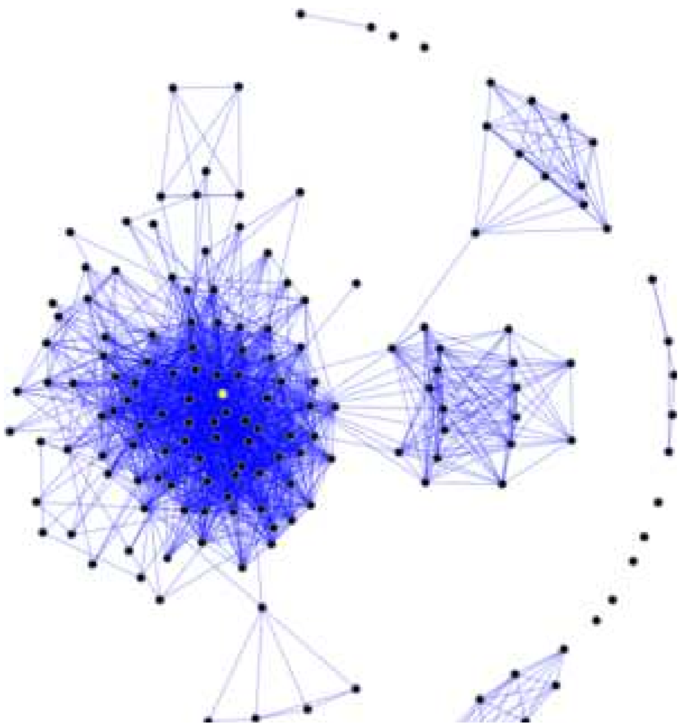
# SOCIAL NETWORK ANALYSIS USING STATA



19 February 2015

Institute for Analytical Sociology

Linköping University

Thomas Grund

thomas.u.grund@gmail.com

Wi Fi

# http://nwcommands.org

**/TUTORIALS AND SLIDES**


KEEP CALM AND CARRY ON
STAY CALM AND THINK STRAIGHT
KEEP CALM AND CALL BATMAN
KEEP CALM AND DRINK BEER
KEEP CALM AND DRINK TEA
KEEP CALM AND RUN LIKE YOU'RE IN THE HUNGER GAMES
STAY CALM & QUILT ON
KEEP CALM. NOBODY ELSE KNOWS WHAT THEY'RE DOING EITHER

Coffee breaks (provided)
Lunch (not provided, reservation nearby)

# OUTLINE

**10.00 – 11.30**

- Session 1: Introduction, Network Data, Dyads, Distance, Triads

**11.30 – 12.30** **Lunch**

**12.30 – 14.00**

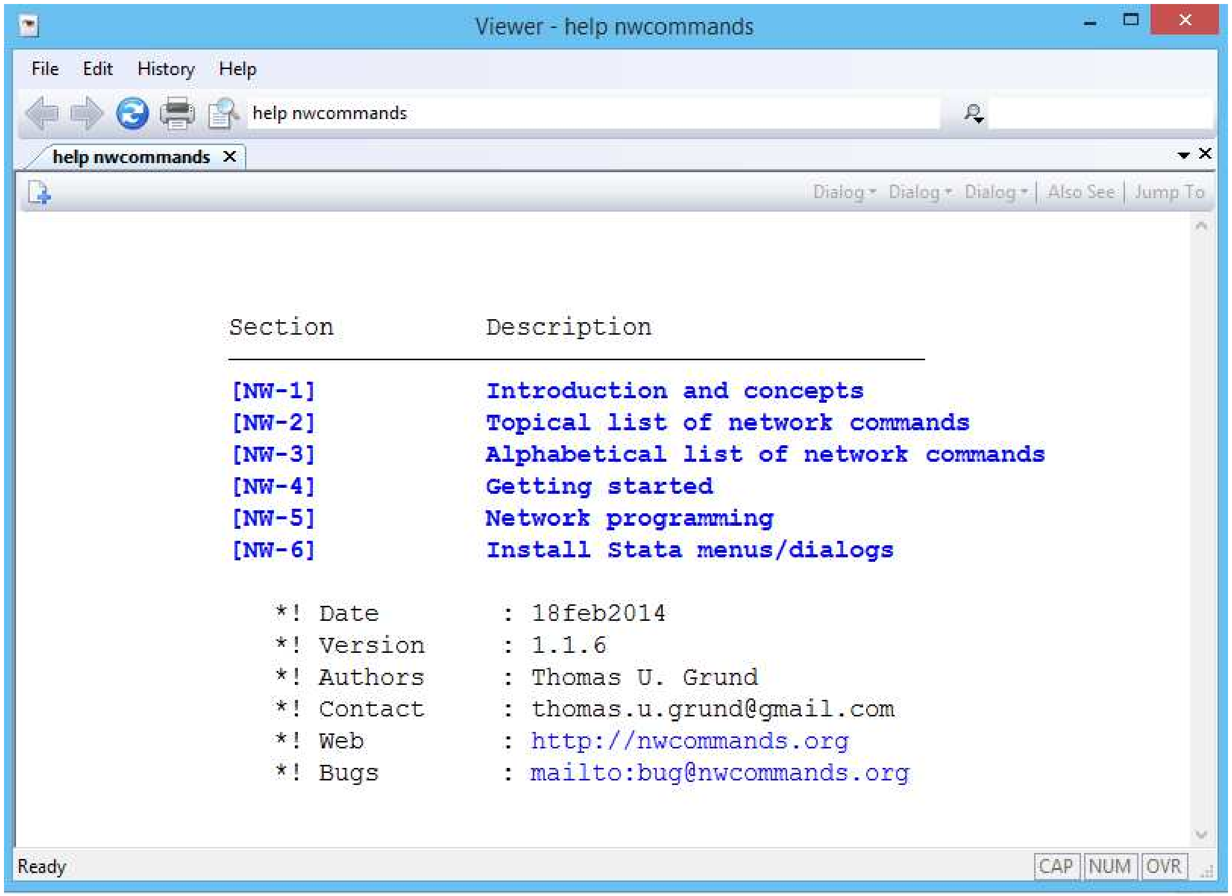
- Session 2: Centrality, Centralization, Simulation, Visualization, Animation

**14.00 – 14.30** **Coffee Break**

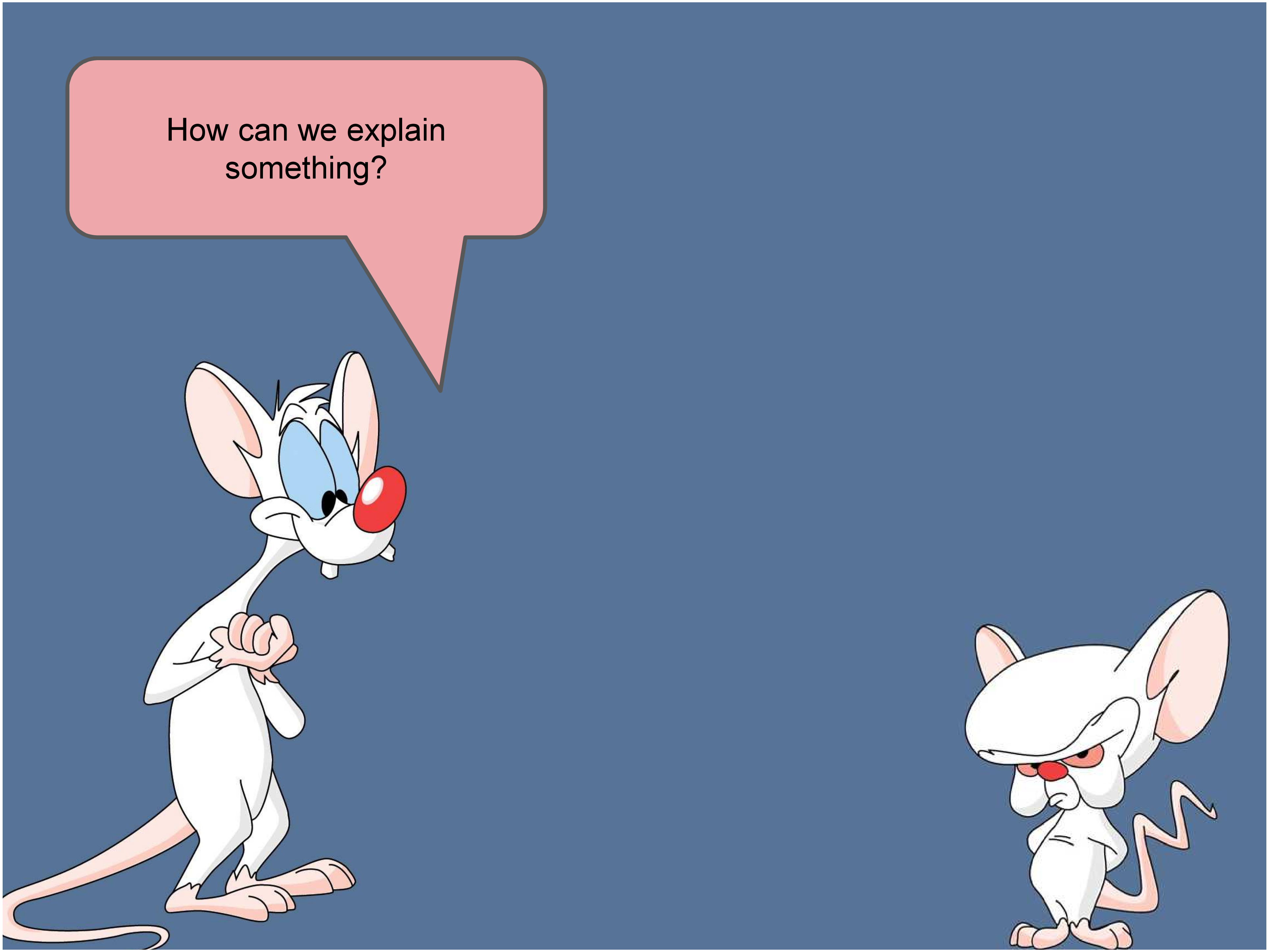
**14.30 – 16.00**

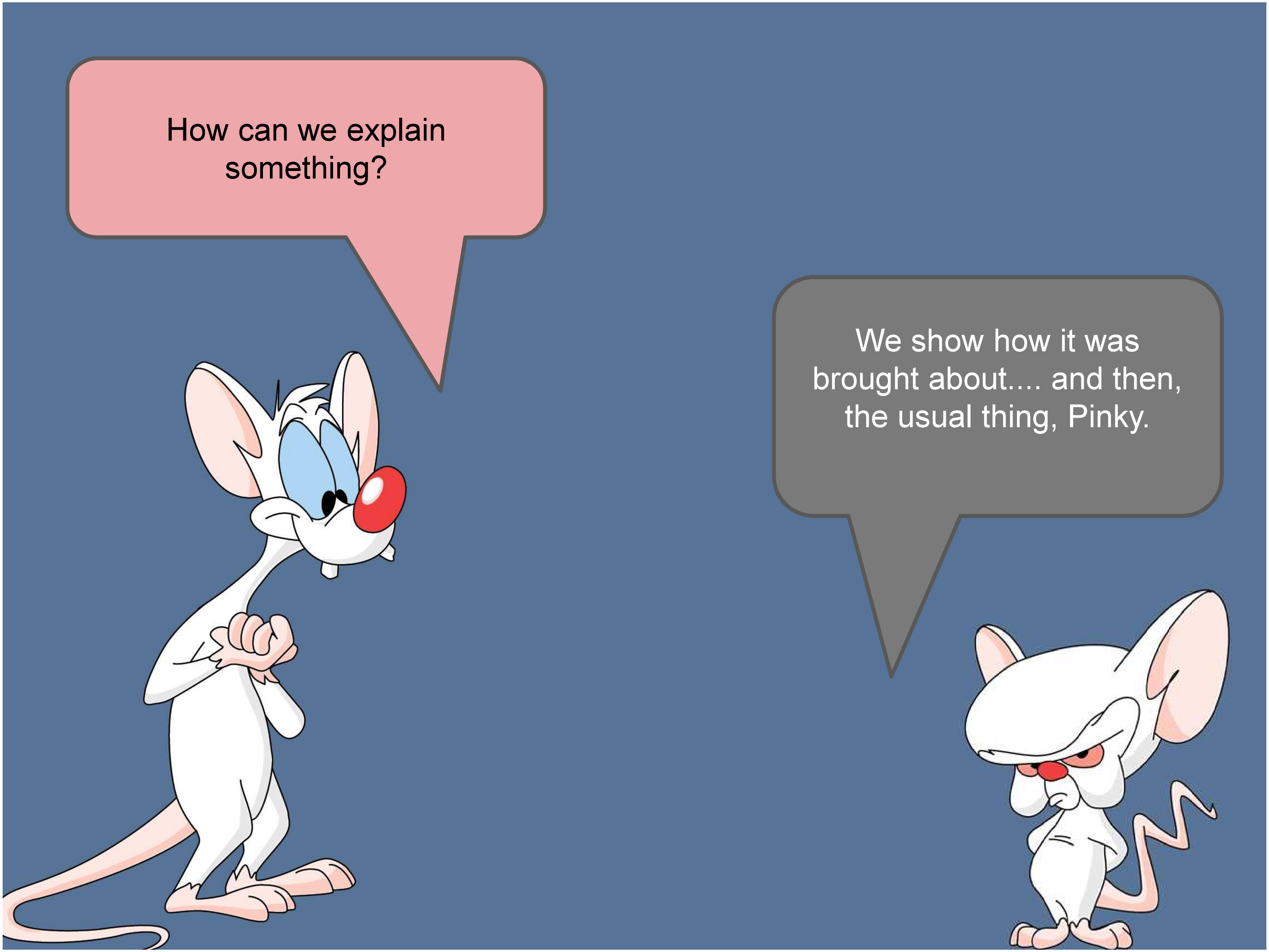
- Session 3: Hypothesis Testing, Conditional Uniform Graphs, Quadratic Assignment Procedure, Regression Approach, ERGM

the big quiz thing



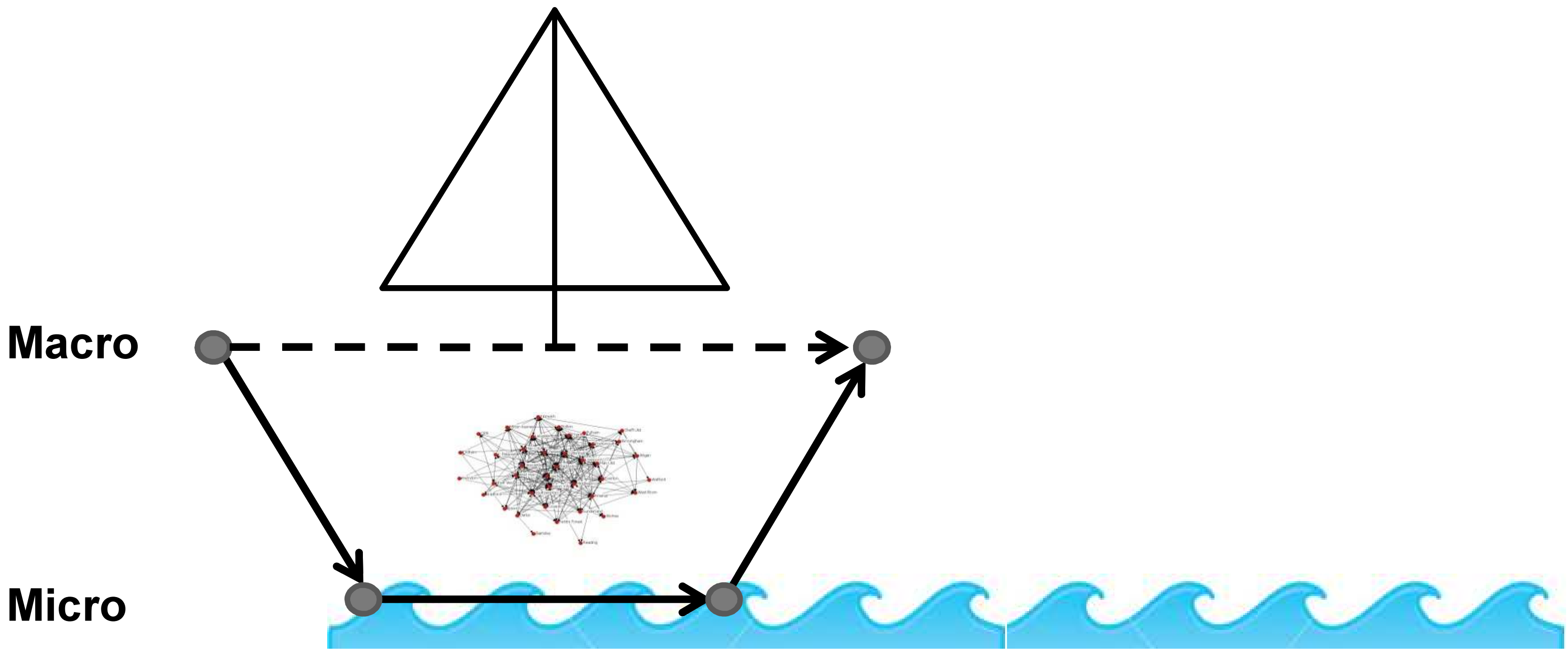

```
. help nwcommands
```


How can we explain something?


How can we explain something?
We show how it was brought about.... and then, the usual thing, Pinky.

# DISSECTING THE SOCIAL

**Coleman's boat**



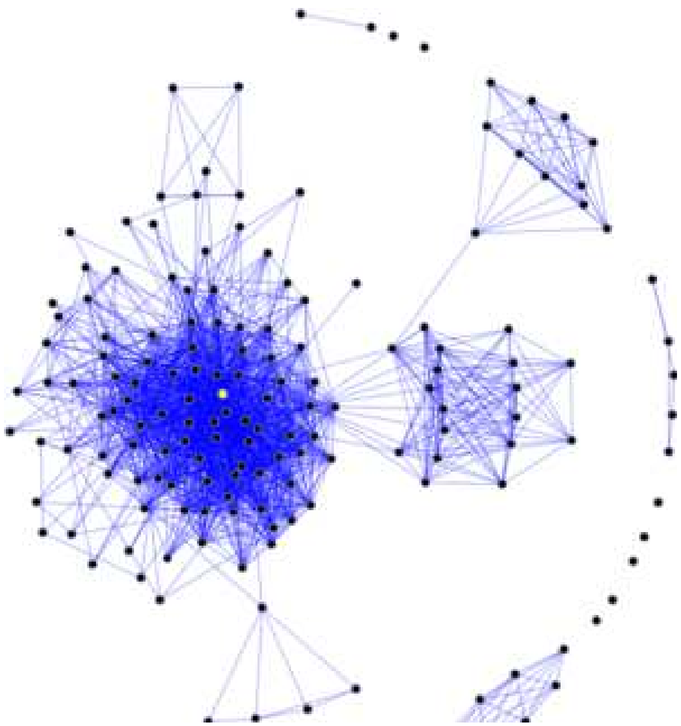

- We cannot simply speak of "aggregations".
- From the principle of dissection immediately follows that one needs to examine explicitly how the entities are arranged and how they interact with each other. Networks are crucial.
- See Coleman (1990), Hedstrom (2006)

# NETWORK PARADIGM

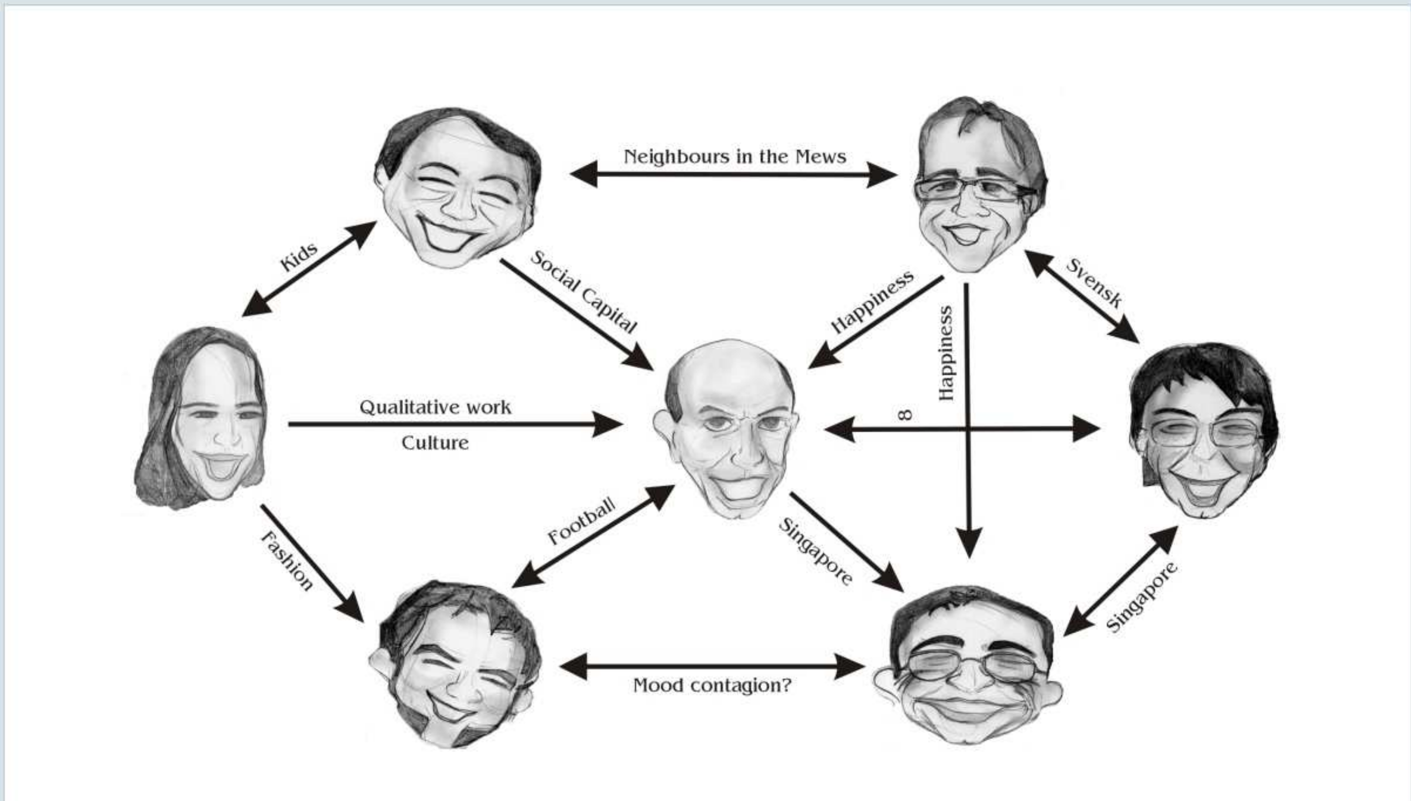
- Not just composition of elements of a system that matters, but also how the elements are arranged and related with each other.
- An individual's position in a network (social structure) determines the opportunities and constraints this individual will encounter.
- Individuals change the social world of others. Individuals are dependent and embedded in a web of relations.

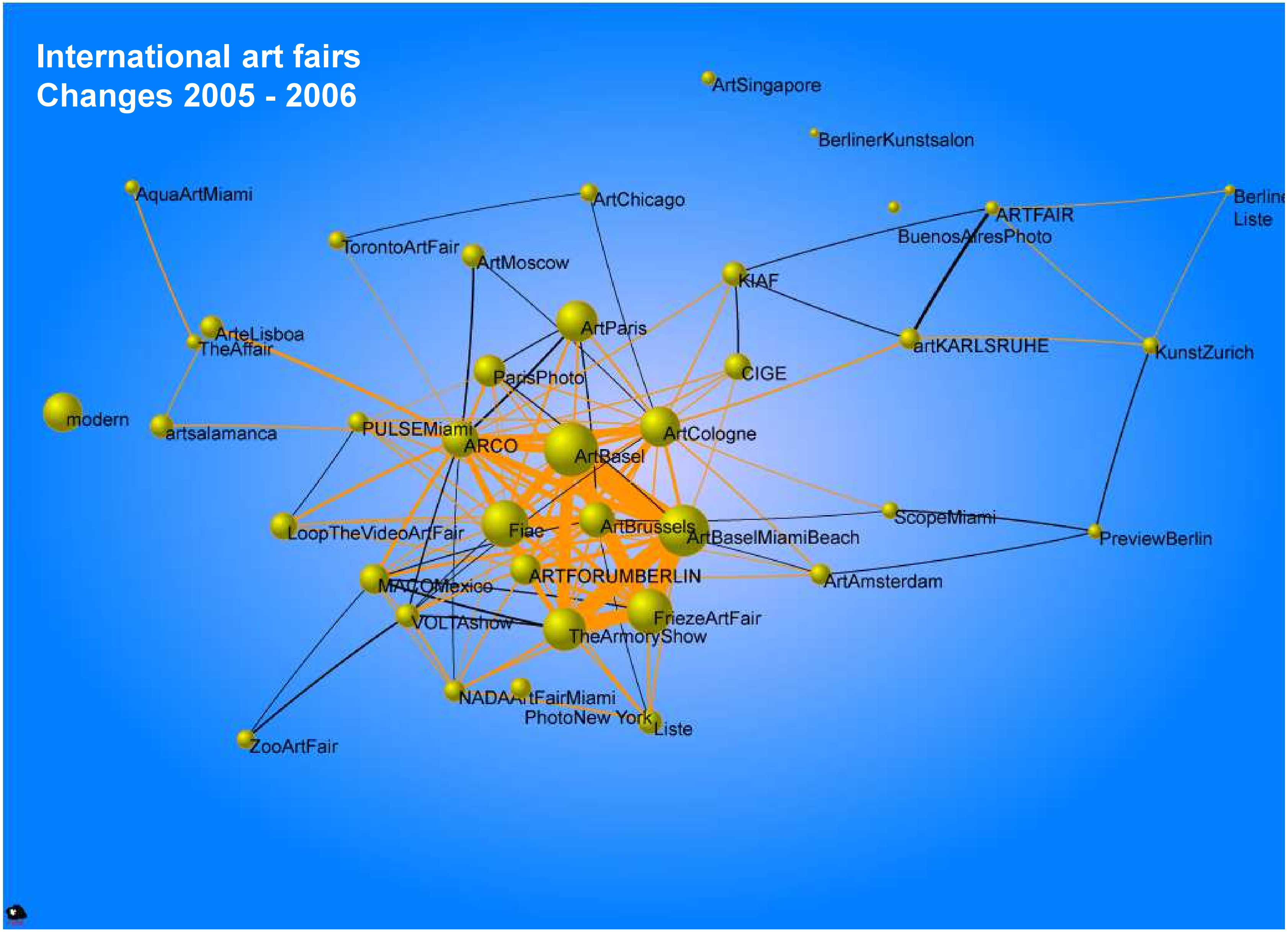
unit of analysis = dyad

# SOCIAL NETWORKS



# Nuffield Network 2008

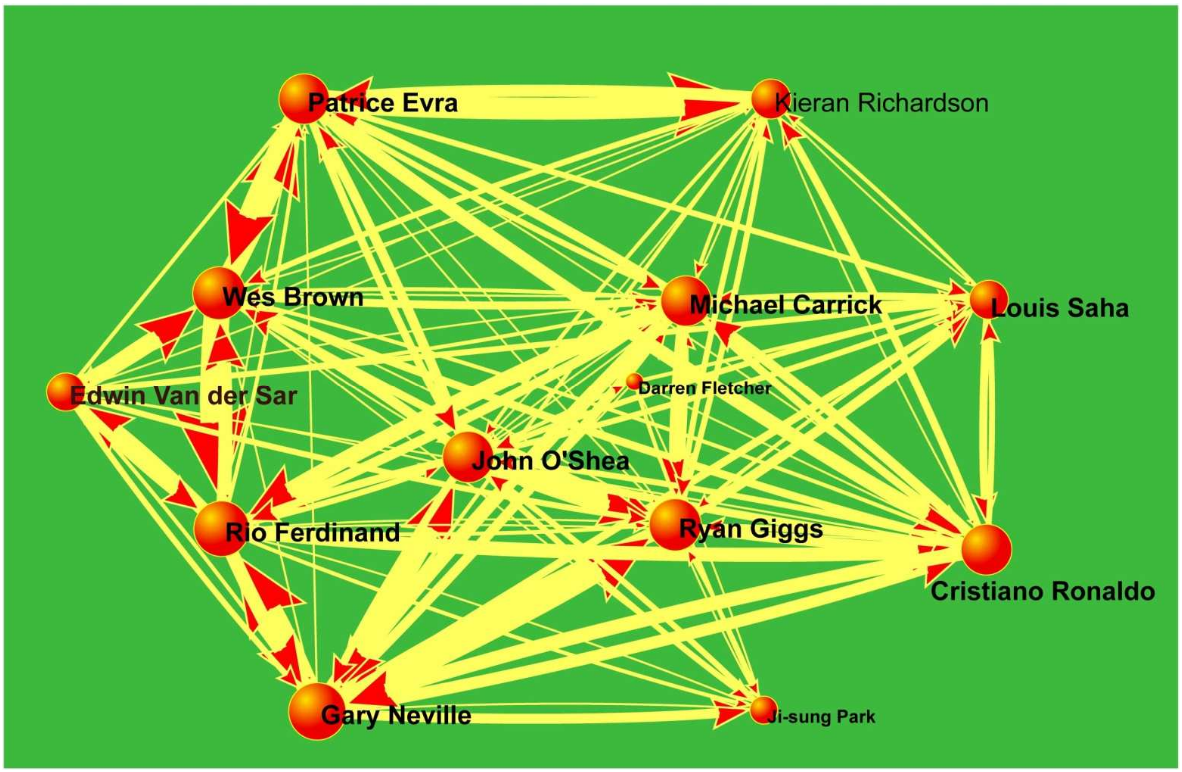




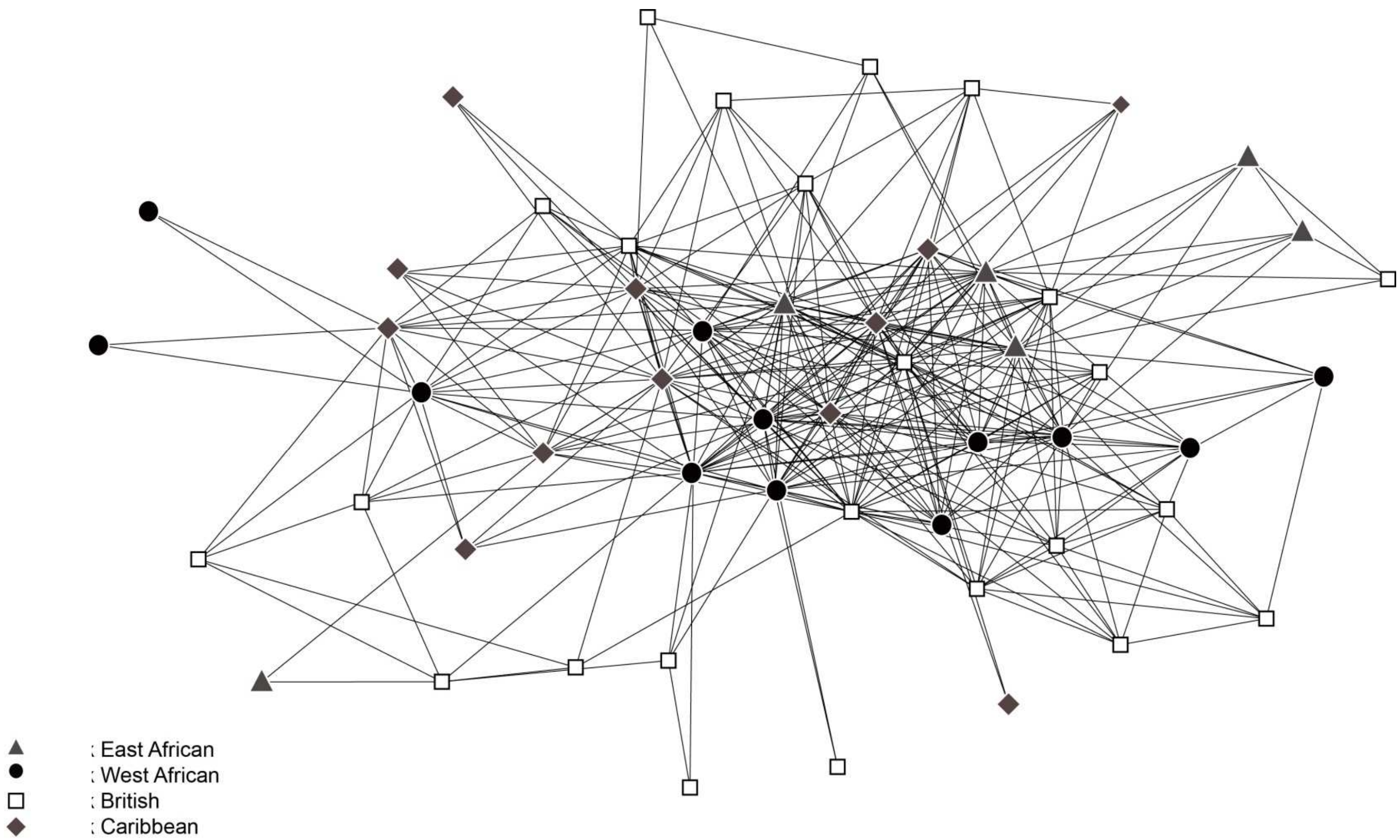

Density = 0.18, Avg Degree = 6.79

# MANCHESTER UTD – TOTTENHAM

9/9/2006, Old Trafford



# CO-OFFENDING IN GANG



# KNOW YOUR TIES

- Social network analysis is not always about “friendship”. Many different forms of relations might matter. Probably they are even influencing each other.
- Be very clear on the type of relationship you are looking at and do not throw obviously different types of relationships in the same pot and treat them equally.

