apa global

[entry]apa/global
3.36pt

# Digital Research Methods II

Francesco Bailo

The University of Sydney

3 May 2018

# Observing behaviour

This section is largely based on Chapter 2 of **salganik_bit_2018** by Matthew Salganik (2018).

Whatever the subject of your research, there are mainly three ways to collect data:

1. Running experiments

Whatever the subject of your research, there are mainly three ways to collect data:

1. Running experiments
2. Asking questions

Whatever the subject of your research, there are mainly three ways to collect data:

1. Running experiments
2. Asking questions
3. Observing behaviour ←

Whatever the subject of your research, there are mainly three ways to collect data:

1. Running experiments
2. Asking questions
3. Observing behaviour ←

- Observational data are collected without interfering with either
  - the subject of the investigation or
  - the environment of the subject of the investigation.

1. Navigate to https://socrative.com
2. ‘Student login’
3. Room name: ‘BAILO’

Observation of something or somebody is the primordial way of investigating what we are interested in (and usually the beginning of an investigation).

The use of instruments and sensors to observe and record what we observe is not new.

What is new is the number of instruments and sensors monitoring and recording human behaviour.

*Figure 1:* **bertini_galileo_1858, bertini_galileo_1858**

The combination of *flow* and the *stock* of data produced by these instruments and sensors is often called **Big Data**.

Big data are *big* on three dimensions:

- **V**olume
- **V**ariety
- **V**elocity

- So... can you think of any example of big data or source of big data?

- Big data are not only data generated by the online activity of users and are not only created by companies.

- Big data are not only data generated by the online activity of users and are not only created by companies.

1. Big data are generated online but also offline every time a sensor records a human behaviour.

- Big data are not only data generated by the online activity of users and are not only created by companies.

1. Big data are generated online but also offline every time a sensor records a human behaviour.
2. Big data are created by companies but also by governments.

'Big data are created and collected by **companies** and **governments** for purposes other than research. Using this data for research therefore requires repurposing.' (**salganik_bit_2018**)

$\neq$

According to (**salganik_bit_2018**), Big Data share 10 characteristics.

1. Big Data are **big**: Rare events, heterogeneity, small differences. *But* how data were created?

According to (**salganik_bit_2018**), Big Data share 10 characteristics.

1. Big Data are **big**: Rare events, heterogeneity, small differences. *But* how data were created?
2. Big Data are **always-on**: Unexpected events and real-time estimates. *But* the systems that collected the data are constantly changing (see *drifting* later)!

According to (**salganik_bit_2018**), Big Data share 10 characteristics.

1. Big Data are **big**: Rare events, heterogeneity, small differences. *But* how data were created?
2. Big Data are **always-on**: Unexpected events and real-time estimates. *But* the systems that collected the data are constantly changing (see *drifting* later)!
3. Big Data are **nonreactive**: Measurement is less likely to change behaviour. *But* a social desirability bias persist.

According to (**salganik_bit_2018**), Big Data share 10 characteristics.

1. Big Data are **big**: Rare events, heterogeneity, small differences. *But* how data were created?
2. Big Data are **always-on**: Unexpected events and real-time estimates. *But* the systems that collected the data are constantly changing (see *drifting* later)!
3. Big Data are **nonreactive**: Measurement is less likely to change behaviour. *But* a social desirability bias persist.
4. Big Data are **incomplete**: No demographic information, no information on behaviour on other platforms, and no data to operationalise theoretical constructs (e.g. 'intelligence').

According to (**salganik_bit_2018**), Big Data share 10 characteristics.

1. Big Data are **big**: Rare events, heterogeneity, small differences. *But* how data were created?
2. Big Data are **always-on**: Unexpected events and real-time estimates. *But* the systems that collected the data are constantly changing (see *drifting* later)!
3. Big Data are **nonreactive**: Measurement is less likely to change behaviour. *But* a social desirability bias persist.
4. Big Data are **incomplete**: No demographic information, no information on behaviour on other platforms, and no data to operationalise theoretical constructs (e.g. 'intelligence').
5. Big Data are **inaccessible**: Access is controlled and conditional.

