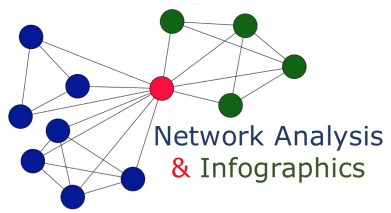
Network Data Collection

Dr Daniele Rotolo

SPRU (Science Policy Research Unit)
Business School
University of Sussex



Week 3

Learning Outcomes

	Learning outcome	Assessment mode
1	Explain the concept of network and list the main network indicators	ESS
2	Describe and apply the major techniques for the collection of network data and their statistical analysis	ESS, GPN + GWS
3	Identify the main characteristics of networks by means of network measures	ESS, GPN + GWS
4	Employ network analysis techniques to produce network data-based infographics	GPN + GWS

Note: ESS: Essay; GPN: Group Presentation; GWS: Group Written Submission

Overview

1. Data structure
2. Data collection
3. Data sources
4. Missing data and measurement challenges

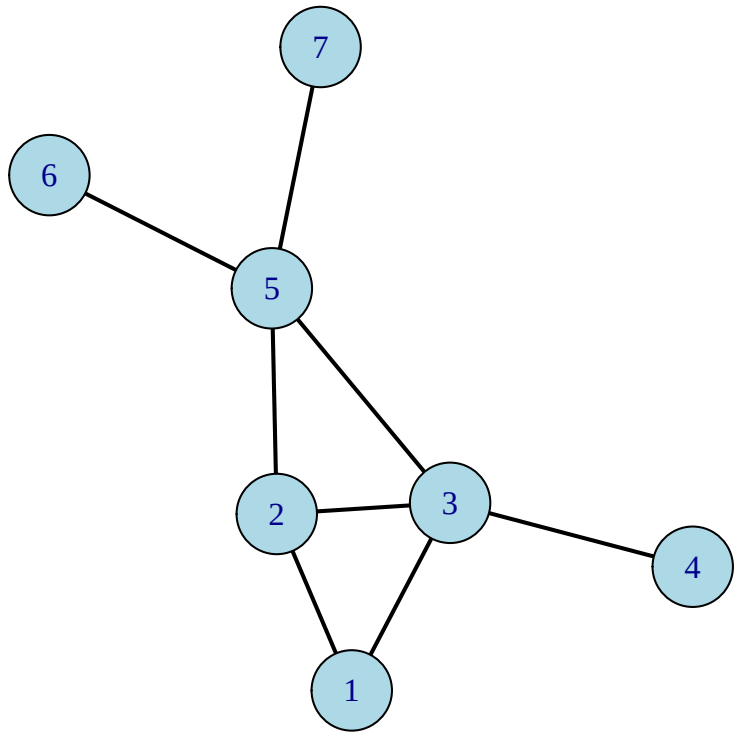
Defining a network [recap]

Defining a network [recap]

- A graph is defined as:

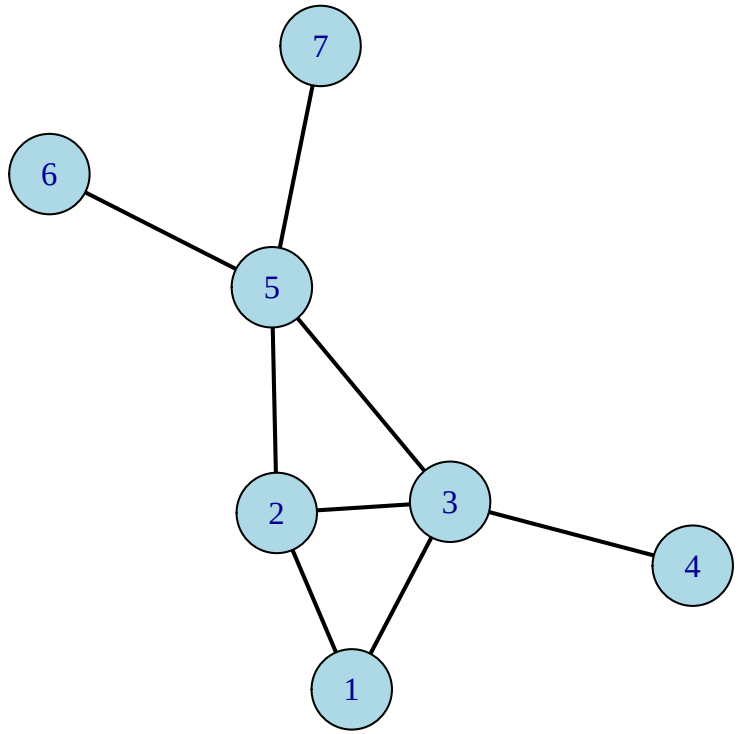
$$G(N, E)$$

- N nodes (or vertices),
  $N = n_1, n_2, ..., n_N$
- E edges (or links, ties),
  $E = e_1, e_2, ..., e_E$
- Example: $G(7, 8)$
- "A network consists of a graph and additional information on the vertices or the lines of the graphs"
  [de Nooy et al., 2005]



Defining a network [recap]

- Tie directionality: Undirected vs. directed networks
- Tie value: Unweighted vs. weighted networks
- Adjacency matrix: The (symmetric or asymmetric) matrix representing the connections among nodes
- Definitions
  - Dyad, triad
  - Subgraph: line- or node-generated
  - Walk, trail, tour, path, shortest path (and geodesic distance)
- Types of networks
  - Bipartite/2-mode networks
  - Multiplex/Multigraph networks
  - Ego-networks



$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & a_{ij} & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}$$

Data structure

Data structure

Variable analysis

	Composition variables (attributes)			
Case	Variable 1	Variable 2	$\cdots$	Variable K
1				
2				
.				
.				
.				
N				

Data structure
Variable analysis: Example

The Orange dataset includes data on the growth of orange trees

- 35 observations (5 trees)
- 3 variables

```
1 data()
2 Orange
3 plot(Orange$age,
4      Orange$circumference)
```

slide_data_collection.R | R data sets

```
Data sets in package 'datasets':

AirPassengers                Monthly Airline Passenger Numbers 1949-1960
BJsales                      Sales Data with Leading Indicator
BJsales.lead (BJsales)       Sales Data with Leading Indicator
BOD                          Biochemical Oxygen Demand
CO2                          Carbon Dioxide Uptake in Grass Plants
ChickWeight                  Weight versus age of chicks on different diets
DNase                        Elisa assay of DNase
EuStockMarkets               Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde                 Determination of Formaldehyde
HairEyeColor                 Hair and Eye Color of Statistics Students
Harman23.cor                 Harman Example 2.3
Harman74.cor                 Harman Example 7.4
Indometh                     Pharmacokinetics of Indomethacin
InsectSprays                 Effectiveness of Insect Sprays
JohnsonJohnson               Quarterly Earnings per Johnson & Johnson Share
LakeHuron                    Level of Lake Huron 1875-1972
LifeCycleSavings             Intercountry Life-Cycle Savings Data
Loblolly                     Growth of Loblolly pine trees
Nile                         Flow of the River Nile
Orange                       Growth of Orange Trees
OrchardSprays                Potency of Orchard Sprays
PlantGrowth                  Results from an Experiment on Plant Growth
Puromycin                    Reaction Velocity of an Enzymatic Reaction
Seatbelts                    Road Casualties in Great Britain 1969-84
Theoph                       Pharmacokinetics of Theophylline
Titanic                      Survival of passengers on the Titanic
ToothGrowth                  The Effect of Vitamin C on Tooth Growth in Guinea Pigs
UCBAdmissions                Student Admissions at UC Berkeley
UKDriverDeaths               Road Casualties in Great Britain 1969-84
UKgas                        UK Quarterly Gas Consumption
USAccDeaths                  Accidental Deaths in the US 1973-1978
USArrests                    Violent Crime Rates by US State
USJudgeRatings               Lawyers' Ratings of State Judges in the US Superior Court
USPersonalExpenditure        Personal Expenditure Data
UScitiesD                    Distances Between European Cities and Between US Cities
VADeaths                     Death Rates in Virginia (1940)
WWWusage                     Internet Usage per Minute
WorldPhones                  The World's Telephones
ability.cov                  Ability and Intelligence Tests
airmiles                     Passenger Miles on Commercial US Airlines, 1937-1960
airquality                   New York Air Quality Measurements
anscombe                     Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu                       The Joyner-Boore Attenuation Data
attitude                     The Chatterjee-Price Attitude Data
austres                      Quarterly Time Series of the Number of Australian Residents
beaver1 (beavers)            Body Temperature Series of Two Beavers
```

