FRAUD DETECTION IN R

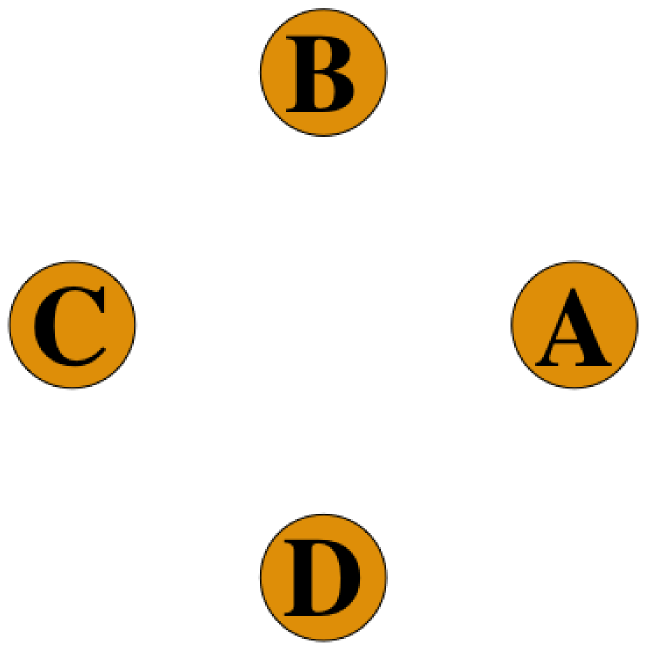
Social network analytics

Bart Baesens
Professor Data Science at KU Leuven

Social network components

Nodes (vertices)

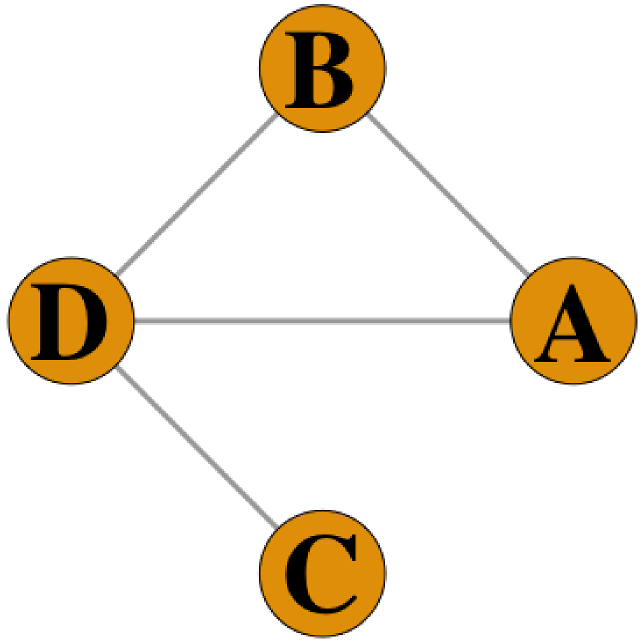
- customers
- companies
- products
- credit cards
- accounts
- web pages



Social network components

Edges

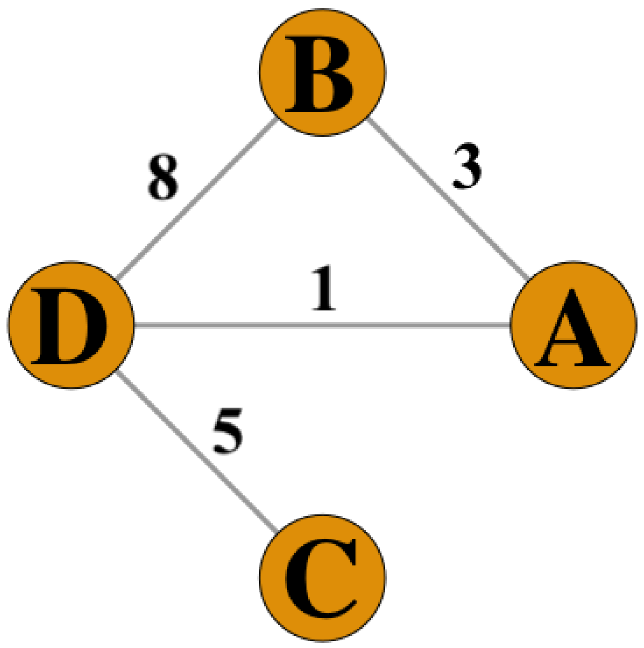
- Different kind of relationships, e.g. money transfer, call, friendship, transmission of a disease, reference



Social network components

Edges

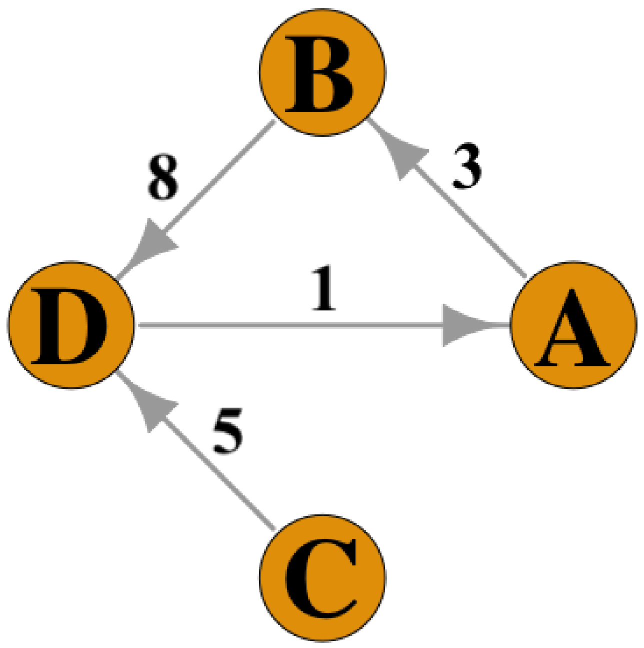
- Different kind of relationships, e.g. money transfer, call, friendship, transmission of a disease, reference
- Weighted based on e.g. interaction frequency, importance of information exchange, intimacy, emotional intensity



Social network components

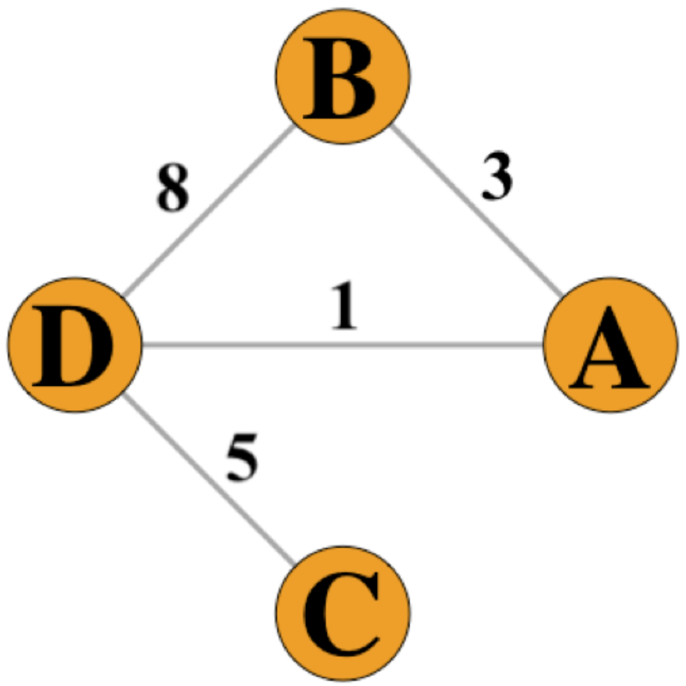
Edges

- Different kind of relationships, e.g. money transfer, call, friendship, transmission of a disease, reference
- Weighted based on e.g. interaction frequency, importance of information exchange, intimacy, emotional intensity
- Directed, e.g. incoming or ougoing



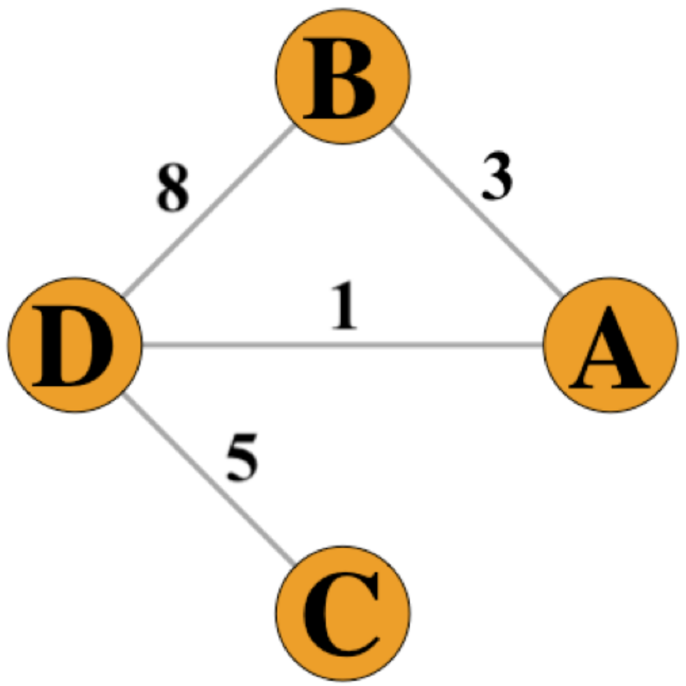
Social network representation

Sociogram



Social network representation

Sociogram

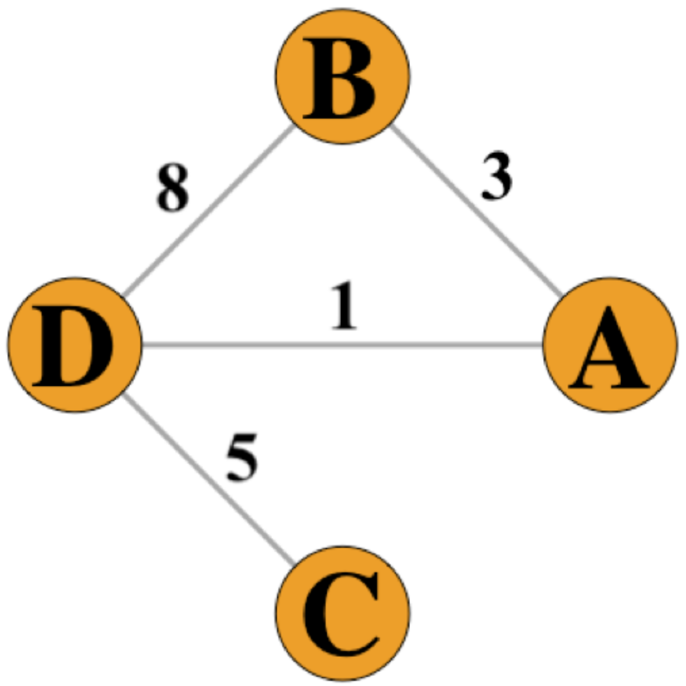


Connectivity Matrix

	A	B	C	D
A	0	1	0	1
B	1	0	0	1
C	0	0	0	1
D	1	1	1	0

Social network representation

<u>Sociogram</u>



<u>Connectivity Matrix</u>

	A	B	C	D
A	0	1	0	1
B	1	0	0	1
C	0	0	0	1
D	1	1	1	0

<u>Adjaceny List</u>

(A, B) (A, D)

