Package 'sand'

March 2, 2017

Version 1.0.3

Title Statistical Analysis of Network Data with R

Author Eric D Kolaczyk, Gábor Csárdi

Maintainer Gábor Csárdi <csardi.gabor@gmail.com>

Depends igraph, igraphdata

Imports utils

Suggests GO.db, GOstats, ROCR, ape, car, eigenmodel, ergm, fdrtool, ggplot2, huge, kernlab, lattice, mixer, network, networkDynamic, networkTomography, ngspatial, org.Sc.sgd.db, sna, vioplot

Description Data sets for the book 'Statistical Analysis of Network Data with R'.

License GPL-3

URL https://github.com/kolaczyk/sand

BugReports https://github.com/kolaczyk/sand/issues

LazyData true

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2017-03-02 08:09:36

R topics documented:

`aidsblog` *AIDS blog citation network*

Description

A snapshot of the pattern of citation among 146 unique blogs related to AIDS, patients, and their support networks, collected by Suchi Gopal (see reference below) over a randomly selected three-day period in August 2005. A directed edge from one blog to another indicates that the former has a link to the latter in their web page (more specifically, the former refers to the latter in their so-called 'blogroll').

Usage

```
aidsblog
```

Format

A directed igraph graph object with 146 vertices and 187 edges.

Source

This dataset was provided to us by Suchi Gopal. Please cite the reference below if you use this dataset in your work.

References

S. Gopal, The evolving social geography of blogs. In Societies and Cities in the Age of Instant Access, ed. by H. Miller (Springer, Berlin, 2007), 139 pp. 275-294.

`calldata` *Austrian phone call network data*

Description

A set of data for phone traffic 60 between 32 telecommunication districts in Austria throughout a period during the 61 year 1991.

Usage

```
calldata
```

Format

A data frame with 32 x 31 flow measurements, 992 rows, and seven columns:

- `Orig`: factor, the origin district.
- `Dest`: factor, the destination district.
- `DistEuc`: numeric, Euclidean distance between the districts.
- `DistRd`: numeric, road distance between districts.
- `O.GRP`: numeric, gross regional product of the origin district, in Austrian schillings.
- `D.GRP`: numeric, gross regional product of the destination district, in Austrian schillings.
- `Flow`: the "amount" of phone calls from the origin district to the destination district, in erlang units (number of phone calls, including faxes, times the average length of the call divided by the duration of the measurement period).

Source

This dataset was provided to us by Suchi Gopal. Please cite the reference below if you use this dataset in your work.

References

M. Fischer, S. Gopal: Artificial neural networks: a new approach to modeling interregional telecommunication flows. J. Reg. Sci. 34(4), 503-527 (1994).

`Ecoli` *E. coli gene expression levels*

Description

Gene expression levels in the bacteria Escherichia coli (E. coli), measured for 153 genes under each of 40 different experimental conditions.

Usage

```
data(Ecoli.data)
Ecoli.expr
regDB.adj
```

Format

`Ecoli.expr` is a 40 by 153 matrix of (log) gene expression levels in the bacteria Escherichia coli (E. coli), measured for 153 transcription factors under each of 40 different experimental conditions, averaged over three replicates of each experiment. The data are a subset of those published in the reference below. The experiments were genetic perturbation experiments,in which a given gene was 'turned off', for each of 40 different genes.

`regDB.adj` is an adjacency matrix of regulatory relationships in E. coli, extracted from the RegulonDB (`http://regulondb.ccg.unam.mx/`) database at the same time the experimental data were collected.

Source

See the reference below. Please cite it if you use this dataset in your work.

References

J. Faith, B. Hayete, J. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. Collins, T. Gardner: Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 5(1), e8 (2007).

`fblog` *Network of French political blogs*

Description

Subnetwork of French political blogs, extracted from a snapshot of over 1,100 such blogs on a single day in October of 2006 and classified by the "Observatoire Presidentielle" project as to political affiliation.

Usage

```
fblog
```

Format

An undirected igraph graph with 192 vertices and 1431 edges. Note that the graph is undirected. The graph has two vertex attributes, 'name' is the URL of the blog, and 'PolParty' is the assigned political affiliation, a political party.

Source

The `mixer` R package, see also http://observatoire-presidentielle.fr/.

`g.bip` *A toy bipartite network*

Description

A toy bipartite network.

