Генерация социальных графов с использованием фреймворка для распределенных вычислений Apache Spark

Институт системного программирования РАН

Кирилл Чихрадзе

chykhradze@ispras.ru

Семинар Московской секции ACM SIGMOD

24 апреля, 2014
Москва, Россия

ИСП РАН

План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Что такое Spark?

- Независимая, быстрая платформа для распределённых вычислений, поддерживающая обработку данных по модели MapReduce, Pregel, GraphX
  - ✓ Хранение данных в оперативной памяти для быстрой обработки интерактивных запросов
  - ✓ Может быть в 100 раз быстрее Hadoop
- Совместим с системой хранения Hadoop (HDFS, Hbase, SequenceFiles и т. п.)

ИСП РАН

Модель программирования Spark

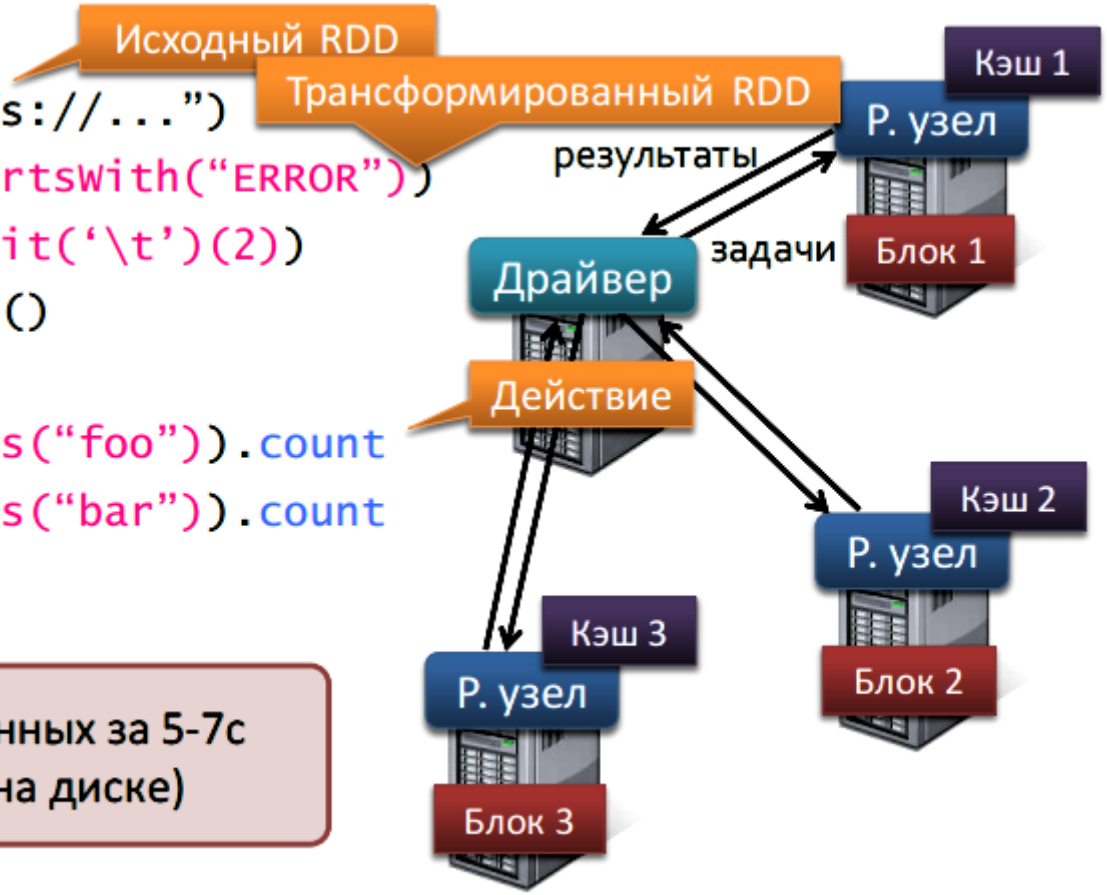
- Основная идея: сбое-устойчивые распределённые наборы данных (RDDs)
  - ✓ Распределённые коллекции объектов, которые могут быть кэшированы в памяти узлов кластера
  - ✓ Можно манипулировать с помощью различных параллельных операций (таких как map и reduce)
  - ✓ Автоматически перестраиваются при сбоях
- Интерфейс:
  - ✓ Элегантный интерфейс, интегрированный в язык Scala
  - ✓ Может использоваться интерактивно из консоли Scala

Пример: анализ логов

1. Загрузить сообщения с ошибками в память
2. Интерактивно выполнять запросы к ним

```
lines = spark.textFile(“hdfs://...”)
errors = lines.filter(_.startsWith(“ERROR”))
messages = errors.map(_.split(‘\t’)(2))
cachedMsgs = messages.cache()

cachedMsgs.filter(_.contains(“foo”)).count
cachedMsgs.filter(_.contains(“bar”)).count
. . .
```



Результат: запросы к 1 ТВ данных за 5-7с
(против 170с для данных на диске)

ИСП РАН

План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Постановка задачи

Сгенерировать случайный граф…

1. …который будет удовлетворять основным свойствам социальных сетей;
2. …за приемлемое время даже из миллиарда вершин;
3. …сообщества пользователей которого подчиняются основным свойствам реальной структуры сообществ.

ИСП РАН

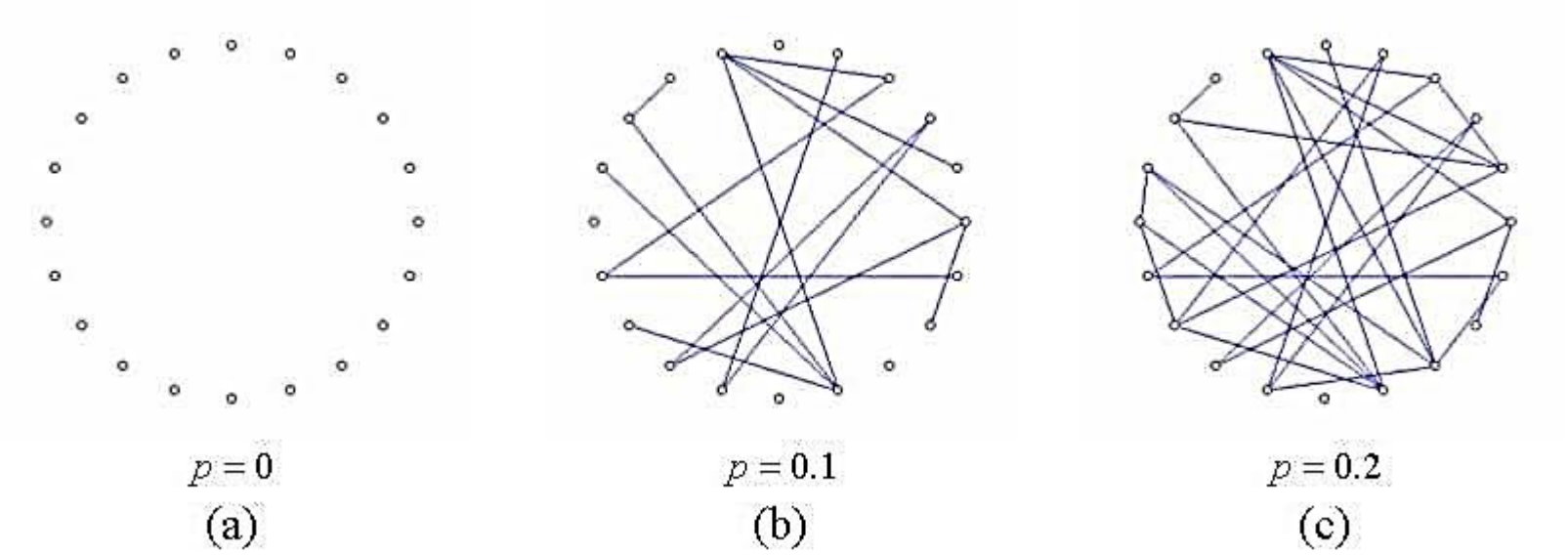
Причины

1. Размеры современных социальных сетей достигаю сотен миллионов вершин.
2. Необходимы алгоритмы поиска сообществ, чья эффективность доказана на больших графах.
3. Стандартный способ оценки эффективности метода поиска сообщества – тестирование на случайных графах с известной структурой сообществ.

Что такое случайный граф?

Граф Эрдёша-Реньи:

- N вершин
- Ребро появляется с вероятностью p



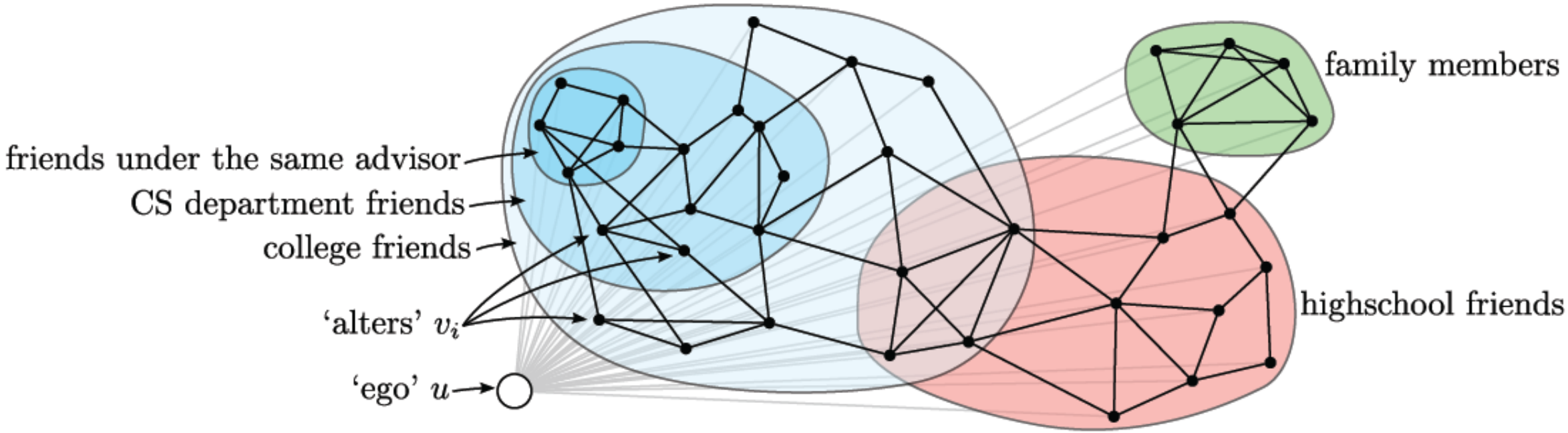
Источник картинки: press.anu.edu.au/cs/chap5Newth-final-5.jpg

ИСП РАН

Что такое сообщество?

ИСП РАН

Что такое сообщество?



