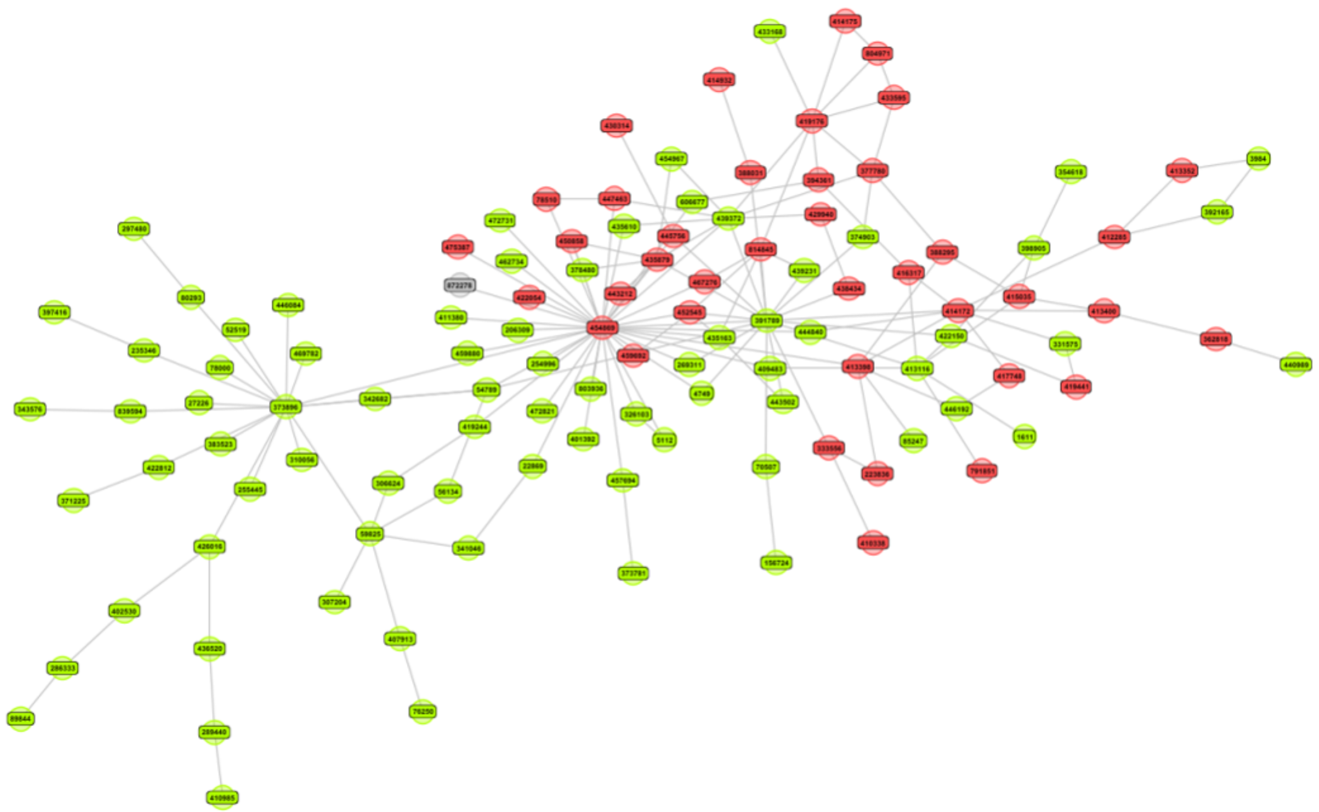
Motivation: social networks and predictive analytics

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Bart Baesens, Ph.D.
Professor of Data Science, KU Leuven
and University of Southampton

