Computational Social Science: Using Big Data as a Societal Microscope

Kinga Makovi

New York University Abu Dhabi

November 4, 2018

Outline

1 Motivation

2 New facts & zooming in

3 Scalable Experiments

4 Conclusion

The Leeuwenhoek microscope – A measurement revolution

Animalcules observed by Anton van Leeuwenhoek, c1795, Wikimedia Commons

The Big Data revolution as a measurement revolution

SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] 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]

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.

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 within 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.

"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."

Science, 2009, 323(5915):721–723.

The theoretical purchase of Big Data (for social science)

- The production of new social facts
- Our ability of zooming in by making big data small & constructing targeted comparisons
- Scalable experiments

Outline

1 Motivation

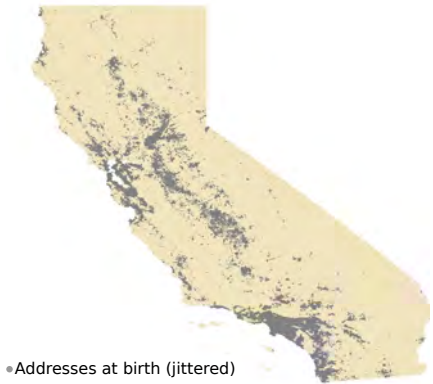
2 New facts & zooming in

3 Scalable Experiments

4 Conclusion

The Understanding Autism project

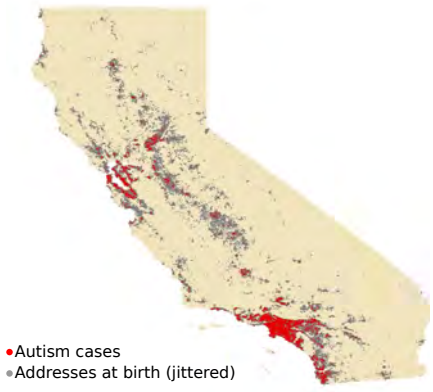
The Understanding Autism project



•Addresses at birth (jittered)

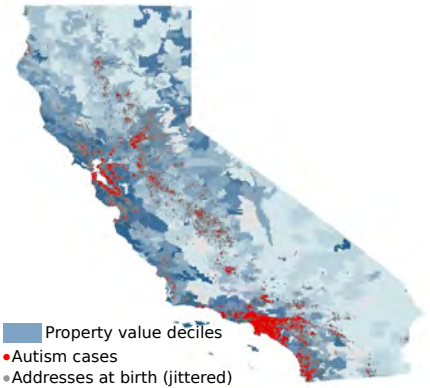
- Every child born in California, 1992–2007 with parental characteristics, birth outcomes & ZIP codes, & addresses from 1997 onward. 8 million births; 500k/year

The Understanding Autism project



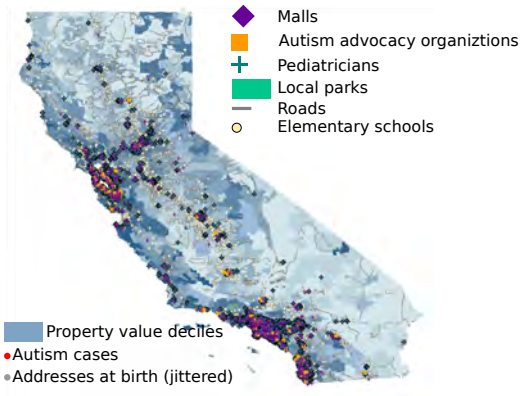
- Every child born in California, 1992–2007 with parental characteristics, birth outcomes & ZIP codes, & addresses from 1997 onward. 8 million births; 500k/year
- Probabilistically matched to: all persons in California with developmental disorders, 1992–2005, with evaluations on diagnostic status, severity every year; 30k autism cases, 140k MR cases

The Understanding Autism project



- Embedded in neighborhoods with socio-demographic characteristics and local pollutant levels – here, property values

The Understanding Autism project



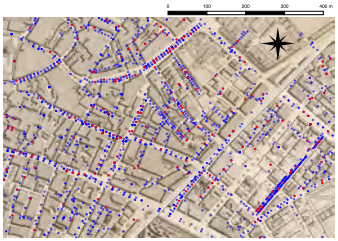
- Embedded in neighborhoods with socio-demographic characteristics and local pollutant levels – here, property values
- Distance to: each other, malls, parks, pediatricians, schools, autism advocacy organizations
- This allows us to see how close every case is to every other case and how close they are to places where people talk

What did we learn?

