Homophily

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Bart Baesens, Ph.D.

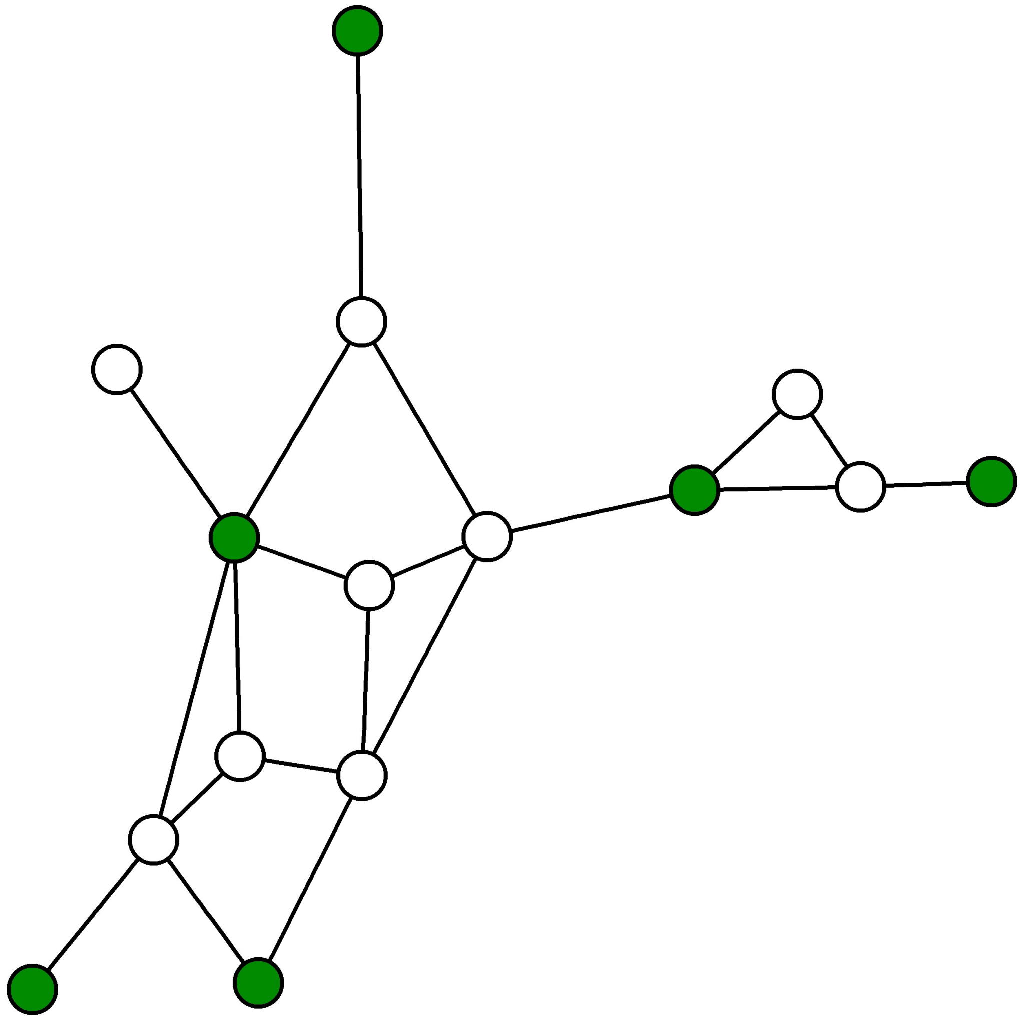
Professor of Data Science, KU Leuven
and University of Southampton

Homophily explained

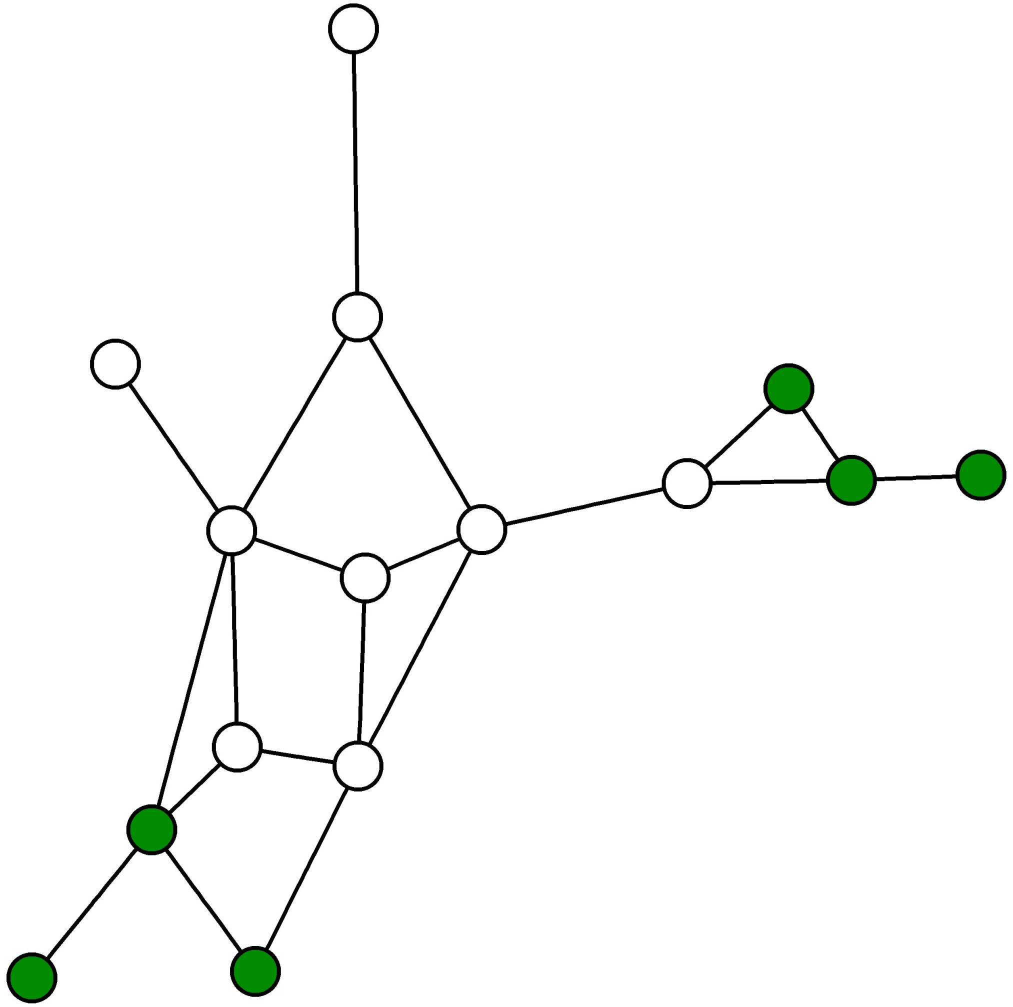
> Birds of a feather flock together

- Share common property, hobbies, interest, origin, etc.
- Depends on:
  - Connectedness between nodes with **same** label
  - Connectedness between nodes with **opposite** labels

Homophilic Networks



- Not Homophilic



- Homophilic

```r
names <- c('A','B','C','D','E','F','G','H','I','J')
tech <- c(rep('R',6),rep('P',4))
DataScientists <- data.frame(name=names,technology=tech)
DataScienceNetwork <- data.frame(
 from=c('A','A','A','A','B','B','C','C','D','D',
        'D','E','F','F','G','G','H','H','I'),
 to=c('B','C','D','E','C','D','D','G','E','F',
        'G','F','G','I','I','H','I','J','J'),
 label=c(rep('rr',7),'rp','rr','rr','rp','rr','rp','rp',rep('pp',5)))

g <- graph_from_data_frame(DataScienceNetwork,directed = FALSE)
```

Add the technology as a node attribute

```r
V(g)$label <- as.character(DataScientists$technology)
V(g)$color <- V(g)$label
V(g)$color <- gsub('R',"blue3",V(g)$color))
V(g)$color <- gsub('P',"green4",V(g)$color)
```

