*Journal of AI and Data Mining*
***Published online:***

# Graph Hybrid Summarization

N. Ashrafi Payaman and M. R. Kangavari*

*Computer Engineering, Iran University of Science & Technology, Narmak, Tehran, Iran.*

Received 06 April 2017; Revised 05 September 2017; Accepted 05 November 2017
**Corresponding author: kangavari@iust.ac.ir (M.Kangavari).*

**Abstract**
One solution for processing and analysis of massive graphs is summarization. Generating a high quality summary is the main challenge of graph summarization. For the aims of generating a summary with a better quality for a given attributed graph, both the structural and attribute-based similarities must be considered. There are two measures, density and entropy, are used to evaluate the quality of structural and attribute-based summaries, respectively. For an attributed graph, a high quality summary is the one that covers the structure and vertex attributes, of-course, with the user-specified degrees of importance. Recently, two methods have been proposed for summarizing/clustering a graph based upon both the structure and vertex attribute similarities. In this paper, a new method is proposed for the hybrid summarization of a given attributed graph, and the quality of the summary generated by the developed method is compared with the quality of summaries generated by the recently proposed method, SGVR, for this purpose. The experimental results showed that the proposed method generates a summary with a better quality.

**Keywords:** *Graph, Summarization, Super-Node, Super-Edge, Structural Similarity, Attribute-based Similarity.*

## 1. Introduction

Graphs are used in a variety of applications for modeling data and their relationships. Social networks, communication networks, web graphs, biological networks, and chemical compounds are examples of data modeled by graphs. These days, many applications generate large scale and massive graphs with billions of nodes and edges, and a lot of research works have been done on the theory and engineering of terra-scale graphs [1, 24]. In fact, we are faced with graphs that are very massive, and their growth rate is also increasing rapidly. For example, Facebook had 1.11 billion members on March 2013, while at the end of 2004, it had only about 1 million members [25].

Graph summarization has been proposed as a solution for processing massive graphs. Graph summarization algorithms [2-5], reduce a massive graph to a smaller one by removing its details but preserving its overall properties. In structural graphs, a dense sub-graph is replaced by a super-node in the summary graph and the edges between two dense sub-graphs are grouped to each other, indicating a super-edge in the summary graph. In attributed graphs, summary can be generated based on similarity of structure, attribute or both. Some other algorithms [9-11] have been proposed for this kind of summarization. Some recently proposed methods [12-18] summarize/cluster a graph based on the spectral summarization/clustering concept. Of-course, spectral-based methods are not very efficient for large-scale graphs. Community detection algorithms [19-22] are related and close to the summarization concept and they can be used in the summarization process.

Although, generating attribute-based summaries is not hard and some algorithms [2] have been proposed for this purpose, generating a summary based on both the graph structure and vertex attribute similarities (hybrid summarization) with the user-specified contributions of structure and attributes is not easy, and this is the main challenge of graph summarization. It is obvious that the importance of structure and attribute similarity in summary is not the same in all

applications, and therefore, considering variable weighting factors for them is more reasonable. Recently, two algorithms [6-7] have been proposed for hybrid summarization/clustering.

There are two measures called density and entropy to measure the quality of a summary. The quality of a hybrid summary is measured based on these two measures.

The rest of this paper is organized as what follows. In Section 2, graph hybrid summarization is reviewed and our proposed method is presented. The evaluation criteria and experimental results are given in Section 3. Finally, Section 4 concludes the paper.

## 2. Graph hybrid summarization

In this section, at first, we review some recently proposed hybrid summarization methods, and then our proposed method is presented for summarizing a graph based on both the structure and attribute similarities.

### 2.1. Recent methods

Recently, some algorithms [6-8], have been proposed for summarizing/clustering attributed graphs. Two of these methods summarize or cluster a graph based on both the structure and attribute similarities. These two methods are selected for review and demonstration. For the aim of evaluation, we compare the quality of summaries generated by our proposed method and with the SGVR method. Two selected methods are briefly reviewed in the following sub-sections.

