# COMP30120 Tutorial

# Social Network Analysis

Derek Greene

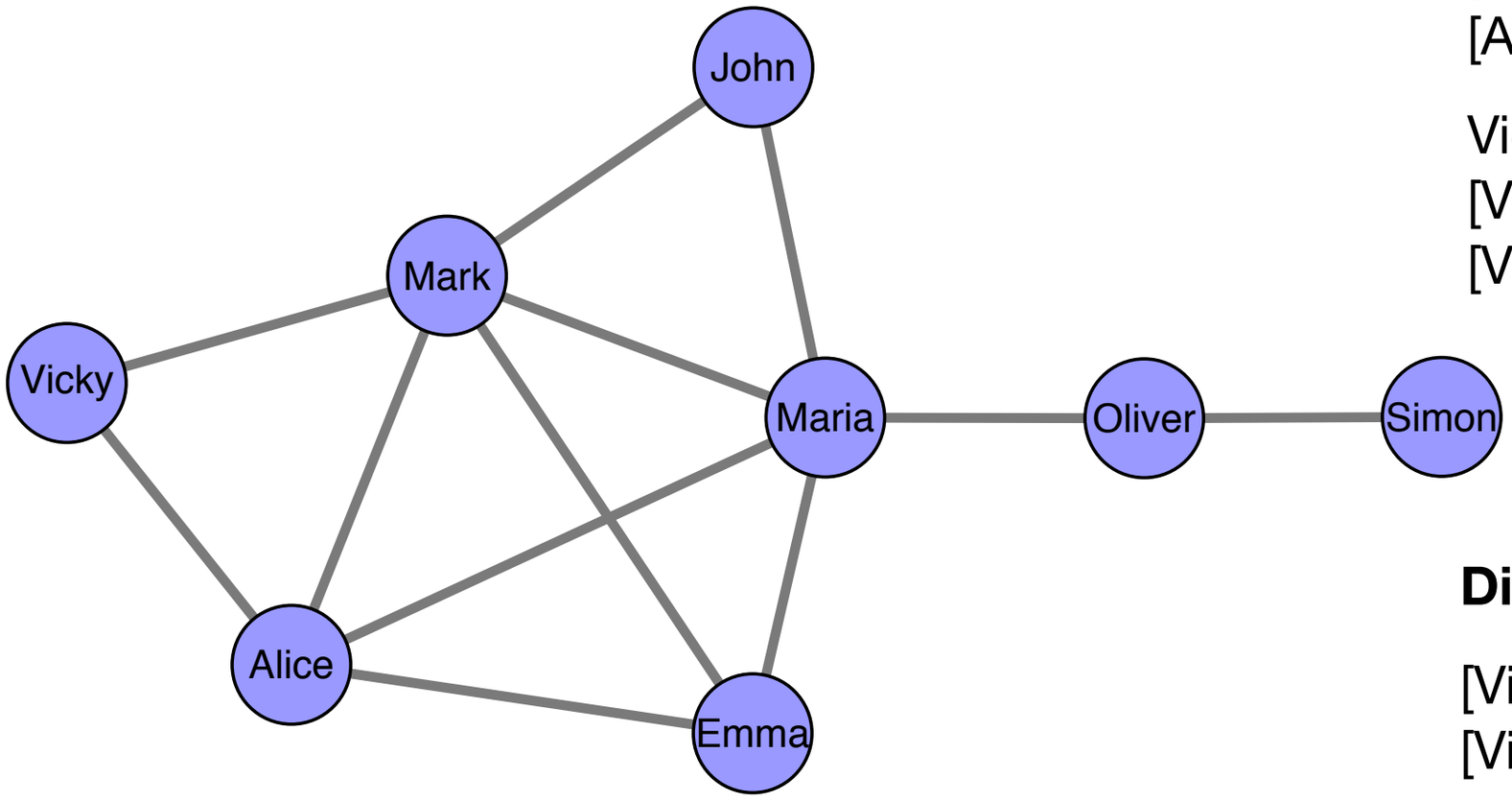
School of Computer Science and Informatics
Autumn 2015

# Tutorial Q1(a)

The diagram below shows a network representing members of a sports club, with 8 nodes and 12 undirected edges. Each node is a member and an edge between two members indicates that they are mutual friends.

Q. What is the shortest path between the pairs:
1) Alice and John; 2) Vicky and Simon?

Q. What is the diameter of this network?




**Shortest Paths**

Alice -> John
[Alice, Maria, John]
[Alice, Mark, John]

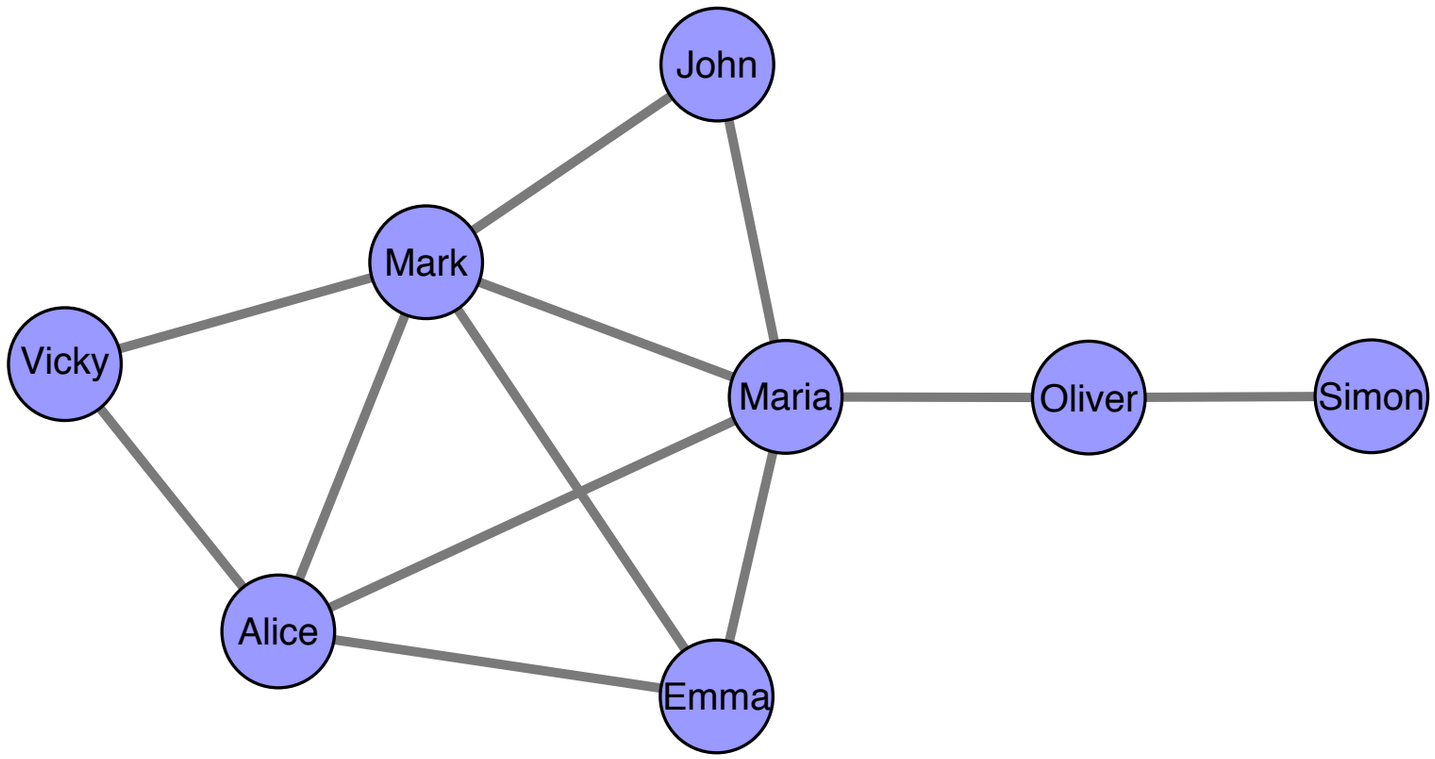
Vicky -> Simon
[Vicky, Alice, Maria, Oliver, Simon]
[Vicky, Mark, Maria, Oliver, Simon]

