Social Network Analysis

Univ. Prof. Dr. Stefanie Rinderle-Ma
Workflow Systems and Technology Group
University of Vienna
stefanie.rinderle-ma@univie.ac.at

Contents

1 Motivation

- Enormous amounts of „social data“ available through, e.g., social networks
- Even coining of a new term „ social data revolution“ → see, for example, Wikipedia
- Possibility for asking new questions:
  - Who is interacting with whom?
  - Whom am I interacting with?
- Where „interacting“ can be any kind of „social relation“, e.g., owe money, hands over work, etc.
- Recall the three BI perspectives
  - Customer
  - Organization
  - Production
- → Social network analysis focuses on organizational perspective

1 Motivation

Questions:

- Which data is suitable?
- How has the data to be prepared?
- What analysis model is typically used?
- Which analysis techniques are there?

Reading and basis for these slides:

- [Scott] John Scott: Social Network Analysis. SAGE (2012)
- [GrRi] Wilfried Grossmann, Stefanie Rinderle-Ma: Fundamentals of Business Intelligence, Springer 2015

Contents

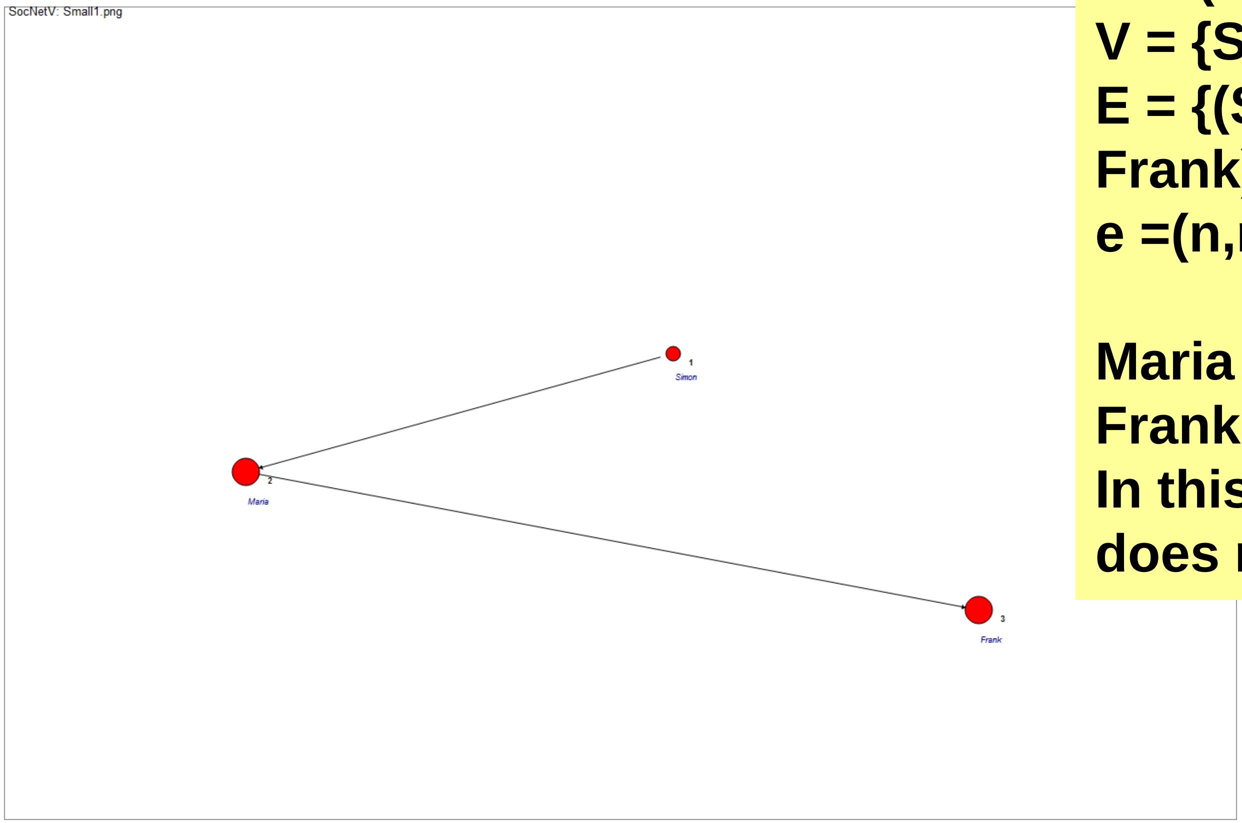
2 Data perspective

- Checking data sources → what is there?
- Checking analysis model → where do we want to go?
- Checking analysis questions → what do we want to know?

- Small lookahead: the analysis model is a sociogram, i.e., a graph $G = (V, E)$ (can be directed or undirected)

- Nodes represent the entities in the social network, e.g., persons
- Edges represent the relation between these entities, e.g., isFriendOf

2 Data perspective



G = (V, E)
V = {Simon, Maria, Frank}
E = {(Simon, Maria), (Maria, Frank)}
e =(n,m) $\in$ E: n is friend of m

Maria has friend
Frank has a friend
In this strict sense: Simon does not have a friend

