MODELS OF DISTRIBUTED PROOF GENERATION FOR ZK-SNARK-BASED BLOCKCHAINS

Yuri Bespalov[1,a], Alberto Garoffolo[2,b], Lyudmila Kovalchuk[3,c], Hanna Nelasa[4,d], Roman Oliynykov[3,e]
[1]*Bogolyubov Institute for Theoretical Physics*
Kiev, Ukraine
[2]*Horizen*
Milan, Italy
[3]*IOHK Research*
Hong Kong
[4]*Zaporizhzhia Polytechnic National University*
Zaporizhzhia, Ukraine
e-mail:
{[a]yu.n.bespalov,[c]lusi.kovalchuk,[d]annanelasa,[e]roliynykov}@gmail.com,
[b]alberto@horizen.global

Abstract

We model distributed proof generation for ZK-SNARKs-based blockchains via discrete Markov chains. Two different types of proof construction models are considered: those in which all the proofs to be built are independent (they can be considered as leaves on the Merkle tree) and those in which the proofs are located at all nodes of the Merkle tree, and hence form a partially ordered set.
Keywords: blockchain, Merkle tree, lumpable Markov chain, Stirling numbers, coupon collector's problem, classical occupancy distribution, Birkhoff duality

1 Introduction

The paper considers the problem of estimating the number of steps to build a complete set of SNARK proofs in the Merkle tree for blockchains. In doing so, we consider two different types of proof construction models: those in which all the proofs to be built are independent (they can be considered as leaves on the Merkle tree) and those in which the proofs are located at all nodes of the Merkle tree, and hence form a partially ordered set. The first one obviously is much more simpler, and we partially solved it.

The article is organized in the next way. The Chapter 3 illustrate the lumping of states technique for Markov chains on the sample of coupon collector's problem. This technique and this sample are used the further sections. Section 3 considers the problem of the number of steps to construct a complete set of proofs that are leaves of the Merkle tree. We proof that the model from Example 5 initially formulated as non-Markovian is stochastically equivalent to the Markov chain from Example 4, and study its lumped form from Example 6. The recurrent formulas for the expectation and variance of the number of steps are received. We show that dependence of expectation on two parameters the number of provers n and the number of leaves m can asymptotically

reduces to a function h of single parameter n/m and describe this function. Section 4 covers the construction of the entire Merkle tree. Moreover, it is convenient to generalize the models from Sections 2 and 3 to the case of a partially ordered set. This generalization leads to some useful ideas, such as a more appropriate probability distribution on poset items. We are interested in the case of a complete Merkle tree with $2^\ell - 1$ nodes. It is hardly possible to expect a complete analytical solution. The corresponding numerical results and their analysis are supposed to be considered in the expanded version of these theses.

This work was supported in part by the National Research Foundation of Ukraine under Grant 2020.01/0351.

2 Preliminary models

The *Stirling numbers of the second kind* can be defined in the context of the Stanley's twelvefold way [Sta11, 1.9]: $S(m,n) = \left\{ {m \atop n} \right\}$ is the number of partitions of the m labeled elements into n non-empty nonlabelled blocks. Denote $\mathbf{m} := \{1, 2, \dots, m\}$. Then the number of surjections $\mathbf{m} \twoheadrightarrow \mathbf{n}$ is $n! \left\{ {m \atop n} \right\}$. It can be calculated as a sum of multinomial coefficients $\binom{m}{m_1,\dots,m_n} := \frac{m!}{m_1!\cdots m_n!}$, using the forward difference operator Δ or the inclusion-exclusion principle:

$$n! \left\{ {m \atop n} \right\} = \sum_{\substack{m_1+\cdots+m_n=m \\ m_i \geqslant 1}} \binom{m}{m_1, \dots, m_n} = \Delta^n 0^m = \sum_{r=0}^{n} (-1)^r \binom{n}{r} (n-r)^m.$$

Here we assume that Markov chains are discrete-time, stationary and with finite or countable state-space S. We write elements of transition matrix in the form

$$p_{ij} = p(i,j) = \mathbf{P}(X(n+1) = j \mid X(n) = i), \qquad i, j \in S.$$

This a stochastic matrix with $\sum_{j \in S} p_{ij} = 1$.

