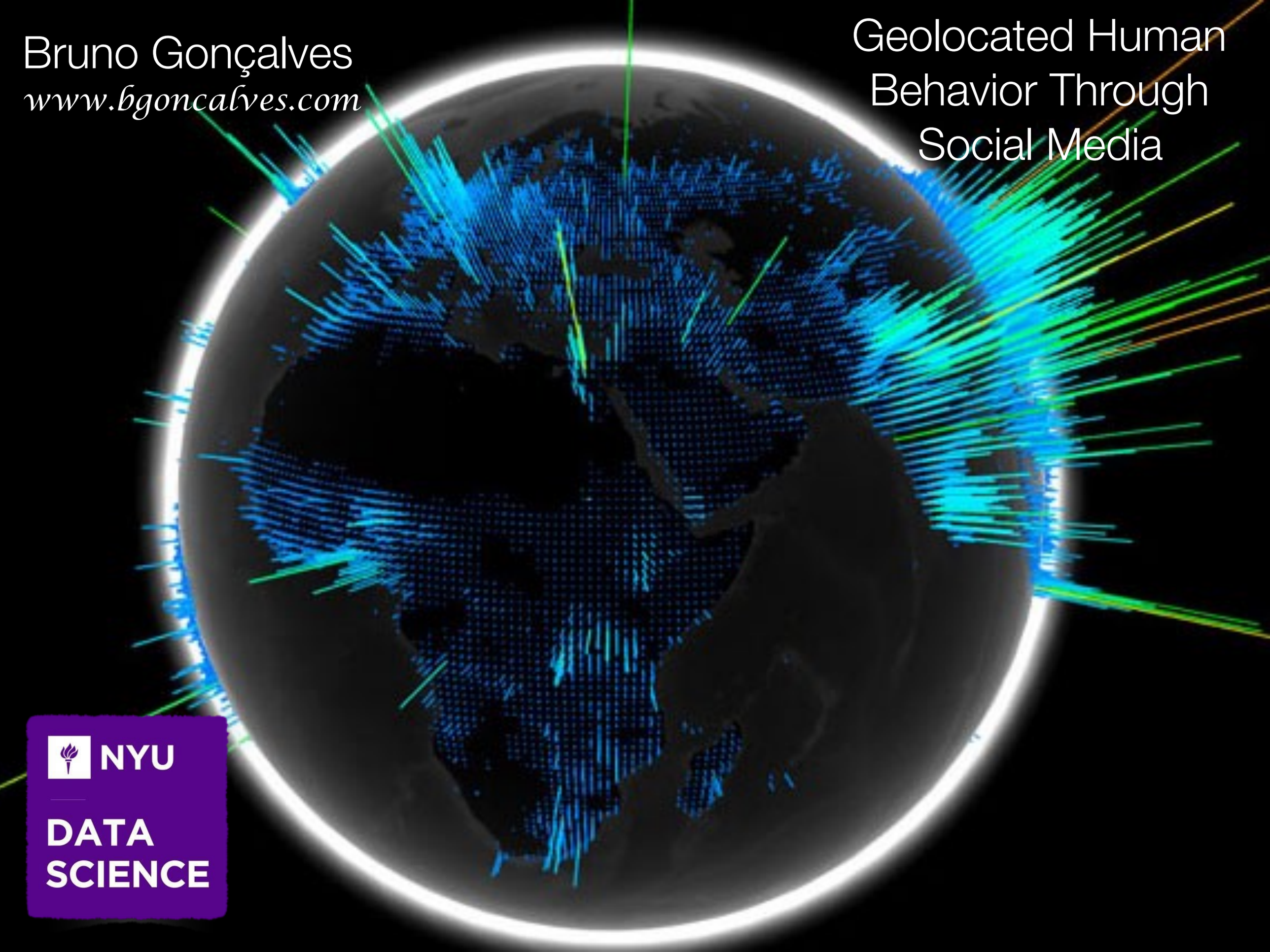
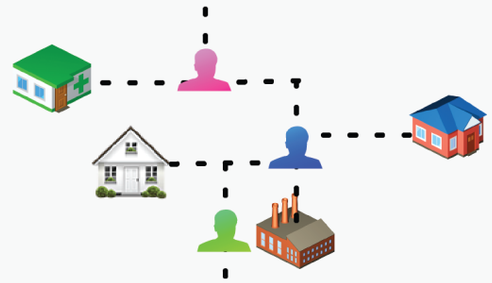
Bruno Gonçalves
www.bgoncalves.com

Geolocated Human Behavior Through Social Media



NYU DATA SCIENCE

Mobility



Airline Flights

Commuting

Realistic Epidemic Spreading

PNAS 106, 21484 (2009)

GLEaMviz.org

Jun 23, 2009

https://www.youtube.com/watch?v=YAf1aXCvvdU

MoBS

GPS-enabled Smartphones

@bgoncalves

www.bgoncalves.com

JAN 2016

GLOBAL DIGITAL SNAPSHOT

A SNAPSHOT OF THE WORLD'S KEY DIGITAL STATISTICAL INDICATORS

TOTAL POPULATION	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS	UNIQUE MOBILE USERS	ACTIVE MOBILE SOCIAL USERS
7.395 BILLION	**3.419 BILLION**	**2.307 BILLION**	**3.790 BILLION**	**1.968 BILLION**
URBANISATION: 54%	PENETRATION: 46%	PENETRATION: 31%	PENETRATION: 51%	PENETRATION: 27%
FIGURE REPRESENTS TOTAL GLOBAL POPULATION, INCLUDING CHILDREN	*FIGURE INCLUDES ACCESS VIA FIXED AND MOBILE CONNECTIONS*	*FIGURE BASED ON ACTIVE USER ACCOUNTS, NOT UNIQUE INDIVIDUALS*	*FIGURE REPRESENTS UNIQUE MOBILE PHONE USERS*	*FIGURE BASED ON ACTIVE USER ACCOUNTS, NOT UNIQUE INDIVIDUALS*

we are social

Sources: Population: UN, US Census Bureau; Internet: ITU, InternetWorldStats, CIA, national government ministries and industry associations; Social & Mobile Social: Facebook, Tencent, VKontakte, LiveInternet.ru, Nikkei, VentureBeat, Niki Aghaei; Mobile: GSMA Intelligence.

All this technology is making us antisocial

@bgoncalves

www.bgoncalves.com

Social Media

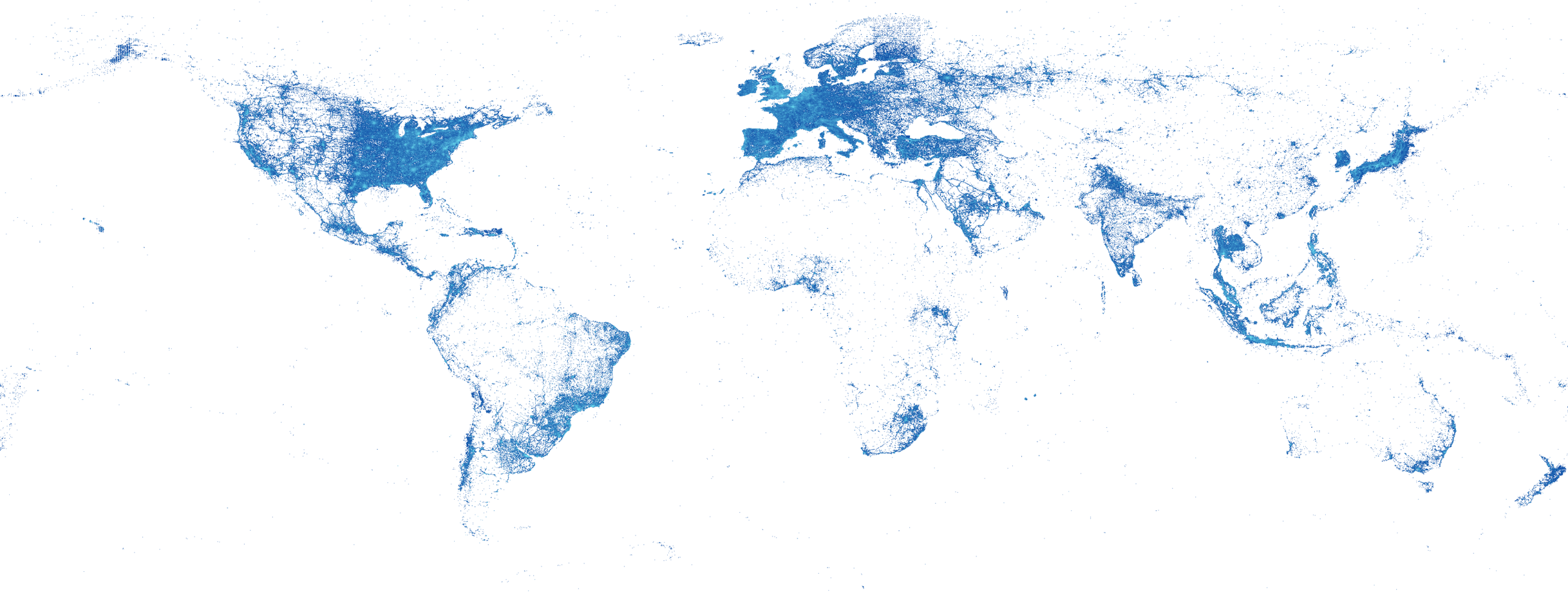
Geolocated Tweets

Geolocated Tweets

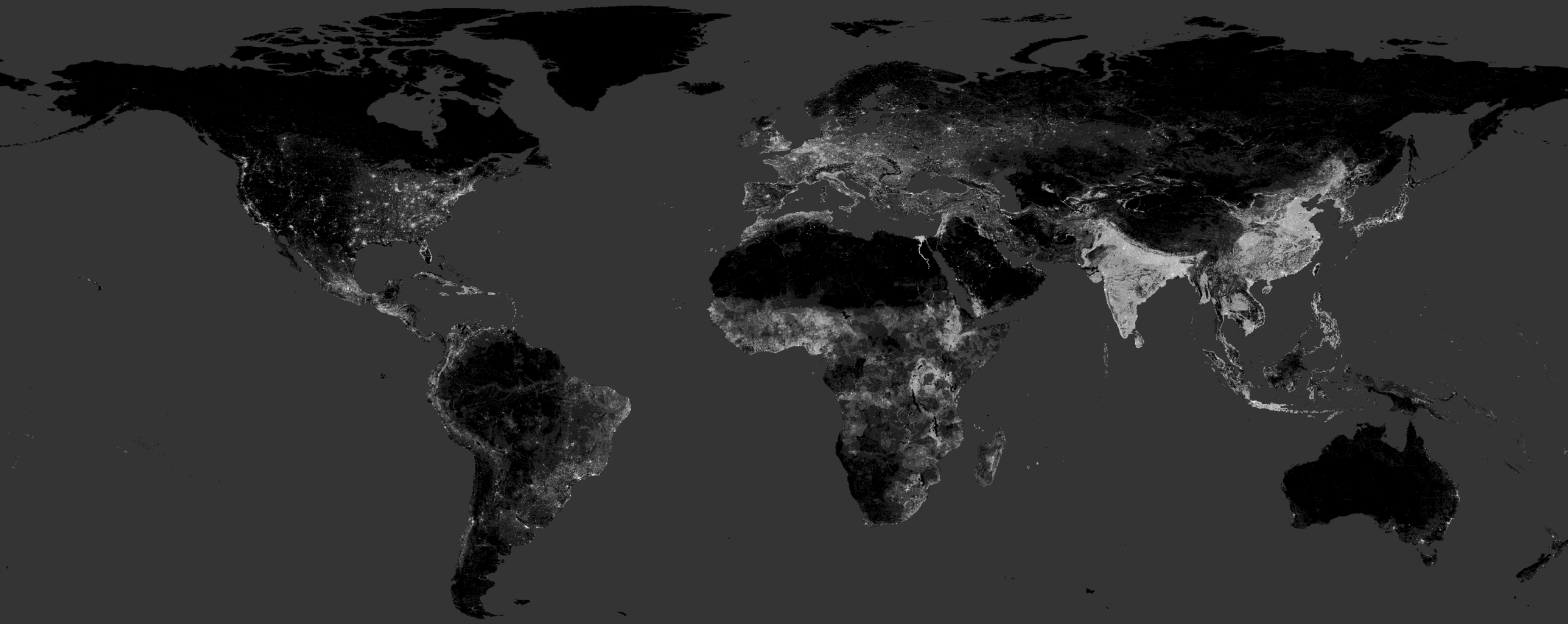
PLoS One 8, E61981 (2013)


Bias
towards
appears
may
exceeds
exhibit
factors
members
area
unbiased
incorrect
whereas
estimator's
type
sampling
ideology
Main
interferes
includes
extreme
interpret
correspondence
systemic
influenced
greater
unconscious
difference
person
generally
diagnostic
people
context
parameters
expectation
influences
estimator
individuals
Labeling
parameter
less
take
regression
refers
preference
biased
term
tend
estimates
new
also
value
null
need
used
way
perspective
prior
types
tendency
information
part
data
beliefs
behavior
independent
misuse
Data-snooping
population
unprejudiced
statistical
significance
hypothesis
awareness
objective
particular
confirmation
measurements
specification
group
describe
brain
Systematic
One
impartial
omits
overestimate
conflicts
ignoring
edit
others
level
techniques
likely
sensitivity
affect
one-sided
groups
patients
estimated
analysis
deeply
prejudice
example
choosing
flexible
Spectrum
denial
attribution
Selection
related
equal
judged
words
implies

GPS Coordinates



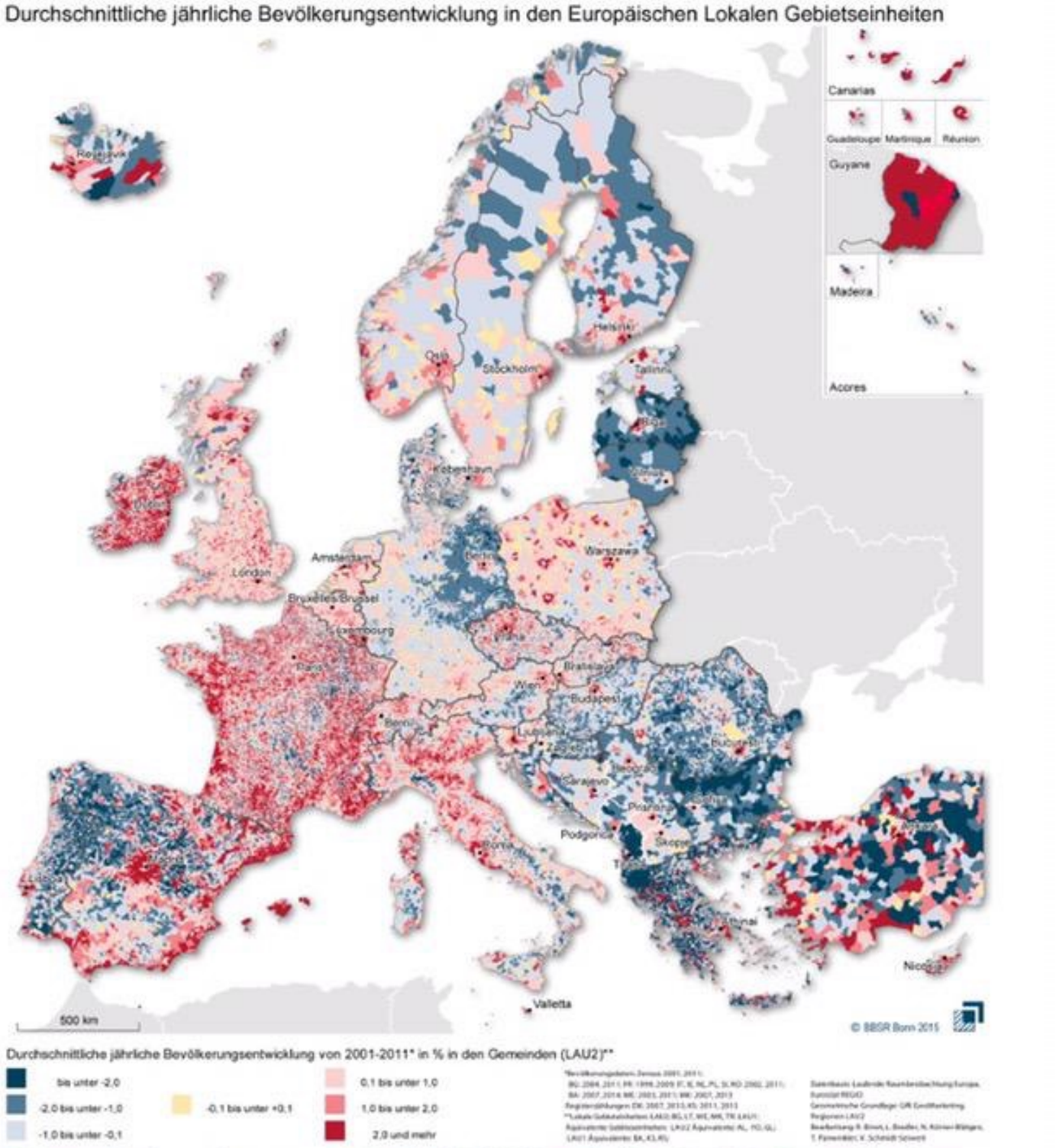
World Population



@bgoncalves

www.bgoncalves.com

Biases



Smartphone Ownership Highest Among Young Adults, Those With High Income/Education Levels

