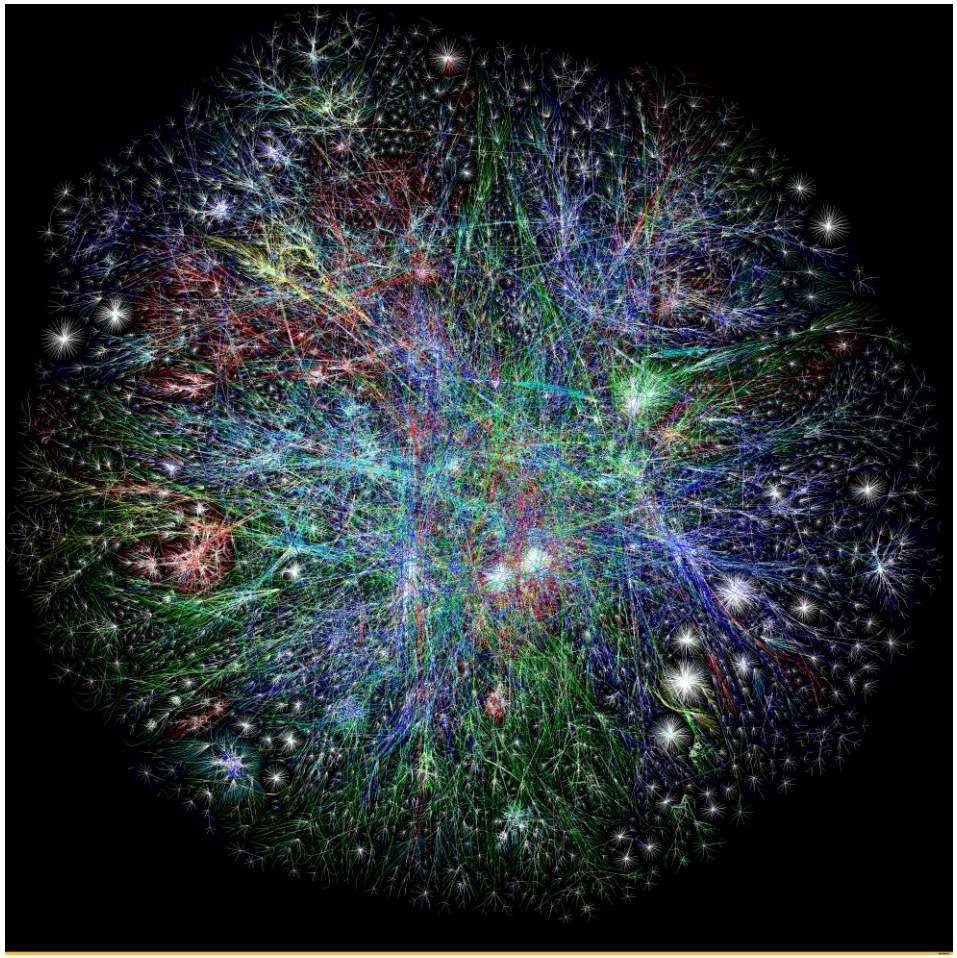
Граф интернета

Граф интернета



Граф интернета

Схема интернета с сайта internet-map.net

Граф интернета (ориентированный мультиграф)

- Вершины (V) – сайты в интернете
- Дуги (E) – ссылки между сайтами

- Экспериментальные свойства графа интернета:
  - закон «шести рукопожатий»
  - исключительная «разреженность»
    (если вершин у веб-графа n, то ребер у него не более mn с некоторым постоянным $m \geq 1$)
  - *степенным законом распределения* степеней вершин: $\#(n,d) = |\{v: \deg(v) = d\}| / n \approx c / d^{2.3}$

