Subgraph Matching via Optimal Transport

Theory, Algorithms, Applications

Wenxin Pan

2023 Master Thesis

Outline

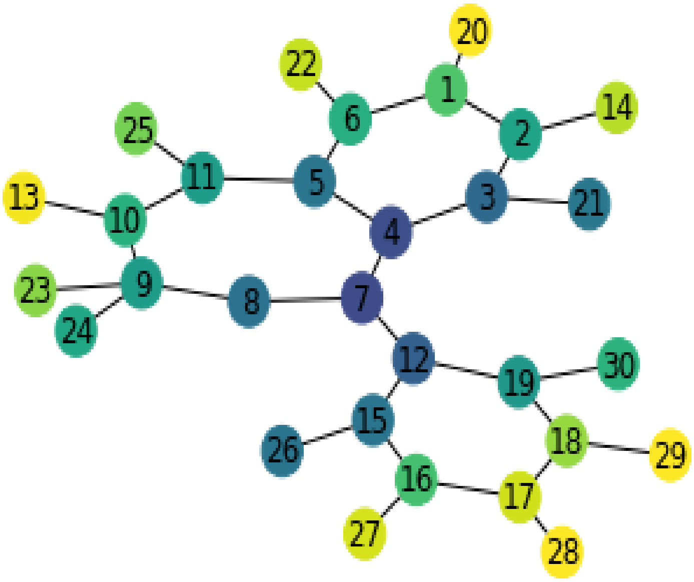
- **Introduction** – Problem formulation & motivation
- **Theory** – Optimal Transport on graphs
- **Algorithms** – Subgraph matching algorithms
- **Applications** – for biological datasets

Outline

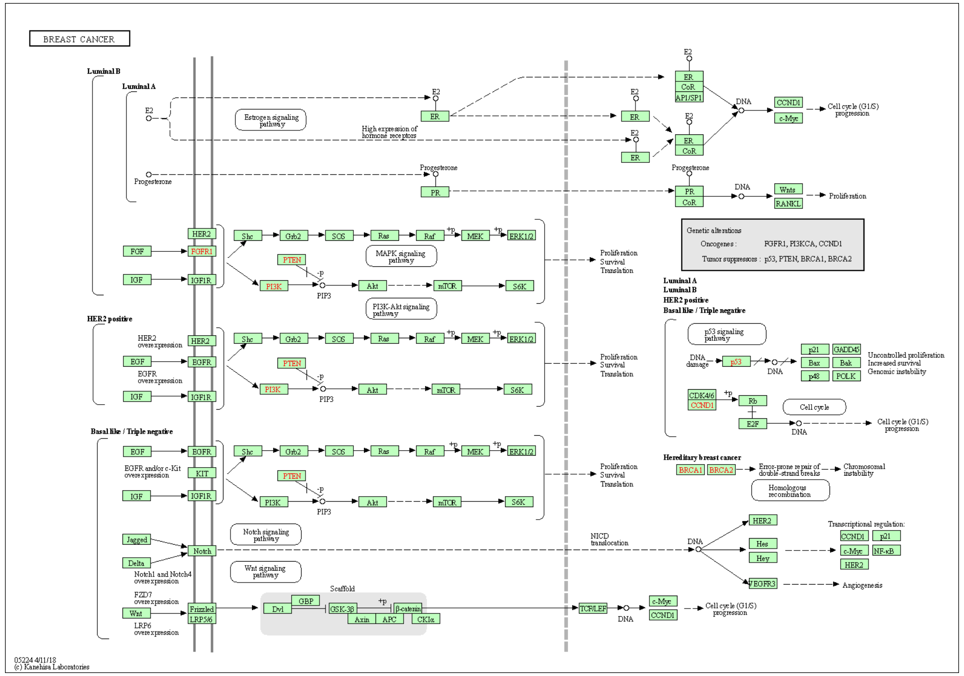
- **Introduction** – Problem formulation & motivation
- **Theory** – Optimal Transport on graphs
- **Algorithms** – Subgraph matching algorithms
- **Applications** – for biological datasets

