Decoding Social Interactions in the Web

Rossano Schifanella

@eBISS 2018

Berg en Dal, The Netherlands

Who Am I?

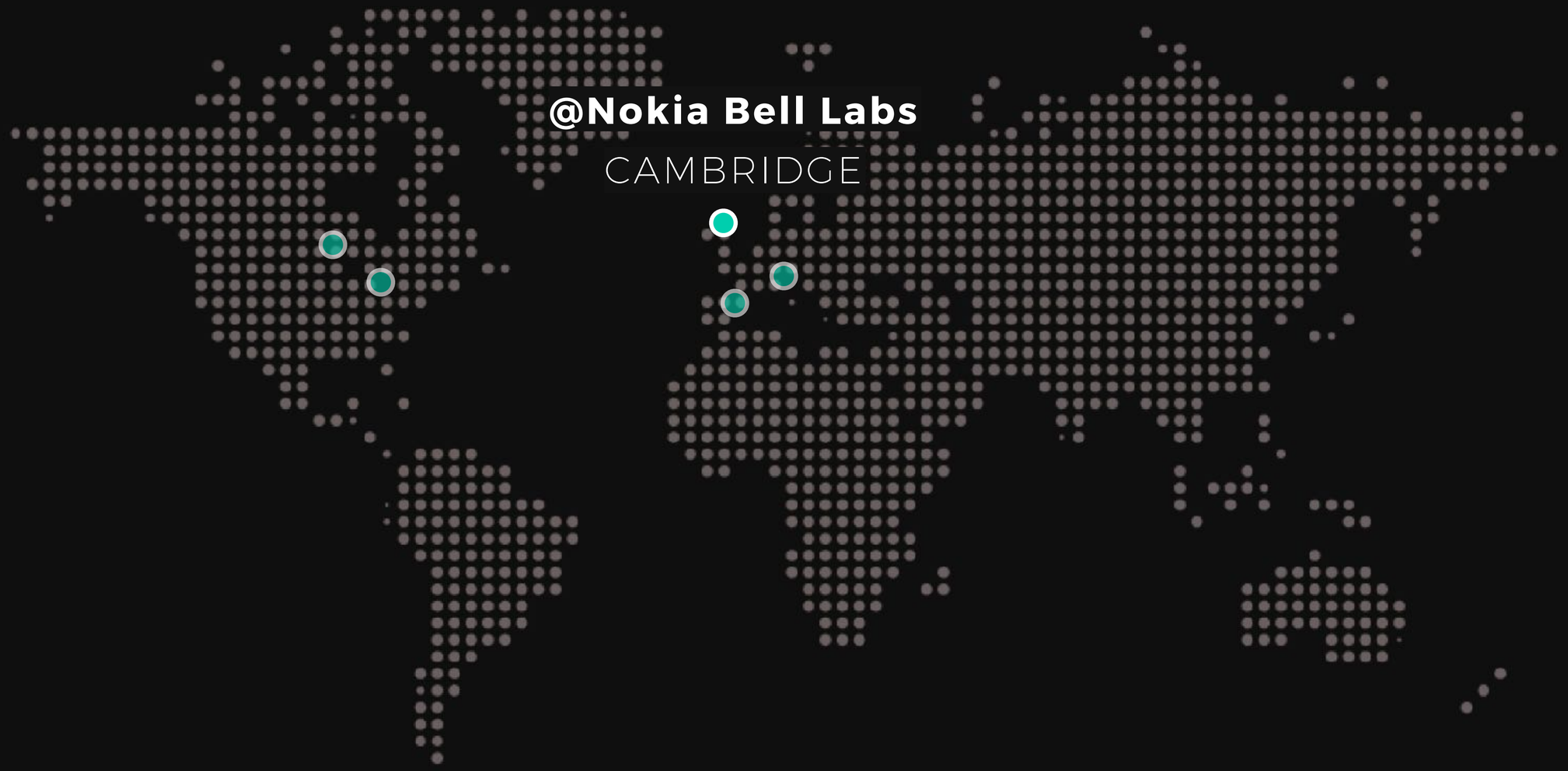
DISTRIBUTED SYSTEMS
RECOMMENDER SYSTEMS

NETWORK SCIENCE
HUMAN COMPUTING

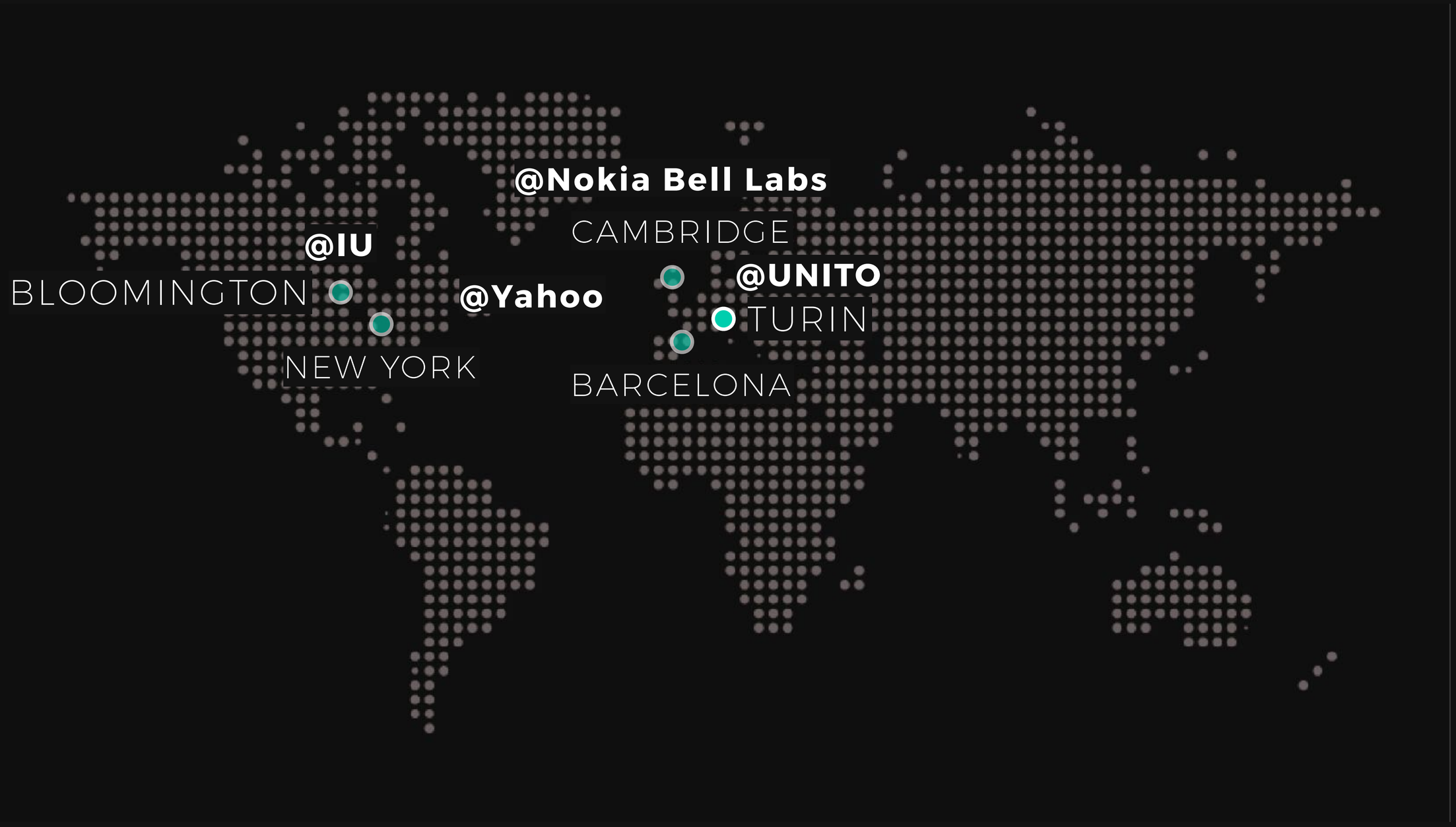
CROWDSOURCING

URBAN COMPUTING
MACHINE LEARNING

BEHAVIOURAL STUDIES
COMPUTATIONAL *



URBAN COMPUTING
COMPUTATIONAL SOCIAL SCIENCE


@Nokia Bell Labs
CAMBRIDGE
@IU
BLOOMINGTON
@Yahoo
@UNITO
TURIN
NEW YORK
BARCELONA

Who Are You?

Acknowledgements

Slides and material have been inspired by:

CSS Tutorial@WWW16

C. Wagner, L. Aiello, M. Strohmaier, Ingmar Weber

Computational Social Science course

Claudia Wagner, Universität Koblenz-Landau

Algorithmic Bias Tutorial@KDD16

H. Hajian, F. Bonchi, C. Castillo

Social Network Course

Dennis M. Feehan, Berkeley (very estensive literature review)

Computational Social Science Course

John McLevey, University of Waterloo

Agenda

- Introduction to Computational Social Science
- The Web as a data source
- Modeling social interactions through networks
- Basic networks metrics
- The importance of social theories

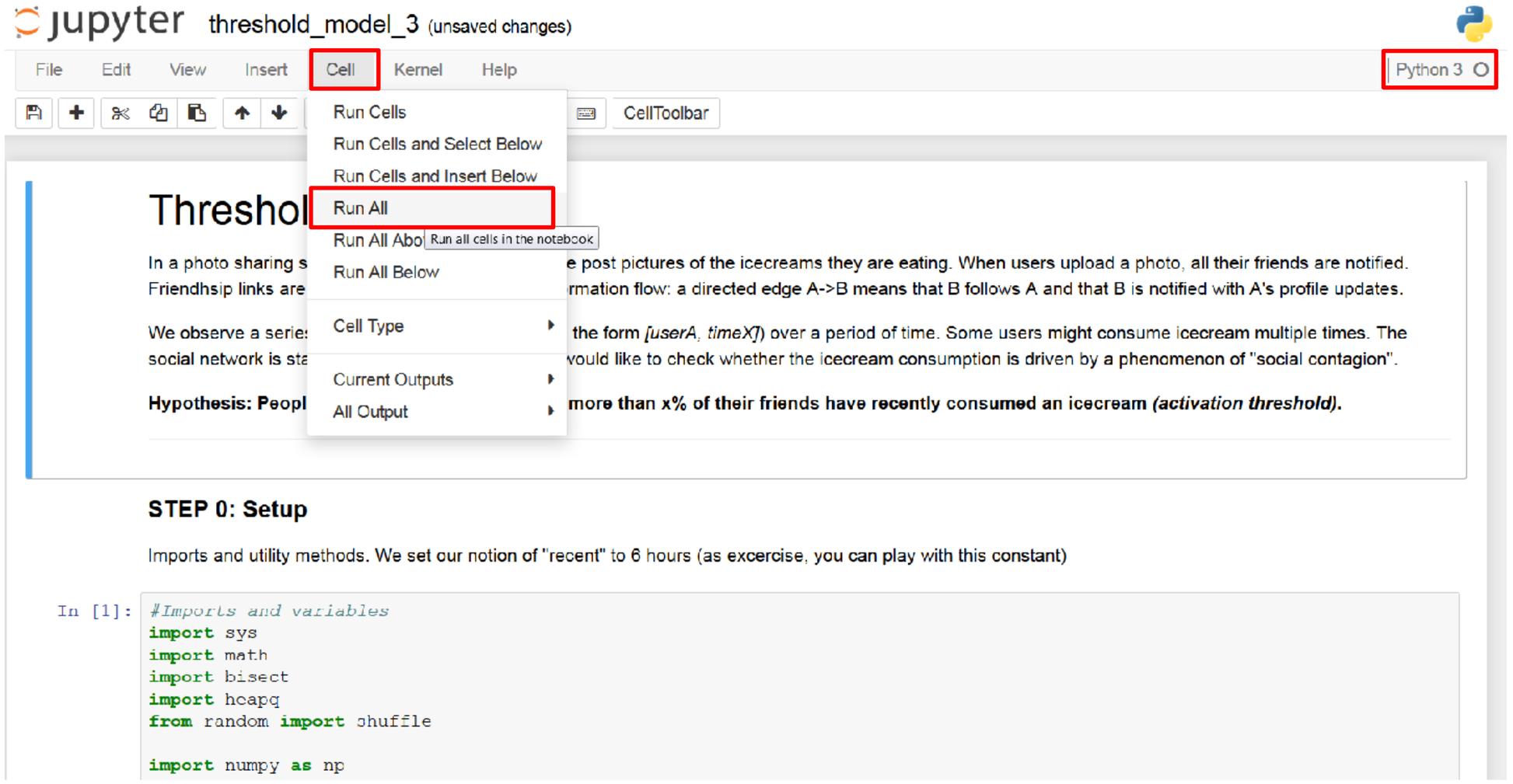
Setup for Practical Session

Setup for Practical Exercises

- We will have practical examples throughout the tutorial
  - Demoing some code. Follow along or try later.
- I have already prepared IPython notebooks for you

- Requirements:
  - Python 3
  - Project Jupyter (https://jupyter.org/)
  - networkx

- Download data and source files from:
  - https://github.com/rschifan/ebiss2018


jupyter threshold_model_3 (unsaved changes)
File Edit View Insert Cell Kernel Help
Python 3
CellToolbar
Run Cells
Run Cells and Select Below
Run Cells and Insert Below
Run All
Run All Abo
Run all cells in the notebook
Run All Below
Cell Type
Current Outputs
All Output
Threshol
In a photo sharing s
e post pictures of the icecreams they are eating. When users upload a photo, all their friends are notified.
Friendhsip links are
rmation flow: a directed edge A->B means that B follows A and that B is notified with A's profile updates.
We observe a serie
the form [userA, timeX]) over a period of time. Some users might consume icecream multiple times. The
social network is sta
would like to check whether the icecream consumption is driven by a phenomenon of "social contagion".
Hypothesis: Peopl
more than x% of their friends have recently consumed an icecream (activation threshold).
STEP 0: Setup
Imports and utility methods. We set our notion of "recent" to 6 hours (as excercise, you can play with this constant)
In [1]:
#Imports and variables
import sys
import math
import bisect
import heapq
from random import shuffle
import numpy as np

Introduction to Computational Social Science

What is Computational Social Science?

Science that investigates **social phenomena** through the medium of **computing** and **algorithmic data processing**.

Go beyond prediction by adding **interpretability** and **explainability.**

Interdisciplinary data science.

biogeographic patterns. Their study, too, is centered on a large database, but in this case it is entirely of living organisms, the marine bivalves. Over 28,000 records of bivalve genera and subgenera from 322 locations around the world have now been compiled by these authors, giving a global record of some 854 genera and subgenera and 5132 species. No fossils are included in the database, but because bivalves have a good fossil record, it is possible to estimate accurately the age of origin of almost all extant genera. It is then possible to plot a backward survivorship curve (8) for each of the 27 global bivalve provinces (9).

On the basis of these curves, Krug *et al.* find that origination rates of marine bivalves increased significantly almost everywhere immediately after the K-Pg mass extinction event. The highest K-Pg origination rates all occurred in tropical and warm-temperate regions. A distinct pulse of bivalve diversification in the early Cenozoic was concentrated mainly in tropical and subtropical regions (see the figure).

The steepest part of the global backward survivorship curve for bivalves lies between 65 and 50 million years ago, pointing to a major biodiversification event in the Paleogene (65 to 23 million years ago) that is perhaps not yet captured in Alroy *et al.*'s database (5, 7). The jury is still out on what may have caused this event. But we should not lose sight of the fact that the steep rise to prominence of many modern floral and faunal groups in the Cenozoic may bear no simple relationship to climate or any other type of environmental change (10, 11).

References

1. G. G. Wittebach *et al.*, *Ecol. Lett.* **10**, 315 (2007).
2. A. Z. Krug, D. Jablonski, J. W. Valentine, *Science* **323**, 767 (2009).
3. P. W. Signor, *Annu. Rev. Ecol. Syst.* **21**, 539 (1990).
4. R. K. Bambach, *Geobios* **32**, 131 (1999).
5. J. Alroy *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6261 (2001).
6. A. M. Bush *et al.*, *Paleobiology* **30**, 666 (2004).
7. J. Alroy *et al.*, *Science* **321**, 97 (2008).
8. M. Foote, in *Evolutionary Patterns*, J. B. C. Jackson *et al.*, Eds. (Univ. of Chicago Press, Chicago, IL, 2001), vol. 245, pp. 245–295.
9. M. D. Spalding *et al.*, *Bioscience* **57**, 573 (2007).
10. S. M. Stanley, *Paleobiology* **33**, 1 (2007).
11. M. J. Benton, B. C. Emerson, *Palaeontology* **50**, 23 (2007).

10.1126/science.1169430

SOCIAL SCIENCE

Computational Social Science

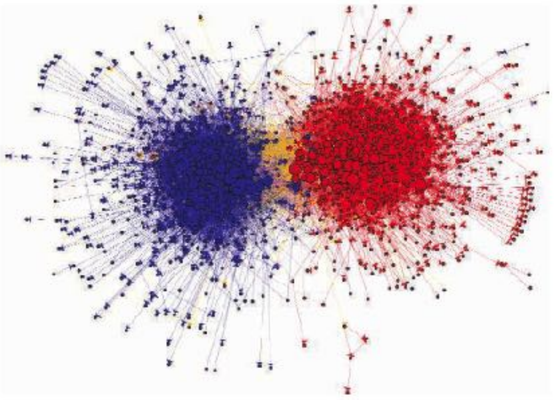
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[7] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the



Data from the blogosphere. Shown is a link structure with n a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (8) with permission from the Association for Computing Machinery]

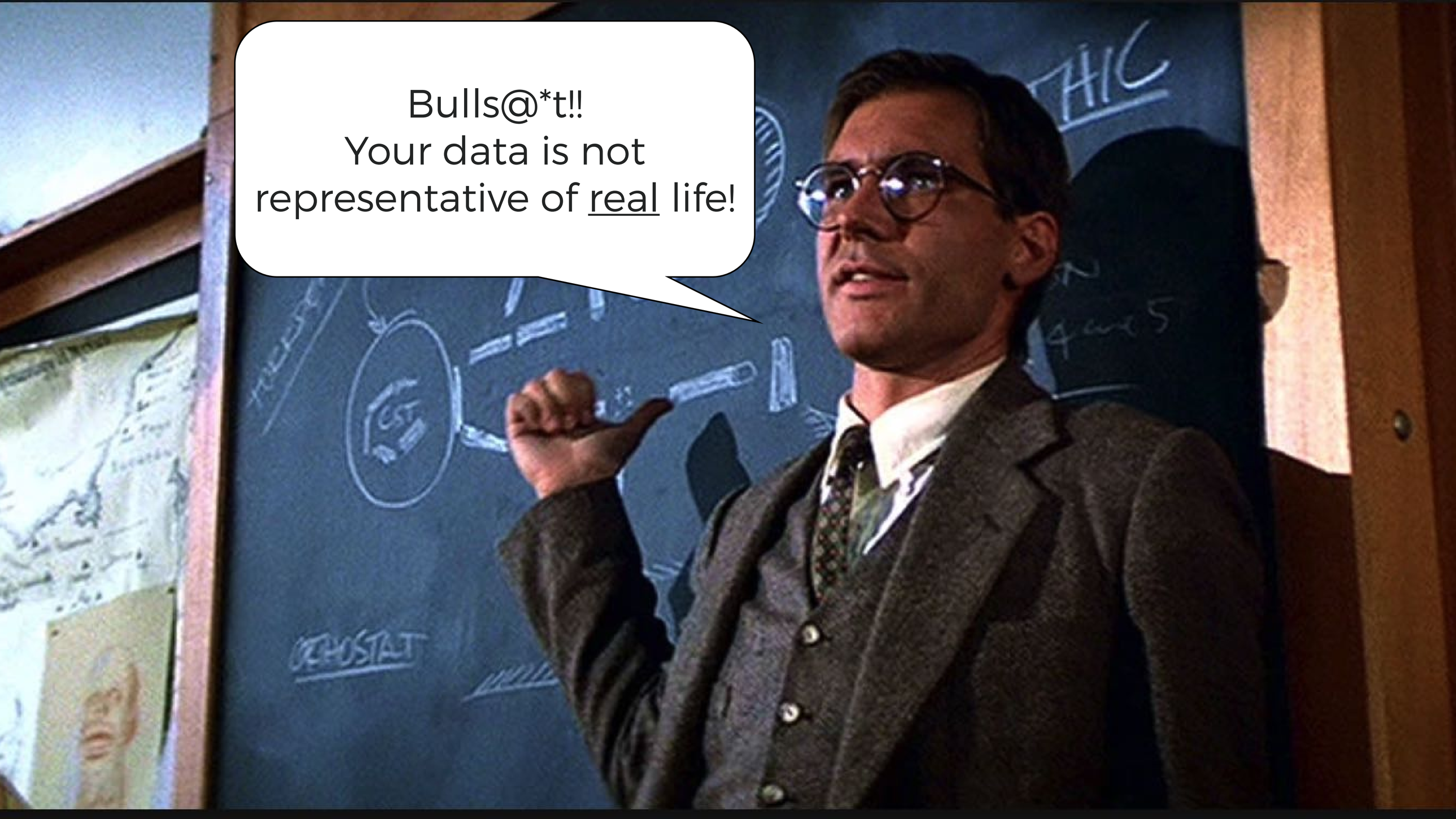
[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA. [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

Published by AAAS

[SCIENCE 2009] http://science.sciencemag.org/content/323/5915/721.long


multidisciplinary (a simplified view)


Bulls@*t!!
Those are all obvious findings!


Bulls@*t!!
Your data is not representative of real life!

What is Computational Social Science?

Social Science Questions

+

Computational Methods and Systems

+

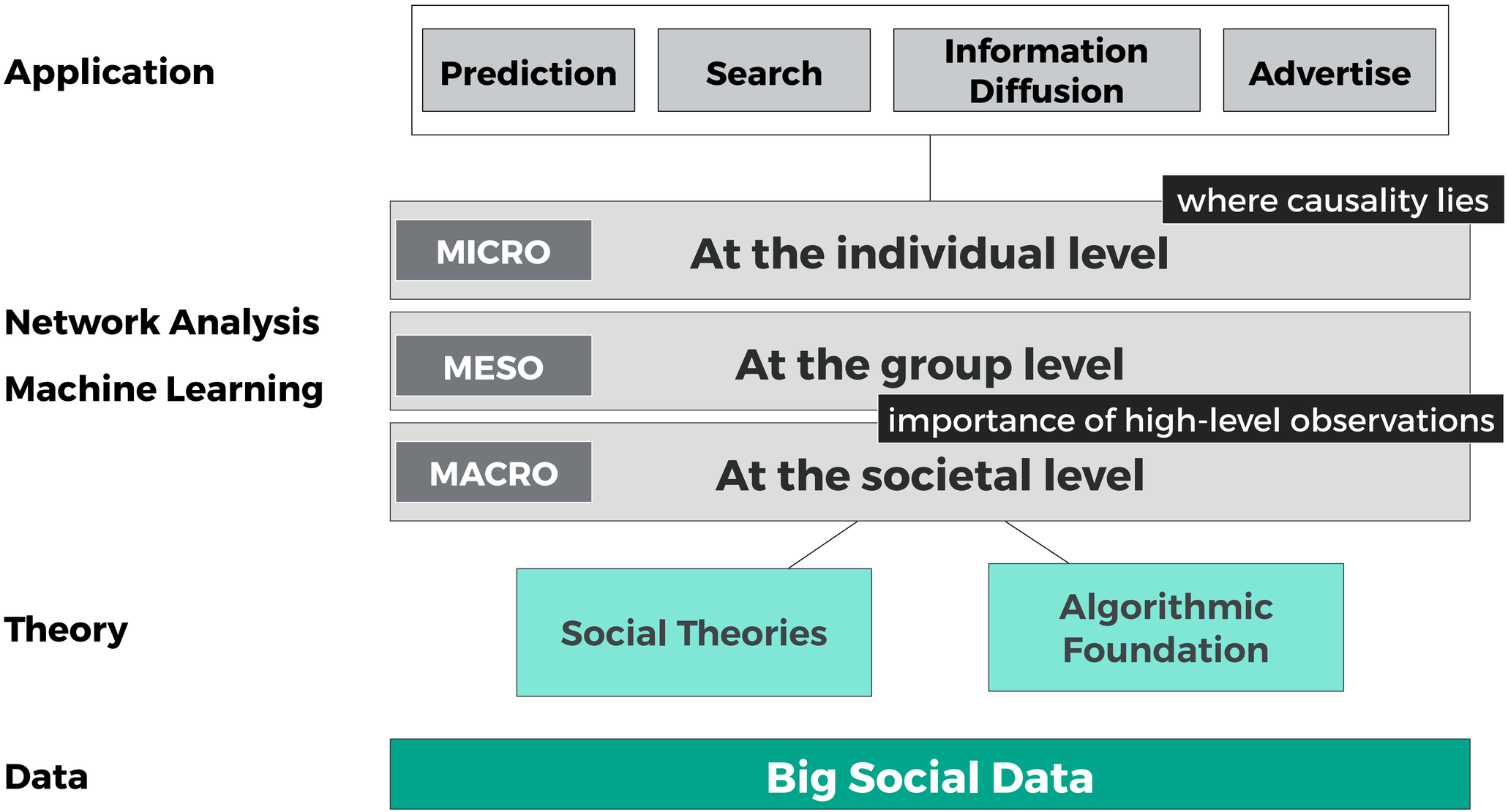
New Types of Data

The Social Sciences (a simplified view)

are interested in understanding how people

- **think/feel/behave in social situations** (social psychology),
- **relate to each other** (sociology),
- **govern themselves** (political science),
- **handle wealth** (socioeconomics), and
- **create culture** (anthropology).

- **understanding human behavior**


Application
Prediction
Search
Information Diffusion
Advertise
where causality lies
Network Analysis
Machine Learning
MICRO
At the individual level
MESO
At the group level
importance of high-level observations
MACRO
At the societal level
Theory
Social Theories
Algorithmic Foundation
Data
Big Social Data

Examples (Micro-Level)

- **Status attainment**
  - why do people obtain their position in society? Status attainment is affected by both achieved factors, such as educational attainment, and ascribed factors, such as family income
- **Radicalization or Crime**
  - why do some individuals radicalize/become criminal?
- **Voter Turnout**
  - what impact who will vote?
- **Voting behavior**
  - which factors explain political leaning?
- **Happiness**
  - what makes people happy? Which factors explain happiness?
- **Opinion Formation**
  - how do people form opinions? Which factors impact the opinion formation process most?
- **Mental Health**
  - why does religion impact mental health? What mechanism explains that?

Examples

Meso-level

- **Team performance/creativity**
  - what makes team successful? Which factors explain team success?
- **Cooperation**
  - what impacts that people cooperate, i.e. contribute to a public good?
- **Social Norms**
  - how do groups of people reach consensus?

Macro-level

- **Inequality**
  - which factors and mechanisms explain the emergence of inequality?
- **Polarization**
  - what leads to polarization in society?

Why is understanding human behavior difficult?

1. **Dealing with fuzzy concepts**

   How do you measure “integration”?

   Often dealing with vague yet fundamental concepts: happiness, intelligence, discrimination

   Potential solution:

   Ensure **construct validity** and **reliability**

Why is understanding human behavior difficult?

2. Can’t put society in a test tube

Randomized controlled trials on society?

- “Let me randomly put into place tougher laws in some locations, but not in others.”
- Illegal, unethical, impractical, …

Potential solution:

Causal inference from observational data

Why is understanding human behavior difficult?

3. Humans aren’t machines

Human behavior is **non-deterministic**

Different people react differently in the same situation.

Think “smoking causes cancer”. Certainly yes, but not for everybody.

Potential solution:

Probabilistic reasoning and **expected effect size**

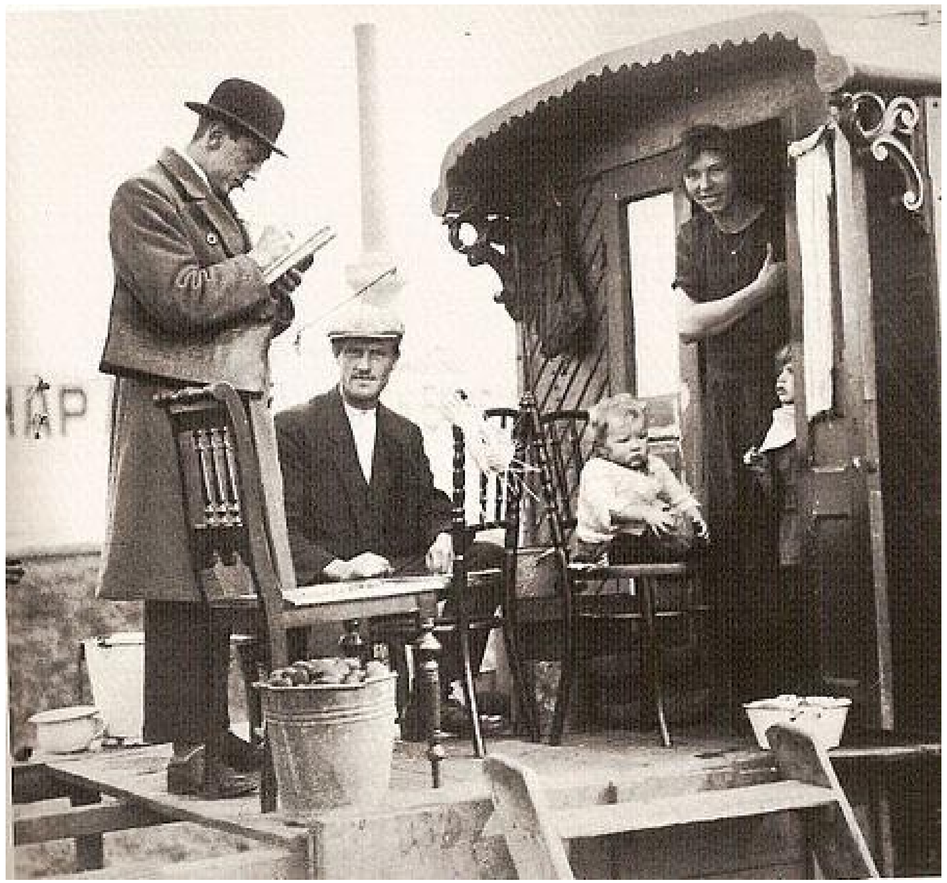
The Web as Data Source

Designed data

Frequently used research methods in social science as

- **Experiments (laboratory)**
- **Surveys**

Surveys



Census taker visits a Romani family living in a caravan, Netherlands 1925

https://en.wikipedia.org/wiki/Census

Surveys are a research method involving the use of **questionnaires** and **interviews** to gather information from individuals.

Questions are designed to measure a **theoretical construct** using a scale

Validity is assessed by experts and by reusing the same scale in different settings

Reliability of the scale is measured by inter-item correlations (Cronbach's Alpha), test-retest reliability

Some Limitations and Problems

- **Researcher Bias**
- **Non-response Bias**
  - Response rate of 10-20% are normal.
- **Sample Bias**
  - some subgroups (e.g. women) are more willing to respond
  - If you don’t know anything about your target population before, you cannot compensate the bias
  - Probabilistic (stratified) samples are not always possible.
- **Social Desirability Bias, Memory issues**
- **Obstrusive**

Found data in the social sciences

- **General types of found data:**
  - **Accretion** - a build-up of physical traces
    - Dust on a library books
  - **Erosion** - the wearing away of material
    - Wear on seats in lecture rooms

- **Unobstrusive**
  - **Observing people without them knowing**
  - Methods of studying social behaviour without affecting it

Unobtrusive Measures, Eugene J. Webb, Lee Sechrest und Donald T. Campbell Raymond Lee (2000). Unobtrusive Methods In Social Research, Buckingham: Open University Press.


