CENTER FOR URBAN SCIENCE+PROGRESS

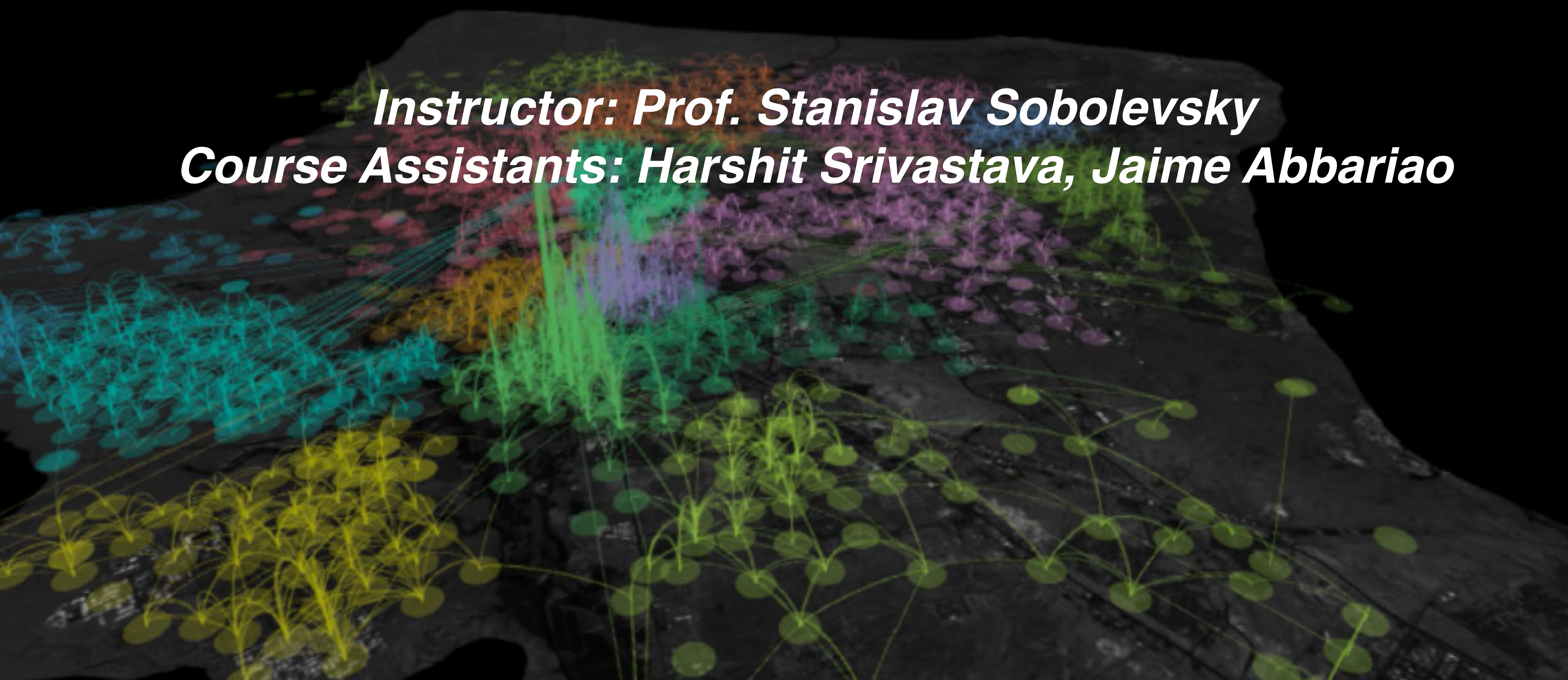
APPLIED DATA SCIENCE

6004.002, Fall 2018

Network Analysis-2. Community detection

Instructor: Prof. Stanislav Sobolevsky

Course Assistants: Harshit Srivastava, Jaime Abbariao

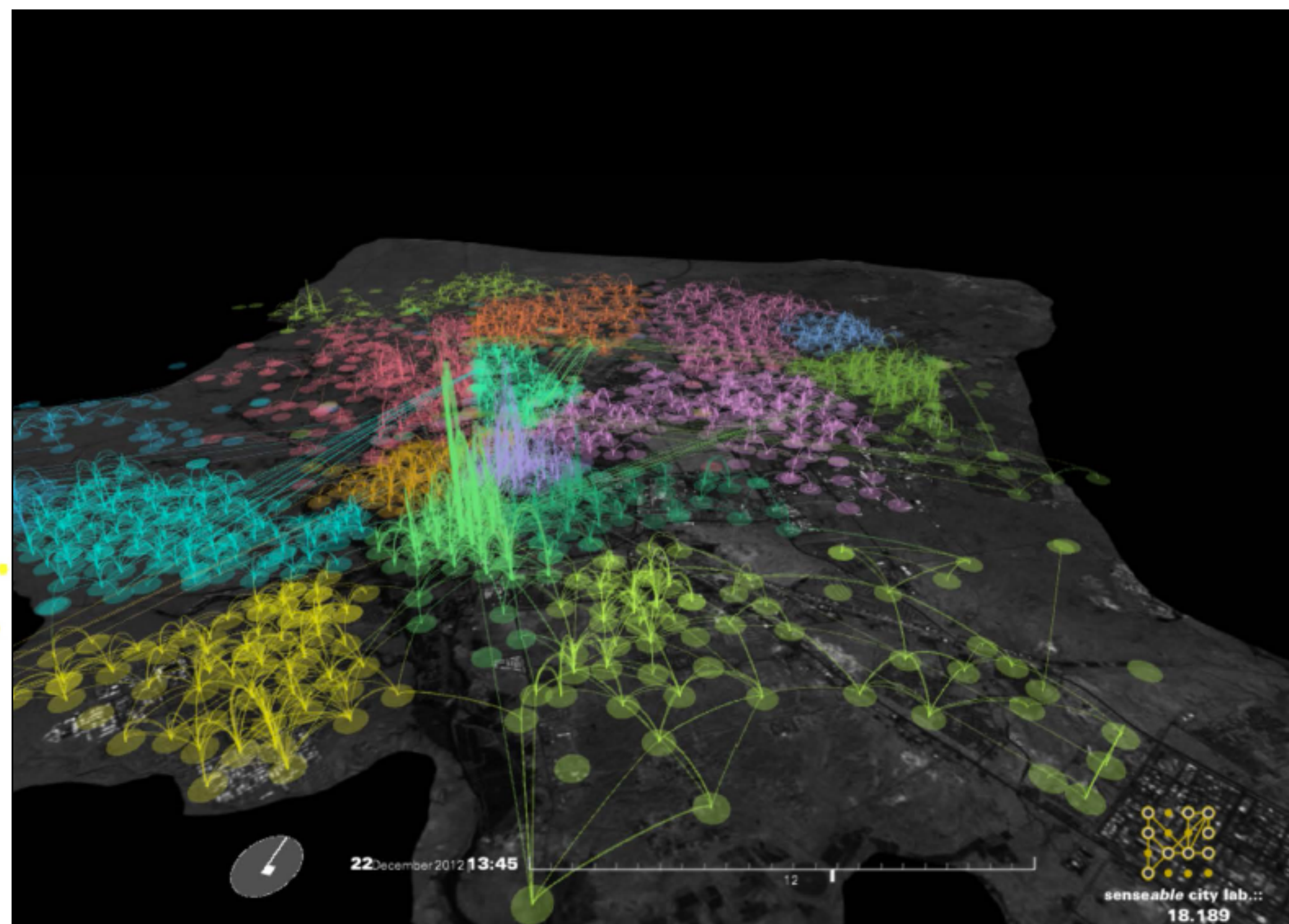


Community detection

Discovering underlying structure connection heterogeneity: some groups of nodes are connected more strongly

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

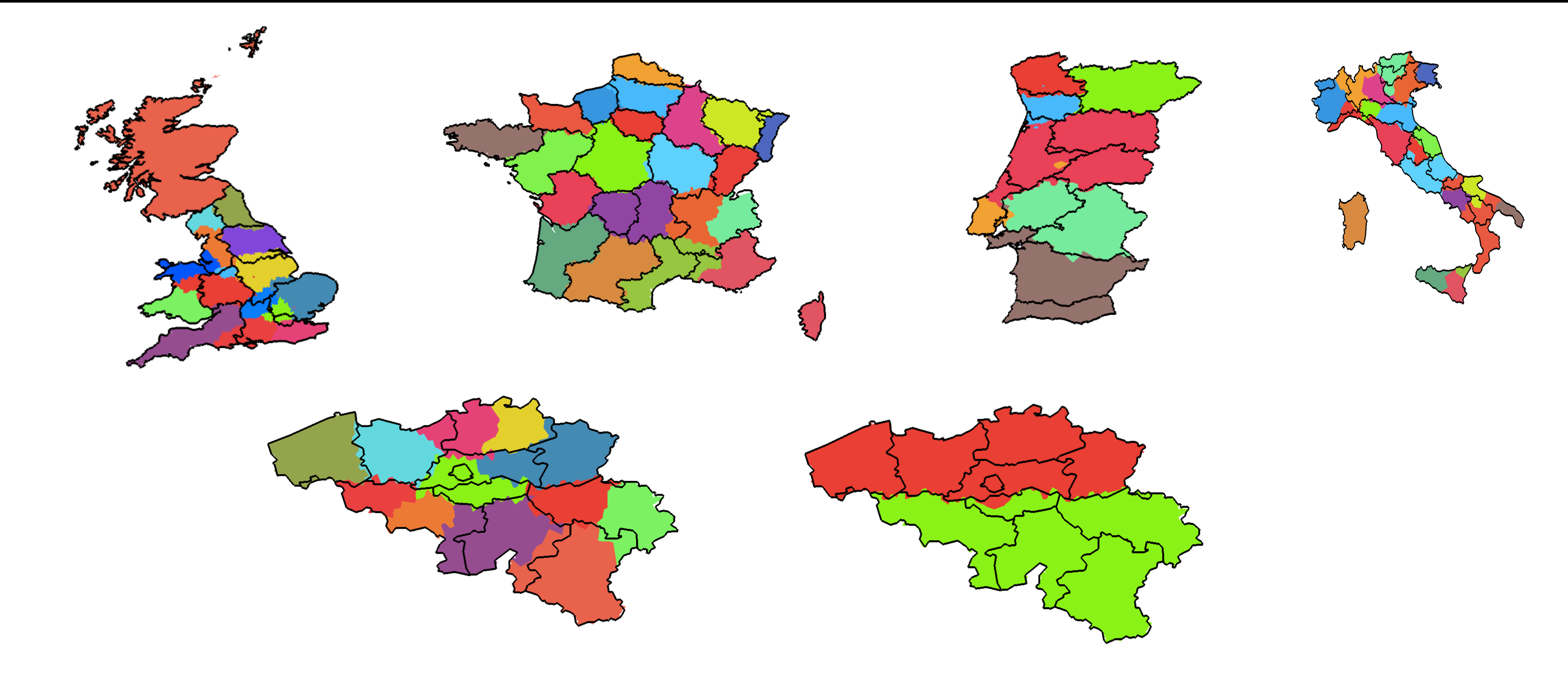
http://sociograph.blogspot.com/2013/05/revealing-community-structure-with.html



Regional delineation

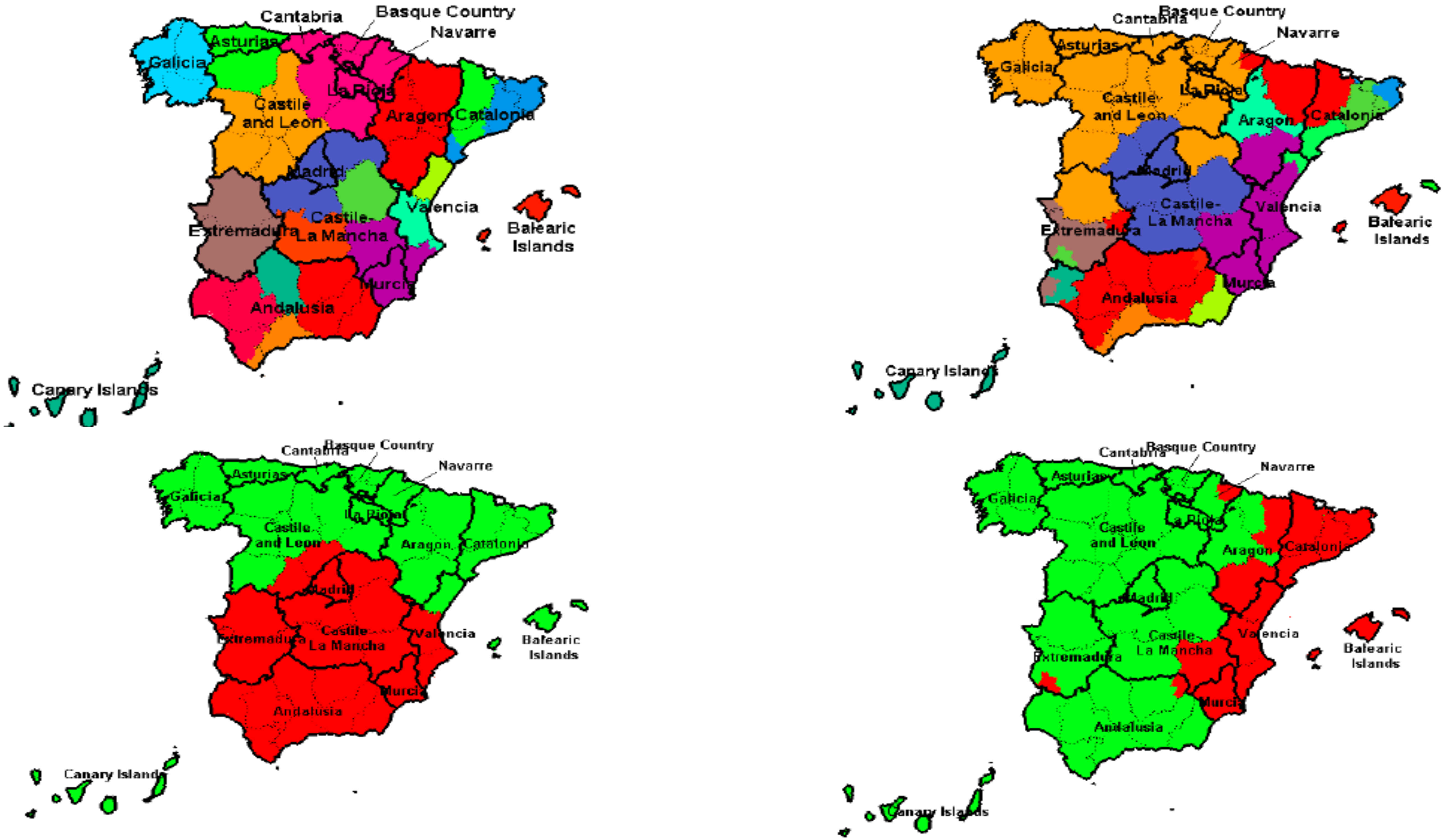
Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., ... & Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PloS one*, *5*(12), e14248.

