Massachusetts Institute of Technology

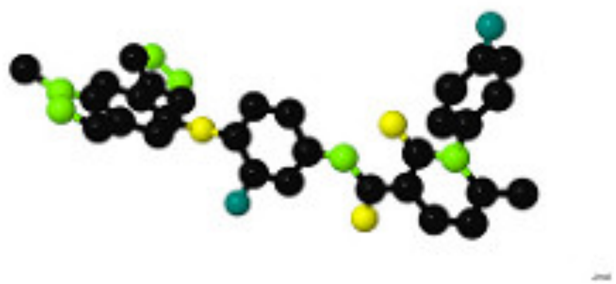
CSAIL

Representational Power of Graph Neural Networks

Stefanie Jegelka
MIT

joint work with
Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, Weihua Hu, Jure Leskovec, Jingling Li, Mozhi Zhang, Simon S. Du

Learning with Graphs

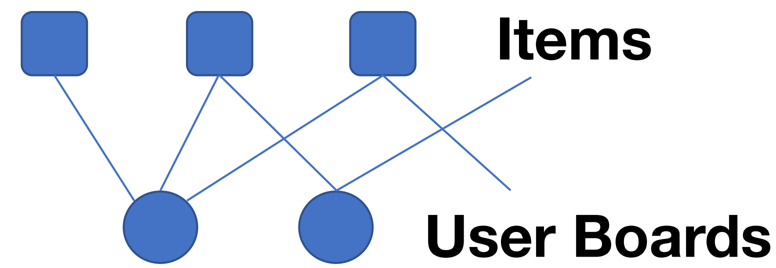


Graph
Molecule

→

Property
Solubility
Toxicity
Drug efficacy
…

(Duvenaud et al, 2015, …)

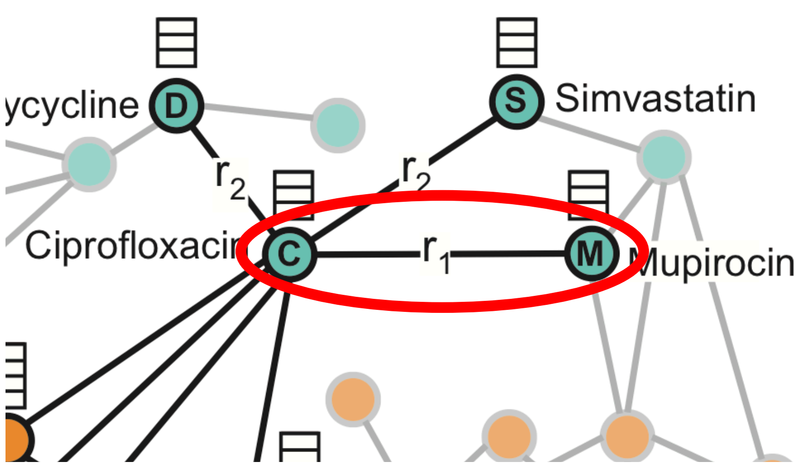


Node
Item

→

Node
Item

(Ying et al, 2018)

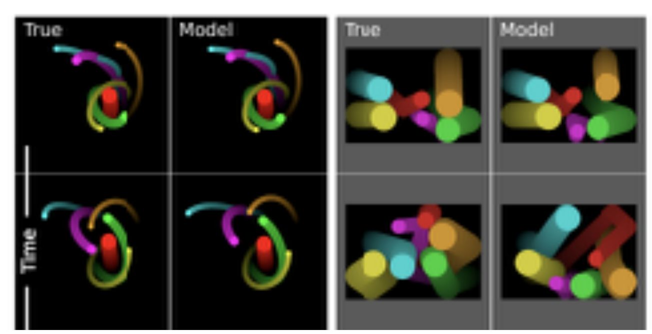


Pair of Nodes
Drugs

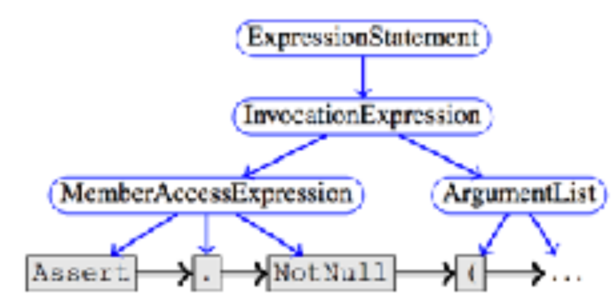
→

Edge
Interaction

(Zitnik et al, 2018)

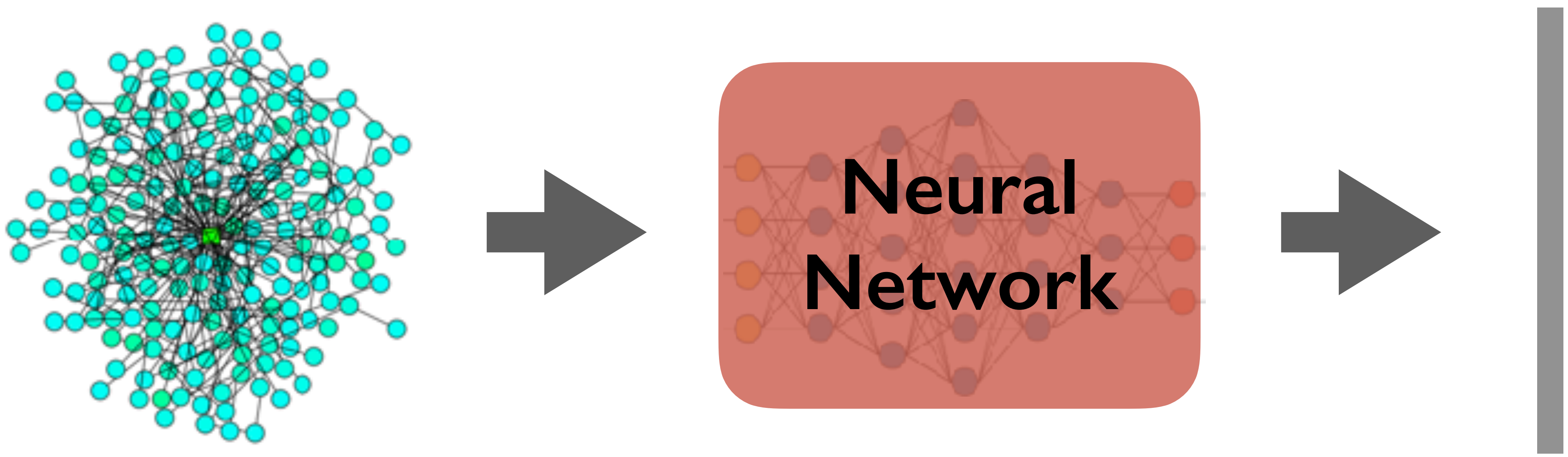


(Battaglia et al. 2016)



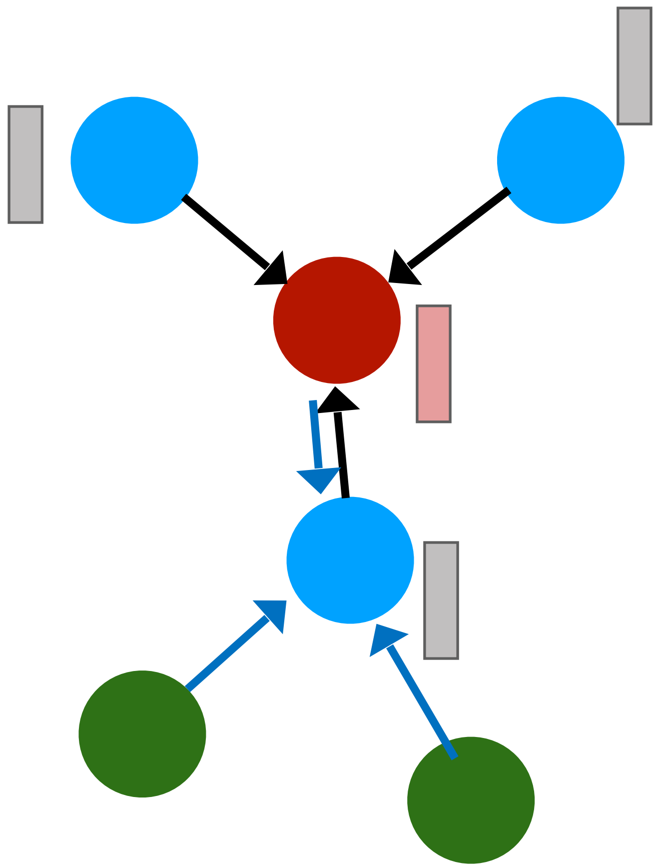
(Allamanis et al. 2018)