6. Big Data are **non-representative**: Data do not come from a probabilistic random sample of the population.

6. Big Data are **non-representative**: Data do not come from a probabilistic random sample of the population.
7. Big Data are **drifting**: Population drift, behavioural drift, system drift. Systems keep changing all the time!

6. Big Data are **non-representative**: Data do not come from a probabilistic random sample of the population.
7. Big Data are **drifting**: Population drift, behavioural drift, system drift. Systems keep changing all the time!
8. Big Data are **algorithmically confounded**: Engineering choices impact user behaviours. Also, performativity issues.

6. Big Data are **non-representative**: Data do not come from a probabilistic random sample of the population.
7. Big Data are **drifting**: Population drift, behavioural drift, system drift. Systems keep changing all the time!
8. Big Data are **algorithmically confounded**: Engineering choices impact user behaviours. Also, performativity issues.
9. Big Data are **dirty**: Dirty data can be created unintentionally or intentionally (e.g. bots).

6. Big Data are **non-representative**: Data do not come from a probabilistic random sample of the population.
7. Big Data are **drifting**: Population drift, behavioural drift, system drift. Systems keep changing all the time!
8. Big Data are **algorithmically confounded**: Engineering choices impact user behaviours. Also, performativity issues.
9. Big Data are **dirty**: Dirty data can be created unintentionally or intentionally (e.g. bots).
10. Big Data are **sensitive**: The potential sensitivity of the data is difficult to always assess.

# Network analysis

**Relations, not attributes. Networks, not groups.**
[S]ocial network analysts argue that causation is not located in the individual, but in the social structure. While people with similar attributes may behave similarly, explaining these similarities by pointing to common attributes misses the reality that *individuals with common attributes often occupy similar positions in the social structure*. That is, *people with similar attributes frequently have similar social network positions*. Their similar outcomes are caused by the **constraints**, **opportunities** and **perceptions** created by these similar network positions. (**marin_social_2011**)

