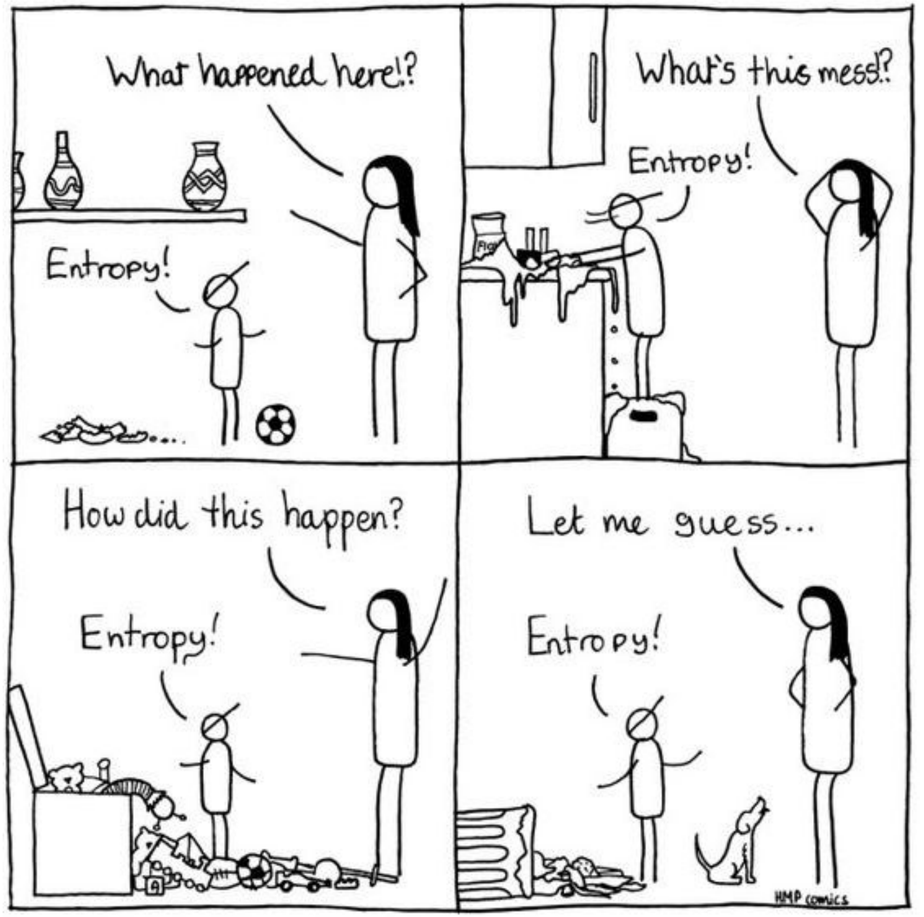
Statistical Entropy Analysis of Network Data

Termeh Shafie



The Women in Network Science (WiNS) seminar
2 May 2022

mrs.schochastics.net
@termehshafie
@termehs

analysing multivariate (social) networks

analysing multivariate (social) networks

framework to assess the interdependence between composition and structure include

- ☑ finding informative dyads and triads (and also higher order configurations)
- ☑ explore the data to inspire further investigations
- ☑ identify social phenomena and processes
- ☑ specify various multivariate models (multiplex, multi-level, etc.)

analysing multivariate (social) networks

framework to assess the interdependence between composition and structure include

- ☑ finding informative dyads and triads (and also higher order configurations)
- ☑ explore the data to inspire further investigations
- ☑ identify social phenomena and processes
- ☑ specify various multivariate models (multiplex, multi-level, etc.)

measures of spread, flatness, association and dependence

that are based on entropy and developed in information theory

analysing multivariate (social) networks

framework to assess the interdependence between composition and structure include

- ☑ finding informative dyads and triads (and also higher order configurations)
- ☑ explore the data to inspire further investigations
- ☑ identify social phenomena and processes
- ☑ specify various multivariate models (multiplex, multi-level, etc.)

measures of spread, flatness, association and dependence

that are based on entropy and developed in information theory

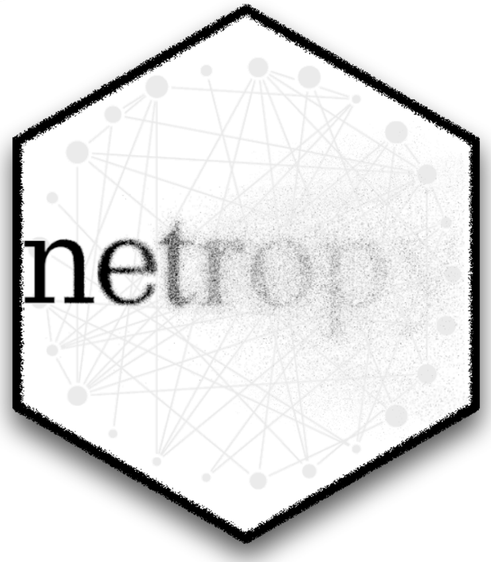
multivariate statistical entropies are used to find

- ☑ redundancies
- ☑ (conditional) dependencies
- ☑ functional dependencies

among a multi-dimensional set of variables measured on different scales

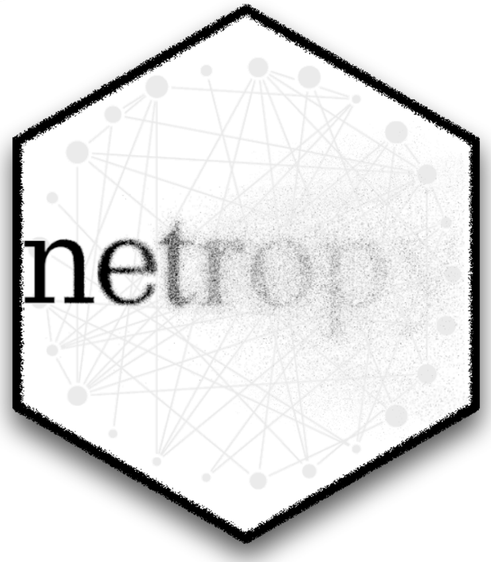
R package {netropy}

```r
install.packages("netropy")

# install.packages("devtools")
devtools::install_github("termehs/netropy")
```



R package {netropy}

```r
install.packages("netropy")

# install.packages("devtools")
devtools::install_github("termehs/netropy")
```



☑ **running example: network study of corporate law firm***

relations between 71 lawyers of a firm:	actor attributes:
- undirected co-work - directed advice - directed friendship	- seniority - formal status - gender - office location - years with the firm - age - practice - law school attended

```r
data(lawdata)
adj.advice <- lawdata[[1]]
adj.friend <- lawdata[[2]]
adj.cowork <-lawdata[[3]]
df.att <- lawdata[[4]]
```

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

variable domains and range spaces

data editing

- ☑ all variables are defined with specified domains and range spaces
- ☑ only consider variables with the same domain together
- ☑ variables on different domains can sometimes be combined

variable domains and range spaces

data editing

- ☑ all variables are defined with specified domains and range spaces
- ☑ only consider variables with the same domain together
- ☑ variables on different domains can sometimes be combined

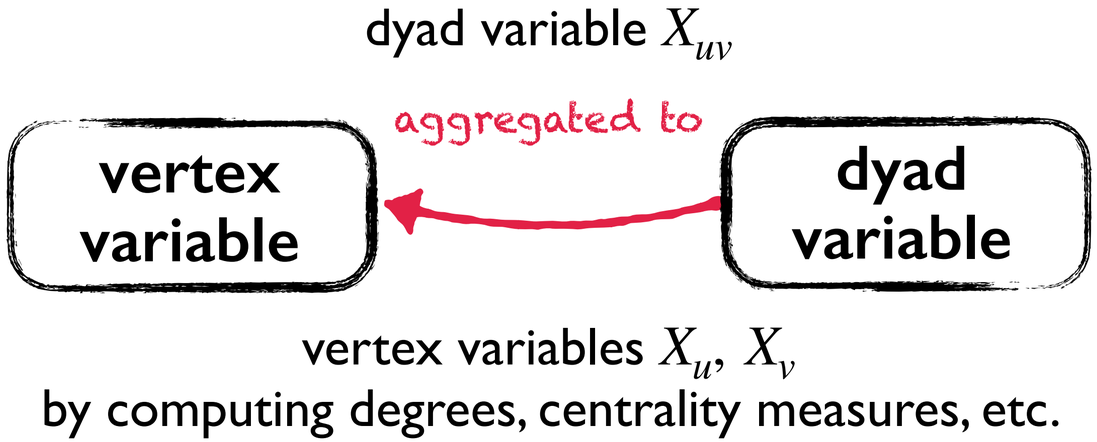
binary vertex variable X_u

trinary dyad variable $X_{uv} = X_u + X_v$

variable domains and range spaces

data editing

- ☑ all variables are defined with specified domains and range spaces
- ☑ only consider variables with the same domain together
- ☑ variables on different domains can sometimes be combined



dyad and triad sequences

node attribute X undirected relation Y directed relation Z

dyad and triad sequences

node attribute X undirected relation Y directed relation Z

dyad sequences

$$S_{uv} = (X_u, X_v, Y_{uv}, Z_{uv}, Z_{vu})$$

example

$S_{uv} = (1,0,\ 1,\ 0,1)$

dyad and triad sequences

node attribute X undirected relation Y directed relation Z

dyad sequences

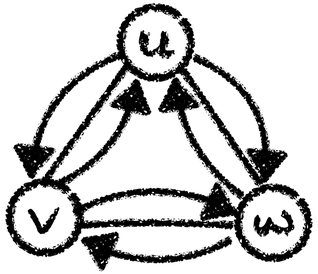
$$S_{uv} = (X_u, X_v, Y_{uv}, Z_{uv}, Z_{vu})$$

example

$S_{uv} = (1,0,\ 1,\ 0,1)$

triad sequences

$$S_{uvw} = (X_{uvw}, Y_{uvw}, Z_{uvw})$$

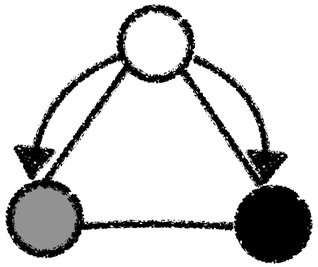


where

$$X_{uvw} = (X_u, X_v, X_w)$$
$$Y_{uvw} = (Y_{uv}, Y_{uw}, Y_{vw})$$
$$Z_{uvw} = (Z_{uv}, Z_{vu}, Z_{uw}, Z_{wu}, Z_{vw}, Z_{wv})$$

example

$S_{uvw} = (0,1,2,\ 1,1,1,\ 1,0,1,0,0,0)$



dyad and triad sequences

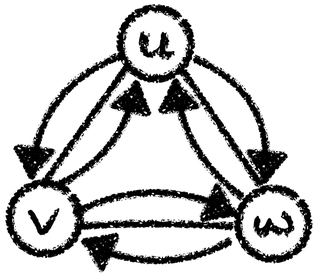
node attribute X undirected relation Y directed relation Z

dyad sequences

$$S_{uv} = (X_u, X_v, Y_{uv}, Z_{uv}, Z_{vu})$$

triad sequences

$$S_{uvw} = (X_{uvw}, Y_{uvw}, Z_{uvw})$$



where

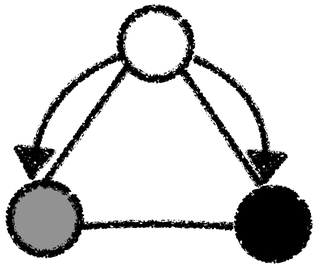
$$X_{uvw} = (X_u, X_v, X_w)$$
$$Y_{uvw} = (Y_{uv}, Y_{uw}, Y_{vw})$$
$$Z_{uvw} = (Z_{uv}, Z_{vu}, Z_{uw}, Z_{wu}, Z_{vw}, Z_{wv})$$

example

$S_{uv} = (1,0,\ 1,\ 0,1)$

example

$S_{uvw} = (0,1,2,\ 1,1,1,\ 1,0,1,0,0,0)$



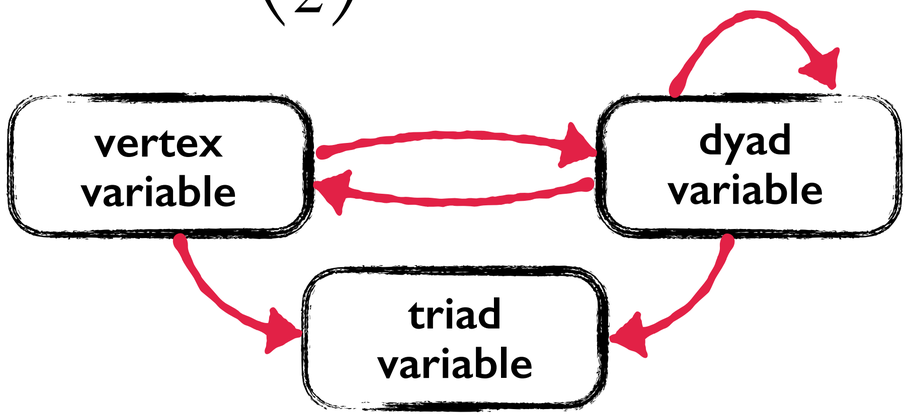
index multiple variables of each kind e.g. for dyad variables:

$$(X_{1u}, X_{1v}, X_{2u}, X_{2v}, \ldots, Y_{1uv}, Y_{2uv}, \ldots, Z_{1uv}, Z_{1vu}, Z_{2uv}, Z_{2vu}, \ldots)$$

example: network study of corporate law firm

number of observations:

