Skoltech
Skolkovo Institute of Science and Technology

Skolkovo Institute of Science and Technology

ON CONNECTION BETWEEN SPARSE GRAPHS AND HYPERBOLIC GEOMETRY

Doctoral Thesis

by

KIRILL POLOVNIKOV

DOCTORAL PROGRAM IN PHYSICS

Supervised by
Professor Mikhail Gelfand, Skoltech
Professor Sergei Nechaev, Interdisciplinary Scientific Center Poncelet

Moscow —

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Kirill Polovnikov)

Supervisors (Prof. Mikhail Gelfand, Prof. Sergei Nechaev)

Abstract

Sparse matrices play an enormous role in physics, biology, finance and in many other fields of science. They correspond to the networks, in which the number of *theoretically possible* connections (relationships) between the modules largely exceeds the *actual* number of connections. In this thesis I report our results showing that the average spectral density of the ensemble of sparse graphs can be approximated by the Dedekind η-function, which is a modular form with respect to the modular group $SL(2, \mathcal{Z})$. The asymptotic behaviour of the spectral density in all rational points within the support is shown to reproduce two-dimensional Lifshitz tails, which turn into one-dimensional at the edge of the spectrum. This finding unravels a fundamental connection between localization in low dimensions and the hyperbolic geometry. The emerging dimension $D = 2$ of eigenvalues of a large sparse matrix revisits the well-known repulsion of eigenvalues in Gaussian invariant ensembles, responding to the two-dimensional Coulomb gas confined to a line.

Ultrametric structure of the spectral density can be explicitly constructed by considering an exponentially growing planar tissue on the Euclidean plane. Strong incompatibility of the local differential growth protocol with the geometric constraints evokes buckling of the tissue into the third dimension. We show that the buckling profile can be described by the eikonal equation and the metric is expressed through the Dedekind η-function. As another example of the system that gets pushed into a regime with strong correlations, we consider two-dimensional random trajectories evading obstacles of different geometrical shapes. We demonstrate that in the strong stretching regime the circular trajectories fluctuate with the PDF that is described by one of the tails of the Tracy-Widom distribution and the one-dimensional KPZ growth exponent.

The isolated spectrum of random walks on a graph frequently becomes a robust tool for the dimensionality reduction of sufficiently dense data. In the sparse case localization on star-like graphs takes place, however, the non-backtracking walks are able to perform community detection in this case. In this thesis, using spectral properties of different stochastic operators we investigate topological structure of two real world networks: core-periphery organization of the cryptocurrencies network and communities in chromatin networks. In the latter case, we propose two operators, based on the non-backtracking walks, whose spectra

reflect biologically significant communities in single cell Hi-C maps. Our approach provides a generalized framework for communities in Erdös-Rényi graphs beyond the conventional stochastic block model.

List of publications

1. Nechaev, S.K., Polovnikov K.E., Rare events statistics and modular invariance, Phys. Usp. 61(1), 2018;
2. S. Nechaev, K. Polovnikov, From geometrical optics to plants: eikonal equation for buckling, Soft Matter 13: 1420-1429, 2017;
3. Nechaev, S., Polovnikov, K., Shlosman, S., Valov, A., Vladimirov, A. Anomalous 1D fluctuations of a simple 2D random walk in a large deviation regime, Physical Review E 99(1), 012110, 2019;
4. K. Polovnikov, V. Kazakov, and S. Syntulsky. Core-periphery organization of the cryptocurrency market inferred by the modularity operator, Physica A: Statistical Mechanics and its Applications 540, 123075 (2020);
5. Sergey V. Ulianov*, Vlada V. Zakharova*, Aleksandra A. Galitsyna*, Pavel I. Kos*, Kirill E. Polovnikov, Ilya M. Flyamer, Elena A. Mikhaleva, Ekaterina E. Khrameeva, Diego Germini, Mariya D. Logacheva, Alexey A. Gavrilov, Aleksander S. Gorsky, Sergey K. Nechaev, Mikhail S. Gelfand, Yegor S. Vassetzky, Alexander V. Chertovich, Yuri Y. Shevelyov, Sergey V. Razin, Order and stochasticity in the folding of individual Drosophila genomes. Submitted to Nature Communications, (2020);
6. K. Polovnikov, A. Gorsky, S. Nechaev, S. V. Razin, S. Ulianov, Non-backtracking walks reveal compartments in sparse chromatin interaction networks, Scientific Reports 10, 11398 (2020).

Contents

Pg.

1. Introduction

The study of random matrices spectra has been greatly inspired by the problem of nuclear interactions, where complexity of the Hamiltonian was first recognized and tackled within this approach by Eugene Wigner and Freeman Dyson [1, 2]. Experimental spectra of heavy nuclei demonstrated the same statistical properties as spectra of *random* matrices. Namely, the spacing distribution of neighboring level states was shown to be in an excellent agreement with the so-called Wigner surmise, $P(s) = \frac{\pi s}{2} \exp\left(-\pi s^2/4\right)$, which is the exact result for the Gaussian orthogonal ensemble [3]. Universality of their approach has made the random matrix theory (RMT) one of the most powerful tools in complex systems research and has revealed a vast number of applications in disordered systems, number theory, quantum information, integrable systems and quantum chromodynamics, as well as in finance and networks [4, 5].

In general, the spectral decomposition of a random matrix is a non-trivial task: while the entries of the random matrix, in the simplest case, are independently distributed, its eigenvalues usually turn out to be strongly correlated. Indeed, it is already seen from the form of the Wigner surmise: for absolutely uncorrelated random variables the distance between the neighboring states would follow the Poisson distribution. In the Wigner level spacing the probability of consecutive eigenvalues to be close becomes arbitrary small, reflecting effective repulsion of the eigenvalues.

1.1 Spectrum of dense random matrices

1.1.1 Gaussian invariant ensemble

Following the classical line of argument, let us consider an ensemble of dense random matrices of size N

$$J = \begin{bmatrix} J_{11} & J_{12} & \dots & J_{1N} \\ J_{21} & J_{22} & \dots & J_{2N} \\ \dots & \dots & \dots & \dots \\ J_{N1} & J_{N2} & \dots & J_{NN} \end{bmatrix}$$

with J_{ij} being i.i.d. random variables drawn from the normal distribution. Then, the probability to observe the matrix J is as follows

$$P[J] \propto \exp\left[-\frac{1}{2}\sum_{i,j}|J_{i,j}|^2\right] \tag{1.1}$$

Gaussian distribution of the entries is chosen not only for the sake of simplicity, but because the resulting measure of the matrix is invariant under rotation of the coordinate system. In fact, if one requires that a random matrix with independent entries (Wigner ensemble) is rotational invariant, then the matrix belongs to the Gaussian ensemble. This is a consequence of a theorem by Porter and Rosenzweig [6, 3]. Weyl's lemma [7] states that the rotational invariant measure (1.24) can only be a function of powers of traces, $P[J] = f(TrJ, TrJ^2, TrJ^3, ...)$, by the cyclic property of the trace. Thus, the Gaussian measure (1.24) needs to possess additionally certain symmetries in order to be rotational invariant. For example, if J is complex, then J needs to be a Hermitian matrix and one can rewrite (1.24) through the trace of $|J|^2$.

In general, the nature of the matrix entries fixes the universality class of rotational invariant measures, which is encoded in the value of the parameter β (Dyson index). If the entries are real numbers, the number of possible components is one and $\beta = 1$. For complex and quaternion entries $\beta = 2$ and $\beta = 4$, accordingly. Therefore, for independent entries with rotational invariance one enjoys three possible classes of universality (i) $\beta = 1$, Gaussian orthogonal ensemble (GOE), (ii) $\beta = 2$, Gaussian unitary ensemble (GUE) and (iii) $\beta = 4$, Gaussian

symplectic ensemble (GSE). Let us rescale the matrix elements by $\sqrt{\beta N}$, then the universal Gaussian measure (1.24) reads

$$P[J] \propto \exp\left[-\beta \frac{N}{2} Tr(J^\dagger J)\right] \tag{1.2}$$

Having the measure of the matrix defined, one is generally interested in the joint probability distribution of N eigenvalues. For rotationally invariant measures this quantity is exactly solvable and all eigenvalues are real. One typically needs to change the variables in order to diagonalize the matrix. Real symmetric matrices can be diagonalized by some orthogonal transformation $J = O\Lambda O^T$, where $\Lambda = diag(\lambda_1,\lambda_2,...,\lambda_N)$ and O is the matrix of the eigenvectors. Then, one have to switch from $N(N+1)/2$ independent variables $\{J_{ij}\}$ to N eigenvalues $\{\lambda_i\}$ and $N(N-1)/2$ independent components of the eigenvectors $\{O_{ij}\}$

$$P(\{J_{ij}\}) \prod_{i \leqslant j} dJ_{ij} =$$

$$P\left(J_{11}\left(\{\lambda_i, O_{ij}\}\right), ..., J_{NN}\left(\{\lambda_i, O_{ij}\}\right)\right) \left|\{J_{ij}\} \to \{\lambda_i, O_{ij}\}\right| d\mathbf{O} \prod_{i=1}^{N} d\lambda_i \tag{1.3}$$

and to compute the Jacobian of the transformation. For the rotationally invariant ensembles this Jacobian depends only on the eigenvalues and is precisely a so-called Vandermonde determinant

$$\left|\{J_{ij}\} \to \{\lambda_i, O_{ij}\}\right| = \prod_{j<k} (\lambda_j - \lambda_k) \tag{1.4}$$

For the other two types of symmetry the Vandermonde is raised to the power β. Finally, for the Gaussian measure (1.2) the joint probability distribution takes the following simple form

$$P(\lambda_1, \lambda_2, ..., \lambda_N) = \frac{1}{Z_N} \exp\left[-\frac{\beta}{2} N \sum_{i=1}^{N} \lambda_i^2\right] \prod_{j<k} |\lambda_j - \lambda_k|^\beta \tag{1.5}$$

where the latter "interaction"term comes from the Jacobian and Z_N is the partition function

$$Z_N = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{i=1}^{N} d\lambda_i \exp\left[-\beta \frac{N}{2} \sum_{i=1}^{N} \lambda_i^2\right] \prod_{j<k} |\lambda_j - \lambda_k|^\beta \tag{1.6}$$

It is straightforward to see from (1.6) that the eigenvalues repel each other as a two-dimensional Coulomb gas confined to a line

$$Z_N = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{i=1}^{N} d\lambda_i \exp\left[-\frac{\beta}{2}\left\{\sum_{i=1}^{N} N\lambda_i^2 - \sum_{j\neq k} \log|\lambda_j - \lambda_k|\right\}\right] \tag{1.7}$$

1.1.2 Average density of states

It is worth to note that apart from Coulomb-gas repulsion between the eigenvalues, the first term under exponent in the partition function (1.7) responds for the external harmonic field accumulating all the eigenvalues at the origin. The two terms are of the same order $O(N^2)$ and balance each other, making the typical eigenvalue $\lambda_{typ} \sim O(1)$. This allows to rewrite the partition function in the following scaling form

$$Z_N = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod d\lambda_i \exp\left[-\beta N^2 E(\{\lambda_i\})\right] \tag{1.8}$$

where the scaled energy of the Coulomb gas $E(\{\lambda_i\})$

$$E(\{\lambda_i\}) = \frac{1}{2N}\sum_{i}^{N} \lambda_i^2 - \frac{1}{2N^2}\sum_{j\neq k} \log|\lambda_j - \lambda_k| \tag{1.9}$$

or, in the continuous limit,

$$E(\rho(\lambda)) = \frac{1}{2}\left[\int \lambda^2 \rho(\lambda) d\lambda - \int\int \log|\lambda - \lambda'|\rho(\lambda)\rho(\lambda')d\lambda d\lambda'\right] \tag{1.10}$$

with $\rho(\lambda) = \frac{1}{N}\sum \delta(\lambda - \lambda_i)$ is the "charge"density. Minimization of the energy subject to the constraint $\int \rho(\lambda)d\lambda = 1$ in the thermodynamic limit $N \to \infty$ allows to recover the famous Wigner semi-circle law for the average density of states. Variation of the action functional with respect to $\rho(\lambda)$ produces an integral equation, which can be solved using the Tricomi theorem [8] for compact supports $\lambda \in [a, b]$ (see, for example, [9] for details). Finally, the scaled semi-circle with $R = \sqrt{2}$ can be obtained

$$n(\lambda) = \langle\rho(\lambda)\rangle = \frac{1}{\pi}\sqrt{2 - \lambda^2}. \tag{1.11}$$

Though the semi-circle law is most easily derived for the Gaussian symmetrical ensembles, it can be shown that it holds for a large class of Wigner matrices (non-invariant) as soon as the distribution of the entries decays sufficiently fast [10]. However, in general, there is no reason to expect the semi-circle for an arbitrary symmetric matrix, especially, when its entries are not guaranteed to be independent. For example, for the Wishart ensemble of covariance matrices $W = X^\dagger X$, where X is a rectangular $M \times N$ random Gaussian matrix (real or complex), $c = N/M \leqslant 1$, the spectral density is replaced by the Marčenko-Pastur distribution [11]

$$n_{MP}(\lambda) = \frac{1}{2\pi\lambda}\sqrt{(\lambda - a)(b - \lambda)} \tag{1.12}$$

where $[a, b]$ defines the support and $a = \left(c^{-1/2} + 1\right)^2$, $b = \left(c^{-1/2} - 1\right)^2$.

The presented Coulomb-gas approach to the semicircle law is based on the mean-field argument and, thus, is valid in the thermodynamic limit. In other words, when $N \to \infty$, there is a *hard* boundary at $\lambda_{max} = \sqrt{2}$, which all eigenvalues of an infinite-size random matrix cannot exceed. However, at finite N some of the leading eigenvalues can overcome the boundary of the semi-circle due to fluctuations. Two natural questions arise: what is the finite-N correction to the typical width of this boundary and how does the largest eigenvalue fluctuate? It turns out that typical fluctuations of the largest eigenvalue are much more universal than the shape of the average spectral density.

1.1.3 Extreme eigenvalue statistics

Though the questions raised are undoubtedly interesting in their own right, there is a deep physical motivation for the extreme eigenvalue statistics. The first example comes from the problem of a light particle moving on a N-dimensional landscape, $V(y_1, y_2, ..., y_N)$

$$\frac{dy_i}{dt} = -\nabla_{y_i} V \tag{1.13}$$

If the landscape is rugged (which is usually the case in the most of interesting physics scenarios, eg. glasses [12, 13] or string landscapes [14]), there are many stationary points y^*, such as $\nabla V|_{y=y^*} = 0$. However, only exponentially small number of them

are stable. Indeed, near stationary points the curvature of the landscape is described by the Hessian matrix $H_{ij} = \frac{\partial^2 V}{\partial y_i \partial y_j}|_{y=y^*}$, which is a symmetric matrix with N real eigenvalues. In particular, eigenvalues of the Hessian matrix determine stability of the stationary points. If all $\lambda_i < 0$, the stationary point is a local maximum; if all $\lambda_i > 0$, the stationary point is a local minimum; as long as some of the N eigenvalues have different sign, the stationary point inevitably becomes a saddle. In the spirit of the RMT a sufficiently rugged and complex landscape can be associated with a random Hessian matrix, belonging to GOE universality class. Then, in the framework of the random Hessian model the fraction of local maxima (minima) is given by statistics of the largest (smallest) eigenvalue of the random matrix

$$q_N = Prob.[\lambda_1 \leqslant 0, \lambda_2 \leqslant 0, ..., \lambda_N \leqslant 0] = Prob.[\lambda_{max} \leqslant 0] \sim \exp\left[-\theta N^2\right] \quad (1.14)$$

which is exponentially small, implying that most of the stationary points of a complex landscape are saddles. Exact result for the stability parameter is derived in [15], $\theta = \frac{1}{4}\log(3) \approx 0.27$.

Another seminal example arises in ecosystems and is provided by the classical work of Robert May [16]. Let us consider a dynamical system of N distinct species, interacting with each other and collectively causing dramatic consequences for some of the species. To begin with, we shall consider the non-interacting system around a fixed point, characterized by stationary densities $c_i^*, i = 1, 2, ..., N$. In the vicinity of the fixed point any small perturbation induces the reverse response that aims at bringing the system back to the equilibrium. The effect of these small perturbations on relaxation can be captured by a harmonic potential, $\frac{dx_i}{dt} = -x_i(t)$, where $x_i = c - c_i^*$ is deviation from the equilibrium density. Now, one may ask, what happens with stability of the fixed points, when one switches on pairwise interactions between the species? Interestingly, more complex the interactions between the species are, the more *universal* is the behavior of the system. Indeed, the evolution of a particular kind i near the equilibrium point is described, in the linear approximation, by the following equation

$$\frac{dx_i(t)}{dt} = -x_i(t) + \alpha \sum_{i=1}^{N} J_{ij} x_j(t) \quad (1.15)$$

where α is the strength of the interactions and matrix elements $J_{ij} = J_{ji}$ can be approximated by normal uncorrelated random variables. We see that this ecological stability problem gets mapped onto the random Hessian model discussed above.

Namely, the population is stable iff $\alpha\lambda_i - 1 \leqslant 1$ for all $i = 1, 2, ..., N$. This condition imposes restriction on the upper bound of the maximal eigenvalue $\lambda_{max} \leqslant 1/\alpha$ of the random matrix J. In his work May has noticed that there is a critical value of the interactions strength α_c, below which the population stays stable, while for larger values $\alpha > \alpha_c = 1/\sqrt{2}$ the population undergoes a sharp transition to instability. This essential result is an achievement of the random matrix theory and follows from the position of the spectral edge of the Wigner semi-circle (1.11).

However, the critical value of α_c is appropriate in the thermodynamic limit, $N \to \infty$, when the sharp transition occurs. At finite population sizes N the average bulk spectral density is not sufficient. One needs the statistics of fluctuations of the largest eigenvalue of a random matrix

$$\mathcal{F}_N(w) = Prob.[\lambda_{max} \leqslant w] \tag{1.16}$$

Note that this largest eigenvalue is *coupled* with all other eigenvalues through the Coulomb-gas interactions discussed in the previous subsection. Thus, as we will see below, fluctuations of the largest eigenvalue in such a simple random matrix model demonstrate *universal* behaviour typical to extreme value statistics of *strongly correlated* random variables in sharp contrast to the universality of i.i.d. random variables [17]. From the joint PDF for all eigenvalues (1.5) one can express the CDF for the largest eigenvalue as follows

$$\mathcal{F}_N(w) = \frac{Z_N(w)}{Z_N(w = \infty)} \tag{1.17}$$

and $Z_N(w)$ is a partition function of the confined 2D Coulomb gas in harmonic potential conditioned to the upper bound (a hard wall) at w

$$Z_N(w) = \int_{-\infty}^{w} ... \int_{-\infty}^{w} \prod_{i=1}^{N} d\lambda_i \exp\left[-\frac{\beta}{2}\left\{\sum_{i=1}^{N} N\lambda_i^2 - \sum_{j\neq k} \log|\lambda_j - \lambda_k|\right\}\right] \tag{1.18}$$

Typical fluctuations occur in the small vicinity $\Delta = \sqrt{2} - \lambda_{max}$ of the spectral edge,

$$\int_{\sqrt{2}-\Delta}^{\sqrt{2}} n(\lambda) d\lambda \sim \frac{1}{N} \tag{1.19}$$

Using Wigner semi-circle asymptotic $n(\lambda) \propto \sqrt{\sqrt{2} - \lambda}$, one arrives at the estimate $\Delta = O(N^{-2/3})$ for the width of the spectral edge [9]. The edge turns from soft

to hard as $N \to \infty$, as it should be. More accurate analysis reveals the following asymptotic behaviour of λ_{max}

$$\lambda_{max} = \sqrt{2} + \frac{1}{\sqrt{2}} N^{-2/3} \chi_\beta \tag{1.20}$$

where χ_β is a random variable with the Tracy-Widom distribution [18, 19]. For the most interesting cases $\beta = 1, 2, 4$ the TW distribution can be expressed through a solution of the Painlevé II equation

$$q''(s) - 2q^3(s) - sq(s) = 0 \tag{1.21}$$

As $s \to \infty$ the second term can be neglected which yields the Schrödinger-type equation in the linear potential, $q(s) \sim Ai(s)$. Then, the CDF of TW distribution $\mathcal{F}_\beta(x)$ is expressed through the certain integrals of $q(s)$, for example, for the unitary ensemble ($\beta = 2$) one has

$$\mathcal{F}_2(x) = \exp\left(-\int_x^\infty (s-x)q^2(s)ds\right) \tag{1.22}$$

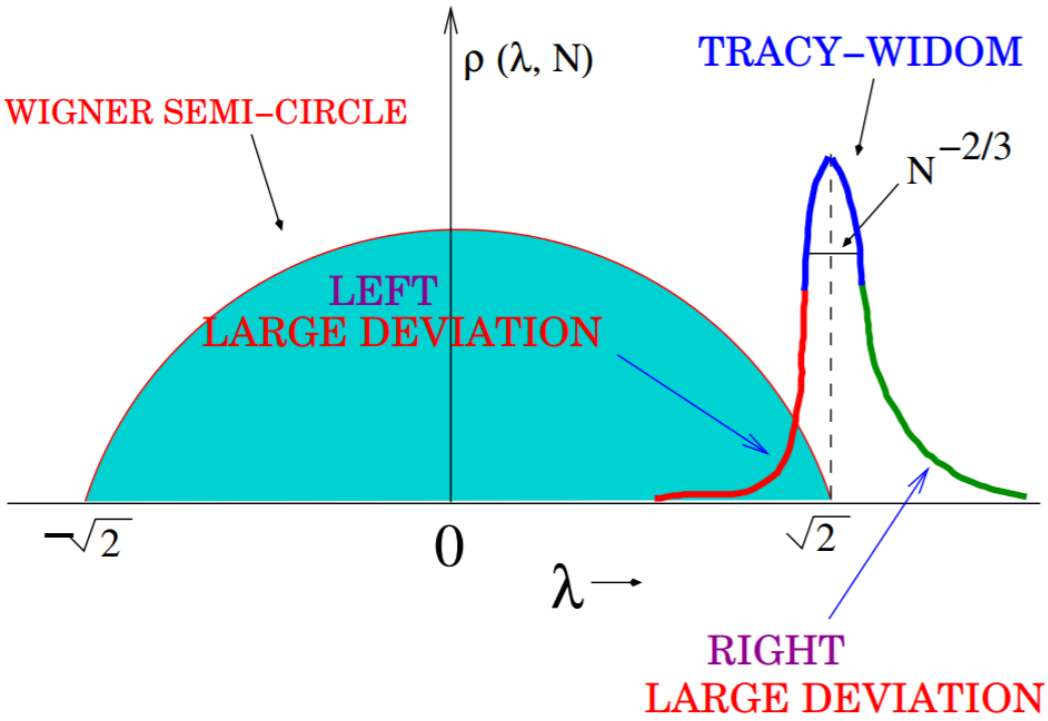

Figure 1.1 — A sketch of the scaled Wigner semi-circle for the average spectral density, the Tracy-Widom distribution for the largest eigenvalue and left and right large deviation tails. The picture is taken from [9].

Importantly, the TW distribution is asymmetric around the position of the spectral edge. The right tail, $x \to \infty$ of the PDF is inherited from the Airy function, while the left one reflects an abrupt decay with cubic dependency under the exponent

$$f_\beta(x) \sim \begin{cases} \exp\left(-\frac{\beta}{24}|x|^3\right), & x \to -\infty \\ \exp\left(-\frac{2\beta}{3}|x|^{3/2}\right), & x \to +\infty \end{cases} \tag{1.23}$$

Noteworthy, fluctuations of the top eigenvalue in the Wishart ensemble [20, 21, 22] are also described by the TW distribution, despite the average spectral density is given by a different law, (1.12). Furthermore, universality of the TW distribution has been demonstrated in a broad variety of seemingly unrelated problems united by the presence of strong correlations. It appears as a distribution function of the maximal height of N 1+1 non-intersecting Brownian motions ("viscous"walkers) [23], which, in turn, are related to 2D quantum chromodynamics [9]; in the problem of the longest increasing subsequence in random permutations [24]; 1D directed polymers in a random environment [25, 26, 22]; area-tilted random walks [27]; traffic models of the TASEP type [28]; growth models in the 1D KPZ universality class [29, 30].

Apart from typical fluctuations taking place in the small vicinity of the spectral edge $\Delta \sim N^{-2/3}$, one can be interested in anomalously large fluctuations of order $O(1)$, see Fig.1.1. Physically such atypical fluctuations correspond to either pulled or pushed Coulomb gas, when the hard wall is either taken away from the spectral boundary or compresses the gas, correspondingly. The respective tails of the distribution are derived in [31, 15, 32]. These tails smoothly approach the tails of the TW distribution as one is moving towards the spectral edge. As it is seen from (1.23) the left tail induces the cubic dependency of the free energy of the pushed Coulomb gas. Thus, as one is approaching the spectral edge from below, the third-order phase transition occurs. The critical zone of the width $N^{-2/3}$ is described by the TW distribution, therefore, the transition from strong to weak coupling is, presumingly, quite universal. In particular, it takes place for the May's model as the strength of pairwise interactions between the species decreases. See [9] for more discussion and relation to 2D QCD.

1.2 Spectrum of sparse random matrices

1.2.1 Replica and cavity methods

In general, $N \times N$ matrix is defined as sparse if the number of non-zero elements in its rows and columns is $O(1)$, i.e. does not grow with the system size. For Erdös-Rényi (ER) graphs this implies that the probability of a random link should behave as $p = q/N$, where $q > 0$ is some constant. When $q = O(N)$ the ER graph is dense and the properties of the corresponding adjacency matrix largely follow the theory of invariant Gaussian ensembles, discussed in the previous section. From the point of the tight binding Anderson model, where bonds are formed only with the neighboring sites on a d-lattice, such dense matrices correspond to $d \to \infty$. However, upon decreasing of q (diluting the system or decreasing the effective d) the average spectral density becomes spiky and does not resemble the classical semi-circle anymore.

The study of sparse matrices spectra had initiated with pioneering works of G. Rodgers and A. Bray [34] and then was advanced by many others [35, 36, 39]. Typically, one considers real symmetric matrices J_{ij} of size N, drawn from the probability distribution

$$P\left[J_{ij}\right] = \left(1 - \frac{p}{N}\right)\delta(J_{ij}) + \frac{p}{N}h(J_{ij}) \tag{1.24}$$

where $h(x)$ is some probability distribution, non-singular at $x = 0$. For the Bernoulli ensemble $h(x) = \delta(x-1)$; in the original work of Rodgers and Bray an even form was used $h(x) = 1/2\,(\delta(x-1) + \delta(x+1))$. The spectral density of a particular random realization reads

$$\rho(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\delta(\lambda - \lambda_i) \tag{1.25}$$

As shown by Edwards and Jones [40], the density (1.25) can be rewritten as

$$\rho(\lambda) = -\lim_{\varepsilon\to 0^+}\frac{2}{\pi N}\Im\left(\frac{\partial}{\partial z}\log\mathcal{Z}_J(z)\right)_{z=\lambda-i\varepsilon} \tag{1.26}$$

where $\mathcal{Z}_J(z)$ is the following partition function

$$\mathcal{Z}_J = \int \prod_{i=1}^{N} \frac{dx_i}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i,j=1}^{N} x_i\,(zI - J)\,x_j\right) \tag{1.27}$$

and I is the identity matrix. Thus, we see that the term under the exponent can be associated with the Hamiltonian

$$\mathcal{H}_J(x,z) = \frac{1}{2}\sum_{i,j=1}^{N} x_i\,(zI - J)\,x_j \tag{1.28}$$

and the problem of spectral density gets reformulated into the statistical mechanics problem of N interacting particles $x_i, i = 1,2,...,N$ with pairwise coupling constants J_{ij} and self-interaction (harmonic oscillators) with strength z.

To compute the averaged spectral density one needs to average (1.25) over the ensemble of matrices (1.24). Using the expression (1.26) one encounters calculation of the $\log \mathcal{Z}_J$, for which the replica trick has been invented

$$\langle \log Z \rangle = \lim_{n\to 0} \frac{1}{n}\left(\langle Z^n \rangle - 1\right) \tag{1.29}$$

To deal with the $n \to 0$ limit correctly, some assumptions about the invariance of the solution (replica symmetry) among replica are required. However, in complex systems with large number of order parameters the number of co-existing phases with small free energy difference is large. This leads to the replica symmetry breaking and corrugated distribution of the overlap between the states [38, 37]. Despite this formal inconsistency, in all cases when the replica solution can be compared with exact one, they give the same result.

Computation of the moments of the partition function generates replica variables $\{x_i^\alpha\}, \alpha = 1,2,...,n$, which need to be decoupled. This is done in [34] by introducing the auxiliary fields through the Hubbard-Stratonovich (HS) transformation and, as a result, a cumbersome integral equation for the spectral density was obtained. In [35] the authors implement a supersymmetrical method of calculation, which was shown to be equivalent to the replica trick and to give identical results. Importantly, at large p (see (1.24)) limit, the authors of [34, 35] recover the semi-circle distribution, however, in the sparse case the obtained results are difficult to analyze analytically. The following exponential tail of the spectral density was obtained in [34] from the leading non-perturbative contribution

$$\log \rho(\lambda) \sim -\lambda^2 \log\left(\lambda^2/ep\right) \tag{1.30}$$

which strikingly differs the sparse system from the dense one, where the spectral density has a finite support in the thermodynamic limit. Simple geometrical arguments [42] explaining the unbound density are as follows. Large eigenvalues correspond to the states (eigenvectors) localized on nodes with extremely high degree, also known as hubs. Local topology of the graph around the hub is the one of a star-graph with the core degree k and the largest eigenvalue $\sqrt{k}$. The last statement is reasonable in the limit of asymptotically large k, for which the environment around the hub becomes not important and weakly contributes to the spectrum. The probability of such hub to appear in the Erdös-Rényi ensemble is $\exp(-p)p^k/k!$. Thus, the amount of states around $\lambda = \sqrt{k}$ is

$$\Delta n(k) \approx \rho\left(\sqrt{k}\right)\left(\sqrt{k} - \sqrt{k-1}\right) \sim \frac{e^{-p}p^k}{k!} \tag{1.31}$$

Now, using the Stirling's approximation, one immediately arrives at the tail (1.30). To obtain full spectral density several approximate schemes, e.g. the effective medium approximation (EMA) or single defect approximation (SDA), were proposed [41, 42, 43]. However, they are not quiet accurate. The EMA scheme assumes that all nodes are equivalent and, thus, does not work well for sufficiently sparse graphs, due to their intrinsic inhomogeneity in local connectivity.

An important piece of study on sparse matrices spectra is provided by the numerical work [36], where the authors were one of the first to demonstrate that the spectrum of a large sparse matrix consists of a family of spikes, arranged in some regular pattern Fig.1.2(a),(b). In particular the singularity at $\lambda \to 0$ was reported to behave as

$$\rho(\lambda) \propto \frac{1}{|\lambda| \log\left(|\lambda|\right)^3} \tag{1.32}$$

provided that $p/N \to 0$. At the same time, at the edges of the spectrum exponential tails were present, confirming the analytical results of Rodgers and Bray. Despite the obvious difference with the semi-circle law, the authors have found that the level spacing distribution is much more universal even for the finite-d lattices (sparse case). The Wigner-Dyson universality remains as long as we are above the Anderson delocalization-localization transition point $p > p_q$. In order to catch the transition point p_q the authors suggest to use relative variance ratio $R = \langle \delta\rho(\lambda)^2\rangle / \langle \rho(\lambda)\rangle$, where $\langle \delta\rho(\lambda)^2\rangle$ is the variance of the energy spectrum. When the fluctuations exceed $R \geqslant 1/2$, the states become localized.

An intuitive and relatively robust approach to the spectra is provided by the cavity method [39]. Its success is relied on a fundamental observation that in the sparse case the graphs can be approximated by a set of trees-like subgraphs. Then if a subgraph is rooted at the node i, one can consider a graph, in which this node is removed $\mathcal{G}_J \to \mathcal{G}_J^i$. Marginal Gibbs-Boltzmann probability distributions with Hamiltonian (1.28) and partition function (1.27)

$$P_J(x) = \frac{1}{Z_J(z)} \exp\left(-\mathcal{H}_J(x,z)\right); \quad x = \{x_1,x_2,...,x_N\} \tag{1.33}$$

are factorised in the neighborhood of the removed node i (Bethe approximation)

$$P^i(x_{\partial i}) = \prod_{l\in\partial i} P^i(x_l) \tag{1.34}$$

where ∂i is the set of neighbors of the node i. In this approximation the set of cavity distributions $\{P^j(x_i)\}$ (1.34) obeys recursive equations that yield Gaussian solutions. Thus, the original distributions $\{P(x_i)\}$, are expressed through the cavity ones as

$$P(x_i) = \frac{\exp\left(-zx_i^2/2\right)}{Z_i} \int dx_{\partial i} \exp\left(x_i \sum_{l\in\partial i} A_{il}x_l\right) \prod_{l\in\partial i} P^i(x_l) \tag{1.35}$$

are also solved by Gaussian functions. As a result, the final expression for the average spectral density can be written as

$$\rho(\lambda) = \lim_{\varepsilon\to 0^+} \frac{1}{\pi N} \sum_{i=1}^{N} \Im\left[\langle x_i^2\rangle_P\right]_{z=\lambda-i\varepsilon} \tag{1.36}$$

with the variance $\langle x_i^2\rangle_P = \Delta_i(z)$ being the solution of a pair of coupled equations, for the cavity variance

$$\Delta_i^j(z) = \frac{1}{z - \sum_{l\in\partial i\setminus j} A_{il}^2 \Delta_l^i(z)} \tag{1.37}$$

and for the variance on the original graph

$$\Delta_i(z) = \frac{1}{z - \sum_{l\in\partial i} A_{il}^2 \Delta_l^i(z)} \tag{1.38}$$

Iterative solution for (1.37) and (1.38) allows to compute the original variance and, finally, the average spectral density, using (1.36). It is done in [39] for the Poisson ensemble of real symmetric matrices of size $N = 1000$ following the distribution

(1.24) with bimodal form of $h(x)$. The result was compared with the direct numerical diagonalization of the matrices. It was shown that the cavity solution is much closer to the exact solution than EMA and SDA. In particular, the spiky behaviour of the spectral density in the central region was reproduced for small ε, see Fig.1.2(c).

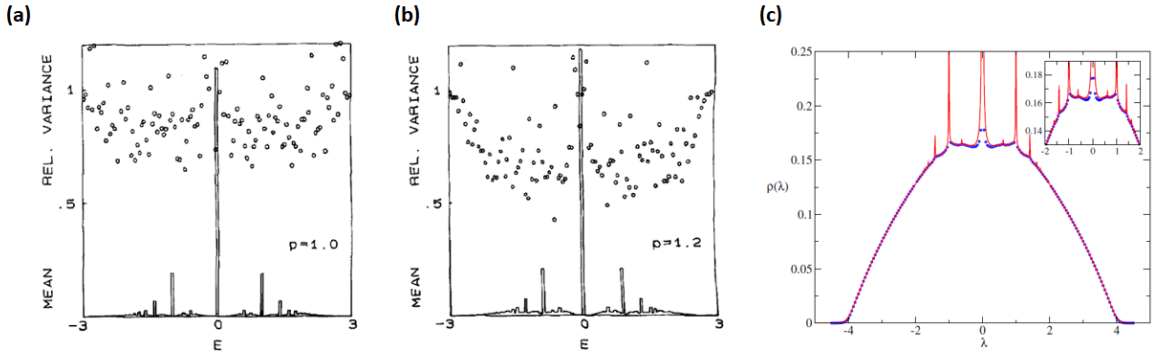

Figure 1.2 — (a), (b): Plots of the average density of states and the relative variance for a set of random realizations; numerical diagonalization of matrices with $N = 1000$ and $p = 1$ (a), $p = 1.2$ (b). The plots are taken from [36]. (c): Iterative solution of cavity equations for $\varepsilon = 0$ (blue circles) and $\varepsilon = 5 * 10^{-3}$ (solid red line) for the bimodal distribution $h(x)$, average connectivity $c = 3$ and $N = 1000$ [39]. The central region consists of a dense collection of δ functions (inset).

1.2.2 Ultrametricity in spectral density

As it was already noted in the abstract, the sparse graphs are internally equipped with the ultrametric. In ultrametric spaces the third requirement for the metric spaces (triangle inequality) is replaced with the *strong* triangle inequality, i.e. $d(x,z) \leqslant \max\{d(x,y), d(y,z)\}$. Due to this property, an ultrametric ball consists of an hierarchy of smaller balls and is isometric to a branching tree of hierarchically nested basins. The basins self-organize into a self-similar energy landscape, so that the ultrametric distance (the barrier) between any two basins can be projected to the distance along the graph to the root of their minimal common subtree.

Ultrametric spaces has been discussed in broad fields of physical, biological and social sciences, in particular, in the context of clustering of big data [44, 45]. Ultrametric spaces in the number theory are exemplified by fields $\mathcal{Q}_p$ of p-adic

numbers and rings $\mathcal{Q}_m$ of m-adic numbers, where m is a composite integer [46, 47, 48]. Low-temperature states of spin glass are organized according to the strong triangle inequality due to the large number of frustrations. Relaxation on the phase landscape takes place via tree-like branching into hierarchically nested domains. Since phase trajectories cannot explore the whole landscape, relaxation in a space with many metastable states is reminiscent to local optimization. Ultrametric organization of equilibrium states has been observed in random models with long-ranged correlations for spin glass, such as Sherrington-Kirkpatrick model [49, 50, 51, 52]. Complex landscapes of protein molecules are described by ultrametric ansatz following the idea that relaxation in proteins occurs locally, i.e. the time to leave the basin is much larger than the equilibration time within the local minimum [53, 54, 55, 56].

Ultrametric branching of basins on a complex landscape is isomorphic to the ensemble of large random trees embedded into a high-dimensional space. In [57] it was shown that a random tree generated recursively in a D-dimensional Euclidean space, $D \to \infty$, is ultrametric in the sense of variances of distance between a pair of points. The procedure consists in consecutive generation of new nodes of the tree in vicinity of the points, belonging to the edge of the tree, drawn from the normal distribution. More generally, it has been observed that a set of random points in a multi-dimensional Euclidean space can be associated with the sparse graph with, on average, ultrametric distances between the points.

In [58] the spectrum of ensembles of sparse graphs was shown to demonstrate ultrametric properties. The authors of [58] consider an ensemble of sparse Erdös-Rényi graphs, parameterized by an edge probability q, whose adjacency matrices are $N \times N$ symmetric matrices A_{ij}

$$A_{ij} = \begin{cases} 1, \text{with probability q} \\ 0, \text{with probability 1-q} \end{cases} \tag{1.39}$$

When $q = O(1)$ the matrices are dense and the average spectral density follows the Wigner semi-circle, while $q = c/N, c = O(1)$ corresponds to the sparse case. The transition point separating the two regimes is the percolation point $q_c = 1/N$, at which the percolation transition occurs. For $q \ll q_c$ there are only few edges and most of the nodes are isolated graph vertices. At $q > q_c$ the graph has a giant component. Below the percolation point $q < q_c$ the giant component is absent and almost all of the disconnected components are trees (see Fig.1.3(a)). This fundamental

topological property of ensembles of sparse graphs allows for a great simplification for the problem of the average spectral density. In particular, this is used in the cavity approach for the spectral density (Bethe approximation) [39]. Furthermore, it turns out that the main contribution to the universal shape of the spectral density comes from linear subgraphs, whose fraction is about 95%.

An important result of the work [58] is calculation of the mass distribution of various subgraphs in the vicinity of the percolation transition point. This is done by considering a kinetic linking of a disjoint graph with a constant rate and following the fraction of clusters of the certain size $c_k = N_k/N$

$$\frac{dc_k}{dt} = \frac{1}{2}\sum_{i+j=k}(ic_i)(jc_j) - kc_k; \quad c_k(0) = \delta_{k,1} \tag{1.40}$$

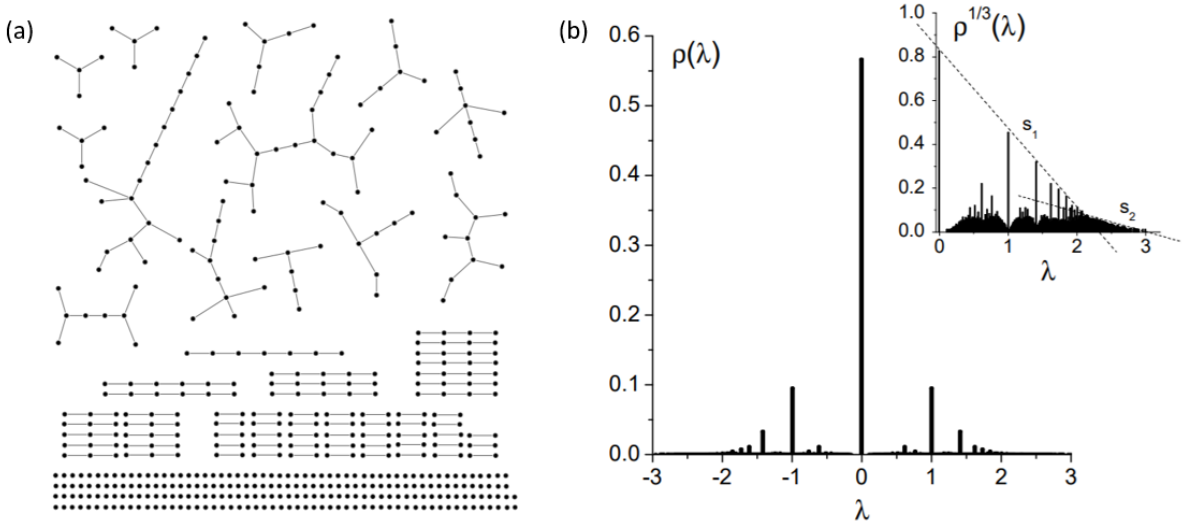

Figure 1.3 — (a): Typical sample of subgraphs in a realization of the sparse graph with size $N = 500$ slightly above the percolation transition point, $q = 2.0028 * 10^{-3}$. (b): Spectral density averaged over 1000 realizations with same parameters as in (a). The inset shows a fracture of the spectral density around $\lambda = 2$ and the tail, related with the density coming from branching trees. Illustrations are taken from [58].

The equation (1.40) states that a new cluster with size $k = i + j$ can form as a result of joining of two disjoint clusters of sizes i and j and can disappear if any of the nodes of the cluster gets linked with a node of other clusters. At time t the average connectivity of the graph is $q = t/N$. Thus, we can relate the sparsity of the graph with a particular mass distribution of the clusters. In the case of linear clusters the corresponding kinetic equation yields exactly exponential distribution

of the chains at the percolation point

$$Q_k = \frac{1}{2}\exp(-k) \tag{1.41}$$

Calculation of the spectra of sparse graphs in the approximation of only linear subgraphs with the exponential mass distribution (1.41) results in peculiar ultrametric profiles for the density. This profiles are found strikingly similar to the spectral densities obtained from the direct numerical matrix diagonalization and subsequent averaging in the ensemble of *all sparse graphs*. The hierarchical structure of the density is reminiscent of the Dedekind η-function, $\eta(z)$, which is defined in the upper half-plane $\Im z > 0$ as follows

$$\eta(z) = \exp\left(\frac{i\pi z}{12}\right)\prod_{n=0}^{\infty}\left(1-\exp\left(2i\pi n z\right)\right) \tag{1.42}$$

Namely, a conjecture that the spectral density of the linear chains distributed exponentially $Q_n \sim q^n$ for $q \to 1^-$ is isomorphic to $\sqrt{-\log|\eta(z)|}$ with $\Im(z) \to 0$ has been proposed in [58]. This conjecture was subsequently proven analytically in [61] and the spectral density for the ensemble of linear chains was shown to express through a so-called popcorn function, which can be regularized by the Dedekind η-function close to the real axis (see Chapter 2).

Despite the number fraction of linear subgraphs in the ensemble of all subgraphs close to the percolation point approaches 0.95, it turns out that mass fraction is only ≈ 0.65, meaning that a small number (about 5%) of disconnected subgraphs in the ensemble are essentially huge trees. However, as the numerical analysis shows [58], their contribution is limited to the edges of the spectral density. It is known that the leading eigenvalue of a tree with a maximal degree d is constrained from above by $2\sqrt{d-1}$, and the upper bound is achieved for d-regular trees. Linear chains can be formally defined as regular trees with $d = 2$, thus, the largest eigenvalue is $\lambda_{max} = 2$. Leading correction to the spectrum of linear chains in vicinity of the percolation point is coming from 3-branching trees, which produces additional density in the interval $[2, 2\sqrt{2}]$, see the inset in Fig.1.3(b). At the same time, they only slightly perturb the spectral density in the middle of the spectrum.

Furthermore, some large symmetric trees share the ultrametric spectra properties with the ensemble of linear subgraphs. In particular, spectrum of a binary tree and spectrum of a star-graph belong to the set of eigenvalues of all the composing linear subgraphs [59]. This is the direct corollary from the theorem of Rojo and Soto

[60], stating that the spectrum of a generalized Bethe tree consists of the eigenvalues of all principle submatrices of some tridiagonal matrix. In the case of a regular Bethe tree with degree $d = 3$ (binary tree) the corresponding tridiagonal matrix is nothing but an adjacency matrix of a linear chain scaled by $\sqrt{2}$; for regular trees of arbitrary degree the matrix and its eigenvalues get scaled by $\sqrt{d-1}$. Therefore, we see that the spectrum of a large Bethe tree possesses the self-averaging property: it is equivalent to the spectral density of the ensemble of exponentially distributed linear chains. Eigenvalues of any tree are calculated in [42] by means of recurrent relations for the characteristic polynomial, in a way similar to the cavity method. For non-regular generalized Bethe trees the correspondence with the ensemble of linear chains is absent. However, we note that the ultrametric peculiarities of the spectrum still persist due to the number-theoretic relations emerging along with the computation of multiplicities of the eigenvalues.

1.3 Community detection in networks

A complex system can be represented as a network with nodes responding to the agents and weights of edges proportional to the pairwise strength of interaction between the agents [62]. The network model allows to extract valuable information on hidden topological structure of the system. One of the most practically important examples of such structure is a mesoscopic organization of the agents into modules or communities [63, 64, 65]. However there are many different definitions of communities in networks, it is qualitatively understood as a group of nodes characterized by reinforced interactions with nodes of the same group, relatively to the other nodes [66]. Typically formation of communities is conjugated with a self-organization evoked by collective interactions of all the nodes in the network and, thus, is irreducible to action of independent agents of the complex system. On practice, the community detection in real networks allows to detect a hidden topological large-scale structure of the system, being an extremely hot topic in various technological [67, 68], biological [69, 70, 71, 72], social [73, 74, 75] and economical [76, 77] contexts.

A widely used approach in the community detection is a spectral decomposition of a linear operator defined on the network: the information on

communities is then encoded in several leading eigenvectors [82, 83]. It has been recently shown that all of the commonly used matrices (adjacency, Laplacian, modularity, non-backtracking) classify well the nodes as long as the network density is sufficient [84, 85]. In particular, the modularity operator has proven itself as one of the most efficient characteristic successfully detecting communities in stochastic networks of various nature [86, 65, 87, 75, 88, 89]. To extract deterministic communities from the fluctuations, the modularity score measures the community-wise weight difference between the observed network and the expected one in the framework of a null generative model, in which the individual degrees of nodes are kept invariant under randomization of the edges. Fixation of the degrees from the sample makes the modularity applicable to scale-free networks, a wide class, including most of the real-world networks [91].

1.3.1 Modularity functional

Modularity functional has been initially proposed by Mark Newman in his seminal paper [65] and, since then, has been vastly used for community detection in networks of various intrinsic nature [86, 65, 87, 75, 88, 89]. The modularity is a functional over a network partition into the n groups G_p, $p = 1, 2, ..., n$, which relates observed weights to expected weights in an annealed ensemble of graphs with fixed strength (or just degree for a non-weighted graph) k_i of each individual node i and $m = \frac{1}{2}\sum_i k_i$ being the total strength of the network. Formally, the modularity functional $Q \equiv Q\{G_1, G_2, ..., G_n\}$ over an arbitrary splitting into the groups G_p can be written as follows

$$Q = \frac{1}{4m}\sum_p \sum_{(i,j)\in G_p} \left(A_{ij} - \frac{k_i k_j}{2m}\right) \tag{1.43}$$

where A_{ij} is the adjacency matrix of the network and the expected matrix is proportional to the product of the respective strengths $\mathbf{k}^T\mathbf{k}$. Maximization of the functional (1.43) yields the "optimal" splitting, which corresponds to the intrinsic community structure provided the network is not very sparse and the communities are sufficiently resolved [84, 85]. Originally, the modularity score (1.43) has been proposed [65] for partition of the scale-free networks, in which the distribution of the degrees is power law and does not follow the Poisson statistics, typical for the

class of the Erdös-Rényi models. Indeed, using any vector $\mathbf{k}$ (for example, from the real data) as a parameter, one can take care of the scale-freeness of a real network. Ensemble of random graphs, produced by randomization of edges with conservation of the nodes strengths is known as the configuration model [92].

The brute-force maximization of the modularity functional is not typically necessary. There is a simplified *spectral approach* based on the leading eigenvectors of the modularity matrix. Suppose for simplicity there are only two communities in the network. One can assign a "spin direction" $s_i = \pm 1$ to each node of the network depending on the group this node belongs to and rewrite the modularity as a quadratic form in the spin space $\mathbf{s}$

$$Q = \frac{1}{4m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j + 1) = \frac{1}{4m} \mathbf{s}^T B \mathbf{s} \tag{1.44}$$

where $B = B_{ij}$ is the modularity operator

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{1.45}$$

and we have used in (1.44) the fact that the rows of the modularity matrix are summed to zero; this is obvious from (1.45). Applying the spectral decomposition of (1.44), one can make use of the principal component approximation, which is well justified for sufficiently resolved communities. In the case of two communities the optimal partition is encoded in the leading eigenvector of the matrix B

$$Q \approx (4m)^{-1} \lambda_1 (\mathbf{u}_1 \mathbf{s})^2 \tag{1.46}$$

where $\mathbf{u}_1$ is the normalized leading eigenvector and λ_1 is the corresponding (largest) eigenvalue. In order to maximize (1.46), one has to choose the most collinear spin vector $\mathbf{s}$ to the given $\mathbf{u}_1$. Therefore, the optimal solution $\mathbf{s}$ takes the value $s_i = +1$, if the corresponding component of $u_{1,i}$ is positive and $s_i = -1$, otherwise.

Note that though modularity maximization is one of the most natural approaches for community detection in scale-free graphs, it has a known limitation when the typical size of communities is small. Namely, it was shown that modularity fails to resolve true communities when their number is larger than $\sqrt{2m}$ [66]. Instead, the optimal solution by modularity yields larger groups and sufficiently small clusters do not get resolved. This maximal number of communities is known as the resolution parameter. Several tricks have been proposed to overcome this issue and effectively

increase the resolution. One of them [93, 94] is to incorporate a tunable parameter γ into the definition of modularity

$$Q(\gamma) = \frac{1}{4m} \sum_{i,j} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta_{g_i, g_j} \tag{1.47}$$

where g_i stands for the group of the node i. When the parameter $\gamma = 1$ one gets back to the traditional modularity (1.43). However, if $\gamma > 1$ one effectively places more expected weight to the nodes, forcing modularity to resolve smaller communities. In the other approach, one can introduce self-loops to the nodes [95]. In fact, this is equivalent to the increase of γ; tuning the weight of the self-loops, it is possible to regulate the size of the communities. Despite its phenomenological introduction the parameter γ has a clear physical interpretation that will be discussed in the next section.