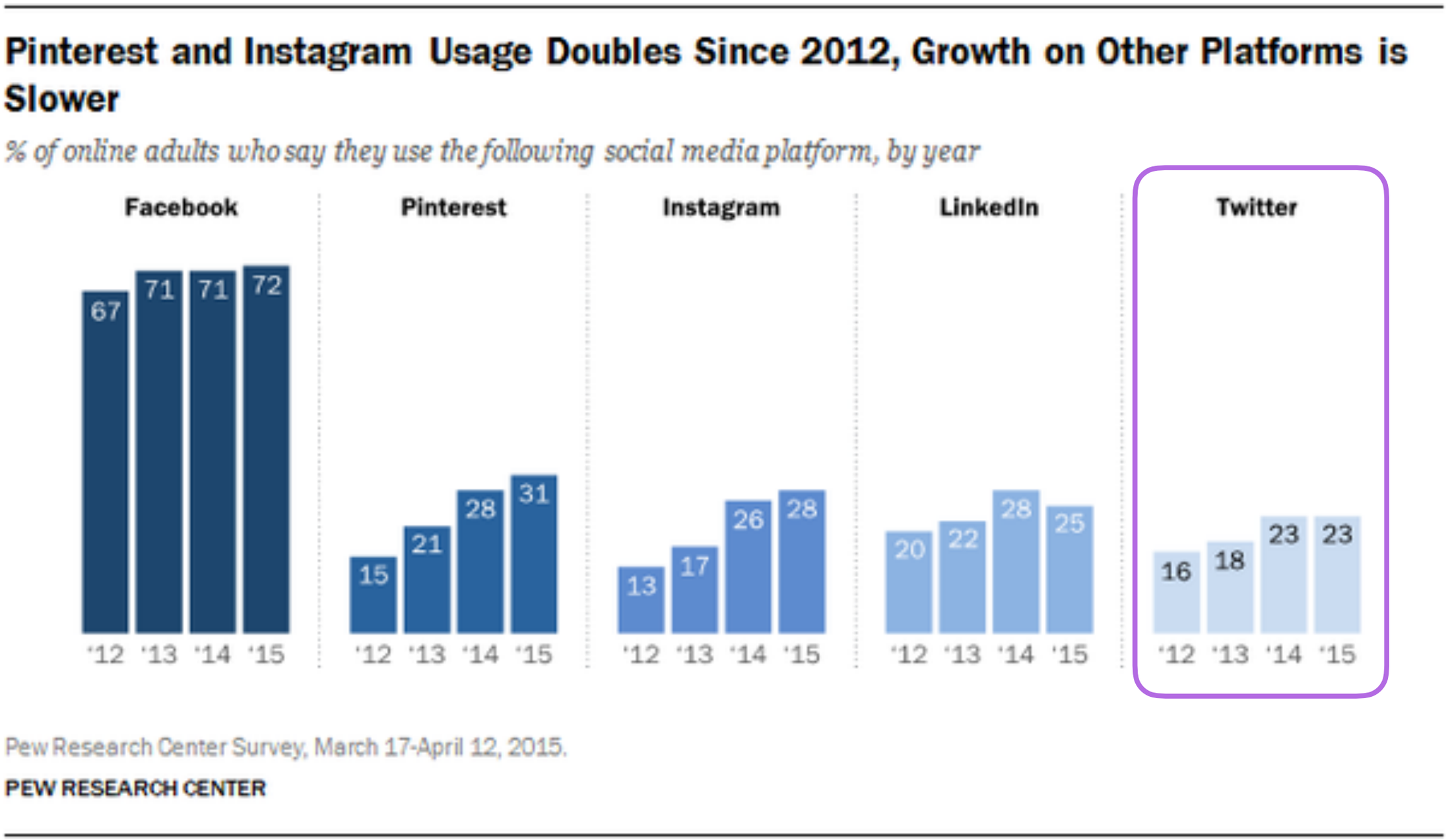
% of U.S. adults in each group who own a smartphone

All adults	**64%**
Male	66
Female	63
18-29	85
30-49	79
50-64	54
65+	27
White, non-Hispanic	61
Black, non-Hispanic	70
Hispanic	71
HS grad or less	52
Some college	69
College+	78
Less than $30,000/yr	50
$30,000-$49,999	71
$50,000-$74,999	72
$75,000 or more	84
Urban	68
Suburban	66
Rural	52

Combined analysis of Pew Research Center surveys conducted Dec. 4-7 and Dec. 18-21, 2014.

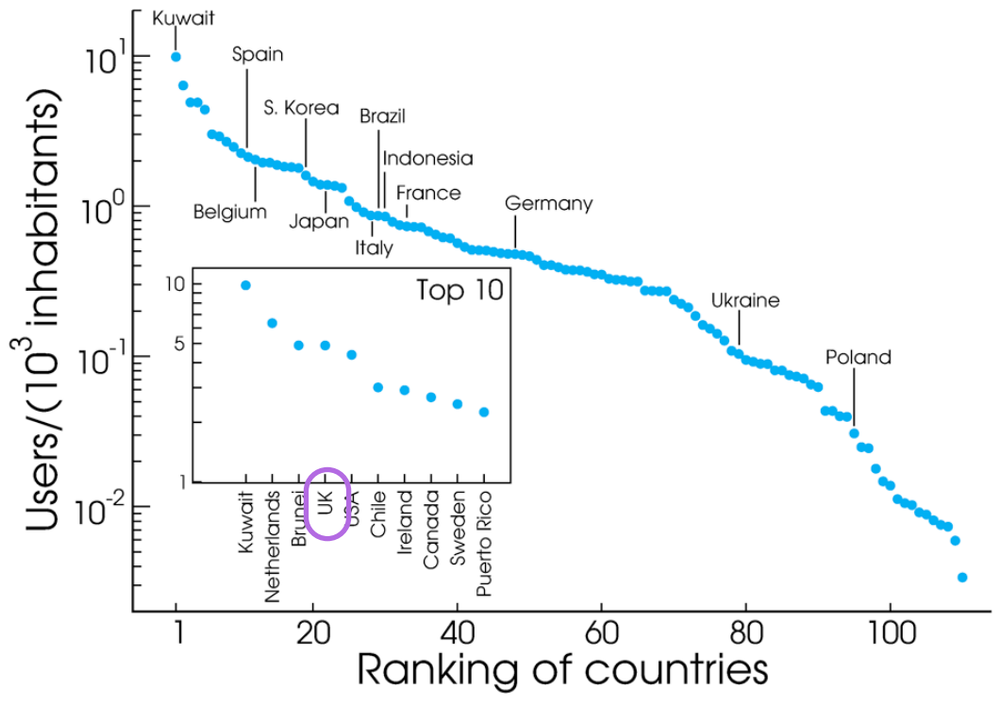
PEW RESEARCH CENTER

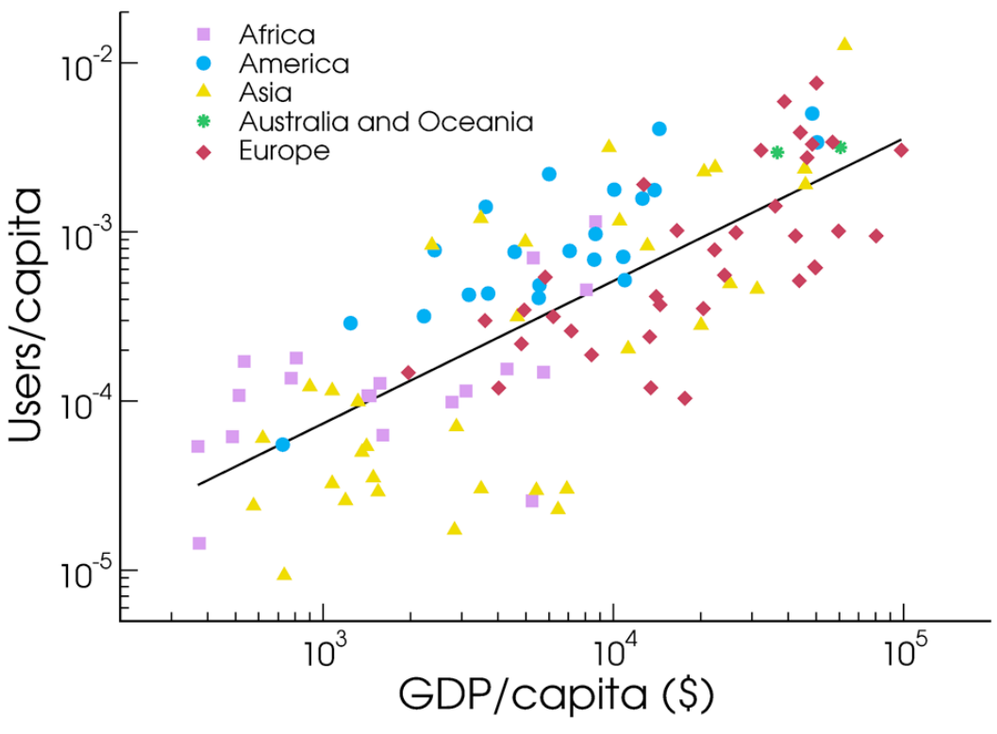
Biases


Pinterest and Instagram Usage Doubles Since 2012, Growth on Other Platforms is Slower
% of online adults who say they use the following social media platform, by year
Facebook
67
71
71
72
'12 '13 '14 '15
Pinterest
15
21
28
31
'12 '13 '14 '15
Instagram
13
17
26
28
'12 '13 '14 '15
LinkedIn
20
22
28
25
'12 '13 '14 '15
Twitter
16
18
23
23
'12 '13 '14 '15
Pew Research Center Survey, March 17-April 12, 2015.
PEW RESEARCH CENTER

Market Penetration

PLoS One 8, E61981 (2013)





www.bgoncalves.com

Age Distribution

PLoS One 10, e0115545 (2015)

Twitter users

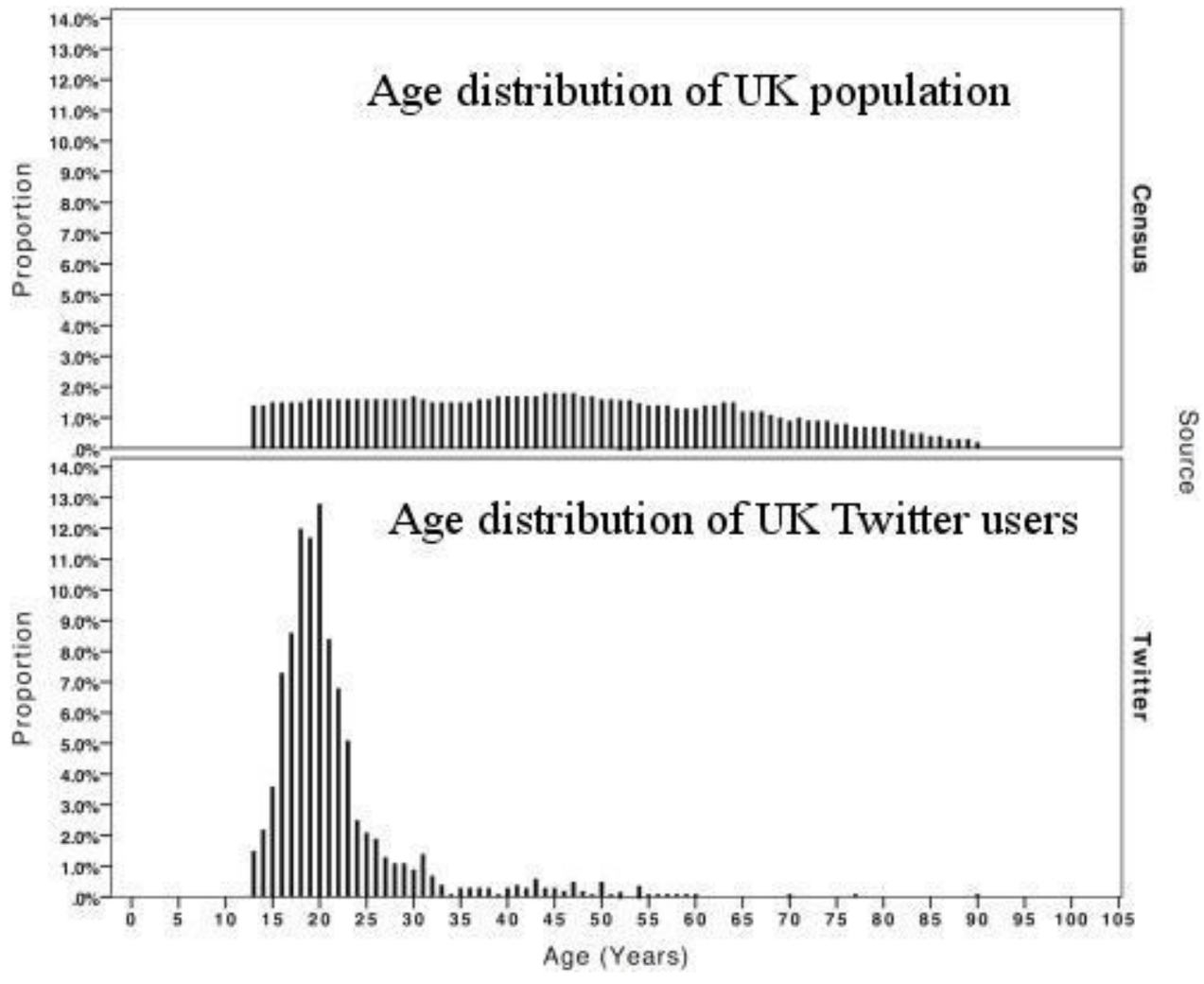
Among online adults, the % who use Twitter

	2013	2014
All internet users	*18%*	*23%**
Men	17	24*
Women	18	21
White, Non-Hispanic	16	21 *
Black, Non-Hispanic	29	27
Hispanic	16	25
18-29	31	37
30-49	19	25
50-64	9	12
65+	5	10*
High school grad or less	17	16
Some college	18	24
College+ (n= 685)	18	30*
Less than $30,000/yr	17	20
$30,000-$49,999	18	21
$50,000-$74,999	15	27*
$75,000+	19	27*
Urban	18	25*
Suburban	19	23
Rural	11	17

Source: Pew Research Center's Internet Project September Combined Omnibus Survey, September 11-14 & September 18-21, 2014. N=1,597 internet users ages 18+. The margin of error for all internet users is +/- 2.9 percentage points. 2013 data from Pew Internet August Tracking Survey, August 07 – September 16, 2013, n= 1,445 internet users ages 18+.

Note: Percentages marked with an asterisk (*) represent a significant change from 2013. Results are significant at the 95% confidence level using an independent z-test.

PEW RESEARCH CENTER



Age Distribution

Twitter users

Among online adults, the % who use Twitter

	2013	2014
All internet users	*18%*	*23%**
Men	17	24*
Women	18	21
White, Non-Hispanic	16	21 *
Black, Non-Hispanic	29	27
Hispanic	16	25
18-29	31	37
30-49	19	25
50-64	9	12
65+	5	10*
High school grad or less	17	16
Some college	18	24
College+ (n= 685)	18	30*
Less than $30,000/yr	17	20
$30,000-$49,999	18	21
$50,000-$74,999	15	27*
$75,000+	19	27*
Urban	18	25*
Suburban	19	23
Rural	11	17

Source: Pew Research Center's Internet Project September Combined Omnibus Survey, September 11-14 & September 18-21, 2014. N=1,597 internet users ages 18+. The margin of error for all internet users is +/- 2.9 percentage points. 2013 data from Pew Internet August Tracking Survey, August 07 – September 16, 2013, n= 1,445 internet users ages 18+.

Note: Percentages marked with an asterisk (*) represent a significant change from 2013. Results are significant at the 95% confidence level using an independent z-test.

