

УДК 519.2:519.67

АНАЛИЗ НЕСМЕЩЕННОСТИ И ЭФФЕКТИВНОСТИ ОЦЕНОК ЧАСТОТ ВСТРЕЧАЕМОСТИ СЕТЕВЫХ МОТИВОВ В СТАТИСТИЧЕСКИХ МЕТОДАХ РАСЧЕТА

М. Н. Юдина

*Омский государственный технический университет,
mg-and-all@mail.ru*

Рассмотрены статистические методы расчета частот встречаемости сетевых мотивов, в частности метод случайного выбора ребра, метод Верника – Раше, метод случайной выборки каркасов и комбинированный метод случайной выборки каркасов. Выполнен сравнительный анализ по показателям качества исследуемых статистических методов. Для комбинированного метода случайной выборки каркасов выведены математические выражения, позволяющие получить состоятельные, несмещенные и эффективные оценки частот встречаемости 4-мотивов.

Ключевые слова: статистические методы расчета, сетевые мотивы, несмещенные эффективные статистические оценки.

ANALYSIS OF UNBIASED AND EFFECTIVE ESTIMATES FOR NETWORK MOTIFS FREQUENCIES BY STATISTICAL METHODS OF CALCULATING

M. N. Yudina

*Omsk State Technical University,
mg-and-all@mail.ru*

The article describes statistical methods for calculating network motifs frequency of occurrences. In particular, the Edge Sampling method, the method by S. Wernicke and F. Rasche, the method of random sampling of frames and the mixed method of random sampling of frames are analyzed. A comparative analysis of the quality indicators of the investigated statistical methods is done. For the mixed method of random sampling of frames, the mathematical expressions that allow obtaining consistent, unbiased, and effective estimates of frequencies for the 4-motifs are derived.

Keywords: statistical methods, network motifs, unbiased effective statistical estimates.

Введение. В «Науке о сетях» большое распространение получил подход, заключающийся в анализе частот встречаемости сетевых мотивов [1–3].

Множество всех возможных попарно неизоморфных связных графов на трех вершинах называется множеством сетевых мотивов на трех вершинах (3-мотивов). Множество всех попарно неизоморфных связных графов на четырех вершинах называется множеством сетевых мотивов на четырех вершинах (4-мотивов). Так, на рис. 1 представлено множество 4-мотивов в неориентированном графе.

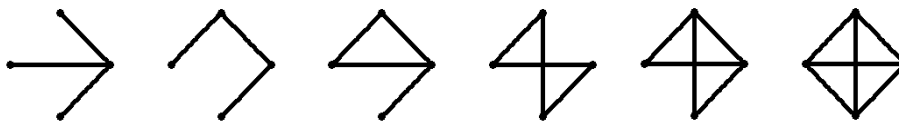


Рис. 1. Все возможные 4-мотивы на в неориентированном графе

Как правило, целью анализа является поиск таких сетевых мотивов, которые встречаются в графе G чаще, чем в рандомизированных версиях этого же графа G , т. е. поиск статистически значимых сетевых мотивов.

Задачу нахождения значимых сетевых k -мотивов в графе G можно разбить на три подзадачи:

1. Получение множества G_R рандомизированных версий графа G .
2. Подсчет в графе G и в каждом графе множества G_R частот встречаемости k -мотивов (при этом решается задача распознавания изоморфизма каждого найденного связного подграфа на k -вершинах одному из k -мотивов).
3. Расчет значимости сетевых мотивов на основе их встречаемости в графе G и в графах множества G_R .

Существует большое разнообразие программ, позволяющих рассчитывать частоты встречаемости сетевых мотивов. Некоторые из этих программ позволяют оценивать значимость сетевых мотивов. Среди таких программ следует назвать MFinder [2] (2003 год создания), Mavisto [4] (2005), Fanmod [5] (2006), NeMoFinder [6] (2006) и более поздняя реализации того же функционала в программе LaMoFinder, Kavosh [7] (2009), библиотека igraph для системы R (2013), AccMotif [8] (2013), MotifNet [3] (2017). Все перечисленные программы не позволяют за приемлемое время рассчитывать частоты встречаемости сетевых мотивов больших графов. Так, в работах [2, 4, 6–8], в которых предлагаются новые алгоритмы точного расчета 4-мотивов, исследуемым графом с наибольшим числом вершин является граф Foldoc. При этом граф Foldoc содержит всего 12 905 вершин и 109 092 дуги и описывает сеть терминов онлайн-библиотеки <http://www.foldoc.org/>. Узлы сети представляют собой термины онлайн-библиотеки, связь (X, Y) от термина X к Y существует, если термин Y используется для описания термина X . При расчетах на моноблоке HP Z1 с тактовой частотой 3,3 ГГц и 8 ГБ оперативной памяти наиболее быстрой из рассмотренных программ потребовалось следующее время для расчета: Kavosh – 559 с, Fanmod – 580 с, функции motifs библиотеки igraph – 18 с, программе AccMotif – 6 с. Из этих результатов следует, что самой быстрой является программа AccMotif. Но более тщательное экспериментальное исследование показало, что программа AccMotif требует значительно большего объема оперативной памяти, чем аналоги. При использовании указанной выше конфигурации оборудования программа AccMotif не позволяет эффективно работать с графами, содержащими более двадцати тысяч вершин, из-за недостатка оперативной памяти, таким образом, AccMotif не может работать с большими графами.

Отметим, что основной вычислительной проблемой при расчете значимых сетевых мотивов является сложность расчета частот встречаемости сетевых мотивов, а не генерация рандомизированных графов. Даже при использовании наиболее быстрых программ, позволяющих рассчитывать частоты встречаемости сетевых мотивов, требуется неприемлемо большое время.