Types of edges

Code to color the edges

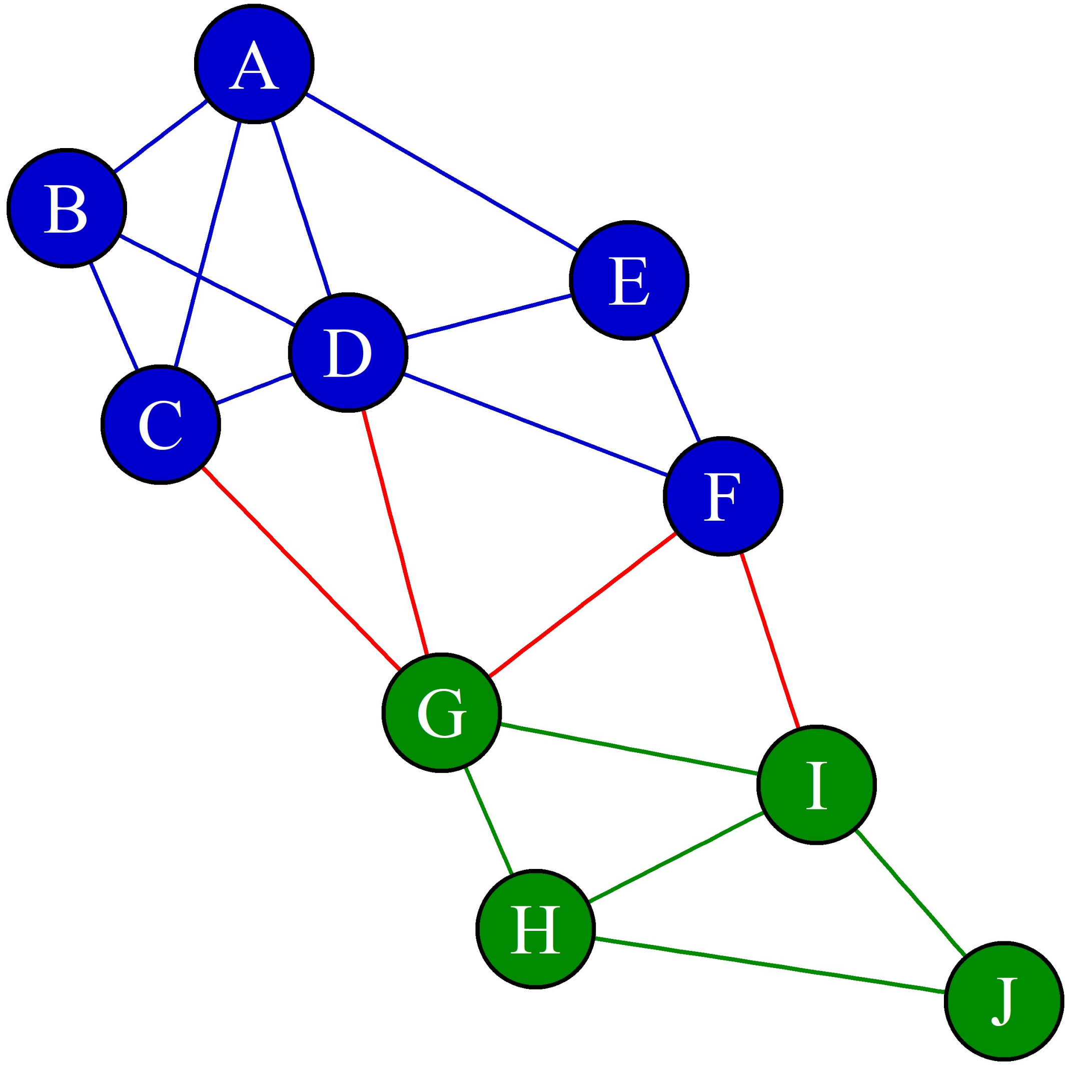
```
E(g)$color<-E(g)$label
E(g)$color=gsub('rp','red',E(g)$color)
E(g)$color=gsub('rr','blue3',E(g)$color)
E(g)$color=gsub('pp','green4',E(g)$color)
```

Code to visualize the network

```
pos<-cbind(c(2,1,1.5,2.5,4,4.5,3,3.5,5,6),
c(10.5,9.5,8,8.5,9,7.5,6,4.5,5.5,4))

plot(g,edge.label=NA,vertex.label.color='white',
layout=pos, vertex.size = 25)
```

Counting edge types



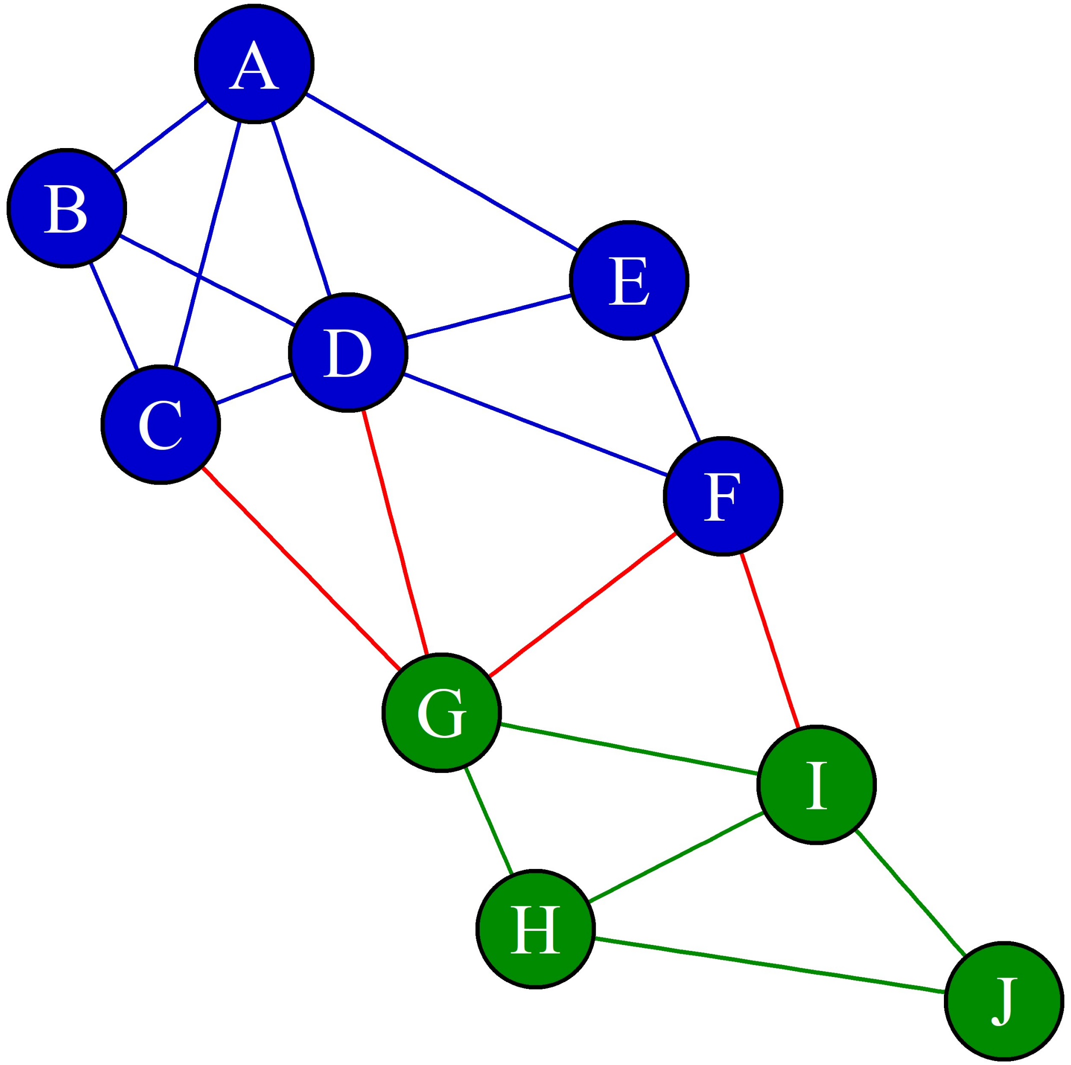
```
# R edges
edge_rr<-sum(E(g)$label=='rr')

# Python edges
edge_pp<-sum(E(g)$label=='pp')

# cross label edges
edge_rp<-sum(E(g)$label=='rp')
```

- `edge_rr=` 10
- `edge_pp=` 5
- `edge_rp=` 4

Network connectance



$$p = \frac{2 \cdot edges}{nodes(nodes-1)}$$

```
p <- 2*edges/nodes*(nodes-1)
```

- `p = 0.42`
- Number of edges in a fully connected network: $\binom{nodes}{2} = \frac{nodes(nodes-1)}{2}$

Let's practice!

