Методы и программные средства моделирования и генерации сложных сетей с сохранением графовых свойств

Дробышевский Михаил Дмитриевич

Научный руководитель:
Турдаков Денис Юрьевич

Специальность:
05.13.11 – “математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”

Институт системного программирования им. В. П. Иванникова РАН

декабрь 2019

Актуальность

Сложные сети: графы *реального мира* с нетривиальными топологическими свойствами.

Насколько надежна сеть Интернет? Как устроены общественные отношения, отраженные в социальных сетях? Какие законы управляют распространением болезней и информационными потоками и как ими можно управлять?

Модель случайного графа: вероятностное распределение над множеством возможных графов $\{\mathbb{P}_\theta(G), G \in \mathcal{G}, \theta \in \Theta\}$. Попытка угадать механизмы формирования структуры сложной сети; воспроизвести графовые свойства (признаки).

Этапы развития теории случайных графов:

- модель Эрдеша-Реньи (1959г)
- безмасштабные сети: Барабаши-Альберт (1999г)
- современные модели: LFR (2009г), CKB (2014г), SKG (2010г)

Проблема: выбор параметров генратора $\rightarrow$ необходимо обучение

Основные приложения: анонимизация графовых данных; создание тестовых выборок; экстраполяция; нулевые модели

Метод сравнения графов — по набору известных характеристик

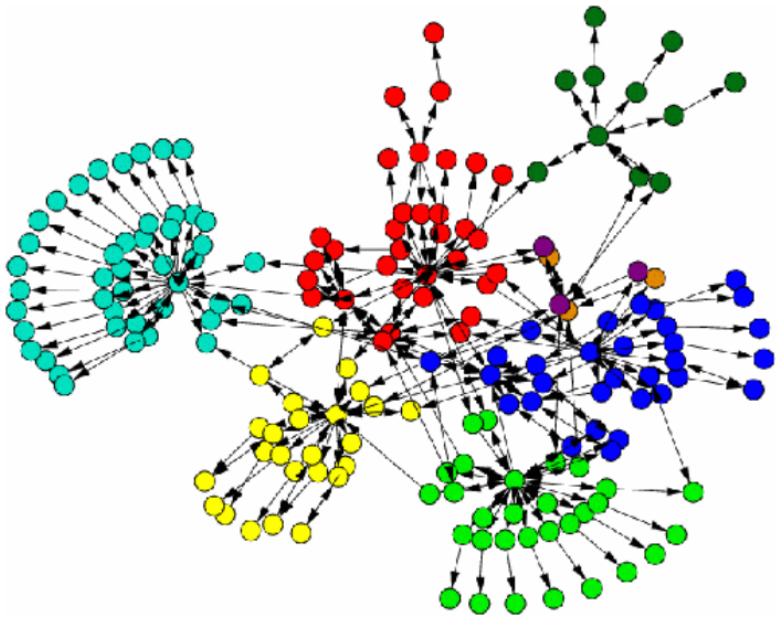
- числовые: средняя степень вершины, взаимность (reciprocity) ребер, ассортативность степеней вершин, средний коэффициент кластеризации, эффективный диаметр гигантской компоненты, спектральный радиус;
- распределения: распределение степеней вершин, кумулятивный средний коэффициент кластеризации, коэффициент кластеризации как функция от степени вершины, распределение подграфов размера 3, достижимость вершин (hop plot).

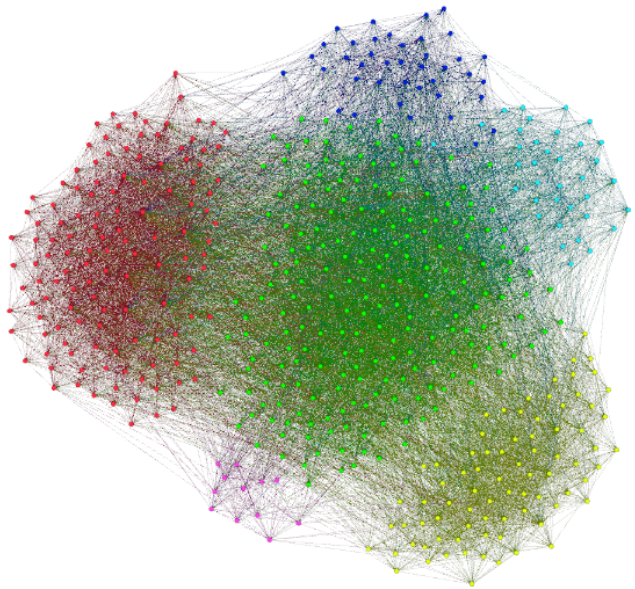
Цель: модель генерации графов контролируемого размера,

1. *похожих* на исходный: отклонения по каждой характеристике не выше соответствующих отклонений у других современных методов, и
2. обладающих необходимой *вариабельностью*: разброс значений числовых характеристик близок к соответствующему разбросу у реальных графов из одного домена.

Дополнительные особенности графов:

- во многих доменах ребра графов **направленные** (звонки, цитирования и т.д.);
- графы могут иметь выделенную **структуру сообществ**, определяющую высокоуровневую организацию вершин в группы со схожими функциями, свойствами, ролями и т.д. (сообщества в соцсетях); качество покрытия сообществами измеряется модулярностью;
- ребра могут иметь **веса** (продолжительность/кол-во звонков, кол-во цитирований), выражающие силу связи вершин.





Цель работы:

- На основе анализа существующих моделей случайных графов разработать и реализовать метод генерации случайных графов, удовлетворяющий указанным требованиям;
  - *автоматическое* обучение на заданном графе;
  - возможность генерировать графы контролируемого размера;
  - одновременная поддержка трех особенностей графа: направленные, взвешенные ребра и структура сообществ;
  - *похожесть* генерируемых графов на исходный: отклонения по каждой характеристике не выше соответствующих отклонений у других современных методов;
  - *вариабельность* генерируемых графов: разброс значений числовых характеристик близок к соответствующему разбросу у реальных графов из одного домена.
- Провести экспериментальное исследование разработанного метода на соответствие требованиям, сравнение его с другими методами.

Научная новизна: впервые предложен подход к генерации случайных графов, основанный на вложении графа в пространство малой размерности.

Среди существующих методов нет удовлетворяющих всем требованиям

Модель	**автоматич. обучение**	**контрол. размер**	**направл. ребра**	**взвеш. ребра**	**структура сообществ**
LFR	−	±	+	+	+
SBM	±	±	+	−	+
MFNG	±	+	+	−	−
ReCoN	±	±	+	−	+
SKG[1]	+	±	+	−	−
Gscaler[2]	+	+	+	−	−
ERGG-dwc	**+**	**+**	**+**	**+**	**+**

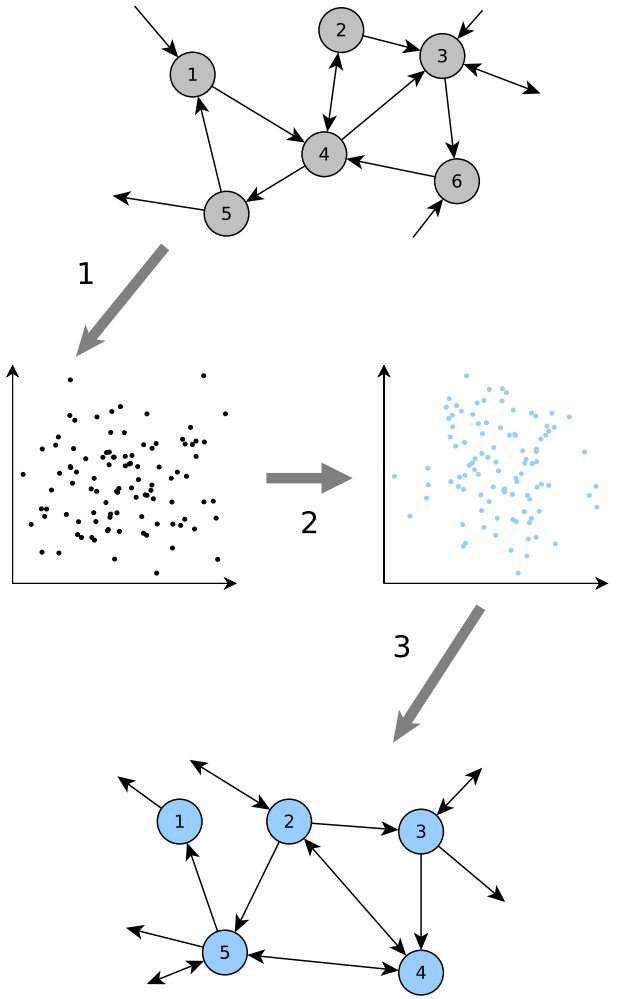
- '+' — поддерживается
- '−' — не поддерживается
- '±' — частично поддерживается

[1] C. L. Staudt, M. Hamann, I. Safro, A. Gutfraind, and H. Meyerhenke. Generating scaled replicas of real-world complex networks. arXiv preprint arXiv:1609.02121, 2016.

[2] Zhang JW, Tay YC. GSCALER: Synthetically Scaling A Given Graph. // EDBT. — 2016. — Pp. 53–64.

Подход на основе вложения графа — ERGG:

1. Получить вложение (*embedding*) графа $G = (N, E)$ в низкоразмерное пространство, так что его узлы $i \in N$ отображаются в вещественные векторы $\{\vec{r}_i\}_{i=1}^{n}$.
2. Аппроксимировать эмпирическое распределение векторов $\{\vec{r}_i\}_{i=1}^{n}$ и сэмплировать набор из n' новых случайных векторов $\{\vec{q}_i\}_{i=1}^{n'}$ из того же вероятностного распределения. Эти векторы будут соответствовать узлам нового графа $(N', \cdot)$.
3. Соединить узлы графа $(N', \cdot)$ ребрами, используя модель вложения из шага 1, получая в результате граф $G' = (N', E')$.



Мотивация:

- вектора вложения кодируют важные *графовые свойства*
- сэмплировать можно *произвольное* число новых вершин

Декомпозиция на 3 независимые подзадачи:

1. Вложение + восстановление

Построить вложение графа, сохраняющее максимальное количество информации: так, что можно восстановить ребра с F_1-мерой 0.99

2. Аппроксимация распределения + сэмплирование

Сэмплировать новые вектора из того же распределения так, чтобы свойства графа сохранялись

3. Атрибуты: метки сообществ и веса ребер

Новая структура сообществ и веса ребер должны быть корректно определены и согласованы между собой

Разработан метод **ERGG-dwc** — решение подзадач в рамках подхода ERGG (от англ. **d**irected **w**eighted **c**ommunities)

Тренировочная коллекция направленных графов

Описание графа	**имя**	**n**	**m**	**d**
Каратэ-клуб Zachary	Karate	34	78	3
Транскрипция генов дрожжей	Yeast	688	1079	9
Мобильные звонки	VAST	400	1562	12
Пищевые цепочки Флорида Бэй	Foods	128	2106	8
Эго-сеть из Твиттер	TW	146	1309	12
Синтетический Кронекеровский граф	Kron	2187	11675	24
Смежность слов в японских текстах	Words	2704	8300	17
Синтетический ER-граф	ER	800	8000	26
Эго-сеть из Google-plus	G+	1243	106485	62

- n — число вершин
- m — число ребер
- d — оптимальная размерность вложения

1. Вложение + восстановление

Цель: разработать метод вложения, такой что F_1-мера восстановленных ребер $\geqslant 0.99$ для любого графа

Решение: модификация COMBO на основе методов BLM[3] и LINE[4].

- Вектор узла $\vec{r_i} = \begin{bmatrix} \vec{u_i} & \vec{v_i} & Z_i \end{bmatrix}^T, i \in N$
- Функция оценки $s_{ij} = s(\vec{r_i}, \vec{r_j}) = \vec{u_i} \cdot \vec{v_j} - Z_i$
- Главное отличие: из шумового распределения отфильтровываются ребра графа. Теперь цель — отличить ребра от не-ребер
  $s_{ij}|(i \to j) \in E \geqslant t_G \geqslant s_{ij}|(i \to j) \notin E$
- Также изменялись другие компоненты алгоритмов для максимизации F_1-меры.

---

[3] *Ivanov Oleg U, Bartunov Sergey O.* Learning representations in directed networks // International Conference on Analysis of Images, Social Networks and Texts / Springer. – 2015. – Pp. 196–207.

