Package ‘igraphdata’

July 13, 2015

Title A Collection of Network Data Sets for the 'igraph' Package

Version 1.0.1

Author Gabor Csardi <csardi.gabor@gmail.com>

Maintainer Gabor Csardi <csardi.gabor@gmail.com>

Depends R (>= 2.10)

Suggests igraph (>= 1.0.0)

Description A small collection of various network data sets,
to use with the 'igraph' package: the Enron email network, various food webs,
interactions in the immunoglobulin protein, the karate club network,
Koenigsberg's bridges, visuotactile brain areas of the macaque monkey,
UK faculty friendship network, domestic US flights network, etc.

License CC BY-SA 4.0 + file LICENSE

URL http://igraph.org

BugReports https://github.com/igraph/igraphdata/issues

NeedsCompilation no

Repository CRAN

Date/Publication 2015-07-13 08:30:34

R topics documented:

`igraphdata-package` *The igraphdata package*

Description

The igraphdata package provides various data sets

How to use the data sets

After loading the `igraphdata` package, the various data sets can be loaded using the regular `data` command.

Type in

```
data(package="igraphdata")
```

to get a list of data sets included in this package.

Author(s)

Gabor Csardi `<csardi.gabor@gmail.com>`

References

`data`

`enron` *Enron Email Network*

Description

An Enron email dataset has been made public by the U.S. Department of Justice.

Usage

```
enron
```

Format

A directed `igraph` graph object.

Graph attributes:

- 'LDC_names' The names of the 32 LDC catagories the emails are classfied into by Michael W. Berry (http://www.cis.jhu.edu/~parky/Enron/Anno_Topic_exp_LDC.pdf)
- 'LDC_desc' Longer descriptions of the 32 LDC categories.
- 'Citation' Additionally, see also the references below.
- 'name'

Vertex attributes:

- 'Email' Email address.
- 'Name' Real name.
- 'Note' E.g. position at Enron.

Edge attributes:

- 'Time' When the email was sent. Note that some time labels are from 1979, these are certainly wrong and you might want to remove them before analyses that include time.
- 'Reciptype' Recipient type, 'to', 'cc' or 'bcc'.
- 'Topic' Assigned based on 3-means clustering of randomly selected 3,120 out of all 125,409 messages, then NN classification for the whole corpus. Note that topic 0 means an outlier, e.g., too few words or all meaningless numbers in the message body.
- 'LDC_topic' Assigned based on Michael W. Berry's 2001 "Annotated (by Topic) Enron Email Data Set." (http://www.cis.jhu.edu/~parky/Enron/Anno_Topic_exp_LDC.pdf) There are 32 topics. Topic "0" means an outlier, e.g., too few words or all meaningless numbers in the message body, etc. Topic "-1" means there is no matching topic.

Source

http://www.cis.jhu.edu/~parky/Enron/

References

C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, Scan Statistics on Enron Graphs Computational and Mathematical Organization Theory, Volume 11, Number 3, p229 - 247, October 2005, Springer Science+Business Media B.V.

C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, Scan Statistics on Enron Graphs, SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, California, April 23, 2005.

Gina Kolata, Enron Offers an Unlikely Boost to E-Mail Surveillance, New York Times, Week in Review, May 22, 2005.

C.E. Priebe, Scan Statistics on Enron Graphs, IPAM Summer Graduate School: Intelligent Extraction of Information from Graphs and High Dimensional Data, UCLA, July 11-29, 2005.

C.E. Priebe, Scan Statistics on Enron Graphs, 2005 Fall Department of Applied Mathematics and Statistics Seminars, September 15, 2005, The Johns Hopkins University.

Y. Park, C.E. Priebe, D.J. Marchette, Scan Statistics on Enron Hypergraphs, Interface 2008, Durham, North Carolina, May 21, 2008,

Y. Park, C.E. Priebe, D.J. Marchette, Anomaly Detection using Scan Statistics on Enron Graphs and Hypergraphs, The Satellite Workshop of the IASC 2008 Conference, Seoul, Korea, December 1-3, 2008.

Y. Park, C.E. Priebe, D.J. Marchette, A. Youssef, Anomaly Detection using Scan Statistics on Time Series of Hypergraphs, Workshop on Link Analysis, Counterterrorism and Security at the SIAM International Conference on Data Mining, Sparks, Nevada, May 1-3, 2009,

Y. Park, C.E. Priebe, A. Youssef, Anomaly Detection in Time Series of Graphs using Fusion of Invariants, Computational and Mathematical Organization Theory, submitted, 2010.

`foodwebs` *A collection of food webs*

Description

A list of graphs. Each one is a food web, i.e. a directed graph of predator-prey relationships.

Usage

```
foodwebs
```

Format

A named list of directed `igraph` graph objects. Here are the list of the graphs included:

'ChesLower' Lower Chesapeake Bay in Summer.

