

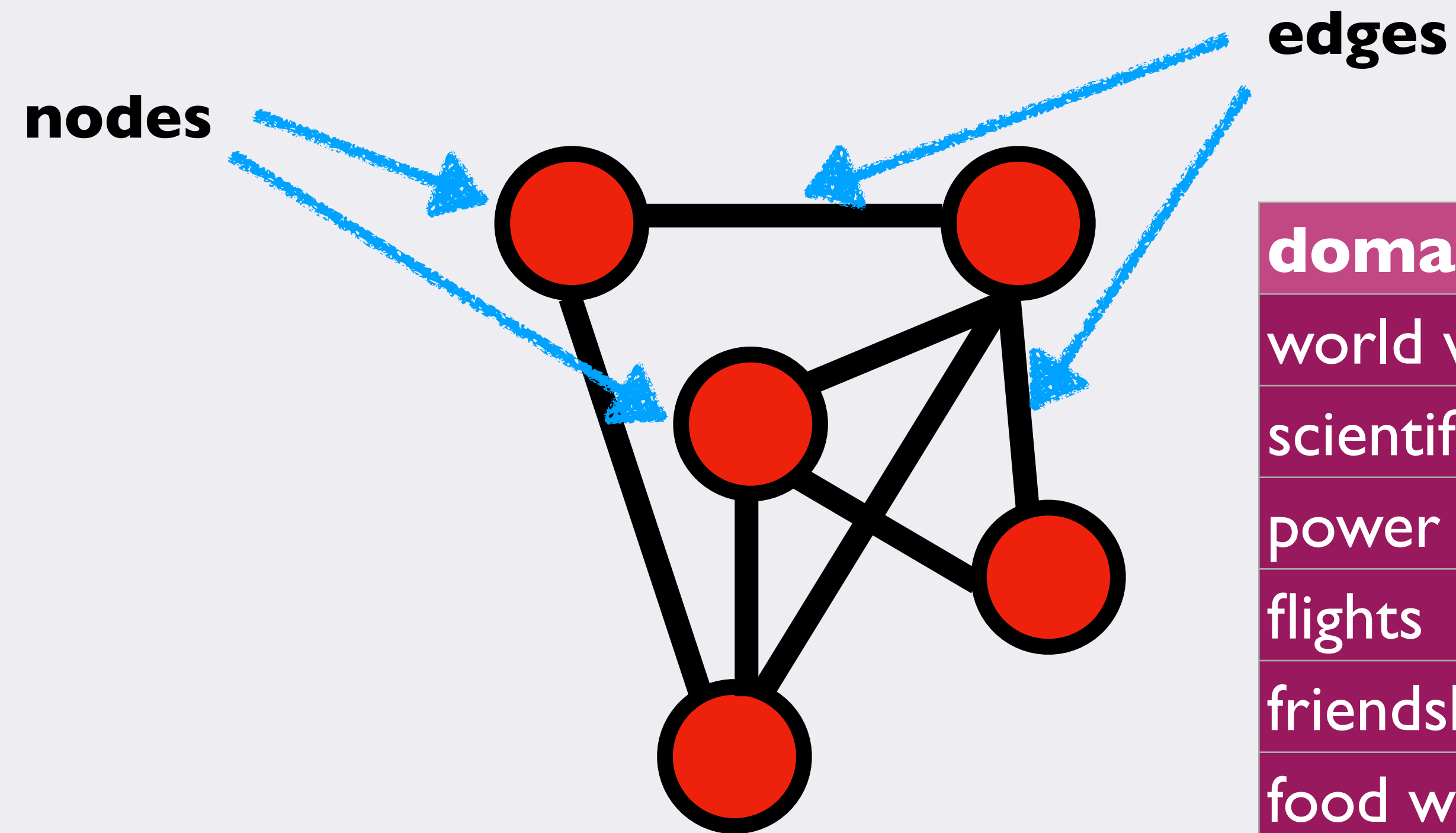


Community Detection

Satyaki Sikdar

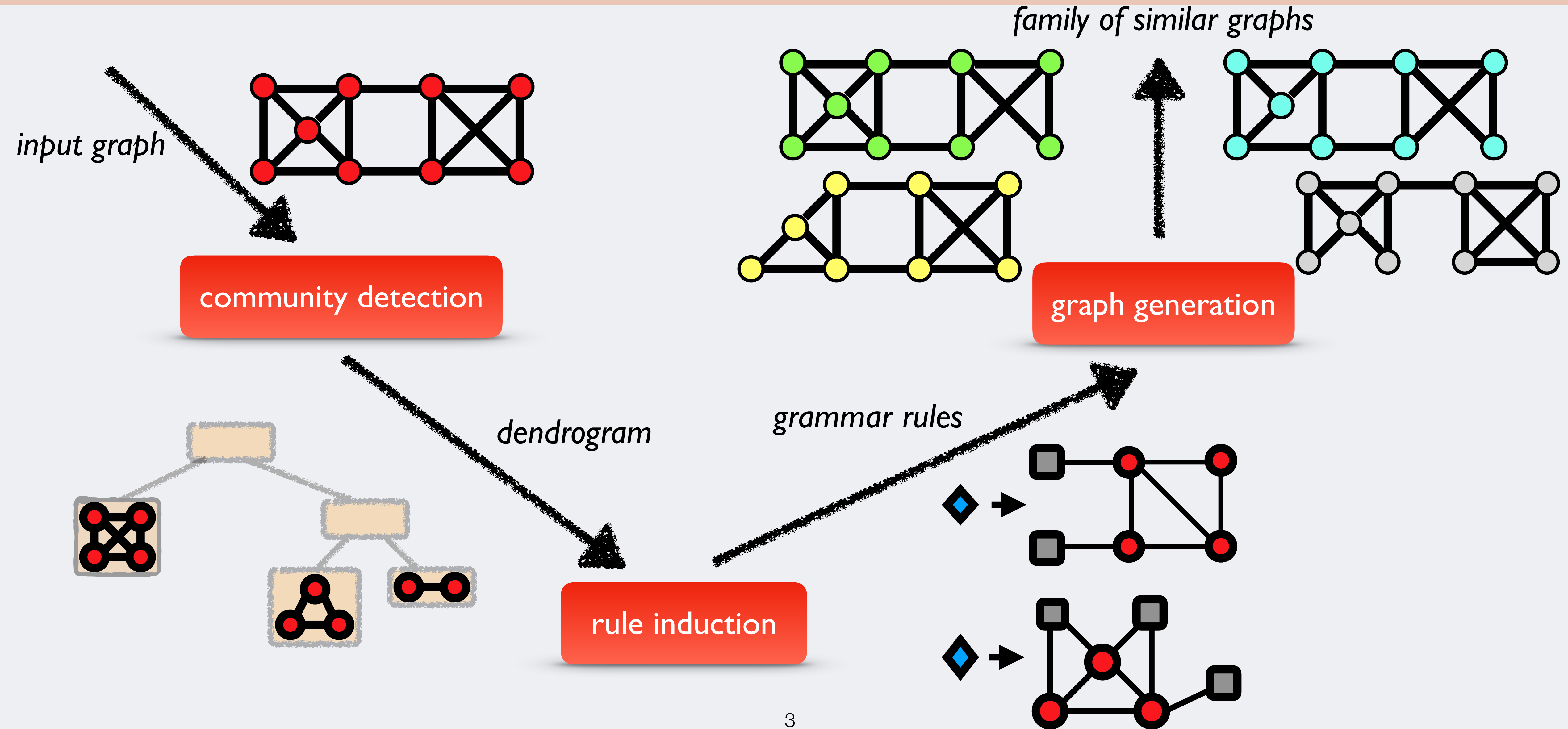
2nd year Ph.D. student

networks - what are they?



domain	nodes	edges
world wide web	webpages	hyperlinks
scientific papers	papers	citations
power grids	generating stations	transmission lines
flights	airports	non-stop flights
friendships	person	friendship
food web	species	predation

