data science: dati, modelli, decisioni

Ciro Cattuto

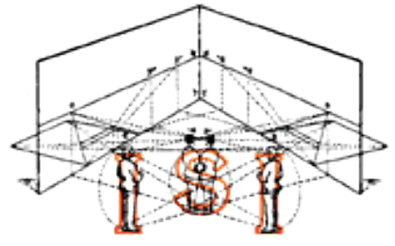
ISI Foundation

ciro.cattuto@isi.it, @ciro

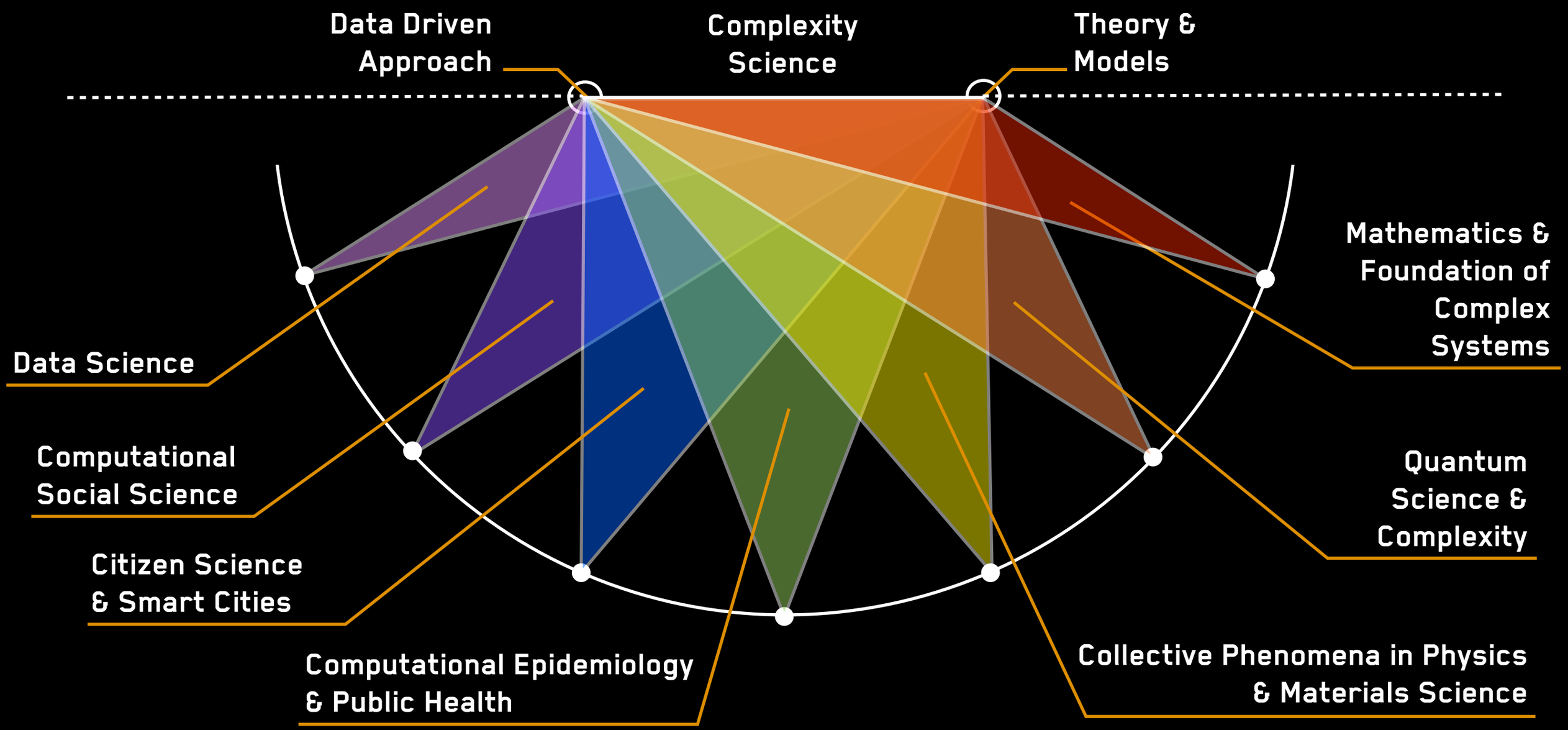
- socio-technical systems
- digital traces, data, metadata, networks
- from data to models to decisions
- challenges and opportunities

Nexa Center for Internet & Society

Torino, October 9th, 2013



ISI Foundation



the big picture

the big picture

▸ **the digital image of the world is tracking the world more and more closely**

• this allows us to use computation to **extract patterns and establish causal inferences** using tools from data mining, machine learning, statistics

• mathematical modeling and forecast now happen on a **data-rich landscape** (e.g., mobility data, OSN data) and are fed by **data streams** from multiple sources

• we can **assess our models against reality** at unprecedented speed and scale, and feed back to models

digital traces of human behavior

digital traces

historical view
temporal horizon
limited reproducibility
limited context
data protection

available as a side effect of many activities
machine processable, pattern discovery
high coverage, can work at scale

methodology

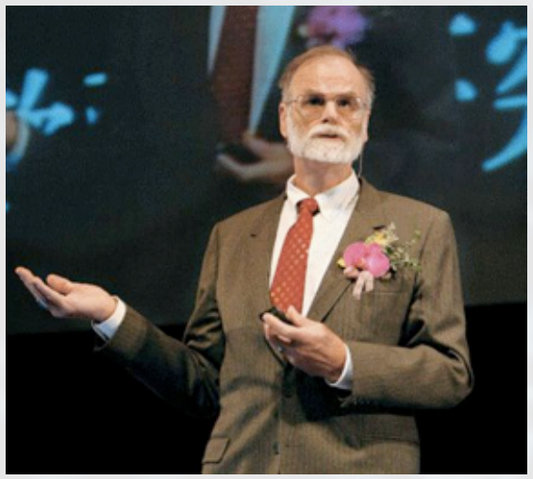
digital traces of human behavior
as first-order objects
for scientific investigation

scientific investigation

complex systems,
network science

data mining,
machine learning,
natural language
processing

digital infrastructures



the 4th paradigm

1. empirical
2. theoretical
3. computational

“The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. ***The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science*** *from computational science as a new, fourth paradigm for scientific exploration.”*

- Jim Gray, 2007

not a paradigm shift

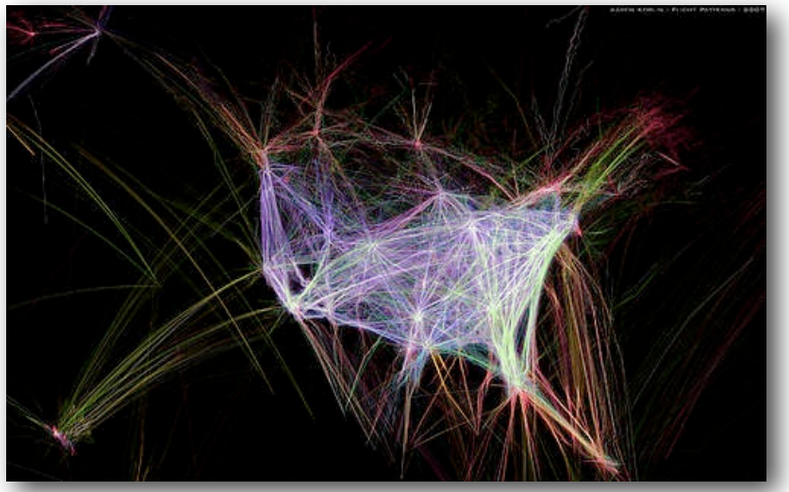
"Thus it is not a paradigm shift in the Kuhnian sense. Data is not sweeping away the old reality. Data is simply placing a set of burdens on the methodologies and social habits we use to deal with and communicate our empiricism and our theory, on the robustness and complexity of our simulations, and on the way we expose, transmit, and integrate our knowledge."

"Data-intensive science, if done right, will mean more paradigm shifts of scientific theory, happening faster, because we can rapidly assess our worldview against the 'objective reality' we can so powerfully measure."

- John Wilbanks (Creative Commons), 2007

the big picture (2)

complex systems science



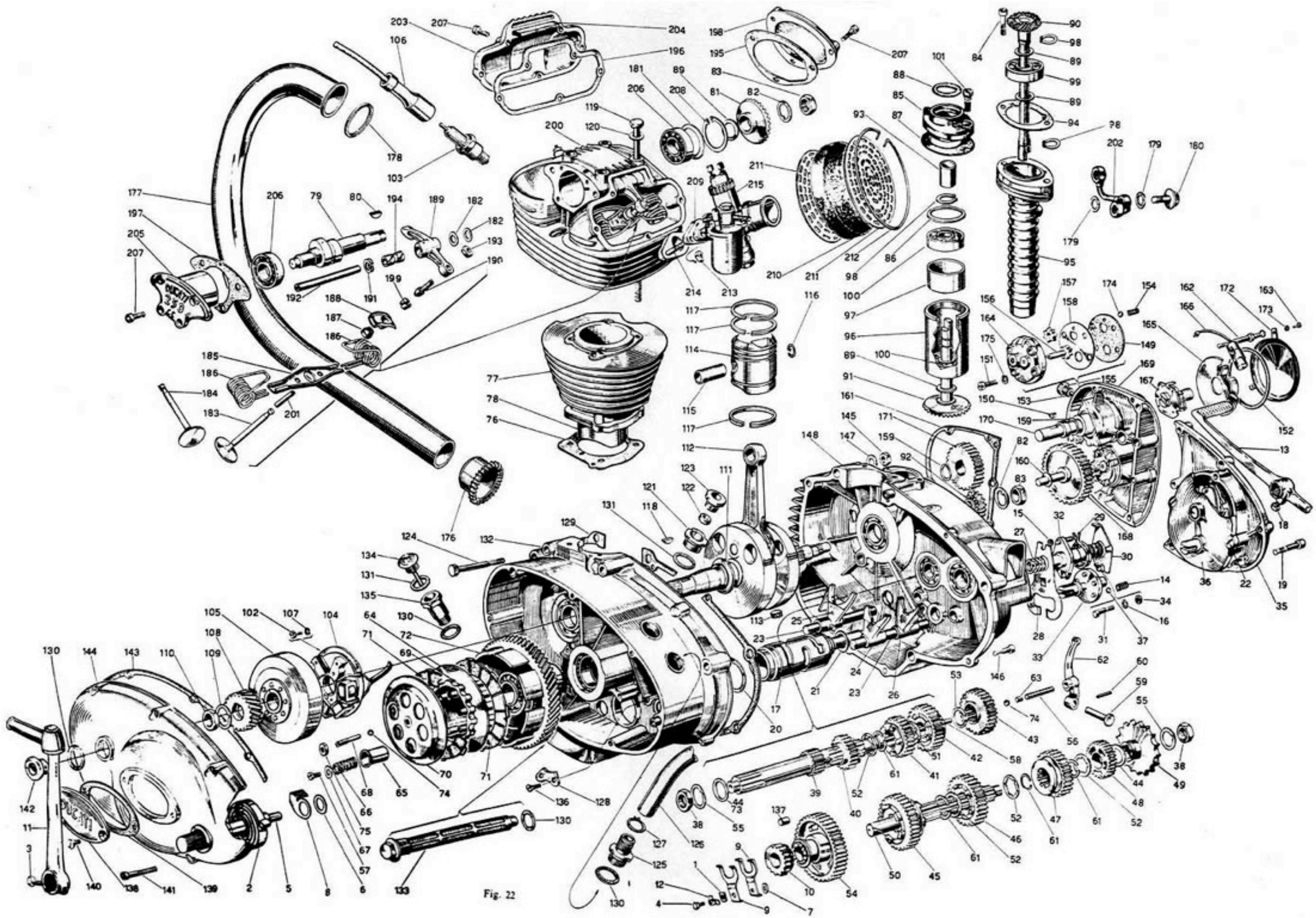
A. Koblin

P. Butler

- ✓ large number of components
- ✓ interactions between components
- ✓ multi-scale hierarchical structures
- ✓ coupling between scales
- ✓ self-organization (*no blueprint*)
- ✓ emergent properties
- ✓ “complex” is more than “complicated”