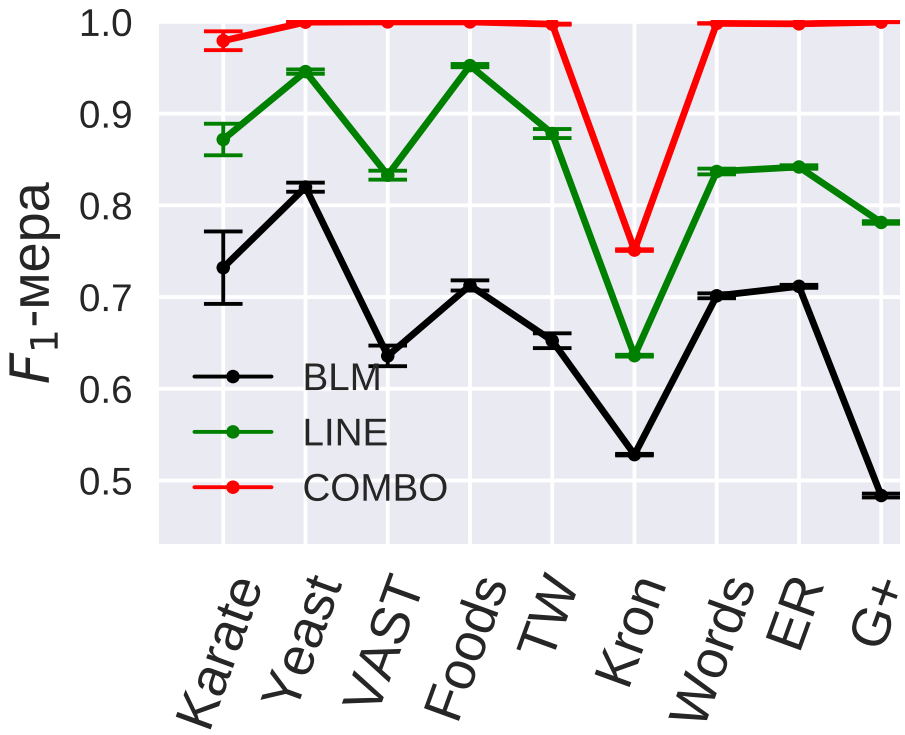
[4] Line: Large-scale information network embedding / Jian Tang, Meng Qu, Mingzhe Wang et al. // Proceedings of the 24th International Conference on World Wide Web / ACM. – 2015. – Pp. 1067–1077.

1. Вложение + восстановление

Сравнение компонентов алгоритмов BLM, LINE и модификации COMBO.

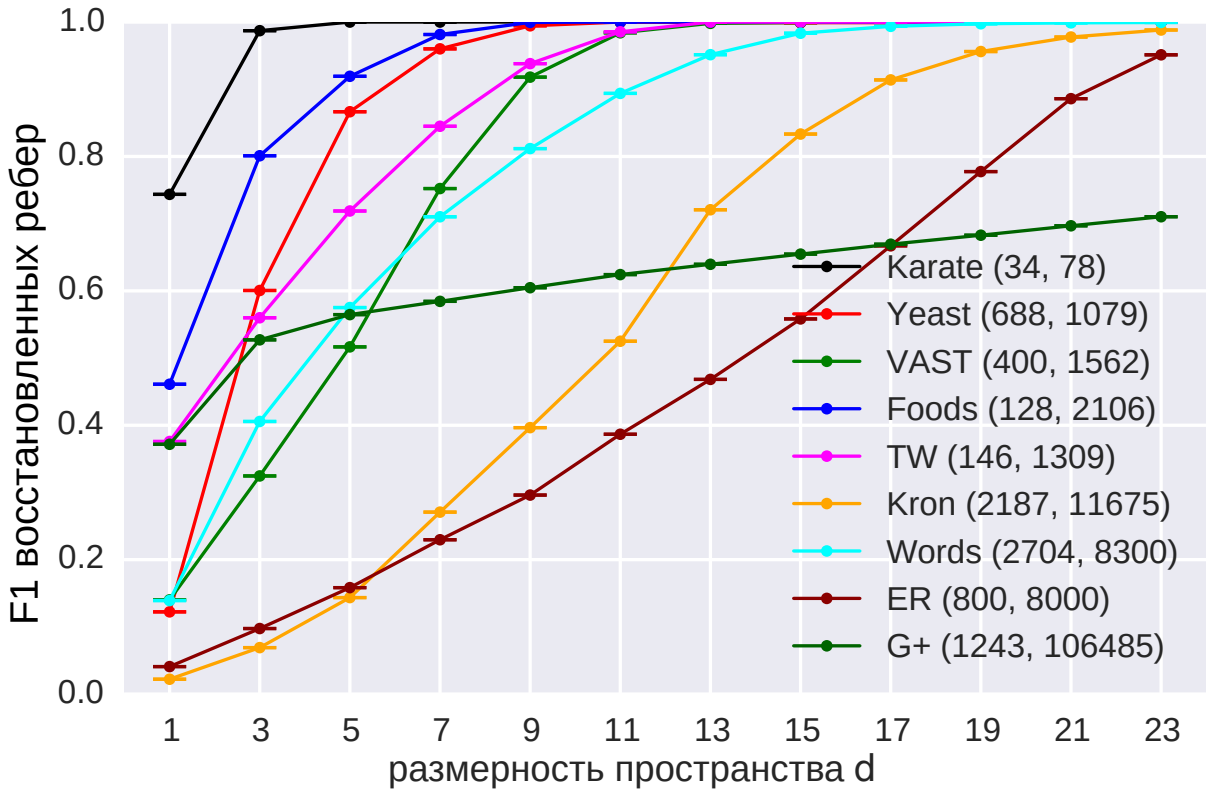
Компонент	**LINE**	**BLM**	**COMBO**
аппроксимация цели J	NEG	NCE	NEG
функция оценки s_{ij}	$(\vec{u}_i, \vec{v}_j)$	$(\vec{u}_i, \vec{v}_j) - Z_i$	$(\vec{u}_i, \vec{v}_j) - Z_i$
инициализация $\lvert\vec{u}_i\rvert$; $\lvert\vec{v}_i\rvert$	$\mathcal{N}(0, diag(\frac{1}{d}, ..., \frac{1}{d}))$; 0	$\mathcal{U}[-\frac{1}{2d}, \frac{1}{2d}]$; $\mathcal{U}[-\frac{1}{2d}, \frac{1}{2d}]$	$\mathcal{U}[-\frac{1}{2\sqrt{d}}, \frac{1}{2\sqrt{d}}]$; $\mathcal{U}[-\frac{1}{2\sqrt{d}}, \frac{1}{2\sqrt{d}}]$
шумовые распределения $p_n(i)$; $p_n(j)$	i с вер-тью 1; $p_n(j) \propto d_j^{3/4}$	i с вер-тью 1; $p_n(j) \propto d_j$	i с вер-тью 1; $p_n(j) \propto d_j^{3/4}$
нормировочный параметр Z_i (для восстановления ребер)	прямой подсчет	из модели / прямой подсчет	из модели / прямой подсчет
фильтр шумовых ребер	НЕТ	НЕТ	ДА
регуляризация	снижение скорости обучения до 0	L^2	НЕТ

1. Вложение + восстановление



Качество вложения при постоянной размерности d=30

1. Вложение + восстановление



Качество вложения при увеличении размерности d

Результат: Для всех графов из коллекции F_1 достигала 0.99 при некоторой размерности d, характеризующей их “сложность”.

2. Аппроксимация распределения + сэмплирование

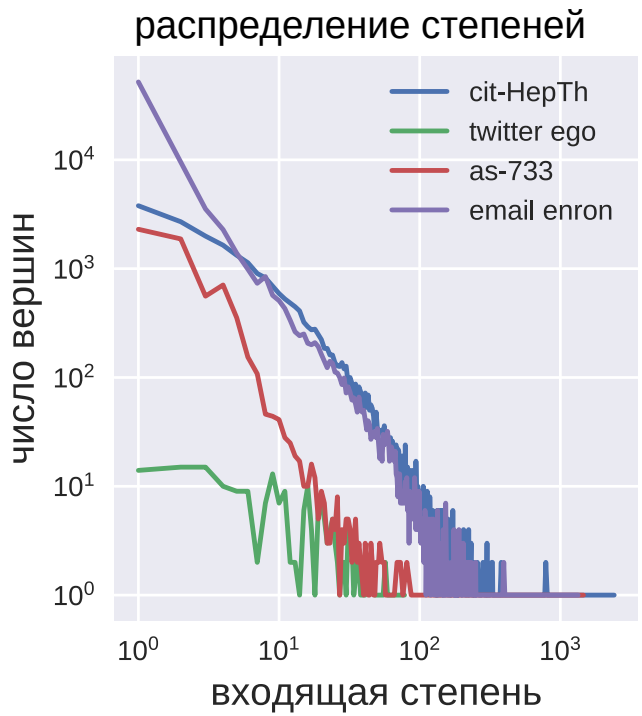
Цель: имея вложение с $F_1 \geqslant 0.99$, аппроксимировать распределение векторов-вершин $\vec{r_i} \sim \mathcal{R}$, так чтобы построенные на основе сэмплирования из него графы были похожи на данный.

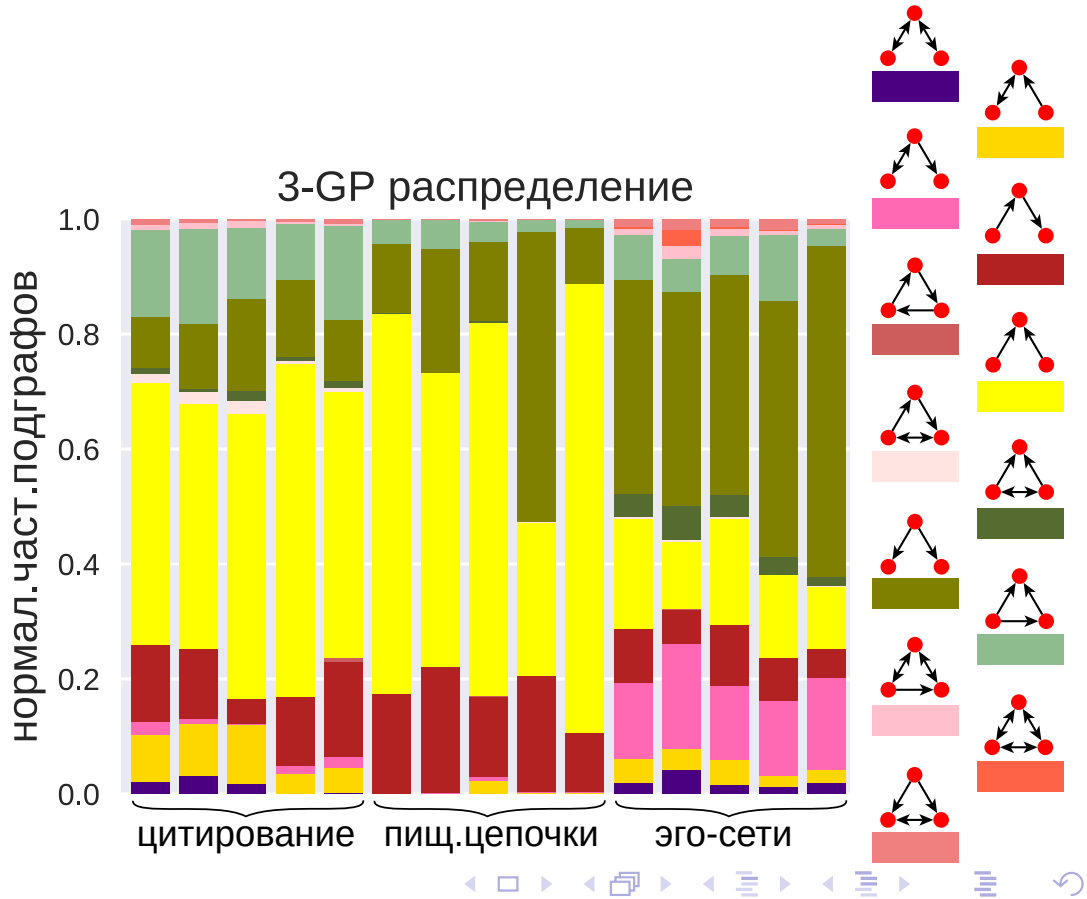
Мера сходства: косинусная близость распределения подграфов размера 3 сгенерированных графов к исходному, отклонение n и m графа от ожидаемых. Схожесть форм распределения степеней в лог-шкале (визуально).

2. Аппроксимация распределения + сэмплирование

Графовые характеристики для установления похожести

- *форма распределения степеней* вершин — определяет малый диаметр, коэффициент кластеризации как функцию степени узла;
- *распределение подграфов размера 3* (3-GP) — определяет свойства кластеризации, позволяет классифицировать графы по их доменам





2. Аппроксимация распределения + сэмплирование

Решение:

Добавление гауссовского шума (GN)

Сэмплирование равномерно из той же выборки векторов с добавлением гауссовского шума малой амплитуды ϵ

Альтернативы:

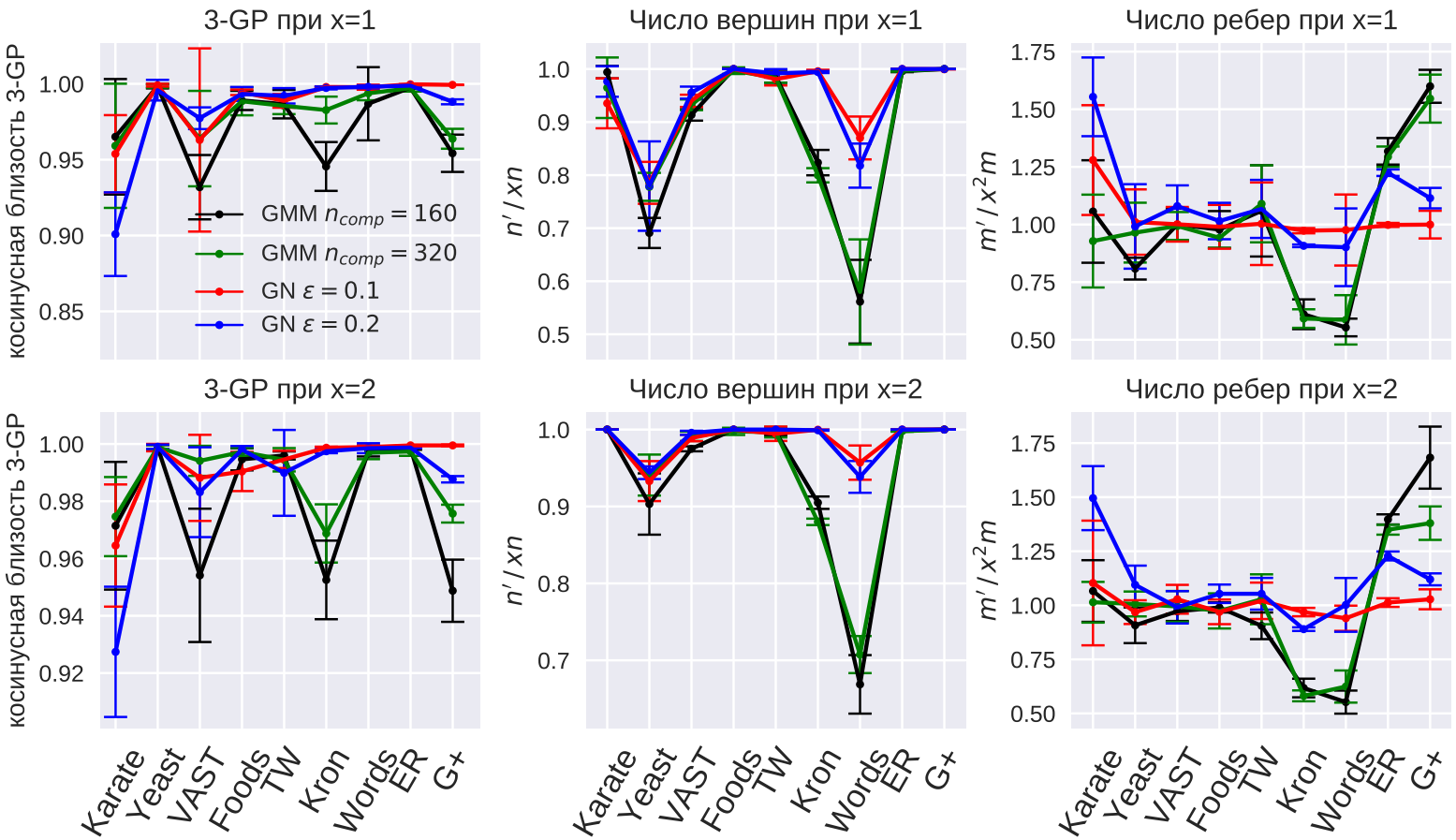
Модель гауссовых смесей (GMM)

Обучение Gaussian Mixture Model с различным кол-вом компонент n_{comp}

Нейросетевые подходы

Генеративно-состязательная сеть, Вариационный автоэнкодер
(Не удалось добиться приемлемой близости генерируемых графов к исходному)

2. Аппроксимация распределения + сэмплирование



Результат: GN с $\epsilon \in [0.1; 0.2]$ показывает лучший результат в смысле близости 3-GP, а также близости n' и m' к ожидаемым.

2. Аппроксимация распределения + сэмплирование

Функция оценки для пары вершин $s_{ij} = s(\vec{r}_i, \vec{r}_j)$
Размерность вложения $d \ll n$

Теорема (о масштабировании графа)
Пусть $\mathcal{R}$ вероятностное распределение в пространстве $\mathbb{R}^d$ и задана функция $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Если из распределения $\mathcal{R}$ получен набор случайных векторов $\{\vec{r}_i\}_{i=1}^n$, соответствующих n вершинам графа, и наличие ребра (i, j) в графе G задается условием $s(\vec{r}_i, \vec{r}_j) > t_G$, то число ребер m графа будет зависеть от числа вершин как $m \propto n^2$.

Теорема задает ограничение метода сэмплирования, однако на практике это несущественно при небольших коэффициентах масштабирования.

3. Атрибуты: метки сообществ и веса ребер

Цель: в сгенерированном графе корректно задать веса ребер и структуру сообществ — согласованно со структурой графа и друг с другом

Мера качества: модулярность структуры сообществ для направленного взвешенного графа[5] должна быть такой же высокой как в исходном графе

Решение: при сэмплинге новых векторов-вершин с помощью GN метки вершин (сообщества) и ребер (веса) наследуются. Если в исходном графе нет соответствующего ребра, назначается вес по умолчанию $w_0 = \min_{(i,j)\in E} w_{ij}$.

[5] M. Drobyshevskiy, A. Korshunov, and D. Turdakov. Parallel modularity computation for directed weighted graphs with overlapping communities. In Proceedings of the Institute for System Programming, volume 28(6), pages 153–170, 2016.

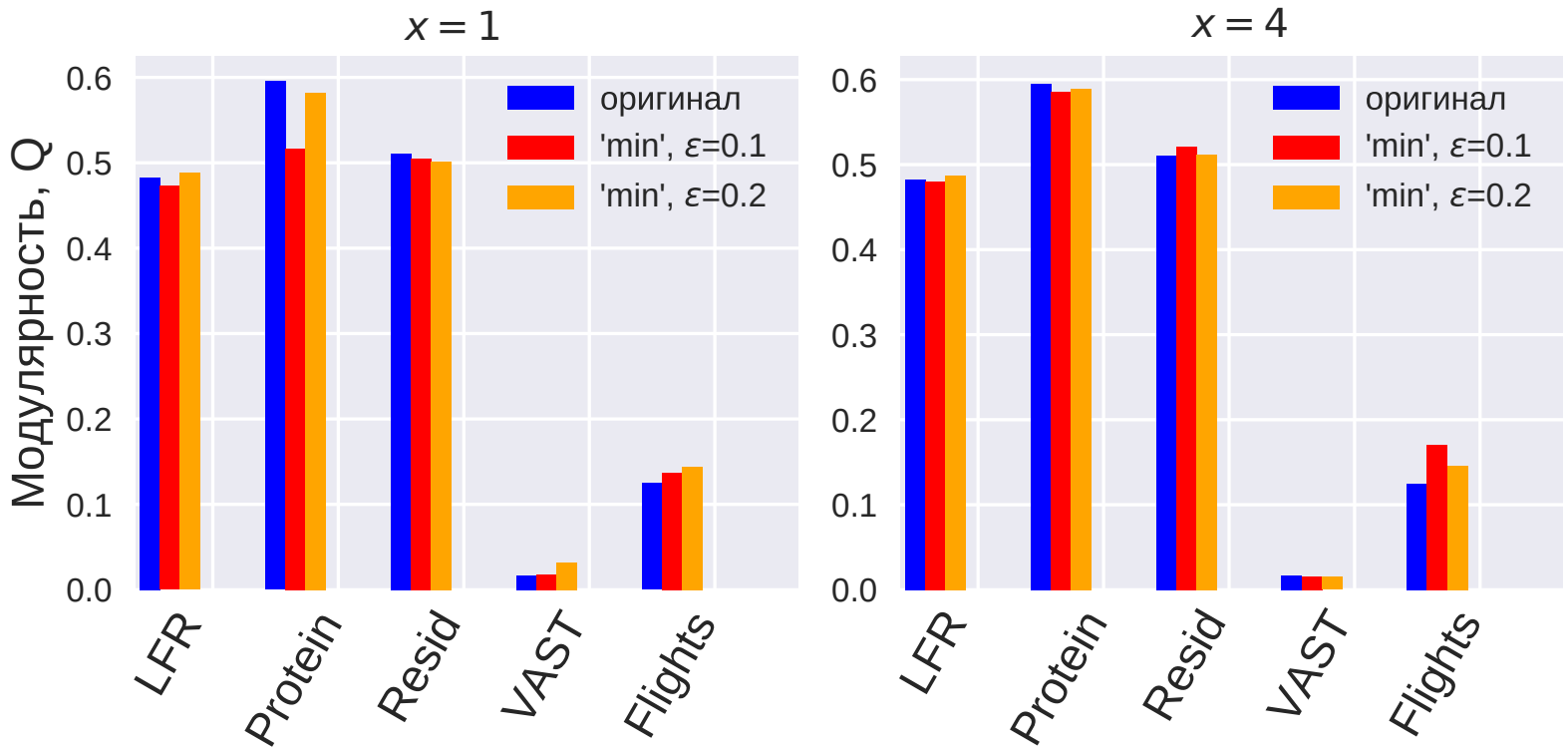
3. Атрибуты: метки сообществ и веса ребер

Тренировочная коллекция направленных взвешенных графов с сообществами, найденными алгоритмом OSLOM

Описание графа	**имя**	**n**	**m**	**Q**	**d**
Структурная смежность иммуноглобулина	Protein	95	213	0.6630	7
Отношения дружбы в группе общежития. Вес — уровень дружбы	Resid	217	2672	0.5106	14
Мобильные звонки; вес ребра — количество звонков	VAST	400	1562	0.5743	12
Перелеты межды аэропортами США	Flights	1574	28236	0.1247	31
Синтетический граф LFR	LFR	1000	14396	0.7209	26

- n — число вершин
- m — число ребер
- d — оптимальная размерность вложения
- Q — модулярность найденных сообществ

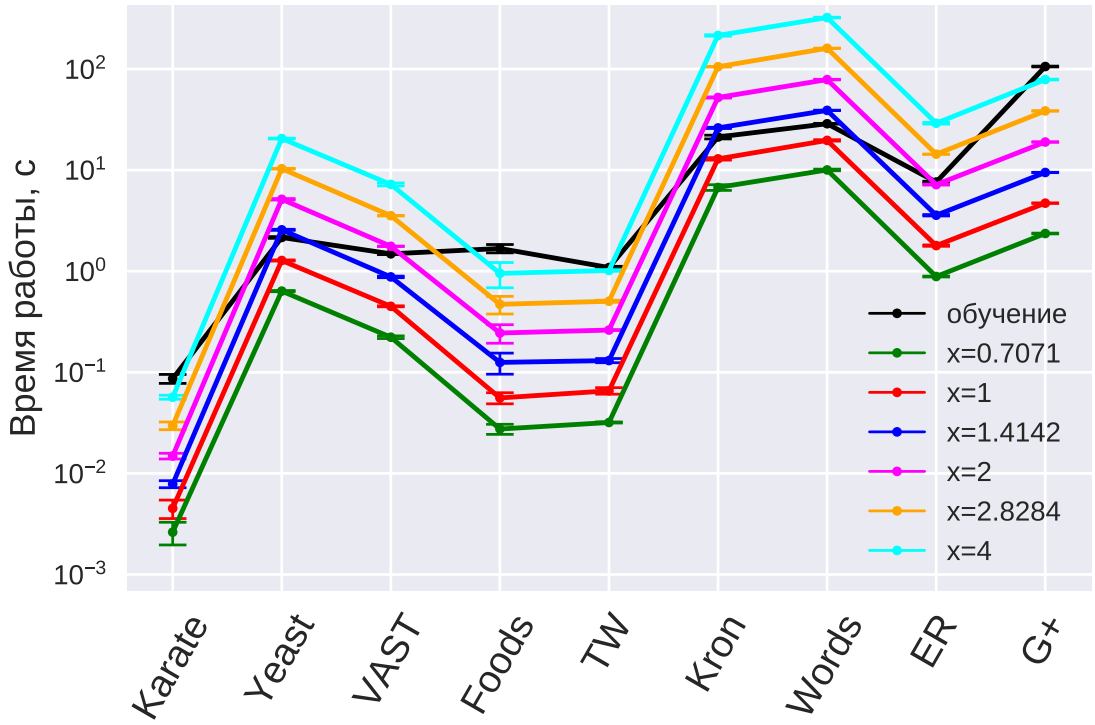
3. Атрибуты: метки сообществ и веса ребер



- при масштабировании модулярность сообществ сохраняет такое же (высокое) значение как в исходном графе
- при $\epsilon \leqslant 0.2$ модулярность остается высокой

Производительность

- n — число вершин
- m — число ребер
- d — размерность пространства вложения
- x — коэффициент масштабирования



Лемма 1. Вычислительная сложность обучения $O((m\frac{m}{n} + n^2)d)$.

Лемма 2. Вычислительная сложность генерации графа $O(x^2n^2d)$.

Теорема 2. Общая вычислительная сложность $O((\frac{m^2}{n} + x^2n^2)d)$.

Программная реализация

Программная система состоит из модулей:

1. Менеджер графовых данных
2. Обучение представления графа
3. Работа с представлением графа
4. Генерация графа
5. Подсчет характеристик
6. Фреймворк для тестирования

Основная часть написана на `python`, обучение представления и генерация графа реализована на `cython` с использованием параллельных вычислений.

Объем кода около 35 000 строк

Программа зарегистрирована в реестре программ для ЭМВ

Веб-демонстрация ERGG-dwc на `http://ergg.at.ispras.ru`

Экспериментальное сравнение с другими подходами

Цель: проанализировать способность моделей имитировать графы из различных графовых доменов.

Требования к генерируемым графам:

1) *похожесть* на исходный: отклонения по каждой характеристике не выше соответствующих отклонений у других современных методов
2) *вариабельность*: разброс значений числовых характеристик близок к соответствующему разбросу у реальных графов из одного домена

Методология

Методы: ERGG-dwc, SKG, Gscaler

Характеристики: число ребер, средняя степень вершин, ассортативность степеней, взаимность ребер, коэф.Джини распределения степеней, средний коэф.кластеризации, косинусная близость 3-GP, эффективный диаметр гигантской компоненты, спектральный радиус.

