Готовые инструменты для анализа данных

Семинар 6

BootCaT

- Поиск текстов по ключевым словам
- Что-то похожее на «Let me google that for you»
- Сохраняет поисковую выдачу и тексты страниц в файлы

Voyant Tools

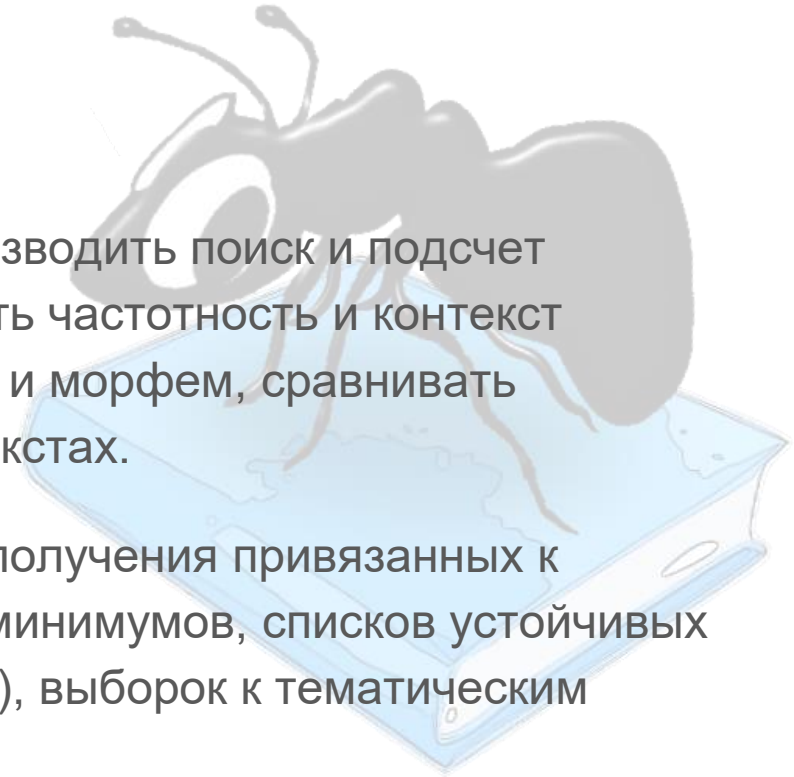
- https://voyant-tools.org
- https://github.com/lilaspourpre/corpus
- Облака тегов
- Встречаемость слов на отрезках текста
- Построение диаграмм
- Частотность
- Фразы, корреляции, контекст
- Возможность загружать свои корпуса

AntConc

http://www.laurenceanthony.net/software.html

С помощью данной программы можно производить поиск и подсчет различных элементов текста, анализировать частотность и контекст употребления словоформ, словосочетаний и морфем, сравнивать употребительность словоформ в разных текстах.

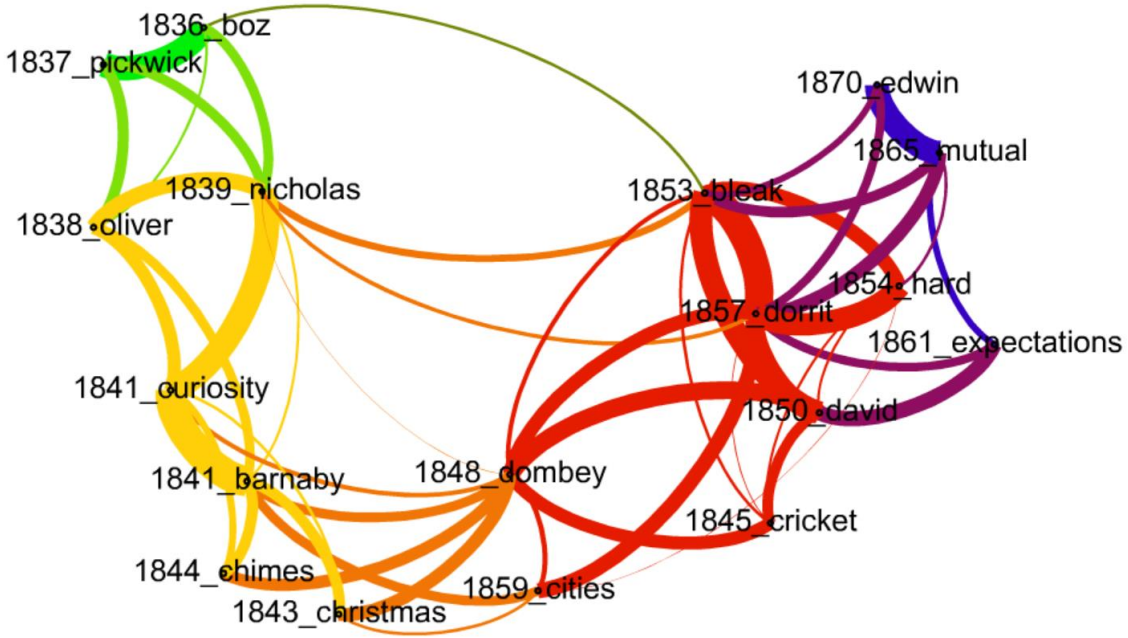
Программа может быть использована для получения привязанных к заданной предметной области словарных минимумов, списков устойчивых сочетаний (в том числе терминологических), выборок к тематическим группам слов.