Основные понятия. Введем следующие понятия. Реализацией i -го сетевого мотива в графе назовем подграф, полученный в результате статистического эксперимента для расчета частот встречаемости сетевых мотивов и изоморфный i -му сетевому мотиву. Пусть S – множество всех реализаций сетевых мотивов, содержащих заданное число k -вершин графа G . Множество S можно разбить на подмножества $S_i(G)$, в каждом из которых содержатся реализации i -го сетевого мотива в графе G , $i \in X$, где X – множество сетевых мотивов. Таким образом, относительная частота встречаемости i -го сетевого мотива:

$$C_i(G) = \frac{|S_i|}{|S(G)|} = |S_i| \times \left(\sum_{j \in X} |S_j| \right)^{-1}.$$

Для множества S реализаций k -мотивов, случайным образом взятых в G по некоторому алгоритму A , отображение $\hat{C}_i^A : (S, G \rightarrow [0,1])$ называется статистической оценкой относительной частоты встречаемости $C_i(G)$ сетевого мотива i , полученной при использовании алгоритма A . Отображение $\hat{n}_i^A : (S, G \rightarrow \{0, 1, \dots\})$ называется статистической оценкой абсолютной частоты $|S_i(G)|$ встречаемости i -го сетевого мотива в графе G .

Будем говорить, что полученные при использовании алгоритма A статистические оценки $\hat{C}_i^A$ или $\hat{n}_i^A$ несмещенные, если математическое ожидание оценки $M(\hat{C}_i^A) = C_i(G)$ или $M(\hat{n}_i^A) = |S_i|$.

Будем говорить, что статистические оценки являются состоятельными при использовании A , если при достаточно больших значениях объема выборки оценки $\hat{C}_i^A$ или $\hat{n}_i^A$ сходятся к истинному значению $C_k^i(G)$ и $|S_i|$ соответственно.

Будем говорить, что статистические оценки $\hat{C}_i^A$ или $\hat{n}_i^A$ являются эффективными, если они имеют наименьшую дисперсию среди всех соответствующих несмещенных оценок, вычисленных по выборкам одного и того же объема.

Метод случайного выбора ребра. Рассмотрим метод случайного выбора ребра (далее – СВР), предложенный Каштаном и соавторами в работе [9], для получения статистических оценок $\hat{C}_k^{CBP}(S, G)$ относительных частот встречаемости сетевых мотивов.

Метод СВР основан на реализации случайного процесса, который начинается с равновероятного выбора ребра. Далее из ближайшего окружения вершин $Nei(\{u, v\})$, инцидентных выбранному ребру $\{u, v\}$, выбирается третья вершина. Потом из ближайшего окружения этих трех вершин – четвертая и т. д. до получения реализации сетевого мотива желаемого размера k . Алгоритм, реализующий метод СВР, представлен на рис. 2.

НАЧАЛО

Ввод: граф $G = (V, E)$,

N – число опытов для расчета относительных частот встречаемости k -мотивов

Вывод: Счетчики $C[i]$ частот встречаемости мотива $i \in X$, где X – множество мотивов

1. Для всех $i \in X$ делай: $C[i] \leftarrow 0$

2. Для всех $j = 1, N$ делай

2.1. $\{u, v\} \leftarrow$ получить случайное ребро из E

2.2. $V' \leftarrow \{u, v\}$

2.3. Пока $|V'| \neq k$ делай

2.4. $\{u, v\} \leftarrow$ получить случайное ребро между V' и $Nei(V')$

2.5. $V' \leftarrow V' \cup \{u, v\}$

2.6. $Num = \text{GetISO}[G[V']]$ // получить класс изоморфизма $Num \in X$

2.7 $C[Num] = C[Num] + 1$

3. Для всех $i \in X$ делай $C[i] = C[i]/N$

КОНЕЦ

Рис. 2. Алгоритм, реализующий метод СВР

Заметим, что метод СВР является первым статистическим методом расчета частот встречаемости сетевых мотивов, но при использовании метода СВР одни сетевые мотивы систематически находятся чаще других [10], что привело к появлению алгоритмов, позволяющих уменьшить величину «смещения» [11].

Рассмотрим «смещение» получаемых оценок относительных частот встречаемости «треугольников» (полный граф на трех вершинах) при расчетах графа, изображенного на рис. 3.

Число реализаций сетевых мотивов на трех вершинах в графе на рис. 3 равно 15, причем только одна из этих реализаций является «треугольником». Очевидно, что истинное значение относительной частоты встречаемости сетевого мотива «треугольник» $C_{\Delta}^{CBP} = 1/15$.

Если же посчитать математическое ожидание относительной частоты встречаемости «треугольников», то получим $M(\hat{C}_{\Delta}^{CBP}) = 1/7 \cdot 1 + (2/7) \cdot (2/6) = 5/21$.