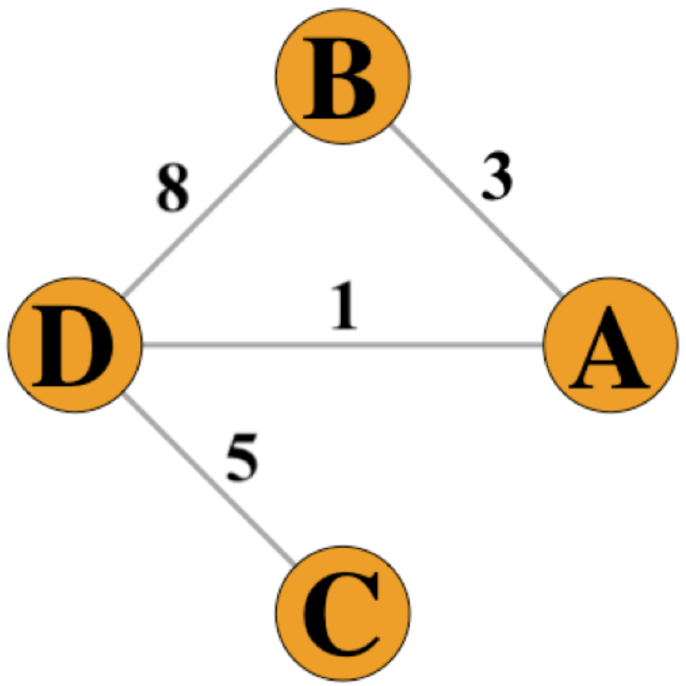
(B, A) (B, D)

(C, D)

(D, A) (D, B) (D, C)

Social network representation

Sociogram



Connectivity Matrix

	A	B	C	D
A	0	3	0	1
B	3	0	0	8
C	0	0	0	5
D	1	8	5	0

Adjaceny List

(A, B, 3) (A, D, 1)

(B, A, 3) (B, D, 8)

(C, D, 5)

(D, A, 1) (D, B, 8) (D, C, 5)

Towards a network

- From a transactional data source ...

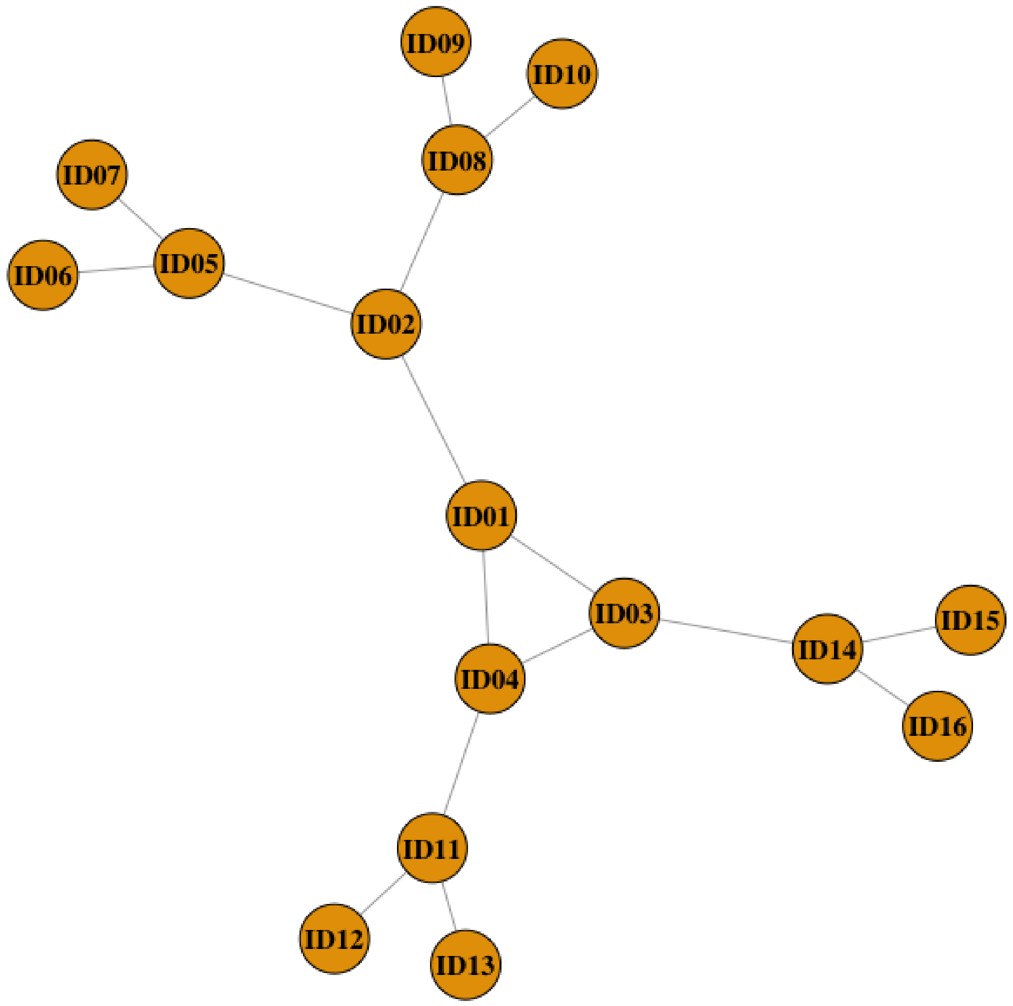
```
> print(transactions)

   originator beneficiary amount  time benef_country payment_channel
1        ID14        ID16    102 22:47           GBR         CHAN_04
2        ID14        ID15    125 20:21           USA         CHAN_02
3        ID02        ID01   1067 10:45           CAN         CHAN_04
4        ID05        ID06     59 15:40           USA         CHAN_02
5        ID05        ID07     99 14:41           USA         CHAN_02
...       ...         ...    ...   ...           ...             ...
15       ID08        ID09    145 18:23           USA         CHAN_01
16       ID03        ID04   1039 21:20           USA         CHAN_02
```

- ... towards a network

```
> library(igraph)
> network <- graph_from_data_frame(transactions, directed = FALSE)
```

Plotting a network

```
> plot(network)
```



A network's edges and nodes

- Edges

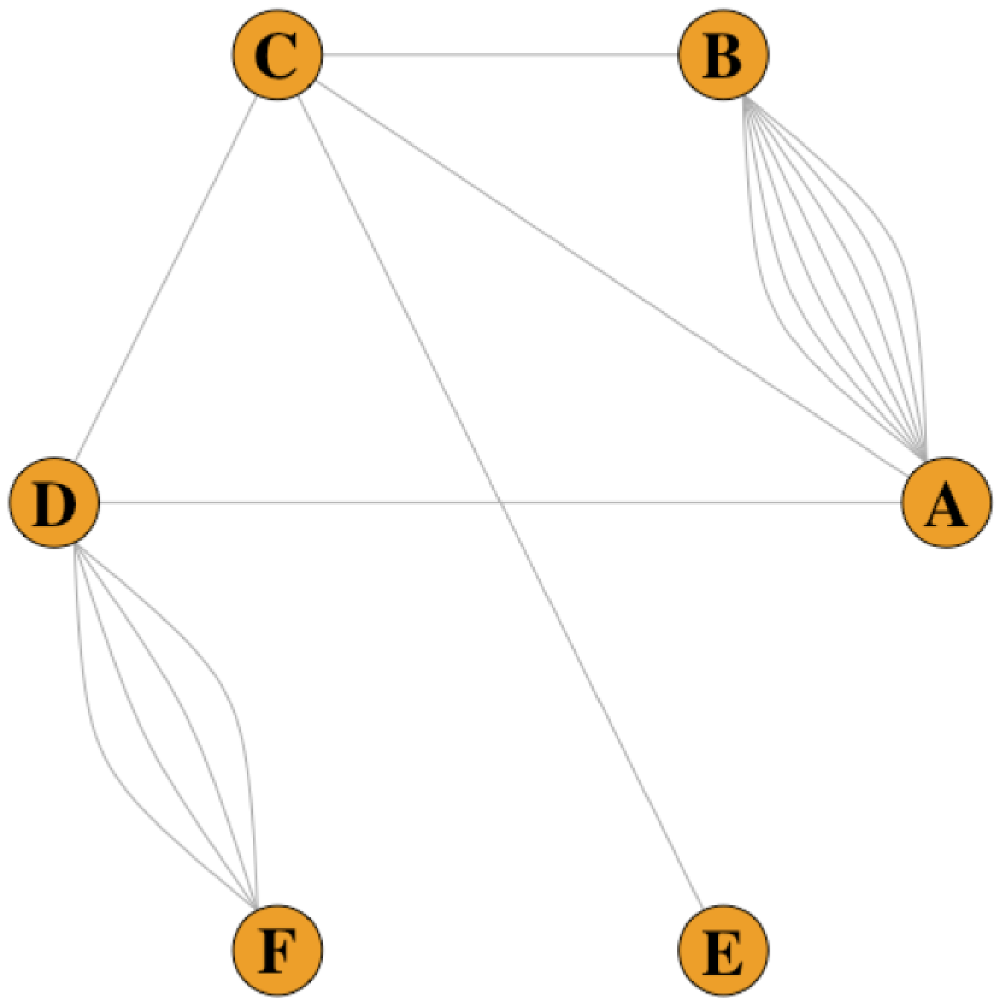
```
> E(network)

+ 16/16 edges from 297af3c (vertex names):
 [1] ID02--ID01 ID11--ID04 ID04--ID01 ID04--ID03 ID03--ID01 ID08--ID09
 [7] ID14--ID15 ID03--ID14 ID05--ID06 ID11--ID12 ID02--ID05 ID11--ID13
[13] ID02--ID08 ID14--ID16 ID08--ID10 ID05--ID07
```

- Vertices (nodes)

```
> V(network)

+ 16/16 vertices, named, from 297af3c:
 [1] ID02 ID11 ID04 ID03 ID08 ID14 ID05 ID01 ID09 ID15 ID06 ID12 ID13 ID16
[15] ID10 ID07

> V(network)$name

 [1] "ID02" "ID11" "ID04" "ID03" "ID08" "ID14" "ID05" "ID01" "ID09" "ID15"
[11] "ID06" "ID12" "ID13" "ID16" "ID10" "ID07"
```

