Использование модели социальной сети с сообществами пользователей для распределённой генерации случайных социальных графов

Институт системного программирования РАН

Кирилл Чихрадзе

10-я Международная конференция
«Интеллектуализация обработки информации-2014»

9 октября, 2014
Крит, Греция

ИСП РАН

План

- Постановка задачи
- Сообщество и его свойства
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

План

- Постановка задачи
- Сообщество и его свойства
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Постановка задачи

Сгенерировать случайный граф…

1. …который будет удовлетворять основным свойствам социальных сетей;
2. …сообщества пользователей которого подчиняются основным свойствам реальной структуры сообществ;
3. …из миллиарда вершин со средней степенью 150 за несколько часов.

ИСП РАН

Актуальность

1. Размеры современных социальных сетей измеряются сотнями миллионов пользователей.
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.

ИСП РАН

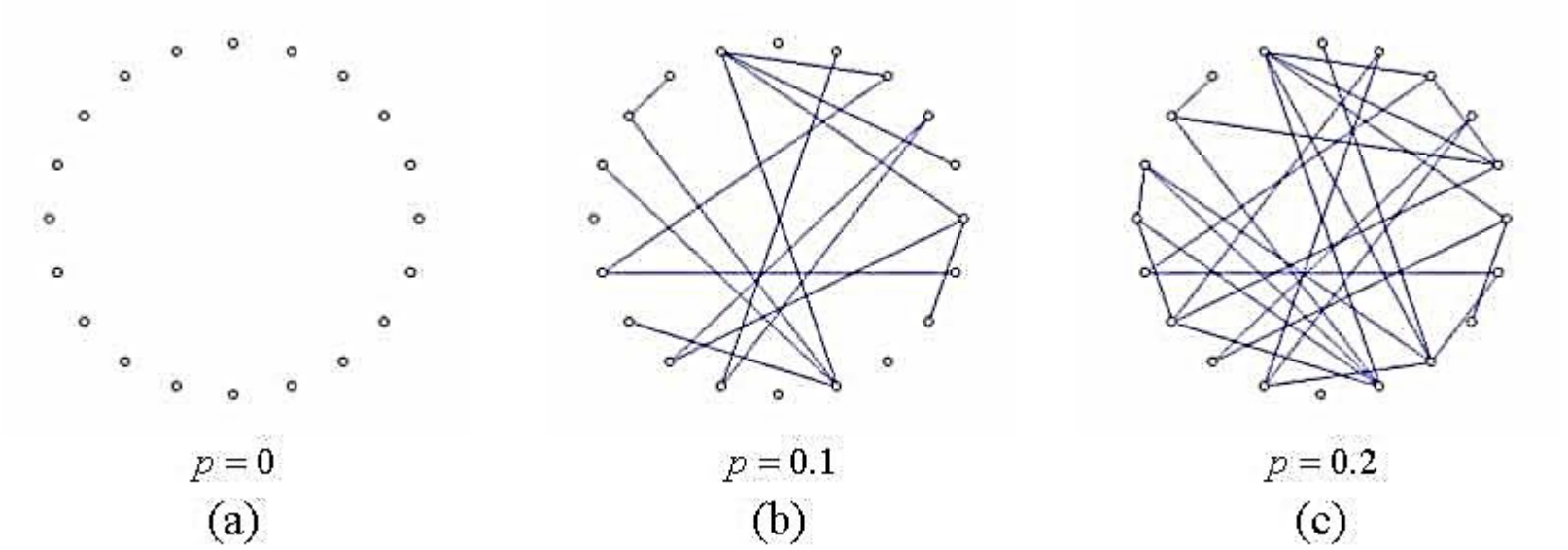
План

- Постановка задачи
- Сообщество и его свойства
- Описание алгоритма
- Оценка точности
- Выводы

Что такое случайный граф?

Граф Эрдёша-Реньи:

- N вершин
- Ребро появляется с вероятностью p



Источник: P. Perez, D.F. Batten, “Complex Science for a Complex World: An Introduction”, 2006

ИСП РАН

Что такое социальный граф?

- Степени вершин распределены по степенному закону*
- Малый диаметр
- «Хорошая» кластеризация

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

ИСП РАН

Что такое сообщество?

ИСП РАН

Что такое сообщество?

Мои Фотографии
Мои Видеозаписи
Мои Аудиозаписи
Мои Сообщения
Мои Группы
Мои Новости
Мои Ответы
Мои Закладки
Мои Настройки

Приложения
Документы

Напоминание

Сегодня день рождения Игоря Понурова, Анастасии Лушниковой.

Люди | Новости | Сообщества | Аудиозаписи | Видеозаписи

Рекомендуемые сообщества

Cosmopolitan
СМИ
200 678 подписчиков
Подписаться

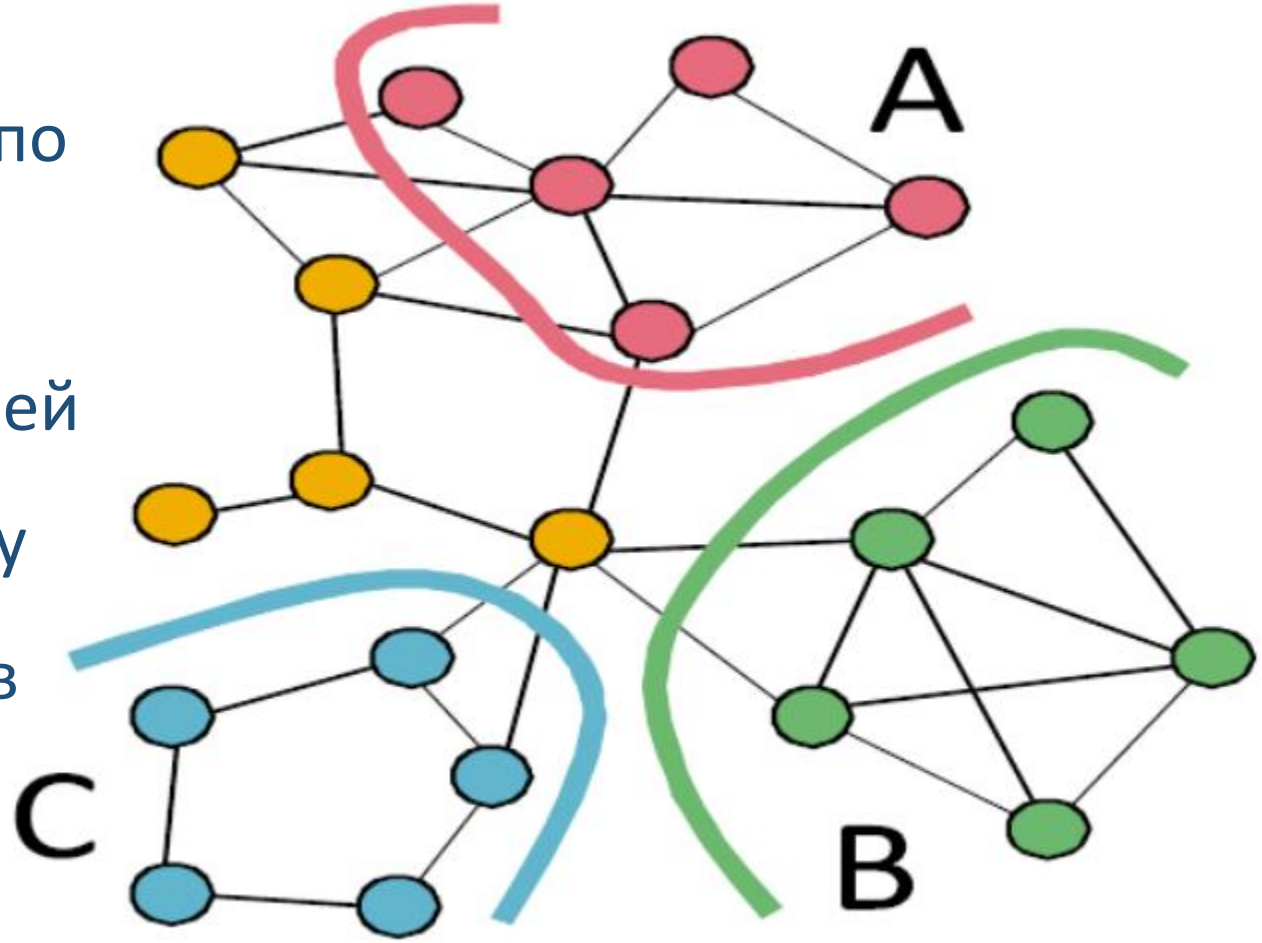
Molodejj.tv — молодёжный интернет-канал
Кинематограф
171 582 подписчика
Подписаться

ФК «Спартак-Москва»
277 911 подписчиков
Подписаться

ИСП РАН

Структурные свойства сообществ

- Размеры сообществ распределены по степенному закону
- Количество вхождений пользователей распределено по степенному закону
- Количество рёбер внутри сообществ растёт суперлинейно с размером сообщества