Applications

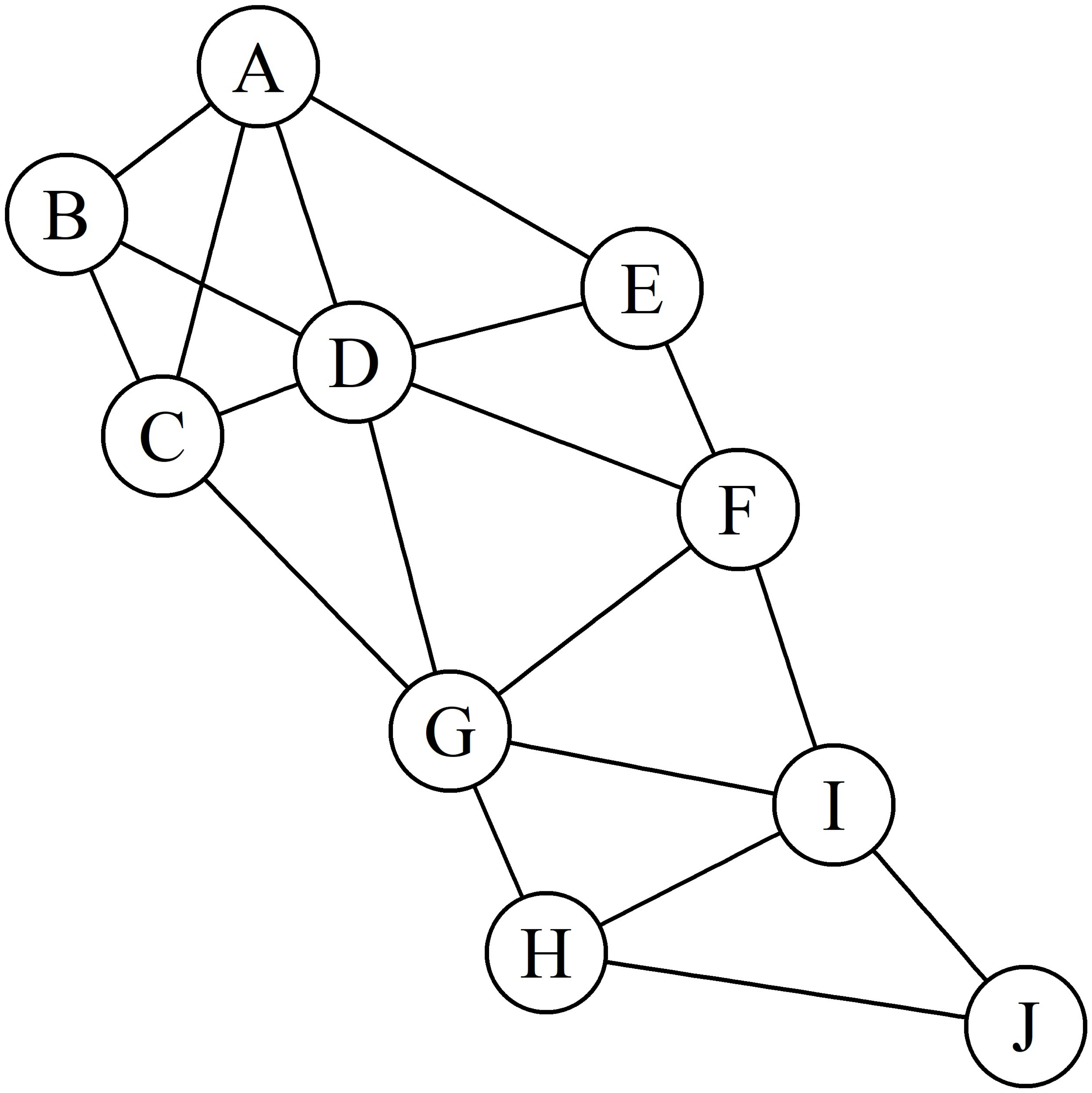
- Age
- Gender
- Fraud



- Churn
  - Customer defection
  - Companies predict who is most likely to churn using
    1. Machine learning techniques
    2. Social networks

Overview

- Labeled social networks
  - Construct and label networks
  - Network learning
- Homophily
  - Measure relational dependency
  - Heterophilicity and dyadicity
- Network featurization
  - Compute node features
- Predictive modeling with networks
  - Turn a network into a flat dataset
  - Predict churn among customers


