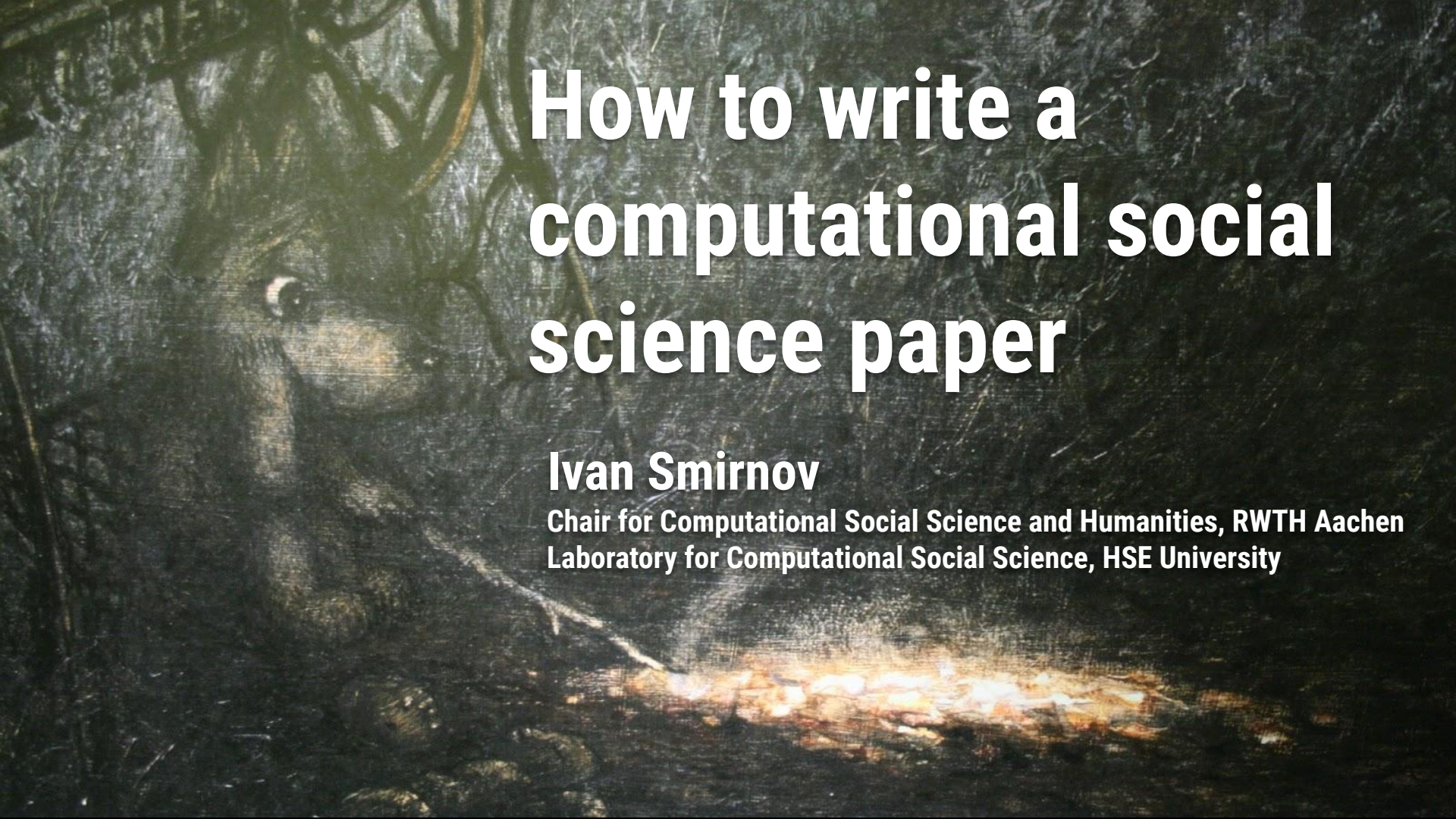


How to write a computational social science paper

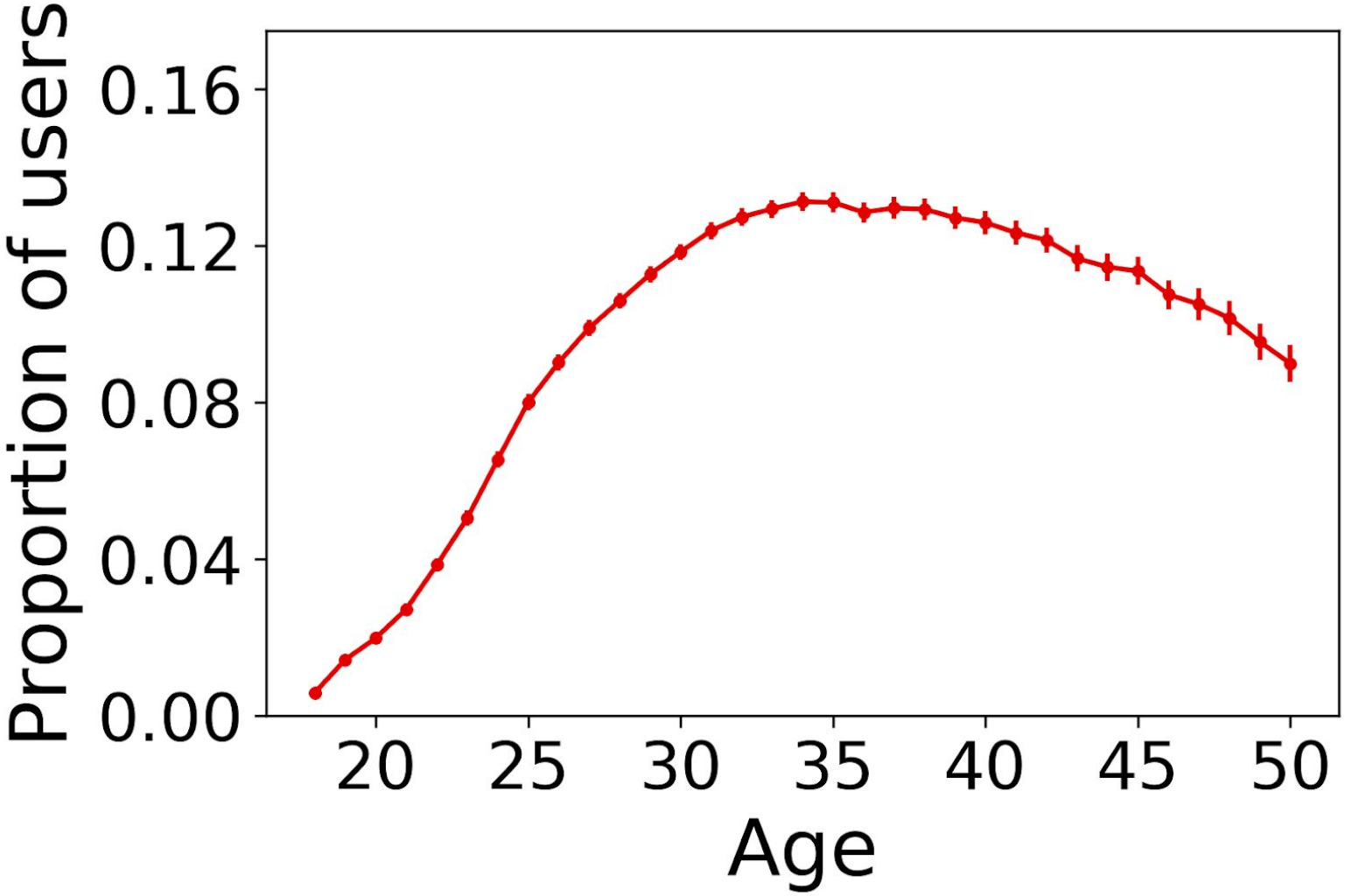
Ivan Smirnov
Chair for Computational Social Science and Humanities, RWTH Aachen
Laboratory for Computational Social Science, HSE University

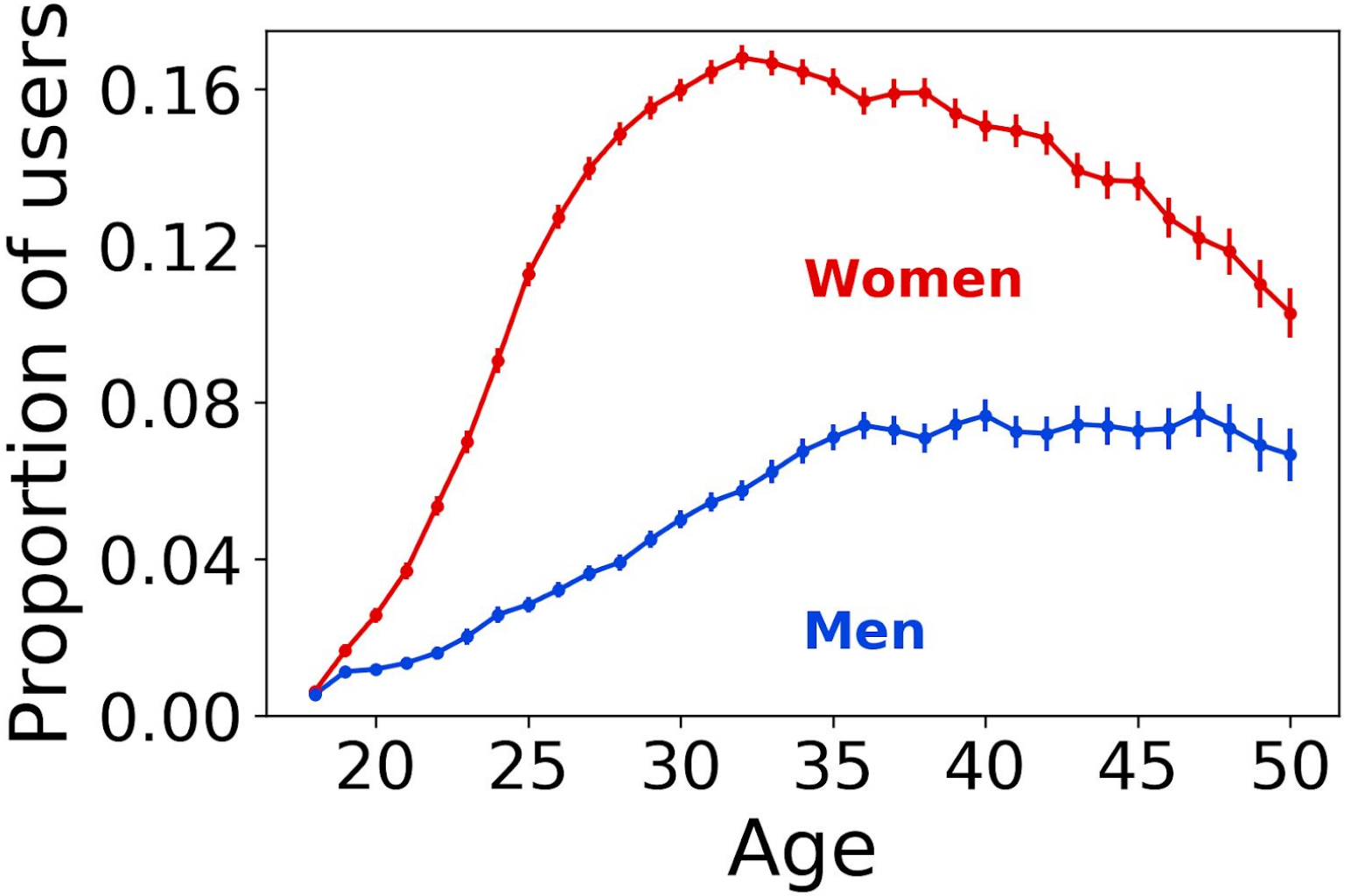
Parents mention sons more often than daughters on social media

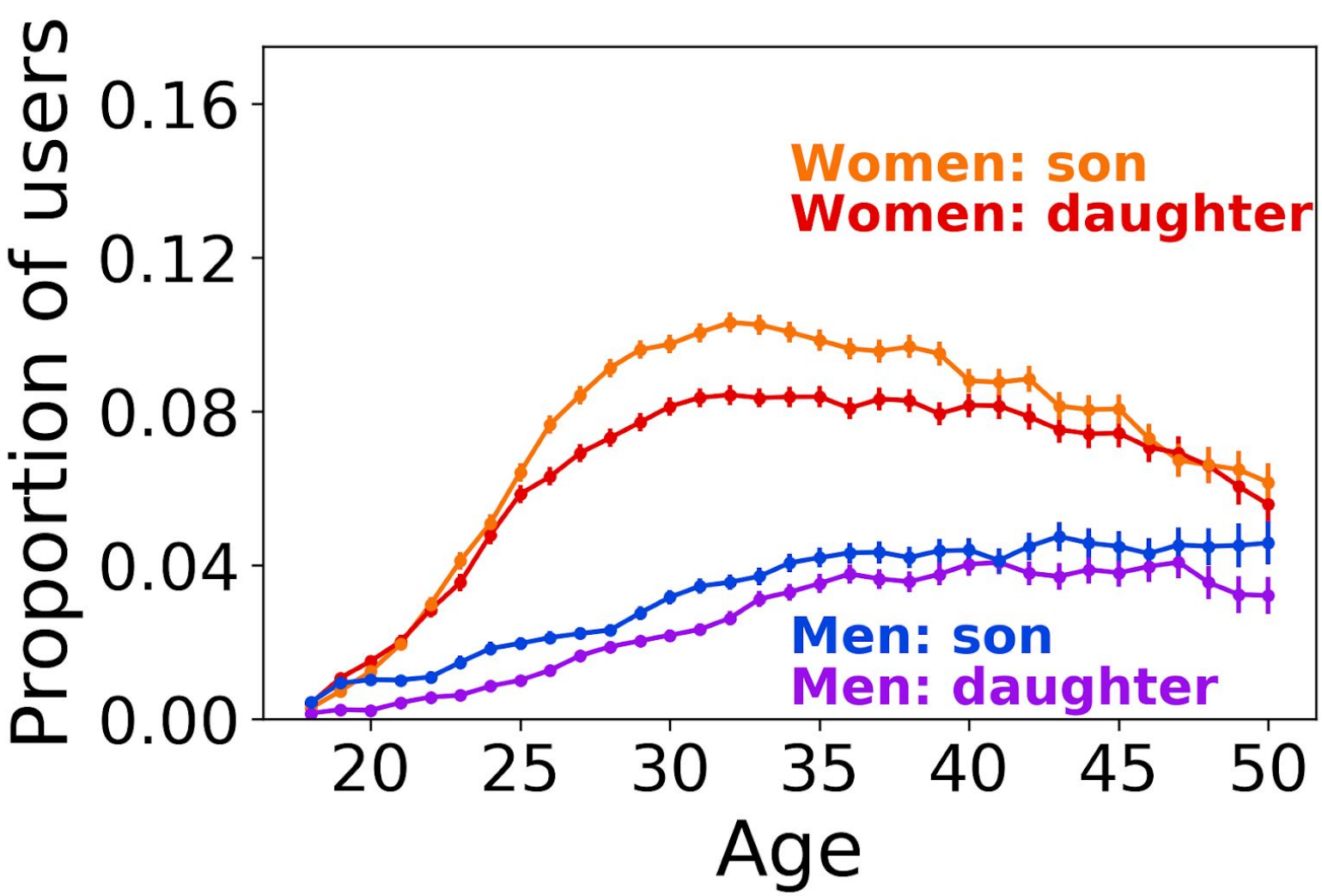
E Sivak, I Smirnov 2019 // *Proceedings of the National Academy of Sciences*

VK Data

“Russian Facebook”, 100M active users


Proportion of users
0.16
0.12
0.08
0.04
0.00
20
25
30
35
40
45
50
Age


Proportion of users
0.16
0.12
0.08
0.04
0.00
Women
Men
20
25
30
35
40
45
50
Age


Proportion of users
0.16
0.12
0.08
0.04
0.00
20
25
30
35
40
45
50
Age
Women: son
Women: daughter
Men: son
Men: daughter

That is it

What make it possible?

What make it possible?