#### 2.1.1. Random walk method

This method [7], clusters large attributed graphs based on a balance between the structural and attribute similarities. In this method, some new attributes, named attributed vertices, are added to the graph due to the existing common attribute values for vertices. In fact, for every two vertices that have the same value for an attribute, an attributed vertex is added to the graph and linked to both vertices by virtual links. Similarity of two vertices is measured based on the number of random shortest paths that exist between those two vertices. Existence of more paths between the two nodes $v_i$ and $v_j$ shows that they have more attribute values in common. Finally, the authors propose some optimization techniques on matrix computation for the aim of measuring the similarity between two vertices.

#### 2.1.2. SGVR method

This method [6], summarizes a graph by introducing real and virtual links to integrate structural and attribute similarities. Exact and similar links are defined for single and multi-valued attributes. At first, the graph is partitioned based on exact and similar links, respectively, and then adjust the resulting summary to the graph topology by moving nodes between super-nodes.

### 2.2. Proposed method

In our proposed method, a graph is summarized by merging similar nodes and repeating this trend to obtain a summary with the right size. We used the following formula (1) to compute the similarity of a pair of nodes:

$$sim(v_i, v_j) = \alpha \times sim_{st}(v_i, v_j) + (1-\alpha) \times sim_{si}(v_i, v_j) \quad (1)$$

where, $sim_{st}$ and $sim_{si}$ are the structural and attribute-based similarities, respectively, and $\alpha$ is the contribution of structure in the resulting summary. The value for $\alpha$ belongs to $[0,1]$. The structural similarity is computed by the following formula (2):

$$sim_{st}(v_i, v_j) = \begin{cases} 0 & if \;\; w[i][j] = 0 \\ 1 & if \;\; w[i][j] = 1 \end{cases} \quad (2)$$

where, $w$ is the adjacency matrix of the given graph. Attribute-based similarity of the two nodes $v_i$ and $v_j$ with $k$ attributes $a_1, a_2, ..., a_k$ and importance degrees of $c_1, c_2, ..., c_k$ is calculated by the following formula (3):

$$sim_{si}(v_i, v_j) = \sum_{h=1}^{k} c_h \times sim_{si}(v_i, v_j, a_h),$$
$$s.t.\ 0 \le c_h \le 1 \ and \sum_{h=1}^{k} c_h = 1, \quad (3)$$

In (3), $c_h$ is the importance degree of the attribute $a_h$ that is given by the user.

Attributes may be single or multi-valued. A single-valued attribute has only one value, while a multi-valued attribute can have more than one value. For example, a node in the Facebook social networks represents an individual that has attributes such as '**gender**' and '**spoken languages**', where the former is a single-valued attribute (Male or Female, only one of these values), and the latter is a multi-valued attribute

(English, Spanish, …, more than one is possible). The similarity of the two vertices based on the given attribute $a_h$ is calculated using the following formula (4):

$$sim_{si}(v_i, v_j, a_h) = \begin{cases} 0 & a_h : single-valued \ and \ \text{val}(v_i, a_h) \neq \text{val}(v_j, a_h) \\ 1 & a_h : single-valued \ and \ \text{val}(v_i, a_h) \neq \text{val}(v_j, a_h) \\ \dfrac{|\text{vals}(v_i, a_h) \cap \text{vals}(v_j, a_h)|}{|\text{vals}(v_i, a_h) \cup \text{vals}(v_j, a_h)|} & a_h : multi-valued \end{cases}, \quad (4)$$

where, $sim_{si}(v_i, v_j, a_h)$ is the similarity of the two vertices $v_i$ and $v_j$ based on attribute $a_h$. The value of $\text{val}(v_i, a_h)$ represents the value of attribute $a_h$ on $v_i$ vertex. According to (4), the attribute-based similarity of two vertices for a multi-valued attribute is computed based on the Jaccard similarity. For example, let $a_h$ be a multi-valued attribute and the values of this attribute on nodes $v_i$ and $v_j$ be $\{a,b\}$ and $\{b,c\}$, respectively. Then $\text{val}(v_i, a_h) = \{a,b\}$ and $\text{val}(v_j, a_h) = \{b,c\}$ and the similarity of these two nodes on attribute $a_h$ is $\frac{|\{a,b\} \cap \{b,c\}|}{|\{a,b\} \cup \{b,c\}|} = \frac{|\{a\}|}{|\{a,b,c\}|} = \frac{1}{3}$.

