Отбор признаков в графе

Обзор "Supervised Feature Selection in Graphs with Path Coding Penalties and Network Flows"[JMLR, 2013]

Вихрева Мария, ВМК МГУ

2 декабря 2014

Формулировка задачи

"Supervised Feature Selection in Graphs with Path Coding Penalties and Network Flows"[2013]

Признаки представимы в виде графа:

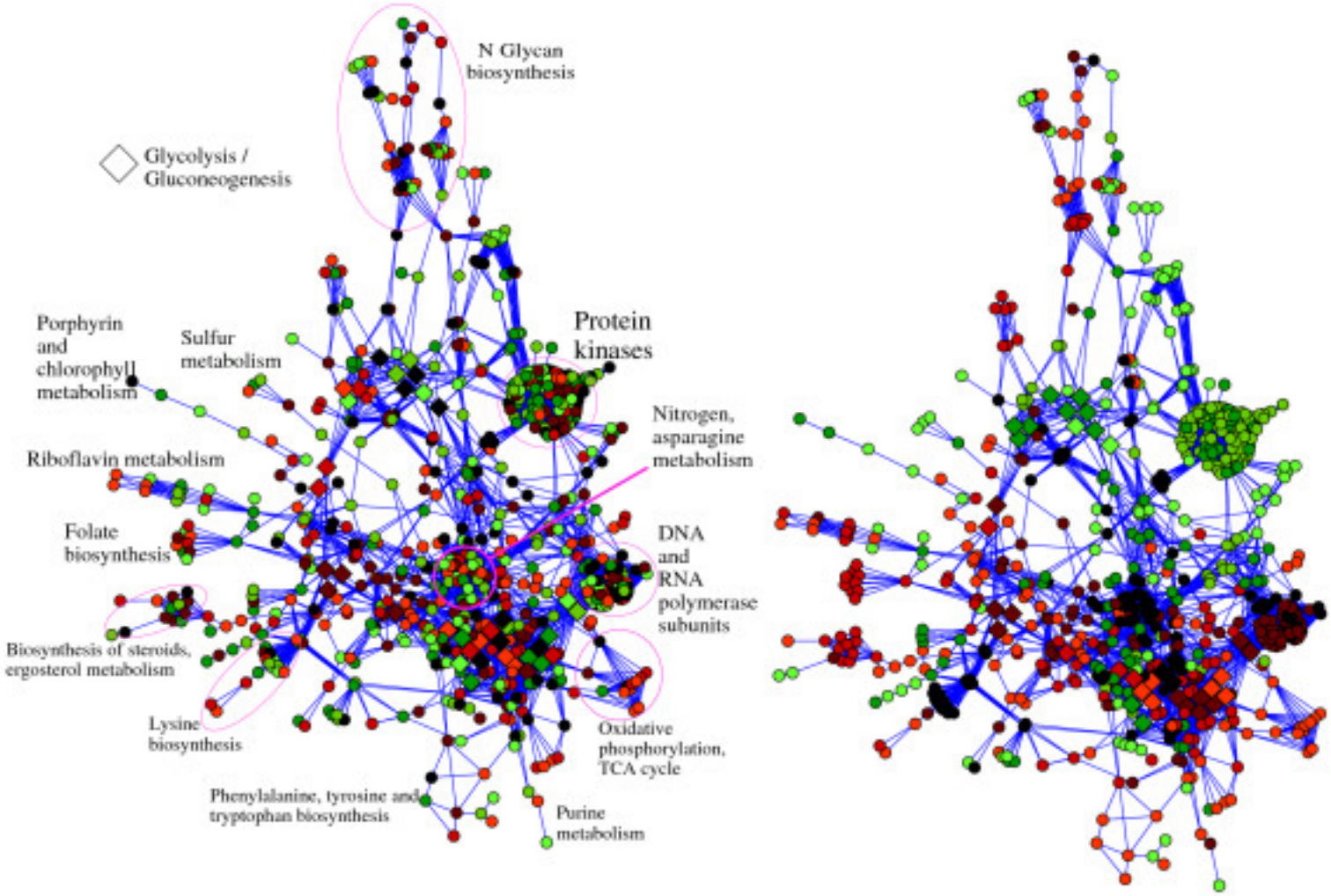
$\exists$ граф $G = (V, U)$,

V – множество вершин, соответствующих признакам $\{1, \ldots, p\}$

$U \subseteq V \times V$ – множество ребер

Граф генов и метаболизмы(процессы) дрожжей

из работы "Classification of microarray data using gene networks"[2007]



Формулировка задачи

$$\min_{w \in \mathbb{R}^p} \Big[\mathcal{L}(w) + \lambda \Omega(w) \Big]$$

$w \in \mathbb{R}^p$ – искомый разреженный вектор весов

$\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$ – выпуклая функция ошибок

