Discrete Kernels

Kernels for Sequences

- Similarity between sequences of different lengths

ACGGTTCAA

ATATCGCGGGAA

Count Kernel

- Inner product between symbol counts

	A	C	G	T
ACGGTTCAA	3	2	2	2
ATATCGCGGGAA	4	2	4	2

- Extension: Spectrum kernels (Leslie et al., 2002)
  - Counts the number of k-mers (k-grams) efficiently

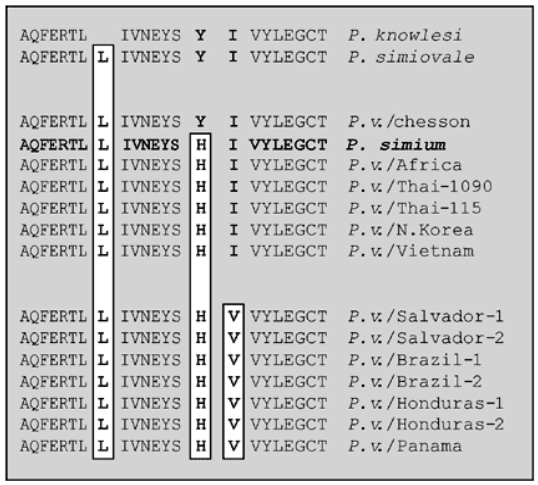
Motivations for graph analysis

- Existing methods assume ” tables”

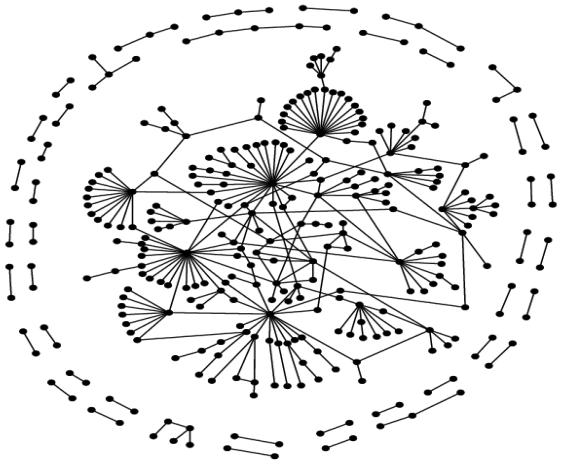
Serial Num	Name	Age	Sex	Address	…
0001	○○	40	Male	Tokyo	…
0002	× ×	31	Female	Osaka	…

