Сложность вычислений
«Дерево Штейнера»

Иванов Вячеслав, группа 699

2 января 2019 г.

Оглавление

1 Аннотация

Рассмотрена задача поиска дерева Штейнера — дерева минимального веса, соединяющего в том числе заданное фиксированное подмножество вершин, называемых терминальными — во взвешенном неориентированном графе. Доказана NP-полнота этой задачи и приведен 2-оптимальный алгоритм её решения. Произведен обзор приложений в таких задачах вычислительной биологии, как построение филогенетических деревьев, а также поиск метаболических путей и предсказание биомаркеров рака на основе сетей биологических взаимодействий.

2 Постановка задачи

$G = (V, E)$ — неориентированный граф, $V_0 \subset V$ — непустое множество *терминальных* вершин $\omega : E \to \mathbb{R}^+$ — весовая функция. Требуется решить задачу оптимизации:

$$\min_{T \subset G} \sum_{e \in E(T)} \omega(e)$$
$$\text{s.t.} \begin{array}{l} T \text{ — дерево} \\ V_0 \subset V(T) \end{array}$$

Т.е. найти дерево минимального веса, покрывающее все терминальные вершины.
В нетривиальных частных случаях задача имеет полиномиальный алгоритм решения:

1. $|V_0| = 2$: задача о кратчайшем пути между выделенными вершинами
2. $V_0 = V$: задача о минимальном остовном дереве

Алгоритмы поиска минимального остовного дерева, как будет показано далее, составляют основу 2-оптимального алгоритма поиска дерева Штейнера в метрическом случае.

3 Практическая значимость

Деревья Штейнера возникают в таком количестве практических задач, что на эту тему пишут целые книги. Автору особенно приглянулись примеры из области его научных интересов:

3.1 Построение филогенетических деревьев

Определение 3.1. *Филогенетическое дерево* — корневое дерево, отражающее эволюционные связи и степень сходства между организмами. Построение таких деревьев — частая задача эволюционной биоинформатики.

Определение 3.2. *Расстояние Левенштейна* или *редакторское расстояние* $L(s_1, s_2)$ — минимальная стоимость получения строки s_2 из строки s_1 применением операций вставки, удаления и замены символов друг на друга, каждая из которых имеет некоторую стоимость.

Определение 3.3. *Расстояние Хэмминга* $H(s_1, s_2)$ определено для пар строк одинаковой длины $n := |s_1| = |s_2|$ и выражается через число позиций, в которых s_1 и s_2 различаются:

$$H(s_1, s_2) = \sum_{i=1}^{n} \mathrm{I}\left(s_1^{(i)} \neq s_2^{(i)}\right)$$

В простейшей постановке из генома нескольких организмов взяты участки одинаковой длины. Каждая строка является терминальной вершиной, метрика — расстояние Хэмминга. Требуется построить дерево Штейнера. Утверждается, что оно является хорошим приближением реального

филогенетического дерева, объединяющего эти организмы. В более общем случае последовательности могут иметь разную длины или состоять из нескольких несвязных частей. Тогда в качестве метрики используют расстояние Левенштейна, а на этапе предобработки данных выполняют множественное выравнивание, с помощью которого добиваются максимального подобия сравниваемых участков, взятых из разных организмов (что необходимо для получения биологически осмысленного результата).

Больше про алгоритмы построения филогенетических деревьев можно прочитать в [UEL].