# SOCIAL NETWORKS

- **Social**
  - Friendship, kinship, romantic relationships
- **Government**
  - Political alliances, government agencies
- **Markets**
  - Trade: flow of goods, supply chains, auctions
  - Labor markets: vacancy chains, getting jobs
- **Organizations and teams**
  - Interlocking directorates
  - Within-team communication, email exchange

# NETWORKS ARE EVEN MORE UNIVERSAL

- Food webs
- Internet
- Power grids, airline networks
- Metabolic networks
- Neural networks
- Economics networks
- …

# DEFINITION

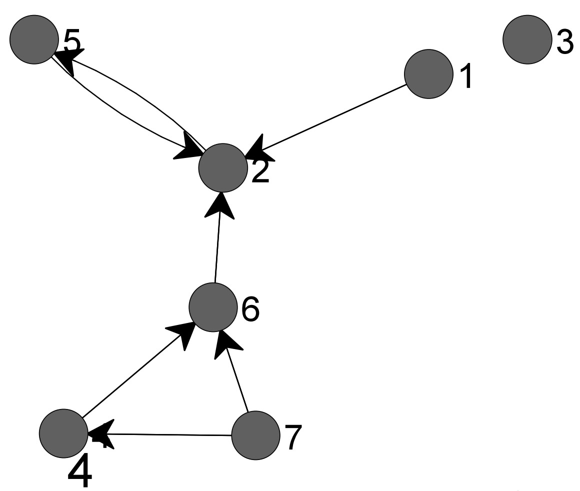
- Mathematically, a (binary) network is defined as $G = (V, E)$ where $V = \{1,2,..,n\}$ is a set of "vertices" (or "nodes") and $E \subseteq \{\langle i,j \rangle |\ i,j \in V\}$ is a set of "edges" (or "ties", "arcs"). Edges are simply pairs of vertices, e.g. $E \subseteq \{(1,2), (2,5) \ldots\}$.
- We write $y_{ij} = 1$ if actors $i$ and $j$ are related to each other (i.e., if $\langle i,j \rangle \in E$), and $y_{ij} = 0$ otherwise.
- In digraphs (or directed networks) it is possible that $y_{ij} \neq y_{ji}$.

# ADJACENCY MATRIX

- We write $y_{ij} = 1$ if actors $i$ and $j$ are related to each other (i.e., if $\langle i, j \rangle \in E$), and $y_{ij} = 0$ otherwise
- The matrix $\boldsymbol{y}$ is called the adjacency matrix and is a convenient representation of a network.

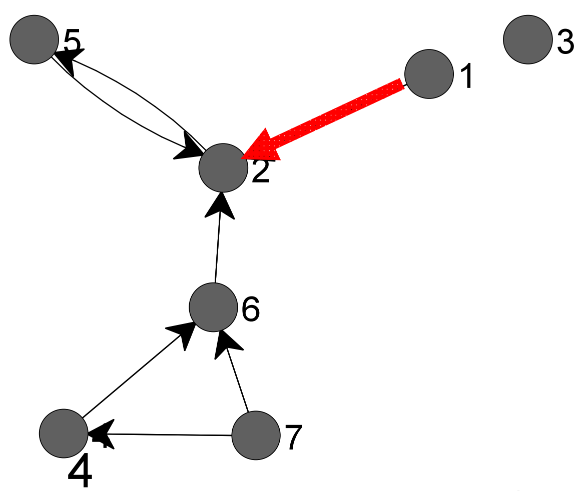
$$\boldsymbol{y} = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{nj} & \cdots & y_{nb} \end{bmatrix}$$

# ADJACENCY MATRIX




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |

# ADJACENCY MATRIX

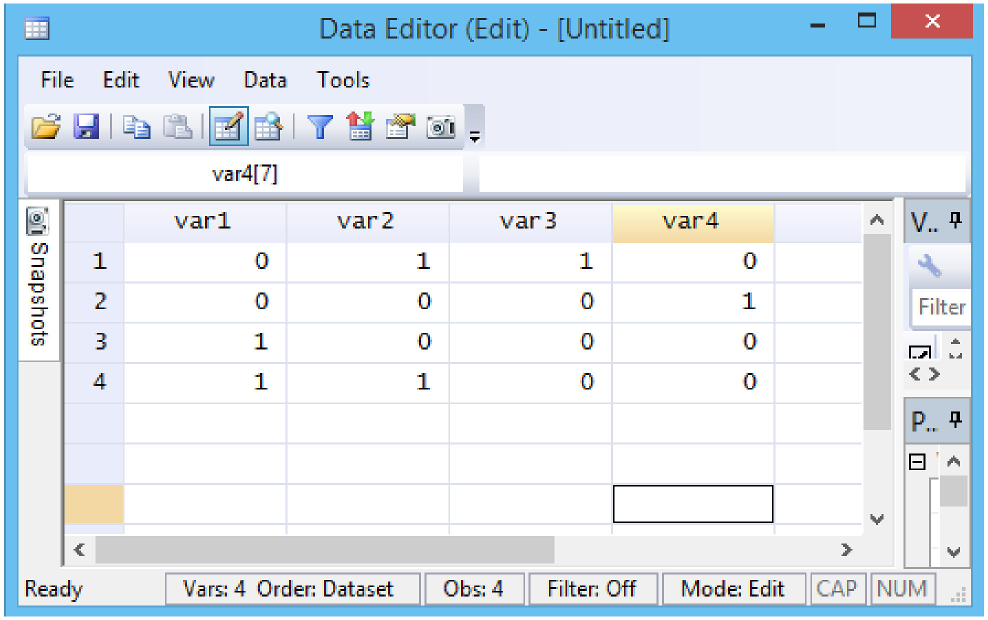


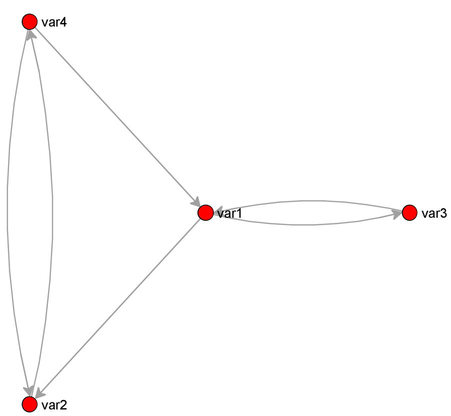

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |

# MORE DEFINITIONS

- In valued networks an edge is a triple of two vertices and one value (i.e. $E \subseteq \{\langle i, j, k\rangle | \; i, j \in V, k \in \mathbb{R}\}$).
- In multiplex networks, nodes can have more than one relationship with each other (e.g. friendship and advice).
- In two-mode/bipartite networks, the nodes are part of subsets and ties can only exist between nodes from different subsets, e.g. “individuals work in organizations”.
- In hypergraphs an edge is a set of more than two nodes.

# NETWORK DATA




```
. nwset _all

. nwds
network

. nwset
(1 network)
------------------------
      network

. nwsummarize network
--------------------------------------

   Network name:  network
   Network id:  1
   Directed:  true
   Nodes:  4
   Arcs:  6
   Minimum value:  0
   Maximum value:  1
   Density:  .5
```

```
. webnwuse florentine

. nwimport http://nwcommands.org/data/madrid.dat, type(ucinet)

. help netexample
```

# NETWORK DATA

```
nwset
nwds
nwsummarize

nwdrop
nwkeep
nwclear
```

```
webnwuse
nwimport
nwexport
nwuse
nwsave
```

# DYADS

# DYAD

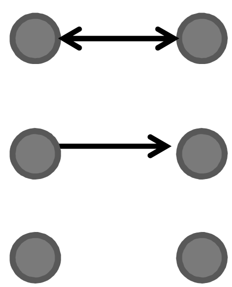
A dyad is a pair of actors $(i, j)$ in the network, plus the configuration of the tie variables $(y_{ij}, y_{ji})$ between them.

- In a directed, binary network, there are $n(n-1)$ tie variables located in $n(n-1)/2$ dyads.
- Dyads can be of three types:

*M: mutual*

*A: asymmetric*

*N: null*



# DYAD

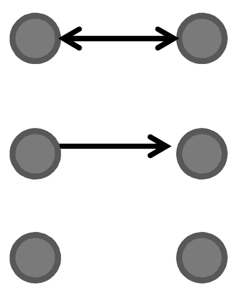
A dyad is a pair of actors $(i, j)$ in the network, plus the configuration of the tie variables $(y_{ij}, y_{ji})$ between them.

- In a directed, binary network, there are $n(n-1)$ tie variables located in $n(n-1)/2$ dyads.
- Dyads can be of three types:

*M: mutual*

*A: asymmetric*

*N: null*



# DYAD

A dyad is a pair of actors $(i, j)$ in the network, plus the configuration of the tie variables $(y_{ij}, y_{ji})$ between them.

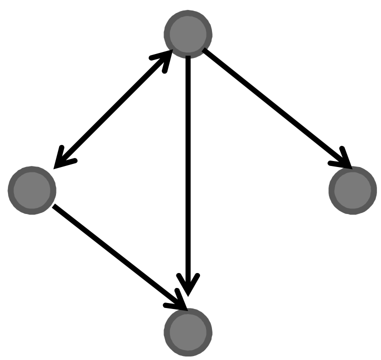
- In an undirected, binary network, there are $n(n-1)/2$ tie variables located in $n(n-1)/2$ dyads.
- Dyads can be of two types:

*M: mutual*

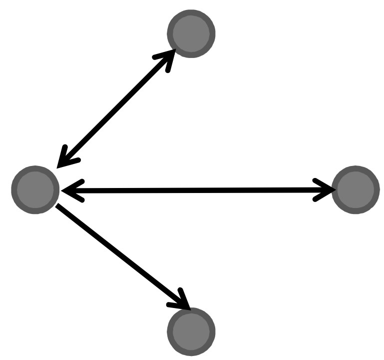
*N: null*

# DYAD CENSUS

We can describe a network by counting the number of **mutual, asymmetric** and **null** dyads. It is like taking a “fingerprint” of a network.



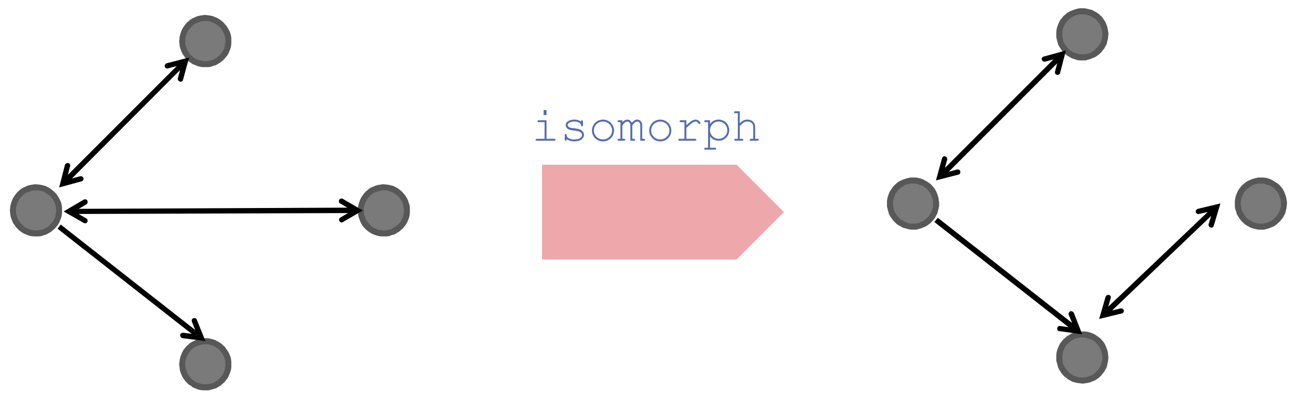
MAN = 132



MAN = 213

# ISOMORPHISM

Two networks are isomorph, when they do not differ according to their “fingerprint”.

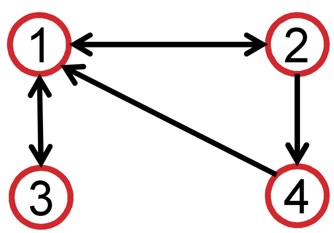



MAN = 213

MAN = 213

# DEGREE

- The indegree $indeg(v)$ of node $v$ is simply the number of ties that point towards $v$.
- The outdegree $outdeg(v)$ of node $v$ is simply the number of ties that point away from $v$.



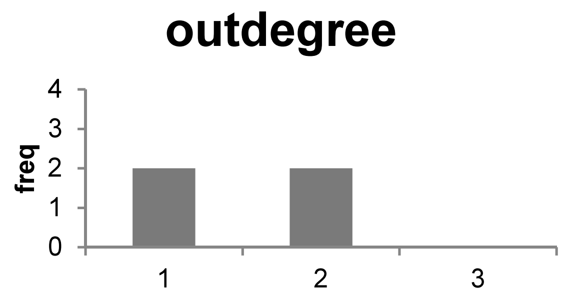

indeg(1) = 3
outdeg(1) = 2

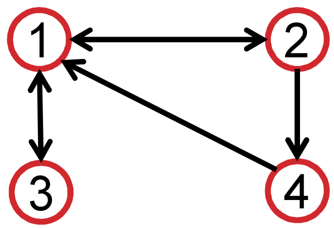
indeg(2) = 1
outdeg(2) = 2

indeg(3) = 1
outdeg(3) = 1

indeg(4) = 1
outdeg(4) = 1

# DEGREE DISTRIBUTION


indeg(1) = 3
outdeg(1) = 2

indeg(2) = 1
outdeg(2) = 2

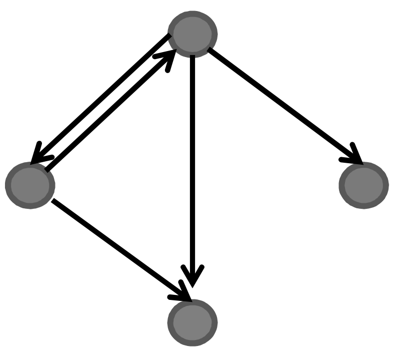
indeg(3) = 1
outdeg(3) = 1

indeg(4) = 1
outdeg(4) = 1

# DENSITY

The density of a network is defined as the proportion of actually observed ties among the potentially observable ones.

- Remember, in a directed, binary network with $n$ actors there could be $n(n-1)$ ties. In an undirected network, there could be $n(n-1)/2$ ties.

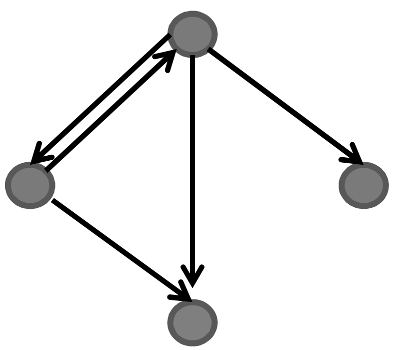


$$density = \frac{5}{4(4-1)} = \frac{5}{12} \approx 0.416$$

# DENSITY

We could also calculate density from the dyad census. Remember, M = mutual dyads, A = asymmetric dyads.

- Actually observed ties are $2M + A$
- Potential ties are $n(n-1)$



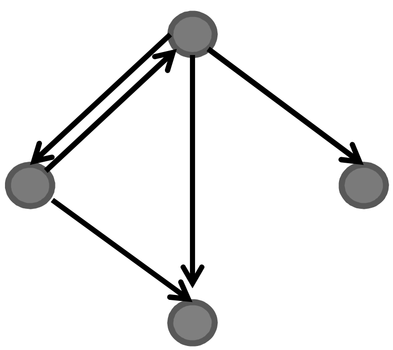
$$density = \frac{2M + A}{n(n-1)}$$

$$density = \frac{5}{4(4-1)} = \frac{5}{12} \approx 0.416$$

# RECIPROCITY

The reciprocity of a network is defined as the proportion of actually reciprocated ties among the potentially reciprocable ones. Remember, M = mutual dyads, A = asymmetric dyads.

- Actually reciprocated dyads are $2M$.
- Potentially reciprocated dyads are $2M + A$



$$reciprocity = \frac{2M}{2M + A}$$

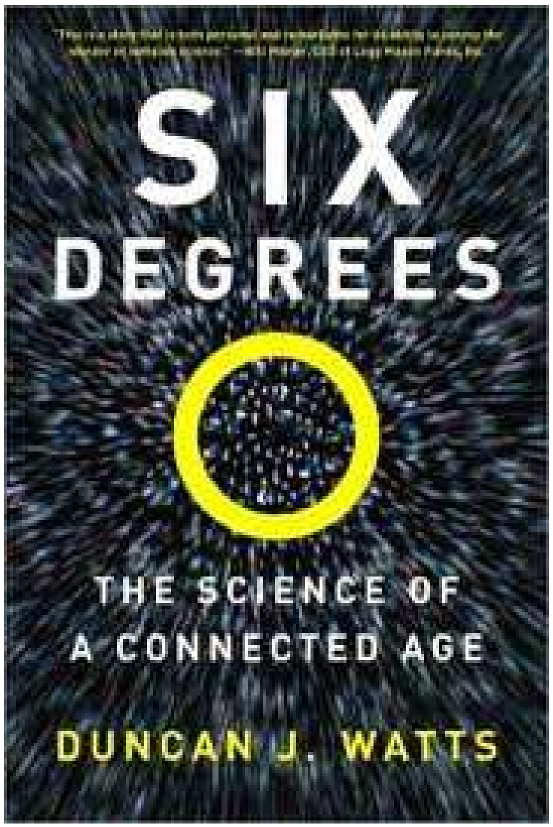
$$reciprocity = \frac{2}{5} = 0.4$$

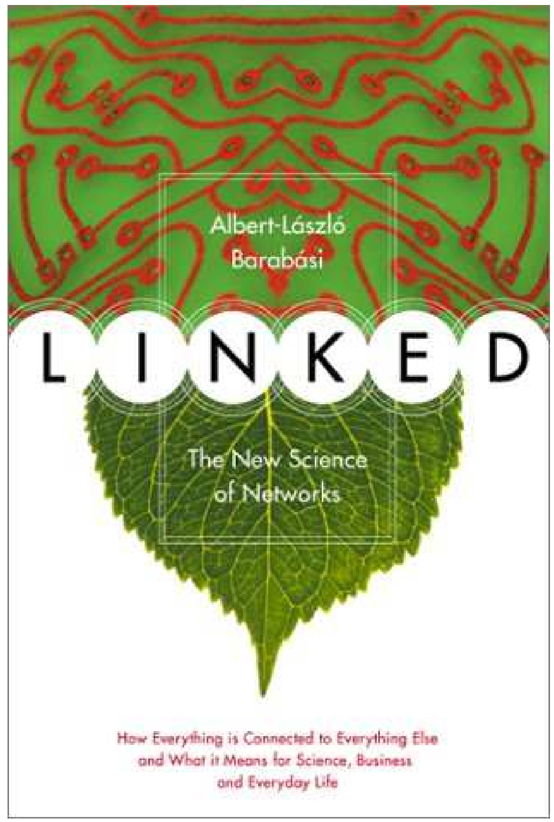
# DYADS

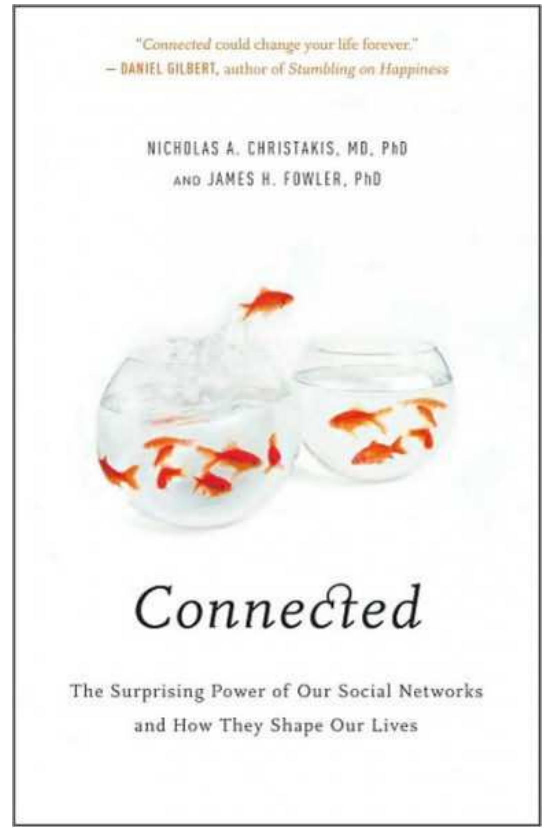
nwdyads
nwdegree
nwsummarize

- Load the **hpotter** data from the nwcommands-Server using `webnwuse`
- List the networks in this file (either `nwds` or `nwset` will do)
- Summarize the network **hpbook1**
- Calculate the dyad census for network **hpbook1**

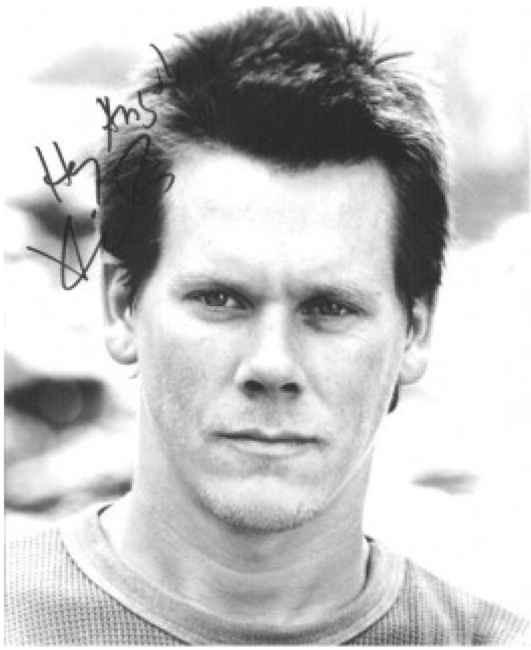
# DISTANCE


SIX
DEGREES
THE SCIENCE OF
A CONNECTED AGE
DUNCAN J. WATTS


Albert-László
Barabási
LINKED
The New Science
of Networks
How Everything is Connected to Everything Else
and What it Means for Science, Business
and Everyday Life


"Connected could change your life forever."
— DANIEL GILBERT, author of Stumbling on Happiness
NICHOLAS A. CHRISTAKIS, MD, PhD
AND JAMES H. FOWLER, PhD
Connected
The Surprising Power of Our Social Networks
and How They Shape Our Lives

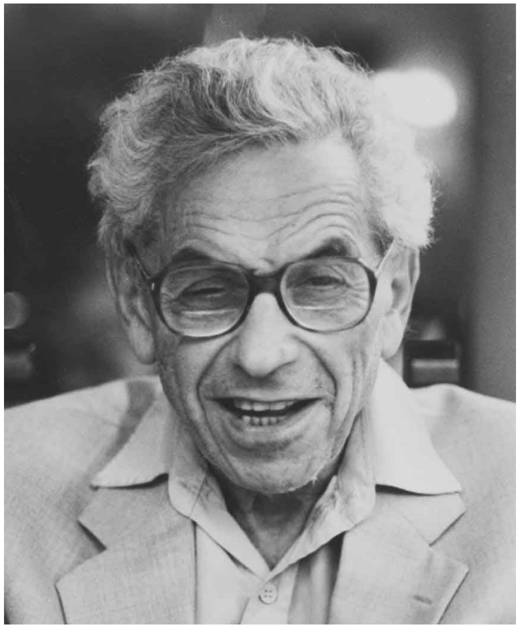
**Kevin Bacon**



http://oracleofbacon.org/

**?**

**Paul  Erdős**

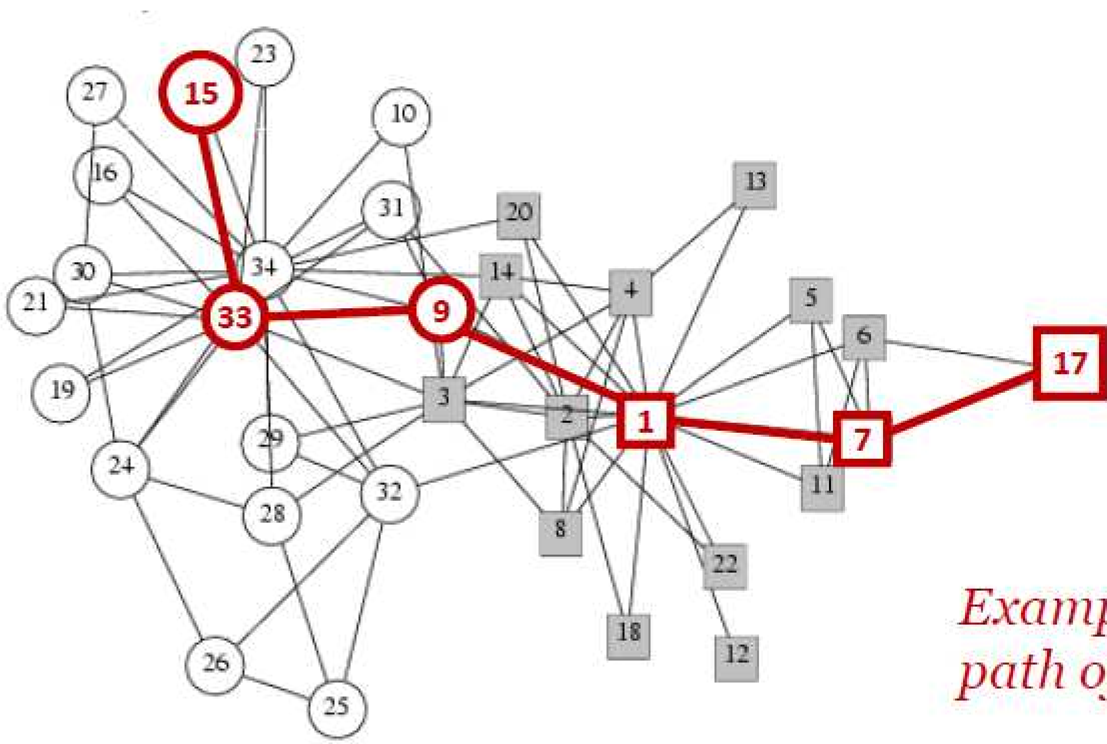


http://academic.research.microsoft.com/VisualExplorer

**?**

# DISTANCE

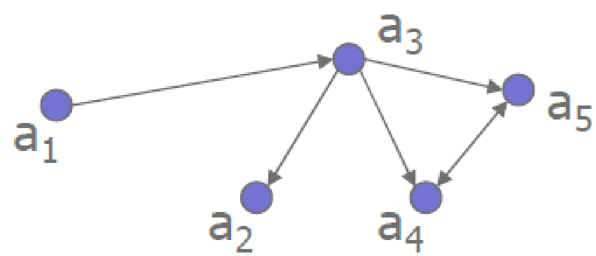
Length of a shortest connecting path defines the (geodesic) distance between two nodes.