Found data is everywhere

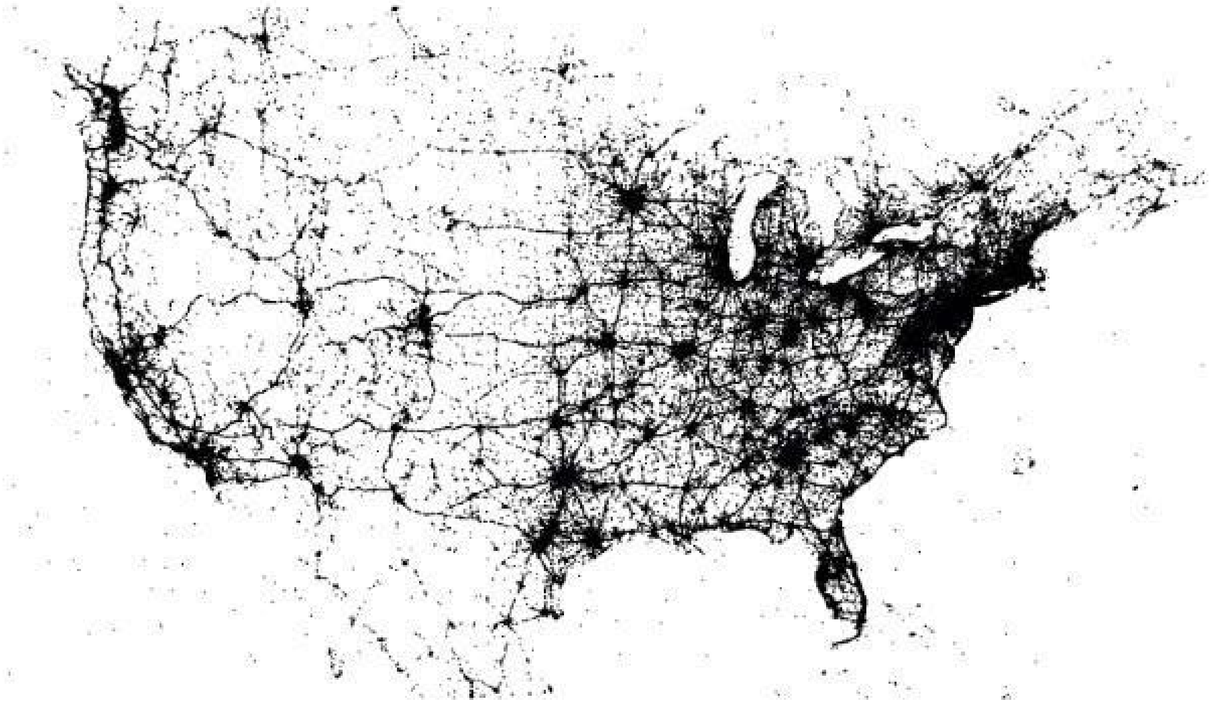
Smarter Devices

Ubiquitous Connectivity

skype

New kinds of data (macro)

Human mobility in societies

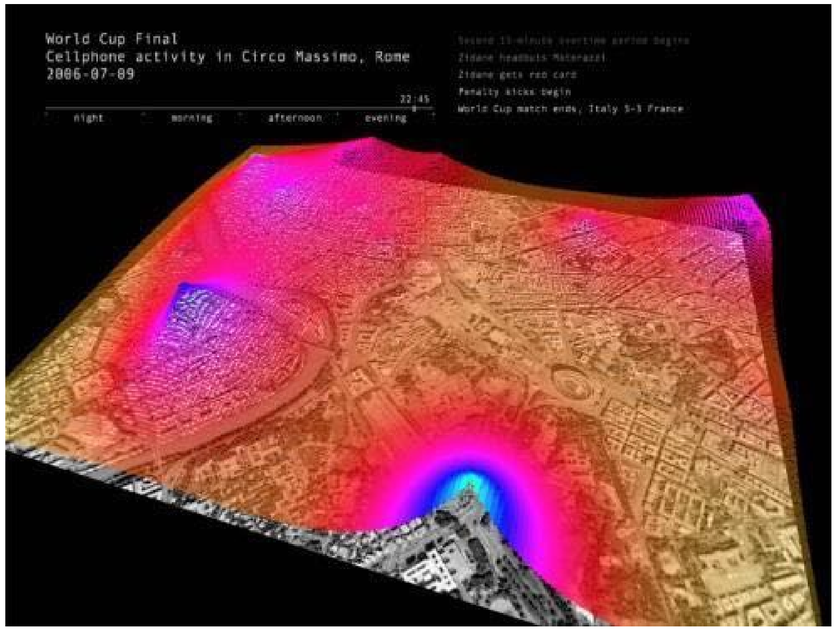


Check-ins (Foursquare, Gowalla, Twitter, ...)

Cheng, Zhiyuan, et al. "Exploring Millions of Footprints in Location Sharing Services." ICWSM 2011 (2011): 81-88.

New kinds of data (meso)

Urban movement analysis from GPS/phone data

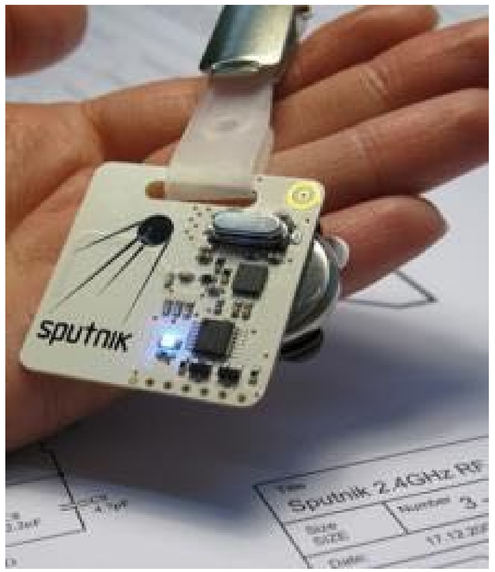


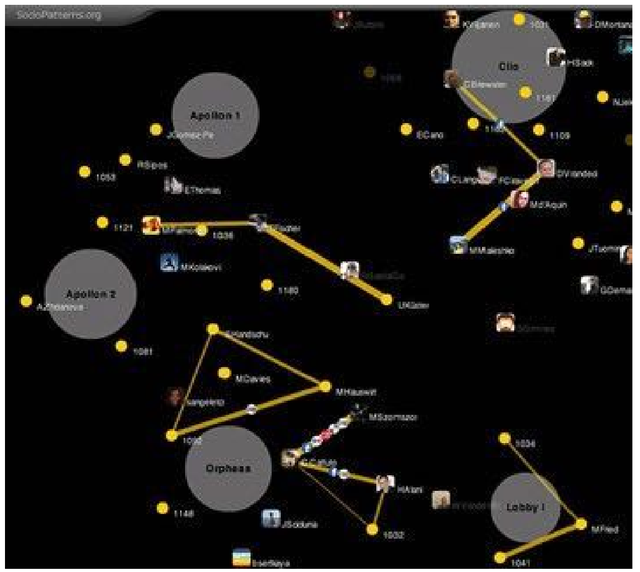
The Amsterdam Real Time Project

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome, IEEE Transactions on Intelligent Transportation Systems, 12(1), 141-151.

New kinds of data (micro)

Social Sensing via RFID





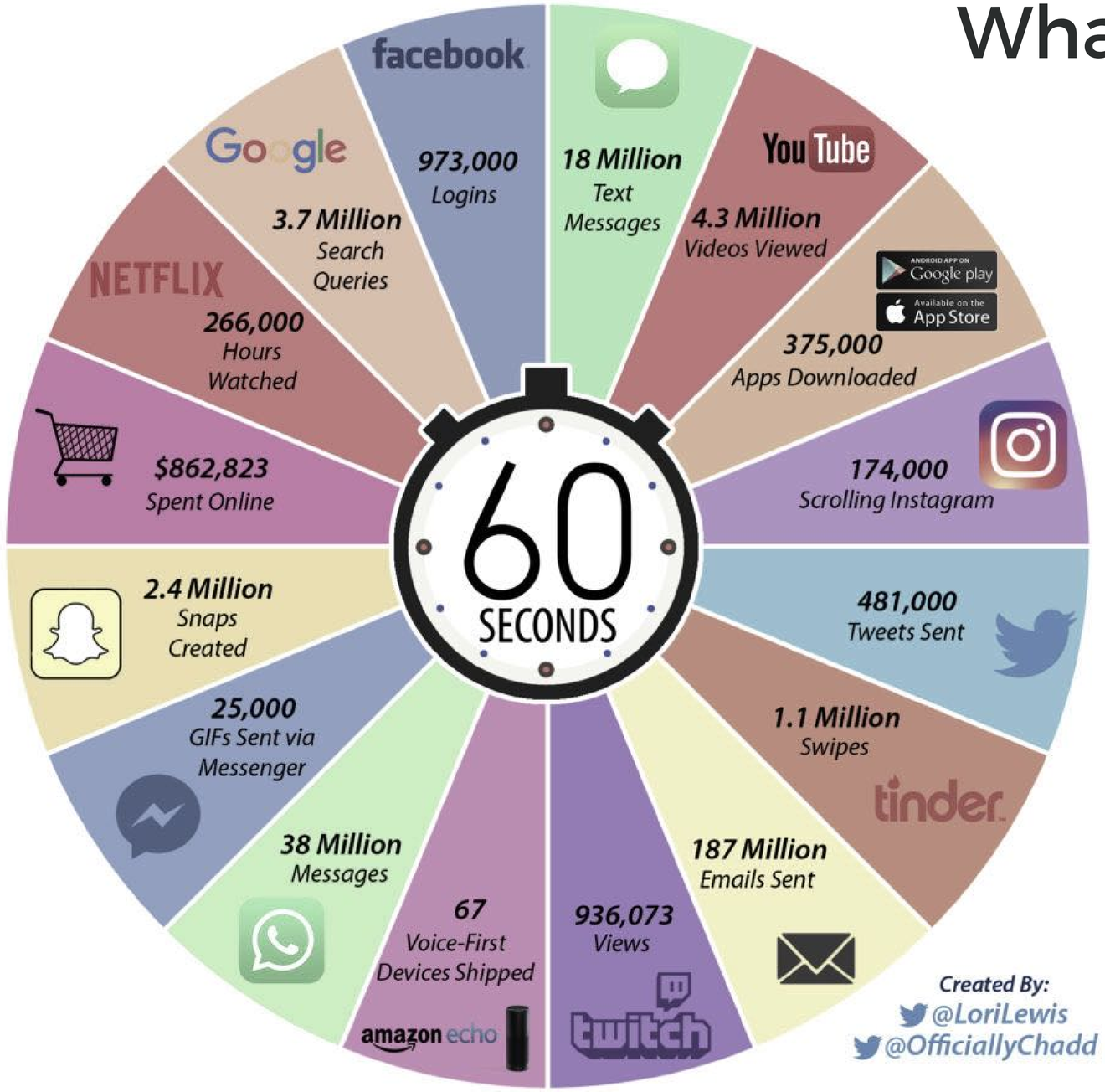
SocioPatterns

Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., & Vespignani, A. (2010). Dynamics ofperson-to-person interactions from distributed RFID sensor networks. PloS one, 5(7), e11596.

Most of this talk features examples using data from the **Web**.


SOCIAL
MEDIA
BLOG
SMS

What is happening in Internet in a minute?




flickr

(lon, lat, t)

WHERE

coordinates

WHEN

timestamp

WHAT

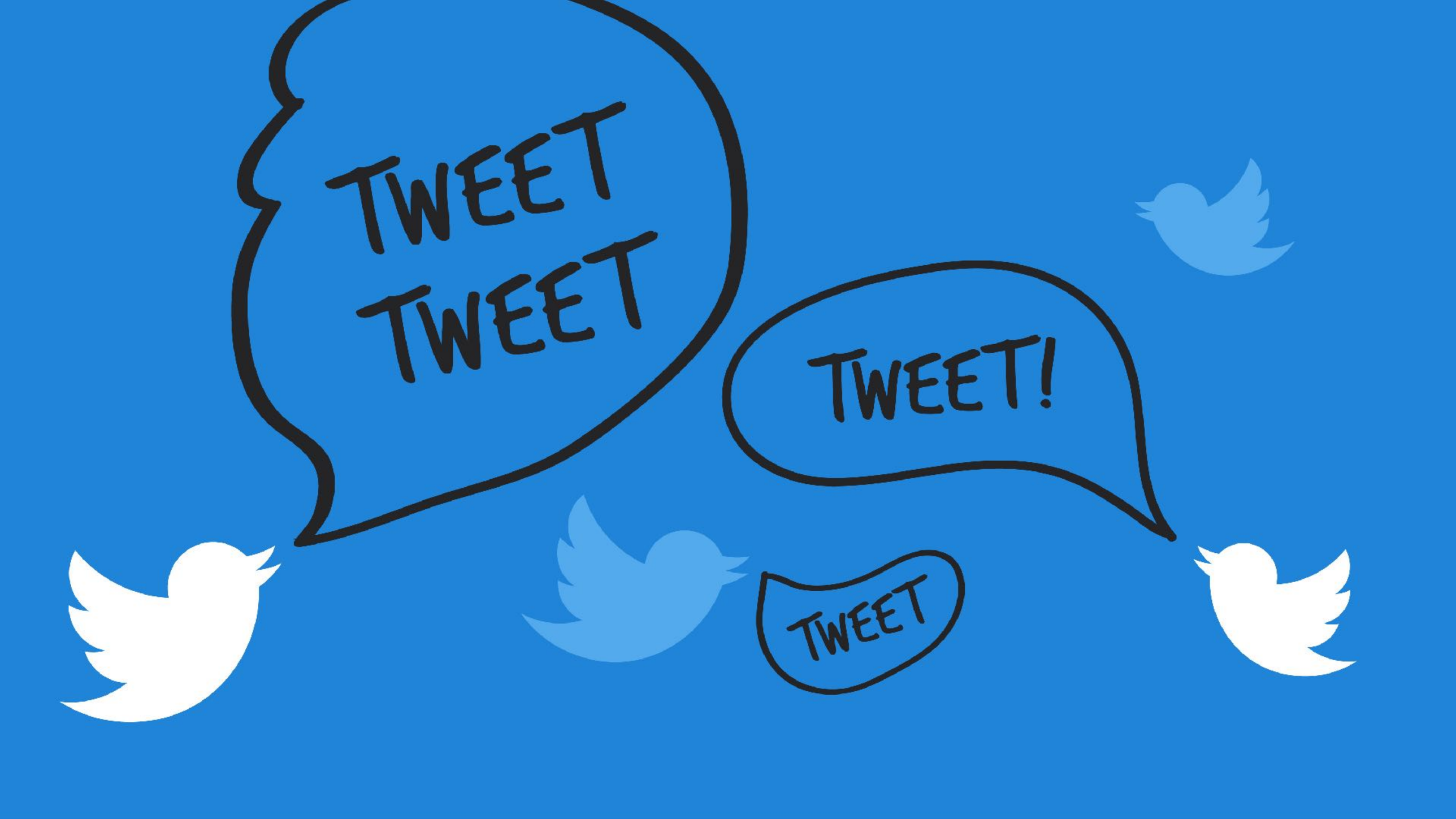
text+visual+audio

WHO

author

WHY

intent, context


TWEET
TWEET
TWEET!
TWEET

Observation

The digital world is tracking the social world more and more closely.

- This enables us to use **computation** to
  - **discover patterns**,
  - **build models**,
  - **validate social theories** and
  - **learn about societies.**

Why are Social Scientists excited about Computational Social Science?

- **Context of Justification**
  - New kinds of data enable to test social theories that we could not test previously
    - Verification, Falsification, Corroboration
- **Context of Discovery**
  - New kinds of data enable to come up with new social theories
    - Induction, Analogies, Abstraction

New types of found data

Potential Limitations

not representative
- population biases

poor in attributes
- unknown demographic attributes

dominated by a few
- power law phenomena
- self-selection

shaped by systems
- algorithmically mediated

noisy
- users != people

Potential Advantages

highly granular
- high temporal resolution
- high spatial resolution

rich in structure
- multi-relational data

rich in sources
- integration of different data types

cheap and big

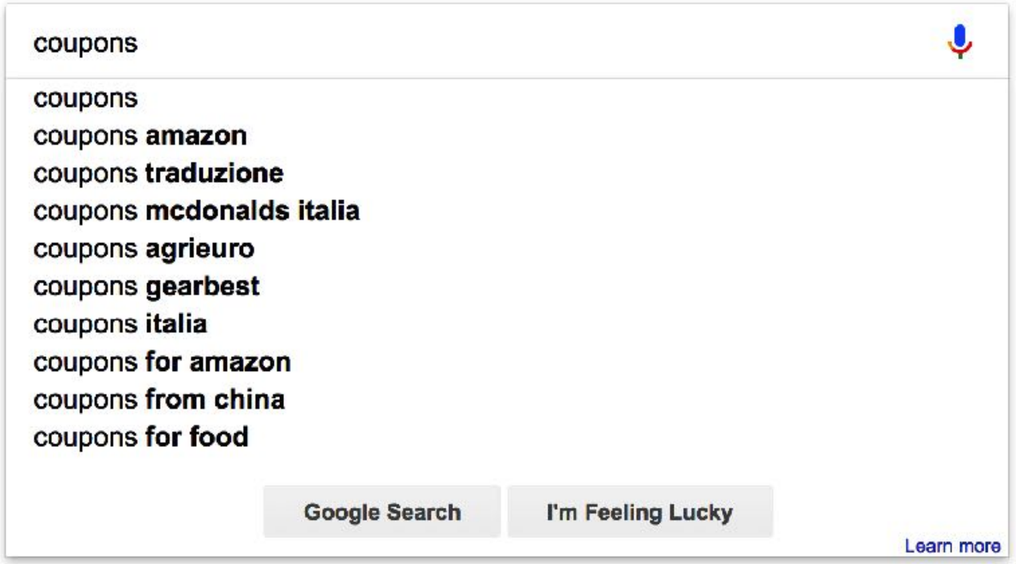
Algorithmic Bias

Two Sides of Mining Web Data

Google

Query suggestions to query logs:
Influences people's searches.

Web as Actor



Query logs to query suggestions:
Data on what people frequently search for.

Web as Data

Are algorithms "neutral"? Should they be neutral?
Popularity-driven algorithms => Tyranny of the majority?
How do algorithms change what we think, whom we befriend?

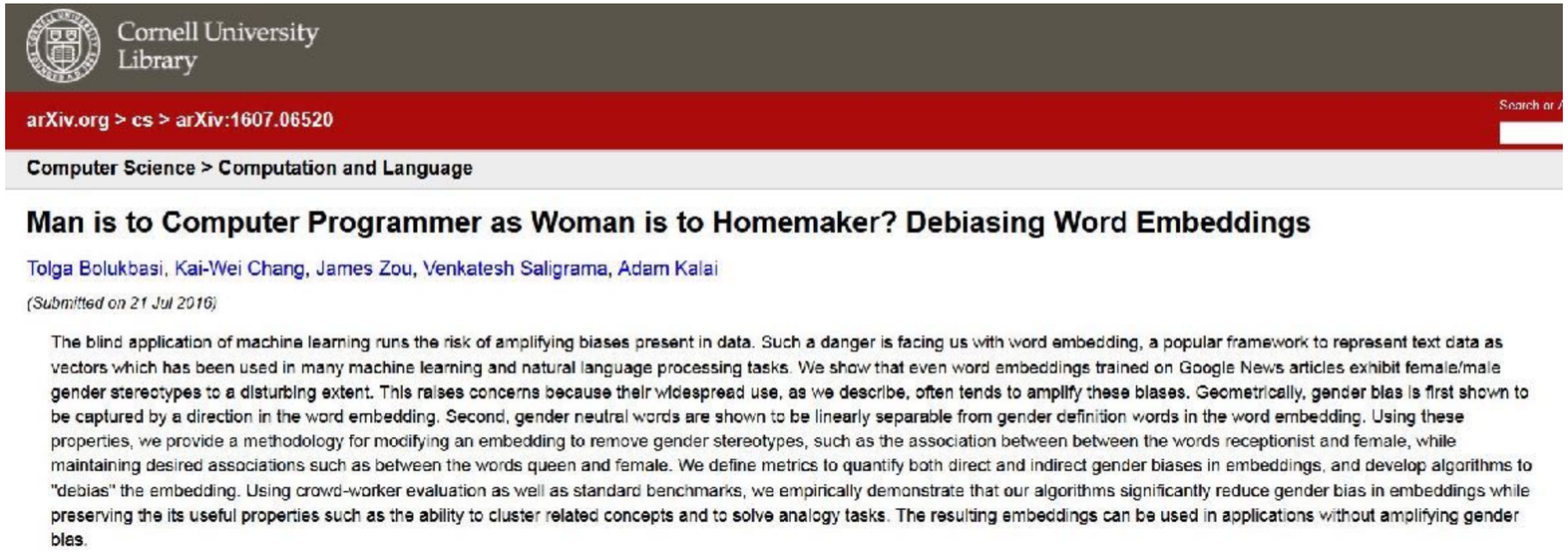
Word2vec and sexism

Man is to king as woman is to ... **queen.** ✓

Sister is to woman as brother is to ... **man.** ✓

Father is to doctor as mother is to ... **nurse.** ✗

Man is to programmer as woman is to ... **homemaker.** ✗



Cornell University Library

arXiv.org > cs > arXiv:1607.06520

Computer Science > Computation and Language

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

(Submitted on 21 Jul 2016)

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Stereotypes

Google image query: **Doctor**

Google image query: **Nurse**

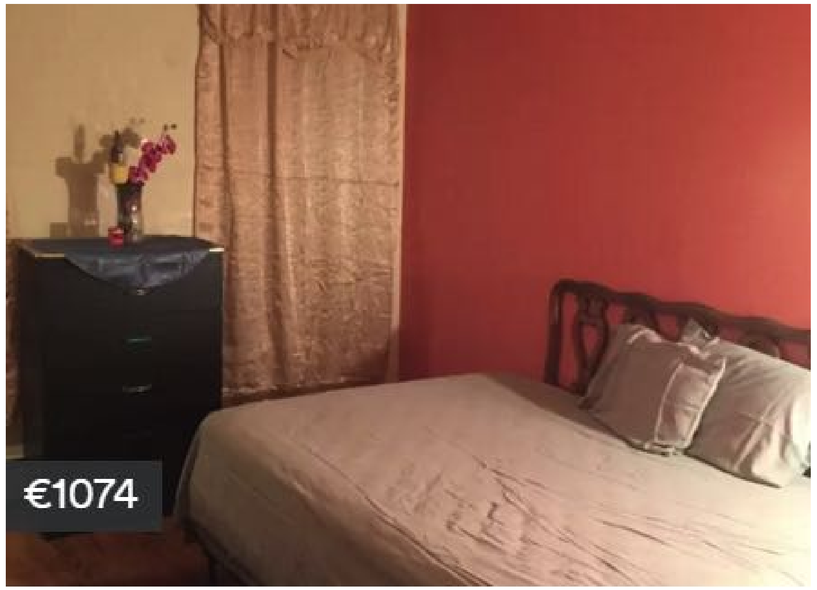
“evidence for **stereotype exaggeration** and systematic **underrepresentation of women**”

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015.

More examples in KDD Tutorial on Algorithmic Bias
http://francescobonchi.com/algorithmic_bias_tutorial.html#slides

Discrimination

“**non-black hosts are able to charge approximately 12% more than black hosts**, holding location, rental characteristics, and quality constant.“



Edelman, Benjamin G. and Luca, Michael, Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054. http://dx.doi.org/10.2139/ssrn.2377353

Racism

Photo app **tagging** black people **as “gorillas”**

https://twitter.com/jackyalcine/status/615331869266157568

Racism

https://twitter.com/geraldmellor/status/712880710328139776/photo/1

Political views

Former Facebook Workers: We Routinely Suppressed Conservative News

Michael Nunez

5/09/16 9:10am · Filed to: FACEBOOK

become a supporter | sign in | subscribe | search | dating | more | International

theguardian

UK world sport football opinion culture business lifestyle fashion environment tech travel | browse all sections

home › tech

Facebook

Facebook accused of censoring conservatives, report says

Curators of the social media company's 'trending news' sidebar purposely leave out stories from rightwing sites, a former employee has alleged

Nellie Bowles and Sam Thielman in New York

Monday 9 May 2016 17.10 BST

This article is 3 months old

Shares 2,967 | Comments 428

Save for later

Facebook has become one of the most important distributors of news online, notably through its 'trending news' sidebar. Photograph: Kennedy Photography/Alamy

Most popular in US

Have we detected an alien megastructure in space? Keep an open mind | Seth Shostak

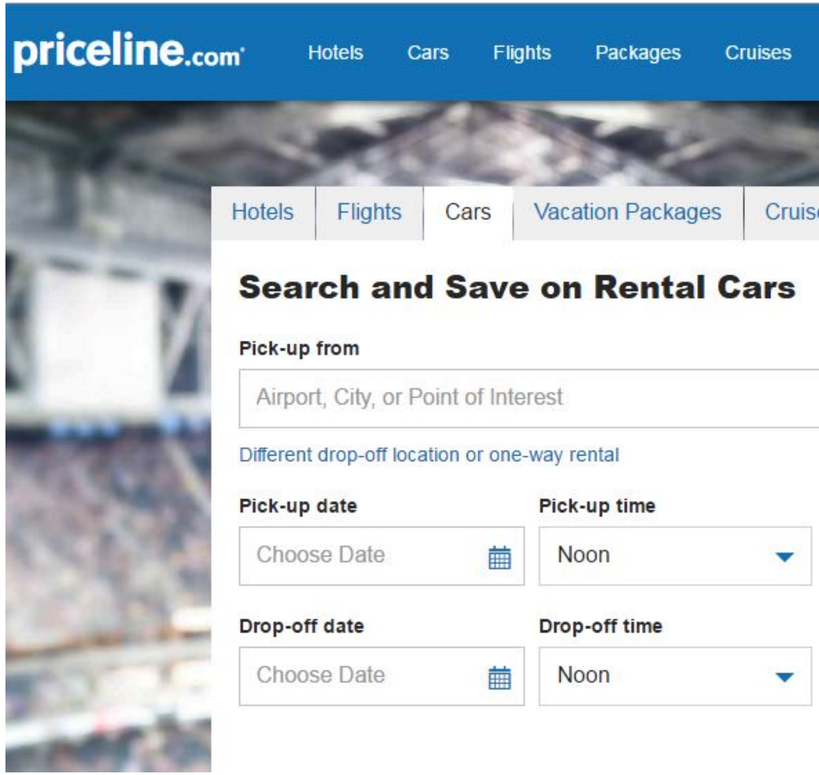
Donald Trump claims 'cheating' is only way he can lose Pennsylvania

Michael Phelps taught a lesson for once by Joseph Schooling | Andy Bull

Hope Solo calls Sweden 'a bunch of cowards' after USA falter at Olympics

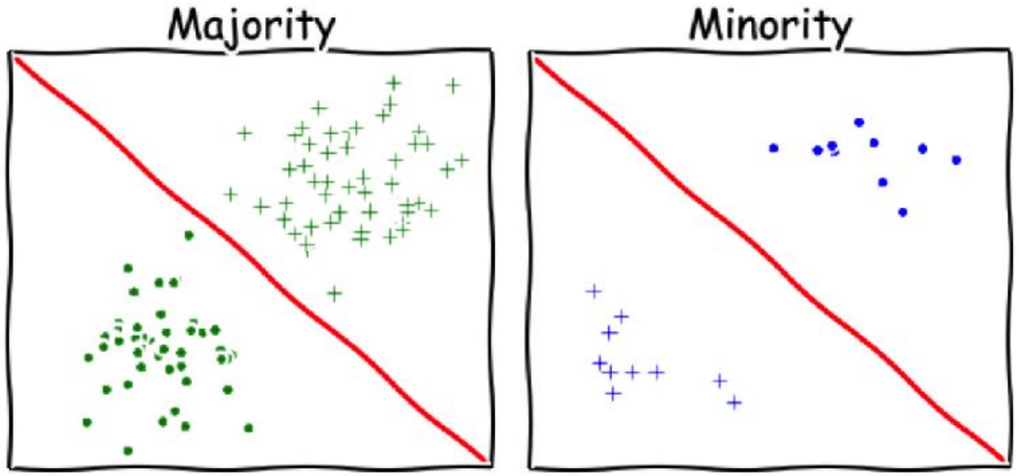
Price discrimination

we find **evidence for price steering and price discrimination** on four general retailers and five travel sites.



Measuring Price Discrimination and Steering on E-commerce Web Sites, Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson, In Proceedings of the 14th ACM/USENIX Internet Measurement Conference (IMC'14)
http://francescobonchi.com/algorithmic_bias_tutorial.html#slides

Some Reasons for Algorithmic Bias



Positively labeled examples are on opposite sides of the classifier for the two groups.

Data tyranny of
the majority class

Data selection bias
- Only cars, no bicycle trips
- Track smartphone users, not others

Algorithmic issues
- Recommender systems that narrow, instead of broaden
- Not compensating for selection bias

https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de
http://francescobonchi.com/KDD2016_Tutorial_Part1&2_web.pdf
https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Network Definition and Basic Measures

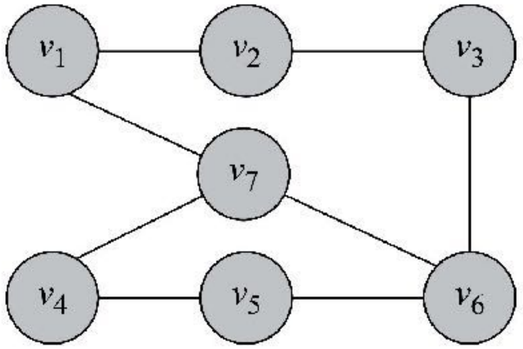
Basic Definitions

- A network is a **graph,** or a collection of points connected by lines
- Points are referred to as **nodes**, actors, or vertices (plural of vertex)
- Connections are referred to as **edges** or ties
- More formally:

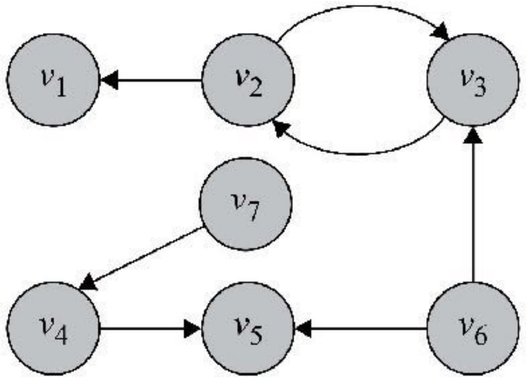
$$G = \{V, E\}$$

$$V = \{v_1, v_2, \ldots, v_n\} \quad \text{Nodes set}$$

$$E = \{e_1, e_2, \ldots, e_m\} \quad \text{Edges set}$$



undirected



directed

Basic Definitions

neighborhood

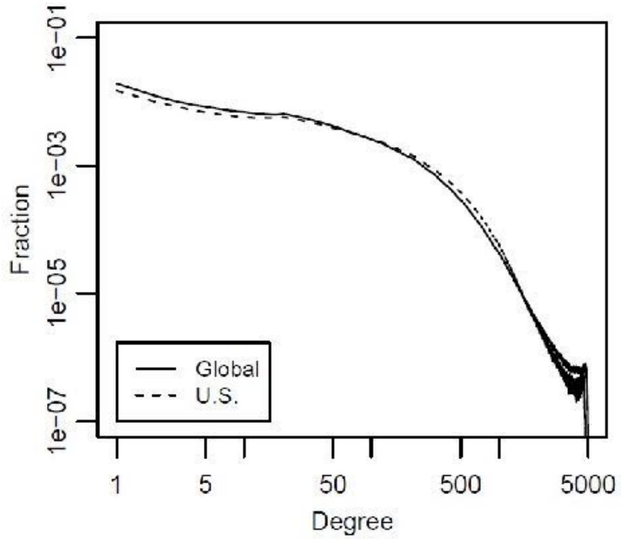
set of connected nodes

degree

number of edges connected to one node

degree distribution

probability distribution of
nodes degrees over the whole network



density

portion of the potential connections in a network that are actual connections

reciprocity

measure of the likelihood of vertices in a directed network to be mutually linked

networkx.classes.function

degree(G[, nbunch, weight])

Return a degree view of single node or of nbunch of nodes.

degree_histogram(G)

Return a list of the frequency of each degree value.

density(G)

Return the density of a graph.

info(G[, n])

Print short summary of information for the graph G or the node n.

networkx.algorithms.reciprocity

reciprocity(G[, nodes])

Compute the reciprocity in a directed graph.

Basic Definitions

walk

sequence of incident edges visited one after another

random walk

a walk that in each step the next node is selected randomly among the neighbors

path

a walk where nodes and edges are distinct

shortest path

path between two nodes that has the shortest length

average shortest path length

average number of steps along the shortest paths for all possible pairs of network nodes

networkx.algorithms.shortest_paths

shortest_path(G[, source, target, weight])

Compute shortest paths in the graph.

all_shortest_paths(G, source, target[, weight])

Compute all shortest paths in the graph.

shortest_path_length(G[, source, target, weight])

Compute shortest path lengths in the graph.

average_shortest_path_length(G[, weight])

Return the average shortest path length.

has_path(G, source, target)

Return True if G has a path from source to target.

Basic Definitions

eccentricity (v)

the length of the maximal shortest path from v to all other nodes

diameter

maximum length of all shortest paths in G (maximal eccentricity in G)

radius

minimum eccentricity in G

periphery

set of nodes with eccentricity equal to the diameter

center

the set of nodes with eccentricity equal to radius

networkx.algorithms.distance_measures

center(G[, e, usebounds])

Return the center of the graph G.

diameter(G[, e, usebounds])

Return the diameter of the graph G.

eccentricity(G[, v, sp])

Return the eccentricity of nodes in G.

periphery(G[, e, usebounds])

Return the periphery of the graph G.

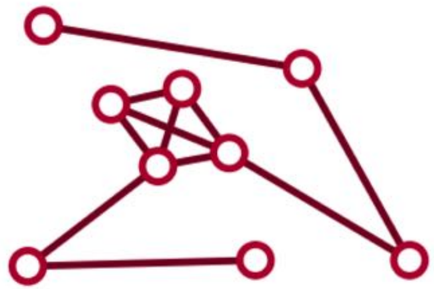
radius(G[, e, usebounds])

Return the radius of the graph G.

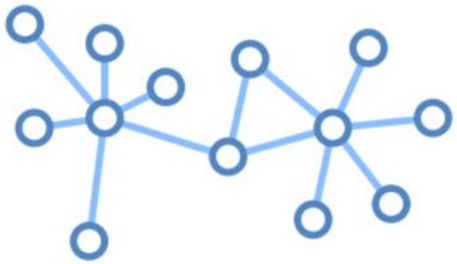
Assortative Mixing (Newman 2003)

- **Preference for a network's nodes to attach to others that are similar in some way**
  - degree - preferential attachment model
  - node properties like gender, race, age

- Nodes of a certain type tend to associate with the same type of nodes (this is also called **homophily**)
- usually in the range [-1, 1]
  - 1 (assortative) and -1 (disassortative)



assortative (r = 0.36)



disassortative (r = -0.91)

networkx.algorithms.assortativity

- **degree_assortativity_coefficient(G[, x, y, ...])**
  - Compute degree assortativity of graph.
- **attribute_assortativity_coefficient(G, attribute)**
  - Compute assortativity for node attributes.
- **numeric_assortativity_coefficient(G, attribute)**
  - Compute assortativity for numerical node attributes.
- **degree_pearson_correlation_coefficient(G[, ...])**
  - Compute degree assortativity of graph.

Clustering coefficient (Watts, Strogatz 1998)

Measure of the degree to which nodes in a graph tend to cluster together.