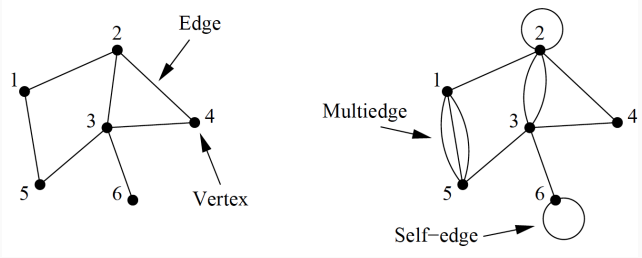



***Figure 2:*** *Traditional visualisation of two small networks...*

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

***Figure 3:*** *... and the adjacency matrix of the left-hand network* (**newman_networks_2010**).

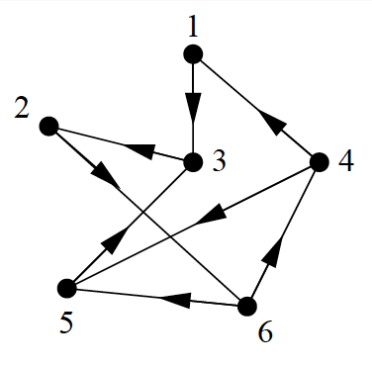



***Figure 4:*** *A directed network...*

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

***Figure 5:*** *... and its adjacency matrix (not symmetric!) (***newman_networks_2010***).*

# Network measures

**Degree of a vertex** number of connections

4->figure/betweenness

5->figure/groups
6->figure/components

# Network measures

**Degree of a vertex** number of connections

**Authority of a vertex** number of important connections

4->figure/betweenness

5->figure/groups

6->figure/components

# Network measures

**Degree of a vertex** number of connections

**Authority of a vertex** number of important connections

**Closeness of a vertex** mean distance to other vertices

4->figure/betweenness

5->figure/groups

6->figure/components

# Network measures

**Degree of a vertex** number of connections

**Authority of a vertex** number of important connections

**Closeness of a vertex** mean distance to other vertices

**Betweenness of a vertex** extent to which a vertex lies on paths between other vertices

4->figure/betweenness

5->figure/groups

6->figure/components

# Network measures

**Degree of a vertex** number of connections

**Authority of a vertex** number of important connections

**Closeness of a vertex** mean distance to other vertices

**Betweenness of a vertex** extent to which a vertex lies on paths between other vertices

**Group of vertices**

4->figure/betweenness

5->figure/groups

6->figure/components

# Network measures

**Transitivity of edges** Alice *friend of* Bob *friend of* Cat *friend of* Alice

1->figure/transitivity
2->figure/reciprocity
3->figure/similarity

# Network measures

**Transitivity of edges** Alice *friend of* Bob *friend of* Cat *friend of* Alice

**Reciprocity of edges** Alice *friend of* Bob *friend of* Alice

1->figure/transitivity
2->figure/reciprocity
3->figure/similarity

# Network measures

**Transitivity of edges** Alice *friend of* Bob *friend of* Cat *friend of* Alice

**Reciprocity of edges** Alice *friend of* Bob *friend of* Alice

**Similarity of vertices** extent to which the *neighbourhood* of vertices is similar

1->figure/transitivity
2->figure/reciprocity
3->figure/similarity

# Network measures

**Transitivity of edges** Alice *friend of* Bob *friend of* Cat *friend of* Alice

**Reciprocity of edges** Alice *friend of* Bob *friend of* Alice

**Similarity of vertices** extent to which the *neighbourhood* of vertices is similar

**Homophily of vertices** tendency to associate with similar vertices

1->figure/transitivity
2->figure/reciprocity
3->figure/similarity

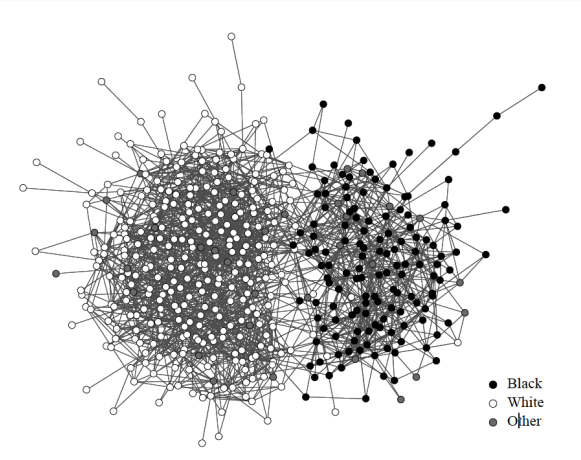



***Figure 6:*** *Friendship network at a US high school (***newman_networks_2010***).*

# Community detection

The goal of a community detection algorithm is simply to separate nodes into groups that have only a few edges *between* them and many edges *within*.

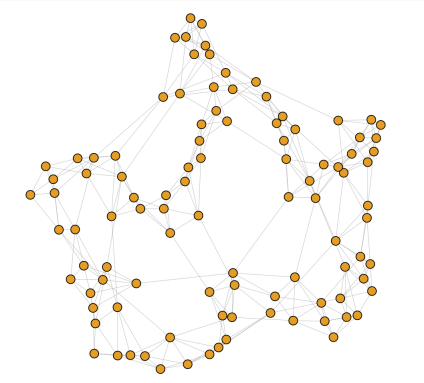
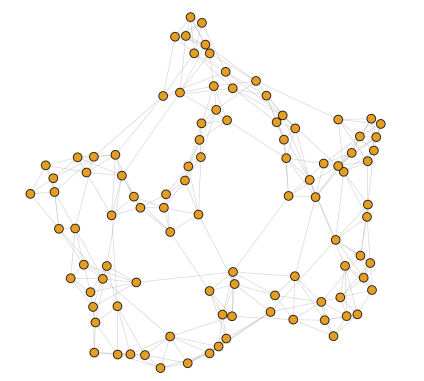