The similarity of super-node $V_i$ and node $v_j$ with $k'$ nodes is calculated as follows:

$$sim(V_i, v_j) = \alpha \times sim_{st}(V_i, v_j) + (1-\alpha) \times sim_{si}(V_i, v_j) \quad (5)$$

where, $sim_{st}(V_i, v_j)$ and $sim_{si}(V_i, v_j)$ are the structural and attribute-based similarities of super-node $V_i$ and node $v_j$ calculated by the formulas (6) and (7), respectively.

$$sim_{st}(V_i, v_j) = \frac{|\{u \mid u \in V_i \ and \ (u, v_j) \in E\}|}{|V_i|}, \quad (6)$$

$$sim_{si}(V_i, v_j) = \frac{1}{|V_i|} \sum_{h=1}^{k} c_h \times sim_{si}(V_i, v_j, a_h), \quad (7)$$

where, in (7), $sim_{si}(V_i, v_j, a_h)$ is the attribute-based similarity of super-node $V_i$ and node $v_j$ based on the given attribute $a_h$, and is calculated by (8).

$$sim_{st}(V_i, v_j, a_h) = \frac{|\{u \mid u \in V_i \ and \ val(u, a_h) = val(v_j, a_h)\}|}{|V_i|}, \quad (8)$$

The similarity of the two super-nodes $V_p$ and $V_q$ is calculated by the following formula (9):

$$sim_{st}(V_p, V_q) = \frac{1}{|V_q|} \sum_{i=1}^{|V_q|} sim(V_p, v) \mid v \in V_q, \quad (9)$$

Based on (1) to (9), our proposed method for summarizing a graph is given in Algorithm 1.

**Algorithm 1:** Summarization( $G, k, A, \alpha, C$ )

**Input:** graph $G$ : graph, $k$ : the right size of the summary, $A$ :user interested attributes, $\alpha$ : structure contribution, $C$ : importance degrees of attributes;

**Output:** $S$ **:** the resulting summary;

1. Calculate the similarity of every pair of vertices;
2. Consider every vertex as a super-node;
3. num =the number of super-nodes;
4. while ( $num > k$ )
5. {
6. Select the pair of vertices or super-nodes with the maximum similarity;
7. Merge two selected vertices or super-nodes;
8. Re_calculate the similarity of vertices;
9. }

## 3. Evaluation

The quality of a hybrid summary is measured based upon density and entropy. For a high quality summary, density is high but entropy is low. In our proposed method, the quality of summary is used as a stopping measure.

### 3.1. Measures

The quality of a summary graph is measured based upon density or entropy, depending on being a structural or an attribute-based summary. The formal definitions of these measures are given by (10) and (11). The quality of a hybrid

summary is measured based on both of these two measures.

**Density:** The density of a summary graph with $k'$ super-nodes is calculated by the following formula (10):

$$density(\{V_i\}_{i=1}^{k'}) = \sum_{i=1}^{k'} \frac{|\{(v_p, v_q)\} | v_p, v_q \in V_i \text{ and } (v_p, v_q) \in E|}{|E|}, \quad (10)$$

where, $E$ is the edges of the graph.

**Entropy:** The entropy of a summary graph with $k'$ super-nodes and $m$ vertex attributes is calculated by the following formula (11):

$$entropy(\{V_i\}_{i=1}^{k'}) = \sum_{i=1}^{m} \frac{w_i}{\sum_{p=1}^{m} w_p} \sum_{j=1}^{k'} \frac{|V_j|}{|V|} entropy(a_i, V_j), \quad (11)$$

where:

$$entropy(a_i, V_j) = -\sum_{n=1}^{n_i} p_{ijn} \log_2^{(p_{ijn})}$$

and $p_{ijn}$ is the percentage of vertices in super-node $V_j$ that has the value $a_{in}$ on attribute $a_i$.