**Diameter = 4**

[Vicky, Alice, Maria, Oliver, Simon]
[Vicky, Mark, Maria, Oliver, Simon]

# Tutorial Q1(b)

Calculate the *degree centrality* for each node in the graph. Which members have the highest and lowest centrality?



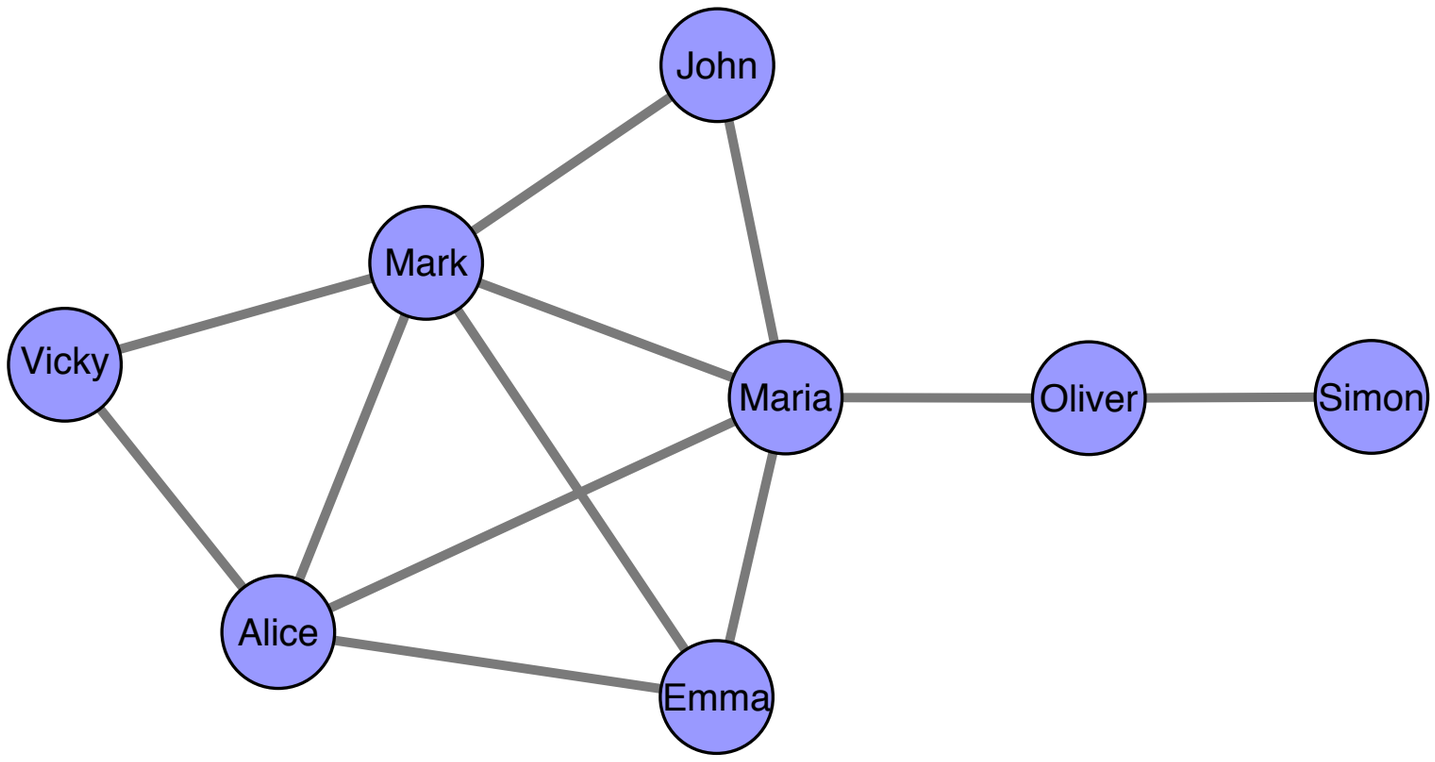

| Node | Degree |
|---|---|
| Alice | 4 |
| Emma | 3 |
| John | 2 |
| Maria | 5 |
| Mark | 5 |
| Oliver | 2 |
| Simon | 1 |
| Vicky | 2 |

**Highest Degree:** Mark & Maria

**Lowest Degree:** Simon

# Tutorial Q1(c)

If we wanted to identify the information brokers in this network, what would be an appropriate measure of centrality to use? Based on inspecting the diagram above, which node would have the highest centrality value according to this measure?