- ★ the end of linear thinking
- ★ interdependence and systemic risk
- ★ the problem of causal inference

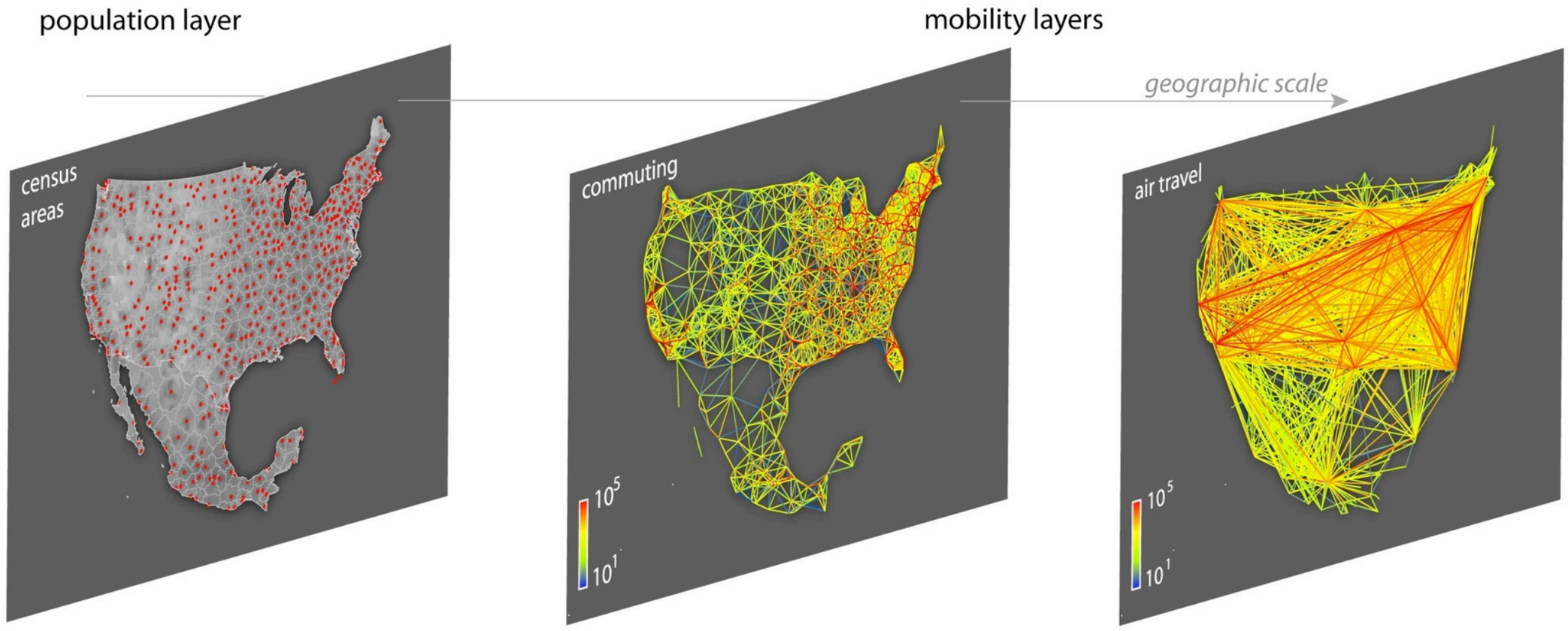
complex ?



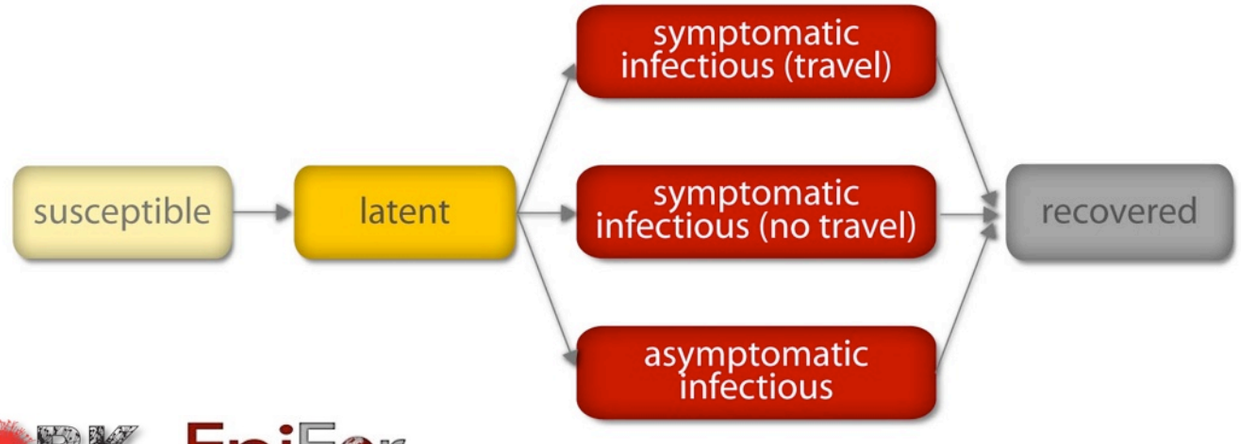
http://www.ducatimeccanica.com/single_engine.jpg

EXAMPLE: GLOBAL EPIDEMIC FORECAST

D.Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, A. Vespignani Multiscale mobility networks and the spatial spreading of infectious diseases **Proc Natl Acad Sci U S A** 106, 21484-21489 (2009).



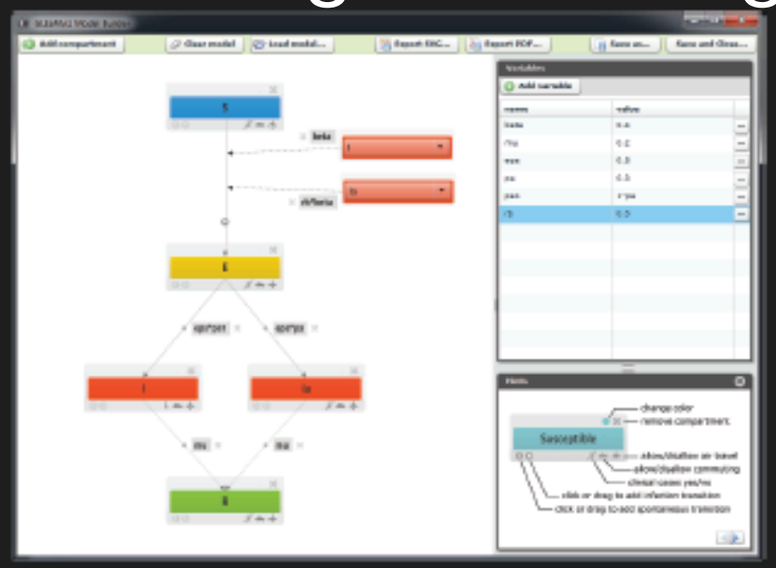
epidemic layer

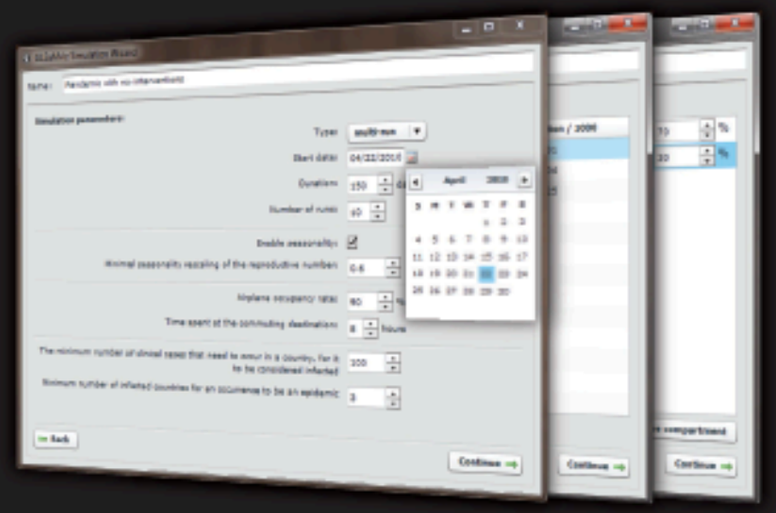


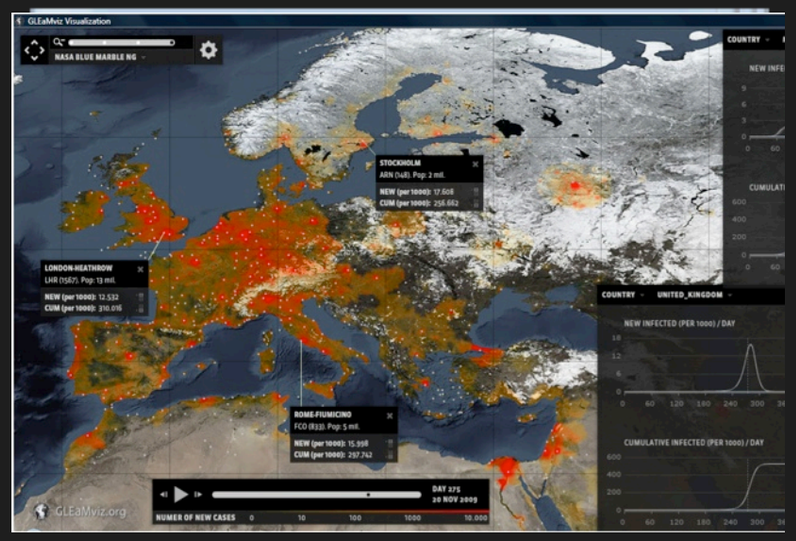
Parameter	Value	Description
β	from R_0	transmission probability
ε^{-1}	1.9 [1.1-2.5] d	average latency period
μ^{-1}	3 [3-5] d	average infectious period
p_t	50%	probability of traveling for infectious individuals
p_a	33%	probability of being asymptomatic
r_β	50%	relative infectiousness of asymptomatic infectious individuals