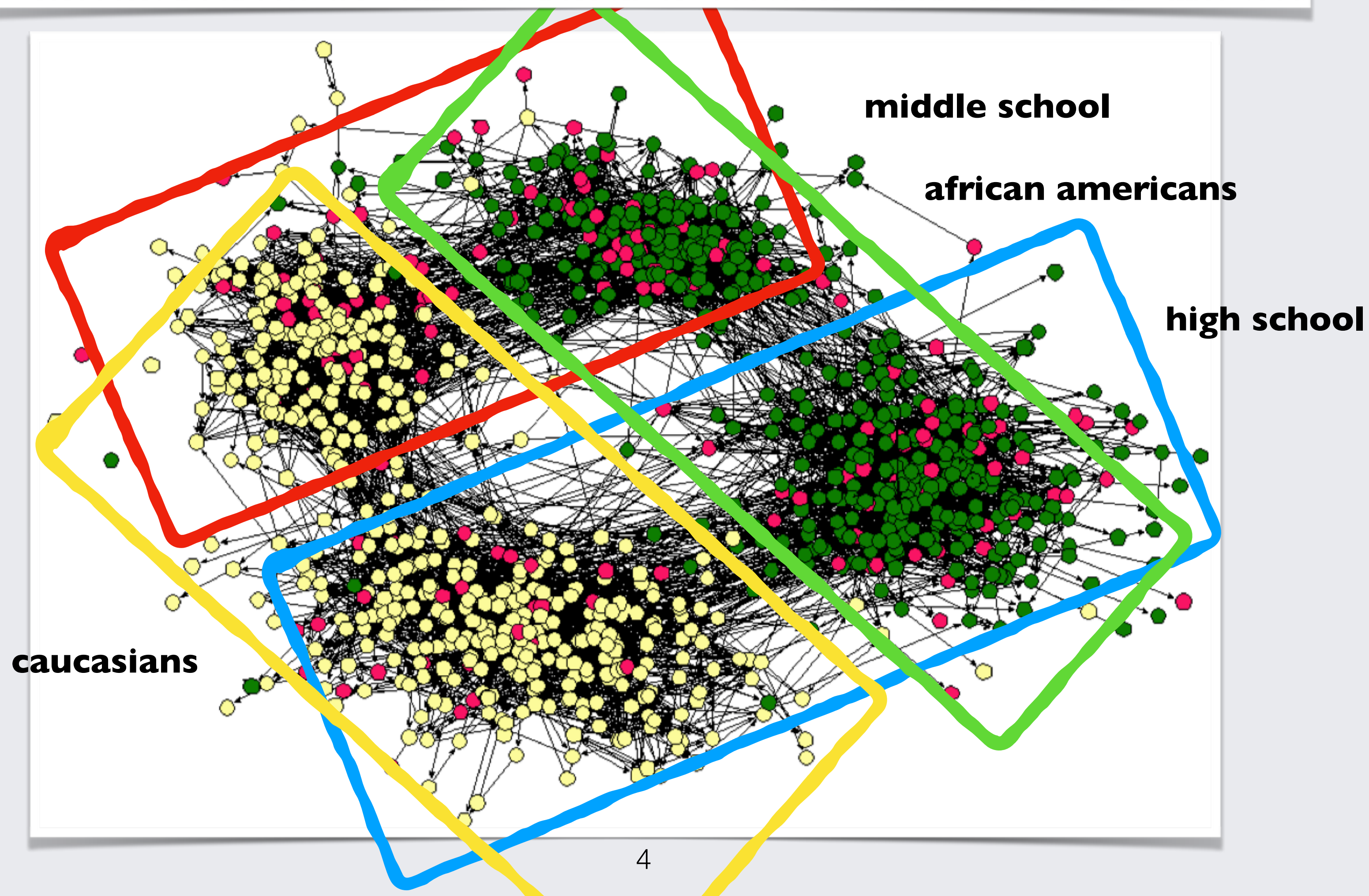
my research pipeline



racial and social homophily

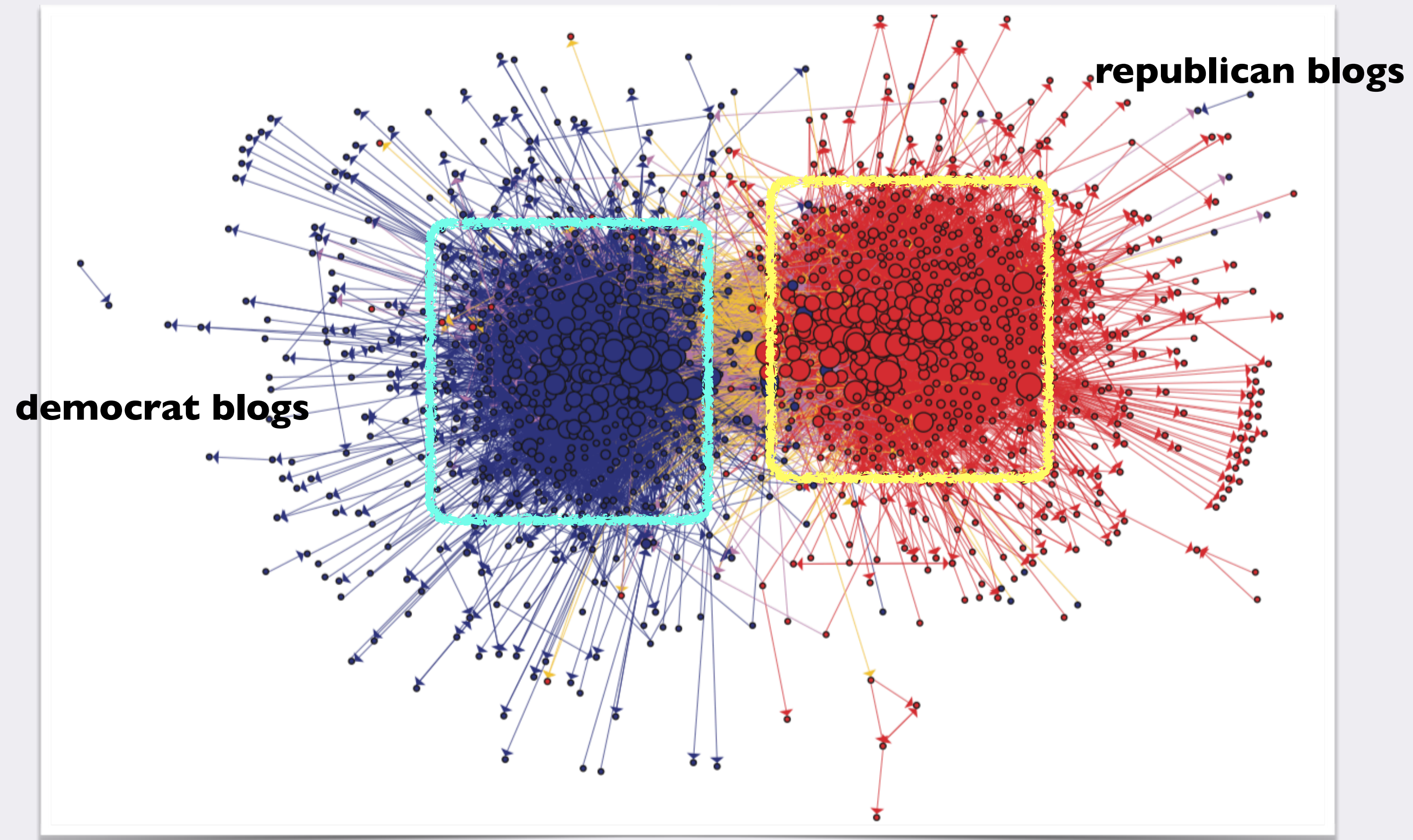
Race, school integration, and friendship segregation in America

[J Moody - American journal of Sociology, 2001 - journals.uchicago.edu](http://journals.uchicago.edu)



