APPLIED DATA SCIENCE

fall 2017

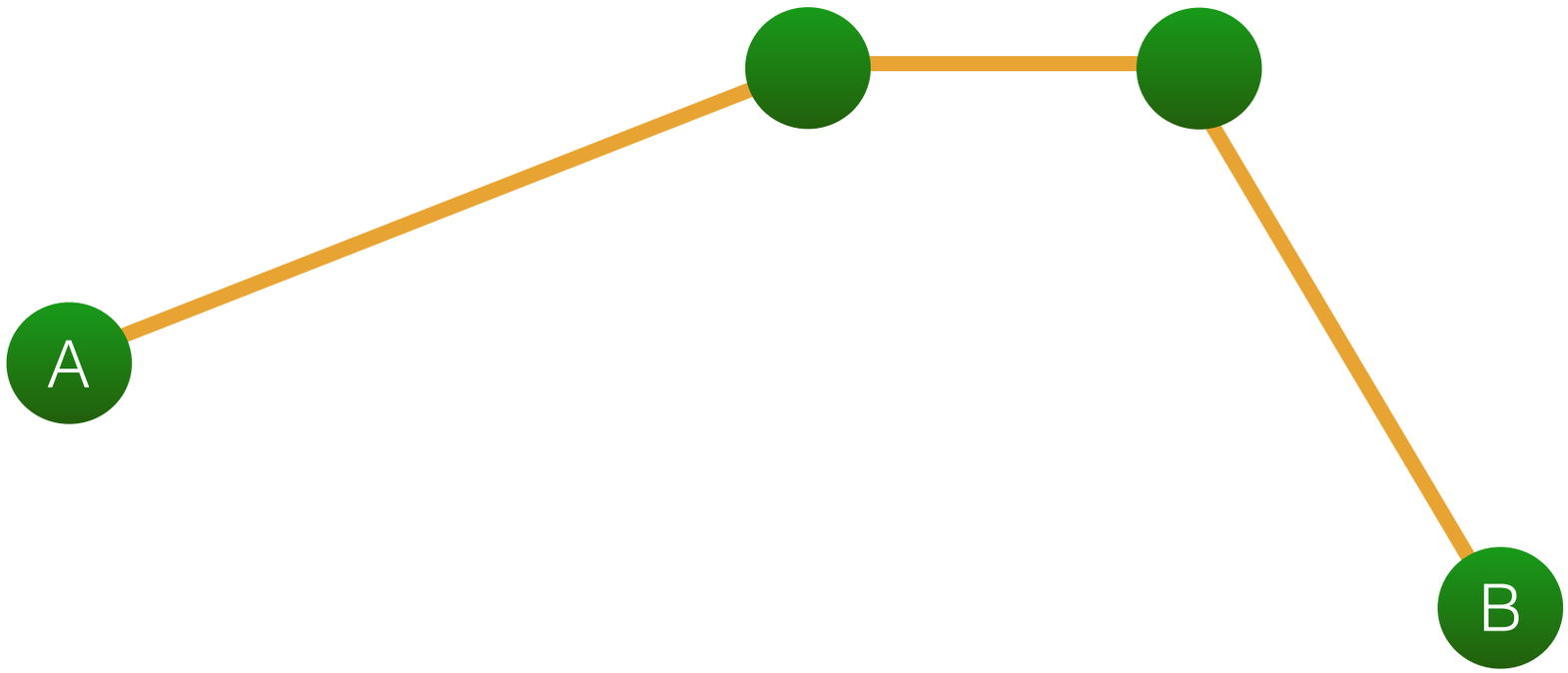
Session 11: Connectivity. Routing in networks. Community detection

Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, Maxim Temnogorod

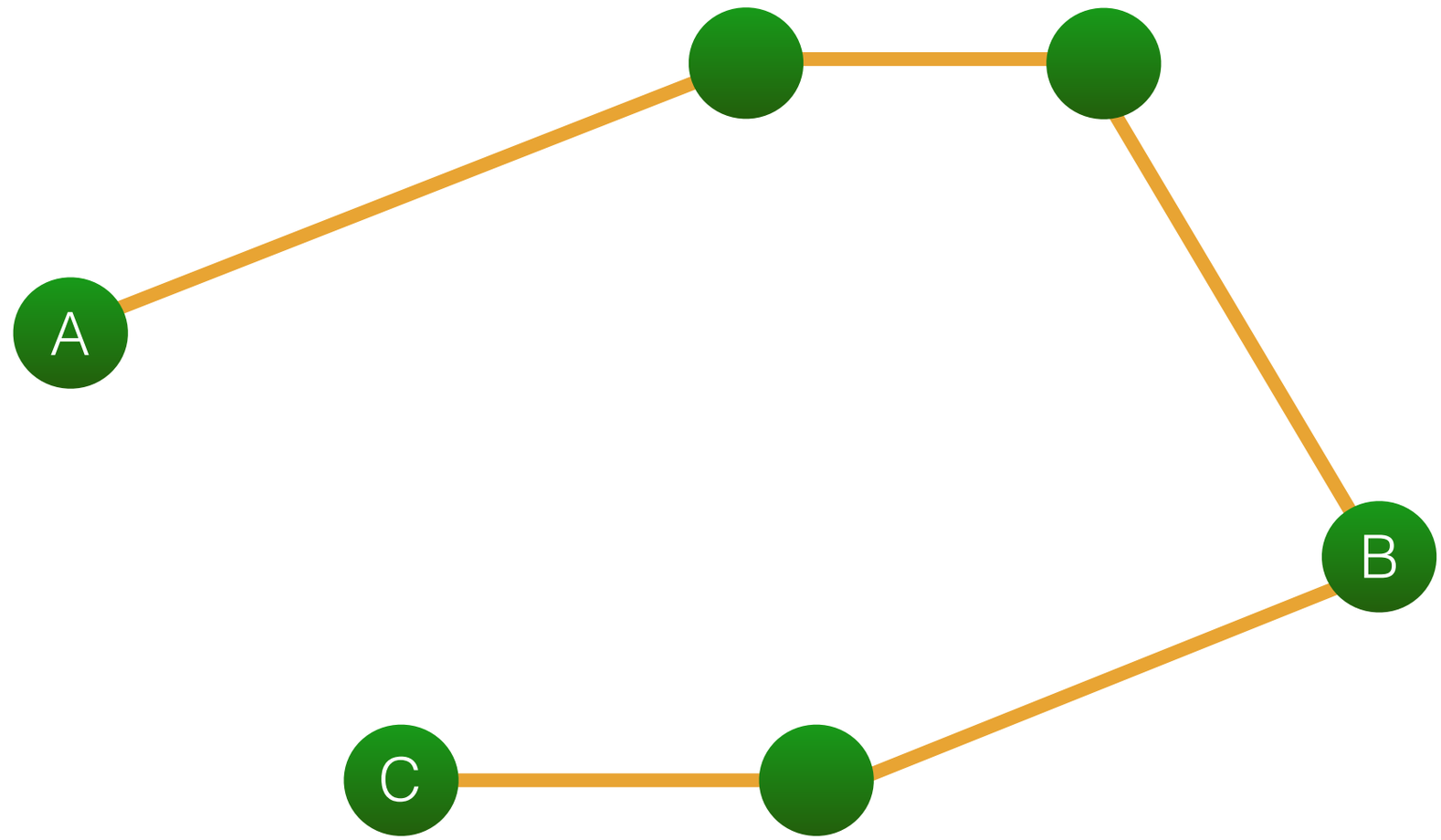
Network routing applications

- Vehicle routing
- Reaching to a person in a social network
- Engineering: Very-large-scale integration (designing integrated circuits to combine transistors into a chip)
- Operations and control: fastest way of reaching target state of the system
- Robotics

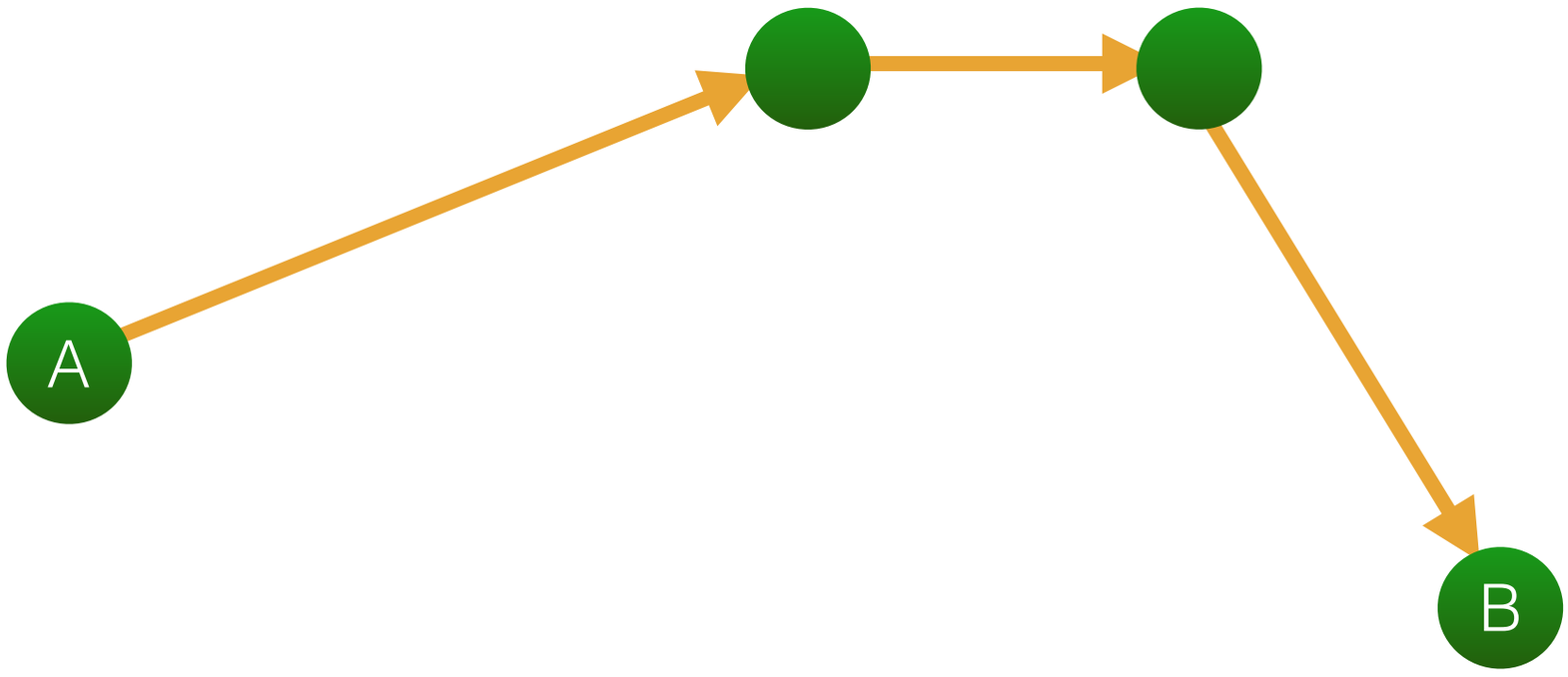
Connectivity



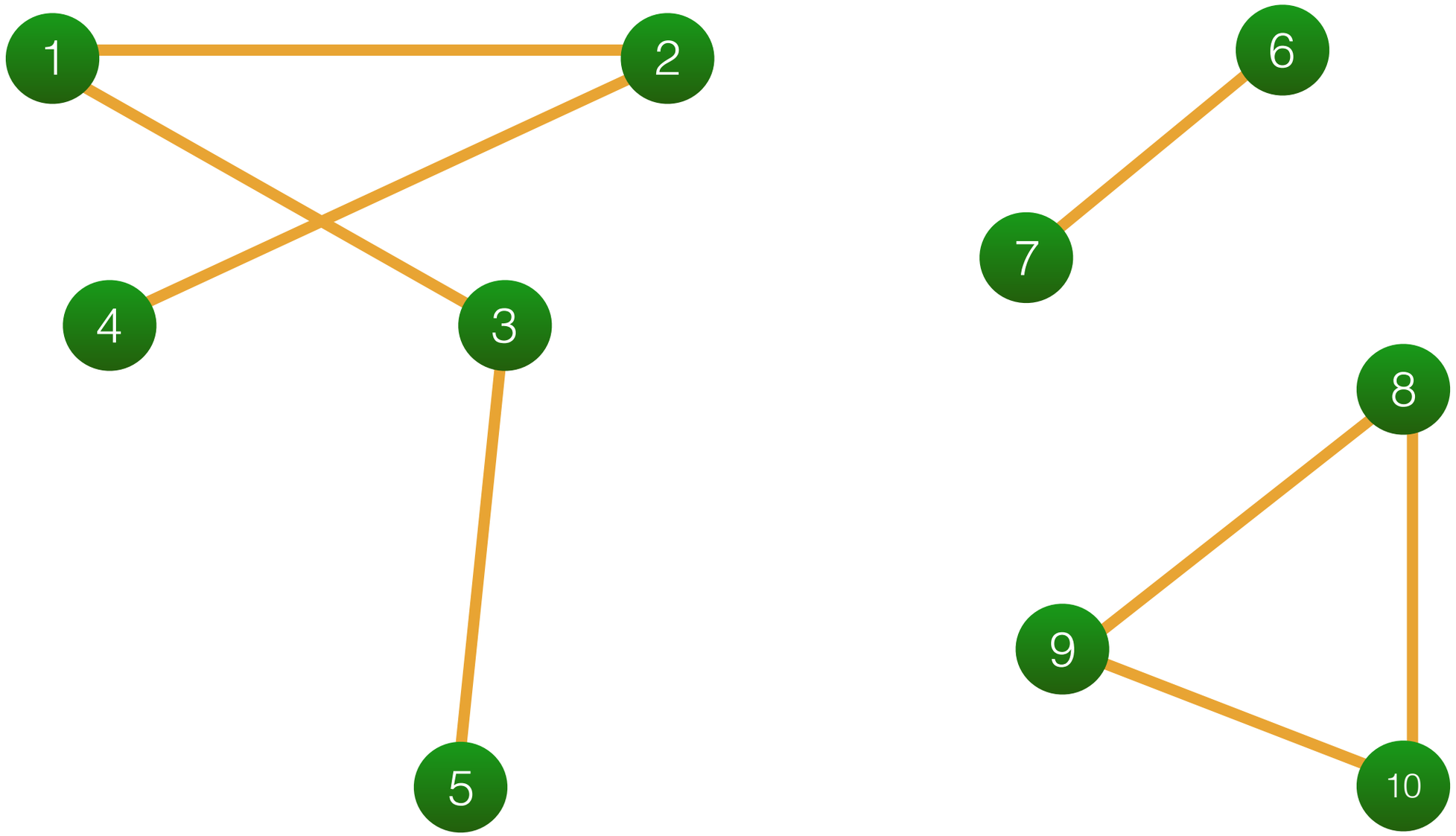
Transitivity of connectivity



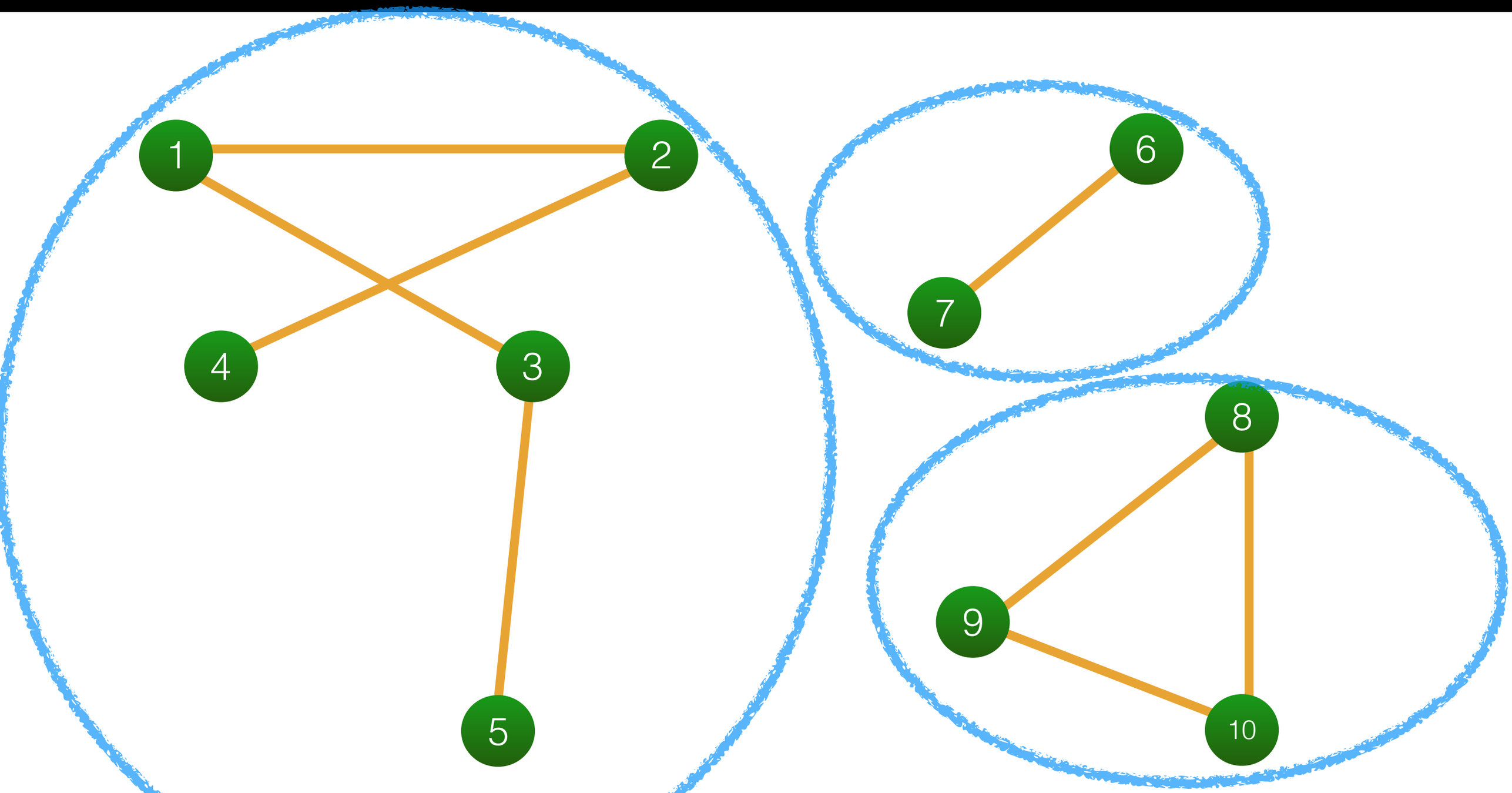
Connectivity - directed case



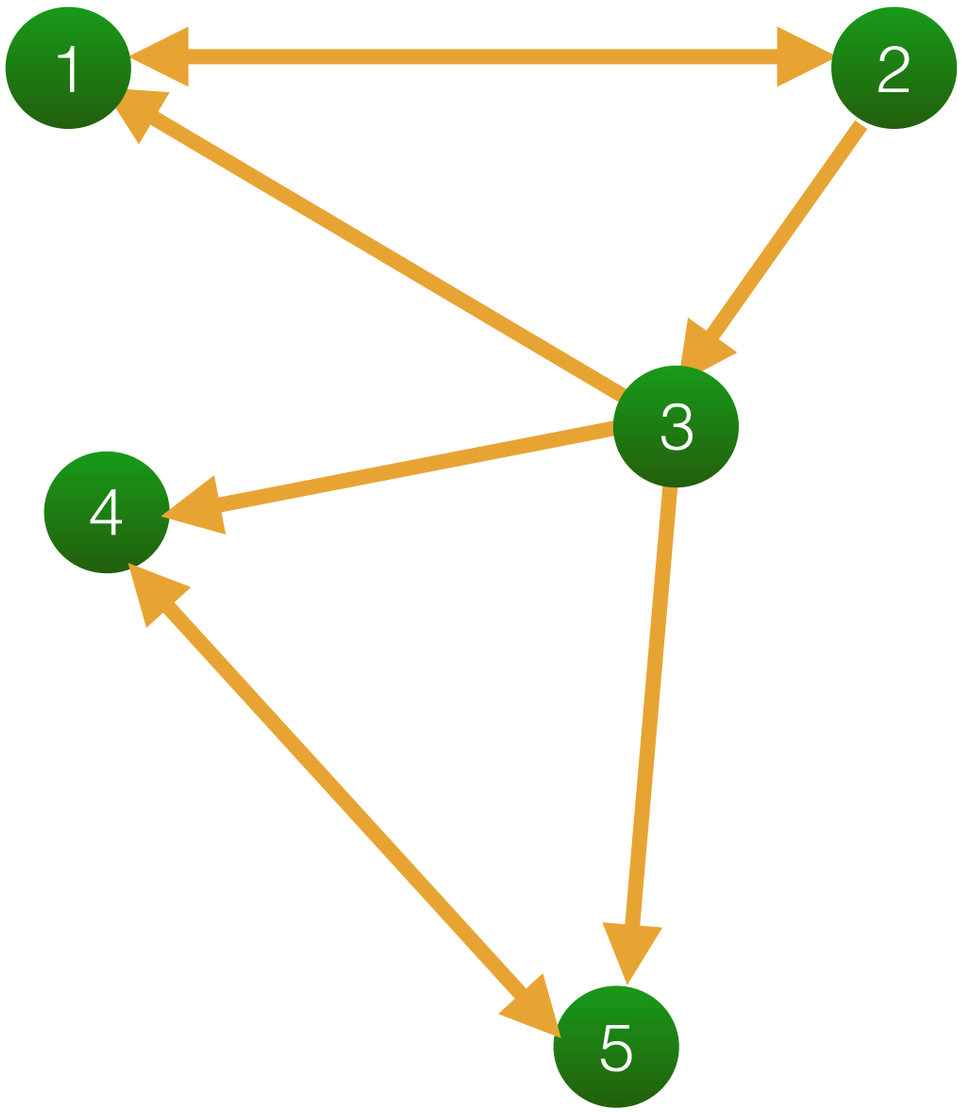
Connected components



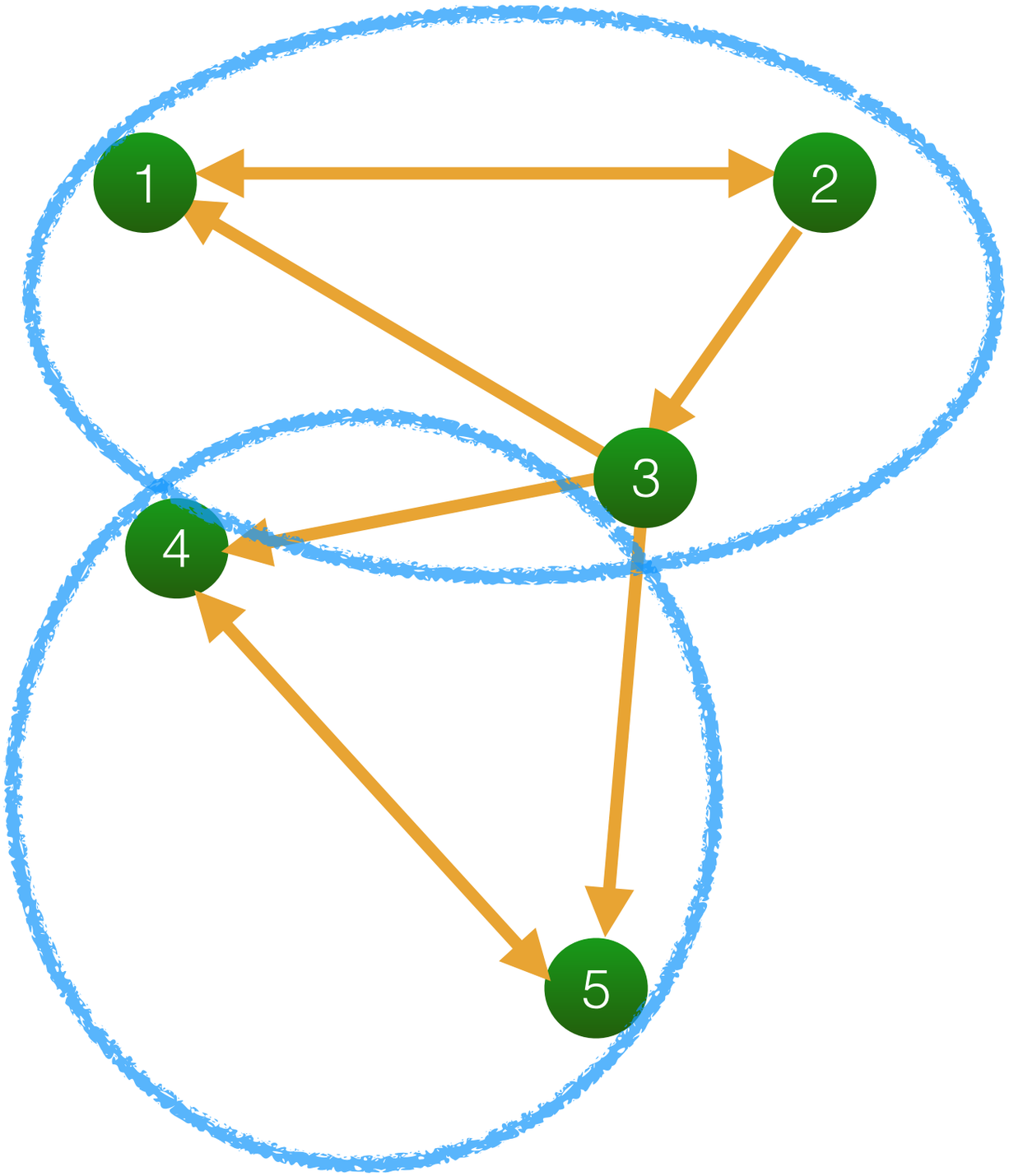
Connected components



Directed networks: strongly and weakly connected components

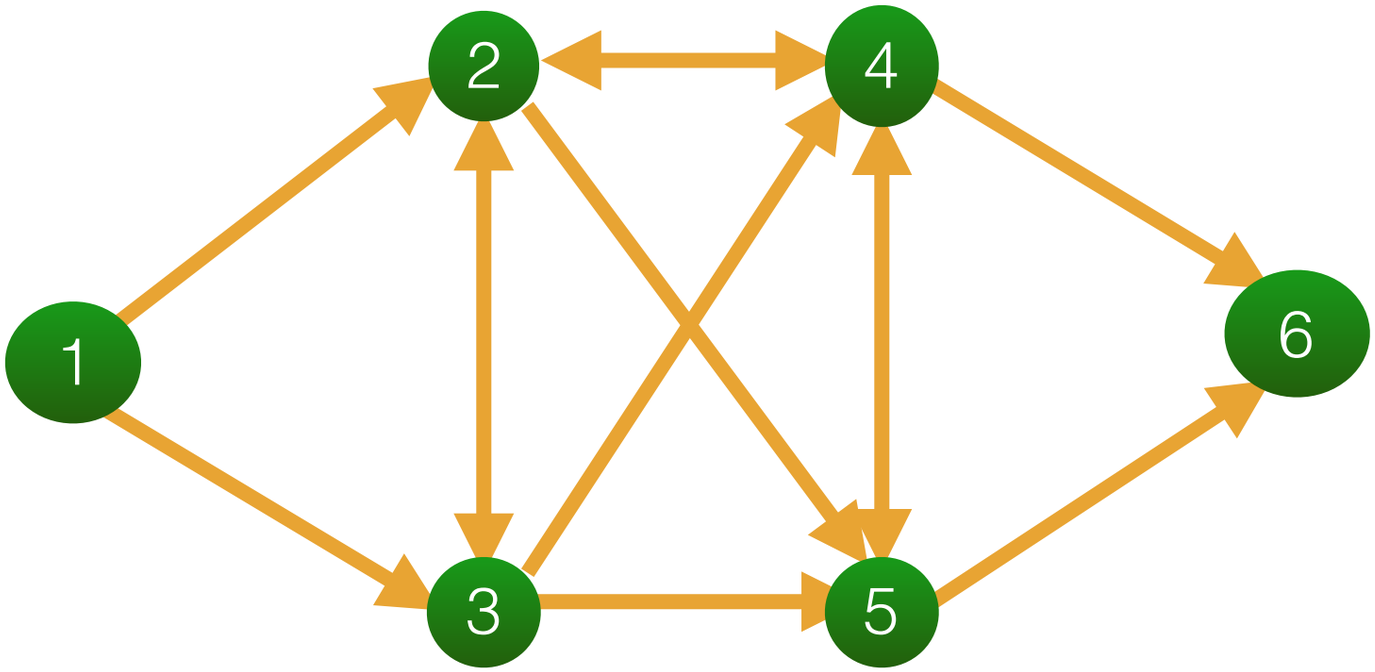


Directed networks: strongly and weakly connected components



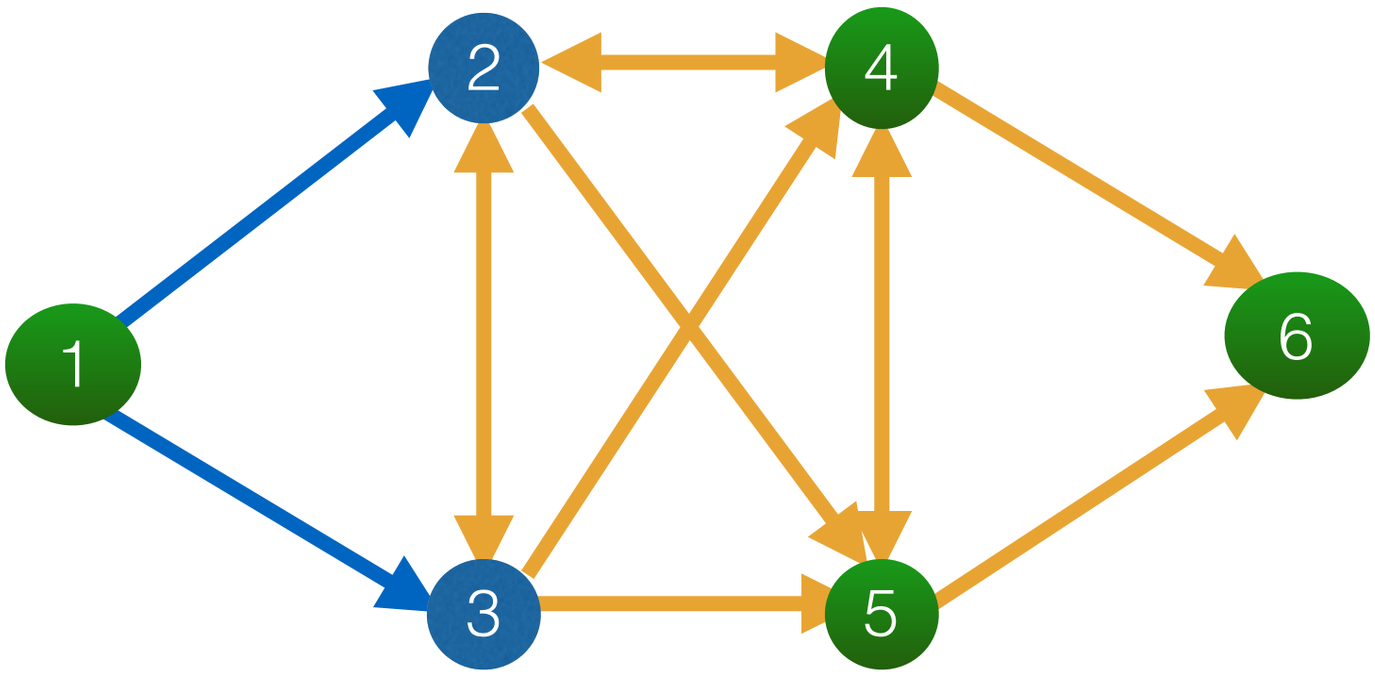
Shortest path routing

Path from 1 to 6?