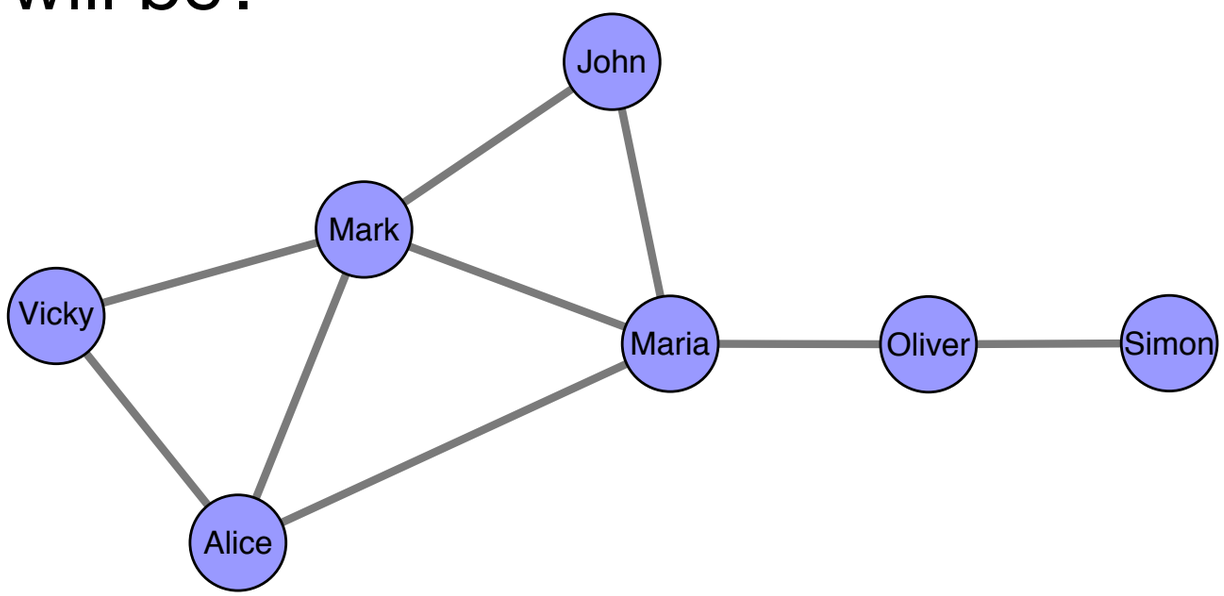

| Node | Between |
|---|---|
| Alice | 0.10 |
| Emma | 0.00 |
| John | 0.00 |
| Maria | 0.52 |
| Mark | 0.19 |
| Oliver | 0.29 |
| Simon | 0.00 |
| Vicky | 0.00 |

**Betweenness Centrality:** Measure to identify strategic linkages, “brokers” or “bridging nodes” in a graph.

**Highest Betweenness Centrality:** Maria, then Oliver

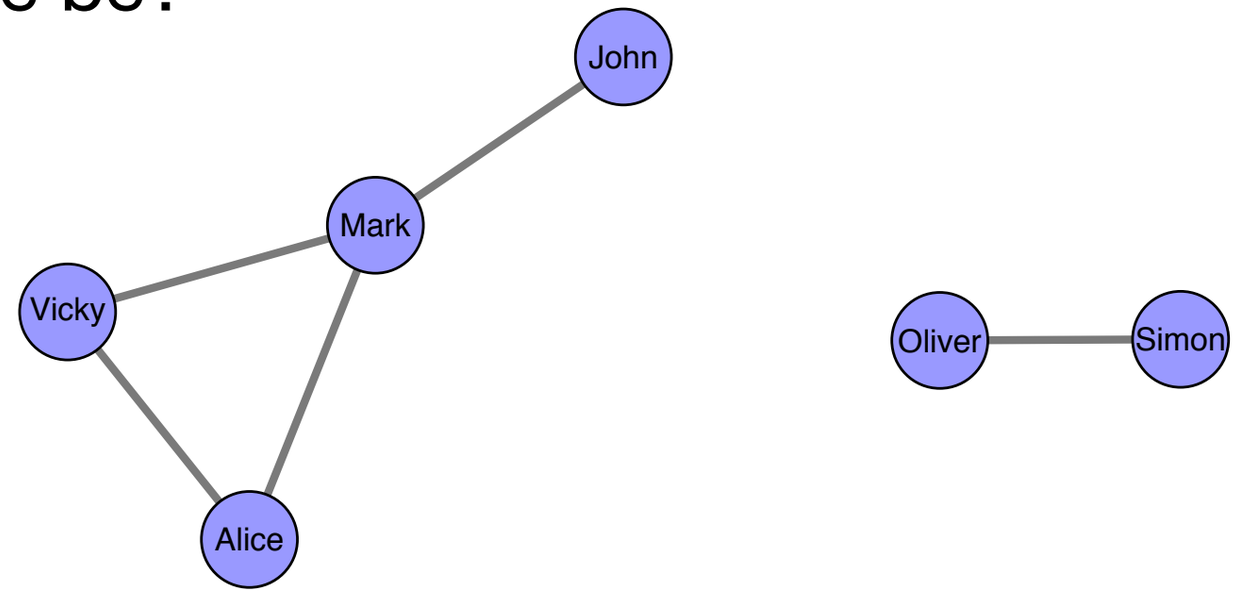
# Tutorial Q1(d)

- If Emma leaves the club (i.e. the network), how many connected components will be?




Still 1 connected component

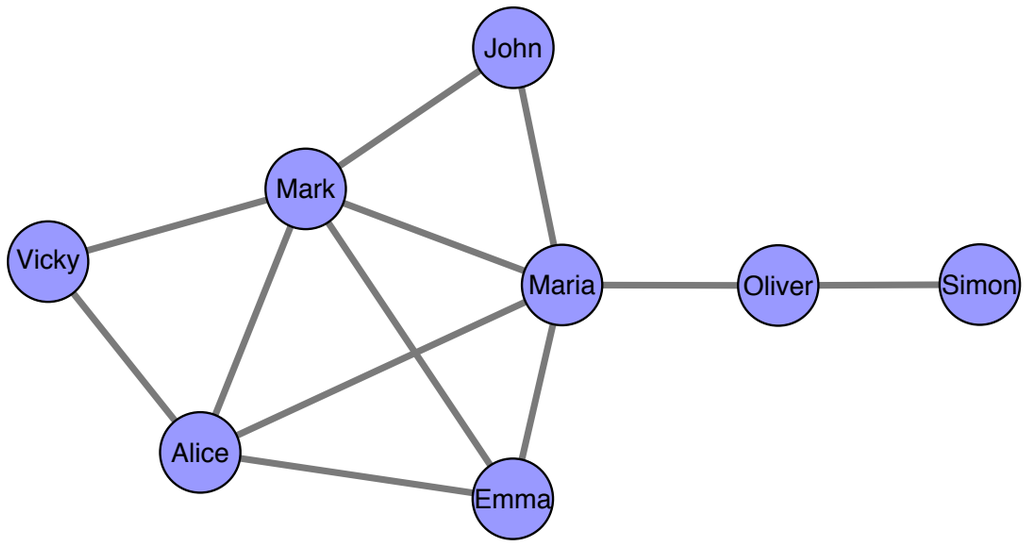
- If Maria then leaves the club, how many connected components will there be?




Now 2 connected components

# Tutorial Q1(e)

- The network above can also be represented as an Adjacency Matrix. Construct this matrix. Will the matrix be symmetric or asymmetric?