Gephi

Gephi – программа для визуализации графов.
С помощью Gephi можно делать очень красивые и наглядные картинки.

Пример из работы польского специалиста по стилометрии Яна Рыбицки – хронология романов Ч. Диккенса, построенная по наиболее частотным словам в тексте.



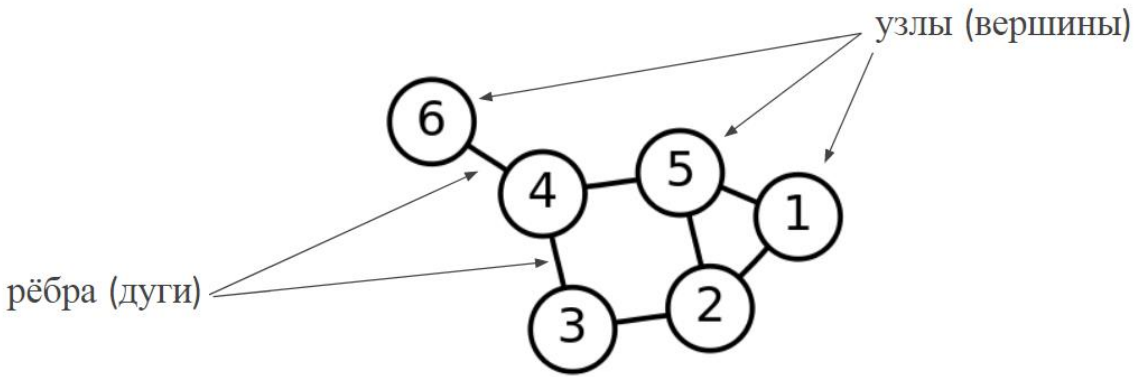
- Скачать
- Инструкция по работе
- Помощь в установке
- Работать онлайн: rollapp.com/app/gephi

Графы и социальные сети

Семинар 6

Что такое граф?

Граф, или сеть – это модель, состоящая из узлов и связей между ними, или **вершин** и **ребер**.
По-английски это называется nodes (vertices) и edges.



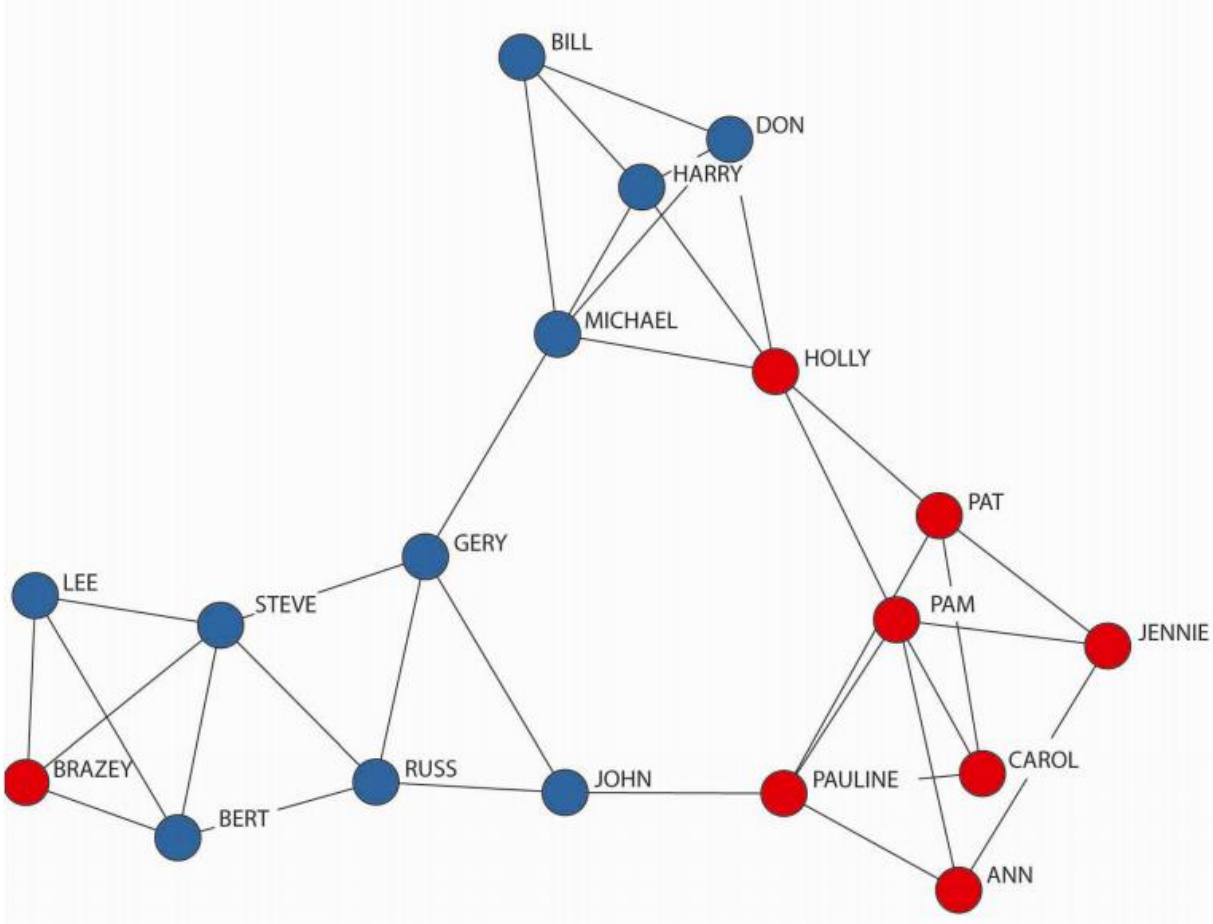
Что такое граф?

