Mining Highly Reliable Dense Subgraphs from Uncertain Graphs

LU Yihong, HUANG Ruizhi* and HUANG Decai

College of Computer Science & Technology, Zhejiang University of Technology
Hangzhou, 310023 - China
[e-mail: lyh@zjut.edu.cn]
*Corresponding author: HUANG Ruizhi

Received Septebmer 27, 2018; revised January 16, 2019; accepted January 23, 2019; published June 30, 2019

Abstract

The uncertainties of the uncertain graph make the traditional definition and algorithms on mining dense graph for certain graph not applicable. The subgraph obtained by maximizing expected density from an uncertain graph always has many low edge-probability data, which makes it low reliable and low expected edge density. Based on the concept of β-subgraph, to overcome the low reliability of the densest subgraph, the concept of optimal β-subgraph is proposed. An efficient greedy algorithm is also developed to find the optimal β-subgraph. Simulation experiments of multiple sets of datasets show that the average edge-possibility of optimal β-subgraph is improved by nearly 40%, and the expected edge density reaches 0.9 on average. The parameter β is scalable and applicable to multiple scenarios.

Keywords: uncertain graph, network reliability, surplus average degree, optimal β-subgraph, graph mining

This research was supported by a research grant from Zhejiang Public Welfare Technology Research Project [GG19E090005], We express our thanks to all the reviewers and Dr. Mohammad Shojafar who carefully checked our manuscript.

http://doi.org/10.3837/tiis.2019.06.012 **ISSN : 1976-7277**

1. Introduction

In recent years, many data mining topics have focused on the problem of mining dense subgraphs from a large graph. A dense subgraph refers to a relatively dense internal sub-area in a graph that is widely used in community-search in social networks [1], detection of DNA sequence structure [2], and identification of real-time reporting information in news [3].

Finding the densest subgraph is an important graph-mining task with many applications [4]. Given a graph $G=(V, E)$, the degree density is defined as $|E|/|V|$. The densest-subgraph problem is to find a subset of vertices $S \subseteq V$ that maximizes the degree density.

Due to experimental errors, noise and other reasons, uncertainty has been recognized to be intrinsic in graph data. This graph is called an uncertain graph [5]. A protein-protein interaction network (hereinafter referred to as a protein network) is a typical uncertain graph, as shown in **Fig. 1**. The vertex represents the protein molecule, the edge represents the interaction relationship between the protein molecules, and the edge-probability represents the credibility of the interaction [6], and its value ranges from 0 to 1. Because of the interference of many factors in protein experiments, proteins with no direct interaction may be detected by mistake, resulting in false positives. Therefore, credibility is used to measure the possibility of true protein interaction.

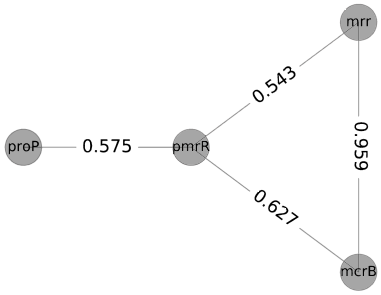


Fig. 1. PPI Network

The dense subgraphs in the protein network often correspond to protein complexes [7], that is, multiple proteins are combined by interaction for a specific function. For example, the Fanconi protein complex [8] is involved in the DNA damage repair process. Therefore, mining dense subgraphs in uncertain graphs is of great significance for protein complex recognition and prediction of unknown protein complexes.

In the study of mining dense subgraphs in an uncertain graph, the maximal clique is used to describe the dense subgraph on the uncertain graph in [9], and the concept of maximal clique probability is proposed based on the uncertainty semantics. Then, an optimized branch-and-bound algorithm, which adopts a new searching strategy, is presented to find top-k maximal cliques. The algorithm in [9] is extended in [10]. Taking advantage of parallelism, a decomposition-based algorithm is proposed to solve the problem on large uncertain graphs. However, the maximal clique requires that any two vertices in the subgraph connected, this definition is too strict to fully reflect the dense subgraph in the real graphs.

Zou et al. [11] first defined the expected density of an uncertain graph and formalized the problem of obtaining the densest subgraph in an uncertain graph. The dense subgraphs found in this way have certain defects. For example, if A and B in **Fig. 2** are two uncertain subgraphs

of an uncertain graph, their expected density is 0.7, and it is impossible to compare these two subgraphs based on the expected density.

In [12], the concept of uncertain graph reliability and reliable subgraphs is introduced, and a sampling-based algorithm is proposed to mine subgraphs with high reliability in uncertain graphs. The reliability of uncertain graphs measures the reliability of a subgraph. Compared with A and B in **Fig. 2**, B' reliability is lower because there are lot of low probability edges in B. Hence, reliable subgraphs are more focused on reliability, and many highly reliable subgraphs are not dense.

To overcome the above weakness of uncertain dense subgraphs that cannot balance reliability and density, this paper introduces the concepts of β-subgraph and optimal β-subgraph, and proposes a greedy approximation algorithm to find the optimal β-subgraph. At the same time, we proved that the surplus average degree of the solution obtained by the algorithm is at least ½ of the surplus average degree of the optimal β-subgraph. The experimental results of set of datasets show that, compared with the densest subgraph, the reliability and the expected edge density are significantly improved, and the dense subgraph is extended well by the parameter β. Thus, the algorithm has broad practical applications.