2 Data perspective

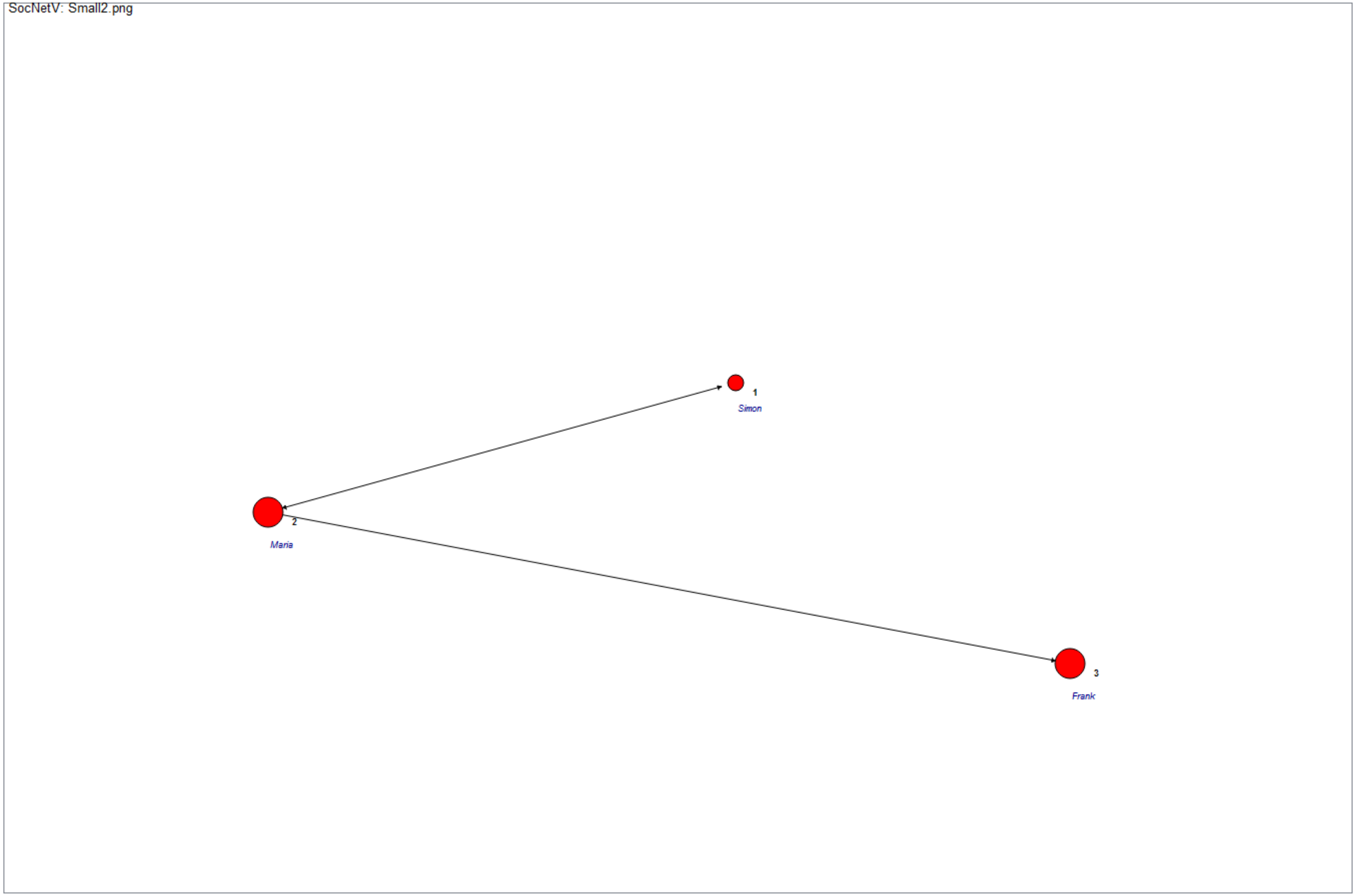
The data for example on previous slide (in .net format)

```
*Network
*Vertices 3
1 "Simon"
2 "Maria"
3 "Frank"
*Arcs
1 2 1
2 3 1
*Edges
```

Difference?

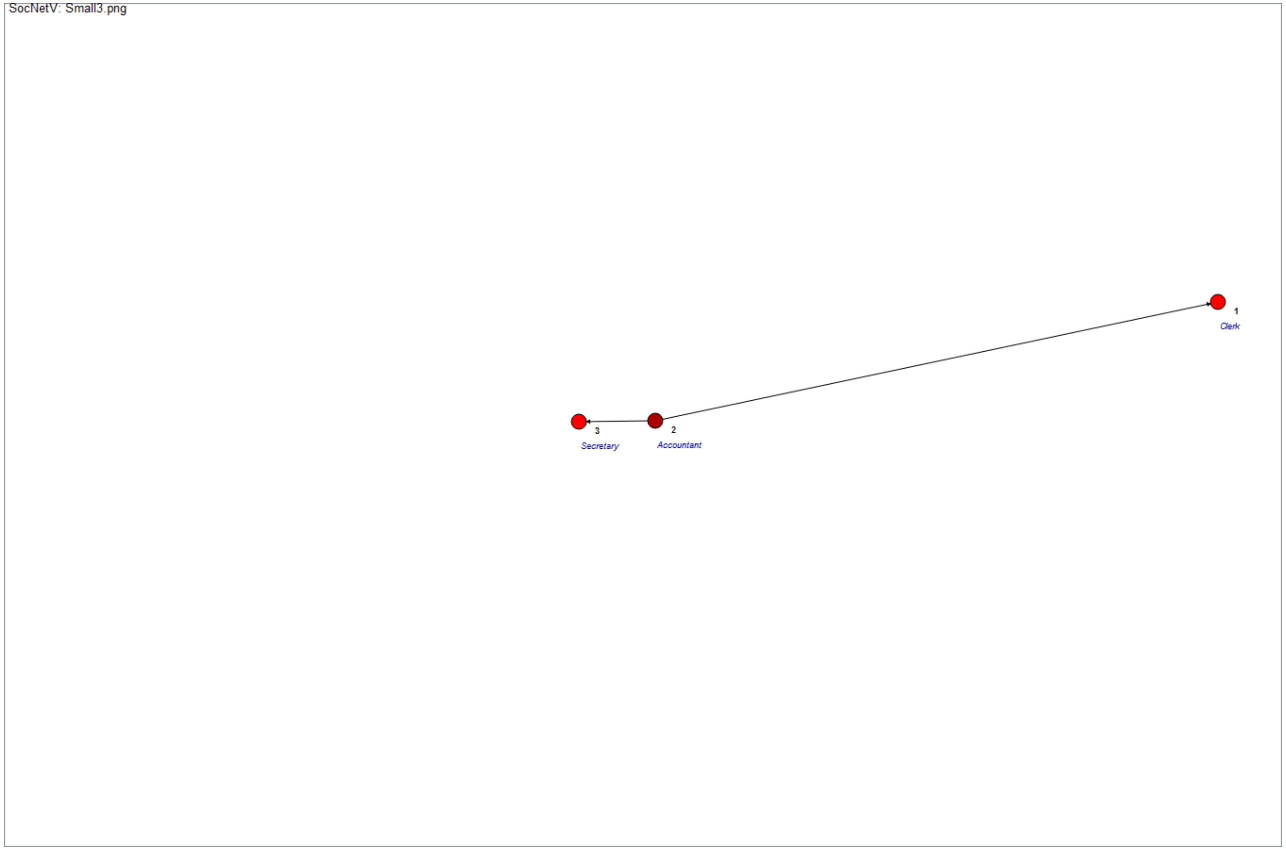
```
*Network
*Vertices 3
1 "Simon"
2 "Maria"
3 "Frank"
*Arcs
2 3 1
*Edges
1 2 1
```

2 Data perspective



2 Data perspective

Derive the data set in .net format for the following sociogram:



2 Data perspective

Other formats:

- Adjacency matrix
- GraphML: xml-based, contains visualization information

```
<graphml> …
<graph id="unnamed" edgedefault="directed">
    <node id="1">
      <data key="d0">Simon</data>
      <data key="d1">0.544782</data>
      <data key="d2">0.429213</data>
      <data key="d5">circle</data>
    </node>
    …
    <edge id="e1" directed="true" source="1" target="2"/>
    <edge id="e2" directed="true" source="2" target="3"/>
  </graph>
</graphml>
```

2 Data perspective

Analysis questions:

- Who or what are identified as entities?
- What are the interesting relations to be analyzed?