Узлы в графах могут группироваться в сообщества.
Сообщество – это плотный подграф, где все (или почти все) узлы связаны между собой.

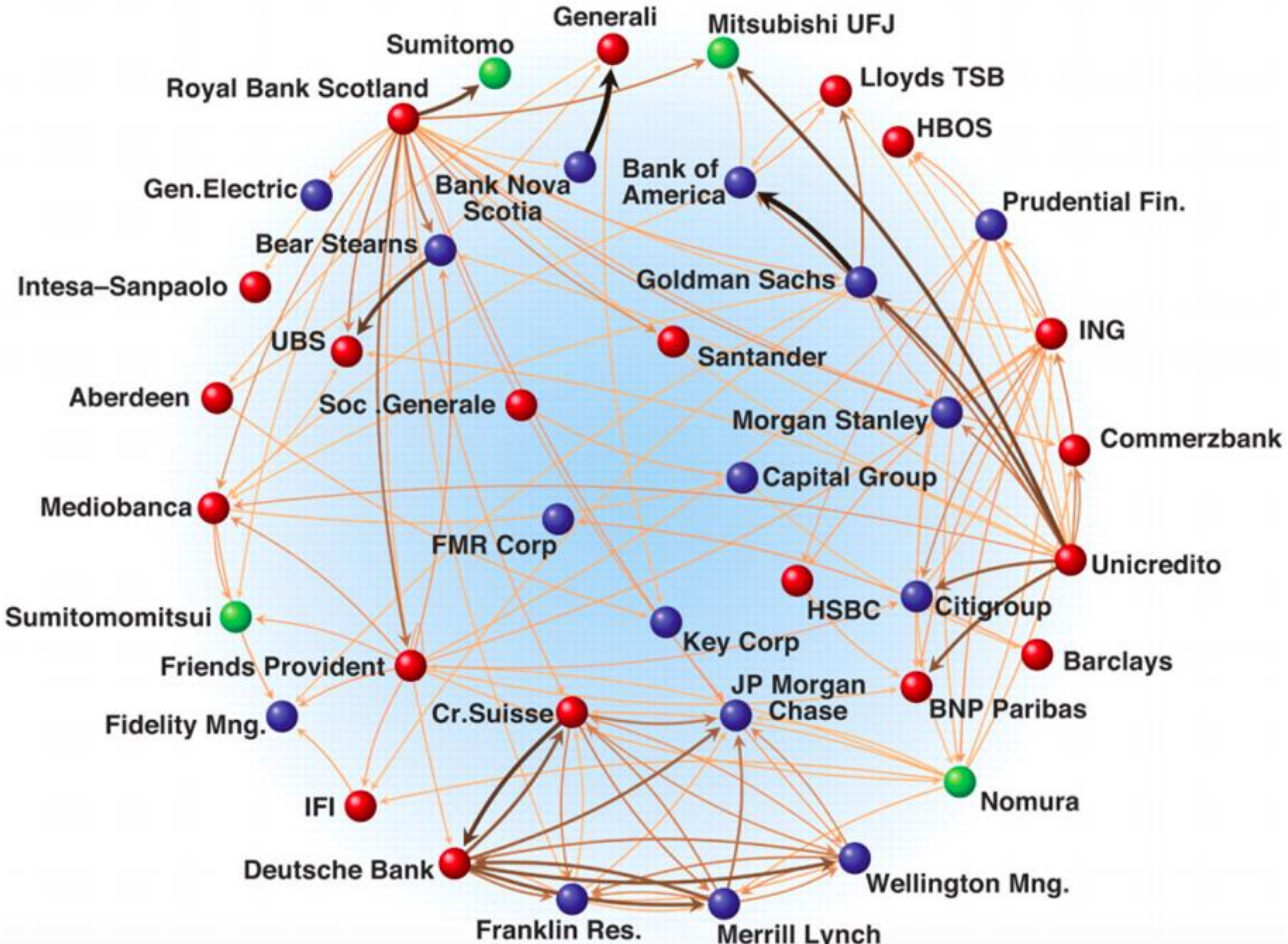
Графы бывают:

- **ориентированные и неориентированные** (связи-стрелочки vs обычные связи): можно пойти только в определенную сторону/обе
- **связные и несвязные** (все узлы связаны vs есть узлы, которые оторваны от основного графа):
- **взвешенные и невзвешенные** (связи имеют разное числовое значение или имеют одинаковое)

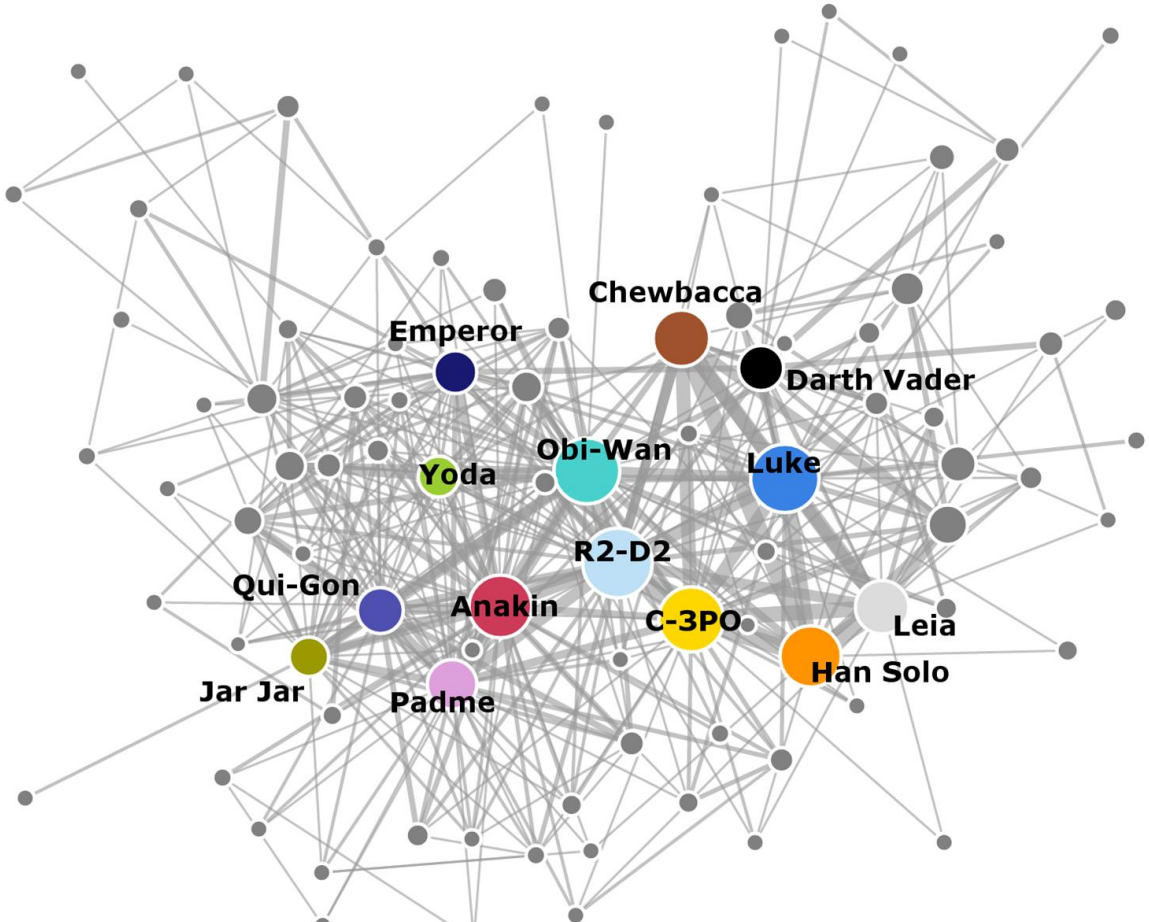
Неориентированный граф

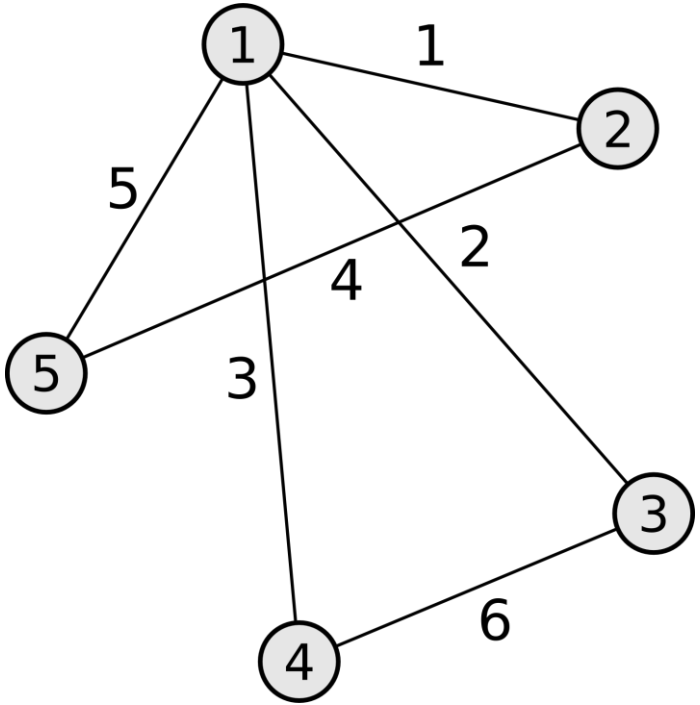


Ориентированный граф

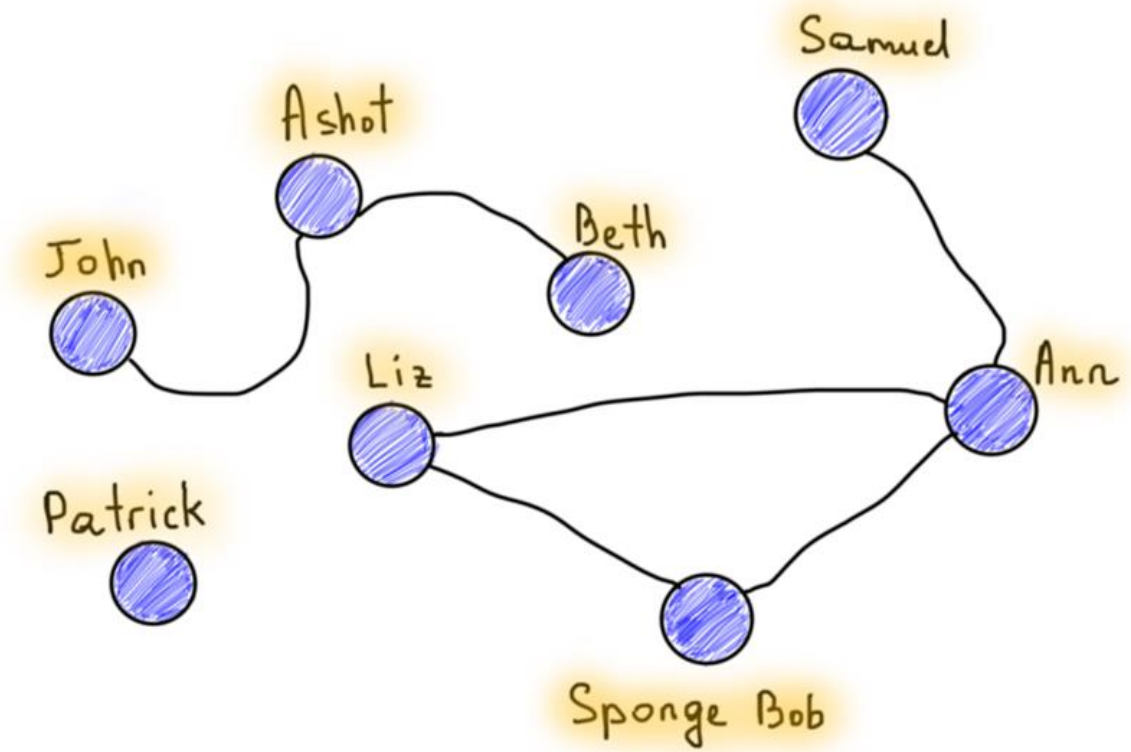


Взвешенные графы

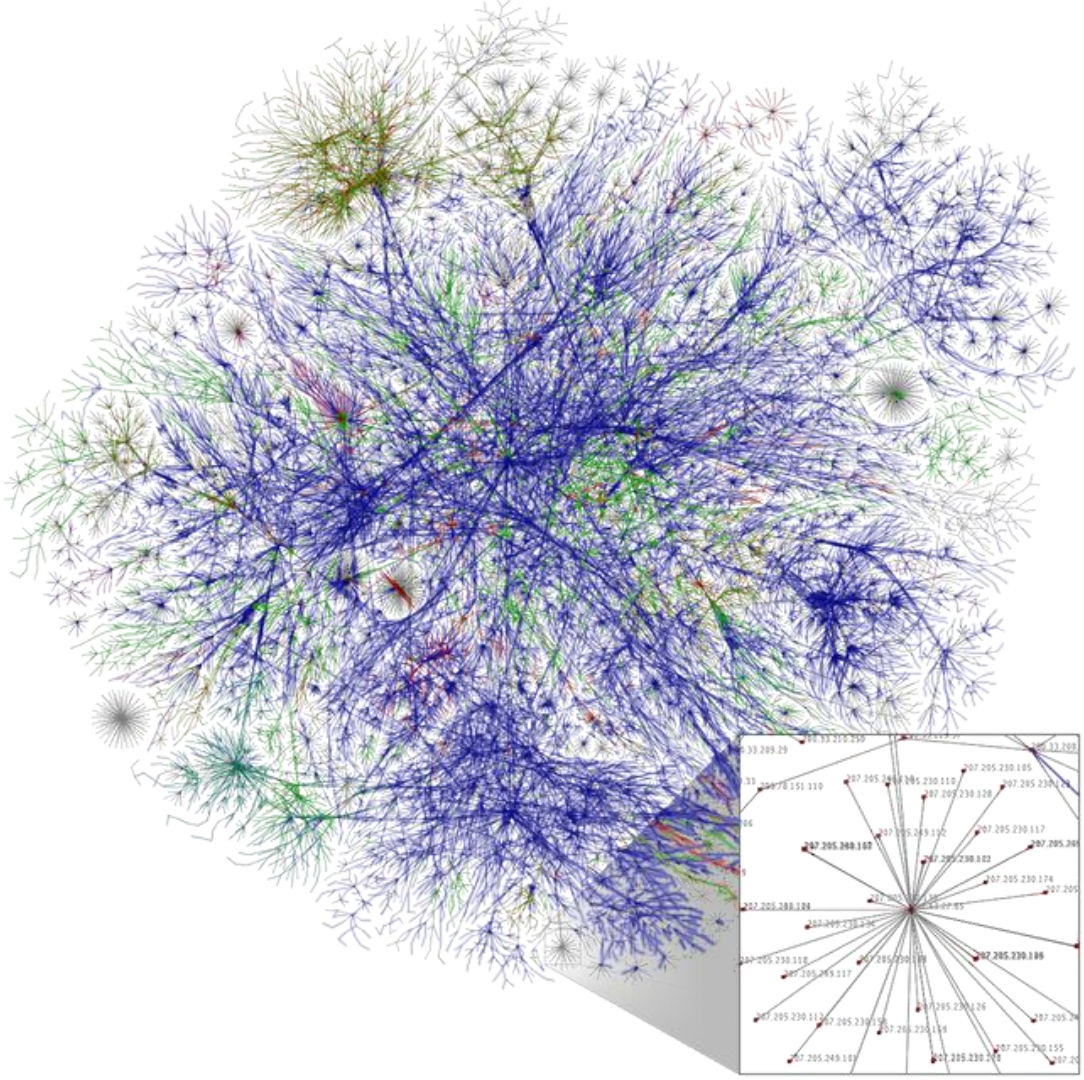




Несвязный граф




Планерная
Сходненская
Тушинская
Спартак
Щукинская
Стрешнево
Панфиловская
Октябрьское Поле
Зорге
Полежаевская
Хорошёво
Шелепиха
Ховрино
Речной вокзал
Водный стадион
Войковская
Балтийская
Сокол
Аэропорт
Динамо
Петровский парк
ЦСКА
Хорошёвская
Селигерская
Верхние Лихоборы
Окружная
Лихоборы
Коптево
Алтуфьево
Бибирево
Отрадное
Владыкино
Петровско-Разумовская
Тимирязевская
Дмитровская
Савёловская
Менделеевская
Фонвизинская
Бутырская
Марьина Роща
Достоевская
Медведково
Бабушкинская
Свиблово
Ботанический сад
ВДНХ
Алексеевская