Outline



- Discriminative Power of Graph Neural Networks
  K. Xu, W. Hu, J. Leskovec, S. Jegelka, ICLR 2019
- Network Depth and Graph Structure
  K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka, ICML 2018
- Neural Networks for Reasoning
  K. Xu, J. Li, M. Zhang, S. Du, K. Kawarabayashi, S. Jegelka. ArXiv 2019

Graph Neural Networks



node embedding

In each round:

feature description of node u in round k-1

Aggregate over neighbors

$$a_v^{(k)} = \text{AGGREGATE}^{(k)}\left(\left\{h_u^{(k-1)} : u \in \mathcal{N}(v)\right\}\right)$$

Combine with current node

$$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, a_v^{(k)}\right)$$

graph embedding

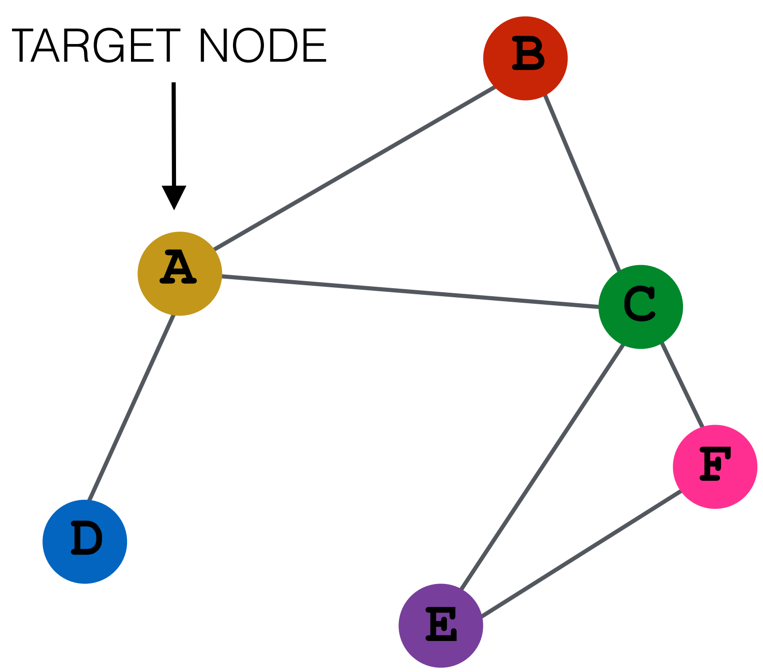
....

Graph-level **readout**

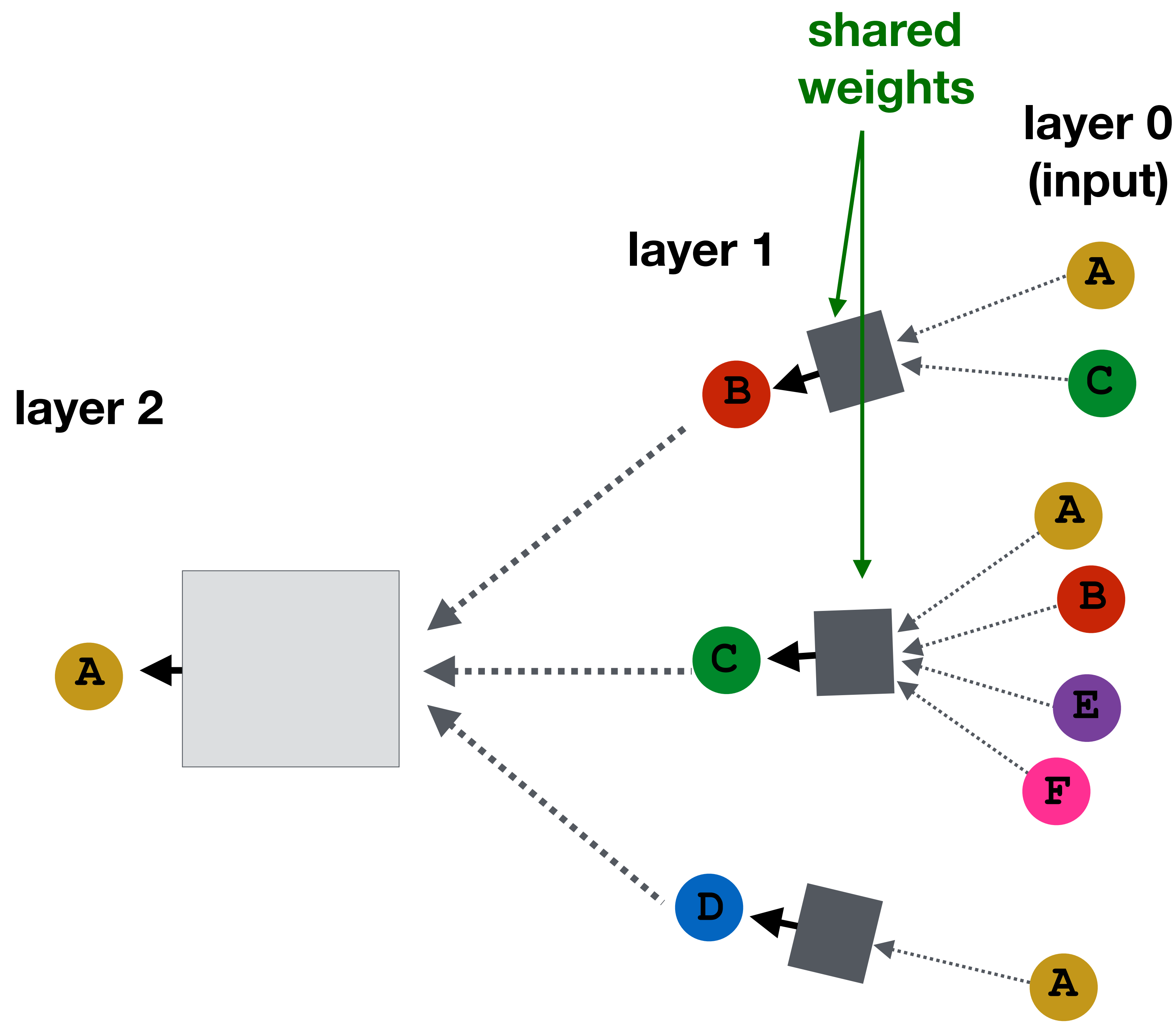
$$h_G = \text{READOUT}\left(\left\{h_v^{(K)} \mid v \in G\right\}\right)$$

(Battaglia et al., 2016; Defferrard et al., 2016; Duvenaud et al., 2015; Hamilton et al., 2017a; Kearnes et al., 2016; Kipf & Welling, 2017; Li et al., 2016; Velickovic et al., 2018; Verma & Zhang, 2018; Ying et al., 2018b; Zhang et al., 2018, …)