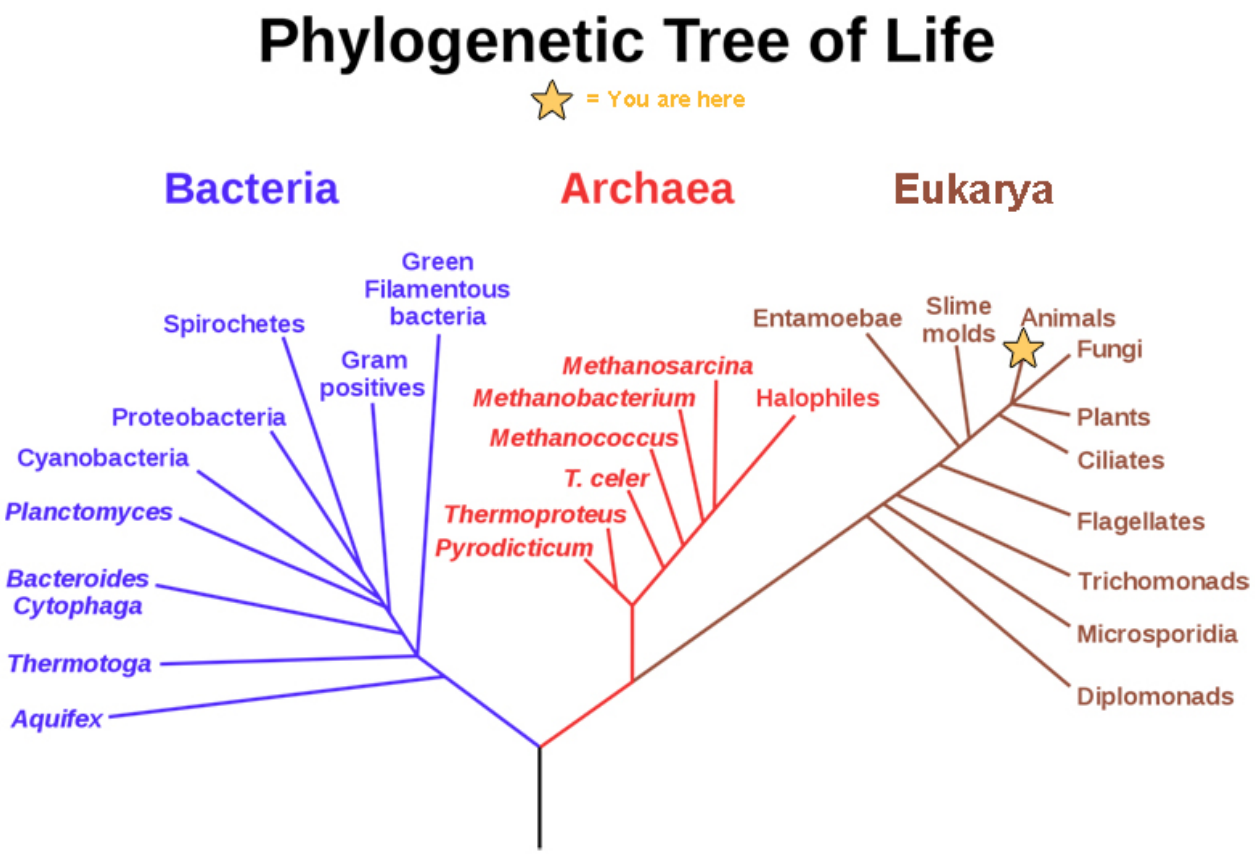

Рис. 1: *Филогенетическое дерево, показывающее общее происхождение организмов из всех трёх доменов: Бактерии, Археи, Эукариоты.*