ИСП РАН

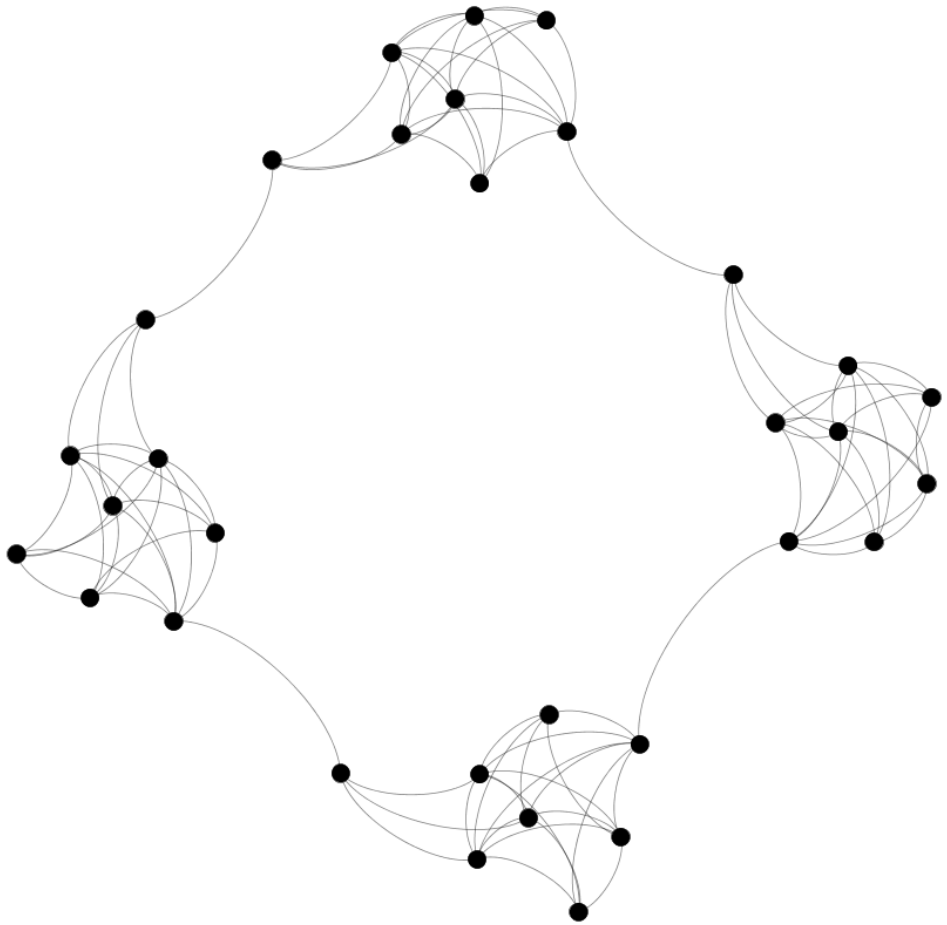
План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

GN

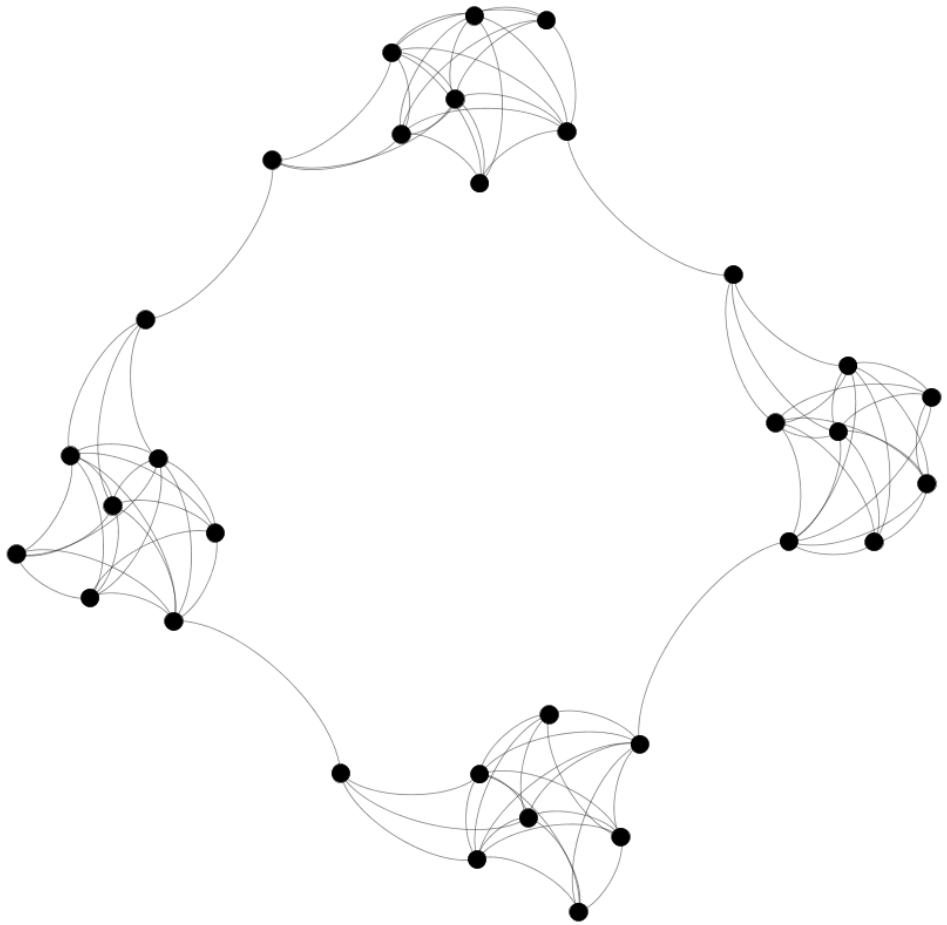
- Граф разбит на *m групп* (m=4)
- В каждой группе *g вершин* (*g*=8)
- Вероятность ребра внутри сообщества – p_{in}
  Вероятность ребра между сообществами – p_{out}



ИСП РАН

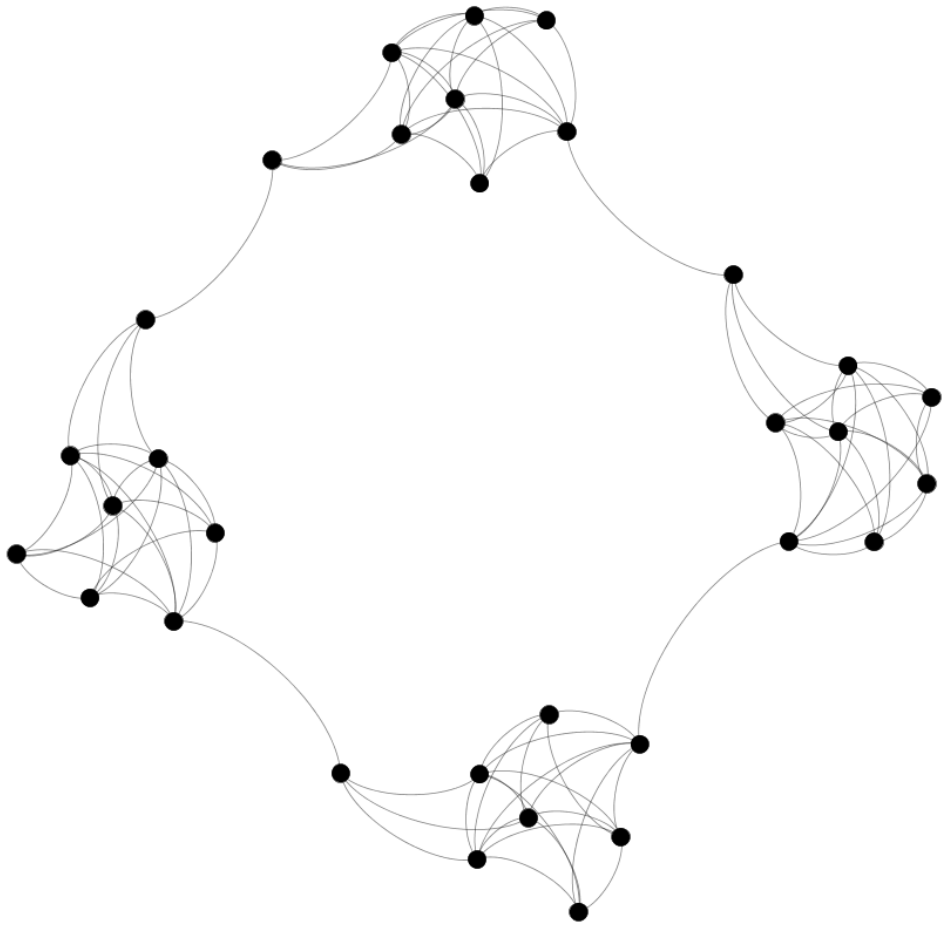
Преимущества GN

- Простая модель
- Быстрая генерация
- Чёткое разграничение сообществ



Недостатки GN

- Сообщества не пересекаются
- Все сообщества одинакового размера
- Распределение степеней не подчиняется степенному закону*



*Степенной закон: $P(x) = Cx^{-\alpha}, \forall\, x \geq x_{min}, \mathrm{C} = (\alpha - 1)x_{min}^{\alpha-1}$

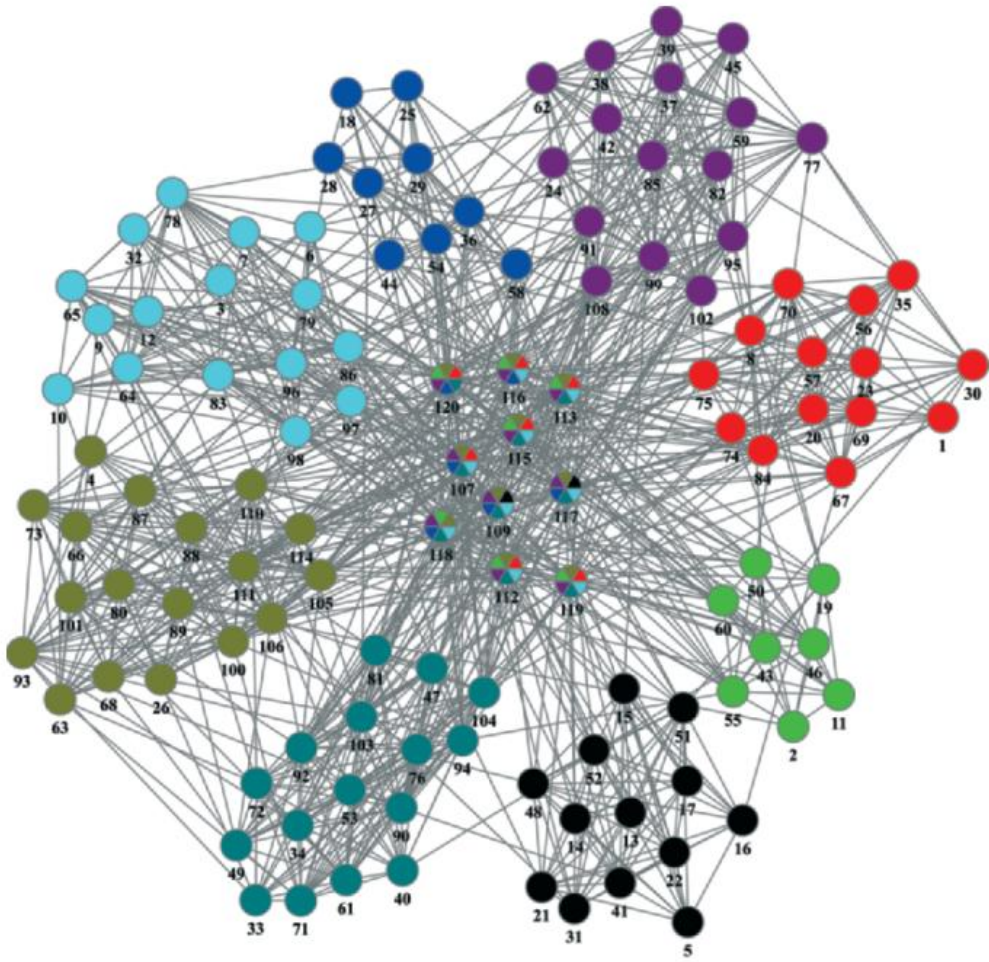
ИСП РАН

LFR

- Степени вершин и размеры сообществ распределяются по степенному закону
- Степень внутри сообщества:

$$k_i^{(in)} = (1 - \mu)k_i$$

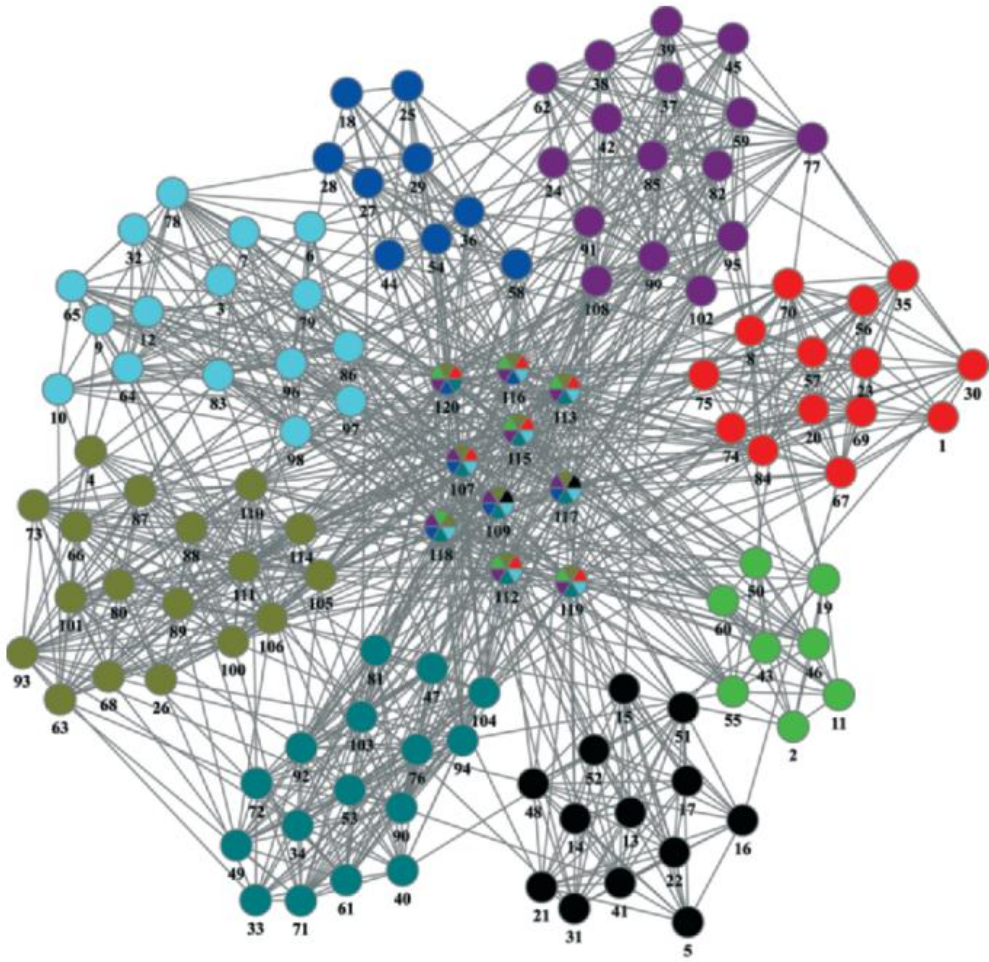
- Rewiring-процесс генерирует связи между разными сообществами