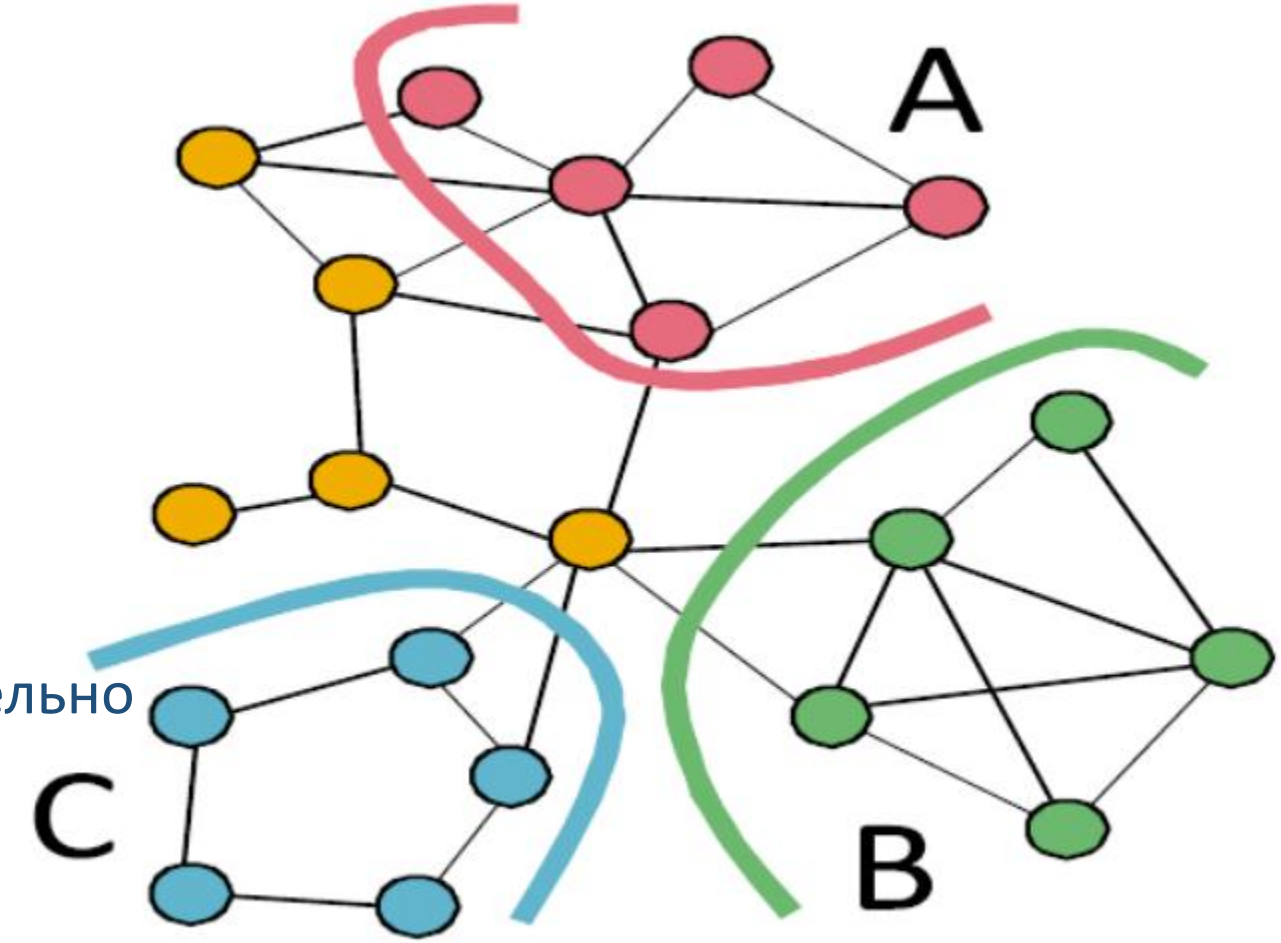


Yang J., Leskovec J., “Structure and overlaps of communities in networks”, arXiv:1205.6228, 2014

ИСП РАН

Структурные свойства сообществ

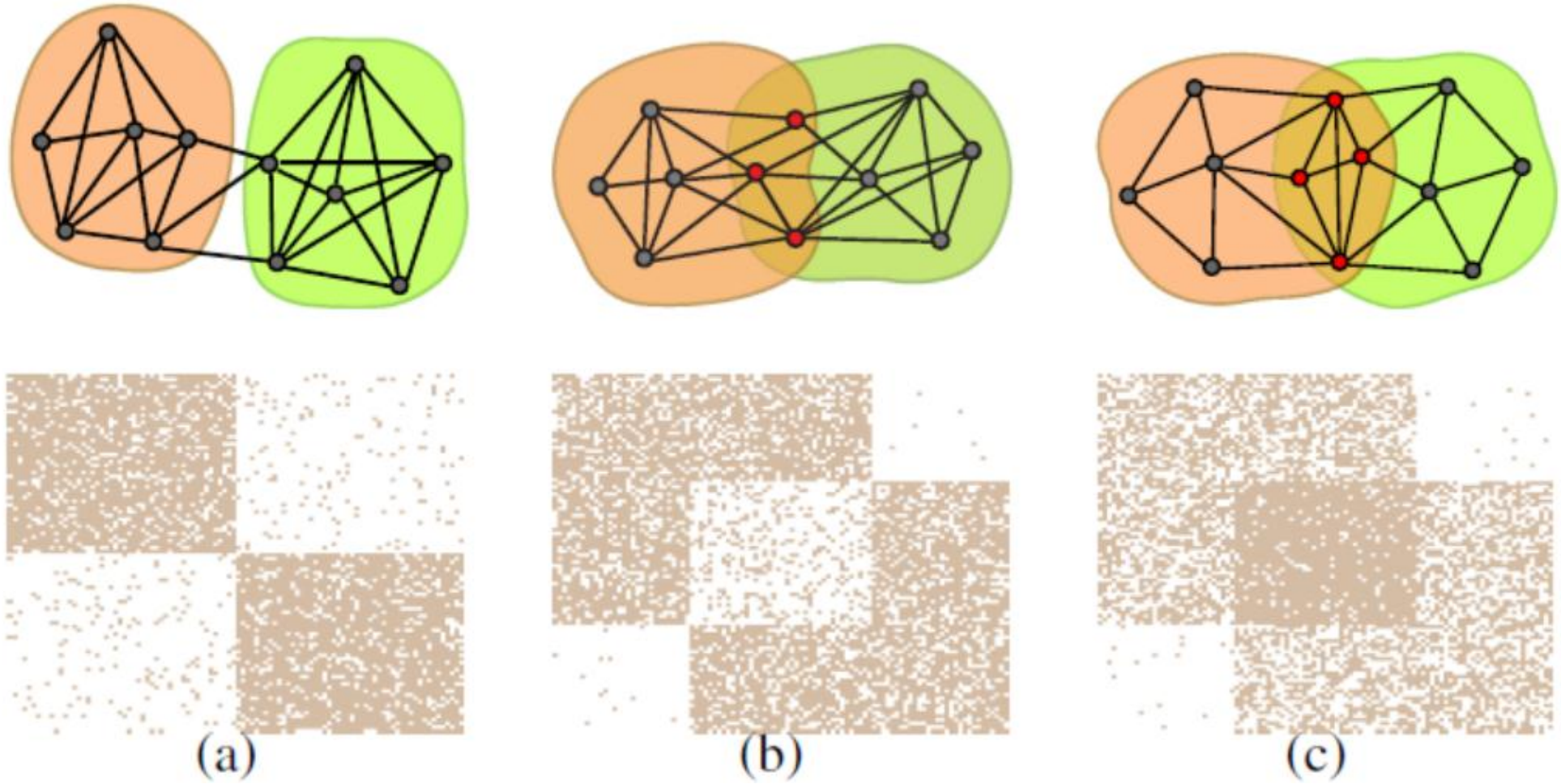
- **Отделимость:** сообщества хорошо отделены от остальной сети
- **Плотность:** сообщества обладают большей связностью
- **Сплочённость:** сообщество относительно трудно разделить на подсообщества



Yang J., Leskovec J., “Structure and overlaps of communities in networks”, arXiv:1205.6228, 2014

ИСП РАН

Структурные свойства сообществ



(a) – нет пересечения
(b) – разреженное пересечение
(c) – плотное пересечение

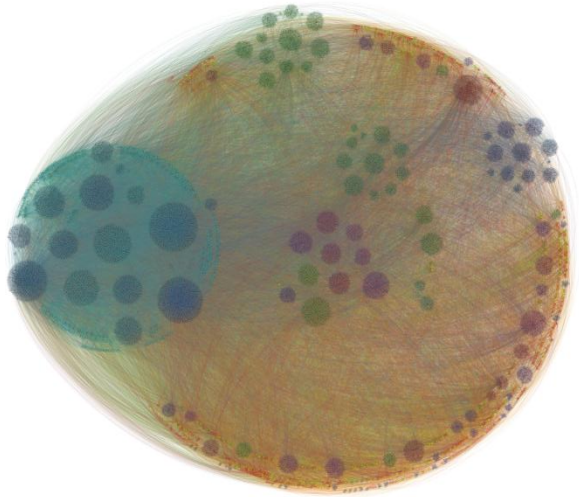
ИСП РАН

План

- Постановка задачи
- Сообщество и его свойства
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Входные параметры

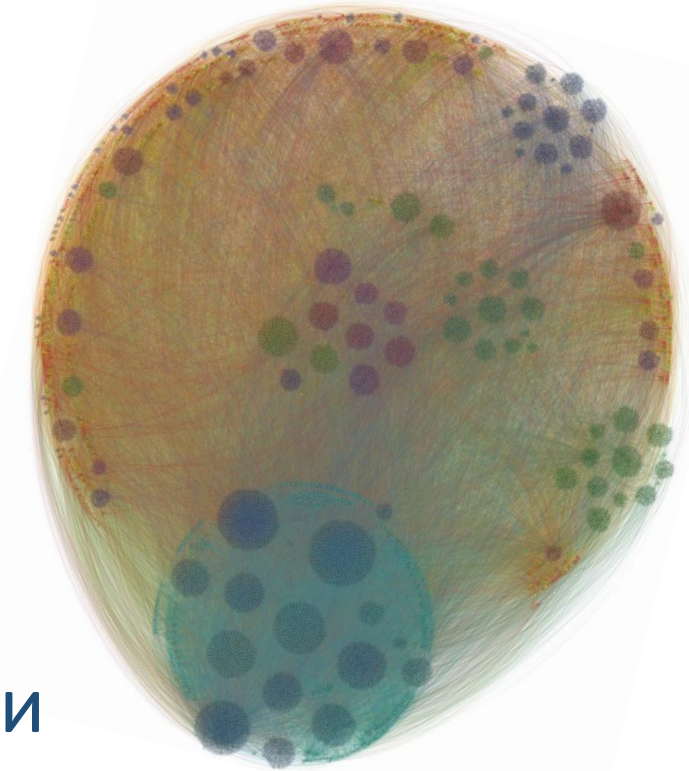


- N_1 — количество вершин
- d_{mean} — средняя степень
- β_1, β_2 — две экспоненты степенного распределения
- Минимальные и максимальные размер сообществ и количество вхождений вершин в сообщество
- α, γ — две константы, определяющие вероятность ребра
- ε — вероятность ребра между сообществами