Definition 1 ([KS76, §6.3])**.** Let $p = (p_{ss'})_{s,s' \in S}$ be a stochastic matrix over a state-space S. A surjection $\pi : S \twoheadrightarrow T$ is called a *lumping map* (and the corresponding partition $S = \coprod_{t \in T} \pi^{-1}(t)$ *lumpable*) if for any $t' \in T$ the sum $\sum_{s' \in \pi^{-1}(t')} p_{ss'}$ is locally constant on $s \in \pi^{-1}(t)$ for each $t \in T$.

Proposition 1. *Let $(p_{ss'})_{s,s' \in S}$ be a stochastic matrix and $\pi : S \twoheadrightarrow T$ a lumping map.*

1. *Then one can define a new stochastic matrix over a state-space T with entries $p^\pi_{tt'} := \sum_{s' \in \pi^{-1}(t')} p_{ss'}$, $s \in \pi^{-1}(t)$.*

2. *Let $v = (\delta_{\pi(s),t})_{\substack{s \in S, \\ t \in T}}$ be the incidence matrix corresponding to the lumping map π, and $u = (v^t v)^{-1} v^t$ the transpose matrix, normalized to stochastic, then the lumped k-fold transition matrix is*

$$(upv)^k = up^k v. \tag{1}$$

We describe a so called coupon collector model as a result of lumping constructions. It is closely related to the our further models, in particular leads to the classical occupancy distribution described via Stirling numbers of the second kind.

Example 1. Consider the asymmetric random walk on the n-dimensional hyperoctant $\mathbb{Z}^n_{\geqslant 0}$ with nonzero transition probabilities $p(a, a+e_i) = 1/n$ for each $a \in \mathbb{Z}^n_{\geqslant 0}$ and basic vectors $e_i = (\underbrace{0,\dots,0}_{i-1}, 1, \underbrace{0,\dots,0}_{n-i})$. Then nonzero entries of m-fold transition matrix are $p^m(a, a+h) = n^{-m}\binom{m}{h_1,\dots,h_n}$, where $h_1 \geqslant 0$ and $h_1 + \dots + h_n = m$.

Example 2. The type map conversion map $\overline{(\text{-})} : \mathbb{Z}_{\geqslant 0} \to \{0,1\}$, $\overline{a} = \begin{cases} 0, & \text{if } a = 0, \\ 1, & \text{if } a > 0, \end{cases}$ applied to each coordinate gets a lumping map $\mathbb{Z}^n_{\geqslant 0} \to \{0,1\}^n$ for the previous Markov chain. According to (1) for the obtained Markov chain on the hypercube $\{0,1\}^n$ m-step transition matrix p^m is the following: if $p^m(a,b)$ then $a_i \leqslant b_i$ for all i; and

$$p^m(a,b) = n^{-m} \sum_{\substack{m_1+\dots+m_n=m \\ m_i \geqslant 1}} \binom{m}{m_1,\dots,m_n} = \frac{r!}{n^m} \left\{ {m \atop r} \right\}, \text{ where } r = \sum_i (b_i - a_i), \text{ if } a \neq b.$$

$$p^m(a,a) = \Big(\sum_i a_i/n \Big)^m.$$

Example 3 (Coupon collector's problem)**.** The projection of hypercube to the main diagonal

$$\{0,1\}^n \to \{0,1,\dots,n\}, \qquad (a_i)_{1 \leqslant i \leqslant n} \mapsto \sum_i a_i$$

is a lumping map. Combining the states we get so called coupon collecting Markov chain [LPW17, 2.2], where nonzero m-step transition probabilities are the following:

$$p^m(k,k) = \frac{k^m}{n^m}, \qquad p^m(k,k+r) = \frac{1}{n^m} \binom{n-k}{r} r! \left\{ {m \atop r} \right\} = \frac{(n-k)_r}{n^m} \left\{ {m \atop r} \right\}, \tag{2}$$

where $(n)_r = n(n-1)\cdots(n-r+1)$ is *the falling* factorial.

There are n distinct coupons in the urn. A collector draw with replacement one random coupon in a step. The number $\xi_m = \xi_0 p^m$ of distinct coupons selected after m steps has the classical occupancy distribution [O'N19]: $\mathbf{P}(\xi_m = r) = p^m(0,r)$.