A
B
C
D
E
F
G
H
I
J

Collaboration Network

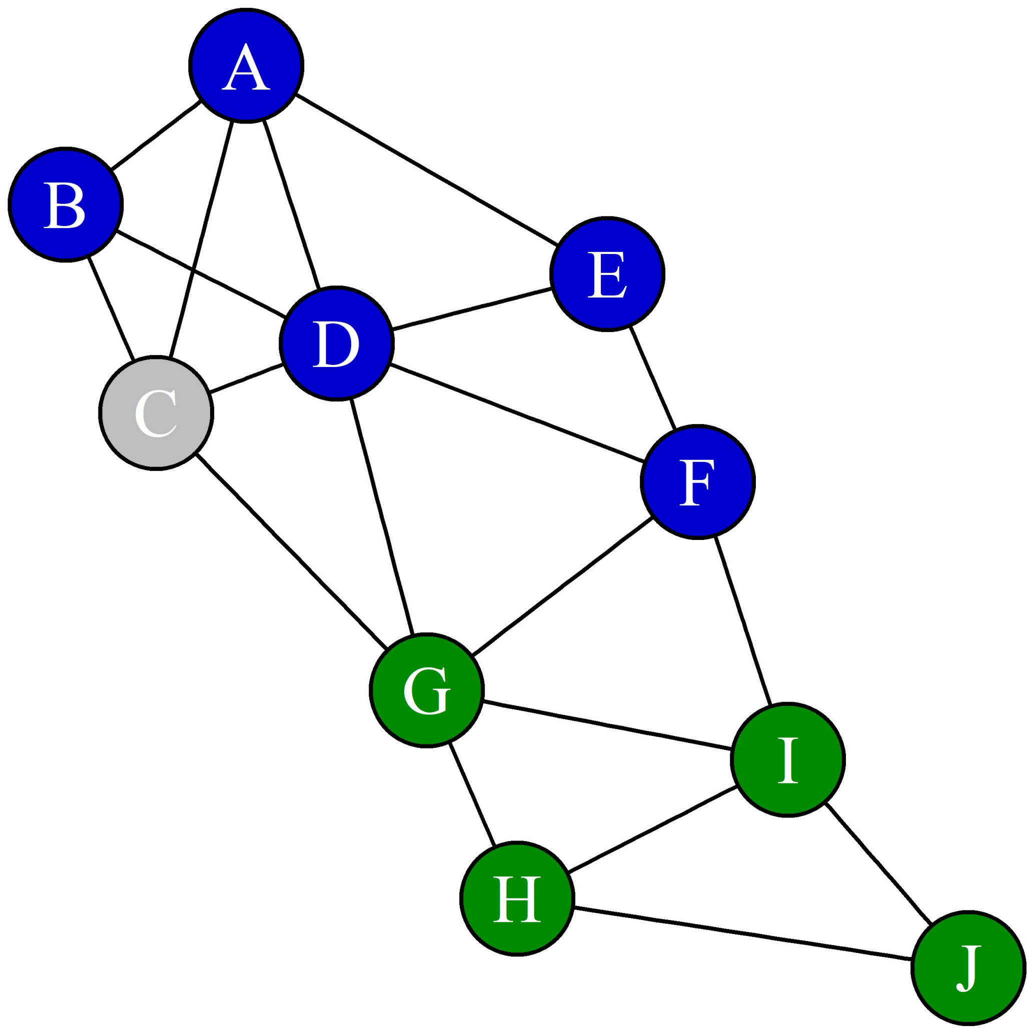
```
library(igraph);
DataScienceNetwork <- data.frame(
  from = c('A', 'A', 'A', 'A', 'B', 'B', 'C', 'C', 'D', 'D', 'D', 'E',
          'F', 'F', 'G', 'G', 'H', 'H', 'I'),
  to = c('B','C','D','E','C','D','D', 'G','E', 'F','G','F','G','I',
          'I','H','I','J','J'))
g <- graph_from_data_frame(DataScienceNetwork, directed = FALSE)
```

```
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.igraph(g, edge.label = NA, edge.color = 'black', layout = pos,
            vertex.label = V(g)$name, vertex.color = 'white',
            vertex.label.color = 'black', vertex.size = 25)
```

Collaboration Network

```
V(g)$technology <-
  c('R','R','?','R','R',
    'R','P','P','P','P')
V(g)$color <- V(g)$technology
```

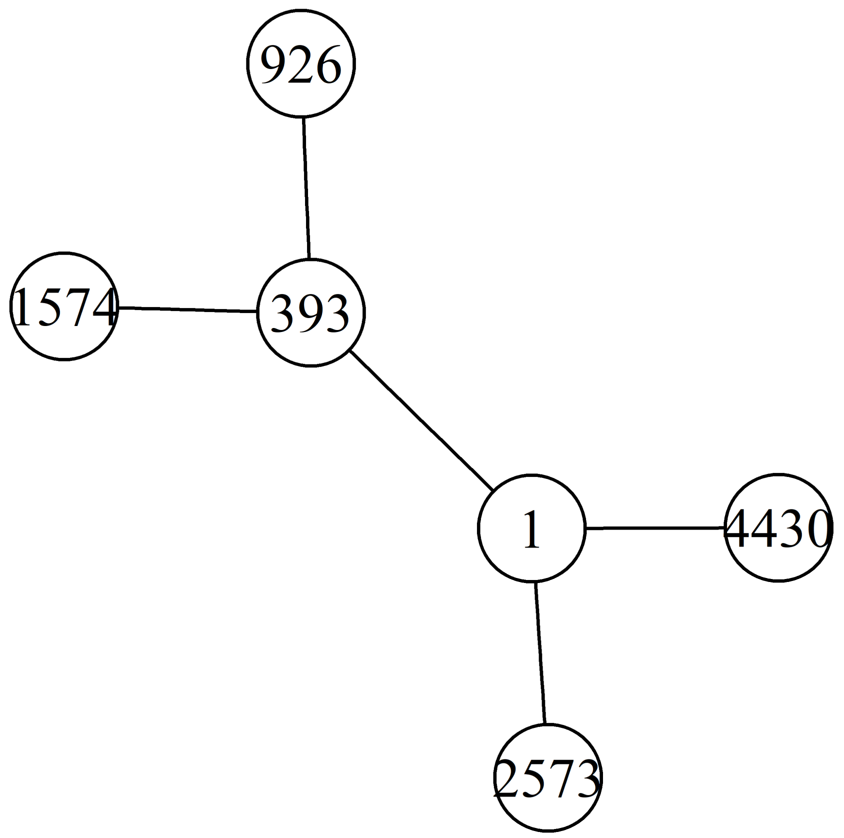
```
V(g)$color <- gsub('R',"blue3", V(g)$color)
V(g)$color <- gsub('P',"green4", V(g)$color)
V(g)$color <- gsub('?',"gray", V(g)$color)
```