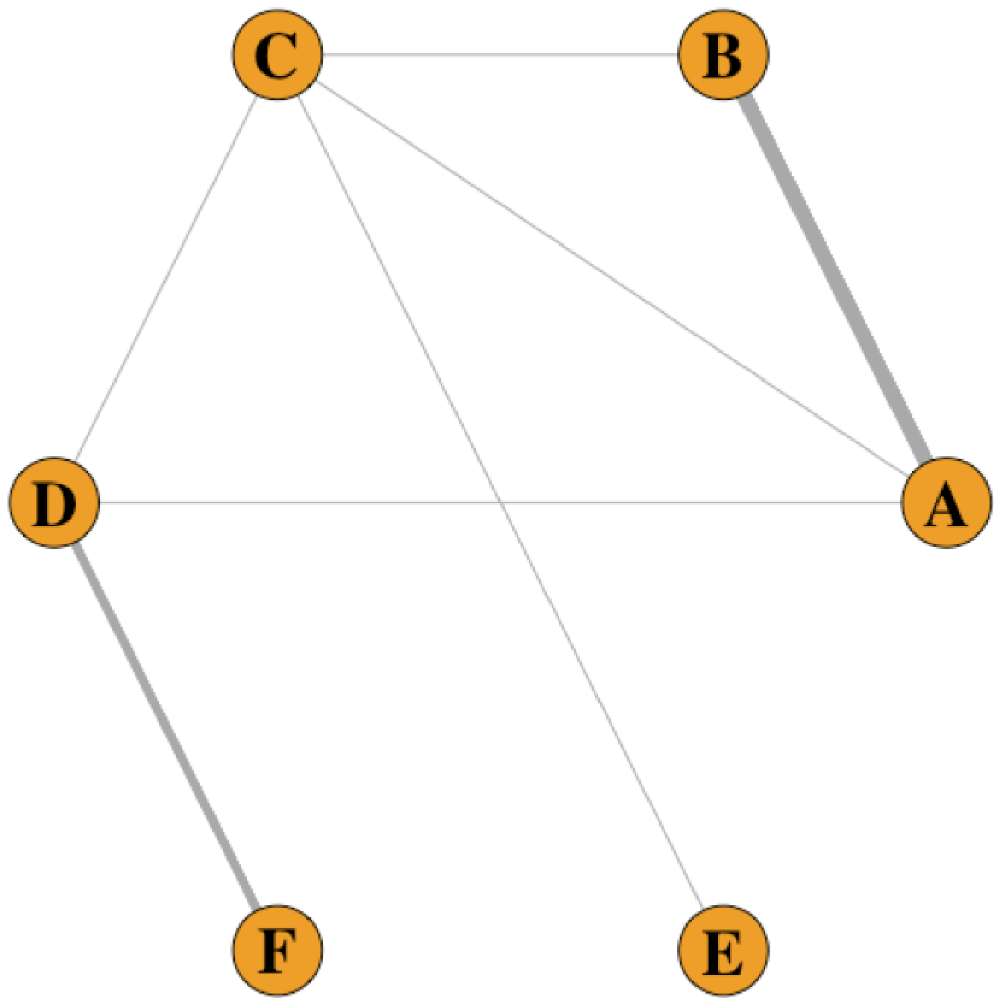
Overlapping edges

```
> plot(net)

> E(net)$width <- count.multiple(net)

> edge_attr(net)

$width
 [1] 7 7 7 7 7 7 7 1 1 1 4 4 4 4 1 1
```



Overlapping edges

```
> E(net)$curved <- FALSE
> plot(net)
```



FRAUD DETECTION IN R

Let's practice!

FRAUD DETECTION IN R

Fraud and social network analysis

Bart Baesens
Professor Data Science at KU Leuven

Is fraud a social phenomenom?

- Intuition: *relationships* between people
- Are there effects indicating that fraud is a social phenomenon?

Is fraud a social phenomenom?

- Fraudsters tend to cluster together:
  - are attending the same events/activities
  - are involved in the same crimes
  - use the same resources
  - are sometimes one and the same person (identity theft)

Homophily

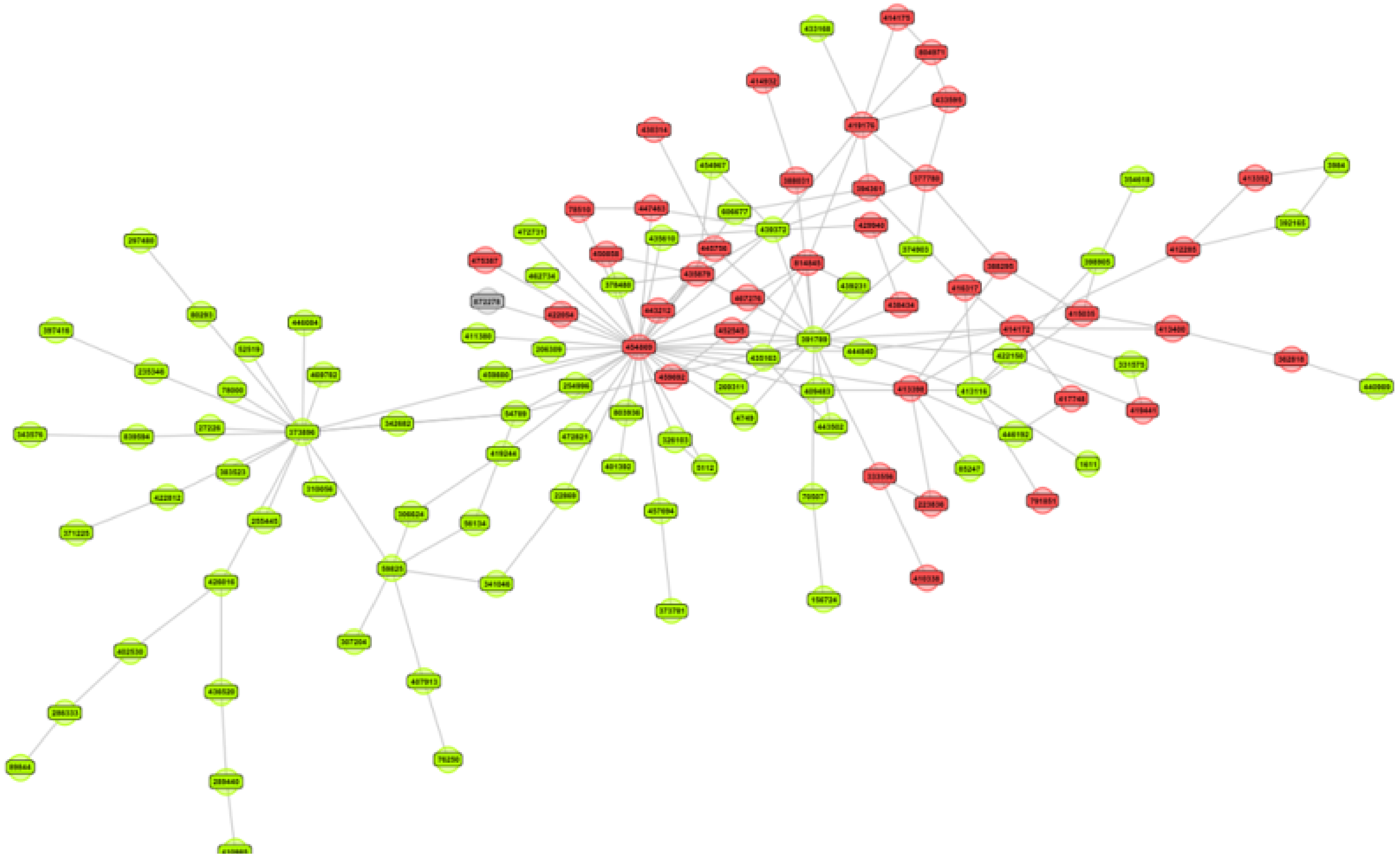
Homophily in social networks (from sociology)
People have a strong tendency to associate with other whom they perceive as being similar to themselves in some way.

Homophily in fraud networks
Fraudsters are more likely to be connected to other fraudsters, and legitimate people are more likely to be connected to other legitimate people.

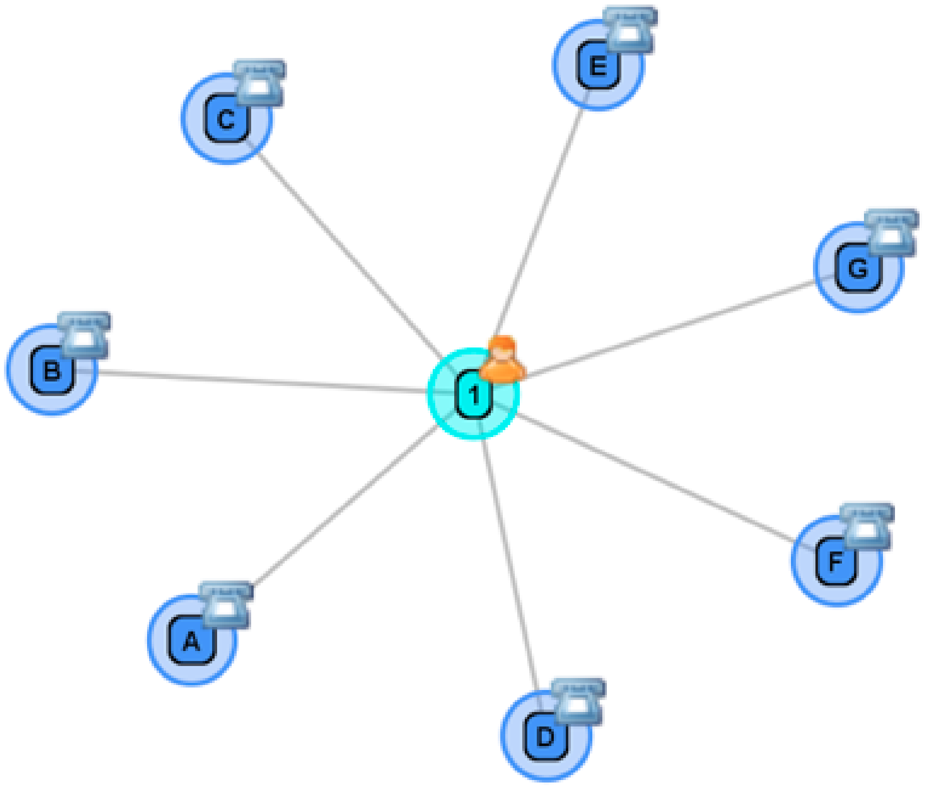
Homophily - social security fraud

Does the network contain statistically significant patterns of homophily?

```
> assortativity_nominal(network, types = V(network)$isFraud, directed = FALSE)
```

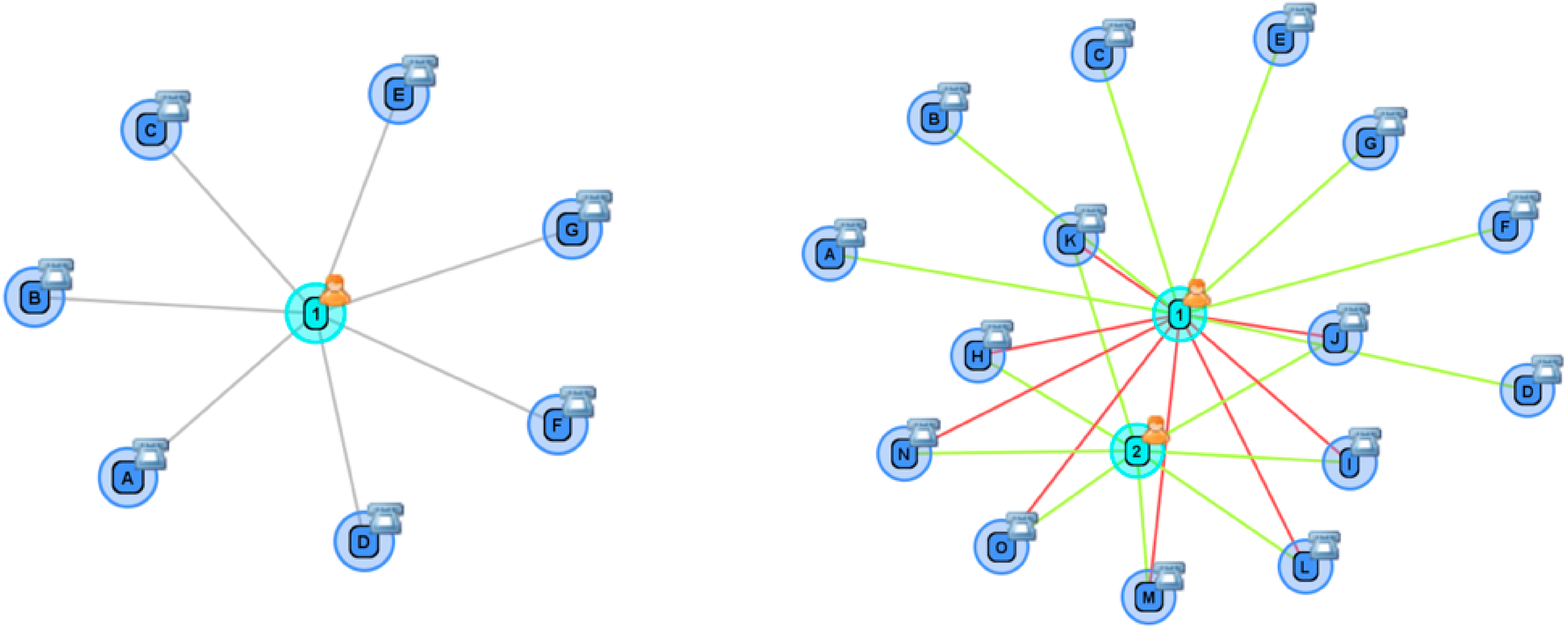


Identity theft



Before: person calls his/her frequent contacts.