The expectation of number ζ^n_r of steps to obtain exactly r distinct coupons is described via harmonic numbers $H_n = 1 + 1/2 + \dots + 1/n$:

$$\mathbf{E}\, \zeta^n_r = n(H_n - H_{n-r}). \tag{3}$$

3 Distributed generation of sets of proofs

Example 4. Suppose that there exist $m > 0$ nodes in a network called *provers* and a finite set N of proof-candidates for which they need to construct proofs. We model

this situation as a Markov chain, where states are subsets of $N' \subseteq N$ of candidates for which proofs are not yet constructed. On each step in the state N' each prover independently selects a single candidate from N' and construct its proof, i.e. selection is given by a function $g : \mathbf{m} \to N'$ uniformly distributed among all functions $\mathbf{m} \to N'$. For given selections the next state is obtained by removing all candidates proved in this step. So nonzero transition probabilities described via number of surjections:

$$p(N', N'') = |N' \setminus N''|! \cdot \left\{ {m \atop |N' \setminus N''|} \right\} \cdot |N'|^{-m}, \qquad N'' \subseteq N', \quad |N' \setminus N''| \leqslant m. \tag{4}$$

Example 5. To force provers to act independently, rules are modified in the following way: Denote $\operatorname{ord} N$ the set of linear orderings of N i.e. bijections $\sigma : \{1, 2, \ldots, |N|\} \xrightarrow{\cong} N$. (Note that $|\operatorname{ord} N| = |N|!$.) Suppose that at the beginning each prover randomly selects its own priority ordering $\sigma_i \in \operatorname{ord} N$, $1 \leqslant i \leqslant m$ (We assume a uniform distribution on $\operatorname{ord} N$). After that the process becomes completely deterministic: In the first step all provers select candidates according to the function $g : \mathbf{m} \to N$ given by $g(i) := \sigma_i(1)$. The next state in $N' = N \setminus \operatorname{Im}(g)$. There is a natural projection $\rho^N_{N'} : \operatorname{ord}(N) \to \operatorname{ord}(N')$, which removes foreign elements from an ordering. And provers can do the next step with priority orderings $\rho^N_{N'}(\sigma_i)$.

Proposition 2. *The model from Example 5 is stochastically equivalent to the Markov chain from Example 4.*

Proof. (Sketch.) Uniform distributions of σ_i imply 1) uniform distribution of the first selection function g, and 2) uniform distributions of $\rho^N_{N'}(\sigma_i)$, because the fiber of $\rho^N_{N'}$ over each point has the same cardinality $|\operatorname{ord} N|/|\operatorname{ord} N'| = |N|!/|N'|!$. □

Example 6. Note that the Markov chain from Example 4 admits a lumping map $N' \mapsto |N|$. For each $m, n > 0$ and we obtain a Markov chain with states $\{0, 1, \ldots, n\}$. Exactly from definition one can see that for fixed m and for $n' \leqslant n$, one Markov chain is included in other. So one can consider the colimit (union) of these chains for all n. So for each $m \in \mathbb{Z}_{>0}$ we obtain a Markov chain, where states are nonnegative integers and the only nonzero elements of transition matrix are the following

$$\begin{aligned} p(0,0) &= 1, \\ p(n, n-r) &= \frac{1}{n^m} \binom{n}{r} r! \left\{ {m \atop r} \right\} = \frac{(n)_r}{n^m} \left\{ {m \atop r} \right\}, \quad n > 0, \ \ 1 \leqslant r \leqslant m. \end{aligned} \tag{5}$$

The formula coincides with the classical occupancy distribution from Example 3.

So if we start from the state $\xi^{mn}_0 \equiv n$, then the evolution on the kth step is defined via kth power of transition matrix:

$$\xi^{mn}_k = \xi^{mn}_0 p^k.$$

The absorbing state is 0. All trajectories are strictly decreasing and $\xi^{mn}_k \equiv 0$ for $k \geqslant n$.

The subject of our interest is the *absorption time* τ^{mn}, a random variable which measure the exact number of steps m provers needs to obtain all n proofs. I.e. $\tau^{mn} = k + 1$ iff $\xi^{mn}_{k+1} = 0$ and $\xi^{mn}_k \neq 0$.