### 3.2. Time complexity

In this method, at first, the weight of every edge of the augmented graph is calculated, and after that, the summary is generated by merging nodes or super-nodes by each other. In the worst case, our proposed method requires at most $|V|$ merging operations to obtain the expected summary. Henceforth, the time complexity of this method is $O(|E| \times |V|)$. Of-course, we can reduce the runtime of the algorithm by removing the isolated and less similar vertices in the initial steps of the algorithm. As this time complexity shows, our proposed method is efficient in comparison with other recently proposed methods such as the random walk and SGVR methods.

### 3.3 Demonstration by example

We will illustrate the proposed method by an example, as shown in figure 1. The given graph is shown in figure 1(a). In this graph, every node that represents a person, has one attribute, **spoken_languages**, a multi-valued attribute. For a node, this attribute indicates languages in which that person can speak. In Figure 1(a), the letters (E, G, P, and S) after the label of the node (person) indicate the languages that the person can speak. The letters E, G, P, and S stand for English, Germany, Persian and Spanish, respectively. The augmented graph is depicted in Figure 1(b), where the real edges are shown by solid lines and the virtual edges by dash lines. In the augmented graph, the most similar pair of nodes to merge is $(v_3, v_4)$ with the weight of 1. By merging these two nodes, the summary graph has 4 nodes, and the weight of edges is calculated again, as shown in figure 1(c). In this summary, two nodes $v_2$ and V, a super-node including $v_3$ and $v_4$, are merged and the resulting graph is shown in figure 1(d).

### 3.4. Dataset

In order to evaluate our proposed method, we generated a synthetic attributed graph. The synthetic graph was summarized by our proposed method and the SGVR method. The qualities of summaries are measured and compared with each other.

#### 3.4.1. Synthetic dataset

We generated a graph with 1,000 nodes and 2,500 edges based on the R-Mat [23] method. Firstly, graph vertices and edges were generated, and then values were assigned to the vertex attributes. Details of attributes are given in table 1. Values were assigned to attributes based on the discovered statistics about the attributed graphs of the context of interest.

**Table 1. Details of vertex attributes of generated graph.**

| Row | Attribute | Single-valued | Multi-valued |
|---|---|---|---|
| 1 | Age | √ | |
| 2 | Education | √ | |
| 3 | Gender | √ | |
| 4 | Country | √ | |
| 5 | Languages | | √ |

### 3.5. Results and discussions

We implemented our proposed method and the SGVR method in python to measure the quality of the generated summaries, and compared these two methods. The values for the density, entropy, and $\frac{density}{entropy}$ measures are shown in table 2.

**Table 2. Comparison of our method and SGVR method.**

| Method | Density | Entropy | Density/Entropy |
|---|---|---|---|
| SGVR | 0.0669 | 0.6829 | 0.0979 |
| Our Method | 0.2572 | 0.38369 | 0.6704 |

As shown in table 2, our proposed method generated a summary with higher density and lower entropy values in comparison with the

SGVR method. This means that our proposed method generates a summary with a cohesive structure and homogenous attribute values. Henceforth, our proposed method, groups closely connected vertices into one super-node.

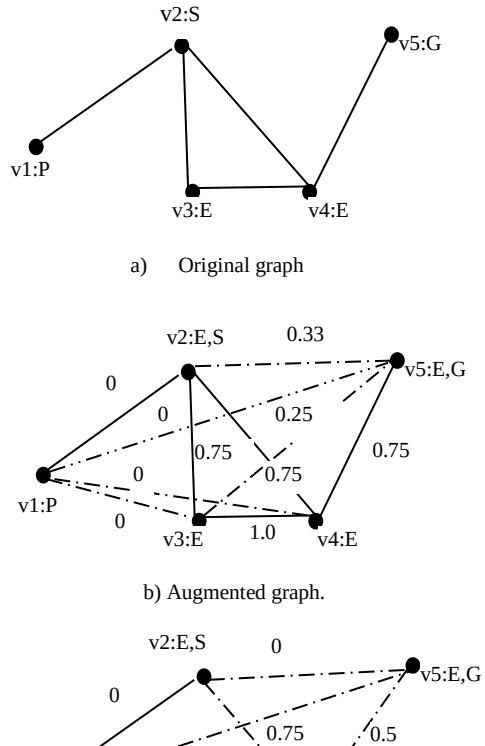


**Figure 1. Proposed method demonstration by example. Main steps of method are depicted by Figures in 1(b), 1(c), and 1(d) for the graph of Figure 1(a).**

## 4. Conclusions

A new method has been proposed for summarizing an attributed graph based on both the structural and attribute similarities. Hybrid summarization aims to generate a summary with a cohesive structure and homogenous attribute values.

