Communications in Information and Systems
Volume 22, Number 1, 147–167, 2022

G-Explorer: a visual exploration and analysis method of subgraph in complex network based on graph embedding

QingHui Zhang, Yi Chen*, MengLu Zhang, and Zeli Guan

As the scale and complexity of networks increase, the mining and analysis of subgraphs in complex network faces great challenges. Graph embedding technique can transform the high-dimensional massive subgraph data into low-dimensional computable embedding vectors while preserving the structural features of subgraphs. In this paper, a graph embedding-based visual exploration and analysis method, G-Explorer, is proposed to help users analyze selected subgraphs from multiple perspectives. The method first constructs a vector representation of the nodes using struc2vec, and then clusters the vectorized nodes to form subgraphs using K-Means. A new visualization design, Force-Radar, is also proposed, in which radar plots reflecting the characteristics of subgraphs are used as nodes and the associations between subgraphs are used as edges to form an overview of the whole network. Detailed information about the selected subgraph is presented in multiple auxiliary views through interaction. The method was applied to analyze the association network of food inspection data for exploring the high-risk foods and hazards. The experimental results illustrate the effectiveness of the G-Explorer method.

1. Introduction

Most entities in the real world are not independent of each other, and they are constructed in different ways into different graphs also known as networks, such as social network, literature mutual citation network, protein network, etc. There are different patterns in different networks [1]. However,

*Research supported is part by National Natural Science Foundation of China (61972010), National key R&D program of China (2018YFC1603602) and Basic Research Project of the Ministry of Science and Technology (2015FY111200).

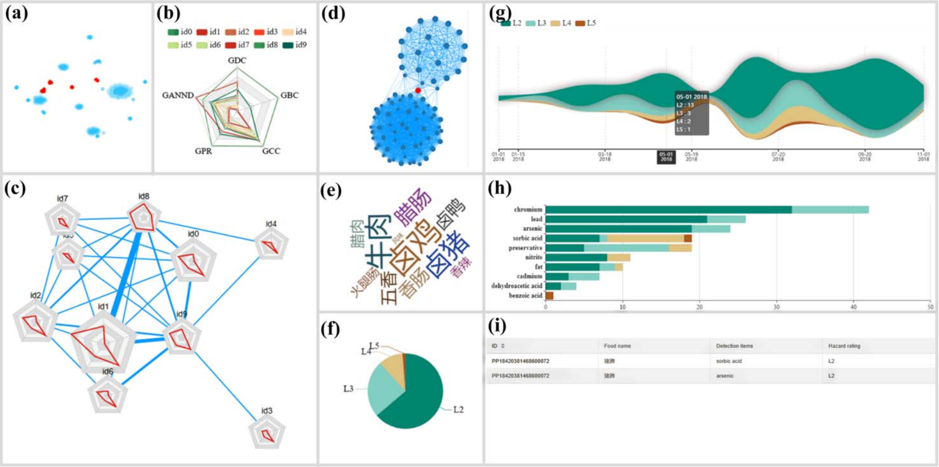
due to the complex relationship between nodes in the network, the exploration of the network needs to consume a lot of computing resources and storage resources. The graph embedding [2] technology tackles this challenge well, it captures the relationship between nodes by the distances between nodes in the vector space, and the topological and structural characteristics of a node are encoded into its embedding vector, so that the approach of machine learning can be applied to graph analysis.

The purpose of mining the different patterns in the graph is to find the interesting subgraphs in the network. By discovering subgraph patterns in different networks, it can effectively help people explore various events. For example, the frequent subgraph in the protein network may be a functional component of the protein. By exploring the frequent subgraph in the protein network, it can provide a basis for predicting the function of unknown proteins. Using visualization and visual analysis technology to explore complex networks can make users understand data better.

Excessive residues of food hazards, such as pesticide residues and heavy metal residues, threaten people health. In order to prevent unqualified food entering the market, the regulatory authority will conduct sampling inspection on the food regularly to determine whether there are excessive residues. They can find out the high-risk foods and high-risk hazards for supporting early warning of food safety. However, due to the variety and complexity of food safety data, this analysis task is extremely difficult. In addition, the existing exploration methods for high-risk foods and high-risk hazards are to compare detection results with the maximum limit of hazards, without considering the correlation between foods, so as to find potentially risky foods and hazardous substances.

Visualization techniques have been widely used in food safety analysis [3] and anomaly detection [4]. In this paper, we propose a new visual method, Force-Rader, which helps to explore the structural features of subgraphs and the associations between subgraphs. And we construct a food association network with food as nodes and whether the same hazards are detected in two foods as edges. The high-dimensional food network data is vectorized by graph embedding technique and the embedded vectors are clustered into subgraphs of different food products. We also designed a multi-view collaborative visual analysis method, G-Explorer, in order to get a better exploration of food network subgraphs. The system's multiple views, including the views of Force-Radar, ThemeRiver, Wordle, etc, to help researchers explore the structural features of food network subgraphs and mine the food network for high-risk hazards and their characteristic through practical interactions. In summary, the main contributions of this paper are as follows:

(1) A new visual design is proposed, called Force-Radar, which shows the network metrics and the relationship between each subgraph.
(2) A food network is constructed, researchers may identify associations between food and commonly detected hazards.
(3) A visualization system, G-Explorer, is designed and implemented. It can analyze networks by visualizing and exploring the structural features of subgraph. The system is applied to the food network we constructed and found that G-Explorer can uncover potential high-risk foods and their hazards, verifying G-Explorer's effectiveness and usability.



