Math-Net.Ru

Общероссийский математический портал

А. М. Зубков, В. И. Круглов, Повторения цепочек на бинарном деревесе случайными метками вершин, *Дискрет. матем.*, 2015, том 27, выпуск 4, 38–48

DOI: <https://doi.org/10.4213/dm1346>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением <http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 52.23.180.231

26 апреля 2023 г., 15:13:20

Повторения цепочек на бинарном дереве со случайными метками вершин

© 2015 г. А. М. Зубков*, В. И. Круглов**

Для полного бинарного дерева, каждой вершине которого случайно, равномерно и независимо от остальных вершин присвоена метка из конечного алфавита, получены формулы для математических ожиданий чисел пар цепочек с одинаково помеченными вершинами. Доказана теорема о сходимости распределений числа пар непересекающихся цепочек с одинаково помеченными вершинами к сложному пуассоновскому распределению.

Исследование выполнено за счет гранта Российского научного фонда (проект № 14-50-00005).

Ключевые слова: бинарные деревья с помеченными вершинами, повторения цепочек, пуассоновская аппроксимация

Задачи, связанные с повторениями цепочек в строках символов, рассматривались разными авторами (см., например, [2], [5], [8], [4], [9]), в частности, в связи с анализом алгоритмов, построением статистических критериев, анализом генетических последовательностей. Решаются также аналогичные задачи путей на ориентированных графах, вершинам которых сопоставлены символы конечного алфавита (например, в [3], [6] рассматриваются алгоритмы поиска заданной комбинации в дереве, а в [7] проводится вероятностный анализ таких алгоритмов и найдены асимптотики вероятности существования заданной комбинации в дереве). Здесь мы рассматриваем простейший случай повторений цепочек знаков на вершинах двоичного дерева, ребра которого ориентированы от корня.

1. Формулировки результатов

Рассмотрим полное бинарное дерево T_2^n высоты n , состоящее из корня $*$ и n слоев вершин; 2^k элементов множества $I^{(k)}$ вершин k -го ($k = 1, 2, \dots, n$) слоя можно пронумеровать двоичными наборами $i = (i_1, i_2, \dots, i_k) \in \{0, 1\}^k$ так, что из вершины $*$ (образующей слой $I^{(0)}$) выходят два ребра, соединяющие ее с вершинами слоя $I^{(1)}$,

*Место работы: Математический институт им. В.А. Стеклова Российской академии наук, e-mail: zubkov@mi.ras.ru

**Место работы: Математический институт им. В.А. Стеклова Российской академии наук, e-mail: kruglov@mi.ras.ru

и любая вершина $i = (i_1, i_2, \dots, i_k) \in I^{(k)}, 1 \leq k \leq n-1$, соединена двумя выходящими из нее ребрами с вершинами $i' = (i_1, i_2, \dots, i_k, 0)$ и $i'' = (i_1, i_2, \dots, i_k, 1)$ слоя $I^{(k+1)}$. В вершину $i = (i_1, i_2, \dots, i_k)$ входит ребро из вершины $i^- = (i_1, i_2, \dots, i_{k-1})$ при $k > 1$ и из корня $* = (0)^- = (1)^-$ при $k = 1$. Таким образом, каждая вершина i дерева T_2^n определяет поддерево, состоящее из этой вершины и вершин следующих слоев, соединенных с i одним или несколькими ребрами. На множестве вершин дерева T_2^n зададим естественное лексикографическое упорядочение, считая, что $i = (i_1, \dots, i_k) \prec j = (j_1, \dots, j_h)$, если либо $i = *, j \neq *$, либо $1 \leq k < h$, либо $1 \leq k = h$ и $\sum_{m=1}^k i_m 2^{k-m} < \sum_{m=1}^h j_m 2^{h-m}$. Множество упорядоченных пар непересекающихся цепочек $(C_i, C_j), i \prec j$, обозначим $\mathcal{P}$.

Цепочкой C_i длины l , начинающейся в вершине $i = (i_1, i_2, \dots, i_k) \in I^{(k)}, k \geq 0$, будем называть любую последовательность l вершин

$$(i_1, i_2, \dots, i_k), (i_1, i_2, \dots, i_k, i_{k+1}), \dots, (i_1, i_2, \dots, i_k, i_{k+1}, \dots, i_{k+l-1}),$$

соединенных ребрами. Будем обозначать эти вершины цепочки C_i соответственно $C_i[0], C_i[1], \dots, C_i[l-1]$. Вершину i будем называть начальной вершиной цепочки C_i , а вершину $(i_1, i_2, \dots, i_k, i_{k+1}, \dots, i_{k+l-1})$ — ее конечной вершиной. Очевидно, конечная вершина и число l однозначно определяют цепочку. Это позволяет определить отношение порядка на множестве цепочек фиксированной длины l : $C_i \prec C_j$ тогда и только тогда, когда $C_i[l-1] \prec C_j[l-1]$.

Легко проверить, что цепочки C_i и C_j пересекаются тогда и только тогда, когда либо $C_i[0] \in C_j$, либо $C_j[0] \in C_i$.

Общее число вершин в дереве T_2^n равно $1 + 2 + \dots + 2^n = 2^{n+1} - 1$, а общее число цепочек длины l в дереве T_2^n равно числу их конечных вершин $|\cup_{j=l-1}^n I^{(j)}| = 2^{l-1} + \dots + 2^n = 2^{n+1} - 2^{l-1}$.