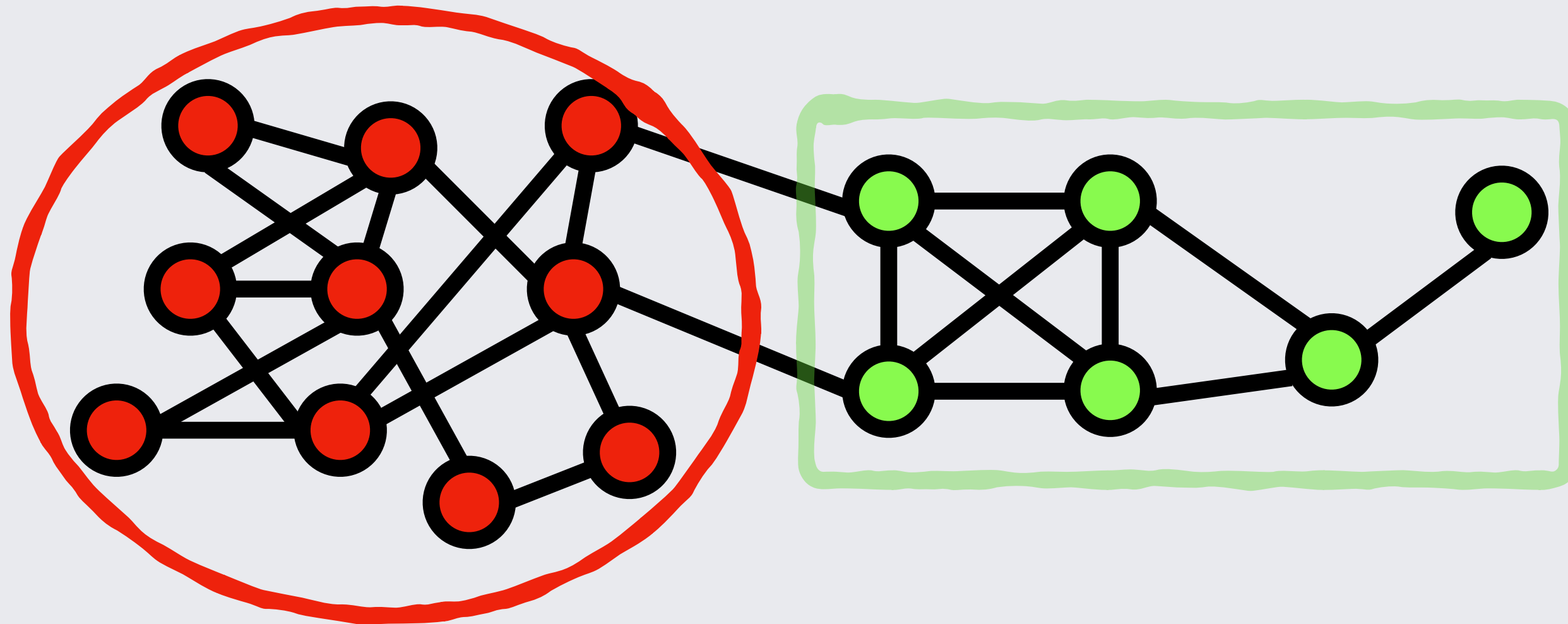
political homophily

The political blogosphere and the 2004 US election: divided they blog
[LA Adamic](#), N Glance - Proceedings of the 3rd international workshop on ..., 2005 - dl.acm.org



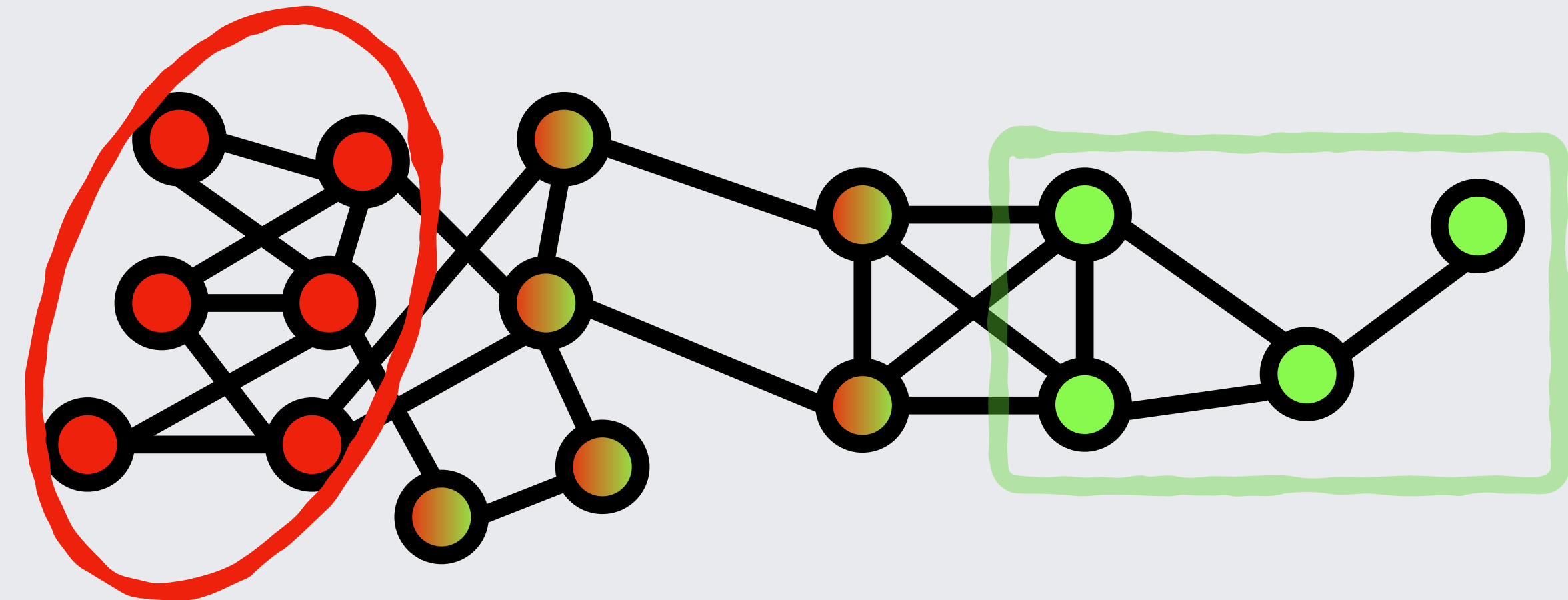
community detection

In a graph $G(V, E)$, find a cover $\mathbb{C} = \{C_1, \dots, C_k\}$ such that $\bigcup_i C_i = V$



disjoint

$$C_i \cap C_j = \emptyset \quad \forall i, j$$



overlapping

$$C_i \cap C_j \neq \emptyset \quad \exists i, j$$

popular techniques

spectral

Normalized cuts and image segmentation

[J Shi](#), [J Malik](#) - IEEE Transactions on pattern analysis and ..., 2000 - [ieeexplore.ieee.org](#)

[PDF] **On spectral clustering: Analysis and an algorithm**

[AY Ng](#), [MJ Jordan](#), [Y Weiss](#) - Advances in neural information ..., 2002 - [papers.nips.cc](#)

- eigenvalues and eigenvectors of adjacency / Laplacian matrix

traversal based

Near linear time algorithm to **detect community** structures in large-scale networks

[UN Raghavan](#), [R Albert](#), [S Kumara](#) - Physical review E, 2007 - APS

Fast detection of community structures using graph traversal in social networks

[P Basuchowdhuri](#), [S Sikdar](#), [V Nagarajan](#)... - ... and Information Systems, 2017 - Springer

- discovery of local neighborhoods and bridges

greedy optimization

Fast unfolding of **communities** in large networks

[VD Blondel](#), [JL Guillaume](#), [R Lambiotte](#)... - Journal of statistical ..., 2008 - [iopscience.iop.org](#)

General optimization technique for high-quality **community detection** in complex networks

[S Sobolevsky](#), [R Campari](#), [A Belyi](#), [C Ratti](#) - Physical Review E, 2014 - APS

- agglomerative or divisive hierarchical clustering

information theory based

The **map equation**

[M Rosvall](#), [D Axelsson](#), [CT Bergstrom](#) - The European Physical Journal ..., 2009 - Springer

Clique percolation in random networks

[I Derényi](#), [G Palla](#), [T Vicsek](#) - Physical review letters, 2005 - APS