3.2 Анализ сетей биологических взаимодействий

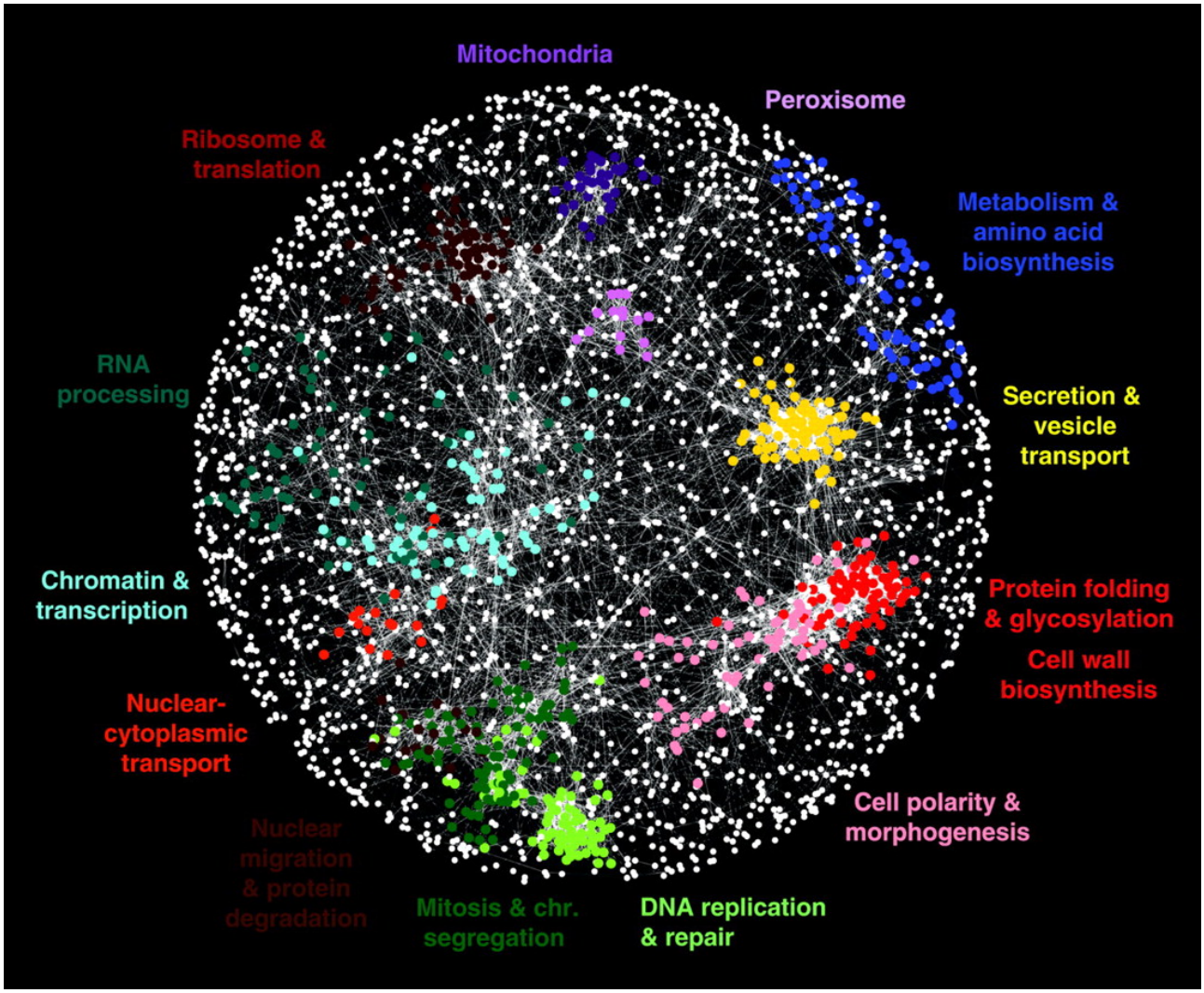

Рис. 2: *Функциональные модули в геноме пекарских дрожжей Saccharomyces cerevisiae. Источник — http://thecellmap.org/*

Определение 3.4. *Экспрессия генов* — процесс преобразования информации, закодированной в ДНК, в клеточные процессы и структуры. В контексте сетей чаще всего говорят про экспрессию матричной РНК (мРНК) и белков, синтезирующихся на её основе. Эти процессы называют *транскрипцией* и *трансляцией* соответственно.

Определение 3.5. *Микрочип (microarray)* — устройство для измерения уровней экспрессии мРНК, представляющее собой кремниевую подложку с множеством ($\approx 10^3 \dots 10^4$) выемок, каждая из которых содержит нить РНК, комплементарную мРНК, транскрибирующейся из исследуемого гена. Микрочипы заложили основу т.н. *high-throughput NGS-методов*[1] — технологий, позволяющих одновременно измерять уровни экспрессии всех (или почти всех) генов в клетке.