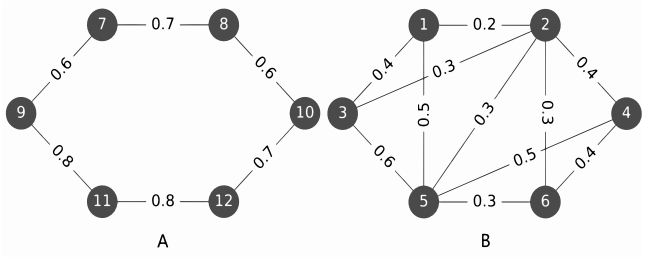


Fig. 2. Two Uncertain Subgraphs

2. Related Work

The problem of determining dense subgraphs on graphs has recently been extensively studied, and various definitions of dense subgraphs have been proposed. Given a graph model $G=(V, E)$ and a vertex subset $S \subseteq V$, $G_S=(S,E(S))$ is a induced subgraph of G for S. Edge density is defined as *edge den*$(S)=|E(S)|/\binom{|S|}{2}$. The subgraph with the maximum edge density of 1 is called a clique and the clique with the most vertices of the graph is the maximal clique [13]. The maximal clique is regarded as a dense area in the graph. However, the problem of the maximal clique is an NP-hard problem, and many graphs have no clique. Therefore, the quasi-clique is proposed in [14] to avoid the NP-hard.

The density of the graph was initiated by Goldberg [4], it is defined as $den(S)= |E(S)|/|S|$. The maximum density subgraph problem (DS-Problem) is to find a sub-graph with the highest density in a G. An algorithm to obtain the approximate solution of the maximum density subgraph is also proposed by using the maximum flow minimum cut theorem in [4]. A greedy algorithm based on the maximum density subgraph is suggested in [15] to solve the maximum density subgraph problem. In [16], the maximum density problem has been further extended by limiting the size of the subgraph. However, it is an NP-hard problem [17]. MapReduce [18], real-time evolving graphs [19] and other techniques are also applied to study the maximum density subgraph problem. In addition, dense subgraphs are defined as k-core[20], k-clique [21], (α, d, L)-decompositions [22], etc.

The definition of uncertain graphs was first studied by Gao & Gao [23]. The research of uncertain graphs has been an active area in recent years, and many related concepts, mining and query algorithms, such as reliable subgraphs [12], frequent pattern mining [24], vertex reachable query [25], graph clustering [26], etc. have been proposed. A number of definitions have been proposed to describe dense subgraphs in a given uncertain graph, e.g., the maximal clique in an uncertain graph in [10]. The k-core subgraph in an uncertain graph [27] and the enumeration of maximal cliques from an uncertain graph is proposed in [28].

For the density definition and the problem of finding the densest subgraph in an uncertain graph, [11] first formalizes the densest subgraph problem on uncertain graphs and introduces the expected density of an uncertain graph. [29] proposes a different definition of expected density of an uncertain graph, we called it *expected edge density*, and investigates the problem that mining a top-k dense subgraph mining problem from uncertain graphs and proves that the problem is NP-Hard problem. A partial order on all induced subgraphs is defined in the paper. Through the partial order, all induced subgraphs are organized as into an enumeration tree, then based on a branch, bound search algorithm is applied to produce top-k dense subgraphs. A definition of the disjoint top-k dense subgraph and a heuristic approximation algorithm are also proposed.

3. Related Definitions

The uncertain graph is represented as a triplet $G=(V, E, p)$, where V is the set of all vertices, $E \subseteq V \times V$ is the set of all edges, and $p: E \rightarrow (0,1]$ is the probability function of the existence of any edge $e \in E$. Let $S \subseteq V$, then $G_S=(S, E(S), p')$ denotes a subgraph derived from S in G, where $E(S)$ indicates a set of edges with both endpoints in S.

Definition 1 Adjoint Graph [30]. Given an uncertain graph $G=(V, E, p)$, the certain graph G^* obtained with all edges' probability equal to 1 is called the adjoint graph of G.

Fig. 3 shows an uncertain graph and its adjoint graphs. In this paper, we assume that the probability of each edge is independent of each other. According to the definition of network reliability in [12], we get:

Definition 2 Adjoint Reliability. Given an uncertain graph $G = (V, E, p)$, its adjoint reliability is defined as:

$$R(G)=\prod_{e \in E} p(e) \tag{1}$$

Definition 3 Average Edge-probability. Given an uncertain graph $G = (V, E, p)$, G_S is an induced subgraph for a subset $S \subseteq V$, and the average edge-probability of G_S is defined as:

$$\overline{p}(G_S(e)) = \sum_{e \in E[S]} p(e) \Big/ |E[S]| \tag{2}$$

Definition 4 Expected Edge Density [29]. Given an uncertain graph $G = (V, E, p)$ and G_S is an induced subgraph for a subset $S \subseteq V$, and the expected edge density of G_S is:

$$\tau(G_S) = \sum_{e \in E[S]} p(e) \Big/ \binom{|S|}{2} \tag{3}$$

Theorem 1 Given an uncertain graph $G = (V, E, p)$ and G_S is an induced subgraph for a subset $S \subseteq V$, the expected edge density of G_S increases as its average edge-probability

increases.

