Subgraph Neural Networks

Emily Alsentzer*, Sam Finlayson*, Michelle Li, Marinka Zitnik
{emilya, sgfin}@mit.edu, michelleli@g.harvard.edu, marinka@hms.harvard.edu

HARVARD MEDICAL SCHOOL

HARVARD–MIT HEALTH SCIENCES AND TECHNOLOGY

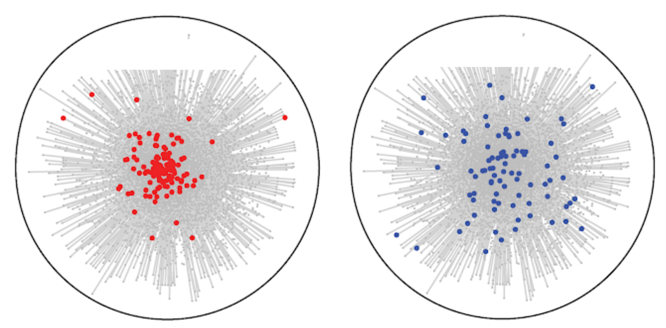
Project Website: https://zitniklab.hms.harvard.edu/projects/SubGNN

Motivation

Limited existing work on subgraph representation learning

Subgraphs present unique challenges for representation learning that do not exist for nodes or entire graphs

- Need to jointly predict over structures of varying size
- Subgraphs can be localized or distributed throughout the graph



Localized Distributed

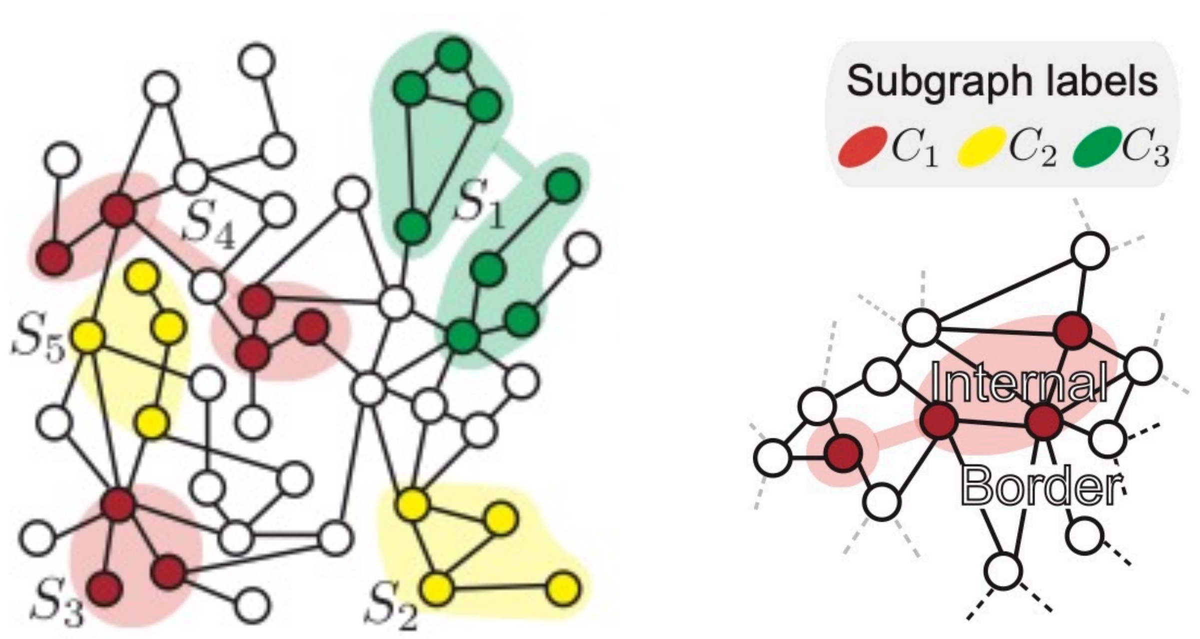
- Subgraphs have rich topology and connectivity patterns, both internally & externally with the rest of G

	Internal (I)	Border (B)
Position (P)	Distance between S_i's components	Distance between S_i and rest of G
Neighborhood (N)	Identity of S_i's internal nodes	Identity of S_i's border nodes
Structure (S)	Internal connectivity of S_i	Border connectivity of S_i