Introduction

- Find predefined patterns within an object



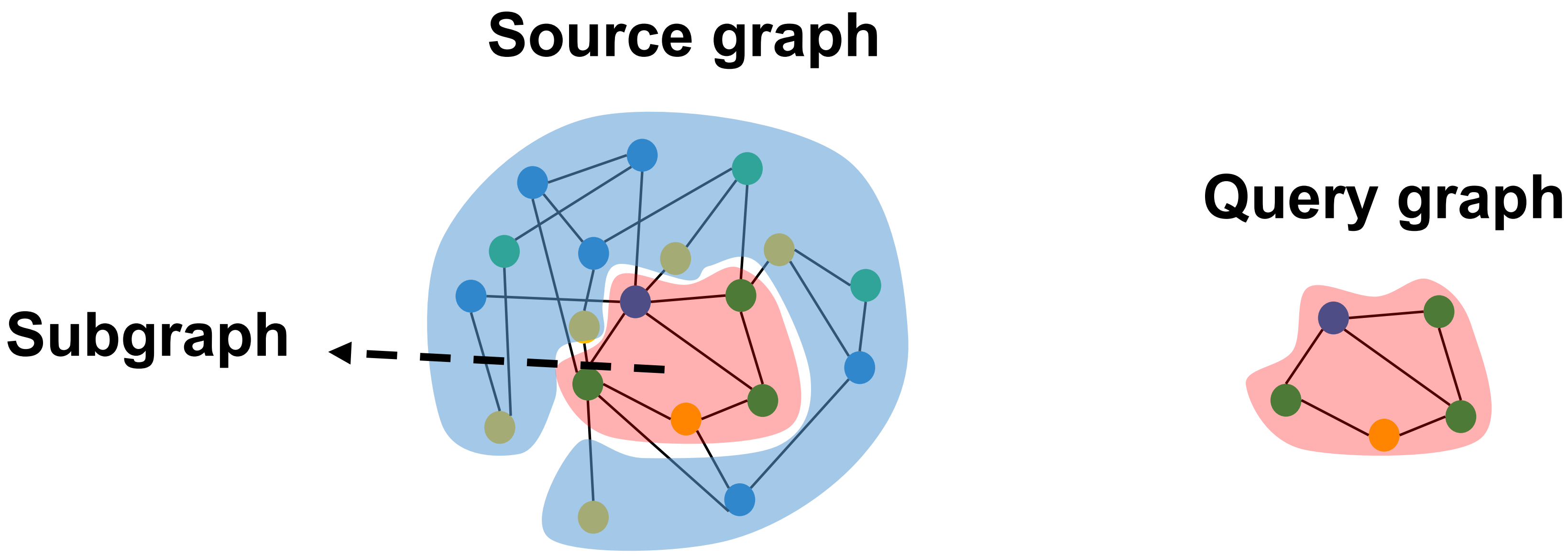
Find **rings** in a chemical compound



Find **signaling pathways** in a biomedical pathway

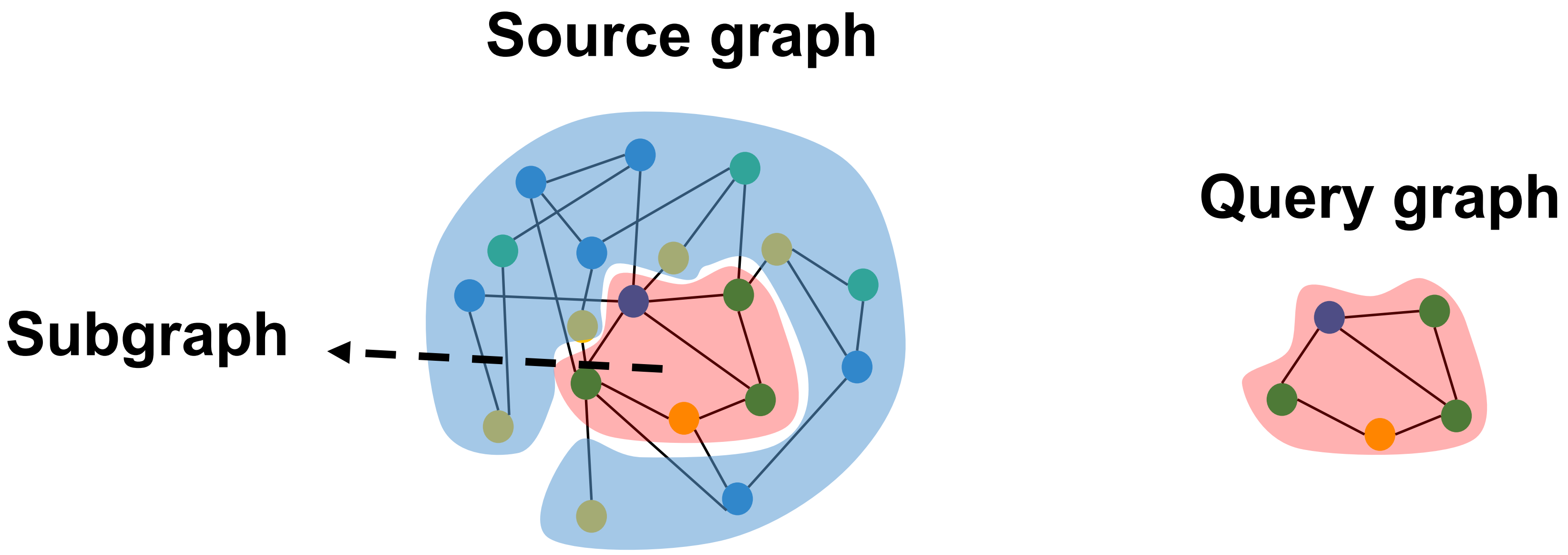
Introduction

- **Problem formulation**
  - Subgraph matching



Introduction

- **Goals**
  - Is there any *subgraph* same/similar to the *query graph* in the *source graph* ?
  - Where is the *subgraph* ?
  - How similar are the *subgraph* and *query graph* ?



Classic Formulation: Quadratic Assignment Problems (QAPs)

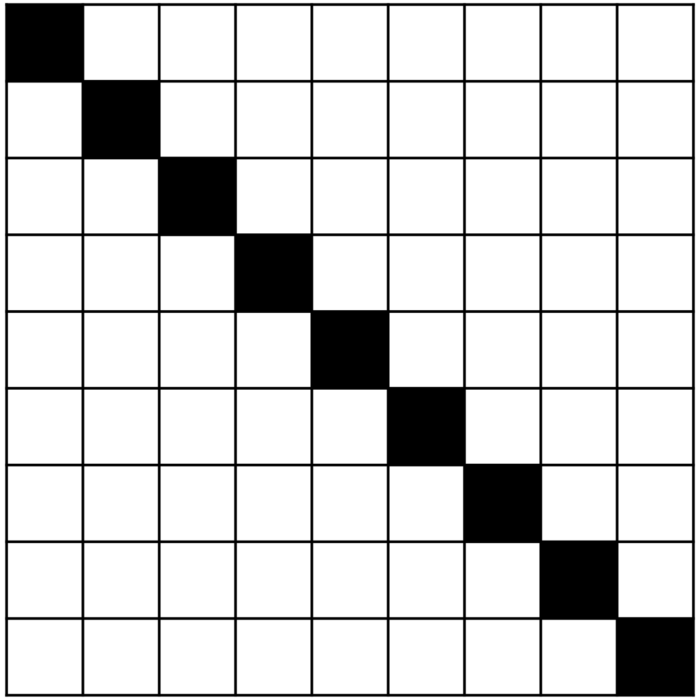
With adjacency matrices $\boldsymbol{A}^s$ and $\boldsymbol{A}^q$,
find a *binary* mapping matrix between source and query

- For two graphs of *equal size:*
  that each column and row has exactly one “1”

$$\boldsymbol{X} \in \{0,1\}^{n\times n}, \boldsymbol{X}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{X}^\top\mathbf{1}_n = \mathbf{1}_n$$

- For two graphs of *different sizes:*
  that each column (or row) has exactly one “1”

$$\boldsymbol{X} \in \{0,1\}^{n\times m}, \boldsymbol{X}\mathbf{1}_m \leq \mathbf{1}_n, \boldsymbol{X}^\top\mathbf{1}_n = \mathbf{1}_m$$



Classic Formulation: Quadratic Assignment Problems (QAPs)

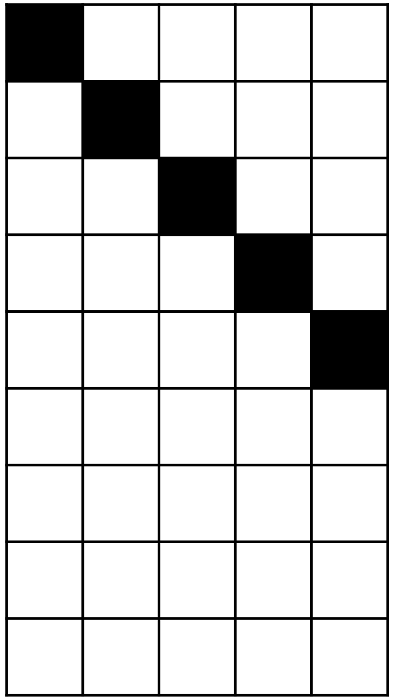
With adjacency matrices $\boldsymbol{A}^s$ and $\boldsymbol{A}^q$,
find a *binary* mapping matrix between source and query

