to this condition

-> and we can check this empirically

$$\#\text{ of internet users} = \frac{\text{total reported connections to internet users}}{\text{average in-reports per internet user}}$$

$$\#\text{ of internet users} = \frac{\text{total reported connections to internet users}}{\text{average in-reports per internet user}}$$

Estimating visibility

Can imagine many different approaches to this

Estimating visibility

Can imagine many different approaches to this

We chose something very simple: assume that people do not pay attention to whether or not they are on Facebook when they share meals with one another

Estimating visibility

Can imagine many different approaches to this

We chose something very simple: assume that people do not pay attention to whether or not they are on Facebook when they share meals with one another

Our approach works if two quantities are equal:

- The rate at which people on the internet share meals with someone on FB
- The rate at which people on FB share meals with someone else on FB

Estimating visibility

Can imagine many different approaches to this

We chose something very simple: assume that people do not pay attention to whether or not they are on Facebook when they share meals with one another

Our approach works if two quantities are equal:

- The rate at which people on the internet share meals with someone on FB
- The rate at which people on FB share meals with someone else on FB

We can estimate the second quantity from our survey responses

Putting it all together

$$\#\text{ of internet users} = \frac{\text{total reported connections to internet users}}{\text{average in-reports per internet user}}$$

Recap: 3 key conditions

- Accurate reporting
- Detailed alters picked at random
- Meals shared between people without paying attention to whether they are on Facebook or not

Our paper has sensitivity framework for understanding what impact violating these conditions would have on estimates

Framework also shows how these conditions can be relaxed or eliminated if additional data can be collected

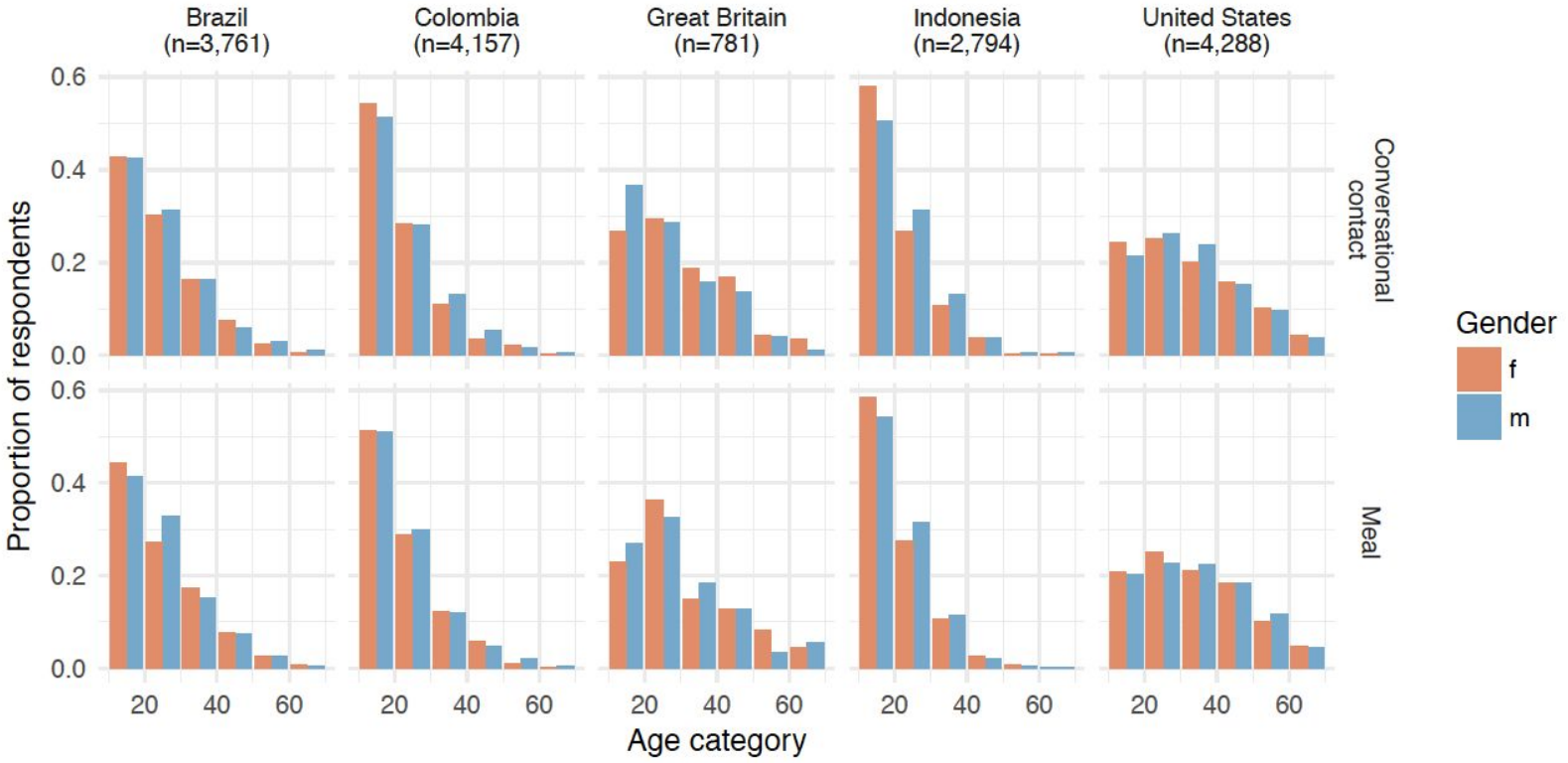
We'll see that the first two conditions can be checked empirically

Results

Sample

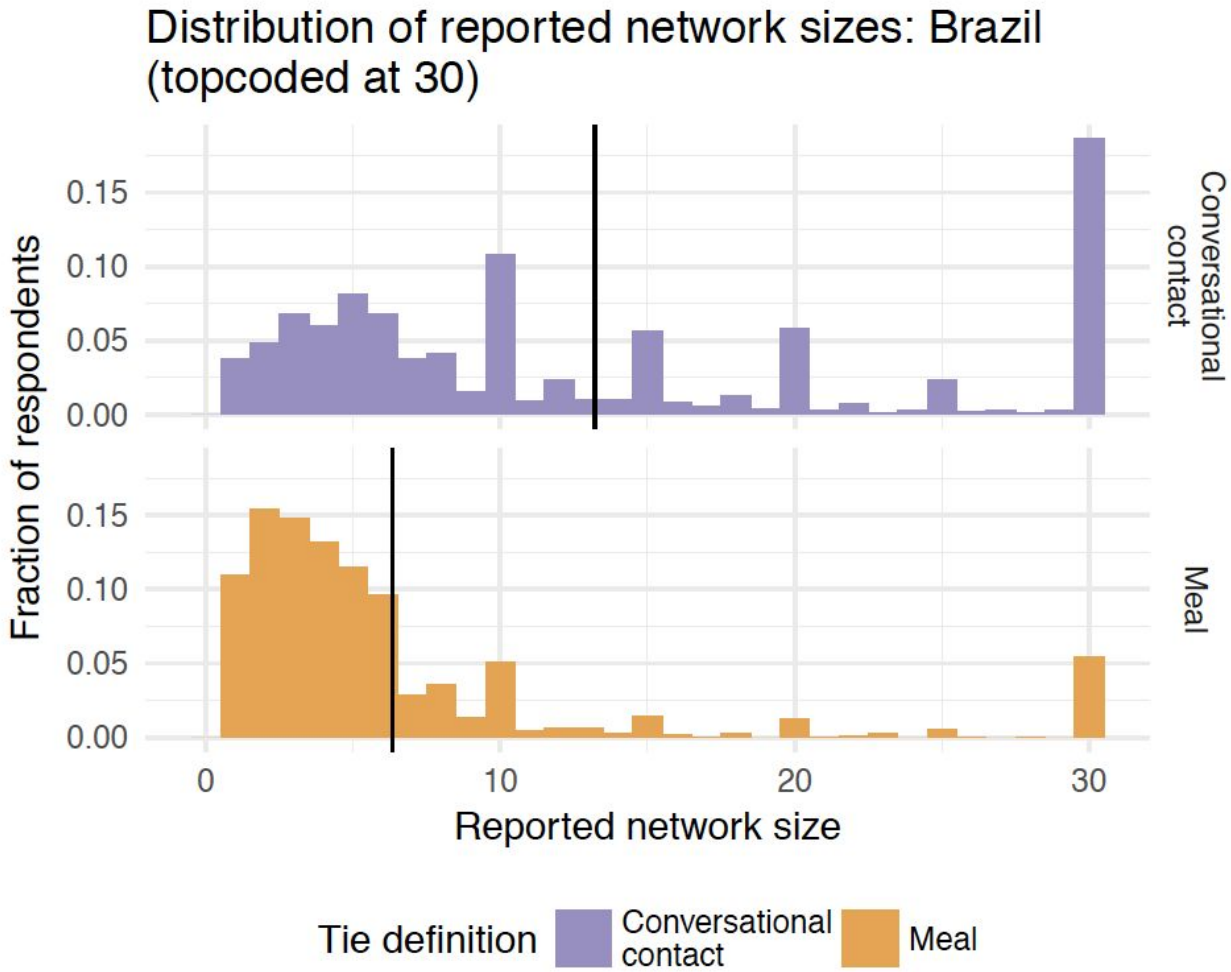
- Random sample of Facebook users, taken using FB's survey infrastructure
- Short survey, taken over web or mobile
- Looked at lots of calibration and post-stratification approaches, found that these mattered very little
- All analyses use rescaled bootstrap to estimate sampling uncertainty

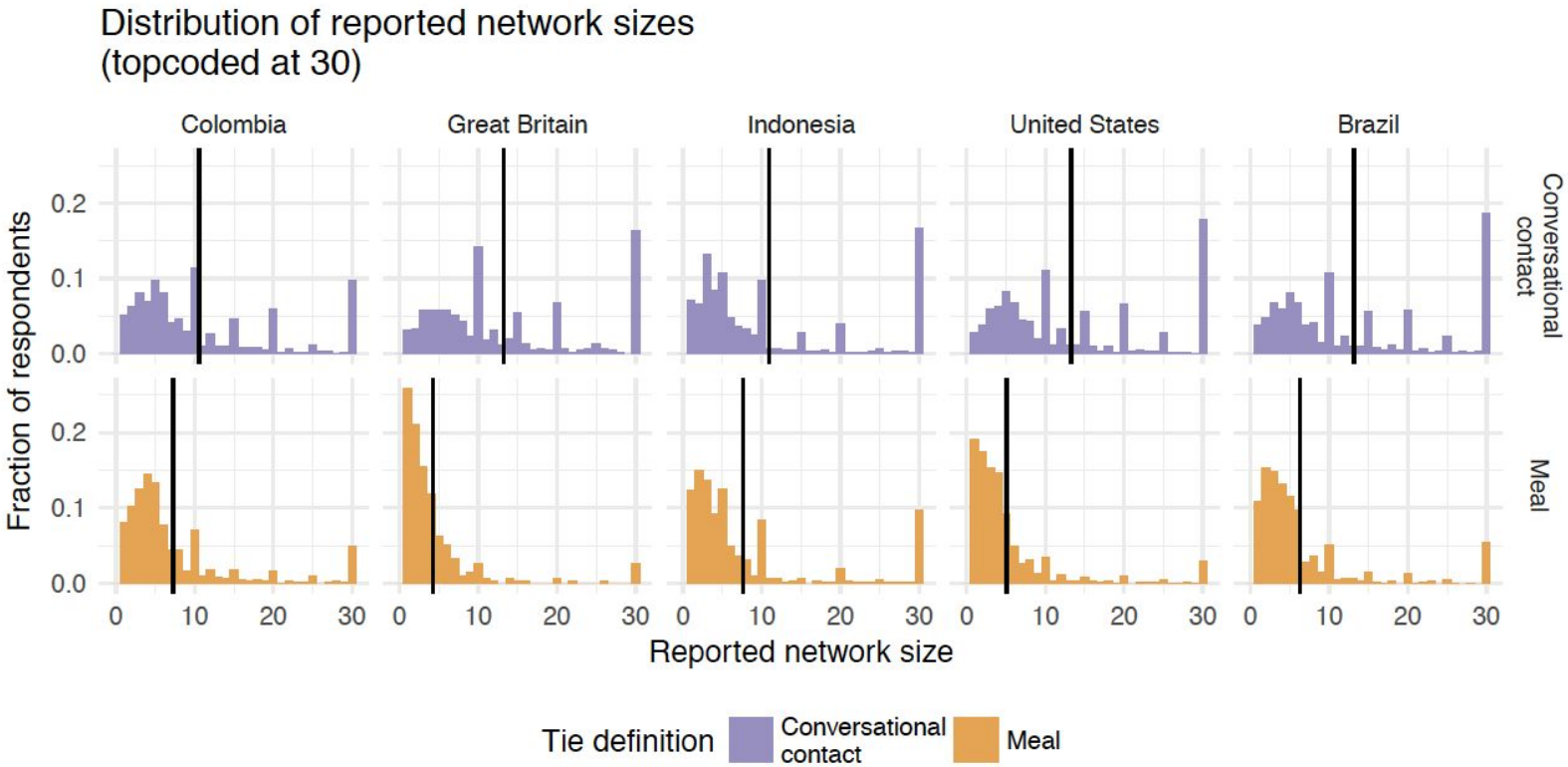
Sample: 5 countries



Degrees

Degrees


Distribution of reported network sizes: Brazil
(topcoded at 30)
Fraction of respondents
0.15
0.10
0.05
0.00
Conversational contact
Meal
0
10
20
30
Reported network size
Tie definition
Conversational contact
Meal