Рижская
Проспект Мира
Ростокино
Белокаменная
Бульвар Рокоссовского
Черкизовская
Локомотив
Преображенская площадь
Сокольники
Красносельская
Комсомольская
Красные Ворота
Щёлковская
Первомайская
Измайловская
Измайлово
Партизанская
Семёновская
Соколиная гора
Электрозаводская
Бауманская
Новокосино
Новогиреево
Перово
Шоссе Энтузиастов
Авиамоторная
Андроновка
Нижегородская
Новохохловская
Угрешская
Пятницкое шоссе
Митино
Волоколамская
Мякинино
Строгино
Крылатское
Молодёжная
Кунцевская
Славянский бульвар
Пионерская
Филёвский парк
Багратионовская
Фили
Международная
Деловой центр
Выставочная
Киевская
Студенческая
Кутузовская
Парк Победы
Минская
Ломоносовский проспект
Раменки
Новослободская
Белорусская
Цветной бульвар
Трубная
Сухаревская
Тургеневская
Чистые пруды
Сретенский бульвар
Краснопресненская
Баррикадная
Улица 1905 года
Беговая
Маяковская
Тверская
Пушкинская
Чеховская
Кузнецкий Мост
Лубянка
Смоленская
Арбатская
Александровский сад
Боровицкая
Библиотека имени Ленина
Охотный Ряд
Театральная
Площадь Революции
Китай-город
Курская
Чкаловская
Римская
Площадь Ильича
Марксистская
Таганская