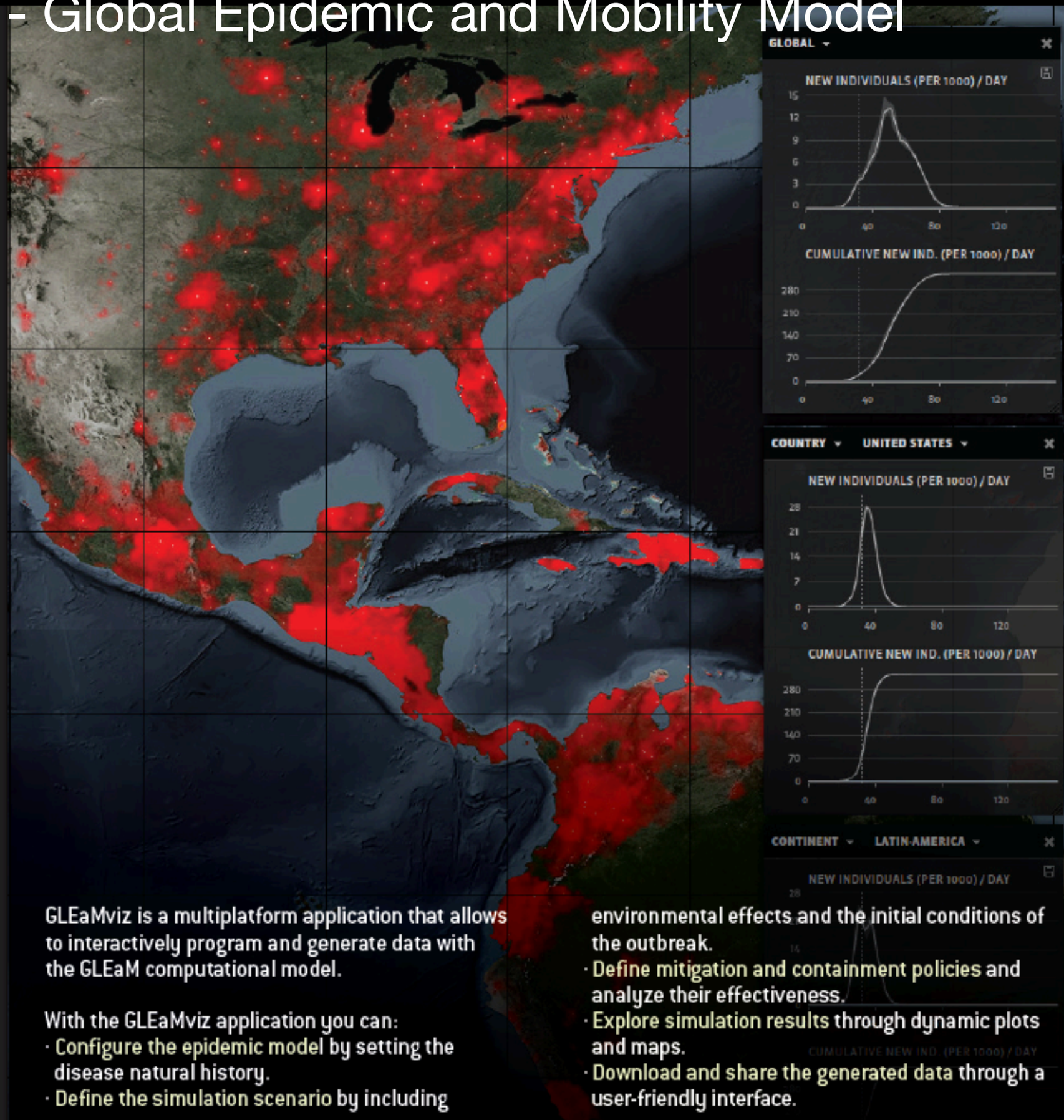
EPIWORK EpiFor

gleamviz.org - Global Epidemic and Mobility Model









GLEaMviz is a multiplatform application that allows to interactively program and generate data with the GLEaM computational model.

With the GLEaMviz application you can:

- Configure the epidemic model by setting the disease natural history.
- Define the simulation scenario by including environmental effects and the initial conditions of the outbreak.
- Define mitigation and containment policies and analyze their effectiveness.
- Explore simulation results through dynamic plots and maps.
- Download and share the generated data through a user-friendly interface.

EPIWORK EpiFor

citizen as (flu) sensors

deGroteGriepMeting.nl
B-NL

Influenzanet

www.influensakoll.se
Kartlägger förekomst och spridning av influensa i Sverige

flusurvey

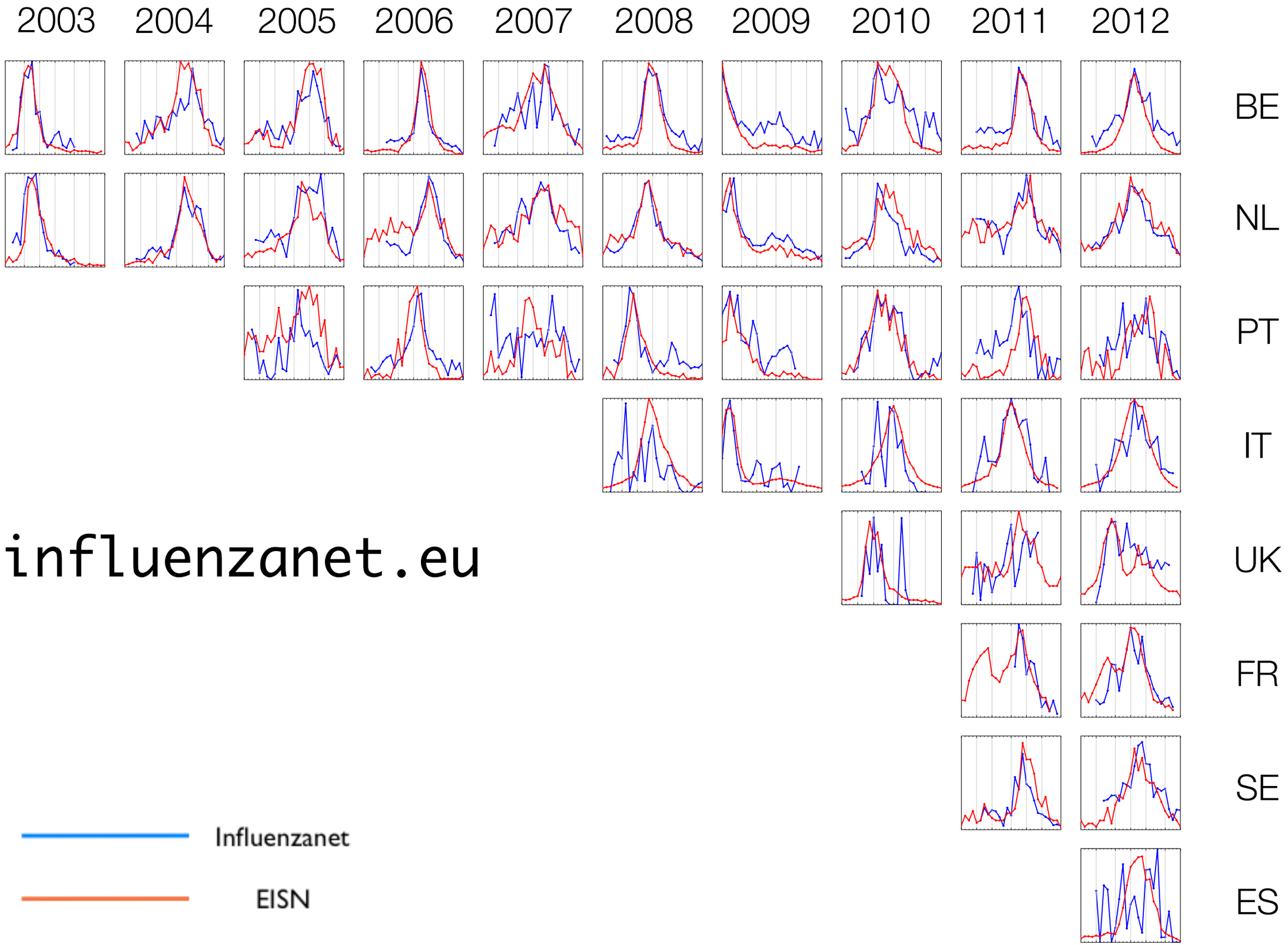
grippenet.fr

gripenet

INFLUWEB

GripeNet.es

Europe-wide participatory surveillance



influenzanet.eu

Influenzanet

EISN

Every Aware

Enhance Environmental Awareness through Social Information Technologies

EVERYAWARE

AIRPROBE International Challenge

Turin

21 October – 01 December

PARTICIPANTS NEEDED

www.everyaware.eu/APIC

ISI Foundation

SAPIENZA UNIVERSITÀ DI ROMA

CSP

UCL

vito vision on technology

L3S Forschungszentrum · Research Center

data & metadata

DEPRECATED

The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

Tweet's creation date.

The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.

The screen name & user ID of replied to tweet author.

Truncated to 140 characters. Only possible from SMS.

The author's user ID.

The author's user name.

The author's biography.

The author's screen name.

The author's URL.

The author's "location". This is a free-form text field, and there are no guarantees on whether it can be geocoded.

Rendering information for the author. Colors are encoded in hex values (RGB).

The creation date for this account.

Whether this account has contributors enabled (http://bit.ly/50npuu).

Number of favorites this user has.

Number of users this user is following.

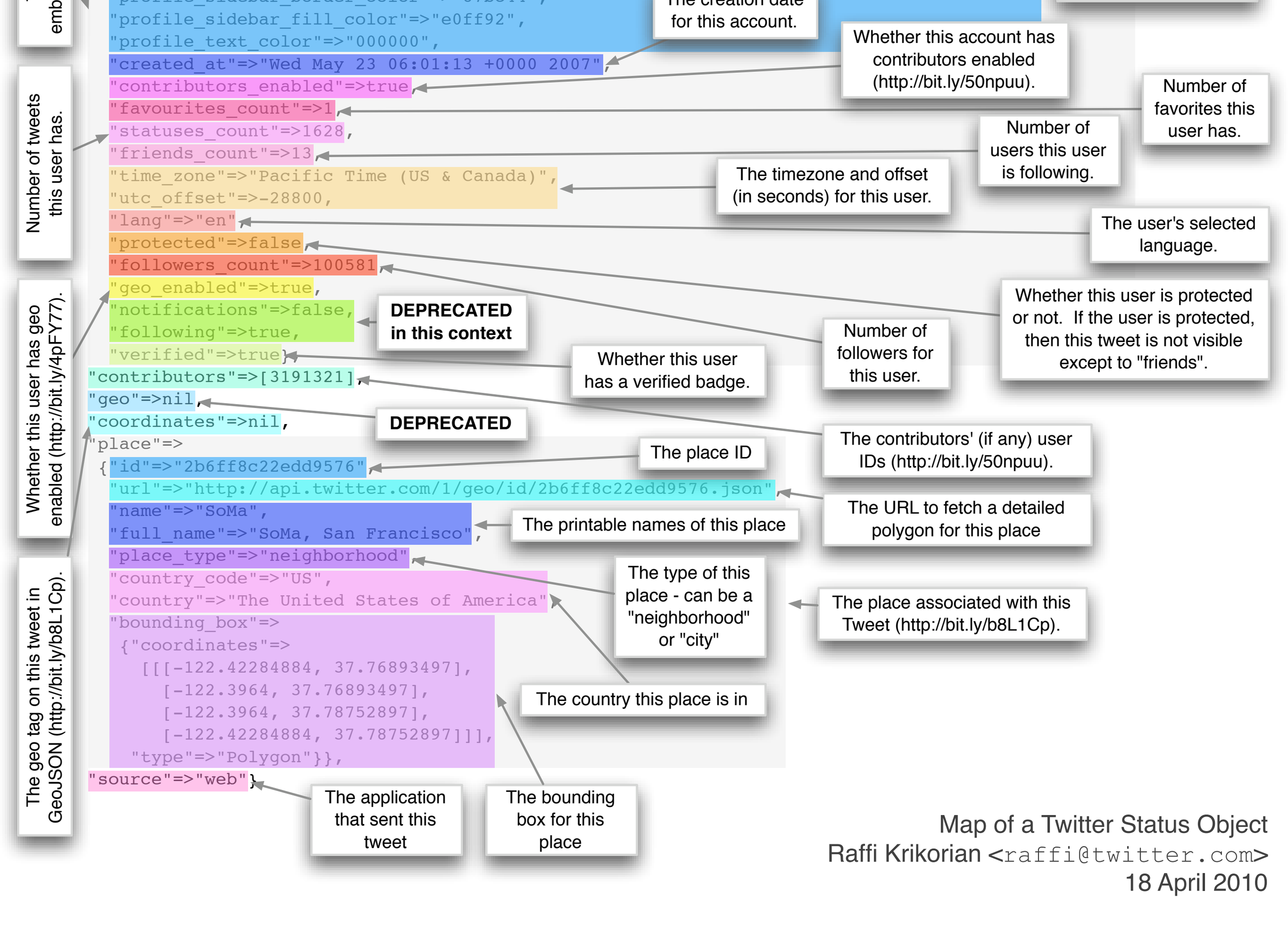
The timezone and offset (in seconds) for this user.

