メンバー間の距離が小さいコミュニティの発見

Finding a Community with Small Diameter

朝廣雄一　宮野英次

Abstract

人を頂点，友人関係を辺で表すことにより，人的ネットワークに対応するグラフ構造を考える．このようなグラフ構造に対して，緊密な関係にある人々のグループの中で大きなものを求めたい．緊密さとして様々な条件を考え得るが，グループに属する人々はお互いに直接の友人である，あるいは例えば友人の友人のように友人関係をたどると距離が小さい（d 以下）という二つの条件に対応するクリーク問題と d-クラブ問題が知られている．本稿では近似アルゴリズムの観点から，これらの二つの問題，特に d-クラブ問題に関する研究動向について紹介する．

キーワード：クリーク，クラブ，近似下界，近似アルゴリズム

1．はじめに――クリークとクラブ――

人を頂点，友人関係を辺で表したグラフ構造を考える．図 1 がその例で，例えば，A さんと B さんや A さんと G さんは直接の友人であるが，A さんと F さんは直接の友人ではないことを表している．ソーシャルネットワークサービスの利用者が持つ友人関係から，このようなグラフ構造が定義できる．今，何らかの理由で（例：ターゲット広告），友人関係が緊密なグループ（頂点集合の部分集合，簡単に頂点部分集合と呼ぶ）のうちできるだけ大きなものを発見したい．緊密であることの定義方法はいろいろあると思うが，その頂点部分集合に属している人たちはお互いに直接の友人であると定義してみる．図 1 では，頂点部分集合 $\{B, C, D, H\}$ が，そのような構造を持っているもののうち最大のものである．この構造はクリークと呼ばれ，最大のクリークを求めることを目的とするクリーク問題は理論計算機科学では超有名である．

クリーク問題が超有名な理由の一つは，この問題が 1950 年頃から知られている[1]とともに，1970 年代に NP-困難性が示された最初の 21 個の問題のうちの一つであることである[2]．NP-困難性の定義は割愛するが，その特徴として，頂点数 n の多項式で表される時間（多項式時間）では最適解が求まりそうにない点が本稿では重要である．計算機科学の分野では，多項式時間で完了できる処理が，コンピュータ上で現実的に利用可能であると考えているからである．ここで「〜そうにない」と歯切れが悪い感じで書いているが，NP-困難な問題に対して多項式時間で最適解が求まるかどうかは，これまた超有名な未解決問題であり，100 万ドルの懸賞金

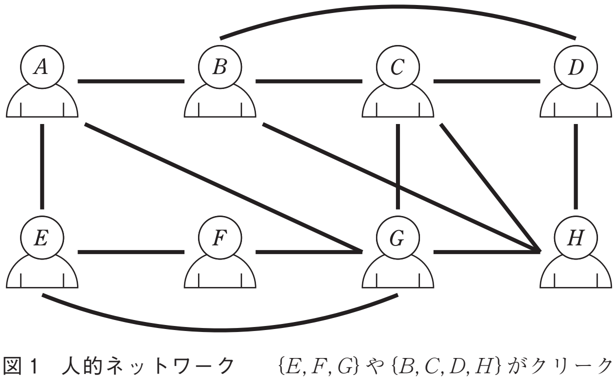

図 1　人的ネットワーク　$\{E, F, G\}$ や $\{B, C, D, H\}$ がクリークで，後者が最大クリークである．また，$\{A, B, C, G, H\}$ は 2-クラブだが $\{A, B, C, E, H\}$ は 2-クラブではない．アルゴリズム 1 は $N[G]=\{A, C, E, F, G, H\}$ を出力する．

朝廣雄一　九州産業大学理工学部情報科学科
E-mail　asahiro@is.kyusan-u.ac.jp
宮野英次　正員　九州工業大学大学院情報工学研究院システム創成情報工学研究系
E-mail　miyano@ces.kyutech.ac.jp
Yuichi ASAHIRO, Nonmember (Faculty of Science and Engineering, Kyushu Sangyo University, Fukuoka-shi, 813-8503 Japan) and Eiji MIYANO, Member (Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Fukuoka-shi, 820-8502 Japan).
電子情報通信学会誌　Vol.101 No.3 pp.262-266 2018 年 3 月