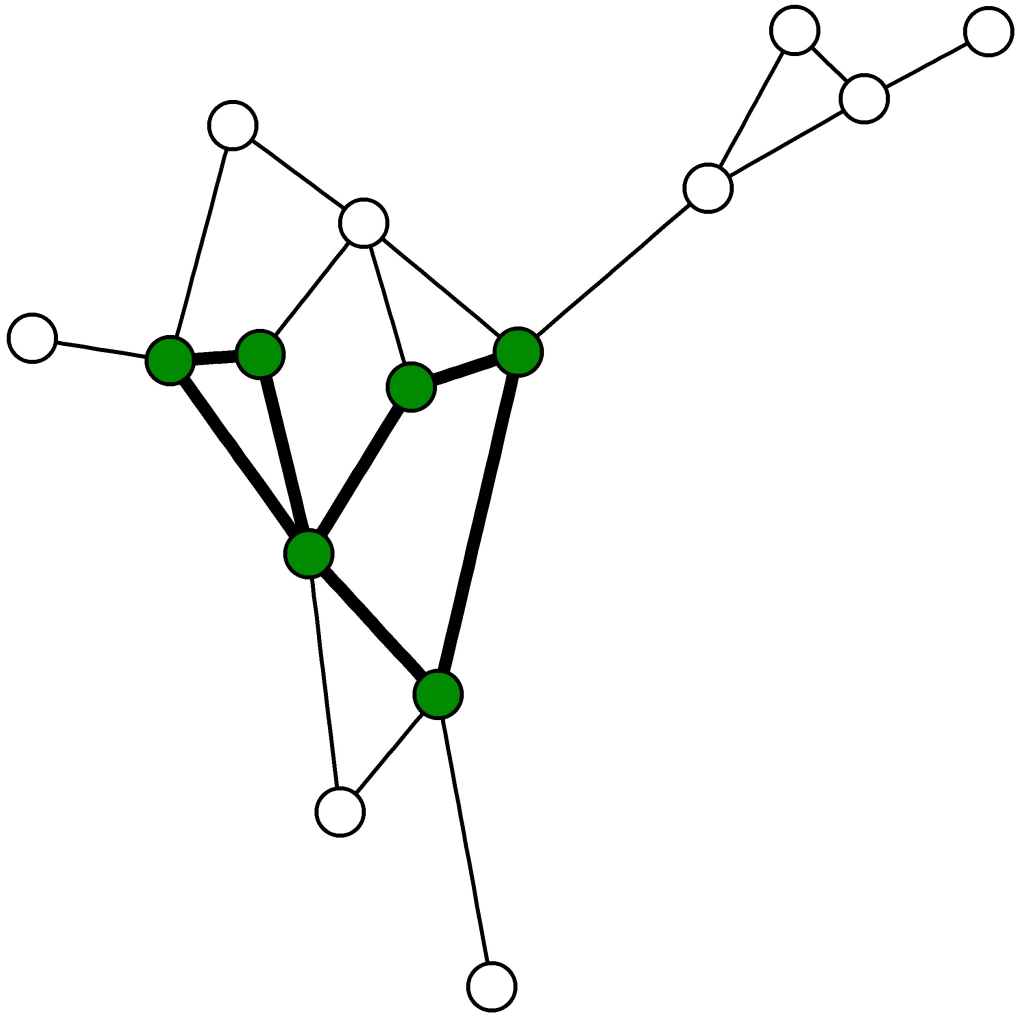
PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Dyadicity

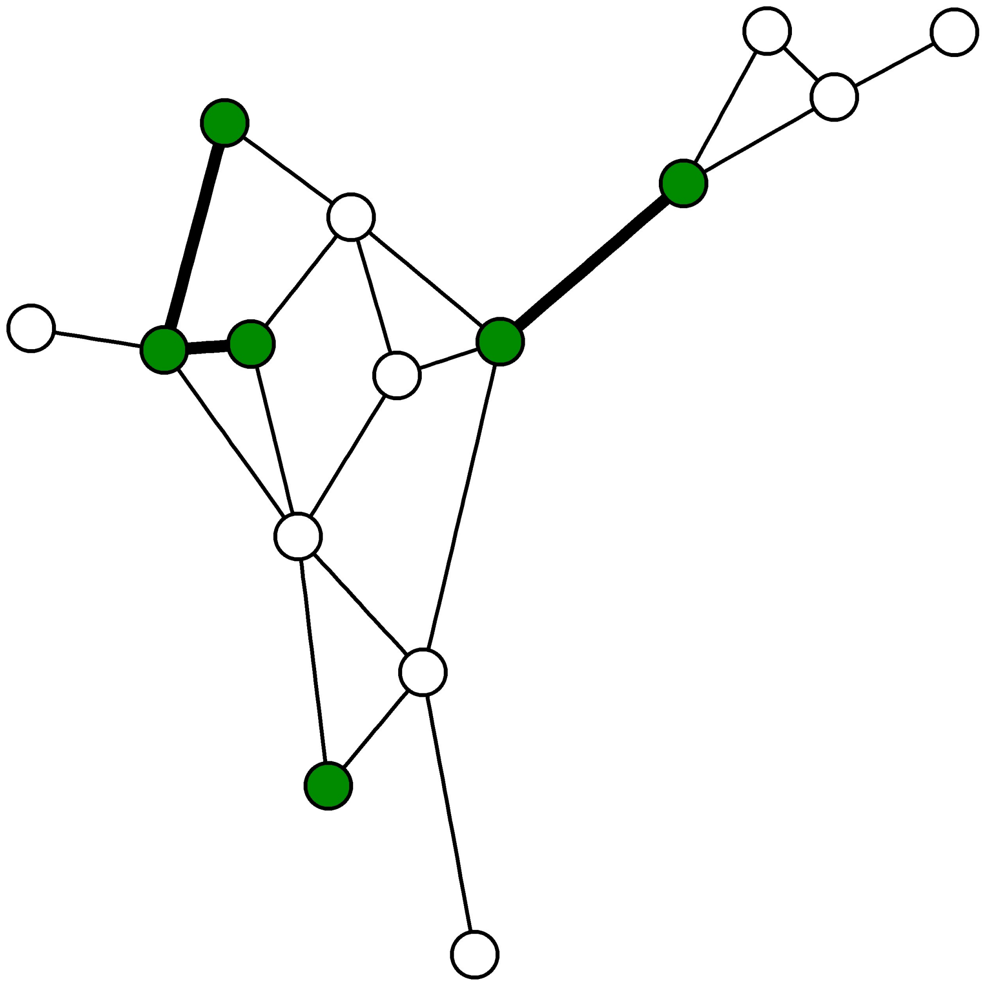
PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