Пусть каждой вершине i дерева T_2^n присвоена случайная метка $m(i)$, принимающая значения в множестве $\{1, \dots, d\}$, причем значения $m(i)$, $i \in T_2^n$, независимы и $\mathbf{P}\{m(i) = j\} = \frac{1}{d}, j \in \{1, \dots, d\}$, для всех $i \in T_2^n$. Таким образом, каждой цепочке C_i длины l сопоставляется случайная последовательность меток

$$M(C_i) = (m(C_i[0]), m(C_i[1]), \dots, m(C_i[l-1])).$$

Очевидно, что если цепочки $C_{i_1}, \dots, C_{i_s}$ попарно не пересекаются, то соответствующие им последовательности меток $M(C_{i_1}), \dots, M(C_{i_s})$ независимы в совокупности и имеют равновероятное распределение на множестве $\{1, \dots, d\}^l$.

Нас интересует распределение числа пар $(C_i, C_j), i \prec j$, цепочек длины l в дереве T_2^n с одинаковыми последовательностями меток $(M(C_i) = M(C_j))$. Общее число таких пар равно, очевидно,

$$V_{n,l} = \sum_{(C_i, C_j) \in \mathcal{P}} \mathbb{I}\{M(C_i) = M(C_j)\}; \quad (1)$$

значение d мы будем считать фиксированным при изменении n и l .

Так как вероятности событий $\{M(C_i) = M(C_j)\}$ зависят от взаимного расположения цепочек C_i и C_j , имеет смысл разбить сумму $V_{n,l}$ на несколько слагаемых

(сумму $V_{n,l}^{(0)}$ по непересекающимся цепочкам, сумму $V'_{n,l}$ по пересекающимся цепочкам с разными началами, суммы $V''_{n,l,k}$ по цепочкам с общими началами):

$$\begin{aligned}
 V_{n,l} &= V_{n,l}^{(0)} + V'_{n,l} + \sum_{k=1}^{l-1} V''_{n,l,k}, \\
 V_{n,l}^{(0)} &= \sum_{(C_i, C_j) \in \mathcal{P}: C_i \cap C_j = \emptyset} \mathbb{I}\{M(C_i) = M(C_j)\}, \\
 V'_{n,l} &= \sum_{(C_i, C_j) \in \mathcal{P}: C_i \cap C_j \neq \emptyset, C_i[0] \neq C_j[0]} \mathbb{I}\{M(C_i) = M(C_j)\}, \\
 V''_{n,l,k} &= \sum_{(C_i, C'_i) \in \mathcal{P}: |C_i \cap C'_i| = k, C_i[0] = C'_i[0]} \mathbb{I}\{M(C_i) = M(C'_i)\}, \quad 1 \leq k < l.
 \end{aligned} \tag{2}$$

Теорема 1. *Справедливы равенства*

$$\mathbf{E}V_{n,l}^{(0)} = \begin{cases} \frac{1}{d^l} (2^{2n+1} - 5 \cdot 2^{n-1+l} + 2^{n+1} + 2^{2l-2l}), & \text{если } 2l - 1 \leq n, \\ \frac{1}{d^l} (2^{2n+1} - 5 \cdot 2^{n-1+l} + 2^{2l-2}(n - l + 4)), & \text{если } 2l - 1 > n, \end{cases} \tag{3}$$

$$\mathbf{E}V'_{n,l} = \begin{cases} \frac{1}{d^l} ((2^{l-1} - 1)2^{n+1} - 2^{2l-2}(l - 1)), & \text{если } 2l - 1 \leq n, \\ \frac{1}{d^l} (2^{l-1}(2^{n+1} - 2^l) - 2^{2l-2}(n - l + 1)), & \text{если } 2l - 1 > n, \end{cases} \tag{4}$$

$$\mathbf{E}V''_{n,l,k} = \frac{1}{d^{l-k}} (2^{n-l+2} - 1)2^{2l-k-3}, \quad 1 \leq k < l, \tag{5}$$

$$\sum_{k=1}^{l-1} \mathbf{E}V''_{n,l,k} = \frac{2^n - 2^{l-2}}{d} \frac{1 - (\frac{2}{d})^{l-1}}{1 - \frac{2}{d}}. \tag{6}$$

Замечание 1. Теорема 1 показывает, что если $n, l, n - l \rightarrow \infty$ так, что среднее число $\mathbf{E}V_{n,l}^{(0)}$ пар непересекающихся цепочек (C_i, C_j) с $M(C_i) = M(C_j)$ стремится к конечной константе, то обязательно $d \geq 4$ и тогда $\mathbf{E}V'_{n,l} \rightarrow 0$, $\mathbf{E}V''_{n,l} \rightarrow \infty$, где $V''_{n,l} = \sum_{k=1}^{l-1} V''_{n,l,k}$. Последнее соотношение довольно очевидно, так как, например, для любых двух цепочек C_i, C'_i с общей начальной вершиной $i = C_i[0] = C'_i[0]$ и $l - 1$ общими элементами (т. е. $C_i[l - 2] = C'_i[l - 2], C_i[l - 1] \neq C'_i[l - 1]$) вероятность совпадения цепочек символов $M(C_i)$ и $M(C'_i)$ равна $\mathbf{P}\{m(C_i[l - 1]) = m(C'_i[l - 1])\} = \frac{1}{d}$ и не зависит от длины цепочки и высоты дерева, так что $\mathbf{E}V''_{n,l}$ заведомо имеет порядок числа вершин в дереве T_2^n , если $n \rightarrow \infty$. Соотношение $\mathbf{E}V'_{n,l} = o(\mathbf{E}V_{n,l}^{(0)})$ при $n, n - l \rightarrow \infty$ тоже естественно, так как число пересекающихся цепочек, длина которых заметно отличается от высоты дерева, по порядку растет значительно медленнее числа пар непересекающихся цепочек. Поэтому имеет смысл рассматривать распределение числа пар непересекающихся цепочек (C_i, C_j) с $M(C_i) = M(C_j)$ в случае, когда математическое ожидание этого числа стремится к константе.

