# 04-630
# Data Structures and Algorithms for Engineers

David Vernon
Carnegie Mellon University Africa

vernon@cmu.edu
www.vernon.eu

# Lecture 25

## Complex Networks

- Communities
  - Fundamental Hypothesis & Connectedness and Density Hypothesis
  - Strong and weak communities
  - Graph partitioning & Community detection
    - Hierarchical clustering
    - Girvan-Newman Algorithm
    - Modularity
    - Random Hypothesis
    - Maximum Modularity Hypothesis
    - Greedy algorithm for community detection by maximizing modularity
  - Overlapping communities
    - Clique percolation algorithm and CFinder

This lecture is based on Chapter 9 of *Network Science* by A.-L. Barabási
(see http://barabasi.com/book/network-science)

# Network Science

by Albert-László Barabási



Start Reading

# Complex Networks

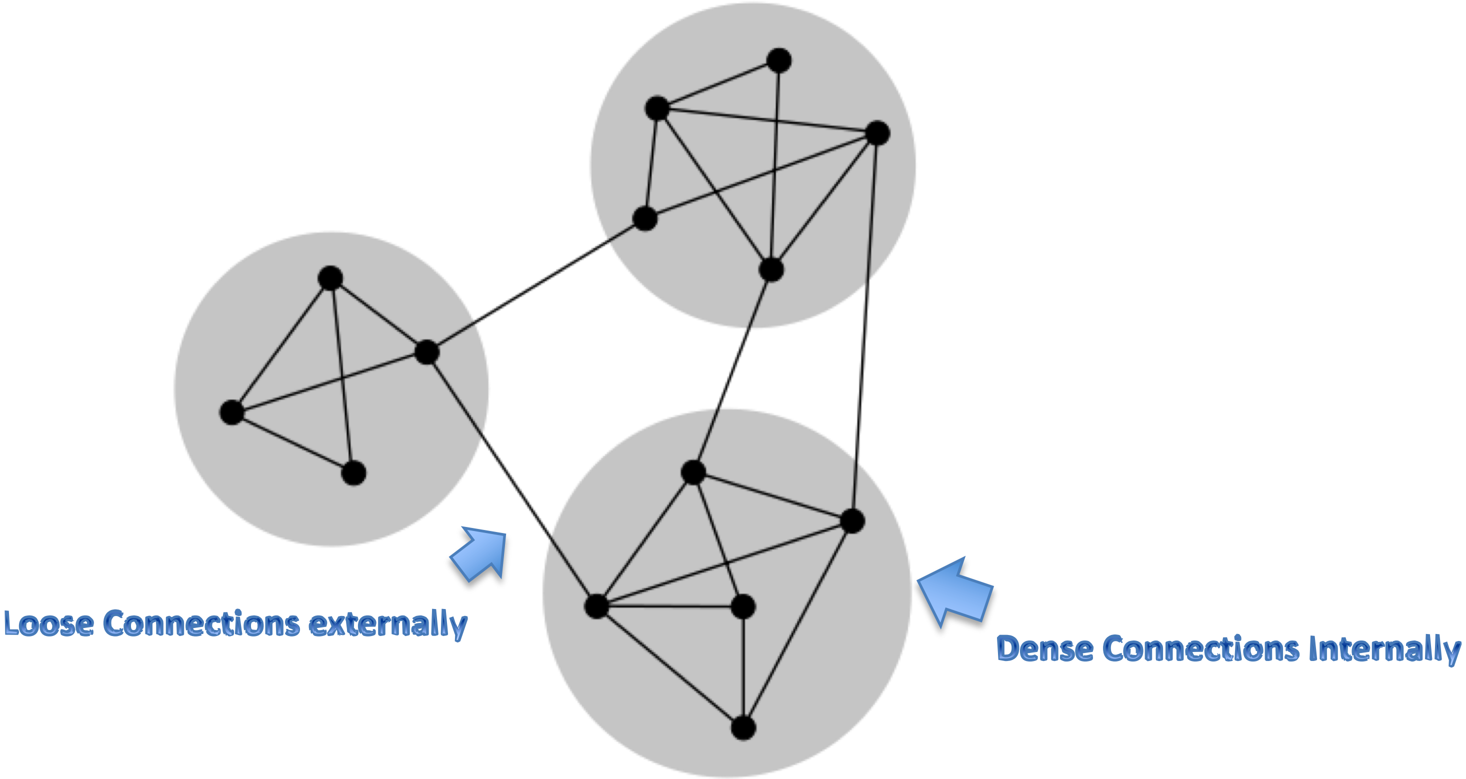
## **Communities**

"In network science we call a *community* a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities."

L.A. Barabási

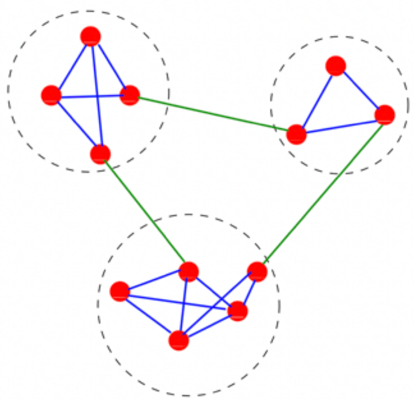
# Complex Networks

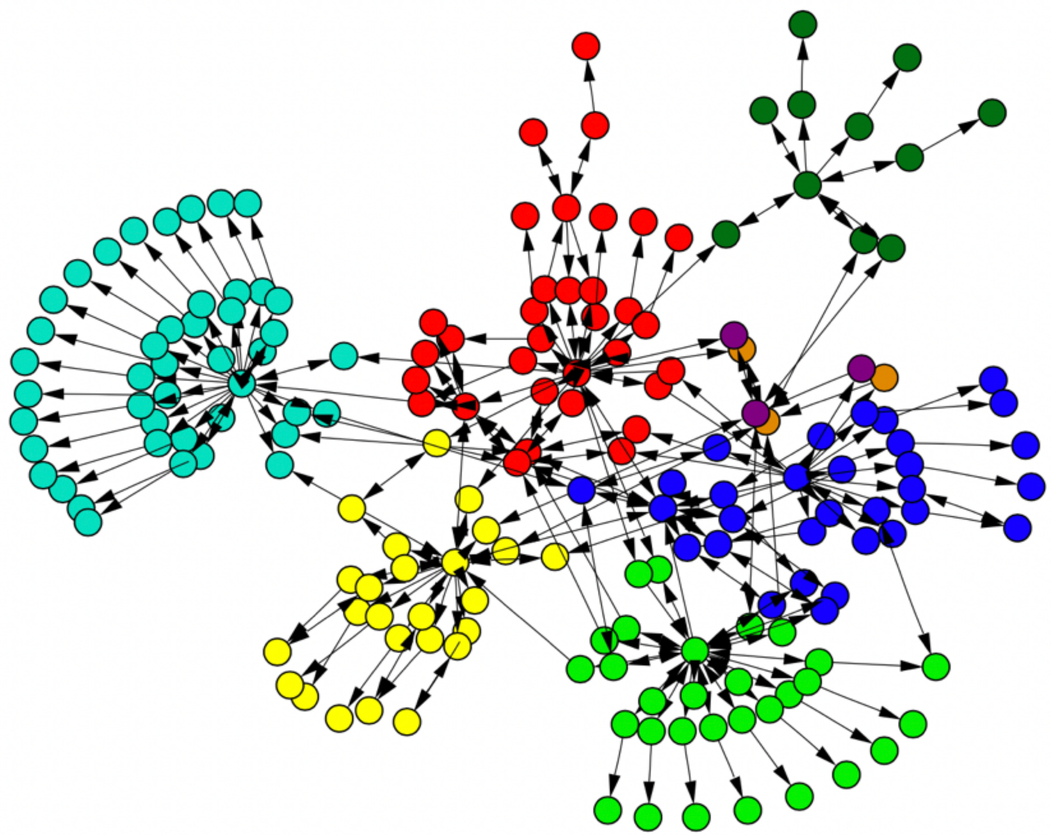
## Communities



# Complex Networks

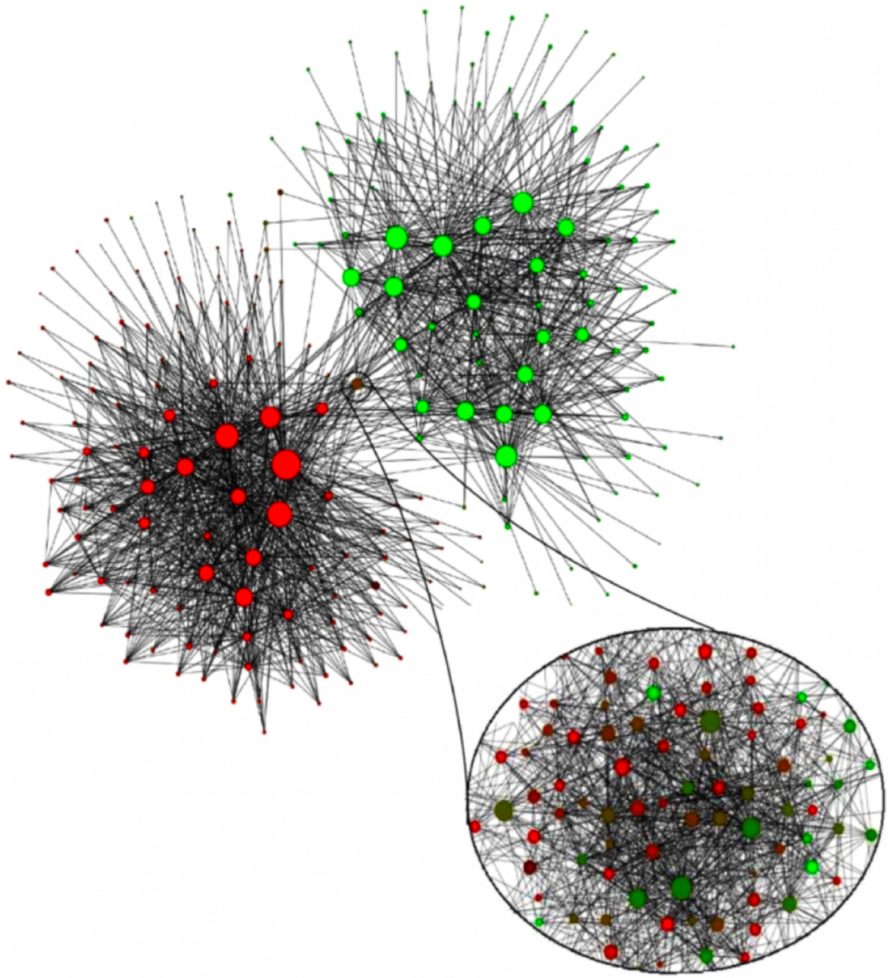
## Communities





# Complex Networks

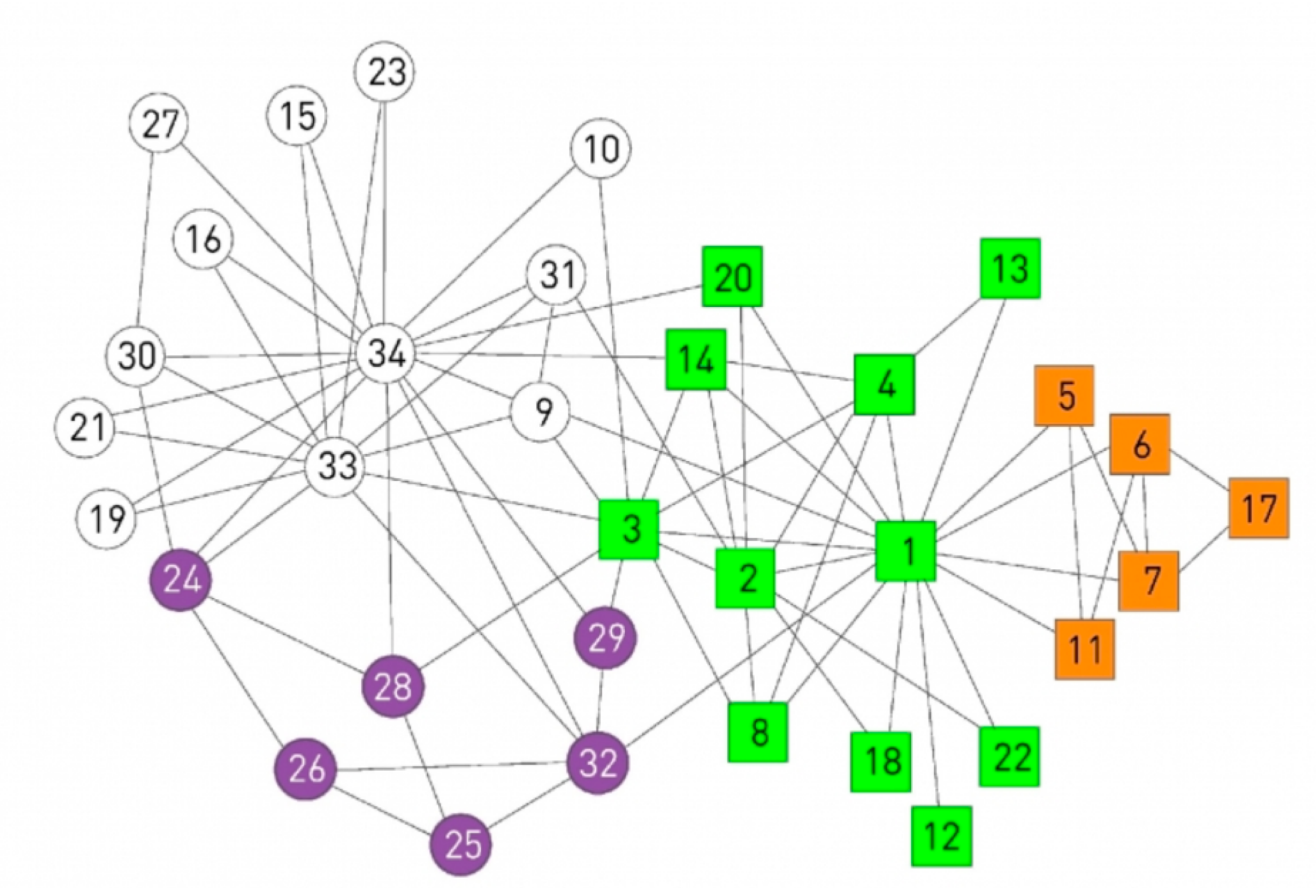
## Communities



Communities in Belgium: red, French-speaking; green, Flemish-speaking
(node size = community size)

