Fundamental Statistics of Graphs

Mark S. Handcock

handcock@ucla.edu

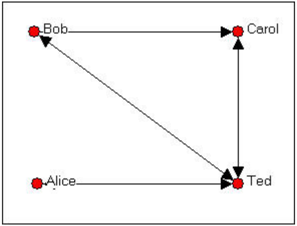
Statistical Analysis of Networks
Stat 218

Matrix form

- The adjacency matrix is often called a *sociomatrix*
- Symmetric in undirected case
- Diagonals represent self-ties, and are often treated as undefined

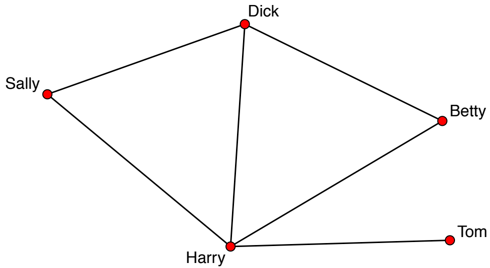
$$\mathsf{Y} = \begin{bmatrix} \mathrm{NA} & y_{12} & y_{13} & \ldots y_{1n} \\ y_{21} & \mathrm{NA} & y_{23} & \ldots y_{2n} \\ y_{31} & y_{32} & \mathrm{NA} & \ldots y_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \ldots & \mathrm{NA} \end{bmatrix}$$

An Example of a Directed Graph



	Bob	Carol	Ted	Alice
Bob	-	1	1	0
Carol	0	-	1	0
Ted	1	1	-	1
Alice	0	0	1	-

An Example of a Undirected Graph



	Betty	Dick	Harry	Sally	Tom
Betty	-	1	1	0	0
Dick	1	-	1	1	0
Harry	1	1	-	1	1
Sally	0	1	1	-	0
Tom	0	0	1	0	-

Summarizing Graph structure

How do we summarize a n node undirected graph?

- It has $n(n-1)/2$ values.
- If they were measured on disjoint units (monads) we would try standard measures
  - sum, mean, median, quartiles, IQR, quantiles, ...
- How do we take into account the *dyadic* structure?
- Start by considering the mean

Summarizing Graph structure

- The *density* of a graph is the mean of the tie values
  - number of ties divided by the number of possible ties

$$\bar{Y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{ij} = \frac{2}{n(n-1)} \sum_{i<j} y_{ij} \quad \text{undirected}$$

$$\bar{Y} = \frac{1}{n(n-1)} \sum_{i,j} y_{ij} \quad \text{directed}$$

Advantages of the mean / density

- The *density* treats all tie values equally.
- The density treats all nodes equally.
- It is a global statistic
    - a *graph statistics* is any function of Y, $g(Y)$.
- It is a global measure of the *sociality* of the graph
- It is an average measure of the sociality of the nodes in the graph

Summarizing the sociality of individual nodes

- Focus on *node level* summarizing
  - The *density* of a node is the mean of the tie values
    - number of its ties divided by the number of possible ties
- measures the sociability of a node.
- some nodes are more social than others

$$\bar{Y}_i = \frac{1}{n-1} \sum_{j:j \neq i} y_{ij} \qquad \text{out-density or undirected}$$

Summarizing the sociality of individual nodes

- Similar ideas for directed networks
- some nodes are more outgoing/send more ties
- some nodes are more popular/receive more ties

$$\bar{Y}_i^{rmo} = \frac{1}{n-1} \sum_{j:j \neq i} y_{ij} \quad \text{directed out}$$

$$\bar{Y}_j^{rmi} = \frac{1}{n-1} \sum_{i:j \neq i} y_{ij} \quad \text{directed in}$$

Summarizing the sociality of networks

- We can summarize the network *heterogeneity* in sociality
  - Standard deviation, IQR of $\bar{Y}_i$, $\bar{Y}^{\mathrm{i}}_j$, or $\bar{Y}^{\mathrm{o}}_j$.
  - Histograms of $\bar{Y}_i$, $\bar{Y}^{\mathrm{i}}_j$, or $\bar{Y}^{\mathrm{o}}_j$.
  - Correlation of $\bar{Y}^{\mathrm{i}}_j$, or $\bar{Y}^{\mathrm{o}}_j$.
  - Scatter plots of $\bar{Y}^{\mathrm{i}}_j$, or $\bar{Y}^{\mathrm{o}}_j$.

