修士論文 要旨

円上の交差グラフ族に対するグラフの個数の評価と簡潔データ構造

数理情報学専攻 48196222 **中島 渓**
指導教員 **定兼 邦彦 教授**

1 はじめに

簡潔データ構造とは，データを情報理論的下限に近いサイズに圧縮しつつ効率的に様々なクエリに回答できるようなデータ構造のことである．近年，WEB ページやゲノム配列など巨大なデータを扱う機会が増えており，簡潔データ構造の重要性が高まっている．

本論文では，グラフクラスを限定して簡潔データ構造を考える．グラフクラスを限定すると，一般のグラフにおいて解くのが困難とされる問題が多項式時間で解けたり近似したりすることができる．

また，グラフクラスに着目することで，そのグラフの性質を利用した簡潔データ構造を考えることができる．このような研究は，木 [5]，コーダルグラフ [6]，区間グラフ [1] など広く行われてきている．

今回簡潔データ構造を考える対象のグラフクラスは，円上の交差グラフ族である．弦や円上の凸多角形など円上の図形による交差グラフであり，circle グラフや circle-trapezoid グラフ，circle-polygon グラフなどが含まれる．図 1 は円上の交差グラフ族のグラフの包含関係を表している．一般化 circle-polygon グラフは本論文で導入するクラスであり，circle-polygon グラフと同値なグラフクラスである．

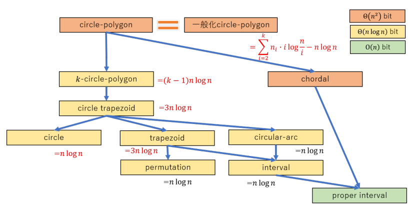

図 1. 矢印は矢の先のグラフクラスを部分クラスとして含むことを表している．

グラフクラスの色は各グラフクラスを表現するのに必要なビット数を表している．グラフクラスの近くの数字は上下界の主要項を記したもので，その中でも赤文字で書いてある数字は本論文で与える上下界である．

2 基本事項

計算モデルは n を入力サイズとしたとき， $\Theta(\log n)$ ビットの word RAM モデルを仮定する．また，扱うグラフは単純無向，重みなし，ラベルなしとする．初めに，簡潔データ構造，0 次経験エントロピーを以下のように定義する．

定義 2.1. 簡潔データ構造
基数 Z の集合の任意の要素について，空間計算量 $\log Z + o(\log Z)$ ビットで効率的にクエリに答えるデータ構造を簡潔データ構造という．

定義 2.2. 0 次経験エントロピー
S をアルファベット $\mathcal{A}$ 上の長さ n の文字列とする．文字 $c \in \mathcal{A}$ の S 中の出現回数を n_c とすると，S の 0 次経験エントロピー $H_0(S)$ は以下の式のように定義される．

$$H_0(S) = \Sigma_{c \in \mathcal{A}} \frac{n_c}{n} \log \frac{n}{n_c}.$$

次に，簡潔データ構造において最も基本的なクエリである rank/select を紹介する．

$S = s_1, s_2, \ldots, s_n$ をアルファベット $\mathcal{A} = \{0, 1, \ldots, \sigma - 1\}$ 上の文字列とする．ここで，$\alpha \in \Sigma$ に対し，$\mathsf{rank}_\alpha(S, i)$ を $S[1..i]$ における α の出現回数，$\mathsf{select}_\alpha(S, i)$ を S 中で i 番目の α が出現する位置と定義する．ビットベクトル上及び文字列上での rank/select について，以下の補題が知られている．

補題 2.3 ([3])**.** 長さ n のビット列 $B[1 \ldots n]$ に対し，B 上での $\mathsf{rank}_\alpha, \mathsf{select}_\alpha$ $(\alpha = \{0, 1\})$ と access が定数時間で処理できる $n + o(n)$ ビットのデータ構造が存在する．

補題 2.4 ([2])**.** 任意の $\alpha > 0, \sigma > 1$ について，$\mathcal{A} = \{0, 1, \ldots, \sigma - 1\}$ 上の文字列上での rank_α, access クエリを $O(1 + \log \log \sigma)$ 時間，select_α クエリを定数時間で回答する，$nH_0 + o(n)(H_0 + 1)$ ビットのデータ構造が存在する．

次に，Range Maximum Query を導入する．全順序集合上の配列 $S = s_1, s_2, \ldots, s_n$ に対し，区間 $[i, j]\,(1 \leq i, j \leq n)$ での Range Maximum Query は $\mathsf{RMax}_S(i, j) := \arg\max_{k \in [i,j]} s_k$ と表される (最大値が複数個出現するときは最も左の位置を返す)．同様にして

Range Minimum Query も定義される．

補題 2.5 ([4])**.** S を長さ n の全順序集合上の配列とする．$2n + o(n)$ ビットで，S は保持せずに，定数時間で $\mathsf{RMax}_S, \mathsf{RMin}_S$ クエリに回答するデータ構造が存在する．