Reference: Hagy, J.D. (2002) Eutrophication, hypoxia and trophic transfer efficiency in Chesapeake Bay PhD Dissertation, University of Maryland at College Park (USA), 446 pp.

'ChesMiddle' Middle Chesapeake Bay in Summer.

Reference: same as for 'ChesLower'.

'ChesUpper' Upper Chesapeake Bay in Summer.

Reference: same as for 'ChesLower'.

'Chesapeake' Chesapeake Bay Mesohaline Network.

Reference: Baird D. & Ulanowicz R.E. (1989) The seasonal dynamics of the Chesapeake Bay ecosystem. Ecological Monographs 59:329-364.

'CrystalC' Crystal River Creek (Control).

Reference: Homer, M. and W.M. Kemp. Unpublished Ms. See also Ulanowicz, R.E. 1986. Growth and Development: Ecosystems Phenomenology. Springer, New York. pp 69-79.

'CrystalD' Crystal River Creek (Delta Temp).

Reference: same as for 'CrystalD'.

'Maspalomas' Charca de Maspalomas.

Reference: Almunia, J., G. Basterretxea, J. Aristegui, and R.E. Ulanowicz. (1999) Benthic-Pelagic switching in a coastal subtropical lagoon. Estuarine, Coastal and Shelf Science 49:363-384.

'Michigan' Lake Michigan Control network.

Reference: Krause, A. and D. Mason. (In preparation.) A. Krause, PhD. Dissertation, Michigan State University. Ann Arbor, MI. USA.

'Mondego' Mondego Estuary - Zostrea site.

Reference: Patricio, J. (In Preparation) Master's Thesis. University of Coimbra, Coimbra, Portugal.

'Narragan' Narragansett Bay Model.

Reference: Monaco, M.E. and R.E. Ulanowicz. (1997) Comparative ecosystem trophic structure of three U.S. Mid-Atlantic estuaries. Mar. Ecol. Prog. Ser. 161:239-254.

'Rhode' Rhode River Watershed - Water Budget.

Reference: Correll, D. (Unpublished manuscript) Smithsonian Institute, Chesapeake Bay Center for Environmental Research, Edgewater, Maryland 21037-0028 USA.

'StMarks' St. Marks River (Florida) Flow network.

Reference: Baird, D., J. Luczkovich and R. R. Christian. (1998) Assessment of spatial and temporal variability in ecosystem attributes of the St Marks National Wildlife Refuge, Apalachee Bay, Florida. Estuarine, Coastal, and Shelf Science 47: 329-349.

'baydry' Florida Bay Trophic Exchange Matrix, dry season.

Reference: Ulanowicz, R. E., C. Bondavalli, and M. S. Egnotovich. 1998. Network analysis of trophic dynamics in South Florida ecosystems, FY 97: the Florida Bay ecosystem. Annual Report to the United States Geological Service Biological Resources Division, University of Miami Coral Gables, [UMCES] CBL 98-123, Maryland System Center for Environmental Science, Chesapeake Biological Laboratory, Maryland, USA.

'baywet' Florida Bay Trophic Exchange Matrix, wet season.

Reference: same as for 'baydry'.

'cypdry' Cypress, dry season.

Reference: Ulanowicz, R. E., C. Bondavalli, and M. S. Egnotovich. 1997. Network analysis of trophic dynamics in South Florida ecosystems, FY 96: the cypress wetland ecosystem. Annual Report to the United States Geological Service Biological Resources Division, University of Miami Coral Gables, [UM-CES] CBL 97-075, Maryland System Center for Environmental Science, Chesapeake Biological Laboratory.

'cypwet' Cypress, wet season.

Reference: same as for 'cypdry'.

'gramdry' Everglades Graminoids - Dry Season.

Reference: Ulanowicz, R. E., J. J. Heymans, and M. S. Egnotovich. 2000. Network analysis of trophic dynamics in South Florida ecosystems, FY 99: the graminoid ecosystem. Technical Report TS-191-99, Maryland System Center for Environmental Science, Chesapeake Biological Laboratory, Maryland, USA.

'gramwet' Everglades Graminoids - Wet Season.

Reference: same as for 'gramdry'.

'mangdry' Mangrove Estuary, Dry Season.

Reference: Ulanowicz, R. E., C. Bondavalli, J. J. Heymans, and M. S. Egnotovich. 1999. Network analysis of trophic dynamics in South Florida ecosystems, FY 98: the mangrove ecosystem. Technical Report TS-191-99, Maryland System Center for Environmental Science, Chesapeake Biological Laboratory, Maryland, USA.

'mangwet' Mangrove Estuary, Wet Season.

Reference: same as for 'mangdry'.