ID ⇕	Food name	Detection items	Hazard rating
PP18420381468600072	猪蹄	sorbic acid	L2
PP18420381468600072	猪蹄	arsenic	L2

Figure 1: The visualization illustrates an example of using our proposed G-Explorer to explore subgraph. (a) a view to show the result of node embedding; (b) a radar map for comparing the network metrics of each subgraph; (c) an overview of the entire structure of the network, called Force-Radar; (d) a node-link view to visualize the subgraphs structure; (e) a wordle view to show the key words of food names; (f) a pie chart is used to show the percentage of different level of hazards in foods; (g) a theme river view to show the detection frequency and the level of hazards in foods at different times; (h) a bar chart to show the food frequency of hazards in different hazardous level; and (i) the table is used to display detailed information about food data.

2. Related work

In this paper, we propose a graph exploration approach based on graph embedding. Thus, we review research work in the areas of and graph exploration and graph embedding.

2.1. Graph embedding

Graph embedding is one of the approaches of graph representation, which represent the nodes in the network (graph) as vectors. According to the different types of graphs, it can be divided into the representation approach of homogeneous graph and heterogeneous graph.

Inspired by word2vec [5], graph embedding algorithm Deepwalk [6] was proposed in 2014. This algorithm first uses a random walk to randomly select the starting point in the graph, generates a random sequence as a training sample, then inputs the random sequence into the SkipGram model [5] for training, and finally outputs the nodes representation in the form of vectors. Since then, many improved algorithms based on Deepwalk have been proposed. Node2vec [7] realized depth-first and breadth-first random walk for graph training sample collection. Struc2vec [8] captures the structural role proximity of nodes according to the neighbors information of different distances. Tang et al. [9] proposed a graph embedding representation approach considering first-order proximity and second-order proximity. In addition to the vectorized representation of nodes, Narayanan et al. also realized the vectorized representation of subgraphs and the entire graph [10, 11]. All above are graph embedding approaches for homogeneous graph.

In 2017, Dong et al. [12] first proposed the graph embedding models of heterogeneous graph, metapath2vce and metapath2vec++, which respectively proposed a meta-path-based random walk considering node and link types, and a SkipGram model considering node and link types. In the same year, the HIN2Vec proposed by Fu et al. [13] carried out vector representation of nodes and edges in the graph respectively. Shi et al. [14] also applied the heterogeneous graph embedding approach to the recommendation system.

2.2. Graph exploration

As an important tool for information transmission, graph can be explored in two ways: top-down and bottom-up.

Top-down techniques allow users to have an overview of data and then explore the subgraphs of interest. It can be realized by graph filtering [15], node clustering [16], edge bunding [17] and other technologies [18]. For example, Zinsmaier et al. [19] simplified the entire network into a heat map based on the density of network nodes, and visualized the network contour. Yoghourdjian et al. [20] proposed an approach called Graph Thumbnails, which visualizes the hierarchical structure of Graph data in the form of thumbnails.

Bottom-up graph exploration techniques are goal oriented, such as nodes and subgraphs. Pienta et al. [21] proposed VIGOR to explore subgraphs constructed by authors, papers and conferences in different fields. By querying, the qualified subgraphs are mapped to a two-dimensional plane in the form of a scatter graph. Graphicle [22] used the technique of combining matrix and node-link graph to realize a unit visualization approach that supports large-scale, multivariate relational data, and considered both attribute and structure exploration. In addition, there are graph exploration tools, which combines the two technologies, visualizes the explored subgraphs in the overview. Yan et al. [23] proposed a visual exploration system which provides some frequent subgraphs extracted from the whole graph, and the relationship between them. Chen et al. [24] proposed a structure-based suggestive exploration approach to support effective exploration of large networks by suggesting appropriate structures upon user request. Our approach is similar to Chen's, and his work has well verified the advantage of using graph embedding for graph exploration. However, his goal is to mine subgraphs with specific structures from large graphs, while our analysis goal is to cluster food nodes with the proximity of structured roles together to form a subgraph, and analyze the differences among different subgraphs.

3. Overview of G-Explorer

The main purpose of this paper is to explore the different subgraph structures and their relationships in the network and to further analyze the node attributes in the subgraphs of different structures. We first extended five network metrics to measure the structure of the subgraphs. Then we used network metrics to represent the structure of different subgraphs and visualize their relationships. Finally, the node attributes in subgraphs with different structures were analysed.

As shown in Fig. 2, our workflow can be summarized in the following three steps: (1) The network nodes were vectorized by graph embedding technology, in which the vectorized nodes should preserve the structural role proximity of nodes. We chose Struc2vec [8] to vectorize network nodes; (2) We used t-SNE [25] to reduce the dimensionality of vectorized nodes and visualized them as a scatter plot to provide support for setting the number of subgraphs K-means [26] was adopted to cluster the vectorized nodes into different subgraph sets. A novel visual design called Force-Radar was proposed, which uses a node-link layout to connect the radar maps, where the radar maps show features about the structure of subgraphs. The structural features of each subgraph were drawn into a separate radar map to compare the differences of the structural features of each subgraph; (3) Then we used some auxiliary views which can help users to explore the details of the subgraph.

According to the above workflow, the structure features of each subgraph, the attributes of network nodes and the relationship between subgraphs can be analyzed effectively.

