# МОДУЛЬ АНАЛИЗА ЧАСТОТ ВСТРЕЧАЕМОСТИ ТИПОВЫХ ПОДГРАФОВ В СИСТЕМЕ АГЕНТНОГО МОДЕЛИРОВАНИЯ SIMBIGRAPH[1]

**Е.Б. Юдин, М.Н. Юдина (Омск)**

## Введение

Моделируемые системы в агентном подходе часто представляются графовыми моделями, это могут быть случайные графы, либо «статичные» графы, моделирующие мгновенные снимки реальных сетей. Вершинами при этом моделируются социальные субъекты, экономические организации и т. д., а ребрами – связи (экономические, социальные и т.д.) между элементами исследуемой системы.

В связи с важностью разработки структур взаимодействия для нужд практики и теории имитационного моделирования разрабатывается система агентного моделирования Simbigraph, в которой реализован ряд оригинальных графовых моделей растущих сетей [1], а также средства анализа графовых моделей с использованием параллельных вычислений. Система Simbigraph была представлена на конференциях «Имитационное моделирование. Теория и практика» в работах [2, 3]. В данной работе представлен модуль для расчета частот встречаемости типовых подграфов на трех и четырех вершинах, реализованный в Simbigraph. Предлагаемый модуль ускоряет расчет, во-первых, за счет использования параллельных вычислений JAVA 8, во-вторых, за счет корректной реализации метода Монте-Карло.

Сама задача расчета частот встречаемости типовых подграфов является актуальной задачей современной теории сетей (Network Science). Частота встречаемости подграфов, как важная структурная характеристика сетей, получила широкое распространение, начиная с ее применения в работе [4], в которой при исследовании генной регуляторной сети бактерии кишечной палочки выделялись такие подграфы, которые встречались чаще в реальной сети, чем в ее модели сети на основе случайных графов. После выявления таких подграфов задача исследователей-прикладников – понять, какую функцию выполняют найденные подграфы, какую роль играют, почему часто встречаются. Само знание частот встречаемости типовых подграфов также дает возможность оценить важные вероятностные характеристики сетей. Так, отношение утроенного числа циклов длины три к числу путей длины два для социальной сети задает вероятность того, что «друзья» случайно выбранного индивида являются «друзьями» между собой.

Задача обнаружения и подсчета подграфов представляет собой задачу высокой вычислительной стоимости. Разработан ряд алгоритмов и программ, решающих эту задачу. Среди разработанных программ, позволяющих рассчитывать частоты встречаемости подграфов, следует выделить:

– программу MFINDER (реализующую точный расчет и «алгоритм случайного выбора ребра» [5]);

– программу FANMOD и библиотеку igraph, разработанную для среды R (обе программы реализуют алгоритмы ESU и RAND-ESU [6]);

– программу AccMotif (использующую комбинаторные вычисления для подсчета подграфов различных видов [7]).

В таблице 1 представлен анализ времени расчета числа подграфов выбранными программами. Расчет выполнялся для таких сетей как сеть автономных систем Internet (22963 вершин, 48436 ребер), сеть пользователей электронной почты Enron (36692 вершин, 183831 ребер), сеть соавторства ученых в области информатики DBLP (317080 вершин, 1049866 ребер), сеть политических блогов Blogs (2224 вершин, 18956 дуг), сеть пользователей компонента

[1] Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 16-31-60023 мол_а_дк.

сети Twitter (81306 вершин, 1768149 дуг), сеть пользователей электронной почты EuAll (265214 вершин, 420045 дуг), сеть пользователей компонента социальной сети Google plus (107614 вершин, 13673453 связей). Данные о сетях предоставлены Юрием Лесковецем (https://snap.stanford.edu/data/) и Марком Ньюманом (http://www-personal.umich.edu/~mejn/netdata/). Программы тестировались на моноблоке HP Z1 с четырехядерным центральным процессором Intel Xeon E3-1245 с поддержкой технологии Hyper-Threading и 8 ГБ оперативной памяти DDR3-1333.