Identity theft

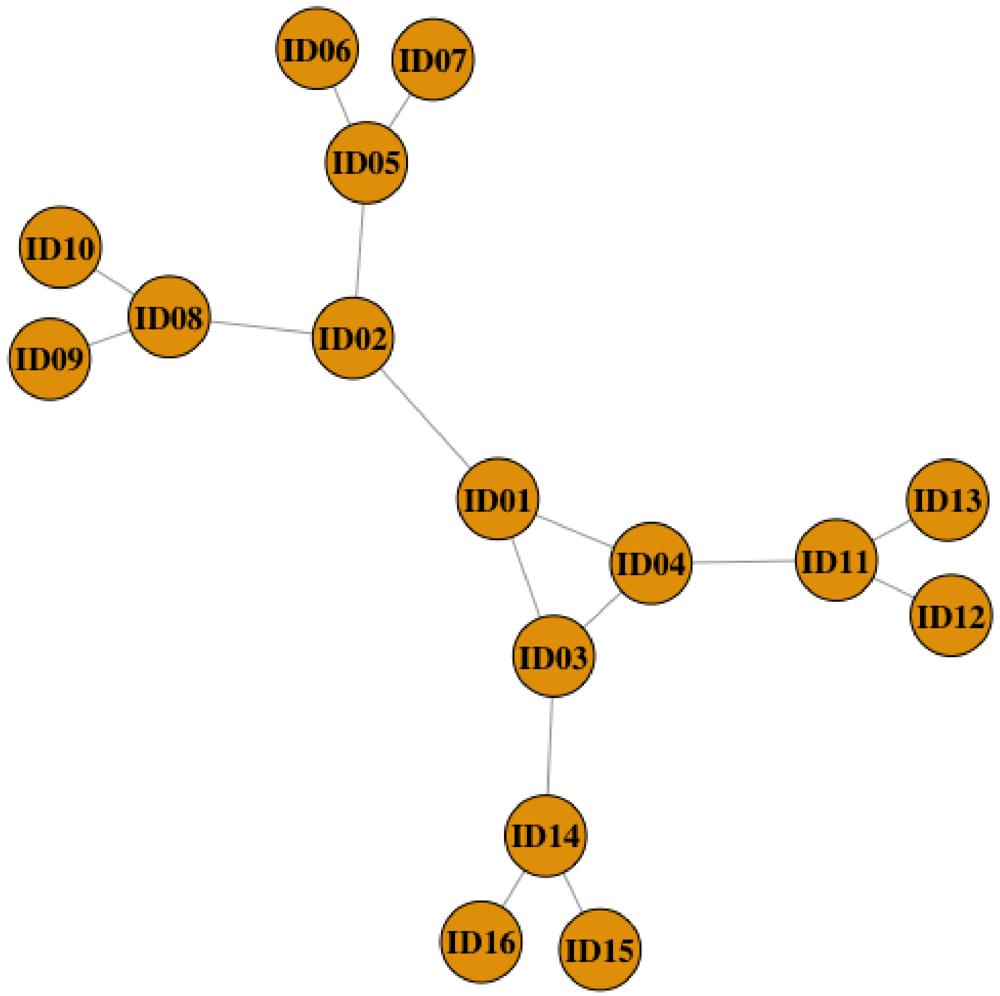


Before: person calls his/her frequent contacts.

After: person calls new contacts which *coincidentally* overlap with another persons contacts.

Money mules

- **Money mule** = person who transfers money acquired illegally (e.g. stolen)
- Beneficiary of fraudulent transaction
- Transfers stolen money on behalf of other (scam operator)

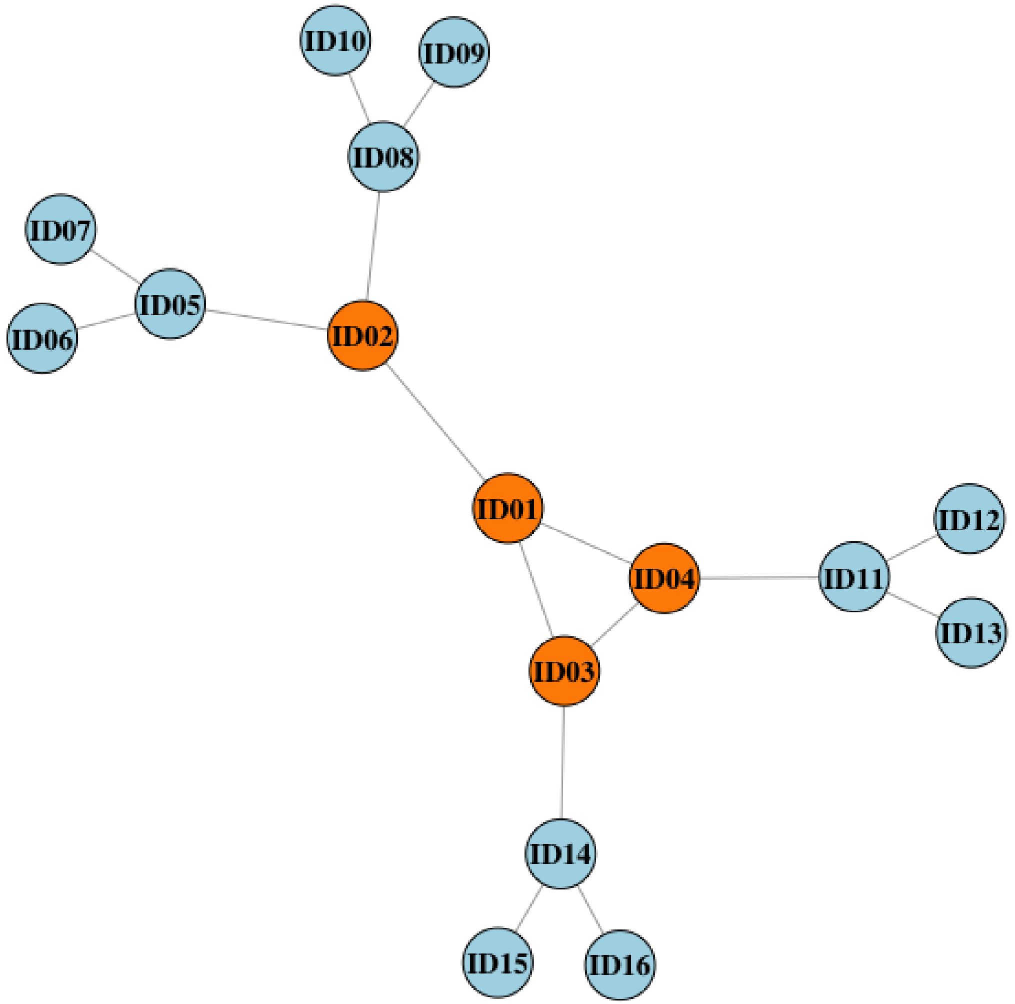


Add attributes to nodes

```
> V(network)$name

 [1] "ID02" "ID11" "ID04" "ID03" "ID08" "ID14" "ID05" "ID01" "ID09" "ID15"
[11] "ID06" "ID12" "ID13" "ID16" "ID10" "ID07"

> print(list_money_mules)

[1] "ID01" "ID02" "ID03" "ID04"

> V(network)$isMoneyMule <- ifelse(V(network)$name %in% list_money_mules,
                                    TRUE, FALSE)

> V(network)$color <- ifelse(V(network)$isMoneyMule,
                              "darkorange", "lightblue")

> vertex_attr(network)

$name
 [1] "ID02" "ID11" "ID04" "ID03" "ID08" ... "ID16" "ID10" "ID07"

$isMoneyMule
 [1]  TRUE FALSE  TRUE  TRUE FALSE ... FALSE FALSE FALSE

$color
 [1] "darkorange" "lightblue" "darkorange" ... "lightblue" "lightblue"
```

Network with highlighted money mules

```
> plot(network)
```



FRAUD DETECTION IN R

Let's practice!