Open Questions

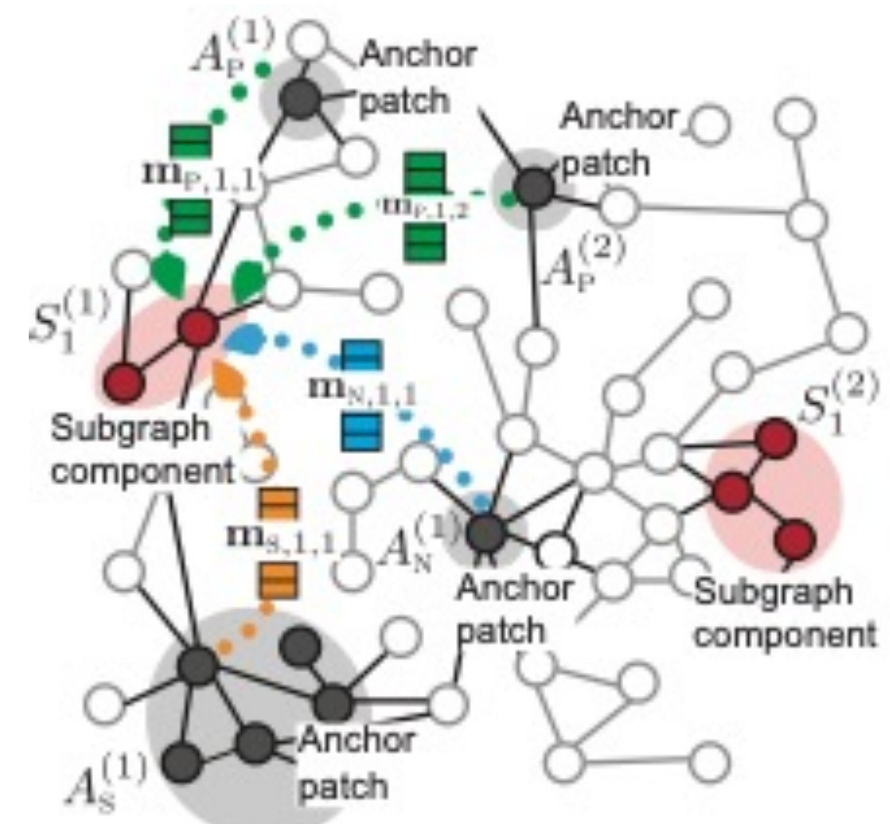
How to represent subgraphs that comprise of multiple disparate components?

How to represent rich internal and border subgraph topology?

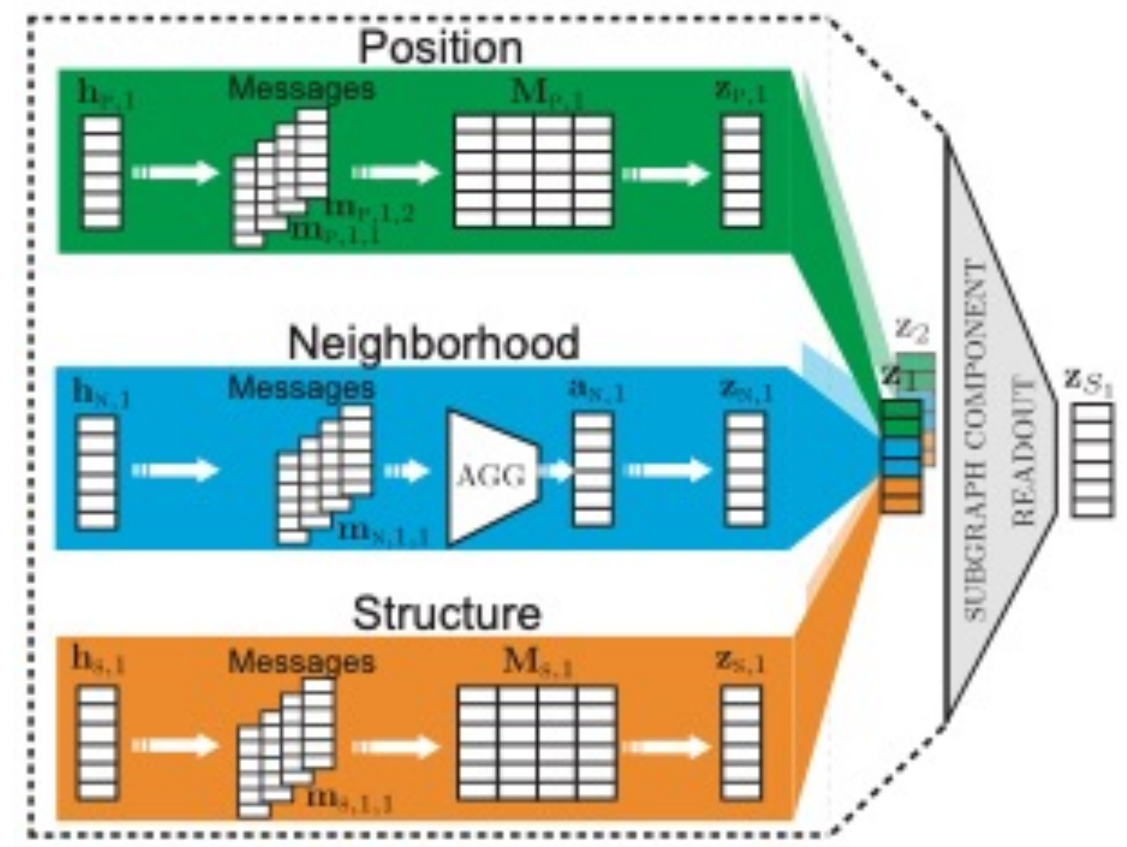


Goal: Learn a representation of each subgraph S_i such that the likelihood of preserving properties of S_i is maximized in the embedding space