Data structure

Variable analysis: Example

The Orange dataset includes data on the growth of orange trees

- 35 observations (5 trees)
- 3 variables

```
1 data()
2 Orange
3 plot(Orange$age,
4      Orange$circumference)
```

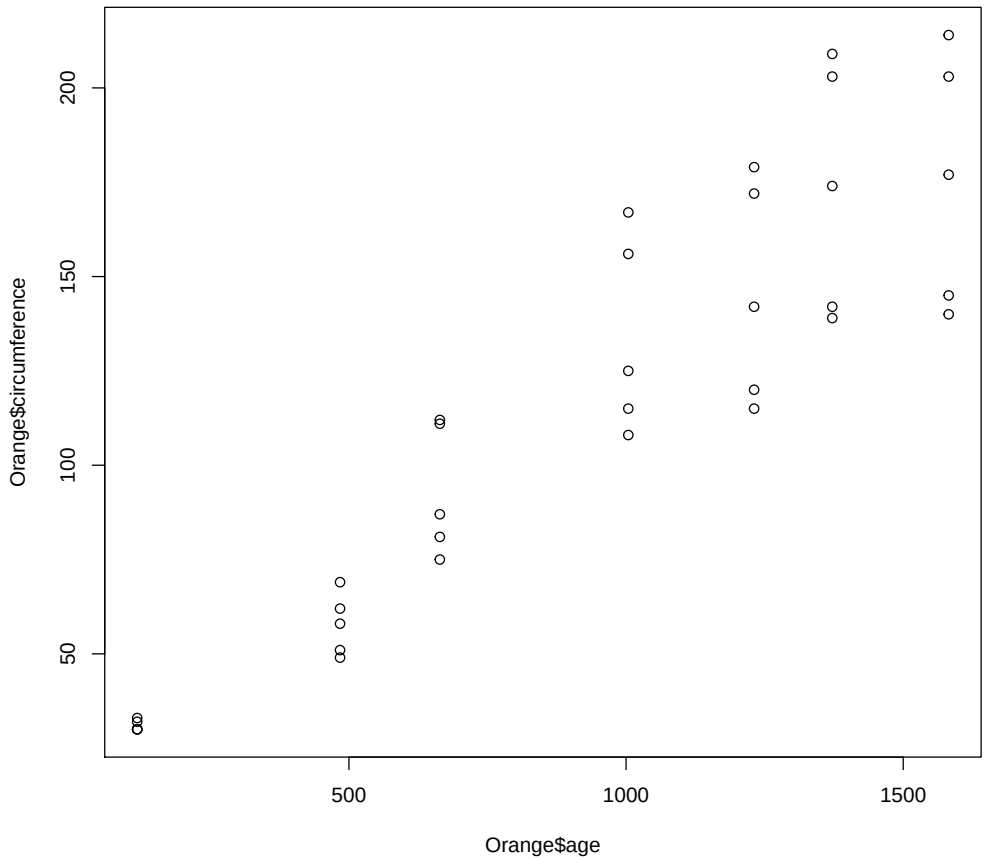
```
> Orange
   Tree  age circumference
1     1  118            30
2     1  484            58
3     1  664            87
4     1 1004           115
5     1 1231           120
6     1 1372           142
7     1 1582           145
8     2  118            33
9     2  484            69
10    2  664           111
11    2 1004           156
12    2 1231           172
13    2 1372           203
14    2 1582           203
15    3  118            30
16    3  484            51
17    3  664            75
18    3 1004           108
19    3 1231           115
20    3 1372           139
21    3 1582           140
22    4  118            32
23    4  484            62
24    4  664           112
25    4 1004           167
26    4 1231           179
27    4 1372           209
28    4 1582           214
29    5  118            30
30    5  484            49
31    5  664            81
32    5 1004           125
33    5 1231           142
34    5 1372           174
35    5 1582           177
```

Data structure
Variable analysis: Example

The Orange dataset includes data on the growth of orange trees

- 35 observations (5 trees)
- 3 variables

```
data()
Orange
plot(Orange$age,
     Orange$circumference)
```



Data structure

Network analysis

	Composition variables (attributes)			
Case	Variable 1	Variable 2	$\cdots$	Variable K
1				
2				
.				
.				
.				
N				

Composition variables (attributes)

Case	Variable 1	Variable 2	$\cdots$	Variable K
1				
2				
$\vdots$				
N				

Structural variables (adjacency matrix)

		Case			
		1	2	$\cdots$	N
Case	1				
	2				
	$\vdots$				
	N				

Data structure

Network analysis: Example

UKfaculty data on personal friendship in a UK faculty

- 81 individuals
- 817 directed and weighted connections
- Affiliation of each individual

```
library(igraphdata)
library(igraph)
data(UKfaculty)
UKfaculty
get.adjacency(UKfaculty)
ll <- layout_with_kk(UKfaculty)
plot(UKfaculty,
     layout = ll,
     vertex.size = 10,
     edge.arrow.size = 0.1)
```

Data structure

Network analysis: Example

UKfaculty data on personal friendship in a UK faculty

- 81 individuals
- 817 directed and weighted connections
- Affiliation of each individual

```
1  library(igraphdata)
2  library(igraph)
3  data(UKfaculty)
4  UKfaculty
5  get.adjacency(UKfaculty)
6  ll <- layout_with_kk(UKfaculty)
7  plot(UKfaculty,
8       layout = ll,
9       vertex.size = 10,
10      edge.arrow.size = 0.1)
```

```
> UKfaculty
IGRAPH 6f42903 D-W- 81 817 --
+ attr: Type (g/c), Date (g/c), Citation (g/c), Author (g/c), Group
| (v/n), weight (e/n)
+ edges from 6f42903:
 [1] 57->52 76->42 12->69 43->34 28->47 58->51  7->29 40->71  5->37 48->55  6->58
[12] 21-> 8 28->69 43->21 67->58 65->42  5->67 52->75 37->64  4->36 12->49 19->46
[23] 37-> 9 74->36 62-> 1 15-> 2 72->49 46->62  2->29 40->12 22->29 71->69  4-> 3
[34] 37->69  5-> 6 77->13 23->49 52->35 20->14 62->70 34->35 76->72  7->42 37->42
[45] 51->80 38->45 62->64 36->53 62->77 17->61  7->68 46->29 44->53 18->58 12->16
[56] 72->42 52->32 58->21 38->17 15->51 22-> 7 22->69  5->13 29-> 2 77->12 37->35
[67] 18->46 10->71 22->47 20->19 19->31 68->13 49->69 30->63  5->49 53->75 62->57
+ ... omitted several edges
```

Data structure

Network analysis: Example

UKfaculty data on personal friendship in a UK faculty

- 81 individuals
- 817 directed and weighted connections
- Affiliation of each individual

```r
library(igraphdata)
library(igraph)
data(UKfaculty)
UKfaculty
get.adjacency(UKfaculty)
ll <- layout_with_kk(UKfaculty)
plot(UKfaculty,
     layout = ll,
     vertex.size = 10,
     edge.arrow.size = 0.1)
```

```
> get.adjacency(UKfaculty)
81 x 81 sparse Matrix of class "dgCMatrix"

 [1,] . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ......
 [2,] . . . . . . . . . . . . . . 1 . . 1 . . 1 . . . 1 . . . 1 . 1 1 . . ......
 [3,] . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . ......
 [4,] 1 . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ......
 [5,] . . . . . 1 1 . 1 1 . 1 1 . . 1 . . . . . 1 1 . . . 1 1 . . . . . . ......
 [6,] . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . 1 1 . . . . ......
 [7,] . . . . . 1 . . . 1 . 1 1 . . 1 . 1 . . . . 1 . . . 1 . 1 1 . . . . ......
 [8,] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1 . . . ......
 [9,] . . 1 1 . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . ......
[10,] . . . . 1 . 1 . . . . . 1 . . . . . . . . 1 1 . . . 1 1 . . . . . . ......
[11,] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ......
[12,] . . . . . . . . . . . . . . . 1 . . . . . 1 1 . . . . . . . . . . . ......
[13,] . . . . 1 . 1 . . . . 1 . . . 1 . . . . 1 . 1 . . . 1 1 . . . . 1 . ......
[14,] . . . . . . . . . . . . . . . . . . . 1 . . . . . 1 . . . . . . . . ......
```