The user's selected language.

The author of the tweet. This is an embedded object can get out of sync.

Number of tweets this user has.

```
{"id"=>12296272736,
 "text"=>
 "An early look at Annotations:
  http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa5da2608865453",
 "created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
 "in_reply_to_user_id"=>nil,
 "in_reply_to_screen_name"=>nil,
 "in_reply_to_status_id"=>nil
 "favorited"=>false,
 "truncated"=>false,
 "user"=>
  {"id"=>6253282,
   "screen_name"=>"twitterapi",
   "name"=>"Twitter API",
   "description"=>
   "The Real Twitter API. I tweet about API changes, service issues and
    happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
   "url"=>"http://apiwiki.twitter.com"
   "location"=>"San Francisco, CA"
   "profile_background_color"=>"c1dfee",
   "profile_background_image_url"=>
   "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
   "profile_background_tile"=>false,
   "profile_image_url"=>"http://a3.twimg.com/profile_images/689684365/api_normal.png",
   "profile_link_color"=>"0000ff",
   "profile_sidebar_border_color"=>"87bc44",
   "profile_sidebar_fill_color"=>"e0ff92",
   "profile_text_color"=>"000000",
   "created_at"=>"Wed May 23 06:01:13 +0000 2007",
   "contributors_enabled"=>true,
   "favourites_count"=>1
   "statuses_count"=>1628,
   "friends_count"=>13
   "time_zone"=>"Pacific Time (US & Canada)",
   "utc_offset"=>-28800,
   "lang"=>"en"
   "protected"=>false
   "followers_count"=>100581
```



Map of a Twitter Status Object

Raffi Krikorian <raffi@twitter.com>

18 April 2010

http://flickr.com

Explore / Tags / nyc


From ... Arjun


From u n c o m m o n


From The Visions of Kai


From orgutcayli


From CiaoChessa

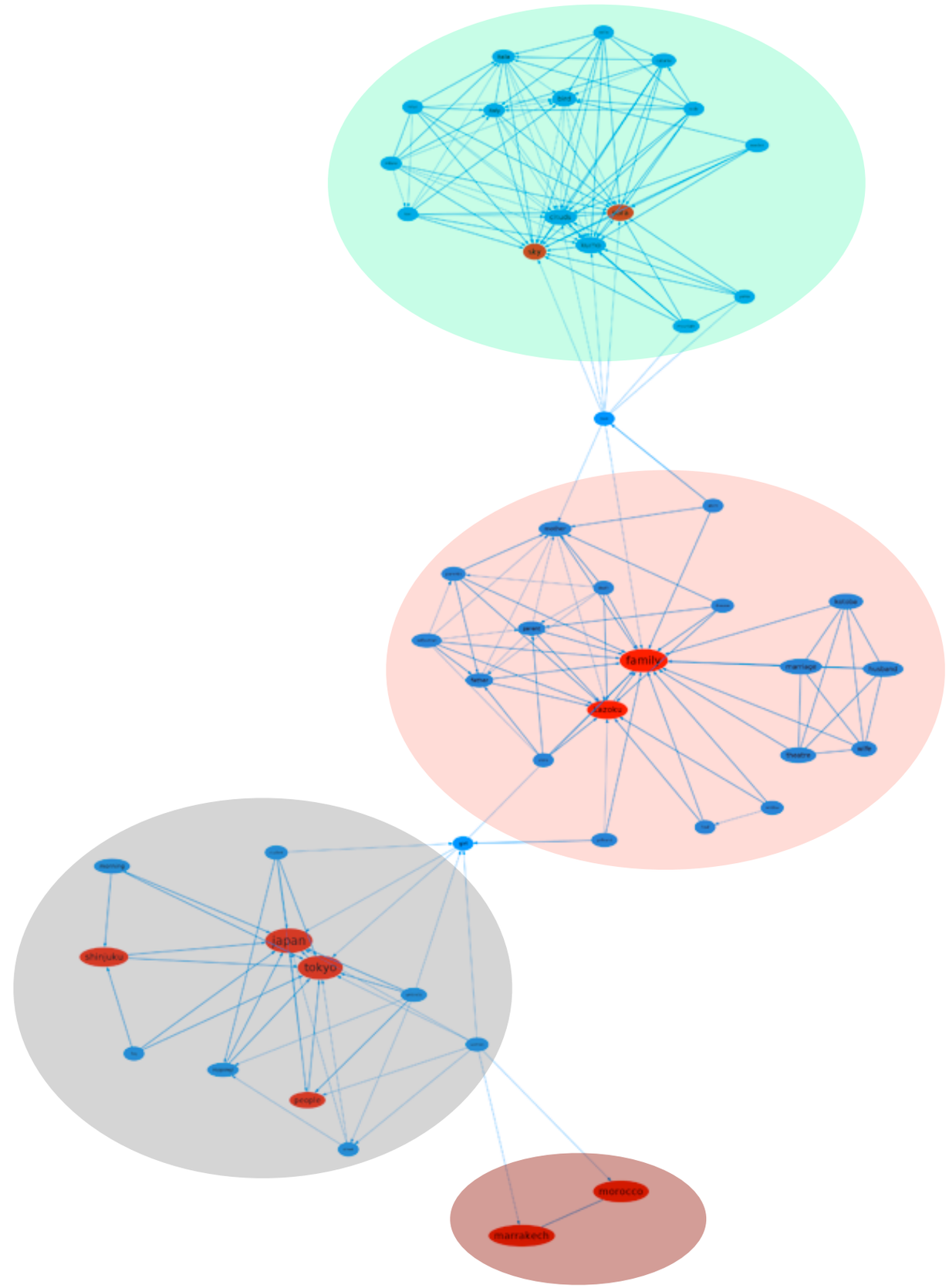

From FrizzText


From CiaoChessa


From C-Monster


From Sir Francis Canker...