Таблица 1

Время расчета частот встречаемости типовых подграфов на трех и четырех узлах в программах MFinder, FanMod, igraph, AccMotif

| Сеть | MFinder | | AccMotif | | igraph | | FanMod | |
|---|---|---|---|---|---|---|---|---|
| | Число узлов в подграфе | | Число узлов в подграфе | | Число узлов в подграфе | | Число узлов в подграфе | |
| | 3 | 4 | 3 | 4 | 3 | 4 | 3 | 4 |
| Internet | 670 с | $>10^4$ с | 0,0084 с | 3,9 с | 6,8 с | 8134.8 с | 9,2 с | $>10^4$ с |
| Enron | 1015 с | $>10^4$ с | 0,31 с | 33 с | 17,9 с | 10097 с | 28,7 с | $>10^4$ с |
| DBLP | $>10^4$ с | $>10^4$ с | невозможно | | 3,51 с | 170,23 с | 10,72 с | 557,1 с |
| Blogs | 10 с | 5364 с | 0,025 с | 2,4 с | 0,9 с | 144 с | 1,1 с | 126,9 с |
| Twitter | $>10^4$ с | $>10^4$ с | невозможно | | 295 с | $>10^4$ с | 121,6 с | $>10^4$ с |
| EuAll | $>10^4$ с | $>10^4$ с | невозможно | | 102 с | $>10^4$ с | 119,9 с | $>10^4$ с |
| Google plus | $>10^4$ с | $>10^4$ с | невозможно | | $>10^4$ с | $>10^4$ с | $>10^4$ с | $>10^4$ с |

Как можно видеть из таблицы 1 программа AccMotif (разработана в 2012 г., использовалась версия 2017 года) работает быстрее MFinder (разработана 2002 году с исправлением расчета частот встречаемости от 2015 года), FANMOD (разработана в 2006 году) и igraph (разработана в 2006, использовалась версия 2017 года). Но ускорение расчета достигается за счет интенсивного использования оперативной памяти, а значит, для больших сетей эта программа непригодна. Поэтому для больших сетей, таких как сеть соавторства ученых в области информатики DBLP или сеть пользователей компонента социальной сети Google plus, использование этой программы невозможно.

Наиболее быстрая программа, которую можно использовать для расчета числа подграфов больших сетей, реализована в пакете igraph (используется алгоритм получивший название ESU [6]). Тем не менее, для больших сетей даже этот алгоритм оказывается неприемлемым. В данной работе предлагается модуль SimbigraphMotif (рис. 1), предназначенный для расчета частот встречаемости подграфов в системе Simbigraph. Использование модуля позволяет уменьшить время расчета по сравнению с известными программными пакетами за счет эффективного сокращения просматриваемой части графа, а также за счет средств распараллеливания на основе использования встроенной поддержки языка Java 8 – Fork-Join Framework и JSR 335: Streams.

**Параллельный алгоритм расчета частот встречаемости подграфов**

Модуль SimbigraphMotif позволяет рассчитывать точные значения частот встречаемости подграфов с использованием параллельных вычислений.

Для ускорения расчета частот встречаемости подграфов на трех вершинах с использованием точного расчета предлагаются следующие шаги:

– скопировать граф в виде списков смежных вершин на каждый вычислитель;

– разбить весь список вершин графа на непересекающиеся подмножества, количество которых равно числу вычислителей;

– разбить список ребер (дуг) графа на непересекающиеся подмножества, количество которых равно числу вычислителей;

– каждый вычислитель рассчитывает частоты подграфов для подграфов «своего» подмножества и передает рассчитанные значения главному вычислителю;

– главный вычислитель, получив данные от всех вычислителей, рассчитывает частоты встречаемости подграфов.