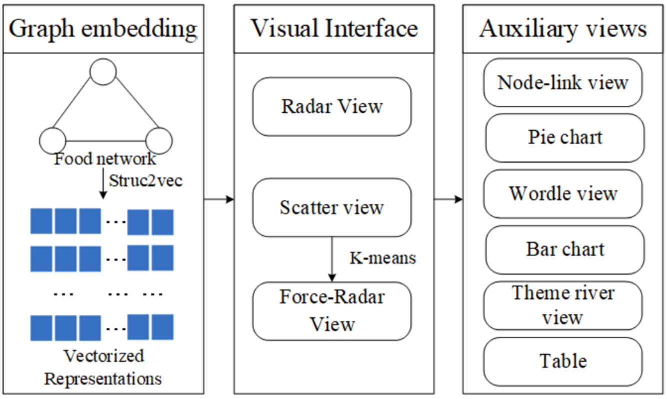


Figure 2: An overview of G-Explorer. First, the network nodes are vectorized. Then, the nodes are clustered into different subgraphs, and the structure features of the subgraphs are displayed. Finally, some auxiliary views are used to explore the details of the subgraphs.

4. Network structure explorer

There are different structures in different types of networks. The importance of nodes varies when we handle the different problems [27]. In order to explore the structure of network, we selected five node metrics (the ranking metrics of nodes' importance) and extended them to measures the network structure which we called network metrics.

In order to introduce the five node metrics clearly, we define the graph as $G = (V, E)$ where $V = (v_1, \ldots, v_n)$ represents the set of nodes, $E = (e_1, \ldots, e_m)$ represents the set of edges, n and m represent the number of nodes and edges respectively.

Definition 1 (Degree centrality)**.** Degree centrality [28] reflects the importance of the node from the number of neighbors. That can be defined as (1)

$$DC(i) = \frac{k_i}{n-1} \tag{1}$$

where n is the number of nodes in the network, and k_i is the number of neighbor points of node v_i.

Definition 2 (Closeness centrality)**.** Closeness centrality [29] is the average distance between the current node and other nodes. The greater the average distance is, the more difficult the node is to spread information in the entire network, and vice versa. That can be defined as (2)

$$CC(i) = \frac{n-1}{\sum_{i \neq j} d_{ij}} \tag{2}$$

where n represents the number of nodes, and d_{ij} represents the distance between node v_i and node v_j.

Definition 3 (Betweenness centrality)**.** Betweenness centrality [30] is determined by the number of shortest paths through a node. It reflects the importance of the node as a medium in the entire process of network transmission. That can be defined as (3)

$$BC(i) = \sum_{i \neq s, i \neq t, s \neq t} \frac{g_{st}^i}{g_{st}} \tag{3}$$

where g_{st} is the number of all shortest paths from node v_s to node v_t, and g_{st}^i is the number of shortest paths through node v_i in the shortest paths from node v_s to node v_t.

Definition 4 (Average nearest-neighbor degree)**.** Average Nearest-Neighbor Degree Average nearestneighbor degree [31] is mainly used to measure the correlation between nodes and can be defined as (4)

$$ANND(i) = \frac{i}{d_i} \sum_{j=1}^{N} a_{ij} d_i \tag{4}$$

where d_i and d_j represent the degree of node i and node j respectively, N represents the total number of neighbor nodes of node i, and a_{ij} represents the weight between the two nodes.

Definition 5 (PageRank)**.** PageRank [32] algorithm was proposed to solve the challenge of web page ranking, and was later applied to many fields, such as food safety [33]. That can be defined as (5)

$$PR_i(t) = \sum_{j=1}^{n} a_{ij} \frac{PR_i(t-1)}{k_j^{\text{out}}} \tag{5}$$

where k_i^{out} is the out-degree of node v_i, and a_{ij} represents the elements of the i row and the j column in the adjacency matrix of the graph. Different from the evaluation approach of node importance, this paper mainly considers different subgraphs and their relationships in complex networks, so we extend the measurement of node to network. We defined five network metrics to measure the structure of subgraphs in the food network, named GDC, GCC, GBC, GPR, and $GANND$, where extend from DC 1, CC 2, BC 3, PR 5, and $ANND$ 4 respectively. GDC can be defined as (6)

$$GDC\left(g_i\right) = \sum_{i=1}^{n} \frac{DC(i)}{N} \tag{6}$$

where g_i represents a subgraph, n represents the total number of nodes in the subgraph, N represents the number of nodes in the entire graph and $DC(i)$ the degree centrality of node v_i. It reflects the average degree centrality of the nodes in the subgraph.

The definition of the other four network metrics follows this principle. We calculated the five metrics for all the nodes in the network, and calculated the average after summation respectively, that is what we defined network metrics.

The reason why we chose these five metrics and expanded them is that after our analysis, these five metrics are more suitable for analyzing the food network. Other different network metrics can also be selected for networks with different features.

5. Visual exploration

This paper proposed a general approach G-Explorer for exploring subgraphs in network and their relationships, as shown in Fig. 1(a–d). To explore the food network, we also added a series of auxiliary views as shown in Fig. 1(e–i). This section discusses only the visual design of G-Explorer.

5.1. Embedding view

We visualized the embedding results to provide support for setting a reasonable K (the number of subgraphs).

In order to mine the subgraphs with different structures in the graph, this study adopts the Struc2vec [8] that can capture the structural role proximity of nodes to vectorize the nodes in the network. In order to intuitively quantity the node relationship, we introduce a reasonable K value which defines the number of subgraphs (radar map) for better display. We used t-SNE [25] which is a simple and effective approach to reduce the dimension of the embedding results to two dimensions. Then we displayed it in the scatter graph as shown in the Fig. 1(a) and the red cluster can be considered as a subgraph.

Although there is information loss in the process of dimension reduction, it can still provide guidance for setting the number of subgraphs.

5.2. Radar view

In order to show the differences of the network metrics in each subgraph, we used radar map to visualize the differences among subgraphs.