Data

Methods

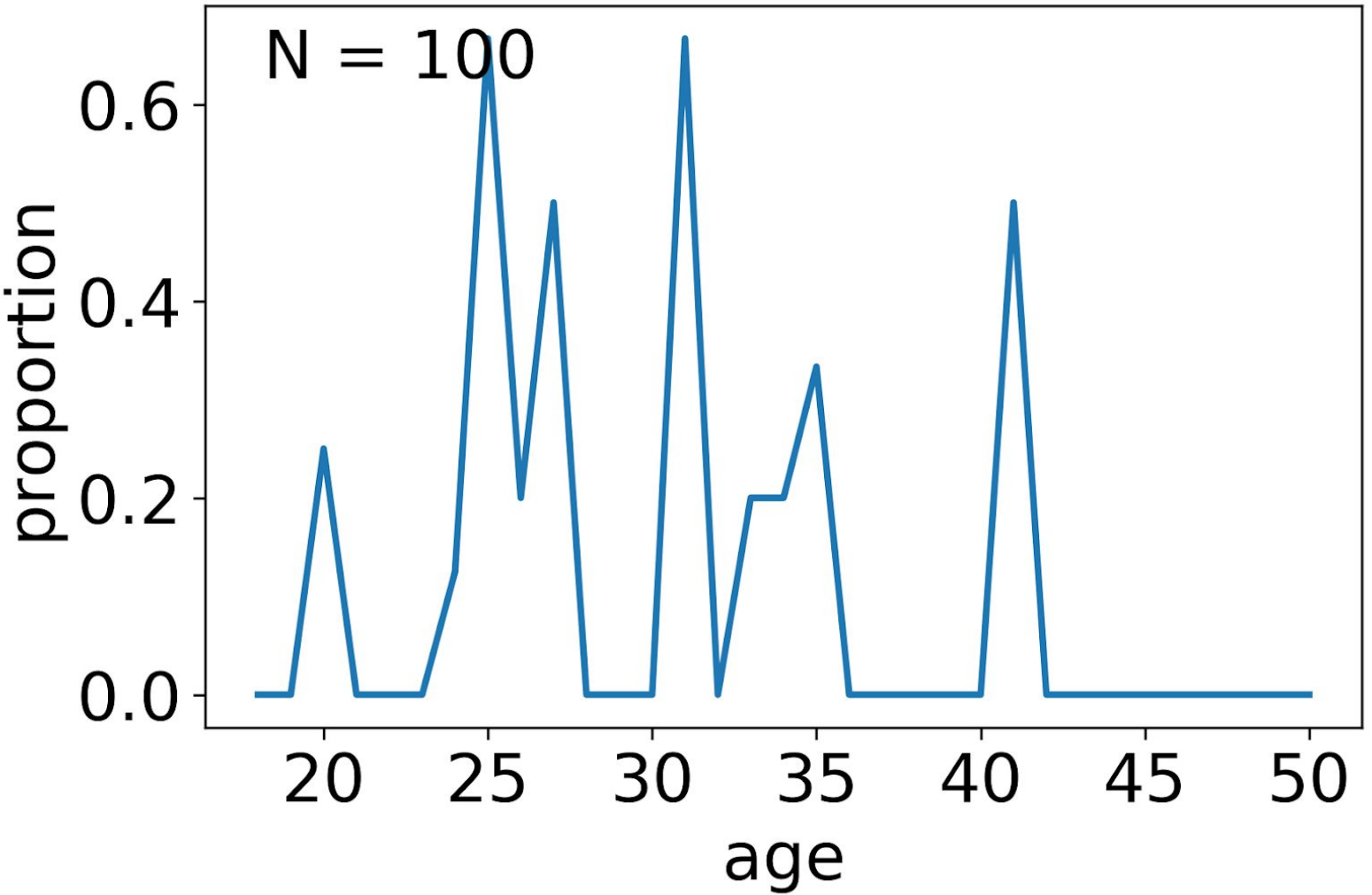
Theory

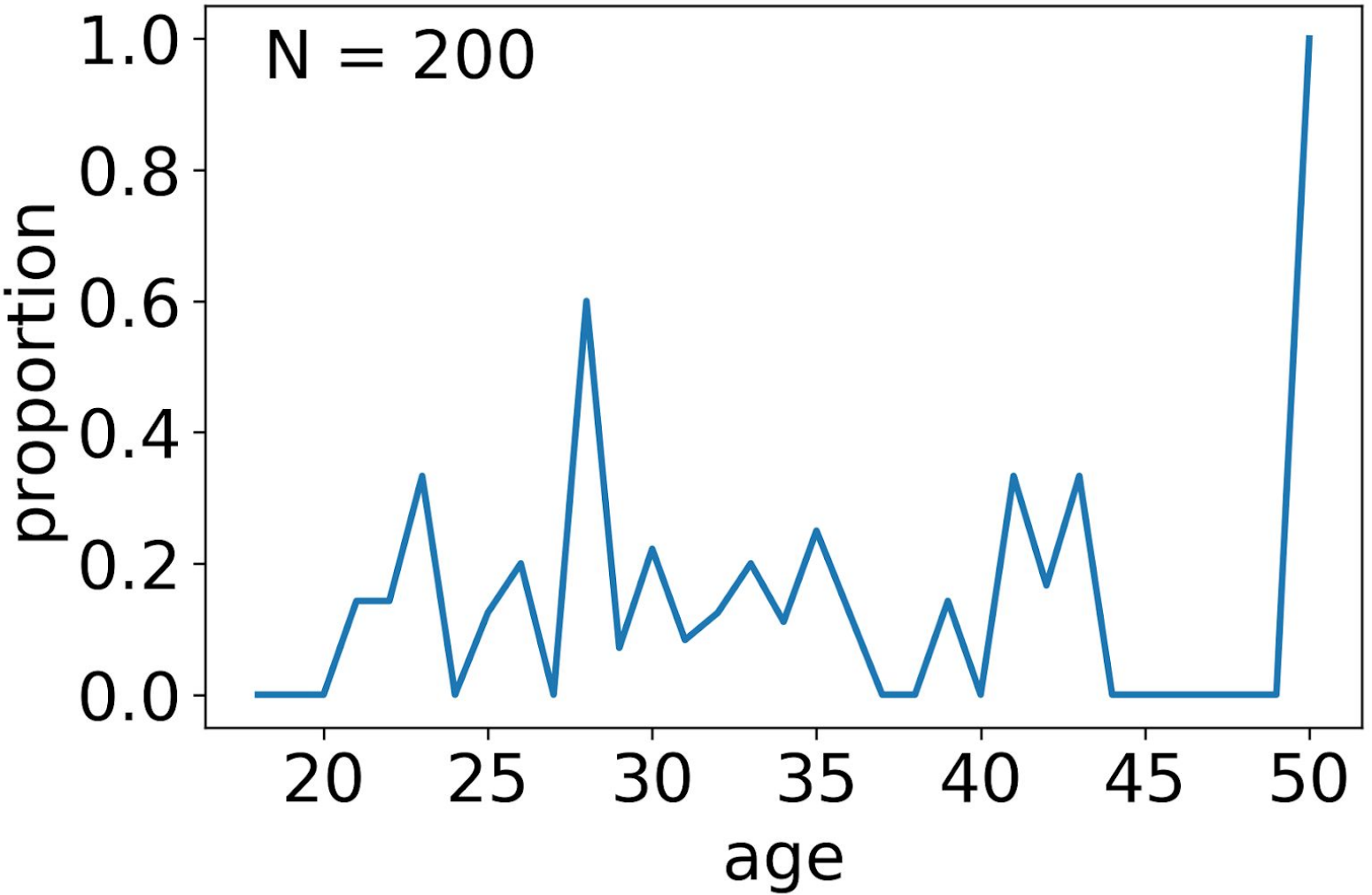
What make it possible?