1.3.2 Stochastic block model

Stochastic block model (SBM) is the simplest and most commonly used Erdös-Rényi graph model with explicit communities. In this model N nodes of a network are split into q different groups G_i, $i = 1,2,...,q$ and the edges between each pair of nodes are distributed independently with a probability that depends on the group labels ("colors") of respective nodes. It is said, there is a matrix of pairwise group probabilities $\Omega = \omega_{rt}$ with $r,t = 1,2,...,q$ and a randomly chosen pair of nodes (i,j) belonging to groups $i \in G_r, j \in G_t$ is linked by an edge with probability ω_{rt}. The corresponding entry in the adjacency matrix A_{ij} is 1 with probability ω_{rt} and 0 otherwise (or A_{ij} is a Poisson variable with $\lambda = \omega_{rt}$ for the weighted version of the model). Often communities can be considered identical (known as a planted stochastic block model); in this case,

$$\Omega_{rt} = \begin{cases} w_{in}, & r = t \\ w_{out}, & r \neq t \end{cases} \tag{1.48}$$

Furthermore, in the simplest scenario all the communities have equal size, $n_c = N/q$, and along with (1.48) they become completely equivalent in the space of parameters. Then, the average internal and external degrees of the nodes are

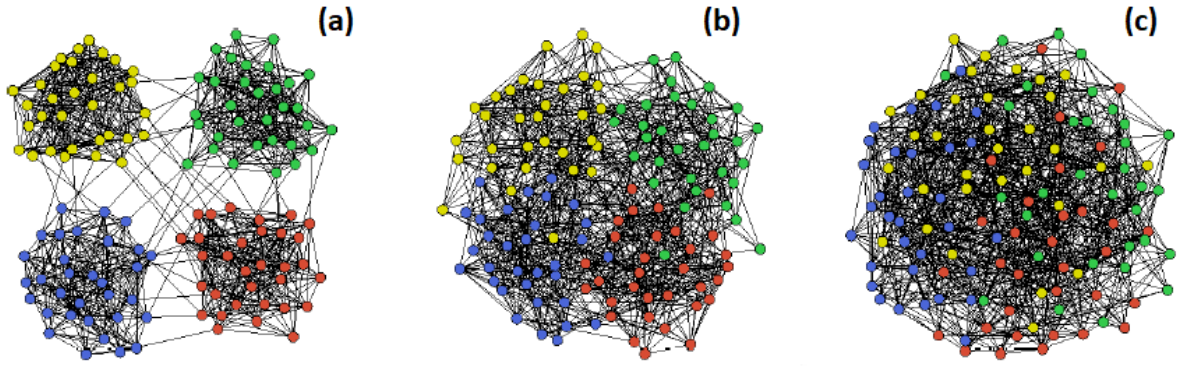

Figure 1.4 — Benchmark of Newman and Girvan. The three networks correspond to three different mean external degrees $\langle k_{out}\rangle = 1$ (a), $\langle k_{out}\rangle = 5$ (b) and $\langle k_{out}\rangle = 8$ (c). The total mean degree $\langle k\rangle = 16$, number of communities $q = 4$ and communities size $n_c = 32$ are fixed. In (c) the groups are indistinguishable by eye and it is hard for the most simplest methods to detect the true community structure (though, they are still detectable, see the next section). The networks are taken from [96].

$\langle k_{in}\rangle = w_{in}n_c$ and $\langle k_{out}\rangle = w_{out}n_c(q-1)$, correspondingly. The total mean degree is $\langle k\rangle = (w_{in} + w_{out}(q-1))\, n_c$. Since the beginning of the 21st century, one of the most popular SBM benchmarks in the literature has been a benchmark of Michelle Girvan and Mark Newman [97]. It fixes the number of communities to $q = 4$, the size $n_c = 32$ and the total average degree to $\langle k\rangle = 16$. Thus, different scenarios of communities resolution can be modelled by changing only one parameter, for example, $\langle k_{out}\rangle$, see Fig.1.4. The last figure (c) has almost mixed communities as we can see by eye. Therefore, a natural question arises: is it possible to resolve the true community structure, in principle? The answer is intrinsically probabilistic. Since the networks are stochastic, a fair question would compare the probability to sort an arbitrary node of the network correctly into its home group with $P_{rand} = 1/q$, i.e. the probability of the correct sorting with help of a $q/2$-dimensional die. This brings us to the idea of statistical inference of the optimal network partition: the optimal is the one that maximizes the likelihood that what we see in the experiment is SBM with a particular set of parameters.

Recently it has been shown [90] that maximization of the generalized modularity functional (1.47) is equivalent to the statistical inference of communities in the framework of the degree corrected version of the planted stochastic block model. Degree corrected SBM ensemble corresponds to the configuration model with fixed strengths $\{k_i\}$, i.e. the expected weight of the edge (i, j) is a product

of the SBM probabilities (1.48) and $P_{ij} = \frac{k_i k_j}{2m}$. Statistical inference approach is formulated as follows. Suppose that an adjacency matrix A from the Poisson degree corrected SBM ensemble is observed as a realization. With given A, what are the optimal parameters of the underlying stochastic model? The statistical weight of A conditioned on the cluster probability matrix Ω, degrees $\{k_i\}$ and group labels of the nodes $\{g_i\}$, reads

$$W(A|\,\Omega,\{k_i\},\{g_i\}) = \prod_{i<j} \frac{\left(P_{ij}\;\omega_{g_i g_j}\right)^{A_{ij}}}{A_{ij}!} \exp\left(-P_{ij}\;\omega_{g_i g_j}\right) \tag{1.49}$$

where the product runs over all pairs of nodes in the network. Since there are no self-edges in the network, all the diagonal elements of the matrix A are zeros and we do not include them into the product (1.49). The corresponding partition entropy of the polymer SBM is

$$\log W(A|\,\Omega,\{k_i\},\{g_i\}) = \sum_{i<j} \left(A_{ij} \log \omega_{g_i g_j} - P_{ij}\omega_{g_i g_j}\right) \tag{1.50}$$

where we have omitted all the constant terms independent of the partition. For identical communities (see (1.48)), we get

$$\begin{cases} \omega_{g_i g_j} = w_{out} + \delta_{g_i g_j}\,(w_{in} - w_{out}) \\ \log \omega_{g_i g_j} = \log w_{out} + \delta_{g_i g_j}\,(\log w_{in} - \log w_{out}) \end{cases} \tag{1.51}$$

Taking into account (1.51) and omitting again all irrelevant constant terms, we arrive at the final expression for the entropy (1.50)

$$T \log W(A|\,\Omega,\{k_i\},\{g_i\}) = \sum_{i<j} \left(A_{ij} - \gamma \frac{k_i k_j}{2m}\right) \delta_{g_i g_j} \tag{1.52}$$

where $T = (\log w_{in} - \log w_{out})^{-1}$ has the sense of temperature and

$$\gamma = \frac{w_{in} - w_{out}}{\log w_{in} - \log w_{out}} \tag{1.53}$$

is a parameter describing the cluster probabilities inherited from the initial definition of the stochastic blocks. We see from (1.52) that maximization of the network entropy with respect to the partition into communities is equivalent to maximization of the generalized modularity (1.47). Furthermore, the parameter γ is connected with the properties of the communities. Let us define $w_{in} = h w_{out}$, where for assortative

communities $h > 1$. Clearly, one of the parameters of the model can be chosen arbitrary, since it just rescales the overall average density of the graph; without loss of generality we may choose $w_{out} = 1$. Then one obtains $h = \exp(1/T)$ and the following dependency of γ on the effective temperature of the system

$$\gamma(T) = T\left(\exp(1/T) - 1\right) \tag{1.54}$$

Physically, the increase of the effective temperature is associated with a weaker structuring into the communities. At $T \to \infty$ we recover the classical modularity case with $\gamma = 1$, which, thus, can be called the "weak modularity". On the contrary, when community structure is sufficiently well pronounced ("strong modularity"), it is more reliable to make the parameter γ free and look for the best γ_{opt}. This can be done using the following renormalization scheme: (i) one takes a trial value of the parameter, e.g. $\gamma_0 = 1$; (ii) maximizes the modularity functional with this $Q(\gamma_0)$; (iii) calculates the sample mean for the pairwise strength inside the obtained communities w_{in} and outside w_{out}; (iv) computes the corrected value of γ_1 according to (1.53); (v) repeats the procedure until convergence $\gamma_\infty = \gamma_{opt}$. In practice, several steps of the iteration is sufficient to achieve the convergence, if the clustering network is fit by the SBM.

Note that for the regular graph $k_i \equiv k$ and in the "weak modularity"regime the functional (1.47) is equivalent to the Laplacian. For non-regular graphs with inhomogeneous distribution of degrees a normalized version of the Laplacian is frequently used, e.g. the symmetric normalized Laplacian, $L^{sym} = D^{-1/2}(D - A)D^{1/2}$, where D is the degree matrix. These operators are considered as classical or traditional in the literature, because they have been widely used for the purposes of clustering (often, spectral) of sufficiently dense networks. However, all of them notably fail when the total density of the network is drastically reduced up to the regime, when the network becomes sparse.

1.3.3 Detectability transition

Detectability of communities in random networks is formulated in the probabilistic sense and in the thermodynamic limit. Suppose one has an ensemble of random networks of total size $N \gg 1$ with q equivalent communities. Then the

communities are called detectable if there is an algorithm that correctly classifies more than $1/q$ nodes in a typical realization of the network. In the $N \to \infty$ limit there is a sharp transition at the edge of the detectability regime. A naive conjecture would be to say that communities are detectable as long as $w_{in} > w_{out}$ or

$$\langle k_{in} \rangle > \frac{\langle k_{out} \rangle}{q-1}, \tag{1.55}$$

i.e. already a very weak tendency towards cluster formation can be successfully exploited by a hypothetic algorithm. It turns out that this is a correct transition point for dense networks, when w_{in} and w_{out} stay constant with increase of the system size [85]. Otherwise, the network is sparse and (1.55) does not provide us the transition anymore. It has been shown that the detectability transition for sparse networks occurs much earlier

$$\langle k_{in} \rangle - \frac{\langle k_{out} \rangle}{q-1} > \sqrt{\langle k_{in} \rangle + \langle k_{out} \rangle} \tag{1.56}$$

In other words, although the groups in the network are still treated as communities, according to the definition (1.48), they are not detectable by any algorithm or the community detection is exponentially hard.

Real-world networks are always finite, thus, there is no proper detectability transition for them; instead, there is usually a smooth crossover from the range of parameters, where the detection is possible, to the range, where it is frequently hardly possible. For the benchmark of Girvan-Newman illustrated in the Fig.1.4 one can readily calculate where the dense-graph transition point is, $\langle k_{out} \rangle = k^{dense} = 12$, while the sparse transition occurs at $\langle k_{out} \rangle = k^{sparse} = 9$. Since the total size of the network is rather small, $N = 128$, these numbers are not quiet appropriate for this benchmark. Numerical analysis of different most powerful community detection methods shows that the true crossover takes place somewhere in between 9 and 12, though, it seems to be rather close to 12 [66].

Since in the sparse case the probability matrix Ω is $O(1/N)$, it is useful to switch to the rescaled variables $c_{in/out} = N w_{in/out}$. We shall also introduce a conjugated variable $c = c_{in}/q + c_{out}(q-1)/q$, which equals to $\langle k_{in} \rangle + \langle k_{out} \rangle$, the mean number of edges per node. In the case of two groups, $c = (c_{in} + c_{out})/2$. In the rescaled variables the detectability transition (1.56) simply reads

$$c_{in} - c_{out} > q\sqrt{c} \tag{1.57}$$

From the condition (1.57) the idea of the spectral clustering can be easily understood, when there are two communities in the network, $q = 2$. For a dense Erdös-Rényi graph $c = O(N)$ the term on the right hand-side of (1.57) is the edge of the Wigner semi-circle, $\lambda_c = 2\sqrt{c}$. The first eigenvector of the adjacency matrix sorts the nodes to their degree, while the second one correlates with the true assignment to communities. For SBM the position of the second eigenvalue is well-known

$$\lambda_2 = \frac{c_{in} - c_{out}}{2} + \frac{c_{in} + c_{out}}{c_{in} - c_{out}} \tag{1.58}$$

Therefore, the detectability condition (1.57) is simply equivalent to $\lambda_2 > \lambda_c$. In other words, the communities are resolved as long as the second eigenvalue of the adjacency matrix is separated by a non-zero gap from the boundary of the Wigner's disk. Condition $\lambda_2 = \lambda_c$ is equivalent to $w_{in} = w_{out}$ for a dense SBM in the thermodynamic limit.

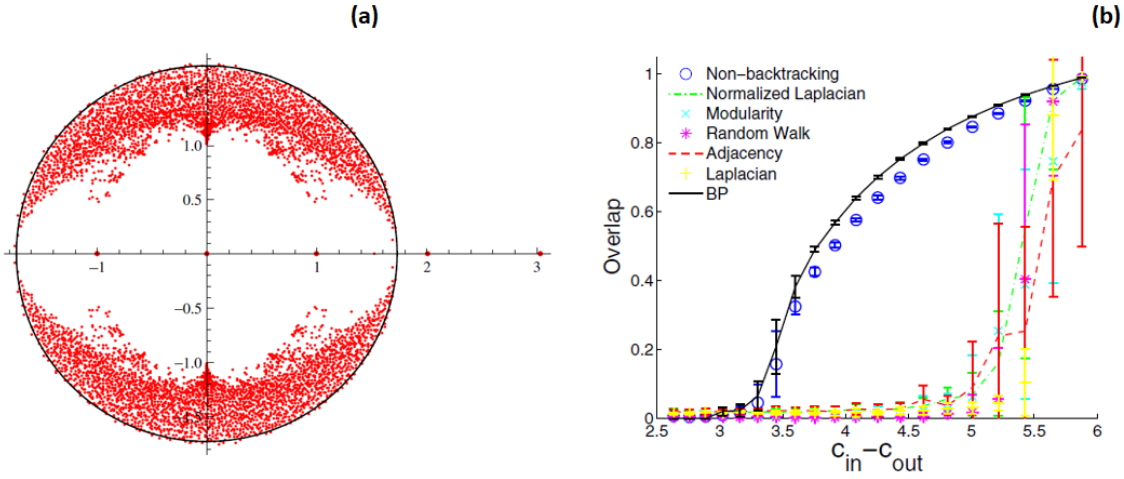

Figure 1.5 — (a) Non-backtracking spectrum of a SBM graph with two communities, $c = 3$ and $c_{in} - c_{out} = 4$. The leading eigenvalue $\lambda_1 = c$, the second one $\lambda_2 = (c_{in} - c_{out})/2$ and the boundary of the disk positions at $\lambda_c = \sqrt{c}$; (b): For the same parameters of the network as in (a), the overlap of the true nodes assignment with the spectral network partition performed by different traditional operators, non-backtracking operator as well as by the belief propagation (BP) method. The detectability transitions occurs very close to the theoretical threshold $2\sqrt{c} \approx 3.46$. Illustrations are taken from [98].

However, in the sparse case c does not grow with N and the spectral density is not a semi-circle anymore. A particular position of the boundary, separating the isolated part of the spectrum from the bulk, is vague. As we have discussed

in the previous section the spectrum of sparse Erdös-Rényi graphs is unbound in the thermodynamic limit and is populated by eigenvalues, which represent star-like subgraphs with arbitrary large degree, appearing with exponentially vanishing probability. In case of real-world sparse graphs with scale-free distribution of degrees $P(k) \sim k^{-\gamma}$ the probability of these subgraphs (hubs) is even larger and the spectral density should demonstrate a heavy tail with exponent $2\gamma - 1$. Thus, in the sparse regime the leading eigenvectors of the adjacency matrix get concentrated on hubs and not on true communities. In fact, all the traditional operators (adjacency, modularity, Laplacian) fail to resolve the clusters above the theoretical threshold (1.57). All such operators are associated with symmetric random walks on graph: the adjacency matrix is the transfer matrix of the walker $t \to t + 1$; normalized Laplacian is a transition operator for the probability flow of the walk; modularity operator evaluates the flow above the expected in the configuration graph model. Random walks entropically favor the hubs and localize on them, throwing large eigenvalues to the isolated part of the spectrum.

To overcome this difficulty, it was proposed to exploit the spectrum of the Hashimoto matrix $\mathbf{B}$, which is a transfer matrix of non-backtracking walks on a graph [99]. It is defined on the edges of the directed graph, $i \to j, k \to l$, as follows

$$\mathbf{B}_{i\to j,k\to l} = \delta_{il}(1 - \delta_{jk}) \tag{1.59}$$

It is seen from (1.59) that the non-backtracking operator prohibits immediate returns to the point which a walker has visited at the previous step, thus, it avoids hubs. As a result, even in the sparse case the leading eigenvectors ignore hubs (in fact, they ignore all hanging trees in the network, take a look at the "reluctant non-backtracking"modification as a way around [100]). Since matrix $\mathbf{B}$ is non-symmetric, its spectrum is complex, and has a clear demarkation between the bulk and the isolated part. For SBM graphs the bulk density of $\mathbf{B}$ is constrained within a circle of radius $\lambda_c = \sqrt{c}$. Isolated eigenvalues lie on the real axis to the right from the circle's boundary, see Fig.1.5(a). The second eigenvalue of the Hashimito operator is

$$\lambda_2 = \frac{c_{in} - c_{out}}{2} > \lambda_c = \sqrt{c} \tag{1.60}$$

This spectral condition immediately brings us to (1.57). Therefore, making use of the leading eigenvectors of B for the network partitioning results in detection of communities all the way down to the theoretical limit (1.57) for sparse graphs. Since the second eigenvector $u^{(2)}_{i\to j}$ of the non-backtracking operator, in contrast to

the adjacency or modularity, is defined on directed edges of the network, in case of two communities one needs to evaluate the spin variables $g_i = \pm 1$ in order to classify the nodes. The contribution to the i-th node g_i comes from the flow along all directed edges pointing to i. Thus, in order to switch from edges to nodes, one needs to evaluate the sign of the sum $v_i = \sum_j A_{ij} u^{(2)}_{j\to i}$ and to assign the node i accordingly, $g_i = sign(v_i)$.

The matrix B can be also derived as a result of linearization of the update equations for belief propagation (BP) [98]. In Fig.1.5(b) we provide a plot from [98], showing the performance of different spectral algorithms based on linear operators, as well as of the BP method. It is seen that the spectral clustering based on the non-backtracing does the job almost perfectly, overlapping with the true assignment similarly to the BP, while all traditional operators break down well above the theoretical threshold and perform not better than chance in a wide range of $c_{in} - c_{out}$. In [101] M. Newman has suggested a modified operator, which conserves non-backtracking probability flow at each step of the walker. Newman's non-backtracking enjoys slightly better behaving spectral boundary and more less spiky eigenvectors. We note that such neutralization towards the expected flow becomes crucial in the case of SBM with inhomogeneous background, such as chromatin graphs equipped with intrinsic linear memory (see Chapter 7).

References

[1] E. Wigner, Proc. Cambridge Philos. Soc. 47, 790 (1951).

[2] F. J. Dyson, J. Math. Phys. 3, 140 (1962).

[3] C.E. Porter and N. Rosenzweig, Ann. Acad. Sci. Fennicae, Serie A VI Physica 6, 44 (1960).

[4] M. L. Mehta, Random Matrices, 2nd Edition, Academic Press (1991).

[5] G. Akemann, J. Baik, P. Di Francesco, eds.The Oxford handbook of random matrix theory, Oxford Univ. Press, Oxford, (2011).

[6] C.E. Porter, Statistical Theories of Spectra: Fluctuations (Academic Press, New York, 1965).

[7] H. Weyl, Classical Groups (Princeton Univ. Press, Princeton, 1946).

[8] F.G. Tricomi, Integral Equations (Dover publications, 1985).

[9] S.N. Majumdar and G. Schehr, Journal of Statistical Mechanics: Theory and Experiment 2014.1, P01012 (2014).

[10] L. Erdös, Russ. Math. Surv. 66, 507 (2011).

[11] V. A. Marchenko, L. A. Pastur, Math. USSR-Sb. 1, 457 (1967).

[12] A. J. Bray, D. S. Dean, Phys. Rev. Lett. 98, 150201 (2007).

[13] Y. V. Fyodorov, I. Williams, J. Stat. Phys., 129, 1081 (2007).

[14] L. Susskind, Universe or multiverse, 247-266, (2003)

[15] D. S. Dean, S. N. Majumdar, Phys. Rev. E 77, 041108 (2008).

[16] R. M. May, Nature 238, 413 (1972).

[17] E. J. Gumbel, Statistics of Extremes, Columbia University Press, (1958).

[18] C. A. Tracy, H. Widom, Commun. Math. Phys. 159, 151 (1994).

[19] C. A. Tracy, H. Widom, Commun. Math. Phys. 177, 727 (1996).

[20] I. M. Johnstone, Ann. Stat. 29, 295 (2001).

[21] A. Soshnikov, J. Stat. Phys. 108, 1033 (2002).

[22] K. Johansson, Commun. Math. Phys. 209, 437 (2000).

[23] P. J. Forrester, S. N. Majumdar, G. Schehr, Nucl. Phys. B 844, 500 (2011).

[24] J. Baik, P. Deift, K. Johansson, J. Am. Math. Soc. 12, 1119 (1999).

[25] V. Dotsenko, Europhys. Lett. 90, 20003 (2010).

[26] V. Dotsenko, Journal of Statistical Mechanics: Theory and Experiment 2012.11, P11014 (2012).

[27] D. Ioffe, S. Shlosman, and Y. Velenik, Commun. Math. Phys. 336, 905 (2015).

[28] B. Derrida, M. R. Evans, V. Hakim, and V. Pasquier, J. Phys. A 26, 1493 (1993).

[29] S. N. Majumdar and S. Nechaev, Phys. Rev. E 72, 020901(R) (2005).

[30] M. Prähofer, H. Spohn, Phys. Rev. Lett. 84, 4882 (2000).

[31] D. S. Dean, S. N. Majumdar, Phys. Rev. Lett. 97, 160201 (2006).

[32] S. N. Majumdar, M. Vergassola, Phys. Rev. Lett. 102, 060601 (2009).

[33] Ravasz, Erzsébet, and Albert-László Barabási, Physical Review E 67.2, 026112 (2003).

[34] G. J. Rodgers and A. J. Bray, Phys. Rev. B 37, 3557 (1988).

[35] Y. V. Fyodorov and A. D. Mirlin, J. Phys. A 24, 2219 (1991).

[36] S. N. Evangelou and E. N. Economou, Phys. Rev. Lett. 68, 361 (1992).

[37] Parisi, Giorgio. "The physical meaning of replica symmetry breaking."arXiv preprint cond-mat/0205387 (2002).

[38] Dotsenko, Viktor. Introduction to the replica theory of disordered statistical systems. Vol. 4. Cambridge University Press, (2005).

[39] T. Rogers, I. P. Castillo, R. Kühn, and K. Takeda, Phys. Rev. E 78, 031116 (2008).

[40] S. F. Edwards and R. C. Jones, J. Phys. A 9, 1595 (1976).

[41] G. Biroli and R. Monasson, J. Phys. A 32, L255 (1999).

[42] G. Semerjian and L. F. Cugliandolo, J. Phys. A 35, 4837 (2002).

[43] T. Nagao and T. Tanaka, J. Phys. A 40, 4973 (2007).

[44] P. Hall, J.S. Marron and A. Neeman, Journal of the Royal Statistical Society B, 67, 424-444, (2005).

[45] F. Murtagh Proc. 2nd International Conference on p-Adic Mathematical Physics, American Institute of Physics, 151-161, (2006).

[46] Schikhof W.H. Ultrametric Calculus. An Introduction to p-adic Analysis, Cambridge University Press, Cambridge, (1984).

[47] Vladimirov V. S., Volovich I. V., Zelenov E. I., p-Adic Analysis and Mathematical Physics, Singapure: World Scientific Publishing, (1994).

[48] Dolgopolov M. V., Zubarev A. P., p-Adic Numbers, Ultrametric Analysis, and Applications, 3 (2011).

[49] Parisi, Giorgio, Journal of Physics A: Mathematical and General 13.4, L115 (1980).

[50] Sherrington, David, and Scott Kirkpatrick, Physical Review Letters 35.26, 1792 (1975).

[51] Katzgraber, Helmut G., and Alexander K. Hartmann, Physical Review Letters 102.3, 037207 (2009).

[52] Leuzzi, Luca, Journal of Physics A: Mathematical and General 32.8, 1417 (1999).

[53] Frauenfelder H., "Complexity in proteins Nature Struct. Biol., V. 2, 821-823 (1995).

[54] Becker O. K., Karplus M., J. Chem. Phys., 106, 1495-1517 (1997).

[55] Avetisov, V. A., Bikulov, A. Kh., Biophysical Reviews and Letters, 3, 387-396 (2008).

[56] Avetisov V. A, Bikulov A. H., Zubarev A. P., J. Phys. A: Math and Theor., V. 42, 85005-85021 (2009).

[57] A. P. Zubarev, p-Adic Numbers, Ultrametric Analysis and Applications, 6, 150 (2014).

[58] Avetisov, V., P. L. Krapivsky, and S. Nechaev, Journal of Physics A: Mathematical and Theoretical 49.3, 035101 (2015).

[59] V Kovaleva, Yu Maximov, S Nechaev and O Valba, J. Stat. Mech. 073402 (2017).

[60] Rojo O and Soto R., Linear Algebr. Appl. Elsevier 403 97 (2005).

[61] Nechaev, S., and K. Polovnikov, Physics-Uspekhi 61.1, 99 (2018).

[62] Newman, Mark Ed, Albert-László Ed Barabási, and Duncan J. Watts. The structure and dynamics of networks. Princeton university press, 2006.

[63] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004). Defining and identifying communities in networks. Proceedings of the national academy of sciences, 101(9), 2658-2663.

[64] Lancichinetti, A., Radicchi, F., Ramasco, J. J., Fortunato, S. (2011). Finding statistically significant communities in networks. PloS one, 6(4), e18961.

[65] Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical review E, 74(3), 036104.

[66] Fortunato, S., Hric, D. (2016). Community detection in networks: A user guide. Physics reports, 659, 1-44.

[67] Albert, R., Jeong, H., Barabási, A. L. (1999). Internet: Diameter of the world-wide web. Nature, 401(6749), 130.

[68] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the web. Computer networks, 33(1-6), 309-320.

[69] Dekker, J., Marti-Renom, M. A., Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics, 14(6), 390.

[70] Pastor-Satorras, R., Vespignani, A. (2001). Epidemic spreading in scale-free networks. Physical review letters, 86(14), 3200.

[71] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabási, A. L. (2000). The large-scale organization of metabolic networks. Nature, 407(6804), 651.

[72] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. science, 297(5586), 1551-1555.

[73] Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B-Condensed Matter and Complex Systems, 4(2), 131-134.

[74] Newman, M. E. (2001). The structure of scientific collaboration networks. Proceedings of the national academy of sciences, 98(2), 404-409.

[75] Chen, J., Zaïane, O. R., Goebel, R. (2009, April). Detecting communities in social networks using max-min modularity. In Proceedings of the 2009 SIAM international conference on data mining (pp. 978-989). Society for Industrial and Applied Mathematics.

[76] Piccardi, C., Calatroni, L., Bertoni, F. (2010). Communities in Italian corporate networks. Physica A: Statistical Mechanics and its Applications, 389(22), 5247-5258.

[77] Corrado, R., Zollo, M. (2006). Small worlds evolving: governance reforms, privatizations, and ownership networks in Italy. Industrial and Corporate Change, 15(2), 319-352.

[78] Borgatti, S. P., Everett, M. G. (2000). Models of core/periphery structures. Social networks, 21(4), 375-395.

[79] Newman, M. E., Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.

[80] Holme, P. (2005). Core-periphery organization of complex networks. Physical Review E, 72(4), 046111.

[81] Rombach, M. P., Porter, M. A., Fowler, J. H., Mucha, P. J. (2014). Core-periphery structure in networks. SIAM Journal on Applied mathematics, 74(1), 167-190.

[82] Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17(4), 395-416.

[83] Krivelevich, M., Sudakov, B. (2003). The largest eigenvalue of sparse random graphs. Combinatorics, Probability and Computing, 12(1), 61-72.

[84] Nadakuditi, R. R., Newman, M. E. (2012). Graph spectra and the detectability of community structure in networks. Physical review letters, 108(18), 188701.

[85] Decelle, A., Krzakala, F., Moore, C., Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Physical Review E, 84(6), 066106.

[86] Norton, H. K., Emerson, D. J., Huang, H., Kim, J., Titus, K. R., Gu, S., et al. (2018). Detecting hierarchical genome folding with network modularity. Nature methods, 15(2), 119.

[87] Grilli, J., Rogers, T., Allesina, S. (2016). Modularity and stability in ecological communities. Nature communications, 7, 12031.

[88] Guimerà, R., Stouffer, D. B., Sales-Pardo, M., Leicht, E. A., Newman, M. E. J., Amaral, L. A. (2010). Origin of compartmentalization in food webs. Ecology, 91(10), 2941-2951.

[89] Sales-Pardo, M., Guimera, R., Moreira, A. A., Amaral, L. A. N. (2007). Extracting the hierarchical organization of complex systems. Proceedings of the National Academy of Sciences, 104(39), 15224-15229.

[90] Newman, M. E. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. Physical Review E, 94(5), 052315.

[91] Barabási, A. L., Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

[92] Molloy, M., and Bruce R. A critical point for random graphs with a given degree sequence. Random structures & algorithms 6.2-3 (1995): 161-180.

[93] Reichardt, Jörg, and Stefan Bornholdt. "Statistical mechanics of community detection."Physical review E 74.1 (2006): 016110.

[94] Arenas, Alex, Alberto Fernandez, and Sergio Gomez. "Analysis of the structure of complex networks at different resolution levels."New journal of physics 10.5 (2008): 053039.

[95] Moscalets, Alexander P., Leonid I. Nazarov, and Mikhail V. Tamm. "Towards a robust algorithm to determine topological domains from colocalization data."arXiv preprint arXiv:1601.01253 (2016).

[96] Guimera, Roger, and Luis A. Nunes Amaral. "Functional cartography of complex metabolic networks."nature 433.7028 (2005): 895-900.

[97] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks."Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.

[98] Krzakala, Florent, et al. "Spectral redemption in clustering sparse networks."Proceedings of the National Academy of Sciences 110.52 (2013): 20935-20940.

[99] Hashimoto, Ki-ichiro. "Zeta functions of finite graphs and representations of p-adic groups."Automorphic forms and geometry of arithmetic varieties. Academic Press, 1989. 211-280.

[100] Singh, Abhinav, and Mark D. Humphries. "Finding communities in sparse networks."Scientific reports 5.1 (2015): 1-7.

[101] Newman, M. E. J. "Spectral community detection in sparse networks."arXiv preprint arXiv:1308.6494 (2013).

2. Rare-event statistics and modular invariance

Introduction

Here, based on the "Euclid orchard"construction, we provide simple geometric arguments that explain the equivalence of various distributions resulting from the rare-event statistics. In particular, we discuss the number-theoretic properties of the spectral density of exponentially weighted ensemble of linear polymer chains. It can be shown that the eigenvalue statistics of corresponding adjacency matrices in the sparse regime demonstrates peculiar hierarchical structure that is described by the popcorn (Thomae) function, discontinuous in the dense set of rational numbers. Moreover, at the edges the spectral density exhibits the Lifshitz tails, reminiscent of the 1D Anderson localization. Finally, based on the Dedekind η-function, we suggest a continuous approximation of the popcorn function and demonstrate that the hierarchical ultrametric structure of the popcorn-like distributions is ultimately connected with hidden $SL(2,Z)$ modular symmetry.

Contribution

I have established the connection of the spectral density of ensemble of linear chains with the popcorn function and, using modular properties of the Eisenstein series, have approximated it by the Dedekind η-function near the real axis. I have demonstrated the ubiquity of popcorn-like distributions using simple toy statistical models, based on the "Euclid orchard"construction.

Rare-event statistics and modular invariance

S. Nechaev[1,2], and K. Polovnikov[3,4]
[1] *Interdisciplinary Scientific Center Poncelet (ISCP), Bolshoy Vlasyevskiy Pereulok 11, 119002, Moscow, Russia*
[2] *P.N. Lebedev Physical Institute RAS, 119991, Moscow, Russia*
[3] *Skolkovo Institute of Science and Technology, 143005 Skolkovo, Russia*
[4] *Physics Department, M.V. Lomonosov Moscow State University, 119992 Moscow, Russia*

Here we provide simple geometric arguments, based on the "Euclid orchard" construction, that explain the equivalence of various distributions, resulting from the rare-event statistics. In particular, we discuss the number-theoretic properties of the spectral density of exponentially weighted ensemble of linear polymer chains. It can be shown that the eigenvalue statistics of corresponding adjacency matrices in the sparse regime demonstrates peculiar hierarchical structure that is described by the popcorn (Thomae) function, discontinuous in the dense set of rational numbers. Moreover, at the edges the spectral density exhibits the Lifshitz tails, reminiscent of the 1D Anderson localization. Finally, we suggest a continuous approximation of the popcorn function, based on the Dedekind η-function, and demonstrate that the hierarchical ultrametric structure of the popcorn-like distributions is ultimately connected with hidden $SL(2,Z)$ modular symmetry.

I. INTRODUCTION

The so-called "popcorn function" [1], $g(x)$, known also as the Thomae function, has also many other names: the raindrop function, the countable cloud function, the modified Dirichlet function, the ruler function, etc. It is one of the simplest number-theoretic functions possessing nontrivial fractal structure (another famous example is the everywhere continuous but never differentiable Weierstrass function). The popcorn function is defined on the open interval $x \in (0,1)$ according to the following rule:

$$g(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q}, \text{ and } (p,q) \text{ are coprime} \\ 0 & \text{if } x \text{ is irrational} \end{cases} \qquad (1)$$

The popcorn function g is discontinuous at every rational point because irrationals come infinitely close to any rational number, while g vanishes at all irrationals. At the same time, g is continuous at irrationals.

One of the most beautiful incarnations of the popcorn function arises in a so-called "Euclid orchard" representation. Consider an orchard of trees of *unit hights* located at every point (an, am) of the two-dimensional square lattice, where n and m are nonnegative integers defining the lattice, and a is the lattice spacing, $a = 1/\sqrt{2}$. Suppose we stay on the line $n = 1 - m$ between the points $A(0,a)$ and $B(a,0)$, and observe the orchard grown in the first quadrant along the rays emitted from the origin $(0,0)$ – see the Fig. 1.

Along these rays we see only the first open tree with coprime coordinates, $M(ap,aq)$, while all other trees are shadowed. Introduce the auxiliary coordinate basis (x,y) with the axis x along the segment AB and y normal to the orchard's plane (as shown in the Fig. 1a). We set the origin of the x axis at the point A, then the point B has the coordinate $x = 1$. It is a nice school geometric problem to establish that: (i) having the focus located at the origin, the tree at the point $M(ap,aq)$ is spotted at the place $x = \frac{p}{p+q}$, (ii) the *visible* height of this tree is $\frac{1}{p+q}$. In other words, the "visibility diagram" of such a lattice orchard is exactly the popcorn function.

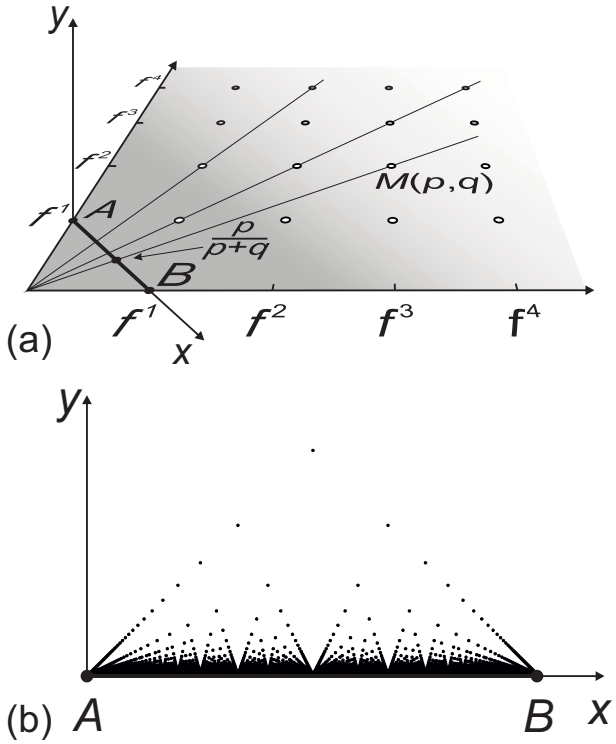

Figure 1: (a) Construction of the Euclid Orchard; (b) Popcorn (Thomae) function.

The popcorn correspondence $\frac{p}{q} \to \frac{1}{q}$ arises in the Euclid orchard problem as a purely geometrical result. However, the same function has appeared as a probability distribution in a plethora of biophysical and fundamen-

tal problems, such as the distribution of quotients of reads in DNA sequencing experiment [2], quantum $1/f$ noise and Frenel-Landau shift [3], interactions of non-relativistic ideal anyons with rational statistics parameter in the magnetic gauge approach [4], or frequency of specific subgraphs counting in the protein-protein network of a *Drosophilla* [5]. Though the extent of similarity with the original popcorn function could vary, and experimental profiles may drastically depend on peculiarities of each particular physical system, a general probabilistic scheme resulting in the popcorn-type manifestation of number-theoretic behavior in nature, definitely survives.

Suppose two random integers, ϕ and ψ, are taken independently from a discrete probability distribution, $Q_n = f^n$, where $f = 1 - \varepsilon > 0$ is a "damping factor". If $\gcd(p, q) = 1$, then the combination $\nu = \frac{\phi}{\phi+\psi}$ has the popcorn-like distribution $P(\nu)$ in the asymptotic limit $\varepsilon \ll 1$:

$$P\left(\nu = \frac{p}{p+q}\right) = \sum_{n=1}^{\infty} f^{n(p+q)} = \frac{(1-\varepsilon)^{p+q}}{1-(1-\varepsilon)^{p+q}} \approx \frac{1}{\varepsilon(p+q)} \quad (2)$$

The formal scheme above can be understood on the basis of the Euclid orchard construction, if one would consider a $1+1$ directed walker on the lattice (see Fig. 1a), who performs ϕ directed steps along one axis of the lattice, following by ψ directed steps along another axis. At every step the walker dies with probability $\varepsilon = 1 - f$. Then, having a number of the walkers starting from the origin of the lattice, one would get an "orchard of walkers", i.e. at every spot ν on the x axis a fraction of survived walkers $P(\nu)$ would be described exactly by the popcorn function.

In order to have a relevant physical picture, consider a toy model of diblock-copolymer polymerization. Without sticking to any specific polymerization mechanism, consider an ensemble of diblock-copolymers AB, polymerized independently from both ends in a cloud of monomers of relevant kind (we assume, only $A-A$ and $B-B$ links to be formed). Termination of polymerization is provided by specific "radicals" of very small concentration, ε: when a radical is attached to the end (irrespectively, A or B), it terminates the polymerization at this extremity forever. Given the environment of infinite capacity, one assigns the probability $f = 1 - \varepsilon$ to a monomer attachment at every elementary act of the polymerization. If N_A and N_B are molecular weights of the blocks A and B, then the composition probability distribution in our ensemble, $P\left(\varphi = \frac{N_A}{N_A+N_B}\right)$, in the limit of small $\varepsilon \ll 1$ is "ultrametric" (see [6] for the definition of the ultrametricity) and is given by the popcorn function:

$$P\left(\varphi = \frac{p}{p+q}\right) \approx \frac{1}{\varepsilon(p+q)} \stackrel{def}{=} \frac{1}{\varepsilon} g(\varphi) \quad (3)$$

In the described process we have assumed identical independent probabilities for the monomers of sorts ("colors") A and B to be attached at both chain ends. Since no preference is implied, one may look at this process as at a homopolymer ("colorless") growth, taking place at two extremities. For this process we are interested in statistical characteristics of the resulting ensemble of the homopolymer chains. What would play the role of "composition" in this case, or in other words, how should one understand the fraction of monomers attached at one end? As we show below, the answer is rather intriguing: the respective analogue of the probability distribution is the spectral density of the ensemble of linear chains with the probability Q_L for the molecular mass distribution, where L is the length of a chain in the ensemble.

To our point of view, the popcorn function has not yet received decent attention among researchers, though its emergence in various physical problems seems impressive, as we demonstrate below. Apparently, the main difficulty deals with the discontinuity of $g(x)$ at every rational point, which often results in a problematic theoretical treatment and interpretation of results for the underlying physical system. Thus, a natural, physically justified "continuous approximation" to the popcorn function is very demanded.

Below we provide such an approximation, showing the generality of the "popcorn-like" distributions for a class of one-dimensional disordered systems. We demonstrate that the popcorn function can be constructed on the basis of the modular Dedekind function, $\eta(x+iy)$, when the imaginary part, y, of the modular parameter $z = x + iy$ tends to 0.

II. SPECTRAL STATISTICS OF EXPONENTIALLY WEIGHTED ENSEMBLE OF LINEAR GRAPHS

A. Spectral density and the popcorn function

The former exercises are deeply related to the spectral statistics of ensembles of linear polymers. In a practical setting, consider an ensemble of noninteracting linear chains with exponential distribution in their lengths. We claim the emergence of the fractal popcorn-like structure in the spectral density of corresponding adjacency matrices describing the connectivity of elementary units (monomers) in linear chains.

The ensemble of exponentially weighted homogeneous chains, is described by the bi-diagonal symmetric $N \times N$

adjacent matrix $B = \{b_{ij}\}$:

$$B = \begin{pmatrix} 0 & x_1 & 0 & 0 & \cdots \\ x_1 & 0 & x_2 & 0 & \\ 0 & x_2 & 0 & x_3 & \\ 0 & 0 & x_3 & 0 & \\ \vdots & & & & \ddots \end{pmatrix} \tag{4}$$

where the distribution of each $b_{i,i+1} = b_{i+1,i} = x_i$ $(i = 1, ..., N)$ is Bernoullian:

$$x_i = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{with probability } \varepsilon = 1 - f \end{cases} \tag{5}$$

We are interested in the spectral density, $\rho_\varepsilon(\lambda)$, of the ensemble of matrices B in the limit $N \to \infty$. Note that at any $x_k = 0$, the matrix B splits into independent blocks. Every $n \times n$ block is a symmetric $n \times n$ bi-diagonal matrix A_n with all $x_k = 1$, $k = 1, ..., n$, which corresponds to a chain of length n. The spectrum of the matrix A_n is

$$\lambda_{k,n} = 2\cos\frac{\pi k}{n+1}; \qquad (k = 1, ..., n) \tag{6}$$

All the eigenvalues $\lambda_{k,n}$ for $k = 1, ..., n-1$ appear with the probability $Q_n = f^n$ in the spectrum of the matrix (4). In the asymptotic limit $\varepsilon \ll 1$, one may deduce an equivalence between the composition distribution in the polymerization problem, discussed in the previous section, and the spectral density of the linear chain ensemble. Namely, the probability of a composition $\varphi = \frac{p}{p+q}$ in the ensemble of the diblock-copolymers can be precisely mapped onto the peak intensity (the degeneracy) of the eigenvalue $\lambda = \lambda_{p,p+q-1} = 2\cos\frac{\pi p}{p+q}$ in the spectrum of the matrix B. In other words, the integer number k in the mode $\lambda_{k,n}$ matches the number of A-monomers, $N_A = kz$, while the number of B-monomers matches $N_B = (n + 1 - k)z$, where $z \in N$, in the respective diblock-copolymer.

The spectral statistics survives if one replaces the ensemble of Bernoullian two-diagonal adjacency matrices B defined by (4)–(5) by the ensemble of random Laplacian matrices. Recall that the Laplacian matrix, $L = \{a_{ij}\}$, can be constructed from adjacency matrix, $B = \{b_{ij}\}$, as follows: $a_{ij} = -b_{ij}$ for $i \neq j$, and $a_{ii} = \sum_{j=1}^{N} b_{ij}$. A search for eigenvalues of the Laplacian matrix L for linear chain, is equivalent to determination its relaxation spectrum. Thus, the density of the relaxation spectrum of the ensemble of noninteracting linear chains with the exponential distribution in lengths, has the signature of the popcorn function.

To derive $\rho_\varepsilon(\lambda)$ for arbitrary values of ε, let us write down the spectral density of the ensemble of $N \times N$ random matrices B with the bimodal distribution of the elements as a resolvent:

$$\begin{aligned} \rho_\varepsilon(\lambda) &= \lim_{N\to\infty} \Big\langle \sum_{k=1}^{n} \delta(\lambda - \lambda_{kn}) \Big\rangle_{Q_n} \\ &= \lim_{\substack{N\to\infty \\ y\to+0}} y \,\mathrm{Im}\, \Big\langle G_n(\lambda - iy) \Big\rangle_{Q_n} \\ &= \lim_{\substack{N\to\infty \\ y\to+0}} y \sum_{n=1}^{N} Q_n \,\mathrm{Im}\, G_n(\lambda - iy) \end{aligned} \tag{7}$$

where $\langle ... \rangle_{Q_n}$ means averaging over the distribution $Q_n = (1-\varepsilon)^n$, and the following regularization of the Kronecker δ-function is used:

$$\delta(\xi) = \lim_{y\to+0} \mathrm{Im}\, \frac{y}{\xi - iy} \tag{8}$$

The function G_n is associated with each particular gapless matrix B of n sequential "1" on the sub-diagonals,

$$G_n(\lambda - iy) = \sum_{k=1}^{n} \frac{1}{\lambda - \lambda_{k,n} - iy} \tag{9}$$

Collecting (6), (7) and (9), we find an explicit expression for the density of eigenvalues:

$$\rho_\varepsilon(\lambda) = \lim_{\substack{N\to\infty \\ y\to+0}} y \sum_{n=1}^{N} (1-\varepsilon)^n \sum_{k=1}^{n} \frac{y}{\left(\lambda - 2\cos\frac{\pi k}{n+1}\right)^2 + y^2} \tag{10}$$

The behavior of the inner sum in the spectral density in the asymptotic limit $y \to 0$ is easy to understand: it is $\frac{1}{y}$ at $\lambda = 2\cos\frac{\pi k}{n+1}$ and zero otherwise. Thus, one can already infer a qualitative similarity with the popcorn function. It turns out, that the correspondence is quantitative for $\varepsilon = 1 - f \ll 1$. Driven by the purpose to show it, we calculate the values of $\rho_\varepsilon(\lambda)$ at the peaks, i.e. at rational points $\lambda = 2\cos\frac{\pi p}{p+q}$ with $\gcd(p, q) = 1$ and end up with the similar geometrical progression, as for the case of diblock-copolymers problem (2):

$$\begin{aligned} \rho_\varepsilon\left(\lambda = 2\cos\frac{\pi p}{p+q}\right) &= \sum_{s=1}^{\infty} (1-\varepsilon)^{(p+q)s-1} \\ &= \left.\frac{(1-\varepsilon)^{p+q-1}}{1-(1-\varepsilon)^{p+q}}\right|_{\varepsilon\to 0} \approx \frac{1}{\varepsilon(p+q)} \\ &\stackrel{def}{=} g\left(\frac{1}{\pi}\arccos\frac{\lambda}{2}\right) \end{aligned} \tag{11}$$

The typical sample plot $\rho_\varepsilon(\lambda)$ for $f = 0.7$ computed numerically via (10) with $\varepsilon = 2\times10^{-3}$ is shown in the Fig. 2 for $N = 10^3$.

B. Enveloping curves and tails of the eigenvalues density

Below we pay attention to some number-theoretic properties of the spectral density of the argument $-\lambda$,

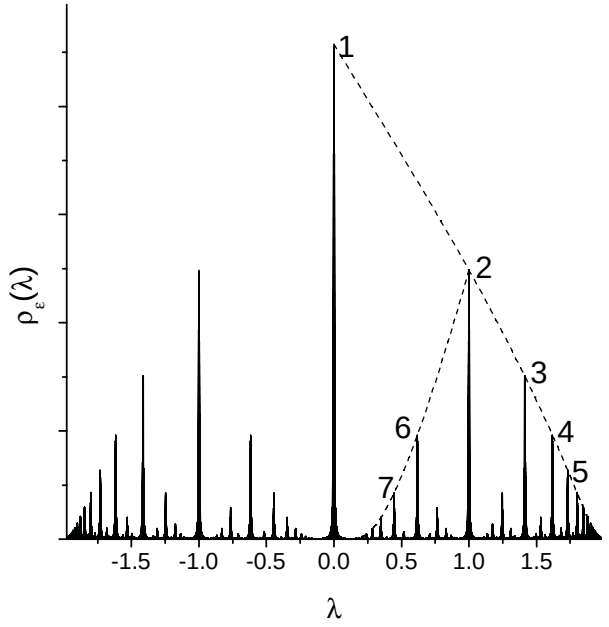

Figure 2: The spectral density $\rho_\varepsilon(\lambda)$ for the ensemble of bidiagonal matrices of size $N = 10^3$ at $f = 0.7$. The regularization parameter ε is taken $\varepsilon = 2 \times 10^{-3}$.

since in this case the correspondence with the composition ratio is precise. One can compute the enveloping curves for any monotonic sequence of peaks depicted in Fig. 2, where we show two series of sequential peaks: $S_1 = \{1\text{–}2\text{–}3\text{–}4\text{–}5\text{–}...\}$ and $S_2 = \{2\text{–}6\text{–}7\text{–}...\}$. Any monotonic sequence of peaks corresponds to the set of eigenvalues $\lambda_{k,n}$ constructed on the basis of a Farey sequence [7]. For example, as shown below, the peaks in the series S_1 are located at:

$$\lambda_k = -\lambda_{k,k} = -2\cos\frac{\pi k}{k+1}, \qquad (k = 1, 2, ...)$$

while the peaks in the series S_2 are located at:

$$\lambda_{k'} = -\lambda_{k',2k'-2} = -2\cos\frac{\pi k'}{2k'-1}, \qquad (k' = 2, 3, ...)$$

Positions of peaks obey the following rule: let $\{\lambda_{k-1}, \lambda_k, \lambda_{k+1}\}$ be three consecutive monotonically ordered peaks (e.g., peaks 2–3–4 in Fig. 2), and let

$$\lambda_{k-1} = -2\cos\frac{\pi p_{k-1}}{q_{k-1}}, \quad \lambda_{k+1} = -2\cos\frac{\pi p_{k+1}}{q_{k+1}}$$

where p_k and q_k $(k = 1, ..., N)$ are coprimes. The position of the intermediate peak, λ_k, is defined as

$$\lambda_k = -2\cos\frac{\pi p_k}{q_k}; \qquad \frac{p_k}{q_k} = \frac{p_{k-1}}{q_{k-1}} \oplus \frac{p_{k+1}}{q_{k+1}} \equiv \frac{p_{k-1} + p_{k+1}}{q_{k-1} + q_{k+1}} \tag{12}$$

The sequences of coprime fractions constructed via the $\oplus$ addition are known as Farey sequences. A simple geometric model behind the Farey sequence, known as Ford circles [8, 9], is shown in Fig. 3a. In brief, the construction goes as follows. Take the segment $[0, 1]$ and draw two circles O_1 and O_2 both of radius $r = \frac{1}{2}$, which touch each other, and the segment at the points 0 and 1. Now inscribe a new circle O_3 touching O_1, O_2 and $[0, 1]$. Where is the position of the new circle along the segment? The generic recursive algorithm constitutes the Farey sequence construction. Note that the same Farey sequence can be sequentially generated by fractional-linear transformations (reflections with respect to the arcs) of the fundamental domain of the modular group $SL(2, Z)$ – the triangle lying in the upper halfplane $\mathrm{Im}\, z > 0$ of the complex plane z (see Fig. 3b).