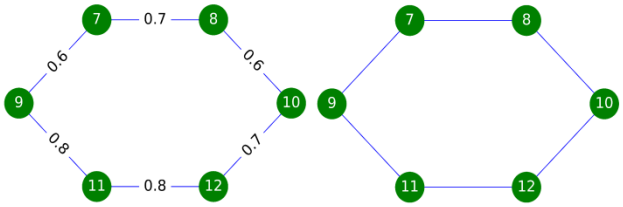


Fig. 3. Uncertain Graph and Adjoint Graphs

Proof by definition 1:

$$\sum_{e\in E[S]} p(e) = \overline{p}((G_S(e)))\cdot |E[S]|$$

$$\Rightarrow \tau(G_S) = \overline{p}(G_S(e))\cdot \frac{|E[S]|}{\binom{|S|}{2}}$$

For a subgraph G_S, the number of edges and vertices are determined, so the expected edge density increases with the increase of the average edge-probability.

Definition 5 Expected Dgree. For a vertex $v\in V$ of the uncertain graph $G=(V, E, p)$, $E(e)=\{e \mid e\in E$ and v is a vertex of $e\}$, the expected dgree of v is :

$$Deg(v) = \sum_{e\in E(e)} p(e) \tag{4}$$

Definition 6 Expected Density [11]. Given an uncertain graph $G = (V, E, p)$ and G_S is an induced subgraph for a subset $S\subseteq V$, the expected density of G_S is:

$$\delta(G_S) = \sum_{e\in E[S]} p(e) \Big/ |S| \tag{5}$$

Definition 7 Densest Subgraph Problem on Uncertertain Graphs. Given an uncertain graph $G=(V, E, p)$, an induced subgraph G_S for a subset $S\subseteq V$ such that $\delta(G_S)=\max\{\delta(G_{S'})|S' \subseteq V\}$ is called densest subgraph problem on uncertertain graphs, or the UDS problem in short. That is, G_S is the UDS of G.

The GreedyUDS algorithm we introduce in this paper is different from the minimum cut idea in [11], and is based on the Charikar algorithm. The relevant modifications are made to the properties of the uncertain graph. The algorithm calculates the expected density of the current graph during each iteration and deletes the vertices with the least expected degree until all vertices are removed, and finally outputs the subgraph with the largest expected density . The details are described in Algorithm 1. The time complexity of the algorithm is O(m + n), where n is the number of vertices and m is the number of edges in the uncertain graph.

Algorithm 1. GreedyUDS algorithm flow.

Input: uncertain graph $G = (V, E, p)$

Output: vertex set $S\subseteq V$

① $n\leftarrow|V|$, $Hn\leftarrow G$;

② for $i=n$ to 2 ;

③ $H_{i-1}=H_i-\{v\}$; /* Where v is the vertice with the least expected degree in H_i */

④ end for

⑤ return H' ; /* H' is the subgraph with the largest expected density in the set $\{H_2, H_3,..., H_n\}$ */

Table 1 shows the experimental results of the GreedyUDS algorithm of several protein network datasets. The specific information of the dataset is listed in **Table 2**. We observed that the average value of the edge-probability of the dense subgraph obtained by the algorithm is small, and the standard deviation is large, that means there are lots of edges with low probability in the subgraph, resulting in low reliability.

Table 1. Densest subgraph

Data set	Δ	τ	$\overline{p}(G_S(e))$	σ
945681.protein	28.153	0.408	0.527	0.213
511145.protein	46.842	0.061	0.338	0.129
8364.protein	64.174	0.221	0.322	0.194

Note:δ represents the expected density of the subgraph, τ denotes the expected edge density of the subgraph, σ is the standard deviation of the edge-probability

Table 2. Experimental datasets

Data set	Number of vertices	Number of edges	Average edge-possibility	Edge-probability standard deviation	Description
Celegans	131	687	0.5	0.291	Neural Network
Email-Enron	36692	183831	0.5	0.289	Email Network
579138.protein	1672	81770	0.322	0.226	Zymomonas mobilis
945681.protein	2369	124544	0.318	0.213	Acetobacter pomorum
511145.protein	4145	568789	0.297	0.190	Escherichia coli
1097668.protein	5617	866990	0.283	0.174	Burkholderia
8364.protein	16745	3117801	0.284	0.183	Xenopus Silurana

4. Optimal β-subgraph

4.1 Problem Analysis

Definition 8 β-subgraph. Given an uncertain graph $G = (V, E, p)$ and G_S is an induced subgraph for a subset $S \subseteq V$, if the average edge-probability of G_S is not less than β, where $\beta \in (0, 1)$, it is called β-subgraph, referred to as β-G_S.