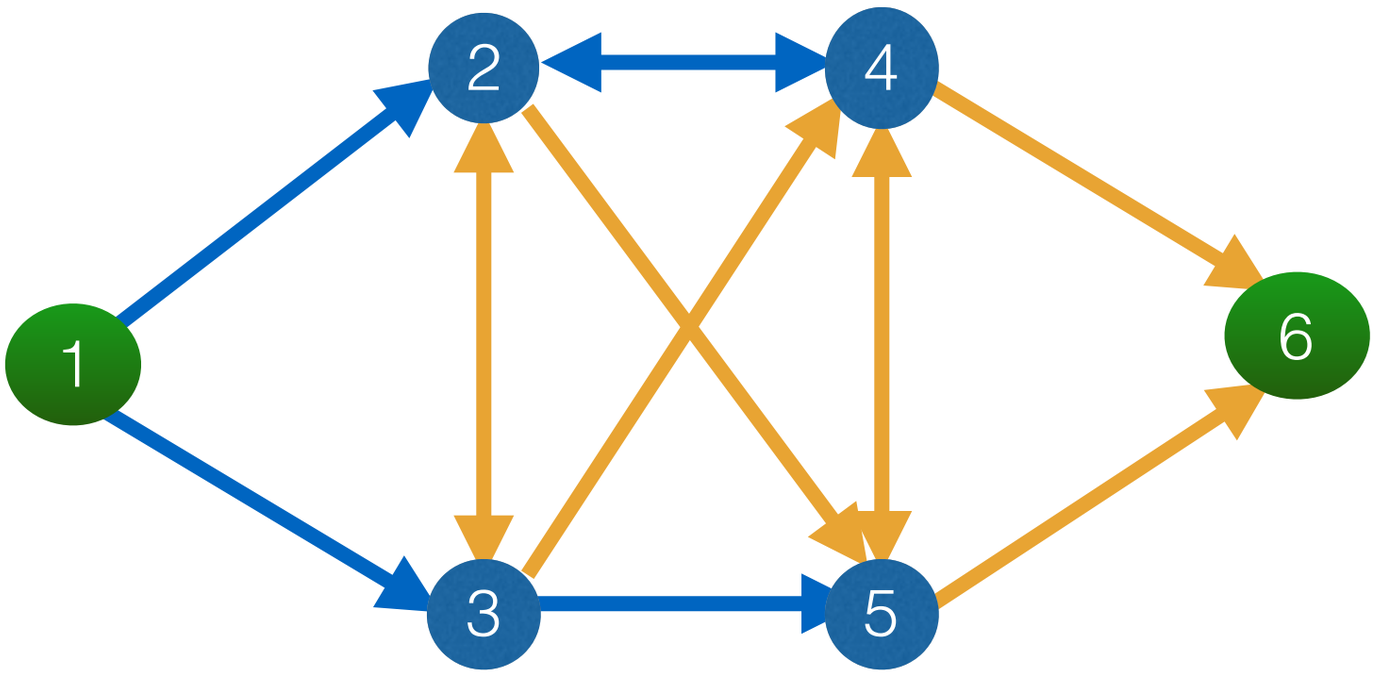
Shortest path routing

Path from 1 to 6?



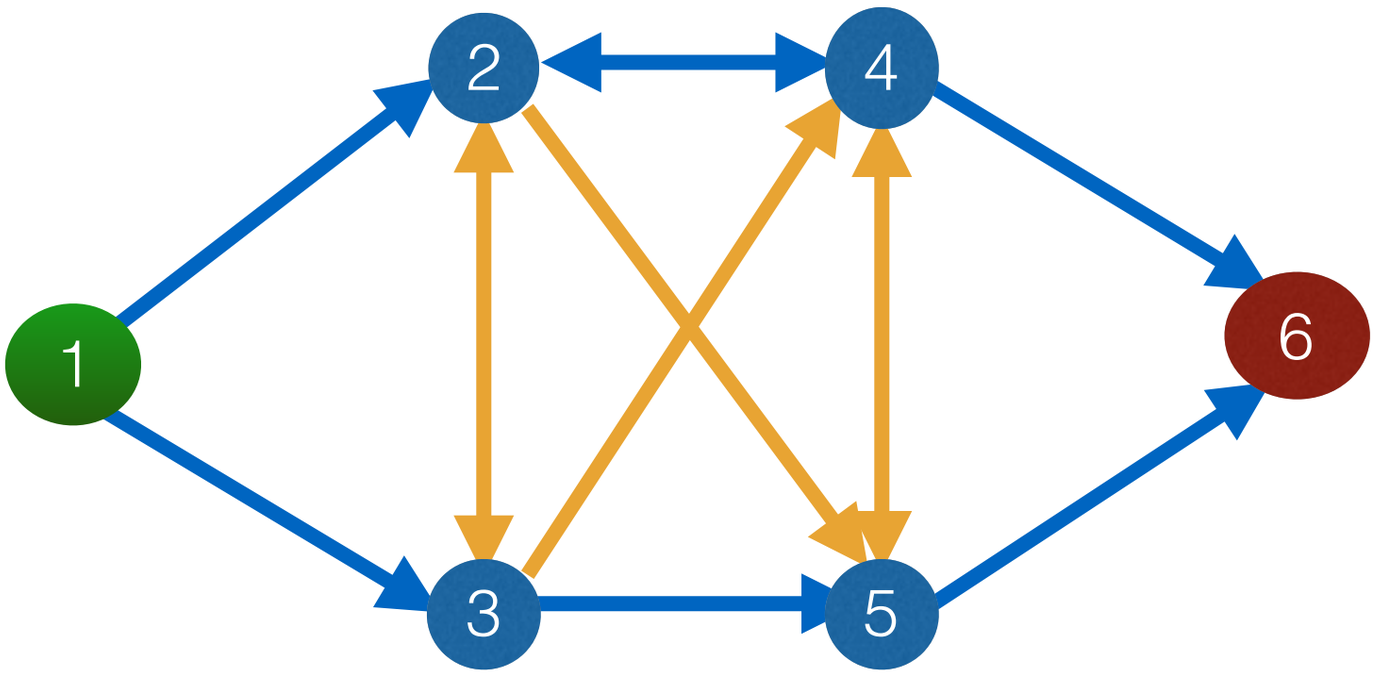
Shortest path routing

Path from 1 to 6?

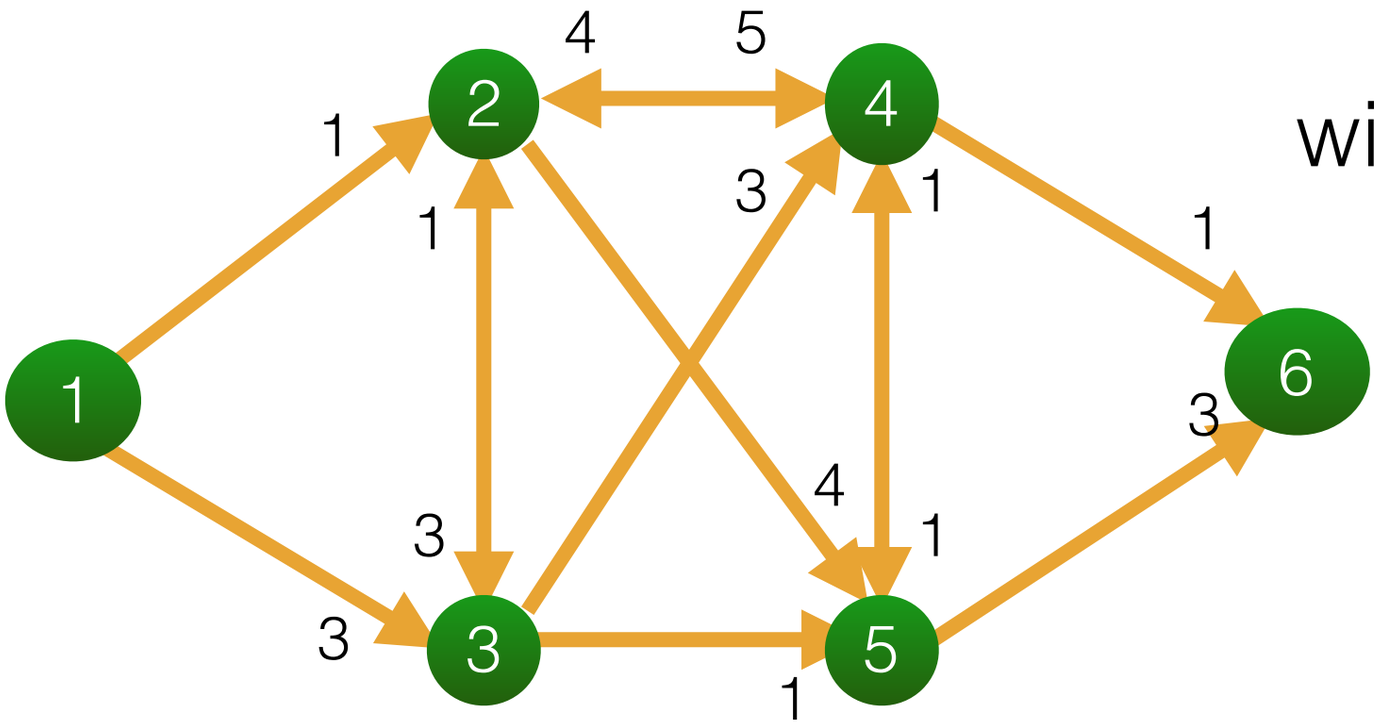


Shortest path routing

Path from 1 to 6?



Shortest path routing

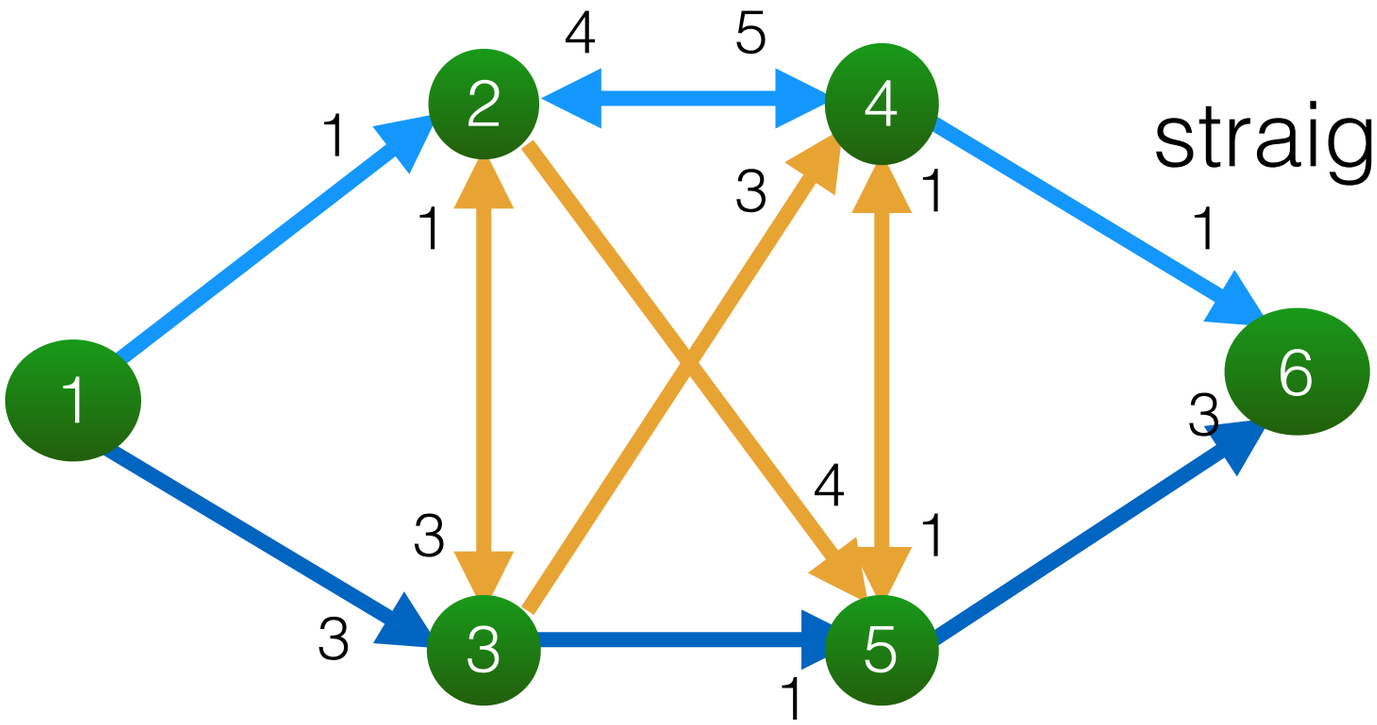


Path from 1 to 6?

with respect to distance

- distance
- time
- cost

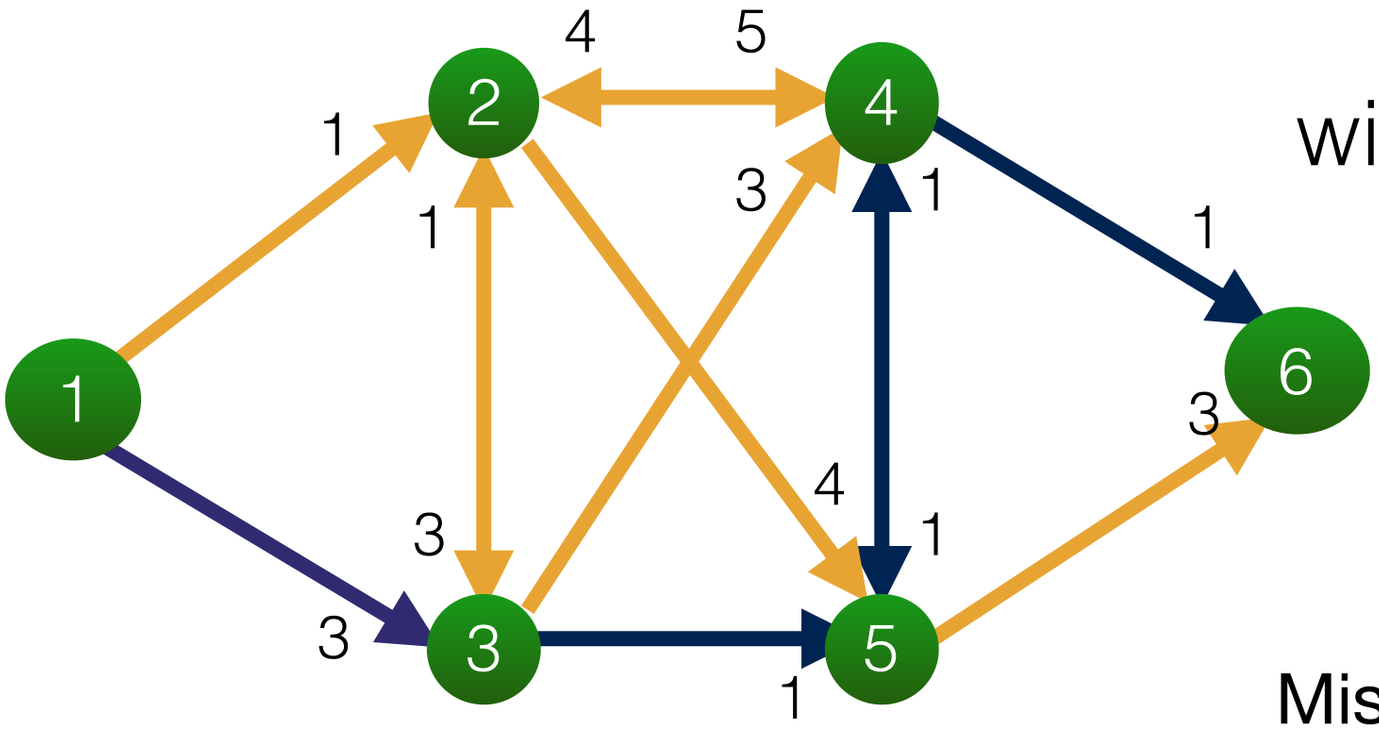
Shortest path routing



Path from 1 to 6?

straightforward solutions - 7

Shortest path routing

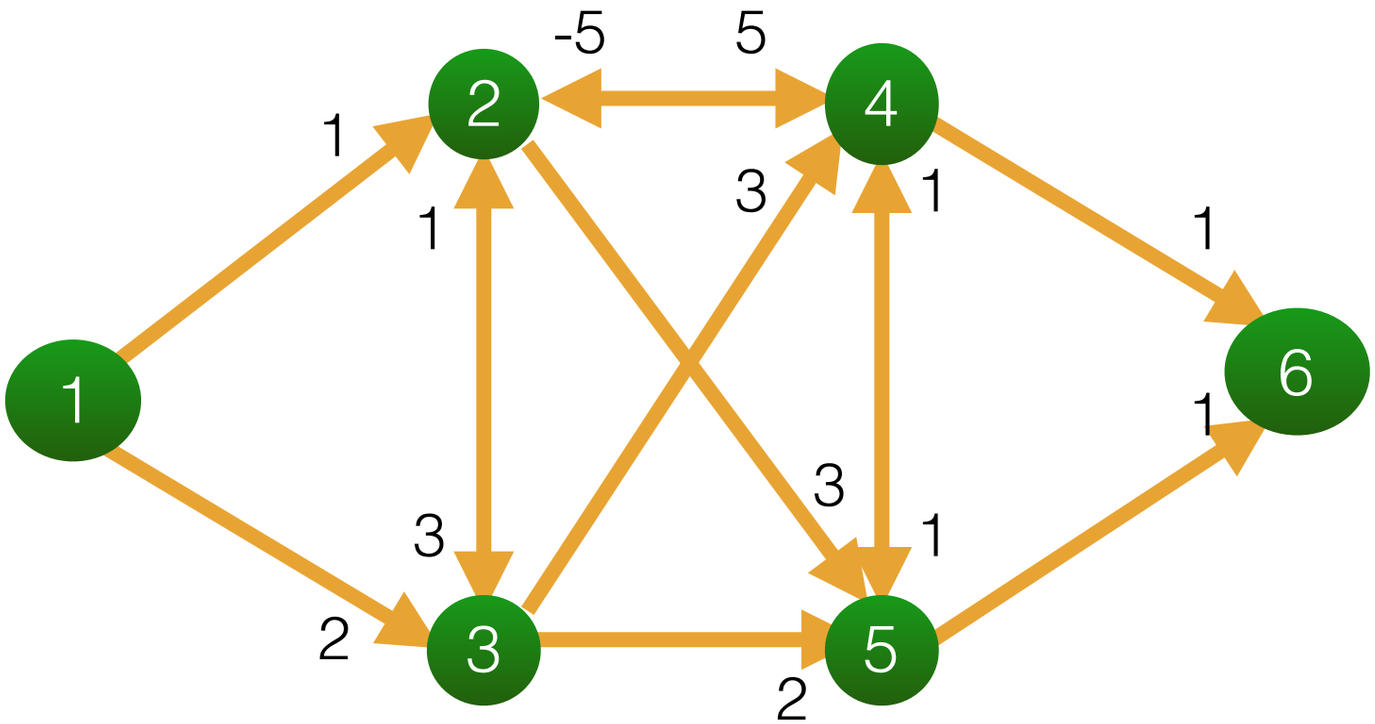


Path from 1 to 6?

with respect to distance

Missed solution with of length 6

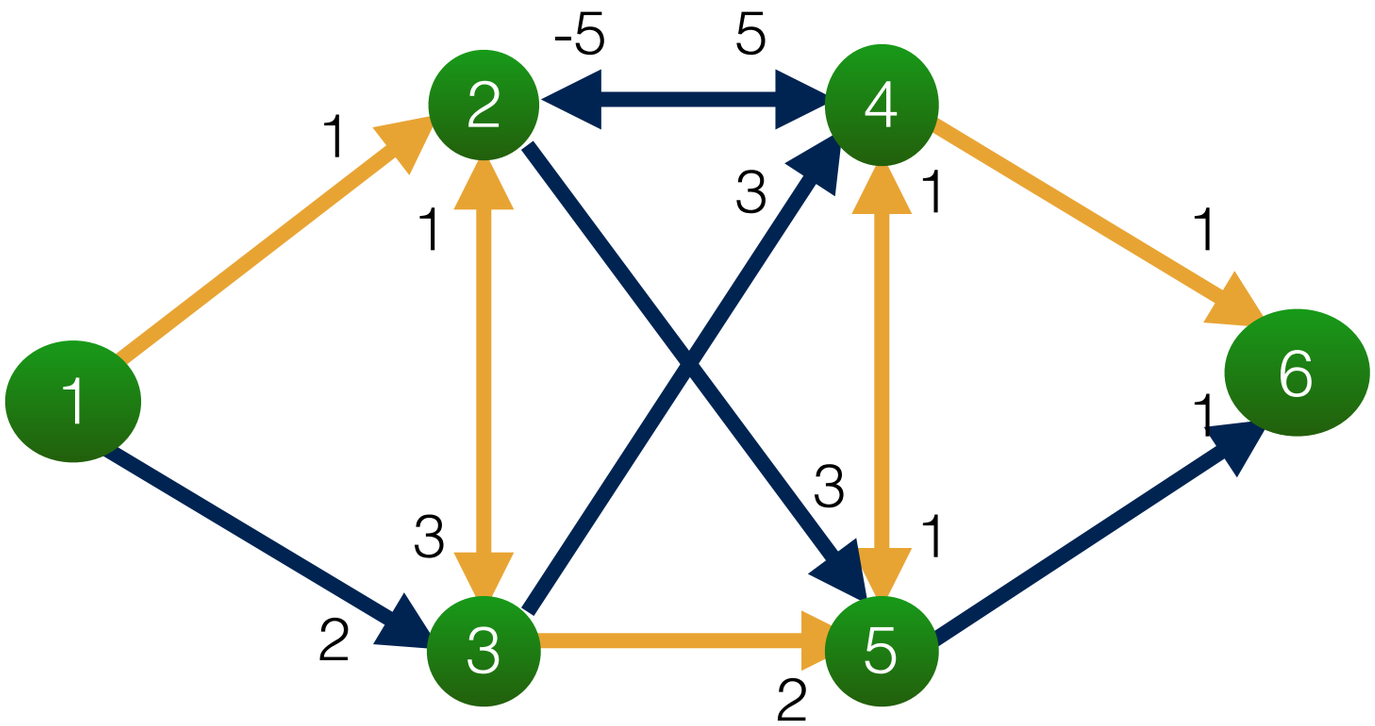
Shortest path routing



Path from 1 to 6?

negative distance

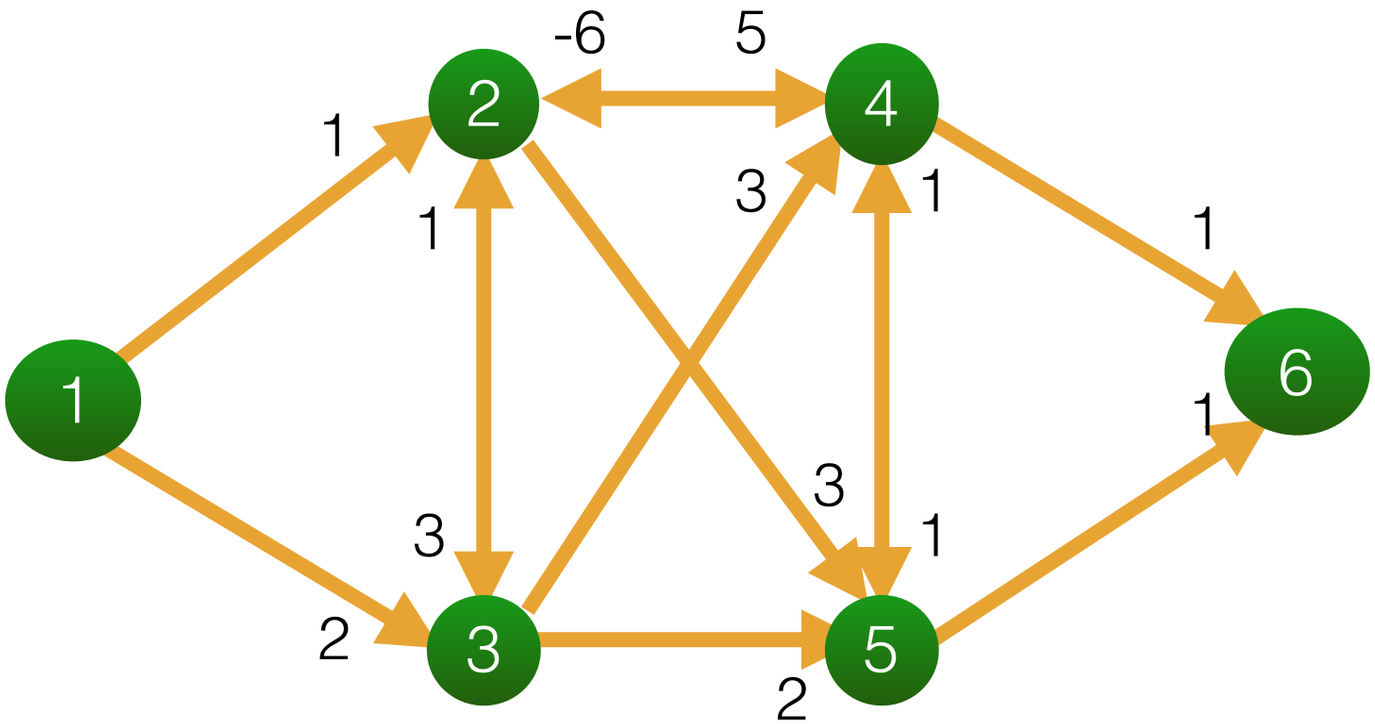
Shortest path routing



Path from 1 to 6?

negative distance

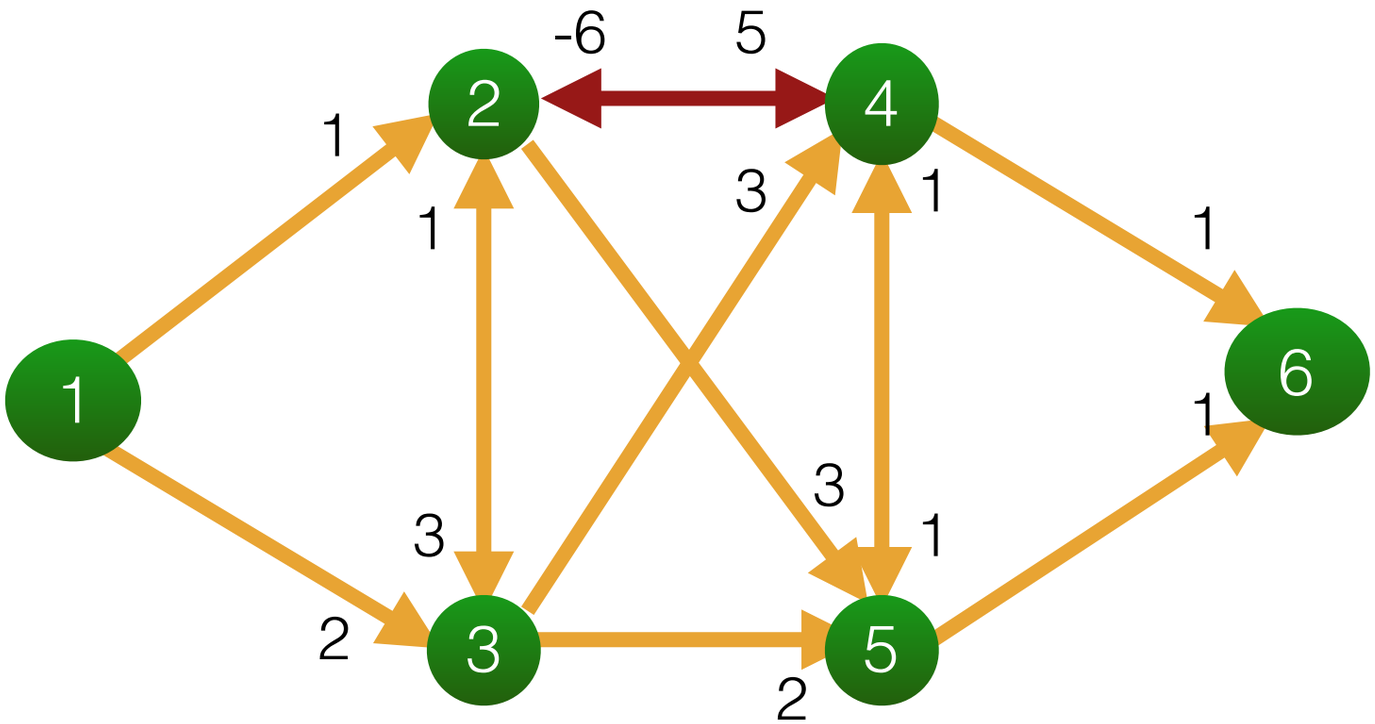
Shortest path routing



Path from 1 to 6?