Distribution of reported network sizes
(topcoded at 30)
Colombia
Great Britain
Indonesia
United States
Brazil
Conversational contact
Meal
Fraction of respondents
0.0
0.1
0.2
0
10
20
30
Reported network size
Tie definition
Conversational contact
Meal

Internal consistency checks

Internal consistency checks

Idea: come up with **two independent ways** of estimating the **same quantity** from network reports

Compare these independent estimates to one another

When all of the technical conditions are satisfied, estimates will agree (up to sampling noise)

Some reporting errors or other violations of conditions can be detected with IC checks

Internal consistency checks

$$\#\ \text{connections from men to women} = \#\ \text{connections from women to men}$$

Normalized difference

$$\Delta_\alpha = \frac{1}{N_F}(\hat{d}_{F_{-\alpha},F_\alpha} - \hat{d}_{F_\alpha,F_{-\alpha}})$$

Example: reported connections to women made by men

Example: reported connections to men made by women

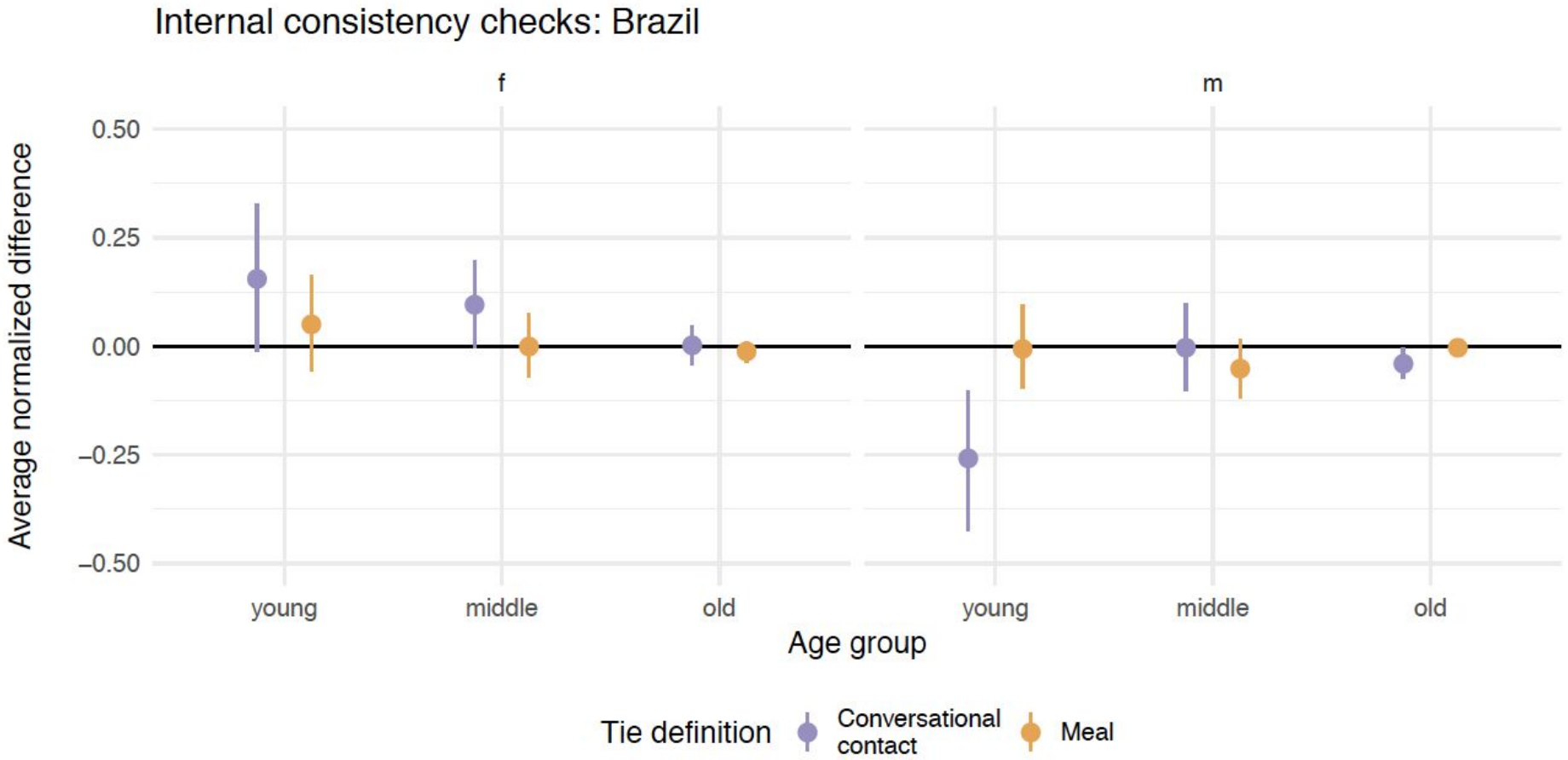
Normalized difference

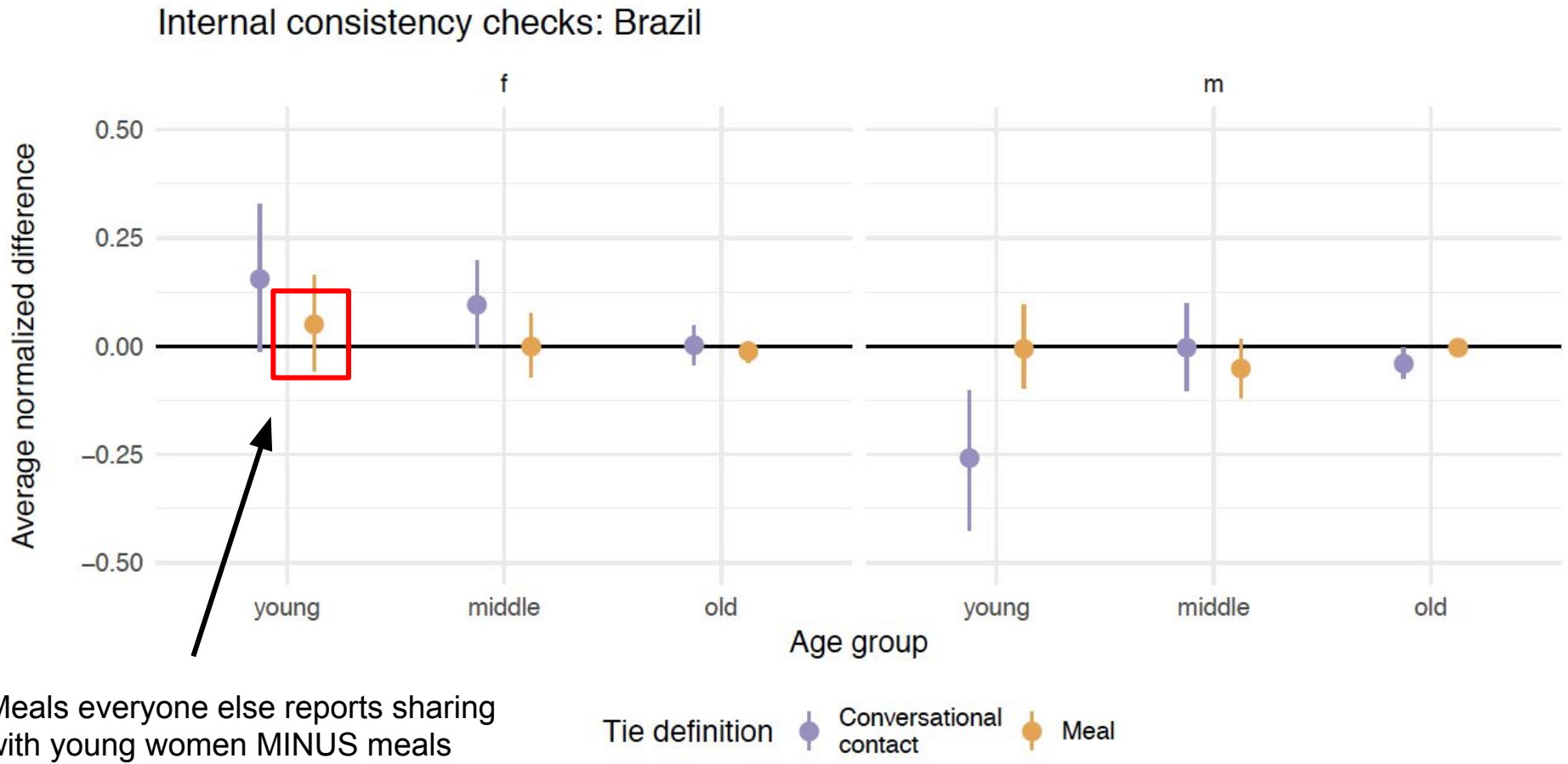
$$\Delta_\alpha = \frac{1}{N_F}(\hat{d}_{F_{-\alpha},F_\alpha} - \hat{d}_{F_\alpha,F_{-\alpha}})$$