Basically:

- Analysis of the entire network
- Analysis for selected nodes (entities)

Job for data preperation:

- Make decisions on the questions above
- Prepare data accordingly
- If data is big, think about sampling

2 Data perspective

		Affiliations		
		A	B	C
Cases	1	1	0	0
	2	1	0	0
	3	1	0	0

What are the entities (nodes) and relations (edges) for this example (taken from [Scott])?

2 Data perspective

According to [Scott] three different representation matrices for SNA exist:

Incidence matrix

		Cases		
		1	2	3
Affiliations	A			
	B			
	C			

Adjacency matrix (→ best for SNA)

		Cases		
		1	2	3
Cases	1			
	2			
	3			

Adjacency matrix

		Affiliations		
		A	B	C
Affiliations	A			
	B			
	C			

2 Data perspective

According to [Scott] three different representation matrices for SNA exist:

Incidence matrix

		Students		
		1	**2**	**3**
Universities	**A**	1	1	0
	B	0	1	0
	C	1	1	1

Adjacency matrix

		Students		
		1	**2**	**3**
Students	**1**	-	2	1
	2	2	-	1
	3	1	1	-

Adjacency matrix

		Universities		
		A	**B**	**C**
Universities	**A**	-	1	2
	B	1	-	1
	C	1	2	-

Contents

3 Model perspective

- As mentioned before, the basic model is the sociogram
- Model structures for SNA (based on [GrRi])
  - *Undirected graphs*: an undirected graph G is defined as G = (V;E) with set of nodes V and set of undirected edges E.
  - *Directed graphs*: Opposed to undirected edges, directed edges establish a relation that reflects a causal relation or a relation that is directed from one to another entity.
  - *Weighted Graphs*: It can be also useful to assign weights to the edges in the graph, i.e., a weight w(e) expressing some kind of quantitative measure for the relation.
  - *Connected Subgraphs*: Special connected subgraphs might be of interest. A subgraph consisting of two nodes (with or without relations between them) describes a *dyad*, a sub-graph consisting of three nodes of interest a *triad* respectively.
  - *Dyad / triad*: Two / three actors who are connected by a relation in the social network

3 Model perspective

How to build the model from the data?

1. Step: create data matrix (as described in Section 2)
2. Step: create models for different analysis tasks

3 Model perspective

Example 1: Building model from relational data

Students	**SID**	**Name**
	S1	Simon
	S2	Maria
	S3	Frank
	S4	Sally
	S5	Bert

enrolled	**SID**	**UID**
	S1	U1
	S2	U1
	S1	U2
	S3	U3
	S3	U2
	S2	U2

University	**UID**	**Name**
	U1	Univie
	U2	TUWien
	U3	WUWien

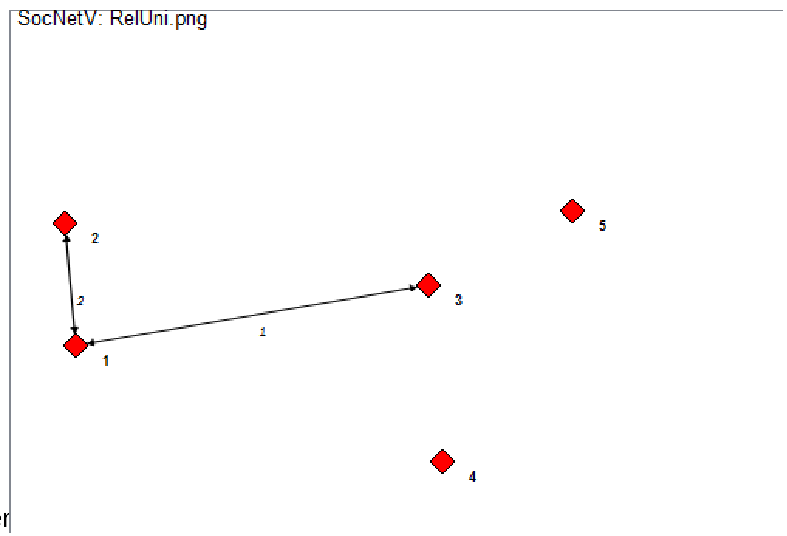
		Cases				
		S1	S2	S3	S4	S5
Cases	S1	-	2	1	-	-
	S2	2	-	1	-	-
	S3	1		-	-	-
	S4	-	1	-	-	-
	S5	-	-	-	-	-

3 Model perspective

Example 1: Building model from relational data

		Cases				
		S1	S2	S3	S4	S5
Cases	S1	-	2	1	0	0
	S2	2	-	1	0	0
	S3	1		-	0	0
	S4	0	1	0	-	0
	S5	0	0	0	0	-

```
*Network
*Vertices 5
1 "Simon"
2 "Maria"
3 "Frank"
4 "Sally"
5 "Bert"
*Edges
1 2 2
1 3 1
```



Example 2: Building model from log data (based on [GrRi])