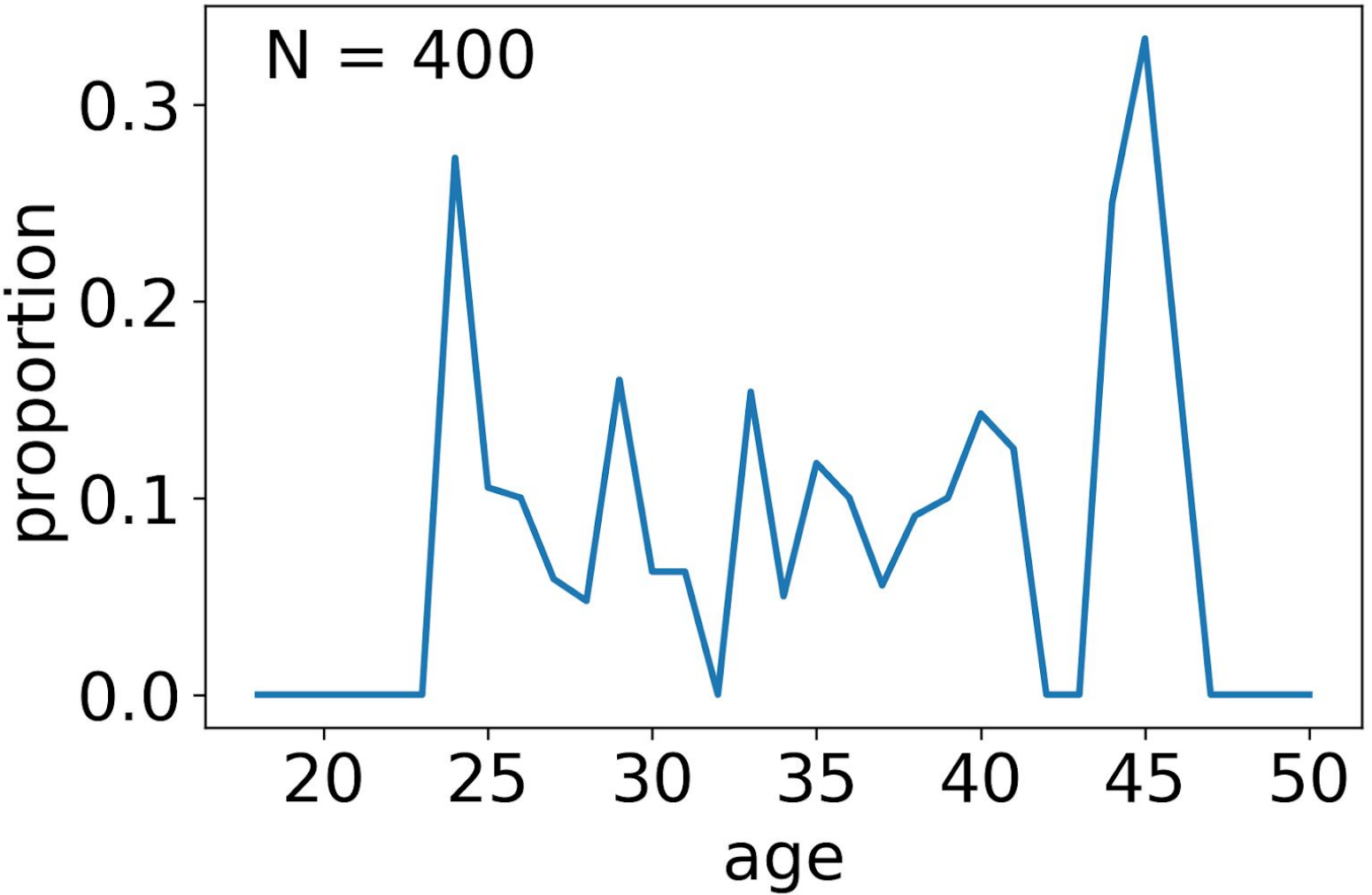
Data

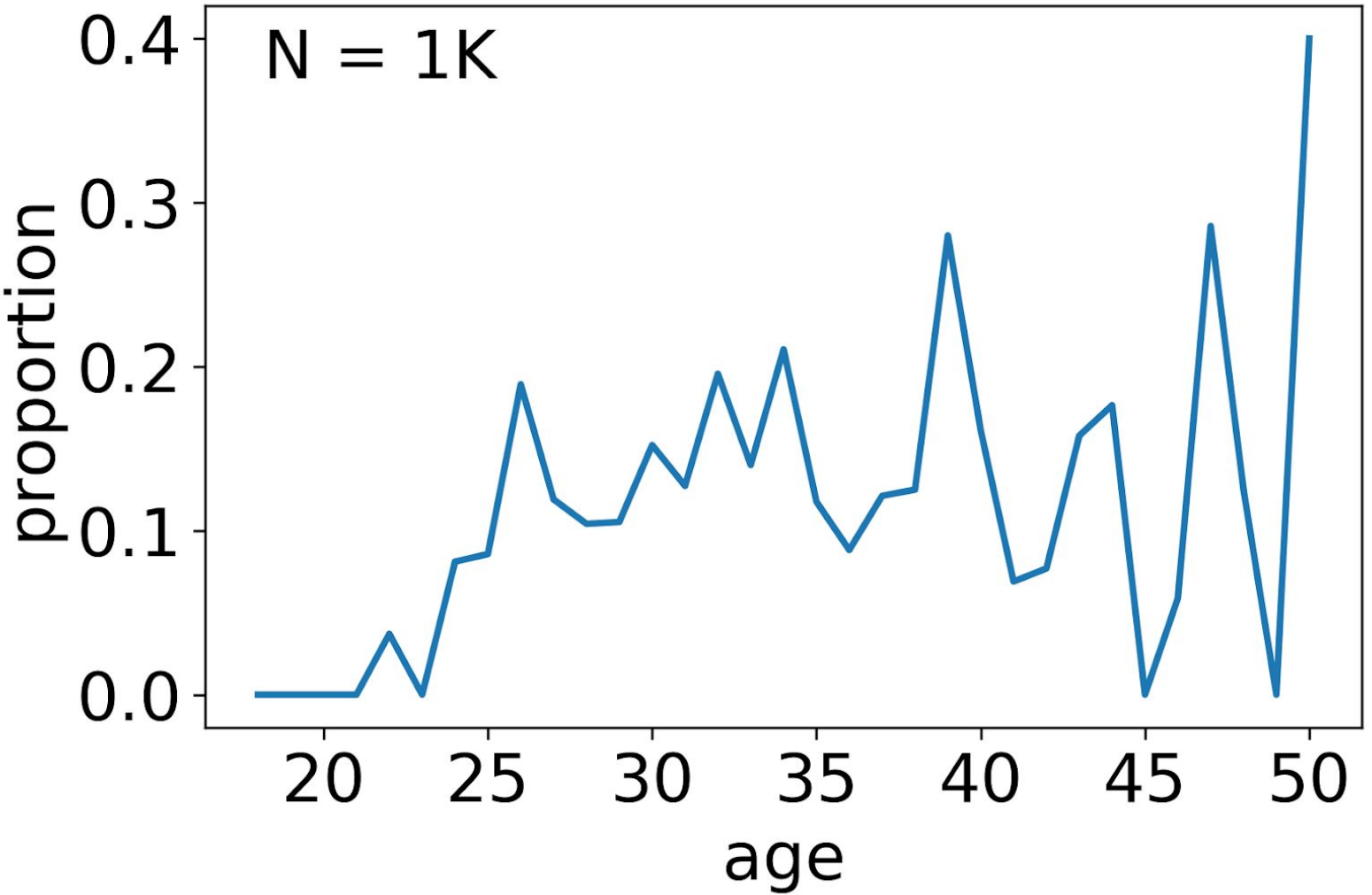
Methods

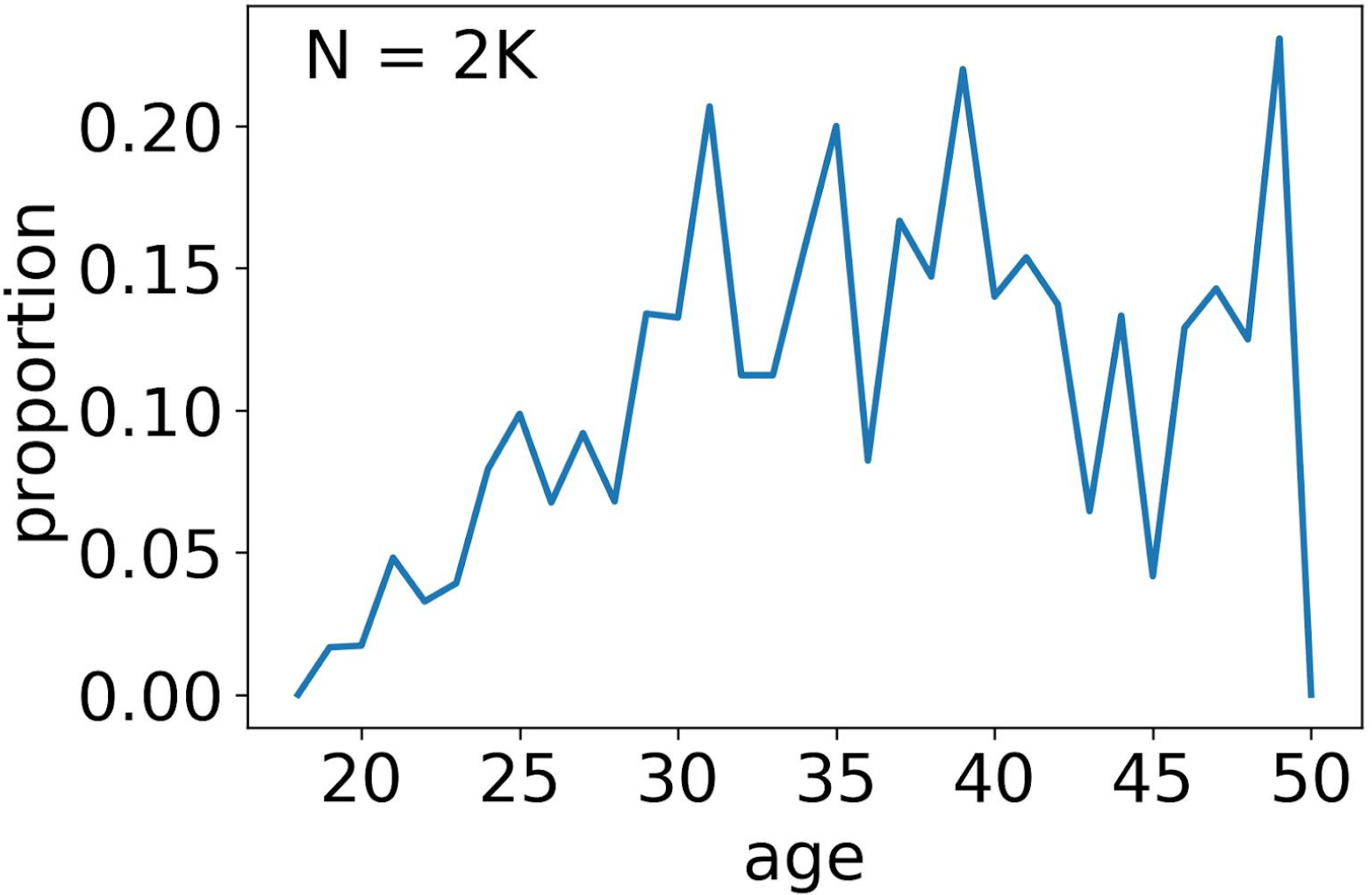
Theory

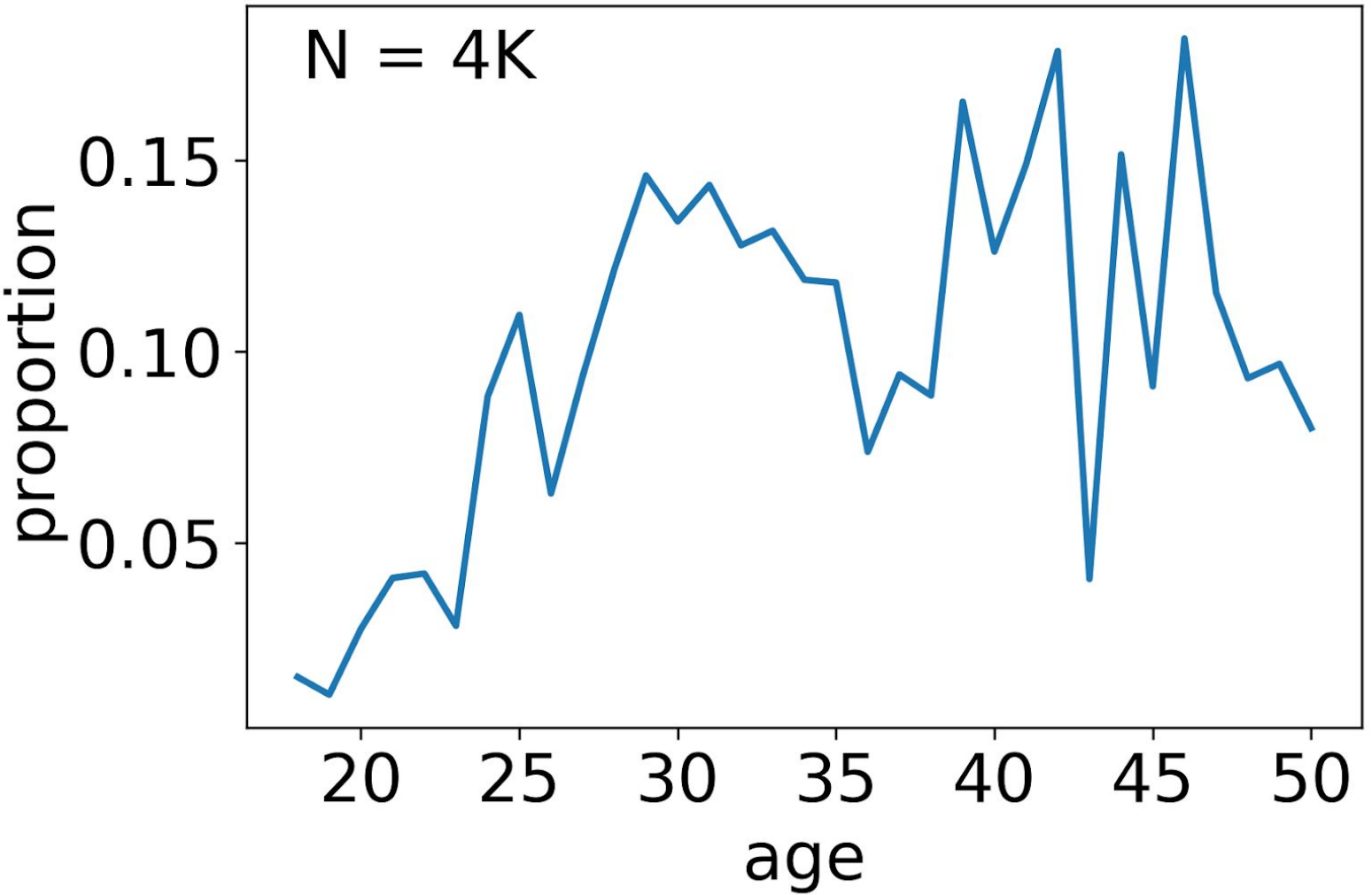

N = 100
proportion
0.0
0.2
0.4
0.6
20
25
30
35
40
45
50
age

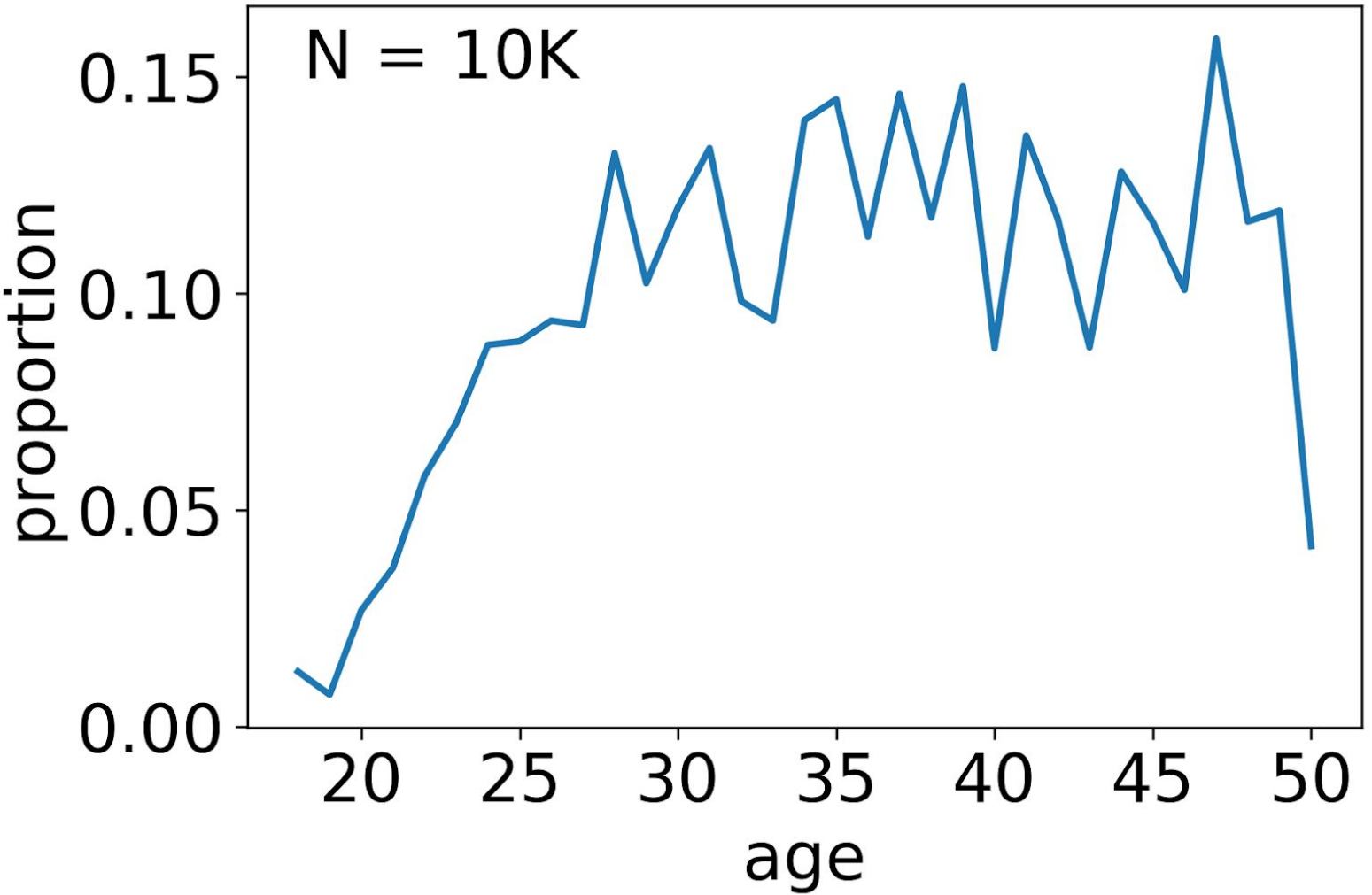

N = 200
proportion
1.0
0.8
0.6
0.4
0.2
0.0
20
25
30
35
40
45
50
age

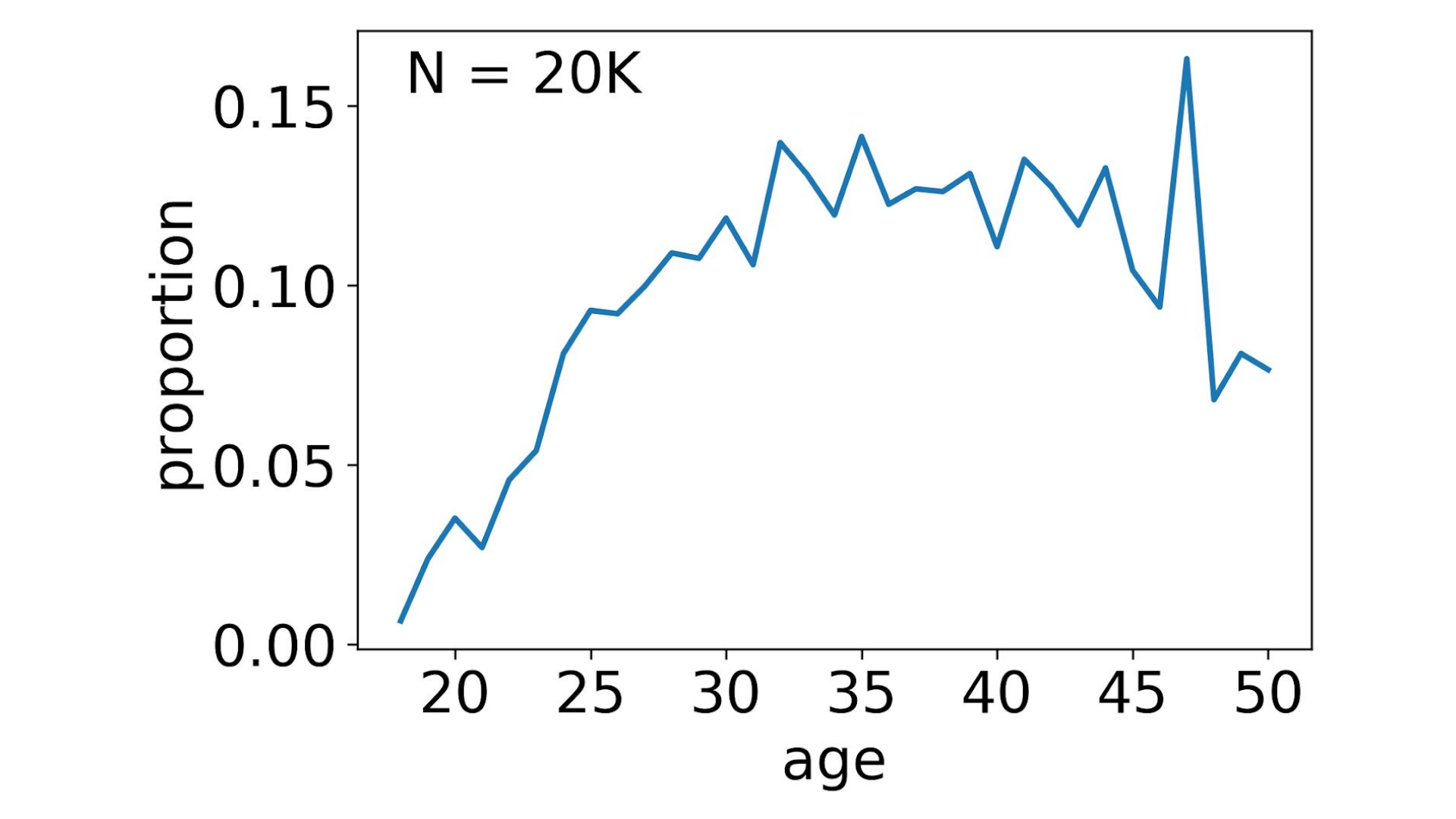

N = 400
proportion
0.0
0.1
0.2
0.3
20
25
30
35
40
45
50
age

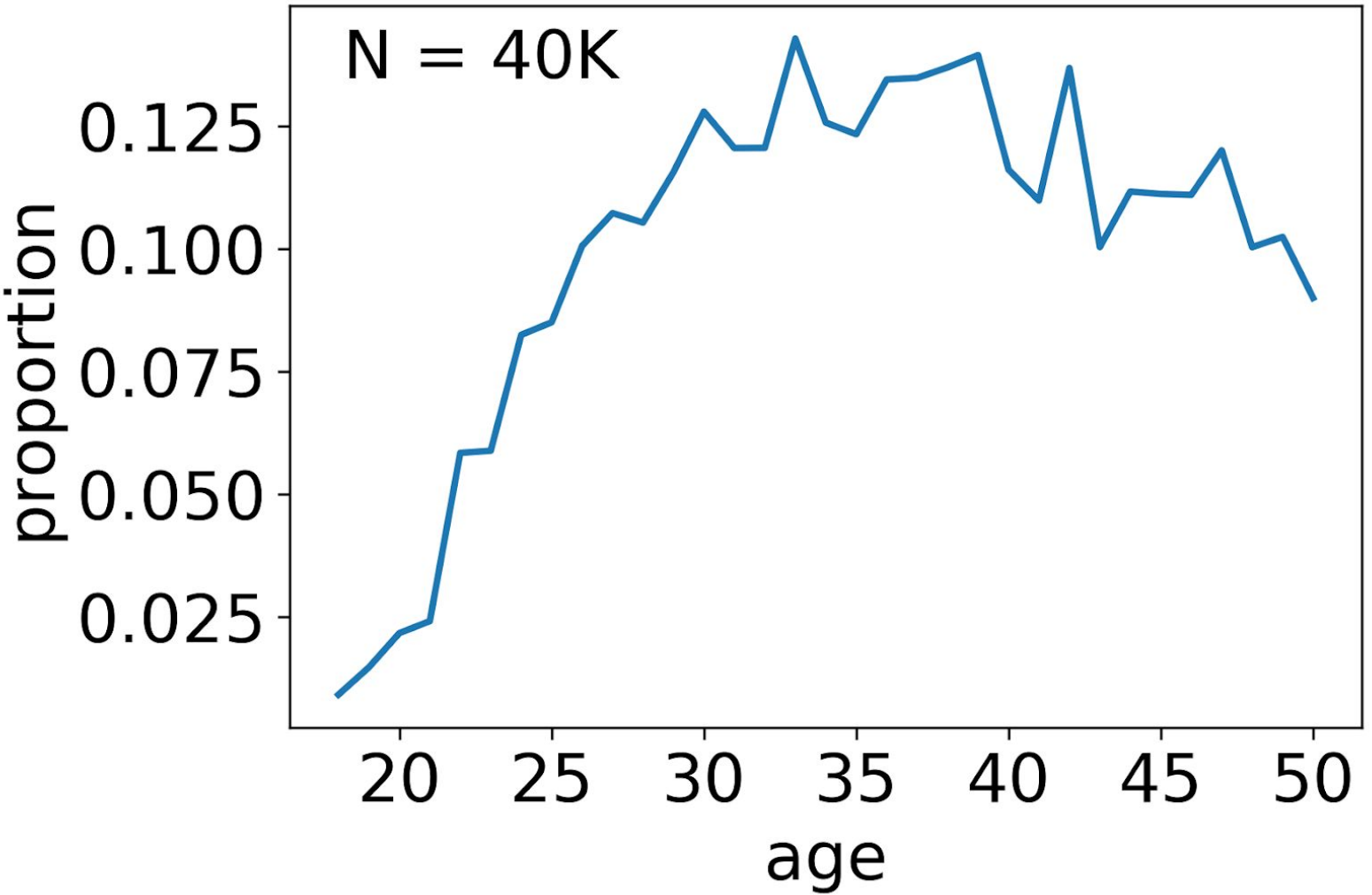

N = 1K
proportion
0.0
0.1
0.2
0.3
0.4
20
25
30
35
40
45
50
age