negative loops

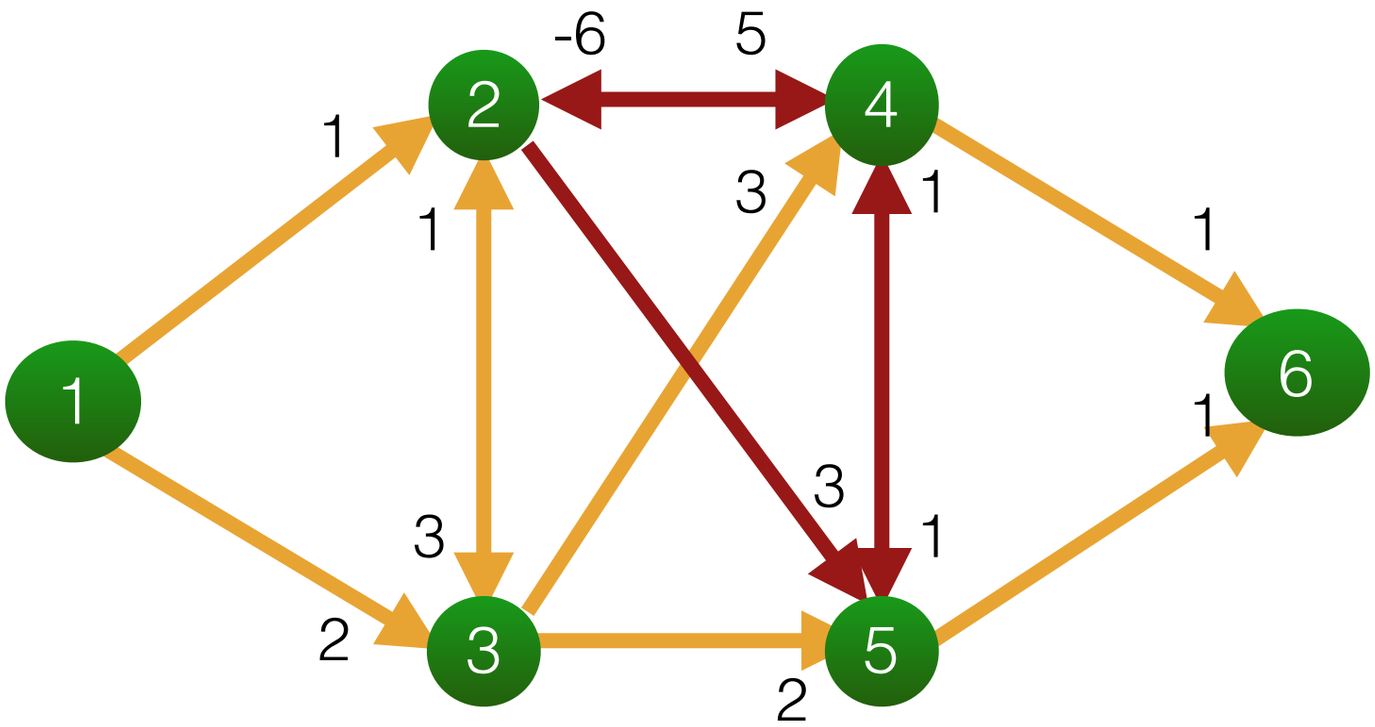
Shortest path routing



Path from 1 to 6?

negative loops

Shortest path routing



Path from 1 to 6?

negative loops

indefinitely small negative distance

Shortest path routing - algorithm types

By what they find:

- Single-pair: from A to B
- Single-source: from A everywhere
- Single-destination: from everywhere to B
- All pairs: from any node to any other one

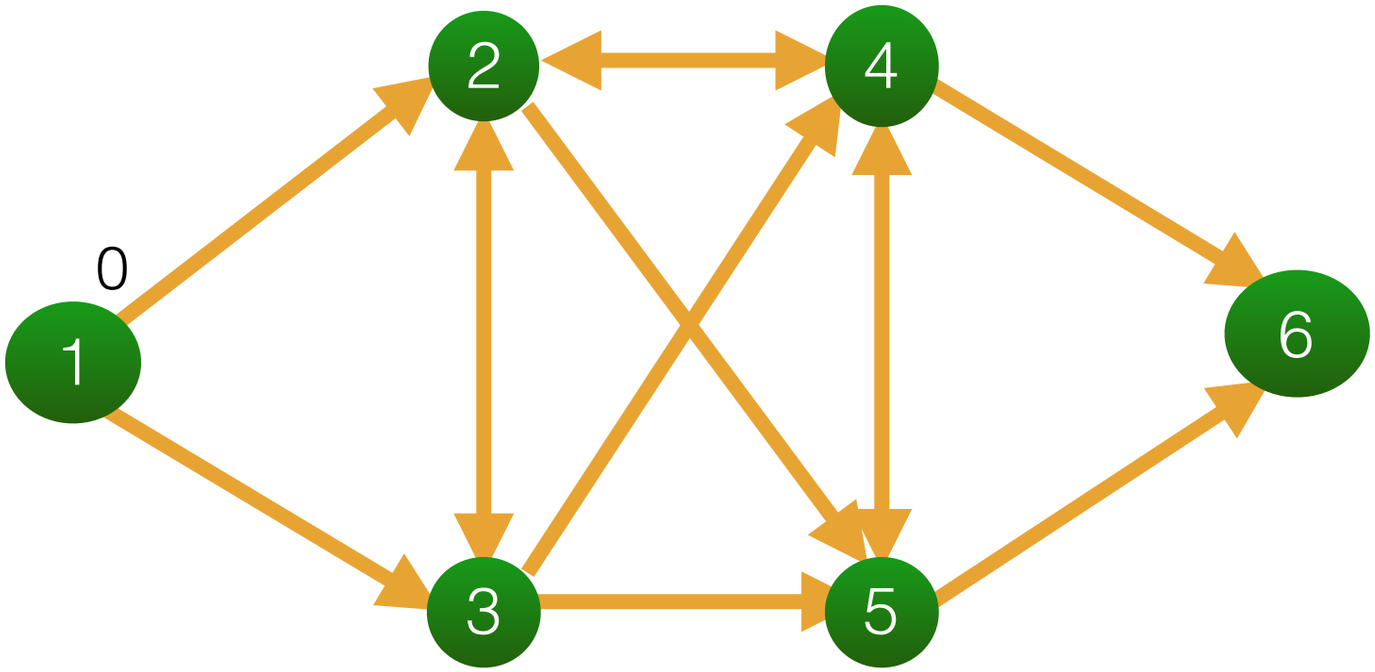
Shortest path routing - algorithm types

By constrains:

- Unweighted graphs
- Undirected graphs
- Positive weights
- Negative, but no loops
- Arbitrary

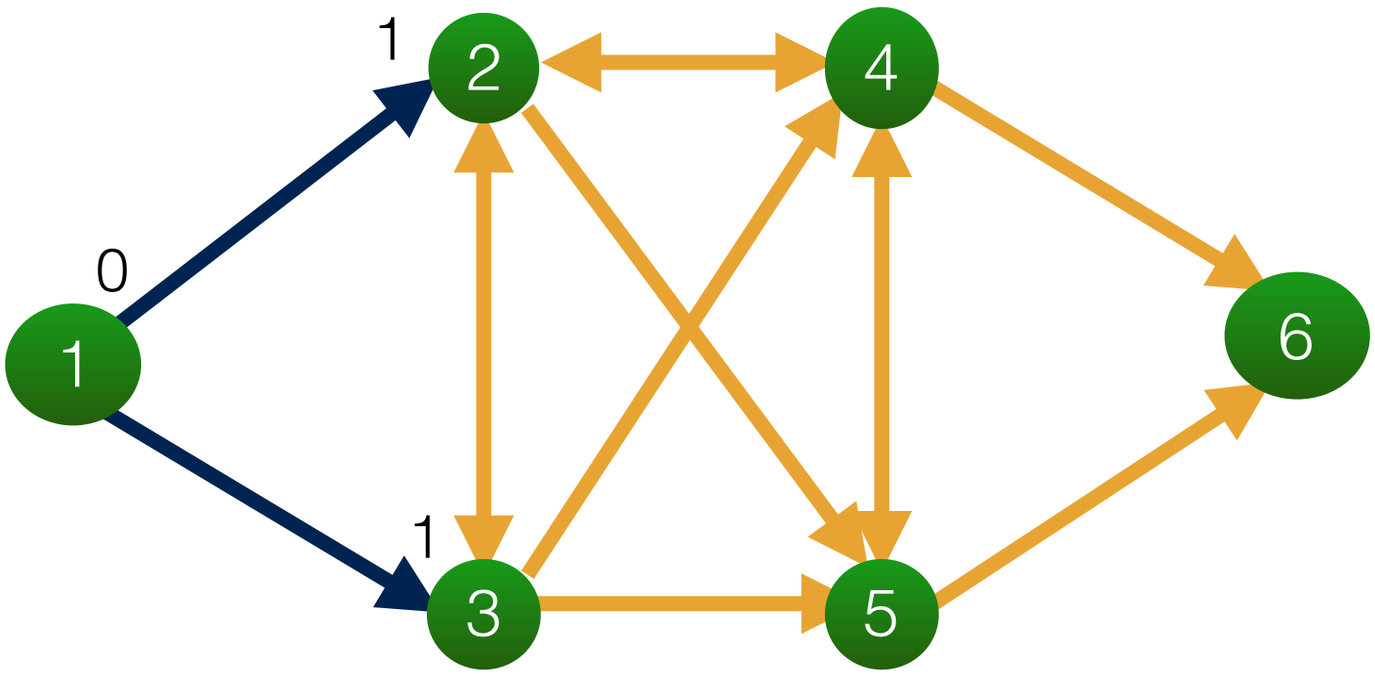
Shortest path routing - unweighted case: breadth-first

Path from 1 to everywhere?



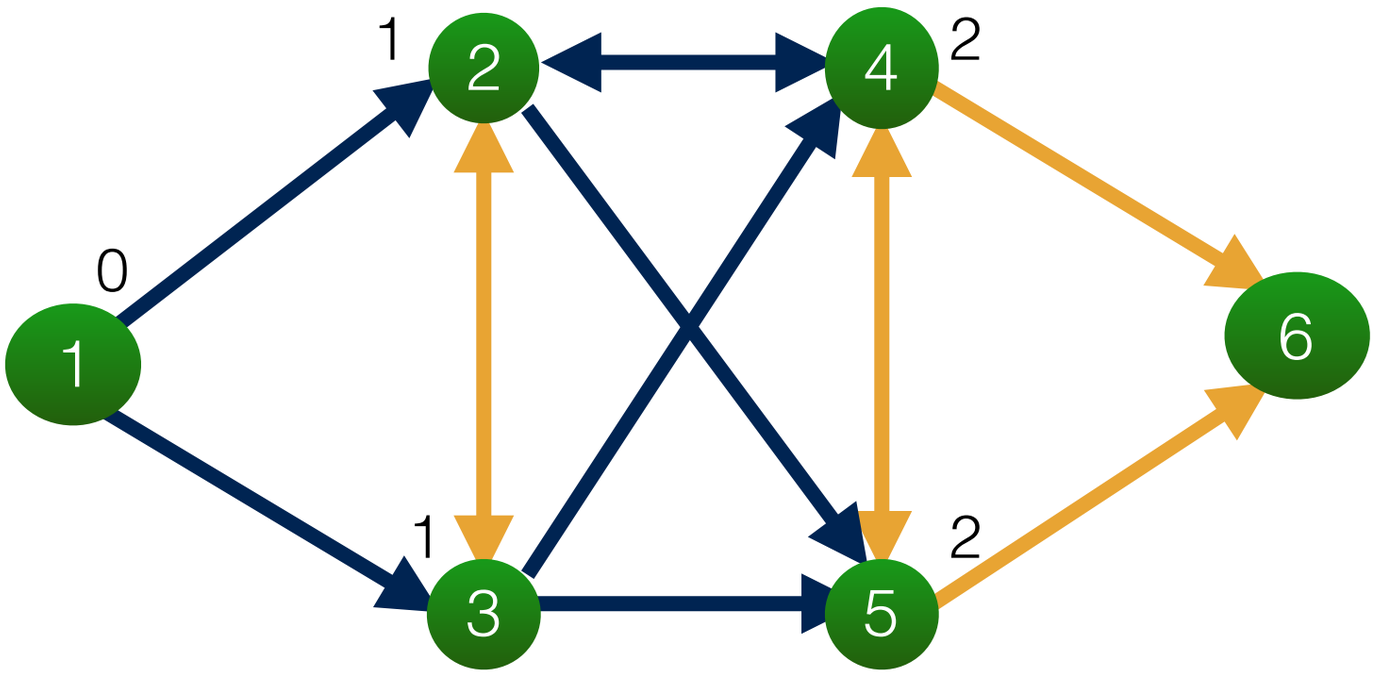
Shortest path routing - unweighted case: breadth-first

Path from 1 to everywhere?

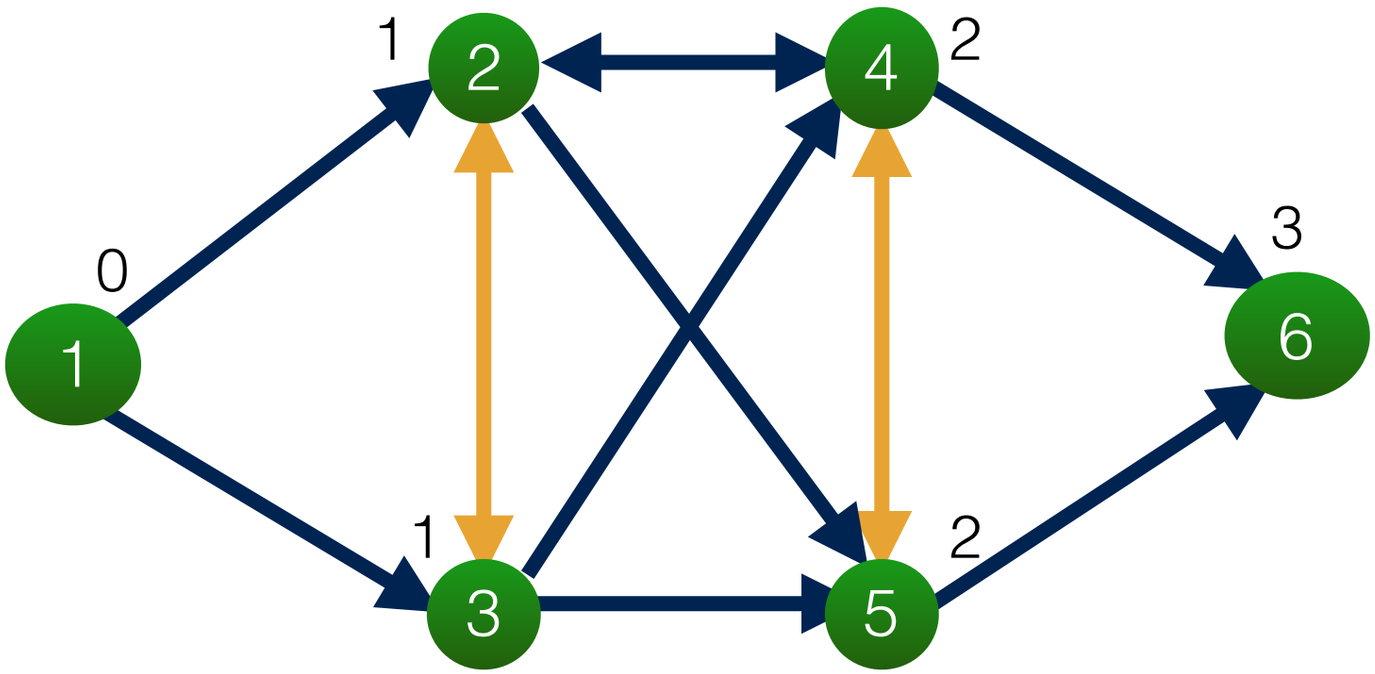


Shortest path routing - unweighted case: breadth-first

Path from 1 to everywhere?



Shortest path routing - unweighted case

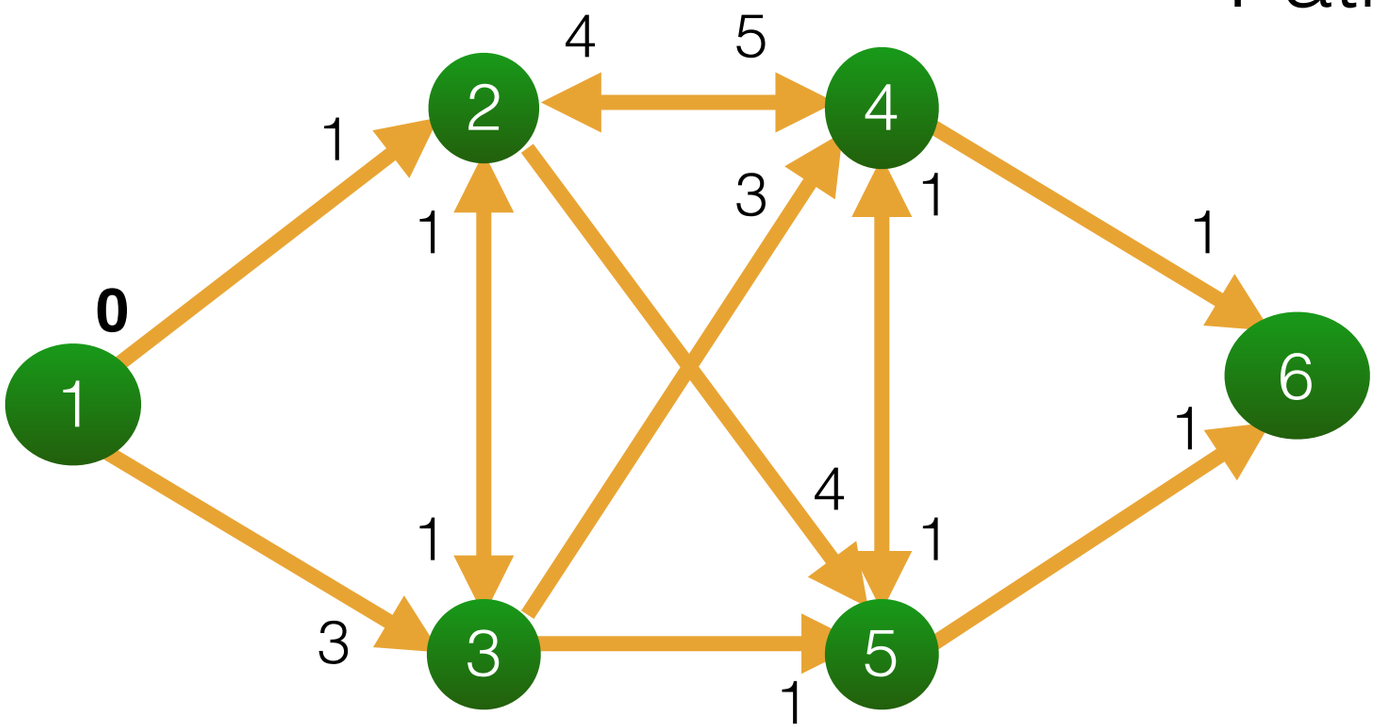


Path from 1 to everywhere?

Path from 1 to 6?

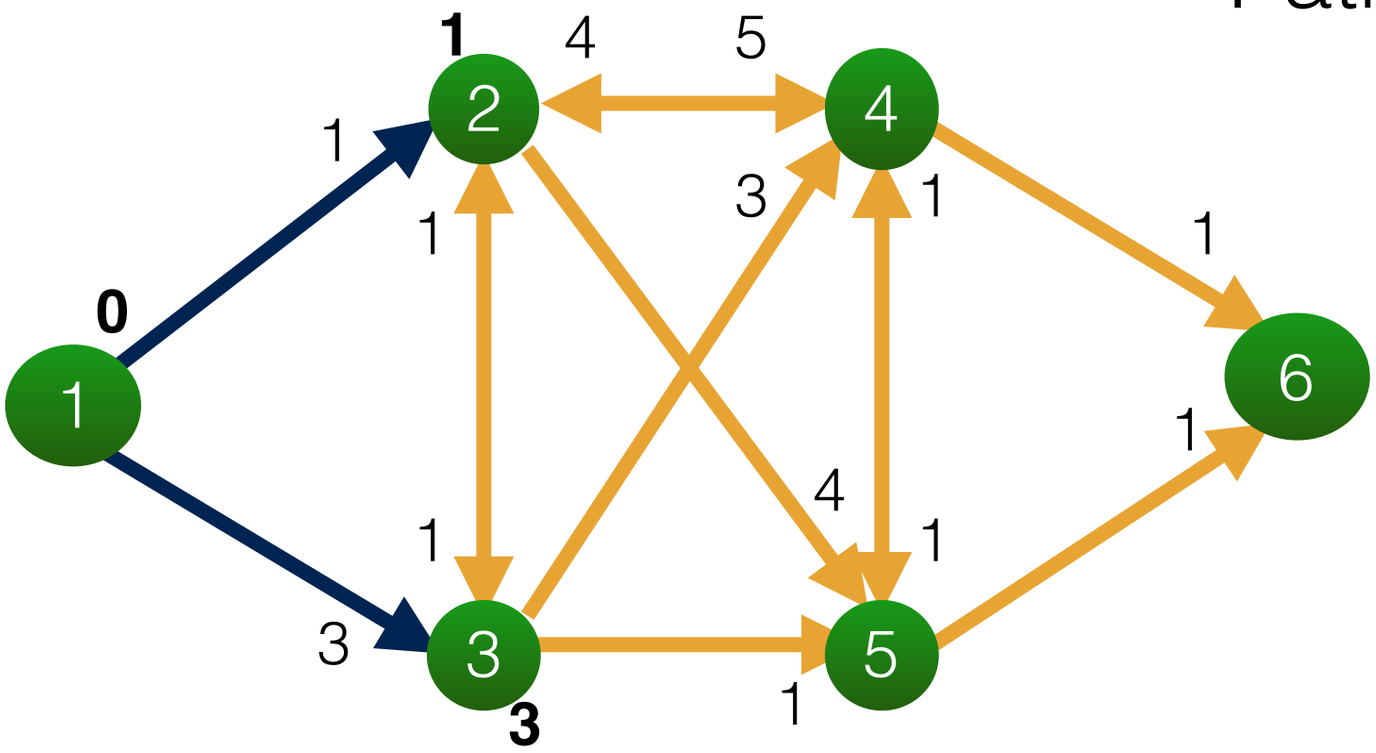
Shortest path routing - weighted case

Path from 1 to everywhere?



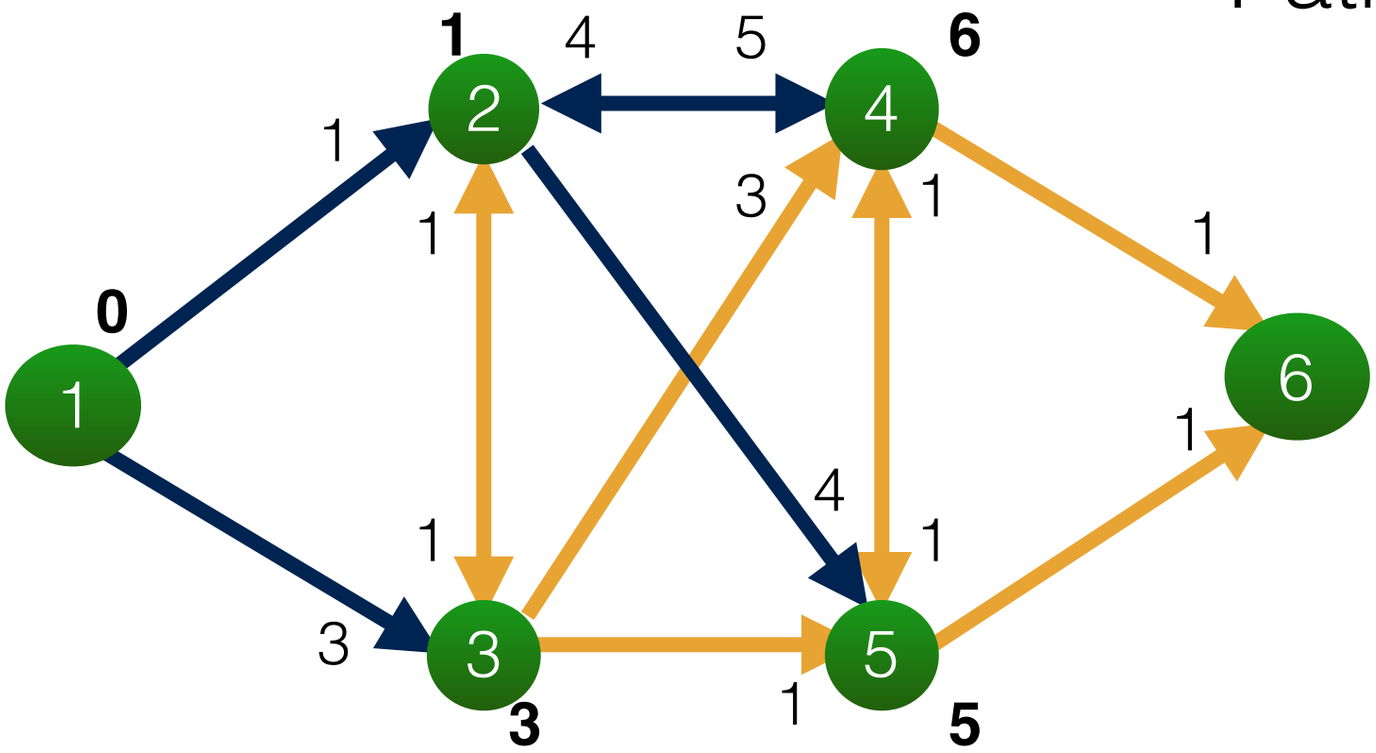
Shortest path routing - weighted case

Path from 1 to everywhere?



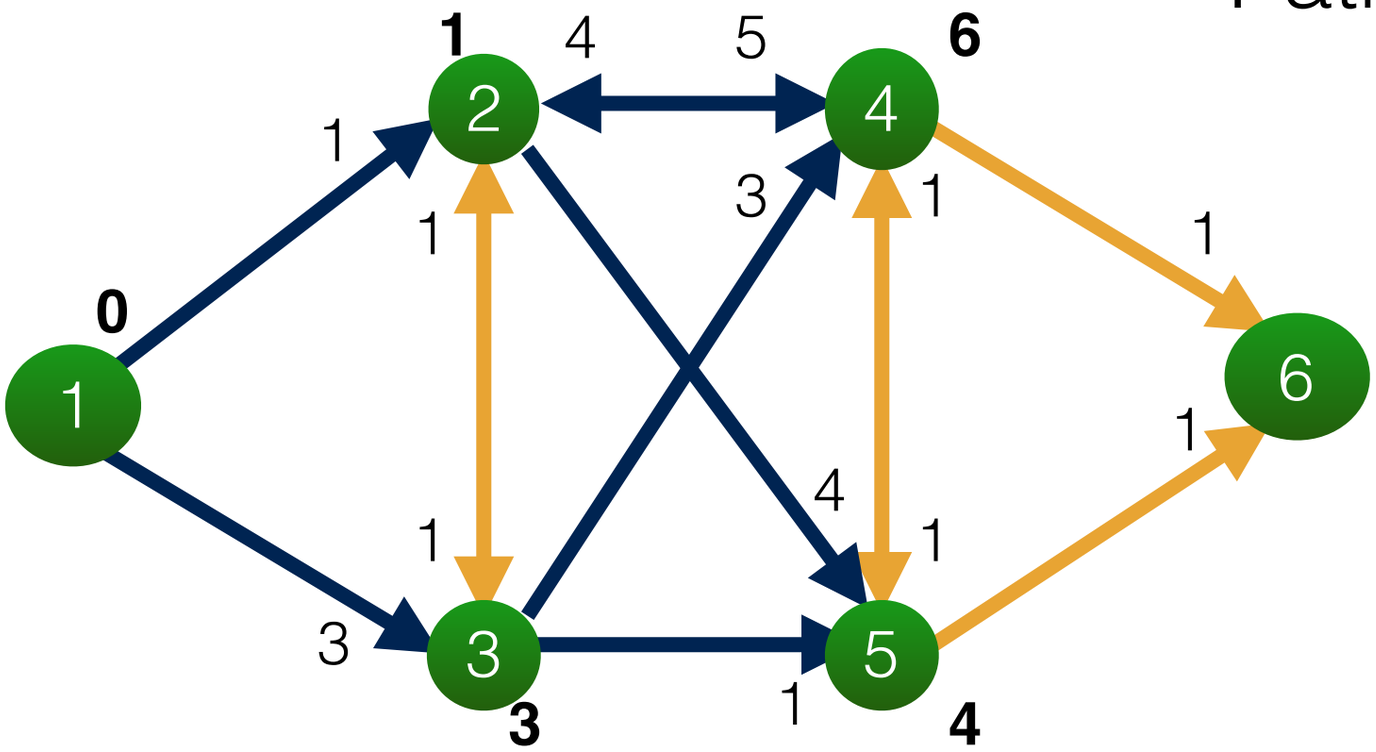
Shortest path routing - weighted case

Path from 1 to everywhere?



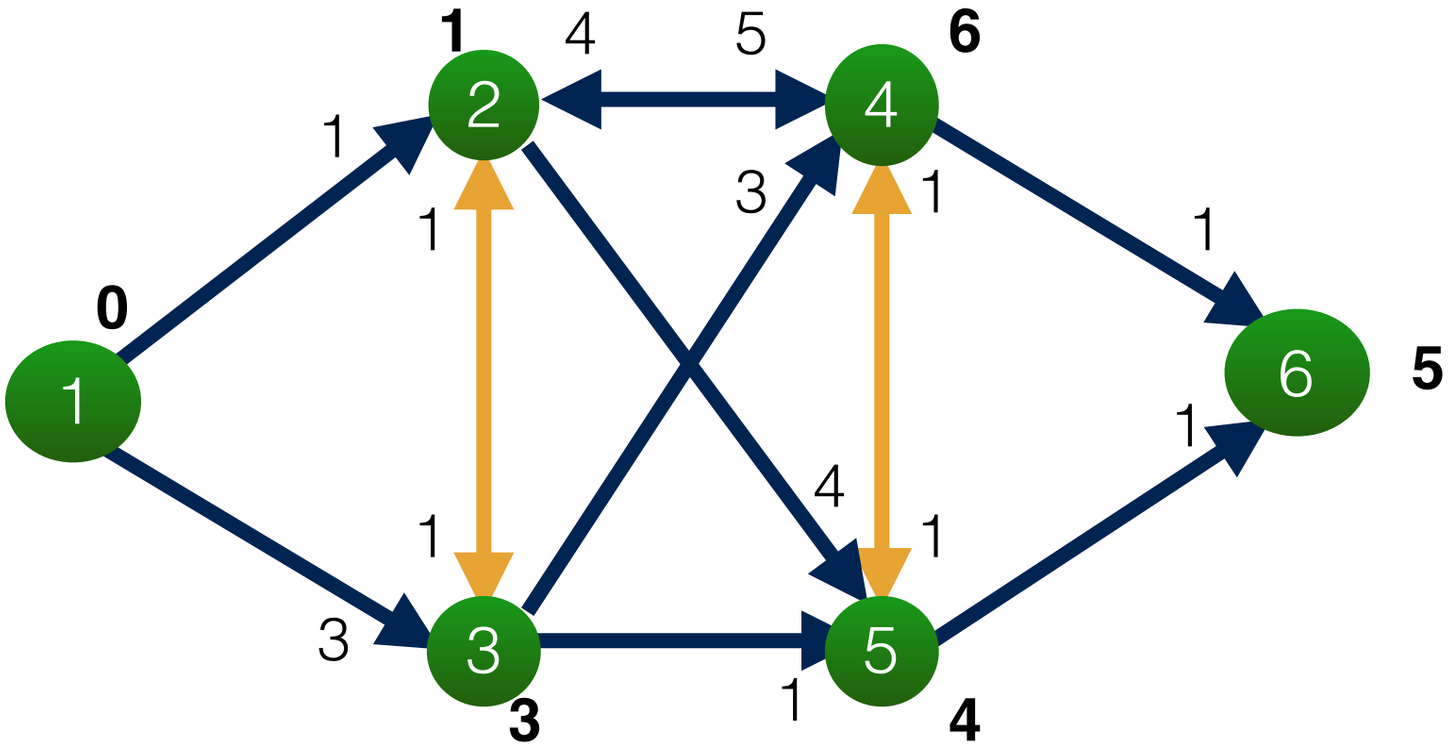
Shortest path routing - weighted case

Path from 1 to everywhere?



