University of St Andrews | FOUNDED 1413

St Andrews Institute for Data-Intensive Research
http://www.idir.st-andrews.ac.uk

Complex networks, real data

Simon Dobson
(Using work by Saray Shai and Aleks Sazonovs)

simon.dobson@st-andrews.ac.uk
http://www.simondobson.org

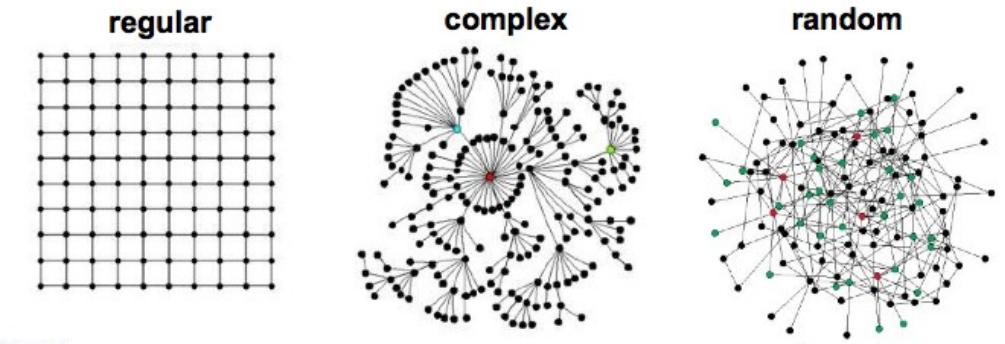
Introduction

- Data-intensive techniques let us do research we otherwise would struggle to do
  - Changes the way we do science...
  - ...and the science we do
  - The computer is the new microscope
- Its variability also raises new problems
  - Data access, data hygiene

- This talk: variability and its outcasts

Background – complex networks

- "Things" and their "relationships"
  - People and their friendships
  - Towns and their road (and rail) connections
- Processes over networks
  - Spread of disease, flow of traffic, …

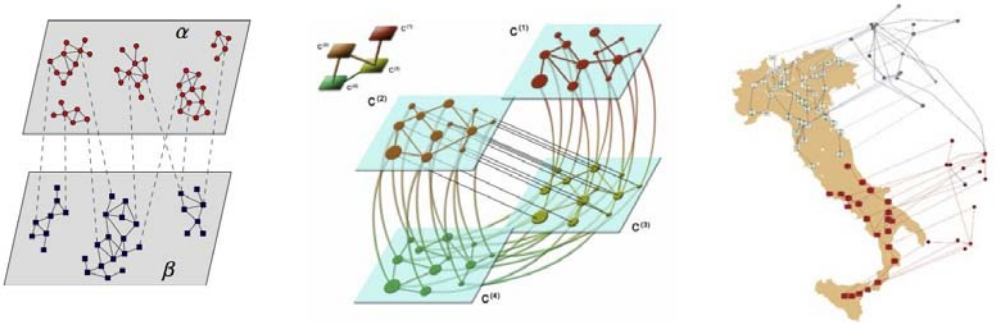


Barabasi *et alia*. Science **286**. 1999.

- How do the properties of the network affect the behaviours of processes?

Where do networks come from?

- Everywhere, it seems
  - Real engineering, real social networks, real protein interactions, real brains, …
  - Often we want to *combine* several networks to see how they interact

