
2022 IEEE International Conference on Big Data (IEEE BigData 2022)

December 17-20, 2022 @ Osaka, Japan

# DPPIN: A Biological Repository of Dynamic Protein-Protein Interaction Network Data

Dongqi Fu
(UIUC)

Jingrui He
(UIUC)

Presenter: Dongqi Fu
Email: dongqif2@illinois.edu

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# Contents

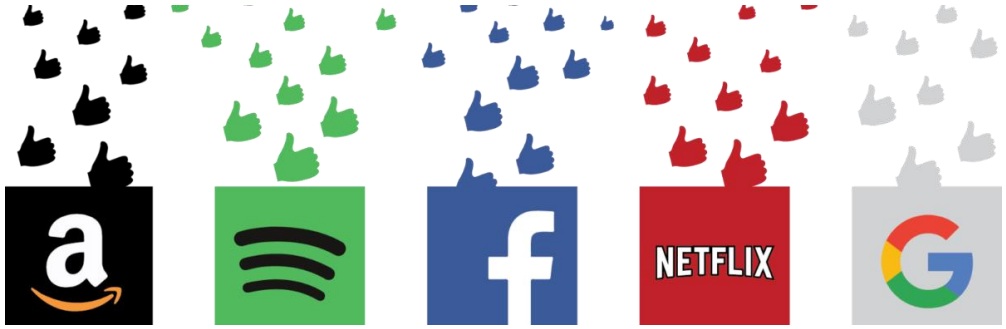
- **Motivation and Overview of DPPIN**
- **Graph Generation Process in DPPIN**
- **Resource Statistics in DPPIN**
- **Experiment**
- **Conclusion**

# Motivation

- Nowadays, graph-based research has served for a wide range of real-world applications, such like

Information Retrieval

Fraud Detection




Recommender Systems

- To some extent, dynamic graph representation learning and mining methods are more suitable for real-world scenarios for modeling the evolving graph structures and attributes.

# Motivation

- Comparing with the publicly available static network data, the volume of dynamic data is not that sufficient.

- Therefore, in this paper, we provide a dynamic network repository, DPPIN, which consists 12 different dynamic network datasets from the biological domain, and they are

  - Label-adequate (i.e., high label rate of nodes)

  - Dynamics-meaningful (i.e., metabolic patterns of yeast cells)

  - Attribute-sufficient (i.e., accessible node and edge features)

# Overview

- Online Repository: https://github.com/DongqiFu/DPPIN
  - Datasets:
    - 12 generated dynamic network datasets, each of which contains
      - Adjacency matrix
      - Node feature, edge weight, node label
  - Code (Python):
    - Generating dynamic networks based on user interests (i.e., specific hyperparameters)
  - Necessary material:
    - Gene expression series

# Contents

- Motivation and Overview of DPPIN
- Graph Generation Process in DPPIN
- Resource Statistics in DPPIN
- Experiment
- Conclusion

# Before Generation

- Collecting static protein interaction data and gene expression
  - Totally, we involve 12 static protein networks from [1] for generating the dynamic ones.
  - The subgraph from 1/12 is illustrated as below.

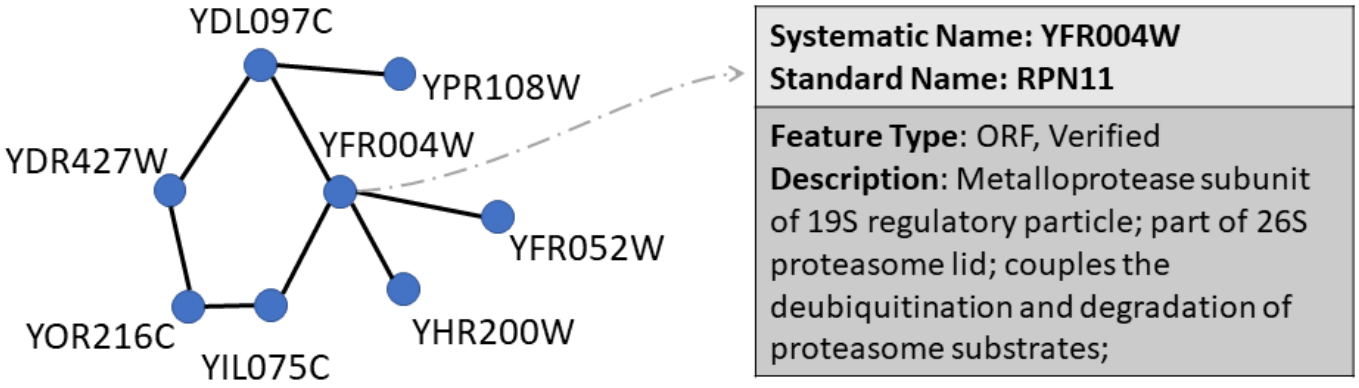



Fig. A Subgraph Extracted from the Static Protein-Protein Interaction Network of Yeast Cells [2]. Each node stands for a gene coding protein, and the description of each protein node can be extracted from the Saccharomyces Genome Database [3].