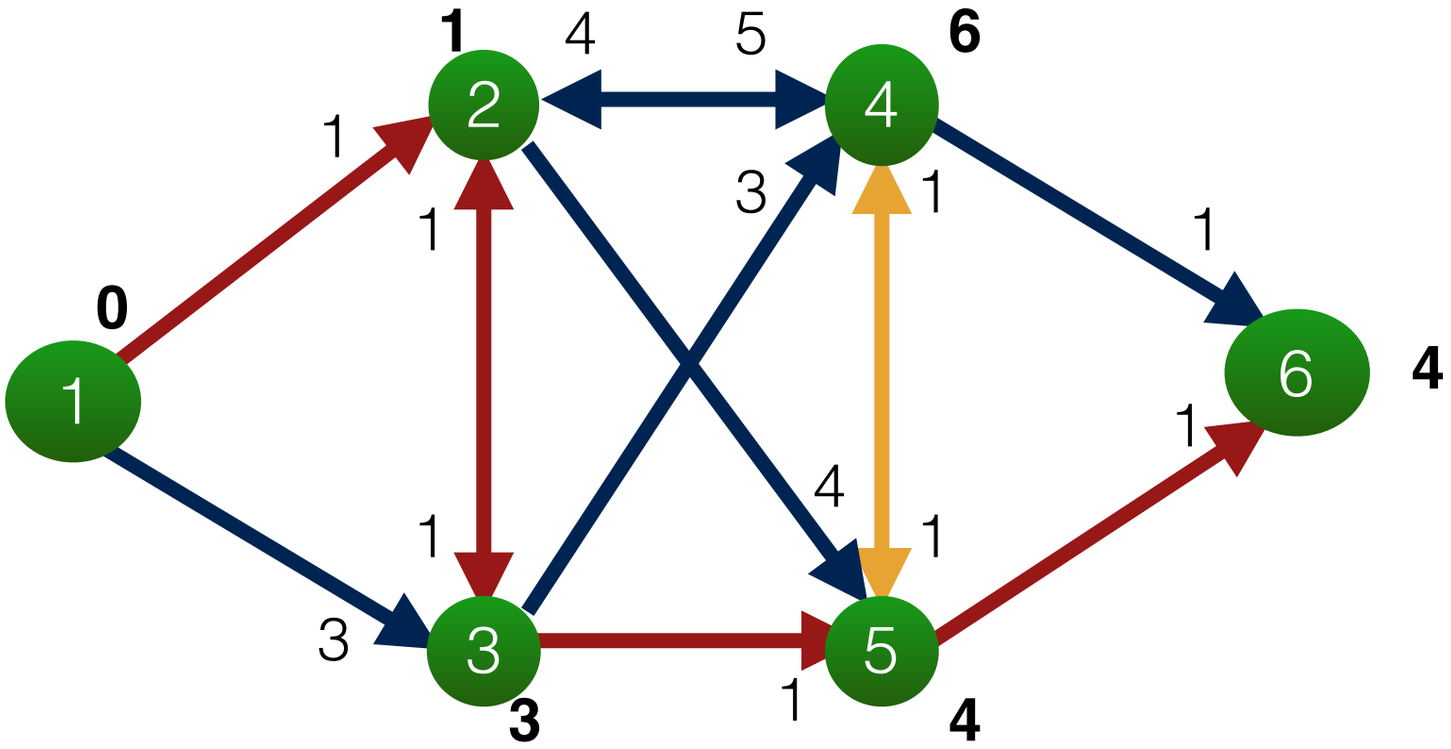
Shortest path routing - weighted case

Have we found a shortest path to 6?



Shortest path routing - weighted case

No - we overlooked a shorter one!



Dijkstra algorithm - 1956: single-source

1. For a given source, assign it an estimated distance 0, all others - infinity; consider estimates for all nodes uncertain

2. Select a node with the smallest estimated distance as current and mark it as certain.

3. If there still exist any uncertain node

No → Stop

Yes

4. Update estimated distances to all the uncertain neighbors of the current node using its certain distance plus corresponding direct connections. Return to 2

Dijkstra algorithm - 1956: single pair

1. For a given source, assign it an estimated distance 0, all others - infinity; consider estimates for all nodes uncertain

2. Select a node with the smallest estimated distance as current and mark it as certain.

3. If target is uncertain

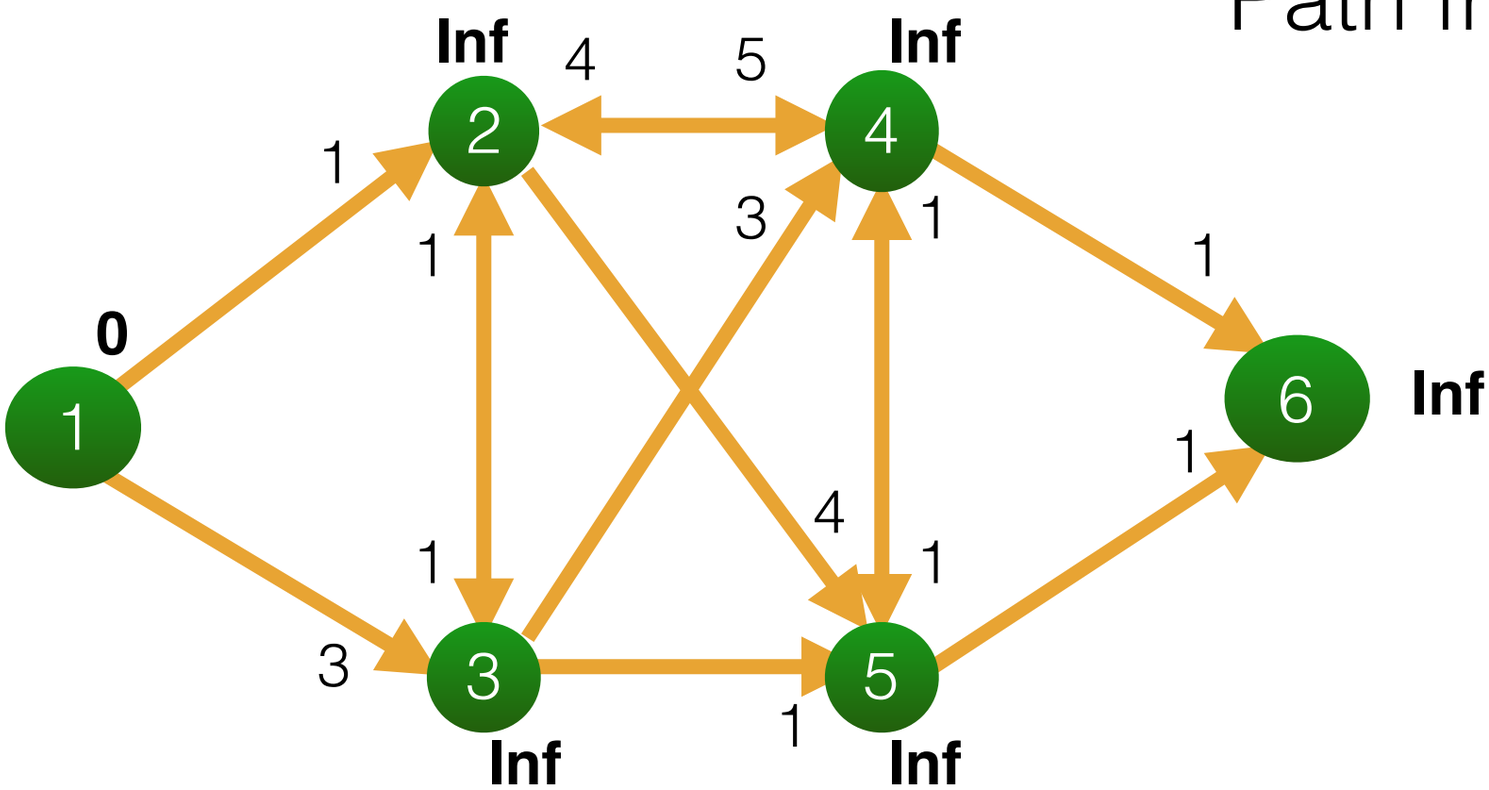
No → Stop

Yes

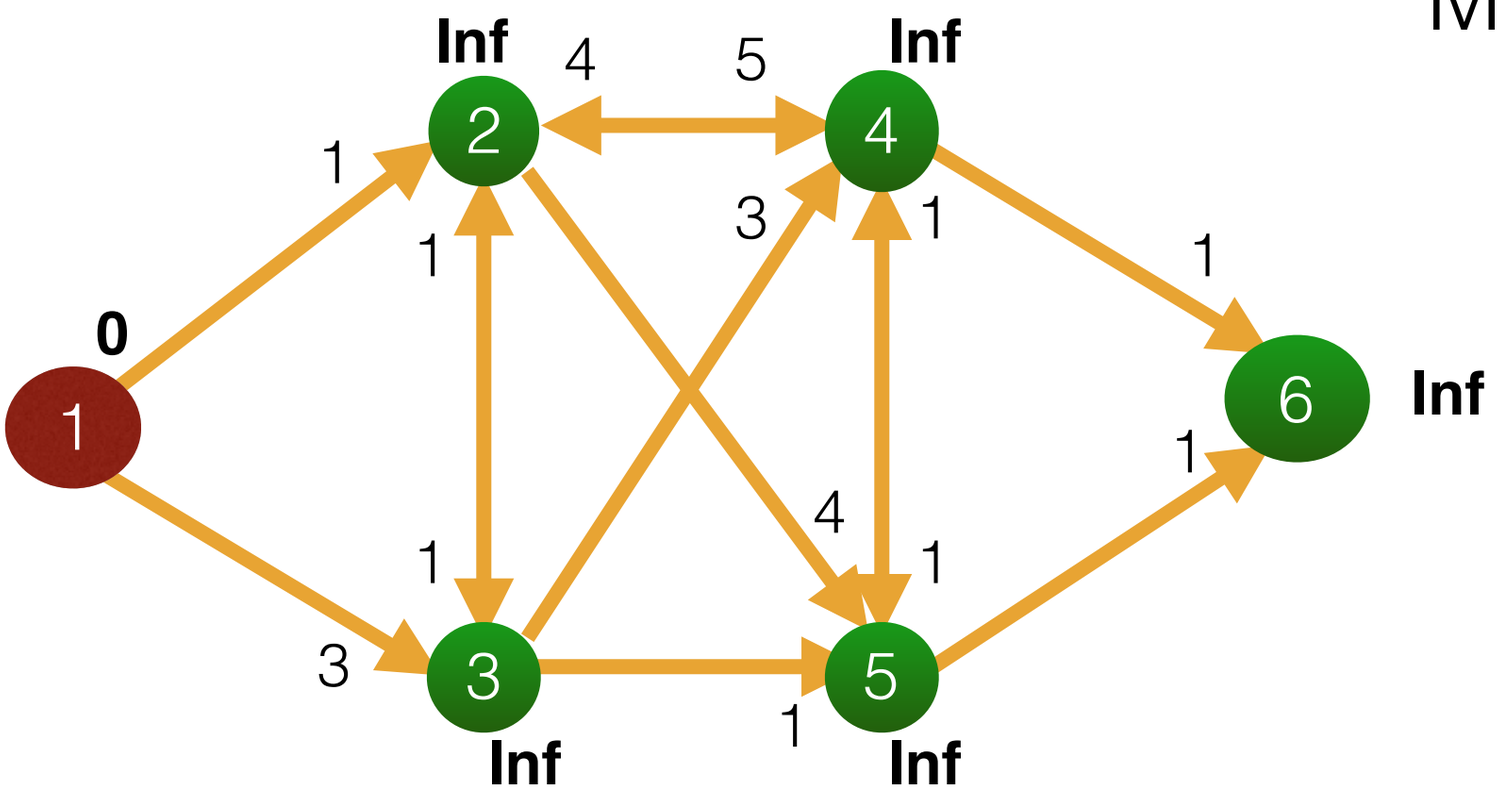
4. Update estimated distances to all the uncertain neighbors of the current node using its certain distance plus corresponding direct connections. Return to 2

Shortest path routing - Dijkstra

Path from 1 to everywhere?