- graph compression through encodings

random walks

Computing **communities** in large networks using **random walks**

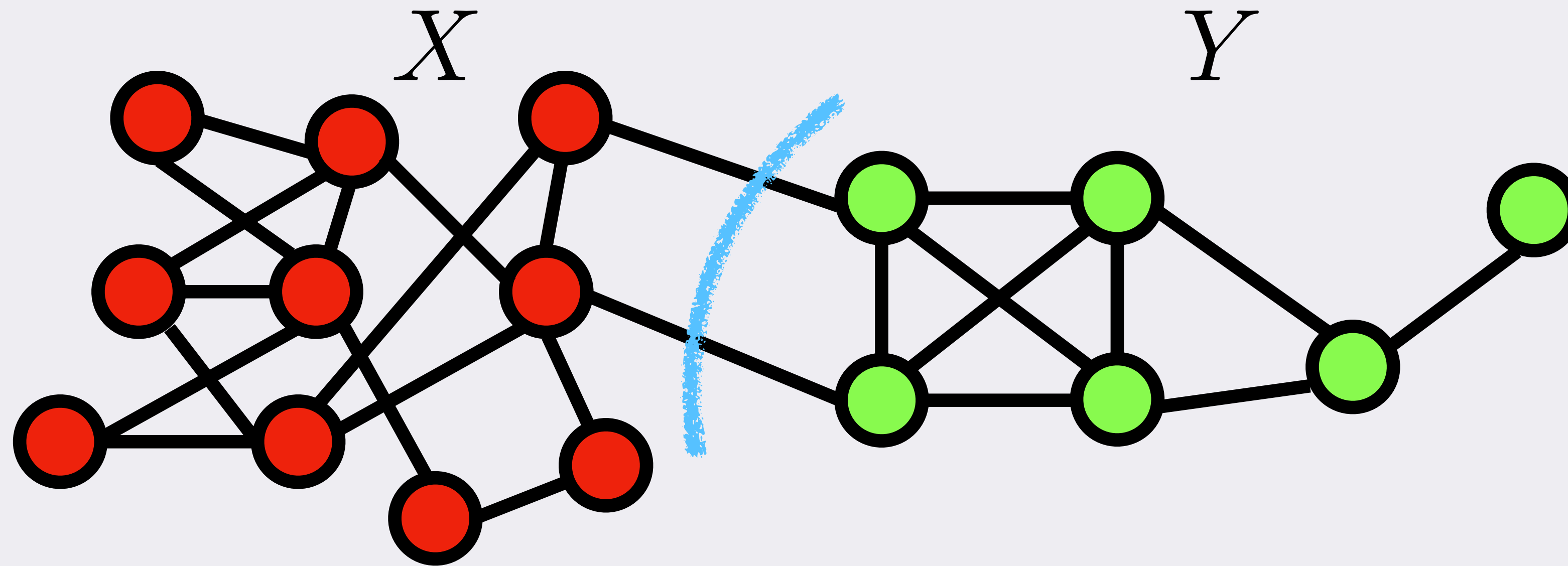
[P Pons](#), [M Latapy](#) - International symposium on computer and information ..., 2005 - Springer