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
```

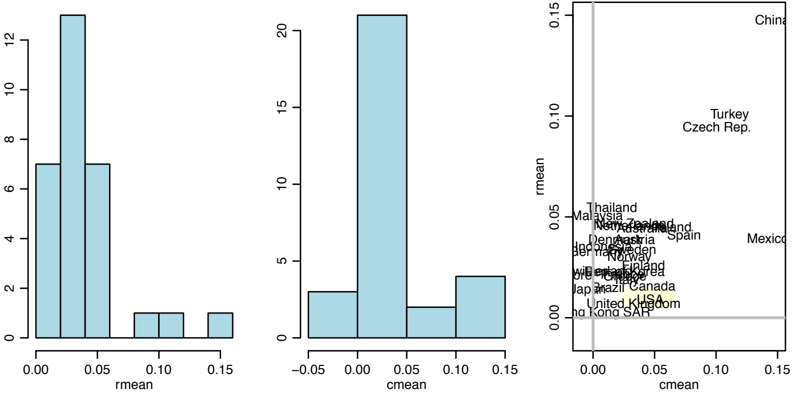
International trade data

```
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



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.

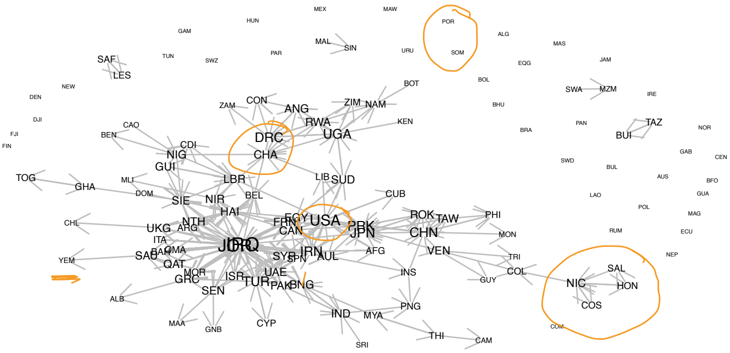
A similar calculation shows that $\bar{y} = \bar{d}^i/(n-1)$. Thus