# Complex Networks

## Communities

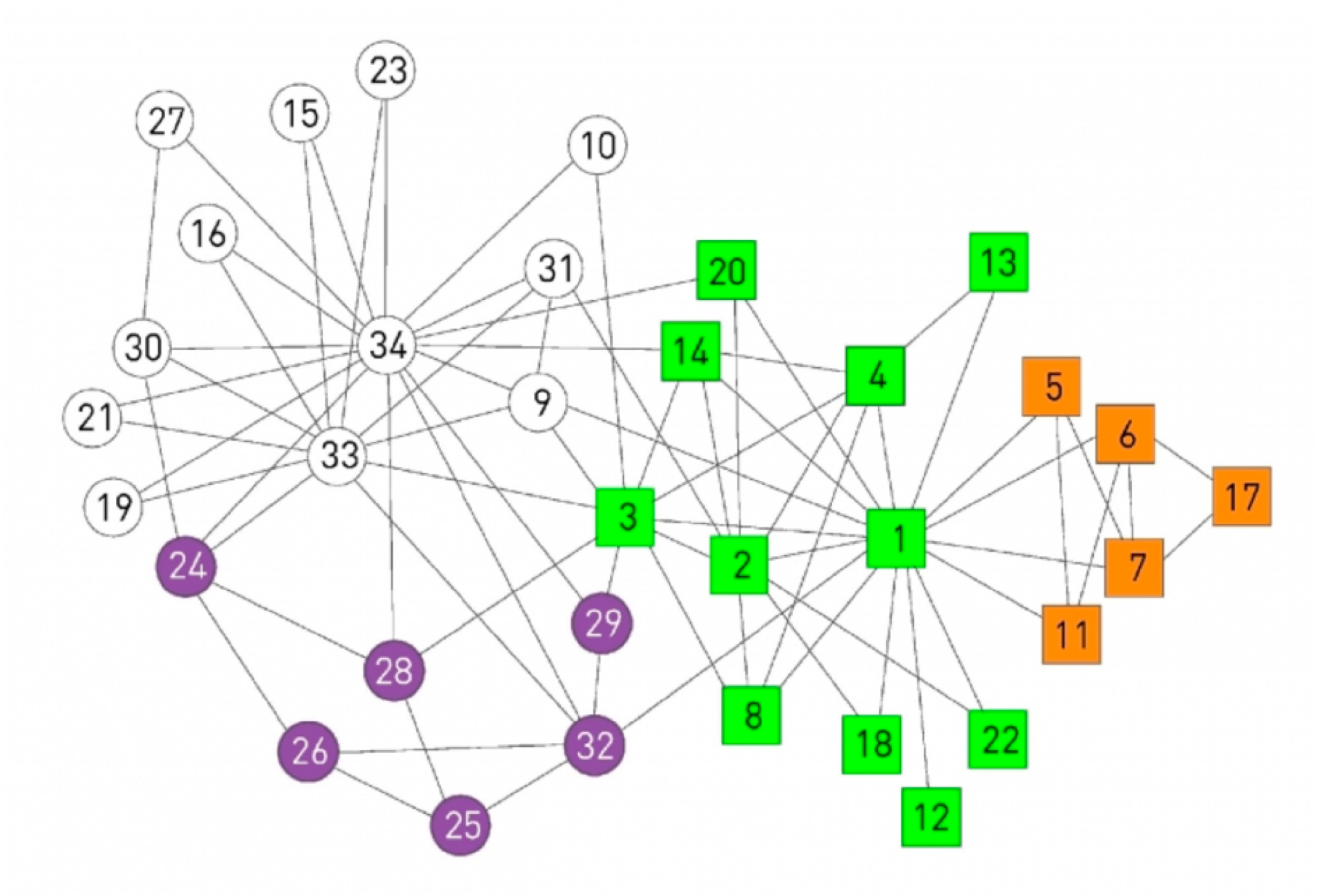



*Zachary's Karate Club*::

A conflict between the club's president and the instructor split the club into two. About half of the members followed the instructor and the other half the president, a breakup that unveiled the ground truth, representing club's underlying community structure

# Complex Networks

## Communities



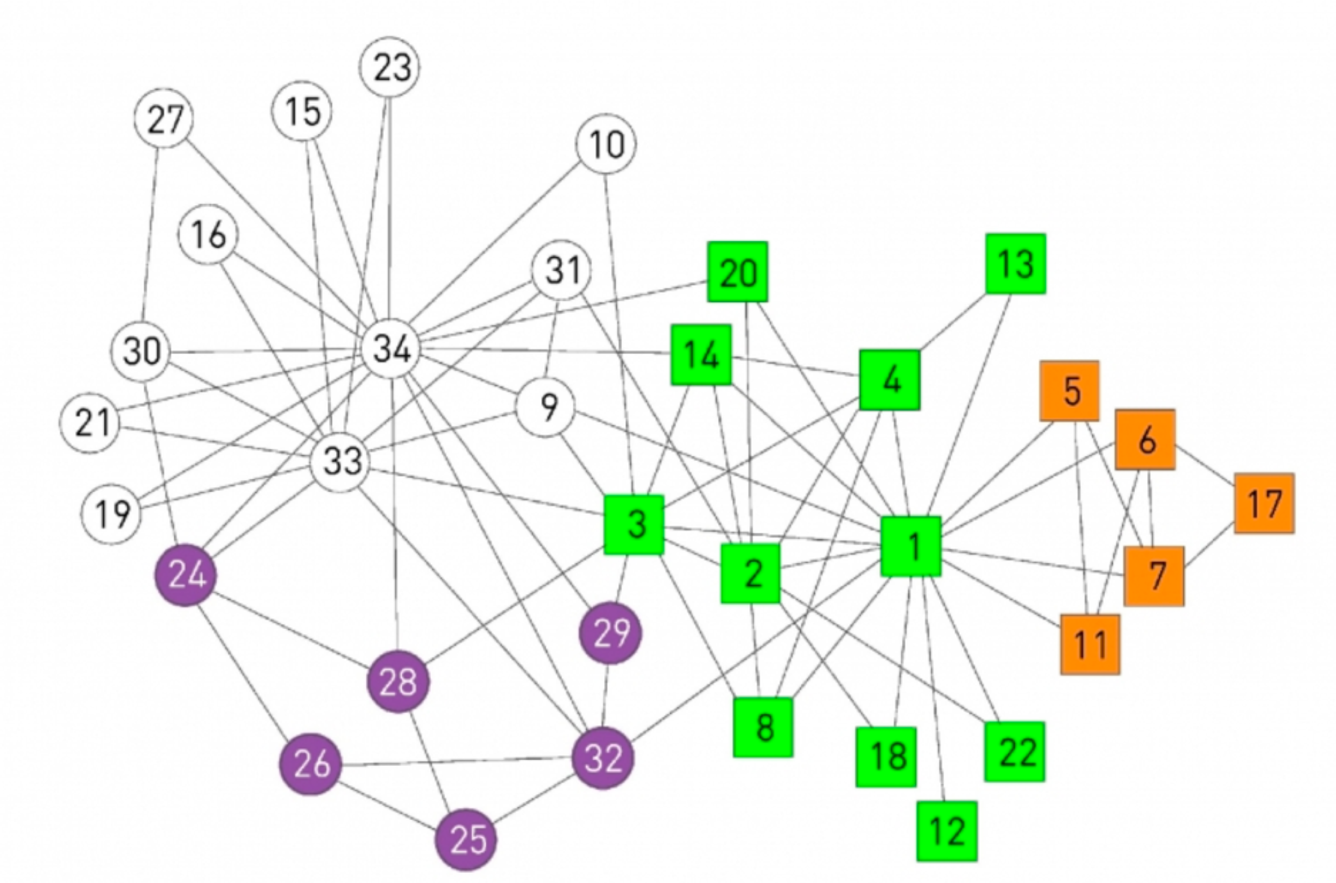

*Zachary's Karate Club*::

Links capture interactions between the club members *outside the club*.

The circles and the squares denote the two factions that emerged after the club split in two.

# Complex Networks

## Communities




*Zachary's Karate Club*::
The colors capture the best community partition predicted by an algorithm that optimizes the modularity coefficient

# Complex Networks

## Communities

**H1: Fundamental Hypothesis**

A network's community structure is uniquely encoded in its wiring diagram.

# Complex Networks

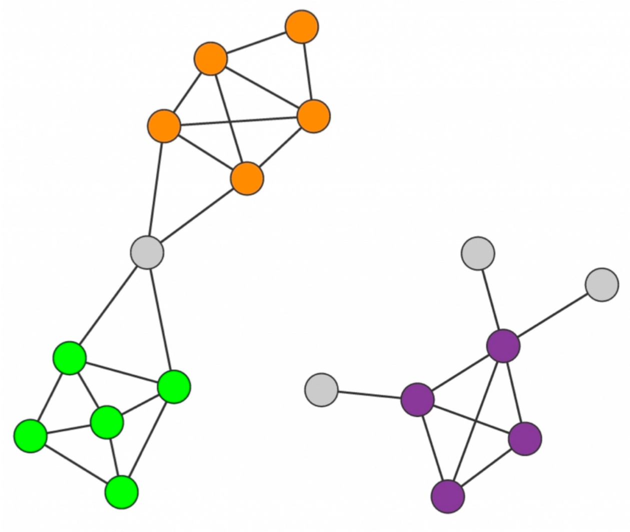
## Communities

**H2: Connectedness and Density Hypothesis**

*A community is a locally dense connected subgraph in a network*

Connected: all members of a community must be reached through other members of the same community

Dense: nodes that belong to a community have a higher probability to link to the other members of that community than to nodes that do not belong to the same community



# Complex Networks

## Communities

**Strong Community**

*C* is a *strong community* if each node within *C* has more links within the community than with the rest of the graph

Specifically, a subgraph *C* forms a strong community if for each node $i \in C$,

$$k_i^{\text{int}}(C) > k_i^{ext}(C)$$

# Complex Networks

## Communities

### Weak Community

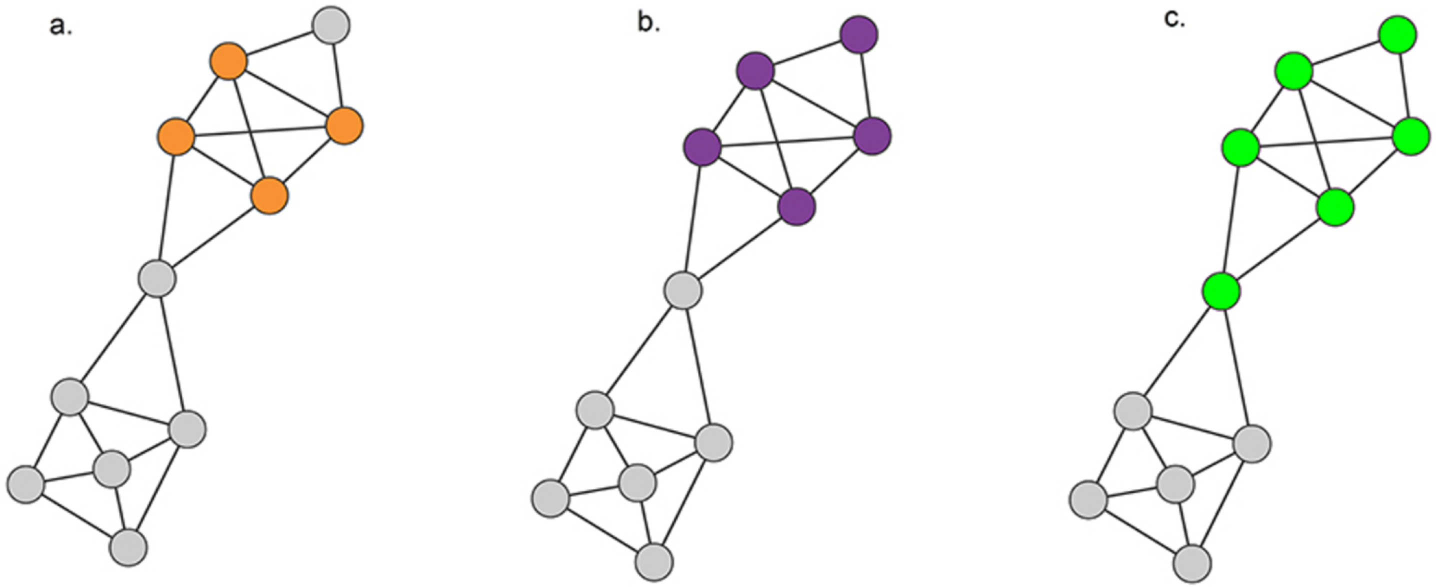
*C* is a *weak community* if the total internal degree of a subgraph exceeds its total external degree

Specifically, a subgraph *C* forms a weak community if

$$\sum_{i \in C} k_i^{\text{int}}(C) > \sum_{i \in C} k_i^{ext}(C)$$

# Complex Networks

## Communities




**a. clique** **b. strong community** **c. weak community**

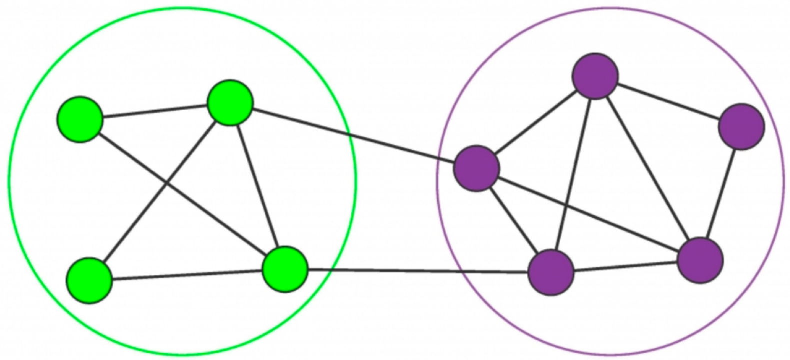
a *clique* corresponds to a complete subgraph (rare)

# Complex Networks
## Communities

### Numbers of communities
How many ways can we group the nodes of a network into communities?

Graph partitioning, also called ***graph bisection***:

We aim to divide a network into two non-overlapping subgraphs, such that the number of links between the nodes in the two groups, called the ***cut size***, is minimized



# Complex Networks

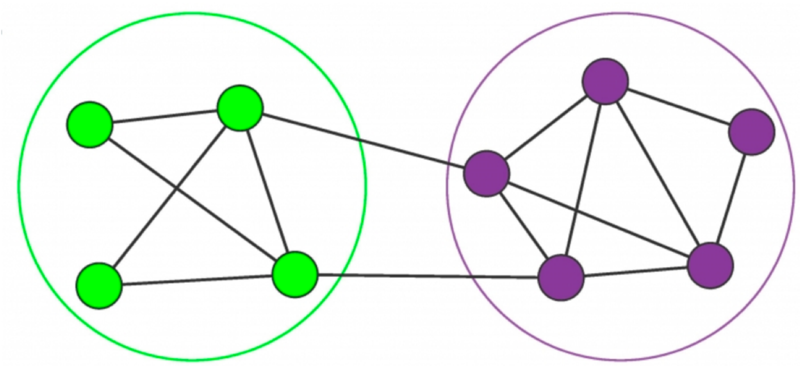
## Communities

### Numbers of communities

How many ways can we group the nodes of a network into communities?

