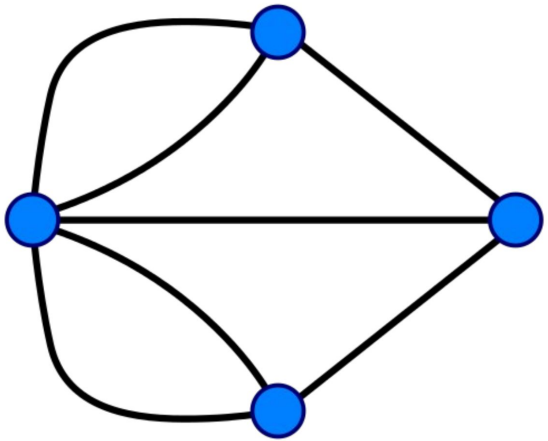
Анализ сетевых данных

Анвар Курмуков, 2019

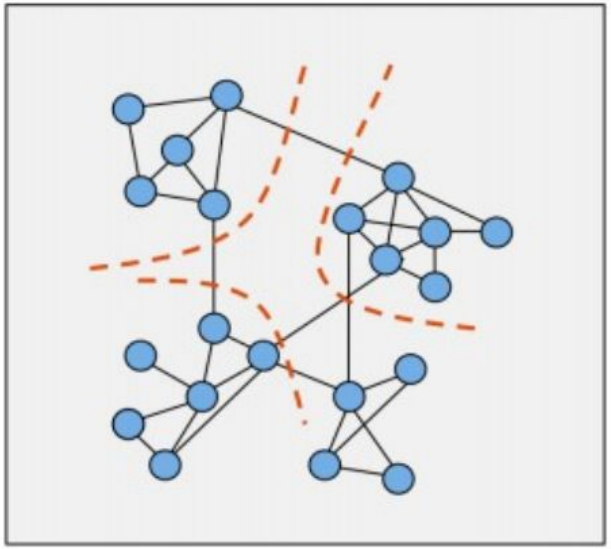
Сети (Графы)

Сетью (графом) называется совокупность объектов (вершин графа) и связей (ребер графа) между этими объектами.



Базовые термины

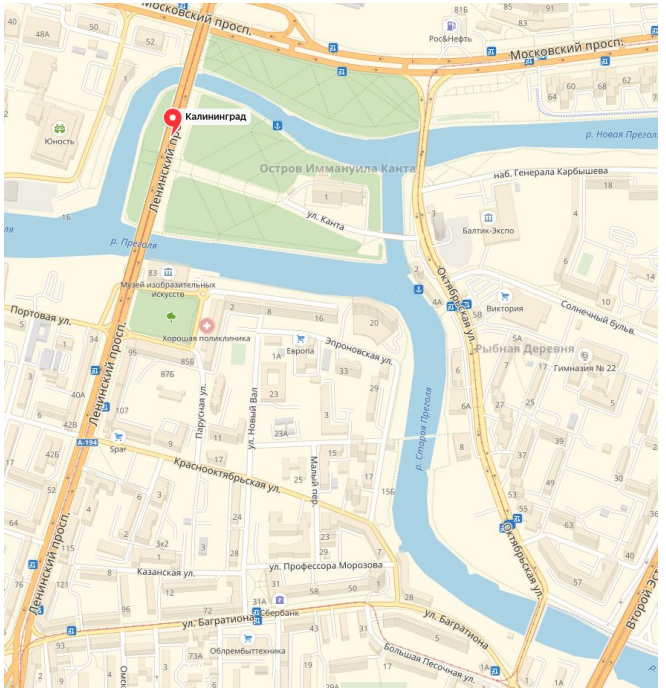
- Граф G(V, E)
  - V - vertices
  - E : V x V - edges
- network = graph (сеть, граф)
- node = vertice, actor (вершина)
- link = edge, relation (ребро, связь)
- cluster = community (кластер, сообщество)
- weighted graph
- directed graph



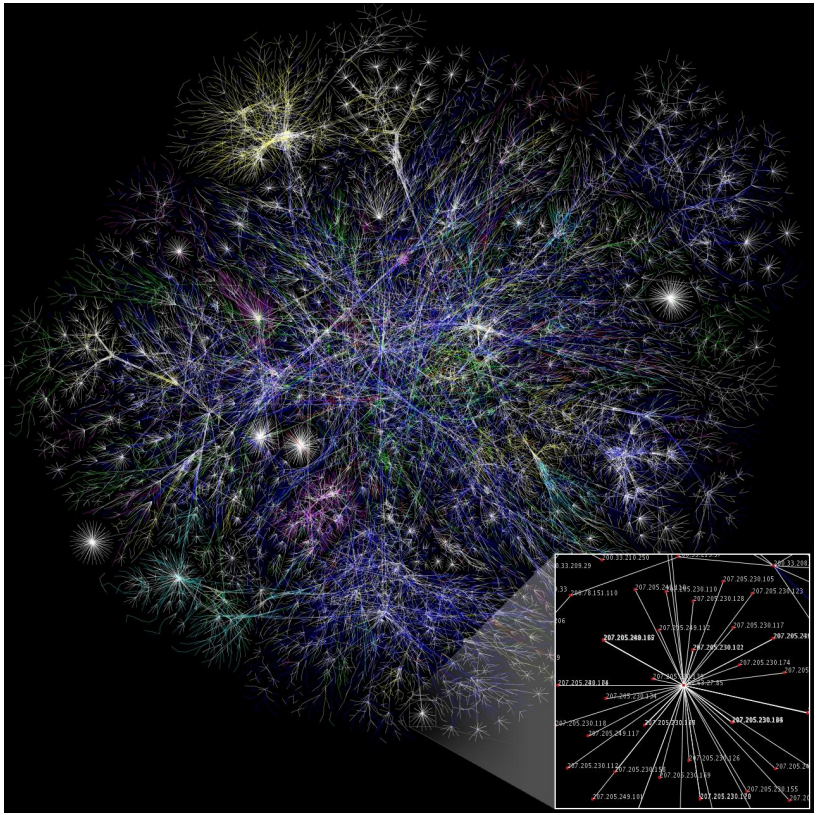
Где возникают сети?