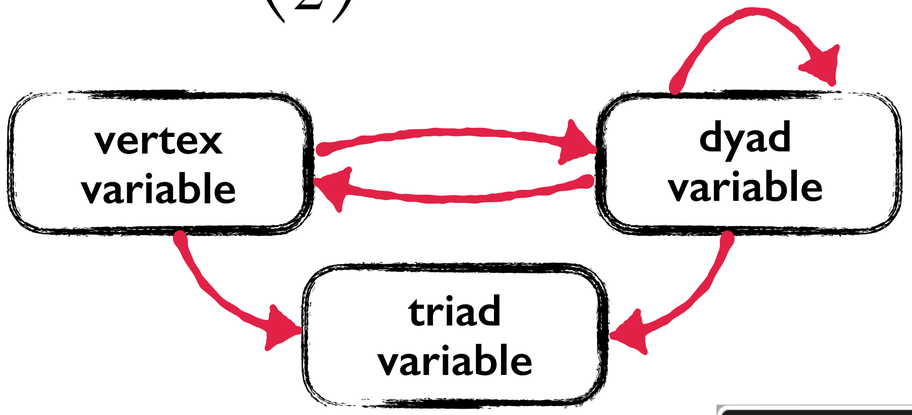
vertices: $n = 71$ # dyads: $\binom{n}{2} = 2485$ # triads: $\binom{n}{3} = 57155$



example: network study of corporate law firm

number of observations:

vertices: $n = 71$ # dyads: $\binom{n}{2} = 2485$ # triads: $\binom{n}{3} = 57155$



dataframe of observed and categorized vertex variables:

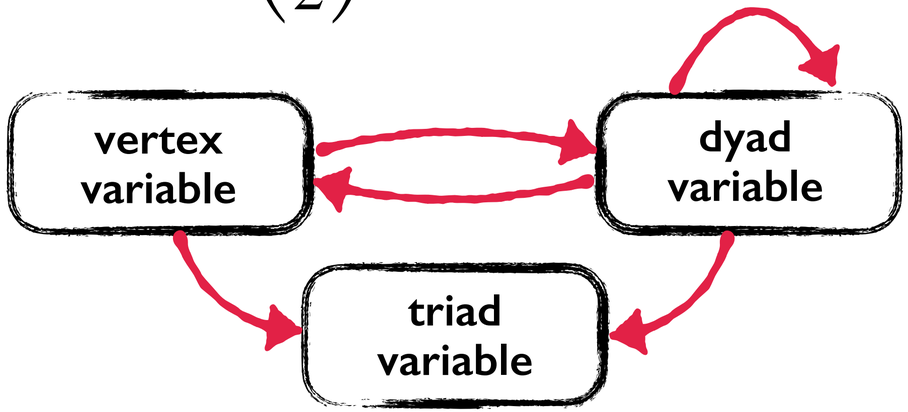
```
##   senior status gender office years age practice lawschool
## 1      1      1      1      0     2   2        1          0
## 2      2      1      1      0     2   2        0          0
## 3      3      1      1      1     1   2        1          0
## 4      4      1      1      0     2   2        0          2
## 5      5      1      1      1     2   2        1          1
```

```
df.att.var <- data.frame(
   senior    = df.att$senior,
   status    = df.att$status,
   gender    = df.att$gender,
   office    = df.att$office-1,
   years     = ifelse(df.att$years<=3,0,
               ifelse(df.att$years<=13,1,2)),
   age       = ifelse(df.att$age<=35,0,
                 ifelse(df.att$age<=45,1,2)),
   practice = df.att$practice,
   lawschool= df.att$lawschool-1)
```

example: network study of corporate law firm

number of observations:

vertices: $n = 71$ # dyads: $\binom{n}{2} = 2485$ # triads: $\binom{n}{3} = 57155$



dataframe of transformed dyad variables:

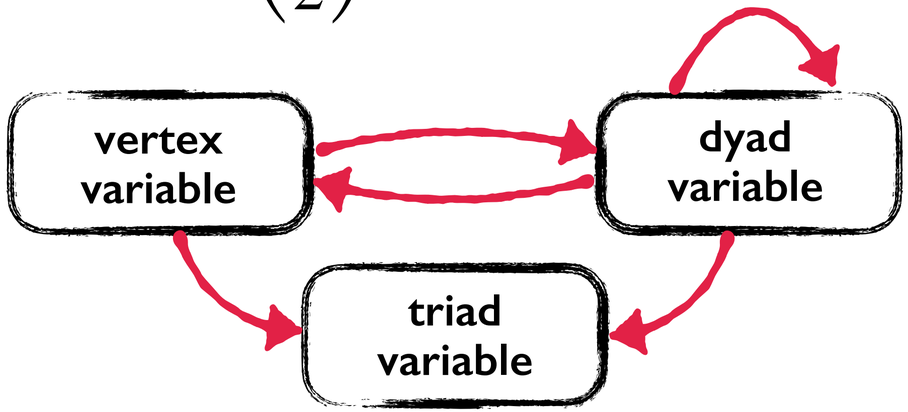
```
##   status gender office years age practice lawschool cowork advice friend
## 1      3      3      0     8   8        1          0      0      3      2
## 2      3      3      3     5   8        3          0      0      0      0
## 3      3      3      3     5   8        2          0      0      1      0
## 4      3      3      0     8   8        1          6      0      1      2
## 5      3      3      0     8   8        0          6      0      1      1
```

```r
# transformed dyad variables
get_dyad_var(var, type = "att")

# transformed triad variables
get_triad_var(var, type = "att")
```

example: network study of corporate law firm

number of observations:

vertices: $n = 71$ # dyads: $\binom{n}{2} = 2485$ # triads: $\binom{n}{3} = 57155$

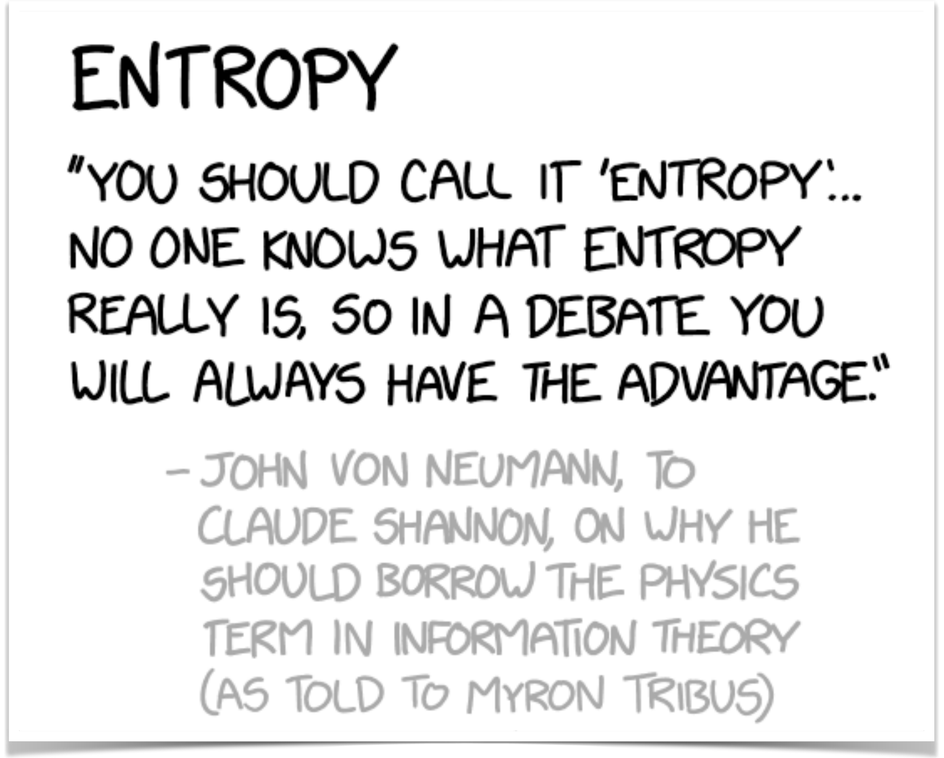


dataframe of transformed triad variables:

	status	gender	office	years	age	practice	lawschool	cowork	advice	friend
1	7	7	9	17	26	5	0	0	35	1
2	7	7	0	26	26	1	18	0	43	37
3	7	7	9	26	26	5	9	0	11	1
4	7	7	9	26	26	5	0	0	19	1
5	7	7	9	26	26	1	18	4	35	1

```r
# transformed dyad variables
get_dyad_var(var, type = "att")

# transformed triad variables
get_triad_var(var, type = "att")
```

what is entropy?



univariate entropy

statistical entropy is a measure of uncertainty of random variables

univariate entropy

statistical entropy is a measure of uncertainty of random variables

for a discrete random variable X with a finite range space of size r_X

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} \qquad p(x) > 0, \ \sum_x p(x) = 1$$

- ☑ minimal zero entropy has no uncertainty
- ☑ maximum entropy $\log_2(r_X) \implies$ uniform distribution

univariate entropy

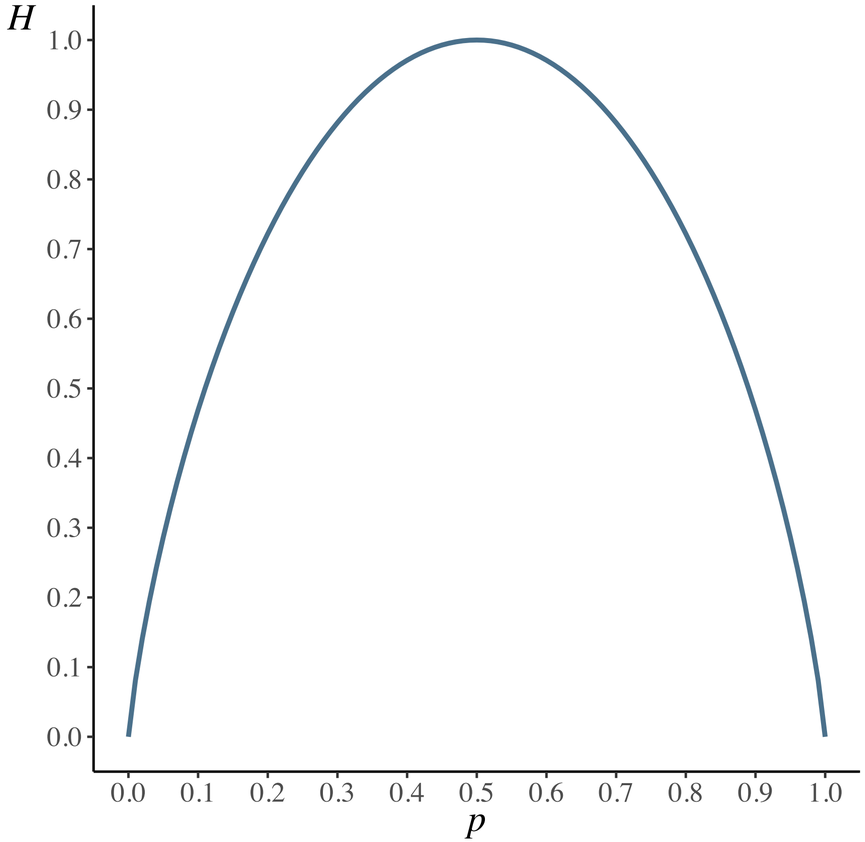
statistical entropy is a measure of uncertainty of random variables

for a discrete random variable X with a finite range space of size r_X

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} \qquad p(x) > 0, \ \sum_x p(x) = 1$$