- As much as 15% increase in autism prevalence can be attributed to the **social diffusion of knowledge about** the symptoms of **autism**, and the **benefits of an autism diagnosis** in terms of resources for children (Liu et al 2010, *AJS*; Liu & Bearman 2012, *SMR*)
- The causal effect of the **autism status of first born children** on subsequent fertility is smaller than previously thought, and **impacts mothers with different educational and class backgrounds in different ways** (Makovi et al 2015, *Sociological Science*)
- This research contributed to public health, sociology, medicine, genetics, and is followed by new research funded by NIH (2018–2021) to study the interlink between autism and artificial reproductive technologies

Big Data techniques shape historical research – abolitionist petitioning in Manchester

Petition (1807), Parliamentary Archive

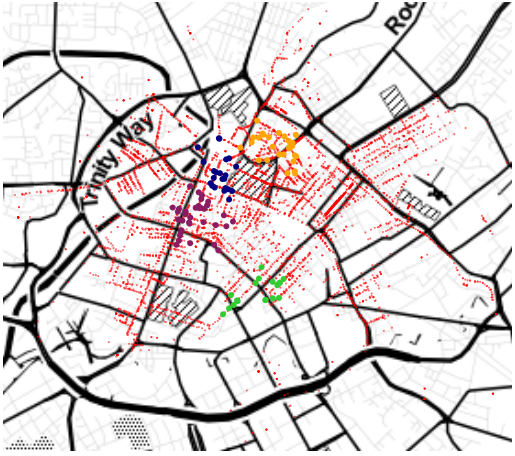


Historic GIS, petitioners & non-petitioners mapped based on a Trade Directory

Inns and taverns mapped

Quaker families identified based on church records of burials

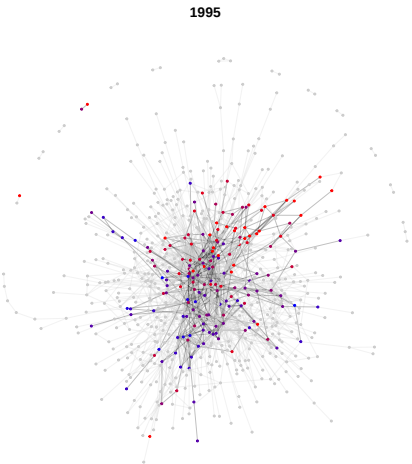
What did we learn?



- Petitioners clustered in space, and that these clusters centered on the inns and taverns that housed merchants regularly from communities that petitioned for abolition at earlier timepoints
- Industrialization changed patterns of social stratification, but also connectivity among communities, and thus information flows

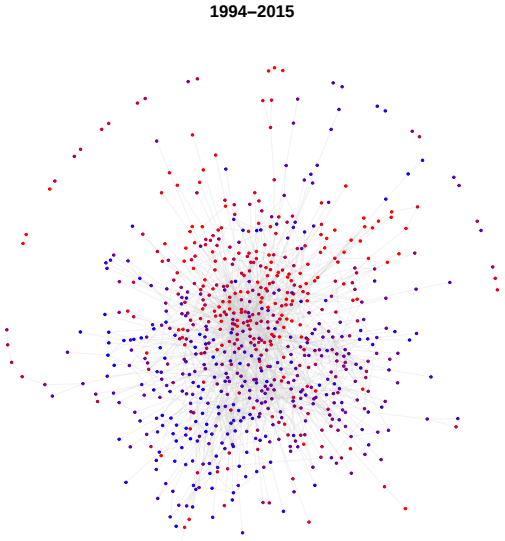
Makovi 2018, *under review*

Big Data approach to gender inequality: Do co-worker networks impact men and women differently on the labor market?