Churn Network

```
edgeList
```

```
  from   to
1    1  393
2    1 2573
3    1 4430
4  393  926
5  393 1574
```



Let's practice!

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Labeled networks and network learning

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

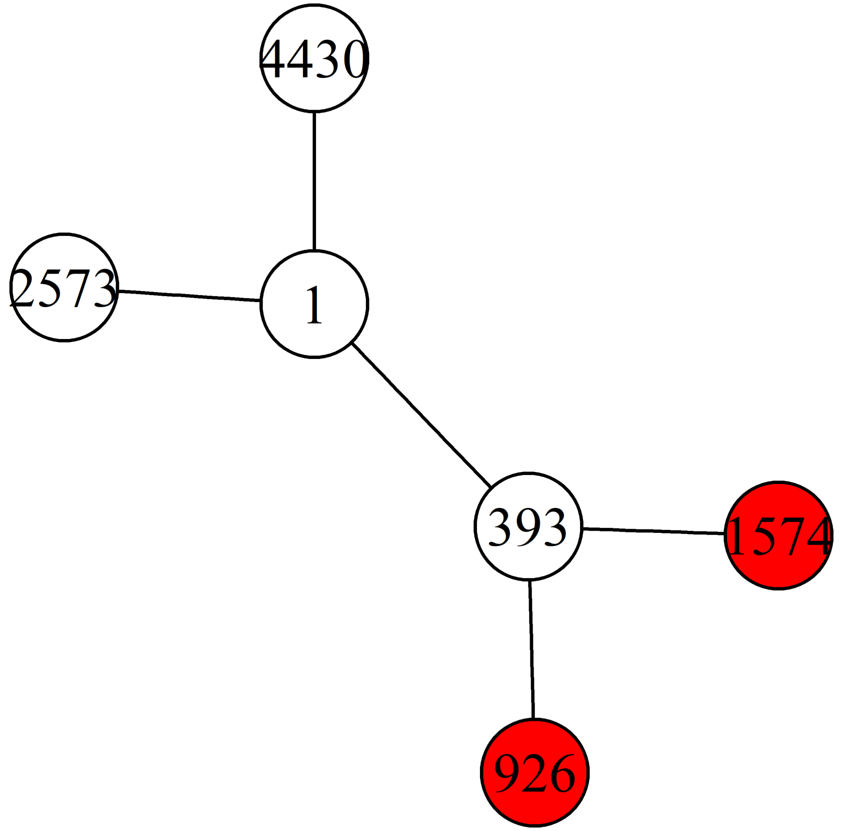
María Óskarsdóttir, Ph.D.
Post-doctoral researcher

```
customers
```

