**A Method for Finding the Highly Reliable Densest Subgraph from an Uncertain Weighted Graph**

Duong Quoc Anh Kiet, Thien Nguyen

Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Corresponding author

Email: [nguyenchithien@tdtu.edu.vn](mailto:nguyenchithien@tdtu.edu.vn)

# Abstract

Đồ thị đã trở thành công cụ mô hình hóa quan trọng cho nhiều loại dữ liệu, nhưng thường chịu ảnh hưởng bởi sự không chắc chắn do nhiễu hoặc không đầy đủ trong quá trình thu thập dữ liệu. Điều này làm tăng tính quan trọng đối với việc khai thác đồ thị con không chắc chắn, đặc biệt là trong trường hợp các đồ thị dày đặc và đáng tin cậy. Trong bối cảnh không chắc chắn, đồ thị con dày đặc thường có mật độ kỳ vọng và độ tin cậy thấp do xác suất tồn tại giữa các cạnh thấp. Hơn nữa, mô hình đồ thị không chắc chắn hiện tại thường không xem xét đến yếu tố trọng số của đồ thị, điều này là không chính xác trong thực tiễn. Do đó, bài báo này đề xuất một mô hình mới cho đồ thị có trọng số không chắc chắn, để khai thác đồ thị con có mật độ và độ tin cậy cao, bài báo còn đề xuất khái niệm về đồ thị con được lọc cao tần (high-pass filter subgraph). Một thuật toán tham lam cũng được phát triển để tìm kiếm đồ thị con trong bối cảnh đồ thị có trọng số không chắc chắn. Cuối cùng, qua thực nghiệm trên bộ dữ liệu, chúng tôi chứng minh được hiệu suất trong việc tìm kiếm đồ thị con có trọng số không chắc chắn dày đặc và đáng tin cậy.

***Keywords:*** uncertain graph, network reliability, surplus average degree, graph mining, weighted graph, uncertain weighted graph

# Introduction

Một phần quan trọng của khoa học tính toán là lý thuyết đồ thị, nghiên cứu về các đặc điểm của đồ thị để mô hình hóa các hệ thống phức tạp. Các nhà nghiên cứu thường tập trung vào phân tích để tìm kiếm những đồ thị con dày đặc từ các đồ thị có kích thước lớn. Một khái niệm quan trọng trong số các khái niệm cơ bản của lĩnh vực này là đồ thị con dày đặc nhất, được định nghĩa là đồ thị con có mật độ cạnh cao nhất, tức là số cạnh có thể có trong đồ thị con nhiều nhất [1]. Việc xác định các đồ thị con dày đặc có vai trò thiết yếu trong nhiều ứng dụng, chẳng hạn như sinh học hoặc phân tích mạng xã hội, đòi hỏi phải xác định các đồ thị con dày đặc. Tuy nhiên, việc thêm yếu tố không chắc chắn vào mô hình đồ thị gây ra nhiều vấn đề hơn. Điều này đặc biệt đúng đối với các đồ thị vô hướng có trọng số không chắc chắn.

Nghiên cứu này sẽ tập trung vào đồ thị vô hướng, nơi các cạnh không có hướng cụ thể tạo ra các đặc điểm và khó khăn riêng biệt. Loại đồ thị mà sự tồn tại của các cạnh mang tính xác suất, đưa yếu tố không chắc chắn vào các kết nối của cạnh được gọi là đồ thị vô hướng không chắc chắn. Bằng cách gắn trọng số cho các cạnh, cả sự tồn tại của cạnh đều được biểu thị bằng trọng số không chắc chắn, khái niệm đồ thị có trọng số không chắc chắn được mở rộng. Nhiễu hoặc quá trình thu thập thông tin không đầy đủ là một trong nhiều nguyên nhân khiến sự không chắc chắn này xảy ra. Các đồ thị Protein-Protein Interaction (PPI) như Rual et al., (2005) đã chỉ ra, nơi mỗi liên kết protein ảnh hưởng đáng kể đến cấu trúc và chức năng sinh học, làm cho sự không chắc chắn này trở nên rõ ràng hơn. Ngoài ra, đồ thị có độ tin cậy cao là những đồ thị có các cạnh kết nối có xác suất cao. Điều này được thực hiện để cân bằng các yếu tố về độ tin cậy và mật độ dày đặc của các đồ thị con, trong đó các cấu trúc quan trọng có thể được lưu trữ để sử dụng với các mục đích khác nhau.

mrr

0.543

0.959

0.575

pmrR

proP

0.575

mcrB

Figure 1 Protein-Protein Network (PPI)

Trọng tâm của nghiên cứu này là phát triển một phương pháp hiệu quả để xác định đồ thị con có đặc tính dày đặc và đáng tin cậy trên đồ thị có trọng số không chắc chắn. Phương pháp của chúng tôi xem xét cả mật độ kết nối của các cạnh và độ tin cậy của chúng, tạo ra sự cân bằng giữa tính dày đặc của đồ thị và độ tin cậy của nó. Điều này rất quan trọng trong việc phân tích các mạng PPI và các ứng dụng liên quan đến tìm kiếm đồ thị con mang dữ liệu quan trọng, vì mật độ và độ tin cậy của đồ thị có thể ảnh hưởng đáng kể đến kết quả của các ứng dụng này. Phương pháp của chúng tôi cung cấp các công cụ giúp giải quyết các vấn đề phức tạp liên quan đến đồ thị có trọng số không chắc chắn, và chúng tôi tin rằng nó sẽ đóng góp tích cực cho lĩnh vực phân tích đồ thị, trong khi các phương pháp truyền thống không đủ hiệu quả trong trường hợp này.