PEW RESEARCH CENTER

Facebook Demographics

Among internet users, the % who use Facebook

	Internet users
Total	*72%*
Men	66
Women	77
White, Non-Hispanic	70
Black, Non-Hispanic (n=85)	67
Hispanic	75
18-29	82
30-49	79
50-64	64
65+	48
High school grad or less	71
Some college	72
College+	72
Less than $30,000/yr	73
$30,000-$49,999	72
$50,000-$74,999	66
$75,000+	78
Urban	74
Suburban	72
Rural	67

Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

Instagram Demographics

Among internet users, the % who use Instagram

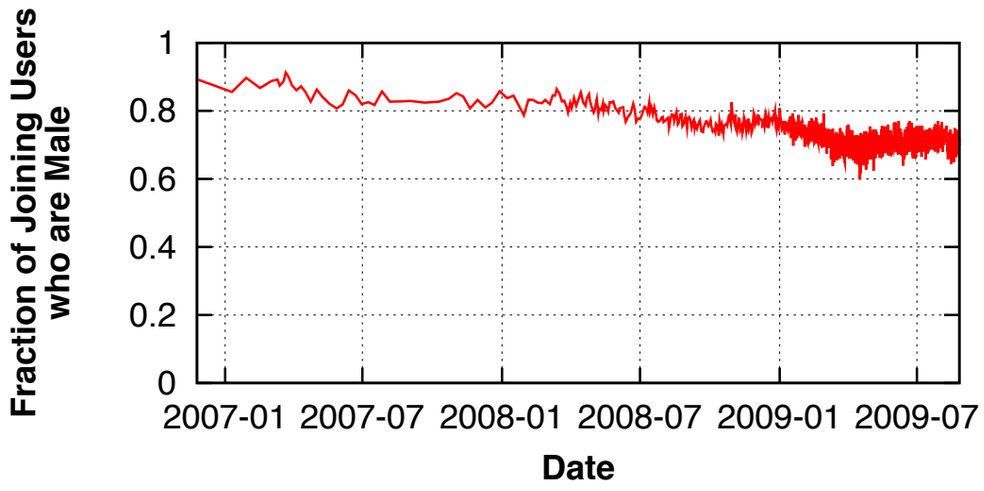
	Internet user
Total	*28%*
Men	24
Women	31
White, Non-Hispanic	21
Black, Non-Hispanic (n=85)	47
Hispanic	38
18-29	55
30-49	28
50-64	11
65+	4
High school grad or less	25
Some college	32
College+	26
Less than $30,000/yr	26
$30,000-$49,999	27
$50,000-$74,999	30
$75,000+	26
Urban	32
Suburban	28
Rural	18

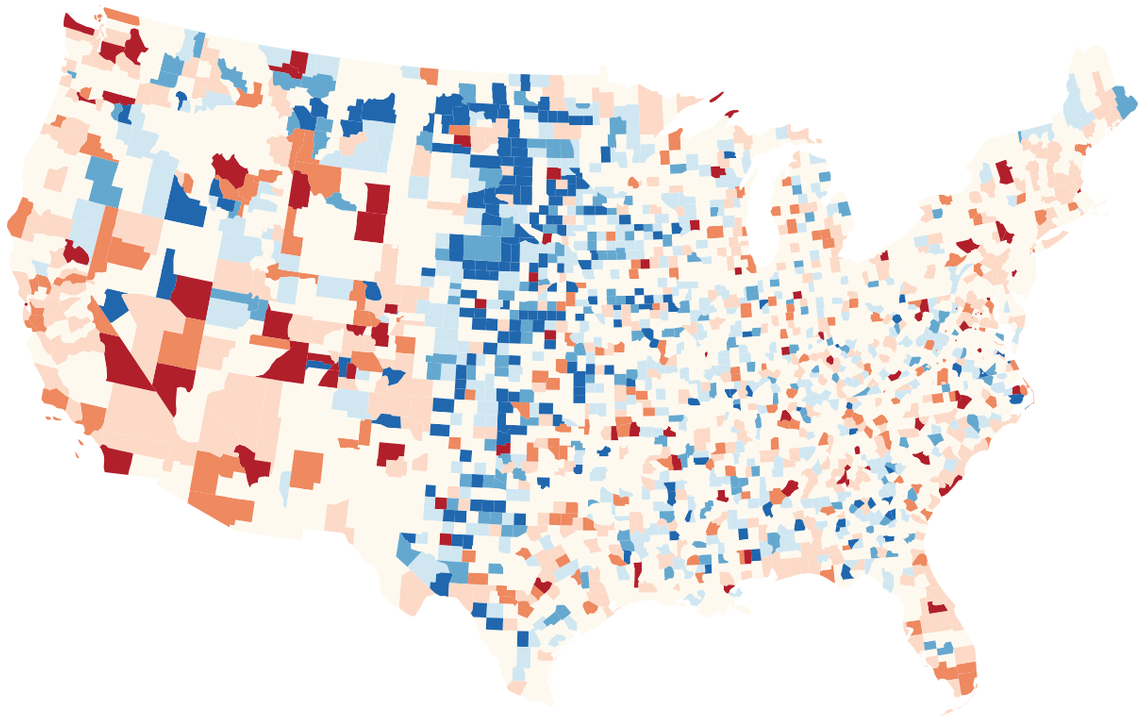
Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

Demographics

ICWSM'11, 375 (2011)

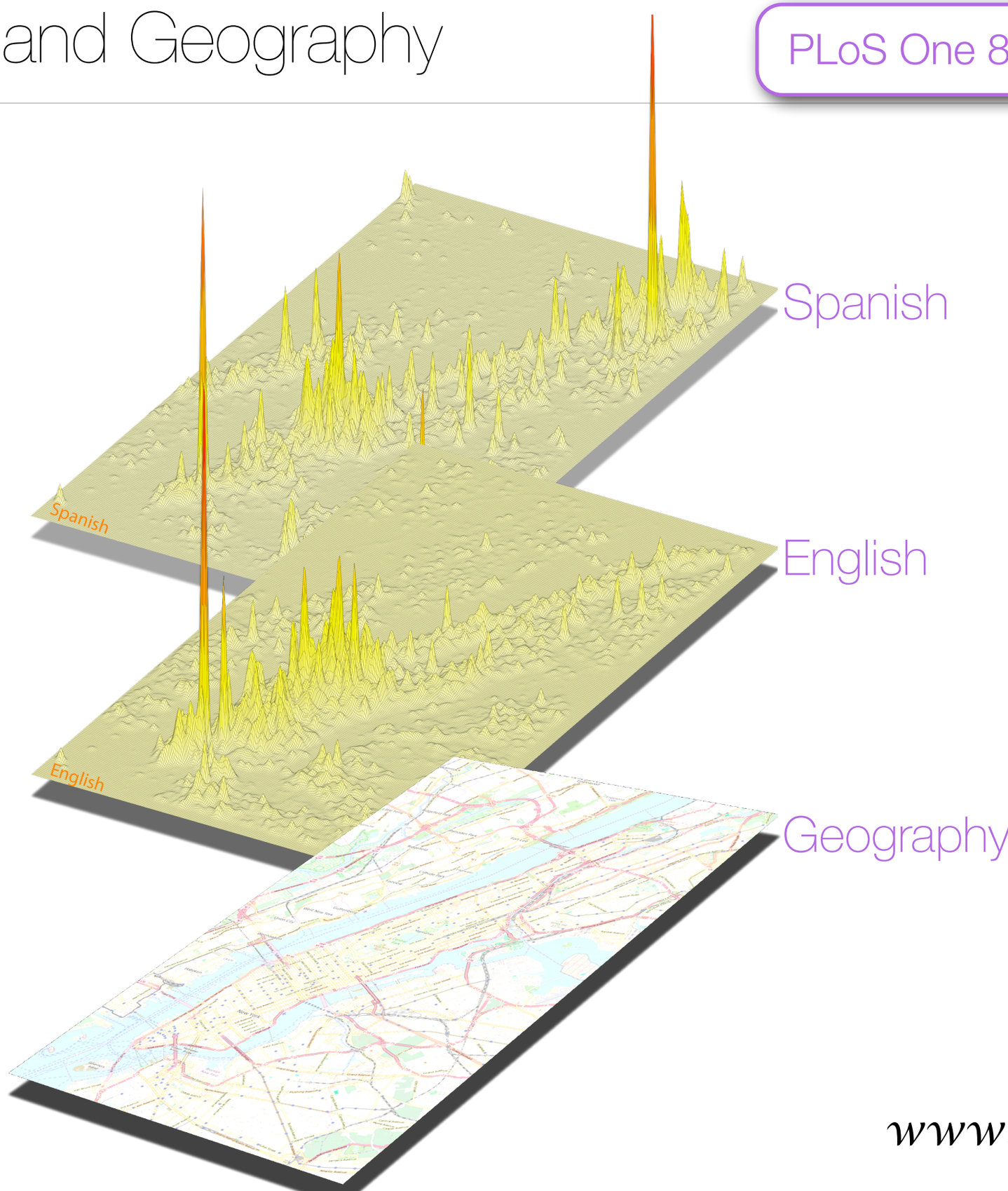




(a) Caucasian (non-hispanic) (b) African-American (c) Asian or Pacific Islander (d) Hispanic

Language and Geography

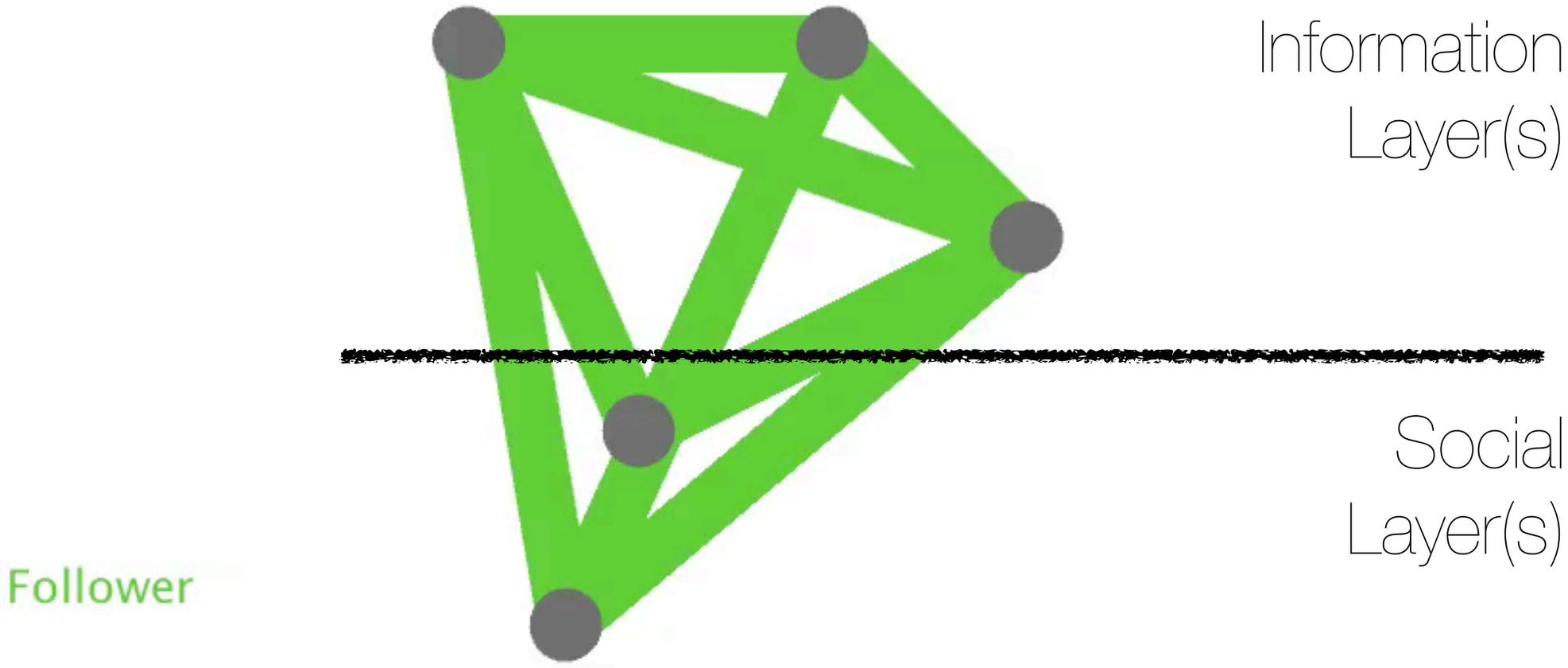
PLoS One 8, E61981 (2013)



@bgoncalves

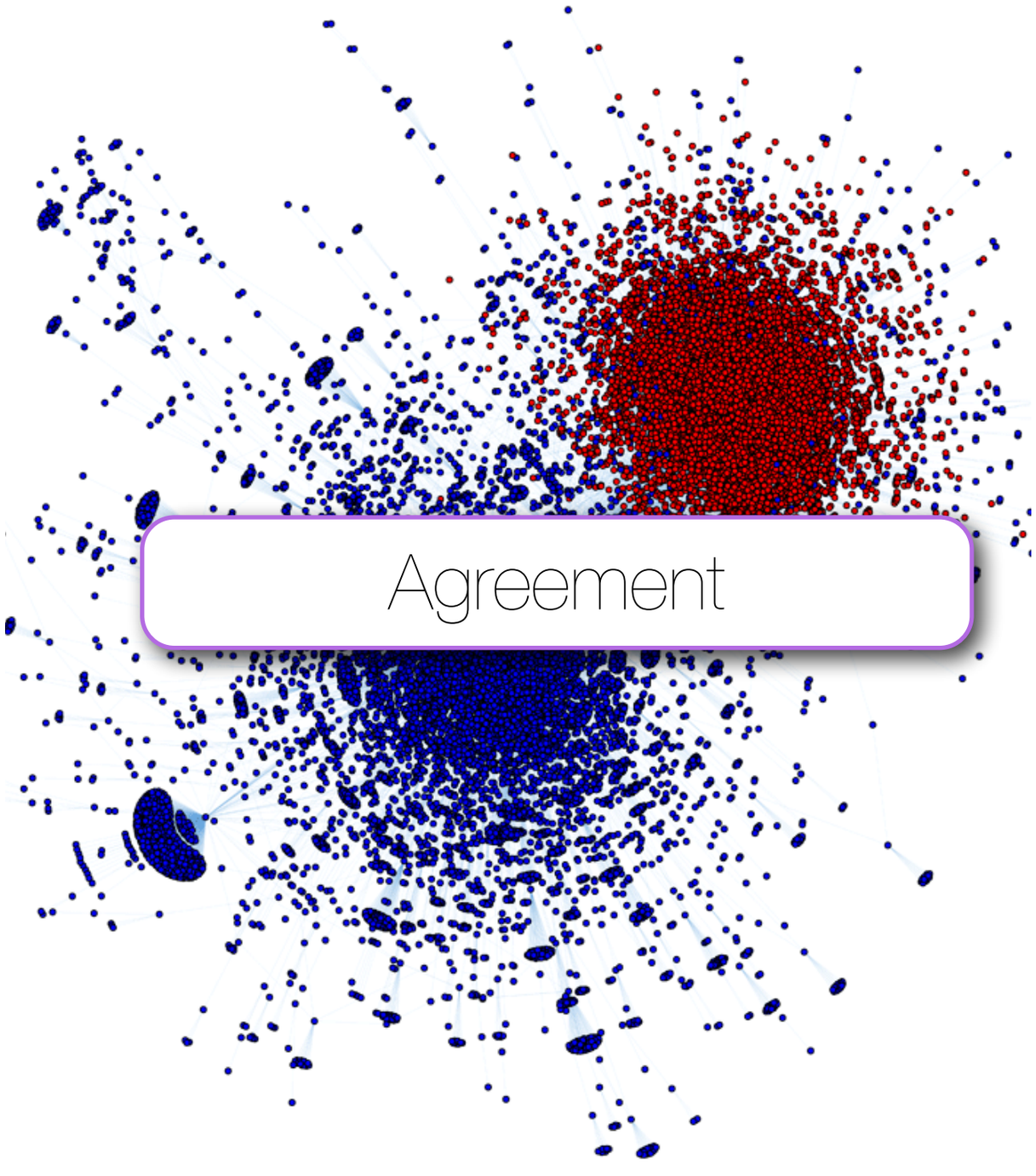
www.bgoncalves.com

Multilayer Network



Link Function

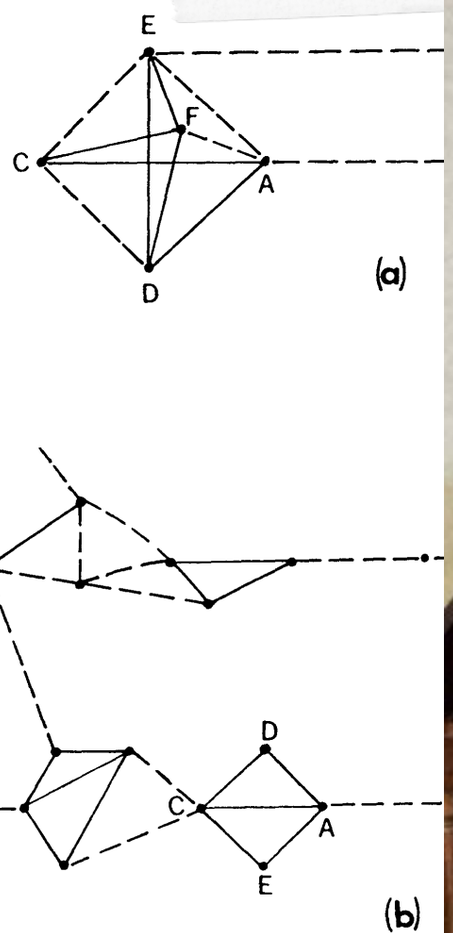
ICWSM'11, 89 (2011)



The Strength of Weak Ties (1973)