**Graph Bisection**

Brute-force solution: inspect all possible divisions into two groups and choosing the one with the smallest cut size (exponential complexity)



# Complex Networks

## Communities

**Graph partitioning vs. community detection**

- Graph partitioning divides a network into a predefined number of smaller subgraphs
- Community detection aims to uncover the inherent community structure of a network

# Complex Networks

## Communities

### Community detection

- Graph partitioning:
  the number and the size of communities is predefined
- Community detection:
  both parameters are unknown
- Idea: *detect communities by investigating all possible partitions*

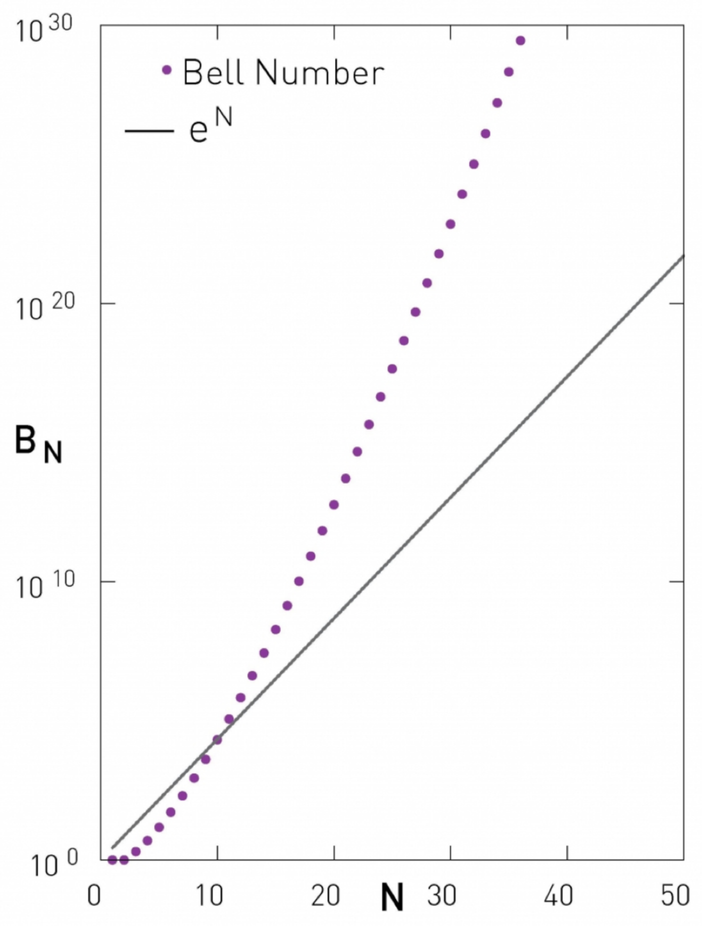
The number of possible partitions is given by $B_N = \frac{1}{e}\sum_{j=0}^{\infty}\frac{j^N}{j!}$

# Complex Networks

## Communities

### Community detection

$$B_N = \frac{1}{e}\sum_{j=0}^{\infty}\frac{j^N}{j!}$$




Brute-force exponential-complexity algorithms that aim to identify communities by inspecting all possible partitions are computationally infeasible

# Complex Networks

## Communities

### Community detection

**We need polynomial-time algorithms that can uncover the community structure of large real networks …**

**Hierarchical Clustering**

Brute-force exponential-complexity algorithms that aim to identify communities by inspecting all possible partitions are computationally infeasible

# Complex Networks

## Community Detection

### Hierarchical Clustering

- Generate a similarity matrix $x_{ij}$ indicating the similarity between vertex/node $i$ and vertex/node $j$
- Iteratively identify groups of nodes with high similarity

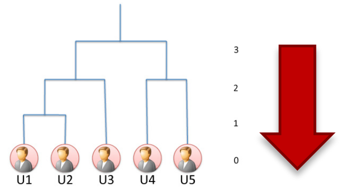
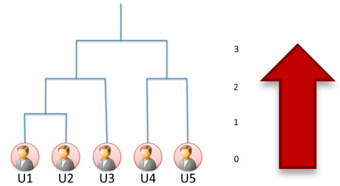
1. *Agglomerative algorithms*
   merge nodes with high similarity into the same community

2. *Divisive algorithms*
   isolate communities by removing low similarity links that tend to connect communities.

Both procedures generate a hierarchical tree, called a dendrogram, that predicts the possible community partitions

# Complex Networks

## Communities

| Publication | Highlights | Example |
| --- | --- | --- |
| Newman and Girvan (2004) | ❑ Divisive Algorithm<br>❑ Remove the edge iteratively from the network |  |
| Newman (2004) | ❑ Agglomerative Algorithm<br>❑ Modularity: measure quality of communities |  |

# Complex Networks

## Community Detection

**Divisive Procedures: the Girvan-Newman Algorithm**

Step 1: Define Centrality

Step 2: Hierarchical Clustering

# Complex Networks

## Community Detection

### Divisive Procedures: the Girvan-Newman Algorithm

Step 1: Define Centrality

The similarity matrix $x_{ij}$ is called centrality and selects node pairs that are in different communities

$x_{ij}$ is high if nodes $i$ and $j$ belong to different communities
$x_{ij}$ is low if they are in the same community

Several options to choose from …

# Complex Networks

## Community Detection

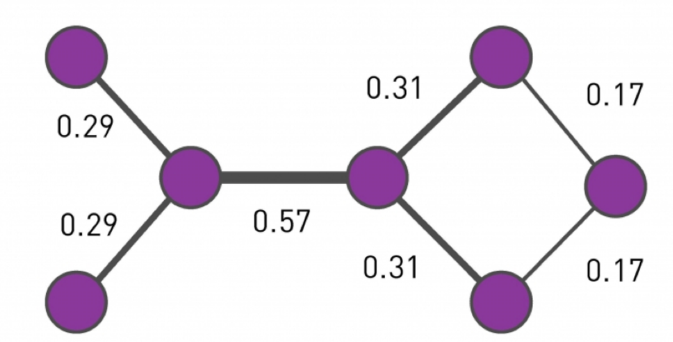
**Divisive Procedures: the Girvan-Newman Algorithm**

Step 1: Define Centrality

**link betweenness**

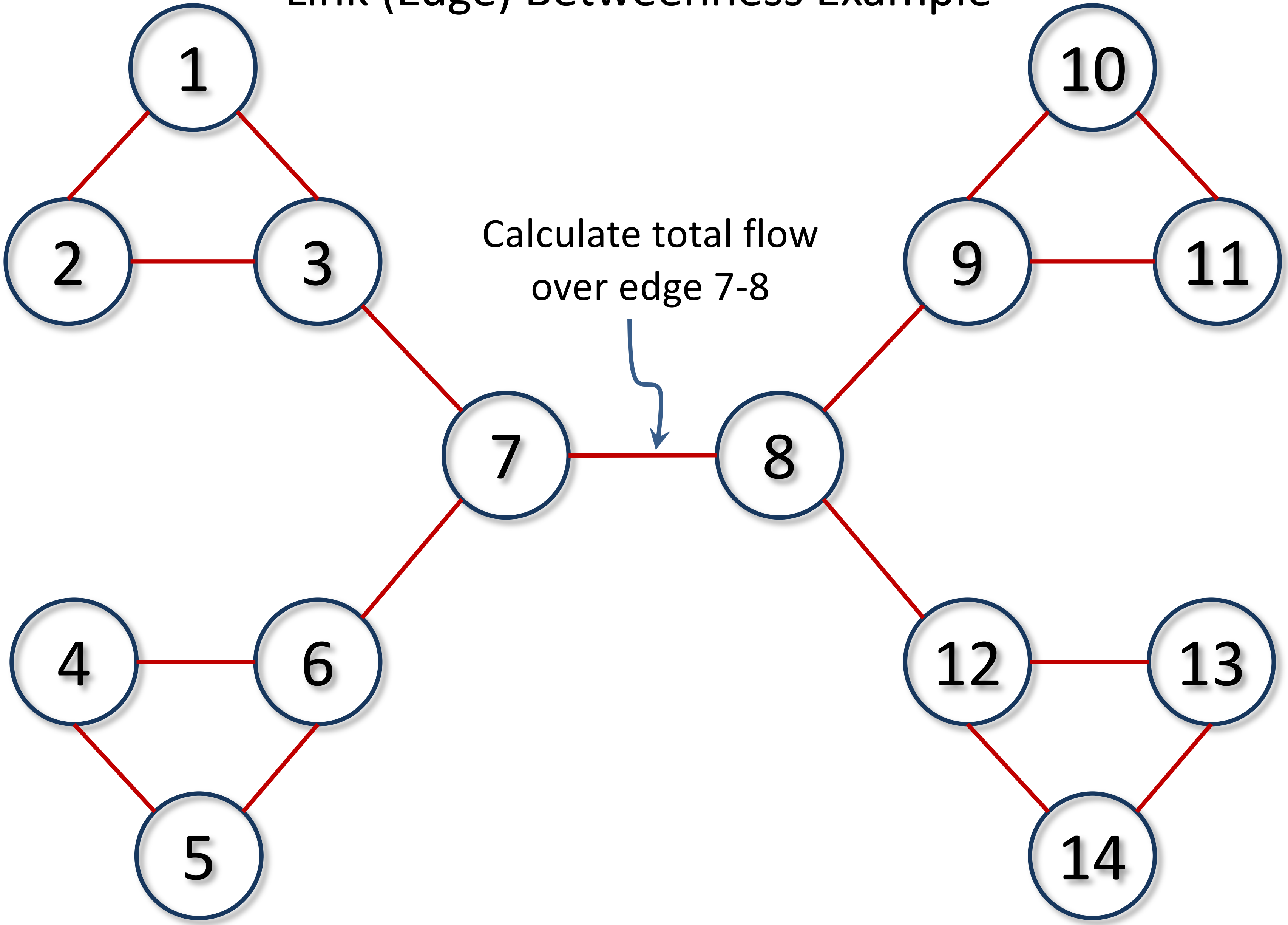
$x_{ij}$ is defined as the number of shortest paths that go through the link $(i, j)$

Links connecting different communities are expected to have large $x_{ij}$ while links within a community have small $x_{ij}$