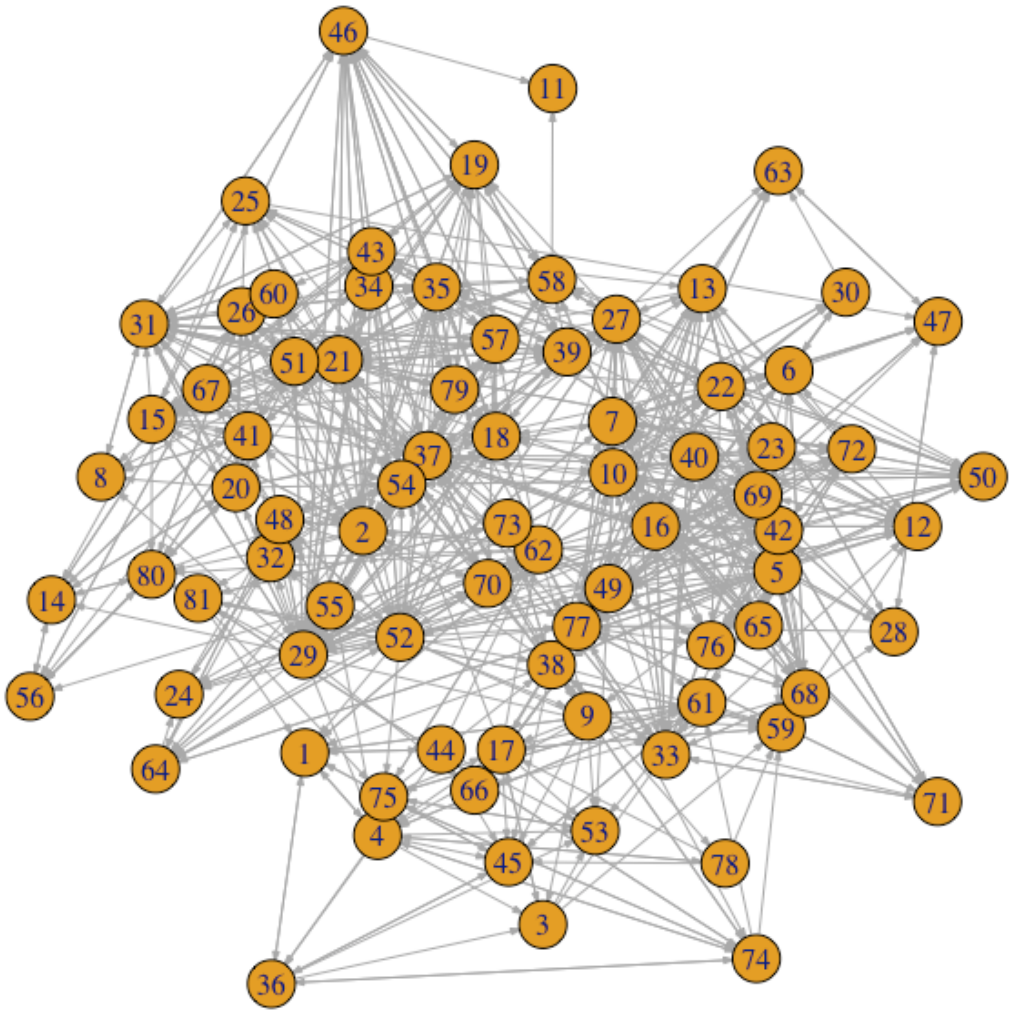
Data structure

Network analysis: Example

UKfaculty data on personal friendship in a UK faculty

- 81 individuals
- 817 directed and weighted connections
- Affiliation of each individual

```
1  library(igraphdata)
2  library(igraph)
3  data(UKfaculty)
4  UKfaculty
5  get.adjacency(UKfaculty)
6  ll <- layout_with_kk(UKfaculty)
7  plot(UKfaculty,
8       layout = ll,
9       vertex.size = 10,
10      edge.arrow.size = 0.1)
```



Data structure

Network analysis: Network data

- Network data are structured as case-case tables: matrices or lists
- Such a format represents ties between cases (i.e. nodes)
- These data are somewhat ready to be analysed

Case-case adjacency matrix

		Case			
		1	2	$\cdots$	N
Case	1				
	2				
	$\vdots$				
	N				

Case-case list

Case	Case
1	2
2	3
$\vdots$	$\vdots$
i	j
$\vdots$	$\vdots$

$i, j = \{1, 2, \cdots, N\}$
$i \neq j$

Data structure

Network analysis: Affiliation networks

- Data are often available in the form of affiliation networks

Case-affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Case	1				
	2				
	$\vdots$				
	N				

Case-affiliation list

Case	Affiliation
1	2
2	3
$\vdots$	$\vdots$
n	k
$\vdots$	$\vdots$

$n = \{1, 2, \cdots, N\}$
$k = \{1, 2, \cdots, K\}$

Data structure

Network analysis: Affiliation networks

- Data are often available in the form of affiliation networks
- Such a format describes 2-mode networks: individual-event, organization-R&D alliance, researcher-publication, etc.

Case-affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Case	1				
	2				
	$\vdots$				
	N				

Case-affiliation list

Case	Affiliation
1	2
2	3
$\vdots$	$\vdots$
n	k
$\vdots$	$\vdots$

$n = \{1, 2, \cdots, N\}$
$k = \{1, 2, \cdots, K\}$

Data structure

Network analysis: Affiliation networks

- Data are often available in the form of affiliation networks
- Such a format describes 2-mode networks: individual-event, organization-R&D alliance, researcher-publication, etc.
- We need to transform these data before we can analyse them

Case-affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Case	1				
	2				
	$\vdots$				
	N				

Case-affiliation list

Case	Affiliation
1	2
2	3
$\vdots$	$\vdots$
n	k
$\vdots$	$\vdots$

$n = \{1, 2, \cdots, N\}$
$k = \{1, 2, \cdots, K\}$

Data structure

Network analysis: Affiliation networks

Case-affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Case	1				
	2				
	$\vdots$				
	N				

Case-affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Case	1				
	2				
	$\vdots$				
	N				

(1)

Case-case adjacency matrix

		Case			
		1	2	$\cdots$	N
Case	1				
	2				
	$\vdots$				
	N				

Data structure

Network analysis: Affiliation networks

Case-affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Case	1				
	2				
	$\vdots$				
	N				

(1)

Case-case adjacency matrix

		Case			
		1	2	$\cdots$	N
Case	1				
	2				
	$\vdots$				
	N				

(2)

Affiliation-Affiliation adjacency matrix

		Affiliation			
		1	2	$\cdots$	K
Affiliation	1				
	2				
	$\vdots$				
	K				

Data structure

Network analysis: Affiliation networks (example)

R&D projects data

- 20 projects
- 21 orgnisations

Project	Partners	Technology
Proj01	U2, F1, NG1	TechA
Proj02	U1, NG4, F1	TechB
Proj03	NG3, NG1, F1	TechC
Proj04	NG3, NG4, F1	TechB
Proj05	U3, F1	TechB
Proj06	U3, F2	TechB
Proj07	U3, F3	TechC
Proj08	U3, U4	TechA
Proj09	F1	TechB
Proj10	U5	TechB
Proj11	U4, U5, U6	TechA
Proj12	U3, U7	TechB
Proj13	U7, G1	TechB
Proj14	U7, O1	TechD
Proj15	U7, G2	TechD
Proj16	G2, F3	TechC
Proj17	F3, O2	TechC
Proj18	O2, F4, NG2	TechB
Proj19	F4, U9, NG2	TechB
Proj20	NG2, U8	TechB

Data structure

Network analysis: Affiliation networks (example)