Taking into account the lower triangular form of our transition matrix we get recurrent and explicit formulas

$$\begin{aligned}
&\mathbf{P}(\tau^{mn}=0)=\delta_{n0},\\
&\mathbf{P}(\tau^{mn}=k+1)=\sum_{r=1}^{\min(m,n)} p_{n\,n-r}\,\mathbf{P}(\tau^{m\,n-r}=k)=\sum_{r=0}^{\min(m,n)}\frac{(n)_r}{n^m}\left\{\begin{matrix}m\\r\end{matrix}\right\}\mathbf{P}(\tau^{m\,n-r}=k) \qquad (6)\\
&\mathbf{P}(\tau^{mn}=k+1)=\sum_{n_k<\cdots<n_1<n_0=n} p_{n_0n_1}\cdots p_{n_{k-1}n_k}p_{n_k0}\\
&\qquad=\sum_{n_k<\cdots<n_1<n_0=n}\frac{n!}{(n_0n_1\cdots n_k)^m}\left\{\begin{matrix}m\\n_0-n_1\end{matrix}\right\}\cdots\left\{\begin{matrix}m\\n_{k-1}-n_k\end{matrix}\right\}\left\{\begin{matrix}m\\n_k\end{matrix}\right\}. \qquad (7)
\end{aligned}$$

Multiplying (6) by k^ℓ and taking a sum over k we get the recurrent formula for ℓth moment:

$$\mathbf{E}(\tau^{mn}-1)^\ell=\sum_{r=1}^{\min(n,m)} p_{m\,n-r}\,\mathbf{E}(\tau^{m\,n-r})^\ell.$$

In particular, this allows to calculate expectation and variance:

Proposition 3. *Let $m>0$. Then $\tau^{m0}\equiv 0$ and for $n>0$*

$$\begin{aligned}
\mathbf{E}\,\tau^{mn}&=1+\sum_{r=1}^{\min(n,m)}\frac{(n)_r}{n^m}\left\{\begin{matrix}m\\r\end{matrix}\right\}\mathbf{E}\,\tau^{m\,n-r},\\
\mathbf{E}(\tau^{mn})^2&=-1+2\,\mathbf{E}\,\tau^{mn}+\sum_{r=1}^{\min(n,m)}\frac{(n)_r}{n^m}\left\{\begin{matrix}m\\r\end{matrix}\right\}\mathbf{E}(\tau^{m\,n-r})^2,\\
\mathbf{V}\mathrm{ar}\,\tau^{mn}&=\mathbf{E}(\tau^{mn})^2-(\mathbf{E}\,\tau^{mn})^2.
\end{aligned}$$

Here we compare the values of $\mathbf{E}\,\tau^{mn}$ as results of analytic calculation using Wolfram Mathematica (3 last digits in numerator) and of 10^5 random tests of model from Example 5 (3 last digits in denominator):

$n\backslash m$	10	20	30	40	50	100	200	300
10	$2.16\frac{869}{787}$	$1.78\frac{542}{357}$	$1.37\frac{086}{051}$	$1.14\frac{190}{100}$	$1.05\frac{090}{099}$	$1.00\frac{027}{027}$	$1.00\frac{000}{000}$	$1.00\frac{000}{000}$
20	$3.47\frac{931}{927}$	$2.34\frac{507}{533}$	$2.00\frac{865}{842}$	$1.96\frac{422}{396}$	$1.83\frac{582}{516}$	$1.11\frac{346}{316}$	$1.00\frac{070}{055}$	$1.00\frac{000}{001}$
30	$4.67\frac{850}{932}$	$3.04\frac{330}{296}$	$2.48\frac{512}{367}$	$2.05\frac{429}{539}$	$2.00\frac{238}{236}$	$1.66\frac{514}{691}$	$1.03\frac{364}{183}$	$1.00\frac{115}{112}$
40	$5.80\frac{575}{489}$	$3.76\frac{690}{573}$	$2.99\frac{443}{496}$	$2.59\frac{552}{423}$	$2.13\frac{500}{559}$	$1.97\frac{687}{744}$	$1.22\frac{719}{433}$	$1.01\frac{995}{928}$
50	$6.89\frac{606}{594}$	$4.20\frac{784}{602}$	$3.27\frac{580}{637}$	$2.98\frac{450}{461}$	$2.68\frac{236}{375}$	$1.99\frac{990}{982}$	$1.60\frac{171}{019}$	$1.11\frac{091}{170}$
100	$12.16\frac{720}{615}$	$7.06\frac{755}{721}$	$5.24\frac{624}{792}$	$4.36\frac{879}{869}$	$3.98\frac{029}{051}$	$2.90\frac{527}{516}$	$2.00\frac{005}{003}$	$1.99\frac{585}{578}$
200	$22.47\frac{230}{252}$	$12.37\frac{230}{229}$	$9.00\frac{099}{246}$	$7.13\frac{489}{424}$	$6.06\frac{816}{896}$	$4.00\frac{045}{029}$	$2.99\frac{159}{165}$	$2.05\frac{752}{897}$
300	$32.65\frac{910}{976}$	$17.60\frac{450}{329}$	$12.50\frac{050}{043}$	$9.98\frac{039}{139}$	$8.27\frac{088}{058}$	$5.02\frac{111}{084}$	$3.30\frac{483}{400}$	$2.99\frac{925}{942}$