Example: reported connections to women made by men

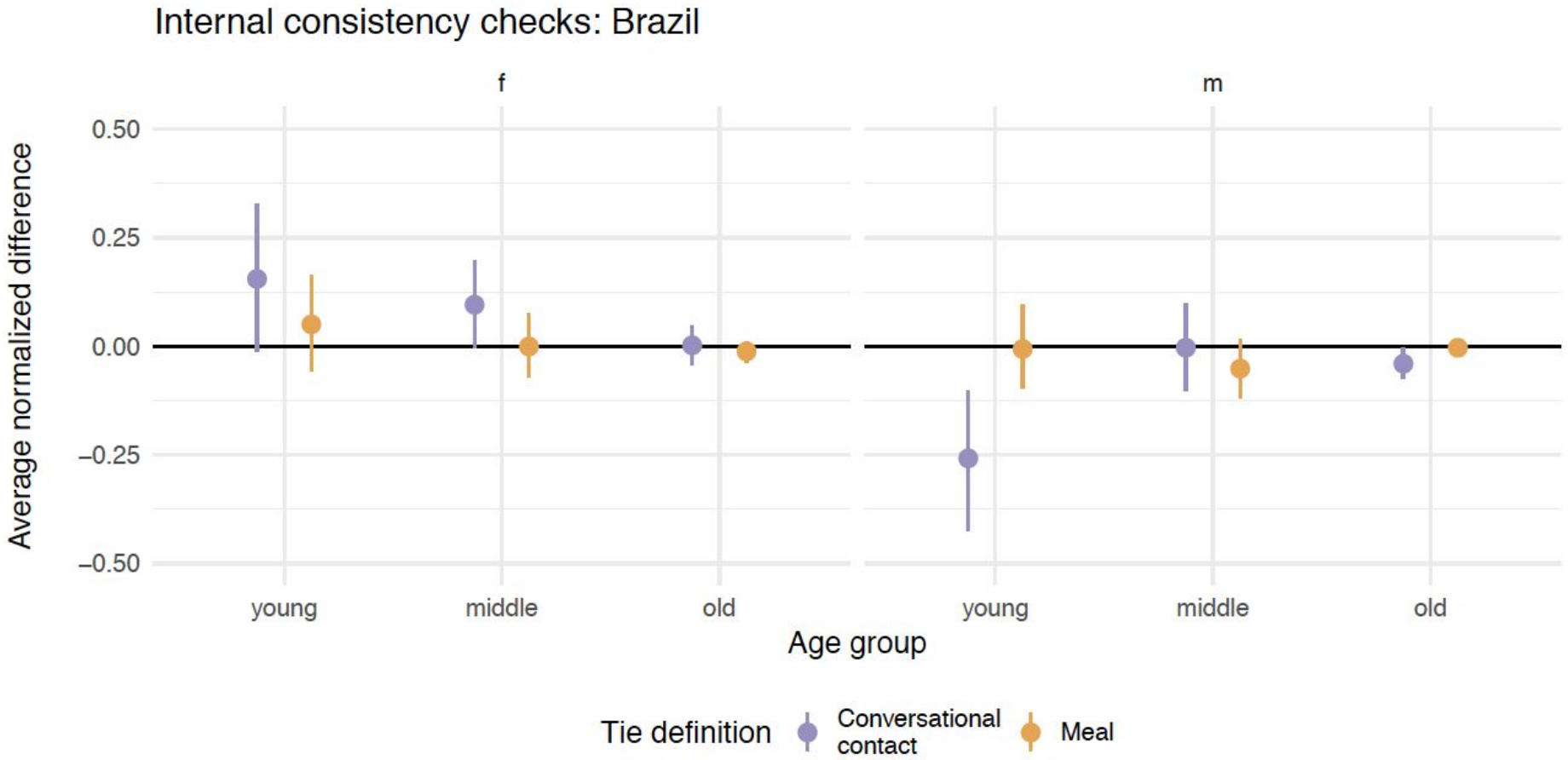
Example: reported connections to men made by women

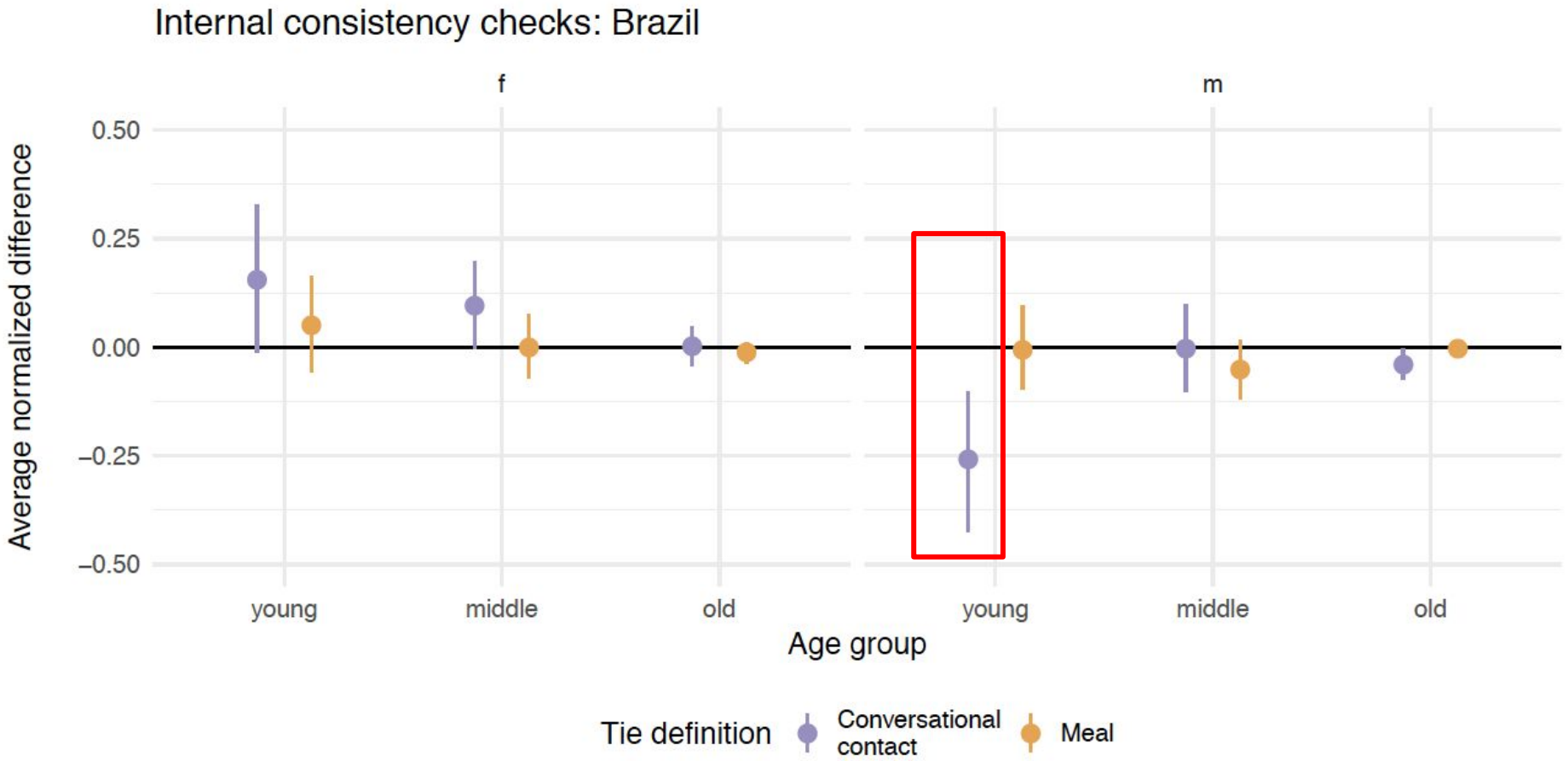
These can be estimated **independently** but they are the **same quantity**

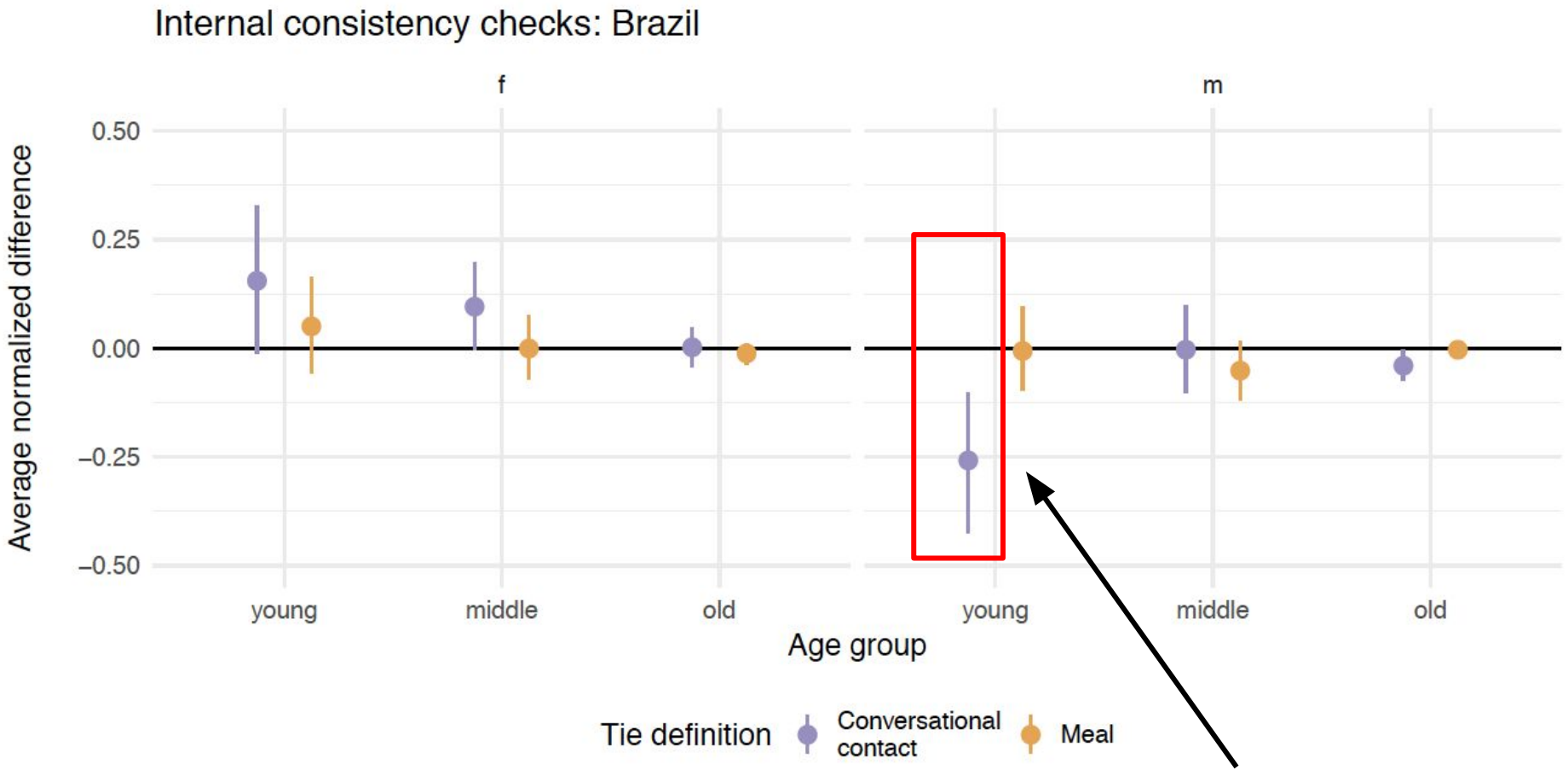

Internal consistency checks: Brazil
f
m
0.50
0.25
0.00
−0.25
−0.50
Average normalized difference
young
middle
old
young
middle
old
Age group
Tie definition
Conversational contact
Meal