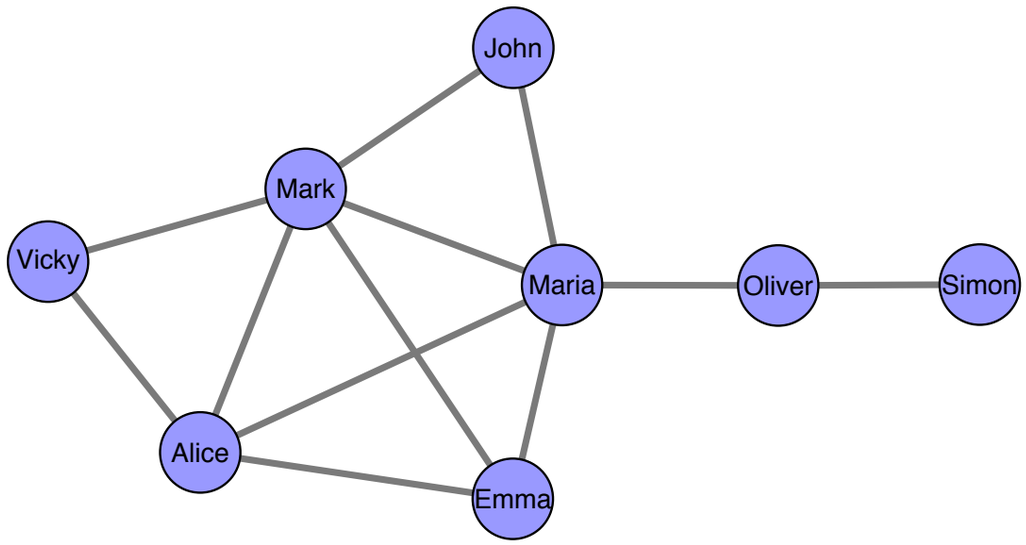

Graph is Undirected, has 8 nodes

8x8 Symmetric Matrix

| | Alice | Emma | John | Mark | Maria | Oliver | Simon | Vicky |
|---|---|---|---|---|---|---|---|---|
| **Alice** | - | | | | | | | |
| **Emma** | | - | | | | | | |
| **John** | | | - | | | | | |
| **Mark** | | | | - | | | | |
| **Maria** | | | | | - | | | |
| **Oliver** | | | | | | - | | |
| **Simon** | | | | | | | - | |
| **Vicky** | | | | | | | | - |

# Tutorial Q1(e)

- The network above can also be represented as an Adjacency Matrix. Construct this matrix. Will the matrix be symmetric or asymmetric?




Graph is Undirected, has 8 nodes

8x8 Symmetric Matrix

| | Alice | Emma | John | Maria | Mark | Oliver | Simon | Vicky |
|---|---|---|---|---|---|---|---|---|
| **Alice** | - | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| **Emma** | 1 | - | 0 | 1 | 1 | 0 | 0 | 0 |
| **John** | 0 | 0 | - | 1 | 1 | 0 | 0 | 0 |
| **Maria** | 1 | 1 | 1 | - | 1 | 1 | 0 | 0 |
| **Mark** | 1 | 1 | 1 | 1 | - | 0 | 0 | 1 |
| **Oliver** | 0 | 0 | 0 | 1 | 0 | - | 1 | 0 |
| **Simon** | 0 | 0 | 0 | 0 | 0 | 1 | - | 0 |
| **Vicky** | 1 | 0 | 0 | 0 | 1 | 0 | 0 | - |

# Tutorial Q2(a)

- The Adjacency Matrix below represents a weighted undirected "co-author network", where each non-zero entry indicates the number of times a pair of researchers have co-authored a paper.

7x7 Symmetric Matrix

| | **Albert** | **Barabasi** | **Collins** | **Jones** | **Lynch** | **Mooney** | **Smith** |
|---|---|---|---|---|---|---|---|
| **Albert** | - | 4 | 1 | 0 | 1 | 0 | 0 |
| **Barabasi** | 4 | - | 0 | 0 | 1 | 0 | 0 |
| **Collins** | 1 | 0 | - | 1 | 0 | 2 | 4 |
| **Jones** | 0 | 0 | 1 | - | 0 | 0 | 0 |
| **Lynch** | 1 | 1 | 0 | 0 | - | 0 | 1 |
| **Mooney** | 0 | 0 | 2 | 0 | 0 | - | 3 |
| **Smith** | 0 | 0 | 4 | 0 | 1 | 3 | - |

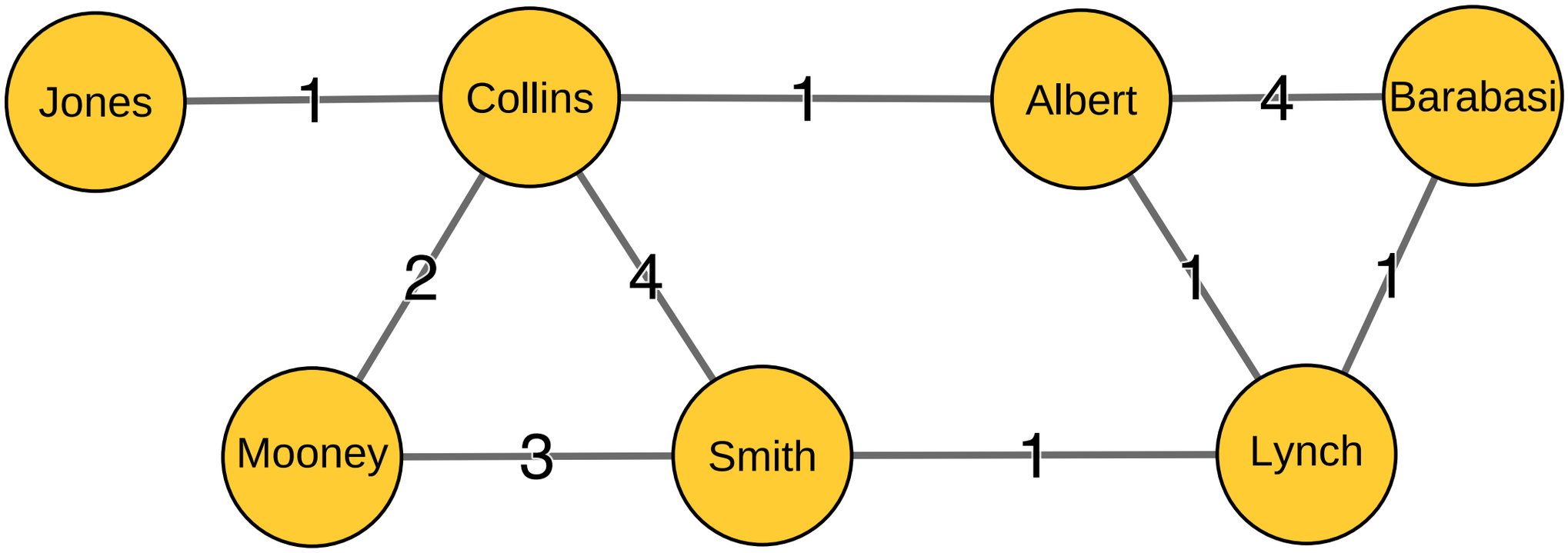
# Tutorial Q2(a)