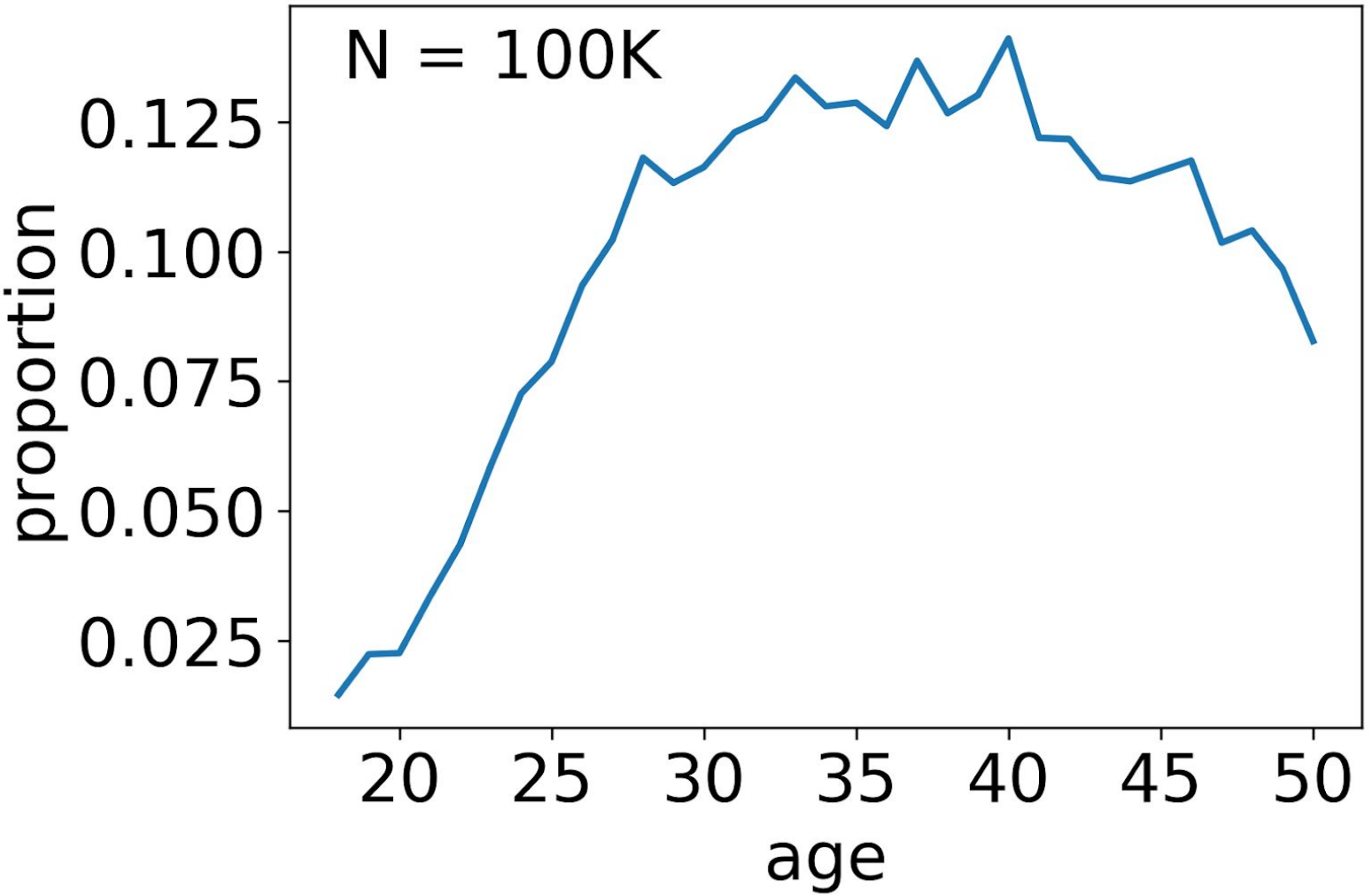

N = 2K
0.20
0.15
0.10
0.05
0.00
proportion
20
25
30
35
40
45
50
age

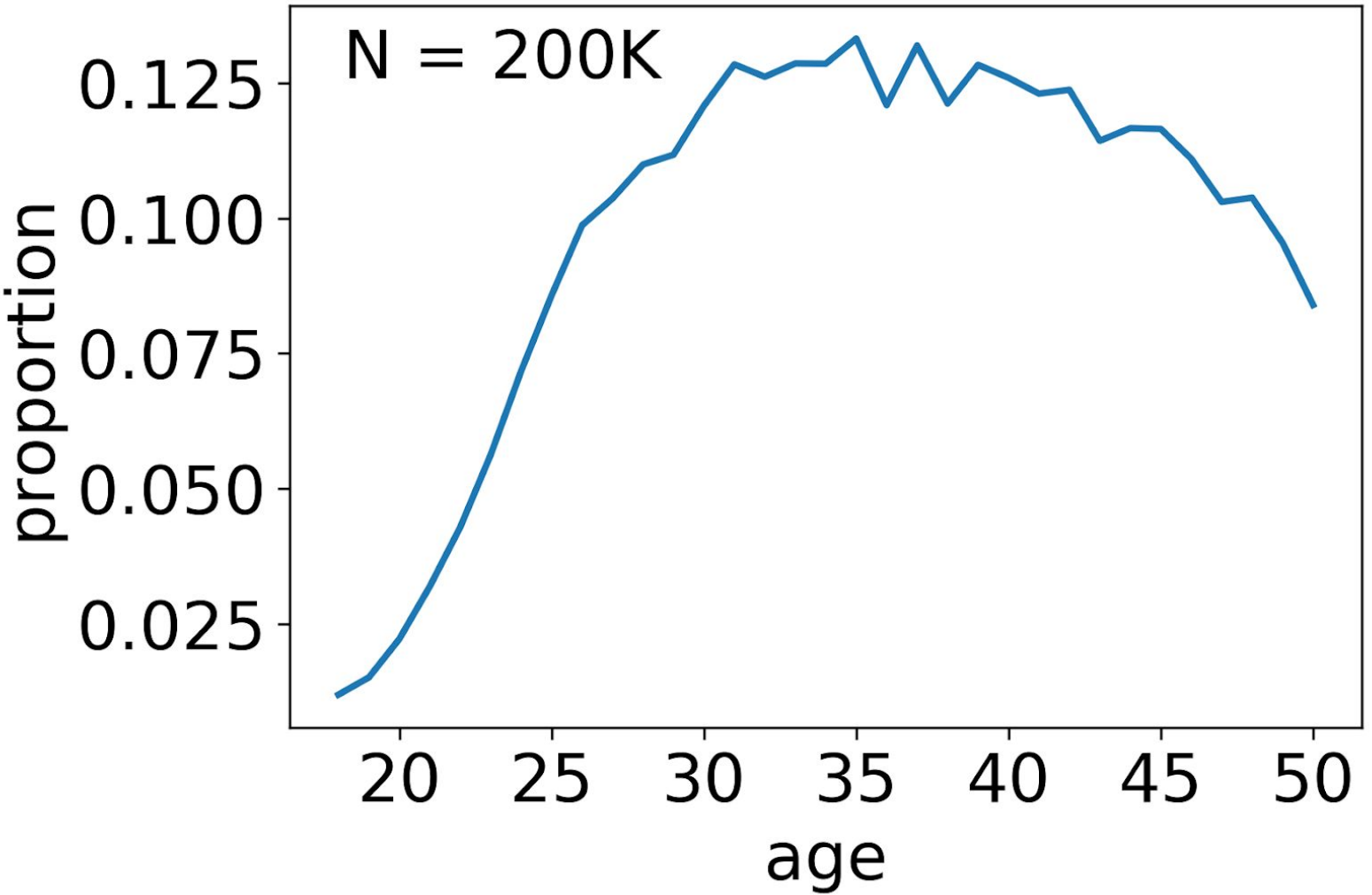

N = 4K
proportion
0.15
0.10
0.05
20
25
30
35
40
45
50
age

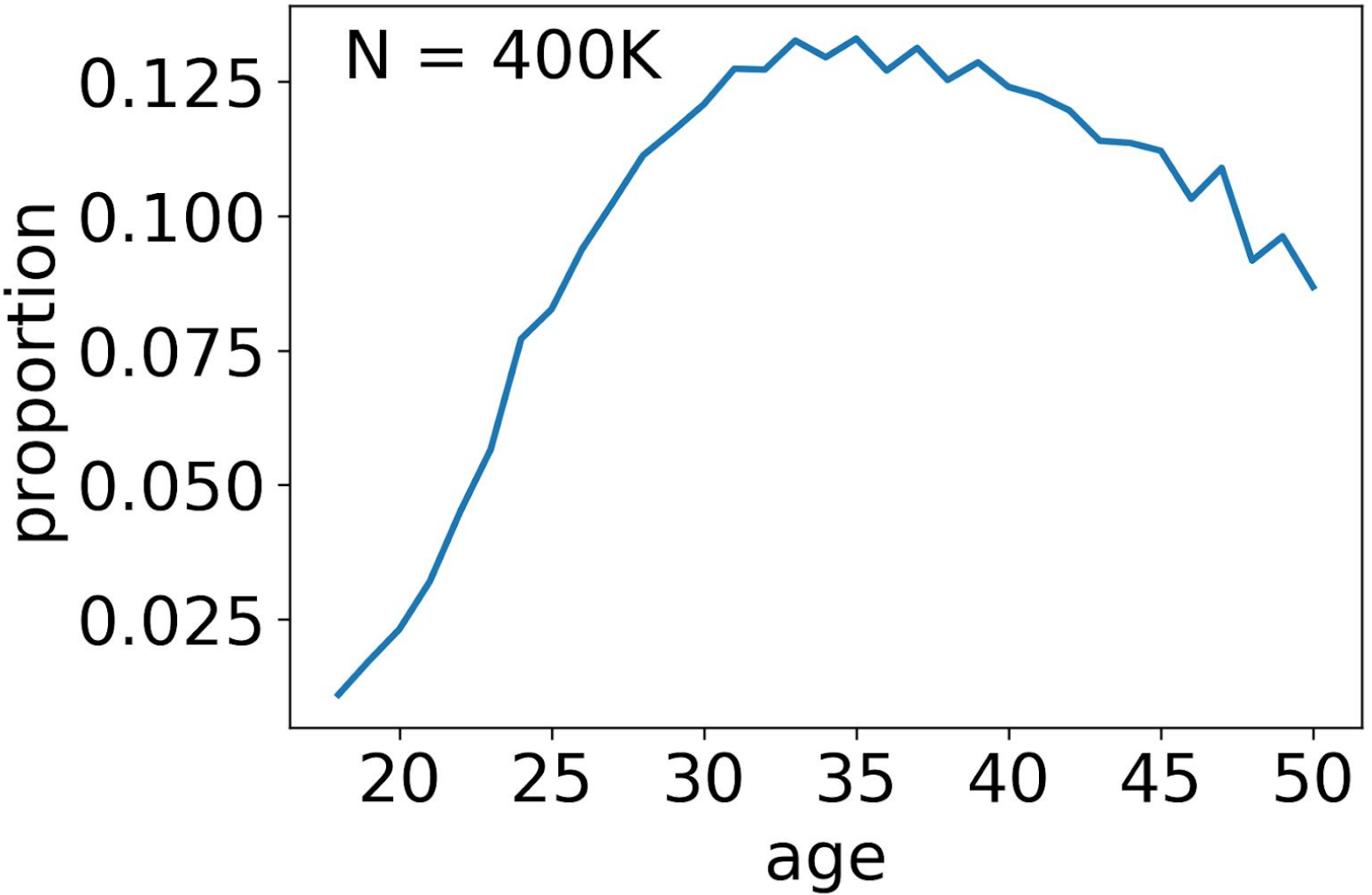

N = 10K
0.15
0.10
0.05
0.00
proportion
20
25
30
35
40
45
50
age

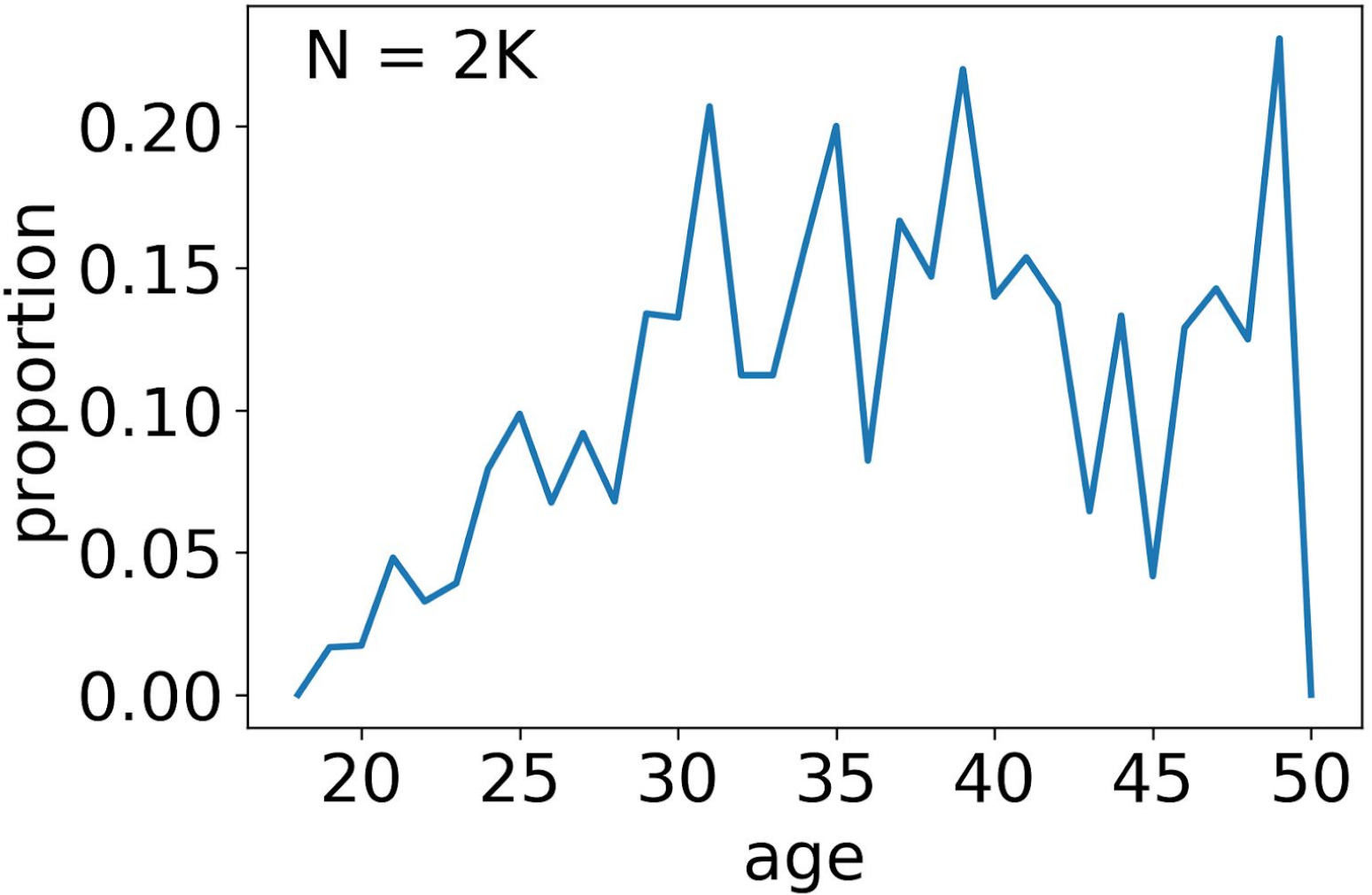

N = 20K
0.15
0.10
0.05
0.00
proportion
20
25
30
35
40
45
50
age


N = 40K
0.125
0.100
0.075
0.050
0.025
proportion
20
25
30
35
40
45
50
age


N = 100K
0.125
0.100
0.075
0.050
0.025
proportion
20
25
30
35
40
45
50
age