Each graph has the following vertex attributes: 'name' is the name of the species, 'ECO' is the type of the node, and integer value between one and five, meaning:

1. Living/producing compartment
2. Other compartment
3. Input
4. Output
5. Respiration.

The 'Biomass' vertex attribute contains the biomass of the species.

Edges are weighted, and the weights denote energy flux between the species involved.

The graphs also contain some informative graph attributes: 'Author', 'Citation', 'URL', and 'name'.

Source

See references for the individual webs above. The data itself was downloaded from `http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm`.

References

See them above.

`immuno` *Immunoglobulin interaction network*

Description

The undirected and connected network of interactions in the immunoglobulin protein. It is made up of 1316 vertices representing amino-acids and an edge is drawn between two amino-acids if the shortest distance between their C_alpha atoms is smaller than the threshold value $\theta = 8$ Angstrom.

Usage

```
immuno
```

Format

An undirected `igraph` graph object.

Graph attributes: 'name', 'Citation', 'Author'.

Source

See reference below.

References

D. Gfeller, Simplifying complex networks: from a clustering to a coarse graining strategy, *PhD Thesis EPFL*, no 3888, 2007. http://library.epfl.ch/theses/?nr=3888

---

karate *Zachary's karate club network*

Description

Social network between members of a university karate club, led by president John A. and karate instructor Mr. Hi (pseudonyms).

The edge weights are the number of common activities the club members took part of. These activities were:

1. Association in and between academic classes at the university.
2. Membership in Mr. Hi's private karate studio on the east side of the city where Mr. Hi taught nights as a part-time instructor.
3. Membership in Mr. Hi's private karate studio on the east side of the city, where many of his supporters worked out on weekends.
4. Student teaching at the east-side karate studio referred to in (2). This is different from (2) in that student teachers interacted with each other, but were prohibited from interacting with their students.
5. Interaction at the university rathskeller, located in the same basement as the karate club's workout area.
6. Interaction at a student-oriented bar located across the street from the university campus.
7. Attendance at open karate tournaments held through the area at private karate studios.
8. Attendance at intercollegiate karate tournaments held at local universities. Since both open and intercollegiate tournaments were held on Saturdays, attendance at both was impossible.

Zachary studied conflict and fission in this network, as the karate club was split into two separate clubs, after long disputes between two factions of the club, one led by John A., the other by Mr. Hi.

The 'Faction' vertex attribute gives the faction memberships of the actors. After the split of the club, club members chose their new clubs based on their factions, except actor no. 9, who was in John A.'s faction but chose Mr. Hi's club.

Usage

```
karate
```

Format

An undirected `igraph` graph object. Vertex no. 1 is Mr. Hi, vertex no. 34 corresponds to John A.

Graph attributes: 'name', 'Citation', 'Author'.

Vertex attributes: 'name', 'Faction', 'color' is the same as 'Faction', 'label' are short labels for plotting.

Edge attribute: 'weight'.

Source

See reference below.

References

Wayne W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research* Vol. 33, No. 4 452-473

`kite` *Krackhardt's kite*

Description

Krackhardt's kite is a fictionary social network with ten actors. It is the smallest graph for which the most central actor is different according to the three classic centality measures: degree, closeness and betweenness.

Usage

```
kite
```

Format

An undirected igraph graph with graph attributes `name`, `layout`, `Citation`, `Author`, `URL`, and vertex attributes `label`, `name` and `Firstname`.

Source

http://www.orgnet.com/sna.html

References

Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. David Krackhardt. *Admin. Sci. Quart.* 35, 342-369, 1990.

`Koenigsberg` *Bridges of Koenigsberg from Euler's times*

Description

The Seven Bridges of Koenigsberg is a notable historical problem in mathematics. Its negative resolution by Leonhard Euler in 1735 laid the foundations of graph theory and presaged the idea of topology.

The city of Koenigsberg in Prussia (now Kaliningrad, Russia) was set on both sides of the Pregel River, and included two large islands which were connected to each other and the mainland by seven bridges

The problem was to find a walk through the city that would cross each bridge once and only once. The islands could not be reached by any route other than the bridges, and every bridge must have been crossed completely every time (one could not walk half way onto the bridge and then turn around and later cross the other half from the other side).

Euler proved that the problem has no solution.

Usage

```
Koenigsberg
```

Format

An undirected `igraph` graph object with vertex attributes 'name' and 'Euler_letter', the latter is the notation from Eulers original paper; and edge attributes `name` (the name of the bridge) and 'Euler_letter', again, Euler's notation from his paper.

This dataset is in the public domain.

Source

Wikipedia, `http://de.wikipedia.org/wiki/K%C3%B6nigsberger_Br%C3%BCcken`

`macaque` *Visuotactile brain areas and connections*

Description

Graph model of the visuotactile brain areas and connections of the macaque monkey. The model consists of 45 areas and 463 directed connections.