By distinguishing the proximities and differences between different subgraphs, users can find the different subgraphs they want to compare. As shown in Fig. 1(b), there are five dimensions on the radar map: GDC, GCC, GBC, GPR, and $GANND$. By reasonable color collocation [34], the structure features of different subgraphs can be displayed, and the network metrics of interested subgraphs can be compared in detail by legend filtering. The advantage of using radar map is that it can comprehensively analyze multiple metrics and has the advantages of integrity, clarity and intuition.

5.3. Force-radar view

In order to give users an overview of the network, a novel visual design called Force-Radar was designed in this study.

Struc2vec [8] was used to vectorize nodes in the network, and nodes were clustered into K subgraphs according to their similar structural roles. Considering that our users are non-computer professionals, and the main parameter that K-means needs to adjust is just the number of clusters K that is easy to operate. Moreover k-means has the advantages of simple

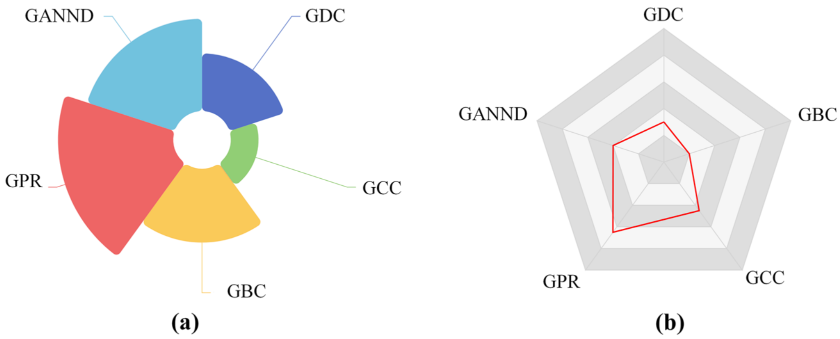


Figure 3: Two structural glyph design alternatives representing subgraphs.

implementation, high time efficiency, fast convergence, etc. Therefore, under the guidance of embedding view (Fig. 1(a)), the nodes were clustered into K sets through K-means. The network metrics of each subgraph are shown by radar map, and the relation between each subgraph is shown by the links between radar map.

Each subgraph was abstracted into a radar map. If there are more nodes in the subgraph, the radar map is larger. If the nodes in the two subgraphs are linked, an edge is connected between the two radar maps. The width of links between radar maps is determined by the number of links between two subgraphs. Then the force-directed model is used to layout each radar map. As in the previous section, each radar map has five dimensions as shown in Fig. 1(c), and it also shows the number of nodes. This design can avoid the excessive overlap of the radar map, and can visualize the relationship between different subgraphs.

We also considered other two glyph design alternatives in the design process (Fig. 3). The first design (Fig. 3(a)) used radial bar charts with different colors to encode the size of different network metrics. However, it is difficult to compare the same networks metrics of different subgraphs. The second design (Fig. 3(b)) used a radar map to encode different network metrics, but the shaded part of the radar map also makes the entire visualization look unclean.

5.4. Node-link view

To show the structure of the subgraphs, we used the node-link graph to display the subgraphs in detail, as shown in the Fig. 1(d). The more edges a node connects, the bigger it is. In addition, the information represented by the node can be further displayed by selecting the node of interest.

6. Case study

In order to evaluate the effectiveness of G-Explorer, we built food network and applied our approach to food network analysis and answered two real-world questions raised food safety experts. (1) The network metrics of different subgraphs are different, so whether they can provide a basis for the exploration of high-risk foods and the high-risk hazardous substances. (2) Whether the attributes of the network in the subgraph with proximal network metrics are proximal or not, and whether it is helpful for the analysis of hazardous substances in food.

6.1. Dataset

Our dataset contains 1571 kinds of meat products, 47 kinds of hazards, a total of 31849 row data, and the dates of manufacture are distributed from January 2017 to December 2018. The data attributes used in this study include: food name, date of manufacture, hazards, maximum residue limit (MRL) in food and determination results of hazardous level. The level of hazard in food is divided into five levels, which can be defined as

$$\text{Level}(p) = \begin{cases} L1 & p \leq MRL \\ L2 & MRL < p \leq 2MRL \\ L3 & 2MRL < p \leq 3MRL \\ L4 & 3MRL < p \leq 4MRL \\ L5 & p > 4MRL \end{cases} \tag{7}$$

where p represents the value of hazards, MRL represents value of maximum residue limit, and $L1$ represents the hazardous level of one.

In order to analyze the hazards in foods with high level hazard, we only consider the detection data with hazardous level greater than L1. After screening, our detection data contains 23 kinds of hazards, 938 kinds of food, a total of 1098 pieces of data. In order to make the nodes correspond to the food one by one, we combine the same food in the detection data. For example, if the level of lead hazard detected from beef A is L2, and the level of lead hazard detected from beef B is L4, then we think that the hazardous level of lead in Beef is L3. Beef A and beef B are the same kind of food, which are represented by a node in the food network. This is a method of calculating the average (rounding).

Taking food as the node, if there is the same hazard between food, an edge is connected between two nodes. As shown in Fig. 4. The nodes with

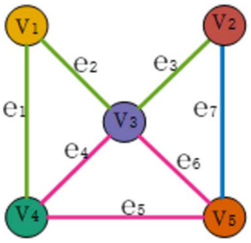


Figure 4: Food network construction approach. The nodes of different colors represent different kinds of foods, and the edges of different colors represent different hazards.

different colors represent the different foods, and the edges with different colors represent different hazards. In this way, we construct a food network with 938 nodes and 105012 edges.