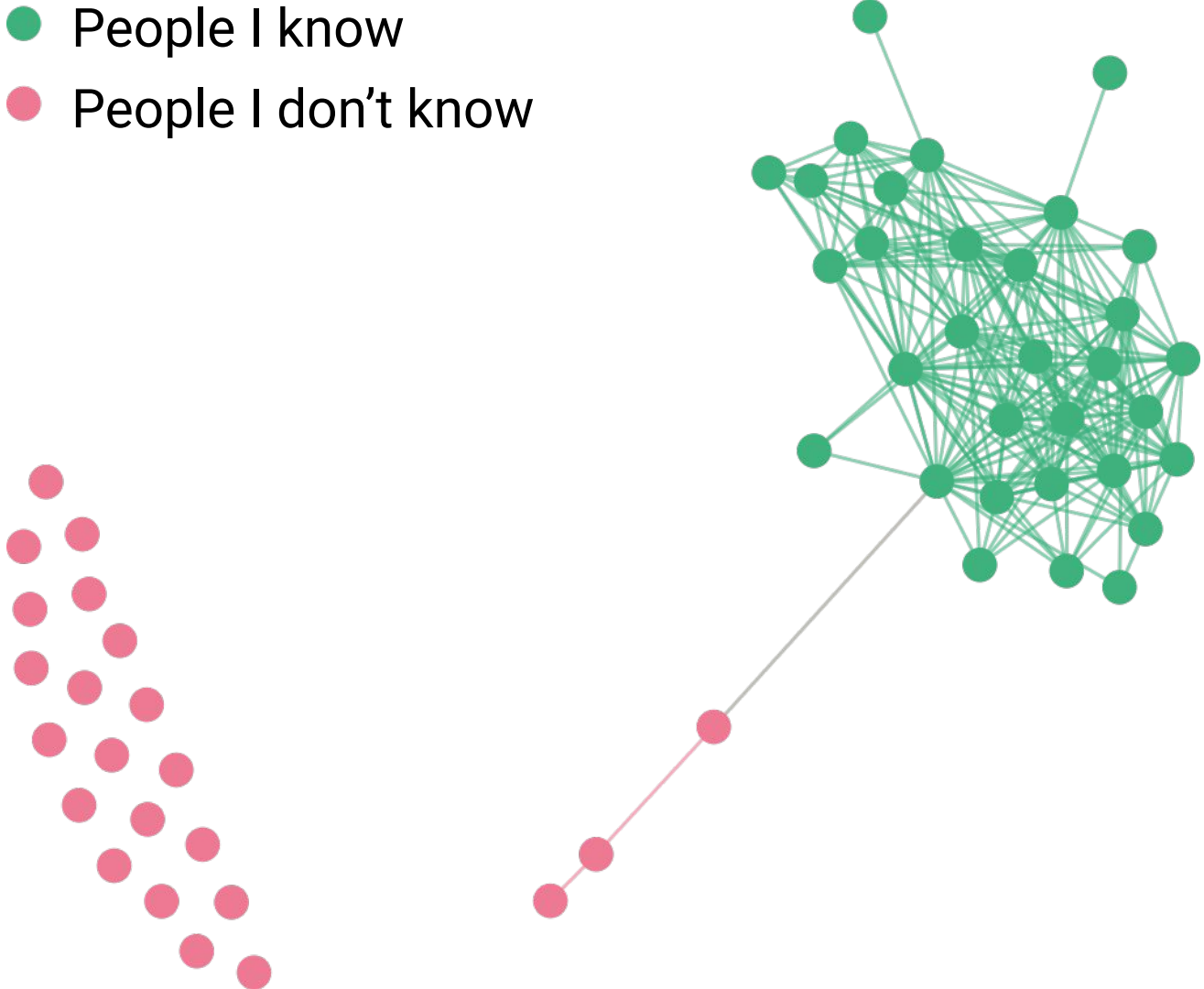

N = 200K
0.125
0.100
0.075
0.050
0.025
proportion
20
25
30
35
40
45
50
age

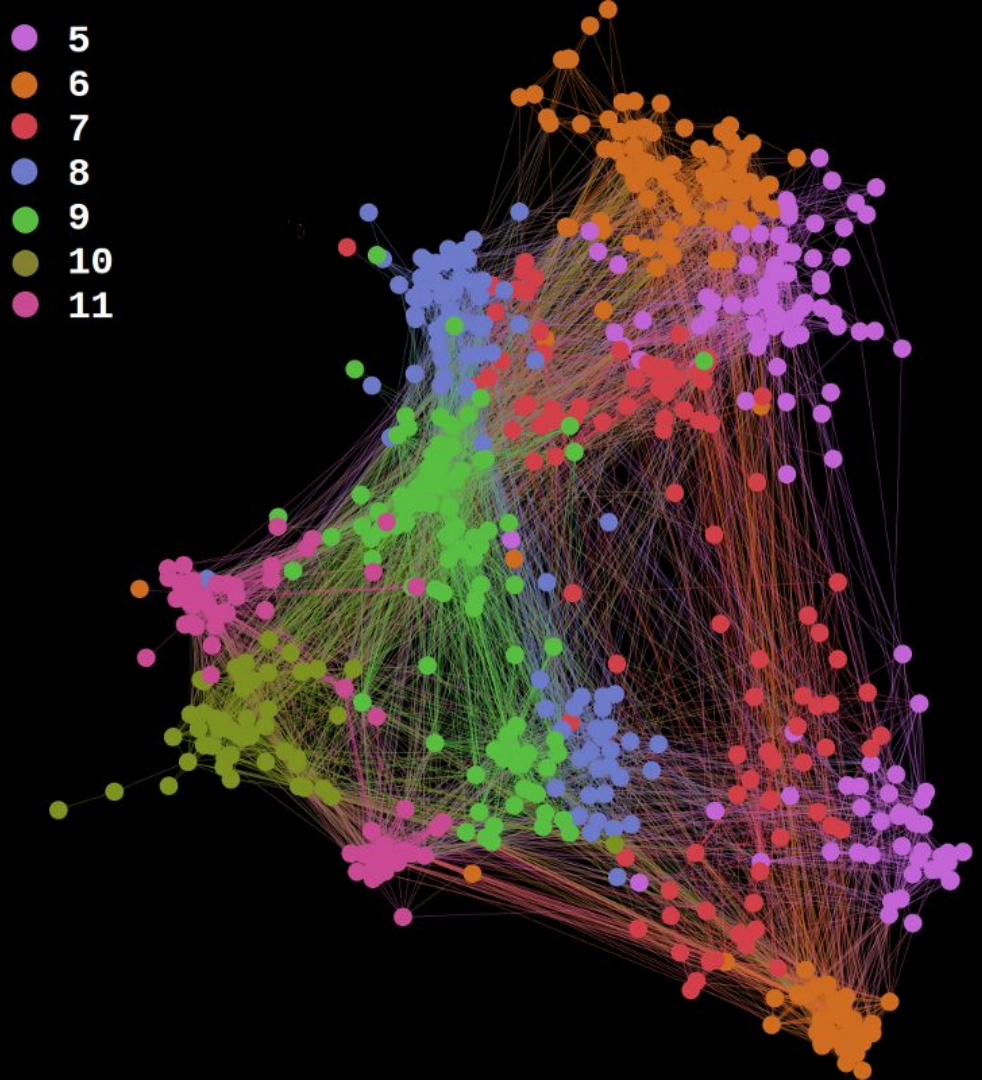

N = 400K
0.125
0.100
0.075
0.050
0.025
proportion
20
25
30
35
40
45
50
age

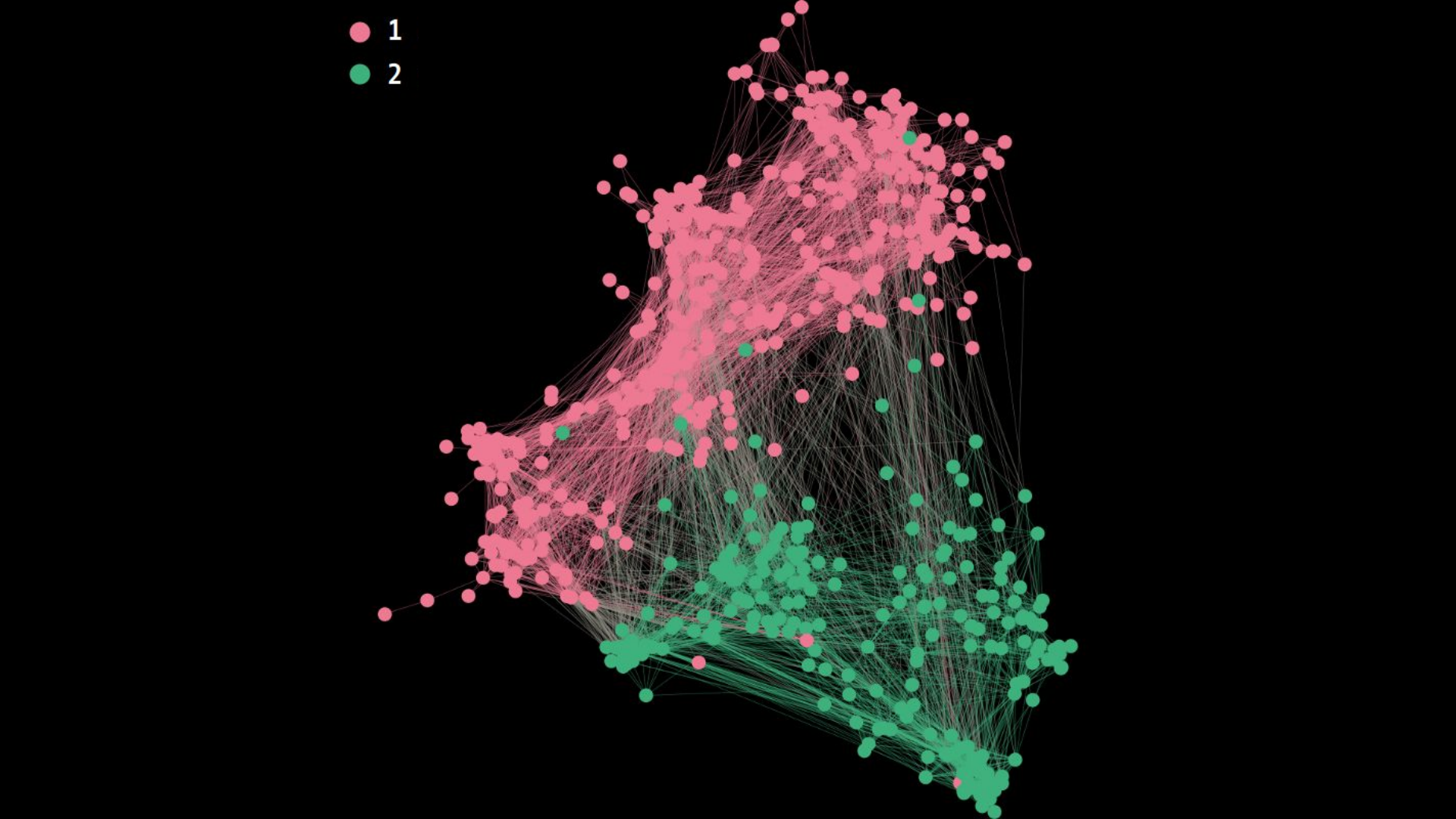

N = 2K
0.20
0.15
0.10
0.05
0.00
proportion
20
25
30
35
40
45
50
age

People I know

People I don’t know

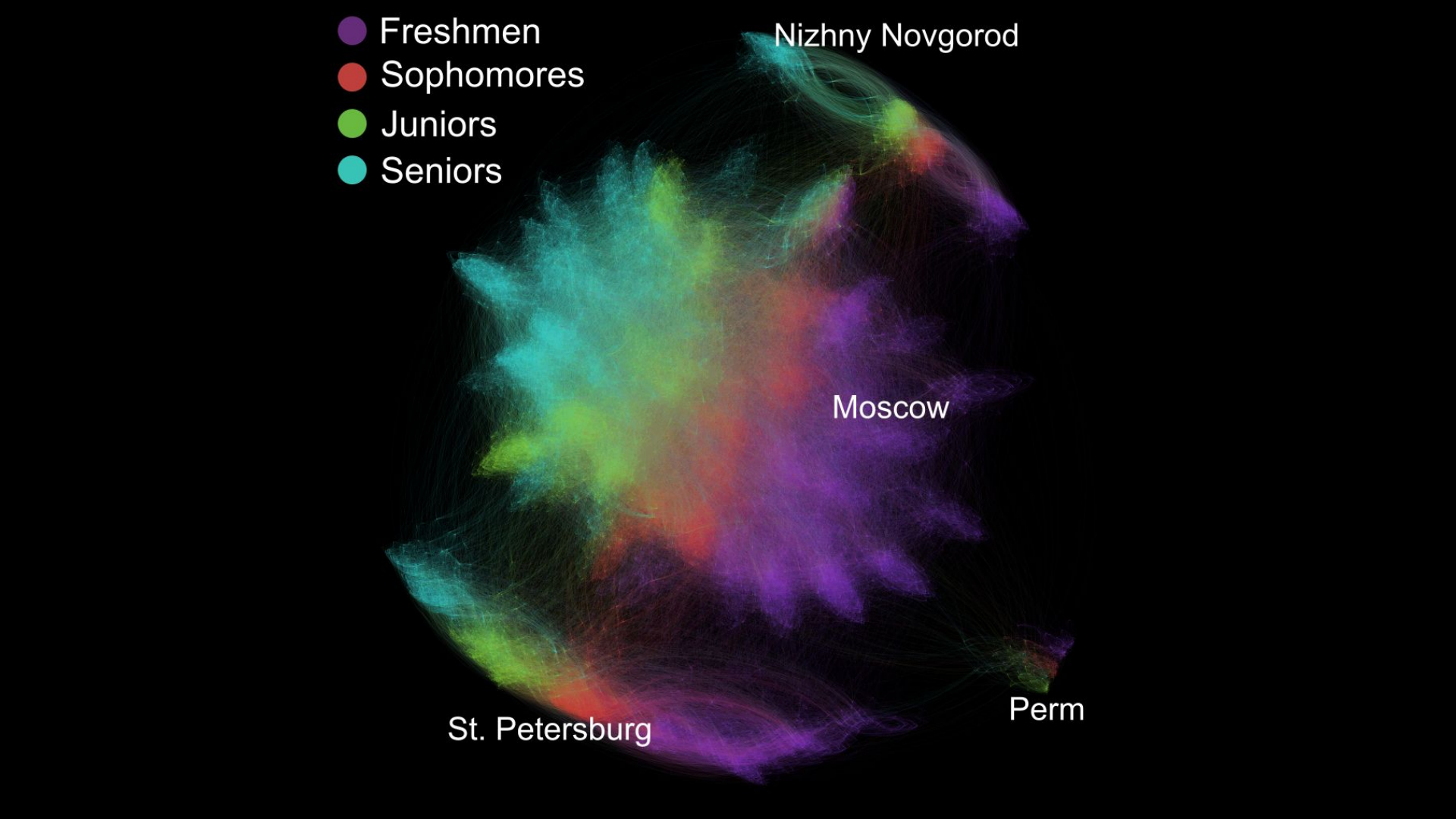


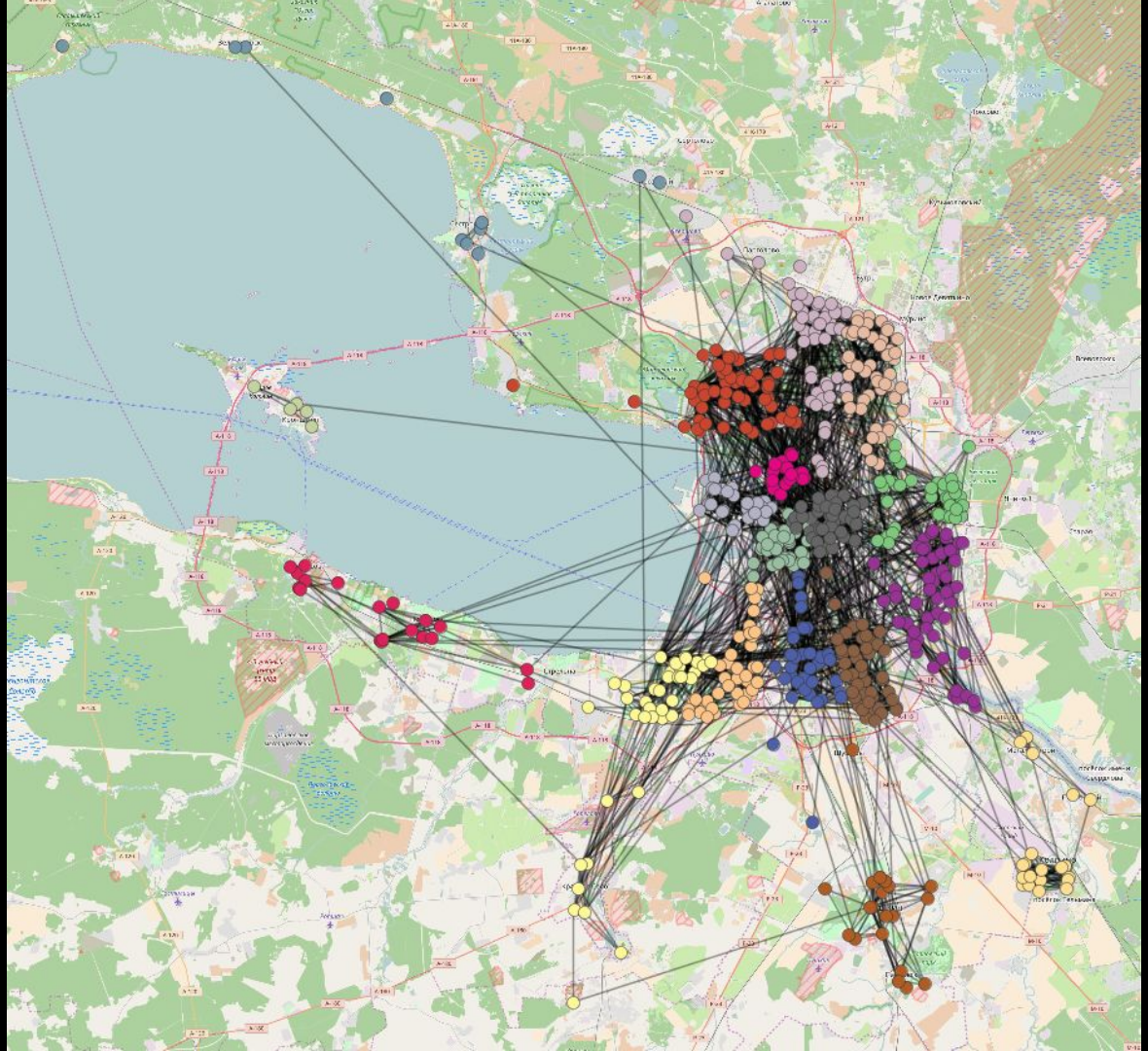

5
6
7
8
9
10
11

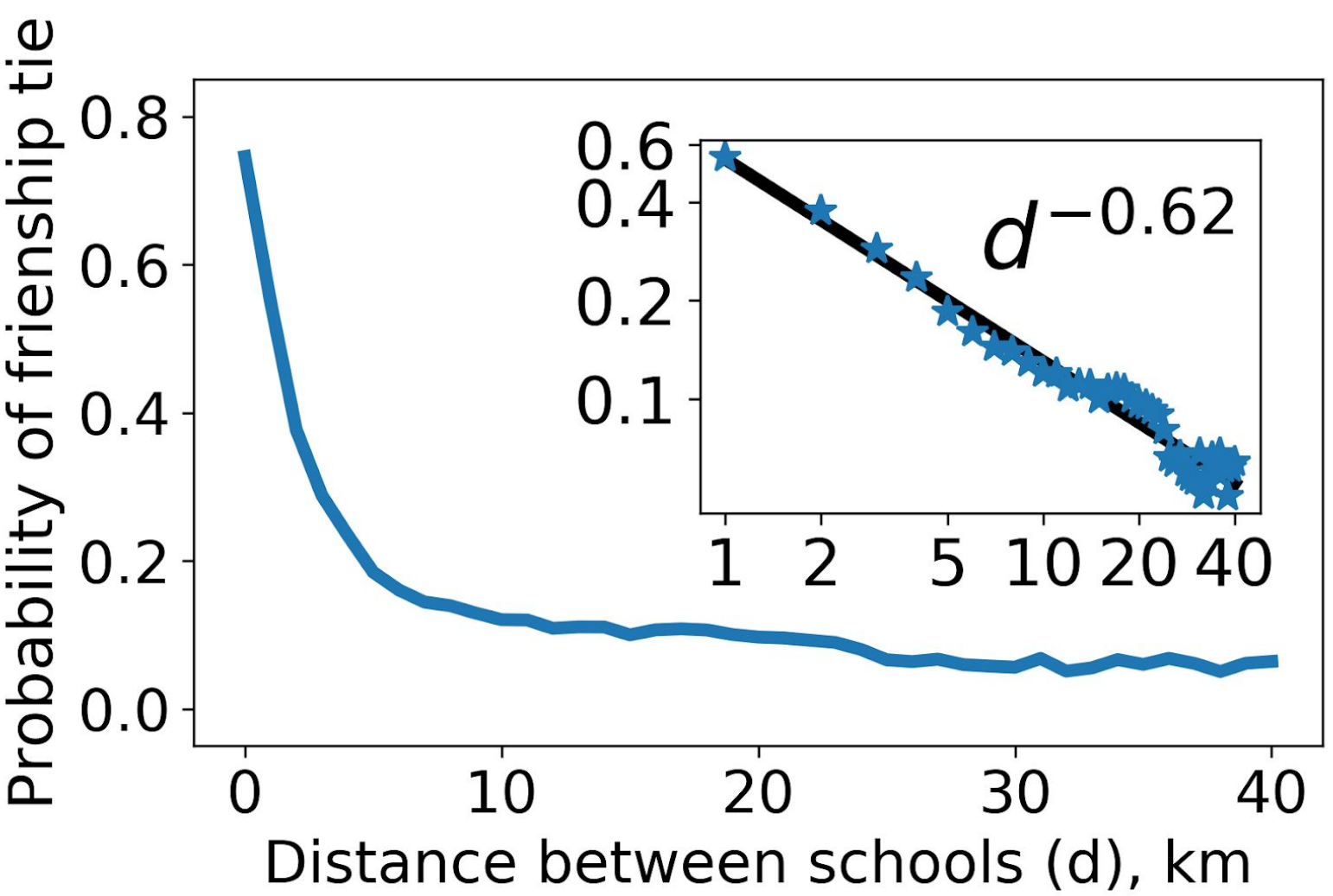

1
2

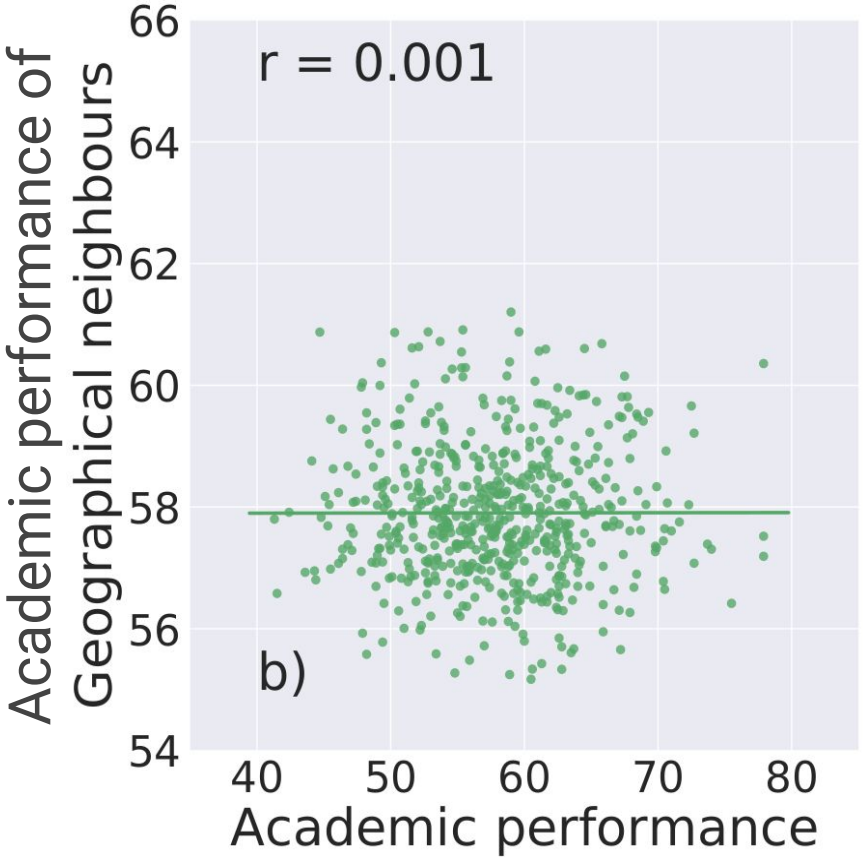