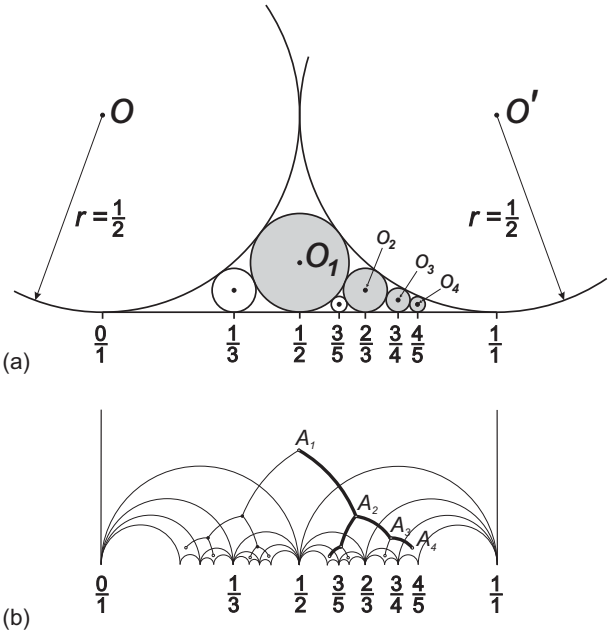

Figure 3: Ford circles as illustration of the Farey sequence construction: (a) Each circle touches two neighbors (right and left) and the segment. The position of newly generated circle is determined via the $\oplus$ addition: $\frac{p_{k-1}}{q_{k-1}} \oplus \frac{p_{k+1}}{q_{k+1}} = \frac{p_{k-1}+p_{k+1}}{q_{k-1}+q_{k+1}}$; (b) The same Farey sequence generated by sequential fractional-linear transformations of the fundamental domain of the modular group $SL(2, Z)$.

Consider the main peaks series, $S_1 = \{1\text{–}2\text{–}3\text{–}4\text{–}5\text{–}...\}$. The explicit expression for their positions reads as:

$$\lambda_k = -2\cos\frac{\pi k}{k+1}; \qquad k = 1, 2, ... \tag{13}$$

One can straightforwardly investigate the asymptotic behavior of the popcorn function in the limit $k \to \infty$. From (11) one has for arbitrary $f < 1$ the set of parametric equations:

$$\begin{cases} \rho_\varepsilon(\lambda_k) = \dfrac{f^k}{1 - f^{k+1}}\Big|_{k\gg 1} \approx f^k \\ \lambda_k = -2\cos\dfrac{\pi k}{k+1}\Big|_{k\gg 1} \approx 2 - \dfrac{\pi^2}{k^2} \end{cases} \tag{14}$$

From the second equation of (14), we get $k \approx \frac{\pi}{\sqrt{2-\lambda}}$. Substituting this expression into the first one of (14), we end up with the following asymptotic behavior of the spectral density near the spectral edge $\lambda \to 2^-$:

$$\rho_\varepsilon(\lambda) \approx \exp\left(\frac{\pi \ln f}{\sqrt{2-\lambda}}\right) \qquad (0 < f < 1) \qquad (15)$$

The behavior (14) is the signature of the Lifshitz tail typical for the 1D Anderson localization:

$$\rho_\varepsilon(E) \approx e^{-CE^{-D/2}}; \qquad (16)$$

where $E = 2 - \lambda$ and $D = 1$.

III. FROM POPKORN TO DEDEKIND η-FUNCTION

A. Some facts about Dedekind η-function and related series

The popcorn function has discontinuous maxima at rational points and continuous valleys at irrationals. We show in this section, that the popcorn function can be regularized on the basis of the everywhere continuous Dedekind function $\eta(x + iy)$ in the asymptotic limit $y \to 0$.

The famous Dedekind η-function is defined as follows:

$$\eta(z) = e^{\pi i z/12} \prod_{n=0}^{\infty} (1 - e^{2\pi i n z}) \qquad (17)$$

The argument $z = x+iy$ is called the modular parameter and $\eta(z)$ is defined for $\mathrm{Im}\, z > 0$ only. The Dedekind η-function is invariant with respect to the action of the modular group $SL(2,\mathbb{Z})$:

$$\begin{aligned} \eta(z+1) &= e^{\pi i z/12}\, \eta(z) \\ \eta\left(-\tfrac{1}{z}\right) &= \sqrt{-i}\, \eta(z) \end{aligned} \qquad (18)$$

And, in general,

$$\eta\left(\frac{az+b}{cz+d}\right) = \omega(a,b,c,d)\, \sqrt{cz+d}\, \eta(z) \qquad (19)$$

where $ad - bc = 1$ and $\omega(a, b, c, d)$ is some root of 24th degree of unity [10].

It is convenient to introduce the following "normalized" function

$$h(z) = |\eta(z)|(\mathrm{Im}\, z)^{1/4} \qquad (20)$$

The real analytic Eisenstein series $E(z, s)$ is defined in the upper half-plane, $H = \{z : \mathrm{Im}\,(z) > 0\}$ for $\mathrm{Re}\,(s) > 1$ as follows:

$$E(z,s) = \frac{1}{2} \sum_{\{m,n\} \in \mathbb{Z}^2 \setminus \{0,0\}} \frac{y^s}{|mz+n|^{2s}}; \qquad z = x + iy \qquad (21)$$

This function can be analytically continued to all s-plane with one simple pole at $s = 1$. Notably it shares the same invariance properties on z as the Dedekind η-function. Moreover, $E(s, z)$, as function of z, is the $SL(2,\mathbb{Z})$–automorphic solution of the hyperbolic Laplace equation:

$$-y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) E(z,s) = s(1-s)\, E(z,s)$$

The Eisenstein series is closely related to the Epstein ζ-function, $\zeta(s, Q)$, namely:

$$\zeta(s,Q) = \sum_{\{m,n\} \in \mathbb{Z}^2 \setminus \{0,0\}} \frac{1}{Q(m,n)^s} = \frac{2}{d^{s/2}} E(z,s), \qquad (22)$$

where $Q(m, n) = am^2 + 2bmn + cn^2$ is a positive definite quadratic form, $d = ac - b^2 > 0$, and $z = \dfrac{-b + i\sqrt{d}}{a}$. Eventually, the logarithm of the Dedekind η-function is known to enter in the Laurent expansion of the Epstein ζ-function. Its residue at $s = 1$ has been calculated by Dirichlet and is known as the first Kronecker limit formula [11–13]. Explicitly, it reads at $s \to 1$:

$$\begin{aligned} \zeta(s,Q) = {} & \frac{\pi}{\sqrt{d}} \frac{1}{s-1} \\ & + \frac{2\pi}{\sqrt{d}} \left(\gamma + \ln\sqrt{\frac{a}{4d}} - 2\ln|\eta(z)|\right) \\ & + O(s-1) \end{aligned} \qquad (23)$$

Equation (23) establishes the important connection between the Dedekind η-function and the respective series, that we substantially exploit below.

B. Relation between the popcorn and Dedekind η functions

Consider an arbitrary quadratic form $Q'(m, n)$ with unit determinant. Since $d = 1$, it can be written in new parameters $\{a, b, c\} \to \{x = \frac{b}{c}, \varepsilon = \frac{1}{c}\}$ as follows:

$$Q'(m,n) = \frac{1}{\varepsilon}(xm - n)^2 + \varepsilon m^2 \qquad (24)$$

Applying the first Kronecker limit formula to the Epstein function with (24) and $s = 1 + \tau$, where $\tau \ll 1$, but finite one gets:

$$\begin{aligned} \zeta(s,Q') = {} & \frac{\pi}{s-1} \\ & + 2\pi \left(\gamma + \ln\sqrt{\frac{1}{4\varepsilon}} - 2\ln|\eta(x+i\varepsilon)|\right) \\ & + O(s-1) \end{aligned} \qquad (25)$$

On the other hand, one can make use of the ε-continuation of the Kronecker δ-function, (8), and assess $\zeta(1+\tau, Q')$ for small $\tau \ll 1$ as follows:

$$\zeta(1+\tau, Q') \approx \frac{1}{\varepsilon} \sum_{\{m,n\}\in\mathbb{Z}^2\setminus\{0,0\}} \frac{\varepsilon^2}{(xm-n)^2+\varepsilon^2 m^2}$$
$$= \frac{2}{\varepsilon} \lim_{N\to\infty} \sum_{m=1}^{N}\sum_{n=1}^{N} \frac{1}{m^2}\delta\left(x-\frac{n}{m}\right) \equiv \theta(x) \quad (26)$$

where $x \in (0,1)$ and the factor 2 reflects the presence of two quadrants on the $\mathbb{Z}^2$-lattice that contribute jointly to the sum at every rational points, while θ assigns 0 to all irrationals. At rational points $\theta\left(\frac{p}{q}\right)$ can be calculated straightforwardly:

$$\theta\left(\frac{p}{q}\right) = \frac{2}{\varepsilon}\sum_{m|q}^{\infty}\frac{1}{m^2} = \frac{\pi^2}{3\varepsilon q^2} \quad (27)$$

Comparing (27) with the definition of the popcorn function, g, one ends up with the following relation at the peaks:

$$g\left(\frac{p}{q}\right) = \sqrt{\frac{3\varepsilon}{\pi^2}\theta\left(\frac{p}{q}\right)} \quad (28)$$

Eventually, collecting (25) and (28), we may write down the regularization of the popcorn function by the Dedekind $\eta(x+i\varepsilon)|_{\varepsilon\to 0}$ in the interval $0<x<1$:

$$g(x) \approx \sqrt{-\frac{12\varepsilon}{\pi}\ln|\eta(x+i\varepsilon)| - o(\varepsilon\ln\varepsilon)}\Bigg|_{\varepsilon\to 0} \quad (29)$$

or

$$-\ln|\eta(x+i\varepsilon)|_{\varepsilon\to 0} = \frac{\pi}{12\varepsilon}g^2(x) + O(\ln\varepsilon) \quad (30)$$

Note, that the asymptotic behavior of the Dedekind η-function can be independently derived through the duality relation, [6]. However, such approach leaves in the dark the underlying structural equivalence of the popcorn and η functions and their series representation on the lattice $\mathbb{Z}^2$. In the Fig. 4 we show two discrete plots of the left and the right-hand sides of (30).

Thus, the spectral density of ensemble of linear chains, (11), in the regime $\varepsilon \ll 1$ is expressed through the Dedekind η-function as follows:

$$\rho_\varepsilon(\lambda) \approx \sqrt{-\frac{12\varepsilon}{\pi}\ln\left|\eta\left(\frac{1}{\pi}\arccos\frac{\lambda}{2}+i\varepsilon\right)\right|} \quad (31)$$

IV. CONCLUSION

We have discussed the number-theoretic properties of distributions appearing in physical systems when an observable is a quotient of two independent exponentially weighted integers. The spectral density of ensemble of linear polymer chains distributed with the law f^L ($0<f<1$), where L is the chain length, serves as a particular example. At $f\to 1$, the spectral density can be expressed through the discontinuous and non-differentiable at all rational points, Thomae ("popcorn") function. We suggest a continuous approximation of the popcorn function, based on the Dedekind η-function near the real axis.

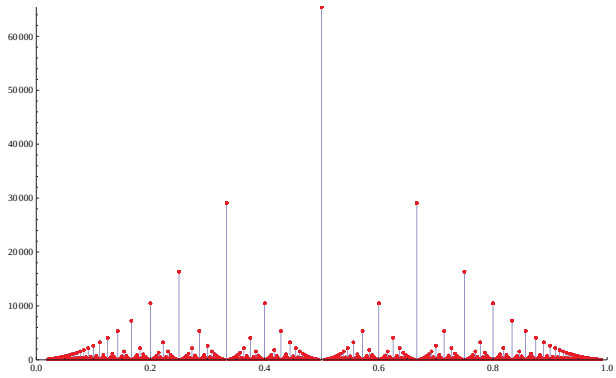

Figure 4: Plots of everywhere continuous $f_1(x) = -\ln|\eta(x+i\varepsilon)|$ (blue) and discrete $f_2(x) = \frac{\pi}{12\varepsilon}g^2(x)$ (red) for $\varepsilon = 10^{-6}$ at rational points in $0<x<1$.

Analysis of the spectrum at the edges reveals the Lifshitz tails, typical for the 1D Anderson localization. The non-trivial feature, related to the asymptotic behavior of the shape of the spectral density of the adjacency matrix, is as follows. The main, enveloping, sequence of peaks $1-2-3-4-5...$ in the Fig. 2 has the asymptotic behavior $\rho(\lambda) \sim q^{\pi/\sqrt{2-\lambda}}$ (at $\lambda\to 2^-$) typical for the 1D Anderson localization, however any internal subsequence of peaks, like $2-6-7-...$, has the behavior $\rho'(\lambda) \sim q^{\pi/|\lambda-\lambda_{cr}|}$ (at $\lambda\to\lambda_{cr}$) which is reminiscent of the Anderson localization in 2D.

We would like to emphasize that the ultrametric structure of the spectral density is ultimately related to number-theoretic properties of modular functions. We also pay attention to the connection of the Dedekind η-function near the real axis to the invariant measures of some continued fractions studied by Borwein and Borwein in 1993 [17]. The notion of ultrametricity deals with the concept of hierarchical organization of energy landscapes [19, 20]. A complex system is assumed to have a large number of metastable states corresponding to local minima in the potential energy landscape. With respect to the transition rates, the minima are suggested to be clustered in hierarchically nested basins, i.e. larger basins consist of smaller basins, each of these consists of even smaller ones, *etc.* The basins of local energy minima are separated by a hierarchically arranged set of barriers: large basins are separated by high barriers, and smaller basins within each larger one are separated by lower barriers. Ultrametric geometry fixes taxonomic (i.e. hierarchical) tree-like relationships between elements and,

speaking figuratively, is closer to Lobachevsky geometry, rather to the Euclidean one.

Acknowledgments

We are very grateful to V. Avetisov, A. Gorsky, Y. Fyodorov and P. Krapivsky for many illuminating discussions. The work is partially supported by the IRSES DIONICOS and RFBR 16-02-00252A grants.

[1] Beanland, K., Roberts, J. W., Stevenson, C. (2009). Modifications of Thomae's function and differentiability. *American Mathematical Monthly, 116*(6), 531-535.

[2] Trifonov, V., Pasqualucci, L., Dalla-Favera, R., Rabadan, R. (2011). Fractal-like distributions over the rational numbers in high-throughput biological and clinical data. *Scientific reports, 1*.

[3] Planat, M., Eckert, C. (2000). On the frequency and amplitude spectrum and the fluctuations at the output of a communication receiver. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 47*(5), 1173-1182.

[4] Lundholm, D. (2016). Many-anyon trial states. *arXiv preprint* arXiv:1608.05067.

[5] Middendorf, M., Ziv, E., Wiggins, C. H. (2005). Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America, 102*(9), 3192-3197.

[6] Avetisov, V., Krapivsky, P. L., Nechaev, S. (2015). Native ultrametricity of sparse random ensembles. *Journal of Physics A: Mathematical and Theoretical, 49(3)*, 035101.

[7] Hardy, G. H., Wright, E. M. (1979). An introduction to the theory of numbers. *Oxford University Press*.

[8] Ford, L. R. (1938). Fractions. *The American Mathematical Monthly, 45*(9), 586-601.

[9] Coxeter, H. S. M. (1968). The problem of Apollonius. *The American Mathematical Monthly, 75*(1), 5-15.

[10] Chandrassekharan, K (1985). Elliptic Functions. *Berlin: Springer*

[11] Epstein, P. (1903). Zur Theorie allgemeiner Zetafunctionen. *Mathematische Annalen, 56*(4), 615-644.

[12] Siegel, C. L. (1961). Lectures on advanced analytic number theory, *Tata Inst. of Fund. Res., Bombay*.

[13] Motohashi, Y. (1968). A new proof of the limit formula of Kronecker. *Proceedings of the Japan Academy, 44*(7), 614-616.

[14] Dyson, F. J. (1953). The dynamics of a disordered linear chain. *Physical Review, 92*(6), 1331.

[15] Domb, C., Maradudin, A. A., Montroll, E. W., Weiss, G. H. (1959). Vibration Frequency Spectra of Disordered Lattices. I. Moments of the Spectra for Disordered Linear Chains. *Physical Review, 115*(1), 18.

[16] Nieuwenhuizen, T. M., Luck, J. M. (1985). Singular behavior of the density of states and the Lyapunov coefficient in binary random harmonic chains. *Journal of statistical physics, 41*(5), 745-771.

[17] Borwein, J. M., Borwein, P. B. (1993). On the generating function of the integer part:$[n\alpha + \gamma]$. *Journal of Number Theory, 43*(3), 293-318.

[18] Comtet, A., Nechaev, S. (1998). Random operator approach for word enumeration in braid groups. *Journal of Physics A: Mathematical and General, 31*(26), 5609.

[19] Mezard, M., Parisi, G., Virasoro, M. (1987). Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications (Vol. 9). *World Scientific Publishing Co Inc*.

[20] Frauenfelder, H. (1987). The connection between low-temperature kinetics and life. In Protein Structure (pp. 245-261). *Springer New York*.

3. From geometric optics to plants: the eikonal equation for buckling

Introduction

Optimal buckling of a tissue, e.g. a plant leaf, growing by means of exponential division of its peripheral cells, is considered in the framework of a conformal approach. It is shown that the boundary profile of a tissue is described by the 2D eikonal equation, which provides the geometric optic approximation for the wavefront propagating in a medium with an inhomogeneous refraction coefficient. By means of a local conformal mapping of the hyperbolic triangle onto the Euclidean one, we demonstrate that the elastic energy of the buckled tissue is expressed through the Dedekind η-function. Thus, the hierarchical organization of soft growing membranes is a natural result due to the number-theoretic properties of the underlying modular form.

Contribution

I have derived the eikonal equation from the condition of the area conservation and performed numerical integration of the profiles with the Dedekind function on the right-hand side in various geometrical settings.

From geometric optics to plants: eikonal equation for buckling

Sergei Nechaev[*a,b] **and Kirill Polovnikov,**[c,d]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

Optimal buckling of a tissue, e.g. plant leaf, growing by means of exponential division of its periphery cells, is considered in the framework of a conformal approach. It is shown that the boundary profile of the tissue is described by the 2D eikonal equation, which provides the geometric optic approximation for the wave front propagating in the media with inhomogeneous refraction coefficient. The variety of optimal surfaces embedded in 3D is controlled by spatial dependence of the refraction coefficient which, in turn, is dictated by the local growth protocol.

1 Introduction

A variety of complex 2D profiles of growing tissues emerges due to the incompatibility of local internal (differential) growth protocol with geometric constraints imposed by embedding of these tissues into the Euclidean space. For example, buckling of a lettuce leaf can be naively explained as a conflict between natural growth due to the periphery cells division (typically, exponential), and growth of circumference of a planar disc with gradually increasing radius. Due to a specific biological mechanism which inhibits growth of the cell experiencing sufficient external pressure, the division of inner cells is insignificant, while periphery cells have less steric restrictions and proliferate easier. Thus, the division of border cells has the major impact on the instabilities in the tissue. Such a differential growth induces an increasing strain in a tissue near its edge and results in two complimentary possibilities: i) in-plane tissue compression and/or redistribution of layer cells accompanied by the in-plane circumference instability, or ii) out-of-plane tissue buckling with the formation of saddle-like surface regions. The latter is typical for various undulant negatively curved shapes which are ubiquitous to many mild plants growing up in air or water where the gravity is of sufficiently small matter[1,2].

A widely used energetic approach to growing patterns exploits a continuous formulation of the differential growth and is based on a rivalry between bending and stretching energies of elastic membranes[2,7–9,12,18–20], reflecting the choice between options (i) and (ii) above. For bending rigidity of a thin membrane, $\mathscr{B}$, one has $\mathscr{B} \sim h^3$, while stretching rigidity, $\mathscr{S}$ behaves as $\mathscr{S} \sim h$, where h is the membrane thickness[14]. Therefore, thin enough tissues, with $h \ll 1$, prefer to bend, i.e. to be negatively curved under relatively small critical strain.

The latter allows one to eliminate the "stretching" regime from consideration, justifying the geometric approach for infinitesimally thin membranes[4,5,10,12,17,21,22] (see also[6]). Here the determination of typical profiles of buckling surfaces relies on an appropriate choice of metric tensor of the non-Euclidean space, and is realized via the optimal embedding of the tissue with certain metrics into the 3D Euclidean space. It should be mentioned, that the formation of wrinkles within this approach seems to be closely related to the description of phyllotaxis via conformal methods[23].

In this letter we suggest a model of a hyperbolic infinitesimally thin tissue, whose periphery cells divide freely with exponential rate, while division of inner cells is absolutely inhibited. Two cases of proliferations, the one-dimensional (directed) and the uniform two-dimensional, are considered. The selection of these two growth models is caused by the intention to describe different symmetries inherent for plants at initial stages of growth. As long as the in-plane deformations are not beneficial, as follows from the relationship between bending and stretching rigidities, all the redundant material of fairly elastic tissue will buckle out. In order to take into account the finite elasticity of growing tissue, resulting from the intrinsic discrete properties of a material, we describe the tissue as a collection of glued elementary plaquettes connected along the hyperbolic graph, γ. The discretization implies the presence of a characteristic scale, of order of the elementary cell (plaquette) size, below which the tissue is locally flat.

As we rely on the absence of in-plane deformations, this graph has to be isometrically embedded into 3D space. The desired smooth surface profile is obtained in two steps: i) isometric mapping of the hyperbolic graph onto the flat domain (rectangular or circular) with hyperbolic metrics, ii) subsequent restoring of the metrics into the 3D Euclidean space above the domain. We demonstrate that such a procedure leads

[a] *J.-V. Poncelet Laboratory, CNRS, UMI 2615, 119002 Moscow, Russia; E-mail: sergei.nechaev@gmail.com*
[b] *P.N. Lebedev Physical Institute, RAS, 119991 Moscow, Russia*
[c] *Physics Department, Moscow State University, 119992 Moscow, Russia*
[d] *The Skolkovo Institute for Science and Technology, 143005 Skolkovo, Russia*

to the "optimal" buckling of the tissue and is described by the eikonal equation for the profile, $f(x,y)$, of growing sample, which by definition, is a variant of the Hamilton-Jacobi equation.

The paper is organized as follows. We introduce necessary definitions in Section 2.1; the model under consideration and the details of the conformal approach are provided in Sections s:2, 2.3 and Appendix; the samples of various typical shapes for two-dimensional uniform and for one-dimensional directed growth, are presented in Section 3; finally, the results of the work are summarized in Section 4, where we also speculate about possible generalizations and rise open questions.

2 Buckling of thin tissues in cylindric and planar geometries

2.1 Basic facts about the eikonal equation

To make the content of the paper as self-contained as possible, it seems instructive to provide some important definitions used throughout the paper. The key ingredient of our consideration is the "eikonal" equation, which is the analogue of the Hamilton-Jacobi equation in geometric optics. As we show below, the eikonal equation provides optimal embedding of an exponentially growing surface into the 3D Euclidean plane. Meaning of the notion "optimal" has two different connotations in our approach:

i) On one hand, from viewpoint of the Hamilton-Jacobi theory, the eikonal equation appears in the minimization of the action $A = \int_\gamma L dt$ with some Lagrangian L. According to the Fermat principle, the time of the ray propagation in the inhomogeneous media with the space-dependent refraction coefficient, $n(x,y)$, should be minimal.

ii) On the other hand, the eikonal equation emerges in our work in a purely geometric setting following directly from the conformal approach.

First attempts to formulate classical mechanics problems in geometric optics terms goes back to the works of Klein[27] in 19th century. His ideas contributed to the corpuscular theory in a short-wavelength regime, as long as the same mechanical formalism applied to massless particles, was consistent with the wave approach. Later, in the context of general relativity, this approach was renewed to treat gravitational field as an optic medium[28].

The Fermat principle states that the time dt for a ray to propagate along a curve γ between two closely located points $M(\mathbf{x})$ and $N(\mathbf{x}+d\mathbf{x})$ in an inhomogeneous media, should be minimal. The total time T can be written in the form $T = \frac{1}{c}\int_M^N n(\mathbf{x}(s))ds$ where $n(\mathbf{x}) = \frac{c}{\mathrm{v}(\mathbf{x})}$ is the refraction coefficient at the point $\mathbf{x} = \{x^i\}$ of a D-dimensional space ($i = 1,...,D$), c and $\mathrm{v}(\mathbf{x})$ are correspondingly the light speeds in vacuum and in the media, and $d|\mathbf{x}| = ds$ is the spatial increment along the ray. Following the optical-mechanical analogy, according to which the action in mechanics corresponds to eikonal in optics, one can write down the "optic length" or eikonal, $S = cT$ in Lagrangian terms: $S = \int_M^N L(\mathbf{x},\dot{\mathbf{x}})ds$ with the Lagrangian $L(\mathbf{x},\dot{\mathbf{x}}) = n(\mathbf{x}(s))\sqrt{\dot{\mathbf{x}}(s)\dot{\mathbf{x}}(s)}$, where $\dot{\mathbf{x}}^2 = \sum_{i=1}^D \left(\frac{dx^i}{ds}\right)^2$. We would like to mention here, that optical properties of the media can be also treated in terms of induced Riemann metrics in vacuum:

$$S = \int_M^N n(\mathbf{x}(s))ds = \int_M^N \sqrt{\dot{\mathbf{x}} g(\mathbf{x}) \dot{\mathbf{x}}} ds \tag{1}$$

where $g_{ij} = n^2(\mathbf{x})\delta_{ij}$ stands for induced metrics components in isotropic media case. Thus, from the geometrical point, the ray trajectory can be understood as a "minimal curve" in a certain Riemann space. This representation suggests to consider optimal ray paths as geodesics in the space with known metrics g.

Stationarity of optic length, S, i.e. $\delta S = 0$, together with the condition $|\dot{\mathbf{x}}| = 1$, defines the Euler equation:

$$\frac{d}{ds}\left(n(\mathbf{x})\frac{d\mathbf{x}}{ds}\right) = \nabla n(\mathbf{x}) \tag{2}$$

from which one can directly proceed to the Huygens principle by integrating (2) over s: $\nabla S(\mathbf{x}) = n(\mathbf{x})\frac{d\mathbf{x}}{ds}$. Squaring both sides of the latter equation we end up with the eikonal equation:

$$(\nabla S(\mathbf{x}))^2 = n^2(\mathbf{x}) \tag{3}$$

The eikonal equation Eq.(3) has the same form as the Hamilton-Jacobi equation in mechanics for action in the $D+1$-dimensional space, which in turn can be understood as the relativistic equation for the light, propagating in the Riemannian space.

2.2 The model: formalization of physical ideas

In our work the eikonal equation arises in the differential growth problem in a purely geometric setting. Consider a tissue, represented by a colony of cells, growing in space without any geometric constraints. The local division protocol is prescribed by nature, being particularly recorded in genes and is accompanied by their mutations[16]. The exponential cell division is implied, as already mentioned above. To make our viewpoint more transparent, suppose that all cells, represented by equilateral triangles, divide independently and their proliferation is initiated by the first "protocell". Connecting the centers of neighboring triangles by nodes, we rise a graph γ. The number of vertices, $P_\gamma(k)$, in the generation k, grows exponentially with k: $P_\gamma(k) \sim c^k$ ($c > 1$). It is known that exponential graphs possess hyperbolic metrics, meaning that they can be isometrically (with fixed branch lengths and angles between adjacent branches) embedded into a hyperbolic plane. Thus, it

is clear, that the corresponding surface, pulled on the isometry of such graph in the 3D Euclidean space, should be negatively curved.

To have a relevant image, suppose that we grow the surface by crocheting it spirally starting from the center[3]. Demanding two nearest neighboring circumference layers, $P(r)$ and $P(r+\Delta r)$, to differ by a factor of c (where $c = \text{const} > 1$), i.e., $P(r+\Delta r)/P(r) = c$, we construct an exponentially growing (hyperbolic) surface, see Fig.1a. The crocheted surface has well-posed properties on large scales, but should be precisely described on the scale of order of the elementary cell. As we have mentioned, the microscopic description is connected with the specific local growth protocol. The simplest way to generate the discrete hyperbolic-like surface out of equilateral triangles, consists in gluing 7 such triangles in each graph vertex and construct a piecewise surface, shown in Fig.1b. On the scale less than the elementary cell ABC this surface is flat. Thus, the size a ($|AB| = |AC| = |BC| = a$) of the triangle ABC stands for the rigidity parameter, playing the role of a characteristic scale in our problem, below which no deviations from the Euclidean metrics can be found. We rely on small enough values of the parameter a, otherwise, it brings the absence of stretching energy of the tissue into question. Later on we shall see that buckling of growing surface essentially depends on this parameter.

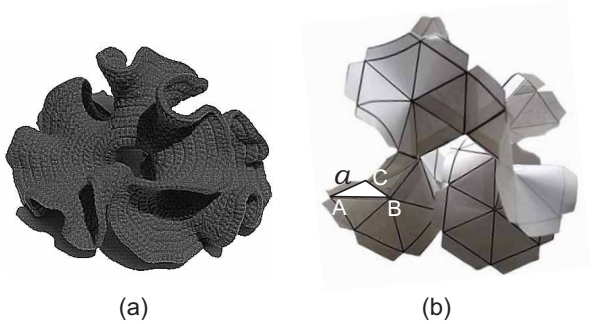

Fig. 1 (a) Hyperbolic surface obtained by spiral crocheting from the center; (b) Hyperbolic piecewise surface constructed by joining 7 equilateral flat triangles (copies of the triangle ABC) in each vertex. The triangle ABC is lying in $z = x + iy$ plane in the 3D Euclidean space, $|AB| = |AC| = |BC| = a$.

We discuss buckling phenomena for two different growth symmetries shown schematically in Fig.2a-b: i) uniform two-dimensional division from the point-like source (Fig.2a), and ii) directed one-dimensional growth from the linear segment (Fig.2b). In Fig.2a-b different generations of cells are shown by the shades of gray. For convenience of perception, sizes of cells in each new generation are decreasing in geometric progression, otherwise it would be impossible to draw them in a 2D flat sheet of paper and the figure would be incomprehensible. In Figs. Fig.2c,d we imitate the protocols of growth depicted above in Figs. Fig.2a,b by embedding the exponentially growing structure in the corresponding plane domain equipped with the hyperbolic metrics. The advantage of such embedding consists in the possibility to continue all functions smoothly through the boundaries of elementary domains, that cover the whole plane without gaps and intersections. Details of this construction and its connection to the growth in the 3D Euclidean space are explained below.

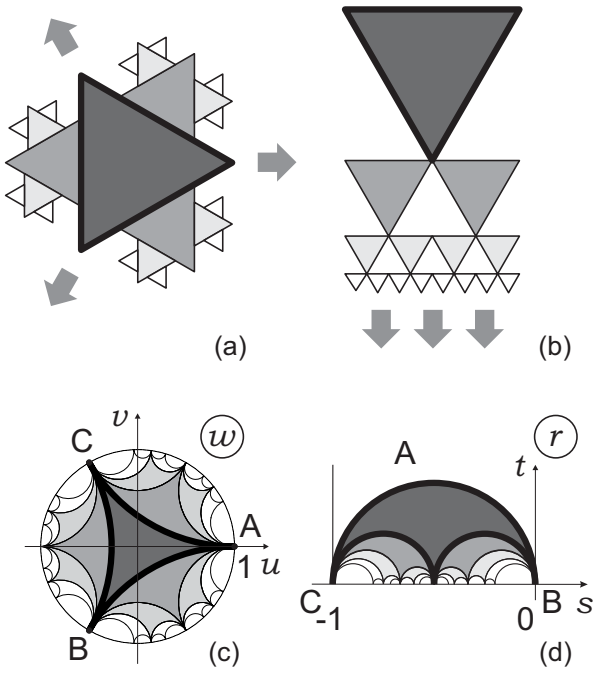

Fig. 2 (a) Uniform two-dimensional hyperbolic growth out of the unit domain in the plane; (b) One-dimensional hyperbolic growth out of the linear segment; c) Tessellation of the hyperbolic Poincaré disc by the images of flat Euclidean triangles; d) Tessellation of the domain in the hyperbolic half-plane by the images of flat Euclidean triangles.

A widely used model (see, for example[17]) suggests that the optimal buckling surface is fully determined by the metric tensor through minimization of a discrete functional of special energetic form. Namely, define the energy of a deformed thin membrane, having buckling profile $f(x,y)$ above the domain, parameterized by (x,y), as:

$$E\{f(x,y)\} \sim \sum_{i,j} \left(\left(f_{ij}\right)^2 - \sum_{\alpha,\beta} \Delta_{ij}^{\alpha} g_{\alpha\beta} \Delta_{ij}^{\beta} \right)^2 \qquad (4)$$

where $g_{\alpha\beta}$ is the induced metrics of the membrane, $f_{ij} \equiv |f(x_i,y_i) - f(x_j,y_j)|$ is the distance between neighboring points and Δ_{ij} is the equilibrium distance between them. The

typical (optimal) shape $\bar{f}(x,y)$ is obtained by minimization of (4) for any rigidity. However, the metric tensor, $g_{\alpha\beta}$ is a priori unknown since its elements depend on specifics of the differential growth protocol, therefore some plausible conjectures concerning its structure should be suggested. For example, in [17] a directed growth of a tissue with one non-Euclidean metrics component, $g_{xx}(y)$, was considered. The diagonal component $g_{xx}(y)$ was supposed to increase exponentially in the direction of the growth, y, and crumpling of a leaf near its edge was finally established and analyzed.

2.3 Conformal approach

The preset rules of uniform exponential cells division determine the structure of the hyperbolic graph, γ, while the infinitesimal membrane thickness allows for the isometrical embedding of the graph γ into the 3D space. We exploit conformal and metric relations between the surface structure in the 3D space and the graph γ embedded into the flat domain with the hyperbolic metrics. The embedding procedure consists of a sequence of conformal transformations with a constraint on area preservation of an elementary plaquette. This eventually yields the knowledge of the Jacobian (the "coefficient of deformation"), $J(x,y)$, for the hyperbolic surface, which is embedded into the 3D space via the orthogonal projection. Equipped by the key assumption, that a smooth yet unknown surface $f(x,y)$ is *function*, our procedure straightforwardly implies a differential equation on the optimal surface. Note, that a version of the crocheted surface cannot be reconstructed in the same way since it is not a function above some planar domain.

To realize our construction explicitly, we first embed isometrically the graph γ: i) into the Poincaré disk ($|w|<1$) for the model of uniform planar growth, and ii) into the strip of the half-plane ($\mathrm{Im}\, r>0, -1<\mathrm{Re}\, r<0$) for the model of one-dimensional growth. In Fig.3 we have drawn the tessellation of the Poincaré disc and of the strip by equilateral curvilinear triangles, which are obtained from the flat triangle *ABC* of the hyperbolic surface (see Fig.1b) by conformal mappings $z(w)$ and $z(r)$ discussed below. Note, that a conformal mapping preserves the angles between adjacent branches of the graph. The graph γ, shown in Fig.3, connects the centers of the triangles and is isometrically embedded into the corresponding hyperbolic domain. Besides, the areas of images of the domain ABC are the same.

For the sake of definiteness consider the graph γ, isometrically embedded into the hyperbolic disk, shown in Fig.3a. Now, we would like to find the surface in the 3D Euclidean space above the w-plane such that its Euclidean metrics coincides with the non-Euclidean metrics in the disk. The Hilbert theorem [35] prohibits to do that for the class of C^2-smooth surfaces. However, since we are interested in the isometric embedding of piecewise surface consisting of glued triangles of fixed area, we can proceed with the standard arguments of differential geometry [36]. The metrics ds^2 of a 2D surface, parameterized by (u,v), is given by the coefficients

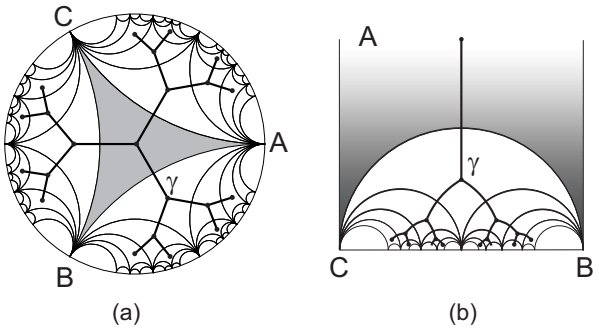

Fig. 3 Tessellation of the hyperbolic plane by the images of the curvilinear triangle *ABC*: (a) for Poincaré disc; (b) for a strip of the upper half-plane. The graph γ connects the centers of images of *ABC*.

$$E=\mathbf{r}_u^2, \quad F=\mathbf{r}_v^2, \quad G=(\mathbf{r}_u,\mathbf{r}_v) \tag{5}$$

of the first quadratic form of this surface:

$$ds^2 = E\,du^2 + 2F\,dudv + G\,dv^2 \tag{6}$$

The surface area then reads $dS=\sqrt{EG-F^2}\,dudv$.

The area S_{ABC} of the planar triangle *ABC* on the plane $z=x+iy$ can be written as:

$$S_{ABC} = \int\limits_{\triangle ABC} dxdy = \mathrm{const} \tag{7}$$

where the integration is restricted by the boundary of the triangle. Since we aimed to conserve the metrics, let us require that the area of the hyperbolic triangle *ABC*, after the conformal mapping, is not changed and, therefore, it reads:

$$S_{ABC} = \int\limits_{\triangle ABC} |J(z,w)|dudv; \quad J(z,w) = \begin{vmatrix} \partial_u x & \partial_u y \\ \partial_v x & \partial_v y \end{vmatrix} \tag{8}$$

where $J(z,w)$ is the Jacobian of transition form z to new coordinates, w. If $z(w)$ is holomorphic function, the Cauchy-Riemann conditions allow to write

$$J(w) = \left|\frac{dz(w)}{dw}\right|^2 \equiv |z'(w)|^2. \tag{9}$$

On the other hand, we may treat the value of the Jacobian, $J(w)$, as a factor relating the change of the surface element under transition to a new metrics, the co-called "coefficient of deformation". Let us note here, that the model of "glued

triangles" should eventually yield the surface from C^1 class in order to be classified as an isometric immersion, since such immersion is allowed by the Nash's theorem, [11]. As long as the metrics in the hyperbollic domain should reproduce the Euclidean metrics of C^1-smooth surface, $f(u,v)$, one should set $J=\sqrt{EG-F^2}$, where E,G,F are the coefficients of the first quadratic form of the surface f. Now, if $f(u,v)$ is function above w-plane, its Jacobian adopts a simple form:

$$J(u,v)=\sqrt{1+(\partial_u f)^2+(\partial_v f)^2} \tag{10}$$

Making use of polar coordinates in our complex w-domain, $\{(\rho,\phi): u=\rho\cos\phi, v=\rho\sin\phi\}$, we eventually arrive at the following nonlinear partial differential equation for the surface profile $f(\rho,\phi)$ above w:

$$\Big(\partial_\rho f(\rho,\phi)\Big)^2+\frac{1}{\rho^2}\Big(\partial_\phi f(\rho,\phi)\Big)^2=|z'(w)|^4-1 \tag{11}$$

In the case of the hyperbolic strip domain, Fig.3b, the equation for the growth profile above the domain can be written in local cartesian coordinates, $r=s+it$:

$$\Big(\partial_s f(s,t)\Big)^2+\Big(\partial_t f(s,t)\Big)^2=|z'(r)|^4-1 \tag{12}$$

Note, that the inequalities $|z'(w)|>1, |z'(r)|>1$, following from (11)-(12), determine the local condition of existence of non-zero real solution and, as we discuss below, can be interpreted as the presence of a finite scale surface rigidity.

To establish a bridge between optic and growth problems, let us mention that, for example, equation (11), coincides with the two-dimensional eikonal equation (3) for the wavefront, $S(w)$, describing the light propagating according the Huygens principle in the unit disk with the refraction coefficient

$$n(w)=\sqrt{|z'(w)|^4-1} \tag{13}$$

We can construct the conformal mappings $z(r)$ and $z(w)$ of the flat equilateral triangle *ABC* in the Euclidean complex plane $z=x+iy$ onto the circular triangle *ABC* in the complex domains $r=s+it$ and $w=\rho(\cos\phi+i\sin\phi)$ correspondingly. The absolute value of the Gaussian curvature is controlled by the number, V, of equilateral triangles glued in one vertex: the surface is hyperbolic only for $V>6$. The surfaces with any $V>6$ have qualitatively similar behavior, however the simplest case for analytical treatment corresponds to $V=\infty$, when the dual graph γ is loopless. The details of the conformal mapping of the flat triangle with side a to the triangle with angles $\{0,0,0\}$ in the unit strip r are given in the Appendix. The Jacobian $J(z(r))$ of conformal mapping $z\to r$ reads:

$$J(r)=|z'(r)|^2=\frac{h^2}{a^2}|\eta(r)|^8 \tag{14}$$

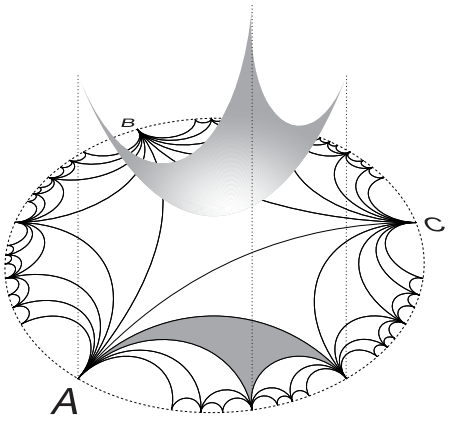

Fig. 4 Orthogonal projection above the Poincaré disc: area of the curvilinear triangle in Euclidean space coincides with the area of the triangle in hyperbolic metrics in Poincaré disc.

and the Jacobian of the mapping $z\to w$, is written through the function $r(w)$ that conformally maps the triangle from the strip onto the Pincaré disk:

$$J(w)=|z'(w)|^2=\frac{3h^2}{a^2}\frac{|\eta(r(w))|^8}{|1-w|^4} \tag{15}$$

where

$$r(w)=e^{-i\pi/3}\frac{e^{2i\pi/3}-w}{1-w}-1;\quad h=\left(\frac{16}{\pi}\right)^{1/3}\frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})}\approx 0.325 \tag{16}$$

In both cases (14) and (15), the function in the right-side of the equation is the Dedekind η-function [25]:

$$\eta(w)=e^{\pi i w/12}\prod_{n=0}^{\infty}(1-e^{2\pi i n w}) \tag{17}$$

3 Results and their interpretation

The eikonal equation, (3), with *constant* refraction index, n, corresponds to optically homogeneous 2D domain, in which the light propagates along straight lines in Euclidean metrics. On the other hand, in this case the eikonal equation yields the action surface with zero Gaussian curvature: a conical surface above the disk, $S(\rho,\phi)\sim\rho$, for the uniform 2D growth and a plane above the strip, $S(s,t)\sim t$, for the directed growth. Note, that at least one family of geodesics of these surfaces consists of lines that are projected to the light propagation paths in the underlying domain. We will show below that the geodesics of the eikonal surface conserve this property even when the media becomes optically inhomogeneous.

For growth, the constant refraction index corresponds to an isometry of a planar growing surface and absence of buck-

ling. The conformal transformation, that results in the corresponding "coefficient of deformation", $J^2(u,v) = n^2(u,v) + 1$, is uniformly compressive and the tissue remains everywhere flat. Thus, it becomes clear, why the essential condition for buckling to appear is the *differential growth*, i.e. the spatial dependence of local rules of cells division.

We solve (11) and (12) numerically with the Jacobian, corresponding to exponentially growing circumference, (14)-(15), for different parameters a. We have chosen the Dirichlet initial conditions along the line (for directed growth above the strip) and along the circle of some small enough radius (for uniform 2D growth above the disk). The right-hand side of the eikonal equations for the specific growth protocol is smooth and nearly constant up some radius and then becomes more and more rugged. The constant plateau in vicinity of initial stages of growth is related with the fact, that exponentially dividing cells can be organized in a Euclidean plane up to some finite generations of growth. However, as the cells proliferate further, the isometry of their mutual disposition becomes incompatible with the Euclidean geometry and buckling of the tissue is observed. Note, that the Jacobian is angular-dependent, that is the artefact of chosen triangular symmetry for the cells in our model. The existence of real solution, $\bar{f}(u,v)$ of the eikonal equation is related to the sign of its right-hand side and is controlled by the parameter a, while the complex solution $f(u,v) = f_R(u,v) + i f_I(u,v)$ can be found for every a.

First, we consider the 2D growth above the Poincaré domain, starting our numerics from low enough values of a, for which the right-hand side of the eikonal equation, (11), is strictly positive on the plateau around the source of growth. Physically that means flexible enough tissues, since, by construction, we require a to be a scale on which the triangulated tissue does not violate flat geometry. The real solution $\bar{f}(u,v)$ for these parameters exists up to late stages of growth, see Fig.5 left. Note, that a conical solution at early stages of growth is related to the plateau in the Jacobian and, as it was discussed above, corresponds to the regime when cells can find places on the surface without violating the flat geometry. From the geometric optics point of view, this corresponds to constant refraction index and straight Fermat geodesic paths in the underlying 2D domain. We show in Fig.5 that under increasing of a the initial area of conical behavior is shrinking, since the critical generation, at which the first buckling mode appears, is lower for larger cells. In course of growth, the surface is getting negatively curved for some angular directions, consistent with chosen triangular symmetry. It is found reminiscent of the shape of bluebells and, in general, many sorts of flowers.

At late stages of growth, as we approach the boundary of the Poincaré disk, $\rho \to 1$ at some fixed value of ϕ, corresponding values of the right hand side of (11) become negative, leading to the complex solution of the eikonal equation. Fortunately, we may infer some useful information from the holomorphic properties of the eikonal equation in this regime, not too close to the boundary of the disc. Applying the Cauchy-Riemann conditions to the solution of the eikonal equation, f, we have: $\partial_u f_R = \partial_v f_I$ and $\partial_v f_R = -\partial_u f_I$. Thus, the function $\bar{f}$ can be analytically continued in the vicinity of points along the curve Γ in the (uv) plane, at which the right hand side of the eikonal equation nullifies. Moreover, using this property, one can show, that the absolute value of the complex solution in the vicinity of Γ smoothly transfers to the real-valued solution, as one approaches the Γ curve:

$$\begin{aligned} \lim_{(u,v)\to\Gamma} (\nabla |f(u,v)|)^2 &= (\nabla f_R(u,v))^2 \,|_\Gamma \equiv \left(\nabla \bar{f}(u,v)\right)^2 \\ |f(u,v)| &= \sqrt{f_R^2(u,v) + f_I^2(u,v)} \end{aligned} \tag{18}$$

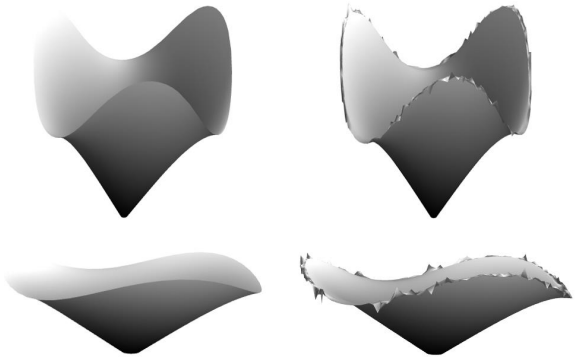

Fig. 5 The bluebell phase, obtained by numerical solution of (11) for two flexible tissues: $a = 0.07$ (first row) and $a = 0.14$ (second row). Figures on the right show appearance of buckling instabilities at the edge with growth.

The non-existence of real solutions of the eikonal equation at late stages is a direct consequence of the presence of finite bending scale, on which the tissue is locally flat. As it was mentioned above and is shown in Fig.5, low values of a lead to elongated conical regime. Since a stands for the scale on which the circumference length of the tissue doubles, in the $a \to 0$ limit the real solution exists everywhere inside the disk, but it is everywhere flat (conical). Hopefully, the analytic continuation allows one to investigate buckling for negative values of $n^2(u,v) = J^2(u,v) - 1$ by taking the absolute value of the solution, at least not far away from the zero-curve Γ. In this regime buckling instabilities on the circumference of the bluebell arise. In Fig.6 we show proliferation of buckling near the critical point. First, the evolution of buckling instabilities at the edge can be understood as a subsequent doubling of peaks

and saddles along the direction of growth. Then some hierarchy in peaks size is seen. We note, that this hierarchical organization is a natural result due to the theoretic-number properties of the Dedekind η-function. Though it is known, that in real plants and flowers buckling instabilities do not proliferate profoundly, since the division process is getting limited at late stages of growth, the formal continuation of the eikonal equation beyond Γ predicts a self-similar buckling profile at the circumference of growing tissues.

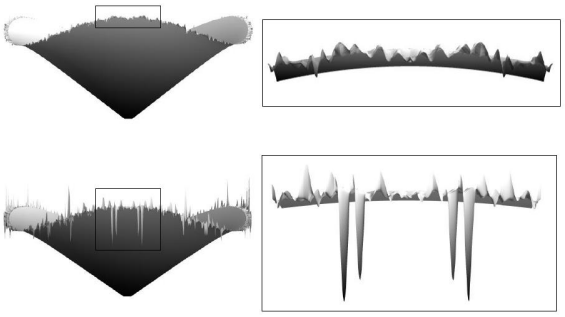

Fig. 6 Development of buckling instabilities at the edge of flower for rigidity parameter $a = 0.14$. The right figures show hierarchical organization of the flower's circumference in detail.

Now we pay attention to the directed growth above the half-plane domain. Here we solve the equation (12) with the Dirichlet boundary conditions, set along the line $t = 1$, and the tissue is growing towards the boundary $t = 0$ in the upper halfplane $\mathrm{Im}\, r > 0$. At low stages of growth the solution is flat until the first buckling mode appear, Fig.7. The subsequent growth is described by taking the absolute value of the solution, since no real solution exists anymore. As in the former case, the behavior is controlled by the value of a.

When the growth approaches the boundary, the edge of the tissue becomes more and more wrinkled. Emergence of new buckling modes is the consequence of the Dedekind η-function properties: doubling of parental peaks at the course of growth. Under the energetic approach for a leaf, very similar fractal structures can be inferred from the interplay between stretching and bending energies in the limit of extremely thin membranes: while the cell density (and the corresponding strain, σ) on the periphery increases, the newly generating wavelengths decrease, $\lambda \sim \sigma^{-1/4}$, [15].