Our proposed method works by merging a pair of similar nodes or super-nodes. A similar function was defined to measure the similarity of two nodes based on both the structure and attribute values. We extended the definition of similarity function to measure the similarity of a node and a super-node or two super-nodes.

We implemented our proposed method and the SGVR method in python in order to evaluate the quality of summary generated by our proposed method. The Experimental results showed that our proposed method results in a better summary based on the density and entropy criteria.

## References

[1] Kang, U (2012). Mining Terra-Scale Graphs : Theory , Engineering and Discoveries Mining Terra-Scale Graphs : Theory , Engineering and, *Ph.D Thesis*, pp. 1–179.

[2] Tian, Y., Hankins, R. A. & Patel, J. M., (*2008*). Efficient Aggregation for Graph Summarization, ACM SIGMOD International Conference on Management of Data, Vancouver, Canada, pp. 567–580, 2008.

[3] Navlakha, S., Rastogi, R. & Shrivastava N., (2008). Graph Summarization with Bounded Error, ACM SIGMOD International Conference on Management of Data, Vancouver, Canada, pp. 419-432, 2008.

[4] LeFevre, K. & Terzi E., (2010). GraSS: Graph Structure Summarization, 2010 SIAM International Conference on Data Mining, Philadelphia, USA, pp. 454-465, 2010.

[5] Zhang, N., Tian, Y. & Patel J. M., (2010). Discovery-Driven Graph Summarization, 2010 IEEE 26th International Conference on Data Engineering, Long Beach, CA, USA, pp. 880-891, 2010.

[6] Bei, Y., Lin, Z. & Chen D., (2016). Summarizing Scale-free Networks based on Virtual and Real Links, Phys. A Stat. Mech. its Appl., vol. 444, no. 2, pp. 360–372.

[7] Cheng, H., Zhou, Y. & Yu J. X., (2011). Clustering Large Attributed Graphs: A Balance between Structural and Attribute Similarities, ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 5, no. 2, article no. 12.

[8] Wu, Y., Zhong, Z., Xiong, W. & Jing N., (2014). Graph Summarization for Attributed Graphs, International Conference on Information Science, Electronics and Electrical Engineering (ISEEE 2014), Sapporo, Japan, vol. 1, pp. 503–507, 2014.

[9] Riondato, M., Garcia-Soriano, D. & Bonchi F., (2015). Graph Summarization with Quality Guarantees, IEEE International Conference on Data Mining, pp. 947–952. Atlantic, USA, 2015.

[10] Liu, X., Tian, Y., He, Q., Lee, W. C. & McPherson J., (2014). Distributed Graph Summarization, 23rd ACM International Conference on Conference on Information and Knowledge

Management, CIKM ’14, Shanghai, China, pp. 799–808, 2014.

[11] Chen, C. & Lin C., (2009). Mining Graph Patterns Efficiently via Randomized Summaries, VLDB Endowment, vol. 2, no. 1, pp. 742–753, 2009.

[12] Dhillon, I., Guan, Y. & Kulis B. (2005). A Unified View of Kernel k-Means, Spectral Clustering and Graph Cuts, Comput. Complex., vol. 25, no. 5, pp. 1–20.

[13] Planck, M. & Von Luxburg U. (2007). A Tutorial on Spectral Clustering, Statistics and Computing, vol. 17, no. 4, pp. 395–416.

[14] Uw, S., Ng, A. Y., Jordan, M. I. & Weiss Y. (2001). On Spectral Clustering: Analysis and an Algorithm, In Advances in neural information processing systems, pp. 849–856.

[15] Liu, J, Wang, C, Danilevsky, M & Han J. (2013). Large-scale spectral clustering on graphs, 23rd International Joint Conference on Artificial Intelligence, Beijing, China, pp. 1486–1492, 2013.