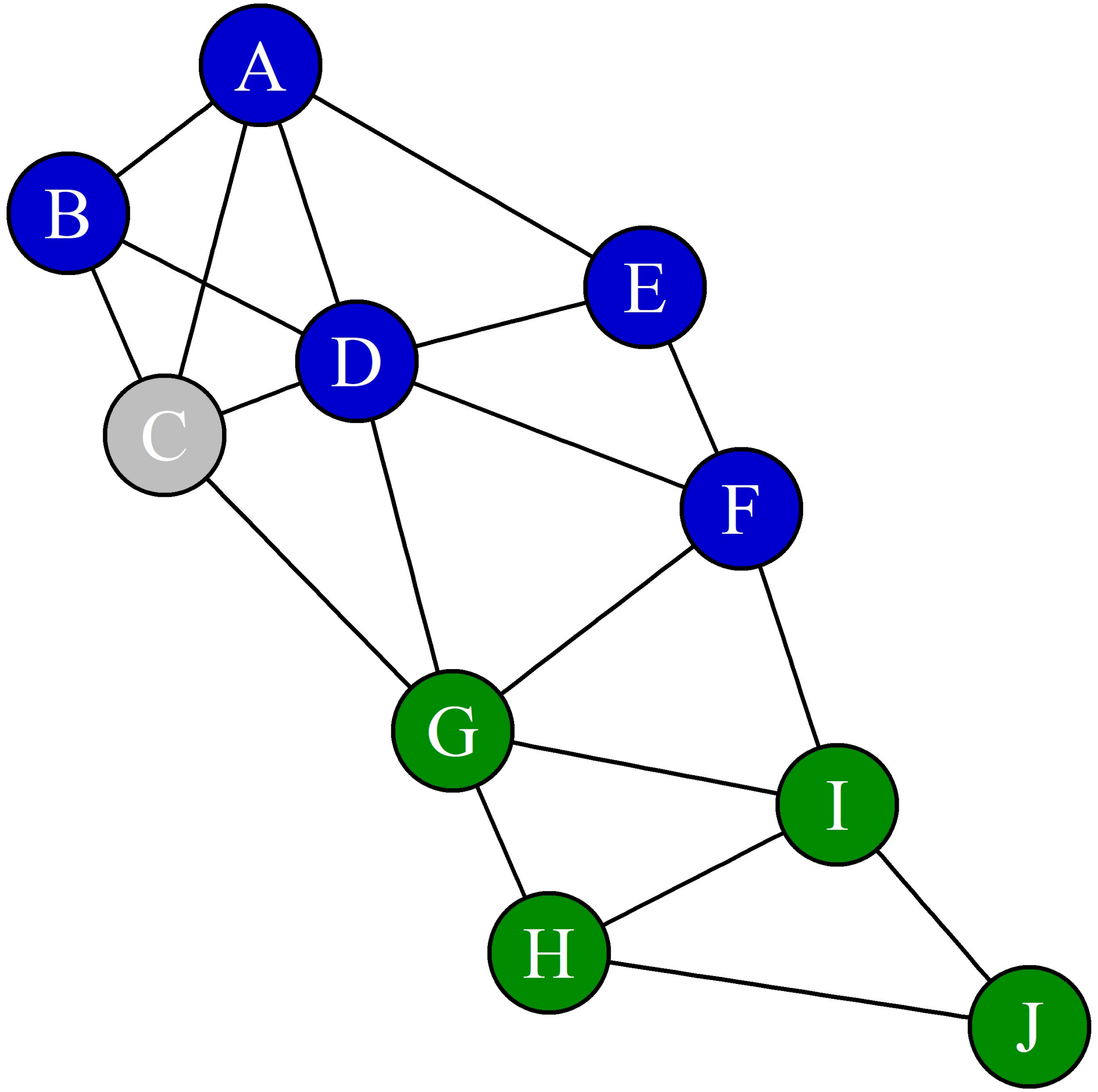
```
     id churn
1     1     0
2   393     0
3  2573     0
4  4430     0
5   926     1
6  1574     1
```

```
edgeList
```

```
  from   to
1    1  393
2    1 2573
3    1 4430
4  393  926
5  393 1574
```

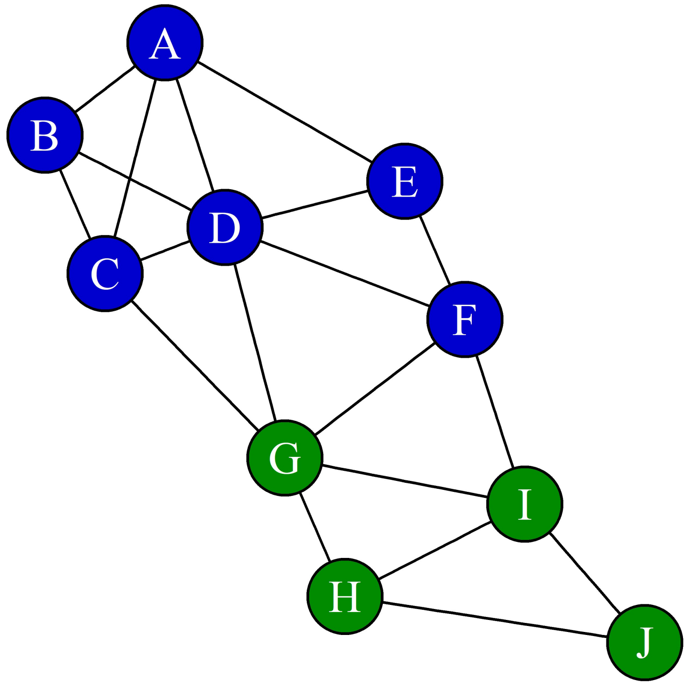


The Relational Neighbor Classifier



- Neighbors of **C**ecelia
  - A,B,D,G
- Neighbors of **C**ecelia that prefer R
  - A, B, D (75%)
- Neighbors of **C**ecelia that prefer Python
  - G (25%)
- **C**ecelia has a higher probability to prefer R

The Relational Neighbor Classifier




```
rNeighbors <- c(4,3,3,5,3,2,3,0,1,0)
pNeighbors <- c(0,0,1,1,0,2,2,3,3,2)
rRelationalNeighbor <- rNeighbors / (rNeighbors + pNeighbors)
rRelationalNeighbor
```

```
1.00 1.00 0.75 0.86 1.00 0.50 0.60 0.00 0.00 0.00
```

Let's practice!

