I work in the field of probability theory, combinatorics, and theoretical data science and machine learning. My research is broadly driven toward rigorous understanding of self-organization in discrete spatial processes, which arise in different fields including neural networks, physics, computer science, and epidemiology. In the context of modern data science, I am interested in convergence of online/distributed algorithms, MCMC sampling, online learning algorithms and inference.

1. Overview

1.1. Markovian theory of network data and machine learning. In the past few decades, a large number of models, methods, and algorithmic frameworks have been developed for understanding large and complex datasets. However, as the size and complexity of data are rapidly increasing, understanding their explicit or hidden structures poses further challenges to existing methods. One of my main research programs is to develop systematic theory of network data and machine learning based on the Markov chain theory. My two main achievements in this program address the problems of 1) Sampling graphs from large networks and computing/testing stable network observables, and 2) convergence of Online Matrix Factorization for Markovian stream of data.

Conventional sampling algorithms in statistical network data analysis such as independent and neighborhood sampling have limitations to dense or sparse networks, respectively. In a joint work with Memoli and Sivakoff [LMS19], we propose two complementary MCMC algorithms for sampling motifs from networks and establish bounds on their mixing times and concentration of their time averages. Based on our sampling algorithms, we propose a novel framework for network data analysis that circumvents some of the drawbacks in conventional sampling methods and provides efficient ways to compute and test network observables that reveal hierarchical structures. When applied for analyzing Word Adjacency Networks of a 45 novels data set, our framework reveals distinct hierarchical patterns depending on the authors (see Figure 2).

Online Matrix Factorization (OMF) is a fundamental tool for dictionary learning problems, giving approximation representation of complex data sets in terms of a reduced number of extracted features. Convergence guarantees for most of the OMF algorithms in the literature assume independence between data matrices, and the case of a dependent data stream remains largely unexplored. In a joint work with Needell and Balzano [LNB19], I have shown that the well-known OMF algorithm for stream of independent and identically distributed data proposed in [MBPS10], in fact converges almost surely to the set of critical points of the expected loss function, even when the data matrices form a Markov chain satisfying a mild mixing condition. This enables learning features of the sample space directly from a MCMC trajectory (see Figure 4). Also, by combining OMF and the network sampling algorithm we proposed in [LMS19], we propose a novel framework of Network Dictionary Learning, which extracts ‘network dictionary patches’ from a given network in an online manner that encodes main features of the network.

Currently I am working on extending the aforementioned Markovian theory of network-matrix data into the hypernetwork-tensor setting. Implications of this project include MCMC algorithms for sampling hypergraphs from hypernetworks, computing and testing various hypernetwork observables, online tensor factorization on Markovian data, learning temporal change of topics or structures in time-varying matrix data, and so on.

1.2. Solitons, box-ball systems, and integrable probability. The box-ball systems (BBS) are integrable cellular automata whose long-time behavior is characterized by the soliton solutions, and have rich connections to other integrable systems such as the celebrated Korteweg-de Veris (KdV) equation. The original form of BBS first invented by Takahashi and Satsuma [TS90] as a discrete counterpart of KdV, and later the explicit limiting procedure from KdV to BBS was discovered [TTMS96]. BBS is known to arise both from the quantum and classical integrable systems by the procedures called crystallization and ultradiscretization, respectively. This double origin of the integrability of BBS lies behind its deep connections to quantum groups, crystal base theory, solvable lattice models, the Bethe ansatz, soliton equations, ultradiscretization of the Korteweg-de Vries equation, tropical geometry and so forth [FY00, HHH01, KOS06, IKT12].