- Data on all workers paying social security contributions on the German labor market
- In Berlin between 1990-2015 500k unique workers, and 700k person-years
- 300k unique men and 200k unique women in 16 industries

What did we learn?



- Preliminary findings suggest that the **increase in salaries after a move between companies** is associated with the **presence of former co-workers**.
- The **salary-bump experienced by men is higher compared to women**.

Outline

1 Motivation

2 New facts & zooming in

3 Scalable Experiments

4 Conclusion

Lying in social networks

- What is the causal effect of social networks on the spread of misinformation?
- What is the effect of verification on the spread of misinformation?

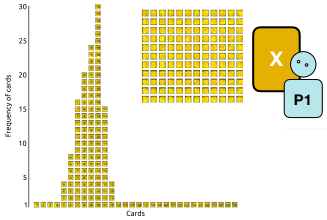
Science, 2018 March

Web of lies – an experimental study of lying

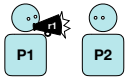
The Web-of-Lies Game

3 players embedded in a directed network

1. P1 clicks on a card to draw a *hidden number* **x={1,30}**. The distribution is public knowledge



2. P1 sends a message to P2 reporting on **x**

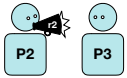


Web of lies – an experimental study of lying

The Web-of-Lies Game

3 players embedded in a directed network

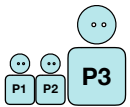
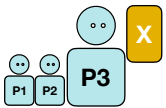
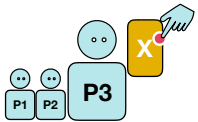
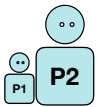
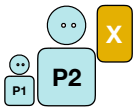
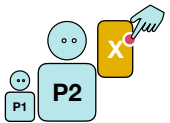
3 P2 observes **r1** and sends a message to P2 reporting on it



4 P3 observes **r2** and sends the final message

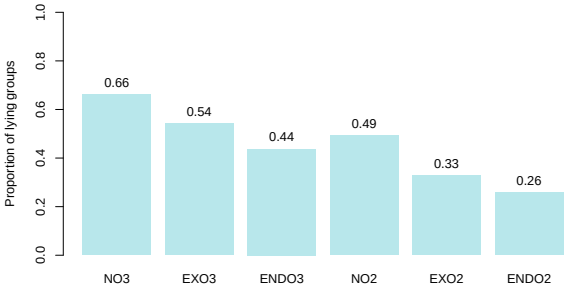
P3's report determines payoffs for all

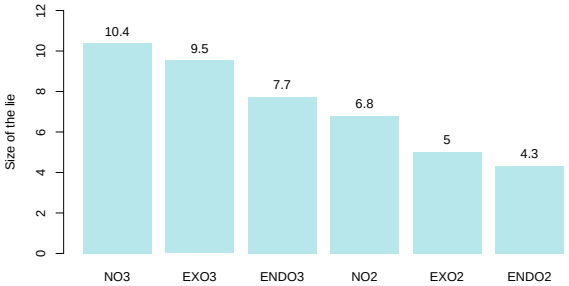
Main treatments: network structure & verification of truth

3 Person	**NO3** No verification 	**EXO3** Verifies with **p=0.8** 	**ENDO3** Verifies by clicking a button 
2 Person	**NO2** No verification 	**EXO2** Verifies with **p=0.8** 	**ENDO2** Verifies by clicking a button 

What did we learn?

Networks allow individuals to hide when they lie, and to lie more. Verification reduces lies.





Outline

1 Motivation

2 New facts & zooming in

3 Scalable Experiments

4 Conclusion

Big Data is a microscope in the social sciences

- NIH Pioneer Award Program #1 DP1 OD003635-01, NSF DDGIR # 1435138
- Peter Bearman, Ka Liu, Manu Munoz, Melina Platas, Malte Reichlet, Alix Winter