# Problem statement

Một thách thức đáng kể trong lĩnh vực lý thuyết đồ thị là tìm ra đồ thị con dày đặc từ đồ thị có trọng số không chắc chắn. Đồ thị con dày đặc, nơi các đỉnh liên kết chặt chẽ, thường chứa thông tin quan trọng về đặc điểm của đồ thị. Do đó, việc sử dụng các đồ thị con dày đặc từ các đồ thị lớn là một nhiệm vụ quan trọng trong nghiên cứu đồ thị. Nhưng việc lấy đồ thị con dày đặc và đáng tin cậy từ môi trường không chắc chắn, đặc biệt khi xác suất tồn tại của mỗi cạnh khác nhau, đòi hỏi phải có những phương pháp phân tích cải tiến. Ngoài ra, việc coi các trọng số của một cạnh bằng nhau thường không cho thấy sự phức tạp của mối quan hệ giữa chúng. Do đó, một phương pháp mới phải được sử dụng để đưa trọng số, một chỉ số thể hiện sức mạnh liên kết của mối quan hệ vào mô hình đồ thị. Đồ thị con dày đặc trong protein thường cho thấy mức độ phức tạp của protein, điều này cho thấy các protein tương tác với nhau để thực hiện các chức năng nhất định. Phức hợp Fanconi protein , ví dụ, có liên quan đến việc sữa chữa DNA. Do đó, việc sử dụng các đồ thị con dày đặc trong các đồ thị không chắc chắn có trọng số là cần thiết để nhận dạng và dự đoán các phức hợp protein chưa từng được biết tới trước đó.

Goldberg (1984) cho biết độ dày đặc của một đồ thị với là . Mục tiêu là xác định tập hợp con các đỉnh S sao cho có độ dày đặc cao nhất. [3] lần đầu định nghĩa về mật độ mong đợi của đồ thị không chắc chắn và viết lại vấn đề cụ thể về việc tìm đồ thị con dày đặc trong đồ thị không chắc chắn. Trong đồ thị không chắc chắn, các cạnh và trọng số có tính xác suất. Điều này khiến việc xác định đồ thị con dày đặc và đáng tin cậy trở nên khó khăn. [4] đề cập đến độ tin cậy của đồ thị không chắc chắn bằng cách xem xét độ tin cậy của đồ thị con. Jin et al., 2011 đã nêu ra khái niệm về độ tin cậy trong đồ thị không chắc chắn, và sử dụng thuật toán sampling-based để khai thác đồ thị con có độ tin cậy cao trong đồ thị không chắc chắn. Các đồ thị con được tìm thấy bằng cách này có một số vấn đề. Ví dụ, đồ thị con không chắc chắn A và B có mật độ mong đợi là 0,7. Rõ ràng là chúng ta không thể so sánh hai đồ thị dựa trên mật độ mong đợi. Độ tin cậy của đồ thị con được sử dụng để đo độ tin cậy của đồ thị không chắc chắn. Đồ thị B, do xác suất cạnh thấp hơn, có độ tin cậy thấp hơn so với đồ thị A trong Figure 2.. Do đó, độ tin cậy của đồ thị con thường chỉ tập trung vào độ tin cậy thay vì độ dày đặc. [5] đề xuất một thuật toán đồ thị con beta và một thuật toán đồ thị con tối ưu để tìm đồ thị con dày đặc có độ tin cậy cao và đạt được kết quả tích cực cho nhiều ứng dụng thực tế. Điều này nhằm giải quyết vấn đề không thể cân bằng được mật độ và độ tin cậy của, nhưng bài báo không xem xét các trọng số không chắc chắn trong đồ thị vì nó giả định rằng các trọng số là bằng nhau.

A diagram of a number and a number

Description automatically generated with medium confidence

Figure 2 Mật độ và độ tin cậy của đồ thị

Nghiên cứu này nhằm phát triển một phương pháp mới để xác định đồ thị con dày đặc và đáng tin cậy trong đồ thị có trọng số không chắc chắn, mục tiêu của chúng tôi là tạo ra một phương pháp có thể giải quyết cả hai yếu tố này đồng thời cung cấp một công cụ đáng tin cậy và hiệu quả để phân tích đồ thị trong môi trường không chắc chắn.. Dựa trên cơ sở lý thuyết đã được thiết lập [6] cùng với các tiến bộ gần đây [[4], [7], [5]], chúng tôi hy vọng phương pháp mới sẽ cung cấp một công cụ mạnh mẽ cho việc phân tích đồ thị trong các ứng dụng thực tế hiện đại.

# Related works

Các nhà nghiên cứu rất quan tâm đến việc xác định đồ thị con dày đặc và đã tập trung vào các mô hình đồ thị và phương pháp khám phá đồ thị con dày đặc trong môi trường không chắc chắn. Xác suất cạnh của đồ thị làm cho việc xác định các đồ thị con dày đặc và có độ tin cậy cao trở nên khó khăn hơn. Nghiên cứu của Goldberg (1984) về định nghĩa của mật độ đồ thị và những đóng góp của Charikar, (2000) về thuật toán xấp xỉ tham lam cung cấp kiến thức cơ bản về mật độ đồ thị trong môi trường chắc chắn, ngoài ra thuật toán tham lam dựa theo mật độ tối đa của đồ thị con được đề xuất để giải quyết vấn đề tìm đồ thị con dày đặc như trong nghiên cứu của [8]và [9] mật độ tối đa được mở rộng bằng cách giới hạn kích thước của đồ thị con nhưng đây lại là vấn đề thuộc NP-hard, Map-Reduce của và các kỹ thuật khác cũng được áp dụng để nghiên cứu bài toán tìm đồ thị con có mật độ cao, tuy nhiên việc chuyển qua các khái niệm này qua môi trường không chắc chắn lại phức tạp hơn rất nhiều.

Nghiên cứu về đồ thị không chắc chắn đã bắt đầu được nghiên cứu bởi [7] và đã phát triển thành một lĩnh vực nghiên cứu rộng rãi, các nghiên cứu này đưa ra nhiều thuật toán và mô hình khác nhau như Jin et al., (2011) về độ tin cậy của đồ thị, khai thác mẫu xuất hiện thường xuyên như Z. Zou et al., (2010) và gom cụm đồ thị của Kollios et al., (2013) . Đáng chú ý hơn cả là khái niệm về maximal clique trong đồ thị không chắc chắn như được đề xuất trong công trình nghiên cứu của [3], đồ thị con có mật độ cạnh tối đa là một được gọi là clique và một clique có nhiều đỉnh nhất gọi là maximal clique, maximal clique được coi là một vùng dày đặc nhất trong đồ thị, tuy nhiên việc tìm maximal clique là một NP-hard và có nhiều đồ thị không có clique, do đó quasi-clique được phát triển để giải quyết vấn đề NP-hard.