R&D projects data

- 20 projects
- 21 orgnisations

(1) Transforming the table into long format

Project	Partners	Technology
Proj01	U2	TechA
Proj01	F1	TechA
Proj01	NG1	TechA
Proj02	U1	TechB
Proj02	NG4	TechB
Proj02	F1	TechB
Proj03	NG3	TechC
Proj03	NG1	TechC
Proj03	F1	TechC
Proj04	NG3	TechB
Proj04	NG4	TechB
Proj04	F1	TechB
...	...	...
Proj17	F3	TechC
Proj17	O2	TechC
Proj18	O2	TechB
Proj18	F4	TechB
Proj18	NG2	TechB
Proj19	F4	TechB
Proj19	U9	TechB
Proj19	NG2	TechB
Proj20	NG2	TechB
Proj20	U8	TechB

Data structure

Network analysis: Affiliation networks (example)

R&D projects data

- 20 projects
- 21 orgnisations

(2) Co-occurence matrix

	F1	F2	F3	F4	G1	G2	NG1	NG2	NG3	NG4	O1	O2	U1	U2	U3	U4	U5	U6	U7	U8	U9
Proj01	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Proj02	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Proj03	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Proj04	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Proj05	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Proj06	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Proj07	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Proj08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Proj09	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Proj10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Proj11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Proj12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
Proj13	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Proj14	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
Proj15	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Proj16	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Proj17	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Proj18	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
Proj19	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1

Data structure

Network analysis: Affiliation networks (example)

R&D projects data

- 20 projects
- 21 orgnisations

(3) Matrix product

	Proj01	Proj02	Proj03	Proj04	Proj05	Proj06	Proj07	Proj08	Proj09	...
F1	1	1	1	1	1	0	0	0	1	...
F2	0	0	0	0	0	1	0	0	0	...
F3	0	0	0	0	0	0	1	0	0	...
F4	0	0	0	0	0	0	0	0	0	...
G1	0	0	0	0	0	0	0	0	0	...
...	...	...	...	...	...					

X

	F1	F2	F3	F4	G1	...
Proj01	1	0	0	0	0	...
Proj02	1	0	0	0	0	...
Proj03	1	0	0	0	0	...
Proj04	1	0	0	0	0	...
Proj05	1	0	0	0	0	...
Proj06	0	1	0	0	0	...
Proj07	0	0	1	0	0	...
Proj08	0	0	0	0	0	...
Proj09	1	0	0	0	0	...
...	...	...	...	...	...	

Data structure

Network analysis: Affiliation networks (example)

R&D projects data

- 20 projects
- 21 orgnisations

```
library(tidyverse)
library(readr)
setwd("YOUR WORKING DIRECTORY")
PR <- read_csv("proj_org.csv") %>%
  separate_rows(Partners,
                sep = ";")
PO <- table(PR$Project,
            PR$Partners)
OO <- t(PO)%*%PO
PP <- PO%*%t(PO)
```

(4) Adjacency matrix

	F1	F2	F3	F4	G1	G2	NG1	NG2	NG3	NG4	O1	O2	U1	U2	U3	U4	U5	U6	U7	U8	U9
F1	0	0	0	0	0	0	2	0	2	2	0	0	1	1	1	0	0	0	0	0	0
F2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F3	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
F4	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	1
G1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
G2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
NG1	2	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
NG2	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
NG3	2	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
NG4	2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
O1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
O2	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
U1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
U2	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
U4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
U5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
U6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
U7	0	0	0	0	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
U8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
U9	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Data collection

Data collection

Key questions

- Define a set of nodes
  - Which actors should be included?
  - Which are the relevant actors for the network?
  - Can we sample nodes/actors?
  - ...
- Define a set of ties between nodes
  - Which relations should be included?
  - Which relations are likely to be excluded?
  - ...

Case-case adjacency matrix

		Case			
		1	2	...	N
Case	1				
	2				
	⋮				
	N				

Data collection
Defining the set of nodes

In the case of relational data, we cannot independently sample actors

- If an actor is selected, all actors to whom the actor is connected should be included
- Network studies tend to focus on whole populations/samples of convenience
- We need to define the boundaries of our population/sample
  - In some cases, we have some *a priori* knowledge to identify the set of actors
    - employees in a firm
    - students in this module
    - ...
  - In other cases, drawing boundaries around a set of nodes is somewhat arbitrary
    - It may be difficult to understand whether a an actor belongs to the set
    - The population may be too large
    - The composition of actors may change over time (joiners and leavers)

[Laumann et al., 1989] and the 'boundary specification problem'

[Laumann et al., 1989] and the 'boundary specification problem'

- Realist approach: focus on actors to identify network boundaries as perceived by the actors themselves

  Example
  With whom you discuss important study related matters?
  Focus on a class/course may leave network ties outside the study

[Laumann et al., 1989] and the 'boundary specification problem'

- Realist approach: focus on actors to identify network boundaries as perceived by the actors themselves

  Example
  With whom you discuss important study related matters?
  Focus on a class/course may leave network ties outside the study

- Nominalist approach: focus on the theoretical concerns of the researcher/analyst

  Example
  How does degree centrality affect scientists' productivity in cancer research?
  Focus on a researchers that published in cancer

[Laumann et al., 1989] and the 'boundary specification problem'

- Realist approach: focus on actors to identify network boundaries as perceived by the actors themselves

  Example
  With whom you discuss important study related matters?
  Focus on a class/course may leave network ties outside the study

- Nominalist approach: focus on the theoretical concerns of the researcher/analyst

  Example
  How does degree centrality affect scientists' productivity in cancer research?
  Focus on a researchers that published in cancer

The research question guides the selection of the approach

Data collection

Defining the set of ties

Once we have identified a set of actors, we need to identify the corresponding set of ties between these actors

- Full-network method
- Snowball method
- Ego-centric method

Data collection
Defining the set of ties

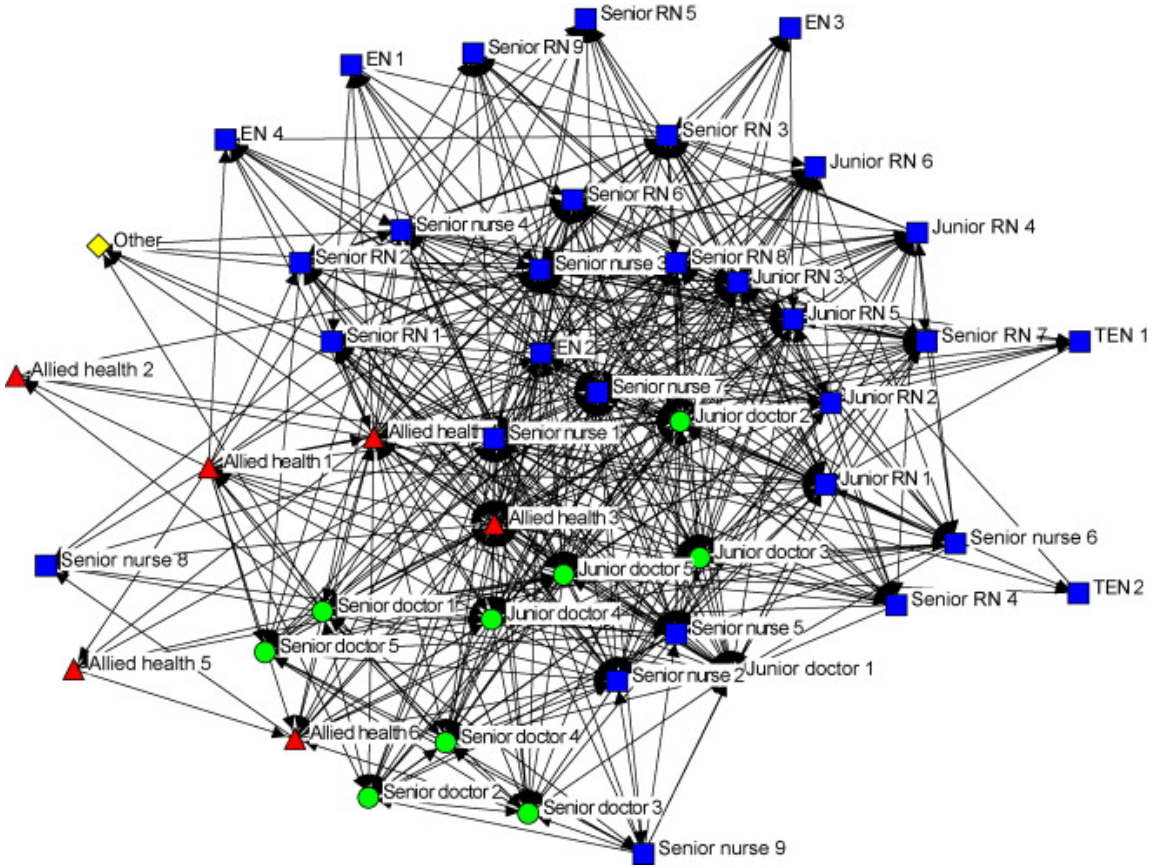
Full-network method

- We collect information about each tie between the actors in our population/sample
- This approach can generate a comprehensive map of a network
- Actors tend to establish a limited number of ties (limited attentional capabilities)
- Feasibility in the case of relatively small networks – time and resources to collect data (e.g. ties between people living in a city)