Spoiler: Везде!

Кёнигсбергские мосты

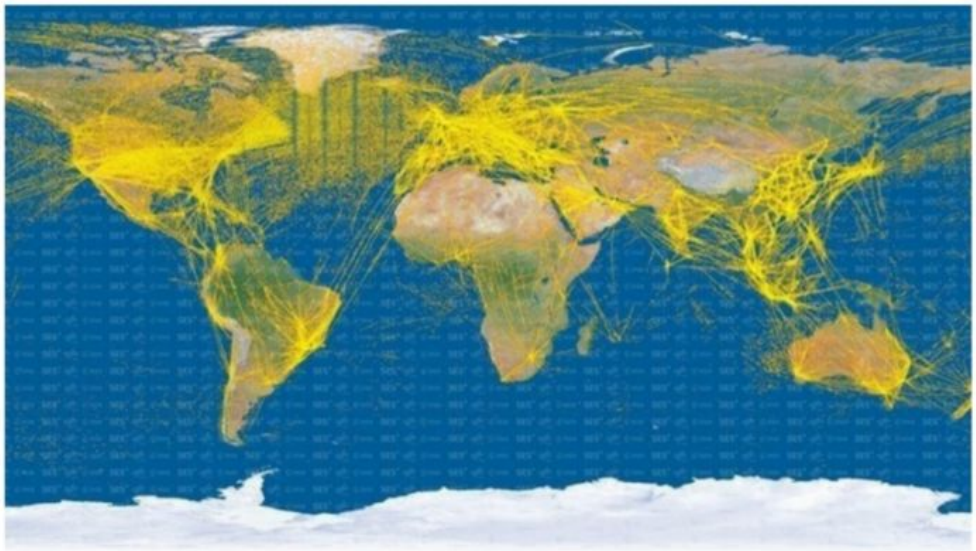


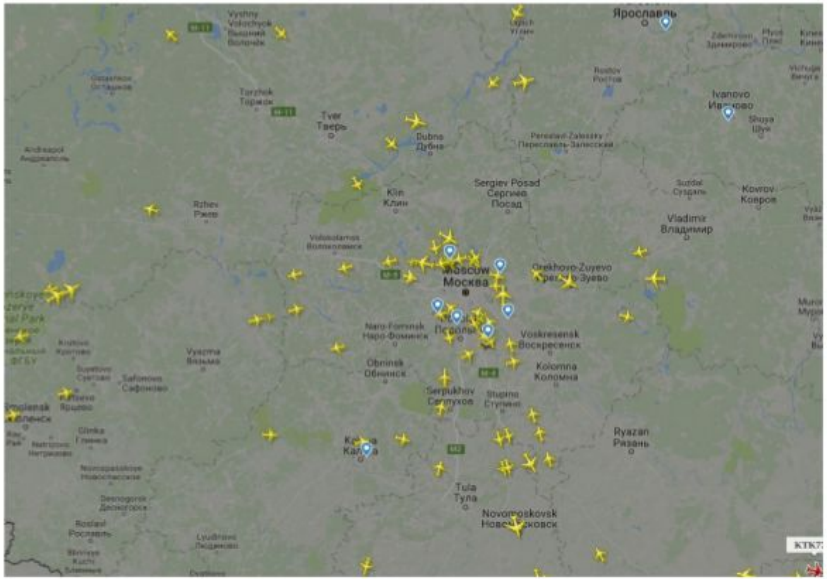
Интернет



Barrett Lyon, Opte Project

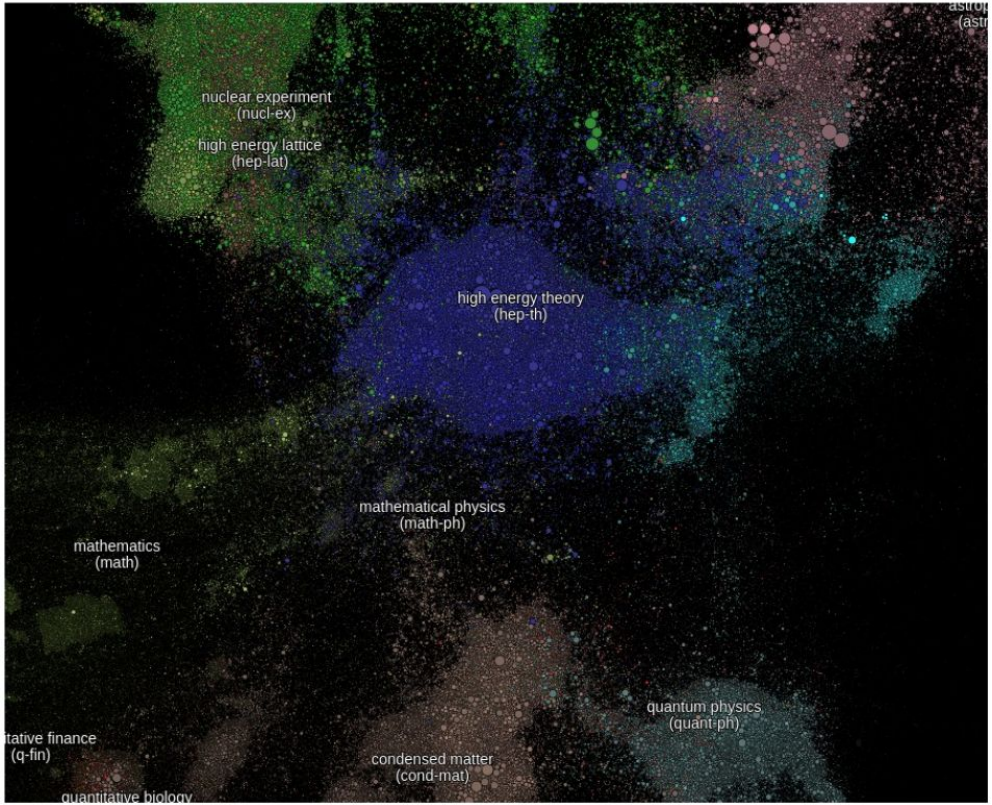