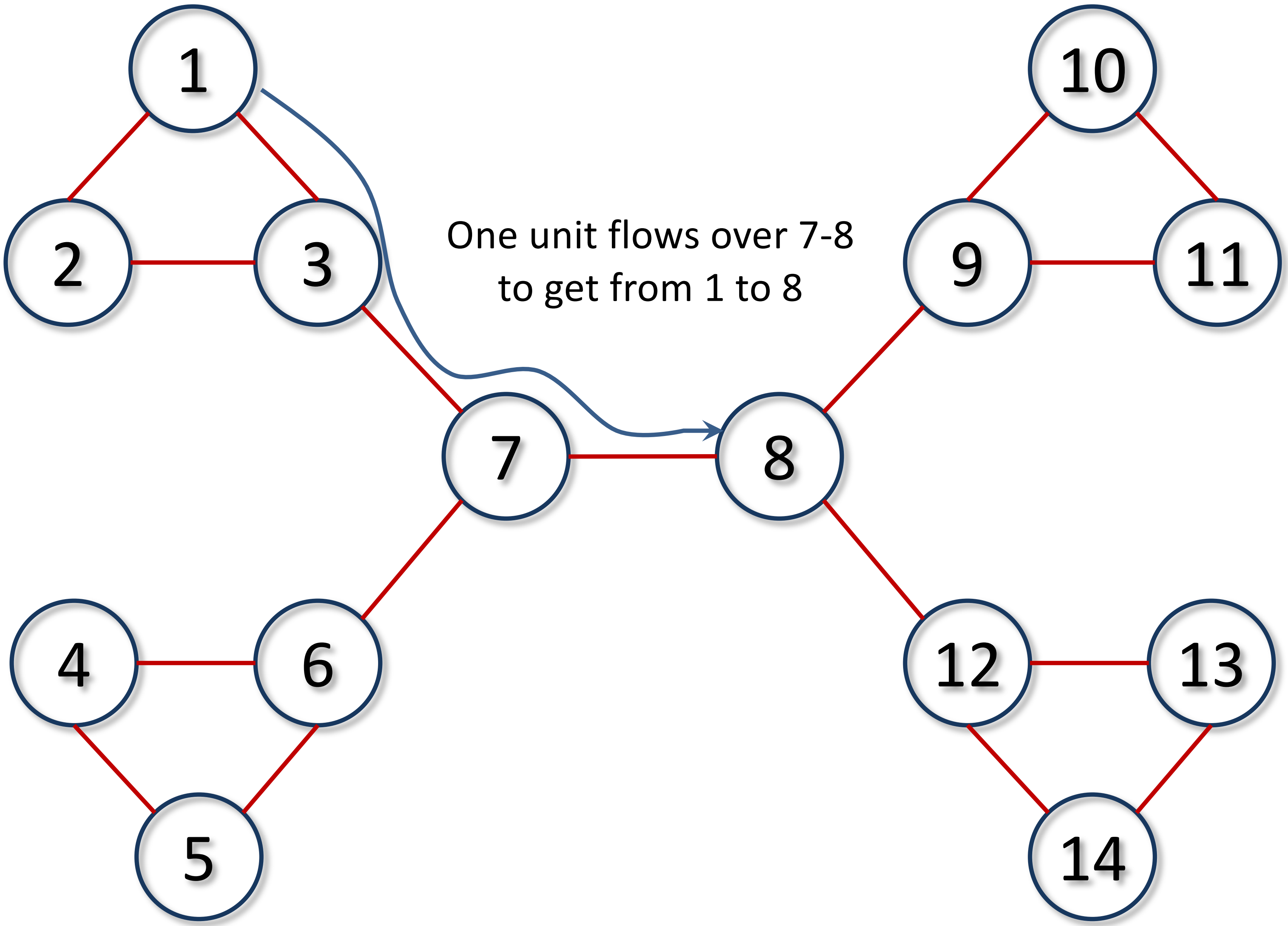


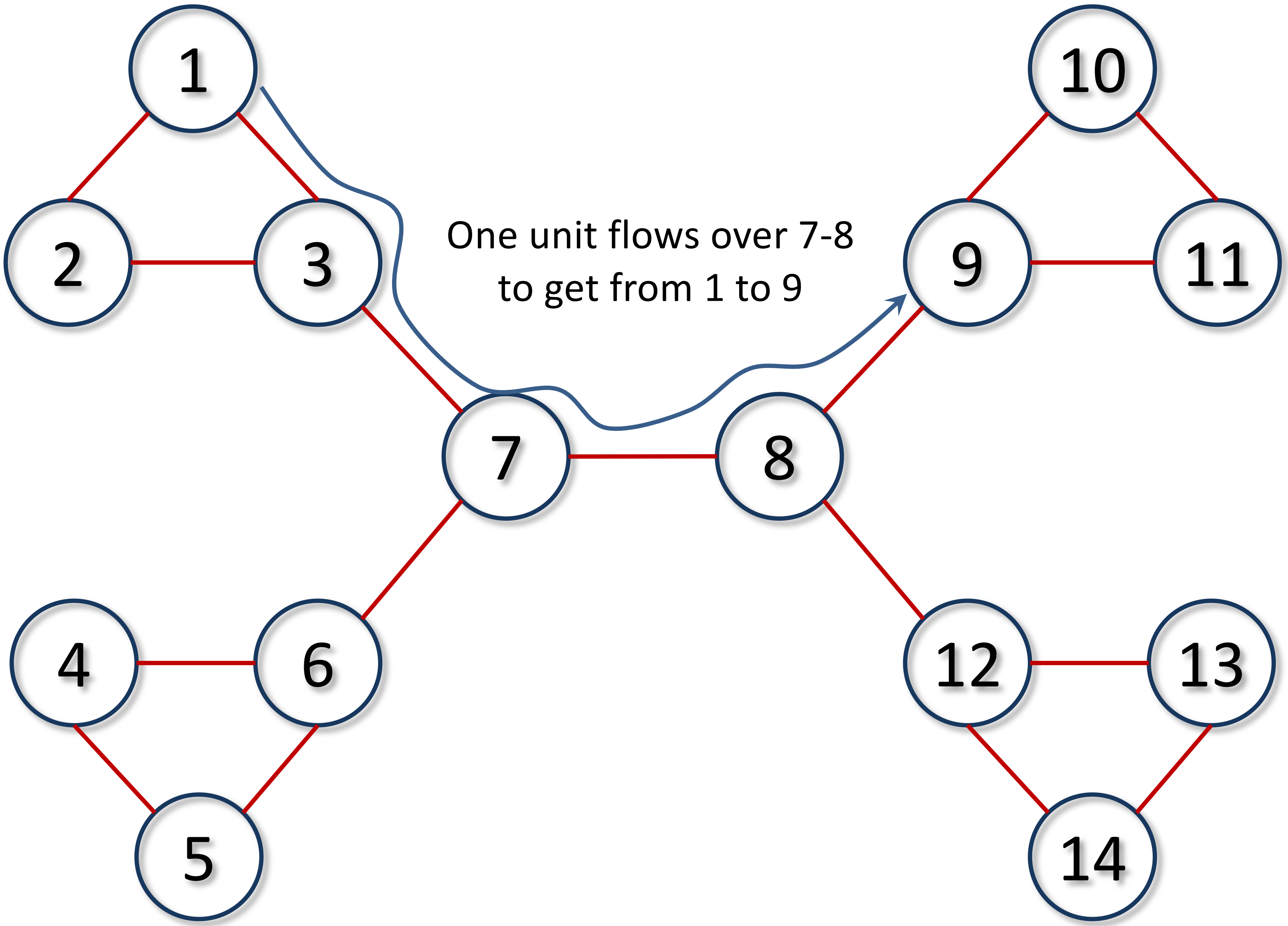

NB: these link betweenness values are based on a single shortest path between two nodes (which is not what the Girvan-Newman algorithm stipulates)