Graph Neural Networks

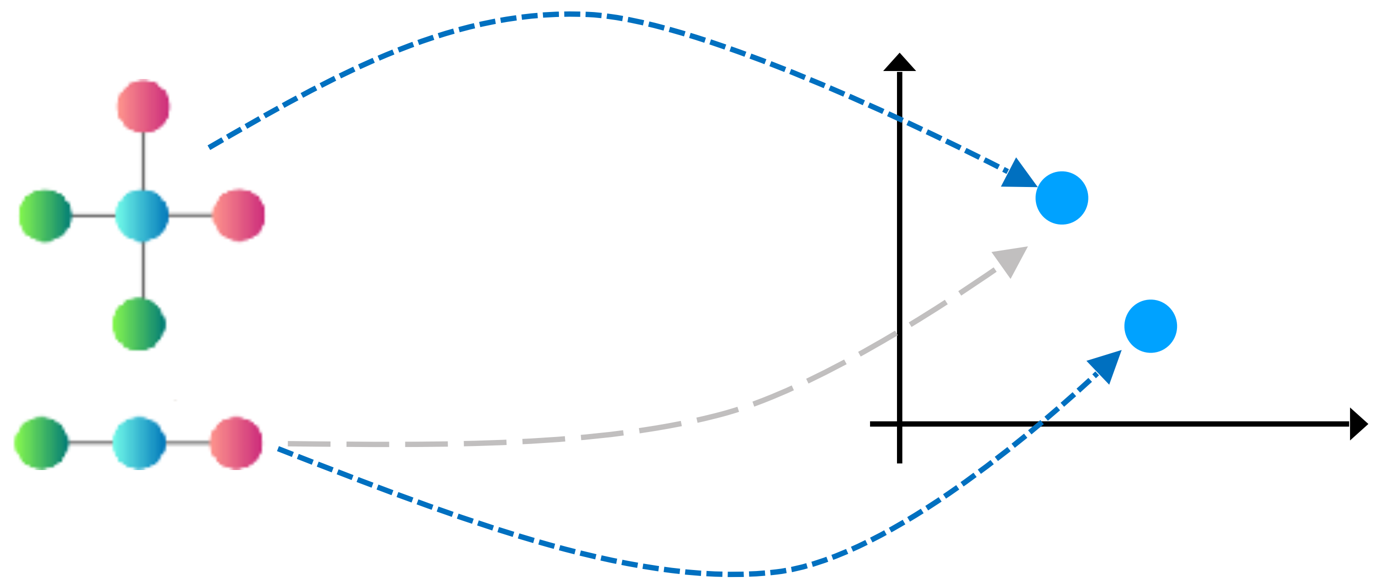


INPUT GRAPH



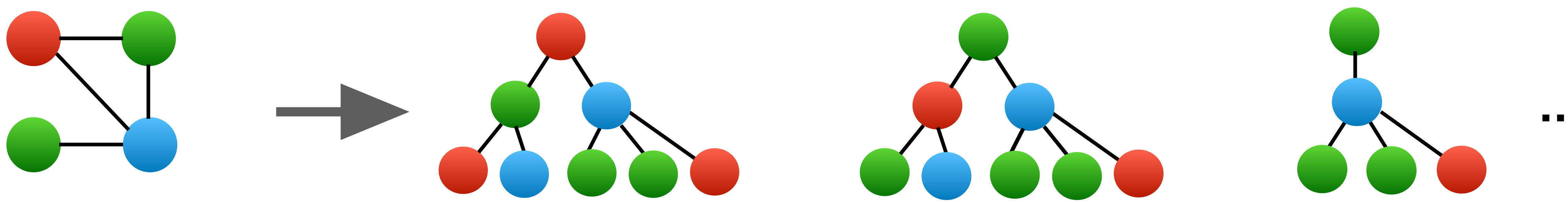
(illustrations: J. Leskovec)

Graph Neural Networks



Which graphs can GNNs distinguish?
What does this depend on?

Discriminative Power



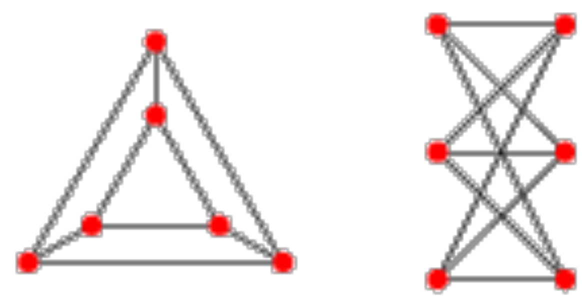
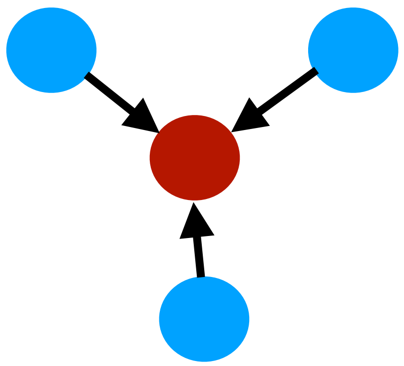


image: Wolfram

Lemma *(XHLJ19)*
Aggregation-based GNNs are at most as powerful as the 1-dim. Weisfeiler-Lehman graph isomorphism test*.

**(Weisfeiler & Lehman 1968, Babai, Erdös, Selkow 1980, Babai & Kucera 1979, Cai, Furer, Immerman 1992, Evdokimov & Ponomarenko 1999, Douglas 2011,…)*

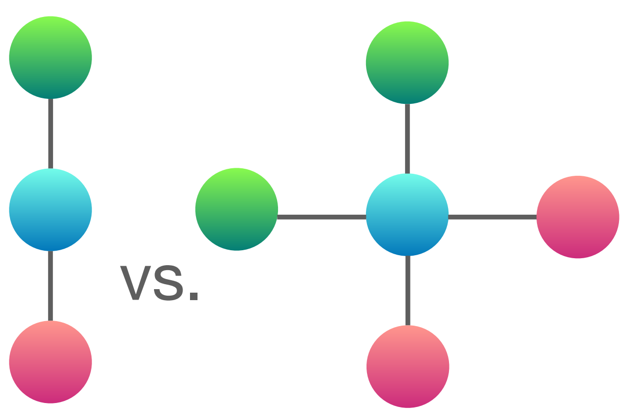
Reach maximum power?



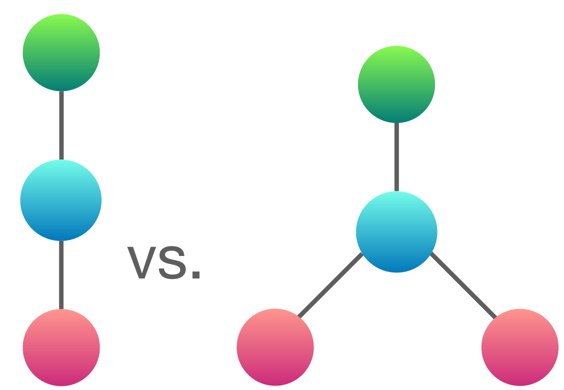
Aggregate from neighbors
e.g. **max/mean pooling**

$$g(X) = \phi\left(\mathrm{MEAN}\{f(x) : x \in X\}\right)$$ *(Kipf et al. 2017)*

$$g(X) = \phi\left(\mathrm{MAX}\{f(x) : x \in X\}\right)$$ *(Hamilton et al. 2017)*



max / mean pooling fail

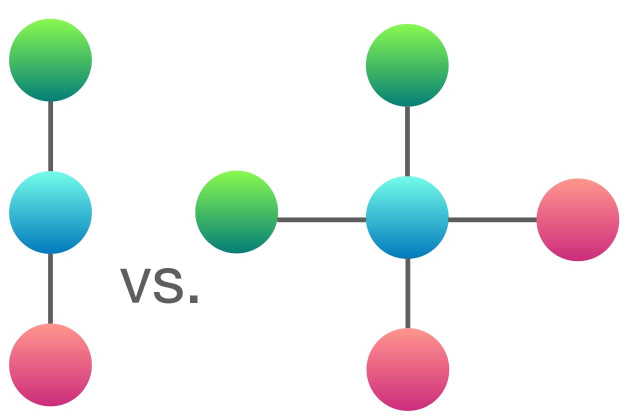


max pooling fails

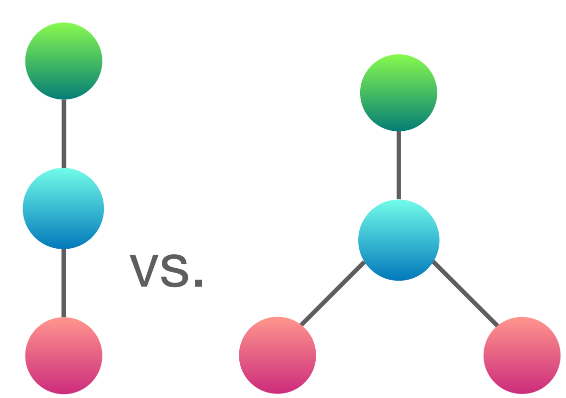
Failure: same node representations

Discriminative Schemes

Theorem *(XHLJ19)*
If Aggregate, Combine and Readout are *injective*, then the network is as discriminative as the 1-dim. WL test.