Ứng dụng của việc khám phá đồ thị con dày đặc không chỉ dừng lại ở lý thuyết mà còn mở rộng sang các ứng dụng thức tế như ứng dụng trong sinh học Protein-Protein Interaction Rual et al., (2005). Hơn thế nữa sự phát triển của kỹ thuật khai thác đồ thị con dày đặc trong môi trường không chắc chắn được mô tả rõ ràng bởi Jin et al., (2011) giới thiệu một chiều hướng mới cho sự hiểu biết của chúng ta về đồ thị con dày đặc. Thêm vào đó Lu et al., (2019) đã đề xuất một phương pháp sử dụng tham số beta như một ngưỡng để tìm đồ thị con có mật độ dày đặc và có độ tin cậy cao, tuy nhiên, họ giả định rằng các cạnh có trọng số bằng nhau, điều mà các nhà nghiên cứu thường giả định, nhưng điều này không đúng. Công trình của [12] đã đặt nền tảng cho lĩnh vực phát hiện đồ thị con dày đặc trong các đồ thị không chắc chắn và có trọng số. Ngoài ra bài nghiên cứu của Cheng et al. 2014 phát triển thuật toán về tìm đường đi ngắn nhất cũng đưa ra cái nhìn về đồ thị có trọng số không chắc chắn, là cơ sở để biểu diễn đồ thị cho bài nghiên cứu của chúng tôi. Để xác định các đồ thị con dày đặc trong các đồ thị có trọng số cạnh không chắc chắn, họ cung cấp các phương pháp nghiên cứu tiên tiến. Chúng tôi muốn mở rộng những phương pháp này để nghiên cứu sự phụ thuộc lẫn nhau giữa trọng số và xác suất. Chúng có thể ảnh hưởng đến độ tin cậy của đồ thị không chắc chắn cũng như mật độ dày đặc của nó. Do đó bài nghiên cứu của chúng tôi dựa vào đó để mở rộng cho bài toán tìm đồ thị con dày đặc có độ tin cậy cao cho đồ thị có trọng số không chắc chắn.

A diagram of a network

Description automatically generated

Figure 3 Uncertain weighted graph

# Preliminaries and definitions

Trong phần này, chúng tôi sẽ giới thiệu một số kí hiệu và định nghĩa. Các công thức toán học này sẽ được sử dụng xuyên suốt bài báo này.

|  |  |
| --- | --- |
| Notion | Definition |
| **G** | Đồ thị có trọng số không chắc chắn |
|  | Tập hợp các đỉnh trong đồ thị |
|  | Tập hợp các cạnh trong đồ thị |
|  | Cạnh giữa và có trong đồ thị |
|  | Trọng số của cạnh E có trong đồ thị |
|  | Xác suất không chắc chắn của trọng số của cạnh E có trọng số bằng giá trị |
|  | Số lượng đỉnh có trong đồ thị |
|  | Số lượng cạnh có trong đồ thị |
|  | Số lượng cạnh tối đa của đồ thị |
|  | Trung bình trọng số của đồ thị |

Table 1 Notion and definition

**Định nghĩa 1** : (Uncertain Weighted Graph) Đồ thị có trọng số không chắc chắn được biểu diễn qua biến ngẫu nhiên , với tập là tập tất cả các đỉnh tất định, tập biến ngẫu nhiên là tập tất cả các cạnh, trong đó mỗi cạnh là một biến ngẫu nhiêncó trọng số Wt(E) tuân theo phân phối Bernouli, cụ thể hoặc . Gọi một tập đỉnh , đồ thị con được sinh ra từ đồ thị có trọng số không chắc chắn trong đó **.** Trong bài báo này, chúng tôi giả định các cạnh là những biến ngẫu nhiên độc lập với nhau từng đôi một.

**Ví dụ** : Trong Figure 3 là một đồ thị không chắc chắn có trọng số, cạnh có trọng số là và xác suất

**Định nghĩa 2** : (Adjoint Weighted Graph) Đồ thị có trọng số liên kết được xác định với là tập tất cả các đỉnh tất định, tập biến ngẫu nhiên là tập tất cả các cạnh, trong đó mỗi cạnh là một biến ngẫu nhiên biết rằng của tất cả các cạnh bằng của nó bằng 1.

**Ví dụ** : Trong Figure 3 giả sử tất cả xác suất của các cạnh bằng 1, nơi các cạnh có liên kết với nhau chỉ phụ thuộc vào trọng số của nó.

**Định nghỉa 3** : (Adjoint Logarit Reliability) Độ tin cậy logarit liên kết của đồ thị không chắc chắn có trọng số , được phát triển từ độ tin cậy liên kết (adjoint reliability) của [5] việc sử dụng logarit cơ số 10 giúp cho việc tính toán và biểu diễn tích của nhiều số floating point number khả thi trên máy tính, do đó được định nghĩa như sau :

**Ví dụ** : Trong Figure 3 ta có độ tin cậy logarit của đồ thị bằng .

**Định nghĩa 4** : (Average edge weighted probability) Xác suất trọng số cạnh trung bình của đồ thị có trọng số không chắc chắn là :

**Ví dụ** : Trong Figure 3 ta có xác suất trọng số cạnh trung bình của đồ thị .

**Định nghĩa 5** : (Edge Weighted Probability Standard Deviation) Độ lệch chuẩn của xác suất trọng số cạnh của đồ thị có trọng số không chắc chắn :

**Ví dụ** : Trong Figure 3 ta có độ lệch chuẩn của xác suất trọng số cạnh đồ thị là

**Định nghĩa 6** : (Expected Edge Density) Mật độ cạnh kì vọng của đồ thị con có trọng số không chắc chắn được sinh ra từ đồ thị có trọng số không chắc chắn định nghĩa như sau :

**Ví dụ** : Trong Figure 3 ta có mật độ cạnh kì vọng của đồ thị .

**Định lí 1** : Đồ thị có trọng số không chắc chắn được biểu diễn qua biến ngẫu nhiên , với tập là tập tất cả các đỉnh tất định, tập biến ngẫu nhiên là tập tất cả các cạnh, trong đó mỗi cạnh là một biến ngẫu nhiêncó trọng số Wt(E) tuân theo phân phối Bernouli, cụ thể hoặc . Gọi một tập đỉnh , đồ thị con được sinh ra từ đồ thị có trọng số không chắc chắn trong đó , mật độ cạnh kì vọng của sẽ tăng lên khi xác suất trọng số cạnh trung bình của nó tăng lên.

