Анализ графа википедии

Захватаев Михаил

Составленные графы и их структура

1. Удалось создать и сохранить графы от 500 до 50000 узлов с шагом в 500 узлов.
2. При обработке графа всего в 5000 на локальной машине или в гугл колабе - многие алгоритмы выполнялись очень долго и для основного анализа был взят граф, состоящий 2000 узлов.
3. Граф направленный и каждым узлом является ссылка на страницу википедии, за исключением ссылок, содержащих слова: “шаблон”, “категория”.
4. Граф в 2000 узлов содержит 8195 ребер, ненаправленный граф с теми же вершинами содержит 6321 ребро.
5. Поля узла графа: все ссылки на странице с доменом “ru.wiki”, заголовок статьи

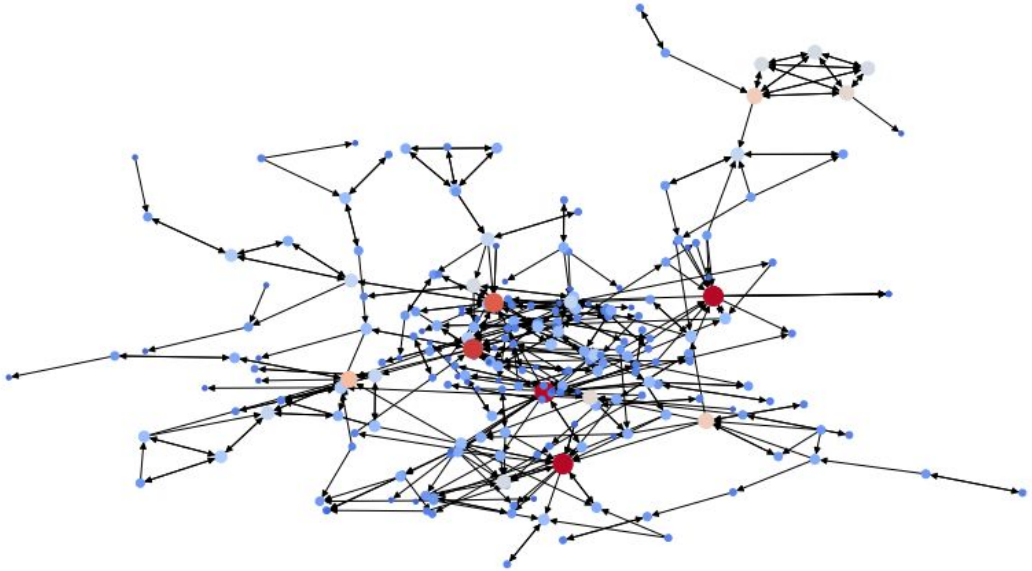
Алгоритм построения графа

- Select some start page of wikipedia (ex. https://ru.wikipedia.org/wiki/MapReduce)
- Add node to Graph
- Add current link to set of busy (added to graph) links
- Add links from page with substring "ru.wikipedia" to set of free links and to node attribute "neighbor_links"
- free links := (free links + links from current page) \ (busy links)
- Add directed edges for current node and already existed nodes if they have a connection with current node
- Randomly select page from set of free links
- Repeat 1.2 from second point N-1 times, N - num of nodes of generating graph

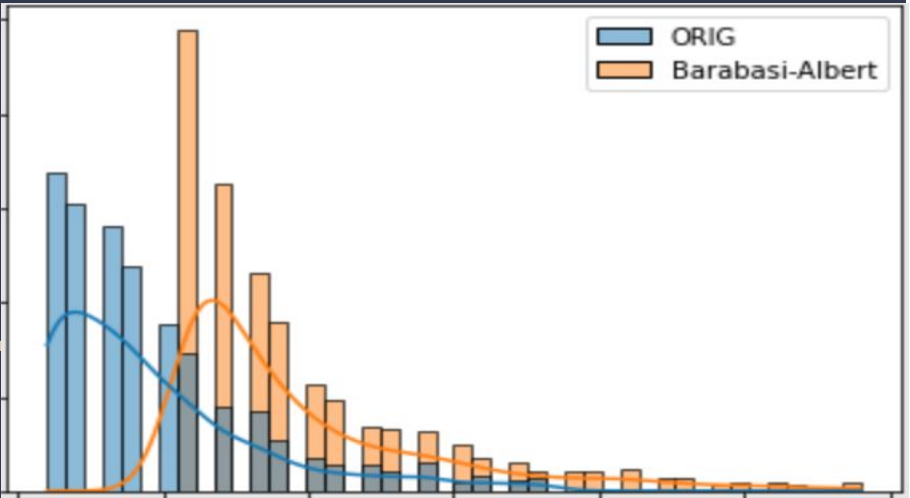
Таким алгоритмом первые 50 тыс. страниц собирались со скоростью от 0.2 сек до 3 сек на добавление одной страницы в граф и нахождение всех связей с уже добавленными