Data collection
Defining the set of ties

Example of full-network method

- Aim: To map medication advice-seeking interactions of doctors, nurses, allied health professionals
- Context: Renal ward of an Australian metropolitan teaching hospital
- Data collection: Questionnaires with the full list of staff(response rate 96%)



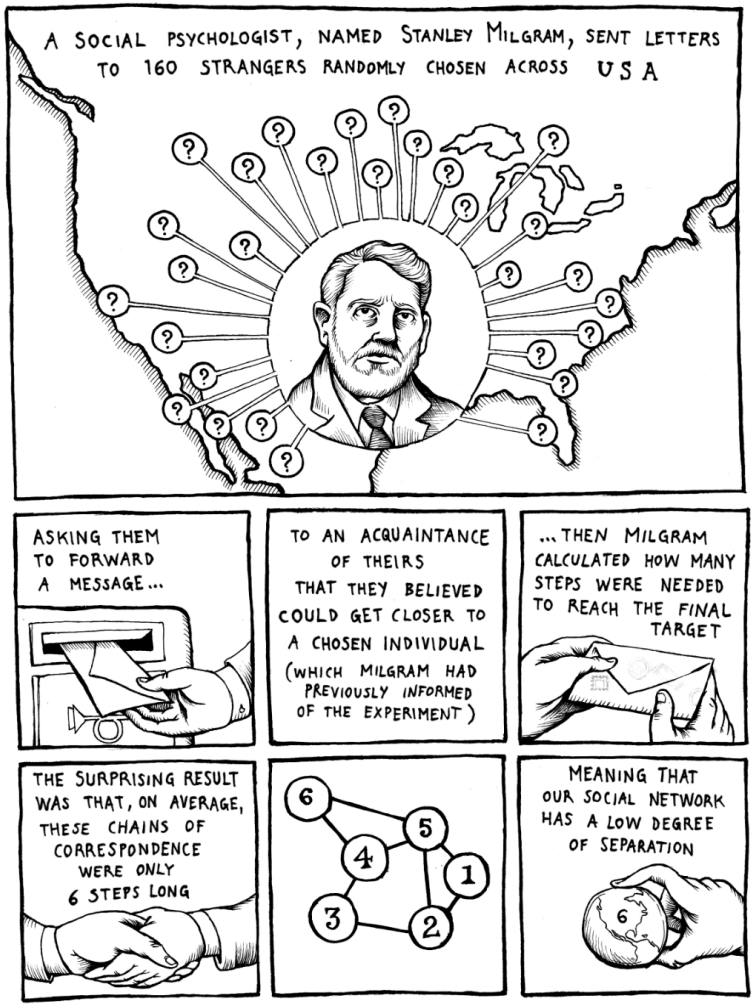
Source: [Creswick and Westbrook, 2010]

Snowball method

- We ask to each actor in our set to list some or all of their ties with other actors
- All the actors listed, but not included in the original set of actors are tracked down and asked for some or all of their ties
- The process stops when
  - no new actors are identified
  - the new actors are very 'marginal' to the set of actors under study
  - we have limitations of time and resources
- It is likely to require less time and resources than the full-network approach
- Isolated actors may not be captured (overestimation of network cohesion)
- A wrong starting set of actors may miss entire sub-sets of actors not linked to our starting set of actors (e.g. multiple components)

Data collection
Defining the set of ties

Example of snowball-network method

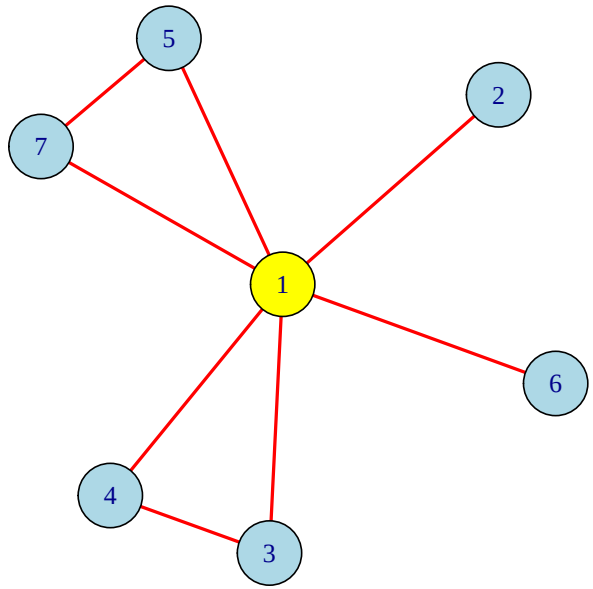


Source: https://matteofarinella.wordpress.com/2010/09/05/a-small-world-theory/
Video: `https://youtu.be/NberyK6kt8c`

Data collection
Defining the set of ties

Ego-centric method

- We start with a set of focal actors (egos), and identify the actors to which they are connected
- Useful when we have feasibility issues with other methods
- We ask to the egos to report which of their direct peers are also tied to one another
- Relatively reliable overview of at least the local neighbourhoods in which the selected actors are embedded (e.g. redundancy and constraint)
- Not useful to estimate network-level measures



Data sources

Data sources

- Questionnaires (or surveys)
- Interviews
- Observations
- Archival data

Data sources
Questionnaires

- Commonly used to map relatively small social networks
- Respondents report their ties with other actors in the network
  - who they like
  - who they seek advice from
  - who they collaborate with
  - ...
- Three main choices
  - Choice 1: predefined or undefined list of names
  - Choice 2: constraints on the number of ties
  - Choice 3: value of the ties

Data sources

Questionnaires - Choice 1: Roster

- All the actors in a network are represented in a roster
- Respondents indicate the existence of ties with all the other actors in the network

Roster

Name	Friend	Advice
Peter Parker	☐	☐
Tony Stark	☐	☐
Bruce Banner	☐	☐
...	...	...

Data sources

Questionnaires - Choice 1: Roster

Advantages

- Respondents recognise names
- Minimum data cleaning (name spelling)

Limitations

- Feasible for small networks
- Need of well-delineated boundaries
  - students in this module
  - employees in a firms' department/business unit
  - ...

Data sources

Questionnaires - Choice 1: Free recall

- The list of names is generated by respondents
- Respondents name the actors with which they have certain ties

Free recall

Name	Friend	Advice
---------	☐	☐
---------	☐	☐
---------	☐	☐
---------	☐	☐
...	...	...

Data sources

Questionnaires - Choice 1: Free recall

Advantages

- Respondents can name actors not included in the initial delineation of the network boundaries (snowball)
- Useful when the boundaries of the network are not clear
  - friendship network in high school
  - researchers working in a discipline
  - ...