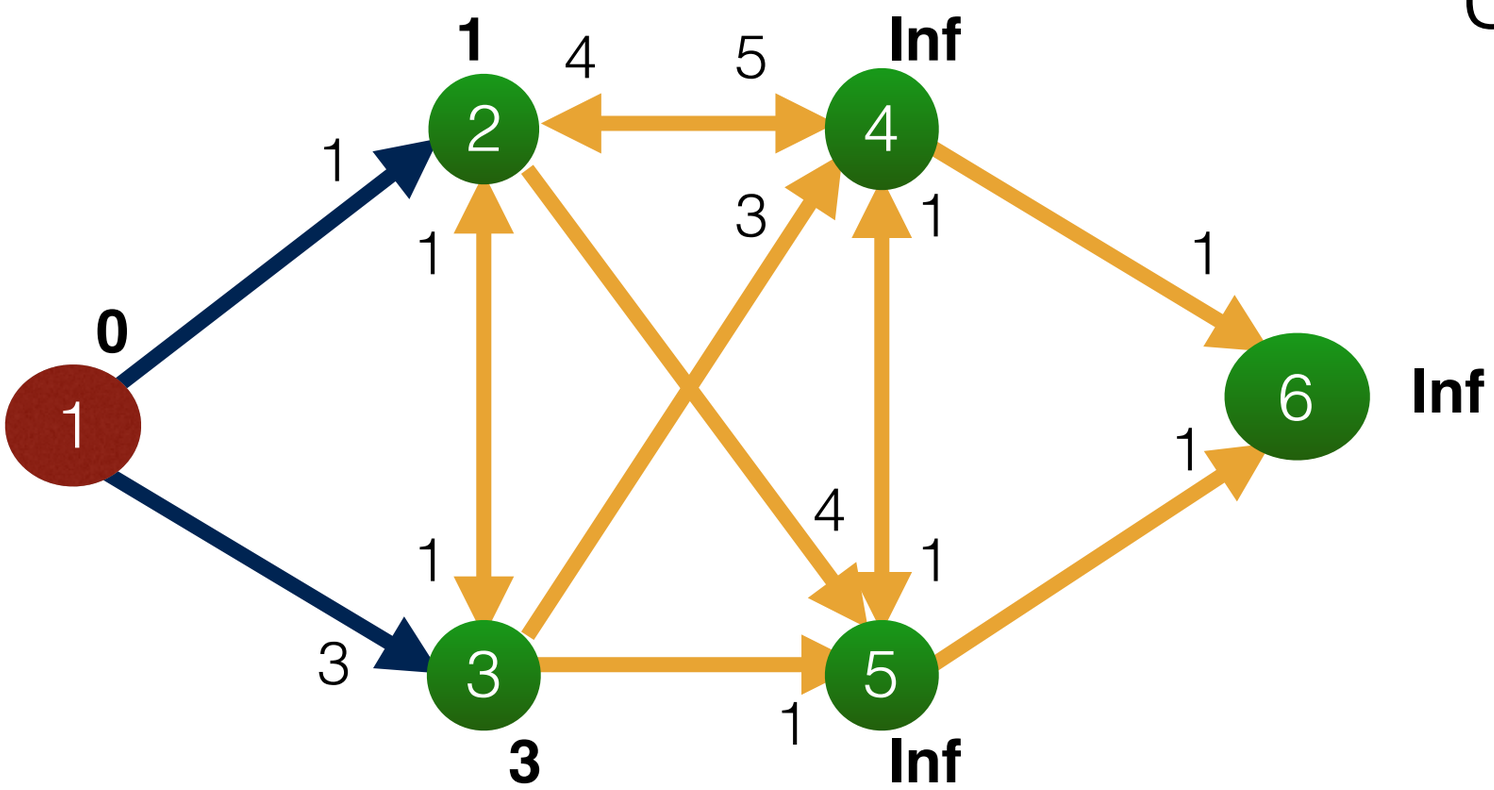


Shortest path routing - Dijkstra



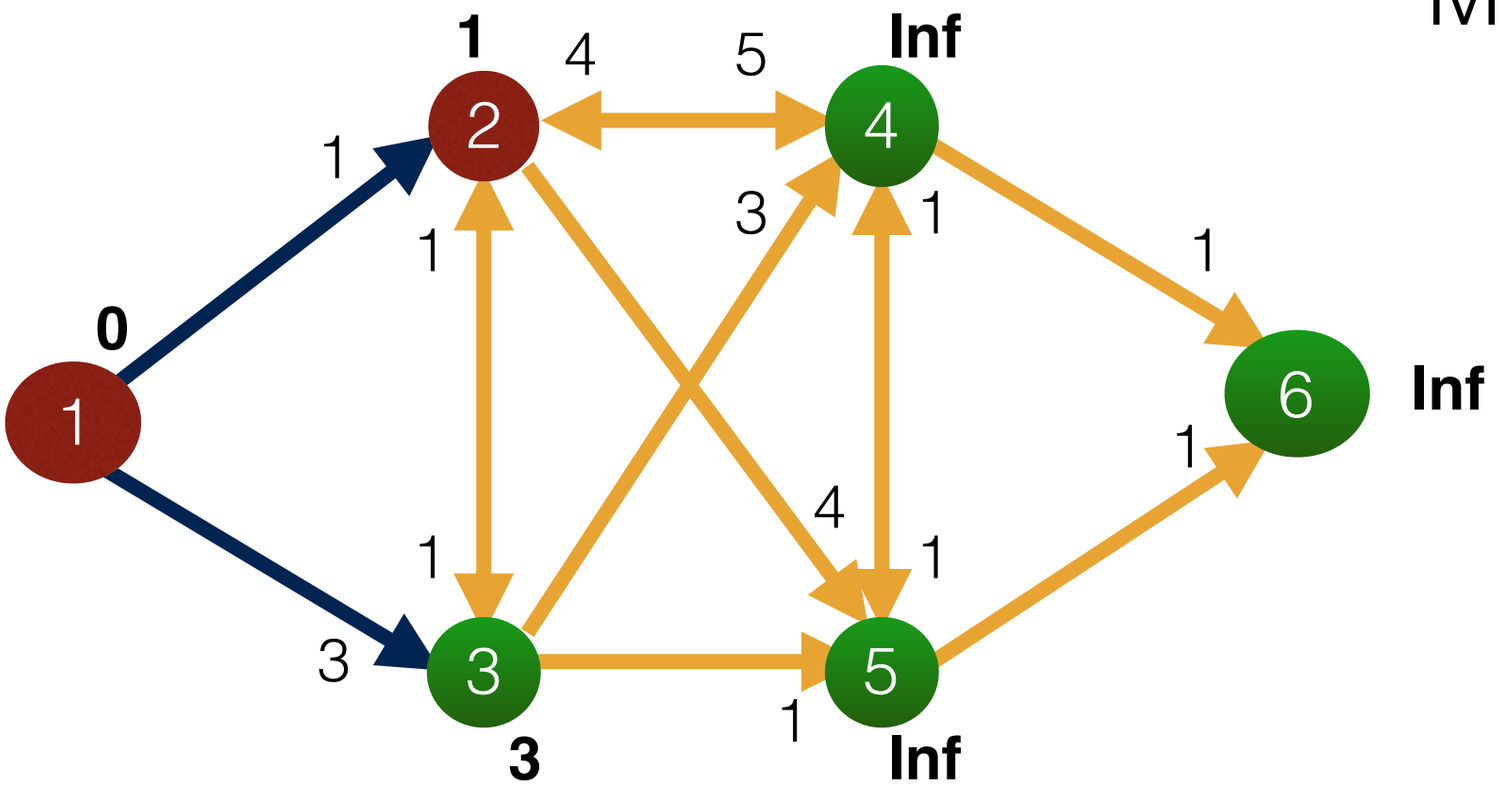
Mark 1 as certain

Shortest path routing - Dijkstra



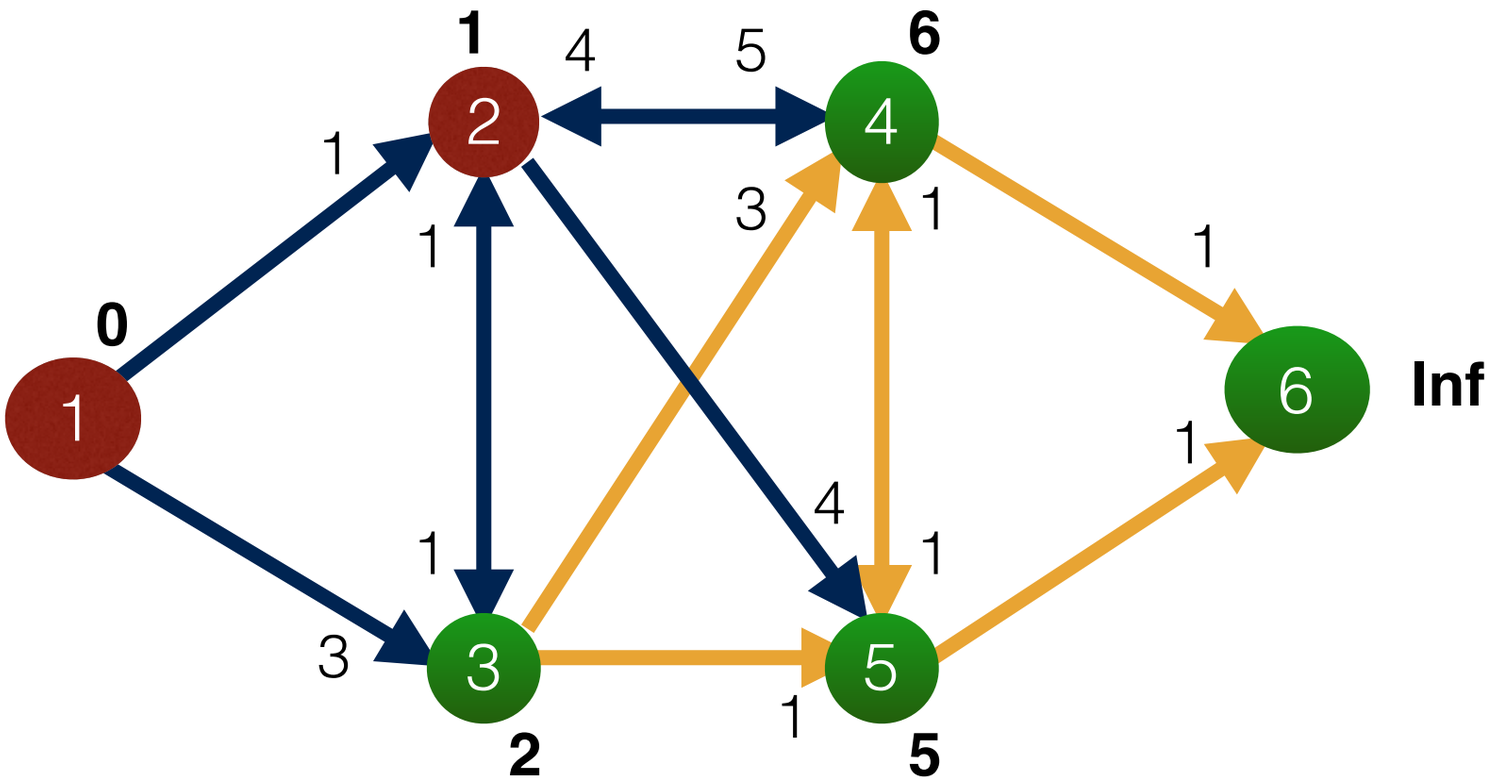
Update 2 and 3

Shortest path routing - Dijkstra



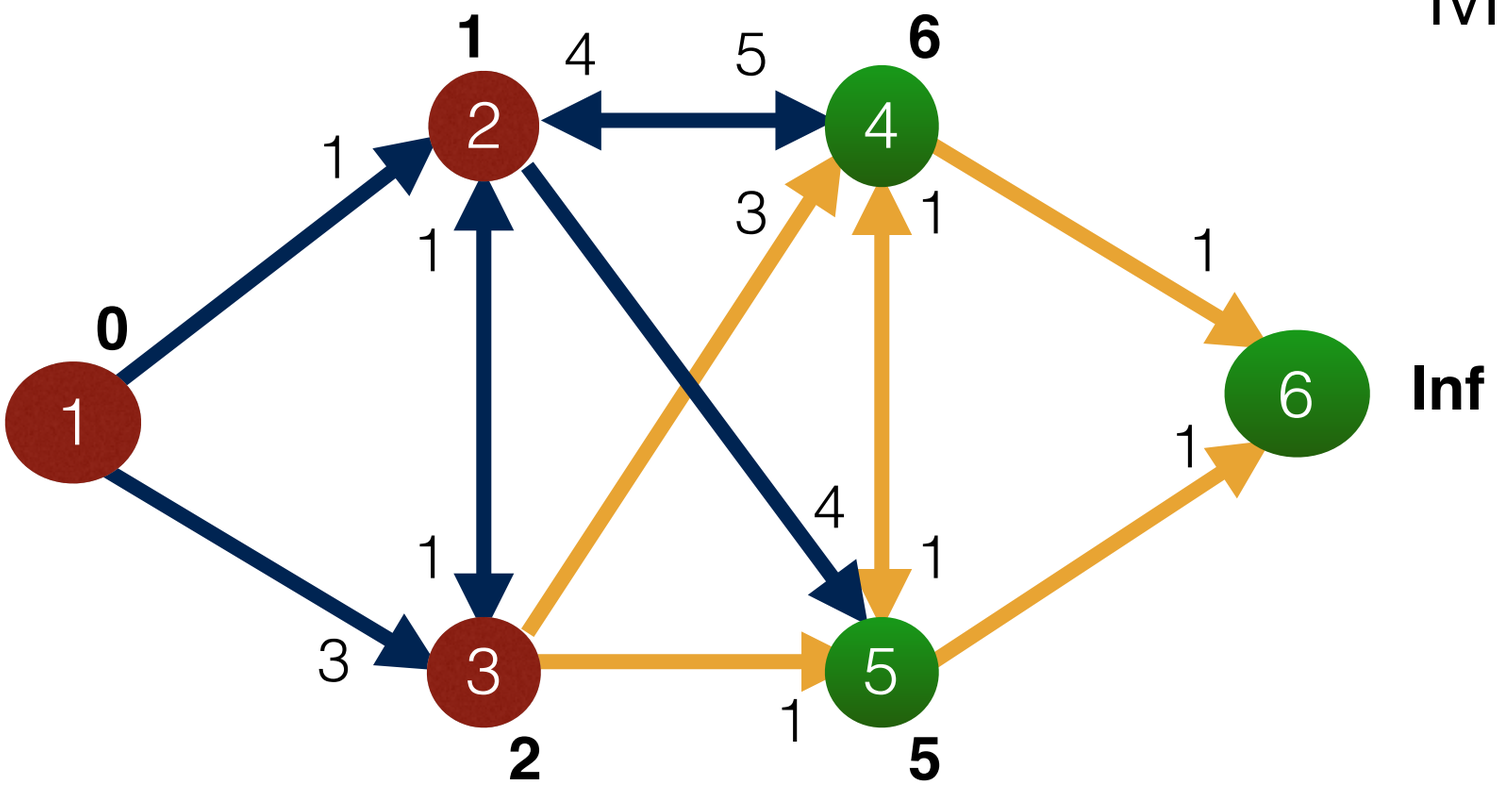
Mark 2 as certain

Shortest path routing - Dijkstra



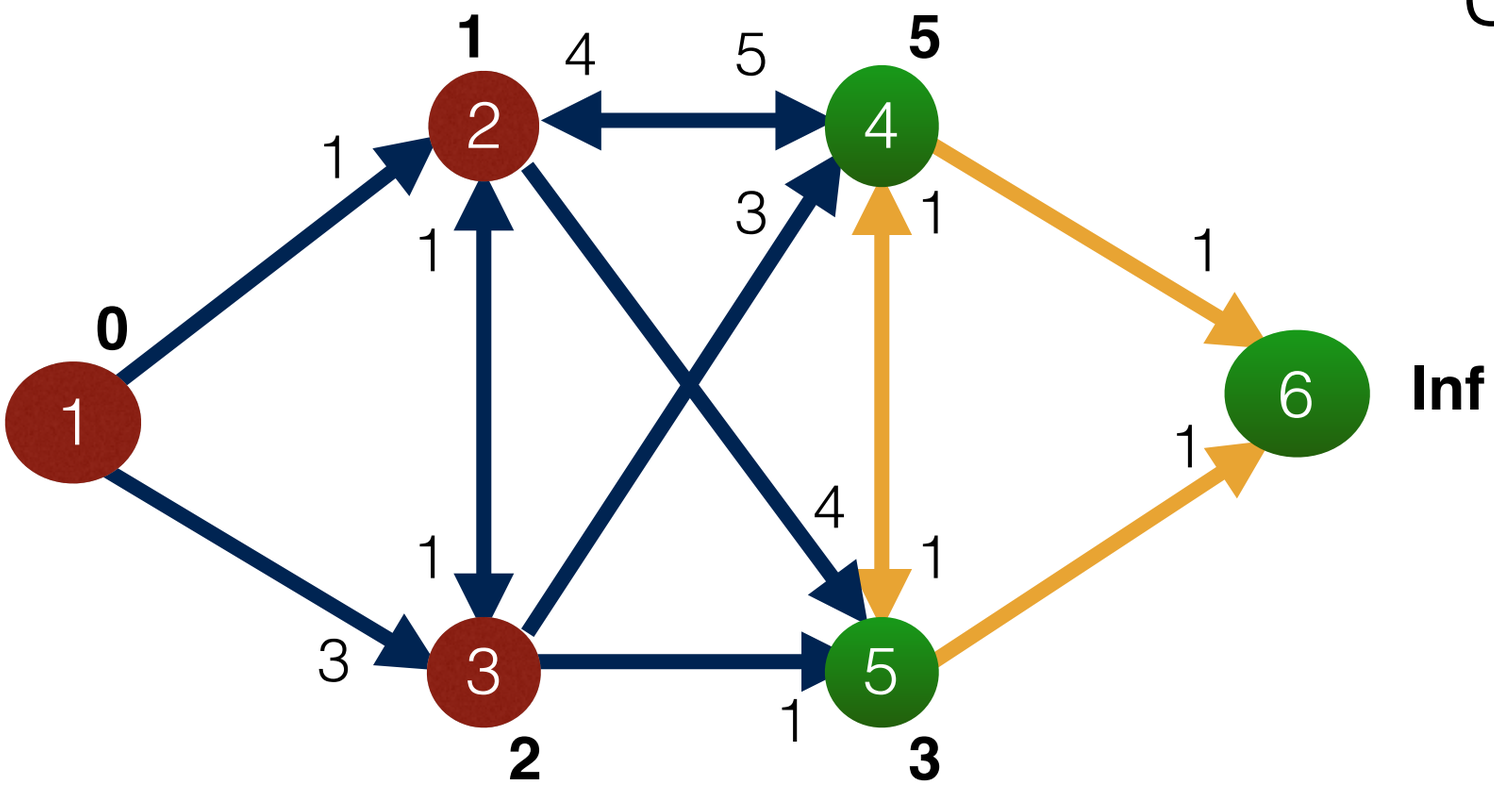
Update 3,4,5

Shortest path routing - Dijkstra



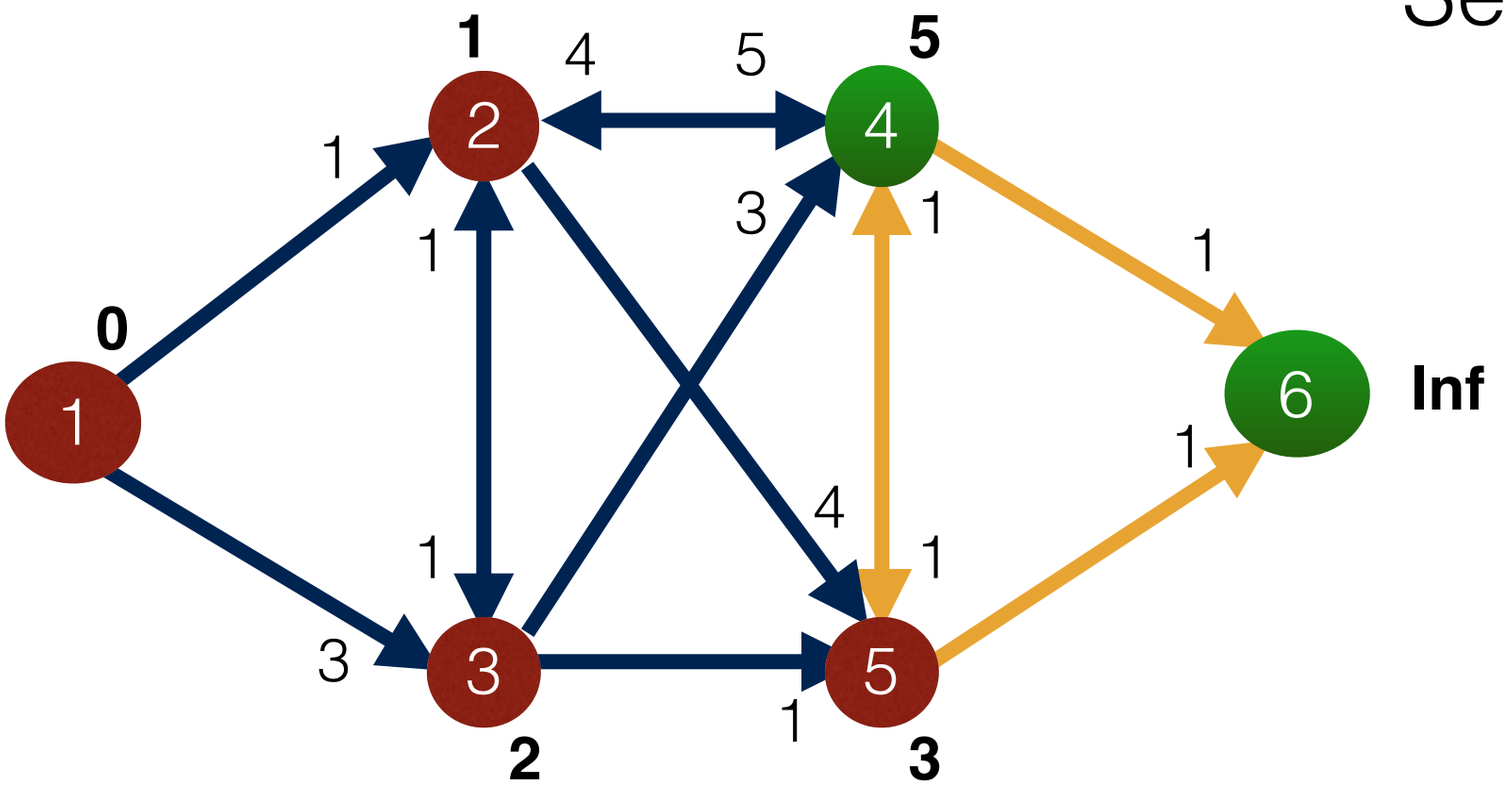
Mark 3 as certain

Shortest path routing - Dijkstra



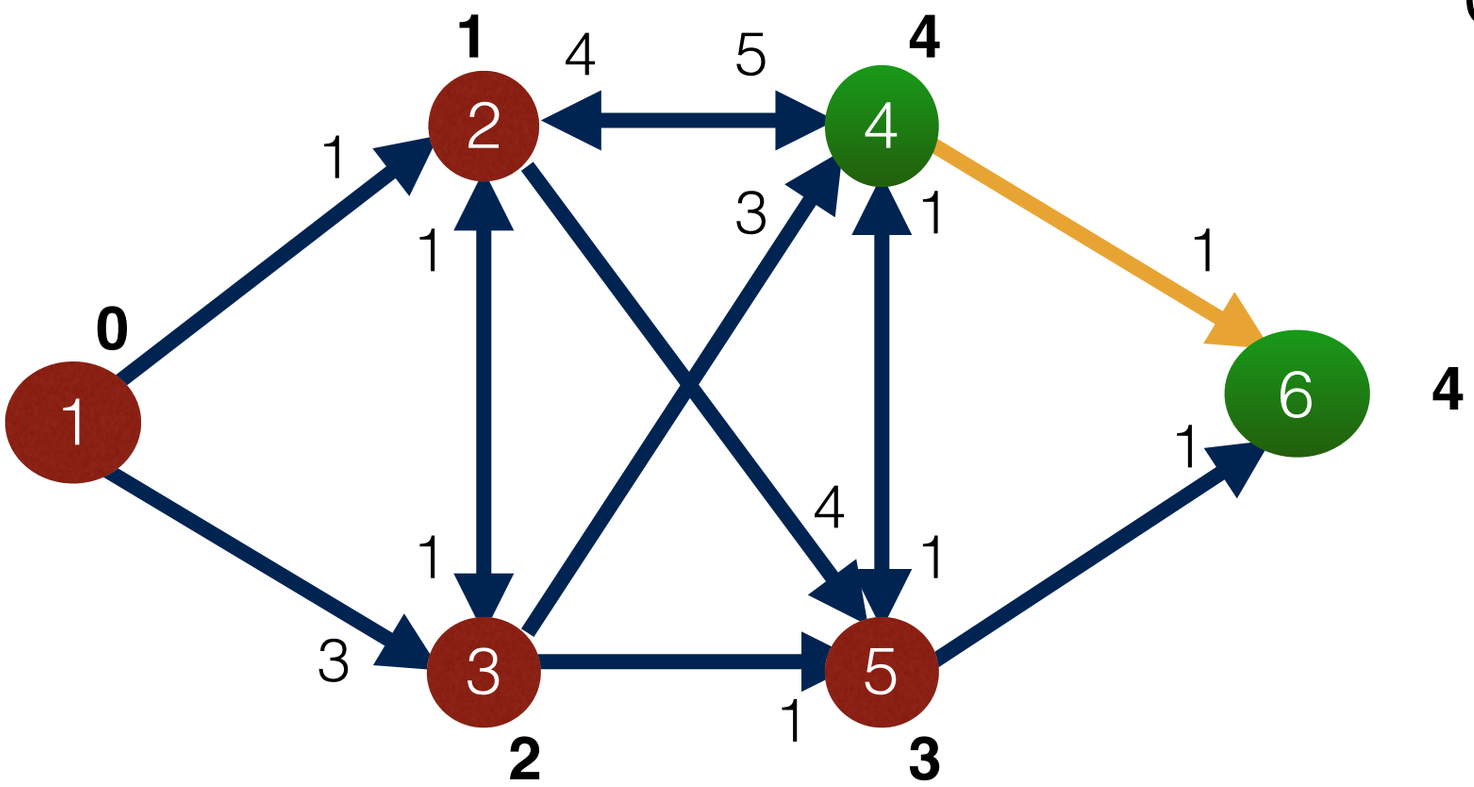
Update 4 and 5

Shortest path routing - Dijkstra



Select 5 as certain

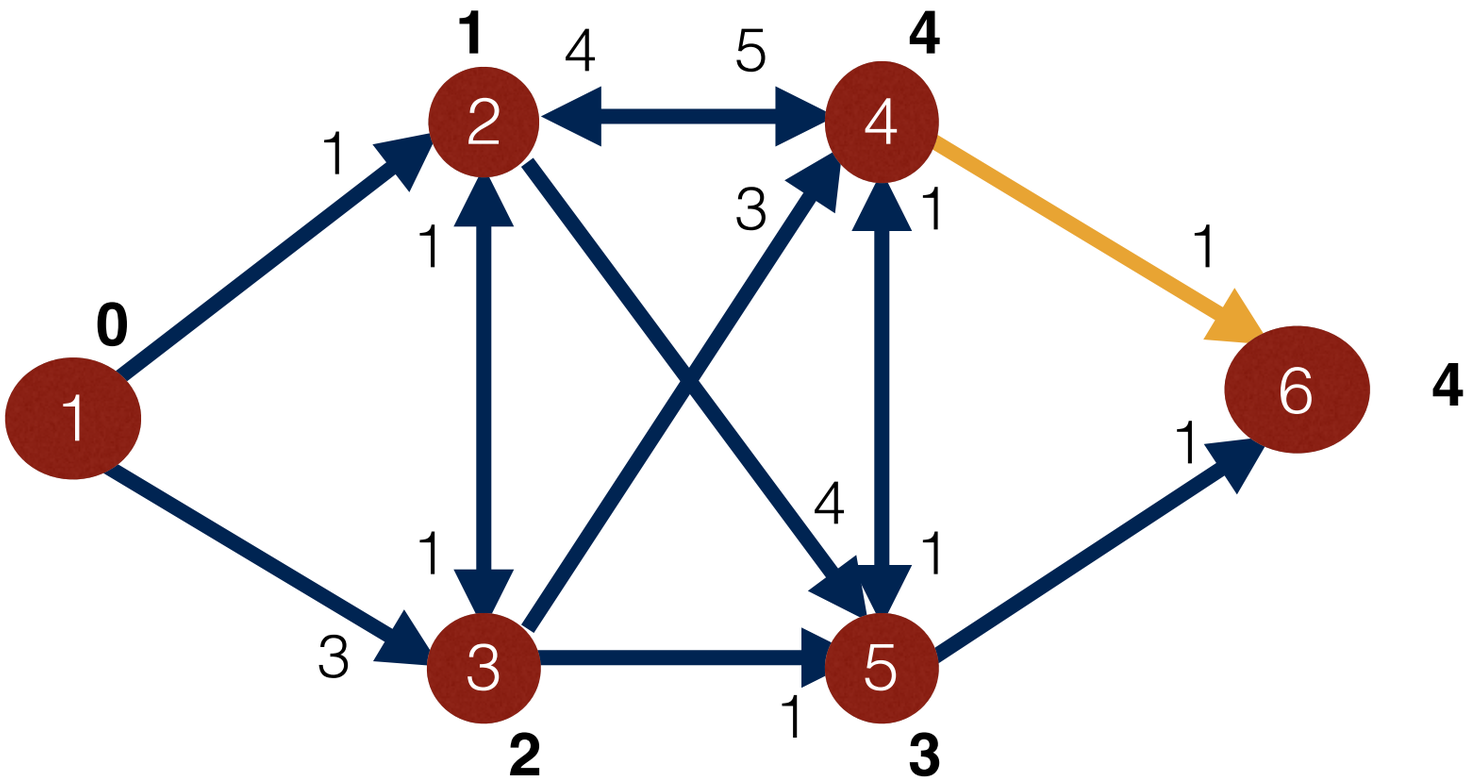
Shortest path routing - Dijkstra



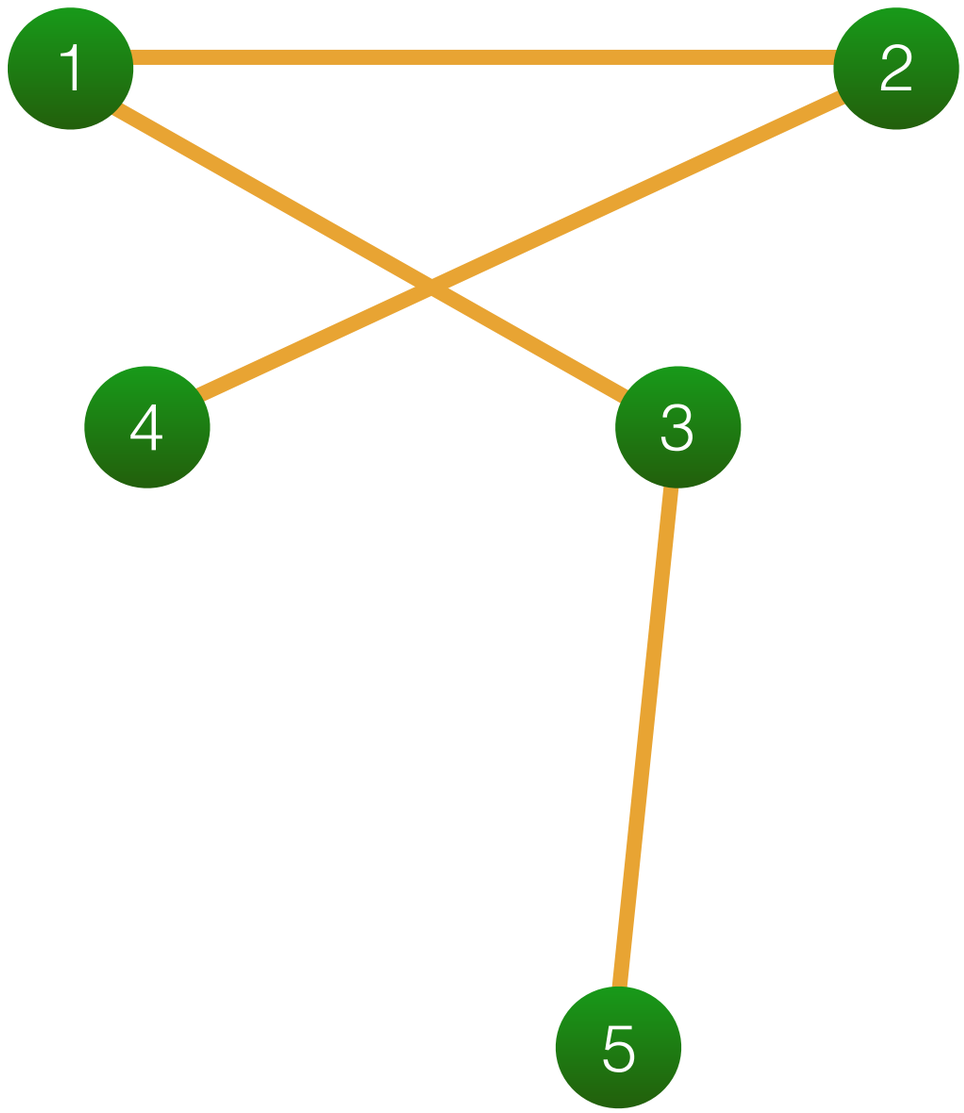
Update 4 and 6

Shortest path routing - Dijkstra

Select both 4 and 6 as certain - stop



Diameter



$$D = \max_{a,b} d(a,b)$$

$$D = 4$$

$$A = \frac{\sum_{a \neq b} d(a,b)}{N^2 - N}$$

Closeness centrality

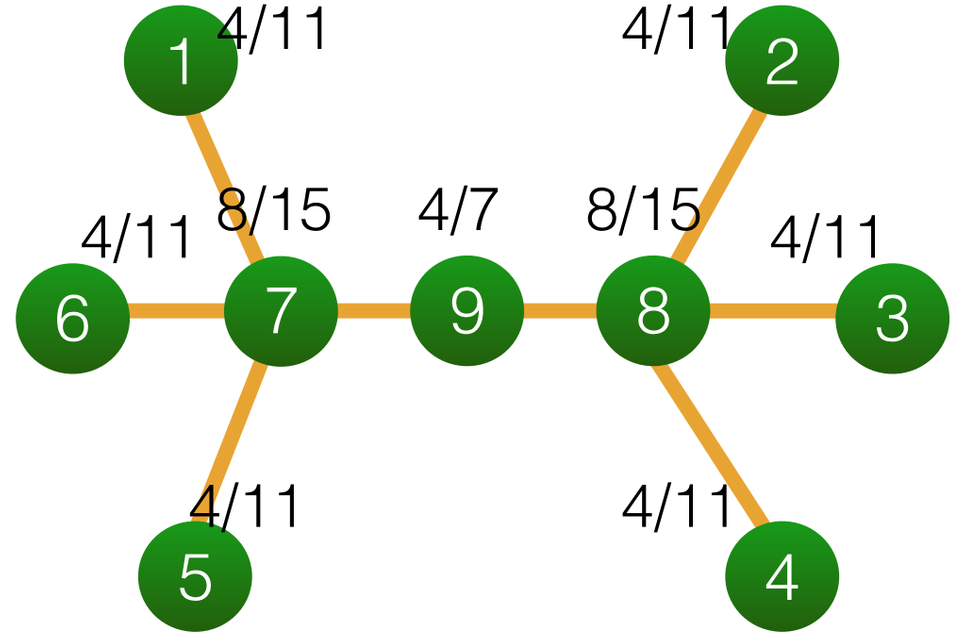
$$N = \{a_i, i = 1..n\},\ E = \{(a_j, b_j)\}$$

$$\sum_{b \in N, b \neq a} d_{a,b}$$

$$\text{closeness\_centrality}(a) = \frac{1}{\sum_{b \in N} d_{a,b}}$$

$$\text{closeness\_centrality}(a) = \frac{n-1}{\sum_{b \in N, b \neq a} d_{a,b}}$$

$$\text{closeness\_centrality}(a) = \frac{n-1}{\sum_{b \in N, b \neq a} d_{b,a}}$$



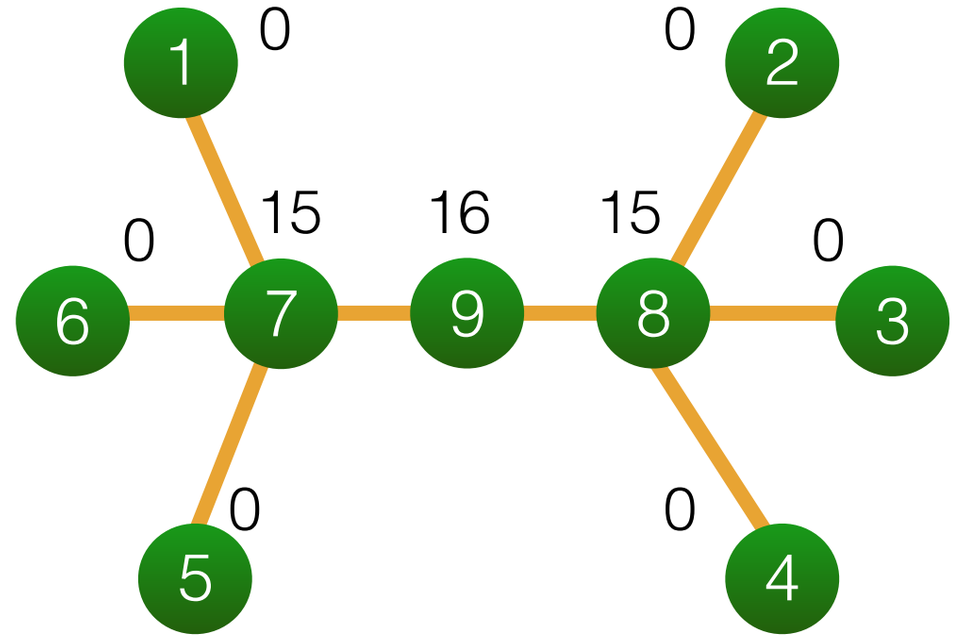
Index of expected time until arrival for given node of whatever is flowing through the network
Gossip network: central player hears things first

Betweenness centrality

$$N = \{a_i, i = 1..n\},\ E = \{(a_j, b_j)\}$$

$$\text{betweenness\_centrality}(a) = \sum_{b,c \in N, b \neq a \neq c} \frac{|g_{b,c} : a \in g_{b,c}|}{|g_{b,c}|}$$

$$\frac{1}{(n-1)(n-2)}$$

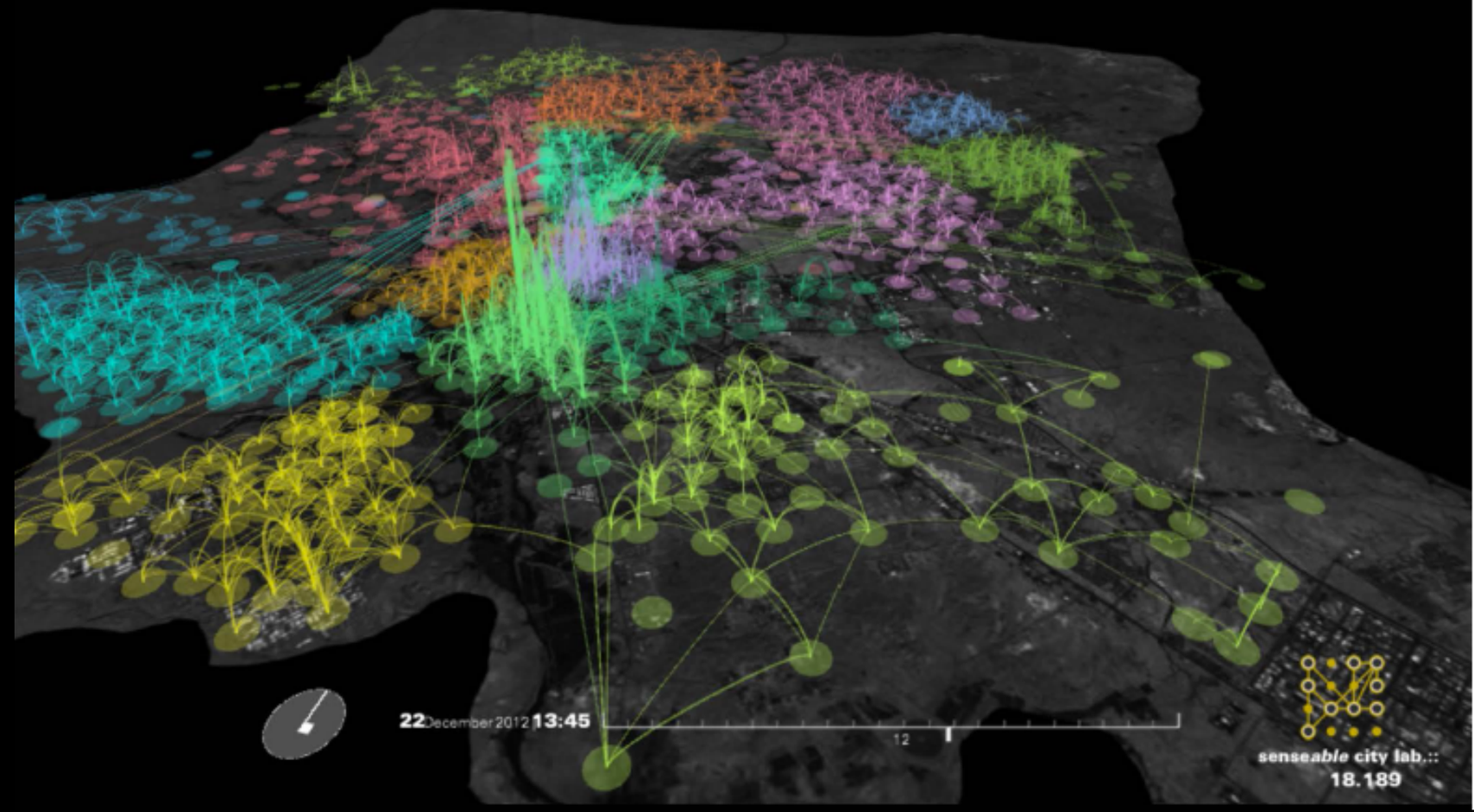


Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

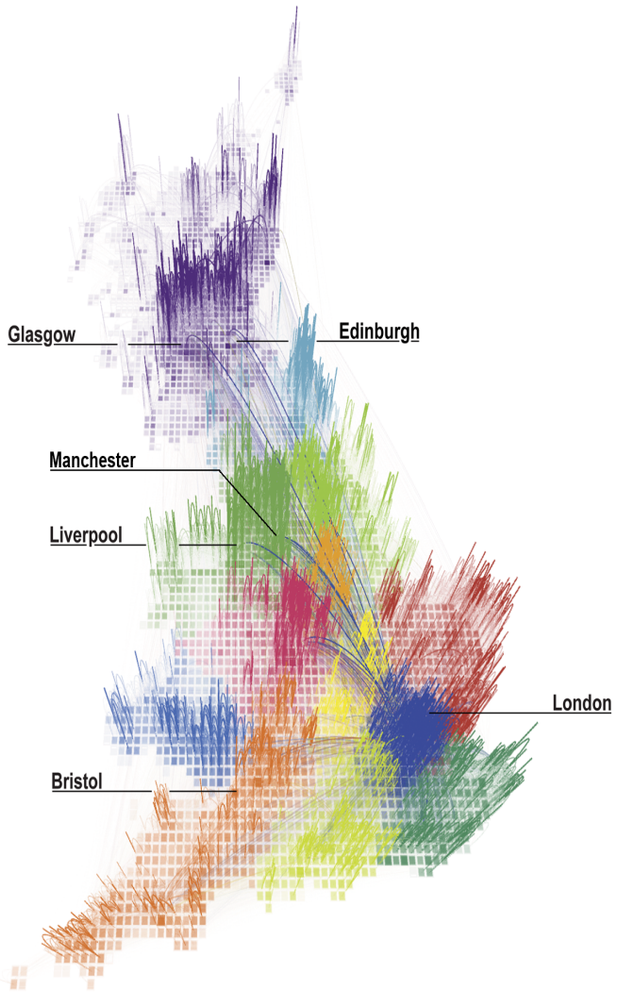
Community detection

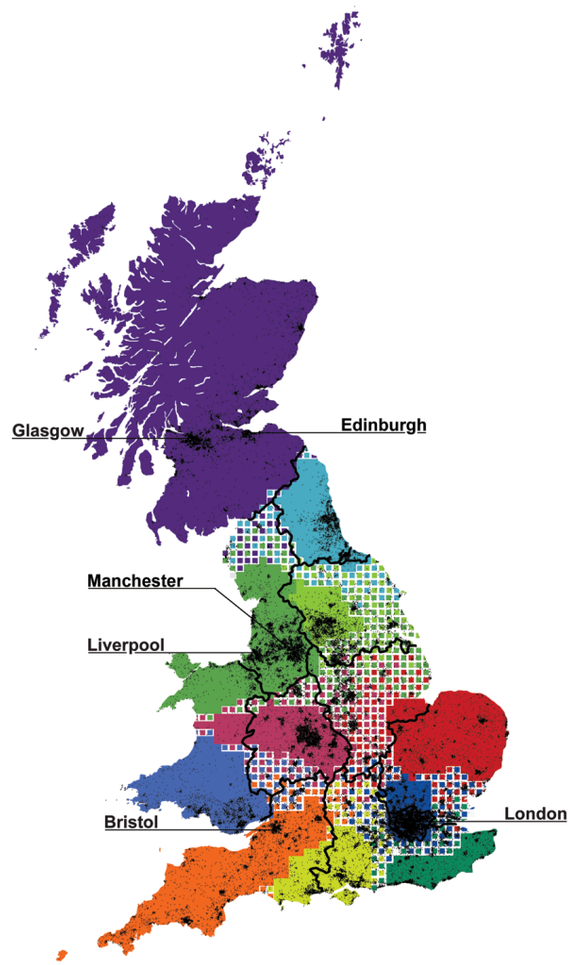
Discovering underlying structure in terms of heterogeneity of connections:
some groups of nodes are connected more strongly

- Social networks
- Biological networks
- Human mobility networks
- Networks of human interactions