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

Challenges of network-based inference

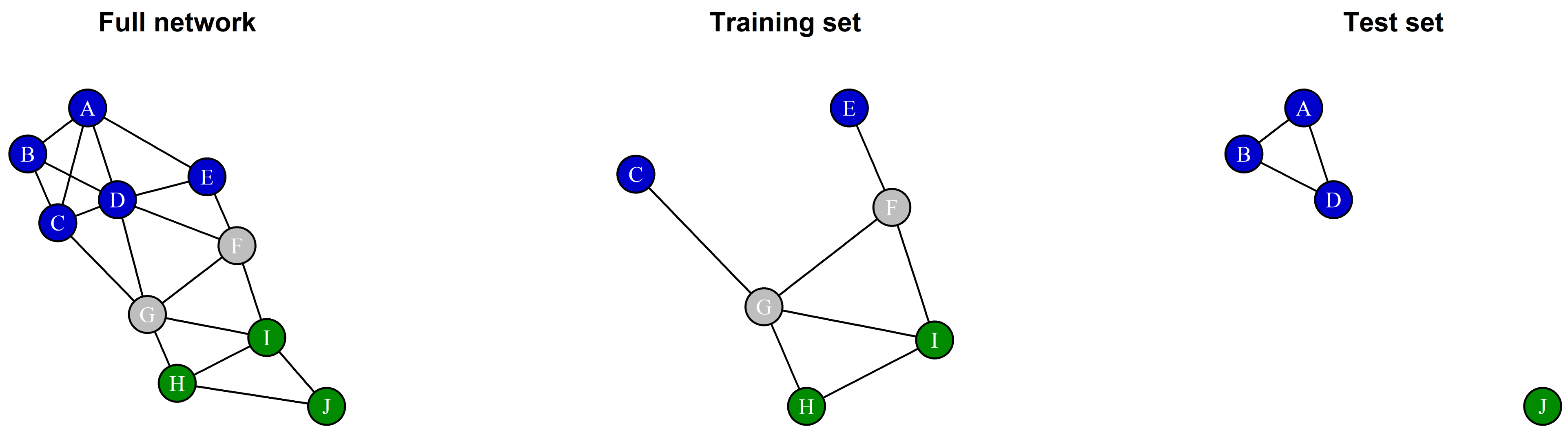
PREDICTIVE ANALYTICS USING NETWORKED DATA IN R

María Óskarsdóttir, Ph.D.
Post-doctoral researcher

First challenge

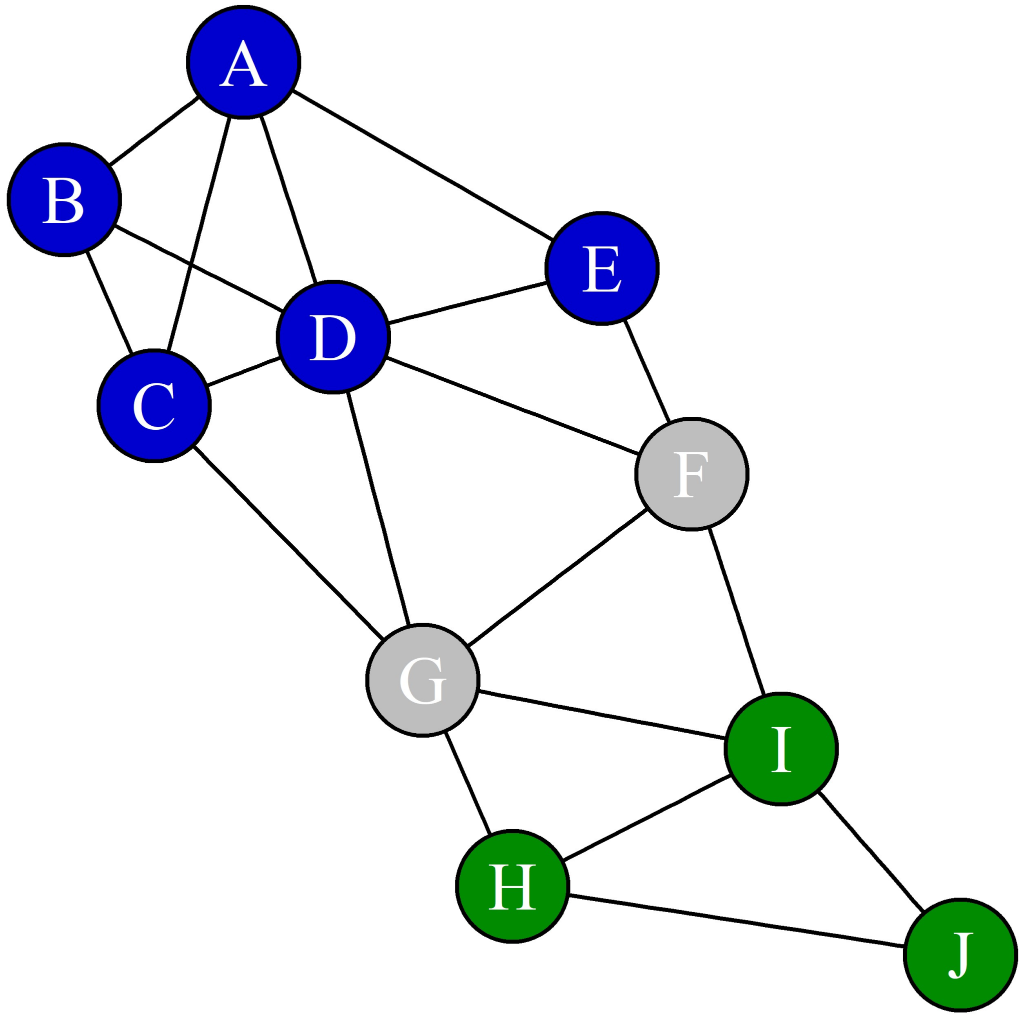
Splitting the data!

```
set.seed(1001)
sampleVertices <- sample(1:10, 6, replace=FALSE)
plot(induced_subgraph(g, V(g)[sampleVertices]))
plot(induced_subgraph(g, V(g)[-sampleVertices]))
```