LOCAL

$$C_i = \frac{number\ of\ connections\ between\ i's\ neighbors}{maximum\ number\ of\ possible\ connections\ between\ i's\ neighbors}$$

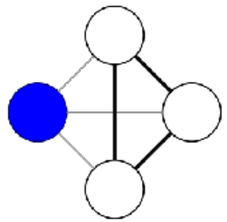
GLOBAL

TRANSITIVITY

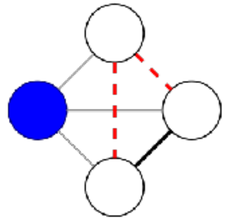
$$C = \frac{number\ of\ closet\ triplets}{number\ of\ all\ triplets} = \frac{3 * number\ of\ triangles}{number\ of\ all\ triplets}$$

AVERAGE CLUSTERING COEFFICIENT

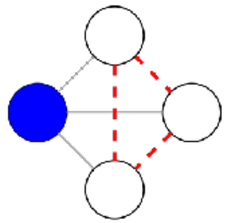
$$C = \frac{\sum_i C_i}{N}$$



c = 1



c = 1/3



c = 0

networkx.algorithms.cluster

- **triangles(G[, nodes])**
  - Compute the number of triangles.
- **transitivity(G)**
  - Compute graph transitivity, the fraction of all possible triangles
- **clustering(G[, nodes, weight])**
  - Compute the clustering coefficient for nodes.
- **average_clustering(G[, nodes, weight, ...])**
  - Compute the average clustering coefficient for the graph G.

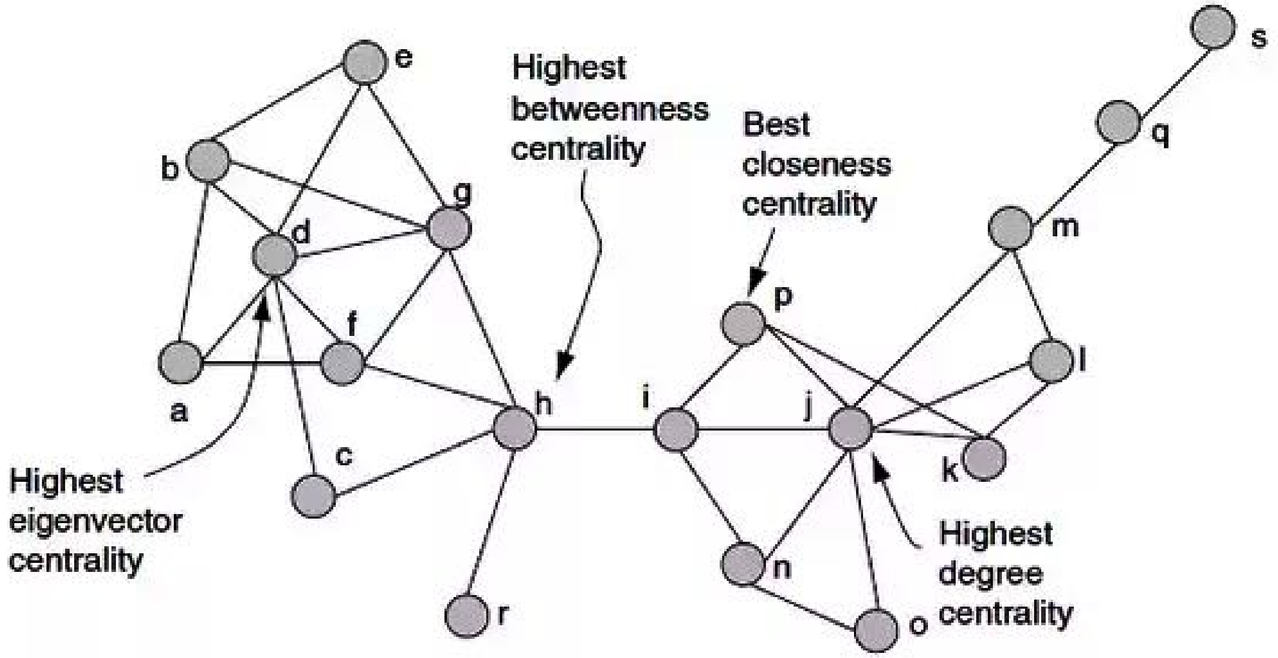
Centrality

- **Identify the most important vertices within a graph**
  - e.g., most influential person(s) in a social network
  - key infrastructure nodes in the Internet or urban networks
  - super-spreaders of disease
- Different interpretations that lead to different measures:
  - **degree**
  - **closeness**
  - **betweenness**
  - **eigenvector**
  - **Page Rank**
  - **Katz**

Centrality

- **Degree**
  - how connected is a node?
- **Closeness**
  - how easy can a node reach other nodes?
- **Betweeness**
  - how much does a node function as a connector, as a bridge for other nodes?
- **Influence, Prestige, Eigenvector**
  - how important are your friends? (Page Rank, HITS)

Centrality


e
b
g
d
f
a
c
h
r
i
p
j
k
n
o
m
q
s
l
Highest
betweenness
centrality
Best
closeness
centrality
Highest
eigenvector
centrality
Highest
degree
centrality

networkx.algorithms.centrality

Degree

degree_centrality(G)
Compute the degree centrality for nodes.

Betweenness

betweenness_centrality(G[, k, normalized, ...])
Compute the shortest-path betweenness centrality for nodes.

edge_betweenness_centrality(G[, k, ...])
Compute betweenness centrality for edges.

betweenness_centrality_subset(G, sources, ...)
Compute betweenness centrality for a subset of nodes.

edge_betweenness_centrality_subset(G, ...[, ...])
Compute betweenness centrality for edges for a subset of nodes.

Closeness

closeness_centrality(G[, u, distance, ...])
Compute closeness centrality for nodes.

Eigenvector

eigenvector_centrality(G[, max_iter, tol, ...])
Compute the eigenvector centrality for the graph G.

katz_centrality(G[, alpha, beta, max_iter, ...])
Compute the Katz centrality for the nodes of the graph G.

networkx.algorithms.link_analysis

pagerank(G[, alpha, personalization, ...])

Return the PageRank of the nodes in the graph.

hits(G[, max_iter, tol, nstart, normalized])

Return HITS hubs and authorities values for nodes.

Social Theories

The Need for Theory

- **Network analysis can be a cool toy.**
- It is easy to get lost in data crunching and forget about **why we care about networks.**
- You must develop a **theory** of why and how networks matter in your case.
- **What are the mechanisms at work?**
  - If larger degrees matter, why is that the case? If centrality helps, why is that the case? If centrality hurts, why is that the case?

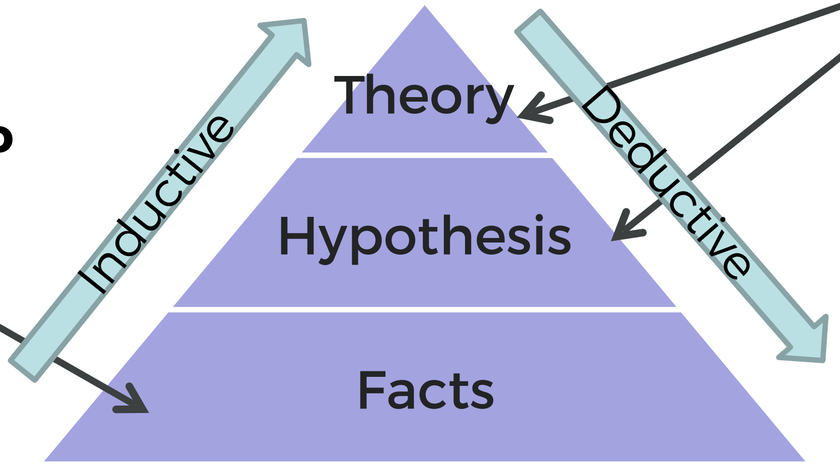
Theory and Hypothesis

Hypothesis: a testable explanation of a fact

Theory: well-substantiated explanation of a phenomenon acquired through the scientific method and repeatedly tested and confirmed through observation and experimentation. A theory can be used to make predictions and explain the why

The gap between the two can vary quite much

Facts of social life can be difficult to measure. Hard to make reliable and valid measurements



Knowledge

Information

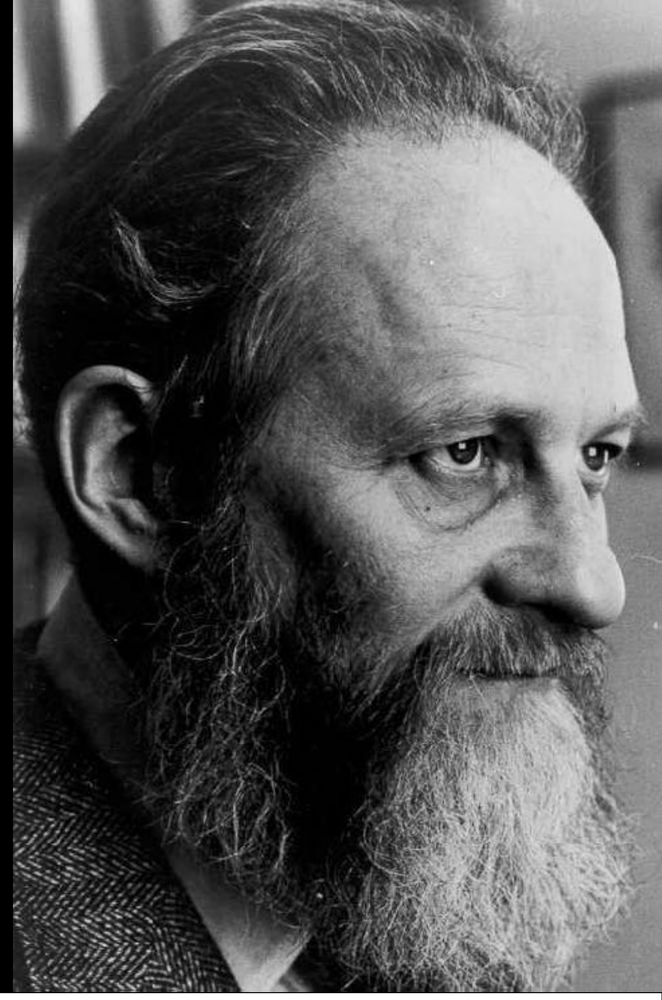
Data

Abundance of Sociological Theories

- Within the domain of Organizational Performance, proliferation of theories over past 100 years
  - Miner (1984) identified **110 distinct theories**
  - Lewin and Minton (1980) i**dentified 13 distinct bodies of theory**
  - No identifiable, stable, generalizable principles
  - Not even agreement on definition of performance
- Question: “What does social science have to say about X?”
- Answer: “Well, hundreds of things!”
- “the field [of organization science] still appears paradigmatically fragmented, ambiguous, and perhaps even irrelevant to those outside academia.” (Schwartz et al. 2007)

The Future of the Social Sciences: Part 3 by Duncan Watts, “Big Questions for 21st Century Social Science”
https://www.youtube.com/watch?v=aXujlQt0das

Why study human interactions?



“Most human pleasures have their roots in social life. Much of human suffering as well as much of human happiness has its source in the actions of other human beings”

Peter Blau

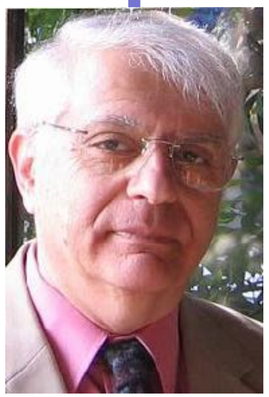
Exchange and Power in Social Life, 1964

The nature of social structures

Small-world hypothesis

Stanley Milgram

Tie strength and selection

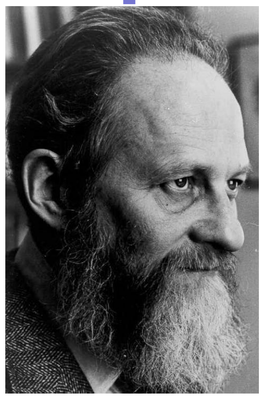


Mark Granovetter

Social brain theory

Robin Dunbar

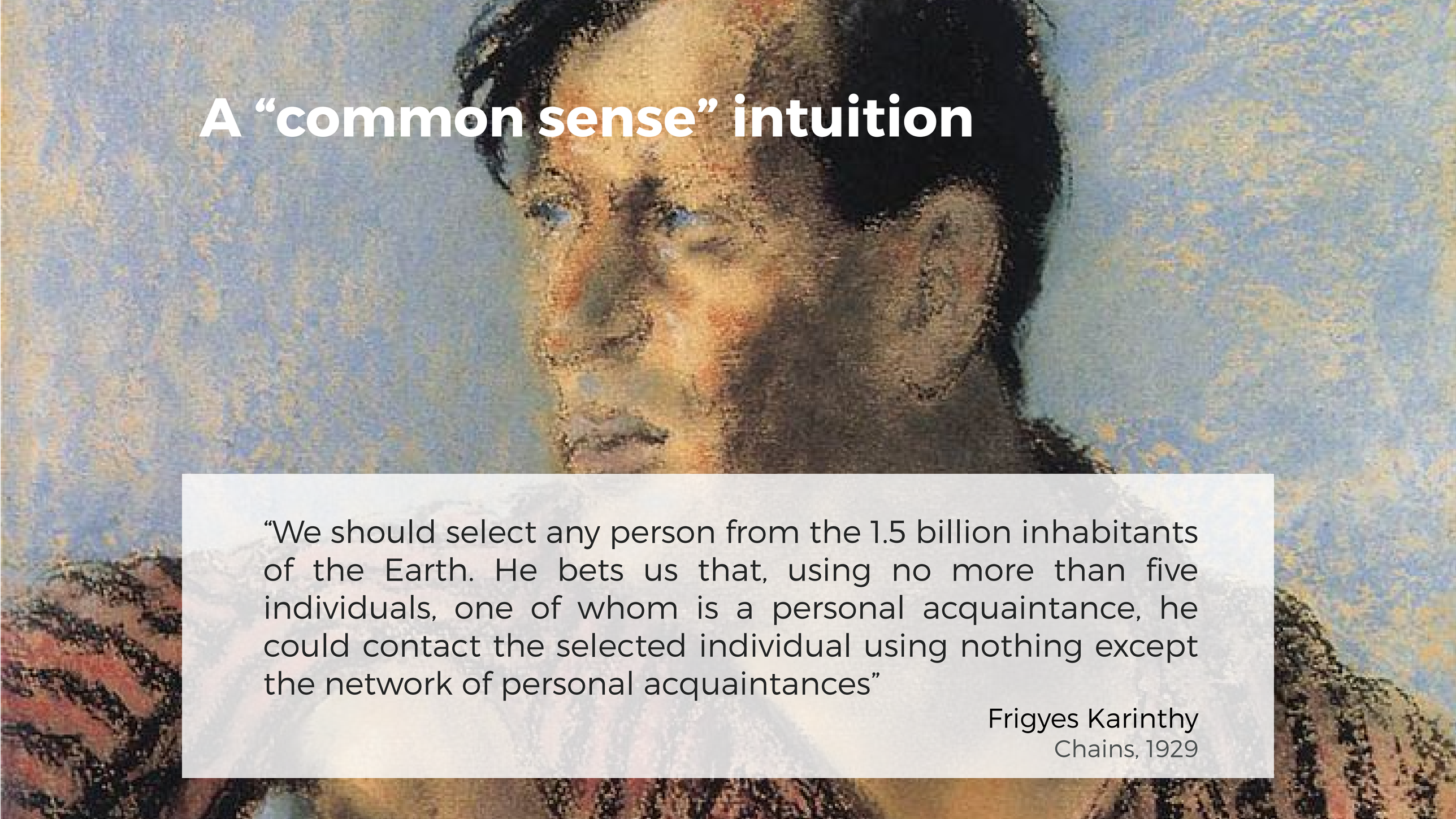
Social exchange



Peter Blau

Small World Hypothesis

A “common sense” intuition

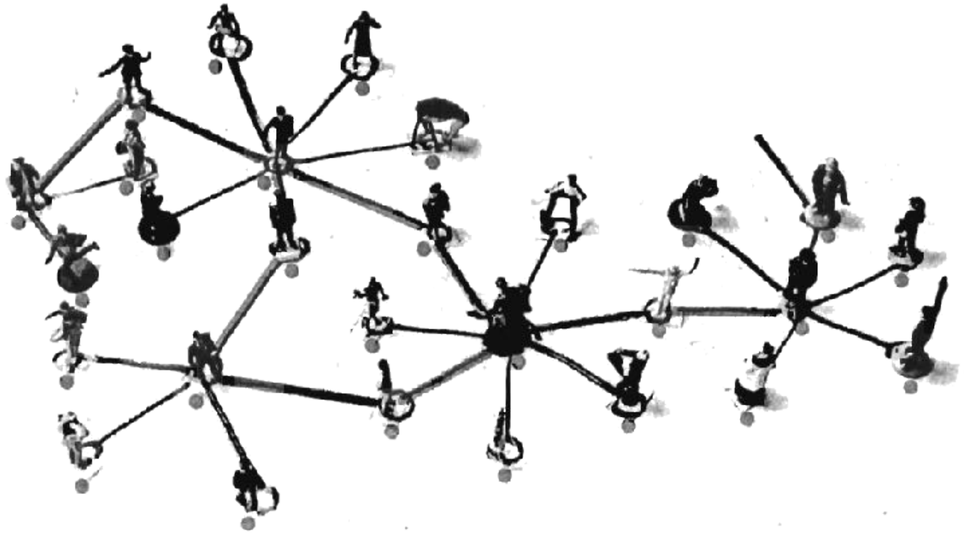


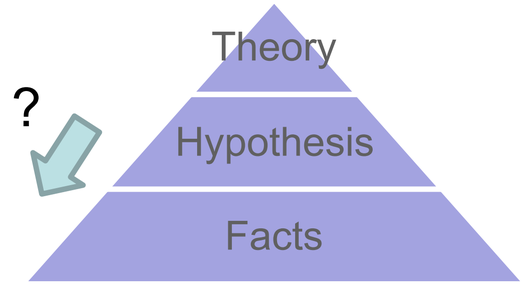
“We should select any person from the 1.5 billion inhabitants of the Earth. He bets us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances”

Frigyes Karinthy
Chains, 1929

The small-world problem

- Formulated by social psychologist **Stanley Milgram**
- Given any two people in the world (X and Y), how many intermediate acquaintance links are needed before X and Y are connected?





Jeffrey & Milgram – An Experimental study of The Small World Problem, 1969

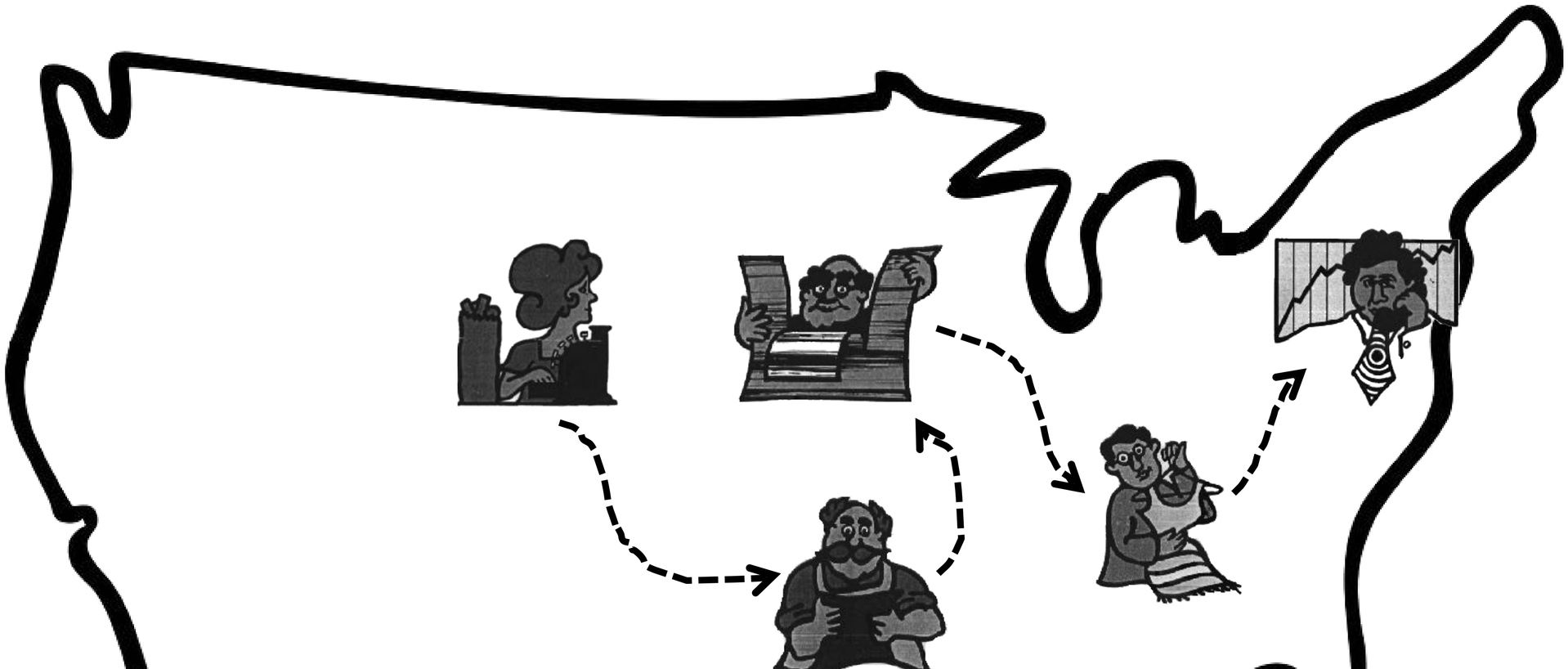
Milgram’s small-world experiment

- **Algorithmic formulation of the problem**
  - Attempt to generate an acquaintance chain from a start person to a target person

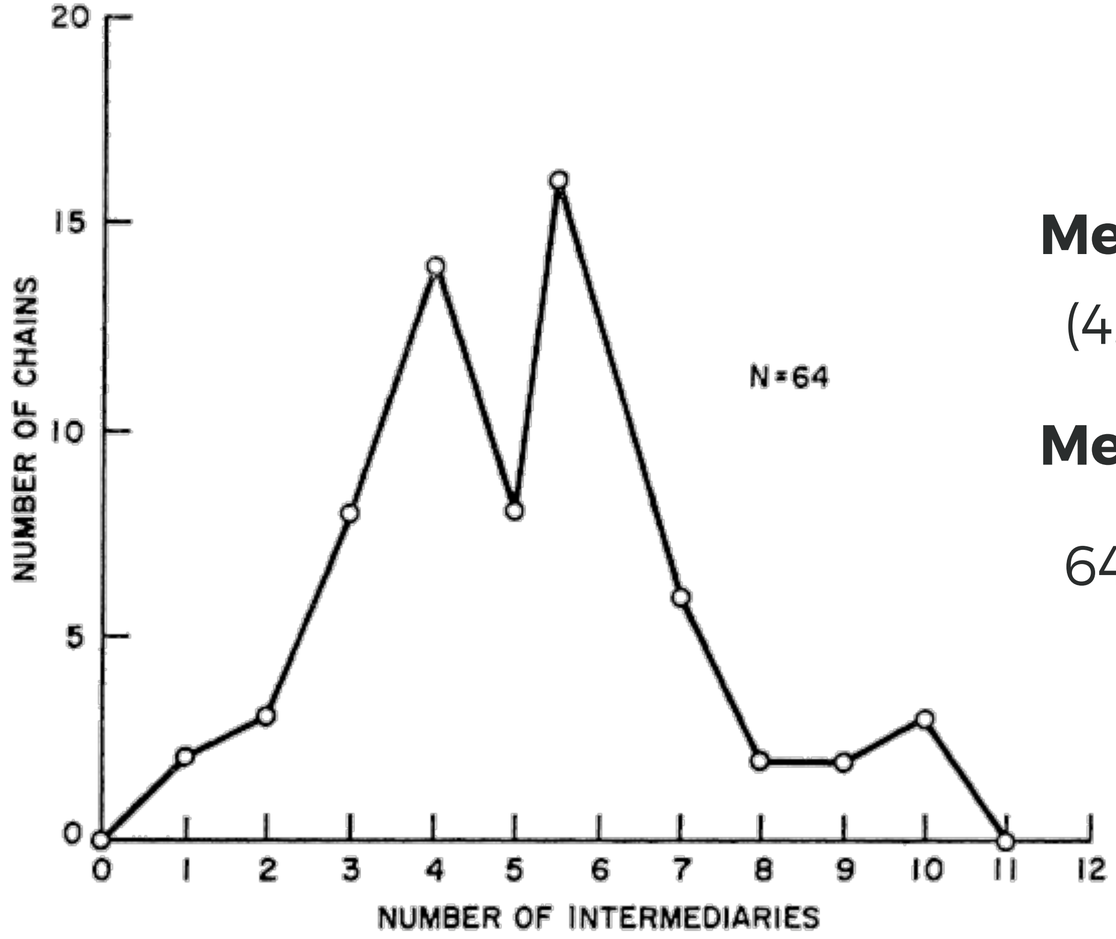
- **A document is marked with information about the target** (name, address, occupation, company, hometown , etc.)
- **The document is delivered to the start person**
- The person with the document must:
  - **Mark it with their name**
  - **Deliver it to target if it’s a personal acquaintance**
  - **Deliver to a personal acquaintance who might have a better chance to know the target**
  - **Send a special tracer card to Harvard (for chain tracking)**

Experimental setup

- 1 target person (a stock broker in Boston, MA)
- 296 start people
  - 100 stockholders in Nebraska
  - 96 random in Nebraska
  - 100 random in Boston



Chain length



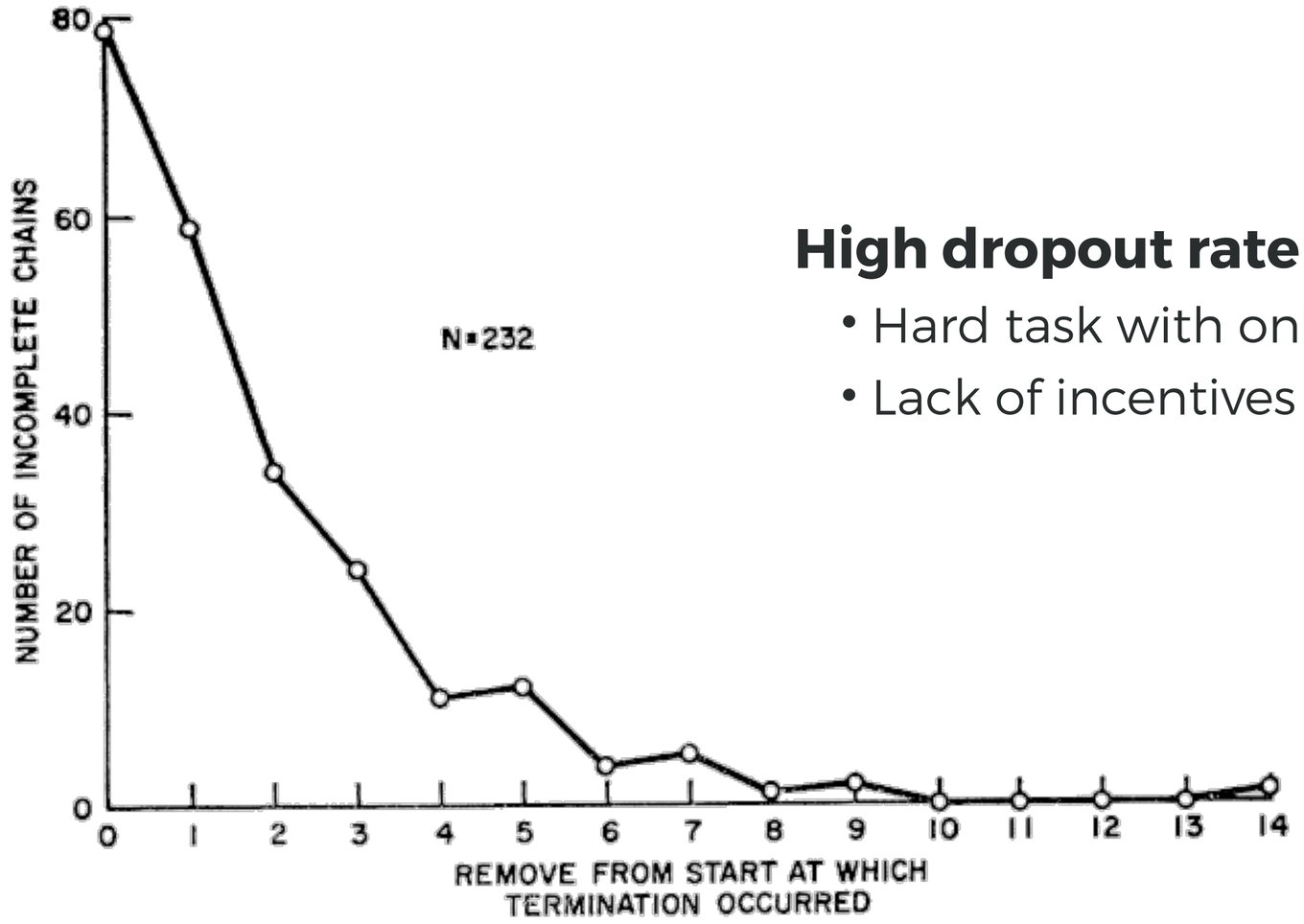
Mean number of intermediaries = 5.2

(4.4 Boston, 6.1 Nebraska)

Mean number of hops = 5 to 7

64 chains reached the target

Length of incomplete chains



High dropout rate

- Hard task with only local knowledge
- Lack of incentives

Reliability and validity

- **Ecological validity [“real life”]**
  - The experiment well approximates a real-world situation (for the 60s)
  - Participants were not incentivized
- **Reliability**
  - Clear algorithmic rules, but…
  - 80% of chains are discarded, longer chains are under-represented: more likely that they will encounter an unwilling participant.
- **External validity [other settings]**
  - Only 1 “special” target person (high social status)
  - Start people recruited with advertisement seeking for “well-connected” individuals

- Less than 100 chains in US only cannot generalize

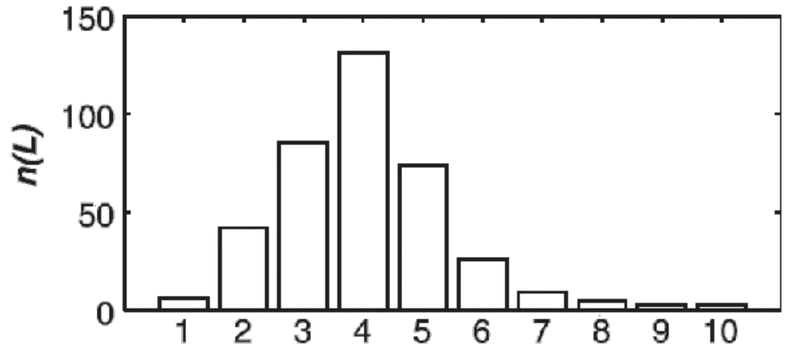
From snail mail to email

- **Similar setup, using email, between 2001 and 2007**
- **Experiment 1:**
  - 18 targets in 13 countries
  - 98,865 start people from 168 countries
  - 106,295 chains
- **Experiment 2:**
  - 21 targets in 13 countries
  - 85,621 start people from 163 countries
  - 56,033 chains

Goel et al. – Social Search in “Small World” Experiments, 2009

Results

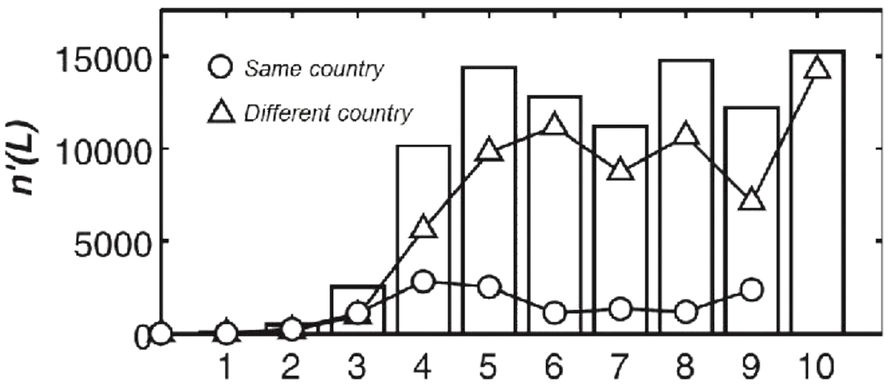
- Less than 1% of chains reached target, dropout rate exponential with length
- 4.05 steps on average
- Dropout for lack of incentive



- Estimating length in a no-attrition scenario

$$n'(L) = \frac{n(L)}{\prod_{i=0}^{L-1}(1 - r_i)}$$

- Median number of steps = 7



"if individuals searching for targets do not have sufficient incentives to proceed, the small-world hypothesis **does not hold**"

Algorithmic vs. Topological definition

- Milgram explored the **algorithmic definition** of the small-world problem
  - Social search: shortest path that ordinary people can find given their **local topological information** about the underlying social graph
- **Large-scale web social data allowed for an exploration of the topological definition** of the same problem
  - Shortest possible path between two individuals, with a **global knowledge** of the social network
- See also “How small is the world, really?” by D. J. Watts
  - bit.ly/1oRZvIX

J. Kleinberg. The small-world phenomenon: An algorithmic perspective, 2000

Topological small-world (MSN)

- **Availability of large-scale online social network data made direct topological measurement possible**
- MSN communication graph with 180M nodes and 1.3B undirected edges
- **Topological distance between 1000 random nodes and all others**. Measure number of hops
  - Mode = 6
  - Median = 7
  - Average = 6.6
- https://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0939v1.pdf

Leskovec & Horvitz – Planetary-scale view on instant messaging network, 2007

Topological small-world (Facebook)

- **Facebook data 721M nodes 69B edges**
- Compute distance between **all the pairs**
- Need a very efficient way to do it (HyperANF): http://webgraph.dsi.unimi.it
  - **Average of 4.74 hops**
- Recently repeated on **1.6B nodes**
  - **Average of 4.57 hops**
  - https://research.fb.com/three-and-a-half-degrees-of-separation

Backstrom et al. – Four degrees of separation, 2012

Open questions

- **Reliability**
  - Are all online social network ties good proxies for acquaintance?
- **External validity**
  - Does the result generalize across platforms?
  - Is the average path length shrinking in time?
- **Construct validity**
  - Is this results in line with expectations? What should we compare it to?
  - Are we measuring “how close the human population” is?
- **Ecological validity**
  - Is the actual graph distance meaningful for any social process?