- Draw the network diagram corresponding to the Adjacency Matrix. Label edges with their corresponding edge weights.

| | Albert | Barabasi | Collins | Jones | Lynch | Mooney | Smith |
|---|---|---|---|---|---|---|---|
| **Albert** | - | 4 | 1 | 0 | 1 | 0 | 0 |
| **Barabasi** | 4 | - | 0 | 0 | 1 | 0 | 0 |
| **Collins** | 1 | 0 | - | 1 | 0 | 2 | 4 |
| **Jones** | 0 | 0 | 1 | - | 0 | 0 | 0 |
| **Lynch** | 1 | 1 | 0 | 0 | - | 0 | 1 |
| **Mooney** | 0 | 0 | 2 | 0 | 0 | - | 3 |
| **Smith** | 0 | 0 | 4 | 0 | 1 | 3 | - |

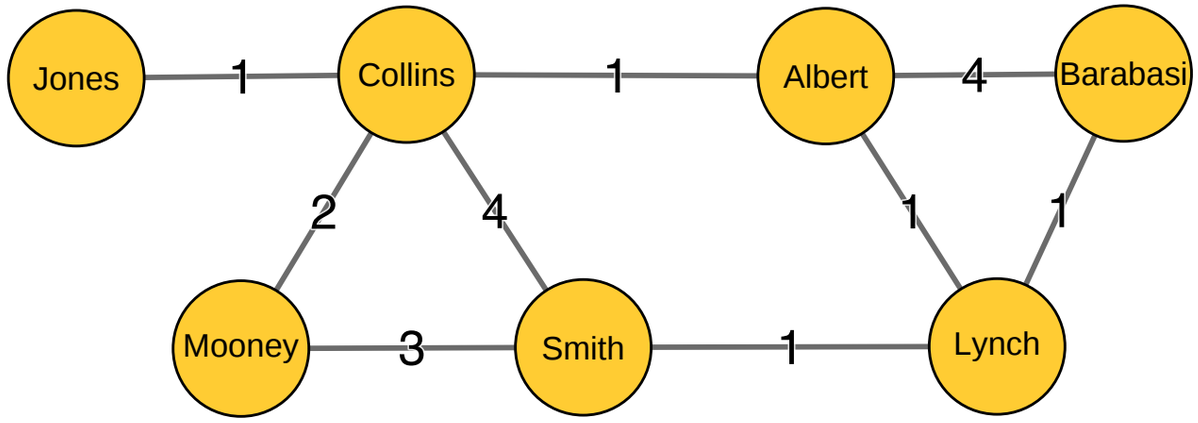
Symmetric 7x7 matrix

⇒ 7 nodes & 9 edges



# Tutorial Q2(b)

- Calculate the *degree centrality* for each node in the graph, ignoring the edge weights. Which authors have the highest and lowest unweighted degree centrality?




| | Albert | Barabasi | Collins | Jones | Lynch | Mooney | Smith |
|---|---|---|---|---|---|---|---|
| **Albert** | - | 4 | 1 | 0 | 1 | 0 | 0 |
| **Barabasi** | 4 | - | 0 | 0 | 1 | 0 | 0 |
| **Collins** | 1 | 0 | - | 1 | 0 | 2 | 4 |
| **Jones** | 0 | 0 | 1 | - | 0 | 0 | 0 |
| **Lynch** | 1 | 1 | 0 | 0 | - | 0 | 1 |
| **Mooney** | 0 | 0 | 2 | 0 | 0 | - | 3 |
| **Smith** | 0 | 0 | 4 | 0 | 1 | 3 | - |

We can count non-zero entries for each row…

| Node | Degree |
|---|---|
| Albert | 3 |
| Barabasi | 2 |
| Collins | 4 |
| Jones | 1 |
| Lynch | 3 |
| Mooney | 2 |
| Smith | 3 |

**Highest Degree:** Collins

**Lowest Degree:** Jones

# Tutorial Q2(c)

- Calculate the *weighted degree centrality* for each node. Which authors have the highest and lowest centrality according to this measure?
- Note: *Weighted degree* is calculated as the sum of the weights on the edges associated with each node.

| | Albert | Barabasi | Collins | Jones | Lynch | Mooney | Smith |
|---|---|---|---|---|---|---|---|
| **Albert** | - | 4 | 1 | 0 | 1 | 0 | 0 |
| **Barabasi** | 4 | - | 0 | 0 | 1 | 0 | 0 |
| **Collins** | 1 | 0 | - | 1 | 0 | 2 | 4 |
| **Jones** | 0 | 0 | 1 | - | 0 | 0 | 0 |
| **Lynch** | 1 | 1 | 0 | 0 | - | 0 | 1 |
| **Mooney** | 0 | 0 | 2 | 0 | 0 | - | 3 |
| **Smith** | 0 | 0 | 4 | 0 | 1 | 3 | - |

We can sum the entries for each row…

| Node | Weighted Degree |
|---|---|
| Albert | 6 |
| Barabasi | 5 |
| Collins | 8 |
| Jones | 1 |
| Lynch | 3 |
| Mooney | 5 |
| Smith | 8 |