FRAUD DETECTION IN R

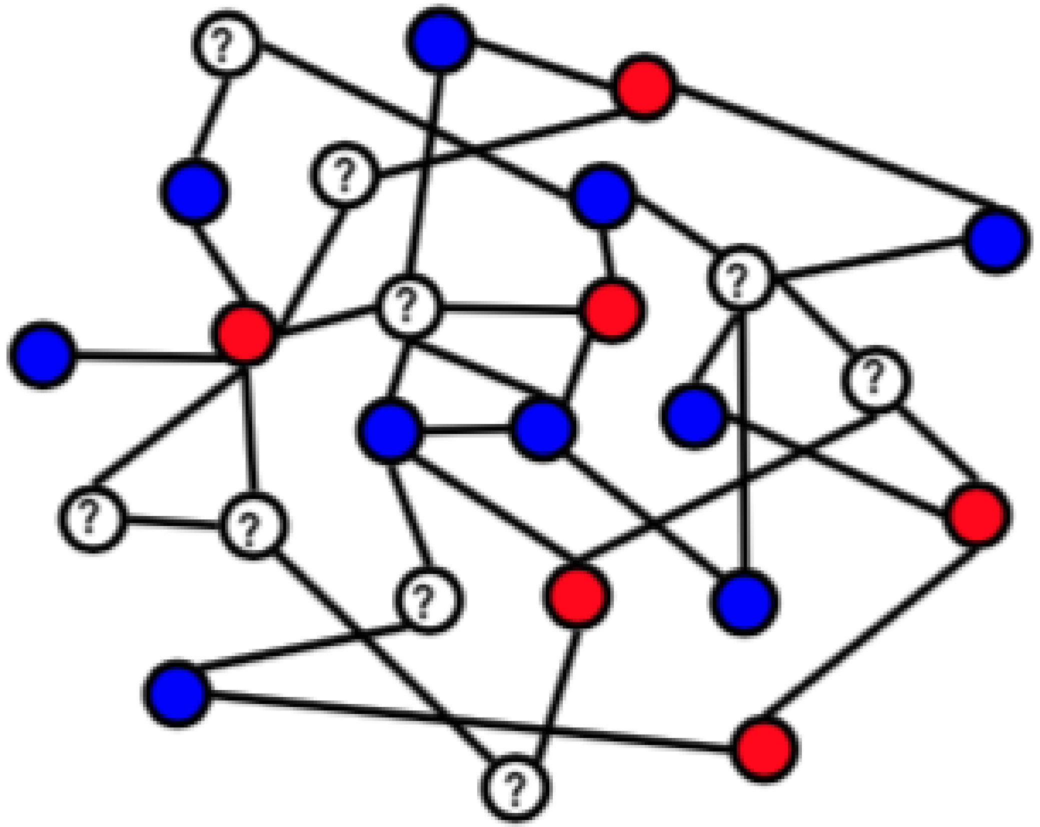
Social network based inference

Tim Verdonck
Professor Data Science at KU Leuven

Social network based inference

Goal

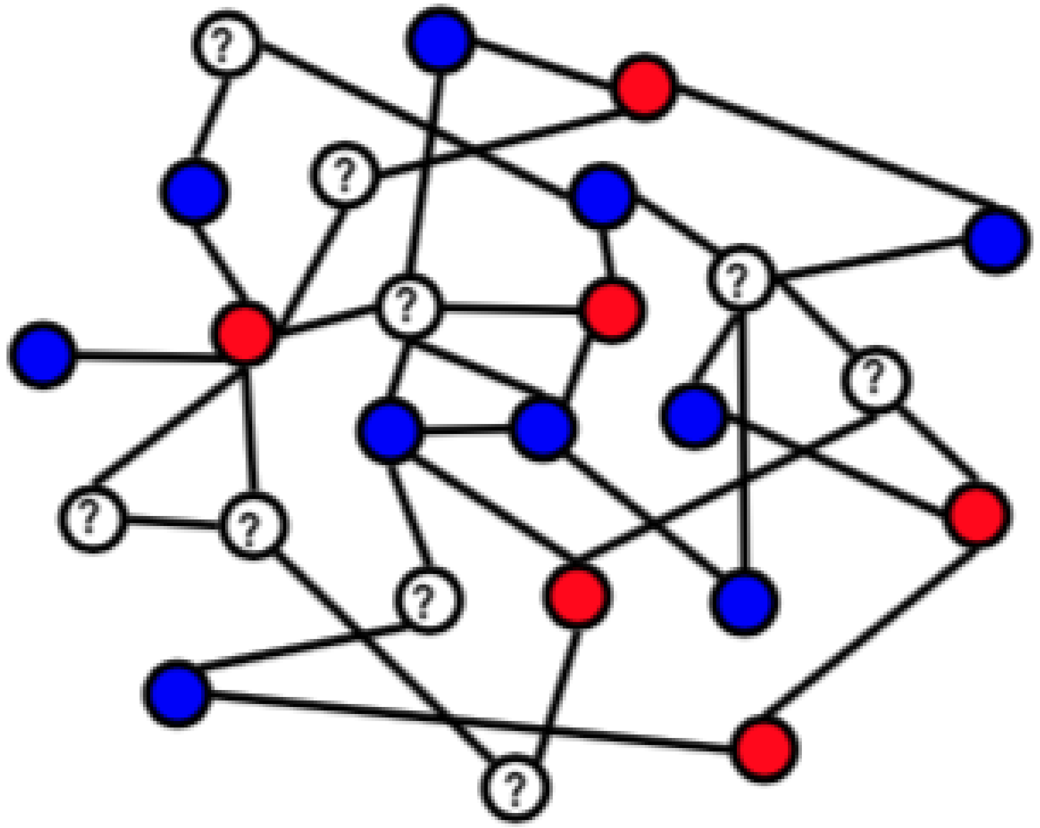
Predict the behavior of a node based on the behavior of other nodes



Social network based inference

Challenges

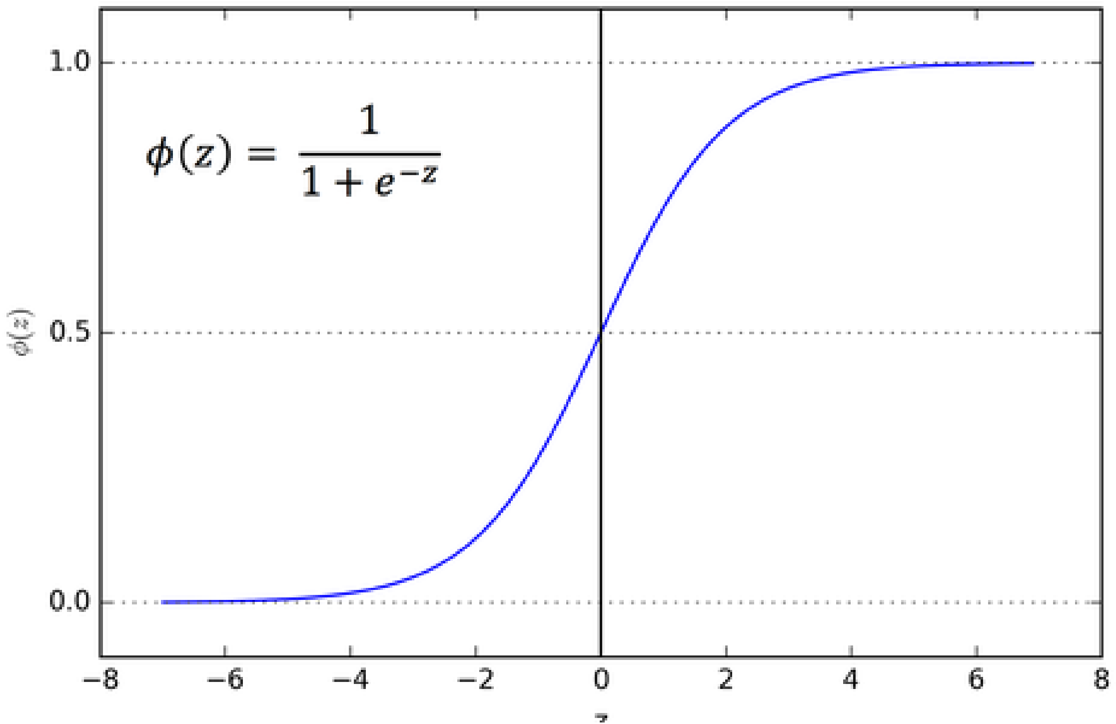
- Data are not independent
  - Behavior of one node might influence behavior of other nodes
  - Correlated behavior between nodes
- Collective inference: inferences about nodes can affect each other



Non-relational vs relational

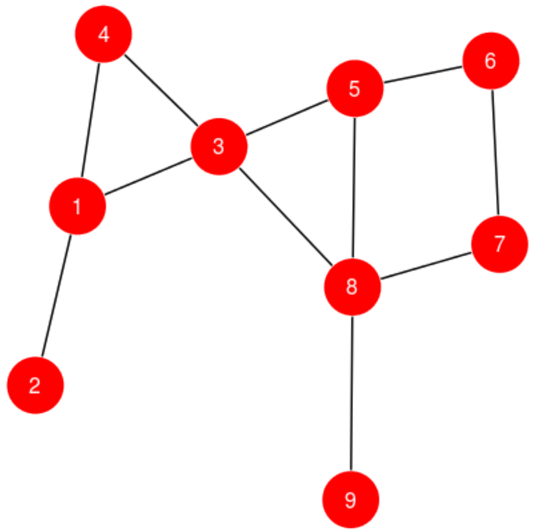
Non-relational model

- Only uses local information
- Traditional methods: logistic regression, decision trees



Relational model

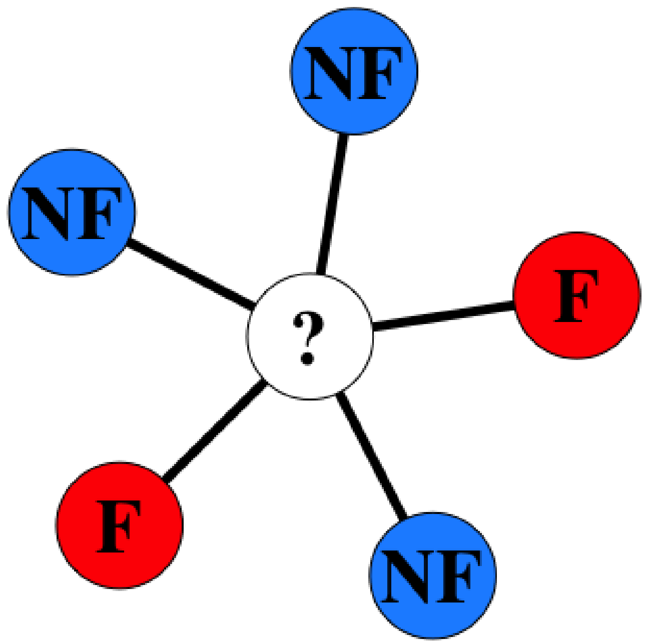
- Makes use of links in the network
- Relational neighbor classifier



Relational neighbor classifier

Assumptions

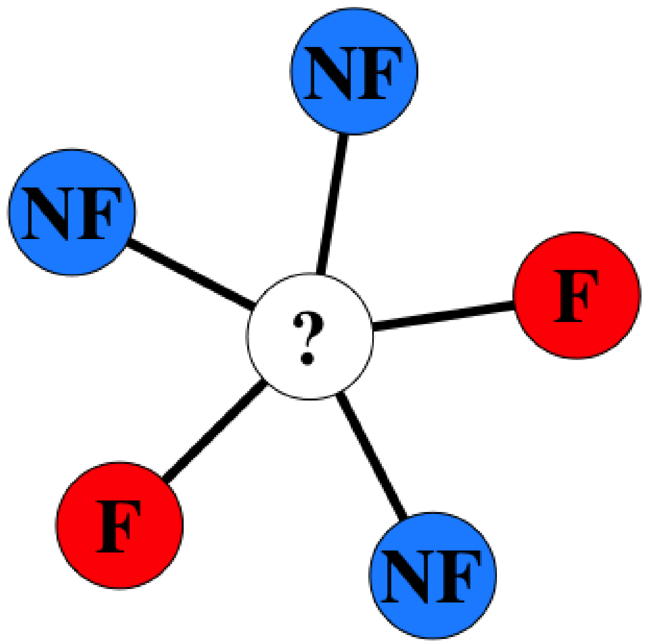
- Homophily: connected nodes have a propensity to belong to the same class ("guilt by association")
- Some class labels are known