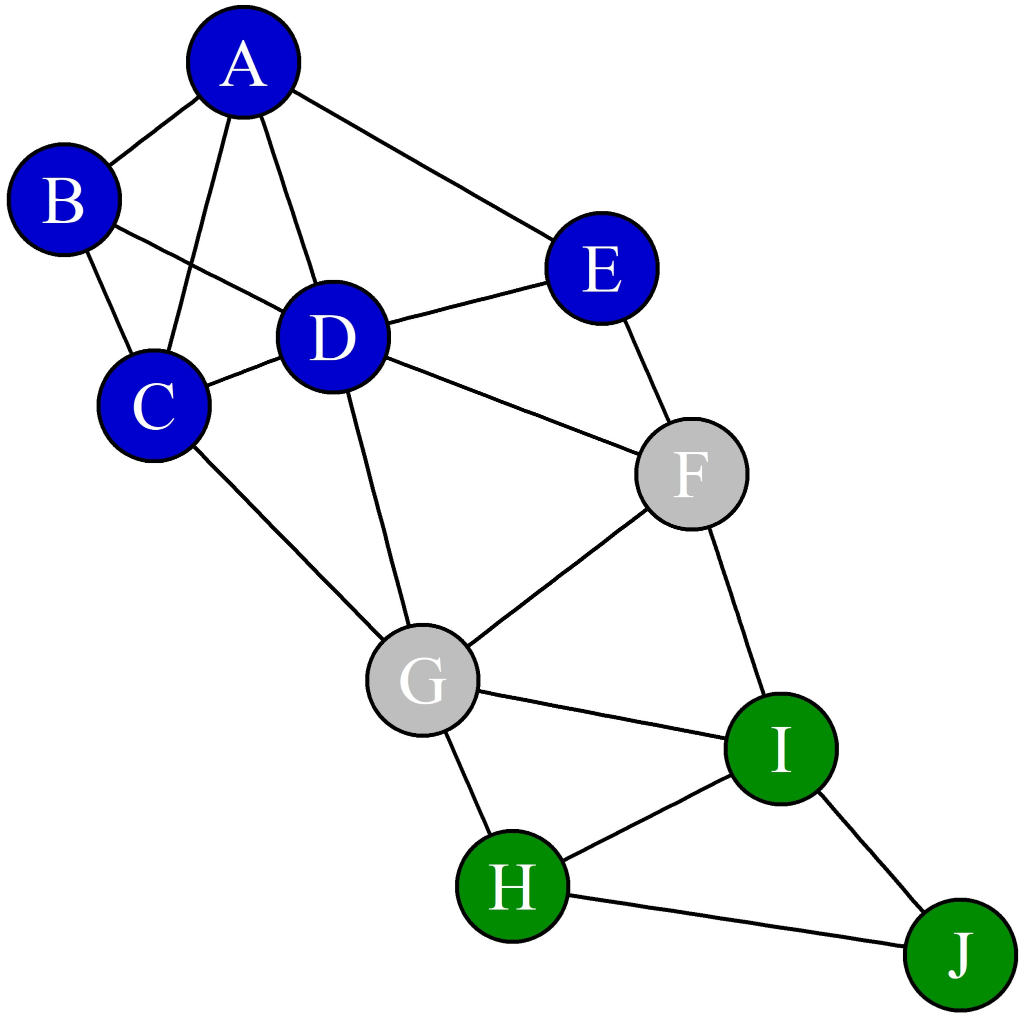
Second challenge

The observations in the dataset are not independent and identically distributed (iid)

