# Modifications of Simon text model

Artyom Kovalevskii (joint work with M. Chebunin)

Novosibirsk State Technical University
Novosibirsk State University

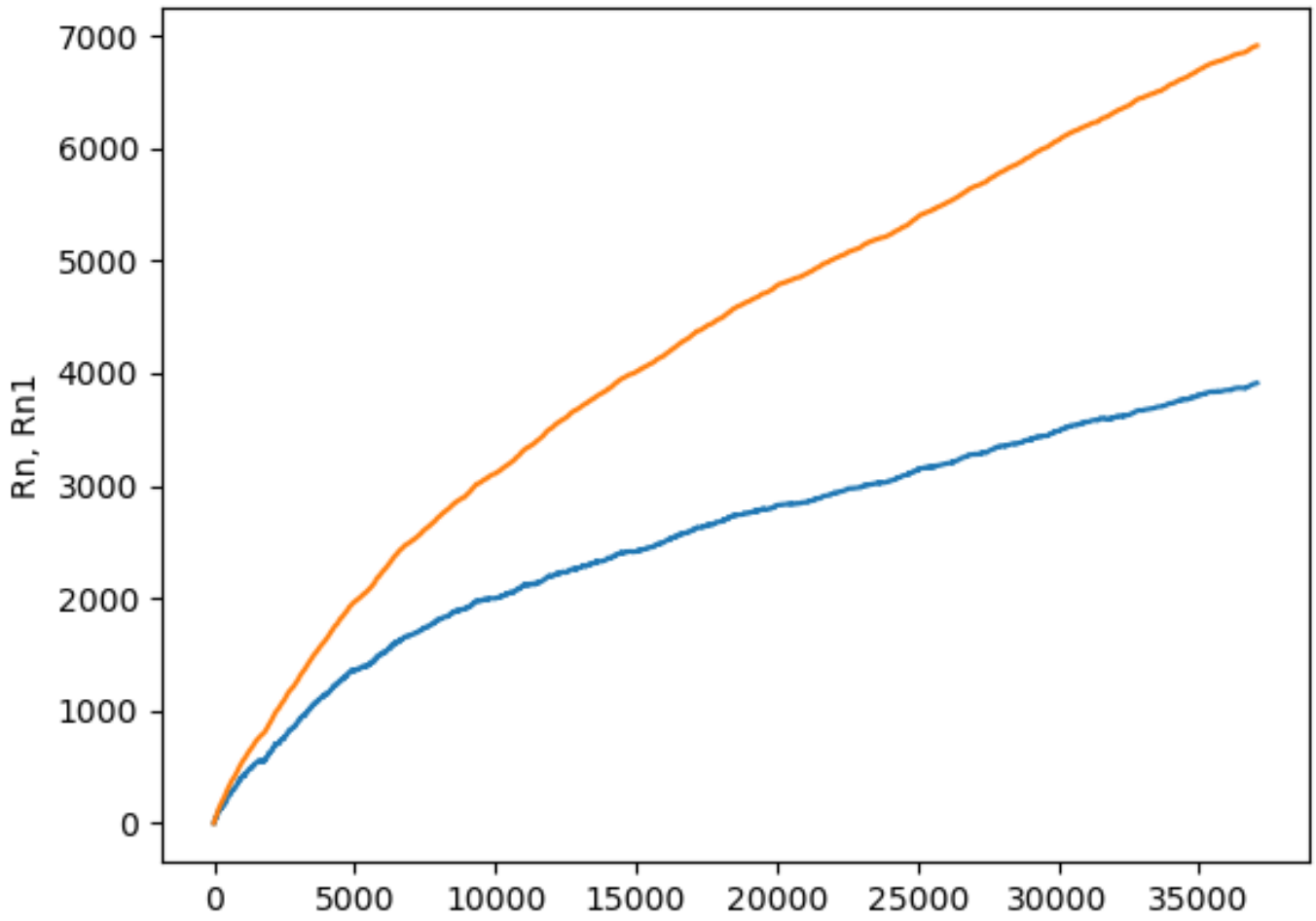
*artyom.kovalevskii@gmail.com*

Novosibirsk, 2020

# Text statistics

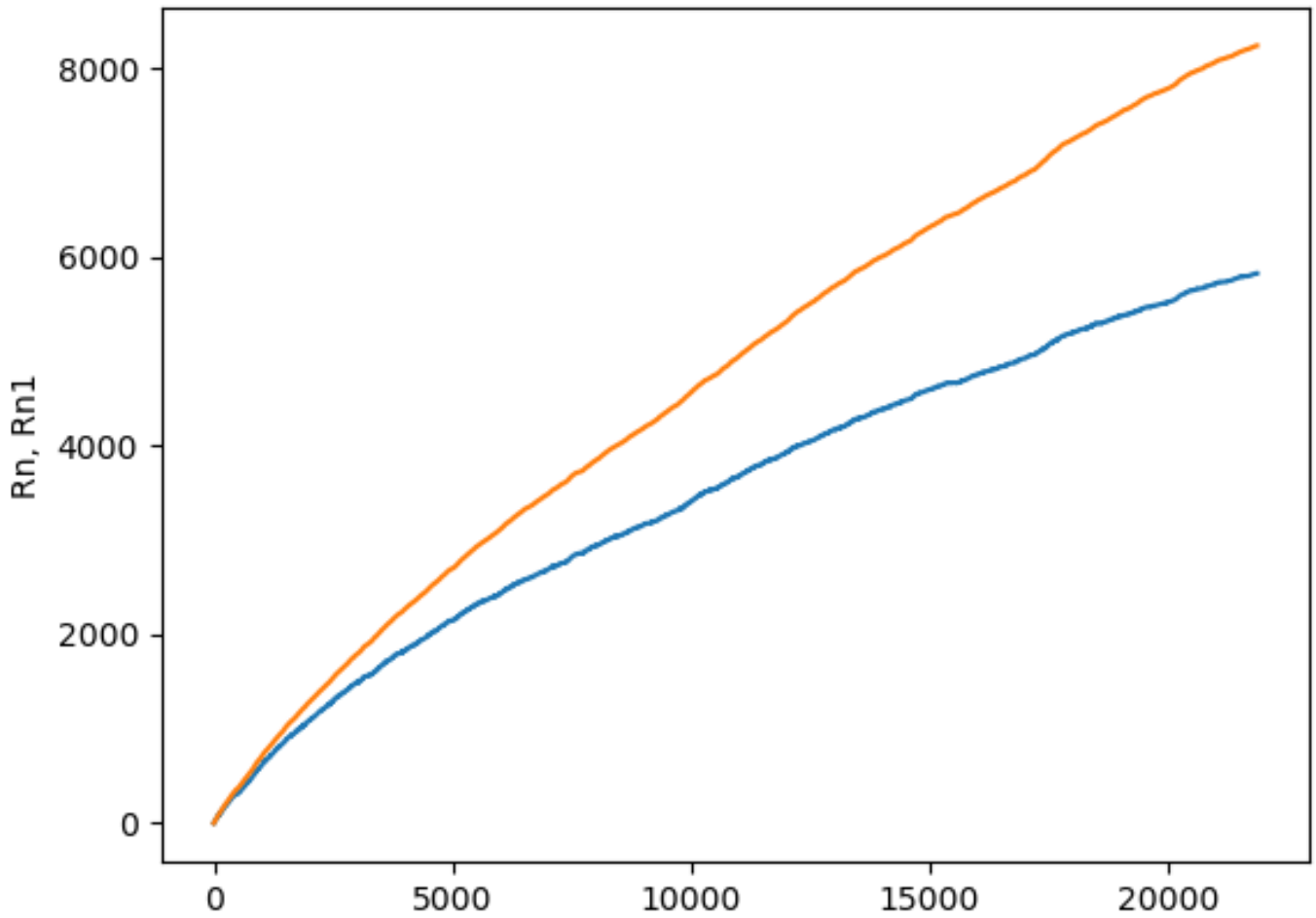
Let $n$ be the number of words in a text.
Denote by $R_k$ the number of different words in the first $k$ words of the text, $k = 1, \ldots, n$.
Denote by $R_{k,1}$ the number of unique (that is, occured only once) words in the first $k$ words of the text, $k = 1, \ldots, n$.
So $R_1 = 1$, $R_{1,1} = 1$.
$R_n$ and $R_{n,1}$ are numbers of different and unique words in all the text, correspondingly.
We are interesting in probabilistic models for sequences $R_k$ and $R_{k,1}$.