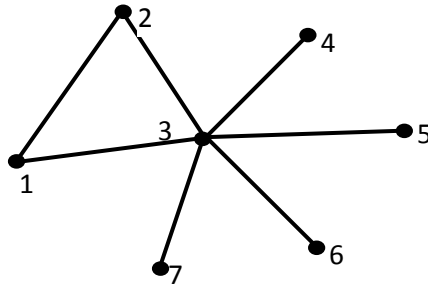


Рис. 3. Изображение графа, содержащего 15 реализаций 3-мотивов, среди которых один «треугольник»

Действительно, если в результате выполнения шага 2.1 алгоритма, реализующего метод СВР (рис. 2), будет выбрано ребро (1, 2), что произойдет с вероятностью $1/7$, то на следующих шагах алгоритма (шаги 2.2–2.6) с вероятностью, равной единице, будет найден сетевой мотив «треугольник». Если будет выбрано ребро (1, 3) или (2, 3), что может произойти с вероятностью $2/7$, то для получения реализации «треугольника» необходимо, чтобы на следующих шагах алгоритма было выбрано ребро (1, 2) или (2, 3) в первом случае и ребро (1, 2) или (1, 3) – во втором. Каждый такой выбор возможен с вероятностью $2/6$.

Поскольку $C_{\Delta}^{CBP} \neq M(\hat{C}_{\Delta}^{CBP})$ оценка частот встречаемости сетевых мотивов методом СВР является смещенной.

Метод Верника – Раше. Для перечисления всех сетевых мотивов размера k в графе $G = (V, E)$ в работе [5] предложен метод, который будем называть методом Верника – Раше по фамилиям его авторов. Также в работе [5] предложена статистическая версия метода, которую будем называть статистическим методом Верника – Раше. Чтобы лучше описать эти методы, необходимо использовать ряд соглашений. Во-первых, положим, что все вершины графа G помечены целыми числами. Для вершин v и множества $V' \subset V$ окрестность исключения $N_{\text{ext}}(v, V')$ определяется как $N_{\text{ext}}(v, V') = \text{Nei}(\{v\}) \setminus \text{Nei}(V')$. Иначе говоря, окрестность исключения – это множество тех вершин, которые входят в окрестность вершины v , но не являются смежными какой-либо вершине из множества V' .

Рассмотрим алгоритм метода Верника – Раше на рис. 4. В алгоритме рассчитываются частоты встречаемости k -мотивов. Работа алгоритма начинается с перебора всех вершин i графа G , $i = 1, |V|$; эти вершины добавляются в соответствующие множества V_{sub} . После выполнения строки 3 алгоритма в каждом таком множестве V_{sub} содержится одна вершина (на рис. 5 обозначение V_{sub}^i означает, что множество V_{sub} содержит вершину i). В дальнейшем во множествах V_{sub} будут содержаться две, три и т. д. вершины, входящие в соответствующие реализации сетевых мотивов. Для добавления новых вершин во множества V_{sub} используется вспомогательное множество V_{ext} , в которое на первом шаге добавляются все вершины из ближайшего окружения вершины i – $\text{Nei}(v_i)$, причем номер добавленной в V_{ext}^i вершины должен быть больше, чем i . Далее работа алгоритма выполняется рекурсивно. На каждом шаге рекурсивного вызова функции (см. функцию `ExtendSubgraphs` на рис. 4) происходит расширение множества V_{sub} . Расширение множества V_{sub} выполняется, пока число элементов в нем не достигнет значения k (т. е. найдена очередная реализация сетевого мотива) или пока множество V_{ext} не станет пустым. Получаемые множества V_{sub} можно представить в виде корневого дерева Верника – Раше. На первом ярусе дерева Верника – Раше в множествах V_{sub}^i задаются все реализации связных подграфов на одной вершине, на втором – на двух вершинах V_{sub}^{ij} (т. е. все ребра $\{i, j\}$), на третьем – на трех вершинах V_{sub}^{ijk} и т. д. (рис. 5).

Начало алгоритма метода Верника – Раше

Ввод: граф $G(V, E)$, целое $1 \leq k \leq |V|$

Вывод: Все сетевые мотивы в графе G , которые содержат k вершин

1. Для всех $v \in V$ делай
2. $V_{\text{ext}} \leftarrow \{u \in \text{Nei}(v) : u > v\}$
3. $V_{\text{sub}} = \{v\}$
4. Выполнить процедуру $\text{ExtendSubgraphs}(V_{\text{sub}}, V_{\text{ext}}, v)$

Конец алгоритма метода Верника-Раше

Начало процедуры $\text{ExtendSubgraphs}(V_{\text{sub}}, V_{\text{ext}}, v)$

E1. Если $|V_{\text{sub}}| = k$ то верни $G[V_{\text{sub}}]$ и Конец

E2. Пока $V_{\text{ext}} \neq \emptyset$ делай

E3. Удалить случайно выбранную вершину w из V_{ext}

E4. $V'_{\text{ext}} \leftarrow V_{\text{ext}} \cup \{u \in \text{Nei}(w, V_{\text{sub}}) : u > v\}$

E5. Выполнить процедуру $\text{ExtendSubgraphs}(V_{\text{sub}} \cup \{w\}, V'_{\text{ext}}, v)$

Конец процедуры $\text{ExtendSubgraphs}(V_{\text{sub}}, V_{\text{ext}}, v)$

Рис. 4. Алгоритм, реализующий метод Верника – Раше

В процедуре $\text{ExtendSubgraphs}(V_{\text{sub}}, V_{\text{ext}}, v)$ множество V_{sub} расширяется путем добавления одной вершины из множества V_{ext} – вершины w . После этого множество V_{ext} изменяется следующим образом. Во-первых, из него удаляется добавленная во множество V_{sub} вершина w , во-вторых, в него добавляются вершины, метка которых больше метки вершины v и которые находятся в окрестности исключения w относительно множества V_{sub} .