Транспортные сети





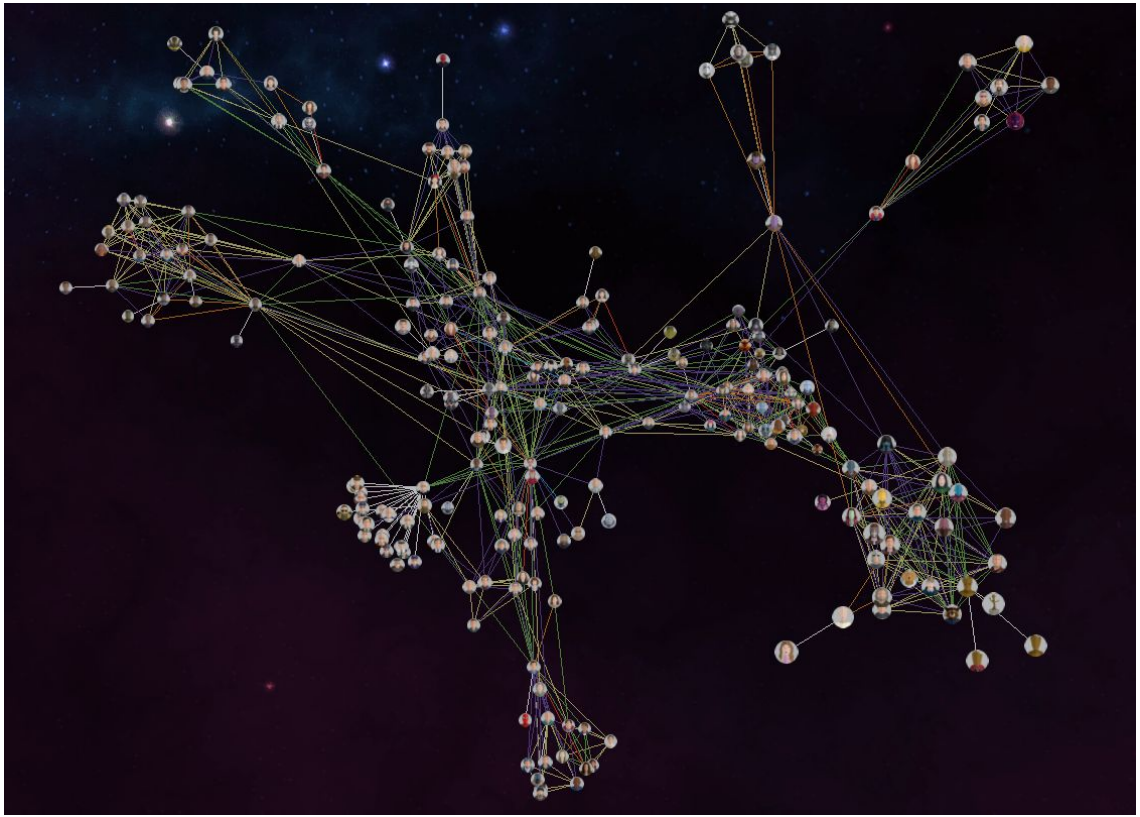
https://www.flightradar24.com/

Сети цитирования



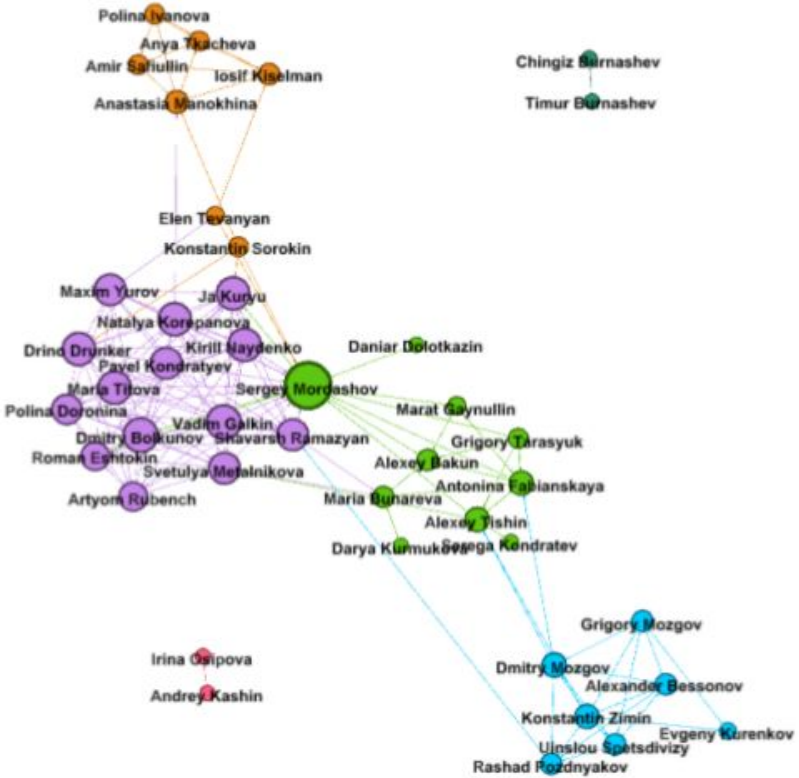
arxiv.org citing graph, http://paperscape.org/