**Chứng minh** :

⟺

Với đồ thị con  có số lượng cạnh và đỉnh đã được xác định, thì mật độ cạnh kì vọng sẽ tăng lên cùng với xác suất trọng số cạnh trung bình của đồ thị.

# Propose Methods

Chúng tôi đã phát triển hai thuật toán chính, Greedy Uncertain Weighted Dense Subgraph (GreedyUWDS) và Greedy Bound Weighted Dense Subgraph (GreedyBWDS), để tìm kiếm đồ thị con dày đặc và đáng tin cậy từ đồ thị không chắc chắn có trọng số. Mỗi thuật toán được thiết kế để giải quyết một phần cụ thể của vấn đề tìm kiếm đồ thị con dày đặc và đáng tin cậy. Ngoài ra, chúng dựa trên các giả định và tiêu chuẩn đánh giá khác nhau để phù hợp với các mục tiêu đa dạng của phân tích đồ thị.

Thuật toán GreedyUWDS được thiết kế để giải quyết sự không chắc chắn về sự tồn tại của các cạnh và tối ưu hóa mật độ của đồ thị con. GreedyUWDS sử dụng cơ sở của thuật toán GreedyUDS, nhưng nó đưa ra các thay đổi trong đồ thị liên quan đến các trọng số không chắc chắn. Thuật toán GreedyUWDS, dựa trên thông số mật độ kì vọng, giúp xác định các khu vực dày đặc có độ tin cậy cao trong đồ thị không chắc chắn. Ngoài ra, nó cung cấp thông tin chi tiết về cấu trúc và mối quan hệ giữa các đối tượng trong đồ thị. Tương tự như Thuật toán GreedyBWDS cũng được tạo ra để tìm đồ thị con dày đặc và có độ tin cậy cao, nhưng nó cũng sử dụng một tham số để kiểm soát phân tích đồ thị con. Theo kết quả của thử nghiệm, phương pháp này có thể được áp dụng cho nhiều ứng dụng khác nhau, Ngoài ra, cách tiếp cận này nhấn mạnh việc đánh giá và lựa chọn các cạnh dựa trên xác suất tồn tại.

Bên cạnh những phương pháp cải tiến này chúng tôi cũng muốn đề cập đến một phương pháp truyền thống để tìm đồ thị con dày đặc và đáng tin cậy dựa theo thông số mật độ kì vọng như thuật toán GreedyUWDS. Phương pháp brute force đơn giản là thử tất cả các khả năng có thể có trong đồ thị và dựa trên mật độ kì vọng để chọn đồ thị con tốt nhất. Brute force là một phương pháp tham chiếu quan trọng để so sánh hiệu suất của các thuật toán tiên tiến như GreedyUWDS và GreedyBWDS, tuy nhiên nó không hiệu quả với các đồ thị lớn hay có thể nói là không nên sử dụng cho những đồ thị phức tạp. GreedyUWDS và GreedyBWDS đã mang lại những tiến bộ đáng kể trong việc giải quyết các vấn đề khó khăn của việc tìm kiếm đồ thị con dày đặc trong môi trường đồ thị không chắc chắn. Cả hai phương pháp đều cho phép hiểu biết sâu sắc hơn về mật độ và độ tin cậy của các đồ thị không chắc chắn. Điều này mở ra những cách mới để tiến hành nghiên cứu và sử dụng trong tương lai.

## **Thuật toán Brute Force**

Trong việc tìm kiếm đồ thị con dày đặc từ đồ thị không chắc chắn có trọng số, phương pháp brute force truyền thống đóng vai trò là một kỹ thuật cơ bản nhưng quan trọng. Phương pháp này thực hiện bằng cách xem xét mọi tập hợp con có thể của các đỉnh và cạnh trong đồ thị, đánh giá từng tập hợp này dựa trên mật độ kỳ vọng - một tiêu chí dựa trên tổng trọng số của các cạnh và xác suất tồn tại của chúng. Mặc dù brute force đảm bảo tìm ra kết quả tối ưu bằng cách đánh giá toàn diện tất cả các khả năng, nhưng phương pháp này lại đối mặt với hạn chế lớn về độ phức tạp tính toán. Đặc biệt trong các đồ thị lớn, số lượng tập hợp con cần xét tăng lên nhanh chóng, khiến cho thời gian chạy trở nên không khả thi. Do đó, dù là một phương pháp tham chiếu quan trọng, brute force thường chỉ được áp dụng trong lý thuyết hoặc như một phương pháp so sánh trong các trường hợp đồ thị có kích thước nhỏ, và nó thường không được khuyến nghị sử dụng cho các đồ thị lớn và phức tạp do giới hạn về hiệu suất. Chính vì vậy hai thuật toán còn lại sẽ là phương pháp để tìm kiếm đồ thị con dày đặc và đáng tin cậy từ đồ thị có trọng số không chắc chắn, trong đó thuật toán GreedyBWDS lại còn có thể kiểm soát được quá trình tìm kiếm của đồ thị con.

## **Thuật toán GreedyUWDS**

**Định nghĩa 7** : (Expected Degree) Độ kì vọng của đỉnh của đồ thị con có trọng số không chắc chắn  được sinh ra từ đồ thị

**Ví dụ** : Trong Figure 3 ta có độ kì vọng của đỉnh B là .

**Định nghĩa 8** : (Expected Density) Đồ thị con có trọng số không chắc chắn được sinh ra từ đồ thị có mật độ kì vọng được định nghĩa như sau :

**Ví dụ** : Trong Figure 3 ta có mật độ kì vọng của đồ thị không chắc chắn có trọng số là .

Định nghĩa 9 :  Đồ thị không chắc chắn có trọng số  được xác định với là tập hợp các đỉnh, **E** là tập hợp các cạnh, trong đó mỗi cạnh là một biến ngẫu nhiêncó trọng số Wt(E) tuân theo phân phối Bernouli, cụ thể hoặc . Gọi một tập đỉnh , đồ thị con được sinh ra từ đồ thị có trọng số không chắc chắn trong đó **.** Trong bài báo này, chúng tôi giả định các cạnh là những biến ngẫu nhiên độc lập với nhau từng đôi một có được gọi là đồ thị con có trọng số dày đặc nhất, hoặc UWDS.