- ☑ minimal zero entropy has no uncertainty
- ☑ maximum entropy $\log_2(r_X) \implies$ uniform distribution

example:
entropy of a binary variable



univariate entropy

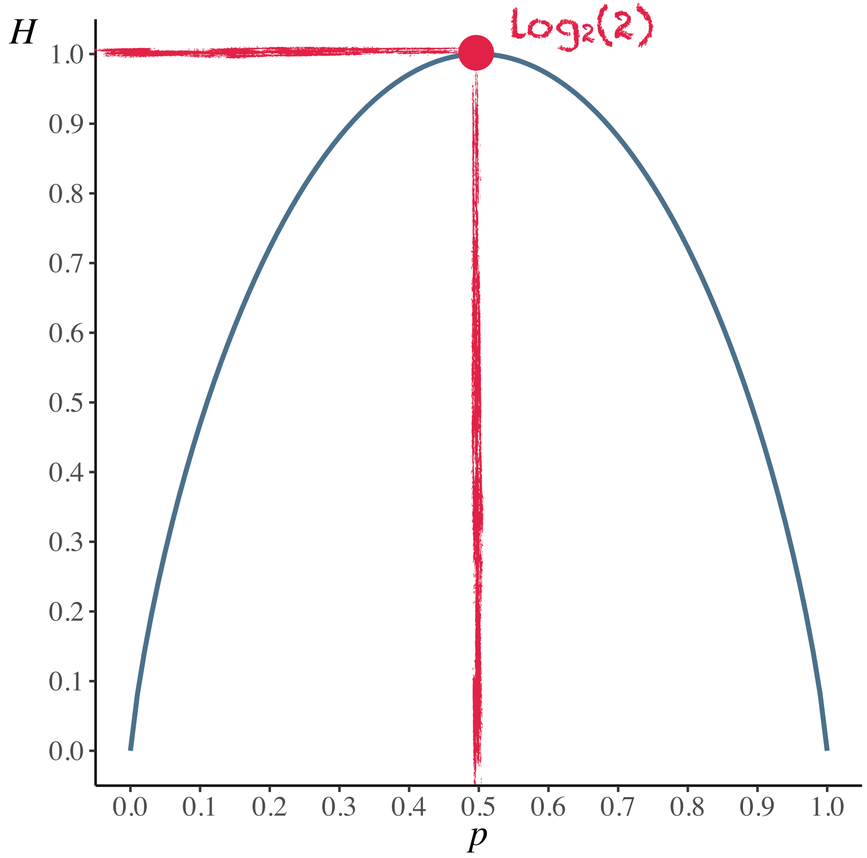
statistical entropy is a measure of uncertainty of random variables

for a discrete random variable X with a finite range space of size r_X

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} \qquad p(x) > 0, \ \sum_x p(x) = 1$$

- ☑ minimal zero entropy has no uncertainty
- ☑ maximum entropy $\log_2(r_X)$ $\Longrightarrow$ uniform distribution

example:
entropy of a binary variable



bivariate and joint entropies

for two discrete random variables X and Y the bivariate entropy is given by

$$H(X, Y) = \sum_x \sum_y p(x, y) \log_2 \frac{1}{p(x, y)}$$

and bounded according to

$$H(X) \leq H(X, Y) \leq H(X) + H(Y)$$

bivariate and joint entropies

for two discrete random variables X and Y the bivariate entropy is given by

$$H(X, Y) = \sum_x \sum_y p(x, y) \log_2 \frac{1}{p(x, y)}$$

and bounded according to

$$H(X) \leq H(X, Y) \leq H(X) + H(Y)$$

equality to the left iff $X \to Y$ equality to the right iff $X \perp Y$

bivariate and joint entropies

for two discrete random variables X and Y the bivariate entropy is given by

$$H(X,Y) = \sum_x \sum_y p(x,y) \log_2 \frac{1}{p(x,y)}$$

and bounded according to

$$H(X) \leq H(X,Y) \leq H(X) + H(Y)$$

equality to the left iff $X \to Y$ equality to the right iff $X \perp Y$

the two increments of the inequalities around $H(X,Y)$:

☑ **joint entropy**

$$J(X,Y) = H(X) + H(Y) - H(X,Y)$$

non-negative and equal to 0 iff $X \perp Y$

☑ **expected conditional entropy**

$$EH(Y|X) = H(X,Y) - H(X)$$

non-negative and equal to 0 iff $X \to Y$

bivariate and joint entropies

for two discrete random variables X and Y the bivariate entropy is given by

$$H(X,Y) = \sum_x \sum_y p(x,y) \log_2 \frac{1}{p(x,y)}$$

} variable selection

and bounded according to

$$H(X) \leq H(X,Y) \leq H(X) + H(Y)$$

equality to the left iff $X \to Y$ equality to the right iff $X \perp Y$

the two increments of the inequalities around $H(X,Y)$:

☑ **joint entropy**

$$J(X,Y) = H(X) + H(Y) - H(X,Y)$$

non-negative and equal to 0 iff $X \perp Y$

} association graphs
divergence statistic

☑ **expected conditional entropy**

$$EH(Y|X) = H(X,Y) - H(X)$$

non-negative and equal to 0 iff $X \to Y$

} prediction power

trivariate and higher order entropies

for three discrete random variables X, Y and Z the trivariate entropy is bounded

$$H(X, Y) \leq H(X, Y, Z) \leq H(X, Z) + H(Y, Z) - H(Z)$$

the two increments of the inequalities around $H(X, Y, Z)$:

☑ **expected joint entropy**

$$EJ(X, Y \mid Z) = H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z)$$

non-negative and equal to 0 iff $(X, Y) \perp Z$

☑ **expected conditional entropy**

$$EH(Z \mid X, Y) = H(X, Y, Z) - H(X, Y)$$

non-negative and equal to 0 iff $X \to Y$

similarily for $H(X, Y, Z, U)$, $H(X, Y, Z, U, V)$, $H(X, Y, Z, U, V, W)$, …

trivariate and higher order entropies

for three discrete random variables X, Y and Z the trivariate entropy is bounded

$$H(X, Y) \leq H(X, Y, Z) \leq H(X, Z) + H(Y, Z) - H(Z)$$

the two increments of the inequalities around $H(X, Y, Z)$:

☑ **expected joint entropy**

$$EJ(X, Y \mid Z) = H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z)$$

non-negative and equal to 0 iff $(X, Y) \perp Z$

} association graph
divergence statistic

☑ **expected conditional entropy**

$$EH(Z \mid X, Y) = H(X, Y, Z) - H(X, Y)$$

non-negative and equal to 0 iff $X \to Y$

} prediction power

similarily for $H(X, Y, Z, U)$, $H(X, Y, Z, U, V)$, $H(X, Y, Z, U, V, W)$, …

example: network study of corporate law firm

☑ univariate and bivariate entropies of vertex variables

```
##           senior status gender office years  age practice lawschool
## senior      6.15   6.15   6.15   6.15  6.15 6.15     6.15       6.15
## status        NA   1.00   1.70   2.08  2.01 2.28     1.98       2.46
## gender        NA     NA   0.82   1.93  2.23 2.38     1.80       2.32
## office        NA     NA     NA   1.12  2.69 2.67     2.09       2.61
## years         NA     NA     NA     NA  1.58 2.75     2.56       3.01
## age           NA     NA     NA     NA    NA 1.58     2.56       2.88
## practice      NA     NA     NA     NA    NA   NA     0.98       2.51
## lawschool     NA     NA     NA     NA    NA   NA       NA       1.53
```

```
# matrix with bivariate entropies
H <- entropy_bivar(dat)
diag(H) # univariate entropies
```

example: network study of corporate law firm

☑ univariate and bivariate entropies of vertex variables

	senior	status	gender	office	years	age	practice	lawschool
senior	6.15	6.15	6.15	6.15	6.15	6.15	6.15	6.15
status	NA	1.00	1.70	2.08	2.01	2.28	1.98	2.46
gender	NA	NA	0.82	1.93	2.23	2.38	1.80	2.32
office	NA	NA	NA	1.12	2.69	2.67	2.09	2.61
years	NA	NA	NA	NA	1.58	2.75	2.56	3.01
age	NA	NA	NA	NA	NA	1.58	2.56	2.88
practice	NA	NA	NA	NA	NA	NA	0.98	2.51
lawschool	NA	NA	NA	NA	NA	NA	NA	1.53

redundancy

```
# matrix with bivariate entropies
H <- entropy_bivar(dat)
diag(H) # univariate entropies
```

example: network study of corporate law firm

☑ univariate and bivariate entropies of vertex variables

	senior	status	gender	office	years	age	practice	lawschool
senior	6.15	6.15	6.15	6.15	6.15	6.15	6.15	6.15
status	NA	1.00	1.70	2.08	2.01	2.28	1.98	2.46
gender	NA	NA	0.82	1.93	2.23	2.38	1.80	2.32
office	NA	NA	NA	1.12	2.69	2.67	2.09	2.61
years	NA	NA	NA	NA	1.58	2.75	2.56	3.01
age	NA	NA	NA	NA	NA	1.58	2.56	2.88
practice	NA	NA	NA	NA	NA	NA	0.98	2.51
lawschool	NA	NA	NA	NA	NA	NA	NA	1.53

redundancy

```
# matrix with bivariate entropies
H <- entropy_bivar(dat)
diag(H) # univariate entropies
```