が用意されている[3]．このため，この「〜そうにない」という表現は，以下でも何度か使うことになるが，御了承頂きたい[注1]．最大クリークを求めるには，頂点部分集合を全部列挙しそれぞれがクリークになっているかどうか確認し，見つけた最大クリークを出力するという簡単な方法がある．しかし，頂点部分集合の候補は 2^n 個近くあるので，この方法は多項式時間では完了できない．

最大クリークを求めるのが難しいとなると，ではどうしようか．「メンバーが直接の友人である」という条件が厳しそうなので，例えば友人の友人でよいという条件に緩めてやれば，簡単に求まるのではないか．友人の友人とは共通の友人を介して近い将来は直接の友人になることが見込まれる，という意味で将来的にはクリークになりそうな頂点部分集合を探そうとしているとも言える．そこで，直径，すなわち二者間の距離の最大値を考え，直径が $d(\geq 2)$ 以下である頂点部分集合を d-クラブと呼ぶ．ここで注意が必要なのは，その頂点部分集合に属している頂点だけを使って距離が d 以下にならねばならない点である[注2]．例えば友人の友人は距離 2 と考え，図 1 では，例えば $\{A,B,C,G,H\}$ が 2-クラブとなっている．また $\{A,B,C,E,H\}$ については，E さんと H さんは元のグラフ中には共通の友人として G さんがいるが，$\{A,B,C,E,H\}$ の中には共通の友人がいないため，2-クラブではないと考える．最大の d-クラブを求めることを目的とする d-クラブ問題はクリーク問題ほど有名ではないものの 1970 年代から知られている[4],[5]．さて，最大の 2-クラブなどなら多項式時間で求まるのであればよかったのだが，残念ながらそんなことにならず相変わらず NP-困難であることが 2002 年に示されており[6]，ちょっと困ってしまう．ちなみにグラフ全体の直径は明らかに $n-1$ 以下であり，こういった自明な値以外の d について d-クラブ問題を考える．

2. 近似アルゴリズムと近似下界

2.1 近似アルゴリズムの概念

NP-困難なことが分かったからといって諦めてしまうわけではなく，様々な方向性で研究が行われるのが常である．本稿では，近似アルゴリズムの分野について主に紹介する．（近似アルゴリズムの参考書籍としては文献(7)が挙げられる．）他の手法による結果なども 4. で簡単にまとめる．

アルゴリズム ALG が σ 近似アルゴリズムである，または ALG の近似度が σ であるとは，全てのグラフ G に対して，$OPT(G)/ALG(G)\leq\sigma$ が成立することを意味する．ここで $ALG(G)$ と $OPT(G)$ はそれぞれ，G に対して ALG と（時間を無限に使ってよい）最適アルゴリズムを適用したときに，それぞれが出力する頂点部分集合の大きさを表す．最大のものは見つけることができないが，$1/\sigma$ の大きさのものならば見つけられるという保証の付いたアルゴリズムと理解してもらえばよい．上記の定義から近似度 σ は 1 よりも大きいわけで，できるだけ小さい近似度を持ち，多項式時間で動作するアルゴリズム（多項式時間近似アルゴリズム）を設計したい．すなわち，処理に必要な時間を現実的なものに短縮する代わりに，解の良さの方を少し諦めようという考え方である．例えば，図 1 に対して $\{E,F,G\}$ を出力すると，最大クリークは $\{B,C,D,H\}$ なので，このグラフに対しては最適解の 3/4 の大きさの解を出力していると考えるわけである．極端な話をすると，辺で接続されている 2 頂点をクリークとして出力する単純なアルゴリズムを考えると，最大クリークはどんなに大きくてもグラフ自体の頂点数 n より大きくはならないので，近似度 $n/2$ のアルゴリズムとなる．以下で近似度を扱う際に，O 記法を用いて，例えば $O(n)$ などと書くのが正確であるが，煩雑になるので O は省略する．