Small Worlds

- We can characterize a network through 2 statistics:
- **Average shortest path length L**
  - this is not the same as the diameter of the graph, which is the maximum shortest path connecting any two nodes
- **The clustering coefficient C**

- A **small world graph** is any graph with a **relatively small L** and a **relatively large C**

1. $L \propto log\ N \approx L_{random}$
2. $C \gg C_{random}$

Found in many real-world phenomena

websites with navigation menus, food webs, electric power grids, metabolite processing networks, networks of brain neurons, voter networks, telephone call graphs, and social influence networks, cultural networks and word co-occurrence

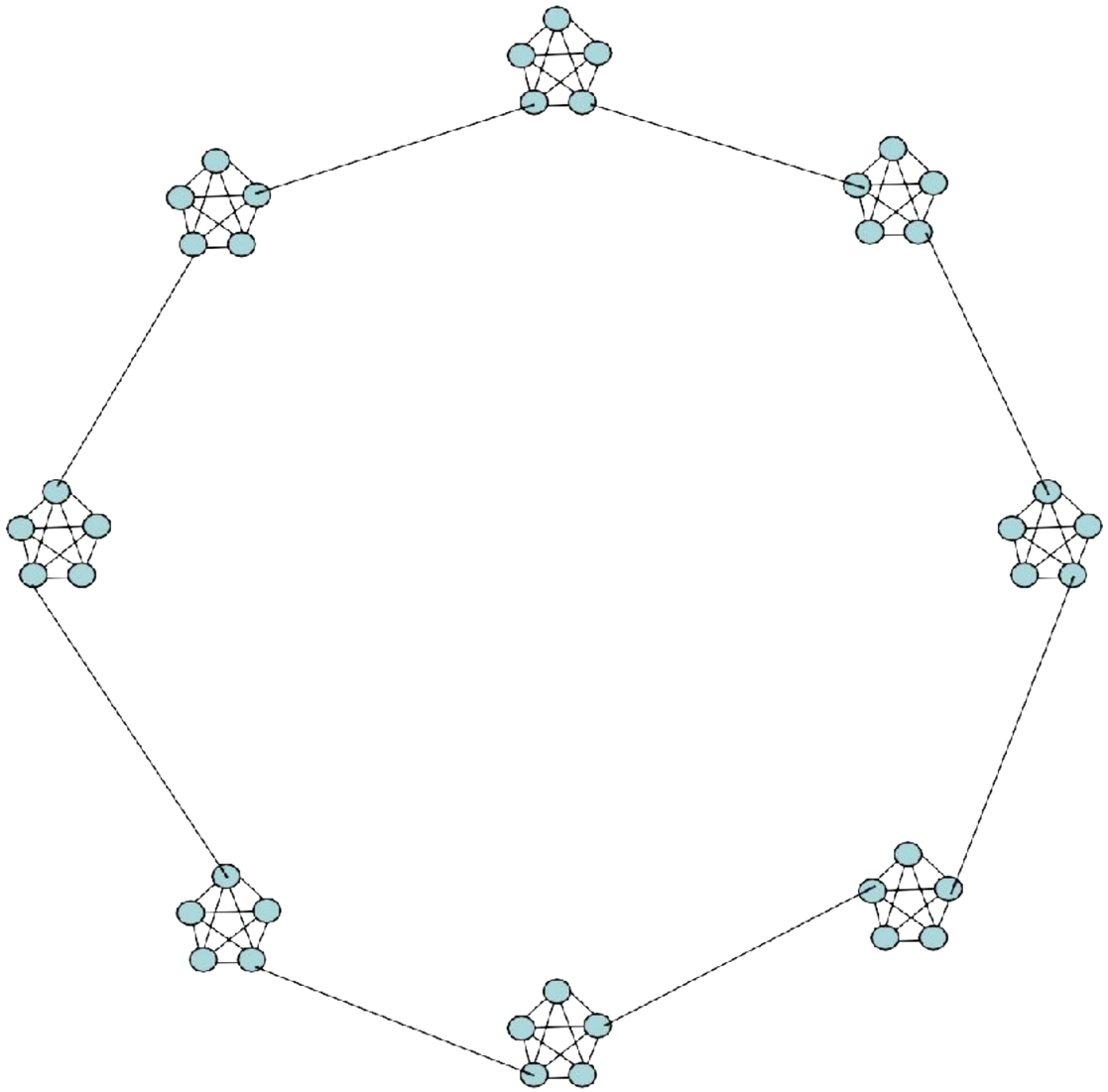
Example

- Everyone in a cave knows each other
- A few people make connections

- **Are C and L high or low?**

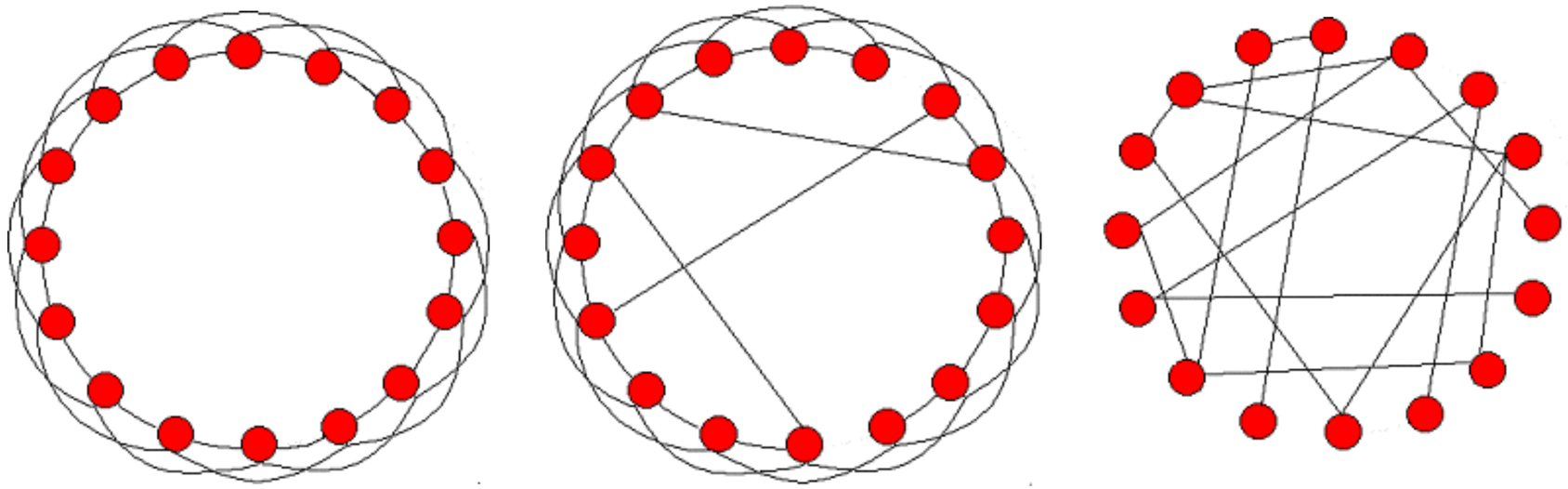
C high

L high



Watts and Strogatz model

- Start with a ring, where **every node is connected to the next z nodes (a regular lattice)**
- **With probability α, rewire every edge (or, add a shortcut) to a uniformly chosen destination**



Three Types of Networks

In a highly clustered, ordered network, **a single random connection will create a shortcut that lowers L dramatically**

Lattice

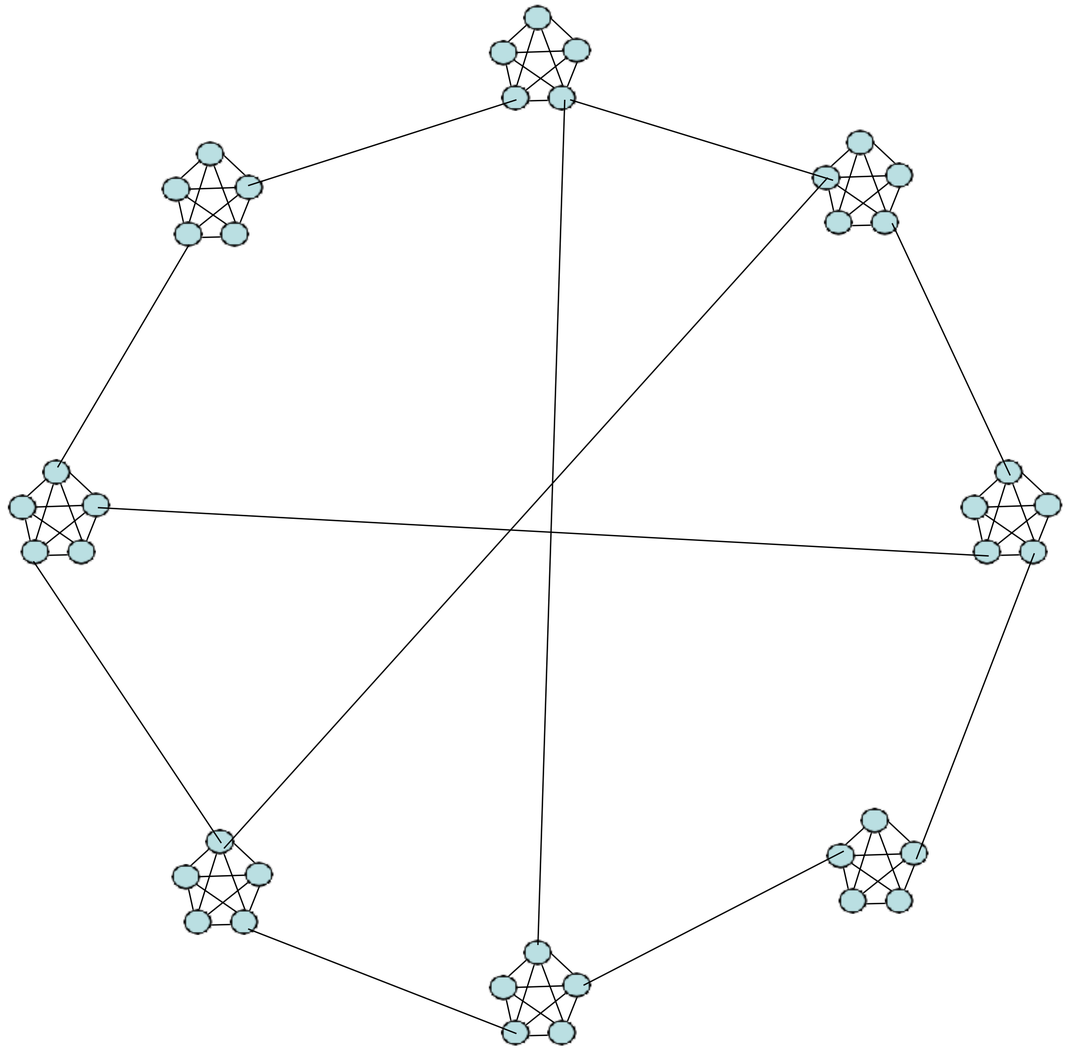
- spatial network where each person knows it‘s neighbours
- **High clustering C**
- **High path length L**

Random Network

- neighbours of each node are randomly selected out of all nodes
- **Low clustering C**
- **Low path length L**

Small World Network

- **High clustering C**
- **Low path length L**



Tie Strength

Tie strength and social search

- In Milgram's Small-World experiment, participants had to report if the next step in the chain was a **"friend"** or **"acquaintance"**
- **Chain completion rate was 26% more likely if the first interracial tie was an "acquaintance"** (target person was black, start persons white)
- Sparked the intuition that **weak ties are more effective to exchange information over a social network**

Granovetter's tie strength

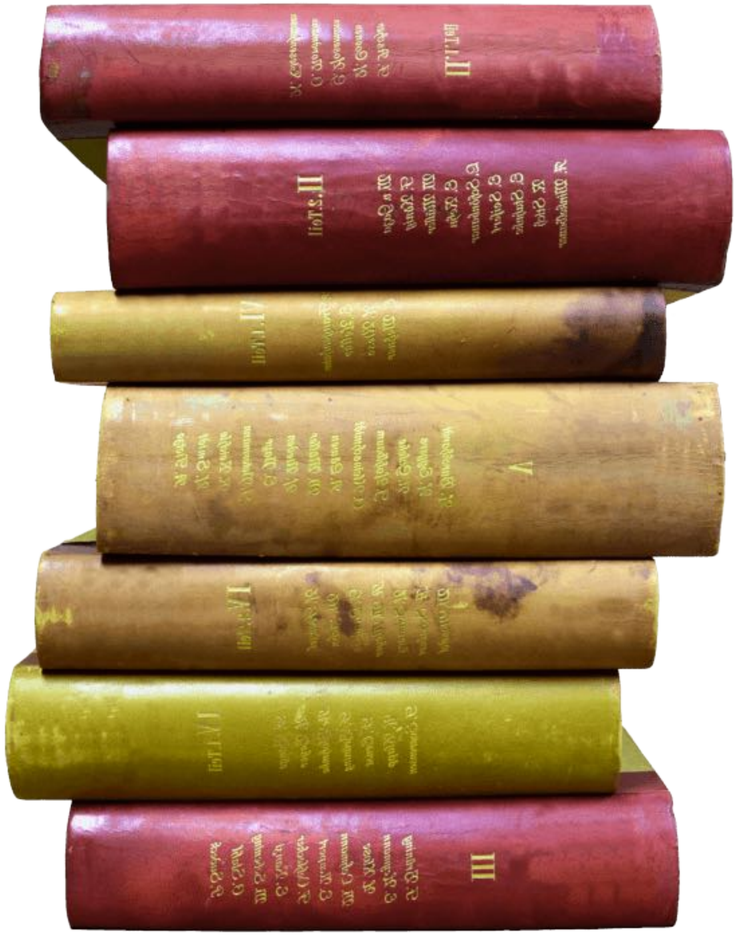
Paper was first rejected. Now has 40k citations.

- The **degree of overlap** of two individuals' friendship networks varies directly with the strength of their social tie. **The strength of a tie is proportional to the similarity of its endpoints**.

"The **strength of a tie** is a (probably linear) combination of the amount of **TIME**, the emotional **INTENSITY**, the **INTIMACY** (mutual confiding), and the reciprocal **SERVICES** which characterize the tie."
— Granovetter

Subjective and hardly measurable concepts! But can be operationalized …

Granovetter, Strength of weak ties, 1973

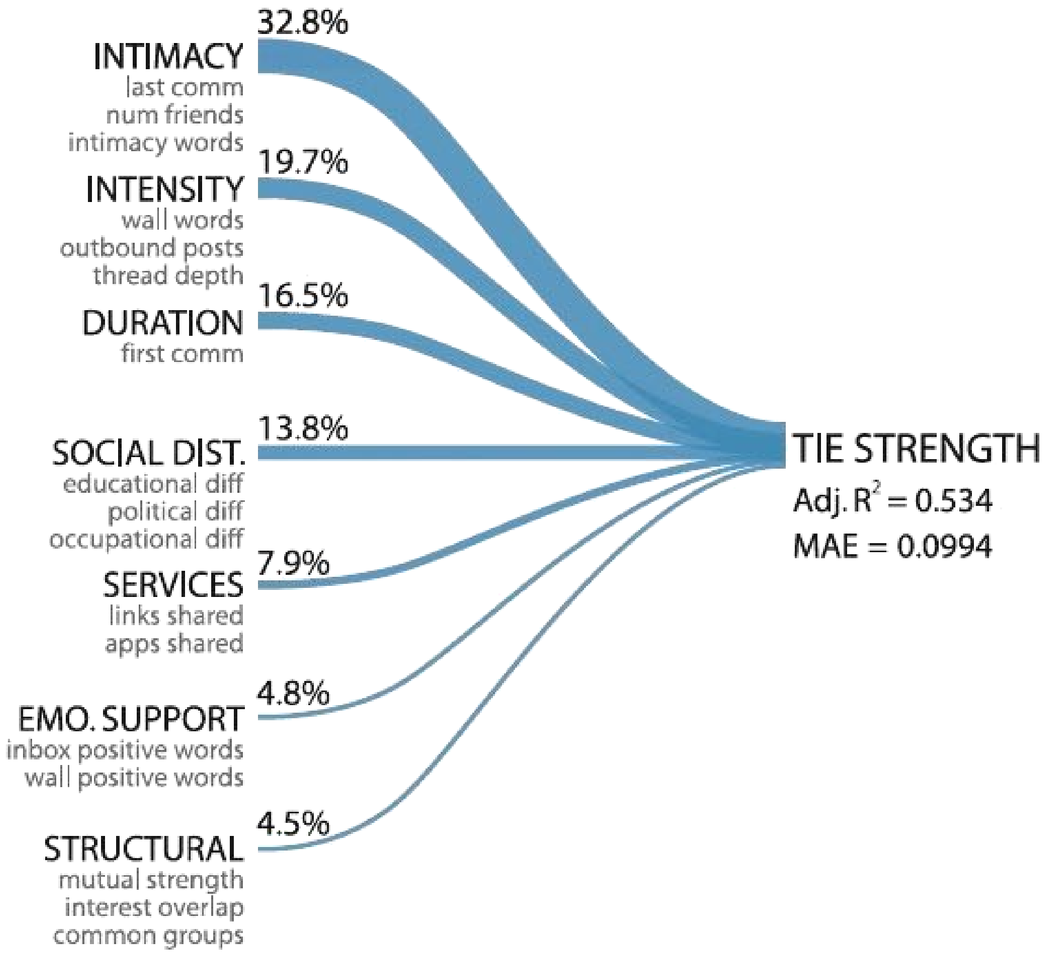


Definition

Operationalization

Objective measure of tie strength

- 2,184 self-rated Facebook friendship tie strength
- Goal: **predict tie strength** using proxies of Granovetter's indicators
- **85% prediction accuracy**



E. Gilbert & C. Karahalios. Predicting Tie Strength With Social Media 2009

Granovetter's tie strength

- **The stronger the tie between A and B**
  - **The higher the number of their shared social contacts**
  - **The more similar they are**

A+B and A+C are friends
A+B spend time together
A+C spend time together

→ C+B will be acquainted

Time spent = tie strength

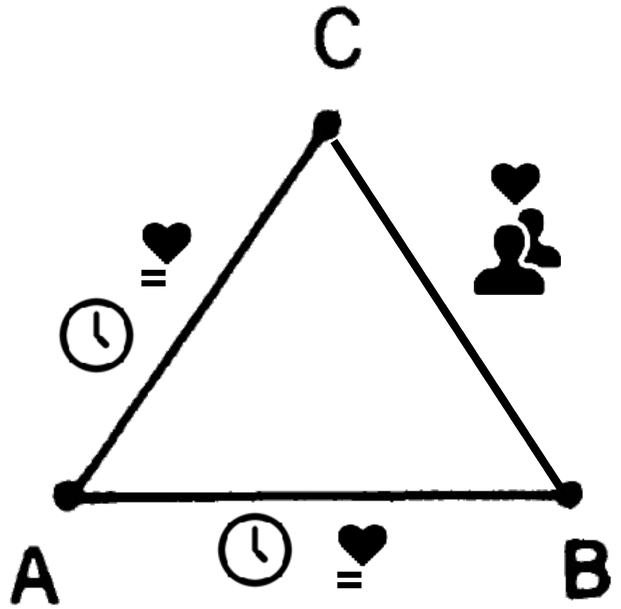


FIG. 1.—Forbidden triad

Granovetter, Strength of weak ties, 1973

Theory -> Hypothesis -> Facts

“Strength means time, intimacy…”

Theory

“The stronger a tie, the more similar its endpoints ”

“… and the higher the number of their shared connections”

People who declare to be close friends spend more time together and exchange more emotion words

Hypothesis

Similarity (profile, behavioral) of people connecting is higher. Links that close triangles are more likely to form

measured on social media exchange

Facts

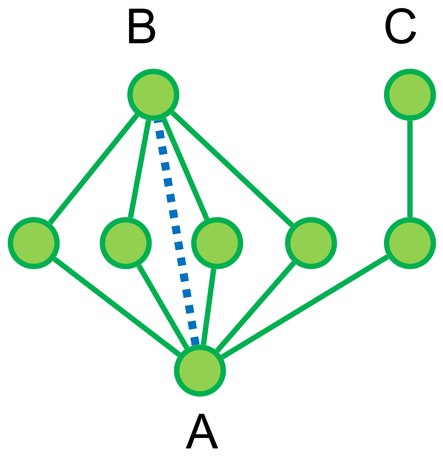
measured on online social graphs

Validity? Do more (and different) experiments

Hypothesis 1: Triangle closure

structural

- **The more triangles the edge closes, the higher the likelihood of its formation**
- Can be tested with **predictive frameworks**
  - Target node A
  - Its distance-2 neighbors at time t: $\Gamma^2_t(A)$
  - Predict the nodes in $\Gamma^2_t(A)$ that will be connected with A at time $t+\Delta$
- **Number of common neighbors (and derivative measures) is highly predictive**
  - e.g., Jaccard coefficient



Kleinberg et al., The link prediction problem for social networks, 2003

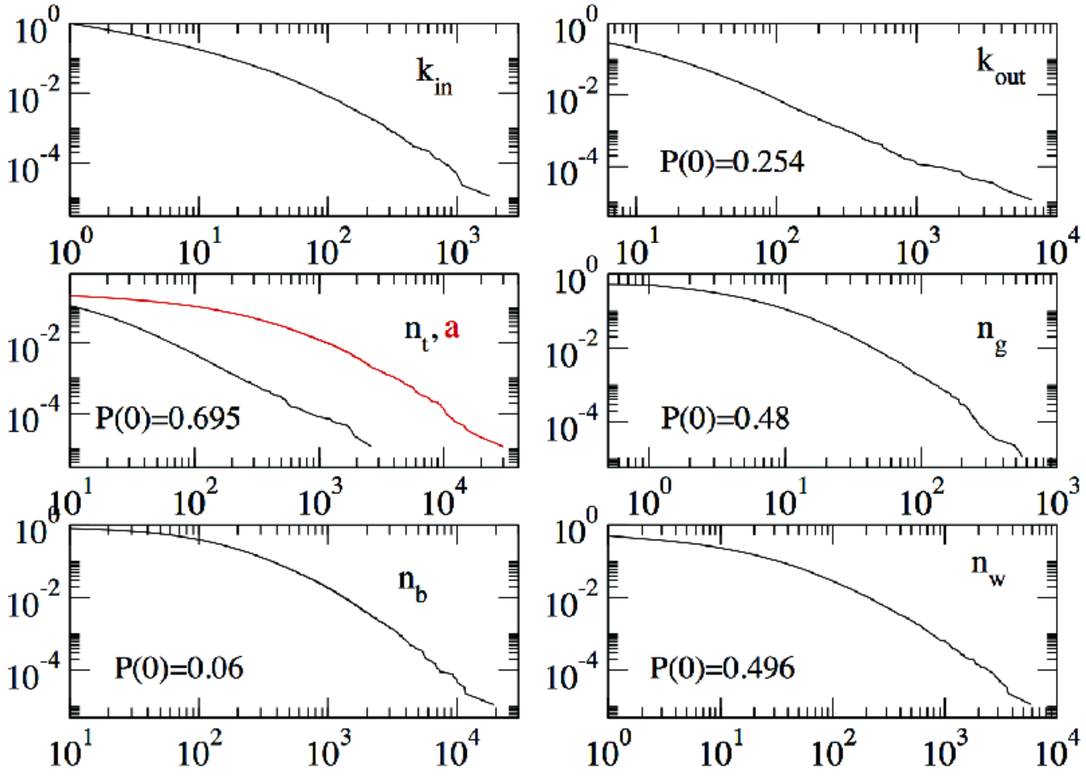
Hypothesis 2: Similarity

homophily

Activity patterns of users are highly heterogeneous

- **k-in** in the friendship network
- **k-out** in the friendship network
- $\mathbf{n_t}$ of distinct tags in the vocabulary
- $\mathbf{n_g}$ groups
- **a** total tagging activity
- $\mathbf{n_b}$ number of books in the library
- $\mathbf{n_w}$ number of books in the wishlist

Activity metrics are all positively correlated

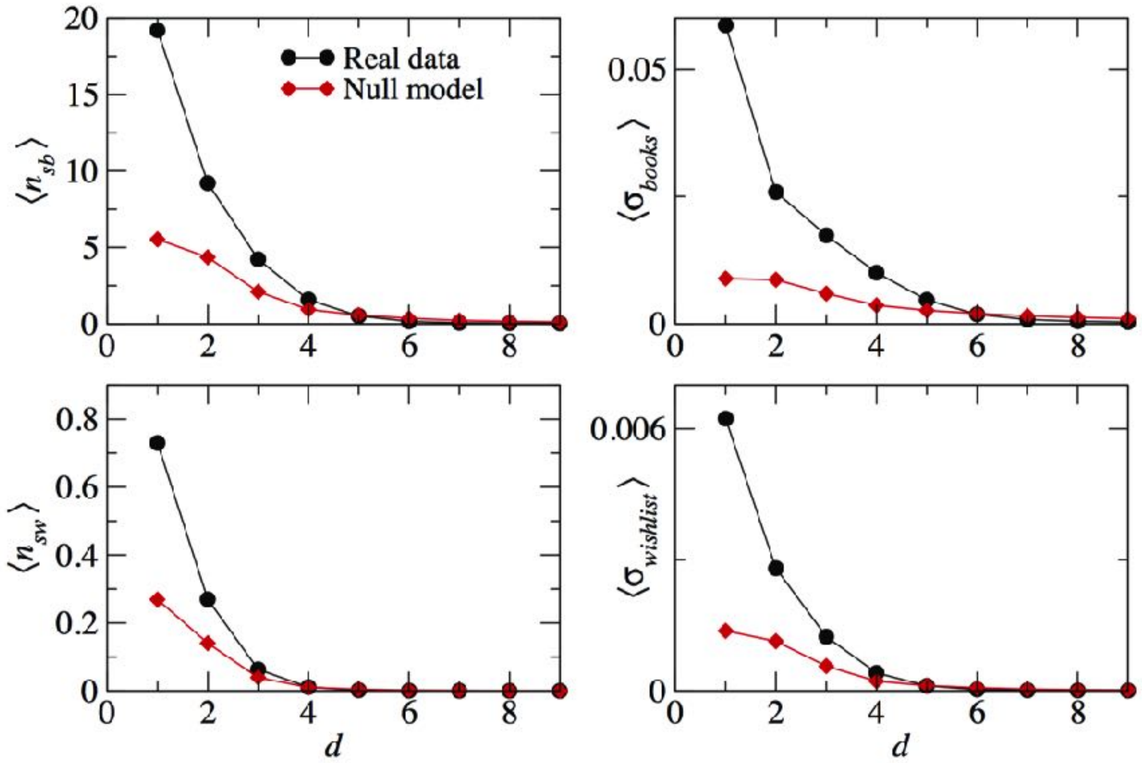


Link creation and information spreading over social and communication ties in aNobii, EPJ Data Science 2012
Friendship prediction and homophily in social media, TWEB 2012

Hypothesis 2: Similarity

homophily

Topical Similarity



n_{sb} number of shared books in the library

n_{sw} number of shared books in the wishlist

δ_{books} cosine similarity books vector in the library

$\delta_{wishlist}$ cosine similarity books vector in the wishlist

δ_{tags} cosine similarity tags vector

δ_{groups} cosine similarity groups vector

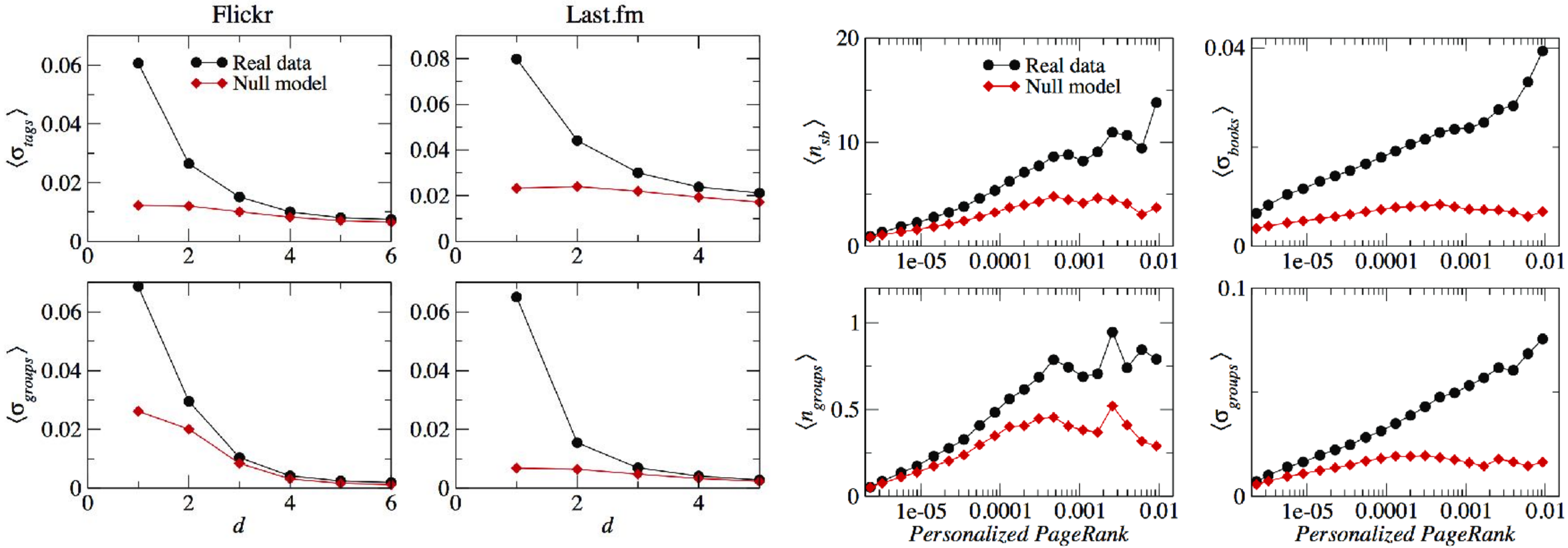
Link creation and information spreading over social and communication ties in aNobii, EPJ Data Science 2012

Friendship prediction and homophily in social media, TWEB 2012

Hypothesis 2: Similarity

homophily

Topical Similarity



Link creation and information spreading over social and communication ties in aNobii, EPJ Data Science 2012

Friendship prediction and homophily in social media, TWEB 2012

Hypothesis 2: Similarity

homophily

Link prediction

		Last.fm		aNobii	
Feature	Similarity	Active	Connected	Active	Connected
Baselines	Tasteometer	0.734	0.759	-	-
	Common neighbors	0.927	-	0.854	-
Items	Distrib cosine	0.663	0.560	**0.915**	**0.655**
	Distrib MIP	**0.749**	0.559	0.878	0.649
	Collab MIP	0.589	**0.613**	0.652	0.561
Tags	Distrib cosine	0.579	**0.625**	0.652	0.554
	Distrib MIP	0.697	0.618	0.651	0.560
	Collab MIP	**0.698**	0.559	**0.916**	**0.648**
Groups	Cosine	**0.810**	**0.677**	**0.662**	**0.690**
Library	Cosine	-	**0.769**	**0.923**	**0.768**

Takeaway

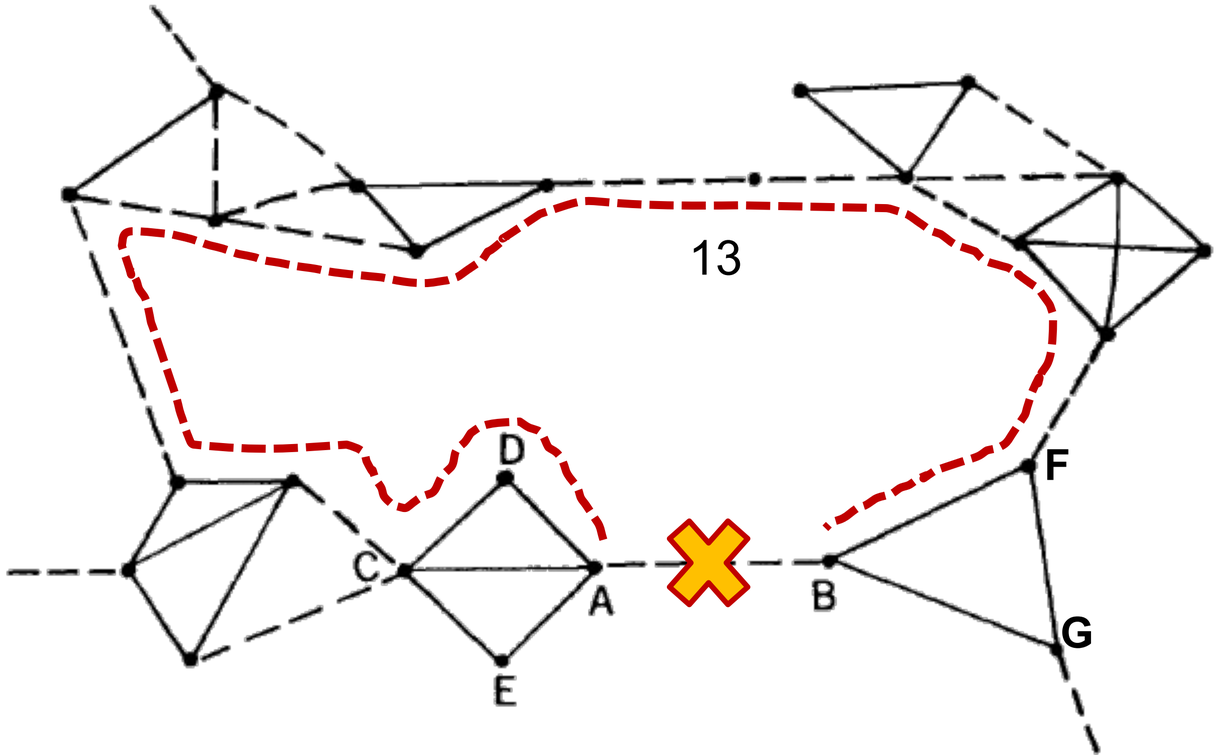
- **Similarity on profile and activity is a good predictor**
- **Structural + Profile features perform the best**