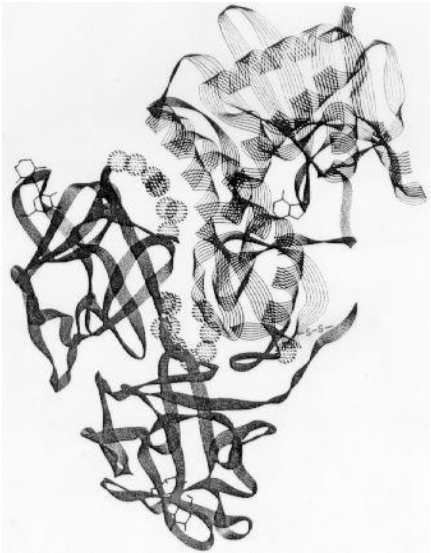
- Structured data beyond this framework

→ New methods for analysis



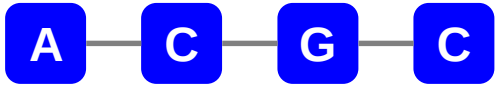
AQFERTL		IVNEYS	Y	I	VYLEGCT	*P. knowlesi*
AQFERTL	L	IVNEYS	Y	I	VYLEGCT	*P. simiovale*
AQFERTL	L	IVNEYS	Y	I	VYLEGCT	*P.v./chesson*
AQFERTL	**L**	**IVNEYS**	**H**	**I**	**VYLEGCT**	***P. simium***
AQFERTL	L	IVNEYS	H	I	VYLEGCT	*P.v./Africa*
AQFERTL	L	IVNEYS	H	I	VYLEGCT	*P.v./Thai-1090*
AQFERTL	L	IVNEYS	H	I	VYLEGCT	*P.v./Thai-115*
AQFERTL	L	IVNEYS	H	I	VYLEGCT	*P.v./N.Korea*
AQFERTL	L	IVNEYS	H	I	VYLEGCT	*P.v./Vietnam*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Salvador-1*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Salvador-2*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Brazil-1*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Brazil-2*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Honduras-1*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Honduras-2*
AQFERTL	L	IVNEYS	H	V	VYLEGCT	*P.v./Panama*