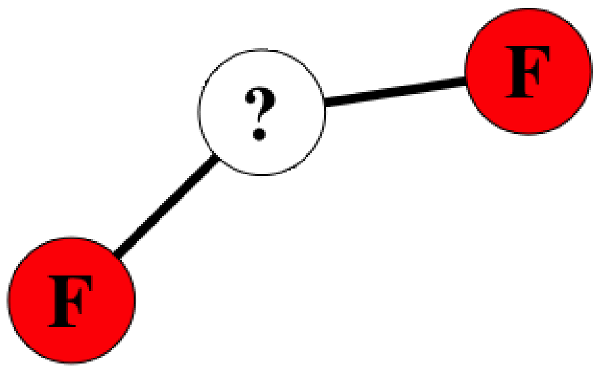


Relational neighbor classifier

Probability of fraud

$$P(F|?) = \frac{1+1}{1+1+1+1+1} = \frac{2}{5} = 40\%$$

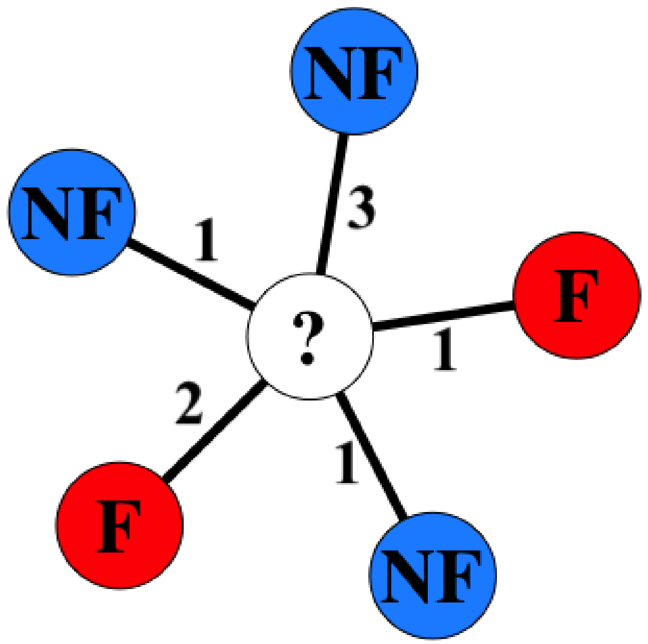


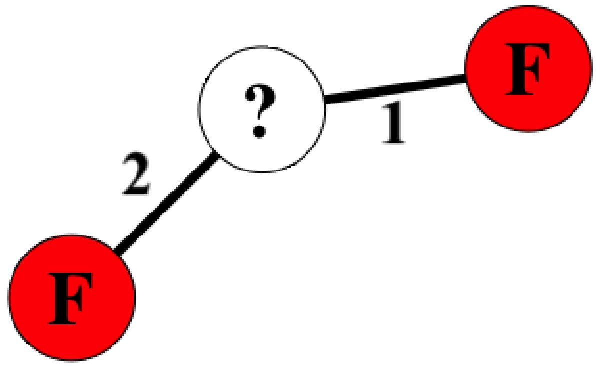


Relational neighbor classifier with weights

Probability of fraud

$$P(F|?) = \frac{1+2}{3+1+1+2+1} = \frac{3}{8} = 37.5\%$$





Relational neighbor classifier

```
# Nodes are labeled as 1 (fraud), 0 (not fraud), or NA (unknown)
> vertex_attr(network)

$name
[1] "?" "B" "C" "D" "E" "A"

$isFraud
[1] NA  1  0  1  0  0

# The edges have a weight
> edge_attr(network)

$weight
[1] 2 3 1 1 1

# Create subgraph containing node "?" and all fraudulent nodes
> subnetwork <- subgraph(network, v = c("?", "B", "D"))

# strength(): sum up the edge weights of the adjacent edges for node "?"
> prob_fraud <- strength(subnetwork, v = "?") / strength(network, v = "?")

> prob_fraud

[1] 0.375
```

FRAUD DETECTION IN R

Let's practice!

FRAUD DETECTION IN R

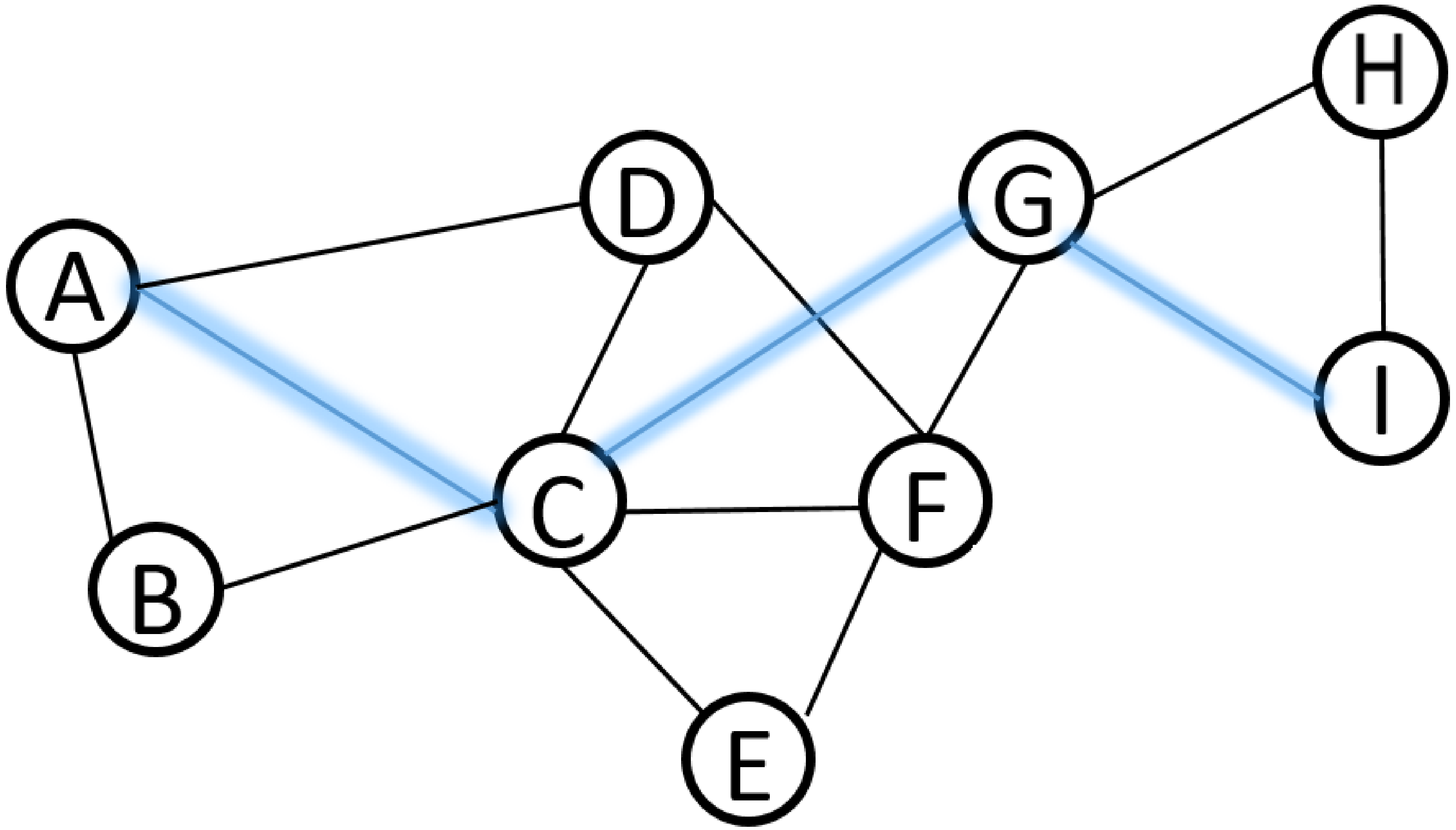
Social network metrics

Tim Verdonck
Professor Data Science at KU Leuven

Geodesic

Shortest path between nodes, e.g. between A and I

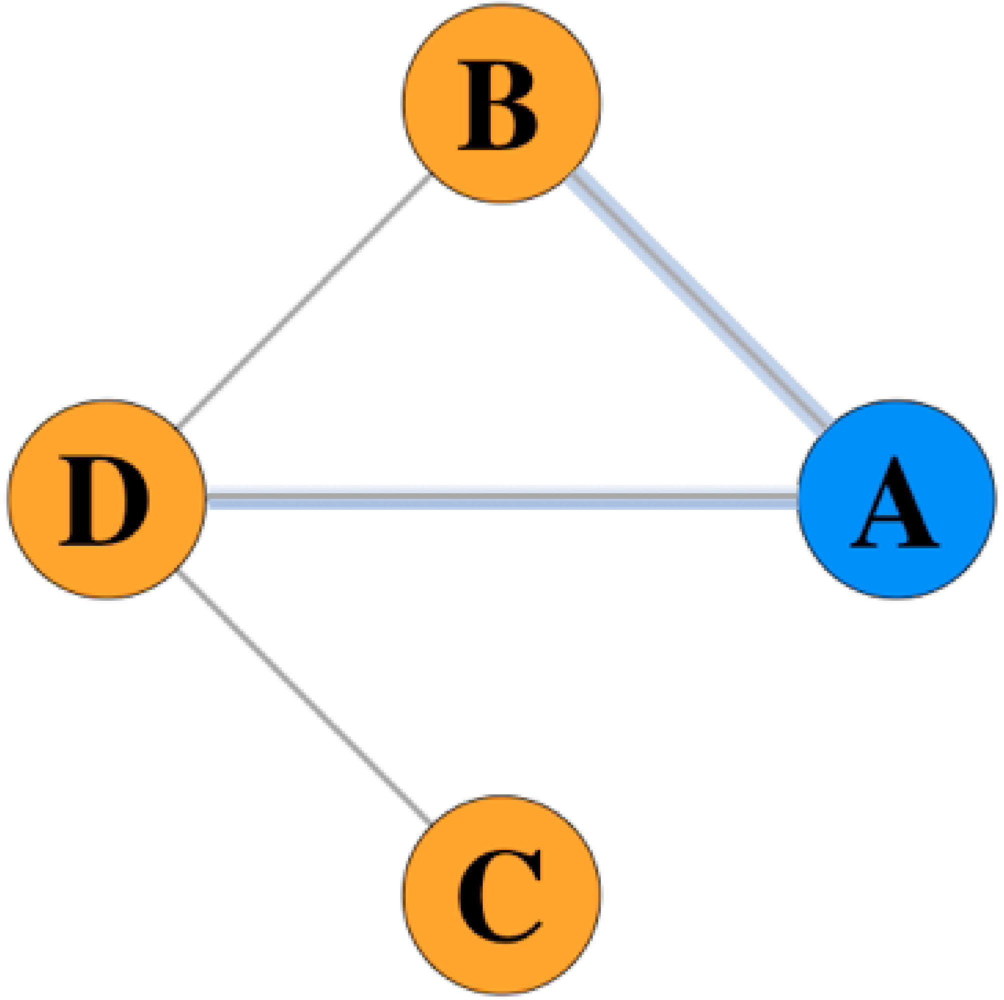
```
> shortest_paths(network, from = "A", to = "I")
[1] A C G I
```