Increasing the size a of the elementary flat triangle domain, we figure out, that for some critical value, a_{cr}, the starting plateau of the corresponding Jacobian crosses the zero level and becomes negative. Our model implies no solutions for such stiff tissues. This limitation is quite natural since we do

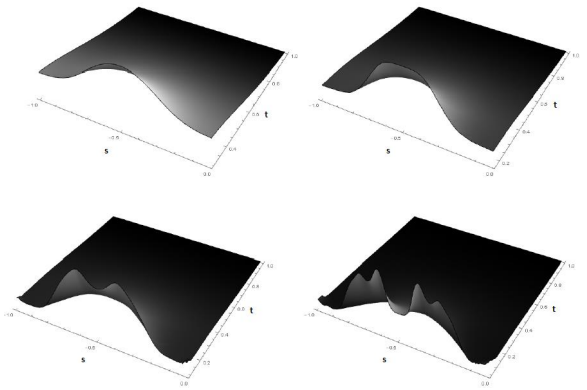

Fig. 7 Numerical solutions of (12) for the directed growth. Figures show enhancing of buckling at the edge.

not consider in-plane deformations of the tissue. In reality, for $a > a_{cr}$ the tissue is so stiff, that it turns beneficial to be squeezed in-plane rather than to buckle out. One may conjecture that a is the analogue of the Young modulus, E, that is known to regulate the rigidity of the tissue in the energetic approach, along with the thickness, h, and the Poisson modulus, μ, in their certain combination, known as bending stiffness, $D = \frac{Eh^3}{12(1-\mu^2)}$.

It is worth mentioning that at first stages of growth, until the instabilities at the circumference have not yet appeared, at certain angles (triangle-like cells) the surface bends similar to the Beltrami's pseudosphere, that has a constant negative curvature at every point of the surface, compare Fig.5 and Fig.8. The similarity is even more striking for very low a, when the triangulating parameter is fairly small. It is known that the pseudosphere locally realizes the Lobachevsky geometry and can be isometrically mapped onto *the finite part* of the half-plane or of the Poincaré disk, Fig.8a-b. According to the Hilbert theorem, [35], no full isometric embedding of the Poincaré disk into the 3D space exists. Thus, in order to organize itself in the 3D space, the plant grows by the cascades of pseudospheres, resembling peaks and saddles, that is an alternative view on essence of buckling. Moreover, it has been shown in the recent work of Gemmer et al., that presence of branch points and lines of inflection lowers the bending energy of the buckling isometry and essentially leads to formation of fractal-like patterns on the edge of a strip with prescribed metric tensor, [13]. These results chord well with our discrete model of glued triangles, where the choice for the metrics is made naturally.

Interestingly, some flowers, such as calla lilies, initially grow psuedospherically, but then crack at some stage of growth and start twisting around in a helix. Apparently, this is another route of dynamic organization of non-Euclidean isom-

etry in the Euclidean space. The Dini's surface, Fig.8c is known in differential geometry as a surface of constant negative curvature and, in comparison with the Beltrami's pseudosphere, is infinite. The problem of sudden cracking of the lilies seems to be purely biological, but as soon as the crack appeared, the flower may relief the stresses caused by subsequent differential growth through twisting its petals in the Dini's fashion.

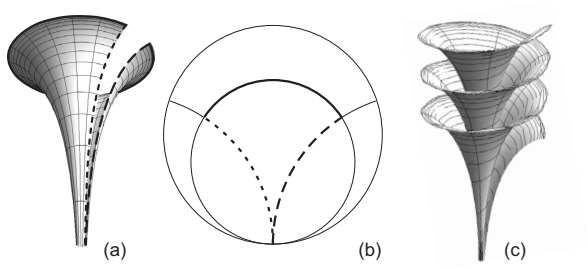

Fig. 8 (a)-(b) Pseudosphere and correspondence of boundaries on the Poincaré disc; (c) Dini surface.

Turn now to the eikonal interpretation of buckling. For the sake of simplicity, we will proceed here in the cartesian coordinates. Seeking the solution of (3) and (12) in the implicit form $H(\mathbf{x}) \equiv H(x^0,x^1,x^2) = H^0$ with $x^0 = if$, $x^1 = u$, $x^2 = v$, we can rewrite (12) as:

$$g^{ij}\frac{\partial H(\mathbf{x})}{\partial x^i}\frac{\partial H(\mathbf{x})}{\partial x^j} = 0; \quad g^{ik} = \begin{pmatrix} n^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{19}$$

Eq.(19) reveals the relativistic nature of the eikonal equation[24] and describes the propagation of light in a (2+1)D space-time in the gravitational field with induced metrics g defined by the metric tensor g^{ik}, where $n \equiv n(x^1,x^2)$, speed of light put $c = 1$. Having g, one can reconstruct geodesics that define the paths of the light propagation in our space-time. The parameterized geodesics family, $x^\lambda(\tau)$, where $\lambda = 0,1,2$, can be found from the equation:

$$\frac{d^2x^\lambda}{d\tau^2} + \Gamma^\lambda_{ij}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau} = 0 \tag{20}$$

where

$$\Gamma^i_{kl} = \frac{1}{2}g^{im}\left(\frac{\partial g_{mk}}{\partial x^l} + \frac{\partial g_{ml}}{\partial x^k} - \frac{\partial g_{kl}}{\partial x^m}\right) \tag{21}$$

are Christoffel symbols and g_{ij} is the covariant form of the metrics ($g_{ij}g^{jk} = \delta^k_i$). Calculating the symbols for the specific metrics (19), we end up with the set of equations for the geodesics in a parametric form:

$$\begin{cases} u_{\tau\tau} - \dfrac{1}{n^3}\dfrac{\partial n}{\partial u}f_\tau^2 = 0, \\ v_{\tau\tau} - \dfrac{1}{n^3}\dfrac{\partial n}{\partial v}f_\tau^2 = 0, \\ f_{\tau\tau} - \dfrac{2}{n}\dfrac{dn}{d\tau}f_\tau = 0 \end{cases} \tag{22}$$

From the first two lines of (22), one gets $\frac{u_{\tau\tau}}{v_{\tau\tau}} = \frac{\partial n}{\partial u}\left(\frac{\partial n}{\partial v}\right)^{-1}$. Note, that the same relation follows directly from (2), if the planar domain is parameterized by the same coordinates $\mathbf{x} = \mathbf{x}(u,v)$. Thus, one may conclude, that the projections of the geodesics from the (2+1)D space-time onto the (uv)-plane coincide with light trajectories in the flat domain with refraction coefficient $n(u,v)$.

4 Conclusion and conjectures

In this paper we discussed the optimal buckling profile formation of growing two-dimensional tissue evoked by the exponential cell division from the point-like source and from the linear segment. Such processes imply excess material generation enforcing the tissue to wrinkle as it approaches the domain boundary. Resulting optimal hyperbolic surface is described by the eikonal equation for the two-dimensional profile, and allows for simple geometric optics analogy. It is shown that the surface height above the domain mimics the eikonal (action) surface of a particle moving in the 2D media with certain refraction index, n, which, in turn, is linked to microscopic rules of elementary cell division and symmetry of the plant. The projected geodesics of this "minimal" optimal surface coincide with Fermat paths in the 2D media, which is the intrinsic feature of the eikonal equation. This result suggests an idea to treat the growth process itself as a propagation of the wavefronts in the media with certain metrics.

We have derived the metrics of the growing plant's surface from microscopic rules of cells division and have shown that the solution of the eikonal equation describes buckling of tissues of different rigidities. Our results, being purely geometric, rhyme well with a number of energetic approaches to buckling of thin membranes, where the stiffness is controlled by the effective bending rigidity. We show that presence of a finite scale on which the tissue remains flat, results in negatively curved growing surfaces and the eikonal equation implies absence of real solution at late stages of the growth. Though, an analytical continuation can be constructed and erratic self-similar patterns along the circumference can be obtained. In reality, there is a biological pressure-governed mechanism that prohibits infinite cell division, thus, intense buckling is rather scarce in flora.

Recall, that the right-hand side of the eikonal equation mimics the squared refraction index, (13), if buckling is interpreted as wavefront propagation in geometric optics. At length of our work it was pointed out, that for the differential growth problem, negative square of refraction index leads to complex solution for f. Does complex solution have any physical meaning for growth? We can provide the following speculation. The complex solution appears for the late stages of growth when the finite bending scale of the tissue prohibits formation of very low-wavelength buckling modes. Since in this regime the tissue would experience in-plane deformations, one may improve the geometric model by letting branches to accumulate the "potential energy". Thereby, the analogy between optics and differential growth can be advanced by noting that the negative squared refraction index means absorbtion properties of the media. The propagating wavefront of a moving particle, dissipates the energy in areas where the refraction index is complex-valued. In the differential growth the proliferation of buckling modes may be limited by the energy losses at branches, that would suppress buckling.

The challenging question concerns the possibility to extend our approach to the growth of three-dimensional objects, for example, of a ball that size R grows faster than R^2. In this case, the redundant material can provoke the surface instabilities. We conjecture that some analogy between the boundary growth and optic wavefronts survives in this case as well.

Authors are grateful to M. Tamm, A. Grosberg, M. Lenz, L. Mirny and L. Truskinovsly for valuable discussions of various aspects of the work and to A. Orlov for invaluable help in numerical solution of the eikonal equation. The work is partially supported by the IRSES DIONICOS and RFBR 16-02-00252A grants.

Appendix: Conformal transformation of the flat triangle to the Poincare domain

The conformal mapping $z(w)$ of the flat equilateral triangle ABC located in z onto the zero-angled triangle ABC in w, used in the derivation of (14), is constructed in four sequential steps, shown in Fig.9.

First, we map the triangle ABC in z onto the upper halfplane ζ of auxiliary complex plane ζ with three branching points at 0, 1 and ∞ – see Fig.9a-b. This mapping is realized by the function $z(\zeta)$:

$$z(\zeta) = \frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})} \int_0^\zeta \frac{d\xi}{\xi^{2/3}(1-\xi)^{2/3}} \tag{23}$$

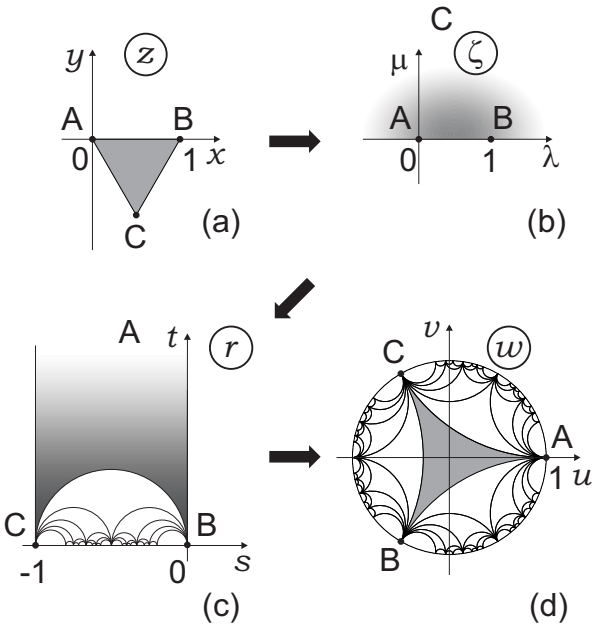

Fig. 9 Conformal mapping $z(w)$ is realized as a composition of three mappings: $z(\zeta)$ [(a)–(b)], $\zeta(r)$ [(b)–(c)], and $r(w)$ [(c)–(d)]. Finally we have $z(\zeta(r(w)))$.

with the following coincidence of branching points:

$$\begin{cases} A(z=0) & \leftrightarrow & A(\zeta=0) \\ B(z=1) & \leftrightarrow & B(\zeta=1) \\ C(z=e^{-i\frac{\pi}{3}}) & \leftrightarrow & C(\zeta=\infty) \end{cases} \tag{24}$$

Second step consists in mapping the auxiliary upper halfplane $\Im\zeta > 0$ onto the circular triangle ABC with angles $\{\alpha,\alpha,0\}$ – the fundamental domain of the Hecke group[26] in r, where we are intersted in the specific case $\{\alpha,\alpha,0\} = \{0,0,0\}$ – see Fig.9b-c. This mapping is realized by the function $\zeta(r)$, constructed as follows[29]. Let $\zeta(r)$ be the inverse function of $r(\zeta)$ written as a quotient

$$r(\zeta) = \frac{\phi_1(\zeta)}{\phi_2(\zeta)} \tag{25}$$

where $\phi_{1,2}(\zeta)$ are the fundamental solutions of the 2nd order differential equation of Picard-Fuchs type:

$$\zeta(\zeta-1)\phi''(\zeta) + \big((a+b+1)\zeta - c\big)\phi'(\zeta) + ab\phi(\zeta) = 0 \tag{26}$$

Following[29,30], the function $r(\zeta)$ conformally maps the generic circular triangle with angles $\{\alpha_0 = \pi|c-1|, \alpha_1 = \pi|a+b-c|, \alpha_\infty = \pi|a-b|\}$ in the upper halfplane of w onto

the upper halfplane of ζ. Choosing $\alpha_\infty = 0$ and $\alpha_0 = \alpha_1 = \alpha$, we can express the parameters (a,b,c) of the equation (26) in terms of α, taking into account that the triangle *ABC* in Fig.9c is parameterized as follows $\{\alpha_0, \alpha_1, \alpha_\infty\} = \{\alpha, \alpha, 0\}$ with $a = b = \frac{\alpha}{\pi} + \frac{1}{2}, c = \frac{\alpha}{\pi} + 1$. This leads us to the following particular form of equation (26)

$$\zeta(\zeta-1)\phi''(\zeta) + \Big(\frac{\alpha}{\pi}+1\Big)\big(2\zeta-1\big)\phi'(\zeta) + \Big(\frac{\alpha}{\pi}+\frac{1}{2}\Big)^2\phi(\zeta) = 0 \quad (27)$$

where $\alpha = \frac{\pi}{m}$ and $m = 3,4,...\infty$. For $\alpha = 0$ Eq.(27) takes an especially simple form, known as Legendre hypergeometric equation[31,32]. The pair of possible fundamental solutions of Legendre equation are

$$\begin{aligned} \phi_1(\zeta) &= F\big(\tfrac{1}{2},\tfrac{1}{2},1,\zeta\big) \\ \phi_2(\zeta) &= iF\big(\tfrac{1}{2},\tfrac{1}{2},1,1-\zeta\big) \end{aligned} \quad (28)$$

where $F(...)$ is the hypergeometric function. From (25) and (28) we get $r(\zeta) = \frac{\phi_1(\zeta)}{\phi_2(\zeta)}$. The inverse function $\zeta(r)$ is the so-called modular function, $k^2(r)$ (see[31?,32] for details). Thus,

$$\zeta(r) \equiv k^2(r) = \frac{\theta_2^4(0,e^{i\pi r})}{\theta_3^4(0,e^{i\pi r})} \quad (29)$$

where θ_2 and θ_3 are the elliptic Jacobi θ-functions[33?],

$$\begin{aligned} \theta_2\left(\chi, e^{i\pi w}\right) &= 2e^{i\frac{\pi}{4}r}\sum_{n=0}^{\infty} e^{i\pi r n(n+1)}\cos(2n+1)\chi \\ \theta_3\left(\chi, e^{i\pi r}\right) &= 1 + 2\sum_{n=1}^{\infty} e^{i\pi r n^2}\cos 2n\chi \end{aligned} \quad (30)$$

and the correspondence of branching points in the mapping $\zeta(r)$ is as follows

$$\begin{cases} A(\zeta=0) & \leftrightarrow & A(r=\infty) \\ B(\zeta=1) & \leftrightarrow & B(r=0) \\ C(\zeta=\infty) & \leftrightarrow & C(r=-1) \end{cases} \quad (31)$$

Third step, realized via the function $r(w)$, consists in mapping the zero-angled triangle *ABC* in r into the symmetric triangle *ABC* located in the unit disc w – see Fig.9c-d. The explicit form of the function $r(w)$ is

$$r(w) = e^{-i\pi/3}\frac{e^{2i\pi/3} - w}{1-w} - 1 \quad (32)$$

with the following correspondence between branching points:

$$\begin{cases} A(r=\infty) & \leftrightarrow & A(w=1) \\ B(r=0) & \leftrightarrow & B(w=e^{-2\pi i/3}) \\ C(r=-1) & \leftrightarrow & C(w=e^{2\pi i/3}) \end{cases} \quad (33)$$

Collecting (23), (29), and (32) we arrive at the following expression for the derivative of composite function,

$$z'(\zeta(r(w))) = z'(\zeta)\,\zeta'(r)\,r'(w) \quad (34)$$

where $'$ stands for the derivative. We have explicitly:

$$z'(\zeta) = \frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})}\,\frac{\theta_3^{16/3}(0,\zeta)}{\theta_2^{8/3}(0,\zeta)\;\theta_0^{8/3}(0,\zeta)}$$

and

$$\zeta'(r)| = i\pi\frac{\theta_2^4\,\theta_0^4}{\theta_3^4}; \qquad i\frac{\pi}{4}\theta_0^4 = \frac{d}{d\zeta}\ln\left(\frac{\theta_2}{\theta_3}\right)$$

The identity

$$\theta_1'(0,e^{i\pi\zeta}) \equiv \left.\frac{d\theta_1(\chi,e^{i\pi\zeta})}{d\chi}\right|_{\chi=0} = \pi\theta_0(\chi,e^{i\pi\zeta})\,\theta_2(\chi,e^{i\pi\zeta})\,\theta_3(\chi,e^{i\pi\zeta})$$

enables us to write

$$\left|z'(r)\right|^2 = h^2\left|\theta_1'\left(0,e^{i\pi r}\right)\right|^{8/3} \quad (35)$$

where $h = \left(\frac{16}{\pi}\right)^{1/3}\frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})}$, and

$$\theta_1(\chi,e^{i\pi r}) = 2e^{i\frac{\pi}{4}r}\sum_{n=0}^{\infty}(-1)^n e^{i\pi n(n+1)r}\sin(2n+1)\chi \quad (36)$$

Differentiating (32), we get

$$r'(w) = \frac{i\sqrt{3}}{(1-w)^2}$$

and using this expression, we obtain the final form of the Jacobian of the composite conformal transformation $J(z(\zeta(r(w))))$:

$$J(z(w)) = |z'(w)|^2 = 3h^2\frac{|\eta(r(w))|^8}{|1-w|^4} \quad (37)$$

where

$$\eta(r) = \left(\theta_1'(0,e^{i\pi r})\right)^{1/3}$$

is the Dedekind η-function (see (15)), and the function $r(w)$ is defined in (32).

References

1 M. A. R. Koehl, W.K. Silk, H. Liang, L. Mahadevan, *Integrative and Comparative Biology*, 2008, **48**, 834.

2 E. Sharon, B. Roman, and H. L. Swinney, *Physical Review E*, 2007, **75**, 046211.
3 D.W. Henderson, D. Taimina, *The Mathematical Intelligencer*, 2001, **23**, 17.
4 S. Nechaev, R. Voituriez, *Journal of Physics A: Math. Gen.*, 2001, **34**, 11069.
5 S. Nechaev, O. Vasilyev, *Journal of Physics A: Math. Gen.*, 2004, **37**, 3783.
6 V. Borrelli, S. Jabrane, F, Lazarus, and B. Thibert, *Proceedings of the National Academy of Sciences*, 2012, **109**, 7218.
7 E. Efrati, E. Sharon, R. Kupferman, 2009, *Journal of the Mechanics and Physics of Solids*, **57**, 762.
8 M. Lewicka, L. Mahadevan, M.R. Pakzad, *Proceedings of Royal Society A*, 2011, **467**, 402.
9 B. Audoly, A. Boudaoud, *Physical Review Letters*, 2003, **91**, 086105.
10 Lewicka M., Pakzad M. R., *ESAIM: Control, Optimisation and Calculus of Variations*, 2011, **17**, 1158.
11 Kuiper N. H. *Indagationes Mathematicae (Proceedings)*, 1955, **58**, 683.
12 J. Gemmer, S.C. Venkataramani, *Soft Matter*, 2013, **9**, 8151.
13 J. Gemmer, E. Sharon, T. Shearman, S. C. Venkataramani, 2016, *Europhysics Letters*, **114**, 24003.
14 J. W. S. Rayleigh, *The theory of sound*, Dover, New York, 1945.
15 E. Cerda, K. Ravi-Chandar, L. Mahadevan, *Nature*, 2002, **419**, 579.
16 Nath, U., Crawford, B. C., Carpenter, R., Coen, E., *Science*, 2003, **299**, 1404.
17 E. Sharon, B. Roman, M. Marder, G.-S. Shin, and H. L. Swinney, *Nature*, 2002, **419**, 579.
18 H. Liang, L. Mahadevan, *Proceedings of the National Academy of Sciences*, 2009, **106**, 22049.
19 A. Goriely, M. Ben Amar, *Phisical Review Letters*, 2005, **94**, 198103.
20 N. Stoop et al., *Physical Review Letters*, 2010, **105**, 068101.
21 M. Marder et al., *Europhysics Letters*, 2003, *62*, 498.
22 M. Marder, N. Papanicolaou, *Journal of Statistical Physics*, 2006, **125**, 1065.
23 L. S. Levitov, *Europhysics Letters*, 1991, **14**, 533.
24 L.D. Landau, E.M. Lifshitz, *Course of theoretical physics, Theory of elasticity*, Pergamon Press, Oxford, 1986.
25 K. Chandrassekharan, *Elliptic Functions*, Springer, Berlin, 1985.
26 Li-Chien Shen, *The Ramanujan Journal*, 2016, **39**, 609.
27 F. Klein, *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 1890, **1**, 35.
28 Y.B. Rumer, *Uspekhi Matematicheskikh Nauk*, 1953, **8**, 55.
29 W. Koppenfels, F. Stallman, *Praxis der konformen Abbildung*, Springer, Berlin, 1959.
30 C. Carathéodory, *Theory of functions, vol. II*, Chelsea Pub. Company, New York, 1954.
31 V.V. Golubev, *Lectures on the analytic theory of differential equations*, Gostekhizdat, Moscow, 1950
32 E. Hille, *Ordinary differential equations in the complex plane*, J. Willey & Sons, New York, 1976.
33 D. Mumford, C. Musili, *Tata lectures on theta II*, Springer, Berlin, 2007.
34 M.H. Amsler, *Mathematische Annalen*, 1955, **130**, 234.
35 D. Hilbert, *Transactions of the American Mathematical Society*, 1901, **2**, 87.
36 G.M. Fihtengolc, *Kurs Diferencialnovo i Integralnovo Iscislenia I*, Nauka, Moscow, 1966.

4. Anomalous one-dimensional fluctuations of a simple two-dimensional random walk in a large-deviation regime

Introduction

The following question is the subject of our work: could a two-dimensional (2D) random path pushed by some constraints to an improbable "large-deviation regime" possess extreme statistics with one-dimensional (1D) Kardar-Parisi-Zhang (KPZ) fluctuations? The answer is positive, though non-universal, since the fluctuations depend on the underlying geometry. We consider in detail two examples of 2D systems for which imposed external constraints force the underlying stationary stochastic process to stay in an atypical regime with anomalous statistics. The first example deals with the fluctuations of a stretched 2D random walk above a semicircle or a triangle. In the second example we consider a 2D biased random walk along a channel with forbidden voids of circular and triangular shapes. In both cases we are interested in the dependence of a typical span $\langle d(t)\rangle \sim t^{\gamma}$ of the trajectory of t steps above the top of the semicircle or the triangle. We show that $\gamma = 1/3$, i.e., $\langle d(t)\rangle$ shares the KPZ statistics for the semicircle, while $\gamma = 0$ for the triangle. We propose heuristic derivations of scaling exponents γ for different geometries, justify them by explicit analytic computations, and compare with numeric simulations. For practical purposes, our results demonstrate that the geometry of voids in a channel might have a crucial impact on the width of the boundary layer and, thus, on the heat transfer in the channel.

Contribution

I have participated in derivation of the scaling relations for different curved geometries and in analytic calculation of the probability distribution of a large fluctuation for the semi-circle setting.

Anomalous 1D fluctuations of a simple 2D random walk in a large deviation regime

Sergei Nechaev[1,2], Kirill Polovnikov[3,4], Senya Shlosman[4,5,6],
Alexander Valov[7], and Alexander Vladimirov[5]

[1] *Interdisciplinary Scientific Center Poncelet, CNRS UMI 2615, 119002 Moscow, Russia*

[2] *P.N. Lebedev Physical Institute RAS, 119991 Moscow, Russia*

[3] *Physics Department, Lomonosov Moscow State University, 119992 Moscow, Russia*

[4] *Skolkovo Institute of Science and Technology, 143005 Skolkovo, Russia*

[5] *Institute of Information Transmission Problems RAS, 127051 Moscow, Russia*

[6] *Aix-Marseille University, Universite of Toulon, CNRS, CPT UMR 7332, 13288, Marseille, France*

[7] *N.N. Semenov Institute of Chemical Physics RAS, 119991 Moscow, Russia*

The following question is the subject of our work: could a two-dimensional random path pushed by some constraints to an improbable "large deviation regime", possess extreme statistics with one-dimensional Kardar-Parisi-Zhang (KPZ) fluctuations? The answer is positive, though non-universal, since the fluctuations depend on the underlying geometry. We consider in details two examples of 2D systems for which imposed external constraints force the underlying stationary stochastic process to stay in an atypical regime with anomalous statistics. The first example deals with the fluctuations of a stretched 2D random walk above a semicircle or a triangle. In the second example we consider a 2D biased random walk along a channel with forbidden voids of circular and triangular shapes. In both cases we are interested in the dependence of a typical span $\langle d(t)\rangle \sim t^{\gamma}$ of the trajectory of t steps above the top of the semicircle or the triangle. We show that $\gamma = \frac{1}{3}$, i.e. $\langle d(t)\rangle$ shares the KPZ statistics for the semicircle, while $\gamma = 0$ for the triangle. We propose heuristic derivations of scaling exponents γ for different geometries, justify them by explicit analytic computations and compare with numeric simulations. For practical purposes, our results demonstrate that the geometry of voids in a channel might have a crucial impact on the width of the boundary layer and, thus, on the heat transfer in the channel.

I. INTRODUCTION

Intensive investigation of extremal problems of correlated random variables in statistical mechanics has eventually led mathematicians, and then, physicists, to understanding that the Gaussian distribution is not as ubiquitous in nature, as it has been thought over the centuries, and shares its omnipresence (at least in one dimension) with another distribution, known as the Tracy-Widom (TW) law. The necessary (though not sufficient) feature of the TW distribution is the width of the distribution, controlled by the critical exponent $\nu = \frac{1}{3}$, the so-called Kardar-Parisi-Zhang (KPZ) exponent. For the first time, the KPZ exponent has appeared in the seminal paper [1] (see [2] for review) as the growth exponent in a non-equilibrium one-dimensional directed stochastic process, for which the theoretical analysis

has been focused mainly on statistical properties of the enveloping surface developing in time.

Nowadays one has accumulated many examples of one-dimensional statistical systems of seemingly different physical nature, whose fluctuations are controlled by the KPZ exponent $\gamma = \frac{1}{3}$, contrary to the exponent $\gamma = \frac{1}{2}$ typical for the distribution of independent random variables. Among such examples it is worth mentioning the restricted solid-on-solid [3] and Eden [4] models, molecular beam epitaxy [5], polynuclear growth [6–10], several ramifications of the ballistic deposition [11–14], alignment of random sequences [15], traffic models of TASEP type [16], (1+1)D vicious walks [17], area-tilted random walks [18], and 1D directed polymer in random environment [19]. Recently, this list has been replenished by the one-dimensional modes describing the fluctuational statistics of cold atoms [20].

Here we study a two-dimensional model demonstrating the one-dimensional KPZ critical behavior. The interest to such systems is inspired by the (1+1)D model proposed by H. Spohn and P. Ferrari in [21] where they discussed the statistics of 1D directed random walks evading the semicircle. As the authors stated in [21], their motivation was as follows. It is known that the fluctuations of a top line in a bunch of n one-dimensional directed "vicious walks" glued at their extremities (ensemble of world lines of free fermions in 1D) are governed by the Tracy-Widom distribution [17]. Proceeding as in [22], define the averaged position of the top line and look at its fluctuations. In such a description, all vicious walks lying below the top line, play a role of a "mean field" of the "bulk", pushing the top line to some equilibrium position. Fluctuations around this position are different from fluctuations of a free random walk in absence of the "bulk". Replacing the effect of the "bulk" by the semicircle, one arrives at the Spohn-Ferrari model where the 1D directed random walk stays above the semicircle, and its interior is inaccessible for the path. In [21] the authors confirmed that this system has a KPZ critical exponent.

In our work we study fluctuations of a two-dimensional random path pushed by some geometric constraints to an improbable "large deviation regime" and ask the question whether it could possess extreme statistics with one-dimensional Kardar-Parisi-Zhang (KPZ) fluctuations. We propose the "minimal" model and in its frameworks formulate the answer to the question posed above.

We consider an ensemble of two-dimensional random paths stretched over some forbidden void with prescribed geometry and characteristic scale, R. Stretching is induced by the restriction on wandering times, t, such that $cR < t \ll R^2$. The resulting paths conformations are "atypical" since their realizations would be highly improbable in the ensemble of unconstrained trajectories which exhibit the Gaussian behavior. Statistics in such a tiny subset of the Gaussian ensemble is naturally controlled by collective behavior of strongly correlated modes, thus, for some geometries one might expect extreme distribution with KPZ scaling for fluctuations, similarly to the (1+1)D model of [21]. Simple dimensional analysis supports this hypothesis. Indeed, consider a realization of the stretched random walk in 2D with the diffusion coefficient D evading a circular void in two distinct regimes. An unconstrained t-step random walk, with $t \gg R^2$ fluctuates freely and does not feel the constraint, thus, the only possible combination of D and t, which has the dimension of length, could be $d \sim (Dt)^{1/2}$ for the typical span of the path. In the opposite regime, $\pi R < t \ll R^2$, the chain statistics is essentially perturbed by the constraint. In the limit of strong stretching, $t \sim R$, these two parameters (t and R) should enter symmetrically in the combination for the span. The suitable dimension is given by the scaling expression $d \sim (DRt)^{\gamma}$ with $\gamma = 1/3$,

which is the unique combination that in the limit $t \ll R^2$ recovers a physically relevant condition $d \ll R$ and at $t \sim R^2$ gives $d \sim R$. Such a dimensional analysis strongly relies on the uniqueness of the scale, characterizing constraint, which is true only for homogeneously curved boundaries and breaks down for more complex algebraic curves, like cubic parabola or boundaries with a local cusp (triangle). In particular, trajectories above triangular obstacles fluctuate irrespectively to the size of the void even in the "strong stretching regime".

The paper is organized as follows. in Section II we formulate the model of a 2D stretched random walk above the semicircle (model "S") and the triangle (model "T") and provide scaling arguments for the averaged span of paths above the top of these voids, supported by numeric simulations. in Section III we solve the diffusion equation in 2D in the limit of stretched trajectories $N = cR$ above the semicircle and the triangle. in Section IV we discuss the results of numeric simulations for fluctuations of biased 2D random walks above forbidden voids of different shapes. in Section V we summarise the obtained results and discuss their possible generalizations and applications.

II. TWO-DIMENSIONAL RANDOM WALK STRETCHED OVER THE VOIDS OF VARIOUS SHAPES

A. The model

We begin with the lattice version of the model. Consider the N-step symmetric random walk, $\mathbf{r}_n = \{x_n, y_n\}$, on a two-dimensional square lattice in a discrete time n ($n = 1, 2, ..., N$). The walk begins at the point A, terminates after N steps at the point B, and satisfies three requirements: (i) for any n one has $y_n \geq 0$, (ii) the random walk evades the semicircle of the diameter $2R$, or the rectangular triangle of the base $2R$, i.e. it remains outside the obstacles shown in Fig. 1 and (iii) the total number of steps is much less than the squared size of the obstacle, $N \ll R^2$. Note that the requirement (i) is not crucial and can be easily relaxed. The points A and B are located in one lattice spacing from left and right extremities of the obstacle (semicircle or triangle) – see Fig. 1.

We are interested in the critical exponents γ of in the dependence $\langle d(R) \rangle \sim R^\gamma$ as $R \to \infty$ for the model "S" and the model "T". In this section we provide qualitative scaling estimates for the mean span of two-dimensional stretched paths above any smooth algebraic curve and support our analysis by numeric simulations.

B. Scaling arguments: from semicircle to algebraic curve

Normally, a stretched path follows the straight line as much as possible, and gets curved only if curving cannot be avoided. A random path which has to travel a horizontal distance, x_S, is localized within a strip of typical width ("span" in a vertical direction), $y_S \sim \sqrt{x_S}$. If the path is forced to travel a distance x_S along some curved arc, and the arc fits this strip, the curving of the arc can be ignored. Consider a path that has to follow a circle of radius R. Note that the arc of that circle of length x_S fits a strip of width x_S^2/R. Therefore the arc length, curving of which can be ignored, is

$$x_S^2/R \leq \sqrt{x_S} \tag{1}$$

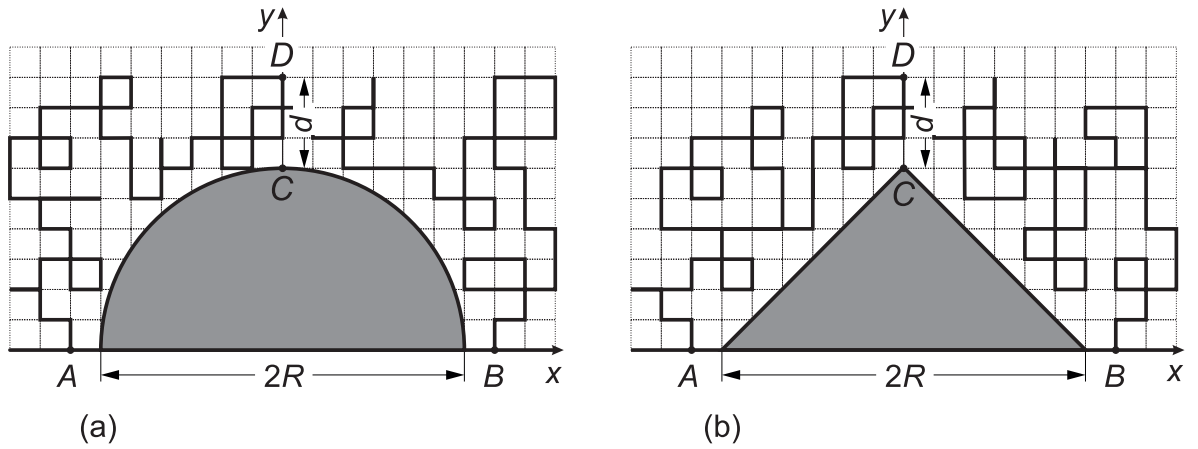

FIG. 1: Two-dimensional random walk on a square lattice in the upper half-plane, that evades: (a) Model "S": the semicircle of radius R, and (b) Model "T": the rectangular triangle of base $2R$. The number of steps $N \ll R^2$.

This puts a limit to x_S: it has to be at most $R^{2/3}$. At shorter distances the stretched path can be considered as an unconstrained random walk. Therefore, the span in vertical direction is of the order of $y_S \sim \sqrt{R^{2/3}} = R^{1/3}$. Beyond this "blob" of length $x_S = R^{2/3}$ the arc itself deviates considerably from a straight segment, and the estimate $\sqrt{x_S}$ for fluctuations above it is no longer applicable.

To add some geometric flavor to these arguments, consider Fig. 2a and denote by y_S an average span of the path in vertical direction above the point C of the semicircle, and by x_S – the typical size of the horizontal segment, along which the semicircle can be considered as nearly flat. We divide the path in three parts: AA', $A'B'$ and $B'B$. The parts AA' and BB' of the trajectory run above essentially curved domains, while the part $A'B'$ constitutes a segment that is mainly flat. Schematically this is shown in Fig. 2b: in the limit $y_S \ll R$, the horizontal segment LM linearly approximates the corresponding arc of the circle. Our goal is to estimate x_S and to provide self-consistent scaling arguments for fluctuations $y_S(R) \sim R^\gamma$ of the stretched path.

From the triangle KLM we have:

$$|LM| = \sqrt{R^2 - |KM|^2} = \sqrt{R^2 - (R - y_S)^2}\,\Big|_{y_S \ll R} \approx \sqrt{2Ry_S} \tag{2}$$

Since $|LM| \equiv x_S$, the condition of stretched trajectories, $y_S \ll R$, implies the relation

$$x_S \sim \sqrt{Ry_S} \tag{3}$$

Consider now a two-dimensional random walk which starts at the point L near the left extremity of the excluded shape and terminates anywhere at the segment MN ($|MN| \equiv y_S$). Since the horizontal support, $|LM| = x_S$, of the path is flat, the span of the trajectory in vertical direction is the same as for an ordinary random walk. Thus, we can estimate the typical span, y_S, as

$$y_S \sim \sqrt{x_S} \tag{4}$$

On the scales larger than x_S the curvature of the semicircle becomes essential and the relation (3) is not valid anymore.

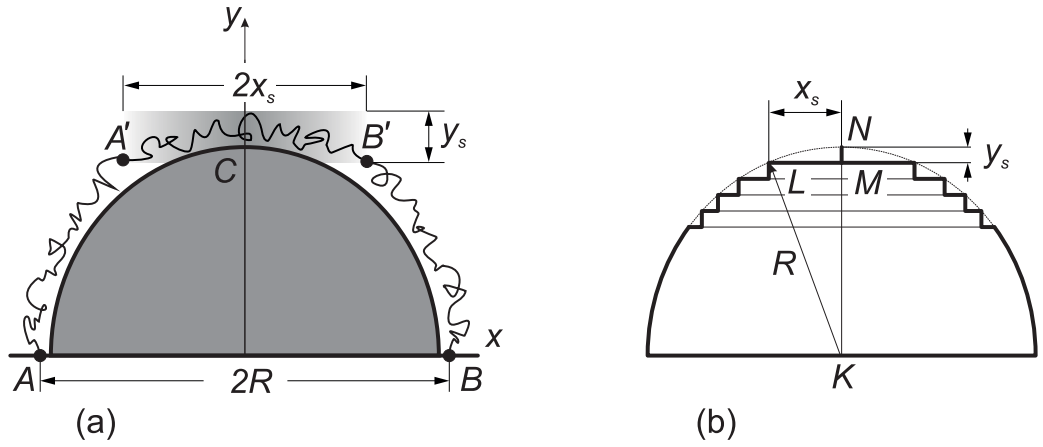

FIG. 2: (a) Two-dimensional random walk evading the semicircle. The part $A'B'$ lies above the essentially flat region of the semicircle. The figure (b) provides an auxiliary geometric construction for Eq.(2).

It should be noted that (4) is insensitive to a specific way of stretching. Eq (4) remains unchanged even if we introduce an asymmetry in random jumps along x–axis while keeping the symmetry of jumps in y direction. Substituting the scaling (4) into (2), we obtain for the semicircle (the model "S"):

$$x_S \sim \sqrt{R\sqrt{x_S}} \tag{5}$$

From the first equation of (5) we get for the semicircle:

$$x_S \sim R^{2/3}; \quad y_S \sim \sqrt{x_S} \sim R^{1/3} \tag{6}$$

which implies that $\gamma = \frac{1}{3}$. The analytic computations presented in Section III for the model "S" support this conclusion. Let us note that the large-scale deviation principle for the constrained 1D random walk process has been discussed recently in [23].

We expect that our scaling can be extended to random walks above any algebraic curve. The critical exponent γ for the fluctuations of the stretched random walk above the curve Γ: $y = x^\eta$ in 2D should be understood as follows. Define the characteristic length scale, R, and represent the curve Γ in dimensionless units:

$$\frac{y}{R} \approx \left(\frac{x}{R}\right)^\eta \tag{7}$$

For $\eta = 2$ we are back to semicircle (3). As in the former case, Eq. (7) should be equipped by (4). Solving these equations self-consistently, we get the following scaling dependence for the span $y_G(R)$ of the path above the curve Γ:

$$y_G(R) \sim R^\gamma; \qquad \gamma = \frac{\eta - 1}{2\eta - 1} \tag{8}$$

Note that for $\eta \to \infty$ the curve is straight and we get the fluctuations with the standard Gaussian exponent, $\gamma = 1/2$, which is the exponent of fluctuations above the straight line. The opposite case of a cusp can be approached in the limit $\eta \to 1$, which gives $\gamma = 0$. This result rhymes well with simulations of paths stretched over the triangle (see below) and analytic solution of the diffusion equation (Section III).

C. Heuristic arguments: triangle

To estimate the fluctuations of the path of N steps stretched over the triangle of base $2R$, the above arguments for the semicircle need to be modified since the curvature of the triangle is non-analytic being concentrated at one single point C at the tip of the obstacle. To proceed, some auxiliary construction should be used – see Fig. 3a and its zoom in Fig. 3b.

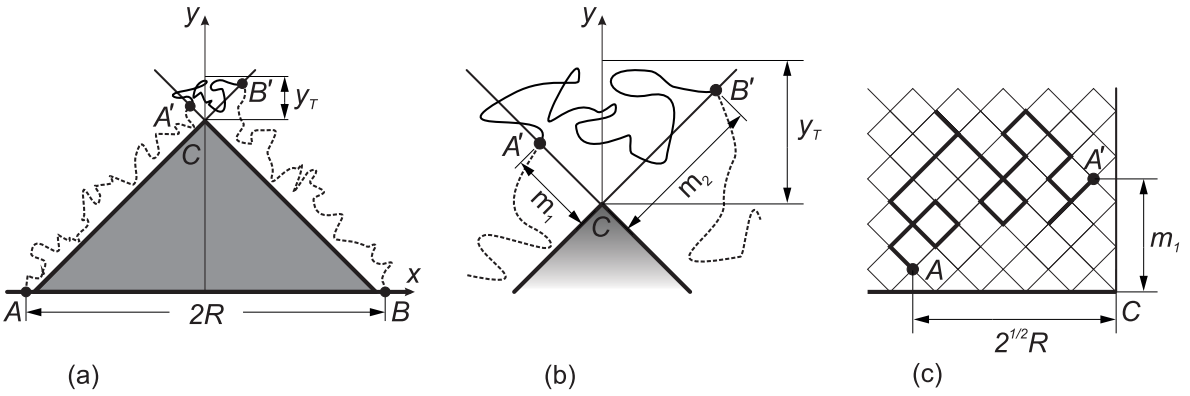

FIG. 3: (a) Two-dimensional random walk evading a triangle, and (b) the magnified part of the system near the tip of the triangle. The points A' and B' are respectively the points of the first entry by the random walk into the wedge above the point O and the last exit from it; (c) subpart of the random walk from A to A' which does not escape the wedge with zero's boundary conditions.

We split the full trajectory between points A and B into three parts: the part of N_1 steps running between point A and first entry to the point A', the part of M steps running between the points A' and B', and the part N_2 running between B and first entry to the point B'. The parts N_1 and N_2 lie above the flat boundaries of the triangle AOB, while the part $A'B'$ is located in the vicinity of the tip of the triangle. The partition function, Z_N, of the full N-step path with the extremities at A and B can be written as follows:

$$Z_N(R) = \sum_{\{N_1+M+N_2=N\}} \sum_{\{m_1,m_2\}} U_{N_1}(m_1,R)\, W_M(m_1,m_2)\, U_{N_2}(m_2,R) \tag{9}$$

where $U_{N_1}(m_1,R)$, $W_M(m_1,m_2)$, $U_{N_2}(m_2,R)$ are, respectively, the partition functions of parts AA', $A'B'$ and $B'B$, the first sum runs over N_1, M, N_2 such that $N_1+M+N_2=N$ and m_1 and m_2 are the positions of the points A' and B' at the edges of the wedge (see Fig. 3b). The partition functions $U_{N_i}(m_i,R)$ $(i=1,2)$ can be computed on the lattice in the geometry shown in Fig. 3c with zero's boundary conditions in the wedge

$$U_{N_i}(m_i,R) = \frac{1}{\pi^2}\int_0^\pi dq_1 \int_0^\pi dq_2 \sin(q_1 R\sqrt{2}) \sin q_1 \sin(q_2 m_i) \sin q_2 (\cos q_1 + \cos q_2)^{N_i} \tag{10}$$

where q_1 and q_2 are the Fourier-transformed coordinates along the wedge sides. In (10) the subpath of N_i steps is not yet stretched, i.e. N_i, m_i and R are all independent.

Our goal now is to estimate the typical length M of the subpath between the points A' and B' as shown in Fig. 3b. Below we show that $M = \text{const}$ which immediately leads to

the conclusion that $y_T = \text{const}$. To proceed, it is convenient to pass to the grand canonical formulation of the problem. Let us define the generating function $Z(s,R) = \sum_{N=0}^{\infty} Z_N(R) s^N$ of the grand canonical ensemble, and introduce the variable $\beta = -\ln s$, which has the sense of an "energy" attributed to each step of the trajectory (note that $\beta > 0$ since $0 < s < 1$). To "stretch" the trajectory, we should imply $\beta \gg 1$. In the stretched regime $\beta \gg 1$ the generating function of $U_{N_i}(m_i, R)$ can be estimated as follows

$$U(\beta, m_i, R) = \int_0^{\infty} U_{N_i}(m_i, R) e^{-\beta N_i} dN_i \sim \frac{m_i R\, \beta_s^{3/4} \exp\left(-2\sqrt{\beta_s}\,\sqrt{m_i^2 + 2R^2}\right)}{(m_i^2 + 2R^2)^{5/4}} \qquad (11)$$

where we also supposed that $R \gg 1$ and introduced $\beta_s = \beta - \ln 4$. The shift by $\ln 4$ in β comes from the fact that the partition function (10) on the square lattice has the exponential prefactor $4^{N_i} \equiv e^{N_i \ln 4}$ which should be properly taken into account in the generation function.

The generation function of $Z_N(R)$ reads:

$$Z(\beta, R) = \sum_{N=0}^{\infty} Z_N(R) s^N = \sum_{\{m_1, m_2\}} U(\beta, m_1, R)\, W(\beta, m_1, m_2)\, U(\beta, m_2, R) \qquad (12)$$

Now we should account for the contribution of $W(\beta, m_1, m_2)$ to (12). Note, that each step of the path of length M between points A' and B' carries the energy $\beta > 0$. To maximize the contribution of $W(\beta, m_1, m_2)$, one should make the corresponding length M between A' and B' as small as possible, since we loose the energy βM for M steps. Thus, M should be of order of $\max(m_1, m_2)$. From (11)–(12) we immediately conclude that at $\beta_s \gg 1$ the major contribution to $Z(\beta)$ comes from m_i which should be as small as possible, i.e. $m_1 \sim m_2 = \text{const}$. This immediately implies that $M = \text{const}$ and the span y_T (for $N = cR$ and $R \gg 1$) becomes independent on R:

$$y_T = \text{const} \qquad (13)$$

The same conclusion follows from the solution of the boundary problem in the open wedge for the model "T" – see Section IV. Note, that putting $\eta = 1$ into (8), we get $\gamma = 0$, thus arriving at the same conclusion of independence of the span of fluctuations of stretched path above the tip of the triangle on R.

D. Numerics

Here we confirm our scaling and heuristic analyses of the mean height of the 2D ensemble of stretched trajectories above the top of the semicircle and the triangle using numeric simulations. Let us emphasize that this part pursues mainly the illustrative goals, while detailed analytic computations for distribution functions are provided in the following Section III.

Specifically, we have enumerated all N-step paths on the square lattice, travelling from the point $A(-R-1, 0)$ to the point $D(0, R+d)$ above the top of the semicircle or triangle, as shown in Fig. 1a,b. Let us emphasize that this is an exact path counting problem. The step length of a path coincides with the lattice spacing. We allow all steps: "up",

"down","right", "left" and set the constraint $N = cR$ on the total number of steps. The values of R and c in the simulations are as follows: $R = \{10, 20, 40, 60, 100, 200, 300, 400\}$ and $c = \{5, 10, 20\}$. Counting ensemble of trajectories from A to D is sufficient for extracting the scaling dependence $\langle d(R)\rangle \sim R^{\gamma}$ since the part of the path from A to C is independent from the part from C to B. The enumeration of trajectories respects boundary conditions and is performed recursively within the box of size $3R \times 3R$ with the bottom left corner located at the point $(-2R, 0)$.

The results of simulations in doubly-logarithmic scale $\log \langle d(R)\rangle$ *vs* $\log R$ for the averaged span $\langle d\rangle$ of paths above the top of the semicircle of radius R and the triangle of base $2R$ are presented in Fig. 4. The physical meaning of the constant c is the effective "stretching" of the path: the less c, the more stretched the path (definitely, on the square lattice $c > 4$.

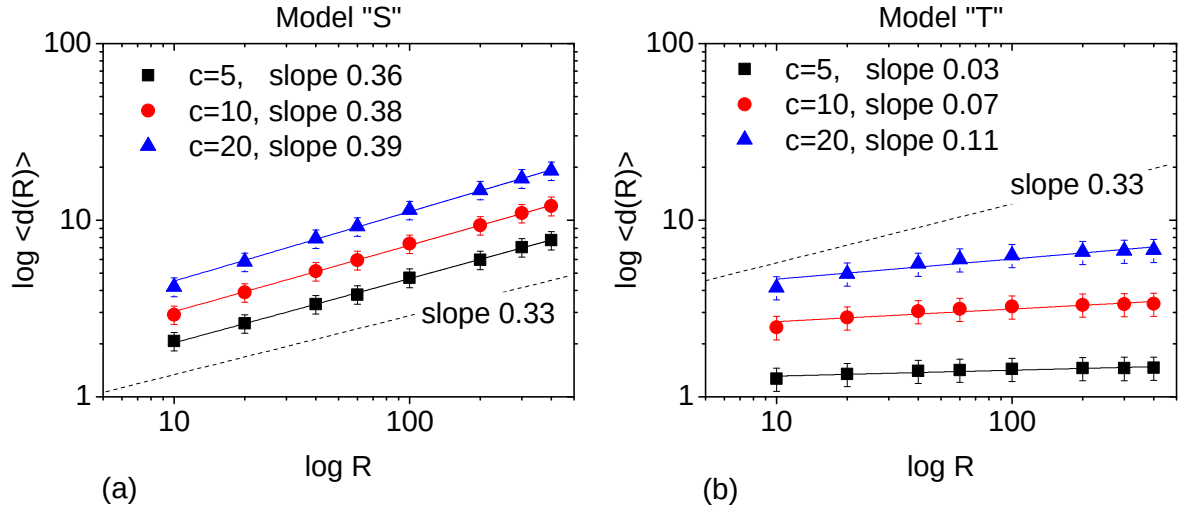

FIG. 4: The mean deviation of the path of N steps above the semicircle (a) and the triangle (b) for different values of the parameter c, which controls "stretching" of the path (the less c the more stretched the path).

As one sees from Fig. 4a, all stretched paths above the semicircle demonstrate the scaling $\langle d(R)\rangle \sim R^{\gamma}$ with the exponent γ close to $1/3$. For less stretched paths (larger values of c) the deviation form the scaling with $\gamma = \frac{1}{3}$ becomes notable. The span of stretched 2D trajectories above the tip of the triangle shown in Fig. 4b are almost independent on R (i.e. the exponent γ is close to 0). This result is consistent with our scaling estimates, as well as with the theoretical arguments presented below. Some conjectures about possible physical consequences of the difference between fluctuations of stretched random trajectories above the semicircle and above the triangle are formulated in Section IV.