- For two graphs of *equal size:*
  that each column and row has exactly one "1"

$$\boldsymbol{X} \in \{0,1\}^{n\times n}, \boldsymbol{X}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{X}^\top\mathbf{1}_n = \mathbf{1}_n$$

- For two graphs of *different sizes:*
  that each column (or row) has exactly one "1"

$$\boldsymbol{X} \in \{0,1\}^{n\times m}, \boldsymbol{X}\mathbf{1}_m \leq \mathbf{1}_n, \boldsymbol{X}^\top\mathbf{1}_n = \mathbf{1}_m$$



Koopmans-Beckmann's QAP

$$\min_{\boldsymbol{X} \in \mathcal{X}} \; -\langle \boldsymbol{A}^s, \boldsymbol{X} \boldsymbol{A}^q \boldsymbol{X}^\top \rangle_F$$

Lawler's QAP

$$\min_{\boldsymbol{X} \in \mathcal{X}} \; \mathrm{vec}(\boldsymbol{X})^\top \boldsymbol{Q} \, \mathrm{vec}(\boldsymbol{X}),$$

$$\boldsymbol{Q}_{n(j-1)+i, n(j'-1)+i'} = d^S(\boldsymbol{A}^s_{i,i'}, \boldsymbol{A}^q_{j,j'})$$

Remarks

The Quadratic Assignment Problems (QAPs)

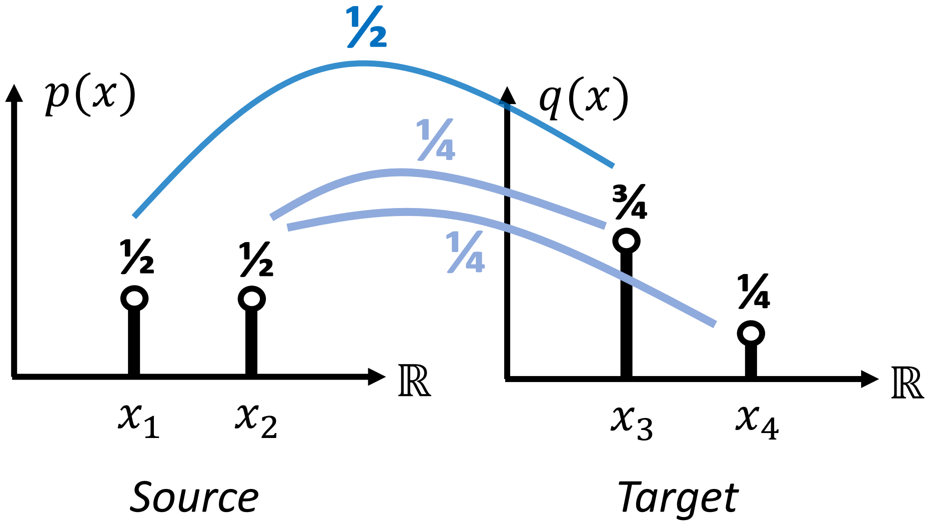
- are complex *combinatorial* optimization problems
- do not consider *node features*

Outline

- **Introduction** – Problem formulation & motivation
- **Theory** – Optimal Transport on graphs
- **Algorithms** – Subgraph matching algorithms
- **Applications** – for biological datasets

Optimal Transport

- **How to transport the *mass* from one probability distribution to another <u>with minimum cost</u> ?**



Cost Matrix $\boldsymbol{M}$:

$\lvert x_3 - x_1\rvert$	$\lvert x_4 - x_1\rvert$
$\lvert x_3 - x_2\rvert$	$\lvert x_4 - x_2\rvert$

A possible Transport Matrix $\boldsymbol{T}$:

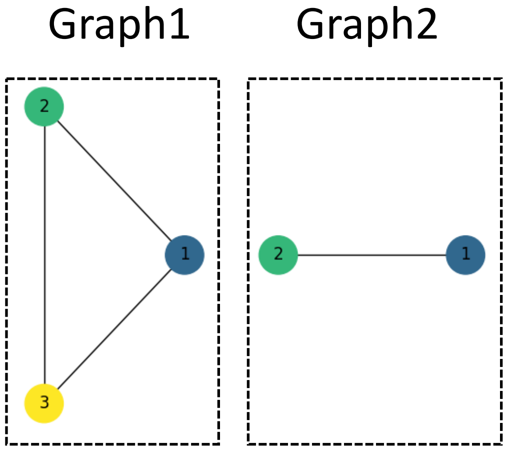
½	**0**
¼	**¼**

- **<u>The total minimum cost</u> is defined as the Wasserstein distance**

$$\mathcal{W}(p(x), q(x)) = \min_{\boldsymbol{T}} \left\{ \langle \boldsymbol{T}, \boldsymbol{M} \rangle_F \overset{\text{def.}}{=} \sum_{i,j} \boldsymbol{T}_{i,j} \boldsymbol{M}_{i,j} \right\}$$

Optimal Transport on Graphs

- **How to transport the *information* from one graph to another <u>with minimum cost</u> ?**
- *Information: Feature & Structure*

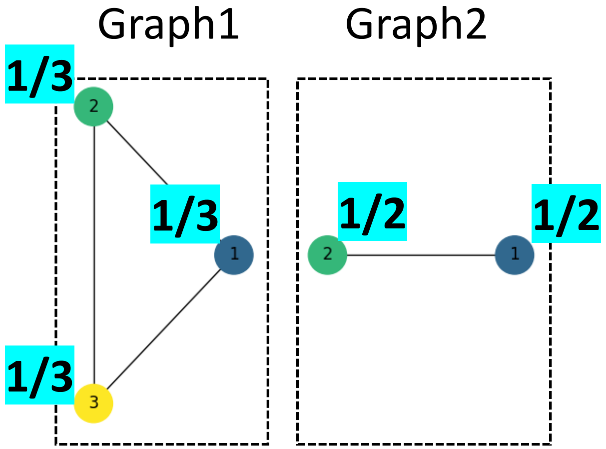


$$\boldsymbol{T} = \Big[\ ?\ \Big]_{3 \times 2}$$

Optimal Transport on Graphs

- **How to transport the *information* from one graph to another <u>with minimum cost</u> ?**
- *Information: Feature & Structure*