BBS with random initial configuration is an emerging topic in the probability literature and has gained considerable attention with a number of recent works [LLP17, CKST18, KL18, FG18, KL18, CS19a, CS19b, LLPS19]. My research on randomized BBS leads a body of this literature, aiming to answer the following question: If the first \( n \) boxes are randomly occupied with balls, what is the limiting shape of the invariant Young diagram corresponding to the soliton decomposition, as the system size \( n \) tends to infinity? In a joint work with Levine and Pike [LLP17],
I have addressed this question for the basic 1-color BBS with i.i.d. initial configuration. In collaborations with Kuniba and Okado [KLO18, KL18], I have investigated the row lengths in the multicolor BBS and obtained Schur polynomials representations of their scaling limit as well as their large deviations principle. Furthermore, with Lewis, Pylyavskyy, and Sen [LLPS19], I have obtained scaling limits of the columns lengths in multicolor BBS using a modified version of Greene-Kleitman invariants for BBS and circular exclusion processes.

Topics for my current and future research on randomized BBS include two-sided scaling limit of the invariant Young diagrams, scaling limit of the higher order Young diagrams in multicolor BBS, generalization to the discrete KdV with random initial configuration, and identifying relation between Grothendieck polynomials, combinatorial $R$, and the BBS.

1.3. What does a random contingency table with non-uniform margins look like? Contingency tables are fundamental objects in statistics for studying dependence structure between two or more variables, see e.g. [Eve92, FLL17, Kat14]. They also correspond to bipartite multi-graphs with given degrees and play an important role in combinatorics and graph theory, see e.g. [Bar09, DG95, DS96].

Hypergeometric (or Fisher-Yates) and uniform distributions are two widely used null models for the set of contingency tables with given row and column margins, in order to test independence and structural hypotheses, respectively [G+63, DE85, MH+13]. Once we reject the null hypothesis of independence using Pearson’s chi-squared or Fisher’s exact tests, we may continue studying the structure of the contingency table by asking some patterns between rows and columns (e.g., in co-occurrence table in ecology [CS79]) or deviation from independence (e.g., Conditional Volume Test [DE85]). However, unlike the case of hypergeometric distribution, sampling a uniformly distributed contingency table is surprisingly hard [DE85, Sul18], and various MCMC algorithms and their mixing times have been proposed and analyzed [DG95, DKM97, BLSY10, BH12, PD19].

Sampling a random uniform contingency tables is closely related to the problem of counting the number of contingency tables with prescribed row and column margins. When the margins are constant or have bounded ratio close to 1, the exact or asymptotically exact asymptotics for this count are known [CM10, GM08, BLSY10], and the uniform and hypergeometric tables behave very similarly. However, when the margins are far from being constant, Barvinok [Bar10] conjectured that the behavior of the random uniform contingency table becomes drastically different from that of hypergeometric table. In a joint work with Dittmer and Pak [DLP19b], I have rigorously shown, for the first time, that there exists a sharp phase transition in the behavior of the uniform contingency table as the size of the table grow, when the margins assume two values and their ratio exceeds an explicit critical ratio $B_c$. This gives a probabilistic answer to the statistical question on why sampling a uniformly distributed contingency table is hard -- the existence of phase transition.

Currently I am working with Pak to generalize our phase transition for two distinct margin values to the general case, aiming to provide a full atlas of phase transition structure in the set of uniform contingency tables.

1.4. Interacting particle systems and discrete spatial processes. Many important phenomena that we would like to understand – formation of public opinion, trending topics on social networks, growth of stock market, development of cancer cells, outbreak of epidemics, and collective computation in distributed systems – are closely related to predicting large-scale behavior of systems of locally interacting agents. Discrete spatial processes provide a simple framework for modeling such systems: A vertex coloring $X_t : V \rightarrow \mathbb{Z}_k$ on a given graph $G = (V,E)$ updates in discrete or continuous time according to a fixed deterministic or random transition rule. In a typical setting in applied probability literature, one draws the initial coloring $X_0$ from some probability measure and asks how the probability $P(X_t$ has property $P)$ behaves. The answer usually depends on details such as topology of the underlying graph and parameters in the model.