Thuật toán GreedyUWDS được chúng tôi giới thiệu trong bài báo này là một biến thể khác từ ý tưởng minimum cut trong [6] và dựa theo thuật toán Charikar. Có điều chỉnh một vài thuộc tính có liên quan để phù hợp với đồ thị không chắc chắn có trọng số. Chi tiết thuật toán được trình bày qua pseudocode trong thuật toán 1

**Thuật toán 1 : GreedyUWDS**

Input : Uncertain Weighted Graph có và với

Ouput : Tập hợp đỉnh của đồ thị con tìm được có expected density cao nhất

1. Khởi tạo best\_subgraph và best\_expected\_density

2. for i = n lần lượt thực hiện các bước sau đến khi i = 2 :

3. Tính expected density hiện tại của tập đỉnh

4. Cập nhật best\_subgraph và best\_expected\_density nếu cần

5. Tính expected degree của từng đỉnh và loại bỏ đỉnh có giá trị expected degree nhỏ nhất khỏi đồ thị

6. end for

7. Trả về best\_subgraph và best\_expected\_density

Thuật toán tính toán mật độ kì vọng của đồ thị hiện tại trong suốt quá trình lặp và xóa các đỉnh có độ kì vọng thấp nhất cho đến khi số đỉnh còn lại trong đồ thị chỉ còn hai và cuối cùng đầu ra của thuật toán là đồ thị con dày đặc đáng tin cậy có mật độ kì vọng cao nhất. Độ phức tạp về thời gian của thuật toán là O(m+n), nơi mà n là số lượng đỉnh và m là số lượng cạnh của đồ thị không chắc chắn có trọng số.

## **Thuật toán GreedyBWDS**

**Định nghĩa 10** : (Surplus Degree) Cho một đỉnh và một cạnh của một đồ thị có trọng số không chắc chắn và một giá trị , bậc dư của đỉnh được định nghĩa là tổng của các tích giữa xác suất trọng số trừ cho giá trị và trọng số cạnh liên kết với đỉnh đó.

**Ví dụ** : Trong Figure 3 giả sử ta có thì độ dư của đỉnh B là

**Định nghĩa 11** : (Surplus Average Degree) Cho một đồ thị có trọng số không chắc chắn , một giá trị với là một đồ thị con được sinh ra từ tập con , thì độ dư trung bình của đồ thị được định nghĩa như sau :

**Ví dụ** : Trong Figure 3 giả sử ta có thì độ dư trung bình của đồ thị là

**Định nghĩa 10** : (High-pass Filtered Subgraph) Cho một đồ thị có trọng số không chắc chắn , một giá trị và là một đồ thị con được sinh ra từ đồ thị có các tập con sao cho:

Chúng tôi gọi đồ thị con là đồ thị con được lọc cao tần (High-pass Filtered Subgraph) và sử dụng tham số làm ngưỡng để lọc.

Thuật toán GreedyBWDS cùng chung ý tưởng với thuật toán GreedyUWDS nhưng có sử dụng tham số để có thể kiếm soát việc tìm kiếm đồng thời sử dụng độ dư trung bình để xác định đồ thị con dày đặc đáng tin cậy và tính toán đỉnh có độ dư thấp nhất để loại bỏ cho đến khi số đỉnh còn lại trong đồ thị còn lại hai đỉnh và đầu ra của thuật toán là đồ thị con dày đặc đáng tin cậy có độ dư trung bình cao nhất. Đồng thời để tối ưu thuật toán chúng tôi còn sử dụng min-heap để có thể luôn lấy ra được giá trị độ dư thấp nhất từ đỉnh, tuy nhiên min-heap từ python không thể chỉnh sửa các giá trị đã được thêm vào heap nên chúng tôi sử dụng một thủ thuật nhỏ để đánh dấu và kiểm soát các đỉnh cũ và đỉnh mới được thêm vào heap. Chi tiết thuật toán được trình bày trong thuật toán 2.

**Thuật toán 2 GreedyBWDS**

Input : Uncertain Weighted Graph có và với và tham số

Ouput : Tập hợp đỉnh của đồ thị con tìm được có surplus average degree cao nhất

1. Khởi tạo best\_subgraph và best\_surplus\_avg\_degree

2. Khởi tạo minHeap và vertex\_to\_marker

3. Duyệt qua mỗi đỉnh v trong đồ thị:

4. Tính surplus\_degree của v

5. Thêm giá trị (surplus\_degree, v, marker hiện tại) vào minHeap

6. Cập nhật vertex\_to\_marker[v] = marker hiện tại

7. for i = n thực hiện các bước sau tới khi i = 2:

8. Tính surplus\_avg\_degree hiện tại của tập đỉnh

9. Cập nhật bestSubgraph và best\_surplus\_avg\_degree nếu cần

10. for len(minHeap) > 0:

11. Lấy ra giá trị (surplus\_degree, v\_to\_remove, marker) từ minHeap

12. Nếu marker của v\_to\_remove khớp với vertex\_to\_marker[v\_to\_remove]:

13. Loại bỏ v\_to\_remove khỏi tập đỉnh và dừng vòng lặp

14. end for

15. Lấy tất cả đỉnh kề của v\_to\_remove

16. Loại bỏ v\_to\_remove khỏi đồ thị

17. Tăng marker hiện tại

18. Duyệt qua mỗi đỉnh kề trong đồ thị:

19. Nếu đỉnh có trong tập đỉnh:

20. Tính surplus\_degree mới và thêm vào minHeap

21. Cập nhật vertex\_to\_marker

22. end for

23. end for

24. Trả về best\_subgraph và best\_surplus\_avg\_degree

Thuật toán yêu cầu tham số như một ngưỡng đầu vào, đây cũng là yếu tố chính trong việc kiểm soát quá trình tìm kiếm đồ thị con dày đặc đáng tin cậy từ thuật toán GreedyBWDS. Thuật toán sử dụng độ dư trung bình để đánh giá đồ thị con, và sử dụng min-heap để lưu trữ độ dư của mỗi đỉnh sau đó loại bỏ đỉnh đầu tiên của heap vì có giá trị nhỏ nhất. Cuối cùng đầu ra của thuật toán GreedyBWDS sẽ cho một tập hợp các đỉnh tạo nên đồ thị con dày đặc và có độ tin cậy cao.