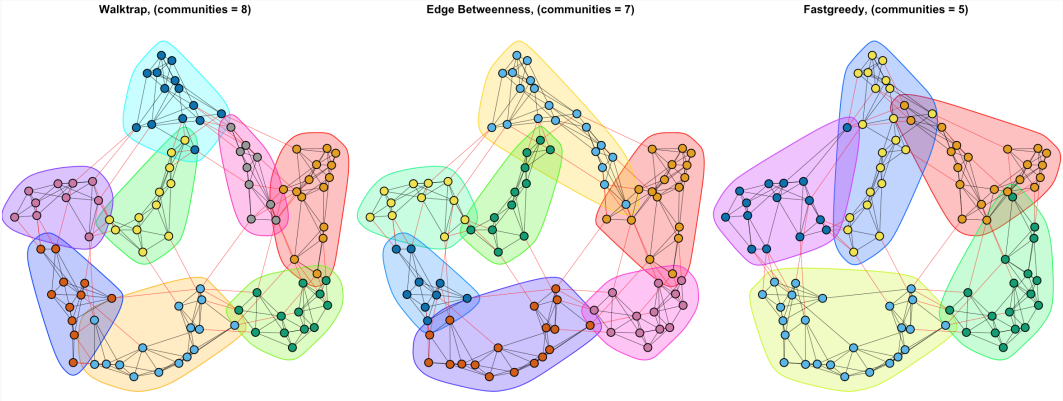
***Figure 7:*** *A randomly generated network with 100 vertices and 300 edges*

# Community detection

How many communities do you see in this network? (Go to: socrative.com, room: BAILO)



# Community detection



INFORMATION, COMMUNICATION & SOCIETY, 2017
VOL. 20, NO. 11, 1660–1679
http://dx.doi.org/10.1080/1369118X.2016.1252410



# Hybrid social and news media protest events: from #MarchinMarch to #BusttheBudget in Australia*

Francesco Bailo and Ariadne Vromen

Department of Government and International Relations, University of Sydney, Sydney, Australia

**ABSTRACT**

Public protest events are now both social media and news media events. They are deeply entangled, with news media actors – such as journalists or news organisations – directly participating in the protest by tweeting about the event using the protest hashtag; and social media actors sharing news items published online by professional news agencies. Protesters have always deployed tactics to engage the media and use news media agencies' resources to amplify their reach, with the dual aim of mobilising new supporters and adding their voice to public, mediatised debate. When protest moves between a physical space and a

**ARTICLE HISTORY**
Received 1 December 2015
Accepted 11 October 2016

**KEYWORDS**
Social movements; social media; news; social networking

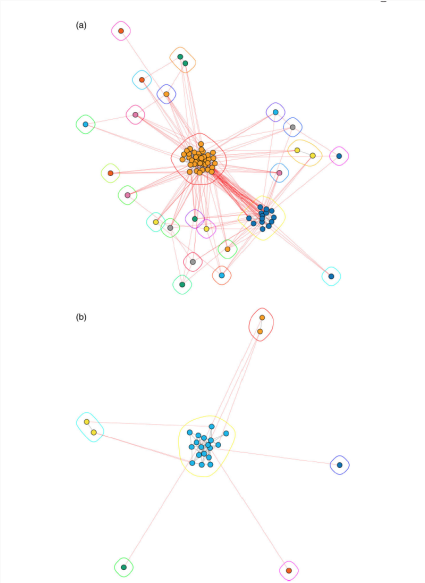
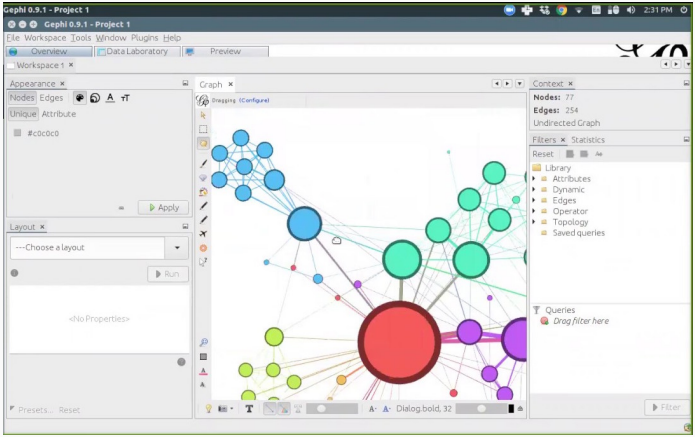
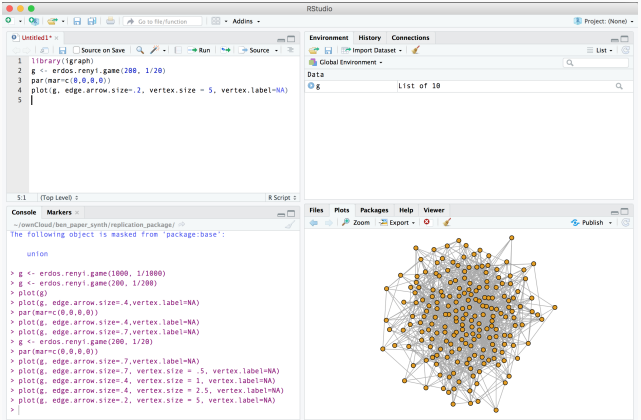