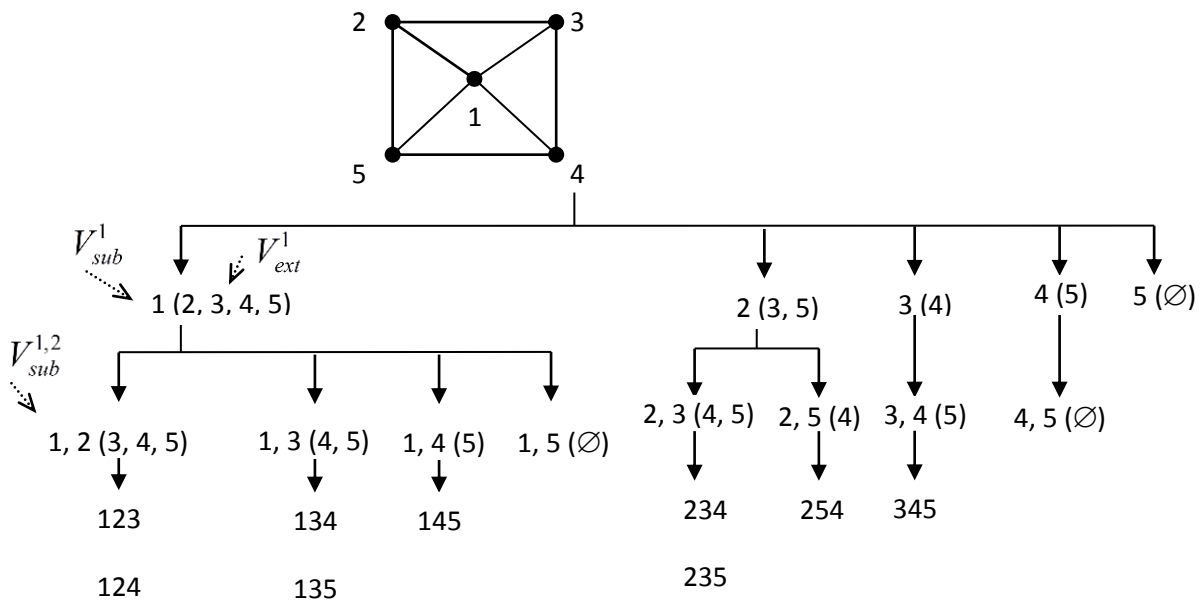


Рис. 5. Граф и соответствующее ему дерево Верника – Раше

Предложенный метод перебирает все сетевые мотивы заданного размера, не допуская перечисления одних и тех же сетевых мотивов больше одного раза. В работе [12] предложена итерационная версия метода Верника – Раше [13].

Если в алгоритме метода Верника – Раше заменить «Выполнить процедуру» в строках 4 и E5 на «Выполнить процедуру с вероятностью p_d », где $d = 1$ в строке 4 и $d = |V_{\text{sub}}| + 1$ в строке E5, то мы получим алгоритм статистического метода Верника – Раше [5].

Однако, вопреки утверждениям авторов [5], метод Верника – Раше в общем случае дает смещенные оценки частот встречаемости сетевых мотивов.

Ранее на большом количестве проведенных расчетов в работе [14] было показано, что использование метода Верника – Раше приводит к получению смещенных оценок частоты встречаемости k -мотивов. Рассмотрим здесь теоретическое доказательство. Действительно, как можно видеть, в графе на рис. 5 имеются 10 различных реализаций 3-мотивов, из них 4 реализации сетевого мотива «треугольник» (выделены фоном на рис. 5). Истинное значение частоты встречаемости сетевого мотива «треугольник» – $C_{\Delta} = 4/10$. При этом если вероятность выполнения строки 4 в алгоритме на рис. 4 равна p , то математическое ожидание частоты встречаемости сетевого мотива «треугольник» по формуле полной вероятности:

$$M(C_{\Delta}^{BP}) = \sum_{i=1}^{N_C} P_i \times C_i, \quad (1)$$

где N_C – количество различных значений относительных частот встречаемости «треугольников», которые могут быть получены в результате проведения статистического эксперимента,

P_i – вероятность получения значения C_i .

При подсчете суммы (1) можно рассматривать только ненулевые множители C_i . Для графа на рис. 5 таких значений всего четыре. Во-первых, относительная частота встречаемости «треугольников» будет ненулевой, если в результате выполнения строки 4 останется только вершина «1», вероятность этого события равна p $(1-p)^2$, а значение относительной частоты встречаемости «треугольников» при этом будет равно $4/6$. Во-вторых, относительная частота встречаемости «треугольников» будет ненулевой, если в результате выполнения строки 4 на рис. 4 останутся вершины «1» и «2», вероятность этого события равна $p^2(1-p)$, а значение относительной частоты встречаемости «треугольников» при этом равно $4/9$. В-третьих, относительная частота встречаемости «треугольников» будет ненулевой, если в результате выполнения строки 4 останутся вершины «1» и «3», вероятность этого события также равна $p^2(1-p)$, а значение относительной частоты встречаемости «треугольников» при этом будет равно $4/7$. Наконец, относительная частота встречаемости «треугольников» будет ненулевой, если в результате выполнения строки 4 останутся вершины «1», «2» и «3», вероятность этого события равна p^3 , а значение относительной частоты встречаемости «треугольников» при этом будет равно $4/10$.

Соответственно, математическое ожидание частоты встречаемости сетевого мотива «треугольник» C_{Δ}^{BP} будет:

$$M(C_{\Delta}^{BP}) = \sum_{i=1}^{N_C} p_i \cdot C_i = (4/6)p(1-p)^2 + (4/9 + 4/7)p^2(1-p) + (4/10)p^3.$$

На рис. 6 можно видеть, что при различных значениях вероятности p частота встречаемости «треугольников» оказывается заниженной, достигая истинного значения $C_{\Delta} = 4/10$ при вероятности выполнения строки 4 алгоритма на рис. 4 равной единице.