☑ redundant variables

	senior	status	gender	office	years	age	practice	lawschool
senior	0	1	1	1	1	1	1	1
status	0	0	0	0	0	0	0	0
gender	0	0	0	0	0	0	0	0
office	0	0	0	0	0	0	0	0
years	0	0	0	0	0	0	0	0
age	0	0	0	0	0	0	0	0
practice	0	0	0	0	0	0	0	0
lawschool	0	0	0	0	0	0	0	0

```
# detect redundancy
redundancy(dat, dec = 3)
```

example: network study of corporate law firm

☑ joint entropies of dyad variables

```r
J <- joint_entropy(dat, dec = 3)
J$matrix # matrix of joint entropies
J$freq # table of joint entropy frequencies
```

```
##           status gender office years  age practice lawschool cowork advice friend
## status      1.49   0.17   0.09  0.79 0.38     0.00       0.08   0.02   0.05   0.05
## gender        NA   1.55   0.03  0.28 0.07     0.00       0.06   0.00   0.01   0.01
## office        NA     NA   2.24  0.08 0.14     0.05       0.13   0.06   0.10   0.08
## years         NA     NA     NA  2.67 0.61     0.05       0.20   0.02   0.05   0.07
## age           NA     NA     NA    NA 2.80     0.02       0.41   0.01   0.02   0.05
## practice      NA     NA     NA    NA   NA     1.96       0.04   0.05   0.08   0.01
## lawschool     NA     NA     NA    NA   NA       NA       2.95   0.00   0.01   0.02
## cowork        NA     NA     NA    NA   NA       NA         NA   0.62   0.18   0.04
## advice        NA     NA     NA    NA   NA       NA         NA     NA   1.25   0.18
## friend        NA     NA     NA    NA   NA       NA         NA     NA     NA   0.88
```

```
##       j  #(J = j) #(J >= j)
## 1  0.79         1         1
## 2  0.61         1         2
## 3  0.41         1         3
## 4  0.38         1         4
## 5  0.28         1         5
## 6   0.2         1         6
## 7  0.18         2         8
## 8  0.17         1         9
## 9  0.14         1        10
## 10 0.13         1        11
## 11  0.1         1        12
## 12 0.09         1        13
## 13 0.08         4        17
## 14 0.07         2        19
## 15 0.06         2        21
## 16 0.05         7        28
## 17 0.04         2        30
## 18 0.03         1        31
## 19 0.02         5        36
## 20 0.01         5        41
## 21    0         4        45
```

example: network study of corporate law firm

☑ joint entropies of dyad variables

```
J <- joint_entropy(dat, dec = 3)
J$matrix # matrix of joint entropies
J$freq # table of joint entropy frequencies
```

	status	gender	office	years	age	practice	lawschool	cowork	advice	friend
status	1.49	0.17	0.09	0.79	0.38	0.00	0.08	0.02	0.05	0.05
gender	NA	1.55	0.03	0.28	0.07	0.00	0.06	0.00	0.01	0.01
office	NA	NA	2.24	0.08	0.14	0.05	0.13	0.06	0.10	0.08
years	NA	NA	NA	2.67	0.61	0.05	0.20	0.02	0.05	0.07
age	NA	NA	NA	NA	2.80	0.02	0.41	0.01	0.02	0.05
practice	NA	NA	NA	NA	NA	1.96	0.04	0.05	0.08	0.01
lawschool	NA	NA	NA	NA	NA	NA	2.95	0.00	0.01	0.02
cowork	NA	NA	NA	NA	NA	NA	NA	0.62	0.18	0.04
advice	NA	NA	NA	NA	NA	NA	NA	NA	1.25	0.18
friend	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.88

strongest dependence

	j	#(J = j)	#(J >= j)
1	0.79	1	1
2	0.61	1	2
3	0.41	1	3
4	0.38	1	4
5	0.28	1	5
6	0.2	1	6
7	0.18	2	8
8	0.17	1	9
9	0.14	1	10
10	0.13	1	11
11	0.1	1	12
12	0.09	1	13
13	0.08	4	17
14	0.07	2	19
15	0.06	2	21
16	0.05	7	28
17	0.04	2	30
18	0.03	1	31
19	0.02	5	36
20	0.01	5	41
21	0	4	45

independence

example: network study of corporate law firm

☑ association graph of dyad variables with $J > 0.15$

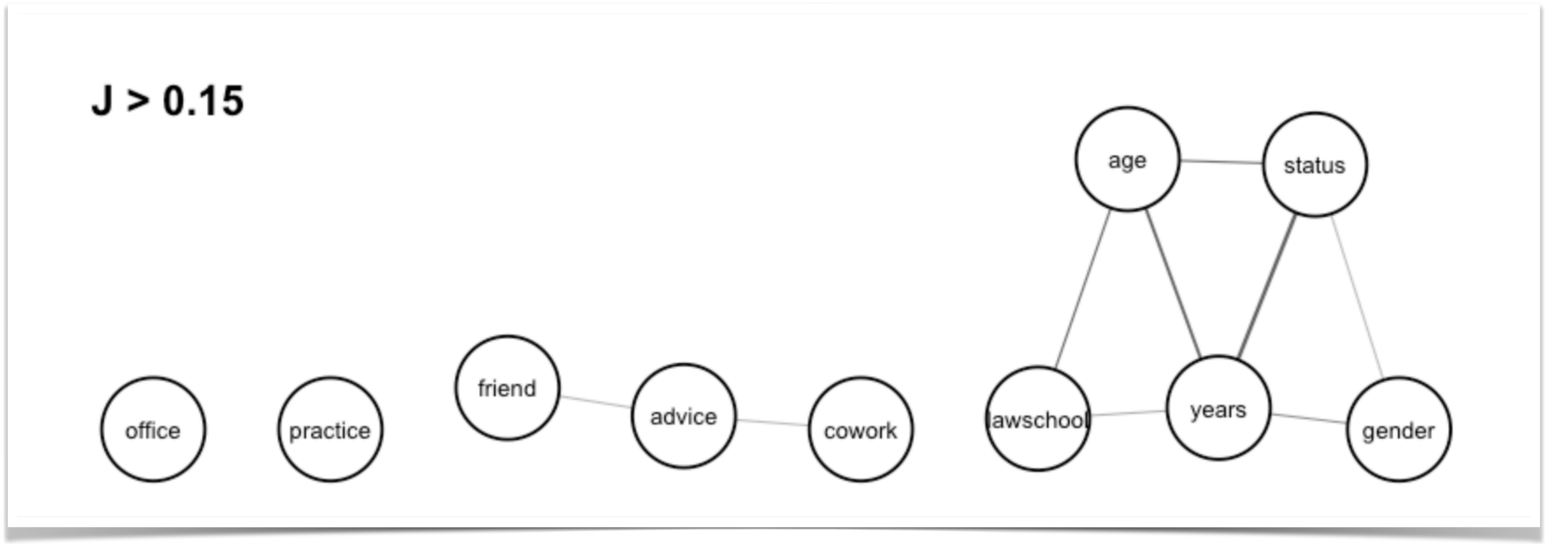
```
library(ggraph)
assoc_graph(dat, cutoff = 0)
```

```
##       j  #(J = j) #(J >= j)
## 1  0.79         1          1
## 2  0.61         1          2
## 3  0.41         1          3
## 4  0.38         1          4
## 5  0.28         1          5
## 6   0.2         1          6
## 7  0.18         2          8
## 8  0.17         1          9
## 9  0.14         1         10
## 10 0.13         1         11
## 11  0.1         1         12
## 12 0.09         1         13
## 13 0.08         4         17
## 14 0.07         2         19
## 15 0.06         2         21
## 16 0.05         7         28
## 17 0.04         2         30
## 18 0.03         1         31
## 19 0.02         5         36
## 20 0.01         5         41
## 21    0         4         45
```

example: network study of corporate law firm

☑ association graph of dyad variables with $J > 0.15$

```
library(ggraph)
assoc_graph(dat, cutoff = 0)
```




```
##       j  #(J = j) #(J >= j)
## 1  0.79         1         1
## 2  0.61         1         2
## 3  0.41         1         3
## 4  0.38         1         4
## 5  0.28         1         5
## 6   0.2         1         6
## 7  0.18         2         8
## 8  0.17         1         9
## 9  0.14         1        10
## 10 0.13         1        11
## 11  0.1         1        12
## 12 0.09         1        13
## 13 0.08         4        17
## 14 0.07         2        19
## 15 0.06         2        21
## 16 0.05         7        28
## 17 0.04         2        30
## 18 0.03         1        31
## 19 0.02         5        36
## 20 0.01         5        41
## 21    0         4        45
```

✓ associations between components and cliques

✓ comparisons and tests of tentative dependence structures

example: network study of corporate law firm

☑ association graph of dyad variables with $J > 0.15$

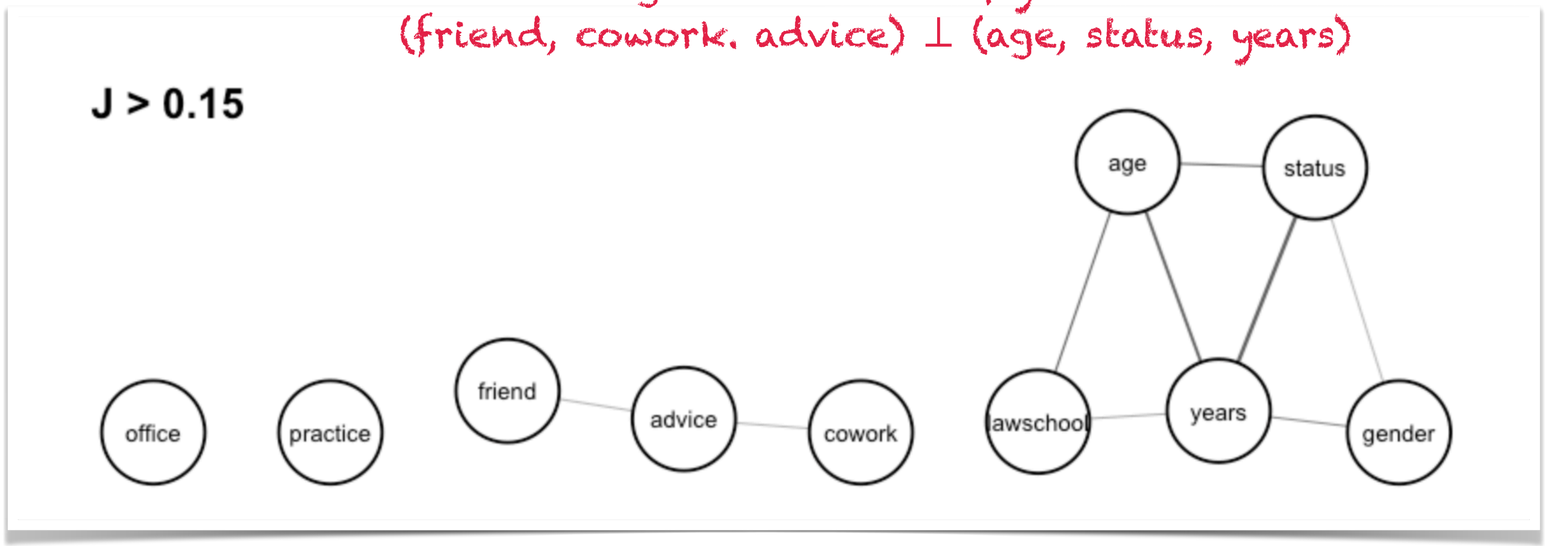
```
library(ggraph)
assoc_graph(dat, cutoff = 0)
```

example of structural models of interest:

friend ⊥ cowork | advice

gender ⊥ status | years

(friend, cowork. advice) ⊥ (age, status, years)



✓ associations between components and cliques

✓ comparisons and tests of tentative dependence structures

```
##        j  #(J = j) #(J >= j)
## 1   0.79         1         1
## 2   0.61         1         2
## 3   0.41         1         3
## 4   0.38         1         4
## 5   0.28         1         5
## 6    0.2         1         6
## 7   0.18         2         8
## 8   0.17         1         9
## 9   0.14         1        10
## 10  0.13         1        11
## 11   0.1         1        12
## 12  0.09         1        13
## 13  0.08         4        17
## 14  0.07         2        19
## 15  0.06         2        21
## 16  0.05         7        28
## 17  0.04         2        30
## 18  0.03         1        31
## 19  0.02         5        36
## 20  0.01         5        41
## 21     0         4        45
```

example: network study of corporate law firm

☑ prediction power based one expected conditional entropy $EH(Z|X,Y)$

finding good predictors:
variables (almost) uniquely determined
by combinations of other

```
# prediction power matrix with E(Z|X,Y)
pp <- prediction_power(var, dat)
diag(pp) # single variable prediction EH(Z|X)
```

example: network study of corporate law firm

☑ prediction power based one expected conditional entropy $EH(Z|X,Y)$

finding good predictors:
variables (almost) uniquely determined
by combinations of other

```
# prediction power matrix with E(Z|X,Y)
pp <- prediction_power(var, dat)
diag(pp) # single variable prediction EH(Z|X)
```

predicting $Z =$ status:

```
##           status gender office years   age practice lawschool cowork advice friend
## status        NA     NA     NA    NA    NA       NA         NA     NA     NA     NA
## gender        NA  1.375  1.180 0.670 0.855    1.304      1.225  1.306  1.263  1.270
## office        NA     NA  2.147 0.493 0.820    1.374      1.245  1.373  1.325  1.334
## years         NA     NA     NA 2.265 0.573    0.682      0.554  0.691  0.667  0.684
## age           NA     NA     NA    NA 1.877    1.089      0.958  1.087  1.052  1.058
## practice      NA     NA     NA    NA    NA    2.446      1.388  1.459  1.410  1.427
## lawschool     NA     NA     NA    NA    NA       NA      3.335  1.390  1.337  1.350
## cowork        NA     NA     NA    NA    NA       NA         NA  2.419  1.400  1.411
## advice        NA     NA     NA    NA    NA       NA         NA     NA  2.781  1.407
## friend        NA     NA     NA    NA    NA       NA         NA     NA     NA  3.408
```

interpretation when EH is rounded to its closest integer:

- ✓ unambiguous prediction of Z when $EH < 0.5$
- ✓ two prediction values for Z when $0.5 \leq EH \leq 1$
- ✓ etc.

example: network study of corporate law firm

☑ prediction power based one expected conditional entropy $EH(Z|X,Y)$

finding good predictors:
variables (almost) uniquely determined
by combinations of other

```
# prediction power matrix with E(Z|X,Y)
pp <- prediction_power(var, dat)
diag(pp) # single variable prediction EH(Z|X)
```

predicting $Z =$ status:

```
##           status gender office years   age practice lawschool cowork advice friend
## status        NA     NA     NA    NA    NA       NA         NA     NA     NA     NA
## gender        NA  1.375  1.180 0.670 0.855    1.304      1.225  1.306  1.263  1.270
## office        NA     NA  2.147 0.493 0.820    1.374      1.245  1.373  1.325  1.334
## years         NA     NA     NA 2.265 0.573    0.682      0.554  0.691  0.667  0.684
## age           NA     NA     NA    NA 1.877    1.089      0.958  1.087  1.052  1.058
## practice      NA     NA     NA    NA    NA    2.446      1.388  1.459  1.410  1.427
## lawschool     NA     NA     NA    NA    NA       NA      3.335  1.390  1.337  1.350
## cowork        NA     NA     NA    NA    NA       NA         NA  2.419  1.400  1.411
## advice        NA     NA     NA    NA    NA       NA         NA     NA  2.781  1.407
## friend        NA     NA     NA    NA    NA       NA         NA     NA     NA  3.408
```

best predictors of 'status':
(years, office)
(age, years)
(lawschool, years)

interpretation when EH is rounded to its closest integer:

- ✓ unambiguous prediction of Z when $EH < 0.5$
- ✓ two prediction values for Z when $0.5 \leq EH \leq 1$
- ✓ etc.

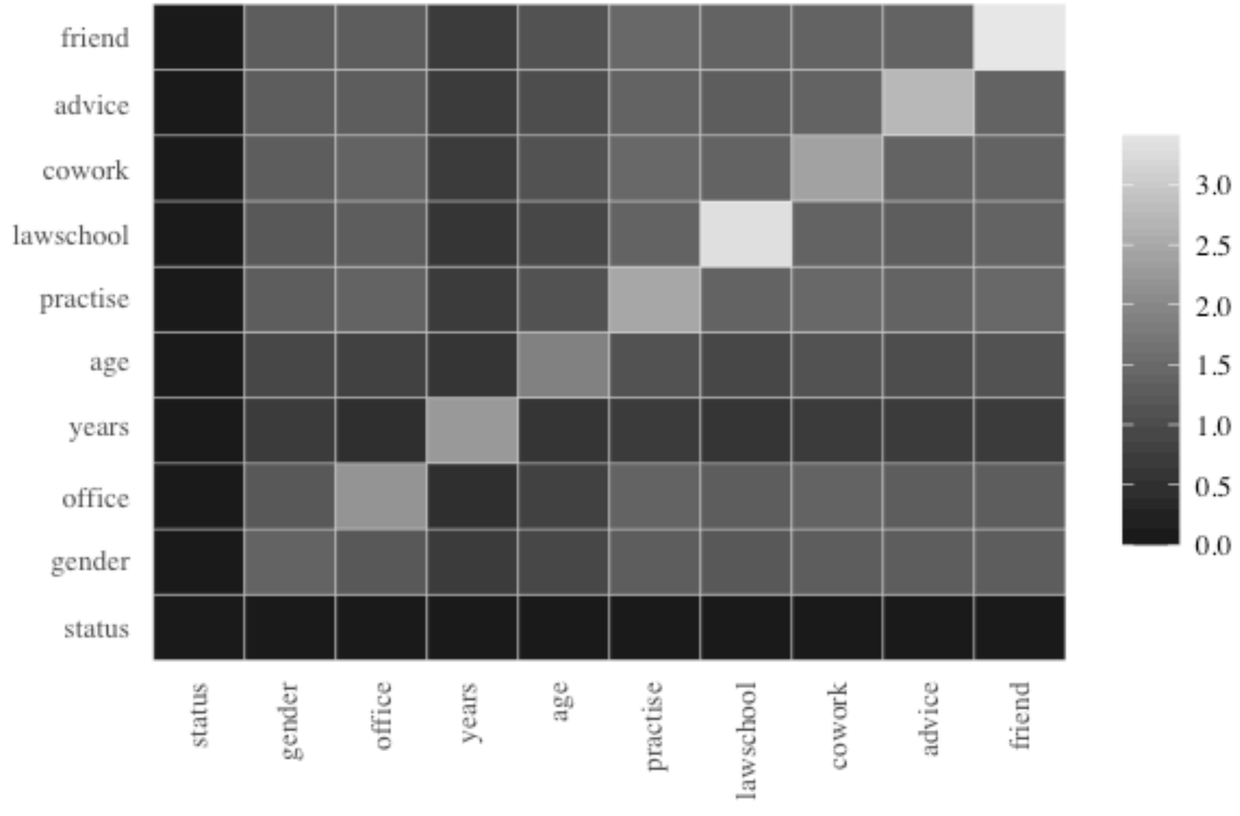
example: network study of corporate law firm

☑ prediction power based one expected conditional entropy $EH(Z|X,Y)$

finding good predictors:
variables (almost) uniquely determined
by combinations of other

```
# prediction power matrix with E(Z|X,Y)
pp <- prediction_power(var, dat)
diag(pp) # single variable prediction EH(Z|X)
```

prediction power visualized using ggplot:



best predictors of 'status':

(years, office)

(age, years)

(lawschool, years)

divergence tests of goodness of fit

goodness of fit tests of hypothetical multivariate discrete distributions
(as suggested by association graphs)

$p =$ general model based on empirical distribution with estimated likelihood function $L(p)$

$q =$ data follows a specified probability model with estimated likelihood function $L(q)$

divergence tests of goodness of fit

goodness of fit tests of hypothetical multivariate discrete distributions
(as suggested by association graphs)

$p =$ general model based on empirical distribution with estimated likelihood function $L(p)$

$q =$ data follows a specified probability model with estimated likelihood function $L(q)$

☑ log likelihood ratio test statistic with d degrees of freedom (for large n)

$$2 \log \frac{L(p)}{L(q)} = 2nD(p,q) \underset{approx}{\sim} \chi^2 \left(d \right)$$

where

$D(p,q)$ is the information divergence (expected log likelihood ratio) with

$d = d(p) - d(q)$ degrees of freedom (numbers of parameters estimated to get p and q)

divergence tests of goodness of fit

goodness of fit tests of hypothetical multivariate discrete distributions
(as suggested by association graphs)

$p =$ general model based on empirical distribution with estimated likelihood function $L(p)$

$q =$ data follows a specified probability model with estimated likelihood function $L(q)$

☑ log likelihood ratio test statistic with d degrees of freedom (for large n)

$$2 \log \frac{L(p)}{L(q)} = 2nD(p,q) \underset{approx}{\sim} \chi^2 \left(d \right)$$

where

$D(p,q)$ is the information divergence (expected log likelihood ratio) with

$d = d(p) - d(q)$ degrees of freedom (numbers of parameters estimated to get p and q)

☑ critical region with approximately 95% confidence level (for large n)

$$\chi^2(d) \geq d + 2\sqrt{2d} = d + \sqrt{8d}$$

divergence tests of goodness of fit

testing uniform distribution

of random variable X with n observations on r_X outcomes

$p =$ model based on empirical distribution $p(x) = n(x)/n$ (the relative frequencies) with $d(p) = r_X - 1$

$q = X$ is uniformly distributed on r_X outcomes with $d(q) = 0$

divergence tests of goodness of fit

testing uniform distribution

of random variable X with n observations on r_X outcomes

p = model based on empirical distribution $p(x) = n(x)/n$ (the relative frequencies) with $d(p) = r_X - 1$

$q = X$ is uniformly distributed on r_X outcomes with $d(q) = 0$

☑ log likelihood ratio test statistic

$$\begin{aligned}\chi^2(r_X - 1) &= 2nD(p, q) \\ &= 2n[\log r_X - H(X)]\end{aligned}$$

where $H(X)$ is the empirical entropy of X

☑ uniformity is rejected if

$$\chi^2(r_X - 1) \geq r_X - 1 + \sqrt{8(r_X - 1)}$$

or if $H(X)$ deviates from its maximum values $\log(r_X)$ by more than $[r_X - 1 + \sqrt{8(r_X - 1)}]/2n$

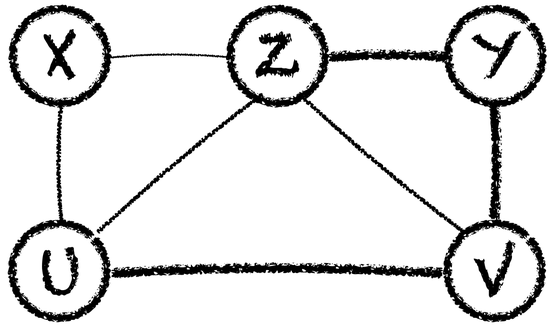
divergence tests of goodness of fit

testing pairwise independence

of random variable X and Y with r_X and r_Y outcomes

$p =$ model based on empirical distribution $p(x, y)$ with $d(p) = r_X r_Y - 1$

$q = X \perp Y$ such that $p(x) \cdot p(y)$ with $d(q) = (r_X - 1) + (r_Y - 1)$



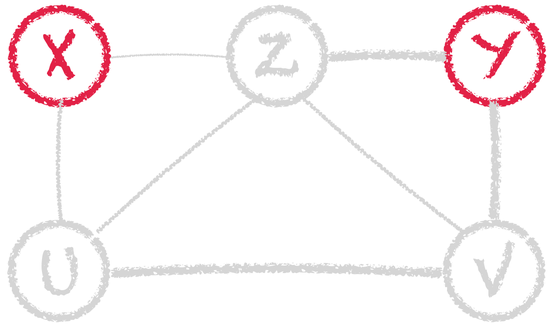
divergence tests of goodness of fit

testing pairwise independence

of random variable X and Y with r_X and r_Y outcomes

$p =$ model based on empirical distribution $p(x, y)$ with $d(p) = r_X r_Y - 1$

$q = X \perp Y$ such that $p(x) \cdot p(y)$ with $d(q) = (r_X - 1) + (r_Y - 1)$



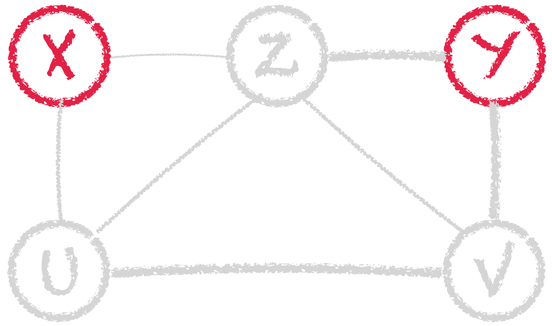
divergence tests of goodness of fit

testing pairwise independence

of random variable X and Y with r_X and r_Y outcomes

$p =$ model based on empirical distribution $p(x, y)$ with $d(p) = r_X r_Y - 1$

$q = X \perp Y$ such that $p(x) \cdot p(y)$ with $d(q) = (r_X - 1) + (r_Y - 1)$



☑ log likelihood ratio test statistic

$$\begin{aligned}\chi^2((r_X - 1)(r_Y - 1)) &= 2nD(p, q) \\ &= 2n[H(X) + H(Y) - H(X, Y)] \\ &= 2nJ(X, Y)\end{aligned}$$