**Figure 4.** Mutual relations among friends and communities of a news media account (a) and regular users (b).

## **Easy, small n**: Gephi (gephi.org)



**Hard, big n** igraph package (igraph.org) in R (www.r-project.org) or Python (www.python.org)



Getting started bibliography:

**Easy** **scott_what_2012, scott_what_2012, scott_what_2012**

**Important** **marin_social_2011, marin_social_2011, marin_social_2011**

**Hard** **newman_networks_2010, newman_networks_2010, newman_networks_2010**

Tutorials for beginners by Katherine Ognyanova (Rutgers University):

- Network visualisation with Gephi (kateto.net/sunbelt2016)
- Network visualization with R (kateto.net/network-visualization)
- Network Analysis and Visualization with R and igraph (kateto.net/networks-r-igraph)

# Text analysis

Quantitative text analysis is necessary when the manual coding of documents in not feasible or acceptable.

When you face a large **corpus** of **documents**, you might want some methods to automatically:

1. Find patterns within the documents,
2. Compare (and maybe group) documents.

# Finding patterns

A textual pattern is as simple as `dog`.

- Finding patterns doesn't involve any statistical analysis.
- But you might need to use of regular expressions (a.k.a. 'regex') if your pattern is complex.

# Finding patterns

Let's say, that you want to find in your corpus all the instances of `dog` and `cat`.

**You want to find** 'I have two dogs and a cat' or 'Cats are felines'

(link to interactive example)

# Finding patterns

Let's say, that you want to find in your corpus all the instances of `dog` and `cat`.

**You want to find** 'I have two dogs and a cat' or 'Cats are felines'

**But you don't want to find** 'the categorization of syntactic categories'

(link to interactive example)

# Finding patterns

Let's say, that you want to find in your corpus all the instances of `dog` and `cat`.

**You want to find** 'I have two dogs and a cat' or 'Cats are felines'

**But you don't want to find** 'the categorization of syntactic categories'

You need a regular expression like: `\b(cats?|dogs?)\b`

(link to interactive example)

# Finding patterns

A few simple regex topics:

**Quantifier** `?`

Exercise: Go to `regexr.com/3os9b` (not with Explorer) and enter a regular expression to match 'France' but also 'French'.

# Finding patterns

A few simple regex topics:

**Quantifier** `?`

- `abc?` matches a string that has ‘ab’ followed by zero or one ‘c’

Exercise: Go to `regexr.com/3os9b` (not with Explorer) and enter a regular expression to match ‘France’ but also ‘French’.

