Package ‘SNscan’

January 19, 2016

Type Package
Title Scan Statistics in Social Networks
Version 1.0
Date 2016-1-9
Author Tai-Chi Wang [aut, cre, ths] <taichi53719@gmail.com>,
Tun-Chieh Hsu [aut] <htj801116@gmail.com>,
Frederick Kin Hing Phoa [ths] <fredphoa@stat.sinica.edu.tw>
Maintainer Tai-Chi Wang <taichi53719@gmail.com>
Description Scan statistics applied in social network data can be used to test the cluster characteristics among a social network.
License GPL-2
Depends R (>= 3.1.0)
Imports igraph, poweRlaw, Rmpfr
NeedsCompilation no
Repository CRAN
Date/Publication 2016-01-19 13:09:28

R topics documented:

SNscan-package .................................................... 2
binom.stat .......................................................... 3
book ................................................................. 4
coauthor ............................................................. 4
conpowerlaw.stat .................................................. 5
graph.rmedge ..................................................... 6
group.graph ....................................................... 7
karate ............................................................... 8
mcpv.function ..................................................... 8
multinom.stat ..................................................... 9
network.mc.scan ................................................ 10
network.scan ...................................................... 11
norm.stat .......................................................... 12
Description

This package is constructed for testing the clustering patterns of structure and attribute among a social network through the scan statistics. Most contents are related to the references mentioned below. Some data sets are presented to be examples in this package as well.

Details

Package: SNscan
Type: Package
Version: 1.0
Date: 2016-1-9

Author(s)

Tai-Chi Wang [aut,cre,ths] <taichi53719@gmail.com>, Tun-Chieh Hsu [aut] <htj801116@gmail.com>, Frederick Kin Hing Phoa [ths] <fredphoa@stat.sinica.edu.tw>

References


See Also

igraph
**binom.stat**  
*Binomial Scan Statistic*

**Description**

The binomial scan statistic evaluate the statistic which compares the node attribute within the subgraph with that outside the subgraph while the node attribute follows the binomial distribution.

**Usage**

\[
\text{binom.stat}(\text{obs}, \text{pop}, \text{zloc})
\]

**Arguments**

- `obs`: Numeric vector of observation values.
- `pop`: Numeric vector of population values.
- `zloc`: Numeric vector of selected nodes.

**Details**

A network with interested attributes is denoted as \( G = (V, E, X) \), where \( X = (x_1, \ldots, x_{|V|}) \) follows a defined distribution. Suppose a subgraph, \( Z \), is selected.

\[
\lambda_A(Z) = n_z \ln \left( \frac{p_{11}}{p_0} \right) + (N_z - n_z) \ln \left( \frac{1 - p_{11}}{1 - p_0} \right) + (n_G - n_z) \ln \left( \frac{p_{10}}{p_0} \right) + [(N_G - N_z) - (n_G - n_z)] \ln \left( \frac{1 - p_{10}}{1 - p_0} \right),
\]

where \( p_0 = n_G/N_G \), \( p_{10} = (n_G - n_z)/(N_G - N_z) \), and \( p_{11} = n_z/N_z \). In addition, \( N_G \) and \( n_G \) are population sizes and number of successful cases of whole graph \( G \). \( N_Z \) and \( n_Z \) are expressed in the same way in the selected subgraph \( Z \).

**Value**

Three values will be returned. The first value is test statistic. The second is the estimated means which estimated outside the selected nodes. The third is estimated means estimated within the selected nodes.

**Author(s)**

Taichi Wang <taichi43@stat.sinica.edu.tw>

**References**


**Examples**

\[
\text{binom.stat}(\text{obs=rbinom(n=100, size=10000, prob=0.0001)}, \text{pop=rep(10000,100)}, \text{zloc=1:5})
\]
**book**  
*Books about US politics*

**Description**
A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent copurchasing of books by the same buyers.

**Usage**
data(book)

**Format**
The data are in the format of igraph object. IGRAPH U—105 441 – + attr: layout (g/n), randomout (g/n), id (v/n), label (v/c), value (v/c)

**Source**

**Examples**
data(book)

---

**coauthor**  
*Co-Authorship Data*

**Description**
This co-authorship network data are extracted from geom.zip (see the source link) with 6158 vertices and 22577 edges. In this network, two authors were connected if they published at least one paper together and the node attribute represents the number of papers of each author.

