**Summary**
- **Goal:** to find high-quality node representations on large-scale hypergraphs
- **Previous Work:** Limited downstream tasks: A few single-entity level downstream tasks have been used to evaluate the hypergraph neural network models. Small-scale benchmark datasets: The evaluation of current hypergraph neural network models has been limited to small datasets (10k-scale). Underdeveloped training strategy for large-scale hypergraphs: Most of the current hypergraph learning approaches are not scalable.
- **Contributions:**
  - We present two new pair-level prediction tasks.
  - We construct and publicly release two large-scale hypergraph datasets.
  - We propose PCL (Partitioning-based Contrastive Learning), a scalable contrastive learning method for hypergraph neural networks (HNNs).

**Background: Hypergraph & Contrastive Learning**
- **Hypergraph:** A set of hyperedges that allow containing any number of nodes.
- **Examples:** Collaborations of researchers, joint interactions of proteins, co-purchases of items.
- **Properties:** A hypergraph naturally models group interactions.
- **Contrastive learning aims to maximize the agreement between differently augmented views of the same input.**
- **Previous research on hypergraph contrastive learning has a scalability issue when dealing with large-scale hypergraphs.**

**Proposed Tasks**
1. **Task1: Hyperedge Disambiguation**
   - **Setting:** every hyperedge is split into two hyperedges
   - **Goal:** To predict whether given two hyperedges are split from one or not
   - **Application:** researcher disambiguation, user identification
2. **Task2: Local Clustering**
   - **Setting:** Each node is involved in one (or more) clusters
   - **Goal:** To predict whether given two nodes belong to the same cluster or not
   - **Application:** sub-field detection, household matching
- **Advantages over formulating these tasks as an entity classification task:**
  1. Does not require the number of labels (split hyperedges or number of clusters).
  2. Does not need to retrain a model whenever the number of labels changes.

**Proposed Datasets**
- **Two 10M co-authorship hypergraph datasets:** MAG & AMiner

**Proposed Learning Method: PCL**
- **Challenge 1:** Large hypergraphs cannot be entirely loaded into GPU memory.
  - **Solution:** we use hypergraph partitioning to divide the large hypergraph into smaller partitions.
- **Challenge 2:** Information loss (e.g., split hyperedges) can be caused by partitioning.
  - **Solution:** we use contrastive learning and propose two additional techniques.
- **We propose PCL (Partitioning-based Contrastive Learning)**
  - We first divide input hypergraphs into several partitions.
  - Then, we train the HNN encoder via contrastive learning, by regarding each partition as a mini-batch of CL.
- **We propose two additional tools for PCL to mitigate information loss.**
  - **P-IOS:** Partitioning-technique that recovers lost topological information.
  - **PINS:** CL-technique that encourages the encoder to learn inter-partition dissimilarity.

**Proposed Datasets**
- **Two 10M co-authorship hypergraph datasets:** MAG & AMiner

**Experiments**
- **Accuracy of PCL on Task1 and Task2**

**Proposed Datasets**
- **Two 10M co-authorship hypergraph datasets:** MAG & AMiner

**Datasets, Tasks, and Training Methods for Large-scale Hypergraph Learning**

**Code and Data:** https://github.com/kswoo97/pcl