Food network can be used to analyze the correlation between foods. The network metrics mentioned in section 4 can be used as examples. High degree centrality indicates that there are many kinds of hazards in the food represented by this node or the node has a common hazard with many kinds of food. The closeness centrality reflects the similarity between foods. It is helpful for hazard analysis in similar food. If the node is connected to a number of densely linked node clusters (with high betweenness centrality), it indicates that the food has a large number of potential risks of a variety of hazards. Average nearest-neighbor degree shows that the food is closely related to high-risk food or hazardous substances, which is also worthy of our attention. PageRank can analyze the central node of the food network and the meaningful structure around it.

6.2. Specific example

In order to show the attributes of food network, we selected five auxiliary views, as shown in Fig. 1(e–i). Wordle is used to show the key words of food names. Pie chart is used to show the percentage of hazardous levels of hazards in foods. Bar char is used to count the hazardous level of each hazards. Theme river view is used to showthe level of food hazards in different months. Table is used to show the foods data.

Guided by the embedding view (Fig. 1(a)), we set K to 10 (network clustering into 10 subgraphs) to get an overview of the food network. Fig. 1(b) can be used to compare the network metrics of each subgraph. Through Fig. 1(c), we can intuitively analyze the subgraphs in the food network. For

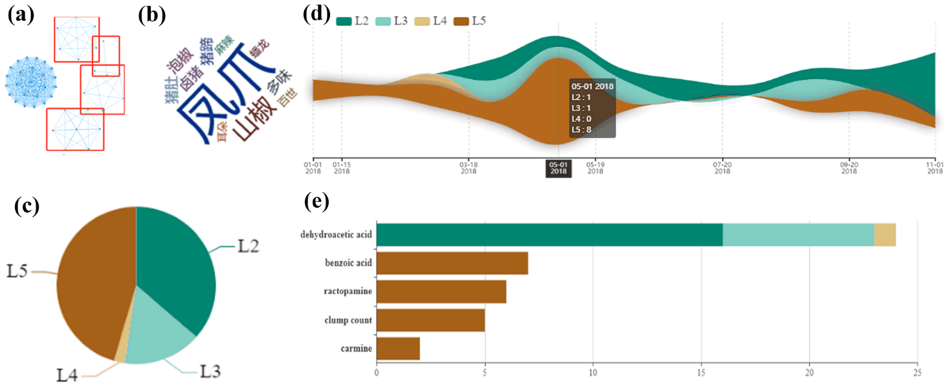


Figure 5: A visual display of the attributes of the food network subgraph with id7.

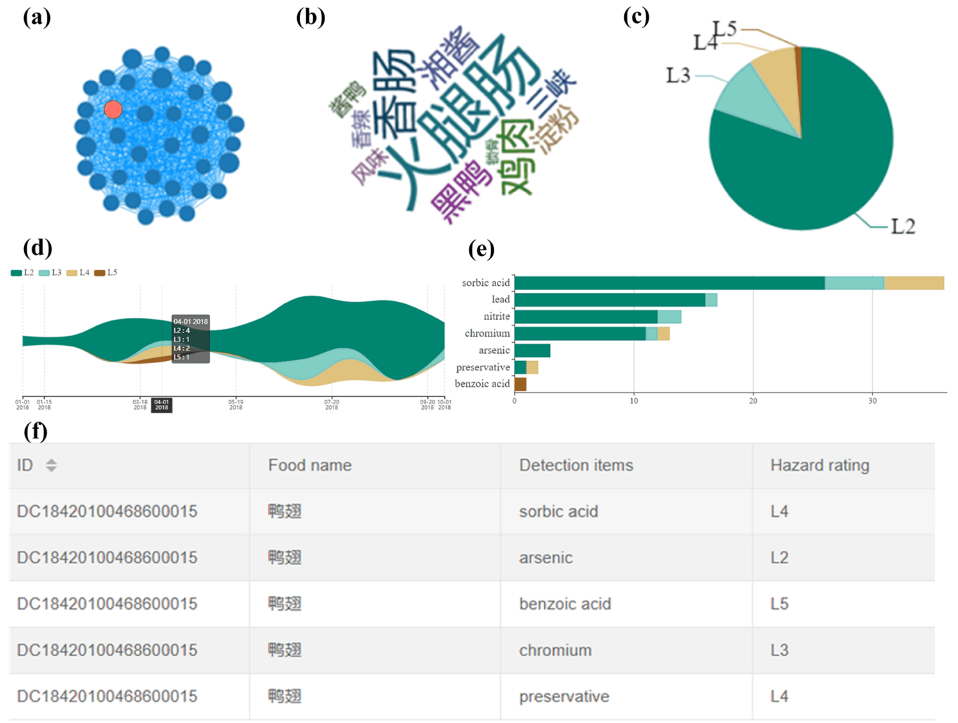
example, the subgraph id1 has the largest number of foods and has many connections with other subgraphs. The connection between subgraph id3 and other subgraphs is the least. In subgraph id2 and subgraph id8, there were the most foods with the same hazards. Fig. 1(d) can intuitively understand the network structure. Then we try to answer the two questions.

In order to answer the first question, we select two subgraphs (id7 and id8) with the largest difference in network metrics from 10 subgraphs. Then we analyze the two subgraphs through the auxiliary views.

We select two subgraphs (id7 and id8) with the largest difference in network metrics, and use auxiliary views to show their network attributes as shown in Fig. 5 and Fig. 6.

It is easy to see that the harm of high hazard in the subgraph id8 is greater than that in the subgraph id7. There are also great differences in the structure of the two subgraphs, id8 is a connected graph and id7's network is composed of five parts as shown in Fig. 5(a) and Fig. 6(a). The subgraph id7 represents is a network constructed by 45 kinds of food as nodes and 5 kinds of hazards as edges, while id8 is a network constructed by 37 kinds of food as nodes and 7 kinds of hazards as edges.