tag co-occurrence

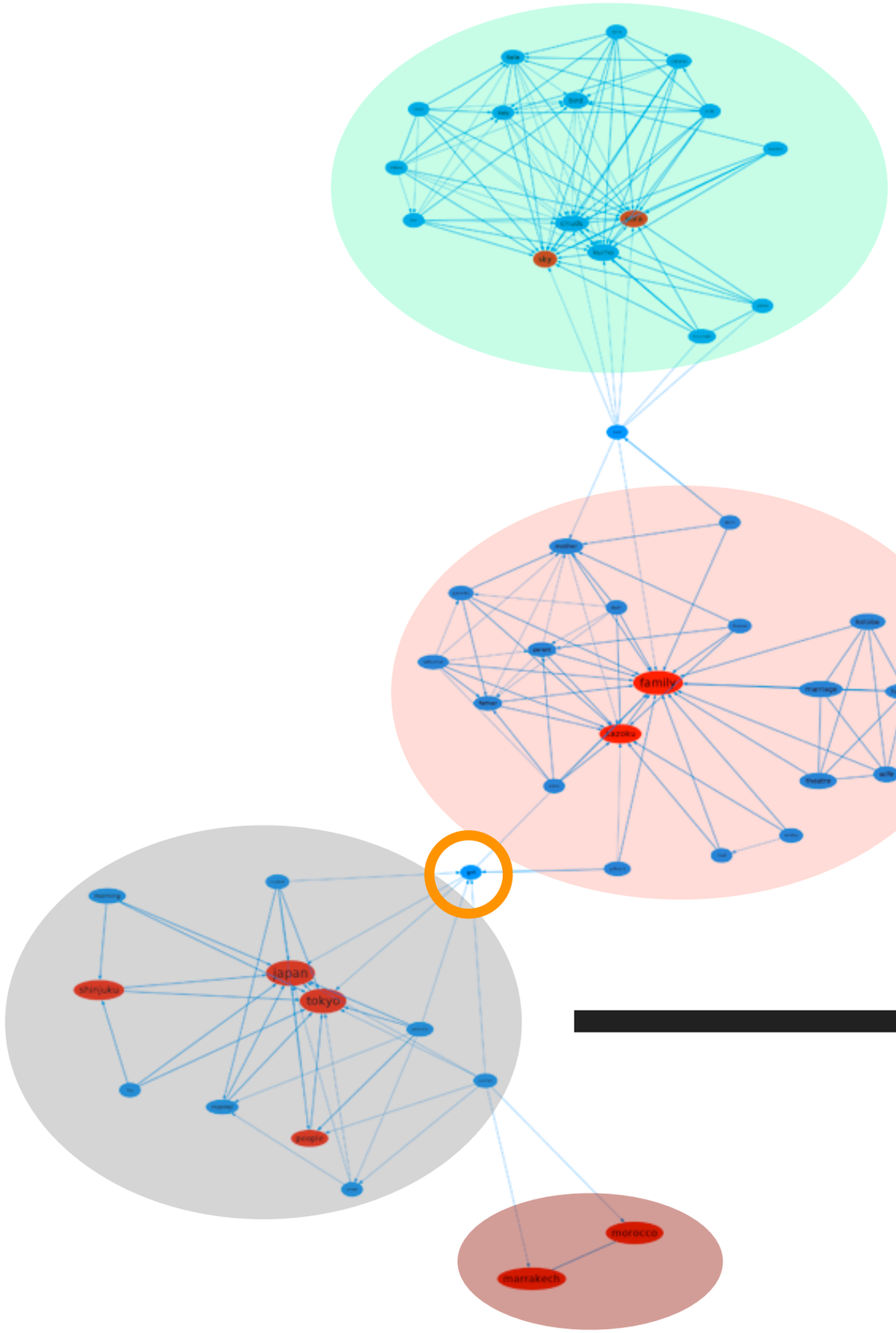


umbrella, red, tokyo, japan, 傘

sushi, tokyo, cuisine, japan

airport, tokyo, japan

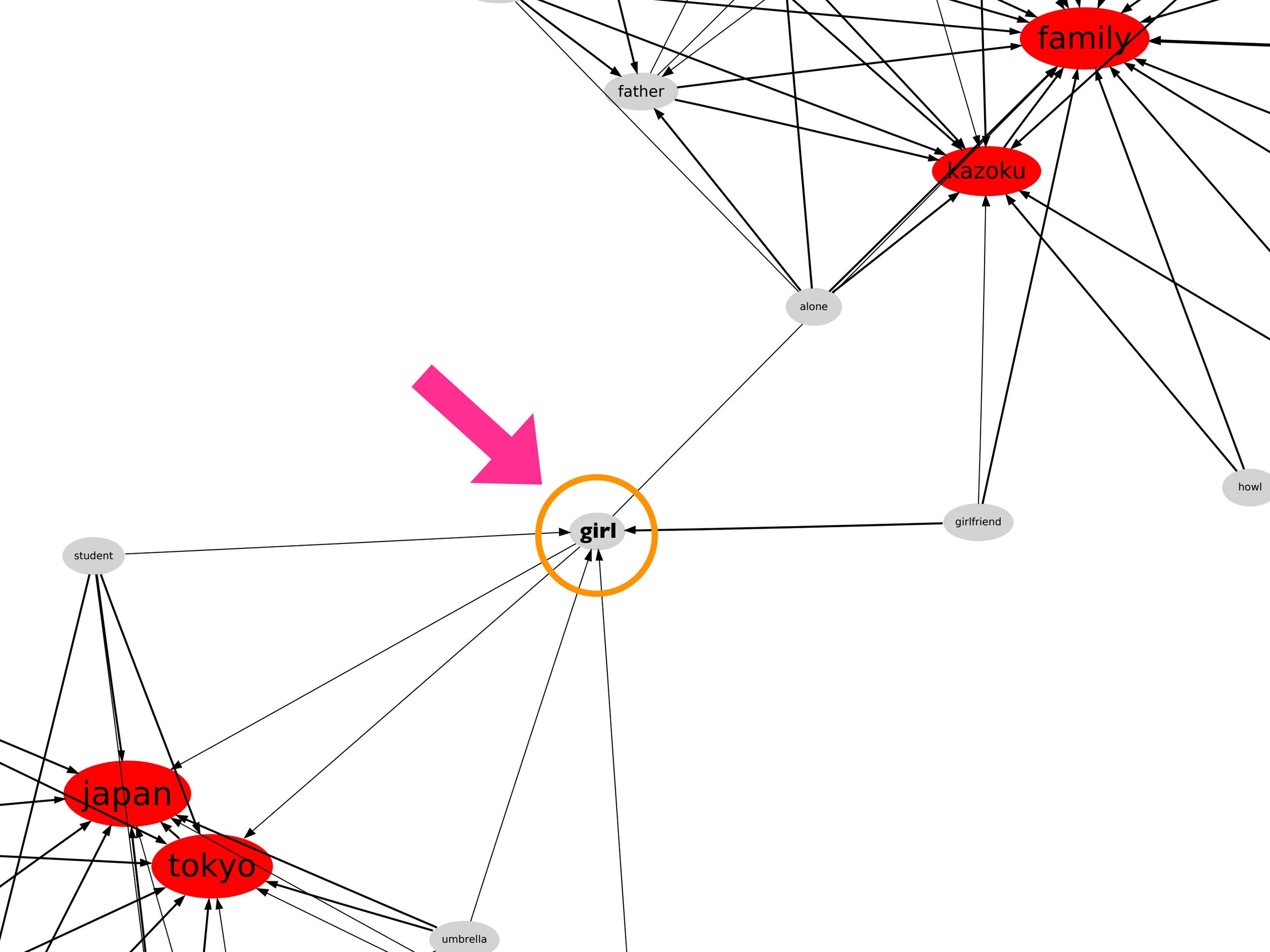
tag networks



→ professional life

→ family

→ life in Japan


family
father
kazoku
alone
howl
girlfriend
girl
student
japan
tokyo
umbrella

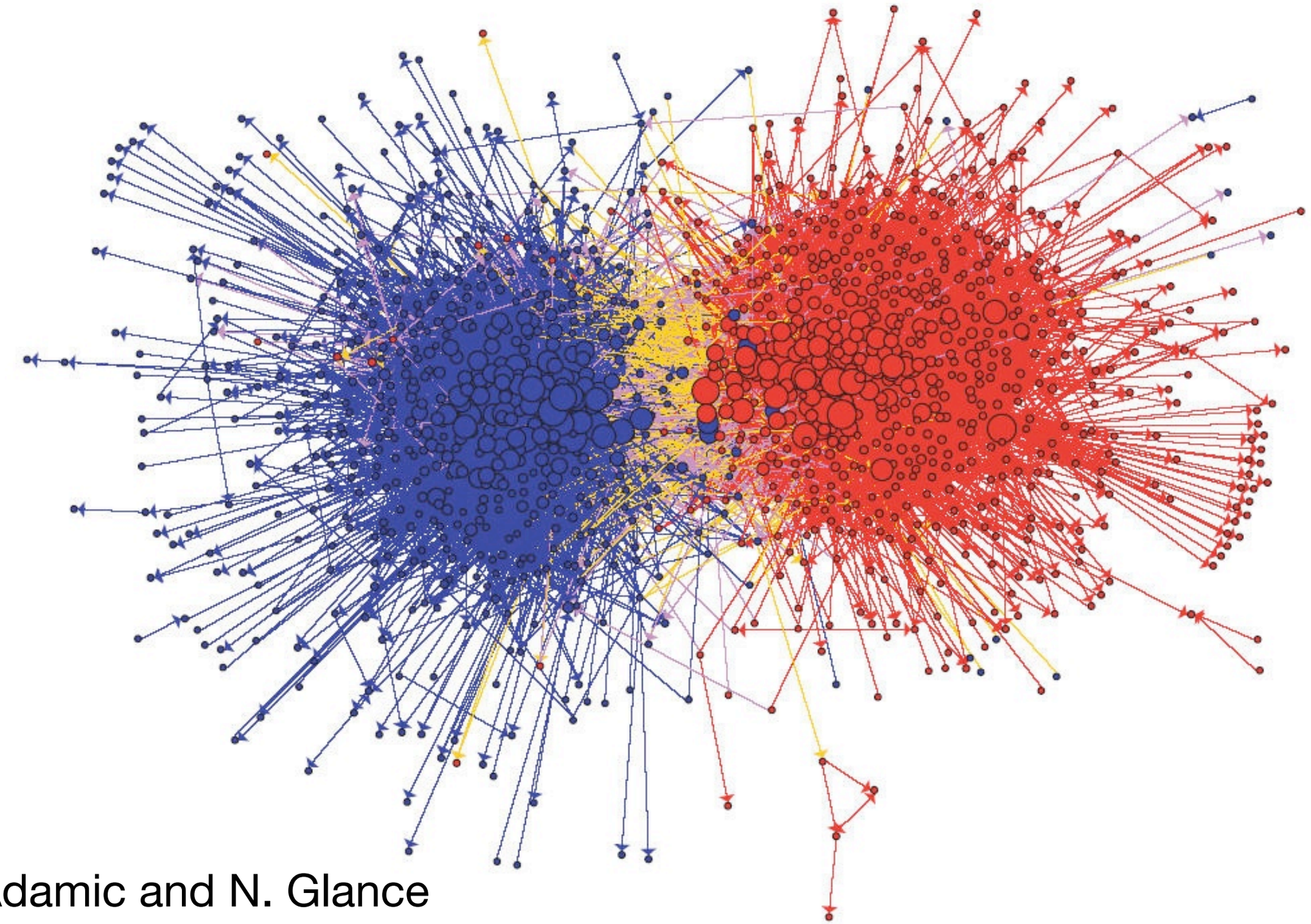
ideas we have just seen:

- **metadata** in place of content
- **co-occurrence** of tags / terms
- co-occurrence network
- **communities** in a network
- (degree) **centrality** of a node
- **bridges** between communities