Таким образом, оценка частот встречаемости, получаемая статистическим методом Верника – Раше, в общем случае является смещенной.

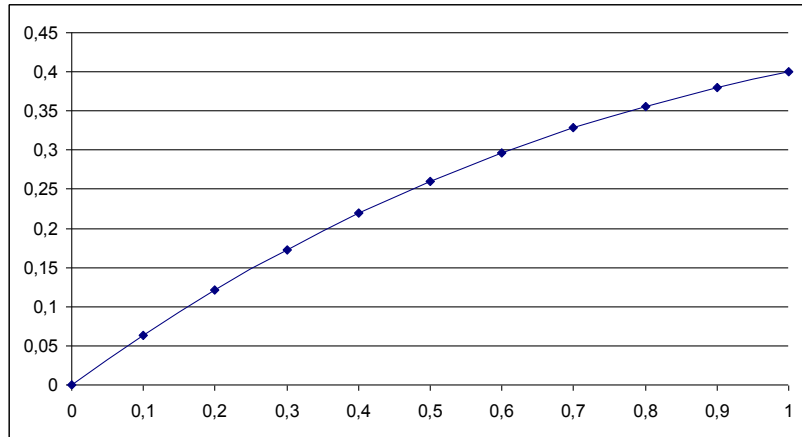


Рис. 6. График зависимости математического ожидания частоты встречаемости «треугольника» от вероятности p выполнения строки 4 алгоритма на рис. 4

Представленное в работе [10] доказательство теоремы о несмещенности оценки в методе Верника – Раше является ошибочным. Это доказательство основывается на том, что вероятность «сохранения» любого k -мотива в соответствии со статистическим методом Верника – Раше одинакова для всех сетевых мотивов на заданном числе вершин, что действительно так. И для графа, представленного на рис. 5, математическое ожидание числа «треугольников» при вероятности p выполнения строки 4 алгоритма на рис. 4 действительно будет $M(\Delta) = p(4 + 0 + 0 + 0 + 0) = 4p$. Математическое ожидание числа всех 3-мотивов $M(N) = p(6 + 3 + 1 + 0 + 0 + 0) = 10p$. Однако ошибочно полагать, что относительную частоту реализации «треугольников» можно считать по формуле $M(C_\Delta) = M(\Delta)/M(N)$, поскольку найденное число «треугольников» и найденное число реализаций всех 3-мотивов в статистическом эксперименте метода Верника – Раше – это зависимые случайные величины. Единственная корректная реализация статистического метода Верника – Раше, при которой отсутствует смещение оценок (поскольку все реализации сетевых мотивов, полученные в результате выполнения метода Верника-Раше, независимы), – это реализация, когда отсекаются ветви дерева Верника – Раше только на последнем ярусе.

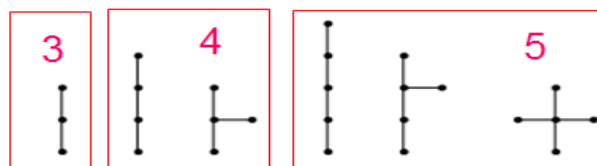


Рис. 7. Остовные деревья на 3, 4 и 5 вершинах

Метод случайной выборки каркасов. Метод случайной выборки каркасов (далее – СВК) является реализацией метода Монте-Карло, в котором выборочное пространство определяется как множество экземпляров одного и того же каркаса [15], содержащихся в рассчитываемом графе и наделенных равными вероятностями выбора. Чтобы получить оценки встречаемости для всех возможных k -мотивов, очевидно, необходимо использовать столько каркасов, сколько существует попарно неизоморфных остовных деревьев на k -вершинах. В неориентированном графе на трех вершинах существует только одно остовное дерево (рис. 7), на четырех вершинах – 2 остовных дерева, на пяти – 3, на шести – 6.

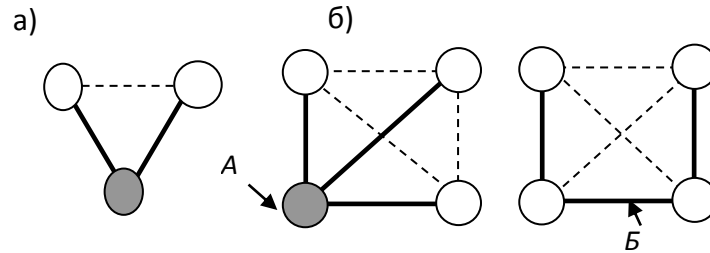


Рис. 8. Каркасы на трех и четырех вершинах,
жирная линия – ребро, присутствующее в экземпляре каркаса;
пунктирная – ребро, которое может быть найдено в сетевом мотиве

Метод СВК, предложенный в работе [15], позволяет рассчитывать частоты встречаемости сетевых мотивов на трех и четырех вершинах в неориентированном графе. Для этого используются три разных каркаса. Первый – каркас «вилка», экземпляры которого представляют собой путь длиной в два ребра. В каркасе «вилка» вершина, инцидентная обоим ребрам каркаса, считается центром каркаса, (рис. 8а). Каркас «вилка» используется для расчета 3-мотивов. Для расчета 4-мотивов используются два каркаса, каждый из которых содержит три связных ребра. Эти каркасы изображены на рис. 8б: слева – каркас «ветвление», справа – каркас «цепочка». Для «ветвления» центром каркаса (рис. 8б слева) называется вершина А, инцидентная всем трем ребрам каркаса. Для «цепочки» центром считается ребро Б, смежное двум другим ребрам каркаса (рис. 8б справа).

