Mean summaries for relational data

567 Statistical analysis of social networks

Peter Hoff

Statistics, University of Washington

Summary statistics

Even in the simplest case of an undirected binary relational data, a sociomatrix can be a complicated and opaque object:

$$\mathbf{Y} = \begin{pmatrix} na & 1 & 0 & 0 & \cdots \\ 1 & na & 0 & 1 & \cdots \\ 0 & 0 & na & 0 & \cdots \\ 0 & 1 & 0 & na & \cdots \\ \vdots & & & & \end{pmatrix}$$

Statistic: A statistic $t(\mathbf{Y})$ is any function of the data.

Descriptive data analysis: A representation of the main features of a dataset via a set of statistics $t_1(\mathbf{Y}), \ldots, t_K(\mathbf{Y})$.

Many important statistics can be computed from the sociomatrix using basic matrix calculations.

Mean

The most basic statistic of a relational dataset is the *mean* or *average*:

Mean: The sum of the relational measurements divided by the number of relational measurements.

For a fully observed directed relation with n nodes:

- the sum of the relational measurements is $\sum_{i \neq j} y_{i,j}$;
- the number of relational measurements is $n \times (n-1)$;
- the mean is

$$\bar{y} = \frac{\sum_{i \neq j} y_{i,j}}{n(n-1)}.$$

For a fully observed undirected relation with n nodes:

- the sum of the relational measurements is $\sum_{i < j} y_{i,j}$;
- the number of relational measurements is $n \times (n-1)/2$;
- the mean is

$$\bar{y} = \frac{\sum_{i < j} y_{i,j}}{n(n-1)/2}.$$

Means via sociomatrices

Means can most easily be computed from sociomatrices.

$$Y_d = \begin{pmatrix} NA & 0 & 2 & 6 \\ 1 & NA & 0 & 0 \\ 0 & 0 & NA & 0 \\ 3 & 0 & 2 & NA \end{pmatrix}$$

First do it “by hand”:

$$\sum_{i \neq j} y_{i,j} = 14$$

$$n(n-1) = 12$$

So the mean is $14/12 \approx 1.67$.

Means via sociomatrices

Now do it “by computer”:

```
Yd

##      [,1] [,2] [,3] [,4]
## [1,]   NA    0    2    6
## [2,]    1   NA    0    0
## [3,]    0    0   NA    0
## [4,]    3    0    2   NA
```

```
sum(Yd)

## [1] NA

sum(Yd,na.rm=TRUE)

## [1] 14
```

```
length(Yd)

## [1] 16

sum(Yd,na.rm=TRUE)/length(Yd) # not correct

## [1] 0.875
```

Means via sociomatrices

Now do it “by computer”:

```
Yd

##      [,1] [,2] [,3] [,4]
## [1,]   NA    0    2    6
## [2,]    1   NA    0    0
## [3,]    0    0   NA    0
## [4,]    3    0    2   NA
```

```
sum(Yd)

## [1] NA

sum(Yd,na.rm=TRUE)

## [1] 14
```

```
sum(!is.na(Yd))

## [1] 12

length(Yd[!is.na(Yd)])

## [1] 12

sum(Yd,na.rm=TRUE)/sum(!is.na(Yd)) # correct

## [1] 1.166667
```

Means via sociomatrices

An easier way:

```
Yd

##      [,1] [,2] [,3] [,4]
## [1,]   NA    0    2    6
## [2,]    1   NA    0    0
## [3,]    0    0   NA    0
## [4,]    3    0    2   NA
```

```
mean(Yd)

## [1] NA

mean(Yd,na.rm=TRUE)

## [1] 1.166667
```

Means via sociomatrices: The undirected case

$$Y_u = \begin{pmatrix} NA & 0 & 2 & 4 \\ 0 & NA & 0 & 0 \\ 2 & 0 & NA & 3 \\ 4 & 0 & 3 & NA \end{pmatrix}$$

First do it "by hand":

$$\sum_{i<j} y_{i,j} = 9$$

$$n(n-1)/2 = 6$$

So the mean is $9/6 = 1.5$.

Means via sociomatrices: The undirected case

Now do it “by computer”:

```
Yu

##      [,1] [,2] [,3] [,4]
## [1,]   NA    0    2    4
## [2,]    0   NA    0    0
## [3,]    2    0   NA    3
## [4,]    4    0    3   NA
```

```
mean(Yu,na.rm=TRUE)

## [1] 1.5
```

How did R do the calculation?

- Did it know that the relation was undirected?
- Did it just use the same procedure as for the directed case?