In collaboration with many researchers in the field, I have addressed the above question for a number of models arising from different contexts: the firefly cellular automata (coupled oscillators), the cyclic cellular automata (BZ chemical reaction), the Greenberg-Hastings model (neural network), the cyclic particle system (multicolor acyclic voter model) [FL17], and lastly, the parking process and ballistic annihilation (annihilating particle systems) [DG+17]. The first model were studied on the one-dimensional integer lattice $\mathbb{Z}$ by constructing integer-valued comparison processes on $\mathbb{Z}$ and using combinatorial correspondences [LS17b, LS17a]. I have also used one of the main techniques to answer an open question on persistence of correlated partial sums [LS17a]. A special case...
(\kappa = 3) of the following two models were studied on arbitrary underlying graphs by constructing a similar comparison process on the universal cover of the underlying graph [GLS16]. For the cyclic particle system, we proved a conjecture of Bramson and Griffeath in 1989 that the 3- and 4-color system clusters on \( \mathbb{Z} \).

My recent contribution to this field concerns multitype annihilating particle systems, namely, ballistic annihilation, the parking process, and diffusion-limited annihilating systems. With Junge [JL18], I have obtained a complete phase transition picture in the asymmetric ballistic annihilation, where only some heuristic was available in the physics literature [DRFP95]. On the other hand, we proposed and studied the parking process on a class of lattice-like graphs (e.g., Cayley graphs) and obtained sharp phase transition in the long-time particle density [DG18]. Extending some of our techniques we developed for the parking process as well as using new techniques, we have answered some open problems on particle densities in diffusion-limited annihilating systems [DF19].

1.5. **Global synchronization of pulse-coupled oscillators on trees.** Systems of coupled oscillators (e.g., blinking fireflies, circadian pacemakers, BZ chemical oscillators) have been a central subject in nonlinear dynamical systems literature for decades [St00], and have found numerous applications in many areas including robotic vehicle networks [NL07] and electric power networks [DB12]. Pulse-coupling is a class of distributed time evolution rule for networked phase oscillators inspired by biological oscillators, which depends only upon event-triggered local pulse communication. Recently, theory of pulse-coupled oscillators (PCOs) are finding applications as a scalable and efficient clock synchronization algorithms for networks of resource-constrained devices such as wireless sensors [PS11].

A fundamental question in systems of coupled oscillators concerns deriving global convergence toward synchrony from all (or almost all under some measure) possible initial configurations. For some well-known models such as Kuramoto’s or Peskin’s, global convergence has been established only on complete graphs and cycles [MS90, KB12, WND13]. In a series solo papers [Ly15, Ly16, Lyu17] I introduced a classes of discrete and continuous PCOs and derived their global convergence on finite trees. Unlike complete graphs or cycles, global synchronization on trees is more versatile since it can easily be extended to general graphs by composing with a spanning tree algorithm. This led to a fast universal clock synchronization algorithm with minimal memory and communication requirement, which is especially suitable for synchronizing wireless sensor networks. More discussion on this topic is given in Section 6.

2. **Markovian theory for network data and machine learning**

My general principle for incorporating the well-established theory of Markov chains into the realm of data theory and machine learning is the following. 1) Construct versatile sampling algorithms by Markov Chain Monte Carlo (MCMC) and obtain bounds on their mixing times; 2) Extend online algorithms for i.i.d. stream of data into the Markovian setting, so that we can learn directly from MCMC trajectories. Use Markov chain limit theorems to guarantee their convergence; 3) Obtain bounds for inference using concentration inequalities for Markov chains and stability inequalities. Below we discuss more details on two of the recently completed projects along this line.