<AuditTrailEntry>

<WorkflowModelElement>Evaluate presentation 1</WorkflowModelElement>…

<Originator>person001-lecturer</Originator>

</AuditTrailEntry>

<AuditTrailEntry>

<WorkflowModelElement>Evaluate presentation 1</WorkflowModelElement>…

<Originator>person003-lecturer</Originator>

</AuditTrailEntry>

<AuditTrailEntry>

<WorkflowModelElement>plus</WorkflowModelElement>…

<Originator>person003-lecturer</Originator>

</AuditTrailEntry>

<AuditTrailEntry>

<WorkflowModelElement>plus</WorkflowModelElement>…

<Originator>person004-lecturer</Originator>

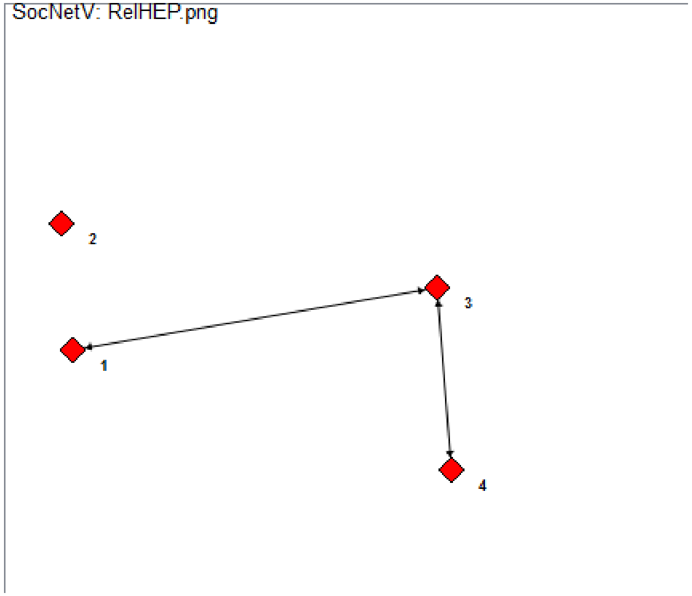
</AuditTrailEntry>.000+01:00</Timestamp>

Event Type and Time Stamp omitted

3 Model perspective

	Evaluate Presentation 1	plus
person001-lecturer	1	0
person002-lecturer	0	0
person003-lecturer	1	1
person004-lecturer	0	1

```
*Network
*Vertices 4
1 "person001-lecturer"
2 "person002-lecturer"
3 "person003-lecturer"
4 "person004-lecturer"
*Edges
1 3 1
3 4 1
```



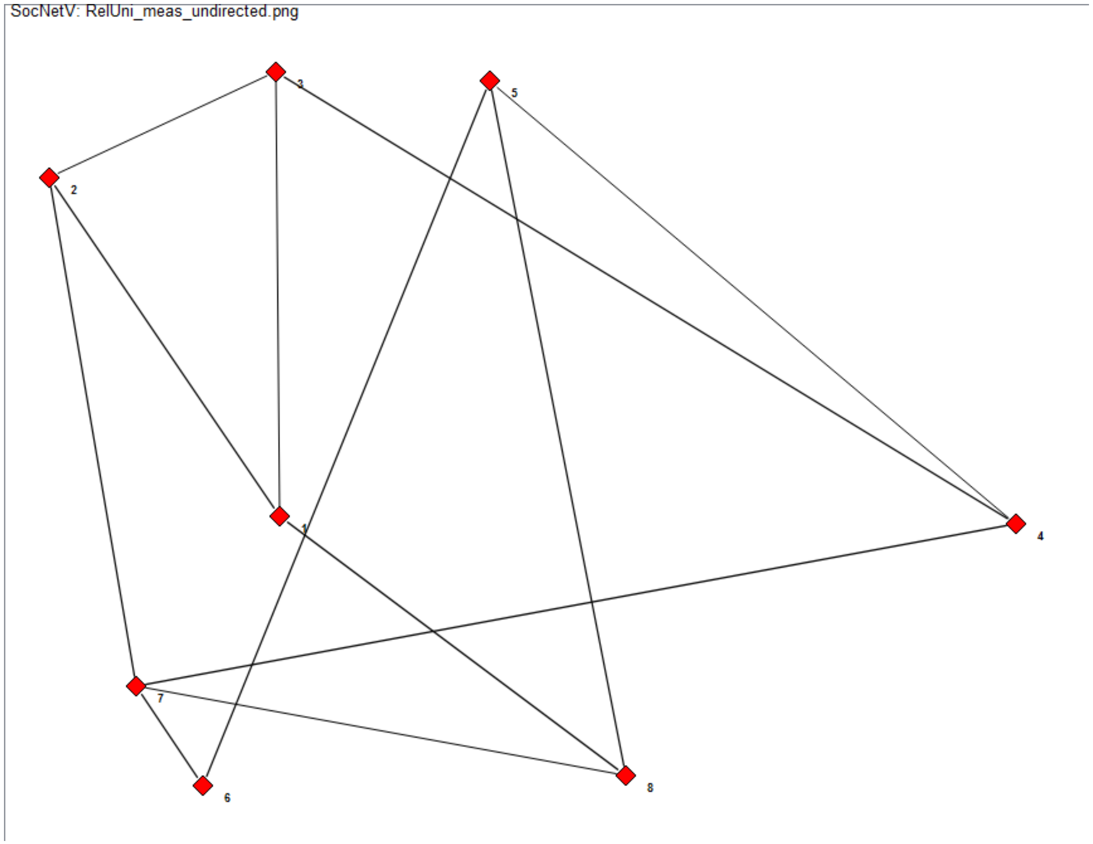
Contents

4 Analytical perspective

- Basically, different measures on the sociogram
  - For the entire network
  - For single nodes
- In addition: local and global measures

Local measures for nodes:

1. degree, in-degree, out-degree

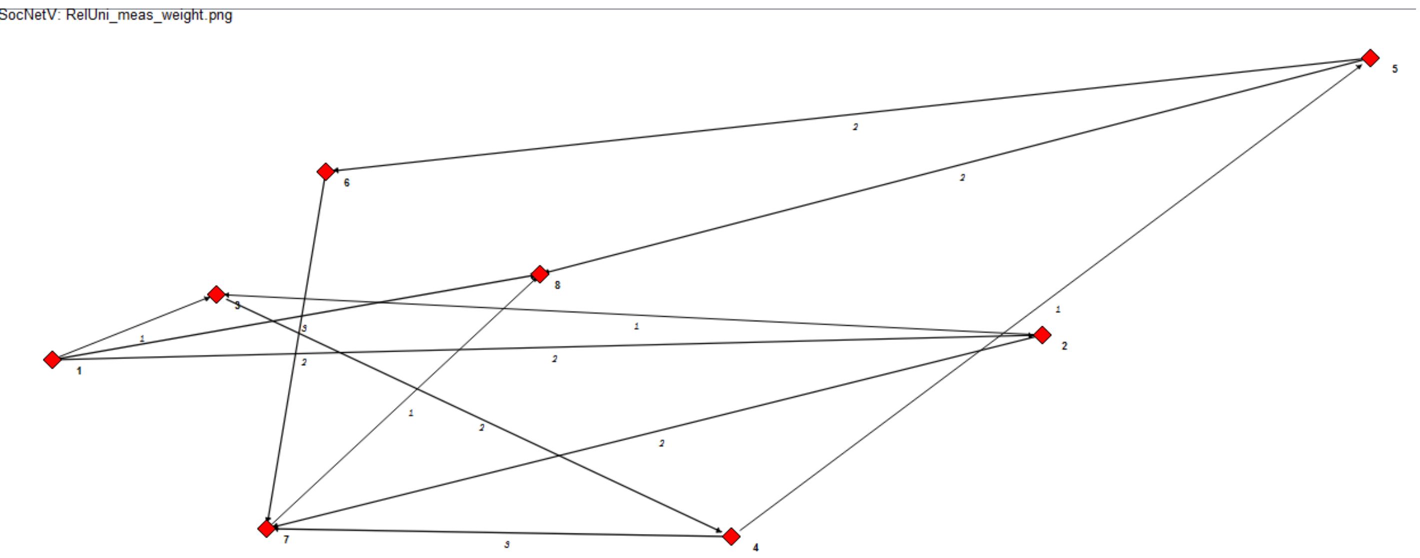


Node	Degree
1	
2	
3	
4	
5	
6	
7	
8	

4 Analytical perspective

Local measures for nodes:

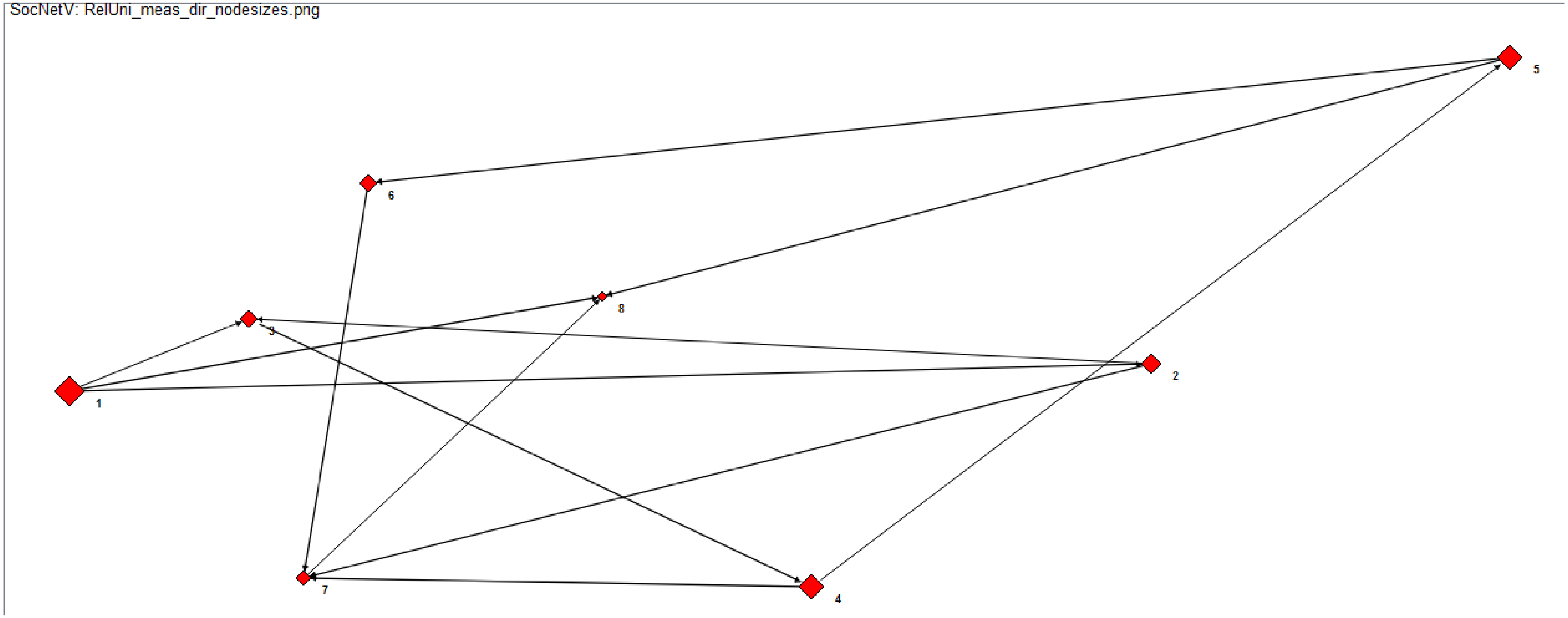
1. degree, in-degree, out-degree



Node	In-degree	Out-degree
1		
2		
3		
4		
5		
6		
7		
8		

Local measures for nodes:

1. Visualization: node sizes by out-degree



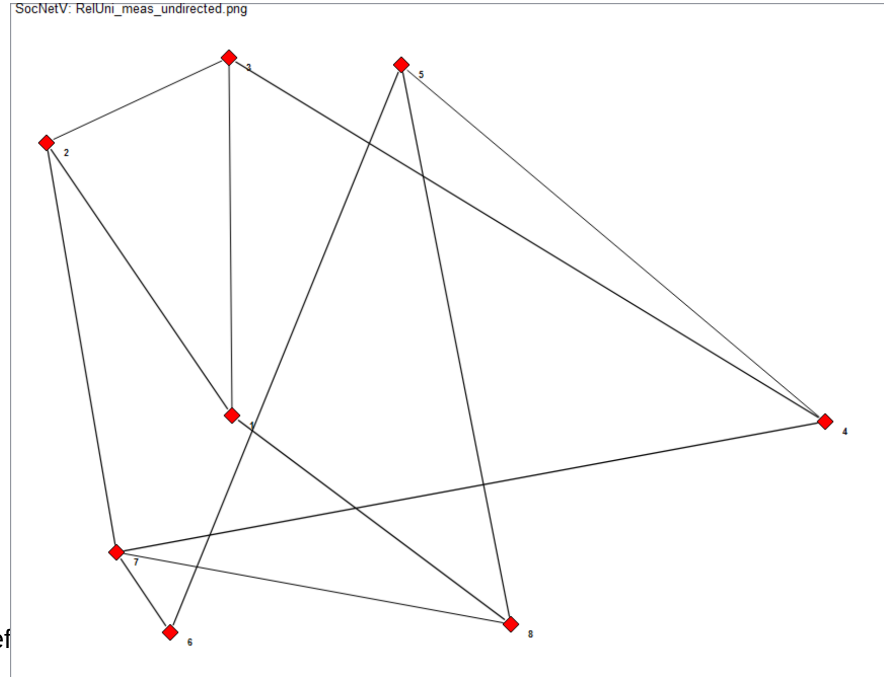
Is the degree meaningful?