Definition 9 Surplus Degree. Given an uncertain graph $G = (V, E, p)$, and its vertex $v \in V$ and the parameter $\beta \in (0,1)$, $E(e)=\{e|e \in E$ and v is one of $e\}$, we define the surplus degree of v as:

$$SDeg(v) = \sum_{e \in E(e)} [p(e) - \beta] \tag{6}$$

Definition 10 Surplus Average Degree. Given an uncertain graph $G = (V, E, p)$, and its subset $S \subseteq V$ and the parameter $\beta \in (0, 1)$, the surplus average degree of the induced subgraph G_S of the vertex set is defined as:

$$f_\beta(G_S) = \sum_{e \in E[S]} p(e) \Big/ |S| - \beta \cdot |E[S]| / |S| \tag{7}$$

Formula (7) is properly converted to:

$$f_\beta(G_S) = [\overline{p}(G_S(e)) - \beta] \cdot \frac{|E[S]|}{|S|} \tag{8}$$

From formula (8), we find that when the surplus average degree is maximized, the parameter β as the influence factor can filter out the subgraphs with more low probability edges, that improves the adjoint reliability of the final subgraph. On the right side of the equation, the dense structure of the adjoint graph of the subgraph is guaranteed.

Theorem 2 Given an uncertain graph $G = (V, E, p)$ and the parameter $\beta \in (0, 1)$, for the induced subgraph G_S of a subset $S \subseteq V$, if $f_\beta(G_S) \geq 0$, then G_S must be a β-subgraph.

Proof From $f_\beta(G_S) \geq 0$, we have:

$$\frac{\sum_{e \in E[S]} p(e)}{|S|} - \beta \cdot \frac{|E[S]|}{|S|} \geq 0$$

$$\Rightarrow \sum_{e \in E[S]} p(e) \geq \beta \cdot |E[S]|$$

$$\Rightarrow \frac{\sum_{e \in E[S]} p(e)}{|E[S]|} \geq \beta$$

Definition 11 Optimal β-subgraph. Given an uncertain graph $G = (V, E, p)$, find a subset of its vertices $S^* \subseteq V$ such that:

$$f_\beta(G_{S^*}) = \max\{ f_\beta(G_S) \mid S \subseteq V \} \tag{9}$$

We call the subgraph G_{S^*} the optimal β-subgraph, and the optimal β-subgraph is referred to as OβS in short.

4.2 GreedyOβS Algorithm

From definition 5 and definition 6, we find that OβS can be approximated by removing the vertices with less surplus degree, so we propose GreedyOβS algorithm by modifying GreedyUDS algorithm, and algorithm 2 gives the details.

Algorithm 2. GreedyOβS.

Input: Uncertain graph $G = (V, E, p)$ and parameter β

Output: vertex set $S \subseteq V$

① Initialize the priority queue Q, calculate the surplus degree of each vertex in the graph, and store it in Q, the elements in Q are sorted by increasing surplus degree.

② Let $n \leftarrow |V|$, $H_n \leftarrow G$

③ With $i=n$, perform the following steps until $i=2$:
1) Calculate the surplus average degree of H_i
2) Remove the first element v of Q and update the relevant vertex surplus degree in Q
3) $H_{i-1}=H_i-\{v\}$

④ Output H', where H' is the subgraph with the largest surplus average degree among $\{H_2, H_3, \ldots, H_n\}$

The algorithm requires a parameter β as an average edge-probability threshold, and maintains a vertex surplus degree queue of an uncertain graph at the beginning. The algorithm removes the vertices with the smallest surplus degree in each iteration and updates the relevant vertices and their surplus degree at the same time. The above is repeated until all vertices are removed. Finally, the algorithm outputs the vertex subset with the largest $f_\beta(G_S)$. The space and time complexity of the algorithm are O(3m + 2n) and O(m + n) respectively, where n is the number of vertices and m is the number of edges in the uncertain graph.

4.3 Algorithm Accuracy

Suppose that for an uncertain graph $G=(V, E, p)$, G_{S^*} is the largest subgraph of the surplus average degree, that is, G_{S^*} is the optimal solution. Let $f_\beta(G_{S^*})=\lambda$, $m_{s^*}=|E[S^*]|$, $n_{s^*}=|S^*|$, the average edge-probability of G_{S^*} is $\bar{p}(G_{S^*}(e))$, then we obtain:

Theorem 3 Given an uncertain graph $G=(V, E, p)$, for any of its vertex $u \in S^*$, there must be $SDeg(u) \geq \lambda$.