	Tags collab MIP	Groups	Library	CN	Profile feaures	All features
AUC	0.785	0.807	0.811	0.844	0.924	0.963
Accuracy	0.786	0.809	0.812	0.846	0.877	0.915
FP Rate	0.177	0.143	0.335	0.031	0.109	0.072
FN Rate	0.251	0.240	0.041	0.277	0.137	0.099

Link creation and information spreading over social and communication ties in aNobii, EPJ Data Science 2012
Friendship prediction and homophily in social media, TWEB 2012

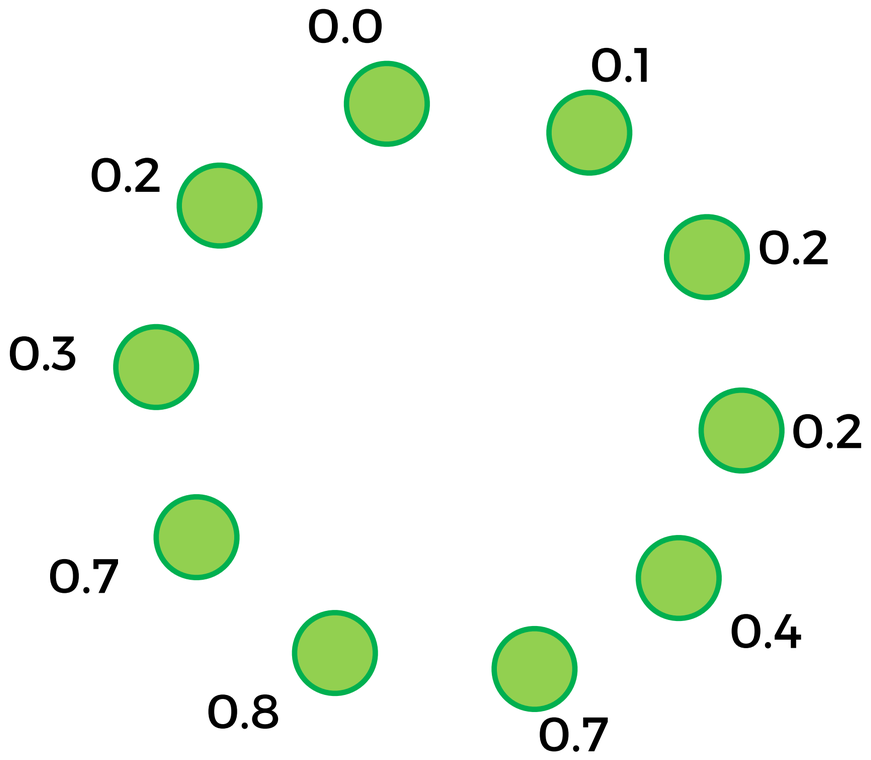
Corollary: Weak ties and bridges

- **Bridges are always weak ties**, never strong ties
- **Strong ties bring fragmentation**: information has way more pathways to stay within dense local clusters
- **Weak ties reduce path length** and create **opportunities to access new information**



Threshold model of diffusion

- Individuals with binary activation status
- Each person i has an activation threshold $\Phi(i)$ in [0,1]
- At time t, i activates if the fraction of active nodes is $\geq \Phi(i)$

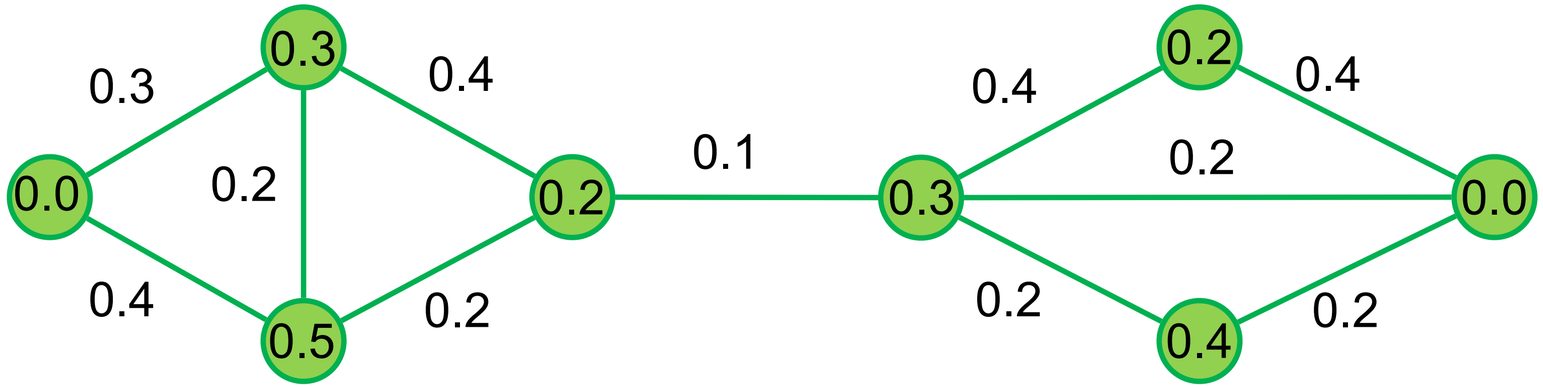


Complex contagion (multiple exposures) vs. simple contagion (contact with disease)

Granovetter, Threshold models of collective behaviour, 1978

Linear model threshold

- Tie strength and individual threshold can be combined in a networked model
- Each person *i* has an **activation threshold** *Φ(i) in [0,1]*
- At time *t*, *i* activates if sum of strengths of ties towards active people is ≥ *Φ(i)*



Kempe et al., Maximizing the Spread of Influence through a Social Network, 2003

Another example

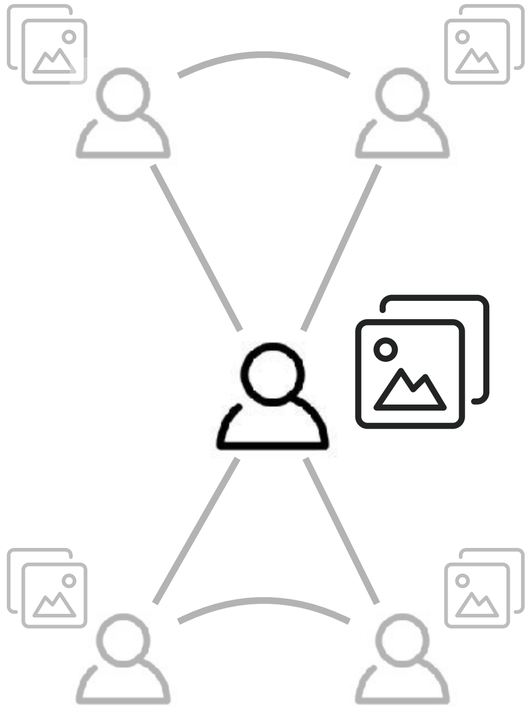
- In a photo sharing social network (Instagram), **people post pictures of the icecreams** they are eating. When users upload a photo, all their friends are notified. Friendhsip links are directioned according to the information flow: a directed edge A->B means that B follows A and that B is notified with A's profile updates.
- We observe a series of "icecream consumptions" (in the form [userA, timeX]) over a period of time. Some users might consume icecream multiple times. The social network is static over that period of time. **We would like to check whether the icecream consumption is driven by a phenomenon of "social contagion".**
- **Hypothesis: People consume an icecream when more than x% of their friends have recently consumed an icecream (activation threshold).**

Social Influence, Quality and Engagement in Social Media

social

popularity

(success)



social network

intrinsic

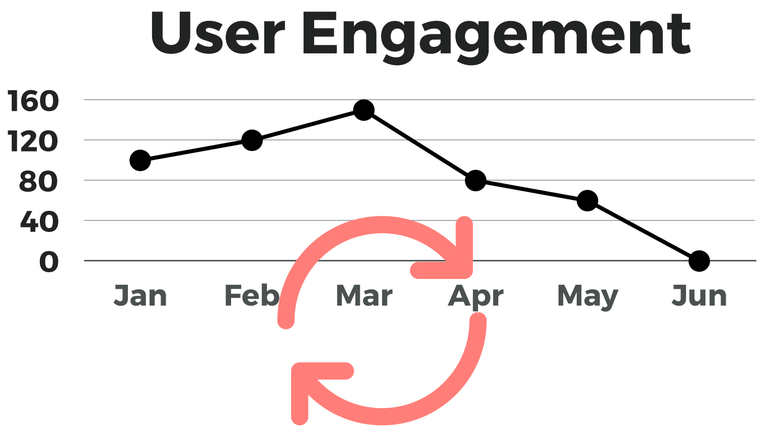
quality

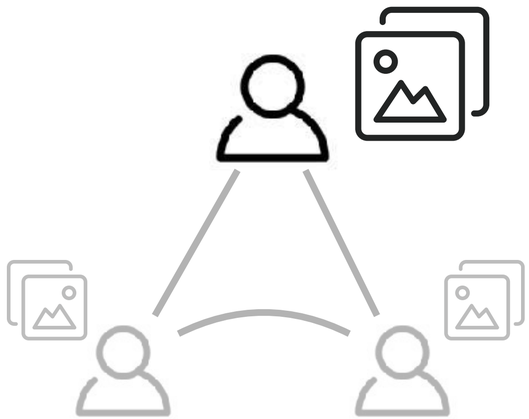
(performance)

social

popularity

(success)





social network

intrinsic

quality

(performance)

Questions

1. Popularity = Quality ?
2. Do social ties affect quality?
3. Are network effects intertwined with engagement?

social

popularity

(success)

Activity (Likes, Views, Comments)

Connectivity (followers, friends)

intrinsic

quality

(performance)

Visual Aesthetics

- **Multidisciplinary**
  - **Philosophy**
  - **Experimental Aesthetics (psychology)**
    - complex interaction between perception, cognition and emotion
  - **Neuroaesthetics (neurology)**
  - **Computational Aesthetics (computer vision)**

How beautiful is this photo?

human score

4.5

Computational
Aesthetics Engine

machine score

4.3

Computational Aesthetics

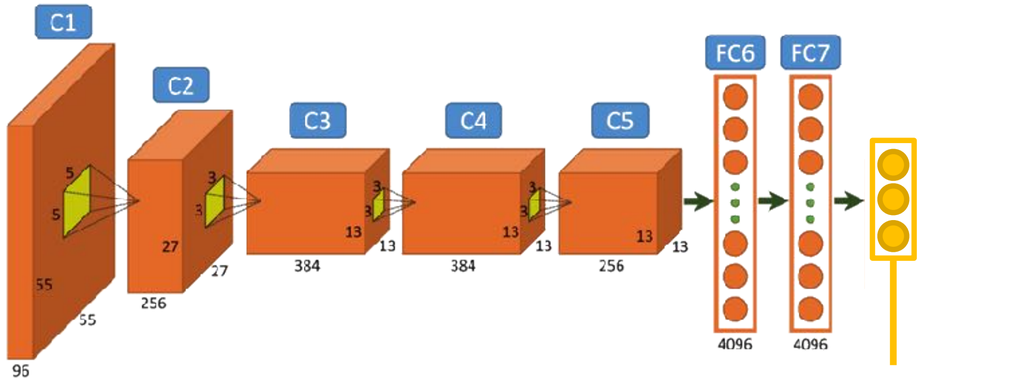
- **Hand-Crafted features**
  - **low level**
    - blur, contrast, colorfulness, saliency, texture, shape, energy
  - **compositional**
    - rule of thirds or depth of field
- **Generic Image Features**
  - **deep learning**
    - content, form, and [blackbox]

Algorithmic Beauty

(Transfer Learning)

OBJECT DETECTION (ImageNet)

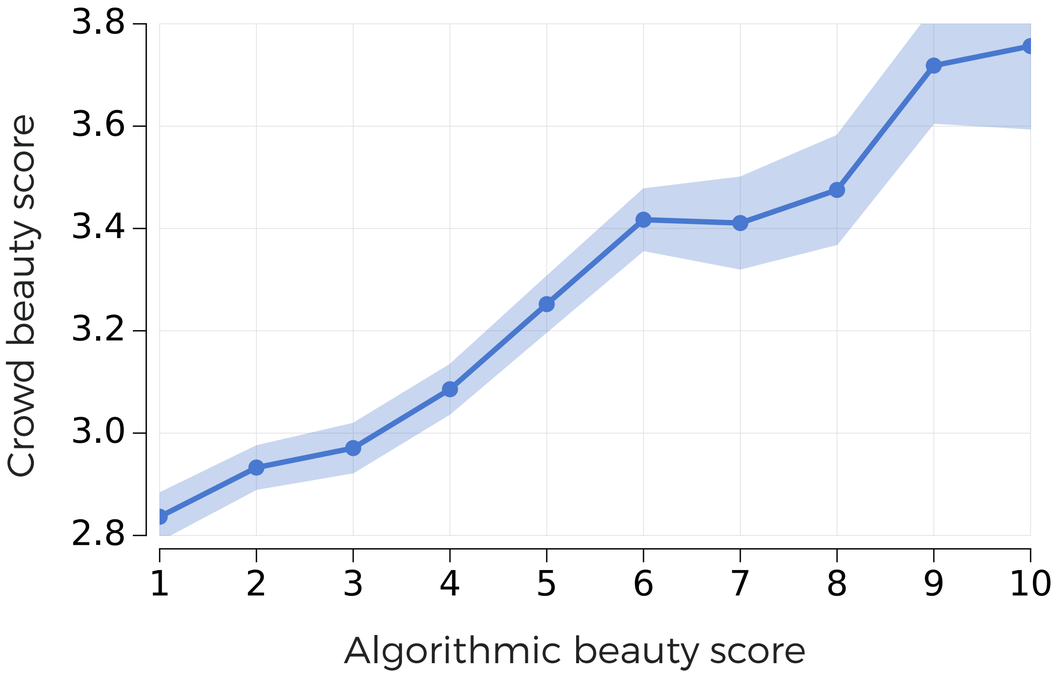
FINE TUNING



AESTHETICS SCORES

[TKDE17] Beautiful and Damned. Combined effect of content quality and social ties on user engagement

Machine eye ≈ Human eye



rank correlation = 0.5

$p < 0.01$

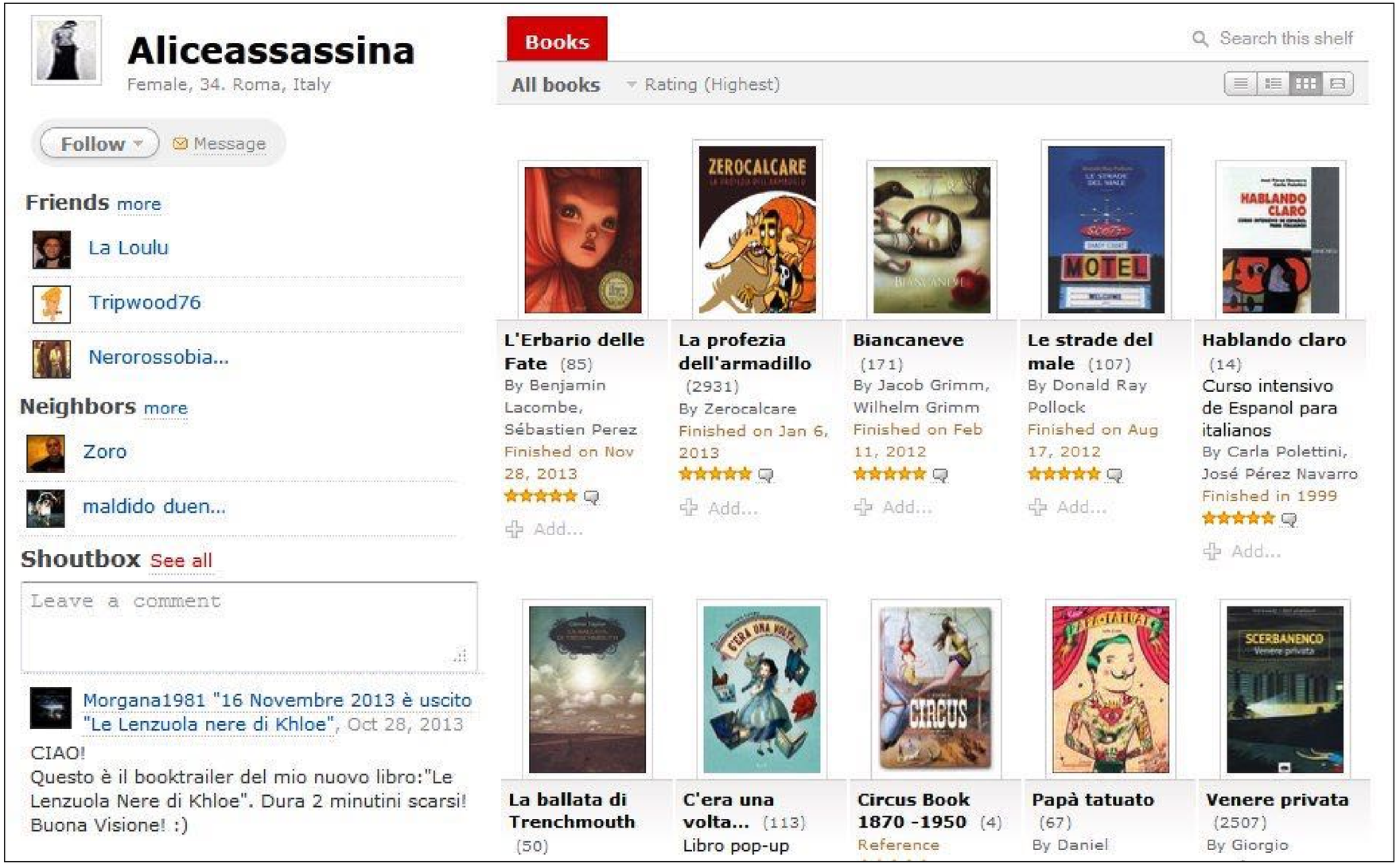
Human evaluation

- **4 Categories**
  - nature, animals, people, urban
- **Full spectrum of popularity**
- **Good agreement**
- **Large ground truth of aesthetic scores made by non-expert**
  - **10K photos**
  - **60K scores**
- **Available at http://di.unito.it/beautyicwsm15**

1.

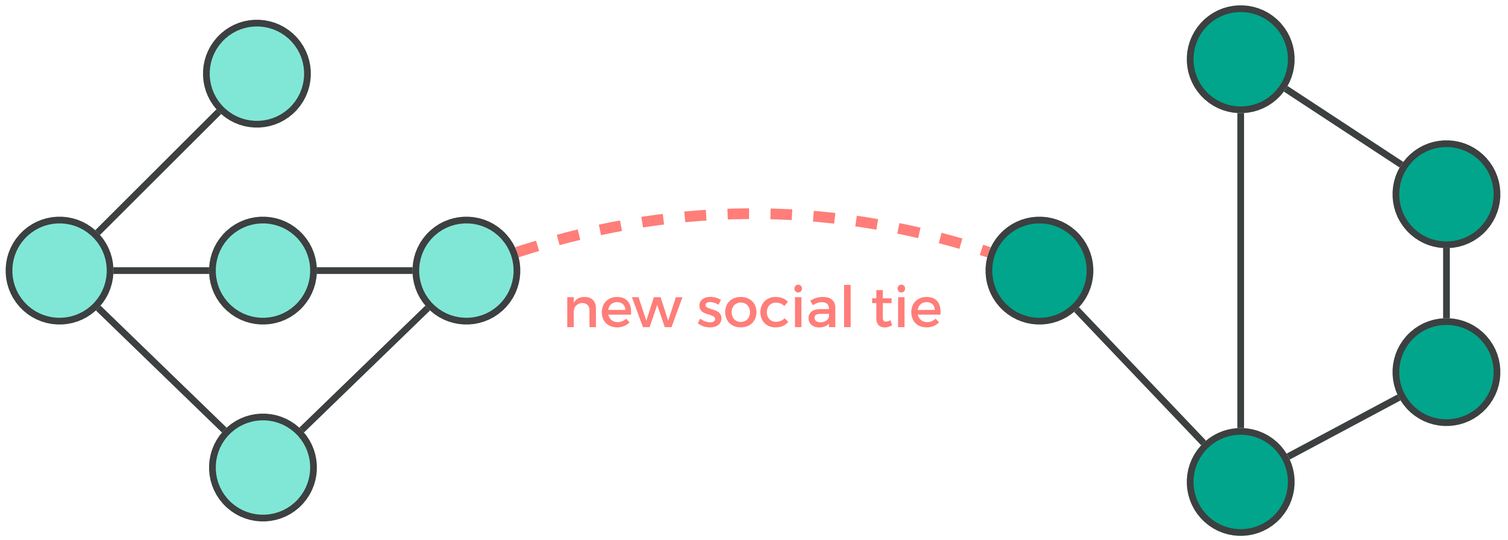
POPULARITY AND QUALITY DO NOT (ALWAYS) ALIGN

aNobii.com

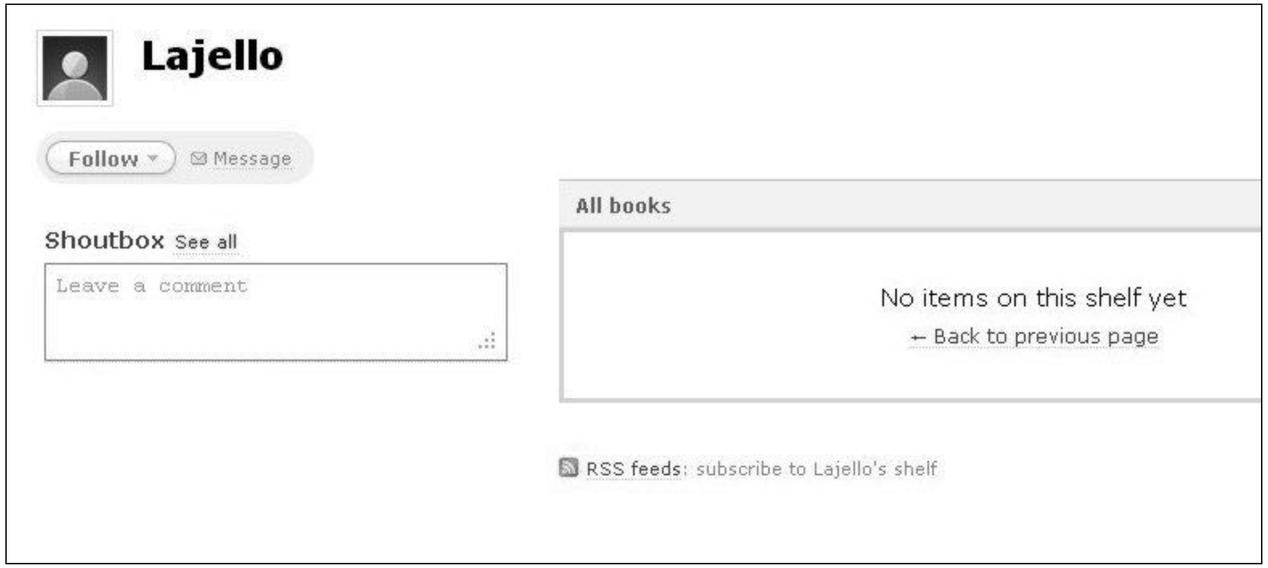


Aiello et al. “People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks”. ICWSM 2012

Our goal was to predict links

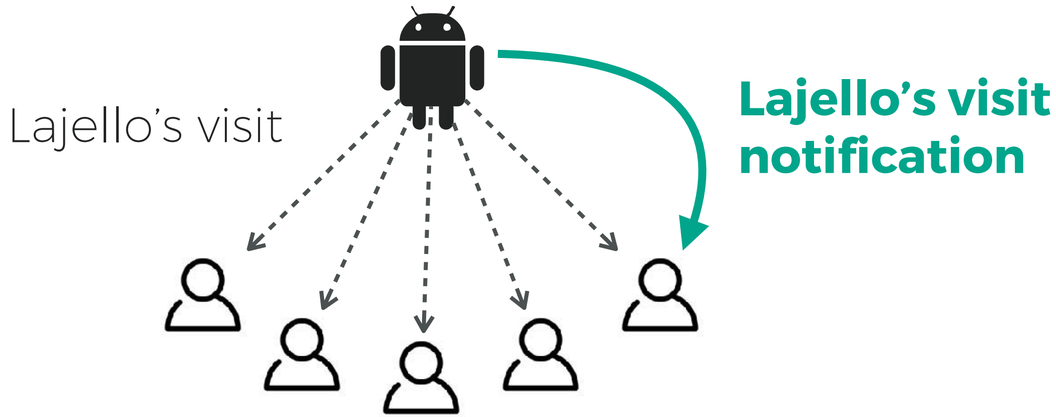


Lajello crawler



totally empty profile!

Iterative probing
(every week)



Gain of popularity

2ND
(OUT OF 100K)

>1.2K

DISTINCT USERS

28TH
(OUT OF 100K)

~2.5K

PUBLIC MESSAGES

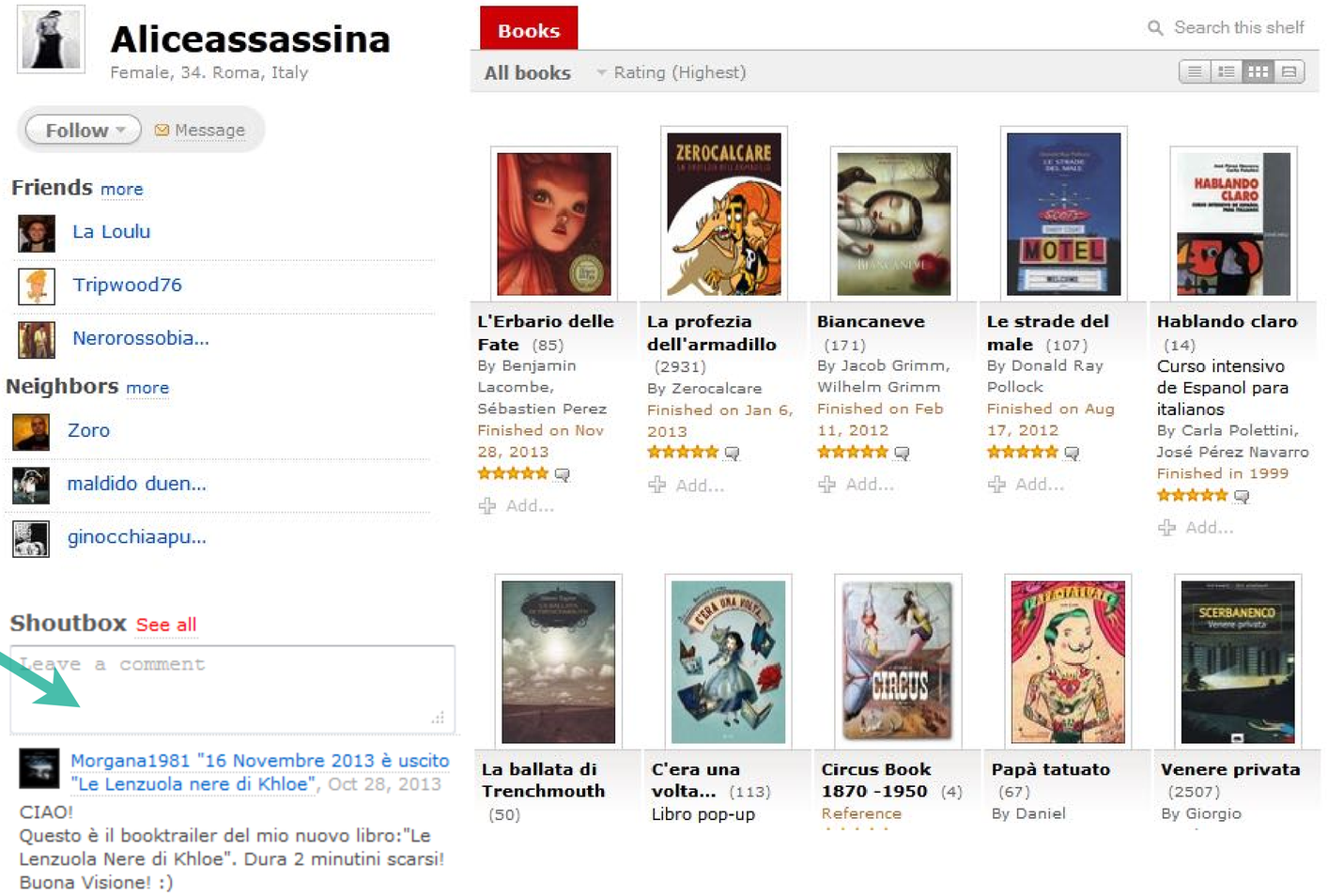
0.3%

>200

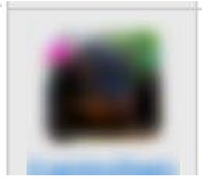
INCOMING LINKS

Influence

Lajello left a personalized comment in the **shout box** of 2K users



chi sei?



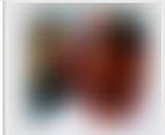
Le tue visite cominciano ad essere inquietanti....

Ieri

Grazie Lajello, mi sono divertita un sacco a leggere i commenti degli altri anobiani. Sembra un esperimento di psicologia sociale, se non ti dispiace ti aggiungo come vicino! e resisti eh...non pubblicare un libro! ;)

Due settimane fa

Ti stimo sinceramente.



LAJELLO... HAI STUFATO..NON SE NE PUO' PIU'...STA ATTENTO/A CHE SONO CAPACE DI ASSOLDARE UN HACKER PER VEDERE CHI SEI..E PO' SONO C...TUOI

Tre settimane

chapeau!!

ahahahhahahaha tu sei un genio!!!!! sei davvero un genio!!! insomma ma quante visualizzazioni hai???? sei un grande!!!!! riesci a farti visitare e a farti scrivere pur non avendo libri!!! ti adoro sei grandissimo :P

ahahahaahah tu sei un genio!!

già che mi ritrovo qui mi faccio pubblicità! Venite a vedere la mia libreria è la più bella -del mondo-. (l'ultima parte andava sottolineata..)

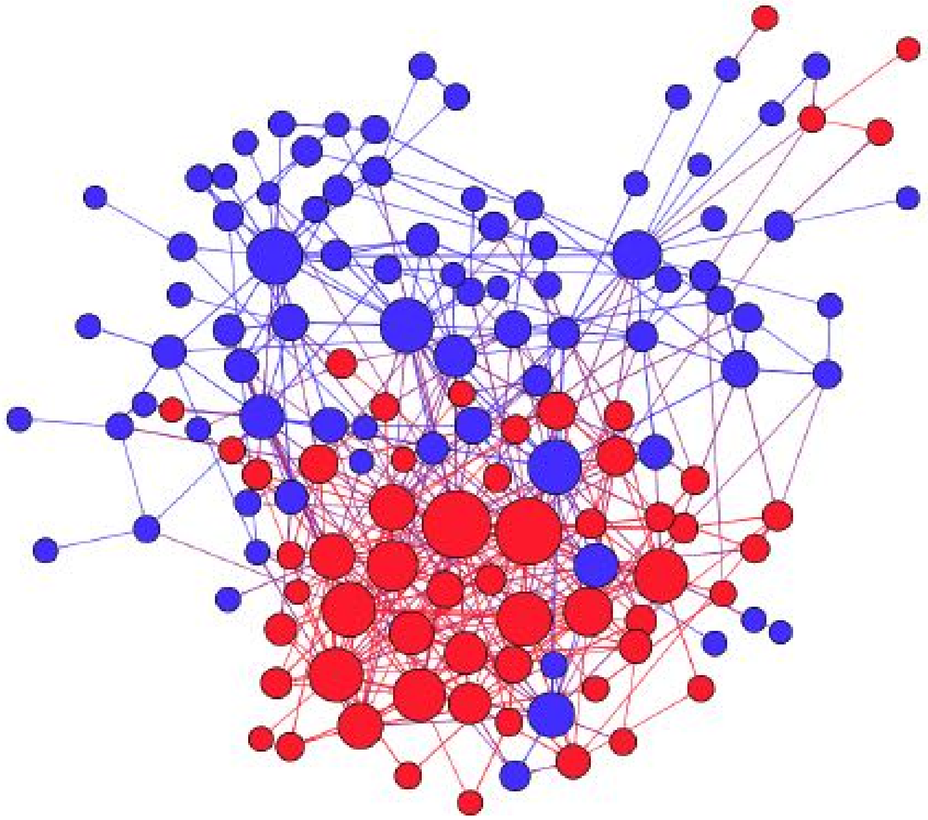
Due settimane fa

un grande.
continua così. Grazie delle visite, si vede che ti sto simpatica....

P.S: propongo di aprire un gruppo the Lajellos fans...

3 giorni fa

From Influence to Chaos



2.