Определение 3.6. Если уровни экспрессии какого-то гена существенно различны в двух проведенных измерениях, то говорят, что имеет место *дифференциальная экспрессия генов (differential gene expression)*. Анализ дифференциальной экспрессии проводят для того, чтобы лучше понять течение заболевания или обнаружить, какие именно гены влияют на наблюдаемые признаки.

[1] NGS = Next Generation Sequencing, секвенирование нового поколения

Определение 3.7. *Интерактом* — ориентированный граф с петлями и кратными рёбрами, кодирующий все взаимодействия между молекулами в клетке. В нём выделяют такие важные подграфы, как:

- Граф белок-белковых взаимодействий, он же *протеом* (чаще всего имеют в виду его).
- Метаболическая сеть (преобразования молекул друг в друга под воздействием энзимов)

Расцвет high-throughput методов привёл к тому, что современная геномика генерирует больше данных, чем астрономия и интернет. Человечество получило возможность построить более-менее полные интерактомы основных модельных организмов и интегрировать это знание в анализ дифференциальной экспрессии генов.

Типичная задача, в контексте которой при этом возникает дерево Штейнера, выглядит так: в некоторой группе генов, играющих роль множества терминальных вершин V_0, обнаружена существенная дифференциальная экспрессия, и есть подозрение, что тому есть общая причина в виде конкретных полиморфизмов[2], метаболических путей и т.д.

В качестве конкретного примера можно привести работу [BMC]. Авторы пишут, что гены, мутации в которых приводят к прогрессии онкозаболеваний, не всегда можно найти при помощи анализа дифференциальной экспрессии. Тем не менее, довольно часто они оказывают скрытое влияние на гены с ярко выраженной картиной дифференциальной экспрессии — так называемые биомаркеры рака, — входя с ними в одни и те же метаболические пути. К сожалению, для большинства генов в геноме человека неизвестно, в каких метаболических путях они участвуют. Тем не менее, сеть взаимодействий между генами довольно хорошо изучена. Более того, есть алгоритмы предсказания новых взаимодействий на основе уже имеющихся данных. Авторы демонстрируют, что поиск дерева Штейнера, соединяющего гены из некоторого поднабора достоверно известных биомаркеров, позволяет найти остальные и сформировать список кандидатов для дальнейшего исследования.

Степень уверенности в том, что для данного гена имеет место дифференциальная экспрессия, обычно выражается в терминах p-значений.[3] Для проверки гипотез о связях между такими генами в графе взаимодействий вводится весовая функция по следующему правилу: если хотя бы один из концов ребра не является терминальной вершиной, то весом будет степень уверенности в достоверности этого ребра[4], иначе учитывают также и вклад корреляции между их p-значениями. Т.е. дерево Штейнера должно максимизировать правдоподобие составляющих его рёбер, причём чем более выражена корреляция между признаками, тем б**о**льший вклад даёт ребро между ними. Построенное дерево позволяет выявить гены или метаболиты, которые оказывают наибольшее влияние на экспрессию генов из V_0, что позволяет выявить скрытые биологические закономерности, недоступные или неочевидные для непосредственного наблюдения.

Задача поиска дерева Штейнера в биологических сетях имеет свои особенности в связи как со статистической природой рёбер в таких сетях, так и с их структурой — большая часть таких сетей являются безмасштабными (scale-free). Вычислительные эксперименты [UT] показывают, что размеры таких графов, равно как и множества V_0, можно значительно (в разы) сократить, убрав из рассмотрения вершины, которые точно не войдут в дерево Штейнера, что позволяет подсчитать точный ответ (с помощью алгоритма Дрейфуса-Вагнера, как в [UT]).

[2] Полиморфизм — стабильный вариант гена, обычно отличающийся в одном нуклетоиде

[3] Конкретные способы подсчёта этих p-значений — отдельная сложная тема, выходящая за рамки данной работы

[4] Большая проблема подходов, использующих интерактом, в том, что все рёбра в нём отражают реальность только лишь с некоторой степенью уверенности. Существенная часть рёбер была предсказана на компьютере и не имеет экспериментального подтверждения, что вызывает критику в научном сообществе.