Degree

Number of edges

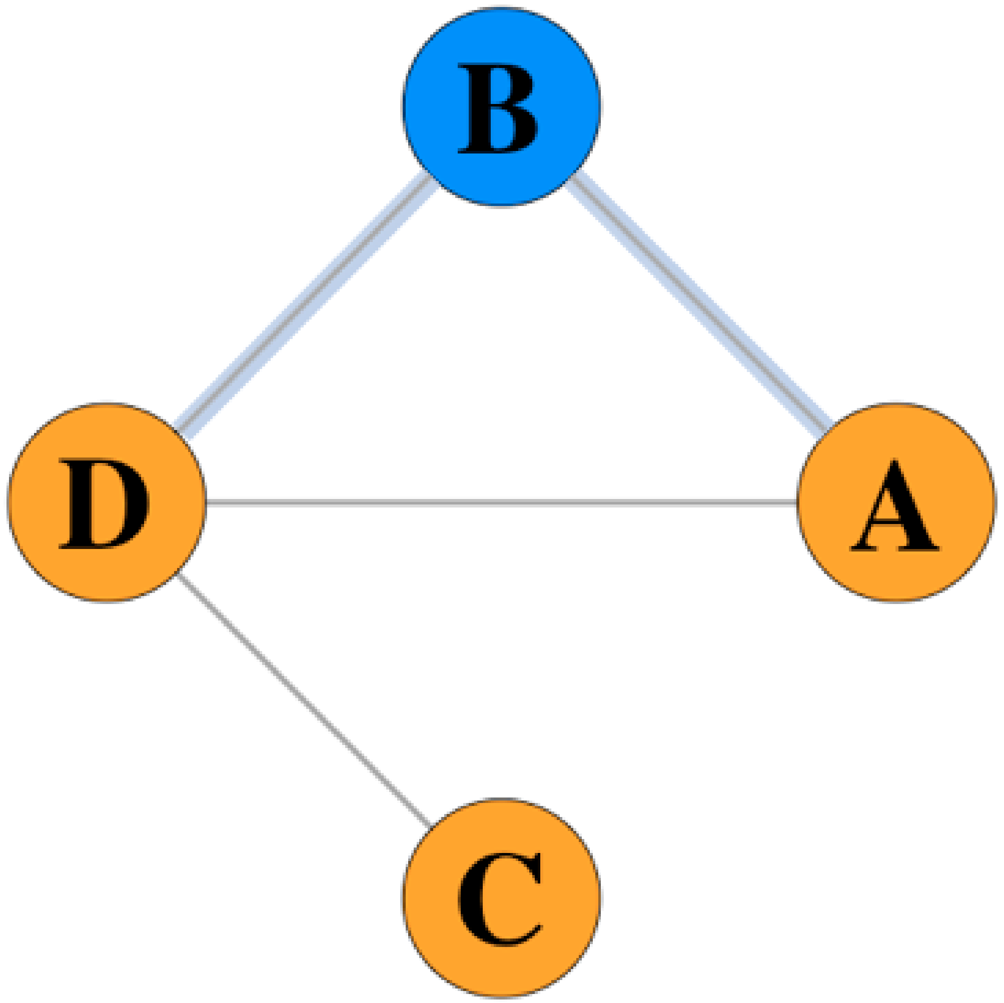
```
> degree(network)

A
2
```



Degree

Number of edges

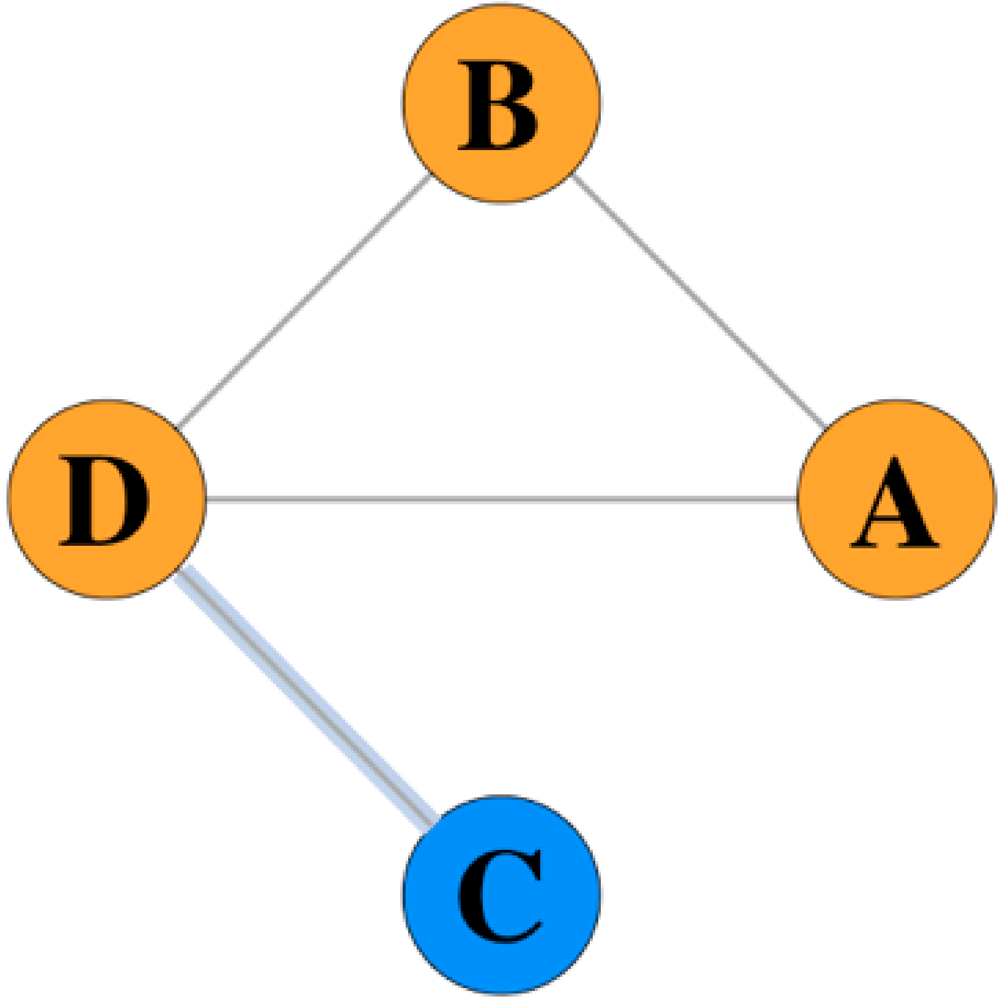
```
> degree(network)

A B
2 2
```



Degree

Number of edges

```
> degree(network)

A B C
2 2 1
```



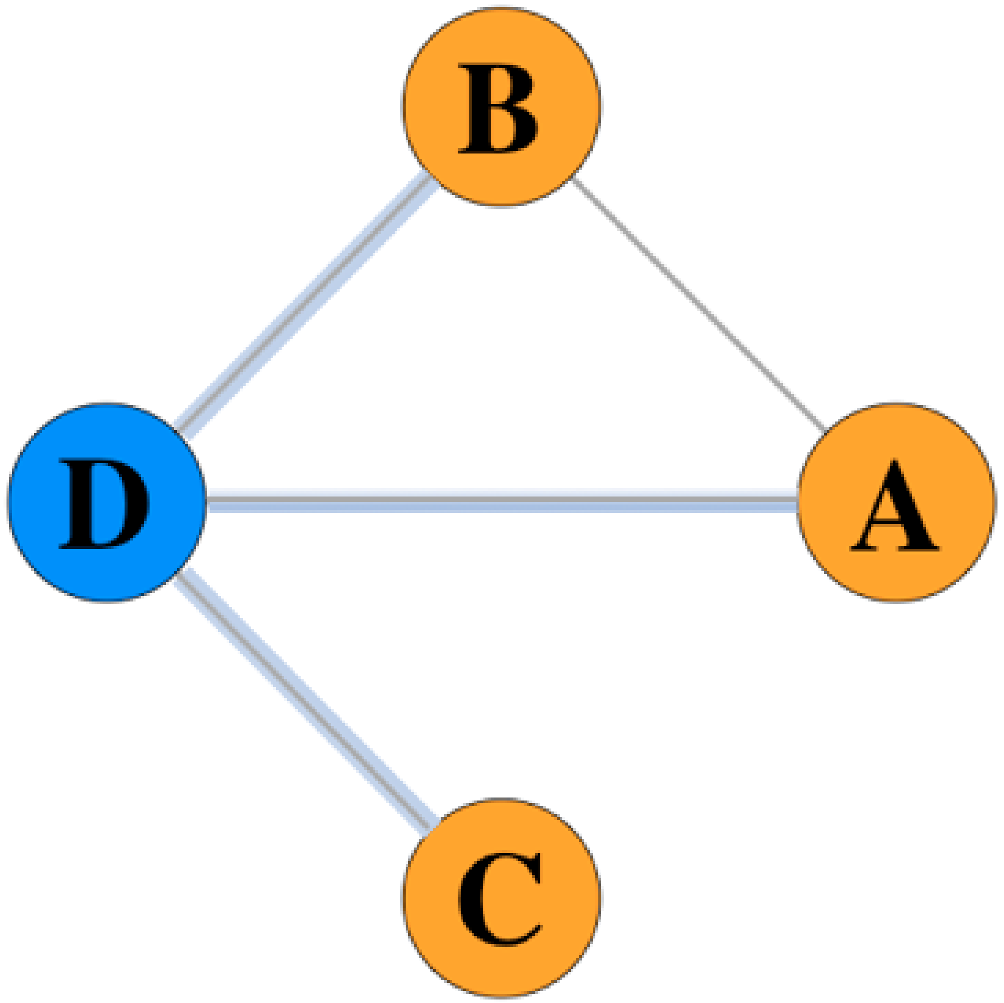
Degree

Number of edges

```
> degree(network)

A B C D
2 2 1 3
```

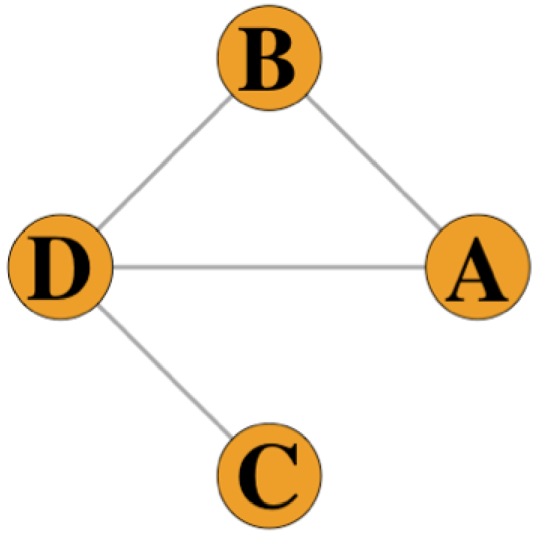
If Network has N nodes, then normalizing means dividing by $N-1$

```
> degree(network, normalized = TRUE)

      A       B       C       D
0.66667 0.66667 0.33333 1.00000
```



Closeness

Inverse distance of a node to all other nodes in the network

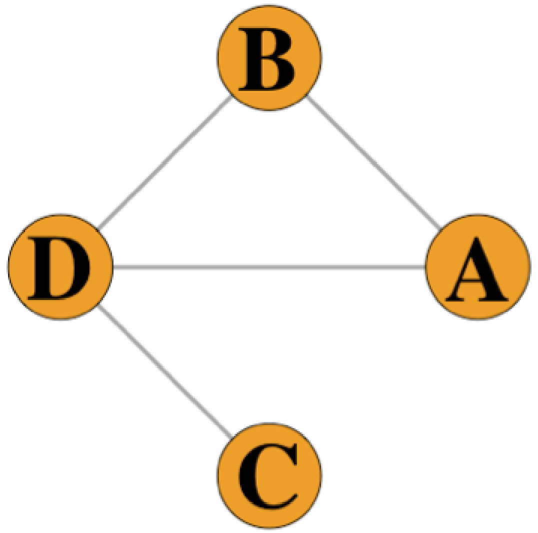


	Closeness	Normalized Closeness

Closeness

Inverse distance of a node to all other nodes in the network

```
> closeness(net)

   A
0.25
```

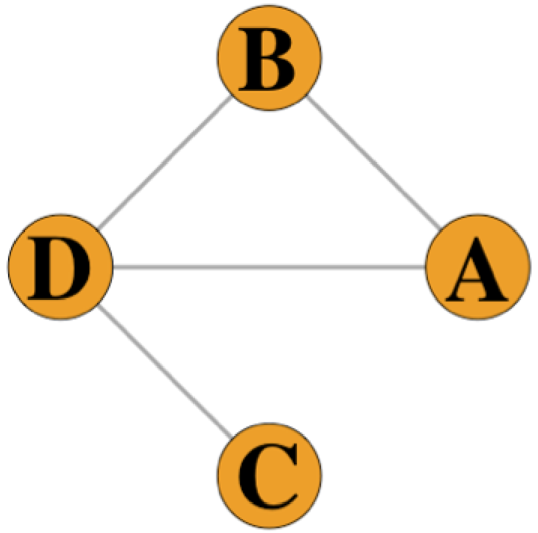


	Closeness	Normalized Closeness
A	$(1 + 1 + 2)^{-1} = 0.25$	

Closeness

Inverse distance of a node to all other nodes in the network

```
> closeness(net)

   A    B
0.25 0.25
```