Grade	5	6	7	8	9	10	11
Found	85%	89%	88%	90%	88%	91%	85%

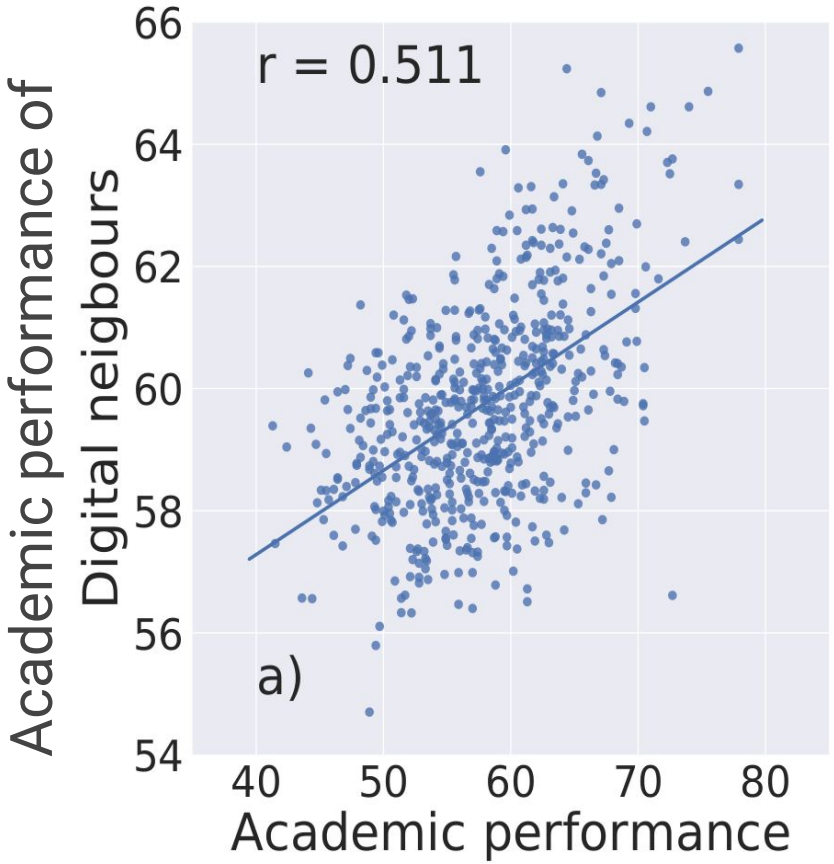
	Girls	GPA
Found on VK	46%	3.80
Not Found on VK	48%	3.79
Found via Friends	48%	3.77
Alternative Names	50%	3.79

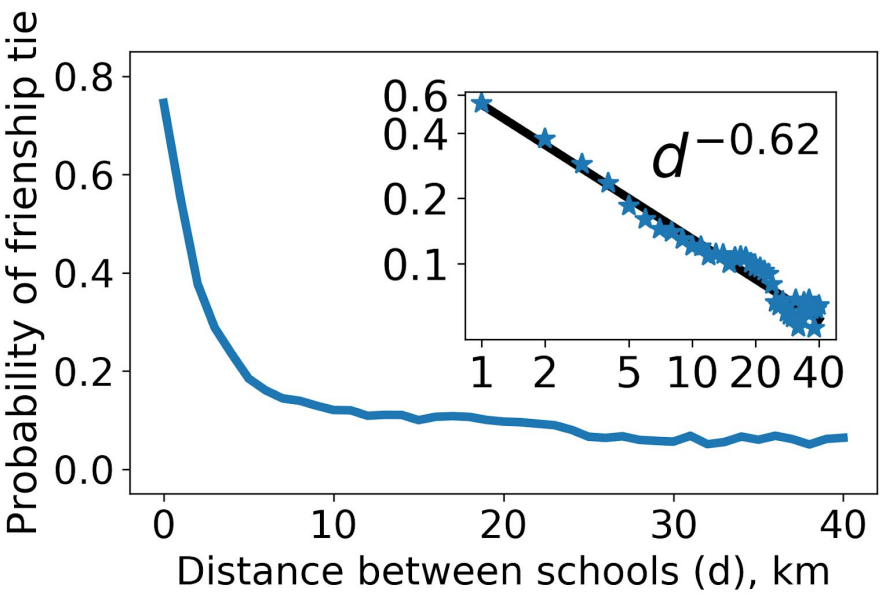

Freshmen
Sophomores
Juniors
Seniors
Nizhny Novgorod
Moscow
Perm
St. Petersburg

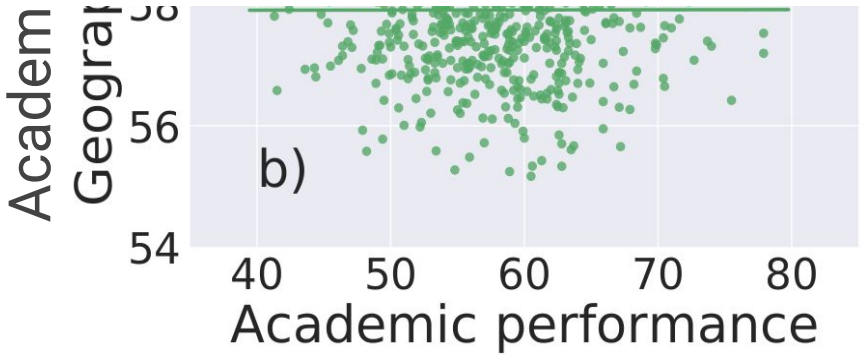


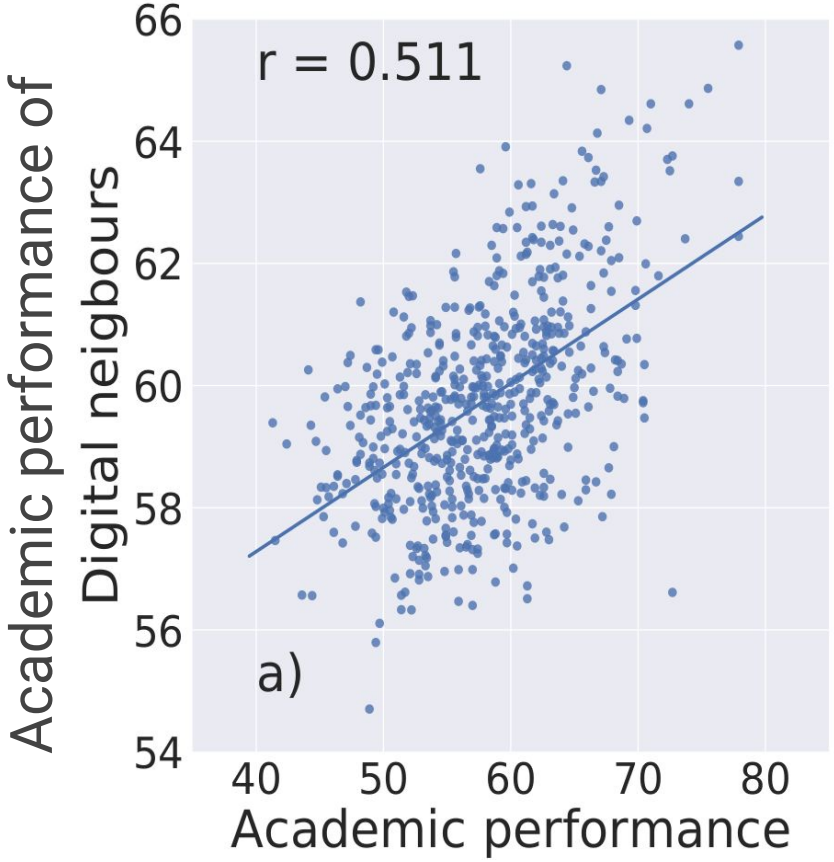

Probability of frienship tie
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Distance between schools (d), km
0.1
0.2
0.4
0.6
1
2
5
10
20
40
$d^{-0.62}$