Больше про различные подходы к предобработке данных при поиске дерева Штейнера и о приложениях последнего в анализе биологических сетей можно прочитать в [UT]. Автор данной работы пока не использовал деревья Штейнера в своей научной практике, но заинтересован.

3.3 Краткий обзор других приложений

С изобилием вариаций задачи о дереве Штейнера можно ознакомиться в обзоре [UB].

4 Доказательство NP-полноты задачи

Теорема 4.1.

$\{(G, k) \mid$ в неориентированном графе G есть дерево Штейнера весом не более чем $k \in \mathbb{Z}\} \in$ NPC

Доказательство [EPFL].

1. **STEINER-TREE** $\in$ **NP**: Сертификат должен проверять, что поданный ему на вход подграф T является деревом, содержит все терминальные вершины и имеет вес $\leqslant k$, причём вторая и третья подзадачи тривиальны. Согласно одному из эквивалентных определений дерева, достаточно проверить связность T и то, что $|E(T)| = |V(T)| - 1$, для чего достаточно обхода в глубину. Т.е. полиномиальный сертификат существует и STEINER-TREE $\in$ NP.

2. **VERTEX-COVER** $\leqslant_p$ **STEINER-TREE**: Полиномиальное сведение устроено так:

   (a) Дополним $G = (V, E)$ до полного графа, после чего произведём подразделение всех исходных рёбер. Множество добавленных при этом вершин обозначим через W, а полученный в результате граф — через $G' := (V', E')$. Всем рёбрам назначим единичные веса, а весовую функцию будем как и раньше обозначать через ω.

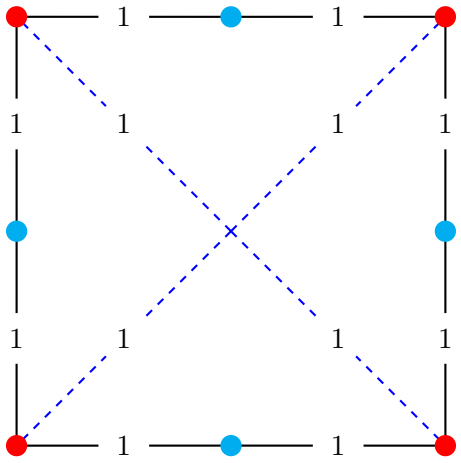

Рис. 3: *Пример конструкции: красные вершины — V, голубые — W, штрихпунктирные рёбра добавлены при дополнении до полного графа, сплошные были изначально. Все веса единичные.*

   (b) Если теперь в качестве множества терминальных вершин V'_0 взять W, то в G' есть дерево Штейнера веса $\leqslant |E| + k - 1 \iff$ в G есть вершинное покрытие мощности $\leqslant k$.

Доказательство.

- $\underline{\Longrightarrow}$: Пусть T — дерево Штейнера для V_0' в G', тогда $C := V(T) \setminus V_0'$ — вершинное покрытие в G. Действительно, $C \subset V$ и накрывает каждое ребро $e \in E$ по построению, ведь терминальные вершины появились при подразделении рёбер из E. Также:

$$|C| = |V(T)| - |V_0'| = \omega(T) + 1 - |V_0'| \leqslant (|E| + k - 1) + 1 - |V_0'| = k,$$

т.к. $|E| = |V_0'|$ по построению.

- $\underline{\Longleftarrow}$: Пусть $C \subset V$ — вершинное покрытие, $|C| \leqslant k$, а T — дерево на вершинах C в G'.[5] Хотим, чтобы это дерево покрывало множество терминальных вершин. Чтобы гарантировать, что $V_0' \subset V(T)$, расширим T, если это необходимо: для каждой вершины $v_0' \in V_0' \setminus V(T)$ добавим к T ребро (v_0', c), где $c \in C$ — вершина покрытия, накрывающая ребро, подразделением которого получена вершина v_0'. Полученный граф содержит $\leqslant |E| + |C| - 1 \leqslant |E| + k - 1$ рёбер. Если при добавлении рёбер в T образовались циклы, их можно раскрыть, уменьшив суммарный вес. По завершении получим дерево Штейнера веса $\leqslant |E| + k - 1$ для V_0' в G'.