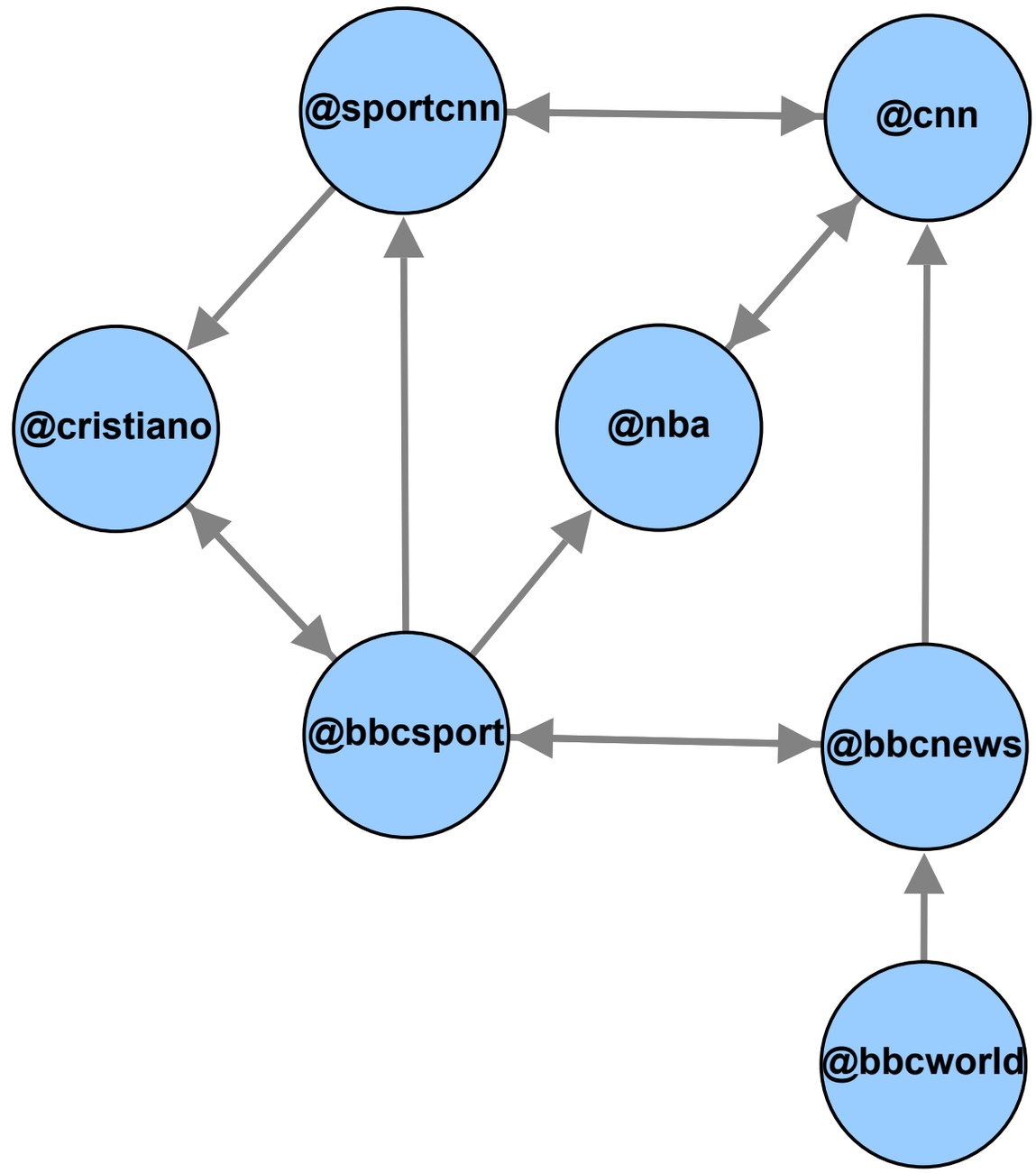
**Highest Degree:** Collins & Smith

**Lowest Degree:** Jones

# Tutorial Q3(a)

- The edge list below represents a small subset of a Twitter *follower* network. An edge *(X, Y)* indicates that account *X* follows account *Y*.

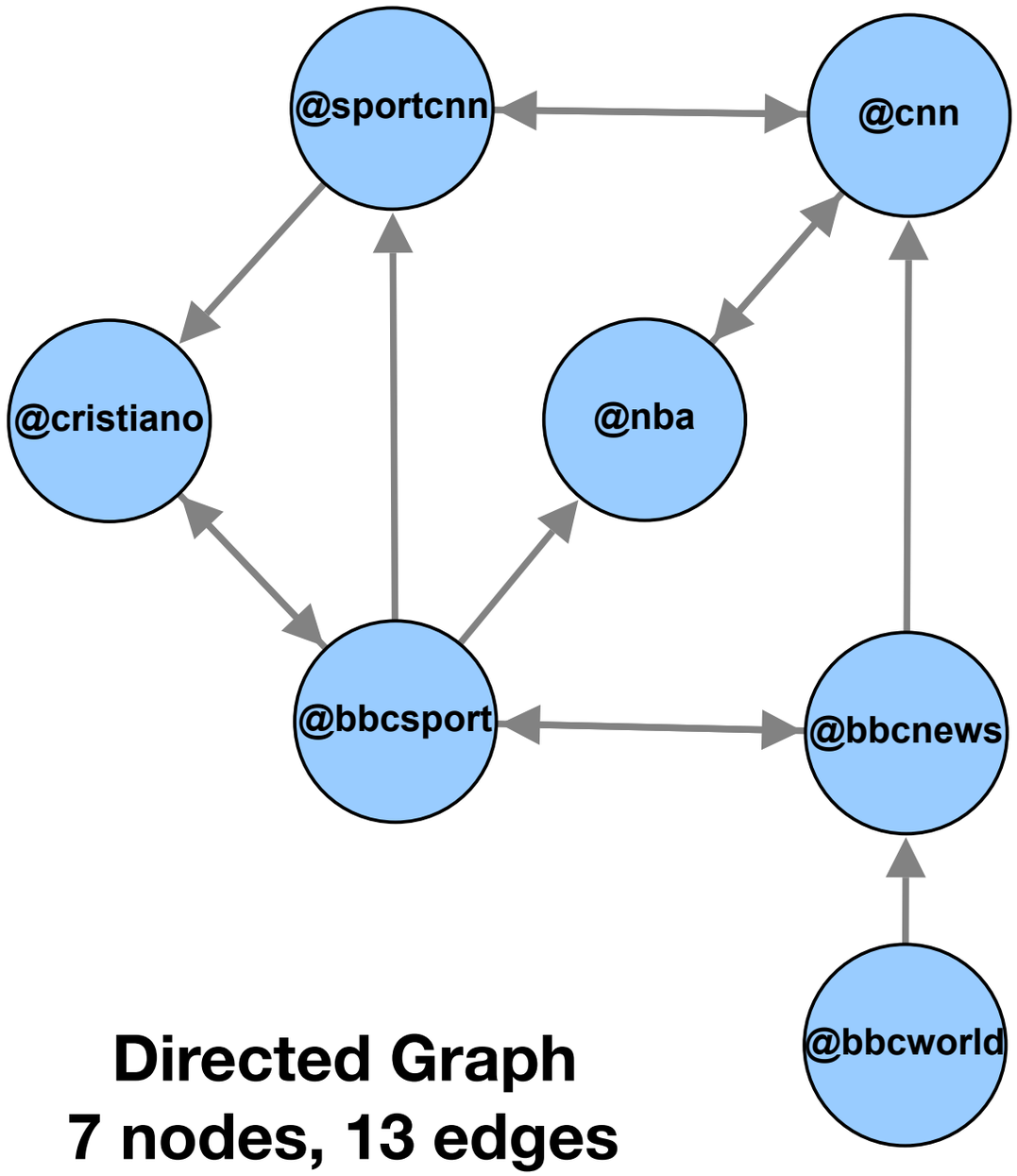
@bbcnews,@bbcsport
@bbcnews,@cnn
@bbcsport,@bbcnews
@bbcsport,@sportcnn
@bbcsport,@nba
@bbcworld,@bbcnews
@cnn,@sportcnn
@cnn,@nba
@sportcnn,@cnn
@sportcnn,@cristiano
@bbcsport,@cristiano
@cristiano,@bbcsport
@nba,@cnn




**Directed Graph**
**7 nodes, 13 edges**

# Tutorial Q3(b)

- Calculate the density of this graph.