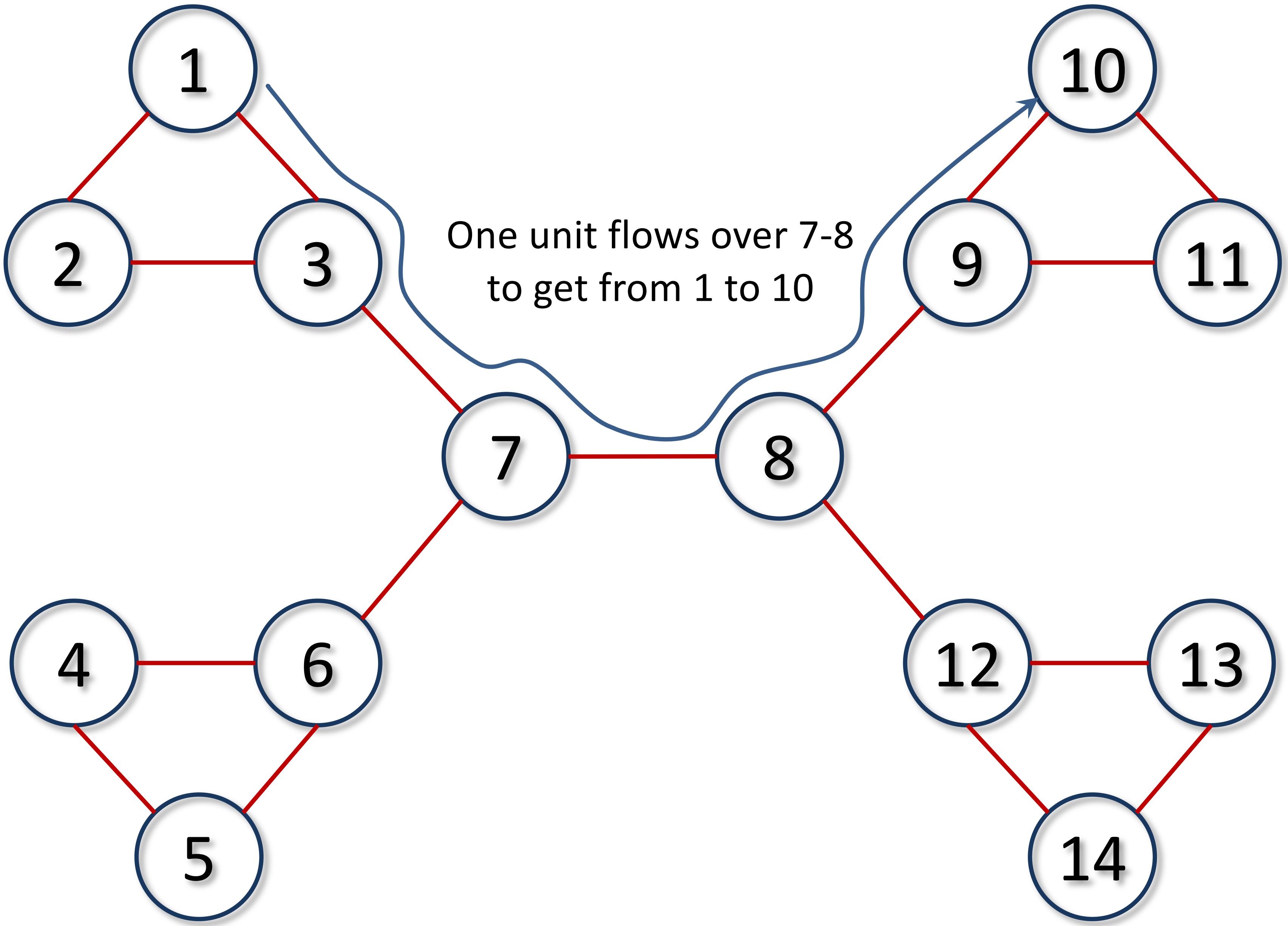
Link (Edge) Betweenness Example

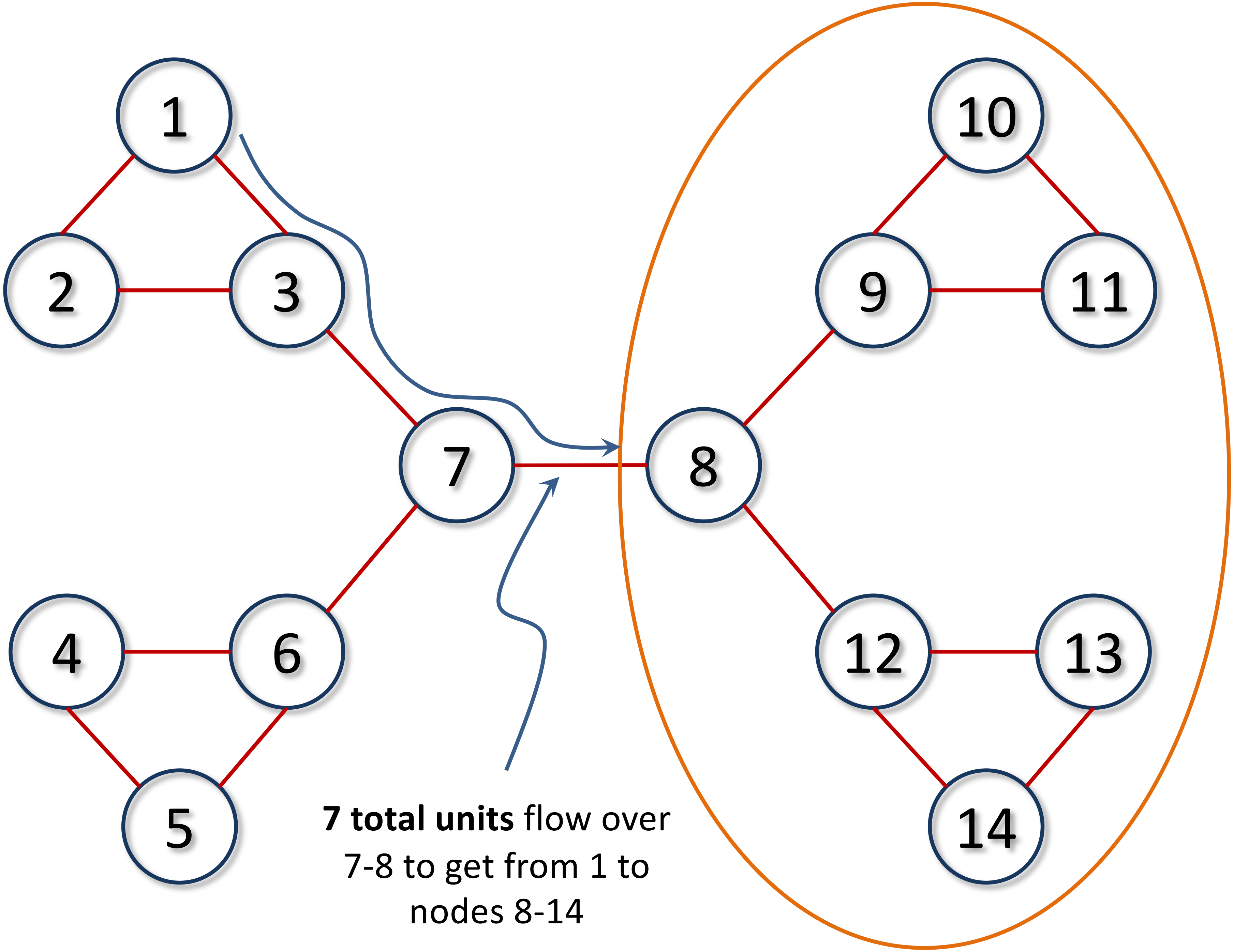



Credit: Frank McCown, Intro to Web Science, Harding University

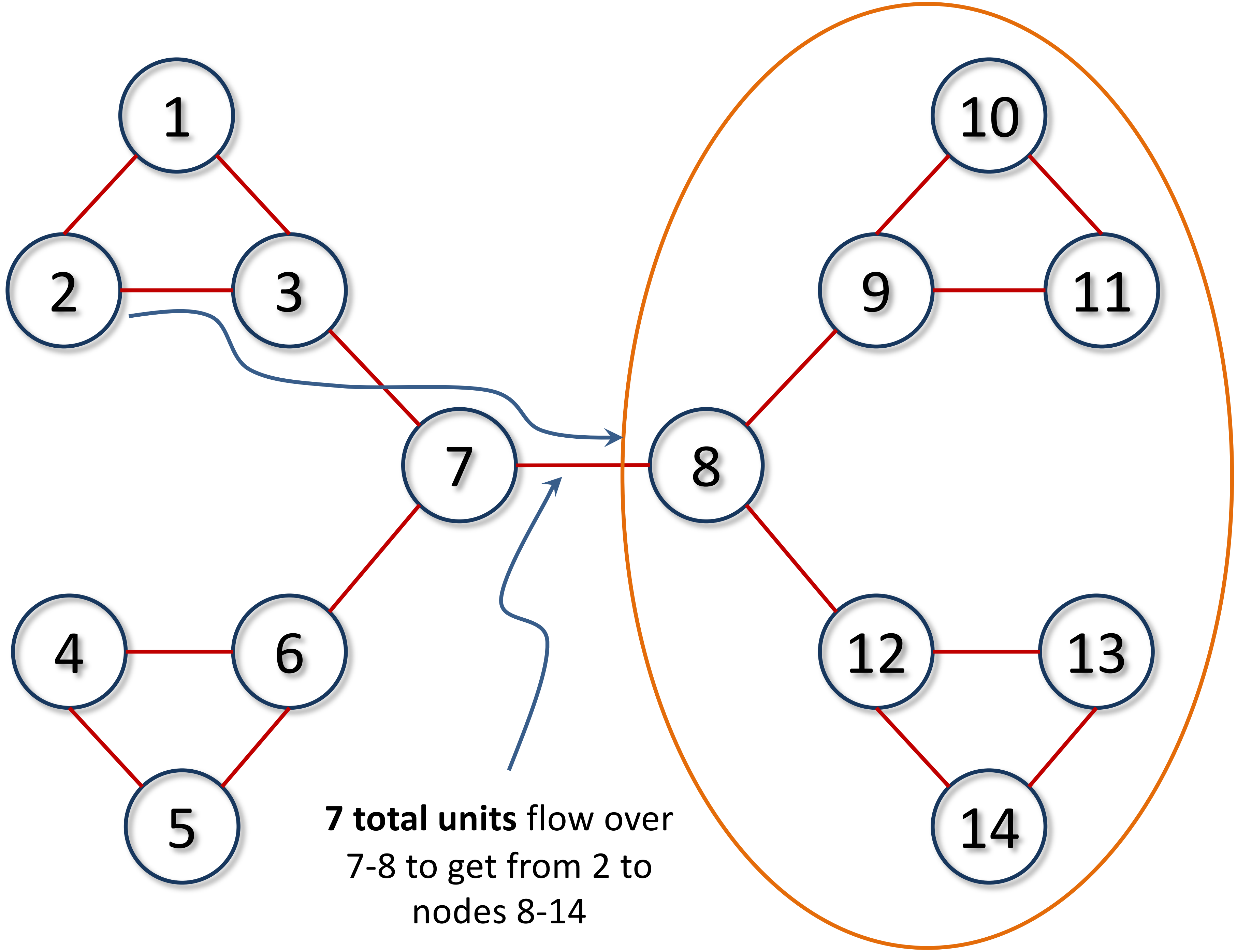

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8
to get from 1 to 8

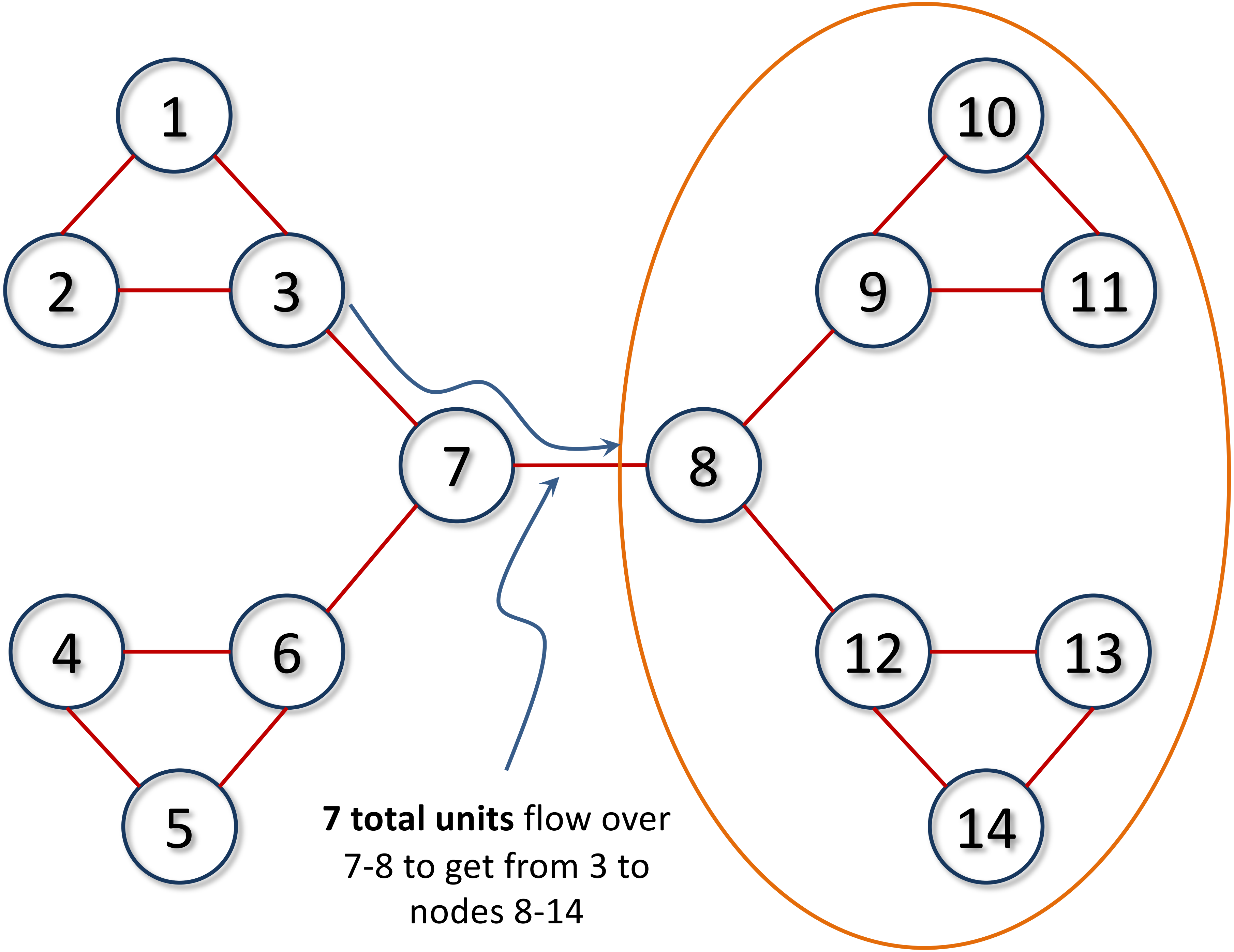

1
2
3
One unit flows over 7-8
to get from 1 to 9
10
9
11
7
8
4
6
12
13
5
14


1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8
to get from 1 to 10



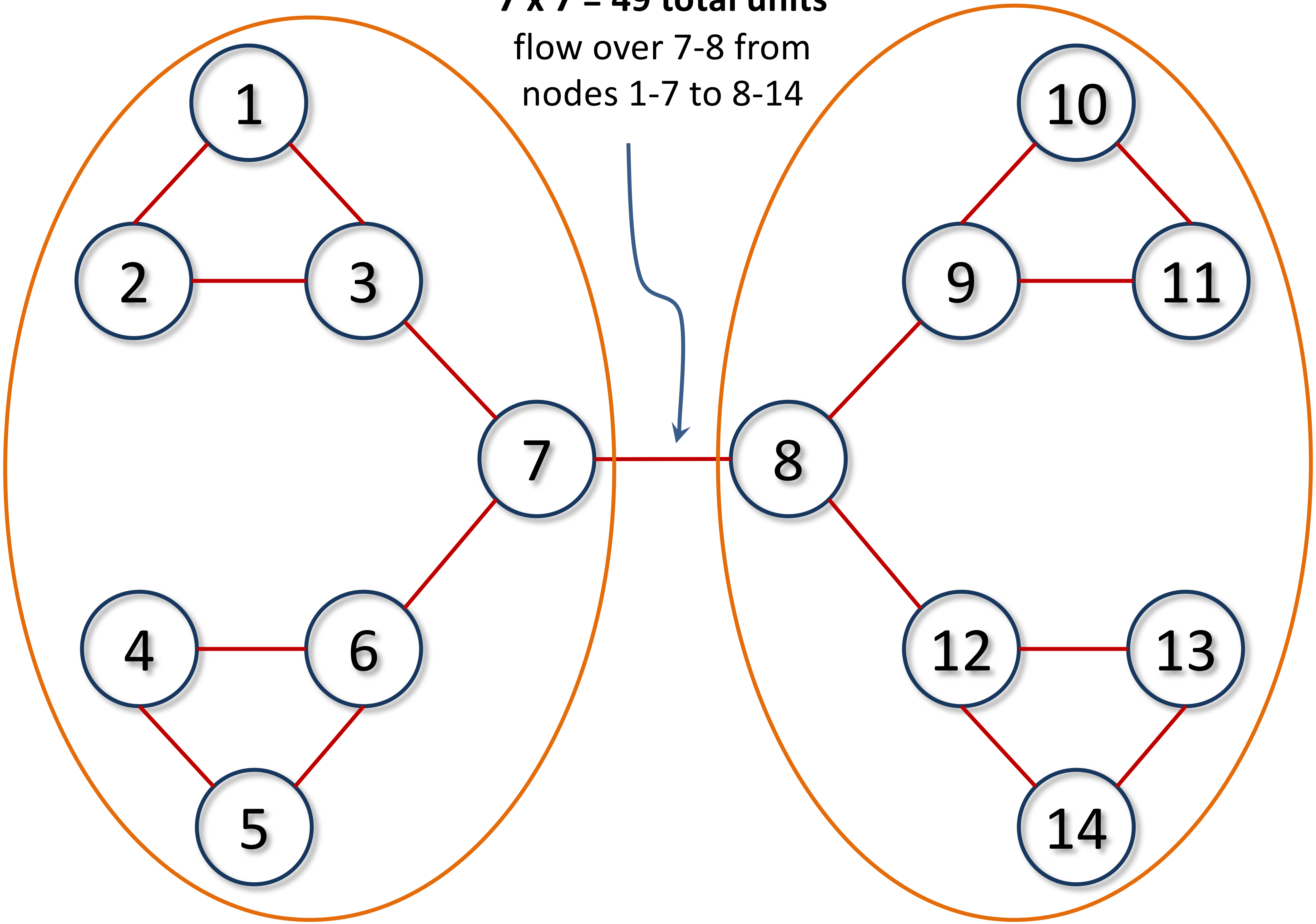

**7 total units** flow over 7-8 to get from 1 to nodes 8-14

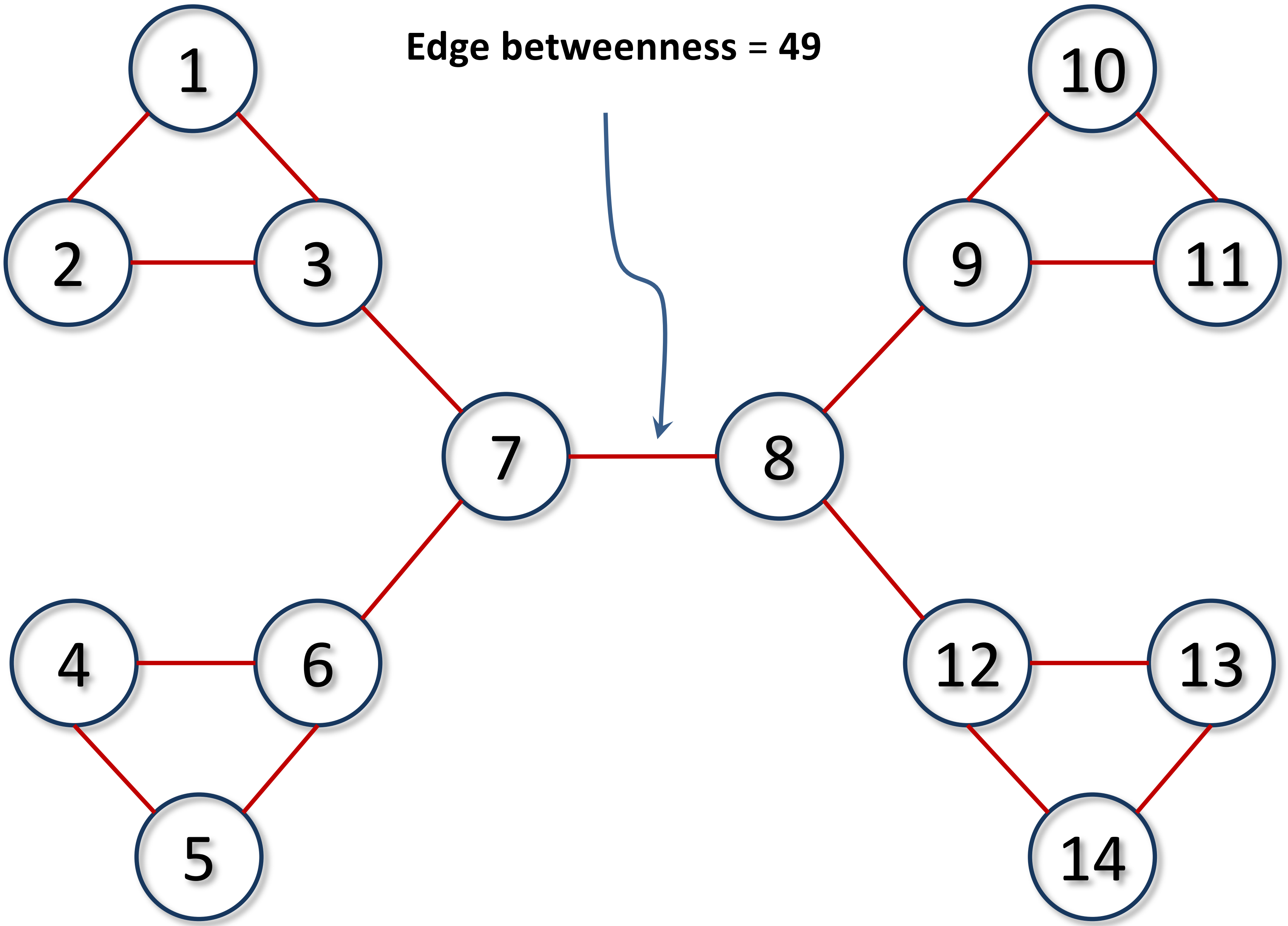

1
2
3
4
5
6
7
8
9
10
11
12
13
14
**7 total units** flow over 7-8 to get from 2 to nodes 8-14