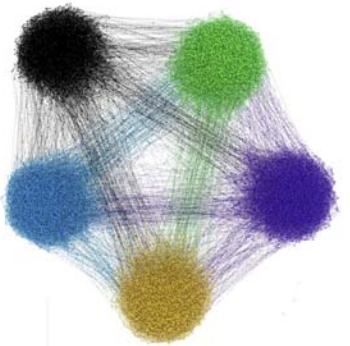


Leicht *et alia*; de Domenico *et alia*; Buldyrev *et alia*

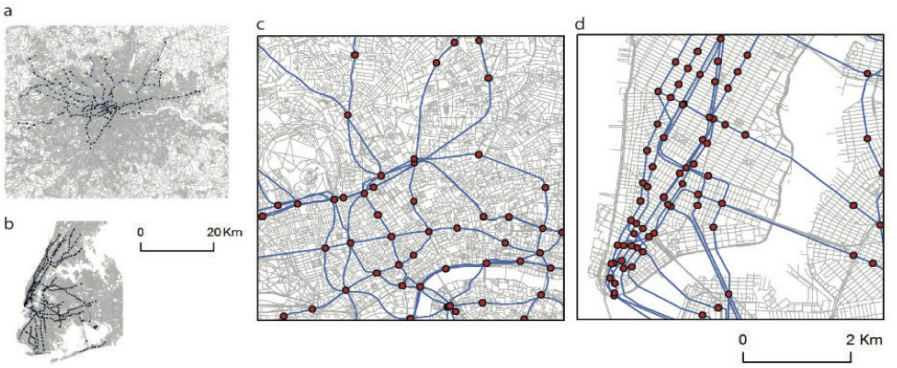
- Variability in the data sources

Variable sources (and implications)

- Simulated
  - Determine the statistical properties we want, and construct by process
  - (Are they the right properties? Are there others we haven't accounted for?)

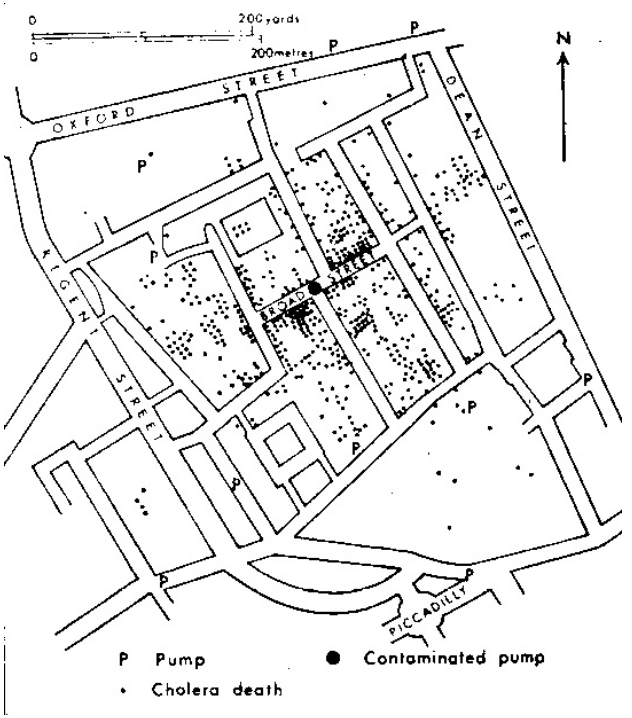


- Real
  - Increasingly widely available
  - (Are they reliable? Timely? Noisy? Complete? Free? Sharable?)



It used to be so simple...

- Collecting primary data
  - Much better control over the methods and care taken
  - ...and you only have yourself to blame if things go wrong
- Re-using public data
  - Often no robust checks
  - May be good reasons it has errors in it, aside from just noise
  - Statistical techniques



Real-world – raw

- Some data is available, some isn't

- Real geographical data from OpenStreetMap
  - Too complex
  - Noisy?
  - No population information

