Deep Learning for Network Biology

Marinka Zitnik and Jure Leskovec

Stanford University
This Tutorial

snap.stanford.edu/deepnetbio-ismb

**ISMB 2018**

July 6, 2018, 2:00 pm - 6:00 pm
Why networks?

Networks are a general language for describing and modeling complex systems.
Network!
Why Networks? Why Now?

- **Question:** How are human genetic diseases and the corresponding disease genes related to each other?
- **Findings:** Genes associated with similar diseases are likely to interact and have similar expression.

**Why Networks? Why Now?**

- **Question:** How to simulate a basic eukaryotic cell?
- **Findings:** Simulations reveal molecular mechanisms of cell growth, drug resistance and synthetic life

**Question:** How to discover heterogeneity of cancer?

**Findings:** Analysis identifies new cancer subtypes with distinct patient survival.

Why Networks? Why Now?

- **Question:** How to study ecological systems?
- **Findings:** Pollinators interact with flowers in one season but not in another, and the same flower species interact with both pollinators and herbivores.

Why Networks? Why Now?

- **Question:** What are features of human microbiome?
- **Findings:** Microbiota reflects the seasonal availability of different types of food and differentiate industrialized and traditional populations.

Many Data are Networks

Patient networks
Hierarchies of cell systems
Disease pathways

Genetic interaction networks
Gene co-expression networks
Cell-cell similarity networks

Deep Learning for Network Biology -- snap.stanford.edu/deepnetbio-ismb -- ISMB 2018
Ways to Analyze Networks

- Predict a type of a given node
  - Node classification
- Predict whether two nodes are linked
  - Link prediction
- Identify densely linked clusters of nodes
  - Community detection
- How similar are two nodes/networks
  - Network similarity
Example: Node Classification

Machine Learning
Example: Node Classification

Classifying the function of proteins in the interactome!

Example: Link Prediction

Machine Learning
Example: Link Prediction

Predicting which diseases a new molecule might treat!

- **Drugs**
  - [Red pill]
  - [Blue pill]
  - [Yellow pill]
  - [Green pill]

- **Diseases**
  - [Headache]
  - [Viruses]
  - [Fever]
  - [Joint pain]

- **“Treats” relationship**
  - [Red pill - Headache]
  - [Blue pill - Viruses]
  - [Yellow pill - Fever]
  - [Green pill - Joint pain]

- **Unknown drug-disease relationship**
  - [Red pill - Joint pain]
  - [Blue pill - Fever]
  - [Yellow pill - Headache]
  - [Green pill - Viruses]
Example: Community Detection

Machine Learning
Example: Community Detection

Identifying disease proteins in the interactome!

Network Analytics Lifecycle

- (Supervised) Machine Learning Lifecycle: This feature, that feature. Every single time!

Raw Data → Structured Data → Learning Algorithm → Model

- Feature Engineering: Automatically learn the features

Downstream prediction task

Deep Learning for Network Biology -- snap.stanford.edu/deepnetbio-ismb -- ISMB 2018
Feature Learning in Graphs

**Goal:** Efficient task-independent feature learning for machine learning in networks.

\[ f: u \rightarrow \mathbb{R}^d \]

node \( u \)

Feature representation, embedding

\( u \rightarrow \mathbb{R}^d \)

vec
Feature Learning in Graphs

\[ f(\text{Input}) = \text{Output} \]

Disease similarity network

2-dimensional node embeddings

How to learn mapping function \( f \)?
Why Is It Hard?

- Modern deep learning toolbox is designed for grids or simple sequences
  - Images have 2D grid structure
  - Can define convolutions (CNN)
Why Is It Hard?

- Modern deep learning toolbox is designed for grids or simple sequences
  - Text and sequences have linear 1D structure
  - Can define sliding window, RNNs, word2vec, etc.
Why Is It Hard?

- But networks are far more complex!
  - Arbitrary size and complex topological structure (i.e., no spatial locality like grids)
  - No fixed node ordering or reference point
  - Often dynamic and have multimodal features

Text
This Tutorial

1) Node embeddings
   - Map nodes to low-dimensional embeddings
   - *Applications*: PPIs, Disease pathways

2) Graph neural networks
   - Deep learning approaches for graphs
   - *Applications*: Gene functions

3) Heterogeneous networks
   - Embedding heterogeneous networks
   - *Applications*: Human tissues, Drug side effects
Tutorial Resources