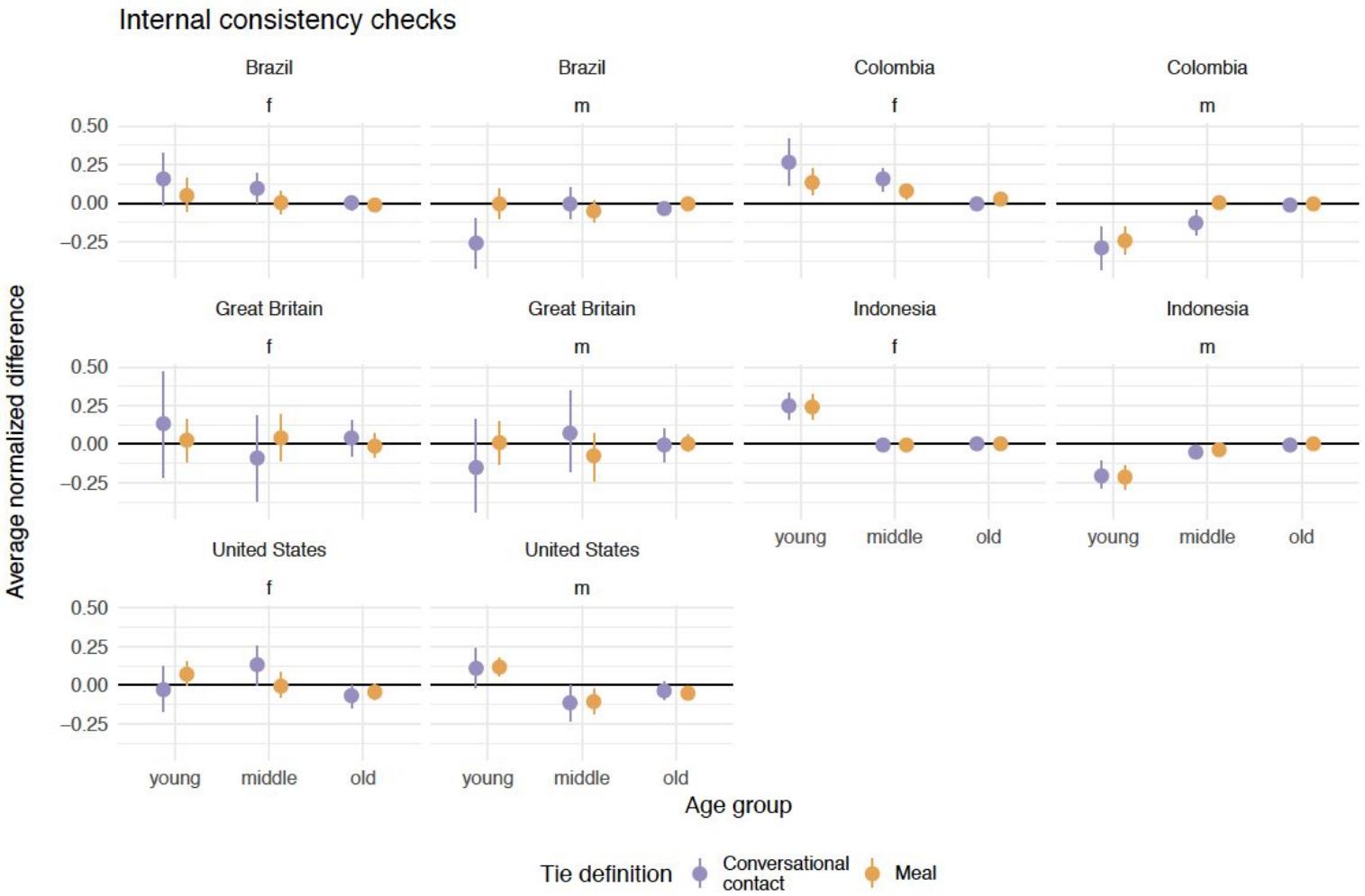


Meals everyone else reports sharing with young women MINUS meals young women report sharing with everyone else


Internal consistency checks: Brazil
f
m
0.50
0.25
0.00
−0.25
−0.50
Average normalized difference
young
middle
old
young
middle
old
Age group
Tie definition
Conversational contact
Meal


Internal consistency checks: Brazil
f
m
0.50
0.25
0.00
−0.25
−0.50
Average normalized difference
young
middle
old
young
middle
old
Age group
Tie definition
Conversational contact
Meal


Internal consistency checks: Brazil
f
m
0.50
0.25
0.00
−0.25
−0.50
Average normalized difference
young
middle
old
young
middle
old
Age group
Tie definition
Conversational contact
Meal
Young men report more connections to everyone else than everyone else reports to young men

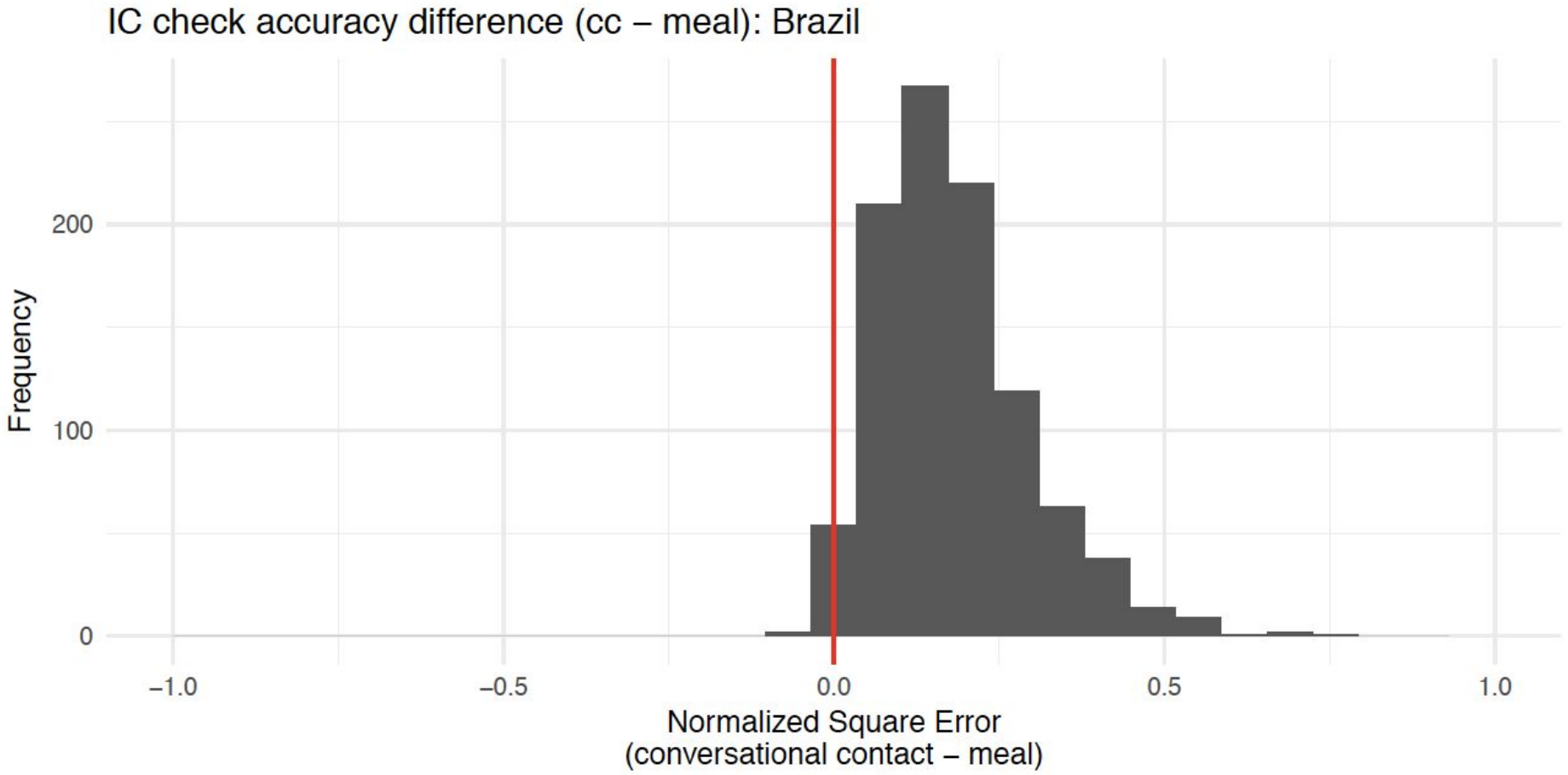

Internal consistency checks
Brazil
f
Brazil
m
Colombia
f
Colombia
m
Great Britain
f
Great Britain
m
Indonesia
f
Indonesia
m
United States
f
United States
m
0.50
0.25
0.00
−0.25
young
middle
old
Average normalized difference
Age group
Tie definition
Conversational contact
Meal

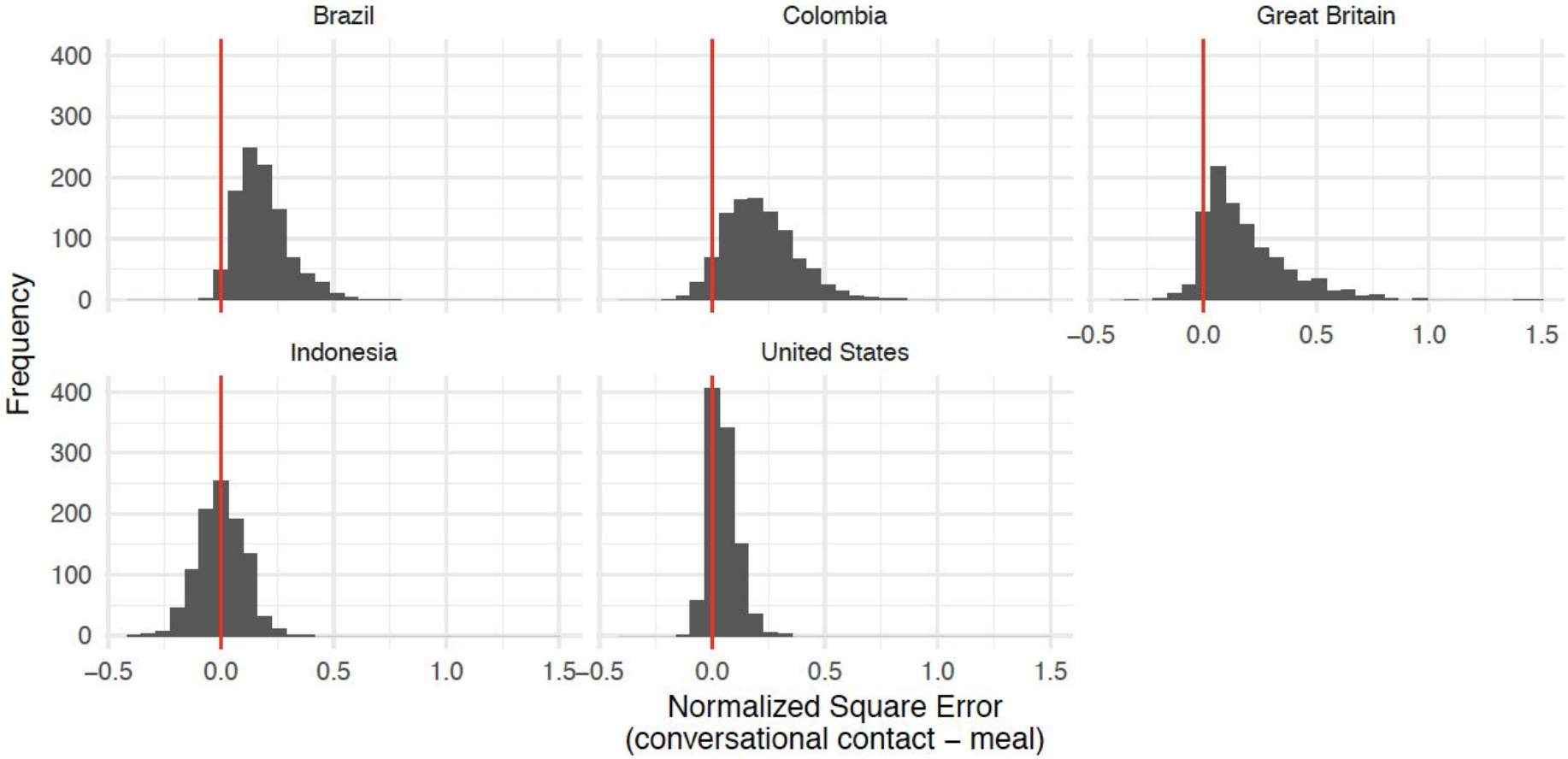

Internal consistency checks
Brazil
f
Brazil
m
Colombia
f
Colombia
m
Great Britain
f
Great Britain
m
Indonesia
f
Indonesia
m
United States
f
United States
m
0.50
0.25
0.00
−0.25
young
middle
old
Average normalized difference
Age group
Tie definition
Conversational contact
Meal

Internal consistency checks

So the IC checks give us a way to detect when conditions are not exactly met

We can also use the IC checks to compare the two different tie definitions to better understand which one is more accurate


IC check accuracy difference (cc – meal): Brazil
Frequency
200
100
0
–1.0
–0.5
0.0
0.5
1.0
Normalized Square Error
(conversational contact – meal)