Proof Since G_{S^*} is the optimal solution, we know that $f_\beta(G_{S^*}) \geq f_\beta(G_{S^*\setminus\{u\}})$, then:

$$\frac{\sum_{e\in E[S^*]} p(e) - \beta \cdot m_{S^*}}{n_{S^*}} \geq \frac{\sum_{e\in E[S^*]} p(e) - \beta \cdot m_{S^*} - SDeg(u)}{n_{S^*}-1}$$

$$\Rightarrow SDeg(u) \geq (1-\frac{n_{S^*}-1}{n_{S^*}})[\sum_{e\in E[S^*]} p(e) - \beta \cdot m_{S^*}]$$

$$\Rightarrow SDeg(u) \geq \frac{1}{n_{S^*}}[\sum_{e\in E[S^*]} p(e) - \beta \cdot m_{S^*}]$$

$$\Rightarrow SDeg(u) \geq \lambda$$

Theorem 4 Given an uncertain graph $G = (V, E, p)$ and a parameter $\beta \in (0, 1)$, G_{S^*} is an optimal β-subgraph of G, that is the optimal solution, it's surplus average degree is $f_\beta(G_{S^*})$. $G_{S'}$ is a near optimal β-subgraph of G, obtained by GreedyOβS, that is the approximate solution, it's surplus average degree is $f_\beta(G_{S'})$. Then,

$$\frac{f_\beta(G_{S'})}{f_\beta(G_{S^*})} \geq \frac{1}{2} \tag{10}$$

that is, the surplus average degree of $G_{S'}$ obtained by the GreedyOβS algorithm is not less than 1/2 of the surplus average degree of the optimal β-subgraph G_{S^*}.

Proof Let the algorithm iterative process proceed until the first vertex v of S^* is removed. The resulting vertex set is $S' \subseteq V$, and the number of vertices of S' is $n_{s'}$, and the number of edges is $m_{s'}$. From Theorem 3, $SDeg(v) \geq \lambda$, and because the algorithm removes the vertices with the smallest surplus degree in every iteration, we have:

$$\sum_{u \in S'} SDeg(u) \geq \lambda \cdot n_{S'}$$

$$\Rightarrow \sum_{u \in S'} SDeg(u) = \frac{1}{2}[\sum_{e \in E[S']} p(e) - \beta \cdot m_{s'}] \geq \lambda \cdot n_{S'}$$

$$\Rightarrow f_\beta(G_{S'}) \geq \frac{\lambda}{2}$$

$$\Rightarrow \frac{f_\beta(G_{S'})}{f_\beta(G_{S^*})} \geq \frac{1}{2}$$

Let $S^+ \subseteq V$ be the result of GreedyOβS. Since the algorithm finally outputs the vertex subset with the largest $f_\beta(G_S)$, then we know that:

$$f_\beta(G_{S^+}) \geq f_\beta(G_{S'}) \geq \frac{\lambda}{2}$$

Therefore, even in the worst case, the surplus average degree of the solution obtained by the algorithm is not less than 1/2 of the surplus average degree of the optimal solution.

5. Experiments and Results

In this section, a lot of simulations are done to compare the optimal β-subgraph and the densest subgraph of the uncertain graph. We use the GreedyUDS algorithm to obtain the densest subgraph and use the GreedyOβS algorithm to obtain the optimal β-subgraph. The differences between UDS and OβS under different evaluation indexes are then compared, and the influence of different parameters β on the obtained OβS is also tested.

All the algorithms in this paper are implemented with Python3.5. The minimum value of float type in Python is 2.225e-308, which is represented by MIN. The experimental environment is a PC with a Core I5 2.30GHz processor and 8GB of RAM and running the win10 operating system. The datasets used in this experiment are shown in **Table 3**. Without loss of generality, this paper assures all edges of the graph dataset Celegans and Email-Enron are assigned with random probability, converted to uncertain graphs, and the remaining uncertain graphs are true datasets. Both are protein networks from the STRING-DB database (http://string-db.org).

Table 3. Comparison between UDS and OβS

Data Set	τ		$\overline{p}(G_S(e))$		σ		R	
	UDS	***OβS***	***UDS***	***OβS***	***UDS***	***OβS***	***UDS***	***OβS***
Celegans	0.075	0.254	0.534	0.881	0.292	0.095	0.177	1.026
Email-Enron	0.069	0.721	0.506	0.901	0.290	0.066	<MIN	0.425
579138.protein	0.642	0.940	0.718	0.947	0.291	0.130	<MIN	2.487e-49
945681.protein	0.408	0.962	0.527	0.967	0.302	0.104	<MIN	6.837e-22
511145.protein	0.061	0.940	0.338	0.944	0.212	0.120	<MIN	2.179e-69
8364.protein	0.080	0.900	0.267	0.900	0.162	0.002	<MIN	1.069e-305

5.1 Evaluation Indicators

The evaluation index of this paper is divided into two parts. The first part evaluates the density of the subgraph and uses the expected edge density τ of formula 3 as the evaluation criterion. The second part evaluates the reliability of the subgraph, using the adjoint reliability R of formula 1 as the direct evaluation criterion. At the same time, the average edge-probability and the edge-probability standard deviation (σ) are used as indirect evaluation indicators, that is, when the average edge-probability of a subgraph is high, and the standard deviation of the edge-probability is low, the adjoint reliability of the subgraph is considered to be higher.