Means via sociomatrices

For an undirected relation,

$$\begin{aligned}\text{mean of the relation} &= \text{mean of the ``upper triangle'' of the sociomatrix}\\ &= \text{mean of the ``lower triangle'' of the sociomatrix}\\ &= \text{mean of the sociomatrix}\end{aligned}$$

So for either directed or undirected relations,

$$\bar{y} = \text{average of the non-missing values of the sociomatrix}$$

Means via edgelists

Directed relation

```
Ed<-sm2el(Yd)

Ed

##      row col w
## [1,]   1   3 2
## [2,]   1   4 6
## [3,]   2   1 1
## [4,]   4   1 3
## [5,]   4   3 2
```

How can we compute the mean?

```
sum( Ed[,3] )/ ( 4*3 )

## [1] 1.166667
```

Means via edgelists

Undirected relation

```
Eu<-sm2el(Yu,directed=FALSE)

Eu

##      row col w
## [1,]   1   3 2
## [2,]   1   4 4
## [3,]   3   4 3
```

How can we compute the mean?

```
sum( Eu[,3] )/ ( 4*3/2 )

## [1] 1.5
```

When using an edgelist, you need to use the formula and be aware if the relation is directed or undirected.

Additionally, it is more difficult to account for missing data with edgelists.

Densities of graphs

Density: the proportion of edges present in a graph.
$=$ (the number of edges)/(the maximum possible number of edges)

The number of edges observed is $|\mathcal{E}|$.
The number of possible edges is

- $n(n-1)$ in a directed graph;
- $n(n-1)/2$ in an undirected graph.

Exercise: Derive the above numbers.

The density is therefore given by

$$|\mathcal{E}|/[n(n-1)]\text{for a directed graph}$$
$$|\mathcal{E}|/[n(n-1)/2]\text{for an undirected graph}$$

Density as an average

Let $y_{i,j}$ be the binary indicator of an edge from i to j.

Then

$$\begin{aligned} |\mathcal{E}| &= \sum_{i<j} y_{i,j} \text{ for an undirected graph} \\ |\mathcal{E}| &= \sum_{i\neq j} y_{i,j} \text{ for a directed graph,} \end{aligned}$$

and so the density of a graph (undirected or directed) is the same the mean of $y_{i,j}$, i.e., the mean of the corresponding sociomatrix.

Computing the density in R

Directed graph

```
Y

##      [,1] [,2] [,3] [,4] [,5]
## [1,]   NA    0    1    0    0
## [2,]    0   NA    0    1    0
## [3,]    0    0   NA    0    0
## [4,]    0    0    1   NA    0
## [5,]    1    1    0    1   NA

sum(Y,na.rm=TRUE)

## [1] 6

nrow(Y)

## [1] 5

sum(Y,na.rm=TRUE) / ( nrow(Y)*(nrow(Y)-1) )

## [1] 0.3

mean(Y,na.rm=TRUE)

## [1] 0.3
```

Computing the density in R

Undirected graph

```
Y

##      [,1] [,2] [,3] [,4] [,5]
## [1,]   NA    0    0    1    0
## [2,]    0   NA    0    1    0
## [3,]    0    0   NA    0    1
## [4,]    1    1    0   NA    0
## [5,]    0    0    1    0   NA

sum(Y[upper.tri(Y)] )

## [1] 3

sum( Y[upper.tri(Y)] )/( nrow(Y)*(nrow(Y)-1)/2 )

## [1] 0.3

mean(Y,na.rm=TRUE)

## [1] 0.3
```

Densities as probability estimates

Densities such as these can be viewed as

- probabilities of the existence of a tie between randomly sampled nodes, or
- estimates of these probabilities,

depending on the context and your perspective.

Let

- i and j be two randomly sampled individuals;
- let θ be the probability that $y_{i,j} = 1$.

$$\Pr(y_{i,j} = 1) = \theta$$

Then

$\bar{y} = \theta$ if your nodeset is the entire population of nodes.

$\bar{y} = \hat{\theta}$ if your nodeset is a random sample of nodes.

Row and column means

The mean $\bar{y}$ is a very coarse description of a relational dataset.

It is a **global statistic:** It doesn't give you infomation about particular nodes.

This is particularly the case in the presence of nodal heterogeneity, e.g.

- some nodes are more outgoing/send more ties , and/or
- some nodes are more popular/receive more ties.

In such cases, we may also want to compute **nodal statistics** such as the row means and column means.

Grand, row and column means

Let $y_{i,j}$ be the (possibly valued) relation from i to j.

Grand mean: the mean of all non-missing observations.

$$\bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/[n(n-1)] = \sum_{i \neq j} y_{i,j}/[n(n-1)]$$

Row mean: the mean of all non-missing observations in each row

$$\bar{y}_{i\cdot} = y_{i\cdot}/(n-1) = \sum_{j:j \neq i} y_{i,j}/(n-1)$$

Column mean: the mean of all non-missing observations in each column

$$\bar{y}_{\cdot j} = y_{\cdot j}/(n-1) = \sum_{i:i \neq j} y_{i,j}/(n-1)$$

Comtrade example

Yearly trade growth: log change in dollars (2000).