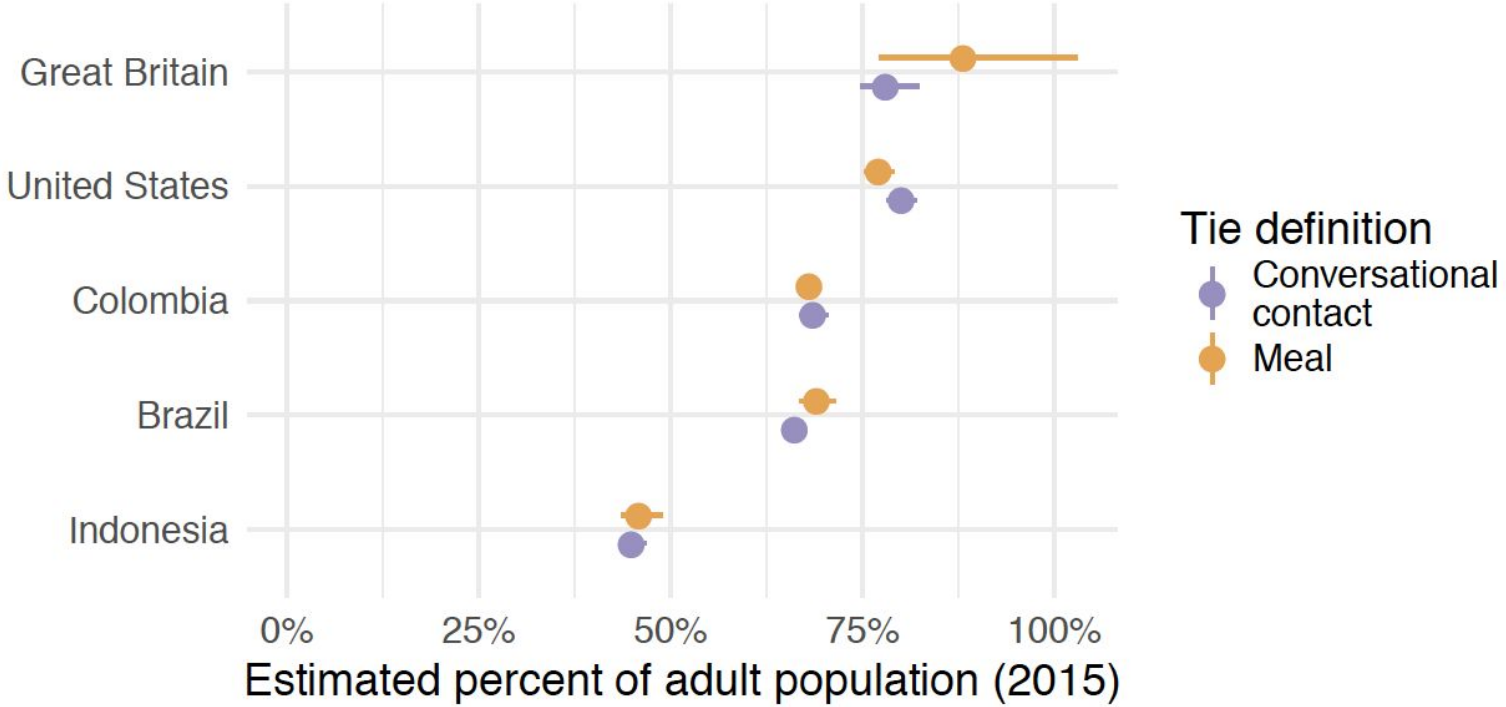

Brazil
Colombia
Great Britain
Indonesia
United States
Frequency
400
300
200
100
0
–0.5
0.0
0.5
1.0
1.5
Normalized Square Error
(conversational contact – meal)

Internal consistency checks: summary

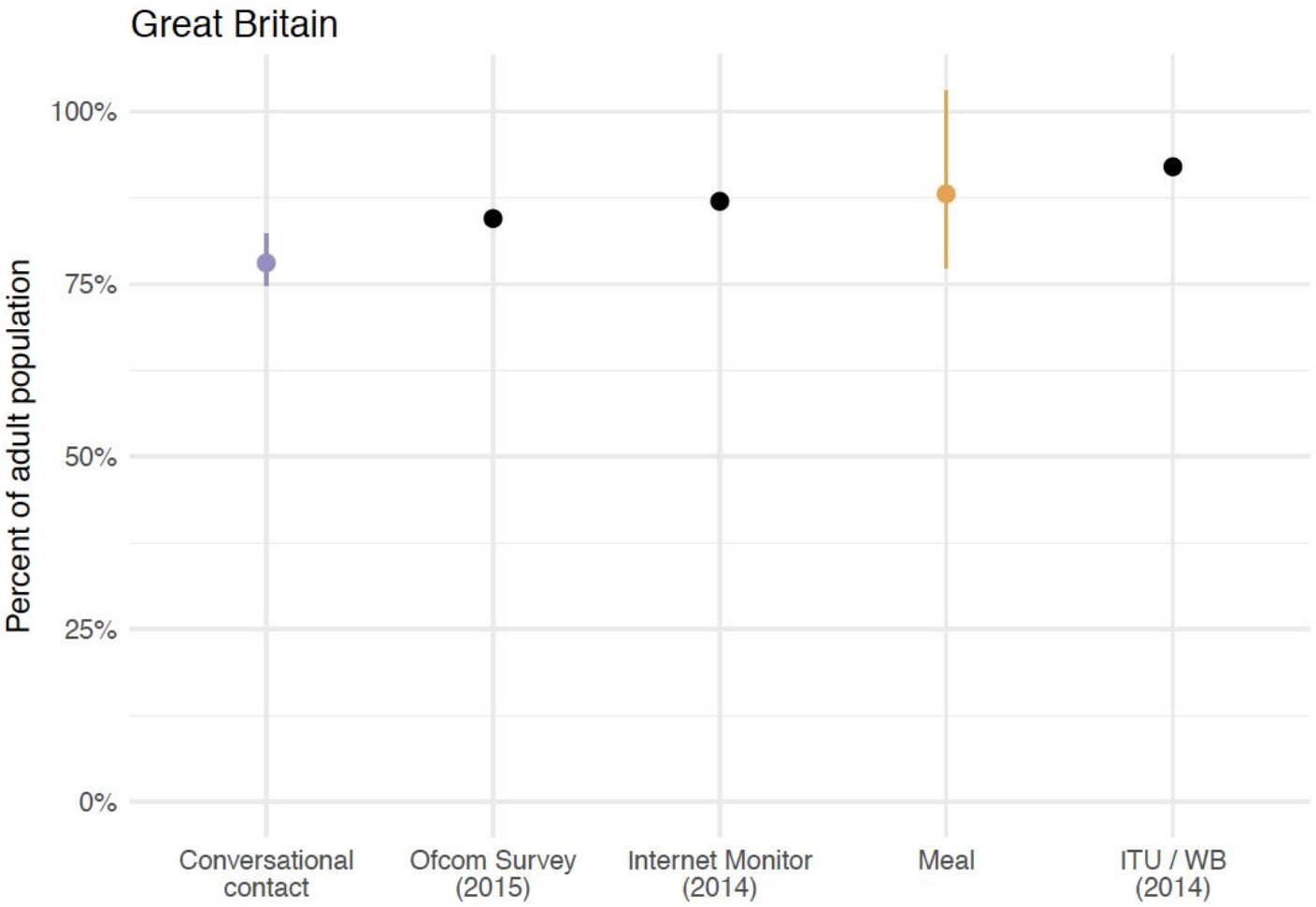
- Built-in way to assess quality of reporting
- This is very useful for building up a picture of what kind of networks people can accurately report about
- Some evidence of reporting error (especially in Indonesia and Colombia); also suggestive of differential social visibility
- They can also form the basis for model-based approaches to improving estimates from a given network
- Results from these five countries and two networks show that meal network reports tend to be more accurate

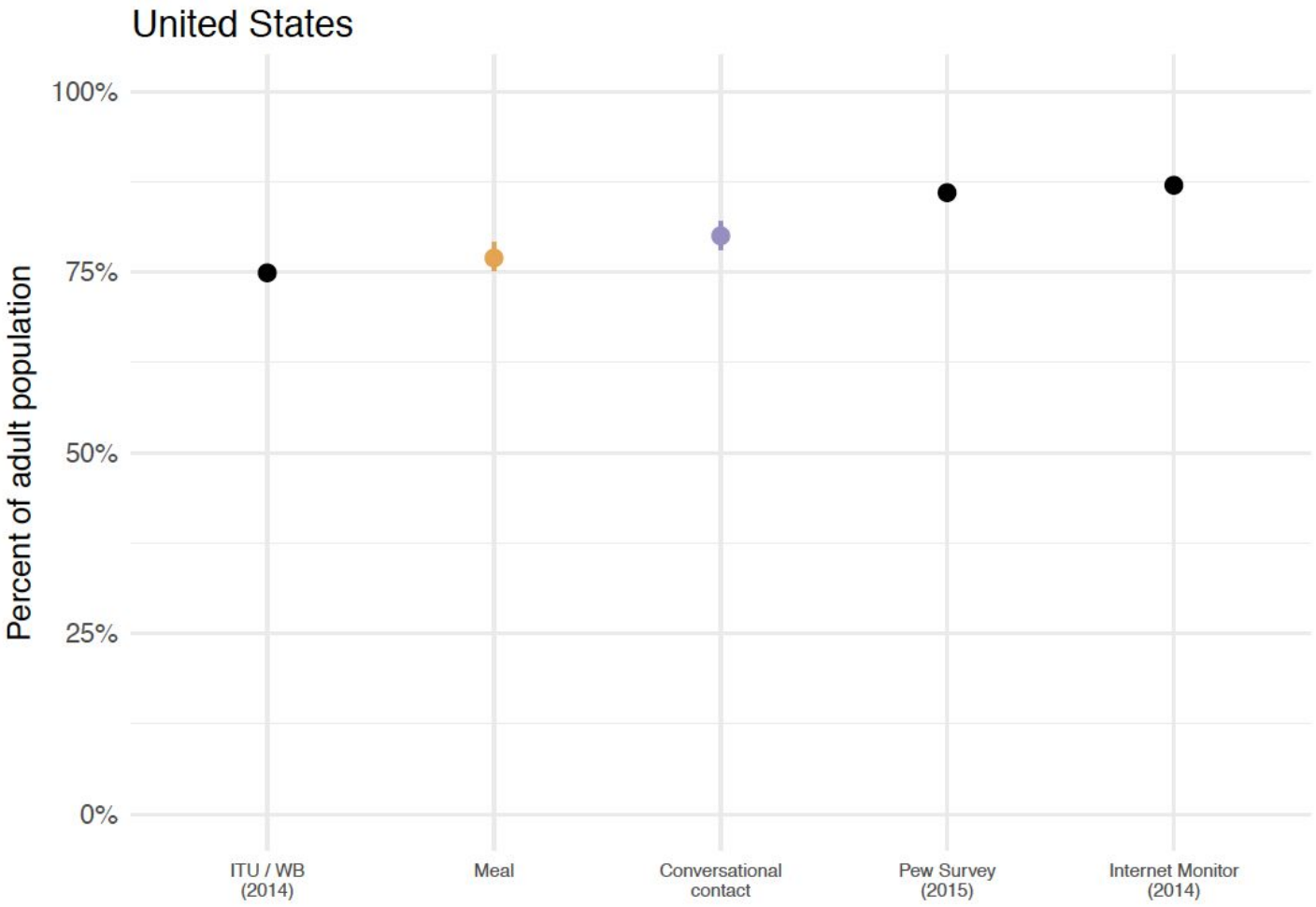
Estimates

Estimates


Great Britain
United States
Colombia
Brazil
Indonesia
0%
25%
50%
75%
100%
Estimated percent of adult population (2015)
Tie definition
Conversational contact
Meal