3　円上の交差グラフ族の個数の上下界

円上の交差グラフとは，弦や円上の凸多角形などの円上の図形による交差グラフのことをいう．

一般化 circle-polygon を，円上に頂点をもち，任意の頂点数で，辺は弦または弧である図形と定義する．一般化 circle-polygon による交差グラフを一般化 circle-polygon グラフと呼ぶこととする．一般化 circle-polygon グラフは circle-polygon グラフと同値なグラフクラスである．

このグラフクラスに対し，個数の上下界を与える．

定理 3.1. k ($k = \mathrm{polylog}(n)$) 個以下の頂点をもつ n 個の一般化 circle-polygon により構成されるグラフクラスを考える．このとき，頂点数 i ($2 \le i \le k$) の一般化 circle-polygon の個数を n_i 個とする．$N = \sum_{i=2}^{k} i \cdot n_i, \boldsymbol{n} = (n_1, n_2, \cdots, n_k)$ とおく．$P_{n,k,\boldsymbol{n}}$ をこのようなグラフの総数とすると，以下の式が成り立つ．

$$\log P_{n,k,\boldsymbol{n}} \ge \sum_{i=2}^{k} n_i \cdot i \log \frac{n}{i} - n \log n - O(N \log\log n)$$

次に，上界を与える．コーナー文字列 S を定義する．空文字列の状態から，円上の任意の点から始まり時計回りに円を一周する．このとき，一般化 circle-polygon の頂点が現れるたびに，S にその頂点が属する一般化 circle-polygon の番号を S の末尾に追加した文字列がコーナー文字列である．また，一般化 circle-polygon の番号はこの一周の間に一般化 circle-polygon が現れた順番とする．

コーナー文字列 S を圧縮する．S の長さは $N = \sum_{i=2}^{k} n_i \cdot i$ で文字 i ($2 \le i \le k$) は n_i 回出現することに注意する．各文字 i について S 中での最初の出現は長さ N のビットベクトル F で表現できる．他の文字については，長さ $N - n$ の文字列 S' に格納する．S' 中には，文字 i は $d_i - 1$ 回出現する．ここで，d_i は一般化 circle-polygon i の頂点数とする．S' を 0 次経験エントロピーに圧縮し，各辺が弧か弦かを表す長さ N のビット列 A も追加すると，空間計算量は，

$$\begin{aligned}&\sum_{j=1}^{n} (d_j - 1) \log \frac{N-n}{d_j - 1} + O(N) \\ &\le \sum_{i=2}^{k} n_i \cdot i \log \frac{n}{i} - n \log n + O(N \log k)\end{aligned}$$

補題 3.2. 以上の表現は $k = o(\log n / \log\log\log n)$ のとき，定理 3.1 の下界と漸近的に一致する．

4　簡潔データ構造

2 頂点が隣接するか判定する adjacent，隣接する頂点集合を求める neighborhood，隣接する頂点数を求める degree クエリを考える．

上記の S', F, A に加えて，補題 2.5 を用いて，注目する頂点に関して出ていく辺の終点/入ってくる辺の始点を管理するデータ構造を追加で持つ．F, S' はそれぞれ補題 2.3, 2.4 のデータ構造で持つとすると，空間計算量 $\mathrm{UB} + o(\mathrm{UB})$ ビット (UB は上述の上界) で，adjacent, neighborhood, degree にそれぞれ $O(k)$, $O(|\mathsf{degree}(u)| \cdot k \log\log n)$, $O(|\mathsf{degree}(u)| \cdot k \log\log n)$ 時間で回答することができる．

5　おわりに

circle-polygon グラフと同値なグラフクラス，一般化 circle-polygon グラフを導入し，個数の上下界と簡潔データ構造を与えた．このグラフクラスは，circle グラフや circle-trapezoid グラフなどを含むグラフクラスであるから，円上の交差グラフ族に対して，統一的なアプローチを提示したといえる．

参考文献

[1] H. Acan, S. Chakraborty, S. Jo and S. R. Satti: Succinct data structures for families of interval graphs. In *WADS*, vol. 11646 of LNCS, Springer, 2019, pp. 1–13.

[2] J. Barbay, F. Claude, T. Gagie, G. Navarro and Y. Nekrich: Efficient Fully-Compressed Sequence Representations. *Algorithmica*, 69(1), 2014, pp. 232–268.

[3] D. R. Clark and J. I Munro: Efficient suffix trees on secondary storage. SODA ’96, 1996, pp. 383–391.

[4] J. Fischer and V. Heun: Space-efficient preprocessing schemes for range minimum queries on static arrays. *SIAM Journal on Computing* 40(2), 2011, pp. 465–492.

[5] J. I. Munro and V. Raman: Succinct representation of balanced parentheses and static trees. *SIAM J. Comput.*, 31(3), 2001, pp. 762–776.

[6] J. I. Munro and K. Wu: Succinct data structures for chordal graphs. In *ISAAC*, 67, 2018, pp. 1–12.