... network science

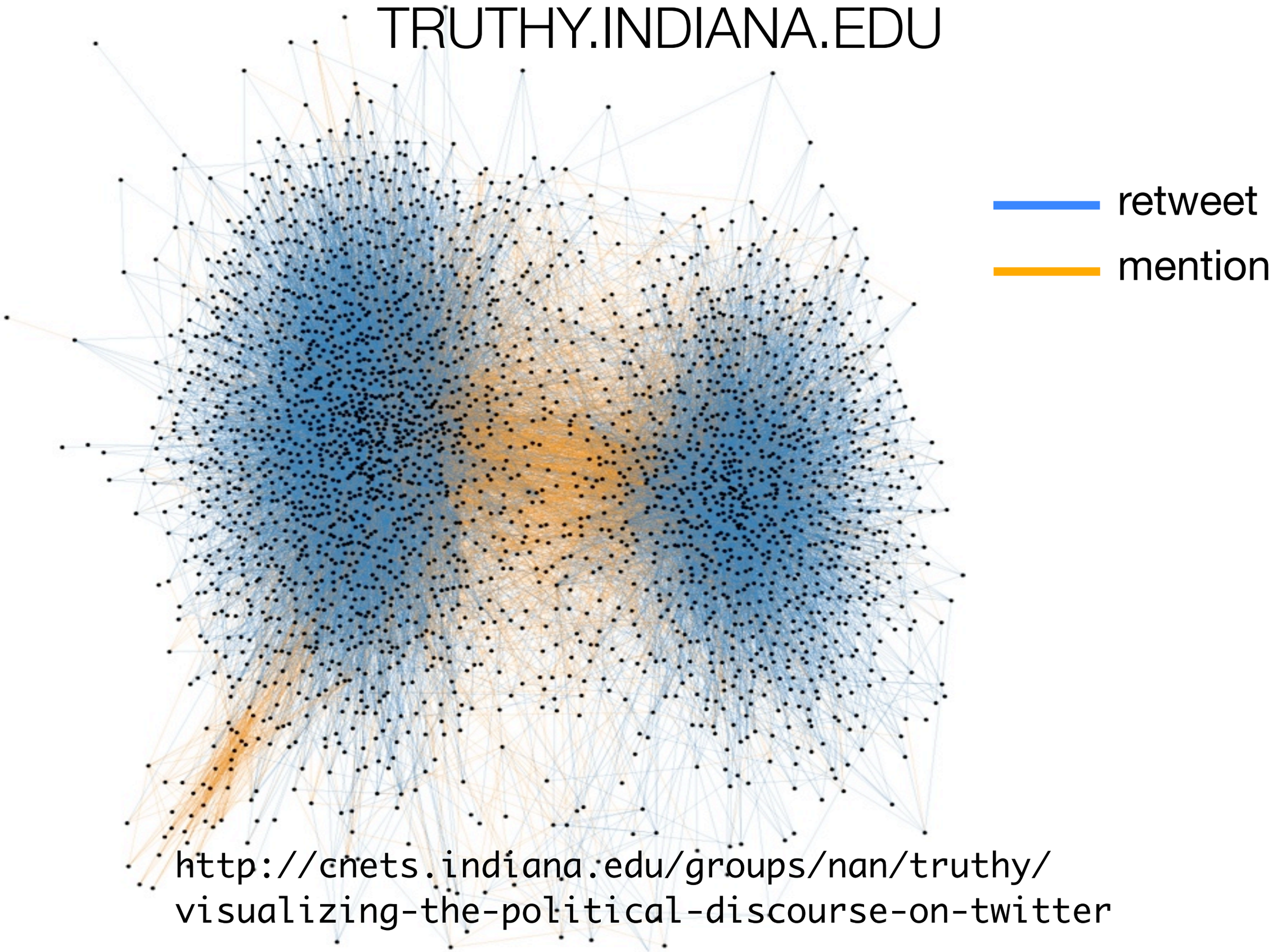
computational social science

a social network of US political bloggers



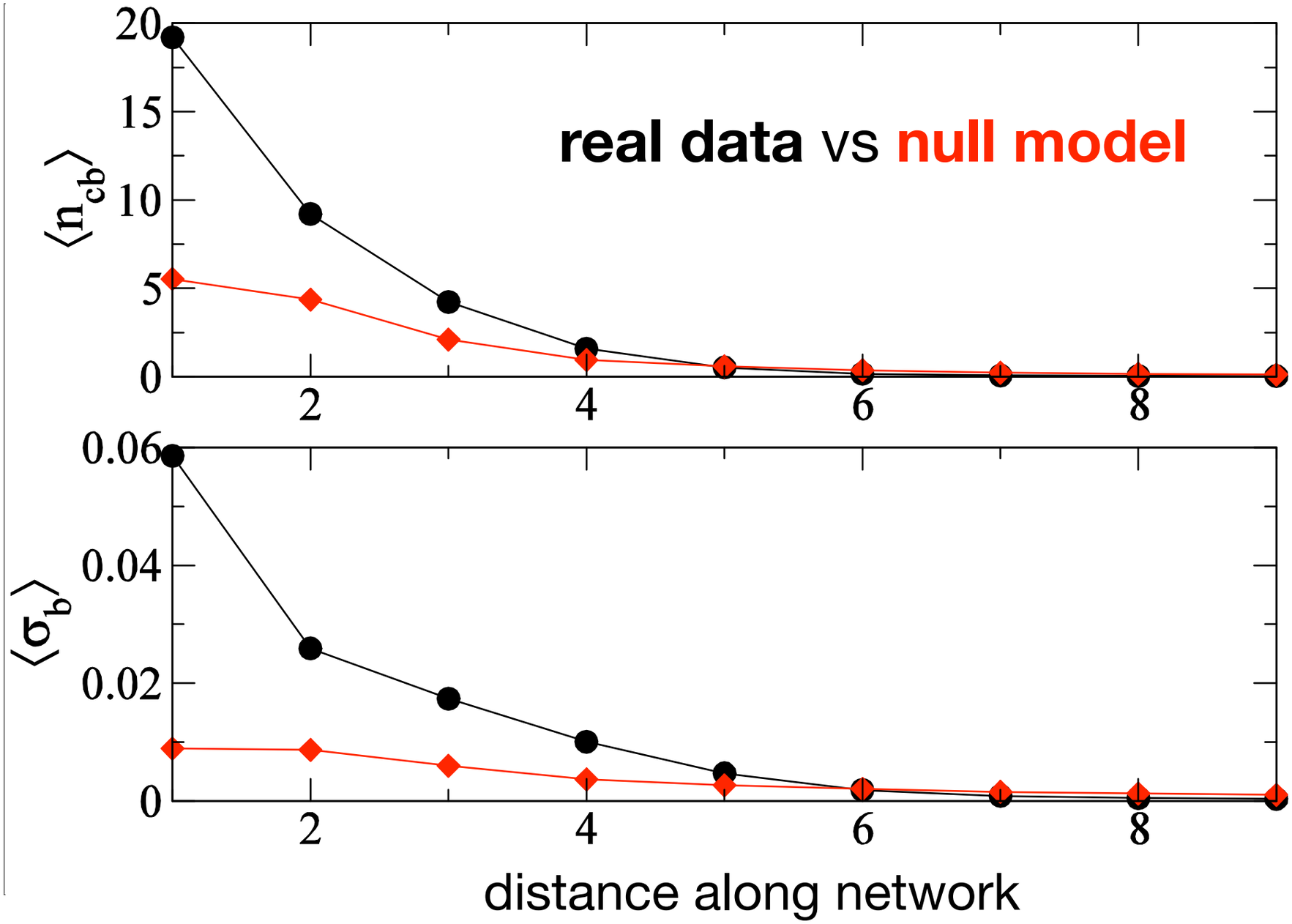
L. Adamic and N. Glance
The political blogosphere and the 2004 US election: divided they blog
Proc. 3rd international workshop on Link discovery, p.36 (2005)

political discourse on Twitter
TRUTHY.INDIANA.EDU



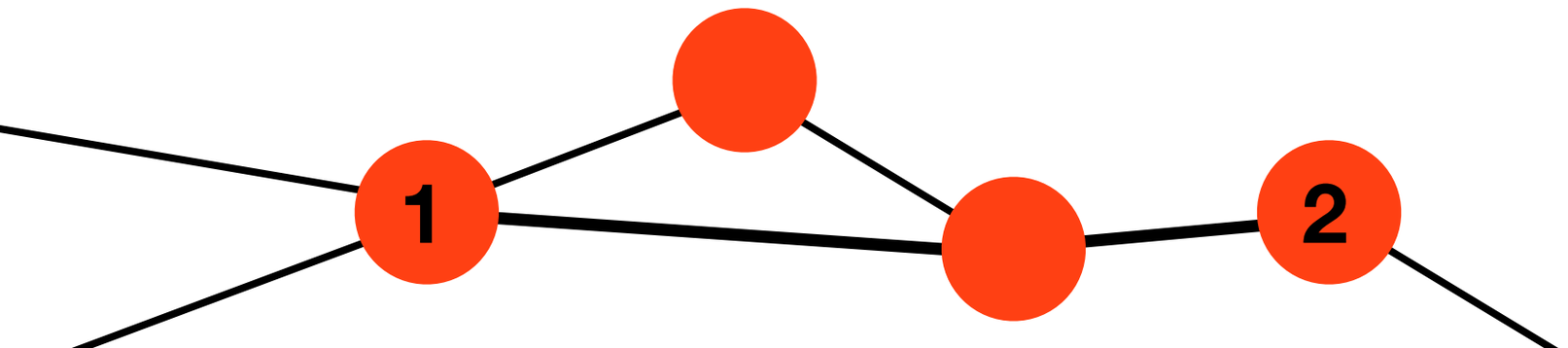
http://cnets.indiana.edu/groups/nan/truthy/visualizing-the-political-discourse-on-twitter

similarity in a social network: the case of aNobii



average number of common books of two users

average similarity between two users



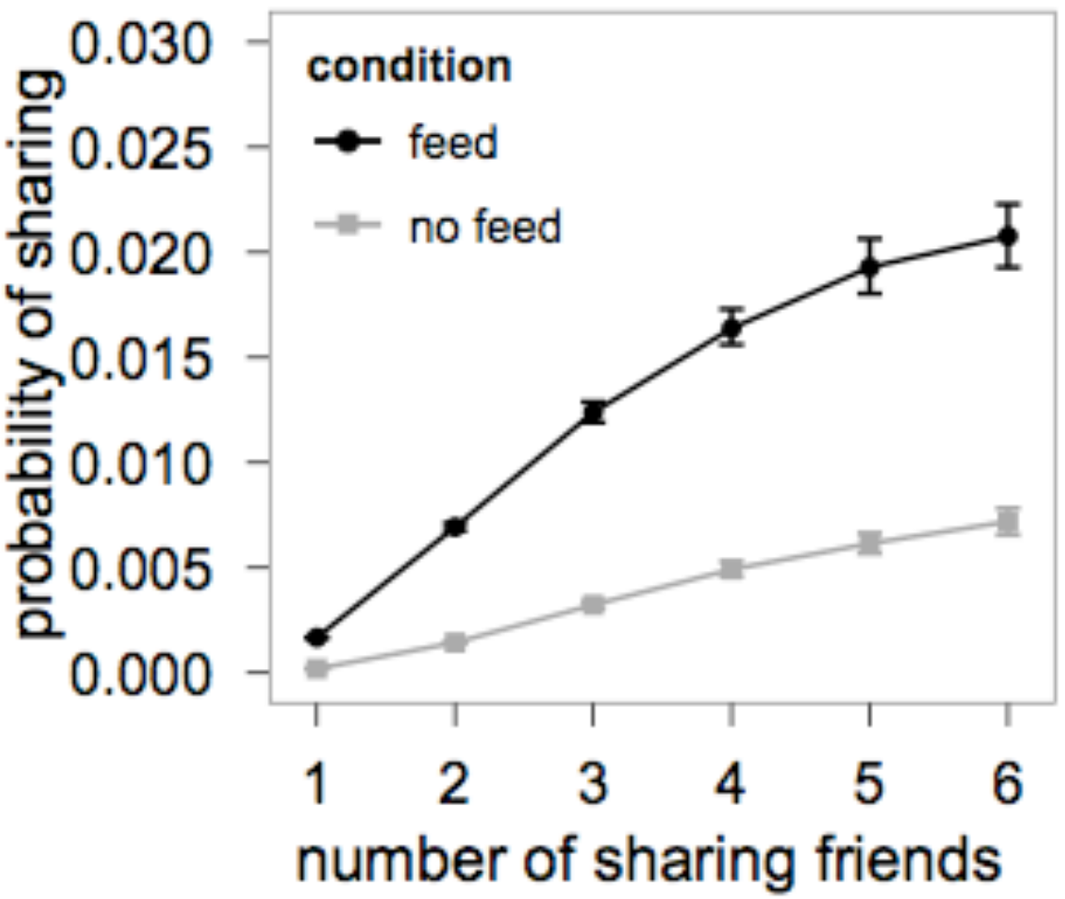
L. Aiello *et al.*,
http://arxiv.org/abs/1006.4966