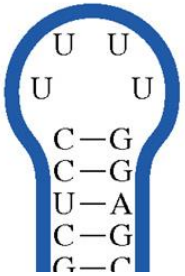


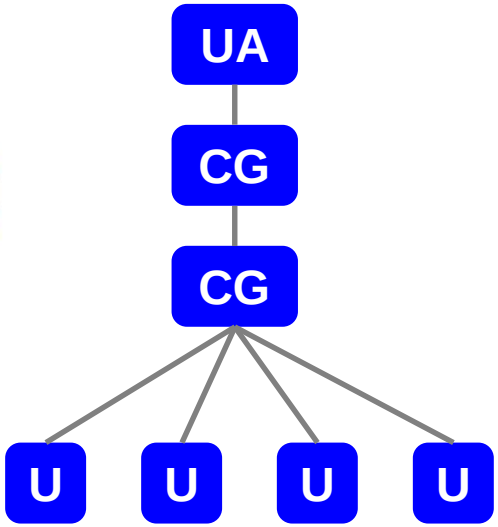
Graph Structures in Biology

- DNA Sequence

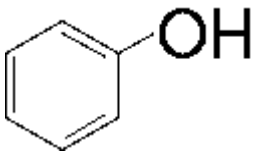


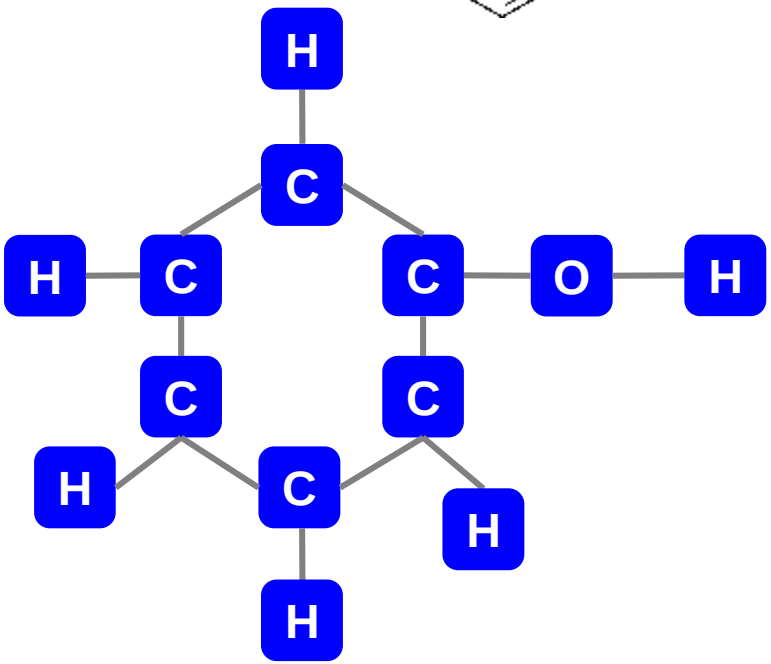
- RNA





- Compounds

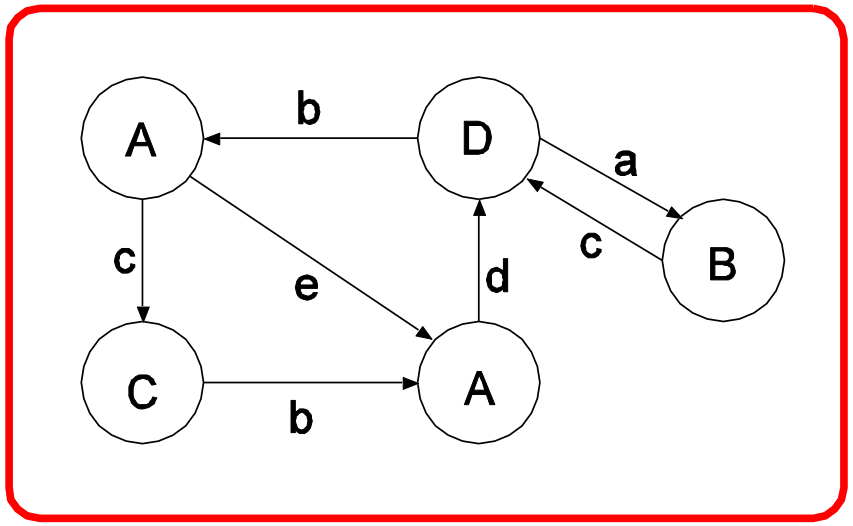




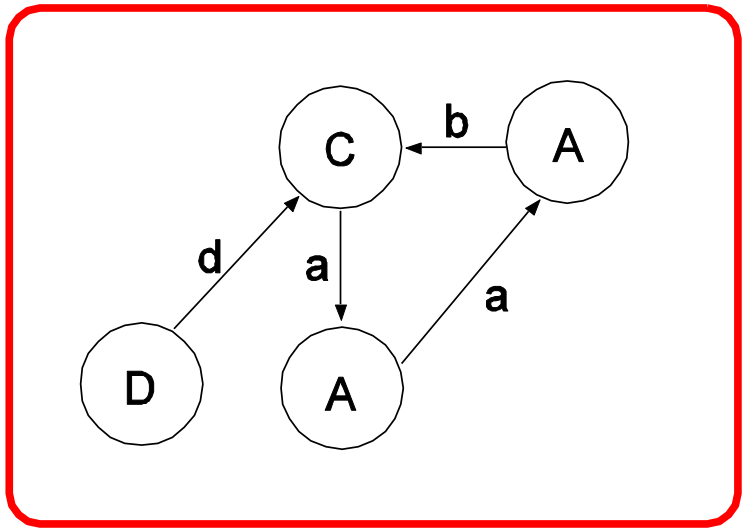
Graph Kernels

(Kashima, Tsuda, Inokuchi, ICML 2003)

- Going to define the kernel function $K(G, G')$
- Both vertex and edges are labeled



G



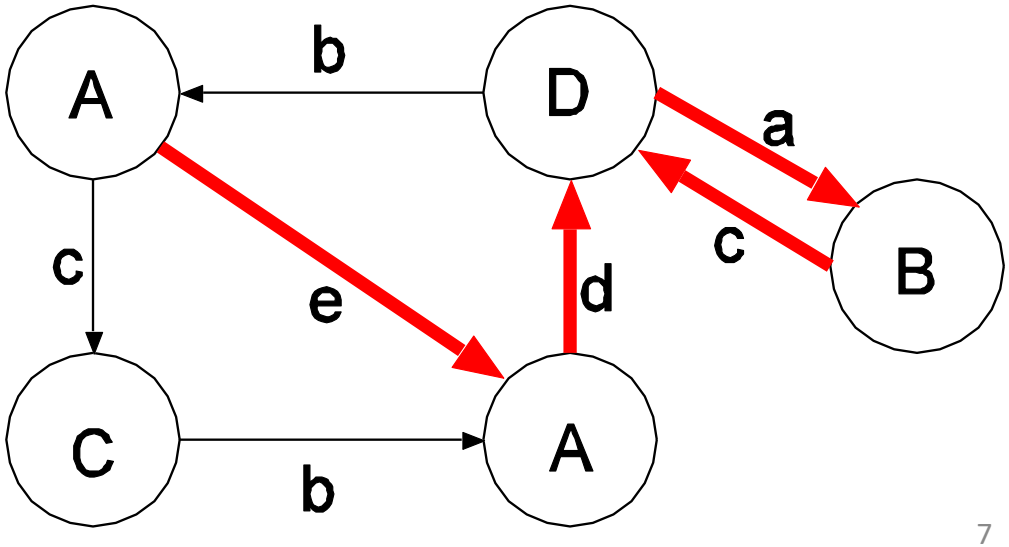
G'