Основой метода СВК является равновероятный выбор в графе экземпляров каркаса и основанный на этом расчет несмещенной оценки доли $\hat{q}$ экземпляров каркаса, «несущих» соответствующий мотив. Поскольку общее число N экземпляров каркаса в графе известно точно (оно легко рассчитывается), несмещенная оценка числа $\hat{n}$ реализаций мотива в графе определяется в виде $\hat{n} = \hat{q} N$.

Алгоритм статистического расчета абсолютной частоты встречаемости сетевых мотивов методом СВК представлен на рис. 9.

НАЧАЛО

Ввод: граф $G(V, E)$, массив поправочных коэффициентов $\text{coef}[i]$, $i \in X$,

где X – множество сетевых мотивов

Вывод: оценки абсолютных частот встречаемости сетевых мотивов $n[i]$, $i \in X$

1. Для всех $i \in X$ делай $n[i] \leftarrow 0$; $C[i] \leftarrow 0$;

2. Рассчитать общее число каркасов N^+

3. Для $j = 1, \dots, N$ выполнить

3.1. Выбрать равновероятно каркас C_j , содержащий множество вершин V'

3.2. $\text{Num} = \text{GetISO}[G[V']]$ // получить класс изоморфизма $\text{Num} \in X$

3.2. $C[\text{Num}] = C[\text{Num}] + 1$;

3. Для всех $i \in M$ делай $n[i] \leftarrow N^+ C[i] / \text{coef}[i]$.

КОНЕЦ

Рис. 9. Статистический расчет для получения оценки абсолютной частоты встречаемости сетевых мотивов методом СВК

В алгоритме на рис. 9 на шаге 2.1 равновероятно выбираются экземпляры каркаса, определяется найденная на каждом из этих экземпляров реализация сетевого мотива и наращивается счетчик найденных реализаций сетевых мотивов. После выполнения статистического эксперимента (на шаге 2 алгоритма на рис. 9) на шаге 3 с учетом поправочных коэффициентов рассчитываются оценки абсолютных частот встречаемости сетевых мотивов.

Метод СВК позволяет получить несмещенные и эффективные оценки только числа встречаемости 3-мотивов. Для расчета 4-мотивов, с одной стороны, использование одного вида каркаса («ветвление» или «цепочка») не позволяет получить оценки для частот встречаемости всех возможных 4-мотивов. С другой стороны, при использовании каркасов «ветвление» и «цепочка» для некоторых 4-мотивов могут быть получены две различные оценки числа реализаций мотива.

Комбинированный метод случайной выборки каркасов. В комбинированном методе случайной выборки каркасов (в методе КСВК) итоговая оценка числа встречаемости мотивов строится виде линейной комбинации оценок, полученных при использовании различных каркасов. Для 4-мотивов в методе КСВК используется формула:

$$\hat{n} = \hat{n}_A + \lambda(\hat{n}_B - \hat{n}_A), \text{ где } 0 < \lambda < 1. \quad (2)$$

Заметим, что случайные величины $\hat{n}_A$ и $\hat{n}_B$ при большом числе испытаний N_A^+ и N_B^+ соответственно имеют нормальное распределение, что вытекает из центральной предельной теоремы, поскольку оценки получены при использовании схемы Бернулли, причем выполняется большое число независимых равновероятных выборов сетевых мотивов. Величина $\hat{n}$ тоже имеет нормальное распределение, поскольку строится как линейная комбинация двух нормальных величин. Для любого нормального закона распределения вероятность, что значение случайной величины отклонится от ее математического ожидания более чем на три среднеквадратических отклонения, не превышает 0,28 %, т. е. пренебрежимо мала. Поэтому в качестве абсолютной погрешности для случайной величины $\hat{n}$ можно рассматривать величину 3σ . В качестве аналога относительной погрешности для величины $\hat{n}$ следует рассматривать ее коэффициент вариации $v(\hat{n}) = \sigma(\hat{n})/M(\hat{n})$, а в качестве самой относительной погрешности – величину $3v$.

В работе [16] приведен критерий выбора λ для минимизации коэффициента вариации (относительной погрешности вычислений) интегральной оценки числа встречаемости сетевого мотива $\hat{n}$. В данной работе рассмотрим такой выбор значения параметра λ , который минимизирует абсолютную погрешность.

Для минимизации абсолютной погрешности расчетов и получения эффективной статистической оценки частот встречаемости 4-мотивов параметр λ в выражении (2) определим из условия минимизации дисперсии $D(\hat{n}(\lambda))$.

Случайная величина $\hat{n}$ определяется через линейную комбинацию независимых случайных величин в выражении (2), ее математическое ожидание $M(\hat{n})$ при любом $0 < \lambda < 1$ будет $M(\hat{n}) = M(\hat{n}_A + \lambda(\hat{n}_B - \hat{n}_A)) = M((1-\lambda)\hat{n}_A + \lambda\hat{n}_B) = (1-\lambda)M(\hat{n}_A) + \lambda M(\hat{n}_B) = (1-\lambda)M_A + \lambda M_B$, а дисперсия:

$$D(\hat{n}(\lambda)) = D((1-\lambda)\hat{n}_A) + D(\lambda\hat{n}_B) = (1-\lambda)D(\hat{n}_A) + \lambda D(\hat{n}_B) = (1-\lambda)^2 D_A + \lambda^2 D_B. \quad (3)$$