Если разность $n/2 - l$ между $n/2$ и длиной l цепочек C_i, C_j с $M(C_i) = M(C_j)$ стремится к бесконечности, то для того, чтобы математическое ожидание числа пар таких цепочек было ограничено, необходимо выполнение условия $d \geq 16$.

Замечание 2. В последовательности $\xi_1, \xi_2, \dots$ независимых одинаково распределенных знаков алфавита из d символов повторения цепочек возникают группами (см., например, [8]): если повторяются цепочки длины l , начинающиеся с ξ_m и ξ_n ,

то, так как $\mathbf{P}\{\xi_{m+l} = \xi_{n+l}\} = \mathbf{P}\{\xi_{m-l} = \xi_{n-l}\} \geq \frac{1}{d}$, с вероятностью не меньше $1/d$ повторяются цепочки, начинающиеся с ξ_{m+1} и ξ_{n+1} , и с такой же вероятностью повторяются цепочки, начинающиеся с ξ_{m-1} и ξ_{n-1} . Поэтому (см. [8]) предельным распределением общего числа пар повторяющихся цепочек является сложное распределение Пуассона, а обычное распределение Пуассона является предельным для числа пар «первых» повторений, т.е. для числа пар таких совпадающих цепочек, начинающихся с ξ_m и ξ_n , что $\xi_{m-1} \neq \xi_{n-1}$.

В нашем случае при условии, что совпадают значения $M(C_i)$ и $M(C_j)$, с вероятностью $1/d = \mathbf{P}\{m(i^-) = m(j^-)\}$ совпадают и значения $M(C_{i-})$ и $M(C_{j-})$. Кроме того, с вероятностью $1/d$ совпадают значения $M(C'_i)$ и $M(C'_j)$ для цепочек C'_i и C'_j с $C'_i[0] = C'_j[0]$, $C'_i[l-2] = C'_j[l-2]$, $C'_i[l-1] \neq C'_j[l-1]$. В теореме 2 указаны достаточные условия и оценка скорости сходимости к сложному распределению Пуассона распределения числа пар непересекающихся цепочек C_i, C_j с $M(C_i) = M(C_j)$, $m(i^-) \neq m(j^-)$. Будем называть такие пары цепочек символов несдвигаемыми к корню.

Для пары непересекающихся цепочек $(C_i, C_j) \in \mathcal{P}$ положим

$$X_{C_i C_j} = \mathbb{I}\{M(C_i) = M(C_j), m(i^-) \neq m(j^-)\};$$

в случае, когда $i = *$, будем считать условие $m(i^-) \neq m(j^-)$ выполненным. Из предположения о независимости и равновероятности меток $m(i)$ следует, что при $(C_i, C_j) \in \mathcal{P}$

$$\mathbf{E}X_{C_i C_j} = \mathbf{E}\mathbb{I}\{M(C_i) = M(C_j)\}\mathbb{I}\{m(i^-) \neq m(j^-)\} = \begin{cases} \frac{d-1}{d^{l+1}}, & \text{если } i^- \neq j^-, \\ 0, & \text{если } i^- = j^-. \end{cases}$$

Положим

$$V_{n,l}^{(0)-} = \sum_{(C_i, C_j) \in \mathcal{P}: C_i \cap C_j = \emptyset} X_{C_i C_j}.$$

Чтобы упростить доказательство предельной теоремы, мы исключим из суммы $V_{n,l}^{(0)-}$ слагаемые, которые дают в нее асимптотически пренебрежимый при $n, l \rightarrow \infty$ вклад.

Пусть $\tilde{\mathcal{P}} \subset \mathcal{P}$ — множество всех таких пар (C_i, C_j) , $i \in I^{(v_i)}$, $j \in I^{(v_j)}$, непересекающихся цепочек, что если вершина j лежит в поддереве, порожденном вершиной i , то $v_j \geq v_i + 2l - 1$. Положим

$$\tilde{V}_{n,l} = \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} X_{C_i C_j}.$$

Лемма 1. *Справедливы соотношения*

$$\mathbf{E}V_{n,l}^{(0)-} = \begin{cases} \frac{d-1}{d^{l+1}} (2^{2n+1} - 6 \cdot 2^{n+l-1} + 2^{n+1} + 2^{2l-2}(l+1)), & \text{если } 2l-1 \leq n, \\ \frac{d-1}{d^{l+1}} (2^{2n+1} - 6 \cdot 2^{n+l-1} + 2^{2l-2}(n-l+5)), & \text{если } 2l-1 > n. \end{cases}$$

$$\mathbf{E}V_{n,l}^{(0)-} - \frac{l}{d^l} 2^{n-l+2} < \mathbf{E}\tilde{V}_{n,l} < \mathbf{E}V_{n,l}^{(0)-}.$$

Следствие 1. Если $n, l \rightarrow \infty$ так, что $\mathbf{E}V_{n,l}^{(0)-}$ остается ограниченным, то $\mathbf{P}\{\tilde{V}_{n,l} = V_{n,l}^{(0)-}\} \rightarrow 1$.

Следствие 1 вытекает из того, что $\tilde{V}_{n,l} \leq V_{n,l}^{(0)-}$ — целочисленные случайные величины и разность между их математическими ожиданиями стремится к 0.

Сравнение формул для $\mathbf{E}V_{n,l}^{(0)}$ и $\mathbf{E}V_{n,l}^{(0)-}$ показывает, что при условиях следствия 1 в среднем на одно несдвигаемое к корню повторение цепочек приходится $\frac{1}{d-1}$ дополнительных (сдвигаемых к корню) повторений.