5.2 Model Comparison

The GreedyUDS and GreedyOβS algorithms were simulated with the same batch of datasets. The default parameter β was 0.6, and the results were compared with the evaluation indicators in the section 5.1. The results obtained are also used to manifest OβS's advantages over UDS. The calculation results of each evaluation index are shown in **Table 3**. Moreover, two experimental datasets were randomly selected, and the same edges were randomly selected from the original image, UDS and OβS. Scatter plots are drawn to visually compare the edge-probability distributions of these three models. The results are depicted in **Fig. 3**. The cross point represents the original image, the diamond represents UDS, and the circle represents OβS.

Analysis of the above experimental results are concluded as follows:

1) The density of OβS is greatly improved compared with UDS. **Table 3** shows that the average expected edge density of UDS calculated by all experimental datasets is 0.223, and the average expected edge density of OβS is 0.786. Additionally, for the experimental set of artificially transformed uncertain graphs, compared with UDS, the expected edge density of OβS is improved and the average increases from 0.072 to 0.488. For the experimental graph of uncertain graphs from the protein network, the expected edge density of OβS is higher than that of UDS, and the average increases from 0.298 to 0.936. Therefore, the density of OβS is significantly higher than that of UDS.
2) The edge-probability distribution of OβS is more uniform than UDS and tends toward 1. **Table 3** shows that the average edge-probability of OβS is improved for all experimental datasets, with the average increased by 0.9 or improved by about 40%. At the same time, the standard deviation of the edge-probability is reduced, and is dropped by about 60%. From **Fig. 3** we observe that most of the edges of the original image have lower probability, and the edge-probability distributions of UDS and OβS are correspondingly improved, and the edge-probability distribution of OβS tends closer to 1 than UDS.
3) The reliability of OβS is greatly improved compared with UDS. Conclusion 2 implies the reliability of OβS is higher than that of UDS. In addition, **Table 3** displays that for each experimental dataset, the adjoint reliability of OβS is significantly improved. Therefore, OβS can be regarded as more reliable than UDS.

5.3 Parameter Selection

The following analyzes the effect of parameter β on OβS through experiments. We select two protein networks as our experimental datasets, and **Table 4** shows the experimental results. From the $f_\beta(G_S)$ equation we know that when β tends to zero, maximizing $f_\beta(G_S)$ is equivalent

to maximizing the expected density. Therefore, when β is 0.1, the result is close to that of UDS. When the value of β is gradually increased, the scale and the expected density of OβS become smaller, and the expected edge density and average edge-probability of OβS become larger.