# Empirical data




*Childe Harold's Pilgrimage* by Byron, $n = 37064$, $R_n = 6911$, $R_{n/2} = 4582$, $R_{n,1} = 3912$.

# Empirical data




*Evgene Onegin* by Pushkin (in Russian), $n = 21882$, $R_n = 8236$, $R_{n/2} = 4916$, $R_{n,1} = 5824$.

# Elementary model

There is a countably infinite dictionary where the words are numbered 1, 2, .... Words are chosen one-by-one independently of each other and accodingly to Zipf-Mandelbrot law [Zipf, 1936], [Mandelbrot, 1965]

$$p_i = c(i+q)^{-1/\theta}, \quad i \geq 1, \quad 0 < \theta < 1, \quad q > -1,$$

$c$ is the normalising constant.
Statistics of the number of different words $R_n$ and the number of unique words $R_{n,1}$ have been studied for generalisations of this model by Bahadur (1960), Karlin (1967), Bogachev, Gnedin, Yakubovich (2008), Barbour (2009), Barbour and Gnedin (2009), Chebunin (2014), Ben-Hamou, Boucheron, Ohannessian (2017), Chebunin and K. (2016, 2019), Zakrevskaya and K. (2001, 2019).

# Properties of the elementary model

$$\mathrm{E}R_n = \sum_{i=1}^{\infty}(1-(1-p_i)^n)$$

$$\mathrm{E}R_n \sim c^{\theta}\Gamma(1-\theta)n^{\theta}$$

where $\Gamma(\cdot)$ is the Euler gamma function [Bahadur, 1960]

$$\mathrm{E}R_{n,1} = n\sum_{i=1}^{\infty} p_i(1-p_i)^{n-1}$$

$$\mathrm{E}R_{n,1} \sim \theta \mathrm{E}R_n$$

$$R_n/\mathrm{E}R_n \overset{a.s.}{\to} 1$$

$$R_{n,1}/\mathrm{E}R_{n,1} \overset{a.s.}{\to} 1$$

[Karlin, 1967]

# Parameter estimates

For the elementary model,

$$\widehat{\theta} = \log_2 \frac{R_n}{R_{[n/2]}} \to \theta \text{ a.s.},$$

$$\theta^* = R_{n,1}/R_n \to \theta \text{ a.s.}$$

For *Childe Harold's Pilgrimage* we have

$$\widehat{\theta} = 0.5929,\ \theta^* = 0.5661.$$

For *Evgene Onegin* we have

$$\widehat{\theta} = 0.7445,\ \theta^* = 0.7071,$$

so the second estimate is smaller than the first one.