- ***Feature***



$$M = \begin{array}{c|cc} & \text{g} & \text{b} \\ \text{g} & 0 & 1 \\ \text{b} & 1 & 0 \\ \text{y} & 1 & 1 \end{array}$$

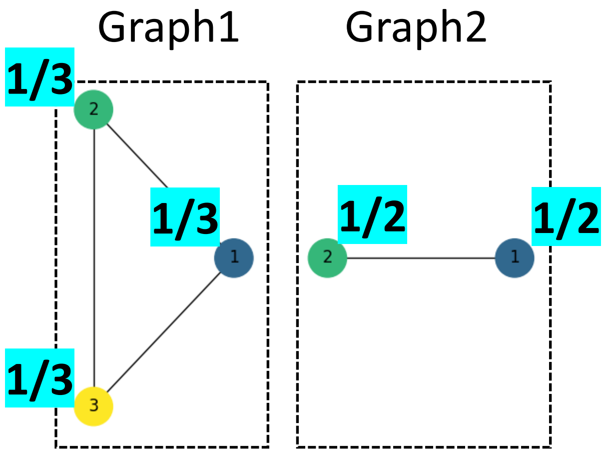
$$\boldsymbol{T}^* = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \\ 1/6 & 1/6 \end{bmatrix}$$

Wasserstein distance: $\mathcal{W}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{M}) = \min_{\boldsymbol{T}} \langle \boldsymbol{T}, \boldsymbol{M} \rangle_F$

Optimal Transport on Graphs

- **How to transport the *information* from one graph to another <u>with minimum cost</u> ?**
- *Information: Feature & Structure*

- ***Structure***



$$C^s = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \qquad C^t = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad L_{i,i',j,j'} = \mathcal{L}\left(C^s_{i,i'}, C^t_{j,j'}\right)$$

6 optimal transport matrices: $T^* = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \\ 1/6 & 1/6 \end{bmatrix}$ $T^* = \begin{bmatrix} 0 & 1/3 \\ 1/3 & 0 \\ 1/6 & 1/6 \end{bmatrix}$ ……

Gromov-Wasserstein distance: $\mathcal{GW}(\boldsymbol{p}, \boldsymbol{q}, C^s, C^t) = \min_{\boldsymbol{T}} \sum_{i,i',j,j'} \mathcal{L}\left(C^s_{i,i'}, C^t_{j,j'}\right) \boldsymbol{T}_{i,j} \boldsymbol{T}_{i',j'}$

Wasserstein Distance

$$\mathcal{W}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{M}) = \min_{\boldsymbol{T}} \left\{ \langle \boldsymbol{T},\boldsymbol{M}\rangle_F \overset{\text{def.}}{=} \sum_{i,j} \boldsymbol{T}_{i,j}\boldsymbol{M}_{i,j} \right\}$$

Gromov-Wasserstein Distance

$$\mathcal{GW}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{C}^s,\boldsymbol{C}^t) = \min_{\boldsymbol{T}} \left\{ \langle \boldsymbol{L}\otimes\boldsymbol{T},\boldsymbol{T}\rangle_F \overset{\text{def.}}{=} \sum_{i,i',j,j'} \mathcal{L}\left(\boldsymbol{C}^s_{i,i'},\boldsymbol{C}^t_{j,j'}\right)\boldsymbol{T}_{i,j}\boldsymbol{T}_{i',j'} \right\}$$

with $(\boldsymbol{L}\otimes\boldsymbol{T})_{i,j} \overset{\text{def.}}{=} \sum_{i',j'} \boldsymbol{L}_{i,i',j,j'}\boldsymbol{T}_{i',j'}$

Fused Gromov-Wasserstein Distance

$$\mathcal{FGW}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{M},\boldsymbol{C}^s,\boldsymbol{C}^t) = \min_{\boldsymbol{T}} (1-\alpha)\langle \boldsymbol{T},\boldsymbol{M}\rangle_F + \alpha\langle \boldsymbol{L}\otimes\boldsymbol{T},\boldsymbol{T}\rangle_F$$

FGW distance

Fused Gromov-Wasserstein Distance

$$\mathcal{FGW}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{M},\boldsymbol{C}^s,\boldsymbol{C}^t) = \min_{\boldsymbol{T}} (1-\alpha)\langle \boldsymbol{T},\boldsymbol{M}\rangle_F + \alpha\langle \boldsymbol{L}\otimes\boldsymbol{T},\boldsymbol{T}\rangle_F$$

with $(\boldsymbol{L}\otimes\boldsymbol{T})_{i,j} \overset{\text{def.}}{=} \sum_{i',j'} \boldsymbol{L}_{i,i',j,j'}\boldsymbol{T}_{i',j'}$

Constraints $\text{s.t.}\ \boldsymbol{T}\mathbf{1}_m = \boldsymbol{p}; \boldsymbol{T}^\top\mathbf{1}_n = \boldsymbol{q}; \boldsymbol{T}\in\mathbb{R}_+^{n\times m}$

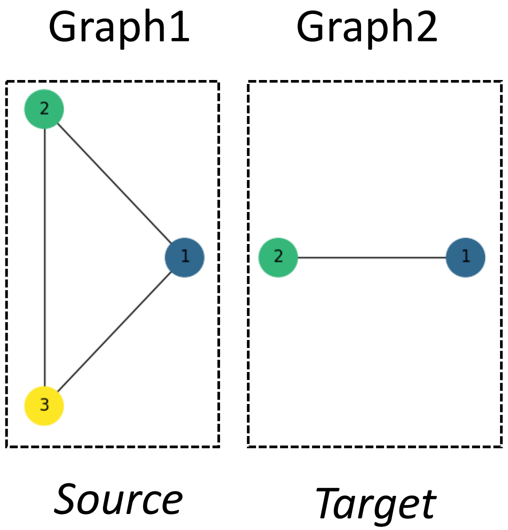
- Probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ indicate *weights/importance* on nodes

Remarks

The FGW distance

- defines a *distance/metric* between two labeled graphs
- is a *generalization* of classic Quadratic Assignment Problems (QAPs)
- is naturally a *continuous* problem
- naturally allows graphs of *different sizes*