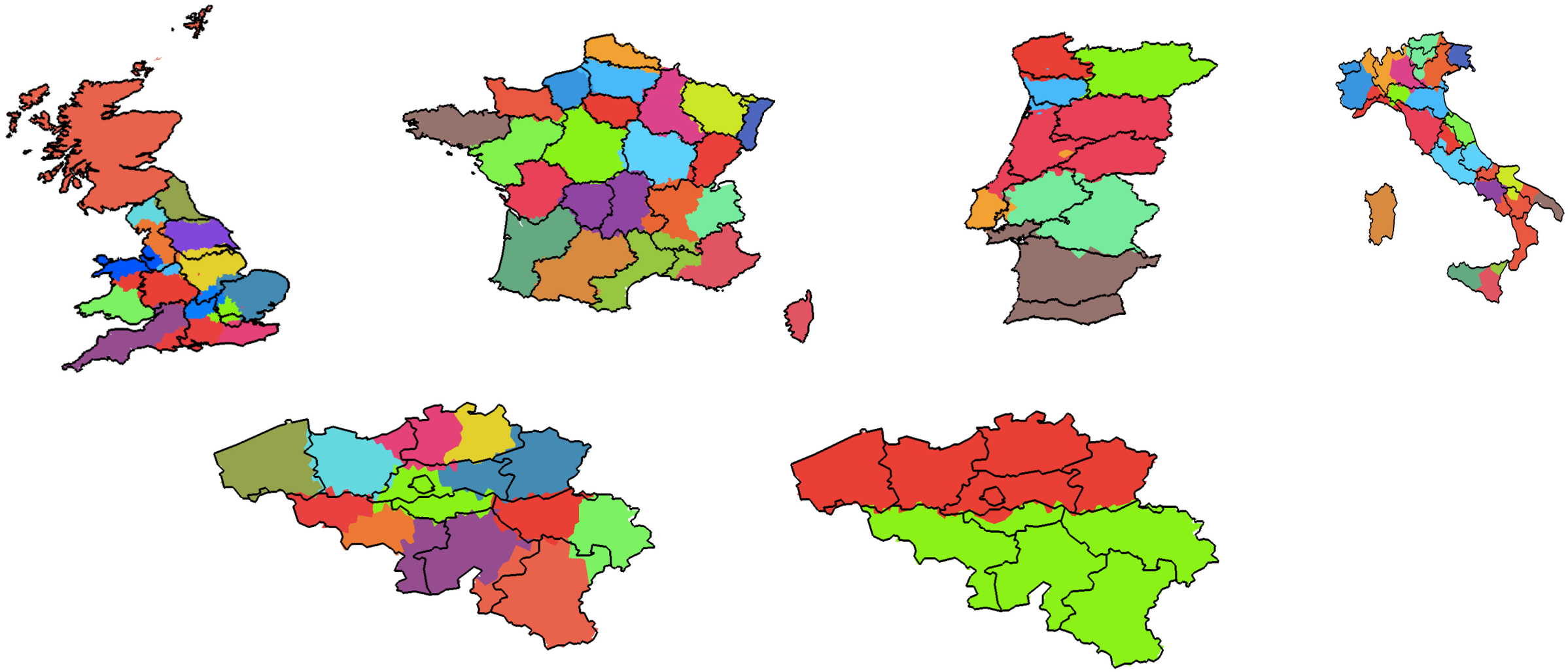
Regional delineation





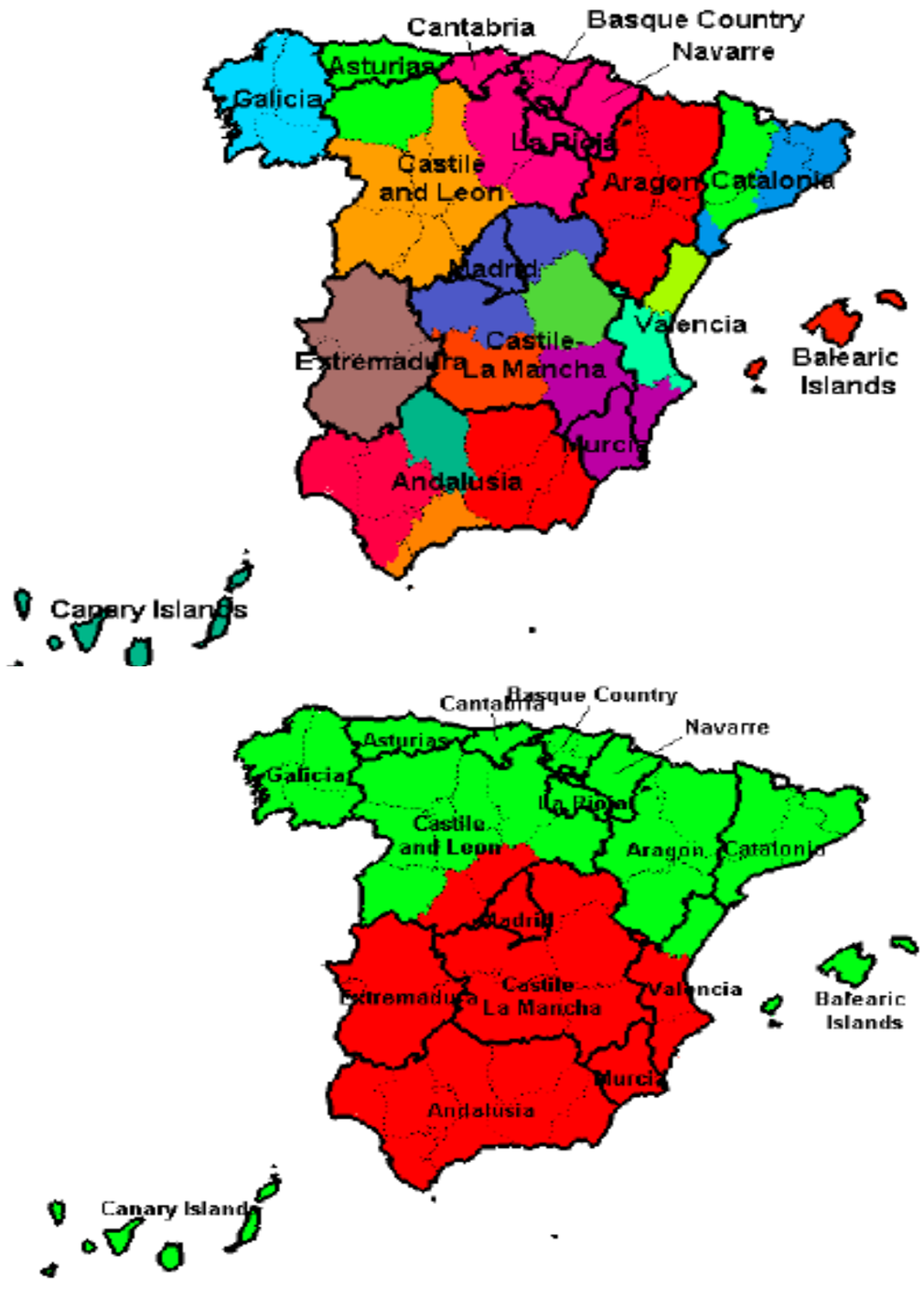
Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., ... & Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PloS one*, *5*(12), e14248.

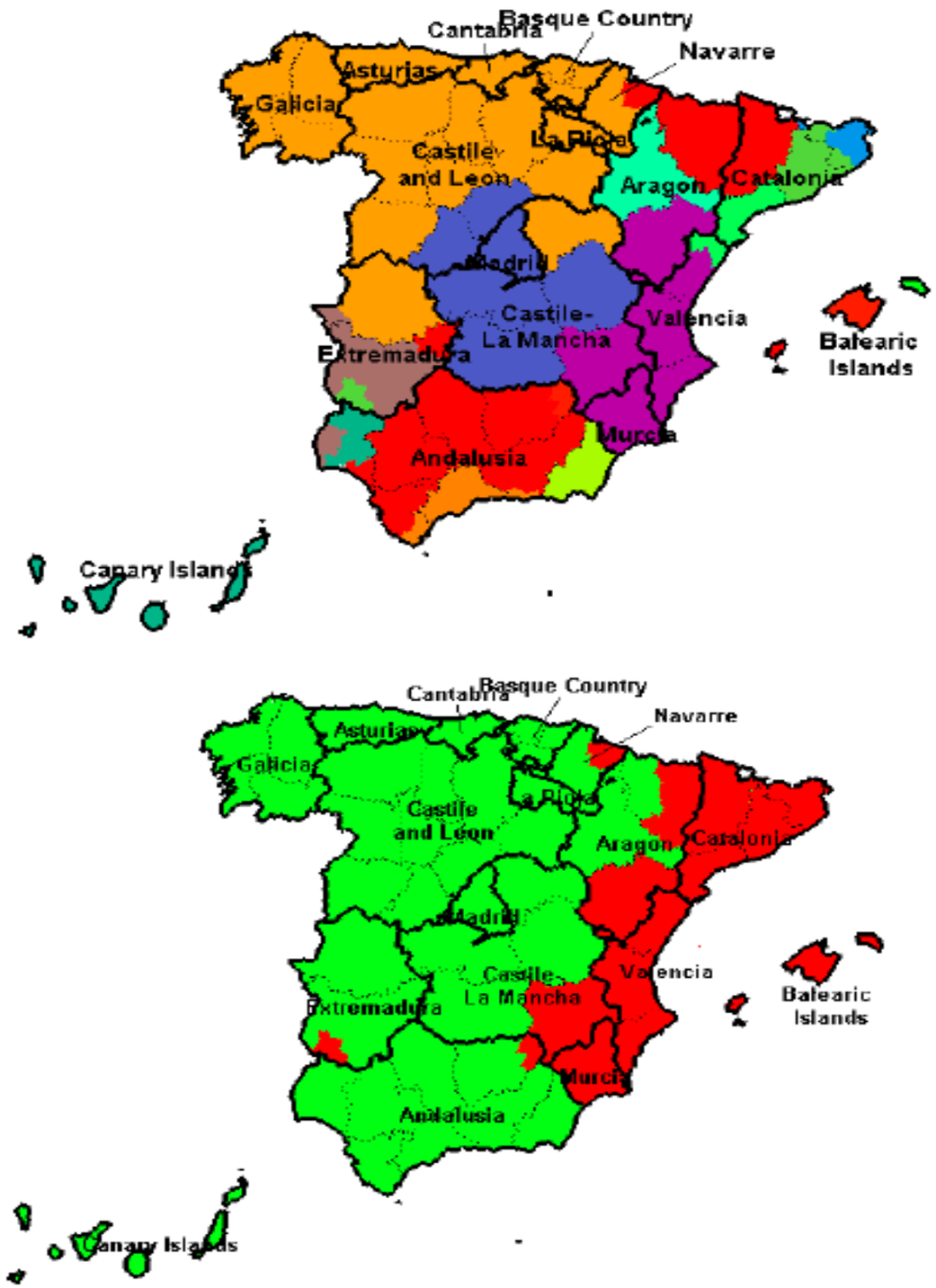
Regional delineation



Sobolevsky S., Szell M., Campari R., Couronne T., Smoreda Z., Ratti C. (2013) Delineating geographical regions with networks of human interactions in an extensive set of countries. PLoS ONE 8 (12), e81707

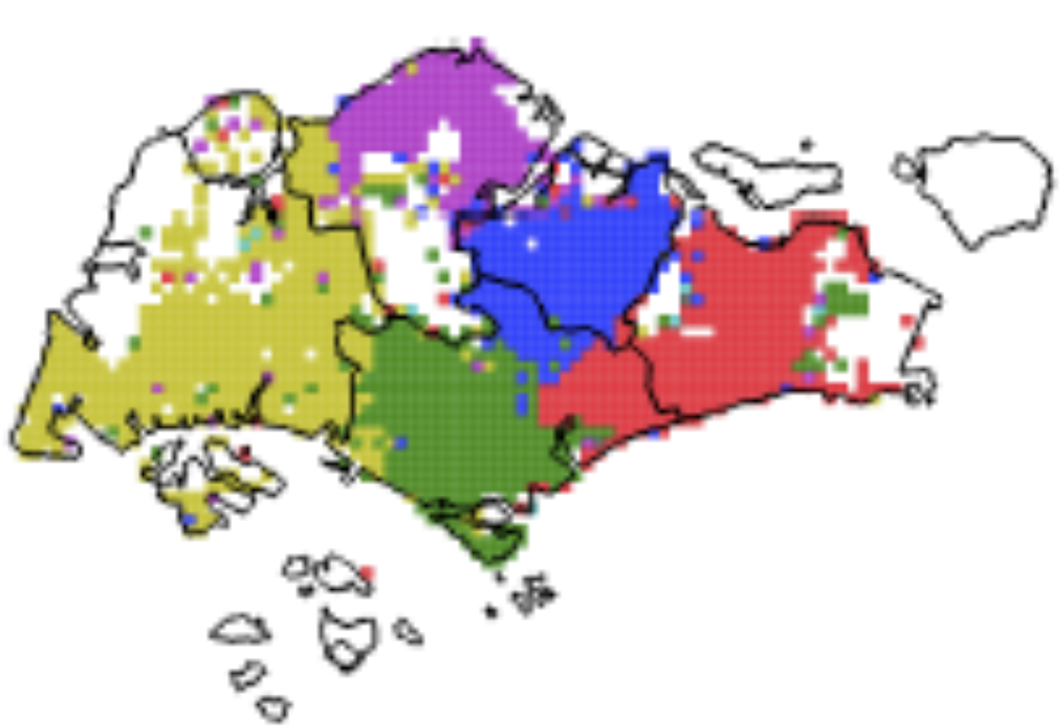
Credit cards

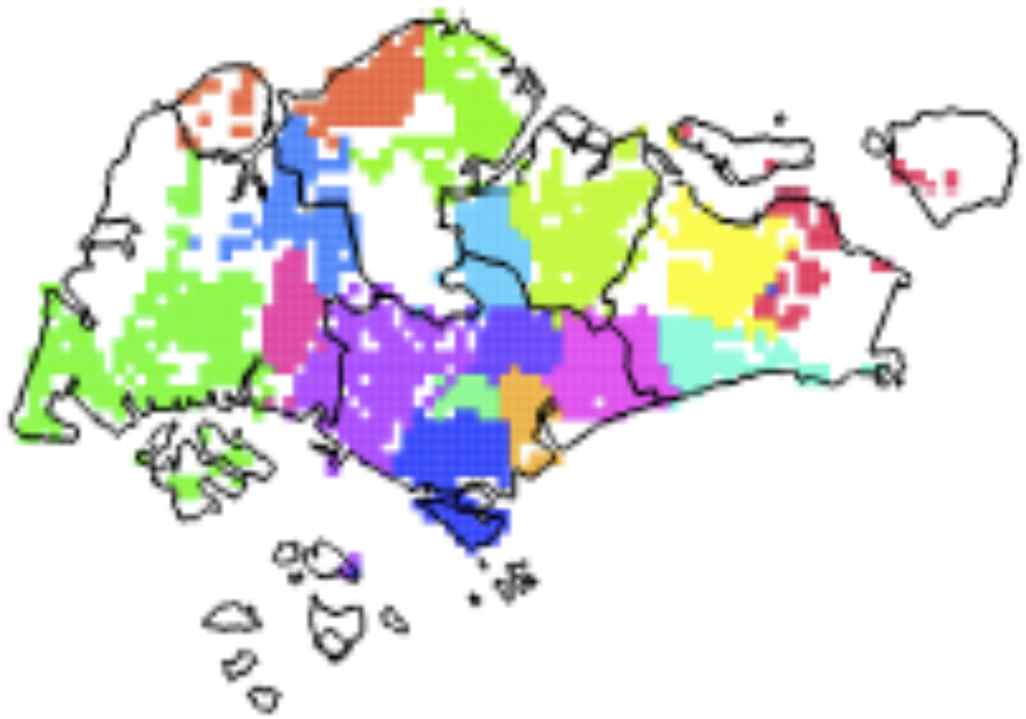




Sobolevsky, Stanislav, et al. "Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain." *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE, 2014.

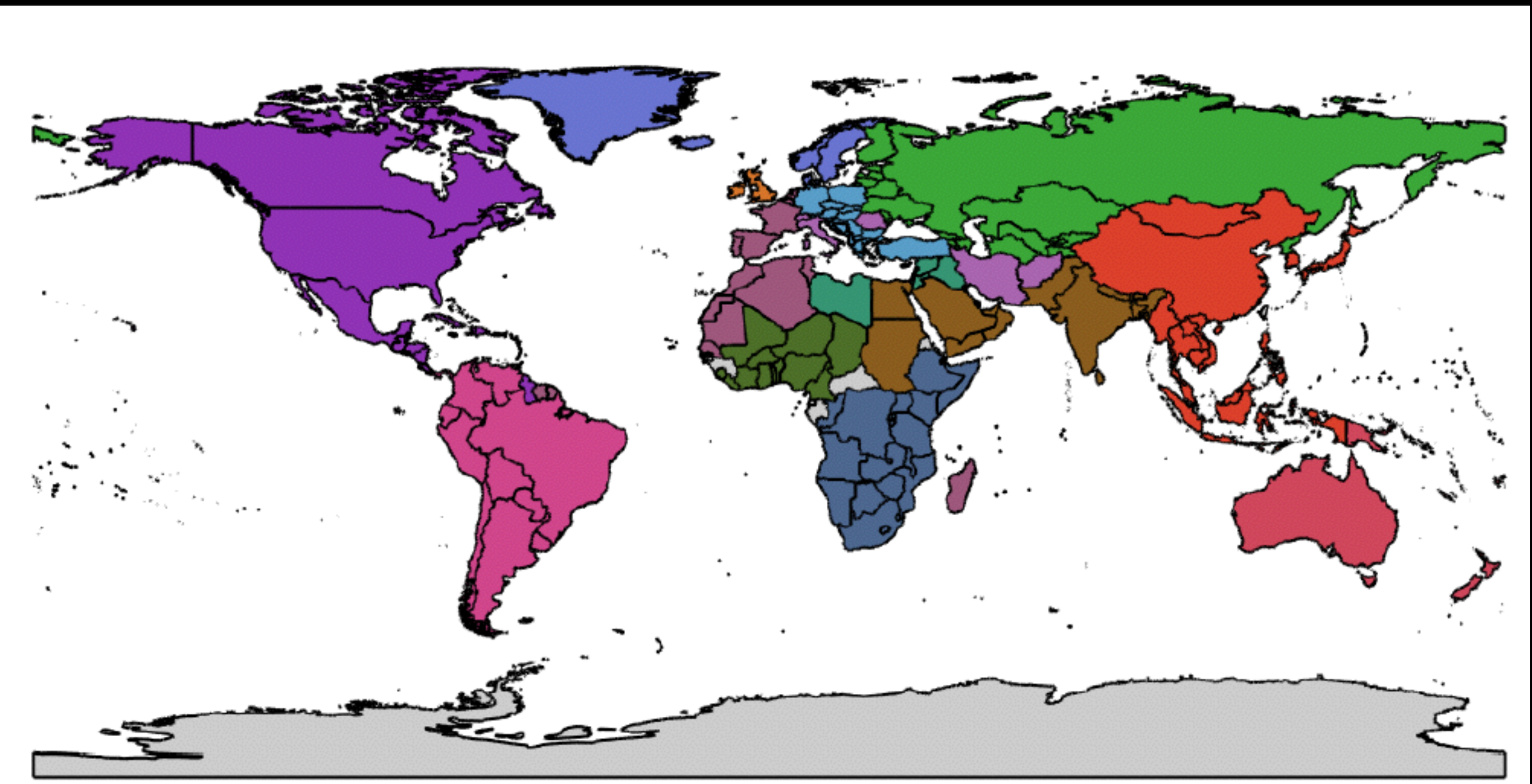
Taxi trips





Kang, C., Sobolevsky, S., Liu, Y., & Ratti, C. (2013, August). Exploring human movements in Singapore: A comparative analysis based on mobile phone and taxicab usages. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing* (p. 1). ACM.

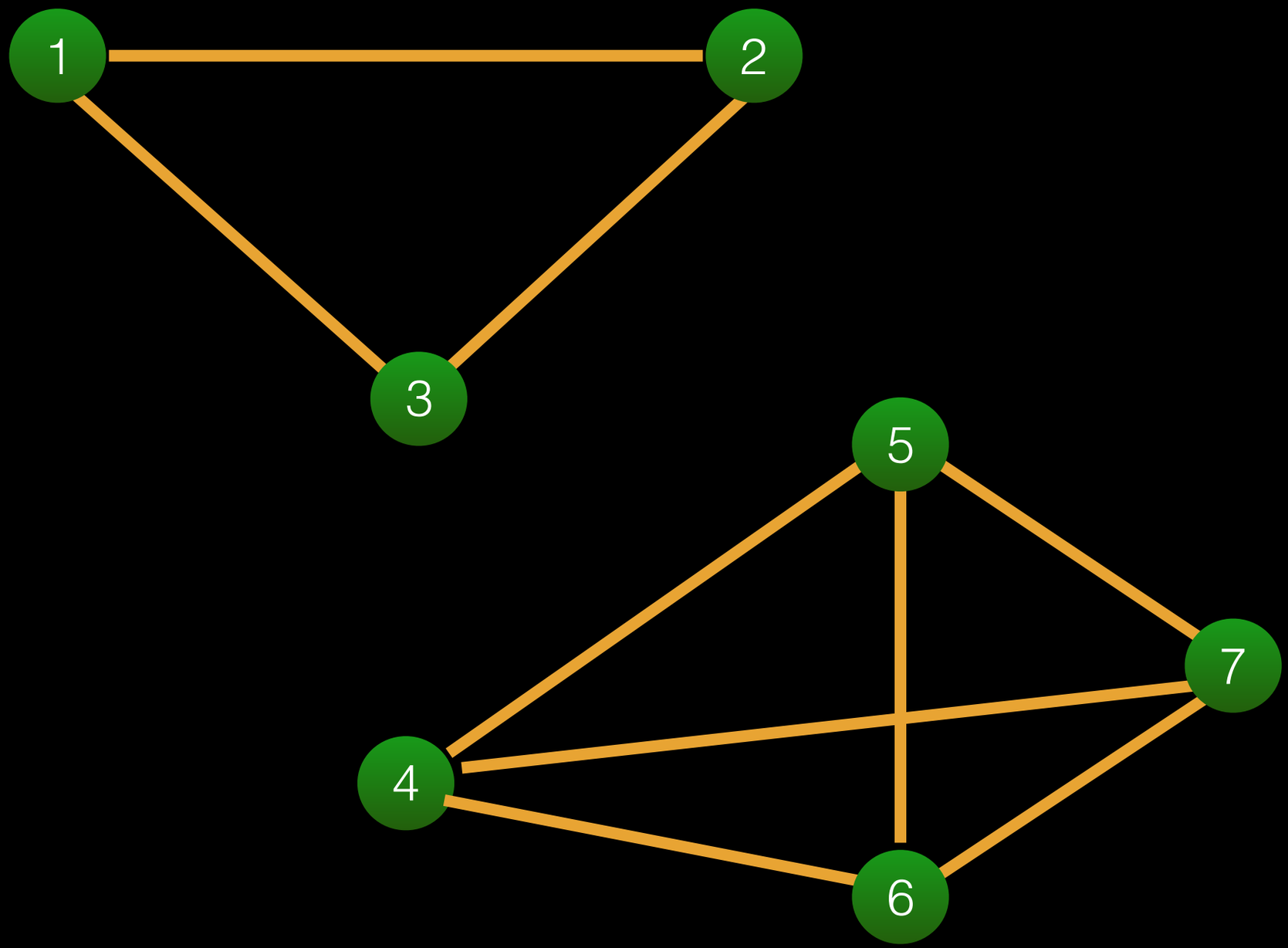
Global: social media



Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260-271.
Belyi A. Bojic I, Sobolevsky S., Sitko I., Hawelka B., Ratti C. Global multi-layer network of human mobility. Submitted