2.2 クリーク問題の近似

クリーク問題に対する多項式時間近似アルゴリズムについて，以下が知られている．

［定理 1］[8]　クリーク問題に対して，多項式時間 $n(\log\log n)^2/(\log n)^3$-近似アルゴリズムが存在する．

ここで log の底は 2 である．例えば $n=2^{16}$ だとすると，$(\log\log n)^2/(\log n)^3=1/256$ なので，上記で考えた単純な $n/2$-近似アルゴリズムより良さそうである．しかし，n は幾らでも大きくなり得るので，できたら n の一次関数ではなく，$\sqrt{n}$ や $\log n$ の関数，もっと言うと定数（例えば 2）の近似度を持つアルゴリズムが設計できるとうれしい．

しかし一方で，以下のような近似度の下界（近似下界）が知られている．

［定理 2］[9]　任意の正定数 ϵ に対して多項式時間 $n^{1-\epsilon}$-近似アルゴリズムはクリーク問題には存在しそうにない．

これによると，例えば $n^{0.999}$ 以下の近似度を持つ多項式時間アルゴリズムは設計できそうにない．つまり，ほとんど何も努力していない $n/2$-近似アルゴリズムでも限界にかなり近い近似度を持っていることになり，近似アルゴリズムの概念を持ち込んでも，クリーク問題はほとんど手も足も出ないような困難な問題であることが分かる．

（注 1）「$P\neq NP$ を仮定すると，多項式時間で最適解は求まらない」がより正確であるが，P や NP の定義を省略するため，このような表現とした．

（注 2）メンバー以外の人を介してでも距離 d 以下であればよいと定義すると，d-クリークと呼ばれるよく似た別の問題になる．

2.3　d–クラブ問題の近似下界

さて，d–クラブ問題については，クリーク問題のように手も足も出ないのであろうか．2002 年には近似下界として以下も知られることとなった．

［定理 3］[(10)]　任意の正定数 ϵ に対して多項式時間 $n^{1/3-\epsilon}$–近似アルゴリズムは d–クラブ問題には存在しそうにない．

クリーク問題の近似下界が $n^{1-\epsilon}$ であったことを考えると，依然 n の関数ではあるものの，アルゴリズムに対する工夫の余地がまだまだありそうである．であるのだが，筆者らが近似下界を少し押し上げてしまった：

［定理 4］[(11)]　任意の正定数 ϵ に対して多項式時間 $n^{1/2-\epsilon}$–近似アルゴリズムは d–クラブ問題には存在しそうにない．

少し余談になる．近似下界を上げることは，大きな d–クラブを求めるのには特に役に立たないように思われるかもしれない．しかし，この近似下界が示せたことにより，後述の近似アルゴリズムを設計できたと言っても過言ではない．この辺りは多分に感覚的な話になるが，近似度 $n^{1/3}$ を目指すとなると妙案はない一方で，目指すところが $n^{1/2}$ なら，ある種のアイデアを利用できる．また，近似下界（や NP–困難性）を示す過程で問題の困難さをつかさどっている部分構造が明らかになるので，それへの対処方法を考えることが，良いアルゴリズムを設計する方向性を与えてくれる．

3．d–クラブ問題の近似アルゴリズム

本章では d–クラブ問題に対する多項式時間近似アルゴリズムとして，3.1 と 3.2 でそれぞれ近似度 $n^{(d-1)/d}$ と近似度 $n^{1/2}$ のものを紹介する．単純なので，アルゴリズムの内容についても記述する．

3.1　近似度 $n^{(d-1)/d}$

幾つか用語を定義する．頂点の次数とは，その頂点と辺でつながっている（隣接する）頂点の個数である．頂点 v 自身と v に隣接する頂点を含む頂点部分集合を $N[v]$ で表す．以下のアルゴリズム 1 が知られている．