□

Все шаги построения G' полиномиальны: добавляется $O(|V|^2)$ рёбер и вершин. Для восстановления вершинного покрытия по дереву Штейнера в G' нужно брать разность множеств — $O(|V|^2)$ операций, а в обратную сторону нужно перебрать V_0' и исходящие из него рёбра (не более двух на каждую вершину) — тоже $O(|V|^2)$ операций.

□

5 Сведение к метрическому случаю

Часто хочется потребовать, чтобы для весовой функции выполнялось правило треугольника:

$$\omega(x, y) \leqslant \omega(x, z) + \omega(z, y),$$

причём она должна быть определена на V^2. Такая весовая функция играет роль метрики на множестве вершин и становится проще для восприятия. В таком случае говорят о поиске *метрического* дерева Штейнера, и именно такой вид задачи был исторически первым.

Утверждение 5.1.

1. Существует полиномиальное сведение задачи о дереве Штейнера к метрическому случаю.
2. Оптимальные ответы к обеим задачам совпадают.

Доказательство. Предложенная конструкция основана на понятии *метрического замыкания*:

Определение 5.1. Пусть $G = (V, E, \omega)$ — неориентированный взвешенный граф, $d : V^2 \to \mathbb{R}^+$ — функция расстояния, сопоставляющая паре вершин длину кратчайшего пути между ними. Тогда граф $G' = (V', E')$ называется *метрическим замыканием* графа G:

$$G' = (V, E', d),\ E' = \{(u, v) \mid u, v \in V,\ u \neq v\}$$

Полученный граф является метрическим, т.к. для d выполнено правило треугольника: если $d(x, y) > d(x, z) + d(z, y)$, то путь $x \to y$ можно было бы прорелаксировать конкатенацией путей $x \to z$ и $z \to y$, что противоречит определению $d(x, y)$ как длины *кратчайшего* пути $x \to y$.
Построить метрическое замыкание можно за $O(|V|^3)$ алгоритмом Флойда-Уоршелла.

[5] Такое дерево всегда существует, т.к. $V^2 \subset V(G')$.

Пусть T, T' — деревья Штейнера для V_0 в G и G' соответственно. Докажем, что $\omega(T) = d(T')$. Очевидно, что $d(T') \leqslant \omega(T)$, т.к. переход к кратчайшим путям не ухудшает ответ. Более того, каждое ребро в T' можно «разжать» в тот кратчайший путь, из которого он был получен, после чего выбрать в полученном графе минимальное остовное дерево T''. $\omega(T'') \leqslant d(T')$, т.к. теперь каждое ребро встречается ровно один раз, а ранее могло вносить свой вклад одновременно в несколько кратчайших путей. Но T'' по построению — дерево Штейнера для V_0 в G! Следовательно, $\omega(T) = \omega(T'') \leqslant d(T') \leqslant \omega(T)$, т.е. $\omega(T) = d(T')$. □

6 2-оптимальный алгоритм

Теорема 6.1.
Существует 2-оптимальный алгоритм, основанный на сведении задачи к метрическому случаю:

1. Построить метрическое замыкание $G' = (V, E', d)$ графа G.
2. Выделить H' — подграф в G', индуцированный терминальными вершинами из V_0.
3. Построить минимальное остовное дерево T_{MST} в H'.
4. Вернуть в качестве ответа дерево в G, восстановленное из T_{MST} заменой рёбер между вершинами на соответствующие кратчайшие пути в графе G.