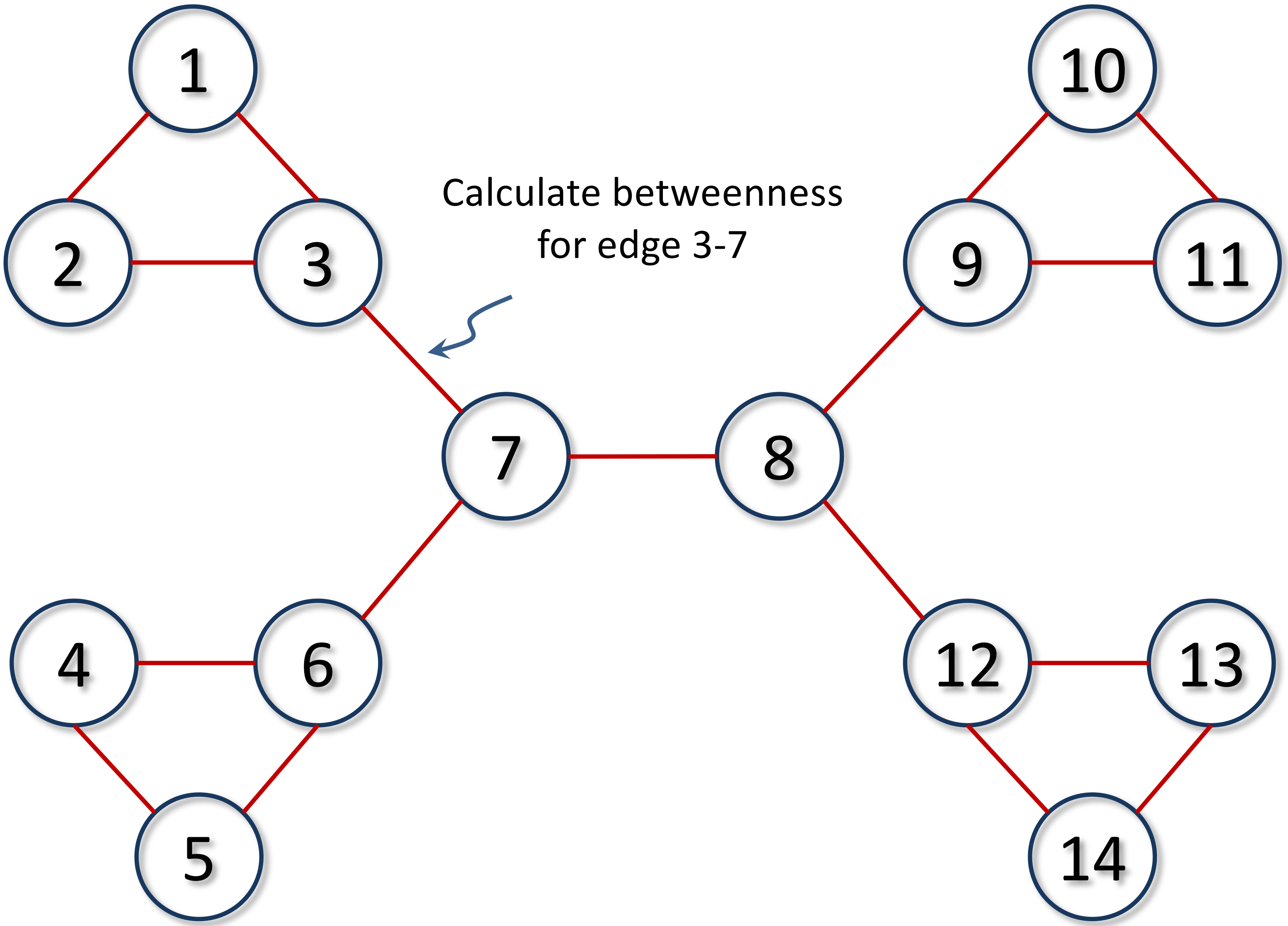


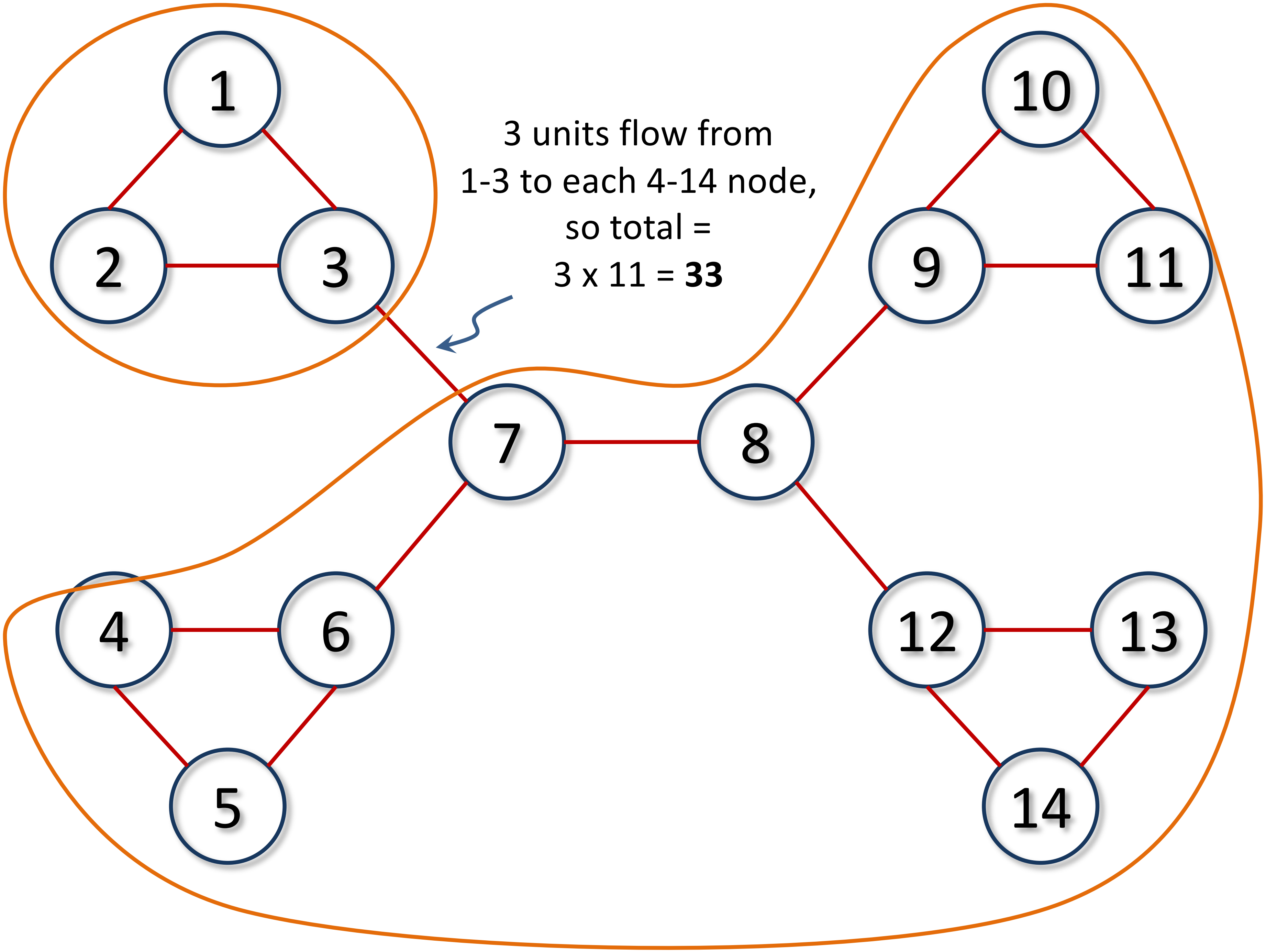

**7 total units** flow over 7-8 to get from 3 to nodes 8-14

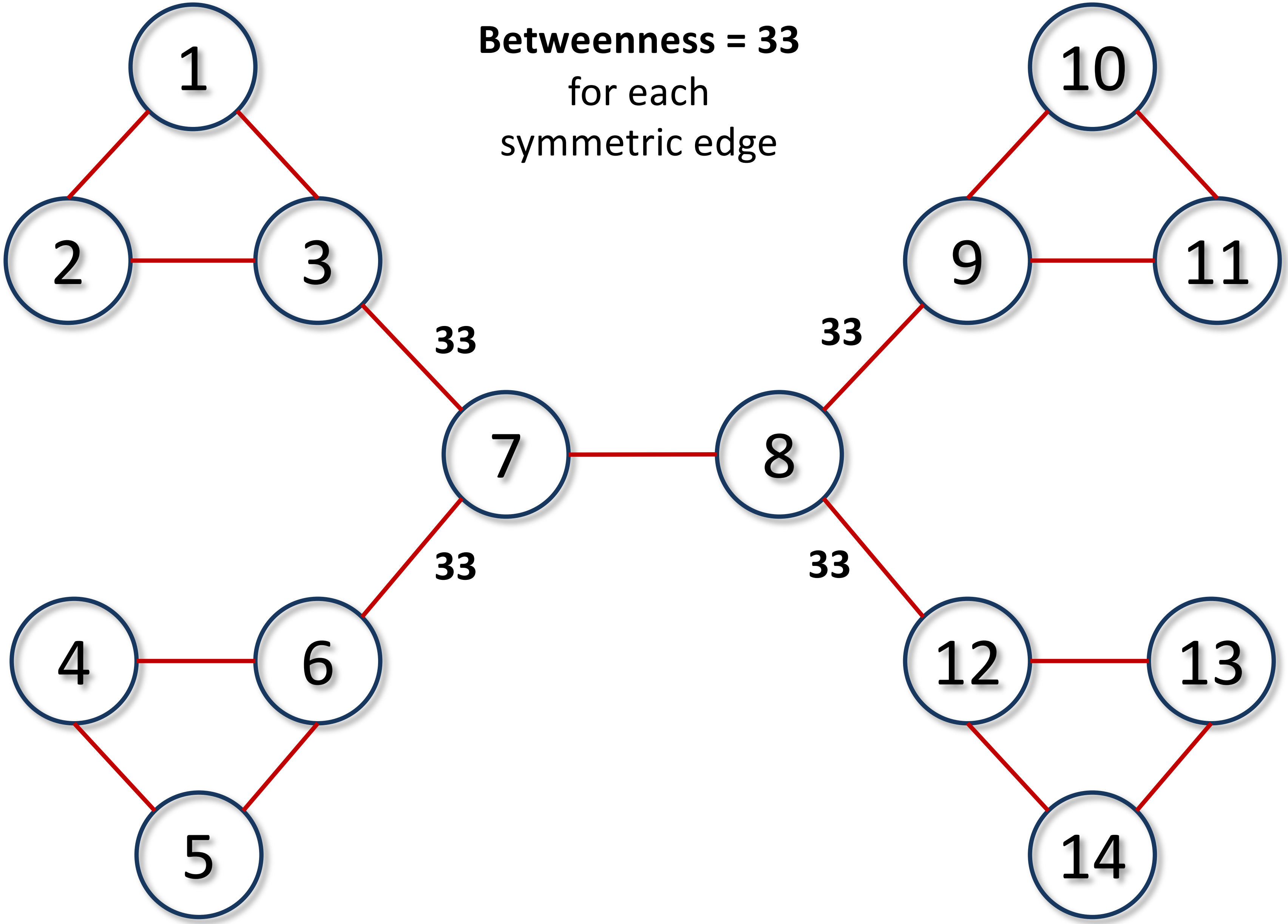
**7 x 7 = 49 total units**
flow over 7-8 from
nodes 1-7 to 8-14