Сети взаимодействия



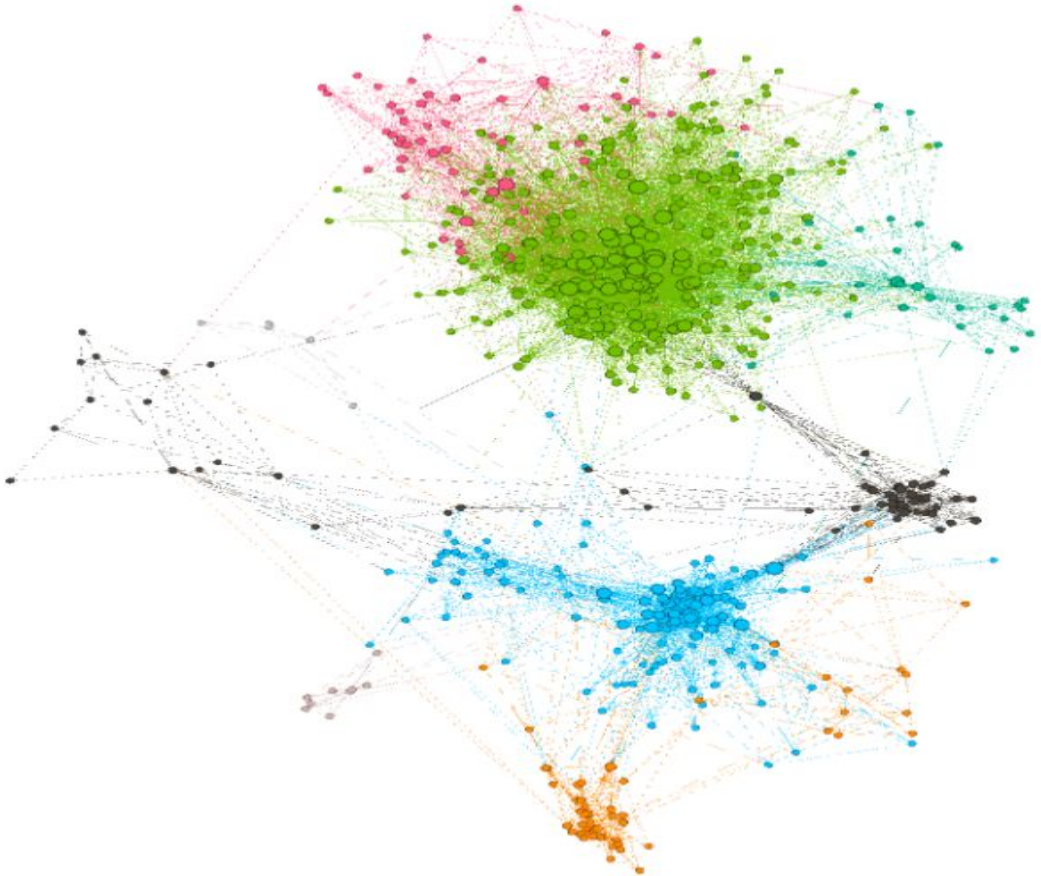
Marvel universe character interaction,
https://www.straitstimes.com

Социальные сети



Anvar’s vkontakte egocentric network, 2014

Социальные сети



Sergey Mordashov vkontakte egocentric network, 2014

Задачи

Какие задачи можно решать с помощью сетей?

- Важность отдельных узлов
- Выявление сообществ
- Связи между отдельными узлами
- Поиск оптимального пути (существование пути)
- Соотношения между разными сетями
- Распространение информации

И многие другие

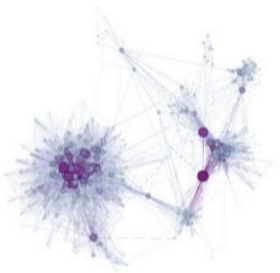
Характеристики вершин сети

- *degree centrality:* степень вершины (число соседей)
- *closeness centrality:* длина пути до других вершин
- *betweenness centrality:* число кратчайших путей проходящих через данную вершину

(a) Betweenness centrality

(b) Closeness centrality

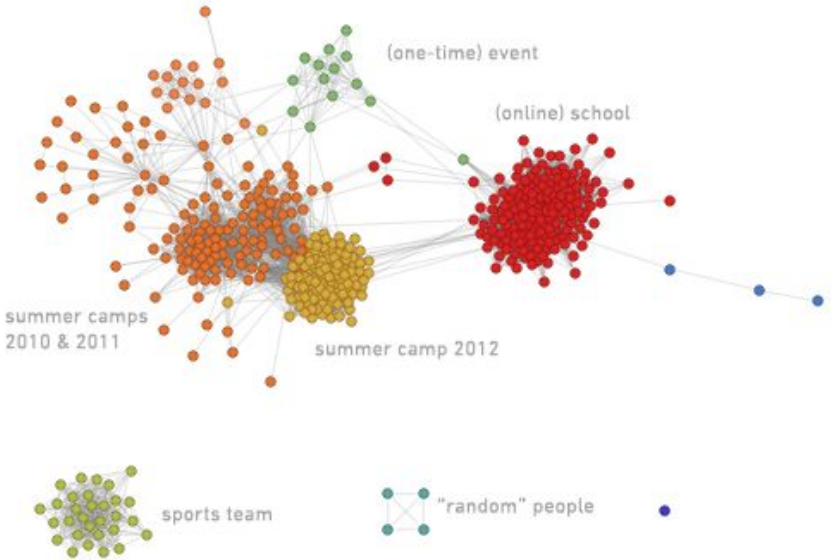
(a) Degree centrality

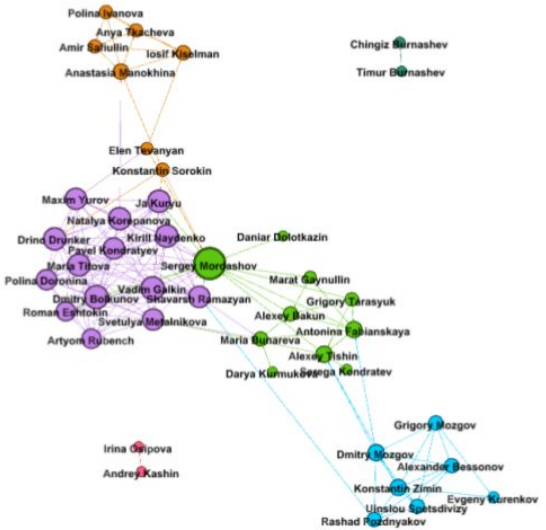


(b) Page Rank

Поиск сообществ

Эгоцентрическим графом социальной сети, называется граф в котором вершинам соответствуют чьи-то друзья, а ребра между ними возникают если они дружат между собой





