アイテム集合付き部分グラフに対する事後処理法の提案

山田 恵　(指導教員：瀬々 潤)

1　はじめに

SNS，遺伝子ネットワークなど大規模なグラフの構造が増加している．これらのグラフの多くにはSNSであれば商品の購買履歴，遺伝子であれば反応した薬剤など頂点にアイテム集合として表される属性が付与されており，アイテム集合付きグラフ (IA グラフ) と呼ばれている [1]．IA グラフはこのように頻出する構造にもかかわらず研究は限定的である [1, 2, 3]．IA グラフで有用な部分グラフは，共通のアイテム集合を有する部分グラフ (ISS) である．この ISS は，共通の商品を購入するコミュニティや共通の薬剤に反応するパスウェイを表し，利用価値が高い．その一方で高速に ISS を求めるアルゴリズムである COPINE [1] を利用して求めた解は，特に辺やアイテムの多いネットワークを扱った際に，互いに類似した解を多数導出してしまう傾向にあり，これが解の可読性を阻害することがあることが知られていた．本研究では，この可読性の低下を防ぐため，COPINE で導出した解に対する後処理法を提案し，有益な ISS を導出する手法を提案する．

2　準備

グラフ G を非連結でラベルや重みのない無向グラフとする．$V(G)$，$E(G)$，$I(G)$ および $I(v)$ は，それぞれ G の頂点集合，辺集合，G 頂点の持つ全アイテムの集合，頂点 $v \in V(G)$ が持つアイテム集合とする．この時，グラフ G を Itemset-Associated graph (IA グラフ) と呼ぶ．図 1(A) に例を示す．

G' を G の部分グラフとする．このとき，グラフ G' の持つ共通アイテム集合は，$I(G') = \bigcap_{v \in V(G')} I(v)$ と表せ，$I(G') \neq \phi$ かつ，G' に隣接する全てのノード v' について $I(G' \cup \{v'\}) \neq I(G')$ のとき，G' を Itemset-Sharing Subgraph (ISS) と定義する．図 1(A) では，頂点 2,3,6,7 はグラフ上連結であり，かつ，これらの頂点は全て共通のアイテム集合 {A,B,C} を有するが，この部分を ISS という．

定義 1 ISS 列挙問題: θ_I，θ_S をユーザ定義の値とする．*ISS* 列挙問題は，与えられた *IA* グラフ G から，共通アイテム集合の大きさが θ_I 以上，グラフの大きさが θ_S 以上の *ISS* を全て列挙する問題である．

図 1(B) は共通アイテムが二個以上の ISS を全て列挙した例である．COPINE は，100 万を超える辺を持つ大規模な IA グラフも解析対象とし，ユーザが与えた部分グラフの大きさと共通アイテム集合の大きさを満たす全 ISS を高速に列挙するアルゴリズムである．

3　実験と問題提起

3.1　大規模データからの ISS 抽出

関らの論文 [1] では ISS の有益性の実証として，辺数 7,564，6,152 遺伝子の酵母のタンパク質相互作用情報と 173 種類のストレス環境下で計測した遺伝子発現量をアイテム集合データとした IA グラフから，特定の環境下で共通して働くネットワークを抽出し，既知の知見との一致を確認している．

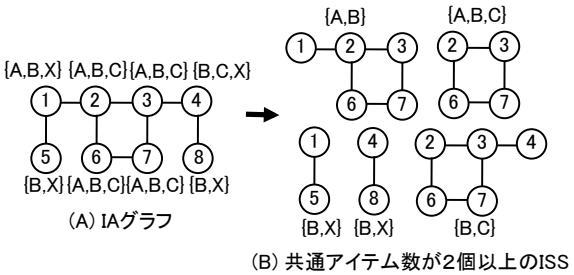

図 1: IA グラフの例，ISS 集合列挙

表 1: ISS が見つかる確率

最小の共通アイテム数 \ 最小のエッジ数	10	15	20
3	1.0000	1.0000	1.0000
4	0.6721	0.3953	0.1734
5	0.0010	<0.0001	<0.0001
6	<0.0001	<0.0001	<0.0001