Определение 1. Пусть пара цепочек $(C_i, C_j) \in \mathcal{P}$ такова, что поддеревья высоты $l-1$, порожденные вершинами i и j , не пересекаются. Положим

$$\pi_k = \frac{1}{k} \mathbf{P} \left\{ \sum_{(C'_i, C'_j) \in \mathcal{P}} X_{C'_i C'_j} = k \mid X_{C_i C_j} = 1 \right\}, \quad k = 1, 2, \dots \quad (7)$$

Определение 2. Сложным распределением Пуассона $CP(\pi)$ называется распределение случайной величины

$$\Xi_\pi = \sum_{k=1}^{\infty} k \xi_k, \quad (8)$$

где $\xi_1, \xi_2, \dots$ независимы и при каждом $k \geq 1$ случайная величина ξ_k имеет распределение Пуассона с параметром π_k .

Укажем условия, при которых стремится к 0 расстояние по вариации между распределением $\mathcal{L}(\tilde{V}_{n,l})$ случайной величины $\tilde{V}_{n,l}$ и сложным пуассоновским распределением $CP(\pi)$, определенным формулами (7), (8):

$$d_{\text{tv}}(\mathcal{L}(\tilde{V}_{n,l}), CP(\pi)) = \frac{1}{2} \sum_{k=0}^{\infty} |\mathbf{P}\{\tilde{V}_{n,l} = k\} - \mathbf{P}\{\Xi_\pi = k\}|.$$

Теорема 2. Пусть $n, l \rightarrow \infty$ так, что $2^{2l} = o(2^n)$ и

$$\mathbf{E}\tilde{V}_{n,l} = \frac{d-1}{d} \cdot \frac{2^{2n+1}}{d^l} (1 + o(1)) \rightarrow \lambda \in (0, \infty),$$

тогда существуют такие $\varepsilon(l, n)$, что $\varepsilon(l, n) = o(1)$ и

$$d_{\text{tv}}(\mathcal{L}(\tilde{V}_{n,l}), CP(\pi)) \leq 2H_1(\pi) \left(\mathbf{E}\tilde{V}_{n,l} \right)^2 \frac{2^{2l}}{2^n} (1 + \varepsilon(l, n)) \rightarrow 0,$$

где $H_1(\pi) \leq \min \left(1, \frac{1}{\pi_1} \right) \cdot \exp \left(\sum_{k=1}^{\infty} \pi_k \right)$.

2. Доказательства

Доказательство теоремы 1. Сперва найдем формулы для чисел слагаемых в суммах $V''_{n,l,k}$. Если $C_i[0] = C'_i[0]$, $|C_i \cap C'_i| = k$, то

$$C_i[0] \in \cup_{v=0}^{n-l+1} I^{(v)}, \quad C_i[s] = C'_i[s] \text{ при } 0 \leq s < k, \quad C_i[k] \neq C'_i[k],$$

при фиксированном $C_i[0]$ существует 2^{l-1} цепочек C_i длины l с началом $C_i[0]$ и для каждой такой цепочки C_i существует ровно $2^{l-(k+1)}$ цепочек $C'_i \neq C_i$ с $C'_i[0] = C_i[0]$, $|C_i \cap C'_i| = k$ ($C_i[k-1] = C'_i[k-1]$, $C_i[k] \neq C'_i[k]$), поэтому число слагаемых в $V''_{n,l,k}$ (с учетом условия $C_i \prec C'_i$) равно

$$\begin{aligned} & |\{(C_i, C'_i) : C_i \neq C'_i, |C_i \cap C'_i| = k, C_i[0] = C'_i[0]\}| = \\ & = \frac{1}{2} \left| \bigcup_{h=0}^{n-l+1} I^{(h)} \right| \cdot 2^{l-1} 2^{l-k-1} = (2^{n-l+2} - 1) 2^{2l-k-3}. \end{aligned} \quad (9)$$

Далее, если $C_i[0] = C'_i[0]$, $|C_i \cap C'_i| = k$, то

$$\mathbf{P}\{M(C_i) = M(C'_i)\} = \mathbf{P}\{m(C_i[h]) = m(C'_i[h]), k \leq h < l\} = d^{-(l-k)},$$

отсюда, из (9) и определений (2) следует формула (5).

Аналогичными рассуждениями докажем равенство (4). Условия $C_i \cap C_j \neq \emptyset$, $C_i[0] \neq C_j[0]$, $C_i \prec C_j$ выполняются, если $C_j[0] \in (C_i \setminus \{C_i[0]\})$. Тогда конечные вершины $C_i[l-1]$ и $C_j[l-1]$ лежат в разных слоях: $C_i[l-1] \in I^{(v_i)}$, $C_j[l-1] \in I^{(v_j)}$ и $l-1 \leq v_i < v_j \leq \min\{n, v_i + l - 1\}$. Вершину $C_j[l-1]$, определяющую цепочку C_j , можно выбрать 2^{v_j} способами; тогда $C_j[0] \in I^{(v_j-l+1)}$ и конечная вершина $C_i[l-1] \in I^{(v_i)}$, $\max\{l-1, v_j-l+1\} \leq v_i < v_j$ цепочки C_i должна лежать в поддереве, порожденном вершиной $C_j[0] = C_i[v_j - v_i]$. Число вершин этого поддерева в слое $I^{(v_i)}$ равно $2^{l-1-v_j+v_i}$; оно равно числу конечных вершин цепочек C_i , удовлетворяющих условиям $C_j[0] = C_i[v_j - v_i]$, $C_i[l-1] \in I^{(v_i)}$. Поэтому число пар $(C_i, C_j) \in \mathcal{P}$, для которых $C_i \cap C_j \neq \emptyset$, $C_i[0] \neq C_j[0]$, равно

$$\begin{aligned} & \sum_{v_j=l}^n 2^{v_j} \sum_{v_i=\max\{l-1, v_j-l+1\}}^{v_j-1} 2^{l-v_j+v_i-1} = \\ & = \sum_{v_j=l}^{\min\{n, 2l-2\}} 2^{v_j} \sum_{v_i=l-1}^{v_j-1} 2^{l-v_j+v_i-1} + \mathbb{I}\{2l-1 \leq n\} \sum_{v_j=2l-1}^n 2^{v_j} \sum_{v_i=v_j-l+1}^{v_j-1} 2^{l-v_j+v_i-1}. \end{aligned}$$