*Example of a shortest path of length 5*

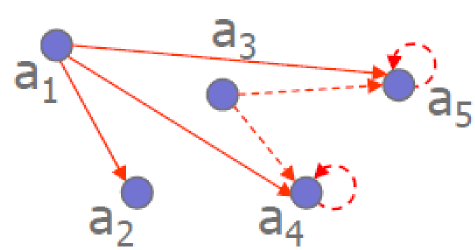
# DISTANCE

**How can we calculate the distance?**




- Matrix $y$ indicates which row actor is directly connected to which column actor.
- The squared matrix $y^2$ indicates which row actor can reach which column actor in two steps.
- The matrix $y^l$ indicates who reaches whom in $l$ steps.

$$y^2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{pmatrix}$$



# DISTANCE

When we take the average of the shortest paths between all nodes (if all are connected) we get the “average shortest path length” $\ell$ of the network.

**Intuition:** If we were to select two nodes at random, how many steps would it take ‘on average’ to connect them?

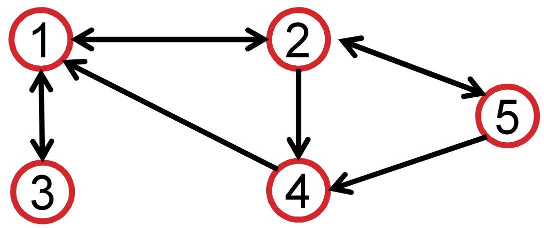
For a random graph one can show that:

$$\ell \approx \frac{\ln(n)}{\ln(k)}$$

$n =$ number of nodes

$k =$ average degree of nodes

# DISTANCE




$$distances = \begin{bmatrix} 0 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 & 1 \\ 1 & 2 & 0 & 3 & 3 \\ 1 & 2 & 2 & 0 & 3 \\ 2 & 1 & 3 & 1 & 0 \end{bmatrix}$$

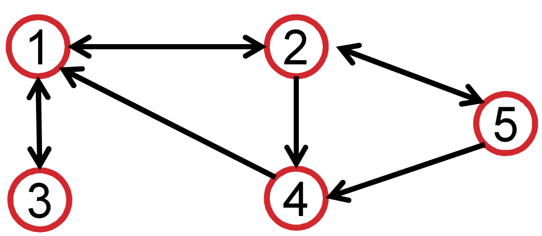
$$avgerage\ shortest\ path\ length = \ \ 1.8$$

# DISTANCE DISTRIBUTION

- Networks can have the same "average shortest path length", but still be vastly different from each other.

- Better, look at the "distribution of shortest paths" instead of the average.
  - Calculate how often each distance occurs.

$$\begin{bmatrix} 0 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 & 1 \\ 1 & 2 & 0 & 3 & 3 \\ 1 & 2 & 3 & 0 & 3 \\ 2 & 1 & 3 & 1 & 0 \end{bmatrix}$$

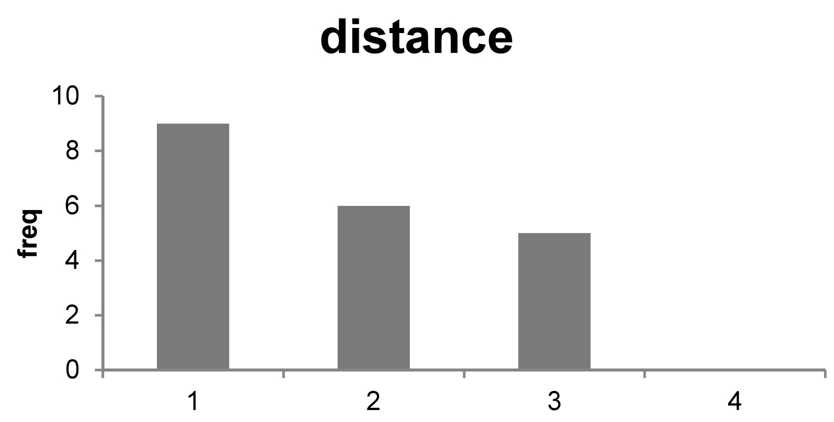


# DISTANCE DISTRIBUTION

- Networks can have the same "average shortest path length", but still be vastly different from each other.

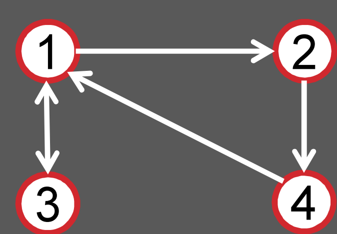
- Better, look at the "distribution of shortest paths" instead of the average.
  - Calculate how often each distance occurs.

$$\begin{bmatrix} 0 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 & 1 \\ 1 & 2 & 0 & 3 & 3 \\ 1 & 2 & 3 & 0 & 3 \\ 2 & 1 & 3 & 1 & 0 \end{bmatrix}$$

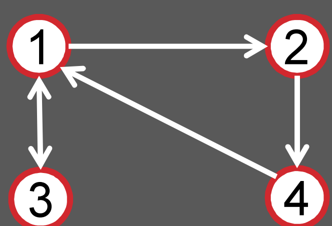


the big quiz thing

What is the "average shortest path" length and what is the distance distribution?



**What is the "average shortest path" length and what is the distance distribution?**



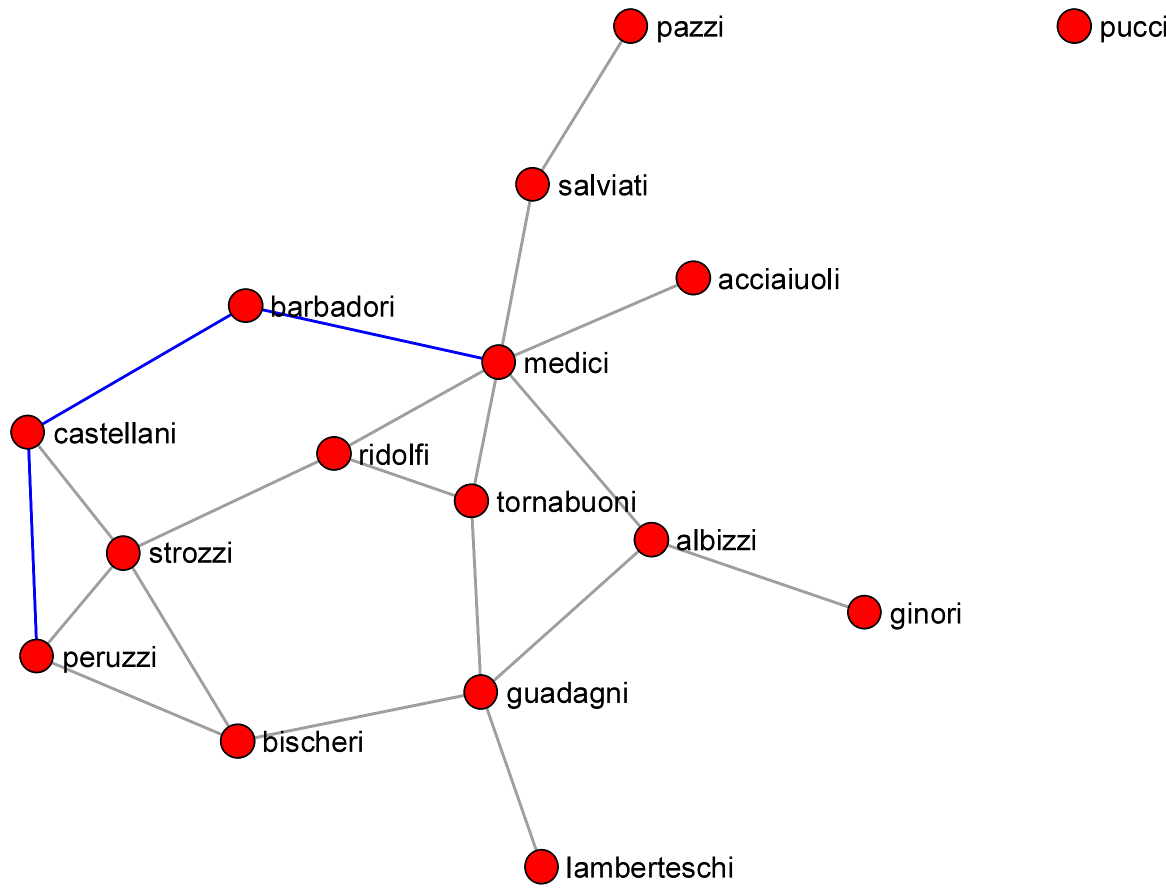

| | |
|---|---|
| From 1 to 2,3,4: | 1,1,2 |
| From 2 to 1,3,4: | 2,3,1 |
| From 3 to 1,2,4: | 1,2,3 |
| From 4 to 1,2,3: | 1,2,2 |

5 x 1 step
5 x 2 steps
2 x 3 steps

$$\ell = \frac{21}{12} = 1.75$$

# DISTANCE

`nwgeodesic`
`nwpath`


pazzi
pucci
salviati
acciaiuoli
barbadori
medici
castellani
ridolfi
tornabuoni
albizzi
strozzi
ginori
peruzzi
guadagni
bischeri
lamberteschi

# TRIADS

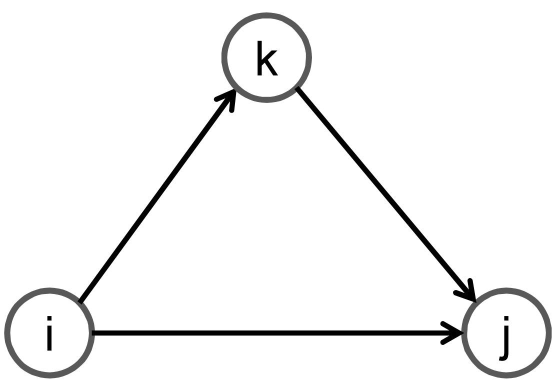
# TRIAD

A triad is a set of three actors $(i, j, k)$ plus the configuration of all tie variables $(y_{ij}, y_{ik}, y_{ji}, y_{jk}, y_{ki}, y_{kj})$ between them.

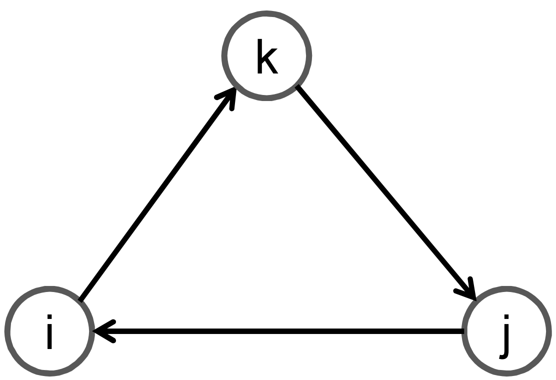
# TRIAD

A triad is transitive when there is a tie $(i, j)$, another one $(i, k)$ and a third one $(k, j)$. Transitivity shows hierarchy.

transitive




non-transitive

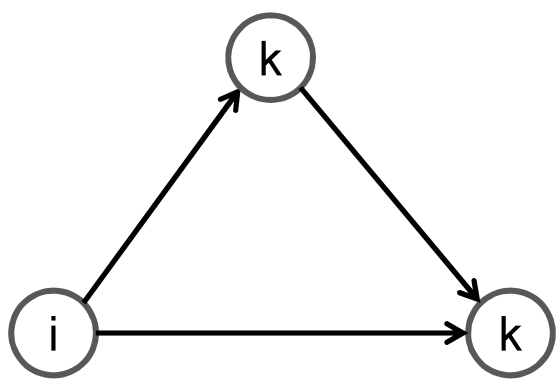


# TRIAD

A triad is cyclical when there is a tie $(i, j)$, another one $(j, k)$ and a third one back $(k, i)$. Cyclicity indicates the absence of hierarchy!

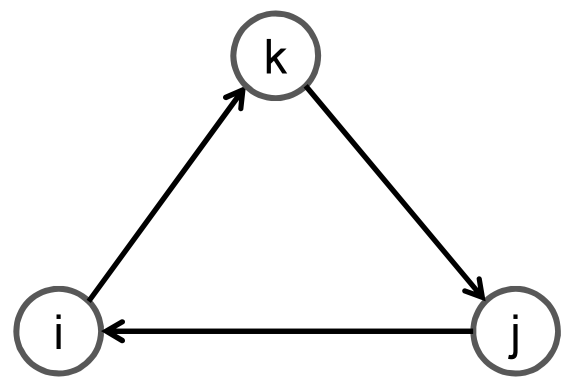
non-cyclical

transitive




cyclical

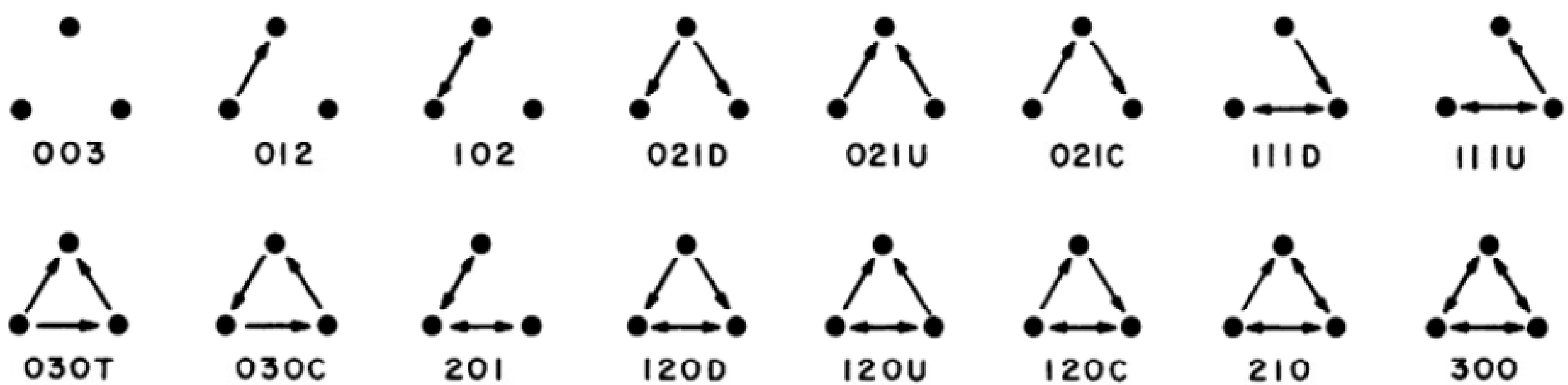
non-transitive



# TRIAD

We can describe a triad as before by counting the number of **mutual, asymmetric**, and **null** dyads plus (where necessary) a distinguishing letter.

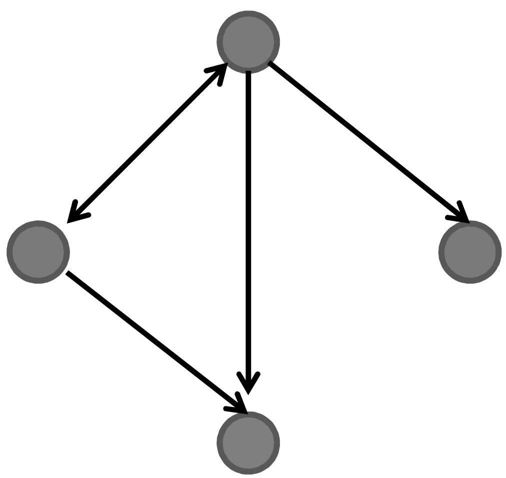
*C = cyclical, T = transitive, U = up, D = down,*



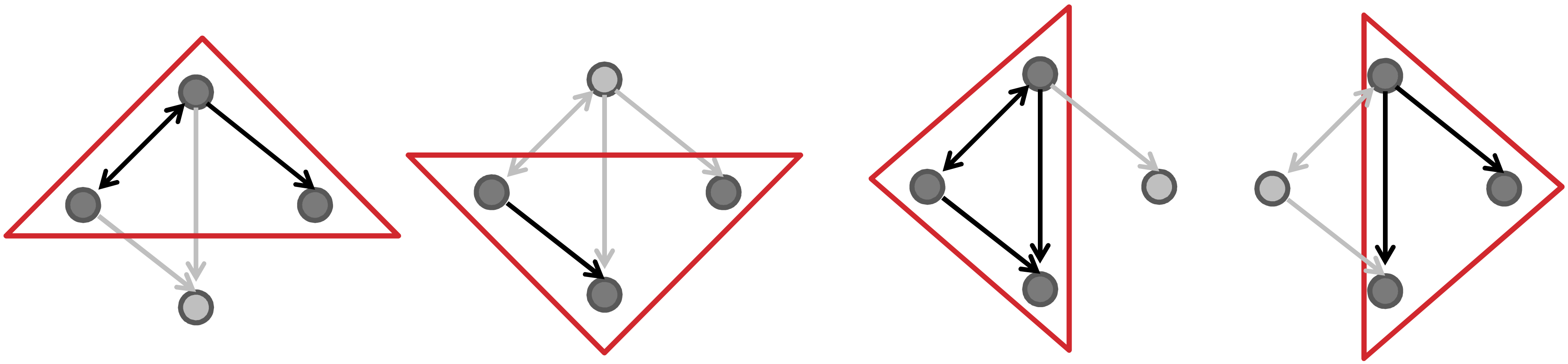

There are 16 possible triad configurations.

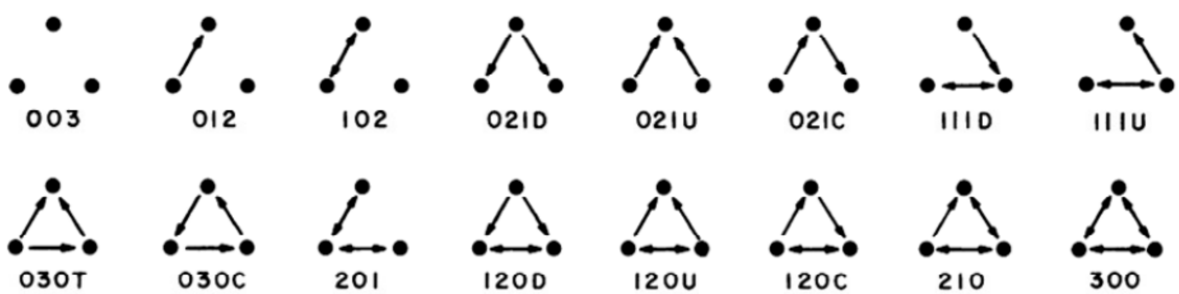
# TRIAD CENSUS

We can now describe a whole network according to its triad census (similar as we did before with the dyad census). Simply count the number of times each of the 16 possible triad configuration appears in the network.

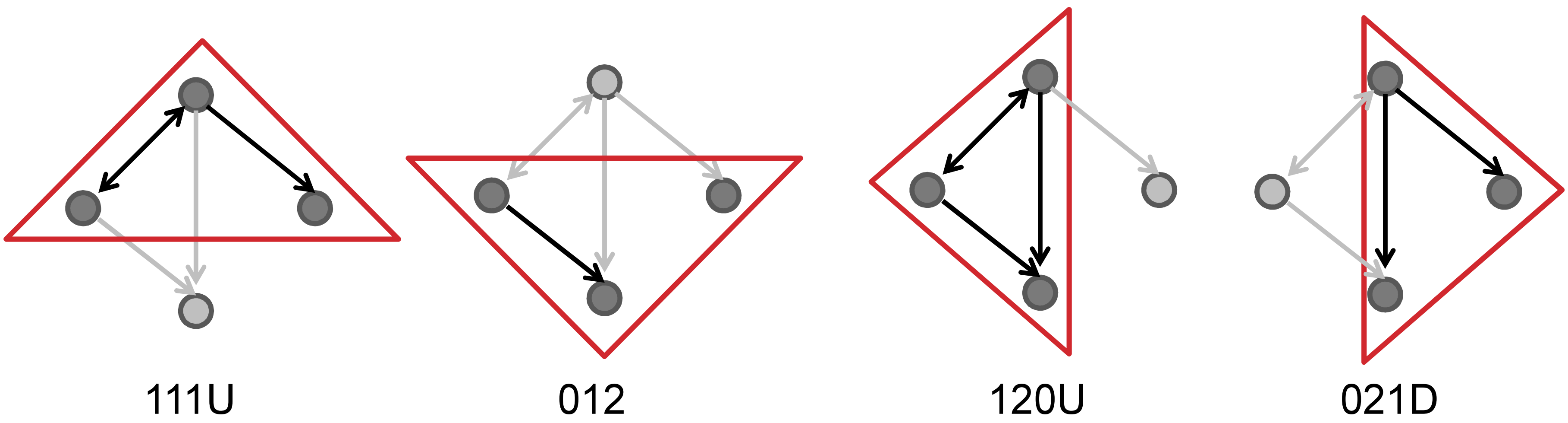


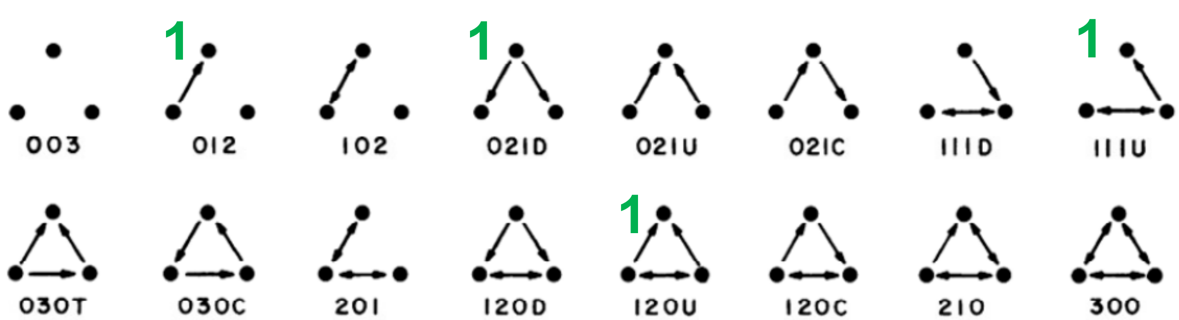
# TRIAD CENSUS





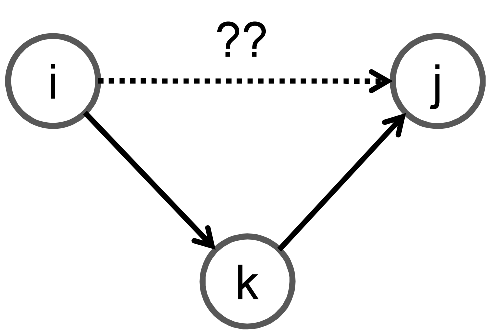
# TRIAD CENSUS



# TRANSITIVITY

The transitivity of a network gives you an idea about you how locally connected the network is. It is defined as the proportion of actually observed transitively closed triples $\langle i, j, k \rangle$ of nodes among the observed potentially closed paths of length 2 from $i$ to $j$ via $k$.



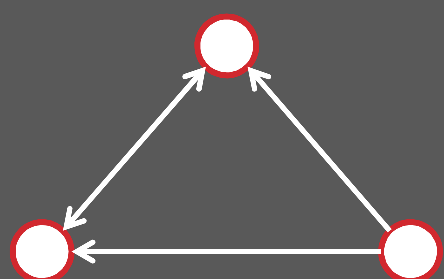

Think of it in this way: Given that two nodes i and j are indirectly connected (via k), what is the probability that there is a direct link from i to j?

the big quiz thing

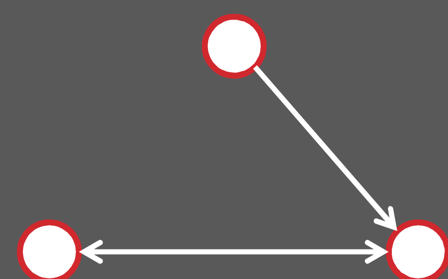
What is the description of the triads below?

the big quiz thing

## What is the description of the triads below?



M = 1, A = 2, N = 0, Down
**120D**



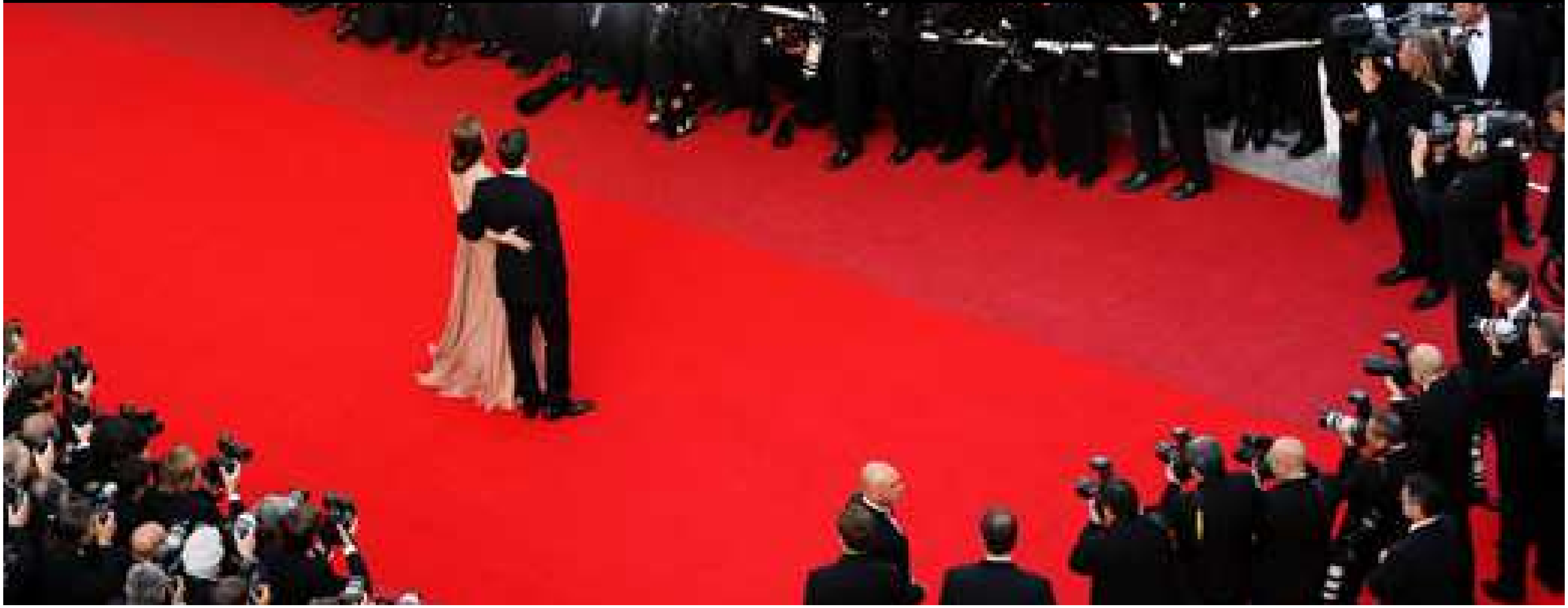
M = 1, A = 1, N = 1 Down
**111D**

# TRIADS

`nwtriads`

- Load the **glasgow** data from the nwcommands-Server using `webnwuse`
- Calculate the triad census for all networks

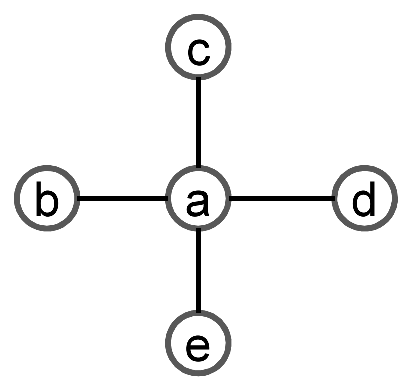
# CENTRALITY



# CENTRALITY

## Well connected actors are in a structurally advantageous position.

- Getting jobs
- Better informed
- Higher status
- …

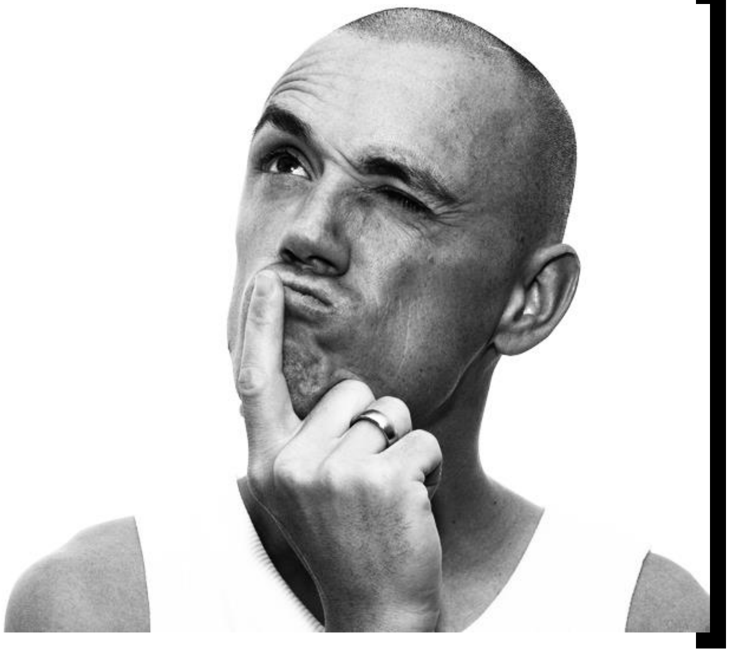


# CENTRALITY

## Well connected actors are in a structurally advantageous position.

- Getting jobs
- Better informed
- Higher status
- …

**What is "well-connected?"**



# DEGREE CENTRALITY

**Degree centrality**

- We already know this. Simply the number of incoming/outgoing ties => indegree centrality, outdegree centrality
- How many ties does an individual have?

$$C_{odegree}(i) = \sum_{j=1}^{N} y_{ij} \qquad C_{idegree}(i) = \sum_{j=1}^{N} y_{ji}$$

# DEGREE CENTRALITY

**Degree centrality**

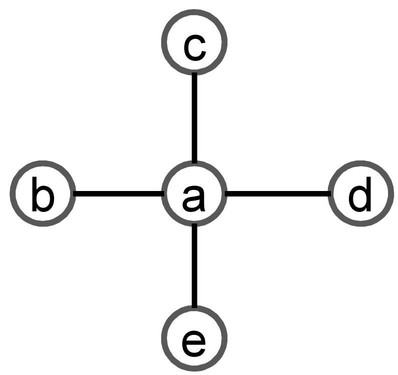
$$C_{degree}(i) = \sum_{j=1}^{N} y_{ij}$$

$C_{degree}(a) = 4$

$C_{degree}(b) = 1$

$C_{degree}(c) = 1$

…



# CLOSENESS CENTRALITY

**Closeness centrality**

- How close is an individual (on average) from all other individuals?