María Óskarsdóttir, Ph.D.
Post-doctoral researcher

Dyadicity



7 edges between green nodes



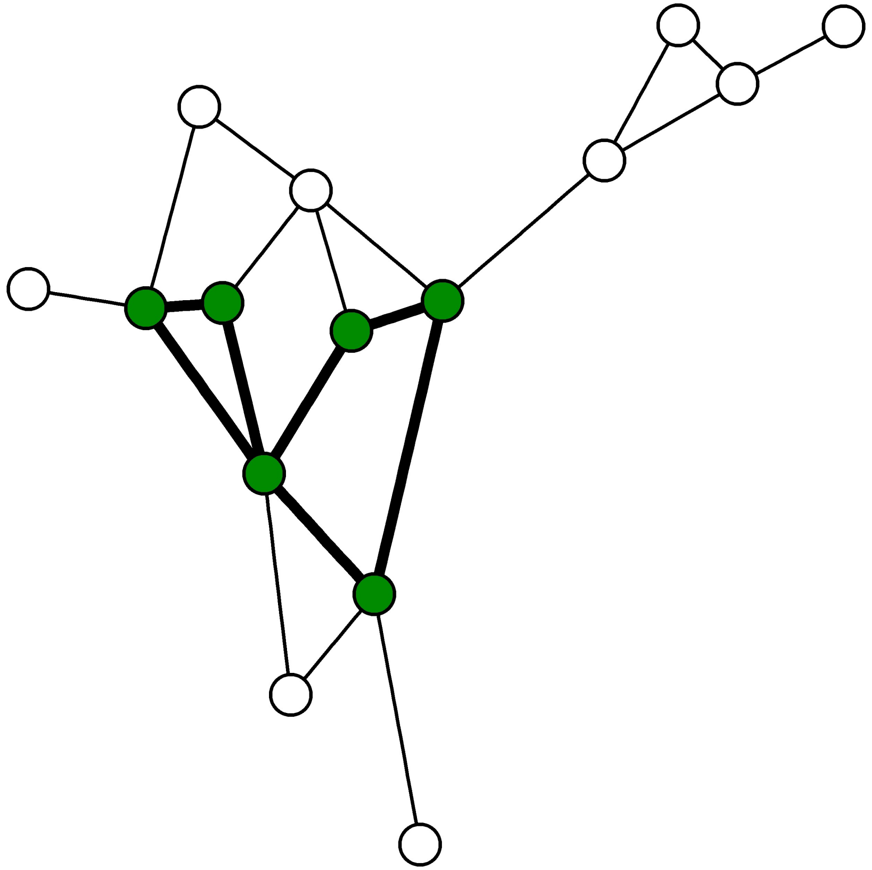
3 edges between green nodes

Dyadicity

Connectedness between nodes with the **same** label compared to what is expected in a random configuration of the network

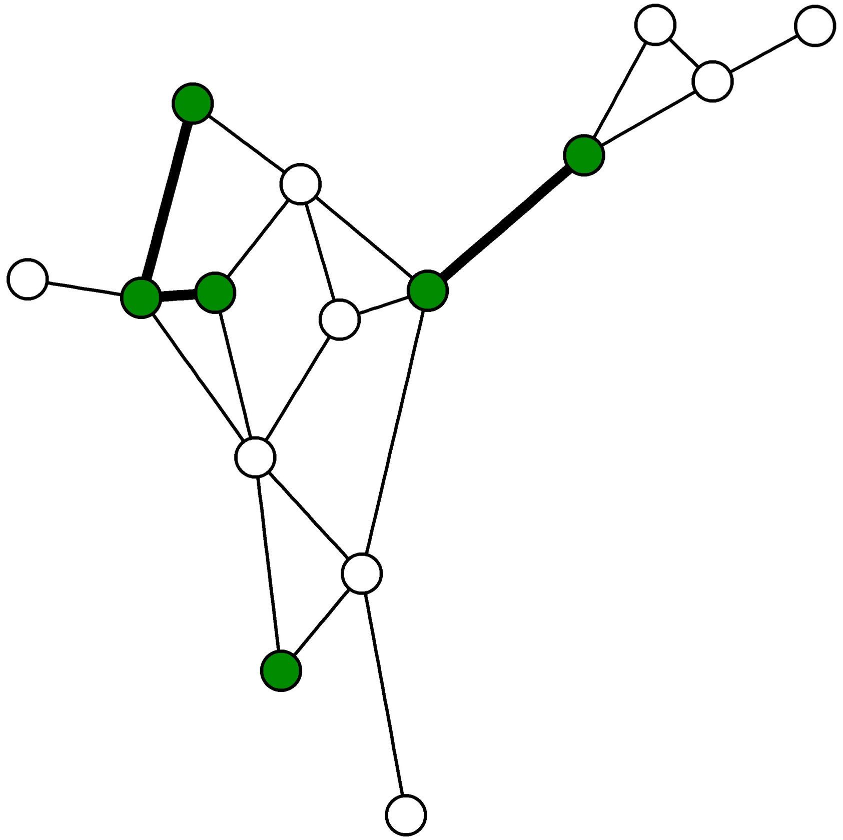
- Expected number of same label edges: $\binom{n_g}{2} \cdot p = \frac{n_g(n_g-1)}{2} \cdot p$
- Example:
  - Network with 9 white nodes, 6 green nodes, 21 edges, and connectance $p = 0.2$
  - Expected number of edges connecting two green nodes is 3 $(= \frac{6 \cdot 5 \cdot p}{2})$
- Dyadicity equals the actual number of same label edges divided by the expected number of same label edge
  - $D = \frac{\text{number of same label edges}}{\text{expected number of same label edges}}$

Dyadicity



7 edges between green nodes

- $D = 7/3 = 2.33$



3 edges between green nodes

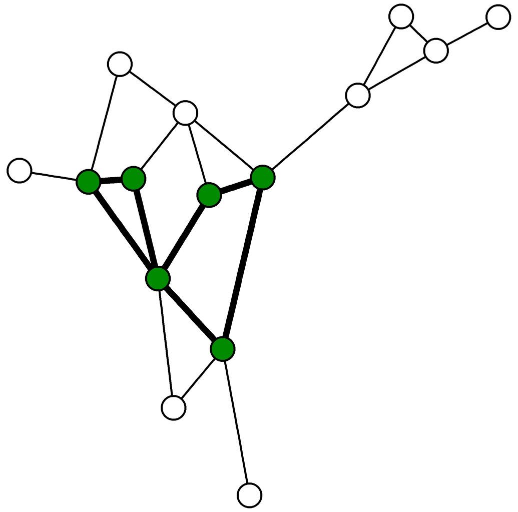
- $D = 3/3 = 1$

Types of Dyadicity

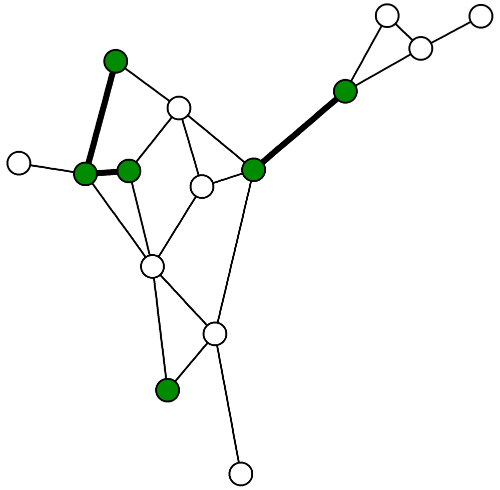
Three scenarios

1. $D > 1 \Rightarrow$ Dyadic
2. $D \simeq 1 \Rightarrow$ Random
3. $D < 1 \Rightarrow$ Anti-Dyadic