QUALITY INSPIRES QUALITY

DATA

15B

PHOTOS

40M

USERS

GROUPS

FAVORITES

550+M

SOCIAL TIES

10+

YEARS

Quality at user-level

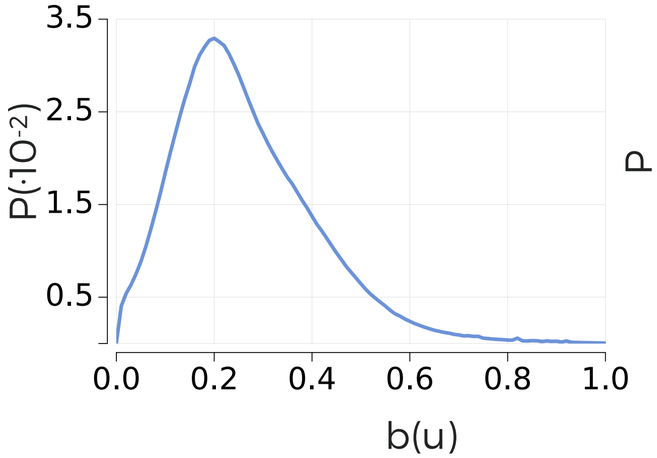
b(u) : average of beauty scores

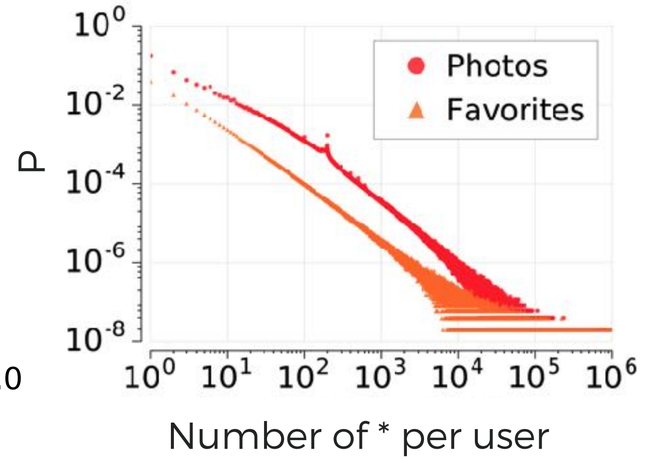
we are currently working on different strategies

evolution of photographers careers

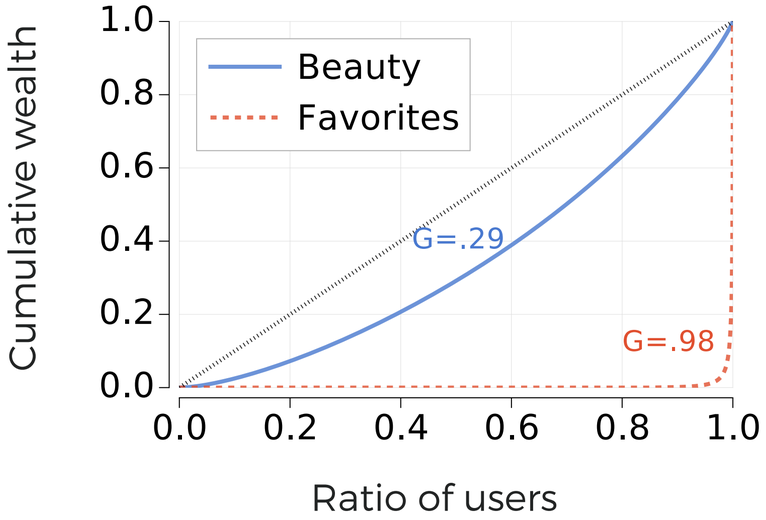
Quality distribution is highly unequal (G=0.98)

measuring beauty at scale (15B pics)





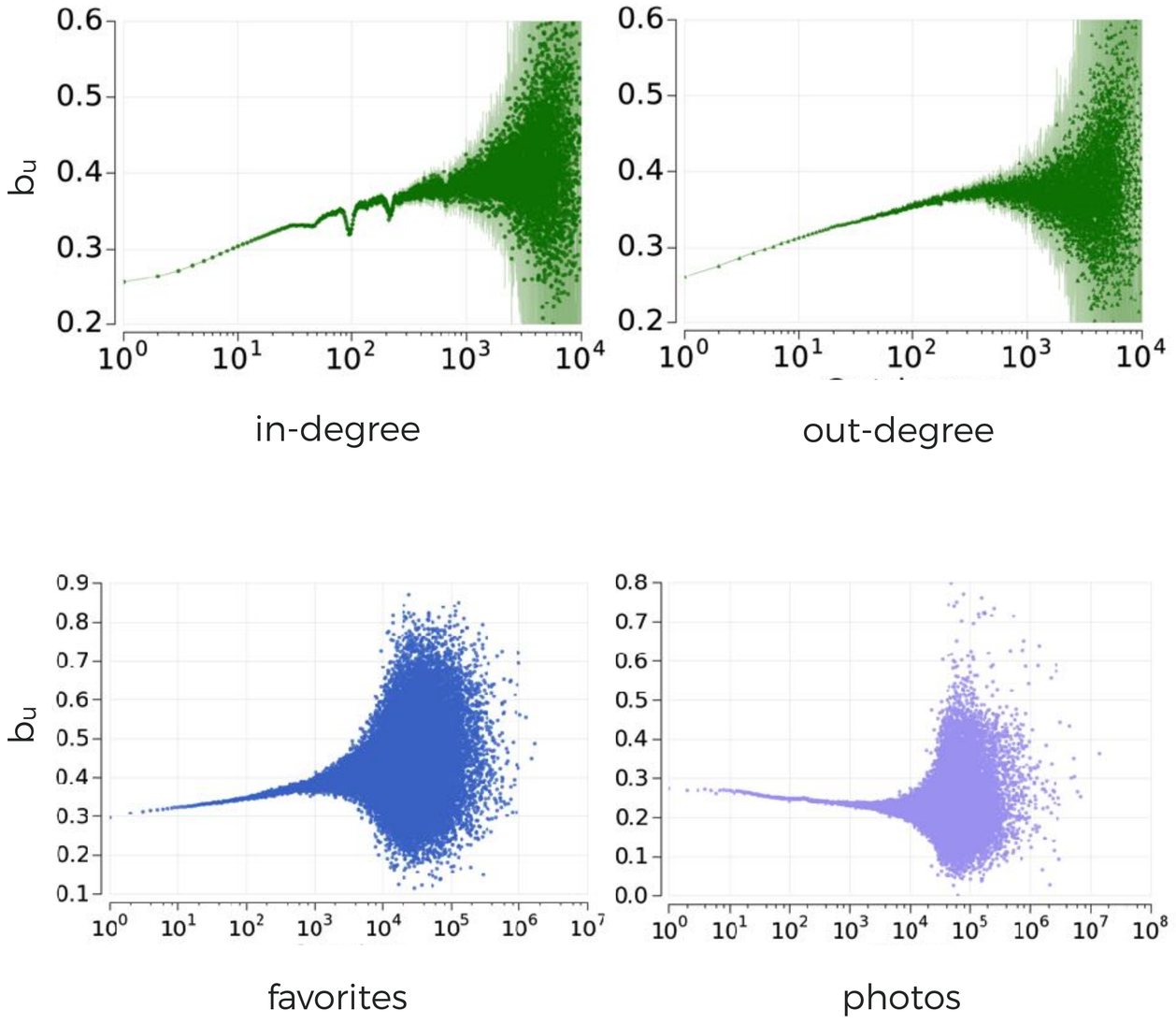
inequality (Gini's coefficient)



QUALITY AND CONNECTIVITY ARE (WEAKLY) CORRELATED

quality(u) is

in-degree	(ρ = .22)
out-degree	(ρ = .24)
#favorites	(ρ = .17)
#photos	(ρ = -.03)



SOCIAL TIES LINK USERS WITH SIMILAR QUALITY

assortative mixing

users tend to be linked to accounts that publish photos with similar quality

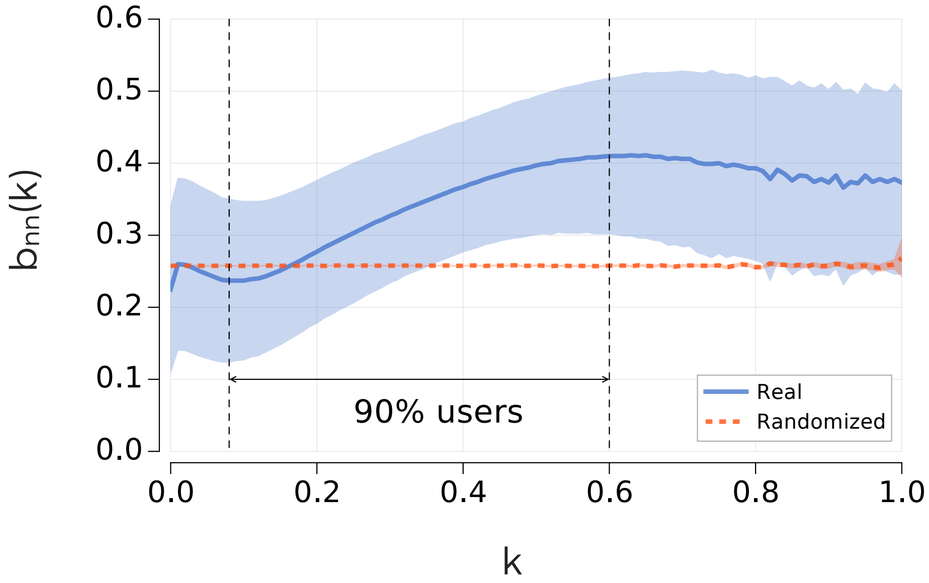
(Spearman ρ = 0.48)

This could be ascribed to

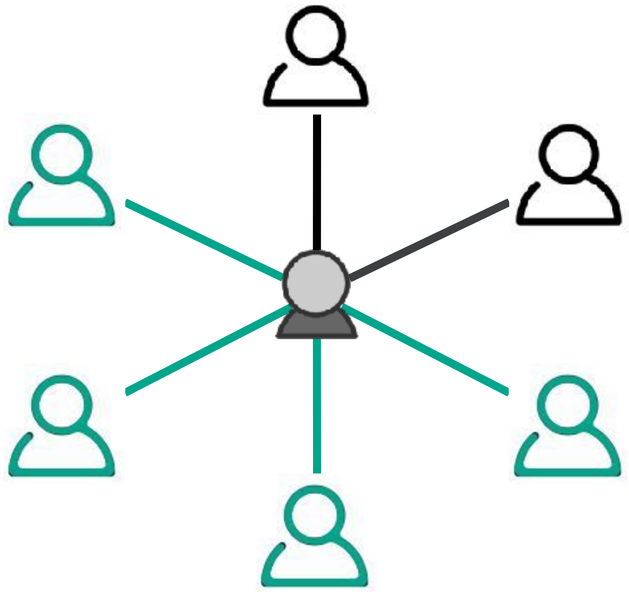
homophily

influence

$$b_{nn}(k) = \frac{1}{|\{i : \bar{b}(i) = k\}|} \cdot \sum_{j \in \Gamma_{out}(i), \forall i : \bar{b}(i) = k} \bar{b}(j)$$



Majority Illusion



0.26
μ

43%
above μ

65%
> 43% friends above μ

To summarise

- High unequal distribution of quality
- Quality and connectivity are weakly correlated
  - no correlation with activity
- Users tend to connect to similar quality users
- Local perception of higher quality of neighbours

Matching experiments

Inferring causality from observational data [Rubin 2001][Stuart 2010] [Althoff et al. 2017]

G_t **Treatment Group** $\cong$ G_c **Control Group**

BALANCED

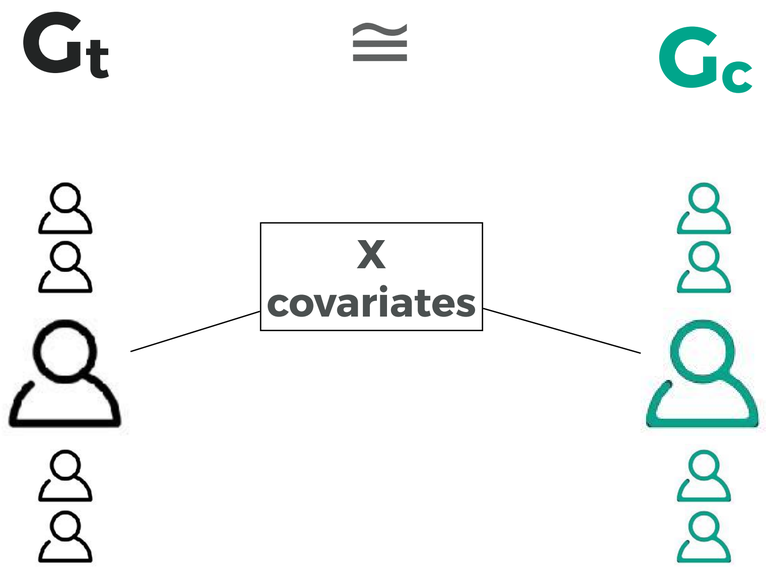
Two groups are **balanced** on a **covariate X** when the **standardized bias SB_X** in under a given threshold (usually set to 0.25)

$$SB_X(G_t, G_c) = \frac{\bar{X}_t - \bar{X}_c}{\sigma(X_t)} \leq 0.25$$

USER
in-out-degree
photos
groups
favorites
μ (neighs)
μ (beauty)
weeks from join date

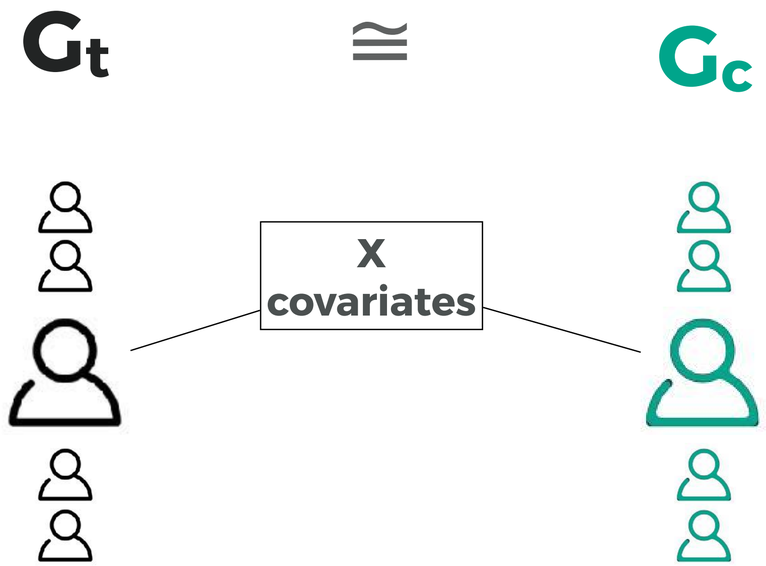
NEIGHBORS
photos
μ (beauty)

Setup on Flickr data



$week_1$ $week_t$ $week_{t+1}$

Setup on Flickr data



$week_1$

$week_t$

$week_{t+1}$

Quality inspires quality

treatment:

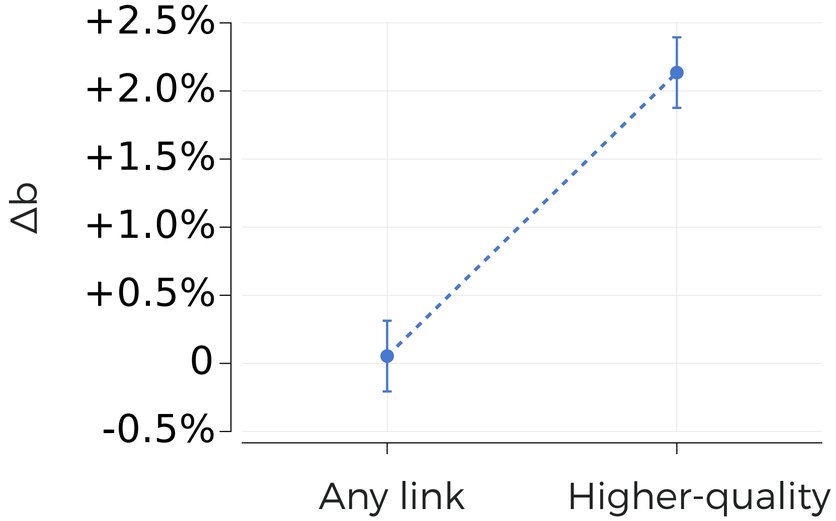
new tie to a **higher-quality** neighbor

outcome variable:

increase in quality in future photos (i.e., photos made in w_{t+1})

NEIGHBOURS QUALITY AFFECT USER QUALITY

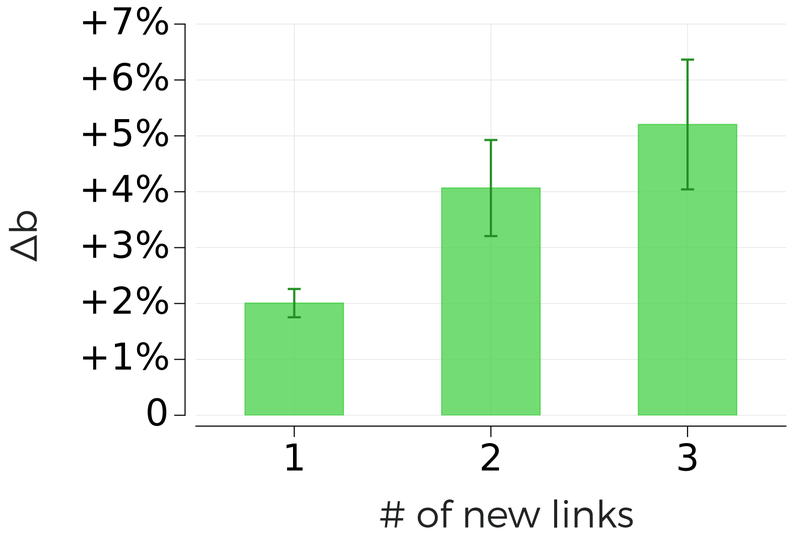
1) G_t experiences an average 2% increase, no significant increase for G_c



NEIGHBOURS QUALITY AFFECT USER QUALITY

1) G_t experiences an average 2% increase, no significant increase for G_c

2) Influence effect accumulates with new connections, with diminishing returns

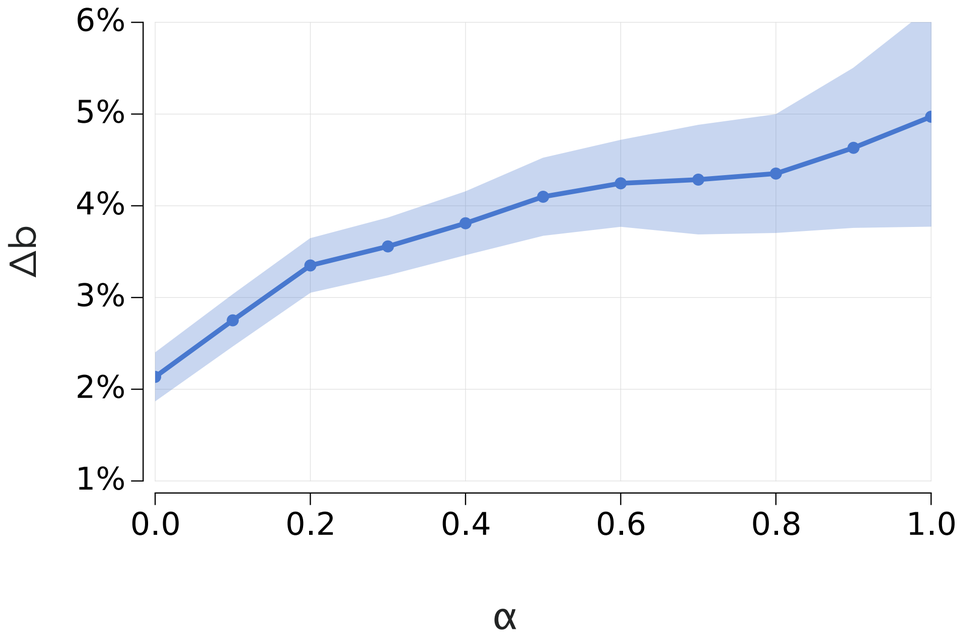


NEIGHBOURS QUALITY AFFECT USER QUALITY

1) G_t experiences an average 2% increase, no significant increase for G_c

2) Influence effect accumulates with new connections, with diminishing returns

3) **The greater the beauty differential, the greater the increase, noticeable until α=0.5**



Imbalance drops engagement

treatment:

average neighbor beauty deviates from the user beauty (**high imbalance**)

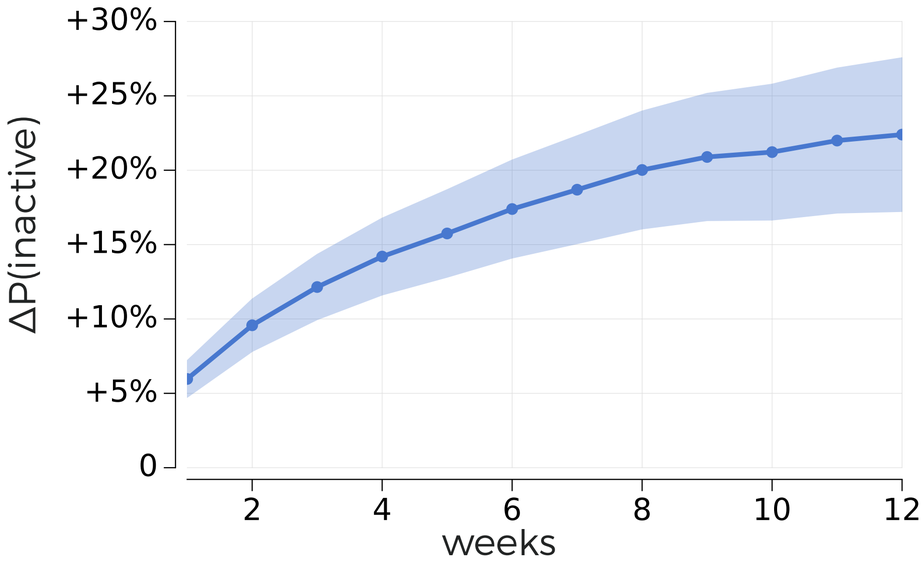
outcome variable:

proportion of users in each group who remain **inactive** (i.e., no photo uploads)

QUALITY IMBALANCE AFFECT ENGAGEMENT

G_t has higher probability of inactivity that grows from +5% to +20% in the first 12 weeks

People exposed to photos that deviate significantly, in terms of quality, from their own contributed content, are more likely to become disengaged in the future



The Social Brain Theory

Densification of the Social Graph

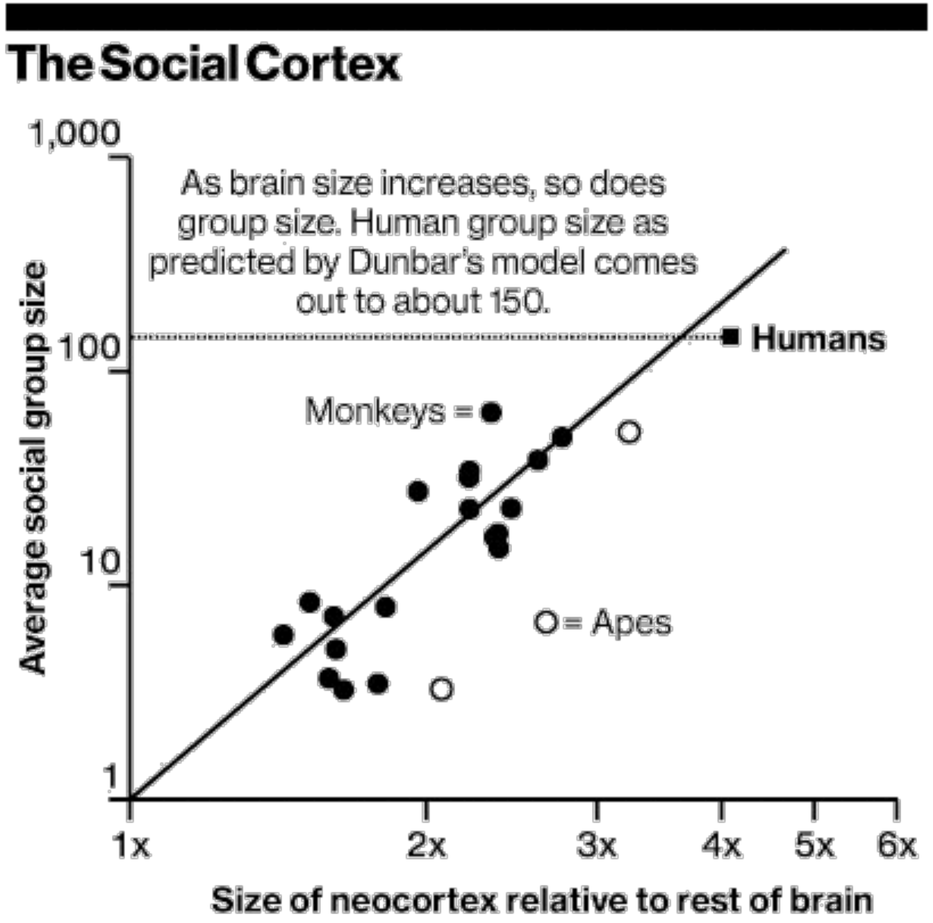
- **Why does the triangle closure process stops?**
  - Time and space constraints
  - Attention and cognitive boundaries

- **How big is a person’s ego network?**

- **Is there a universal limit to human connectivity?**

The Social Brain Theory

- **Big brains evolved to solve the problem of social life**
  - Cognitive effort
  - Time constraints (social grooming)
  - Face to face interaction
  - Emotional intensity
- **Correlation between neocortical volume and typical social group size**
- **Limits of human social network size between 100 and 200 individuals** (150 is the anecdotal number – Dunbar number)



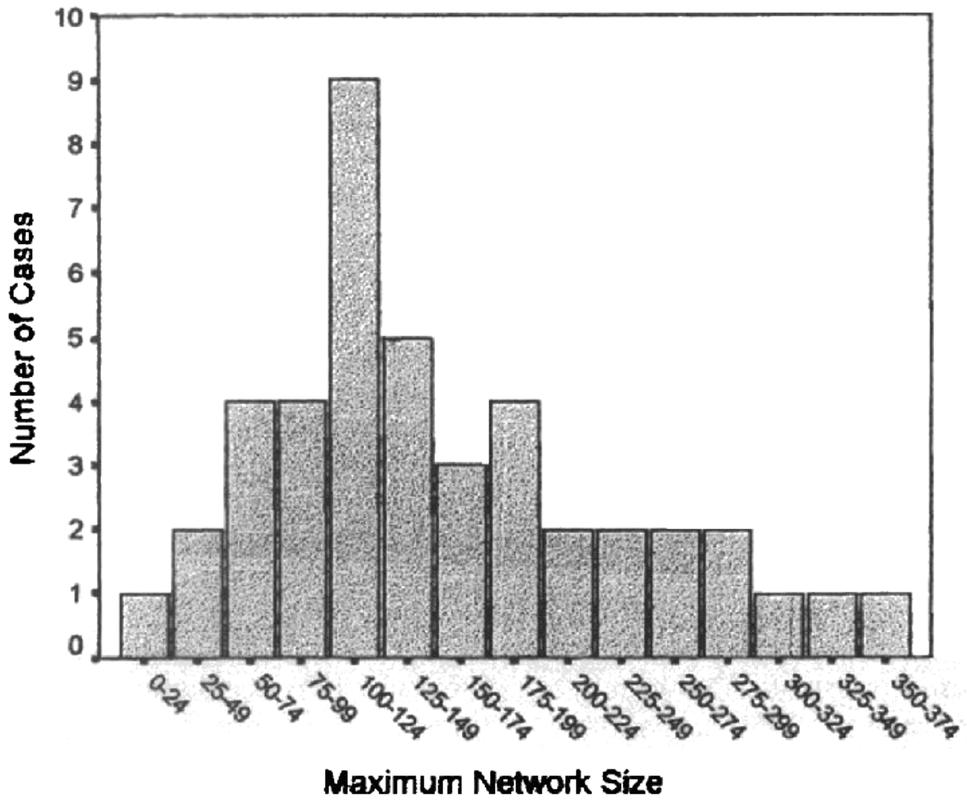
R. Dunbar – Neocortex size as a constraint on group size in primates, 1992

Exchange of Xmas cards

- **Sending Christmas cards individuals make an effort to contact their network of acquaintances they value**

  - 43 questionnaires from UK households
  - 2,984 Christmas cards sent
  - Number of contacts =
    - # family members +
    - # card recipients (household) +
    - # people seen on Christmas
  - Results
    - Mean cards = 68.19
    - **Mean contacts = 153.5**

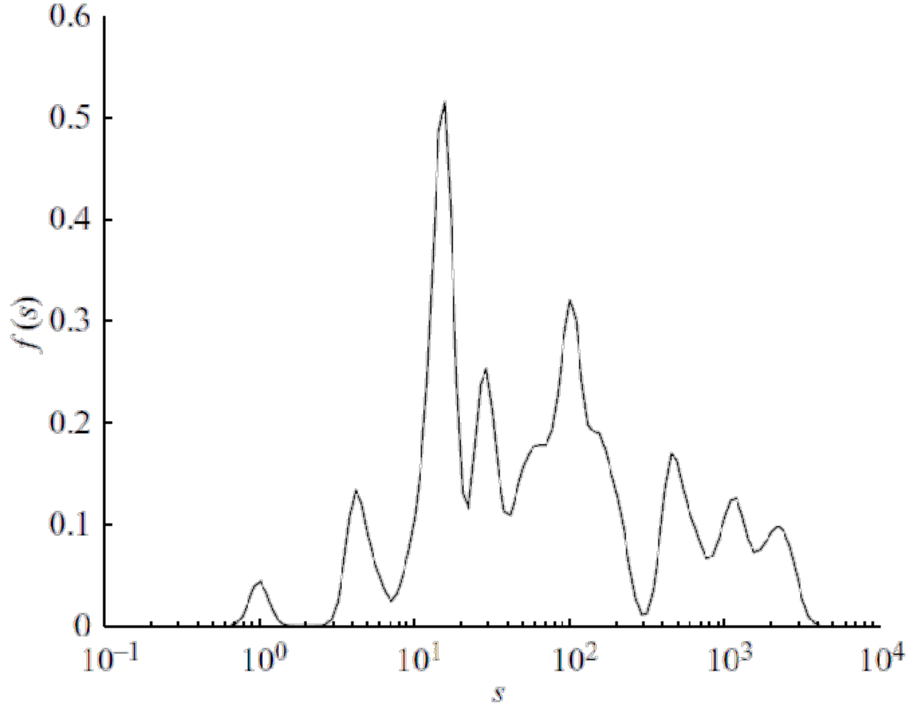
Experimental limitations ?



Hill & Dunbar. Social Network Size in Humans, 2002

Social Circles

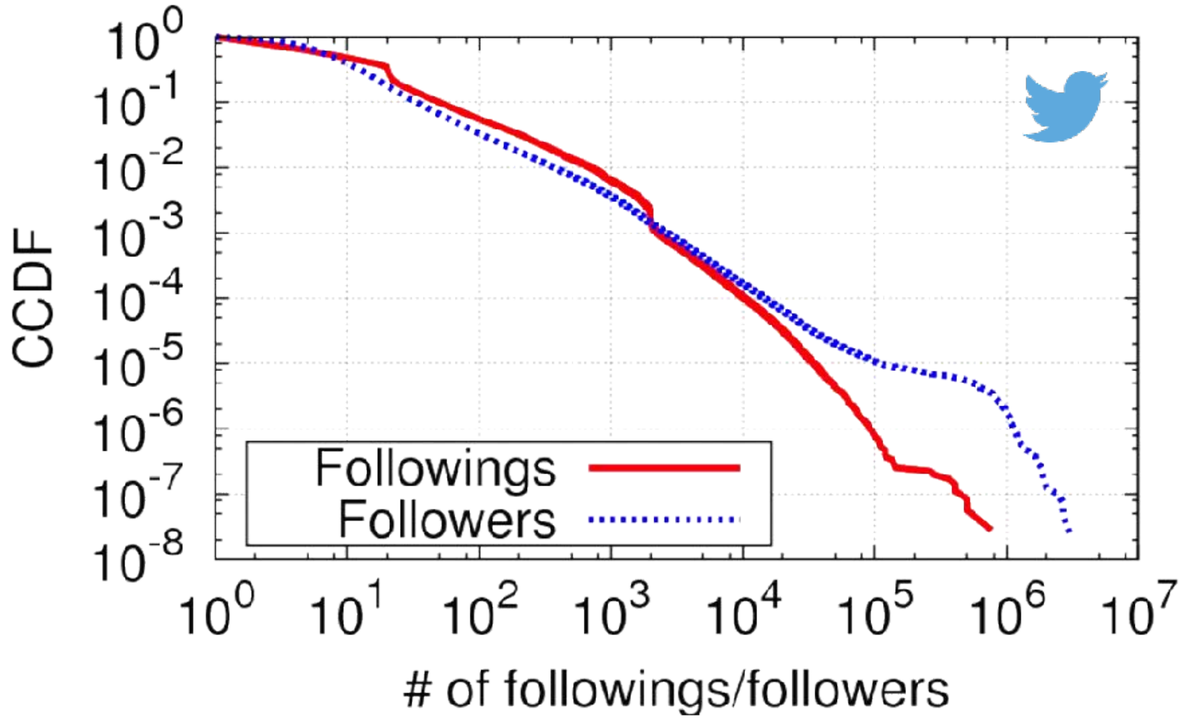
- **61 group datasets from literature**
- **Expanding intimacy circles**
  - Support cliques (5)
  - Sympathy circle (15)
  - Close relationships (50)
  - Stable relationships (150)
  - Acquaintances (500)
  - People we can name (1500)
- **Scaling factor of ~3**



Zhou et al. Discrete hierarchical organization of social group sizes, 2005

Grooming in digital media?