より大規模なネットワークを有するヒトなどの高等な生物に解析が適用可能か判断するため，IA グラフとして，iRefIndex [1]の遺伝子ネットワーク，アイテム集合として BioGPS [4] のヒトの 79 組織に対し 2 回ずつ実験を行ったものを用いた．遺伝子発現量は，各遺伝子に正規分布を当てはめた時に p 値が 0.05 より低い場合に高発現であるとみなした．本データから生成される IA グラフは，頂点数 15,519 個, アイテム数 $79 \times 2 = 158$ 種, 辺の数 235,407 本であった．ここから COPINE を用い ISS の最小サイズ $\theta_S = 20$, 共通アイテム最小サイズ $\theta_I = 5$ として全 ISS を列挙した結果，7,064 個の ISS が抽出された．

3.2　ランダムデータから ISS が得られる確率

大規模複雑な IA グラフのため，ISS が偶発的に発生している可能性もあり得る．この可能性を確認するため，疑似的に IA グラフを作成し，今回求めた ISS が偶発的に現れたものである確率が低いことを確認した．

疑似データの作成は以下の手順で行った．(1) 元データと頂点数が同一でノード数が近いスケールフリーなグラフを作成．[5] (2) 各アイテムが全頂点に含まれる確率が同じになる様に頂点にアイテムをランダムに配置．(3) 作成したグラフで COPINE を実行し解の有無を確認．(1) でスケールフリーなグラフを作成しているのは，元のグラフにその性質が認められた為である．以上の手順を COPINE のパラメータを動かしながら，10,000 回の試行を行い，解が現れる確率を計算した．結果を表 1 に示す．<0.0001 は 10,000 回の試行では 1 度もそのパラメータを満たす解が見つからなかった事を示している．この結果より，今回の共通アイテム 5 個以上，辺 20 本以上のグラフが現れる確率は非常に低く，得られた解が偶発的とは言えないことが分かる．

[1] http://irefindex.uio.no/wiki/iRefIndex

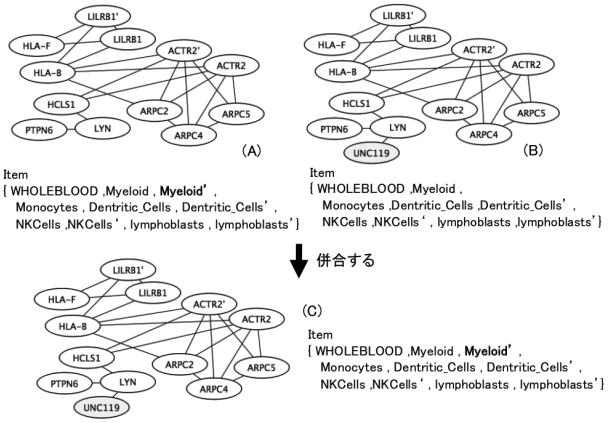

図 2: 併合による遺伝子が追加される例

3.3 ISS の重複

偶発的な解ではないながら，今回得られた解は数が多い上に，結果の一部には共通する頂点かつ共通するアイテム集合を持つ ISS 群が存在したため，可読性が低かった．その例が図 2 である．この図は抽出された解から選んだ 2 個の ISS である．頂点内の文字は遺伝子名を表し，グラフに付与した集合は共通アイテム集合を表している．(A)(B) は UNC119 以外の頂点は共通し，共通アイテム集合も Myeloid' 以外は共通である．この類似した異なる 2 つの解が現れた原因は，UNC119 のアイテム集合に Myeloid' が含まれていなかった事である．この調査から実データでは頻繁に起こる観測ノイズにより，ISS の重複が引き起こされている可能性が高い事が分かった．以後，多くの頂点もアイテムも共通する ISS の事を「重複している」と呼ぶ．

次章ではこの問題点を乗り越え，複雑な IA グラフからでも可読性の高い解を得る手法を提案する．

4 提案手法

本章では重複の多い ISS を併合することで重要な ISS を求める手法を提案する．

IA グラフ間の重複の多さを測る指標を定義し，その定義を利用して併合する IA グラフを再帰的に決め，併合する．

定義 2 *(IA グラフの併合)* 2 つの *IA* グラフを G_i, G_j とする．頂点 $V(G_i) \cup V(G_j)$ 及び辺 $E(G_i) \cup E(G_j)$ を有し，アイテム集合 $I(G_i) \cup I(G_j)$ が付与されている *IA* グラフ G' を考える．この時，G' を G_i と G_j を併合したグラフと呼ぶ．なお，併合により作成した G' は，作成方法により必ず連結グラフである．