ИСП РАН

Преимущества LFR

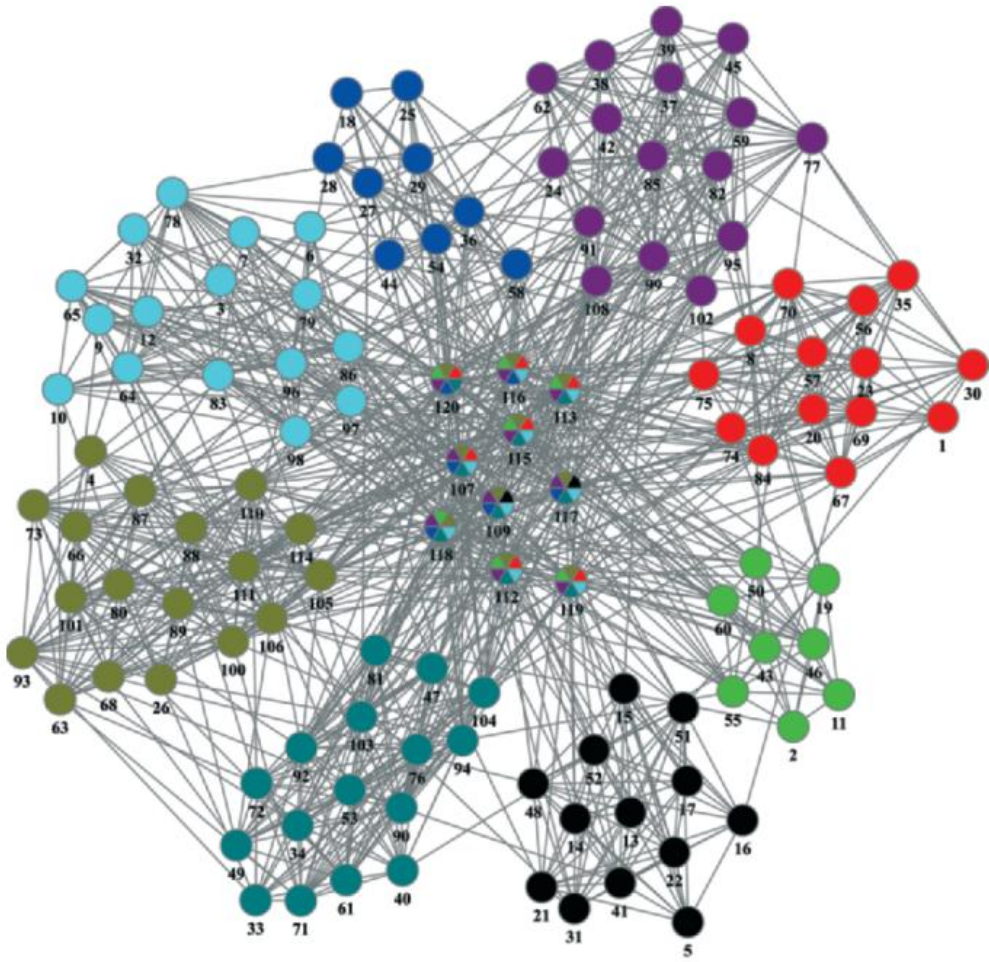
- Более реалистичные сети
  со степенным распределением
- Пересекающаяся и/или иерархическая
  структура сообществ
- Поддержка вариации коэффициента
  кластеризации



ИСП РАН

Недостатки LFR

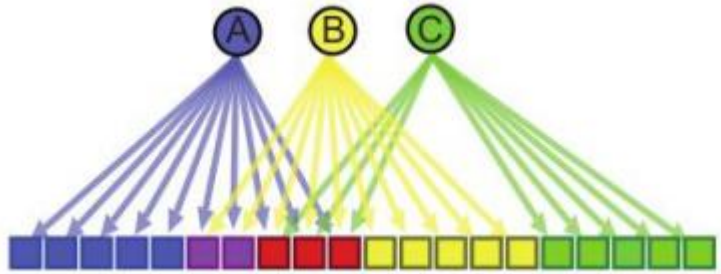
- Большое время генерации (500,000 вершин за 1 час)
- Нераспределённая реализация алгоритма
- Фиксированное количество вхождений вершины в сообщества

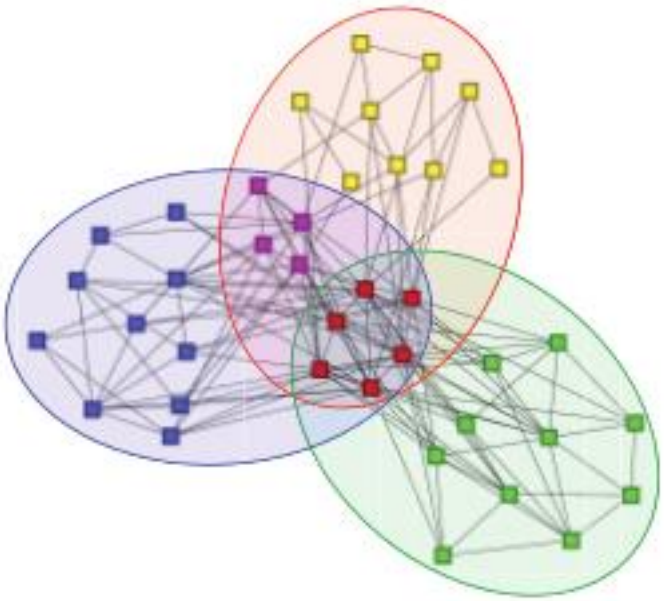


ИСП РАН

agmgen

- Вероятностная генеративная модель
- Граф распределения вершин по сообществам задан
- Рёбра в сообществах появляются независимо и с одинаковой вероятностью

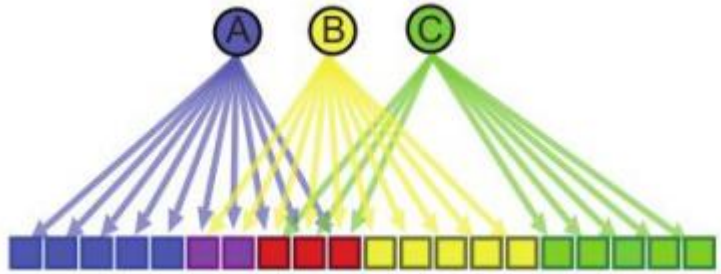




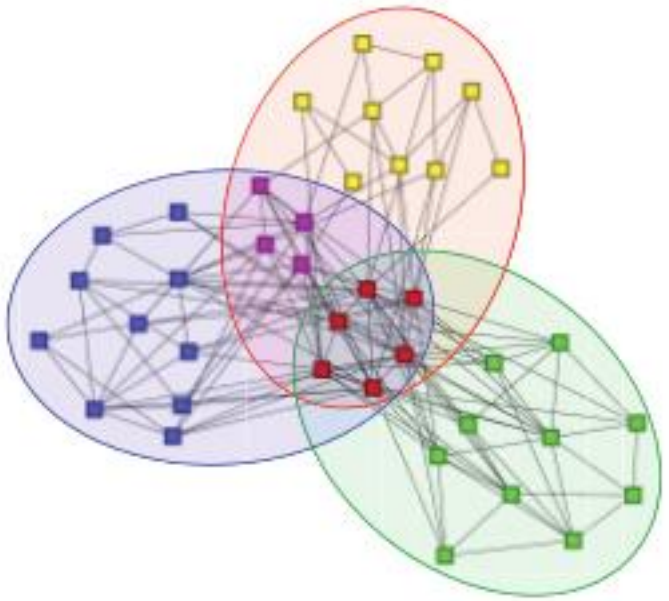
ИСП РАН

Преимущества agmgen

- Степенной распределение степени вершин



- Правдоподобная вероятность ребра в каждом сообществе



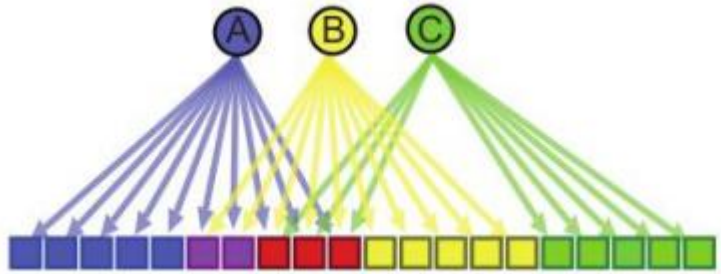
- Сообщества подчиняются основным свойствам структуры сообществ

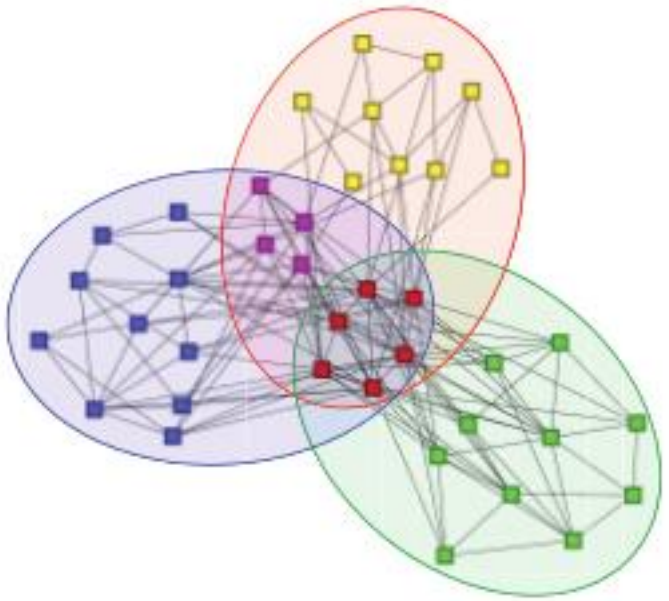
ИСП РАН

Недостатки agmgen

- Необходимо задать граф распределения вершин по сообществам

- Нераспределённая реализация алгоритма





ИСП РАН

Сравнение существующих решений

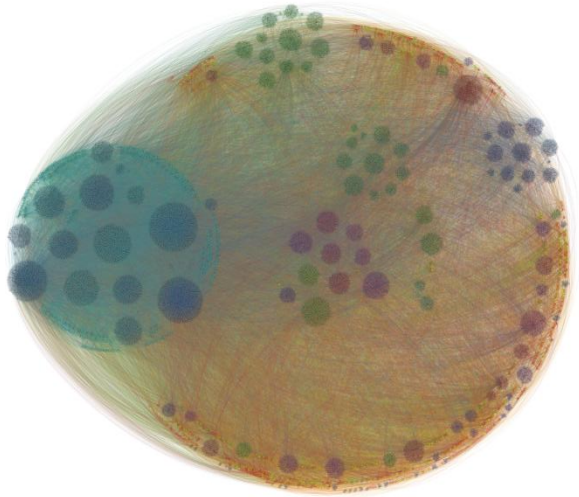
Свойство	GN	LFR	agmgen
Способность генерировать графы со степенным распределением вершин	–	+	+
Способность создавать реалистичную структуру сообществ	–	+	–
Коэффициент кластеризации	–	+	–
Степенное распределение размеров сообществ	–	+	–
Степенное распределение числа вхождений вершин в сообщества	–	–	–
Плотные пересечения сообществ	–	–	+