III. 2D STRETCHED RANDOM WALKS ABOVE THE SEMICIRCLE AND TRIANGLE: ANALYTIC RESULTS

A. Semicircle

The symmetric two-dimensional random walk on a lattice depicted in Fig. 1a in the limit $N \to \infty$, $a \to 0$ (where a is the lattice spacing) where $Na = t$, converges to the two-dimensional Brownian motion of time t with diffusion coefficient $D = \frac{a^2}{4}$, that evades the semicircular void of radius R. Let $P(\rho, \phi; \rho_0, \phi_0; t)$ be the probability density to find the random walk of length (time) t at the point (ρ, ϕ) above the void under the condition that the path begins at the point (ρ_0, ϕ_0). The function $P(\rho, \phi; \rho_0, \phi_0; t) \equiv P(\rho, \phi, t)$ satisfies the diffusion equation in polar coordinates

$$\begin{cases} \dfrac{\partial P(\rho,\phi,t)}{\partial t} = D\left[\dfrac{1}{\rho}\dfrac{\partial}{\partial \rho}\left(\rho \dfrac{\partial P(\rho,\phi,t)}{\partial \rho}\right) + \dfrac{1}{\rho^2}\dfrac{\partial^2 P(\rho,\phi,t)}{\partial \phi^2}\right] \\ P(\rho = R, \phi, t) = P(\rho \to \infty, \phi, t) = P(\rho, \phi = 0, t) = P(\rho, \phi = \pi, t) = 0 \\ P(\rho, \phi, 0) = \delta(\rho - \rho_0)\delta(\phi - \phi_0) \end{cases} \tag{14}$$

The explicit solution of (14) reads

$$P(\rho, \phi, t) = \sum_{k=1}^{\infty} \frac{2\rho_0}{\pi} \sin(k\phi_0)\sin(k\phi) \int_0^{\infty} e^{-\lambda^2 Dt} Z_k(\lambda\rho, \lambda R) Z_k(\lambda\rho_0, \lambda R) \lambda d\lambda \tag{15}$$

where

$$Z_k(\lambda\rho, \lambda R) = \frac{-J_k(\lambda\rho)N_k(\lambda R) + J_k(\lambda R)N_k(\lambda\rho)}{\sqrt{J_k^2(\lambda R) + N_k^2(\lambda R)}} \tag{16}$$

and J and N denote correspondingly the Bessel and the Neumann functions. Introducing the new variables, μ and r, and making in (16) the substitution

$$\lambda = \frac{\mu}{R}, \qquad \rho = R + r, \tag{17}$$

we arrive at the following expression for $P(\rho, \phi, t)$:

$$P(r, \phi, t) = \frac{2\rho_0}{\pi R^2} \sum_{k=1}^{\infty} \sin(k\phi_0)\sin(k\phi) \int_0^{\infty} e^{-\frac{\mu^2 Dt}{R^2}} Z_k\left(\mu + \frac{\mu r}{R}, \mu\right) Z_k\left(\mu + \frac{\mu r_0}{R}, \mu\right) \mu d\mu \tag{18}$$

The probability to stay above the top of the semicircle consists of two parts: the probability $P' = P\left(r, \phi = \frac{\pi}{2}, t'\right)$ to run from the point A to the point $(r, \phi = \frac{\pi}{2})$ during the time t' and the probability $P'' = P\left(r, \phi = \frac{\pi}{2}, t''\right)$ to run from the point $(r, \phi = \frac{\pi}{2})$ to the point B during the time $t'' = t - t'$. Obviously, P' and P'' are independent, thus the total probability to find path at the point $(r, \phi = \frac{\pi}{2})$ above the semicircle can be estimated as $Q = P' \times P''$ where $t' = t'' = t/2$, namely

$$Q\left(r, \phi = \frac{\pi}{2}, t\right) = \frac{1}{\mathcal{N}} P^2\left(r, \phi = \frac{\pi}{2}, t = cR\right); \qquad \mathcal{N} = \int_0^{\infty} P^2\left(r, \phi = \frac{\pi}{2}, t\right) dr \tag{19}$$

Recall that we are interested in *stretched* trajectories only, meaning that we should impose the condition $t = cR$ and consider the typical width, $d(R)$ of the distribution $Q(r, R)$, where $d^2(R)$ is defined as follows:

$$\left\langle d^2(R) \right\rangle = \int_0^\infty r^2\, Q\left(r, \phi = \frac{\pi}{2}, cR\right) dr - \left(\int_0^\infty r\, Q\left(r, \phi = \frac{\pi}{2}, cR\right) dr \right)^2 \qquad (20)$$

at large R. By the condition $t = cR$ to deal with stretched trajectories, our consideration differs from the standard diffusion process above the impenetrable disc, which was exhaustively discussed in many papers, for example, in [24]. In the figure Fig. 5 we have plotted (for $D = 1$):

(a) The expectation $\bar{d}(R) = \sqrt{\langle d^2(R) \rangle}$ as a function of R in doubly logarithmic coordinates which enables us to extract the critical exponent γ in the dependence $\bar{d}(R) \sim R^\gamma$ (Fig. 5a),

(b) The distribution function $Q\left(r, \phi = \frac{\pi}{2}, cR\right)$ of r at some fixed c ($c = 5$) and R in comparison with the function $b\,\mathrm{Ai}^2(a_1 + \ell r)$, where $\mathrm{Ai}(z) = \frac{1}{\pi}\int_0^\infty \cos(\xi^3/3 + \xi z)\, d\xi$ is the Airy function (see, for example, [25]), $a_1 \approx -2.3381$ is the first zero of Ai, $b = \left[\int_0^\infty \mathrm{Ai}^2(a_1 + \ell r) dr\right]^{-1}$, and $\ell(c)$ is the c-dependent numeric constant (Fig. 5b).

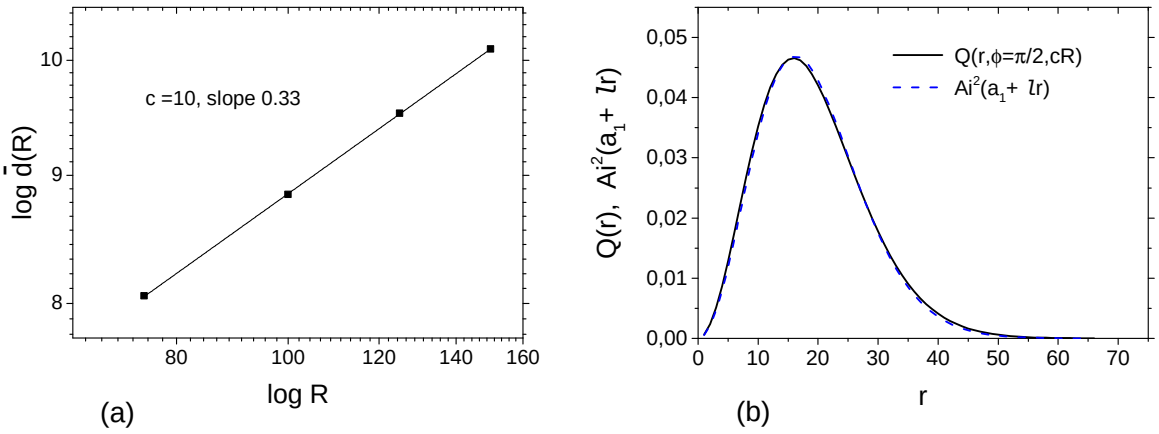

FIG. 5: (a) Expectation $\bar{d}$ as a function of R is doubly logarithmic coordinates for stretched trajectories above the semicircle; (b) Comparison of the distribution $Q(r)$ with $\mathrm{Ai}^2(a_1 + \ell r)$ for the radius of a semicircle $R = 100$, where $a_1 \approx -2.3381$ is the first zero of Ai and $\ell \approx 0.0811$.

As one sees from Fig. 5, the function $\mathrm{Ai}^2(a_1 + \ell r)$ perfectly matches the probability distribution $Q\left(r, \phi = \frac{\pi}{2}, cR\right)$. The detailed analysis of this correspondence is postponed to the paper [26], which will be devoted to the discussion of the statistics of closed stretched random "flights" above the circle.

B. Triangle

The statistics of random paths above the triangle can be treated in polar coordinates centered at the tip C of the triangle as shown in Fig. 1b. The random walk is free in the

outer sector ACB with the angle $\frac{3\pi}{2}$ and zero boundary conditions at the sides AC and BC are applied. Seeking the solution for the corresponding diffusion equation in the form $P(r,v,t) = T(t)\mathcal{P}(r,v)$, we have:

$$\begin{cases} \nu^2\mathcal{P}(r,v) + \left(\partial^2_{rr} + \dfrac{\partial_r}{r} + \dfrac{\partial^2_{vv}}{r^2}\right)\mathcal{P}(r,v) = 0 \\ \mathcal{P}(r=0,v) = \mathcal{P}(r\to\infty,v) = \mathcal{P}\left(r,0\right) = \mathcal{P}\left(r,\dfrac{3\pi}{2}\right) = 0 \\ \partial_t T(t) + \nu^2 D T(t) = 0 \end{cases} \tag{21}$$

Separating variables, we can write $\mathcal{P}(r,v) = Q(r)V(v)$ and get a set of coupled eigenvalue problems for the "angular", v, and "radial", r, coordinates.

$$\begin{cases} \partial^2_{vv}V(v) + \lambda_n^2 V(v) = 0 \\ V\left(0\right) = V\left(\frac{3\pi}{2}\right) = 0 \end{cases} \quad ; \quad \begin{cases} \left(r^2\partial^2_{rr} + r\partial_r + \left(\nu^2 r^2 - \lambda_n^2\right)\right)Q(r) = 0 \\ Q(r=0) = Q(r\to\infty) = 0 \end{cases} \tag{22}$$

The particular solutions to the "angular" and "radial" boundary problems read as follows:

$$\begin{cases} V_n \propto \sin\left(\frac{2nv}{3}\right) \\ Q_n \propto J_{\frac{2n}{3}}(\nu r) \end{cases} \tag{23}$$

The function $P(r,v,t)$ can be written now as follows:

$$P(r,v,t) = \sum_{n=1}^{\infty}\int_0^{\infty} A_n(\nu)J_{\frac{2n}{3}}(\nu r)\sin\left(\frac{2nv}{3}\right)e^{-\nu^2 Dt}d\nu \tag{24}$$

where constants $A_n(\nu)$ satisfy the initial conditions:

$$\sum_{n=1}^{\infty}\int_0^{\infty} A_n(\nu)J_{\frac{2n}{3}}(\nu r)\sin\left(\frac{2nv}{3}\right)d\nu = \delta(r-R)\delta(v-v_0) \tag{25}$$

and

$$A_n(\nu) = \frac{4R}{3\pi}\sin\left(\frac{2nv_0}{3}\right)\nu J_{\frac{2n}{3}}(\nu R) \tag{26}$$

Rewrite the sum in (24) as follows:

$$P(r,v,t) = \sum_{n=1}^{\infty}\frac{4R}{3\pi}\sin\left(\frac{2nv_0}{3}\right)\sin\left(\frac{2nv}{3}\right)\int_0^{\infty}\nu J_{\frac{2n}{3}}(\nu R)J_{\frac{2n}{3}}(\nu r)e^{-\nu^2 Dt}d\nu \tag{27}$$

Evaluating the integral in (27):

$$\int_0^{\infty}\nu J_{\frac{2n}{3}}(\nu R)J_{\frac{2n}{3}}(\nu r)e^{-\nu^2 Dt}d\nu = \frac{1}{2Dt}e^{-\frac{r^2+R^2}{4Dt}}I_{\frac{2n}{3}}\left(\frac{rR}{2Dt}\right) \tag{28}$$

we arrive finally at the following expression for the probability distribution:

$$P(r,v,t)=\frac{4R}{3\pi}\frac{1}{2Dt}e^{-\frac{r^2+R^2}{4Dt}}\sum_{n=1}^{\infty}\sin\left(\frac{2nv_0}{3}\right)\sin\left(\frac{2nv}{3}\right)I_{\frac{2n}{3}}\left(\frac{rR}{2Dt}\right) \tag{29}$$

Consider a conditional probability distribution for the trajectory passing from A to B above the triangle through the point D:

$$P(A\to D\to B)=\frac{P(A\to D)P(B\to D)}{\int_0^{\infty}P(A\to D)P(B\to D)dr} \tag{30}$$

where $P(X\to D)$ is the probability to run from the point X to the point $D(d,\frac{3\pi}{4})$ above the tip of the triangle. The sum in (28) has the following asymptotic behavior

$$\sum_{n=1}^{\infty}\sin\left(\frac{2nv_0}{3}\right)\sin\left(\frac{n\pi}{2}\right)I_{\frac{2n}{3}}(x)\sim xe^{-x^{6/7}} \tag{31}$$

Collecting (29)–(31), we find the behavior of $\langle d\rangle$ for $t=cR$

$$\langle d\rangle=\int_0^{\infty}rP(A\to D\to B)dr\sim\text{const} \tag{32}$$

which means that the fluctuations of stretched trajectories above the tip C of the triangle are bounded and do not depend on R. This result supports the simple scaling consideration exposed in Section II.

IV. BIASED 2D RANDOM WALKS IN A CHANNEL WITH FORBIDDEN VOIDS

As a further development of the problem of 2D random walk statistics above the semicircle and triangle, we numerically consider an ensemble of 2D random walks with a horizontal drift in a presence of forbidden voids of different shapes, as it is shown in Fig. 6. The setting of this model slightly differs from the one discussed above. We regard an ensemble of long trajectories ($t\gg R$) starting at the point A located to the left from the semicircle of the triangle, however we do not fix the terminal point of the path, allowing it to be everywhere. Instead of controlling the lengths of the path, t, we have fixed the value of the horizontal drift, ε. Thus, the coordinates of the tadpole of a growing lattice path obey the following recursive transformations:

$$(x_{t+1},y_{t+1})=\begin{cases}(x_t-1,y_t) & \text{with probability } \frac{1}{4}-\varepsilon\\ (x_t+1,y_t) & \text{with probability } \frac{1}{4}+\frac{\varepsilon}{3}\\ (x_t,y_t+1) & \text{with probability } \frac{1}{4}+\frac{\varepsilon}{3}\\ (x_t,y_t-1) & \text{with probability } \frac{1}{4}+\frac{\varepsilon}{3}\end{cases} \tag{33}$$

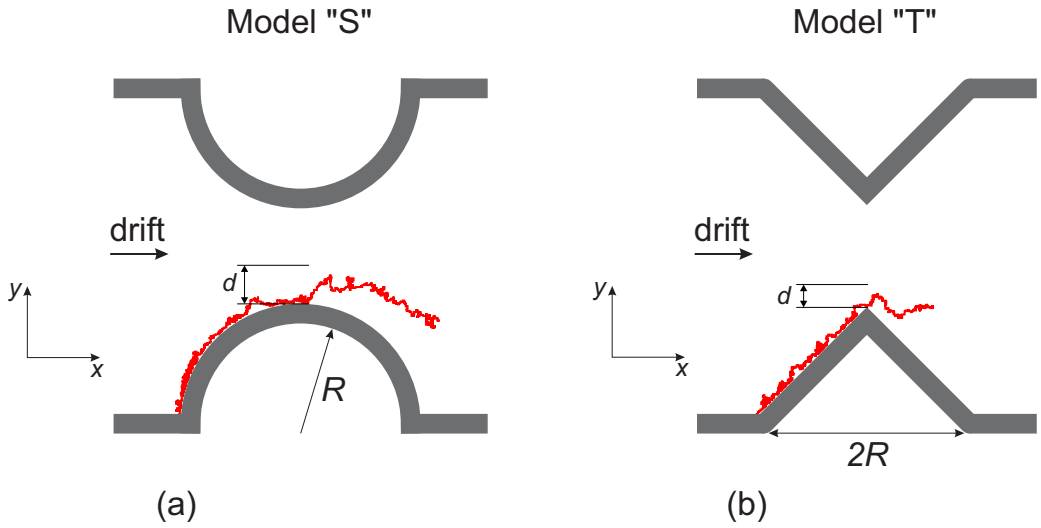

FIG. 6: Biased 2D random walk in a channel with forbidden voids in a form of semicircle (a) and a triangle (b).

at $\varepsilon = 0$ we return to the symmetric two-dimensional random walk, while at $\varepsilon = \frac{1}{4}$ the backward steps are completely forbidden.

We have performed Monte-Carlo simulations to determine the fluctuations of 2D trajectories with the drift ε ($\varepsilon \geq 0$) above the top of the semicircle (triangle). The corresponding results are presented in Fig. 7 for $\varepsilon = \frac{3}{28}$, for which the quotient of forward to backward horizontal jump rates is equal to 2. In the case of a semicircle, the KPZ scaling for the expectation, $\langle d(R) \rangle \sim R^{1/3}$, holds, while for the case of the triangle the fluctuations do not depend on R, and the behavior $\langle d(R) \rangle = \text{const}$ is clearly seen. We have simulated of order of 10^3 lattice trajectories up to the length $t_{max} = 2 \times 10^3$ in the presence of voids characterized by $R = \{250, 500, 750, 1000, 1250, 1500\}$ (measured in the units of lattice spacing). Thus, the statistics of biased 2D random walks in presence of forbidden voids of semicircular and triangular shapes matches the fluctuations of stretched 2D random walks above the same shapes discussed at length of the Section II.

Found behavior of biased random walks in vicinity of excluded voids of various shapes, allows us to make a conjecture about possible thermodynamic properties of laminar flows in tubes with periodic contractions. The combination of the drift and geometry pushes the laminar flow lines which spread near the boundary, into a large deviation regime with the extreme value statistics, typical for 1D systems with spatial correlations. Since the width of the fluctuational (skin) layer near the boundary is shape-dependent, one may expect different heat emission of laminar flows in presence of excluded voids of different geometries.

V. DISCUSSION

In this work we considered simple two-dimensional systems in which imposed external constraints push the underlying stochastic processes into the "improbable" (i.e. large deviation) regime possessing the anomalous statistics. Specifically, we dealt with the fluctuations of a two-dimensional random walk above the semicircle and the triangle in a special case of

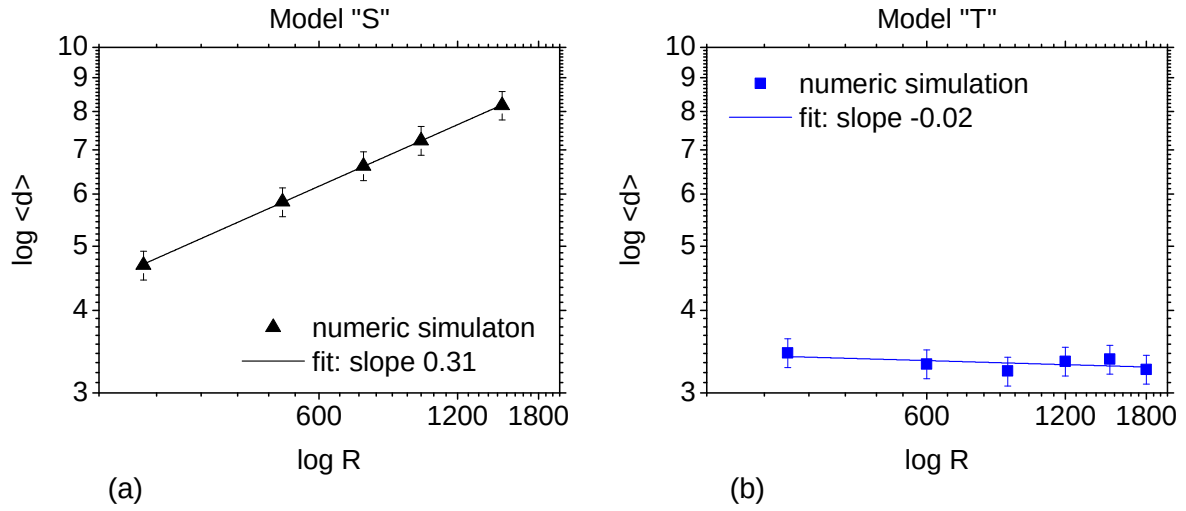

FIG. 7: Mean deviations of open random paths shown in Fig. 6 for $\varepsilon = \frac{3}{28}$: (a) above the top of the semicircle; (b) above the tip of the triangle.

"stretched" trajectories. We proposed the simple scaling arguments supported by the analytic consideration. As a brief outline of the results, it is worth highlighting three important points:

- Imposing constraints on a conformational space, which cut off a tiny region of available ensemble of trajectories, we can push the sub-ensemble of random walks into the atypical large deviation regime possessing anomalous fluctuations, which could have some similarities with the statistics of correlated random variables;
- Stretching 2D random paths above the semicircle, we may effectively reduce the space dimension: in specific geometries we force the system to display the 1D KPZ fluctuations;
- Strong dependence of the fluctuation exponent γ on the geometry of the excluded area, manifests the non-universality in the underlying reduction of the dimension. We outline three archetypical geometries: stretching above the plane (Gaussian, with $\gamma = 1/2$), above the semicircle (KPZ-type, with $\gamma = 1/3$) and above the triangle or the cusp (finite, with $\gamma = 0$). For an algebraic curve of order η the fluctuation exponent is $\gamma = \frac{\eta-1}{2\eta-1}$.

Our results demonstrate that geometry has a crucial impact on the width of the boundary layer in which the laminar flow lines diffuse. We could speculate that such an effect is important for some technical applications in rheology of viscous liquids, for instance, for cooling of laminar flows in channels with periodically displaced excluded voids of various shapes (like shown in Fig. 5). Such a conjecture is based on the following obvious fact. The heat transfer through walls depends not only on the total contact surface of the flow with the wall, but also on a width of a mixing skin layer: the bigger a mixing layer near the boundary, the better cooling. However as we have seen throughout the paper, the width of the mixing layer is shape-dependent, and hence, it might control the "optimal" channel geometry for cooling of laminar viscous liquids flows.

The 1D KPZ-type behavior in a 2D restricted random walk goes far beyond the pure academic interest. Two important relevant applications should be mentioned. First, by this model we provide an explicit example of the two-dimensional statistical system which, being pushed to the large-deviation ("atypical") region, mimics the behavior of some one-dimensional correlated stochastic process. Second, our study deals with the manifestations of a 1D KPZ-type scaling in the localization phenomena of 2D constrained disordered systems. Namely, let us estimate the free energy, $F(N)$ of an ensemble of N-step paths stretched above the semicircle as shown in Fig. 8a.

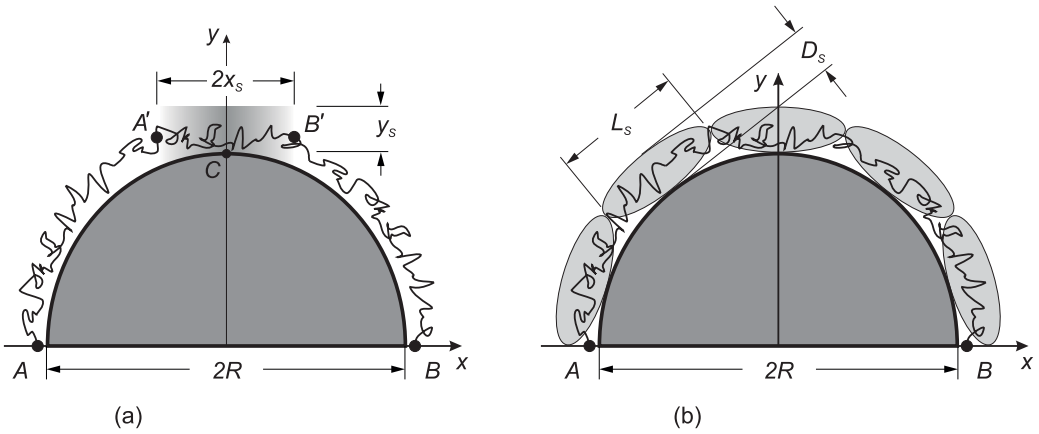

FIG. 8: (a) Two-dimensional random walk evading the semicircle. The part $A'B'$ lies above the essentially flat region of the semicircle; (b) splitting in blobs of a trajectory evading curved surface (semicircle).

One can split the entire stretched path of length N running from A to B above the semicircle into the sequence of independent "blobs" with the longitudinal size $L_S = x_S \sim R^{2/3}$ and the transversal size $D_S = y_s \sim R^{1/3}$ – see Fig. 8b. Thus, taking into account the additive character of the free energy, we can estimate $F(N)$ of ensemble of $N = cR$–step paths as

$$F(R) \sim \frac{N}{L_S} \sim \frac{R}{R^{2/3}} \sim R^{1/3} \tag{34}$$

Therefore, the Gibs measure, which provides expression of the "survival probability" in the curved channel of length $N \sim R$ and diameter $\sim R^{1/3}$, can be estimated as follows

$$P(R) = e^{-F(R)} \sim e^{-\alpha R^{1/3}} \tag{35}$$

where α is some model-dependent numerical constant. Passing to the grand canonical formulation of the problem, i.e. attributing the energy E to each step of the path (remembering that $N = cR$), one can rewrite the expression for $P(R)$ in (35) as follows

$$P(E) = \int_0^\infty P(R)\, e^{-ER}\, dR \sim \varphi(E)\, e^{-b/\sqrt{E}} \tag{36}$$

where $b = \frac{2\alpha^{3/2}}{3^{3/2}}$ and $\varphi(E)$ is a power-law function of E.

To provide some speculations behind the behavior (36), recall that the density of states, $r(E)$, of the 1D Anderson model (the tight-binding model with the randomness on the

main diagonal) at $E \to 0$, has the asymptotics (36), known as the "Lifshitz singularity", $r(E) \sim e^{-a/\sqrt{E}}$, where E is the energy of the system and a is some positive constant (see [27, 28] for more details).

The asymptotics (35), has appeared in the literature under various names, like "stretched exponent", "Griffiths singularity", "Balagurov-Waks trapping exponent", however, as mentioned in [29], in all cases this is nothing else as the inverse Laplace-transformed Lifshitz tail of the one-dimensional disordered systems possessing Anderson localization (36). We claim that the KPZ-type behavior with the critical exponent $\gamma = \frac{1}{3}$ can also be regarded as an incarnation of a specific "optimal fluctuation in a large deviations regime" for the one-dimensional Anderson localization. Finding in some 2D systems a behavior typical for 1D localization, seems to be a challenging problem of connecting localization in constrained 2D and 1D systems. In details this issue will be discussed in a forthcoming publication.

Acknowledgments

We are grateful to V. Avetisov, A. Gorsky, A. Grosberg, B. Meerson, S. Pirogov and M. Tamm for number of fruitful discussions and useful critical remarks. The work of S.N. is partially supported by the RFBR grant No. 16-02-00252; K.P. acknowledges the support of the Foundation for the Support of Theoretical Physics and Mathematics "BASIS" (grant 17-12-278); The work of A. Vladimirov was supported by RFBR grant 16-29-09497. The work of A. Valov was performed within frameworks of the state task for ICP RAS 0082-2014-0001 (registration #AAAA-A17-117040610310-6). The work of S.S. was supported by the RFBR–CNRS grant No. 17-51-150006.

[1] M. Kardar, G. Parisi, and Y.-C. Zhang, Dynamic Scaling of Growing Interfaces, Phys. Rev. Lett. **56** 889 (1986)

[2] T. Halpin-Healy and Y.-C. Zhang, Kinetic roughening phenomena, stochastic growth, directed polymers and all that, Physics Reports **254** 215 (1995)

[3] J.M. Kim and J. M. Kosterlitz, Growth in a restricted solid-on-solid model, Phys. Rev. Lett. **62** 2289 (1989)

[4] F. Family and T. Vicsek, Scaling of the active zone in the Eden process on percolation networks and the ballistic deposition model, J. Phys. A: Math. Gen. **18** L75 (1985)

[5] M.A. Herman and H. Sitter, *Molecular Beam Epitaxy: Fundamentals and Current* (Springer: Berlin, 1996)

[6] P. Meakin, *Fractals, scaling, and growth far from equilibrium* (Cambridge University Press: Cambridge, 1998)

[7] M. Prähofer and H. Spohn, Universal Distributions for Growth Processes in 1+1 Dimensions and Random Matrices, Phys. Rev. Lett. **84** 4882 (2000)

[8] M. Prähofer, H. Spohn, Scale Invariance of the PNG Droplet and the Airy Process, J. Stat. Phys. **108** 1071 (2002)

[9] J. Baik and E.M. Rains, Limiting distributions for a polynuclear growth model with external sources, J. Stat. Phys. **100** 523 (2000)

[10] K. Johansson, Discrete Polynuclear Growth and Determinantal Processes, Comm. Math. Phys. **242** 277 (2003)
[11] B.B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982)
[12] P. Meakin, P. Ramanlal, L. M. Sander, and R. C. Ball, Ballistic deposition on surfaces, Phys. Rev. A **34** 5091 (1986)
[13] J. Krug and P. Meakin, Microstructure and surface scaling in ballistic deposition at oblique incidence, Phys. Rev. A **40** 2064 (1989)
[14] D. Blomker, S. Maier-Paape, and T. Wanner, Interfaces and Free Boundaries **3** 465 (2001)
[15] S.N. Majumdar and S. Nechaev, Exact asymptotic results for the Bernoulli matching model of sequence alignment, Phys. Rev. E **72** 020901(R) (2005); S.N. Majumdar, K. Mallick, and S. Nechaev, Bethe Ansatz in the Bernoulli Matching Model of Random Sequence Alignment, Phys. Rev. E **77** 011110 (2008)
[16] B. Derrida, M. R. Evans, V. Hakim, and V. Pasquier, Exact solution of a 1D asymmetric exclusion model using a matrix formulation, J. Phys. A: Math. Gen. **26** 1493 (1993)
[17] G. Schehr, S.N. Majumdar, A. Comtet, and J. Randon-Furling, Exact Distribution of the Maximal Height of p Vicious Walkers, Phys. Rev. Lett. **101** 150601 (2008)
[18] D. Ioffe, S. Shlosman, and Y. Velenik, An invariance principle to Ferrari-Spohn diffusions, Commun. Math. Phys. **336** 905 (2015)
[19] V. Dotsenko, Bethe ansatz derivation of the Tracy-Widom distribution for one-dimensional directed polymers, Europhys. Lett. **90** 20003 (2010); V. Dotsenko, Bethe ansatz replica derivation of the GOE TracyWidom distribution in one-dimensional directed polymers with free endpoints, J. Stat. Mech. P11014 (2012)
[20] D.S. Dean, P. Le Doussal, S.N. Majumdar, and G. Schehr, Statistics of the maximal distance and momentum in a trapped Fermi gas at low temperature, J. Stat. Mech.: Theory and Experiment, 063301 (2017)
[21] P.L. Ferrari and H. Spohn, Constrained brownian motion: fluctuations away from circular and parabolic barriers, Annals of Probability, **33** 1302 (2005)
[22] P.L. Ferrari, M. Praehofer, and H. Spohn, Stochastic Growth in One Dimension and Gaussian Multi-Matrix Models, In proceedings of the *14th International Congress on Mathematical Physics* (ICMP 2003), World Scientific (Ed. J.-C. Zambrini) 404 (2006)
[23] N.R. Smith and B. Meerson, Geometrical optics of constrained Brownian excursion: from the KPZ scaling to dynamical phase transitions, arXiv:1811.01565
[24] A. Grosberg and H. Frisch, Winding angle distribution for planar random walk, polymer ring entangled with an obstacle, and all that: Spitzer-Edwards-Prager-Frisch model revisited, J. Phys. A: Math. Gen., **37** 3071 (2004)
[25] O. Vallée, M. Soares, *Airy functions and applications to physics* (London: Imperial College Press, 2004)
[26] S. Nechaev, K. Polovnikov, S. Shlosman, A. Valov, A. Vladimirov, Random flights over a disc, in preparation
[27] I. M. Lifshitz, Theory of fluctuation levels in disordered systems, Sov. Phys. JETP, **26** 462 (1968)
[28] I. M. Lifshitz, S. A. Gredeskul, and L. A. Pastur, *Introduction to the theory of disordered systems* (Wiley-Interscience: 1988)
[29] Th. M. Nieuwenhuizen, Trapping and Lifshitz Tails in Random Media, Self-Attracting Polymers, and the Number of Distinct Sites Visited: A Renormalized Instanton Approach in Three Dimensions, Phys. Rev. Lett. **62** 357 (1989)

5. Core-periphery organization of the cryptocurrency market inferred by the modularity operator

Introduction

Modularity matrix has long been used for inferring modular structure of stochastic networks of different scale-free nature. In this paper we show efficiency of the modularity to detect the core-periphery organization on the example of the cryptocurrency correlation-based network. The cryptocurrencies exemplify assets with dual macroeconomical background sharing properties of currency and stock markets with a non-obvious topological organization. We demonstrate that the modularity operator applied to a daily correlation-based network rules out community structure of the cryptocurrency market, simultaneously revealing stratification into a core and a periphery. Classification of tokens into two groups is shown to be day-dependent, however, stable tokens with statistically significant participation ratio can be easily identified. To approve the core-periphery organization of the stable assets, we compute the centrality measure of the two groups and show that it is considerably less for the periphery than for the core. Embedding of a subgraph of the stable tokens into the Euclidean space demonstrates clear spatial core-shell segregation. Furthermore, we show that the degree distribution of the minimal spanning tree has a distinctive power-law tail with exponent $\gamma \approx -2.6$ which makes the cryptomarket an archetypal example of the scale-free network. Economical reasoning suggests that the revealed topological motif is in the full agreement with the outliers hypothesis. The core is driven by traditionally liquid and highly capitalized tokens, resembling blockchain and payment systems, while the periphery is marked by the stable tokens with little exposure to the market. We report that the very center of the core is populated by tokens with strong financial usage, while main drivers of the market (such as

ETH or XRP) turn out to locate in the middle layers. This is an clear evidence of speculative processes underlying formation and evolution of the market.

Contribution

I have performed the spectral clustering on the cryptocurrencies data using the Newman's modularity, have calculated the closeness centrality measure, have visualized the network in the Euclidean space and have constructed the mean spanning tree on the most stable nodes.

Core-periphery organization of the cryptocurrency market inferred by the modularity operator

Kirill Polovnikov[1,2], Vlad Kazakov[3] and Sergei Syntulsky[3,4]
[1] *Skolkovo Institute of Science and Technology, 143005 Skolkovo, Russia*
[2] *Physics Department, M.V. Lomonosov Moscow State University, 119992 Moscow, Russia*
[3] *Cindicator LLC, 191186 St. Petersburg, Russia*
[4] *New Economic School, 121353 Moscow, Russia*

(Dated: September 19, 2019)

Modularity matrix has long been used for inferring modular structure of stochastic networks of different scale-free nature. In this paper we show efficiency of the modularity to detect the core-periphery organization on the example of the cryptocurrency correlation-based network. The cryptocurrencies exemplify assets with dual macroeconomical background sharing properties of currency and stock markets with a non-obvious topological organization. We demonstrate that the modularity operator applied to a daily correlation-based network rules out community structure of the cryptocurrency market, simultaneously revealing stratification into a core and a periphery. Classification of tokens into two groups is shown to be day-dependent, however, stable tokens with statistically significant participation ratio can be easily identified. To approve the core-periphery organization of the stable assets, we compute the centrality measure of the two groups and show that it is considerably less for the periphery than for the core. Embedding of a subgraph of the stable tokens into the Euclidean space demonstrates clear spatial core-shell segregation. Furthermore, we show that the degree distribution of the minimal spanning tree has a distinctive power-law tail with exponent ≈ -2.6 which makes the cryptomarket an archetypal example of the scale-free network. Economical reasoning suggests that the revealed topological motif is in the full agreement with the outliers hypothesis. The core is driven by traditionally liquid and highly capitalized tokens, resembling blockchain and payment systems, while the periphery is marked by the stable tokens with little exposure to the market. We report that the very center of the core is populated by tokens with strong financial usage, while main drivers of the market (such as ETH or XRP) turn out to locate in the middle layers. This is an clear evidence of speculative processes underlying formation and evolution of the market.

I. INTRODUCTION

Many core properties of a complex system (such as a market) can be captured in the network representation [1], in which the nodes respond to agents (say, buyers and vendors) and weights of the edges are defined by some quantitative measure of pairwise interactions between the agents (say, amount of goods traded). Resulting dimensionality reduction allows to extract valuable information on hidden topological structure of the system. One of the most striking and practically important examples of such structure is a mesoscopic organization of the agents into modules or communities [2–4]. Though the precise definition of a community depends on a generative stochastic model of the network [5], it is generally understood as a group characterized by reinforced interactions within itself, while having significantly less interaction strength with other nodes of the network. Whether it is a result of self-organization or intrinsic heterogeneity, such modular structuring is an important signature of collective behaviour, i.e. irreducible to action of independent agents, in the complex system. Practically speaking, the community detection problem allows to infer hidden relationships in the system and is an extremely hot topic in various technological [6, 7], biological [8–11], social [12–14] and economical [15, 16] contexts.

Not all of the real-world stochastic networks self-organize in modules, i.e. have a deterministic community structure. A stochastic network *per se* might have a different topological organization or be statistically indistinguishable from a random preferential attachment [39]. An alternative scenario includes formation of a dense core and of a relatively sparse periphery (shell). Such organization manifests the existence of a strongly interacting group of agents pulling the strength over the other players, which, in turn, are left to relatively weakly interact with each other being still strongly connected to the core. Notably, a network is capable of changing its topological mode from the "core-periphery" to the "communities" if the interaction strength of the periphery with itself overcomes the interaction strength between the periphery and the core. Myriad methods have been proposed to separate the cores from the peripheries in real stochastic networks on the basis of different quantitative measures and definitions of a core [40–43]. The most popular

approach has been suggested by Borgatti and Everett [40] and is constructed on a generative block model of a network with a fully-connected subgraph (the core). The rest subgraph (the periphery) is assumed to have no internal edges, however, is fully-connected to the core. Nevertheless, frequently one does not know from the very beginning the intrinsic topological structure of the network and what family of methods is appropriate to use. Thus, it is desirable to have an approach that would determine an "optimal" splitting of a stochastic network into statistically significant groups of arbitrary mutual topological relationship.

A widely used approach in the community detection is a spectral decomposition of a linear operator defined on the network: the information on communities is then encoded in several leading eigenvectors [17, 18]. It has been recently shown that all of the commonly used matrices (adjacency, Laplacian, modularity, non-backtracking) classify well the nodes as long as the network density is sufficient [19, 20]. In particular, the modularity operator has proven itself as one of the most efficient characteristic successfully detecting communities in stochastic networks of various nature [4, 14, 21–24]. To extract deterministic communities from the fluctuations, the modularity score measures the community-wise weight difference between the observed network and the expected one in the framework of a null generative model, in which the individual degrees of nodes are kept invariant under randomization of the edges. Fixation of the degrees from the sample makes the modularity applicable to scale-free networks, a wide class including most of the real-world networks [39].

In this paper we show the efficiency of the communities-specialized modularity operator in splitting the stochastic network into two groups with distinct centrality measures on the example of the cryptocurrency market, which exemplifies a youngling complex system with an unexplored topological motif. Cryptocurrencies have gained sufficient popularity over last several years due to their decentralized nature and sudden boost in capitalization in 2017. Still developing, maturity of the cryptomarket has been recently revealed from statistical characteristics of the bitcoin (BTC) time series [26] and of the bitcoin/etherium (BTC/ETH) rates [27], such as multifractality and volatility autocorrelations. Operation of cryptocurrencies (tokens) does not require a central authority and is sustained through a blockchain. Many of the tokens are functional units of the blockchain-based framework of technological companies which are issued during initial coin offerings (ICOs) and subsequently distributed to public through the crowd-funding mechanism. In contrast to stock markets, where clustering of the stocks is economically pre-determinant, in currency markets the reasoning behind formation of communities is more vague. A peculiarity of the tokens is that their nature is in between the two. On the one hand, the tokens represent monetary units, but, one the other hand, they are associated with a business model belonging to a certain technological sector.

Structure of stock and foreign exchange markets has previously been probed by ultrametric hierarchical and minimal spanning trees [28–32]. In particular, this technical approach to the US stock market was shown to be consistent with the standard of S&P500, classifying stocks into sectors or industries. One typically builds up the metric space based on cross-correlations emerging between the stocks and studies its temporal evolution. The minimal spanning tree then corresponds to a shortest path graph connecting all the nodes in the network, so that the leaves correspond to isolated communities [33]. The ultrametric tree approach fits the metric space of the market with an ultrametric model [31], i.e. aims to establish hierarchical relationships between the communities by organizing them into self-nested basins. Valuable evidence of collective behaviour comes from comparison of the leading eigenvalues of the network with spectral density the corresponding random matrix ensemble. Thus, coherent movements of the market as a whole are reflected in the magnitude of the largest eigenvalues of the metric tensor [34–37]. In [38] a magnification of the largest eigenvalue of the correlation matrix during market crashes has been demonstrated, implying strong coupling of the stocks during economical crises.

Can one identify a community-structure in the market of cryptocurrencies, analogous to sectors, industries and sub-industries existing in the stock market? Or does the core-periphery organization fit the cryptomarket better? There is a need to simultaneously examine both scenarios, which, as we show, can be met using the spectral modularity approach. To reveal the hidden topological structure of the cryptomarket, we model it by a correlation-based network, in which the tokens represent the nodes and the correlations between the vectors of the log-returns set up weights of the edges. The base currency was chosen to be the historical market leader, bitcoin (BTC), so that the network is insensitive, at least explicitly, to the volatility with respect to stable fiat currencies (ex., US dollar). In order to rule out the scam assets notoriously flooding the cryptomarket over the last two years [44], we perform a preliminary low-volume filtering. We assume that, because of their fraudulent essence, the scam tokens would hardly participate in formation of a stable long-term topological organization of the market. Having cleansed the ensemble of tokens from the scam, we maximize the modularity score in the principal eigenvector approximation for different base days and show that the topology of the cryptomarket responds to a stable core-periphery structure with a diffusive ring layer of non-stable tokens. We verify the robustness of the splitting done by the modularity using the closeness centrality measure [46, 47], as well as using the Euclidean metric tensor. The latter allows us to visualize the stable core-periphery organization of the network using the multi-dimensional scaling algorithm [48]. To ensure consistence

between different approaches, we additionally grow the mean spanning tree on the Euclidean manifold showing that the outer part of the tree corresponds to the periphery while the central part connects the tokens from the core in full agreement with predictions of the modularity.

The structure of the paper is as follows. In the Section II we describe the cryptomarket filtering from the scam assets, construct the correlation-based network of the true tokens and discuss the spectral modularity approach. In the Section III we report main results of the paper and in the Section IV we make the conclusion.

II. DATA AND METHODS

A. Price-volume data and scam filtering

All technical price-volume data on tokens has been taken from the Binance exchange [51]. As a quantitative measure of each particular token in the space, we have naturally chosen to calculate the logarithmic daily returns

$$r_i^{(T,\,t)} = \ln P_i(T-t) - \ln P_i(T-t-1); \quad t = 0, 1, 2, ..., D-1 \tag{1}$$

where $P_i(\tau)$ is the price of the i-th token at the day τ; T denotes the base day the log-returns are computed for, t enumerates the components and D is dimensionality of the log-returns vector.

Importantly, we calculate the price time series $P_i(\tau)$ in the base currency Bitcoin (BTC) in order to discard global movements of the cryptomarket from our analyses. Bitcoin has been a global market leader and it is assumed that most of the tokens are positively correlated with BTC. These coherent shifts are known to be the result of a strong collective response of the market to external macro-economical impulses and news, such as SEC regulations, large investments of institutional organizations etc.

The market of cryptocurrencies has experienced a strong inflow of macro-economically inappropriate tokens over the last several years, called "scam tokens" [44]. These are units resembling fake ventures launched for a short-term speculative purpose in order to execute "pump and dump" schemes. Soon after the release such assets accidentally loose their attractiveness for the investors, following by an abrupt drop in their price. Therefore, it seems naturally to assume that scam tokens do not reflect the long-standing network structure of the market, but rather make the data on "true tokens" (non-scam) noisy. Though a precise definition of a scam token would need to account for a complex combination of different factors, such as their price-volume data, white papers, founding teams, ICO information etc. [45], our aim here is to provide a simple general scheme eliminating assets that are non-representable for the global topological structuring. We assume that low-traded tokens are likely to be non-representable for our purpose, thus, one can roughly define a token as a scam if its daily volumes have appeared to be less than 10 BTC more than in 5% of days in the dataset. From Fig.1 it is seen that the amount of all tokens (the true and scam) is somewhat 3 times larger than the amount of true tokens. In other words, the size of the scam market is approximately twice the number of the true tokens. This unambiguously shows the importance of the filtration preceding the analyses.

B. Construction of the correlation-based network

A key step of the current study is construction of the weighted network. In what follows we assume that the nodes of the network represent the true tokens and the weight of an edge ("strength" of the connection) is associated with their pairwise correlation. Such networks are called correlation-based (CB) and are commonly used when an agent of a complex system can be characterized by a vector of a multi-dimensional space. This approach inevitably neglects many-agents interactions existing in the real market, such as a state of a pair of currencies conditioned by a certain value of a third one. However, it is believed that such reduction of complexity allows to unravel basic principles, driving the market's structuring.

Instantaneous correlations between a pair of tokens i, j at the base day T are characterized by the matrix $a_{ij}^{(T)}$ of the Pearson's correlation coefficients between the vectors $r_i^{(T,\,.)}$ and $r_j^{(T,\,.)}$ according to the following relation

$$a_{ij}^{(T)} = \frac{\left(r_i^{(T,\,.)}, r_j^{(T,\,.)}\right)}{\sqrt{\left(r_i^{(T,\,.)}, r_i^{(T,\,.)}\right)}\sqrt{\left(r_j^{(T,\,.)}, r_j^{(T,\,.)}\right)}} \tag{2}$$

where the scalar product is understood in the Euclidean sense. However, a straightforward use of the matrix $a_{ij}^{(T)}$ as weights of edges for the respective CB-network is not convenient for our further analyses because of its negative values. Thus, we will use either

(i) the scaled correlation matrix

$$\tilde{A}^{(T)} = \left(A^{(T)} + J_N\right)\frac{1}{2} \tag{3}$$

where J_N is the all-ones matrix and N is the amount of true tokens in our set (dimensionality of the matrix A; note that N does not depend on the base day T, see discussion below); or

(ii) the pairwise distances matrix d_{ij}

$$d_{ij}^{(T)} = \sqrt{2(1 - a_{ij}^{(T)})} \tag{4}$$

The distances $d_{ij}^{(T)}$ respond to the Euclidean metric space and the base day T. In particular, one can prove that (4) satisfies the triangle inequality. In particular, the elements d_{ij} can be associated with the shortest path distance between the assets i and j in this metric space and can be used for construction of the minimal-spanning tree.

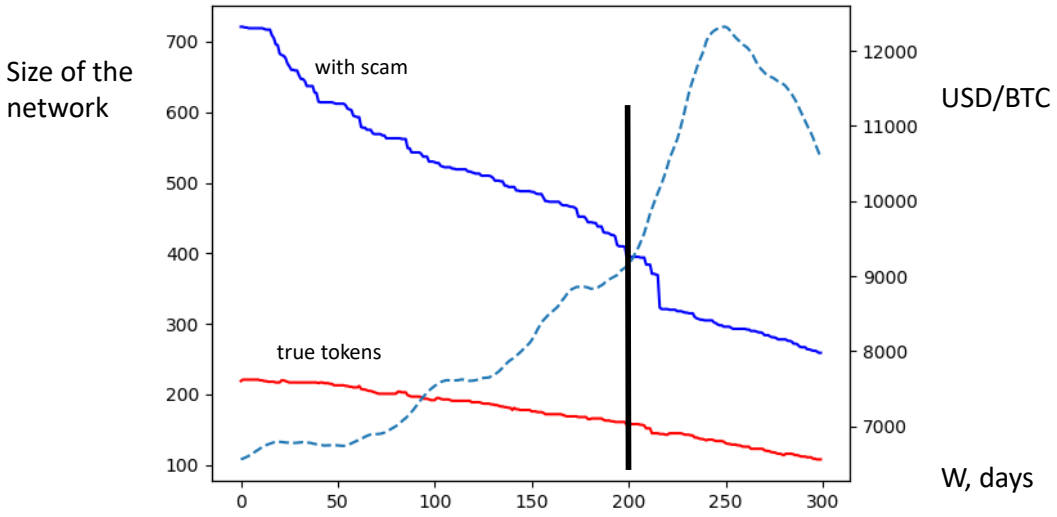

Figure 1: How should one choose the window size W? The amount of all tokens (blue) and of the true tokens (red) as functions of the window size and fixed $D = 100$. Moving average (100 days) of the BTC price (dashed): the window size of the moving average coincides with the number of days involved in calculation of the correlations, i.e. D.

In order to study evolution of the market's structure, we investigate temporal variations of the weight matrices (i) and (ii) as the base day T is sliding in a window of the size W starting from the last base day T_0, so that $T = \{T_0,\ T_0 - 1,\ T_0 - 2,\ ...,\ T_0 - W + 1\}$. In what follows, we fix the dimensionality of the returns vector $D = 100$. We collect the historical data starting from November 11th, 2018, which, therefore, stands for the last base day T_0 in our dataset. As for the window size W, it should be chosen accordingly to a trade-off between the size of the network and considerable significance of the forthcoming statistical analyses. On the one hand, the window W needs to be large enough, in order to keep track of conservative features in the fluctuating market's structure. On the other hand, the market of cryptocurrencies has not yet shaped, therefore, the lifetime of many of important tokens in the market essentially restricts the window size from above. In the Fig.1 we show how fast the number of available tokens in the market decreases with increase of the window size W. For instance, if one takes a one-year window, the respective set of tokens would consist of as few as one hundred true tokens. The next important circumstance to take into account is the state of the cryptomarket, which is illustrated in the same figure by the moving-averaged USD price of the bitcoin

(MA is taken for 100 days, which coincides with the depth D of the log-returns vectors), the main actor in the market. The turn of 2017 was marked by a rapid boost of the cryptomarket and by a subsequent drop in January 2018. We assume that the seismic regime of the market in the beginning of the year might unbalance emerging bonds between the tokens and, eventually, might obscure the topological structure of the whole market. Taking all the considerations above into account, we have chosen the window size to be equal to $W = 200$, which implies that $W + D = 300$ days are involved in computation of the correlations (from January 14th, 2017, to November 11th, 2017). This is a safe range that does not enter into the quaky regime of the market and accumulates $N = 157$ non-scam tokens in our ensemble. Note that the size N of the ensemble is determined only by the window size W, not by the choice of the base day T.