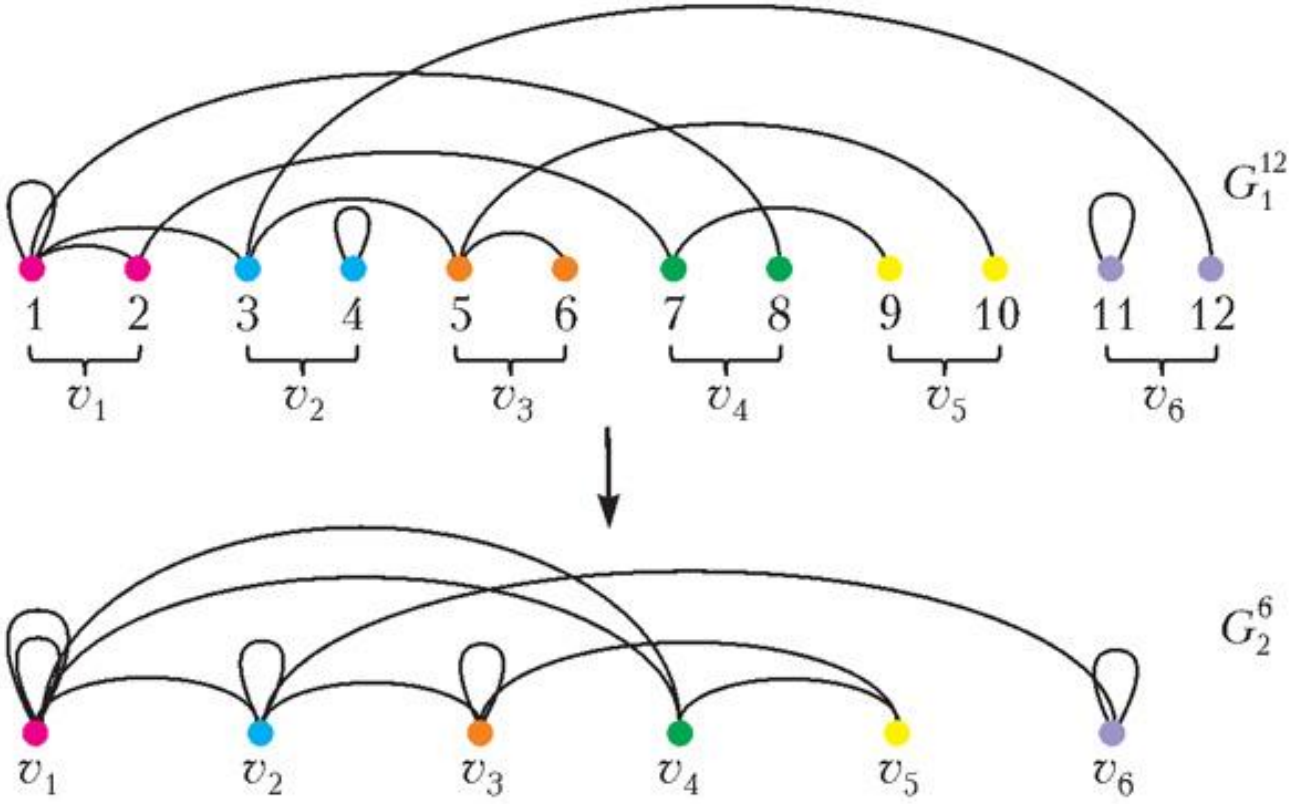
Модель веб-графа

- Зная свойства «настоящего» интернета, можно создавать модели его разного размера.
  - тестировать алгоритмы обхода интернета роботами
  - выявлять спамовую активность

Случайные веб-графы

- 1999, А. Л. Барабаши и Р. Альберт
  - **Идея!** Когда появляется новый сайт, он, скорее всего, «предпочитает» сослаться на те сайты, которые и без того уже многими цитированы. Более точно, вероятность, с которой новый сайт ставит ссылку на сайт-предшественник, пропорциональна (входящей) степени вершины веб-графа, отвечающей этому сайту.
- 2000, модель Боллобаша–Риордана
  - желание избавиться от очевидных проблем первой модели

модель Боллобаша–Риордана



Построим последовательность графов G(n, 1), а из них G(n, m).