**Farness**

- How many steps (on average) does it take an individual to reach all other individuals?

$$Farness(i) = \frac{1}{N-1}\sum_{j=1}^{N} l_{ij}$$

$j \neq i$

$l_{ij} =$ shortest path between i and j

# FARNESS

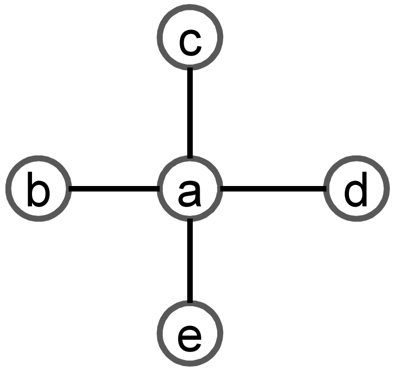
**Farness**

$$Farness(i) = \frac{1}{N-1}\sum_{j=1}^{N} l_{ij}$$

$$Farness(a) = \frac{1}{4}(1+1+1+1) = 1$$

$$Farness(b) = \frac{1}{4}(1+2+2+2) = \frac{7}{4}$$

...



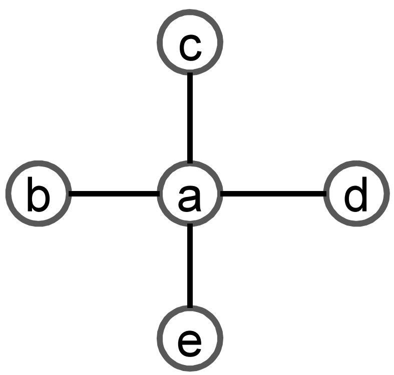
# CLOSENESS CENTRALITY

$$C_{closeness}(i) = \frac{1}{Farness(i)}$$

$$C_{closeness}(a) = 1/\left[\frac{1}{4}(1+1+1+1)\right] = 1$$

$$C_{closeness}(b) = 1/\left[\frac{1}{4}(1+2+2+2)\right] = \frac{4}{7}$$

…



# BETWEENNESS CENTRALITY

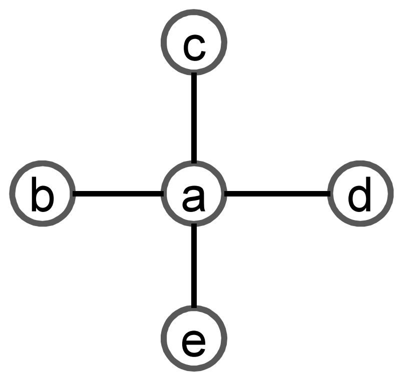
**Betweeness centrality**

- How many shortest paths go through an individual?

$C_{betweenness}(a) = 6$

$C_{betweenness}(b) = 0$

…

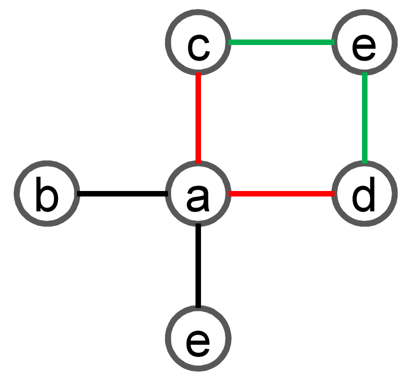


# BETWEENNESS CENTRALITY

**Betweeness centrality**

- How many shortest paths go through an individual?

What about multiple shortest paths? E.g. there are two shortest paths from c to d (one via a and another one via e)




Give each shortest path a weight inverse to how many shortest paths there are between two nodes.

# CENTRALITY

nwdegree
nwbetween
nwevcent
nwcloseness
nwkatz

# CENTRALIZATION

# CENTRALIZATION

How equally/unequally distributed are the centrality scores of all individuals?

- A network is highly centralized when one individual is very central and all others are not.
- A network is not centralized when all individuals have the same centrality score.

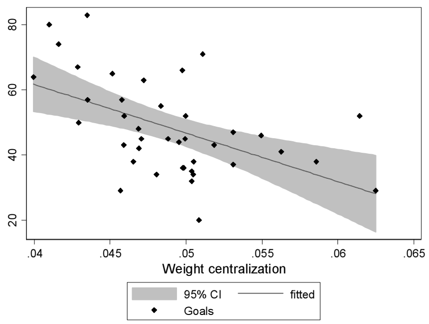
# CENTRALIZATION

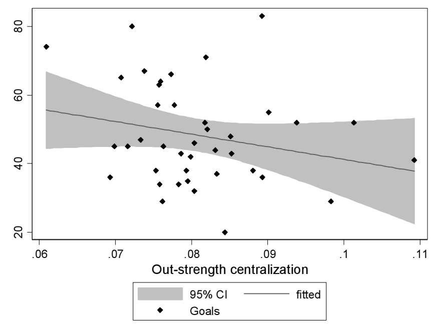
The general definition of centralization for non-weighted networks was proposed by Linton Freeman (1979).

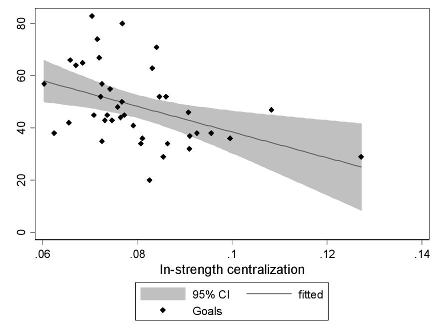
1. Calculate the sum in differences in centrality between the most central node in a network and all other nodes; and
2. Divide this quantity by the theoretically largest such sum of differences in any network of the same size.

$$C_x = \frac{\sum_i C_x(max) - C_x(i)}{\max\_sum}$$

# CENTRALIZATION



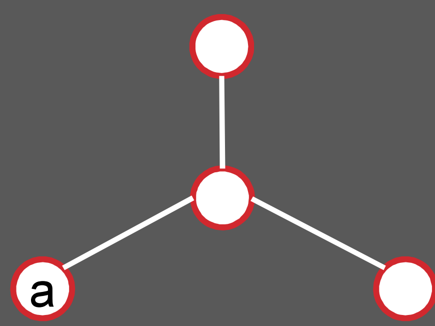



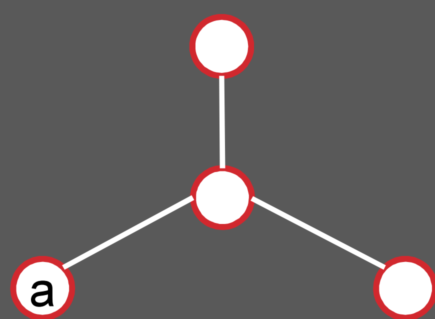

Grund, T. (2012) Network Structure and Team Performance: The Case of English Premier League Soccer Teams. *Social Networks,* Vol. 34, Issue 4, pp. 682-690.

the big quiz thing

What is the closeness centrality of node a?



What is the closeness centrality of node a?



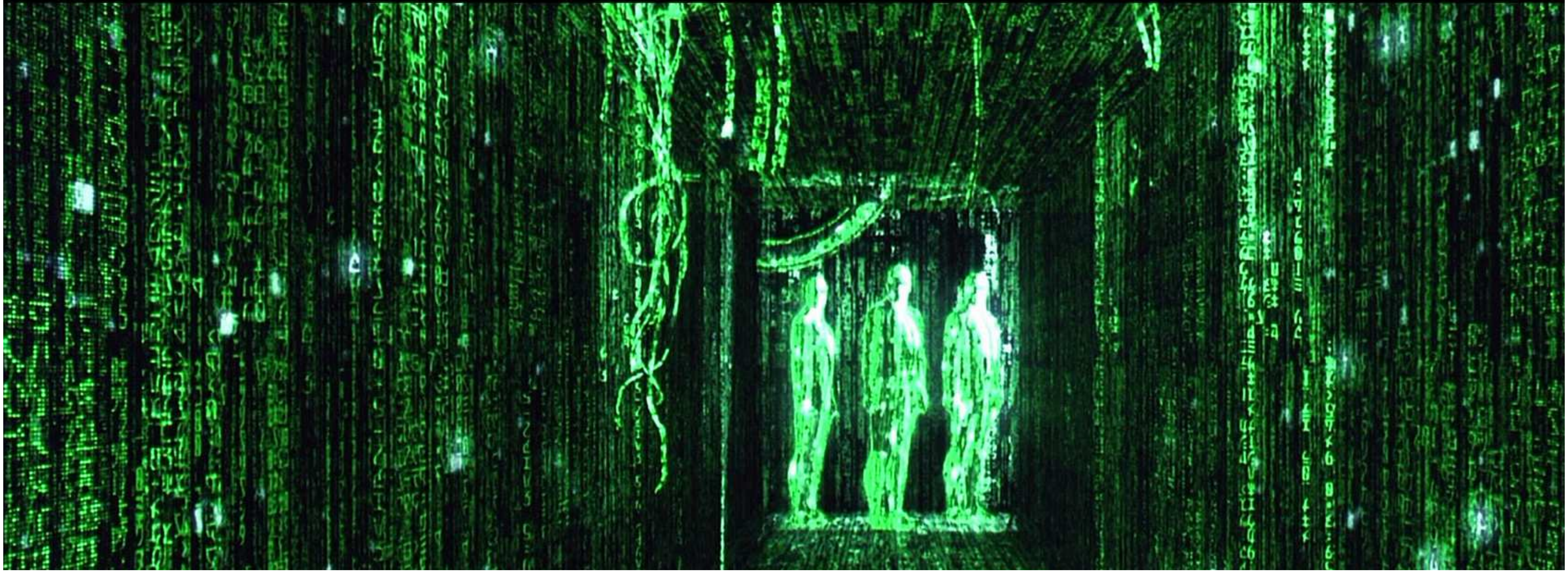

$$\frac{1}{\left[\frac{(1+2+2)}{3}\right]} = \frac{3}{5}$$

# CENTRALIZATION

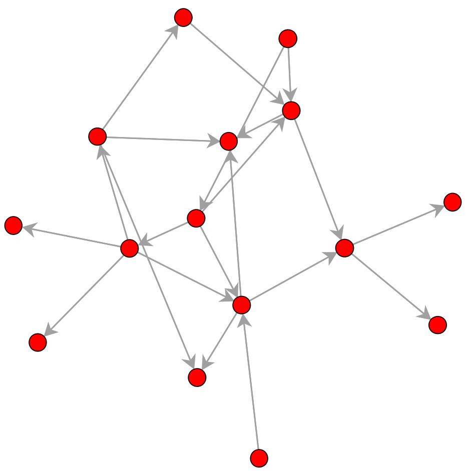
```
nwdegree
nwbetween
nwsummarize
```

- Load the **glasgow** data from the nwcommands-Server using `webnwuse`
- Calculate indegree centralization
- Tabulate and visualize the indegree distribution (check out `tab` and `his`
- Calculate betweenness centralization

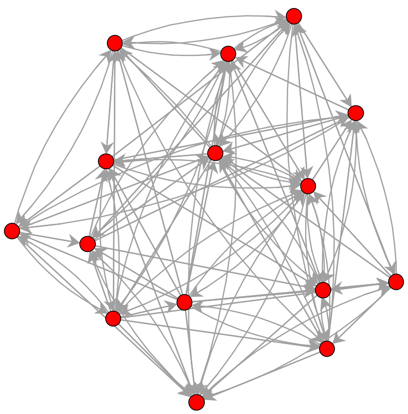
# SIMULATION



# RANDOM NETWORK



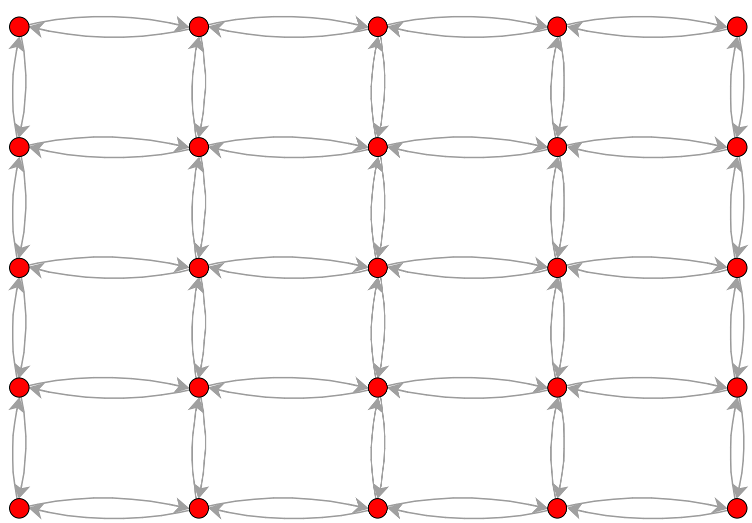
`nwrandom 15, prob(.1)`



`nwrandom 15, prob(.5)`

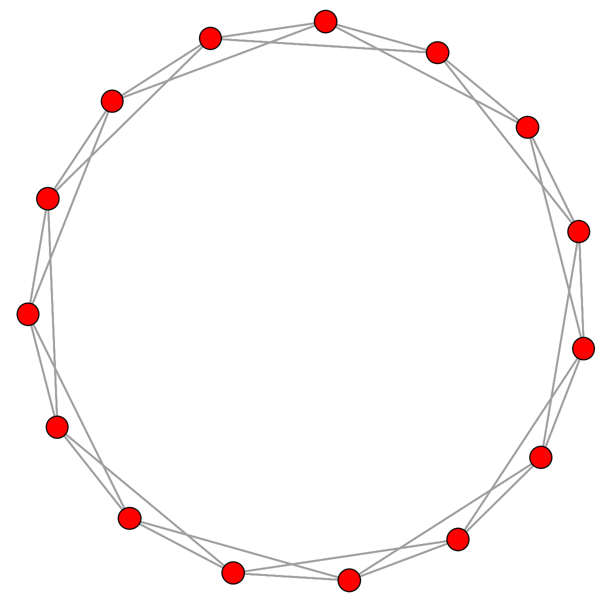
Each tie has the same probability to exist, regardless of any other ties.

# LATTICE



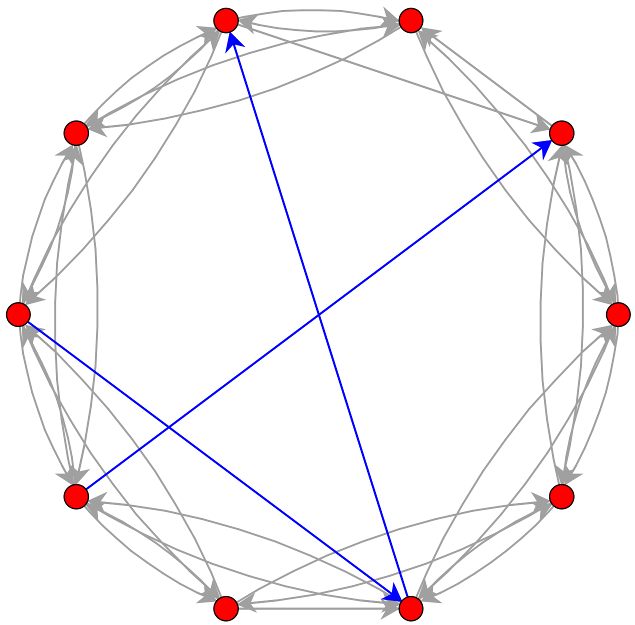
`nwlattice 5 5`

# RING LATTICE



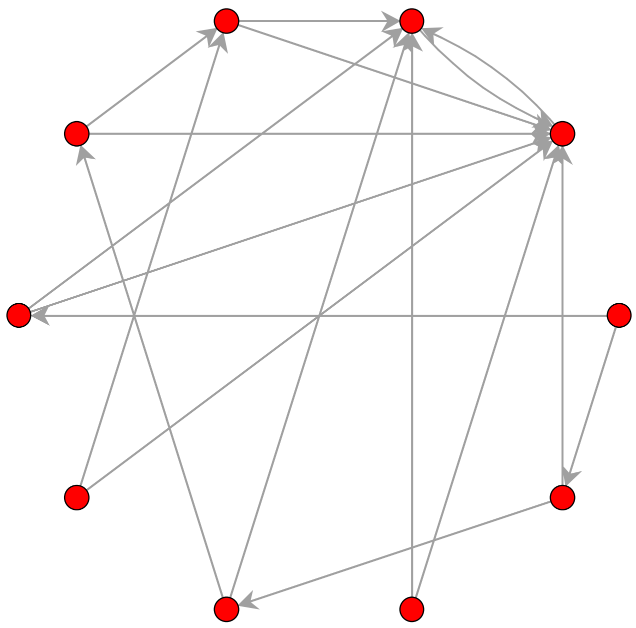
`nwring 15, k(2)`

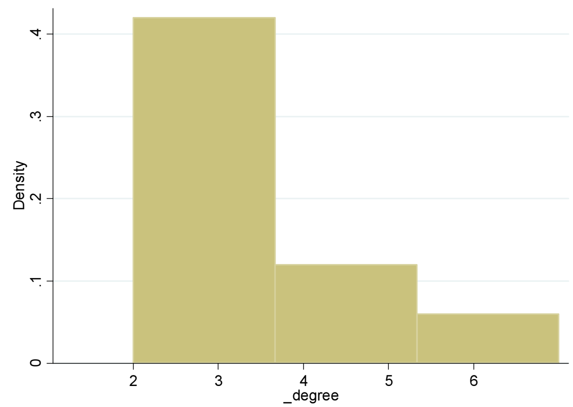
# SMALL WORLD NETWORK



`nwsmall 10, k(2) shortcuts(3)`

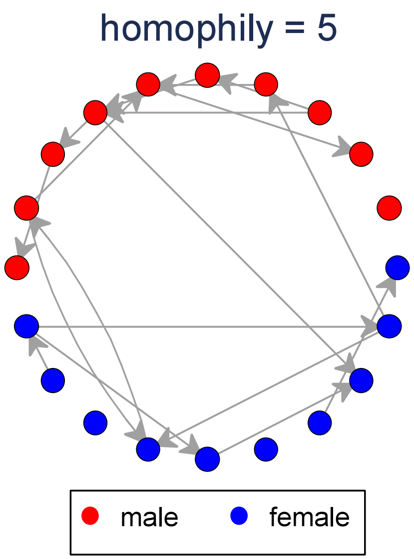
# PREFERENTIAL ATTACHMENT NETWORK

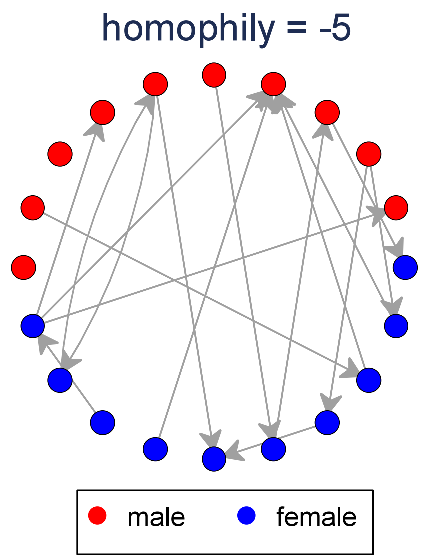





```
nwpref 10, prob(.5)
```

# HOMOPHILY NETWORK




```
nwhomophily gender, density(0.05) homophily(5)
```

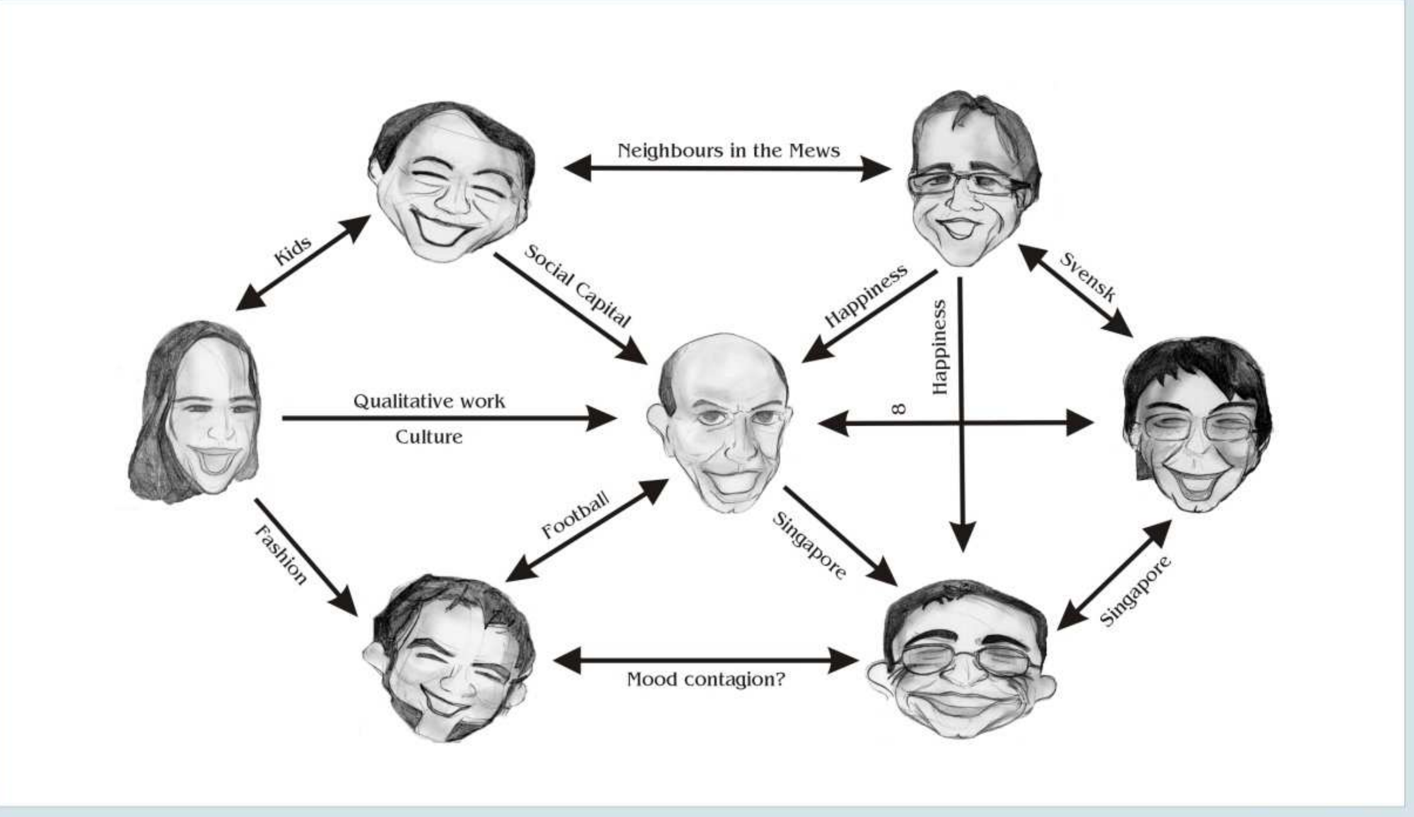
# SIMULATION

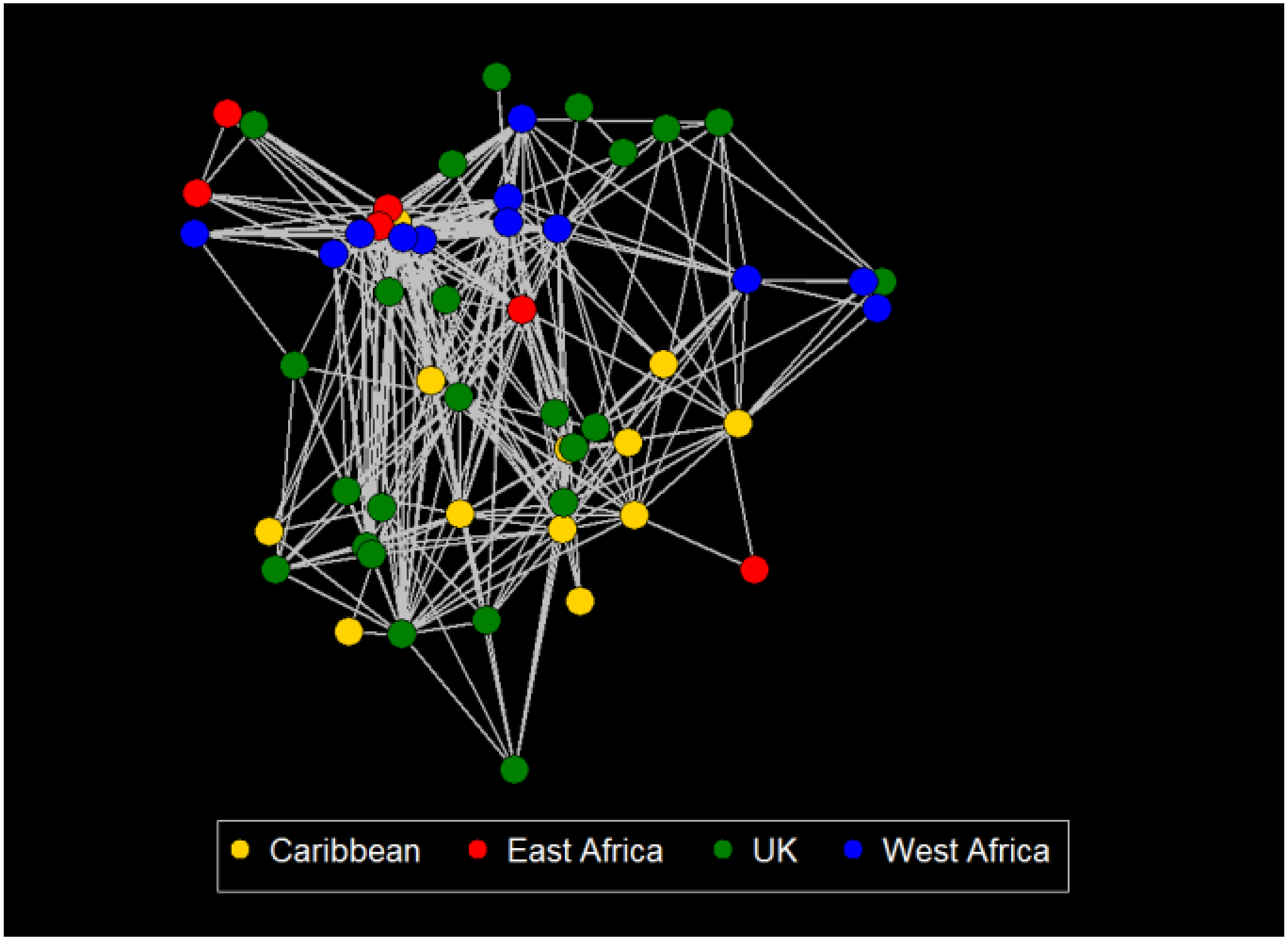
```
nwrandom  nwlattice
nwsmall   nwpref
nwring
nwhomophily
nwdyadprob
```

- Generate a random network with 50 nodes and <u>exactly</u> 245 ties.
- Summarize this network to check that you succeeded.