Третьяковская
Новокузнецкая
Кропоткинская
Парк культуры
Фрунзенская
Октябрьская
Полянка
Добрынинская
Серпуховская
Павелецкая
Крестьянская Застава
Пролетарская
Волгоградский проспект
Текстильщики
Кузьминки
Рязанский проспект
Выхино
Лермонтовский проспект
Жулебино
Котельники
Дубровка
Лужники
Спортивная
Воробьёвы горы
Университет
Проспект Вернадского
Юго-Западная
Тропарёво
Румянцево
Саларьево
Шаболовская
Тульская
Площадь Гагарина
Ленинский проспект
Крымская
Верхние Котлы
ЗИЛ
Автозаводская
Технопарк
Коломенская
Каширская
Варшавская
Академическая
Профсоюзная
Новые Черёмушки
Калужская
Беляево
Коньково
Тёплый Стан
Ясенево
Новоясеневская
Битцевский парк
Нагатинская
Нагорная
Нахимовский проспект
Каховская
Севастопольская
Чертановская
Южная
Пражская
Улица Академика Янгеля
Аннино
Кантемировская
Царицыно
Орехово
Домодедовская
Красногвардейская
Алма-Атинская
Кожуховская
Печатники
Волжская
Люблино
Братиславская
Марьино
Борисово
Шипиловская
Зябликово
Лесопарковая
Бульвар Дмитрия Донского
Улица Старокачаловская
Улица Скобелевская
Бульвар Адмирала Ушакова
Улица Горчакова
Бунинская аллея