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University



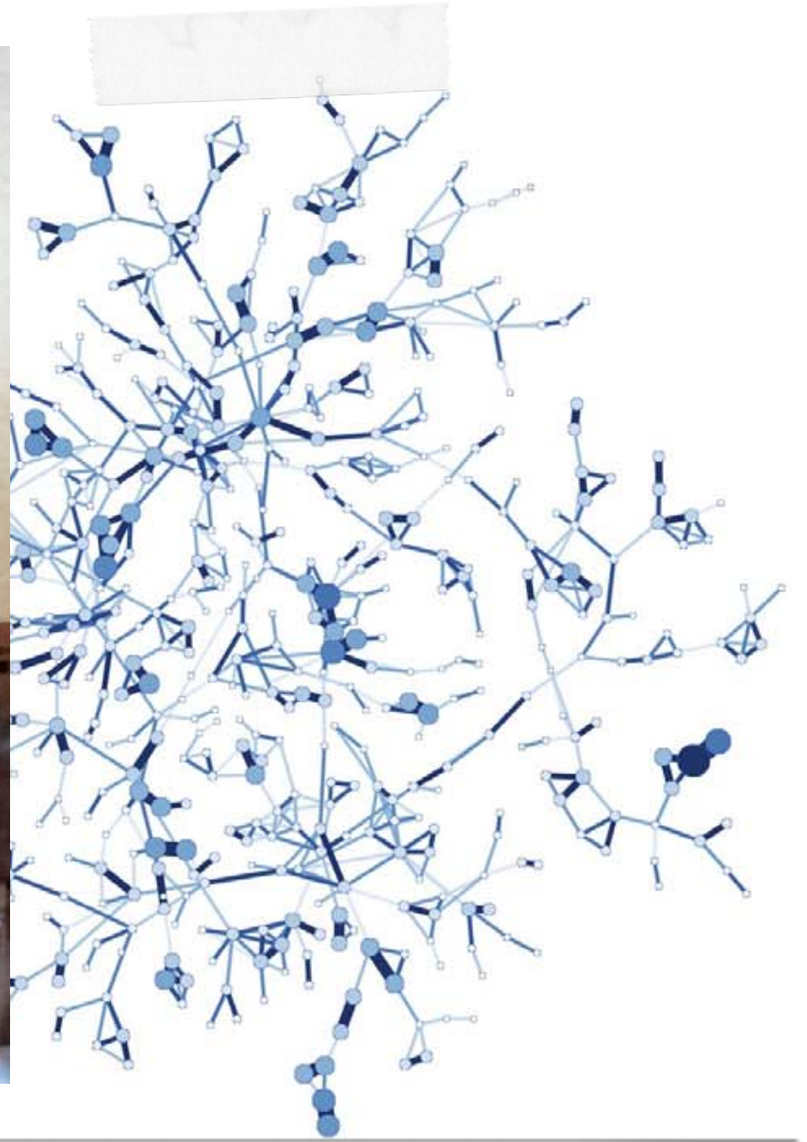
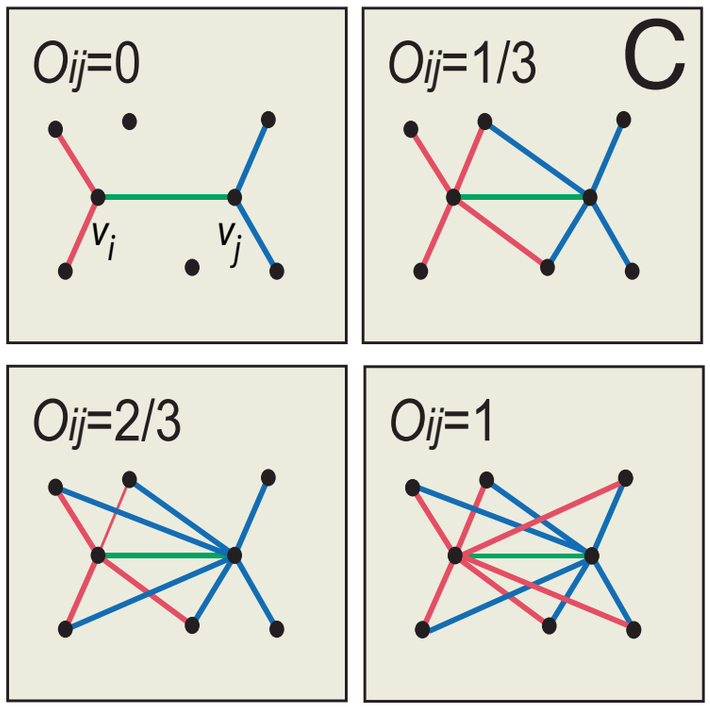
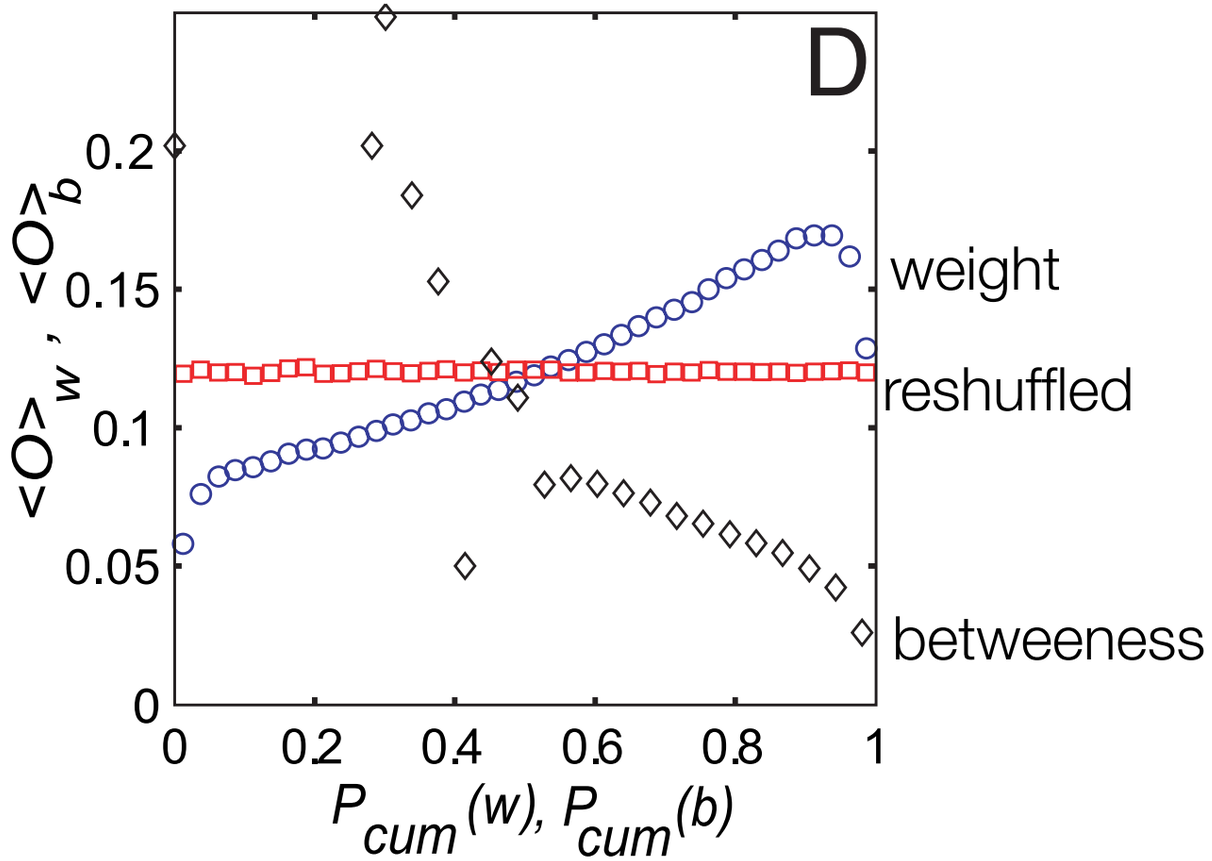


FIG. 2.—Local bridges. *a*, Degree 3; *b*, Degree 13. ——— strong tie, — — — weak tie.

Neighborhood Overlap

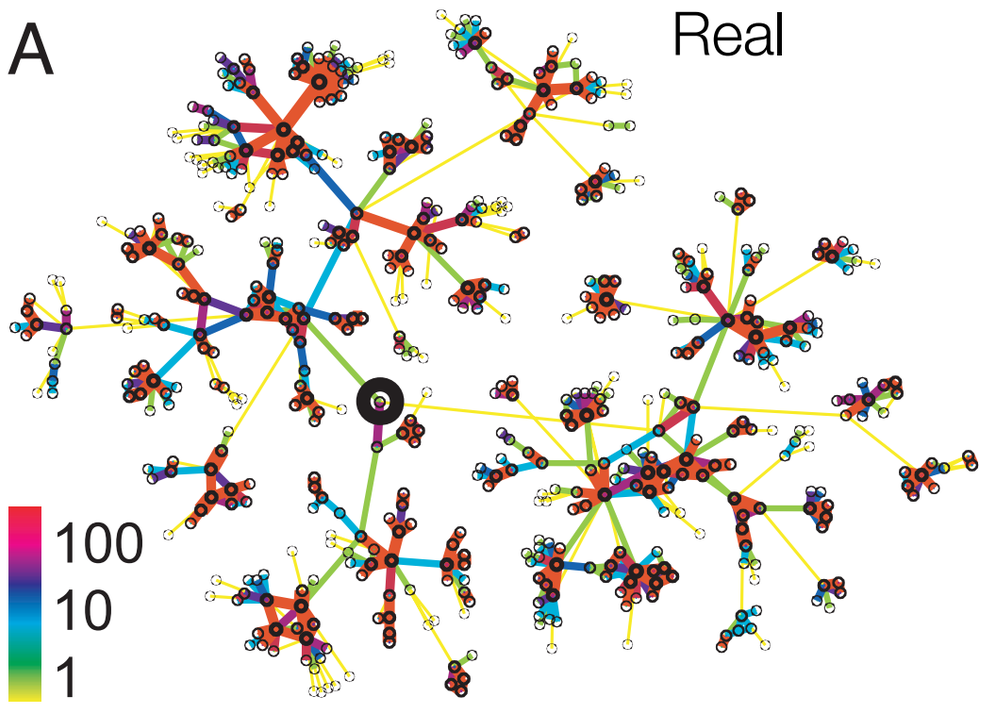
PNAS 104, 7333 (2007)

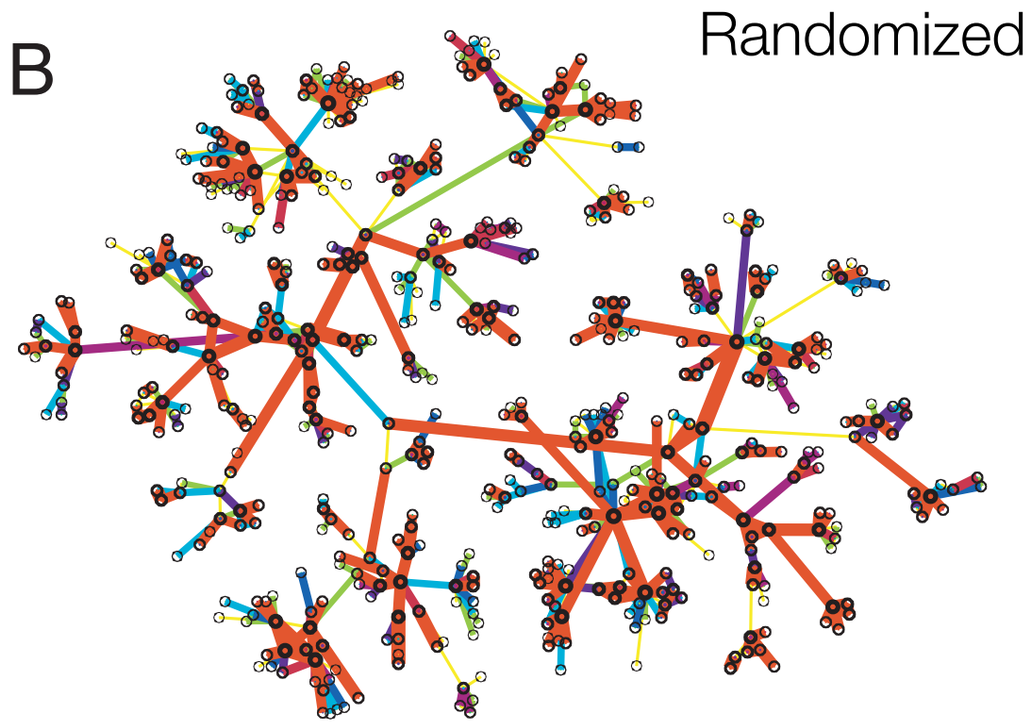


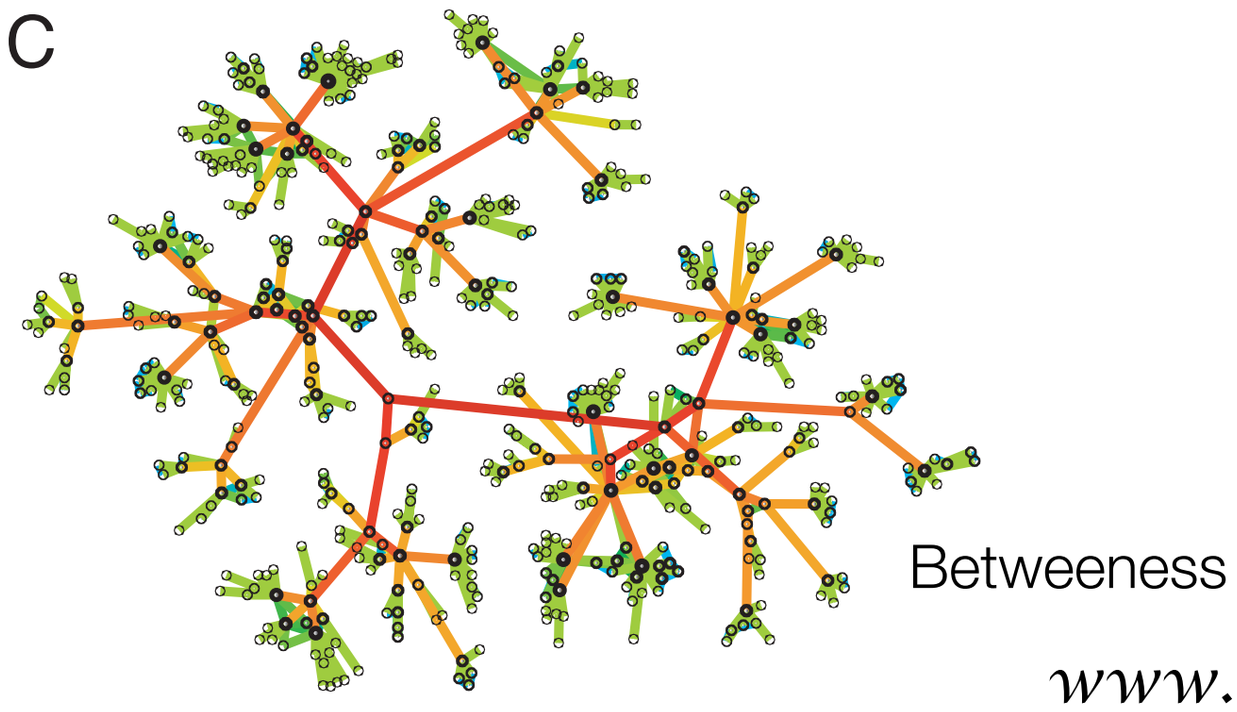


Strong Ties have higher overlaps

PNAS 104, 7333 (2007)

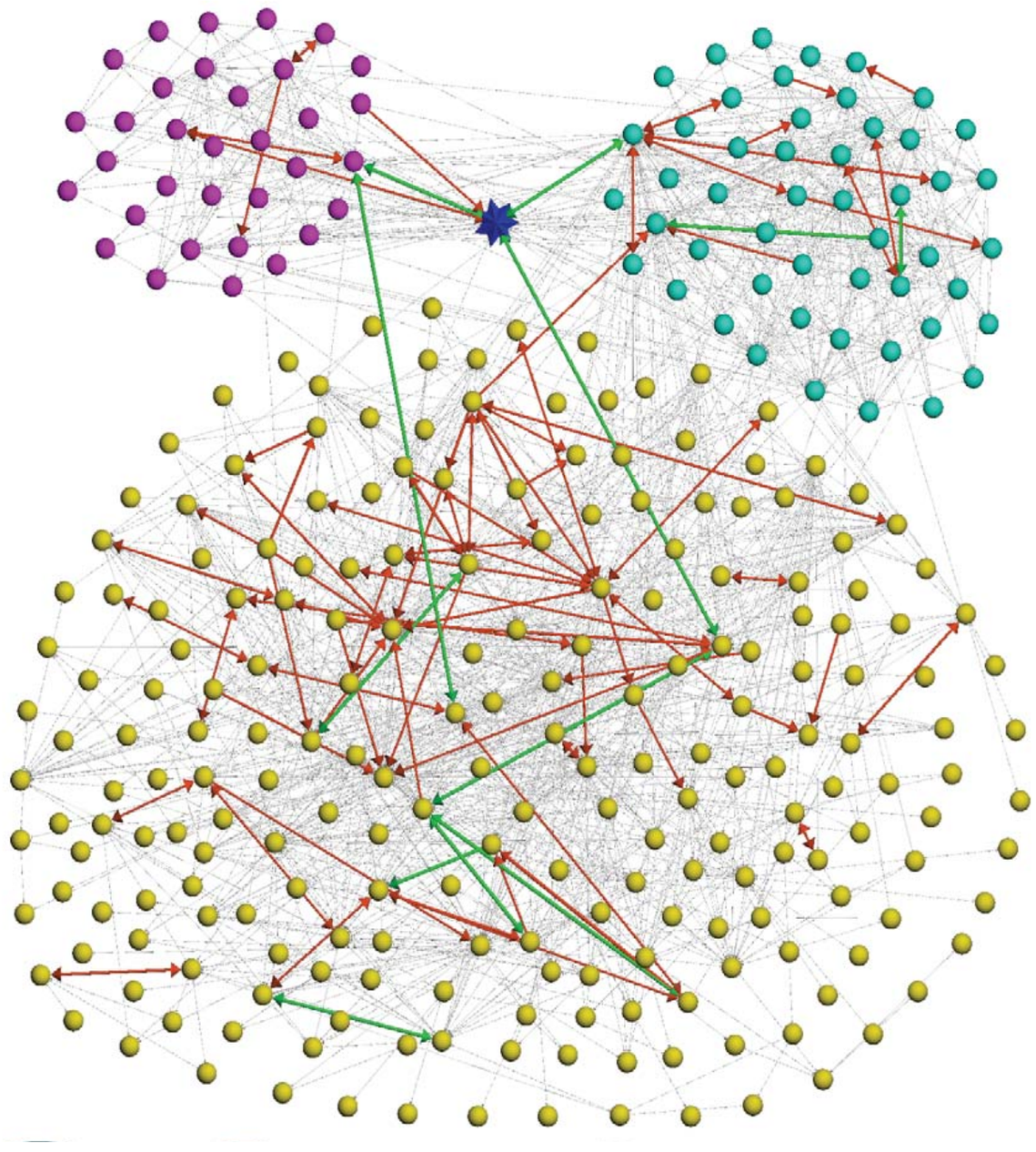






Network Structure

PLoS One 7, e29358 (2012)



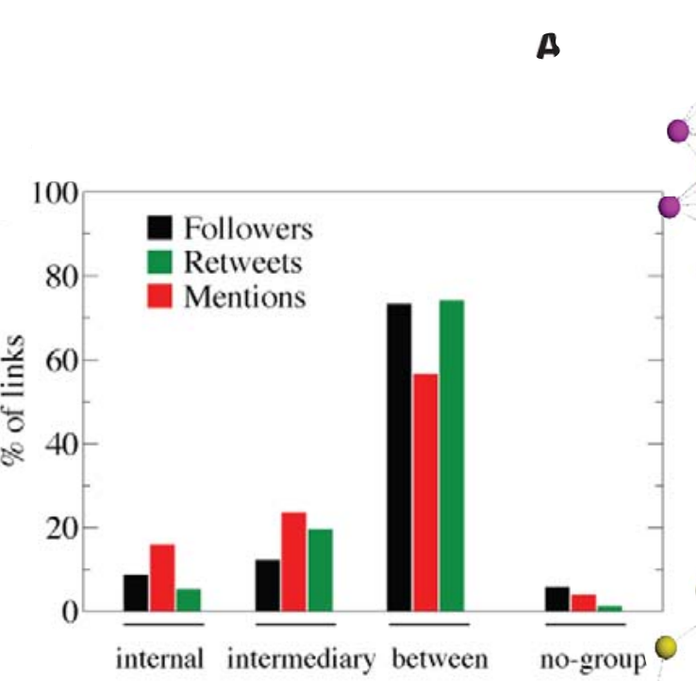
Structural Holes and Good Ideas[1]