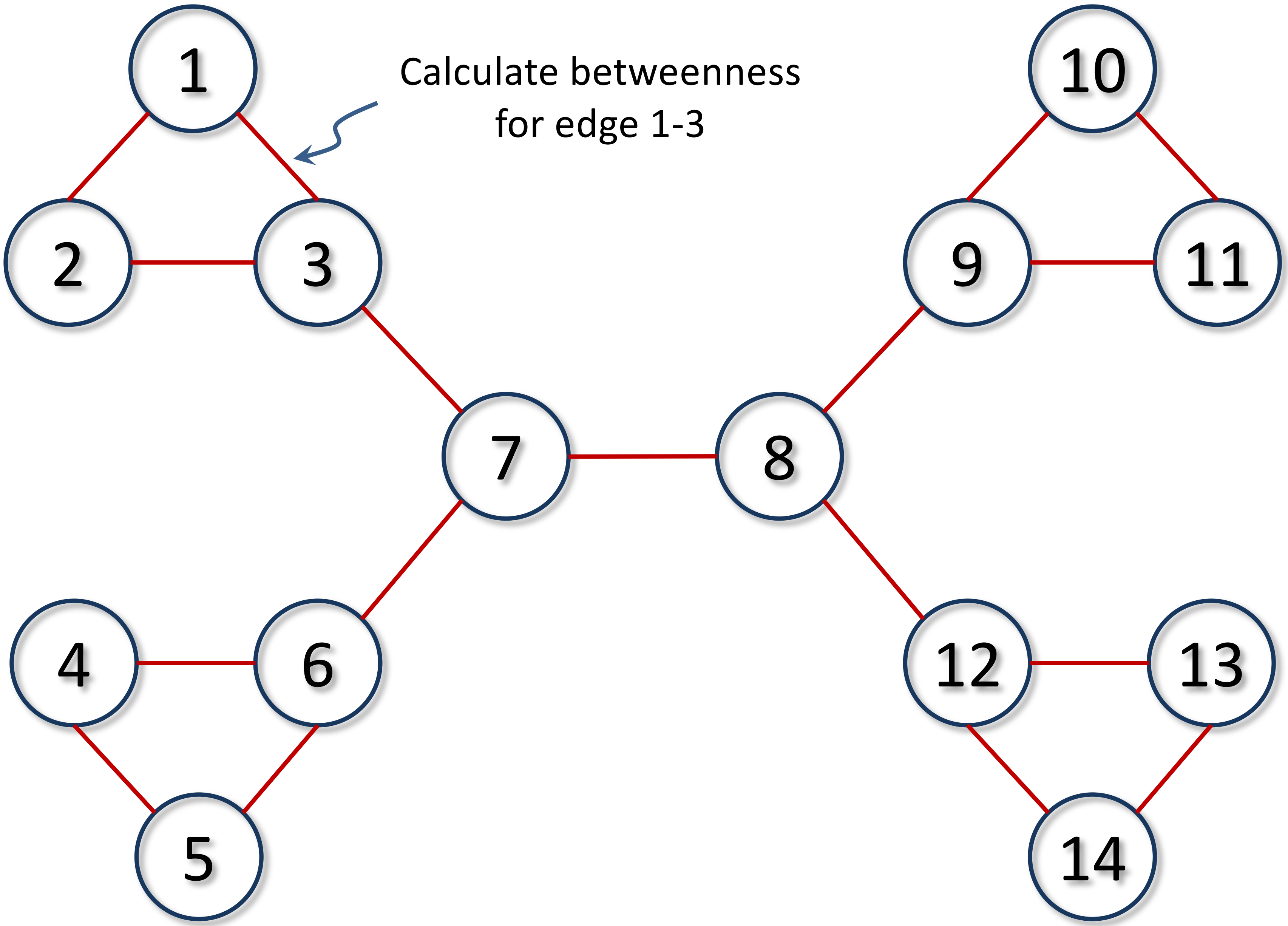


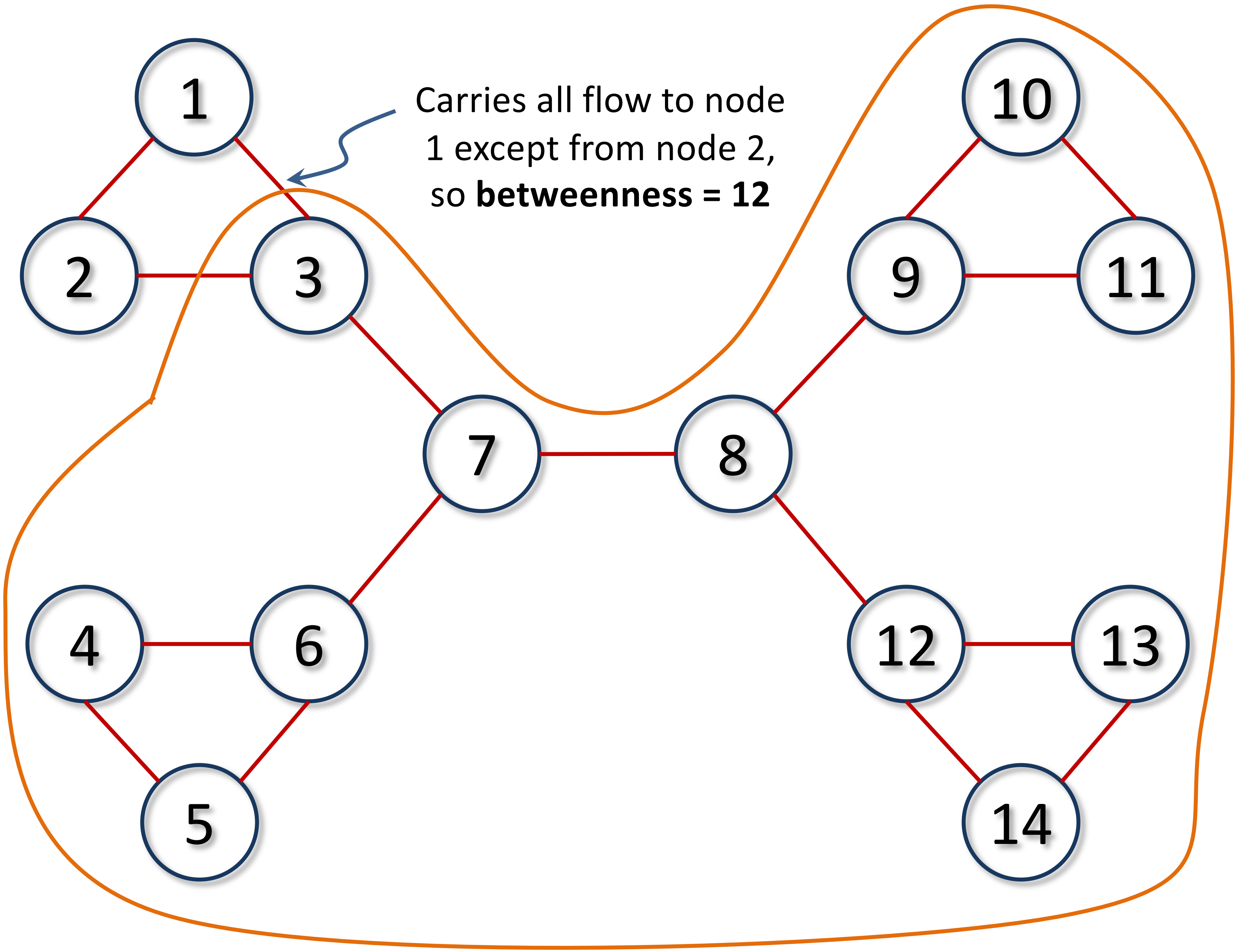

Edge betweenness = 49
1
2
3
7
8
9
10
11
4
6
5
12
13
14

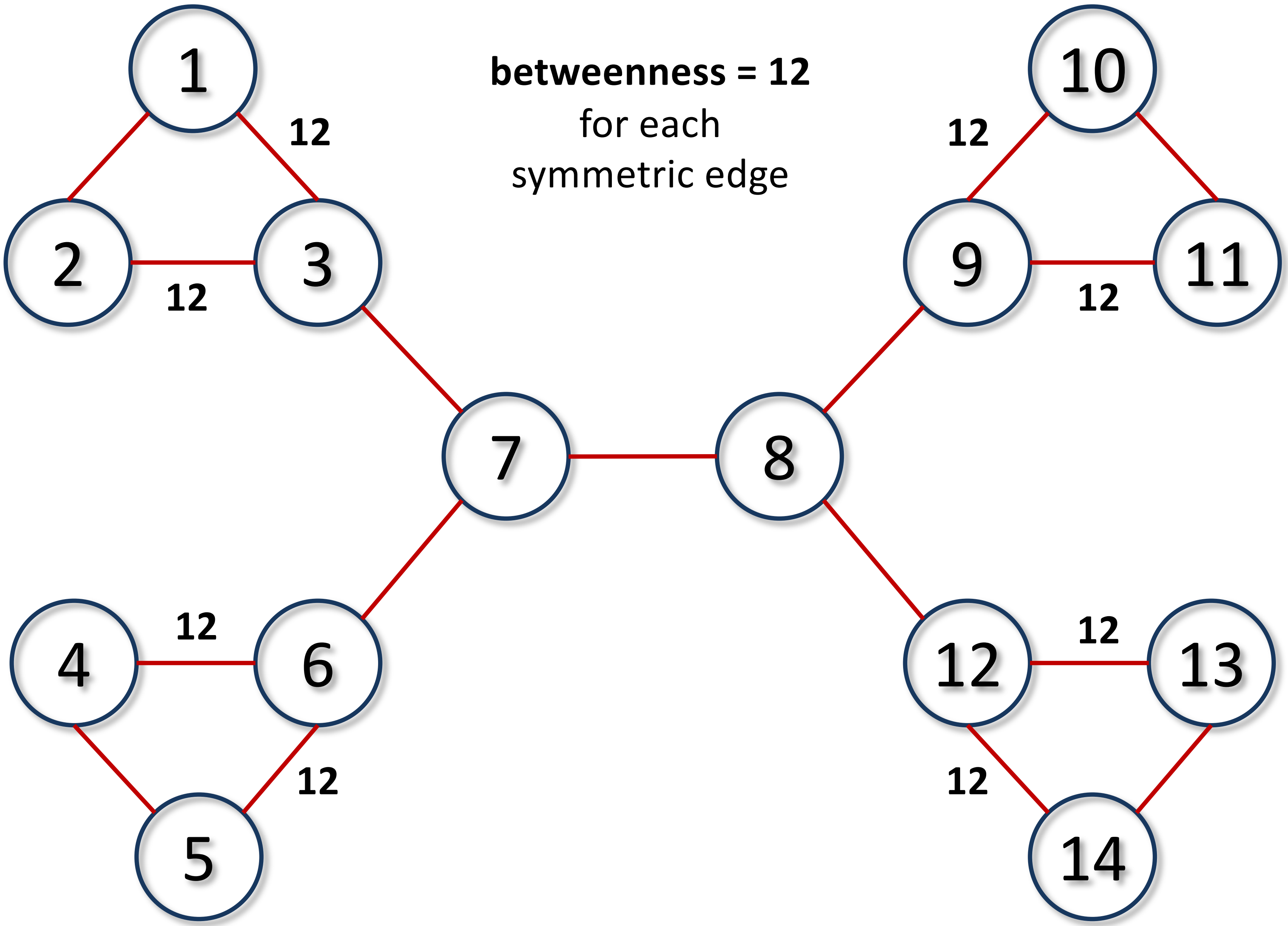

1
2
3
Calculate betweenness for edge 3-7
4
5
6
7
8
9
10
11
12
13
14


1
2
3
3 units flow from
1-3 to each 4-14 node,
so total =
3 x 11 = **33**
10
9
11
7
8
4
6
5
12
13
14

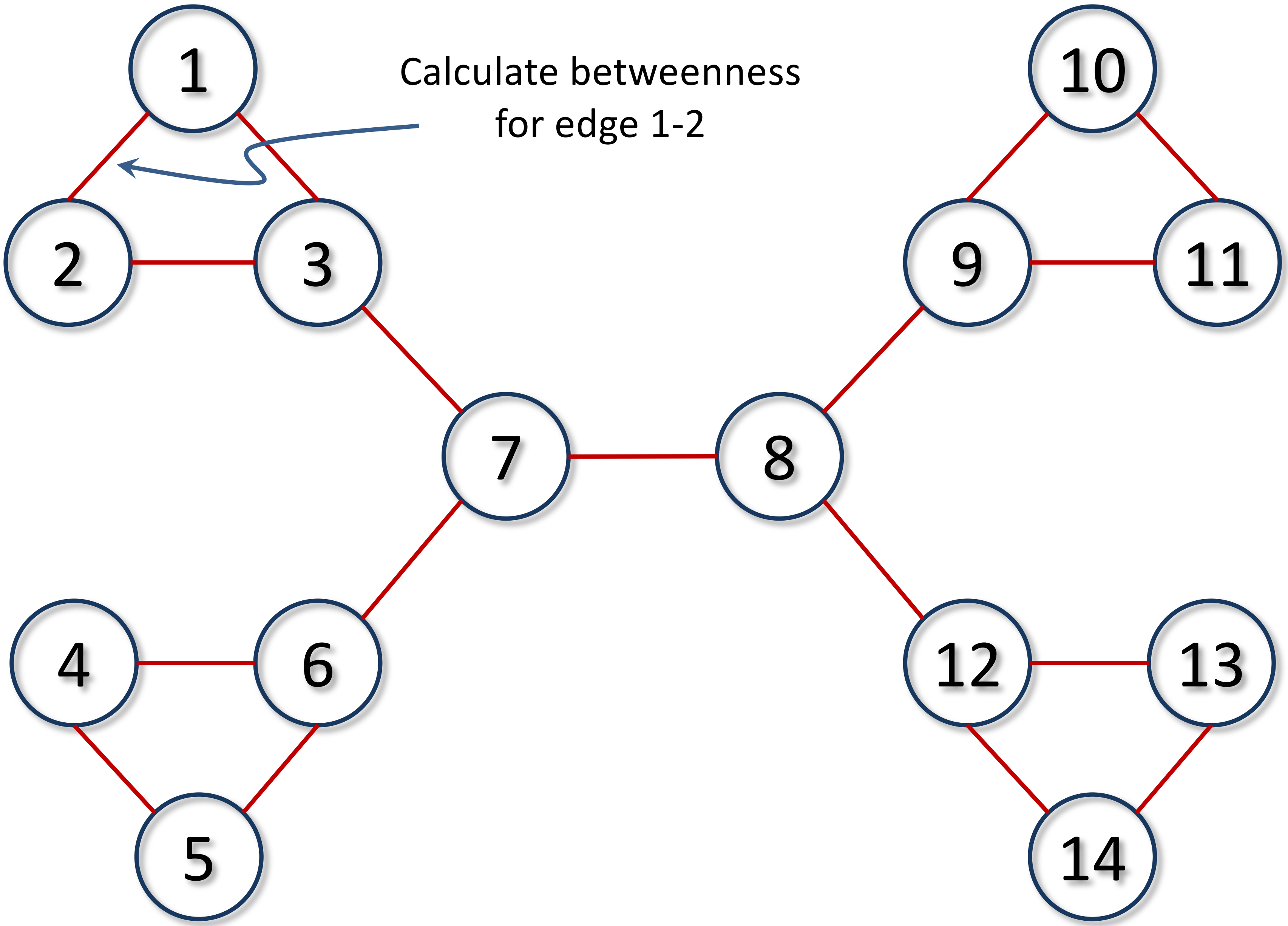

Betweenness = 33
for each
symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
33
33
33
33


1
2
3
4
5
6
7
8
9
10
11
12
13
14
Calculate betweenness
for edge 1-3

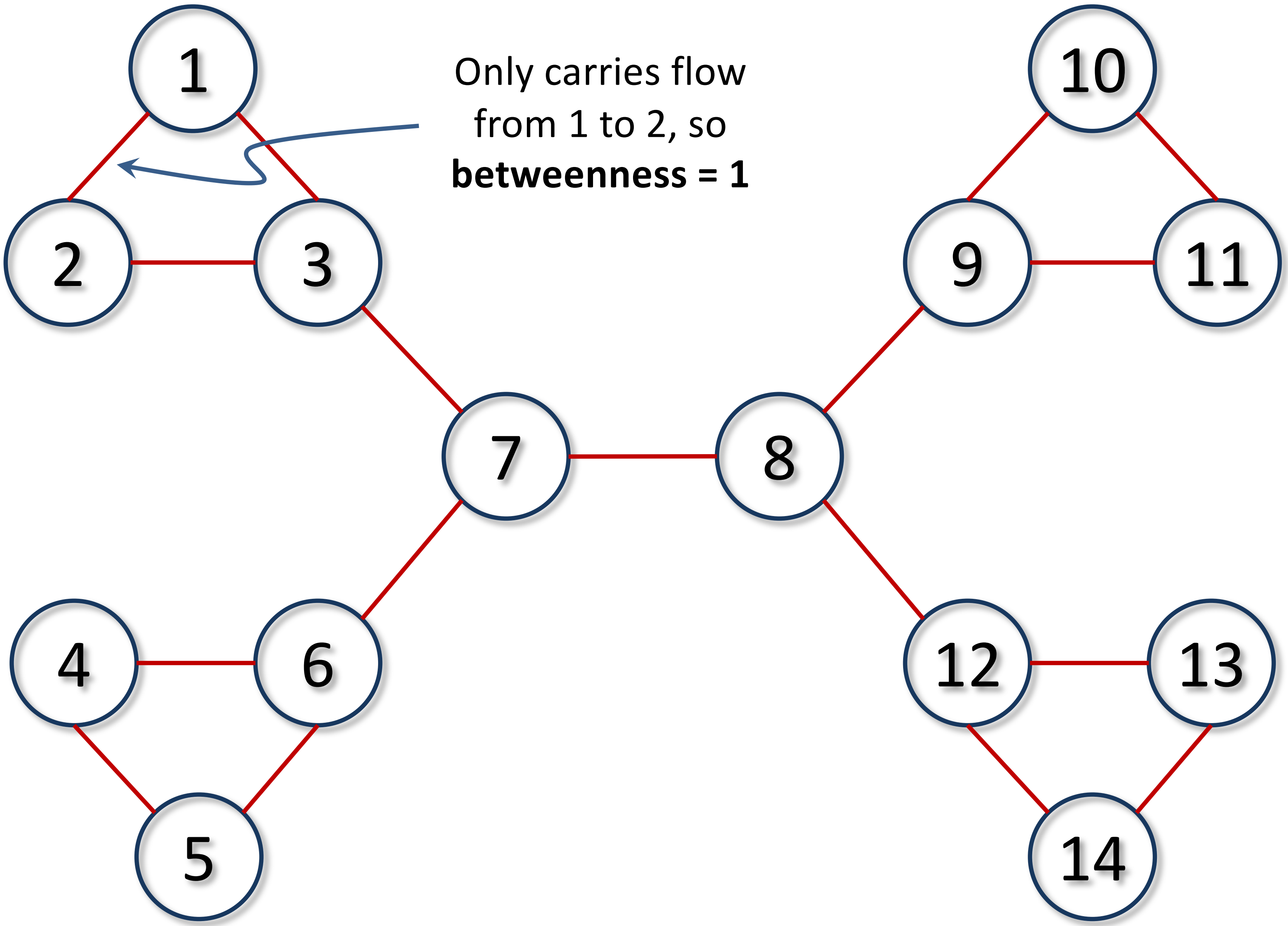

Carries all flow to node
1 except from node 2,
so **betweenness = 12**
1
2
3
4
5
6
7
8
9
10
11
12
13
14