Efficient and principled method for detecting communities in networks

[B Ball](#), [B Karrer](#), [MEJ Newman](#) - Physical Review E, 2011 - APS

- distribution of node visits through multiple random walks

graph partitioning - cuts



degree

$$d_i = \sum_j A_{ij}$$

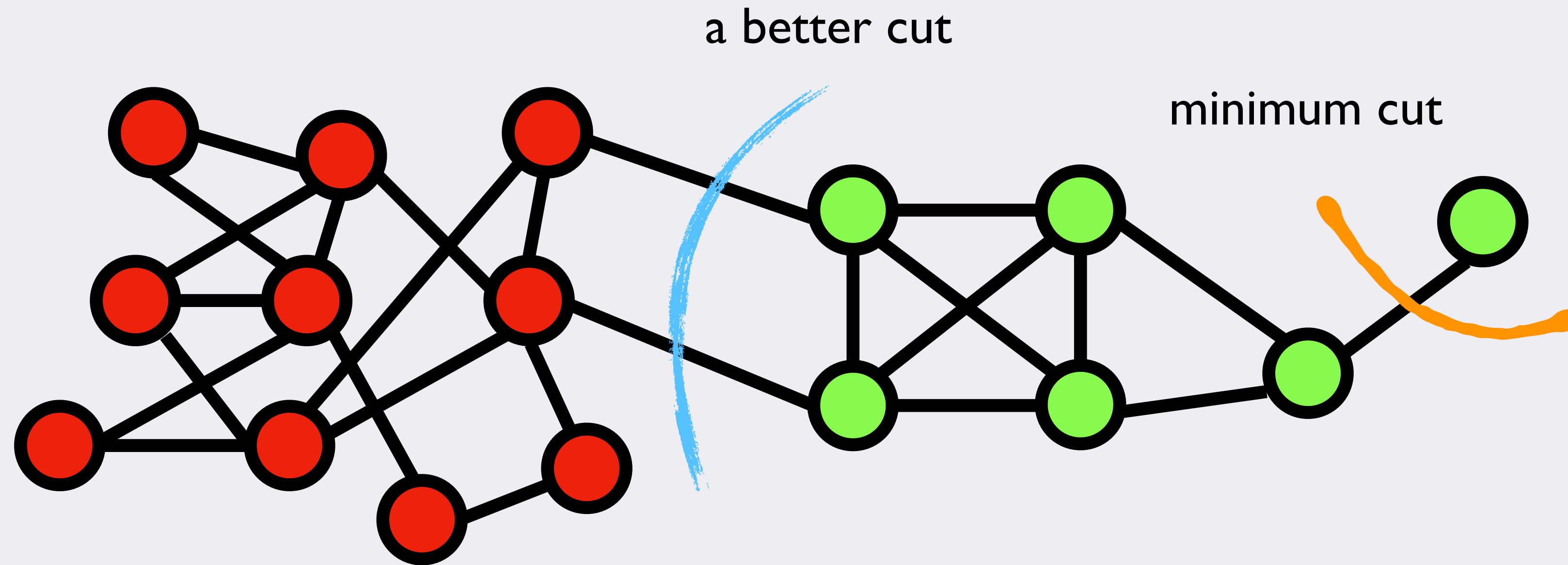
volume

$$\text{vol}(X) = \sum_{j \in X} d_j$$

cut

$$\text{cut}(X, Y) = \sum_{i \in X, j \in Y} A_{ij}$$

graph partitioning - cuts



min cut is not necessarily the **best** cut

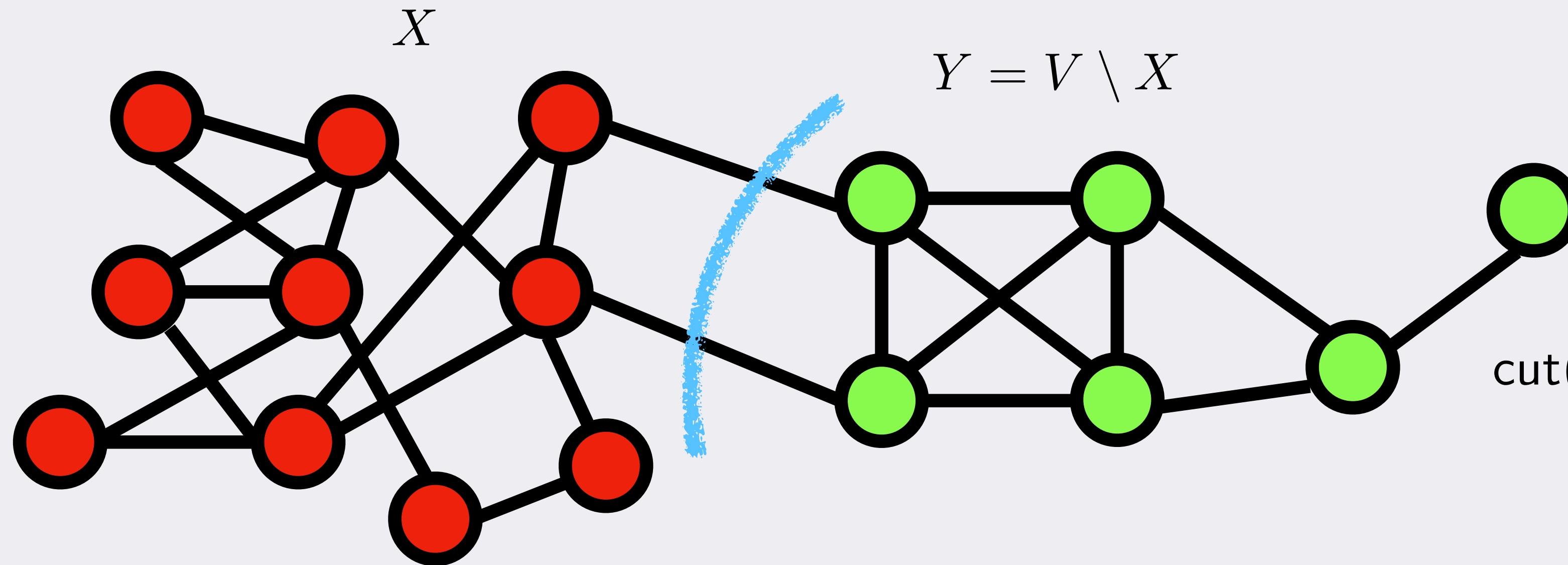
graph partitioning - quality measures

degree

$$d_i = \sum_j A_{ij}$$

volume

$$\text{vol}(X) = \sum_{j \in X} d_j$$



cut

$$\text{cut}(X, Y) = \sum_{i \in X, j \in Y} A_{ij}$$

conductance

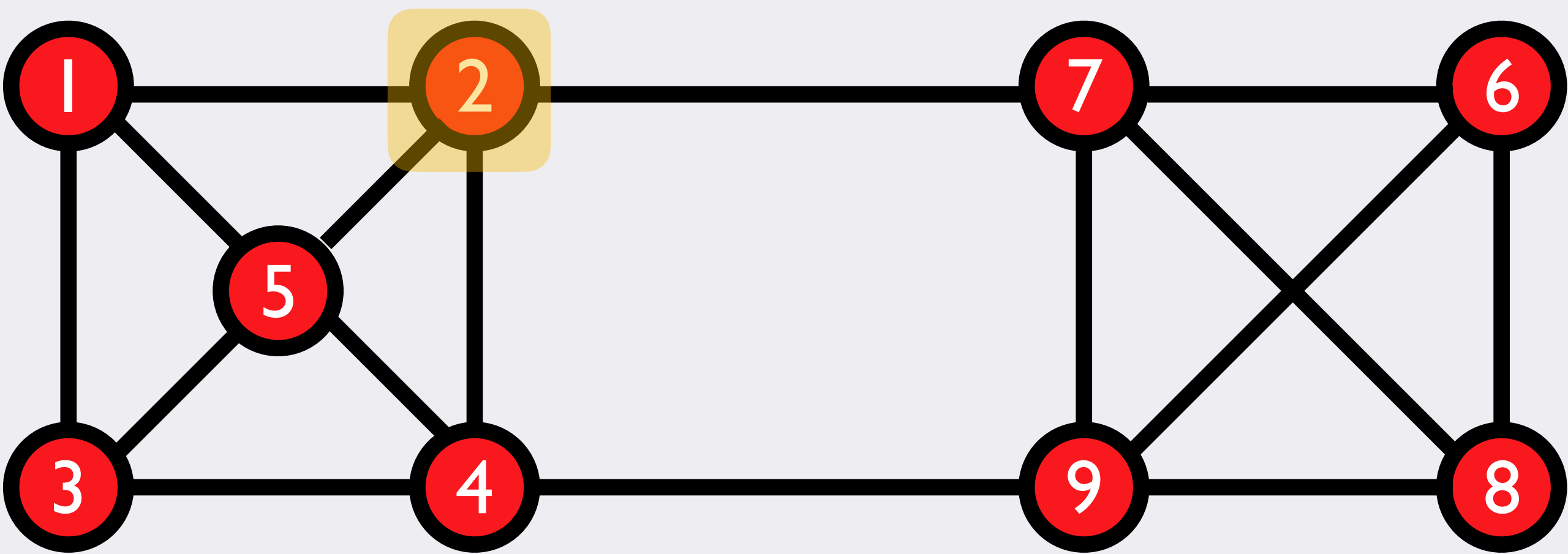
$$\phi(X) = \frac{\text{cut}(X, V \setminus X)}{\min\{\text{vol}(X), \text{vol}(V \setminus X)\}}$$

normalized cut

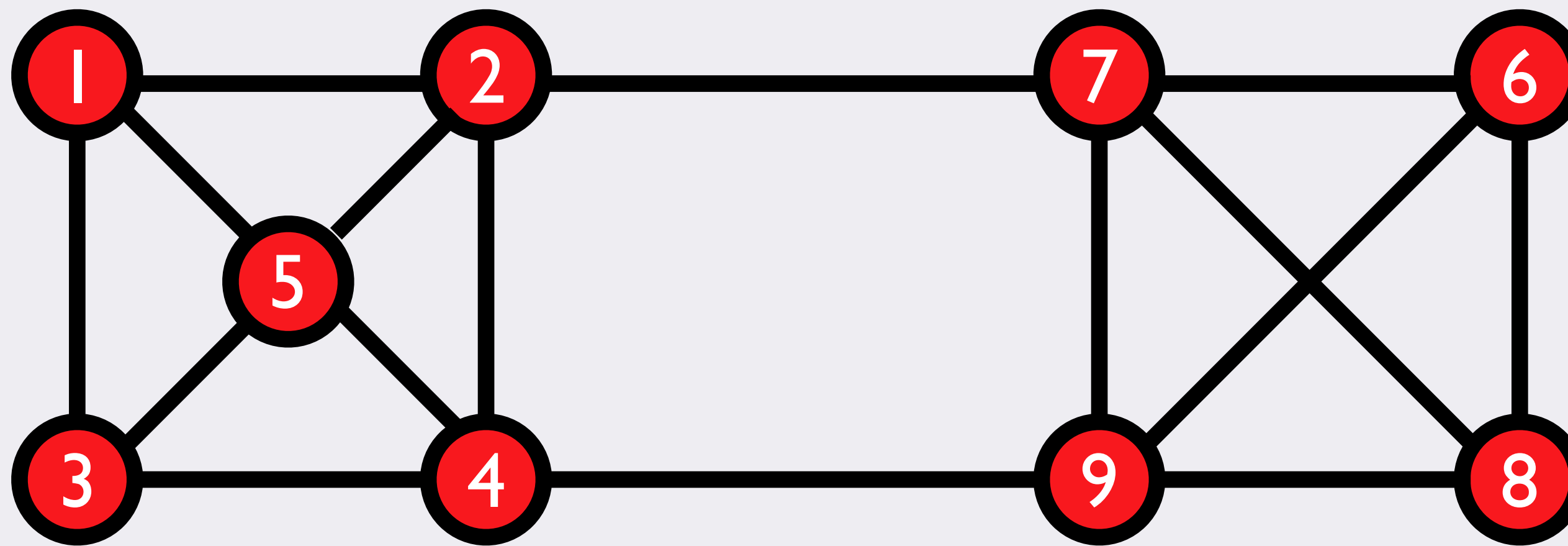
$$\text{ncut}(X) = \frac{\text{cut}(X, V \setminus X)}{\text{vol}(X)} + \frac{\text{cut}(X, V \setminus X)}{\text{vol}(V \setminus X)}$$

Kannan, R., Vempala, S., & Vetta, A. (2004). On clusterings: Good, bad and spectral. Journal of the ACM (JACM), 51(3), 497-515
 Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.

example network



<i>A</i>	1	2	3	4	5	6	7	8	9
1	0	1	1	0	1	0	0	0	0
2	1	0	0	1	1	0	1	0	0
3	1	0	0	1	1	0	0	0	0
4	0	1	1	0	1	0	0	0	1
5	1	1	1	1	0	0	0	0	0
6	0	0	0	0	0	0	1	1	1
7	0	1	0	0	0	1	0	1	1
8	0	0	0	0	0	1	1	0	1
9	0	0	0	1	0	1	1	1	0