Коллекция: графы разного размера из разных доменов

Тестовая коллекция направленных (взвешенных) графов

Описание графа	домен	имя	n	m
Белок-белковые взаимодействия	биол.	PPI	2 239	6 452
Сеть доверия Epinion	социал.	Epinion	49 288	487 183
E-мейлы сотрудников Enron	социал.	Enron	87 273	321 918
Мобильные звонки	социал.	WU	72 146	100 974
Цитирования статей по физике высоких энергий	информ.	CitHepTh	27 770	352 807
Смежность слов в англоязычных текстах	информ.	Words	7 381	46 281
Перелеты между аэропортами США	технол.	Flights	1 574	28 236
Зависимости между классами софта JDK 1.6.0.7	технол.	JDK	6 434	53 892

- n — число вершин
- m — число ребер

Числовые характеристики. Отклонение значения (среднее по 5 запускам) характеристики в сгенерированном графе от оригинального. В каждой ячейке символ генератора ('E' – ERGG-dwc, 'G' – Gscaler, 'S' – SKG), если отклонение меньше 10%, или прочерк. **Результат:** ERGG-dwc: 37, Gscaler: 49, SKG: 8.

характеристика	PPI	Epinions	CitHepTh	Words	WU	Flights	JDK	Enron
число ребер	-GS	EG-	-G-	EGS	EG-	EG-	EG-	-GS
средняя степень	EG-	EG-	-GS	-GS	-G-	EG-	EG-	-GS
ассорт. степени	EG-	—	-G-	EG-	-G-	E–	EG-	-G-
взаимность ребер	—	—	—	-G-	—	—	—	—
коэф. Джини распр. степ.	EG-	EG-	EG-	EG-	EG-	EG-	EG-	EG-
ср. коэф. кластеризации	—	E–	—	-G-	—	—	—	—
косинусная близость 3-GP	EG-	EG-	EG-	EGS	—	—	EG-	EG-
эфф. диаметр	EG-	-GS	E–	EG-	—	—	—	—
спектр. норма	-G-	EG-	EG-	EG-	-G-	EG-	EG-	EG-

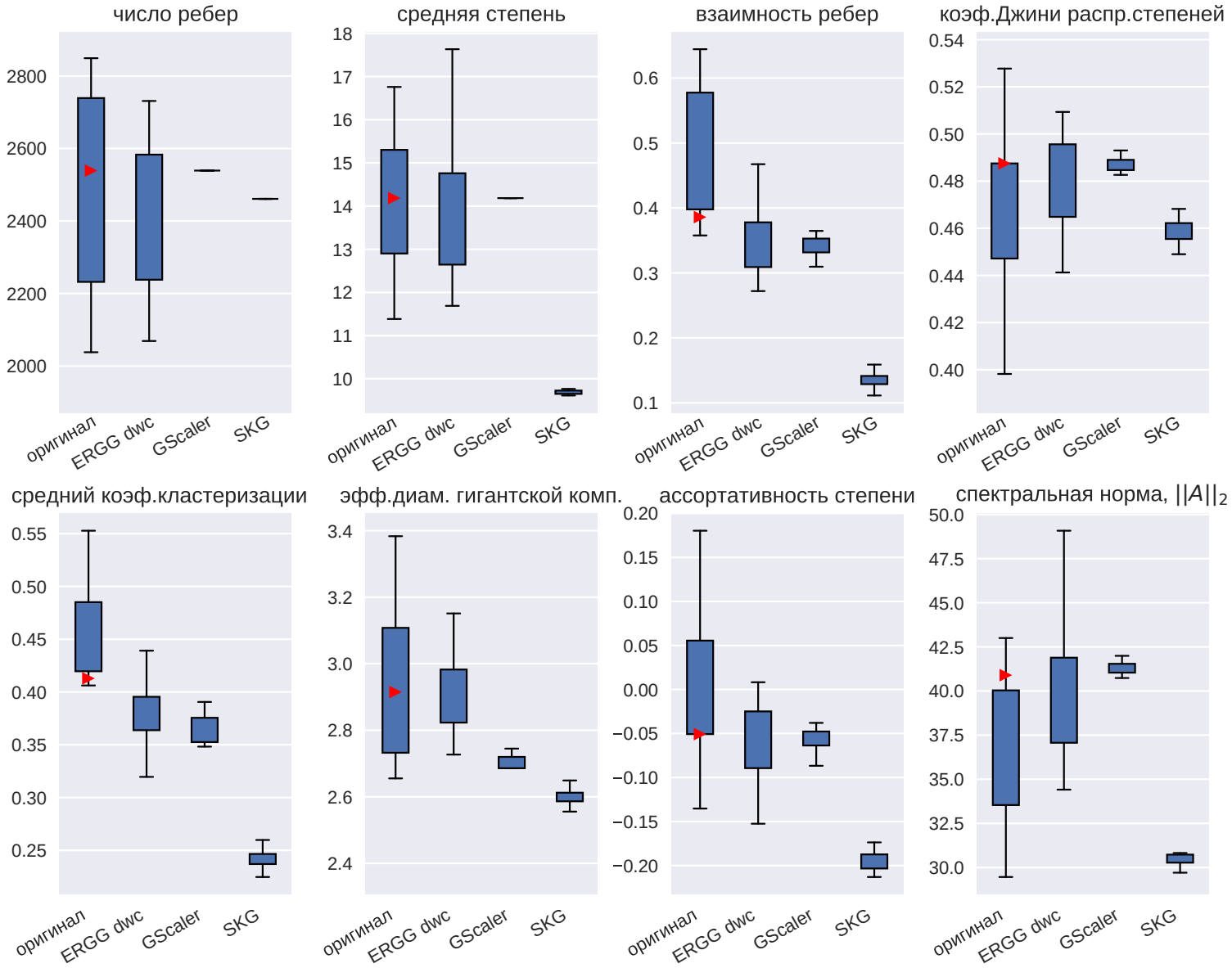
Измерение вариабельности

Мотивация:

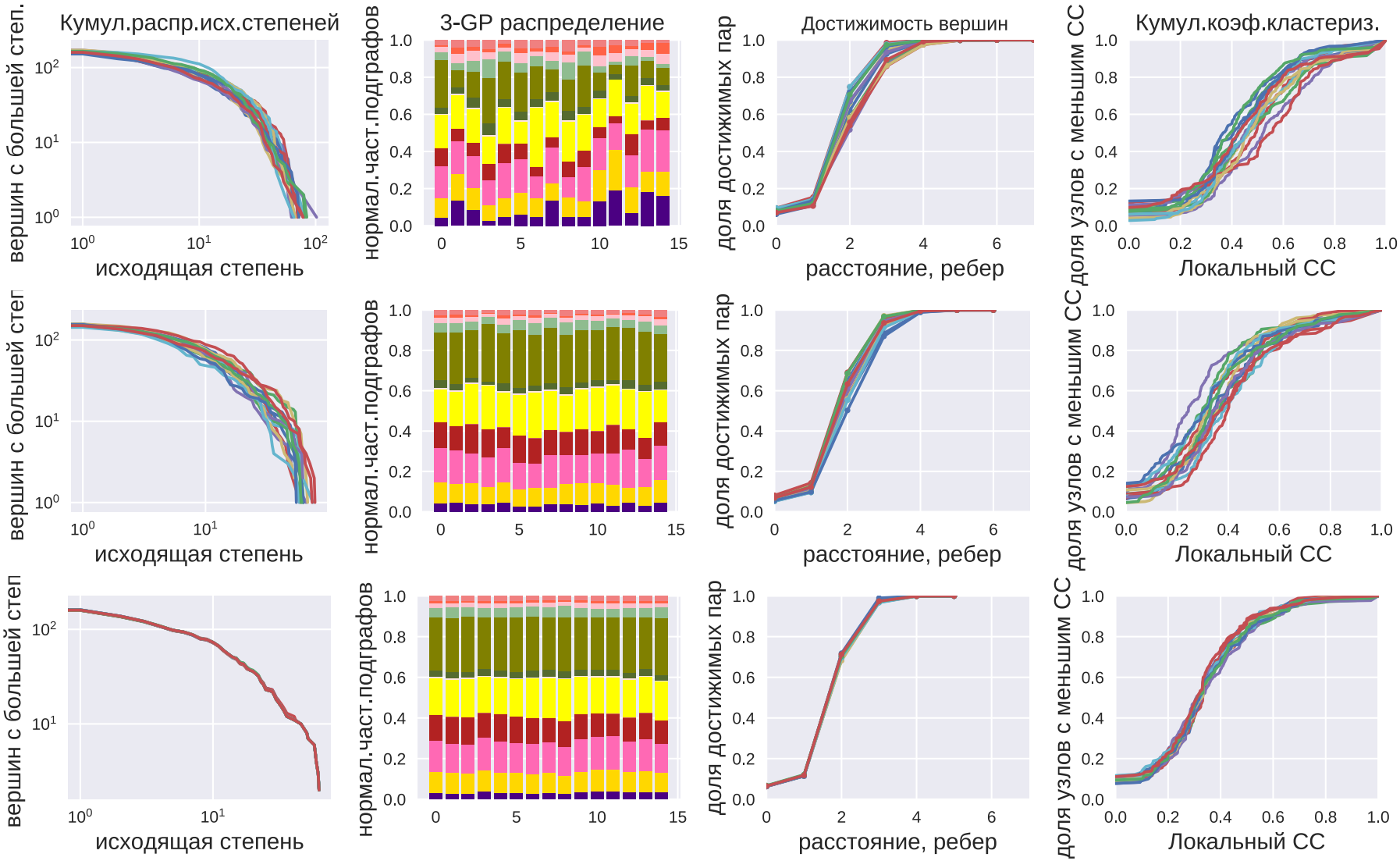
- Графы в одном домене похожи, но имеют разброс в признаках
- Для тестирования статистической значимости алгоритмов нужна представительная выборка графов

Эксперимент:
1. Реальный домен — 15 эго-сетей twitter близких по числу узлов ($n \in [170; 180]$) и числу ребер ($m \in [2000; 3000]$)
2. Обучение модели на одном из графов
3. Искусственный домен — 15 результатов запуска генерации

Оценка:
Сравнение разброса числовых признаков в реальном домене и в имитации. Метод должен обеспечивать вариабельность сравнимую с реальной.



Вариабельность числовых признаков. Красной стрелкой отмечено значение характеристики для графа, на котором происходило обучение.



Вариабельность признаков-распределений. Сверху вниз: оригинал; ERGG-dwc: разброс достаточно широкий, кроме 3-GP; GScaler: разброс заметно меньше.

Выводы

Показана способность ERGG-dwc имитировать графы, похожие на данный, с двумя условиями:

1. похожесть на исходный граф по ряду известных характеристик;
2. вариабельность по ряду характеристик, отражающая вариабельность внутри одного домена.

Учитывая эти две особенности, ERGG-dwc может применяться на практике для создания искусственных наборов данных для надежной проверки качества работы алгоритмов майнинга графов. Возможность генерирования графов контролируемого размера с сохранением исходных признаков позволяет тестировать масштабируемость алгоритмов.

Основные результаты

1. Предложен новый подход ERGG к генерации случайных направленных графов, похожих на данный, основанный на вложении графа в пространство размерности, много меньшей числа его вершин.
2. В рамках подхода ERGG предложен метод ERGG-dwc, решающий задачу генерации графов, похожих на данный и удовлетворяющих требованиям: автоматическое обучение на заданном графе, контролируемый размер генерируемых графов, поддержка направленных ребер, взвешенных ребер и структуры сообществ. Соответствие метода ERGG-dwc требованиям похожести и вариабельности генерируемых графов подтверждено экспериментально, также показана корректность назначения структуры сообществ и весов ребер. Доказаны теоремы о вычислительной сложности и масштабировании. Налагаемые ими ограничения не являются существенными для применимости метода.
3. Создана программная система, в которой реализован прототип ERGG-dwc и проведено его экспериментальное сравнение с другими современными методами. Показано, что ERGG-dwc не уступает другим методам в похожести генерируемых графов, но превосходит их по вариабельности.

Публикации

1. *Drobyshevskiy Mikhail, Korshunov Anton, Turdakov Denis*. Parallel modularity computation for directed weighted graphs with overlapping communities // *Труды Института системного программирования РАН*. – 2016. – Vol. 28, no. 6. – Pp. 153–170.
2. *Drobyshevskiy Mikhail, Korshunov Anton, Turdakov Denis*. Learning and scaling directed networks via graph embedding // Joint European Conference on Machine Learning and Knowledge Discovery in Databases / Springer. – 2017. – Pp. 634–650.
   получена награда "best student paper award"
3. *Drobyshevskiy Mikhail, Turdakov Denis, Kuznetsov Sergey*. Reproducing Network Structure: A Comparative Study of Random Graph Generators // Ivannikov ISPRAS Open Conference (ISPRAS), 2017 / IEEE. – 2017. – Pp. 83–89.
4. *Filippov A., Drobyshevsky M., Korshunov A. et al*. Network analysis tool testing. — 23.08.2018. — Патент РФ 2018/151619. URL: `https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2018151619`.