Conjecture 1. • There exists a monotone increasing function $h:\mathbb{R}_{>0}\to\mathbb{R}_{>0}$ given by the limit

$$h(z):=\lim_{\substack{m,n\to\infty\\ n/m\to z}}\mathbf{E}\,\tau^{mn};$$

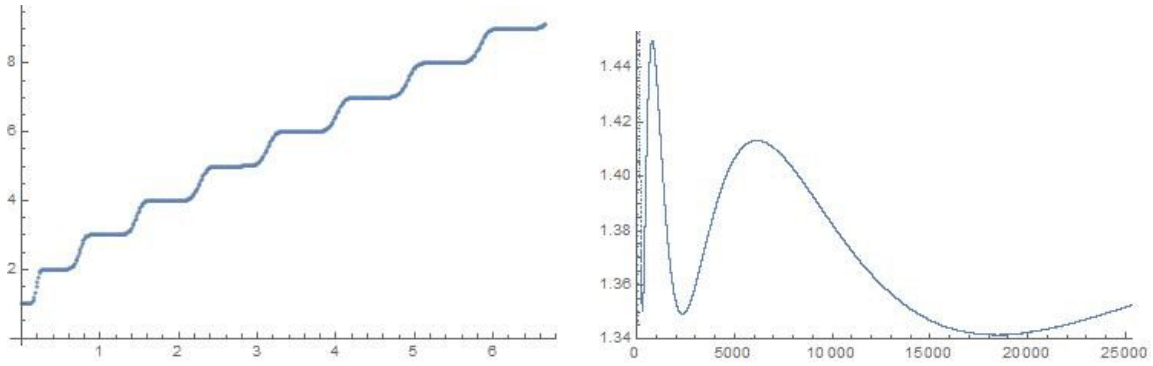
- $h(z) \searrow 1$ when $z \searrow 0$;
- $h(1) = 3$ (for example $\mathbf{E}\,\tau^{750\,750} \approx 3 - 1.1 \cdot 10^{-8}$);
- $h(z)/z \searrow 1$ and $h(z) = z + \frac{1}{2}\log z + O(1)$ when $z \to +\infty$;

Proof. We can proof only part of the above statements, other are results of numerical experiments. 1. For fixed $m, n \in \mathbb{Z}_{>0}$ the function $\mathbb{Z}_{>0} \to \mathbb{R}_{>0}$, $a \mapsto \mathbf{E}\,\tau^{am\,an}$ is monotone increasing. 2. The expectation $\mathbf{E}\,\tau^{mn}$ can be majorized by $\mathbf{E}\,\zeta_r^n$ form coupon collector model:

$$\begin{aligned}\mathbf{E}\,\tau^{m\,bm} - \mathbf{E}\,\tau^{m\,m} &\leqslant (\mathbf{E}\,\zeta_m^{bm} + \mathbf{E}\,\zeta_m^{(b-1)m} + \cdots + \mathbf{E}\,\zeta_m^{2m})/m \\ &= b(H_{bm} - H_{(b-1)m}) + (b-1)(H_{(b-1)m} - H_{(b-2)m}) + \cdots + 2(H_{2m} - H_m) \\ &\approx \log\frac{b^b}{(b-1)!} \underset{b\gg 1}{\approx} b + \frac{1}{2}\log\frac{b}{2\pi}.\end{aligned}$$

□

The graph of $h(x)$ for small x looks like a ladder. From other hand we can see the asymptotic of $h(x)$ for $x \to \infty$. On the figures bellow graphs of the functions $n/750 \mapsto \mathbf{E}\,\tau^{750\,n}$ and $n/50 \mapsto \mathbf{E}\,\tau^{50\,n} - n/50 - \ln(n/50)/2$ give suitable approximations:



The behaviour of the variance $\mathbf{V}\mathrm{ar}\,\tau^{mn}$ is more complicated. Our numerical calculations allows to suppose that $\mathbf{V}\mathrm{ar}\,\tau^{mn} < 1$ if $m \geqslant 10$ and $n/m < 10000$.

