Overlapping Community Detection via Edge-Space Representation

Yikun Zhang, Fengjie Chen, and Ruojin He
Department of Statistics, University of Washington

Problem

Detecting overlapping communities in real-world networks has proved to be a challenging task. The Link Clustering (LC) method proposed by Ahn et al.[1] deals with this problem by **clustering the edges** in a network into communities and deriving the corresponding node communities from edge clusters. However, the edge similarity metric used in the original LC only considers a narrow scope of networks.

Our main goal is to ameliorate the edge similarity metric in LC so that information in a broader neighborhood of an edge pair can be explored and incorporated.

Data Description

Network	Nodes	Edges	C	ACS	CM
Amazon	334863	925872	75149	30.22	6.78
DBLP	317080	1049865	13477	53.40	2.27

C: number of communities, ACS: average community size, CM: community memberships per node.

- **Amazon product co-purchasing network**: *Nodes*: products; *Edges*: co-purchased product connections; *Communities*: product categories.
- **DBLP collaboration network**: *Nodes*: Authors; *Edges*: paper co-authorships; *Communities*: publication venues.

Method

***Edge2vec* Similarity Metric:**

Let $G = (V, E)$ be a given network and $f : V \to \mathbb{R}$ be the mapping function from nodes to node-space feature representions. The *edge2vec* representation for $e_{ij} \in E$ is defined to be

$$vec(e_{ij}) = f(i) \circ f(j),$$

where $\circ$ is a binary operator (*e.g.*, arithmetic mean). Then the *edge2vec* similarity metric between pair of adjacent edges e_{ik}, e_{jk} can be computed as

$$S(e_{ik}, e_{jk}) = \frac{vec(e_{ik})^T vec(e_{jk})}{||vec(e_{ik})||_2 \cdot ||vec(e_{jk})||_2},$$

and $S(e_{ij}, e_{kh}) = 0$ if two edges e_{ij}, e_{kh} share no common nodes.

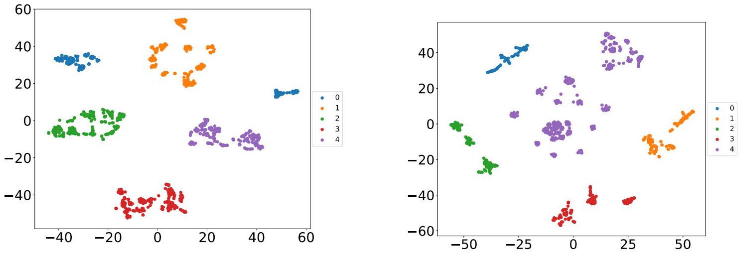
Evaluation Metrics: Let C^* be a set of ground-truth communities and $\hat{C}$ be a set of detected communities.

- **Average F1 Score:**[2]

$$\frac{1}{2}\left[\frac{1}{|C^*|}\sum_{C_i \in C^*} F1(C_i, \hat{C}_{g(i)}) + \frac{1}{|\hat{C}|}\sum_{\hat{C}_i \in \hat{C}} F1(C_{g'(i)}, \hat{C}_i)\right]$$

where g and g' are the best-matching defined as $g(i) = \arg\max_j F1(C_i, \hat{C}_j)$, $g'(i) = \arg\max_j F1(C_j, \hat{C}_i)$.

- **Omega Index:** $\frac{1}{|V|^2}\sum_{u,v \in V} \mathbf{1}\{|C_{uv}| = |\hat{C}_{uv}|\}$, where C_{uv} and $\hat{C}_{uv}$ are the sets of common ground-truth and detected communities for $u, v \in V$, respectively.

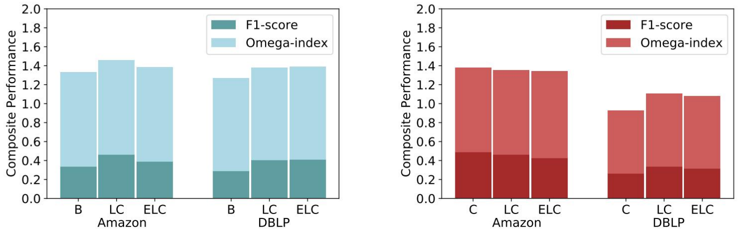


(a) Amazon

(b) DBLP

Figure 1: Visualizations of *edge2vec* representations on Amazon and DBLP Networks via t-SNE

Results



(a) Metrics on the whole network

(b) Average metrics on the sampled subnetworks

Figure 1: Composite Performances of Methods on Amazon and DBLP Network. B: BIGCLAM[2]; LC: Link Clustering; ELC: edge2vec Link Clustering; C: Clique percolation method (CPM).

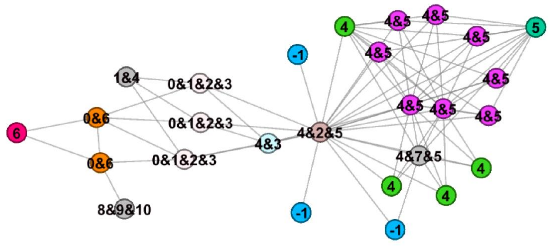


Figure 2: Partial Detected Communities via our *edge2vec* LC Method on the Amazon Network

Conclusion and Future Work

Conclusion: We propose a new framework that transforms task-independent feature learning for nodes into edge-space representations and apply it to overlapping community detection. The *edge2vec* Link Clustering method

- preserves the community structure in the original network after transformations
- is comparable to other state-of-the-art algorithms
- is scalable to large networks with millions of nodes and edges.

Future Work: The sementics of edges (*e.g.* biomedical information) can be considered during *edge2vec* feature representation learning.

References

[1] Y.-Y. Ahn, J. Bagrow, and S. Lehmann (2010). *Link Communities Reveal Multiscale Complexity in Networks.* Nature, 466(7307): 761-764.

[2] J. Yang and J. Leskovec (2015). *Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach.* In WSDM'13, pages 587-596. ACM.