- 30 different countries;
- 10 years from 1996-2005;
- 6 different commodity classes.

```
dimnames(comtrade)[c(1,3,4)]

## [[1]]
##  [1] "Australia"            "Austria"              "Brazil"
##  [4] "Canada"               "China"                "China, Hong Kong SAR"
##  [7] "Czech Rep."           "Denmark"              "Finland"
## [10] "France"               "Germany"              "Greece"
## [13] "Indonesia"            "Ireland"              "Italy"
## [16] "Japan"                "Malaysia"             "Mexico"
## [19] "Netherlands"          "New Zealand"          "Norway"
## [22] "Rep. of Korea"        "Singapore"            "Spain"
## [25] "Sweden"               "Switzerland"          "Thailand"
## [28] "Turkey"               "United Kingdom"       "USA"
##
## [[2]]
## [1] "Chemicals"
## [2] "Crude materials, inedible, except fuels"
## [3] "Food and live animals"
## [4] "Machinery and transport equipment"
## [5] "Manufact goods classified chiefly by material"
## [6] "Miscellaneous manufactured articles"
##
## [[3]]
##  [1] "1996" "1997" "1998" "1999" "2000" "2001" "2002" "2003" "2004" "2005"
```

Comtrade example

Compute 10-year mean increase in manufactured goods:

```
Y<-apply(comtrade[,,c(5,6),],c(1,2),mean)

dim(Y)

## [1] 30 30

round( Y[1:5,1:5] ,2 )

##           Australia Austria Brazil Canada China
## Australia        NA    0.10   0.08   0.03  0.08
## Austria        0.08      NA   0.06   0.06  0.09
## Brazil        -0.06    0.03     NA   0.07  0.14
## Canada         0.00    0.05  -0.03     NA  0.10
## China          0.13    0.12   0.14   0.16    NA
```

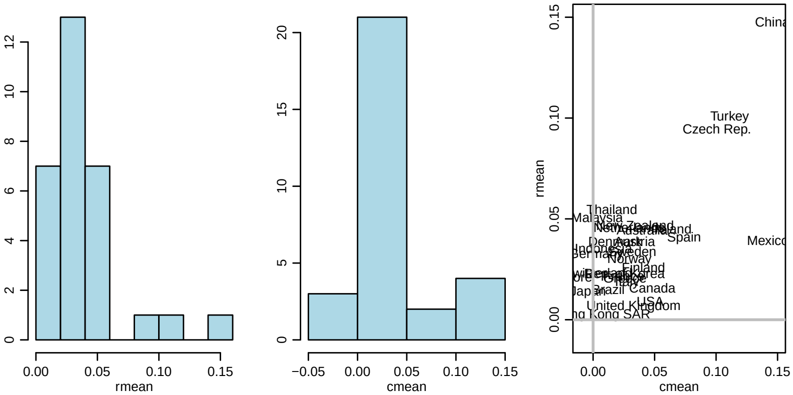
International trade data

```
mean(Y,na.rm=TRUE)

## [1] 0.03778362

rmean<-rowMeans(Y,na.rm=TRUE)
cmean<-colMeans(Y,na.rm=TRUE)
```

```
mean(rmean) ; sd(rmean)

## [1] 0.03778362
## [1] 0.03019967

mean(cmean) ; sd(cmean)

## [1] 0.03778362
## [1] 0.04101555

cor(rmean,cmean)

## [1] 0.7002526
```

Exercise: Derive the fact that the mean of the row means is the overall mean.

Comtrade example



Mean statistics

Grand mean:

- a **global statistic**;
- an across-dyad average of the relation.

Row and column means:

- are **nodal statistics**;
- characterize differing levels of network activity across nodes.

Summarizing heterogeneity in row and column means:

- Univariate:
  - histograms of row and column means;
  - standard deviations of row mean vector, of column mean vector;
- Bivariate:
  - bivariate scatterplots;
  - correlation of row mean and column mean.

Additive decomposition of a socimatrix

We've discussed summarizing the socimatrix with a variety of means

- $\bar{y}_{\cdot\cdot}$ to represent the overall mean or density;
- $\{\bar{y}_{i\cdot}, i = 1, \ldots, n\}$ to represent row heterogeneity;
- $\{\bar{y}_{\cdot j}, j = 1, \ldots, n\}$ to represent column heterogeneity.

Let $\hat{a}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$. Then

- $\hat{a}_i$ above/below 0 $\iff$ $\bar{y}_{i\cdot}$ is above/below the average.
- $\hat{a}_i$ is a centered measure of **sociability**;
- $\sum_i \hat{a}_i = 0$.