→ Degree centrality of node x (*point centrality*):

DC(x) = degree(x)/(N-1)

where N is the number of nodes in the sociogram

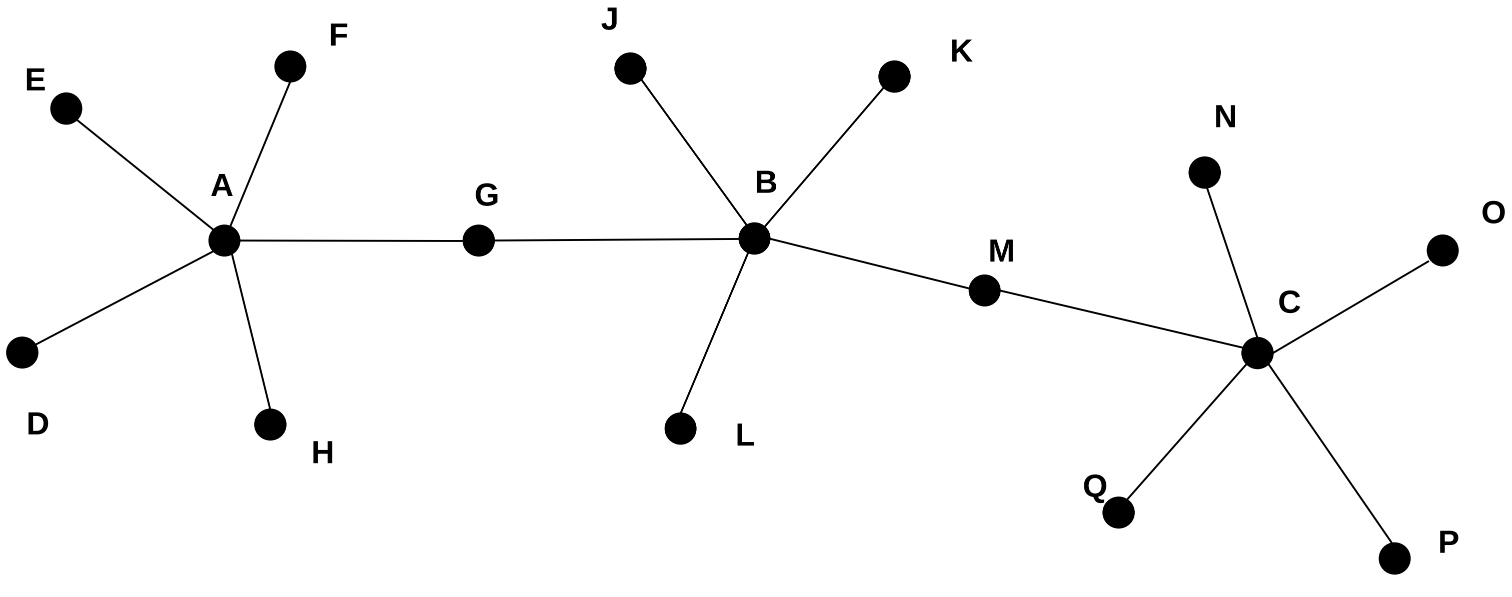
→ Undirected: degree; directed: out-degree; weighted: sum of all weights of outgoing edges



Node	DC
1	
2	
3	
4	
5	
6	
7	
8	

Interpretation degree centrality:

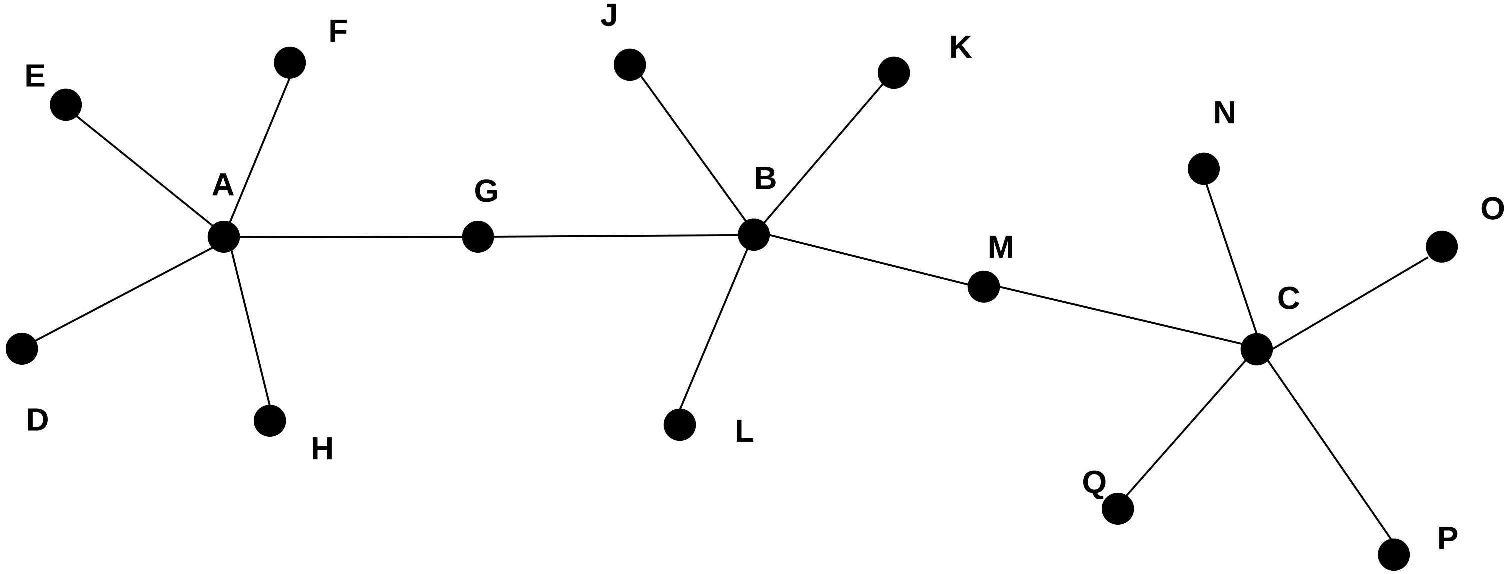
- When is this a useful measure? In which situations probably not?
- Example taken from [Scott]:
- Degree centrality is a local (node) measure



To come to a global measure, take paths instead of edges:

k-path centrality of node x = $\sum_n path\ (x, n)$

where n $\in$ N\{x} and path(x,n) denotes the shortest path from x to n



(based on [Scott])	A, C	B	G, M	J, K. L	others
Local centrality (abs)					
Local centrality (rel)					
Global centrality					