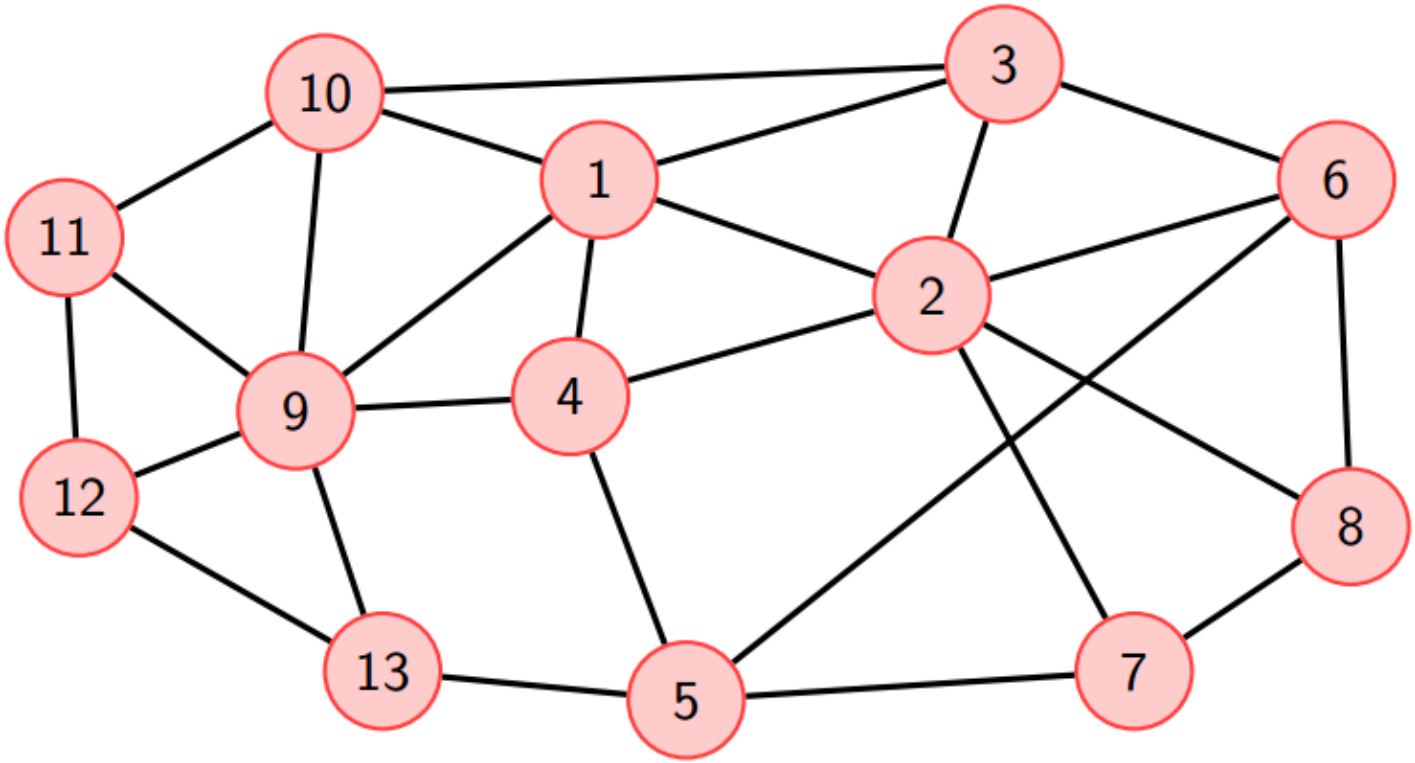
$\Omega : \mathbb{R}^p \to \mathbb{R}$ – функция регуляризации

Вопрос

Как учесть структуру графа в слагаемом регуляризации? Конкретнее, хотим, чтобы регуляризация поощряла признаки, образующие сильно связные подграфы.

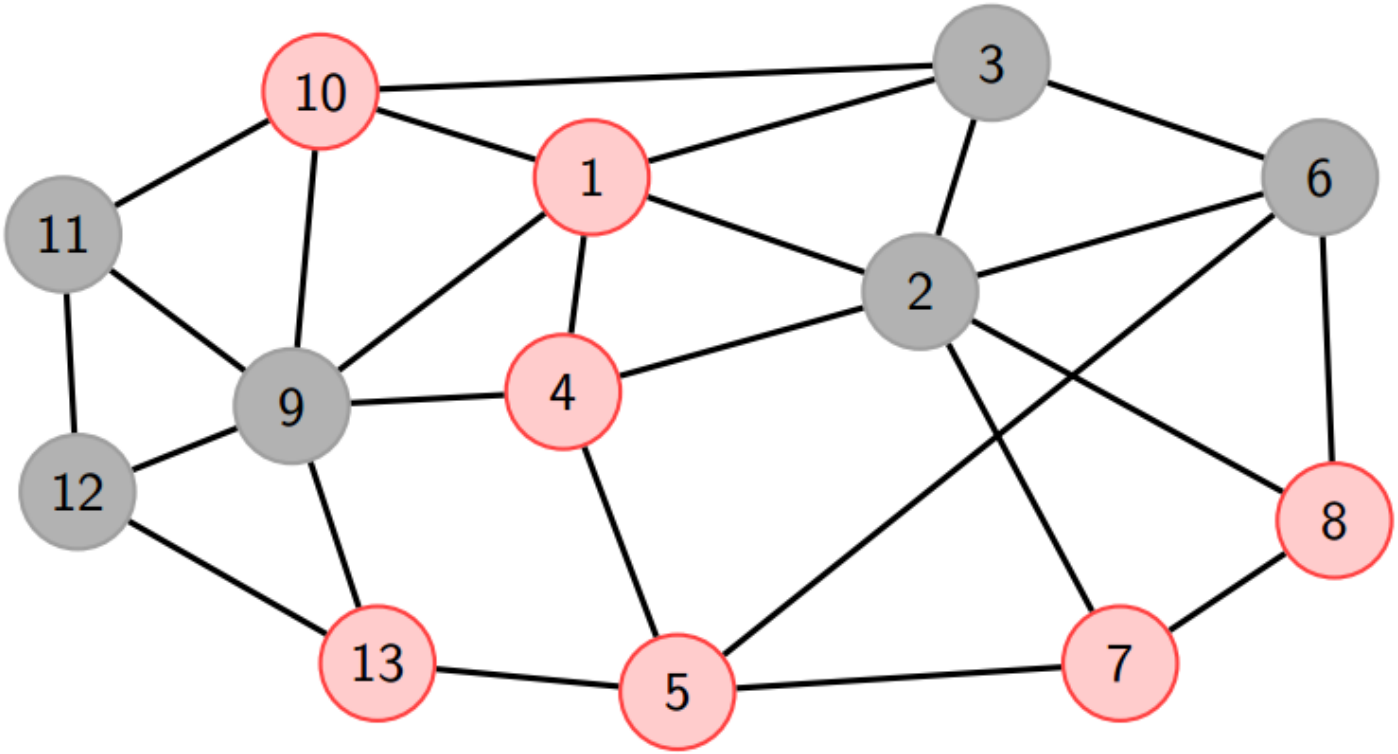
Граф признаков

$G = (V, U)$, $V = \{1, \ldots, p\}$



Граф признаков

Поощрение признаков, образующих подграф с небольшим количеством компонент связности



Структуированная разреженность в графе

невыпуклая регуляризация Huang, Zhang, and Metaxas [2009]

Определение

$$\varphi_{\mathcal{G}}(w) \triangleq \min_{\mathcal{J} \subseteq \mathcal{G}} \left\{ \sum_{g \in \mathcal{G}} \eta_g \mid \quad \mathrm{Supp}(w) \subseteq \bigcup_{g \in \mathcal{J}} g \right\}.$$

$\mathrm{Supp}(w) \triangleq \{j \in \{1, \ldots, p\} : w_j \neq 0\}$ - признаки с ненулевыми весами

$\mathcal{G}$ – предопределенное множество групп признаков

η_g – плата за выбор группы $g \in \mathcal{G}$

- штрафующая функция не выпуклая
- NP-полная задача
- структура ненулевых w – объединение нескольких групп из $\mathcal{G}$

Структуированная разреженность в графе

выпуклая релаксация и регуляризация Jacob, Obozinski, and Vert [2009]

$\varphi_{\mathcal{G}}$ может быть переписана в матричном виде:

$$\varphi_{\mathcal{G}}(\mathrm{w}) = \min_{\mathrm{x}\in\{0,1\}^{|\mathcal{G}|}} \left\{ \eta^{\mathrm{T}}\mathrm{x} \;\middle|\; \mathrm{N}\mathrm{x} \geq \mathrm{Supp}(\mathrm{w}) \right\}.$$

И ее выпуклая LP-релаксация:

$$\psi_{\mathcal{G}}(\mathrm{w}) \triangleq \min_{\mathrm{x}\in\mathbb{R}_{+}^{\mathcal{G}}} \left\{ \eta^{\mathrm{T}}\mathrm{x} \;\middle|\; \mathrm{N}\mathrm{x} \geq |\mathrm{w}| \right\}.$$

Структуированная разреженность в графе

Стуктуры групп

Очевидные примеры множеств групп $\mathcal{G}$, поощряющих связанность графов:

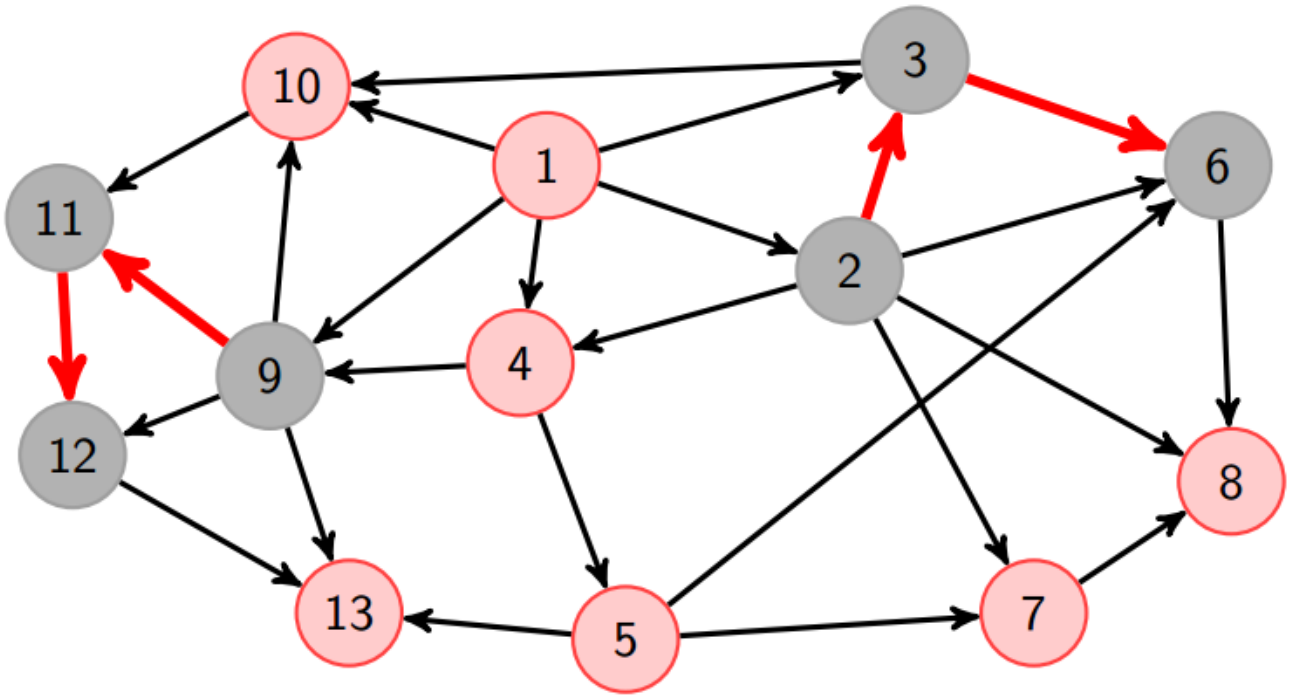
- всевозможные пары соседних вершин. не учитывает дальних соседей;
- всевозможные связанные подграфы размера не более L. трудоемкие вычисления;
- всевозможные связанные подграфы. трудоемкие вычисления.

Вопрос

Можем ли мы заменить связные подграфы другой структурой, которая (i) учитывала бы всех соседей (соседей соседей и т. д.) в графе и (ii) сводилась бы к быстро вычислимой функции регуляризации?

Решение в случае ориентированного ацикличного графа (DAG)

1. Определим $\mathcal{G}$ как множество всех путей в DAG.
2. Опреределим $\eta_{\mathrm{g}} = \gamma + |\mathrm{g}|$ (плата за выбор группы g).



$$\psi_{\mathcal{G}}(\mathrm{w}) = (\gamma + 3) + (\gamma + 3)$$

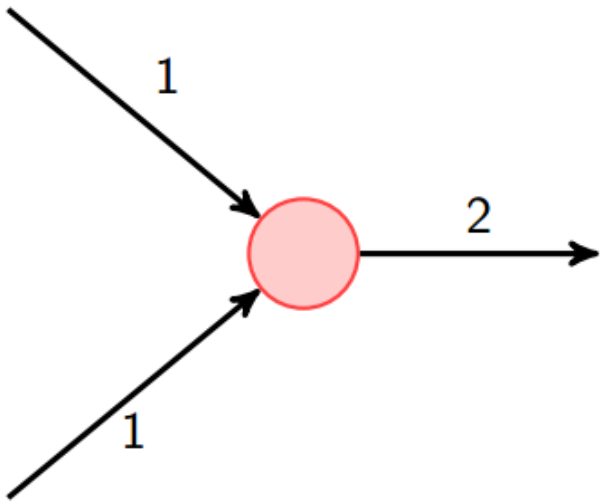
Кратко о транспортных сетях

Транспортная сеть (flow network) – ориентированный граф $G(V, E)$, в котором

- каждому ребру $(u, v) \in E$ приписана неотрицательная пропускная способность $\delta(u, v) \geq 0$
- выделены две вершины: источник (source) s и сток (sink) t

Поток $f : V \times V \to \mathbb{R}$ – неотрицательная функция, определенная на ребрах:

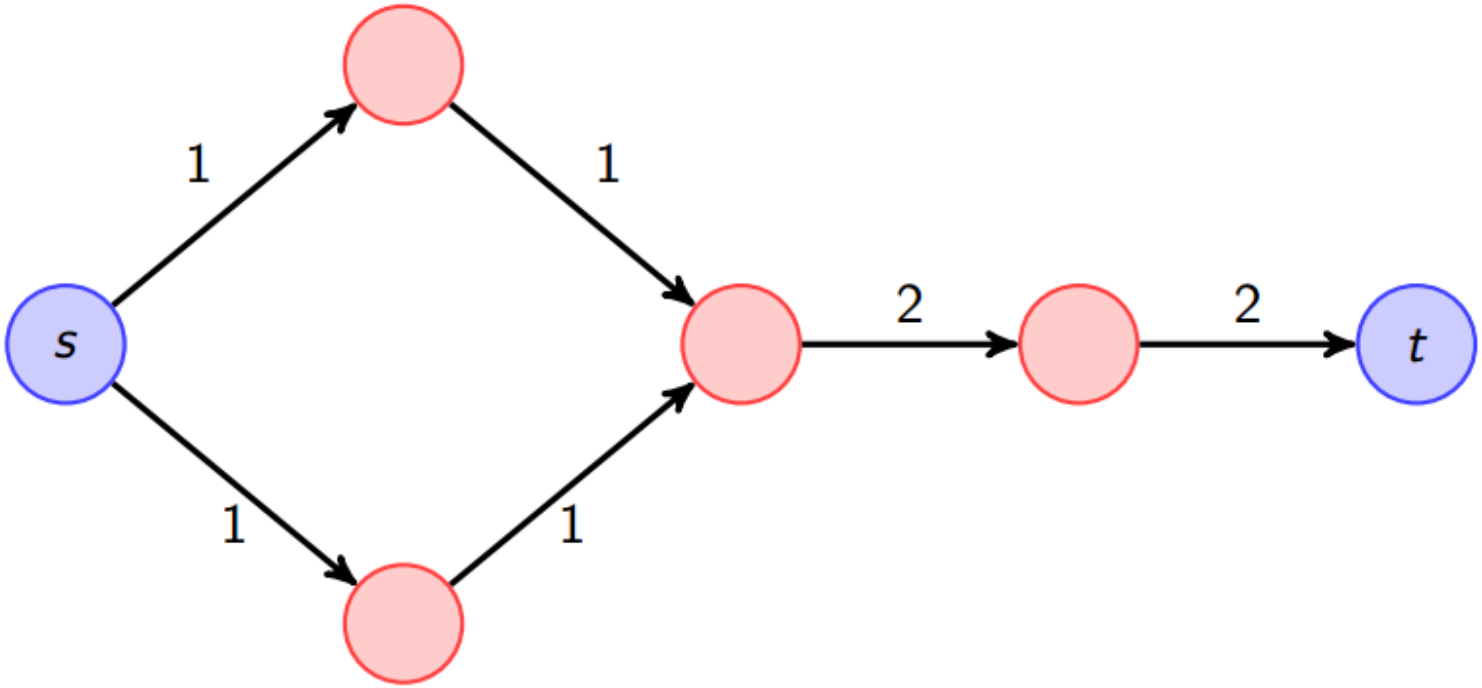
- ограничения по пропускной способности $f_{uv} \leq \delta_{uv}$
- сохранение потока $\sum\limits_{(u,v)\in E} = \sum\limits_{(v,z)\in E}$ для $\forall v \in V \backslash \{s, t\}$



Кратко о транспортных сетях

Свойства

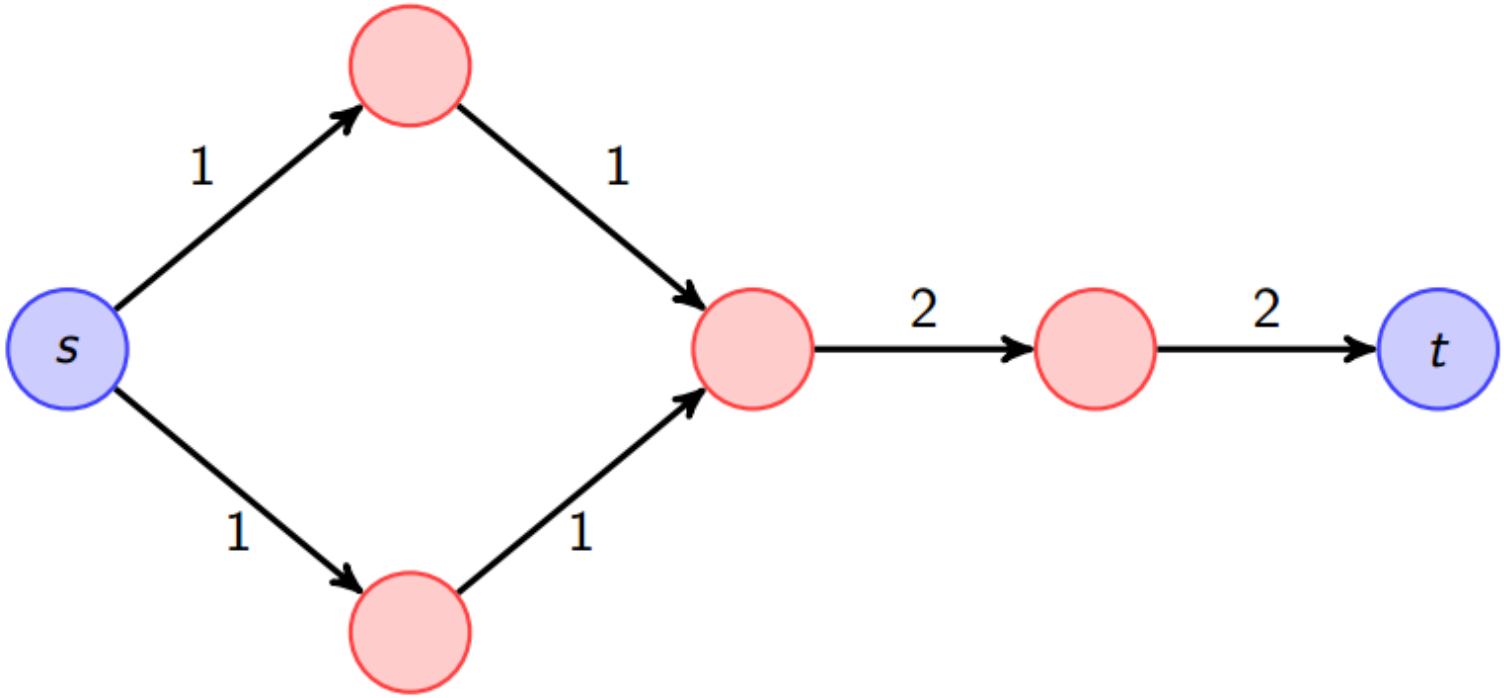
Поток идет от источника s к стоку t.



Кратко о транспортных сетях

Свойства

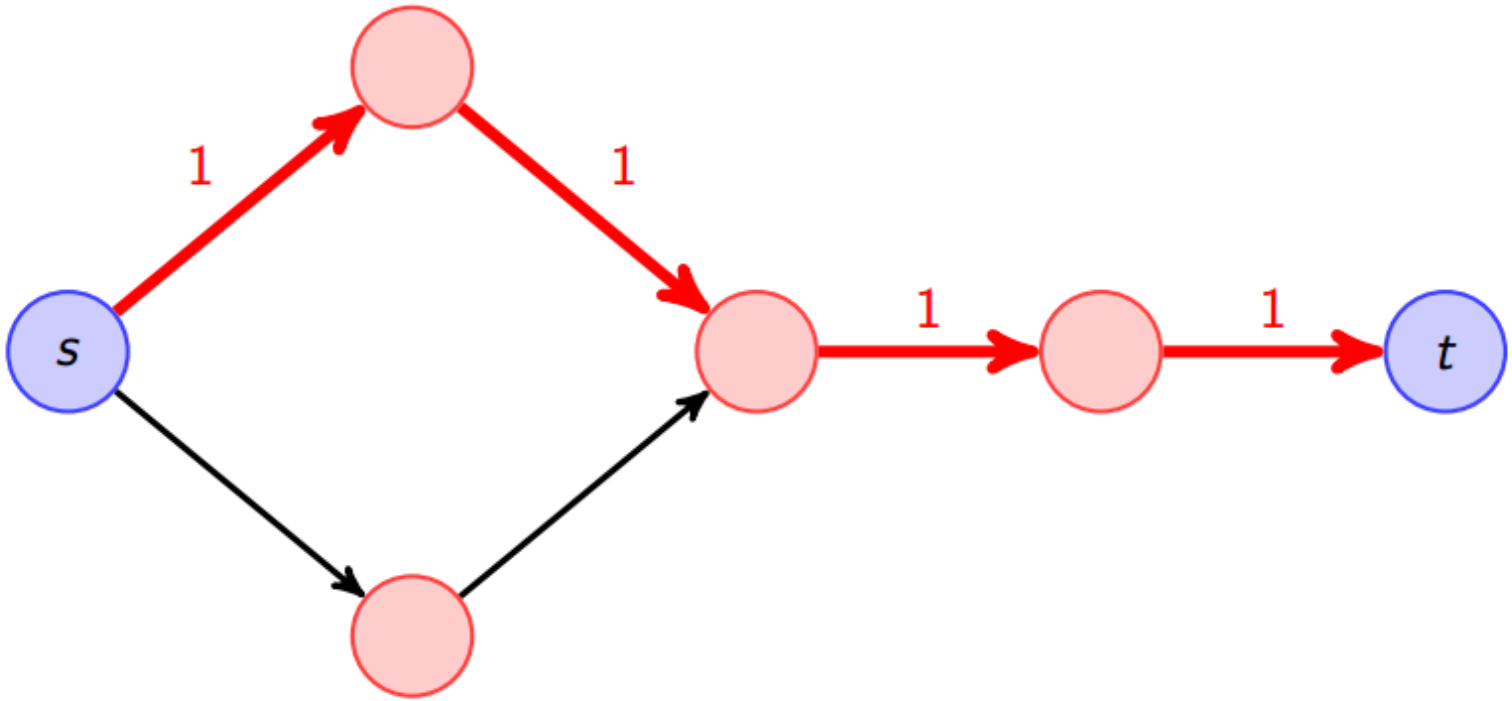
Любой поток представим в виде суммы "потоков-путей".



Кратко о транспортных сетях

Свойства

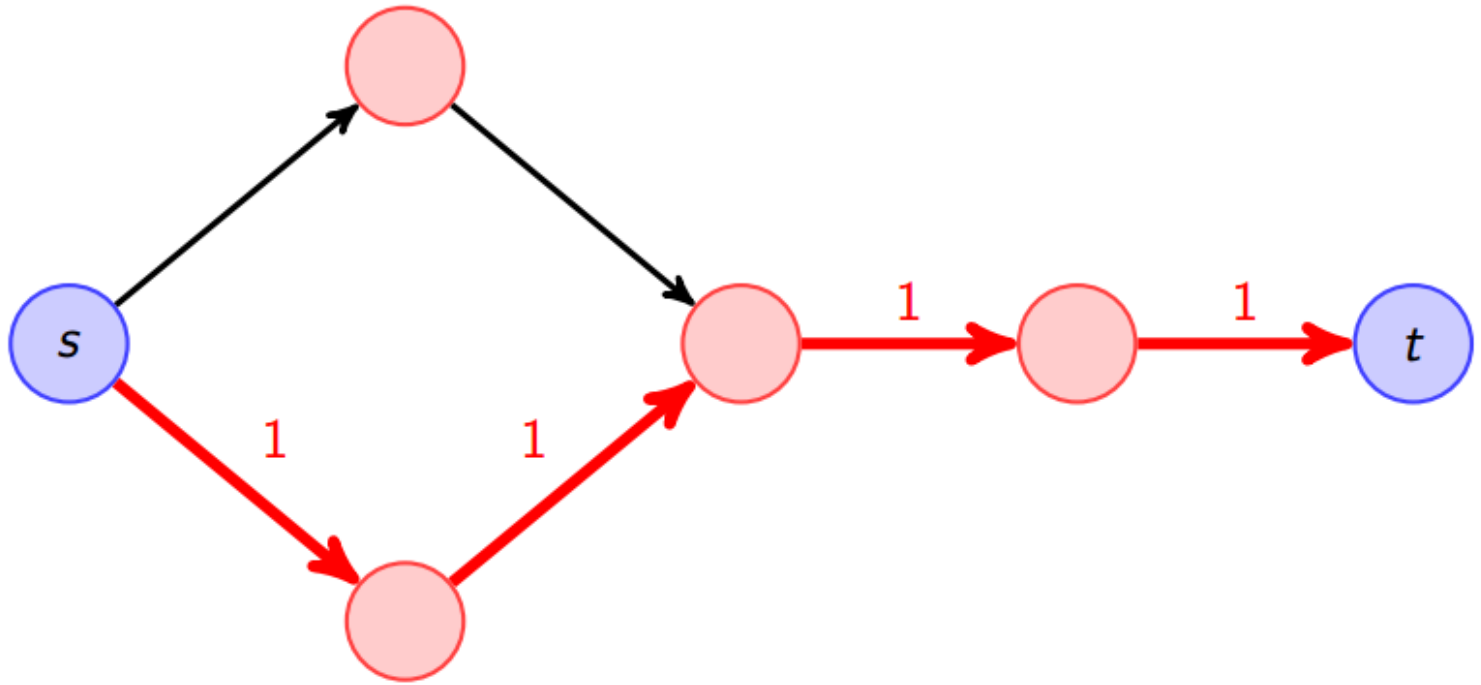
Любой поток представим в виде суммы "потоков-путей".



Кратко о транспортных сетях

Свойства

Любой поток представим в виде суммы "потоков-путей".



Кратко о транспортных сетях

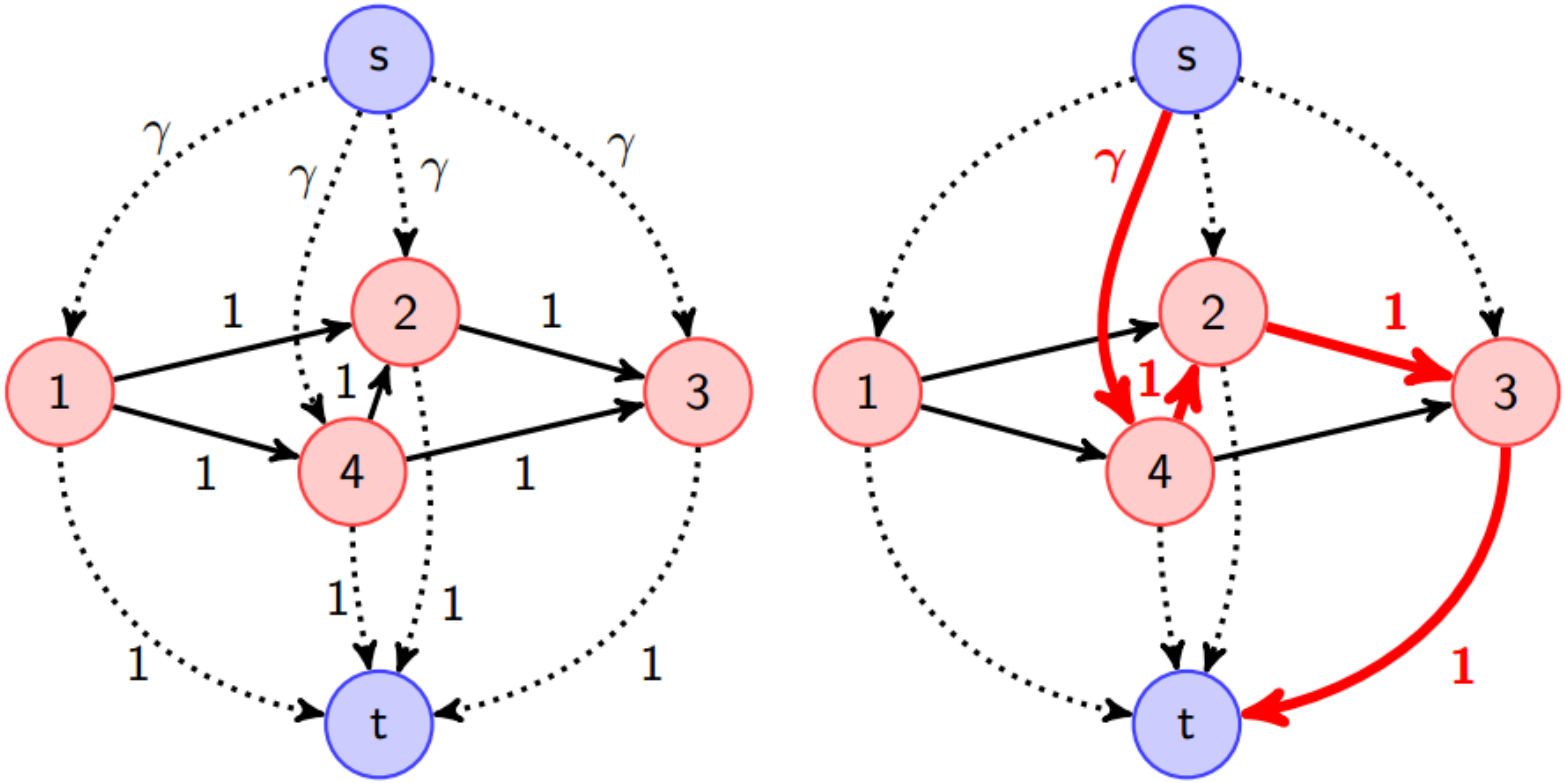
Задача о потоке минимальной стоимости

$$\min_{f} \sum_{(u,v)\in E} c_{uv} f_{uv}$$

- $c_{uv} \in \mathbb{R}$ – стоимость каждого ребра $(u, v) \in E$
- находит самый дешевый способ передачи определённого количества потока через транспортную сеть
- существуют быстрые алгоритмы решения

Решение для DAG

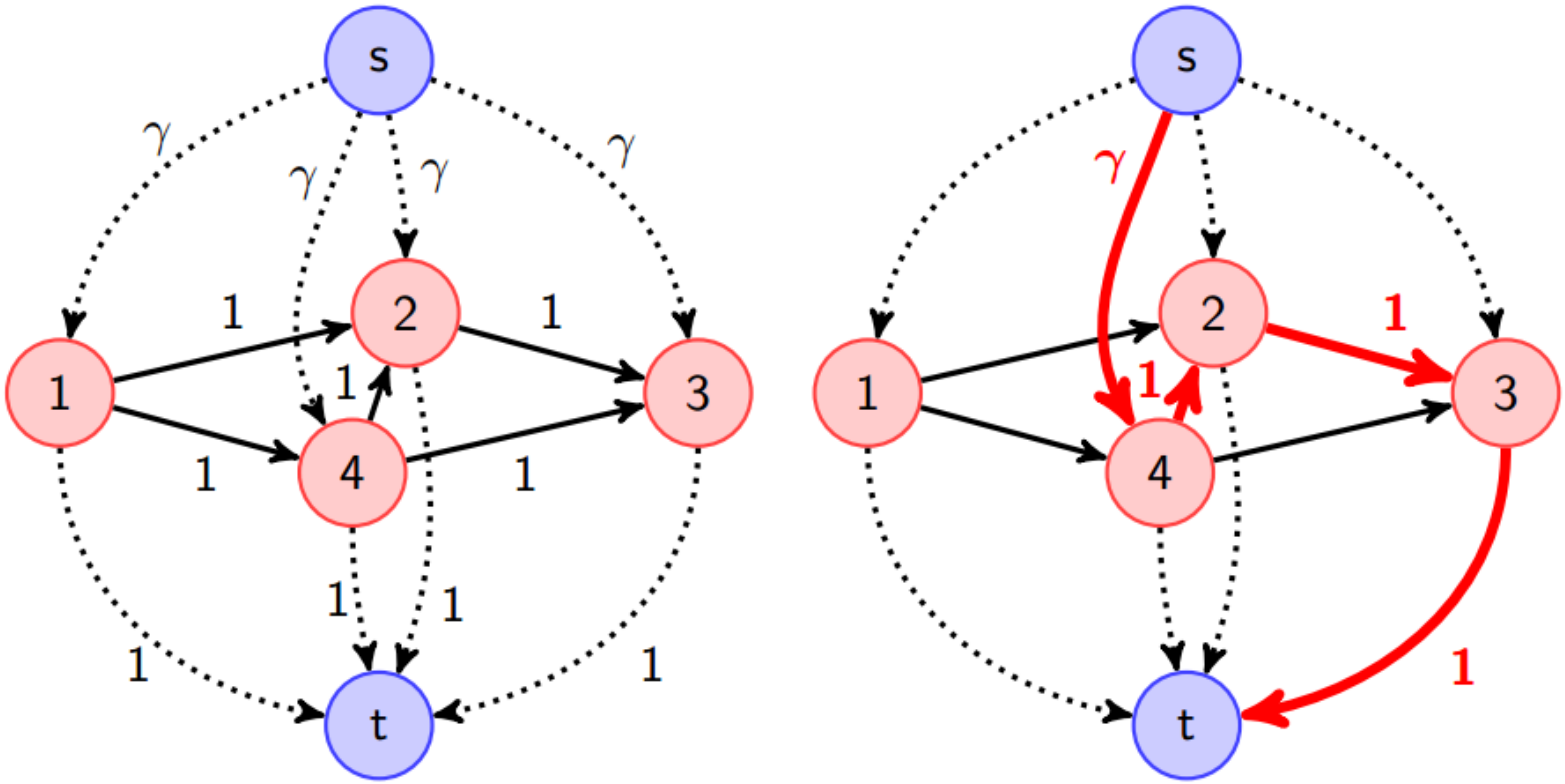
$V' \triangleq V \cup \{s, t\}$,
$E' \triangleq E \cup \{(s, v) : v \in V\} \cup \{(u, t) : u \in V\}$
$\eta_g \triangleq \gamma + |g|$



Решение для DAG

Обобщим веса η_g, используя стоимости ребер:

$$\eta_g \triangleq c_{su_1} + \left(\sum_{i=1}^{k-1} c_{u_i u_{i+1}}\right) + c_{u_k t} = \sum_{(u,v)\in(s,g,t)} c_{uv}$$



Сведение к задаче транспортных потоков

Задача оптимизации может быть трансформирована в эквивалентную задачу о потоке минимальной стоимости:

$$\varphi_{\mathcal{G}}(w) = \min_{f \in \mathcal{F}} \sum_{(u,v) \in E'} f_{uv} c_{uv} \quad | \quad s_j(f) \geq 1, \forall j \in \mathrm{Supp}(w),$$

$$s_j(f) \triangleq \sum_{u \in V' : (u,v) \in E'} f_{uj} \text{ – поток, входящий в j-ую вершину}$$

И выпуклый вариант задачи оптимизации:

$$\psi_{\mathcal{G}}(w) = \min_{f \in \mathcal{F}} \sum_{(u,v) \in E'} f_{uv} c_{uv} \quad | \quad s_j(f) \geq |w|, \forall j \in \{1, \ldots, p\}$$

Важно

$\varphi_{\mathcal{G}}$, $\psi_{\mathcal{G}}$ могут быть посчитаны за полиномиальное время с помощью транспортных сетей.

Задача классификации рака груди

Описание данных:

- цепочки генов с p = 7910 генами.
- n = 295 опухолей, 78 злокачественных, 217 доброкачественных.
- граф генов был сгенерирован Chuang et al. [2007].

Для преобразования графа в DAG, направления ребер выбирались случайно, циклы удалялись.
20% данных для теста, 80% для 10-fold кросс-валидации.

Задача классификации рака груди

Результаты

	Ridge	Lasso	Elastic-Net	Группы-пары	ψ(вып)
error in %	31.0	36.0	31.5	35.9	30.2
error std.	6.1	6.5	6.7	6.8	6.8
nnz	7910	32.6	929	68.4	69.9
connect	58	30.9	355	13.1	1.3
stab	100	7.9	30.9	6.1	32

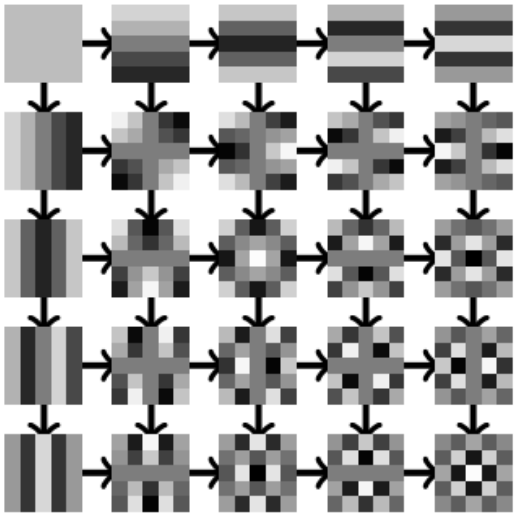
stab – процент генов, повторяющихся в разных запусках.

чем больше stab, тем стабильнее алгоритм.

Задача шумоподавления изображения

- Режем изображение на 10×10 перекрывающихся частей
- Апроксимируем каждую часть разреженной матрицей
- Усредняем перекрывающиеся пиксели для чистого изображения

Матрица признаков – разложение по косинусам разных частот (DCT dictionary):



Задача классификации рака груди

Результаты

- 7 видов шума
- параметры оптимизируются на первых трех изображениях

σ	5	10	15	20	25	50	100
l_0	37.04	33.15	31.03	29.59	28.48	25.26	22.44
l_1	36.42	32.28	30.06	28.59	27.51	24.48	21.96
$\varphi_{\mathcal{G}}$	37.01	33.22	31.21	29.82	28.77	25.73	22.97
$\psi_{\mathcal{G}}$	36.32	32.17	29.99	28.54	27.49	24.54	22.12

Привет!