Similarly, we can define $\hat{b}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$ as a centered measure of **popularity**.

Additive decomposition of a socimatrix

Convince yourself that the following is true:

$$y_{i,j} = \hat{\mu} + \hat{a}_i + \hat{b}_j + \hat{\epsilon}_{i,j}$$

where

- $\hat{\mu} = \bar{y}_{\cdot\cdot}$ and
- $\hat{\epsilon}_{i,j} = (y_{i,j} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})$.

Exercise: Derive this identity.

Exercise: Obtain the value of $\sum_{i \neq j} \hat{\epsilon}_{i,j}$.

Additive decomposition of a socimatrix

ANOVA decomposition

$$y_{i,j} = \hat{\mu} + \hat{a}_i + \hat{b}_j + \hat{\epsilon}_{i,j}$$

This is simply the additive ANOVA decomposition of a (socio)matrix.

Grand mean: $\hat{\mu} = \bar{y}_{\cdot\cdot}$ is the overall (grand) mean of the relation;

Variance components: $\hat{a}_i$'s,$\hat{b}_j$'s,$\hat{\epsilon}_{i,j}$'s represent variability of $y_{i,j}$'s around $\hat{\mu}$:

- $\{(\hat{a}_i, \hat{b}_i) : i = 1, \ldots, n\}$ describe **additive** components of variability;
- $\{\hat{\epsilon}_{i,j} : i \neq j\}$ represents non-additive variability.

Additive model:

$$y_{i,j} \approx \hat{\mu} + \hat{a}_i + \hat{b}_j$$

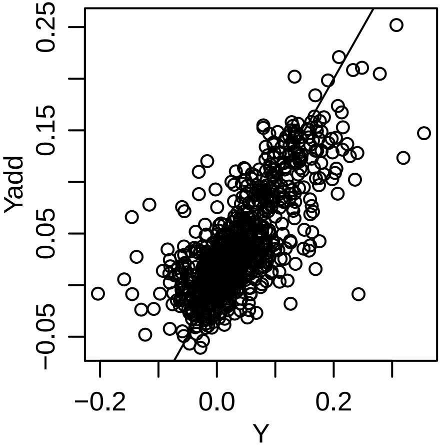
- A good model if variability is approximately additive/$\hat{\epsilon}_{i,j}$'s are small or patternless.
- A bad model otherwise.

Example: International trade growth

```
Y<- apply( comtrade[,,5:6,],c(1,2),mean)

mu<-mean(Y,na.rm=TRUE)
a<-rowMeans(Y,na.rm=TRUE) - mu
b<-colMeans(Y,na.rm=TRUE) - mu

mean(a)

## [1] 1.618736e-18

mean(b)

## [1] 2.313501e-18
```

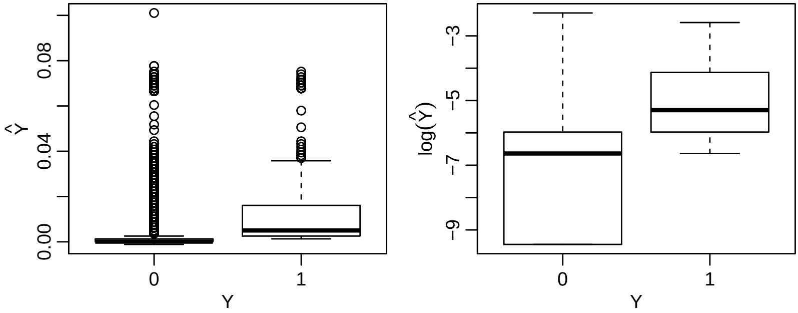
Example: International trade growth

```
Yadd<- mu + outer(a,b,"+")

mean( (Y-Yadd)^2,na.rm=TRUE ) / mean( (Y - mu)^2,na.rm=TRUE )

## [1] 0.4369656
```



Example: Protein interactions

```r
Y<- yeast

mu<-mean(Y,na.rm=TRUE)
a<-rowMeans(Y,na.rm=TRUE) - mu
b<-colMeans(Y,na.rm=TRUE) - mu


Yadd<- mu + outer(a,b,"+")

mean( (Y-Yadd)^2,na.rm=TRUE ) / mean( (Y - mu)^2,na.rm=TRUE )

## [1] 0.9880097
```



Degrees for binary relations

For binary relations, nodal heterogeneity can be described by **nodal degrees**.

- Undirected relation:
  - The **degree** of a node is the node's number of ties.
- Directed relation:
  - The **outdegree** of a node is the node's number of outgoing ties.
  - The **indegree** of a node is the node's number of incoming ties.