**Directed Graph**
**7 nodes, 13 edges**

Density for directed graph with *n* nodes and *m* edges

$$\frac{m}{n \times (n-1)}$$

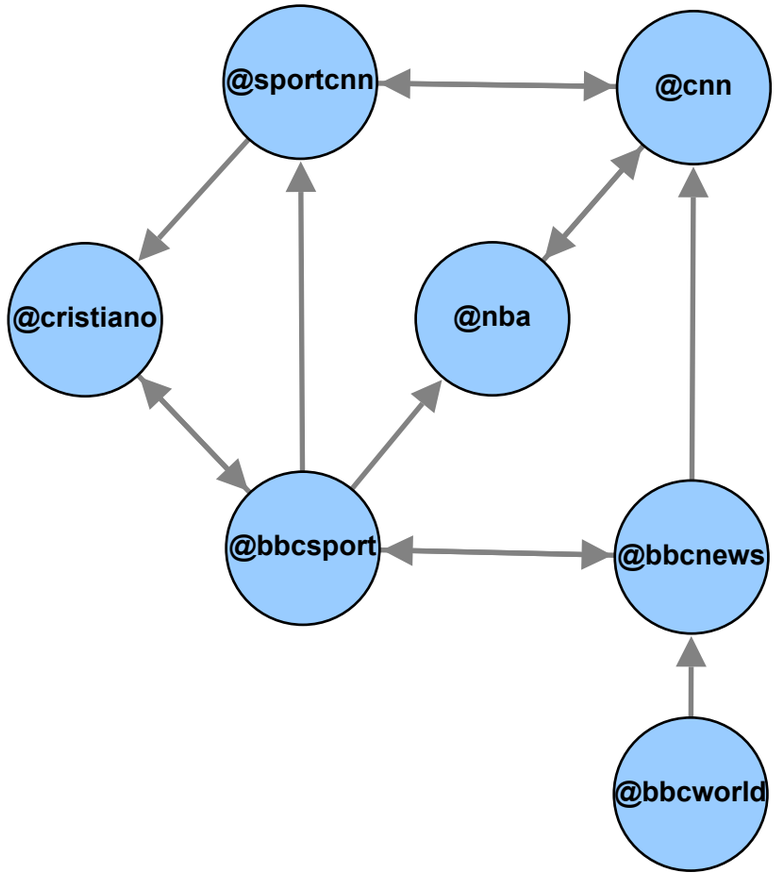
$$density = \frac{13}{7 \times 6} = 0.31$$

# Tutorial Q3(c)

- One proposed measure of “prestige” on Twitter has been to calculate an account’s *Followers-Following* ratio. For node **X**, this is equivalent to:

$$\frac{\text{In-Degree(X)}}{\text{Out-Degree(X)}}$$

- Calculate this score for each node in the graph. Which accounts have the highest and lowest “prestige” score?



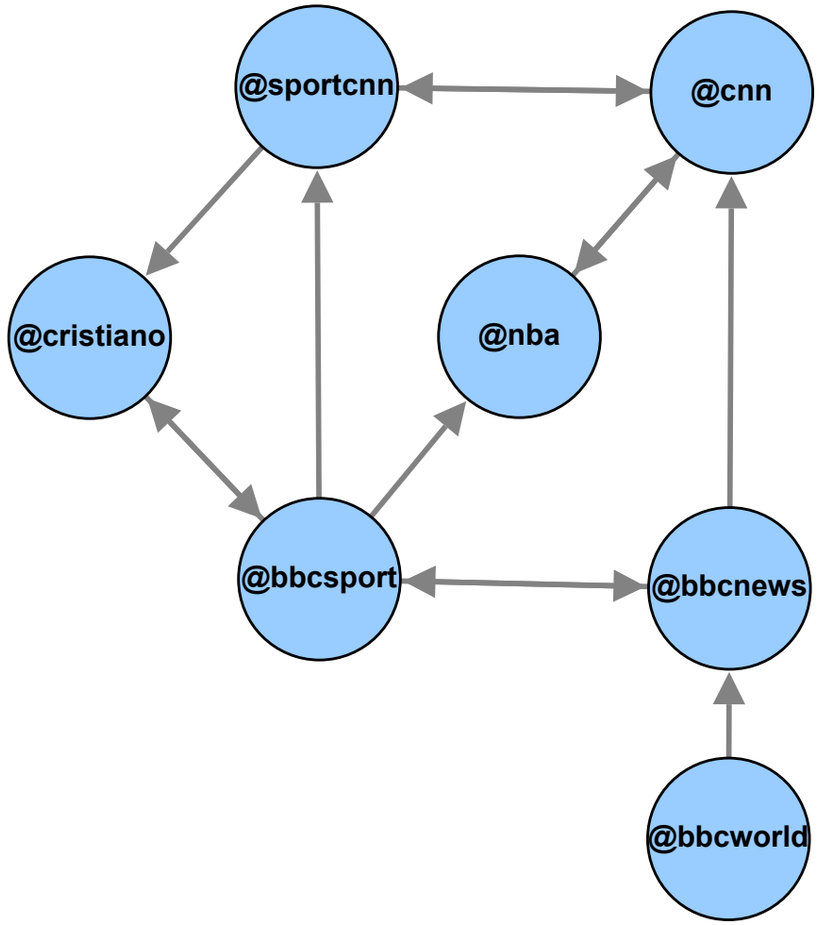

| Node | In-Degree | Out-Degree | Score |
|---|---|---|---|
| @bbcnews | 2 | 2 | 1.00 |
| @bbcsport | 2 | 4 | 0.50 |
| @bbcworld | 0 | 1 | 0.00 |
| @cnn | 3 | 2 | 1.50 |
| @cristiano | 2 | 1 | 2.00 |
| @nba | 2 | 1 | 2.00 |
| @sportcnn | 2 | 2 | 1.00 |

**Highest:** @cristiano, @nba

**Lowest:** @bbcworld

# Tutorial Q3(d)

- One way to convert a directed graph into an undirected graph is to only include those edges that are *reciprocated* (i.e. only create an undirected edge between nodes *A* and *B*, if there exist both edges $A \rightarrow B$ and $B \rightarrow A$ in the original graph).
- Draw the undirected reciprocal network from the undirected network, removing any isolated nodes which have no edges. How many nodes and edges will exist in the new graph?