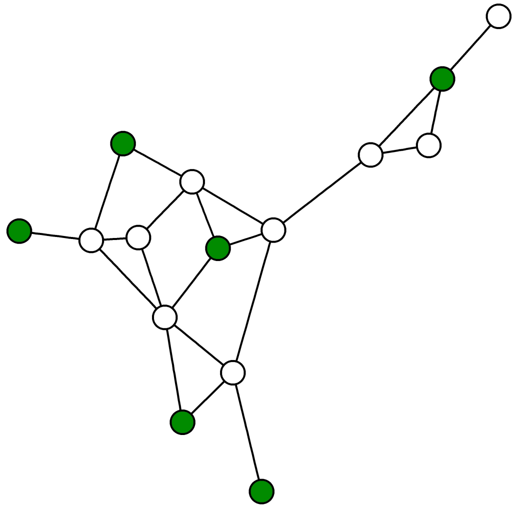
$D = 2.33$



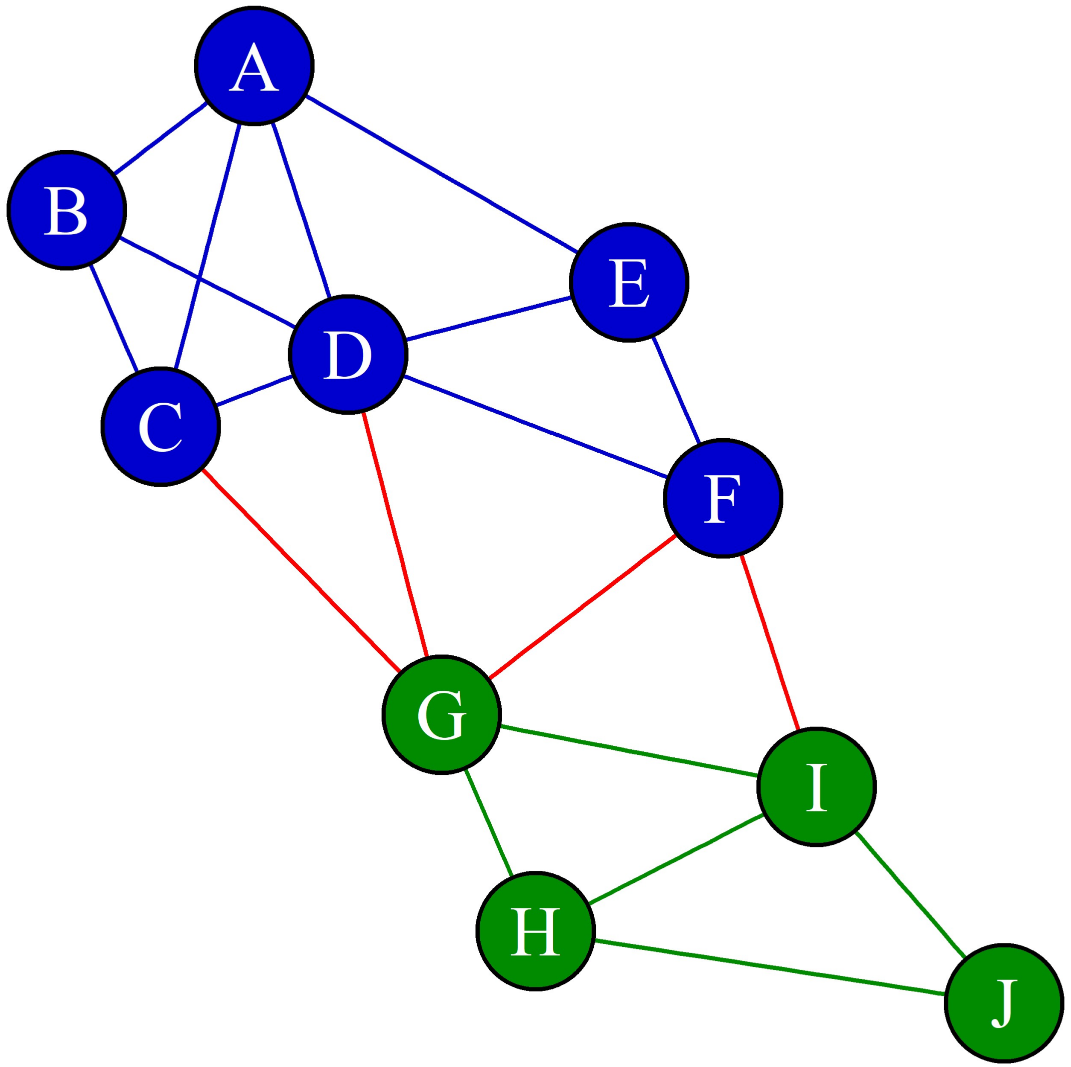
$D = 1$



$D = 0$



Dyadicity in the Network of Data Scientists



```
p <- 2 * 19 / (10 * 9)
expectedREdges <- 6 * 5 / 2 * p
expectedPEdges <- 4 * 3 / 2 * p
dyadicityR <- rEdges / expectedREdges
dyadicityP <- pEdges / expectedPEdges
dyadicityR
```

```
1.578947
```

```
dyadicityP
```

```
1.973684
```

Let's practice!

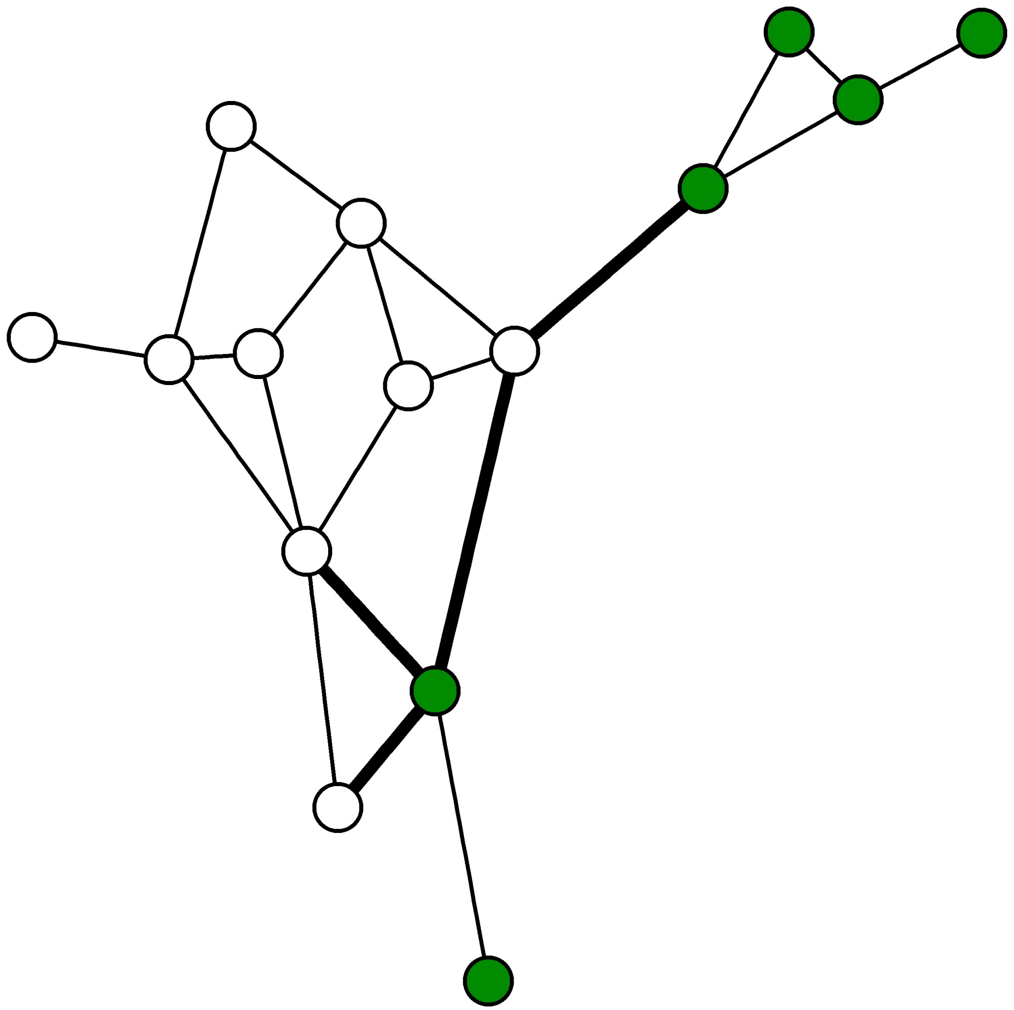
PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Heterophilicity