Regional delineation



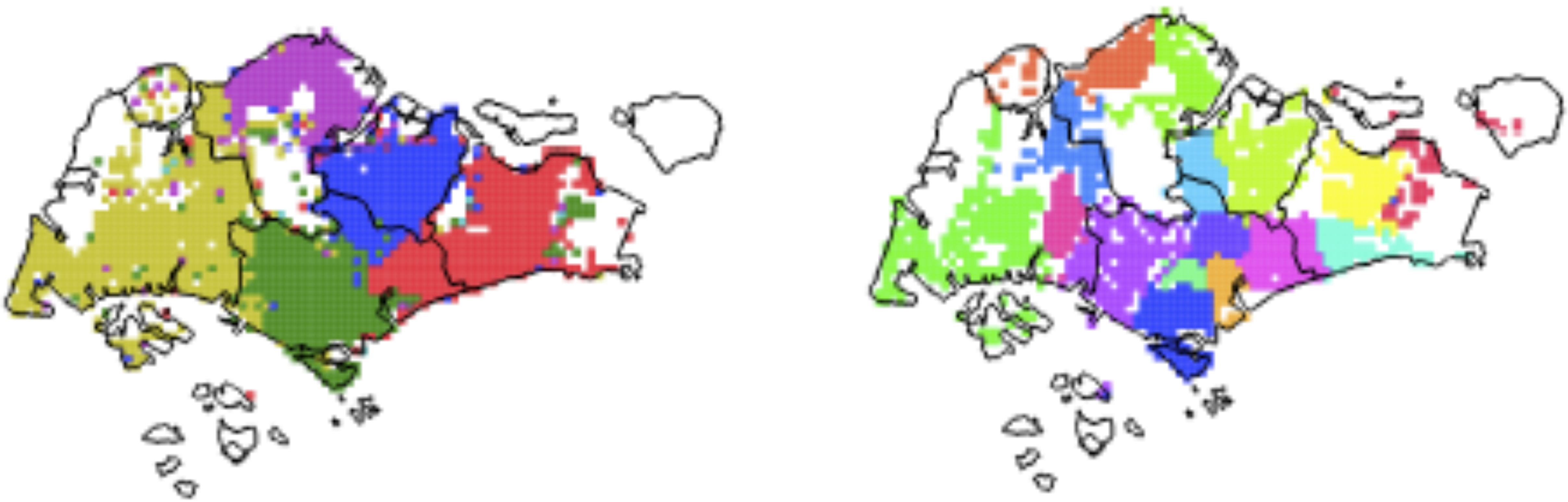
Sobolevsky S., Szell M., Campari R., Couronne T., Smoreda Z., Ratti C. (2013) Delineating geographical regions with networks of human interactions in an extensive set of countries. PLoS ONE 8 (12), e81707

Credit cards



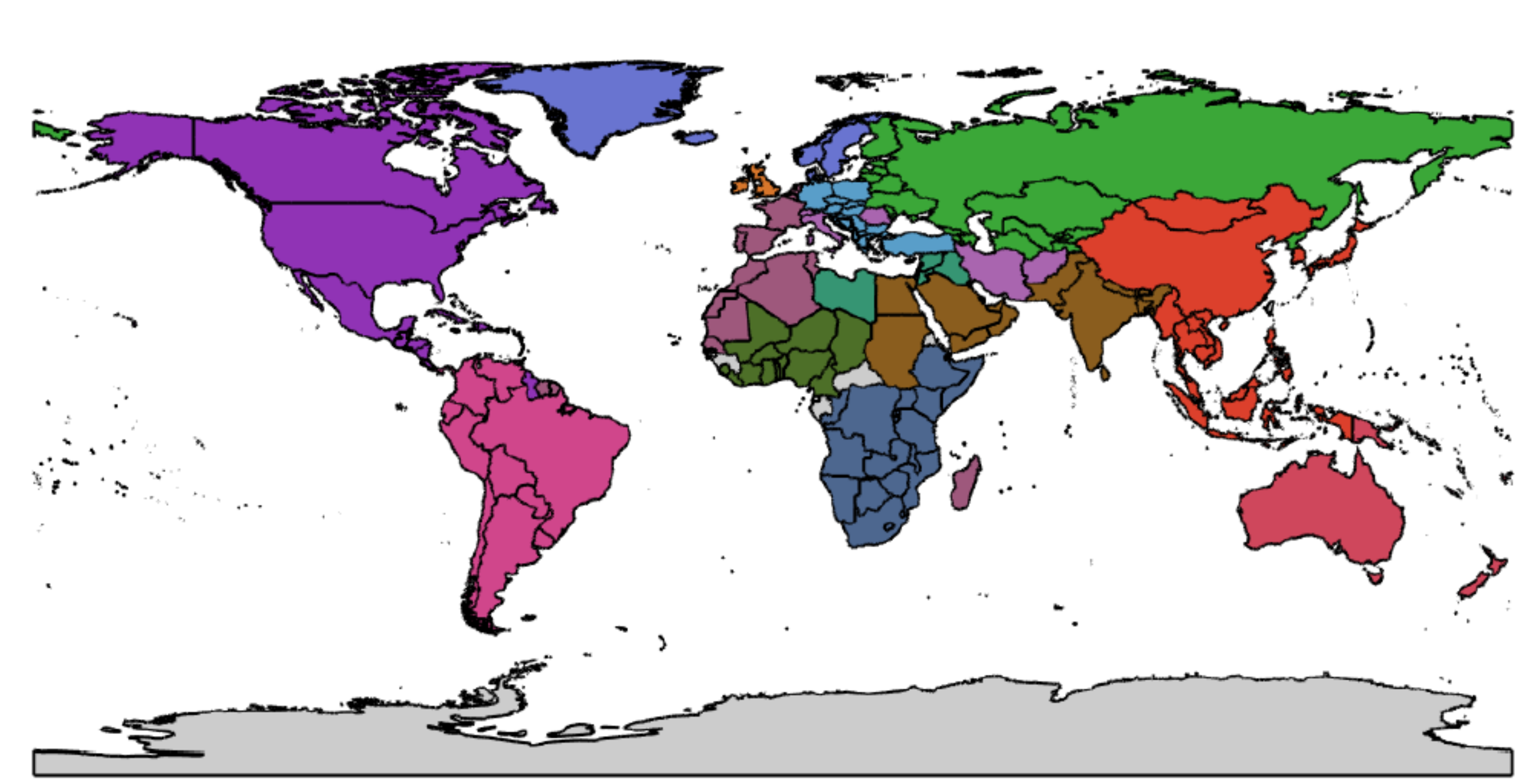
Sobolevsky, Stanislav, et al. "Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain." *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE, 2014.

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.