- SNSs may circumvented at least some of the real-life constraints
- Apparently no limit … at least for some people



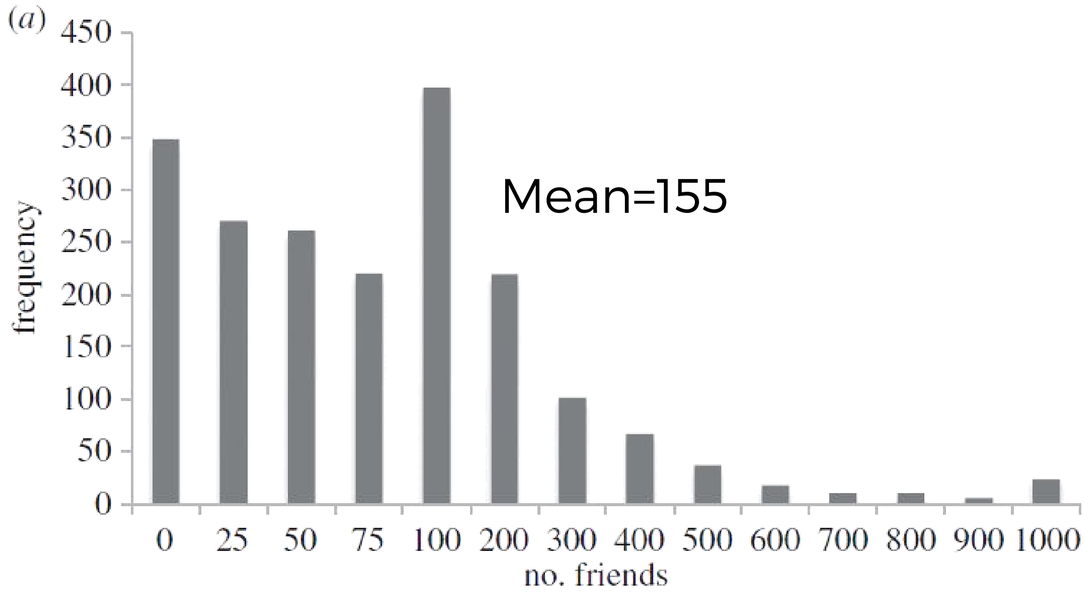
Kwak et al. WWW 2010

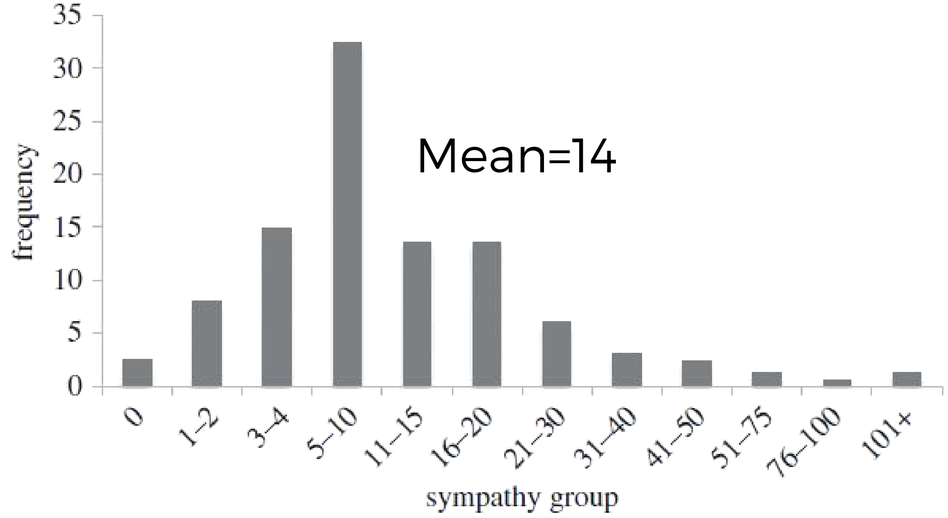
Probing digital barriers

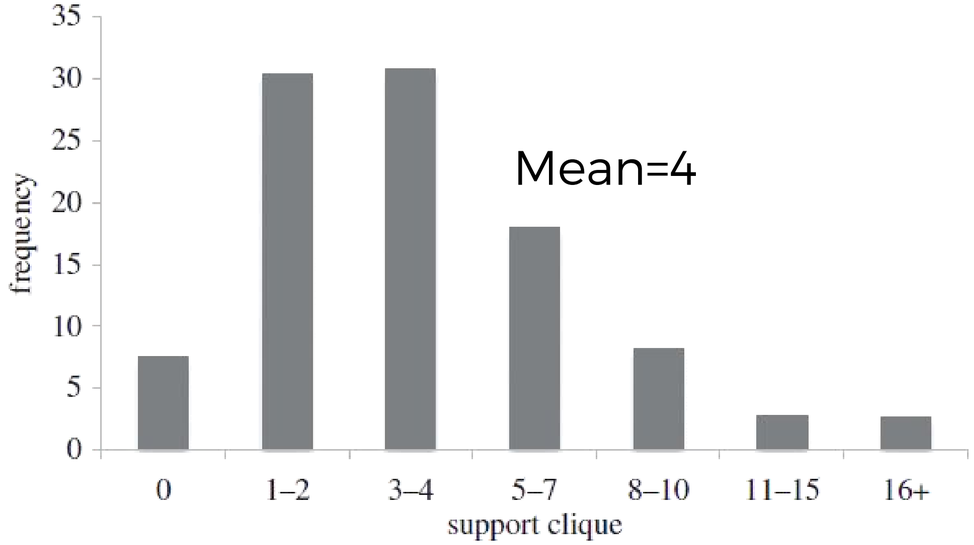
- **Survey to 2,000 UK Facebook users**, stratified sample (age, gender and other cofounding)
- How many Facebook...
  - **Friends [0, 1.000+]**
  - **Close friends [0, 100+]**
  - Support clique members (advice/sympathy in time of distress) [0, 16+]

R. Dunbar. Do online social media cut through the constraints that limit the size of offline social networks?, 2016

SNSs do not circumvent real-life constraints





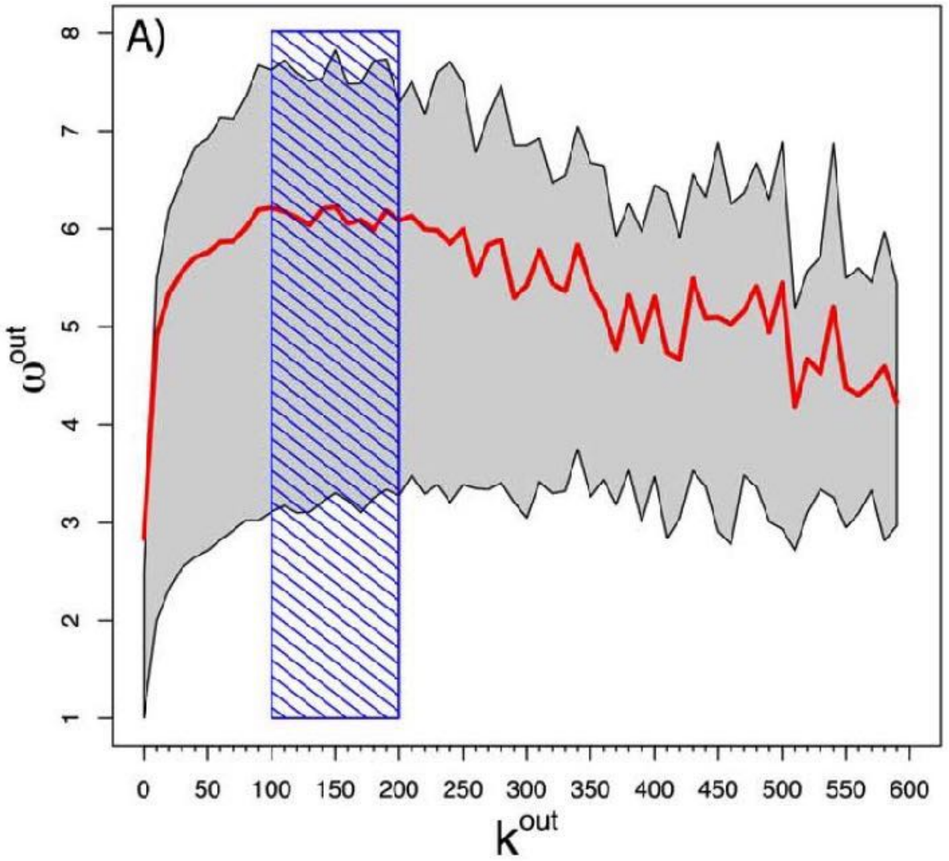


Dunbar number on Twitter

- **Twitter dataset, 380M tweets, 1.7M users**
- **Network of conversations**
- Direct measurement of the interaction strength:

$$\omega_i^{out}(T) \equiv \frac{\sum_j w_{ij}(T)}{k_i^{out}}$$

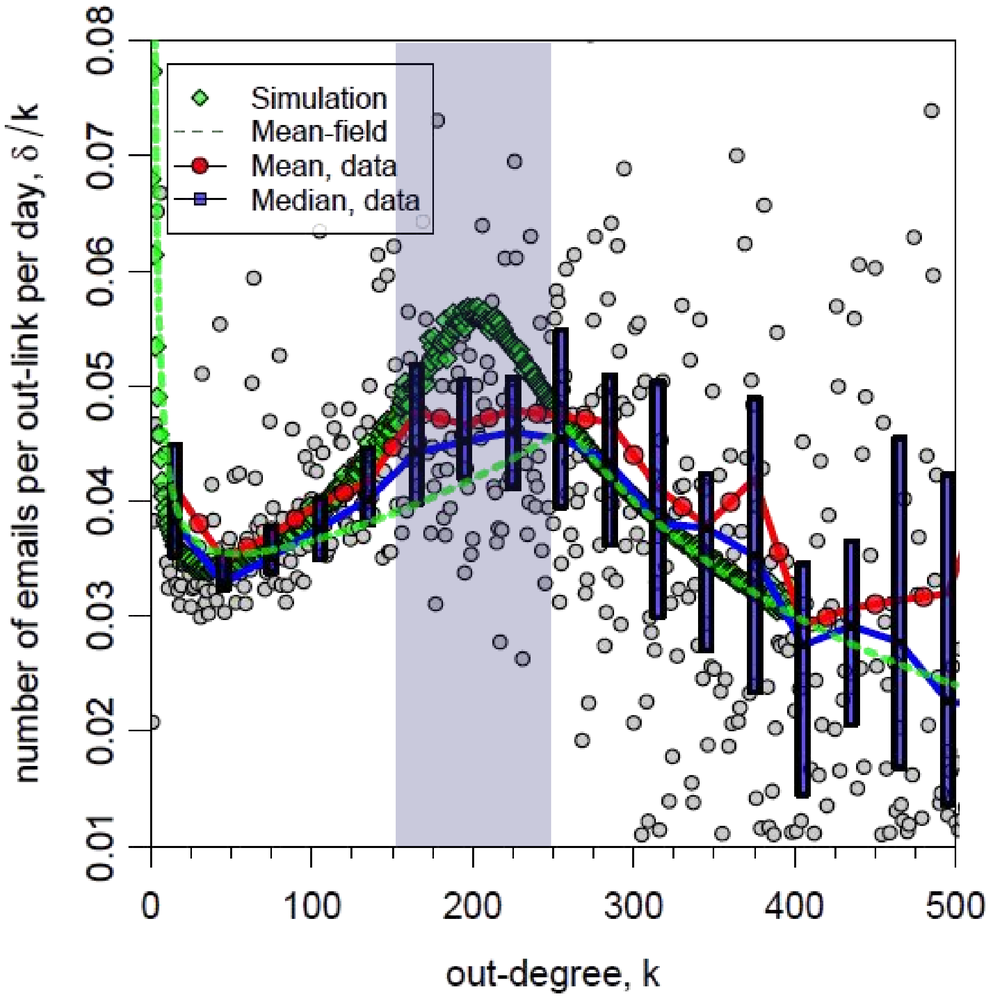
- w_{ij} number of replies
- T time window
- K_{out} reciprocated connections
- Plot ω^{out} (av. link strength) vs. k^{out} (number of bi-links)



B. Gonçalves et al., Validation of Dunbar's number in Twitter conversations 2011

Dunbar number on email

- Email corpus of Univ. Oslo employees and students (35,600)
- ~1M contacts in the outside world

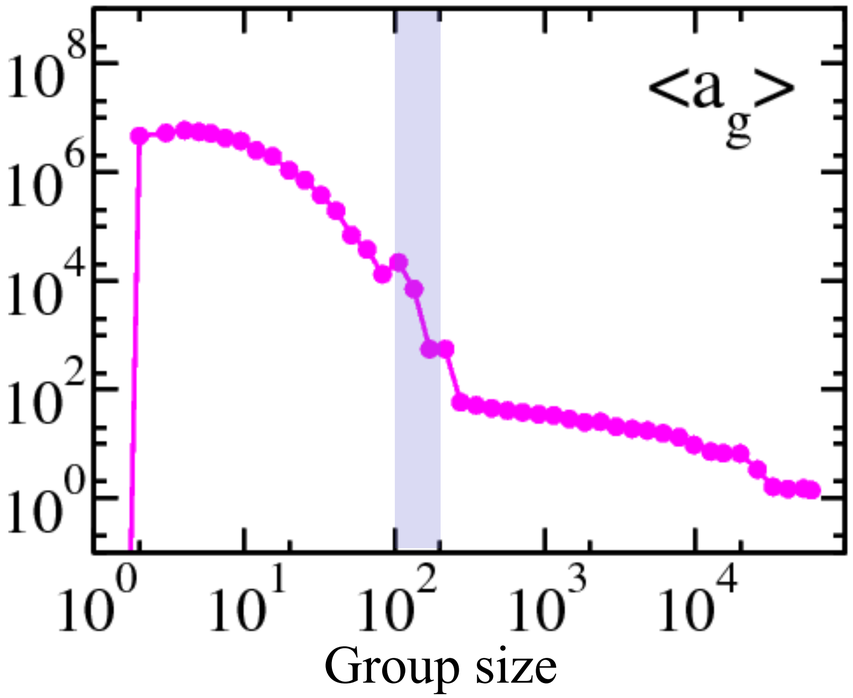


Group perspective

- **Flickr dataset, 238M comments, 504K groups, 71M users**
- **Measure activity inside the group**:

$$a_g = \frac{E_g^{\text{int}}}{(D_g^{\text{in}} D_g^{\text{out}})/E}$$

E=tot #edges
E_g^{int}=#internal group edges
D_g^{in}=#edges incoming to group
D_g^{out}=#edges outgoing from group

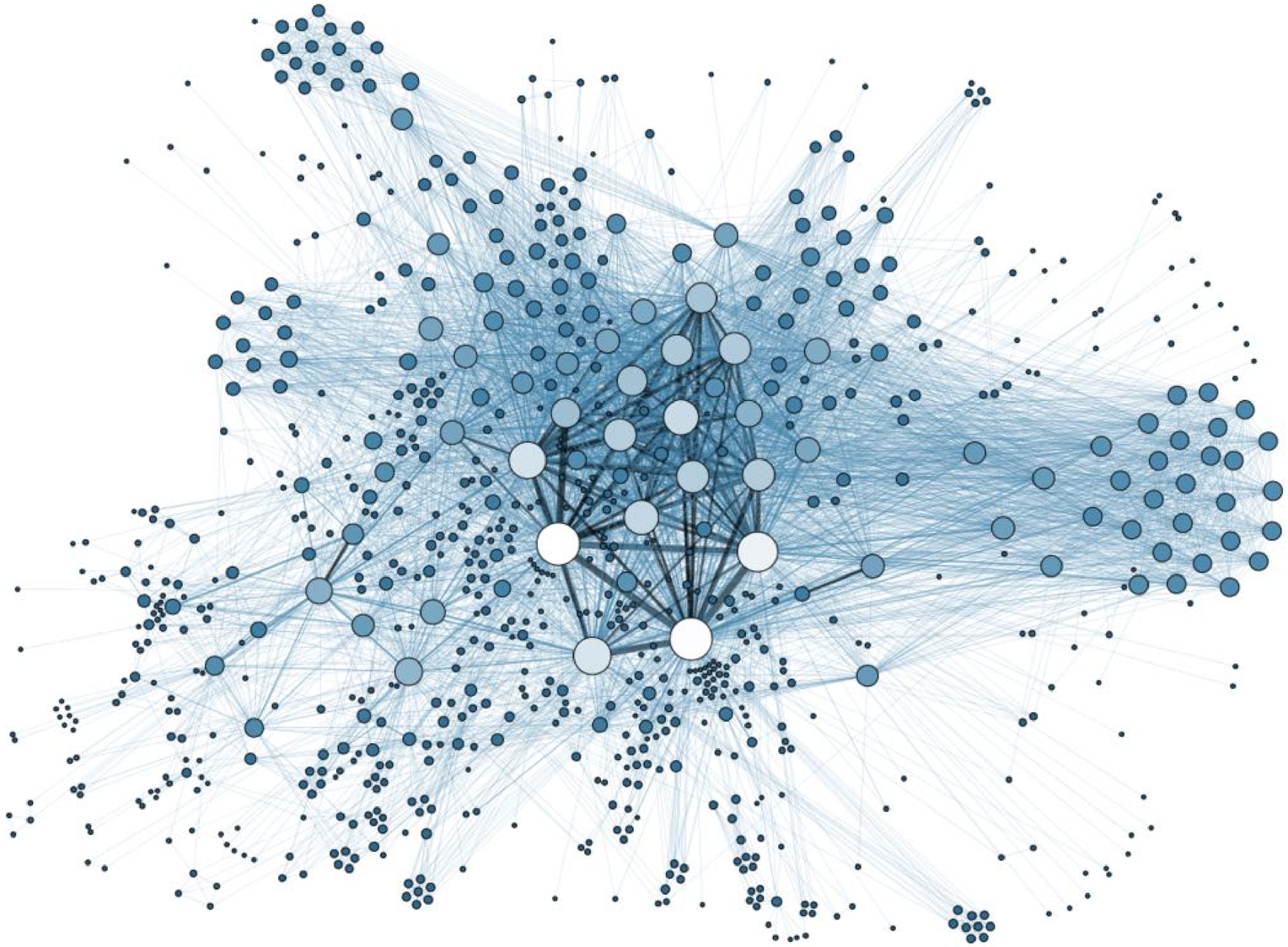


- Plot avg. ag vs. group size

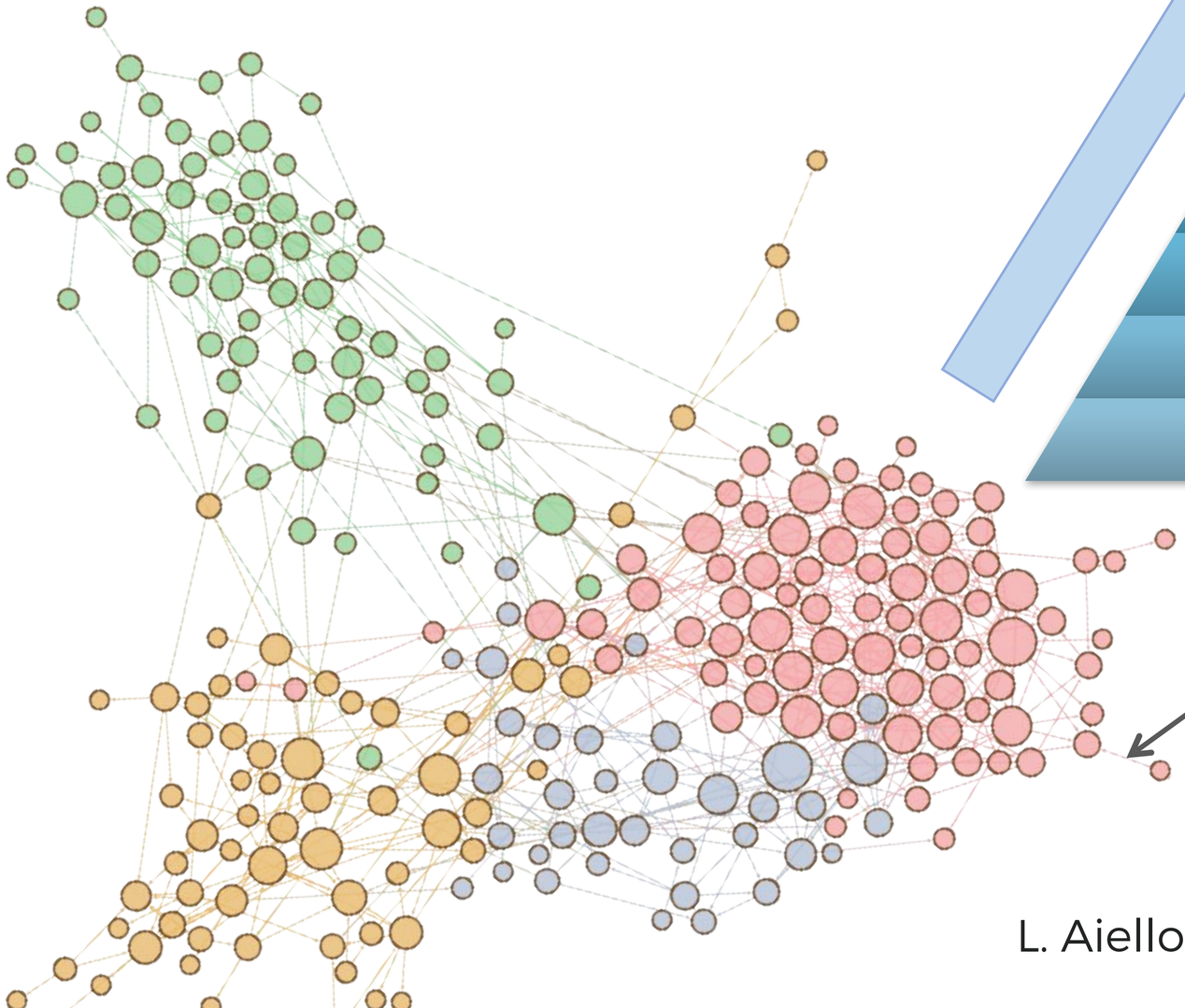
P. Grabowicz et al., Distinguishing topical and social groups..., 2013

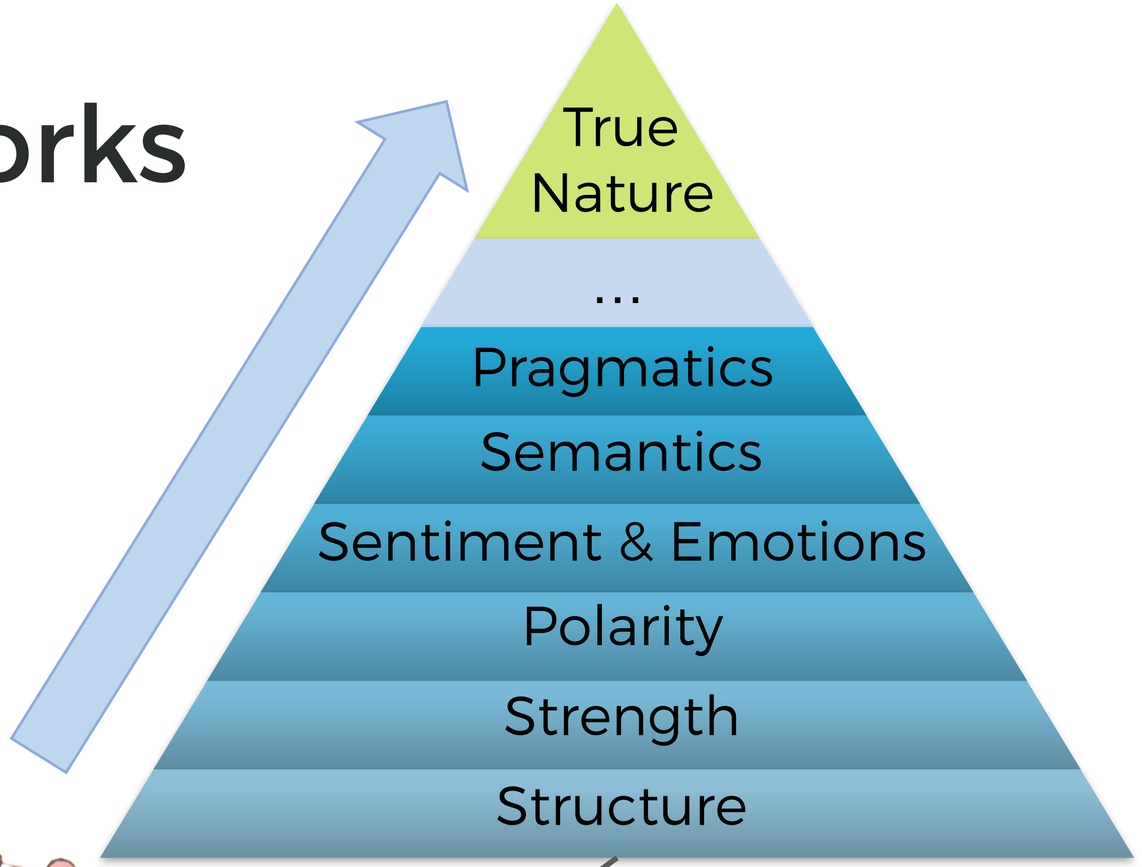
Social Exchange Theory

How we can model conversations?



Modeling social networks





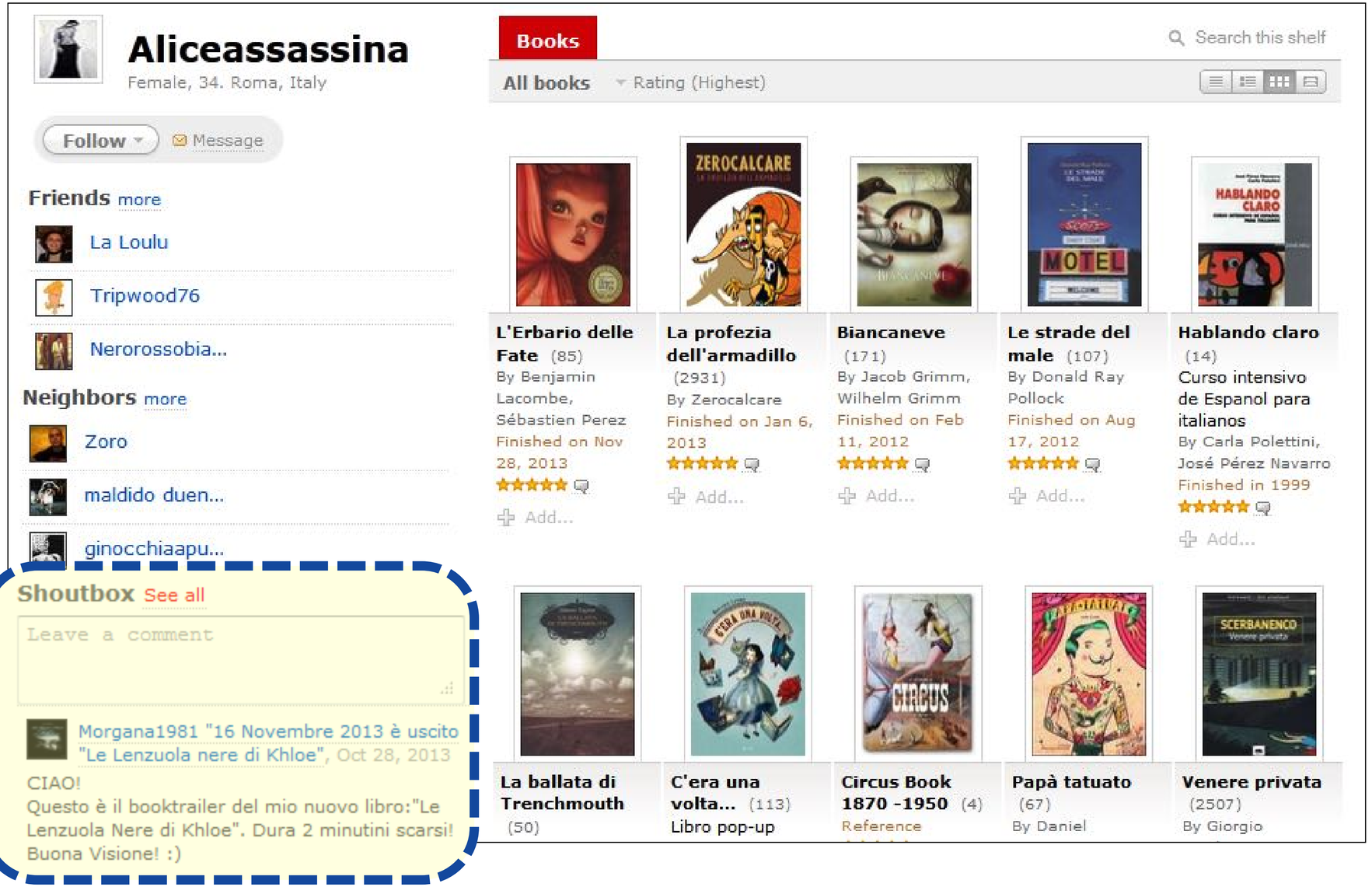
L. Aiello The Nature of Social Structures, 2017

Messages Are Communication Acts That Define the Type of Social Relationship

+PRAGMATICS

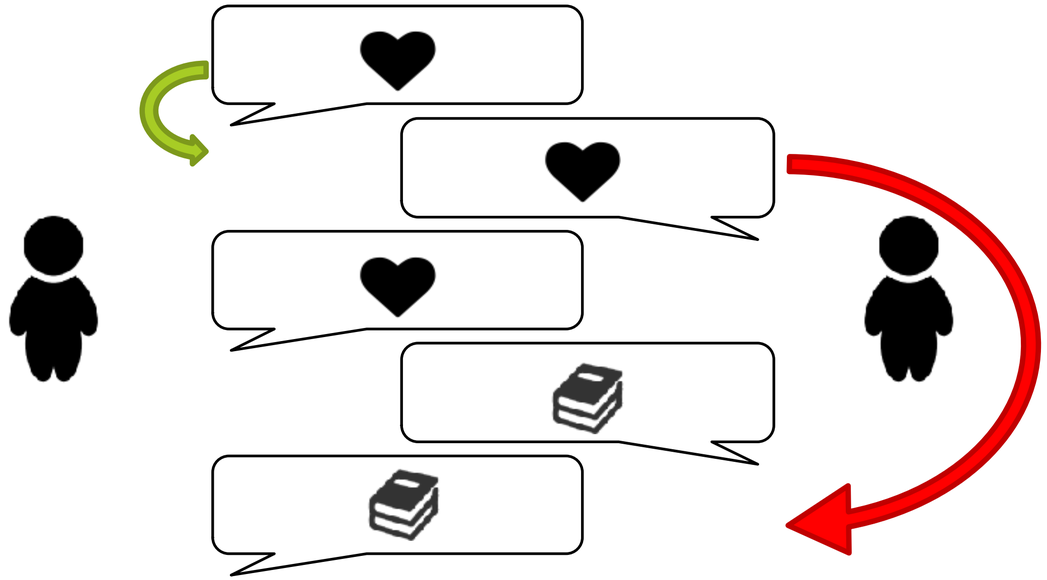
(BEYOND SAYING)

Anobii.com dataset

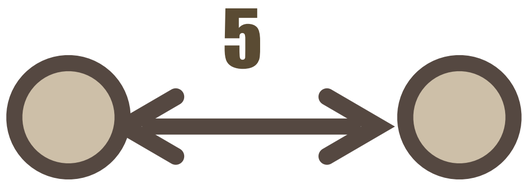


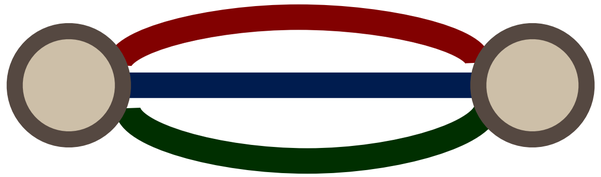
Blau's Social Exchange Theory

Repeated set of **exchanges** of **different types of non-material resources** transacted in an interpersonal situation, such as **knowledge**, **social support** or **manifestation of approval** [1]



[1] P. Blau. Exchange and Power in Social Life. Transaction Publishers, 1964.