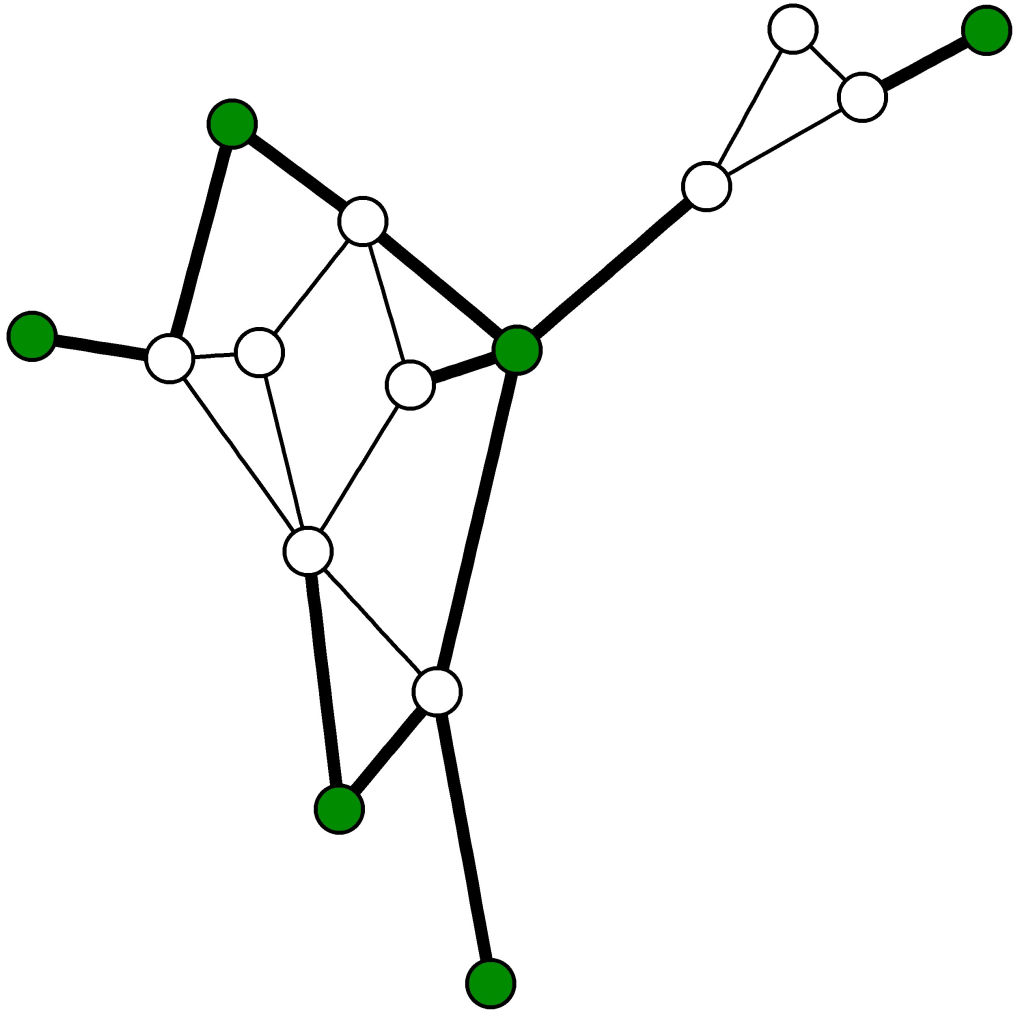
PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

María Óskarsdóttir, Ph.D.
Post-doctoral researcher

Heterophilicity



4 cross label edges



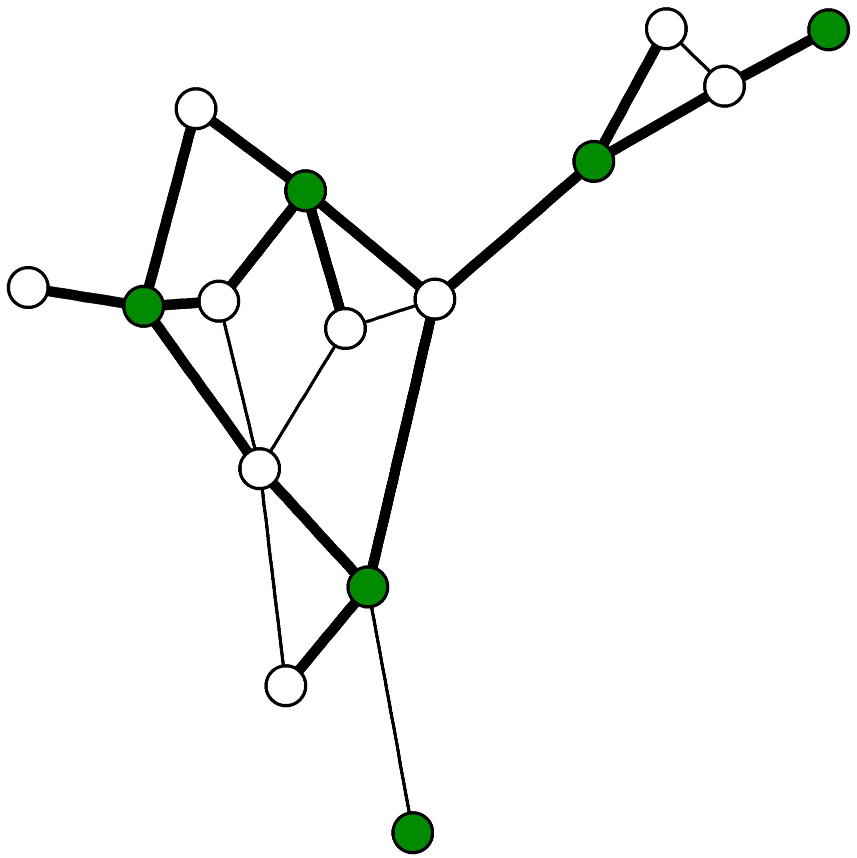
11 cross label edges

Heterophilicity

Connectedness between nodes with **different** labels compared to what is expected for a random configuration of the network

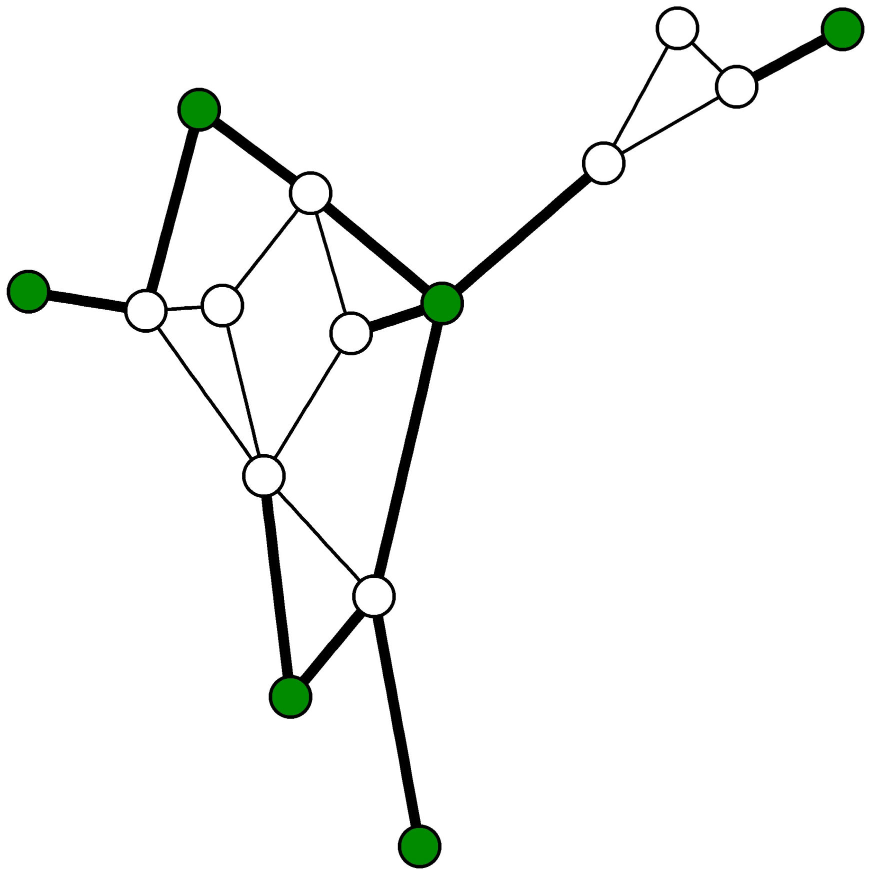
- Expected number of cross label edges $= n_w n_g p$
- Example:
  - Network with 9 white nodes, 6 green nodes, 21 edges, and connectance $p = 0.2$
  - Expected number of cross label edges is 11 $(= 9 \cdot 6 \cdot p)$
- Heterophilicty equals the actual number of cross label edges divided by the expected number of cross label edges
  - $H = \frac{\text{number of cross label edges}}{\text{expected number of cross label edges}}$