# FCLT for the elementary model

**Theorem** (Chebunin, K., 2016)
There is weak convergence of the process $(Z_n,\ Z_{n,1})$,

$$Z_n(t) = (R_{[nt]} - \mathsf{E}R_{[nt]})/\sqrt{\mathsf{E}R_n},$$

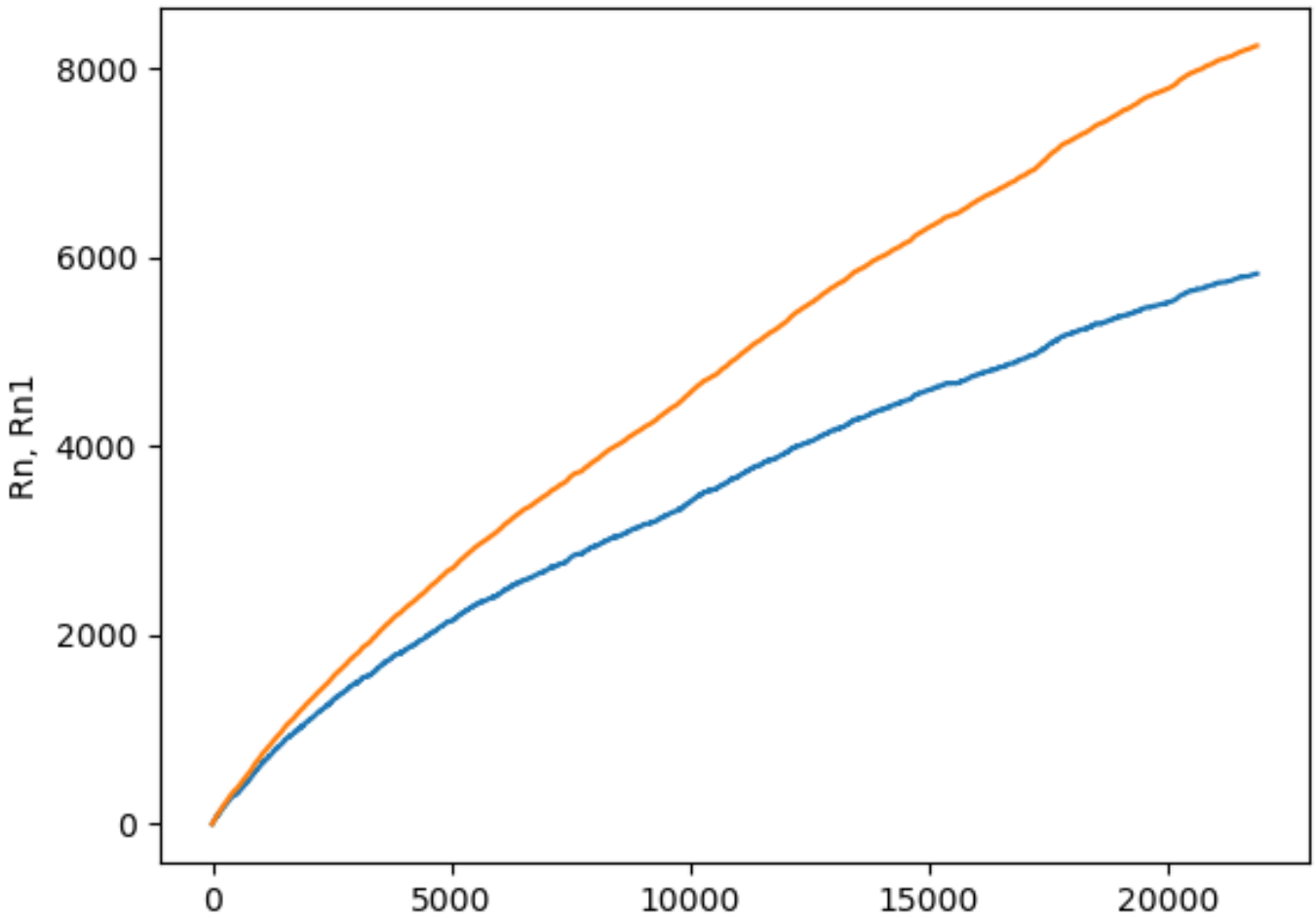
$$Z_{n,1}(t) = (R_{[nt],1} - \mathsf{E}R_{[nt],1})/\sqrt{\mathsf{E}R_n},\quad 0 \le t \le 1,$$

in $D(0,1)$ with uniform metrics to a centered Gaussian process $(Z_\theta,\ Z_{\theta,1})$ with continuous a.s. sample paths and covariance matrix function that depends on $\theta$ only.
**Corollary** There is $\sigma(\theta) > 0$ such that