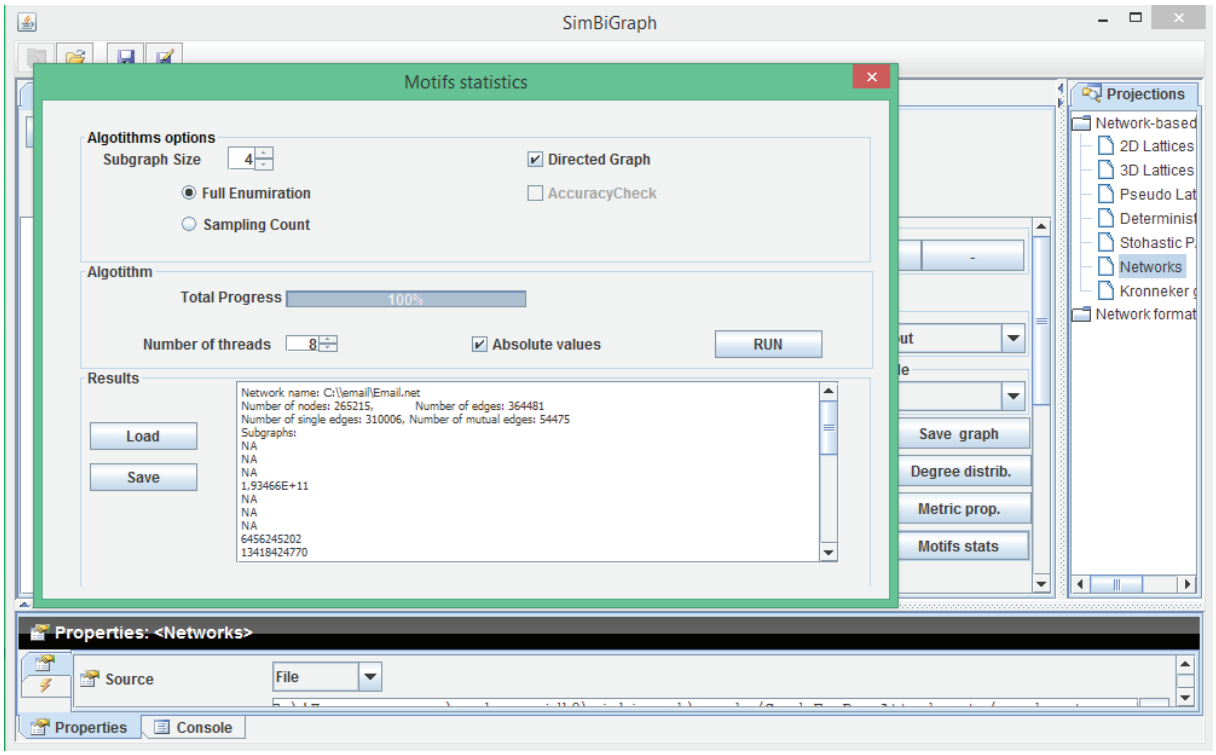


Рис. 1. Интерфейс модуля SimbigraphMotif для расчета частот встречаемости подграфов на трех и четырех вершинах

На рис. 2 приведены усредненные значения времени выполнения последовательного (1 поток) и параллельного (2, 4, 8 потоков) расчетов частот встречаемости типовых подграфов на трех и четырех узлах для сети соавторства ученых в области информатики DBLP. Расчеты выполнялись с использованием модуля SimbigraphMotif, библиотеки igraph и программы FANMOD.