**Reciprocated Edges**

(@bbcnews,@bbcsport) & (@bbcsport,@bbcnews)

(@bbcsport,@cristiano) & (@cristiano,@bbcsport)

(@cnn,@sportcnn) & (@sportcnn,@cnn)

(@cnn,@nba) & (@nba,@cnn)

# Tutorial Q3(d)

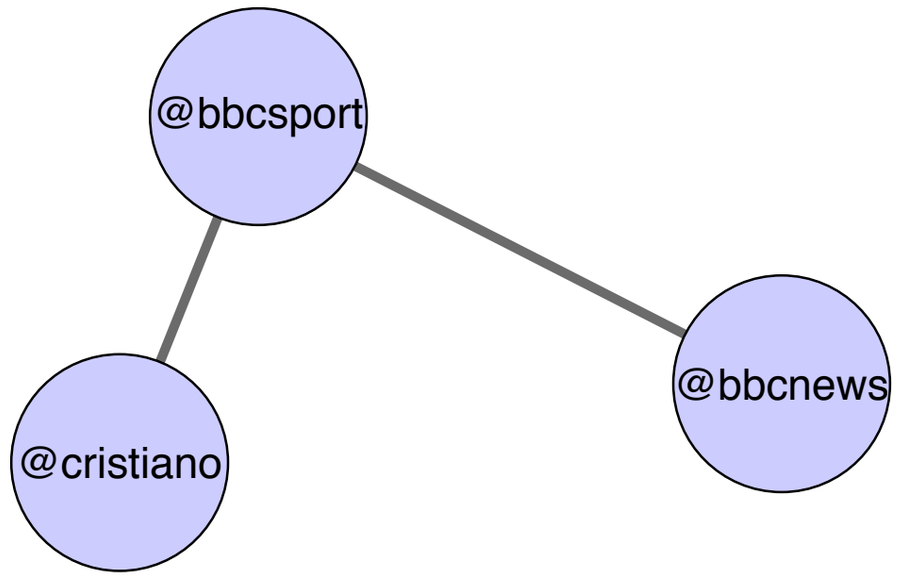
**Reciprocated Edges**

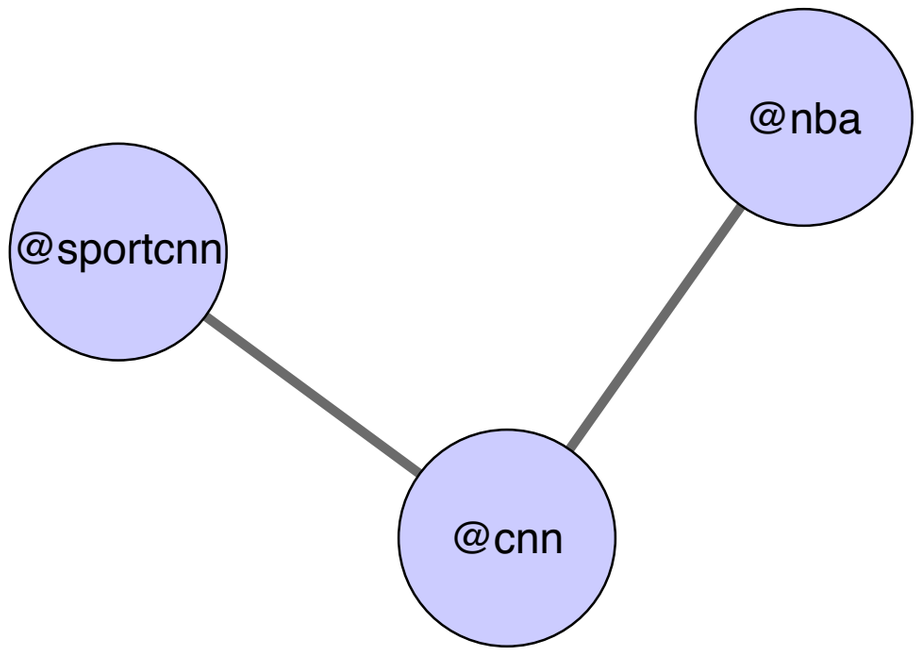
(@bbcnews,@bbcsport) & (@bbcsport,@bbcnews)

(@bbcsport,@cristiano) & (@cristiano,@bbcsport)

(@cnn,@sportcnn) & (@sportcnn,@cnn)

(@cnn,@nba) & (@nba,@cnn)




**New undirected graph:**
6 nodes, 4 edges

**Remove nodes:** @bbcworld

# Assignment 3 - Ensembles in Weka

**Objectives:**

Objective is to use the ensemble functionality in Weka to identify the extent to which classification performance can be improved through the combination of multiple classifiers.

**Data:**

Dataset of patients undergoing voice rehabilitation treatment. Binary classification task is to predict whether treatment will be successful (class=yes) or unsuccessful (class=no).

You should download your personal dataset for the assignment from the URL:

http://mlg.ucd.ie/datasets/comp30120/voice/<STUDENT_NUMBER>.arff

For example, if your student number is 126023491, your dataset is at the URL:

http://mlg.ucd.ie/datasets/comp30120/voice/126023491.arff

**Tasks:**

Complete all tasks on your dataset, discuss the results in your report. See full assignment PDF on Moodle for task details.

# Assignment 3 - Ensembles in Weka

**Guidelines:**

1. When downloading the dataset, please ensure your student number is correct. Only use your assigned dataset. Submissions using an incorrect dataset will receive a 0 grade.
2. This is an individual assignment. Plagiarism will be treated seriously. Evidence of plagiarism in the assignment will result in a 0 mark.
3. Submit your report as **a single PDF** via the COMP30120 CS Moodle page. Include your full name and student ID number in the report. Remember to **submit fully** on Moodle.
4. Assignment should be submitted on or before **Monday November 30th**. Note this is a **hard deadline**.

**See full assignment PDF on Moodle**

**15% of overall grade**

# COMP30120 - Theory Exam

- **Tuesday 24th November 9am B004, arrive early.**
- 35% of overall module mark
- 45 minutes
- Choice: 8 questions, answer any 6
- All questions carry equal marks
- Written exam, use exam booklets
- Printed course notes and personal notes can be used
- **No use of laptop, phones, or tablets etc**
- **Scientific calculator required for exam, not phone**

# COMP30120 - Theory Exam

- **Exam Topics:**
  - Classification: KNN, Decision Trees, Naive Bayes
  - Evaluation in Machine Learning
  - Feature Selection & Dimension Reduction
  - Clustering: Partitional, Hierarchical, Validation
  - Ensembles for Classification
  - Social Network Analysis
  - Recommender Systems
- No Weka related questions
- See tutorial solutions and lecture notes for relevant examples.