4 Analytical perspective

(based on [Scott])	A, C	B	G, M	J, K. L	others
Local centrality (abs)	5	5	2	1	1
Local centrality (rel)	0,33	0,33	0,13	0,07	0,07
Global centrality	43	33	37	48	57

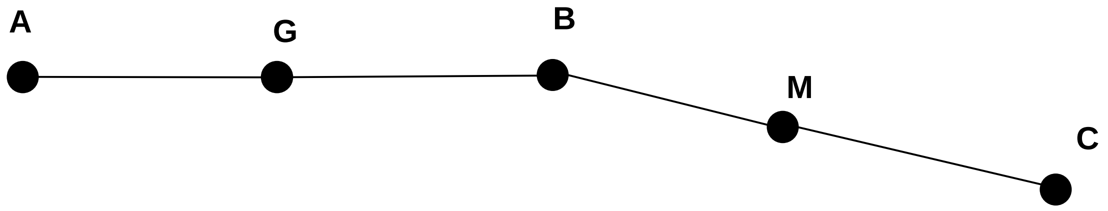
- Which nodes are locally central?
- Which nodes are globally central?
- Interpretation:

4 Analytical perspective

- Another point centrality measure: *betweenness centrality*
- Betweenness centrality BC of a node x:

$$BC(x) = \sum_{i \neq j} path(i,j,x)/path(i,j)$$

- Where path(i, j, x) denotes the shortest path from i to j through x.



- BC(B) = 3/3+4/4 + 2/2 + 3/3 = 4
- BC(G) = 2/2+3/3+4/4 = 3
- Interpretation: betweenness centrality estimates the role of an intermediary in a SNA, e.g., a broker

4 Analytical perspective

Result Social Network Visualizer:

BETWEENESS CENTRALITY (BC)
The BC index of a node u is the sum of delta (s,t,u) for all s,t in V where delta (s,t,u) is the ratio of all geodesics between s and t which run through u. Read the Manual for more.
BC' is the standardized BC.

BC range: 0 < BC < 12 (Number of pairs of nodes excluding u)
BC' range: 0 < BC'< 1 (C' is 1 when the node falls on all geodesics)

Node	BC	BC'	%BC'
1	0	0	0
2	3	0.25	25
3	4	0.333	33.3
4	3	0.25	25
5	0	0	0

Normalization with factor number of all pairs: (n-1)*(n-2)/2

Max BC' = 0.333 (node 3)
Min BC' = 0 (node 1)
BC classes = 3

BC' sum = 0.833
BC' Mean = 0.167
BC' Variance = 0.0194

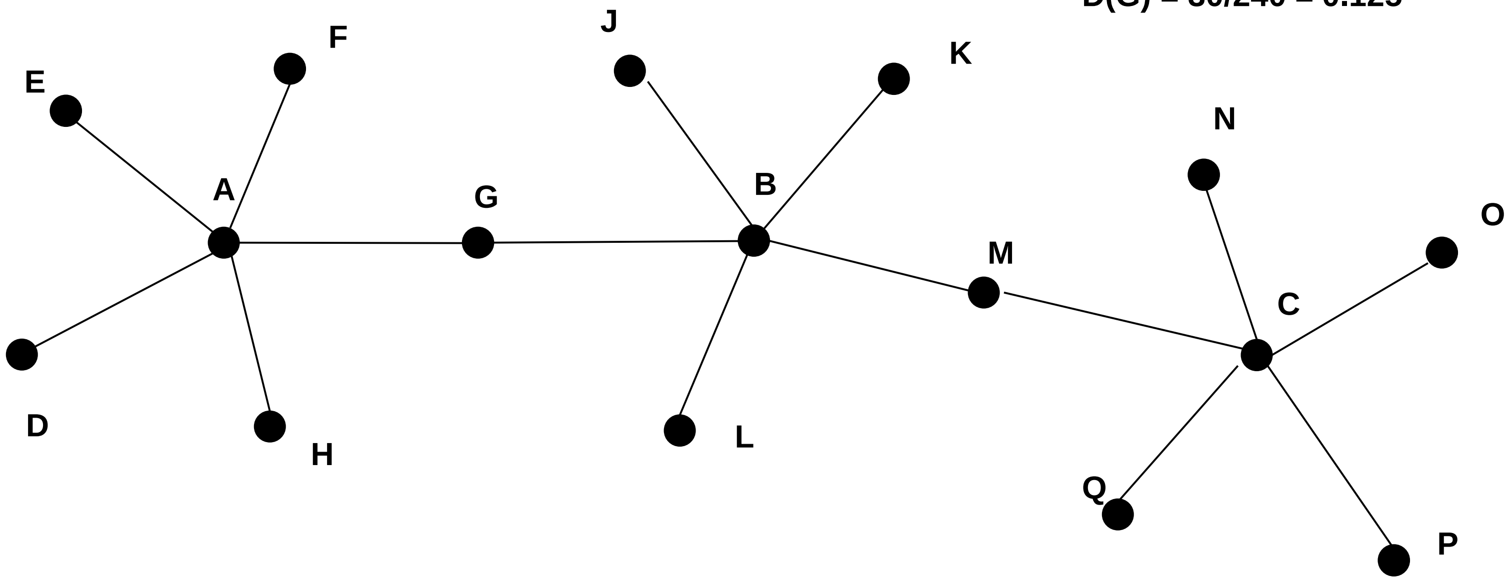
4 Analytical perspective

Graph metrics

- *density* D of a graph / sociogram G=(V,E):