C. Spectral modularity approach

Having the network defined, we probe its topological structure in the framework of the classical cluster analyses. For this purpose we use the spectral modularity approach. Modularity quality function has been used vastly for communities detection in networks of various intrinsic nature [4, 14, 21–24]. The modularity is a functional over a network partition into the n groups G_p, $p = 1, 2, ..., n$, which relates observed weights to expected weights in an annealed ensemble of graphs with invariant strengths k_i of each individual node i and $m = \frac{1}{2}\sum_i k_i$ being the total strength of the network. Formally, the modularity functional $Q \equiv \{G_1, G_2, ..., G_n\}$ over different splitting into groups G_p can be written as follows

$$Q = \frac{1}{4m} \sum_p \sum_{(i,j)\in G_p} \left(\tilde{a}_{ij} - \frac{k_i k_j}{2m} \right) \tag{5}$$

where $\tilde{a}_{ij}$ are the weights of the respective edges of the network, taken from the scaled correlation matrix $\tilde{A}$, see (3) [52]. Maximization of the functional (5) yields the "optimal" splitting, which corresponds to the intrinsic community structure, subject to the network is not very sparse and the communities are sufficiently resolved [19, 20]. Originally, the modularity score (5) has been proposed [4] for partition of the scale-free networks, in which the distribution of the nodes degrees is some power law and does not follow the Poisson statistics, typical for the class of the Erdos-Renyi models. Recently it has been shown [25] that maximization of the modularity functional is equivalent to the maximum likelihood for a degree corrected version of the planted stochastic block model. The latter observation implies that the modularity maximization is qualified for the networks consisting of statistically indistinguishable communities, which for many of the real-world networks is, of course, an approximation.

In the case of two groups there is a simplified spectral approach, based on the principal eigenvector of the modularity matrix. One can assign a "spin direction" $s_i = \pm 1$ to each node of the network, depending on the group this node belongs to, and rewrite the modularity as a quadratic form in the spin space $\mathbf{s}$

$$Q = \frac{1}{4m} \sum_{i,j} \left(\tilde{a}_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j + 1) = \frac{1}{4m} \mathbf{s}^T B \mathbf{s} \tag{6}$$

where $B = b_{ij}$ is the modularity operator

$$b_{ij} = \tilde{a}_{ij} - \frac{k_i k_j}{2m} \tag{7}$$

and we have used in (6) the fact that the rows of the modularity matrix are summed to zero. Applying the spectral decomposition of (6), one can make use of the principal component approximation, which is is justified for sufficiently resolved communities. In this case the optimal partition becomes encoded in the leading eigenvector of the matrix B

$$Q \approx (4m)^{-1} \lambda_1 (\mathbf{u}_1 \mathbf{s})^2 \tag{8}$$

where $\mathbf{u}_1$ is the normalized leading eigenvector and λ_1 is the corresponding (largest) eigenvalue. In order to maximize (8), one has to choose the most collinear spin vector s to the given u_1. Therefore, the optimal solution $\mathbf{s}$ takes the value $s_i = +1$, if the corresponding component of $\mathbf{u}_1(i)$ is positive and $s_i = -1$ otherwise [53].

Below, on a particular example of the crypto-network, we show that the modularity operator might give meaningful information even when the network lacks intrinsic community structure. In particular, the spectral approach described

above is able to efficiently determine the core and the periphery of a network. Indeed, it is seen from (5) that, absence of communities in the network forces the modularity to optimize itself in order to accumulate the maximal weight in the one group. Thus, a sufficiently dense core could be resolved: nodes within the core have stronger connections with each other comparatively to their mean external strength with the second group, which becomes associated with the periphery.

III. RESULTS

A. Stability of splitting inferred by the modularity

We analyse the log-returns correlations emerging between the tokens of the cryptomarket for $W = 200$ base days starting from November 11 ($T = 0$) to April 26 ($T = 200$) of 2018. The ensemble of tokens actively trading in this time window (non-scam tokens according to our volume-based definition in the previous section) has the size $N = 157$. Scaled correlations comprise weights of the respective edges in the CB-network representation of the cryptomarket. At each base day T the spectral properties of the modularity operator defined on a spin-space (assuming two groups in the network) have been studied in approximation of the leading eigenvector: the edge of the spectrum of the modularity matrix $b_{ij}^{(T)}$ describes the topological pattern of the network. In particular, the signs of components of the leading eigenvector determine the group the corresponding token gets assigned into. Information on the splitting, produced at each base day T is recorded so that the first group is the one containing another crypto leader, Etherium (ETH, one of the most popular blockchain platform). As a matter of fact, the main outlier of the market, USDT (analogue of USD), is then always classified to the second group. Such group labeling allows one to keep track of the topological difference between the groups produced by the modularity and is discussed below.

Assignment of tokens to the groups has found to be day-dependent, which implies stochastic character of the market. To infer conservative properties of the market's structure, we aim at determining *stable tokens* that might be associated with a certain group with sufficient statistical significance. In order to quantify the stability of tokens, we introduce a participation ratio ϕ that equals to the fraction of days a token spends in the ETH group (and $1 - \phi$ days in the USDT group, correspondingly). The Fig.2a demonstrates that the distribution of tokens by the number of days a token spends in the ETH group is pronouncedly bimodal. Peaks located at the edges of this histogram infer tokens persisting in their respective groups over the course of time. A value $\phi^* = 0.8$ is chosen to provide a threshold for a token to be classified as a stable token of the first group. Accordingly, $1 - \phi^* = 0.2$ is a maximal fraction of days a token is allowed to spend in the first group to be classified as a stable token of the second group. The two clauses above might be combined in a following single one: the token is stable only if its participation ratio ϕ satisfies the following condition

$$|\phi - 0.5| > 0.3 \qquad (= \phi^* - 0.5) \tag{9}$$

Such definition results in the identification of $N_c = 57$ stable ETH-coupled tokens, $N_p = 38$ stable USDT-coupled tokens and $N_{ns} = 62$ non-stable ones, which correspond to the central part of the histogram Fig.2a. The top ten most liquid tokens of the first group are {ETH, BCH, EOS, XRP, TRX, NEO, QTUM}, while the top ten ones of the second group are {USDT, MCO, DOGE, BCD, PRO, ARDR, MAID, ZEN, R, GUP}, see Fig. S1. The first group is comprised of the most liquid, highly capitalized blockchains and payment systems like Ethereum, Ripple, EOS and Bitcoin Cash. As for the second group, no general principle under its formation can be ruled out apart from that it includes the main market's outlier USDT. We will see below that is simply because the second group topologically is not a community, but a periphery of the ETH group.

The Fig.2b illustrates that population of the two groups fluctuates, though these fluctuations are mostly related to the non-stable tokens. In the same figure the daily sizes of the intersections with the stable sets are shown to be much more characteristic, being slightly below the level lines, denoting the sizes of the stable sets (N_c and N_p). The deviations from the level lines become more considerable both for the two groups as one approaches the end of the window: this is a signature of the market rally that was taking place in the beginning of the year. Notably, if at the base day $T = 0$ one assigns two colors to the groups of sizes $N_c' \approx 85$ (red) and $N_p' \approx 70$ (blue) assuming the market is typically getting split into, then the expected number of red and blue tokens falling by chance into these groups N_c' and N_p' is less than the sizes of the stable groups ($46 < N_c = 57$; $31 < N_p = 38$). This is a supportive argument in favour of non-random distribution of tokens between the two groups and points out to the intrinsic connection between the essence of an asset and its affinity either to ETH or to USDT.

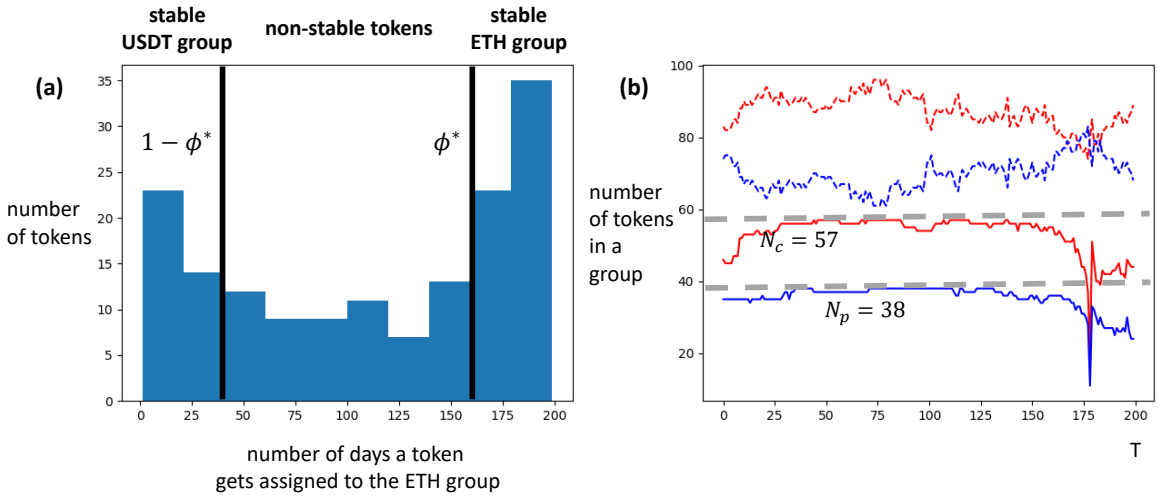

Figure 2: Inferring stability of the splitting. (a): distribution of tokens by the number of days they are assigned to the ETH group by modularity. Black vertical lines resemble critical fractions of days in the window of size $W = 200$, tokens need to spend to be classified in the stable groups. (b): Daily sizes of the two groups, ETH-coupled (red) and USDT-coupled (blue), as determined by modularity (dashed); sizes of the intersection between the two groups and the corresponding stable groups (solid). Dashed gray lines show the sizes of the stable groups.

In order to determine a topological structure of the splitting produced by the modularity we compute the, so-called, closeness centrality [46, 47] for the two groups. The closeness centrality $C_i^{(T)}$ of an asset i is a measure of its centrality with respect to a subgraph G in a certain metric space, computed at the base day T

$$C_i^{(T)} = \frac{1}{\langle d_{ij}^{(T)} \rangle_{j \in G}} \tag{10}$$

where $d_{ij} \equiv d_{ij}^{(T)}$ are the Euclidean distances between the nodes i and j, (4). Here we compute the centralities of the tokens with respect to the ETH group.

Time series of the closeness centrality, as the base day changes, are shown for all $N = 157$ tokens in the Fig.3(a). It is seen that up to fluctuations in the intermediate layer populated by the non-stable assets the centrality of the ETH-coupled tokens remarkably exceeds the centrality of the USDT-coupled ones. It is said that all the tokens assigned by modularity into the first group are sufficiently closer to each other than to the tokens from the second group.

Importantly, the non-stable tokens do not greatly affect stratification of the tokens in the Fig.3(a). It implies that the non-stability transition at the points ϕ^* and $1 - \phi^*$ is smooth, i.e. the proportion of days the token spends in either group decreases non-abruptly. This result is much more evident from the Fig.3(b) where we have provided a correlation plot of the mean token's centrality $\langle C^{(T)} \rangle_T$ versus its participation ratio ϕ. Pronounced correlations between the two quantities might be readily noticed. The non-stable assets transit continuously between red and blue phases, evenly filling the correlation strip. Though the transition between the ETH and the USDT groups is smooth, we will se below that the properties of the two phases are quiet different.

Evolution of the strip in the Fig.3(a) demonstrates that the centrality of the first group is slightly increasing as T decreases (the historical time increases). This implies kinetic fortification of the market structuring: the stable tokens of the first group are interacting stronger with each other in the course of the historical time. A bimodal distribution of the ETH centrality, see Fig.3(c), is an evidence of a qualitative change in the cryptomarket behaviour, possibly related to relaxation after the peak in January 2018.

In the Fig.3(a) we depict the time series for ETH separately in order to illustrate that, contrary to the intuitive thinking, the ETH does not constitute the center of the first group. The leadership in the first group considerably

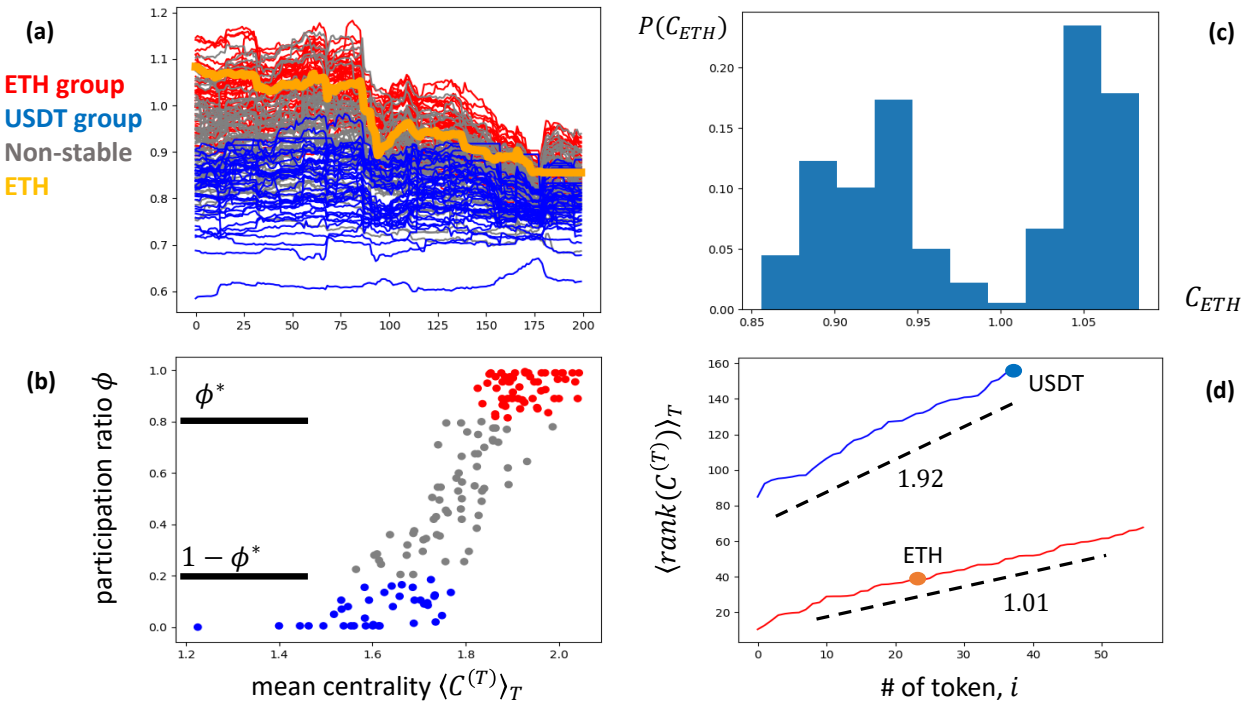

Figure 3: The closeness centrality rationalizes the groups, produced by modularity. Red marks denote tokens from the ETH group, blue ones denote tokens from the USDT group and gray ones stand for the non-stable tokens. (a): closeness centrality of the tokens at each particular base day; (b): Scatter plot of the mean closeness centrality of a token and its participation ratio; (c): Distribution of centrality measure of ETH token; (d): Sorted mean rank of the centrality versus the index of token in the sequence is illustrated for the two groups separately.

fluctuates, though ETH never takes it in the window studied: the orange line resembling ETH never crosses the envelope of the red strip. The mean rank of the centrality Fig.3(d) supports this statement. It illustrates the average location of the token relatively to the other tokens of the group. The tokens are sorted according to their rank along the x-axis. It is seen that ETH is fairly close to the middle of this sequence, having the mean rank of centrality $\langle rank\left(C^{(T)}\right)_{ETH}\rangle_T = 39$, which brings it to the 24th place among all the tokens in the first group. In other words, almost half of the tokens in the ETH group are found to correlate with the rest of the group stronger than ETH does. A possible explanation is the following. One can assume that crypto market prices are driven mostly by financial usage (as speculation and investment). Thus, tokens with many non-speculative use cases are, in general, less correlated with the others. At the same time, tokens that represent purely investment instruments, given the same level of price manipulation activity, are most mutually correlated along with their share investors sentiment. Therefore, ETH as a token which is actively used for functional purposes (e.x., for ICOs) like most other high cap tokens is not in the very center of the universe.

The second group does not show such amplification of the centrality in the course of the historical time, implying that its interactions with the first group are marginally weak. The bottom curve in the Fig.3(a) corresponds to USDT, i.e. the dollar analogue has the lowest centrality in the market, permanently. The centrality rank of USDT equals the size of the network $\langle rank\left(C^{(T)}\right)_{USDT}\rangle_T = N = 157$. Such a strong opposition of USDT to the first group constitutes a major factor driving segregation in the market. Repulsion of USDT draws off a part of the market from the first group and, thus, leads to formation of the second group.

Interestingly, as we see from the Fig.3(d), the sorted mean rank of the centrality almost coincides with the index number of the token in the sequence for the ETH group (the corresponding coefficient is ≈ 1.01). Note that the unit coefficient would naturally arise in the static sequence or in case when one forbids "overtakes" between the tokens (e.x., for self-excluding particles diffusing on a line). One can infer that fluctuations of mutual positions of the tokens in the first group are only local and do not lead to the global redistribution of the assets' locations inside the group. At the same time, the structure of the second group is much more dynamic: the sorted rank as a function of the index number demonstrates the slope ≈ 1.92. This is almost twice larger than it would be for the static sequence and

implies relative boosting of the second group outwards the first group. A reason for this internal boosting is repulsion between the outliers and the ETH group.

B. Embedding the network into a metric space: the core-periphery structure

In the previous section we have studied stability of the topological splitting into two groups, produced by the modularity at different base days. Here we prove that the metric structure of these two groups corresponds to the core-periphery organization of the network. A self-obvious core-shell metric profile straightforwardly follows from visualization of the averaged characteristic matrices, see heatmaps in the Fig.4. Indeed, the mean internal correlations of the first group (the core) with itself is notably larger than the mean external correlations with the second group (the periphery). In the language of the Euclidean metrics d_{ij}, tokens of the first group are closer to each other, than to tokens of the second group, which is consonant with the closeness centrality, discussed in the previous section, Fig.3. At the same time the periphery (i.e. the USDT group) interacts considerably weaker with itself than with the core, see Fig.4. The bright line in the averaged distances matrix and, correspondingly, the dark one in the averaged correlation matrix resemble USDT, which demonstrates the lowest negative correlations with the rest of the market, around -0.4. As we have observed, qualitatively these patterns describe well the networks at all base days T from the window, i.e. the same structure emerges in all realizations of daily matrices. Accordingly, communities in the classical sense have never formed in the days analyzed.

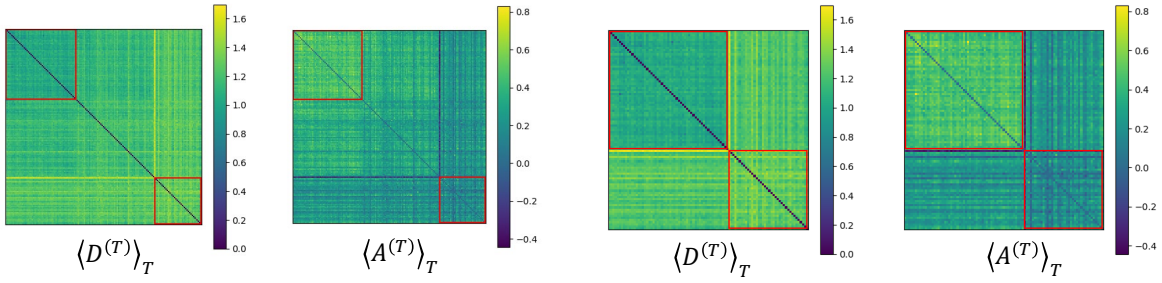

Figure 4: Averaged matrices of the pairwise distances D and of the pairwise correlations A between the tokens; stable groups are marked with red. The tokens are sorted by their typical volume separately inside each group: in the core, in the periphery and in the non-stable layer.

Additionally, we embed the core-periphery structure of the cryptomarket into the Euclidean space, using the characteristic matrix of pairwise distances d_{ij}. This was done using the sklearn library, which realizes the multi-dimensional scaling (MDS) algorithm [48]. Generally speaking, reconstruction of the coordinates in the Euclidean metric space by the pairwise distances is an ill-posed problem. Thus, the MDS algorithm seeks an optimal fitting determined by the minimal residual "stress" of the manifold. Having the coordinates of tokens obtained we have performed a series of the following transformations: (i) first, we translate the network so that USDT token gets positioned at the origin, (ii) second, we rotate the structure at some angle so that the polar angle of ETH equals $\pi/4$.

These two transforms allow to run the MSD algorithm for different base days and visualize the network's organization at the first quadrant. And finally, (iii) we inverse the network over the line, connecting USDT and ETH, if a third token of our choice (we have chosen ADA) is found to be located in the upper half-plane relatively to this line. Applying the transformations listed above to the sets of coordinates computed at different base days, we uniquely set up the structure of the rigid network in the metric space.

Analyzing evolution of the network with the change of the base day T we have found that the overall core-periphery structure persists. The periphery is inhomogeneously distributed around the core, often significantly shifted to the origin, where USDT is localized by construction. Generally, the fluctuating structure of the network is best described by the matrix of mean pairwise distances $\langle d_{ij} \rangle$. The corresponding typical embedding is illustrated in the Fig.5(a). One can notice a two-phase geometrical separation of the network into the core and the shell with a clear boundary between them. Thus, it can be visually verified that the spectral modularity algorithm has successfully identified the tokens belonging to the core and to the periphery.

In order to infer additional information on structuring taking place in the network, we investigate the growth of the minimal spanning tree (MST) on it. The MST is a connected subgraph of a given graph, which has no cycles and collects the minimum possible weight on the graph. If weights of the edges of a graph, for instance, are described by the matrix d_{ij}, then the corresponding MST sets up the shortest paths along the graph between any two nodes. For a CB-network the MST characterizes routes the correlations propagate. In the Fig.5(a) we show the minimal spanning tree constructed *via* the Kruskal's algorithm [49]. Different fractions of the tree from $\varphi = 0.25$ to $\varphi = 1$ are shown in the Fig.5(a), mimicking the growth of the tree (the process of its construction).

It can be noticed that the first two stages of the growth $\phi < 0.5$ correspond to the core of the network, i.e. the core comprise of TOP-50% of the most strong interactions of the MST, while the outer part of the tree, $\phi > 0.5$, connects tokens from the periphery. Second, as the tree is growing, its leaves do not form ultrametrically nested modules as it has been reflected in the subindustrial structure of the stock market [28–31]. Abundance of intersection of links rather implies absence of the secondary structuring in sub-modules and communities within on top of the core-periphery organization. The same conclusion is drawn by the modularity which, as reported above, has failed to split the two core and the periphery into smaller groups.

We report formation of hubs in the MST of stable cryptotokens such as STRAT, REQ, NEO and LEND within the core and PTOY, OK, BITB, GUP within the periphery which is a manifestation of the scale-freeness of the network. In order to further investigate this phenomenon we plot the degree distribution of the MST and establish a power-law tail with exponent $\gamma = -2.58$, see Fig.5(b). This result is a qualitative indicator of the scale-freeness inherent to the cryptomarket. Notedly, the foreign exchange market has been shown to demonstrate a generally wider distribution of the MST degree with exponents which are dependent on the base currency but generally do not exceed $\gamma < 2$ [50].

IV. CONCLUSION

In this note we have shown that the modularity functional is capable of splitting the network into the core and the periphery. The functional has been defined on a correlation-based network of cryptocurrencies and has subsequently been optimized in the largest eigenvalue approximation. The method is robust in the sense that it both discards the hypothesis about intrinsic community structure of the network and establishes the core-periphery organization. This organization has been shown to be consonant with the centrality measure with respect to the core, i.e. tokens from the periphery have considerably less centrality than ones from the core. Splitting provided by modularity depends on the base day which exhibits stochastic nature of the market. However, distribution of the participation ratio with respect to one of the groups is bimodal, which is a signature of the two phases present in the system. Associating the tails of the distribution with the two phases, the two stable groups of tokens can be established.

In order to study the topological motif of the network, we have visualized the mean matrix of pairwise correlations and the mean matrix of the pairwise distances with rows and columns sorted according to the outcome of the spectral modularity. It is clear that the average correlation-based weight accumulated in the core within itself exceeds both the weight of the periphery-periphery and of the core-periphery interaction, supporting the core-periphery topological organization of the network. By means of the multidimensional scaling algorithm we have embedded the network to the metric space and have demonstrated stratification into the dense community of tokens in the center of the universe and the periphery which gets positioned around the core. This organization in the metric space is found to be stable to variations of the base day. An averaged pairwise distances matrix produces a clear separation of the network into two phases corresponding to the two stable sets of tokens found with the spectral modularity approach. Evolution of the minimal spanning tree grown on the core-periphery metric structure demonstrates no formation of isolated

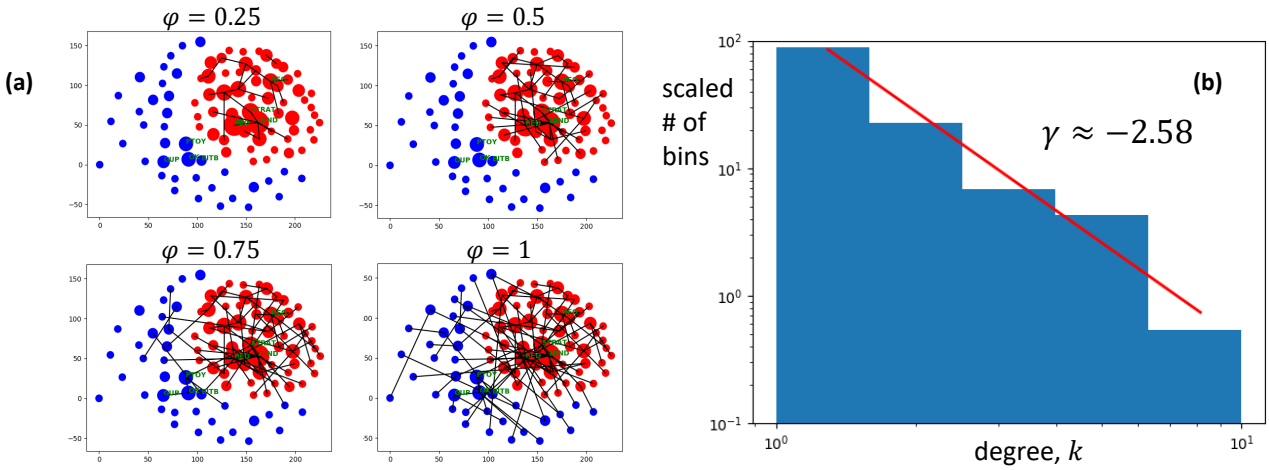

Figure 5: (a): Visualization of the average core-periphery metric structure and the growing minimal spanning tree embedded into it. The network is constructed by its mean matrix of pairwise distances $\langle d_{ij} \rangle$ using the multi-dimensional scaling algorithm. The minimal spanning tree is determined using the Kruskal's algorithm; a subsequent growth of the MST occurs in four stages corresponding to the four quantiles its edges are sorted into by their weight: (i) TOP-25%, (ii) TOP-50%, (iii) TOP-75% and (iv) all the edges belonging to the tree. Tokens belonging to the core and to the periphery are marked by red and blue correspondingly. Sizes of the nodes are proportional to their strengths. Four strongest nodes of the core and of the periphery are annotated. (b): Degree distribution of the MST in the double-log scales demonstrates a power-law tail $P(k) \sim k^{-\gamma}$ with exponent $\gamma \approx -2.58$. The histogram is plotted for the logarithmically spaced bins and the y-axis stands for the amount of nodes divided by the corresponding width of the bin.

leaves, which might have been referred as sub-communities, corroborating the conclusions drawn by the modularity. This result puts the the cryptocurrency market in contrast both with the stock market and with the foreign exchange market, where hierarchical modular organization takes place and macro-economically determined. We also note that the scale-free exponent γ of the cryptocurrency market, reconstructed based on its minimal spanning tree, places it into the archetypal class of the scale-free networks with $2 \leq \gamma \leq 3$.

Economical explanation of the core-periphery topological structure of the crypto market is straightforward. Stable coins (like USDT and NBT) are filtered to the periphery, which is expected as these coins have no exposure to the crypto market and have low correlation with other crypto currencies. Most actively traded peripheral coins excluding stable coins are Dogecoin and Bitcoin Diamond, which are the bitcoin forks with very high volatility and relatively controversial reputation. Periphery tokens comprise opposition to traditional crypto market and, due to quiet low correlation with the rest of the market, can be named outliers. Most liquid, high cap blockchains and payment systems (like Ethereum, Ripple, EOS and Bitcoin Cash) belong to the core cluster. It matches expectations as these tokens show less uncorrelated behavior (pumps and dumps) than smaller tokens. These tokens drive the whole market behavior so they are in the center of it. At the same time the very center of the core is filled not by the leader tokens, but rather by much less known small-cap ones (like Stratis, Request, SALT). These tokens form hubs in the metric structure and have high centrality measure, which is explained by their significant financial usage (as speculation and investment). In contrast, the high cap tokens of the core (ETH, XRP, EOS, BCH etc.) are not purely investment instruments, their financial manipulation is moderate and, as a result, that are shifted from the very center of the core.

Acknowledgments

We are grateful to Mikhail Tamm for discussions and to the referee for useful suggestions on the manuscript. KP acknowledges the support of the Foundation for the Advancement of Theoretical Physics and Mathematics "BASIS" (grant 17-12-278).

[1] M. Newman, *Networks: an introduction* (Oxford University Press, 2010).
[2] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, V. and D. Parisi, Defining and identifying communities in networks. PNAS 101, 2658 (2004).
[3] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, Finding statistically significant communities in networks. PloS one 6, e18961 (2011).
[4] M. E. Newman, Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104 (2006).
[5] S. Fortunato and D. Hric, Community detection in networks: A user guide. Physics reports 659 (2016).
[6] R. Albert, J. Hawoong and A.-L. Barabasi, Internet: Diameter of the world-wide web. Nature 401, 130 (1999).
[7] A. Broder, et al., Graph structure in the web. Computer networks 33, 309 (2000).
[8] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Gen. 14, 390 (2013).
[9] R. Pastor-Satorras, and A. Vespignani, Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200 (2001).
[10] H. Jeong, et al., The large-scale organization of metabolic networks. Nature 407, 651 (2000).
[11] E. Ravasz, et al., Hierarchical organization of modularity in metabolic networks. Science 297, 1551 (2002).
[12] S. Redner, How popular is your paper? An empirical study of the citation distribution. The Europ. Phys. J. B-Cond. Matt. and Compl. Sys. 4, 131 (1998).
[13] M. Newman. The structure of scientific collaboration networks. PNAS 98, 404 (2001).
[14] J. Chen, O.R. Zaiane, and R. Goebel, Detecting communities in social networks using max-min modularity. Proceedings of the 2009 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, (2009).
[15] C. Piccardi, L. Calatroni, and F. Bertoni, Communities in Italian corporate networks. Physica A: Statistical Mechanics and its Applications 389, 5247 (2010).
[16] R. Corrado and M. Zollo, Small worlds evolving: governance reforms, privatizations, and ownership networks in Italy, Ind. Corp. Change 15, 2 (2006).
[17] U. Von Luxburg, A tutorial on spectral clustering. Stat. and comp. 17, 395 (2007).
[18] M. Krivelevich and B. Sudakov, The largest eigenvalue of sparse random graphs. Comb., Prob. and Comp. 12, 61 (2003).
[19] R. R. Nadakuditi, and M. Newman, Graph spectra and the detectability of community structure in networks. Phys. Rev. Lett. 108, 188701 (2012).
[20] A. Decelle, et al., Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Phys. Rev. E 84, 066106 (2011).
[21] H. K. Norton, et al., Detecting hierarchical genome folding with network modularity. Nat. Met. 15, 119 (2018).
[22] J. Grilli, T. Rogers, and S. Allesina. Modularity and stability in ecological communities. Nat. Comm. 7, 12031 (2016).
[23] R. Guimera, et al., Origin of compartmentalization in food webs. Ecology 91, 2941 (2010).
[24] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, Extracting the hierarchical organization of complex systems. PNAS 104, 15224 (2007).
[25] M. E. J. Newman, Equivalence between modularity optimization and maximum likelihood methods for community detection. Phys. Rev. E 94, 052315 (2016).
[26] S. Drozdz, et al. Bitcoin market route to maturity? Evidence from return fluctuations, temporal correlations and multiscaling effects. Chaos 28.7 (2018).
[27] S. Drozdz, et al. Signatures of crypto-currency market decoupling from the Forex. arXiv preprint arXiv:1906.07834 (2019).
[28] J.-P. Onnela, et al., Dynamics of market correlations: Taxonomy and portfolio analysis. Phys. Rev. E 68, 056110 (2003).
[29] R. B. Roy and U. K. Sarkar, A social network approach to change detection in the interdependence structure of global stock markets. Soc. Net. Anal. and Mining 3, 269 (2013).
[30] V. Boginski, S. Butenko, and P. M. Pardalos, Mining market data: a network approach. Comp. & Oper. Res. 33, 3171 (2006).
[31] R. N. Mantegna, Hierarchical structure in financial markets. The Europ. Phys. J. B-Cond. Matt. and Comp. Sys. 11, 193 (1999).
[32] R. N. Mantegna, and H. E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, 1999.
[33] C. H. Papadimitriou, and K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity. Printice-Hall.Inc., Englewood Cli s, New Jersey (1982).
[34] L. Laloux, et al., Random matrix theory and financial correlations. Int. J. of Theor. and Appl. Fin. 3, 391 (2000).
[35] V. Plerou, et al., Random matrix approach to cross correlations in financial data. Phys. Rev. E 65, 066126 (2002).

[36] S. Valeyre, D. S. Grebenkov, and S. Aboura, Emergence of correlations between securities at short time scales. Physica A: Stat. Mech. and its Appl. (2019).
[37] D. M. Song, et al., Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. Phys. Rev. E 84, 026108 (2011).
[38] S. Drozdz, et al., Dynamics of competition between collectivity and noise in the stock market. Physica A: Stat. Mech. and its Appl. 287, 440 (2000).
[39] A. L. Barabasi, and R. Albert, Emergence of scaling in random networks, Science 286, 509 (1999).
[40] S. P. Borgatti and M. G. Everett, Models of core/periphery structures. Social Networks 21, 4 (2000).
[41] M. Newman and M. Girvan, Finding and evaluating community structure in networks. Phys. Rev. E 69, 2 (2004).
[42] P. Holme, Core-periphery organization of complex networks. Phys. Rev. E 72, 4 (2005).
[43] M. P. Rombach et al., Core-periphery structure in networks. SIAM J. on Appl. Math. 74, 1 (2014).
[44] U. W. Chohan, Initial coin offerings (ICOs): Risks, regulation, and accountability (2017).
[45] S. Bian, et al., ICOrating: A deep-learning system for scam ICO identification. arXiv preprint arXiv: 1803.03670 (2018).
[46] S. Wasserman and K. Faust, Social network analysis: Methods and applications, Vol. 8. Cambridge university press, (1994).
[47] L. C. Freeman, Centrality in social networks conceptual clarification. Social networks 1.3 (1978).
[48] I. Borg and P. Groenen, Modern multidimensional scaling: Theory and applications. J. of Edu. Meas. 40.3 (2003).
[49] J. B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society 7.1 (1956).
[50] J. Kwapie, et al. Analysis of a network structure of the foreign currency exchange market. Journal of Economic Interaction and Coordination 4.1 (2009).
[51] http://www.binance.com.
[52] Here in (5) and below all characteristic tensors are computed for the particular base day, but we do not denote this fact explicitly for the brevity of respective formulas.
[53] Note that the optimal spin vector $s \equiv s^{(T)}$ is different for different base days T.

99

6. Order and stochasticity in the folding of individual Drosophila genomes

Introduction

Mammalian and Drosophila genomes are partitioned into topologically associating domains (TADs). Although this partitioning was reported to be functionally relevant, it is unclear whether TADs represent true physical units located at the same genomic positions in each cell nucleus or emerge as an average of numerous alternative chromatin folding patterns in a cell population. Here, applying an improved single-nucleus Hi-C technique (snHi-C), we constructed Hi-C maps in individual Drosophila genomes with a 10 kb resolution. These maps demonstrate chromatin compartmentalization at the megabase scale and partitioning of the genome into non hierarchical TADs at a scale of 100 kb, which closely resembles the TAD profile in the bulk in situ Hi-C data. Over 40nuclei, and these boundaries possess a high level of active epigenetic marks. Polymer simulations demonstrate that chromatin folding is best described by the random walk model within TADs and is best approximated by a crumpled globule build of Gaussian blobs at longer distances. We observed prominent cell-to-cell variability in the long range contacts between either active genome loci or between Polycomb-bound regions, arguing for an important contribution of stochastic processes to the formation of the Drosophila 3D genome.

Contribution

I have verified that the experimental sparse Hi-C matrices are not equivalent to random realizations of the configuration model graphs with conserved contact probability. I have proposed a method to annotate TADs in sparse Hi-C matrices based on the non-backtracking walks. I have demonstrated the efficacy of the method on the ensemble of single cell matrices and have proved that the found domains are epigenetically significant.

Order and stochasticity in the folding of individual *Drosophila* genomes

Sergey V. Ulianov[1,2]*, Vlada V. Zakharova[1,2,3]*, Aleksandra A. Galitsyna[4]*, Pavel I. Kos[6]*, Kirill E. Polovnikov[4,7], Ilya M. Flyamer[8], Elena A. Mikhaleva[9], Ekaterina E. Khrameeva[4], Diego Germini[3], Mariya D. Logacheva[4], Alexey A. Gavrilov[1,10], Alexander S. Gorsky[5,15], Sergey K. Nechaev[11,12], Mikhail S. Gelfand[4,5], Yegor S. Vassetzky[3,13], Alexander V. Chertovich[6,14], Yuri Y. Shevelyov[9], Sergey V. Razin[1,2,#].

[1]Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia;

[2]Faculty of Biology, M.V. Lomonosov Moscow State University, Moscow, Russia;

[3]UMR9018, CNRS, Université Paris-Sud Paris-Saclay, Institut Gustave Roussy, Villejuif, France;

[4]Skolkovo Institute of Science and Technology, Moscow, Russia;

[5]Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia;

[6]Faculty of Physics, M.V. Lomonosov Moscow State University, Moscow, Russia;

[7]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139;

[8]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK;

[9]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia;

[10]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia;

[11]Interdisciplinary Scientific Center Poncelet (CNRS UMI 2615), Moscow, Russia;

[12]P.N. Lebedev Physical Institute, Russian Academy of Sciences, Moscow, Russia;

[13]Koltzov Institute of Developmental Biology, Russian Academy of Sciences, Moscow, Russia;

[14]Semenov Federal Research Center for Chemical Physics;

[15]Moscow Institute for Physics and Technology, Dolgoprudnyi, Russia.

*These authors contributed equally to this work.

[#]Correspondence should be addressed to S.V.R (email: sergey.v.razin@usa.net).

Abstract

Mammalian and *Drosophila* genomes are partitioned into topologically associating domains (TADs). Although this partitioning has been reported to be functionally relevant, it is unclear whether TADs represent true physical units located at the same genomic positions in each cell nucleus or emerge as an average of numerous alternative chromatin folding patterns in a cell population. Here, we used a single-nucleus Hi-C technique (snHi-C) to construct high-resolution Hi-C maps in individual *Drosophila* genomes. These maps demonstrate chromatin compartmentalization at the megabase scale and partitioning of the genome into non-hierarchical TADs at the scale of 100 kb, which closely resembles the TAD profile in the bulk *in situ* Hi-C data. Over 40% of TAD boundaries are conserved between individual nuclei; these boundaries possess a high level of active epigenetic marks. Polymer simulations demonstrate that chromatin folding is best described by the random walk model within TADs and is most suitably approximated by a crumpled globule build of Gaussian blobs at longer distances. We observed prominent cell-to-cell variability in the long-range contacts between either active genome loci or between Polycomb-bound regions, arguing for an important contribution of stochastic processes to the formation of the *Drosophila* 3D genome.

INTRODUCTION

The principles of higher-order chromatin folding in the eukaryotic cell nucleus have been disclosed thanks to the development of chromosome conformation capture techniques, or C-methods[1,2]. High-throughput chromosome conformation capture (Hi-C) studies demonstrated that chromosomal territories were partitioned into partially insulated topologically associating domains (TADs)[3-5]. TADs likely coincide with functional domains of the genome[6-8], although the results concerning the role of TADs in the transcriptional control are still conflicting[6,9-12]. Analysis performed at low resolution suggested that active and repressed TADs were spatially segregated within A and B chromatin compartments[13,14]. However, high-resolution studies demonstrated that the genome was partitioned into relatively small compartmental domains bearing distinct chromatin marks and comparable in sizes with TADs[15]. In mammals, the formation of TADs by active DNA loop extrusion partially overrides the profile of compartmental domains[15,16]. Of note, TADs identified in studies of cell populations are highly hierarchical (i.e. comprising smaller subdomains, some of which are represented by DNA loops[5,17]).

Partitioning of the genome into TADs is relatively stable across cell types of the same species[3,4]. The recent data suggest that mammalian TADs are formed by active DNA loop extrusion[18,19]. The boundaries of mammalian TADs frequently contain convergent binding sites for the insulator protein CTCF that are thought to block the progression of loop extrusion[19-21]. Contribution of DNA loop extrusion in the assembly of *Drosophila* TADs has not been demonstrated yet[22]; thus, *Drosophila* TADs might represent pure compartmental domains[23]. Large TADs in the *Drosophila* genome are mostly inactive and are separated by transcribed regions characterized by the presence of a set of active histone marks, including hyperacetylated histones[5,24]. Some insulator/architectural proteins are also overrepresented in *Drosophila* TAD boundaries[24-26], but their contribution to the formation of these boundaries has not been directly tested. The results of computer simulations suggest that *Drosophila* TADs are assembled by the condensation of nucleosomes of inactive chromatin[24].

The current view of genome folding is based on the population Hi-C data that present integrated interaction maps of millions of individual cells. It is not clear, however, whether and to what extent the 3D genome organization in individual cells differs from this population average. Even the existence of TADs in individual cells may be questioned. Indeed, the DNA loop extrusion model considers TADs as a population average representing a superimposition of various extruded DNA loops in individual cells[18]. Heterogeneity in patterns of epigenetic modifications and transcriptomes in single cells of the same population was shown by different single-cell techniques, such as single-cell RNA-seq[27], ATAC-seq[28], and DNA-methylation analysis[29]. Studies performed using FISH demonstrated that the relative positions of individual genomic loci varied significantly in individual cells[30]. The first single-cell Hi-C study captured a low number of unique contacts per individual cell[31] and allowed only the demonstration of a significant variability of DNA path at the level of a chromosome territory. Improved single-cell Hi-C protocols[32,33] allowed to achieve single-cell Hi-C maps with a resolution of up to 40 kb per individual cell[32,34] and investigate local and global chromatin spatial variability in mammalian cells, driven by various factors, including cell cycle progression[33]. Of note, TAD profiles directly annotated in individual cells demonstrated prominent variability in individual mouse cells[32]. The possible contribution of stochastic fluctuations of captured contacts in sparse single-cell Hi-C matrices into this apparent variability was not analyzed[32]. More comprehensive observations were made when super-resolution microscopy (Hi-M, 3D-SIM) coupled with high-throughput hybridization was used to analyze chromatin folding in individual cells at a kilobase-scale resolution. These studies demonstrated chromosome partitioning into TADs in individual mammalian cells and confirmed a trend for colocalization of CTCF and cohesin at TAD boundaries, although the positions of boundaries again demonstrated significant cell-to-cell variability[35]. Condensed chromatin domains coinciding with population TADs were also observed in *Drosophila* cells[36,37]. In accordance with previous observations made in cell population Hi-C studies[24], the obtained results suggested that partitioning of the *Drosophila* genome into TADs

was driven by the stochastic contacts of chromosome regions with similar epigenetic states at different folding levels[38].

Although studies performed using FISH and multiplex hybridization allowed to construct chromatin interaction maps with a very high resolution[35], they cannot provide genome-wide information. Here, we present single-nucleus Hi-C (snHi-C) maps of individual *Drosophila* cells with a 10-kb resolution. These maps allow direct annotation of TADs that appear to be non-hierarchical. At least 50% of TAD boundaries identified in each individual cell bear active chromatin marks and are highly reproducible between individual cells.

To comply with the Thesis content, below I propose only the results of the group related to:

(i) statistical analyses of the snHi-C maps aimed to prove that the maps are not random realizations provided some average characteristics of the spatial chromatin folding and answer the question of the correlation length of the genome at which the correlations between the pair contacts vanish (Section I);

(ii) annotation of the maps into TADs by means of the representation of a snHi-C map as a sparse network with intrinsic contiguous communities and using the non-backtracking operator specialized in community detection tasks in sparse random networks (Section II).

RESULTS

Section I. Marginal scaling (MS) and marginal scaling and stickiness (MSS) models

We carried out the statistical analysis of the single-cell Hi-C maps to provide statistical arguments supporting the premise that the clustering observed in snHi-C contact matrices "is not random". For this, we used two different models of a polymer network based on Erdos-Renyi graphs, where bins of the contact map resemble graph vertices, and contacts between bins are graph edges[39]:

a) In the MS model, we require the probability of contact between nodes to respect the contact probability of the experimental contact map, i.e. $\mathrm{P(s)} = \mathrm{P_c}(|\mathrm{i}-\mathrm{j}|)$. Decay of the contact probability originates from the intrinsic linear connectivity of the chromatin nodes; therefore, it is an important ingredient for studying fluctuations in a polymer network. The probability of the link between i and j in the random graph $\mathrm{i,j} = 1, 2 \ldots, \mathrm{N}$ is, thus, defined as follows:

$$\mathrm{p_{ij}} = \frac{\mathrm{P_c}(|\mathrm{i}-\mathrm{j}|)}{\sum_{\mathrm{s}=1}^{\mathrm{N}-1}(\mathrm{N}-\mathrm{s})\mathrm{P_c(s)}}\mathrm{N_c}, \qquad (1)$$

where the normalization factor in the denominator guarantees that the mean number of links in the graph equals $\mathrm{N_c}$, (i.e., the number of experimentally observed links in each single cell).

To obtain the average scaling, we merge all contacts from the available single cells and compute the average $\mathrm{P_c(s)}$. Given the probability p_{ij} by Eq. 1, we randomly generate adjacency matrices that have a homogenous distribution of contacts along the diagonals and do not respect local peculiarities of the bins, such as insulation score, acetylation, and protein affinity. Nevertheless, some non-homogeneity (clustering) of contacts still emerges as a result of stochasticity in each realization of this graph (Fig. 1e).

b) the MSS model introduces probabilistic non-homogeneity along the diagonals of the adjacency matrices through definition of the "stickiness" of bins. Specifically, under "stickiness", we understand a non-selective affinity $\mathrm{k_i}$ of a bin *i*

to other bins; the probability that the bin i forms a link with any other bin in the polymer graph is proportional to its stickiness. Thus, the clusters of contacts close to the main diagonal of contact matrices form as a result of different “stickiness" of bins in the MSS model. Stickiness might effectively emerge as a result of a particular distribution of “sticky” proteins, such as PcG proteins known to mediate bridging interactions between nucleosomes and to participate in stabilization of the repressed chromatin state.

Assuming that the stickiness is distributed independently of the polymer scaling $P_c(|i-j|)$, we use the following expression for the probability of the link, p_{ij}, in the MSS model:

$$p_{ij} = \frac{k_i k_j P_c(|i-j|)}{\sum_{i<j} k_i k_j P_c(|i-j|)} N_c \quad (2)$$

To derive the values of stickiness, we calculated the coverage at each bin in the merged contact map $\widetilde{k_l}$, which stands for the average number of contacts at a particular bin. Due to the polymer scaling, the rates of contacts along each row (column) vary. Thus, $\widetilde{k_l}$ is not equal to stickiness, $\widetilde{k_l} \neq k_i$. To determine the stickiness values k_i, one should correlate the experimental coverage $\widetilde{k_l}$ with the theoretical mean number of contacts per bin, according to Eq. 2:

$$\widetilde{k_l} = \sum_j p_{ij} = k_i \alpha_i \quad (3)$$

where is “activity” of surrounding bins, measured for the i-th bin:

$$\alpha_i = \frac{1}{Z}\sum_j k_j P_c(|i-j|), \; Z = \frac{1}{N_c}\sum_{i<j} k_i k_j P_c(|i-j|) \quad (4)$$

Eq. 3 sets a system of N non-linear equations that cannot be solved analytically. To determine the stickiness values, we implement the numerical method of iterative approximations. Namely, we start with:

$$k_i^{(0)} = \widetilde{k_l}, \alpha_i^{(0)} = \alpha_i(\widetilde{k_l}) \quad (5)$$

and recalculate $k_i^{(1)}$ using equations (3)–(4) at the second step. After several recursive steps, we find good convergence of the stickiness and activity to their limiting values k_i^{∞} and α_i^{∞}. In particular, the derived values of the stickiness provide a good estimate for the averaged theoretical coverage $\widetilde{k_l}$ as compared to the

experimental coverage; see Fig. 1f,g. Therefore, the derived null-model of single-cell maps reproduces, on average, the observed coverage of contacts of each bin by means of the individual stickiness assignment. We would like to point out the difference between the limiting values of the stickiness and $\widetilde{k_i}$, used as a starting approximation in the iterative procedure; Fig. 1h. This difference is a result of the non-homogeneous redistribution of contacts at each particular row in accordance with the marginal polymeric scaling $P_c(|i-j|)$.

Number of contacts in windows

The MS and MSS models introduced above demonstrate apparent clustering of generated contacts close to the main diagonal in realizations of adjacency matrices. In the MS model, this is purely due to fluctuations: the mean weight of the link $w_{ij} = p_{ij} = p_s$ depends only on the genomic distance between the bins $s = |i-j|$ in the respective Poisson version of the weighted network. In contrast, in the MSS model, the non-homogeneity of bin sicknesses allows for a deterministic non-homogeneous distribution of contacts along the main diagonal.

To statistically compare the clustering of contacts generated by the two models with the clustering in experimental single cell Hi-C maps, we studied distributions of the number of contacts in certain "windows" of different sizes. The inspected windows are isosceles triangles with the base located on the main diagonal and having the angle with the congruent sides. These windows look like TADs but, in contrast to the latter, have a fixed size throughout the genome.

At a given window size W, we sampled the number of contacts falling in the defined windows in each snHi-C map. We compared the samples originating from 100 random MS-generated maps and 100 random MSS-generated maps with derived limiting values of stickiness (see the previous section for discussion of the models).

Note that in the theoretical models (MS and MSS), all contacts are statistically independent: in both models, the number of contacts falling in a window of size can be interpreted as a number of "successes" occurring independently in a certain fixed interval. In the MS model, the "success" rate is constant along each diagonal; thus,

for rather sparse MS maps (i.e. sufficiently small rates), one would expect the observed contacts in the windows to follow the Poisson distribution. In the MSS maps, the stickiness distributions introduce non-homogeneity to "success" rates along the diagonals; however, as our analyses suggest, the random MSS maps exhibit much more satisfactory Poisson statistics than their original experimental counterparts; Fig. 1j,k.

Deviations from the Poisson statistics of the snHi-C contact maps are evaluated by the *p*-value of the χ^2 goodness of fit test (Fig. 1k). The heatmaps of the common logarithm of *p*-values for the top-10 single cells and the corresponding MS and MSS maps are presented in Fig. 1j. The random maps (the second and third rows) demonstrate reasonably even distributions of the *p*-values across distinct single cells that rarely enter below the significance level $\alpha = 10^{-5}$. Several atypically low *p*-values correspond either to extremely dense single cells and small window sizes (upper-left corner), for which the sparse Poisson limit is violated, or to a quite uneven distribution of stickiness for a given chromosome. Notably, the snHi-C maps demonstrate remarkable deviations from the Poisson statistics for small window size $W < 40$ bins (< 400 kb). As can be seen from the heatmaps (Fig. 1j) the χ^2 test rejects the null hypothesis at the significance level $\alpha = 10^{-5}$ for most of the single cells at small scales. Therefore, the probability that the experimental contact maps are described by the Poisson statistics is significantly low (α).