[16] Auffarth, B., (2007). Spectral Graph Clustering, Universitat de Barcelona, course report for Technicas Avanzadas de Aprendizaj, at Universitat Politecnica de Catalunya, 2007.

[17] Zhou, D. & Burges, C. J. C. (2007). Spectral Clustering and Transductive Learning with Multiple Views, 24th international conference on Machine Learning, Corvallis, USA, pp. 1159–1166, 2007.

[18] Smyth, S. & White S. (2005) A Spectral Clustering Approach to Finding Communities in Graphs, 5th SIAM International Conference on Data Mining, Newport Beach, CA, pp. 76–84, 2005.

[19] Wang, C. D., Lai, J. H. & Yu, P. S. (2013). Dynamic Community Detection in Weighted Graph Streams, 2013 SIAM International Conference on Data Mining, Austin, Texas, USA, pp. 151–161, 2013.

[20] Xie, J., Kelley, S. & Szymanski B. K. (2013). Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study, ACM Computing Surveys, vol. 45, no. 4, pp. 43:1–43:35.

[21] Lancichinetti, A. & Fortunato S. (2009). Community Detection Algorithms: A Comparative Analysis, Physical Review E, vol. 80, no. 5, pp. 1–12.

[22] Wang, W. & Street W. N. (2014). A novel Algorithm for Community Detection and Influence Ranking in Social Networks, International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Beijing, China, pp. 555-560, 2014.

[23] Chakrabarti, D., Zhan, Y. & Faloutsos C. (2004). R-MAT: A Recursive Model for Graph Mining, 2004 SIAM International Conference on Data Mining, Florida, USA, pp. 442-446, 2004.

[24] Aghaei, M. & Dastfan A., (2015). A Graph Search Algorithm: Optimal Placement of Passive Harmonic Filters in a Power System, AI and Data Mining, vol. 3, no. 2, pp. 217-224.

[25] The Yahoo website (2017), http://news.yahoo.com/number-active-users-facebook-over-230449748.html.

نشریه هوش مصنوعی و داده کاوی

# خلاصه‌سازی هیبریدی گراف

**نصرت‌علی اشرفی پیامن و محمدرضا کنگاوری**[×]

**گروه آموزشی مهندسی نرم‌افزار، دانشکده کامپیوتر، دانشگاه علم و صنعت، تهران، ایران.**

ارسال ۲۰۱۷/۰۴/۰۶؛ بازنگری ۲۰۱۷/۰۹/۰۵؛ پذیرش ۲۰۱۷/۱۱/۰۵

**چکیده:**

یک روش برای پردازش و تحلیل گراف‌های حجیم، خلاصه‌سازی می‌باشد. چالش اصلی خلاصه‌سازی گراف، تولید خلاصه‌ای با کیفیّت بالاتر، می‌باشد. به منظور تولید خلاصه‌ای با کیفیّت بالاتر برای یک گراف دارای ویژگی، هر دوی شباهت‌های ساختاری و مبتنی بر ویژگی رئوس باید در محاسبه شباهت رئوس در نظر گرفته شوند. برای ارزیابی خلاصه‌های ساختاری و مبتنی بر ویژگی به ترتیب معیارهای دانسیته و انتروپی وجود دارند. برای یک گراف دارای ویژگی، خلاصه‌ای مطلوب است که هردوی ساختار و ویژگی رئوس را، البته با درجات اهمیّت مشخص، پوشش دهد. اخیراً دو روش برای خلاصه‌سازی/خوشه‌بندی یک گراف دارای ویژگی بر حسب هردوی ساختار و ویژگی رئوس پیشنهاد شده است. در این مقاله، یک روش جدید برای خلاصه‌سازی هیبریدی ارائه می‌شود که خلاصه‌هایی با کیفیّت بالاتر در مقایسه با روش‌های موجود مثل SGVR تولید می‌کند. نتایج تجربی نشان می‌دهد که که کیفیت خلاصه‌های تولید شده با روش پیشنهادی، بالاتر است.

**کلمات کلیدی:** گراف، خلاصه‌سازی گراف، ابررأس، ابریال، شباهت ساختاری، شباهت مبتنی بر ویژگی.