Remark 1. Example 5 allows to obtain rough but very quick estimation of proof construction success. Note that $\xi_k^{mn} \geqslant n - r$ implies that $\#\{\sigma_i(j) | 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant r\}$. So calculating probabilities we have $\mathbf{P}(\xi_k^{mn} \geqslant n - r) \leqslant \binom{n}{\ell}\left(\frac{(r)_k}{(n)_k}\right)^m$. In particular, $\mathbf{P}(\tau^{mn} > k) = \mathbf{P}(\xi_k^{mn} \geqslant 1) \leqslant n\,(1 - k/n)^m \underset{k/n\ll 1}{\approx} ne^{-k\frac{m}{n}}$.

4 Distributed generation of Merkle trees

Our practical task is to generate proofs for nodes of Merkle tree. The nodes form a partially ordered set (poset) whose Hasse diagram is the tree itself.

Some basic facts about posets can be found in [Sta11, ch.3]. Let P be a poset. A subset $I \subseteq P$ is called a *down-set* (resp. *up-set*) if for each $x \in I$ and $y \in P$ with $y \leqslant x$ (resp. $y \geqslant x$) we have $y \in I$. Note that down-sets in P are up-sets in the

opposite poset P^{op} and vice versa. And $I \subseteq P$ is a down-set iff its complement $P \setminus I$ is an up-set. The set of up-sets in P form a distributive lattice ordered by inclusion (this statement is a part of Birkhoff's representation theorem). Denote $\min P$ the set of minimal elements in P.

A Merkle tree M_ℓ with $2^\ell - 1$ nodes as a poset consists of words of length $< \ell$ in alphabet of two letters, say $\{0, 1\}$; and $w \geqslant w'$ iff w' start with w. So the empty word corresponds to the greatest element, the root. The number u_ℓ of up-sets in this poset satisfies the recurrent relation $u_{\ell+1} = u_\ell^2 + 1$ (the sequence A003095).

One can generalise Makov chains from Examples 1,2 to the case of poset N. In particular, for poset-guided analog of coupon collector Markov chain: a graph (not mentioning loops) is the Hasse diagram for the lattice of down-sets in N. The further lumping like in Example 3 exists only for special posets.

Let N be a poset. We consider a Markov chain, where states are up-sets in N. Non-zero elements of transition matrix are

$$p(N', N'') = |N' \setminus N''|! \cdot S(m, |N' \setminus N''|) \cdot |\min N'|^{-m},$$

where $N' \setminus \min N' \subseteq N'' \subseteq N'$ and $|N' \setminus N''| \leqslant m$. If N is a discrete poset we obtain a Markov chain from Example 4.

Note that very similar constructions around Birkhoff's representation theorem describe shapes of cells of higher categories in [Bes19].

Remark 2. We can extend the model from Example 5 to the case of poset N. The only modification is to define a linear ordering of N as a monotone bijections $\sigma : N \xrightarrow{\cong} \{1 < 2 < \cdots < |N|\}$.

The number of linear orderings of a Merkle tree: $|\operatorname{ord}(M_{\ell+1})| = |\operatorname{ord}(M_{\ell+1})|^2 \binom{2^\ell - 2}{2^{\ell-1}-1}$ and $|\operatorname{ord}(M_{\ell+1})| = \prod_{k=1}^{\ell-1} \binom{2^{k+1}-2}{2^k-1}^{2^{\ell-k-1}} = (2^\ell - 1)! / \prod_{k=2}^{\ell} (2^k - 1)^{2^{\ell-k}}$.

In this more general situation Proposition 2 is broken if we use equiprobability distributions. An analog of this proposition remains true if we consider a system of agreed probability distributions in general different from uniform.

References

[Bes19] Yuri Bespalov, *Categories: Between cubes and globes. Sketch I*, Ukrainian Journal of Physics **64** (2019), no. 12, 1125–1128.

[KS76] John G. Kemeny and J. Laurie Snell, *Finite Markov chains*, - Undergraduate Texts in Mathematics, Springer-Verlag, 1976.

[LPW17] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer, *Markov chains and mixing times*, 2nd ed., AMS, 2017.

[O'N19] Ben O'Neill, *The classical occupancy distribution: Computation and approximation*, The American Statistician (2019).

[Sta11] Richard P. Stanley, *Enumerative combinatorics*, 2nd ed., - Cambridge studies in advanced mathematics, 49, vol. 1, Cambridge University Press, 2011.