Our Solution: SubGNN



Message Passing at **Subgraph-Level**



Property-Aware Routing Channels

- SubGNN learns subgraph representations in a hierarchical fashion by propagating neural messages from anchor patches sampled throughout the graph to subgraph components and aggregating the resulting representations into a final subgraph embedding
- SubGNN specifies three channels, each designed to capture a distinct property: position, neighborhood, structure
- Message passing occurs separately within each channel: messages are sent from anchor patches and weighted by a property-specific similarity function
- Each channel x has three key elements:
  - **Sampling function ϕ_x** to sample anchor patches (helper subgraphs randomly sampled from G)
  - **Anchor patch encoder ψ_x** to embed the anchor patches
  - **Similarity function γ_x** to weight messages sent from anchor patches to connected components

Strong performance on **8 novel tasks** for subgraph prediction

- 4 synthetic datasets designed to test ability of methods to capture subgraph properties. Binned values of each metric serve as subgraph labels
- 4 real world datasets that consist of a base graph and subgraphs with associated labels. HPO-METAB and HPO-NEURO are clinical diagnostic tasks that ask the following: *what is the subcategory of metabolic/neurological disease consistent with the phenotypes (i.e. phenotype subgraph)?*
- We compare against 7 baselines that represent common approaches to encode subgraphs, but that fail to encode all subgraph topology.
- We find that SubGNN outperforms baselines by an average of 77% on synthetic and 125% on real-world datasets and that SubGNN channels encode their intended properties.

Ablation study over SubGNN channels on 4 synthetic datasets

SUB-GNN Channel	DENSITY	CUT RATIO	CORENESS	COMPONENT
Position (P)	0.758±0.046	0.516±0.083	0.581±0.044 ✓	**0.958**±0.098 ✓
Neighborhood (N)	0.777±0.057	0.313±0.087	0.485±0.075	0.823±0.089
Structure (S)	**0.919**±0.016 ✓	**0.629**±0.039 ✓	**0.663**±0.058 ✓	0.600±0.170
All (P+N+S)	0.894±0.025	0.458±0.101	0.659±0.092	0.726±0.120

Channels designed to encode relevant properties yield best performance (e.g. structure channel performs best on DENSITY (internal structure dataset)

Micro F1 on Synthetic Datasets

Method	DENSITY	CUT RATIO	CORENESS	COMPONENT
SUBGNN (Ours)	**0.919**±0.016	**0.629**±0.039	**0.659**±0.092	**0.958**±0.098
Node Averaging	0.429±0.041	0.358±0.055	0.530±0.050	0.516±<0.001
Meta Node (GIN)	0.442±0.052	0.423±0.057	0.611±0.050	0.784±0.046
Meta Node (GAT)	0.690±0.021	0.284±0.052	0.519±0.076	0.935±<0.001
Sub2Vec Neighborhood	0.345±0.066	0.339±0.058	0.381±0.047	0.568±0.039
Sub2Vec Structure	0.339±0.036	0.345±0.121	0.404±0.097	0.510±0.013
Sub2Vec N & S Concat	0.352±0.071	0.303±0.062	0.356±0.050	0.568±0.021
Graph-level GNN	0.803±0.039	0.329±0.073	0.370±0.091	0.500±0.068

Micro F1 on Real-World Datasets

Method	PPI-BP	HPO-NEURO	HPO-METAB	EM-USER
SUBGNN (+ GIN)	**0.599**±0.024	0.632±0.010	**0.537**±0.023	0.814±0.046
SUBGNN (+ GraphSAINT)	0.583±0.017	**0.644**±0.019	0.428±0.035	0.816±0.040
Node Averaging	0.297±0.027	0.490±0.059	0.443±0.063	0.808±0.138
Meta Node (GIN)	0.306±0.025	0.233±0.086	0.151±0.073	0.480±0.089
Meta Node (GAT)	0.307±0.021	0.259±0.063	0.138±0.034	0.471±0.048
Sub2Vec Neighborhood	0.306±0.009	0.211±0.068	0.132±0.047	0.520±0.090
Sub2Vec Structure	0.306±0.021	0.223±0.065	0.124±0.025	**0.859**±0.014
Sub2Vec N & S Concat	0.309±0.023	0.206±0.073	0.114±0.021	0.522±0.043
Graph-level GNN	0.398±0.058	0.535±0.032	0.452±0.025	0.561±0.059

Standard deviations from runs with 10 random seeds