Таксономия подходов моделирования случайных графов

Таксономия	Описание
Генеративные	
Классические	каждое ребро появляется независимо
Локальные правила	
PA принцип	новый узел присоединяется к узлу с большей степенью
копирование	граф растет засчет репликации уже имеющихся структур
другие	правила присоединения вовлекают соседей узла
Рекурсивные	рекурсивная процедура формирует структуру графа
Скрытые атрибуты	
геометрические	узлам ассоциированы вектора в "скрытом" пространстве
метки узлов	узлы соединяются на основе схожести их атрибутов
Топология из оптимизации	решение оптимизационной задачи дает топологию графа
Управляемые признаками	
Аналитические	желаемые признаки выразимы через параметры модели
Оптимизация функции	
оценка параметров	статистический метод оценки параметров модели
экспоненциальные	сэмплинг графов из заданного вероятностного пространства
Редактирование графа	
переключ. ребер	рандомизация ребер графа при заданных ограничениях
другие	модификация графа путем рандомизации представления
Предметно-специфичные	
Со структурой сообществ	есть группы узлов, более связные внутри, чем между собой
С весами на ребрах	сила связи выражена весом ребра

- Многие модели комбинируют несколько подходов

Стохастические Кронекеровские графы: **SKG**[6]

- Рекурсивная генеративная модель: вероятностная матрица смежности $A_k = A_1^{\otimes k}$
- Процедура *автоматического* подгона параметров Kronfit

	u_1	u_2
u_1	a	b
u_2	c	d

	v_1	v_2	v_3	v_4
v_1	a·a	a·b	b·a	b·b
v_2	a·c	a·d	b·c	b·d
v_3	c·a	c·b	d·a	d·b
v_4	c·c	c·d	d·c	d·d

Свойства: мультиномиальные распр.степеней, собственных значений и собственных векторов, степенной закон уплотнения $m \propto n^{\alpha}$, малый диаметр

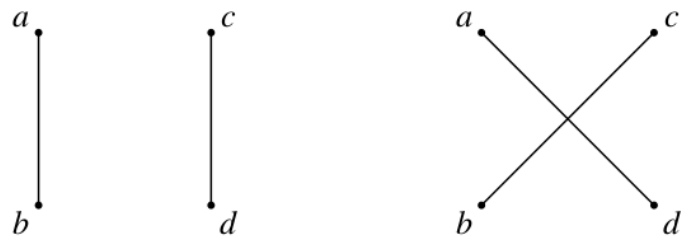
Недостатки:

- Размер графа только $n = 2^k$ вершин
- Осцилляции в распр.степеней, низкий коэф.кластеризации
- Не поддерживаются ни структура сообществ, ни веса ребер

[6] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. Journal of Machine Learning Research, 11(Feb):985–1042, 2010

Репликатор сложных сетей: **ReCoN**[7]

- реплицируется входной граф с сообществами
- переключение ребер внутри и между сообществами с сохранением степеней



Свойства: воспроизводимость признаков: средняя и максимальная степень, коэф.Джини распределения степеней вершин, средний коэф.кластеризации, диаметр, кол-во компонент связности и кол-во сообществ.

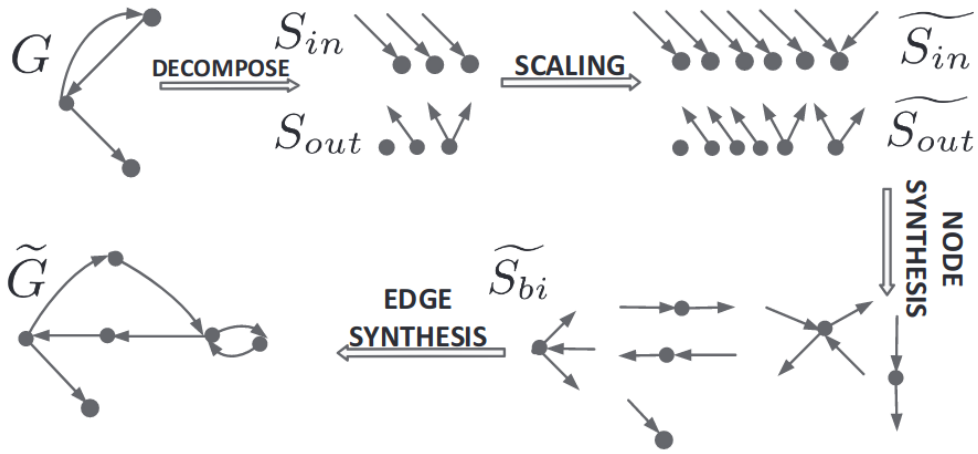
Недостатки:

- Размер графа только $n = k \cdot n_0$ вершин, $k \in \mathbb{N}$
- Переключение ребер нарушает все остальные признаки

[7] C. L. Staudt, M. Hamann, I. Safro, A. Gutfraind, and H. Meyerhenke. Generating scaled replicas of real-world complex networks. arXiv preprint arXiv:1609.02121, 2016.

Gscaler[8]

- разбиение и мультиплицирование на уровне отдельных узлов
- при соединении сохраняется распределение бистепеней узлов $f_{bi}(d_{in}, d_{out})$, и корреляция ребер $f_{corr}(\alpha_1, \alpha_2)$, α_i — бистепень узла (d_i^{in}, d_i^{out})



Свойства: воспроизводит распределения входящих/исходящих степеней и бистепеней, а также близость генерирумых графов по эффективному диаметру, коэффициенту кластеризации, отношению размеров наибольших сильно связных компонент и среднему минимальному пути.

Недостатки: Не поддерживаются ни структура сообществ, ни веса ребер

[8] Zhang JW, Tay YC. GSCALER: Synthetically Scaling A Given Graph. // EDBT. — 2016. — Pp. 53–64.

Распределение степеней вершин: Gscaler повторяет распределение степеней почти идеально (в силу алгоритма), ERGG-dwc воспроизводит форму распределения не точно, но намного ближе к оригиналу, чем SKG

Коэффициент Джини

граф	оригинал	ERGG	GScaler	SKG
PPI	0.6717	0.6266	**0.6719**	0.5984
Epinions	0.7993	0.7567	**0.7994**	0.5695
CitHepTh	0.5702	0.5591	**0.5704**	0.5155
Words	0.7466	0.7110	**0.7466**	0.6295
WU	0.5138	0.2896	**0.5139**	0.4535
Flights	0.7534	0.6952	**0.7534**	0.5831
JDK	0.6381	0.6865	**0.6384**	0.5565
Enron	0.8037	0.7700	**0.8037**	0.5981

Ассортативность:
хорошо улавливается Gscaler, хуже ERGG-dwc, результаты SKG еще дальше от оригинальных.

Ассортативность входящих степеней

граф	оригинал	ERGG	GScaler	SKG
PPI	0.0125	-0.0099	**0.0117**	-0.0790
Epinions	0.0525	0.0344	**0.0506**	-0.0405
CitHepTh	0.0378	0.0514	**0.0364**	0.1327
Words	-0.2394	-0.1765	**-0.2392**	-0.0634
WU	-0.0382	0.0072	**-0.0384**	0.0052
Flights	-0.1186	-0.1235	**-0.1186**	-0.1442
JDK	-0.0132	-0.0293	**-0.0181**	0.0082
Enron	-0.0180	-0.0363	**-0.0218**	-0.0612

То же можно сказать для ассортативности входящих и исходящих степеней

Средний коэффициент кластеризации:
SKG не в состоянии смоделировать высокий коэффициент кластеризации; результаты ERGG ближе к оригинальным чем Gscaler в 6 из 8 доменах

Средний коэффициент кластеризации

граф	оригинал	ERGG	GScaler	SKG
PPI	0.0400	**0.0293**	0.0144	0.0289
Epinions	0.1808	**0.1584**	0.0253	0.0019
CitHepTh	0.3120	**0.2646**	0.0181	0.0030
Words	0.4083	0.2191	**0.3659**	0.0276
WU	0.0535	**0.0011**	0.0004	0.0001
Flights	0.5042	**0.4370**	0.3052	0.1236
JDK	0.6707	0.3647	**0.5203**	0.0423
Enron	0.1193	**0.2209**	0.0159	0.0010

Распределение подграфов размера 3:
воспроизводится ERGG и Gscaler практически для всех графов, кроме Flights (есть неточности) и WU (не воспроизводится — возможно из-за его низкой плотности (средняя степень 1.4)).
SKG не улавливает распределение подграфов.

Косинусная близость 3-GP к оригинальному

граф	ERGG	GScaler	SKG
PPI	0.999953	**0.999994**	0.127098
Epinions	**0.982114**	0.973139	0.646164
CitHepTh	**0.999368**	0.998012	0.575088
Words	0.991189	**0.999839**	0.937990
WU	0.189969	**0.743930**	0.145881
Flights	**0.849542**	0.848725	0.194089
JDK	0.999967	**0.999998**	0.083487
Enron	0.995244	**0.996218**	0.424371

Достижимость вершин: Точность воспроизведения сильно варьируется для разных доменов.

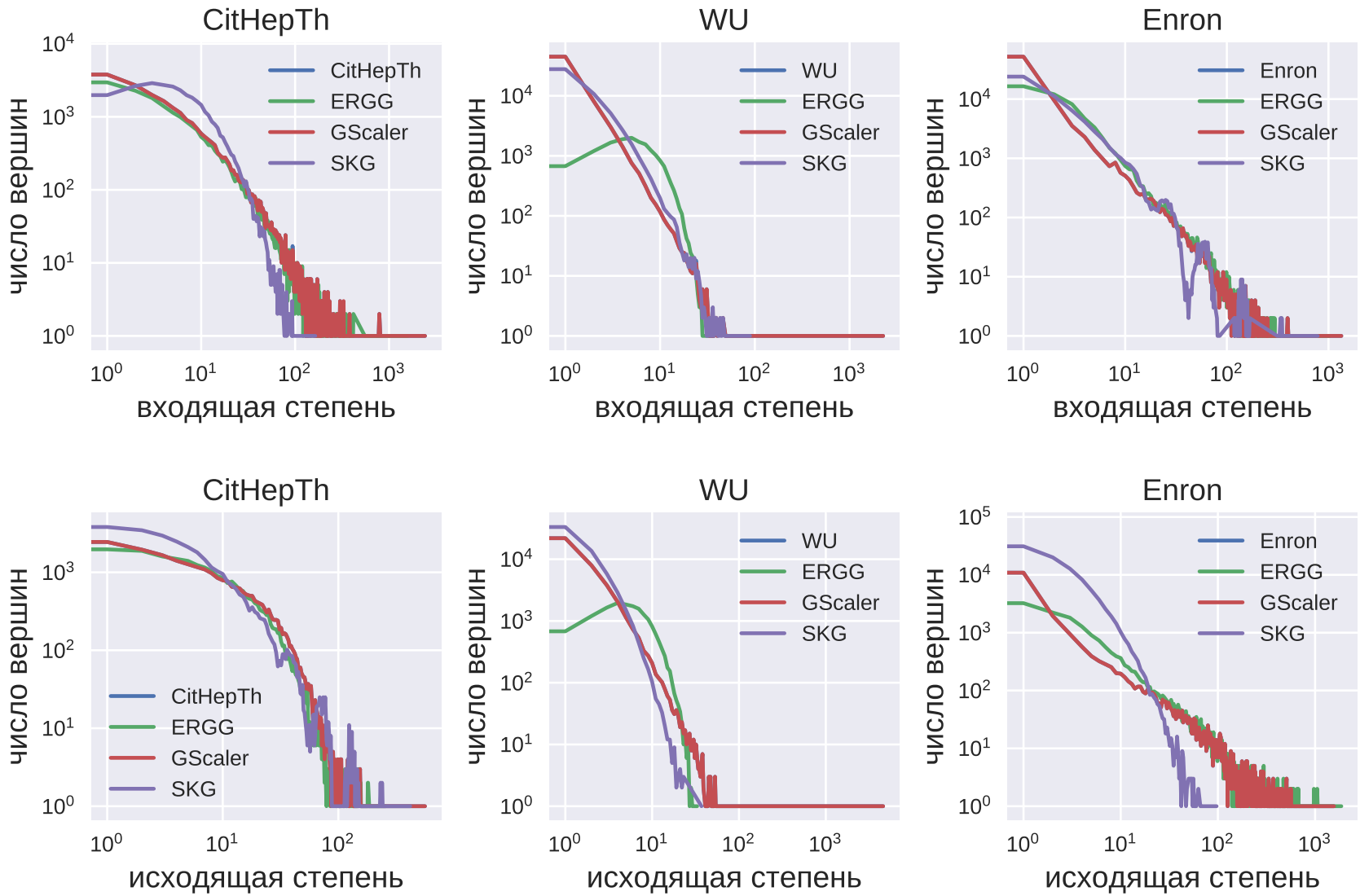
Максимальное расстояние между распределениями

граф	ERGG	GScaler	SKG
PPI	**0.027**	0.054	0.053
Epinions	0.106	0.094	**0.093**
CitHepTh	**0.104**	0.251	0.320
Words	0.085	**0.003**	0.330
WU	0.778	**0.156**	0.170
Flights	**0.194**	0.231	0.270
JDK	0.194	**0.006**	0.750
Enron	0.349	0.176	**0.119**