$$\frac{\sqrt{R_n}(\widehat{\theta} - \theta^*)}{\sigma(\theta)}$$

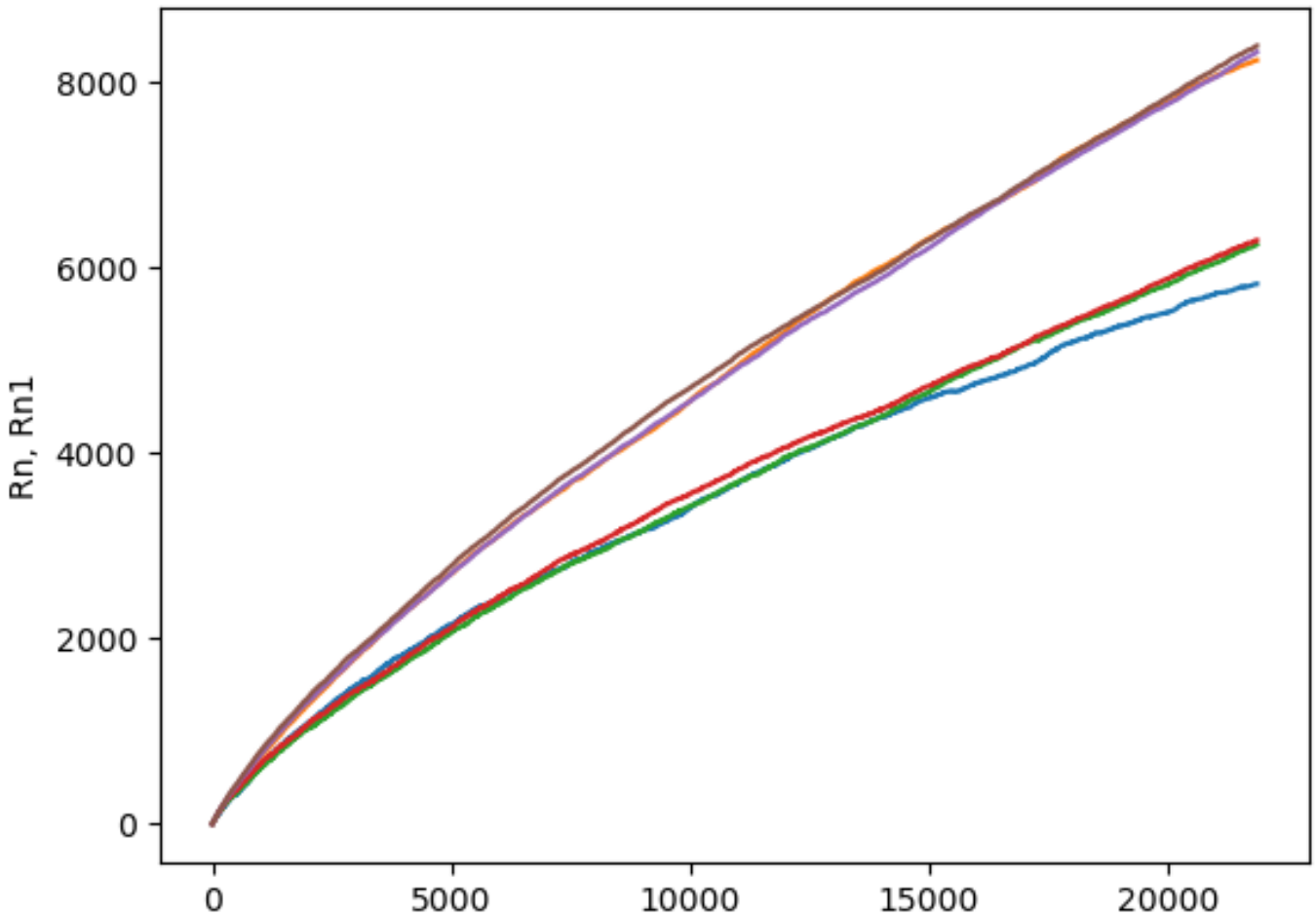
converges weakly to standard normal distribution.
Typically, the number of words that occur once is significantly less than $\widehat{\theta} R_n$.