Метрики

Метрика – это результат измерений, проведенных определенным способом.

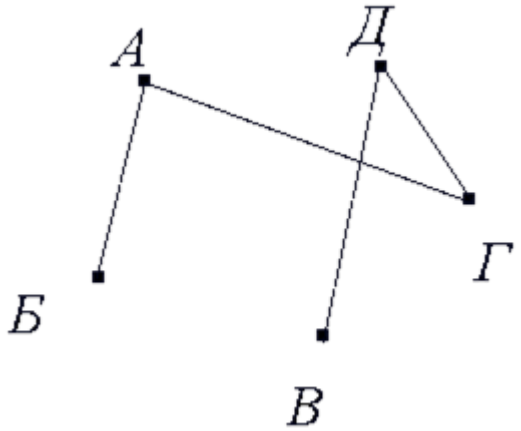
Представьте, что вы выбираете **материал для реферата** по “Философским крохам” Кьеркегора. У вас есть **оригинальный текст в 50 страниц**, **современная книга** “Cumulative Index to Kierkegaard's Writings: The Works of Søren Kierkegaard” H.J. Hong и **статья** “Идея спасения в псевдонимных произведениях С. Керкегора. Очерк первый. Лестница Йоханнеса Климакуса” Д.А. Лунгиной **в 100 страниц.**

Если метрикой для вас является **количество страниц**, то вы выберете оригинальный текст, а если **простота чтения** – то статью.

Метрики

- **Степень, или мощность узла (degree)** – это количество его связей.

- **Взвешенная степень (weighed degree)** – это количество связей узла, разделенное на общее количество связей в графе.