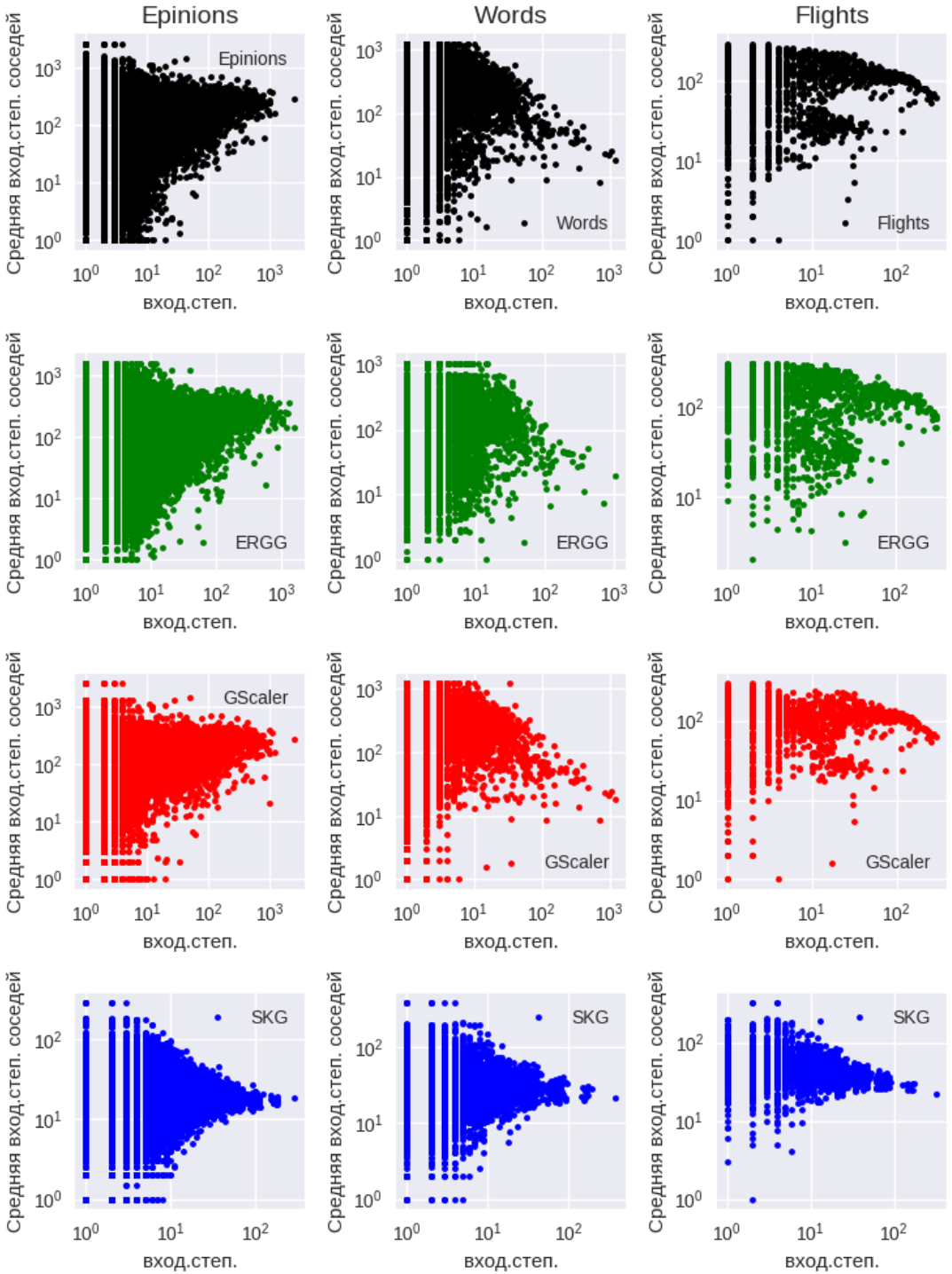
90%-эффективный диаметр гигантской компоненты:
результаты малоинформативны. Все алгоритмы плохо отработали на WU (слишком малая компонента) и Flights (неточно).

90%-эффективный диаметр

граф	оригинал	ERGG	GScaler	SKG
PPI	4.63	**4.69**	4.36	4.48
Epinions	4.70	4.36	**4.41**	4.55
CitHepTh	5.38	**5.70**	4.35	3.96
Words	2.97	3.42	**2.95**	3.84
WU	12.17	4.37	8.45	**8.99**
Flights	3.82	**3.06**	2.92	2.87
JDK	2.13	2.80	**2.04**	3.45
Enron	5.80	4.83	4.99	**5.45**

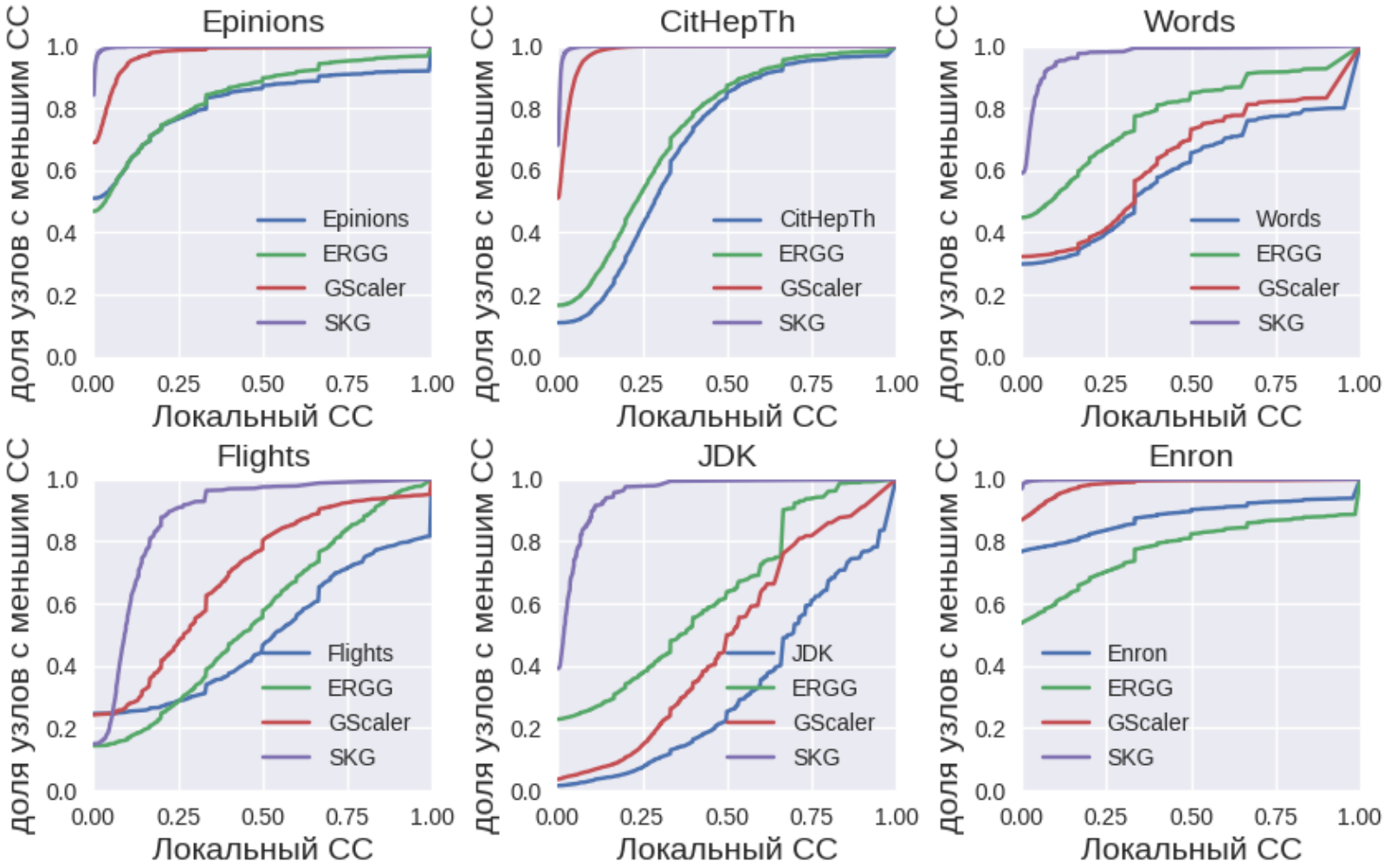


Распределение степеней: Gscaler повторяет почти идеально, ERGG-dwc воспроизводит форму не точно, но намного ближе к оригиналу, чем SKG

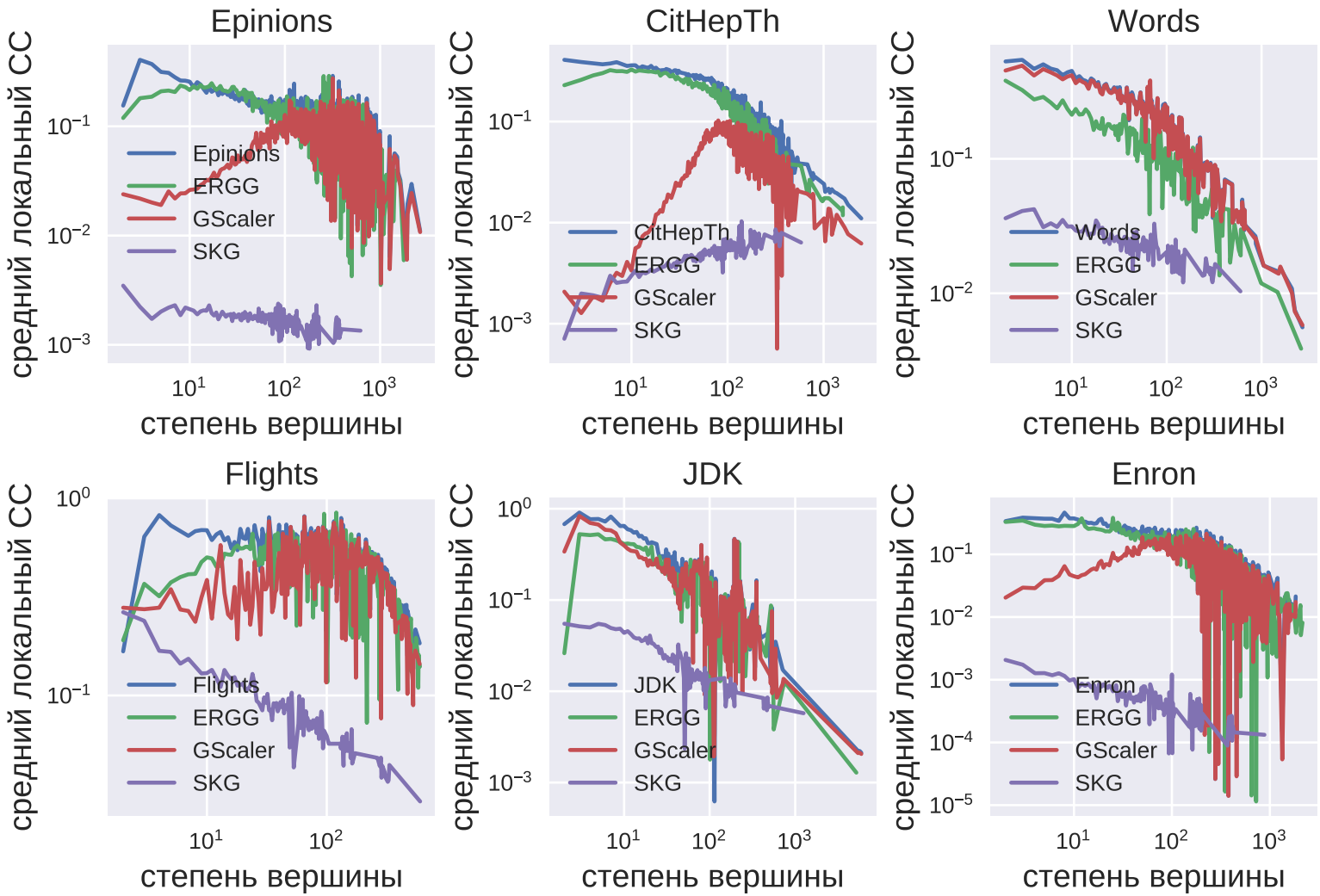


Ассортативность: хорошо улавливается как ERGG-dwc, так и Gscaler, тогда как для SKG графики ассортативности больше похожи друг на друга, чем на оригинальные.

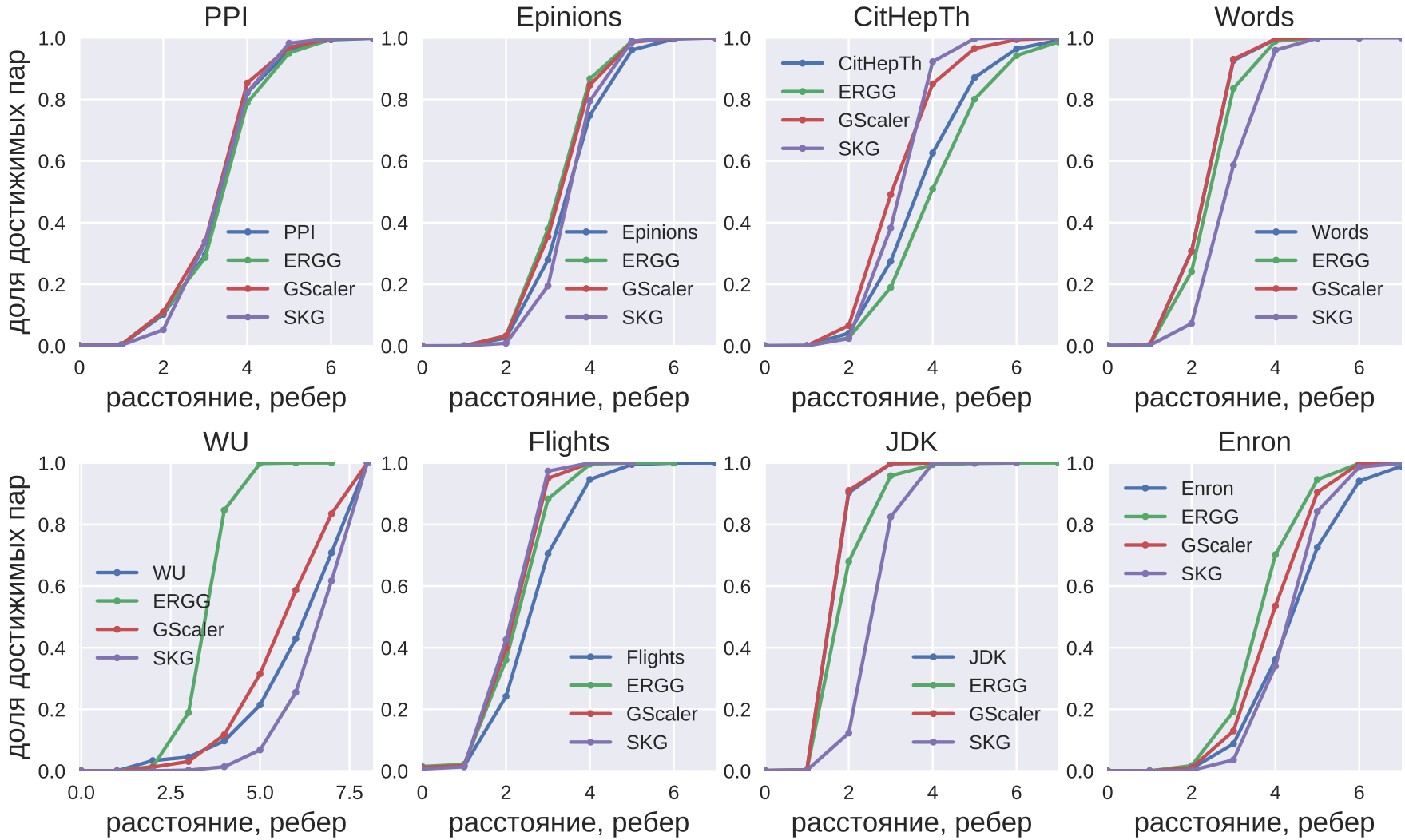
То же можно сказать для ассортативности входящих и исходящих степеней