Through the analysis of the nodes attributes of the subgraph id7, we found that in this subgraph, the hazardous level of the food is generally high, and almost half of the food with hazardous level of 5, as shown in Fig. 5(c). And we found that the hazards in the four connected subgraphs with the lowest number of nodes all had a hazardous level of L5 in the food. In addition, there are altogether five food hazards in this subgraph, and according to four of them, the hazardous level of the food is all L5,



ID	Food name	Detection items	Hazard rating
DC18420100468600015	鸭翅	sorbic acid	L4
DC18420100468600015	鸭翅	arsenic	L2
DC18420100468600015	鸭翅	benzoic acid	L5
DC18420100468600015	鸭翅	chromium	L3
DC18420100468600015	鸭翅	preservative	L4

Figure 6: A visual display of the attributes of the food network subgraph with id8.

as shown in Fig. 5(e). In other words, the four subgraphs with the least nodes in id7 are connected by these four hazards. We conclude that the hazards included in this subgraph are high-risk and have been confirmed by food safety experts. It is worth noting that the end of the theme river view, which represents the level of food hazards and the amount of food hazards, tends to get larger. It suggests that in the future there may be more cases of substandard food being detected.

Then we analyze the subgraph id8. Through the auxiliary view, we found that the food in this subgraph had a relatively low hazardous level be detected. Foods with a hazardous level of 5 were detected only in April 2018, as shown in Fig. 6(d). However, we find from Fig. 6(a) that the nodes in the node-link graph are relatively large (there are many connections among nodes). After selecting some nodes and analyzing them, we find that a variety of hazardous substances have been detected in the food corresponding to each node, as shown in Fig. 6(f). We can infer that the food in this subgraph is high-risk food, and the food safety experts also believe that these

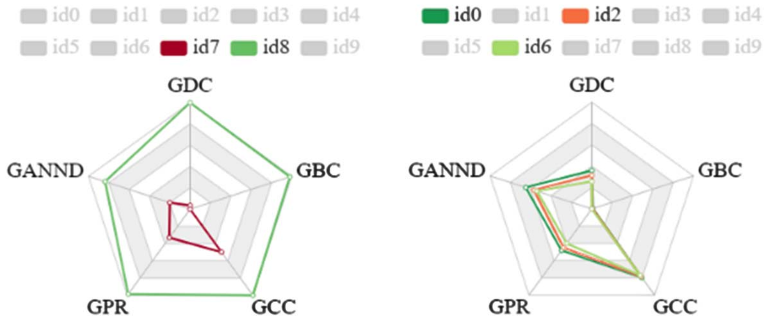


Figure 7: Visualization of the largest and smallest differences in network metrics.

foods need to be strictly monitored. Because the food name contains some trademark information, we do not show the name of the food here, but count all the food and extract their keywords. From the wordle view, we can see that the keywords extracted from the food names contain not only the information of the food ingredients but also some geographical information.

In order to answer the second question, we want to select two subgraphs with the largest proximity. However, there are three subgraphs with very proximal network metrics (id0, id2 and id6), as shown in Fig. 7. Here, the subgraphs represented by id0, id2 and id6 are analyzed as shown in Fig. 8.

Through the analysis, we find that these three subgraphs are very dense connected graph and contain a lot of nodes. The subgraph represented by id0 consists of 141 kinds of food and 4 kinds of hazards, the subgraph represented by id2 consists of 133 kinds of food and 2 kinds of hazards, and the subgraph represented by id6 consists of 80 kinds of food and 1 kind of hazard.

We found that there is only one kind of hazardous substances be detected in the food subgraph represented by id6, which proves that id6 is a complete graph, while the food subgraphs represented by id0 and id2 are not complete graphs, but their edges are mainly composed of one kind of hazardous substances. We can infer that these hazardous substances exist in specific food or only the hazardous substances in the subgraph are detected in these foods frequently. We should remind the supervision organizations to strictly control the use of such hazardous substances in these foods.

From the perspective of food hazardous level, the food hazardous level in these three subgraphs is concentrated in 2 and 3, and there is no food with a particularly high hazardous level. It is a helpful for the supervision

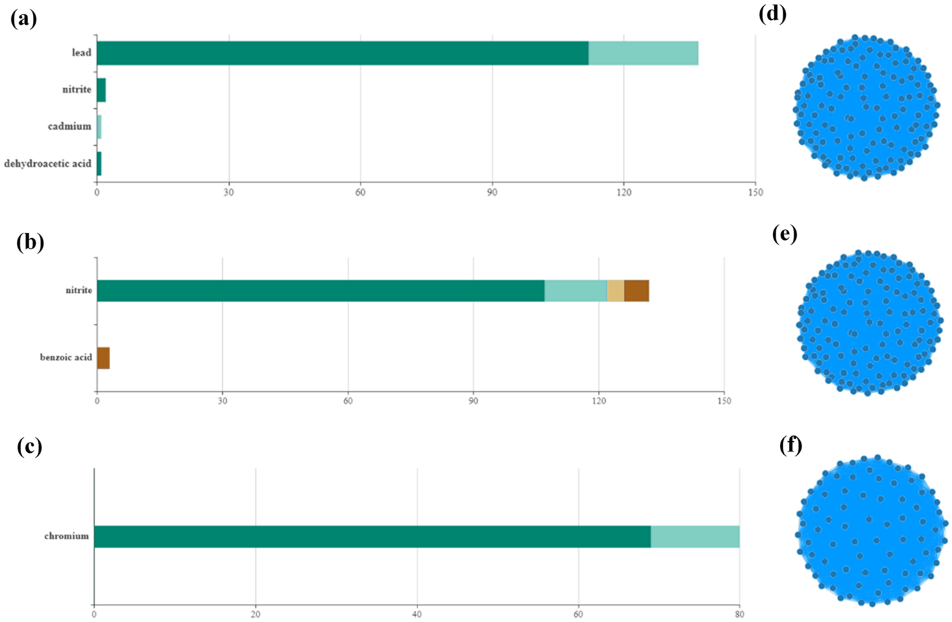


Figure 8: The types and quantities of hazards detected in the food subgraphs represented by id0, id2 and id6 are shown by (a), (b) and (c) respectively, and the network structure views are shown by (d), (e) and (f) respectively.

organizations to have a judgment on the current safety environment of food. Therefore, we have reason to believe that the hazards in these two subgraphs have low-risk levels in these foods. Food safety experts agree with us very much.