модель Боллобаша–Риордана. 1 этап

- *G(1,1)* — это граф с одной вершиной и одной петлей (1, 1).
- Граф *G(n,1)* получим путем добавления к графу *G(n-1,1)* одной вершины (одного сайта) с «именем» *n* и одного ребра
  - с вероятностью *1/(2n-1)* ссылка из *n* пойдет на само *n*
  - с вероятностью *deg(v)/(2n-1)* сайт *n* процитирует сайт v

модель Боллобаша–Риордана. 2 этап

- Зафиксируем натуральное $m \geq 2$. Рассмотрим найденный на первом этапе граф *G(nm,1)*.
- Обозначим v_1 группу из первых *m* его вершин, т.е. множество {1, ..., *m*}; v_2 следующую группу его вершин {*m* + 1, ..., 2*m*} …
- Схлопнем вершины из *G(nm,1)* в своего рода «метасайты», а все прежние ссылки сохраняем.

Свойства модели

- при $m \geq 2$ и при любом $\varepsilon > 0$ с увеличением числа вершин графа G(n,m) всё ближе к единице становится вероятность того, что диаметр графа G(n,m) заключен в пределах

  $$(1 - \varepsilon)\ln(n)/\ln(\ln(n)) \ldots (1 + \varepsilon)\ln(n)/\ln(\ln(n))$$

- Е. Гречников (2011) доказал, что для всех m и d с ростом n всё ближе к единице становится вероятность того, что величина # (n, d), определенная на графе $G(n,m)$, практически не отличается от величины c/d^3, где c зависит лишь от m.

Уточнение модели

- П. Бакли и Д. Остгуса
  - возьмем произвольное число $a > 0$
  - $1/(2n-1) \rightarrow a/((a+1)n-1)$
  - $\deg(v) / (2n-1) \rightarrow (\deg(v) + a - 1) / ((a+1)n - 1)$

  * при a=1 имеем предыдущую модель
  - величина # (*n, d*), определенная на новой последовательности графов, становится почти наверняка приближенно равной c/d^{2+a}

Модель копирования

- Идея! когда появляется новый сайт, он либо цитирует какого-то «случайного» предшественника, либо *копирует* ссылки с некоторого сайта, чья тематика близка его автору. Идея призвана объяснить не только степенной закон, но и факт наличия в интернете плотных сообществ, участники которых объединены общими интересами.

Ссылки:

- http://elementy.ru/lib/431792
  - 1. *А. М. Райгородский*. Модели случайных графов. — М: МЦНМО, 2011.
  - 2. *B. Bollobás*. Random Graphs, Second Edition. — Cambridge Univ. Press, 2001.
  - 3. *E. A. Grechnikov*. An estimate for the number of edges between vertices of given degrees in random graphs in the Bollobas–Riordan model. — Moscow Journal of Combinatorics and Number Theory, 1 (2011), №2, p. 40–73.

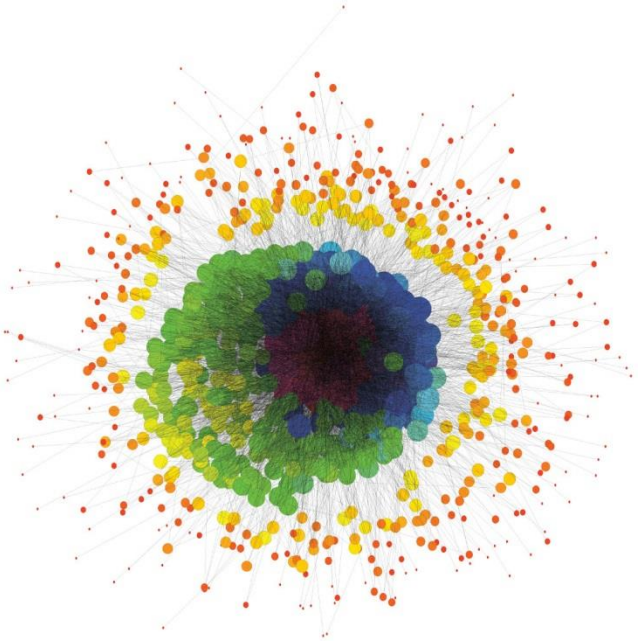
Graph 500 Benchmark

- **Graph500** (с 2010) — рейтинг суперкомпьютеров, ориентированный на задачи класса Data intensive. 211 записи (июнь 2016)
  - Бенчмарк в большей степени нагружает коммуникационную подсистему компьютера, и не зависит от количества исполняемых в секунду операций над числами с плавающей запятой
  - состоит из двух вычислительно сложных частей:
    - в первой происходит генерация графа и его сжатие в разреженные структуры CSR или CSC (Compressed Sparse Row/Column)
    - во второй происходит параллельный BFS-поиск из 64 вершин графа, выбранных случайно