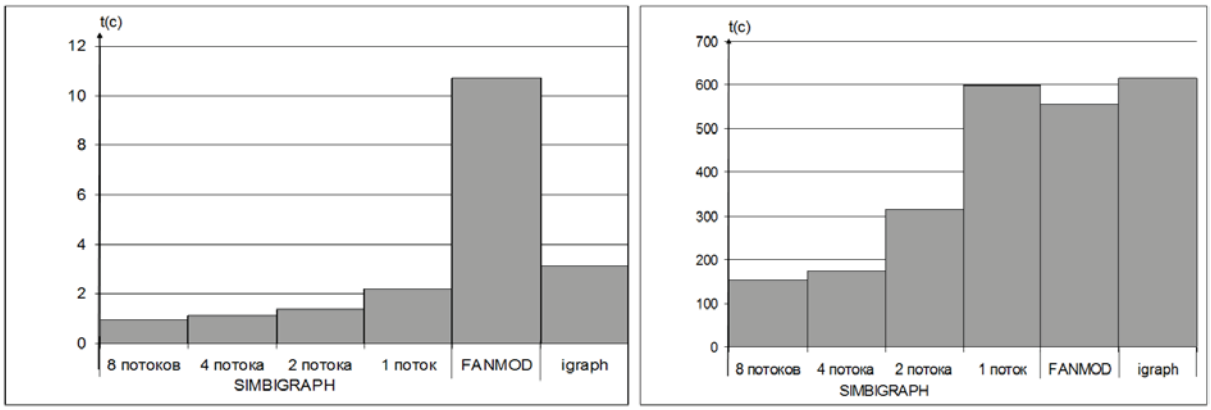


Рис. 2. Усредненное время расчета частот встречаемости подграфов на трех (слева) и четырех (справа) узлах в сети сотрудничества DBLP

Как можно видеть на рис. 2, в реализованном модуле удалось ускорить расчет не только по сравнению с программой FANMOD и с программой igraph.

**Расчет частот встречаемости методом случайной выборки каркасов**

В модуле SimbigraphMotif для расчета частот встречаемости подграфов реализована также корректная реализация метода Монте-Карло – метод случайной выборки каркаса (МВК). Подробный анализ эффективности метода, а также его описание представлено в работе [8]. Предложенный численный метод основан на получении большого числа N реализаций случайного процесса, который формируется таким образом, чтобы его вероятностные характеристики совпадали с искомыми величинами решаемой задачи. Используемые в модуле SimbigraphMotif алгоритмы дают несмещенную оценку числа подграфов с относительно небольшой дисперсией, что выгодно их отличает от алгоритмов, реализованных в системах MFINDER, FANMOD, igraph. Результаты тестирования показали, что разработанные алгоритмы расчета числа подграфов, использующие МВК, выполняют расчет значительно быстрее алгоритма Rand-ESU [6], реализованного в существующих программах igraph и Fanmod. Так, по сравнению с менее точным алгоритмом Rand-ESU (igraph) при расчетах частот встречаемости подграфы на трех вершинах в сети Twitter коэффициент ускорения равен 13,4, для сети пользователей электронной почты Email-EuAll – 1,7, а для сети пользователей социальной сетью G+ коэффициент ускорения достигает 3800.

Таблица 2
Время расчета частот подграфов на трех узлах, с помощью алгоритма Rand_ESU и алгоритма на основе МВК, N = 100 000

| Сеть | Rand-ESU FANMOD | Rand-ESU igraph | Время работы алгоритма, использующего МВК, с | | | |
|---|---|---|---|---|---|---|
| | | | 1 поток | 2 потока | 4 потока | 8 потоков |
| Political Blogs | 0,846 с | 0,118 с | 2,246 с | 1,393 с | 1,044 с | 0,881 с |
| Twitter | 294,971 с | 43,22 с | 8,535 с | 5,104 с | 3,716 с | 3,221 с |
| Email-EuAll | 102,434 с | 12,67 с | 33,930 с | 16,882 с | 12,015 с | 7,479 с |
| G+ | 2025698 с | 108074 с | 101,378 с | 61,589 с | 35,126 с | 28,318 с |
| Internet | 6,846 с | 0,696 с | 2,70 с | 1,51 с | 1,00 с | 0,81 с |
| Email-Enron | 4,3 с | 2,96 с | 1,87 с | 0,11 с | 0,74 с | 0,65 с |

В таблице 3 приведены затраты времени при использовании разработанных алгоритмов, использующих МВК, при числе опытов N = 100000 и N = 1000000. Для сравнения и оценки эффективности разработанных алгоритмов в табл. 2 приведено усредненное время расчета частот встречаемости подграфов на четырех вершинах, затраченное программой FANMOD при использовании алгоритма RAND-ESU с параметрами «по умолчанию».