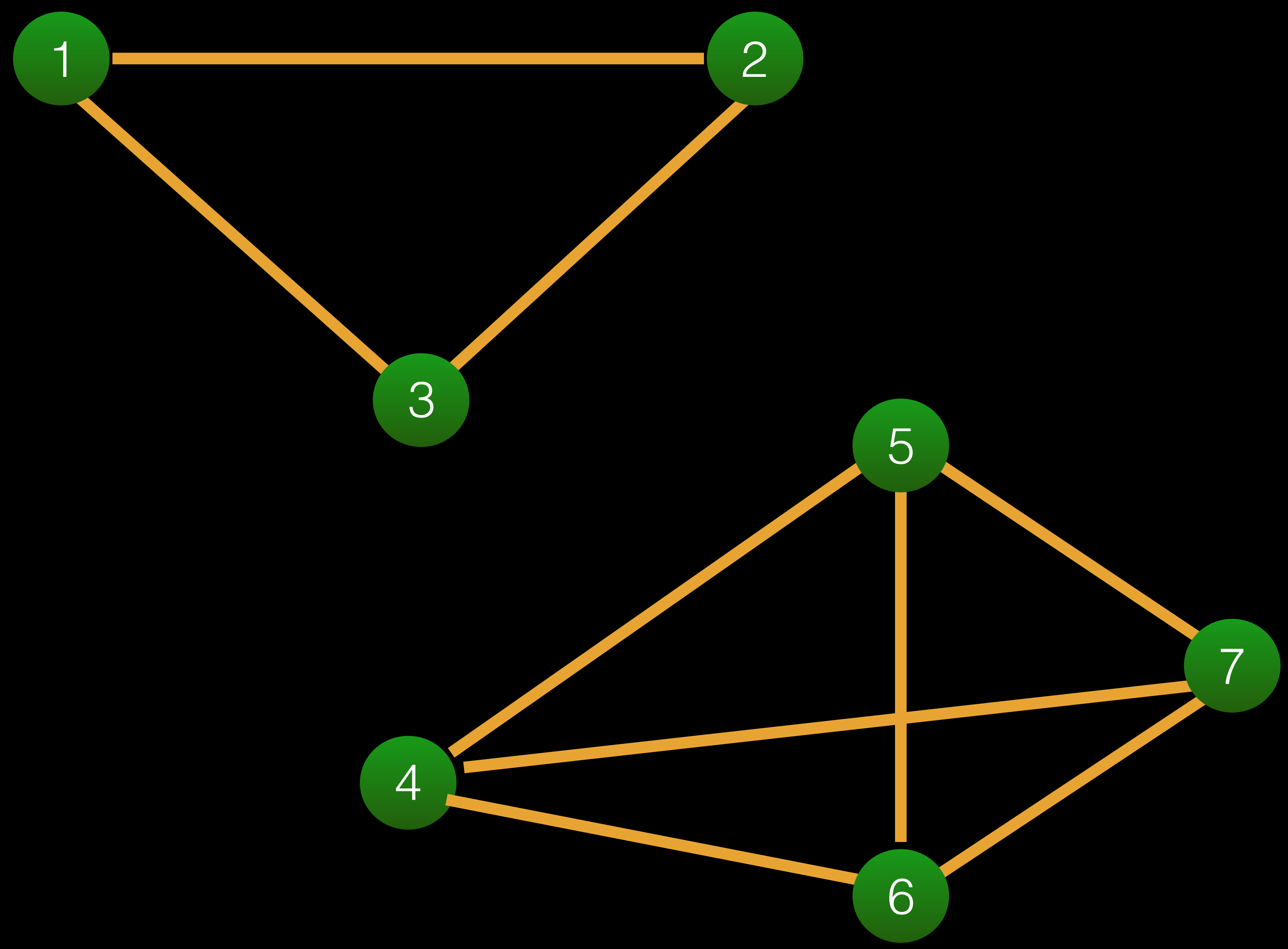
Global: social media



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

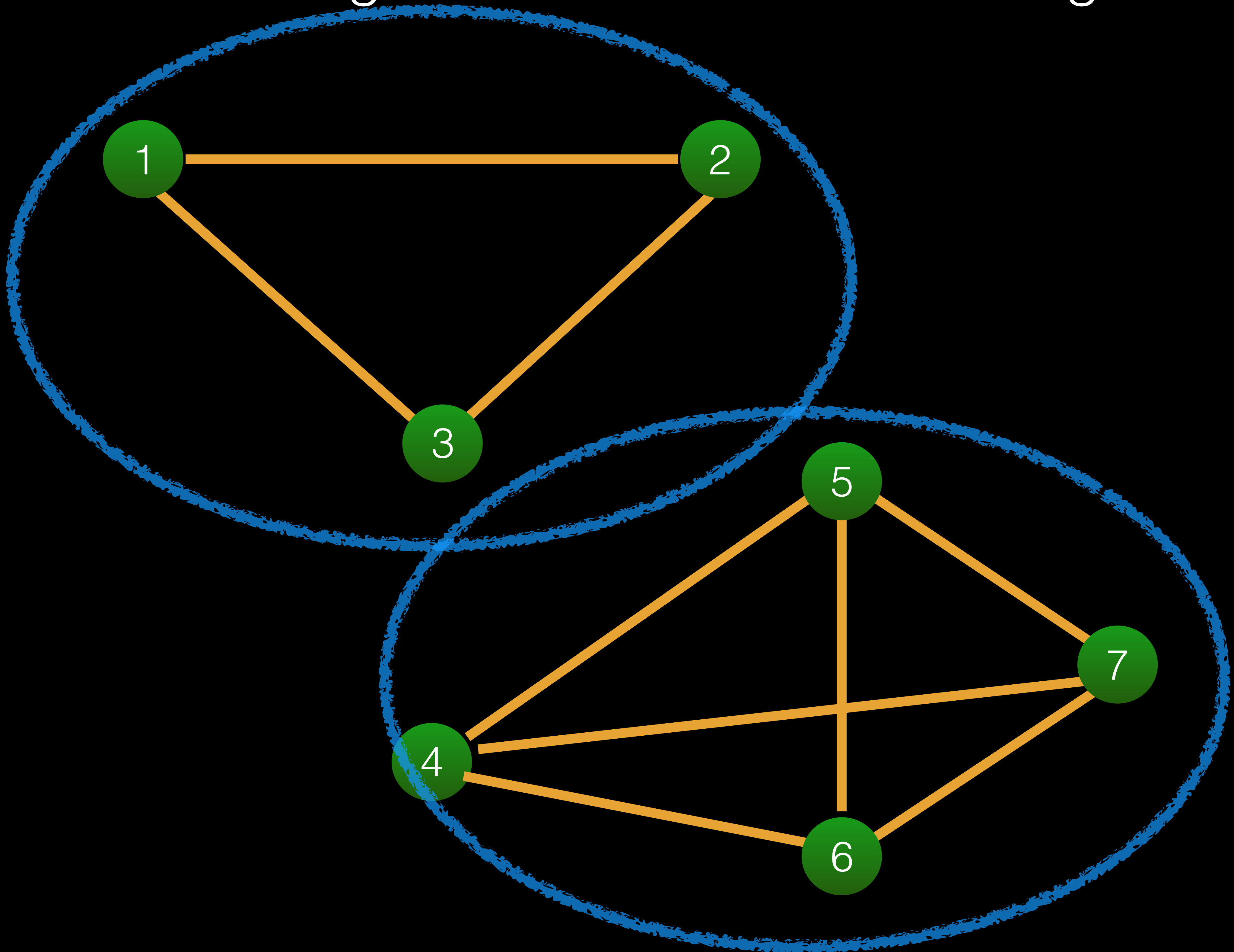
Community

Group of nodes having internal connections stronger vs external



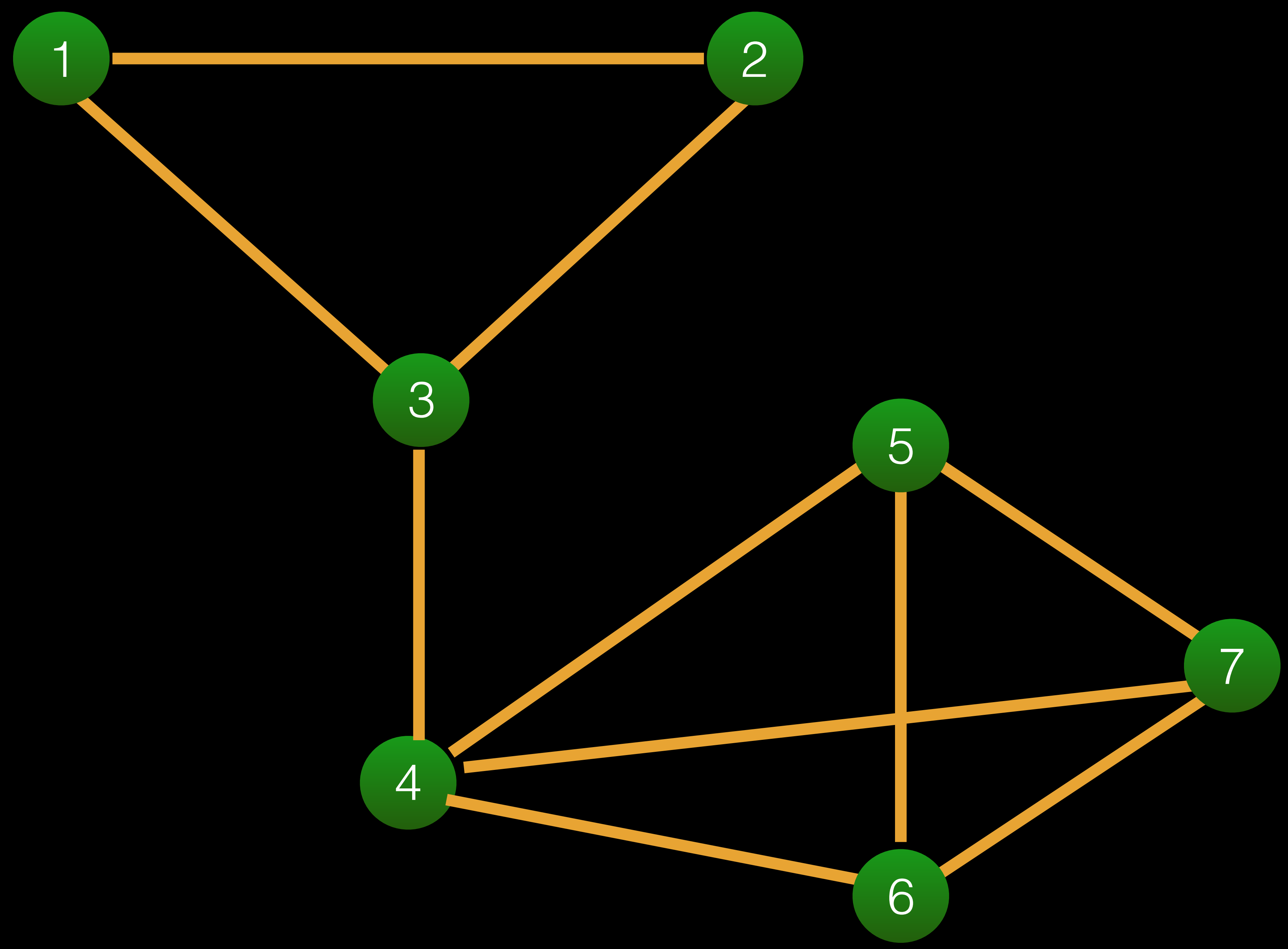
Community

Group of nodes having internal connections stronger vs external



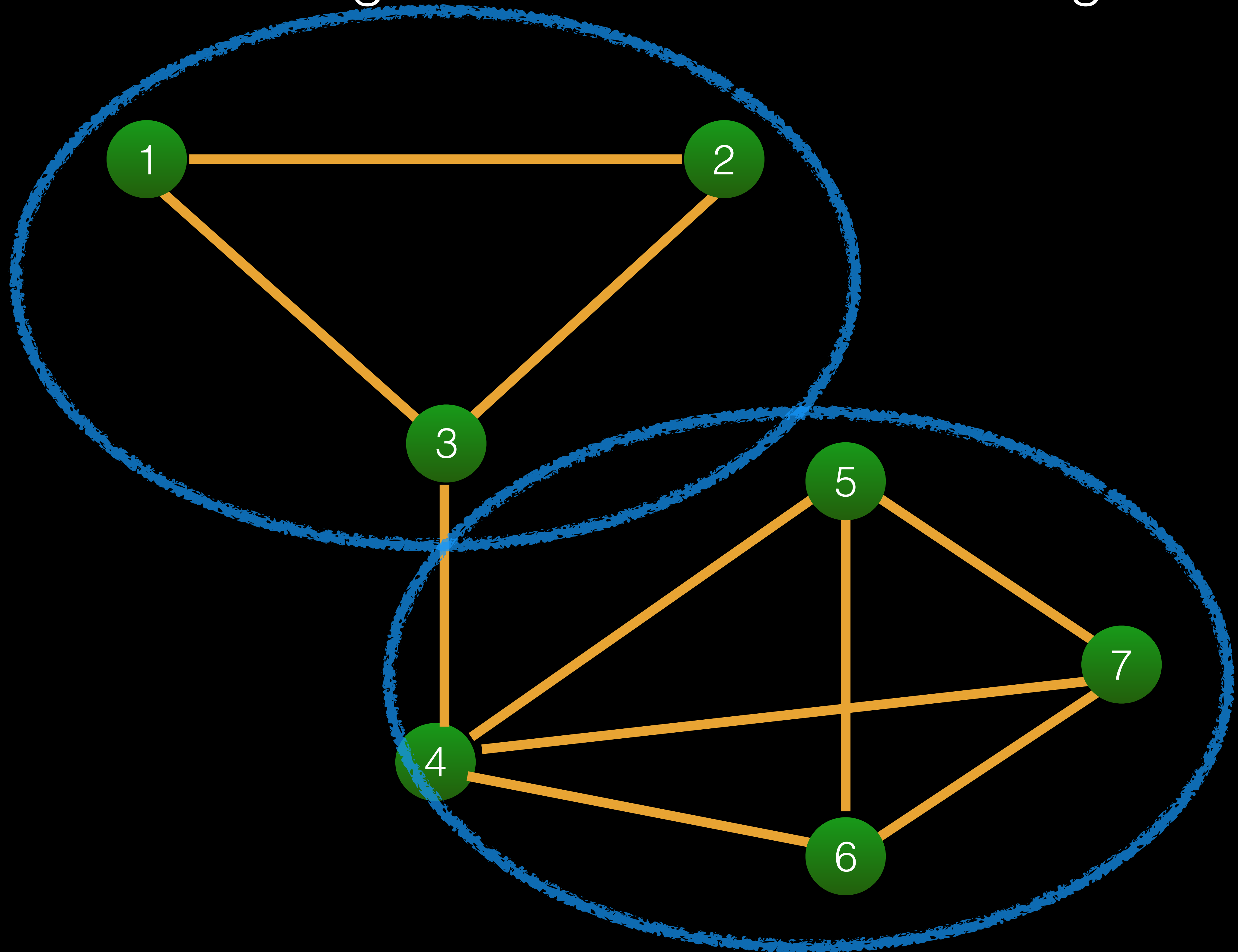
Community

Group of nodes having internal connections stronger vs external