To understand the source of inconsistency between the experimental and Poisson distributions, we plotted the histograms of the number of contacts along with their best Poisson-fit for $W = 10$ (Fig. 1k, left) and $W = 40$ (Fig. 1k, right). The presence of large-scale heavy tails and low-scale shoulders in the experimental histograms results in the rejection of the null hypothesis.

Finally, the samples corresponding to larger windows are notably better described by the Poisson distribution, exhibiting a level of *p*-values similar to the random maps. The crossover $W_0 \approx 40$ (400 kb) corresponds to the scale of 3–4 typical TADs; this implies that the positioning of the contacts inside a single TAD is sufficiently correlated. Correlations between the contacts of different pairs of loci

can originate from a specific non-ideal folding of chromatin (e.g., fractal globule) or be a signature of active processes (e.g., loop extrusion) operating at the scale of one TAD. Larger window sizes accumulate contacts from different TADs, whereas most of the inter-TADs contacts are much less correlated. As a result, we see reasonable Poisson statistics of the number of contacts from larger windows with $W > W_0$. Taken together, we conclude that correlations in contacts is a structural feature of experimental single cell maps and that clusters (TADs) identified in the maps cannot be reduced to random fluctuations imposed by the white noise or imperfections of the experimental setup.

Figure 1

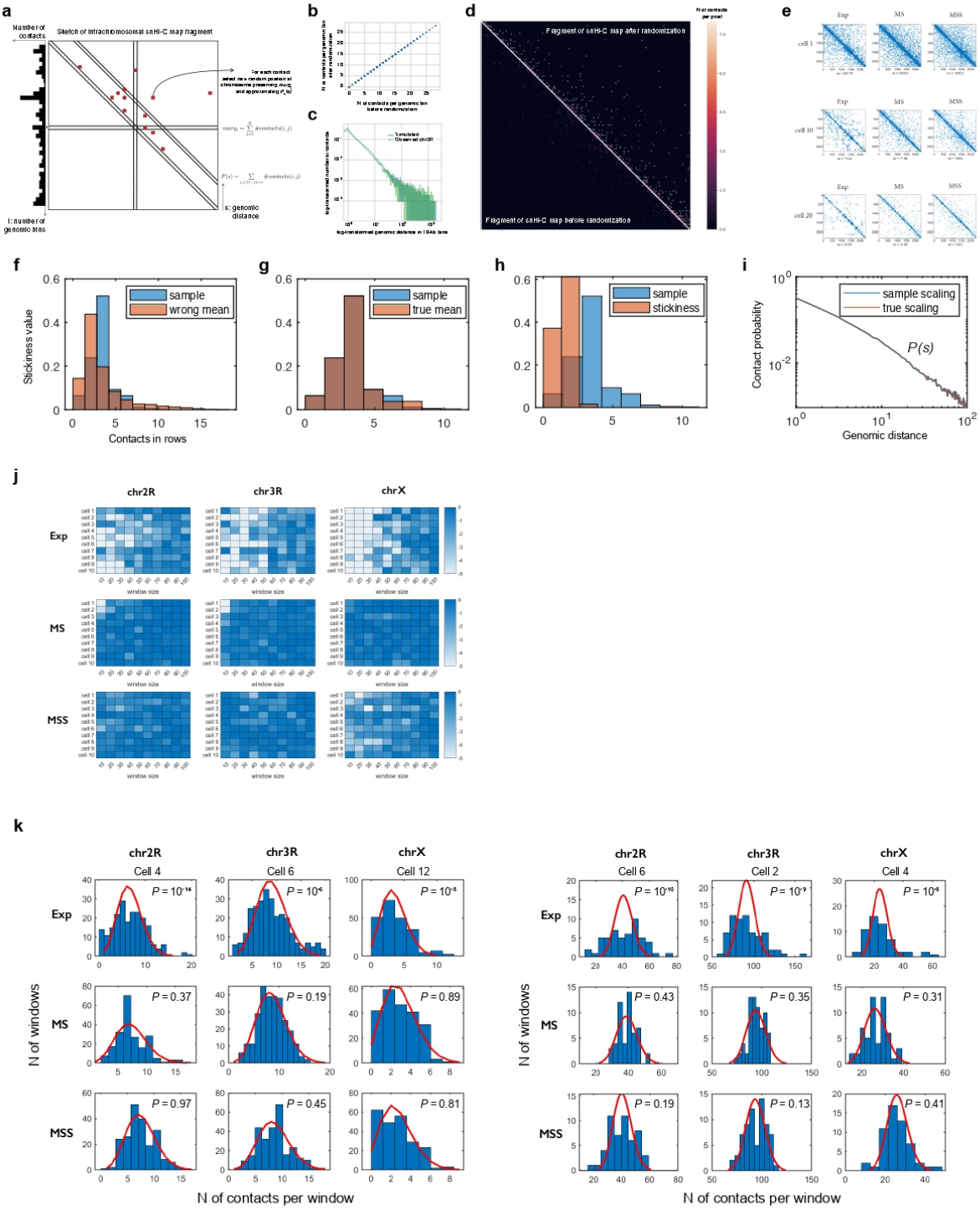
a
Sketch of intrachromosomal snHi-C map fragment
Number of contacts
For each contact select new random position of chromosome preserving marg. and approximating P(s)
s: genomic distance
I: number of genomic bins
b
N of contacts per genomic bin after randomization
N of contacts per genomic bin before randomization
c
Simulated
Observed chr2R
log-transformed number of contacts
log-transformed genomic distance in 10Kb bins
d
Fragment of snHi-C map after randomization
Fragment of snHi-C map before randomization
N of contacts per pixel
e
Exp
MS
MSS
cell 1
cell 10
cell 20
f
Stickiness value
sample
wrong mean
Contacts in rows
g
sample
true mean
h
sample
stickiness
i
Contact probability
sample scaling
true scaling
P(s)
Genomic distance
j
chr2R
chr3R
chrX
Exp
MS
MSS
window size
k
chr2R
chr3R
chrX
Cell 4
Cell 6
Cell 12
Cell 2
Exp
MS
MSS
N of windows
P = 10^-14
P = 10^-6
P = 10^-5
P = 10^-10
P = 10^-7
P = 10^-6
P = 0.37
P = 0.19
P = 0.89
P = 0.43
P = 0.35
P = 0.31
P = 0.97
P = 0.45
P = 0.81
P = 0.19
P = 0.13
P = 0.41
N of contacts per window
N of contacts per window

Figure 1. snHi-C maps do not follow the rules of random distribution of contacts.

a Background model of snHi-C interactions (MSS; Marginal Scaling with Stickiness Model). For each intrachromosomal map of the single nucleus, the number of contacts in each row or column (marginal distribution of the number of contacts) and the probability of contact for certain genomic distance $P_c(s)$ are calculated. Then, the positions of observed contacts are randomly selected from all possible positions on the same chromosome so that the marginal distribution is exactly the same, and $P_c(s)$ is approximately the same.

b Scatter plot of the initial number of contacts per genomic bin in snHi-C map and after randomization for chr2R of Cell 1 with 107,823 unique contacts.

c Contact probability $P_c(s)$ for chr2R of Cell 1 before (green) and after (blue) randomization.

d Cell 1 snHi-C interactions map for a region of chr2R (lower triangle) and randomized background control (upper triangle). Note the presence of contact clusters at the diagonal both in original and reshuffled data.

e Examples of experimental (Exp) single-cell Hi-C maps with those simulated using the MSS and MS models.

f-h Derivation of the stickiness values (Y axis) given the coverage of bins (numbers of contacts in rows, X axis) obtained by iterative approximations for the MSS model and chr2L (merged snHi-C data were used). At each step, the theoretical average for the coverage $\widetilde{k_i}$ at each particular bin i is recomputed, and the stickiness values k_i are corrected until convergence with experimental coverage is achieved.

f Histograms of observed coverage from merged snHi-C map (blue) and of theoretical values (brown) calculated with $k_i = \widetilde{k_i}$ (red) at the first step of the iterative procedure; wrong mean – computed with wrong stickiness.

g The same histogram as in (**f**) after a series of iterative corrections of the stickiness values which led to convergence towards the limiting values $k_i = k_i^{\infty}$. The resulting distribution of the coverage (red) reproduces the experimental values;

true mean – computed with true stickiness, which is the outcome of the iterative procedure.

h Distributions of the experimental coverage $\widetilde{k_i}$ (blue) and of the limiting stickiness k_i^{∞}(red) are significantly different. Notably, the stickiness values k_i^{∞} have lower variance than the experimental coverage because the latter incorporate fluctuations of the contact probability, $P_c(s)$.

i Initial and limiting scaling probability functions remain unchanged after the iterative approach.

j Heatmaps of $\log_{10}$ of p-values for the χ^2 test for the top 10 cells sorted to their contact densities. Experimental, MS, and MSS distributions of the number of contacts in windows of different size $\mathrm{W} = 10, 20, \ldots, 100$ bins at the main diagonal are statistically compared with the corresponding Poisson distributions. The original single cells demonstrate higher deviations from the Poisson statistics than the random models for small window sizes $\mathrm{W} \leq 40$ bins (400 kb). The χ^2 test rejects the null hypothesis at the significance level $\alpha = 10^{-5}$ for most of the top-density single cells at the TAD scale. Clustering of contacts at the scale of TADs cannot be explained by the random models at the significance level $\alpha = 10^{-5}$.

k Experimental, MS and MSS distributions of the number of contacts in windows of the size $W = 10$ bins (100 kb) (left) and $W = 40$ bins (400 kb) (right) displaced at the main diagonal $\Delta = 0$ and their best Poisson distribution (in red). Three cells and three chromosomes are considered. The corresponding p-values of the χ^2 test are shown in each plot.

Section II. Non-backtracking approach for annotation of TADs in single cells contact maps

The chromatin network, constructed on the basis of the single-cell Hi-C data, can be classified as sparse (i.e. the number of actual contacts per bin in a single-cell contact matrix (adjacency matrix of the network) is much less than the matrix size N). The sparsity of the data significantly complicates the community detection problem in single cells. It is known that upon dilution of the network, there is a fundamental resolution threshold for all community detection methods[40]. Furthermore, traditional operators (adjacency, Laplacian, modularity) fail far above this resolution limit (i.e. their leading eigenvectors become uncorrelated with the true community structure above the threshold)[41]. That is explained by the emergence of tree-like subgraphs (hubs) overlapping with true clusters in the isolated part of the spectrum for these operators. Localization on the hubs, but not on true communities in the network, is a drawback of all conventional spectral methods in the sparse regime.

To overcome the sparsity issue and to make spectral methods useful in the sparse regime, Krzakala et al.[41] proposed to construct the transfer-matrix of non-backtracking random walks (NBT) on a directed network. The NBT operator B is defined on the edges $i \to j, k \to l$ as follows:

$$B_{i\to j,k\to l} = \delta_{il}(1 - \delta_{jk}) \qquad (6)$$

By construction, NBT walks cannot revisit the same node on the subsequent step and, thus, they do not concentrate on hubs. It has been shown that the non-backtracking operator is able to resolve the community structure in a sparse stochastic block model up to the theoretical resolution limit. In recently published paper[42], we have proposed the neutralized towards the expected contact probability NBT operator for the sake of a large-scale splitting of a sparse polymer network into two compartments.

Here, we are interested in the small-scale clustering into TADs, for which the conventional NBT operator is appropriate. To eliminate the compartmental signal from the data, we first cleansed all chromosome contact matrices starting from the

diagonal, corresponding to 1 Mb separation distance (100th diagonal in the 10-kb resolution). To respect the polymeric nature of the contact matrices, we have filled all empty cells on the leading sub-diagonals with 1. Then, the NBT spectra of all single-cell contact matrices were computed. The majority of eigenvalues of the non-Hermitian NBT operator are located inside the disc in a complex plane, and some number of isolated eigenvalues with large amplitudes lie on the real axis. The edge of the isolated part of the spectrum was defined as the real part of the largest in absolute value eigenvalue with a non-zero imaginary part. All eigenvalues λ_i such that $Re(\lambda_i) > r_c$ are isolated, and the corresponding eigenvectors correlate with annotation into the TADs. The position of the spectral edge, determined by the procedure above, has been found to be very close to the edge of the disk for the stochastic block model $r_c = \sqrt{\langle d \rangle^{-1} \langle \frac{d}{d-1} \rangle}$, where d is the vector of degrees[43]. The typical number of the isolated eigenvalues was around 100 for dense contact matrices and somewhat less for sparser ones. The leading eigenvectors define the coordinates $u_j^{(i)}, j = 1,2, \dots, N$ of the nodes (bins) of the network in the space of reduced dimension $k \ll N$. At the second step, the clustering of the data was performed using the spherical k-means method, realized in the Python library *spherecluster*[44]. The dimension of the space k establishes a lower bound on the number of clusters because the leading eigenvectors are linearly independent. To take into account the hierarchical organization of TADs, we have communicated to the spherical k-means the number of clusters somewhat larger than the lower bound. Although the final splitting was found to be not particularly sensitive to this number, we have chosen to split the network into *2.5*k* clusters in order to obtain the same mean amount of TADs per chromosome as with the modularity method (171 TADs).

The annotations produced by the spherical k-means on the single-cell Hi-C matrices were contiguous (i.e. the clusters were sequence respective, thus resembling TADs). The clusters (i) of size less than 30 kb and (ii) of size with amount of contacts equal to $2(l-1)$ (i.e. with no contacts other than on the sub-diagonals) were excluded from the set as the inter-TADs regions. The ultimate

median size of the TADs across all single cells obtained by this algorithm was 110 kb (from 60 kb to 260 kb), and the mean chromosome coverage was 82% (from 57% to 93%). The same analyses of shuffled contact maps have revealed a similar number, size, and coverage of the domains, formed purely due to fluctuations. The boundaries of the NBT TADs in single cells were significantly conserved from cell to cell: the mean pairwise fraction of matched boundaries was 44% for all the cells and 59% for the five densest ones (for the shuffled cells with preservation of stickiness and scaling, see the MSS model; the mean pairwise fraction was 38% and 50% for the five densest cells).

Regarding the comparison of TAD boundaries with the modularity approach, the mean fraction of conserved modularity boundaries is somewhat less—42% for all pairs of cells in the analyses and 52% for the five densest cells, whereas the number of TADs per chromosome is the same in the two methods (171). Between the two methods, the mean number of matched boundaries for the corresponding cells is 61%.

Figure 2

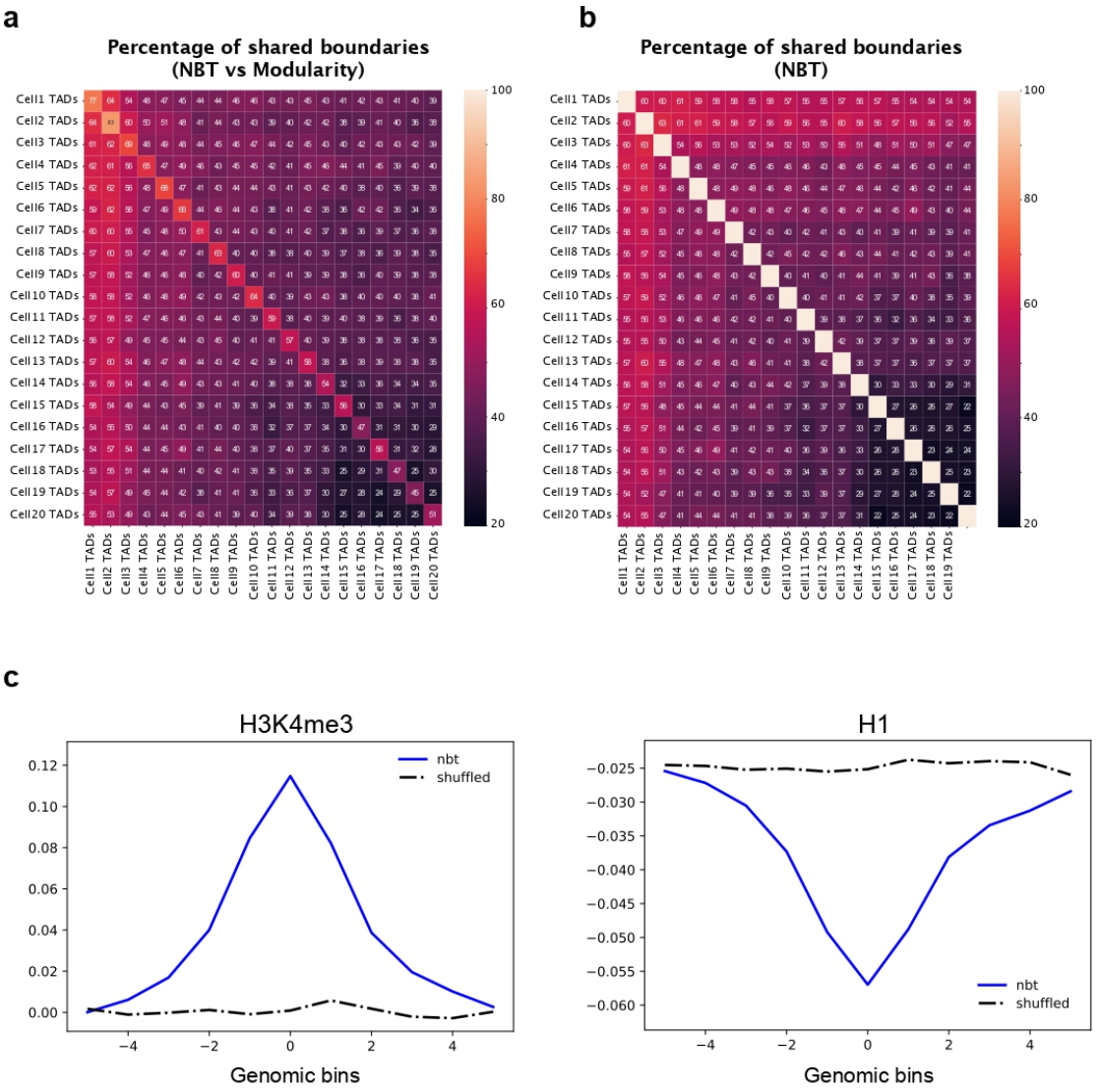
a
Percentage of shared boundaries
(NBT vs Modularity)
b
Percentage of shared boundaries
(NBT)
Cell1 TADs
Cell2 TADs
Cell3 TADs
Cell4 TADs
Cell5 TADs
Cell6 TADs
Cell7 TADs
Cell8 TADs
Cell9 TADs
Cell10 TADs
Cell11 TADs
Cell12 TADs
Cell13 TADs
Cell14 TADs
Cell15 TADs
Cell16 TADs
Cell17 TADs
Cell18 TADs
Cell19 TADs
Cell20 TADs
100
80
60
40
20
c
H3K4me3
H1
nbt
shuffled
Genomic bins

Figure 2. NBT as an approach for identification of TAD boundaries.

a Percentage of TAD boundaries shared between NBT- and modularity-derived TAD segmentations in individual cells. The mean percentage of shared boundaries is 61%.

b Percentage of TAD boundaries shared between single cells for the NBT TAD calling procedure. The mean percentage of shared boundaries is 42%.

c Epigenetic profiles around the NBT-identified TAD boundaries.

DISCUSSION

Folding of interphase chromatin in eukaryotes is driven by multiple mechanisms operating at different genome scales and generating distinct types of the 3D genome features[16,20]. In mammalian cells, cohesin-mediated chromatin fiber extrusion mainly impacts the genome topology at the scale of approximately 100–1000 kb by producing loops, resulting in the formation of TADs[18,19] and establishing enhancer-promoter communication. Chromatin loop formation by the loop extrusion complex (LEC) in mammalian cells is a substantially deterministic process due to the preferential positioning of loop anchors encoded in DNA by CTCF binding sites (CBS). The cohesin-CTCF molecular tandem modulates folding of intrinsically disordered chromatin fiber[16,23]. On the other hand, association of active and repressed gene loci in chromatin compartments[13,14], and formation of Polycomb and transcription-related nuclear bodies in both mammalian and *Drosophila* cells shape the 3D genome at the scale of the whole chromosome. These associations appear to be stochastic: a particular Polycomb-bound or transcriptionally active region in individual cells interacts with different partners located across a wide range of genomic distances.

Here, for the first time, we applied the single-nucleus Hi-C to probe the 3D genome in individual *Drosophila* cells at a relatively high resolution that was not achieved previously in single-cell Hi-C studies. Based on our observations, we suggest that, in *Drosophila*, both deterministic and stochastic forces govern the chromatin spatial organization.

We found that the entire individual *Drosophila* genomes were partitioned into TADs; this observation supports the results of recent super-resolution microscopy studies[37]. TAD profiles are highly similar between individual *Drosophila* cells and demonstrate lower cell-to-cell variability as compared to mammalian TADs. According to our model[24], large inactive TADs in *Drosophila* are assembled by multiple transient electrostatic interactions between non-acetylated nucleosomes in transcriptionally silent genome regions. Conversely, TAD boundaries and inter-TAD regions at the 10-kb resolution of Hi-C maps in *Drosophila* were found to be

formed by transcriptionally active chromatin. This result may explain why TADs in individual cells occupy virtually the same genomic positions. Gene expression profile is a characteristic feature of a particular cell type, and, thus, should be relatively stable in individual cells within the population. In agreement with this, we demonstrated that invariant TAD boundaries present in a major portion of individual cells were highly enriched in active chromatin marks. Moreover, stable boundaries were also largely conserved in other cell types, possibly due to the fact that TAD boundaries were frequently formed at the position of housekeeping genes.

In contrast to stable TAD boundaries, the boundaries that demonstrate cell-to-cell variability bear silent chromatin. Some cell-specific TAD boundaries may originate at various positions due to a putative size limit of large inactive TADs or other restrictions in chromatin fiber folding. Indeed, it appears that the assembly of randomly distributed TAD-sized self-interacting domains is an intrinsic property of chromatin fiber folding[35]. In mammals, the positioning of these domains is modulated by cohesin-mediated DNA loop extrusion[35], whereas in *Drosophila*, it may be modulated by segregation of chromatin domains bearing distinct epigenetic marks[16,23]. Even if cell-specific and unstable TAD boundaries are distributed in a random fashion, they should be depleted in active chromatin marks because active chromatin regions are mainly occupied by stable TAD boundaries. We also cannot exclude that variable boundaries and the TAD boundary shifts are caused by local variations in gene expression and active chromatin profiles in individual cells that we cannot assess simultaneously with constructing snHi-C maps.

Our results are also compatible with an alternative mechanism of TAD formation. Given that the above-mentioned cohesin-driven loop extrusion is evolutionarily conserved from bacteria to mammals, it is compelling to assume that extrusion works in *Drosophila* as well. Despite the presence of all potential components of LEC (cohesin, its loading and releasing factors), TAD boundaries in *Drosophila* are not significantly enriched with CTCF[24,25] and do not form CTCF-enriched interactions or TAD corner peaks. These observations suggest that the binding sites of CTCF or other distinct proteins do not constitute barrier elements

for the *Drosophila* LEC even if these proteins are enriched in TAD boundaries; this may be due to some other properties of a genomic region. For example, stably bound cohesins were proposed to act as the barriers for cohesin extrusion in yeast.

Active transcription interferes with DNA loop extrusion. Because TAD boundaries in *Drosophila* are highly transcribed, we propose that open chromatin with actively transcribing polymerase and/or a high density of chromatin remodeling complexes could serve as a barrier for the *Drosophila* LEC. Contrary to the strictly positioned and short CBSs in mammals, active loci flanking *Drosophila* TADs represent relatively extended regions up to several dozens of kb in length. Probabilistic termination of LEC at varying points within such regions in different cells of the population could explain the absence of canonical loop signals and the presence of strong compartment-like interactions between active regions flanking a TAD. This model also provides a potential explanation for the relatively high stability of TAD positioning in individual *Drosophila* cells in comparison to mammals. A relative permeability of CBSs in mammalian cells allows LEC to proceed through thousands of kilobases and to produce large contact domains[17]. Extended active regions acting as "blurry" barrier elements where LEC termination occurs at multiple points, should stop the LEC more efficiently, making the TAD pattern more stable and pronounced.

Taken together, the order in the *Drosophila* chromatin 3D organization is manifested in a TAD profile that is relatively stable between individual cells and likely dictated by the distribution of active genes along the genome. On the other hand, our molecular simulations of individual haploid X chromosomes indicate a prominent stochasticity in both the form of individual TADs and the overall folding of the entire chromosome territory. According to our data, the active A-compartment is radially detectable in individual cells, and the profiles of interaction between individual active regions are highly variable between individual cells. Notably, this also holds true for Polycomb-occupied loci that are known to shape chromatin fiber in living cells.

Although these highly variable long-range interactions of active regions and Polycomb-occupied loci are closely related to the shape of chromosome territory (CT), the cause-and-effect relationships between them and the stochastic nature of the cell-specific chromatin chain path are currently unclear. The main question to be answered by future studies is whether these interactions are fully stochastic or at least partially specific. The possible molecular mechanisms that may provide specific communication between remote genomic loci separated by up to megabases of DNA are not known. In a scenario of the absence of any specificity, the pattern of contacts inside A-compartment and within Polycomb bodies in a particular cell is established by stochastic fluctuations of the large-scale chromatin fiber folding. In this case, the large-scale chromatin fiber folding dictates the cell-specific location of Polycomb-enriched and active chromatin regions in the 3D nuclear space. The formation of Polycomb bodies and transcription-related chromatin hubs is achieved by confined diffusion of these regions and might be further stabilized by specific protein-protein interactions and liquid-liquid phase separation. This mechanism allows to sort through alternative configurations of the 3D genome and to transiently stabilize those that are functionally relevant under specific conditions. A balance between the order and the stochasticity appears to be an intrinsic property of nuclear organization that enables rapid adaptation to changing environmental conditions.

REFERENCES

1 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).

2 Kim, T. H. & Dekker, J. 3C-Based Chromatin Interaction Analyses. *Cold Spring Harbor protocols* **2018**, doi:10.1101/pdb.top097832 (2018).

3 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

4 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).

5 Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472, doi:10.1016/j.cell.2012.01.010 (2012).

6 Lupianez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025, doi:10.1016/j.cell.2015.04.004 (2015).

7 Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res* **24**, 390-400, doi:10.1101/gr.163519.113 (2014).

8 Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol CeLL* **62**, 668-680, doi:10.1016/j.molcel.2016.05.018 (2016).

9 Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269, doi:10.1038/nature19800 (2016).

10 Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat Genet* **52**, 294-305, doi:10.1038/s41588-019-0564-y (2020).

11 Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56, doi:10.1038/nature24281 (2017).

12 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324, doi:10.1016/j.cell.2017.09.026 (2017).

13 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

14 Hildebrand, E. M. & Dekker, J. Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem Sci* **45**, 385-396, doi:10.1016/j.tibs.2020.01.002 (2020).

15 Drucker, J. L. & King, D. H. Management of viral infections in AIDS patients. *Infection* **15 Suppl 1**, S32-33, doi:10.1007/BF01650109 (1987).

16 Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci U S A* **115**, E6697-E6706, doi:10.1073/pnas.1717730115 (2018).

17 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

18 Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell reports* **15**, 2038-2049, doi:10.1016/j.celrep.2016.04.085 (2016).

19 Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465, doi:10.1073/pnas.1518552112 (2015).

20 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**, 789-800, doi:10.1038/s41576-018-0060-8 (2018).

21 Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573-3599, doi:10.15252/embj.201798004 (2017).

22 Matthews, N. E. & White, R. Chromatin Architecture in the Fly: Living without CTCF/Cohesin Loop Extrusion?: Alternating Chromatin States Provide a Basis for Domain Architecture in Drosophila. *BioEssays* **41**, e1900048, doi:10.1002/bies.201900048 (2019).

23 Rowley, M. J. *et al.* Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol CeLL* **67**, 837-852 e837, doi:10.1016/j.molcel.2017.07.022 (2017).

24 Ulianov, S. V. *et al.* Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* **26**, 70-84, doi:10.1101/gr.196006.115 (2016).

25 Wang, Q., Sun, Q., Czajkowsky, D. M. & Shao, Z. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature communications* **9**, 188, doi:10.1038/s41467-017-02526-9 (2018).

26 Ramirez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature communications* **9**, 189, doi:10.1038/s41467-017-02525-w (2018).

27 Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol CeLL* **58**, 610-620, doi:10.1016/j.molcel.2015.04.005 (2015).

28 Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914, doi:10.1126/science.aab1601 (2015).

29 Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* **17**, 72, doi:10.1186/s13059-016-0944-x (2016).

30 Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiol Mol Biol Rev* **79**, 347-372, doi:10.1128/MMBR.00006-15 (2015).

31 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).

32 Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114, doi:10.1038/nature21711 (2017).

33 Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61-67, doi:10.1038/nature23001 (2017).

34 Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600-3618, doi:10.15252/embj.201798083 (2017).

35 Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, doi:10.1126/science.aau1783 (2018).

36 Cardozo Gizzi, A. M. *et al.* Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol CeLL* **74**, 212-222 e215, doi:10.1016/j.molcel.2019.01.011 (2019).

37 Szabo, Q. *et al.* TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science advances* **4**, eaar8082, doi:10.1126/sciadv.aar8082 (2018).

38 Cattoni, D. I. *et al.* Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nature communications* **8**, 1753, doi:10.1038/s41467-017-01962-x (2017).

39 Anderson, G. W., Guionnet, A. & Zeitouni, O. *An introduction to random matrices*. (Cambridge University Press, 2010).

40 Decelle, A., Krzakala, F., Moore, C. & Zdeborova, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical review. E*, Statistical, nonlinear, and soft matter physics **84**, 066106, doi:10.1103/PhysRevE.84.066106 (2011).

41 Krzakala, F. et al. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences* **110,** 20935-20940, doi:10.1073/pnas.1312486110 (2013).

42 Polovnikov, K., Gorsky, A., Nechaev, S., Razin, S. V. & Ulianov, S. V. Non-backtracking walks reveal compartments in sparse chromatin interaction networks. *Scientific reports* **10**, doi:10.1038/s41598-020-68182-0 (2020).

43 Newman, M. E. J. Spectral methods for community detection and graph partitioning. Physical Review E **88**, doi:10.1103/PhysRevE.88.042822 (2013).

44 Banerjee, A., Dhillon, I. S., Ghosh, J. & Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J Mach Learn Res* **6**, 1345-1382 (2005).

7. Non-backtracking walks reveal compartments in sparse chromatin interaction networks

Introduction

Chromatin communities stabilized by protein machinery play essential role in gene regulation and refine global polymeric folding of the chromatin fiber. However, treatment of these communities in the framework of the classical network theory (stochastic block model, SBM) does not take into account intrinsic linear connectivity of the chromatin loci. Here we propose the "polymer"block model, paving the way for community detection in polymer networks. On the basis of this new model we modify the non-backtracking flow operator and suggest the first protocol for annotation of compartmental domains in sparse single cell Hi-C matrices. In particular, we prove that our approach corresponds to the maximum entropy principle. The benchmark analyses demonstrates that the spectrum of the polymer non-backtracking operator resolves the true compartmental structure up to the theoretical detectability threshold, while all commonly used operators fail above it. We test various operators on real data and conclude that the sizes of the non-backtracking single cell domains are most close to the sizes of compartments from the population data. Moreover, the found domains clearly segregate in the gene density and correlate with the population compartmental mask, corroborating biological significance of our annotation of the single cells into active and inactive compartments.

Contribution

I have developed the polymer stochastic block model and the non-backtracking flow operator, neutralized to the polymer contact probability. I have established the connection with the generalized modularity and have proved that partition of a chromatin network into two compartments by means of the leading eigenvector of the proposed operator responds to the maximum entropy principle. I have tested

the suggested framework on the benchmark, emulating compartmentalization in single cells. I have realized the approach on real sparse data and have demonstrated biological significance of the annotation by profiling the single cell domains using the GC content and the leading eigenvector of the population-averaged Hi-C matrix.

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Non-backtracking walks reveal compartments in sparse chromatin interaction networks

K. Polovnikov[1,2]✉, A. Gorsky[5,6], S. Nechaev[3,4], S. V. Razin[7,8] & S. V. Ulianov[7,8]

Chromatin communities stabilized by protein machinery play essential role in gene regulation and refine global polymeric folding of the chromatin fiber. However, treatment of these communities in the framework of the classical network theory (stochastic block model, SBM) does not take into account intrinsic linear connectivity of the chromatin loci. Here we propose the polymer block model, paving the way for community detection in polymer networks. On the basis of this new model we modify the non-backtracking flow operator and suggest the first protocol for annotation of compartmental domains in sparse single cell Hi-C matrices. In particular, we prove that our approach corresponds to the maximum entropy principle. The benchmark analyses demonstrates that the spectrum of the polymer non-backtracking operator resolves the true compartmental structure up to the theoretical detectability threshold, while all commonly used operators fail above it. We test various operators on real data and conclude that the sizes of the non-backtracking single cell domains are most close to the sizes of compartments from the population data. Moreover, the found domains clearly segregate in the gene density and correlate with the population compartmental mask, corroborating biological significance of our annotation of the chromatin compartmental domains in single cells Hi-C matrices.

Many real-world stochastic networks split into self-organized communities. Social networks feature circles of friends[1–3], colleagues[2], members of a karate club[1], communities of dolphins[4] etc. Cellular networks demonstrate modular organization, which optimizes crucial biological processes and relationships, such as synchronization of neurons in the connectome[5,6], efficiency of metabolic pathways[7,8]], genes specialization[9] or interaction between enhancers and promoters[10].

Interest to polymer modular networks has appeared recently in the context of genome spatial folding. Proximity of chromatin loci in space is believed to be deeply connected with gene regulation and function. Hi-C experiments[11–13] provide the genome-wide colocalization data of chromatin loci. As the main outcome of the experiment, large genome-wide matrices of contacts from each individual cell or from the population are produced. Analyses of these matrices has revealed that the eukaryotic genome is organized in various and biologically relevant communities, whose main function is to insulate some regions of DNA and to provide easy access to the others. In particular, the data collected from a population of cells suggest that transcribed ("active") chromatin segregates from the, "inactive" one, forming two compartments in the bulk of the nucleus[12,14]. Within compartments chromatin is organized further as a set of topologically-associated domains (TADs)[15–17] that regulate chromatin folding at finer scales. However, interpretation and validation of communities in individual cells remains vaguely defined due to sparsity of respective data.

The broad field of applications of stochastic modular networks has initiated the boost development of community detection methods. Spectral algorithms exploit the spectrum of various operators (adjacency, Laplacian, modularity) defined on a network to identify the number of communities and to infer the optimal network partition[18–22]. Typically, leading eigenvectors of these operators positively correlate with the true community

[1]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Skolkovo Institute of Science and Technology, Skolkovo, Russia 143026. [3]Interdisciplinary Scientific Center Poncelet (UMI 2615 CNRS), Moscow, Russia 119002. [4]Lebedev Physical Institute RAS, Moscow, Russia 119991. [5]Moscow Institute for Physics and Technology, Dolgoprudnyi, Russia. [6]Institute for Information Transmission Problems of RAS, Moscow, Russia. [7]Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia. [8]Faculty of Biology, M.V. Lomonosov Moscow State University, Moscow, Russia. ✉email: kipolovnikov@gmail.com

structure or with the underlying core-periphery organization of the network[23]. These algorithms, along with the majority of theoretical results in the field, are derived for the stochastic block model (SBM)[18,22] as an extension of Erdös-Rényi graphs[24] to graphs with explicitly defined communities. One of the strongest limitations of the SBM is that edges between vertices belonging to the same cluster inevitably attain equal weights. At the same time, biological networks typically have several levels of organization within their communities[25]. In particular, identification of several hierarchical levels in the network becomes tremendously important in the case of polymer networks, where different pairs of loci have marginally different probabilities to form a contact in space[26], caused by the frozen linear connectivity along the chain.

Even for simplest polymer systems the contact probability demonstrates a power-law behavior with the dimensional-dependent scaling exponent characterizing universal long-ranged behavior of polymer folding[27]. In this work we propose the "polymer stochastic block model" which reflects a specific global polymer network organization with explicit structuring into communities. The main new ingredient of the model under consideration is the average contact probability $P(s = |i - j|)$ between the pairs of loci (i, j) which is constant for standard non-polymeric networks, however cannot be neglected for polymers.

Chromatin single cell networks are not only polymeric, but also sparse[13,28]. It is known that upon reduction of the total number of edges in the network, there is a fundamental resolution limit for all community detection methods[22,29]. Furthermore, traditional operators (adjacency, Laplacian, modularity) fail far above this resolution limit, i.e. their leading eigenvectors become uncorrelated with the true community structure above the threshold[30]. That is explained by emergence of tree-like subgraphs (hubs) overlapping with true clusters in the isolated part of the spectrum for these operators. The edge of the spectral density of sparse networks is universal and demonstrates the so-called "Lifshitz tail"[31–34]. Localization on hubs, but not on true communities is a drawback of all conventional spectral methods in the sparse regime.

To prevent the effect of localization on hubs and to make spectral methods useful in sparse regime, Krzakala et al. proposed to deal with non-backtracking random walks on a directed graph that cannot revisit the same node on the subsequent step[30]. The crucial property of non-backtracking walks[35] is that they do not concentrate on hubs. It has been shown that the non-backtracking operator is able to resolve the community structure in sparse stochastic block model up to the theoretical resolution limit. Typically, the majority of eigenvalues of the non-backtracking operator (which is a non-symmetric matrix with complex eigenvalues) are located inside a disc in a complex plane, and a number of isolated eigenvalues lie on the real axis.

For the sake of community detection in sparse polymer networks we construct the polymer-type non-backtracking walks, appropriate for community detection in graphs with hidden linear memory ("polymeric background"). We establish the connection between this operator and the generalized polymer modularity, thus, bridging a gap with the maximum entropy principle. We test the performance of different spectral methods (with and without polymer background) on sparse artificial benchmarks of polymer networks that mimic compartmentalization in single cell Hi-C graphs. We show that polymer non-backtracking walks resolve the structure of communities up to the detectability threshold, while all other operators fail above it. In order to demonstrate efficiency of the method on real data, we partition a set of single cell Hi-C contact maps of mouse oocytes into active (A) and inactive (B) compartments by different operators. Found domains are shown to have similar sizes to the compartmental domains and correlate with the compartmental mask from the population-averaged data. Analyses of the GC content within the domains demonstrates enrichment and depression of the genes density in the two clusters, thus, corroborating their biological significance.

The structure of the paper is as follows. In Section "Stochastic block model with polymer contact probability" we propose the polymer stochastic block model, derive the entropy and the corresponding generalized modularity functional. In Section "Polymer non-backtracking flow operator" we discuss polymer non-backtracking walks, prove their robustness on the benchmarks emulating compartments, and, finally, test them on the real single cell data. In Section "Conclusion" we draw the conclusions.

Stochastic block model with polymer contact probability

Definition of the model. Characterize a N-bead polymer chain by coordinates $\{x_1, x_2, \ldots, x_N\}$ of monomers $i = 1, 2, \ldots, N$ and construct a corresponding topological graph $\mathcal{G} = (V, E)$ with the adjacency matrix A_{ij} (accounting for the bead's proximity in space). Such graphs are typically constructed upon processing of chromatin single cell Hi-C data and in computer simulations of DNA folding[11,12]. A graph $\mathcal{G}$ does not contain pairwise spatial distances of the polymer configuration, however, provides information on spatial proximity of monomers (or groups of monomers), which is usually of major biological relevance. For the 1-bin resolution of $\mathcal{G}$ the polymer beads (bins) are the nodes V. The edge between a pair of nodes (i, j) is defined by the condition $(i, j) \in E$ iff $|x_i - x_j| < \varepsilon$, where the threshold ε is some cutoff radius with which the contacts between the two loci are registered in Hi-C. Due to finite excluded volume of chromatin, the theoretical number of contacts per monomer that can be registered in single cell experiments is of order of few units, while the total size of the polymer chain, measured in number of beads, is huge ($N \sim 10^5$ in the 1-kb resolution for human chromosomes). Thus, the single cell contact matrices are essentially sparse[13,28]. Summation over realizations of adjacency matrices A_{ij} obtained from different cells results in a "population-averaged" matrix $\mathcal{A}_{ij}$. By construction, entries of the weight matrix $\mathcal{A}_{ij}$ are proportional to the probability that the spatial distance between monomers (i, j) is less than ε.

Already for the simplest configurations, such as a conformation of ideal polymer chain isomorphic to the random walk, the matrix $\mathcal{A}_{ij}$ is not expected to be uniform. This is due to a polymeric power-law behaviour of a contact probability,

$$P(s) \sim s^{-\alpha}, \quad \text{for } s = |i - j| \tag{1}$$

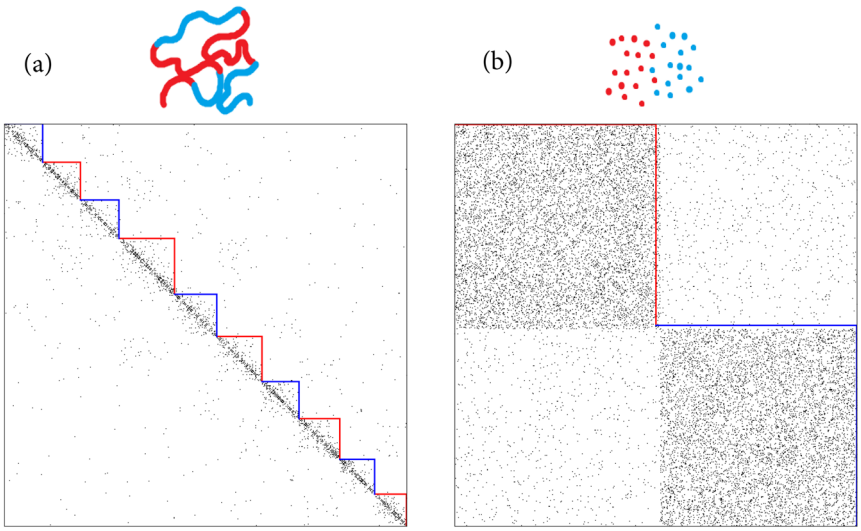

Figure 1. Adjacency matrices of $N = 1000$ with *two clusters* generated according to the (**a**) polymer stochastic block model ($w_{in} = 1, w_{out} = 0.1, P(s) = s^{-1}, \lambda = 100$) and (**b**) canonical stochastic block model ($w_{in} = 0.1, w_{out} = 0.01, \lambda = 500$). Vertices in the graph are enumerated by the polymer coordinate (**a**) and first all red, then all blue ones (**b**).

By definition, $P(s)$ is probability to find two beads of a linear chain, separated by a chemical distance s, close to each other in space. The critical exponent, α, is an important parameter, which characterizes the "memory" about the embedding of a polymer loop of length s in a D-dimensional space[27]. Such a memory can arise due to some equilibrium topological state of chromatin, or could be a result of partial relaxation of mitotic chromosomes[36]. Notable examples of α, typically appearing in the chromatin context for chain embedding in a three-dimensional space, are $\alpha = 3/2$ for ideal chain and $\alpha \approx 1$ for the crumpled globule[12,37–39].

Communities of folded chromatin refine the background (polymeric) contact probability at small scales and are biologically significant. We treat communities as canonical stochastic blocks[18,22] superimposed over the background. Stochastic block model is a network model in which N nodes of a network are split into q different groups G_i, $i = 1, 2, \ldots, q$ and the edges between each pair of nodes are distributed independently with a probability, depending on the group labels ("colors") of respective nodes. In a matrix of pairwise group probabilities $\Omega = \{\omega_{rt}\}$ with $(r, t) = 1, 2, \ldots, q$, any randomly chosen pair of nodes (i, j) (where $i \in G_r, j \in G_t$) is linked by an edge with probability ω_{rt}. The corresponding entry in the adjacency matrix A_{ij} is 1 with probability ω_{rt} and 0 otherwise. The sum of many such "single-cell" Bernoulli matrices generates an analogue of the "population-averaged" Hi-C matrix $\mathcal{A}_{ij}$ with Poisson distributed number of contacts with the mean $\langle \mathcal{A}_{ij} \rangle = \omega_{rt}$ where $i \in G_r, j \in G_t$. To the first approximation, the communities can be considered identical (known as a "planted" version of the model)

$$\Omega_{rt} = \begin{cases} w_{in}, & r = t \\ w_{out}, & r \neq t \end{cases} \tag{2}$$

Having (1) and (2), the simplest assumption one can come up with is that formation of compartments in chromatin is independent of the global memory of folding. Indeed, phenomenon of compartments is likely related to preferential interactions of nodes of the same epigenetic type (e.g., "active" or "inactive") and is modelled as a phase separation of block-copolymers[40]. This allows to suggest the factorization of (1) and (2), so that the final probability for the edge (i, j) reads

$$Prob_{ij} = P(|i - j|) \begin{cases} \omega_{in}, & r = t \\ \omega_{out}, & r \neq t \end{cases}, \quad i \in G_r,\ j \in G_t \tag{3}$$

To emulate A and B compartments in a single cell Hi-C network, we consider a simple adjacency benchmark of a polymer with two communities. Namely, we represent the chain as a sequence of alternating segments of A and B type (painted in red and blue), whose lengths are Poisson-distributed with the mean length λ. An example of the resulting adjacency matrix is depicted in Fig. 1a. Note that due to decay of the contact probability, the "checkerboard" compartmentalization pattern is hardly seen in single cells Hi-C data[28]. Since segments of the same type are surrounded in space by segments of the other type, they form local "blob-like" clusters along the main diagonal of the adjacency matrix reminiscent to topologically-associated domains[15]. However, they are likely formed by a different mechanism and have an order of magnitude larger size than TADs[40]. Such a multi-domain blob structure in Fig. 1a is a manifestation of the polymeric nature of the network and it cannot be reproduced with communities of general memory-less networks, i.e. in the framework of the canonical stochastic block model with two clusters—see Fig. 1b for comparison.

Statistical inference of polymer SBM and generalized modularity functional. Suppose that a population-averaged matrix $\mathcal{A}$ is observed. By definition, each entry $\mathcal{A}_{ij}$ of this matrix counts the amount of reads between the bins i and j coming from a population of single cells. Thus, after proper normalization, A_{ij} is a Poisson variable with the mean dictated by (3), $\langle A_{ij}\rangle = P_{ij}\ \omega_{g_ig_j}$, and $\omega_{g_ig_j} = \Omega_{ij}$ are the pairwise group probabilities (at the moment we do not require all the groups to be identical). Neglecting correlations between the matrix entries, the statistical weight of $\mathcal{A}$ conditioned on the cluster probability matrix Ω, background contact probability P and group labels of the nodes $\{g_i\}$, can be factorized into the product of the Poisson probabilities for the entries A_{ij}

$$Z(\mathcal{A}|\ \Omega, P, \{g_i\}) = \prod_{i<j} \frac{\left(P_{ij}\ \omega_{g_ig_j}\right)^{\mathcal{A}_{ij}}}{\mathcal{A}_{ij}!} \exp\left(-P_{ij}\ \omega_{g_ig_j}\right) \tag{4}$$

where the product runs over all pairs of nodes in the network. Since there are no self-edges in the network, all the diagonal elements of the matrix $\mathcal{A}_{ij}$ are zeros and we do not include them into the product (4). The corresponding partitioning entropy of the polymer SBM is

$$\log Z(\mathcal{A}|\ \Omega, P, \{g_i\}) = \sum_{i<j} \left(\mathcal{A}_{ij} \log \omega_{g_ig_j} - P_{ij}\ \omega_{g_ig_j}\right) \tag{5}$$

where we have omitted the constant terms $-\log \mathcal{A}_{ij}!$ and $\mathcal{A}_{ij} \log P_{ij}$, independent of the partitioning. For identical communities (see (2)), we get

$$\begin{cases} \omega_{g_ig_j} = w_{out} + \delta_{g_ig_j}(w_{in} - w_{out}) \\ \log \omega_{g_ig_j} = \log w_{out} + \delta_{g_ig_j}\left(\log w_{in} - \log w_{out}\right) \end{cases} \tag{6}$$

Taking (6) into the account and omitting again all irrelevant constant terms, we arrive at the final expression for the entropy (5)

$$T \log Z(\mathcal{A}|\ \Omega, P, \{g_i\}) = \sum_{i<j} \left(\mathcal{A}_{ij} - \gamma P_{ij}\right)\delta_{g_ig_j} \tag{7}$$

where $T = \left(\log w_{in} - \log w_{out}\right)^{-1}$ is the effective temperature and

$$\gamma = \frac{w_{in} - w_{out}}{\log w_{in} - \log w_{out}} \tag{8}$$

is a parameter describing the cluster probabilities inherited from the initial definition of stochastic blocks.

The entropic functional (7), up to normalization coefficients and constant terms, is the generalized modularity functional. For $P_{ij} = d_i d_j / \sum_i d_i$, where d is the vector of degrees, (7) reduces to the modularity proposed by Newman[3,41] for the sake of spectral community detection in scale-free networks. Recently it has been shown that the same functional can be used to partition a network with the core-periphery organization[23]. The operator of the generalized modularity reads

$$\mathbf{Q} = A - \gamma P \tag{9}$$

The second term in (9) can be understood as an expectation number of contacts between nodes (i, j) in the population-averaged data, or as a probability of the link in the single cell graph. Indeed, in absence of the stochastic blocks, this value equals P_{ij} by definition. The factor γ responds for the clustering structure superimposed over the background. In the limit of "weak" communities, when $w_{in} = w_{out} \to 1$, the partitioning yields $\gamma \to 1$, which corresponds to the pure background. To determine the optimal value of γ, one can run a recursive procedure, which consists of iterative maximization of the generalized modularity and renormalization of γ according to (8). We realize this approach in our numerical analyses below.

Polymer non-backtracking flow operator

Non-backtracking walks on a directed polymer network. Search for the global maximum to the modularity functional is a very hard computational problem. One of most promising approaches which avoids a brute force, is to suggest that if the community structure is significantly strong, there is an operator whose eigenvectors encode the network partitioning in these communities[3,22]. However, as it was first noted by Krzakala et al.[30], for sparse networks leading eigenvectors become uncorrelated with true community structure well above the theoretical threshold. As a result, all conventional operators such as adjacency, Laplacian and modularity fail to find communities in rather sparse networks.