The degees are easy to calculate from the sociomatrix $\mathbf{Y} = \{y_{i,j} : i \neq j\}$:

$$d_i^o = \sum_{j:j\neq i} y_{i,j} \quad , \quad d_i^i = \sum_{j:j\neq i} y_{j,i}$$

This calculation works for both directed and undirected relations.
Specifically, for an undirected relation,

$$\begin{aligned} d_i^o &= \sum_{j:j\neq i} y_{i,j} \\ &= \sum_{j:j\neq i} y_{j,i} = d_i^i = d_i \end{aligned}$$

Nodal degree

$$\mathbf{Y} = \begin{pmatrix} na & 0 & 1 & 1 & 0 & 1 \\ 1 & na & 1 & 0 & 0 & 1 \\ 0 & 0 & na & 1 & 0 & 1 \\ 0 & 0 & 1 & na & 0 & 1 \\ 1 & 0 & 1 & 1 & na & 1 \\ 0 & 0 & 1 & 1 & 0 & na \end{pmatrix}$$

$$d_4^o = \sum_{j:j\neq 4} y_{4,j} = 2$$

$$d_4^i = \sum_{i:i\neq 4} y_{i,4} = 4$$

Nodal degree

For an undirected relation:

$$\mathbf{Y} = \begin{pmatrix} na & 0 & 1 & 1 & 0 & 1 \\ 0 & na & 0 & 0 & 0 & 1 \\ 1 & 0 & na & 1 & 1 & 1 \\ 1 & 0 & 1 & na & 0 & 1 \\ 0 & 0 & 1 & 0 & na & 0 \\ 1 & 1 & 1 & 1 & 0 & na \end{pmatrix}$$

$$\begin{aligned} d_4 = d_4^o &= \sum_{j:j\neq 4} y_{4,j} = 3 \\ &= d_4^i = \sum_{i:i\neq 4} y_{i,4} = 3 \end{aligned}$$

Degrees and density

Recall that the formula for the density of a graph, directed or undirected, is

$$
\begin{aligned}
\bar{y} &= \frac{1}{n(n-1)} \sum_{i \neq j} y_{i,j} \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j:j \neq i} y_{i,j} \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} d_i^o = \bar{d}^o/(n-1),
\end{aligned}
$$

and so the average degree is $n-1$ times the density.

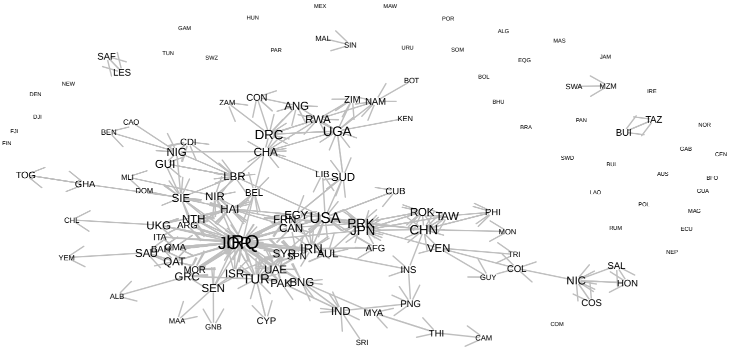
A similar calculation shows that $\bar{y} = \bar{d}^i/(n-1)$. Thus

- the average indegree equals the average outdegree;
- the average degree equals $n-1$ times the density.

Example: 1990-2000 international conflict

$y_{i,j}$ is the binary indicator that i initiated an action against j.

```r
Y<-1*( conflict90s$conflict > 0 )    # dichotomize the data
```



Computing degrees in R

```
odeg<-rowSums(Y,na.rm=TRUE)
ideg<-colSums(Y,na.rm=TRUE)


odeg[1:10]

## AFG ALB ALG ANG ARG AUL AUS BAH BEL BEN
##   1   0   0   2   1   1   0   1   0   0

ideg[1:10]

## AFG ALB ALG ANG ARG AUL AUS BAH BEL BEN
##   2   1   0   3   2   3   0   2   3   1
```

Degree distributions

For an undirected relation, the set of degrees is an $n \times 2$ matrix.

It is generally desirable to summarize the data further

This can be done by summarizing the **joint degree distribution**:

- mean degree, standard deviation of in- and outdegrees
- correlation of in- and outdegrees
- empirical marginal distributions of each set of degrees.

Univariate summaries of degrees

Let $\mathbf{d} = \{d_1, \ldots, d_n\}$ be a set of nodal degrees

(either outdegrees, indegrees, or undirected degrees)

The entries of $\mathbf{d}$ are often summarized with the

- mean: $\bar{d} = \sum d_i/n = (n-1)\bar{y}$,
- variance: $s_d^2 = \sum(d_i - \bar{d})^2/(n-1)$,
- degree distribution.