**Usage**
data(coauthor)

**Format**
The data are in the format of igraph object. IGRAPH U—6158 22577 – + attr: count (v/n)

**Source**
http://vlado.fmf.uni-lj.si/pub/networks/data/collab/Geom.zip
Examples

data(coauthor)

conpowerlaw.stat  Continuous Power-Law Scan Statistic

Description

The continuous power law scan statistic evaluate the statistic which compare the node attribute within the subgraph with that outside the subgraph while the node attribute follows the continuous power-law distribution.

Usage

conpowerlaw.stat(obs, pop = 1, zloc, xmin = 1, zetatable = NULL)

Arguments

- obs: Numeric vector of observation values.
- pop: Numeric vector of population values; default = 1.
- zloc: Numeric vector of selected nodes.
- xmin: The minmum value of power law distribution; defaut=1.
- zetatable: When xmin=1, set up the data(zetatable).

Value

Three values will be returned. The first value is test statistic. The second is the estimated means which estimated outside the selected nodes. The third is the estimated means estimated within the selected nodes.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

References


See Also

rplcon,zetatable,powerlaw.stat

Examples

library(powerlaw)
x=rplcon(n=100, xmin=1, alpha=3)#function from powerlaw
conpowerlaw.stat(obs=x, zloc=1:5)
**graph.rmedge**

**Random Graph with Expected Edges**

**Description**

Generate random graph with expected edges.

**Usage**

```r
graph.rmedge(n, g, fix.edge = TRUE)
```

**Arguments**

- `n`: Integer value, the number of random graph.
- `g`: An igraph object, a baseline graph for generating random graph.
- `fix.edge`: When `fix.edge` = TRUE, the number of edges is fixed and equal to the number of edges in the original graph `g`. When `fix.edge` = FALSE, the number of edges is randomly generated according to the connection probability of the original graph `g`.

**Value**

A data list in which each component is an igraph object.

**Author(s)**

Taichi Wang <taichi43@stat.sinica.edu.tw>

**See Also**

- `erdos.renyi.game`

**Examples**

```r
library(igraph)
g = graph.ring(10)
graph.rmedge(n=1,g=g,fix.edge = TRUE)
graph.rmedge(n=1,g=g,fix.edge = FALSE)
```
**group.graph**

*Generate igraph Objects with Different Connection Characteristics.*

**Description**

Generate two groups of igraph object with different connection probabilities.

**Usage**

```r
group.graph(V, cv = NULL, p1, p2 = NULL)
```

**Arguments**

- `V`: The number of vertices.
- `cv`: The assigned nodes with connection probability `p2`.
- `p1`: The connection probability for group 1.
- `p2`: The connection probability for group 2.

**Details**

If there is only one group, it is equivalent to the Erdos and Renyi model.

**Value**

An igraph object.

**Author(s)**

Taichi Wang <taichi43@stat.sinica.edu.tw>

**See Also**

- `erdos.renyi.game`

**Examples**

```r
group.graph(V=10, cv =1:3 , p1=1/10, p2 = 1/2)
```
karate  

Karate Club Data

Description
Social network of friendships between 34 members of a karate club at a US university in the 1970s

Usage
data(karate)

Format
The data are in the format of igraph object. IGRAPH U— 34 78 – + attr: layout (g/n)

Source
http://www-personal.umich.edu/~mejn/netdata/

References

Examples
data(karate)

mcpv.function  

Monte Carlo p-value

Description
Compute the Monte Carlo p-value.

Usage
mcpv.function(obs.stat, ms.stat, direction)

Arguments
obs.stat  The observed testing statistics.
ms.stat  The Monte Carlo testing statistics.
direction  The comparative direction of observed testing statistics and the Monte Carlo testing statistics.
Value
The p-values of observations will be returned.

Author(s)
Taichi Wang <taichi43@stat.sinica.edu.tw>

See Also
network.scan

Examples
# Please refer to the page of network.scan.

```
#m <- matrix(c(0,1,1,0,0), nrow=2, byrow=TRUE)
#d <- network.scan(m)
```

Description
The multinomial scan statistic evaluates the statistic which compares the node attribute within the subgraph with that outside the subgraph while the node attribute follows the multinomial distribution.

Usage
```
multinom.stat(obs, pop = 1, zloc)
```

Arguments
- **obs**: Numeric vector of observation values.
- **pop**: Numeric vector of population values (default is 1).
- **zloc**: Numeric vector of selected nodes.