Твиттер

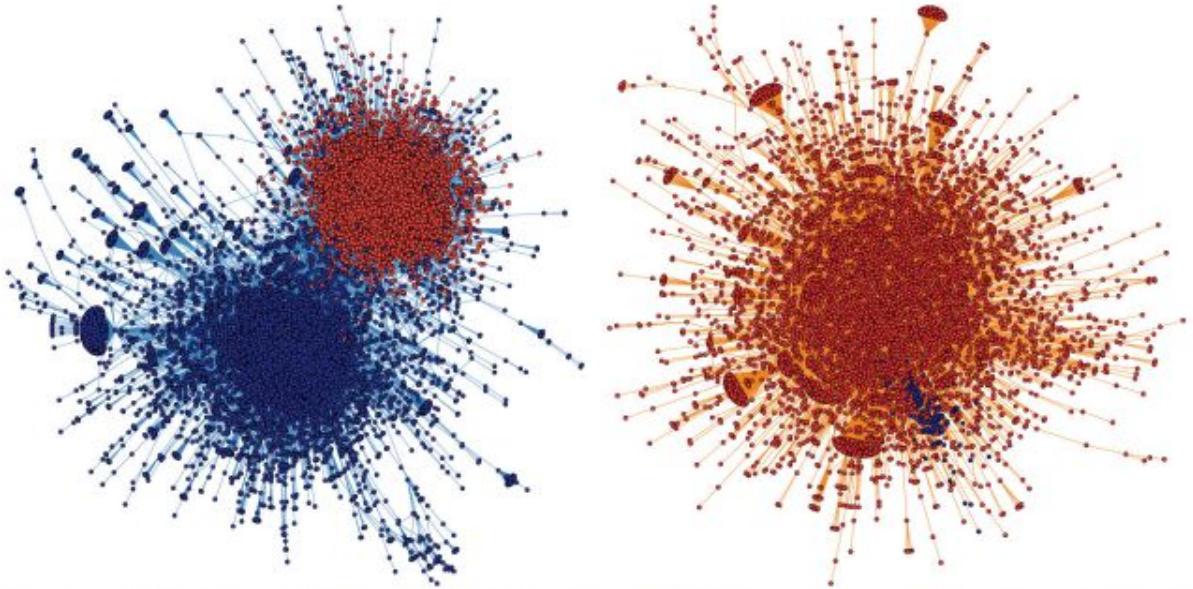


Figure 1: The political retweet (left) and mention (right) networks, laid out using a force-directed algorithm. Node colors reflect cluster assignments (see § 3.1). Community structure is evident in the retweet network, but less so in the mention network. We show in § 3.3 that in the retweet network, the red cluster A is made of 93% right-leaning users, while the blue cluster B is made of 80% left-leaning users.

Conover, Michael, et al. "Political polarization on twitter." *Icwsm* 133 (2011): 89-96.

Ранжирование

theorem

Harmonic centrality	Indegree	PageRank	Page views
0. Poincaré conjecture	0. Pythagorean theorem	0. Strahler number	0. Bayes' theorem
1. Pythagorean theorem	1. Gödel's incompleteness theorems	1. Pythagorean theorem	1. Pythagorean theorem
2. Fermat's Last Theorem	2. Fermat's Last Theorem	2. Fermat's Last Theorem	2. Euler's formula
3. Bell's theorem	3. Bayes' theorem	3. Gödel's incompleteness theorems	3. Central limit theorem
4. Reductio ad absurdum	4. Central limit theorem	4. Central limit theorem	4. Binomial theorem
5. Gödel's incompleteness theorems	5. Reductio ad absurdum	5. Bayes' theorem	5. Fermat's Last Theorem
6. Vafa–Witten theorem	6. Chain rule	6. Law of large numbers	6. De Morgan's laws
7. Binomial theorem	7. Law of large numbers	7. Euler's formula	7. Gödel's incompleteness theorems
8. Bayes' theorem	8. Triangle inequality	8. Reductio ad absurdum	8. Chain rule
9. Arrow's impossibility theorem	9. Binomial theorem	9. Chain rule	9. Euler's identity

http://wikirank-2018.di.unimi.it

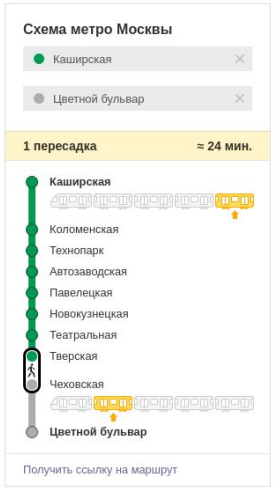
Ранжирование

human

Harmonic centrality	Indegree	PageRank	Page views
0. Barack Obama	0. Barack Obama	0. Barack Obama	0. Donald Trump
1. Napoleon	1. George W. Bush	1. Napoleon	1. Lali Espósito
2. Franklin D. Roosevelt	2. William Shakespeare	2. George W. Bush	2. Elizabeth II
3. George W. Bush	3. Jesus	3. Jesus	3. Dulce María
4. Ronald Reagan	4. Elizabeth II	4. Elizabeth II	4. Barack Obama
5. Donald Trump	5. Napoleon	5. William Shakespeare	5. Meghan Markle
6. Elizabeth II	6. Bill Clinton	6. Franklin D. Roosevelt	6. Elon Musk
7. Adolf Hitler	7. Adolf Hitler	7. Adolf Hitler	7. Adolf Hitler
8. William Shakespeare	8. Ronald Reagan	8. Bill Clinton	8. Ed Sheeran
9. Winston Churchill	9. Franklin D. Roosevelt	9. Stephan von Breuning (entomologist)	9. Queen Victoria

http://wikirank-2018.di.unimi.it

Поиск кратчайшего пути



Moscow metro map, https://metro.yandex.ru/

Модель малого мира

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; subpopulations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed.