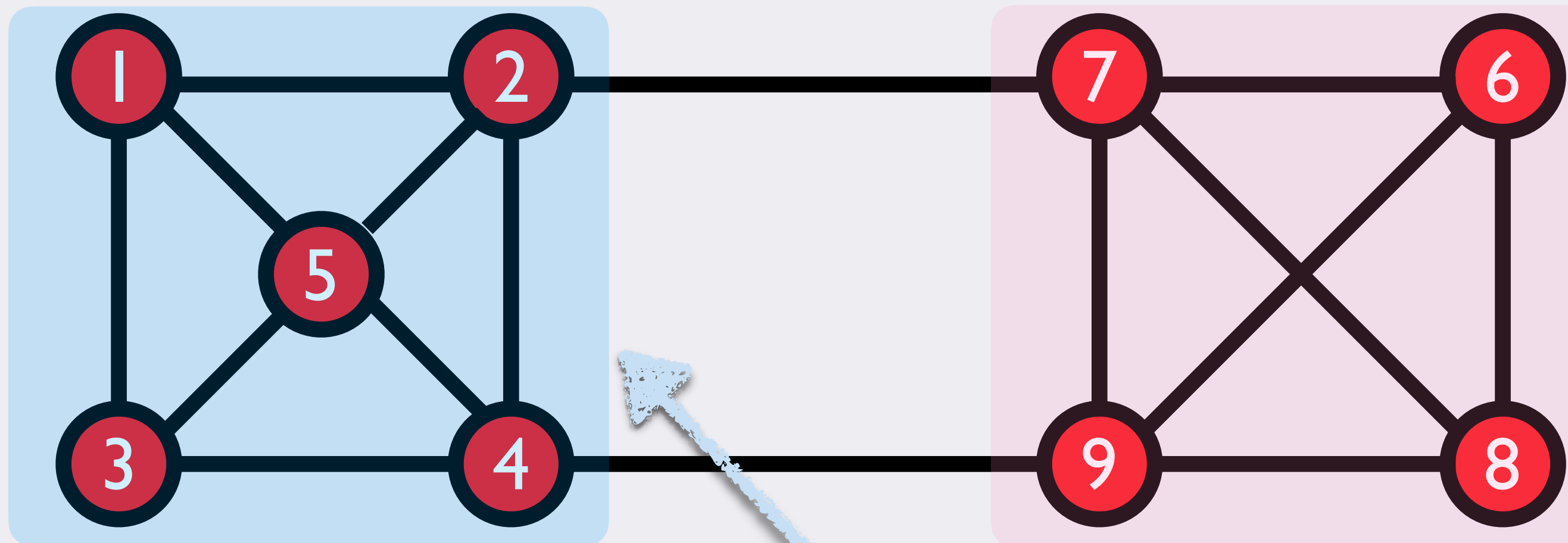
$$L = D - A$$

D	1	2	3	4	5	6	7	8	9
1	3								
2		4							
3			3						
4				4					
5					4				
6						3			
7							4		
8								3	
9									4

A	1	2	3	4	5	6	7	8	9
1		1	1		1				
2				1	1		1		
3				1	1				
4		1	1		1				1
5	1	1	1	1					
6							1	1	1
7		1				1		1	1
8						1	1		1
9				1		1	1	1	

L	1	2	3	4	5	6	7	8	9
1	3	-1	-1		-1				
2	-1	4		-1	-1			-1	
3	-1		3	-1	-1				
4		-1	-1	4	-1				-1
5	-1	-1	-1	-1	4				
6						3	-1	-1	-1
7		-1				-1	4	-1	-1
8						-1	-1	3	-1
9				-1		-1	-1	-1	4

$$Lx = \lambda x$$



L	1	2	3	4	5	6	7	8	9
1	3	-1	-1		-1				
2	-1	4		-1	-1		-1		
3	-1		3	-1	-1				
4		-1	-1	4	-1				-1
5	-1	-1	-1	-1	4				
6						3	-1	-1	-1
7		-1				-1	4	-1	-1
8						-1	-1	3	-1
9				-1		-1	-1	-1	4

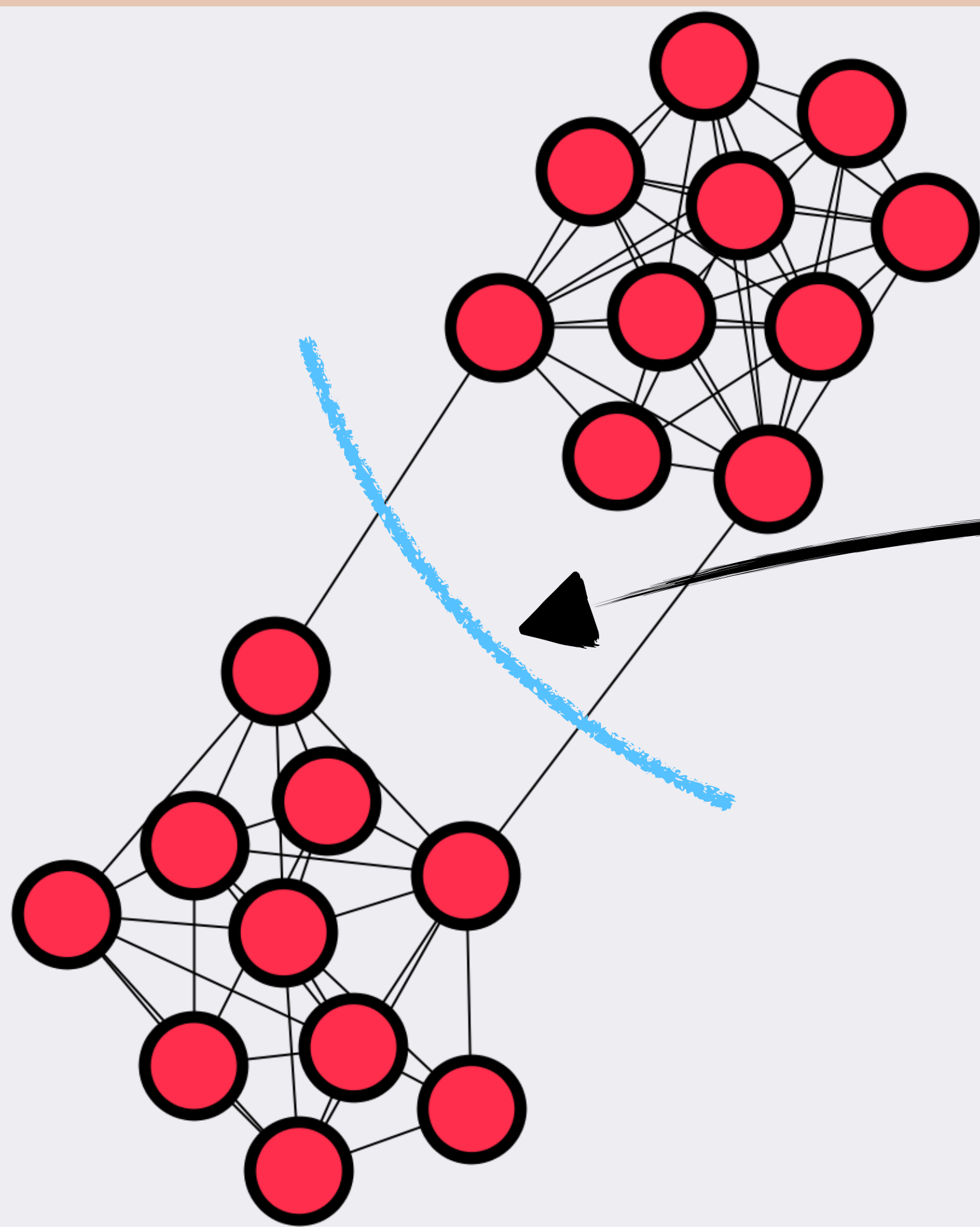
sorted eigenvalues

$9.540 \text{ e} - 17$
$6.498 \text{ e} - 01$
3.198
3.326
4
4.554
4.641
5.382
6.246