max / mean fail



max fails

A powerful GNN (GIN)

Lemma

Any multi-set function g can be decomposed as

$$g\left(X\right) = \phi\left(\sum_{x \in X} f\left(x\right)\right)$$

- Aggregation: sum & appropriate nonlinearity:

$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1 + \epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right)$$

A powerful GNN (GIN)

Lemma

Any multi-set function g can be decomposed as

$$g\left(X\right) = \phi\left(\sum_{x \in X} f(x)\right)$$

- Aggregation: sum & appropriate nonlinearity:

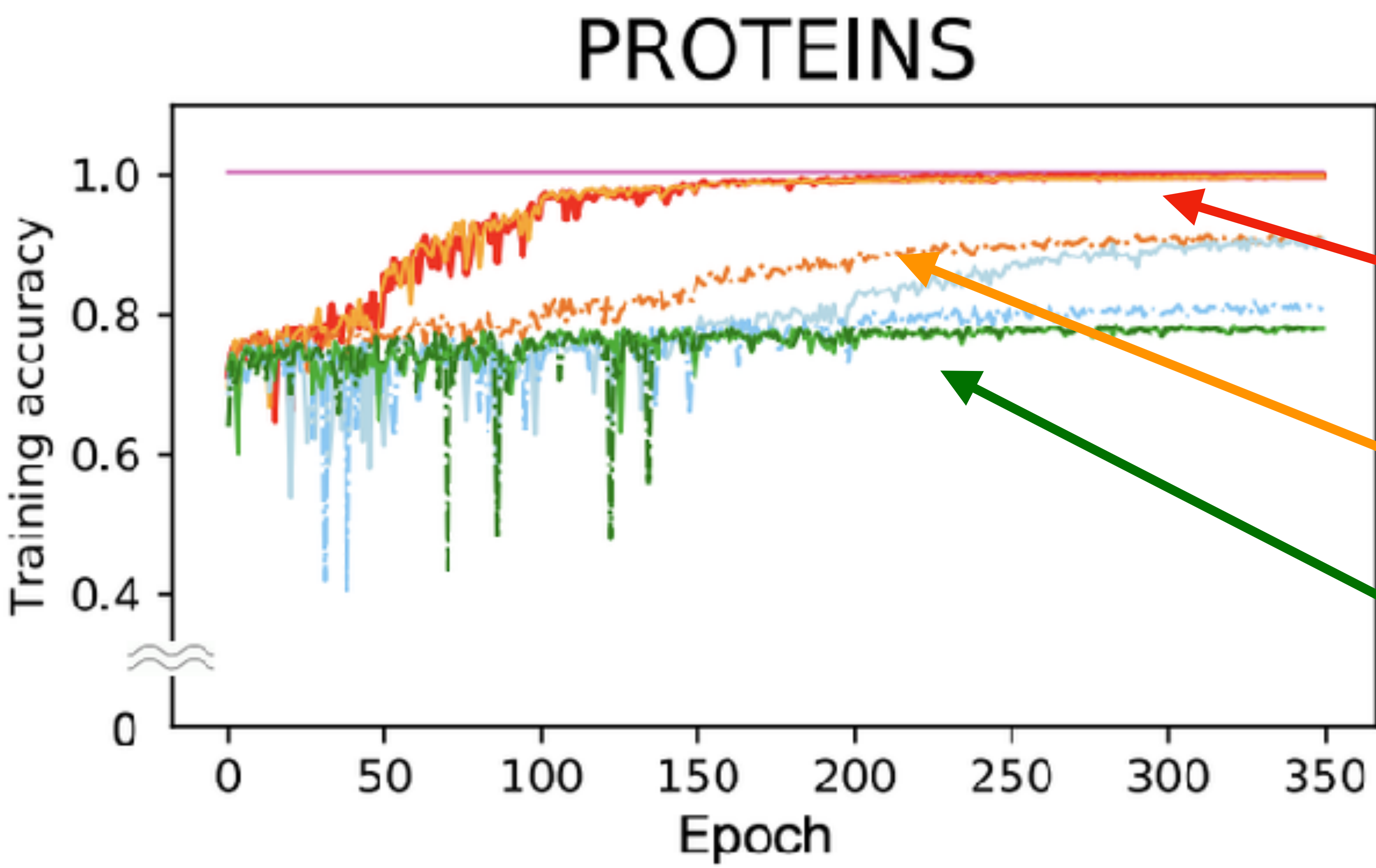
$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1 + \epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right)$$

- Readout: concatenation or sum

$$h_G = \mathrm{CONCAT}\left(\mathrm{READOUT}\left(\left\{h_v^{(k)} | v \in G\right\}\right) \mid k = 0, 1, \ldots, K\right)$$

Example Empirical Results

$$g\left(X\right)=\phi\left(\sum_{x\in X}f\left(x\right)\right)$$

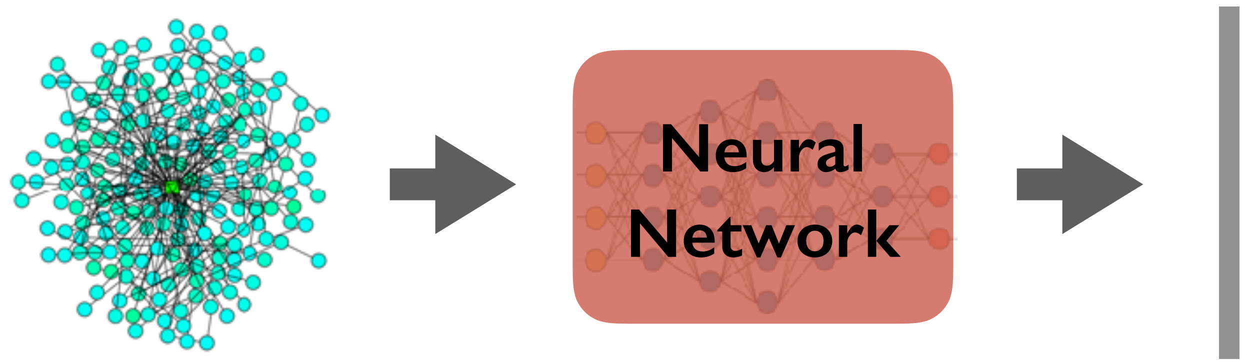


Sum — MLP (injective)

Sum — linear+ReLu

Mean/Max —
MLP/linear+ReLu

Towards Understanding Graph Neural Networks



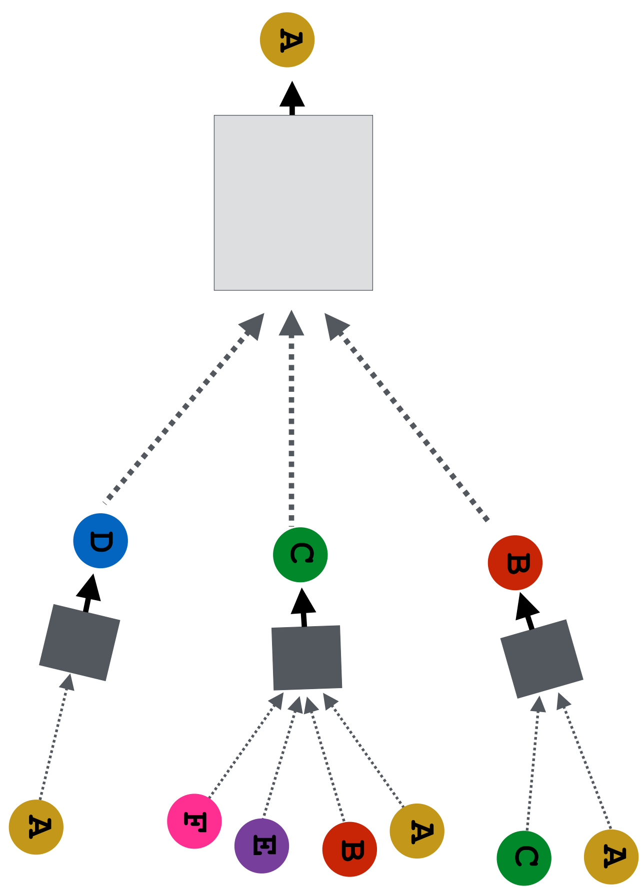
- Discriminative Power of Graph Neural Networks
  K. Xu, W. Hu, J. Leskovec, S. Jegelka, ICLR 2019