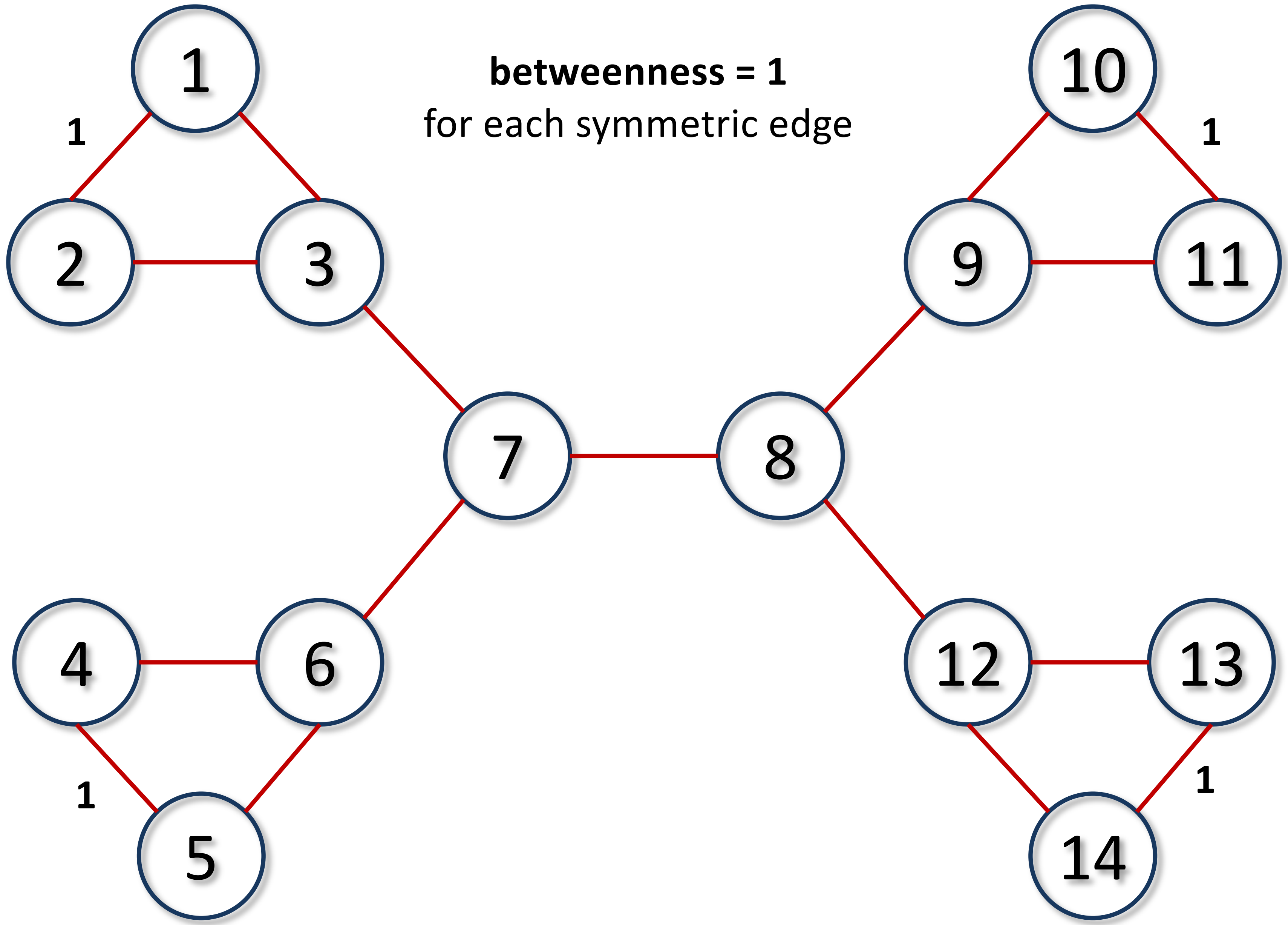

**betweenness = 12**
for each
symmetric edge



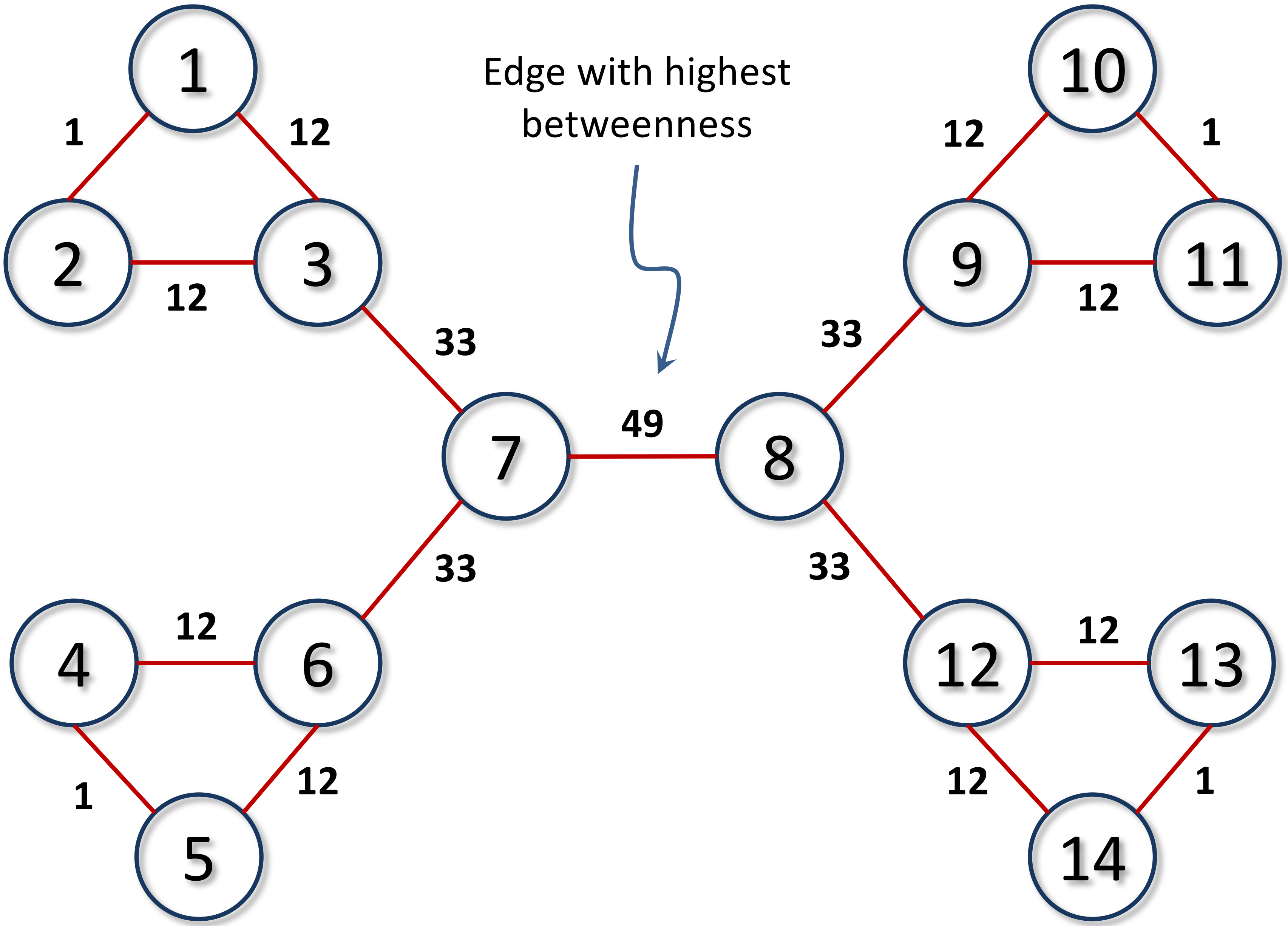

Calculate betweenness for edge 1-2




Only carries flow
from 1 to 2, so
**betweenness = 1**



betweenness = 1
for each symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1
1
1
1


Edge with highest betweenness
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1
12
12
33
49
33
12
1
12
33
33
12
1
12
12
1
12
12

# Complex Networks

## Community Detection

**Divisive Procedures: the Girvan-Newman Algorithm**

Step 2: Hierarchical Clustering

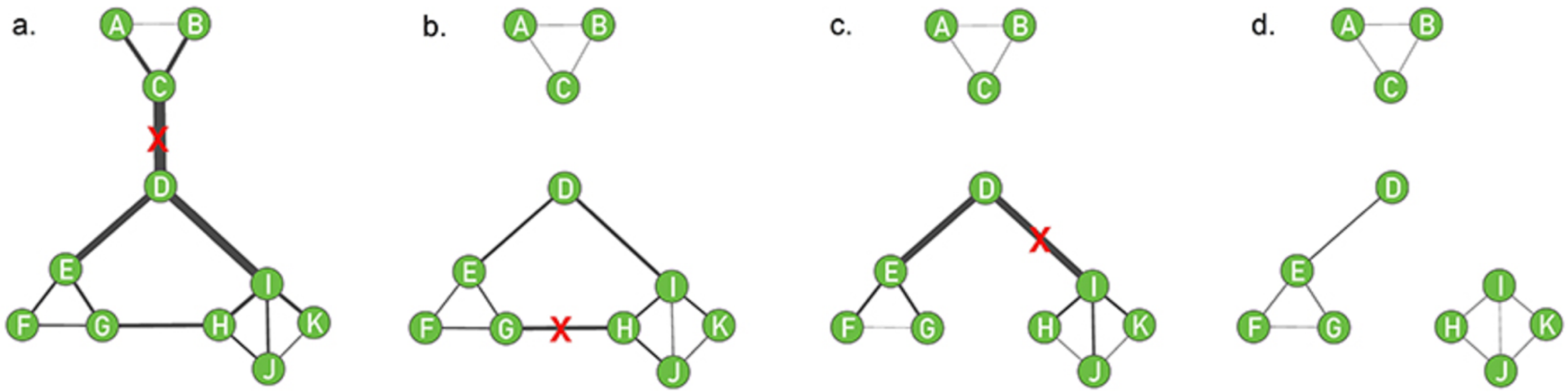
1. Compute the centrality $x_{ij}$ of each link

2. Remove the link with the largest centrality.
   In case of a tie, choose one link randomly

3. Recalculate the centrality of each link for the altered network

4. Repeat steps 2 and 3 until all links are removed

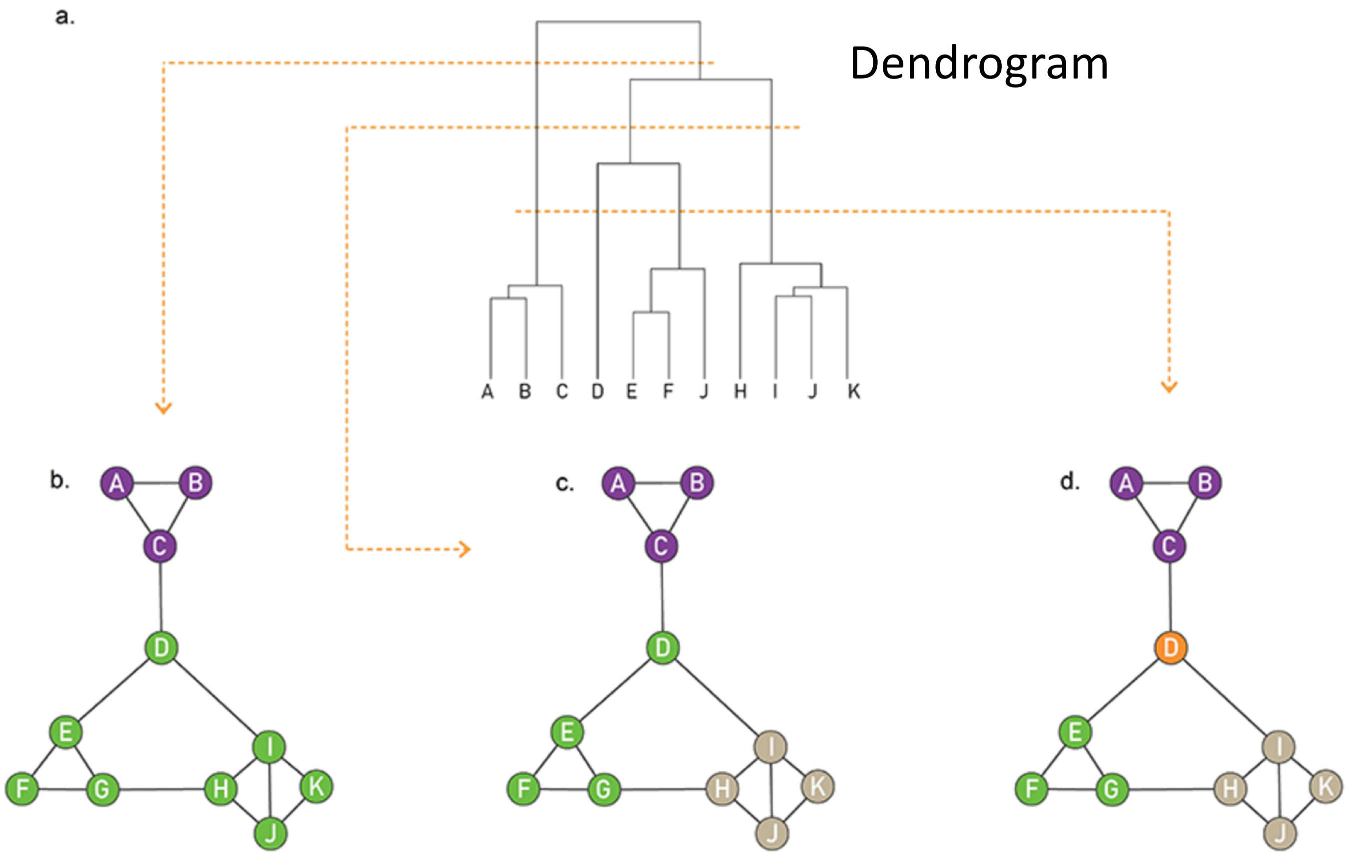
### Divisive Procedures: the Girvan-Newman Algorithm



# Complex Networks

## Community Detection

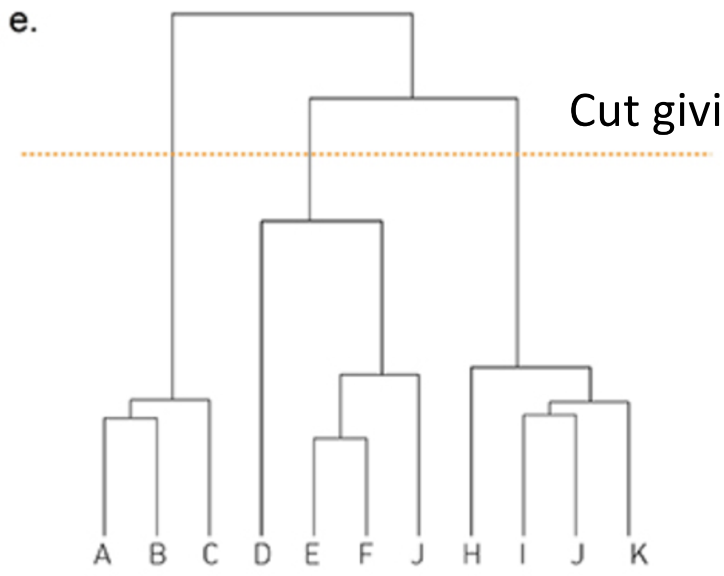
### Divisive Procedures: the Girvan-Newman Algorithm



# Complex Networks

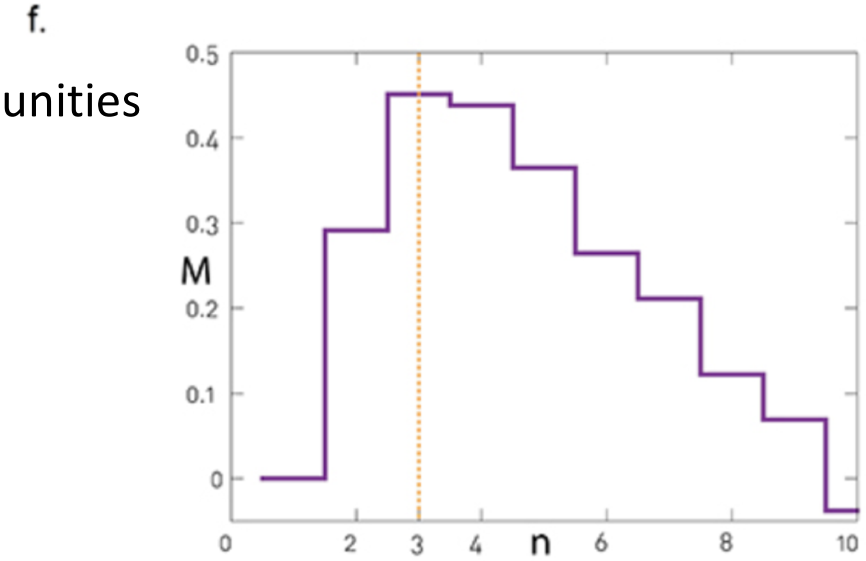
## Community Detection

### Divisive Procedures: the Girvan-Newman Algorithm




Cut giving 3 communities

Dendrogram




Cut is determined using a Modularity measure M

# Complex Networks

## Community Detection

**Divisive Procedures: the Girvan-Newman Algorithm**

Computational complexity depends on the centrality metric

For link betweenness: $O(LN)$

Including Modularity: $O(L^2N)$
$O(N^3)$ for sparse graph

# Complex Networks

## Community Detection

### Divisive Procedures: the Girvan-Newman Algorithm

The Girvan-Newman algorithm predicted communities in Zachary's Karate Club that the matched almost perfectly two groups after the break-up.