Usage

```
g.bip
```

Format

An undirected bipartite igraph graph object, with vertex attributes 'name' and 'type'.

hc *Hospital encounter network data*

Description

Records of contacts among patients and various types of health care workers in the geriatric unit of a hospital in Lyon, France, in 2010, from 1pm on Monday, December 6 to 2pm on Friday, December 10. Each of the 75 people in this study consented to wear RFID sensors on small identification badges during this period, which made it possible to record when any two of them were in face-to-face contact with each other (i.e., within 1-1.5 m of each other) during a 20-second interval of time.

Usage

```
hc
```

Format

A data frame, where each row is an interaction. It has five columns:

- `Time`: integer, time in seconds when the 20 second encounter terminated.
- `ID1`: integer, numeric ID of the first person.
- `ID2`: integer, numeric ID of the second person.
- `S1`: factor, the status of the first person, see below.
- `S2`: factor, the status of the second person, see below.

Status codes: administrative staff (ADM), medical doctor (MED), paramedical staff, such as nurses or nurses' aides (NUR), and patients (PAT).

Source

See the reference below. Please cite the it if you use this dataset in your work.

References

P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Regis, B.-a. Kim, B. Comte, N. Voirin: Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PloS One 8(9), e73970 306 (2013).

install_sand_packages *Install all packages used in the book*

Description

This function makes it easy to download and install all R packages that are used in the book 'Statistical Analysis of Network Data with R'.

Usage

```
install_sand_packages()
```

Details

The function uses the BioConductor installer, as this can install both all required BioConductor and CRAN packages.

Value

Returns the names of the installed packages, invisibly.

Author(s)

Gabor Csardi <csardi.gabor@gmail.com>

lazega *Lazega lawyers network data*

Description

This data set comes from a network study of corporate law partnership that was carried out in a Northeastern US corporate law firm, referred to as SG&R, 1988-1991 in New England. It includes (among others) measurements of networks among the 71 attorneys (partners and associates) of this firm, i.e. their strong-coworker network, advice network, friendship network, and indirect control networks. Various members' attributes are also part of the dataset, including seniority, formal status, office in which they work, gender, lawschool attended, individual performance measurements (hours worked, fees brought in), attitudes concerning various management policy options, etc.

Note that this is only a subset of the originally collected data, including the 36 partners of the firm.

Usage

```
lazega
elist.lazega
v.attr.lazega
```

Format

`lazega` is an igraph graph object, undirected. It has the following vertex attributes: 'name', 'Seniority', 'Status' (all 1, meaning partner), 'Gender' (1 is man, 2 is woman), 'Office' (1 is Boston, 2 is Hartford, 3 is Providence), 'Years' (years with the firm), 'Age', 'Practice' (1 is litigation, 2 is corporate), and 'School' (1 is Harvard or Yale, 2 is University of Connecticut, 3 is other). See the reference below for more.

`elist.lazega` is a data frame containing an edge list of the network.

`v.attr.lazega` is a data frame containing the vertex attributes only.

Source

Provided to us by Emmanuel Lazega. Please cite the reference below if you use this dataset in your work.

References

E. Lazega, The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership. Oxford University Press, Oxford (2001).

`ppi.CC` *Yeast protein interaction network*

Description

A network of 241 interactions among 134 proteins. They were assembled by Jiang et al. (see below), from various sources, and pertain to only those proteins annotated, as of January 2007, with the term "cell communication" in the gene ontology (GO) database.

Usage

```
ppi.CC
```

Format

An undirected igraph graph object, with vertex attributes:

- 'name': the name of the protein.
- 'ICSC': whether the protein is annotated with the "intracellular signaling cascade" GO term, zero or one.
- 'IPR000198': whether the protein contains the 'rho GTPase-activating protein domain' (IPR000198) motif.
- 'IPR000403': whether the protein contains the 'phosphatidylinositol 3-/4-kinase, catalytic domain' (IPR000403) motif.
- 'IPR001806': whether the protein contains the 'small GTPase superfamily' (IPR001806) motif.

- 'IPR001849': whether the protein contains the 'pleckstrin homology domain' (IPR001849) motif.
- 'IPR002041': whether the protein contains the 'ran GTPase' (IPR002041) motif.
- 'IPR003527': whether the protein contains the 'mitogen-activated protein (MAP) kinase, conserved site' (IPR003527) motif.

Source

See the reference below. Please cite it if you use this dataset in your work.