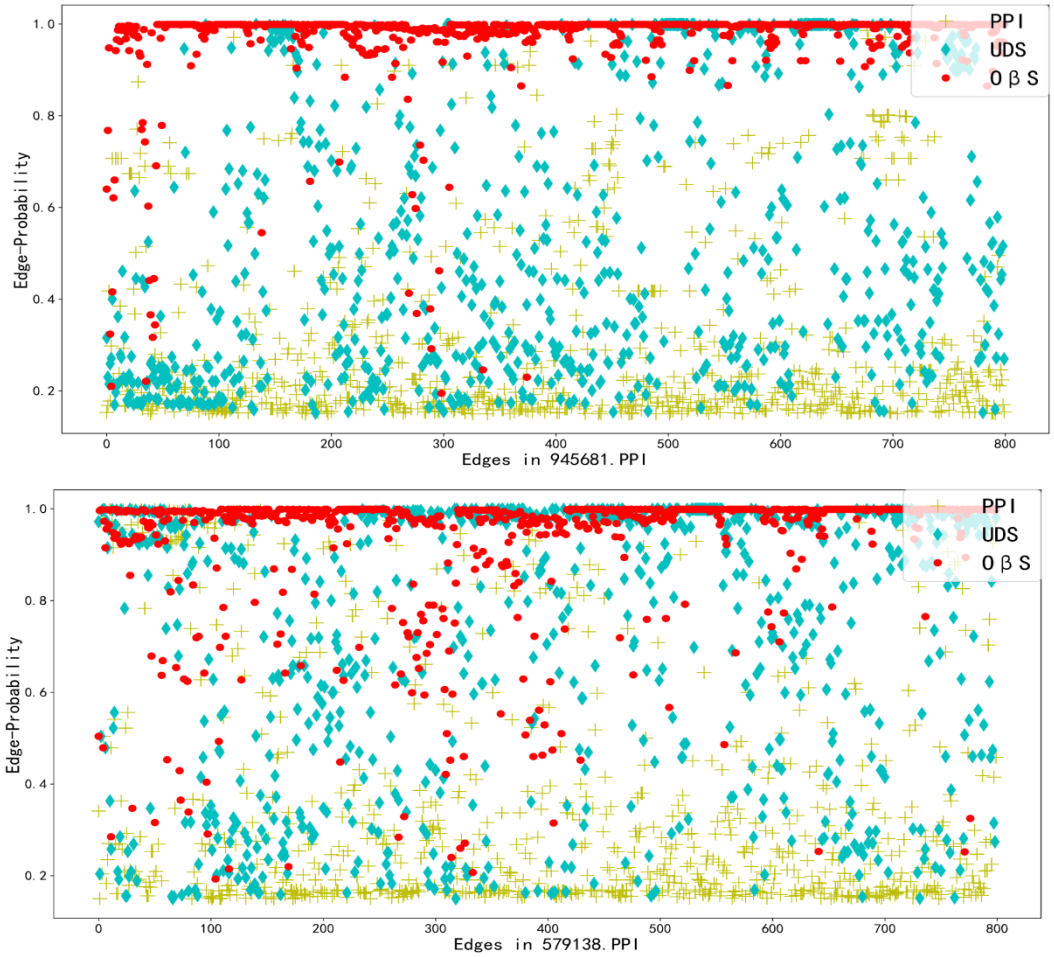


Fig. 3. Edge-Probability Distribution

Table 4. Comparison between different β

Data set	β	Number of edges	δ	τ	$\overline{p}(G_S(e))$	σ
579138.protein	0.1	88	30.721	0.706	0.771	0.271
	0.2	78	30.203	0.785	0.829	0.242
	0.4	70	29.371	0.851	0.877	0.207
	0.6	57	26.322	0.940	0.947	0.130
	0.8	50	23.896	0.975	0.977	0.072
8364.protein	0.1	3579	110.129	0.062	0.288	0.190
	0.2	593	58.101	0.196	0.842	0.203
	0.4	466	55.094	0.237	0.892	0.119
	0.6	116	51.754	0.900	0.900	0.002
	0.8	115	51.303	0.900	0.900	0.001

From above we find that by controlling the value of β, we can control the size of OβS. In the study of the actual uncertain dense subgraph, this is achieved by appropriately adjusting the size of the parameter β. Therefore, the OβS model has high scalability in practical applications.

6. Conclusion

This paper extensively understands the modern dense subgraphs and reliable subgraph mining methods of uncertain graphs. Based on the characteristics of uncertain graphs and requirements of dense subgraphs, a novel concept of optimal β-subgraph is proposed. A greedy algorithm is also devised to approximately mine the optimal β-subgraph. Experiments demonstrate that the optimal β-subgraph has many advantages over previous dense subgraph models. Therefore, the algorithm has broad practical applications.

References

[1] Sozio M, Gionis A, "The community-search problem and how to plan a successful cocktail party," in *Proc. of the 16th ACM SIGKDD Int Conf on Knowledge discovery and data mining*, Washington: ACM, pp.939-948, July 25-28, 2010. Article (CrossRef Link).

[2] Fratkin E, Naughton B T, Brutlag D L, and Batzoglou S. Motifcut, "MotifCut: regulatory motifs finding with maximum density subgraphs," *Bioinformatics*, vol. 22, no. 14, pp.150-157, July, 2006. Article (CrossRef Link).

[3] Angel A, Sarkas N, Koudas N, and Srivastava D, "Dense subgraph maintenance under streaming edge weight updates for real-time story identication," *VLDB Endowment*, vol. 5, no. 6, pp. 574-585, February, 2012. Article (CrossRef Link).

[4] Gao X, Gao Y, "Connectedness Index of Uncertain Graphs," *International Journal of Uncertainty,Fuzziness and Knowledge-Based Systems*, vol. 21, no. 1, pp. 127-137, 2013. Article (CrossRef Link).

[5] Szklarczyk D, Franceschini A, et al, "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43,pp. 447-452, January 28, 2015. Article (CrossRef Link).

[6] Rual J F, Venkatesan K, Hao T, et al, "Towards a proteome-scale map of the human protein-protein interacrion network," *Nature*, vol. 437, no. 7062, pp. 1173-1178, October 20, 2005. Article (CrossRef Link).

[7] Wang W, "Emergence of a DNA-damage response network consisting of Fanconianaemia and BRCA proteins," *Nature Reviews Genetics*, vol. 8, no. 10, pp. 735-748, 2007. Article (CrossRef Link).

[8] Zou Z, "Polynomial-time algorithm for finding densest subgraphs in uncertain graphs," in *Proc. of the 11th workshop on mining and learning with graph*, Chicago: MLG, August 11, 2013. Article（CrossRef Link).

[9] Zou Z, Li J, Gao H, et al, "Finding top-k maximal cliques in an uncertain graph," in *Proc. of the 26th International Conference on Data Enginering*, California, USA, pp. 649-652, March 1-6, 2010.

[10] Zou Zhaonian, Zhu Rong, "Mining Top-k maximal cliques from large uncertain graphs," *Chinese Journal of Computers*, vol. 36, no. 10, pp. 2146-2155, 2013. (in Chinese). Article (CrossRef Link).

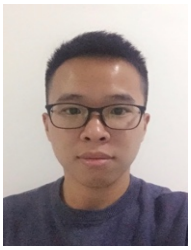
[11] Jin R, Liu L, Aggarwal C C, "Discovering highly reliable subgraphs in uncertain graphs," in *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego:ACM, pp. 992-1000, August 21-24, 2011. Article (CrossRef Link).

[12] Cheng J, Ke Y, Fu A W C, et al, "Finding maximal cliques in massive networks by h*-graph," in *Proc. of the 2010 ACM SIGMOD Int Conf on Management of Data*, Indianapolis: ACM, pp.