Degree distribution

The **degree distribution** is a set of counts $\{f_0, \ldots, f_n\}$ where

$$f_k = \#\{d_i = k\} = \text{number of nodes with degree equal to } k$$

For example, if

$$\mathbf{d} = (2, 1, 0, 3, 2, 3, 0, 2, 3, 1)$$

then

$$\mathbf{f} = (2, 2, 3, 3, 0, 0, 0, 0, 0, 0, 0),$$

which we might write more informatively as

$$\mathbf{f} = \left(\begin{array}{ccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hline 2 & 2 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right)$$

Bivariate summaries of degrees

Let $\mathbf{d}^o = (d_1^o, \ldots, d_n^o)$ and $\mathbf{d}^i = (d_1^i, \ldots, d_n^i)$ be vectors of out and indegrees.

The joint distribution of $\mathbf{d}^o$ and $\mathbf{d}^i$ are often described with

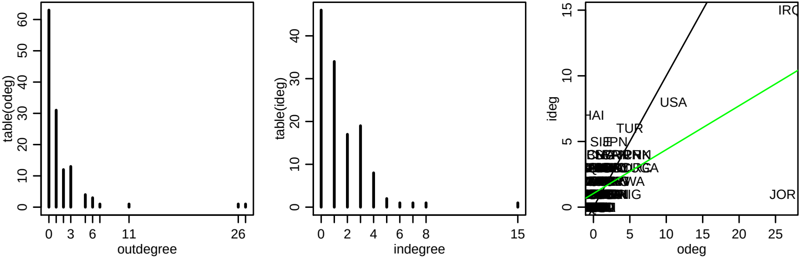
- the correlation between $\mathbf{d}^o$ and $\mathbf{d}^i$
- a scatterplot of $\mathbf{d}^o$ versus $\mathbf{d}^i$.

These are all straightforward to obtain in R.

Example: 1990-2000 international conflict

```
mean(odeg)

## [1] 1.561538

mean(ideg)

## [1] 1.561538

sd(odeg)

## [1] 3.589398

sd(ideg)

## [1] 1.984451

cor(odeg,ideg)

## [1] 0.6040145

table(odeg)

## odeg
##  0  1  2  3  5  6  7 11 26 27
## 63 31 12 13  4  3  1  1  1  1

table(ideg)

## ideg
##  0  1  2  3  4  5  6  7  8 15
## 46 34 17 19  8  2  1  1  1  1
```

Example: 1990-2000 international conflict



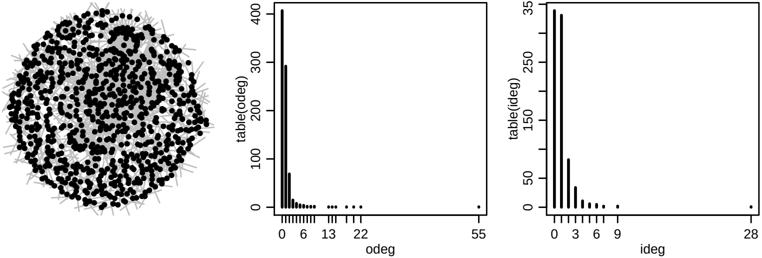
Example: 1990-2000 international conflict

Descriptive degree analysis: For the 1990-2000 conflict data:

- The probability that any pair of countries were in conflict at some point is around 1%. ($\bar{y} = 0.018$).
- Countries were more heterogeneous in terms of initiating conflict than being the target ($sd(\mathbf{d}^o) = 3.59 > 1.98 = sd(\mathbf{d}^i)$).
- Countries that initiatied more conflicts tended to be the target of more conflicts ($cor(\mathbf{d}^o, \mathbf{d}^i) = 0.60$).
- USA, IRQ, JOR, TUR HAI were the most active nodes:
    - JOR has a very high outdegree and a low indegree.
    - HAI has a high indegree and a low outdegree.

More degree distributions

Yeast protein interaction network (n=813)



Degree variation

Note that for the conflict and protein networks,

- most nodes have small degrees,
- few nodes have large degrees.

Recall the degree distribution $\mathbf{f} = \{f(k), k = 0, \ldots, n\}$, where

$$f(k) = f_k = \#\{d_i = k\}.$$

For the two networks above, the degree distribution $f(k)$ is roughly a decreasing function of k.

Power law behavior

Some researchers have posited an explicit form for $f(k)$:

$$f(k) = ak^{-b}, \quad a > 0, b > 0.$$

A distribution for which this (roughly) holds is said to follow a **power law**.

A network (or network model) whose degree distribution follows a power law is said to be **scale free**.