In order to answer the two questions raised by food safety experts, we designed two groups of experiments respectively to analyze the performance of subgraph network metrics in the food network. It is proved by examples that according to the differences of network metrics of different subgraphs, users can find high-risk food and high-risk hazards. When the network metrics of the subgraphs is proximately, their structures are also proximately, and some properties of the contained food are proximal too, which can help users analyze the harmful substances in foods.

7. Discussion

Advantages. Our approach uses the graph embedding to vectorize the nodes in the network, so the topological relation between nodes in the network is transferred to the distance information in the high-dimensional space, which avoids the time and storage resources spent in searching the graph in the traditional approach. In the process of exploring the subgraph in the network, we first abstract the subgraph into a radar map, and express the relationship between the subgraphs by the relation between the radar maps. In addition, we also plot the structure feature of the subgraphs on the radar maps to avoid the visual clutter caused by the complex network. Presenting a complex network in the form of an overview can effectively help users grasp the structure of the entire network. We constructed food network and our approach is applied in food network analysis, which can help users explore potential relationships between foods from the perspective of graph analysis, and find high-risk foods and hazards.

Limitations. Our approach needs to manually determine the number of clusters, and the setting of the number of clusters has a great impact on the effectiveness of network exploration, which is unfriendly to users. We use the node-link graph to show the structure of the subgraphs, but when the size of the edges in the subgraph is very large, it is difficult to visualize the specific structure of the subgraph.

Future works. The approach of network node vectorization in this work is based on homogeneous graph, only the information between nodes was preserved, and the influence of different types of edges on node vectorization is not considered. In the future works, we will consider the influence of different types of edges on the network. In addition, we will also explore how to automatically set the number of clusters to make our system more user-friendly and we will also explore the structural composition of the subgraph. Finally, we hope to apply our approach to more fields.

8. Conclusion

In this paper, we design a general approach to visualize the structure of subgraphs and study the relationship between subgraphs of complex networks. Firstly, the network nodes are vectorized by graph embedding technology, and then the nodes are clustered into several subgraphs based on the proximity of structured roles. Radar map is used to represent the subgraph after clustering, and five networks metrics which we defined to measure the proximity of graph structure are displayed on radar map. In order to prove the effectiveness of our approach, we constructed a food network with food as the node and same hazards among foods as the edge. We proposed G-Explorer to analyze the food network and proved the effectiveness of the approach by answering two questions from experts in the field of food safety.

References

[1] M. Behrisch, T. Schreck and H. Pfister, *GUIRO: User-guided matrix reordering.* IEEE Transactions on Visualization and Computer Graphics, **26**(1):184–194, Jan. 2020.

[2] P. Cui, X. Wang, J. Pei and W. Zhu, *A survey on network embedding.* IEEE Transactions on Knowledge and Data Engineering, **31**(5):833–852, 2019.

[3] Y. Chen, X. Zhang, Y. Feng, J. Liang and H. Chen, *Sunburst with Ordered Nodes based on Hierarchical Clustering: A Visual Analyzing Method for Associated Hierarchical Pesticide Residue Data.* Journal of Visualization, **18**(2):237–254, 2015.

[4] Y. Zhao, X. Luo, A. Lin, H. Wang, X. Kui, F. Zhou, J. Wang, Y. Chen and W. Chen, *Visual Analytics for Electromagnetic Situation Awareness in Radio Monitoring and Management.* IEEE Transactions on Visualization and Computer Graphics, **26**(1):590–600, 2019.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, *Distributed Representations of Words and Phrases and their Compositionality.* In "Advances in neural information processing system", ISSAC'2013, pages 3111–3119, 2013. ACM.

[6] B. Perozzi, R. Alrfou and S. Skiena, *DeepWalk: Online learning of social representations.* In "Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining", ISSAC'2014, pages 701–710, 2014.

[7] A. Grover and J. Leskovec, *node2vec: Scalable feature learning for networks.* In "Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining", ISSAC'2016, pages 855–864, 2016. ACM.

[8] D. R. Figueiredo, L. F. R. Ribeiro and P. H. P. Saverese, *struc2vec: Learning node representations from structural identity.* In "Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining", ISSAC'2017, pages 385–394, 2017.

[9] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan and Q. Mei, *Line: Large-scale information network embedding.* In "Proceedings of the 24th international conference on world wide web", ISSAC'2015, pages 1067–1077, 2015.

[10] A. Narayanan, M. Chandramohan, L. Chen, Y. Liu and S. Saminathan, *Subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs.* arXiv preprint arXiv:1606.08928, 2016.

[11] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu and S. Jaiswal, *Graph2vec: Learning distributed representations of graphs.* arXiv preprint arXiv:1707.05005, 2017.

[12] Y. Dong, N. V. Chawla and A. Swami, *Metapath2vec: Scalable representation learning for heterogeneous networks.* In "The 23th. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", ISSAC'2017, pages 135–144, Halifax, 2017. ACM.