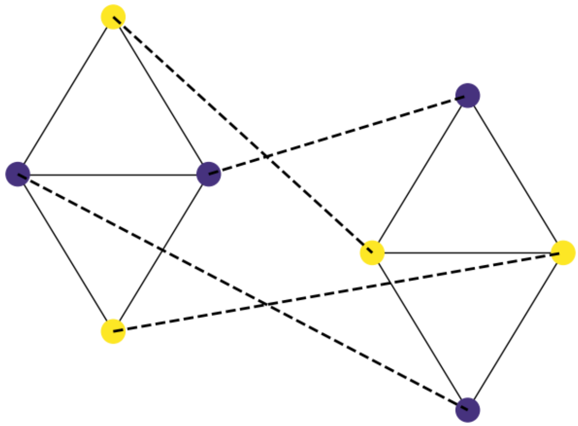
Examples



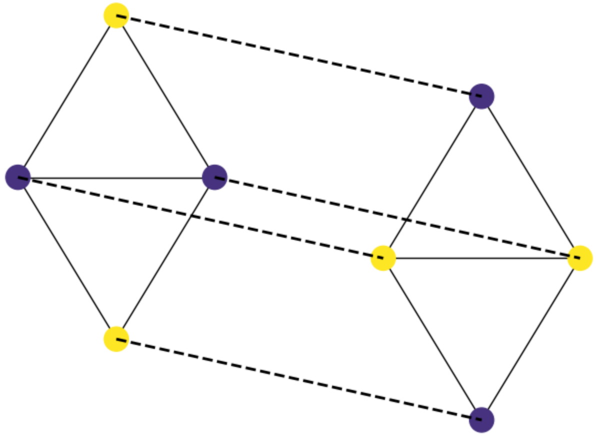
$$\boldsymbol{T}^* = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \\ 1/6 & 1/6 \end{bmatrix}$$

$$\mathcal{FGW}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{M}, \boldsymbol{C}^s, \boldsymbol{C}^t) = 0.306$$

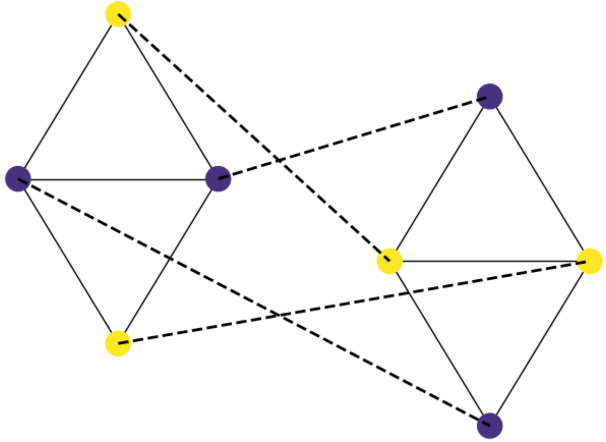
Examples



Wasserstein distance = 0



GW distance = 0

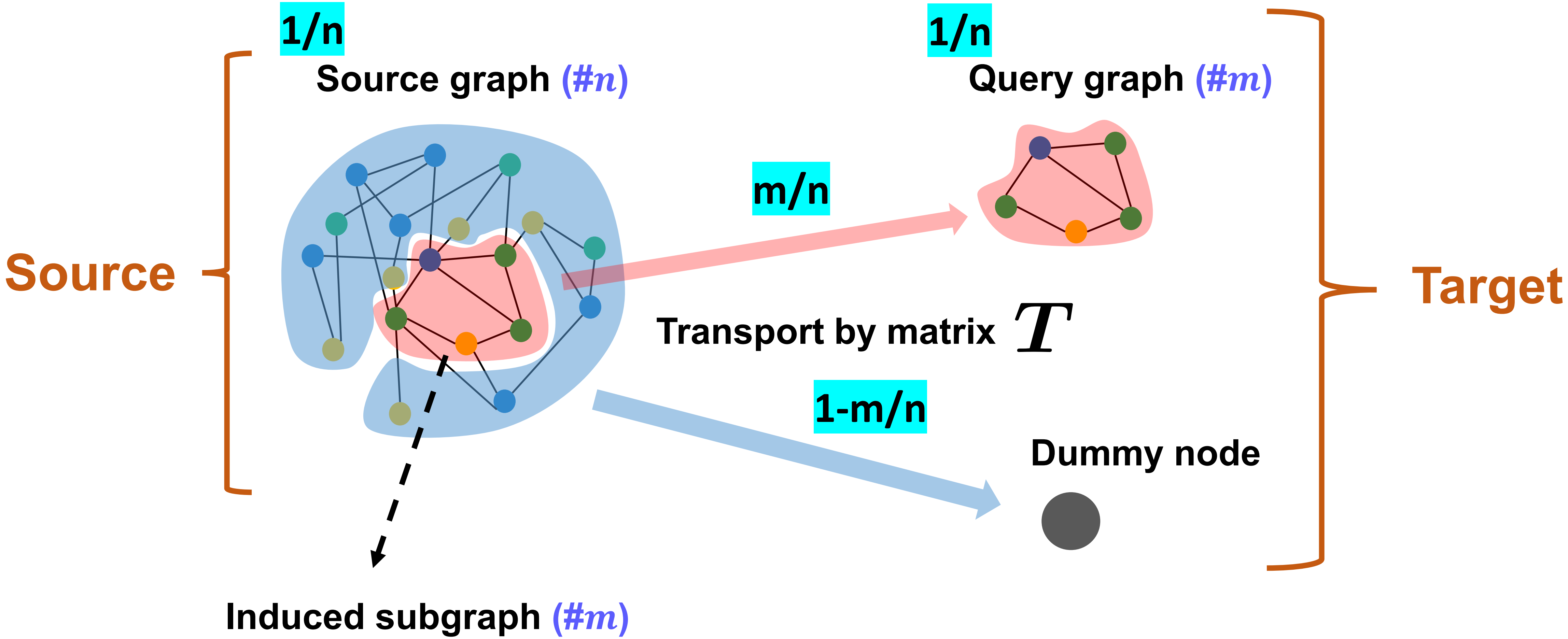


FGW distance = $0.125\ \ (\alpha = 0.5)$

Outline

- **Introduction** – Problem formulation & motivation
- **Theory** – Optimal Transport on graphs
- **Algorithms** – Subgraph matching algorithms
- **Applications** – for biological datasets