# Finding patterns

A few simple regex topics:

**Quantifier** `?`

- `abc?` matches a string that has ‘ab’ followed by zero or one ‘c’

**OR operator** `|`

Exercise: Go to `regexr.com/3os9b` (not with Explorer) and enter a regular expression to match ‘France’ but also ‘French’.

# Finding patterns

A few simple regex topics:

**Quantifier** `?`

- `abc?` matches a string that has ‘ab’ followed by zero or one ‘c’

**OR operator** `|`

- `a(b|c)` matches a string that has ‘a’ followed by ‘b’ or ‘c’

Exercise: Go to `regexr.com/3os9b` (not with Explorer) and enter a regular expression to match ‘France’ but also ‘French’.

# Finding patterns

A few simple regex topics:

**Quantifier** `?`

- `abc?` matches a string that has ‘ab’ followed by zero or one ‘c’

**OR operator** `|`

- `a(b|c)` matches a string that has ‘a’ followed by ‘b’ or ‘c’

**Boundaries** `\b`

Exercise: Go to `regexr.com/3os9b` (not with Explorer) and enter a regular expression to match ‘France’ but also ‘French’.

# Finding patterns

A few simple regex topics:

**Quantifier** `?`

- `abc?` matches a string that has ‘ab’ followed by zero or one ‘c’

**OR operator** `|`

- `a(b|c)` matches a string that has ‘a’ followed by ‘b’ or ‘c’

**Boundaries** `\b`

- `\babc\b` matches only a whole word

Exercise: Go to `regexr.com/3os9b` (not with Explorer) and enter a regular expression to match ‘France’ but also ‘French’.

# Comparing documents

Comparing documents involves statistical analysis and matrix algebra (while finding patterns doesn't). It usually relies on Natural-language processing (NLP), the branch of computer science that studies the human language and its interactions with the machines.

In its most primordial application, NLP treats documents as **bag-of-words**:

- The *position* of terms within the document is disregarded,
- What counts is the *frequency* of the terms.

# Comparing documents

Let's see how we process documents in a common NLP application.

- We remove from the documents all the stop-words;

# Comparing documents

Let's see how we process documents in a common NLP application.

- We remove from the documents all the stop-words;
- doc1 = "drugs hospitals doctors"
  doc2 = "smog pollution environment"
  doc3 = "doctors hospitals healthcare"
  doc4 = "pollution environment water"

# Comparing documents

Let's see how we process documents in a common NLP application.

- We remove from the documents all the stop-words;
- doc1 = "drugs hospitals doctors"
  doc2 = "smog pollution environment"
  doc3 = "doctors hospitals healthcare"
  doc4 = "pollution environment water"
- We count the frequency of each term in each document an we produce a term-document matrix

# Comparing documents

| | doc1 | doc2 | doc3 | doc4 |
|---|---|---|---|---|
| doctor | 1 | 0 | 1 | 0 |
| drug | 1 | 0 | 0 | 0 |
| environ | 0 | 1 | 0 | 1 |
| healthcar | 0 | 0 | 1 | 0 |
| hospit | 1 | 0 | 1 | 0 |
| pollut | 0 | 1 | 0 | 1 |
| smog | 0 | 1 | 0 | 0 |
| water | 0 | 0 | 0 | 1 |

***Table 1:*** *Term-document matrix. Terms were stemmed.*

- Nvivo (www.qsrinternational.com/nvivo)
- Regular Expression (regexr.com)
- R (www.r-project.org) or Python (www.python.org)

**Introductory** jockers_text_2014, jockers_text_2014, jockers_text_2014

**Introductory** bird_natural_2009, bird_natural_2009, bird_natural_2009

**Hard** manning_introduction_2008, manning_introduction_2008, manning_introduction_2008

# Ethics

# Issues with relational data

**Online Privacy as a Collective Phenomenon**

Emre Sarigol
ETH Zurich
Weinbergstrasse 56/58
Zurich, Switzerland
semre@ethz.ch

David Garcia
ETH Zurich
Weinbergstrasse 56/58
Zurich, Switzerland
dgarcia@ethz.ch