Limitations

- Respondents may be unwilling or uncomfortable to name actors
- Data cleaning (name spelling)
- Time and resources to contact 'snowballed' actors

Free choice

- Respondents have no constraints on the number of actors they can name
- No constraints on the maximum number of ties of respondents' ego-network
- *Please identify people who you have exchanged ideas with most often ...*

Fixed choice

- Respondents have constraints on the number of actors they can nominate
- There is a maximum number of ties that each actor can have in the network
- *Please identify up to N people who you have exchanged ideas with most often ...*

Rating

- Respondents are asked to assign a value to each tie

Complete ranking

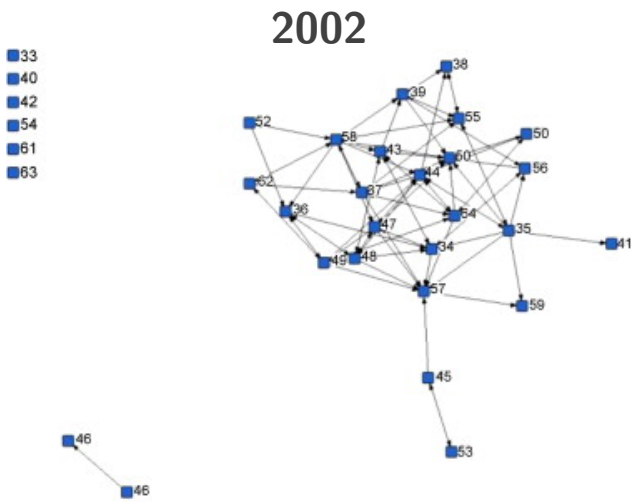
- Respondents are asked to rank all their ties

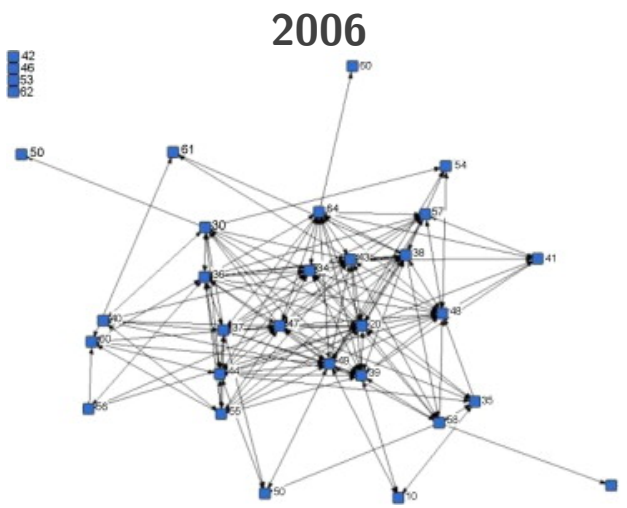
Data sources
Interviews

- Face-to-face, over the phone, videoconferencing
- Used when survey are not feasible (willing to participate)
- Gather ego-network data
- Three main choices (as in the case of questionnaires)
  - Choice 1: predefined or undefined list of names
  - Choice 2: constraints on the number of ties
  - Choice 3: value of the ties

Data sources
Interviews (Example)

- Study of a wine cluster in Chile [Giuliani, 2013]
- Network data collected with interviews (50 interviews)
- Examples of questions in the interviews
  - If you are in a critical situation and need technical advice, to which of the local firms mentioned in the roster do you turn?
  - Which of the following firms do you think have benefited from technical support provided by your firm?





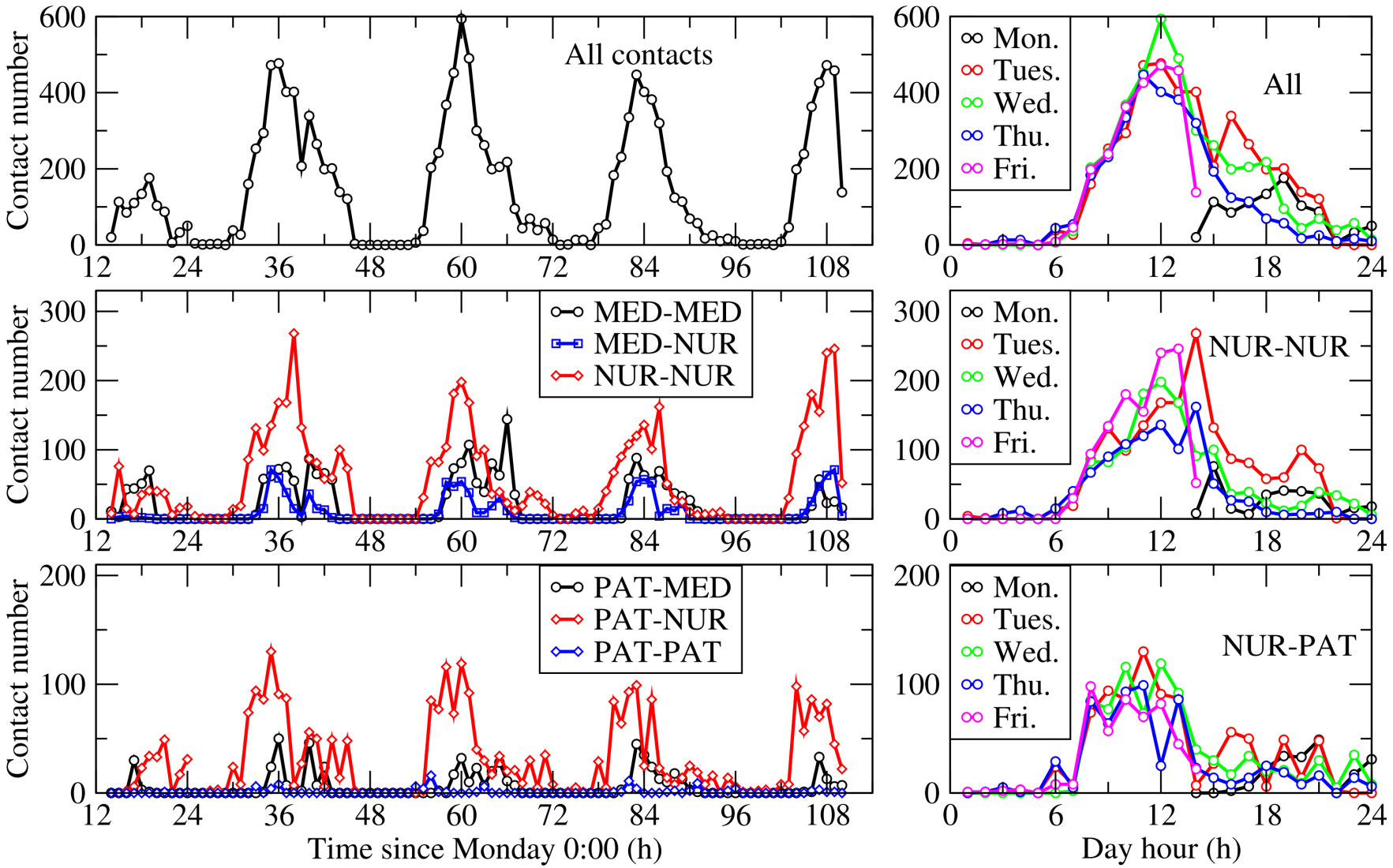
Source:[Giuliani, 2013]

Data sources
Observations

- Direct observation of the interaction among actors
- Used in field research to study relatively small groups
  - non-human primates
  - attending events (e.g. SPRU seminars)
  - ...
- Challenges in observing multiple actors simultaneously