References

X. Jiang, N. Nariai, M. Steffen, S. Kasif, E. Kolaczyk: Integration of relational and hierarchical network information for protein function prediction. BMC Bioinform. 9, 350 (2008).

`sand` *The sand package*

Description

This R package accompanies the book 'Statistical Analysis of Network Data with R'. It contains some of the data sets used in the book (the others are in the `igraphdata` package). It also contains the code from the book, and some simple functions to run the code without the need for typing it in.

In brief

Type in `N<enter>` to run the next chunk of code, and `C<x>` to jump to Chapter `x`, where `x` is between 2 and 10. E.g. `C6<enter>` resets R and "loads" Chapter 6. `P<enter>` prints the next code chunk to be run (without actually running it).

The data sets

The various data sets are loaded from the code chunks in the book. The sand package contains the following data sets, each is documented in its on manual page: `Ecoli`, `aidsblog`, `calldata`, `elist.lazega`, `fblog`, `g.bip`, `hc`, `lazega`, `ppi.CC`, `sandwichprobe`, `v.attr.lazega`.

Code chunks

Code chunks of the book are numbered by chapter and each chunk is identified the chapter number and the chunk number connected by a dot.

The reader is supposed to run the code chapter by chapter, ideally, starting from a clean, new R session. This might not be critical, but it is not always possible to unload packages in R, so it is the only way to make sure that the code works correctly.

To make it easy to step through the code, the sand package define some "commands". Note that these are are not functions, and also q that they are meant to be used interactively, and not programatically.

The cursor

The cursor marks the point the reader is at in the book, and commands discussed below move the cursor and run the code the cursor is at.

The 'C' commands clear R, i.e. unload all loaded packages except for sand and its dependencies, and delete all objects from the global workspace. They also set the cursor to the first chunk of the given chapter: there are nine 'C' commands, from 'C2' to 'C10', one for each Chapter of the book. (Chapter 1 has no code to run.)

The command 'N' runs the chunk at the cursor, and steps the cursor to the next chunk. It is possible to run multiple chunks at once, with the form 'N + x' (with or without the spaces), where 'x' is the number of additional chunks to run. (I.e. 'N + 2' runs three chunks.)

The command 'P' prints the chunk at the cursor, without running it. It is possible to print other chunks as well: 'P - 1' prints the previous chunk, 'P - 2' the one before that, etc., 'P + 1' prints the next chunk, etc.

The reader is welcome to inspect R objects, or run arbitrary R code between the 'N' and 'P' commands.

Author(s)

Gabor Csardi <csardi.gabor@gmail.com>

See Also

install_sand_packages to install all R packages needed for the book.

Examples

```
## Start with Chapter 2
C2

## Run first code chunk
N

## Run next code chunk
N

## Jump to Chapter 5
C5

## Run first code chunk in Chapter 5
## It will create a plot
N
```

`sandwichprobe` *Internet packet probes data*

Description

These data correspond to an experiment conducted by Coates et al. to measure the difference in delay experienced by packet probes sent over the Internet during a short period in 2001, from a desktop computer in the ECE department at Rice University to similar machines at ten other university locations. The data were intended for use with a newly proposed method of Internet topology inference.

Usage

```
delaydata
host.locs
```

Format

The data is provided in two files. `delaydata` is a three-column data frame. The first columnf is the difference in delay of the small packets (in milliseconds). The second column is the numeric code of the destination of small packets. The third column is the numeric code of the destination of large packet.

`host.locs` contains the character code of the destinations:

1. 'IST' Instituto Superior Tecnico (Portugal)
2. 'IT' Instituto de Telecomunicacoes (Portugal)
3. 'Bkly' University of California, Berkeley
4. 'MSU1' Michigan State University (Host 1)
5. 'MSU2' Michigan State University (Host 2)
6. 'UIUC' University of Illinois, Urbana-Champaign
7. 'UWisc1' University of Wisconsin, Madison (Host 1)
8. 'UWisc2' University of Wisconsin, Madison (Host 2)
9. 'RiceU1' Rice University (Host 1)
10. 'RiceU2' Rice University (Host 2)

Source

Provided by Mark Coates, see reference below. Please cite the reference below if you use this dataset in your work.

References

M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, Y. Tsang, Maximum likelihood network topology identification from edge-based unicast measurements. Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2002, pp. 11-20.

Index