# Simulation




*Evgene Onegin* by Pushkin
(in Russian)

# Simulation




Simulations of *Evgene Onegin* by Pushkin (in Russian),
$\theta = 0.7392$, $q = 90$

# Simon model

Simon (1955) proposed the next stochastic model: the $(n+1)$-th word in the text is new with probability $p$; it coincides with each of the previous words with probability $(1-p)/n$.
The drawback of Simon's model is that the number of different words grows linearly.
Yule showed that

$$\mathrm{E}R_{n,i}/\mathrm{E}R_n \to f(i), \quad i \geq 1, \tag{1}$$

$$f(i) = \rho B(i, 1+\rho),$$

$\rho = (1-p)^{-1}$, $B(\cdot,\cdot)$ is Beta function.
We analyze stochastic aspects of this convergence. The limiting distribution is named Yule-Simon distribution.

# Simon model

There are many ramifications and applications of Yule-Simon model. Haight & Jones (1974) gave special references to word associations tests. Lansky & Radill-Weiss (1980) proposed a generalization of the model for better correspondence to applications.
This model can be embedded in more general context of random cutting of recursive trees. In this context, statistics under study are most frequent words. See Aldous & Pitman (1998) for its limiting distribution and convergence, Baur & Bertoin (2014, 2015) for an overview and new results.
Aldous (1996) proposed a generalization of the limiting distribution but without an underlying process.

# Simon model

Janson (2004) considered generalized Polya urns and proved SLLN, CLT and FCLT for it. Finite-dimentional vectors

$$(R_{n,1}, \ldots, R_{n,m-1}, n - \sum_{i<m} iR_{n,i})$$

can be studied using these models. So we have componentwise SLLN and finite-dimentional CLT and FCLT for these statistics.

# FCLT for Simon model

## Theorem

*For any $p \in (0,1)$ in Simon model and any $m \geq 1$*

$$\left( \frac{R_{n,1}}{n}, \dots, \frac{R_{n,m}}{n} \right) \to \frac{p}{1-p} \left( B\left(1, \frac{2-p}{1-p}\right), \dots B\left(m, \frac{2-p}{1-p}\right) \right)$$

*a.s.,*

$$\left\{ n^{-1/2} \left( R_{[nt],j} - tn \frac{p}{1-p} B\left(j, \frac{2-p}{1-p}\right), \ 1 \leq j \leq m \right), \ t \geq 0 \right\} \to_d V$$

*in $D(0,\infty)$, $V$ is the centered $m$-dimensional Gaussian process with continuous a.s. trajectories, its covariance matrix-function $\mathrm{E}V(x)V^T(y)$ depends on $p, x, y$ only.*

# Modification of Simon model

The disadvantage of Simon model is the linear growth of $R_n$. We need a power growth with exponent lesser than 1.
Baur and Bertoin (2020) proposed a process that give a two-parameter Yule-Simon distribution in a limit. Their model proposes dependence $p = p_n$.
Our idea is to use classical infinite urn model with re-distribution: any ball takes an urn independently with some discrete power law; then with probability $1 - p$ it is re-distributed uniformly on all previous balls.