Легко проверить, что $\sum_{v_i=m}^{v_j-1} 2^{l-v_j+v_i-1} = 2^{l-1} - 2^{l-v_j+m-1}$. Если $2l-1 \leq n$, то

$$\begin{aligned} & \sum_{v_j=l}^{2l-2} 2^{v_j} \sum_{v_i=l-1}^{v_j-1} 2^{l-v_j+v_i-1} + \mathbb{I}\{2l-1 \leq n\} \sum_{v_j=2l-1}^n 2^{v_j} \sum_{v_i=v_j-l+1}^{v_j-1} 2^{l-v_j+v_i-1} = \\ & = \sum_{v_j=l}^{2l-2} 2^{v_j} (2^{l-1} - 2^{2l-v_j-2}) + \sum_{v_j=2l-1}^n 2^{v_j} (2^{l-1} - 1) = \\ & = 2^{l-1} (2^{n+1} - 2^l) - (2^{n+1} - 2^{2l-1}) - 2^{2l-2} (l-1) = \\ & = (2^{l-1} - 1) 2^{n+1} - 2^{2l-2} (l-1), \end{aligned} \quad (10)$$

если же $2l-1 > n$, то

$$\begin{aligned} & \sum_{v_j=l}^n 2^{v_j} \sum_{v_i=l-1}^{v_j-1} 2^{l-v_j+v_i-1} = \sum_{v_j=l}^n 2^{v_j} (2^{l-1} - 2^{2l-v_j-2}) = \\ & = 2^{l-1} (2^{n+1} - 2^l) - 2^{2l-2} (n-l+1). \end{aligned} \quad (11)$$

Математическое ожидание каждого слагаемого в сумме $V'_{n,l}$ с $C_j[0] = C_i[h]$, $|C_i \cap C_j| = k$ есть

$$\begin{aligned} & \mathbf{P}\{M(C_i) = M(C_j) \mid C_j[0] = C_i[h], |C_i \cap C_j| = k\} = \\ & = \mathbf{P}\left\{ \begin{array}{ll} m(C_j[s]) (= m(C_i[h+s])) = m(C_i[s]) & (0 \leq s < k) \\ m(C_j[w]) = m(C_i[w]) & (k \leq w < l) \end{array} \right\}. \end{aligned}$$

В последовательности событий $\{m(C_j[w]) = m(C_i[w]), w = 0, 1, \dots, l-1\}$ каждое следующее содержит случайную величину $m(C_j[w])$, не входящую в предыдущие события и поэтому не зависящую от них. Следовательно, эти события независимы, вероятность каждого равна $\frac{1}{d}$, поэтому

$$\mathbf{P}\{M(C_i) = M(C_j) \mid C_j[0] = C_i[h], |C_i \cap C_j| = k\} = \frac{1}{d^l}$$

что вместе с (10), (11) и (2) доказывает (4).

Наконец, сумма, соответствующая $V_{n,l}^{(0)}$, состоит из тех и только тех слагаемых, которые соответствуют парам $(C_i, C_j) \in \mathcal{P}$, не вошедшим в суммы $V'_{n,l}$ и $V''_{n,l,k}$, т.е. парам непересекающихся цепочек. Общее число слагаемых в суммах $V'_{n,l}$ и $V''_{n,l,k}$ равно

$$\begin{aligned} & (2^{l-1} - 1)2^{n+1} - 2^{2l-2}(l-1) + \sum_{k=1}^{l-1} (2^{n-l+2} - 1)2^{2l-k-3} = \\ & = (2^{l-1} - 1)2^{n+1} - 2^{2l-2}(l-1) + (2^{n-l+2} - 1)(2^{2l-3} - 2^{l-2}) = \\ & = 3(2^{n-1+l} - 2^n) - (2l-1)2^{2l-3} \end{aligned}$$

при $2l-1 \leq n$, и равно

$$\begin{aligned} & 2^{l-1}(2^{n+1} - 2^l) - 2^{2l-2}(n-l+1) + \sum_{k=1}^{l-1} (2^{n-l+2} - 1)2^{2l-k-3} = \\ & = 2^{l-1}(2^{n+1} - 2^l) - 2^{2l-2}(n-l+1) + (2^{n-l+2} - 1)(2^{2l-3} - 2^{l-2}) = \\ & = 3 \cdot 2^{n-1+l} - 7 \cdot 2^{2l-3} - 2^{2l-2}(n-l) - 2^n + 2^{l-2} \end{aligned}$$

при $2l-1 > n$. Вычитая эти числа из $|\mathcal{P}| = \frac{1}{2}(2^{n+1} - 2^{l-1})(2^{n+1} - 2^{l-1} - 1)$, находим, что сумма $V_{n,l}^{(0)}$ содержит

$$2^{2n+1} - 5 \cdot 2^{n-1+l} + 2^{n+1} + 2^{2l-2}l \quad (12)$$

слагаемых при $2l-1 \leq n$ и

$$2^{2n+1} - 5 \cdot 2^{n-1+l} + 2^{2l-2}(n-l+4) \quad (13)$$

слагаемых при $2l-1 > n$. Математическое ожидание каждого слагаемого равно $\frac{1}{d^l}$, откуда следует (3). Теорема 1 доказана.