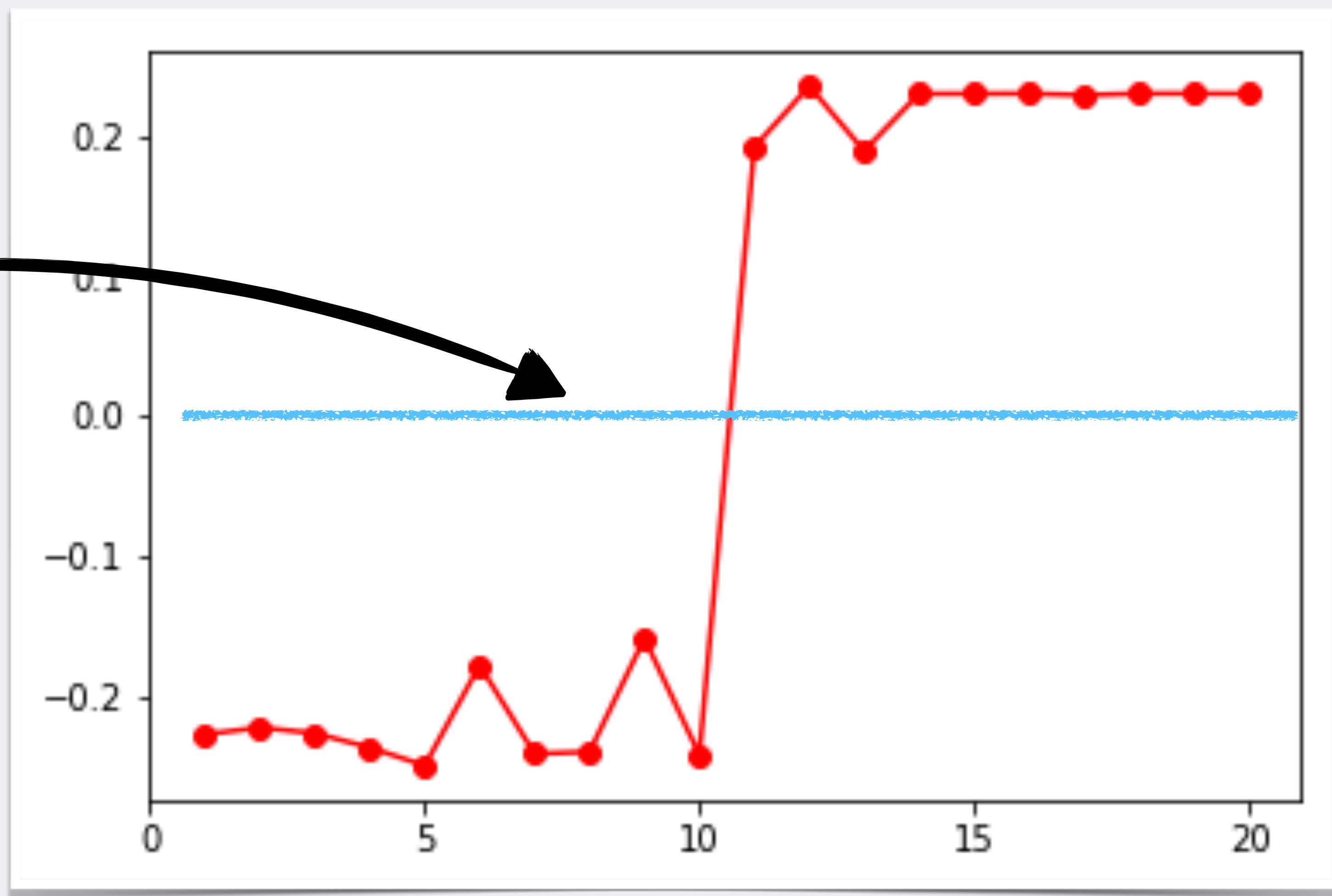
fiedler vector

1	-0.378
2	-0.178
3	-0.378
4	-0.332
5	-0.178
6	0.291
7	0.291
8	0.431
9	0.431

bipartition



fiedler vector value



node id

k-way partitions

recursive bipartitions

New spectral methods for ratio cut partitioning and clustering

[L Hagen](#), [AB Kahng](#) - ... transactions on computer-aided design of ..., 1992 - ieeexplore.ieee.org

- hierarchical divisive clustering based on median / fixed value
- disadvantages: unstable, computationally expensive

Normalized cuts and image segmentation

[J Shi](#), [J Malik](#) - IEEE Transactions on pattern analysis and ..., 2000 - ieeexplore.ieee.org

clustering multiple eigenvectors

[PDF] On spectral clustering: Analysis and an algorithm

[AY Ng](#), [MI Jordan](#), [Y Weiss](#) - Advances in neural information ..., 2002 - papers.nips.cc

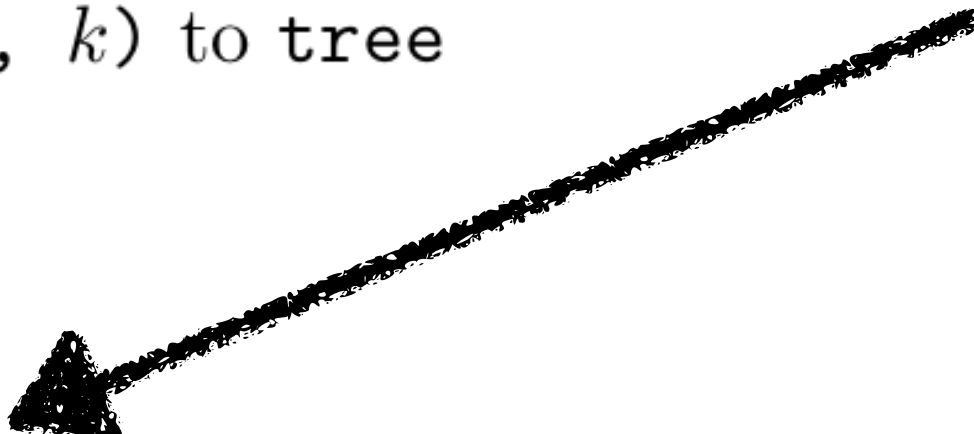
- spectral embedding of nodes and then k-means
- more efficient than recursive bipartitions

pseudocode

Algorithm 1: `approx_min_cut(G, k)`

```
1 begin
2   tree ← empty list
3   if  $G$  has  $\leq k$  nodes then
4     | return nodes in  $G$ 
5   end
6   if  $G$  is not connected then
7     | foreach connected component  $p$  of  $G$  do
8       | Append approx_min_cut( $p, k$ ) to tree
9     | end
10    return tree
11  end
12  fiedler ← Fiedler vector of  $G$ 
13   $p_1$  ← node ids with value less than the median of fiedler
14   $p_2$  ← rest of the nodes in  $G$ 
15   $SG_1$  ← subgraph induced by  $p_1$  on  $G$ 
16   $SG_2$  ← subgraph induced by  $p_2$  on  $G$ 
17  Append approx_min_cut( $SG_1, k$ ) to tree
18  Append approx_min_cut( $SG_2, k$ ) to tree
19  return tree
20 end
```