Дифференцируя (3) по λ и приравнявая полученное выражение к нулю, получим:

$$((1-\lambda_{opt})^2 D_A + \lambda_{opt}^2 D_B)' = -2D_A + 2\lambda_{opt} D_A + 2\lambda_{opt} D_B = 0.$$

Откуда выразим искомое λ_{opt} :

$$\lambda_{opt} = \frac{D_A}{D_B + D_A}. \quad (4)$$

Найденное значение λ_{opt} действительно является минимумом, поскольку вторая производная $D(\hat{n}(\lambda))'' = 2D_A + 2D_B$ положительна, в том числе в точке λ_{opt} . Случай, когда дисперсии

D_A и D_B одновременно равны нулю следует рассмотреть отдельно. Эта ситуация возможна только тогда, когда все связные подграфы на четырех вершинах в исследуемом графе являются реализацией искомого 4-мотива и находятся одновременно как при использовании каркаса «цепочка», так и при использовании каркаса «ветвление»; в этом случае $M(\hat{n})=M(\hat{n}_A)=M(\hat{n}_B)$. Иначе, при $0 < \lambda < 1$ $M(\hat{n}(\lambda))=(1-\lambda_{opt})M_A+\lambda_{opt}M_B$, где λ_{opt} вычисляется по формуле (4).

Сравнение статистических методов расчета по показателям качества. При анализе статистических методов расчета частот встречаемости сетевых мотивов выявлено, что метод Верника – Раше (в общем случае) не позволяет получить несмещенные оценки частот встречаемости сетевых мотивов. Выявлен единственный режим работы статистического метода Верника – Раше, позволяющий получить состоятельную оценку относительных частот встречаемости сетевых мотивов, причем для его реализации необходимо построение всех ярусов дерева Верника – Раше, за исключением последнего его яруса, который достраивается частично. Назовем этот режим работы методом Верника – Раше (*). Метод СВК дает несмещенные оценки частот встречаемости сетевых мотивов, однако при использовании различных каркасов могут быть получены различные оценки, а использование одного из каркасов не позволяет получить статистические оценки числа всех возможных 4-мотивов. Метод КСВК решает эту проблему, представляя оценку числа 4-мотивов в виде линейной комбинации оценок, полученных при использовании различных каркасов. В таблице представлены основные характеристики статистических методов расчета 4-мотивов.

Таблица

Характеристики статистических методов расчета числа встречаемости 4-мотивов

Название метода	Оценка относительных частот встречаемости 4-мотивов			Оценка числа 4-мотивов		
	Состоятельность	Несмещенность	Эффективность	Состоятельность	Несмещенность	Эффективность
Метод СВР	нет	нет	нет	–	–	–
Метод Верника – Раше	–	нет	нет	–	–	–
Метод Верника – Раше (*)	да	да	да	–	–	–
Метод СВК	да	да/нет	да/нет	да	да/нет	да/нет
Метод КСВК	да	да	да	да	да	да

Выводы. Возникший в рамках «Науки о сетях» подход к исследованию больших сетей путем анализа частот встречаемости сетевых мотивов используется для сетей белковых взаимодействий и геномных сетей [1], трофических сетей [2], сетей химических реакций [17–18] и сетей передачи электроэнергии [19], сетей телекоммуникаций и транспортных сетей [20–21]. Этот топологический подход позволяет прогнозировать белок-белковые взаимодействия [22], пути конструирования новых химических взаимодействий [23]. Сетевые мотивы широко используются при классификации сетевых структур [2, 21].

Тем не менее при большом числе вершин и ребер анализируемых графов точный расчет частот встречаемости 4-мотивов становится практически нереализуемым. Поэтому широкое распространение получили статистические методы расчета: метод случайного выбора ребра (СВР), метод Верника – Раше, метод случайной выборки каркасов (СВК), комбинированный метод случайной выборки каркасов (КСВК). Однако характеристики (состоятельность, несмещенность, эффективность) получаемых статистическими методами оценок все еще требуют углубленного анализа. В данной работе получены следующие новые результаты:

1. Проведен анализ известных статистических методов расчета частот встречаемости сетевых мотивов, таких как метод СВР, метод Верника – Раше, метод СВК и КСВК. Впервые показывается, что статистический метод Верника – Раше не гарантирует получения несмещенных оценок частот встречаемости сетевых мотивов. Выявлен режим работы метода Вер-

ника – Раше, при использовании которого будут получены несмещенные оценки частот встречаемости сетевых мотивов.

2. Метод СВК позволяет получить несмещенные и эффективные оценки частот встречаемости только 3-мотивов. Для расчета 4-мотивов, с одной стороны, использование одного вида каркаса («ветвление» или «цепочка») не позволяет получить оценки для частот встречаемости всех возможных 4-мотивов. С другой стороны, при использовании каркасов «ветвление» и «цепочка» для некоторых 4-мотивов могут быть получены две различные оценки числа реализаций мотива. Впервые для метода КСВК приводится режим работы для получения состоятельных, несмещенных и эффективных оценок частот встречаемости 4-мотивов.

Следует выделить следующие перспективные направления исследований в области разработки статистических методов расчета сетевых мотивов:

- разработка алгоритмов, обеспечивающих остановку расчетов при достижении заданной точности расчета;
- разработка программных комплексов специализированных программ [24–25];
- использование параллельных и распределенных алгоритмов для ускорения расчетов;
- разработка веб-сервиса для расчета сетевых мотивов методом КСВК (по аналогии с веб-сервисом MotifNet [3], реализующим статистический метод Верника – Раше);
- доработка метода КСВК для расчета частот встречаемости сетевых мотивов более высокого порядка.