- Network **analytics tools** in SNAP
- Network data:
  - [snap.stanford.edu/projects.html](http://snap.stanford.edu/projects.html):
    - CRank, Decagon, MAMBO, NE, OhmNet, Pathways, and many others
- Deep learning **code bases**:
  - End-to-end examples in Tensorflow/PyTorch
  - Popular code bases for graph neural nets
  - Easy to adapt and extend for your application
Network Analytics in SNAP

- **Stanford Network Analysis Platform (SNAP)** is our general purpose, high-performance system for analysis and manipulation of large networks
  - [http://snap.stanford.edu](http://snap.stanford.edu)
  - Scales to massive networks with hundreds of millions of nodes and billions of edges

- **SNAP software**: C++, Python

- **Software requirements**: none
Biomedical network dataset collection:

- Different types of biomedical networks
- Ready to use for:
  - Algorithm benchmarking
  - Method development
  - Knowledge discovery
- Easy to link entities across datasets

Total: 250M entities, 2.2TB raw network data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Items</th>
<th>Raw Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DisGeNet</td>
<td>30K</td>
<td>10MB</td>
</tr>
<tr>
<td>STRING</td>
<td>10M</td>
<td>1TB</td>
</tr>
<tr>
<td>OMIM</td>
<td>25K</td>
<td>100MB</td>
</tr>
<tr>
<td>CTD</td>
<td>55K</td>
<td>1.2GB</td>
</tr>
<tr>
<td>HPRD</td>
<td>30K</td>
<td>30MB</td>
</tr>
<tr>
<td>BioGRID</td>
<td>64K</td>
<td>100MB</td>
</tr>
<tr>
<td>DrugBank</td>
<td>7K</td>
<td>60MB</td>
</tr>
<tr>
<td>Disease Ontology</td>
<td>10K</td>
<td>5MB</td>
</tr>
<tr>
<td>Protein Ontology</td>
<td>200K</td>
<td>130MB</td>
</tr>
<tr>
<td>Mesh Hierarchy</td>
<td>30K</td>
<td>40MB</td>
</tr>
<tr>
<td>PubChem</td>
<td>90M</td>
<td>1GB</td>
</tr>
<tr>
<td>DGIdb</td>
<td>5K</td>
<td>30MB</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>45K</td>
<td>10MB</td>
</tr>
<tr>
<td>MSigDB</td>
<td>14K</td>
<td>70MB</td>
</tr>
<tr>
<td>Reactome</td>
<td>20K</td>
<td>100MB</td>
</tr>
<tr>
<td>GEO</td>
<td>1.7M</td>
<td>80GB</td>
</tr>
<tr>
<td>GEOC (66 cancer projects)</td>
<td>40M</td>
<td>1TB</td>
</tr>
<tr>
<td>GTEx</td>
<td>50M</td>
<td>100GB</td>
</tr>
<tr>
<td>Many more...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Deep Learning for Network Biology -- snap.stanford.edu/deepnetbio-ismb -- ISMB 2018
Deep Learning Code Bases

This tutorial: Using graph neural networks:

- End-to-end examples in Tensorflow/PyTorch
- Popular code bases for graph neural nets
- Easy to adapt and extend for your application
PhD Students

Claire Donnat
Mitchell Gordon
David Hallac
Emma Pierson
Geet Sethi

Himabindu Lakkaraju
Rex Ying
Tim Althoff
Will Hamilton
Alex Porter

Post-Doctoral Fellows

Baharan Mirzasoleiman
Marinka Zitnik
Michele Catasta
Srijan Kumar

Research Staff

Stephen Bach
Adrijan Bradaschia
Rok Sosic

Industry Partnerships

Funding

Collaborators

Dan Jurafsky, Linguistics, Stanford University
Christian Danescu-Miculescu-Mizil, Information Science, Cornell University
Stephen Boyd, Electrical Engineering, Stanford University
David Gleich, Computer Science, Purdue University
VS Subrahmanian, Computer Science, University of Maryland
Sarah Kunz, Medicine, Harvard University
Russ Altman, Medicine, Stanford University
Jochen Profit, Medicine, Stanford University
Eric Horvitz, Microsoft Research
Jon Kleinberg, Computer Science, Cornell University
Sendhil Mullainathan, Economics, Harvard University
Scott Delp, Bioengineering, Stanford University
Jens Ludwig, Harris Public Policy, University of Chicago
WE’RE HIRING!

Many interesting high-impact projects in Machine Learning and Large Biomedical Data

Applications: Precision Medicine & Health, Drug Repurposing, Drug Side Effect modeling, Network Biology, and many more