［アルゴリズム 1］　次数が最も大きい頂点 u を探し，$N[u]$ を出力する．

アルゴリズム 1 は，人的ネットワークで言うと，友人のとても多い一人とその友人を解として出力するものである．解として出力する $N[u]$ は，u 以外の人同士の間の辺を考慮しなければ，（u を中心とする）星グラフと呼ばれる構造を持つ．例えば図 1 のグラフに対して，アルゴリズム 1 は，最大次数を持つ頂点 G を見つけ，$N[G]=\{A, C, E, F, G, H\}$ を出力する．

アルゴリズム 1 が出力する頂点部分集合においては，u を仲介することで u 以外のどの 2 頂点間の距離も 2 以下であり，また，u 以外のどの頂点も u との距離は 1 であるので，2–クラブとなっている．よって，$d \geq 3$ の場合の問題に対しても，これをそのまま d–クラブとして出力してしまってもよい．アルゴリズム 1 は，現実的なグラフ構造に対する実験的な性能が優れており（例えば文献(12)），理論的には最良な発見的手法（ヒューリスティックス）である[(注3)]，すなわち，アルゴリズム 1 の出力する解を常に上回れるようなヒューリスティックスは存在しそうにないことが知られている[(13)]．

アルゴリズム 1 に対して，筆者らは近似度の保証を与えた：

［定理 5］[(11)]　d–クラブ問題に対して，アルゴリズム 1 の近似度は $n^{(d-1)/d}$ である．

$d=2$ の場合には，この単純なアルゴリズムの近似度が $n^{1/2}$ となり，定理 4 で示される近似下界から，理論的にほぼ最良の性能を持つことが分かる[(注4)]．一方で $d \geq 3$ の場合には，近似下界との差が存在するので，更に良いアルゴリズムを設計したい．

3.2　近似度 $n^{1/2}$

本節で紹介するアルゴリズム 2 は筆者らのグループによるものであり[(11)]，べきグラフの概念を用いる．グラフ G の d 次べきグラフ G^d とは，頂点集合としては G と同じ頂点集合を持ち，辺集合としては G 中で距離が 2 以上 d 以下である頂点間に辺を追加したグラフ（元々あった辺はそのまま持つ）である．

［アルゴリズム 2］

ステップ 1．入力グラフ G の各辺に頂点を 1 個挿入し，それぞれ 2 本の辺に置き換える．G に元々あった頂点を黒頂点，辺に挿入した頂点を白頂点と呼び，出来上がったグラフを H とする．

ステップ 2．H から d 次べきグラフ H^d を作る．

ステップ 3．最多の黒頂点を含む星グラフを H^d から探し，それを構成する黒頂点の集合を出力する．

例えば，$d=3$ の場合を考えてみよう．図 2 に示すように，黒頂点から成るグラフ G の全ての辺に白頂点を追加することで得たグラフ H の三次べきグラフ H^3 を求める．その後，黒頂点の数が最大（この場合は 6 頂点）となるような頂点 u を中心とする星グラフを探し，それに含まれる黒頂点の集合（図中では中黒の二重丸）を出力することになる．アルゴリズム 2 の出力が 3–クラブであることは，アルゴリズム 1 の議論と同様の（少しだけ複雑な）議論で証明できる．ちなみに同じグラフにアルゴリズム 1 を適用すると，G 中の次数 3 の頂点とそれに隣接する頂点の合計 4 頂点を解として出力するの

（注 3）　提唱者らはこの概念に provably best heuristics と名付けている．適切な日本語訳を筆者らは寡聞にして知らないのでこのように訳した．