an experiment within Facebook

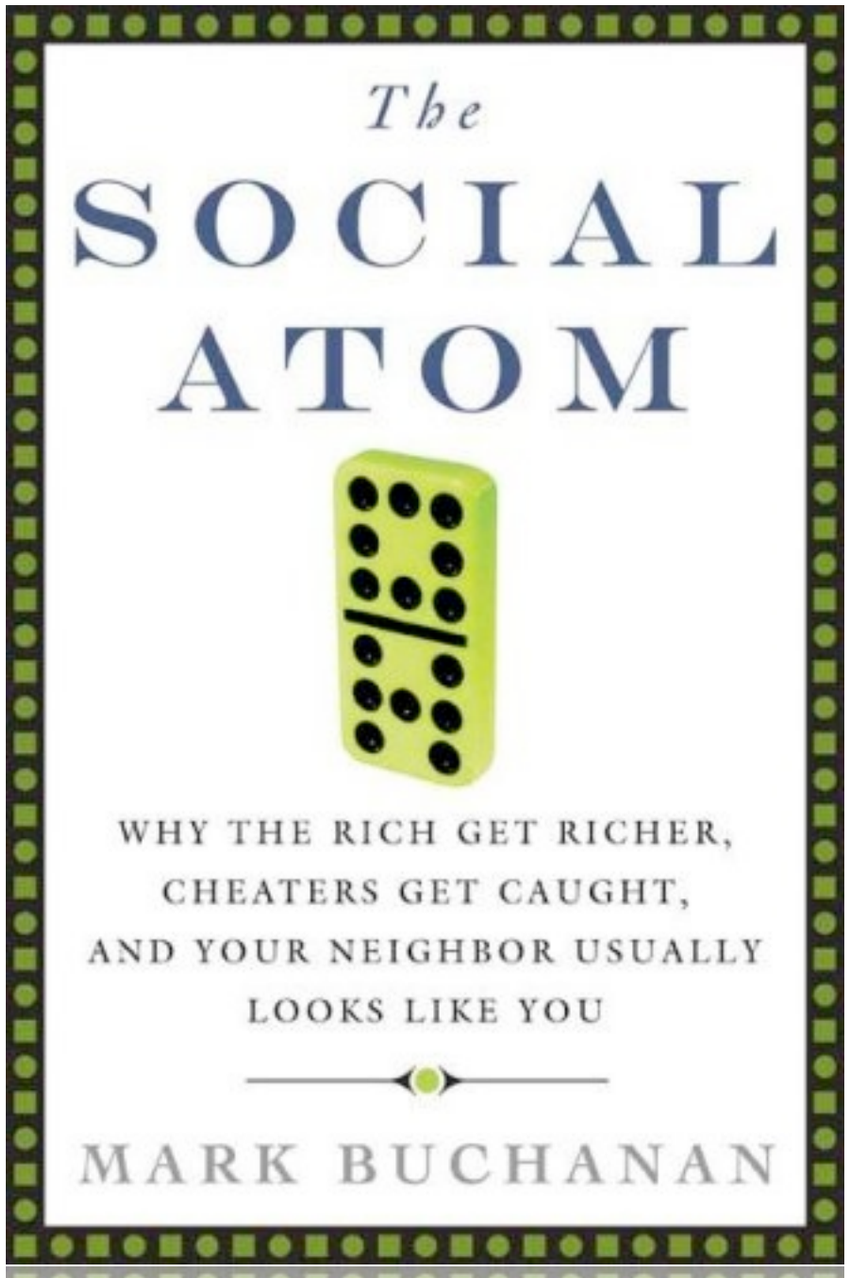
feed **no-feed**

E. Bakshy *et al.*, WWW2012, `http://arxiv.org/abs/1201.4145`

an experiment within Facebook



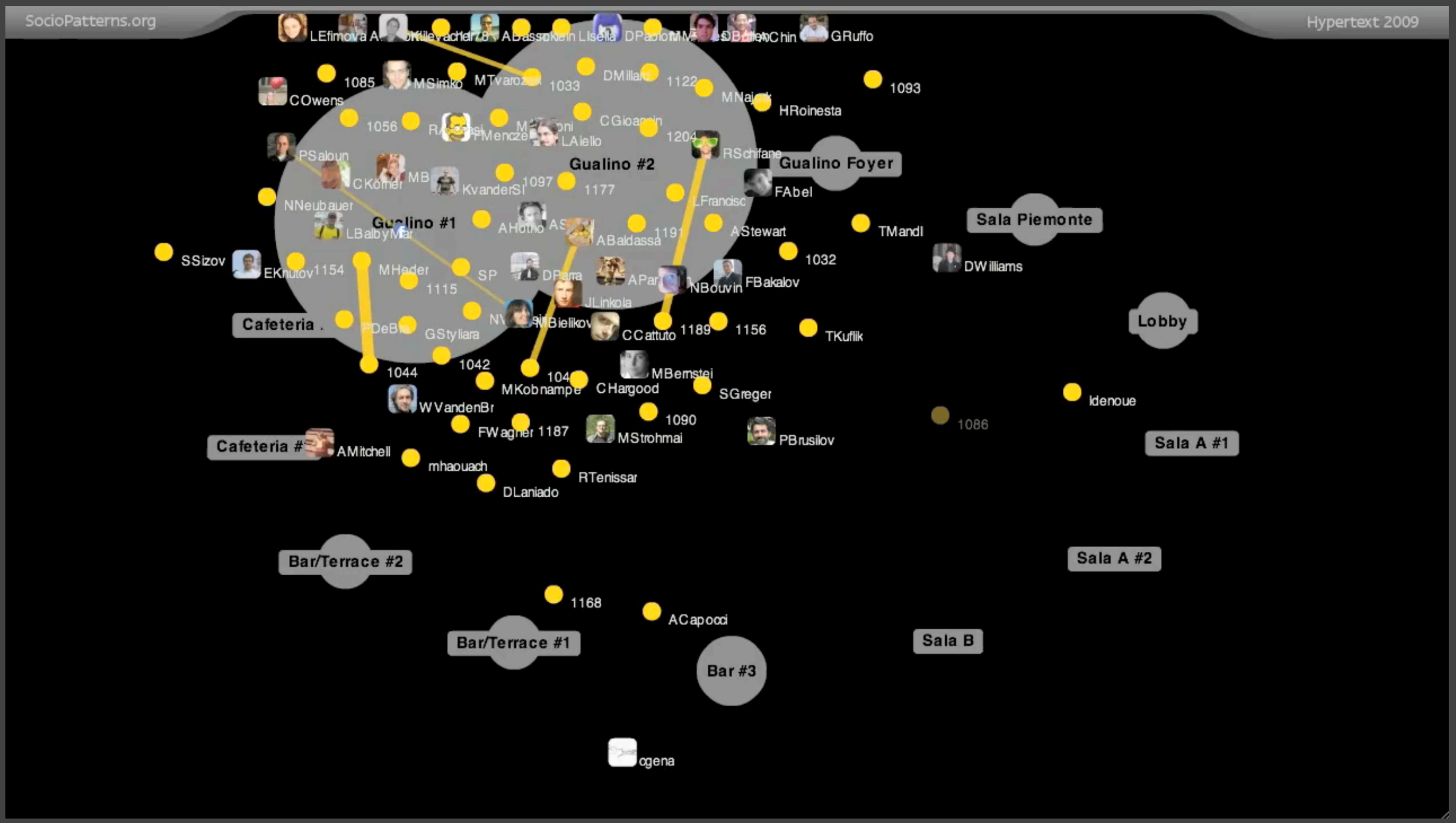
E. Bakshy *et al.*, WWW2012,
The Role of Social Networks in Information Diffusion
`http://arxiv.org/abs/1201.4145`


