Paper “Mining Highly Reliable Dense Subgraphs from Uncertain Graphs”

1/ Tóm tắt

Bài báo "Mining Highly Reliable Dense Subgraphs from Uncertain Graphs" nghiên cứu cách tìm dense subgraph và reliable trong đồ thị không chắc chắn, việc phát hiện các cấu trúc dày đặc trong đồ thị không chỉ đơn giản là tìm các khu vực có mật độ cao, mà còn cần xét đến độ tin cậy của các liên kết (cạnh). Bài báo giới thiệu hai thuật toán chính là GreedyUDS và GreedyOβS. GreedyUDS tối ưu hóa expected density (mật độ dự kiến) của subgraph bằng cách loại bỏ các đỉnh dựa trên expected degree, trong khi GreedyOβS tập trung vào việc optimal β-subgraph thông qua surplus degree (độ dư thừa). Cả hai phương pháp này đều được chứng minh qua thử nghiệm và phân tích để cải thiện đáng kể cả về mật độ và độ tin cậy của subgraph, đặc biệt là trong các đồ thị lớn và phức tạp. Bài báo cung cấp các định nghĩa cơ bản và chỉ số đánh giá, chứng minh rằng các phương pháp đề xuất cải thiện đáng kể cả về density lẫn reliable của subgraph, mở ra hướng tiếp cận mới trong việc khai phá và phân tích dữ liệu đồ thị trong nhiều lĩnh vực ứng dụng thực tế.

2/ Chi tiết thuật toán

2.1/ GreedyUDS

Tìm ra các dense subgraph trong đồ thị không chắc chắn

1. **Input**: Một đồ thị không chắc chắn $G = (V, E, p)$, trong đó V là tập hợp các đỉnh, E là tập hợp các cạnh, và p biểu thị xác suất liên kết với mỗi cạnh.
2. **Quá Trình**:
    - Thuật toán bắt đầu với toàn bộ đồ thị G.

- Nó tính toán expected density của đồ thị hiện tại. Expected density được tính toán dựa vào xác suất của mỗi cạnh.
- Trong mỗi lần lặp, đỉnh có độ mong đợi thấp nhất sẽ bị loại bỏ khỏi đồ thị. Expected degree của một đỉnh thường là tổng xác suất của các cạnh kết nối với nó.
- Quá trình này được lặp lại cho đến khi tất cả các đỉnh đều được loại bỏ. Trong mỗi lần lặp, một subgraph mới được hình thành bằng cách loại bỏ đỉnh.

3. **Output**: Subgraph có expected density cao nhất trong quá trình được chọn là densest subgraph.

2.2/ GreedyOβS

Tìm ra tiểu optimal β-subgraph trong đồ thị không chắc chắn

1. **Input**: Một đồ thị không chắc chắn $G = (V, E, p)$ và một tham số β. Tham số β này xác định ngưỡng xác suất trung bình cho các cạnh trong subgraph.
2. **Quá Trình**:
   - Thuật toán bắt đầu bằng cách khởi tạo một hàng đợi ưu tiên và tính toán surplus degree (độ dư thừa) của mỗi đỉnh. Độ dư thừa là hiệu giữa xác suất của mỗi cạnh và β.
   - Trong mỗi lần lặp, thuật toán loại bỏ đỉnh có độ dư thừa thấp nhất và cập nhật độ dư thừa của các đỉnh liên quan.
   - Quá trình này được tiếp tục cho đến khi tất cả các đỉnh đều được xét qua.
3. **Output**: Subgraph có surplus degree trung bình lớn nhất được chọn là optimal β-subgraph.

3/ Các thông số đánh giá

Expected Edge Density (τ)

Chỉ số đánh giá mật độ cạnh trong một subgraph. Chỉ số này giúp đánh giá mức độ kết nối chặt chẽ của một subgraph. Giá trị τ càng cao cho thấy đồ thị con đó càng dày đặc về mặt liên kết giữa các đỉnh.

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

Probability of edges $p(G_S(e))$

Đo lường xác suất trung bình của các cạnh trong subgraph G_S. Trong môi trường đồ thị không chắc chắn, xác suất này phản ánh độ tin cậy của mỗi cạnh.

$$P(G_S(e)) = \frac{\sum_{e \in G_S} p(e)}{|E(G_S)|}$$

Edge-Probability Standard Deviation (σ)

Độ lệch chuẩn của xác suất cạnh trong subgraph. Đo lường mức độ phân tán của xác suất cạnh so với xác suất trung bình của các cạnh trong subgraph. Giúp phát hiện sự không đồng đều trong xác suất của các cạnh.

Một giá trị σ thấp chỉ ra rằng các cạnh trong tiểu đồ thị có xác suất tương đối ổn định và đồng đều, trong khi một giá trị cao chỉ ra sự không đồng nhất lớn trong xác suất cạnh, điều này có thể là dấu hiệu của sự không chắc chắn cao hoặc nhiễu.

$$\sigma = \sqrt{\frac{1}{N} \sum_{e \in E[G_S]} (p(e) - \bar{p})^2}$$

Adjoint Reliability (R)

Chỉ số đánh giá độ tin cậy tổng thể của subgraph. Trong đồ thị không chắc chắn, độ tin cậy này được tính toán dựa trên xác suất của các cạnh.

Một giá trị R cao cho thấy subgraph có khả năng tồn tại cao trong bối cảnh của mô hình đồ thị không chắc chắn, đồng nghĩa với việc các cạnh trong subgraph có xác suất cao.

$$R(G_S) = \prod_{e \in E[G_S]} p(e)$$