Subgraph Matching – Framework 1



Source

1/n

Source graph $(\#n)$

Induced subgraph $(\#m)$

m/n

Transport by matrix T

1-m/n

Target

1/n

Query graph $(\#m)$

Dummy node

Subgraph Matching – Framework 1

$$\mathcal{FGW}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{M}, \boldsymbol{C}^s, \boldsymbol{C}^t) = \min_{\boldsymbol{T}} (1-\alpha)\langle \boldsymbol{T}, \boldsymbol{M}\rangle_F + \alpha\langle \boldsymbol{L} \otimes \boldsymbol{T}, \boldsymbol{T}\rangle_F$$

- mass of *source graph* are transported to
    - *query graph* with **original costs**
    - *dummy node* **"for free”**

$$\boldsymbol{M}_{i,j} = \begin{cases} d(a_i, b_j), & \text{for} \quad j = 1, \dots, m \\ 0, & \text{for} \quad j = m+1. \end{cases}$$

$$\boldsymbol{L}_{i,i',j,j'} = \begin{cases} \mathcal{L}\left(\boldsymbol{C}^s_{i,i'}, \boldsymbol{C}^t_{j,j'}\right), & \text{for} \quad j = 1, \dots, m \quad \text{and} \quad j' = 1, \dots, m \\ 0, & \text{for} \quad j = m+1 \quad \text{or} \quad j' = m+1 \end{cases}$$

- Minimum value of FGW distance is still 0.

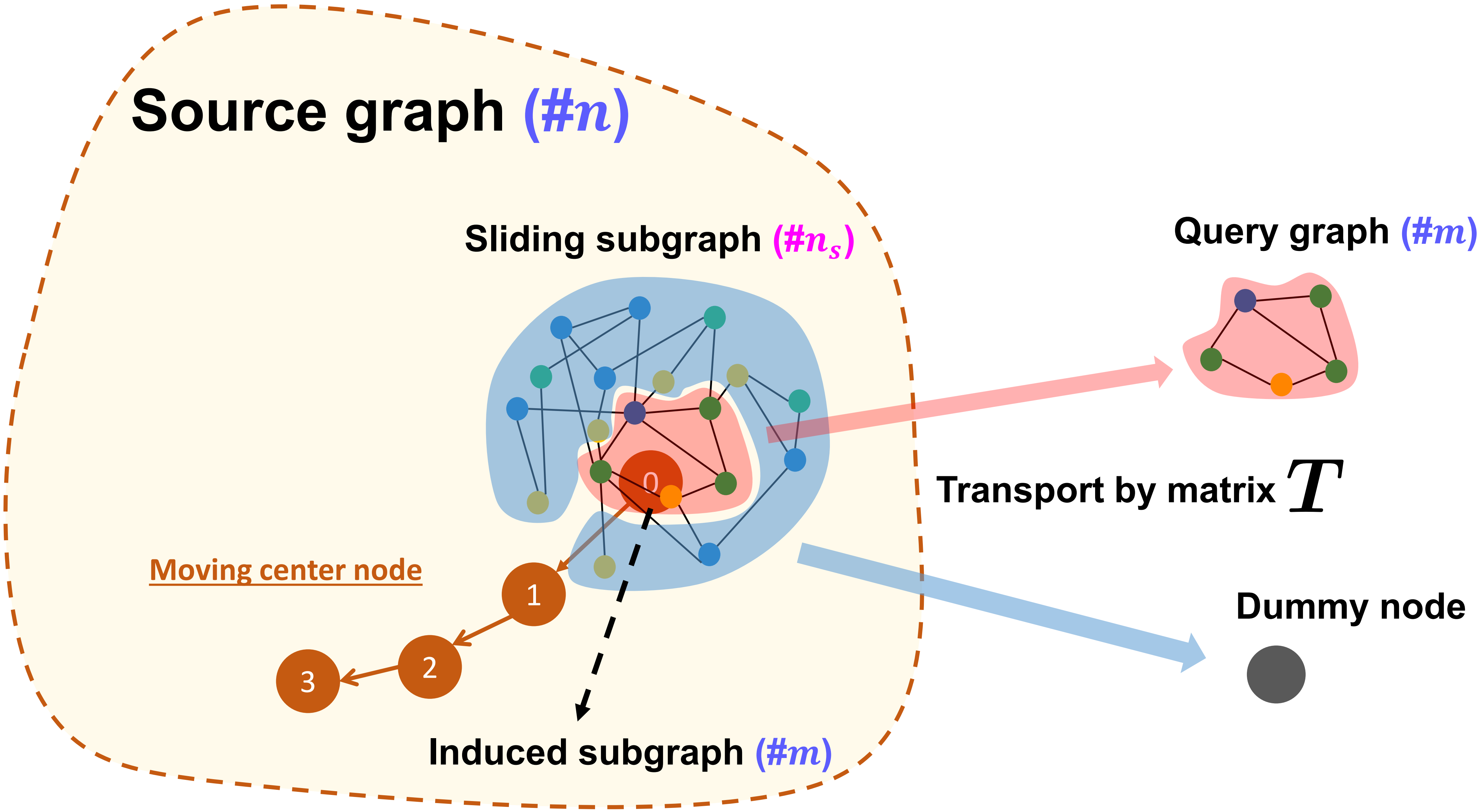
Issues raised in large graphs

$$\mathcal{FGW}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{M},\boldsymbol{C}^s,\boldsymbol{C}^t) = \min_{\boldsymbol{T}} (1-\alpha)\langle \boldsymbol{T},\boldsymbol{M}\rangle_F + \alpha\langle \boldsymbol{L}\otimes\boldsymbol{T},\boldsymbol{T}\rangle_F$$

- **Large Computation** of the tensor-matrix product $\boldsymbol{L}\otimes\boldsymbol{T}$
- **Non-convexity** due to the product $\boldsymbol{T}_{i,j}\boldsymbol{T}_{i',j'}$
  - larger graphs create more local minima

Potential solution: “shrink” the graph that is directly used in the FGW distance

Subgraph Matching – Framework 2



Source graph (#n)

Sliding subgraph (#n_s)

Query graph (#m)

Transport by matrix T

Dummy node

Moving center node

Induced subgraph (#m)

Subgraph Matching – Framework 2

```
For node in Source graph:
    Create a sliding subgraph
    Framework 1
End
```

Outline

- **Introduction** – Problem formulation & motivation
- **Theory** – Optimal Transport on graphs
- **Algorithms** – Subgraph matching algorithms
- **Applications** – for biological datasets