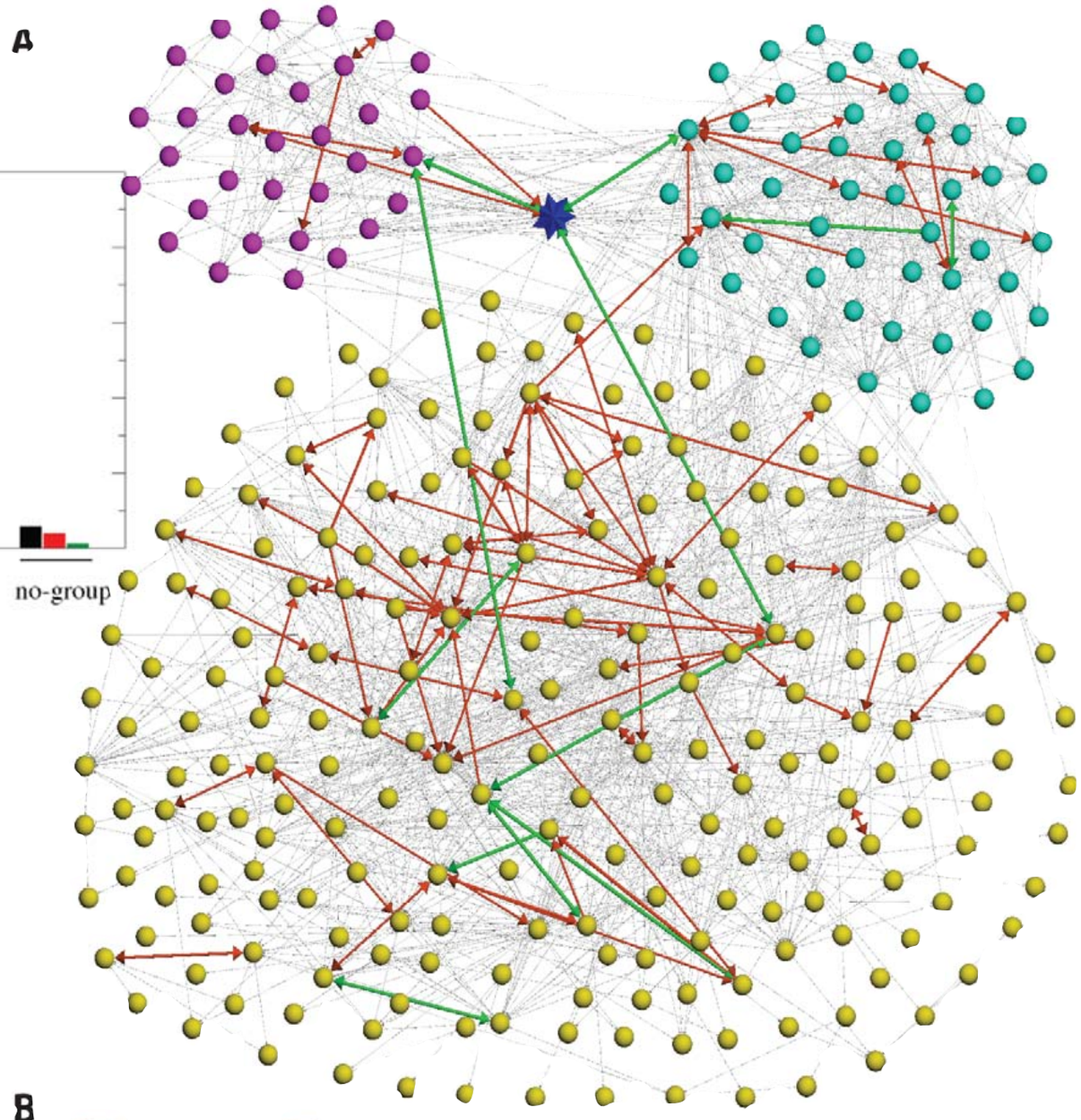
Ronald S. Burt
University of Chicago

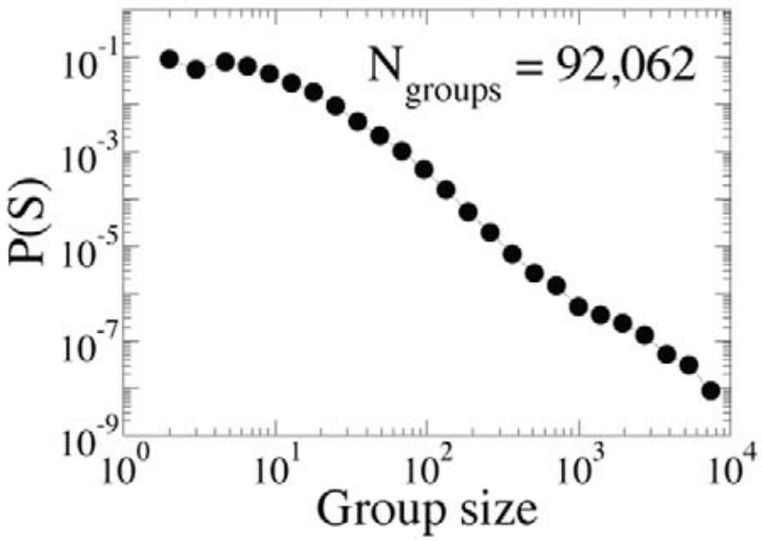
"People whose networks bridge the structural holes between groups have an advantage in detecting and developing rewarding opportunities. Information arbitrage is their advantage. They are able to see early, see more broadly, and translate information across groups."

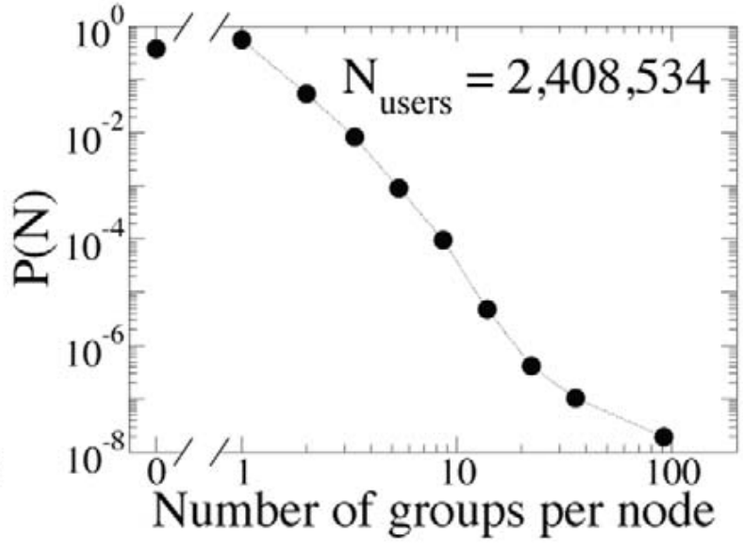
Network Structure

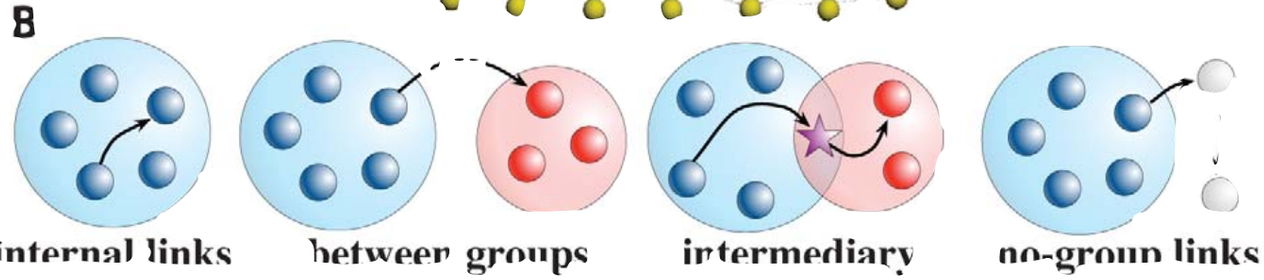
PLoS One 7, e29358 (2012)





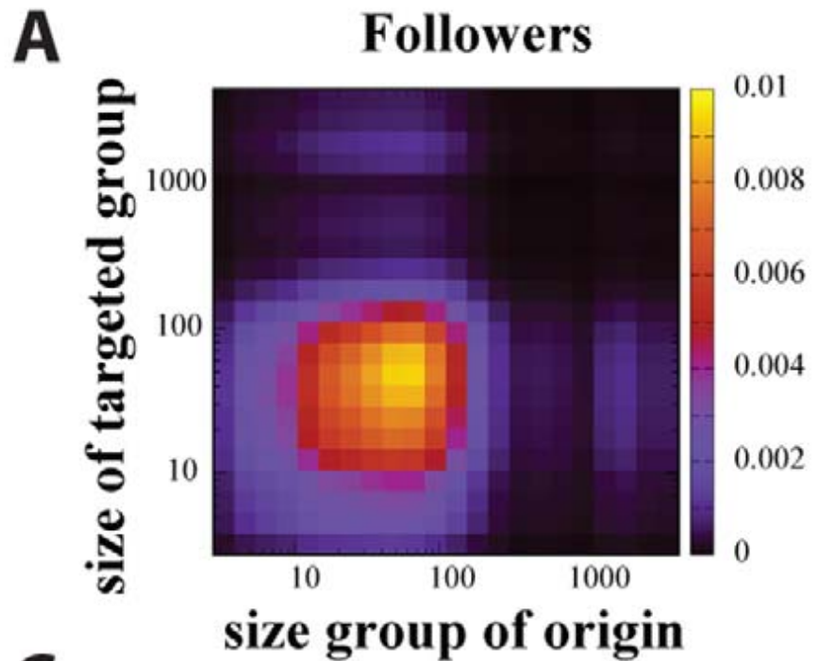


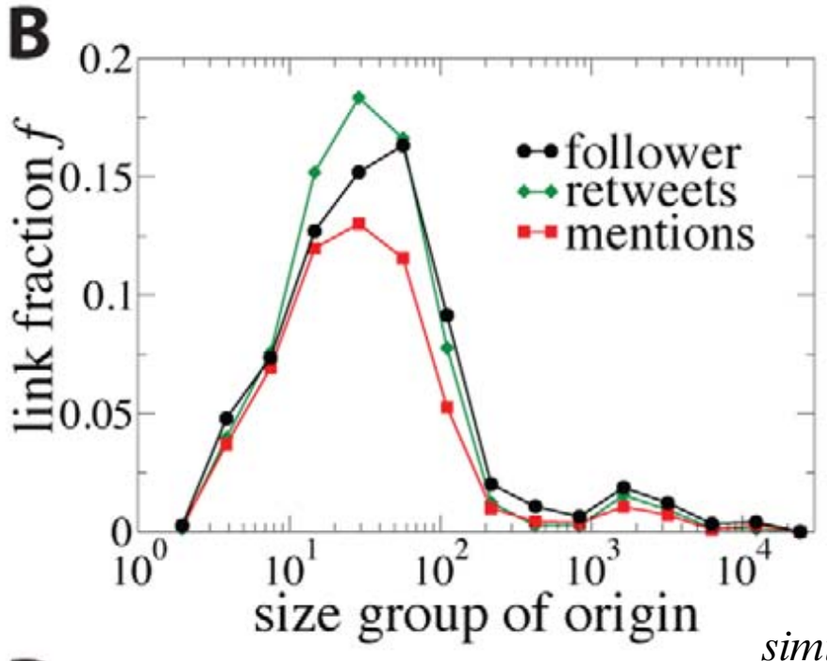




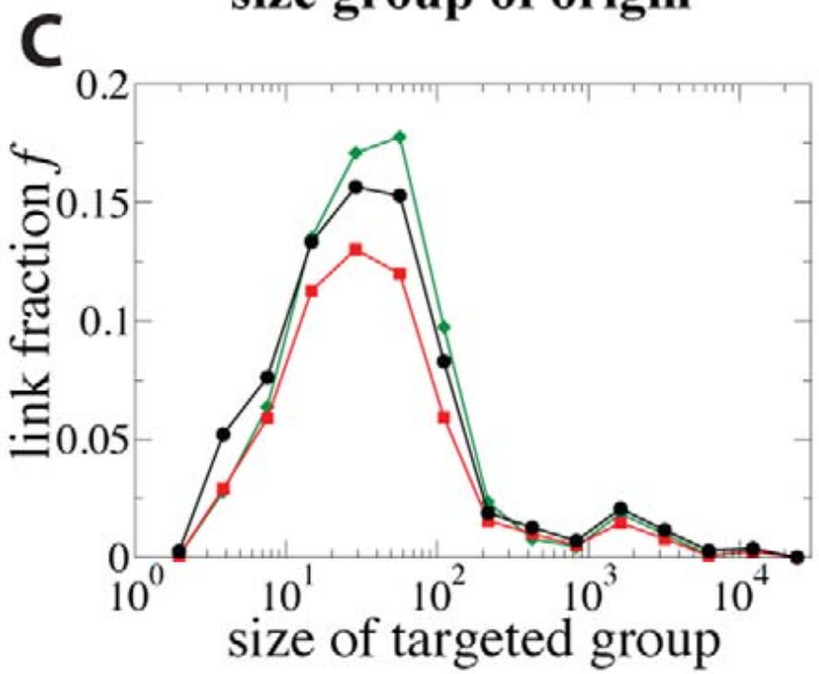
Groups

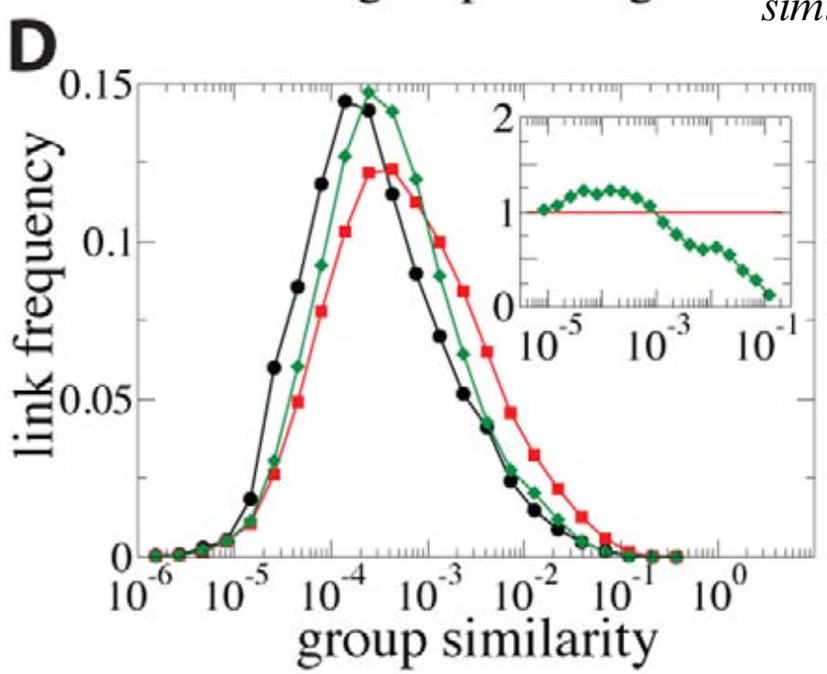
PLoS One 7, e29358 (2012)





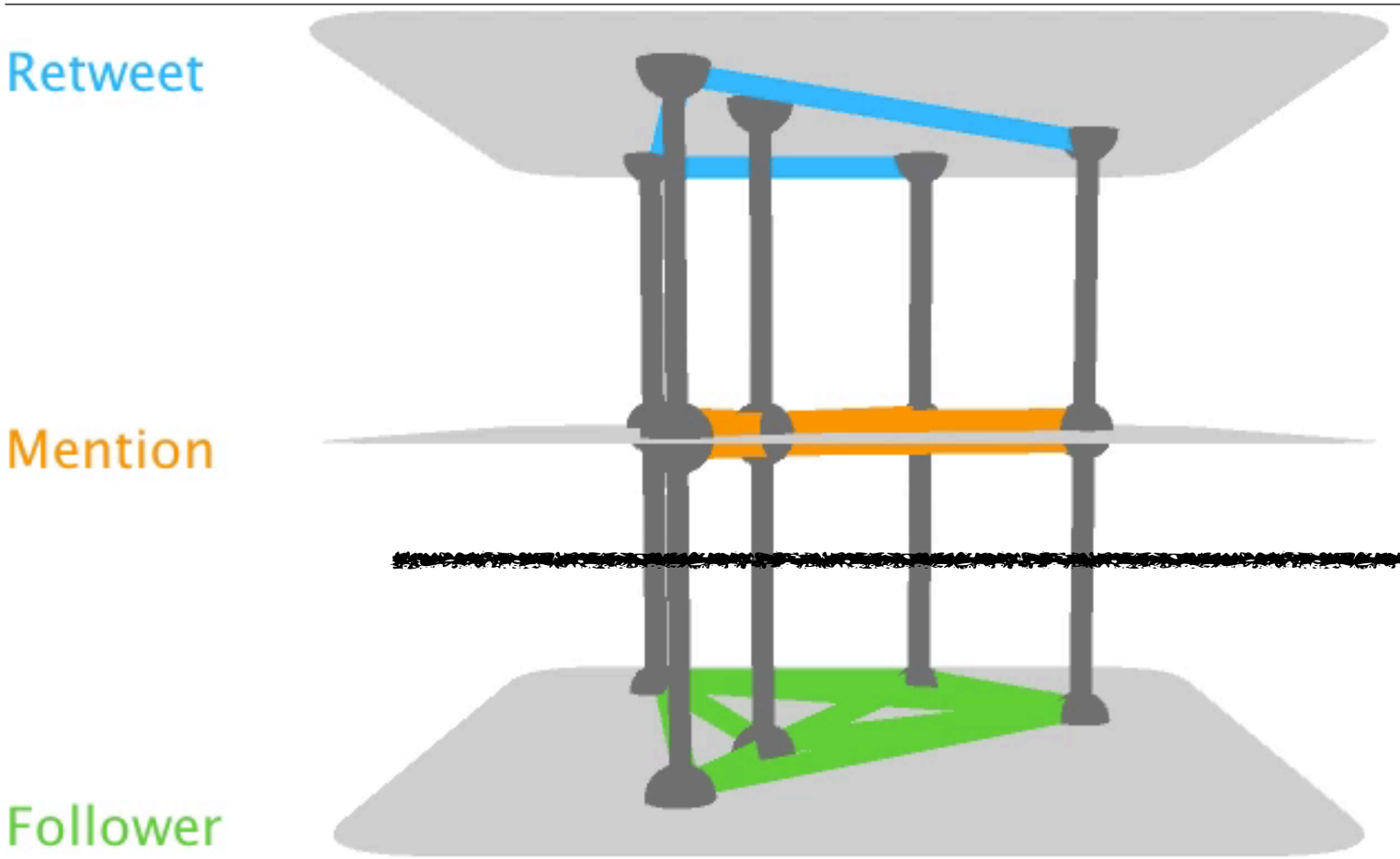
$$similarity(A,B) = \frac{|\cap \text{links of } A \text{ and } B|}{|\cup \text{links of } A \text{ and } B|}$$