遺伝子ネットワークや SNS の解析においては利用者がデータに関して知識を持っていることが多いため，過小評価した結果より，多少のミスを含んだとしても大きな範囲で取られたデータの方が利用しやすい点を考慮し，IA グラフの併合として和集合を取っている．

定義 3 *(重複度)* 現在存在している *IA* グラフの集合を $\mathcal{G} = \{G_1, ..., G_n\}$ とする．$G_i, G_j \in \mathcal{G}(i \neq j)$ に対し重複度 $intersec(G_i, G_j)$ を

$$intersec(G_i, G_j) = \frac{|V(G_i) \cap V(G_j)|}{|V(G_i) \cup V(G_j)|} \times \frac{|I(G_i) \cap I(G_j)|}{|I(G_i) \cup I(G_j)|}$$

と定義する．

この指標は，G_i, G_j の両方に関し頂点及びアイテム集合の両者の重複が大きい場合に値が大きくなる．以上の定義をふまえ，併合手順を Algorithm1 に示す．

Algorithm 1 併合

Require: $\mathcal{G} = \{G_1, ..., G_n\}$，ユーザ指定の閾値 th
  $\mathcal{S} \leftarrow \mathcal{G}$ の重複度が th より大きい ISS ペア
  while $\mathcal{S} \neq \{\}$ **do**
    $(G_a, G_b) \leftarrow \mathcal{S}$ のうち最大の重複度をもつ ISS ペア
    $G' \leftarrow G_a$ と G_b を併合した ISS
    $\mathcal{G}' \leftarrow \{G \in \mathcal{G} | V(G) \subseteq V(G') \& I(G) \subseteq I(G')\}$
    $\mathcal{G}$ から $\mathcal{G}'$を削除し,G' を追加
    $\mathcal{S} \leftarrow \mathcal{G}$ の重複度が th より大きい ISS ペア
  end while
  return $\mathcal{S}$ の中に含まれる G の集合

4.1 併合した結果

特に，閾値 0.2 以上の ISS を併合した結果，164 個の ISS 集合が抽出できた．併合の一例を図 2(C) に示す．これは図 2(A) と (B) が併合されたものである．今回使用したデータでは，同一のサンプルから 2 度同じ実験を行ったものであり，同じ実験については，2 回目の実験には' をつけて 1 回目と区別している．(A),(B) の IA グラフとも，2 回の内 1 回の Myeloid での実験がアイテムとして含まれており，併合後の ISS の共通アイテム集合に Myeloid' を入れることで，実験誤差等による値のぶれで発見できなかったネットワークを補完できた可能性が高いことが分かる．

また，併合が生物学的に妥当であったかを調べる為，併合前後の遺伝子集合に対し対応する機能を遺伝子オントロジーを用いて調べた．ランダムに 10 クラスタを選択し，二項検定による検証の結果，7 クラスタについて併合前より併合後の方がより既知の機能情報に有意な関係が認められた．つまり，より既知の情報に妥当な遺伝子群となっていると考えられる．

5 まとめ

大規模 IA グラフから抽出された重複の多い ISS を併合し簡潔な結果を得るための事後処理法を提案した．冗長な ISS を併合する事で，より可読性の高い ISS を抽出する事が出来た．さらに，生物学的な知見からの検証の結果，併合の妥当性が認められた．本提案手法により，より有用な ISS を抽出することが可能となった．

参考文献

[1] M. Seki and J. Sese, Identification of Active Biological Networks and Common Expression Conditions. *IEEE BIBE 2008*, pp. 1–6, 2008.

[2] T. Itoh, C. Muelder, K. Ma, and J. Sese. A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs. *IEEE Pacific Visualization Symposium 2009*, pp.121-128, 2009.

[3] M. Fukuzaki, M. Seki, H. Kashima and J. Sese, Side Effect Prediction Using Cooperative Pathways. *IEEE BIBM 2009*, pp. 142–147, 2009.

[4] A. Su, T. Wiltshire, *et al.*, A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci*, Vol. 101, No 16, pp 6062-7, 2004.

[5] N. Masuda, H. Miwa, N. Konno, Analysis of scale-free networks based on a threshold graph with intrinsic vertex weights *Phys. Rev. E*, 70(3), Article No. 036124, 2004.