Travers, Jeffrey, and Stanley Milgram. "An experimental study of the small world problem." *Social Networks*. 1977. 179-197.

Эксперимент Стэнли Милгрэма

HOW TO TAKE PART IN THIS STUDY

1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

2. DETACH ONE POSTCARD. FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.

3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

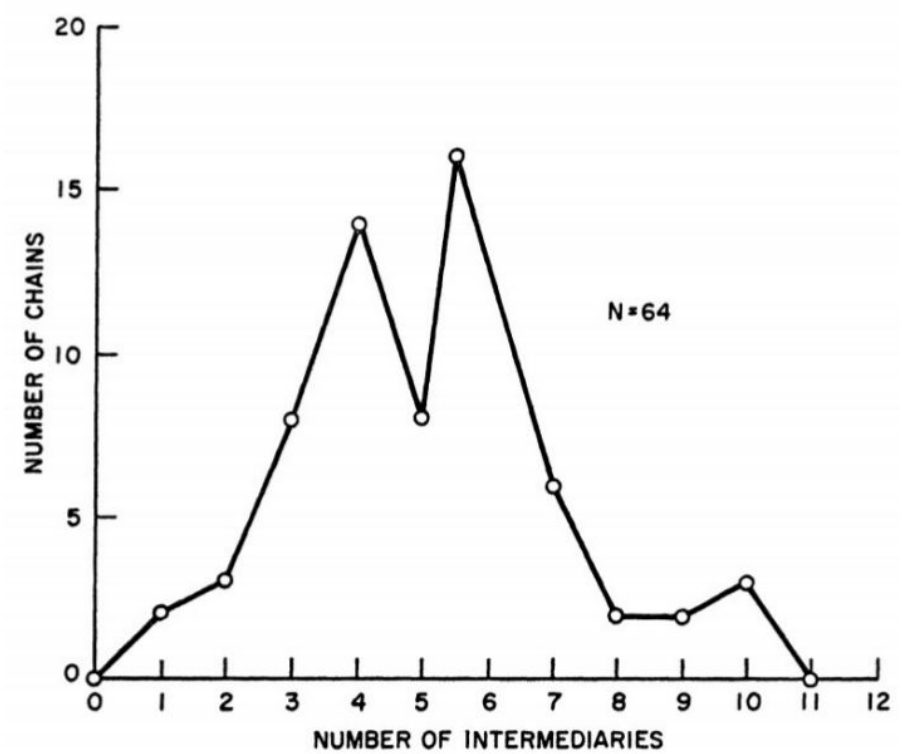
4. IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POSTCARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder

Результат: 6 degree of separation

- 296 испытуемых (217 разослали первое письмо)
- 196 в Небраске
- 100 в Бостоне
- Конечный получатель в Бостоне
- Имя, адрес, город рождения

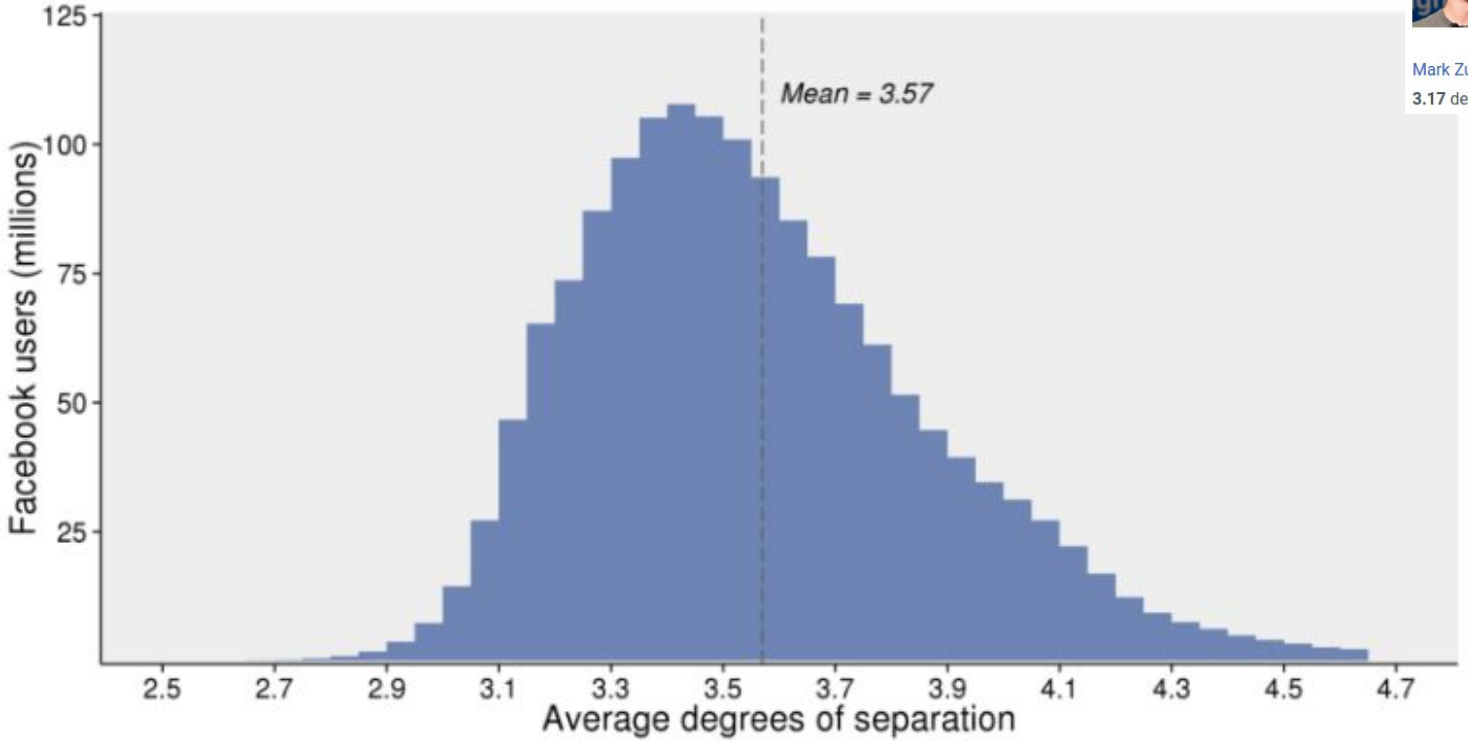
Результат: 6 degree of separation

- Достигли адресата 64 (29%)
- Средняя длина пути 5.2
- Из Бостона 4.4
- Из Небраски 5.7
- Имя, место работы, город рождения
- 25% went through the same last person, and 50% through one of three persons (“hubs”)