（注 4）　クリーク問題のように，例えば log などのレベルで近似度を更に小さくできるのかもしれない．

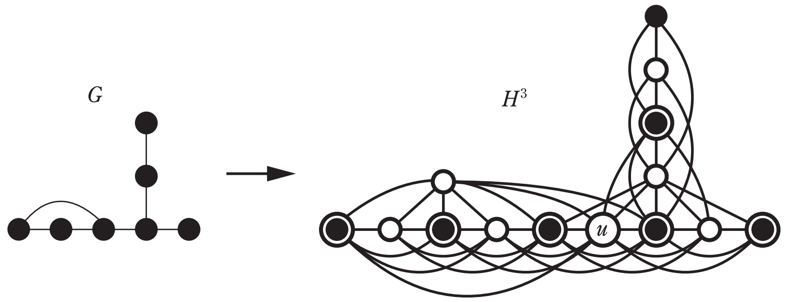

図2　グラフ G とグラフ H^3　　アルゴリズム2は G から H^3 を作り，その中から最多の黒頂点を含む星グラフ $N[u]$ に含まれる黒頂点の集合（中黒の二重丸）を出力する．

で，このグラフに対してはアルゴリズム2の方が良い解を出力することが分かる．アルゴリズム2は多項式時間で動作し，$d \geq 3$ の場合でも近似度 $n^{1/2}$ を達成する：

［定理6］[11]　d-クラブ問題に対して，アルゴリズム2の近似度は $n^{1/2}$ である．

4. クラブを求める他のアルゴリズム

近似解ではなく最適解を求めるアルゴリズムとして，単純なしらみ潰し法によれば 2^n 程度の時間が掛かってしまうところ，1.62^n 程度の時間で実行できるアルゴリズムが知られている[14]．また，入力グラフ中に大きさ k 以上の d-クラブが存在するか否かを問う問題に対しては，$(k-2)^k \cdot k! \cdot kn + nm$ 程度の時間で実行できるアルゴリズムも知られており[15]，k が定数などの n に比べて小さい値ならば現実的な時間で答えを得ることができる．

一般のグラフを対象とするならば，上で述べてきたように d-クラブ問題は困難である．しかしながら特殊な構造を持つグラフ，定義は省略するが，例えば，区間交差グラフ，台形交差グラフ，強弦グラフなどに対しては，多項式時間で最適解を求められるアルゴリズムも知られている[16]～[18]．なお，特殊なグラフ構造に対して多項式時間で最適解を求められるという性質はクリーク問題にも見られ，例えば平面グラフや弦グラフに対するアルゴリズムが知られている[19]．

5. おわりに――関連の話題――

誌面が尽きたので関連の研究テーマについて，ごく簡単に紹介する．d-クラブ問題としては，生化学ネットワーク内で似た構造を持つ一群を見つけるという別の応用例もある（例えば文献(20)）．また，入力グラフ中に存在するクリークを列挙する手法も開発されており（例えば文献(21)），d-クラブ問題に対しても，同様の方向性の研究が考えられるだろう．更に，本稿で紹介した d-クラブ問題は，距離に着目してクリーク問題を拡張したものと捉えられるが，クリーク問題の拡張方法としては様々なもの（例えば次数に着目した文献(22)など）もある．

謝辞　本稿は JSPS 科研費 JP17K00016 と JP17K00024 の助成を受けたものである．

文　　献

（1） R.D. Luce and A.D. Perry, “A method of matrix analysis of group structure,” Psychometrika, vol. 14, no. 2, pp. 95-116, 1949.

（2） R.M. Karp, “Reducibility among combinatorial problems,” Complexity of Computer Computations, R.E. Miller and J.W. Thatcher, eds., pp. 85-103, Plenum Press, New York, 1972.

（3） Clay Mathematics Institute, P vs NP Problem. http://www.claymath.org/millennium-problems/p-vs-np-problem （2017年9月1日閲覧）

（4） R. Alba, “A graph-theoretic definition of a sociometric clique,” J. Math. Sociol., vol. 3, no. 1, pp. 113-126, 1973.

（5） R.J. Mokken, “Cliques, clubs and clans,” Qual. Quant., vol. 13, no. 2, pp. 161-173, 1979.

（6） J.-M. Bourjolly, G. Laporte, and G. Pesant, “An exact algorithm for the maximum k-club problem in an undirected graph,” Eur. J. Oper. Res., vol. 138, no. 1, pp. 21-28, 2002.