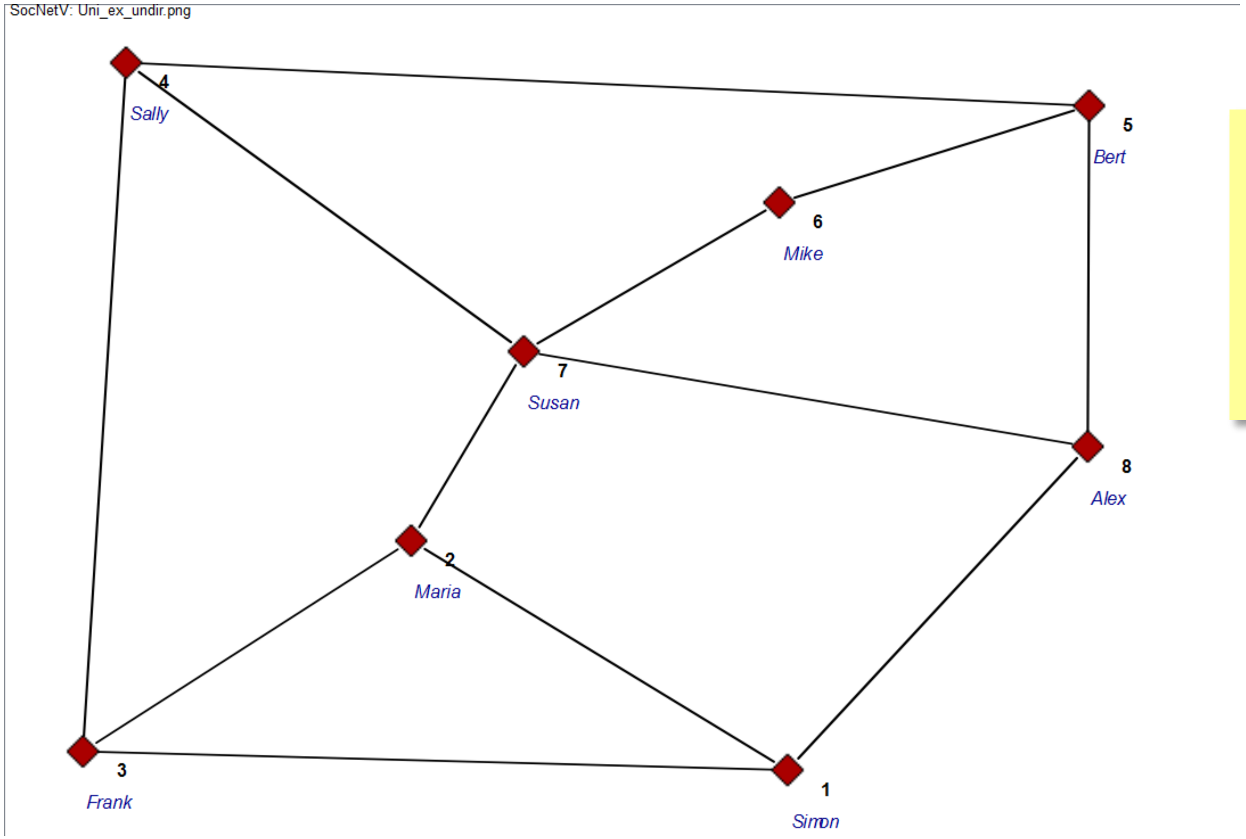
$$D(G) := \frac{2*|E|}{|V|*(|V|-1)}$$

D(G) = 30/240 = 0.125



Interpretation?

4 Analytical perspective



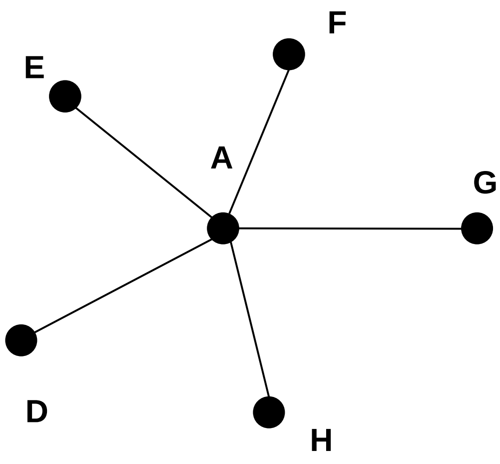
Exercise:
Analyse the SNA with the instruments we have at hand now

Graph centrality:

Measures the centrality of the nodes in the graph in relation to the most central point

Let x* be the node with the highest centrality in the SNA G. Then:

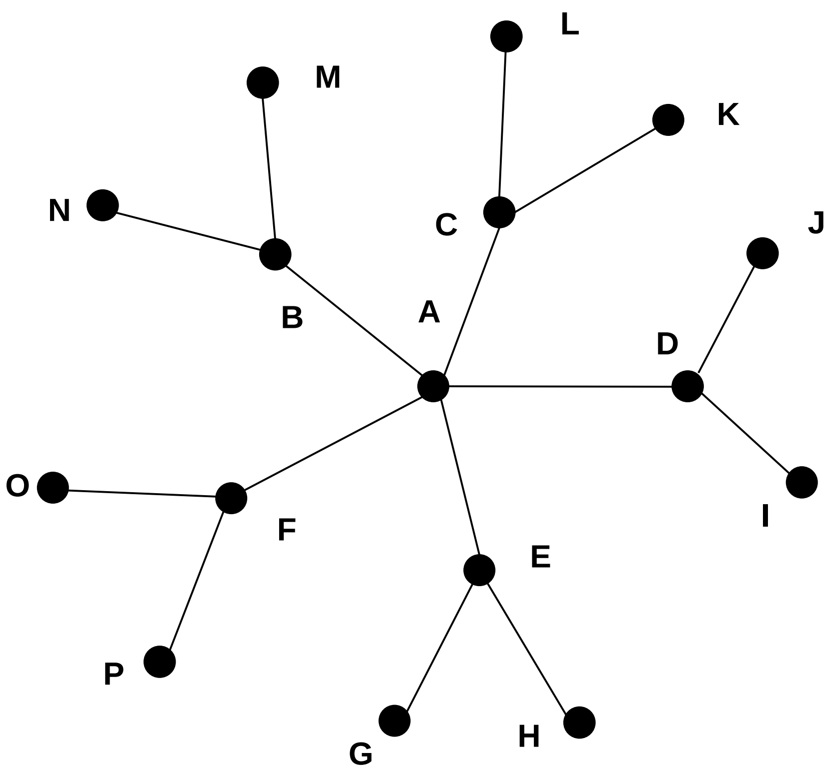
$$GC(G) = \frac{\sum_{n,n \neq x} C(x*) - C(n)}{(n-1)*(n-2)}$$



Centrality?
Assuming degree centrality
DC(A) = 5
DC(D) = DC(E) = DC(F) = DC(G) = DC(H) = 1
GC(G) = 5*4/5*4 = 1

Graph centrality:

Another example based on [Scott]



node	DC
A	5
B	3
C	3
D	3
E	3
F	3
G	1
H	1
I	1
J	1
K	1
L	1
M	1
N	1
O	1
P	1

Contents

5 Summary

- There are many more metrics to analyze SNA
  - Closeness
  - Cliques in the graph
- Tools:
  - Pajek
  - Social Network Visualizer
  - R
- Organizational mining:
  - Lies at the interface between process mining and social network mining
  - Hence at the interface between production and organization perspective