Постановка задачи

Problem class definitions and required storage for the edge list assuming 64-bit integers.

Problem class	Scale	Edge factor	Approx. storage size in TB
Toy (level 10)	26	16	0.0172
Mini (level 11)	29	16	0.1374
Small (level 12)	32	16	1.0995
Medium (level 13)	36	16	17.5922
Large (level 14)	39	16	140.7375
Huge (level 15)	42	16	1125.8999



Генерация графа и валидация

- Для проверки созданной версии в тесте существует валидация, которая проверяет корректность построенного дерева при поиске в ширину.
- Для этого при выполнении поиска заполняется специальный массив, в котором хранятся вершины-родители для каждой вершины в построенном дереве.

- Проверятся корректность только косвенно:
  - Вывод является деревом, не содержит циклов
  - Метки у вершин, соединенных ребром дерева, отличаются ровно на 1
  - Все ребра в выходных данных либо обеими вершинами в дереве и их метки отличаются не более чем на 1, либо обеими вершинами вне его
  - Вершина и ее родитель соединены ребром из входных данных

- При генерации графов в Graph500 используется генератор Кронекера, очень похожий на генератор графов типа Recursive MATrix (R-MAT), который в процессе работы использует матрицу смежности создаваемого графа.
- При добавлении каждой дуги матрица смежности NxN рекурсивно дробится до тех пор, пока не будет получена матрица из одного элемента — это и есть выбранная дуга. Такой процесс повторяется M раз.
- Матрица на каждом шаге такого рекурсивного процесса дробится на четыре равные части: A, B, C и D.
- Для каждой из этих частей изначально задана вероятность, с которой происходит выбор именно ее при добавлении новой дуги. По умолчанию вероятности выбора частей матрицы равны: P(A) = 0,57; P(B) = 0,19; P(C) = 0,19; P(D) = 1-(A+B+C) = 0,05.

June 2016

No.	Rank	Machine	Installation Site	Number of nodes	Number of cores	Problem scale	GTEPS
1	1	K computer (Fujitsu - Custom)	RIKEN Advanced Institute for Computational Science (AICS)	82944	663552	40	38621.4
2	2	Sunway TaihuLight (NRCPC - Sunway MPP)	National Supercomputing Center in Wuxi	40768	**10 599 680**	40	23755.7
3	3	DOE/NNSA/LLNL Sequoia (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Lawrence Livermore National Laboratory	98304	1572864	41	23751
4	4	DOE/SC/Argonne National Laboratory Mira (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Argonne National Laboratory	49152	786432	40	14982
5	5	JUQUEEN (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Forschungszentrum Juelich (FZJ)	16384	262144	38	5848
6	6	Fermi (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	CINECA	8192	131072	37	2567

Ссылки

- http://www.graph500.org
- http://www.graph500.org/referencecode

Задания

- **[8 баллов]** 1) на вашем ноутбуке запустить тест Graph500 (референс код), определить характеристики теста
- **[10 баллов (или +5 к 1)]** 2) на каком-либо кластере запустить параллельную версию теста
- **[+5 баллов к 1)-2)]** 3) определить параметры графа Кронекера (построив 1000+ случайных графов scale=12+): диаметр, среднее расстояние между вершинами в одной компоненте связности, построить гистограмму #(n, d), оценить вероятность быть связным
- **[15 баллов]** 4) реализовать алгоритм построения случайного веб-графа в модели Боллобаша–Риордана (различные n=4000+, m=10..16). Определить параметры графов в этой модели (описаны выше).