ИСП РАН

План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

СКВ: входные параметры



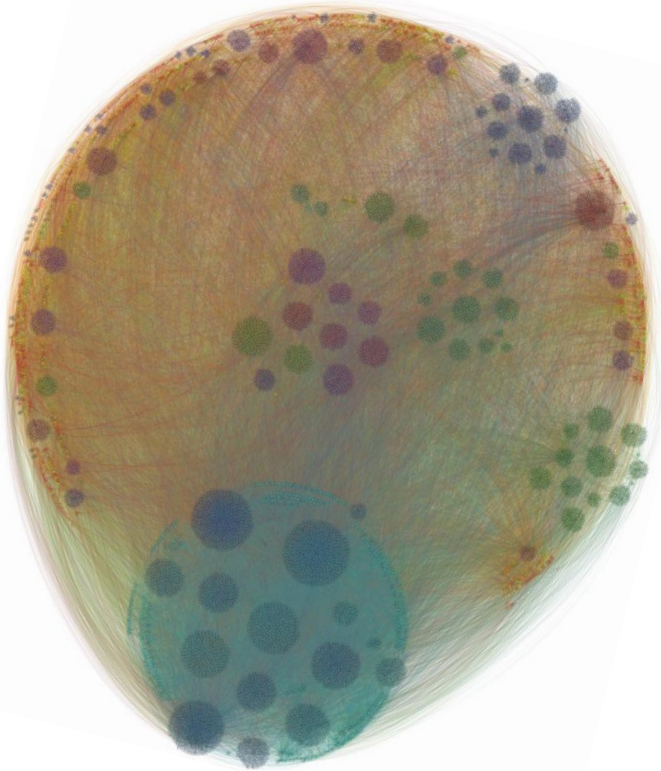
- N_1 — количество вершин
- d_{mean} — средняя степень
- β_1, β_2 — две экспоненты степенного распределения
- Минимальные и максимальные размер сообществ и количество вхождений вершин в сообщество
- α, γ — две константы, определяющие вероятность ребра
- ε — вероятность ребра между сообществами

ИСП РАН

СКВ

Основные шаги:

1. Генерация двудольного графа вершина-сообщество
2. Генерируются рёбра внутри сообществ
3. Генерируются рёбра между сообществами

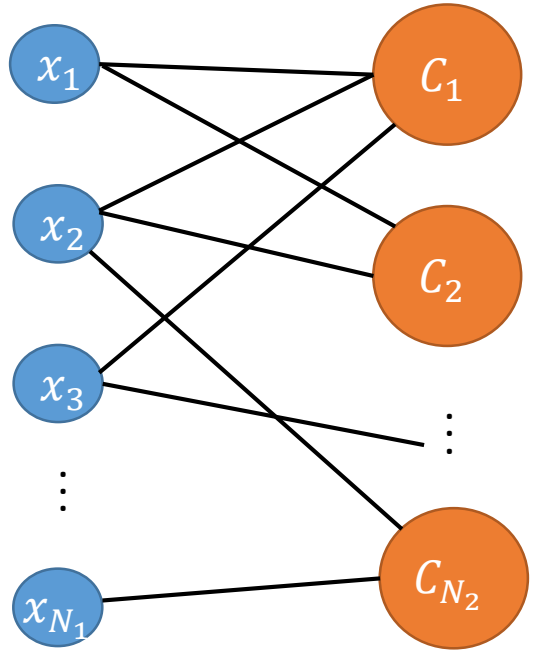


СКВ: вершина-сообщество

- Количество сообществ N_2 вычисляется из*

$$N_1 \cdot E[X_1] = N_2 \cdot E[X_2]$$

- Распределения количества вхождений вершин в сообщества (m_i) и размера сообществ (x_c) генерируется по степенному закону с экспонентами β_1 и β_2
- Рёбра в двудольном графе создаются в соответствии с моделью конфигураций



*$E[X_1]$ и $E[X_2]$ – средние значения числа вхождений пользователя в сообщество и размера сообществ соответственно

СКВ: генерация рёбер

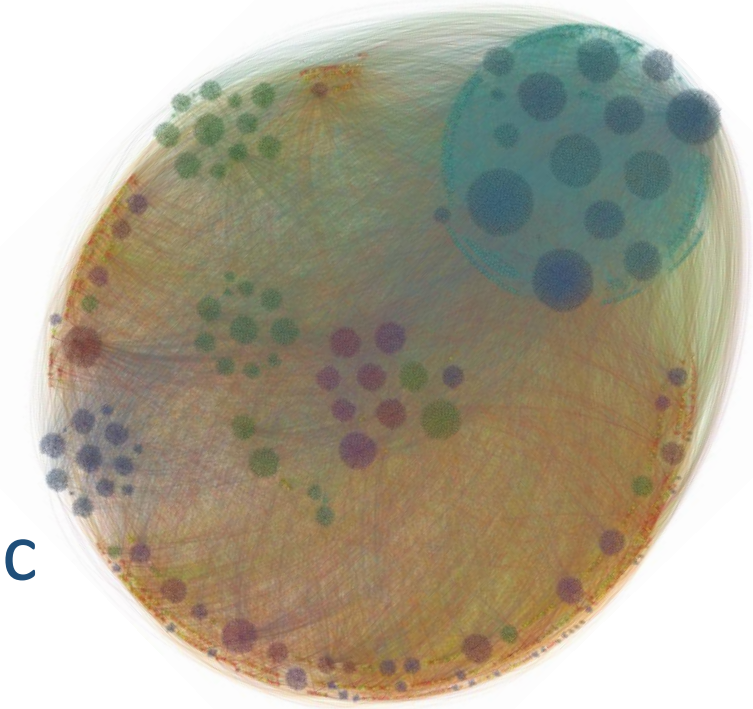
- Рёбра в сообществах генерируются с вероятностью*:

$$p_c = \frac{\alpha}{x_c^{\gamma}}$$

где x_c – размер сообщества

- Рёбра между сообществами генерируются с вероятностью**:

$$p_{out} = \varepsilon$$



* Yang, J., and Leskovec, J. Structure and overlaps of communities in networks.
**Yang, J., and Leskovec, J. Community-affiliation graph model for overlapping network community detection.

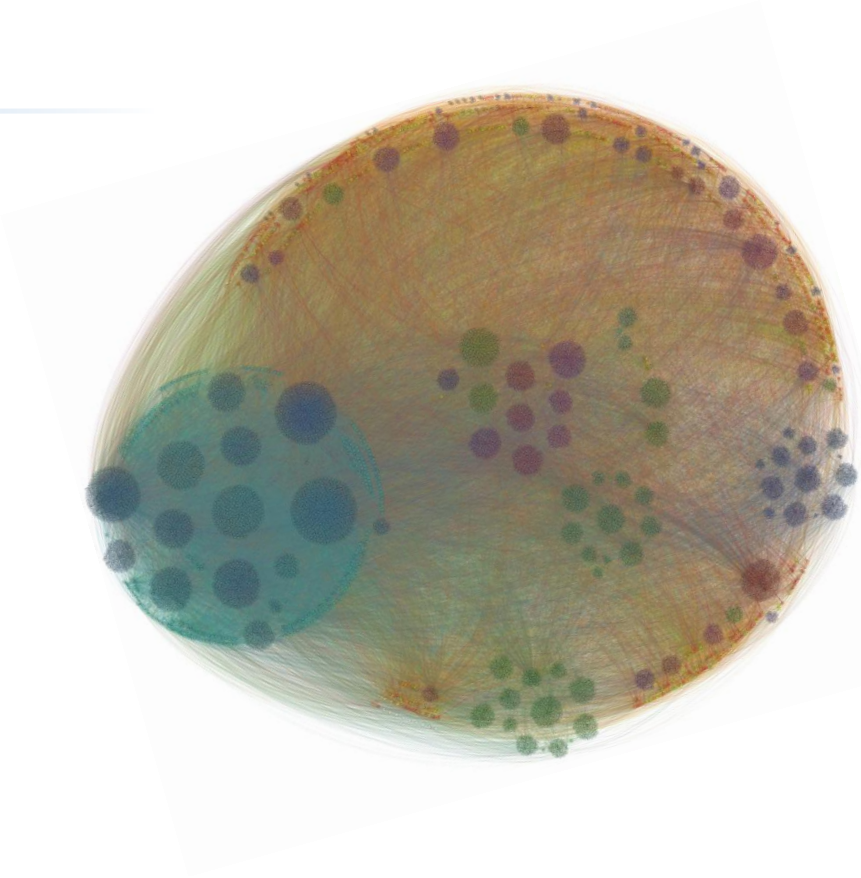
ИСП РАН

СКВ: генерация рёбер

- Количетсво рёбер в сообществе:

$$M_c \sim (1 + \mathbb{P}) Bin\left(\frac{x_c(x_c - 1)}{2}, p_c\right),$$

где $\mathbb{P}$ – вероятность кратного ребра

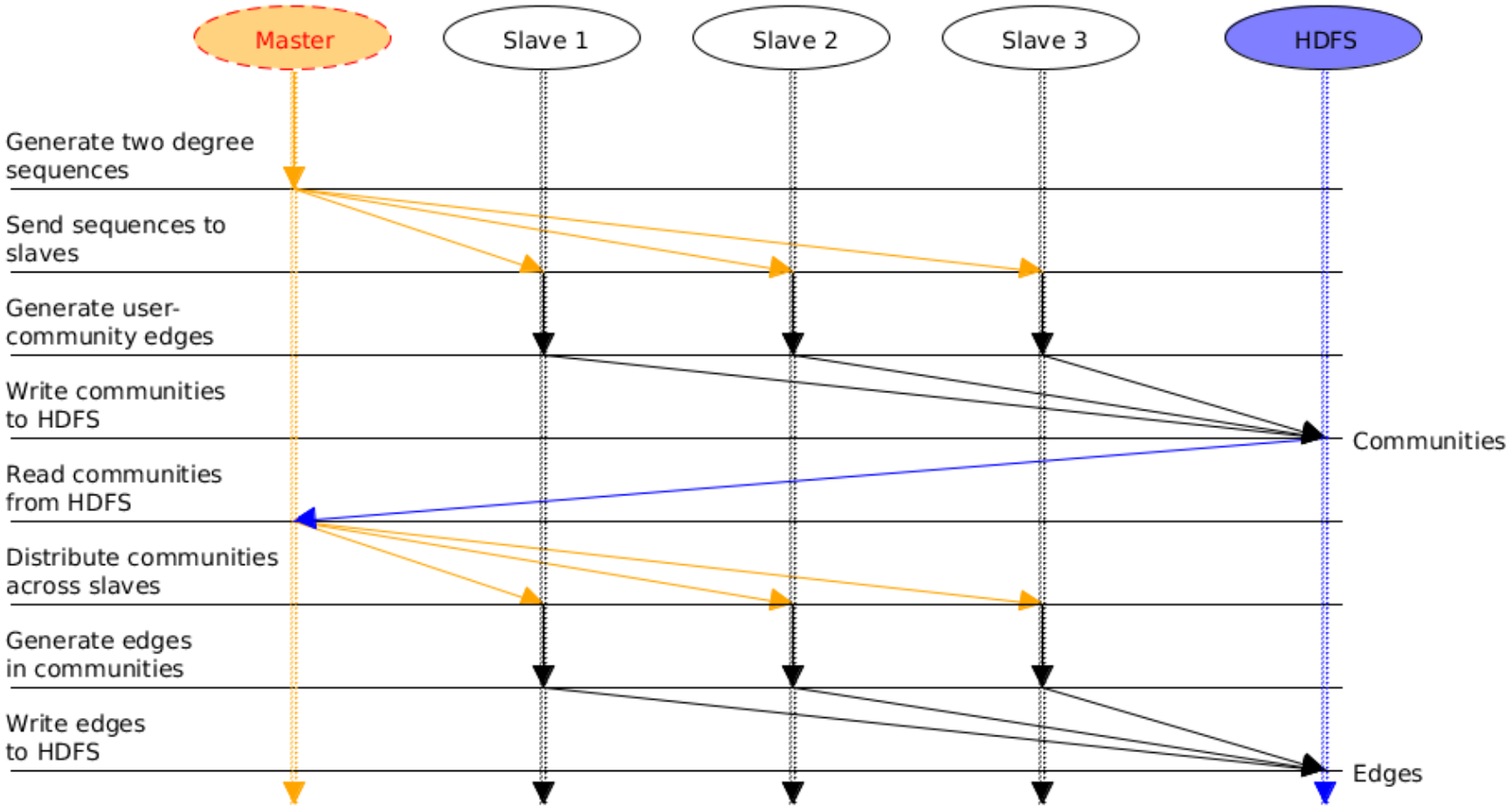


- Количество рёбер между сообществами:

$$M_o \sim Bin(\frac{N_1(N_1 - 1)}{2}, \varepsilon)$$

ИСП РАН

СКВ: реализация Apache Spark



ИСП РАН

План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Оценка точности

	LiveJournal	CKB
Количество вершин	$\sim 4 \cdot 10^6$	$\sim 4.2 \cdot 10^6$
Количество рёбер	$\sim 34.6 \cdot 10^6$	$\sim 38.2 \cdot 10^6$
$\beta_{\text{вершины}}$	2.14	2.15
$\beta_{\text{сообщества}}$	2.22	2.26
$\beta_{\text{степени}}$	2.15	2.15
Медиана распределения размеров сообществ (x_c)	10	8
Медиана распределения количества вхождения вершин в сообщества (m_i)	2	2
Доля вершин с $m_i > 1$	63%	66%
Средний коэффициент кластеризации	0.3538	0.0134
Эффективный диаметр	6.4	5.16

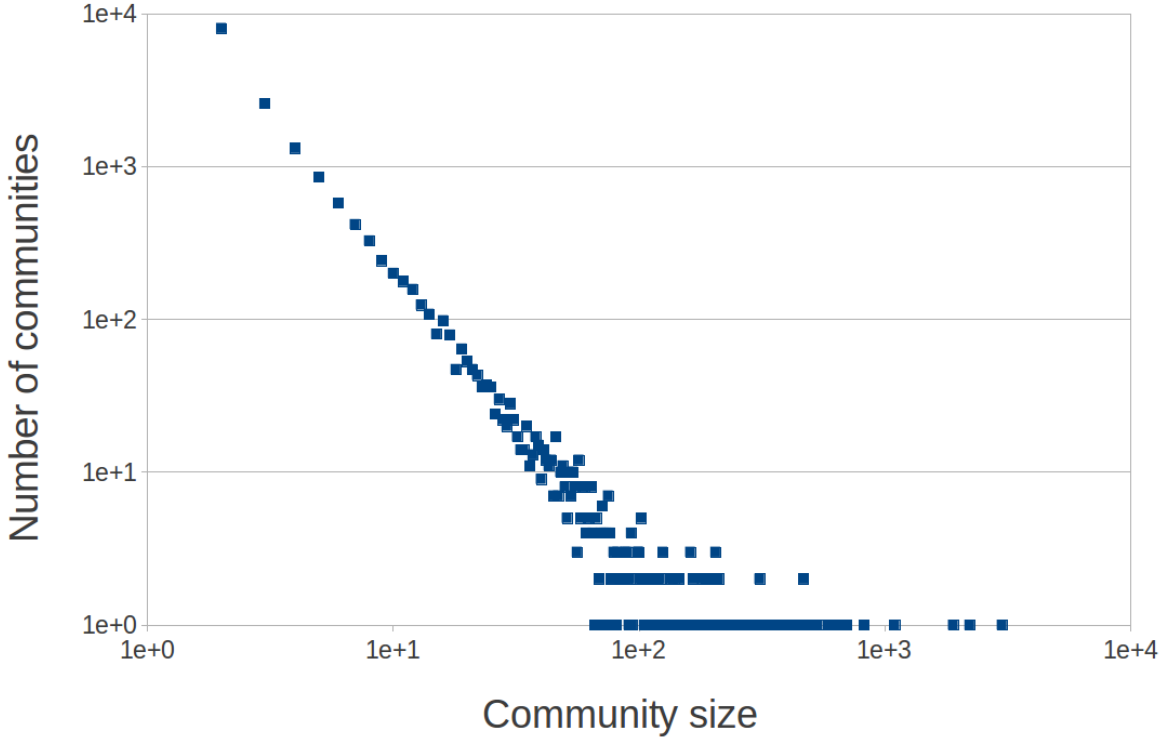
ИСП РАН

Оценка точности

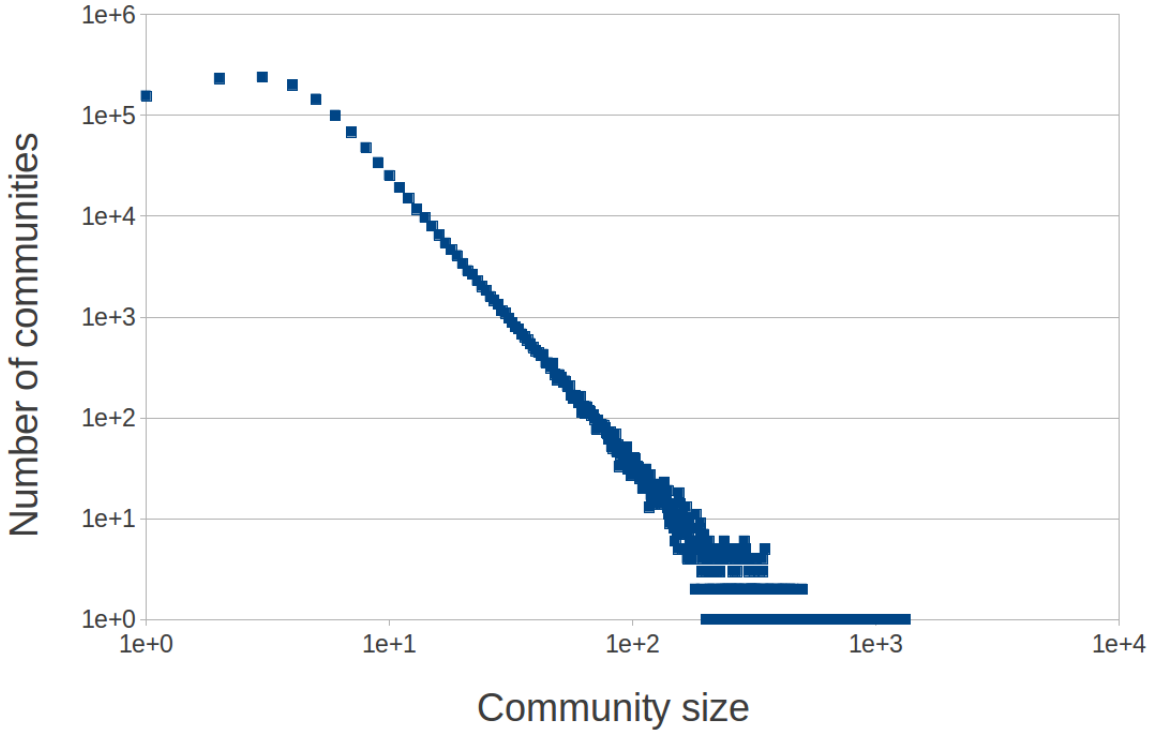
	YouTube	СКВ
Количество вершин	$\sim 1.1 \cdot 10^6$	$\sim 1.1 \cdot 10^6$
Количество рёбер	$\sim 3 \cdot 10^6$	$\sim 3 \cdot 10^6$
$\beta_{\text{вершины}}$	2.36	2.41
$\beta_{\text{сообщества}}$	2.83	2.95
$\beta_{\text{степени}}$	2.53	2.45
Медиана распределения размеров сообществ (x_c)	3	4
Медиана распределения количества вхождения вершин в сообщества (m_i)	2	2
Доля вершин с $m_i > 1$	38%	68%
Средний коэффициент кластеризации	0.1723	0.0166
Эффективный диаметр	6.5	6.2