Label path

- Sequence of vertex and edge labels

$$\boldsymbol{h} = (A, e, A, d, D, a, B, c, D)$$

- Generated by random walking
- Uniform initial, transition, terminal probabilities



Path-probability vector

Label path $\boldsymbol{h}$	Probability $p(\boldsymbol{h}\|G)$
AaA	0.001
$\vdots$	$\vdots$
AcDbE	0.000003
$\vdots$	$\vdots$
AeAdDaBcD	0.00000007
$\vdots$	$\vdots$

Kernel definition

- Kernels for paths

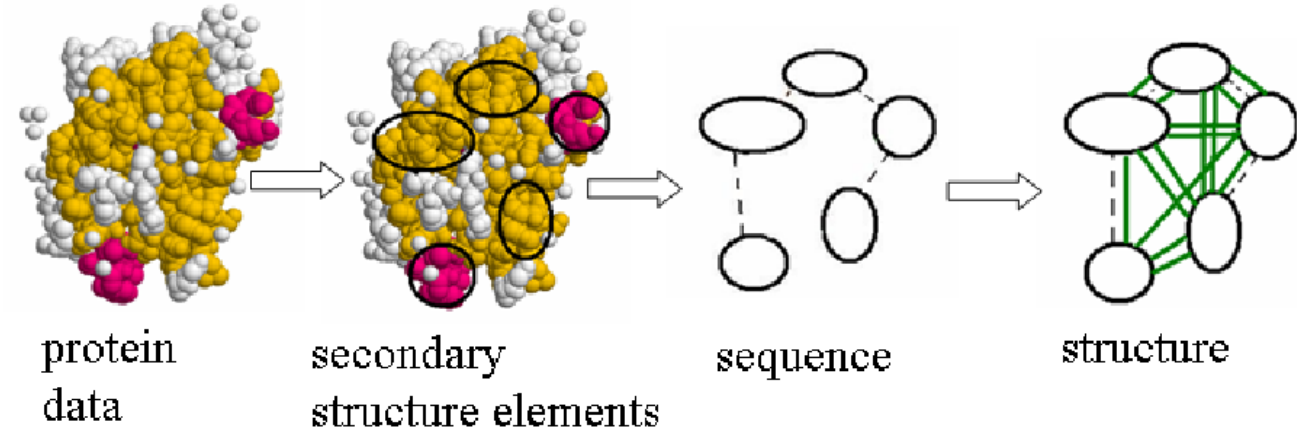
$$K(\boldsymbol{h}, \boldsymbol{h}') = \begin{cases} 0 & (|\boldsymbol{h}| \neq |\boldsymbol{h}'|) \\ k_v(h_1, h_1')k_e(h_2, h_2') \cdots k_v(h_\ell, h_\ell') & (|\boldsymbol{h}| = |\boldsymbol{h}'|) \end{cases}$$

- Take expectation over all possible paths!
- Marginalized kernels for graphs

$$K(G, G') = \sum_{\boldsymbol{h}} \sum_{\boldsymbol{h}'} p(\boldsymbol{h}|G) p(\boldsymbol{h}'|G') K(\boldsymbol{h}, \boldsymbol{h}')$$

Classification of Protein 3D structures

- Graphs for protein 3D structures
  - Node: Secondary structure elements
  - Edge: Distance of two elements
- Calculate the similarity by graph kernels