Details
A network with interested attributes is denoted as \( G = (V, E, X) \), where \( X = (x_1, \ldots, x_{|V|}) \) follows a defined distribution. Suppose a subgraph, \( Z \), is selected. Suppose there are \( k \) categories in an interested data, the multinomial test statistic is expressed as

\[
\lambda_A(Z) = \sum_k (n_{zk} \ln \left( \frac{n_{zk}}{n_z} \right) + (n_k - n_{zk}) \ln \left( \frac{n_k - n_{zk}}{n - n_z} \right) - n_k \ln \left( \frac{n_k}{n} \right)),
\]

where \( n \) is the total number of observations (nodes), \( n_z \) is total number of observations in \( z \), \( n_{zk} \) is total number of \( k \) category in \( z \), and \( n_k \) is total number of \( k \) category in all data.
Value

Three values will be returned. The first value is test statistic. The second is the estimated means which estimated outside the selected nodes. The third is the estimated means estimated within the selected nodes.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

References


Examples

```r
multinom.stat(obs=rep(1:5,each=10),zloc=1:5)
```

**network.mc.scan**  
*Monte Carlo scan statistic in social network*

Description

Evaluate scan statistics based on Monte Carlo data.

Usage

```r
network.mc.scan(n, g, radius, attribute, model, pattern,  
fix.edge= FALSE, max.prop = 0.5, xmin = NULL, zetatable =NULL)
```

Arguments

- `n`: The size of the generated Monte Carlo data.
- `g`: An igraph object.
- `radius`: The radius of scanning windows. Default is 3.
- `attribute`: The interested attribute which should be a data list including observations (obs) and population (pop).
- `model`: The distribution of attribute which can be "norm.stat", "pois.stat", "binom.stat", and "multinom.stat".
- `pattern`: The testing pattern of the network which can be "structure", "attribute", and "both".
- `fix.edge`: Logical term: TRUE for generating fixed number of edges; FALSE for random number of edges. Default is FALSE.
- `max.prop`: Numeric value, the maximum proportion of selecting graph. Default is 0.5.
- `xmin`: Numeric value, the minimum value only for powerlaw stat. Default is 1.
- `zetatable`: Zetatable is applied when power-law distribution is used. Default is NULL.
network.scan

Details

All arguments should be exactly the same as the set applied in network.scan.

Value

A matrix will be returned. Each meaning of the row is equal to to network.scan. The test statistic in network.mc.scan is maximum in each Monte Carlo sample.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

See Also

network.scan; graph.rmedge

Examples

#Please refer to the page of network.scan.

<table>
<thead>
<tr>
<th>network.scan</th>
<th>Network Scan Statistic</th>
</tr>
</thead>
</table>

Description

Evaluate scan statistics in social network.

Usage

network.scan(g, radius, attribute, model, pattern, max.prop = 0.5, xmin = NULL, zetatable = NULL)

Arguments

g An igraph object.

radius The radius of scanning windows. Default is 3.

attribute The interested attribute which should be a data list including observations (obs) and populations (pop).

model The distribution of attribute which can be "norm.stat", "pois.stat", "binom.stat", "multinom.stat", and "powerlaw.stat".

pattern The testing pattern of the network which can be "structure", "attribute", and "both".

max.prop Numeric value, the maximum proportion of selecting graph. Default is 0.5.

xmin Numeric value, the minimum value only for powerlaw stat. Default is 1.

zetatable Zatable is applied when power-law distribution is used. Default is NULL.
Value

A matrix will be returned. The values include C: the center of a scanning window, D: The radius of a scanning window. test.L: the test statistic, S0: Indicating the testing information within the selected nodes. Sz: Indicating the testing information outside the nodes. The final column, z.length, is the number of the nodes in the cluster with corresponding center and radius.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

See Also

network.mc.scan

Examples

data(karate)

ks=network.scan(g=karate, radius=3, attribute=NULL, model="pois.stat", pattern="structure")
mc.ks=network.mc.scan(n=9, g=karate, radius=3, attribute=NULL, model="pois.stat", pattern="structure")
pv=mcpv.function(obs.stat=ks[,]3, ms.stat=mc.ks[,]3, direction">=")

-------

norm.stat

Normal Scan Statistic

Description

The normal scan statistic evaluates the statistic which compares the node attribute within the subgraph with that outside the subgraph while the node attribute follows the normal distribution.