Оценка точности: YouTube vs. CKB

YouTube

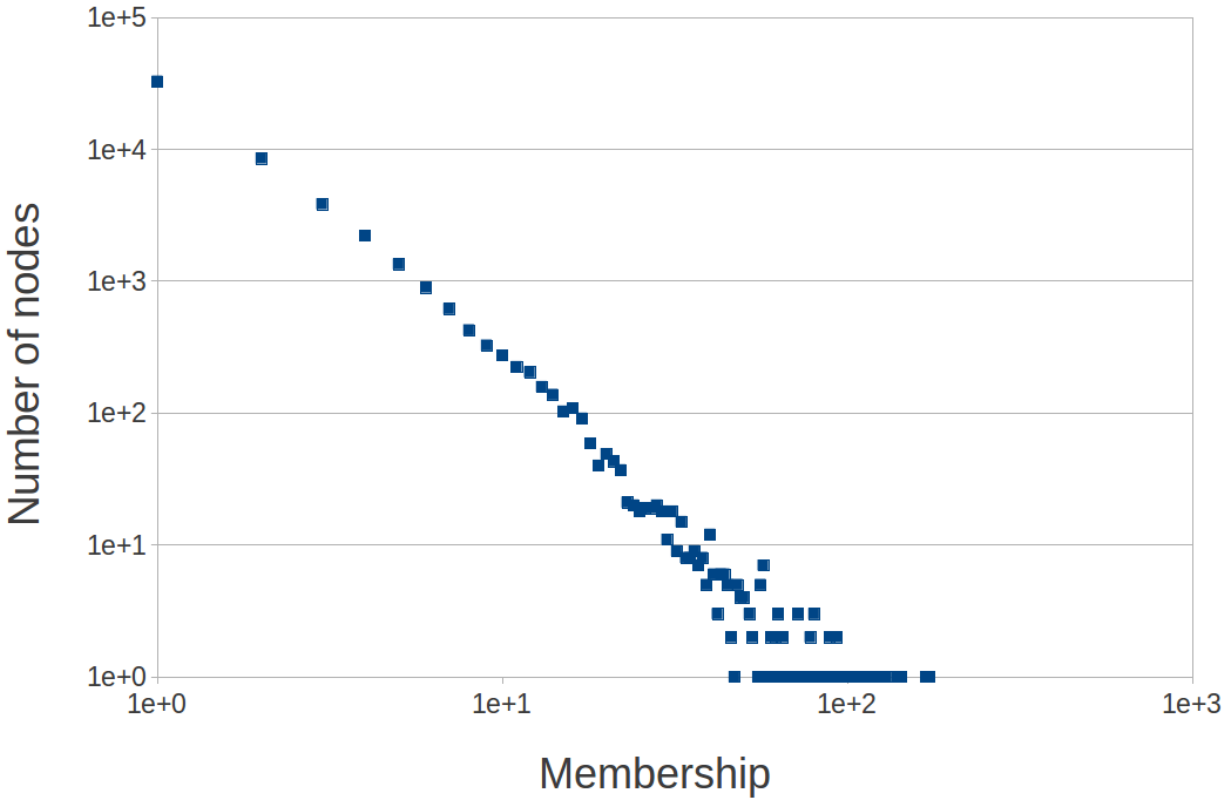


CKB

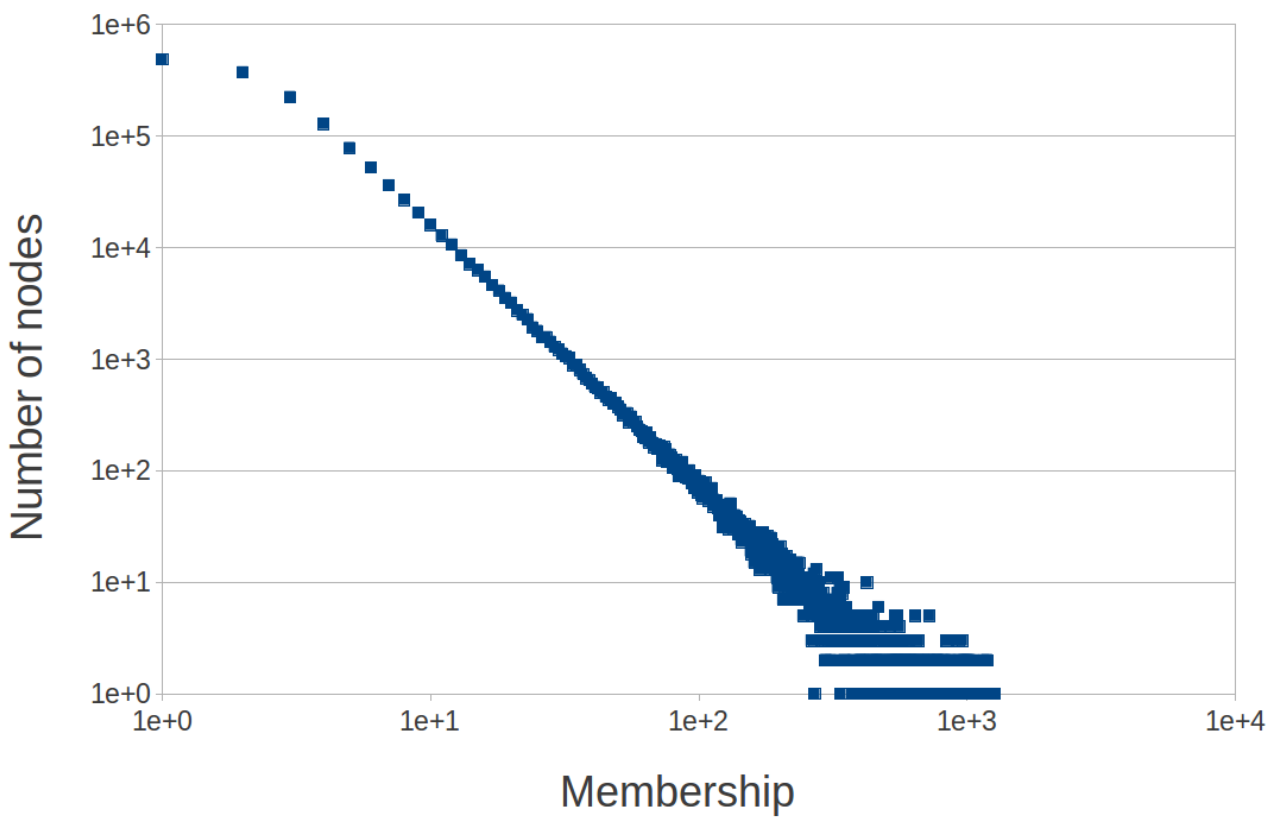


Оценка точности: YouTube vs. CKB

YouTube

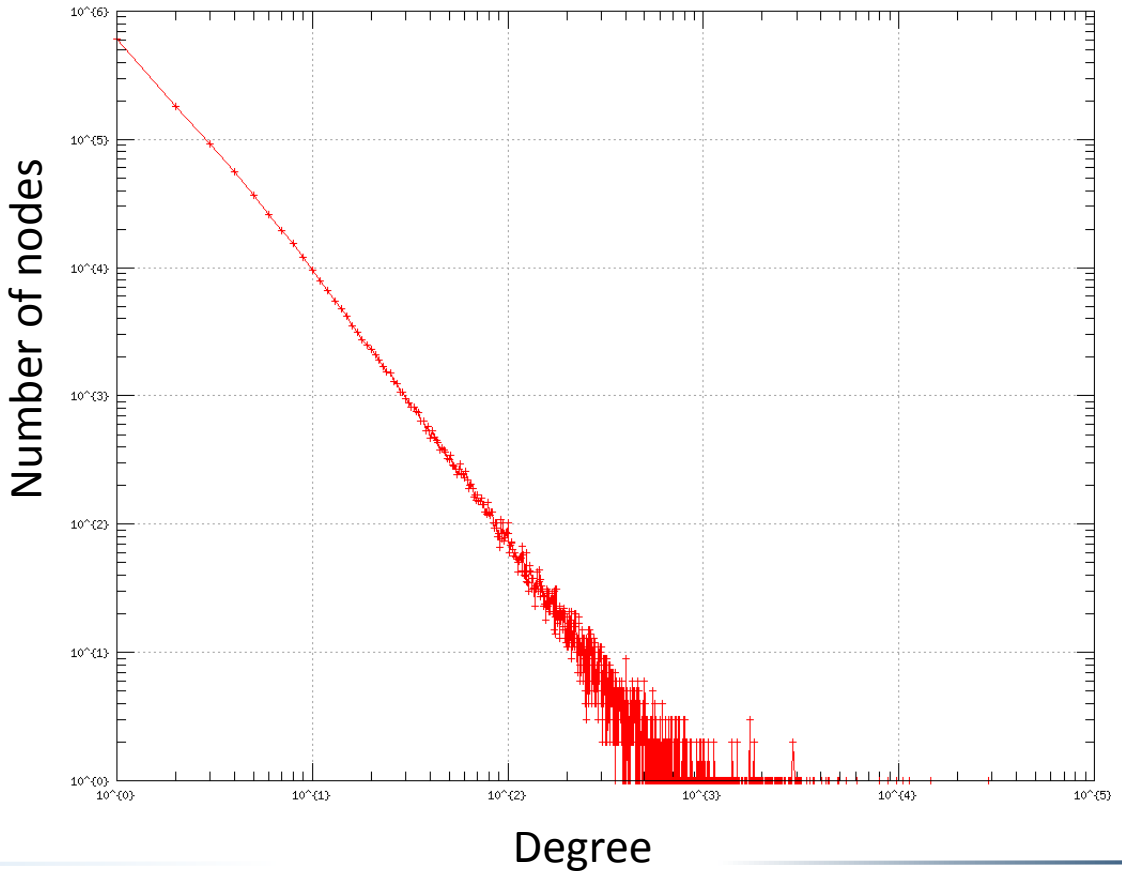


CKB

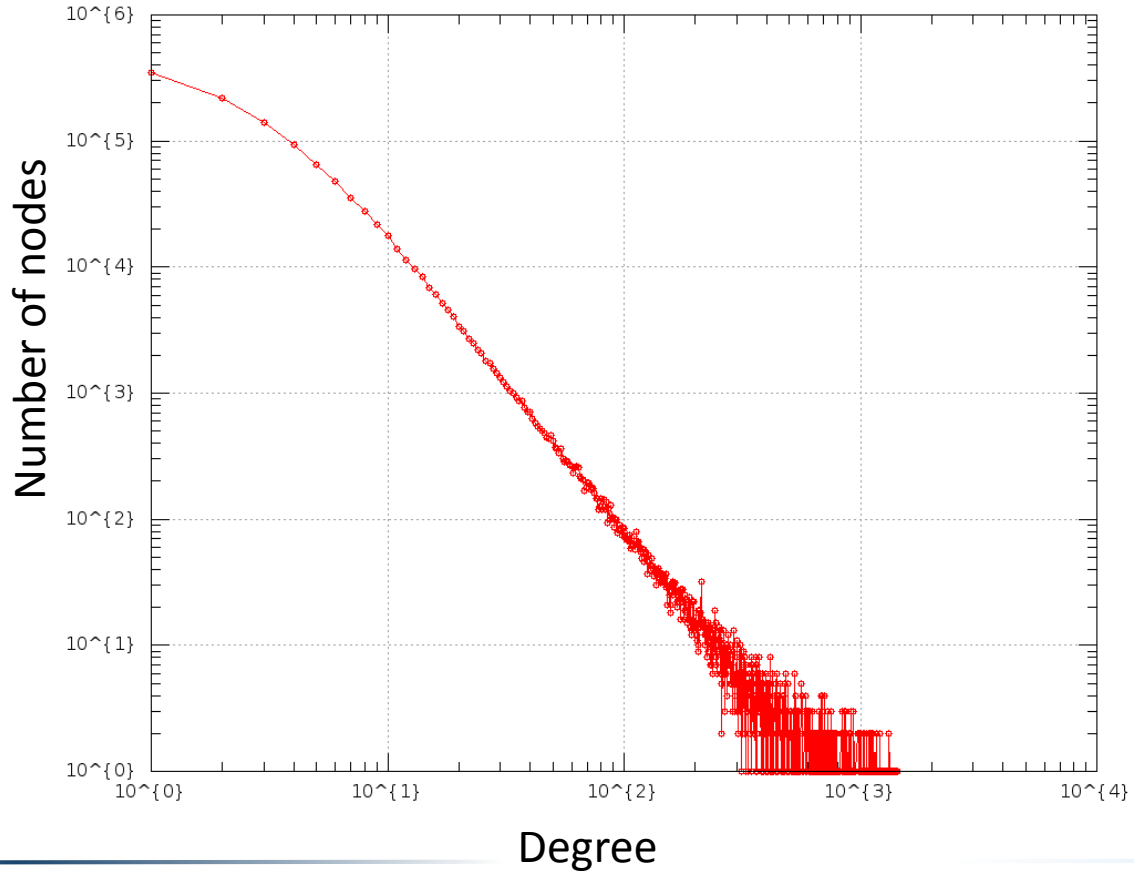


Оценка точности: YouTube vs. СКВ

YouTube

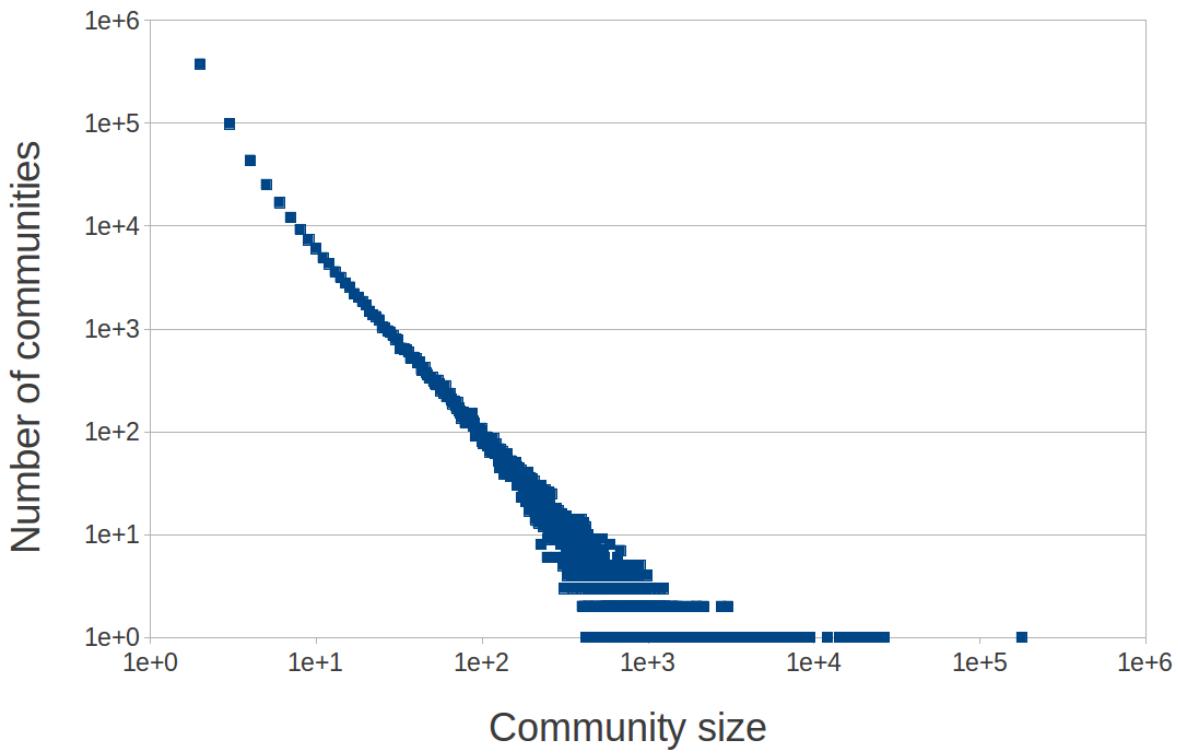


СКВ

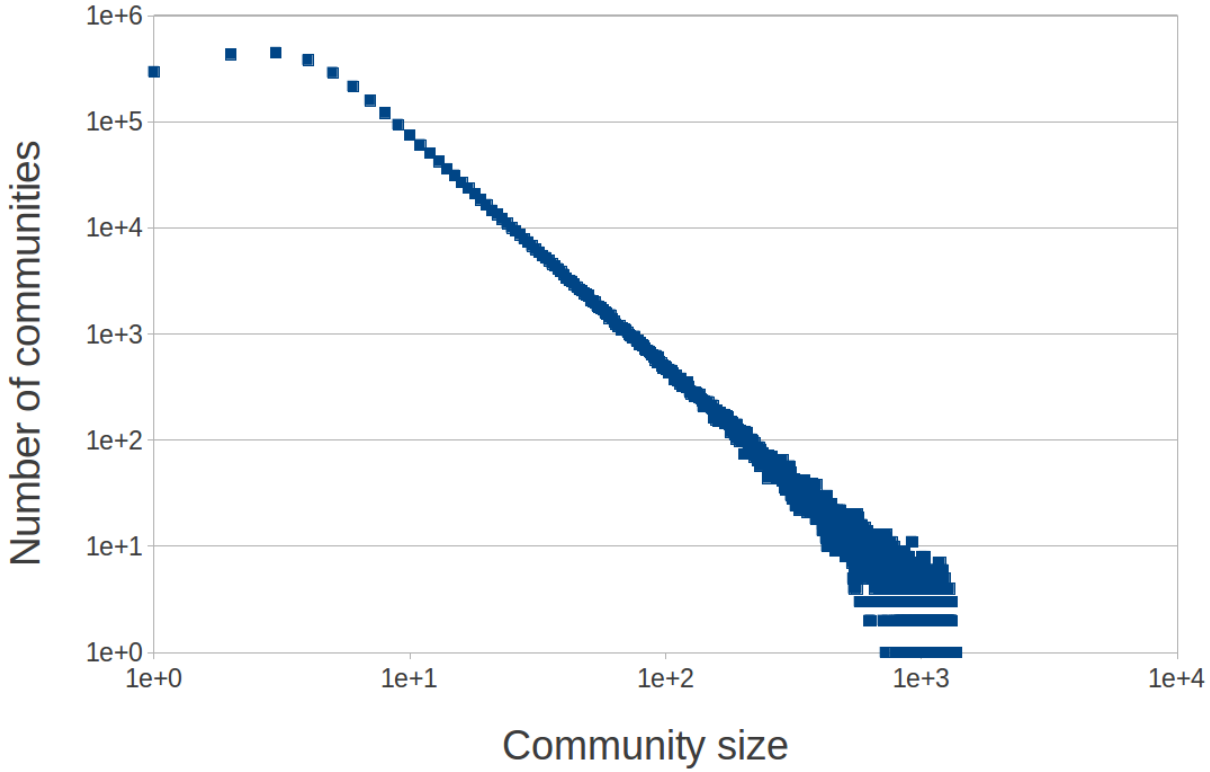


Оценка точности: LiveJournal vs. CKB

LiveJournal

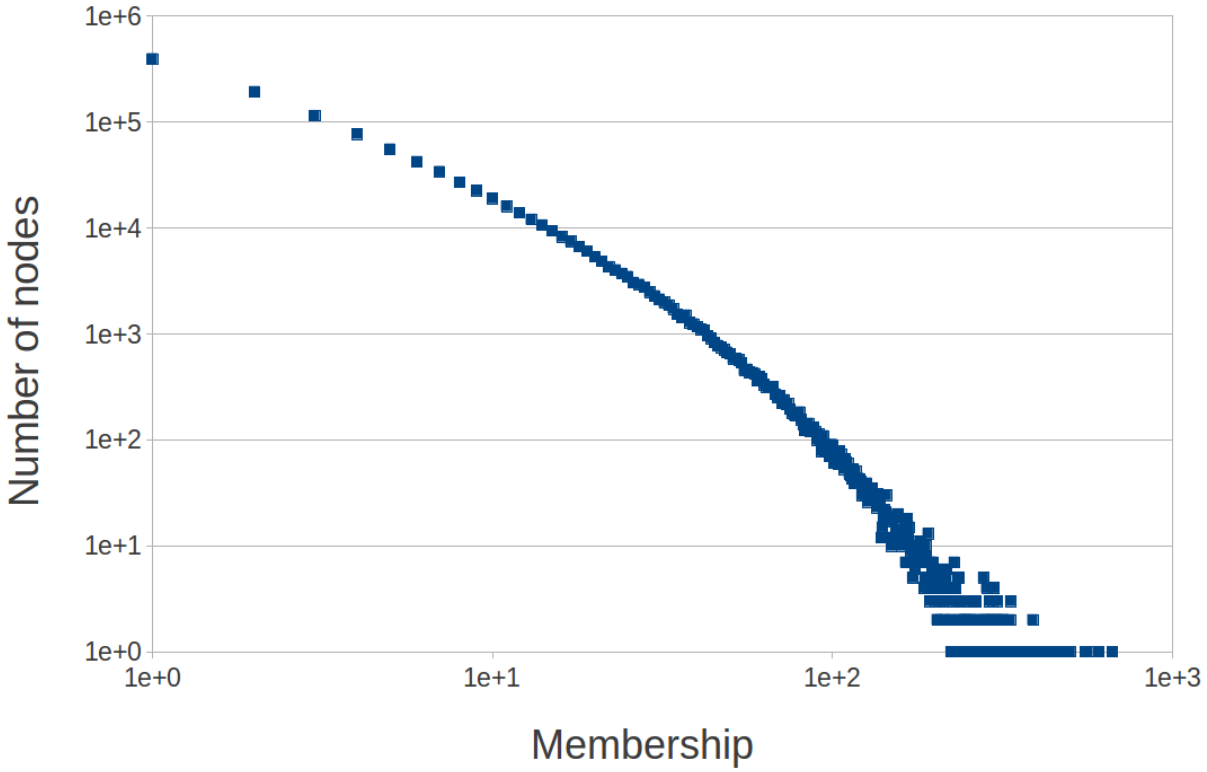


CKB

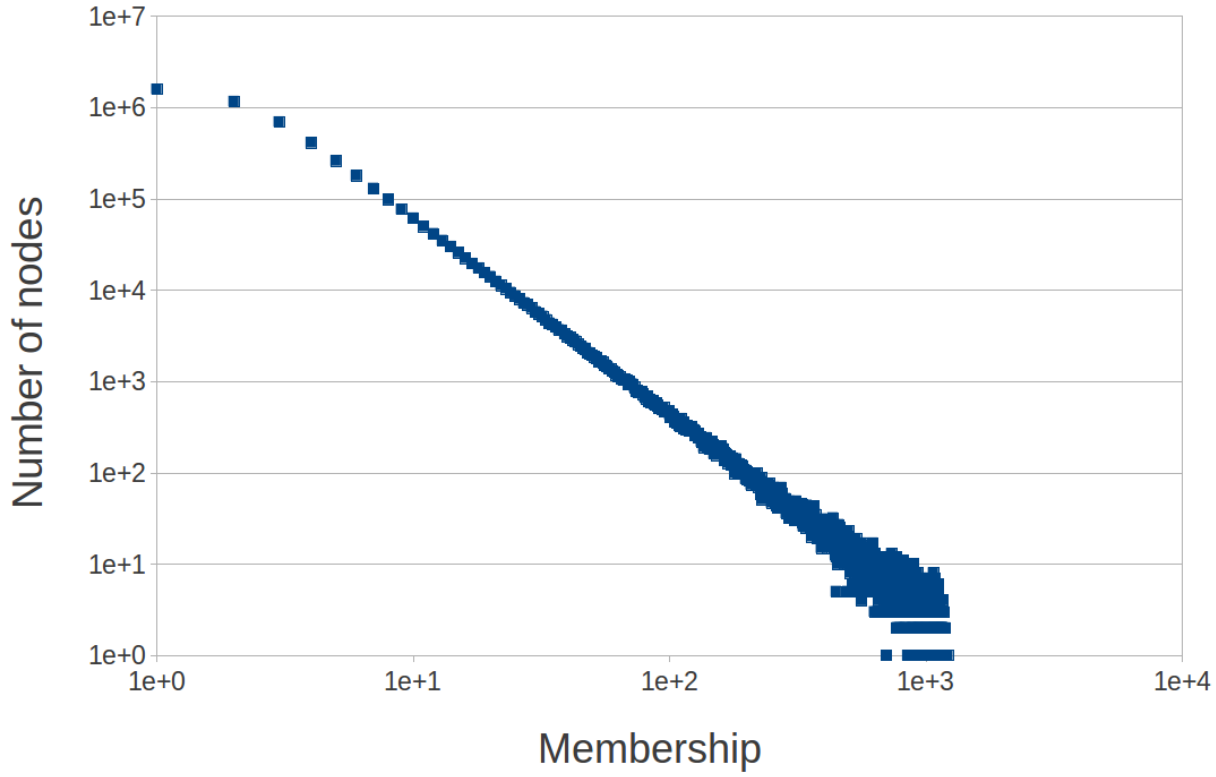


Оценка точности: LiveJournal vs. CKB

LiveJournal

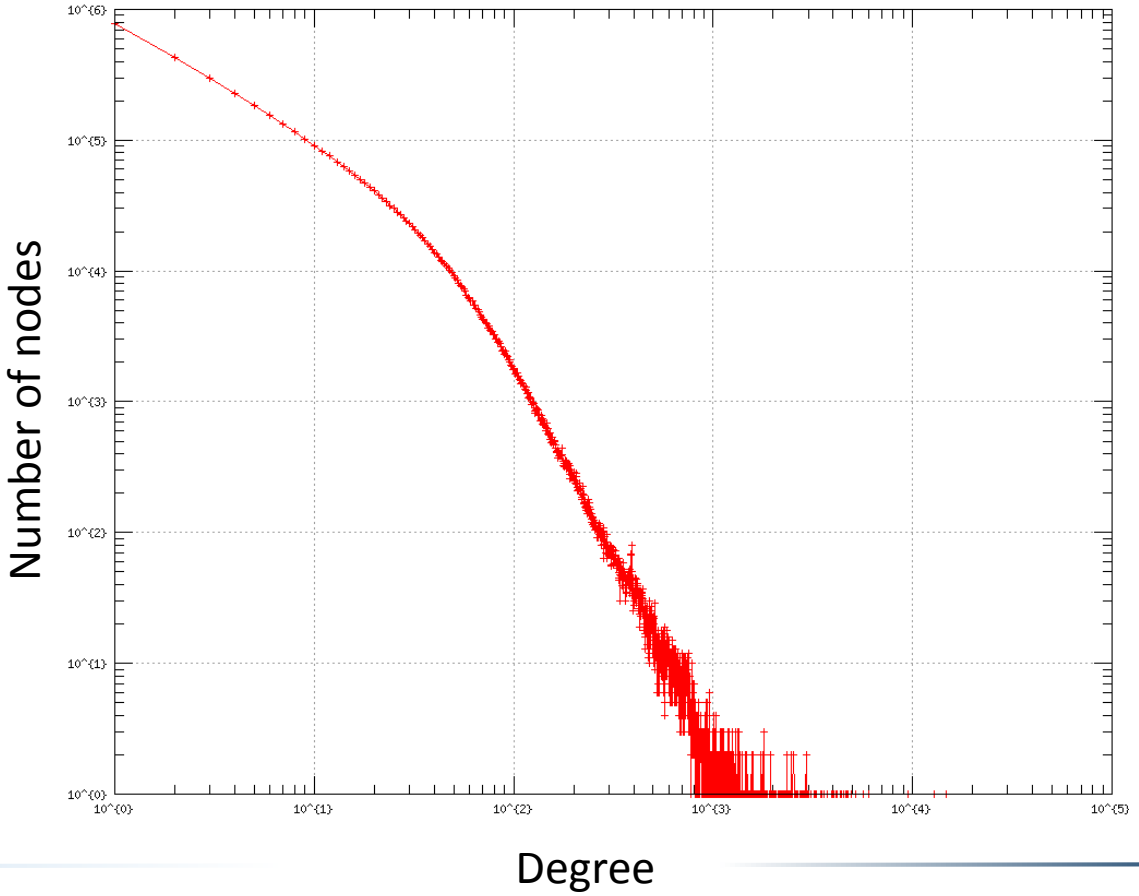


CKB

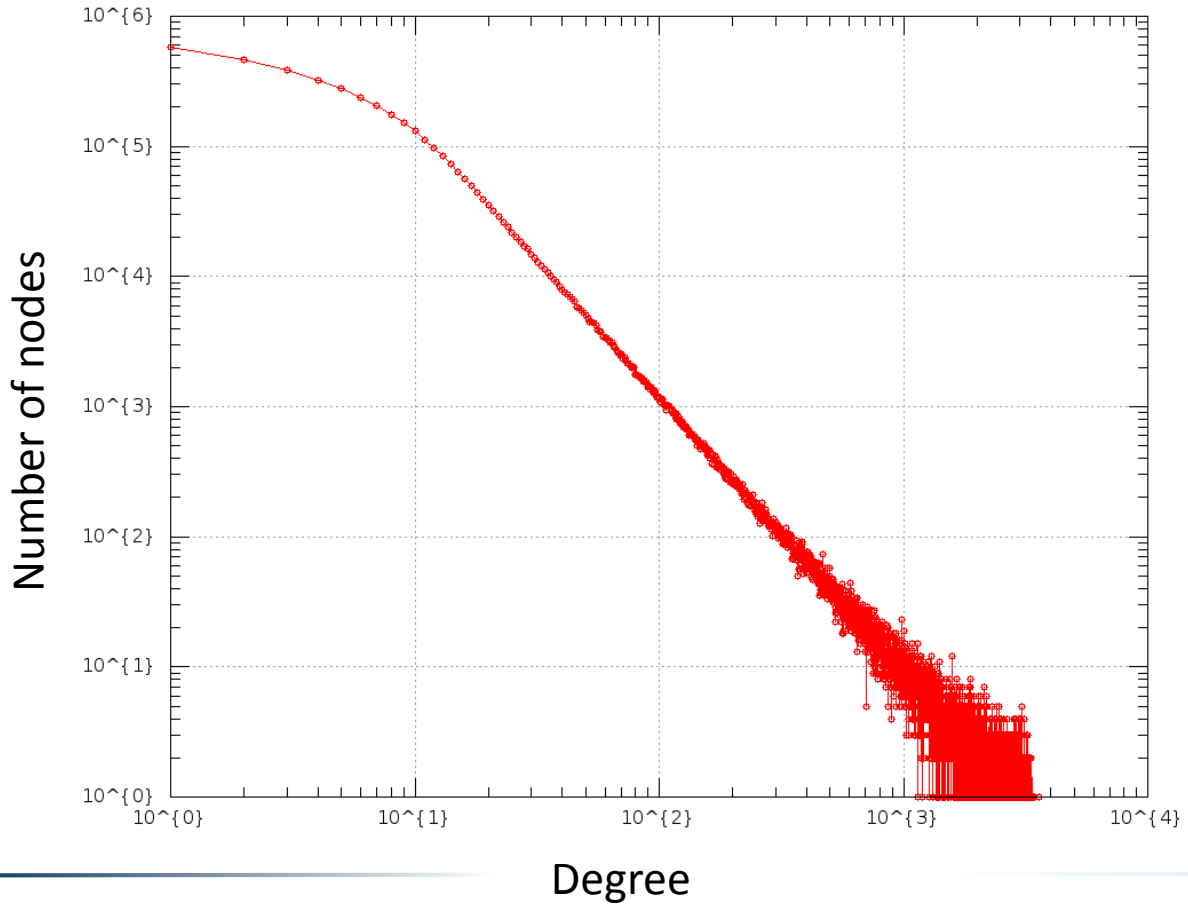