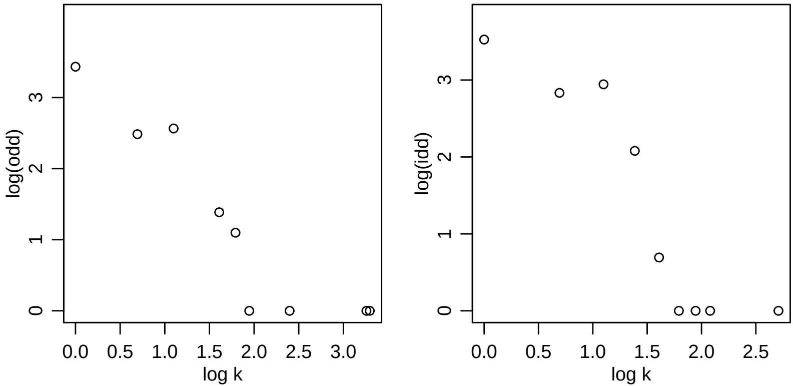
For such a degree distribution,

$$\log f(k) = \log a - b \log k,$$

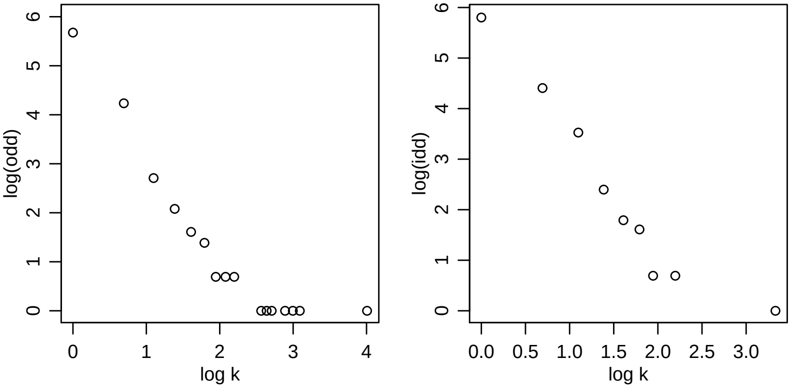
so that the *logged value of* $f(k)$ should be *linearly decreasing in log* k.

This can be checked empirically by plotting the log degree distribution versus k, and assessing whether or not the relationship is linear.

Assessing power law behavior: conflict network



Assessing power law behavior: protein network



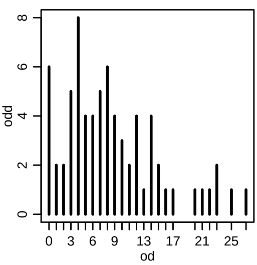
Lawyer friendship network

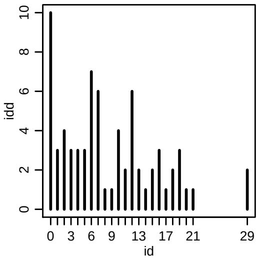
Lazega's law firm data:
Several nodal and dyadic variables measured on 71 attorneys in a law firm.

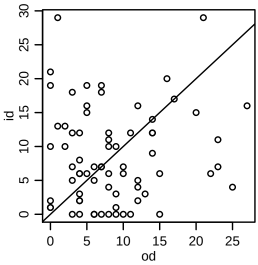
```
dim(lazegalaw$X)

## [1] 71  7

colnames(lazegalaw$X)

## [1] "status"     "female"     "office"     "seniority" "age"        "practice"  
## [7] "school"

dim(lazegalaw$Y)

## [1] 71 71  3

dimnames(lazegalaw$Y)[[3]]

## [1] "advice"      "friendship" "cowork"
```

Lawyer friendship network

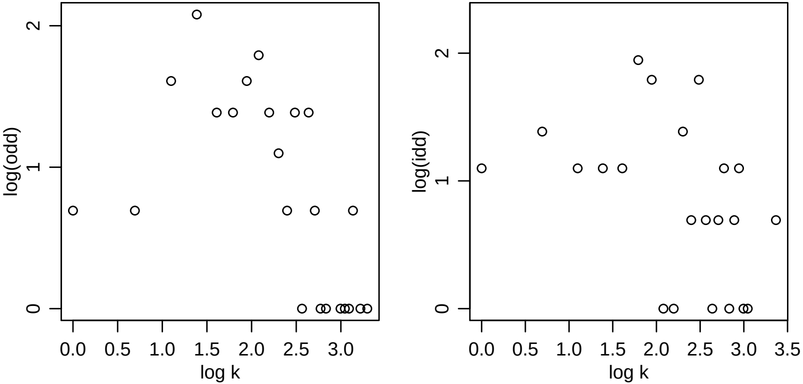
```
advice<-(lazegalaw$Y)[,,1]
od<-rowSums(advice,na.rm=TRUE)
id<-colSums(advice,na.rm=TRUE)

table(od)

## od
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 20 21 22 23 25 27
##  6  2  2  5  8  4  4  5  6  4  3  2  4  1  4  2  1  1  1  1  1  2  1  1

table(id)

## id
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 29
## 10  3  4  3  3  3  7  6  1  1  4  2  6  2  1  2  3  1  2  3  1  1  2
```







Lawyer friendship network



Assessing power law behavior

For the first two networks, the trend is arguably linear, except for large k.