447-458, June 6-10, 2010. Article (CrossRef Link).
[13] Tsourakakis C, Bonchi F, Gionis A, Gullo F, Tsiarli M, "Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees," in *Proc. of the 19th ACM SIGKDD Int Conf on Knowledge discovery and data mining*, Chicago: ACM, pp. 104-112, August 11-14, 2013. Article (CrossRef Link).
[14] Goldberg A V, "Finding a maximum density subgraph," *CA: University of California at Berkeley*, 1984.
[15] Charikar M, "Greedy approximation algorithms for finding dense components in a graph," in *Proc. of the 3rd Int Workshop on Approximation Algorithms for Combinatorial Optimization*, Saarbrucken: ACM, pp. 84-95, September 5-8, 2000. Article (CrossRef Link).
[16] Andersen R, Chellapilla K, "Finding Dense Subgraphs with Size Bounds," in *Proc. of the International Work on Algorithms and Models for the Web-graph (WAW 2009)*, Barcelona, Spain, pp. 25-37, Feb. 12-13, 2009. Article (CrossRef Link).
[17] Khuller S, Saha B, "On Finding Dense Subgraphs," in *Proc. of ICALP 2009*, Rhodes, Greece, pp. 597-608, July 5-12, 2009. Article (CrossRef Link).
[18] Bahmani B, Kumar R, Vassilvitskii S, "Densest subgraph in streaming and MapReduce," in *Proc. of the VLDB Endowment*, vol.5, no. 5, pp. 454-465, 2012. Article (CrossRef Link).
[19] Epasto A, Lattanzi S, Sozio M, "Efficient Densest Subgraph Computation in Evolving Graphs," in *Proc. of International Conference on World Wide Web, International World Wide Web Conferences Steering Committee*, pp. 300-310, May 18-22, 2015. Article (CrossRef Link).
[20] Seidman S B, "Network structure and minimum degree," *Social Networks*, pp. 269–287, vol. 5, no. 3, 1983. Article (CrossRef Link).
[21] Tsourakakis C, "The k-cliques densest subgraph problem," in *Proc. of the 24th Int Conf on World Wide Web*, Florence: ACM, pp. 1122-1132, May 18-22, 2015. Article (CrossRef Link).
[22] Bhattacharya S, Henzinger M, Nanongkai D, et al, "Space- and Time-Efficient Algorithm for Maintaining Dense Subgraphs on One-Pass Dynamic Streams," in *Proc. of 47th ACM Symposium on Theory of Computing*, pp. 173-182, June 14-17, 2015. Article (CrossRef Link).
[23] XIULIAN GAO, YUAN GAO, "CONNECTEDNESS INDEX OF UNCERTAIN GRAPH," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 21, no. 1, pp. 127-137, 2013. Article (CrossRef Link).
[24] Zou Z, Li J, Gao H, et al, "Mining Frequent Subgraph Patterns from Uncertain Graph Data," *IEEE Transactions on Knowledge & Data Engineering*, vol. 22, no. 9, pp. 1203-1218, 2010. Article (CrossRef Link).
[25] Jin R, Liu L, Ding B, et al, "Distance-constraint reachability computation in uncertain graphs," in *Proc. of the VLDB Endowment (PVLDB 2011)*, vol. 4, no. 9, pp. 551-562, 2011. Article (CrossRef Link).
[26] Kollios G, Potamias M, Terzi E, "Clustering Large Probabilistic Graphs," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 2, pp. 325-336, 2013. Article (CrossRef Link).
[27] Bonchi F, Gullo F, Kaltenbrunner A, Volkovich Y, "Core de-composition of uncertain graphs," in *Proc. of the 20th ACM SIGKDD Int Conf on Knowledge discovery and data mining*, New York: ACM, pp. 1316-1325, August 24 – 27, 2014. Article (CrossRef Link).
[28] Provo A, Xu P, Tirthapura S, "Enumeration of maximal cliques from an uncertain graph," *IEEE Trans on Knowledge and Data Engineering*, vol. 29, no. 3, pp. 543-555, 2017. Article (CrossRef Link).
[29] Zhu R, Zou Zhaonian, Li Jianzhong, "Mining top-k dense subgraphs from uncertain graphs," *Chinese Journal of Computers*, vol. 39, no. 8, pp. 1570-1582, 2016. (in Chinese)
[30] Gao Yuan, "Uncertain Graph and Uncertain Network," *Doctor-Tsinghua University*, Beijing, 2013. (in Chinese) Article（CrossRef Link).

Lu Yihong received her master degree in 2003 from Hangzhou Institute of Electronics Engineering. Now she is an associate professor in College of Computer Science & Technology, Zhejiang University of Technology. Her main research interests include software theory and data mining.

E-mail：lyh@zjut.edu.cn



Huang Ruizhi receicved his master degree in Computer Science & Technology from Zhejiang University of Technology in 2018. His main research interests include graph mining and data mining

E-mail：huanggw1@gmail.com

Huang Decai received his doctor's degree in 1994 from Chongqing University. Now he is a professor in College of Computer Science & Technology, Zhejiang University of Technology. His main research interests include data mining and software theory.

E-mail：hdc@zjut.edu.cn