Borgwardt et al. "*Protein function prediction via graph kernels*", ISMB2005

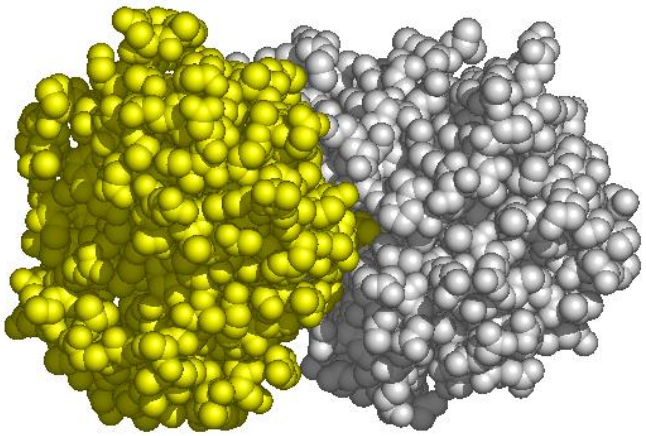
Biological Networks

- Protein-protein physical interaction
- Metabolic networks
- Gene regulatory networks
- Network induced from sequence similarity

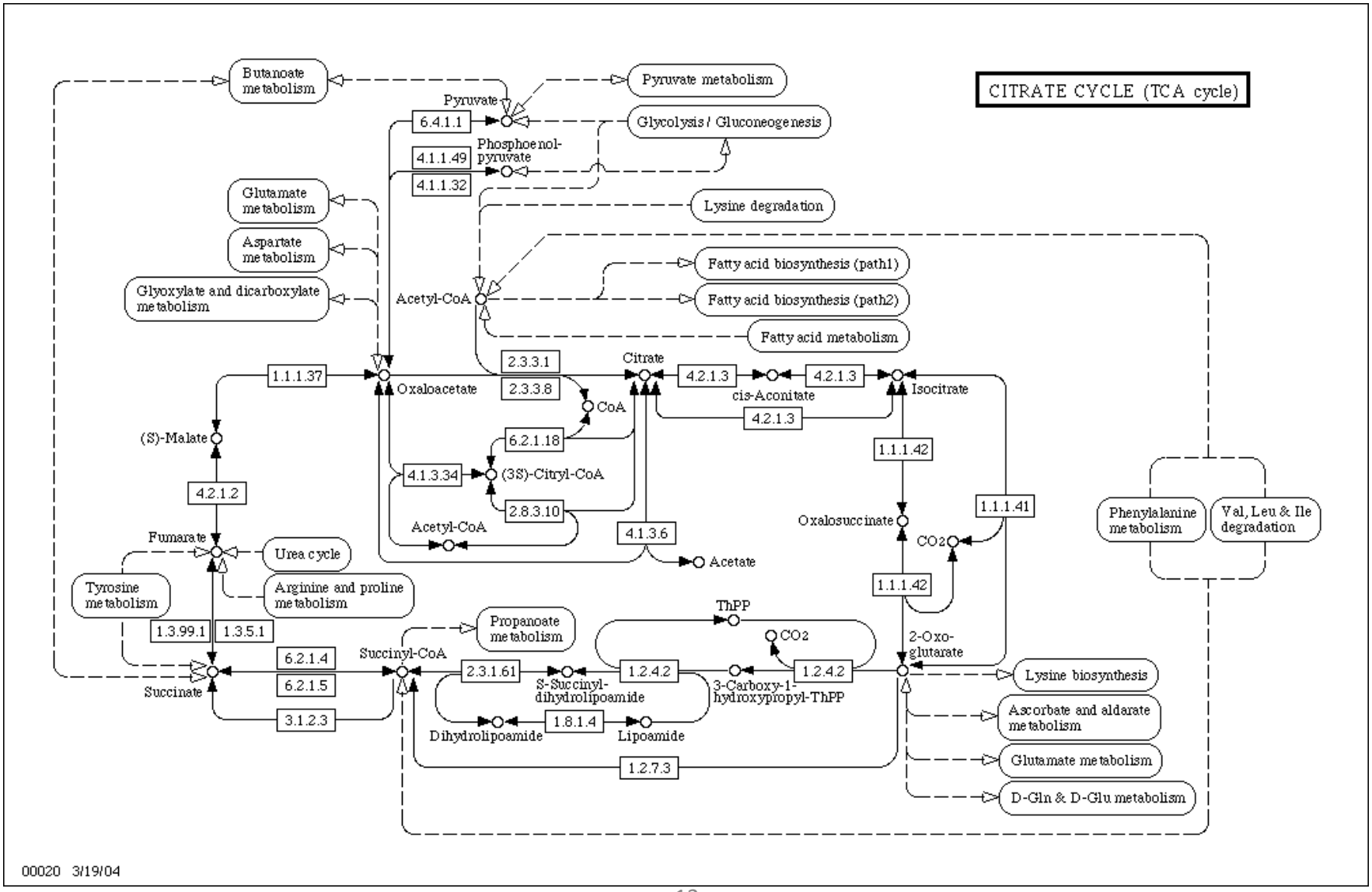
- Thousands of nodes (genes/proteins)
- 100000s of edges (interactions)