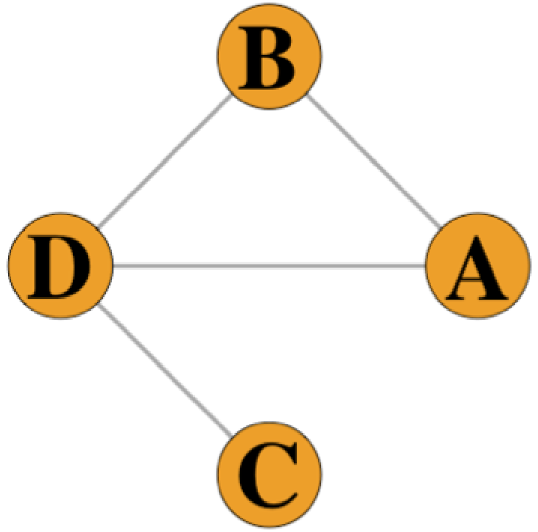


	Closeness	Normalized Closeness
A	$(1+1+2)^{-1} = 0.25$	
B	$(1+1+2)^{-1} = 0.25$	

Closeness

Inverse distance of a node to all other nodes in the network

```
> closeness(net)

   A    B    C
0.25 0.25 0.20
```

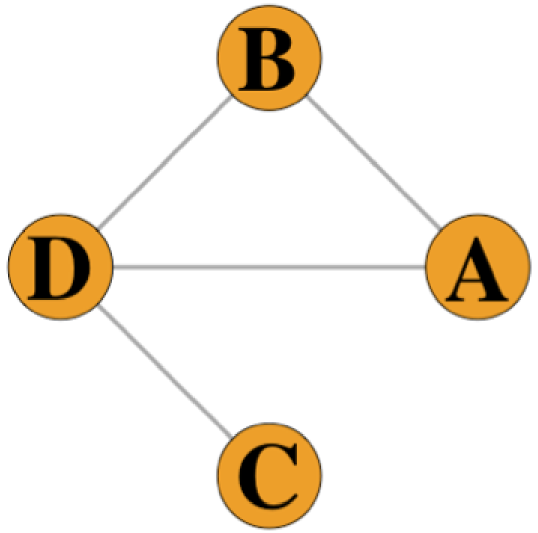


	Closeness	Normalized Closeness
A	$(1+1+2)^{-1} = 0.25$	
B	$(1+1+2)^{-1} = 0.25$	
C	$(1+2+2)^{-1} = 0.20$	

Closeness

Inverse distance of a node to all other nodes in the network

```
> closeness(net)

   A    B    C    D
0.25 0.25 0.20 0.33
```

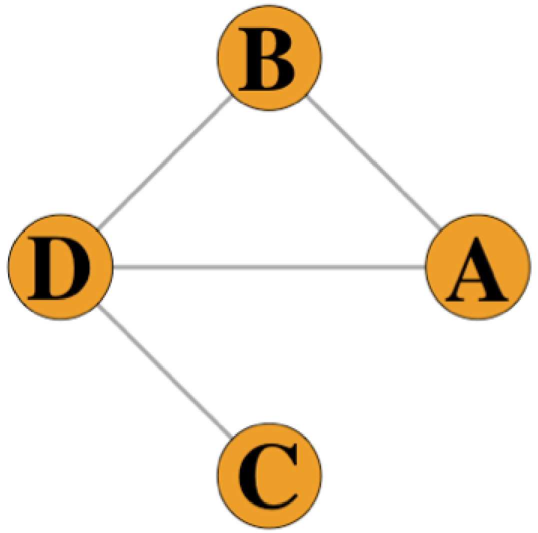


	Closeness	Normalized Closeness
A	$(1+1+2)^{-1} = 0.25$	
B	$(1+1+2)^{-1} = 0.25$	
C	$(1+2+2)^{-1} = 0.20$	
D	$(1+1+1)^{-1} = 0.33$	

Closeness

Inverse distance of a node to all other nodes in the network

```
> closeness(net)

   A    B    C    D
0.25 0.25 0.20 0.33

> closeness(net, normalized = TRUE)

   A    B    C    D
0.75 0.75 0.60 1.00
```



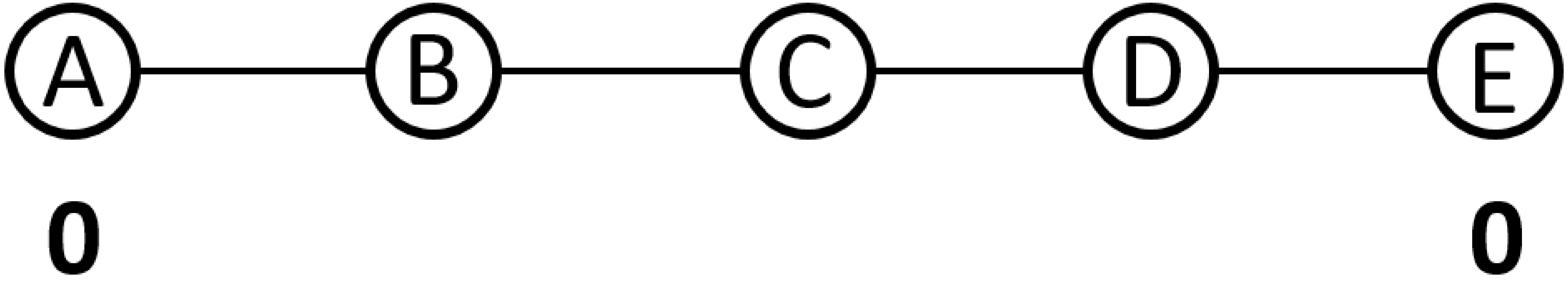
	Closeness	Normalized Closeness
A	$(1 + 1 + 2)^{-1} = 0.25$	$((1 + 1 + 2) / 3)^{-1} = 0.75$
B	$(1 + 1 + 2)^{-1} = 0.25$	$((1 + 1 + 2) / 3)^{-1} = 0.75$
C	$(1 + 2 + 2)^{-1} = 0.20$	$((1 + 2 + 2) / 3)^{-1} = 0.60$
D	$(1 + 1 + 1)^{-1} = 0.33$	$((1 + 1 + 1) / 3)^{-1} = 1.00$

Betweenness

Number of times that a node or edge occurs in the geodesics of the network

Betweenness

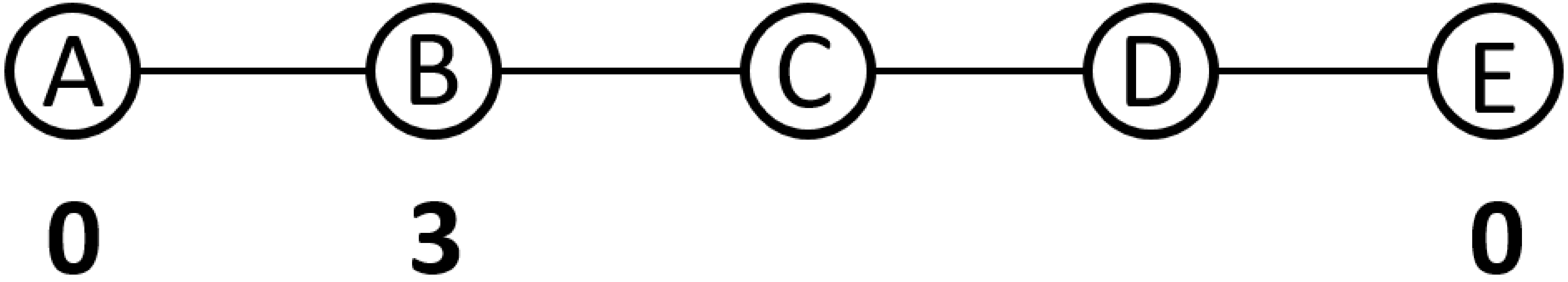
Number of times that a node or edge occurs in the geodesics of the network




```
> betweenness(network)

A        E
0        0
```

Betweenness

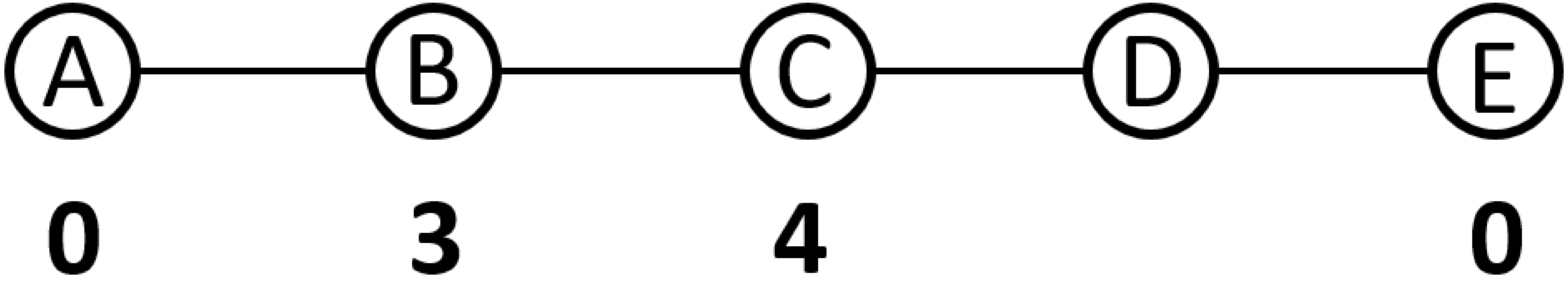
Number of times that a node or edge occurs in the geodesics of the network




```
> betweenness(network)

A  B       E
0  3       0
```

Betweenness

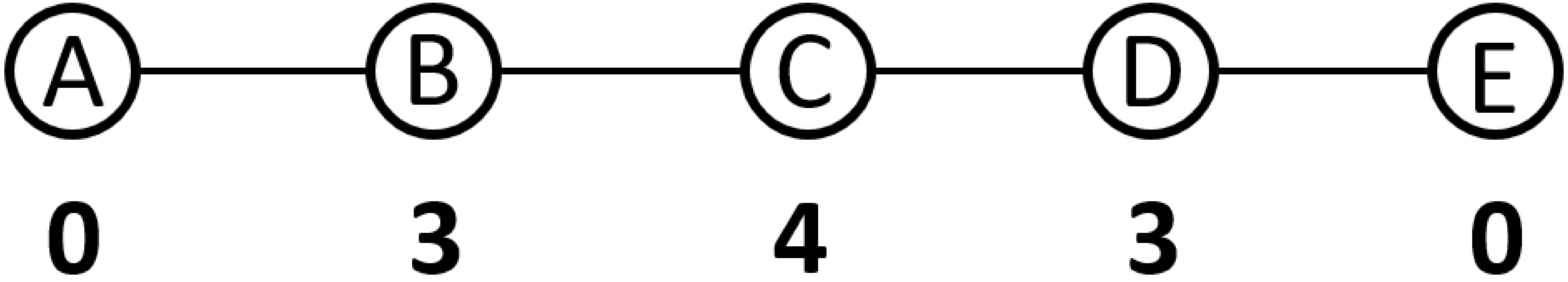
Number of times that a node or edge occurs in the geodesics of the network




```
> betweenness(network)

A  B  C     E
0  3  4     0
```

Betweenness

Number of times that a node or edge occurs in the geodesics of the network




```
> betweenness(network)

A B C D E
0 3 4 3 0

> betweenness(network, normalized = TRUE)

  A   B   C   D   E
0.0 0.6 0.8 0.6 0.0
```

Featurization

	Payment channel	...	Amount	Freq_auth	...	Rec_auth	Fraud degree	Legit degree	Closeness	...	Betweenness	Fraud
	$\underbrace{\hspace{6em}}$ Traditional features			$\underbrace{\hspace{6em}}$ Features based on recency, frequency, timestamps,...			$\underbrace{\hspace{6em}}$ Features based on social network analysis					
1	Mobile		102	3		0.02	1	4	2.73		13	No
2	ATM		125	1		0.59	0	5	2.32		29	No
3	Web		1067	0		0.86	3	2	3.05		63	No
...												
n	Mobile		1039	2		0.12	0	3	1.89		31	No

FRAUD DETECTION IN R

Let's practice!