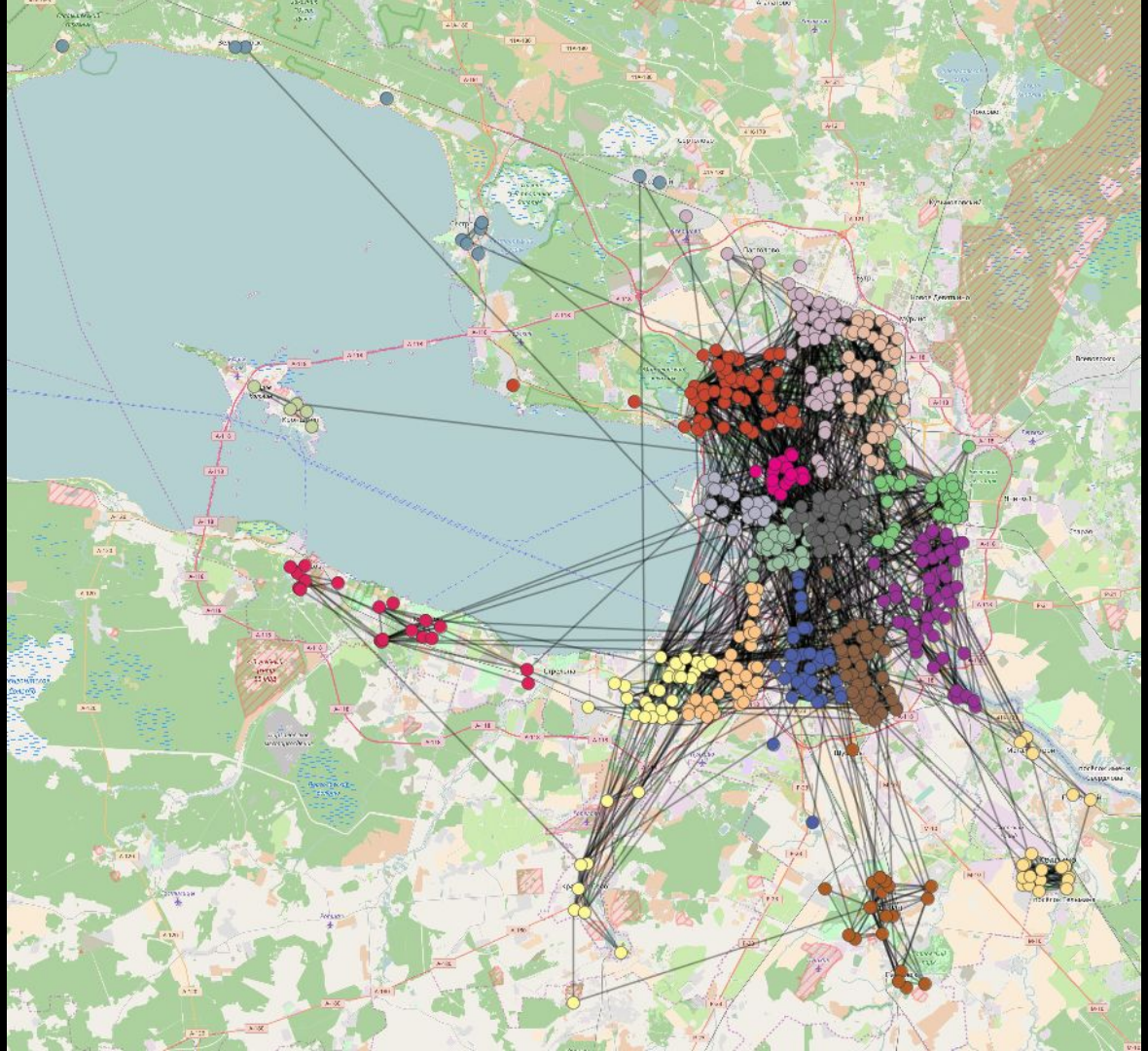

r = 0.001
b)
Academic performance of Geographical neighbours
Academic performance
66
64
62
60
58
56
54
40
50
60
70
80


r = 0.511
a)
Academic performance of Digital neigbours
Academic performance
66
64
62
60
58
56
54
40
50
60
70
80


Probability of friendship tie
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Distance between schools (d), km
$d^{-0.62}$
0.1
0.2
0.4
0.6
1
2
5
10
20
40


Academ
Geograp
54
56
b)
40
50
60
70
80
Academic performance


Academic performance of
Digital neigbours
54
56
58
60
62
64
66
r = 0.511
a)
40
50
60
70
80
Academic performance

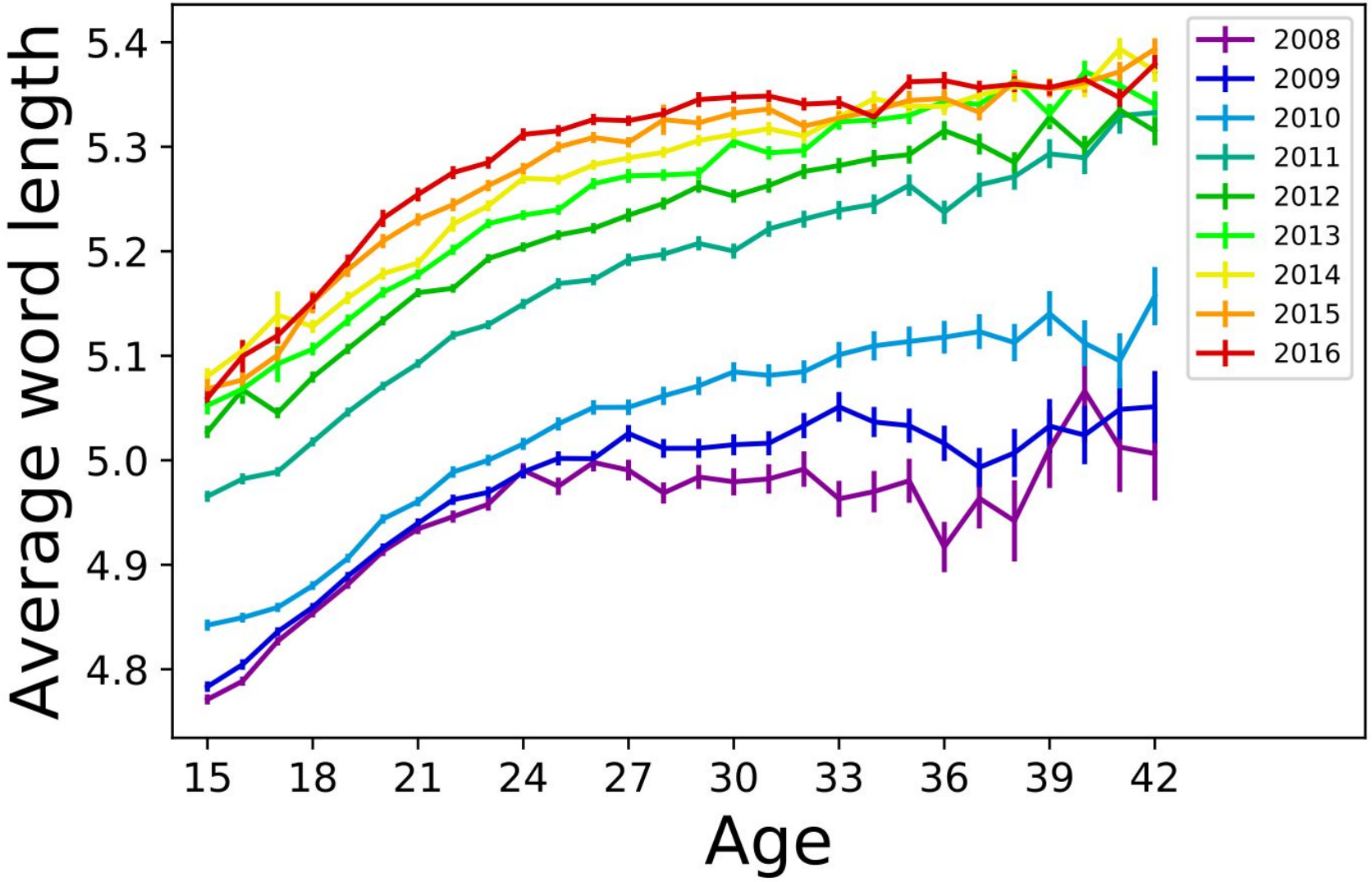


Data set

942,336 users

1B words

Systematic vs Random


Average word length
5.4
5.3
5.2
5.1
5.0
4.9
4.8
15
18
21
24
27
30
33
36
39
42
Age
2008
2009
2010
2011
2012
2013
2014
2015
2016

What make it possible?

Data

Methods

Theory

False positive

“Gazprom’s daughter”

False positive

“Gazprom’s daughter”

10,000 posts

False positive

“Gazprom’s daughter”

False negative

“so-o-o-o-o-on”, spelling mistakes, ...

False positive

“Gazprom’s daughter”

False negative

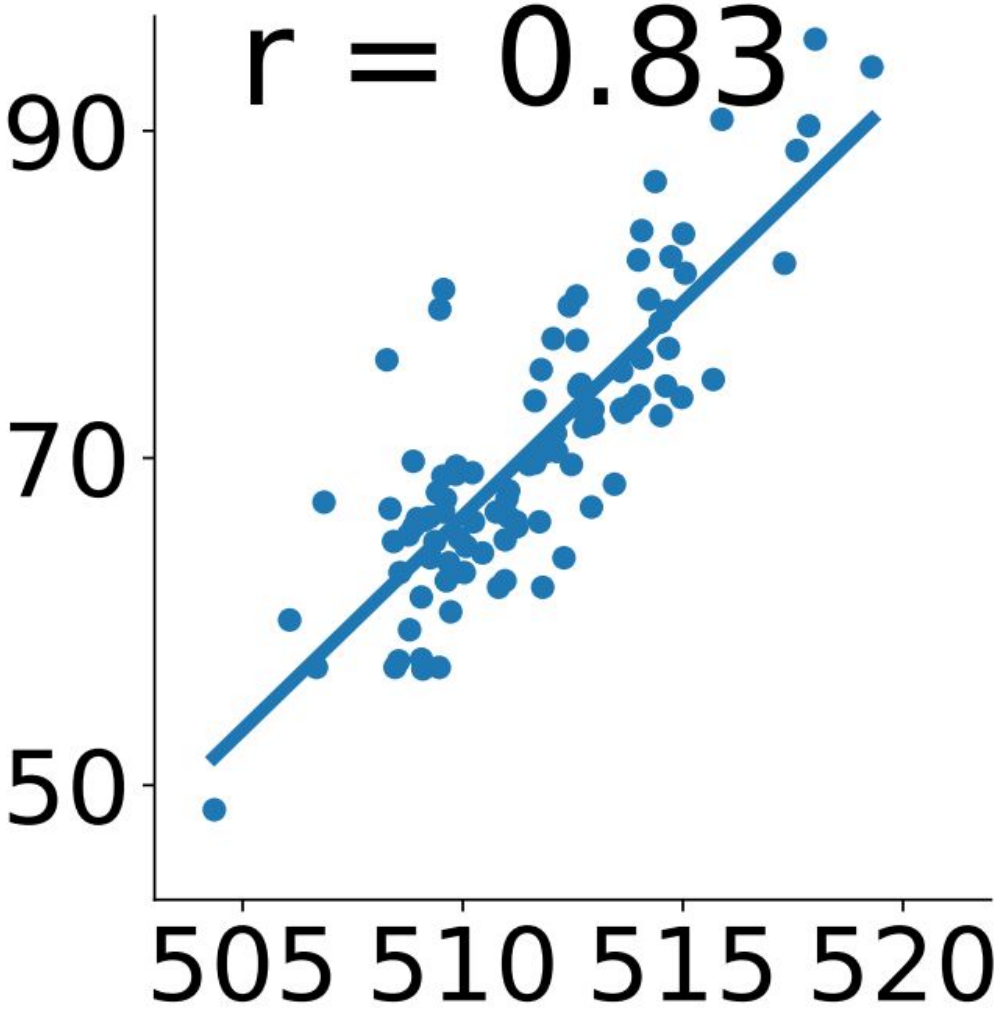
“so-o-o-o-o-on”, spelling mistakes, …

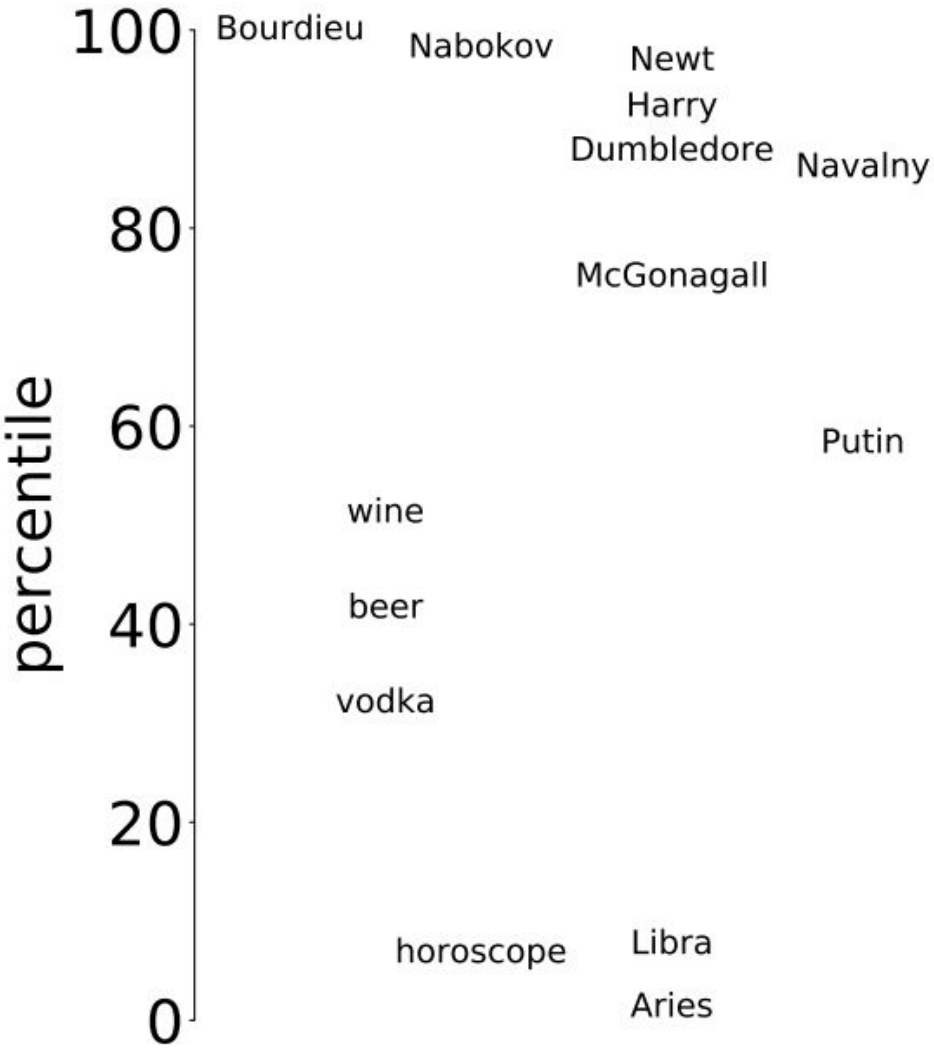
Word embeddings trained on VK

Example

Predicting academic performance from posts

Applying to 115K users from 100 universities


r = 0.83
90
70
50
505
510
515
520


100
80
60
40
20
0
percentile
Bourdieu
Nabokov
Newt
Harry
Dumbledore
Navalny
McGonagall
Putin
wine
beer
vodka
horoscope
Libra
Aries

What make it possible?

Data

Methods

Theory

Theory

where to look?

infinite number of potential actions

Theory

where to look?

how to interpret?

Son bias

Sex-selective abortions

Breastfeeding

Pastime

Marriage

Investment

Female characters underrepresented

Books, movies, textbooks, etc.

FRAMED by GENDER

How Gender Inequality Persists in the Modern World

CECILIA L. RIDGEWAY

Why physicists/computer scientists often fail in social science?

Why physicists/computer scientists often fail in social science?

Explicit vs implicit assumptions

Nobel Prize in Economic Sciences

Nobel Prize in Economic Sciences

Behavioral economics

Nobel Prize in Economic Sciences

Behavioral economics

Signaling theory

Nobel Prize in Economic Sciences

Behavioral economics

Signaling theory

...

Ivan Smirnov

@ibsmirnov

#computational social science #Fair AI
#education #inequality #social networks
#emotional well-being #gender #ML

General interest journals

Nature, Science, PNAS, Science Advances, Nature Human Behaviour, etc.

Old

Nature (1869), Science (1880), PNAS (1914)

And new

PLOS One (2006), Scientific Reports (2011), Science Advances (2015)

Selective

Nature, Science ~7%

Selective

Nature, Science ~7%

And not so selective

PLOS One, Scientific Reports ~50%

Any scientific field

Should be interesting and understandable to general scientific audience

Special format

Prestigious selective journals

A lot of citations

Media attention

Very important to career

DORA

San Francisco Declaration on Research Assessment

General Recommendation

1. Do not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist’s contributions, or in hiring, promotion, or funding decisions.

Main text

Science (up to ~2500 words, including references, notes and captions)

+ Supplementary Materials/Information

People with more friends and more social ties in their community tend to live longer (1–4). Many researchers interpret this association as evidence that greater social support and social network integration lead to better health outcomes (4). For example, social integration is thought to improve health by motivating engagement in healthy behaviors (5, 6), improving immunity (7), and reducing inflammation (8).

Abstract

Logical order: describe your results — summarize them in abstract

Abstract

Logical order: describe your results — summarize them in abstract

Result: so-so

Abstract

Logical order: describe your results — summarize them in abstract

Result: so-so

But should be perfect!

How people read papers:

Title -> Abstract -> Figures -> Everything else

Losing readers at each step!