To overcome this difficulty, it was proposed to exploit the spectrum of the Hashimoto matrix $\mathbf{B}$, which is a transfer matrix of non-backtracking walks on a graph[35]. It is defined on the edges of the directed graph, $i \to j, k \to l$, as follows

$$\mathbf{B}_{i\to j,k\to l} = \delta_{il}(1 - \delta_{jk}) \tag{10}$$

It is seen from (10) that the non-backtracking operator prohibits returns to the point which a walker has visited at the previous step. Since matrix $\mathbf{B}$ is non-symmetric, its spectrum is complex. For Poissonian graphs the spectral density of $\mathbf{B}$ is constrained within a circle of radius $\sqrt{\langle d\rangle}$ in the complex plain and exhibits no "Lifshits

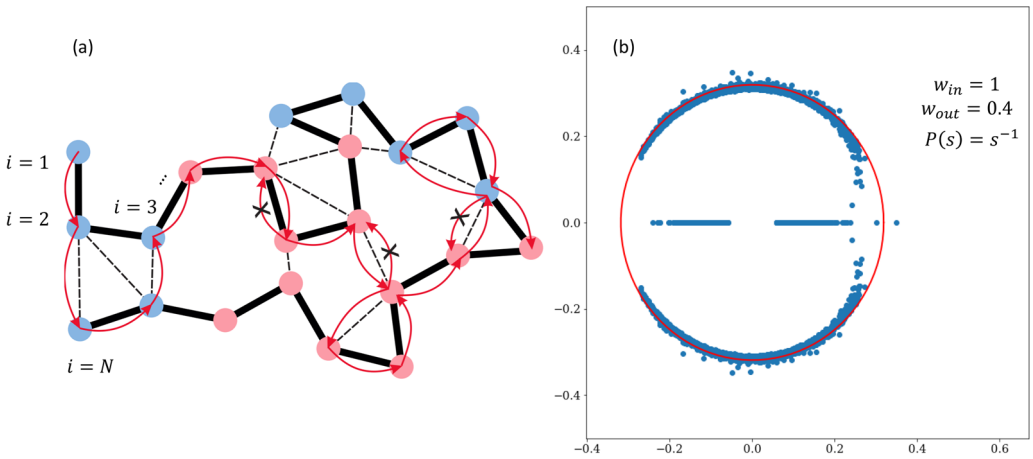

Figure 2. (**a**) Depiction of the polymer SBM network: the backbone (bold), contacts between genomically distant monomers (dashed) and two chemical sorts of the monomers (red and blue), arranged into contiguous alternating segments. An example of the non-backtracking walk on such graph is shown by arrows. Immediate returns are forbidden, preventing localization on hubs; (**b**) Spectrum of the polymer non-backtracking flow (11) for the fractal globular ($P(s) = s^{-1}$) large-scale organization of the chain with two overlaid compartments with the mean length $\lambda = 100$.

tail" singularities near the spectral edge, in contrast to other conventional operators[30,31]. Real eigenvalues of **B** lying out of the circle become relevant to the community structure even in sparse networks. Associating the corresponding eigenvectors with the network partitioning permits to detect communities all the way down to the theoretical limit. In[19] M. Newman suggested a normalized operator, that conserves the probability flow at each step of the walker.

For the sake of community detection in sparse polymer graphs, we propose a conceptually similar operator that describes the evolution of the non-backtracking probability flow on a graph with intrinsic linear memory

$$\mathbf{R}_{i\to j,k\to l} = \frac{\delta_{il}\left(1-\delta_{jk}\right)}{d_i - 1} - \gamma\left(d_j d_l\right)^{-1} P_{jl} \tag{11}$$

In "Appendix" we establish the connection between the non-backtracking operator and the generalized modularity, derived in the previous Section from the statistical inference of the polymer SBM. Thus, partitioning of a polymer network into two communities according to the leading eigenvector of the polymer non-backtracking flow operator (11) responds to the maximum entropy principle.

An example of the non-backtracking walk on a polymer graph is illustrated in the Fig. 2a. Note that despite immediate revisiting of nodes is forbidden, the walker is allowed to make loops. The second term in (11) plays a role of neutralization towards the contact probability, arising from the linear organization of the network. This compensation provides a measure for the non-backtracking operator to tell apart the true communities from the fluctuations, evoked by the polymeric scaling. Trivially, the proposed non-backtracking operator is converged to the Newman's flow operator, when the background is non-polymeric, but rather corresponds to the configuration model with fixed degrees $P_{ij} = d_i d_j / 2m$[19]. For a pure polymeric graph without contamination by communities, the spectrum of (11) lies inside a circle of radius $r = \sqrt{\langle d(d-1)^{-1}\rangle}$. As sufficiently resolved communities are formed in the network, isolated eigenvalues pop up at the real axis.

In Fig. 2b we depict the non-backtracking spectrum of a polymer SBM, corresponding to the fractal globule polymer network with $P(s) = s^{-1}$ of the size $N = 1000$ with two compartments, organized as contiguous alternating segments with the mean length $\lambda = 100$. For the parameters w_{in}, w_{out} used, the two compartments are well resolved that is provided by the isolated eigenvalue separated from the circle. Since the leading eigenvector $u^{(1)}$ of the polymer non-backtracking flow, in contrast to the adjacency or modularity, is defined on directed edges of the network, one needs to evaluate the Potts spin variables $g_i = \pm 1$ in order to classify the nodes. From the correspondence between the modularity and polymer flow operator one sees that contribution to the i-th node g_i comes from the flow along all the directed edges pointing to i. Thus, in order to switch from edges to nodes, one needs to evaluate the sign of the sum $v_i = \sum_j A_{ij} u^{(1)}_{j\to i}$ and to assign the node i accordingly, $g_i = sign(v_i)$.

Spectral clustering of the polymer stochastic block model. In this section we investigate spectral properties of the polymer non-backtracking flow and compare performance of various linear operators in partition the polymer SBM. The two compartments with $\lambda = 100$ are superimposed over the fractal globule, $P(s) = s^{-1}$, with total size of the network, $N = 1000$. We fix the weight of internal edges at $w_{in} = 1$ and change the resolution of compartments by tuning the weight of external edges, $w_{out} = 0.1 - 0.8$. Efficiency of splitting is assessed by the fraction of correctly classified nodes.

In Fig. 3a we compare the performance of adjacency, normalized Laplacian, M. Newman's non-backtracking flow operator, polymer modularity and polymer non-backtracking flow matrices. For the latter two, the optimal

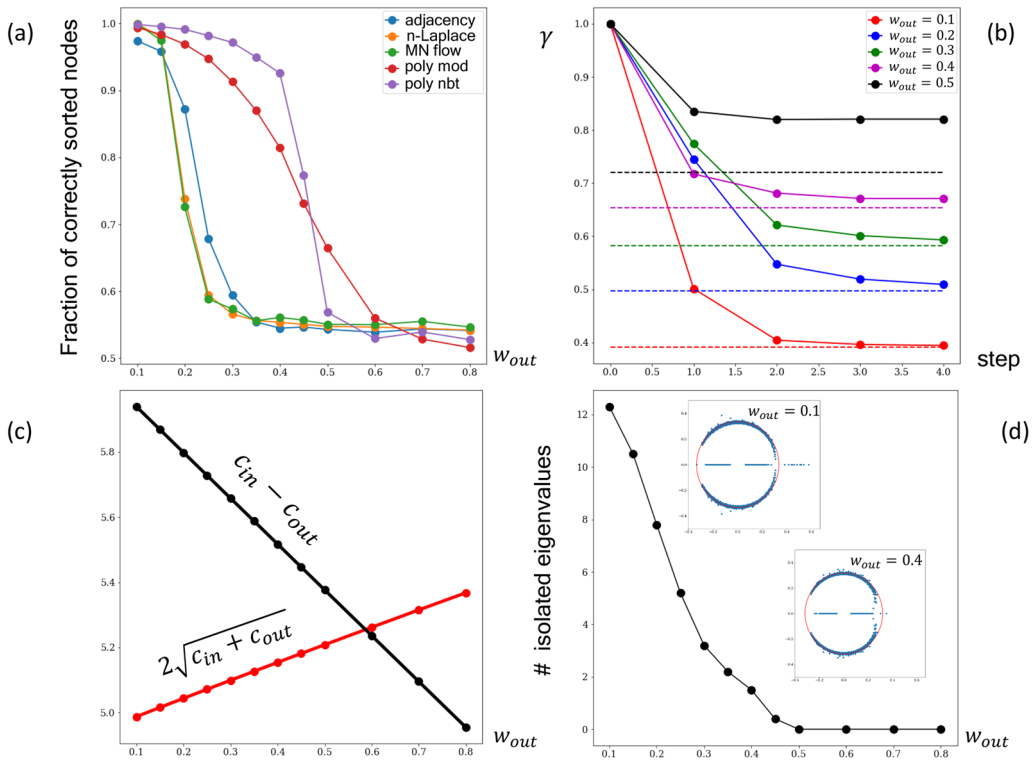

Figure 3. (**a**) Comparison of performance of different classical operators without background, polymer modularity and polymer non-backtracking flow operators ($N = 1000, P(s) = s^{-1}, w_{in} = 1, \lambda = 100$); (**b**) The iterative approach that can be used to determine the optimal value of γ for five values of w_{out}; the true optimal values of γ calculated from (8) are shown by dash; (**c**) The mean numbers of inner c_{in} and outer c_{out} edges are calculated for each value of w_{out} in order to estimate the detectability threshold for the corresponding regular network. (**d**) Amount of isolated eigenvalues of the polymer flow operator plotted against w_{out}. Full spectra of the polymer flow operator for the two values of w_{out} are shown in the insets.

value (8) of the parameter γ was chosen. It is evident that the polymer flow operator surpasses all conventional operators without the background, as well as the polymer modularity everywhere below $w_{out} \approx 0.5$. Qualitatively similar behaviour was demonstrated by the traditional non-backtracking operator without the background, when it was compared to other operators in[30]. Therefore, our analyses (i) underscores the importance of taking into account the contact probability (polymer background) when dealing with polymer graphs, and (ii) recapitulates efficiency of non-backtracking walks in resolving communities in sparse networks.

It is worth noting that the abrupt fall in performance of the polymer flow operator coincides with the leveling of its amount of isolated eigenvalues at zero, see Fig. 3d. Values around $w_{out} \approx 0.5$ define the detectability transition, above which the leading eigenvector becomes uncorrelated with the true nodes assignment. To understand whether it corresponds to the theoretical detectability limit, we translate w_{out} into the average amount of inner, $c_{in} = Nw_{in}/2$, and outer, $c_{out} = Nw_{out}/2$, edges and plot them as functions of w_{out}. As it is shown in Fig. 3c, the polymer flow operator drops close to the theoretical detectability transition for regular stochastic block models[42] (i.e. each node has *exactly* c_{in} random links with other nodes in its community and *exactly* c_{out} randomly pointed links to nodes from the other community)

$$c_{in} - c_{out} > 2\sqrt{c_{in} + c_{out}} \tag{12}$$

For the stochastic block model the number of isolated eigenvalues of **B** exceeds the number of communities by one[30]. However, in case of the polymer operator **R** the number of isolated eigenvalues can be much larger and "apparent" clusters might be formed "locally" at the main diagonal due to the frozen linear connectivity, see Fig. 1a. This is evident from the Fig. 3d, which shows that the number of isolated eigenvalues for the polymer flow operator can be of order of the amount of the segments (N/λ), if w_{out} is sufficiently low. Indeed, for the fractal globule probability of the edge between two distant segments of the same type is s times smaller than probability of the link for two close monomers ($s = |k - m|$ is the genomic distance between segments k and m). Due to the overall small number of contacts in the network, the polymer non-backtracking flow ends up rationalizing them as separate clusters.

The value of γ cannot be chosen arbitrary since it characterizes optimal parameters of stochastic blocks. Thus, one may propose the following iterative approach:

1. begin with the initial value $\gamma_0 = 1$, for which we obtain the network partition;
2. use the amount of inner and outer edges for estimating w_{in}, w_{out};
3. recalculate γ_1 according to (8);
4. repeat the procedure iteratively until γ converges to γ_{opt}.

Results of this procedure are demonstrated in the Fig. 3b for five different values of w_{out}. It is seen that just several steps of iteration is sufficient to obtain a reasonable convergence towards the theoretical values provided by (8). A drawback of this iterative procedure is that at each step one needs to evaluate the spectrum of the operator $2m \times 2m$, which could become a hard computational task for large and dense networks. As a reasonable approximation to the optimal value of γ for the polymer flow operator, one can evaluate γ_{opt} similarly for the polymer modularity, which is smaller in size and symmetric.

Polymer non-backtracking flow resolves compartments in a single cell Hi-C network. To check robustness of the polymer non-backtracking flow operator on real Hi-C data we run it on a set of individual oocyte cells of mouse[28]. From the public repository we have taken the single cells Hi-C data on cis-contacts of 20 chromosomes from 13 single cells (260 adjacency matrices, in total). While single cells matrices with sufficiently large number of contacts are not sparse and can be split into compartments using conventional methods largely used for the bulk data (e.g., the leading eigenvectors of observed/expected transformation of a population-averaged Hi-C map,[12]), here we take the cells with low to moderate amount of contacts for the sake of comparative analyses of clustering performance of different spectral methods on *sparse* polymer graphs.

Before proceeding with the analyses of compartments in single cells, the raw data must be preliminary processed. In order to extract compartmentalization signal from the maps, we have coarse-grained them to the resolution 200 kb. At this resolution all finer genome folding structuring (like topologically-associated domains) is encoded within the coarse-grained blobs and does not communicate with two large-scale *A* and *B* compartments. We note, that, in principle, the method is applicable at higher resolution as well. However, there are two important considerations. The non-backtracking operator is defined on the edges, therefore, the leading eigenvectors need to be computed for much larger matrix than in case of traditional operators, which are defined on the nodes (e.g., modularity). This means that the computation time of the method is very sensitive to the resolution. Furthermore, one needs to be very careful with the overall network density: it decreases by several times upon decreasing of the bin size, so that one can occasionally cross the detectability limit (12). In each particular case the resolution for the annotation should be chosen with respect to the sparsity of the experimental single cell contact maps. According to this logic, we have decided to use the resolution 200 kb for the data of Flyamer et al.

Most of the contacts in the cells have degeneracy 1 at the chosen resolution, however, several pairs of bins have more than 1 contact. To preserve this feature of enhanced connectivity, we consider the counts of contacts between the pairs as weights of the corresponding edges. Furthermore, the single-cell maps are noisy and some of really existing contacts get lost due to technical shortcomings of the experimental protocol. As long as the neighboring blobs in the chromatin chain are connected with probability 1, all lost contacts $A_{i,i+1}$ need be added to the adjacency matrix manually; we assign the weight 1 to such edges. We also cleans the coarse-grained data from the self-edges, assigning $A_{ii} = 0$.

To determine the background model for our analyses we calculate the contact probability $P(s) = \frac{1}{N-s}\sum_{i=1}^{N-s} A_{i,i+s}$ for each individual single cell and for the merged cell (summing single cells matrices), see Fig. 4a. Resulting dependence turns out to be fairly close to the fractal globule contact probability, $P(s) \sim s^{-\alpha}$ with $\alpha \approx 1$ at scales from $\approx$ 1-2Mb to the end of the chromosome. A shoulder at lower scales around 1 Mb reflects enhancement of the contact probability due to the compartmentalization. Importantly, the fractal globule scaling at the megabase scale is universal across different species and cell types; it is evident in the population-averaged contact matrices in mouse oocytes[28], human lymphoblastoid cells[12] and *Drosophila* cells[43]. As it was shown in previous Section, in order to extract compartmentalization profile overlaying a specific long-ranged folding, it is crucial to incorporate the respective background contact probability into the polymer model of the stochastic blocks.

Having the background model determined, we construct the polymer non-backtracking flow operator with the variable parameter γ and run the iterative clustering procedure to derive the optimal value γ_0. Similarly to the analyses on the benchmarks, see Fig. 3b, a swift convergence to the optimal value is observed here. The spectrum of the polymer flow operator for the cell 29749, chromosome 3 at $\gamma_0 \approx 0.9$ is shown in the inset of the Fig. 4b. Nineteen isolated eigenvalues on the real axis are separated from the bulk spectrum. As we have shown in the previous Section, this is a quite typical scenario for sparse polymer stochastic block models. In the sparse limit of the polymer SBM, the number of isolated eigenvalues could be much larger than the number of compartments.

The partition of the single cells in two compartments has been performed in the leading eigenvector approximation of the different operators. The boundaries of active and inactive domains are determined according to the sign of the respective compartmental signal (see Fig. 4b and Supplementary Fig. S1 online). It is known that the gene density is higher in the actively transcribed *A* compartment, thus, the fraction of GC letters in bins of active compartmental domains needs to be larger than in inactive domains. To validate that the clusters found in single cells respond to the transcriptional domains and are biologically significant, we calculate the GC content profiles around the centers of all *A* and *B* domains separately and then take the average of these profiles in each group. The types of the domains were phased in accordance with the leading eigenvector of the bulk data (population Hi-C on embryonic stem cells was used[44]; the eigenvector was computed on the observed-over-expected map). We also plot analogous profiles for the leading eigenvector of the bulk data. In absence of direct annotation methods for single cells due to their sparsity, these two measures have been of use to approximate positions of the compartmental domains in single cell Hi-C data[28].

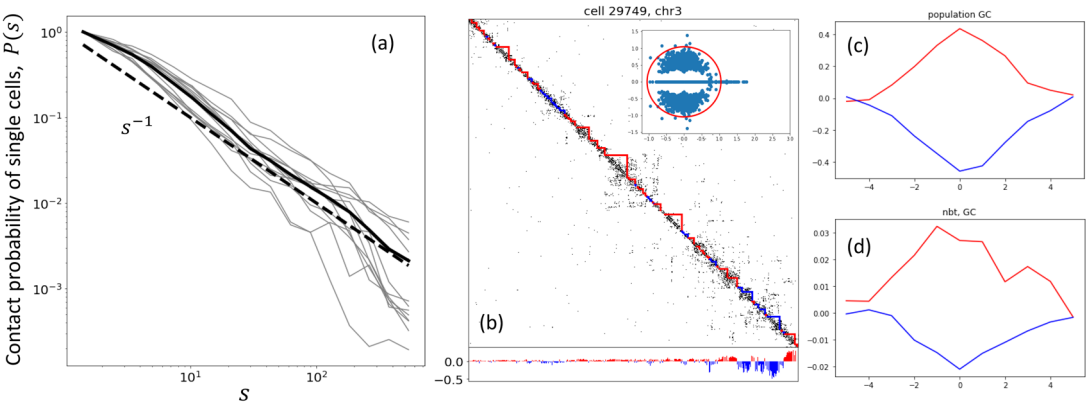

Figure 4. (**a** The average contact probability $P(s)$ of single cells (gray) and of the merged cell (solid, black) computed for logarithmically spaced bins with the logfactor 1.4; the fractal globule scaling $P(s) \sim s^{-1}$ is also shown by dashed line for comparison. (**b**) Annotation of active (red) and inactive (blue) compartmental domains for one of the contact maps (cell 29749, chromosome 3, length $N = 492$, 200kb resolution) by the polymer non-backtracking flow operator. Below the map the compartmental signal from the corresponding leading eigenvector of the polymer non-backtracking flow matrix is shown. Inset: the full spectrum of the polymer flow for the same contact map. (**c**,**d**) Averaged profiles of the GC content (z-scores) plotted around the centers of the compartmental domains (active—red, inactive—blue) for the population of cells and for a pool of single cells.

As expected, the GC content for the population-averaged map and the bulk E1 vector both have pronounced peaks at the center of A domains and symmetrical dips at the center of B domains with the z-score amplitude equal to 0.4 (GC) and 0.7 (E1), correspondingly. Single cells profiles demonstrate notably lower amplitudes (see Fig. 4c,d and Supplementary Figs. S2, S3 online). However, only the polymer non-backtracking flow yields the annotation with the similar shape and span. Both profiles (for A and for B) of the polymer non-backtracking flow fall symmetrically to zero at the same genomic distance, around $4-5$ bins from the centers of domains, which also strikingly coincides with the span of the bulk profiles. This is also compliment to the similarity of the characteristic sizes of compartmental domains determined by the non-backtracking flow operator ($\langle l \rangle \approx$ 2.2 Mb) and domains from the bulk data ($\langle l \rangle \approx 1.7$ Mb). To test the effect of different α, we additionally run the polymer non-backtracking for $\alpha = 3/2$, which is the scaling exponent of the contact probability for the ideal chain packing. Comparison of the two values of the parameter is demonstrated in Supplementary Fig. S4 online: the profiles with $\alpha = 3/2$ show significantly worse correlation with both GC content and the E1 bulk vector. This is consistent with the slope $\alpha \approx 1$ of $P(s)$ for the set of single cells, Fig. 4a, underscoring the importance of neutralization on the appropriate average polymeric scaling before the clustering.

Note that the partitions of the polymeric operators (non-backtracking, modularity) are visibly much more adequate to apparent clustering of contacts in a particular cell (Supplementary Fig. S1 online). Despite the similarity in compartmental signals from the polymer modularity and from the polymer non-backtracking flow, the sizes of modularity domains are almost twice larger ($\langle l \rangle \approx 4.1$ Mb) and show negative z-scores of GC content both for the active and inactive compartments. The profile of the E1 vector plotted for the polymer modularity has a similar bell shape, however, it levels at ≈ -0.07 and stays negative throughout the whole range of the compartmental interval. This is a consequence of sparsity, which results in a limited performance of all traditional spectral methods.

Conclusion

In this paper we have developed theoretical grounds for spectral community detection in sparse polymer networks. On the basis of suggested polymeric extension of the stochastic block model, we have proposed the polymer non-backtracking flow operator and have proven that its leading eigenvector performs partitioning of a polymeric network into two clusters according the maximum entropy principle. The established connection with the modularity functional provides a computationally efficient tool for the network partitioning and search for the optimal resolution parameter of the partition in polymer networks, which, however, is inferior to the non-backtracking in efficiency for sparse networks.

The proposed theoretical framework is verified by extensive numerical simulations of polymer benchmarks, constructed in order to emulate compartmentalization in sparse chromatin networks. Comparative analyses of different operators on the benchmark has suggested that the polymer flow detects the communities up to the theoretical detectability limit, while all other operators fail above it. At the same time, the amount of isolated eigenvalues of the polymer flow operator can be larger than amount of true communities present in the network, due to frozen linear connectivity that forces the chain to form "blobs" along the chain contour. This result distinguishes the polymer system with thespect the canonical stochastic block model, where the number of isolated eigenvalues of the non-backtracking exactly matches the number of communities.

Analyses of the single cell Hi-C data of mouse oocytes suggests that the non-backtracking walks efficiently split experimental sparse networks into biologically significant communities, characterized by enrichment and

depression of the genes density. The sizes of the compartmental domains are fairly close to the sizes of the population-averaged domains. Comparison with characteristics of the domains, inferred by other operators, underscores superiority of the non-backtracking walks in partitioning sparse polymer networks.

In this study we have exploited for the polymer network analysis only the simplest spectral characteristics. More involved ones, e.g. spectral correlators and the level spacing distribution, carry additional information about the propagation of excitations in network. The spectral statistics and non-ergodicity have been discussed in clustered networks in[45,46]. In the context of the gene interactions the spectral statistics has been discussed in[47] for the matrices with the real spectrum. The non-backtracking matrices enjoy complex spectrum hence the special means are required to analyze the level spacing in this case. The corresponding tool has been invented recently[48,49], therefore, the spectral statistics of the polymer non-backtracking flow operator certainly deserves a separate study.

Appendix

Methods. *Quadratic form of the polymer non-backtracking operator.* Let us consider a quadratic form involving the operator over the Potts spin variables $g_i, i = 1, 2, \ldots, N$ and introduce the $2m$-dimensional ($2m$ is the number of edges in the network) vector u, such as $u_{i\to j} = g_j$. Then,

$$R = u^T \mathbf{R}\, u = \sum_{\substack{(i,j)\in E\\(k,l)\in E}} \mathbf{R}_{i\to j,k\to l}\, g_j g_l \tag{13}$$

It can be shown that (13) coincides with the quadratic form of the generalized modularity. Let us consider the terms separately. The quadratic form of the first, non-backtracking term, yields

$$\sum_{\substack{i\to j\\k\to l}} \frac{\delta_{il}\left(1-\delta_{jk}\right)}{d_i-1}\, g_j g_l = \sum_{i,j} \frac{g_i g_j}{d_i-1} A_{ij} \sum_k \left(1-\delta_{0A_{ik}}\right)\left(1-\delta_{jk}\right) = \sum_{ij} A_{ij}\, g_i g_j \tag{14}$$

where the sum over k enumerates the edges of the node i except of the edge (i, j) and, thus, equals $d_i - 1$. Expanding the quadratic form of the second term similarly, we get

$$\gamma \sum_{\substack{i\to j\\k\to l}} \left(d_j d_l\right)^{-1} P_{jl}\, g_j\, g_l = \gamma \sum_{jl} \left(d_j d_l\right)^{-1} P_{jl}\, g_j\, g_l \sum_{ik} \left(1-\delta_{0A_{ij}}\right)\left(1-\delta_{0A_{kl}}\right) = \gamma \sum_{jl} P_{jl}\, g_j\, g_l \tag{15}$$

Collecting (14) and (15) together one arrives at

$$R = u^T R\, u = \sum_{ij} \left(A_{ij} - \gamma P_{ij}\right) g_i\, g_j; \quad P_{ij} = \frac{1}{|i-j|^\alpha} \tag{16}$$

which is the quadratic form of the generalized modularity functional, proportional to the entropy of the polymer SBM.

Received: 1 April 2020; Accepted: 19 June 2020
Published online: 09 July 2020

References

1. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
2. Girvan, M. & Newman, M. Community structure in social and biological networks.*Proc. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
3. Newman, M. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
4. Lusseau, D. & Newman, M. Identifying the role that animals play in their social networks. *Proc. R. Soc. Lond. B Biol.***271**, S477–S481 (2004).
5. Harris, K. D. *et al.* Organization of cell assemblies in the hippocampus. *Nature* **424**, 552–556 (2003).
6. Humphries, M. Spike-train communities: Finding groups of similar spike trains. *J. Neurosci.* **31**, 2321–2336 (2011).
7. Jeong, H. *et al.* The large-scale organization of metabolic networks. *Nature* **407** (6804), 651–654 (2000).
8. Ravasz, E. *et al.* Hierarchical organization of modularity in metabolic networks. *Science* **297** (5586), 1551–1555 (2002).
9. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**(1), 17 (2005).
10. Doyle, B. *et al.* Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Computat. Biol.* **10**(10), e1003867 (2014).
11. Dekker, J. *et al.* Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
12. Lieberman-Aiden E., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
13. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502** (7469), 59–64 (2013).
14. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.***16**(1), 180 (2015).
15. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
16. Sexton, T., et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
17. Szabo, Q. *et al.* Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**(4), eaaw1668 (2019).

18. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**(3–5), 75–174 (2010).
19. Newman, M. E. J. Spectral methods for community detection and graph partitioning. *Phys. Rev. E* **88**(4), 042822 (2013).
20. Newman, M. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113 (2004).
21. Shen, H.-W. & Cheng, X. Spectral methods for the detection of network community structure: a comparative analysis. *J. Stat. Mech. Theory Exp.* **2010**(10), P10020 (2010).
22. Decelle, A. *et al.* Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**(6), 066106 (2011).
23. Polovnikov, K., Kazakov, V. & Syntulsky, S. Core-periphery organization of the cryptocurrency market inferred by the modularity operator. *Physica A Stat. Mech. Appl.* **540**, 123075 (2020).
24. Erdos, P. & Renyi, R. On pseudoprimes and Carmichael numbers. *Publ. Math. Debrecen* **4**, 201–206 (1956).
25. Ravasz, E. & Barabasi, A.-L. Hierarchical organization in complex networks. *Phys. Rev. E* **67**, 026112 (2003).
26. Lee, S. H. *et al.* Mapping the spectrum of 3D communities in human chromosome conformation capture data. *Sci. Rep.***9**(1), 1–7 (2019).
27. Grosberg, A. Yu. & Khokhlov, A. R. *Statistical Physics of Macromolecules* (American Institute of Physics, New York, 1994).
28. Flyamer, I. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544** (7648), 110–114 (2017).
29. Zhang, P. *et al.* Comparative study for inference of hidden classes in stochastic block models. *J. Stat. Mech.* **12**, P12021 (2012).
30. Krzakala, F. *et al.* Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci.* **110** (52), 20935–20940 (2013).
31. Nechaev, S. K. & Polovnikov, K. Rare-event statistics and modular invariance. *Physics-Uspekhi* **61** (1), 99 (2018).
32. Lifshitz, I. M. Theory of fluctuation levels in disordered systems. *Sov. Phys. JETP* **26**, 462 (1968).
33. Goh, K.-I. et al. Spectra and eigenvectors of scale-free networks. *Phys. Rev. E* **64**, 051903 (2001).
34. Nadakuditi, R. R. & Newman, M. E. J. Spectra of random graphs with arbitrary expected degrees. *Phys. Rev. E* **87**, 012803 (2013).
35. Hashimoto, K. Zeta functions of finite graphs and representations of p-adic groups. *Adv. Stud. Pure Math.* **15**, 211–280 (1989).
36. Rosa, A. & Everaers, R. Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.* **4**(8), e1000153 (2008).
37. Polovnikov, K., Nechaev, S., & Tamm, M. Effective Hamiltonian of topologically stabilized polymer states. *Soft Matter* **14**, 6561–6570 (2018).
38. Grosberg, A. Yu. *et al.* Crumpled globule model of the three-dimensional structure of DNA. *EPL (Europhys. Lett.)* **23** (5), 373 (1993).
39. Grosberg, A. Yu. *et al.* The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de physique* **49** (12), 2095–2100 (1988).
40. Nuebler, J. *et al.* Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci.* **115** (29), E6697–E6706 (2018).
41. Newman, M. E. J. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E* **94** (5), 052315 (2016).
42. Radicchi, F. Detectability of communities in heterogeneous networks. *Phys. Rev. E* **88** (1), 010801 (2013).
43. Ulianov, S. V. *et al.* Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* **26** (1), 70–84 (2016).
44. Bonev, B. *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* **171** (3), 557–572 (2017).
45. Avetisov, V., Hovhannisyan, M., Gorsky, A., Nechaev, S., Tamm, M., & Valba, O. Eigenvalue tunneling and decay of quenched random network. *Phys. Rev. E* **94**, 062313 (2016).
46. Avetisov, V., Gorsky, A., Nechaev, S. & Valba, O. Localization and non-ergodicity in clustered random networks. *J. Complex Netw.* https://doi.org/10.1093/comnet/cnz026 (2018).
47. Kikkawa, A. Random matrix analysis for gene interaction networks in cancer cells. *Sci. Rep.* **8**, 10607 (2018).
48. Zhang, G. H. & Nelson, D. R. Eigenvalue repulsion and eigenfunction localization in sparse non-Hermitian random matrices. *Phys. Rev. E* **100**, 052315 (2019).
49. Lucas, S., Ribeiro, P. & Prosen, T. Complex spacing ratios: a signature of dissipative quantum chaos. *Phys. Rev. X* **10** (2), 021019 (2020).

Acknowledgements

We are grateful to Leonid Mirny, Mikhail Gelfand, Mikhail Tamm, Maxim Imakaev and Nezar Abdennur for valuable discussions on the subject of the paper. KP, AG and SN acknowledge supports of the Foundation for the Support of Theoretical Physics and Mathematics "BASIS" (respectively 17-1-2-27-8 for KP, 17-11-122-1 for AG and 19-1-1-48-1 for SN). This work was supported by Grants RFBR 18-29-13013 (KP, AG, SN) and RSF 19-14-00016 (SVR, SVU). The authors are grateful to CNRS for the organizational and financial support of the publication.

Author contributions

K.P. designed the study, performed the benchmark and single cell analyses and wrote the manuscript. A.G., S.N., S.V.R. and S.V.U. supervised the work and participated in writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68182-0.

Correspondence and requests for materials should be addressed to K.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Non-backtracking walks reveal compartments in sparse chromatin interaction networks. Supplementary information

K. Polovnikov[1,2*], A. Gorsky[5,6], S. Nechaev[3,4], S. V. Razin[7,8], S. Ulyanov[7,8]

[1] *Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2] *Skolkovo Institute of Science and Technology, 143026 Skolkovo, Russia*

[3] *Interdisciplinary Scientific Center Poncelet (ISCP), 119002, Moscow, Russia*

[4] *Lebedev Physical Institute RAS, 119991, Moscow, Russia*

[5] *Moscow Institute for Physics and Technology, Dolgoprudnyi, Russia*

[6] *Institute for Information Transmission Problems of RAS, Moscow, Russia*

[7] *Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia*

[8] *Faculty of Biology, M.V. Lomonosov Moscow State University, Moscow, Russia*

(Dated: May 20, 2020)

* To whom correspondence should be addressed. Email: kipolovnikov@gmail.com

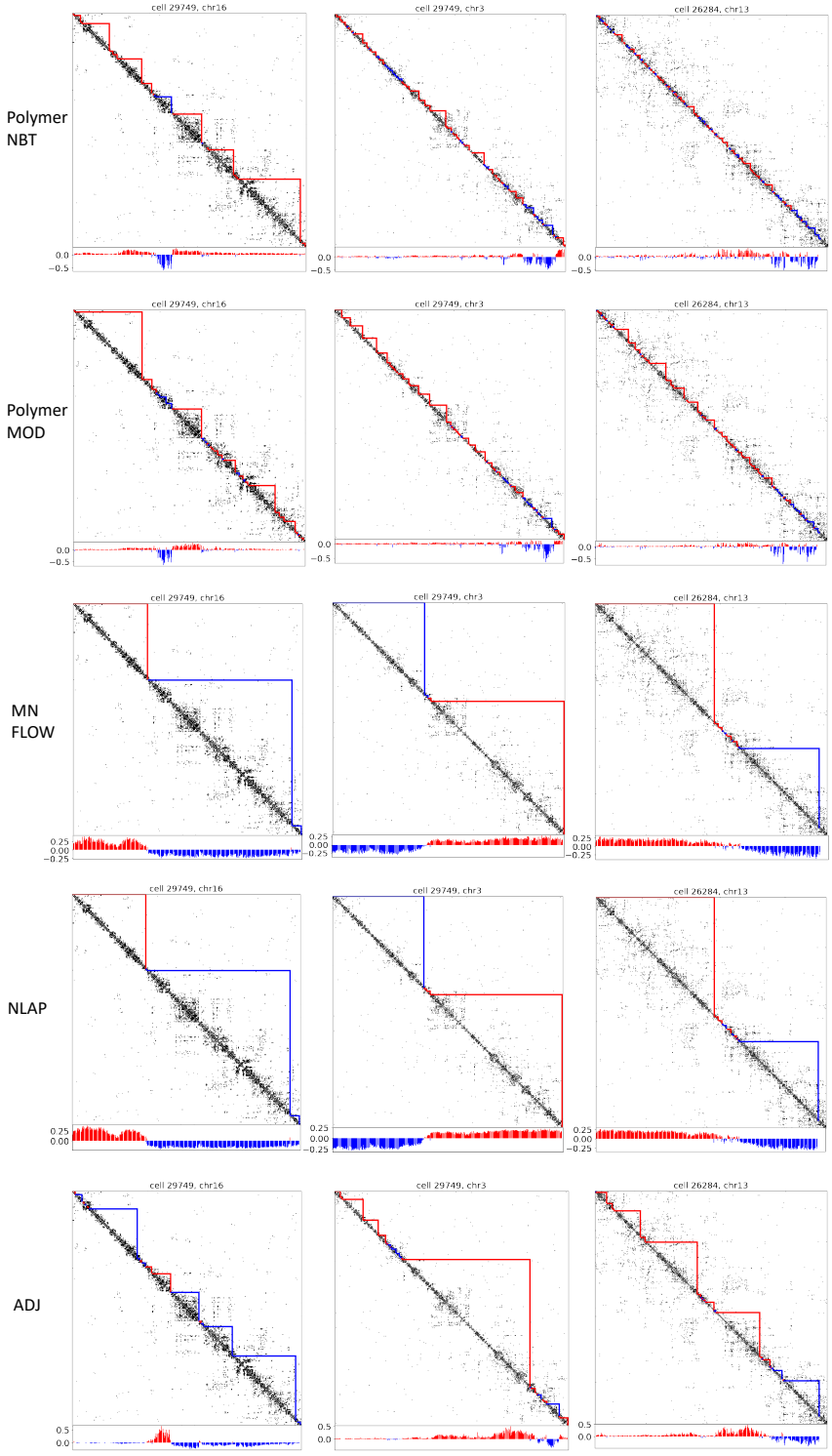

Figure S1: Annotations of active (red) and inactive (blue) compartmental domains for three chromosomes (16, 3 and 13; resolution 200kb) of the cell 29749 by the polymer non-backtracking flow operator, polymer modularity, M. Newman's non-backtracking flow, normalized Laplacian and adjacency. Below each map the compartmental signal from the leading eigenvector of the corresponding operator is provided. Hi-C data is taken from Flyamer et al.

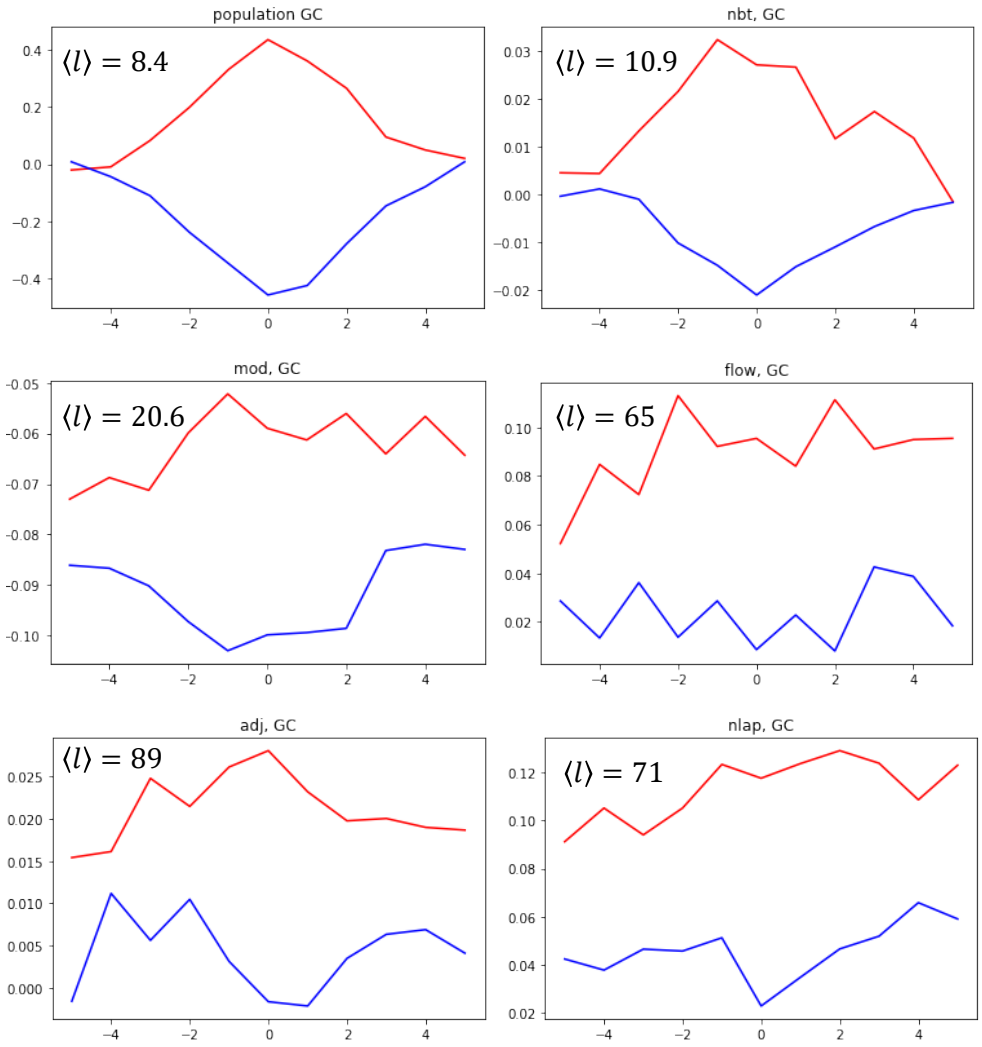

Figure S2: Averaged profiles of the GC content (z-scores) plotted around the centers of the compartmental domains (active - red, inactive - blue) for the population (embryonic stem cells, data taken from Bonev et al.), polymer non-backtracking flow operator, polymer modularity, M. Newman's non-backtracking flow, normalized Laplacian and adjacency. In case of single cells the average is taken over all compartmental domains of respective type from 260 contact maps. Mean sizes of the domains in bins (200 kb), inferred by each operator, are labeled on the plots.

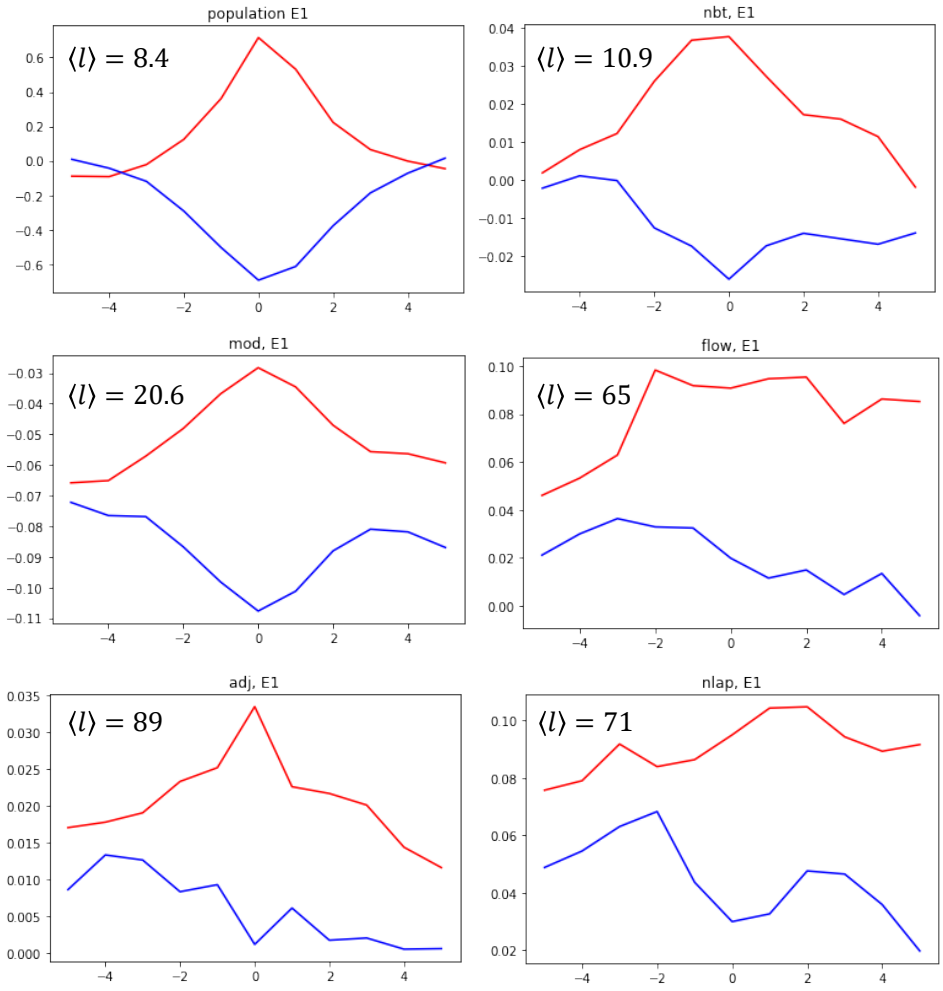

Figure S3: Averaged profiles of the leading eigenvector of the population-averaged Hi-C map (z-scores, ES cells, data taken from Bonev et al.) plotted around the centers of the compartmental domains (active - red, inactive - blue). The bulk matrices are preliminary normalized over expected and the eigenvector is phased with respect to the GC content, as usual. The profiles are demonstrated for the same population (ES cells) and for the domains determined in single cells by means of the polymer non-backtracking flow operator, polymer modularity, M. Newman's non-backtracking flow, normalized Laplacian and adjacency. In case of single cells the average is taken over all compartmental domains of respective type from 260 contact maps. Mean sizes of the domains in bins (200 kb), inferred by each operator, are labeled on the plots.

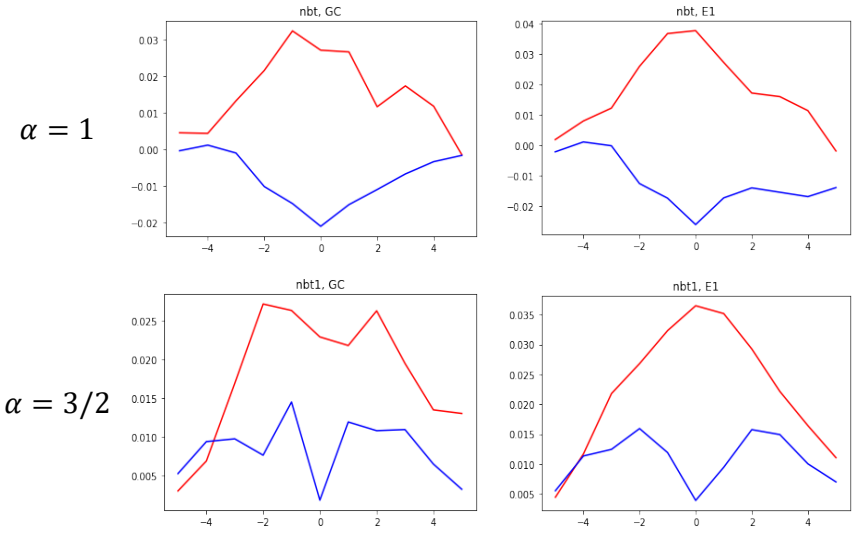

Figure S4: Comparison of the compartmental domains, inferred by the polymer non-backtracking flow for two values of $\alpha = 1$ (fractal globule) and $\alpha = 3/2$ (ideal chain). The profiles for the GC content and for the bulk leading eigenvector E1 are demonstrated.

Conclusion

In this thesis I have demonstrated a fundamental connection between the average spectral density of an ensemble of sparse Erdös-Rényi graphs and action of $SL(2, Z)$ modular group in the hyperbolic space. The spectral densities of linear and regular Bethe tree subgraphs comprise of an hierarchy of peaks located at all rational points governed by number-theoretic relationships which can be analytically approximated by the modular form (Dedekind η-function) close to the real axis. Manifestation of the hyperbolic geometry in sparse graphs becomes particularly evident when one is looking for a C^1 immersion of the Poincare disk to the three-dimensional Euclidean space. In our everyday life these immersions appear as shapes of plants and leaves and feature surface undulations which are evoked by the incompatibility of the local growth protocol with the ambient Euclidean metric. One can say that naturally grown surfaces are often found buckled because their growth generates the abundant material that cannot be disposed properly on the Euclidean plane. Emerging undulations can be mapped onto the optimal path problem of the light propagating in the media with a certain refraction index. In the strong metric incompatibility regime (expected to realize in tumour growth) the self-similar buckling patterns emerge at the boundary of the growing material. We have demonstrated that this pattern can be described by the eikonal equation with the refraction index that is expressed through the Dedekind η-function. Our purely geometric arguments agree well with a number of energetic approaches to buckling of thin membranes, where the stiffness is controlled by the effective bending rigidity. However, the geometric approach allows to understand two important features of observed buckling patterns: (i) self-similarity, as a result of the ultrametricity inherited from the modular relations of the η-function and (ii) singularities, which are an effect of non-analyticity of the immersion of surfaces with constant negative curvature to $\mathcal{R}^3$ (Hilbert's theorem).

As an important feature of the spectral density of sparse graphs, we note the emergence of the one-dimensional Lifshitz tail at the edge of the spectrum.

This is a distinctive feature of the density of states in the sparse regime, which is shown to be related to the KPZ behaviour in the grand canonical ensemble. As an example, we have demonstrated that stretched two-dimensional random paths over a semi-circular boundary become effectively one-dimensional and exhibit the KPZ fluctuations with exponent $\gamma = 1/3$, contrary to $\gamma = 1/2$ for unconstrained random paths. The Laplace transform of the corresponding Gibbs measure (or survival probability in a curvilinear channel) produces the Lifshitz tail with $D = 1$ for stretched paths and with $D = 2$ for the unconstrained ones. This mathematical procedure physically corresponds to the grand canonical ensemble of stretched random walks evading circular boundaries of different sizes.

Spectral random matrix theory is a hot topic in contemporary big data analyses, since it provides effective tools for probing the topological structure of real-world random networks. To this aim, I have studied an emergent correlation-based network of cryptocurrencies and have demonstrated that though it has a non-traditional core-periphery organization, it still can be captured by the modularity operator, which, therefore, provides a universal means for studying networks with undefined topological motif. Theoretical advancements in the field of sparse random matrices are of even more practical importance, as long as most of the real networks are huge with the size, much exceeding the number of effective "connections"at each node. Community detection in sparse single cell Hi-C matrices is one of the most actual and data-inspired problem in contemporary biology. Our contribution to the field is the development of two conceptually novel algorithms, which identify communities in sparse Erdös-Rényi graphs using non-backtracking random walks. The respective operator is non-Hermitian and manages to resolve communities in a sparse network up to the detectability threshold. Annotation of contiguous communities in single cell networks allows to detect topologically-associated domains, which are one of the most universal structural units of the genome folding. The boundaries of the domains are shown to be filled with various epigenetic markers, arguing in favour of the biological relevance of the found communities. The other algorithm is elaborated to grasp another widely-known structural element of the genome folding, so-called compartments, associated with different activity of transcription. To this aim, we have proposed a modified non-backtracking operator that is neutralized to the polymer contact probability in order to take into account polymer connectivity of chromatin; we note that the modified polymer-NBT is applicable to large-scale analyses of arbitrary networks

with intrinsic linear memory or another known background. Our approach provides the first method in chromatin bioinformatics for revelation of compartments directly from sparse single cells Hi-C data.

Emergence of Lobachevsky geometry and hierarchical modularity in nature can often be understood as being a result of optimal dynamics on the underlying ultrametric landscape. Scale-free sparse graphs are known to be very deeply related with the hyperbolic spaces, for example, in the sense of the geometric random graphs in the Poincare disk. Thus, it is tempting to describe universalities inherent to many complex systems by hidden ultrametric relationships between the agents. We hope that the analytical results and practical tools collected under the roof of this Thesis would serve a basis for and argue in favor of these attempts.

Acknowledgements

I am grateful to my supervisors Sergei Nechaev and Mikhail Gelfand, discussion with whom is always pleasant and intellectually challenging. Though Skoltech is a way more flexible and bureaucratically friendly place to work in than other universities in Russia, I am deeply thankful to my supervisors, as well as to Grigory Kabatiansky and to Mikhail Skvortsov, who helped me to resolve occasionally occurring administrative issues. Since it is the second PhD thesis in my life (and I plan to take a break here), I thank my former supervisor at the department of physics of MSU, Mikhail Tamm, for valuable and motivating discussions on the topics falling into this thesis. I thank the organizers and lecturers of two brilliant physics schools, in Les Houches and in Bangalore (ICTP), where I learned several new approaches in statistical physics, random matrix theory and extreme values statistics, which accelerated production of new results by my own. Also I would like to emphasize the shared essence of the all the results reported in this thesis and the contribution of all my dear colleagues and co-authors is deeply appreciated. With the hope not to forget anyone, here is a list of all whom I have collaborated with in the last four years: N. Abdennur, A. Astakhov, V. Avetisov, S. Belan, H. Brandao, A. Chertovich, A. Galitsyna, A. Gavrilov, A. Grosberg, M. Gelfand, A. Gorsky, M. Imakaev, V. Kazakov, E. Khrameeva, P. Kos, M. Lenz, S. Majumdar, Y. Maximov, L. Mirny, B. Meerson, R. Metzler, S. Nechaev, A. Orlov, I. Potemkin, S. Razin, V. Scholari, S. Shlosman, S. Syntulsky, M. Tamm, L. Truskinovsky, S. Ulianov, A. Valov, Y. Vassetzky, A. Vladimirov, V. Zakharova.