- 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.

```
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

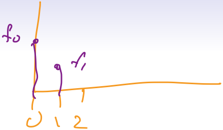
Let $\mathbf{d} = \{d_1, \ldots, d_n\}$ be a set of nodal degrees

(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

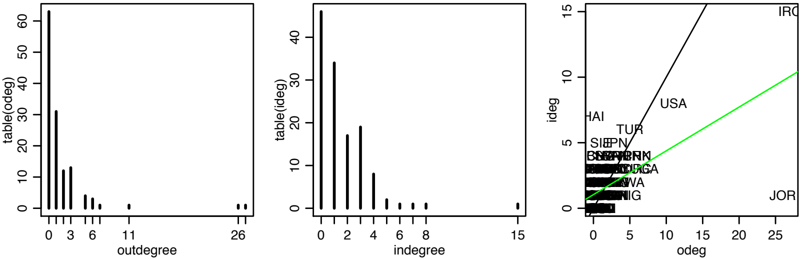
- the correlation between $\mathbf{d}^o$ and $\mathbf{d}^i$
- a scatterplot of $\mathbf{d}^o$ versus $\mathbf{d}^i$.

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

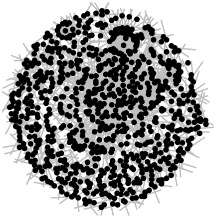


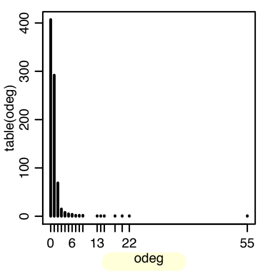
Example: 1990-2000 international conflict

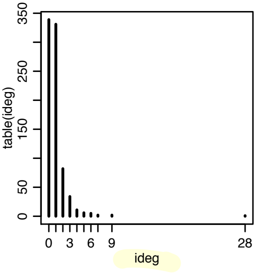
Descriptive degree analysis: For the 1990-2000 conflict data:

- 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.

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**.

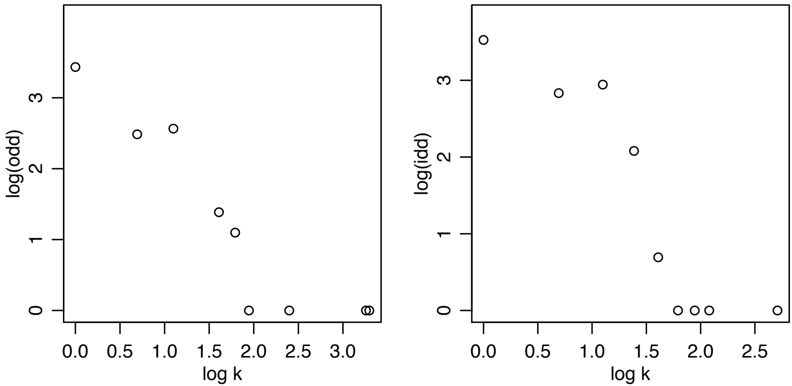
For such a degree distribution,

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

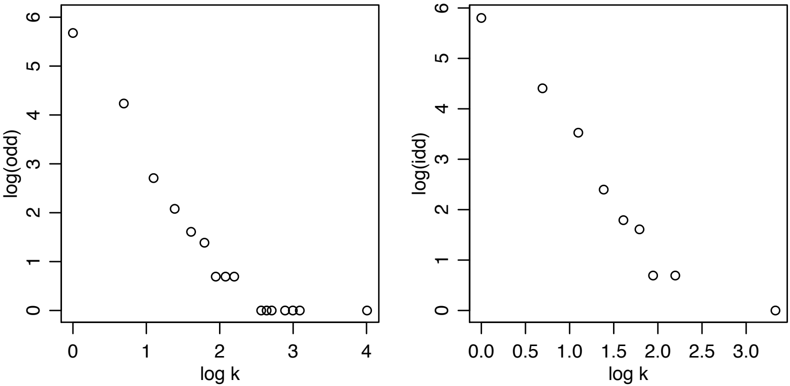
so that the *logged value of* $f(k)$ should be *linearly decreasing in log* k.

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



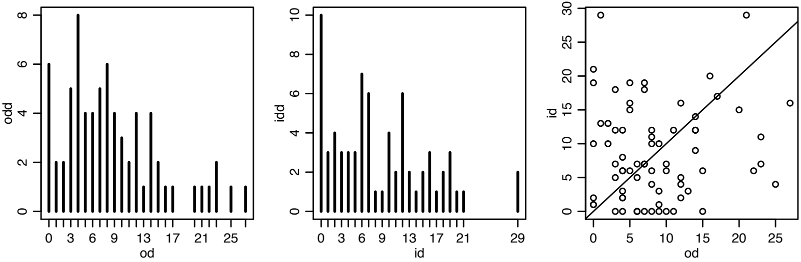
Lawyer friendship network

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

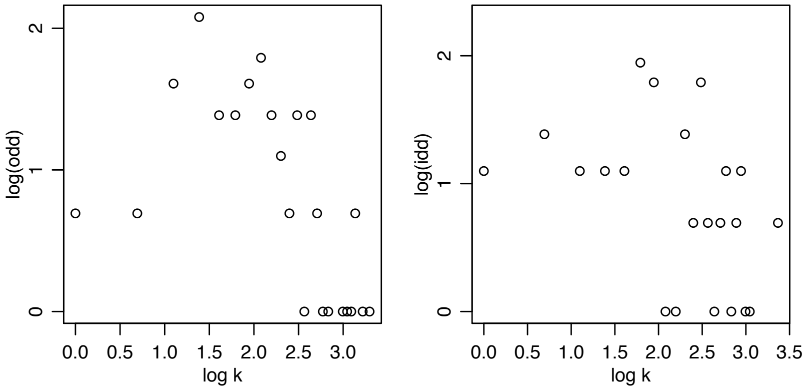
```
dim(lazegalaw$X)

## [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 ith 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.