The
SOCIAL
ATOM
WHY THE RICH GET RICHER,
CHEATERS GET CAUGHT,
AND YOUR NEIGHBOR USUALLY
LOOKS LIKE YOU
MARK BUCHANAN

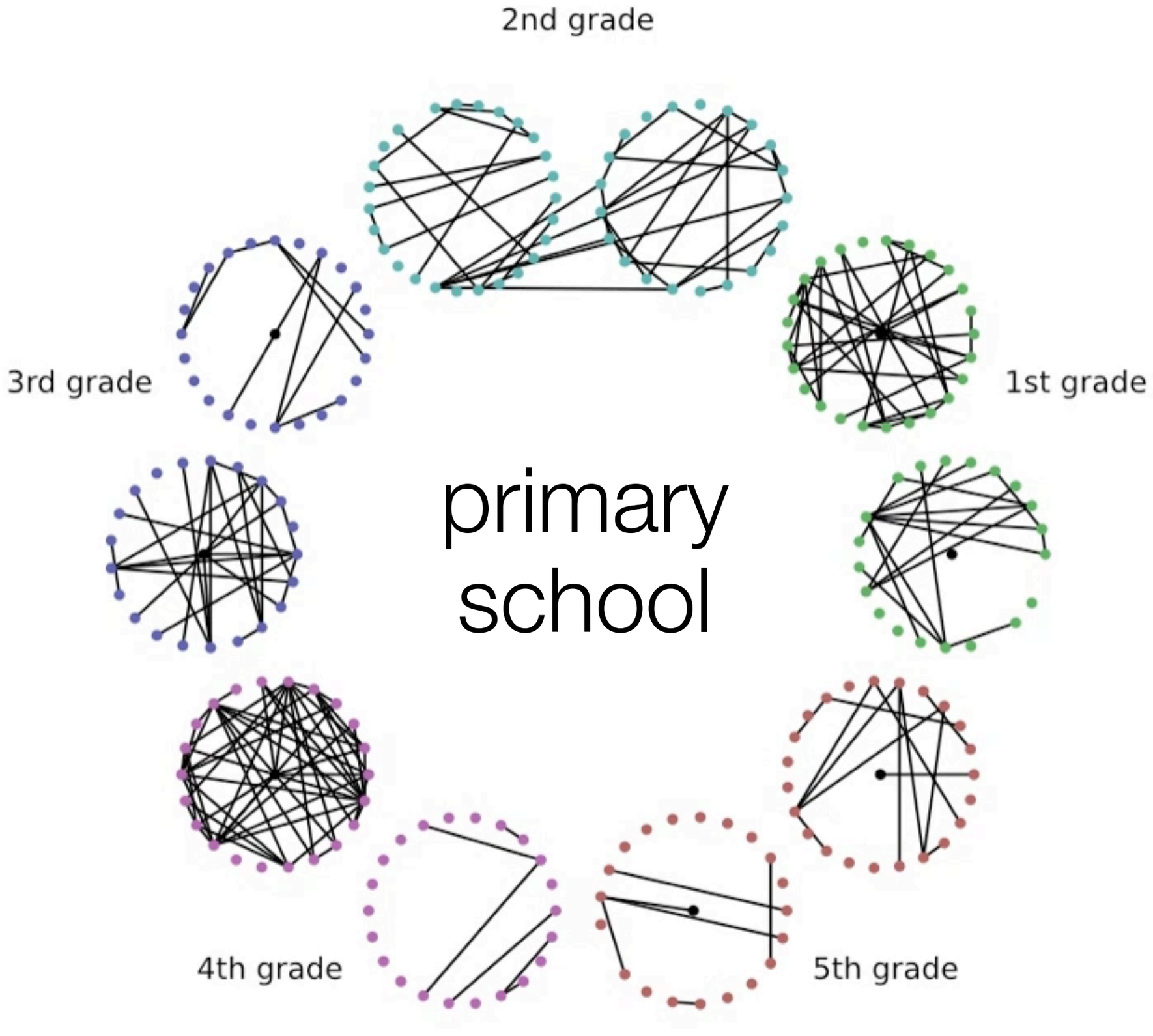
wearable sensors

www.sociopatterns.org

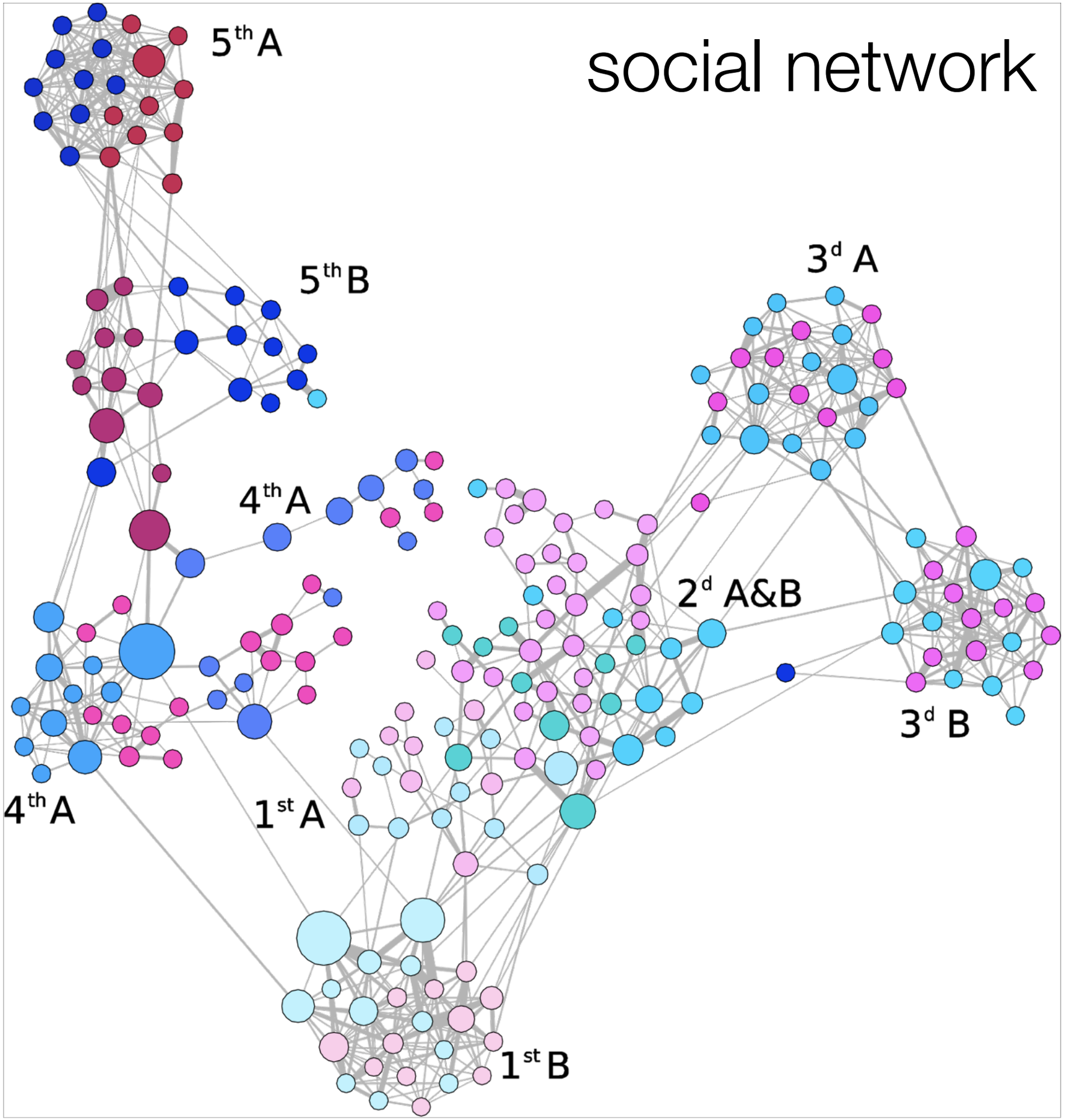
in vivo social networks



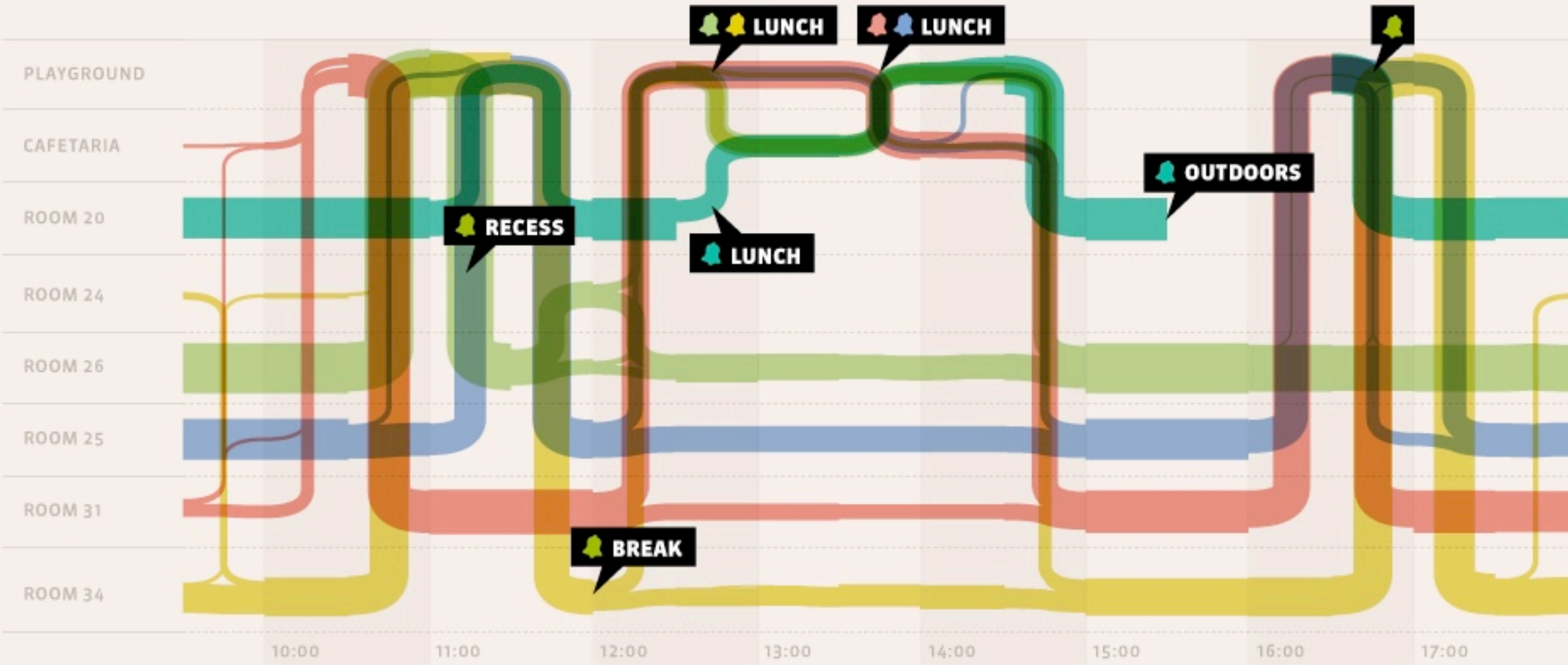
http://www.vimeo.com/6590604



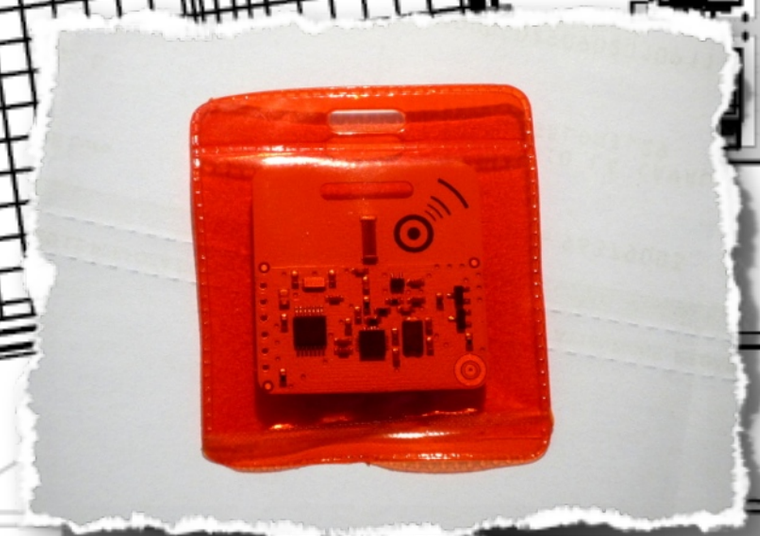
www.sociopatterns.org


social network
5th A
5th B
4th A
3d A
2d A&B
3d B
4th A
1st A
1st B

groups and trajectories



hospital

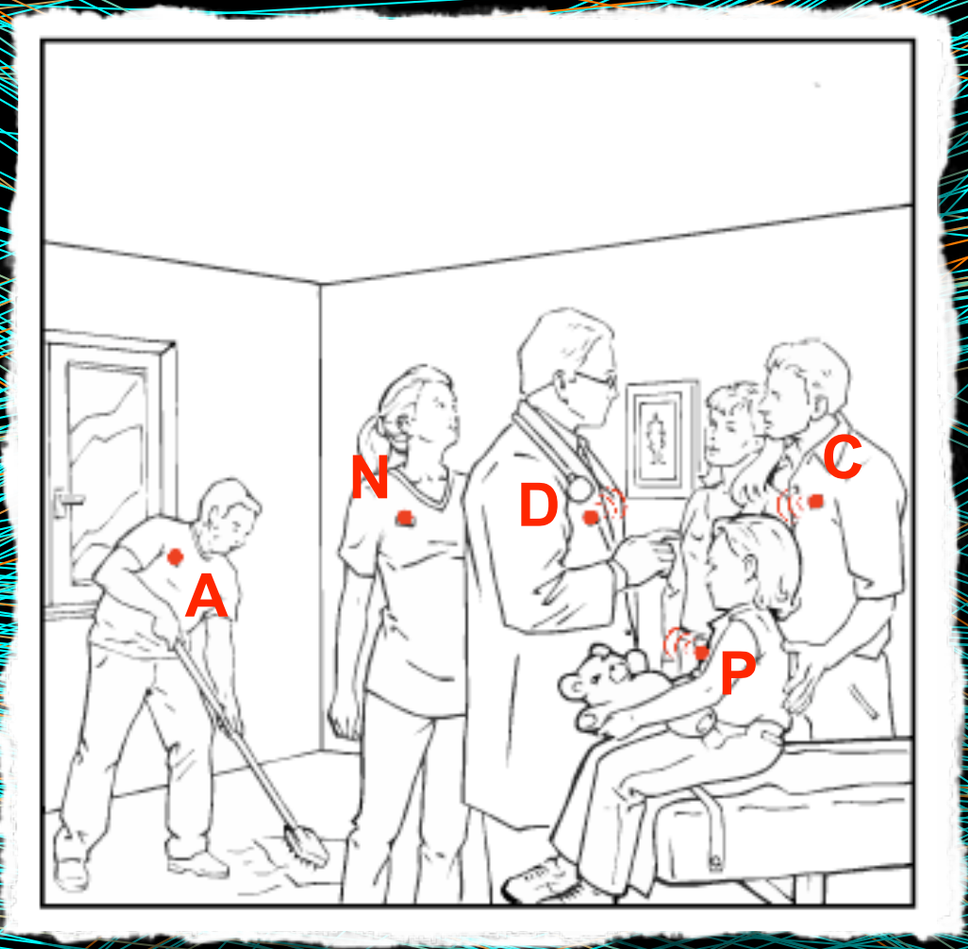


Text

sociopatterns.org

doctors

children



nurses

auxiliaries

parents

models

many types of models

- mathematical models
- statistical models
- generative models
- agent-based models
- machine learning models
- descriptive models vs predictive models vs dynamical models
- ...

>> forward

decision and policy making

human-machine compositionality

complex systems,
network science

data mining,
machine learning,
natural language
processing

digital platforms

BIG DIVE | HACKING DEVELOPMENT, VISUALIZATION & SCIENCE

A four-week training program to boost the technical skills needed to dive into the BIG DATA universe.

The idea behind **BIG DIVE** is to boost the growth of a new generation of developers.

A **street-fighting gym** where **high value datasets** are the raw material

20 candidates admitted

3 cross-disciplinary study paths

3 special events with national/international experts

Starting date: **October 1st**, 2012

Ending date: **October 26th**, 2012

Lessons from **Monday to Thursday**

We are looking for:

Hackers

Designers skilled in coding

the new role of data

TRADITIONAL APPROACH	NEW PERSPECTIVE
Data actively collected with user awareness	Most data from machine to machine transactions and passive collection – difficult to notify individuals
Definition of personal data is predetermined and binary	Definition of personal data is contextual and dependent on social norms
Data collected for specified use	Economic value and innovation come from combining data sets and subsequent uses
User is the data subject	User can be the data subject, the data controller, and/or data processor
Individual provides legal consent but is not truly engaged	Individuals engage and understand how data is used and how value is created
Policy framework focuses on minimizing risks to the individual	Policy focuses on balancing protection with innovation and economic growth

Source: World Economic Forum and The Boston Consulting Group

http://www.weforum.org/issues/rethinking-personal-data

reflecting on big data

- Bigger data are not always better data
- Just Because it is Accessible Doesn’t Make it Ethical
- Limited Access to Big Data Creates New Digital Divides

d. boyd, K. Crawford
Six Provocations for Big Data
Oxford Internet Institute’s *“A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society”*