Доказательство леммы 1. Хотя области суммирования в суммах $V_{n,l}^{(0)-}$ и $V_{n,l}^{(0)}$ совпадают, формулы для их математических ожиданий отличаются не только множителем $\frac{d-1}{d}$, так как ряд слагаемых в $V_{n,l}^{(0)-}$ равен 0. Количество нулевых слагаемых

равно числу таких непересекающихся пар $(C_i, C_j) \in \mathcal{P}$, что $i^- = j^-$. Общей вершиной $i^- = j^-$ может быть любая вершина дерева T_2^n из слоев $I^{(v)}$, $0 \leq v \leq n-l$. Число таких вершин равно $2^{n-l+1} - 1$. Для каждой такой вершины существует $(2^{l-1})^2$ пар непересекающихся цепочек (C_i, C_j) с $i^- = j^-$. Значит, число ненулевых слагаемых в $V_{n,l}^{(0)-}$ меньше числа слагаемых в $V_{n,l}^{(0)}$ на $(2^{n-l+1} - 1)(2^{l-1})^2 = 2^{n+l-1} - 2^{2l-2}$. Вычитая это число из чисел слагаемых в (12) и (13) и умножая эти разности на $\frac{d-1}{d}$, получаем формулы для $\mathbf{E}V_{n,l}^{(0)-}$.

Чтобы оценить число пар непересекающихся цепочек $(C_i, C_j) \in \mathcal{P} \setminus \tilde{\mathcal{P}}$, заметим, что если $i \in I^{(v_i)}$, то число таких цепочек C_j , что вершина $j \in T_2^n$ содержится в слое $I^{(v_j)}$ в поддереве, порожденном вершиной i , и при этом $v_i < v_j \leq \min\{n-l+1, v_i + 2l-2\}$, меньше 2^{2l-2} при $v_i \leq n-3l+3$ и меньше 2^{n-l-v_i+1} при $n-3l+4 \leq v_i \leq n-l$. Значит,

$$|\mathcal{P} \setminus \tilde{\mathcal{P}}| < 2^{2l-2} 2^{n-3l+4} \mathbb{I}\{n-3l+4 \geq 0\} + \sum_{v_i=n-3l+4}^{n-l} 2^{v_i} 2^{n-l-v_i+1} \leq \\ \leq \min\{n-l, 2l-1\} 2^{n-l+1}.$$

Отсюда и из формулы для $\mathbf{E}X_{C_i C_j}$ следует второе соотношение в формулировке леммы 1.

Доказательство теоремы 2. Используем метод Чена–Стейна (см., например, [1]). Для каждой пары $(C_i, C_j) \in \tilde{\mathcal{P}}$, $C_i \cap C_j = \emptyset$, определим подмножества в $\tilde{\mathcal{P}} \setminus \{(C_i, C_j)\}$:

$$\tilde{\mathcal{P}}_{C_i C_j}^{vs} = \left\{ (C'_i, C'_j) \in \tilde{\mathcal{P}} \setminus \{(C_i, C_j)\} \mid C'_i[0] = C_i[0], C'_j[0] = C_j[0] \right\},$$

$$\tilde{\mathcal{P}}_{C_i C_j}^w = \left\{ (C_u, C_v) \in \tilde{\mathcal{P}} \mid (C_u \cup C_v) \cap (C'_i \cup C'_j) = \emptyset \quad \forall (C'_i, C'_j) \in \tilde{\mathcal{P}}_{C_i C_j}^{vs} \cup \{(C_i, C_j)\} \right\},$$

$$\tilde{\mathcal{P}}_{C_i C_j}^b = \tilde{\mathcal{P}} \setminus \left(\{(C_i, C_j)\} \cup \tilde{\mathcal{P}}_{C_i C_j}^{vs} \cup \tilde{\mathcal{P}}_{C_i C_j}^w \right)$$

и соответствующие суммы индикаторов

$$Z_{C_i C_j} = \sum_{(C'_i, C'_j) \in \tilde{\mathcal{P}}_{C_i C_j}^{vs}} X_{C'_i C'_j}, \quad W_{C_i C_j} = \sum_{(C'_i, C'_j) \in \tilde{\mathcal{P}}_{C_i C_j}^w} X_{C'_i C'_j}, \\ U_{C_i C_j} = \sum_{(C'_i, C'_j) \in \tilde{\mathcal{P}}_{C_i C_j}^b} X_{C'_i C'_j}.$$

Для всех пар цепочек $(C_i, C_j) \in \tilde{\mathcal{P}}$ совместные распределения $(X_{C_i C_j}, Z_{C_i C_j})$ одинаковы. В этом случае согласно (см. [1], теоремы 3.5 и 3.6) для расстояния полной

вариации $d_{tv}(\mathcal{L}(\tilde{V}_{n,l}), CP(\pi))$ имеет место оценка

$$\begin{aligned} d_{tv}(\mathcal{L}(\tilde{V}_{n,l}), CP(\pi)) &\leq \\ &\leq H_1(\pi) \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} (\mathbf{E}X_{C_i C_j} \mathbf{E}(X_{C_i C_j} + Z_{C_i C_j} + U_{C_i C_j}) + \mathbf{E}(X_{C_i C_j} U_{C_i C_j})) + \\ &+ H_0(\pi) \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} \sum_{k=1}^{\infty} \mathbf{E}|\mathbf{P}\{X_{C_i C_j} = 1, X_{C_i C_j} + Z_{C_i C_j} = k\} - \\ &- \mathbf{P}\{X_{C_i C_j} = 1, X_{C_i C_j} + Z_{C_i C_j} = k | W_{C_i C_j}\}|, \end{aligned}$$

где

$$\max\{H_0(\pi), H_1(\pi)\} \leq \min\left\{1, \frac{1}{\pi_1}\right\} \cdot \exp\left\{\sum_{k=1}^{\infty} \pi_k\right\}.$$

Так как при любой паре $(C_i, C_j) \in \tilde{\mathcal{P}}$ случайная величина $W_{C_i C_j}$ зависит только от меток вершин, не входящих в множество меток вершин, определяющих случайные величины $X_{C_i C_j}$ и $Z_{C_i C_j}$, то случайные величины $(X_{C_i C_j}, Z_{C_i C_j})$ и $W_{C_i C_j}$ независимы, поэтому второе слагаемое в оценке расстояния по вариации равно нулю и

$$\begin{aligned} d_{tv}(\mathcal{L}(\tilde{V}_{n,l}), CP(\pi)) &\leq \\ &\leq H_1(\pi) \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} \{\mathbf{E}X_{C_i C_j} \mathbf{E}\{X_{C_i C_j} + Z_{C_i C_j} + U_{C_i C_j}\} + \mathbf{E}\{X_{C_i C_j} U_{C_i C_j}\}\}. \end{aligned}$$

Доказательства следующих двух лемм будут приведены позднее.