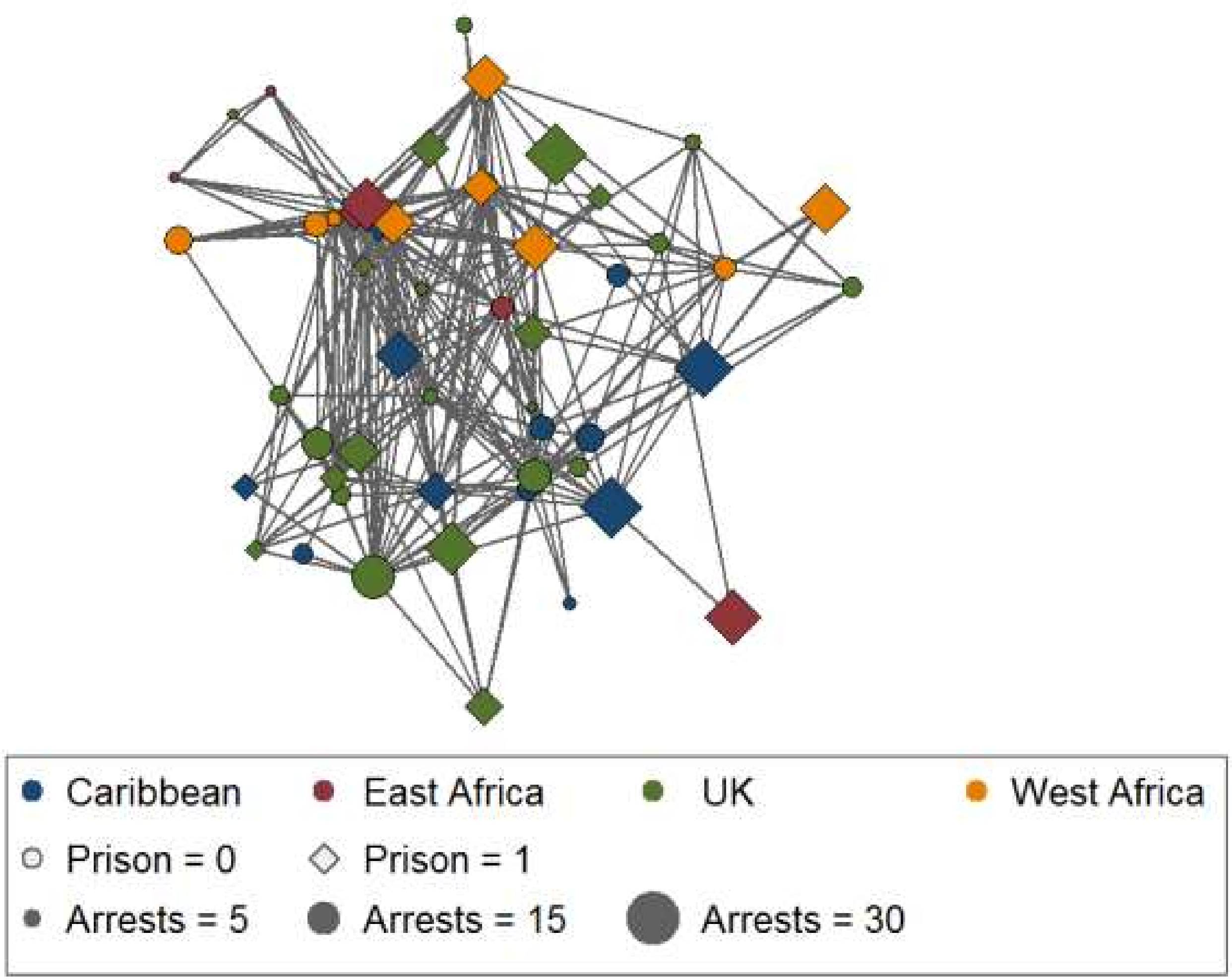
# VISUALIZATION

# Nuffield Network 2008

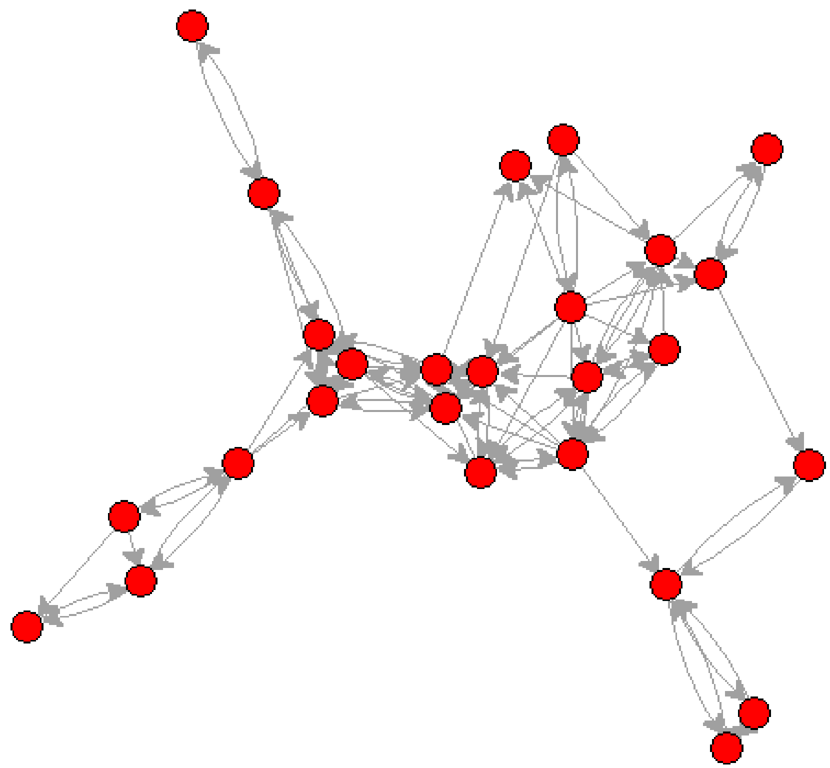




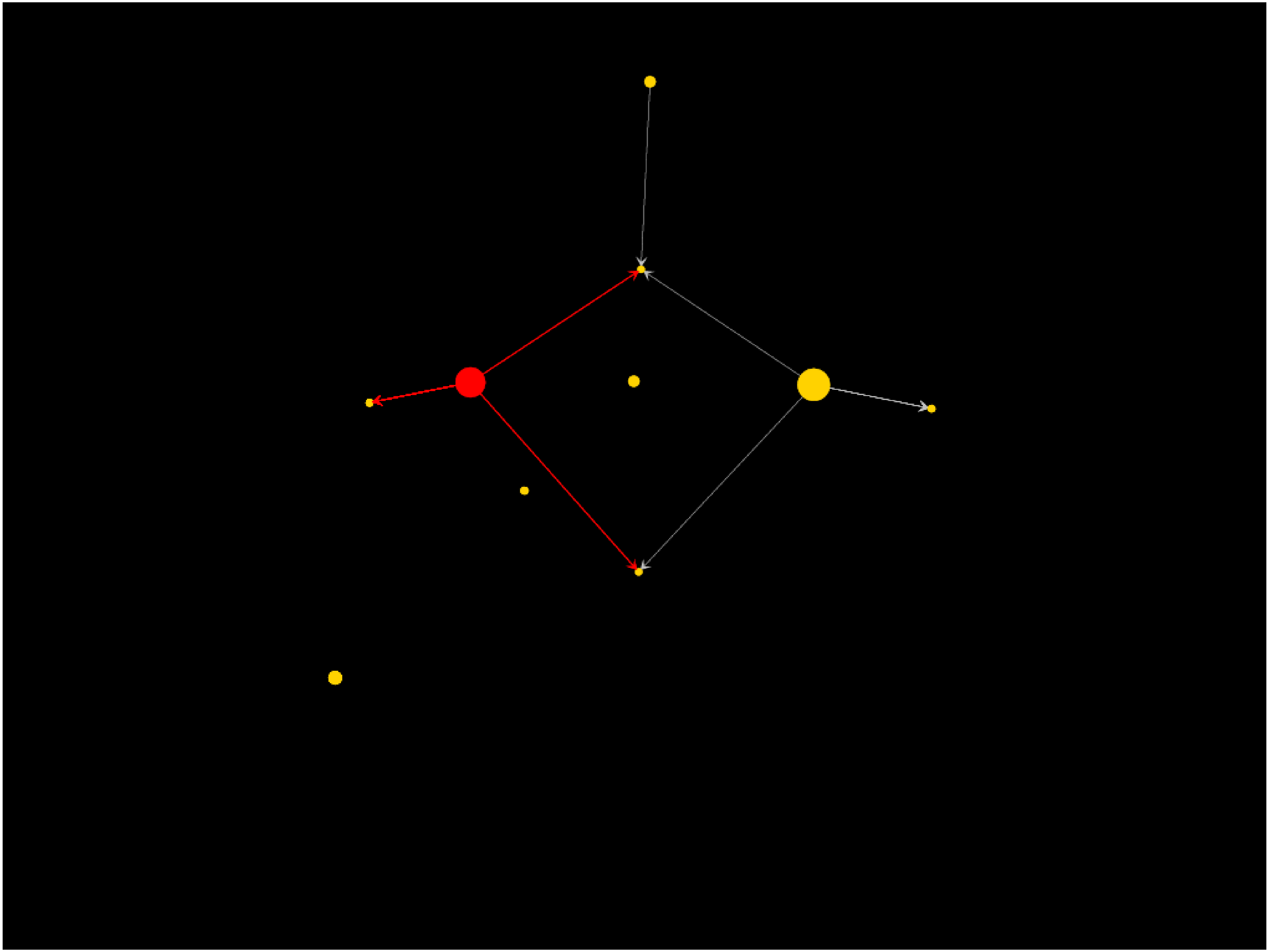

```
. webnwuse gang
. nwplot gang, color(Birthplace)
```

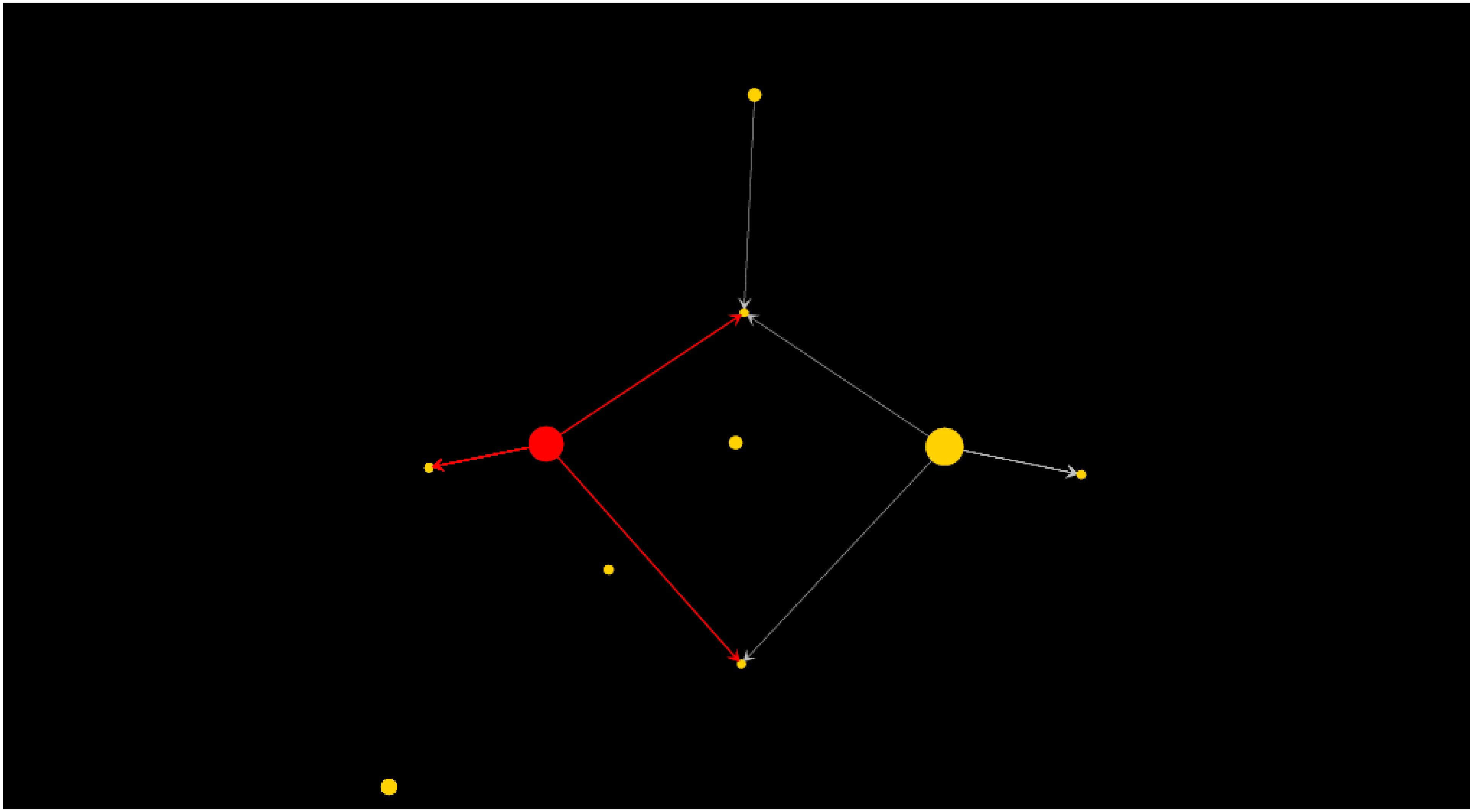



```
nwplot gang, color(Birthplace) symbol(Prison) size(Arrests)
```

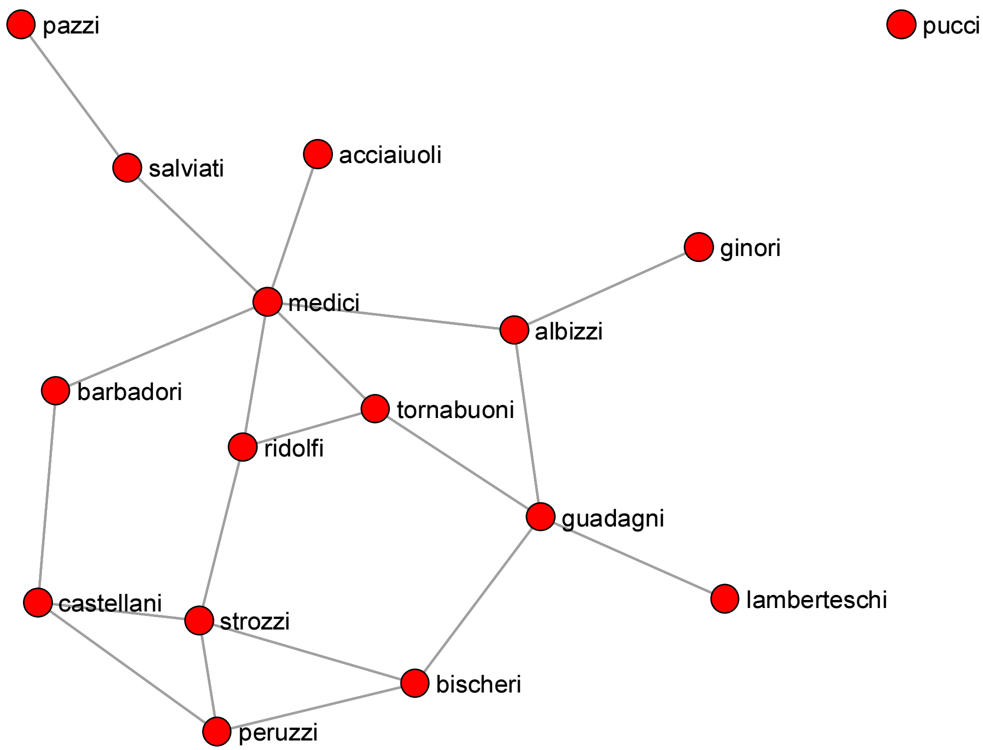


```
. webnwuse klas12
. nwmovie klas12_wave1-klas12_wave4
```

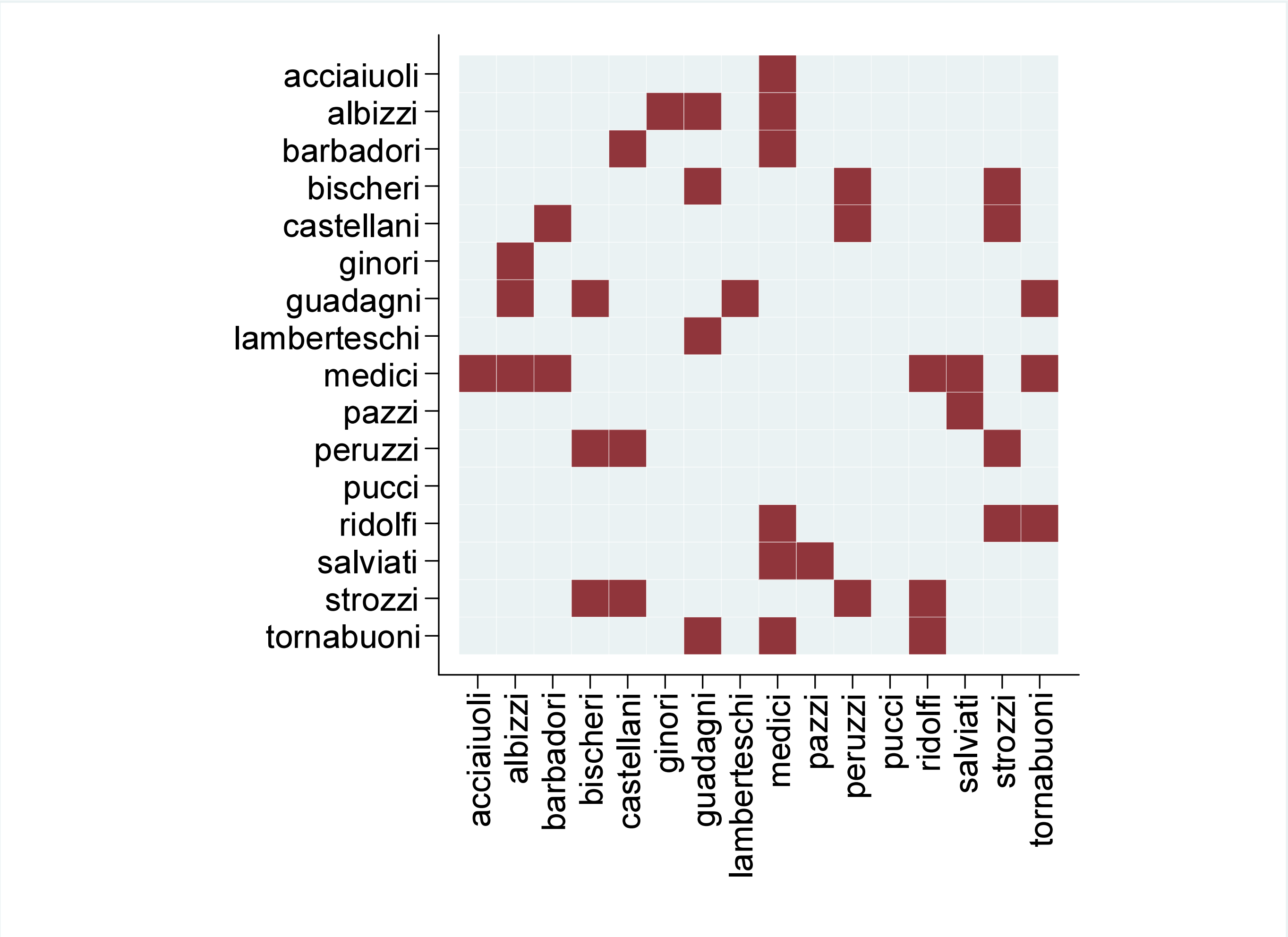




```
. nwmovie _all, colors(col_t*) sizes(siz_t*) edgecolors(edge_t*)
```



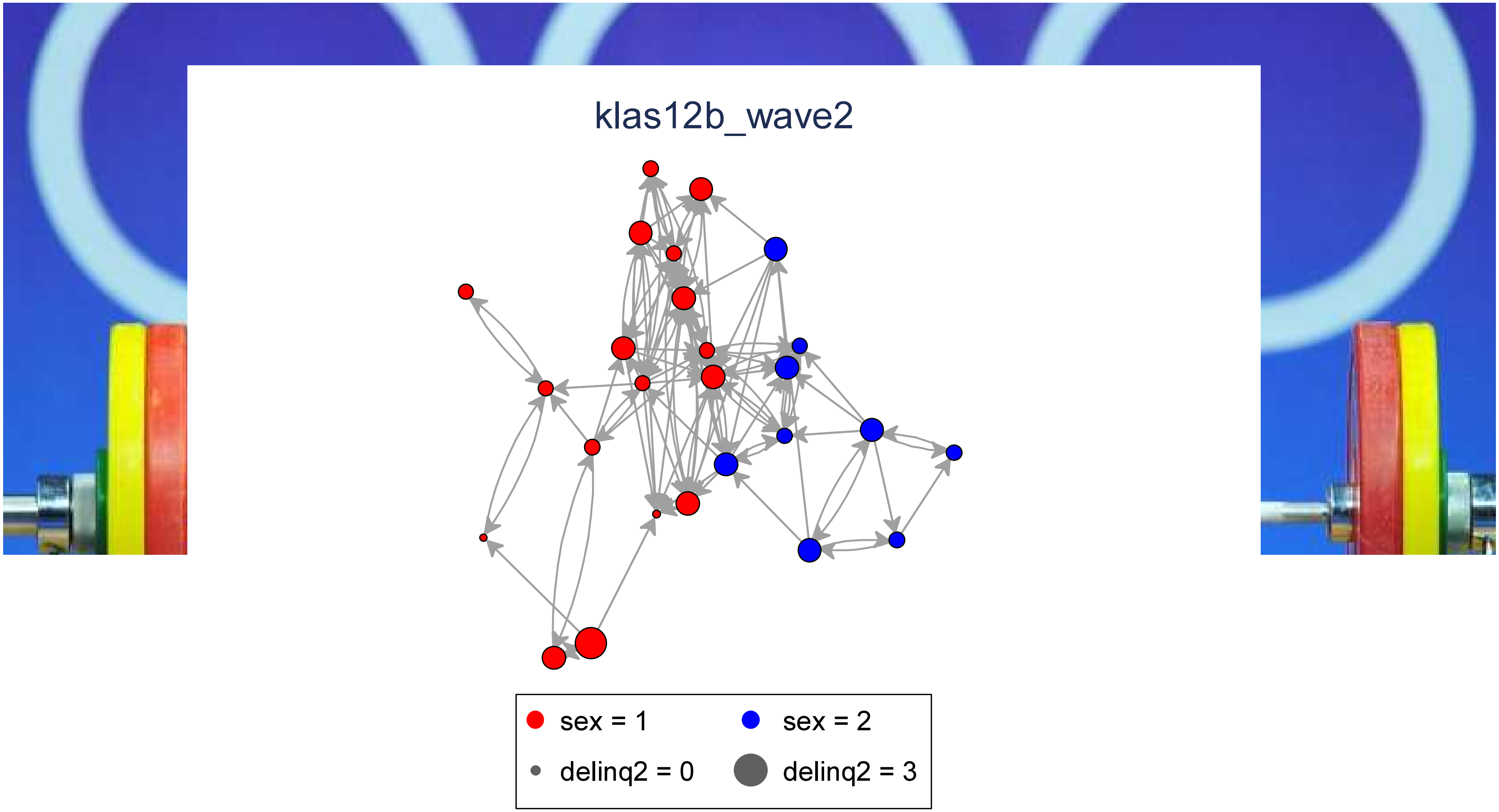

```
. webnwuse florentine
. nwplot flomarriage, lab
```




```
. nwplotmatrix flomarriage, lab
```

# VISUALIZATION

nwplot
nwplotmatrix
nwmovie




- Load the **klas12b** network data with `webnuse` and make the plot above using `nwplot`

# HYPOTHESIS TESTING

Is a particular network pattern more (or less) prominent than **expected**?

# (NON-)RANDOMNESS

Even random networks have non-random features. What is random anyway?

**Question:** how many triads is one likely to observe in a network with 3 nodes and 2 undirected ties?

**Question:** How many triads is one likely to observe in a network with 3 nodes and 3 undirected ties?

# (NON-)RANDOMNESS

Even random networks have non-random features. What is random anyway?

**Question:** how many triads is one likely to observe in a network with 3 nodes and 2 undirected ties?

= 0

**Question:** How many triads is one likely to observe in a network with 3 nodes and 3 undirected ties?

=1

# PROBLEM

Even complete randomness produces certain network patterns in networks, e.g. because of the size of sub-groups.

**General strategy:**

1. Calculate a network statistic that you are interested in.
2. Think about the properties of the network that you want to conserve.
3. Generate many “random” networks that have the same properties as the “observed” network.
4. Calculate the network statistic on these “conditional random networks” and compare this baseline distribution against the actually observed network statistic in the “observed” network.

**1** **Test-statistic**

e.g. number of triads, number of reciprocal ties, number of ties between similar individuals…

**2** **Distribution of test-statistic under null hypothesis**

e.g. distribution of triads we can expect when there is no clustering…

1

**Test-statistic**

e.g. number of triads, number of reciprocal ties, number of ties between similar individuals…

2

**Distribution of test-statistic under null hypothesis**

e.g. distribution of triads we can expect when there is no clustering…

# CONDITIONAL UNIFORM GRAPHS

# CONDITIONAL UNIFORM GRAPHS

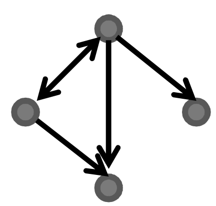
- Generate random networks with the same size, density, or dyad census as the observed network and then calculate the test-statistic (transitivity) on these ‘conditional uniform graphs’.

- Force the network to have certain properties, e.g. density = condition on density.

# TIE PROBABILITIES

Density of the network = $\frac{2M+A}{n(n-1)}$

Reciprocity of the network = $\frac{2M}{2M+A}$



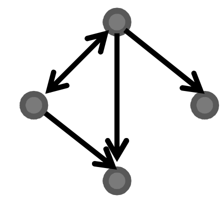
M=1, A=2, N=1 (121)

Density = $\frac{2*1+3}{4(4-1)} = \frac{5}{12} = 0.416$

Reciprocity = $\frac{2*1}{2*1+3} = \frac{2}{5} = 0.4$

# TIE PROBABILITIES


Now, assume two nodes $i, j$ are randomly sampled from this little network and look at tie $y_{ij}$.

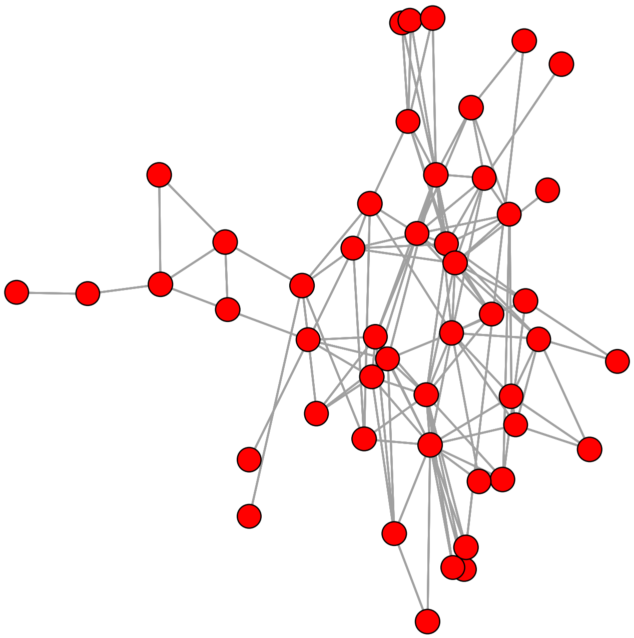


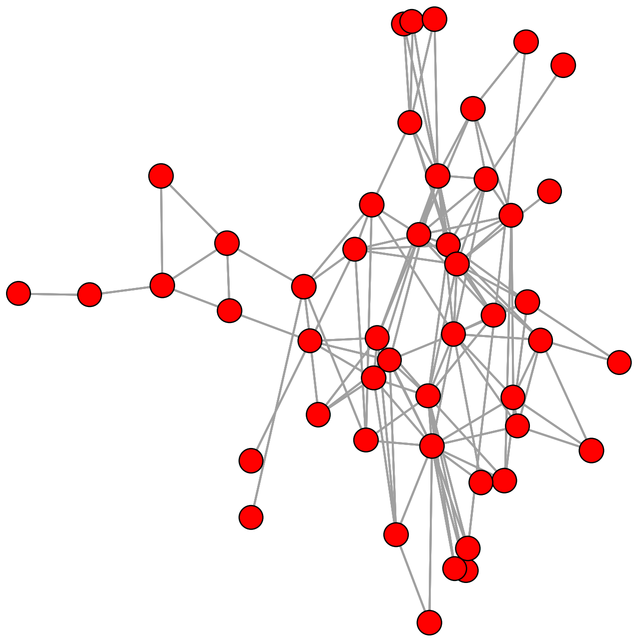
M=1, A=2, N=1 (121)

- What is the probability for $Pr(y_{ij} = 1)$?
  - There are 12 possible ties. And 5 of these 12 are realized. That means:

$$Pr(y_{ij} = 1) = \frac{5}{12} = 0.416$$

**Question:** Is there more or less clustering (triads) than expected?





transitivity score is 0.36

**Is this a lot?**

Problem: We do not know how much transitivity we should expect by chance given a certain number of ties in the network.