ИСП РАН

Основные шаги

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

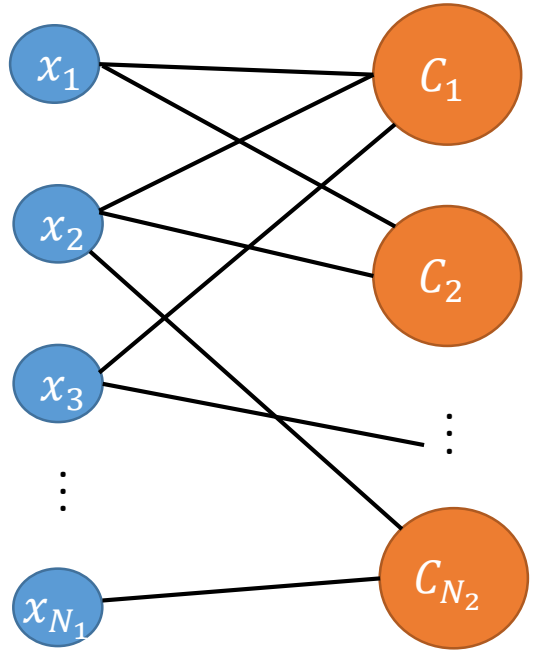


Вершина-сообщество

- 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

ИСП РАН

Генерация рёбер

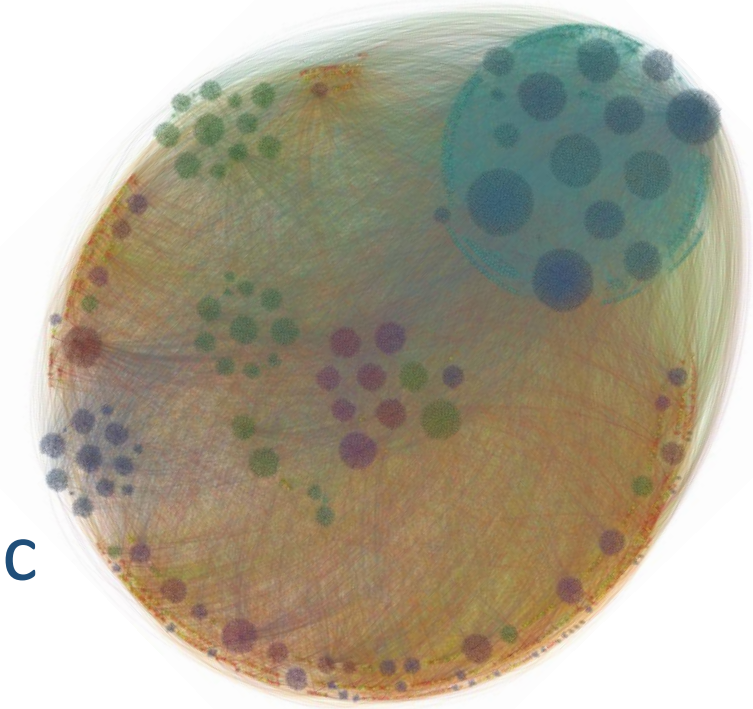
- Рёбра в сообществах генерируются с вероятностью*:

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

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

Генерация рёбер

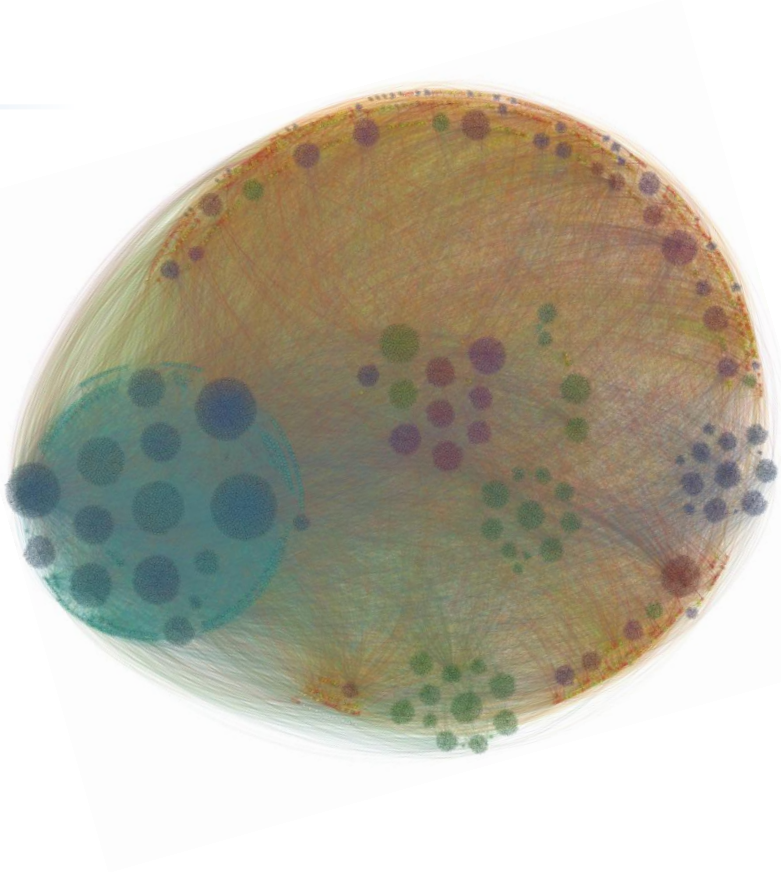
- Количетсво рёбер в сообществе:

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

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

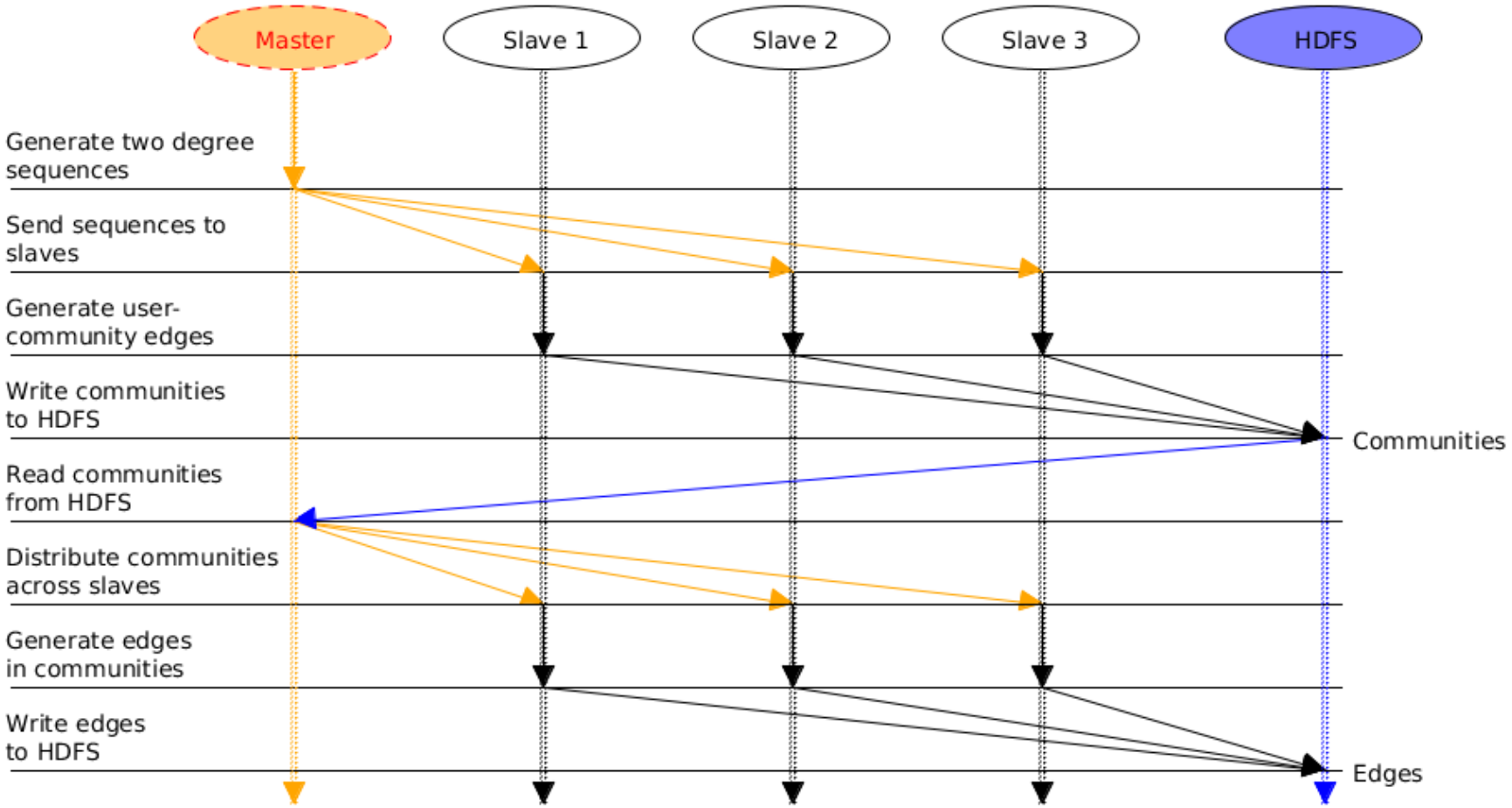
- Количество рёбер между сообществами:

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



ИСП РАН

Реализация Apache Spark



ИСП РАН

План

- Постановка задачи
- Сообщество и его свойства
- Описание алгоритма
- Оценка точности
- Выводы