Heterophilicity



15 cross label edges

- $H = 15/11 = 1.39$



11 cross label edges

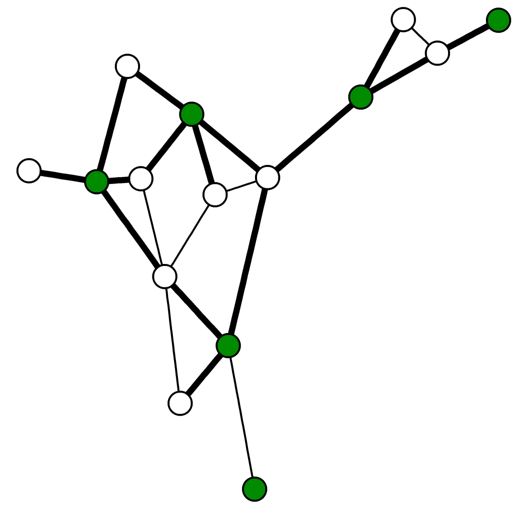
- $H = 11/11 = 1.02$

Types of Heterophilicity

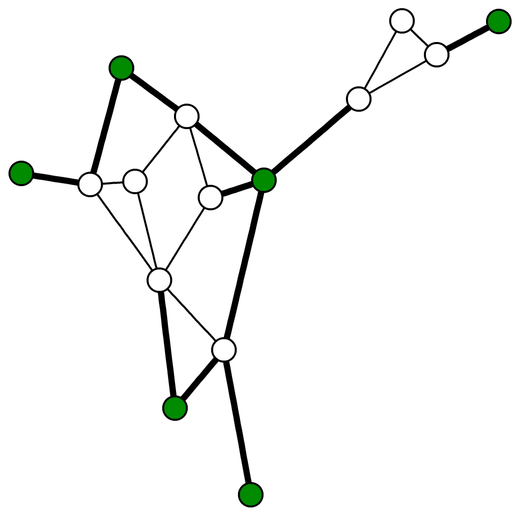
Three scenarios

1. $H > 1 \Rightarrow$ Heterophilic
2. $H \simeq 1 \Rightarrow$ Random
3. $H < 1 \Rightarrow$ Heterophobic

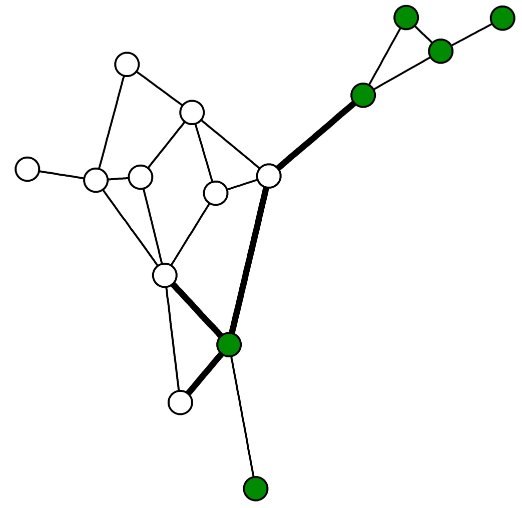
$H = 1.39$



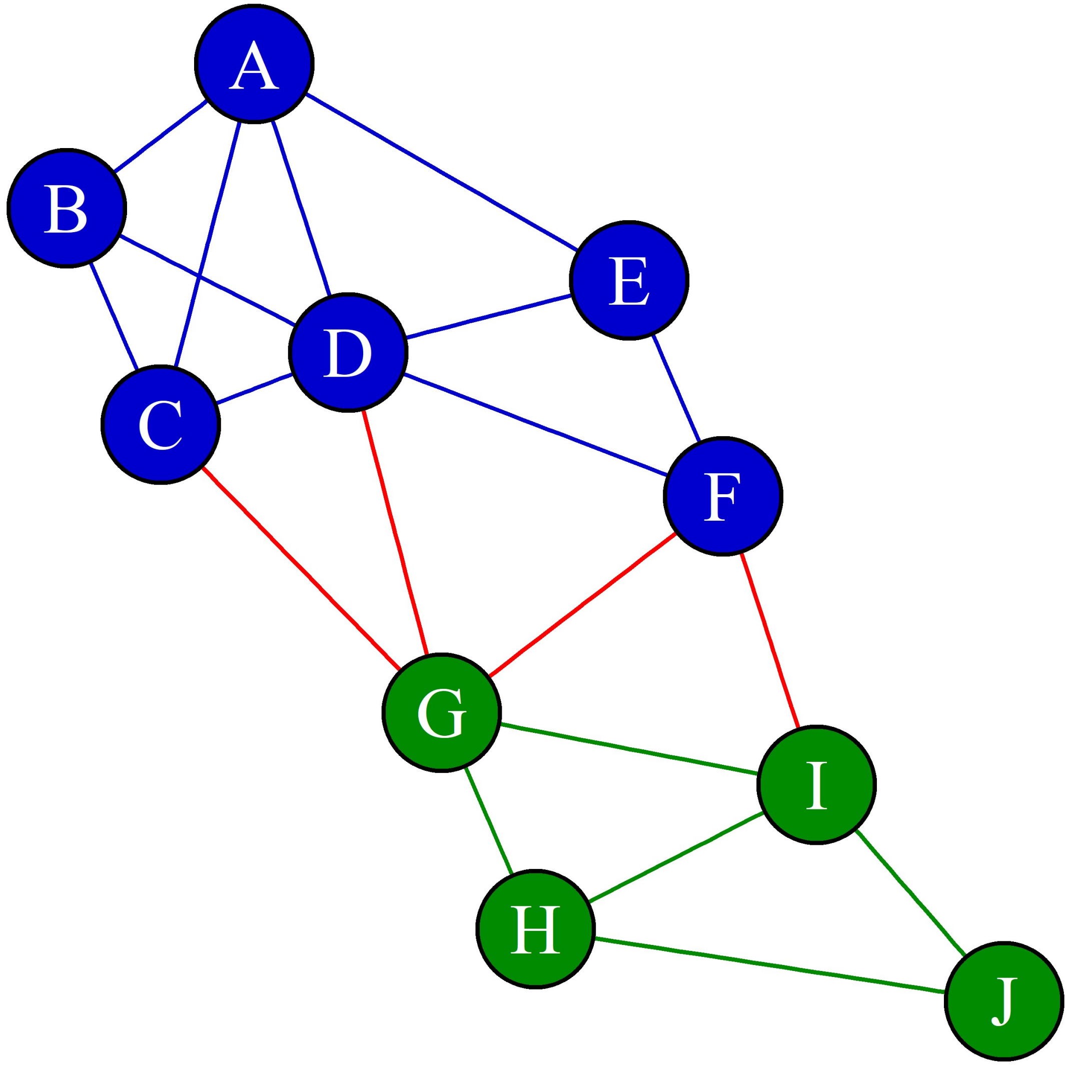
$H = 1.02$



$H = 0.37$



Heterophilicity in the network of data scientists



```
p<-2*19/(10*9)
m_rp<-6*4*p


H_rp<-edge_rp/m_rp
```

```
H_rp
```

```
0.3947368
```

Let's practice!