1

**Test-statistic**

Here, the transitivity score is 0.36

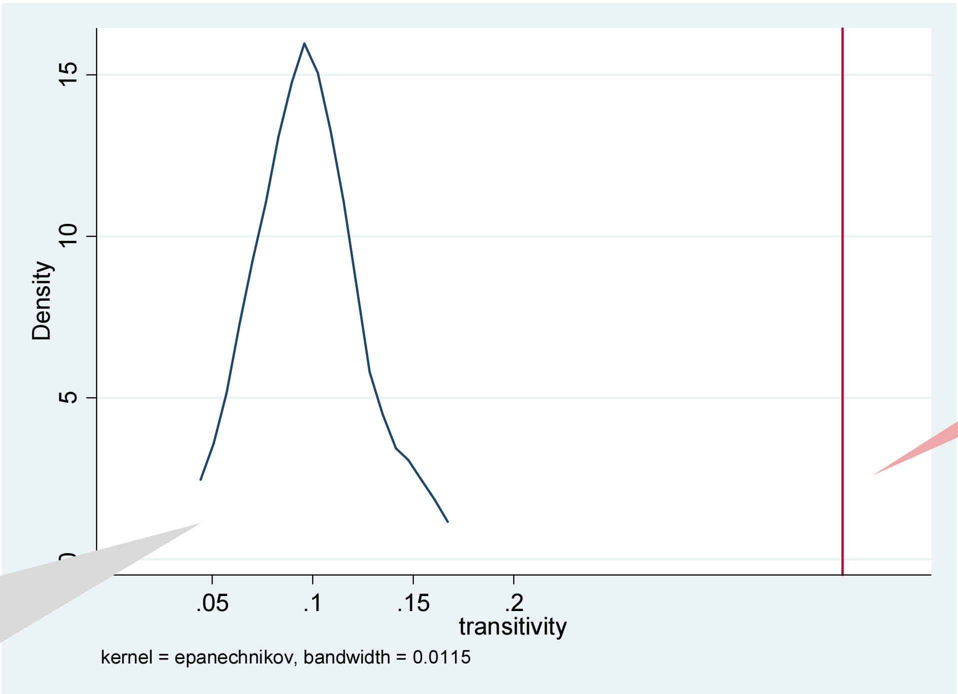
2

**Distribution of test-statistic under null hypothesis**

**???**

# CONDITIONAL UNIFORM GRAPHS

- Generate random networks with the same size, density, or dyad census as the observed network and then calculate the test-statistic (transitivity) on these 'conditional uniform graphs'.




transitivity of random networks that have same density as the gang network

transitivity of the gang network

```
. webnwuse gang2
. nwsummarize gang, detail
```

```
Network name:  gang
Network id:  21
Directed: false
Nodes: 54
Edges: 133
Minimum value:  0
Maximum value:  1
Density:  .092941998602376
Reciprocity:  .092941998602376
Transitivity:  .3635371179039302
Betweenness centralization:  .1066829629553841
Degree centralization: .1973875181422351
```

transitivity of the **gang** network

```
. webnwuse gang2
. nwsummarize gang, detail
```

```
Network name:  gang
Network id:  21
Directed: false
Nodes: 54
Edges: 133
Minimum value:  0
Maximum value:  1
Density:  .092941998602376
Reciprocity:  .092941998602376
Transitivity:  .3635371179039302
Betweenness centralization:  .1066829629553841
Degree centralization: .1973875181422351
```

density of the observed network

```
. nwclear
. nwrandom 54, prob(.092) undirected ntimes(20)
. nwsummarize _all, detail save(myfile)

...

. use myfile, clear
. kdensity transitivity, xline(.363) xscale(range(0 .4))
```

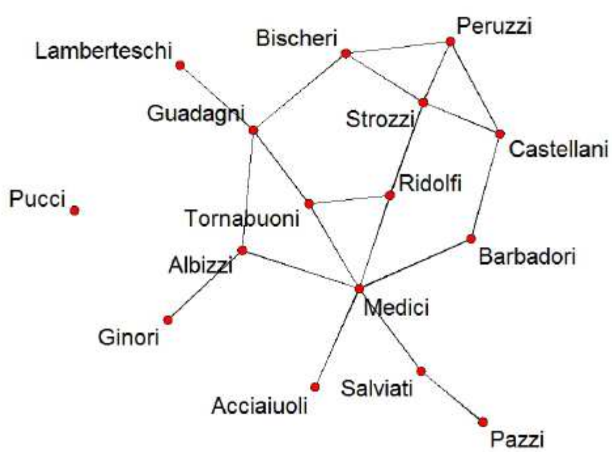
# CONDITIONAL UNIFORM GRAPHS

`nwrandom`

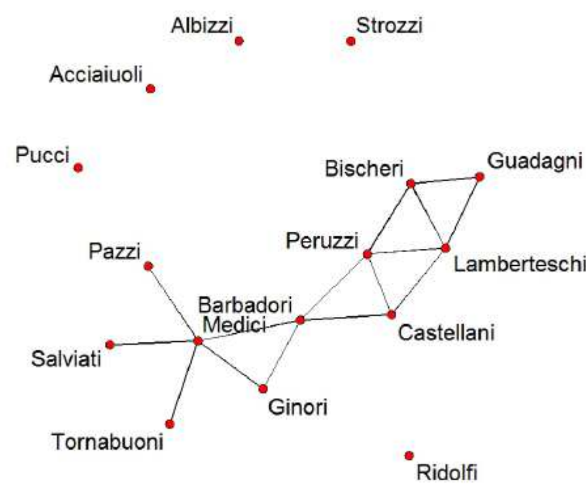
`nwsummarize`

# PERMUTATION TESTS
# QAP

**Question:** Is there more or less correlation between these two networks than expected?




Florentine Marriage Network

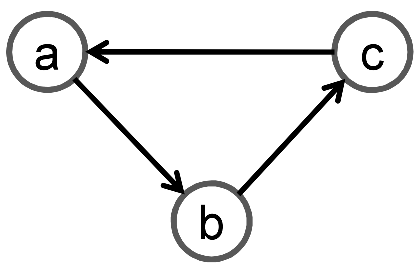



Florentine Business Ties

Padgett, J. and Ansell, C. (1993) Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology* 98: 1259-1319

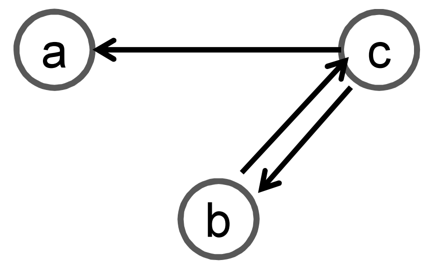
# GRAPH CORRELATION

**Network 1**




$$
\begin{array}{c c}
 & \begin{array}{ccc} a & b & c \end{array} \\
\begin{array}{c} a \\ b \\ c \end{array} & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}
\end{array}
$$

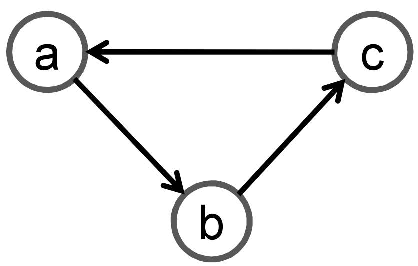
**Network 2**




$$
\begin{array}{c c}
 & \begin{array}{ccc} a & b & c \end{array} \\
\begin{array}{c} a \\ b \\ c \end{array} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}
\end{array}
$$

# GRAPH CORRELATION

**Network 1**



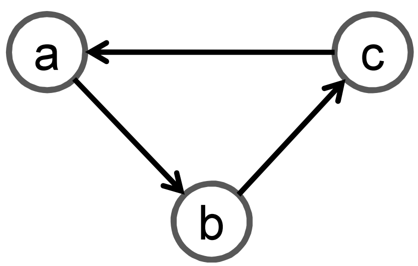

$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array} =$$

Transform adjacency matrix in a dataset of dyads.

| row | col | net1 |
|---|---|---|
| a | b | 1 |
| a | c | 0 |
| b | a | 0 |
| b | c | 1 |
| c | a | 1 |
| c | b | 0 |

# GRAPH CORRELATION

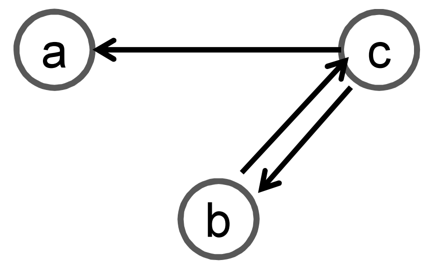
**Network 1**




$$\begin{array}{c c} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array} =$$

| net1 |
|---|
| 1 |
| 0 |
| 0 |
| 1 |
| 1 |
| 0 |

**Network 2**

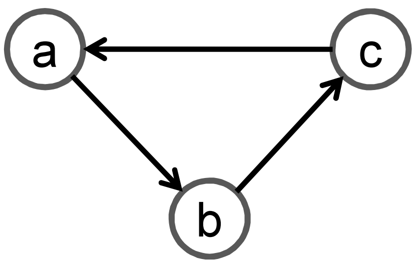



$$\begin{array}{c c} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \end{array} =$$

| net2 |
|---|
| 0 |
| 0 |
| 0 |
| 1 |
| 1 |
| 1 |

# GRAPH CORRELATION

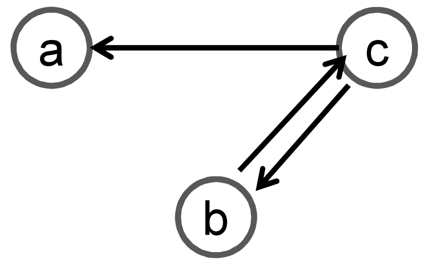
**Network 1**




$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array} =$$

| net1 |
|---|
| 1 |
| 0 |
| 0 |
| 1 |
| 1 |
| 0 |

**Network 2**

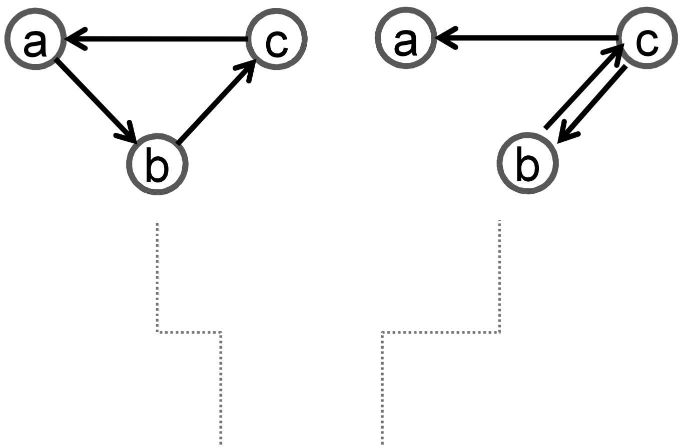



$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \end{array} =$$

| net2 |
|---|
| 0 |
| 0 |
| 0 |
| 1 |
| 1 |
| 1 |

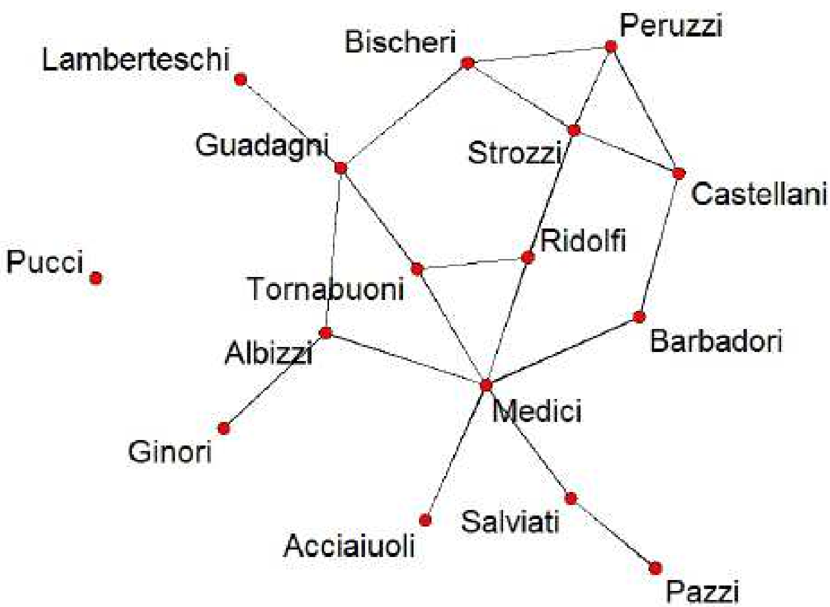
# GRAPH CORRELATION

| row | col | net1 | net2 |
|---|---|---|---|
| a | b | 1 | 0 |
| a | c | 0 | 0 |
| b | a | 0 | 0 |
| b | c | 1 | 1 |
| c | a | 1 | 1 |
| c | b | 0 | 1 |

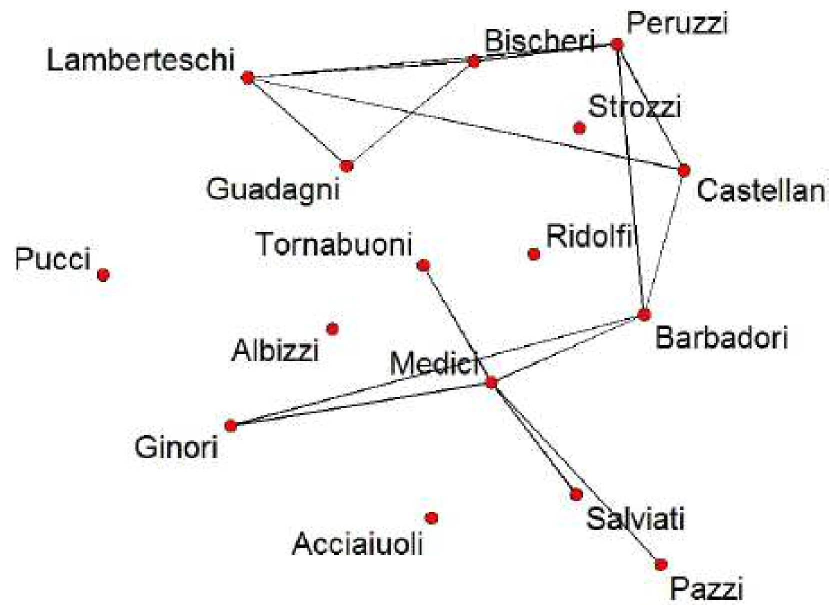



$$corr(net1, net2) = 0.333$$

# GRAPH CORRELATION




Florentine Marriage Network




Florentine Business Ties

$$corr_{obs} = 0.372$$

## Is this a lot?

Problem: We do not know how much correlation we should expect by chance given the marriage and the business network!

1 **Test-statistic**

$$corr_{obs} = 0.372$$

2 **Distribution of test-statistic under null hypothesis**

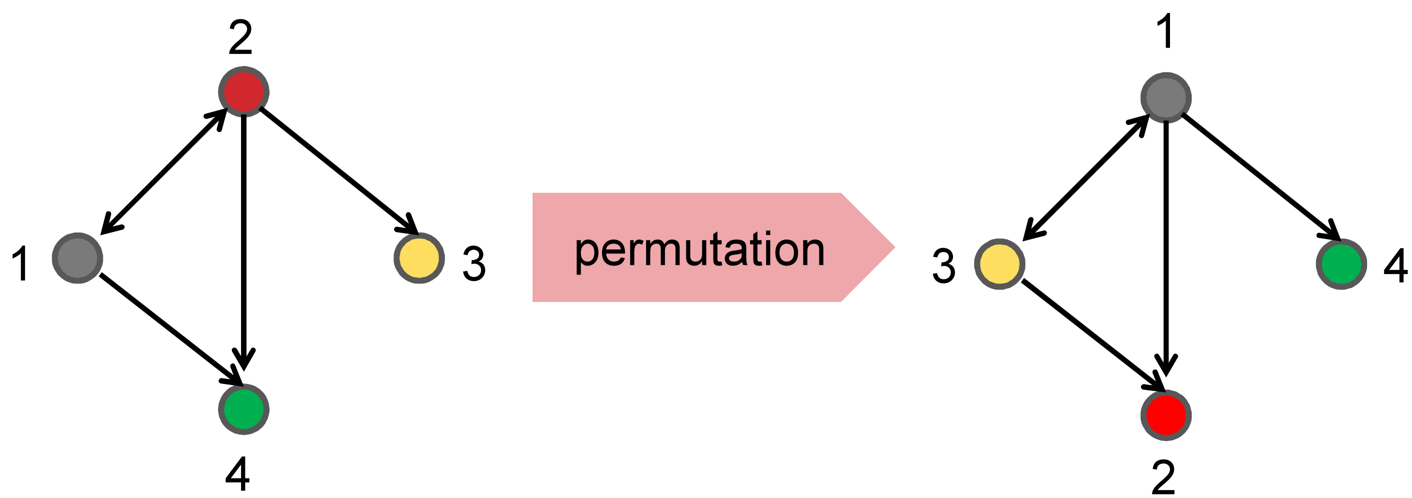
$$corr_{random} = ??$$

# QUADRATIC ASSIGNMENT PROCEDURE

- Scramble the network by permuting the actors (randomly re-label the nodes), i.e. the actual network does not change, however, the position each node takes does.
- Re-calculate the test-static on the permuted networks and compare it with test-statistic on the unscrambled network.

***Network structure is 'controlled' for. Keeps dependencies.***

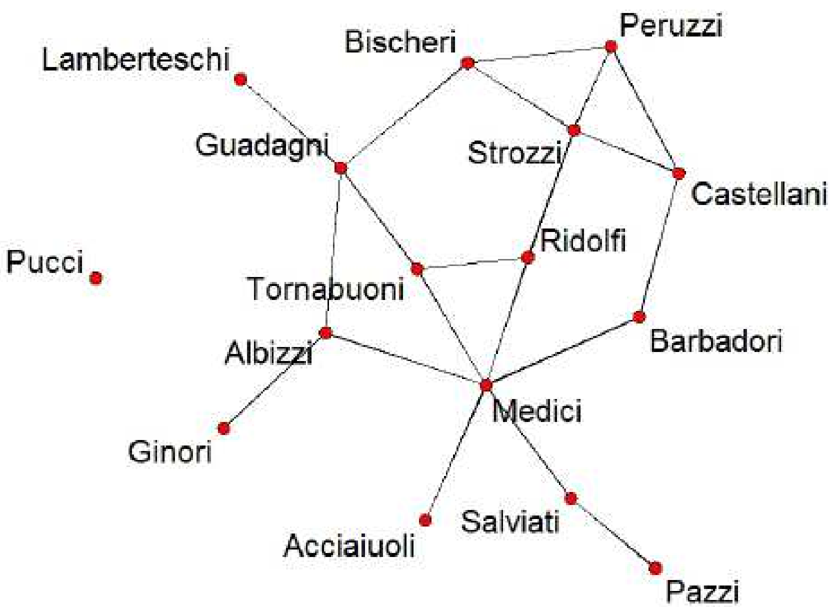
# PERMUTATION TEST



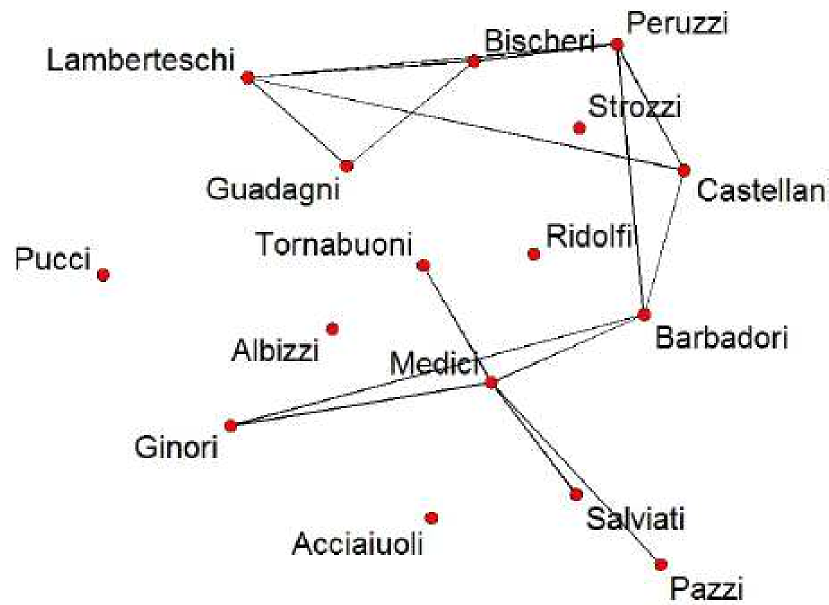

| | | | |
|---|---|---|---|
| - | 1 | 0 | 1 |
| 1 | - | 1 | 1 |
| 0 | 0 | - | 0 |
| 0 | 0 | 0 | - |

| | | | |
|---|---|---|---|
| - | 1 | 1 | 1 |
| 0 | - | 0 | 0 |
| 1 | 1 | - | 0 |
| 0 | 0 | 0 | - |

# GRAPH CORRELATION



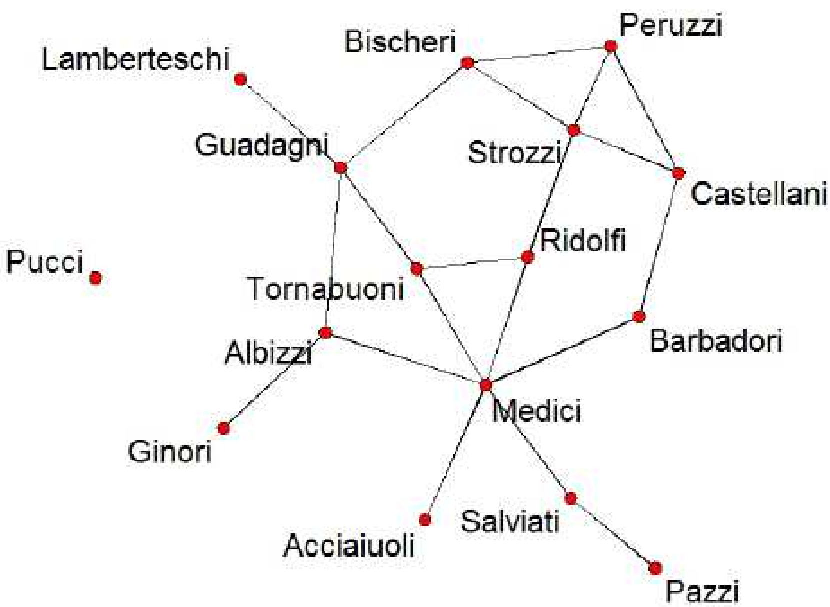

Florentine Marriage Network



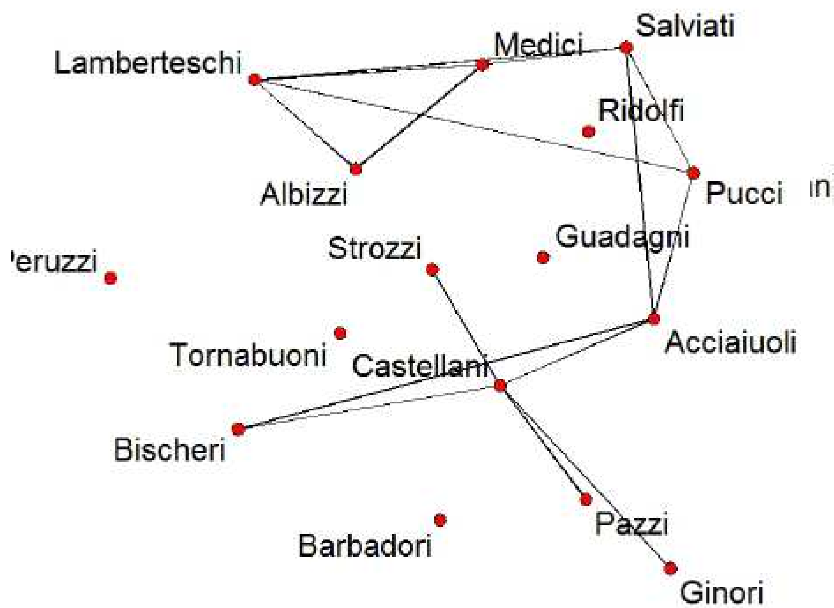

Florentine Business Ties

$$corr_{obs} = 0.372$$

# GRAPH CORRELATION



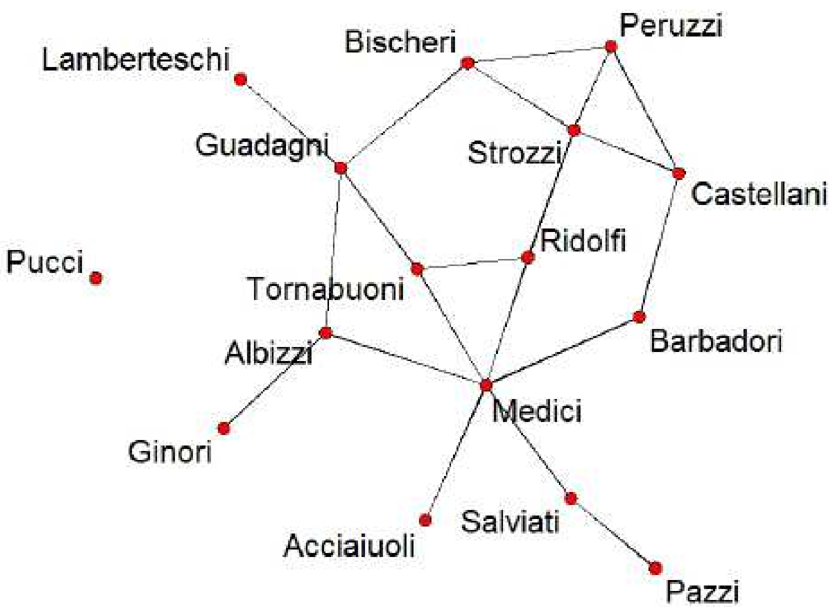

Florentine Marriage Network



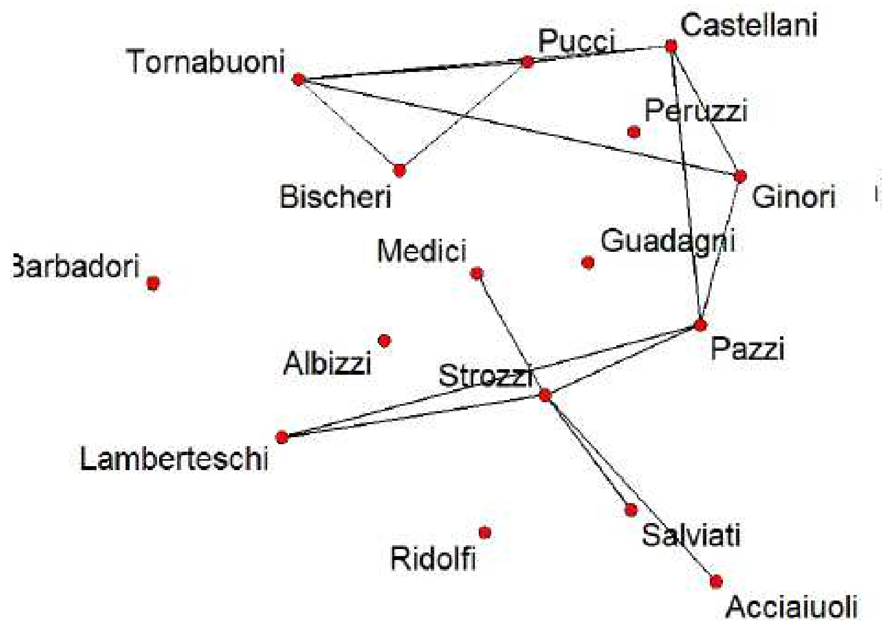

Florentine Business Ties

$$corr = -0.034$$

# GRAPH CORRELATION



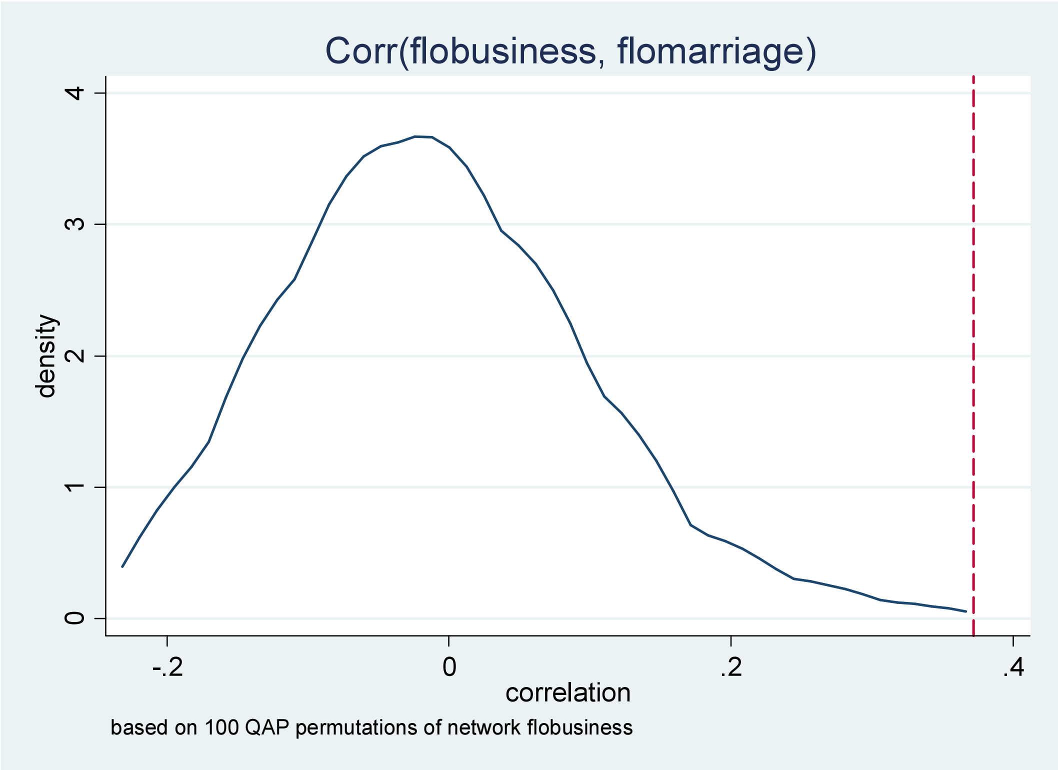

Florentine Marriage Network




Florentine Business Ties

$$corr = -0.101$$

# GRAPH CORRELATION

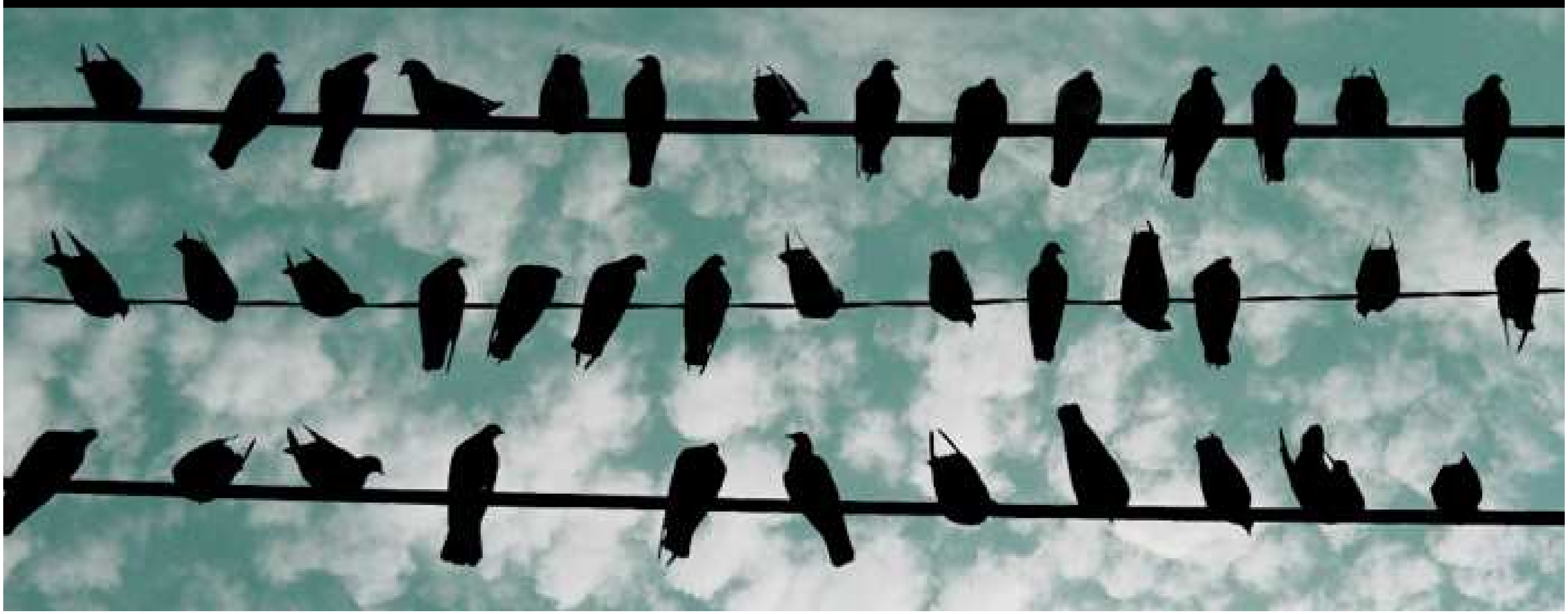



```
nwcorrelate flobusiness flomarriage, permutations(100)
```

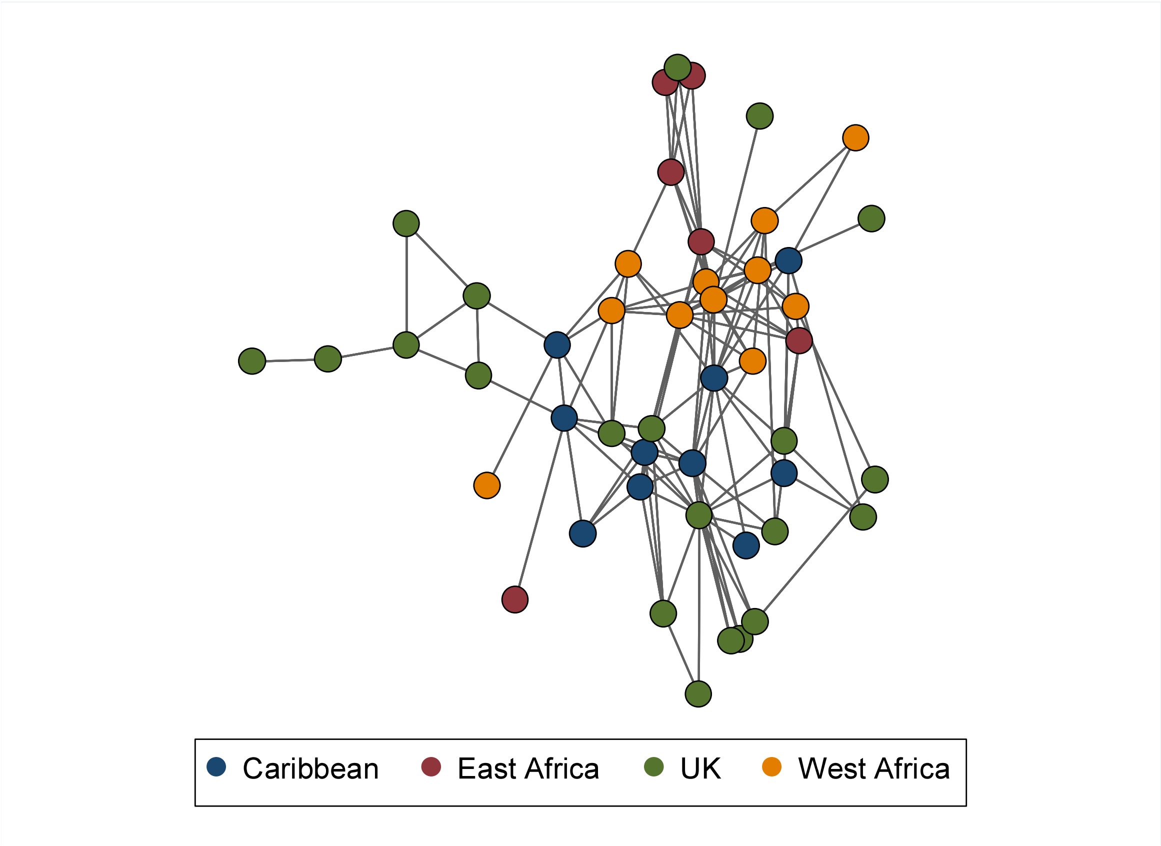
# PERMUTATION TESTS QAP

```
nwcorrelate
nwpermute
```

# HOMOPHILY



**Question:** Are co-offending ties between gang members from the same ethnicity more likely than ties between gang members from different ethnicities?