Доказательство [St].
Пусть T_S — дерево Штейнера для V_0 в G', $|V_0| = k + 1$.
Выпишем вершины T_S в порядке обхода в глубину. Будем добавлять вершину в список и при каждом выходе из неё по исходящему ребру, и по возвращению из рекурсивного вызова. Получим массив, называемый эйлеровым обходом дерева T_S:

$$u_0, u_1, \ldots, u_m = u_0$$

Такой обход задаёт Эйлеров цикл, поскольку в алгоритм обхода запрещает проходить по одному ребру дважды. Значит:

$$\sum_{i=0}^{m-1} d(u_i, u_{i+1}) = 2 \cdot d(T_S)$$

Если исключить из рассмотрения все нетерминальные вершины и оставить только первое вхождение всех терминальных, не меняя порядок вхождения вершин, то получим путь[6]:

$$v_0, v_1, \ldots, v_k$$

содержащий все терминальные вершины (т.к. изначально это был обход связного графа G').
По неравенству треугольника:

$$d(v_i, v_{i+1}) \leqslant \sum_{j=1}^{t} d(u_{i_j}, u_{i_{j+1}})$$

Здесь $u_{i_1}, \ldots, u_{i_{t+1}}$ — подотрезок в эйлеровом обходе, т.е. $v_i = u_{i_1}, v_{i+1} = u_{i_{t+1}}$, причём подотрезки для каждой пары (v_i, v_{i+1}) не пересекаются, т.к. в исходном эйлеровом обходе *первые* вхождения этих вершин встречаются именно в таком порядке.
Поскольку $v_0, \ldots, v_k$ — путь, это также дерево. Более того, в силу того, как задана весовая функция, это ещё и минимальное остовное дерево T_{MST} в H', причём:

$$d(T_{\text{MST}}) = \sum_{i=0}^{k-1} d(v_i, v_{i+1}) \leqslant \sum_{i=0}^{m-1} d(u_i, u_{i+1}) = 2 \cdot d(T_S)$$

$$1 \leqslant \frac{d(T_{\text{MST}})}{d(T_S)} \leqslant 2$$

[6] Поскольку мы работаем с метрическим замыканием, рёбра (v_i, v_{i+1}) в графе G' гарантированно есть.

По *утверждению 3.1.*, по нему восстанавливается дерево Штейнера в G. □

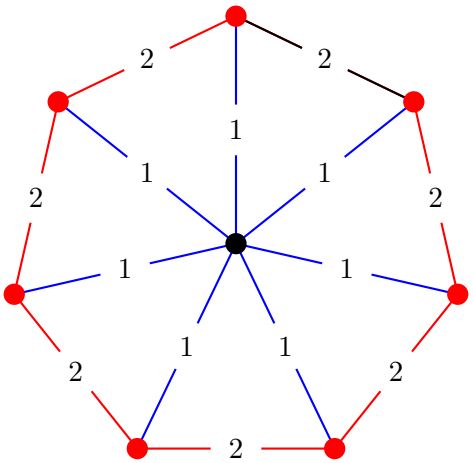

Рис. 4: Пример графа, на котором оценка достигается (в пределе). Терминальные вершины выделены красным. Реальное дерево Штейнера T_S выделено синим, а T_{MST}, построенное алгоритмом, — красным. Если внешний цикл имеет вид n-угольника, то верно, что $\frac{d(T_{\text{MST}})}{d(T_S)} = \frac{2n-2}{n} = 2 - \frac{2}{n} \to 2$.

7 Реализация

7.1 Асимптотика работы алгоритма и ограничения по его применению

1. Алгоритм Флойда-Уоршелла поиска кратчайших путей между всеми парами вершин: $\Theta(|V|^3)$ операций, $\Theta(|V|^2)$ памяти на хранение матрицы расстояний.

2. Алгоритм Краскала поиска минимального остовного дерева с помощью системы непересекающихся множеств: $O(|E| \log |V|)$ операций, $\Theta(|V|)$ памяти.

Т.к. метрическое замыкание делает граф полным, итого имеем $\Theta(|V|^3)$ операций и $\Theta(|V|^2)$ памяти. Следовательно, алгоритм оправданно применять при $|V| \leqslant 1000$, что является существенным недостатком предложенного подхода.

Также стоит отметить, что алгоритмы Флойда-Уоршелла и Краскала по своей природе последовательные, в связи с чем воспользоваться преимуществами вычислительного кластера МФТИ, к которому автор имеет авторизованный доступ, не представляется возможным при таком подходе.