ИСП РАН

Сравнение с реальными графами

	LiveJournal	СКВ
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

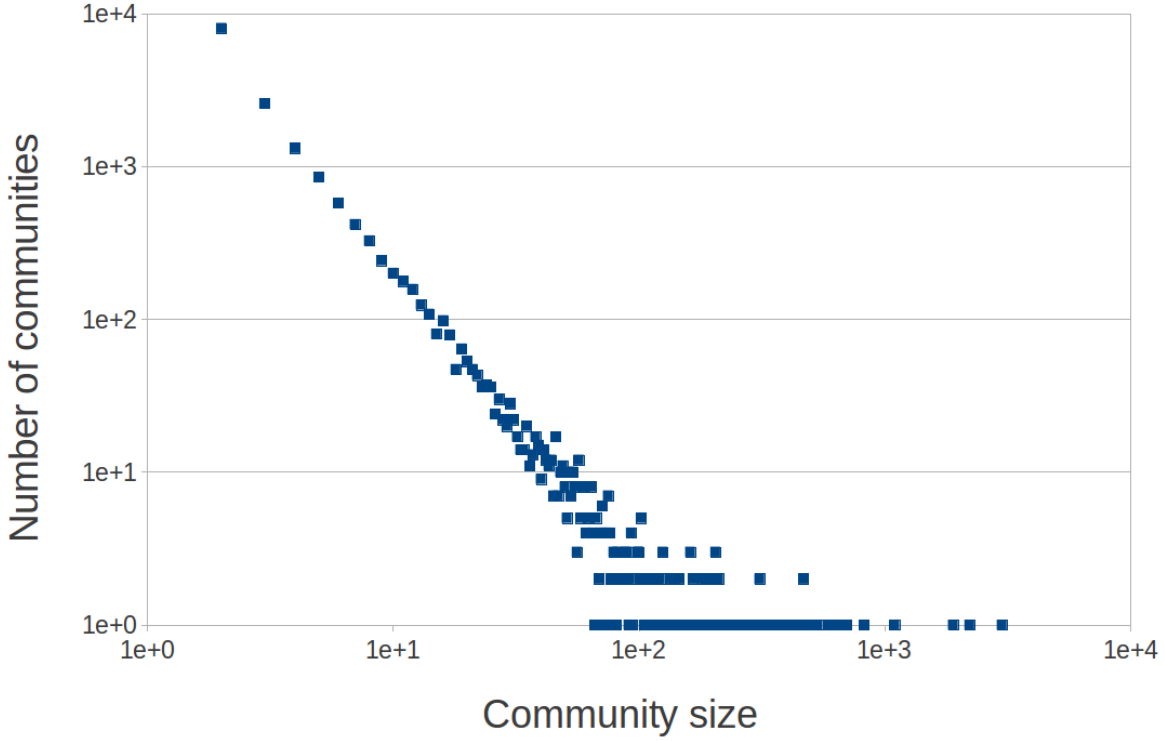
Сравнение с реальными графами

	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

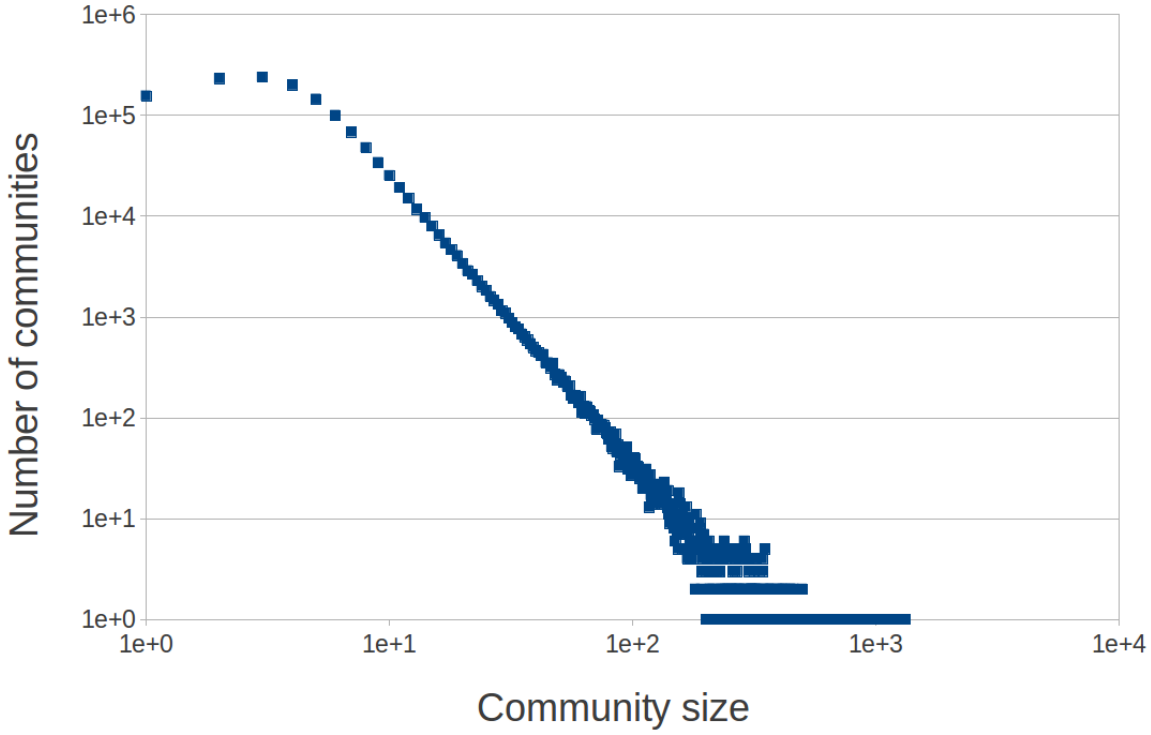
ИСП РАН

Сравнение с реальными графами

YouTube



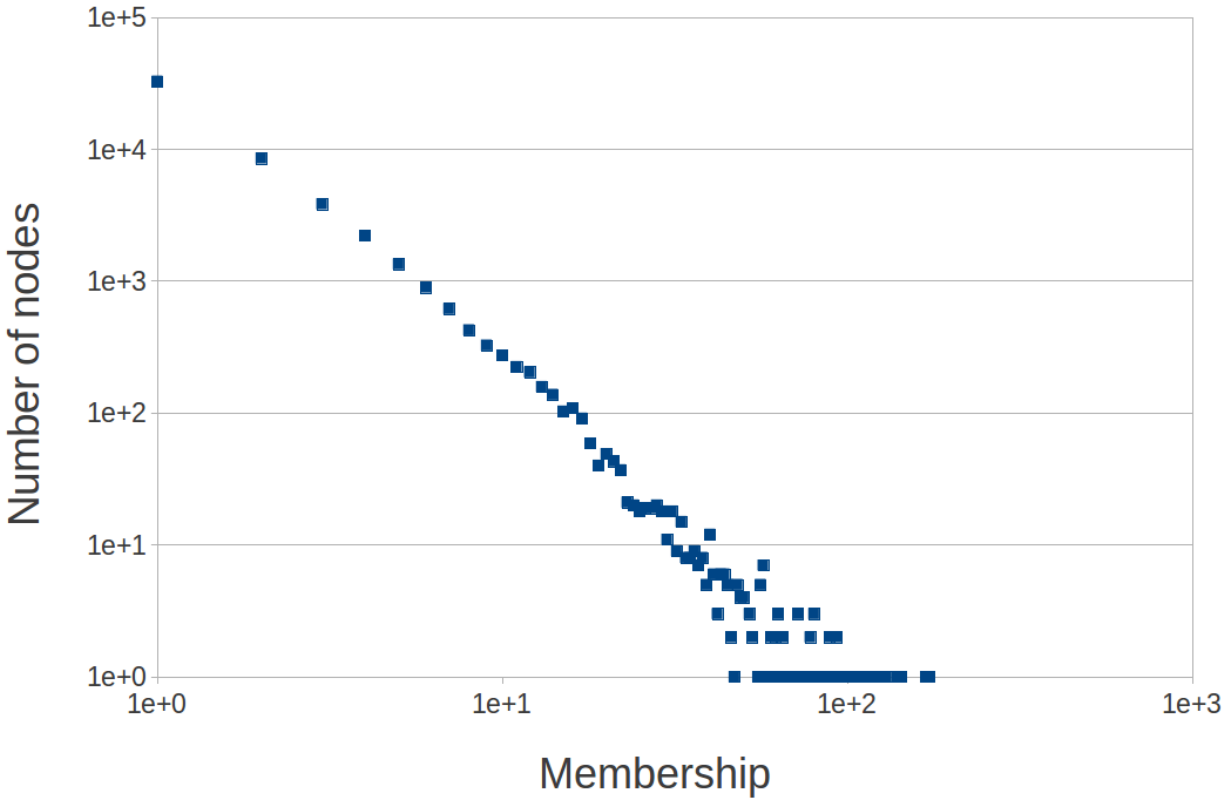
CKB