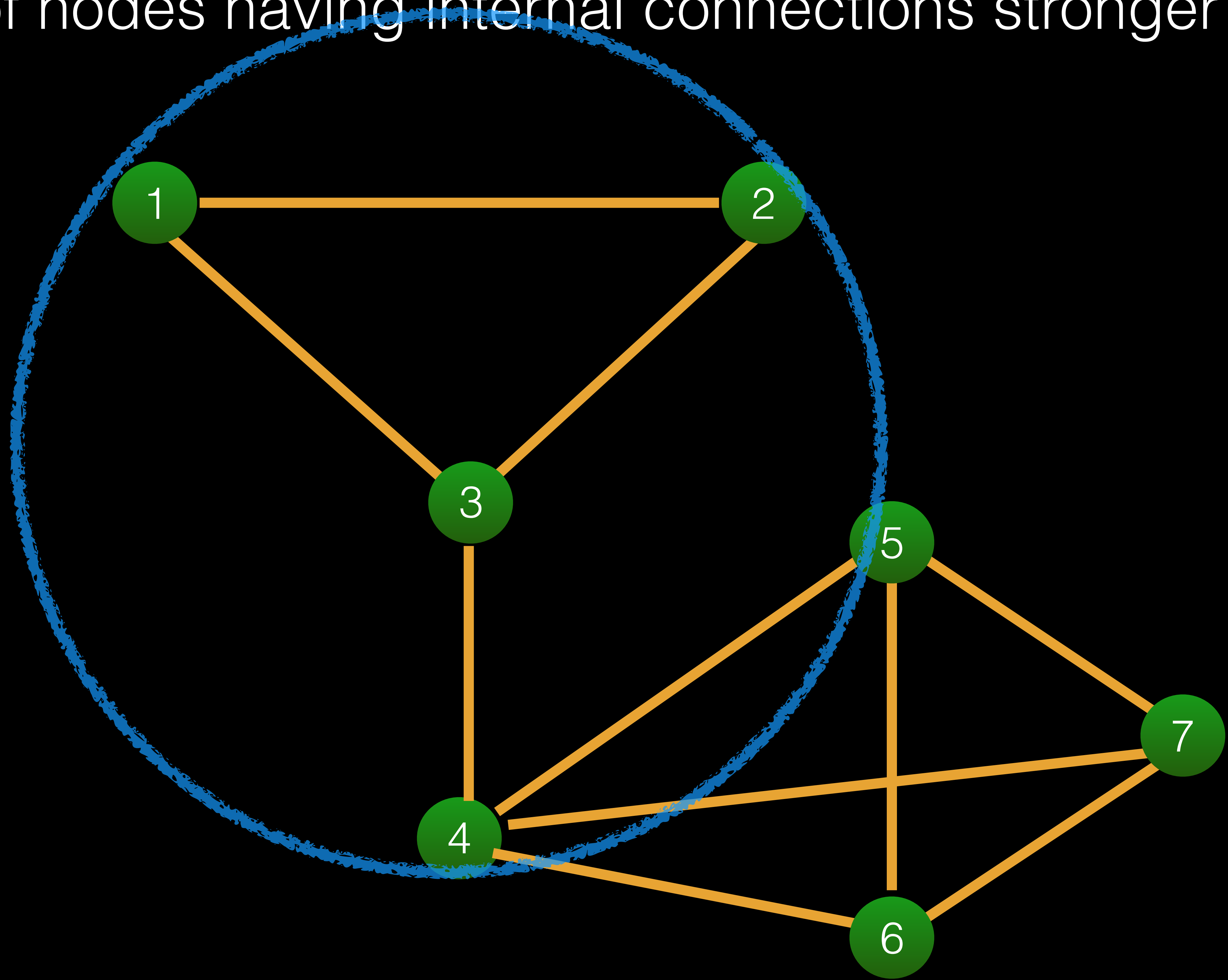
Community

Group of nodes having internal connections stronger vs external



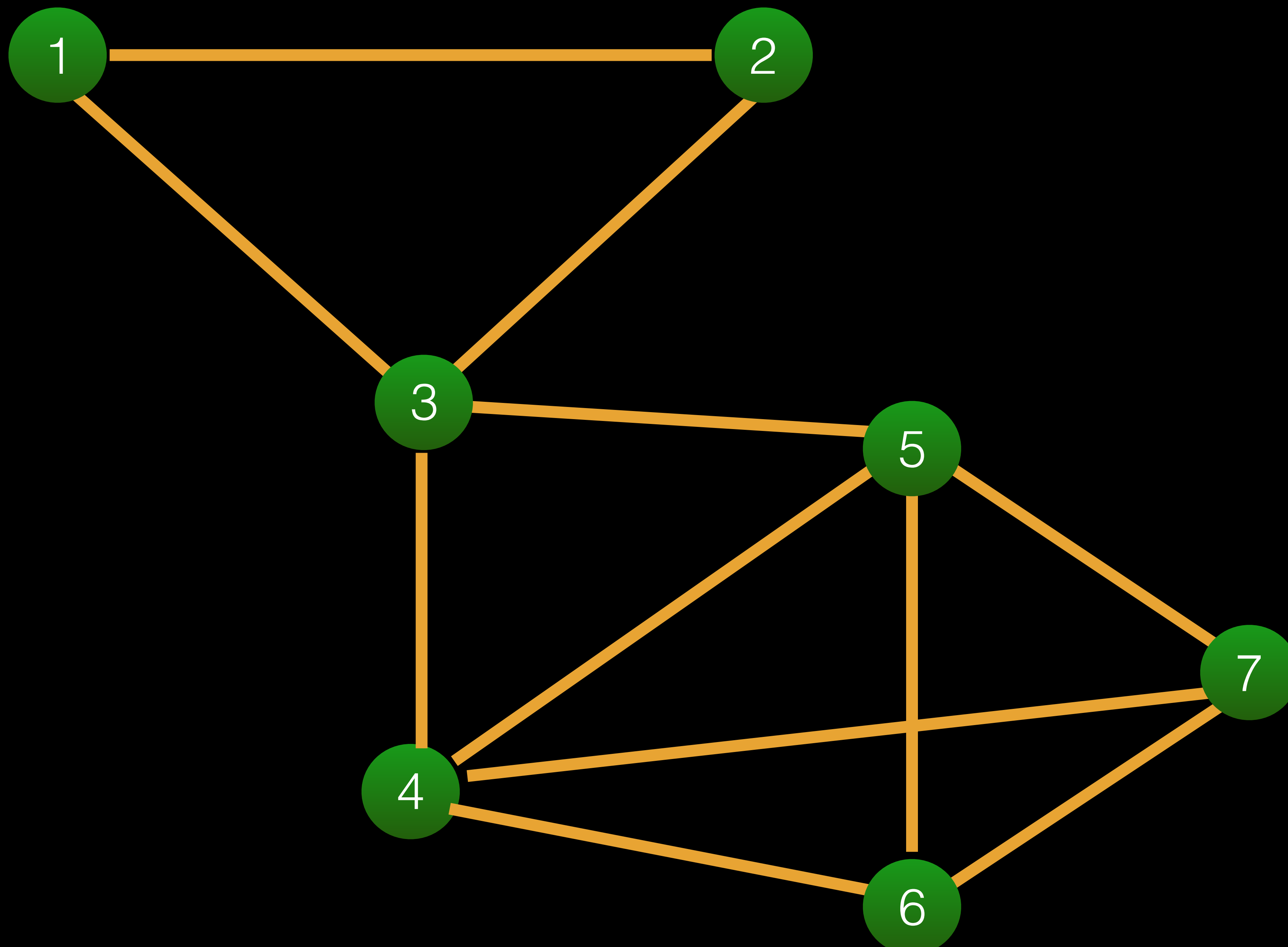
Community

Group of nodes having internal connections stronger vs external



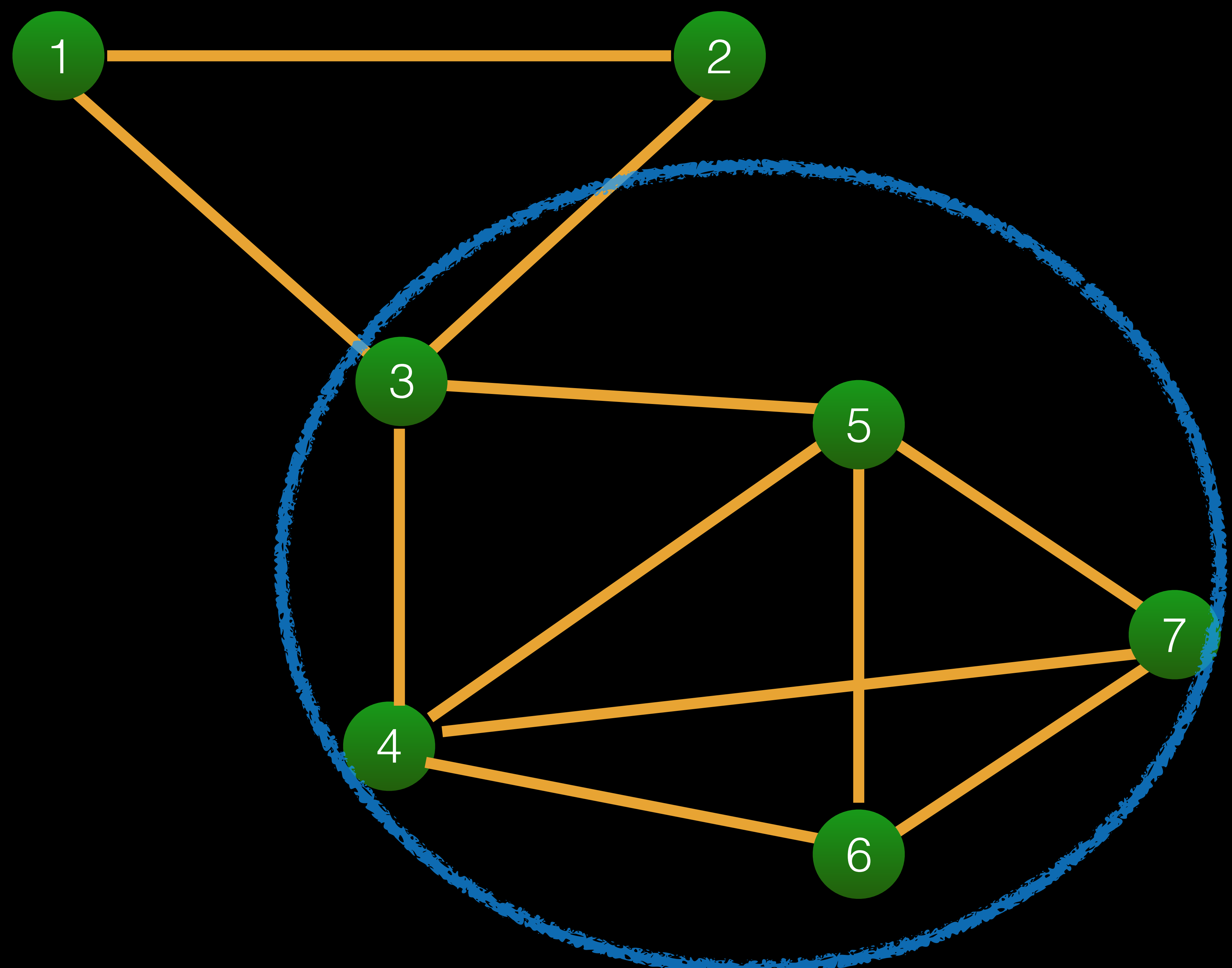
Community

Group of nodes having internal connections stronger vs external



Community

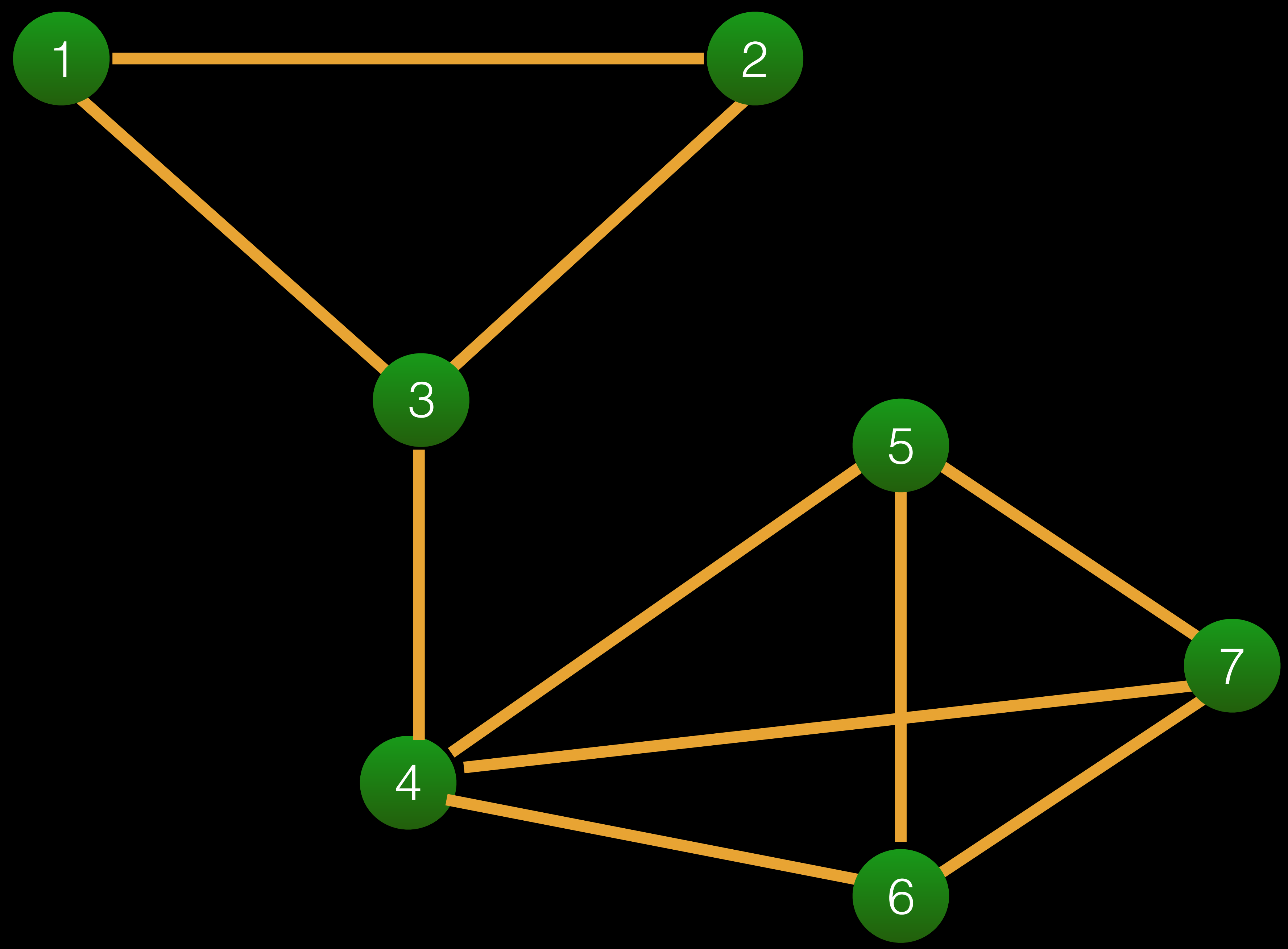
Group of nodes having internal connections stronger vs external