Более того, алгоритм Флойда-Уоршелла хорош тем, что не совершает больших прыжков по памяти при хранении матрицы кратчайших расстояний в виде последовательно расположенной в памяти двумерной таблицы, что уменьшает число промахов по кэшу процессора и оказывает положительный эффект на скорость работы алгоритма по сравнению с параллельными запусками алгоритмов поиска кратчайших путей из данной вершины во все остальные (в духе алгоритма Дейкстры) при малых $|V|$.

Тем не менее, при $|V| \geqslant 10^3$ не лишено смысла хранить исходную матрицу расстояний в базе данных (на жёстком диске или распределённо) и действительно выполнять параллельные запуски, скажем, алгоритма Дейкстры на доступных процессорах для параллельного подсчёта матрицы кратчайших расстояний. Эта неасимптотическая оптимизация делает возможной обработку существенно б**о**льших графов, чем было заявлено изначально, т.к. это типичная map-reduce задача,

что с точки зрения автора является довольно дешёвой в реализации, но полезной идеей, пусть и выходящей за рамки требований поставленной учебной задачи.

7.2 Использованные технологии

Для работы с графами в Python была использована библиотека graph-tool. По опыту автора, альтернативы в лице networkx и igraph имеют слишком много нетривиальных недостатков: networkx написана на Pure Python, из-за чего производительность становится неприемлемой уже на маленьких графах, а разработка Python-API для написанной на C библиотеки igraph уже больше года как приостановлена, а по состоянию на данный момент пользоваться ей очень уж неудобно по причинам, расписывать которые слишком долго для того, чтобы приводить их здесь. Справедливости ради, R-API для igraph гораздо лучше, но Python для меня предпочтительнее.

7.3 Проверка корректности

Тесты разделены на несколько групп:

1. Маленькие ($|V| \leqslant 20$) случайные графы в модели Эрдеша-Реньи, реальное дерево Штейнера в которых ищется полным перебором (возможно, с некоторыми эвристиками).

2. Некоторые последовательности графов с известным точным ответом.

Написание тестов, иллюстрирующих биологические приложения алгоритма, затруднительно в силу необходимости знаний предметной области для интерпретации результатов. Заинтересованный читатель может ознакомиться с ними в [BMC].

7.4 Результаты

Ознакомиться с реализацией можно по ссылке.
Поскольку построение метрического замыкания — ключевой этап алгоритма, оно же ограничивает его применимость. Даже при параллельном поиске кратчайших путей алгоритмом Дейкстры или его модификациями и с хранением графа и матрицы расстояний в разделяемой памяти, не удастся добиться асимптотики лучше, чем $O(|V|^3)$ по времени и $O(|V|^2)$ по памяти, в связи с чем использовать алгоритм при $|V| \geqslant 10^4$ нерационально и непрактично.
Тем не менее, алгоритм прост в реализации, а его показатель аппроксимации отличается от оптимального на данный момент показателя в ≈ 1.55 [CSV] всего на ≈ 0.45.

8 Список литературы

[1] [EPFL] Chair of Combinatorial Geometry, EPFL — Lectures on Computational Complexity Theory

[2] [St] Luca Trevisan; Stanford University — "Approximating the Metric Steiner Tree Problem"

[3] [UB] M.Hauptmann, M.Karpinski; Universität Bonn - A Compendium On Steiner Tree Problems

[4] [UT] N. Betzler; Universität Tübingen — "Steiner Tree Problems in the Analysis of Biological Networks"

[5] [UEL] Regina Krisztina Bíró, Eötvös Loránd University — "Constructing Phylogenetic Trees"

[6] [BMC] Md Jamiul Jahid, Jianhua Ruan; BMC Genomics. 2012; 13(Suppl 6): S8. — "A Steiner tree-based method for biomarker discovery and classification in breast cancer metastasis"

[7] [CSV] Gabriel Robinsy, Alexander Zelikovsky; — "Improved Steiner Tree Approximation in Graphs"