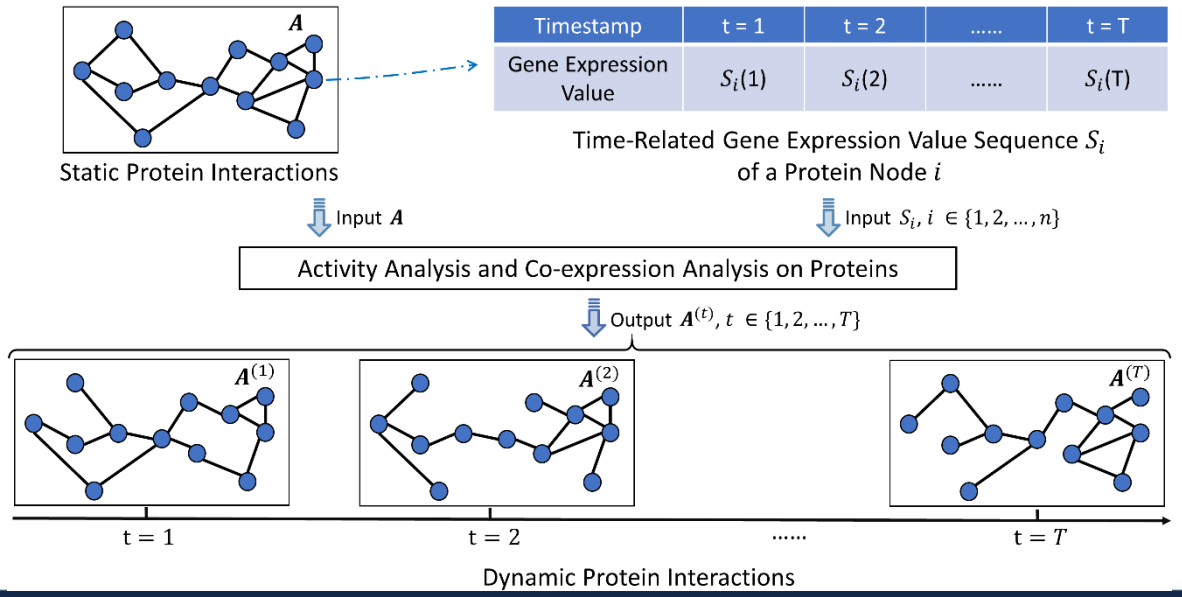
  - Gene expression series: GSE3431 [4]


[1] YeastNet: https://www.inetbio.org/yeastnet/downloadnetwork.php
[2] Babu et al., Interaction landscape of membrane-protein complexes in saccharomyces cerevisiae. Nature, 2012
[3] Saccharomyces Genome Database: https://www.yeastgenome.org/
[4] Tu et al., Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science, 2005

# Generation Process

- The generation process is three-folded, adopted from [5]
  - (1) Determine active proteins $\boldsymbol{A}_{act}^{(t)}$
  - (2) Determine co-expressed protein pairs $\boldsymbol{A}_{coe}^{(t)}$
  - (3) Determine active and co-expressed protein interactions

$$\boldsymbol{A}^{(t)} = \boldsymbol{A}_{act}^{(t)} \odot \boldsymbol{A}_{coe}^{(t)} \odot \boldsymbol{A}$$ (weighted adjacency matrix at each timestamp)




[5] Zhang et al., A method for predicting protein complex in dynamic ppi networks. BMC Bioinformatics, 2016.

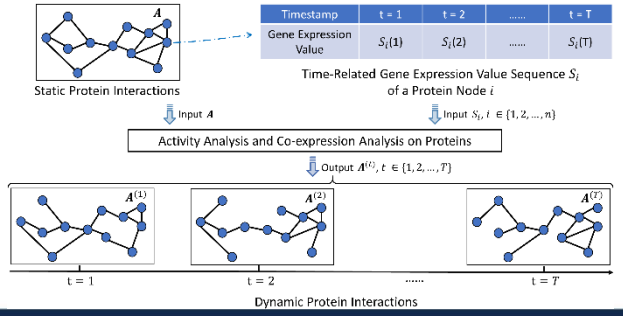
# Generation Process

- The generation process is three-folded, adopted from [5]
  - (1) Determine active proteins
  - (2) Determine co-expressed protein pairs
  - (3) Determine active and co-expressed protein interactions

- The detailed equations and the pseudo algorithm are illustrated in the paper.