Таблица 3
Усредненное время расчета частот подграфов на четырех вершинах в сети DBLP, полученное с помощью алгоритма RAND-ESU (FANMOD с параметрами «по умолчанию») и разработанных алгоритмов, использующих МВК

| Rand-ESU FANMOD | Время работы алгоритма, использующего МВК (N=100000) | | | | Время работы алгоритма, использующего МВК (N=1000000) | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 4 | 8 | 1 | 2 | 4 | 8 |
| 29,244 с | 4,370 с | 2,923 с | 2,312 с | 2,157 с | 30,815 с | 16,225 с | 9,803 с | 8,374 с |

**Заключение**

В статье предложен модуль SimbigraphMotif для расчета частот встречаемости подграфов в системе агентного моделирования SIMBIGRAPH. В предложенном модуле реализована корректная реализация метода Монте-Карло [8], а также использовались параллельные вычисления средствами Java 8 – Fork-Join Framework и JSR 335: Streams.

Тестирование модуля SimbigraphMotif для расчета точного числа частот встречаемости типовых подграфов показало, что разработанный функционал работает быстрее «прототипов», реализованных в программе FANMOD и библиотеки igraph для математической системы R. Полученное ускорение расчета зависит от размера и структуры графа.

Тестирование разработанного модуля для расчета частот встречаемости типовых подграфов на основе использования реализации метода Монте-Карло [8] показало, что разработанные алгоритмы работают быстрее и с большей точностью по сравнению с алгоритмом RAND-ESU (FANMOD и igraph). Так, при расчете числа подграфов на трех узлах для компонента социальной сети Google plus скорость работы модуля SimbigraphMotif при соизмеримой точности в 3800 раз быстрее, чем при использовании алгоритма RandESU, реализованного в библиотеке igraph. Большее ускорение можно достичь за счет распределенных вычислений, как это реализовано в работах [9-11].

**Литература**


1. **Задорожный В.Н., Долгушин Д.Ю., Юдин Е.Б.** Аналитико-имитационные методы решения актуальных задач системного анализа больших сетей, Омск: Изд-во ОмГТУ, 2013. 324 с.
2. **Юдин Е.Б., Задорожный В. Н., Пендер Е. А.** Случайные графы с нелинейным правилом предпочтительного связывания в системе агентного моделирования SIMBIGRAPH // Имитационное моделирование. Теория и практика / материалы 5-й Всерос. Конф. СПб: ФГУП ЦНИИТС, 2011. Т. 1. С. 425–429.
3. **Юдин Е.Б., Курчанов А.А.** Система агентного моделирования SimbiGraph // Сборник докладов шестой всероссийской научно-практической конференции «Имитационное моделирование. Теория и практика» (ИММОД-2013). Том 1. // ISBN 978-5-9690-0221-0 // Издательство «ФЭН» Академии наук РТ, Казань, 2013. С. 361–365.
4. **Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U.** Network motifs: simple building blocks of complex networks // Science. Oct 2002. Vol. 298 (5594). P. 824–827.
5. **Kashtan N., Itzkovitz S., Milo R., Alon U.** Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs // Bioinformatics. Jul. 2004. Vol. 20. № 11. P. 1746–1758.
6. **Wernicke S., Rasche F.** Fanmod: a tool for fast network motif detection // Bioinformatics. 2006. Vol. 22. №. 9. P. 1152–1153.
7. **Luis A. Meira A.**, et al acc-Motif: Accelerated Network Motif Detection // IEEE / ACM Trans. Comput. Biology Bioinform. 2014. Vol. 11 (5). P. 853–862.
8. **Yudin E. B., Zadorozhnyi V. N.** Statistical approach to calculation of number of network motifs // 2015 International Siberian Conference on Control and Communications, SIBCON 2015 – Proceedings : conference proceeding, 21-23 May 2015 / Omsk State Technical University. Omsk, 2015. P. 714–729. – DOI: 10.1109/SIBCON.2015.7147296.
9. **Pagh R. and Tsourakakis C. E.** Colorful triangle counting and a MapReduce implementation. Information Processing Letters, 112(7):277–281, Mar. 2012.
10. **Cordella L, Foggia P, Sansone C, and Vento M.** An improved algorithm for matching large graphs // Proc. of the 3rd IAPR TC-15 Workshop on Graph based Representations in Pattern Recognition, 149-159, 2001.
11. **Solnon C.** All Different-based Filtering for Subgraph Isomorphism, Artificial Intelligence 174 (12-13):850–864, 2010.