# HOMOPHILY

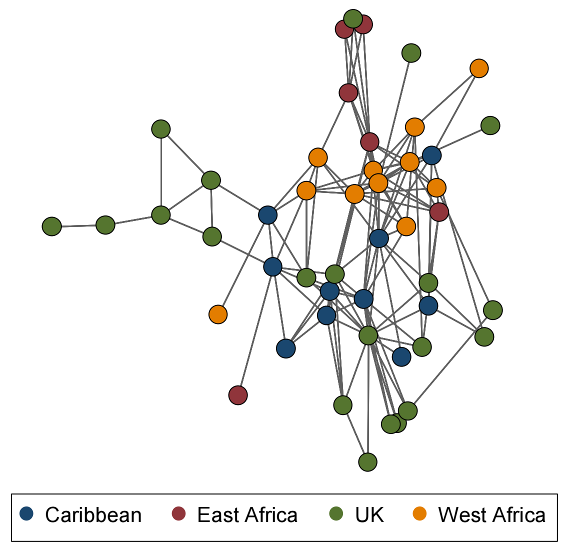
**Homophily = tendency of similar people to associate with each other**

- For example ties are more likely to form between people with the same age, education, occupation, religion, income and so on (McPherson, Smith-Lovin, and Cook 2001), .
- One of the most striking empirical regularities in social life.

# HOMOPHILY

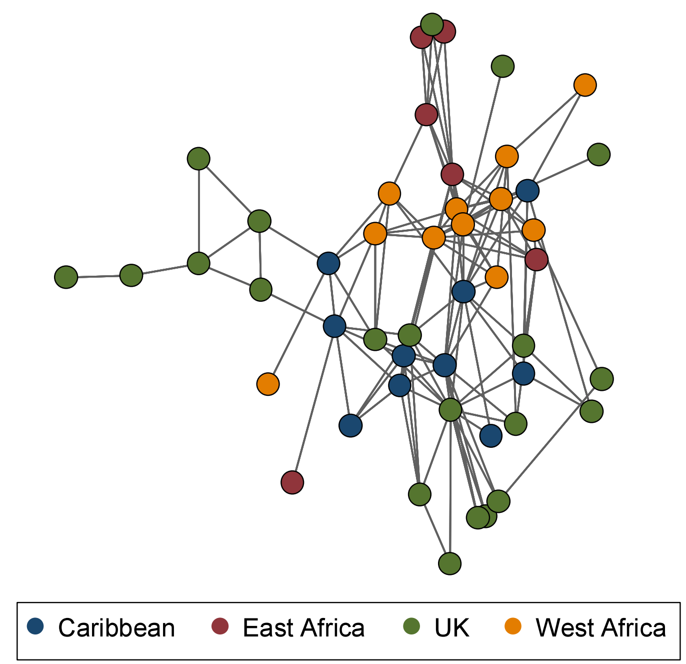
Empirically observed homophily statistics

| | Observed count |
|---|---|
| | |
| Match(ethnicity) | 63 |
| Match(British) | 23 |
| Match(Jamaican) | 14 |
| Match(Somali) | 6 |
| Match(West African) | 20 |

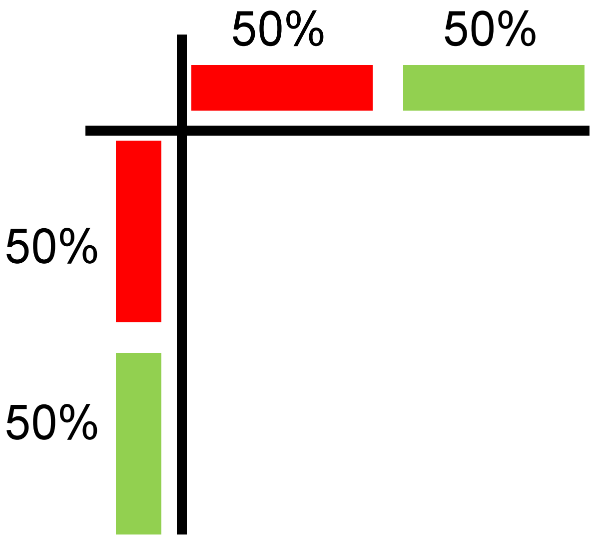


# ADJACENCY MATRIX

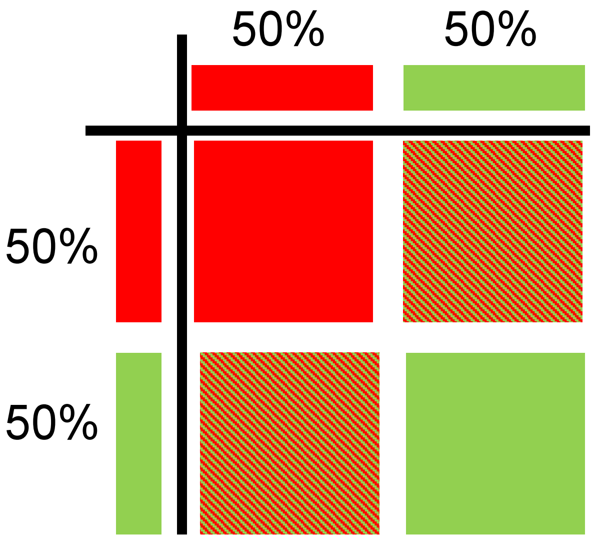
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



# GROUP SIZE MATTERS

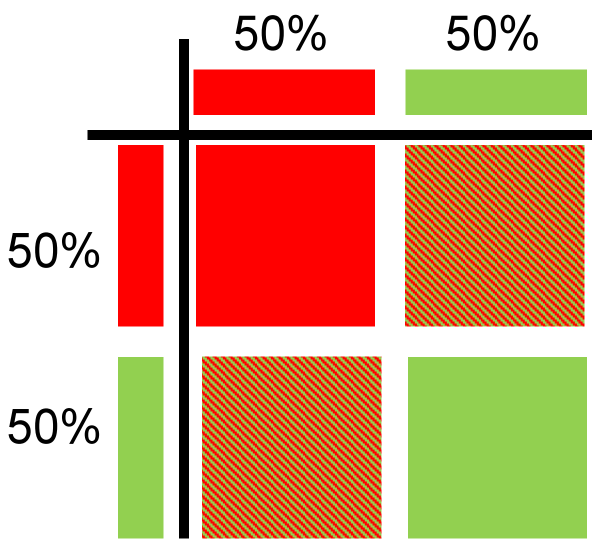


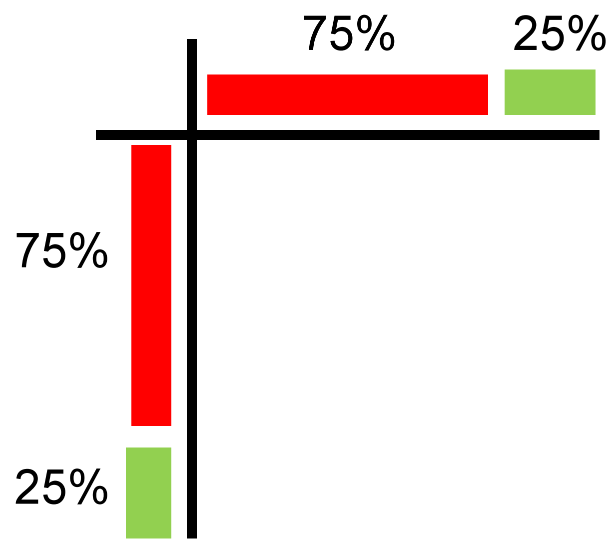
# GROUP SIZE MATTERS




If ties are assigned at random, 50% of the ties should be between similar individuals.

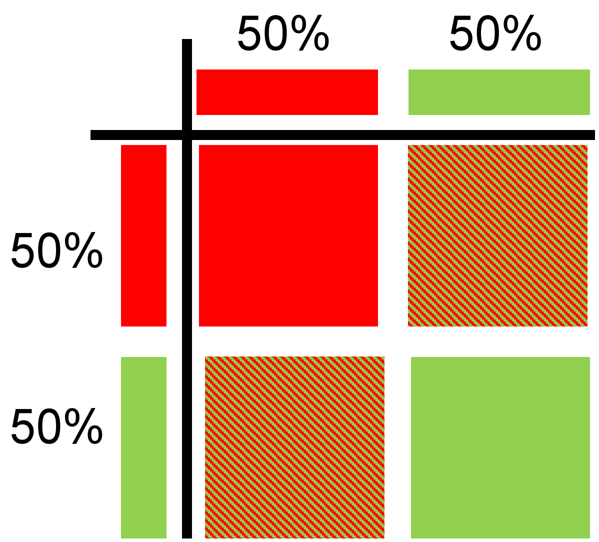
# GROUP SIZE MATTERS



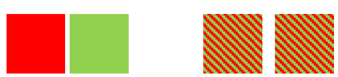



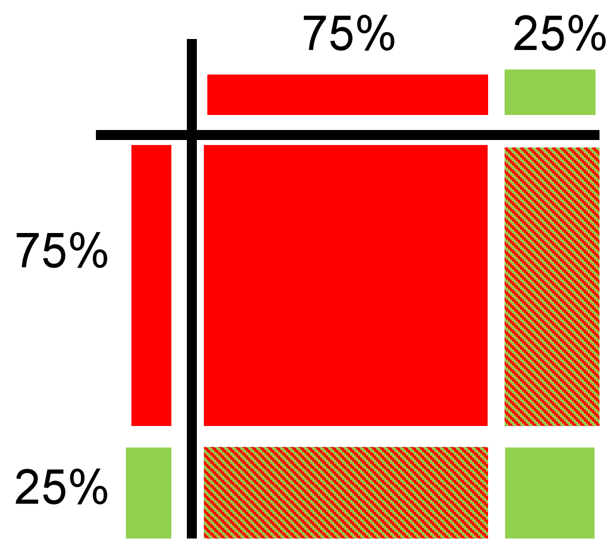

If ties are assigned at random, 50% of the ties should be between similar individuals.

# GROUP SIZE MATTERS

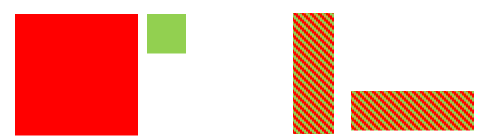



If ties are assigned at random, 50% of the ties should be between similar individuals.





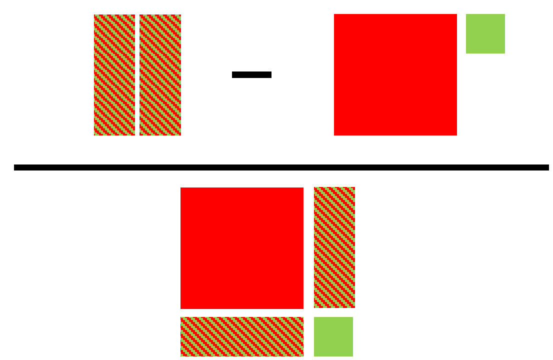

If ties are assigned at random, 62.5% of the ties should be between similar individuals.



# E-I INDEX

- E-I index is the number of ties external to the groups minus the number of ties that are internal to the group divided by the total number of ties.
- This value can range from 1 to -1.

$$EI = \frac{E - I}{T} =$$

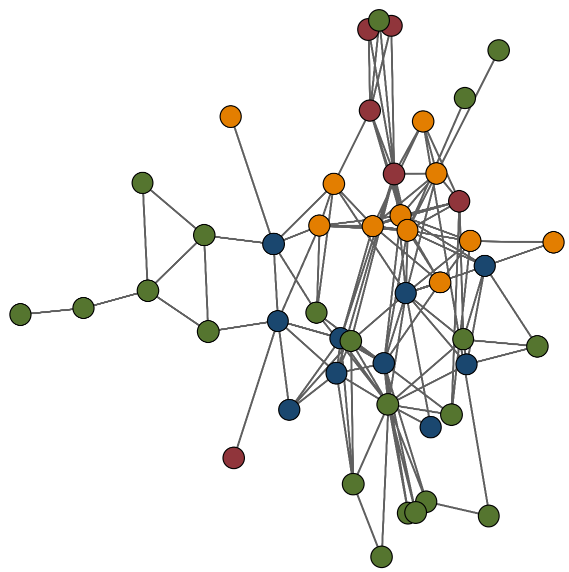


```
. nwtabulate gang Birthplace

   Network:   gang              Directed: false
   Attribute:   Birthplace

       The network is undirected.
       The table shows two entries for each edge.
```

| Birthplace | Caribbean | East Afri | UK | West Afri | Total |
|---|---|---|---|---|---|
| Caribbean | 28 | 4 | 27 | 16 | 75 |
| East Africa | 4 | 12 | 5 | 12 | 33 |
| UK | 27 | 5 | 46 | 6 | 84 |
| West Africa | 16 | 12 | 6 | 40 | 74 |
| Total | 75 | 33 | 84 | 74 | 266 |

E-I Index: .0526315789473684

# ARE THERE MORE TIES BETWEEN MEMBERS WITH THE SAME ETHNICITY?

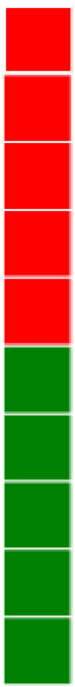


Network name: **gang**
Attribute: **same_Birthplace**

Correlation: **.124927833546448**

# EXPAND VARIABLE TO NETWORK

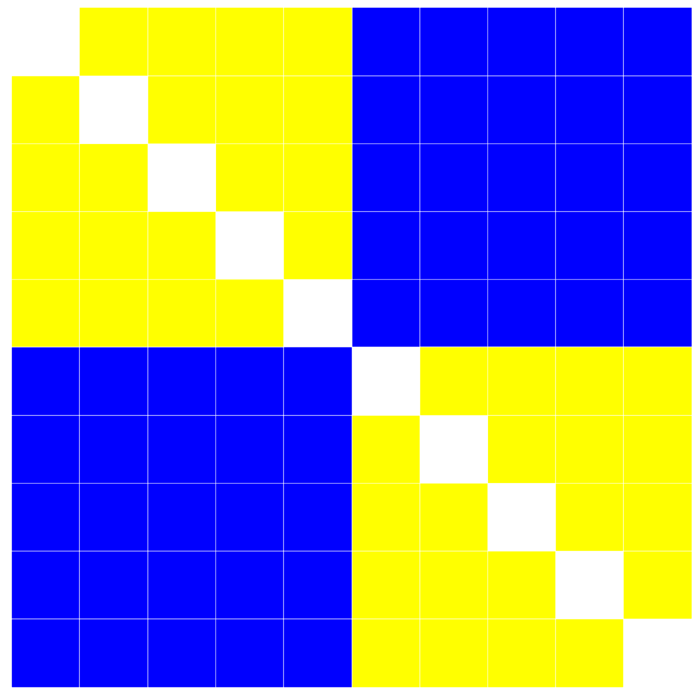
Variable



$$y_{ij} = [var(i) == var(j)]$$

**mode** = “same”

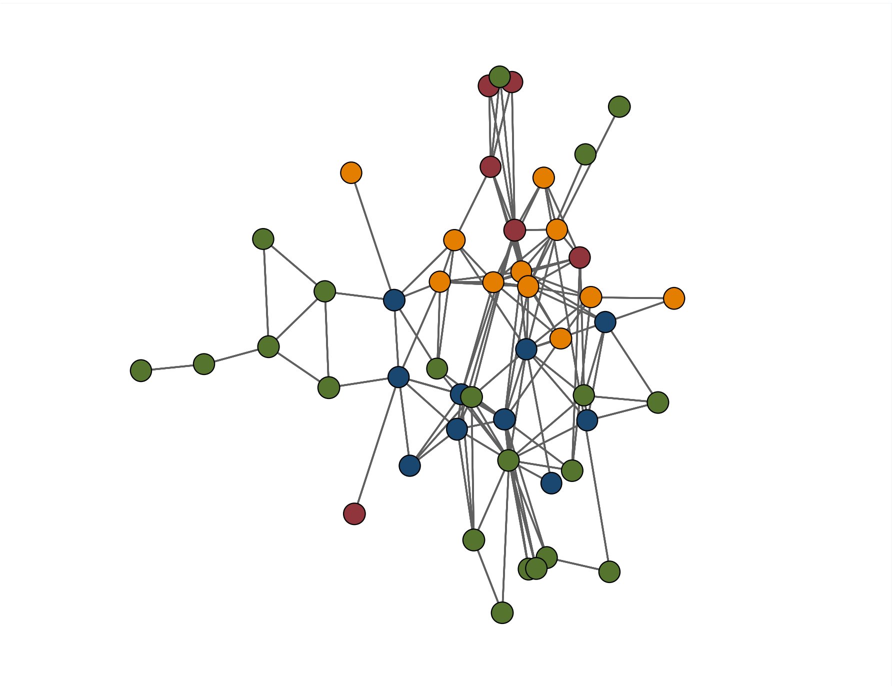
Expanded network

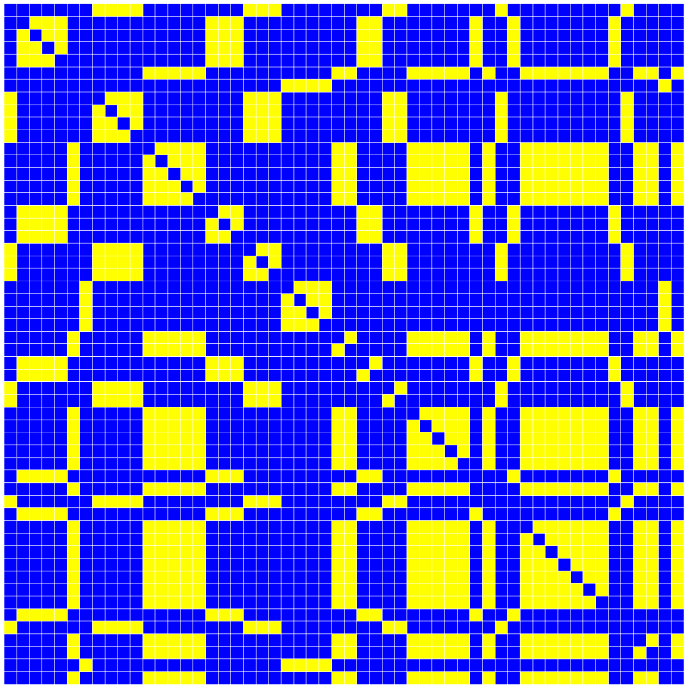


i and j same

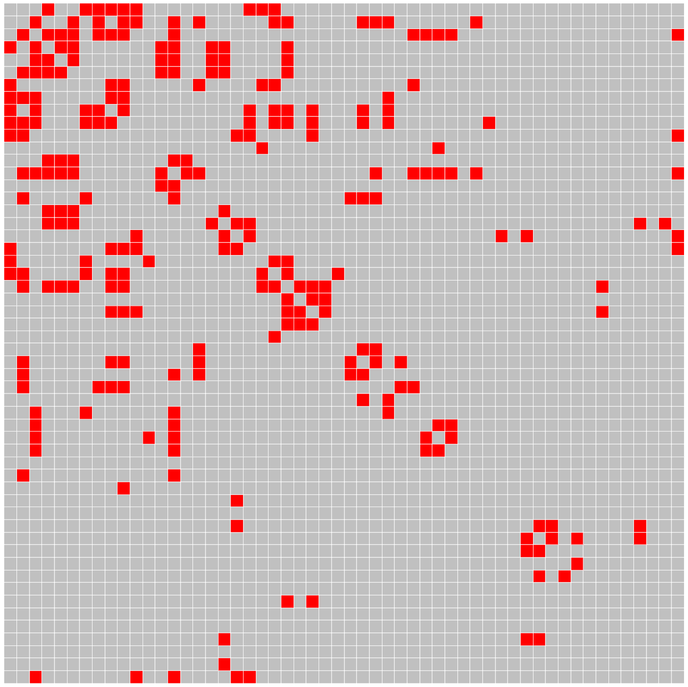
i and j different

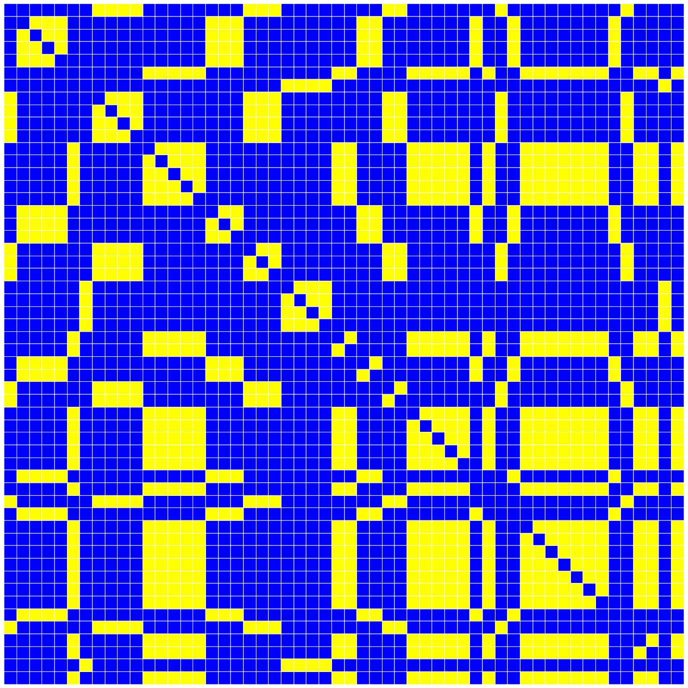
# ARE THERE MORE TIES BETWEEN MEMBERS WITH THE SAME ETHNICITY?





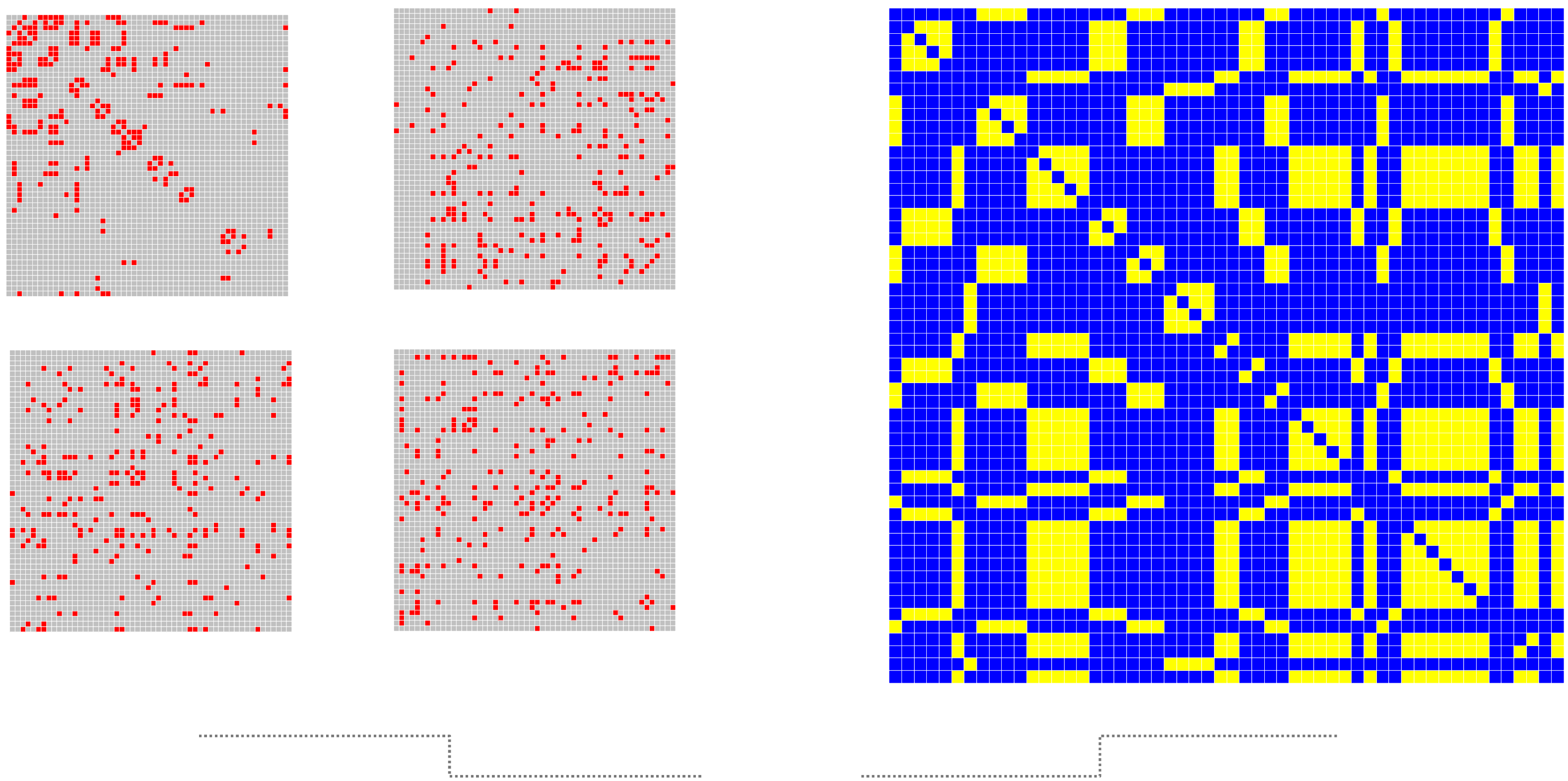
# ARE THERE MORE TIES BETWEEN MEMBERS WITH THE SAME ETHNICITY?