2.1. **Motif sampling and network data analysis.** Sampling is an indispensable tool in statistical analysis of large graphs and networks, which provides means of computing some of the essential network observables such as average degree, mean shortest path length, and clustering coefficient [WS98, KC14]. In [LMS19], we proposed a novel graphical sampling method called motif sampling that circumvents some of the main drawbacks of the independent and neighborhood samplings. The key idea is first to fix ‘template graph’ (motif) \( F \) of \( k \) nodes, and then to sample \( k \) nodes from a given network \( \mathcal{G} \) so that the induced subnetwork always contains a copy of \( F \). This is equivalent to conditioning the independent sampling to contain a ‘homomorphic copy’ of \( F \) so that we always sample some meaningful portion of the network, where the prescribed graph \( F \) serves as a backbone. One can then study properties of subnetworks of \( \mathcal{G} \) induced on this random copy of \( F \). An immediate advantage is that we have the freedom to change the motif \( F \) to capture different aspects of the network.

\[
\begin{array}{c}
\text{Figure 1. Independent sampling (left), neighborhood sampling (middle), and motif sampling (right).}
\end{array}
\]
Motif sampling can be formulated as sampling a random vertex map $\mathbf{x}$ from a motif $F = ([k], A_F)$ into a network $\mathcal{G} = ([n], A)$ according to the following probability distribution

$$\pi_{F \to \mathcal{G}}(\mathbf{x}) = \frac{1}{Z} \prod_{1 \leq i,j \leq k} A(x(i), x(j))^{A_F(i,j)} \frac{1}{n^k}, \tag{1}$$

where $Z$ denotes the partition function. Note that rejection and importance samplings based on independent sampling would be too costly for motif sampling, especially when the network $\mathcal{G}$ is sparse. In [LMS19], we provide two complementary Markov Chain Monte Carlo algorithms for motif sampling inspired by Glauber chain and random walk on networks, and obtain bounds on their mixing times. In practice, this tells us how long we should run the MCMC algorithms to obtain desired precision.

Given the MCMC algorithms, we can take various time averages along the MCMC trajectory to compute the expectation of various functions of the network. For instance, consider the ‘wedge motif’ $F = ([3], 1_{([1,2],[1,3])})$, where node 1 has connections to nodes 2 and 3. Given a filtration parameter $\theta \geq 0$ and a MCMC trajectory $\{\mathbf{x}_t\}_{t \geq 0}$ of embeddings $F \to \mathcal{G}$, Markov chain ergodic theorem implies

$$\mathbb{P}_{\pi_{F \to \mathcal{G}}} (A(\mathbf{x}(2), \mathbf{x}(3)) > \theta) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}(A(\mathbf{x}_t(2), \mathbf{x}_t(3)) > \theta).\tag{2}$$

When $\theta = 0$, the left hand side asks the probability that individuals $\mathbf{x}(2)$ and $\mathbf{x}(3)$ with a common friend $\mathbf{x}(0)$ in network $\mathcal{G}$ know each other, when the three are chosen randomly according to the measure (1). Notice that this is a probabilistic re-formulation of the celebrated network clustering coefficient [WS98]. Hence if we increase the filtration parameter $\theta$ from 0, we obtain a hierarchical profile of clustering coefficient of $\mathcal{G}$. Taking finite partial sum in the right hand side of (2) gives an approximate algorithm to compute this profile, where mixing time bounds and concentration inequalities for Markov chains give control on the statistical error.

Figure 2 shows example of such profiles corresponding to the Function Word Adjacency Networks associated with four novels, two by William Shakespeare and other two by Mark Twain. The profiles seem to indicate that there is a clear distinction between the two authors (with the interpretation that Shakespeare tend to repeat function words within a sentence less frequently). But how much can we trust such distinction given by the profiles? A quantitative answer to this question about reliability of inference is given by establishing what is called a ‘stability inequality’ for the observable. For the ‘wedge-profile’ in (2), one of the main result in [LMS19] asserts that for any two networks $\mathcal{G}_1$ and $\mathcal{G}_2$,

$$\left\| \mathbb{P}_{\pi_{F \to \mathcal{G}_1}} (A(\mathbf{x}(2), \mathbf{x}(3)) > \cdot) - \mathbb{P}_{\pi_{F \to \mathcal{G}_2}} (A(\mathbf{x}(2), \mathbf{x}(3)) > \cdot) \right\|_1 \leq d(\mathcal{G}_1, \mathcal{G}_2), \tag{3}$$

where the left hand side denotes the $L_1$ distance between the corresponding profiles and the right hand side is a suitable metric distance between the networks. This means that small change in the network results in a small change in the profile, so we know quantitatively how much we can infer about the networks from the profiles.