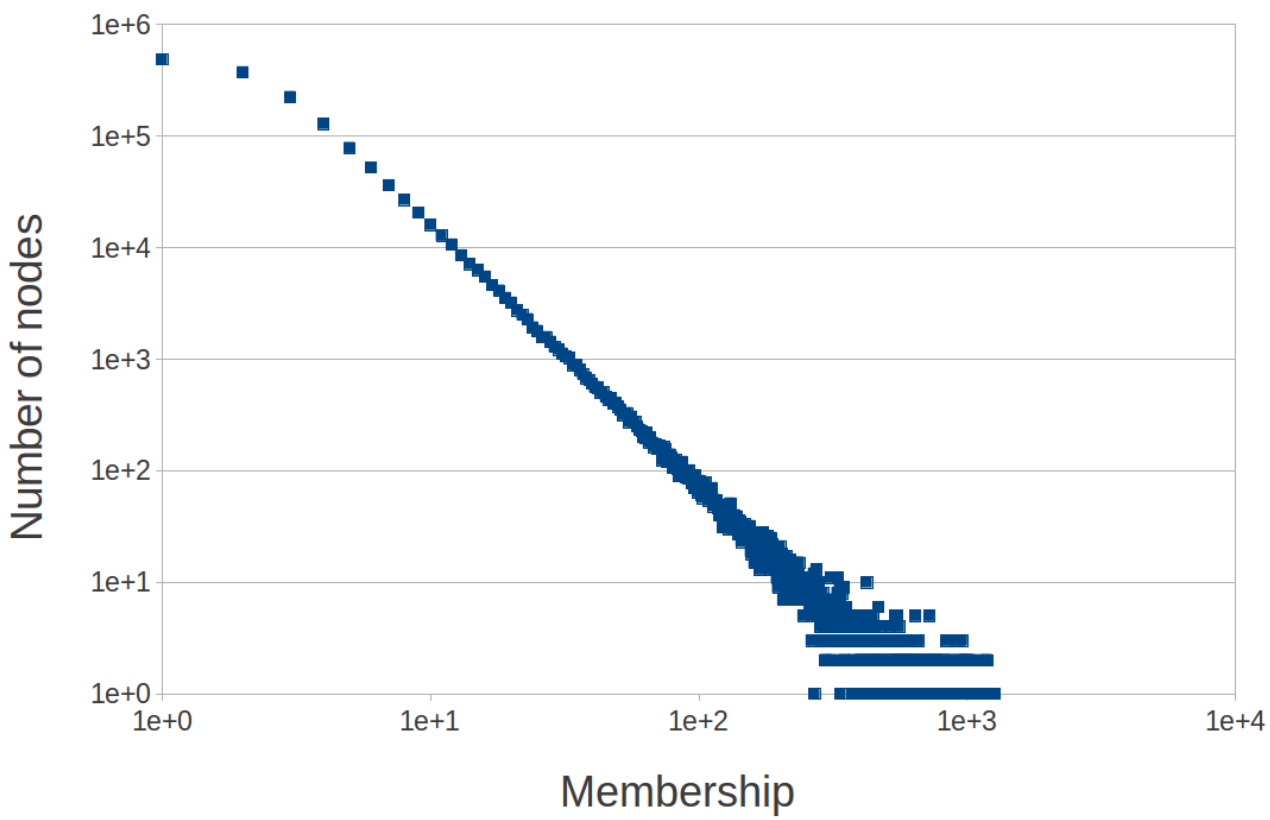
ИСП РАН

Сравнение с реальными графами

YouTube

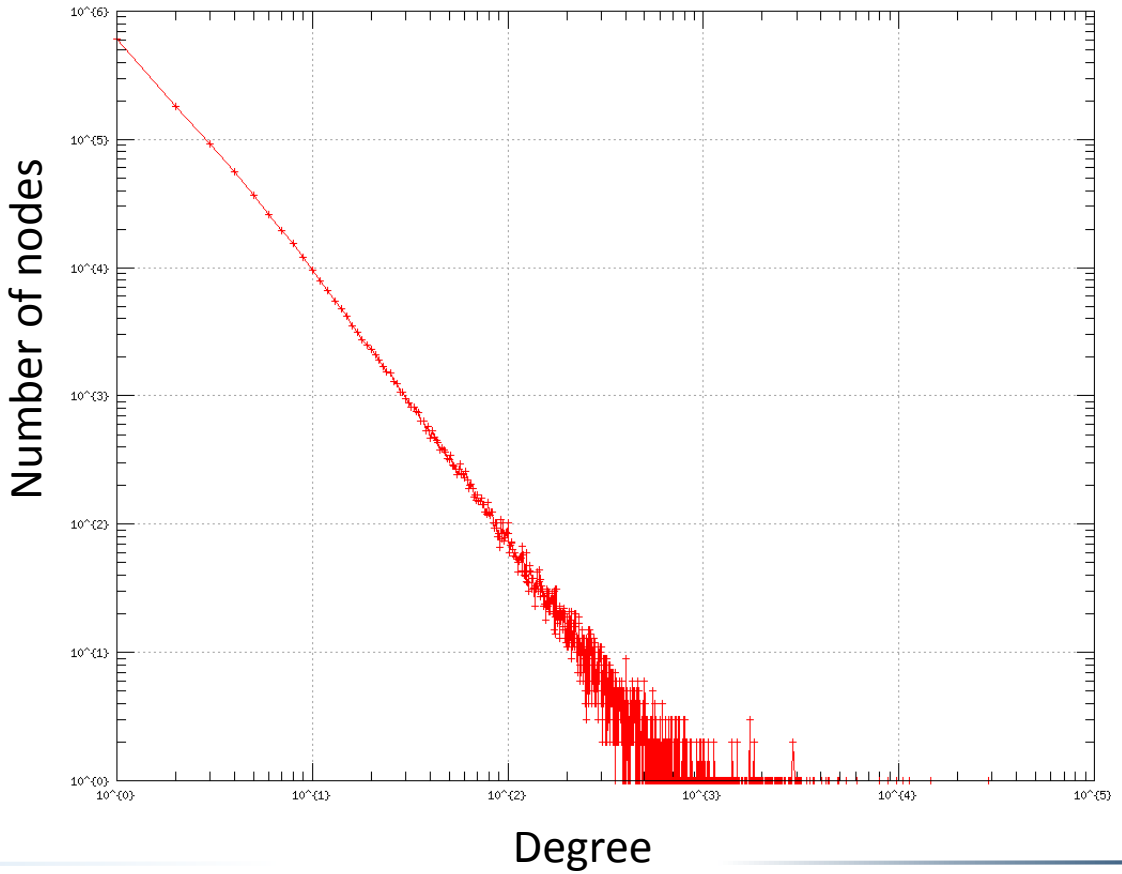


CKB

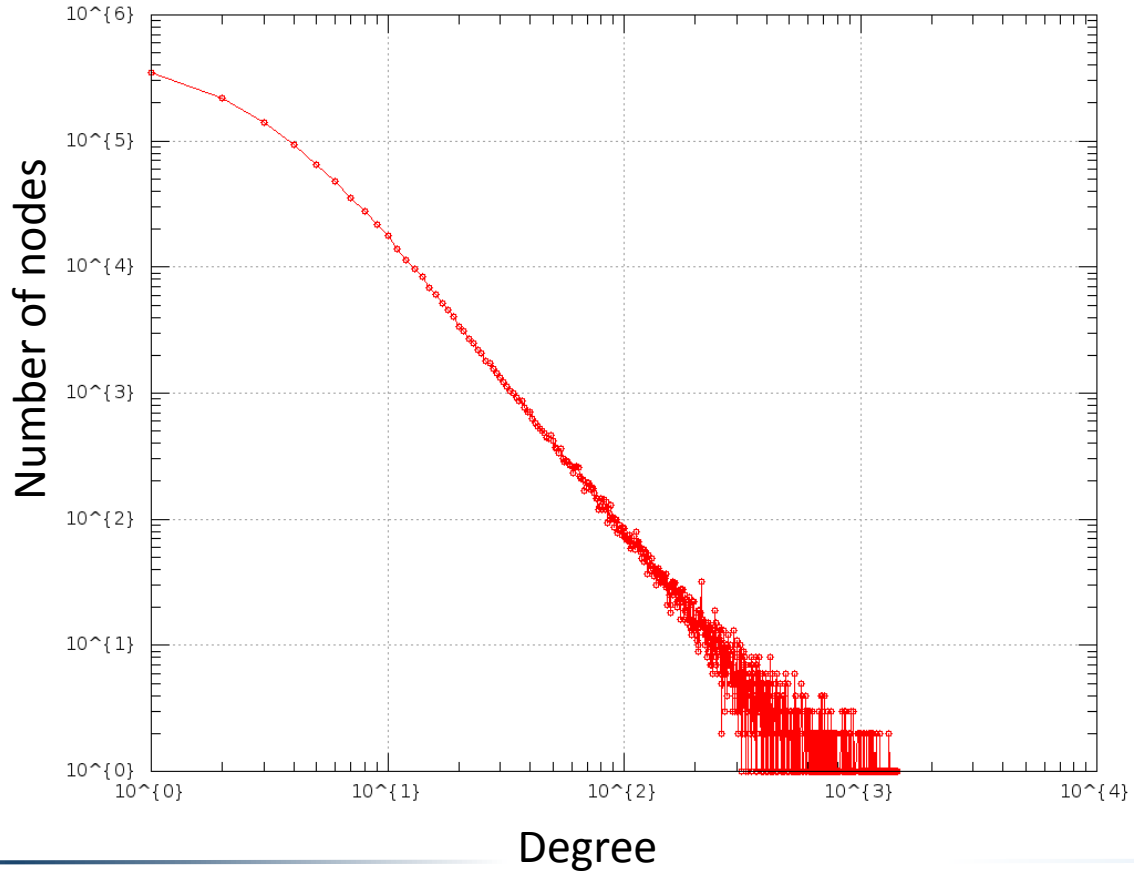


Сравнение с реальными графами

YouTube



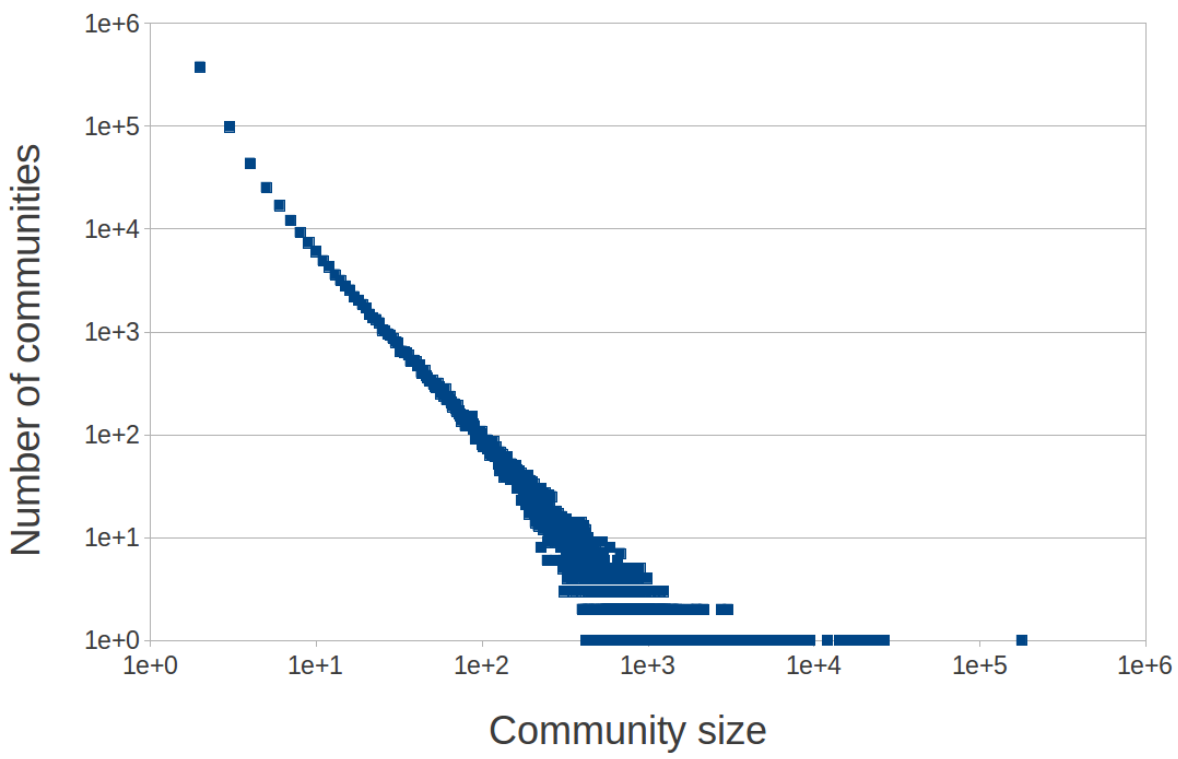
СКВ



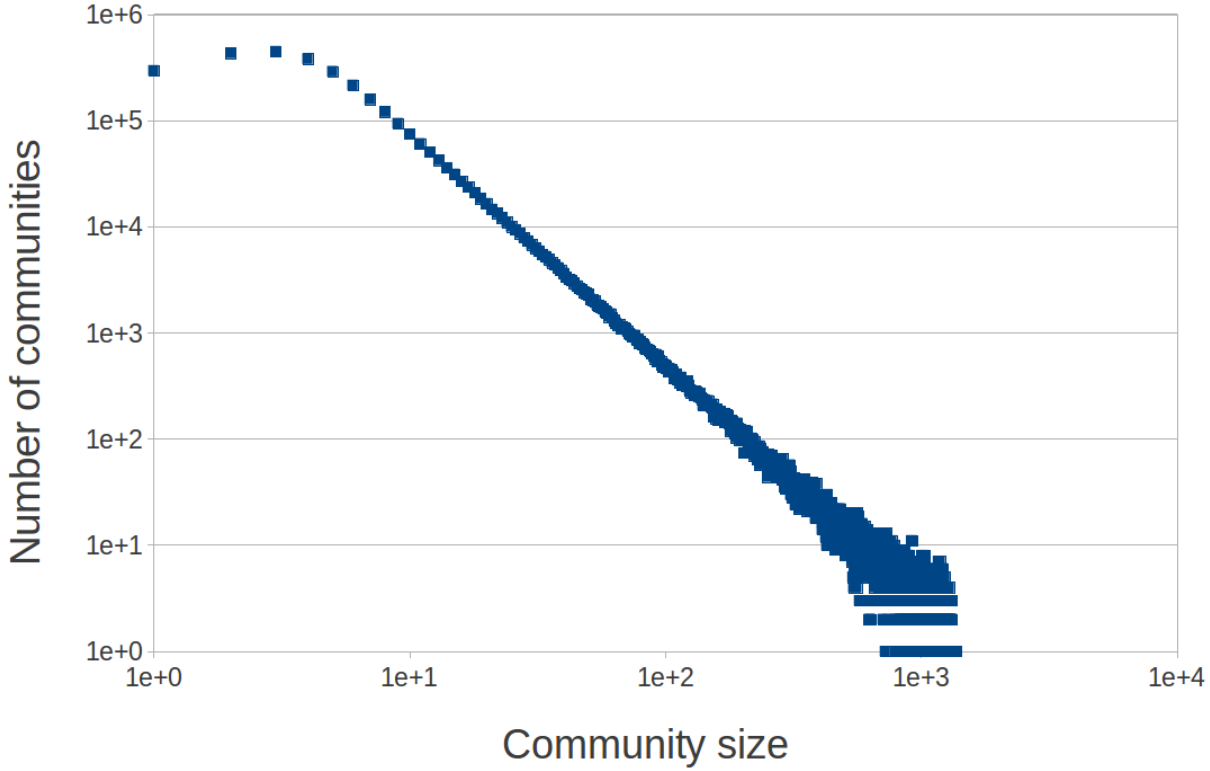
ИСП РАН

Сравнение с реальными графами

LiveJournal



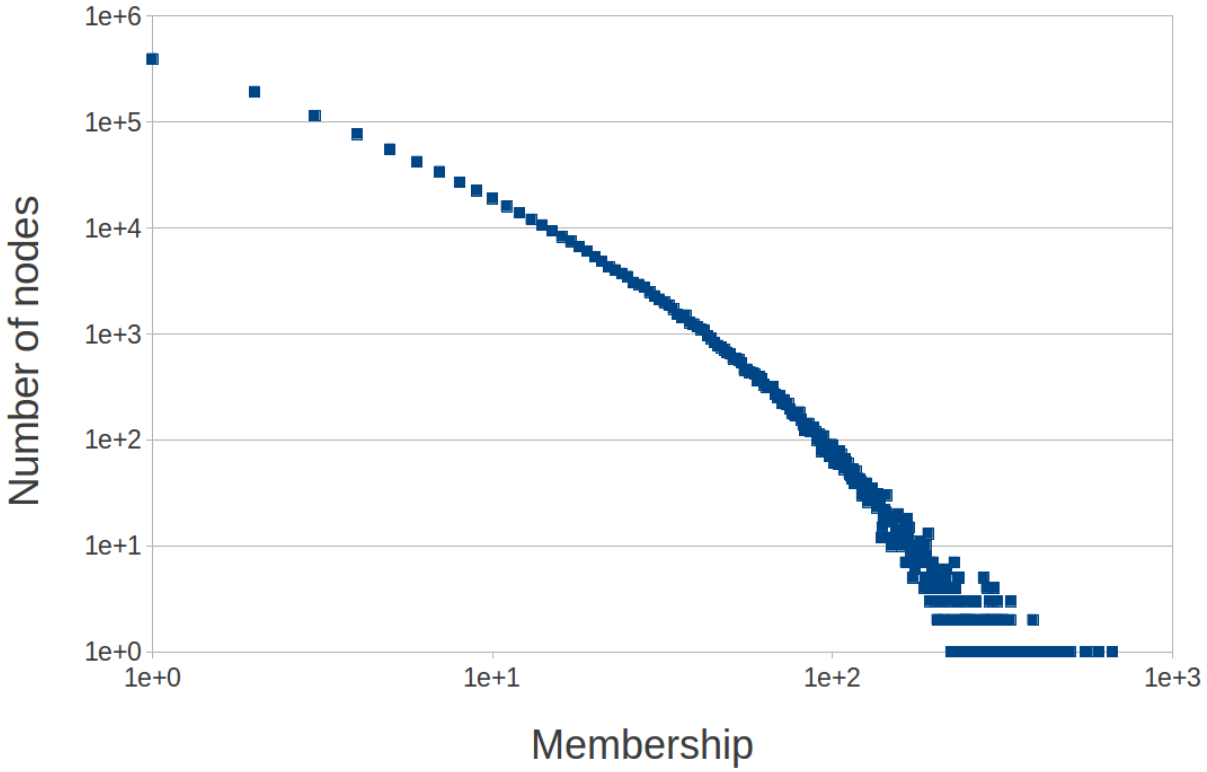