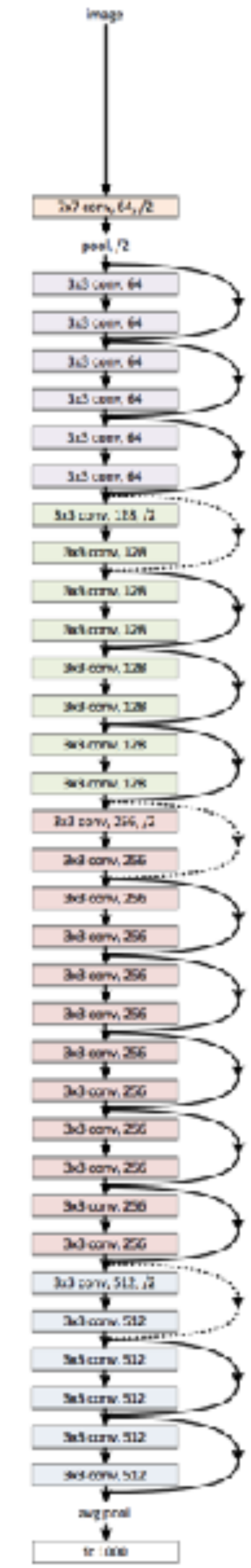
- Network Depth and Graph Structure
  K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka, ICML 2018

- Neural Networks for Reasoning

Depth

Why is deeper not always better here?





comparison: 34-layer

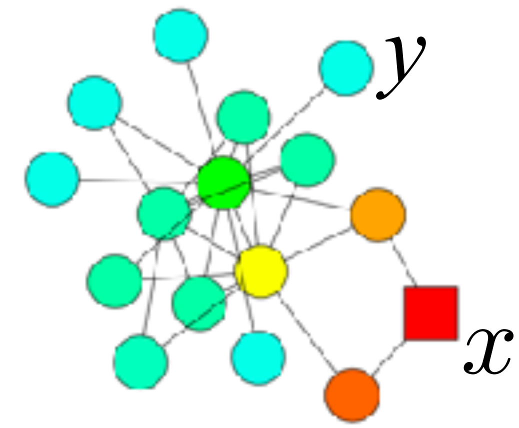
ResNet for **Images**

(He et al., 2016)

Neighborhood Aggregation

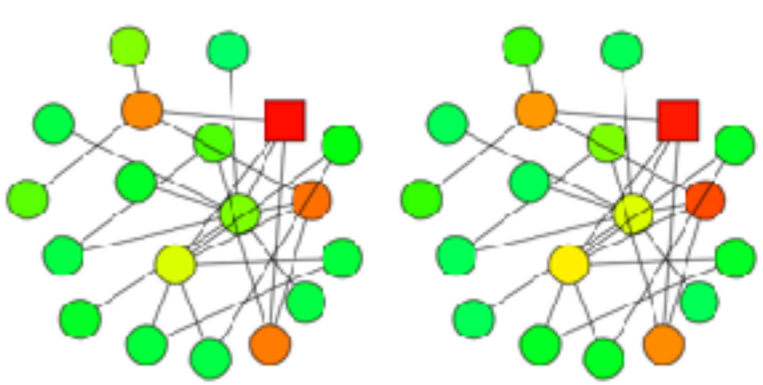
Influence distribution

$$I_x(y) = e^T \left[\frac{\partial h_x^{(k)}}{\partial h_y^{(0)}} \right] e \Big/ \left(\sum_{z \in V} e^T \left[\frac{\partial h_x^{(k)}}{\partial h_z^{(0)}} \right] e \right)$$



how much is node x's feature influenced by node y

Theorem *(XLTSKJ18)* Under simplifying assumptions*:
Influence distribution of a *k-layer* GCN equals a *k-step* random walk distribution in expectation. ResNet: Lazy random walk.

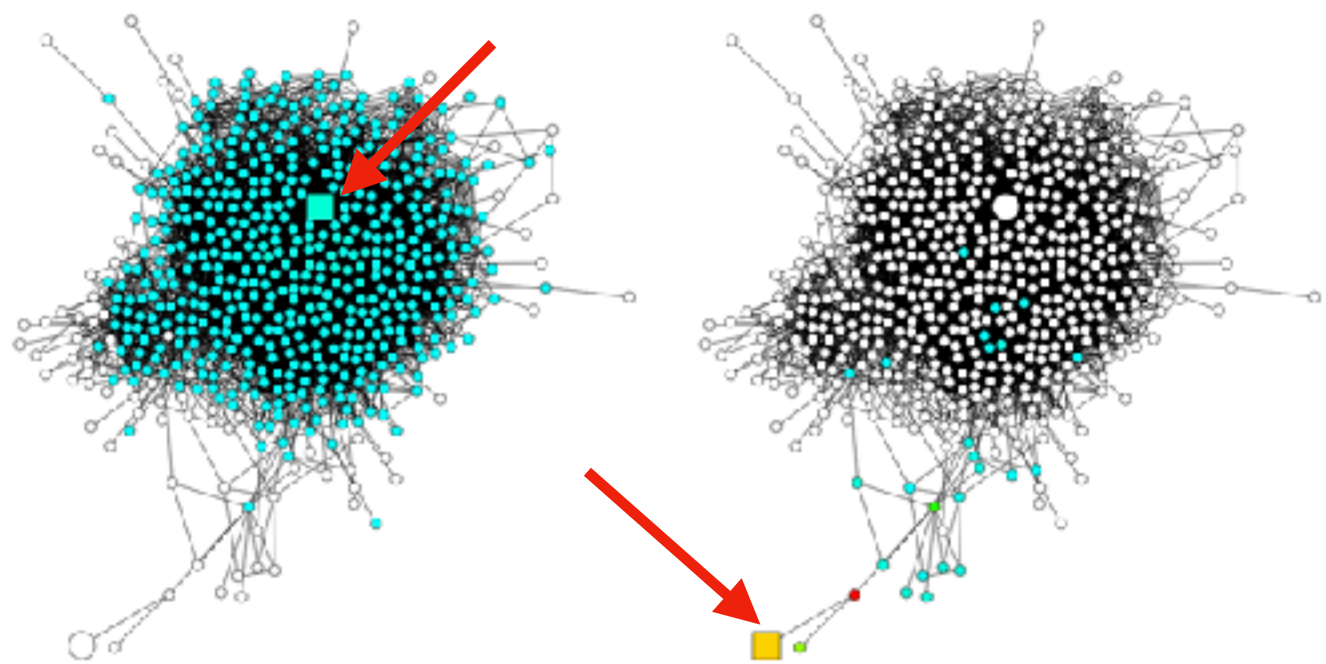


(a) 2 layer GCN (b) 2 step r.w.