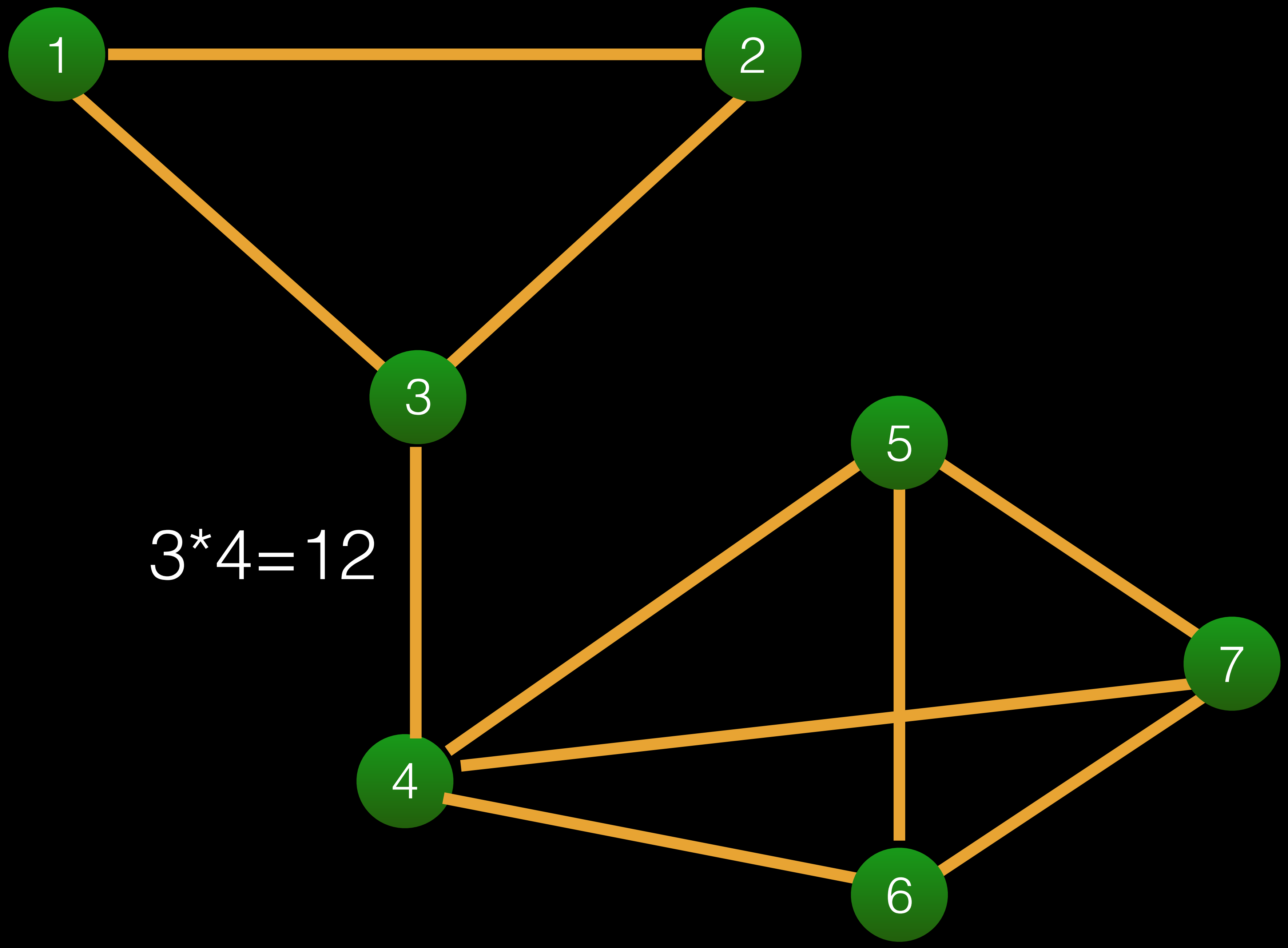
Community detection

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

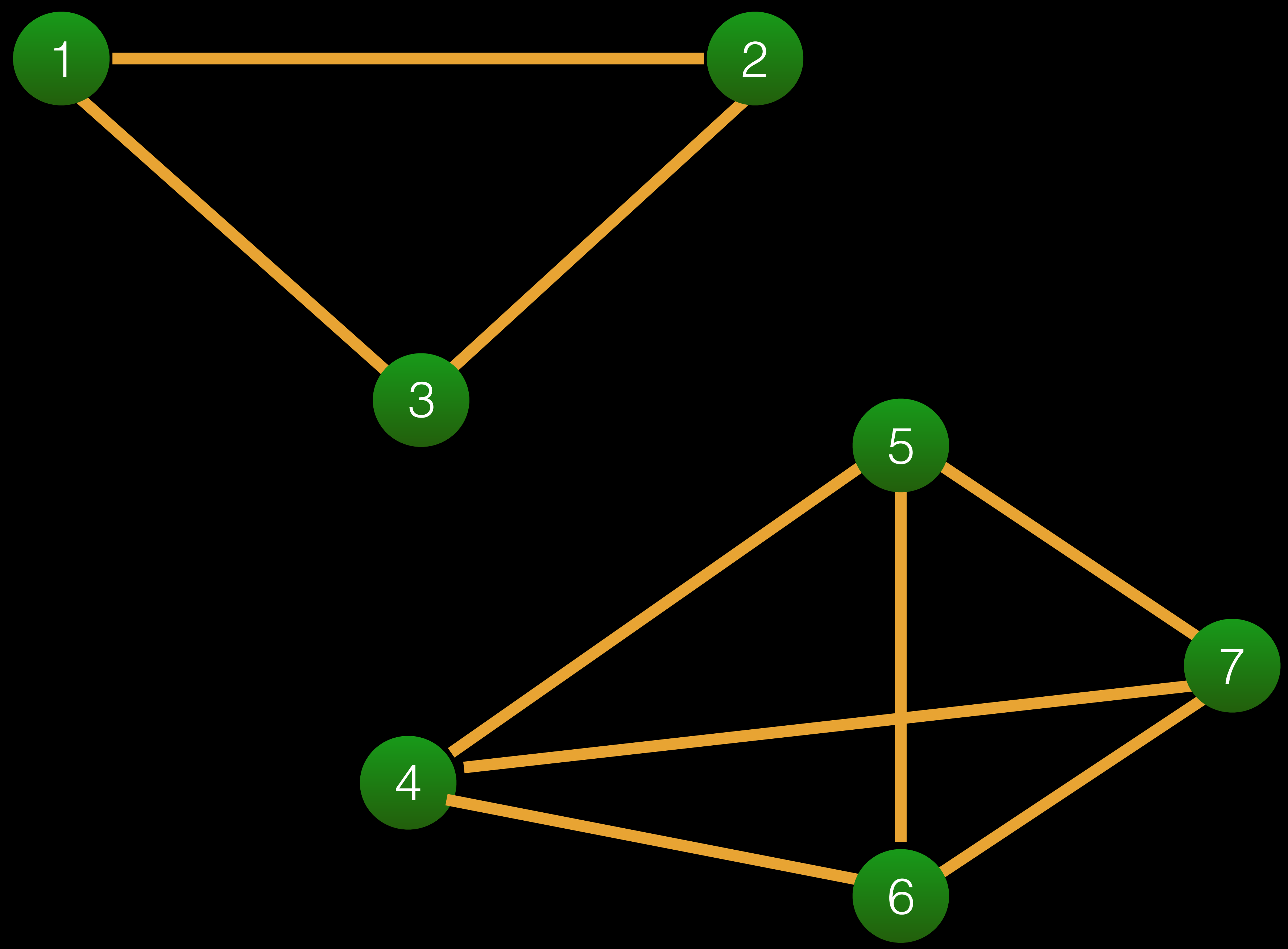
Girvan-Newman algorithm, 2002



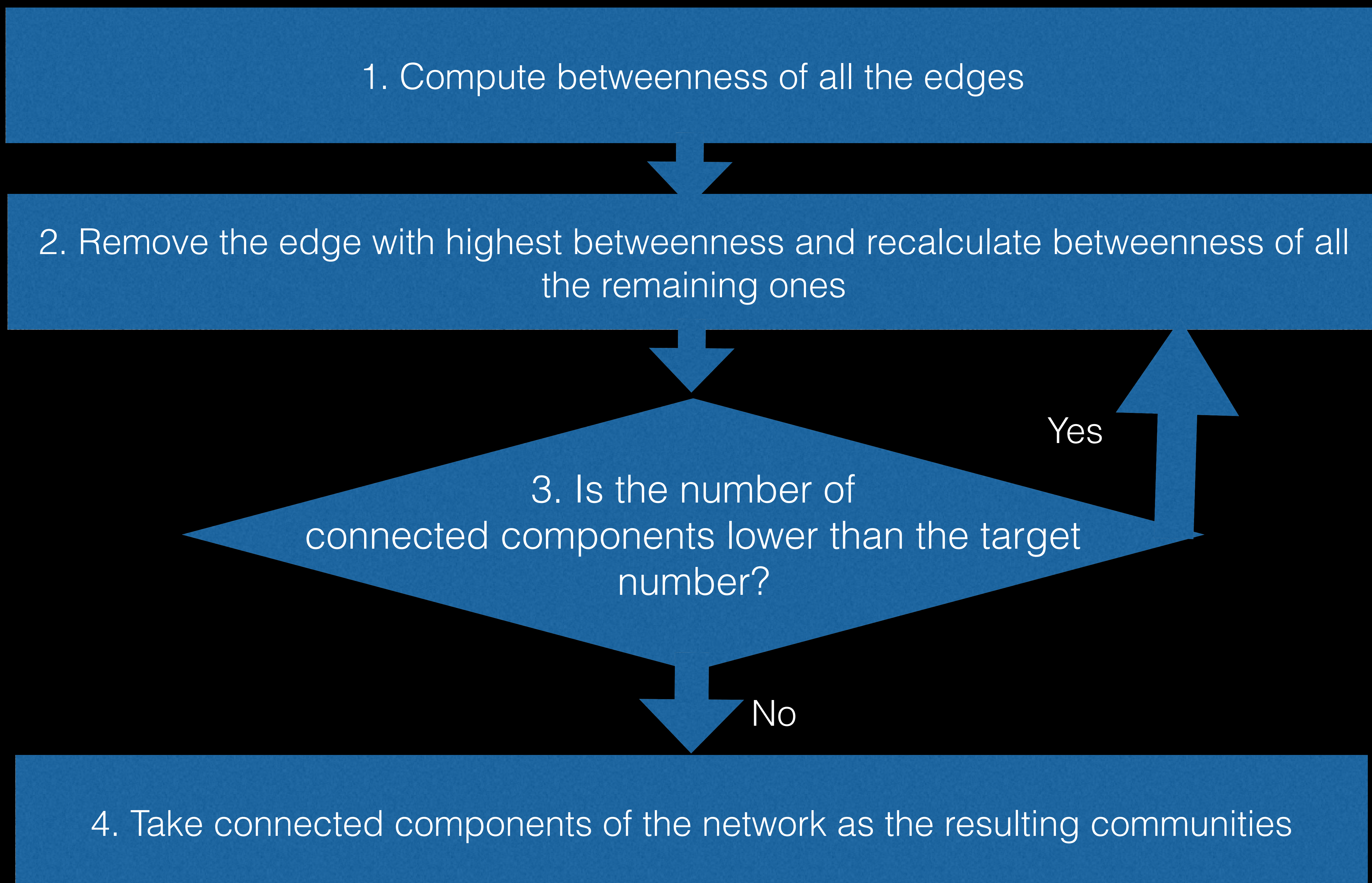
Girvan-Newman algorithm, 2002



Girvan-Newman algorithm, 2002

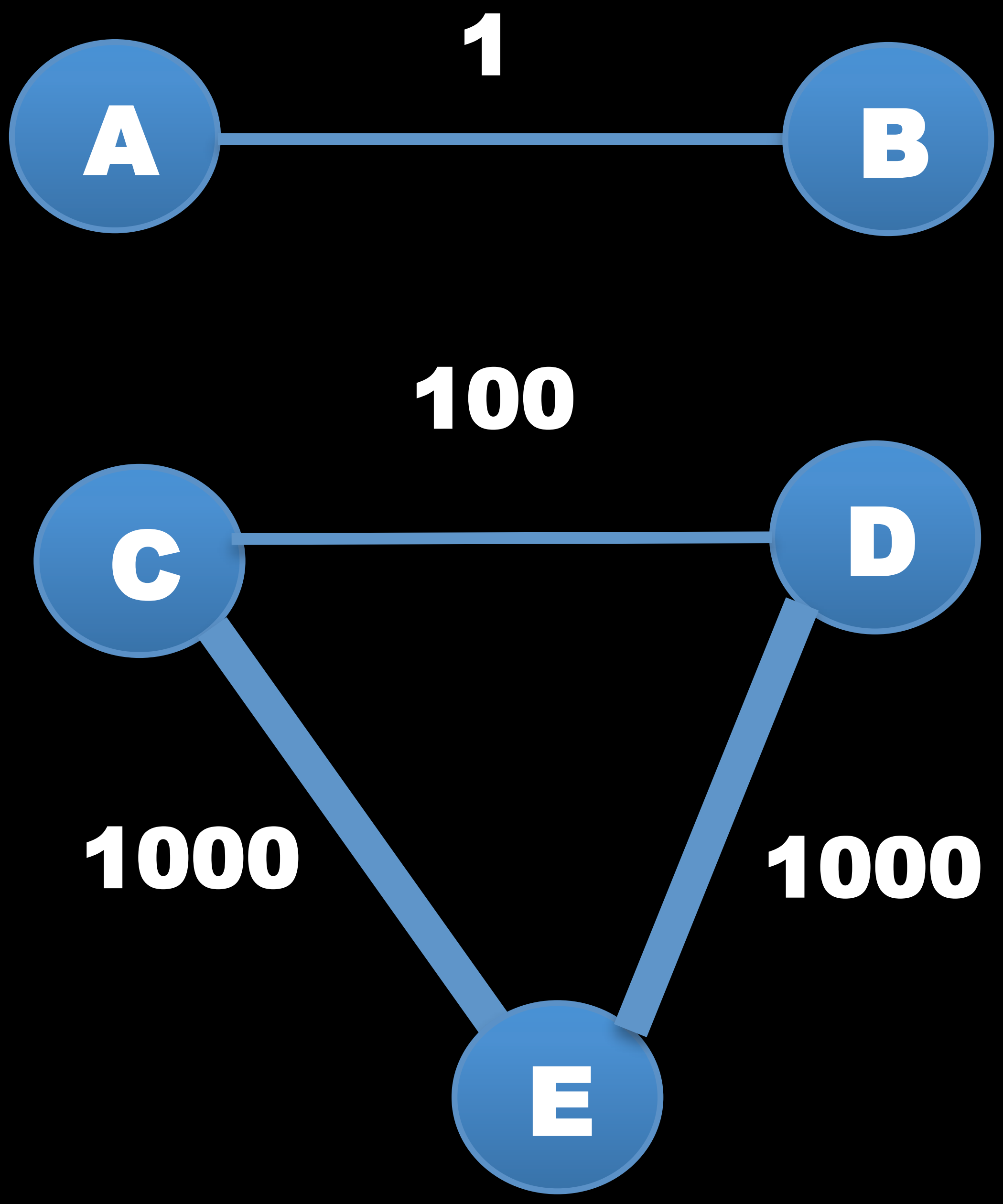


Girvan-Newman algorithm, 2003

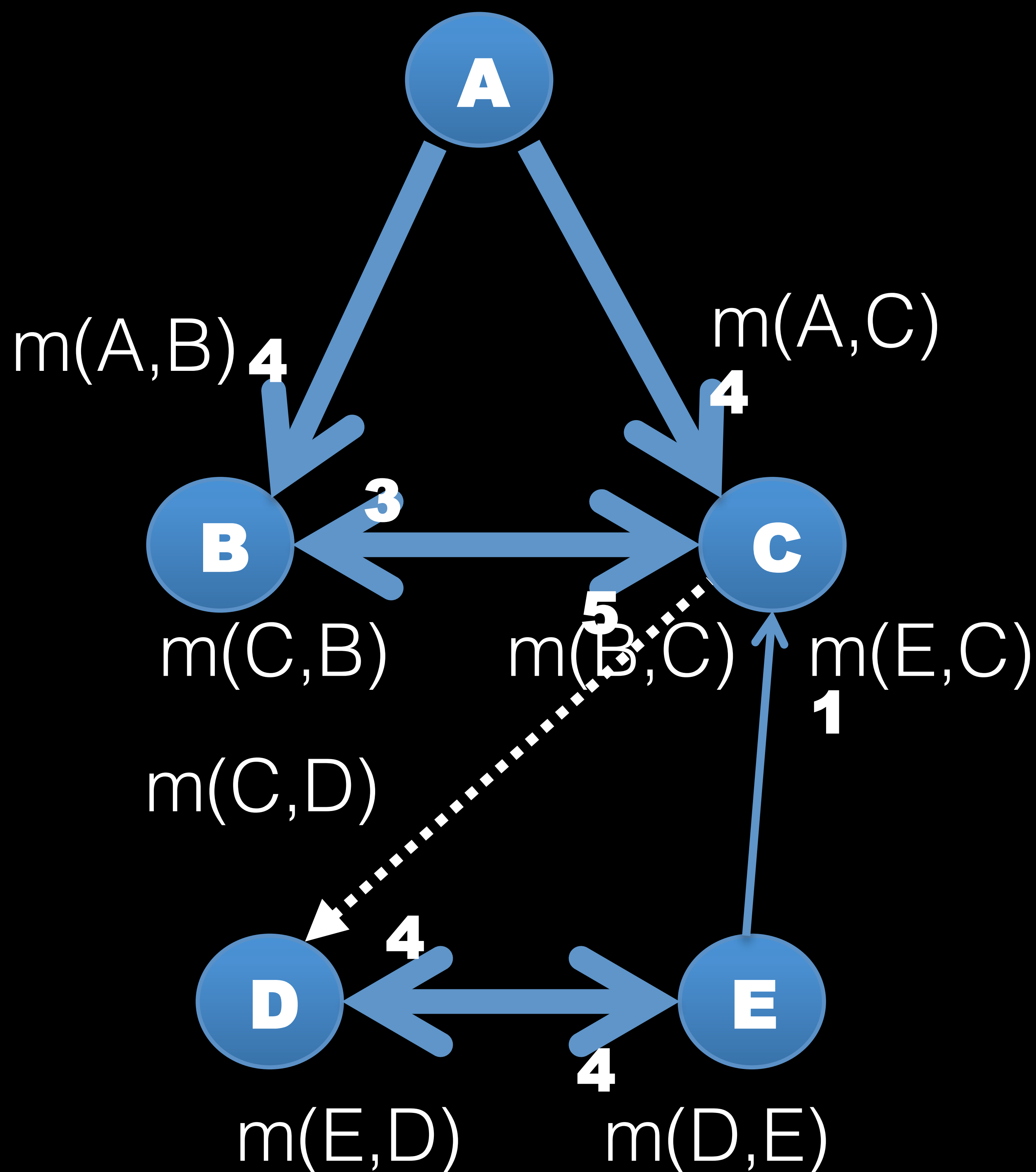


Modularity

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



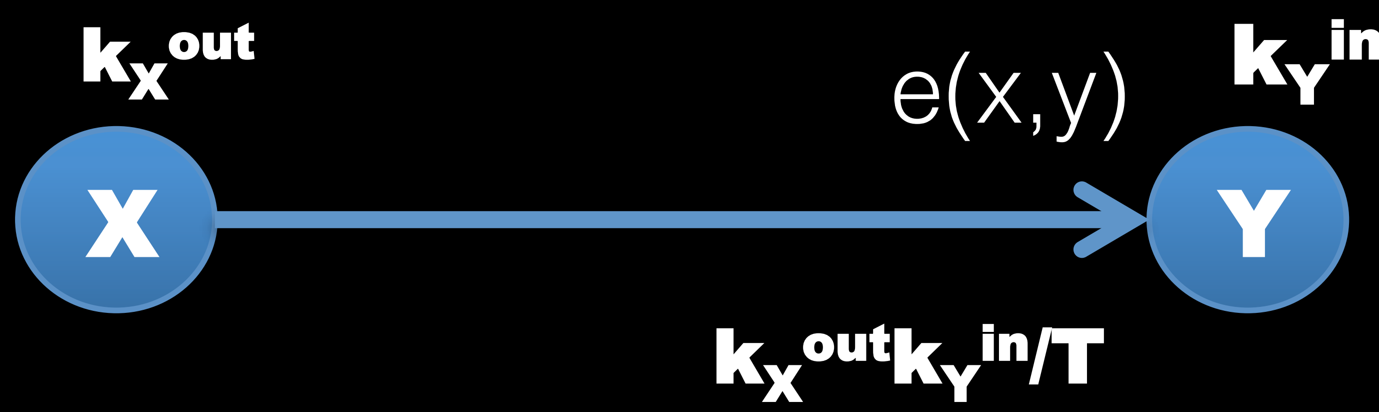
Modularity



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

Modularity

Relative strength is estimated by comparing the actual edge weight against the average expectation of a network having same node strength but randomly/homogenously wired edges, normalized by the total network weight



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

Modularity