Лемма 2. Для любой пары $(C_i, C_j) \in \tilde{\mathcal{P}}$

$$|\tilde{\mathcal{P}}_{C_i C_j}^{vs}| < 2^{2l-2}, \quad |\tilde{\mathcal{P}}_{C_i C_j}^b| < 2^{n+l+2}(2^l + l).$$

Лемма 3. Для любой пары $(C_u, C_v) \in \tilde{\mathcal{P}} \setminus \left(\{(C_i, C_j)\} \cup \tilde{\mathcal{P}}_{C_i C_j}^{vs}\right)$ вероятность $\mathbf{P}\{X_{C_i C_j} = X_{C_u C_v} = 1\}$ принимает значения либо 0, либо $\left(\frac{d-1}{d^{l+1}}\right)^2$.

Пользуясь этими двумя леммами, соотношениями $\mathbf{E}\tilde{V}_{n,l} = \frac{d-1}{d^{l+1}} |\tilde{\mathcal{P}}|$, $|\tilde{\mathcal{P}}| = 2^{2n+1}(1 + o(1))$ и $2^{2l} = o(2^n)$, мы получаем, что

$$\begin{aligned} d_{tv}(\mathcal{L}(\tilde{V}_{n,l}), CP(\pi)) &\leq \\ &\leq H_1(\pi) \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} (\mathbf{E}X_{C_i C_j} \mathbf{E}(X_{C_i C_j} + Z_{C_i C_j} + U_{C_i C_j}) + \mathbf{E}(X_{C_i C_j} U_{C_i C_j})) \leq \\ &\leq H_1(\pi) \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} \left(\left(\frac{d-1}{d^{l+1}}\right)^2 (1 + |\tilde{\mathcal{P}}_{C_i C_j}^{vs}| + |\tilde{\mathcal{P}}_{C_i C_j}^b|) + \left(\frac{d-1}{d^{l+1}}\right)^2 |\tilde{\mathcal{P}}_{C_i C_j}^b| \right) = \\ &= H_1(\pi) \left(\frac{d-1}{d^{l+1}}\right)^2 \sum_{(C_i, C_j) \in \tilde{\mathcal{P}}} (1 + |\tilde{\mathcal{P}}_{C_i C_j}^{vs}| + 2|\tilde{\mathcal{P}}_{C_i C_j}^b|) < \\ &< H_1(\pi) \left(\frac{d-1}{d^{l+1}}\right)^2 |\tilde{\mathcal{P}}| (1 + 2^{2l-2} + 2 \cdot 2^{n+l+2}(2^l + 1)) = \\ &= H_1(\pi) \left(\left(\frac{d-1}{d^{l+1}}\right)^2 |\tilde{\mathcal{P}}| \right)^2 \frac{1+2^{2l-2}+2^{n+l+3}+2^{n+2l+2}}{|\tilde{\mathcal{P}}|} = \\ &= 2H_1(\pi) \left(\mathbf{E}\tilde{V}_{n,l}\right)^2 \frac{2^{2l}}{2^n} (1 + o(1)) \rightarrow 0. \end{aligned}$$

Доказательство леммы 2. Оценим мощность множества $\tilde{\mathcal{P}}_{C_i C_j}^{vs}$. Так как оно состоит из всех пар цепочек с фиксированными начальными вершинами i и j , кроме пары (C_i, C_j) , то $|\tilde{\mathcal{P}}_{C_i C_j}^{vs}| < 2^{2l-2}$.

Оценим мощность множества $\tilde{\mathcal{P}}_{C_i C_j}^b$; оно состоит из таких пар цепочек (C_u, C_v) , что хотя бы одна из цепочек C_u, C_v пересекается с поддеревом высоты $l-1$, порожденным вершиной i или вершиной j . Пусть для определенности это цепочка C_u . В таком случае u должна либо принадлежать одному из двух поддеревьев высоты $l-1$ (число таких цепочек не больше $2 \cdot 2^l \cdot 2^{l-1} = 2^{2l}$), либо удовлетворять одному из условий $C_u[h] = i, C_u[h] = j, h \in \{0, \dots, l-1\}$, а это эквивалентно тому, что конечная вершина $C_u[l-1]$ принадлежит одному из двух поддеревьев высоты $l-1$; число таких цепочек C_u меньше $2 \cdot 2^l$. Значит, число рассматриваемых цепочек C_u не больше $2^l(2^l + 1)$, а число возможных значений второй цепочки меньше 2^{n+1} ; отсюда следует оценка $|\tilde{\mathcal{P}}_{C_i C_j}^b| < 2 \cdot 2^{n+1+1}(2^l + 1)$.

Доказательство леммы 3. Чтобы вычислить $\mathbf{P}\{X_{C_i C_j} = X_{C_u C_v} = 1\}$ для пар непересекающихся цепочек $(C_i, C_j), (C_u, C_v) \in \tilde{P}$, удовлетворяющих условию $\{i, j\} \neq \{u, v\}$, заметим, что подграф дерева T_2^n , образованный этими четырьмя цепочками, может иметь 2, 3 или 4 связных компоненты; этот граф не может быть связным в силу определения множества $\tilde{P}$.