Data sources

Observations: Example

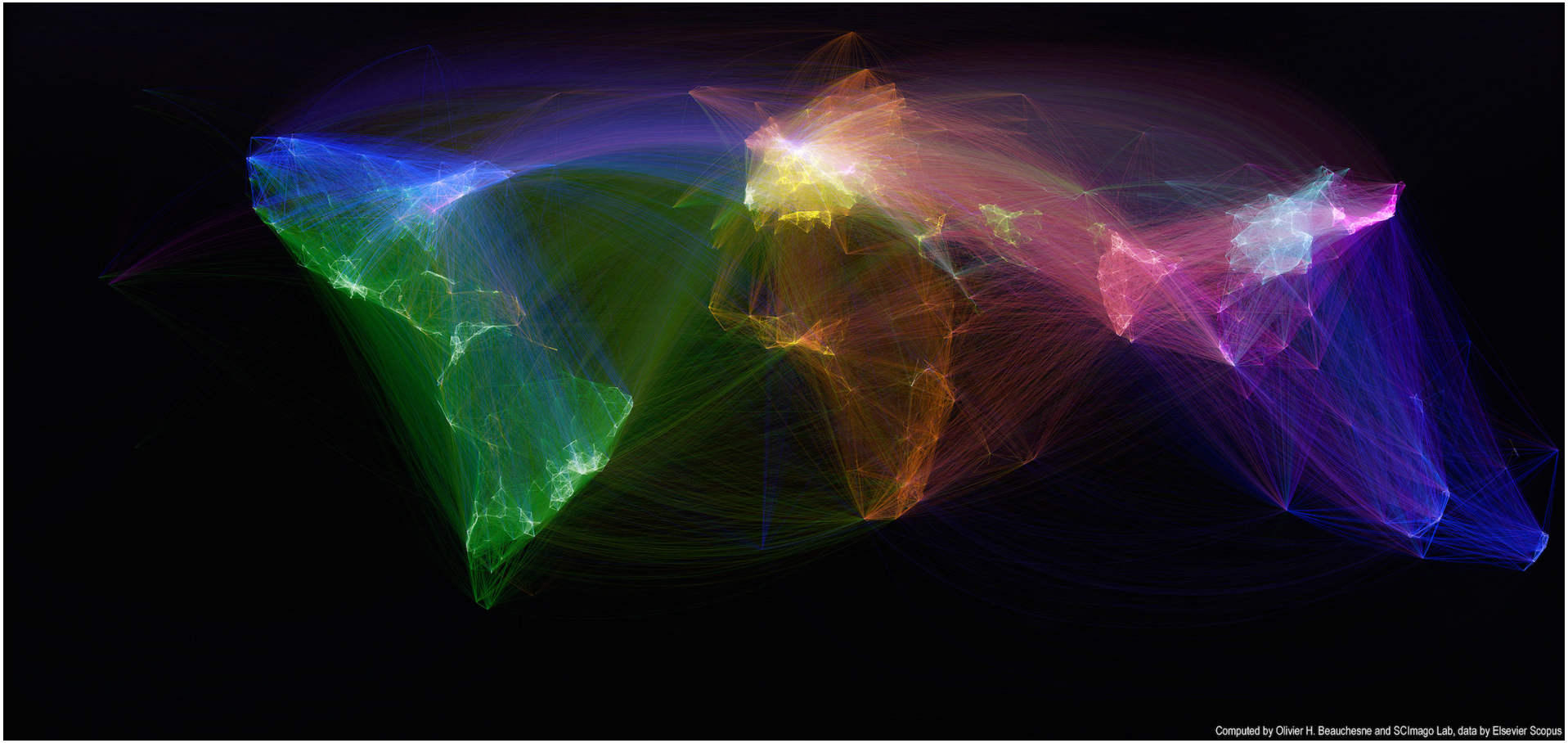


Source: Patient, medical doctor, nurse interactions and infection transmission [Vanhems et al., 2013]

Data sources
Archival data

- Records of interactions between actors
  - Political interactions
  - Published articles
  - Collaboration on patents
  - Citation patterns (publications, patents, scientometrics)
  - Bank transactions
  - ...
- Access to large datasets
- Data proxy interactions (validity)

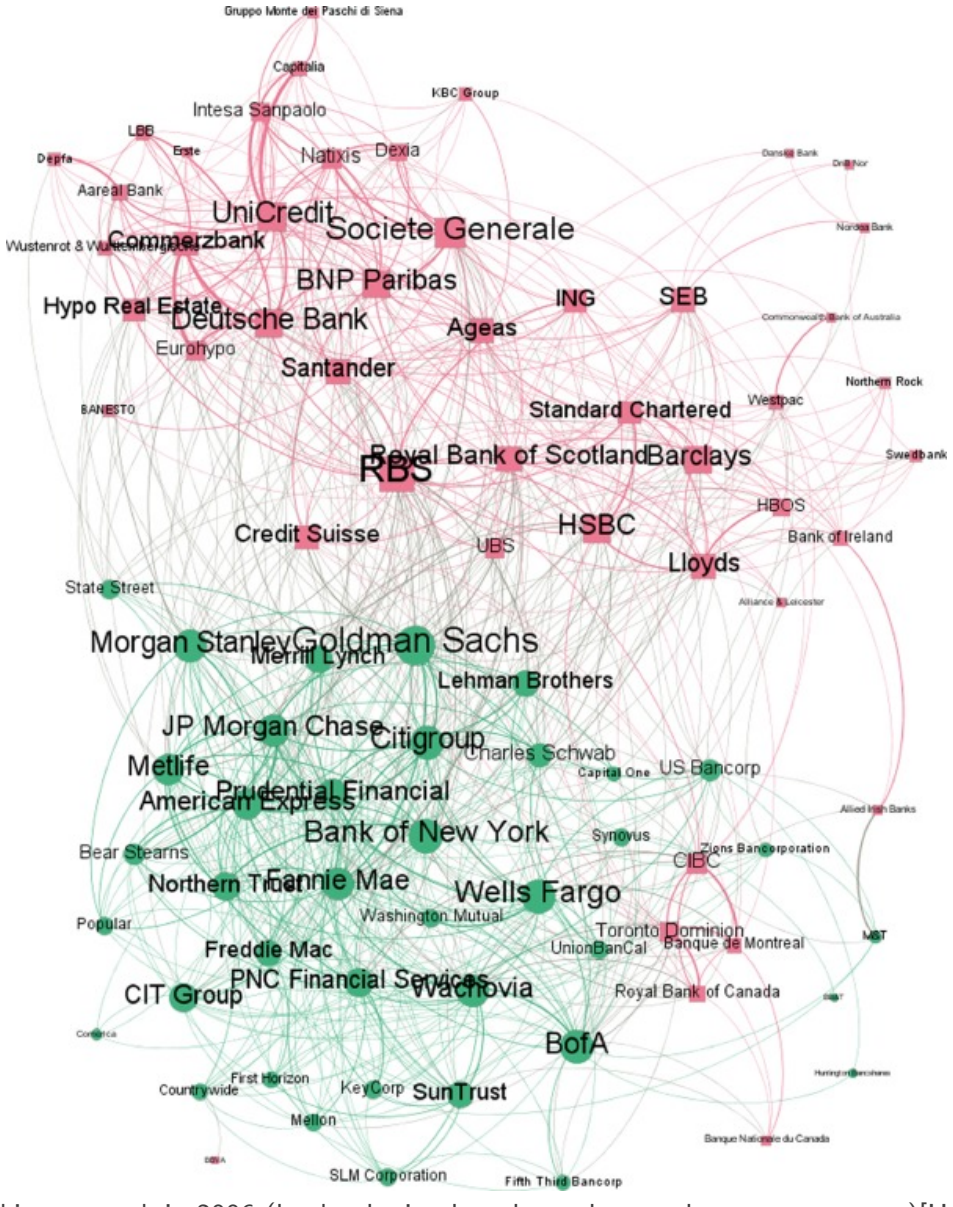
Data sources
Archival data: Example



Source: Co-authorship at the city level (SCOPUS 2008-2012) [http://olihb.com]

Data sources

Archival data: Example



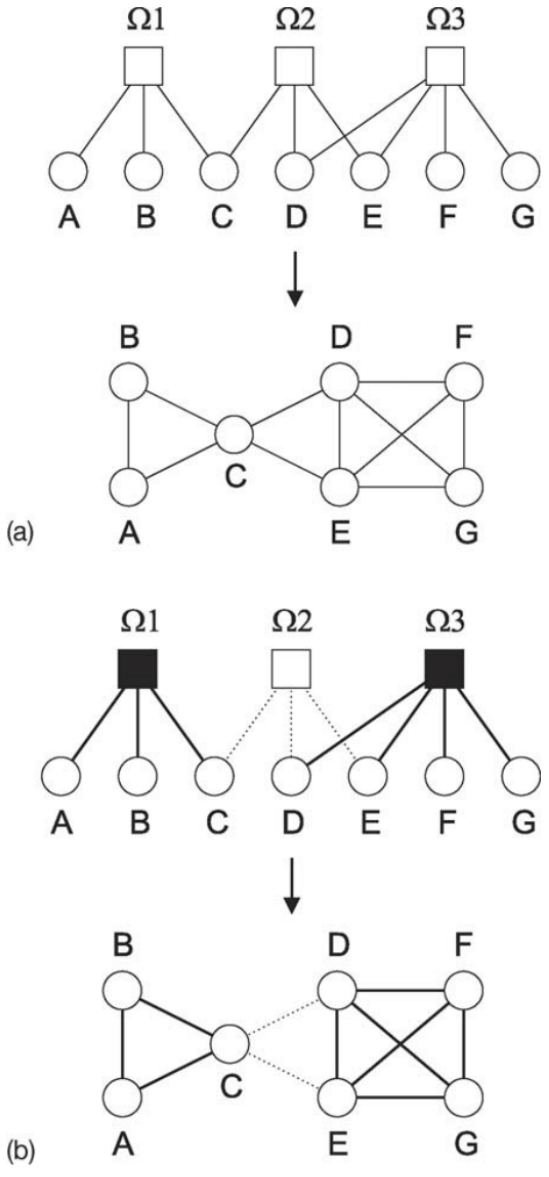
Source: Global banking network in 2006 (banks sharing board members and top management)[Houston et al., 2018]