Визуализация подграфа

На визуализации подграфа (400 узлов) уже видно, что в сети появляются некоторые кластеры (пятиугольник-клика справа сверху) и более значимые чем все остальные страницы (чем краснее, тем больше ссылок на страницу), на которые ссылается достаточно большое количество источников.



Структурный анализ графа: сравнение со случайными графами



Характеристики графа

- Радиус - 6
- Диаметр - 11
- Средний коэффициент кластеризации - 0.33
- Средний наикратчайший путь - 3.92
- Одна связная компонента

По коэффициенту кластеризации граф ближе всего к модели случайного графа Эрдёша-Реньи.

По остальным параметрам граф ближе всего к модели случайного графа Барабаши-Альберт

Лидеры центральных мер

Почти во всех центральных мерах лидером является VIAF - Википедия. В лидеры (топ 5) попадают достаточно общие понятия, названия стран и даты: Столица — Википедия, 1988 год — Википедия, Россия — Википедия.

На фото топ 5 узлов по степени посредничества (betweenness centrality)

Корреляции мер

На матрице корреляций видим, что центральность узлов по Кацу очень сильно коррелирует со степенью близости (closeness), со степенью посредничества (betweenness). Это значит, что многие узлы в нашем графе выполняют несколько ролей одновременно.

	Closeness	Betwenness	Katz	Eigen
Closeness	1.000000	0.348726	0.541646	0.141816
Betwenness	0.348726	1.000000	0.768714	0.067546
Katz	0.541646	0.768714	1.000000	0.327918
Eigen	0.141816	0.067546	0.327918	1.000000

PageRank & HITS

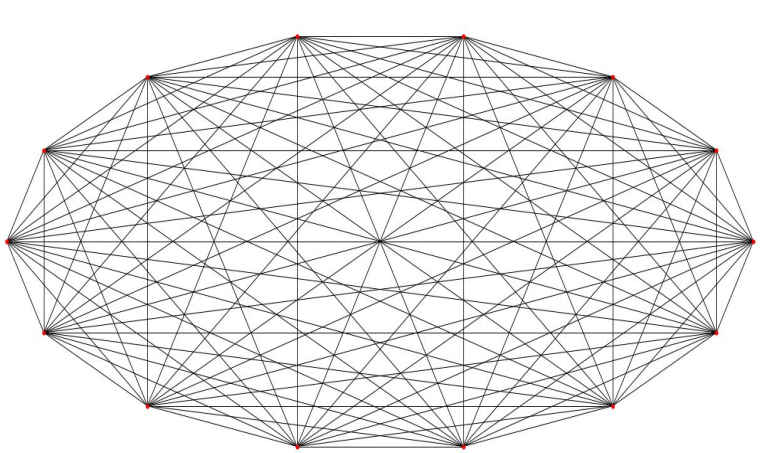
```
#1: Россия — Википедия
PageRank score: 0.0037566722217781076
https://ru.wikipedia.org/wiki/%D0%A0%D0%BE%D1%81%D1%81%D0%B8%D0%B9%D1%
81%D0%BA%D0%B0%D1%8F_%D0%A4%D0%B5%D0%B4%D0%B5%D1%80%D0%B0%D1%86%D0%B8%
D1%8F
--------------------------------------------------
#2: Союз Советских Социалистических Республик — Википедия
PageRank score: 0.0033131384944123034
https://ru.wikipedia.org/wiki/%D0%A1%D0%BE%D0%B2%D0%B5%D1%82%D1%81%D0%
BA%D0%B8%D0%B9_%D0%A1%D0%BE%D1%8E%D0%B7
--------------------------------------------------
#3: Германия — Википедия
PageRank score: 0.0032147745495289463
https://ru.wikipedia.org/wiki/%D0%A4%D0%A0%D0%93
--------------------------------------------------
#4: Казахстан — Википедия
PageRank score: 0.0032146159218564754
https://ru.wikipedia.org/wiki/%D0%9A%D0%B0%D0%B7%D0%B0%D1%85%D1%81%D1%
82%D0%B0%D0%BD
--------------------------------------------------
#5: Иран — Википедия
PageRank score: 0.0031953995074056655
https://ru.wikipedia.org/wiki/%D0%9F%D0%B5%D1%80%D1%81%D0%B8%D1%8F
--------------------------------------------------
```