Should master concise writing

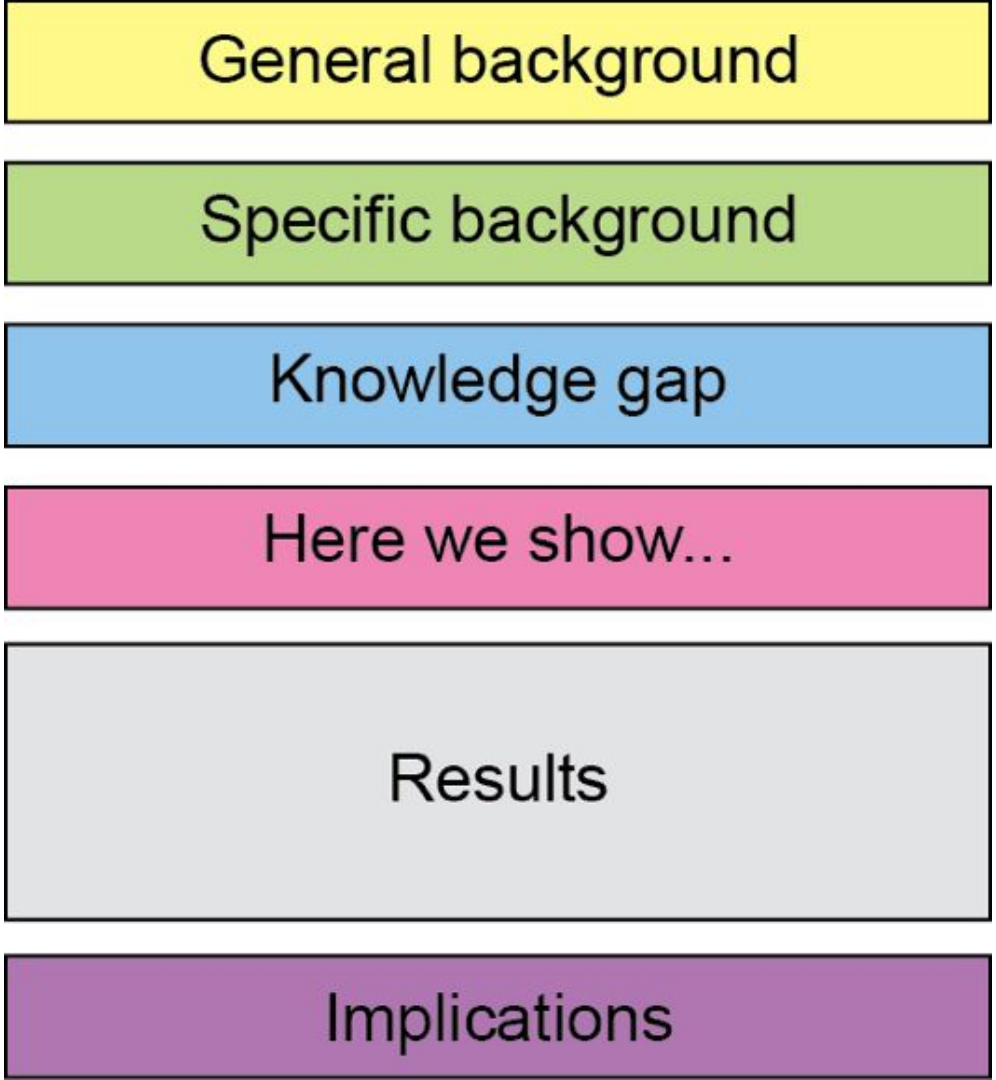
PNAS: 250 words

Science: 125 words

PNAS Significance Statement: 120 words

Should master concise writing

There is a typical structure



https://mitcommlab.mit.edu/broad/commkit/journal-article-abstract/

Social interactions increasingly take place online. Friendships and other offline social ties have been repeatedly associated with human longevity, but online interactions might have different properties. Here, we reference 12 million social media profiles against California Department of Public Health vital records and use longitudinal statistical models to assess whether social media use is associated with longer life. The results show that receiving requests to connect as friends online is associated with reduced mortality but initiating friendships is not. Additionally, online behaviors that indicate face-to-face social activity (like posting photos) are associated with reduced mortality, but online-only behaviors (like sending messages) have a nonlinear relationship, where moderate use is associated with the lowest mortality. These results suggest that online social integration is linked to lower risk for a wide variety of critical health problems. Although this is an associational study, it may be an important step in understanding how, on a global scale, online social networks might be adapted to improve modern populations' social and physical health.

Social interactions increasingly take place online. Friendships and other offline social ties have been repeatedly associated with human longevity, but online interactions might have different properties.

General background, specific background, knowledge gap

Here, we reference 12 million social media profiles against California Department of Public Health vital records and use longitudinal statistical models to assess whether social media use is associated with longer life. The results show that receiving requests to connect as friends online is associated with reduced mortality but initiating friendships is not.

Here we show

Additionally, online behaviors that indicate face-to-face social activity (like posting photos) are associated with reduced mortality, but online-only behaviors (like sending messages) have a nonlinear relationship, where moderate use is associated with the lowest mortality. These results suggest that online social integration is linked to lower risk for a wide variety of critical health problems.

Results

Although this is an associational study, it may be an important step in understanding how, on a global scale, online social networks might be adapted to improve modern populations' social and physical health.

Implications

Online social media are information resources that can have a transformative power in society. While the Web was envisioned as an equalizing force that allows everyone to access information, the digital divide prevents large amounts of people from being present online. Online social media, in particular, are prone to gender inequality, an important issue given the link between social media use and employment. Understanding gender inequality in social media is a challenging task due to the necessity of data sources that can provide large-scale measurements across multiple countries. Here, we show how the Facebook Gender Divide (FGD), a metric based on aggregated statistics of more than 1.4 billion users in 217 countries, explains various aspects of worldwide gender inequality. Our analysis shows that the FGD encodes gender equality indices in education, health, and economic opportunity. We find gender differences in network externalities that suggest that using social media has an added value for women. Furthermore, we find that low values of the FGD are associated with increases in economic gender equality. Our results suggest that online social networks, while suffering evident gender imbalance, may lower the barriers that women have to access to informational resources and help to narrow the economic gender gap.

Online social media are information resources that can have a transformative power in society. While the Web was envisioned as an equalizing force that allows everyone to access information, the digital divide prevents large amounts of people from being present online. Online social media, in particular, are prone to gender inequality, an important issue given the link between social media use and employment.

General background, specific background

Understanding gender inequality in social media is a challenging task due to the necessity of data sources that can provide large-scale measurements across multiple countries.

Knowledge gap

Here, we show how the Facebook Gender Divide (FGD), a metric based on aggregated statistics of more than 1.4 billion users in 217 countries, explains various aspects of worldwide gender inequality.

Here we show

Our analysis shows that the FGD encodes gender equality indices in education, health, and economic opportunity. We find gender differences in network externalities that suggest that using social media has an added value for women. Furthermore, we find that low values of the FGD are associated with increases in economic gender equality.

Results

Our results suggest that online social networks, while suffering evident gender imbalance, may lower the barriers that women have to access to informational resources and help to narrow the economic gender gap.

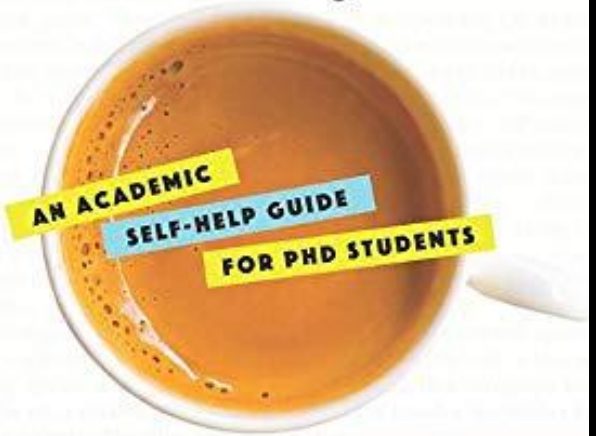
Implications

Success =
Research skills +
Hidden knowledge +
Prestige or luck

Hidden knowledge

Books, online resources

How to Write a Scientific Paper



AN ACADEMIC SELF-HELP GUIDE FOR PHD STUDENTS

Jari Saramäki

Hidden knowledge

Books, online resources

Read a lot of papers

Hidden knowledge

Books, online resources

Read a lot of papers

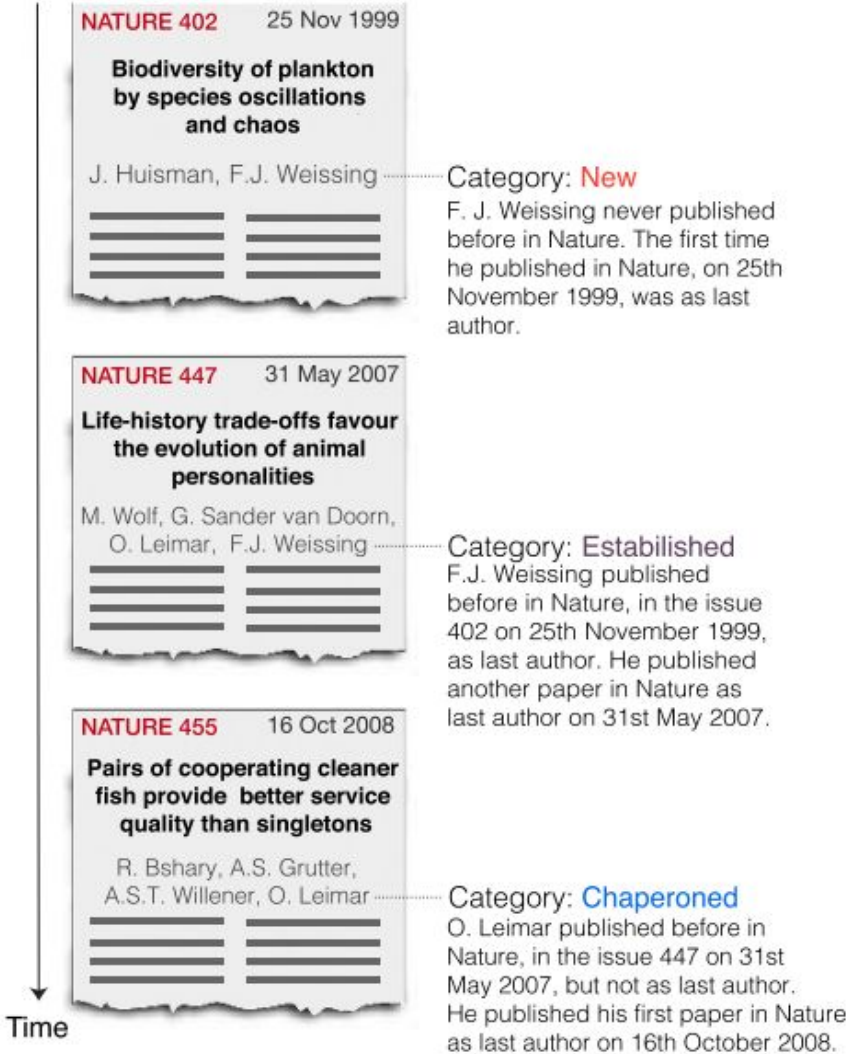
Find a mentor

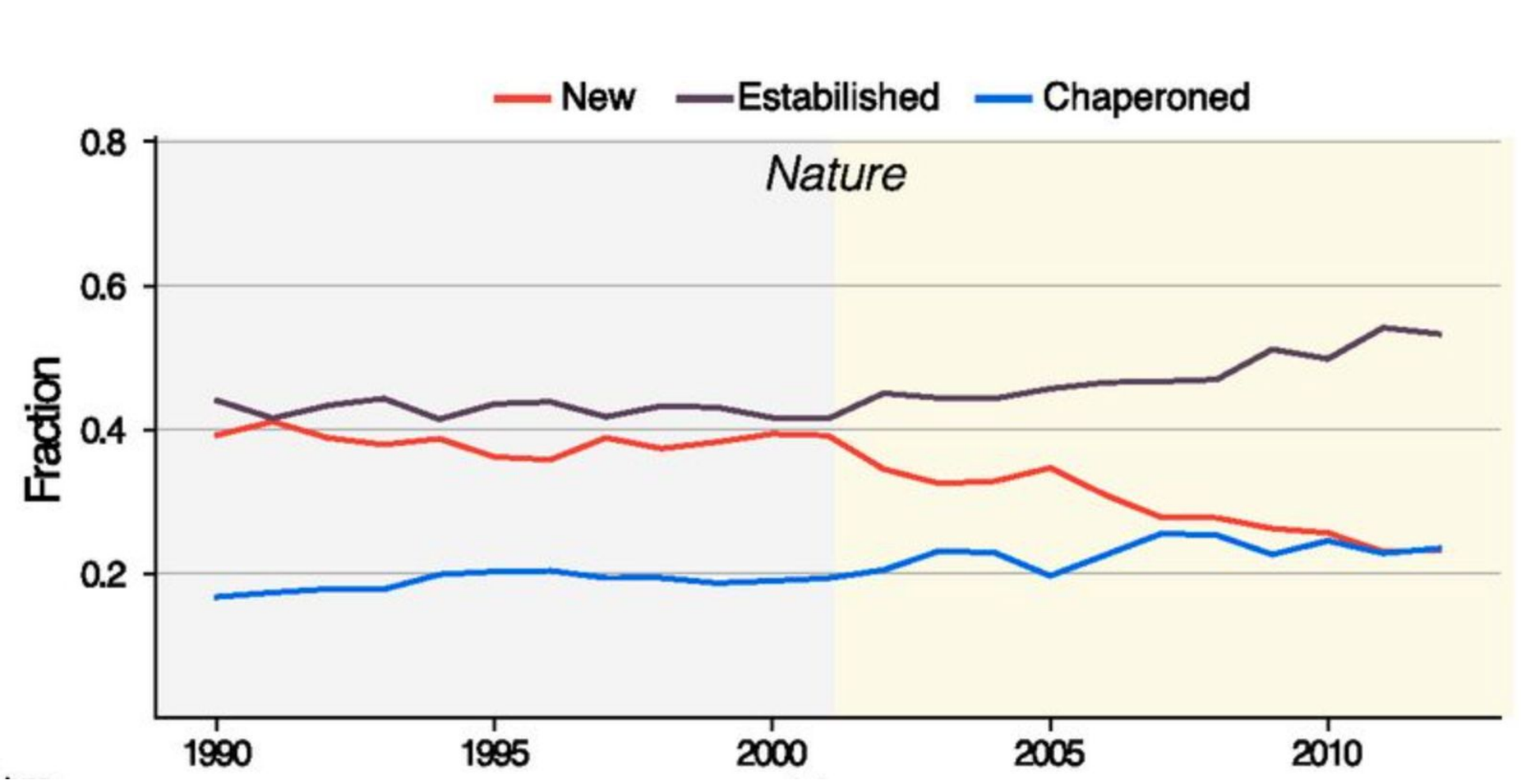
Success =
Research skills +
Hidden knowledge +
Prestige or luck

The chaperone effect in scientific publishing

Sekara V, Deville P, Ahnert SE, Barabási AL, Sinatra R, & Lehmann, S. (2018)

Proceedings of the National Academy of Sciences


NATURE 402
25 Nov 1999
Biodiversity of plankton by species oscillations and chaos
J. Huisman, F.J. Weissing
Category: New
F. J. Weissing never published before in Nature. The first time he published in Nature, on 25th November 1999, was as last author.
NATURE 447
31 May 2007
Life-history trade-offs favour the evolution of animal personalities
M. Wolf, G. Sander van Doorn, O. Leimar, F.J. Weissing
Category: Estabilished
F.J. Weissing published before in Nature, in the issue 402 on 25th November 1999, as last author. He published another paper in Nature as last author on 31st May 2007.
NATURE 455
16 Oct 2008
Pairs of cooperating cleaner fish provide better service quality than singletons
R. Bshary, A.S. Grutter, A.S.T. Willener, O. Leimar
Category: Chaperoned
O. Leimar published before in Nature, in the issue 447 on 31st May 2007, but not as last author. He published his first paper in Nature as last author on 16th October 2008.
Time


New
Established
Chaperoned
Nature
Fraction
0.8
0.6
0.4
0.2
1990
1995
2000
2005
2010

Editors are gatekeepers

Thank you for submitting your manuscript to Science Advances. Because your manuscript was not given a high priority rating during our initial assessment, we have decided not to send your paper for further review. We recommend that you consider a more specialized publication venue for this work.

In this case, while we do not question the validity of your work, I am afraid we are not persuaded that your findings represent a sufficiently striking advance to justify publication in Nature Communications.

This paper has a lucid discussion and a neat stylized model relating to <...>. It is not enough of an advance to be appropriate for a publication at PNAS, but it is a useful contribution.

Disadvantage

Not part of editors’ network

Not from prestigious institution

Not from US-UK

N = 1,000,000
in Russia, evidence from China

40 US undergrads
Universal law of human behaviour

Professional designers

Other paid job

Ivan Smirnov

@ibsmirnov

#computational social science #Fair AI
#education #inequality #social networks
#emotional well-being #gender #ML