* *as in (Choromanska et al 2015, Kawaguchi 2016)*

Implications

Random Walk convergence / size of feature neighborhood depends on **expansion properties of the graph**



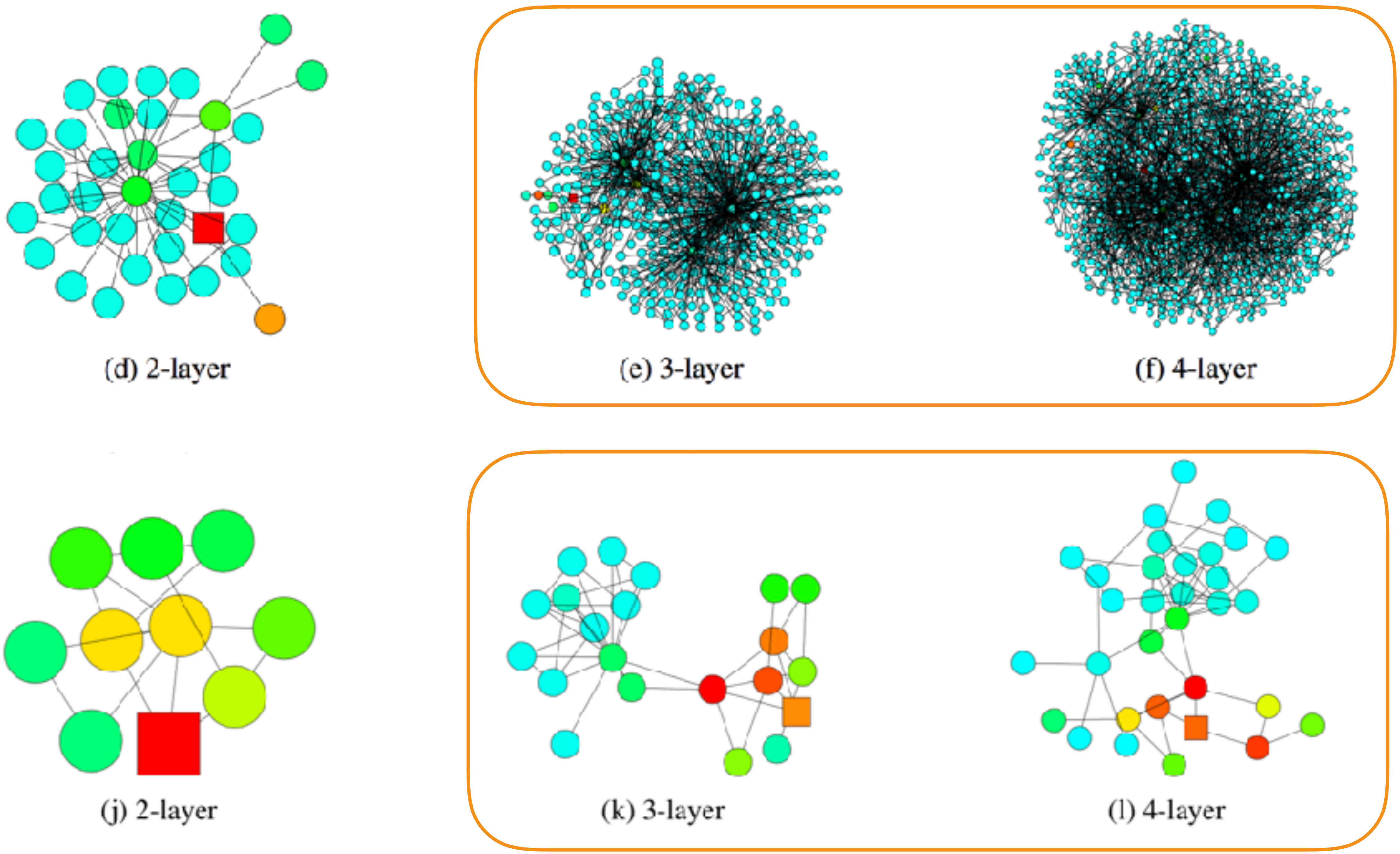
(a) 4 steps at core (b) 4 steps at tree

Conjecture:

Different depths work for different subgraph structures

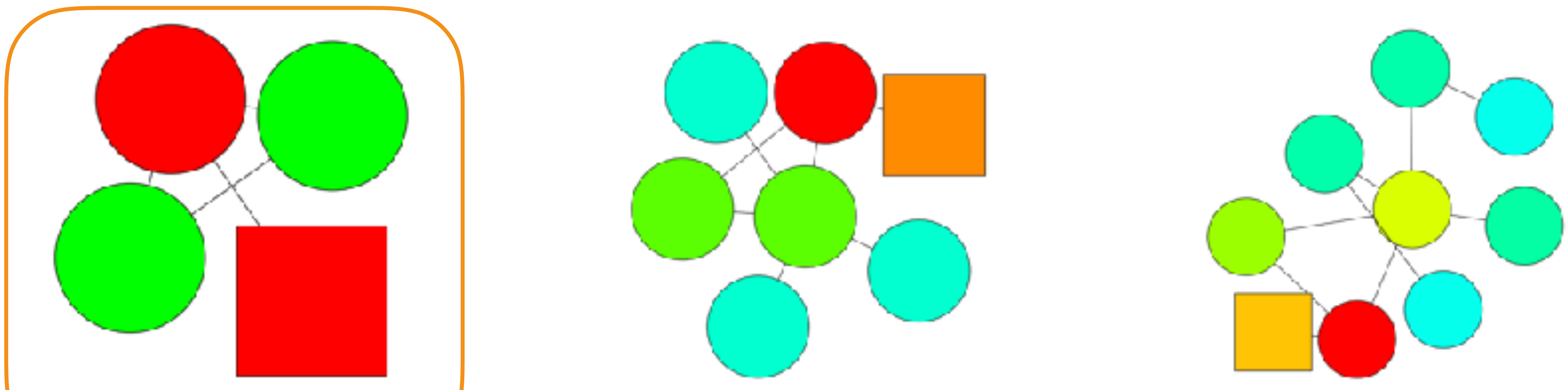
Examples

3 and 4-layer models make incorrect prediction

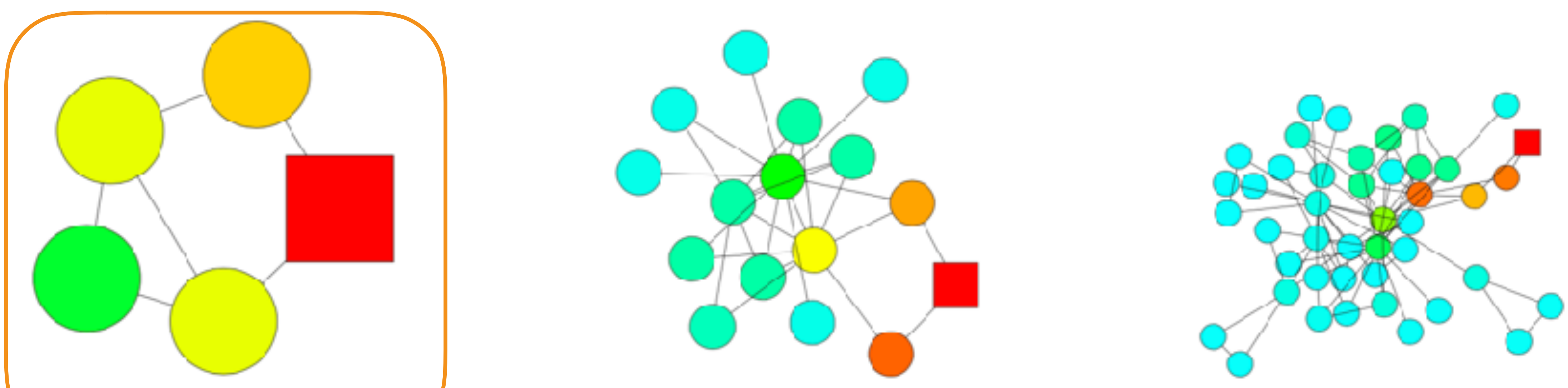


Examples

2-layer models make incorrect prediction

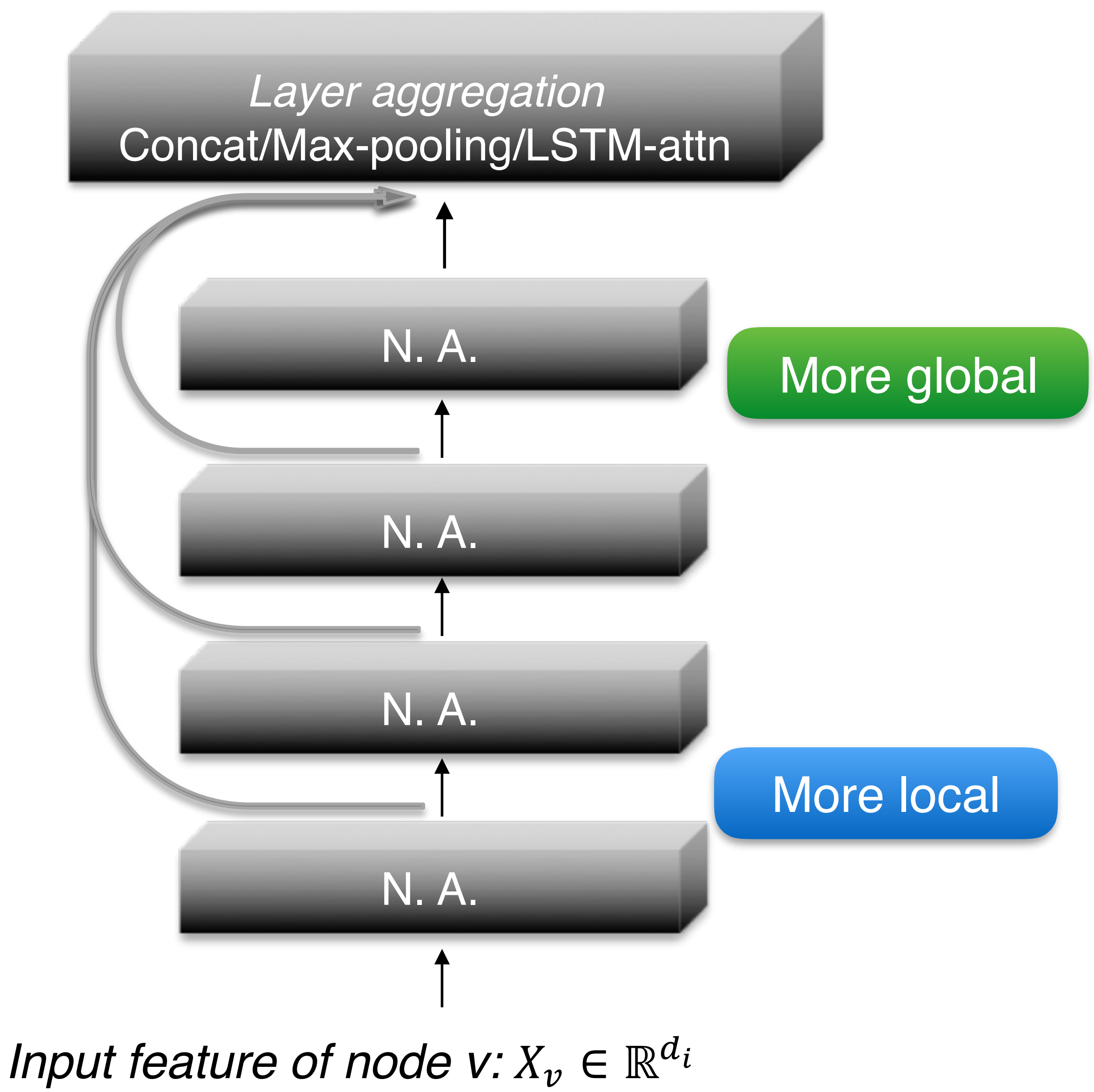


(d) 2-layer (e) 3-layer (f) 4-layer



(a) 2-layer (b) 3-layer (c) 4-layer

Adaptive depth



adapting to subgraph structures empirically improves performance by 2-30%

JK-Concat adapts once for entire dataset/graph:
best for more regular graphs (images: DenseNet)

JK-LSTM adapts for each node/subgraph:
best for large graphs with diverse structures

Towards Understanding Graph Neural Networks

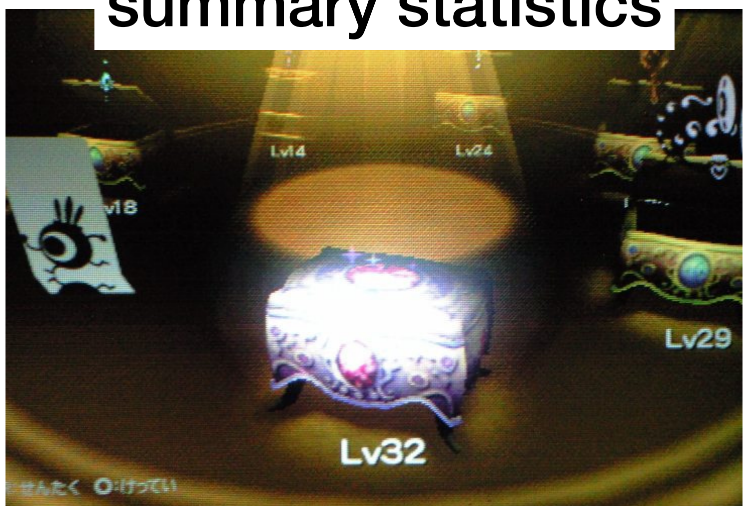
- Discriminative Power of Graph Neural Networks
  K. Xu, W. Hu, J. Leskovec, S. Jegelka, ICLR 2019

- Network Depth and Graph Structure
  K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka, ICML 2018

- Neural Networks for Reasoning
  K. Xu, J. Li, M. Zhang, S. Du, K. Kawarabayashi, S. Jegelka. ArXiv 2019

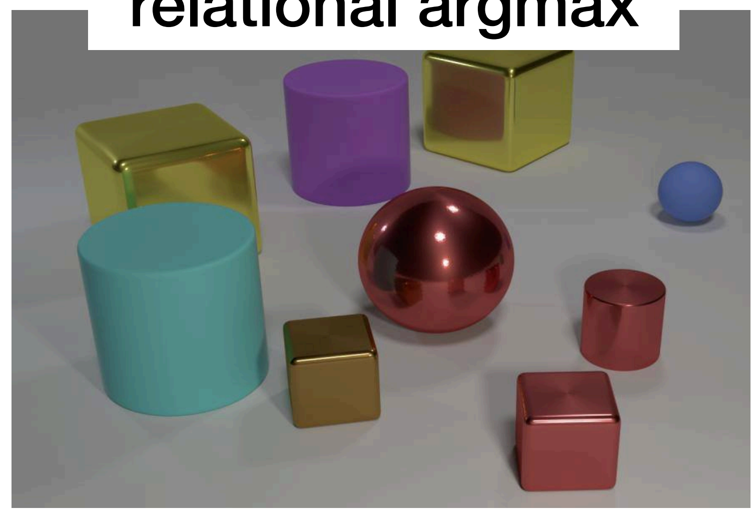
Reasoning Tasks

summary statistics



What is the maximum value difference among treasures?

"relational argmax"

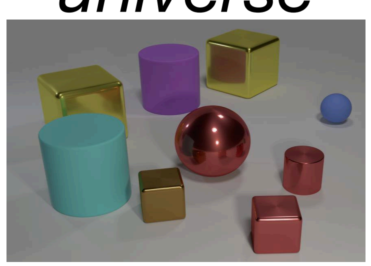


What are the colors of the farthest pair of objects?

dynamic programming

What is the cost to defeat monster X by following the optimal path?

universe



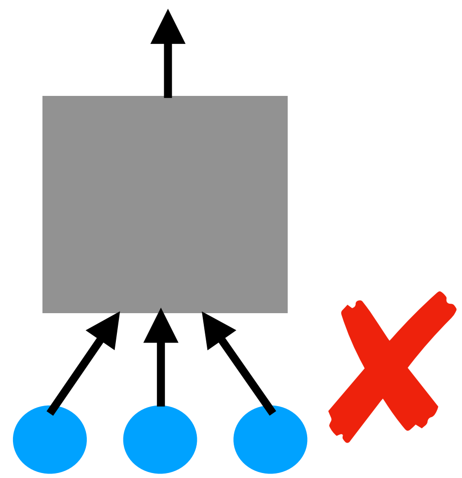
?