Литература

1. Ma'ayan A., Jenkins S. L., Neves S., Hasseldine A., Grace E., Dubin-Thaler B., Eungdamrong N. J., Weng G., Ram P. T., Rice J. J., Kershenbaum A., Stolovitzky G. A., Blitzer R. D., Iyengar R. Formation of Regulatory Patterns During Signal Propagation in a Mammalian Cellular Network // *Science*. 2005. Vol. 310. P. 1078–1083.
2. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U. Network Motifs: Simple Building Blocks of Complex Networks // *Science*. 2002. Vol. 594, No. 298. P. 824–827.
3. Smoly I. Y., Lerman E., Ziv-Ukelson M., Yeager-Lotem E. MotifNet: A Web-Server for Network Motif Analysis // *Bioinformatics*. 2017. Vol. 33, No. 12. P. 1907–1909.
4. Schreiber F., Schwobbermeyer H. Frequency Concepts and Pattern Detection for the Analysis of Motifs in Networks // *Transactions on Computational Systems Biology III*. 2005. Vol. 3737. P. 89–104.
5. Wernicke S., Rasche F. FANMOD: a Tool for Fast Network Motif Detection // *Bioinformatics*. 2006. Vol. 22, No. 9. P. 1152–1153.
6. Chen J., Hsu W., Lee M. L., Ng S.-K. NeMoFinder: Dissecting Genome-Wide Protein-Protein Interactions with Meso-Scale Network Motifs // *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Philadelphia, Pennsylvania, USA, 2006. P. 106–115.
7. Kashani Z. R., Ahrabian H., Elahi E., Nowzari-Dalini A., Ansari E. S., Asadi S., Mohammadi S., Schreiber F., Masoudi-Neja A. Kavosh: a New Algorithm for Finding Network Motifs // *BMC Bioinformatics*. 2009. No. 10. P. 318.
8. Meira L. A. A., Máximo V. R., Fazenda Á. L., da Conceição A.F. acc-Motif: Accelerated Network Motif Detection // *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Vol. 11, No. 5. P. 853–862.
9. Kashtan N., Itzkovitz S., Milo R., Alon U. Efficient Sampling Algorithm for Estimating Subgraph Concentrations and Detecting Network Motifs // *Bioinformatics*. 2004. Vol. 20, No. 11. P. 1746–1758.
10. Wernicke S. A Faster Algorithm for Detecting Network Motifs // *Lecture Notes in Bioinformatics*. 2005. Vol. 3692. P. 165–177.
11. Ray A., Holder L., Bifet A. Efficient Frequent Subgraph Mining on Large Streaming Graphs // *Intelligent Data Analysis*. 2019. Vol. 23, No. 1. P. 103–132.

12. Itzhack R., Mogilevski Y., Louzoun Y. An Optimal Algorithm for Counting Network Motifs // *Physica A*. 2007. Vol. 381. P. 482–490.
13. Wernicke S. Comment on “An optimal algorithm for counting networks motifs” // *Physica A*. 2011. Vol. 390. P. 143–145.
14. Yudina M. N. Assessment of Accuracy in Calculations of Network Motif Concentration by Rand ESU algorithm // *Journal of Physics: Conference Series*. 2019. Vol. 1260. P. 022012.
15. Yudin E. B., Zadorozhnyi V. N. Statistical Approach to Calculation of Number of Network Motifs // *Proceedings – International Siberian Conference on Control and Communications, SIBCON*. Omsk, 2015. P. 1–4.
16. Yudina M. N., Zadorozhnyi V. N., Yudin E. B. Mixed Random Sampling of Frames Method for Counting Number of Motifs // *Journal of Physics: Conference Series*. 2019. Vol. 1260. P. 022013.
17. Jacob P. M., Lapkin A. Statistics of the Network of Organic Chemistry // *Reaction Chemistry & Engineering*. 2018. No. 3. P. 102–118.
18. Menon A., Krdzavac N. B., Kraft M. From Database to Knowledge Graph – Using Data in Chemistry // *Current Opinion in Chemical Engineering*. 2019. Vol. 26. P. 33–37.
19. Dey A. K., Gel Y. R., Poor H. V. What Network Motifs Tell us About Resilience and Reliability of Complex Networks // *Proceedings of the National Academy of Sciences*. 2019. Vol. 116, No. 39. P. 19368–19373.
20. Stone L., Simberloff D., Artzy-Randrup Y. Network Motifs and Their Origins // *PLoS Computational Biology*. 2019. Vol, 15. No. 4. P. e1006749.
21. Jain D., Patgiri R. Network Motifs: A Survey // *3rd International Conference on Advances in Computing and Data Sciences, ICACDS 2019*. Ghazibad; India; April 2019. Vol. 1046. P. 80–91.
22. Albert I., Albert R. Conserved Network Motifs Allow Protein–Protein Interaction Prediction // *Bioinformatics*. 2004. Vol. 20, No. 18. P. 3346–3352.
23. Müller I. E., Rubens J. R., Jun T., Graham D., Xavier R., Lu T. K. Gene Networks that Compensate for Crosstalk with Crosstalk // *Nature Communications*. 2019. Vol. 10. 8 p.
24. Юдина М. Н. Узлы в социальных сетях: меры центральности и роль в сетевых процессах // *Омск. науч. вестн.. Сер. Приборы, машины и технологии*. 2016. Т. 148, № 4. С. 161–165.
25. Юдина М. Н. Комплекс программных библиотек для анализа молекулярных сетей клетки // *Омск. науч. вестн.. Сер. Приборы, машины и технологии*. 2018. Т. 162, № 6. С. 265–217.