☑ independence is rejected if

$$\chi^2((r_X - 1)(r_Y - 1)) \geq (r_X - 1)(r_Y - 1) + \sqrt{8(r_X - 1)(r_Y - 1)}$$

or if the empirical joint entropy $J(X, Y)$ is larger than $[(r_X - 1)(r_Y - 1) + \sqrt{8(r_X - 1)(r_Y - 1)}]/2n$

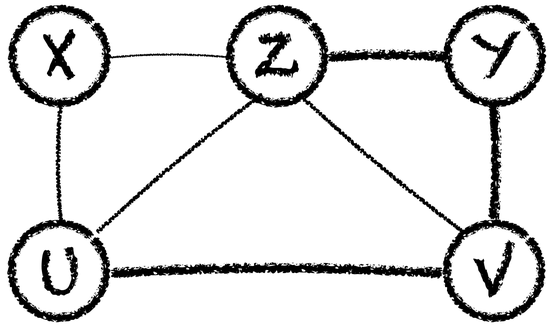
divergence tests of goodness of fit

testing conditional independence

of random variable X, Y and Z with r_X, r_Y and r_Z outcomes

$p =$ model based on empirical distribution $p(x, y, z)$ with $d(p) = r_X r_Y r_Z - 1$

$q = X \perp Y | Z$ such that $p(x, z)p(y, z)/p(z)$ with $d(q) = r_Z - 1 + r_Z(r_X - 1 + r_Y - 1)$



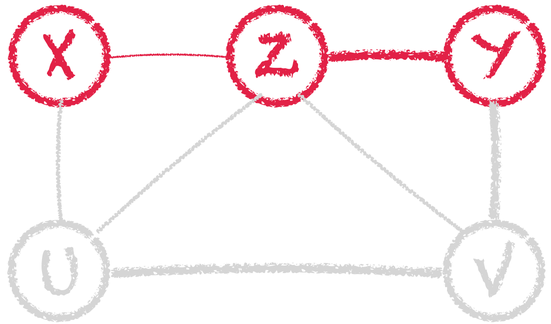
divergence tests of goodness of fit

testing conditional independence

of random variable X, Y and Z with r_X, r_Y and r_Z outcomes

$p =$ model based on empirical distribution $p(x, y, z)$ with $d(p) = r_X r_Y r_Z - 1$

$q = X \perp Y | Z$ such that $p(x, z)p(y, z)/p(z)$ with $d(q) = r_Z - 1 + r_Z(r_X - 1 + r_Y - 1)$

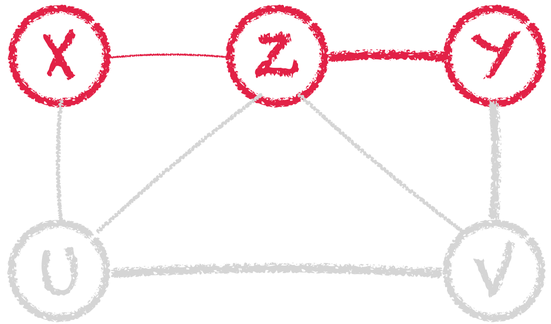


divergence tests of goodness of fit

testing conditional independence
of random variable X, Y and Z with r_X, r_Y and r_Z outcomes

$p =$ model based on empirical distribution $p(x, y, z)$ with $d(p) = r_X r_Y r_Z - 1$

$q = X \perp Y \,|\, Z$ such that $p(x, z)p(y, z)/p(z)$ with $d(q) = r_Z - 1 + r_Z(r_X - 1 + r_Y - 1)$



☑ log likelihood ratio test statistic

$$\begin{aligned}\chi^2((r_X - 1)(r_Y - 1)r_Z) &= 2nD(p, q) \\ &= 2n[H(X, Z) + H(Y, Z) - H(Z) - H(X, Y)] \\ &= 2nEJ(X, Y \,|\, Z)\end{aligned}$$

☑ independence is rejected if

$$\chi^2((r_X - 1)(r_Y - 1)r_Z) \geq (r_X - 1)(r_Y - 1)r_Z + \sqrt{8(r_X - 1)(r_Y - 1)r_Z}$$

or if the empirical expected joint entropy $J(X, Y)$ is larger than

$$[(r_X - 1)(r_Y - 1)r_Z + \sqrt{8(r_X - 1)(r_Y - 1)r_Z}]/2n$$

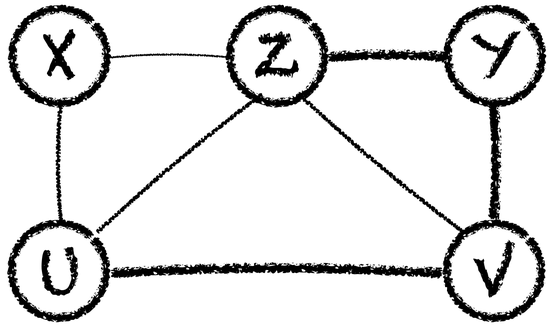
divergence tests of goodness of fit

testing nested model specifications

example: five dimensional data (X, Y, Z, U, V) with r_X, r_Y, r_Z, r_U, r_V outcomes

$p =$ model based on empirical distribution $p(x, y, z, u, v)$ with $d(p) = r_X r_Y r_Z r_U r_V - 1$

$q =$ model with listed imposed independence and conditional independence assumptions



divergence tests of goodness of fit

testing nested model specifications

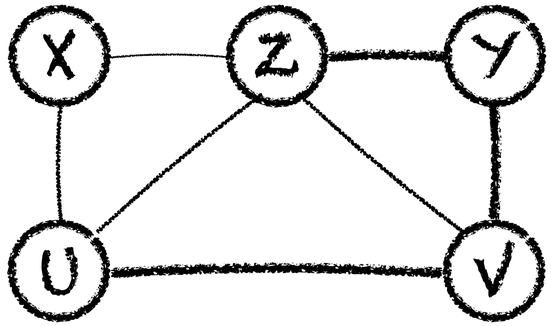
example: five dimensional data (X, Y, Z, U, V) with r_X, r_Y, r_Z, r_U, r_V outcomes

$p =$ model based on empirical distribution $p(x, y, z, u, v)$ with $d(p) = r_X r_Y r_Z r_U r_V - 1$

$q =$ model with listed imposed independence and conditional independence assumptions

examples: $q_1 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$

$q_2 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$ and $Z \perp V \,|\, Y$



divergence tests of goodness of fit

testing nested model specifications

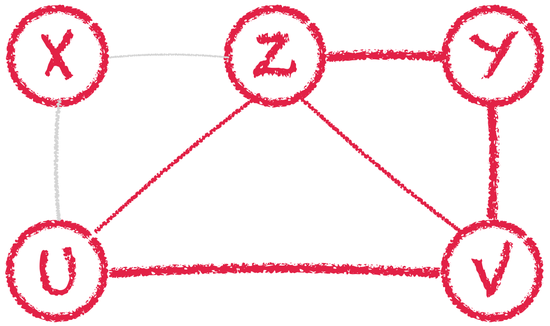
example: five dimensional data (X, Y, Z, U, V) with r_X, r_Y, r_Z, r_U, r_V outcomes

$p =$ model based on empirical distribution $p(x, y, z, u, v)$ with $d(p) = r_X r_Y r_Z r_U r_V - 1$

$q =$ model with listed imposed independence and conditional independence assumptions

examples: $q_1 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$

$q_2 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$ and $Z \perp V \,|\, Y$



divergence tests of goodness of fit

testing nested model specifications

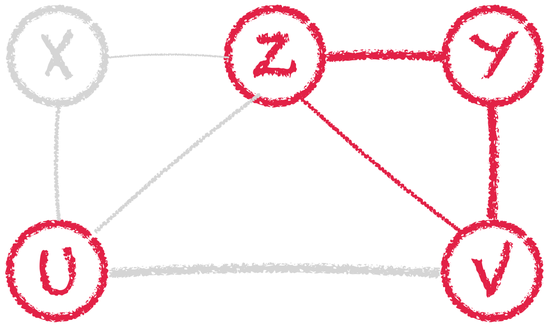
example: five dimensional data (X, Y, Z, U, V) with r_X, r_Y, r_Z, r_U, r_V outcomes

$p =$ model based on empirical distribution $p(x, y, z, u, v)$ with $d(p) = r_X r_Y r_Z r_U r_V - 1$

$q =$ model with listed imposed independence and conditional independence assumptions

examples: $q_1 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$

$q_2 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$ and $Z \perp V \,|\, Y$



divergence tests of goodness of fit

testing nested model specifications

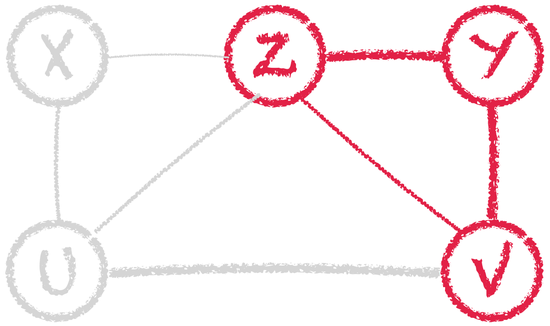
example: five dimensional data (X, Y, Z, U, V) with r_X, r_Y, r_Z, r_U, r_V outcomes

$p =$ model based on empirical distribution $p(x, y, z, u, v)$ with $d(p) = r_X r_Y r_Z r_U r_V - 1$

$q =$ model with listed imposed independence and conditional independence assumptions

examples: $q_1 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$

$q_2 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$ and $Z \perp V \mid Y$



divergence tests of goodness of fit

testing nested model specifications

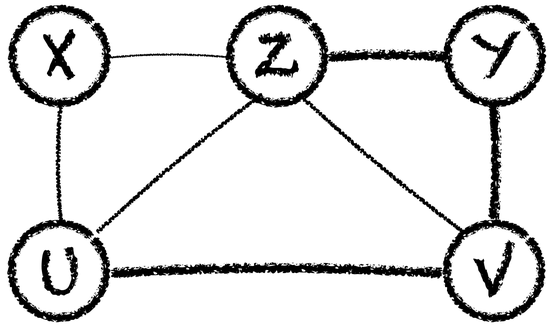
example: five dimensional data (X, Y, Z, U, V) with r_X, r_Y, r_Z, r_U, r_V outcomes

$p =$ model based on empirical distribution $p(x, y, z, u, v)$ with $d(p) = r_X r_Y r_Z r_U r_V - 1$

$q =$ model with listed imposed independence and conditional independence assumptions

examples: $q_1 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$

$q_2 = X \perp (Y, Z, U, V)$ and $U \perp (Y, Z, V)$ and $Z \perp V \,|\, Y$



☑ $d(q)$ is sum of the degrees of freedom of its independent components

☑ divergence $D(p, q)$ of each model is the sums of the divergences of its nested specifications

$$\chi^2(d) = 2nD(p, q_1) = 2n[D(X \perp (Y, Z, U, V)) + D(U \perp Y, Z, V)]$$

$$\chi^2(d) = 2nD(p, q_2) = 2n[D(X \perp (Y, Z, U, V)) + D(U \perp Y, Z, V) + D(Z \perp V \,|\, Y)]$$