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

For a given partition P estimated the cumulative relative strength of all the internal edges

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

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

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

Modularity

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

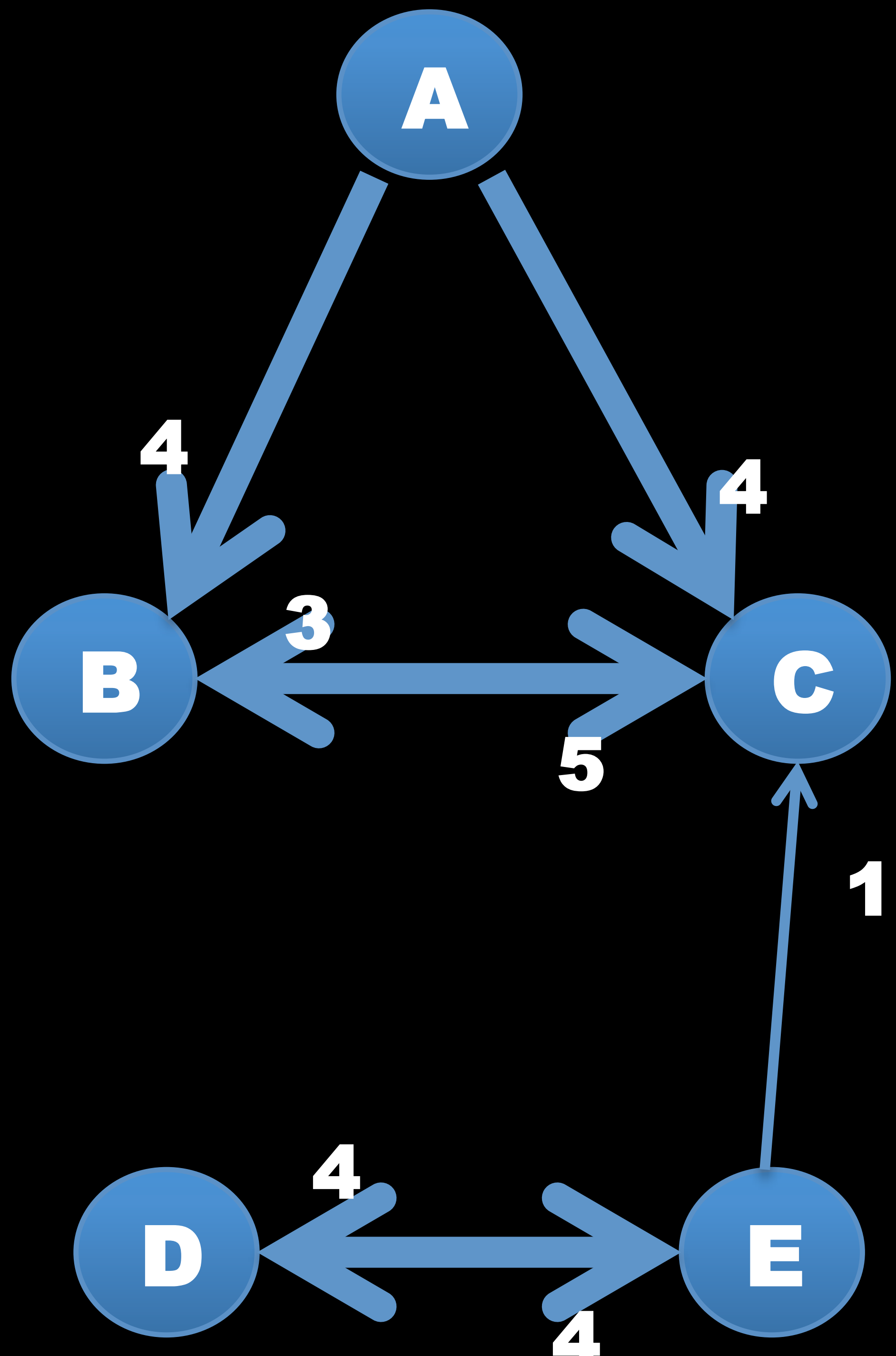
$$Q(P_0) = \sum_{x,y}\left[\frac{e(x,y)}{T} - \frac{k_x^{out}k_y^{in}}{T^2}\right] = \sum_{x,y}\frac{e(x,y)}{T} - \sum_{x,y}\frac{k_x^{out}k_y^{in}}{T^2} = 1 - 1 = 0.$$

Modularity is a normalized metric between -1 and 1. It is always zero for a trivial partition where all nodes form a single community

So we are looking for partitions having positive modularity score, i.e. better than trivial

With modularity as an objective function community detection becomes an optimization problem