CKB



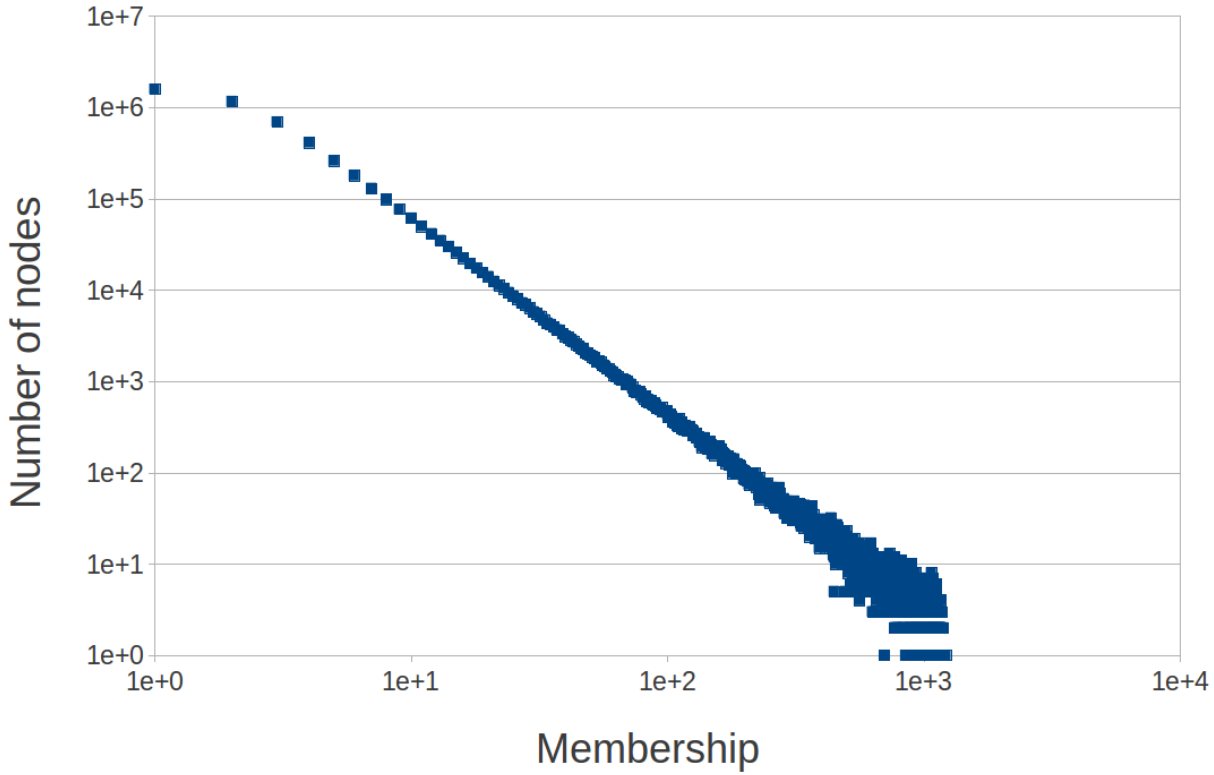
ИСП РАН

Сравнение с реальными графами

LiveJournal



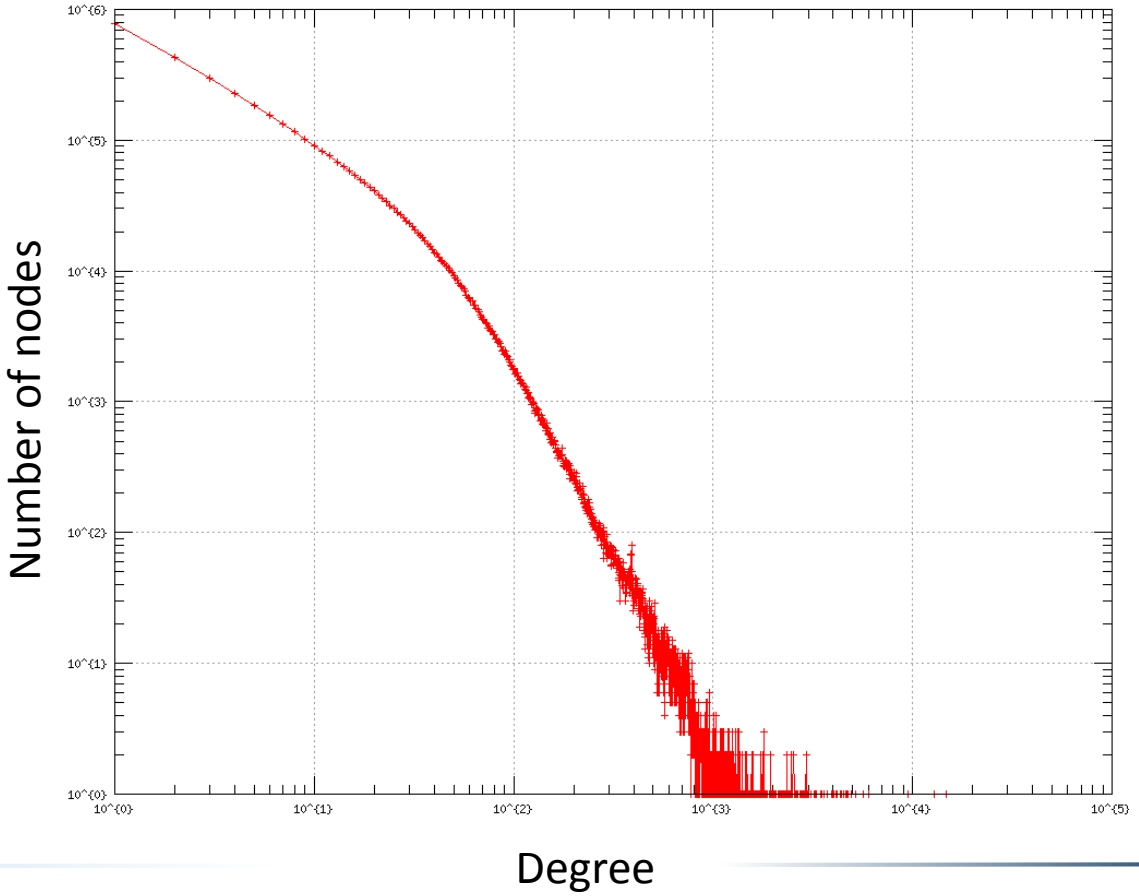
СКВ



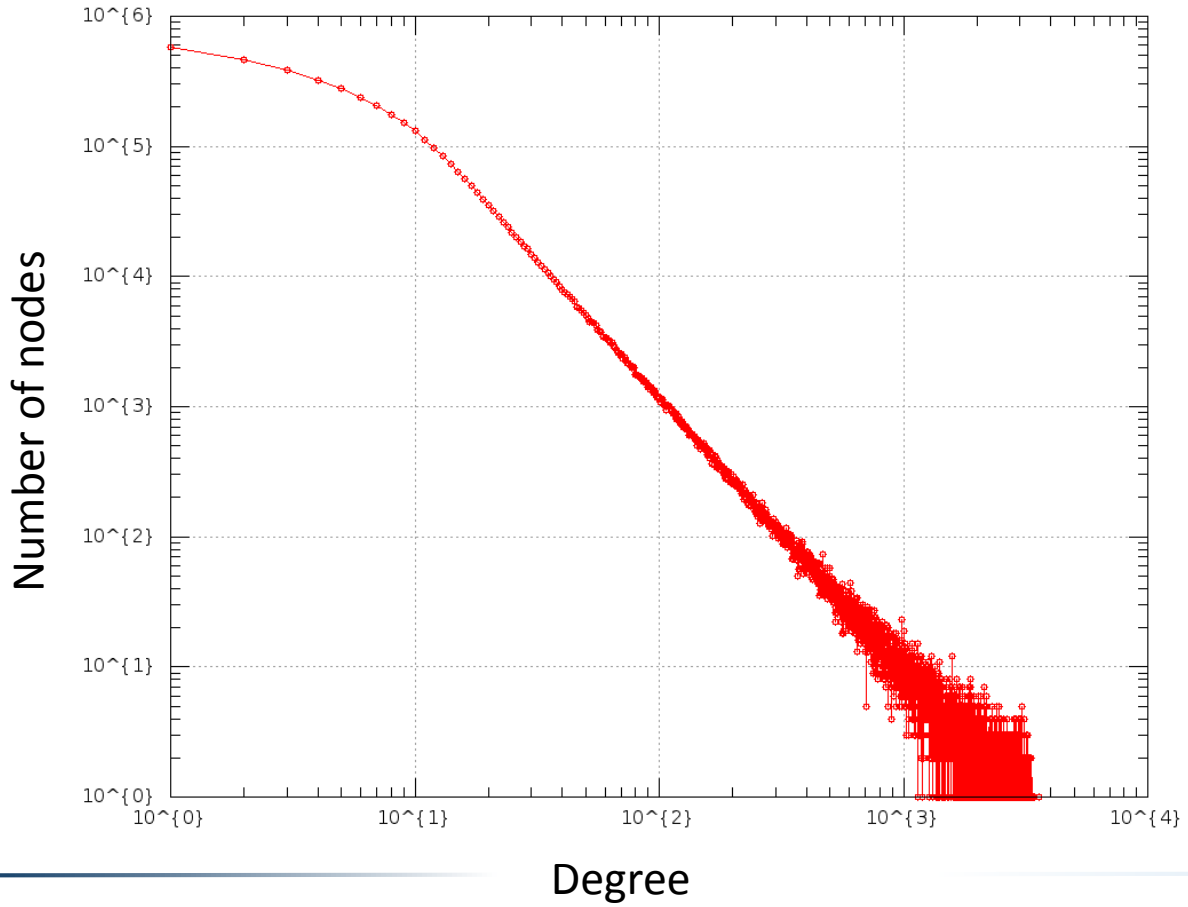
ИСП РАН

Сравнение с реальными графами

LiveJournal

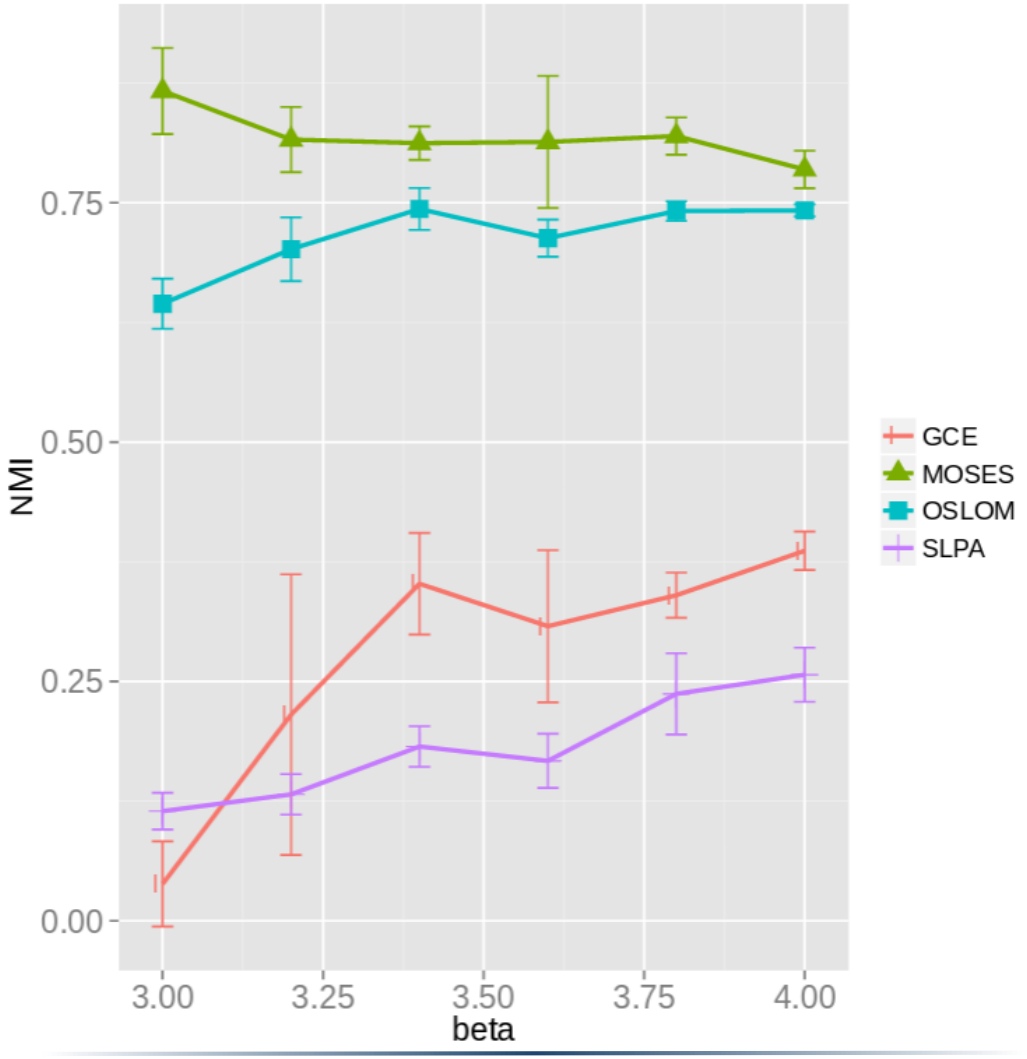


СКВ



ИСП РАН

Восстановление структуры сообщества



ИСП РАН

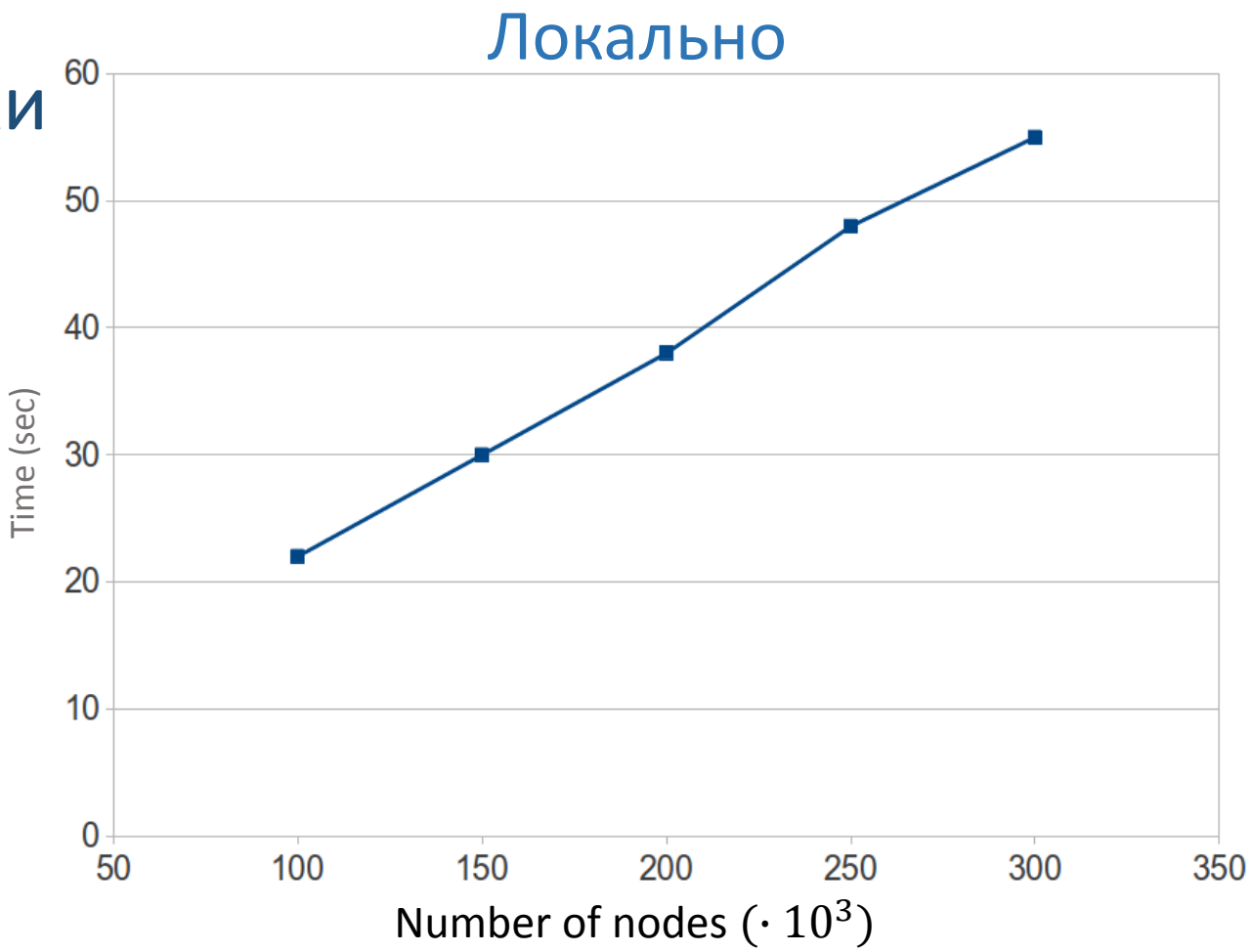