（7） V.V. Vazirani, Approximation Algorithms. Springer, 2003.

（8） U. Feige, “Approximating maximum clique by removing subgraphs,” SIAM J. Discrete Math., vol. 18, no. 2, pp. 219-225, 2004.

（9） D. Zuckerman, “Linear degree extractors and the inapproximability of max clique and chromatic number,” Theory Comput., vol. 3, no. 1, pp. 103-128, 2007.

（10） J. Marinček and B. Mohar, “On approximating the maximum diameter ratio of graphs,” Discrete Math., vol. 244, no. 1-3, pp. 323-330, 2002.

（11） Y. Asahiro, Y. Doi, E. Miyano, K. Samizo, and H. Shimizu, “Optimal approximation algorithms for maximum distance-bounded sugraph problems,” Algorithmica, vol. 79, no. 289, pp. 1-23, 2017.

（12） S. Hartung, C. Komusiewicz, A Nichterlein, and O. Suchý, “On structural parameterizations for the 2-club problem,” Discrete Appl. Math., vol. 185, pp. 79-92, 2015.

（13） S. Kahruman-Anderoglu, A. Buchanan, and S. Butenko, “On provably best construction heuristics for hard combinatorial optimization problems,” Networks, vol. 67, no. 3, pp. 238-245, 2016.

（14） M.-S. Chang, L.-J. Hung, C.-R. Lin, and P.-C. Su, “Finding large k-clubs in undirected graphs,” Computing, vol. 95, no. 9, pp. 739-758, 2013.

（15） A. Schäfer, C. Komusiewicz, H. Moser, and R. Niedermeier, “Parameterized computational complexity of finding small-diameter subgraphs,” Optim. Lett., vol. 6, no. 5, pp. 883-891, 2012.

（16） Y. Asahiro, E. Miyano, and K. Samizo, “Approximating maximum diameter-bounded subgraphs," Proc. LATIN 2010, Lect. Notes Comput. Sci., vol. 6034, pp. 615-626, 2010.

(17) P.A. Golovach, P. Heggernes, D. Kratsch, and A. Rafiey, "Finding clubs in graph classes," Discrete Appl. Math., vol. 174, pp. 57-65, 2014.

(18) Y. Asahiro, Y. Doi, E. Miyano, and H. Shimizu, "Optimal approximation algorithms for maximum distance-bounded subgraph problems," Proc. COCOA 2015, Lect. Notes Comput. Sci., vol. 9486, pp. 586-600, 2015.

(19) F. Gavril, "Algorithms for minimum coloring, maximum clique, minimum covering by cliques, and maximum independent set of a chordal graph," SIAM J. Comput., vol. 1, no. 2, pp. 180-187, 1972.

(20) B. Balasundaram, S. Butenko, and S. Trukhanov, "Novel approaches for analyzing biological networks," J. Comb. Optim., vol. 10, no. 1, pp. 23-39, 2005.

(21) K. Makino and T. Uno, "New algorithms for enumerating all maximal cliques," Proc. SWAT2004, Lect. Notes Comput. Sci., vol. 3111, pp. 250-272, 2004.

(22) S.B. Seidman and B.L. Foster, "A graph-theoretic generalization of the clique concept," J, Math. Sociol., vol. 6, no. 1, pp. 139-154, 1978.

（平成 29 年 9 月 23 日受付）

あさひろ　ゆういち
朝廣　雄一

平 6 九大・工・情報卒，平 10 同大学院博士課程了．工博．九大を経て現在は九州産大・理工・教授．組合せ最適化アルゴリズムと計算複雑さの研究に従事．IPSJ，ACM 各会員．

みやの　えいじ
宮野　英次（正員）

平 3 九大・工・情報卒，平 7 同大学院博士課程了．工博．九大，九州芸工大を経て現在は九工大・情報工・教授．アルゴリズムと計算複雑さの研究に従事．IPSJ，ORSJ，ACM の各会員．