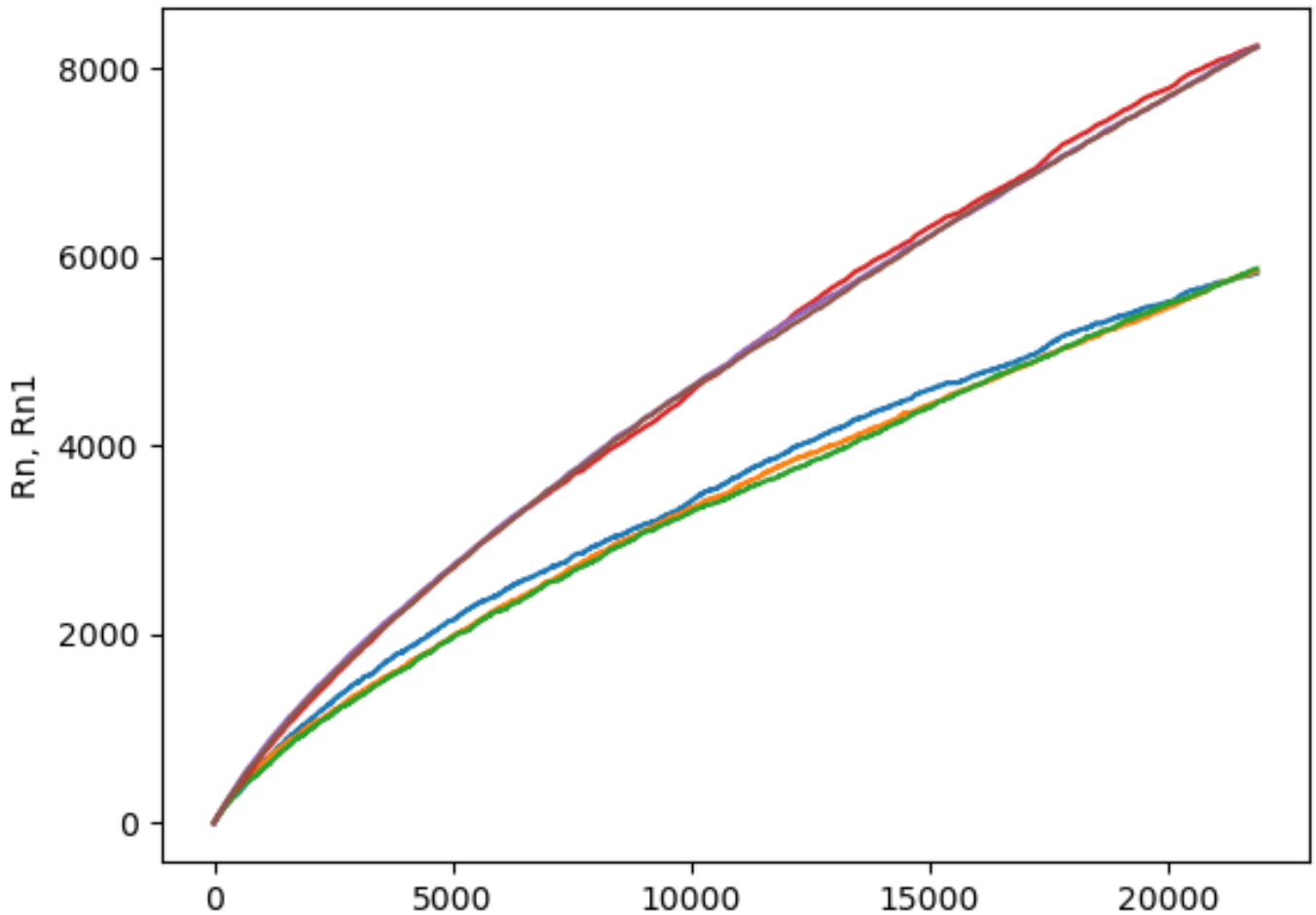
# Modification of Simon model

The model starts from the infinite sequence of empty urns. First ball takes one of the urns with the integer-valued power law with exponent $-1/\beta$, $0 < \beta < 1$. Each next ball takes one of the urns with the same law, independently of previous balls. After this, if this ball entered an empty urn, then this ball is re-tossed indepedently, that is, with probability $1 - p$ selects one of the previous balls at random and joins it, like in Simon model. All other balls stay in selected urns.

# Modification of Simon model

For *Evgene Onegin* we need some nonzero $p$, that is, $\beta = 0.73$, $q = 130$, $p = 0.92$.
The simulation shows that statistics $\widehat{\theta}$ and $\theta^*$ converge to different limits under these models, in contrast to the elementary urn model. However, analytical dependencies of these limits on the parameters of the models remain unclear.

# Simulation by Simon-Zipf model




Simulations of *Evgene Onegin* by Pushkin (in Russian),
$\beta = 0.73$, $q = 130$, $p = 0.92$