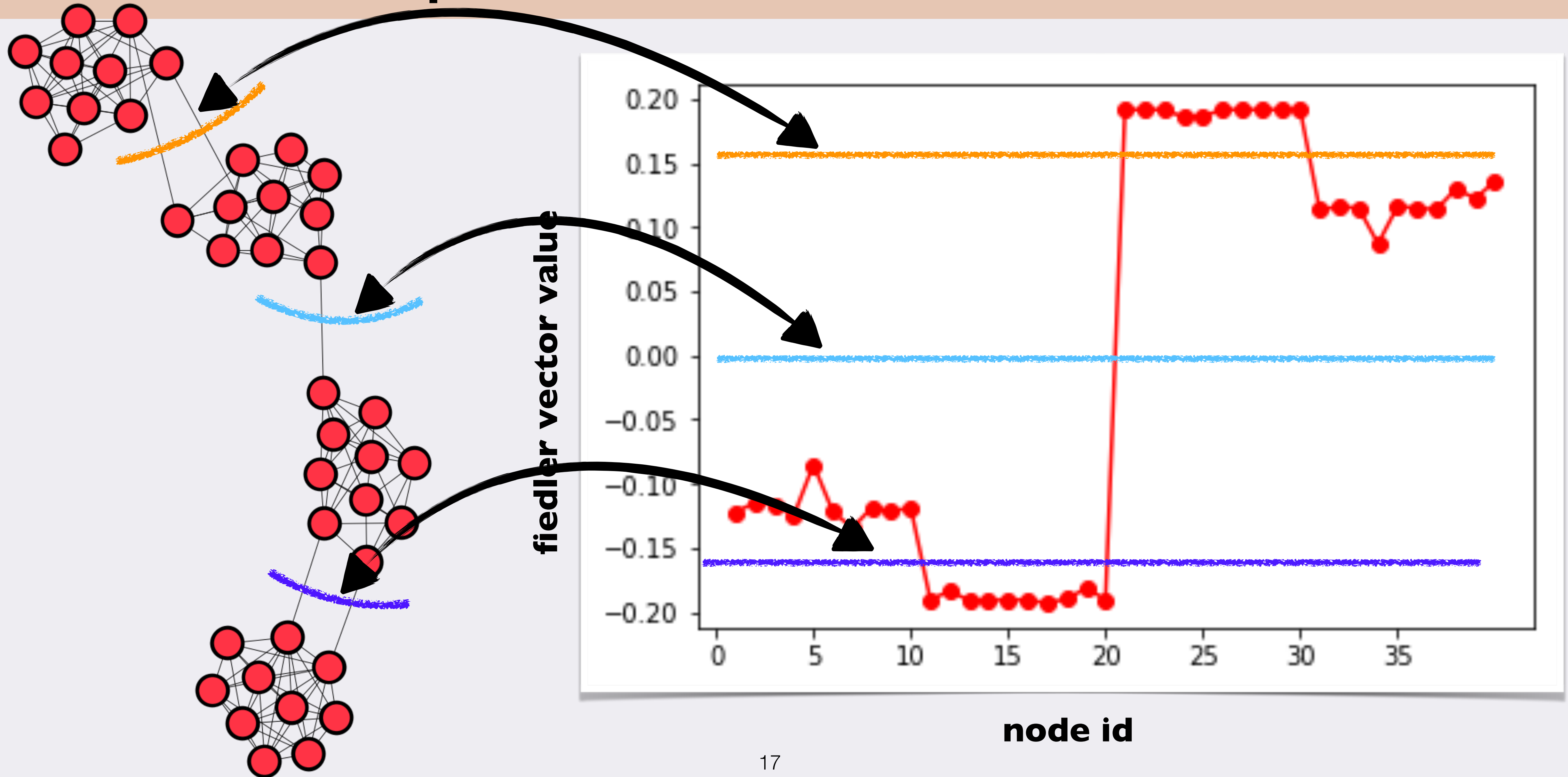
$\mathcal{O}(n^3)$



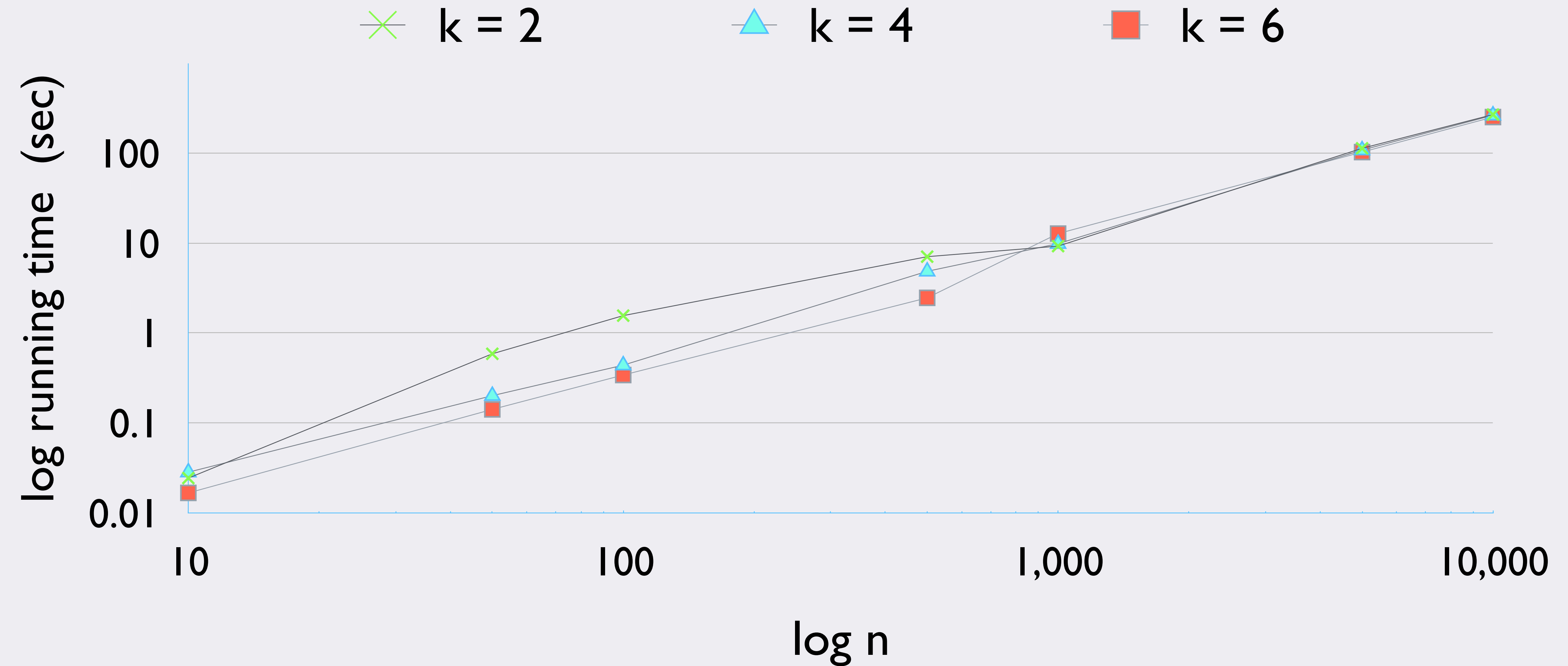
$$T(n) \approx 2T\left(\frac{n}{2}\right) + \mathcal{O}(n^3)$$

$$T(n) = \Theta(n^3)$$

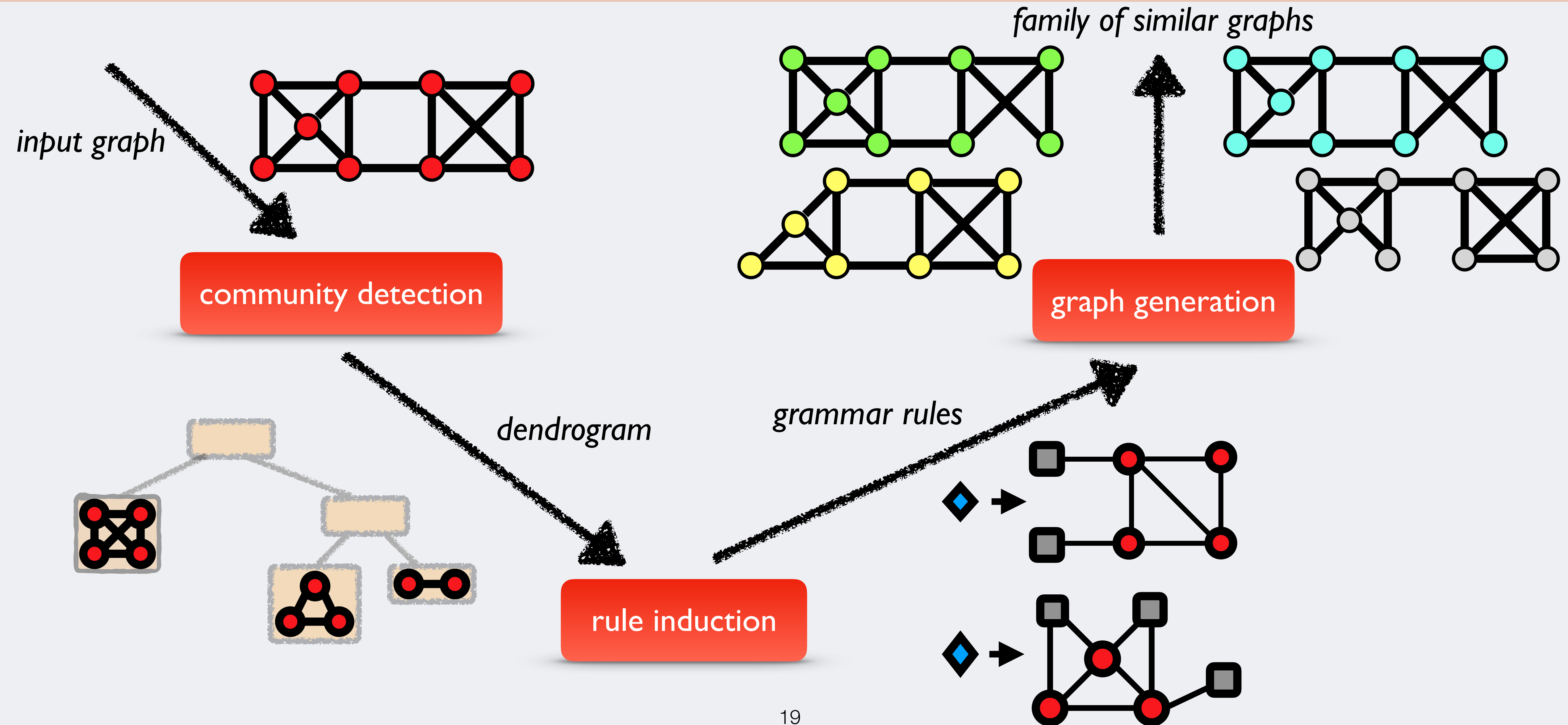
recursive bipartition



running times



pipeline revisited



thanks!