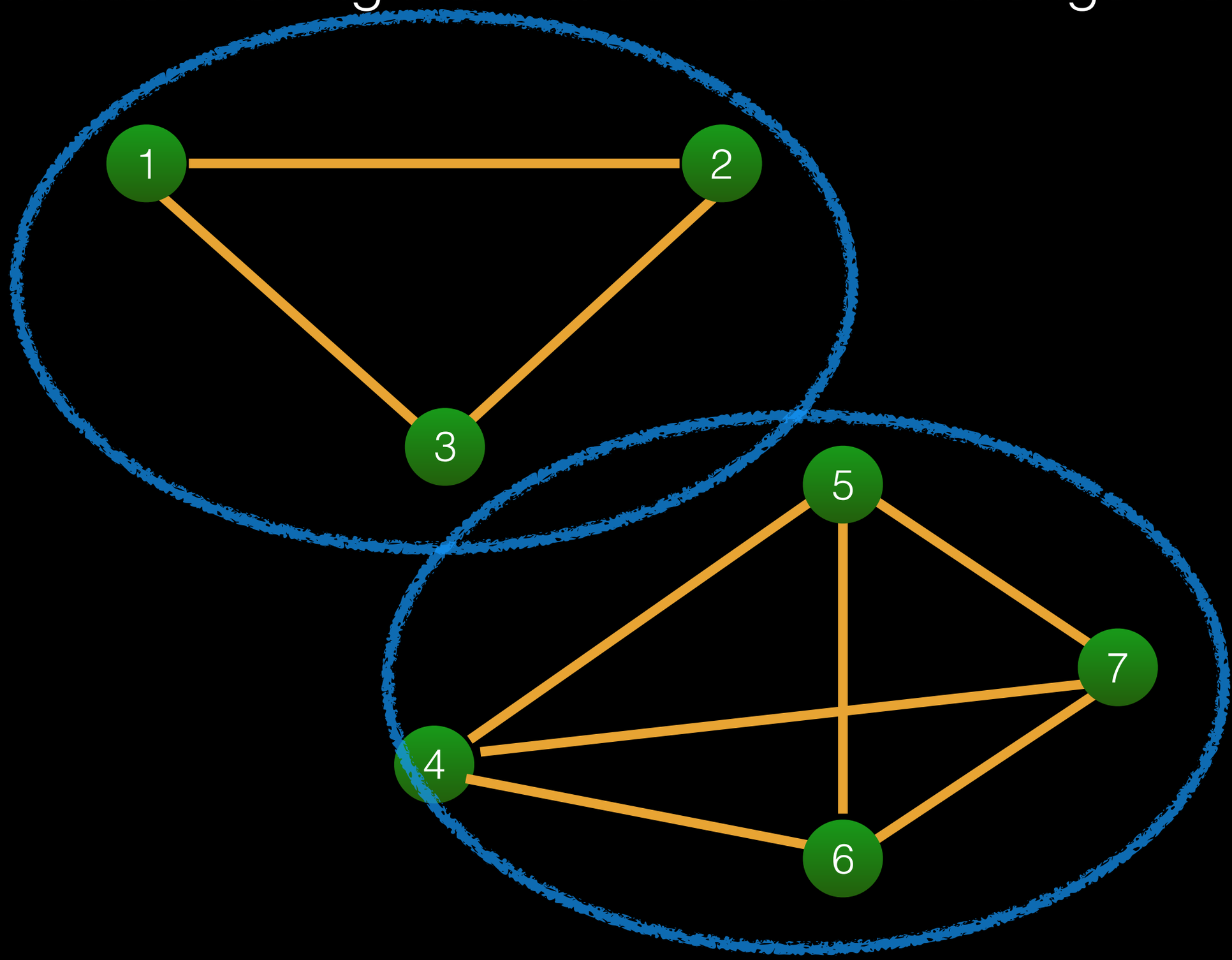
Community

Group of nodes having internal connections stronger vs external



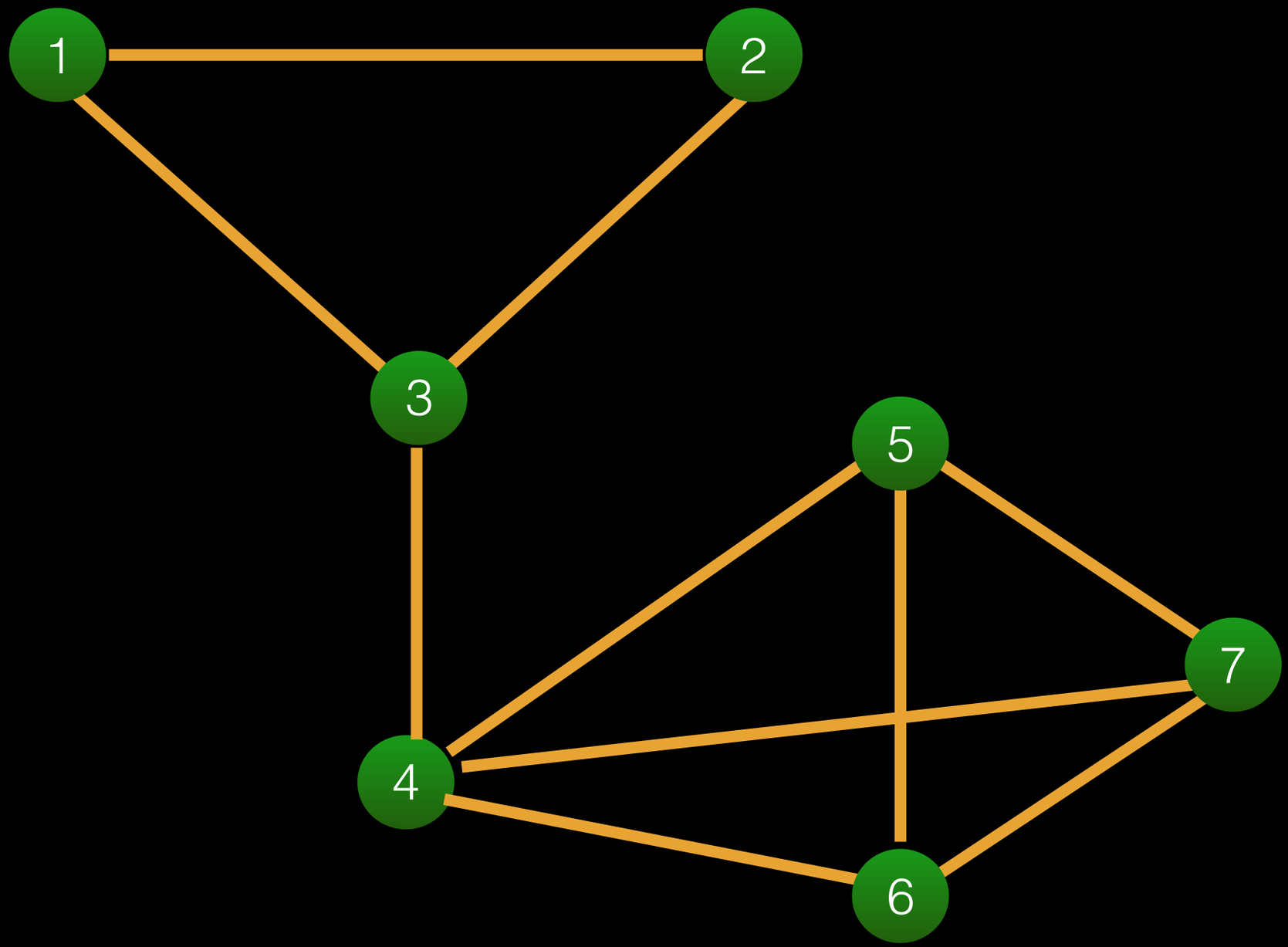
Community

Group of nodes having internal connections stronger vs external



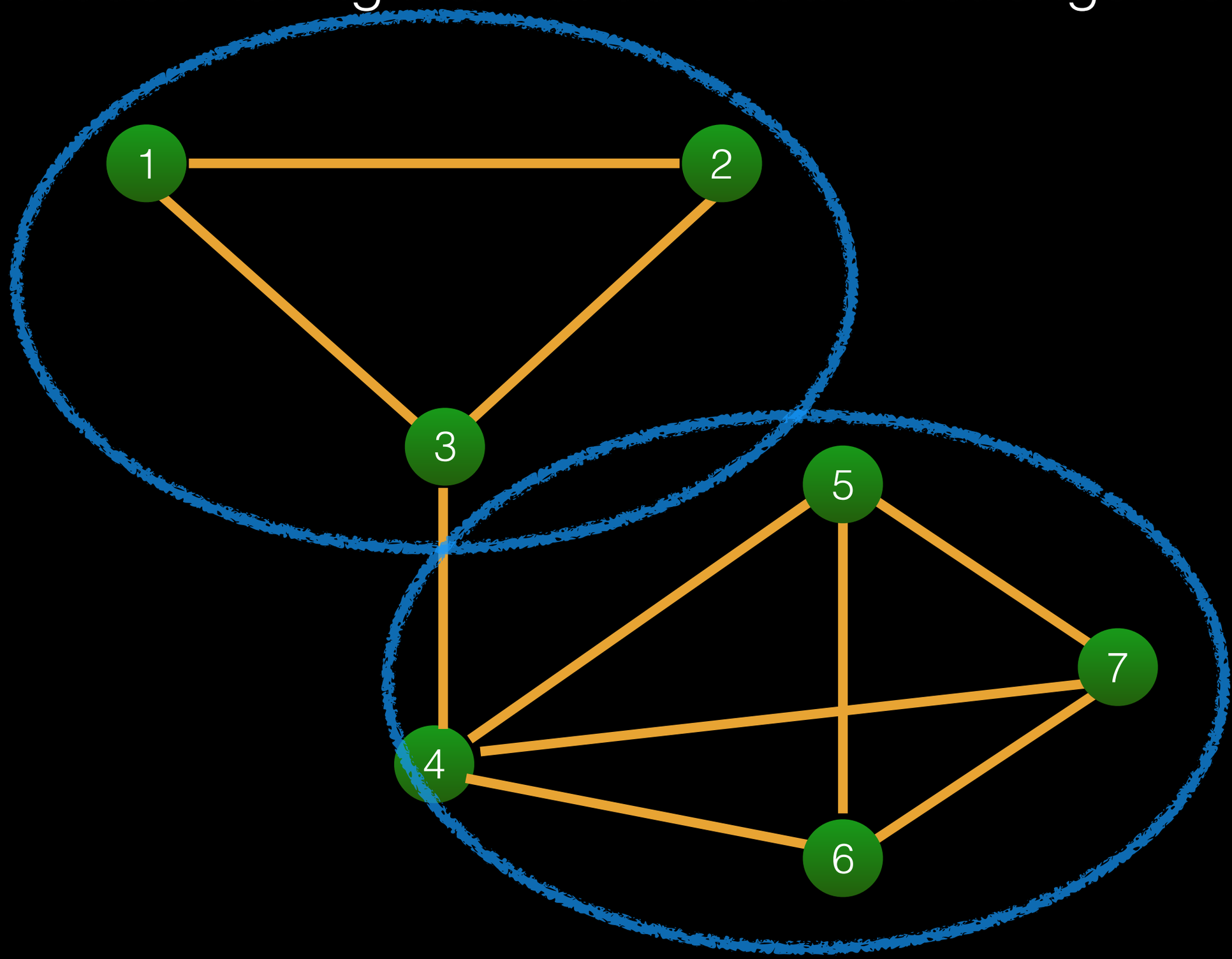
Community

Group of nodes having internal connections stronger vs external



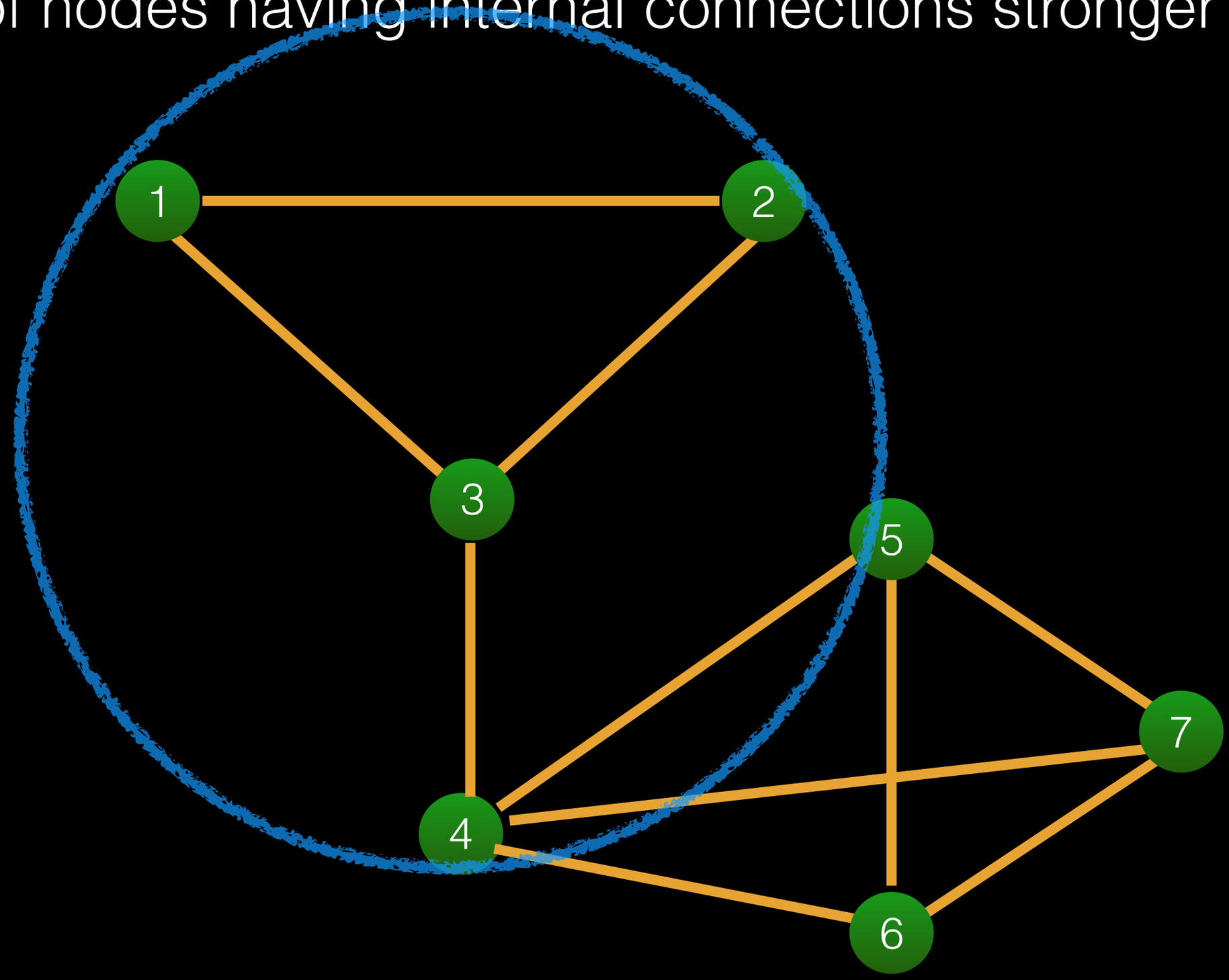
Community

Group of nodes having internal connections stronger vs external



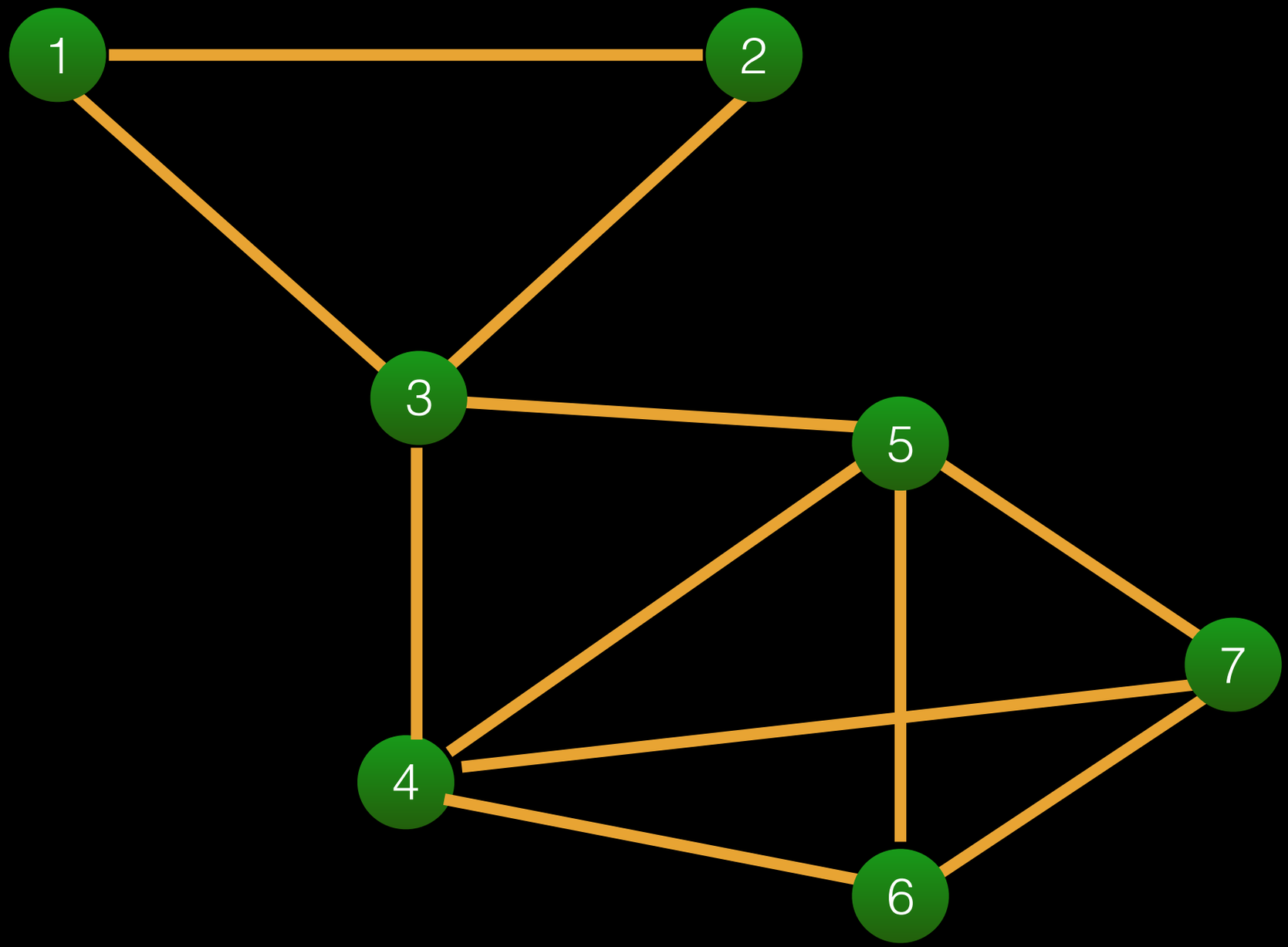
Community

Group of nodes having internal connections stronger vs external



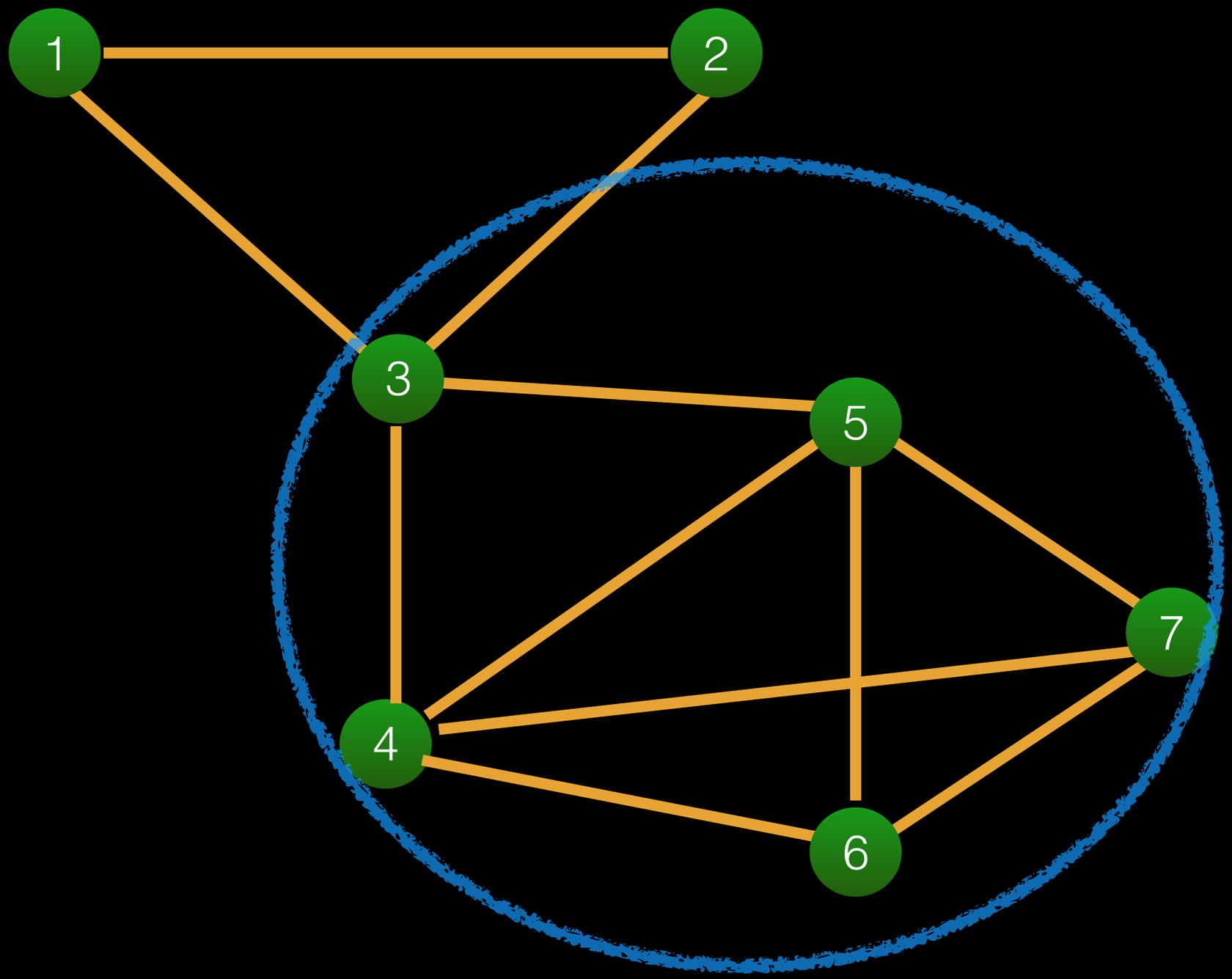
Community

Group of nodes having internal connections stronger vs external



Community

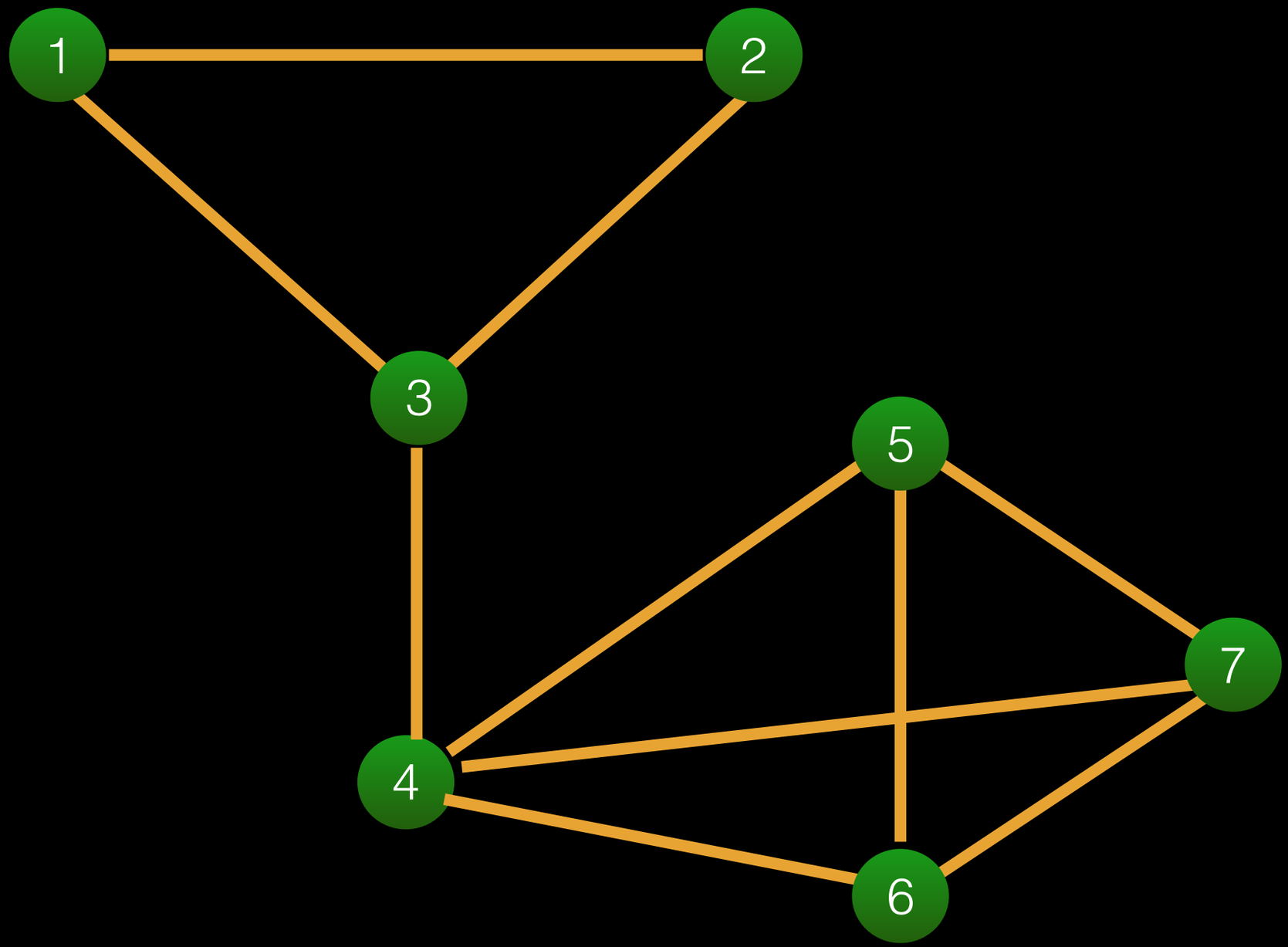
Group of nodes having internal connections stronger vs external



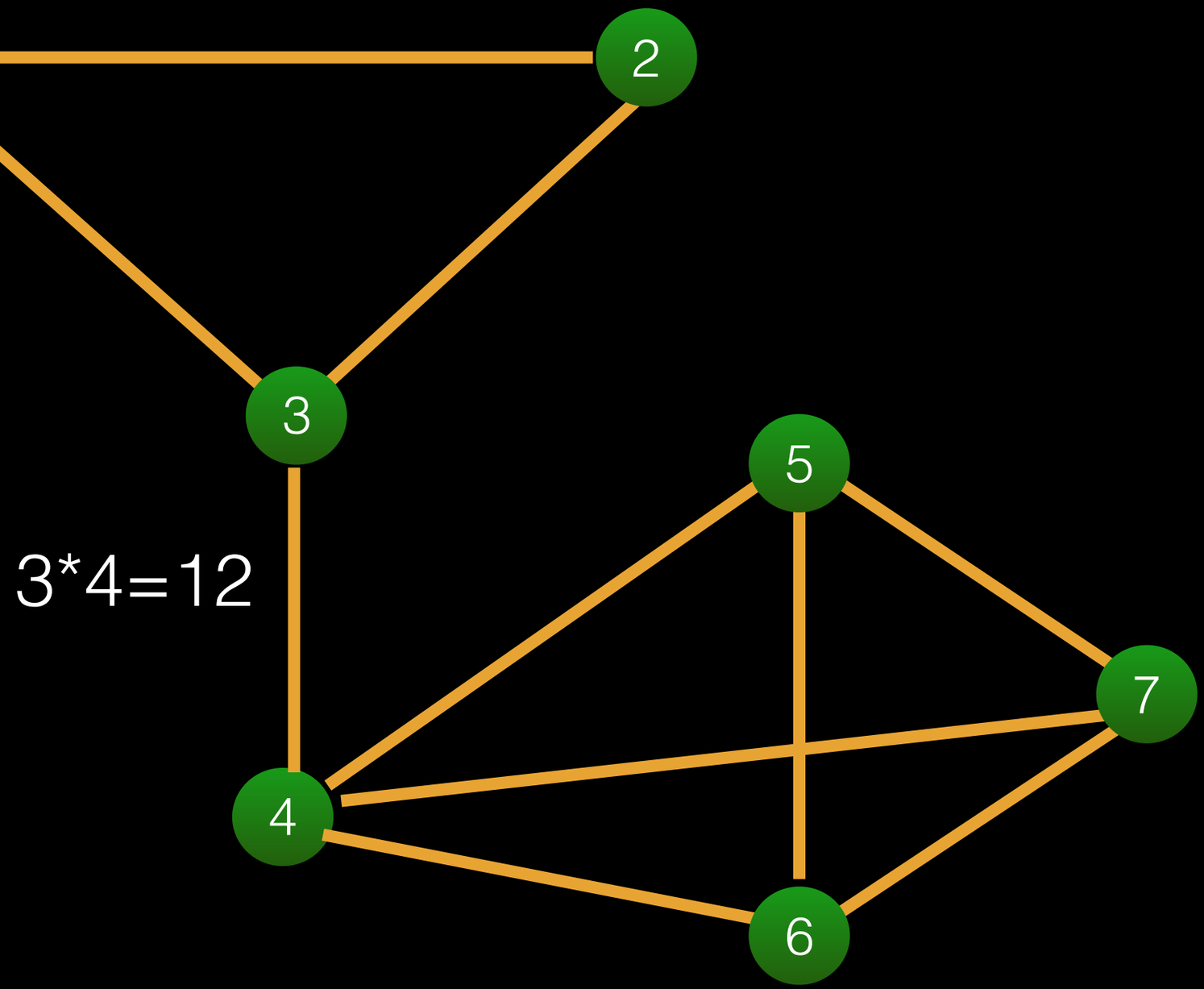
Community detection

- Straightforward algorithms
  - Girvan-Newman
  - Hierarchical clustering
- Optimization algorithms (objective function)
  - Modularity optimization
  - Infomap
  - Block-model

Girvan-Newman algorithm, 2002

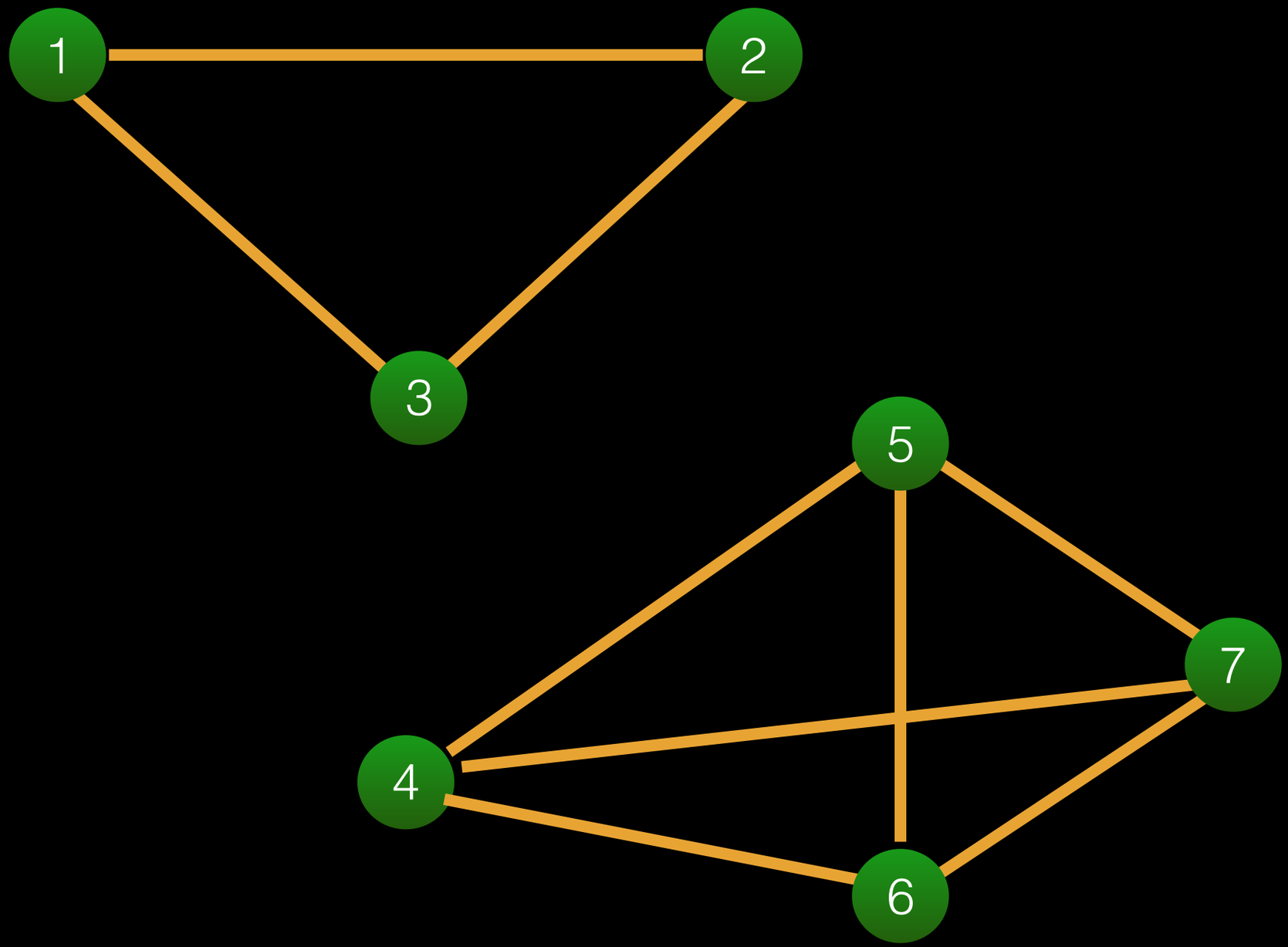


Girvan-Newman algorithm, 2002



3*4=12

Girvan-Newman algorithm, 2002



Girvan-Newman algorithm, 2003

1. Compute betweenness of all the edges

2. Remove the edge with highest betweenness and recalculate betweenness of all the remaining ones

3. Is the number of connected components lower than the target number?

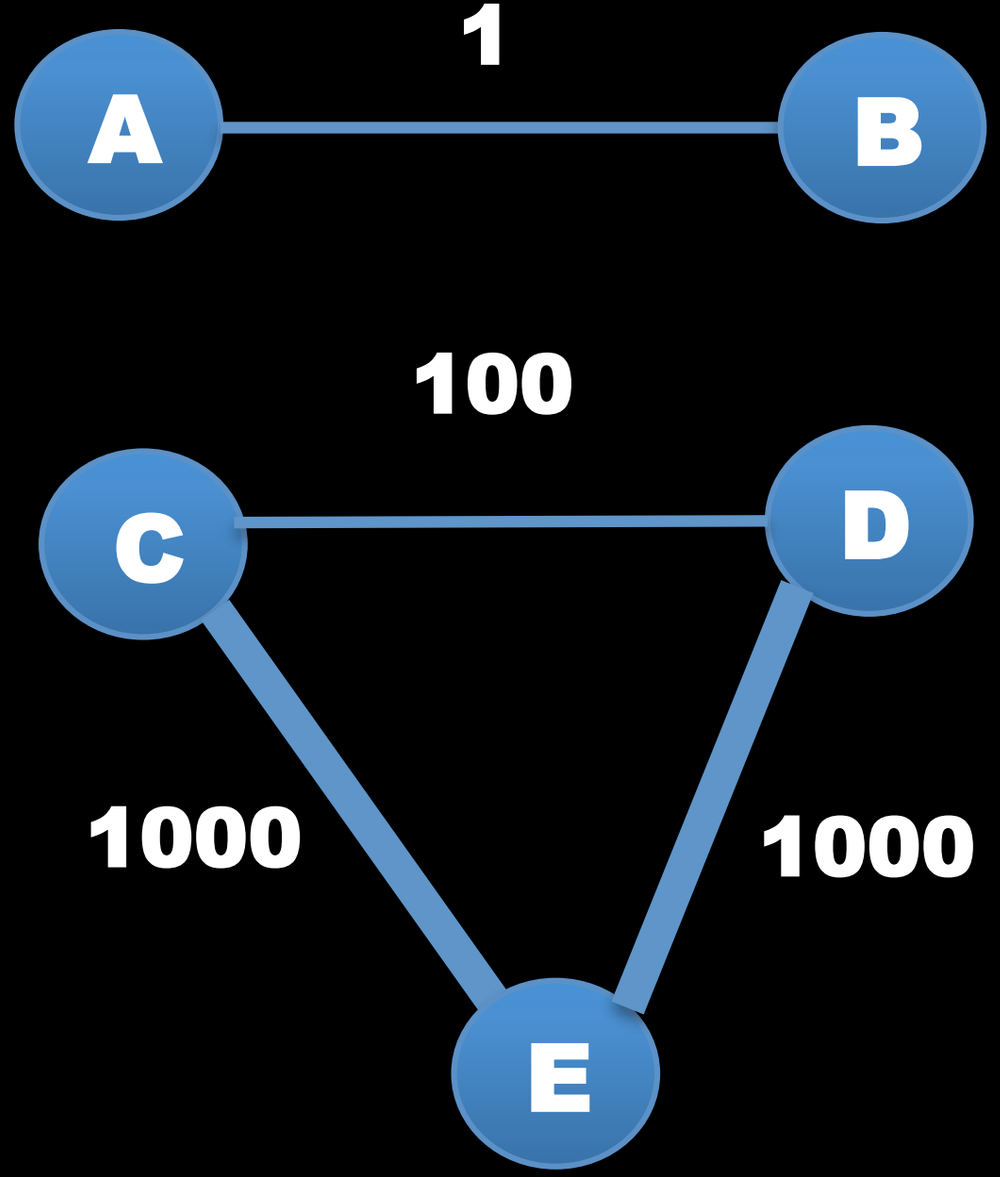
Yes

No

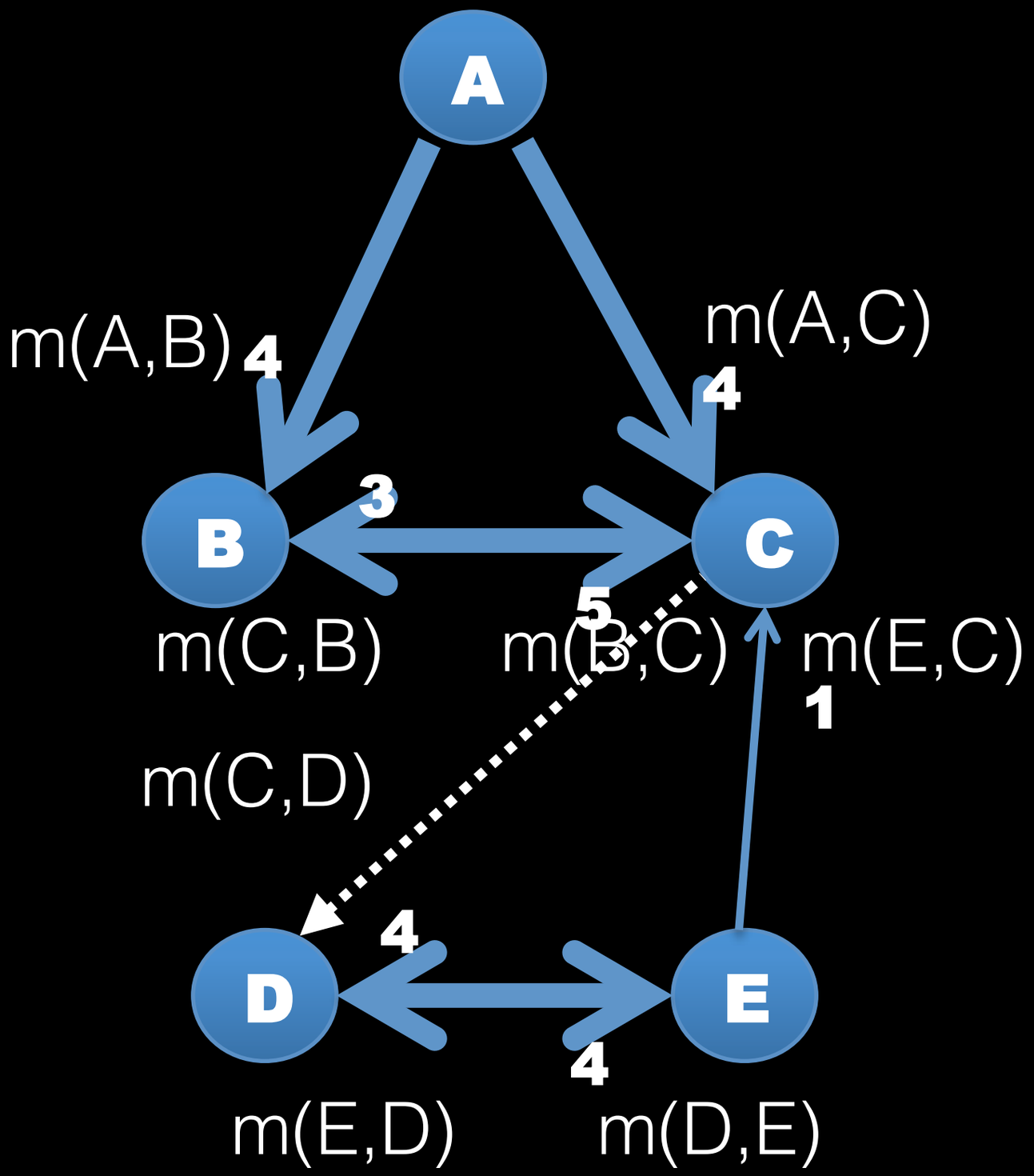
4. Take connected components of the network as the resulting communities