- PageRank выдал ссылку на VIAF, Чешскую и Лихтенштейнскую библиотеки, Варшаву и Казахстан
- HITS (hubs) - Россия, СССР, Германия, Казахстан и Иран. Это большие статьи, в которых содержится много ссылок, поэтому неудивительно, что они заняли весь топ хабов
- HITS (authorities) - VIAF, Чешская библиотека, Казахстан, Столица, Первый канал. Больше похоже на результаты PageRank.

Казахстан попал во все топы!

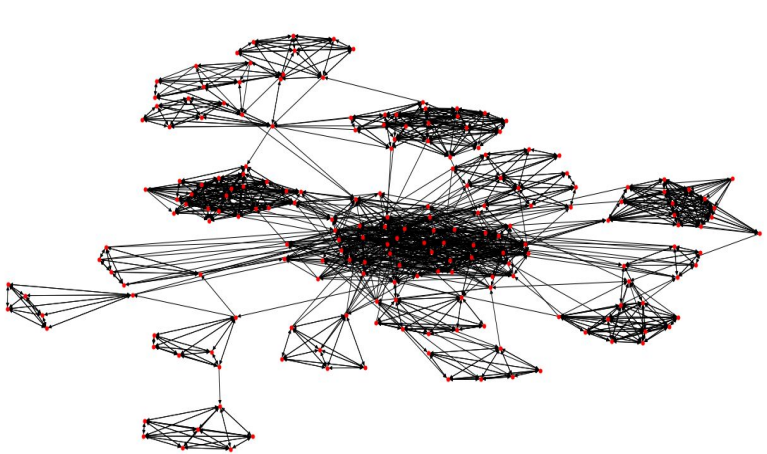
Ищем максимальную клику

Максимальная клика состоит из 14 узлов и, по всей видимости, посвящена теме колоний



Барбуда
Джеймс (остров, Гамбия)
Южная Каролина (провинция)
Мьянма
Королевство Раротонга
Малайская Федерация
Джохор
Новые Гебриды
Канадская конфедерация
Доминион
Колониальная история США
Британские Соломоновы острова
Колония и протекторат Нигерия
Уолфиш-Бей

k core (k=10)

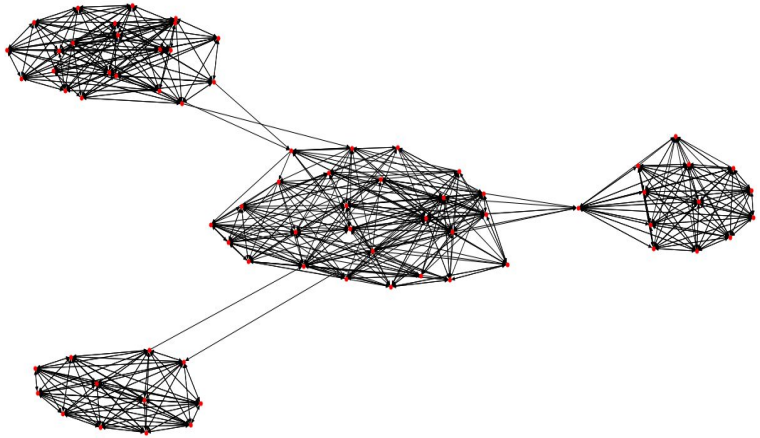


При k=10 хорошо видна структура графа.

Из-за алгоритма добавления новых ссылок в граф у нас есть некоторое смещение в сторону начальной страницы, которая образует вокруг себя более плотный кластер.

Тем не менее видно, что появляются уверенные ответвления и достаточно плотные кластеры уже очень далеко от центра.

k core (k=15)



При k=15 видим 4 самых плотных кластера. При k=20 остается только центральный кластер и гипотеза о том, что этот кластер вырос вокруг стартовой страницы (PageRank) - подтверждается.

Ядро данного кластера составляют следующие страницы: PageRank, Google Hangouts, Google Планета Земля, Google Ads, История Google, Список поглощений Google, Knol, Picnik, Мастер сообществ Google, Google Now, Google Analytics

Видим, что основа данного кластера - страницы, связанные с компанией Google

Спасибо за внимание!

Google кластер на графе википедии от самой википедии :)