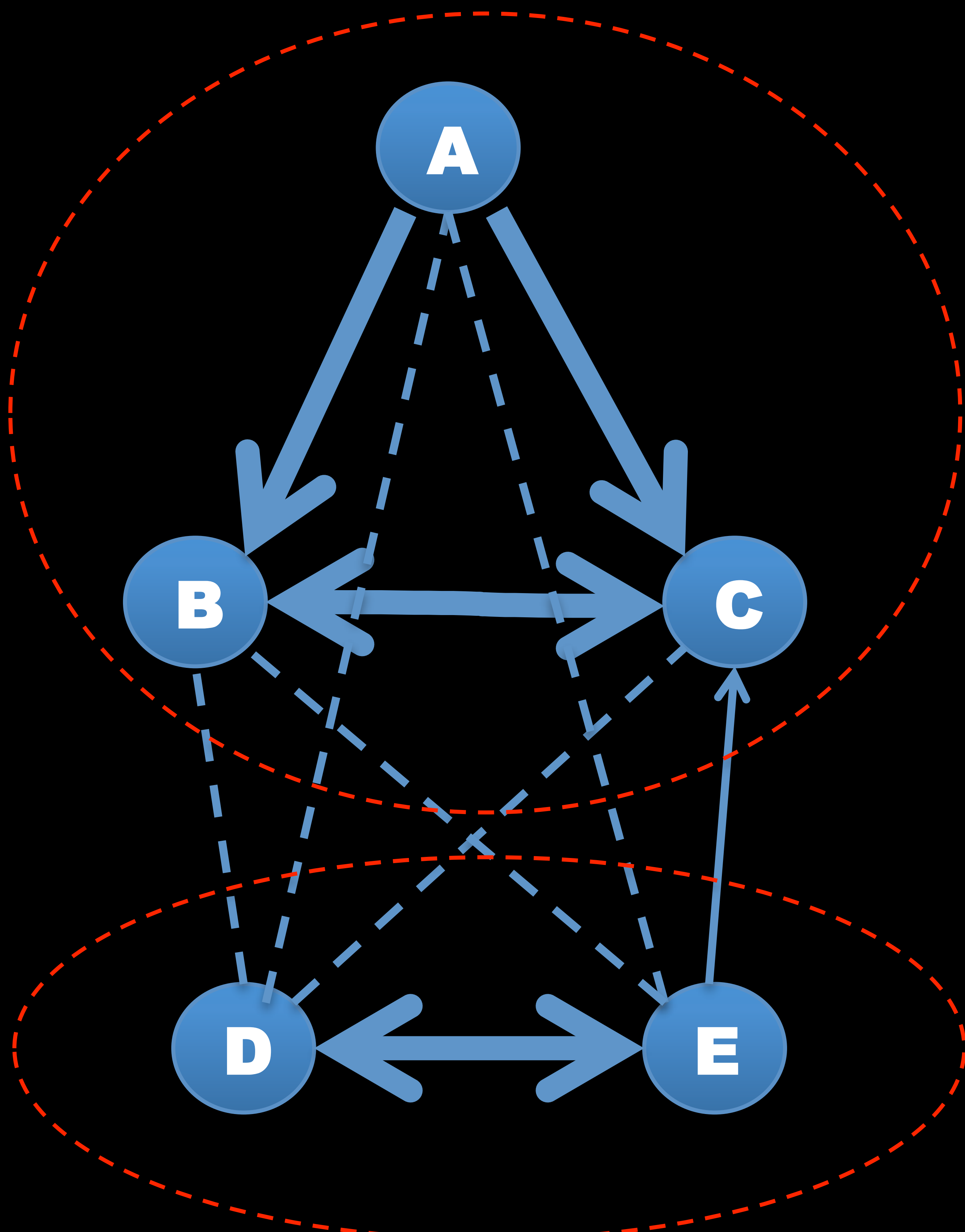
Modularity - illustration



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

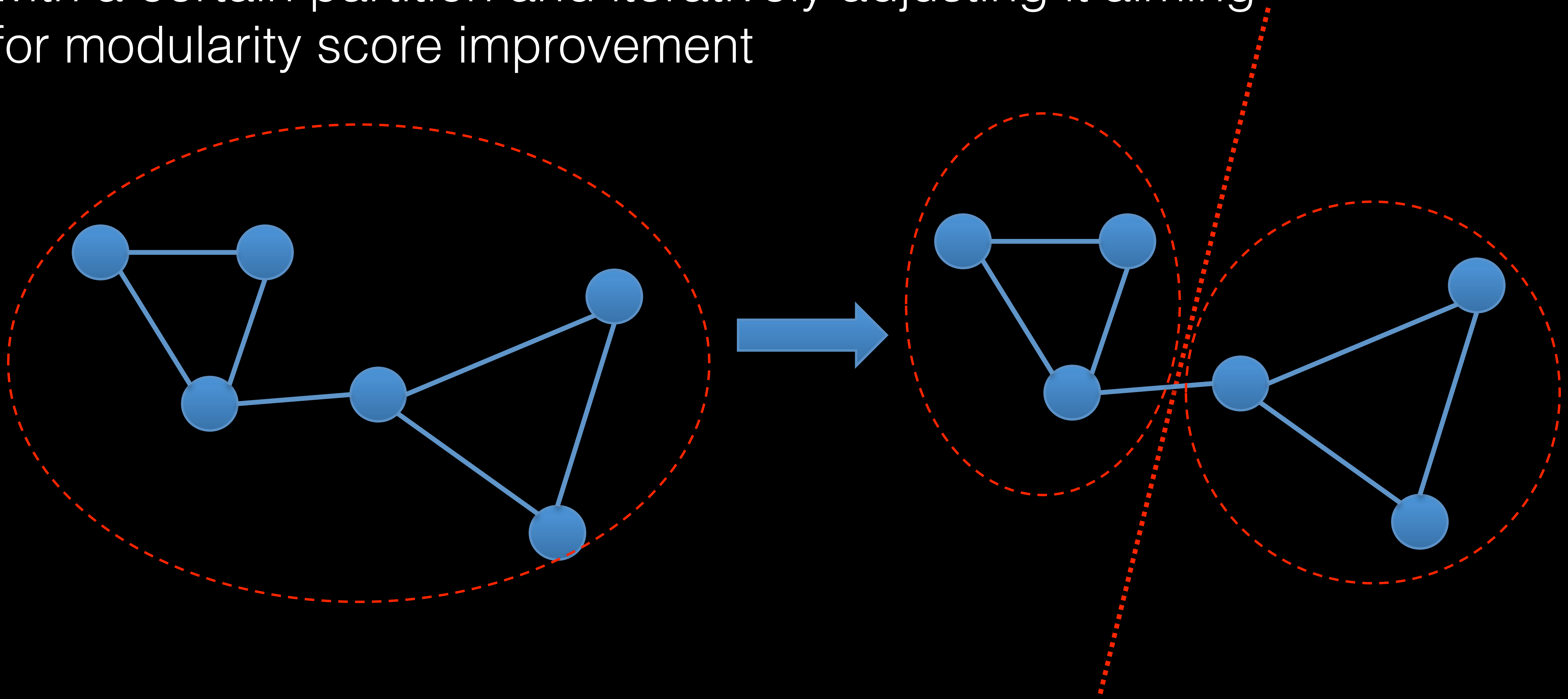
Modularity



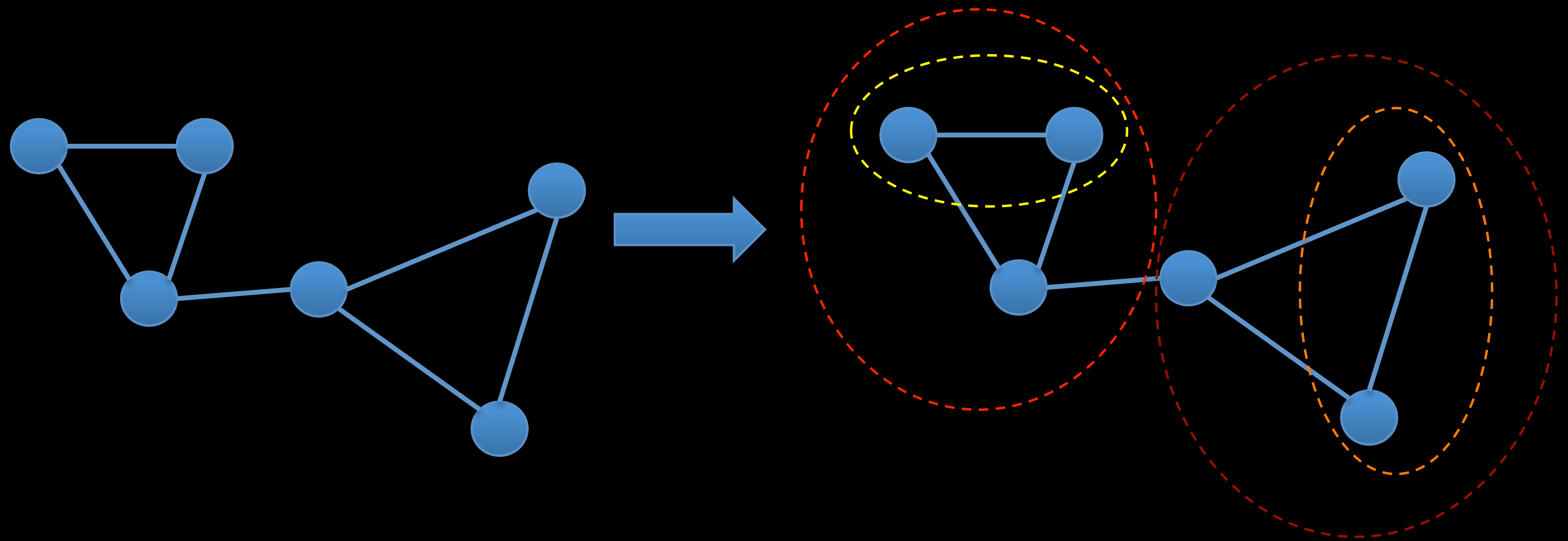
With all edges between A,B,C and between D,E having positive score and all edges between the two groups having negative score, the best partition will be to cut between A,B,C and D,E

Modularity optimization - splitting

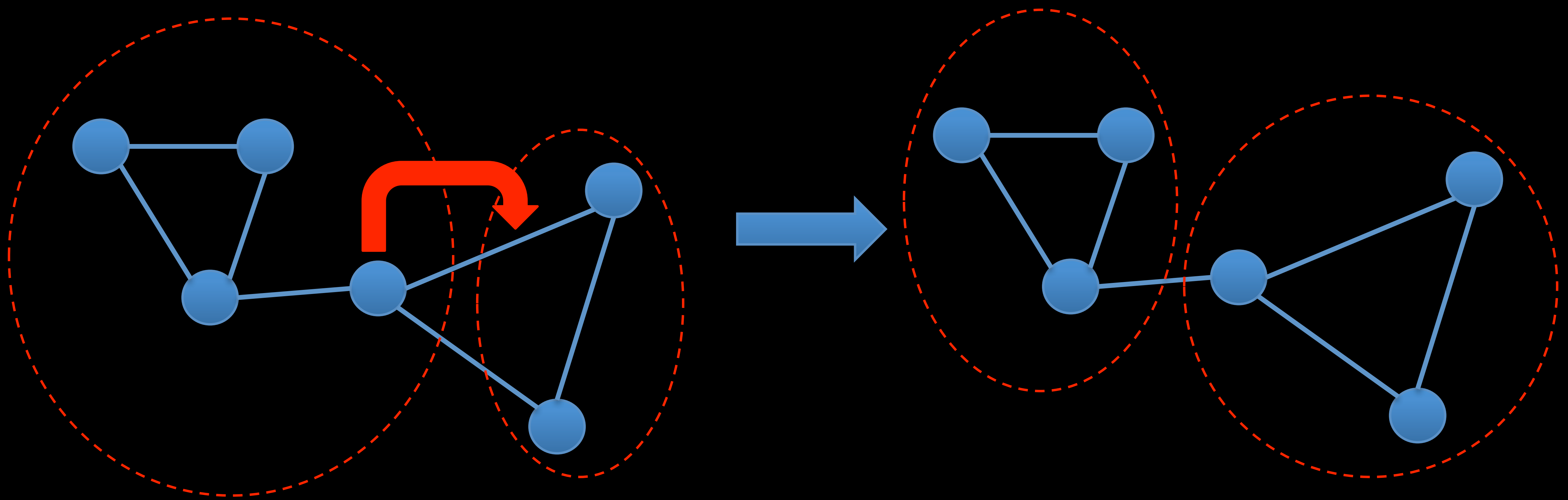
Modularity optimization is usually performed by starting with a certain partition and iteratively adjusting it aiming for modularity score improvement



Modularity optimization - merging



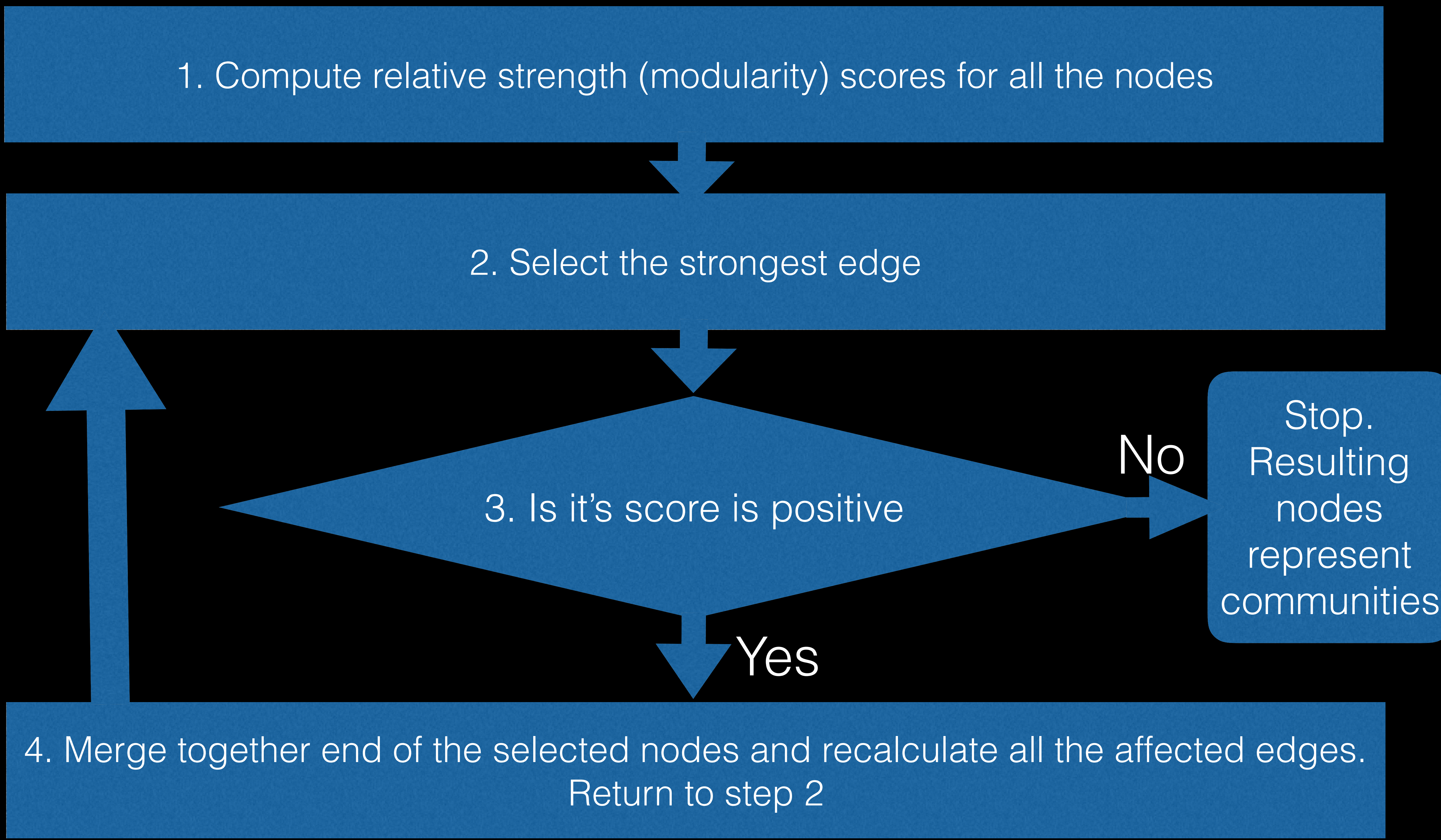
Modularity optimization - shifting



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.