Physical Interaction Network

- Undirected graphs of proteins
- Edge exists if two proteins physically interact
  - Docking (Key – Keyhole)
- Interacting proteins tend to have the same biologic

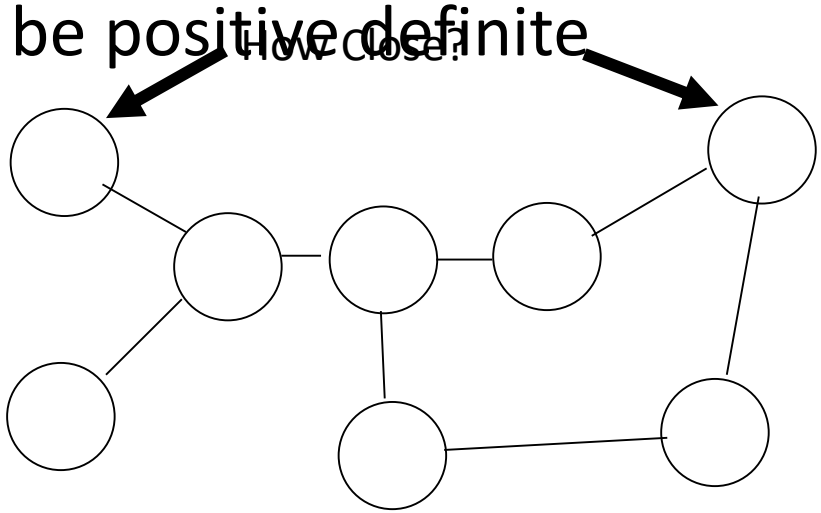


Metabolic Network



Diffusion kernels (Kondor and Lafferty, 2002)

- Function prediction by SVM using a network
  - Kernels are needed !
- Define closeness of two nodes
  - Has to be positive definite



Definition of Diffusion Kernel

- A: Adjacency matrix,
- D: Diagonal matrix of Degrees
- L = D-A: Graph Laplacian Matrix
- *Diffusion kernel matrix*

$$K = \exp(-\beta L)$$

  - β : Diffusion paramater

- Matrix exponential, not elementwise exponential

Computation of Matrix Exponential

- Definition

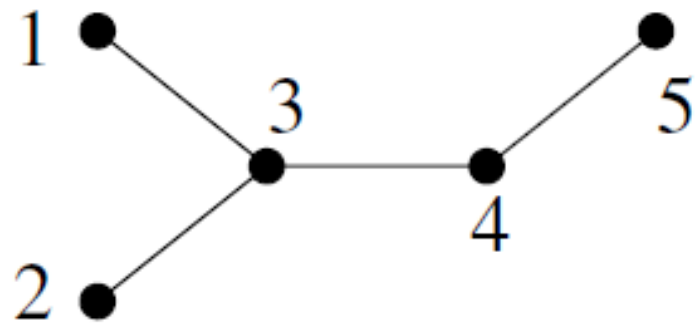
$$\exp(A) = \lim_{s \to \infty} (I + \frac{A}{s})^s$$