Missing data and measurement challenges

Missing data and measurement challenges

Mechanisms

- Network boundary specification (non-inclusion of actors or affiliations)

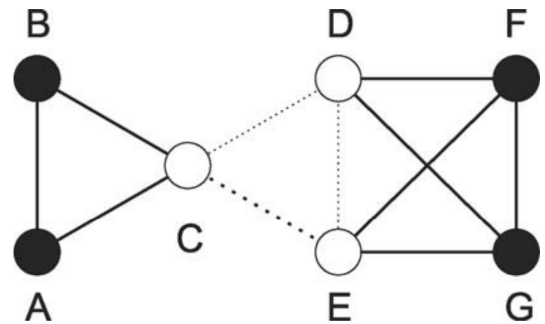


Source: Failure to include a context of interaction or affiliation [Kossinets, 2006]

Missing data and measurement challenges

Mechanisms

- Network boundary specification (non-inclusion of actors or affiliations)
- Survey non-response

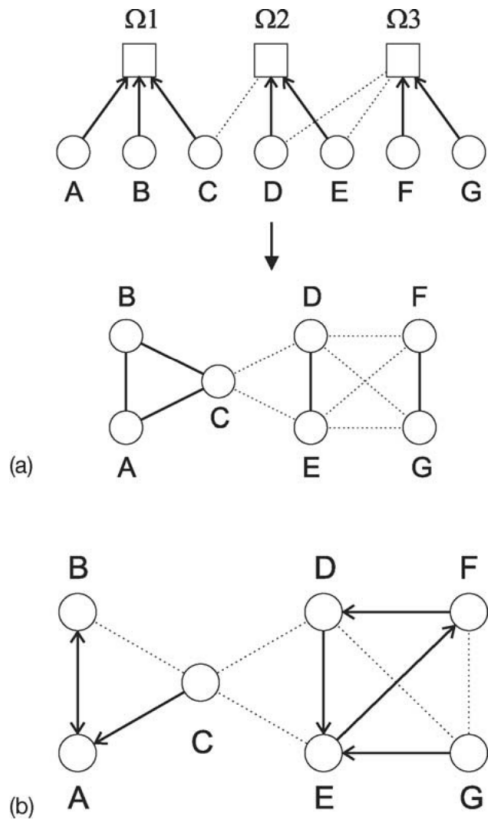


Source: Non response of C, D and E links [Kossinets, 2006]

Missing data and measurement challenges

Mechanisms

- Network boundary specification (non-inclusion of actors or affiliations)
- Survey non-response
- Censoring by vertex degree (fixed choice design)



Source: Fixed number of nominations of affilaitions (a) or acquaintances (b) [Kossinets, 2006]

Missing data and measurement challenges

- Accuracy
  - Respondents are often asked to recall their interactions with other actors
    - ★ Organisational network defined on the basis of data collected from individuals
    - ★ Ego-network and perception of the ties between the peers of the ego
    - ★ ...
  - Data triangulation (when applicable)

- Validity
  - A measure of a concept is 'valid' to the extent that it actually measures what it is intended to measure
    - ★ Collaboration $\Rightarrow$ co-authorship publications
    - ★ Friendship $\Rightarrow$ Facebook
    - ★ Citations in patent data $\Rightarrow$ knowledge flows
    - ★ ...
  - Construct validity: match between measures of concepts and theoretical predictions

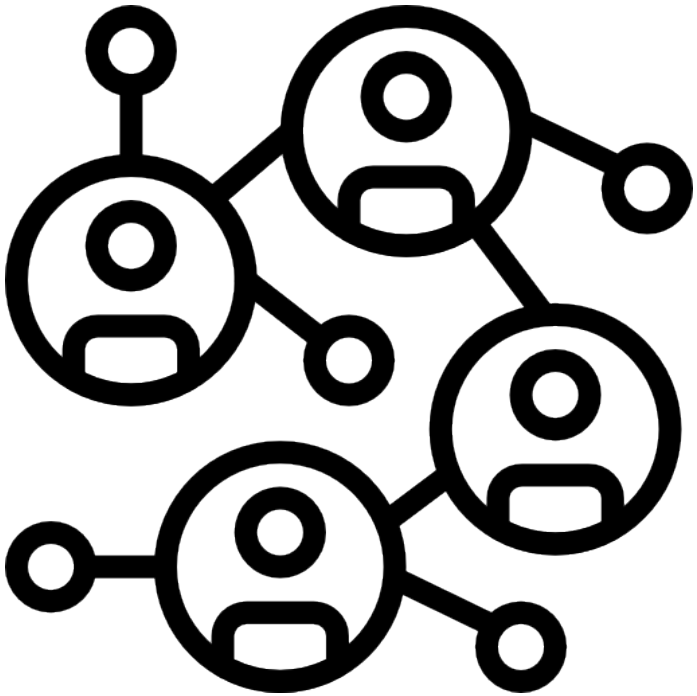
Missing data and measurement challenges

- Reliability
  - A measure of a concepts is 'reliable' if repeated measurements produce the same outcome or estimates
  - Assessing reliability is challenging in the case of social networks (interdependency between observations)
  - There is evidence that complete ranking questionnaires tend to produce reliable measures
- Error
  - The error is the difference between the 'true' value and the observed value
  - Fixed choice questionnaires tend to introduce error: it is unlikely that each actor has the same number of ties in a network

Design a network data collection

Groupwork

1. You will be allocated to a group
2. Nominate a group leader that will be reporting on the activity of the group
3. Identify a network about which you would like to collect data and a question you would like to address with these data
4. Design a data collection approach and provide a rationale for choosing the selected approach
5. Identify and discuss potential challenges



Design a network data collection

Groupwork

1. You will be allocated to a group
2. Nominate a group leader that will be reporting on the activity of the group
3. Identify a network about which you would like to collect data and a question you would like to address with these data
4. Design a data collection approach and provide a rationale for choosing the selected approach
5. Identify and discuss potential challenges

Group 1	Akanksha Belen Ross Maria Ayesha
Group 2	Charunan Oscar Hiroki Saradha
Group 3	Anas Ananya Evi Perizat Daniela
Group 4	Samuel Jongho America Noemie
Group 5	Johanna Hatty Poojani Tanya

Next time ...

Next time ...

- **Seminar: Network data collection**
  - Importa data in R
  - Network file formats
  - How to import and manipulate network data in igraph

- **Lecture: Descriptive network analysis A**
  - Network measures at the level of the whole network

Questions

References I

Creswick, N. and Westbrook, J. I. (2010).

Social network analysis of medication advice-seeking interactions among staff in an Australian hospital.
International Journal of Medical Informatics, 79(6):e116–e125.

de Nooy, W., Mrvar, A., and Batagelj, V. (2005).

Exploratory Social Network Analysis with Pajek, volume 53.
Cambridge University Press, Cambridge, UK.

Giuliani, E. (2013).

Network dynamics in regional clusters: Evidence from Chile.
Research Policy, 42(8):1406–1419.

Houston, J. F., Lee, J., and Suntheim, F. (2018).

Social networks in the global banking sector.
Journal of Accounting and Economics, 65(2-3):237–269.

Kossinets, G. (2006).

Effects of missing data in social networks.
Social Networks, 28(3):247–268.

Laumann, E. O., Marsden, P. V., and Prensky, D. (1989).

The boundary specification problem in network analysis.
In Freeman, L. C., White, D. R., and Romney, A. K., editors, *Research Methods in Social Network Analysis*, pages 61–87. George Mason University Press, Fairfax, VA.

Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-a., Comte, B., and Voirin, N. (2013).

Estimating potential infection transmission routes in hospital wards using wearable proximity sensors.
PloS one, 8(9):e73970.