Оценка точности: LiveJournal vs. CKB

LiveJournal



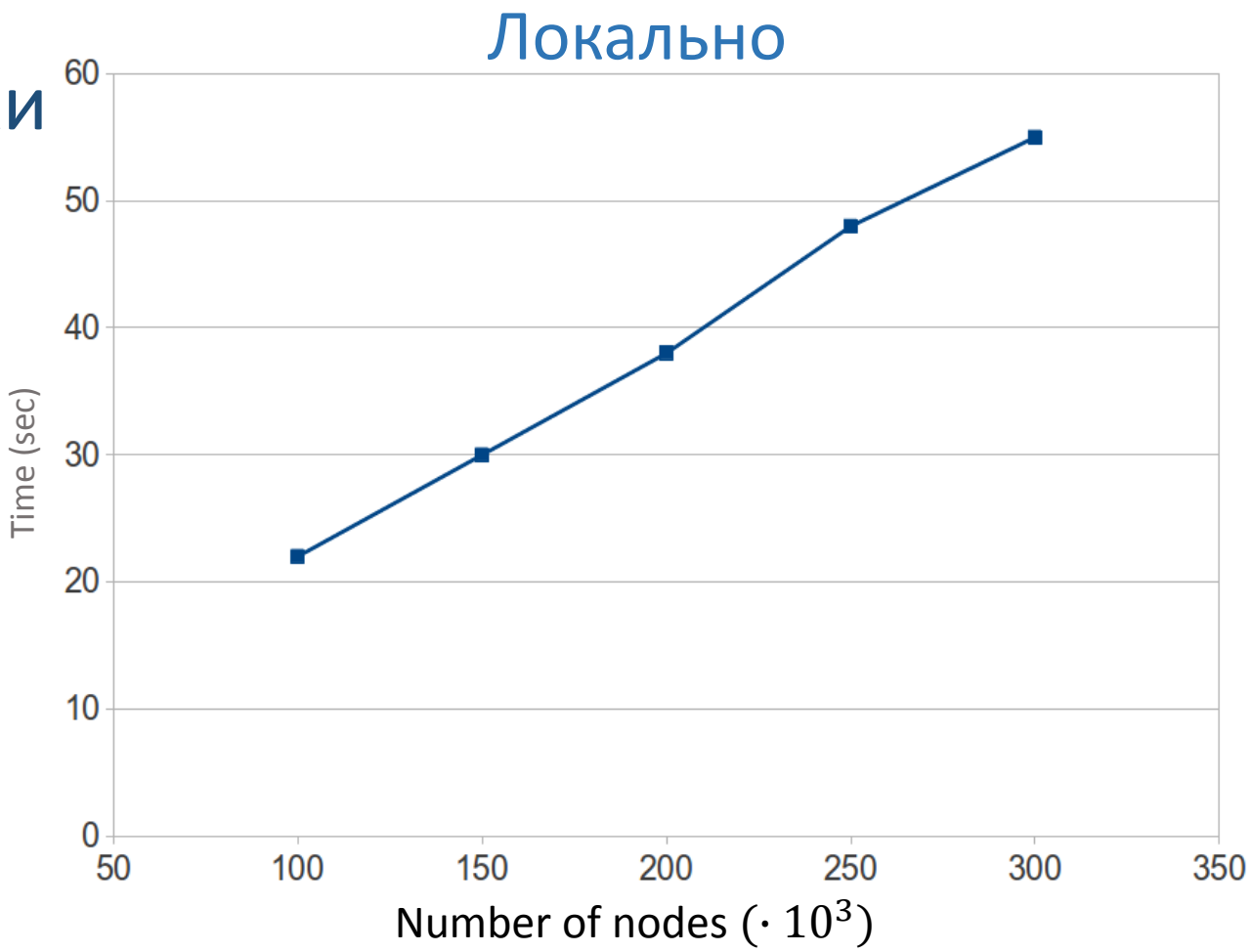
CKB



Масштабируемость

Параметры генерации для оценки масштабируемости:

- $\beta_1 = \beta_2 = 2.5$
- $\alpha = 4, \gamma = 0.5$
- $\min\{m_i\} = 1$
- $\min\{c_i\} = 2$
- $\max\{c_i\} = \max\{m_i\} = 10^4$

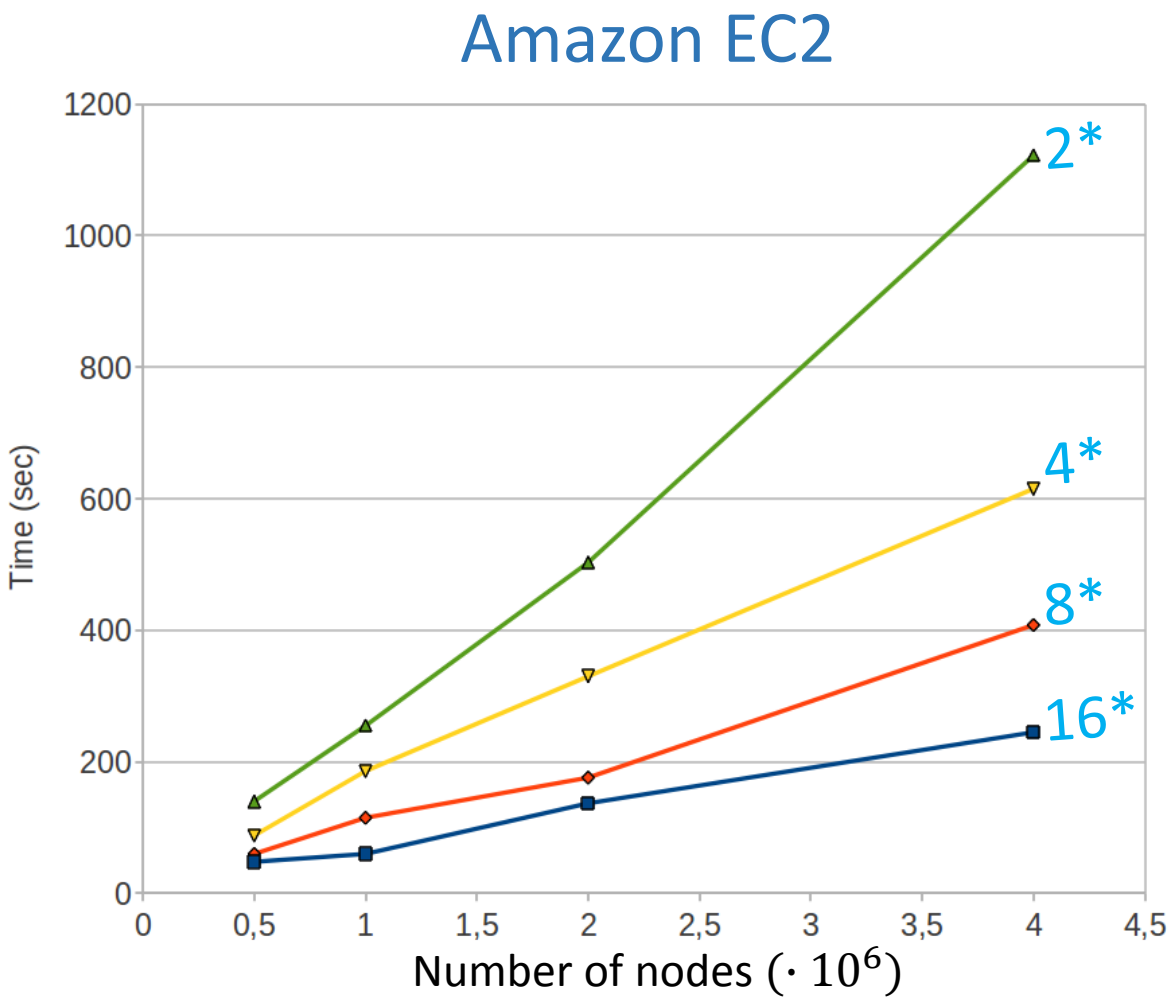


ИСП РАН

Масштабируемость

Параметры генерации для оценки масштабируемости:

- $\beta_1 = \beta_2 = 2.5$
- $\alpha = 4, \gamma = 0.5$
- $\min\{m_i\} = 1$
- $\min\{c_i\} = 2$
- $\max\{c_i\} = \max\{m_i\} = 10^4$



*Числа на конце линий обозначают количество машин m1.large** в кластере, использовавшимся для генерации.

**m1.large – тип машины на кластере Amazon EC2 (2 vCPU, 7.5 GiB memory, 2x420GB instance storage).

ИСП РАН

План

- Немного о Spark
- Постановка задачи
- Предыдущие работы
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Выводы

СКВ генерирует реальные сети и имеет ряд преимуществ:

- Реалистичная структура сообществ
- Линейная масштабируемость
- Генерация миллиардного графа занимает приемлемое время (150 минут на 150 машинах)
- Гибкость модели

ИСП РАН

Открытые вопросы

- Тестирование различных алгоритмов поиска сообществ
- Поддержка вариации коэффициента кластеризации
- Направленная, взвешенная и иерархическая модификации
- Генерации атрибутов пользователей

ИСП РАН

Заключение

Спасибо за внимание!

Вопросы?

chykhradze@ispras.ru