Если число компонент равно 4, то цепочки попарно не пересекаются и

$$\begin{aligned} & \mathbf{P}\{X_{C_i C_j} = X_{C_u C_v} = 1\} = \\ & = \mathbf{P}\{M(C_i) = M(C_j), m(i^-) \neq m(j^-), M(C_u) = M(C_v), m(u^-) \neq m(v^-)\} = \\ & = \mathbf{P}\{M(C_i) = M(C_j), M(C_u) = M(C_v)\} \times \\ & \times \mathbf{P}\{m(i^-) \neq m(j^-), m(u^-) \neq m(v^-) \mid M(C_i) = M(C_j), M(C_u) = M(C_v)\}. \end{aligned}$$

Первый сомножитель равен $\frac{1}{d^{2l}}$, а второй равен либо $\left(\frac{d-1}{d}\right)^2$, либо 0 (если, например, $i^- = C_u[h], j^- = C_v[h]$).

Пусть число связных компонент равно 3. Тогда какие-то две цепочки из разных пар $(C_i, C_j), (C_u, C_v)$ пересекаются, а остальные две не пересекаются между собой и с пересекающимися. Пусть для определенности $C_i \cap C_u \neq \emptyset$, тогда

$$\begin{aligned} & \mathbf{P}\{X_{C_i C_j} = X_{C_u C_v} = 1\} = \\ & = \mathbf{P}\{M(C_i) = M(C_j), m(i^-) \neq m(j^-), M(C_u) = M(C_v), m(u^-) \neq m(v^-)\} = \\ & = \mathbf{P}\{M(C_i) = M(C_j)\} \mathbf{P}\{M(C_u) = M(C_v) \mid M(C_i) = M(C_j)\} \times \\ & \times \mathbf{P}\{m(i^-) \neq m(j^-), m(u^-) \neq m(v^-) \mid M(C_i) = M(C_j), M(C_u) = M(C_v)\}. \end{aligned}$$

Аналогично предыдущему случаю первые два сомножителя равны $\frac{1}{d^l}$, так как цепочка C_v не пересекается с C_i, C_j, C_u , а последний равен либо $\left(\frac{d-1}{d}\right)^2$, либо 0 (если, например, $u^- = C_i[h], v^- = C_j[h]$).

Пусть число связных компонент, образованных цепочками $(C_i, C_j), (C_u, C_v)$, равно 2. Без ограничения общности можно считать, что $C_i[0] = C_u[h]$ и либо $C_j[0] =$

$C_v[k]$, либо $C_j[k] = C_v[0]$, $0 \leq h < k < l$. Рассмотрим первый случай, предполагая дополнительно, что $|C_i \cap C_u| = x, |C_j \cap C_v| = y, 0 < x, y < l$. Тогда

$$\begin{aligned} & \mathbf{P}\{X_{C_i C_j} = X_{C_u C_v} = 1\} = \\ & = \mathbf{P} \left\{ \begin{array}{l} m(i^-) \neq m(j^-), \quad m(C_i[a]) = m(C_j[a]), 0 \leq a < l \\ m(u^-) \neq m(v^-), \quad m(C_u[a]) = m(C_v[a]), 0 \leq a < l \end{array} \middle| \begin{array}{l} C_i[0] = C_u[h] \\ |C_i \cap C_u| = x \\ C_j[0] = C_v[k] \\ |C_j \cap C_v| = y \end{array} \right\}. \end{aligned}$$

В последовательности событий

$$\begin{aligned} & m(i^-) \neq m(j^-), m(C_i[0]) = m(C_j[0]), \dots, m(C_i[l-1]) = m(C_j[l-1]), \\ & m(u^-) \neq m(v^-), m(C_u[0]) = m(C_v[0]), \dots, m(C_u[l-1]) = m(C_v[l-1]) \end{aligned}$$

каждое следующее не зависит от предыдущих в силу равновероятности распределения меток, условия $h \neq k$ и того, что в каждое событие входит хотя бы одна метка вершины, отсутствующей в предшествующих событиях. Поэтому

$$\mathbf{P}\{X_{C_i C_j} = X_{C_u C_v} = 1\} = \left(\frac{d-1}{d^{l+1}}\right)^2.$$

Второй случай рассматривается аналогично.

Список литературы

1. Erhardsson T., “Stein’s method for Poisson and compound Poisson approximation”, In “*An introduction to Stein’s method*”, eds. Barbour A. D., Chen L. H. Y., Singapore Univ. Press, 2005, 61–113.
2. Guibas L. J., Odlyzko A. M., “Long repetitive patterns in random sequences”, *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **53** (1980), 241–262.
3. Hoffmann C. M., O’Donnell M. J., “Pattern matching in trees”, *J. ACM*, **29**:1 (1982), 68–95.
4. Karlin S., Ost F., “Counts of long aligned word matches among random letter sequences”, *Adv. Appl. Probab.*, **19**:2 (1987), 293–351.
5. Karnin E. D., “The first repetition of a pattern in a symmetric Bernoulli sequence”, *J. Appl. Prob.*, **20**:3 (1983), 413–418.
6. Rowland E. S., “Pattern avoidance in binary trees”, arXiv: 0809.0488.
7. Steyaert J.-M., Flajolet P., “Patterns and pattern-matching in trees: an analysis”, *Inf. & Control*, **58**:1 (1983), 19–58.
8. Зубков А. М., Михайлов В. Г., “Предельные распределения случайных величин, связанных с длинными повторениями в последовательности независимых испытаний”, *Теория вероятн. и ее примен.*, **XIX**:1 (1974), 173–181.
9. Михайлов В. Г., “Оценка точности сложной пуассоновской аппроксимации для распределения числа совпадающих цепочек”, *Теория вероятн. и ее примен.*, **46**:4 (2001), 713–723.

Статья поступила 30.10.2015.