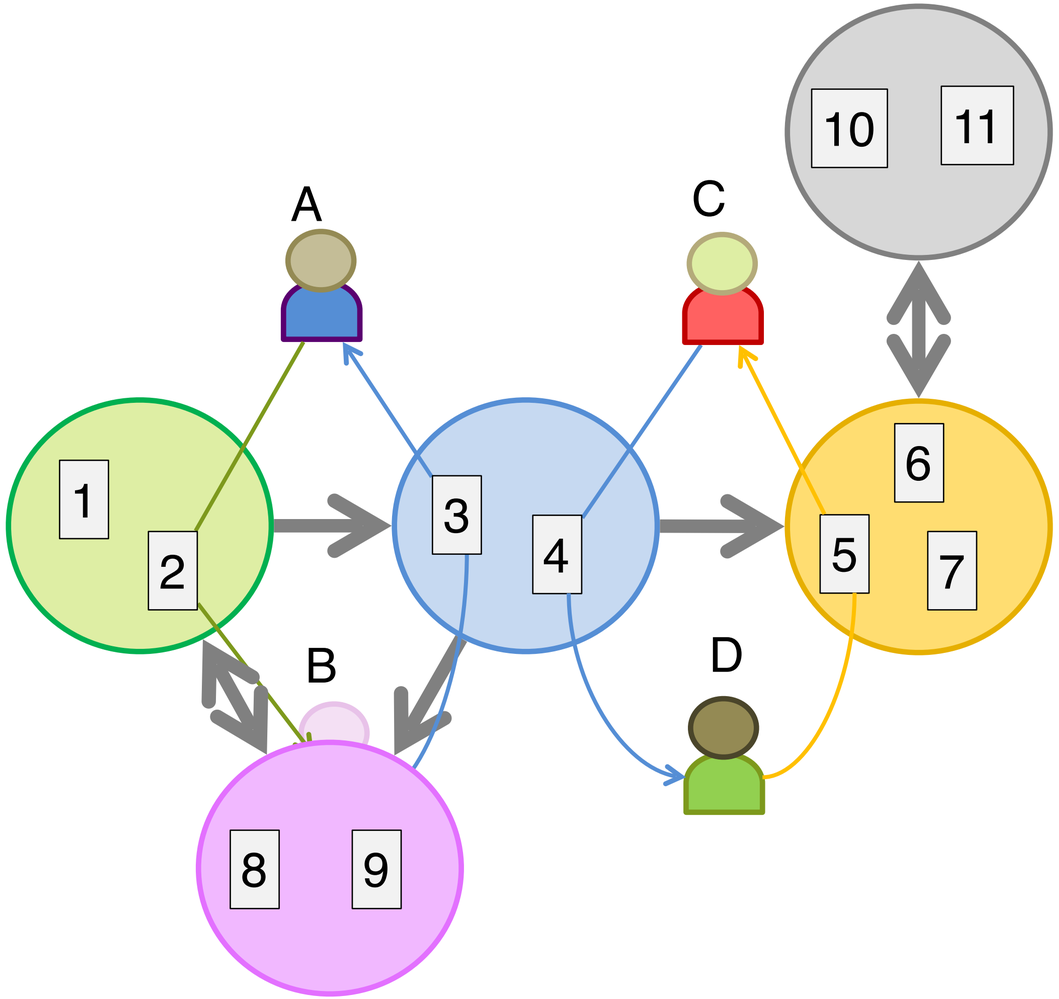
DECOMPOSITION OF A SOCIAL TIE INTO

the resources it conveys

How?

Input: directed comm. multigraph, arcs labeled with time and text
Output: (probabilistic) assignment message -> resource

- **Preprocessing**
  - Stopwords, stemming
- **Message bucketing**
  - NMF, LDA, ...
- **Transition graph**
  - Buckets as nodes transitions as edges
- **Intuition:** conversations tend to stick to the same resource ("You're very good at it" -> "You are pretty good as well")
- **Resource extraction**
  - Community detection on transition graph



Emerging resource types

Knowledge exchange

Technical knowledge of a domain (stackoverflow)

Request for knowledge

"I read a very good review of that book"

Status exchange

Expression of admiration or esteem

Recognition of the partner's higher status

"You are very smart!"

Social support

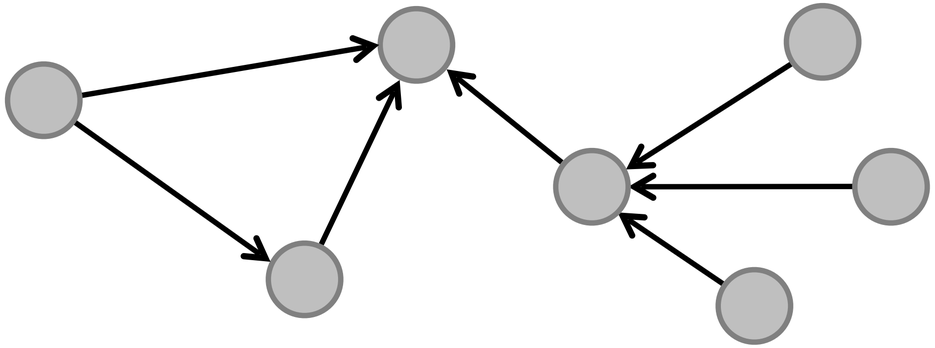
Emotional valuation

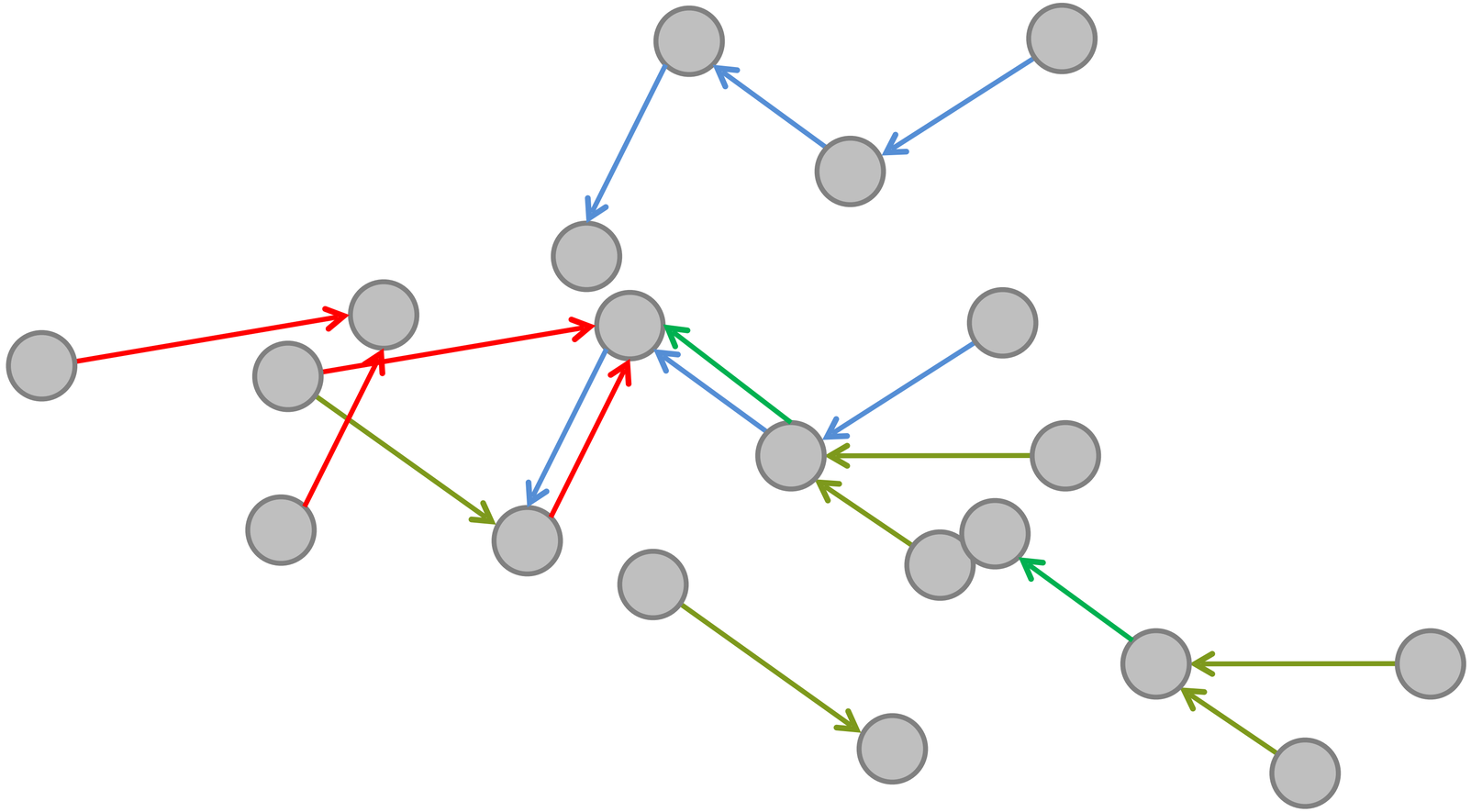
Everyday minute exchanges

"Hope your dad is feeling better now"

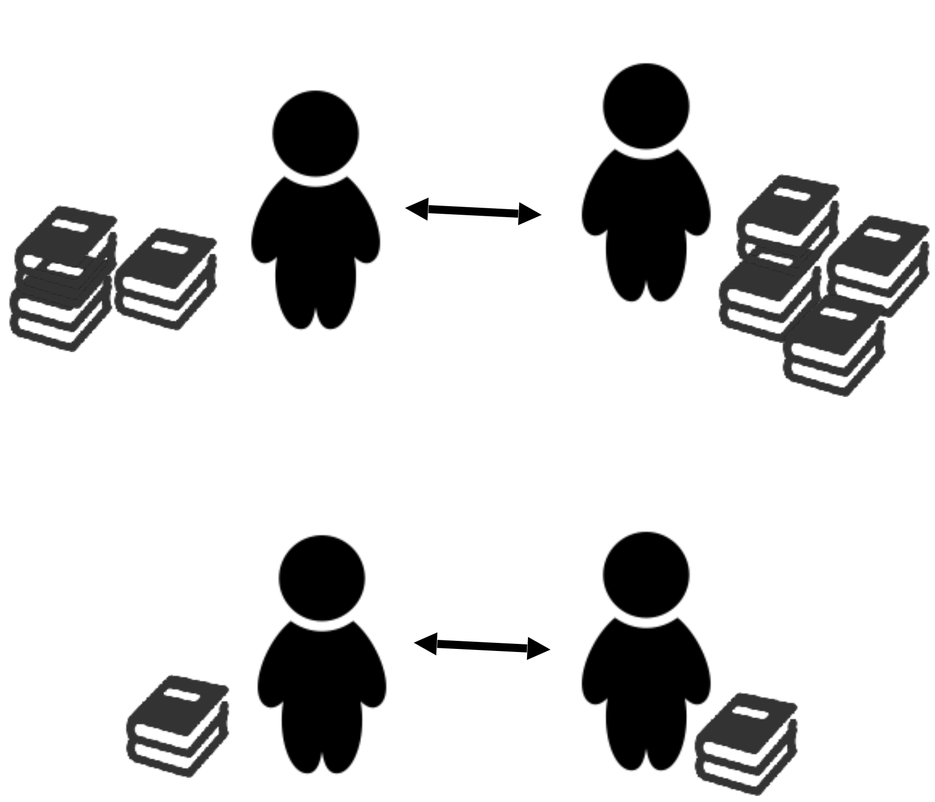
	Most representative tokens
Status	neighbor · library · return · read · bye thanks · much · hi · good read · thanks · interest · add · like · congratulations · visit
Support	good · morning · wish · dear · pass · good day · many · greeting · well · friend · soon · here · done · year · week · kiss
Knowledge	book · read · arrive · see · know · what · now · when · first · end · think · then · advice · only · again · have to · anobii · why

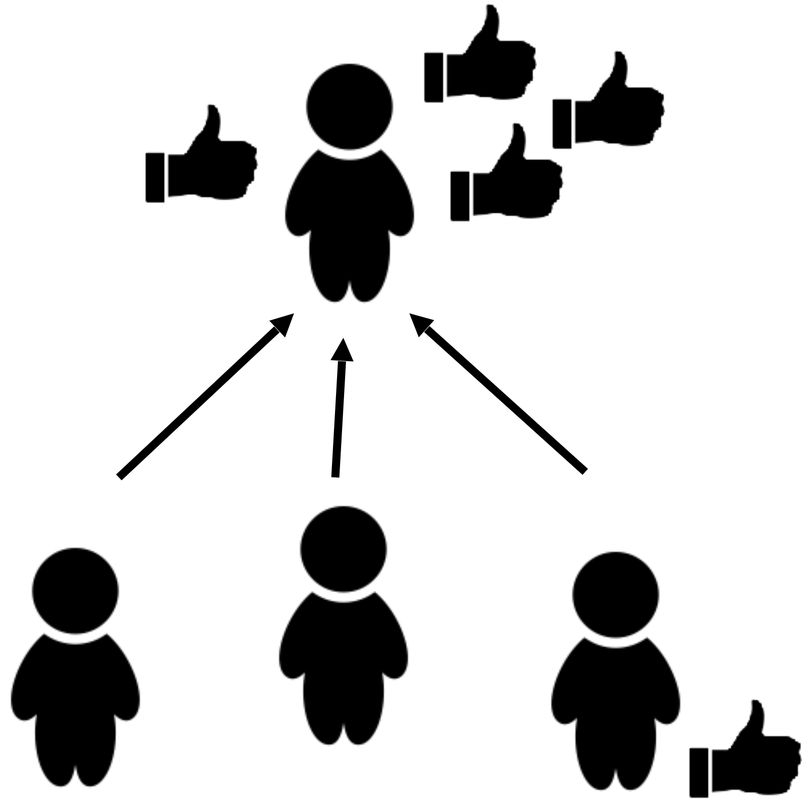
80% of messages are correctly assigned (human coders)



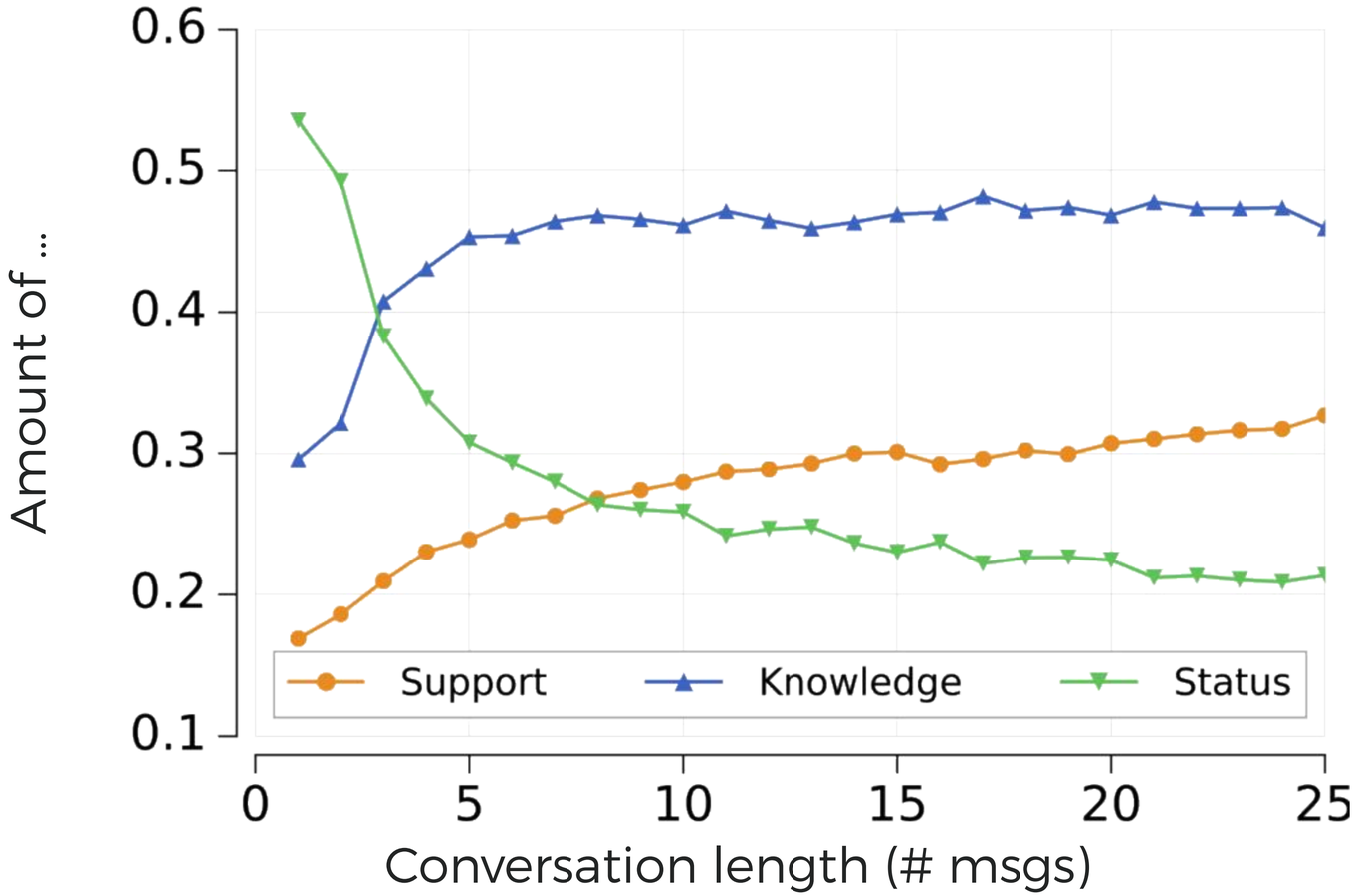


Knowledge net VS. Status net





Resources over time



Aiello et al. Reading the Source code of Social Ties, 2014

Literature review

Literature review

Psychology

Social Psychology

Sociology

Economics

Biology

~1k papers surveyed

Similarity
shared interests, hobbies, motivations, outlooks on life

Trust
high degree of dependability, and reliability

Romance
physical, emotional, and sexual attraction

Social support
high degree of emotional warmth, closeness

Identity
forms of meaningful common group memberships

Status
admiration, respect, prestige

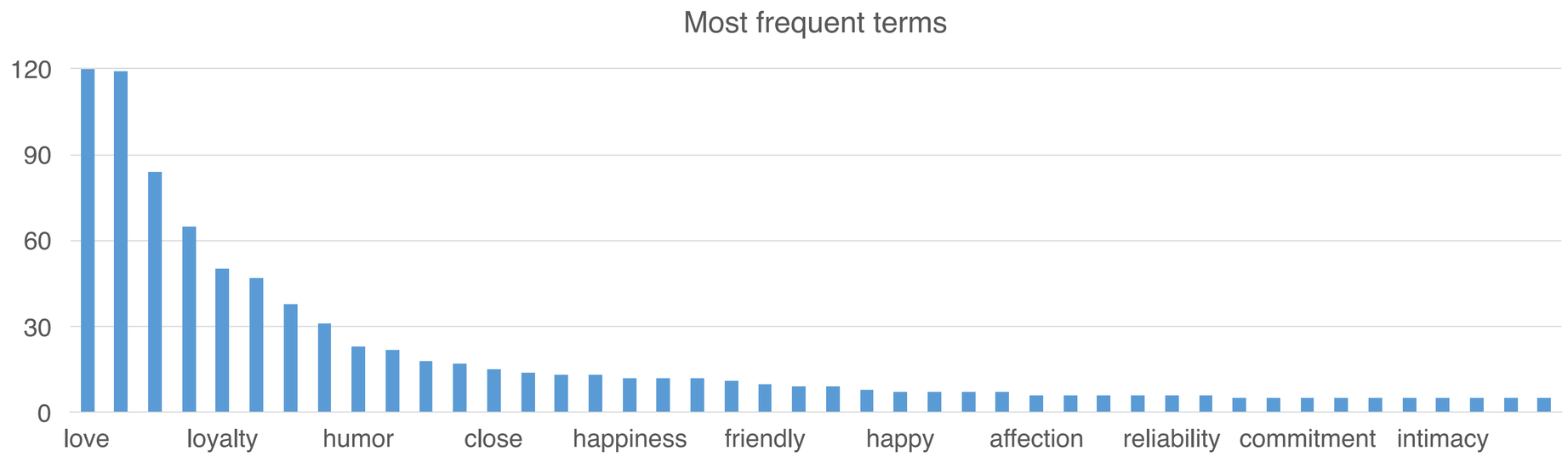
Knowledge transfer
exchange of ideas, learning, and mentorship

Power
having the "upper hand," more resources, imbalance

crowdsourcing

Word-collection on MTurk

“Write the words the [best describe | matter most to you] in a social relationship”



362 Unique words, 151 non-redundant
200 workers. 45% women, 81% White from Canada, UK, US

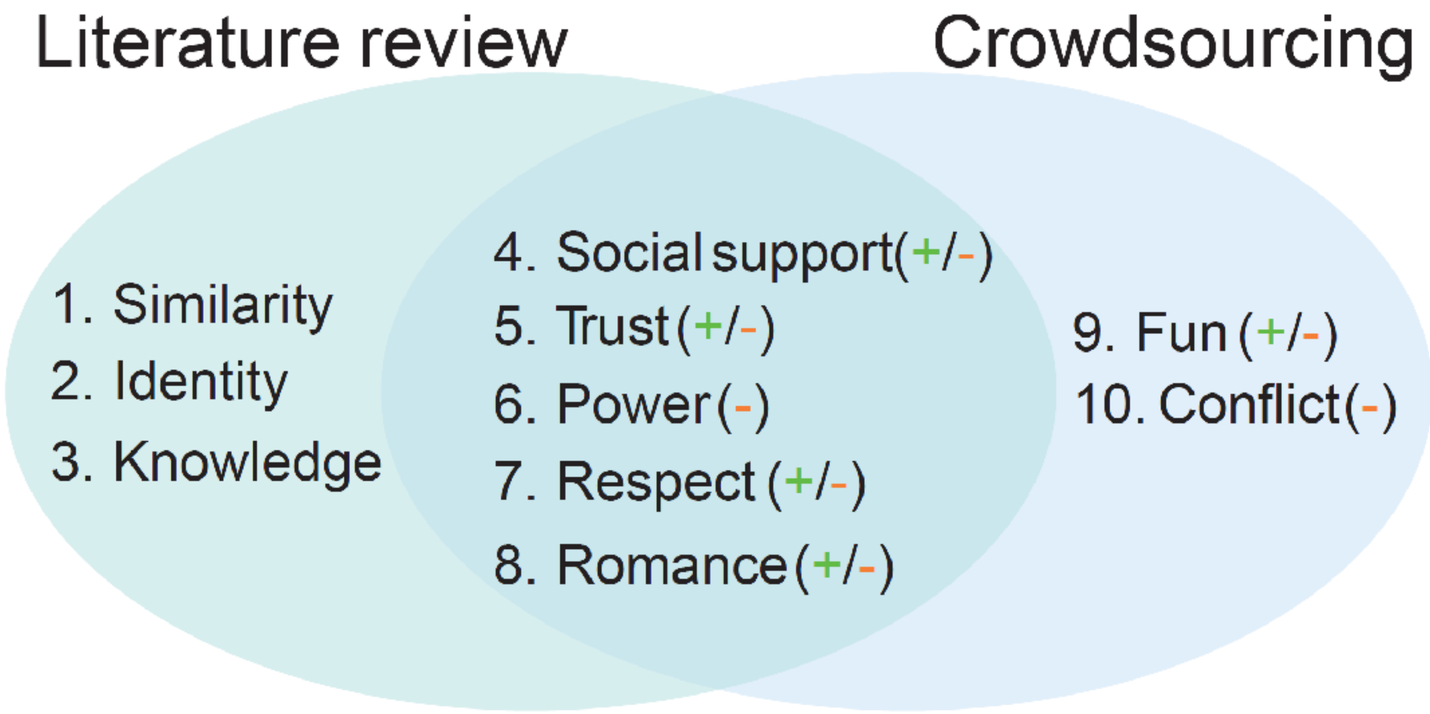
Word-scoring on MTurk

Think about the relationships in your life, in general (e.g. friends, family, coworkers, romantic partners). Now look at the words below. How well do the words below describe those relationships?

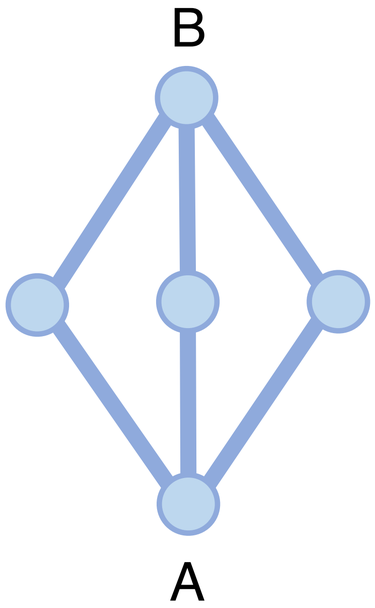
	1 Does not describe well at all	2	3 Describes fairly well	4	5 Describes very well
abusive	○	○	○	○	○
accountability	○	○	○	○	○
active	○	○	○	○	○
admiration	○	○	○	○	○
advantageous	○	○	○	○	○
adventurous	○	○	○	○	○
affection	○	○	○	○	○
amazing	○	○	○	○	○
amicable	○	○	○	○	○
attraction	○	○	○	○	○

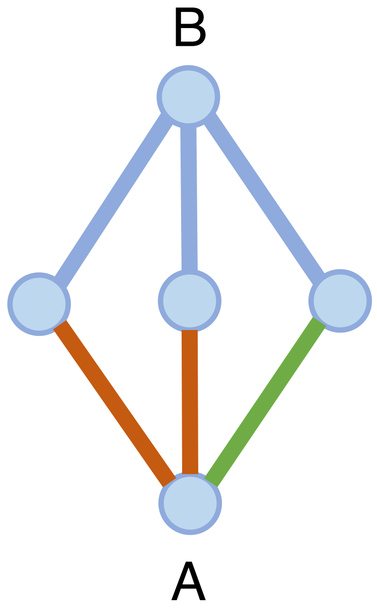
Each word is represented by a 200-dimensional vector of scores (1 entry per worker)

Word clustering



Link prediction





Simple word matching

$$TO(u,v) = \frac{\Gamma_{out}(u) \cap \Gamma_{in}(v)}{|\Gamma_{out}(u)|}$$

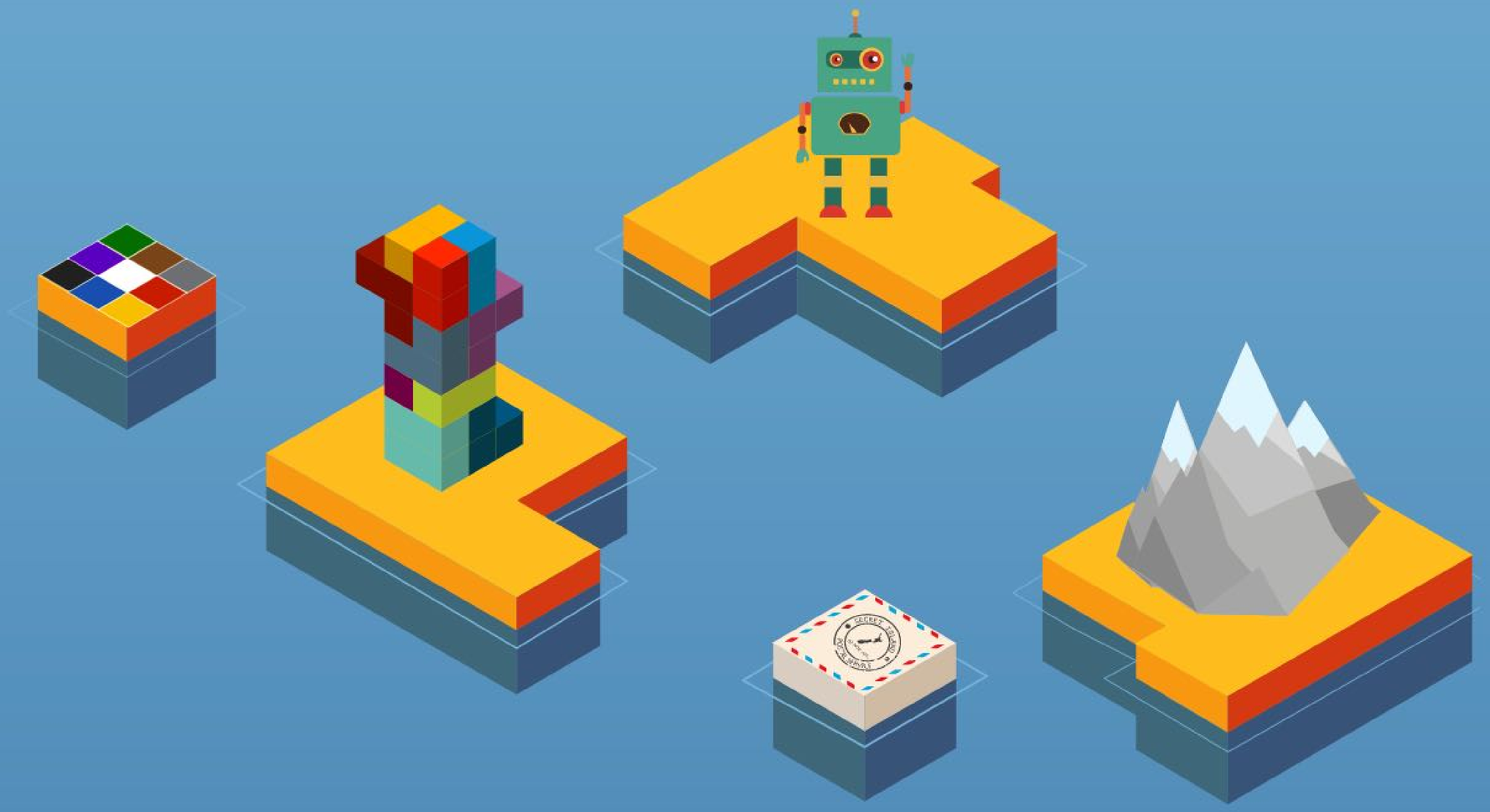
Features	Precision	AUC
Triangle overlap	0.749	0.755
Relationship dimensions	0.800 (+7%)	0.803 (+6%)
All	0.783 (+5%)	0.841 (+11%)

Deri et al. Coloring the social ties, 2018

tinghy

login

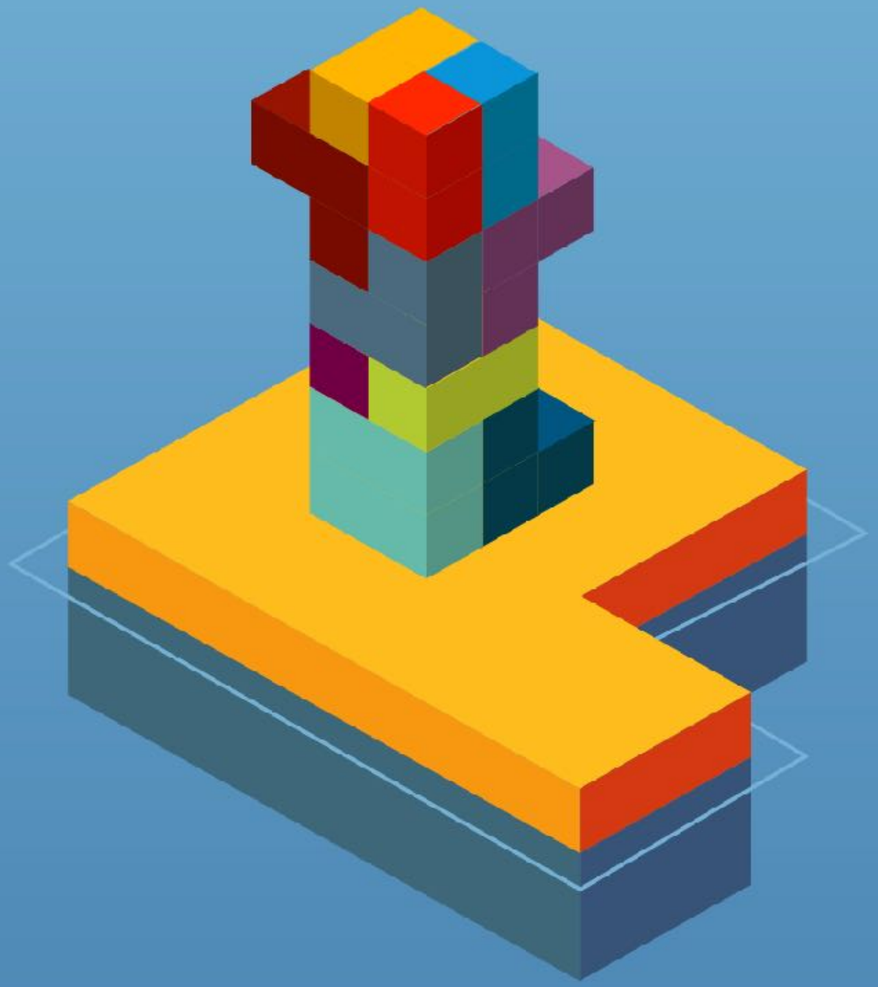
Click on an island and discover who you are



tinghy

Click on an island and discover who you are

login



Isle of Ties

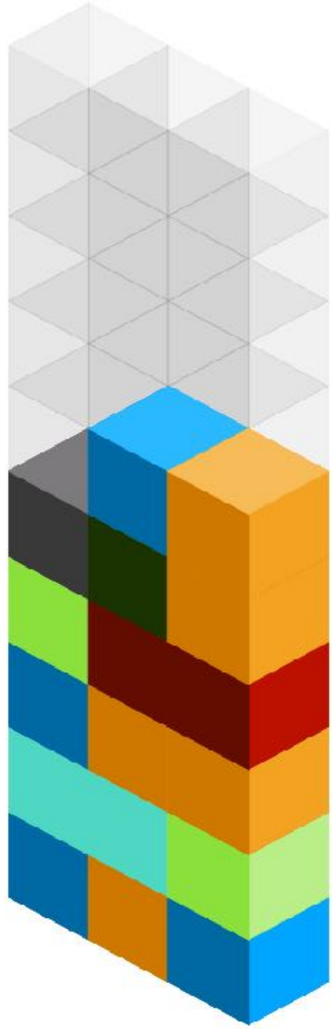
You likely know many people, but what are the relationships in your life really about? Play this game to find out what you really value in your relationships and how you compares to others.

Connect to play

I'd rather play another game.

Pick three blocks describing your relation with *Alessandra Sala*
If you can think about one thing only, then pick the same block multiple times.

- Similarity
- Trust
- Romance
- Social support
- Identity
- Respect
- Knowledge transfer
- Power
- Fun
- Conflict



5/10

Undo

Skip

or restart

Isle of Ties

You likely know many people, but what are the relationships in your life really about? Play this game to find out what you really value in your relationships and how you compares to others.

Questions?

@rschifan

schifane@di.unito.it

http://www.di.unito.it/~schifane