Geography

Multilayer Network



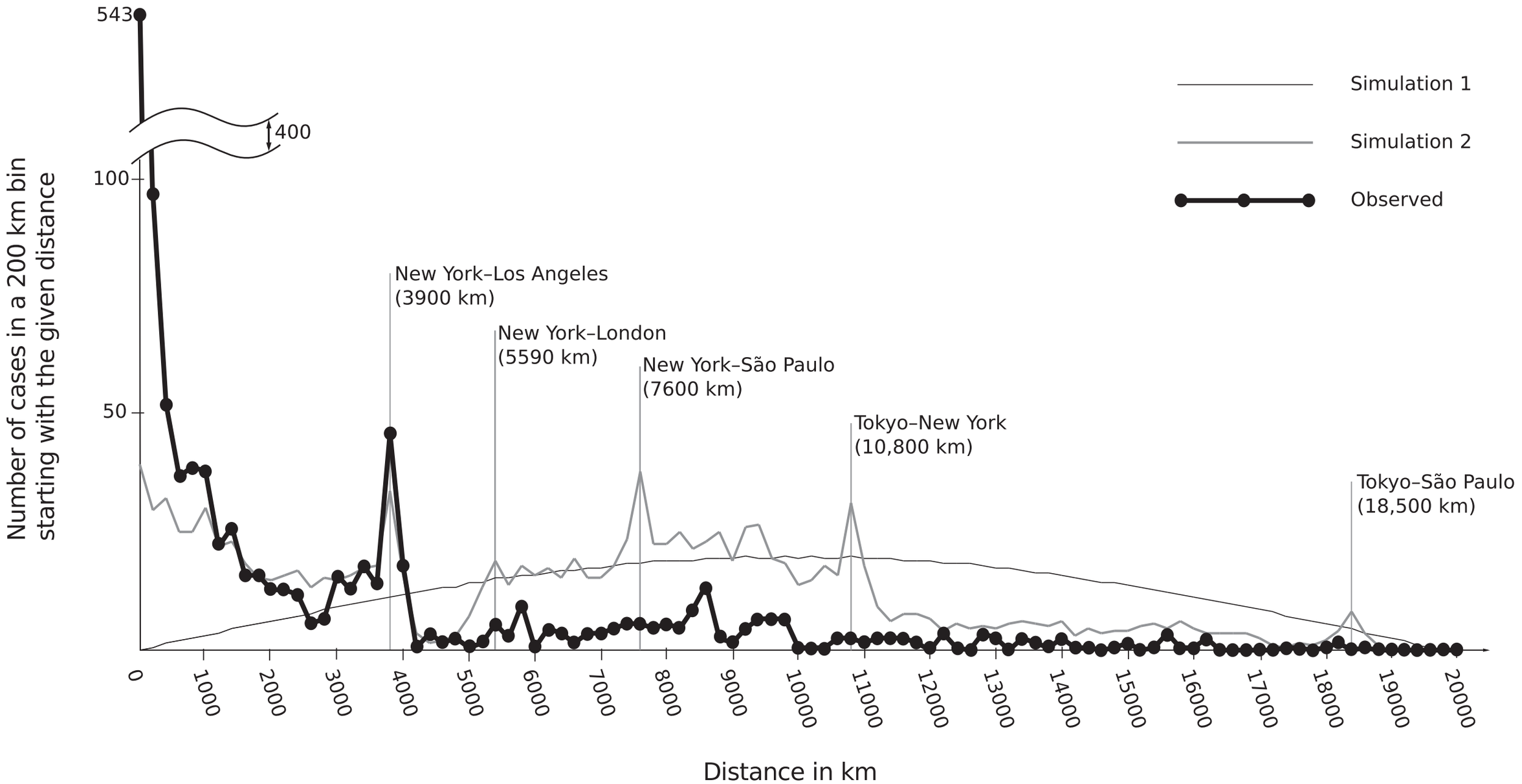
Information Layer(s)

Social Layer(s)

Geographical Layer(s)

Twitter Follower Distance

Social Networks 34, 73 (2012)



Locality

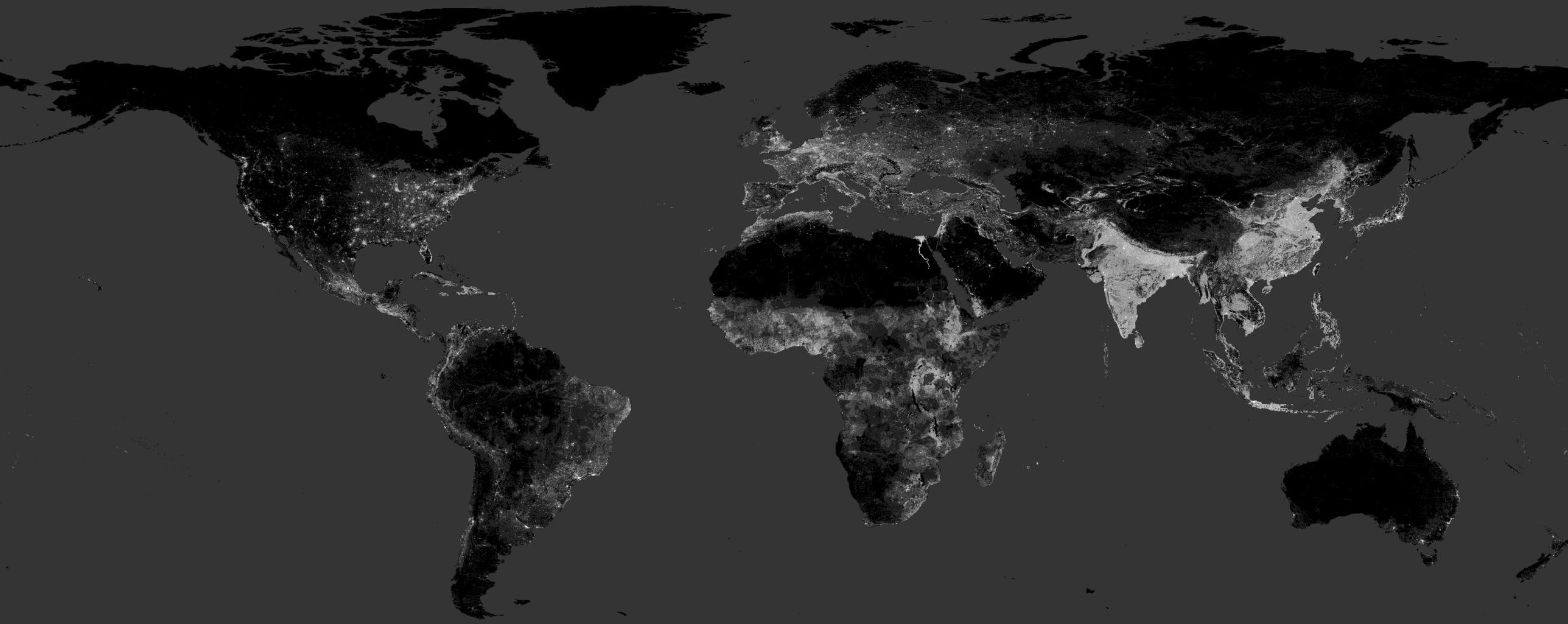
	Share of egos (%)[a]	Share of egos (%) for egos in dyads[b]	Share of alters (%)[c]	Percentage of domestic ties[d]	Percentage of domestic ties among non-local ties[d]
USA	48.5	45.7	54.5	91.6	89.3
Brazil	10.6	12.1	10.5	83.5	72.5
UK	7.6	8.3	7.6	50.6	33.3
Japan	5.5	6.5	6.3	92.1	86.0
Canada	3.7	3.8	2.9	33.3	23.1
Australia	2.7	2.7	1.9	50.0	32.0
Indonesia	2.6	1.8	1.2	60.0	25.0
Germany	2.1	1.8	1.3	62.9	58.8
Netherlands	1.4	1.4	1.2	66.7	22.2
Mexico	1.2	1.3	0.7	44.0	8.3

[a] Out of the 2852 egos located at the level of country or better.
[b] Out of the egos included in 1953 dyads with both parties located at the level of country or better.
[c] Out of the 1953 alters located at the level of country or better.
[d] The number of ties with the ego and the alter in the given country as a share of all ties for egos in that country.

Rank	Cluster[a]	Share of egos (%)[b]	Share of egos (%) for egos in dyads[c]	Share of alters (%)[d]	Locality[e]
1	“New York”	8.5	8.3	10.2	54.3
2	“Los Angeles, CA”	5.1	5.6	10.4	53.3
3	“東京都” (Tokyo)	4.1	4.8	5.0	62.9
4	“London”	3.6	3.3	4.9	48.8
5	“São Paulo”	3.5	3.0	3.6	78.4
6	“San Francisco”	2.8	2.7	4.1	41.2
7	“New Jersey”[f]	2.5	2.8	2.1	20.0
8	“Chicago”	2.2	2.0	1.7	32.0
9	“Washington, DC”	2.1	2.8	2.6	34.3
10	“Manchester, UK”	1.9	2.0	1.1	30.8
11	“Atlanta”	1.7	2.1	2.1	46.2
12	“San Diego”	1.5	1.5	1.1	26.3
13	“Toronto, Canada”	1.3	1.1	1.5	42.9
14	“Seattle”	1.3	1.4	1.2	58.8
15	“Houston”	1.2	1.2	1.0	40.0
16	“Dallas, Texas”	1.2	1.0	1.4	61.5
17	“Rio de Janeiro”	1.2	1.0	1.1	30.8
18	“Boston, MA”	1.2	1.2	1.1	20.0
19	“Amsterdam”	1.1	1.1	0.9	50.0
20	“Jakarta, Indonesia”	1.1	0.6	0.3	42.9
21	“Austin, TX”	1.0	1.0	1.3	50.0
22	“Sydney”	0.9	1.0	0.8	38.5
23	“Orlando, Forida”	0.9	1.0	0.6	16.7
24	“Phoenix, AZ”	0.8	0.7	0.6	11.1
25	“兵庫” (Hyōgo)[g]	0.8	1.0	1.0	25.0

[a] Each cluster is labeled with the name most frequently used for locations assigned to the cluster.
[b] Out of the 2167 egos located with precision of <25,000 km^2.
[c] Out of the 1259 egos included in dyads with both parties located with precision of <25,000 km^2.
[d] Out of the 1259 alters included in dyads with both parties located with precision of <25,000 km^2.
[e] Defined as the share of local of ties among all ties for egos in a cluster.
[f] Centered between Philadelphia and Trenton, NJ and includes all locations identified as just “New Jersey”.
[g] Centered near the boundary between Hyōgo and Osaka prefectures, in the Kansai region of Japan.

World Population



@bgoncalves

www.bgoncalves.com

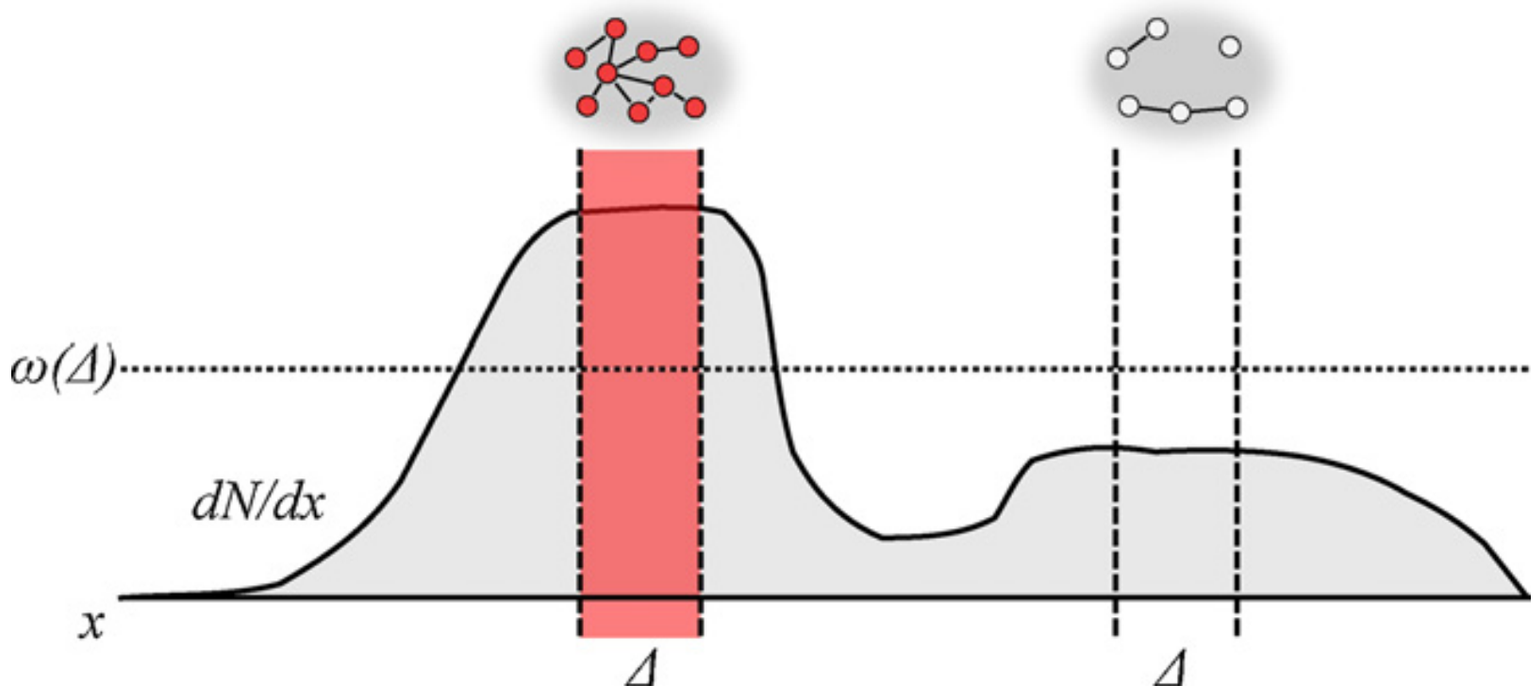
Population Heterogeneity

Social Networks 34, 82 (2012)

- Bernoulli process to generate adjacency matrix given a distance matrix between nodes

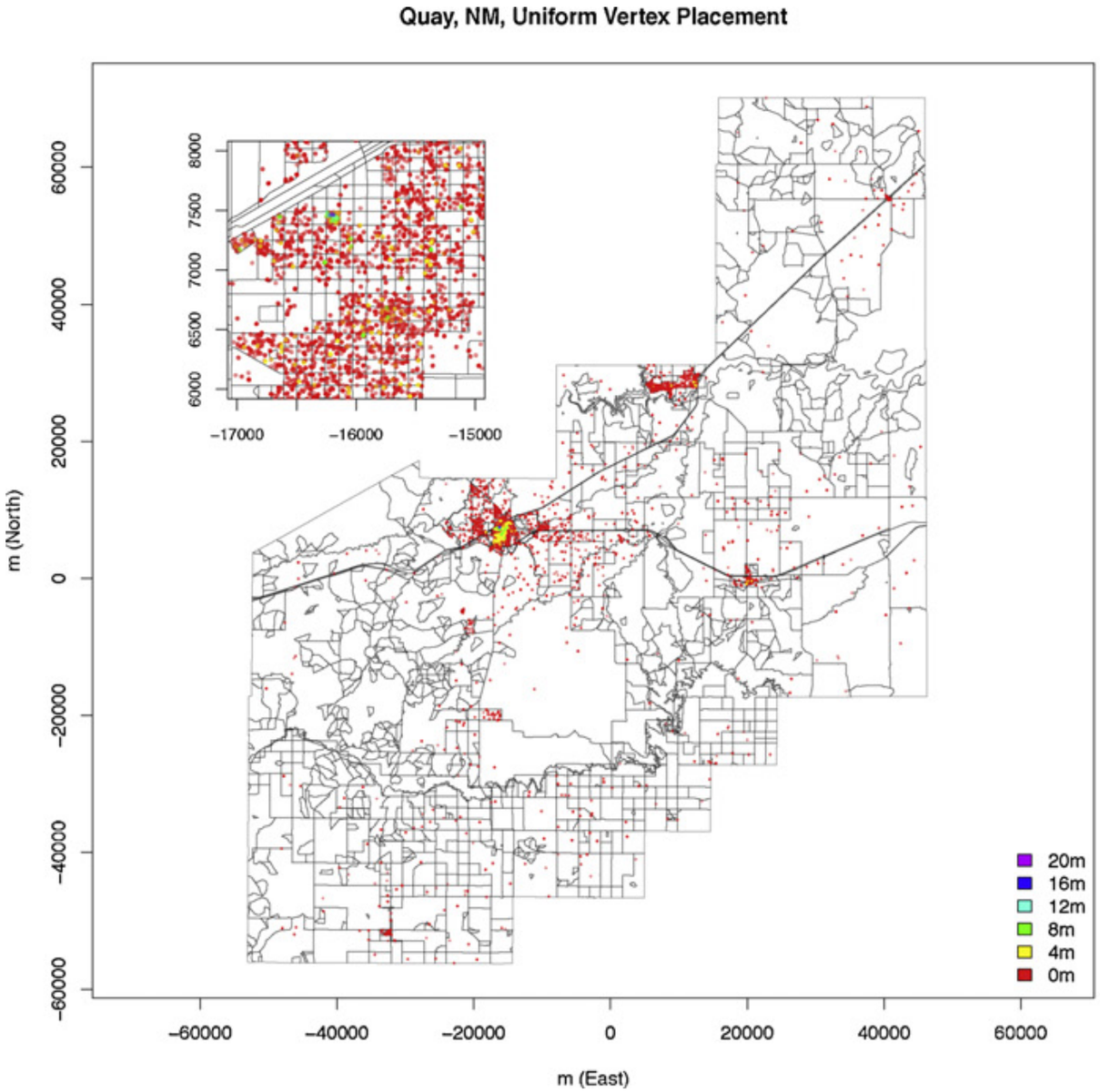
$$P\left(A=a|D\right)=\prod_{\{i,j\}}B\left(A_{ij}=a_{ij}|\mathcal{F}\left(D_{ij},\theta\right)\right)$$