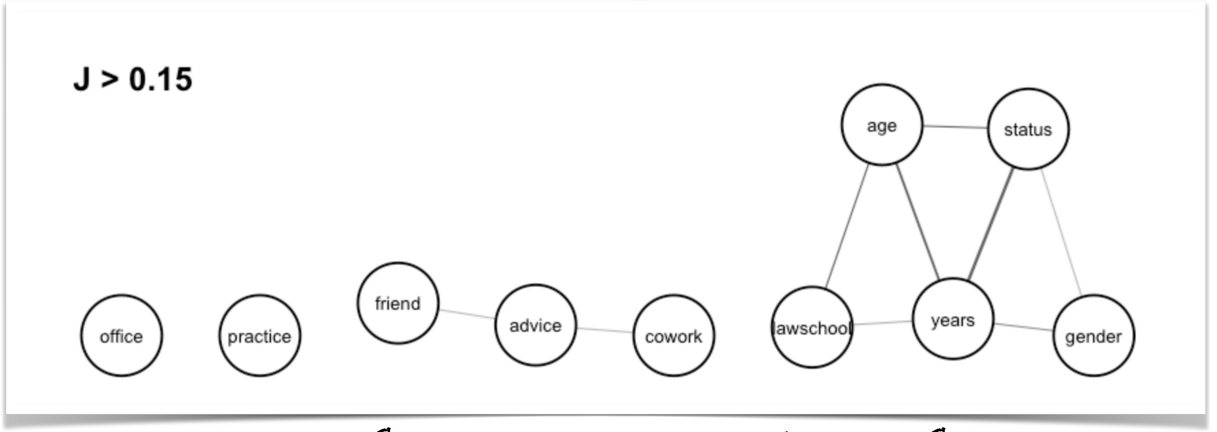
where d can be obtained as either

✓ the sums of degrees of freedom for the divergences of the nested specifications

✓ the difference between degrees of freedom of the general and the specified model $d(p) - d(q)$

example: network study of corporate law firm

☑ divergence tests of goodness of fit: dyad variables



example of structural models of interest:

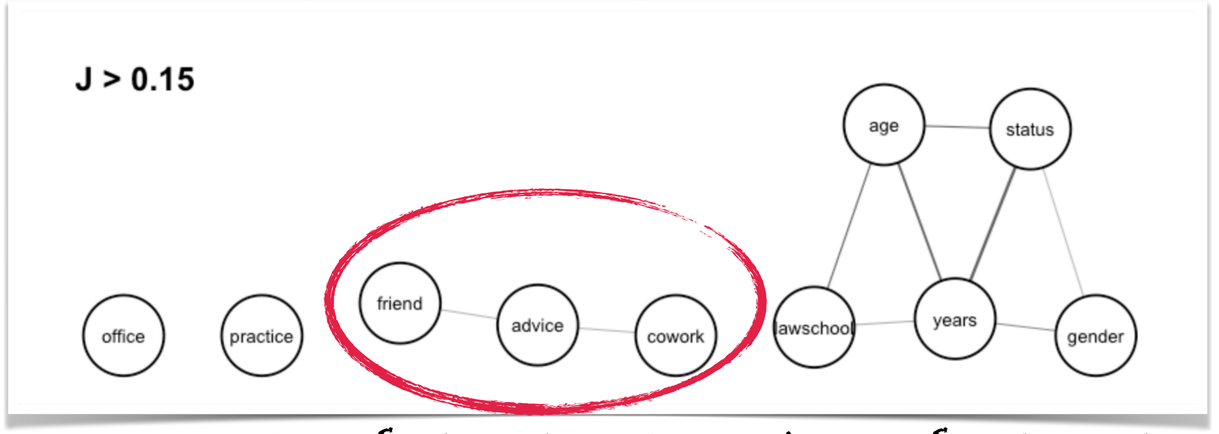
friend ⊥ cowork | advice

gender ⊥ status | years

(friend, cowork. advice) ⊥ (age, status, years)

example: network study of corporate law firm

☑ divergence tests of goodness of fit: dyad variables



example of structural models of interest:

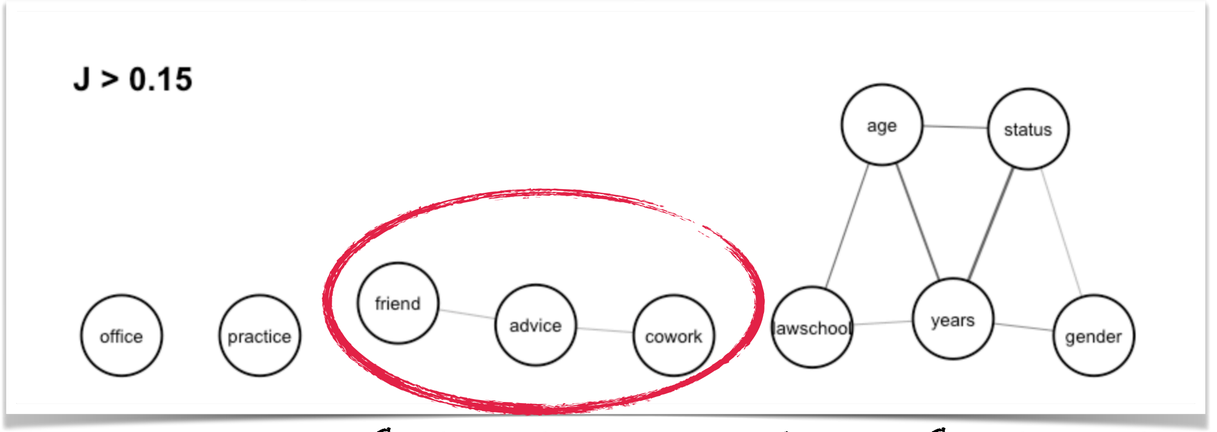
friend ⊥ cowork | advice

gender ⊥ status | years

(friend, cowork. advice) ⊥ (age, status, years)

example: network study of corporate law firm

☑ divergence tests of goodness of fit: dyad variables



example of structural models of interest:

friend ⊥ cowork | advice

gender ⊥ status | years

(friend, cowork. advice) ⊥ (age, status, years)

```
# install development version from GitHub
# install.packages("devtools")
devtools::install_github("termehs/netropy")
```

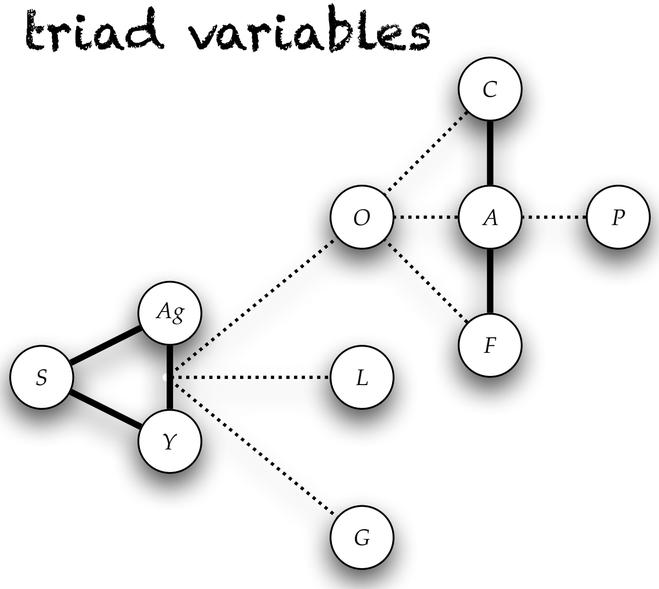
```
div_gof(dat, var1, var2, var_cond = NULL)
```

```
div_gof(dat = dyad.var, var1 = "friend", var2 = "cowork", var_cond = "advice")
```

```
## the specified model of conditional independence cannot be rejected
```

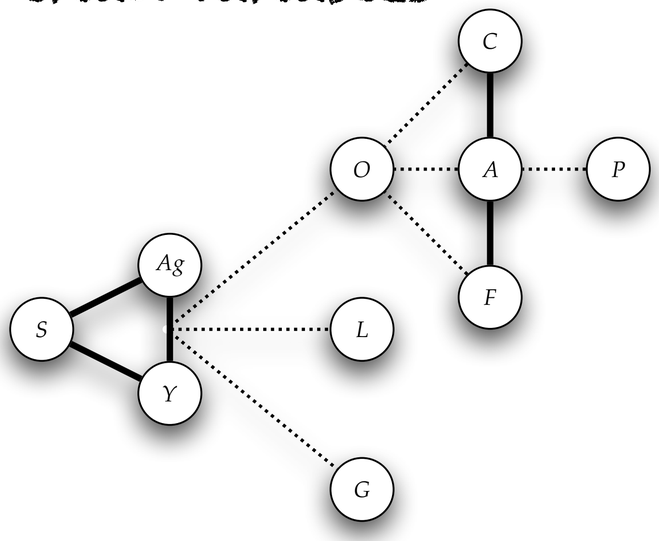
```
##      D df(D)
## 1 0.94    12
```

example: network study of corporate law firm



example: network study of corporate law firm

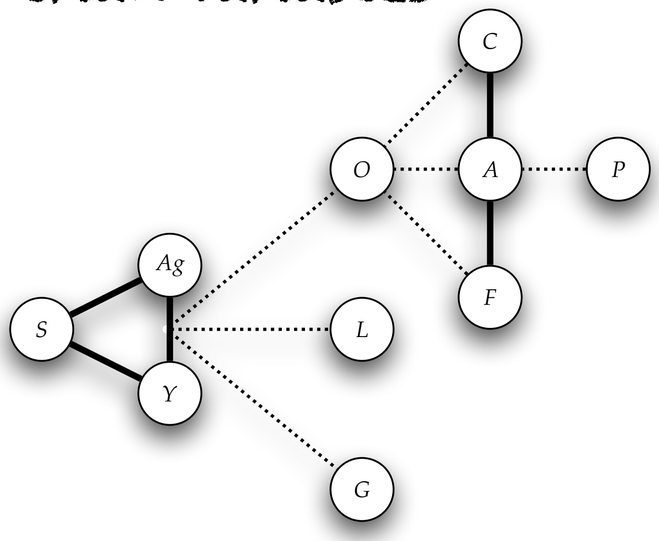
triad variables



dyad independence models
not sufficient to explain triadic behavior

example: network study of corporate law firm

triad variables

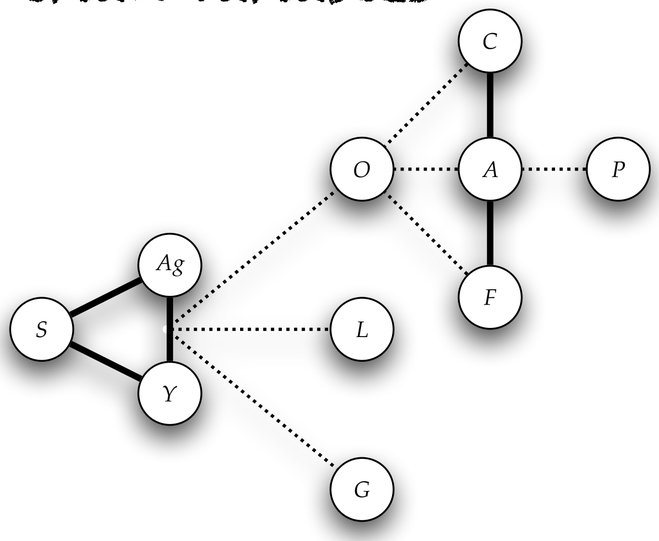


dyad independence models
not sufficient to explain triadic behavior

can attributes of adjacent vertices
explain dyad dependence?