Real-world – simplify and infer

- Clean up
  - Remove unnecessary topographic information

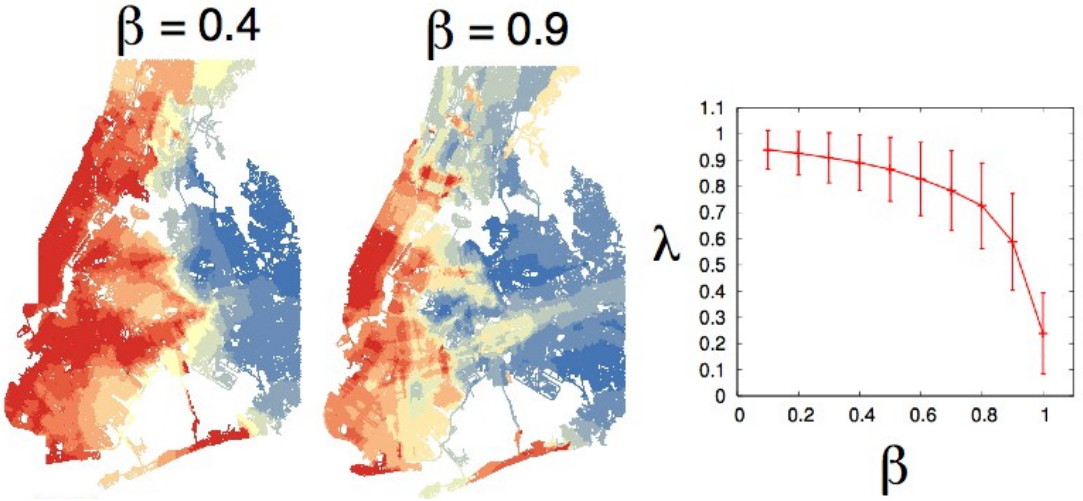
This road got straighter as we removed the unnecessary bends that affect its topography but not its topology

Sazonovs. A metapopulation model for predicting the success of genetic control measures for malaria. University of St Andrews. 2015. Senior honours project.

- Infer
  - Clusters of roads indicate settlements?

Real-world – real meets simulation

- A stack of variable quality
  - Real base data
  - Inferred structure
  - Simulated process

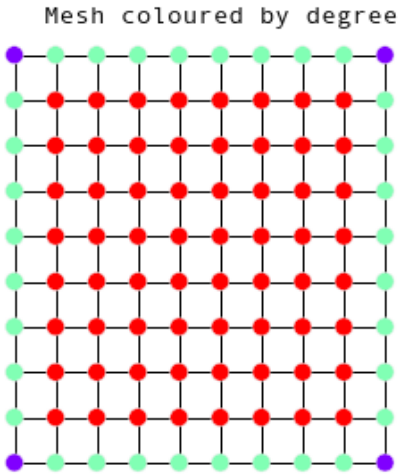


Strano, Shai, *et alia*. Multiplex networks in metropolitan areas: universal features and local effects. Submitted to Journal of the Royal Society Interface.

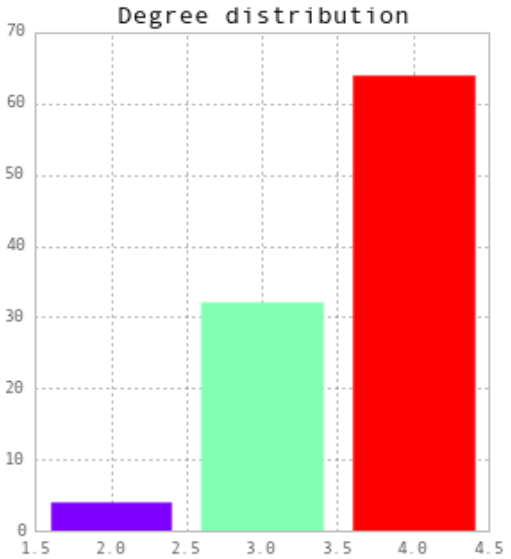
- I worry about the interactions between these layers
  - What are the right statistical properties to measure? The right confidence metrics?

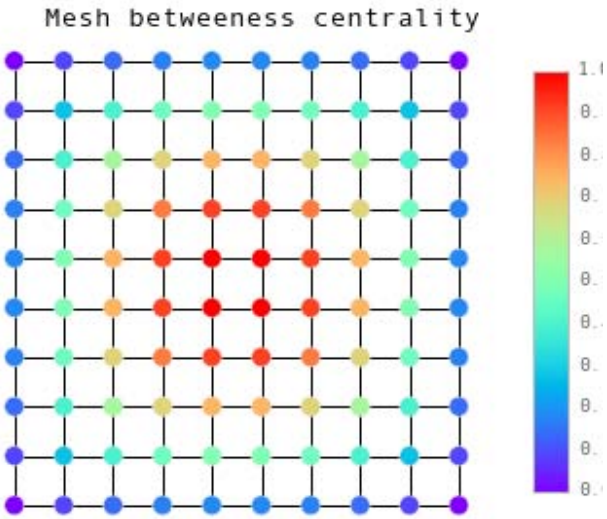
Sampling – 1

- Strip to essentials
  - What happens for *uncomplex* networks?
  - How would they be "sampled"?