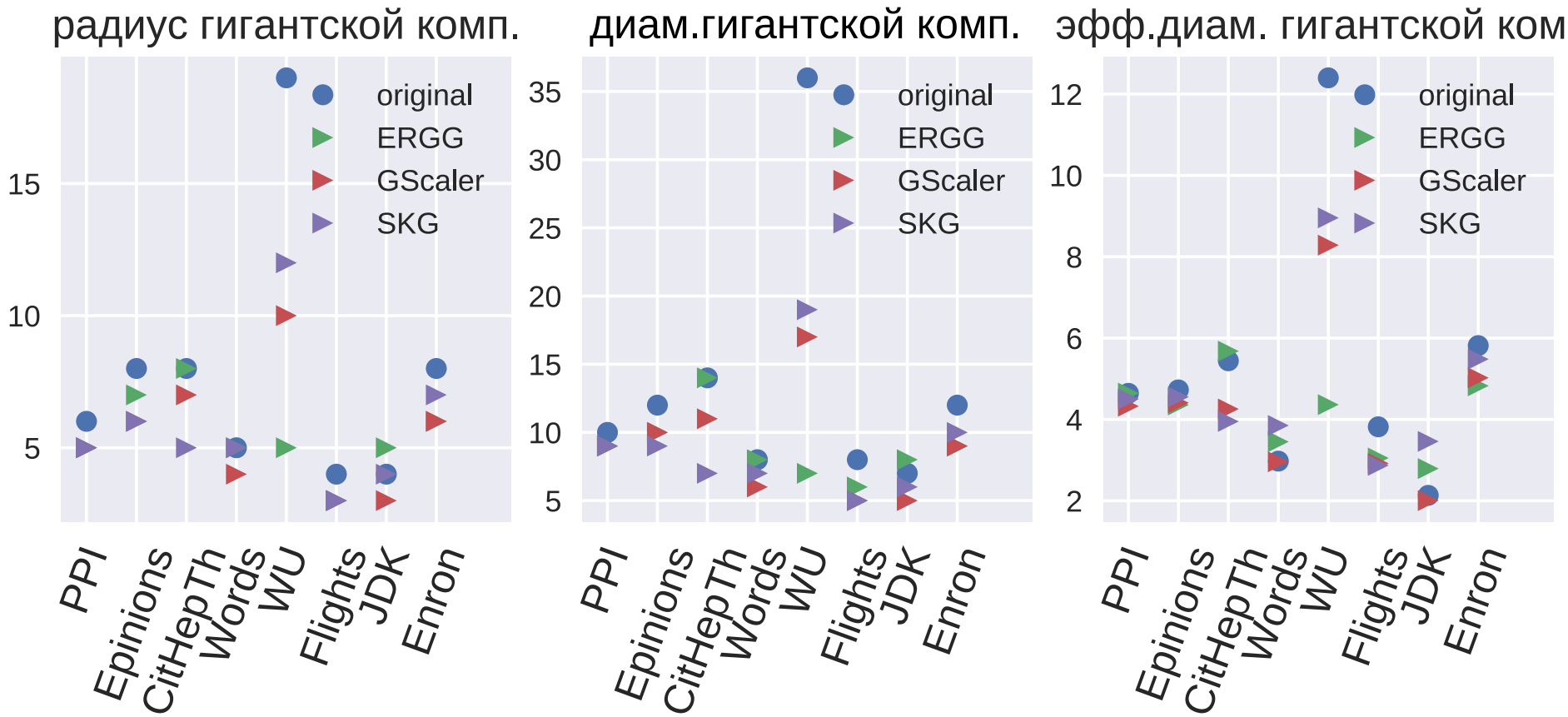
Кумул. коэфф. кластеризации: SKG не в состоянии смоделировать высокий коэфф.кластер.; результаты Gscaler и ERGG сильно различаются по доменам. WU, Flights, JDK, Enron — плохо воспроизведены



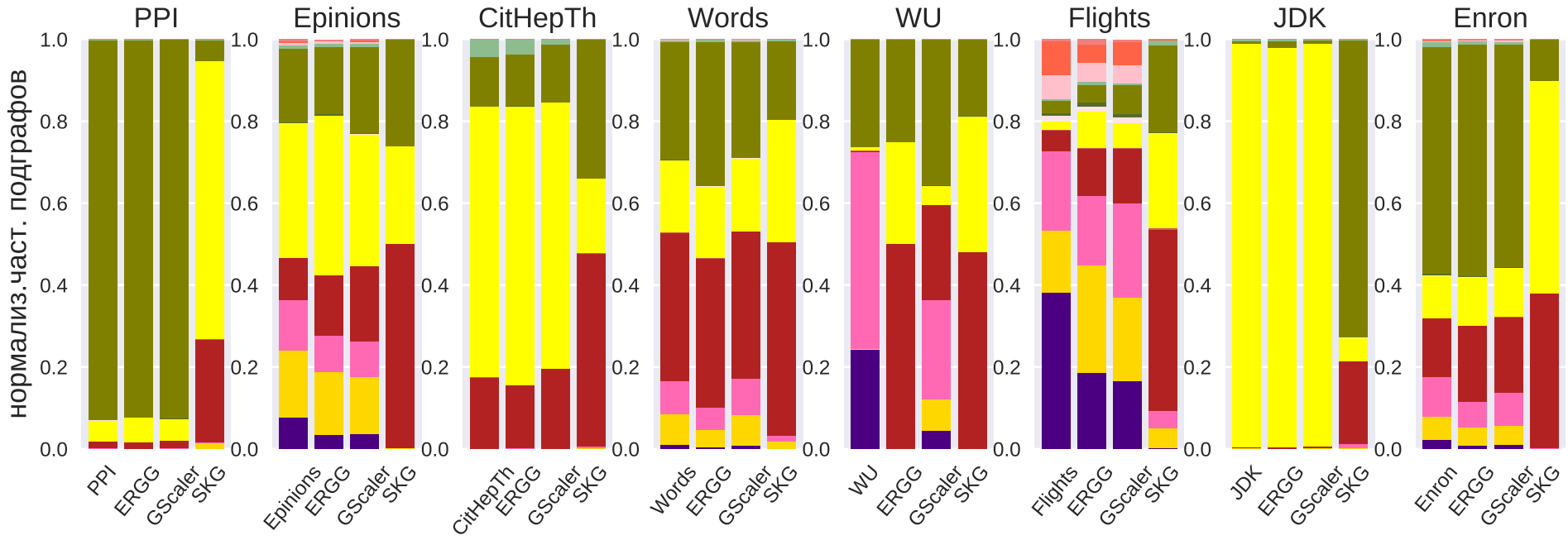
Коэфф. кластеризации от степени: аналогично.
Gscaler часто не улавливает монотонную зависимость



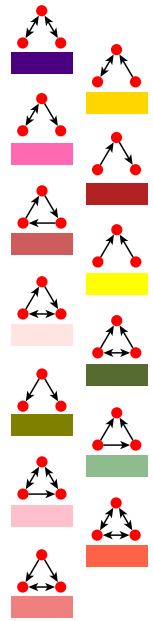
Достижимость вершин: Точность воспроизведения варьируется для разных доменов. Gscaler чаще ближе к оригиналу чем ERGG-dwc (в 5 из 8 случаев)



Радиус, диаметр и 90%-эффективный диаметр гигантской компоненты: результаты не очень информативны. Все алгоритмы плохо отработали на WU (слишком малая компонента) и Flights (неточно).

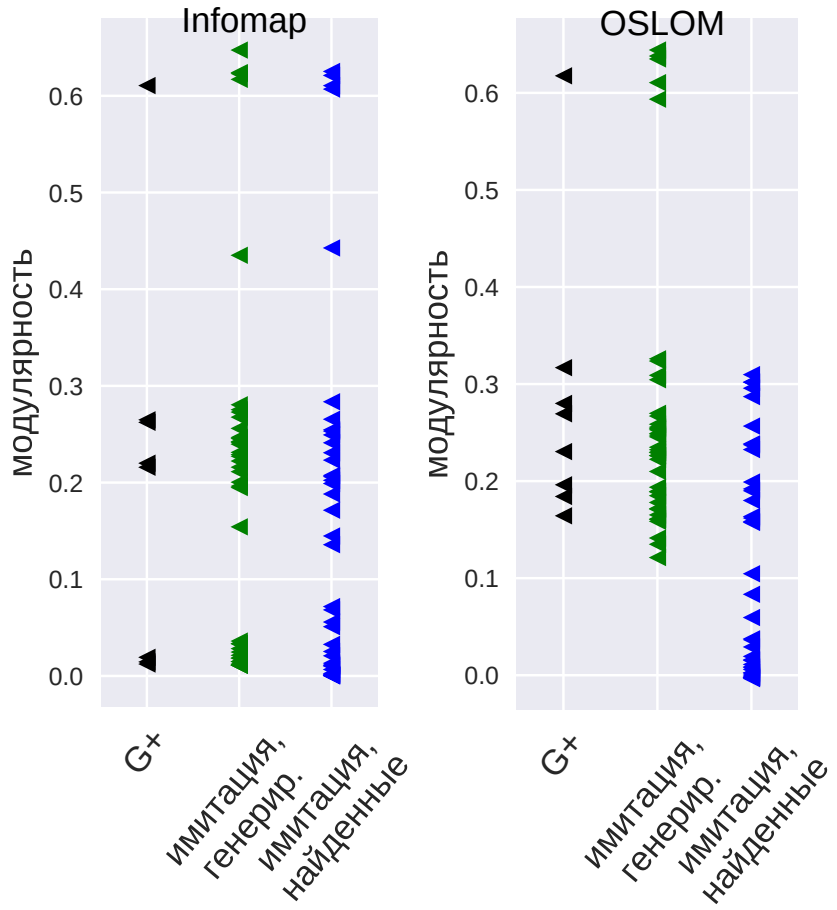


Распределение подграфов размера 3: воспроизводится ERGG и Gscaler практически для всех графов, кроме Flights (есть неточности) и WU (не воспроизводится — возможно из-за его низкой плотности (средняя степень 1.4)).
SKG не улавливает распределение подграфов.