Usage

norm.stat(obs, pop = 1, zloc)

Arguments

obs Numeric vector of observation values.
pop Numeric vector of population values; default = 1.
zloc Numeric vector of selected nodes.
Details

A network with interested attributes is denoted as \( G = (V, E, X) \), where \( X = (x_1, \ldots, x_{|V|}) \) follows a defined distribution. Suppose a subgraph, \( Z \), is selected.

\[
\lambda_A(Z) = n \ln\left(\sqrt{\hat{\sigma}^2}\right) - n \ln\left(\sqrt{2\hat{\sigma}_Z^2}\right),
\]

where \( \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \), and \( \hat{\sigma}_Z^2 = \frac{[\sum_{i \in Z} (x_i - \bar{x}_z)^2 - \sum_{j \notin Z} (x_j - \bar{x}_c)^2]}{n} \), in which \( n \) is the number of nodes, and \( \bar{x}_z = \frac{\sum_{i \in Z} x_i}{n} \) and \( \bar{x}_c = \frac{\sum_{j \notin Z} x_j}{(n - n_z)} \). It is equivalent to minimize the variance within the subgraph \( Z \).

Value

Three values will be returned. The first value is test statistic. The second is the estimated means which estimated outside the selected nodes. The third is the estimated means estimated within the selected nodes.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

References


Examples

```r
norm.stat(obs=rnorm(100,10,1),zloc=1:5)
```

---

**pois.stat**

*The Poisson Scan Statistic*

**Description**

The Poisson scan statistic evaluates the statistic which compares the node attribute within the subgraph with that outside the subgraph while the node attribute follows the Poisson distribution.

**Usage**

`pois.stat(obs, pop, zloc)`

**Arguments**

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>obs</code></td>
<td>Numeric vector of observation values.</td>
</tr>
<tr>
<td><code>pop</code></td>
<td>Numeric vector of population values.</td>
</tr>
<tr>
<td><code>zloc</code></td>
<td>Numeric vector of selected nodes.</td>
</tr>
</tbody>
</table>
Details

A network with interested attributes is denoted as $G = (V, E, X)$, where $X = (x_1, \ldots, x_{|V|})$ follows a defined distribution. Suppose a subgraph, $Z$, is selected.

$$\lambda_A(Z) = n_Z \ln \left( \frac{p_{11}}{p_0} \right) + (n_G - n_z) \ln \left( \frac{p_{10}}{p_0} \right),$$

where the estimated parameters are equivalent to those in the binomial distribution.

Value

Three values will be returned. The first value is test statistic. The second is the estimated means which estimated outside the selected nodes. The third is estimated means estimated within the selected nodes.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

References


See Also

*binom.stat*

Examples

```r
pois.stat(obs=rpois(n=100, lambda=10), pop=rep(1,100), zloc=1:5)
```

---

**powerlaw.stat**  
*Discrete Power-Law Scan Statistic*

**Description**

The discrete power-law scan statistic evaluates the statistic which compares the node attribute within the subgraph with that outside the subgraph while the node attribute follows the power-law distribution.

**Usage**

```r
powerlaw.stat(obs, pop = 1, zloc, xmin = 1, zetatable = NULL)
```
Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>obs</td>
<td>Numeric vector of observation values.</td>
</tr>
<tr>
<td>pop</td>
<td>Numeric vector of population values; default = 1.</td>
</tr>
<tr>
<td>zloc</td>
<td>Numeric vector of selected nodes.</td>
</tr>
<tr>
<td>xmin</td>
<td>The minimum value of power law distribution; default = 1.</td>
</tr>
<tr>
<td>zetatable</td>
<td>When xmin=1, set up the data(zetatable).</td>
</tr>
</tbody>
</table>

Value

Three values will be returned. The first value is test statistic. The second is the estimated means which estimated outside the selected nodes. The third is estimated means which estimated within the selected nodes.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

References


See Also

rplcon, zetatable, conpowerlaw.stat

Examples

```r
library(powerlaw)
data(zetatable)
x=rpldis(n=100, xmin=1, alpha=1.5) #function from powerlaw
powerlaw.stat(\text{obs}=x, zloc=1:5, \text{zetatable} = \text{zetatable})
```

---

rmulti.one

Generate Random \((0, 1)\) Multinomial Data

Description

Generate random multinomial data in which all value no more than 1.

Usage

```
rmulti.one(size, p)
```
The Structure Scan Statistic

Description

The structure scan statistic evaluate the statistic which compare the log likelihood within the subgraph with that outside the subgraph when the distribution of the graph follows Poisson random graph.