Usage

```
macaque
```

Format

A directed `igraph` graph object with vertex attributes 'name' and 'shape'.

This dataset is licensed under a Creative Commons Attribution-Share Alike 2.0 UK: England & Wales License, see `http://creativecommons.org/licenses/by-sa/2.0/uk/` for details. Please cite the reference below if you use this dataset.

Source

See reference below.

References

Negyessy L., Nepusz T., Kocsis L., Bazso F.: Prediction of the main cortical areas and connections involved in the tactile function of the visual cortex by network analysis. *European Journal of Neuroscience*, 23(7): 1919-1930, 2006.

`rfid` *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

```
rfid
```

Format

An igraph graph with graph attributes 'name' and 'Citation', vertex attribute 'Status' and edge attribute 'Time'.

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

'Time' is the time of the encounter, it is the second when the 20 second encounter terminated.

Source

See the reference below. Please cite 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).

`UKfaculty` *Friendship network of a UK university faculty*

Description

The personal friendship network of a faculty of a UK university, consisting of 81 vertices (individuals) and 817 directed and weighted connections. The school affiliation of each individual is stored as a vertex attribute. This dataset can serve as a testbed for community detection algorithms.

Usage

```
UKfaculty
```

Format

A directed `igraph` graph object with vertex attribute 'Group', the numeric id of the school affiliation, and edge attribute 'weight', i.e. the graph is weighted.

Source

See reference below.

References

Nepusz T., Petroczi A., Negyessy L., Bazso F.: Fuzzy communities and the concept of bridgeness in complex networks. Physical Review E 77:016107, 2008.

`USairports` *US airport network, 2010 December*

Description

The network of passanger flights between airports in the United States. The data set was compiled based on flights in 2010 December. This network is directed and edge directions correspond to flight directions. Each edge is specific to a single carrier aircraft type. Multiple carriers between the same two airports are denoted by multiple edges.

See information about the included meta-data below.

Usage

```
USairports
```

Format

A directed `igraph` graph object, with multiple edges. It has a 'name' graph attribute, and several vertex and edge attributes. The vertex attributes:

name Symbolic vertex name, this is the three letter IATA airport code.

City City and state, where the airport is located.

Position Position of the airport, in WGS coordinates.

Edge attributes:

Carrier Name of the airline. The network includes both domestic and international carriers that performed at least one flight in December of 2010.

Departures The number of departures (for a given airline and aircraft type.

Seats The total number of seats available on the flights carried out by a given airline, using a given aircraft type.

Passengers The total number of passangers on the flights carried out by a given airline, using a given aircraft type.

Aircraft Type of the aircraft.

Distance The distance between the two airports, in miles.

Source

Most of this information was downloaded from The Research and Innovative Technology Administration (RITA). See `http://www.rita.dot.gov/about_rita/` for details. The airport position information was collected from Wikipedia and other public online sources.

`yeast` *Yeast protein interaction network*

Description

Comprehensive protein-protein interaction maps promise to reveal many aspects of the complex regulatory network underlying cellular function.

This data set was compiled by von Mering et al. (see reference below), combining various sources. Only the interactions that have 'high' and 'medium' confidence are included here.

Usage

```
yeast
```

Format

An undirected `igraph` graph object. Its graph attributes: 'name', 'Citation', 'Author', 'URL'. 'Classes'. The 'Classes' attribute contain the key for the classification labels of the proteins, in a data frame, the original MIPS categories are given after the semicolon:

E energy production; energy

G aminoacid metabolism; aminoacid metabolism

M other metabolism; all remaining metabolism categories

P translation; protein synthesis

T transcription; transcription, but without subcategory 'transcriptional control'

B transcriptional control; subcategory 'transcriptional control'

F protein fate; protein fate (folding, modification, destination)

O cellular organization; cellular transport and transport mechanisms

A transport and sensing; categories 'transport facilitation' and 'regulation of / interaction with cellular environment'

R stress and defense; cell rescue, defense and virulence

D genome maintenance; DNA processing and cell cycle

C cellular fate / organization; categories 'cell fate' and 'cellular communication / signal transduction' and 'control of cellular organization'

U uncharacterized; categories 'not yet clear-cut' and 'uncharacterized'

Vertex attributes: 'name', 'Description', 'Class', the last one contains the class of the protein, accoring to the classification above.

Note that some proteins in the network did not appear in the annotation files, the 'Class' and 'Description' attributes are `NA` for these.

Source

The data was downloaded from http://www.nature.com/nature/journal/v417/n6887/suppinfo/nature750.html.

References

Comparative assessment of large-scale data sets of protein-protein interactions. Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields and Peer Bork. *Nature* 417, 399-403 (2002)

Index