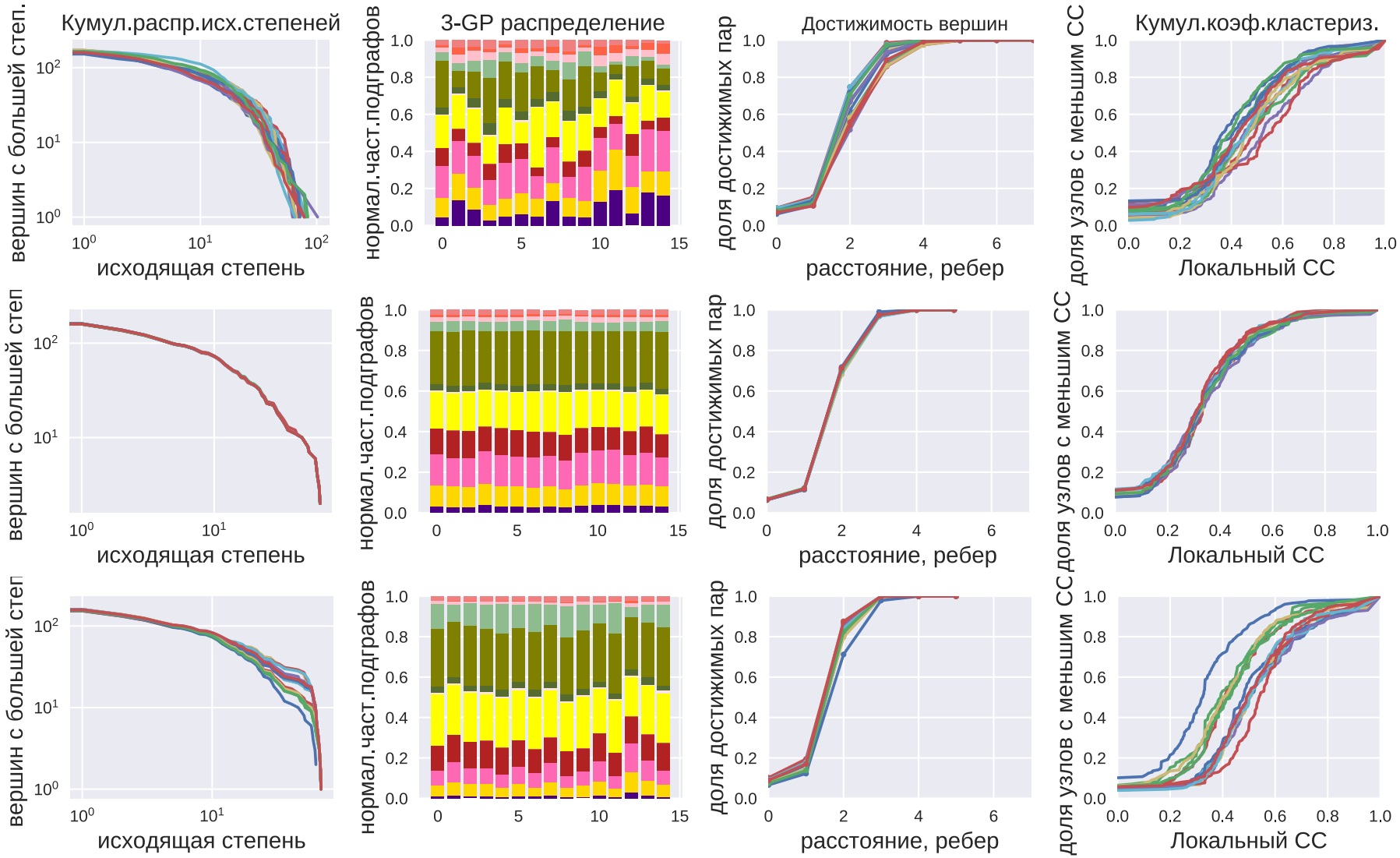
Приложение: алгоритм поиска сообществ

Эксперимент:
1. Домен — 8 эго-сетей Google-plus близких по числу узлов ($n \in [450; 500]$)
Алгоритмы — два метода поиска сообществ: OSLOM и Infomap
2. Запускается алгоритм на исходных графах (**черный**)
3. Для каждого графа генерируется 5 похожих (с сообществами) с помощью ERGG-dwc (**зеленый**)
4. Запускается алгоритм на искусственных графах (**синий**)



Оценка:
Модулярность — качество покрытия сообществами.
Результат:
Infomap дает стабильный результат, OSLOM — нет

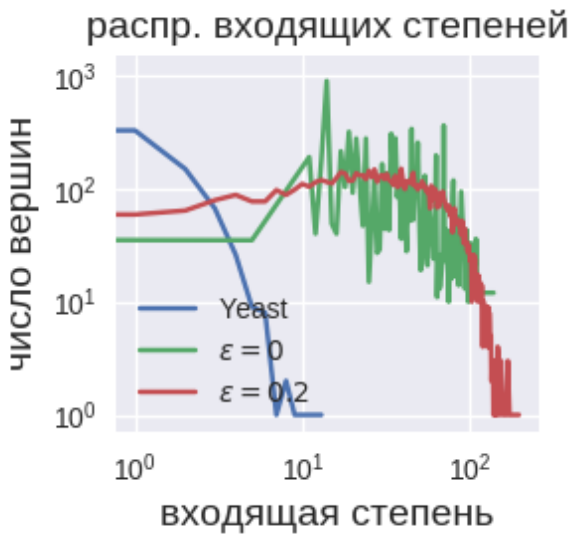
lfnf

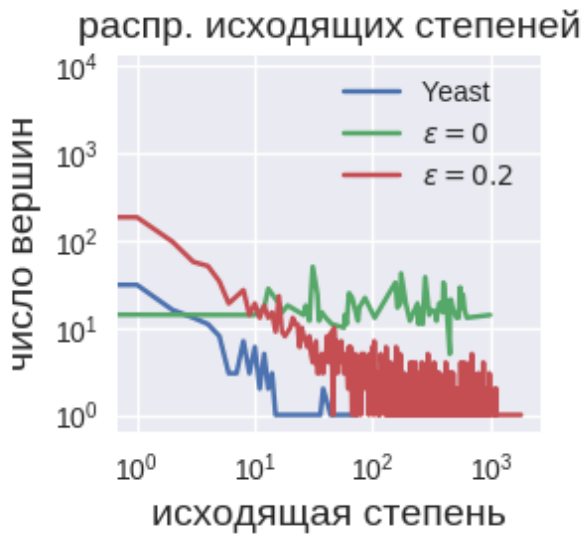

Кумул.распр.исх.степеней
3-GP распределение
Достижимость вершин
Кумул.коэф.кластериз.
вершин с большей степ.
нормал.част.подграфов
доля достижимых пар
доля узлов с меньшим CC
исходящая степень
расстояние, ребер
Локальный CC

Случай отсутствия шума

Пусть $\epsilon = 0$

1. Тогда новые вектора-вершины просто выбираются из множества существующих;
2. при больших x вершины образуют группы, либо связанные между собой всеми возможными ребрами, либо несвязанные вовсе;
3. в результате — перекос в распределении степеней

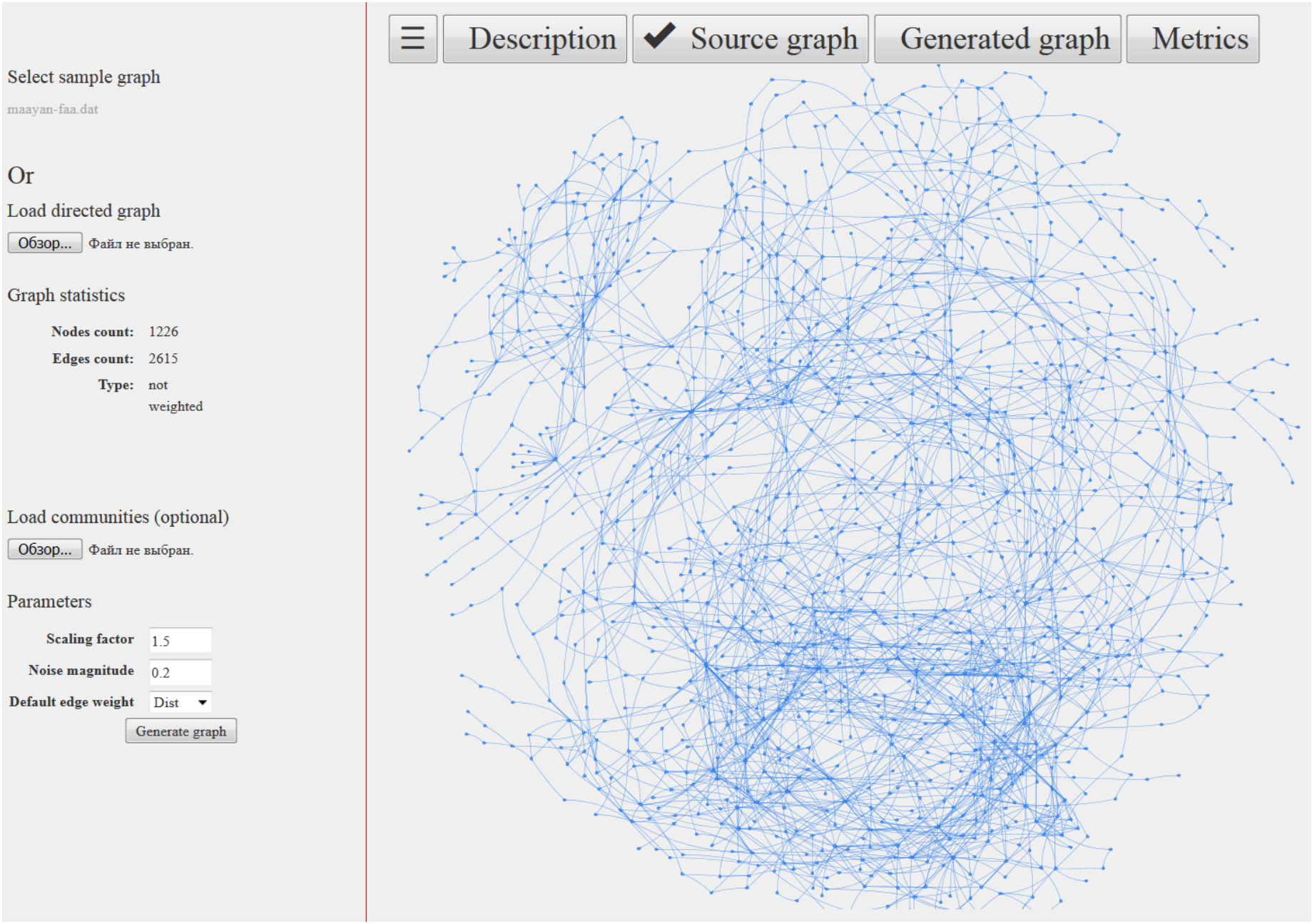




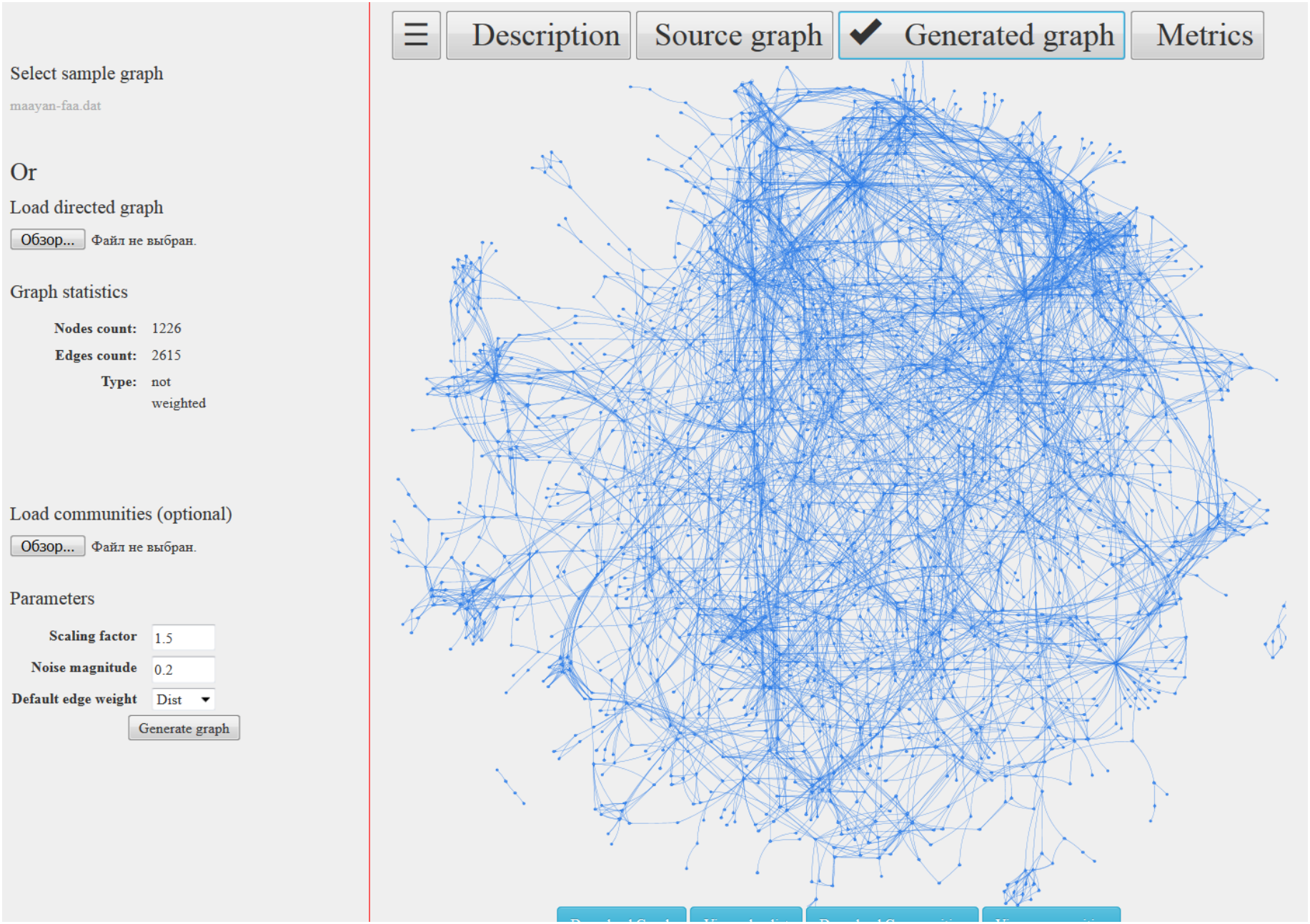
Устранение недостатков метода ERGG-dwc

1. Преодоление квадратичной от числа вершин сложности генерации ребер в графе.
2. Достижение более гибкого закона масштабирования генерируемых графов: в текущей схеме число ребер растет пропорционально квадрату от числа вершин, в то время как на практике наблюдается степенная зависимость с различными показателями степени.

Веб-демонстрация ERGG-dwc — http://ergg.at.ispras.ru



Веб-демонстрация ERGG-dwc — http://ergg.at.ispras.ru



Веб-демонстрация ERGG-dwc — http://ergg.at.ispras.ru