Восстановление структуры сообщества: MOSES

N_1	m_{min}	x_{min}	x_{max}	m_{max}	β_1	β_2	α	γ	ε	NMI	d_{mean}
10,000	1	2	1000	1000	2.5	2.5	1	0.3	0	0.59	67.4
10,000	1	2	1000	1000	3.1	3.1	1	0.3	0	0.72	32.3
10,000	1	2	1000	1000	3.1	3.1	2	0.3	0	0.75	57.3
10,000	1	10	1000	1000	2.5	2.5	1	0.5	0	0.88	90.8
10,000	1	10	1000	1000	2.5	2.5	1	0.5	0.0005	0.84	107.1
10,000	2	10	1000	1000	2.5	2.5	0.5	0.5	0	0.52	72.0
10,000	1	10	1000	1000	2.5	2.5	0.2	0.2	0	0.65	88.9

Производительность

Параметры генерации для оценки масштабируемости:

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

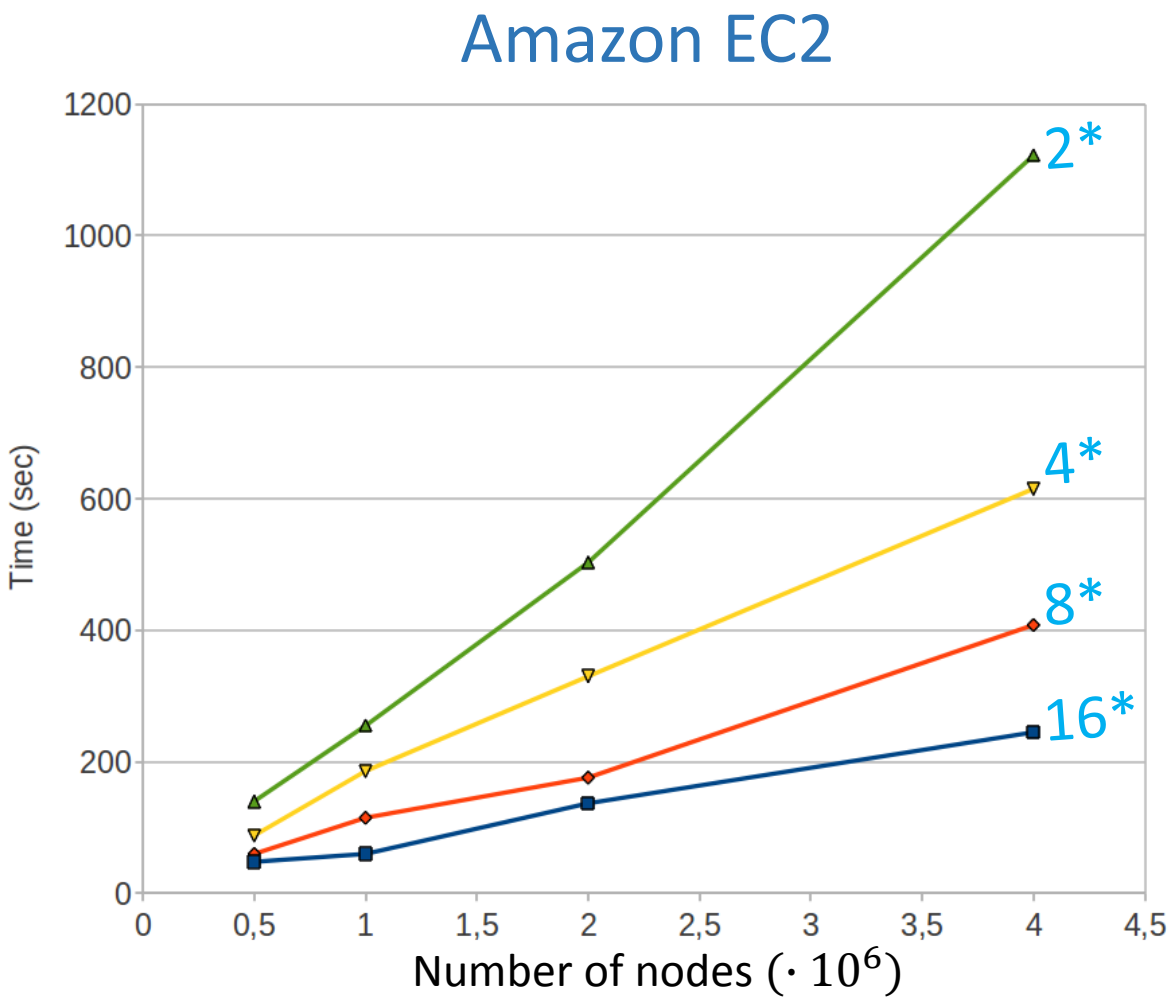


ИСП РАН

Производительность

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

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