Estimates: comparisons

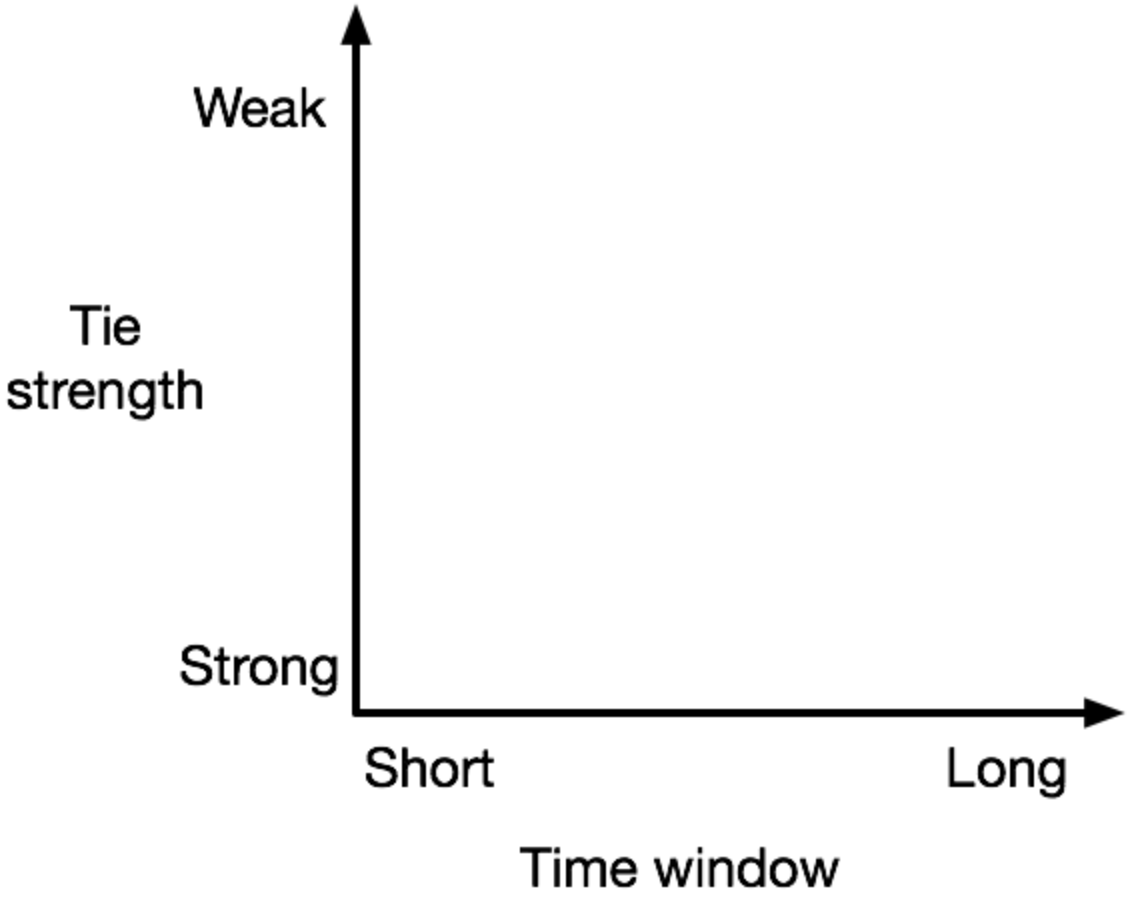

Great Britain
100%
75%
50%
25%
0%
Percent of adult population
Conversational contact
Ofcom Survey (2015)
Internet Monitor (2014)
Meal
ITU / WB (2014)

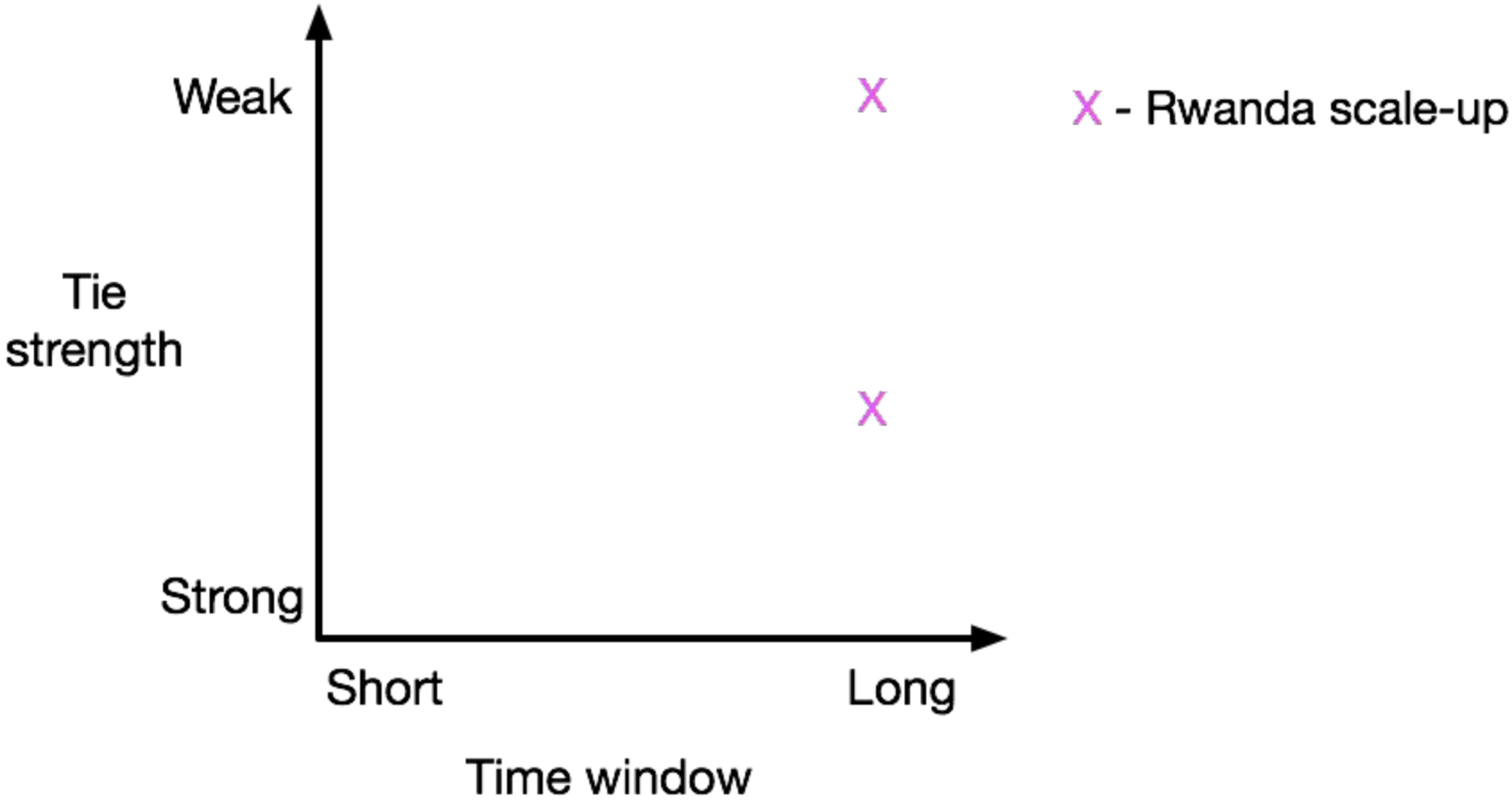

United States
100%
75%
50%
25%
0%
Percent of adult population
ITU / WB (2014)
Meal
Conversational contact
Pew Survey (2015)
Internet Monitor (2014)

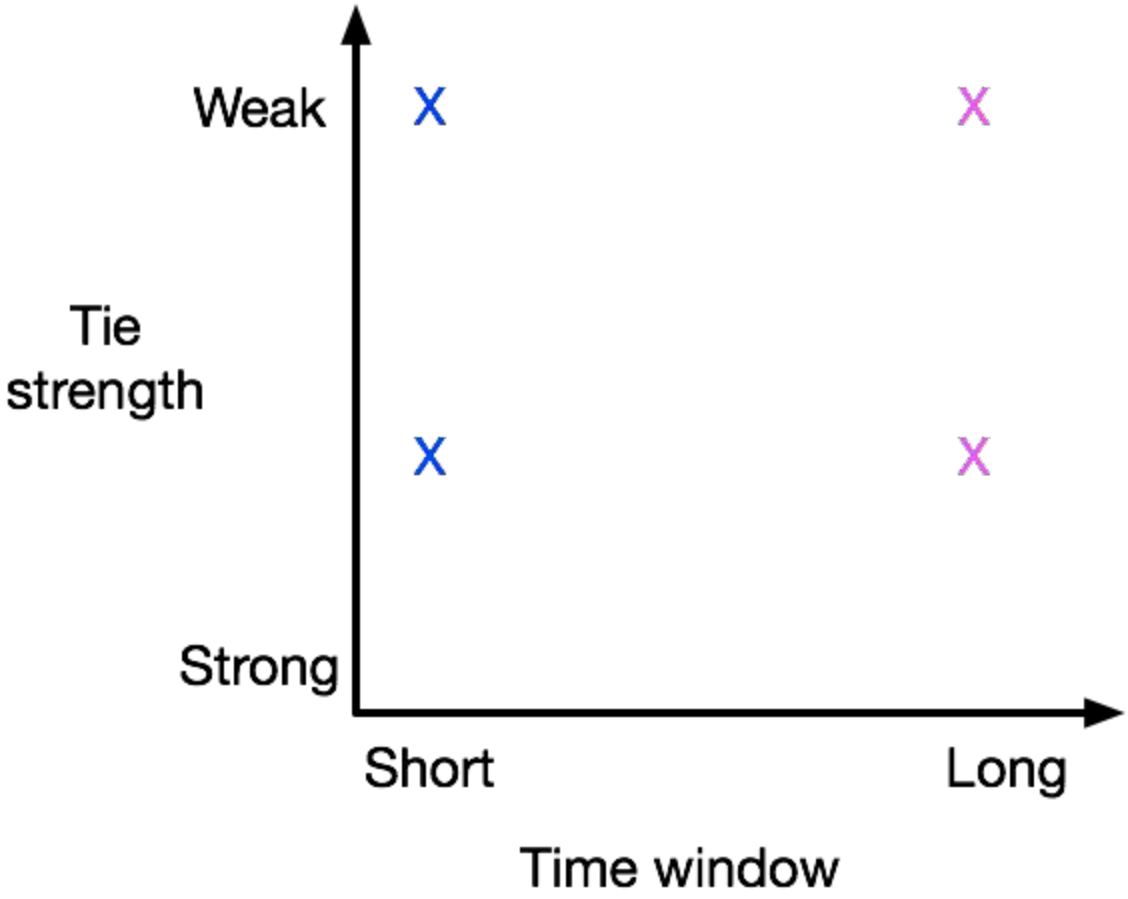
Estimates: summary

- No gold standard data to compare against, so we can’t assess estimates directly
- Comparisons to other estimates in US and GB suggest our estimates are similar to other approaches, maybe slightly low
- Internal consistency checks show some evidence of reporting error (and modeling may help with this)
- Paper has sensitivity framework that can be used to formally understand what impact violating different conditions would have on estimates