Тоже для Facebook

Mark Zuckerberg
3.17 degrees of separation



https://research.fb.com/three-and-a-half-degrees-of-separation/

Сетевой анализ данных

Какой то обобщающий слайд

Какого рода задачи помогает решать сетевой анализ данных. Какого рода вопросы можно задавать. Какие ответы можно получать.

- Алгоритмы
- Характеристики/Свойства
- Модели

Структура курса

План курса

1. Введение в сетевой анализ
2. Модели случайных графов (2)
3. Метрики центральности

4. Динамические сети
5. Распространение информации в сетях (2)

6. Поиск сообществ вершин (2-3)
7. Машинное обучение на графах (2)
8. Визуализация сетей

Информация о курсе

- Лекторы: Максим Панов, Анвар Курмуков
- Продолжительность: 14 недель
- e-mails: kurmukovai@gmail.com, panov.maxim@gmail.com
- github: https://github.com/kurmukovai/iitp_networks1

Пререквизиты

- Дискретная математика (базовые понятия теории графов)
- Алгебра (умножение матриц, собственные числа/векторы)
- Теория вероятности (функция распределения вероятности)
- Программирование (Python, Jupyter Notebook)

Формы контроля

- Домашние задания (60%)
- Проект (40%)
- Активность на занятиях (10%)

Формы контроля: Домашние задания

- 4 домашних задания
  - Модели случайных графов; метрики центральности (~25.09)
  - Динамические сети и распространение информации (~9.10)
  - Community detection (~30.10)
  - Машинное обучение на графах (~20.11)
- Вес каждого задания 15%
- Programming assignment + writing task
- Продолжительность выполнения 2 недели
- Выполняются индивидуально

Формы контроля: Проект

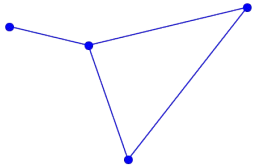
- Вес проекта 30%
- Выполняется в группах до 3 человек
- Темы проектов могут (и должны) быть предложены студентами и утверждены одним из лекторов
    - Анализ реальных данных
    - Теоретический проект (модель/задача + алгоритм + выводы)
- Утверждение проектов до 30.10

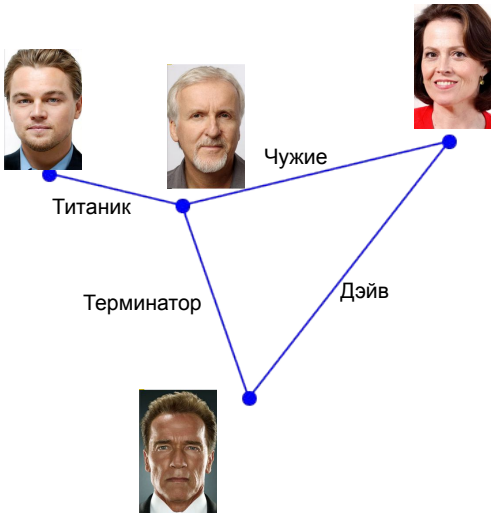
Введение

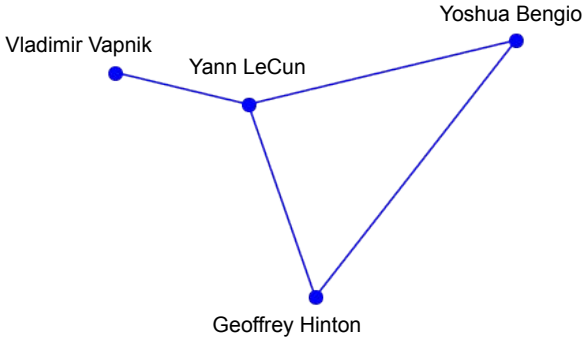
Network vs Graph

Граф - математический объект

Сеть - объект реального мира







Построение сетей

На основе одних и тех же данных могут быть построены различные сети.

На основе данных о публикациях работ можно построить:

- Сеть цитирования (**вершина** - автор, **ребро** - ссылка на другого автора)
- Сеть соавторств (**вершина** - автор, **ребро** возникает в случае наличия общей публикации)

На основе данных социальной сети можно построить:

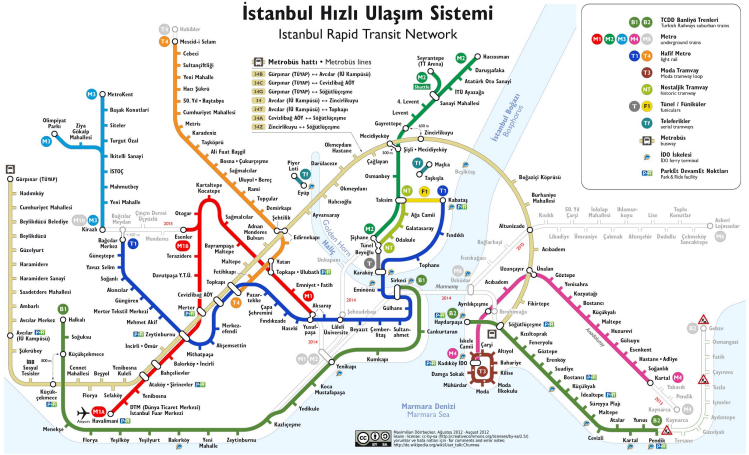
- Индивидуальные эгоцентрические сети
- Полный граф социальной сети
- Граф взаимодействия “групп”

На сколько осмысленно изучение сети **вершинами** которой являются публикации, а **ребра** возникают между публикациями с одинаковым словом в названии?