Frank Schweitzer
ETH Zurich
Weinbergstrasse 56/58
Zurich, Switzerland
fschweitzer@ethz.ch

**ABSTRACT**

The problem of online privacy is often reduced to individual decisions to hide or reveal personal information in online social networks (OSNs). However, with the increasing use of OSNs, it becomes more important to understand the role of the social network in disclosing personal information that a user has not revealed voluntarily: How much of our private information do our friends disclose about us, and how much of our privacy is lost simply because of online social interaction? Without strong technical effort, an OSN may be able to exploit the assortativity of human private features, this way constructing shadow profiles with information that users chose not to share. Furthermore, because many users share their phone and email contact lists, this allows an OSN to create full shadow profiles for people who do not even have an account for this OSN.

We empirically test the feasibility of constructing shadow profiles of sexual orientation for users and non-users, using

**Categories and Subject Descriptors**

H.1.2 [Information Systems]: Models and principles—*User/machine Systems*

**General Terms**

Data mining, Privacy, Social Systems

**Keywords**

Privacy; Shadow Profiles; Prediction

**1. INTRODUCTION**

Our society is increasingly grounded on information and communication technologies, in which protecting one's privacy might not be an individual choice [12]. In online social networks (OSNs), the characteristics of each user is determined primarily by its connections, rather than by its in-

*Figure 8:* **sarigol_online_2014, sarigol_online_2014, sarigol_online_2014**

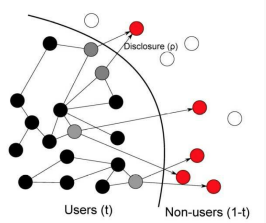



Figure 4: Schema of the full shadow profile construction problem.

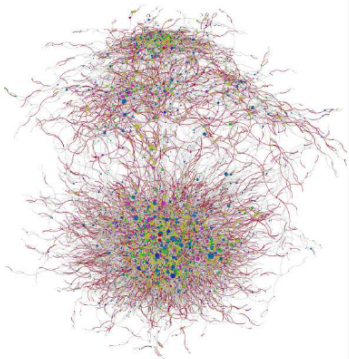


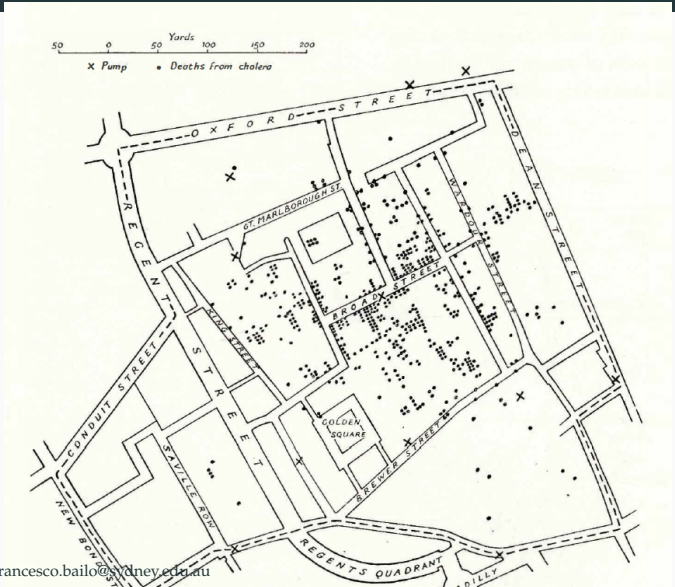
**Figure 1: The network for a subset of Friendster users. The red edges represent assortativity, where the endpoint nodes are in the same sexual orientation class. The node colors correspond to the sexual orientation class.**

# Ethics in the digital age: Open issues

- Public and Private space. What about online fora (e.g. Facebook public pages?)
- Informed consent.
- Right to privacy. But who owns the data?

# Bonus: Spatial analysis

# Spatial analysis




Redrawing of John Snow's map of cases of cholera during the London outbreak of 1854 (**tufte_visual_2001**)

# Tool for spatial analysis

- QGIS (qgis.org)