Quality Metrics and Implementation Details

1 DETAILS OF COMMUNITY QUALITY EVALUATION METRICS

Since the vertices in the searched community are not directly connected, the existing quality evaluation metrics become inappropriate; thus, we propose an extension to overcome this challenge. Specifically, we assume that any two end vertices connected by a meta-structure instance are connected, and set the distance between them to one. We extend the following evaluation metrics to measure the quality of the searched community:

Closeness of Communities: The closeness of communities is measured using two widely accepted metrics [2], namely the **community diameter** and the **average path length**. The community diameter is calculated by finding the largest shortest distance between any pair of vertices in the community, while the average path length is calculated as the average of the shortest distance between all possible pairs of vertices. Although the former may be influenced by individual vertices, the latter provides a more accurate reflection of the distance between pairs of vertices. To redefine “the distance”, we set the length of a meta-structure instance to one. If the distance between two vertices is greater than one, it means that the meta-structure cannot connect them directly. Smaller values of these two metrics indicate higher overall connectedness of the community.

2. **Density of Connection:** The density of connection is defined as the number of edges over the number of vertices [4]. However, this may cause a problem that the density increases with the number of vertices, such as in complete graphs. Taking the complete graph as an example. According to the traditional definition, the density D should be $D = n \times (n-1)/2n = (n-1)/2$. The density increases with the number of vertices, but obviously the complete graph is already a very tightly connected community independent of the number of vertices. A good evaluation metric should minimise the impact of graph size. Therefore, we define the connection density as the average degree over the number of vertices, which can be used to reduce the effect of graph size.

3. **Clustering Coefficient:** Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together [1]. Graphs with higher clustering coefficients are found to have significant modular structures, and the average distance between different vertex pairs is smaller.

2 IMPLEMENTATION DETAILS

Evaluation Settings. In line with existing works about meta-structures [3], we focus on meta-structures with diameters at most four. We select meta-structures with more connected vertices as expert suggest, so as to ensure that our query is meaningful. Our dataset contains four vertex types that, coincidentally, constitute a meta-structure. To ensure the validity of our experiments, we randomly selected 20 vertices as the set of query vertices. The topic similarity threshold θ is set from 0.5 to 0.95, and the k is set from 1 to 12. In the results reported in the following, each data point is the average result for these queries. From the parameter analysis section in paper, We have analyzed that a larger k and θ mean denser topology. In order to identify intermediate-sized dense communities surrounding query vertices in the DBLP and ASN datasets, we determined appropriate parameter configurations to be $k = 9$ and $\theta = 0.60$, and 0.95, respectively (**the settings of table 3 in paper**).

For the parameters settings of **Table 4 in paper**, we utilized the same vertices set as query vertices and set k equal to 9 for k-core. For $k\mathcal{KP}$-core, we employed the exact keywords set that corresponded with the query vertices, while for SNCS, we used the topic vectors that had been extracted from these vertices. Topic constraints for $k\mathcal{KP}$-core and SNCS were established at 0.08 and 0.60, respectively. Setting topic constraints for $k\mathcal{KP}$-core that cover a certain proportion of the keyword set is challenging, so we increased θ starting from 0.01 in increments of 0.01 until the optimal community was achieved.

REFERENCES

[1] Holland, P.W., Leinhardt, S.: Transitivity in structural models of small groups. Comparative group studies **2**(2), 107–124 (1971)

[2] Huang, X., Lakshmanan, L.V.S., Yu, J.X., Cheng, H.: Approximate closest community search in networks. Proc. VLDB Endow. **9**(4), 276–287 (2015). https://doi.org/10.14778/2856318.2856323, http://www.vldb.org/pvldb/vol9/p276-huang.pdf

[3] Huang, Z., Zheng, Y., Cheng, R., Sun, Y., Mamoulis, N., Li, X.: Meta structure: Computing relevance in large heterogeneous information networks. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining. pp. 1595–1604 (2016)

[4] Wu, Y., Jin, R., Li, J., Zhang, X.: Robust local community detection: on free rider effect and its elimination. Proceedings of the VLDB Endowment **8**(7), 798–809 (2015)