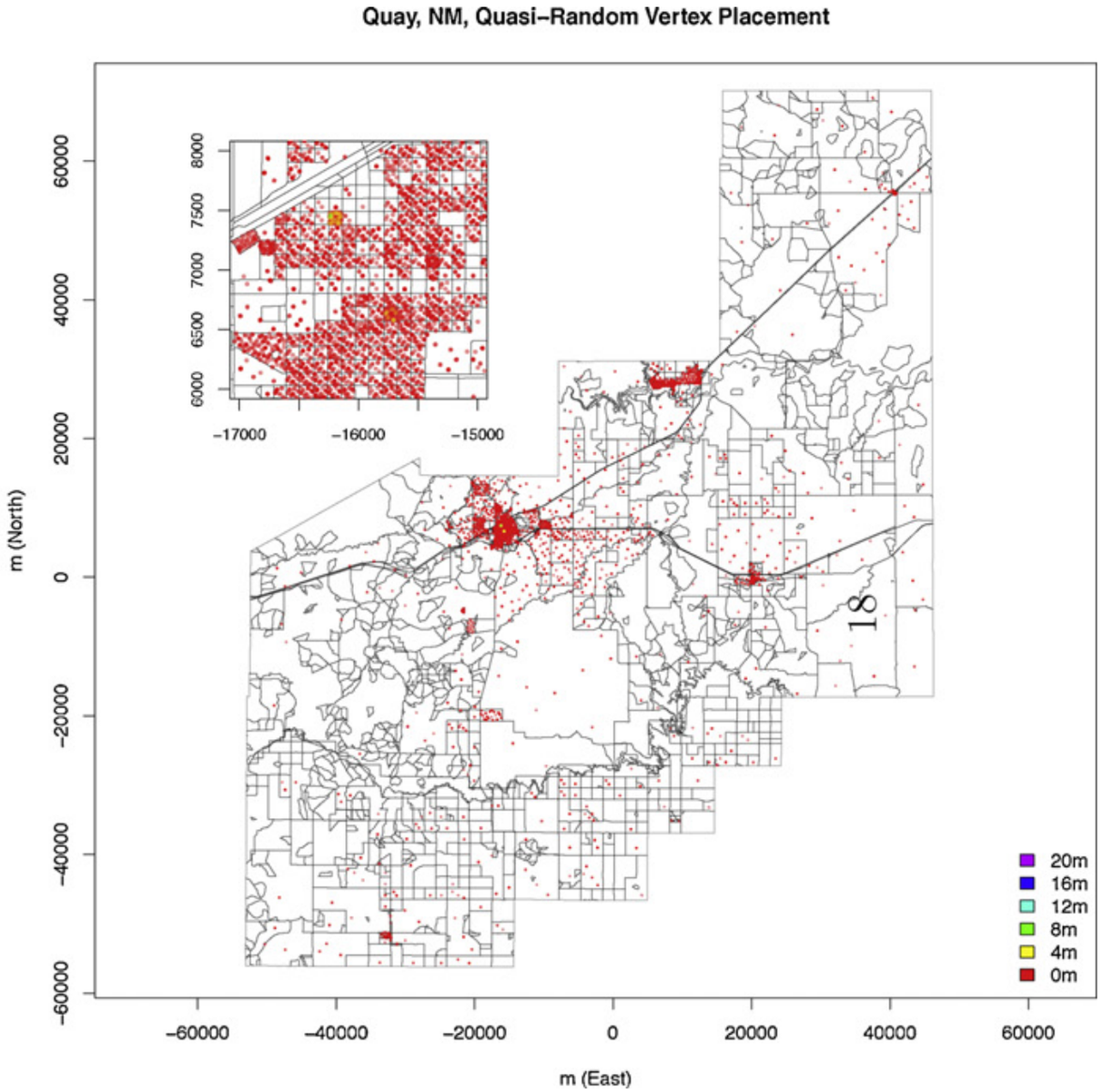
- Above some density threshold, $\omega\left(\Delta\right)$ networks is naturally connected.



Vertex Placement

Social Networks 34, 82 (2012)





Friendship probability

Social Networks 34, 82 (2012)

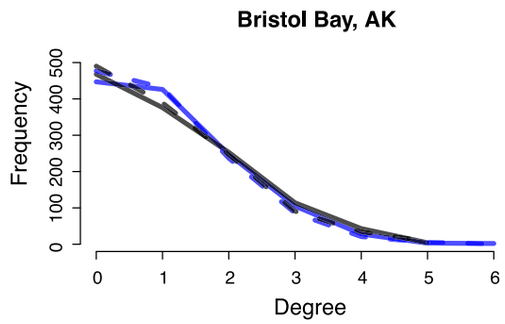
- Probability that two people are friends as a function of distance:

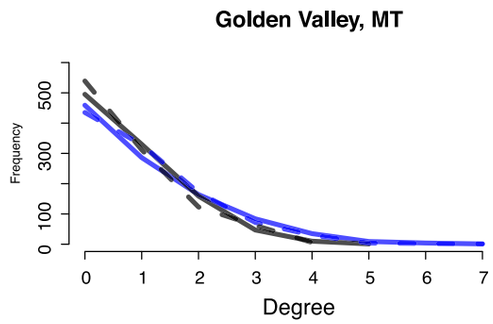
$$\mathcal{F}(d) = \frac{\theta_1}{(1+\theta_2 d)^{\theta_3}}$$

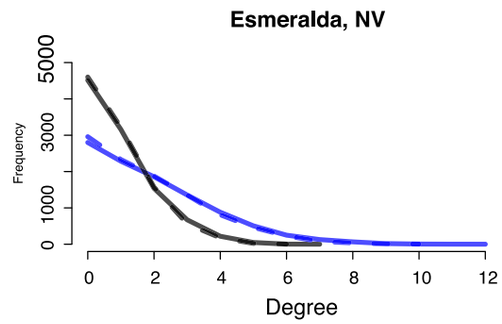
- with (0.533, 0.032, 2.788) for "social friendships" and (0.859, 0.035, 6.437) for "face-to-face interactions".

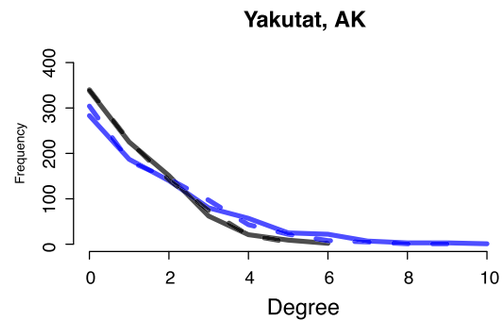
Social Network Properties

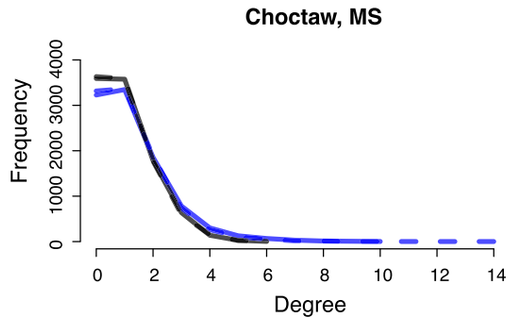
Social Networks 34, 82 (2012)

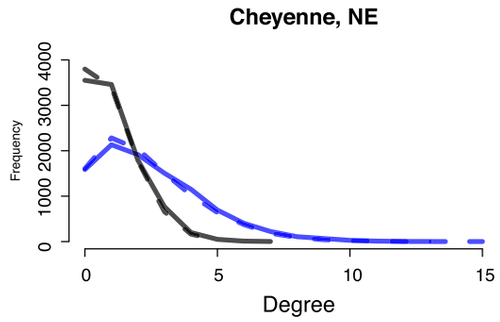


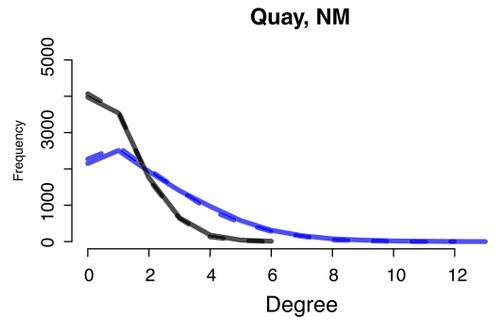


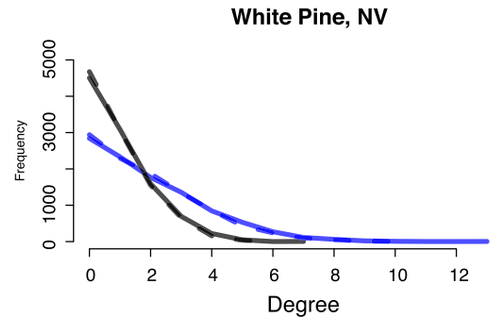


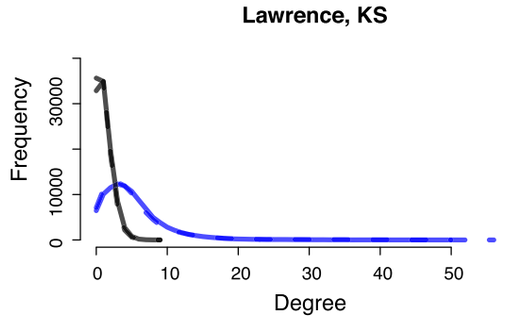


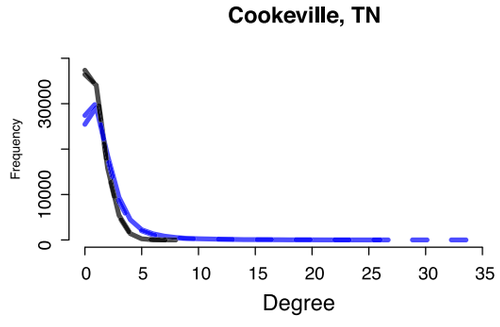


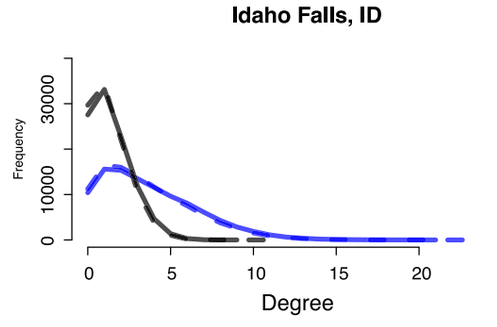


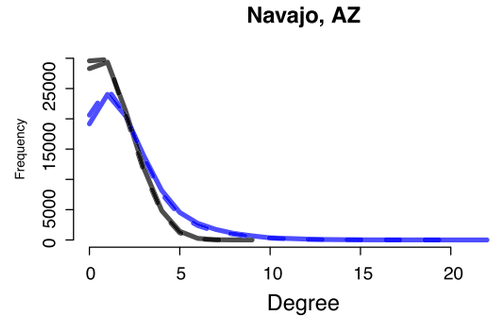


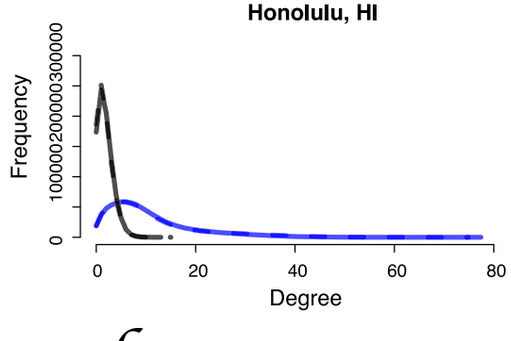


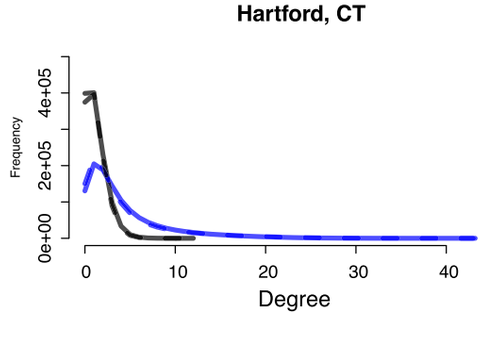


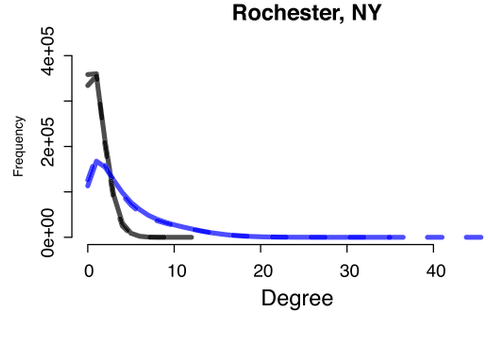


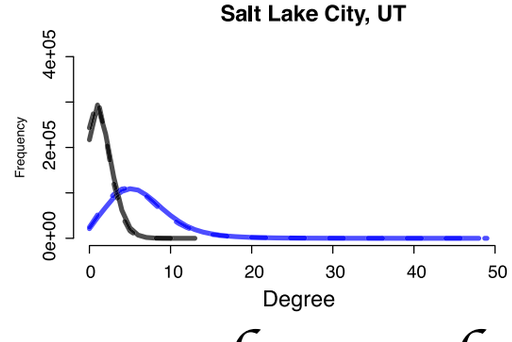










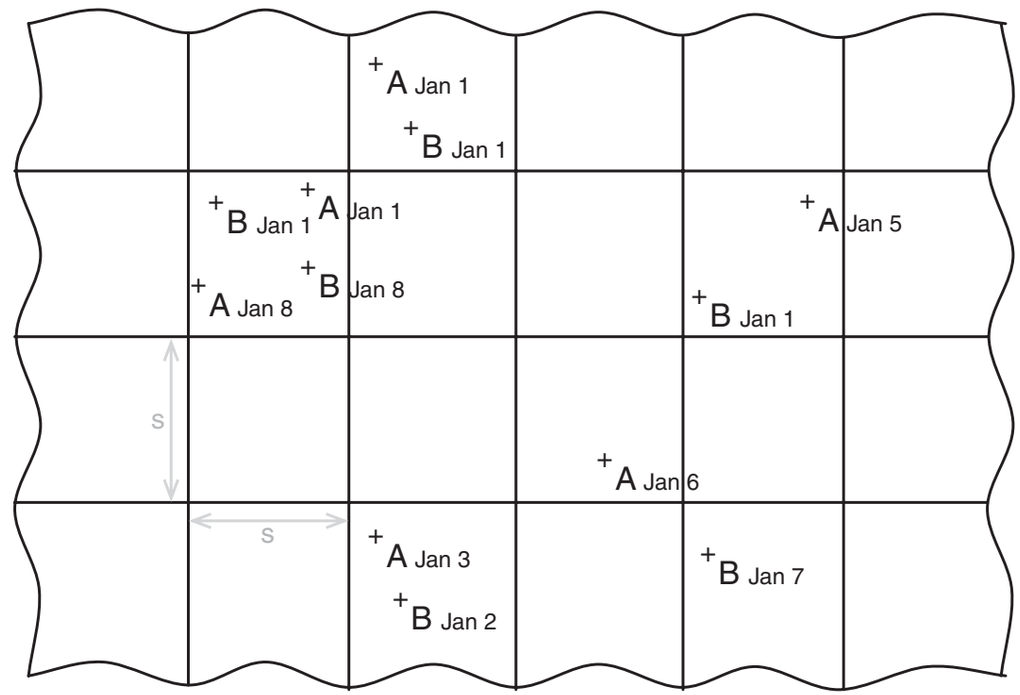


Co-occurences and Social Ties

PNAS 107, 22436 (2010)