Верно ли тоже для аналогичной сети на основе постов в твиттере/реддите?

Построение сетей

Istanbul subway



London underground map

- Узловые станции Лондонского метро зачастую имеют только 1 платформу
- Переходы между узловыми станциями Стамбульского метро могут занимать 10-15 минут

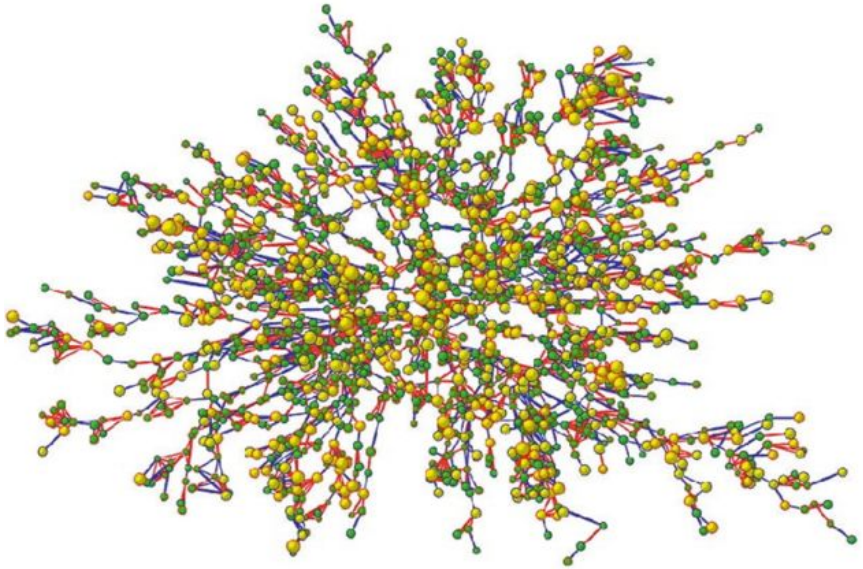
Метки на ребрах

- Вес
    - Расстояние (географическое, возраст)
    - Сходство (похожесть, корреляция)
- Категория (друг, родственник, коллега)
- Аналоги вершинных центральностей
- Знак (положительная/отрицательная корреляция)

Атрибуты на вершинах

- Пол
- Возраст
- Сексуальная ориентация
- Покупательская способность
- Наличие семьи
- Проблемы со здоровьем

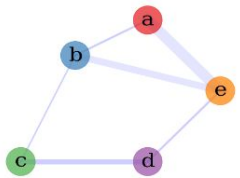
Basically anything!



Christakis, Nicholas A., and James H. Fowler. "The spread of obesity in a large social network over 32 years." *New England journal of medicine* 357.4 (2007): 370-379.

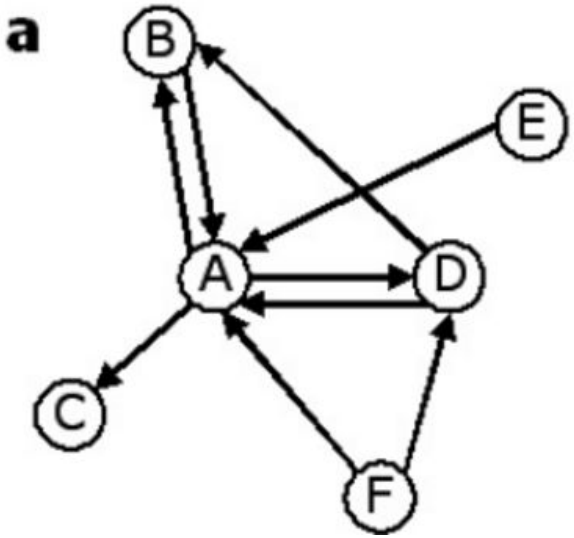
Представление графов

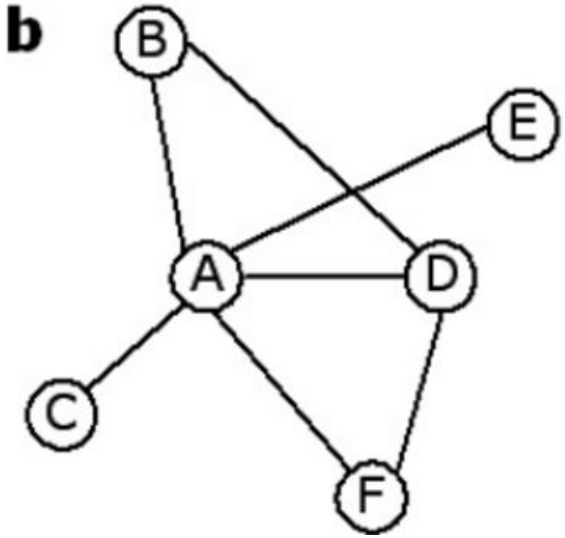
- Матрица смежности (|V||V|)
- Матрица инцидентности (|V||E|)
- Список ребер (|E|)
- Список смежности (|V| + |E|)



	a	b	c	d	e
a	0	**6**	**0**	**0**	**30**
b	6	0	**4**	**0**	**20**
c	0	4	0	**15**	**0**
d	0	0	15	0	**6**
e	30	20	0	6	0

Directed vs Undirected





Направленные:

- Цитирование
- Подписки в социальной сети

Ненаправленные:

- Друзья в социальной сети
- Соавторы

Что еще?

- Связность
- Сильная связность
- Компоненты связности
- Разреженные графы (sparse)
- Степень вершины
- Полный граф
- Двудольный граф

Что установить

- Python (Jupyter)
    - Networkx
    - Igraph
    - Pytorch (November)
- SNAP (optional) http://snap.stanford.edu/
- Gephi (GUI for visualization, optional) https://gephi.org/