- Eigen-decomposition

$$A = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

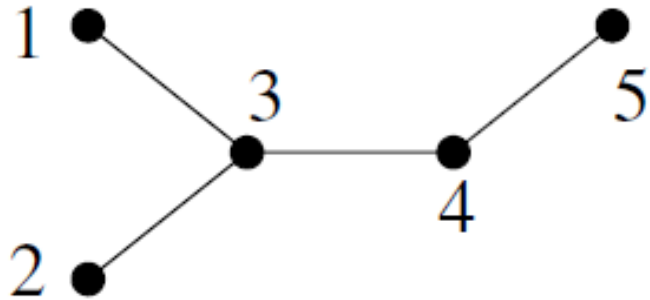
$$\exp(A) = \sum_{i=1}^{n} \exp(\lambda_i) \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

Adjacency Matrix and Degree Matrix



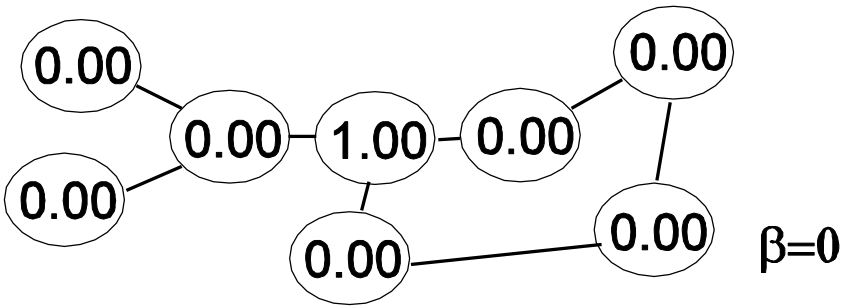
$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Graph Laplacian Matrix L

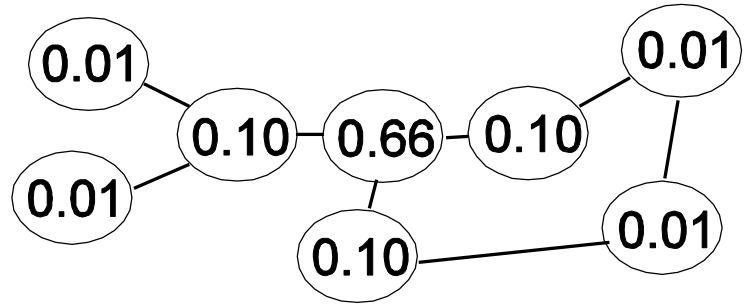


$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

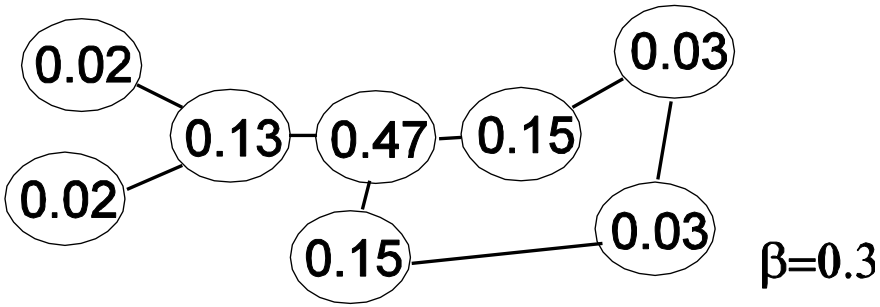
Actual Values of Diffusion Kernels



$\beta=0$



$\beta=0.15$



$\beta=0.3$

Closeness from the "central node"