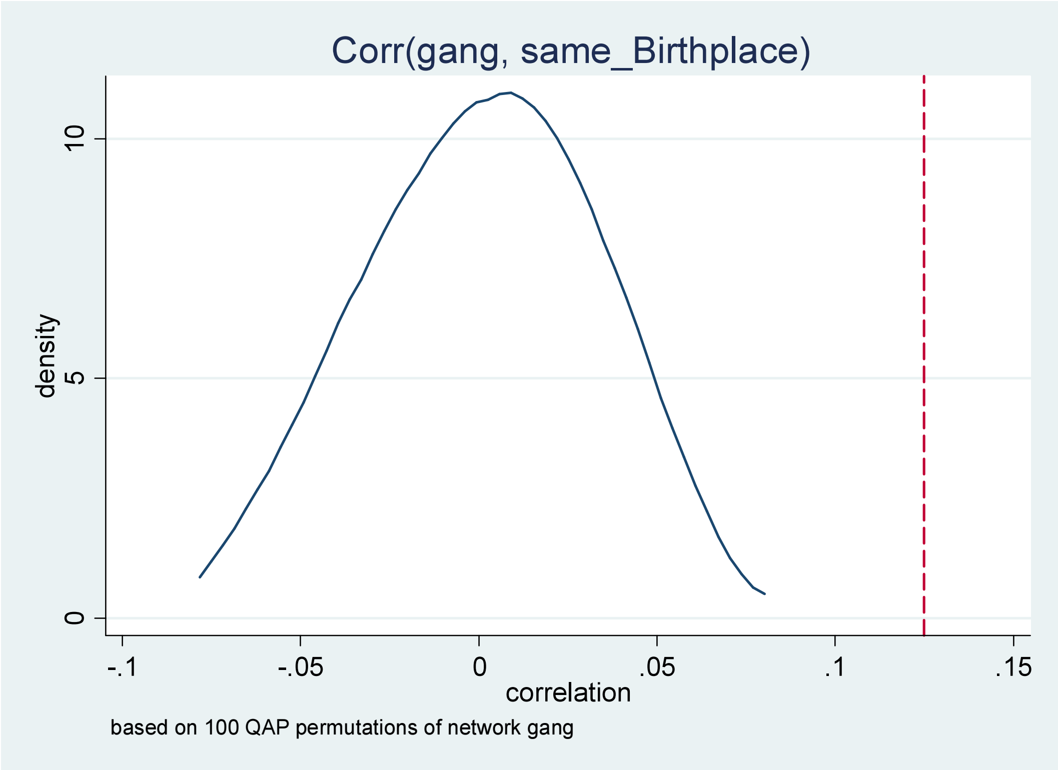
$$corr(gang, Birthplace) = 0.124$$

# ARE THERE MORE TIES BETWEEN MEMBERS WITH THE SAME ETHNICITY?



$$corr[perm(gang), Birthplace]$$

# ARE THERE MORE TIES BETWEEN MEMBERS WITH THE SAME ETHNICITY?

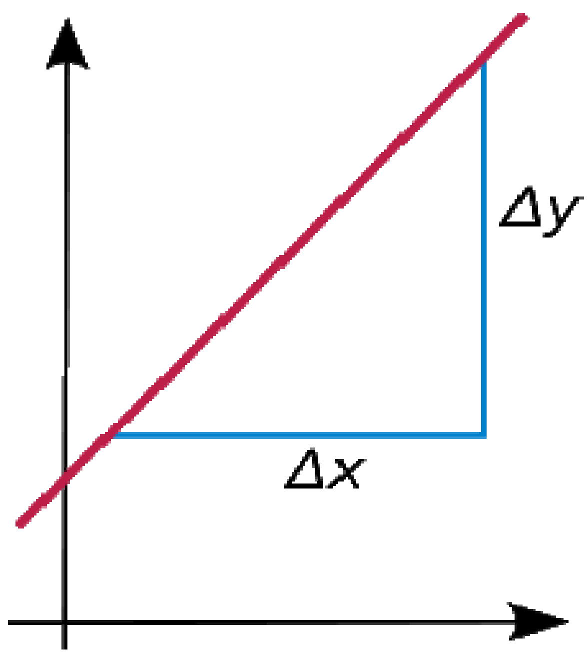



```
nwcorrelate gang, attribute(Birthplace) permutations(100)
```

# HOMOPHILY

nwtabulate
nwcorrelate
nwpermute

# REGRESSION-BASED APPROACH



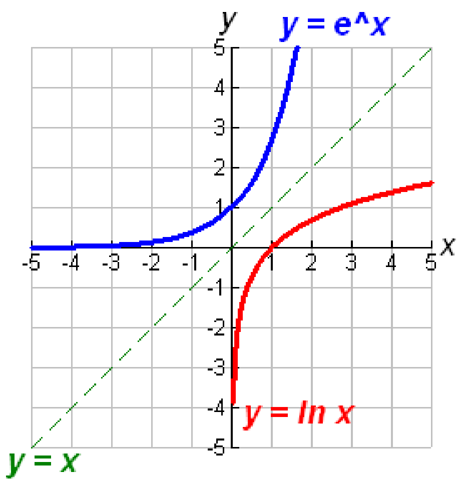
# LOGISTIC REGRESSION

**Dependent variable = binary (1 or 0)**

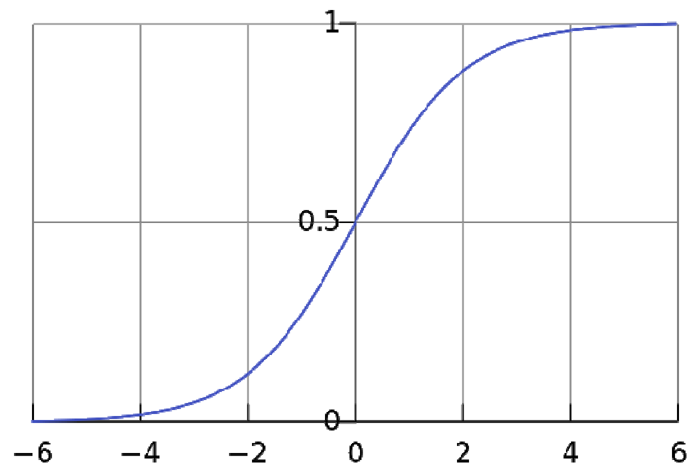
**Logistic regression:** $Pr(y_i = 1) = p_i$

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum \beta_k x_{ki}$$

**Logistic function:** $$\text{logistic}(x) = \frac{1}{1 + e^x}$$




exponential function



logistic function

# LOGISTIC REGRESSION

**Dependent variable = binary (1 or 0)**

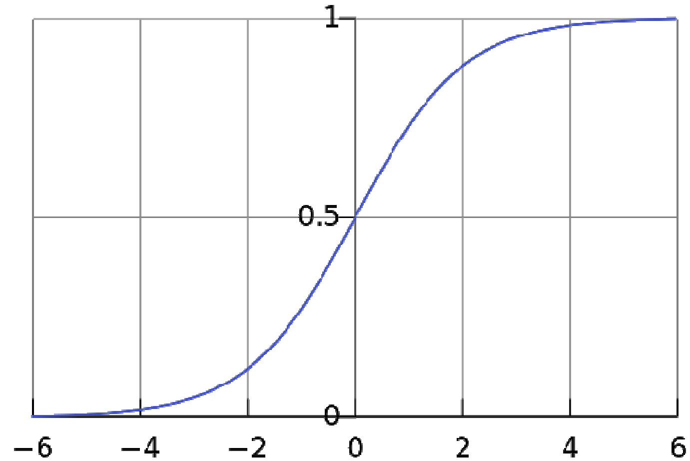
**Logistic regression:** $\Pr(y_i = 1) = p_i$

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum \beta_k x_{ki}$$

**Logistic function:** $$\operatorname{logistic}(x) = \frac{1}{1+e^{x}}$$

**Odds-ratio:** $e^{\beta_k}$

By which factor do the odds increase?




logistic function

# QAP REGRESSION

- We can use the QAP principle to run
  1. Dyad-level logistic regression on dyadic dataset
  2. Permute network many times
  3. Run dyad-level logistic regression on permuted networks
  4. Compare regression estimate from unscrambled network with regression estimates obtained with permuted networks to derive standard errors.

For example:. Grund, T. and Densley, J. (2012) Ethnic Heterogeneity in the Activity and Structure of a Black Street Gang. ***European Journal of Criminology*****, , Vol. 9, Issue 3, pp. 388-406.**

```
. nwqap gang Birthplace Residence Arrests, mode(same same absdist) permutations(200)

Permutation: 1 out of 200
Permutation: 50 out of 200
Permutation: 100 out of 200
Permutation: 150 out of 200
Permutation: 200 out of 200


Multiple Regression Quadratic Assignment Procedure

  Estimation             =  QAP
  Regression             =  logit
  Permutations           =  200
  Number of vertices     =  54
  Number of edges        =  133
```

| gang | Coef. | P-value |
|---|---|---|
| same_Birthplace | .859192 | .005 |
| same_Residence | .186923 | .41 |
| absdist_Arrests | -.036064 | .095 |
| _cons | -2.447445 | |

# REGRESSION BASED APROACH

nwqap

- Load the **hpotter** data from the nwcommands-Server using `webnwuse`
- Use `nwtabulate` to show the support ties by **gender**
- Do same sex persons support each other more? (`nwcorrelate`)
- Is this significant?


DANCING
WITH THE STARS

# DYADIC (IN-)DEPENDENCE

# INDEPENDENCE

- Independence means that whatever you observe as outcome variable in one case does not depend on the value the outcome variable has in other cases.

- **Statistical independence** is one of the most crucial and common assumptions in statistics (and practically you make it all the time).

  - $P(A \cap B) = P(A)\ P(A)$

  - Somehow defeats the purpose of ‘sociology’ (in my view)

# LOOKING BACK

**How do the different approaches we looked at…**

a) **allow to test hypotheses on network data?**

Typically, the "actual hypotheses" are not so much about structure at all, but about who is central, who links up with whom and so on.

a) **address the issue of interdependent data?**

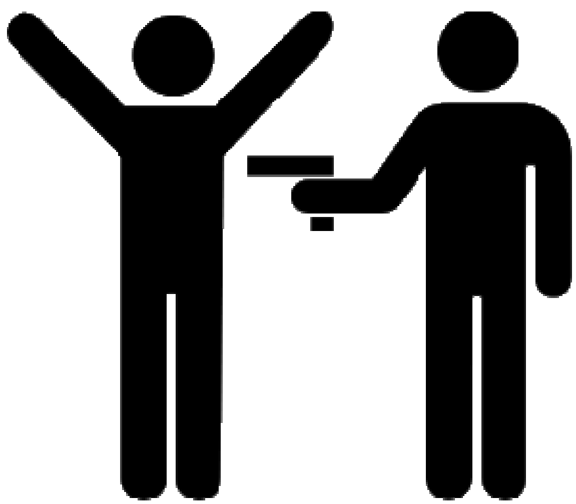
Interdependence is seen as a "nuisance", something that needs to be taken into account, but not as something of focal interest.

# CONDITIONAL UNIFORM MODELS

(e.g. tie independence given the dyad census)

a) Hypothesis are tested by working with a network distribution that enforces a selection of structural constraints.
b) Interdependence is taken into account as far as the structural constraints already imply it.

Here, structure is not treated as ***endogenous*** (part of the dependent variable, “to be explained”), but as ***exogenous*** (here even: enforced). It is not always clear what to enforce. Approach is not very flexible.



# PERMUTATION-BASED MODELLING

a) Hypothesis are tested (like in conditional uniform tests) by working with a network distribution that enforces structural constraints – here even the “total structure” (actual structure remains the same).

b) Interdependence is taken into account by completely fixing the network structure.

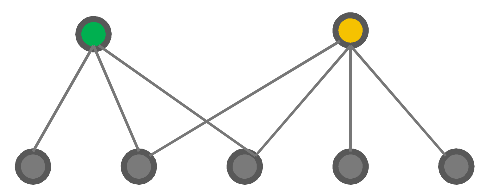
Also here, structure is treated as ***exogenous***. This means it is difficult to study how dependencies due to structure and dependencies due to explanatory variables interact.

# P2 MODEL

a) Hypothesis are done based on parameter estimation/model fitting. The distribution of networks is not fixed (as in previous approaches) but modelled by a parametric family of models.

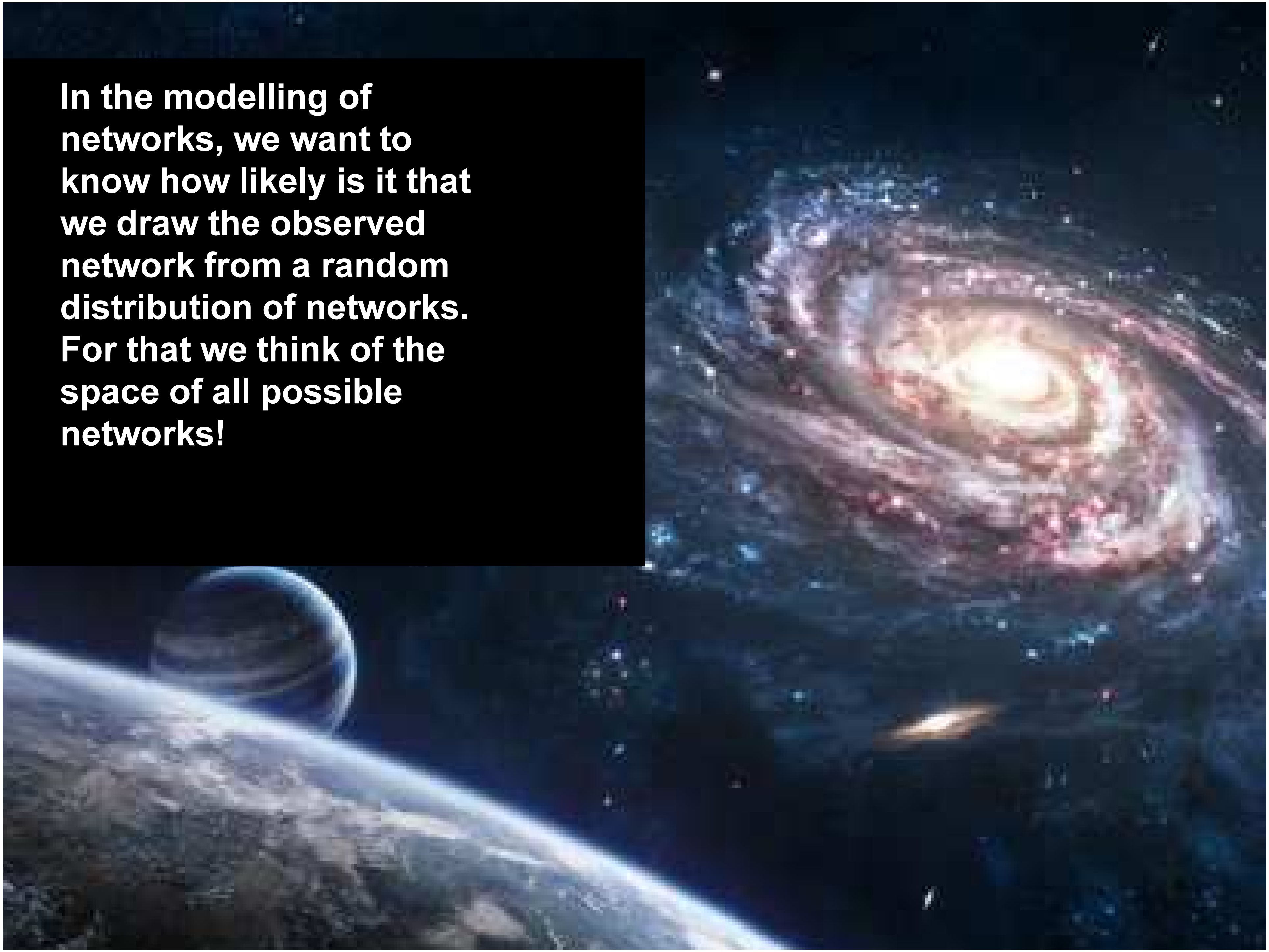
b) Within-dyad dependence is modelled through correlation, between-dyad dependence through random effects (common sender and/or receiver effects)

Here, structure is treated as ***endogenous*** – but in a limited sense. Triad level dependencies that are not due to common sender or common receiver effects cannot be expressed: transitivity, social balance, preferential attachment… ERGM allow for this!



# EXPONENTIAL RANDOM GRAPH MODELS

**What is the probability to ‘observe’ our network?**


In the modelling of networks, we want to know how likely is it that we draw the observed network from a random distribution of networks. For that we think of the space of all possible networks!


The space of ‘possible’ networks/worlds is huge! If you would count all possible networks for the people in this room you would be busy until the end of time.

**Build a model that is good in producing the network features that we observed.**

Instead of thinking about the probability of a particular network, we think of the probability of networks having certain features.

# ERGM

$\boldsymbol{Y = random\ variable}$, a randomly selected network from the pool of all potential networks

$\boldsymbol{y = observed\ variable}$, here observed network

$\boldsymbol{\theta = parameters}$, to be estimated

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) = \frac{e^{\left(\theta^T s(\boldsymbol{y})\right)}}{c(\theta)}$$

A score given to our network **y** using some parameters $\theta$ and the network features **s** of **y**

Probability to draw 'our' observed network **y** from all potential networks

A score given to all other networks we could have observed

# ERGM

$\boldsymbol{Y} = \boldsymbol{random\ variable}$, a randomly selected network from the pool of all potential networks

$\boldsymbol{y} = \boldsymbol{observed\ variable}$, here observed network

$\boldsymbol{\theta} = \boldsymbol{parameters}$, to be estimated

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) = \frac{\boxed{e^{\left(\theta^T s(\boldsymbol{y})\right)}}}{c(\theta)}$$

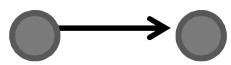
A score given to our network **y** using some parameters

$s(\boldsymbol{y}) = \boldsymbol{independent\ variable}$, prevalence of micro structures in the network, e.g. number ties, number of reciprocal ties, number of triads…

# ERGM: POSSIBLE MICROSTRCUTURES

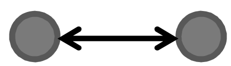
$s(\boldsymbol{y}) = \boldsymbol{independent\ variable}$, prevalence of micro structures in the network, e.g. number ties, number of reciprocal ties, number of triads…

*tie count statistic*



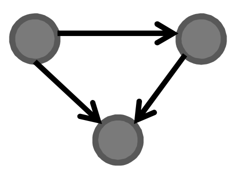
$$s_k(y) = \sum y_{ij}$$

*reciprocity statistic*



$$s_k(y) = \sum y_{ij}\, y_{ji}$$

*transitive triplets*



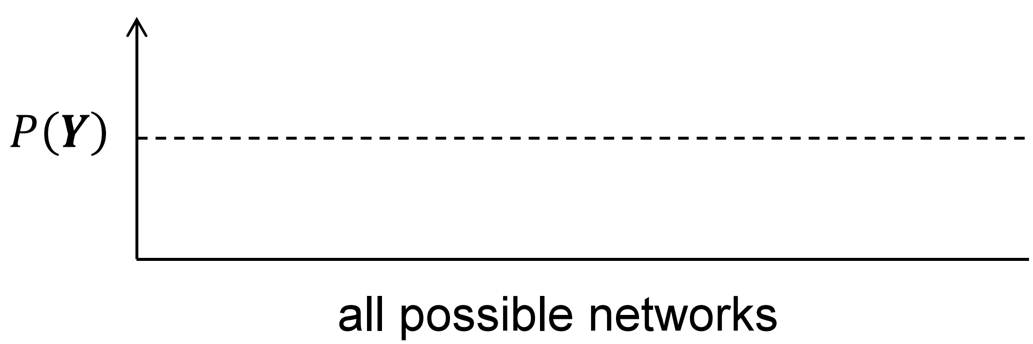
$$s_k(y) = \sum y_{ij}\, y_{jk} y_{ik}$$

# ERGM: LINK FEATURES AND PROBABILITIES

Think of the space of all possible isomorph networks (networks with unique features).

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) = \frac{\boxed{e^{\left(\theta^T s(\boldsymbol{y})\right)}}}{c(\theta)}$$

If parameter vector $\theta$ is set to [0,0,0…], then, all unique classes of isomorph networks get the same score. All possible combination of network features have the same probability

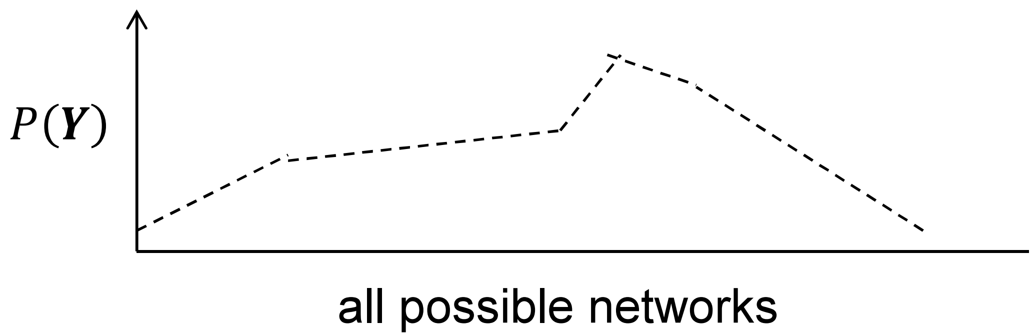


# ERGM: LINK FEATURES AND PROBABILITIES

Think of the space of all possible isomorph networks (networks with unique features).

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) = \frac{\boxed{e^{\left(\theta^T s(\boldsymbol{y})\right)}}}{c(\theta)}$$

If parameter vector $\theta$ is **not** [0,0,0…], then, some networks get a higher score than others, which means, we assume that they are more likely to be drawn.



# ERGM

By changing the parameter vector $\theta$ we can alter the link between all possible networks and the chances of these networks to be drawn in a stochastic process (depending on the network features the possible networks have).

In the estimation we try to find a vector $\theta$ which changes the probability distribution for all potential networks to be drawn in such a way that the one network **y** that we actually did draw, is the most likely one.

Find $\theta$ so that: $$\max_{\theta} P(\boldsymbol{Y} = \boldsymbol{y} | \theta)$$

# ERGM: REFORMULATION

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) = \frac{e^{\left(\theta^T s(\boldsymbol{y})\right)}}{c(\theta)}$$

Although we only look at classes of networks defined by their features, there are still too many of them to calculate this.

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) \propto e^{\left(\theta^T s(\boldsymbol{y})\right)}$$

‘proportional to’. The actual proportionality constant is uncalculable.

# ERGM: REFORMULATION

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) = \frac{e^{\left(\theta^T s(\boldsymbol{y})\right)}}{c(\theta)}$$

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) \propto e^{\left(\theta^T s(\boldsymbol{y})\right)}$$

$$\text{logit}\left[P\left(Y_{ij=1}\middle|n\ actors, Y_{ij}^{c}\ \right)\right] = \sum_{k=1}^{K} \theta_k \delta s_k(\boldsymbol{y})$$

# ERGM: REFORMULATION

$Y_{ij}^{c}$ = all dyads other than $Y_{ij}$

Amount by which the feature $s_k(y)$ changes when $Y_{ij}$ is toggled from 0 to 1.

$$\mathrm{logit}\left[P\left(Y_{ij} = 1 \middle| n\ actors, Y_{ij}^{c}\right)\right] = \sum_{k=1}^{K} \theta_k \boxed{\delta s_k(\boldsymbol{y})}$$

Probability that there is a tie from *i* to *j*.

Given, *n* actors AND the rest of the network, excluding the dyad in question!

# ERGM: INTEPRETATION

**ERGM's ultimately give you an estimate for various parameters $\theta_k$, which mean…**

If a potential tie $Y_{ij} = 1$ (between *i* and *j*) would change the network statistic $s_k$ by one unit.

This changes the log-odds for the tie $Y_{ij}$ to actually exist by $\theta_k$.

# ERGM: INTEPRETATION

**ERGM's ultimately give you an estimate for various parameters $\theta_k$, which mean…**

If a potential tie $Y_{ij} = 1$ (between *i* and *j*) would change the network statistic $s_k$ by one unit.

This changes the log-odds for the tie $Y_{ij}$ to actually exist by $\theta_k$.

When $\theta_k$ is significant, it means that…

There is a tendency in the network for the underlying micro-structure defined by $s_k$ to be important in 'generating' the observed network.

**(more precisely, in generating networks with similar features than the one that we observe)**

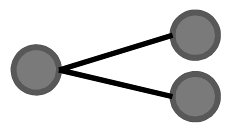
# EXAMPLE

Consider an ERGM for an undirected network with parameters for these three statistics:

1) *number of edges*

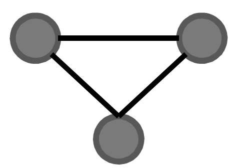
$$s_{edges}(y) = \sum y_{ij}$$

2) *number of 2-stars*



$$s_{2stars}(y) = \sum y_{ij}\, y_{ik}$$

3) *number of triangles*



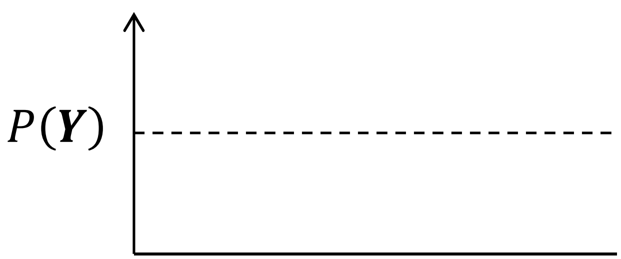
$$s_{triangles}(y) = \sum y_{ij}\, y_{jk} y_{ik}$$

Then the 3-parameter ERG distribution function is:

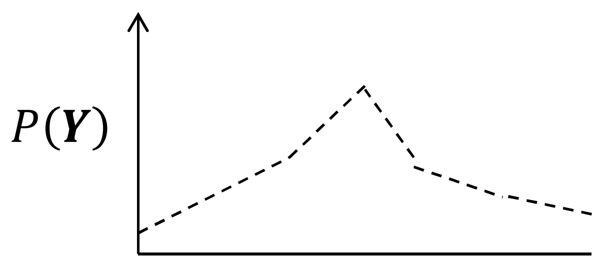
$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) \propto e^{\left(\theta_{edges} s_{edges}(y) + \theta_{2stars} s_{2stars}(y) + \theta_{triangles} s_{triangles}(y)\right)}$$

# EXAMPLE

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta) \propto e^{\left(\theta_{edges}s_{edges}(y) + \theta_{2stars}s_{2stars}(y) + \theta_{triangles}s_{triangles}(y)\right)}$$




all possible networks




all possible networks

The ERG distribution function (the combination of $s_k$ ’s and $\theta_k$’s) defines how the ‘hypothetically assumed’ underlying distribution of all possible networks to be drawn looks like. It says which networks are more likely to be randomly drawn.

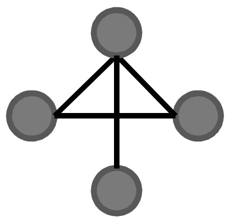
# EXAMPLE

…and consider the following two 4-node networks and their statistics:

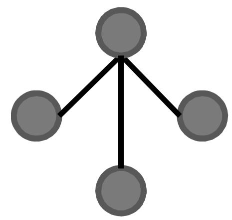
| | $\boldsymbol{y}_a$ | $\boldsymbol{y}_b$ |
|---|---|---|
| $s_{edges}$ | 4 | 3 |
| $s_{2stars}$ | 5 | 3 |
| $s_{triangles}$ | 1 | 0 |

# EXAMPLE

…we do not know the proportionality constant …



$\boldsymbol{y}_a$



$\boldsymbol{y}_b$

$$P(\boldsymbol{y}_a|\theta) \propto e^{\left(4 \times \theta_{edges} + 5 \times \theta_{2stars} + 1 \times \theta_{triangles}\right)}$$

$$P(\boldsymbol{y}_b|\theta) \propto e^{\left(3 \times \theta_{edges} + 3 \times \theta_{2stars}\right)}$$

# EXAMPLE

…although we do not know the proportionality constant we can calculate the **ratio** between the two probabilities!


$$\frac{P(\boldsymbol{y}_a|\theta)}{P(\boldsymbol{y}_b|\theta)} = \frac{e^{\left(4\times\theta_{edges}+5\times\theta_{2stars}+1\times\theta_{triangles}\right)}}{e^{\left(3\times\theta_{edges}+3\times\theta_{2stars}\right)}}$$

$$= e^{\left(\theta_{edges}+2\times\theta_{2stars}+\theta_{triangles}\right)}$$

How much more likely is $\boldsymbol{y}_a$ in contrast to $\boldsymbol{y}_b$?

# EXAMPLE

… so, suppose in a larger network the estimation gave the following parameters:

low density: $\theta_{edges} = -1.5$

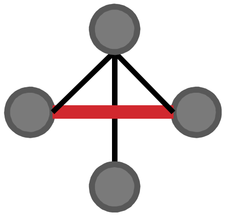
positive degree variance: $\theta_{2stars} = 0.1$

redundant ties are avoided: $\theta_{triangles} = -0.4$

$$\frac{P(\boldsymbol{y}_a|\theta)}{P(\boldsymbol{y}_b|\theta)} = e^{(\theta_{edges}+2\times\theta_{2stars} + \theta_{triangles})}$$

$$= e^{(-1.5\,+2\,\times 0.1\,-0.4)} = e^{-1.7} \approx \frac{1}{5.5}$$

i.e. the middle tie is about 5.5 times likely NOT to exist as to exist (given the rest of the network)



# ERGM FEATURES

- Think of ERG models as a probability distribution on a (huge) space of all possible networks.

- The observed network is modelled as if it has been drawn from this distribution.

- The model parameters $\theta$ are
  - Attached to network statstictis $s$
  - These statistics in general correspond to subgraph counts (local patterns, ’motifs’)
  - The parameters describe the relative prevalence of the corresponding subgraph in ’generating’ the total graph.

- The parameters $\theta$ are estimated in such a way that each change of a tie (during the process of ’generating’ a network) is considered for the next ties that could change. Structure is ***endogenous*** **=> dyadic dependence model**

# ERGM FEATURES

- High flexibility due to the many possibilities of choosing statistic and controlling effects for each other.
- Estimation, model specification and interpretation can be difficult!

# EXPONENTIAL RANDOM GRAPH MODEL

nwergm