Diameter of this network is 18





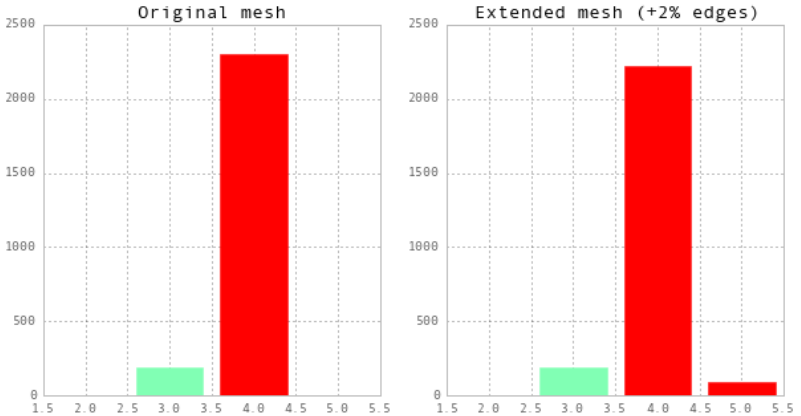
Dobson. Complex networks, complex processes. 2015. *Open textbook in preparation.* http://www.simondobson.org/research/complex-networks-complex-processes/

Sampling – 2

- Now introduce some “sampling error”
  - 2% more edges, added at random

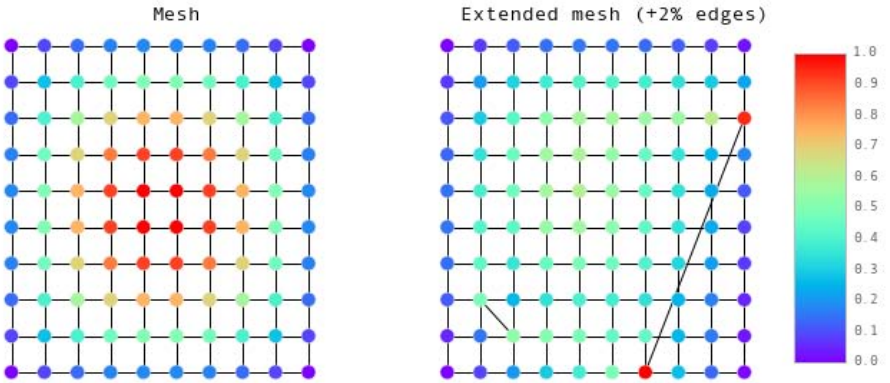
Sampling – 3

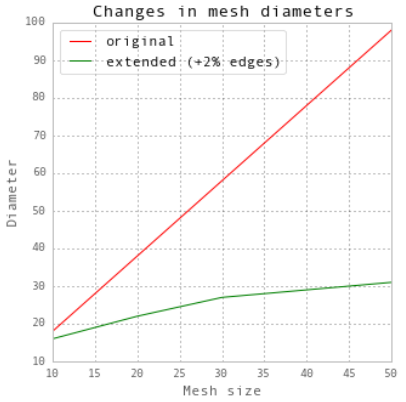
- Now introduce some "sampling error"
  - 2% more edges, added at random



Would you even notice if you didn't sample these extra edges?

Betweenness centrality collapses





For 50x50, 2% more edges collapses diameter by >65%

Impact

- Complex processes can be vey sensitive to the topology of the network they run on
  - ...and if we're sampling, this is something else to worry about
- There may not *be* a better source
  - ...so our cleaning is unavoidable
- There may be only *one* network
  - ...which is a problem if we need to do multiple runs for statistical purposes, *i.e.*, to reduce variance

Three things to take away

- Being awash with data is not an undiluted blessing
- We need to understand the interactions between the different aspects of our models
- This might also let us build better, fake, real networks as a basis for experiments