However, the frequencies for large k depend on only a few nodes, so maybe a power law is a reasonable **model** for the degree distributions for these networks.

For the friendship network, the trend is nonlinear.

Implications:

- some very simple models of network formation imply a power law;
- other very simple models imply something other than a power law.

The degree distribution may then help us discriminate between *classes* of models (scale free versus non-scale free).

We will return to this when we discuss **hypothesis testing**.

Summary: grand means, row means, density and degree

Grand means and density:

- The grand mean is the average of all observed relations.
- Density is just a term for the mean when the relations are binary.

Means and degrees:

- The *i*th row mean is the average of the observed relations in row i.
- The outdegree of node i is the
  - total number of outgoing links of node i;
  - the sum of $y_{i,j}$ across $j : j \neq i$.

Therefore, for a completely observed binary relation,

$$\bar{y}_i = \frac{\sum_{j:j\neq i} y_{i,j}}{n-1} = \frac{\text{odeg}_i}{n-1}$$

The row means are the outdegrees divided by $n-1$.

(similarly for column means and indegrees)

Discuss: In the presence of missing data, which do you think would be a better summary, row means or outdegrees?

Means for various data types

Means or sums may not be appropriate for every type of relationship:

- categorical, non-ordinal relationships
  - $y_{i,j} \in \{$ mother, father, sibling, uncle, $\ldots\}$.
  - $y_{i,j} \in \{$ red , blue , green$\}$
- ordinal non-metric relationships
  - $y_{i,j} \in \{$ dislike, neutral, like $\}$
  - $y_{i,j} \in \{$ none, some, many $\}$
- sparse valued data
  - $y_{i,j} = \{$ number of minutes of communication $\}$
  - $y_{i,j} = \{$ number of emails sent $\}$

Means for categorical relations

One strategy for such data is to decompose the relation:

$y_{i,j} \in \{$ red , blue , green $\}$

- $y_{i,j,r} = 1 \times (y_{i,j} =$ red $)$.
- $y_{i,j,b} = 1 \times (y_{i,j} =$ blue $)$.
- $y_{i,j,g} = 1 \times (y_{i,j} =$ green $)$.

Define $\tilde{y}_{i,j} = y_{i,j,r} + y_{i,j,b} + y_{i,j,g}$,
i.e. $\tilde{y}_{i,j}$ indicates the presence of any relationship.

- Grand mean: $\bar{\tilde{y}}_{\cdot\cdot} = \bar{y}_{\cdot\cdot r} + \bar{y}_{\cdot\cdot b} + \bar{y}_{\cdot\cdot g}$.
- Row means: $\bar{\tilde{y}}_{i\cdot} = \bar{y}_{i\cdot r} + \bar{y}_{i\cdot b} + \bar{y}_{i\cdot g}$.
- Column means: $\bar{\tilde{y}}_{\cdot j} = \bar{y}_{\cdot jr} + \bar{y}_{\cdot jb} + \bar{y}_{\cdot jg}$.

Conditional means

If $y_{i,j}$ is valued but sparse, it can be useful to decompose $y_{i,j}$ as follows:

$$x_{i,j} = \begin{cases} 0 & \text{if } y_{i,j} = 0 \\ 1 & \text{if } y_{i,j} \neq 0 \end{cases} \qquad w_{i,j} = \begin{cases} NA & \text{if } y_{i,j} = 0 \\ y_{i,j} & \text{if } y_{i,j} \neq 0 \end{cases}$$

$x_{i,j}$ can be analyzed as with a binary relation:

- density, out and indegrees
- grand, row and column means

$w_{i,j}$ can be analyzed with means, but the interpretation is subtle:

- $\bar{w}_{\cdot\cdot}$ is the mean of non-zero relations;
- $\bar{w}_{i\cdot}$ is the mean of i's non-zero outgoing relations;
- $\bar{w}_{\cdot j}$ is the mean of j's non-zero incoming relations.

Summary

- Grand and nodal means are a starting point for relational data analysis:
  - represent the overall level of relations and heterogeneity among the nodes;
  - correspond to the well-known ANOVA decompostion of two-way data;
  - for binary data, they are equivalent to density, outdegree and indegree.
- Nodal Heterogeneity can be explored with row and column means:
  - standard deviations, histograms or tables of means or degrees;
  - correlations and scatterplots of row versus column means or degrees.
- Modifcations may be necessary for different data types:
  - non-binary categorical relations;
  - sparse, valued relations.