

МЕТОДЫ ОПТИМИЗАЦИИ СТРУКТУРЫ СЛОЖНЫХ СИСТЕМ И ИХ ПРИМЕНЕНИЕ В МАТЕМАТИЧЕСКОЙ ХИМИИ

Михаил Губко

ИПУ РАН

23 декабря 2014 года

Аннотация: Задачи оптимизации структуры системы возникают в самых разных предметных областях. Многие из них сводятся к поиску допустимой структуры, минимизирующей или максимизирующей заданный критерий. Множество допустимых структур обычно настолько обширно, что его полный перебор исключен. Конструктивные аналитические и алгоритмические методы оптимизации иерархических структур удается разработать в рамках концепции секционных функций затрат (Воронин, Мишин 2003). К оптимизации секционных функций сводится широкий круг прикладных задач, однако желание применить теорию к неиерархическим, сетевым структурам потребовало развития этой модели. Недавно разработанная модель оптимальной связывающей сети обобщает концепцию секционных функций и также имеет многочисленные приложения.

В докладе описывается применение разработанных методов поиска оптимальных связывающих сетей в математической химии, для решения задач оптимизации топологических молекулярных индексов. Приводится точная нижняя оценка значений обобщенного первого загребского индекса, второго загребского индекса, ABC-индекса для деревьев с заданным числом листьев, характеризуются экстремальные деревья. С помощью обобщенного алгоритма Хаффмана строится дерево, минимизирующее индекс Винера для взвешенных вершин (Klavzar, Gutman 1997) на множестве деревьев с заданными весами и степенями вершин. Характеризуются деревья, минимизирующие индекс Винера на множестве химических деревьев с заданными весами вершин. Обсуждается применение этих результатов в задачах поиска веществ с экстремальными физико-химическими свойствами.

План

1) Задачи оптимизации структуры

- Общая постановка задачи
- Оптимизация иерархий и секционные функции затрат
- Приложения
- Оптимальные связывающие сети

2) Приложения в математической химии

- Задача молекулярного дизайна как обратная задача управления
- Оптимизация индексов, зависящих от степеней вершин
- Оптимизация индексов, зависящих от расстояний
- Примеры приложений: спирт с минимальной T кипения

3) Бонусы:

- Метамоделли оптимизации иерархий
- Перспективные приложения

Общая постановка задачи оптимизации структуры

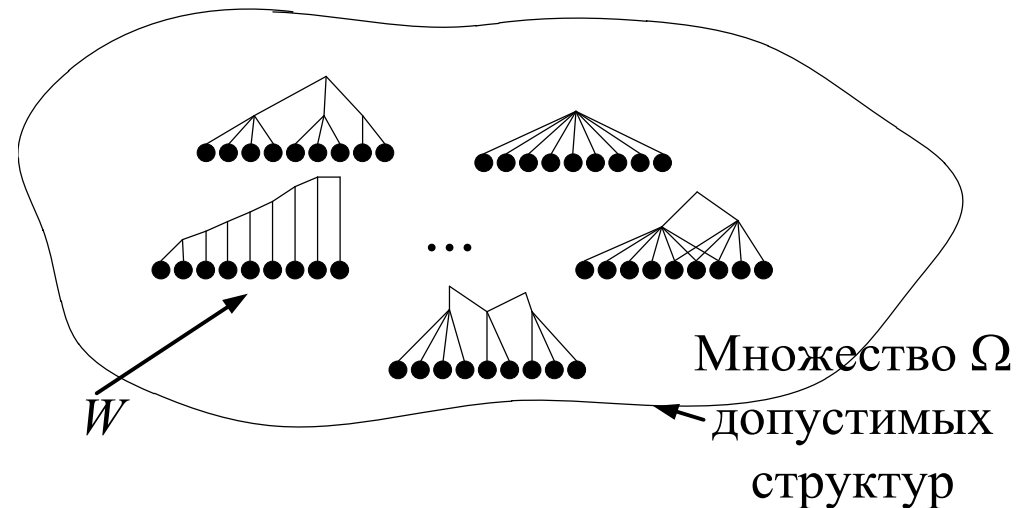
Дано:

- 1) множество базовых вершин W
- 2) множество допустимых структур Ω – ориентированных или неориентированных графов
 - Обычно рассматриваются только связные графы
 - Если $V(G)$ – множество вершин допустимой структуры $G \in \Omega$, то $W \subseteq V(G)$
- 3) критерий эффективности или функция затрат $C : \Omega \rightarrow \mathbb{R}$.

Задача:

найти $H^* \in \underset{H \in \Omega}{\text{Arg min}} C(H)$.

- Проблема обычно в том, что множество Ω допустимых структур слишком обширно, чтобы задачу можно было решать перебором.
- Первоначально задача формулировалась для иерархий (ациклических направленных графов с заданным множеством начальных вершин).



Секционные функции затрат

$W = \{w_1, \dots, w_n\}$ – конечное множество терминальных вершин, $n > 1$.

Подчиненная группа $s_H(m)$ – подмножество терминальных вершин, непосредственно или опосредовано подчиненных вершине m в иерархии H .

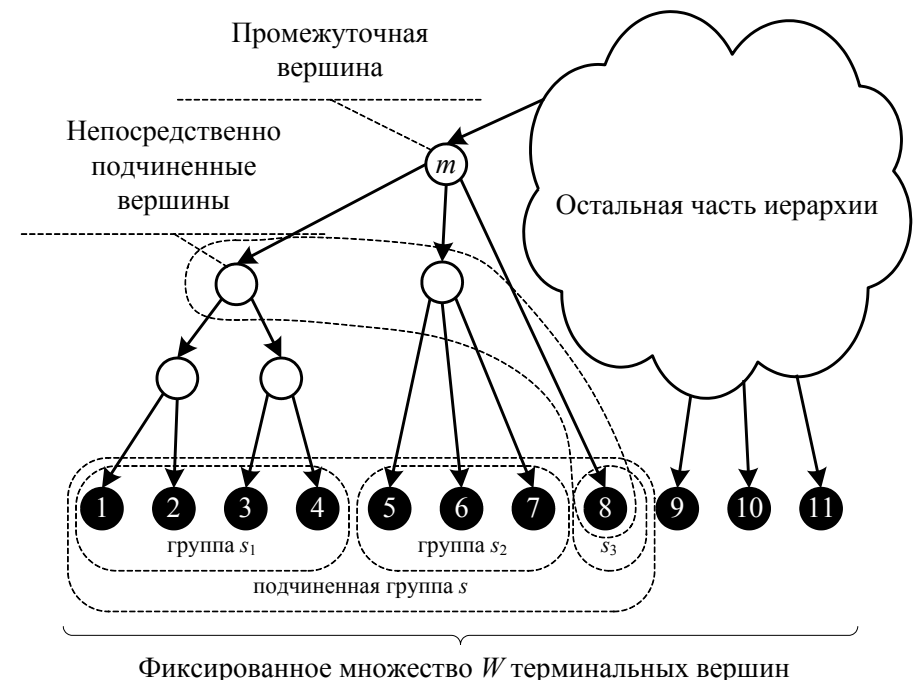
При **секционной функции** затраты вершины зависят от набора групп $s_H(v_1), \dots, s_H(v_r)$, подчиненных ее непосредственным подчиненным v_1, \dots, v_r :

$$c_H(m) = c(s_H(v_1), \dots, s_H(v_r))$$

Например, $c(\{1, 2, 3, 4\}, \{3, 4, 5, 6, 7\}, \{8\})$

Затраты иерархии равны сумме затрат вершин

- **Анонимность** затрат по нетерминальным вершинам (но не по терминальным вершинам)
- **Локальность** затрат вершины:
 - Зависят только от групп, подчиненных непосредственным подчиненным вершины
 - Не зависят от того, как организована иерархия «под» непосредственными подчиненными
 - Не зависят от вышестоящей части иерархии



Приложения

Информационные системы

- Оптимальное кодирование
- Структура алгоритмов
- Иерархическая декомпозиция задач и параллельные вычисления
- Оптимизация пользовательских меню
- Разработка структуры классификаторов
- Построение деревьев решений для анализа данных
- Структура иерархических индексов СУБД

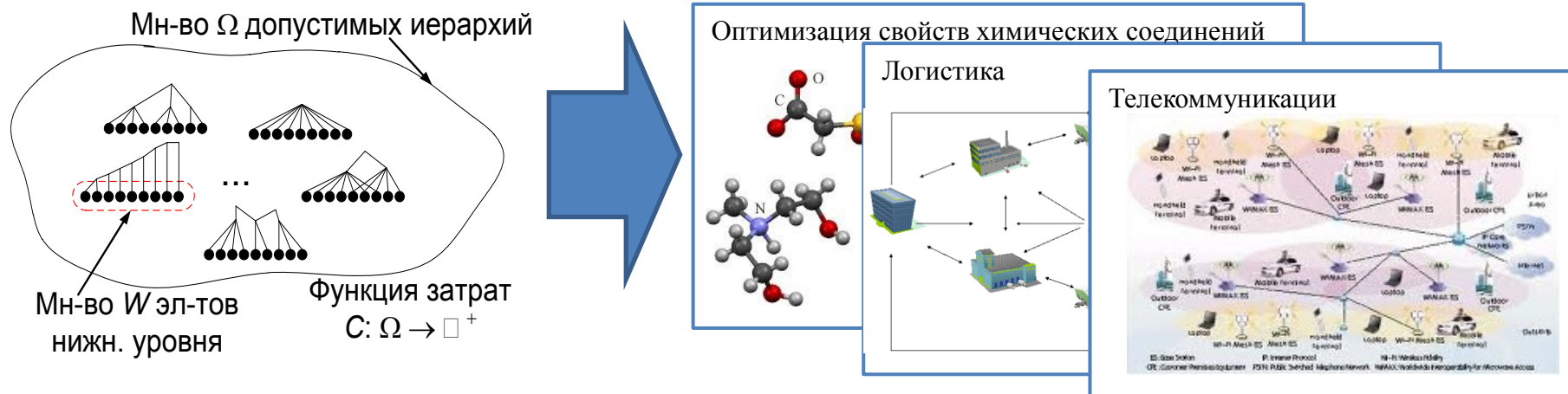
Организационные системы

- Реинжиниринг организационно-штатного расписания
- Модели иерархической организации для теории фирмы
- Сети сбора данных
- Оптимизация структуры сетей поставки (supply networks)

Технические системы

- Структура иерархических сетей сотовой связи
- Планирование производства (балансировка сборочной линии)
- Организация вычислений и агрегирование в реальном времени
- Структура вычислительных сетей

От оптимизации иерархий – к оптимизации сетевых структур



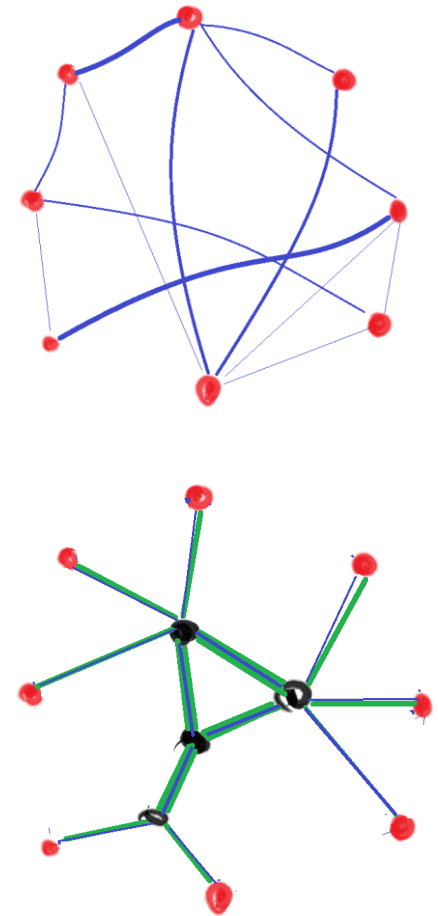
- Для моделирования иерархических структур используются направленные ациклические графы. Но чтобы оптимизировать сетевые структуры, основанные на неиерархических «горизонтальных» связях, имеющиеся методы нужно дорабатывать, что и является основной задачей настоящего проекта.
- Оказалось, что многие ранее разработанные методы оптимизации иерархических структур могут быть переформулированы в терминах неориентированных графов.

Модель связывающей сети

- Задан *граф связей* R на множестве W из n вершин нижнего уровня. В частности, его дуги могут быть взвешены p -мерными векторами $x_R(w, w')$ – объемами потоков.
- Для обеспечения потоков создается *сеть* (орграф) H , в которой они являются висячими вершинами, а внутренние вершины коммутируют потоки.
- Движение потоков определяется *функциями маршрутизации* $\lambda_{(v, v')}(w, w') \rightarrow [0, 1]$, определяющими долю каждого потока (w, w') в дуге (v, v') .
- В *допустимой сети* выполняются требования баланса потоков в каждой вершине.
- Затраты внутренней вершины m определяются ее входящими и исходящими потоками.
- Если u_1m, \dots, u_km – входящие дуги, а mv_1, \dots, mv_r – исходящие, то

$$c(m, H) = c(\lambda_{u_1m}(\cdot), \dots, \lambda_{u_km}(\cdot), \lambda_{mv_1}(\cdot), \dots, \lambda_{mv_r}(\cdot)).$$
- Задача – найти допустимую сеть с минимальными затратами.

Для сетей древовидной структуры задача сводится к поиску оптимальной древовидной иерархии для **секционной функции** затрат.



Аддитивная функция затрат

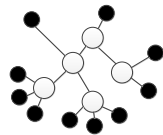
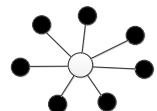
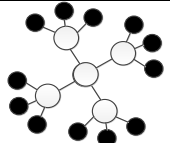
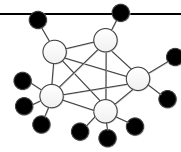
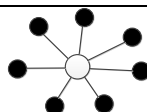
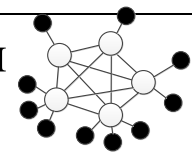
Определение. Аддитивная функция затрат $c(m, H) = c_1(d_H(m)) + c_2(x_H(m))$, где $d_H(m)$ – степень вершины m в сети H , а $x_H(m)$ – поток через вершину m . Обозначим

$$C_1(H) := \sum_{m \in M} c_1(d_H(m)), \quad C_2(H) := \sum_{m \in M} c_2(x_H(m)).$$

Лемма. Пусть $\Omega_1 := \text{Argmin}_{H \in \Omega} C_1(H)$, \underline{C}_1 – нижняя оценка $C_1(H)$, \underline{C}_2 – нижняя оценка $C_2(H)$ при $H \in \Omega$, и сеть $H_1 \in \Omega_1$. Тогда $\underline{C} = \underline{C}_1 + \underline{C}_2$ – нижняя оценка затрат сети для функции затрат вершины $c_1(\cdot) + c_2(\cdot)$, относительная погрешность затрат сети H_1 $\varepsilon(H_1) = (C_2(H_1) - \underline{C}_2) / (\underline{C}_1 + \underline{C}_2) = \underline{\varepsilon}$, где $\underline{\varepsilon}$ – относительная погрешность \underline{C} .

Для поиска H_1 и вычисления ее погрешности $\varepsilon(H_1)$ необходимо уметь:

1. Решать оценочную задачу для функции затрат сети $C_2(H) = \sum_{m \in M} c_2(x_H(m))$,
2. Находить множество Ω_1 сетей, доставляющих минимум $C_1(H) = \sum_{m \in M} c_1(d_H(m))$ и искать на нем оптимальное или приближенно оптимальное дерево H_1 : $C_2(H_1) \cong \min_{H \in \Omega_1} C_2(H)$.

	Монотонная	Субаддитивная	Выпуклая
$c_1(d_H(m))$	Дерево 	«Звезда» 	Однородное дерево 
$c_2(x_H(m))$	Двухуровневая сеть 	«Звезда» 	Двухуровневая сеть 

Оптимальные сети для функции затрат зависящей от потока

Чтобы найти оптимальную сеть, нужно определить:

1. Количество внутренних вершин q ,
2. их степени r_1, \dots, r_q ,
3. структуру сети, связывающей внутренние вершины,
4. число вершин нижнего уровня k_m , инцидентных внутренней вершине $m = 1, \dots, q$,
5. распределение вершин нижнего уровня по внутренним вершинам,
6. функцию маршрутизации потоков.

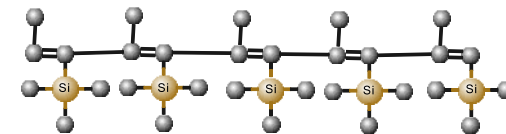
Утверждение. Пусть древовидная сеть H содержит $q \geq 1$ внутренних вершин, первые p из них имеют по k_m инцидентных вершин нижнего уровня ($m = 1, \dots, p$), $D(\alpha_1, \dots, \alpha_q)$ – матрица вершинных расстояний между p этими вершинами в сети H , внутренние вершины которой взвешены числами $\alpha_1, \dots, \alpha_q$, и $K = \text{diag}(k_1, \dots, k_q)$.

Тогда для выпуклой функции $c_2(\cdot)$ затраты сети H оцениваются снизу выражением

$$C_2(H) \geq \sup_{\alpha_1, \dots, \alpha_q} \left\{ \sum_{m=1}^q \left\langle c_2([c_2']^{-1}(\alpha_m)) - \alpha_m [c_2']^{-1}(\alpha_m) + \alpha_m \mathbf{1}_n^T R \cdot \mathbf{1}_n \right\rangle - \sum_{i=1}^p \lambda_i(R) \lambda_i \left(K^{1/2} \left[\sum_{m=1}^q \alpha_m \mathbf{1}_p \mathbf{1}_p^T - D(\alpha_1, \dots, \alpha_q) \right] K^{1/2} \right) \right\},$$

где $\lambda_1(A) \geq \dots \geq \lambda_n(\cdot)$ – собственные числа матрицы A .

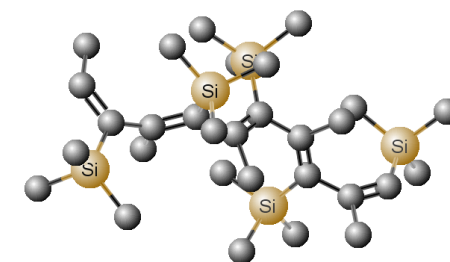
Задача молекулярного дизайна



- Физико-химические свойства вещества P_i , $i = 0, \dots, k$, определяются структурой G его молекулы через числовые характеристики – молекулярные дескрипторы.

$$\tilde{P}_i(G) = \tilde{P}_i(I_1(G), \dots, I_m(G)), i = 0, \dots, k,$$

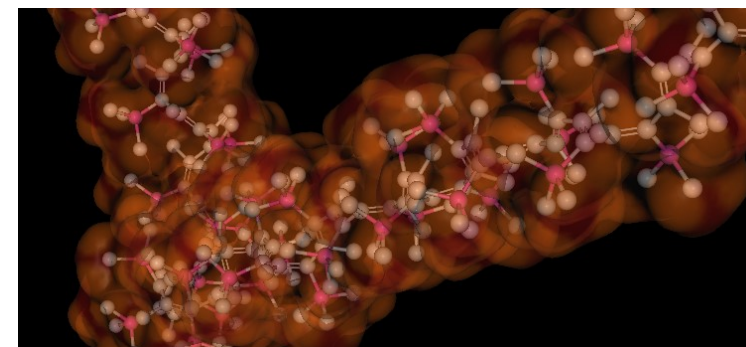
- За последние десятилетия разработано большое число топологических, геометрических, квантово-механических индексов.
- Поиск материалов с заданными свойствами сводится к оптимизации одних индексов при ограничениях на другие.



$$P_0(G) \rightarrow \min_{G \in \Omega} (\max_{G \in \Omega})$$

$$P_i^{min} \leq P_i(G) \leq P_i^{max}, i = 1, \dots, k.$$

- Наиболее интенсивно разрабатывались методы оптимизации топологических индексов.



Топологические индексы

- Зависящие от степеней вершин
 - a. $M_1(G) := \sum_{v \in V(G)} d_G(v)^2$ – первый загребский индекс, где $V(G)$ – множество вершин графа G , $d_G(v)$ – степень вершины v в G .
 - b. $M_2(G) := \sum_{uv \in E(G)} d_G(u)d_G(v)$ – второй загребский индекс, где $E(G)$ – множество дуг графа G .
- Зависящие от расстояний в графах
 - a. $WI(G) := \frac{1}{2} \sum_{u,v \in V(G)} d_G(u,v)$ – индекс Винера, где $d_G(u,v)$ – расстояние между вершинами u и v в графе G
- Спектральные (зависящие от спектров матриц графов)
 - a. $EE(G) = \text{tr } e^{A(G)} = \sum_{i=1}^n e^{\lambda_i(A(G))}$ – индекс Эстрады, где $A(G)$ – матрица смежности графа G , $\lambda_1(A) \geq \lambda_2(A) \geq \dots \geq \lambda_n(A)$ – собственные числа матрицы A .
- Фрагментные
 - a. Число $n_i(G)$ вершин степени i , число путей длины 3, циклов длины 3, ...
- Информационные индексы



Обобщенный первый загребский индекс

$$C_1(G) := \sum_{v \in V(G)} c(d_G(v)) = \sum_{i=1}^{|V(G)|} n_i(G) c(i)$$

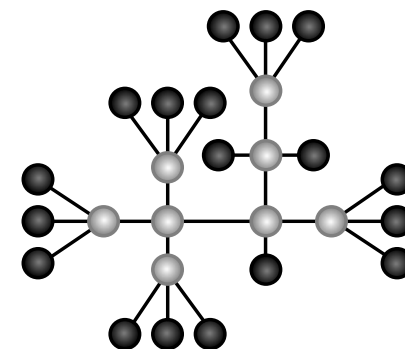
Теорема 1. Для любого дерева T с n листьями

$$C_1(T) \geq nc(1) + (n-2) \frac{c(\Delta(n))}{\Delta(n)-2} \text{ где } \Delta(2) = 3, \Delta(n) \in \text{Arg} \min_{d=3, \dots, n} \frac{c(d)}{d-2} \text{ при } n > 2.$$

Следствие 1. Для любого дерева T с n листьями $M_1(T) \geq 9n - 16$ для четного n , $M_1(T) \geq 9n - 15$ для нечетного n . Равенство достигается на 4-деревьях и «почти» 4-деревьях соответственно.

В дальнейшем было доказано, что равенство достигается **только** на таких деревьях, и для произвольной выпуклой функции $c(\cdot)$ были полностью охарактеризованы деревья с n листьями, доставляющие минимум $C_1(\cdot)$

Теорема 2 [Гутман 2014]. Для любого графа G с n листьями и $k = 0, 1, 2$ циклами $M_1(G) \geq 9n - 16 + k$ для четного n , , причем равенство достигается, если все внутренние вершины имеют степень 4, $M_1(G) \geq 9n - 15 + k$ для нечетного n .



Gutman I., Goubko M.V. Trees with fixed number of pendent vertices with minimal first Zagreb Index // Bulletin of International Mathematical Virtual Institute, 2013. V. 3(2). P.161-164.
 Goubko M. Minimizing Degree-Based Topological Indices for Trees with Given Number of Pendent Vertices // MATCH Commun. Math. Comput. Chem. 2014. V. 71, No 1. P. 33-46.
 Goubko M., Gutman I. Degree-based topological indices: Optimal trees with given number of pendants // Applied Mathematics and Computation. V. 240, 1 August 2014, P. 387-398.

Второй загребский индекс

$$M_2(G) := \sum_{uv \in E(G)} d_G(u)d_G(v).$$

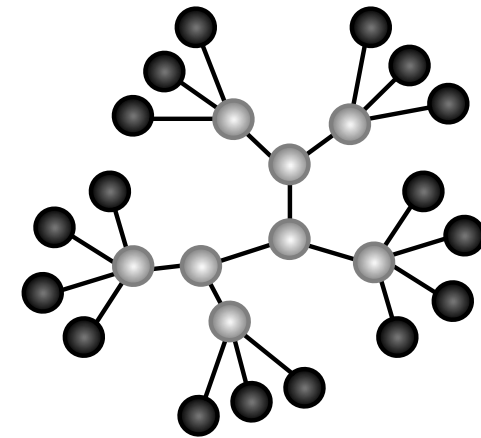
Теорема 3. Для любого дерева T с $n \geq 9$ листьями $M_2(T) \geq 11n - 27$, причем равенство достигается на дереве, изображенном на рисунке.

Впоследствии были найдены оптимальные химические деревья (степень вершины ≤ 4), охарактеризованы все деревья с $n \geq 2$ листьями, доставляющие минимум $M_2(\cdot)$.

Обобщенный второй загребский индекс

$$C_2(G) := \sum_{uv \in E(G)} c(d_G(u), d_G(v)).$$

Для произвольной $c(\cdot, \cdot)$ в принципе, можно доказывать аналоги теоремы 3, но не всегда.

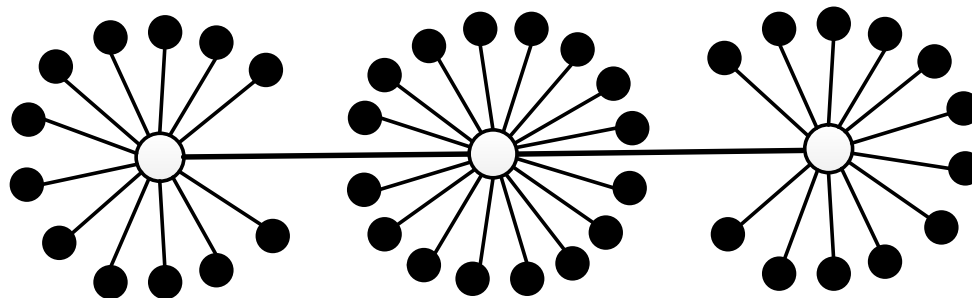


M. Goubko, T. Reti. Note on Minimizing Degree-Based Topological Indices of Trees with Given Number of Pendent Vertices // MATCH ... 2014. V. 72, No 3. pp. 633-639.

Goubko M., Gutman I. Degree-based topological indices: Optimal trees with given number of pendants // Applied Mathematics and Computation. V. 240, 1 August 2014, P. 387-398.

ABC-индекс (Atom Bond Connectivity Index)

$$ABC(G) := \sum_{uv \in E(G)} \sqrt{\frac{d_G(u) + d_G(v) - 2}{d_G(u)d_G(v)}}$$



Дерево с 40 листьями, доставляющее минимум $ABC(\cdot)$

Теорема 4. Для любого химического дерева T с $n \geq 4$ листьями, если n – четное, то

$$ABC(T) \geq \left(\frac{\sqrt{3}}{2} + \frac{\sqrt{6}}{8} \right) n - \frac{\sqrt{6}}{2},$$

причем равенство достигается на 4-дереве. Если n

нечетно, то $ABC(T) \geq \left(\frac{\sqrt{3}}{2} + \frac{\sqrt{6}}{8} \right) n - \sqrt{3} - \frac{5\sqrt{6}}{8} + \frac{4\sqrt{2} + \sqrt{5}}{2\sqrt{3}}$, причем равенство

достигается на «почти 4-дереве».

Оптимизация линейных комбинаций индексов

$$C_1(G) + bM_2(G) = \sum_{v \in V(G)} c(d_G(v)) + b \sum_{uv \in E(G)} d_G(u)d_G(v)$$

Задача – минимизация $C_1 + bM_2$ на множестве химических деревьев с n вершинами.

Определение. Экстремально разветвленное химическое дерево – не более одной нелистьевой вершины имеет степень $\neq 4$.

Теорема 5. Пусть выполнены неравенства

$$c(1) + c(4) + 18b < c(2) + c(3),$$

$$c(1) + c(3) + 8b < 2c(2),$$

$$c(2) + c(4) + 8b < 2c(3).$$

Тогда **любое** экстремально разветвленное дерево доставляет минимум $C_1(\cdot) + bM_2(\cdot)$ на множестве химических деревьев с $n \geq 3$ вершинами.

Индекс Винера для графов с взвешенными вершинами

$$VWWI(G) := \frac{1}{2} \sum_{u,v \in V(G)} \mu_u \mu_v d_G(u,v)$$

Задача 1. Минимум $VWWI$ для деревьев с множеством вершин V с фиксированными весами μ_v и степенями d_v , $v \in V$.

Определение. Веса μ_v монотонны по степеням d_v , если

$$d_u > d_v > 1 \Rightarrow \mu_u \geq \mu_v.$$

Теорема 6. Если веса монотонны по степеням, то дерево Хаффмана решает задачу 1.

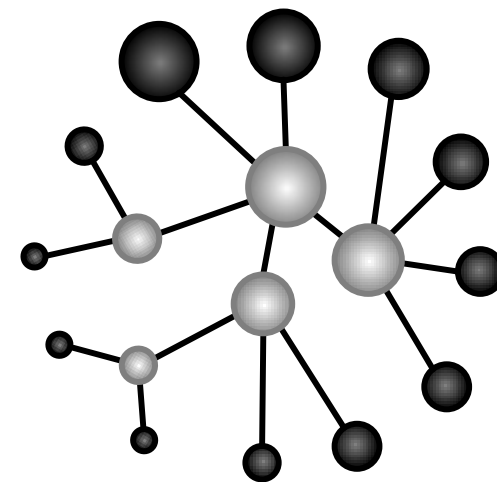
Задача 2. Минимум $VWWI$ для химических деревьев с множеством вершин V с фиксированными весами μ_v .

Определение. Пусть по всем допустимых химических деревьях вершины некоторого множества L – листья. *Экстремальная последовательность степеней* назначает внутренними $\lceil (|V|-2)/3 \rceil$ самых «тяжелых» вершин $v \in V \setminus L$, причем лишь самая легкая из них может иметь степень $d_v < 4$.

Теорема 7. Д. Хаффмана для экстремальной последовательности степеней решает задачу 2.

Goubko M. Minimizing Wiener Index for Vertex-Weighted Trees with Given Weight and Degree Sequences // upcoming in MATCH Commun. Math. Comput. Chem, 2015

Goubko M., Miloserdov O. Simple Alcohols with the Lowest Normal Boiling Point Using Topological Indices // upcoming in MATCH Commun. Math. Comput. Chem, 2015



Индекс Винера для графов с взвешенными парами вершин

$$PWWI(G) := \frac{1}{2} \sum_{u,v \in V(G)} \mu_{uv} d_G(u,v)$$

- Имеется матрица $R = (\mu_{uv})$ весов пар вершин.
- Частный случай – **vertex distance** $\Delta_G(v)$, где единичный вес имеют только пары, включающие заданную вершину.
- Если вершине $v \in V(G)$ назначить вес $\mu_v := 1/\varepsilon$, а остальным вершинам – вес $\mu_u := \varepsilon$, то

$$\lim_{\varepsilon \rightarrow 0} PWWI(G) = \Delta_G(v)$$

- Задача сводится к **модели связывающей сети** и полностью пока не решена.

Алифатические насыщенные спирты с минимальной T кипения

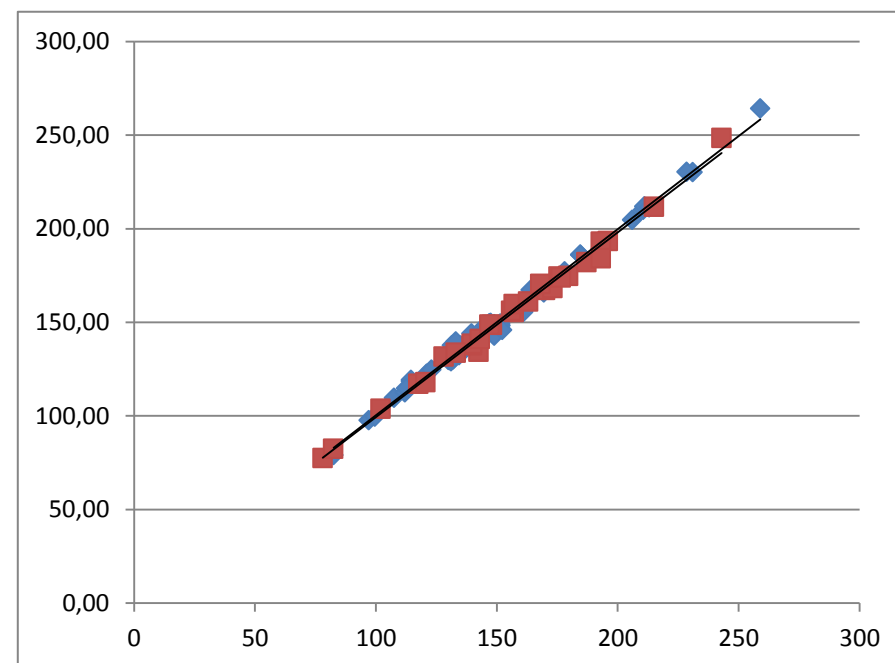
Факторы, влияющие на температуру кипения: разветвленность изомера, эффект экранирования гидроксильной группы ОН

$$BP^0 = b_0^0 + b_1^0 WI_O(G)^{1/3} + c^0(2)n_2(G) + c^0(3)n_3(G) + b_2^0 S_2(G) + b_3^0 M_2(G)$$

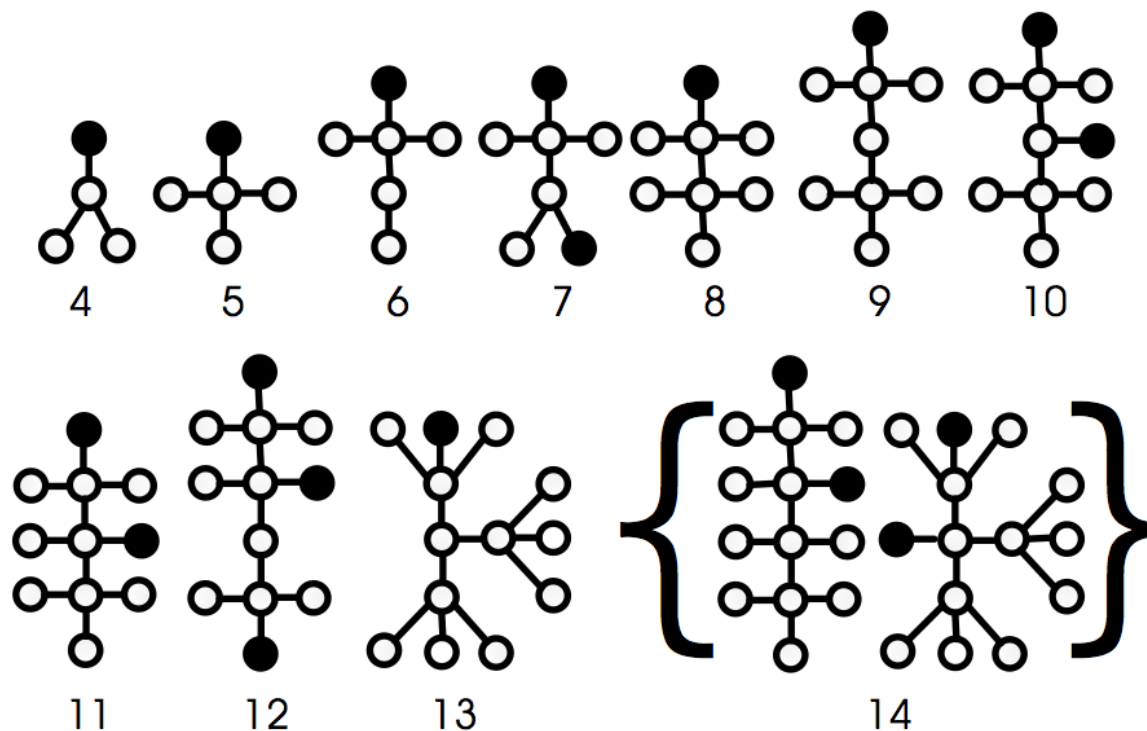
$$BP^I = b_0^I + c^I(2)n_2(G) + c^I(3)n_3(G) + c^I(4)n_4(G) + b_2^I S_2(G) + b_3^I M_2(G)$$

$$BP^{II} = b_0^{II} + b_1^{II} WI_O(G)^{1/3} + c^{II}(2)n_2(G) + c^{II}(3)n_3(G) + c^I(4)n_4(G) + b_2^I S_2(G)$$

где $S_i(G)$, $i = 2, \dots, 4$ – индикатор степени атома углерода инцидентного ОН-группе.

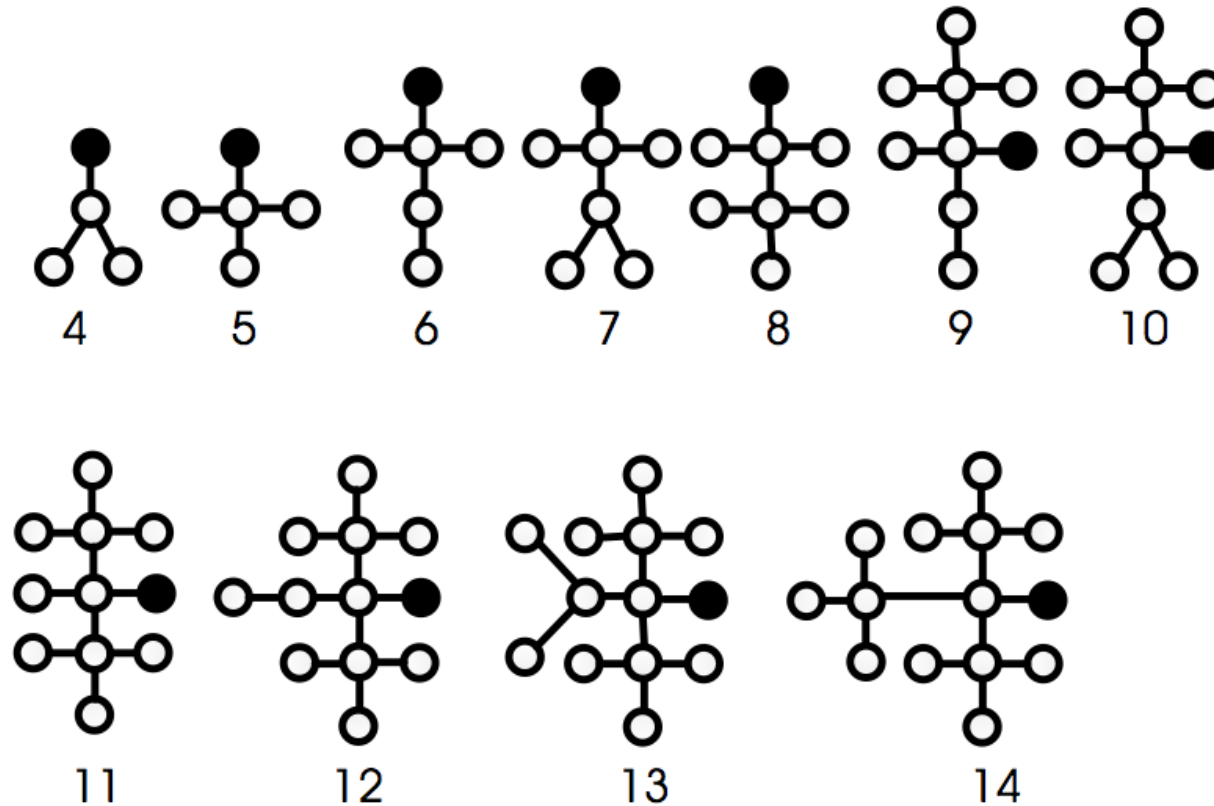


Утверждение 1. Для BP^I выполняются условия Теоремы 5. Структуры спиртов, изображенные на рисунке, минимизируют BP^I .



Структуры спиртов, минимизирующих BP^I для $n = 4, \dots, 14$ (возможные положения атома кислорода обозначены черной точкой)

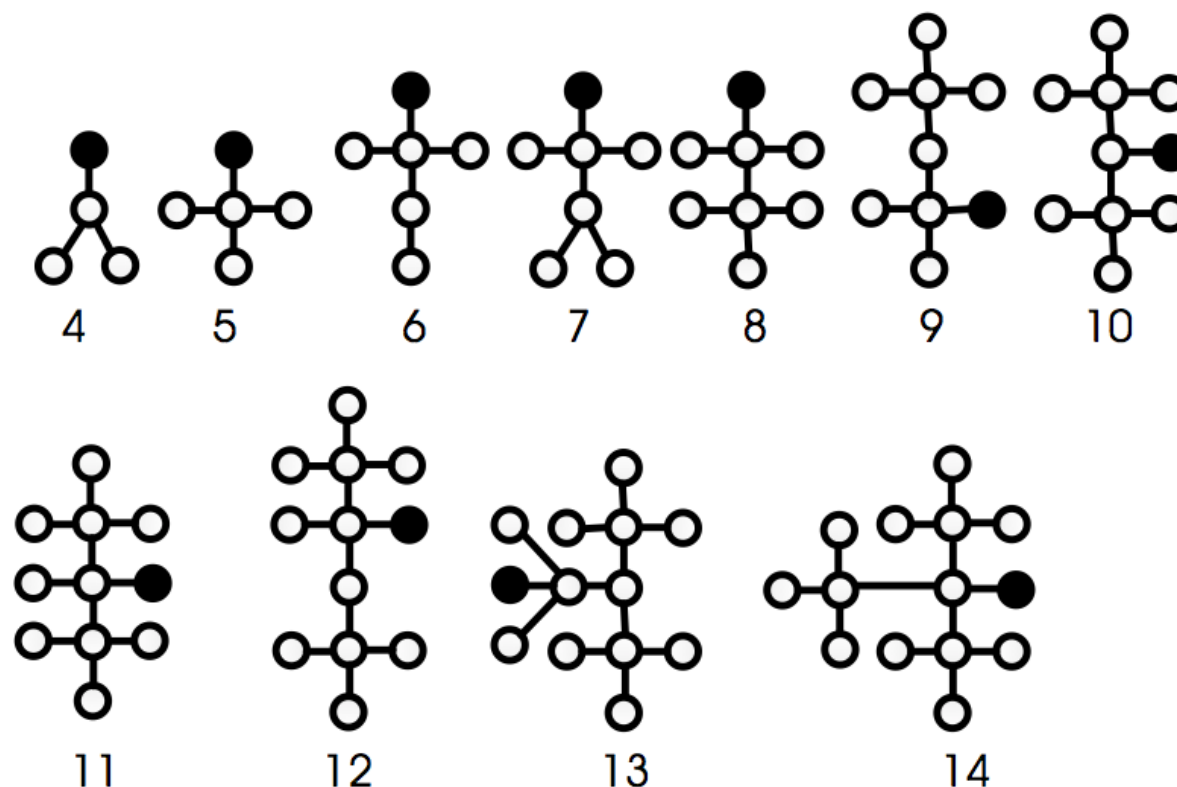
Утверждение 2. Для BP^{II} выполняются условия Теоремы 7. Структуры спиртов, изображенные на Рисунке, минимизируют BP^{II} .



Структуры спиртов, минимизирующих BP^{II}

Утверждение 3. Из Теорем 5 и 7 следует, что оптимальное дерево является экстремально разветвленным.

Оптимальное экстремально разветвленное дерево находим перебором.

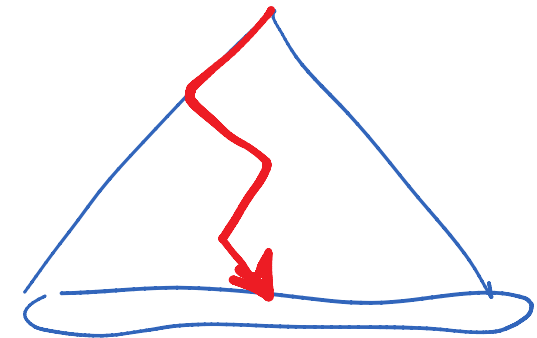


Структуры спиртов, минимизирующих BP^0

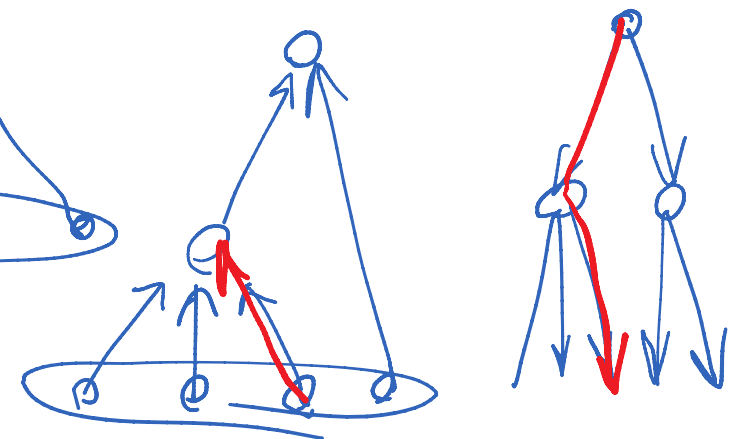
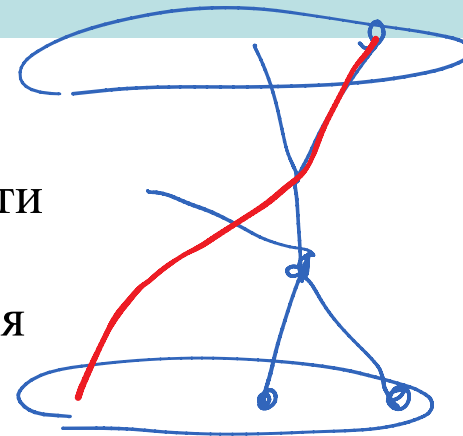
Метамоделли

1. Деревья поиска. Листья – искомые объекты, а дерево описывает порядок поиска, где ветвления – это принимаемые решения. Каждое использование дерева порождает путь от корня к одному из листьев.

- Структура алгоритмов поиска (оптимизация иерархии ветвлений)
- Деревья поиска (индексы)
- Иерархические пользовательские меню
- Оптимальные вопросы и деревья решений
- Оптимальное дерево вычисления логической функции



2. Сети управления потоками. По дугам перемещаются материальные или информационные потоки, а вершины сети занимают маршрутизацию. Частный случай – задачи сбора и распространения информации (один приемник или источник).



- Модель делегирования решения проблем (поток проблем)
- Иерархии сбора информации (минимизация максимального времени)
- Иерархическая структура сетей мобильной связи
- Оптимизация цепочек поставок и дистрибьюторские сети
- Структура системных сетей

3. Сети операций. Вершины сети – операции, связи – ограничения причинности

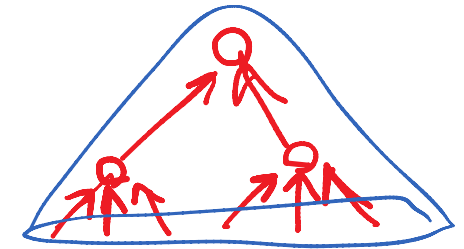
- Сетевые графики проектов
- Организационная структура как вычислительная иерархия (сборка)
- Иерархические вычисления - пример иерархии сумматоров (сборка)

• Оптимизация дихотомического представления функций (сборка)

• Иерархическая декомпозиция решения разреженных систем ЛУ (сборка)

• Проектирование структуры сборочного производства (сборка)

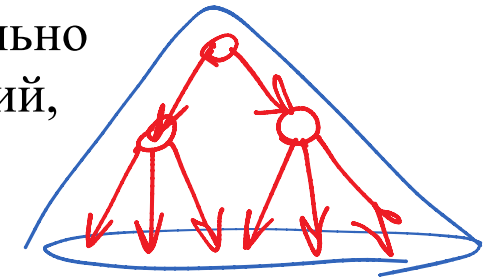
• Иерархия исполнения приказов и детализации планов (разборка)



4. Иерархии декомпозиции. Корень – целое, которое последовательно декомпозируется до «элементарных» частей (детализация функций, целей, пространственной структуры).

• Классификация и декомпозиция объекта, декомпозиция бизнес-процессов

• Иерархическая кластеризация



5. Иерархии контроля. Вершины вышестоящего уровня контролируют непосредственных подчиненных. Потребность в контроле присуща всем вершинам.

• Модели потери контроля и однородные функции

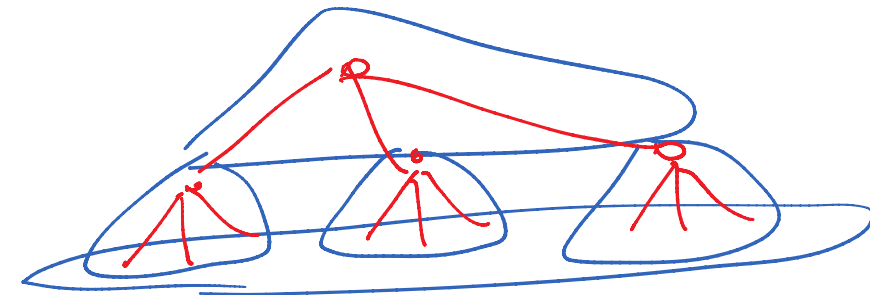
• Скорость роста функции затрат иерархии

• Механизмы стимулирования менеджеров

• Идентификация функции затрат на содержание менеджеров

• Модель совместной оптимизации иерархии и объема выпуска

• Оптимизация иерархии контроля исполнения бизнес-процессов



Перспективные приложения

Предсказание и оптимизация конформаций полимеров

Включение методов оптимизации свойств веществ в автоматизированные системы молекулярного дизайна