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Summary of homophily

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

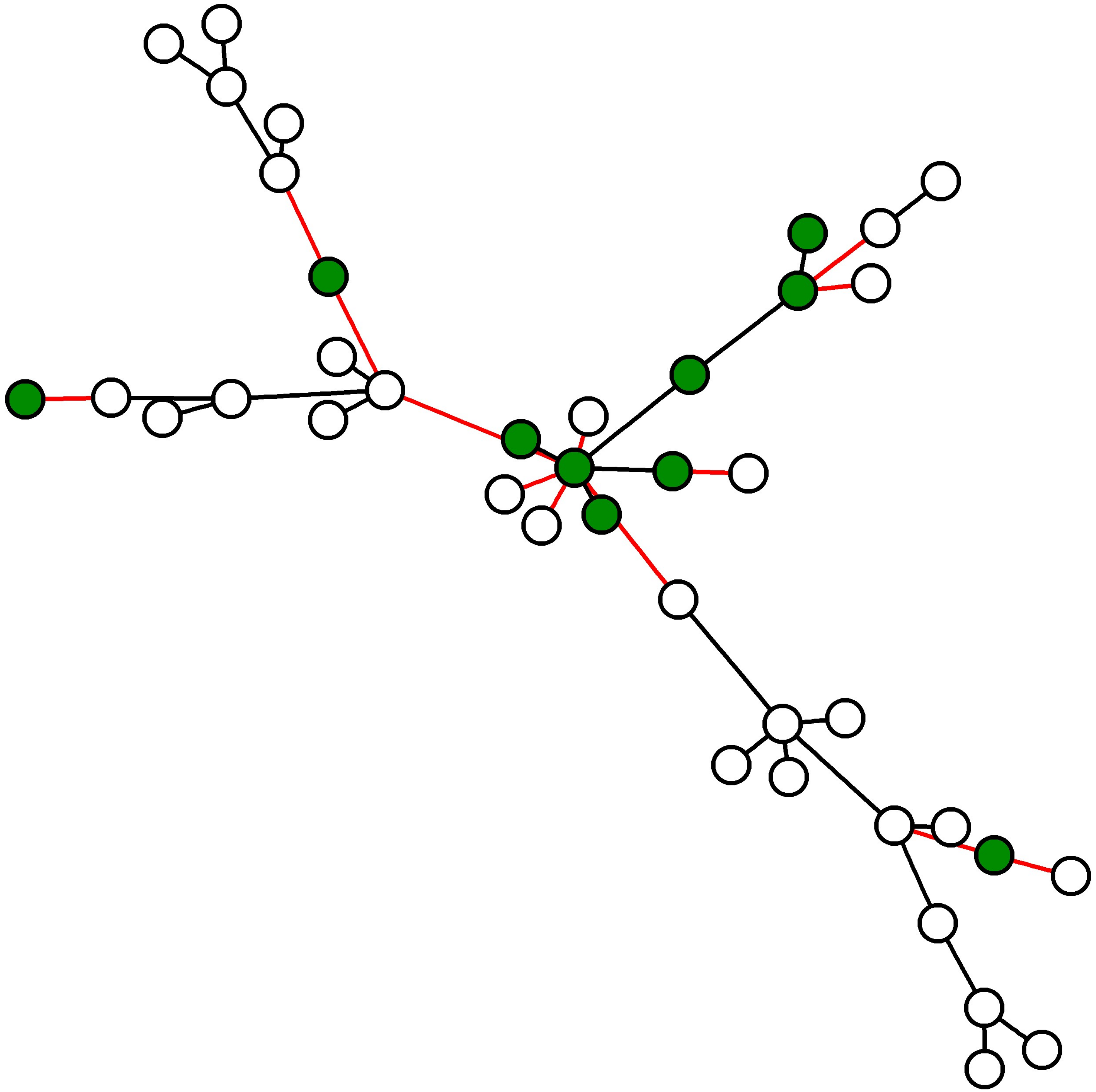
María Óskarsdóttir, Ph.D.
Postdoctoral researcher

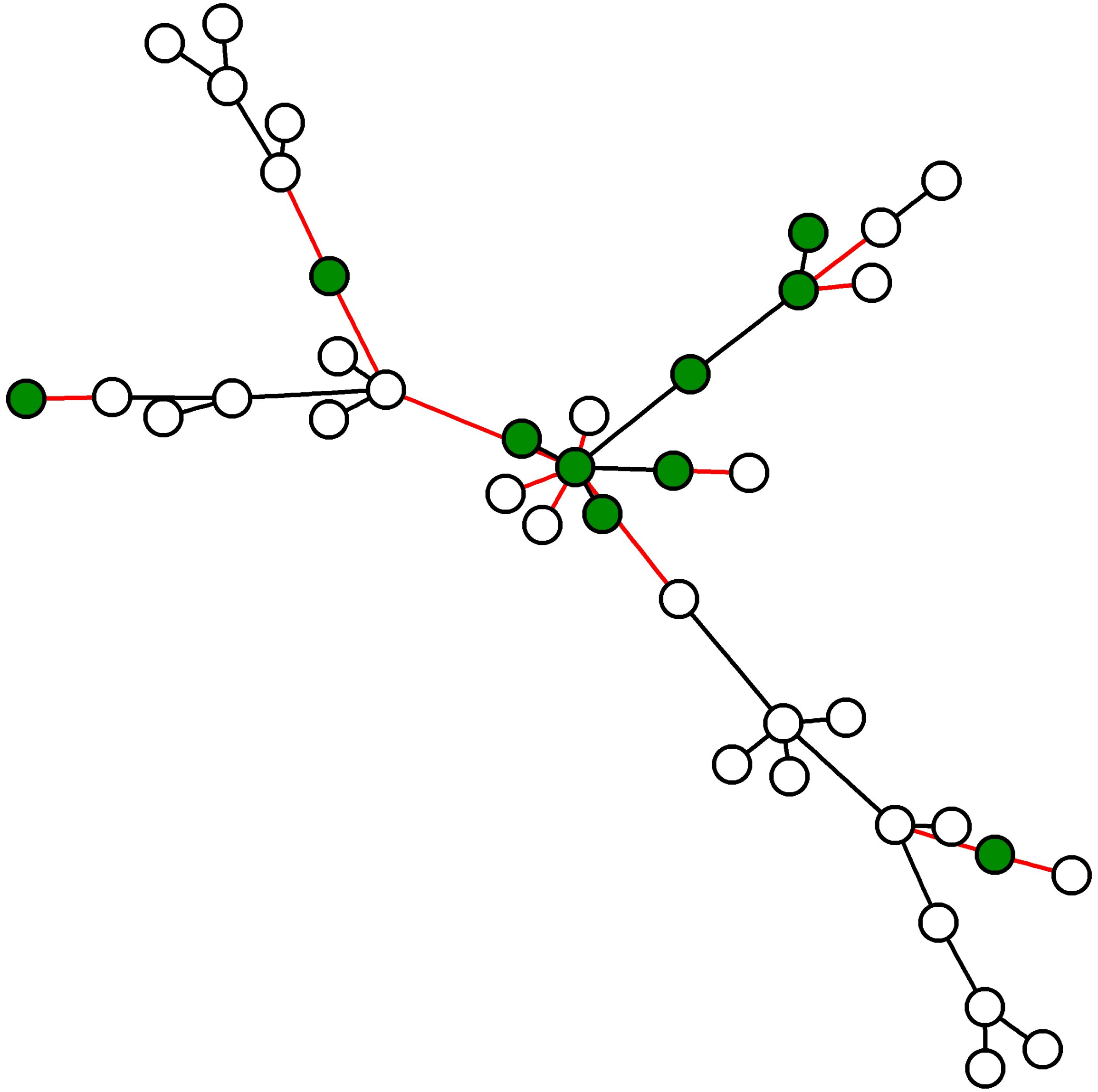
Can I do predictive analytics with my network?

Are the relationships between nodes important?

Are the labels randomly spread through the network or is there some structure?

Is the network homophilic?





```
# Heterophilicity
e_mixed / m_mixed
```

```
N <- 40
E <- 39
n_green <- 10
n_white <- 30
e_green <- 6
e_mixed <- 13
p <- 2 * E / N / (N-1)
m_green <- n_green * (n_green-1)/2 * p
m_mixed <- n_green * n_white * p
# Dyadicity
e_green / m_green
```

```
2.666667
```

```
0.8666667
```

⇒ Homophilic

Let's practice!

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R