Future directions

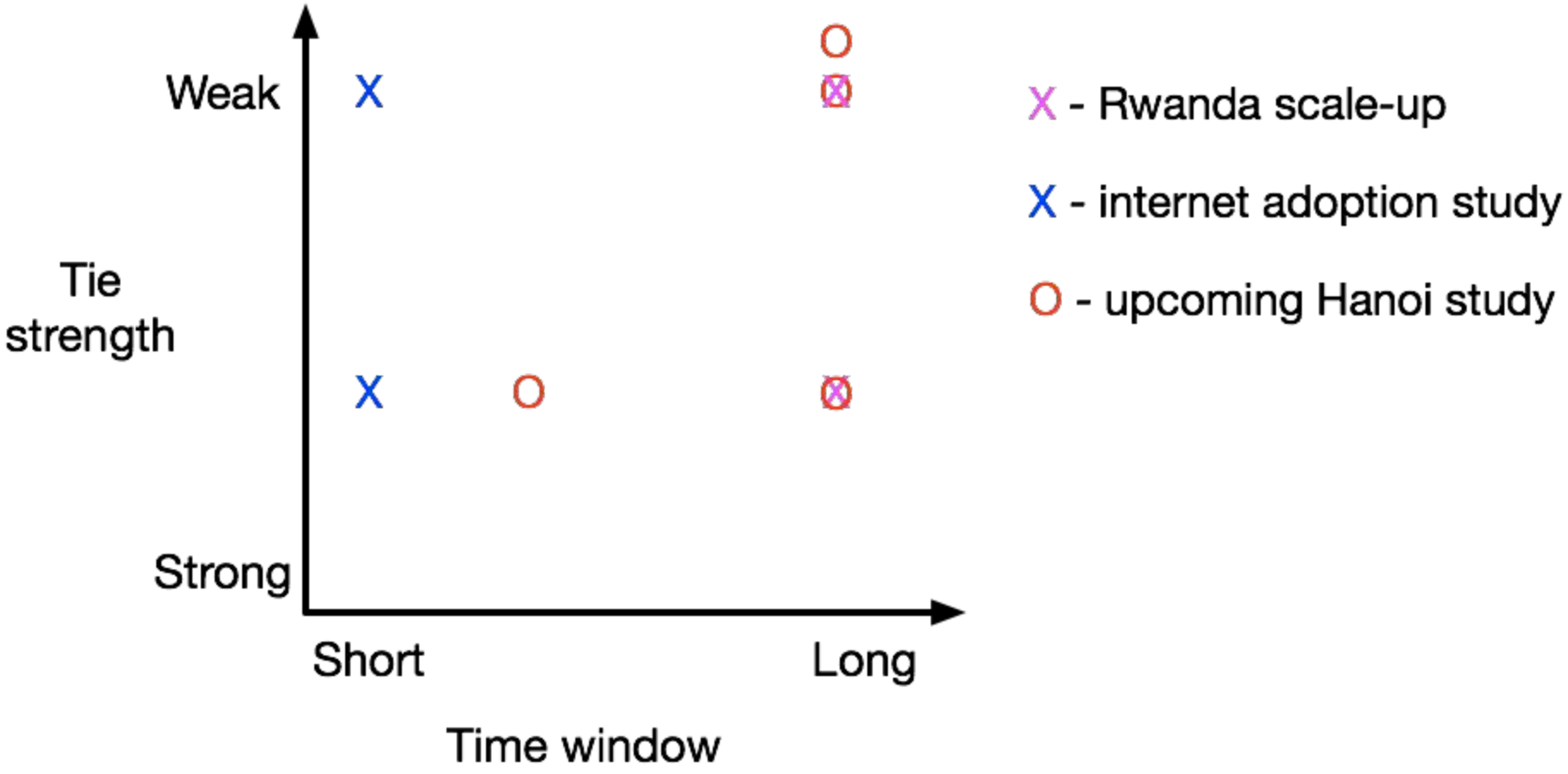

Weak
Tie
strength
Strong
Short
Long
Time window

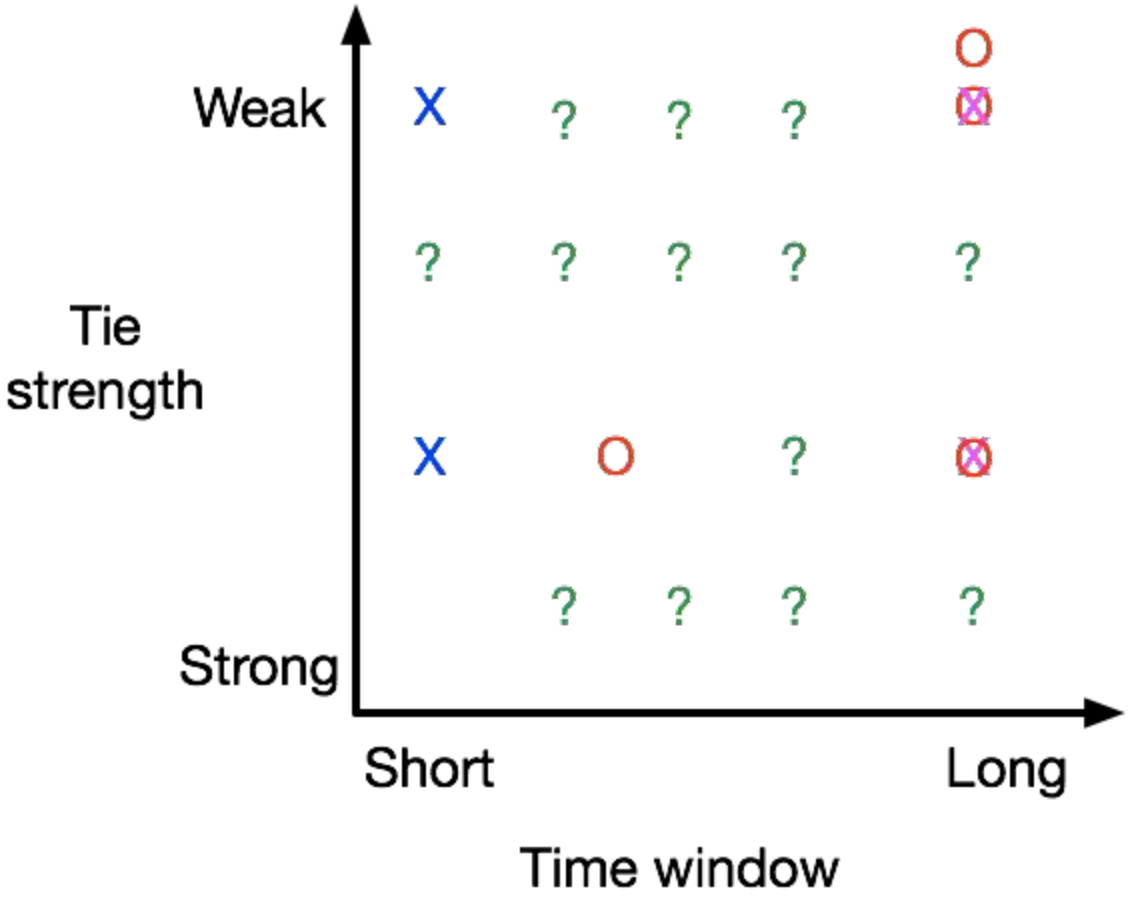

Weak
Tie strength
Strong
Short
Long
Time window
X
X
X - Rwanda scale-up

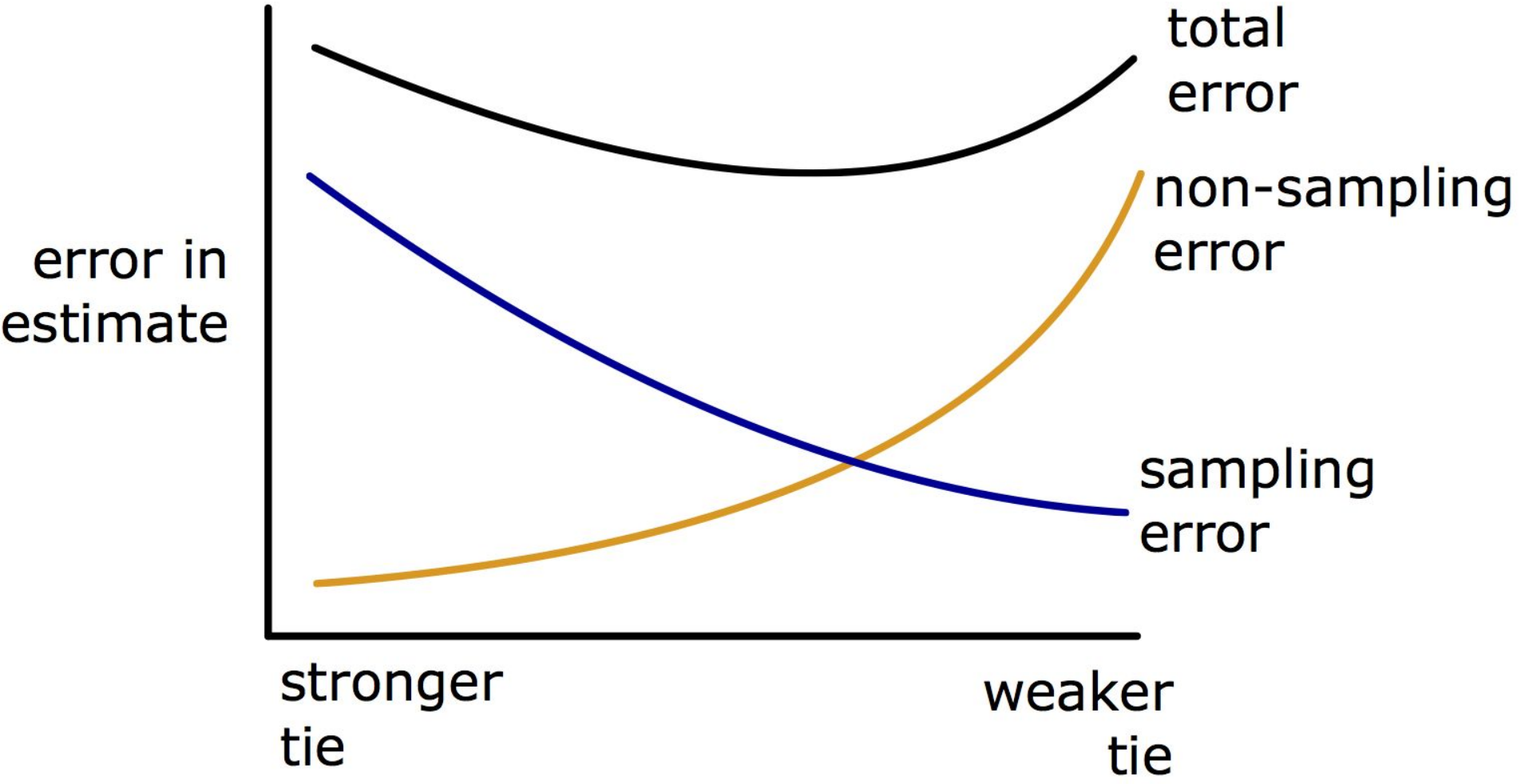


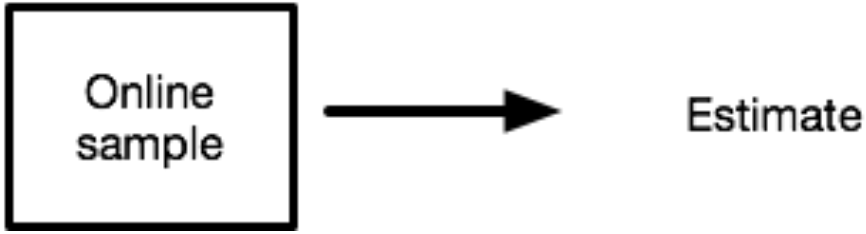
X - Rwanda scale-up

X - internet adoption study

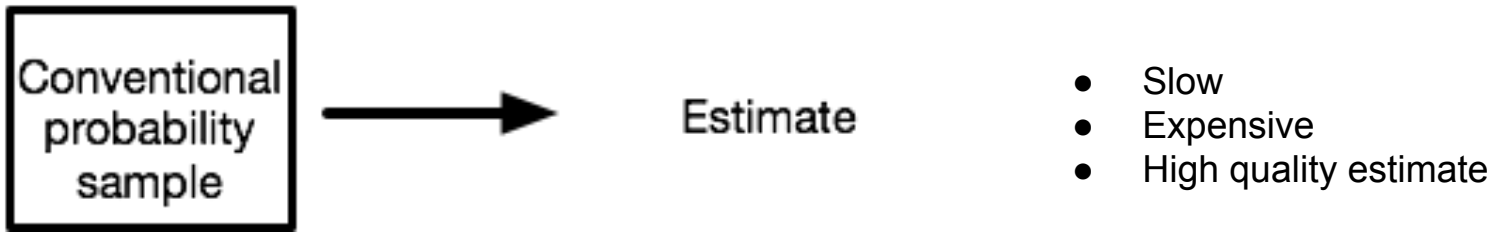

Weak
Strong
Tie strength
Short
Long
Time window
X - Rwanda scale-up
X - internet adoption study
O - upcoming Hanoi study


Weak
Strong
Tie strength
Short
Long
Time window
X - Rwanda scale-up
X - internet adoption study
O - upcoming Hanoi study
? - your study?