Application - example

Frequent subgraph matching

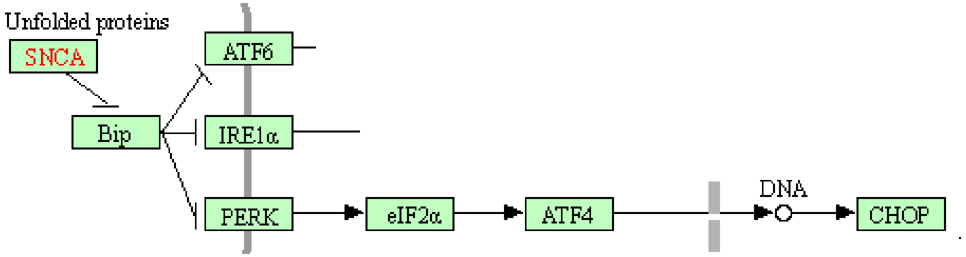
- Detect the similarities within a set of graphs

Application - example

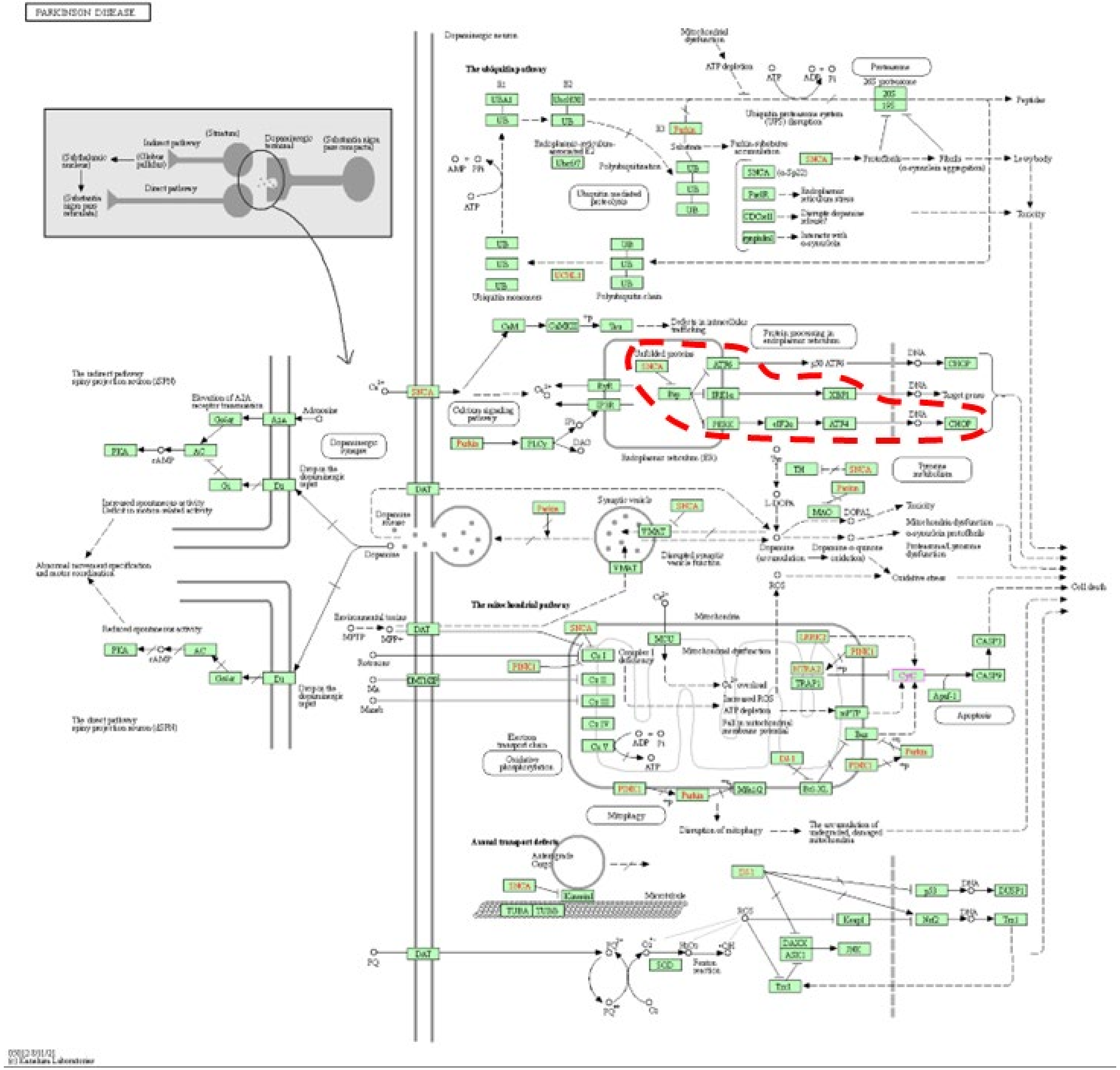
Detect signaling pathways within Neurodegenerative disease pathways