## **Algorithm Accuracy**

Giả sử ta có đồ thị có trọng số không chắc chắn , ta tìm được là đồ thị con có độ dư trung bình cao nhất, nghĩa là là đồ thị có phương pháp tốt nhất.

**Định lí 2** : Cho một đồ thị có trọng số không chắc chắn , với bất kì đỉnh nào mà , thì .

**Chứng minh** : Vì là đồ thị có phương pháp tốt nhất, chúng ta đều biết rằng , thì :

⟺

⟺

⟺

**Định lí 3** : Cho đồ thị có trọng số không chắc chắn và tham số , có là đồ thị con được lọc cao tần của , phương pháp tốt nhất, có độ dư trung bình cao nhất. là đồ thị con gần với đồ thị con được lọc cao tần của đồ thị , thu được từ thuật toán GreedyBWDS, nghĩa là nó là một phương pháp xấp xỉ có độ dư trung bình ta có :

Độ dư trung bình của thu được từ thuật toán GreedyBWDS không nhỏ hơn độ dư trung bình của đồ thị con được lọc cao tần , điều này được chứng minh từ thuật toán xấp xỉ của [6].

# Experiment Setup

Dữ liệu protein được sử dụng trong phần này là từ cơ sở dữ liệu STRING-DB (<http://string-db.org>) có tên 579138.protein. Tuy nhiên, vì dữ liệu không có trọng số giữa các cạnh, chúng tôi sẽ lấy các cột thuộc tính từ dữ liệu làm trọng số liên kết của protein. Hai cột đầu tiên trong dữ liệu sẽ là các cột protein đại diện cho các đỉnh và sự liên kết của chúng, "combine\_scored" mô tả xác suất, thể hiện trọng số không chắc chắn về sự liên kết của protein, các cột dữ liệu còn lại sẽ đại diện cho trọng số của các cạnh tương ứng và thay các giá trị bằng hằng số từ dữ liệu sinh học protein. Thời gian chạy của các thuật toán được tính dựa vào trung bình các thời gian chạy riêng biệt. Các kết quả thập phân sẽ được làm tròn tới chữ số thập phân thứ ba và sử dụng tham số mặc định để so sánh của thuật toán GreedyBWDS là 0.6.

Để tìm đồ thị con có mật độ dày đặc có độ tin cậy cao và hiệu quả, chúng tôi sử dụng thuật toán GreedyUWDS và GreedyBWDS để so sánh hiệu suất của hai thuật toán. Sau đó, chúng tôi sẽ so sánh hiệu suất của các trong thuật toán GreedyBWDS để chứng minh rằng tham số này có thể được chỉnh sửa cho nhiều ứng dụng. Bài nghiên cứu này đã sử dụng tất cả thuật toán và thử nghiệm trên Python3.9.16 trên một máy tính thử nghiệm có CPU AMD Ryzen 5 5600H 3.3 GHz, 8GB RAM và Windows 10.

Do sử dụng trên cùng một bộ dữ liệu protein và thay đổi các trọng số giữa các cạnh bằng các cột thuộc tính khác nhau nên sẽ có cùng số lượng đỉnh là và số lượng cạnh là .

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Xác suất trọng số cạnh trung bình | Độ lệch chuẩn của xác suất trọng số cạnh |
| neighborhood | 36.721 | 219.572 |
| fusion | 14.684 | 137.995 |
| cooccurence | 53.865 | 184.996 |
| coexpression | 56.314 | 362.262 |
| experimental | 55.889 | 355.513 |
| database | 26.28 | 333.74 |
| textmining | 35.231 | 91.835 |

Table 2 Tổng quan về bộ dữ liệu sử dụng để thực nghiệm

**Results and discussion**

Trong phần này ta sẽ so sánh kết quả và đánh giá kết quả thu được từ thực nghiệm với bộ dữ liệu protein, có hai phần so sánh, đầu tiên là so sánh hai thuật toán GreedyUWDS và thuật toán GreedyBWDS, phần sau là so sánh thuật toán GreedyBWDS khi sử dụng các khác nhau.

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Thời gian thực thi | |
| UWDS | BWDS |
| neighborhood | 133.731 | 55.951 |
| fusion | 155.252 | 55.381 |
| cooccurence | 158.06 | 68.385 |
| coexpression | 160.907 | 67.853 |
| experimental | 160.892 | 67.039 |
| database | 157.174 | 69.746 |
| textmining | 165.555 | 62.023 |

Table 3 So sánh thời gian thực thi của hai thuật toán

Figure 4 Biểu đồ so sánh thời gian thực thi UWDS và BWDS

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Thuộc tính | Mật độ cạnh kì vọng | | Xác suất trọng số cạnh trung bình | | Độ lệch chuẩn của xác suất trọng số cạnh | | Độ tin cậy logarit liên kết | |
| UWDS | BWDS | UWDS | BWDS | UWDS | BWDS | UWDS | BWDS |
| neighborhood | 388.321 | 616.659 | 438.329 | 575.923 | 215.412 | 176.195 | -209.163 | -8.518 |
| fusion | 35.123 | 640.159 | 32.665 | 654.708 | 279.519 | 143.502 | -4822.897 | -1.874 |
| cooccurence | 508.576 | 1016.148 | 278.939 | 508.074 | 175.252 | 162.742 | -2298.398 | -19.189 |
| coexpression | 1401.044 | 1547.738 | 714.772 | 776.682 | 204.707 | 167.679 | -447.093 | -174.462 |
| experimental | 1445.787 | 1548.561 | 749.406 | 790.524 | 134.945 | 107.097 | -350.654 | -218.437 |
| database | 1350.766 | 1341.696 | 695.726 | 670.848 | 70.619 | 102.233 | -3.826 | -0.648 |
| textmining | 339.654 | 401.997 | 176.414 | 201.095 | 72.708 | 52.356 | -603.928 | -77.534 |

Table 4 So sánh kết quả hai thuật toán GreedyUWDS và GreedyBWDS

Figure 5 Biểu đồ so sánh mật độ cạnh kì vọng của UWDS và BWDS

## **Evaluation Indicators**

Trong quá trình đánh giá hiệu suất của các thuật toán trên dữ liệu đồ thị có trọng số không chắc chắn, ta xem xét một số chỉ số đánh giá quan trọng. Đầu tiên là mật độ cạnh dự kiến đây là một thước đo giúp ta hiểu rõ về mức độ dày đặc của cấu trúc trong đồ thị con mà thuật toán tìm ra. Tiếp theo là độ tin cậy logarit liên kết, chỉ số thể hiện mức độ tin cậy của các mối quan hệ trong đồ thị con. Ngoài ra, ta xét đến xác suất trọng số cạnh trung bình và độ lệch chuẩn của xác suất trọng số cạnh để có cái nhìn tổng quan và sâu sắc hơn về đồ thị con thu được.