- The generation program is coded by Python and publicly available.

- Node labels are retrieved from [3].





[3] Saccharomyces Genome Database: https://www.yeastgenome.org/

[5] Zhang et al., A method for predicting protein complex in dynamic ppi networks. BMC Bioinformatics, 2016.

# Contents

- Motivation and Overview of DPPIN
- Graph Generation Process in DPPIN
- Resource Statistics in DPPIN
- Experiment
- Conclusion

# Datasets Statistics

- Statistics of each generated dynamic network dataset in DPPIN

| Generated Dynamic PPINs | #Nodes | #Edges | Node Features | Edge Features | Node Label Rate | #Timestamps |
|---|---|---|---|---|---|---|
| DPPIN-Uetz | 922 | 2,159 | ✓ | ✓ | 921/922 (99.89%) | 36 |
| DPPIN-Ito | 2,856 | 8,638 | ✓ | ✓ | 2854/2856 (99.93%) | 36 |
| DPPIN-Ho | 1,548 | 42,220 | ✓ | ✓ | 1547/1548 (99.93%) | 36 |
| DPPIN-Gavin | 2,541 | 140,040 | ✓ | ✓ | 2538/2541 (99.88%) | 36 |
| DPPIN-Krogan (LCMS) | 2,211 | 85,133 | ✓ | ✓ | 2208/2211 (99.86%) | 36 |
| DPPIN-Krogan (MALDI) | 2,099 | 78,297 | ✓ | ✓ | 2097/2099 (99.90%) | 36 |
| DPPIN-Yu | 1,163 | 3,602 | ✓ | ✓ | 1160/1163 (99.74%) | 36 |
| DPPIN-Breitkreutz | 869 | 39,250 | ✓ | ✓ | 869/869 (100.00%) | 36 |
| DPPIN-Babu | 5,003 | 111,466 | ✓ | ✓ | 4997/5003 (99.88%) | 36 |
| DPPIN-Lambert | 697 | 6,654 | ✓ | ✓ | 697/697 (100.00%) | 36 |
| DPPIN-Tarassov | 1,053 | 4,826 | ✓ | ✓ | 1051/1053 (99.81%) | 36 |
| DPPIN-Hazbun | 143 | 1,959 | ✓ | ✓ | 143/143 (100.00%) | 36 |

# Contents

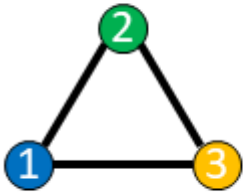
- Motivation and Overview of DPPIN
- Graph Generation Process in DPPIN
- Resource Statistics in DPPIN
- Experiment
- Conclusion

# Experiment

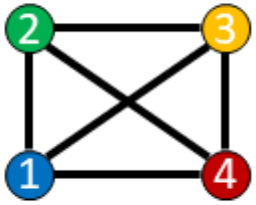
- Problem Definition: Dynamic Spectral Clustering
  - Input: (1) a dynamic graph $\tilde{G} = \{G^{(1)}, G^{(2)}, \ldots, G^{(T)}\}$, and (2) the desired number of disjoint clusters $q$
  - Output: (2) $q$ disjoint clusters $\{C_1^{(t)}, C_2^{(t)}, \ldots, C_q^{(t)}\}$ minimizing the normalized cut, and $G^{(t)} = \sum_{i=1}^{q} C_i^{(t)}$ at each timestamp $t \in \{1, 2, \ldots, T\}$.

# Experiment

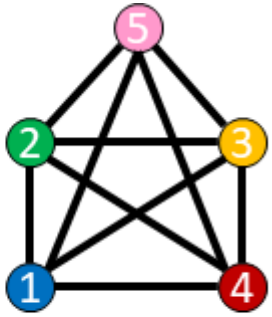
- Problem Definition: Dynamic Spectral Clustering (**Clique-Preserved**)
  - Intuition: Clique patterns are important in indicating protein properties [6], how could we preserve clique patterns during clustering on protein networks?




3-clique




4-clique




5-clique

[6] Adamcsek et al., CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics, 2006.

# Experiment

- To be best of our knowledge, there is no existing work for dynamic clique-preserving spectral clustering.

- In the paper, we contribute a simple trial by tracking eigen-decomposition to realize that goal.