[13] T. Y. Fu, W. C. Lee and Z. Lei, *HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning.* In "Proceedings of the 2017 ACM on Conference on Information and Knowledge Management", ISSAC'2017, pages 1797–1806, Singapore, 2017. ACM.

[14] C. Shi, B. Hu, W. X. Zhao and S. Y. Philip, *Heterogeneous information network embedding for recommendation.* IEEE Transactions on Knowledge and Data Engineering, **31**(2):357–370, Feb. 2018.

[15] Y. Wu, N. Cao, D. Archambault, Q. Shen, H. Qu and Y. Wu, *Evaluation of graph sampling: A visualization perspective.* IEEE Transactions on Visualization and Computer Graphics, **23**(1):401–410, Jan. 2017.

[16] Y. Zhao, F. Luo, M. Chen, Y. Wang, J. Xia, F. Zhou, Y. Wang, Y. Chen and W. Chen, *Evaluating multi-dimensional visualizations for understanding fuzzy clusters.* IEEE transactions on visualization and computer graphics, **25**(1):12–21, Jan. 2019.

[17] A. Lhuillier, C. Hurter and A. Telea, *State of the art in edge and trail bundling techniques.* Computer Graphics forum, **36**(3):619–645, Jun. 2017.

[18] Y. Chen, Z. Guan, R. Zhang, X. Du and Y. Wang, *A Survey on Visualization Approaches for Exploring Association Relationships in Graph Data.* Journal of Visualization, **33**(3):625–639, Jun. 2019.

[19] M. Zinsmaier, U. Brandes, O. Deussen and H. Strobelt, *Interactive level-of-detail rendering of large graphs.* IEEE Transactions on Visualization and Computer Graphics, **18**(12):2486–2495, Dec. 2012.

[20] V. Yoghourdjian, T. Dwyer, K. Klein, K. Marriott and M. Wybrow, *Graph thumbnails: Identifying and comparing multiple graphs at a*

glance. IEEE Transactions on Visualization and Computer Graphics, **24**(12):3081–3095, Dec. 2018.

[21] R. Pienta, F. Hohman, A. Endert, A. Tamersoy, C. Gates, S. B. Navathe and D. H. Chau, *VIGOR: Interactive visual exploration of graph query results*. IEEE Transactions on Visualization and Computer Graphics, **24**(1):215–225, Jan. 2017.

[22] T.Major and R. C. Basole, *Graphicle: Exploring Units, Networks, and Context in a Blended Visualization Approach*. IEEE Transactions on Visualization and Computer Graphics, **25**(1):576–585, Jan. 2019.

[23] K. Yan, W. Cui and T. Zhao, *Frequent pattern-based graph exploration*. In "Proceedings of the 12th International Symposium on Visual Information Communication and Interaction", ISSAC'2019, pages 1–8, Shanghai, China, 2019. ACM.

[24] W. Chen, F. Guo, D. Han, J. Pan, X. Nie, J. Xia and X. Zhang, *Structure-based suggestive exploration: A new approach for effective exploration of large networks*. IEEE Transactions on Visualization and Computer Graphics, **25**(1):555–565, Jan. 2019.

[25] L. Van der Maaten and G. Hinton, *Visualizing data using t-SNE*. Journal of Machine Learning Research, **9**(11):2579–2605, Nov. 2008.

[26] J. A. Hartigan and M. A. Wong, *A K-means clustering algorithm*. Journal of the Royal Statistical Society, **28**(1):100–108, Jan. 1979. MR0405726

[27] X. Ren and L. Lv, *Review of ranking nodes in complex networks*. Review of ranking nodes in complex networks, **59**(13):1175–1197, May. 2014.

[28] P. Bonacich, *Factoring and weighting approaches to status scores and clique identifcation*. Journal of Mathematical Sociology, **2**(1):113–120, Aug. 1972.

[29] L. C. Freeman, *Centrality in social networks conceptual clarifcation*. Social Networks, **1**(3):215–239, 1978.

[30] L. C. Freeman, *A Set of Measures of Centrality Based on Betweenness*. Sociometry, **40**(1):35–41, 1977.

[31] R. Pastorsatorras, A. Vázquez and A. Vespignani, *Dynamical and correlation properties of the Internet*. Physical Review Letters, **87**(25):258701, Nov. 2001.

[32] S. Brin and L. Page, *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems, **30**(1–7):107–117, 1998.

[33] Y. Chen, C. Lv, Y. Li, W. Chen and K. Ma, *Ordered matrix representation supporting the visual analysis of associated data*. Science China Information Sciences, **63**(8):1–3, Aug. 2020.

[34] M. Harrower and C. A. Brewer, *ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps*. The Cartographic Journal, **40**(1):27–37, Jun. 2003.

QingHui Zhang
Beijing Key Laboratory of Big Data Technology for Food Safety
School of Computer Science and Engineering
Beijing Technology and Business University
Beijing, 100048
China
E-mail address: 528997486@qq.com

Yi Chen
Beijing Key Laboratory of Big Data Technology for Food Safety
School of Computer Science and Engineering
Beijing Technology and Business University
Beijing, 100048
China
E-mail address: chenyi@btbu.edu.cn

MengLu Zhang
Beijing Key Laboratory of Big Data Technology for Food Safety
School of Computer Science and Engineering
Beijing Technology and Business University
Beijing, 100048
China
E-mail address: 251230950@qq.com

Zeli Guan
Beijing Key Laboratory of Big Data Technology for Food Safety
School of Computer Science and Engineering
Beijing Technology and Business University
Beijing, 100048
China
E-mail address: 445151865@qq.com

Received October 25, 2020