- Parkinson
- Alzheimer

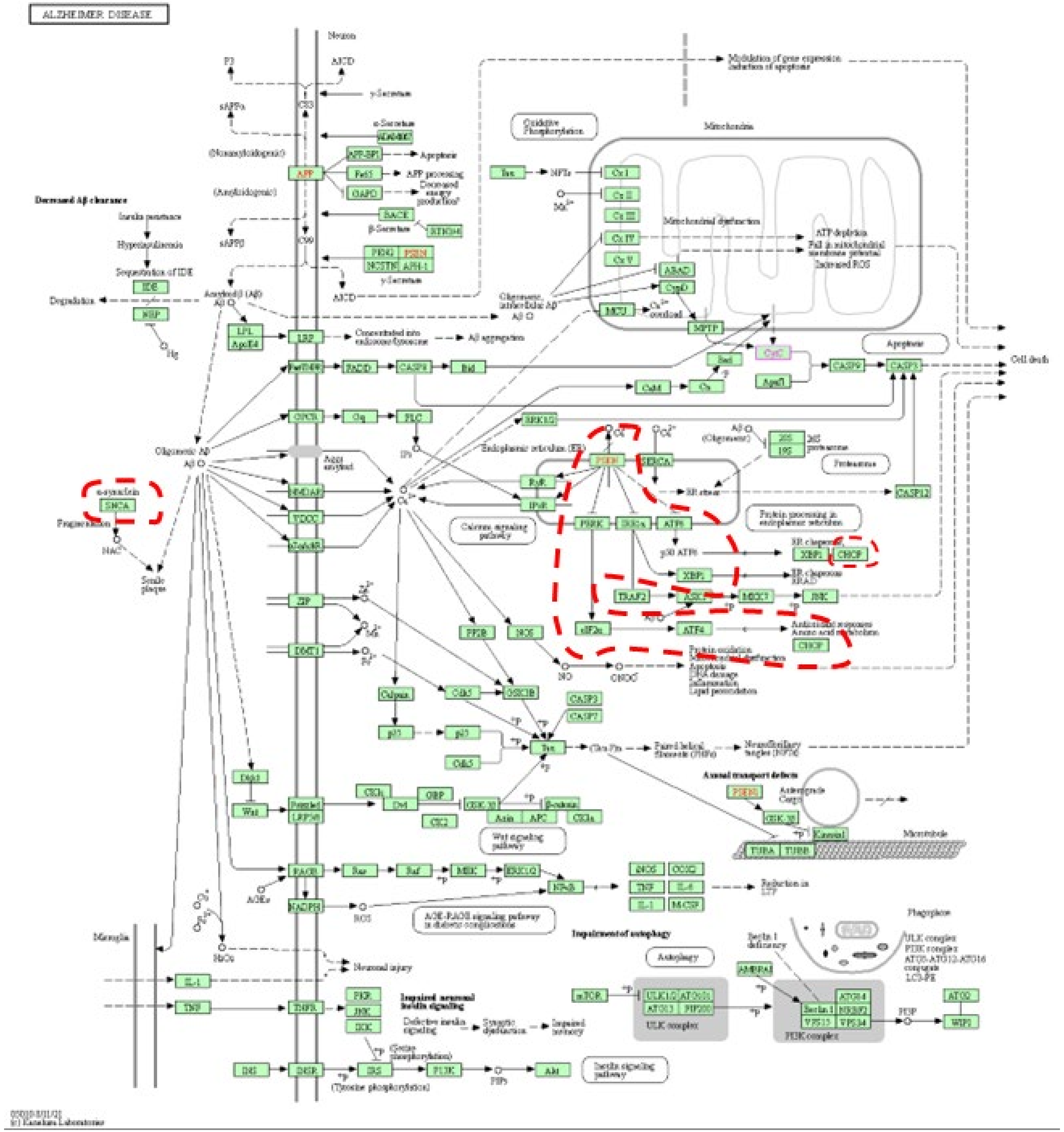
Query graph: unfolded protein response (UPR)



Parkinson


PARKINSON DISEASE
Dopaminergic neuron
The ubiquitin pathway
UBA1
UB
Parkin
SNCA
UCHL1
Ubiquitin monomers
Polyubiquitin chain
Proteasome
26S proteasome
20S
19S
Peptides
Protofibrils
Fibrils
Lewy body
Toxicity
CaM
Unfolded proteins
ATF6
CHOP
XBP1
PERK
eIF2a
ATF4
Bip
IRE1a
DNA
Target genes
Endoplasmic reticulum (ER)
TH
L-DOPA
MAO
DOPAL
Dopamine
DAT
VMAT
Synaptic vesicle
Calcium signaling pathway
Dopaminergic synapse
The indirect pathway spiny projection neuron (iSPN)
The direct pathway spiny projection neuron (dSPN)
PKA
AC
cAMP
Adenosine
A2A
D1
D2
The mitochondrial pathway
Mitochondria
MPTP
MPP+
Rotenone
Cx I
Cx II
Cx III
Cx IV
Cx V
ADP
Pi
ATP
CytC
CASP3
CASP9
Apaf-1
Apoptosis
PINK1
Bax
Mitophagy
Oxidative stress
Cell death
Axonal transport defects
Kinesin
TUBA
TUBB
Microtubule
ROS
SOD
Fenton reaction
DJ-1
p53
Nrf2
Keap1
DAXX
ASK1
JNK
Trx1
PQ
05012 8/01/21
(c) Kanehisa Laboratories

Alzheimer


ALZHEIMER DISEASE
Neuron
Mitochondria
Decreased Aβ clearance
Oligomeric Aβ
α-synuclein
SNCA
Senile plaque
Endoplasmic reticulum (ER)
PSEN
SERCA
ER stress
XBP1
CHOP
ATF4
eIF2α
PERK
IRE1α
ATF6
TRAF2
ASK1
MKK7
JNK
Apoptosis
Cell death
CASP9
CASP3
CytC
Apaf1
FADD
CASP8
Bid
Tau
Neurofibrillary tangles (NFTs)
Axonal transport defects
Microtubule
Impairment of autophagy
Autophagy
ULK complex
PI3K complex
Phagophore
Neuronal injury
Microglia
IL-1
TNF
TNFR
INS
INSR
IRS
PI3K
Akt
Impaired neuronal insulin signaling
Insulin signaling pathway
Wnt signaling pathway
Calcium signaling pathway
AGE-RAGE signaling pathway in diabetic complications
05010 8/11/21
(c) Kanehisa Laboratories

Conclusion

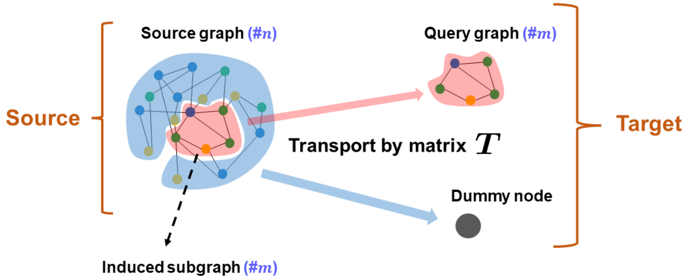
Theory

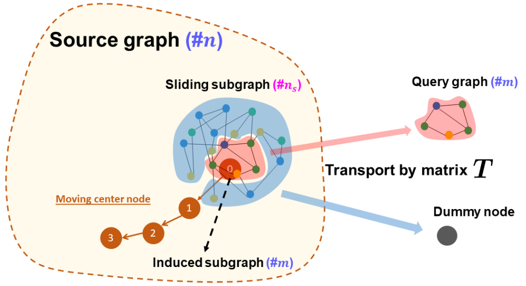
Algorithms

Applications

Fused Gromov-Wasserstein Distance

$$\mathcal{FGW}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{M}, \boldsymbol{C}^s, \boldsymbol{C}^t) = \min_{\boldsymbol{T}} \ (1-\alpha)\langle \boldsymbol{T}, \boldsymbol{M} \rangle_F + \alpha \langle \boldsymbol{L} \otimes \boldsymbol{T}, \boldsymbol{T} \rangle_F$$





Frequent subgraph matching

Thank you!