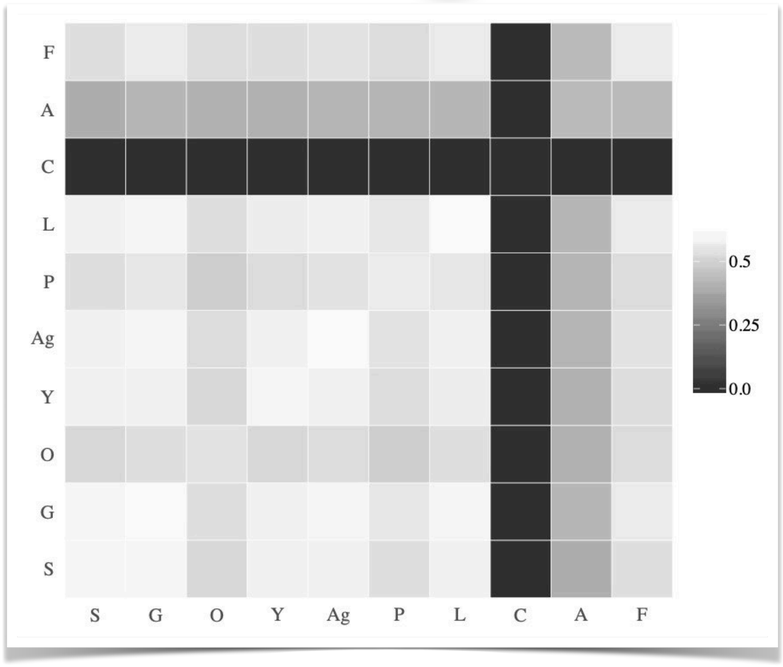
example: network study of corporate law firm

triad variables



dyad independence models
not sufficient to explain triadic behavior

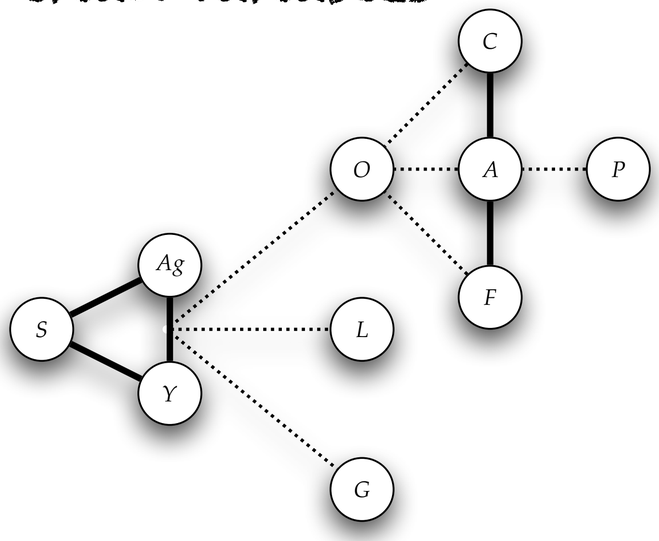
can attributes of adjacent vertices explain dyad dependence?



blockmodel
not sufficient to explain triadic behavior

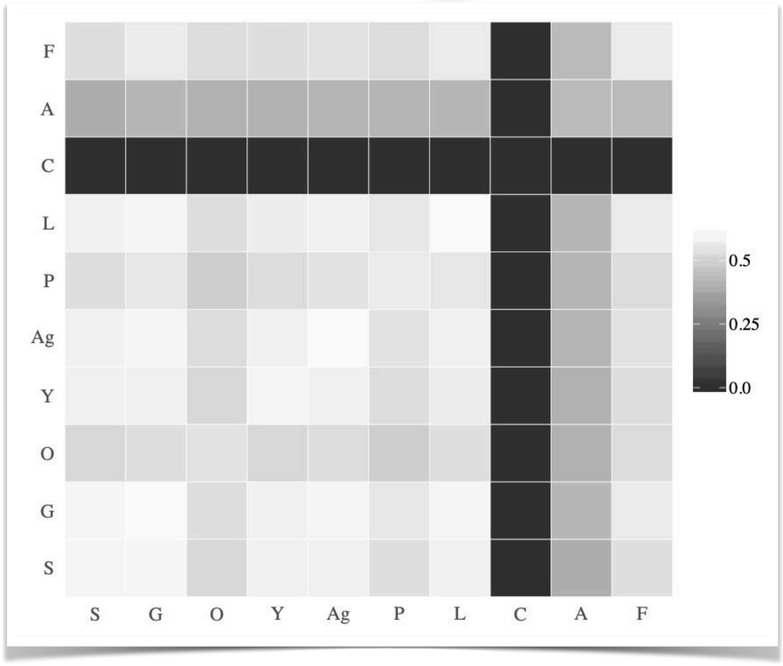
example: network study of corporate law firm

triad variables



dyad independence models
not sufficient to explain triadic behavior

can attributes of adjacent vertices explain dyad dependence?

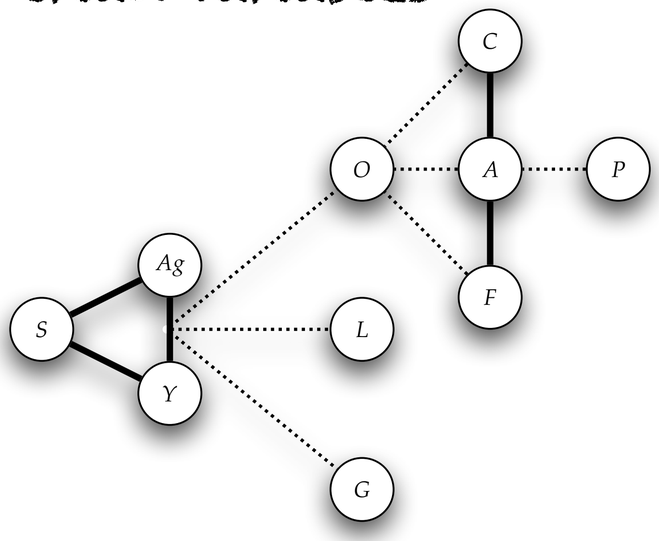


blockmodel
not sufficient to explain triadic behavior

is there a dependence between the different relations?

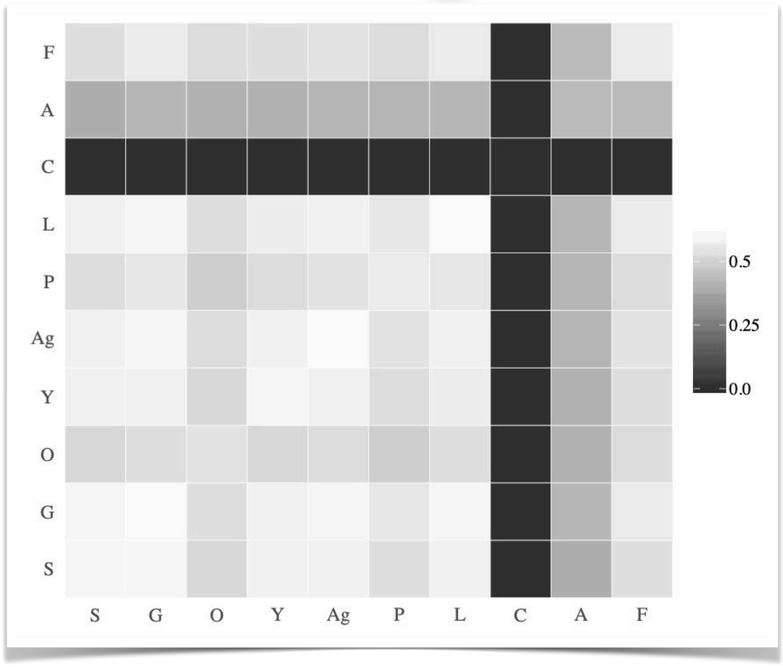
example: network study of corporate law firm

triad variables



dyad independence models
not sufficient to explain triadic behavior

can attributes of adjacent vertices explain dyad dependence?



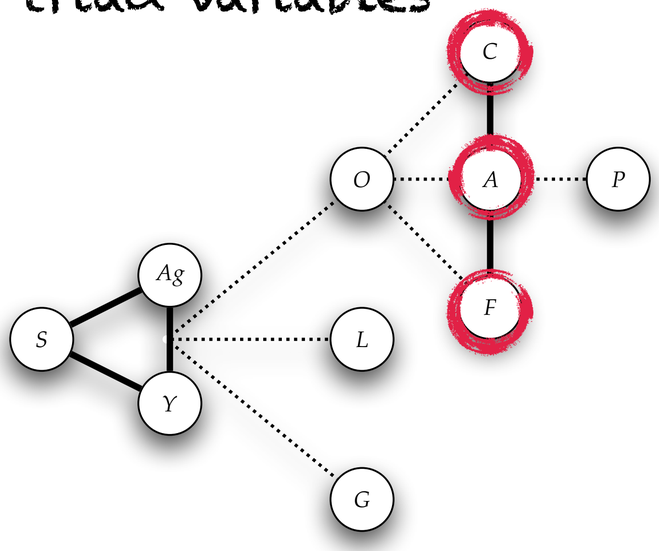
blockmodel
not sufficient to explain triadic behavior

is there a dependence between the different relations?

multiplexity model
almost perfect fit for model $C \perp F \mid A$

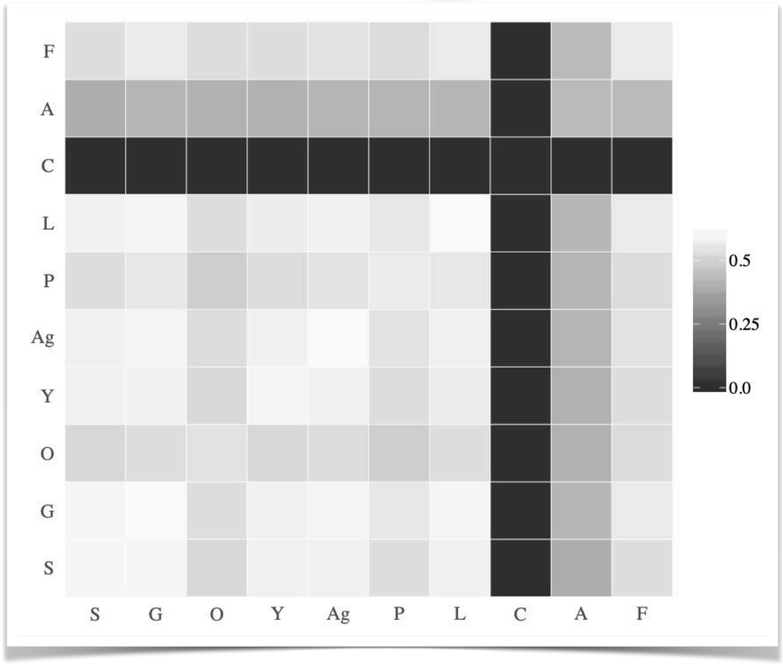
example: network study of corporate law firm

triad variables



dyad independence models
not sufficient to explain triadic behavior

can attributes of adjacent vertices explain dyad dependence?



blockmodel
not sufficient to explain triadic behavior

is there a dependence between the different relations?

multiplexity model
almost perfect fit for model $C \perp F \mid A$

final remarks

information-based screening methods for data editing and variable selection should be more accessible and common in applied statistics

a systematic use of such measures is beneficial
for both exploratory and confirmatory statistical analyses of data
on nominal and ordinal as well as numerical scales

references and resources

☑ Frank, O., & Shafie, T. (2016). Multivariate entropy analysis of network data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 129(1), 45-63.

☑ Nowicki, K., Shafie, T., & Frank, O. (Forthcoming 2022). Statistical Entropy Analysis of Network Data.

☑ Shafie, T. (2022). netropy: Statistical Entropy Analysis of Network Data. R package version 0.1.0, https://CRAN.R-project.org/package=netropy

vignettes available

☑ the R codes to replicate the examples in this presentation are available on my website http://mrs.schochastics.net/