Modularity

Motivation - absolute value of edge weight is not enough to judge on the strength of the connection

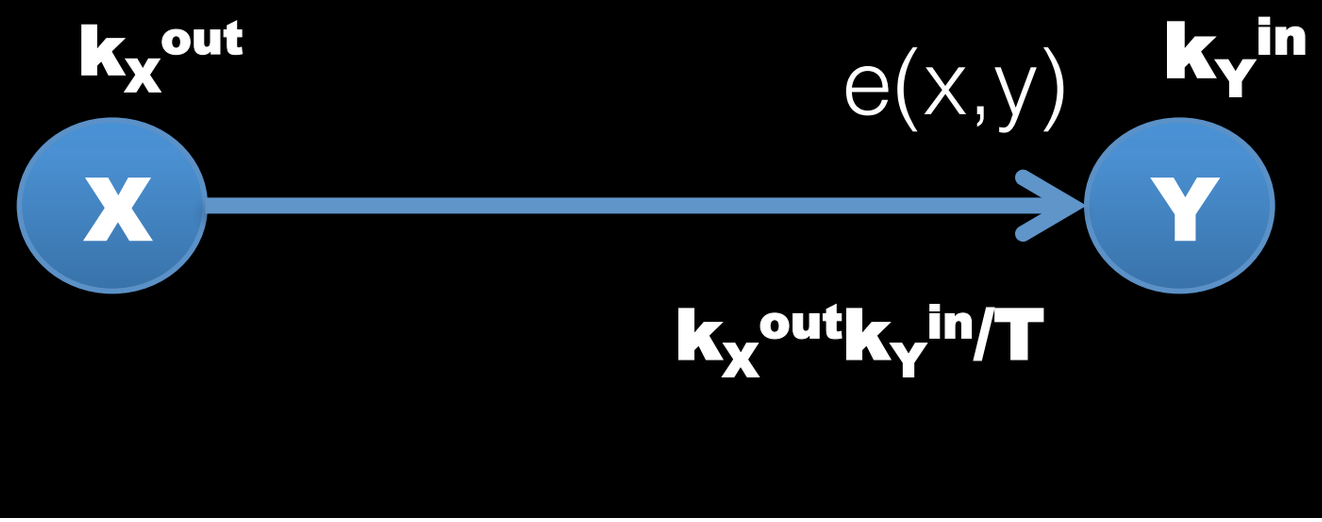


Modularity



Given the network and edge weights, provide an estimate of how relatively strong they are

Modularity



$$q(x,y) = \frac{e(x,y)}{T} - \frac{k_x^{out} k_y^{in}}{T^2}$$

Modularity

$$P = (c_x, x \in N)$$

$$Q(P) = \sum_{x,y,\ c_x=c_y} q(x,y)$$

$$Q(P) = \sum_{x,y,\ c_x=c_y} \left[\frac{e(x,y)}{T} - \frac{k_x^{out} k_y^{in}}{T^2} \right]$$

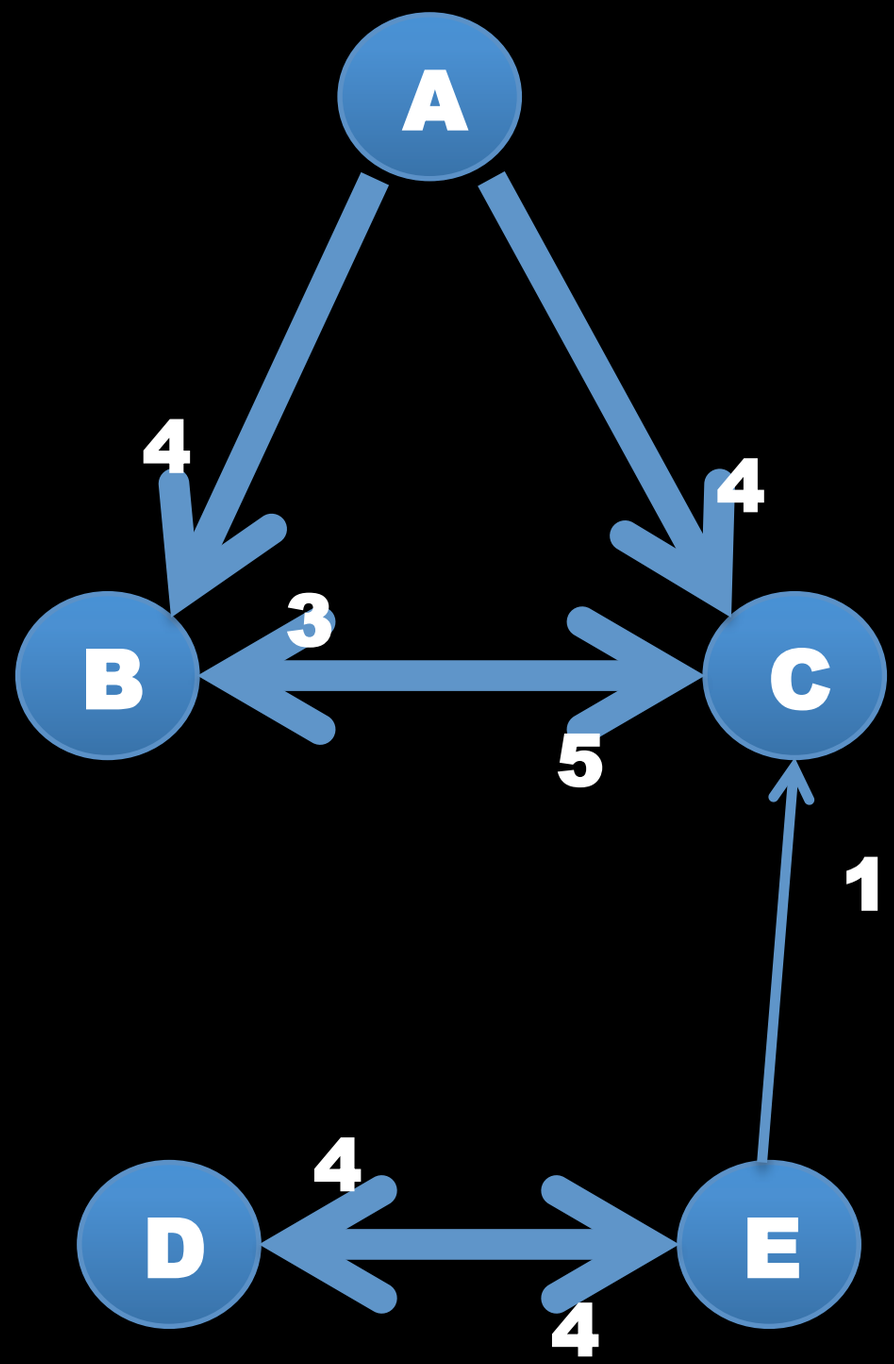
$$Q(P) = \sum_{x,y,\ c_x=c_y} \left[\frac{e(x,y)}{T} - \frac{k_x k_y}{T^2} \right]$$

Modularity

$$-1 = -\sum_{x,y} \frac{k_x^{out} k_y^{in}}{T^2} < Q = \sum_{x,y,\ c_x = c_y} \left[\frac{e(x,y)}{T} - \frac{k_x^{out} k_y^{in}}{T^2} \right] < \sum_{x,y} \frac{e(x,y)}{T} = 1$$

$$Q(P_0) = \sum_{x,y} \left[\frac{e(x,y)}{T} - \frac{k_x^{out} k_y^{in}}{T^2} \right] = \sum_{x,y} \frac{e(x,y)}{T} - \sum_{x,y} \frac{k_x^{out} k_y^{in}}{T^2} = 1 - 1 = 0.$$

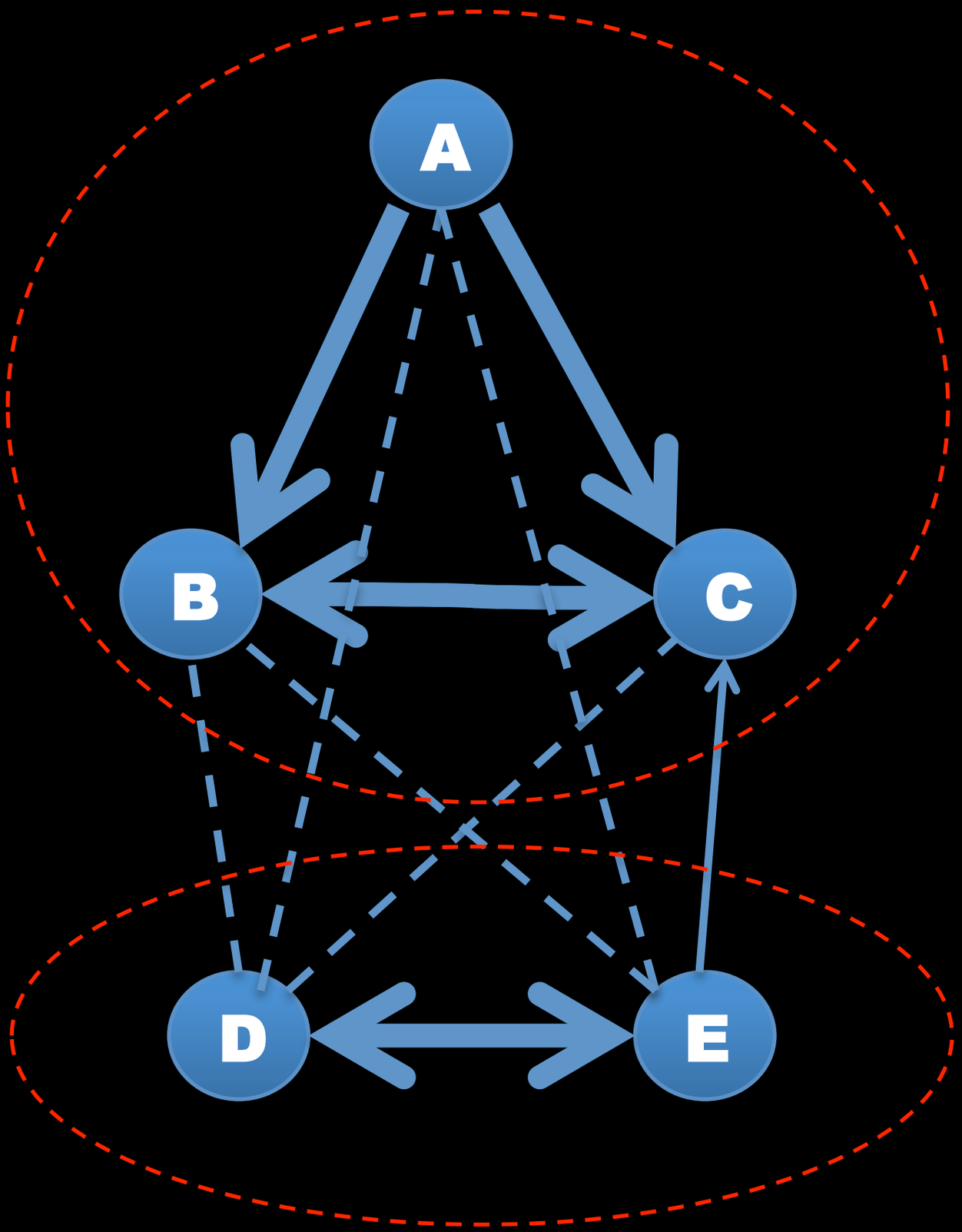
Modularity



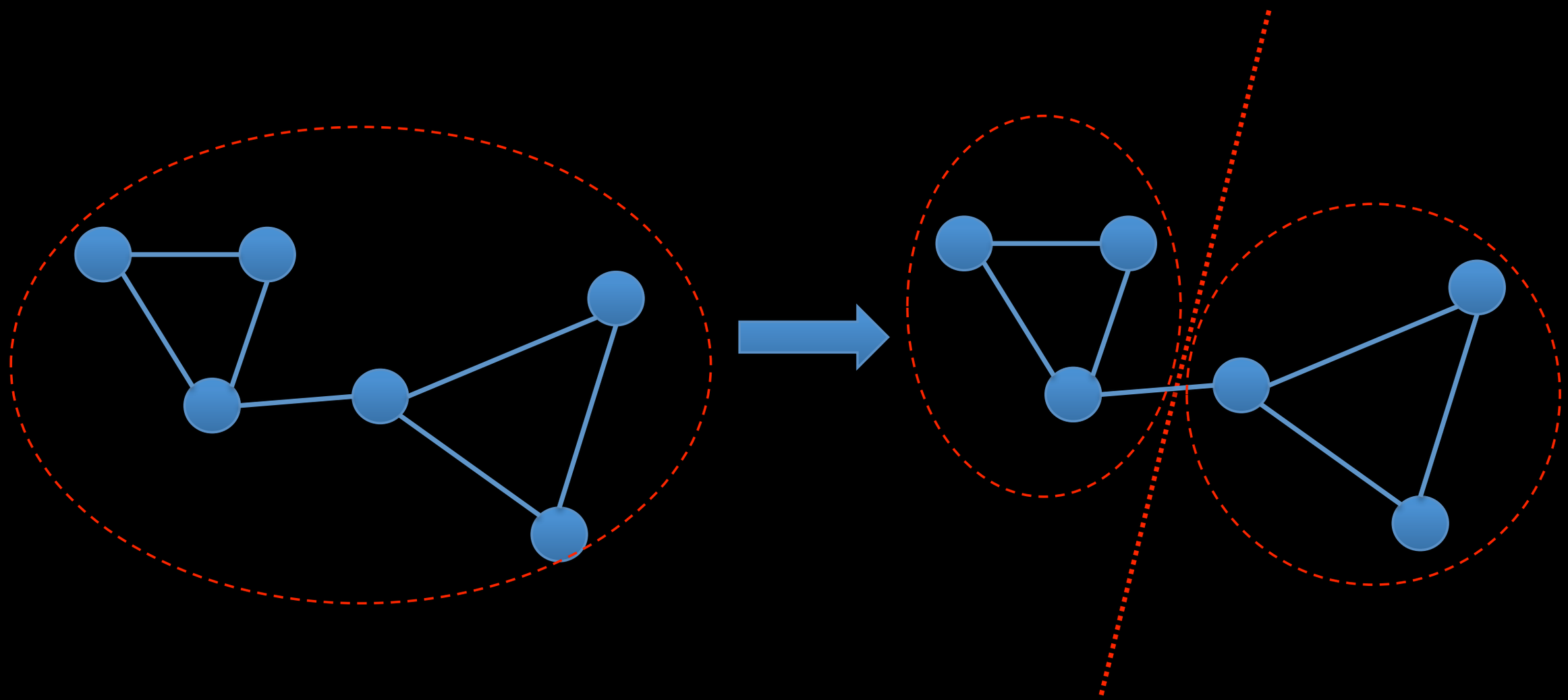
Node	Outgoing weight	Incoming weight
A	8	0
B	5	7
C	3	10
D	4	4
E	5	4
Total	25	25

q(E,C)=1/25-5*10/25/25=1/25-2/25=-1/25

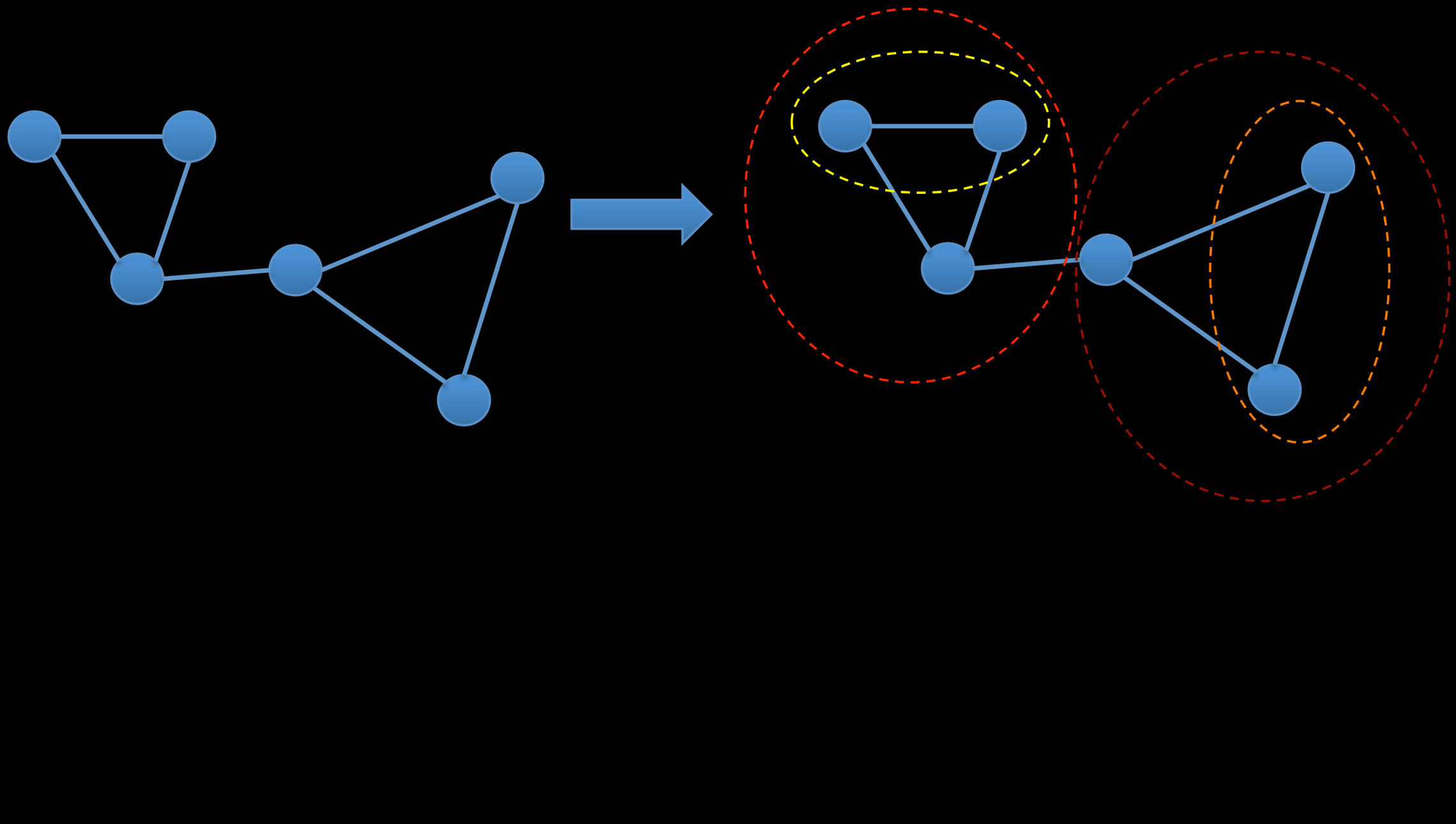
Modularity



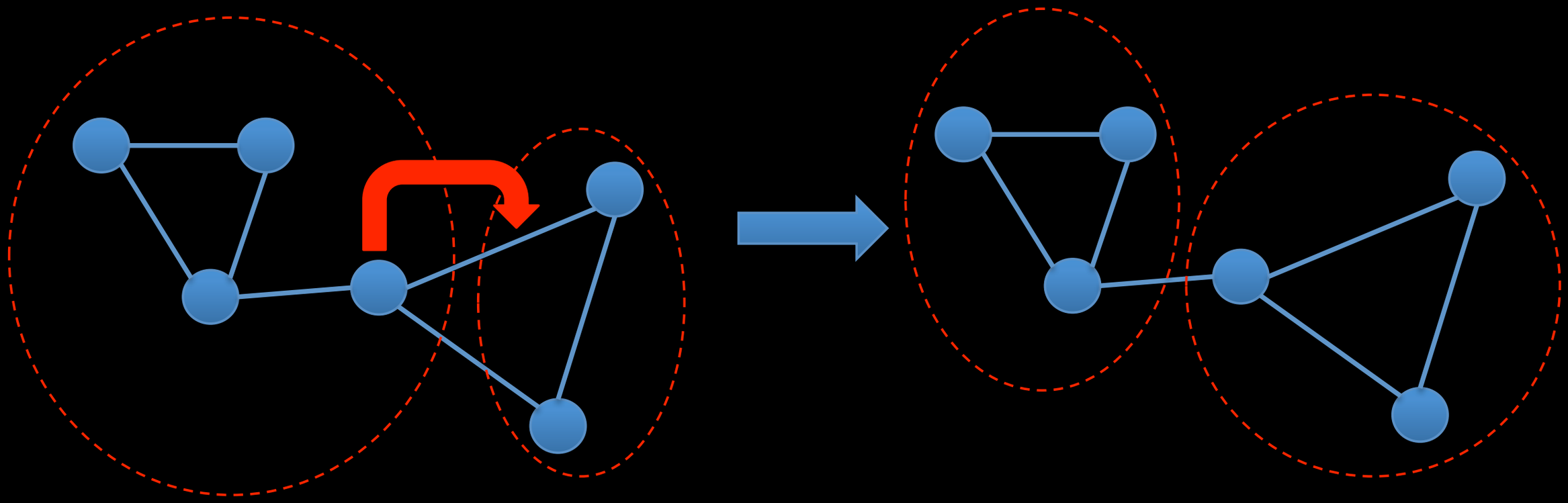
Modularity optimization - splitting



Modularity optimization - merging



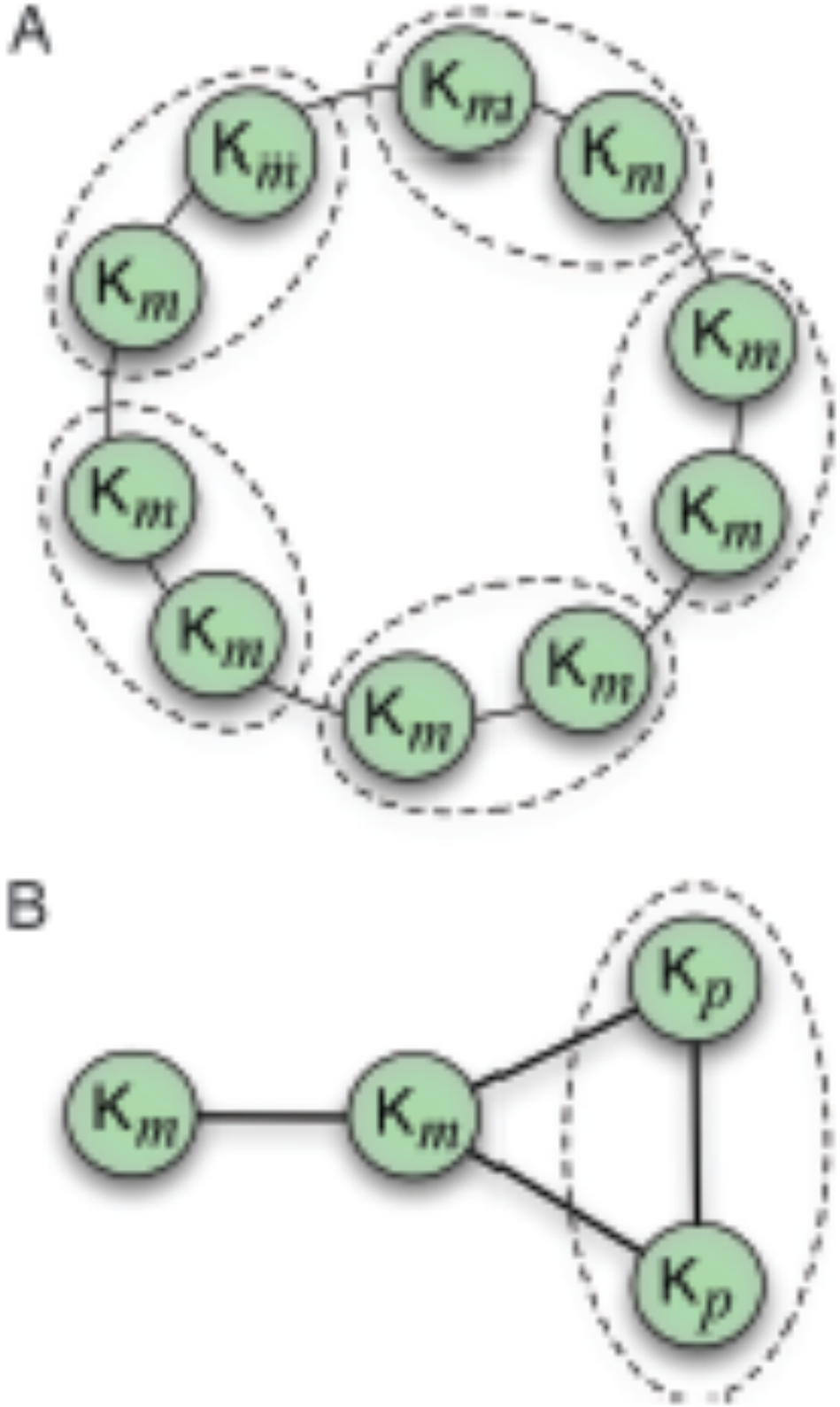
Modularity optimization - shifting



Modularity optimization approaches

Newman's greedy heuristic	M. E. J. Newman, Phys. Rev. E 69, 066133 (2004), URL http://link.aps.org/doi/10.1103/PhysRevE.69.066133
Clauset-Newman-Moore	A. Clauset, M. Newman, and C. Moore, Phys. Rev. E70 (6), 066111 (2004).
Newman's spectral method with refinement	M. Newman, Proceedings of the National Academy of Sciences 103, 8577 (2006).
Louvain method	V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech 10008 (2008)
Le-Martelot's method	E. Le Martelot and C. Hankin, in Proceedings of the 2011 International Conference on Knowledge Discovery and Information Retrieval (KDIR 2011) (SciTePress, Paris, 2011), pp. 216–225.
Extremal optimization	J. Duch and A. Arenas, Phys. Rev. E 72, 027104 (2005),
Simulated Annealing	L. A. R. Guimer`a, M. Sales-Pardo, Phys, Rev. E70(2), 025101 (2004). B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E 81, 046106 (2010)
COMBO	Sobolevsky, S., Campari, R., Belyi, A. and Ratti, C., 2014. General optimization technique for high-quality community detection in complex networks. *Physical Review E*, *90*(1), p.012811.

Modularity resolution limit



Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, *104*(1), 36-41.

Alternatives to modularity

Infomap	M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences 104, 7327 (2007) M. Rosvall and C. Bergstrom, Proc. Natl. Acad. Sci. USA 105, 1118 (2008).
Surprise	R. Aldecoa and I. Mar`ın, PLoS ONE 6, e24195 (2011),
Block-model likelihood	B. Karrer and M. E. J. Newman, Phys. Rev. E 83 B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E 84, 036103 (2011),