- Geotagged Flickr Photos

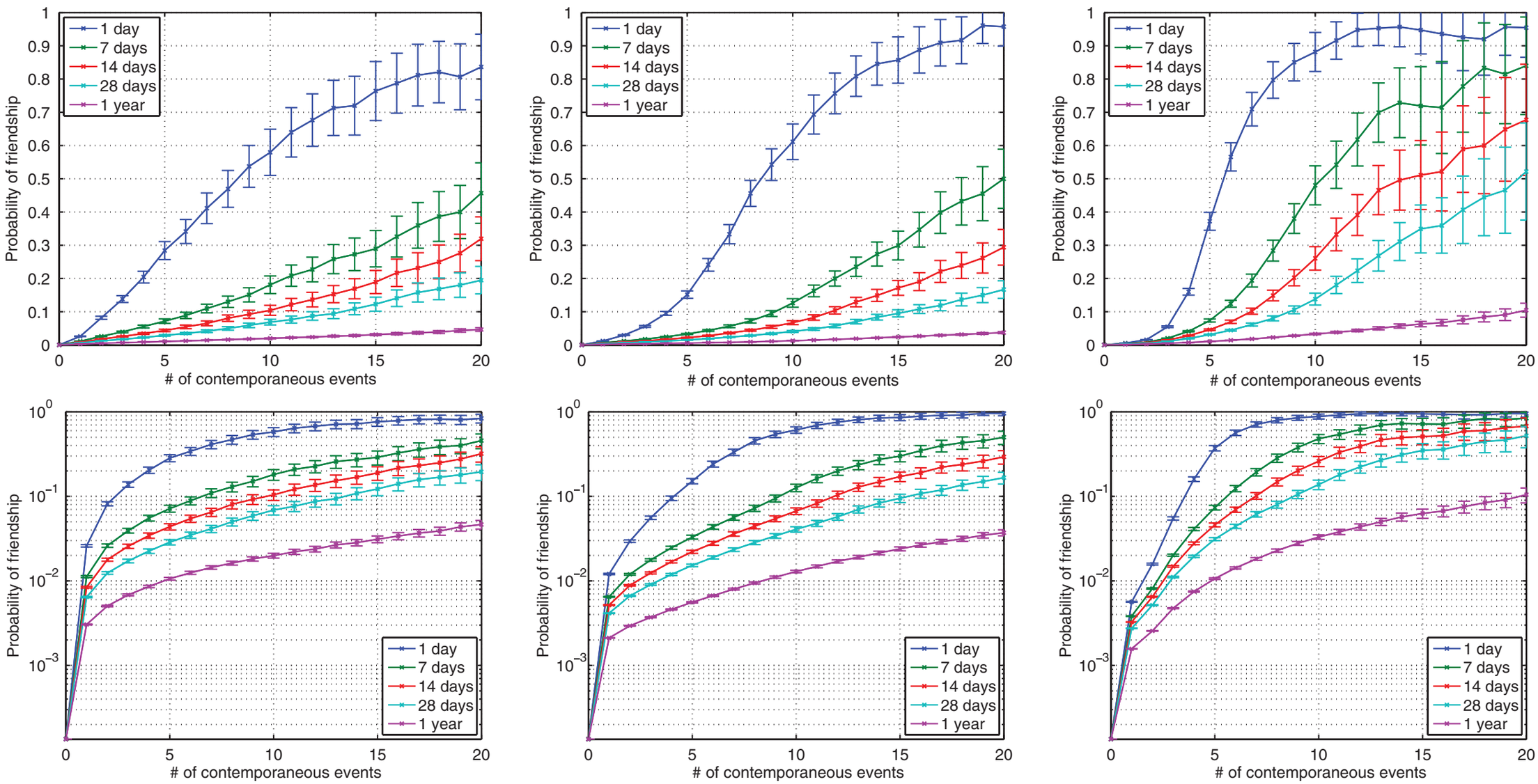
- Divide the world into a grid



Count number of cells on which two individuals were within a given interval

Co-occurences and Social Ties

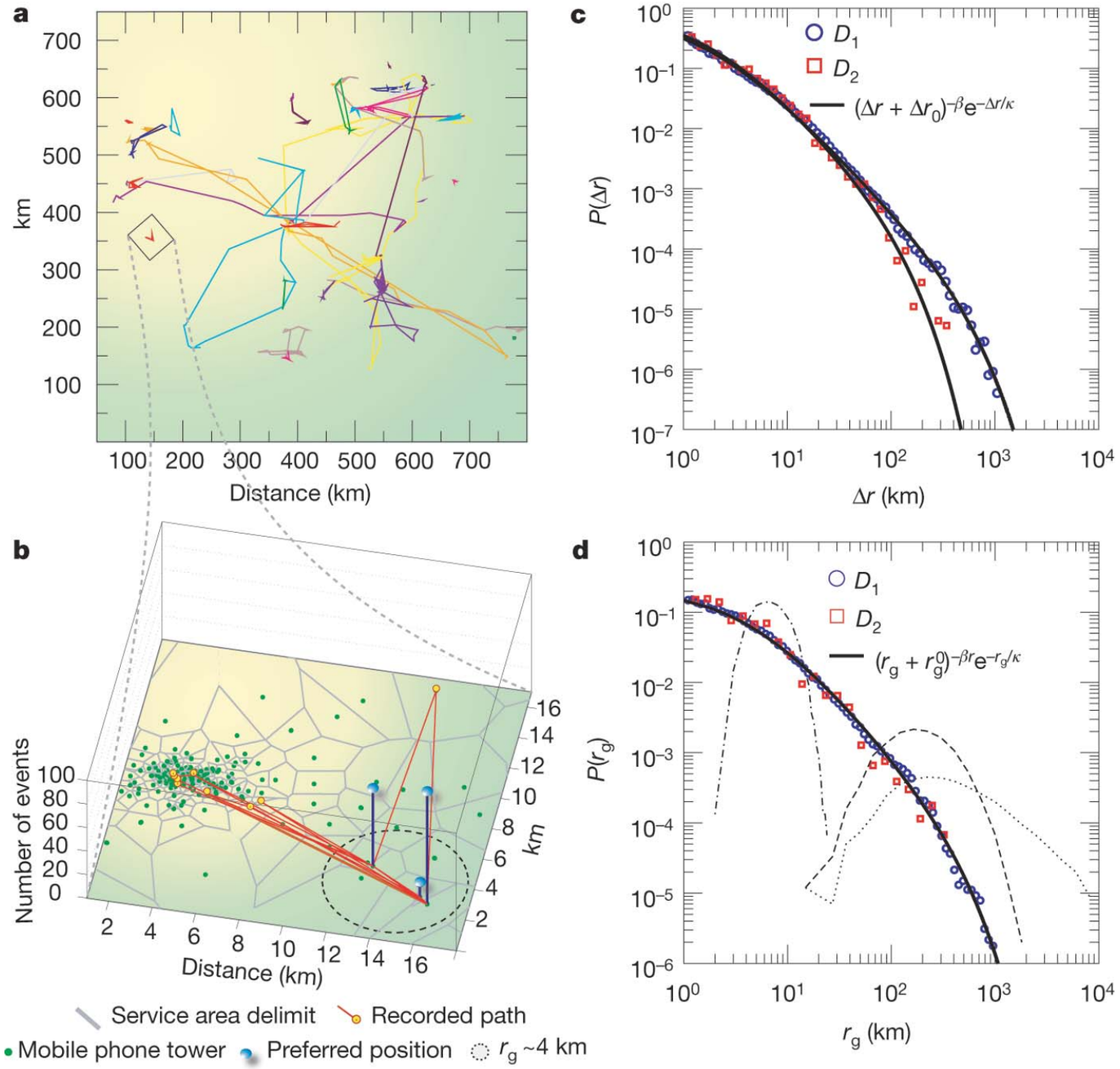
PNAS 107, 22436 (2010)



Geolocated Tweets

Human Mobility

Nature 453, 779 (2008)



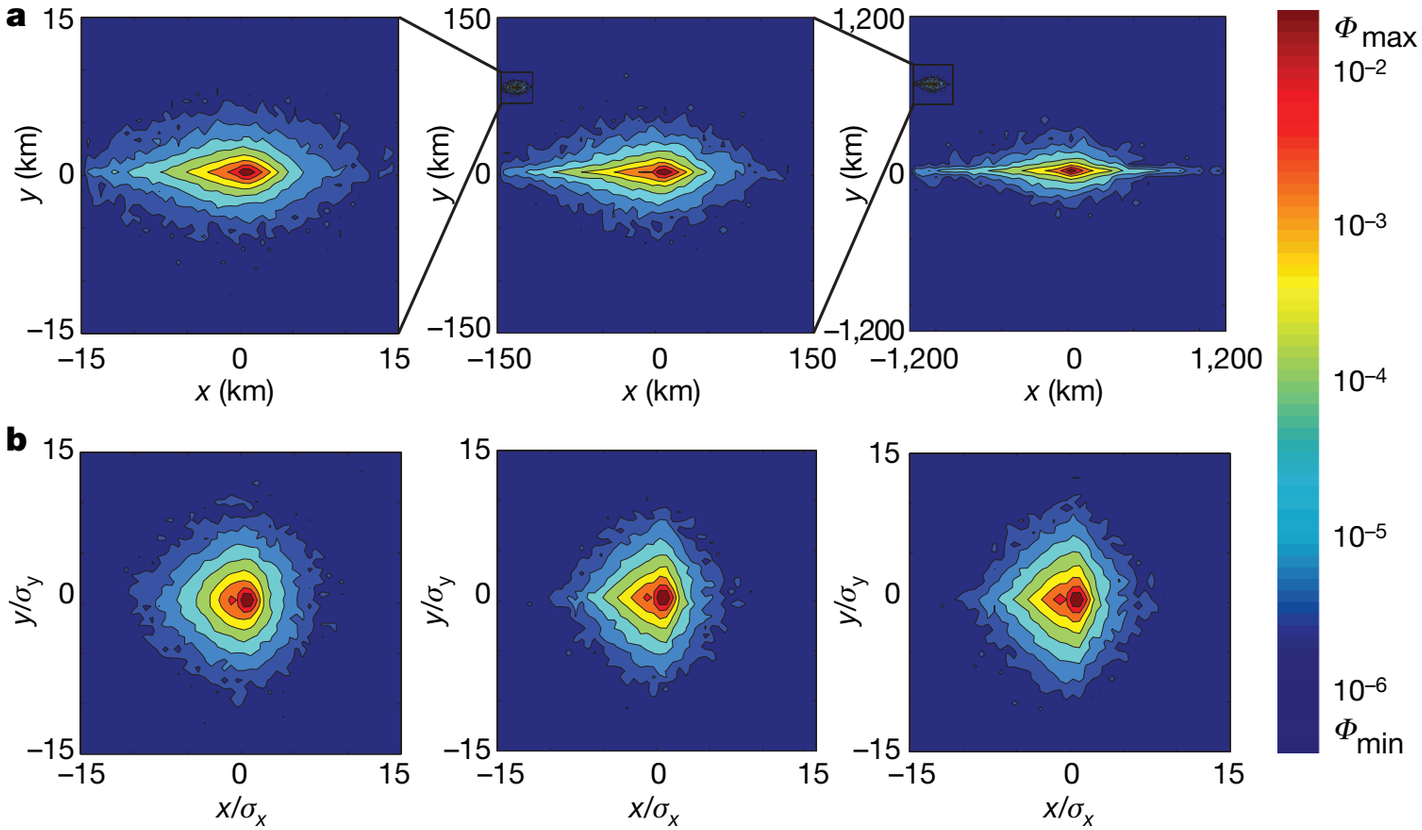
$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-r_g/\kappa)$$

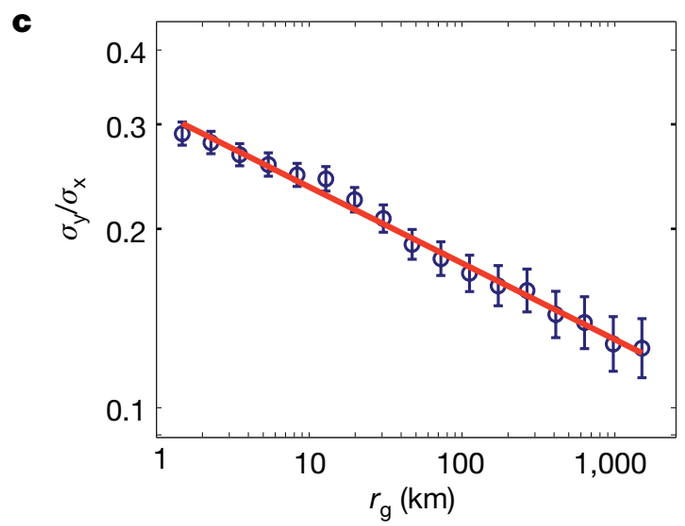
@bgoncalves

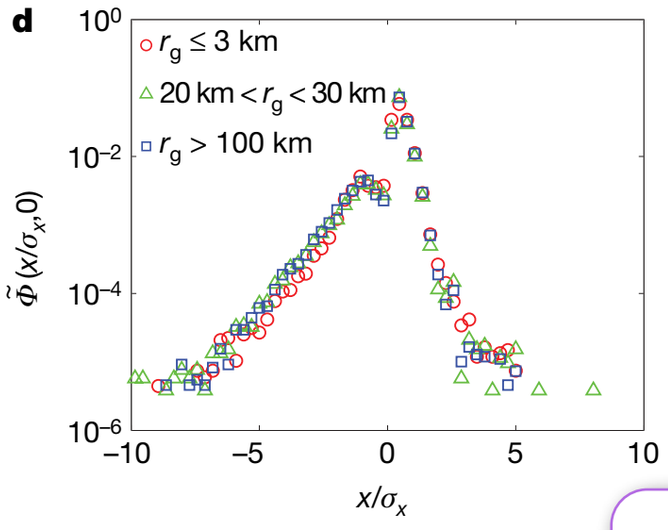
Cell Phones

Human Mobility

Nature 453, 779 (2008)





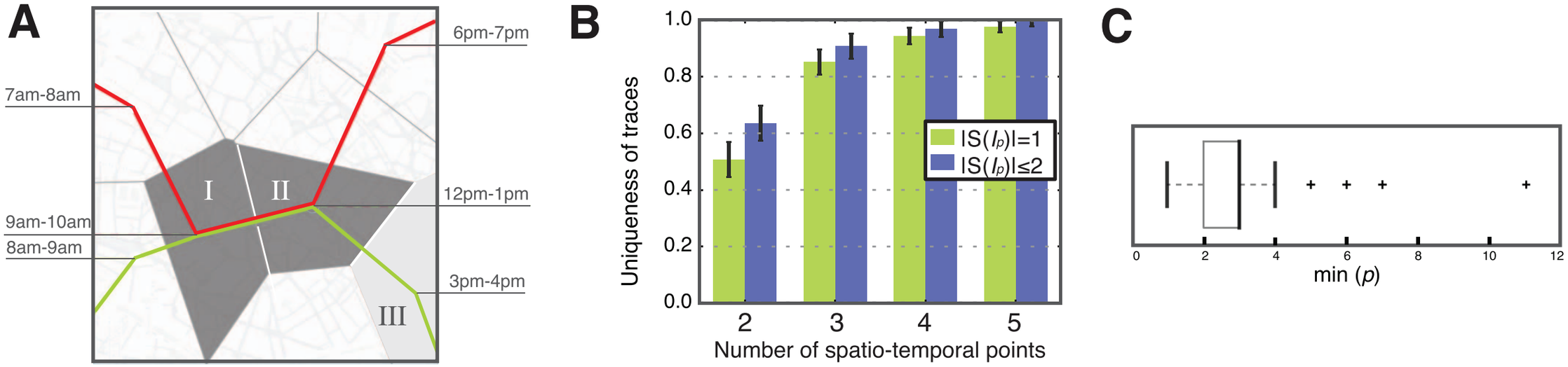


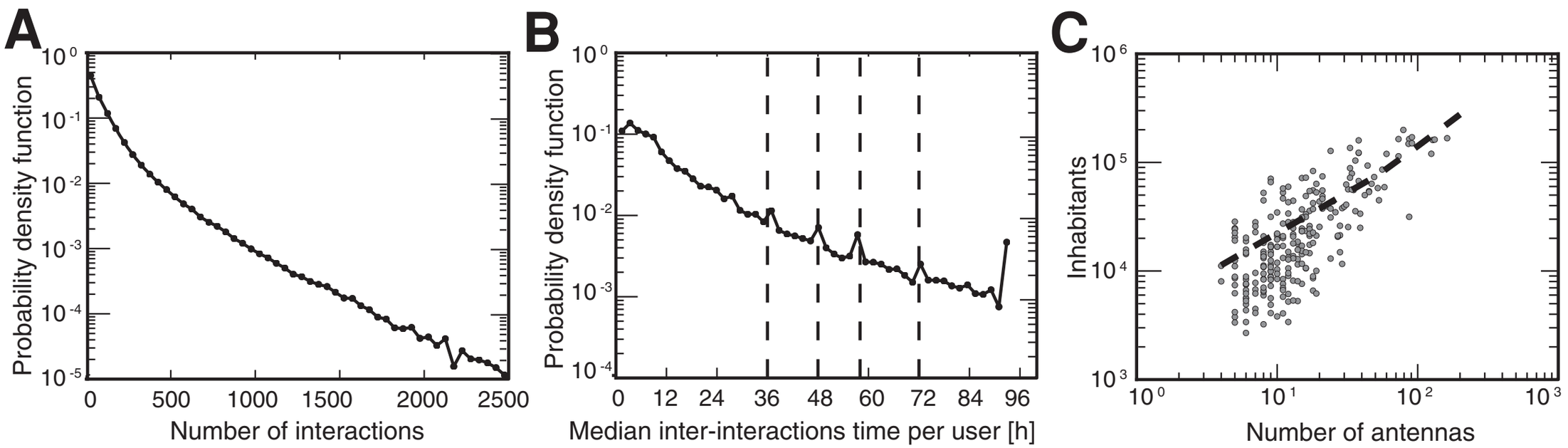
@bgoncalves

Cell Phones

Privacy

Sci Rep 3, 1376 (2013)





@bgoncalves

Cell Phones

Sci Rep 3, 1376 (2013)