(1) Motif is set to be 3-clique;
(2) Motif conductance is the lower the better, indicating more cliques are preserved during the graph partitioning.

MSC: Motif-aware spectral clustering method [7]

| Methods | DPPIN-Ho | |
|---|---|---|
| | Motif Conductance | Time Consumption |
| MSC | 0.0000 | 0.7354s |
| MSC+T | 0.7500 | 0.0469s |

+T: Eigen-decomposition tracking method [8]

[7] Benson et al., Higher-order organization of complex networks. Science, 2016.
[8] Chen Chen and Hanghang Tong. Fast eigen-functions tracking on dynamic graphs. In SDM, 2015

# Potential Application

- Research opportunity:
  - MSC achieves the better clustering compactness but consumes a larger amount of time.
  - Although MSC+T outputs the solution in a fast manner, the compactness of clusters is not always ideal.

(1) Motif is set to be 3-clique;
(2) Motif conductance is the lower the better, indicating more cliques are preserved during the graph partitioning.

MSC: Motif-aware spectral clustering method [7]

| Methods | DPPIN-Ho | |
|---|---|---|
| | Motif Conductance | Time Consumption |
| MSC | 0.0000 | 0.7354s |
| MSC+T | 0.7500 | 0.0469s |

+T: Eigen-decomposition tracking method [8]

[7] Benson et al., Higher-order organization of complex networks. Science, 2016.
[8] Chen Chen and Hanghang Tong. Fast eigen-functions tracking on dynamic graphs. In SDM, 2015

# Contents

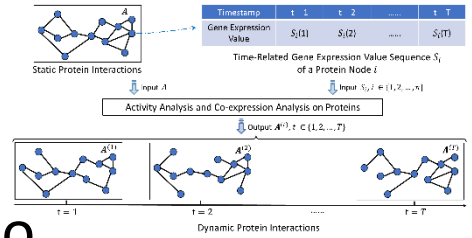
- Motivation and Overview of DPPIN
- Graph Generation Process in DPPIN
- Resource Statistics in DPPIN
- Experiment
- Conclusion

# Conclusion

- DPPIN Repository: https://github.com/DongqiFu/DPPIN
  - 12 dynamic network datasets from the biological domain
  - Graph structures, node labels, node and edge features

| Generated Dynamic PPINs | #Nodes | #Edges | Node Features | Edge Features | Node Label Rate | #Timestamps |
|---|---|---|---|---|---|---|
| DPPIN-Uetz | 922 | 2,159 | ✓ | ✓ | 921/922 (99.89%) | 36 |
| DPPIN-Ito | 2,856 | 8,638 | ✓ | ✓ | 2854/2856 (99.93%) | 36 |
| DPPIN-Ho | 1,548 | 42,220 | ✓ | ✓ | 1547/1548 (99.93%) | 36 |
| DPPIN-Gavin | 2,541 | 140,040 | ✓ | ✓ | 2538/2541 (99.88%) | 36 |
| DPPIN-Krogan (LCMS) | 2,211 | 85,133 | ✓ | ✓ | 2208/2211 (99.86%) | 36 |
| DPPIN-Krogan (MALDI) | 2,099 | 78,297 | ✓ | ✓ | 2097/2099 (99.90%) | 36 |
| DPPIN-Yu | 1,163 | 3,602 | ✓ | ✓ | 1160/1163 (99.74%) | 36 |
| DPPIN-Breitkreutz | 869 | 39,250 | ✓ | ✓ | 869/869 (100.00%) | 36 |
| DPPIN-Babu | 5,003 | 111,466 | ✓ | ✓ | 4997/5003 (99.88%) | 36 |
| DPPIN-Lambert | 697 | 6,654 | ✓ | ✓ | 697/697 (100.00%) | 36 |
| DPPIN-Tarassov | 1,053 | 4,826 | ✓ | ✓ | 1051/1053 (99.81%) | 36 |
| DPPIN-Hazbun | 143 | 1,959 | ✓ | ✓ | 143/143 (100.00%) | 36 |

- Algorithm:
  - Generation equations and algorithm
  - Generation program by Python




- Evaluation:
  - Link DPPIN with real-world application scenario
  - Provide research opportunities and future improvement directions


2022 IEEE International Conference on Big Data (IEEE BigData 2022)

December 17-20, 2022 @ Osaka, Japan

# Thanks!

**Dongqi Fu**
**(UIUC)**

**Jingrui He**
**(UIUC)**

**Presenter: Dongqi Fu**
**Email: dongqif2@illinois.edu**

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN