ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №5

Элементы теории графов

Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

Сайт курса http://intbio.org/bioinf2019-2020

Теория Графов

Alexander Lazarev

Institute of Control Sciences of Russian Academy of Sciences

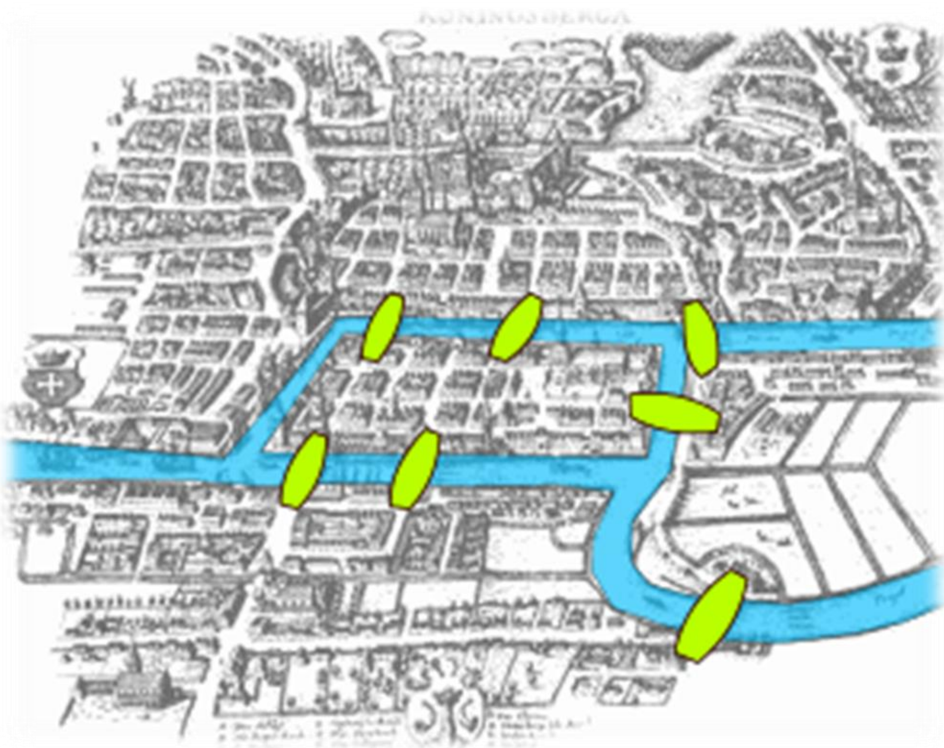
2009-2010 учебный год

Задача о кёнигсбергских мостах

(лат. *Problema Regiomontanum de septem pontibus*)

Можно ли пройти по всем мостам Кёнигсберга, не проходя ни по одному из них дважды?

Была решена Эйлером в 1736 году и послужила основой для создания теории графов.



Леонард Эйлер
(1707 – 1783)

При чём тут биология?

Best matches for graph theory:

Graph Theory and Brain Connectivity in Alzheimer's Disease.
delEtoile J et al. Neuroscientist. (2017)

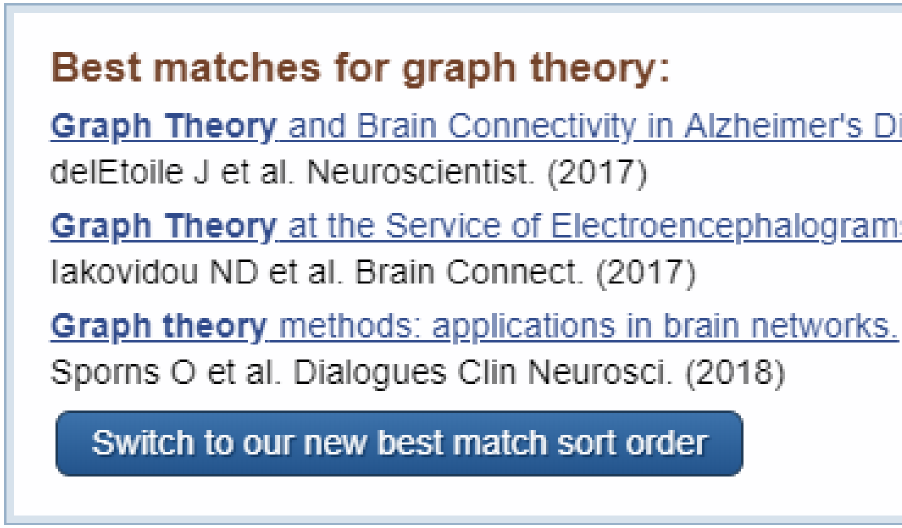
Graph Theory at the Service of Electroencephalograms.
Iakovidou ND et al. Brain Connect. (2017)

Graph theory methods: applications in brain networks.
Sporns O et al. Dialogues Clin Neurosci. (2018)

Switch to our new best match sort order

Results by year

При чём тут биология?

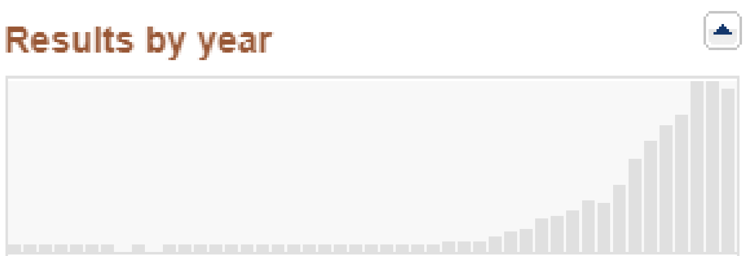

Best matches for graph theory:

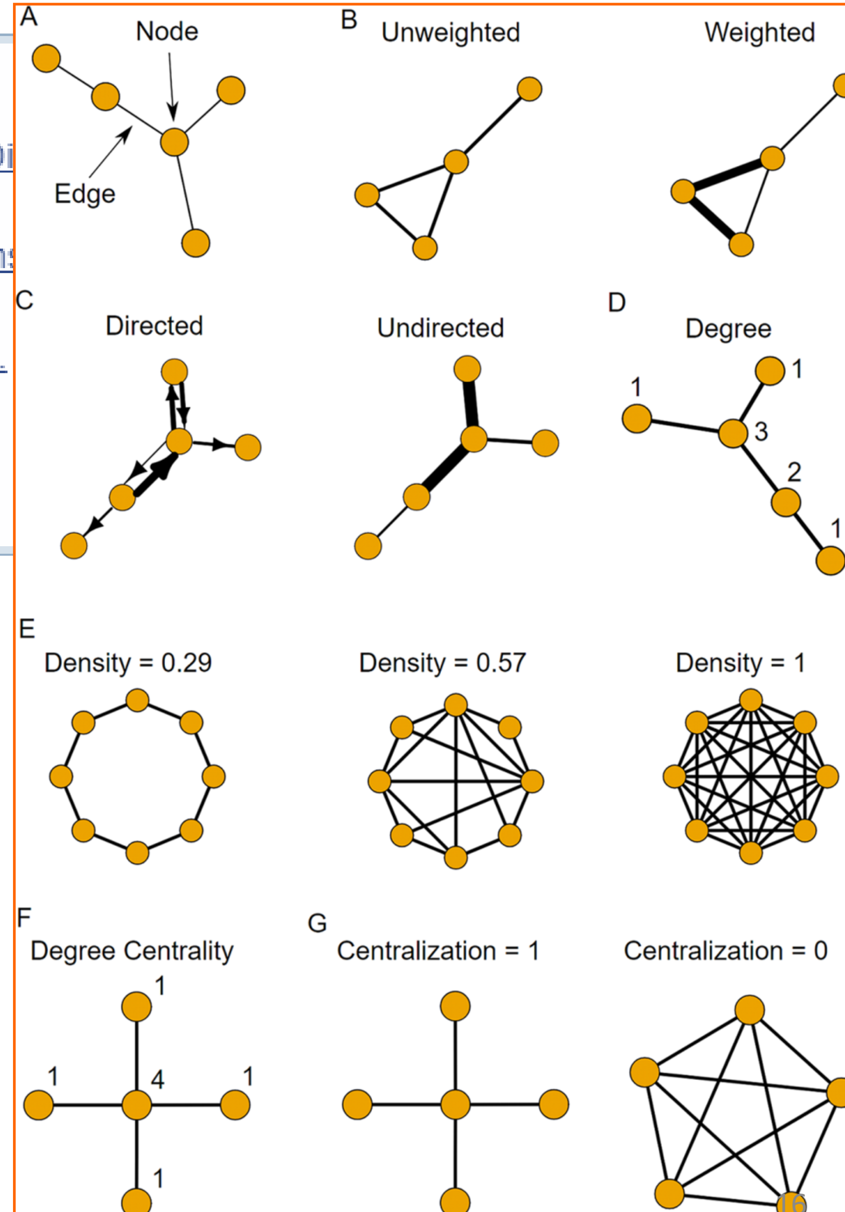
Graph Theory and Brain Connectivity in Alzheimer's Di
delEtoile J et al. Neuroscientist. (2017)

Graph Theory at the Service of Electroencephalogram
Iakovidou ND et al. Brain Connect. (2017)

Graph theory methods: applications in brain networks.
Sporns O et al. Dialogues Clin Neurosci. (2018)

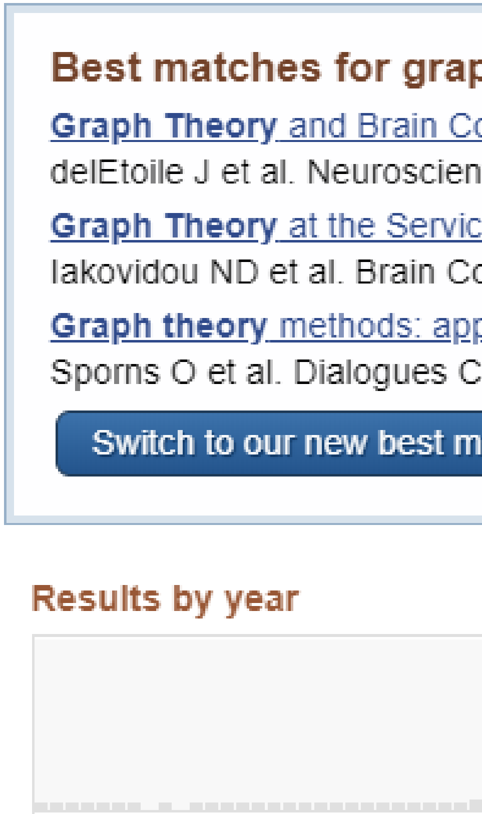
Switch to our new best match sort order

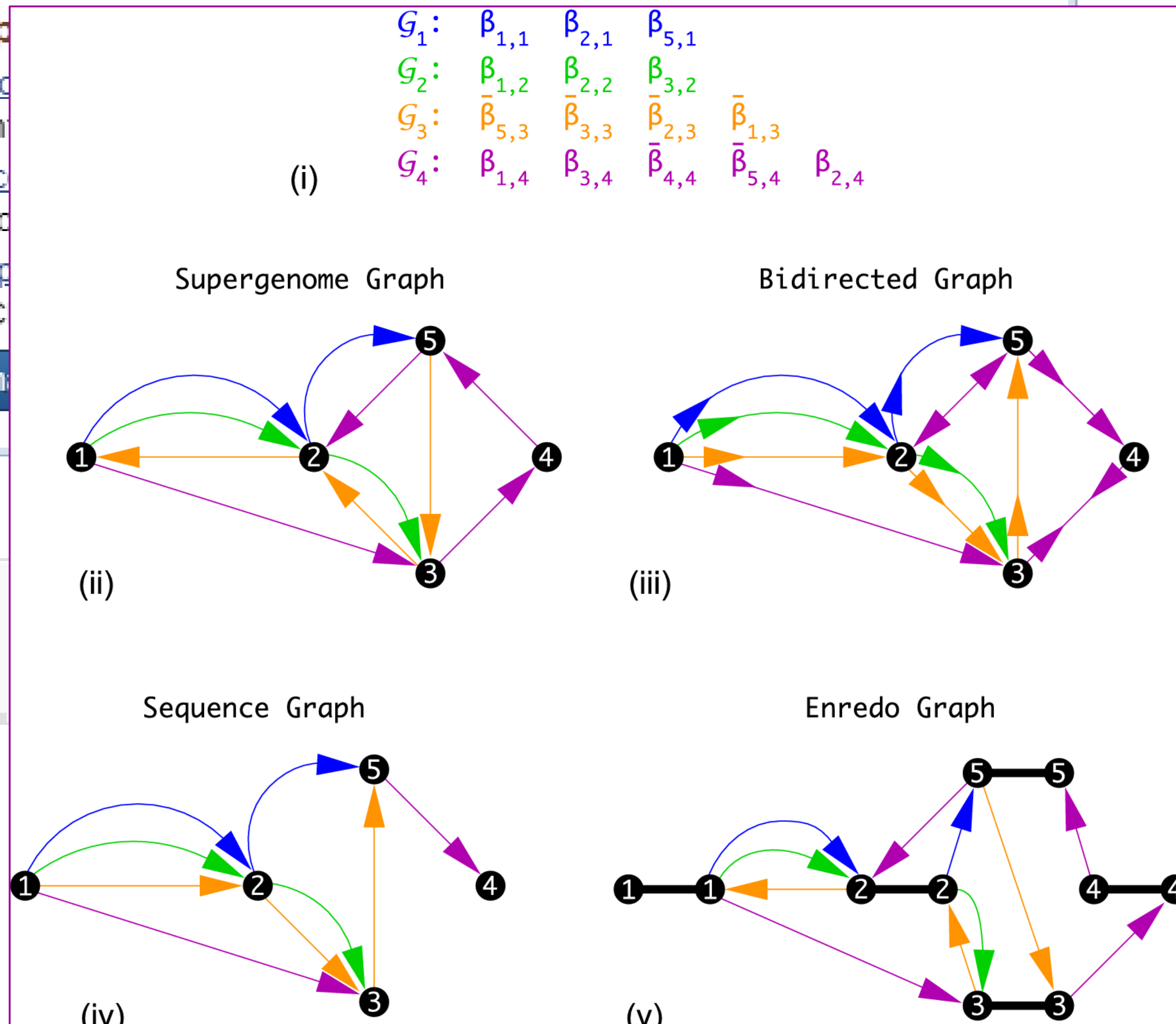




Applying Graph Theory to Examine the Dynamics of Student Discussions in Small Group Learning (2019)

При чём тут биология?

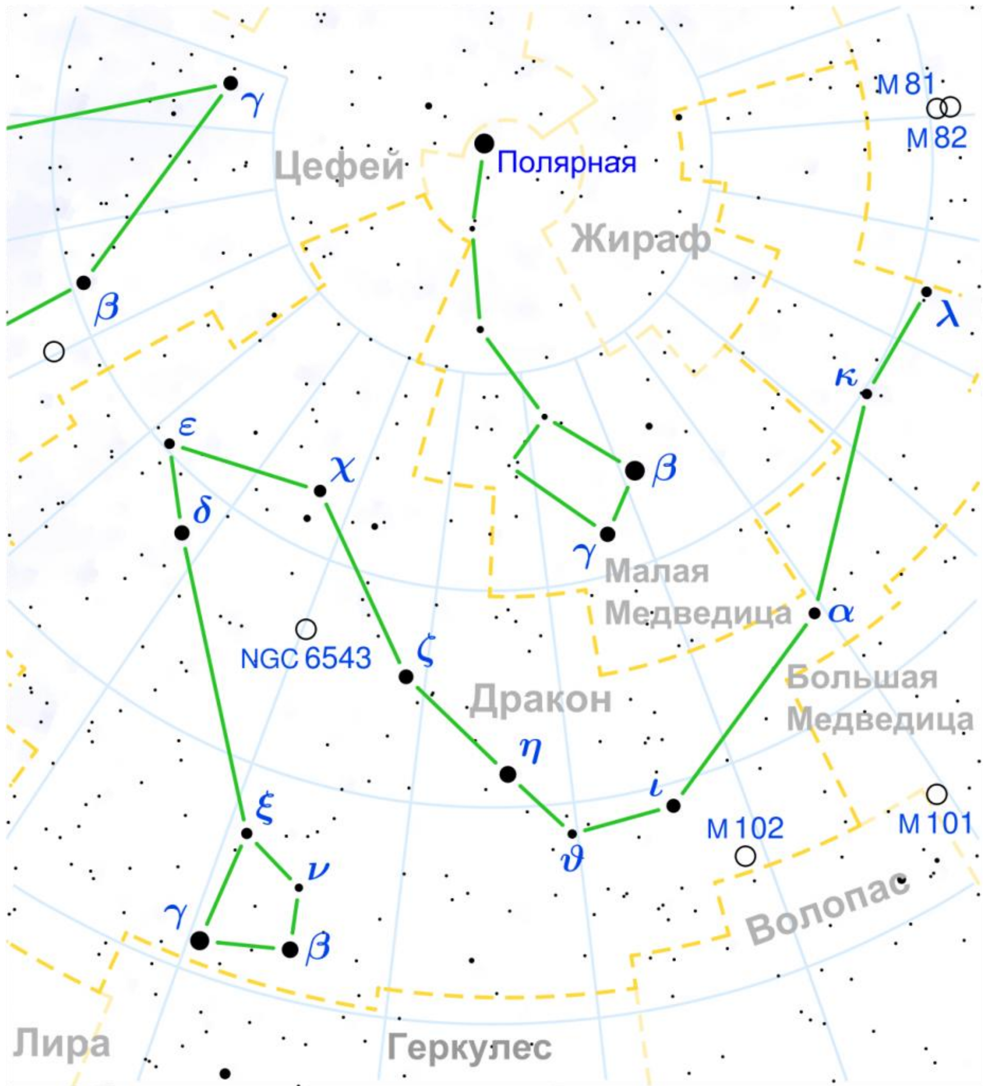




Coordinate systems for supergenomes (2018)

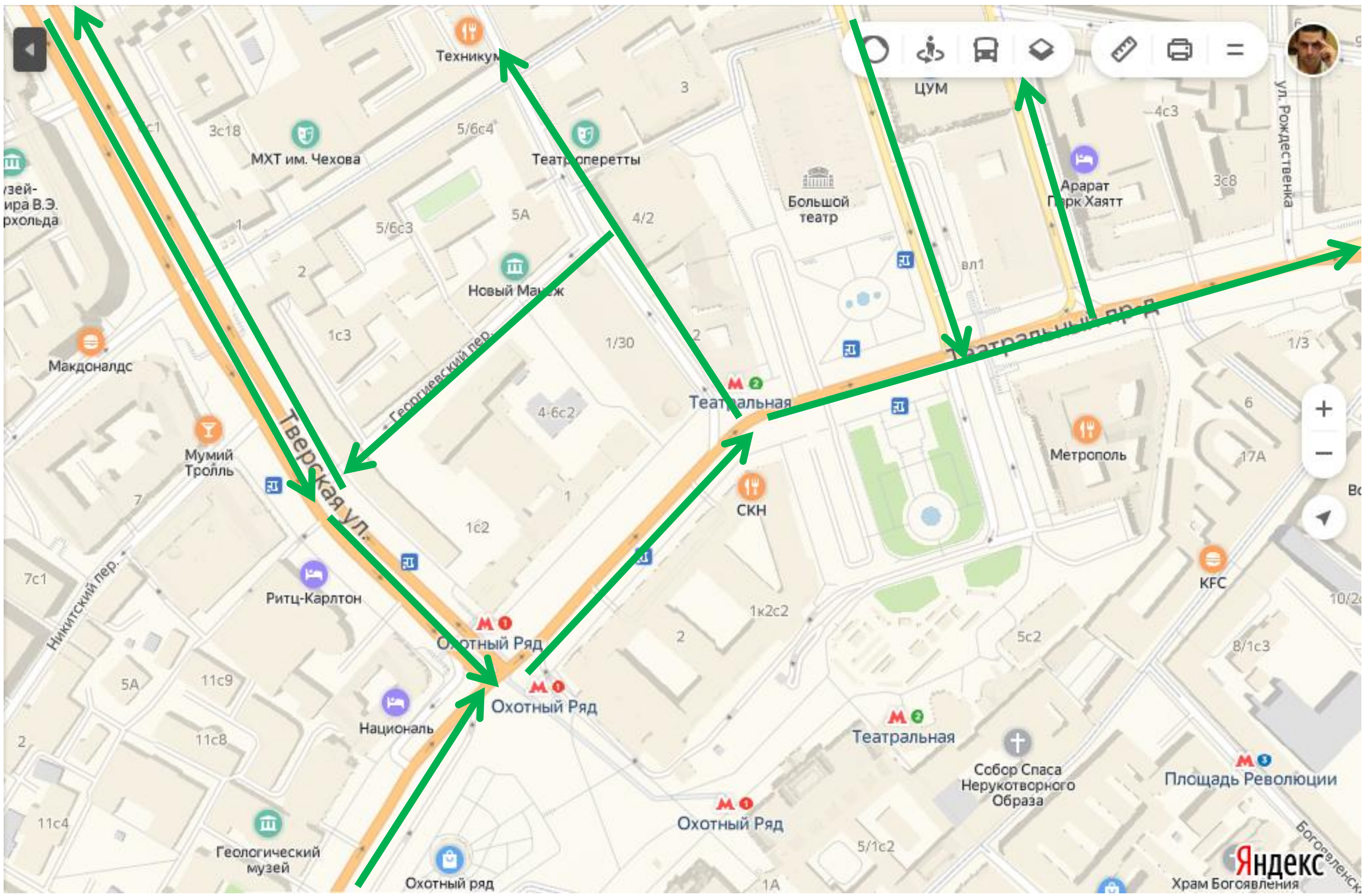
Определения

Граф *G = <V, E>* **есть совокупность множества вершин** *V* **и множества рёбер** (дуг) *E*.



Определения

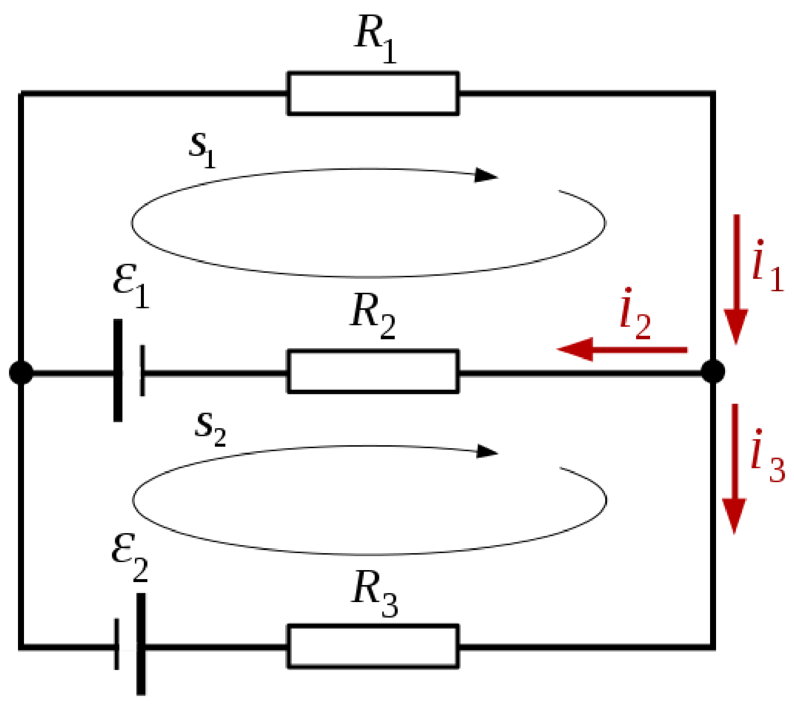
Граф называется **неориентированным** (неограф), если все его ребра {x; y} неориентированы, и **ориентированным** (орграф), если все его ребра <x; y> ориентированы. В случае произвольного графа ребра (x; y).



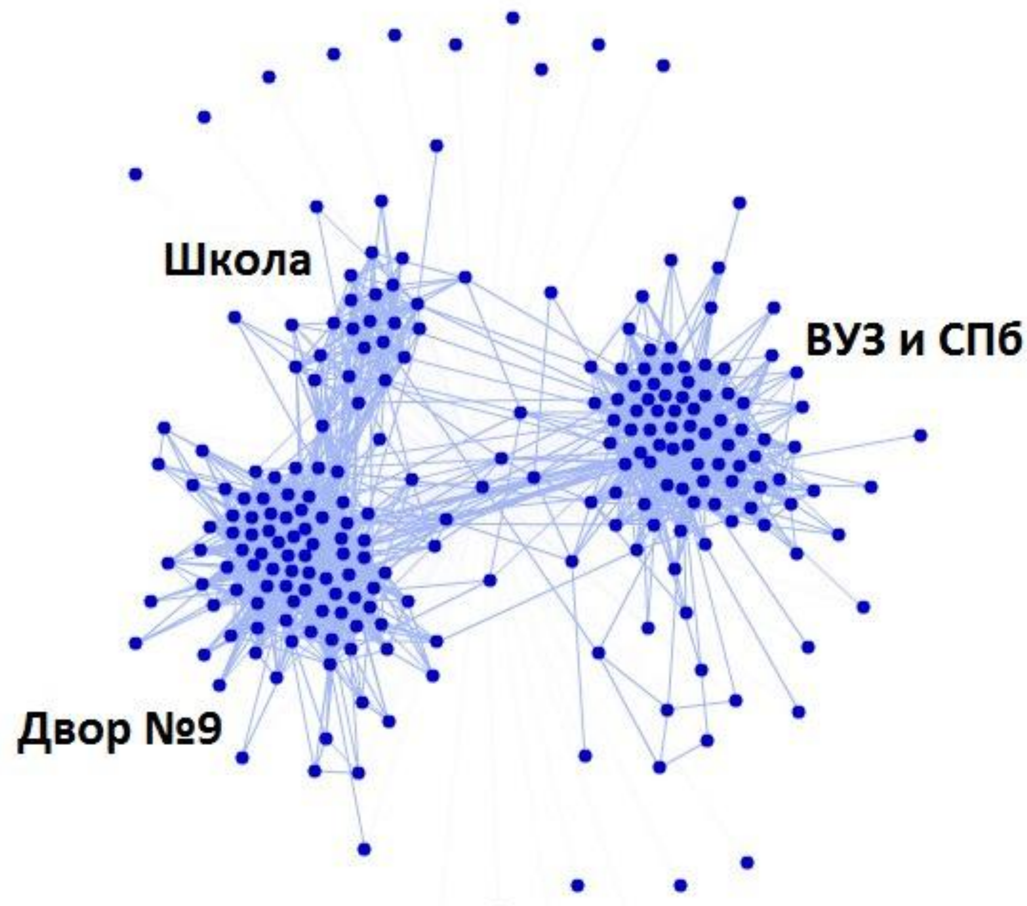
Правила Кирхгофа

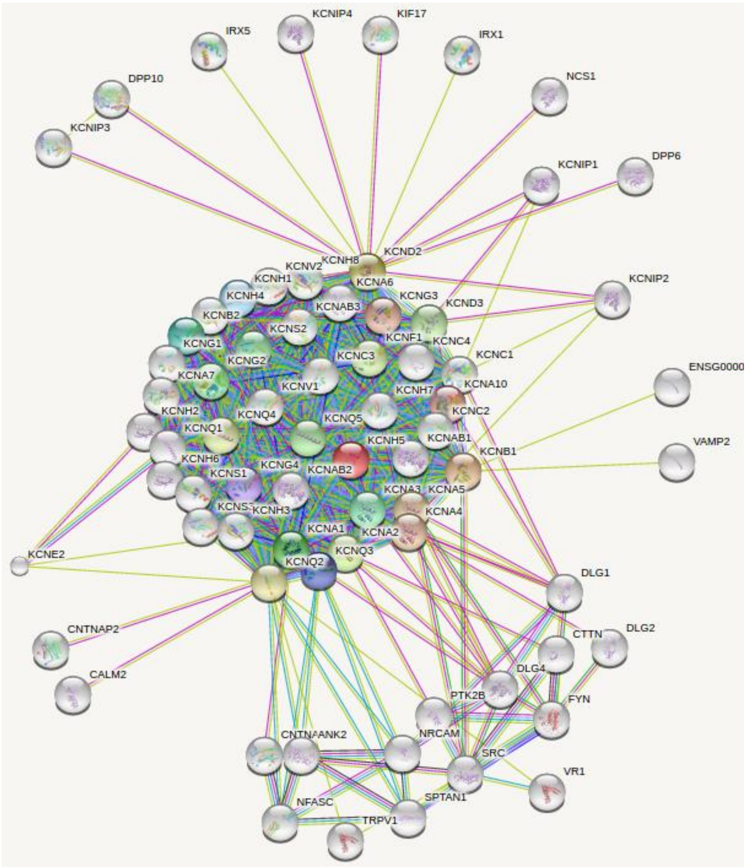
1) алгебраическая сумма токов ветвей, сходящихся в каждом узле любой цепи, равна нулю.

2) алгебраическая сумма напряжений на резистивных элементах замкнутого контура равна алгебраической сумме ЭДС, входящих в этот контур

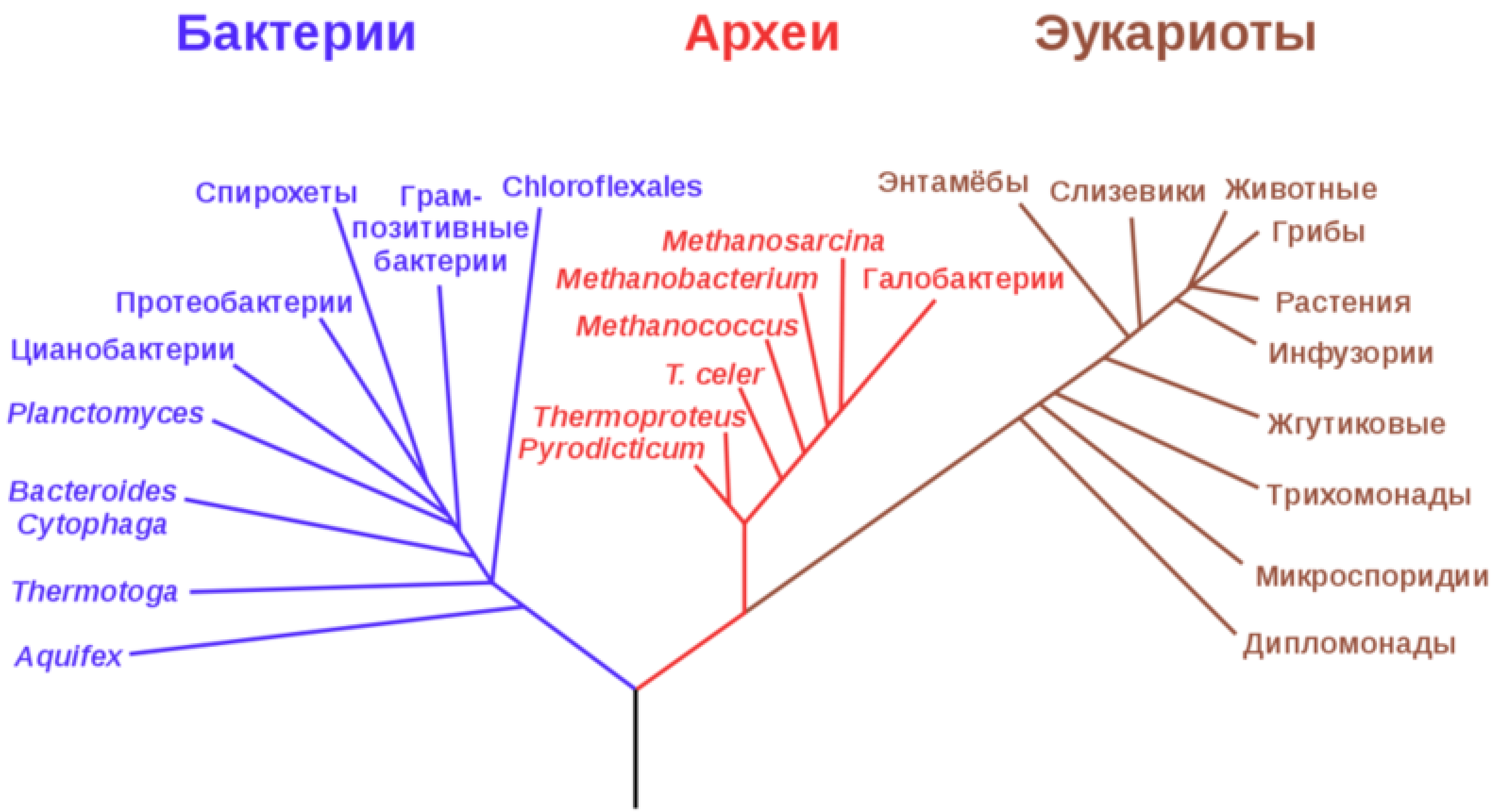


Примеры графов





Примеры графов



Определения

Ребро e **инцидентно** вершинам x и y, а вершины x и y инцидентны ребру e. Вершина, не инцидентная никакому ребру, называется **изолированной**.

Вершины x и y **смежны**, если (x, y) является ребром. Два ребра смежны, если имеют общую вершину.

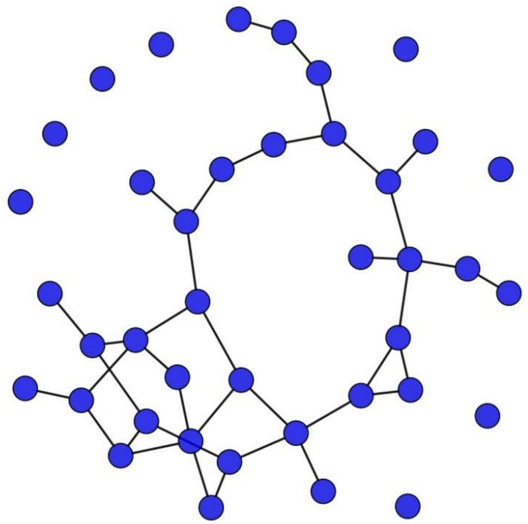
Простой граф – граф без петель и кратных рёбер.

Связный граф — граф, содержащий ровно одну компоненту связности. Это означает, что между любой парой вершин этого графа существует как минимум один путь.

Случайный граф – совокупность *n* изолированных вершин и случайного количества соединяющих их рёбер.

Впервые рассмотрены Эрдёшем и Реньи, и, независимо, Гилбертом в 1959 году.

ER(40,0.05)



Степени вершин

Пусть G – неориентированный граф. Число *ρ(x)* ребер, инцидентных вершине x, называется **степенью вершины**. Отдельно следует оговаривать, считать ли петли однократными или двойными.

Пусть *ρ(x, y) = ρ(y, x)* – число ребер, соединяющих вершины x и y. Тогда для *ρ(x)* верно равенство:

$$\rho(x) = \sum_{y \in V} \rho(x, y)$$

Обозначим через *ne(G)* число ребер в этом графе. Поскольку каждое ребро при подсчете общего числа степеней учитывается дважды (при вершине x и при вершине y), то

$$2ne(G) = \sum_{x \in V} \rho(x) = \sum_{x,y \in V} \rho(x, y)$$

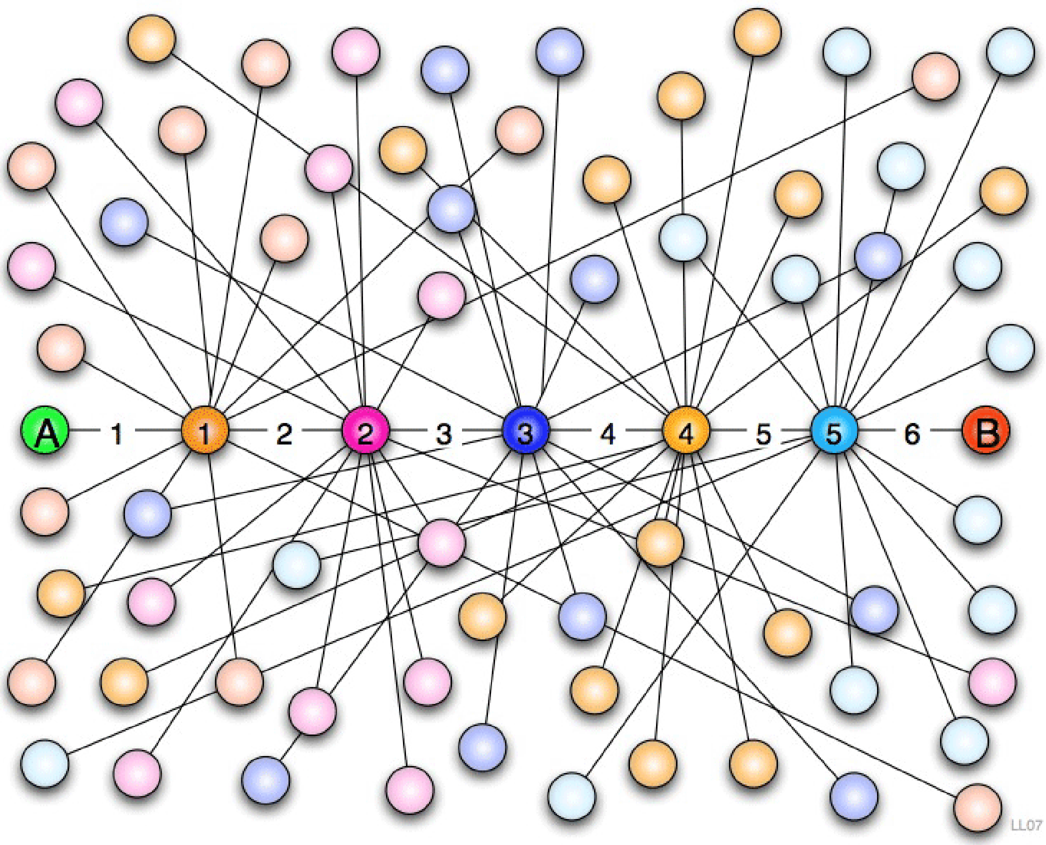
Формула справедлива и при наличии петель, если в степенях вершин учитывать их дважды.

Теорема 1 («Лемма о рукопожатиях»):

В конечном графе число вершин нечетной степени чётно.

Мир тесен

«Теория шести рукопожатий» - недоказанная теория, согласно которой любые два человека на Земле разделены не более чем пятью уровнями общих знакомых (и, соответственно, шестью уровнями связей).



«Звенья цепи» (Фридьеш Каринти, Венгрия, 1929)?

Мир тесен

Эксперимент «Мир тесен»

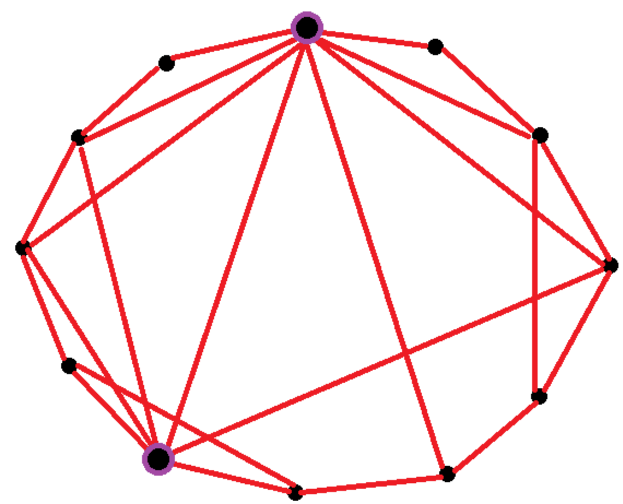
(Стэнли Милгрэм, США, 1967):

296 писем отправлено,

64 дошло до адресатов.

Средняя длина цепочки:

6 пересылок.



Граф «Мир тесен»: две произвольно взятые вершины с большой вероятностью не смежны, но могут быть соединены небольшим числом рёбер.

Мир тесен

Число Эрдёша – длина кратчайшего пути соавторства по совместным научным публикациям от какого-либо учёного до венгерского математика Пала Эрдёша.

EN (Эрдёш) = 0

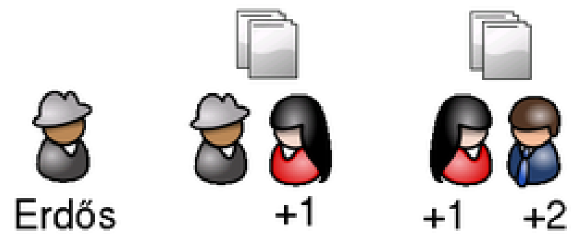
EN (соавторы Эрдёша) = 1

EN (Эйнштейн) = 2

EN (Фейнман) = 3

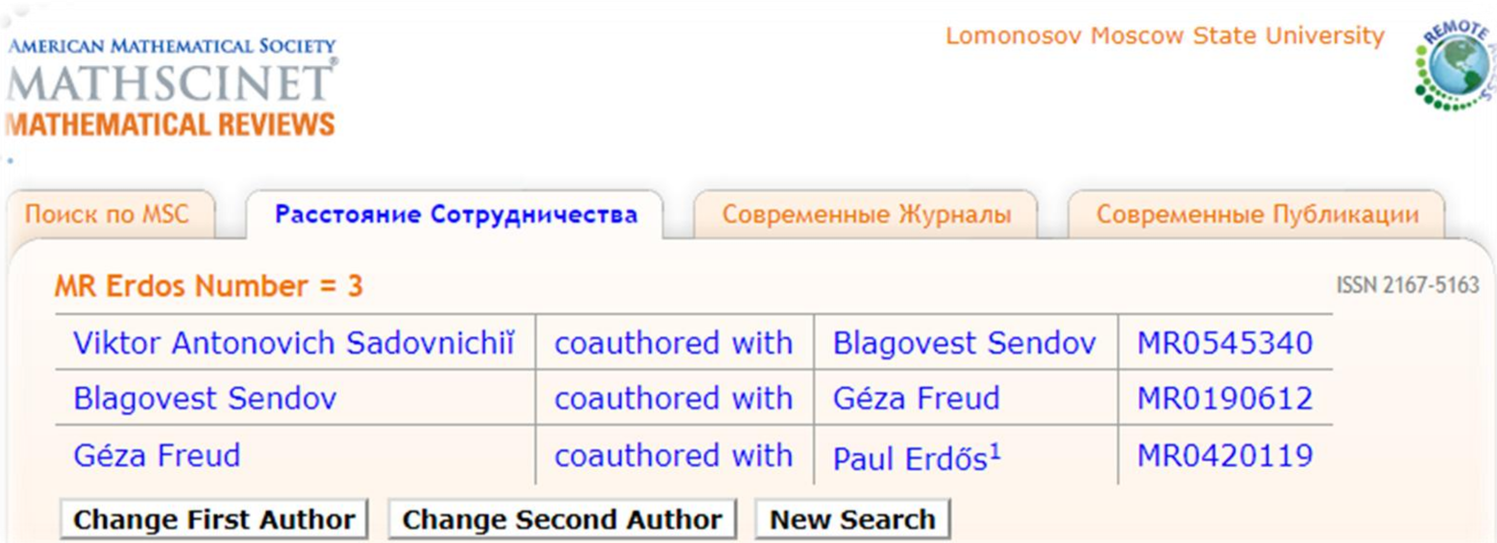
…

EN (Лаплас) = 14



Эрдёш Пал

(1913 – 1996)



AMERICAN MATHEMATICAL SOCIETY
MATHSCINET
MATHEMATICAL REVIEWS

Lomonosov Moscow State University

Поиск по MSC | Расстояние Сотрудничества | Современные Журналы | Современные Публикации

MR Erdos Number = 3

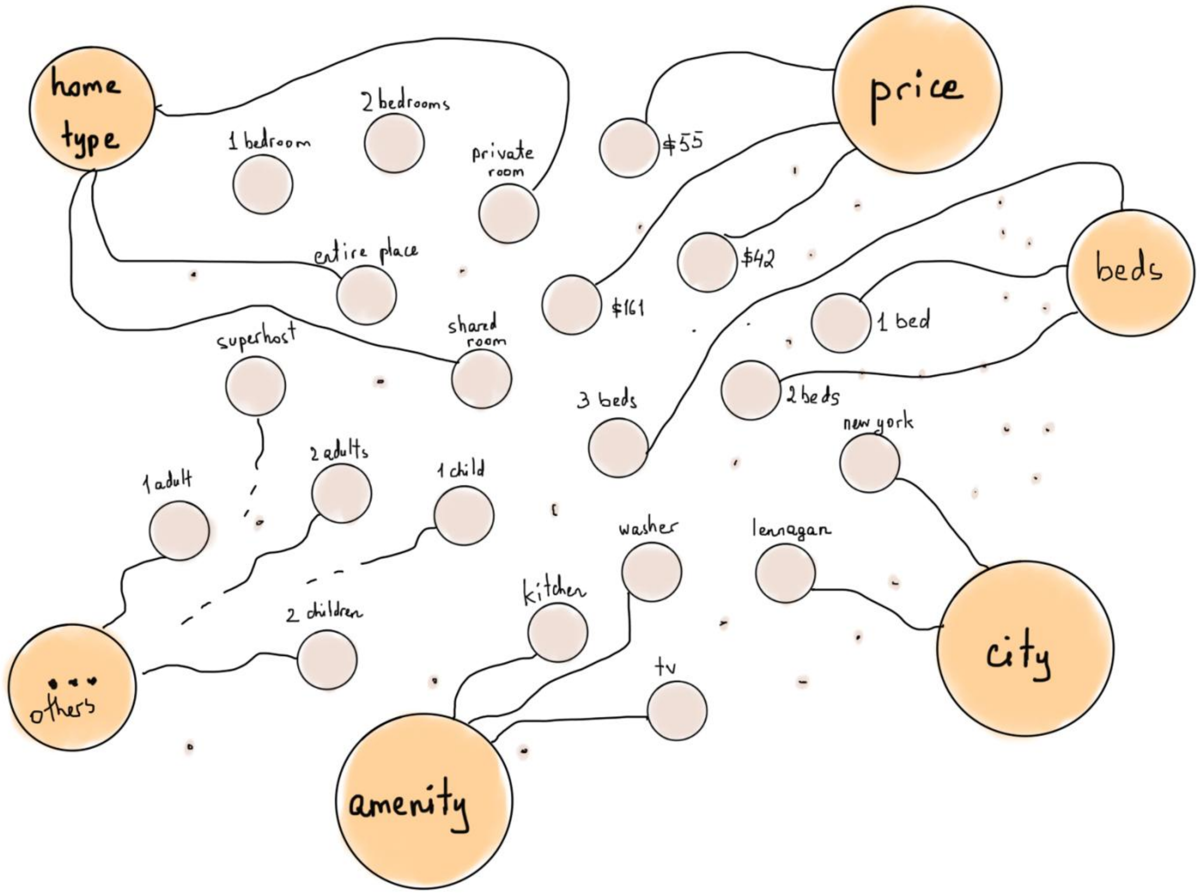
ISSN 2167-5163

Viktor Antonovich Sadovnichiĭ	coauthored with	Blagovest Sendov	MR0545340
Blagovest Sendov	coauthored with	Géza Freud	MR0190612
Géza Freud	coauthored with	Paul Erdős[1]	MR0420119

Change First Author | Change Second Author | New Search

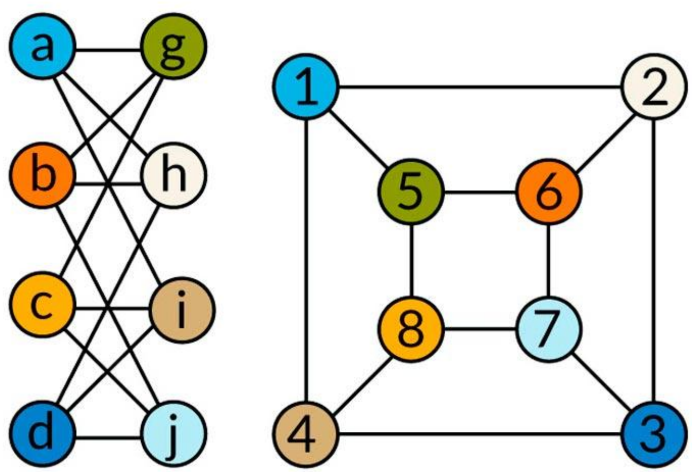
Двудольный граф

Граф *G = <V, E>* называется **двудольным**, если множество его рёбер разбито на два подмножества: *V = V1 U V2, V1 ∩ V2 = Ø*, и рёбрами связаны только вершины из разных подмножеств.

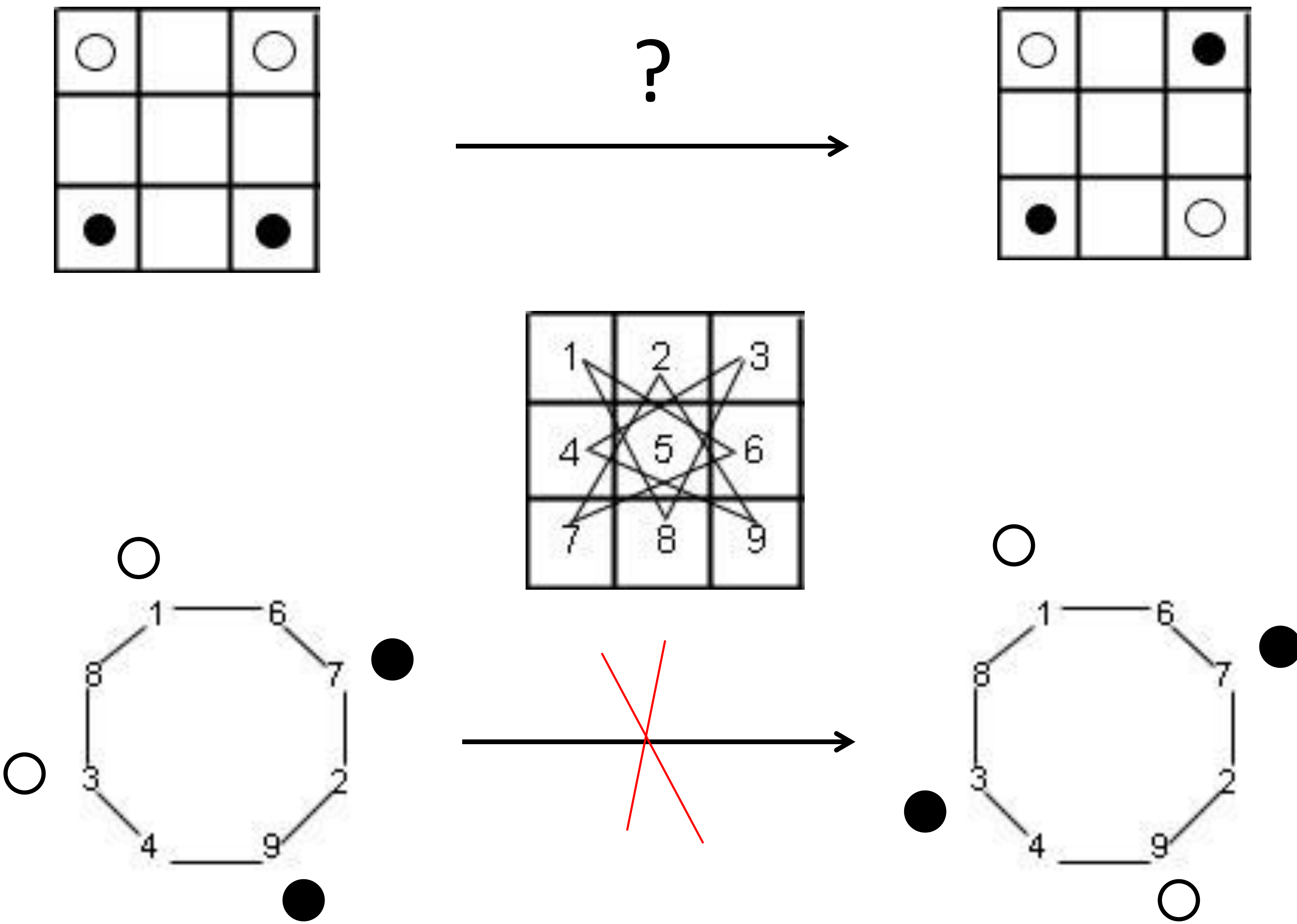


Изоморфизм

Графы $G = \langle V, E \rangle$ и $G' = \langle V', E' \rangle$ **изоморфны**, если существует такое взаимно-однозначное соответствие между множествами их вершин V и V', что вершины соединены ребрами в одном графе тогда и только тогда, когда соответствующие им вершины соединены ребрами в другом графе.



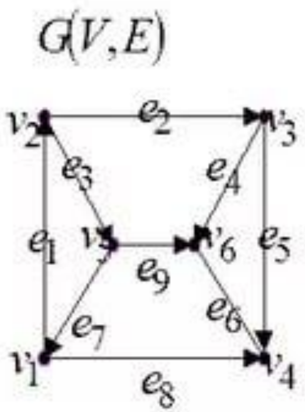
Задача о конях



Машинное представление графов

Матрица инцидентности – матрица со строками, соответствующими вершинам, и столбцами, соответствующими ребрам.

Для орграфов разные авторы используют разные варианты представления ребер *<x; y>*: либо (-1, 1), либо наоборот.



матрица инциденций графа

$$
A = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array}
\begin{array}{ccccccccc}
e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & e_9 \\
1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
-1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & -1
\end{array}
$$

Рисунок 3.3

Худший способ представления графа:

- требует m x n ячеек памяти
- неудобный доступ.

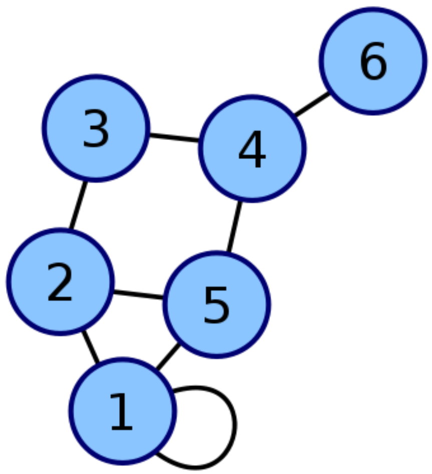
Проверка существования ребра между вершинами требует перебора всех столбцов

Машинное представление графов

Матрица смежности (вершин) – матрица *n x n* с индексами *Bij = 1,* если существует ребро из вершины *vi* в вершину *vj,* и нулю в остальных случаях. Для неографа симметрична.

Проверка существования ребра между вершинами выполняется за один шаг.

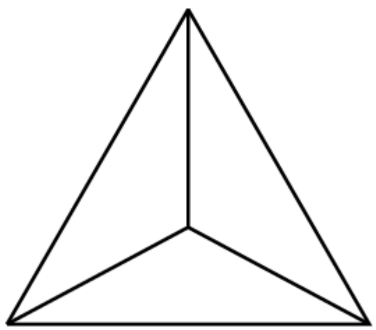
Прежний недостаток: требует *n x n* ячеек памяти.



$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

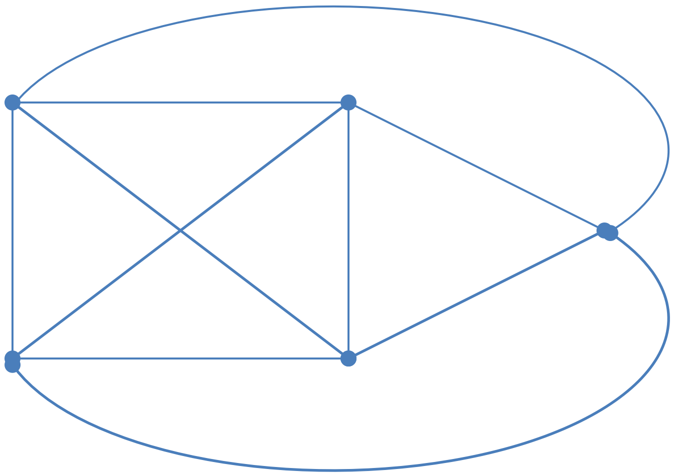
Машинное представление графов

Матрица смежности (вершин) – матрица *n x n* с индексами *Bij = 1*, если существует ребро из вершины *vi* в вершину *vj,* и нулю в остальных случаях. Для неографа симметрична.



Полный граф!

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

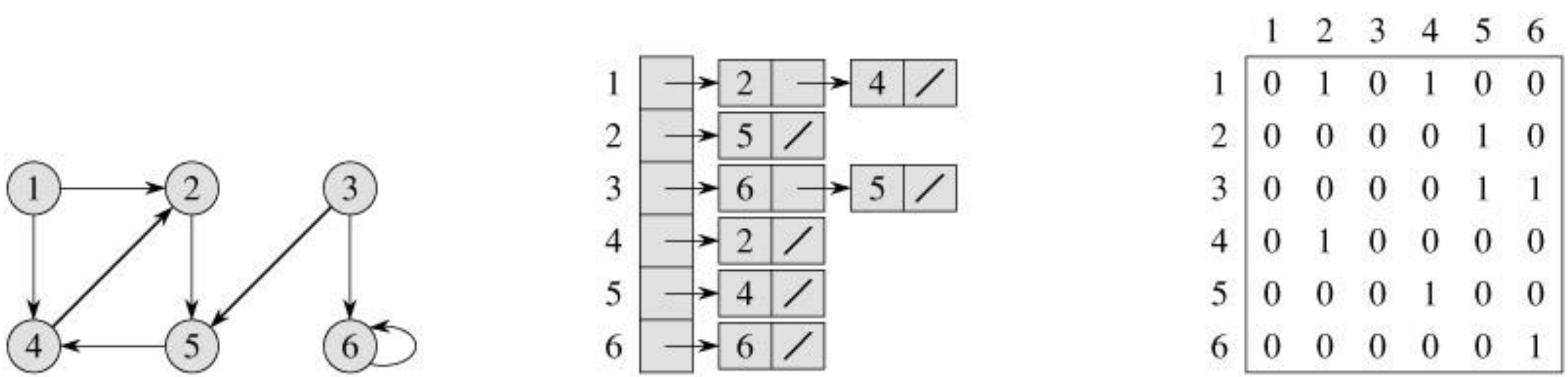


?

Машинное представление графов

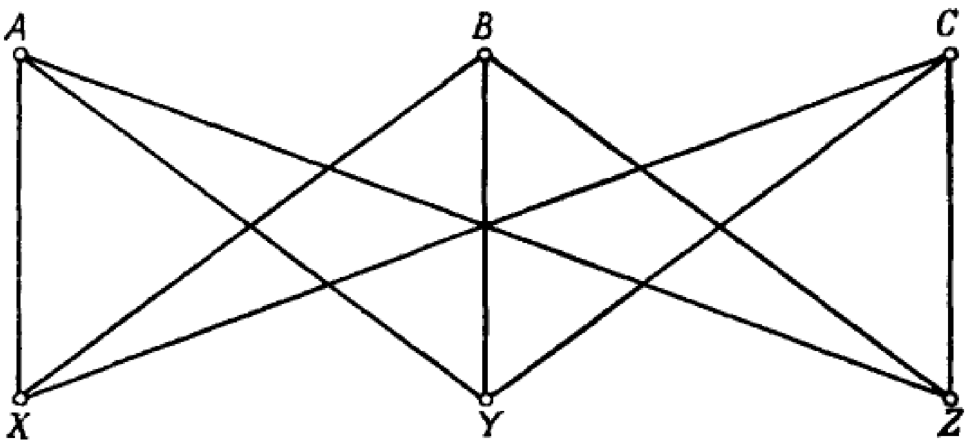
Список инцидентности (ребер) – 2 столбца по m ячеек с указанием вершин, инцидентных данному ребру. Наиболее компактный способ представления графов.

Список смежности (вершин) – список вершин, смежных с данной. Требует O(m+n) ячеек. Для разреженных графов - наиболее удобный способ хранения.

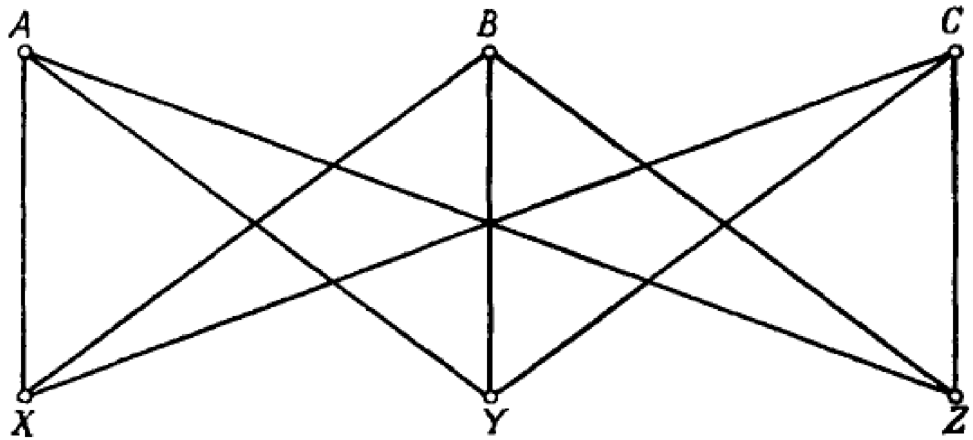


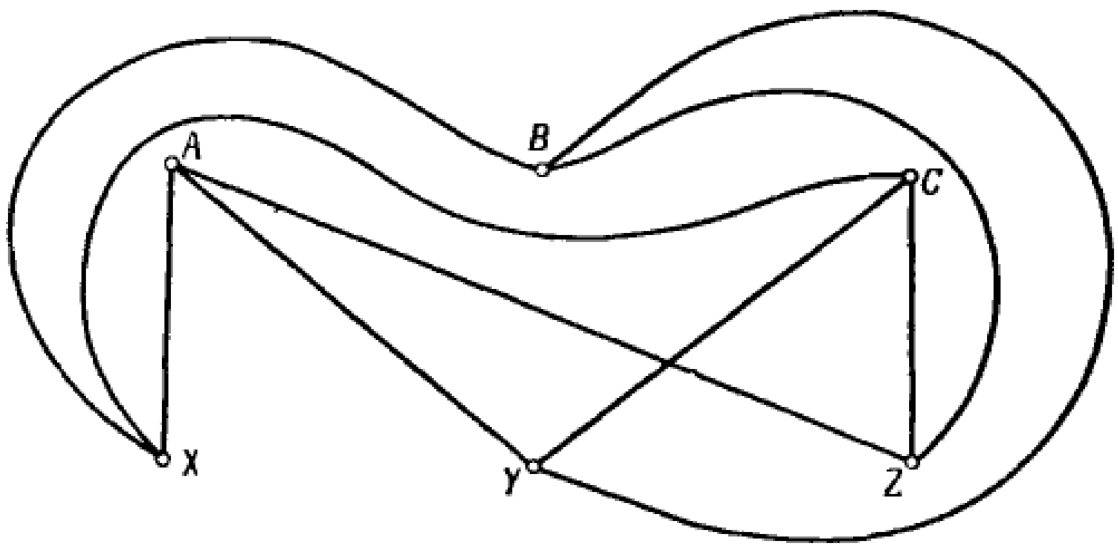
(a) (b) (c)

Задача о домиках и колодцах



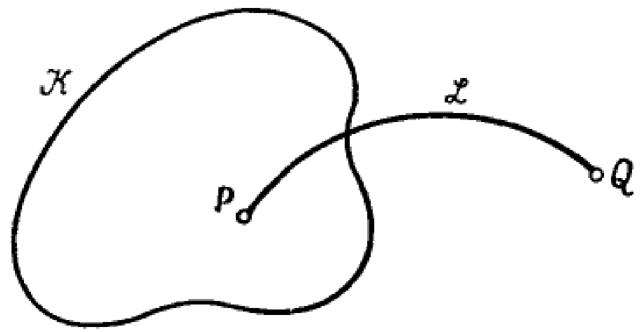
Задача о домиках и колодцах



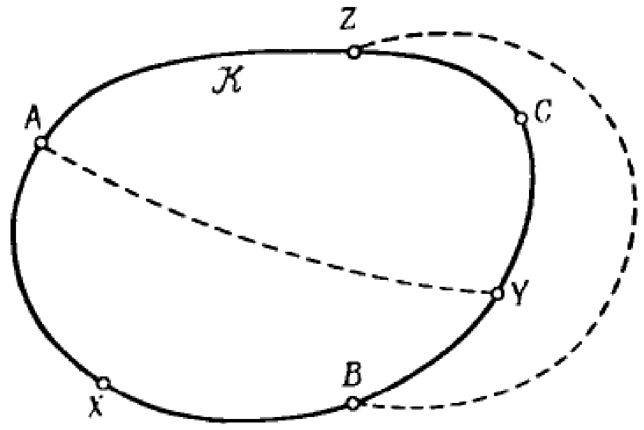


Задача о домиках и колодцах

Теорема Жордана о кривых. Пусть $\mathscr{K}$ - непрерывная замкнутая линия на плоскости. Тогда любая непрерывная линия $\mathscr{L}$, соединяющая произвольную точку $\mathcal{P}$ из внутренней области и произвольную точку $\mathcal{Q}$ из внешней области, пересекает $\mathscr{K}$.

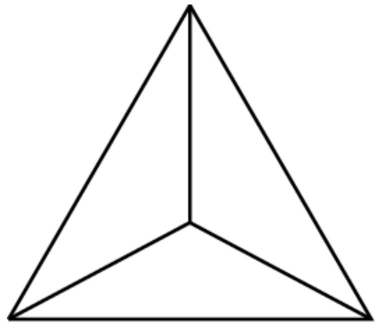


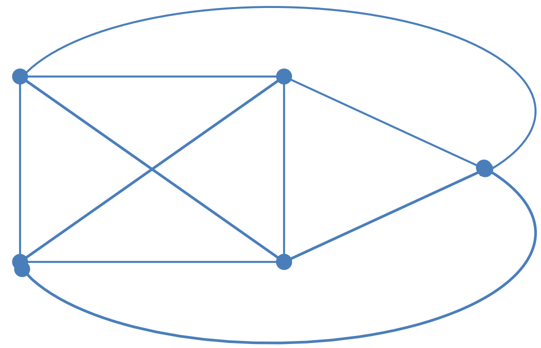
Следствие. Если точки $\mathcal{A}$ и $\mathcal{Y}$, лежащие на кривой $\mathscr{K}$, соединить кривой $\mathscr{AY}$, не имеющей с $\mathscr{K}$ общих точек, кроме $\mathcal{A}$ и $\mathcal{Y}$, то эта кривая будет лежать либо внутри $\mathscr{K}$, либо вне.



Плоские графы

Планарный граф — граф, который может быть изображён на плоскости без пересечения рёбер. Иначе говоря, граф планарен, если он изоморфен некоторому плоскому графу, то есть графу, изображённому на плоскости так, что его вершины — это точки плоскости, а рёбра — непересекающиеся кривые на ней. Области, на которые граф разбивает плоскость, называются его гранями. Неограниченная часть плоскости — тоже грань, так называемая внешняя грань.



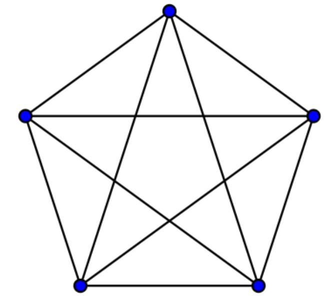


Формула Эйлера для плоского графа. Число вершин *V*, ребер *E* и граней *F* для плоского графа G связаны соотношением

$$|V(G)| - |E(G)| + |F(G)| = 2.$$

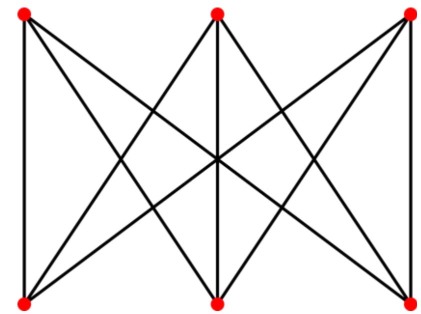
Эта же формула справедлива и для многогранников, топологически эквивалентных сфере.

Плоские графы



Следствие. Если каждая грань ограничена не менее чем тремя рёбрами и каждое ребро разделяет две грани, то $3\ |F(G)| \le 2\ |E(G)|$, и, подставляя в предыдущее выражение, получаем

$$E(G)| \le 3\ |V(G)| - 6.$$



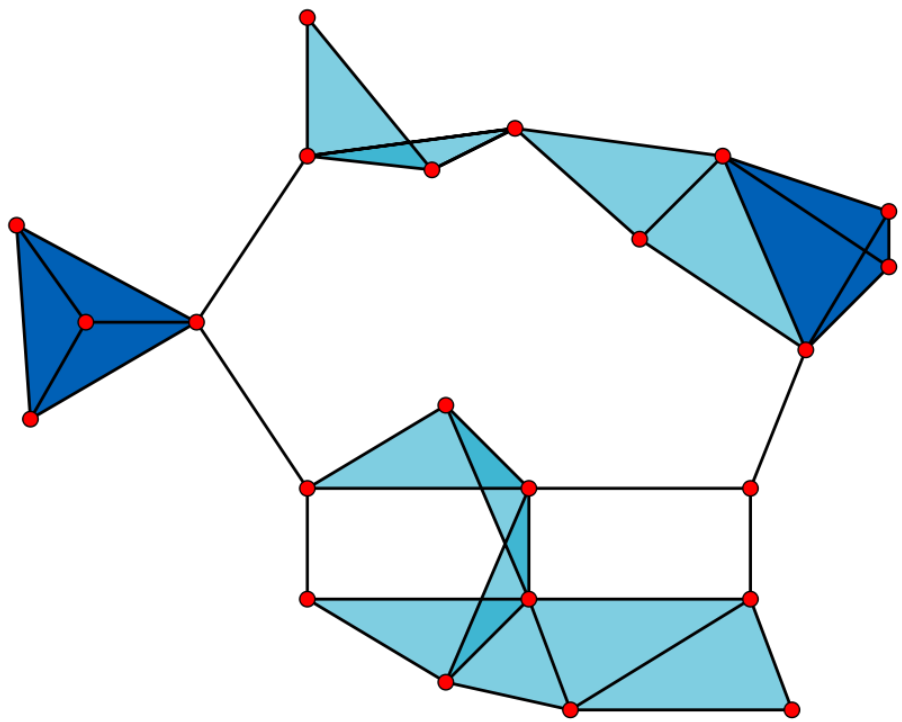
т.е. **при достаточно большом числе рёбер граф заведомо не планарен**.

Задача о раскраске графа (1852). Плоский граф всегда можно раскрасить четырьмя красками так, чтобы никакие две соседние грани не были одного цвета (доказана вычислительно в 1976 г.).

Задача о клике

Кликой в неориентированном графе *G = <V, E>* называется подмножество вершин *C*, принадлежащее *V*, такое что для любых двух вершин в *C* существует ребро, их соединяющее. Это эквивалентно утверждению, что подграф, порождённый подмножеством *C*, является полным.

Задача о клике: определить, существует ли в заданном графе *G* клика размера *k*, *или* найти в заданном графе *G* клику максимального размера. Относится к классу **NP-полных задач**. Используется, в частности, для проведения кластеризации.



Generalized gene co-expression analysis via subspace clustering using low-rank representation (2019)

Пути и циклы в графах

Путем в графе *G = <V, E>* называют последовательность рёбер вида *<e(1), e(2), …, e(n-1)> = S = <(v(1), v(2)), (v(2), v(3)), …, (v(n-1), v(n))>*. Говорят, что этот путь идёт из *v(1)* в *v(n)* и имеет длину (n-1) (если веса рёбер равны 1).

Путь называют **простым**, если все рёбра и вершины на нём различны, кроме быть может первой и последней.

Цикл – это простой путь длины не менее 1, который начинается и заканчивается в одной вершине. В простом неографе длина цикла не меньше 3**.**

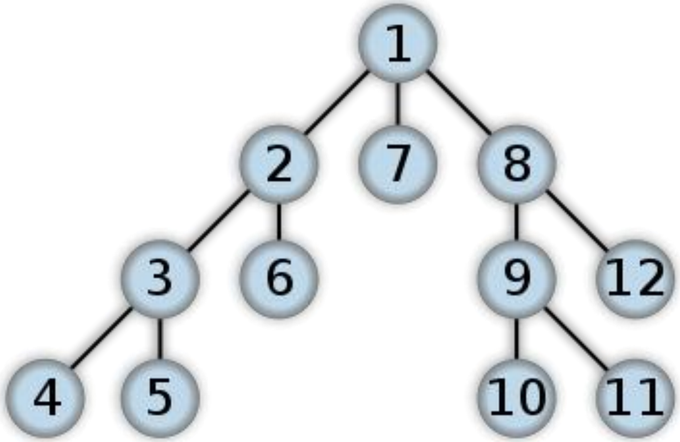
Для нахождения пути между заданными вершинами *u* и *v* могут быть использованы два вида поиска: **в глубину и в ширину**. При поиске в глубину последовательность вершин, определяющих путь между *u* и *v* уже содержится в памяти алгоритма на момент обнаружения вершины v. Однако этот путь может быть не кратчайшим.

Поиск в графе

Поиск в глубину:

Начинаем с некоторой вершины *v(0)*, затем переходим к произвольной вершине *u*, смежной с *v(0)*, и повторяем процесс. В общем случае, при нахождении в вершине v проверяем наличие ещё не посещённых смежных вершин. Если такие существуют, то переходим к одной из них и повторяем процедуру. В противном случае отмечаем вершину *v*, как посещённую, и возвращаемся в ту вершину, откуда попали в *v*.

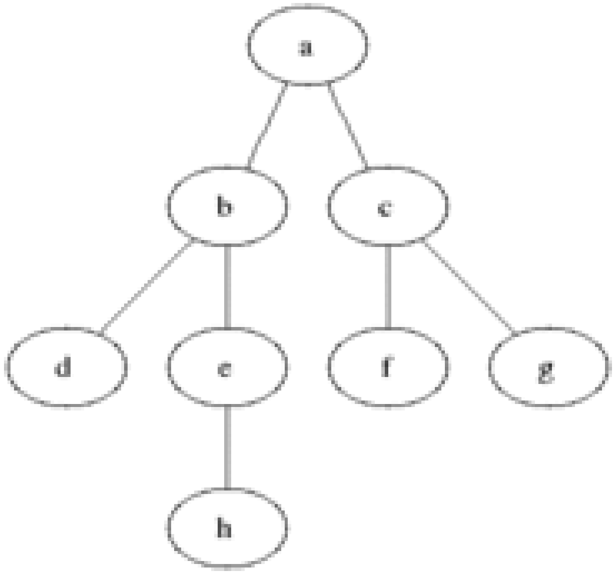
Каждая вершина посещается не более одного раза и можно показать, что сложность этого алгоритма *O(n + m)*.



Поиск в графе

Поиск в ширину:

Начиная с вершины-источника *u*, рассмотрим все рёбра *(u,v)*, выходящие из этой вершины. Если очередная вершина *v* является искомой, то поиск завершается; в противном случае вершина *v* добавляется в очередь. После того, как будут проверены все рёбра, выходящие из вершины *u*, из очереди извлекается следующая вершина *u’*, и процесс повторяется. Сложность этого алгоритма также *O(n + m)*.



Белый — вершина, которая еще не обнаружена. Серый — вершина, уже обнаруженная и добавленная в очередь. Черный — вершина, извлечённая из очереди.

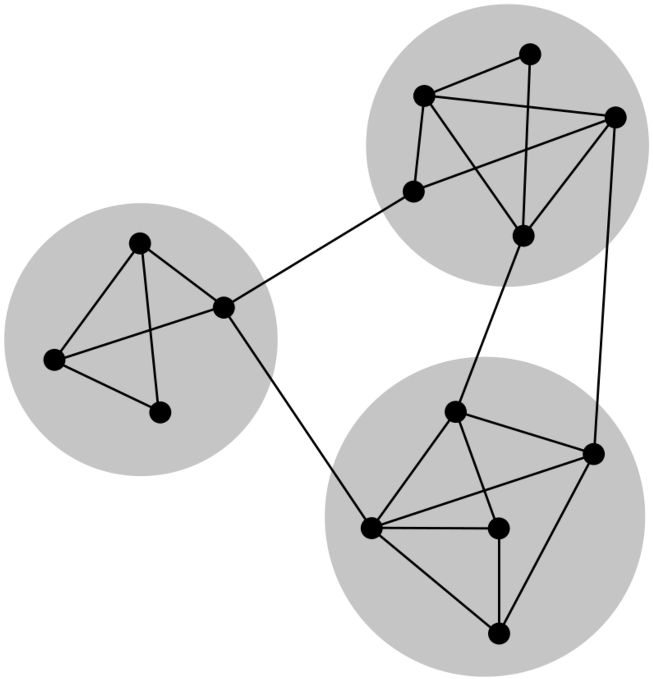
Связность графов

Пусть G – неограф. **Две вершины *a* и *b* называются связанными, если существует путь *S* с начальной вершиной *a* и конечной вершиной *b*.**

Граф называется **связным**, если связана любая пара его вершин.

Для всякого графа существует такое разбиение множества его вершин на попарно непересекающиеся множества V_i

$$V = \bigcup_{i=1}^{k} V_i$$



что вершины в каждом V_i связаны, а вершины из различных V_i не связаны. В этом случае граф состоит из k непересекающихся связных подграфов – k *компонентов связности*.

Связность графов

Теорема 2: Если в конечном неориентированном простом графе G ровно две вершины a и b имеют нечетную степень, то они связаны.

По теореме 1 каждый конечный граф имеет четное число вершин нечетной степени, это же верно и для компонент.

Теорема 3: Если неориентированный простой граф G имеет n вершин и k связных компонент, то максимальное число рёбер в G

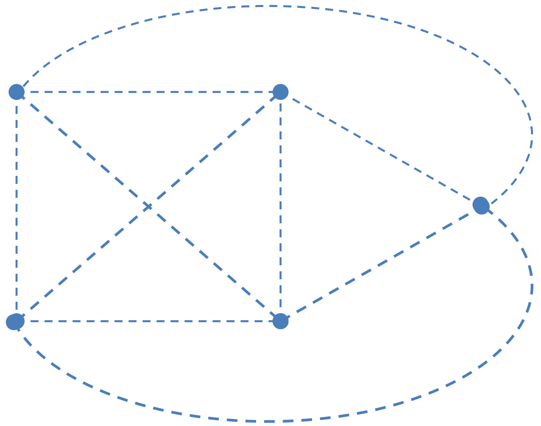
$$N(n, k) = \tfrac{1}{2}\,(n - k)\,(n - k + 1)$$

Доказательство опустим.

Теорема 4 (следствие): Простой неограф с n вершинами и числом рёбер m, большим, чем $N(n, 2) = \tfrac{1}{2}\,(n - 2)\,(n - 1)$, связан.

Обязательно ли будет ли связан граф G

при n = 5 и m = 5? m = 6? m = 7?



Деревья

Связный неограф называется деревом, если он не имеет циклов. Кроме того, дерево не имеет петель и кратных ребер.

Дерево может быть укорененным и неукорененным.

Лес – упорядоченное множество упорядоченных деревьев.

Двоичное дерево – ориентированное дерево, в котором исходящие степени вершин (число исходящих рёбер) не превосходят 2.

Длина ребра – число, соотнесенное с каждым ребром и обозначающее, **в каком-то смысле**, расстояние между двумя вершинами, соединенными этим ребром.

Длина пути – сумма всех длин ребер, составляющих путь.

Теорема 5: В дереве любые две вершины связаны единственным простым путем.

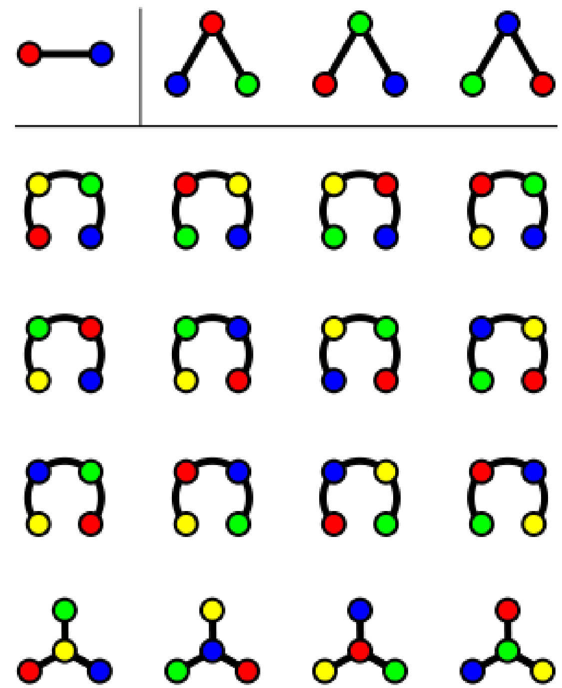
Док-во: Если бы путей было 2, то был бы цикл.

Деревья

Теорема 7: Любое дерево с n вершинами содержит *(n – 1)* ребро.

Док-во: По индукции по числу вершин. Для *n = 1* очевидно. Пусть *n > 1*. Тогда в дереве существует концевая вершина v, удаляя которую вместе с инцидентным ребром (*u, v*), получим дерево с *(n – 1)* вершиной, которое, по предположению, имеет *(n – 2)* ребра. Значит, в исходном дереве было *(n – 2 + 1) = (n – 1)* ребро.

Теорема 8: Число различных деревьев, которые можно построить на *n* маркированных вершинах, равно *n^(n-2)* (**Теорема Кэли**).



Для деревьев с немаркированными вершинами формулы не существует (?).

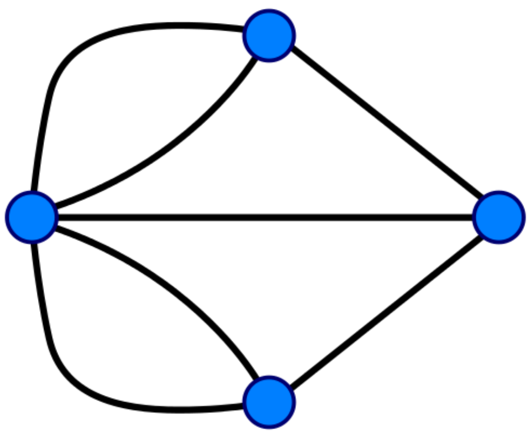
Эйлеровы пути и циклы

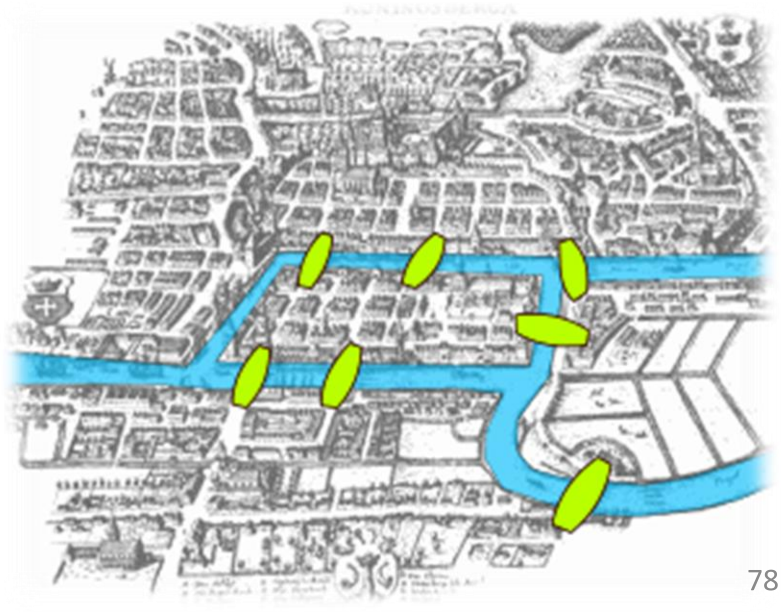
Эйлеровым путем в графе G называется произвольный путь, проходящий **через каждое ребро** графа в точности один раз. Если начальная и конечная вершины совпадают, то путь называется **эйлеровым циклом**.

Теорема 11: Эйлеров **путь** в простом неографе существует тогда и только тогда, когда граф связный и содержит не более 2 вершин нечетной степени.

Теорема 12: Эйлеров **цикл** в простом неографе существует тогда и только тогда, когда граф связный и степени всех его вершин четные.

Доказательство традиционно опустим ☺





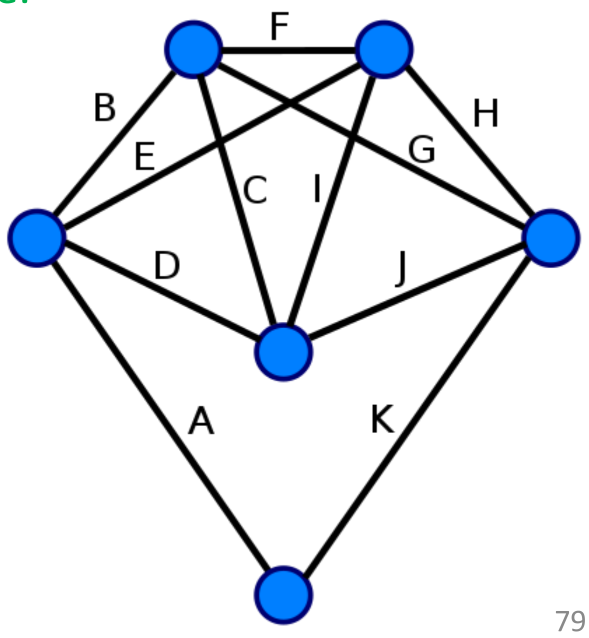
Эйлеровы пути и циклы

Эйлеровым путем в графе G называется произвольный путь, проходящий **через каждое ребро** графа в точности один раз. Если начальная и конечная вершины совпадают, то путь называется **эйлеровым циклом**.

Теорема 11: Эйлеров **путь** в простом неографе существует тогда и только тогда, когда граф связный и содержит не более 2 вершин нечетной степени.

Теорема 12: Эйлеров **цикл** в простом неографе существует тогда и только тогда, когда граф связный и степени всех его вершин четные.

Доказательство традиционно опустим ☺



Задача коммивояжера
(Travelling salesman problem, TSP)

Поиск самого выгодного маршрута, проходящего через ряд городов хотя бы по одному разу с последующим возвратом в исходный город.

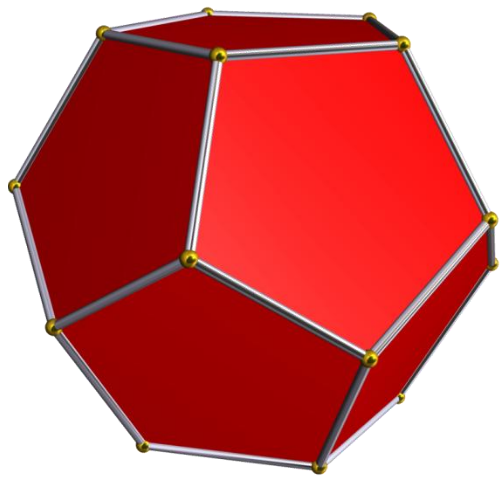
~ (n – 1)! маршрутов для n городов

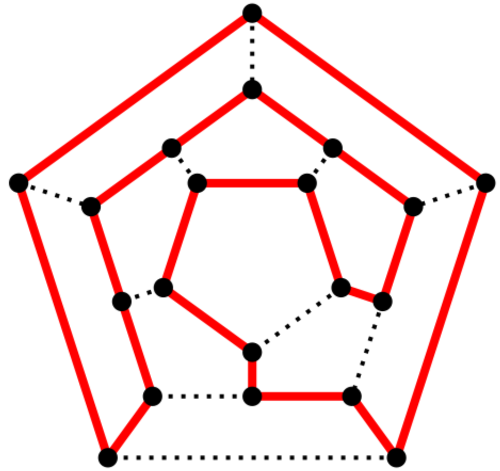
Оптимальный маршрут коммивояжёра через 15 крупнейших городов Германии. Указанный маршрут является самым коротким из всех возможных 43 589 145 600 вариантов.

Для 66 городов задача в настоящее время неразрешима (?).

Гамильтоновы пути и циклы

«Путешествие вокруг света» (1859)





Уильям Роуэн Гамильтон (1805 — 1865)

Простой путь (цикл) называется **гамильтоновым путем (циклом)**, если он проходит **через каждую вершину** графа ровно один раз.

В отличие от эйлеровых путей, неизвестно ни одного простого необходимого и достаточного условия для существования гамильтоновых путей и циклов.

Неизвестен и алгоритм, проверяющий существование такого пути в произвольном графе за полиномиальное время от числа вершин *n* (NP-полная задача).

Гамильтоновы пути и циклы

Теорема 17: Если в графе G с n вершинами для любой пары вершин vi и vj имеет место соотношение $\rho(vi) + \rho(vj) \geq (n - 1)$, то граф G имеет **гамильтонов путь**. Если $\rho(vi) + \rho(vj) \geq n$, то имеет **гамильтонов цикл**.

Условие Поша: Пусть граф G имеет $p > 2$ вершин. Если для всякого n, $0 < n < (p-1)/2$, число вершин со степенями меньшими или равными n меньше, чем n, и для нечетного p число вершин со степенью $(p-1)/2$ не превосходит $(p-1)/2$, то G — гамильтонов граф. **Это достаточное условие не является необходимым**.

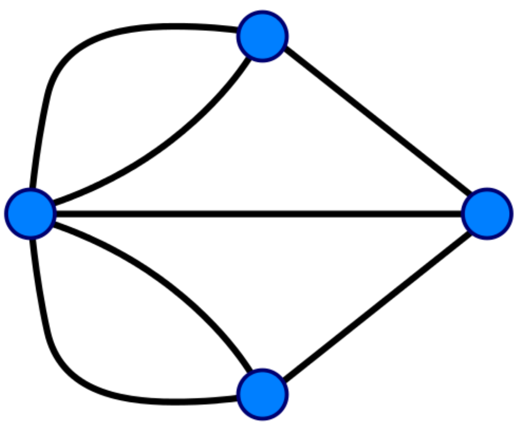
Условие Дирака: пусть p — число вершин в данном графе и $p > 3$; если степень каждой вершины не меньше, чем p/2, то данный граф — гамильтонов.

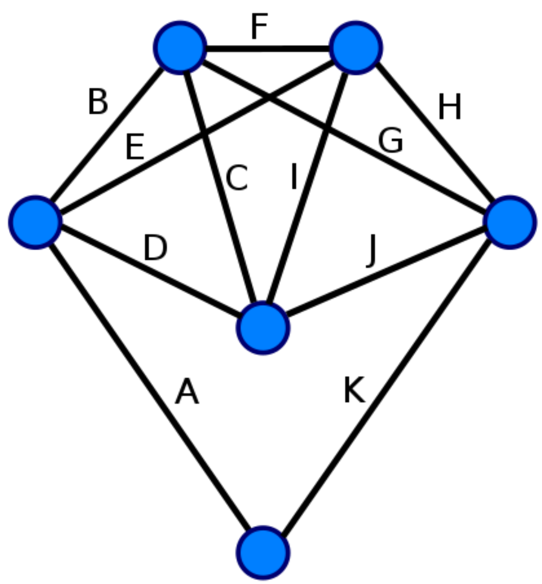
Условие Оре: пусть p — число вершин в данном графе и $p > 2$. Если для любой пары несмежных вершин (x, y) выполнено неравенство $\deg x + \deg y \geq p$, то данный граф — гамильтонов.

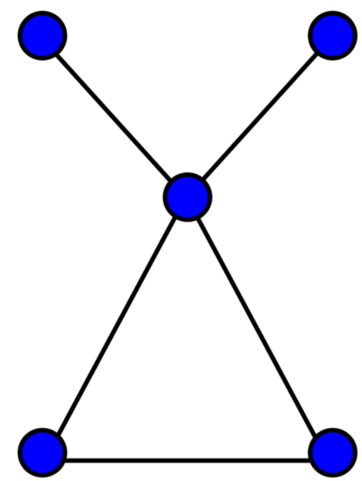
Гамильтоновы пути и циклы

Теорема 17: Если в графе G с n вершинами для любой пары вершин vi и vj имеет место соотношение $\rho(vi) + \rho(vj) \geq (n - 1)$, то граф G имеет **гамильтонов путь**. Если $\rho(vi) + \rho(vj) \geq n$, то имеет **гамильтонов цикл**.

Условие Поша: Пусть граф G имеет $p > 2$ вершин. Если для всякого n, $0 < n < (p-1)/2$, число вершин со степенями меньшими или равными n меньше, чем n, и для нечетного p число вершин со степенью $(p-1)/2$ не превосходит $(p-1)/2$, то G — гамильтонов граф. **Это достаточное условие не является необходимым**.







Последовательности де Брёйна

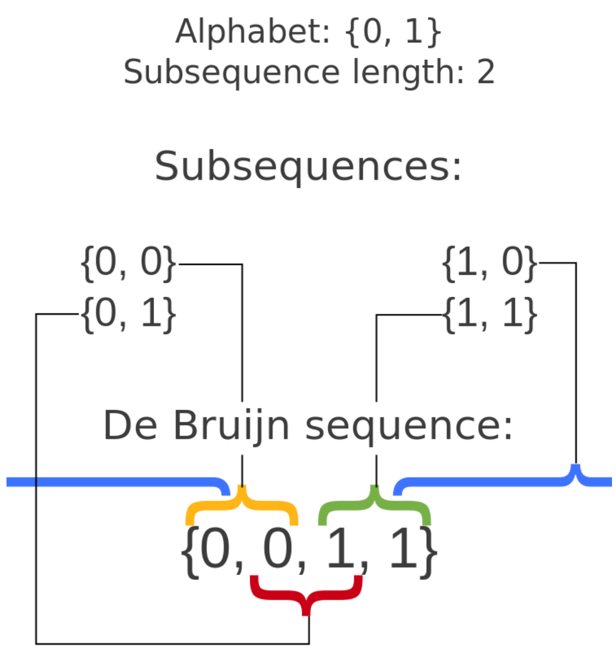
Последовательностью де Брёйна порядка *n* на алфавите *A* размера *k* называется такая циклическая последовательность *B(k, n)*, в которой единожды встречаются все возможные последовательности длины *n* составленные из букв алфавита *A*.

Длина такой последовательности *k^n*, что также является числом различных строк длины *n* из алфавита *A*.

Николас де Брёйн

(1918 – 2012)

EN = 1



Число различных последовательностей *B(k, n)* =

$$\frac{(k!)^{k^{n-1}}}{k^n}$$

Последовательности де Брёйна

Последовательность де Брёйна может быть получена посредством нахождения **гамильтонова цикла** в n-мерном графе де Брёйна или, что то же, **эйлерова цикла** в (n – 1)-мерном графе де Брёйна.

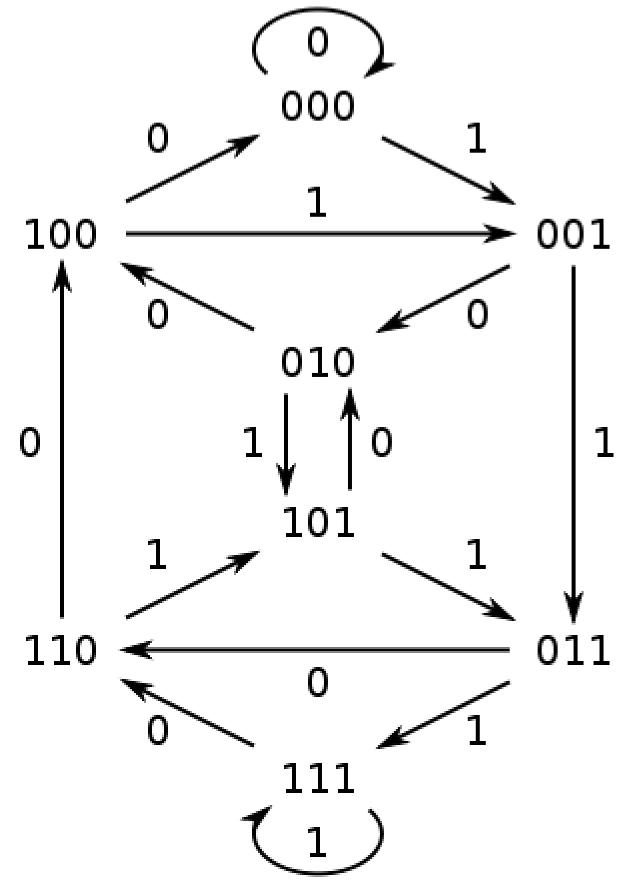
Рассмотрим построение последовательности B(k = 2, n = 4) длиной 2^4 = 16 с использованием эйлерова цикла на (n – 1) = 3-мерном графе.

Существует ли он?

Да! Ибо у всех вершин чётная степень.

Например, список вершин цикла может быть таков: 000, 000, 001, 011, 111, 111, 110, 101, 011, **110**, 100, 001, **010**, 101, 010, 100, 000.

Тогда B(2, 4) = 0 0 0 0 1 1 1 1 0 **1 1 0** 0 1 0 1 (0 0…)



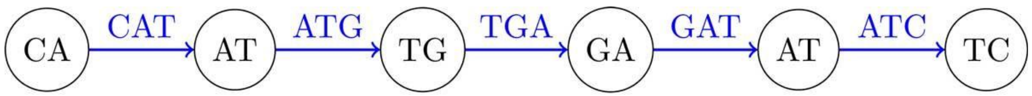
Последовательности де Брёйна

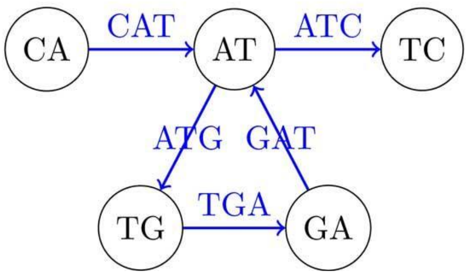
Зачем они нужны? Можно подбирать пин-код в домофоне (если тот реагирует на 4 последних нажатия) ☺

Длина (B (10, 4)) = 10^4 = 10000 => требуется не более 10000 + 3 нажатия для того, чтобы встретилась нужная комбинация цифр.

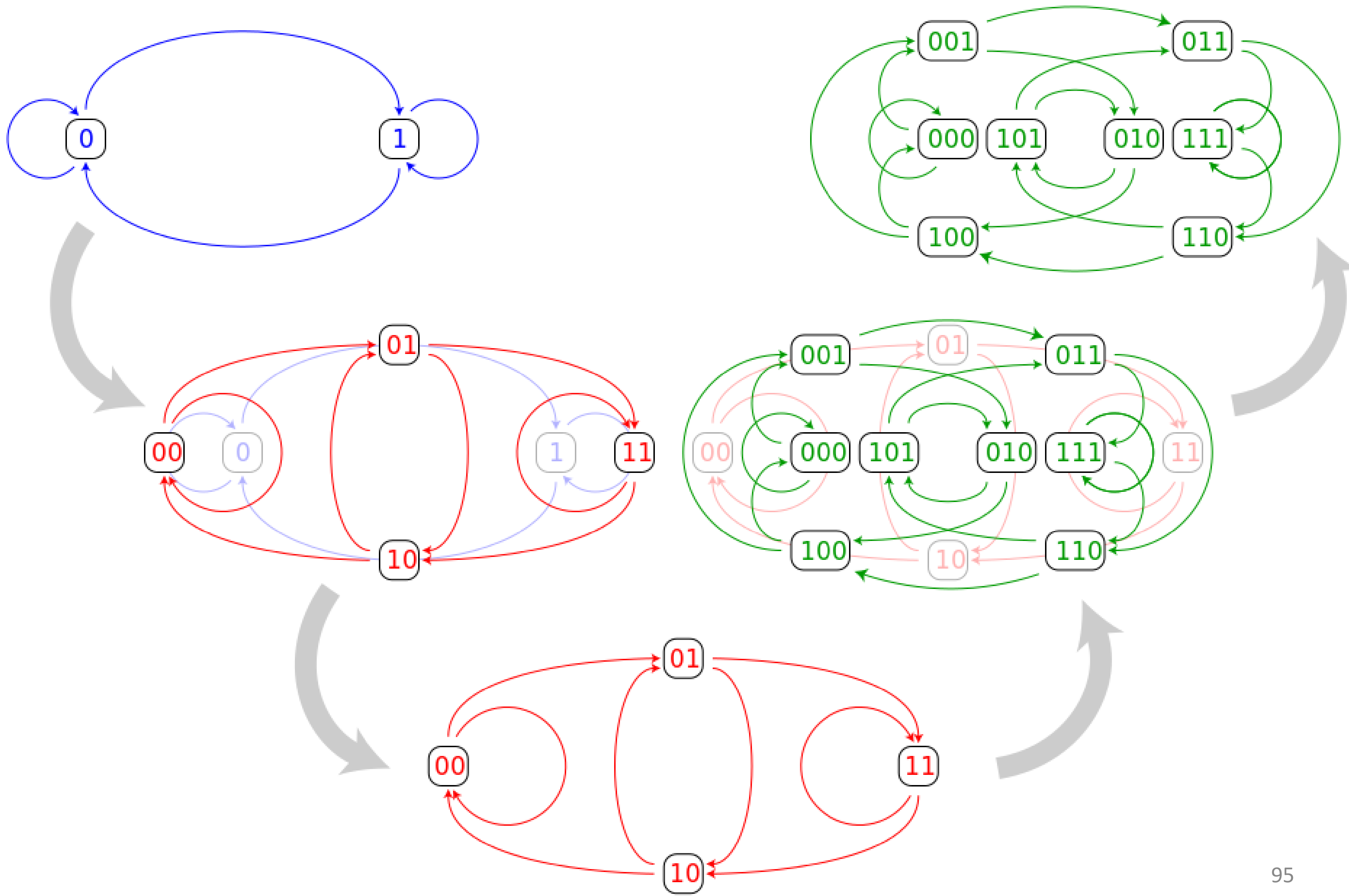
Классический подход дает 10000 * 4 = 40000 нажатий.

Граф де Брёйна — ориентированный граф с *k^n* вершинами, соответствующими *k^n* различных наборов длины n с элементами из алфавита *A* размера *k*, в котором из вершины *(x(1), …, x(n))* в вершину *(y(1), …, y(n))* ребро ведёт в том и только том случае, когда *x(i) = y(i-1)*; при этом самому ребру можно сопоставить набор длины *(x(1), x(2), …, x(n), y(n)) = (x(1), y(1), …, y(n-1), y(n))*.

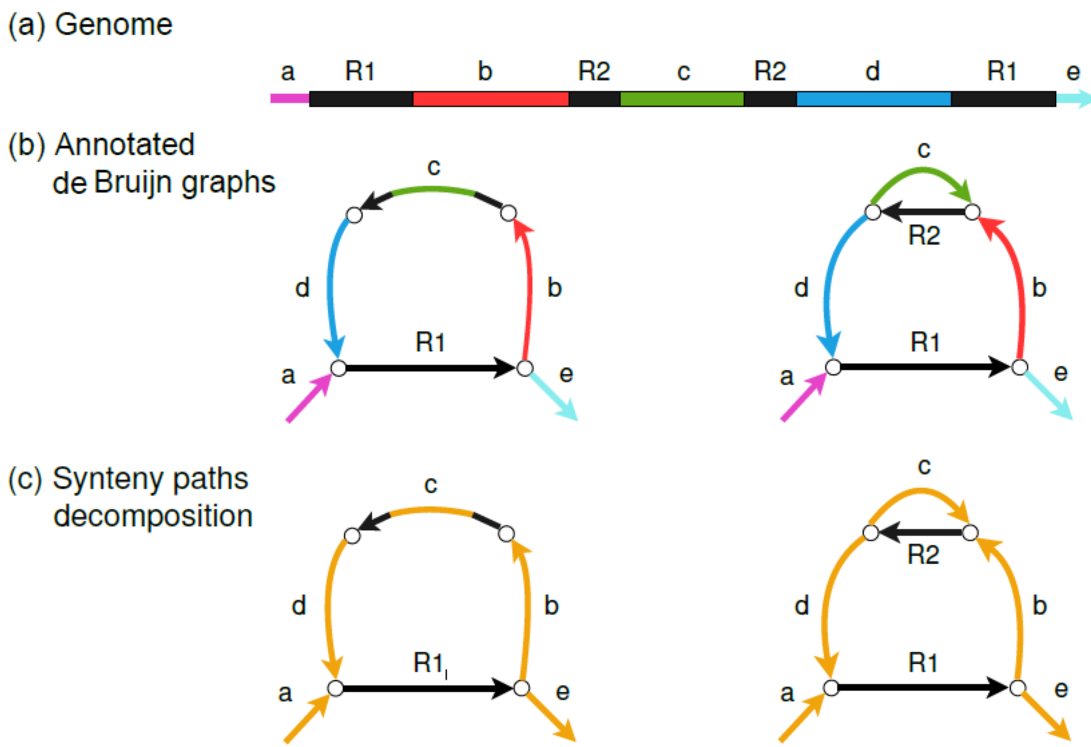




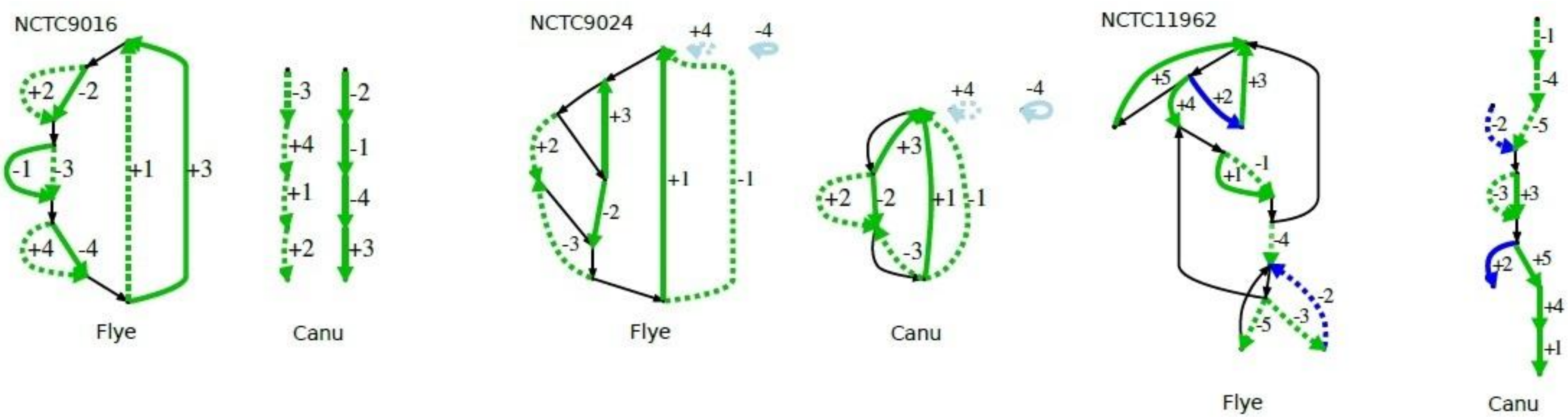
Графы де Брёйна



Графы де Брёйна



Synteny Paths for Assembly Graphs Comparison (2019)



Ойстин Оре

(1899 – 1968)

EN = 3