Usage

structure.stat(g, subnodes)

Arguments

g The original graph.
subnodes Numeric vector, the vertices of the original graph which will separate the original graph into two partitions.
Details

Suppose subnodes are selected and form a subgraph $Z$. The likelihood ratio statistic of a selected subgraph $Z$ is

$$\lambda_S(Z) = \ln LR(Z) = \ln \frac{L_Z}{L_0} = |E_Z| \ln \left( \frac{|E_Z|}{\mu(Z)} \right) + (|E_G| - |E_Z|) \ln \left( \frac{|E_G| - |E_Z|}{\mu(G) - \mu(Z)} \right)$$

when $\hat{\alpha} > \hat{\beta}$, where $\hat{\alpha} = \frac{|E_Z|}{\mu(Z)}$ and $\hat{\beta} = \frac{|E_G| - |E_Z|}{\mu(G) - \mu(Z)}$, in which

$$\mu(G) = \frac{k_G^2}{4|E_G|}, \mu(Z) = \frac{k_Z^2}{4|E_G|}, \text{and } \mu(Z^C) = \mu(G) - \mu(Z)$$

with $k_G = \sum_{i \in G} k_i$, the total sum of degrees. When $\hat{\alpha} \leq \hat{\beta}$, $\lambda_S(Z) = 0$.

Value

Three values will be returned. The first value is the test statistic. The rest of are estimated ratios which equal to observed edges divided by expected edges while the former was estimated outside and the later was estimated within the selected subgraph.

Author(s)

Taichi Wang <taichi43@stat.sinica.edu.tw>

References


See Also

igraph

Examples

library(igraph)
g <- graph.ring(10)
structure.stat(g, subnodes=c(1:3))
stubs.sampling  Generate Random Graph

Description
Generate random graph by randomly permuting the stubs of nodes

Usage
stubs.sampling(s, g)

Arguments
s  Number of samplings.
g  An igraph object, a baseline graph for generating random graph.

Value
A data matrix in which each row is the edge expression of graph.

Author(s)
Taichi Wang <taichi43@stat.sinica.edu.tw>

See Also
c

Examples
library(igraph)
par(mfrow=c(2,1))
g <- graph.ring(10); plot(g)
Sg=stubs.sampling(s=10, g=g)
sg=graph(Sg[1,],n=V(g),directed=FALSE);plot(sg)

usag  USA Poverty Data

Description
The data were estimated by U.S. Census Bureau, 2007–2011 American Community Survey, and were reported in [http://www.census.gov/hhes/www/poverty/](http://www.census.gov/hhes/www/poverty/).

Usage
data(usag)
**zetatable**

**Format**

The data are in the format of igraph object. IGRAPH UN- 49 109 – + attr: name (v/c), white.pop (v/n), black.pop (v/n), white.poverty (v/n), black.poverty (v/n), coor.x (v/n), coor.y (v/n)

**Details**

The vertex attributes including name: US state names; white.pop: white people population size for each state; black.pop: black people population size for each state; white.poverty: number of white people under the poverty line for each state; black.poverty: number of black people under the poverty line for each state; coor.x: longitudes of state centres; coor.y: latitudes of state centres.

**Source**


**Examples**

data(usag)

---

**zetatable**

**Zeta Function Table**

---

**Description**

Each row of zeta function table equals to the derivation of Riemann zeta function divided by zeta function.

**Usage**

data(zetatable)

**Format**

A data frame with 99999 observations on the following 2 variables; The first column is the given value of alpha in zeta function, and the second is the value of the derivation of Riemann zeta function divided by zeta function.

**Examples**

data(zetatable)
str(zetatable)
Index

*Topic datasets
  book, 4
  coauthor, 4
  karate, 8
  usag, 18
  zetatable, 19

*Topic graph sampling
  graph.rmedge, 6
  group.graph, 7
  stubs.sampling, 18

*Topic package
  SNscan-package, 2

*Topic scanning method
  network.mc.scan, 10
  network.scan, 11

*Topic statistics
  binom.stat, 3
  conpowerlaw.stat, 5
  multinom.stat, 9
  norm.stat, 12
  pois.stat, 13
  powerlaw.stat, 14
  structure.stat, 16
  snscan
    SNscan-package, 2
  stubs.sampling, 18
  usag, 18
  zetatable, 5, 15, 19

binom.stat, 3, 14
book, 4
coauthor, 4
conpowerlaw.stat, 5, 15
erdos.renyi.game, 6, 7
graph, 18
graph.rmedge, 6, 11
group.graph, 7
igraph, 2, 17
karate, 8
mcpv.function, 8
multinom.stat, 9
network.mc.scan, 10, 12
network.scan, 9, 11, 11
norm.stat, 12
pois.stat, 13
powerlaw.stat, 5, 14
rmulti.one, 15
rmultinom, 16
rplcon, 5, 15
SNscan (SNscan-package), 2
SNscan-package, 2
structure.stat, 16
usag, 18
zetatable, 5, 15, 19