Kết quả thử nghiệm cho thấy mối quan hệ tích cực giữa các thông số đánh giá được nêu trên: khi mật độ cạnh kì vọng tăng thì xác suất trọng số cạnh trung bình tăng lên, độ lệch chuẩn giảm xuống, ngoài ra độ tin cậy logarit liên kết cao hơn, cung cấp một đồ thị con có độ tin cậy cao và dày đặc.

## **Model Comparison**

Khi áp dụng thuật toán GreedyUWDS và GreedyBWDS trên cùng một bộ dữ liệu và đặt tham số ở mức mặc định là 0.6, ta nhận thấy rằng GreedyBWDS vượt trội hơn so với GreedyUWDS ở cả hai tiêu chí quan trọng là mật độ cạnh dự kiến và độ tin cậy logarit liên kết. Cụ thể, GreedyBWDS không chỉ tạo ra đồ thị con có mật độ cao hơn mà còn đảm bảo được độ tin cậy cao hơn, cho thấy sức mạnh của nó trong việc tìm ra các đồ thị con dày đặc và đáng tin cậy từ dữ liệu không chắc chắn. Ngoài các thông số đánh giá đã nêu trên chúng tôi còn tính thời gian thực thi của hai thuật toán, kết quả từ Table 3 cho thấy thời gian thực thi của GreedyBWDS hoàn toàn tốt hơn so với thuật toán GreedyUWDS và nhanh hơn xấp xỉ 2.5 lần, điều này cho thấy thuật toán GreedyBWDS ngoài khả năng tìm kiếm đồ thị con dày đặc đáng tin cậy mà còn cho tốc độ chạy nhanh hơn rất nhiều so với thuật toán GreedyUWDS.

Phân tích chi tiết thực nghiệm trên thu được từ :

1/ Mật độ dày đặc của thuật toán GreedyBWDS tốt hơn rất nhiều so với thuật toán GreedyUWDS. Table 3 có trung bình mật độ cạnh kì vọng của GreedyUWDS là 781.324 còn trung bình mật độ cạnh kì vọng của GreedyBWDS là 1037.258. Chính vỉ vậy mật độ dày đặc của thuật toán GreedyBWDS cho kết quả tìm kiếm đồ thị con dày đặc tốt hơn so với thuật toán GreedyUWDS.

2/ Độ tin cậy logarit liên kết của GreedyBWDS cũng được cải thiện rất nhiều so với GreedyUWDS. Table 3 cho thấy tất cả với tất cả thuộc tính trong tập dữ liệu đều cho thấy độ tin cậy logarit liên kết của GreedyBWDS tốt hơn GreedyUWDS.

## **Parameter Selection**

Trong phần này chúng tôi sử dụng hai thuộc tính “experimental” và “text mining” để thử nghiệm và so sánh sự tác động của các tham số khác nhau lên kết quả tìm kiếm đồ thị con dày đặc và đáng tin cậy . Quá trình thực nghiệm được thiết kế để cung cấp cái nhìn sâu rộng về cách các tham số khác nhau có thể tối ưu hóa hiệu suất tìm kiếm và định hình chất lượng của đồ thị con được trích xuất. Điều này không chỉ giúp chúng tôi hiểu rõ hơn về ảnh hưởng của tham số mà còn là nền tảng quan trọng để xác định được rằng thuật toán GreedyBWDS là một phương pháp tiếp cận tối ưu và cách cấu hình tham số của các ứng dụng thực tế khác nhau.

Figure 6 Biểu đồ mật độ cạnh kì vọng các b khác nhau

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Thuộc tính | b | Số lượng cạnh | Mật độ cạnh kì vọng | Xác suất trọng số cạnh trung bình | Độ lệch chuẩn của xác suất trọng số cạnh | Độ tin cậy logarit liên kết | Thời gian thực thi |
| experimental | 0.1 | 2972 | 1445.787 | 749.406 | 134.945 | -350.654 | 120.982 |
| 0.2 | 2906 | 1463.720 | 756.289 | 128.665 | -320.305 | 142.639 |
| 0.3 | 2906 | 1463.720 | 756.289 | 128.665 | -320.305 | 95.102 |
| 0.4 | 2906 | 1463.720 | 756.289 | 128.665 | -320.305 | 77.279 |
| 0.5 | 2771 | 1498.980 | 770.858 | 119.400 | -283.647 | 64.667 |
| 0.6 | 2574 | 1548.561 | 790.524 | 107.097 | -218.437 | 60.415 |
| 0.8 | 1950 | 1692.501 | 847.553 | 77.576 | -64.973 | 55.393 |
| text mining | 0.1 | 2935 | 356.219 | 182.236 | 69.523 | -446.475 | 137.015 |
| 0.2 | 2613 | 377.517 | 189.842 | 65.475 | -254.904 | 134.338 |
| 0.3 | 2477 | 387.107 | 194.179 | 63.047 | -192.838 | 120.862 |
| 0.4 | 2409 | 391.308 | 196.141 | 61.961 | -170.589 | 101.346 |
| 0.5 | 2340 | 395.420 | 198.217 | 60.821 | -150.769 | 82.937 |
| 0.6 | 2079 | 401.997 | 201.095 | 52.356 | -77.534 | 72.327 |
| 0.8 | 1595 | 424.933 | 212.600 | 44.862 | -21.541 | 68.831 |