Neural Network

answer

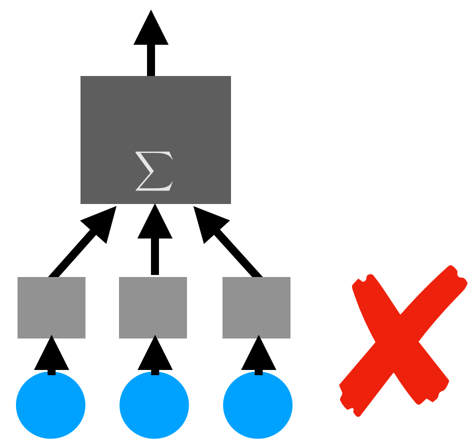
(Johnson et al., 2017a; Weston et al., 2015; Hu et al., 2017; Fleuret et al., 2011; Antol et al., 2015; Battaglia et al., 2016; Watters et al., 2017; Fragkiadaki et al., 2016; Chang et al., 2017; Saxton et al., 2019; Chang et al., 2019; Santoro et al., 2018; Zhang et al., 2019, …)

MLP

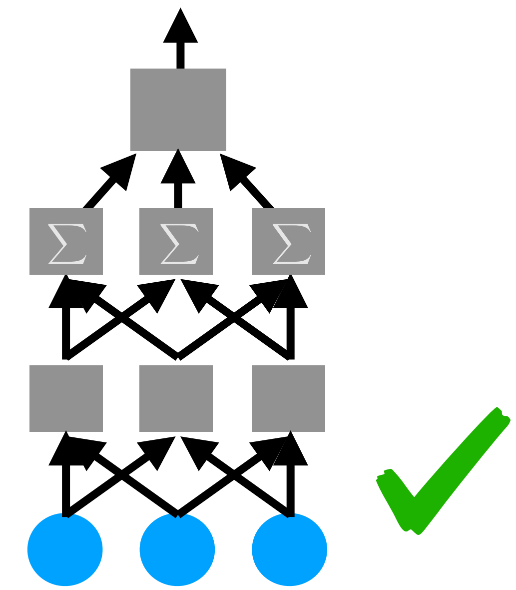


Deep Sets

(Zaheer et al 2017)



GNN



Algorithmic Alignment: Network can mimic algorithm via *few, easy-to-learn* modules

Bellman-Ford

for k = 1 … |S| - 1:

  for u in S:

    d[k][u] = $\min_v$ d[k-1][v] + cost (v, u)

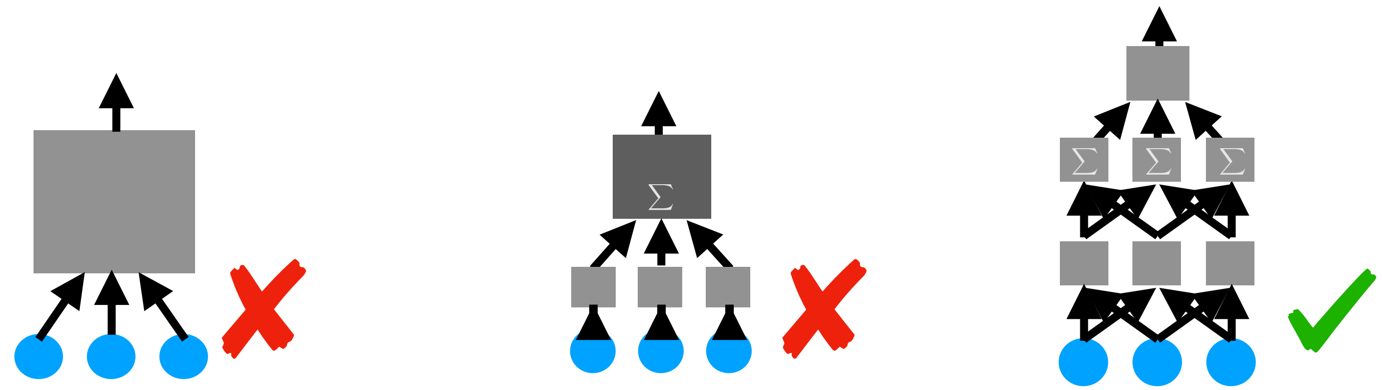
GNN

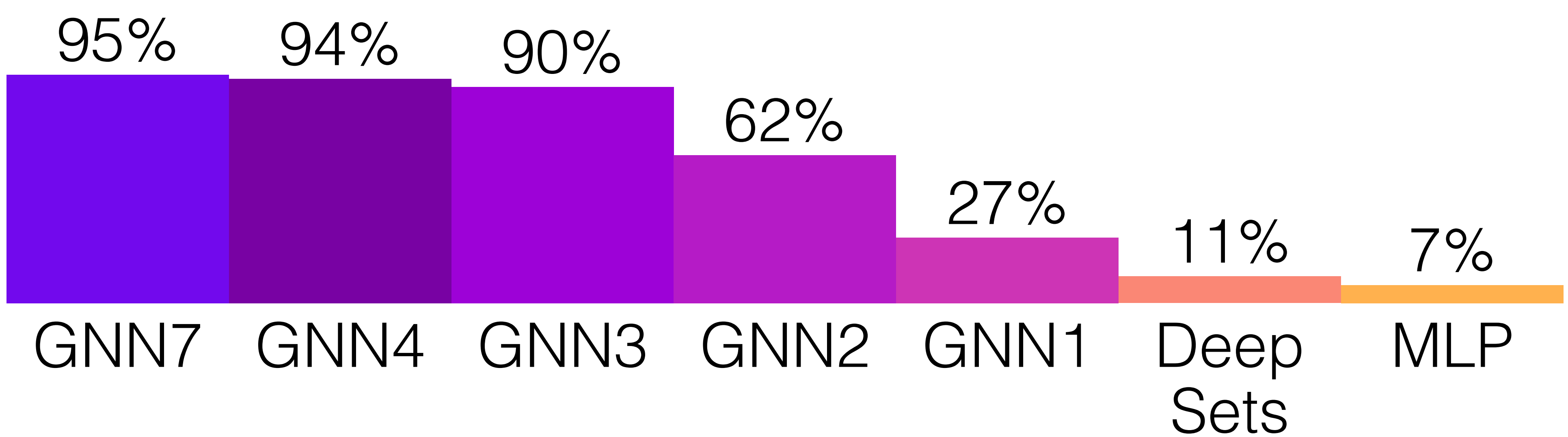
for k = 1 … GNN iter:

  for u in S:

    $h_u^{(k)} = \Sigma_v \text{ MLP}(h_v^{(k-1)}, h_u^{(k-1)})$

Alignment and Learning





More generally: GNNs align with a class of DP

$$\text{Answer}[k][i] = \text{DP-Update}(\{\text{Answer}[k-1][j],\ j = 1 \ldots n\})$$

Includes many reasoning tasks: visual question answering, physical reasoning,…

Alignment & Implications

A neural network (M, ϵ, δ)-aligns with an algorithm if it can mimic the algorithm via *n* different (shared) network modules, each of which can be learned with at most M/n samples.

Theorem

If a network and task algorithm (M, ϵ, δ) -align, then, under assumptions, the task is $(M, O(\epsilon), O(\delta))$ -learnable by the network.

Representational Power of GNNs

- **Discriminative power of graph neural networks**
  - Aggregation schemes important!
- **Local expansion and network depth**
  - Adaptive depths for different graph structures
- **Alignment for reasoning tasks**

References

- K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka. Representation Learning on Graphs with Jumping Knowledge Networks. International Conference on Machine Learning (ICML), 2018.
- K. Xu, W. Hu, J. Leskovec and S. Jegelka. How Powerful are Graph Neural Networks? International Conference on Learning Representations (ICLR), 2019.
- K. Xu, J. Li, M. Zhang, S. Du, K. Kawarabayashi, S. Jegelka. What Can Neural Networks Reason About? arXiv 2019