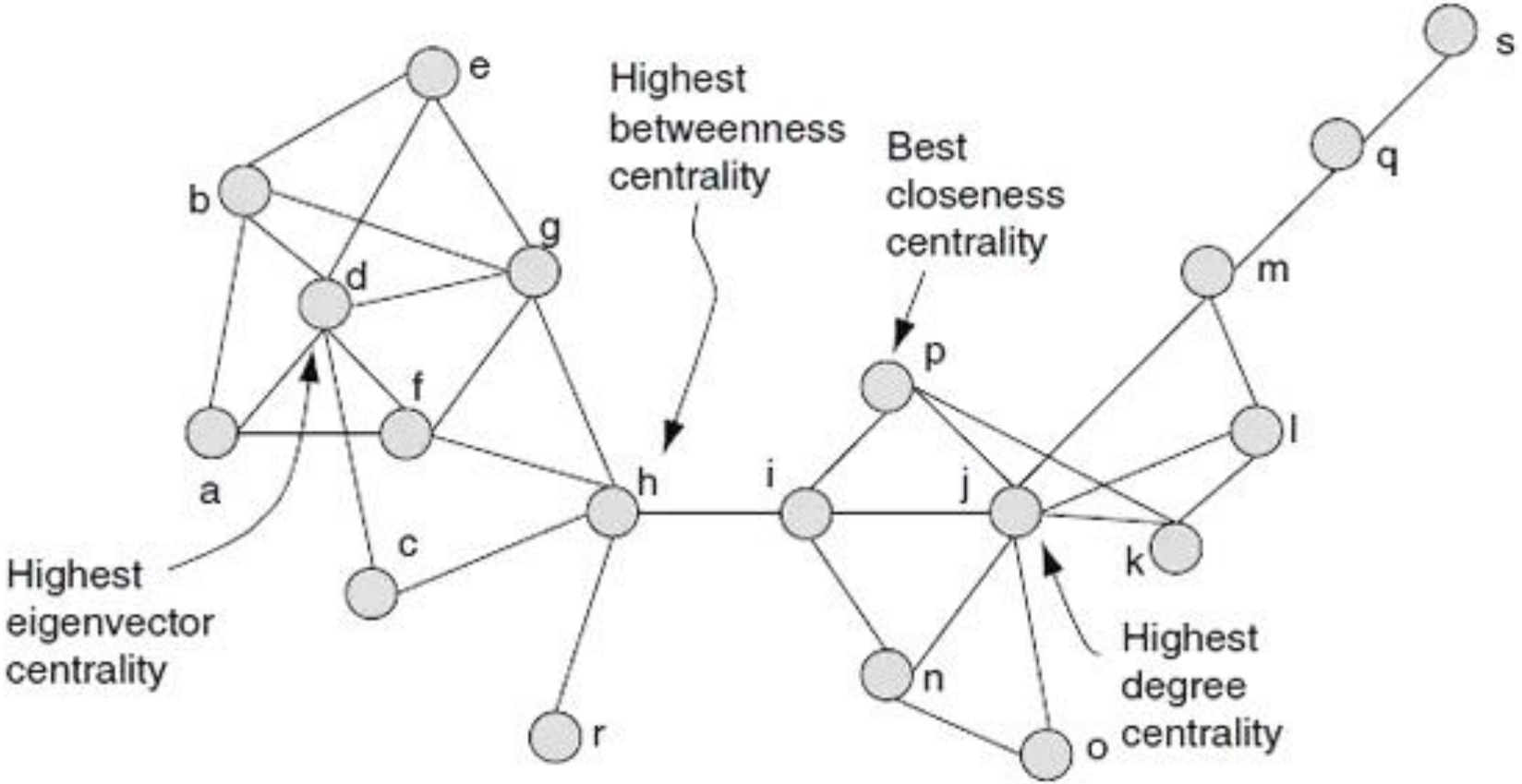


Метрики

Важность узла можно определять разными способами:

- **degree centrality:** у кого больше связей, тот и важнее
- **closeness centrality:** чем центральнее узел (т.е. чем короче путь от него до всех остальных узлов), тем он важнее
- **betweenness centrality:** количество кратчайших путей, проходящих через узел
- **eigencentrality:** чем больше друзей у твоих друзей, тем ты важнее

Метрики графа



Коэффициент ассортативности (assortativity coefficient) определяет, с кем связаны "важные" узлы: если с другими "важными" узлами, то значение коэффициента высокое, а если нет – низкое.

Коэффициент кластеризации (clustering coefficient) – степень взаимодействия между собой ближайших соседей узла, т.е. вероятность того, что ближайшие соседи узла будут связаны не только с ним, но и между собой.

Плотность графа (density) – отношение числа ребер к максимально возможному. В сообществах высокий коэффициент кластеризации и высокая плотность.

Модулярность (modularity) показывает, насколько при заданном разбиении графа на группы плотность связей внутри группы больше плотности связей между группами. С помощью этой метрики граф разбивается на сообщества.

Форматы графов

Граф записывается в виде текстового (.gml) или XML-файла (.graphml, .gexf), где перечисляются все узлы, ребра и их атрибуты – например, название узла или вес ребра.

```
node
[
  id 76
  label "MmeHucheloup"
]
edge
[
  source 1
  target 0
  value 1
]
```

```
    <node id="388894866" label="Катерина Пикъ">
        <attvalues>
            <attvalue for="0" value="Катерина" />
            <attvalue for="1" value="Пикъ" />
            <attvalue for="2" value="" />
            <attvalue for="3" value="turclubpik" />
            <attvalue for="4" value="1" />
            <attvalue for="5" value="https://pp.user
            <attvalue for="6" value="0" />
            <attvalue for="7" value="1" />
            <attvalue for="8" value="2" />
        </attvalues>
    </node>
</nodes>
<edges>
    <edge id="1" source="1257" target="2177279" />
    <edge id="2" source="1257" target="7919359" />
    <edge id="3" source="1257" target="13075598" />
    <edge id="4" source="1257" target="13840887" />
    <edge id="5" source="1257" target="14691867" />
    <edge id="6" source="1257" target="22583724" />
    <edge id="7" source="5284" target="7586" />
    <edge id="8" source="5284" target="70108" />
```