Table 5 So sánh các tham số b

Figure 7 So sánh thời gian thực thi khi sử dụng các tham số b khác nhau

Qua quá trình thử nghiệm với các giá trị tham số khác nhau, chúng tôi đã rút ra được những nhận xét quan trọng về thuật toán GreedyBWDS. Cụ thể, tham số đã chứng minh là một tham số hiệu quả để điều chỉnh và kiểm soát kết quả của thuật toán. Ví dụ, khi giá trị được đặt ở mức thấp , kết quả thu được từ GreedyBWDS rất gần với những gì thuật toán GreedyUWDS đem lại, phản ánh một mức độ dày đặc và một mức độ tin cậy thấp.Tuy nhiên, khi tăng giá trị của , ta nhận thấy một sự cải thiện đáng kể không chỉ về mật độ của đồ thị con mà còn ở cả xác suất trọng số cạnh trung bình và độ tin cậy. Thêm vào đó thời gian thực thi khi sử dụng các tham số b cũng rất quan trọng, khi tham số b càng lớn thời gian thực thi của thuật toán cũng nhanh hơn. Điều này chỉ ra rằng GreedyBWDS có khả năng thích nghi để phù hợp với các nhu cầu cụ thể trong các ứng dụng thực tế, từ đó mở ra khả năng tùy chỉnh thuật toán một cách linh hoạt dựa trên các yêu cầu đặc thù của từng bộ dữ liệu và mục tiêu nghiên cứu.

Mặc dù nghiên cứu này cung cấp kiến thức chuyên sâu và phương pháp mới để xác định đồ thị con dày đặc và đáng tin cậy từ đồ thị có trọng số không chắc chắn, song vẫn còn một số hạn chế cần được khắc phục. Trước hết, thuật toán của chúng tôi cho thấy hiệu quả với bộ dữ liệu thử nghiệm protein, song khả năng xử lý các loại dữ liệu khác chưa được kiểm chứng. Nghiên cứu của chúng tôi chủ yếu tập trung vào đồ thị vô hướng có trọng số không chắc chắn, song có thể không thể áp dụng trực tiếp cho các loại đồ thị khác hoặc đồ thị có hướng, đặt ra một giới hạn về tính toàn diện của phương pháp.

# Conclusion

Bài báo này nghiên cứu về phương pháp khai thác đồ thị con dày đặc và đáng tin cậy từ đồ thị có trọng số không chắc chắn, thông qua việc phát triển một thuật toán tham lam mới. Thuật toán này sử dụng tham số như một công cụ điều chỉnh ngưỡng để xác định đồ thị con tốt nhất thỏa điều kiện. Qua thử nghiệm thực tiễn, chúng tôi đã chứng minh rằng đồ thị con được tìm ra bằng thuật toán này vượt trội hơn các mô hình trước đó về cả mật độ và độ tin cậy. Sự phát triển này mở ra khả năng ứng dụng rộng rãi trong việc phân tích các hệ thống có cấu trúc đồ thị phức tạp, đánh dấu một bước tiến quan trọng trong nghiên cứu và áp dụng các đồ thị có trọng số không chắc chắn. Trong tương lai, bài báo sẽ mở rộng hơn về khả năng áp dụng các bộ dữ liệu khác nhau để có cái nhìn chính xác hơn về việc áp dụng thuật toán. Bên cạnh đó, chúng tôi kỳ vọng sẽ phát triển phương pháp để có thể tìm thấy nhiều hơn một đồ thị con dày đặc và đáng tin cậy từ đồ thị lớn.

# References

[1] R. Dondi, M. M. Hosseinzadeh, G. Mauri, and I. Zoppis, “Top-k overlapping densest subgraphs: approximation algorithms and computational complexity,” *J Comb Optim*, vol. 41, no. 1, 2021, doi: 10.1007/s10878-020-00664-3.

[2] J. F. Rual *et al.*, “Towards a proteome-scale map of the human protein-protein interaction network,” *Nature*, vol. 437, no. 7062, 2005, doi: 10.1038/nature04209.

[3] Z. N. Zou and R. Zhu, “Mining top-k maximal cliques from large uncertain graphs,” *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 36, no. 10, 2013, doi: 10.3724/SP.J.1016.2013.02146.

[4] R. Jin, L. Liu, and C. C. Aggarwal, “Discovering highly reliable subgraphs in uncertain graphs,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011. doi: 10.1145/2020408.2020569.

[5] Y. Lu, R. Huang, and D. Huang, “Mining highly reliable dense subgraphs from uncertain graphs,” *KSII Transactions on Internet and Information Systems*, vol. 13, no. 6, 2019, doi: 10.3837/tiis.2019.06.012.

[6] M. Charikar, “Greedy approximation algorithms for finding dense components in a graph,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2000. doi: 10.1007/3-540-44436-x\_10.

[7] X. Gao and Y. Gao, “CONnectedness index of uncertain graph,” *International Journal of Uncertainty, Fuzziness and Knowldege-Based Systems*, vol. 21, no. 1, 2013, doi: 10.1142/S0218488513500074.

[8] S. Khuller and B. Saha, “On finding dense subgraphs,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009. doi: 10.1007/978-3-642-02927-1\_50.

[9] R. Andersen and K. Chellapilla, “Finding dense subgraphs with size bounds,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009. doi: 10.1007/978-3-540-95995-3\_3.

[10] Z. Zou, J. Li, H. Gao, and S. Zhang, “Mining frequent subgraph patterns from uncertain graph data,” *IEEE Trans Knowl Data Eng*, vol. 22, no. 9, 2010, doi: 10.1109/TKDE.2010.80.

[11] G. Kollios, M. Potamias, and E. Terzi, “Clustering large probabilistic graphs,” *IEEE Trans Knowl Data Eng*, vol. 25, no. 2, 2013, doi: 10.1109/TKDE.2011.243.

[12] A. Miyauchi and A. Takeda, “Robust Densest Subgraph Discovery,” in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2018. doi: 10.1109/ICDM.2018.00157.

[13] Y. Cheng, Y. Yuan, G. Wang, B. Qiao, and Z. Wang, “Efficient sampling methods for shortest path query over uncertain graphs,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. doi: 10.1007/978-3-319-05813-9\_9.