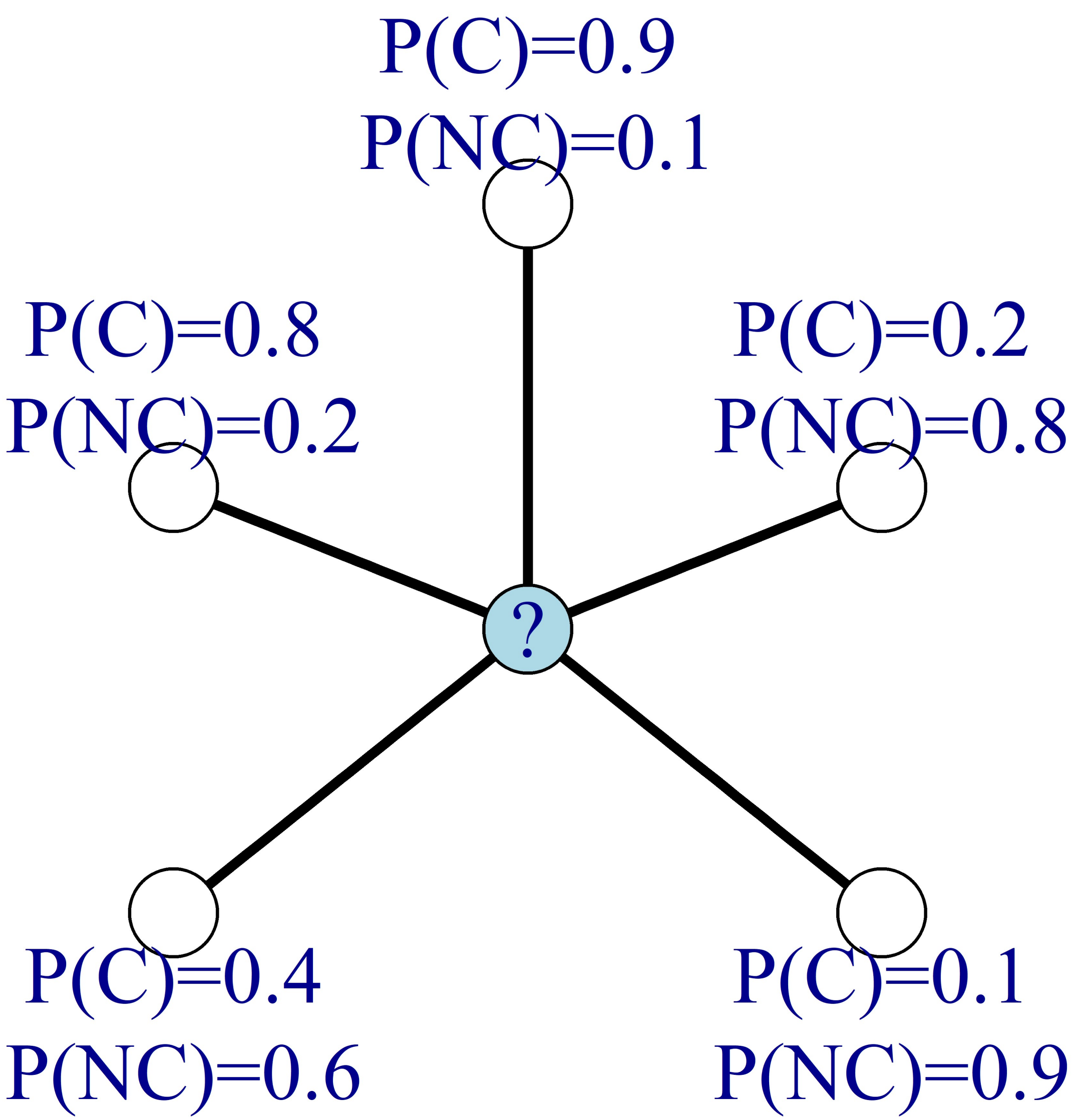


Third challenge

Collective Inference!



Probabilistic relational neighbor classifier




```
# probability churn (C)
(0.9 + 0.2 + 0.1 + 0.4 + 0.8) / 5
```

```
0.48
```

```
# probability non-churn (NC)
(0.1 + 0.8 + 0.9 + 0.6 + 0.2) / 5
```

```
0.52
```

Let's practice!

PREDICTIVE ANALYTICS USING NETWORKED DATA IN R