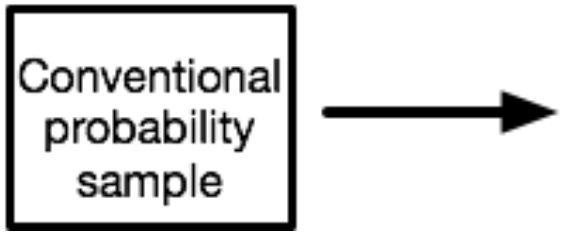

total
error
non-sampling
error
error in
estimate
sampling
error
stronger
tie
weaker
tie



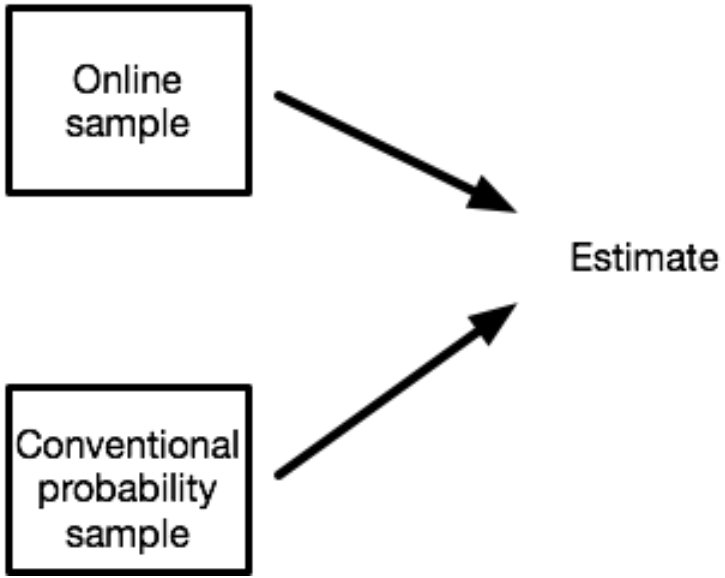
- Fast
- Inexpensive
- Easy to experiment
- High time resolution
- … but approximate



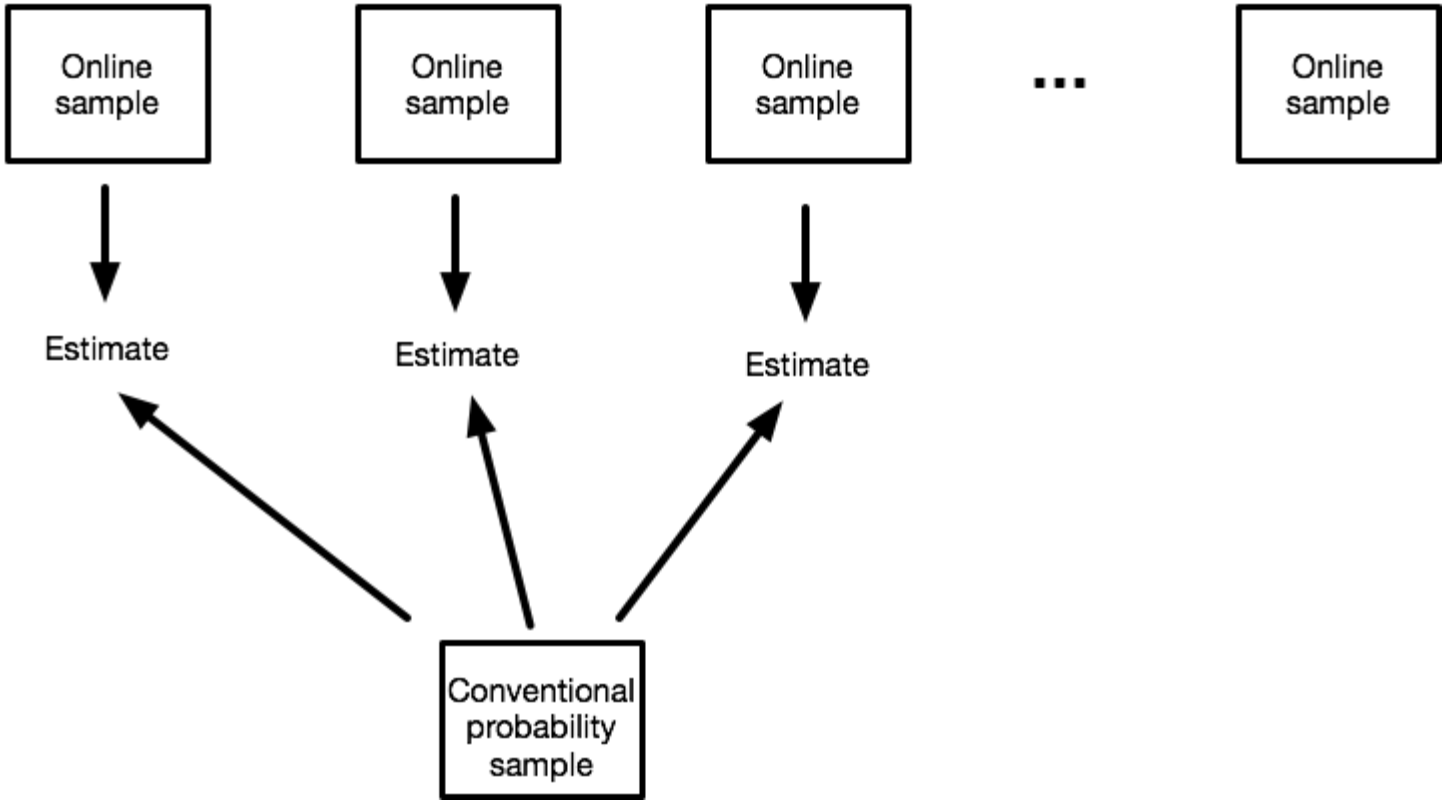
- Slow
- Expensive
- High quality estimate

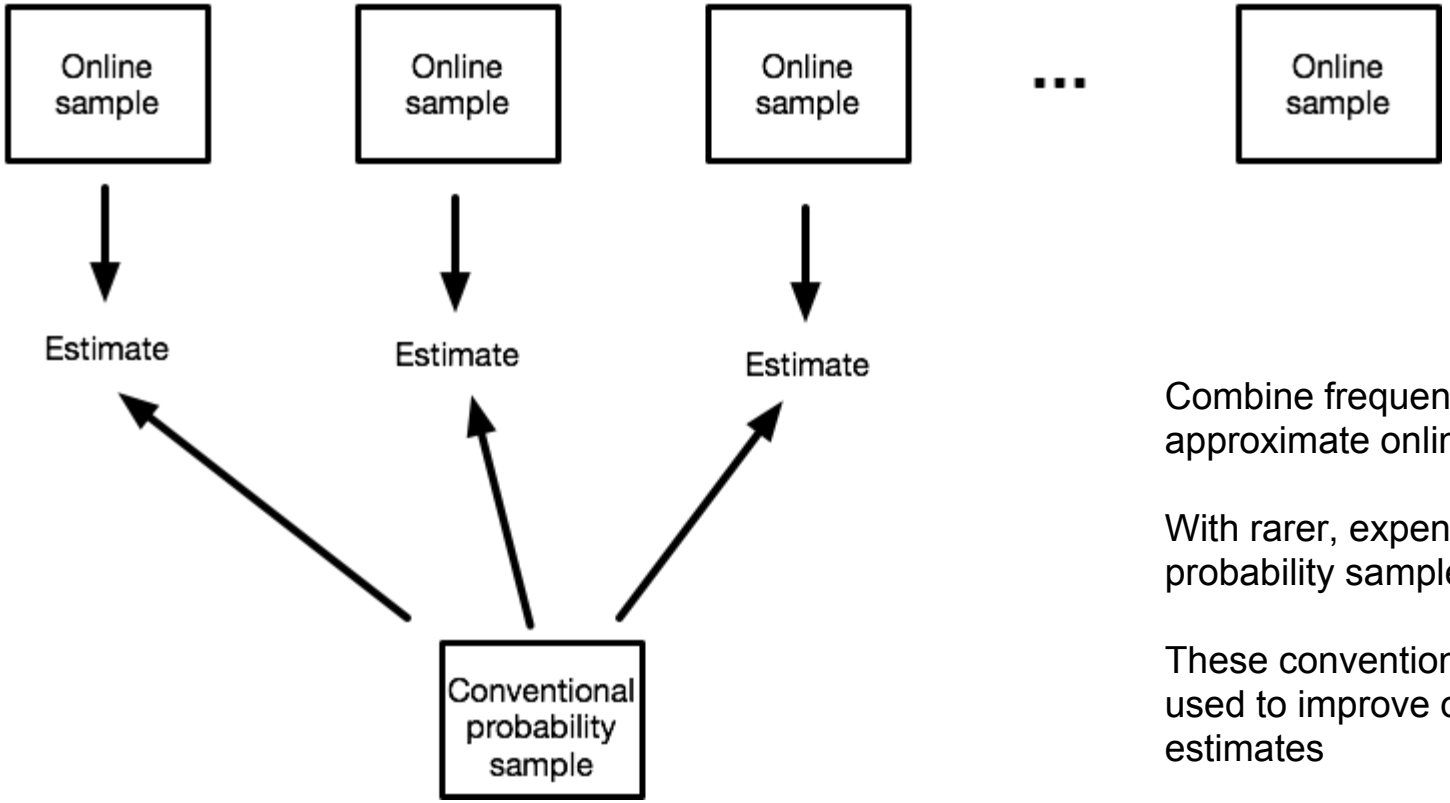


- Slow
- Expensive
- High quality estimate
- Can collect information to help online estimates



- Network reporting framework can be used to understand how to measure things in conventional sample to improve online estimates


Online sample
Online sample
Online sample
...
Online sample
Estimate
Estimate
Estimate
Conventional probability sample



Combine frequent, inexpensive, approximate online-based estimates

With rarer, expensive conventional probability samples

These conventional samples can be used to improve online-based estimates

Coming next...

- Internet adoption
  - Full sensitivity framework
  - Explore models to adjust for IC checks
  - Can also calculate estimated adoption by age and gender
  - And it’s possible to do some reporting adjustments from data we collected

Coming next...

- Internet adoption
  - Full sensitivity framework
  - Explore models to adjust for IC checks
  - Can also calculate estimated adoption by age and gender
  - And it’s possible to do some reporting adjustments from data we collected
- Sibling histories (PAA 2018 session 68-4, Thurs)
- Brazil: probability sample of 25,000 respondents
  - Validate network survival methods for adult mortality
  - Test estimating out-migration using network reports
- Hanoi network scale-up for key populations at risk of HIV
- Guidance on sampling and study design

Thanks!

- Collaborator, Curtiss Cobb

- My R packages *networkreporting* and *surveybootstrap* are available on CRAN
- Rwanda data are downloadable from the DHS website

- Feehan, Umubyeyi, Mahy, Hladik, and Salganik (2016) “Quantity vs quality: a survey experiment to improve the network scale-up method”, *American Journal of Epidemiology*
- Feehan and Salganik “Generalizing the network scaleup method”, *Sociological Methodology*.
- Feehan, Mahy, and Salganik “The network survival estimator for adult mortality: evidence from Rwanda”, *Demography*

See http://www.dennisfeehan.org for more information.