Newman’s fast greedy algorithm

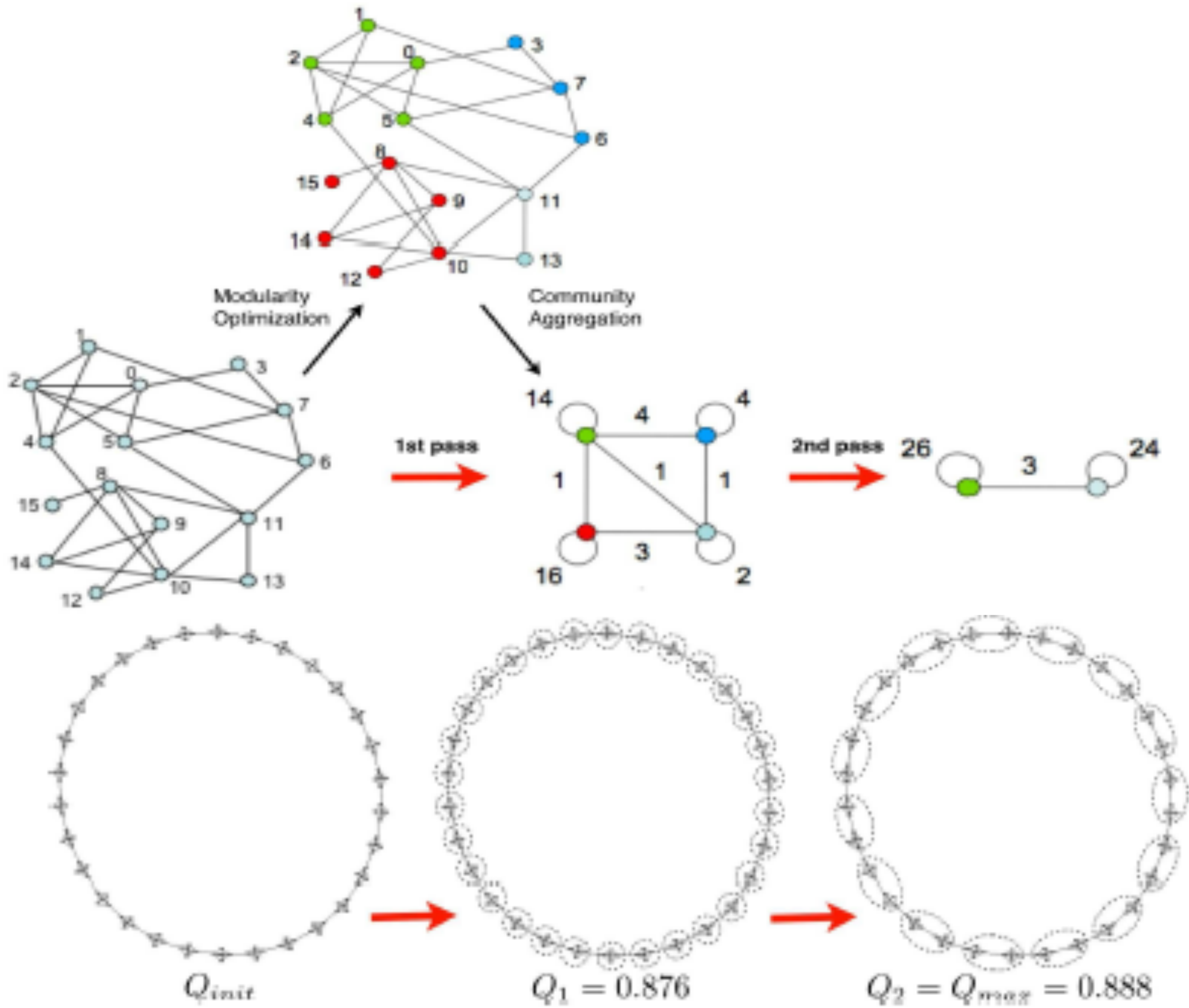


Simulated annealing

- local moves
- splits and merges
- accept or reject each random transformation depending on modularity improvement and changing acceptance criteria

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)

Louvain method

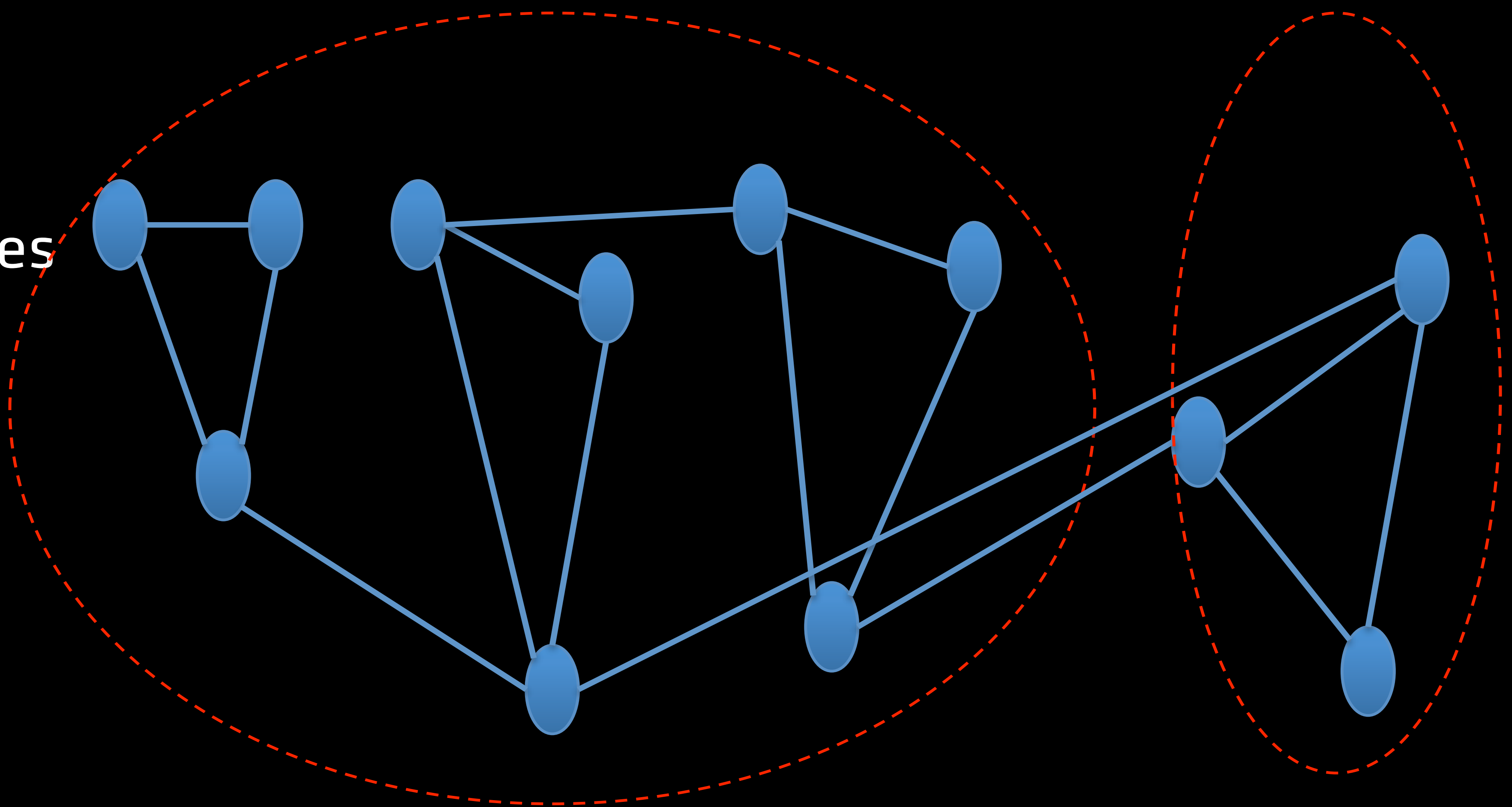


V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech 10008 (2008)

Combo

Iterating best

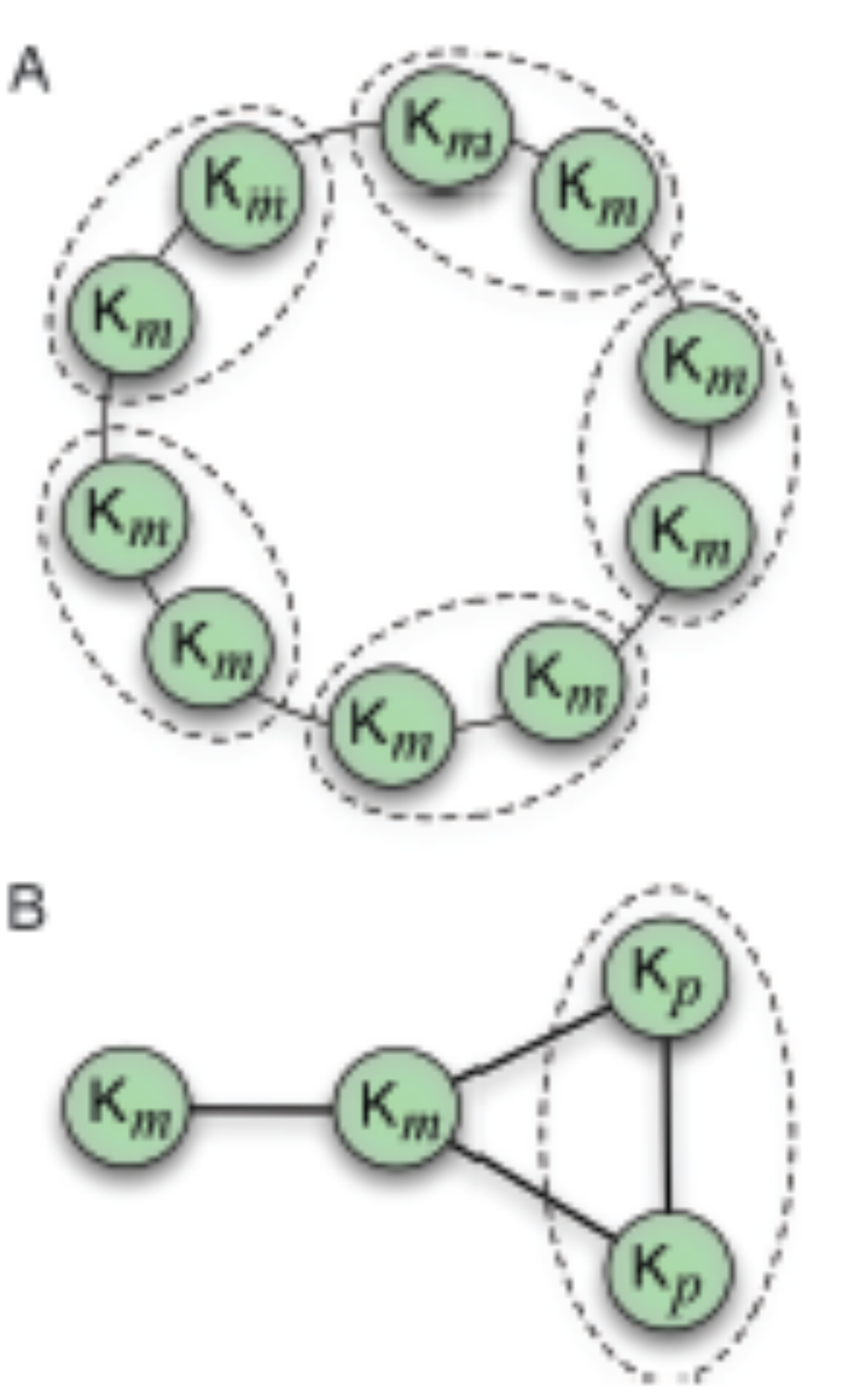
- Splits
- Merges
- Global moves



Sobolevsky, S., Campari, R., Belyi, A., & Ratti, C. (2013). A general optimization technique for high quality community detection in complex networks. *arXiv preprint arXiv:1308.3508*. Accepted in Physical Review E

Algorithm implementation can be downloaded from http://senseable.mit.edu/community_detection/

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),