Inequalities of this type is known for the persistence diagram Topological Data Analysis [CSEH07, CCSG+09] and for homomorphism density graphon theory [Lov12], respectively.

2.2. Online Matrics Factorization for Markovian data. In modern data analysis, a central step is to find a low-dimensional representation to better understand, compress, or convey the key phenomena captured in the data. Matrix factorization provides a powerful setting for one to describe data in terms of a linear combination of factors or atoms. In this setting, we have a data matrix $X \in \mathbb{R}^{d \times n}$, and we seek a factorization of $X$ into the product $WH$ for
Online matrix factorization is a problem setting where data are accessed in a streaming manner and the matrix factors should be updated each time. That is, we get draws of $X$ from some distribution $\pi$ and seek the best factorization such that the expected loss $\mathbb{E}_{X \sim \pi} [\|X - WH\|^2_F]$ is small. This is a relevant setting in today’s data world, where large companies, scientific instruments, and healthcare systems are collecting massive amounts of data every day. One cannot compute with the entire dataset, and so we must develop online algorithms to perform the computation of interest while accessing them sequentially. There are several algorithms for computing factorizations of various kinds in an online context. Many of them have algorithmic convergence guarantees, however, all these guarantees require that data are sampled at each iteration i.i.d. with respect to previous iterations. In all of the application examples mentioned above, one may make an argument for (nearly) identical distributions, but never for independence. This assumption is critical to the analysis of previous works (see., e.g., [MBPS10, GTLY12, ZTX16]).

A natural way to relax the assumption of independence in this online context is through the Markovian assumption. In many cases one may assume that the data are not independent, but independent conditioned on the previous iteration. The central contribution of our work is to extend the analysis of online matrix factorization in [MBPS10] to the setting where the sequential data form a Markov chain. This is naturally motivated by the fact that the Markov chain Monte Carlo (MCMC) method is one of the most versatile sampling techniques across many disciplines, where one designs a Markov chain exploring the sample space that converges to the target distribution.

In a joint work with Needell and Balzano [LNB19], I have shown that the well-known OMF algorithm for i.i.d. stream of data proposed in [MBPS10], in fact converges almost surely to the set of critical points of the expected loss function, even when the data matrices form a Markov chain satisfying a mild mixing condition. One of the key ideas in our analysis is to condition the future state of the Markov chain on a distant past so that the conditional expectation of the future state is very close to its stationary expectation. This allows us to control the difference between the new and the average losses by concentration of Markov chains.

We discuss two application contexts for our result on Markovian OMF. First, we apply dictionary learning and reconstruction for images using online Non-negative Matrix Factorization for a sequence of Ising spin configurations generated by the Gibbs sampler (see Figure 4). This illustrates that we can learn dictionary image patches
from an MCMC trajectory of images. Second, we propose a novel framework for network data analysis that we call *Network Dictionary Learning* (see Figure 5). This allows one to extract ‘network dictionary patches’ from a given network to see its fundamental features and to reconstruct the network using them. Two fundamental building blocks are online NMF on Markovian data, which is the main subject in this paper, and the MCMC motif sampling algorithms that I have developed in [LMS19] (see Subsection 2.1).

3. SOLITONS, BOX-BALL SYSTEMS, AND INTEGRABLE PROBABILITY

The κ-color BBS [Tak93] is an integrable cellular automaton on the half-integer lattice \( \mathbb{N}_\frac{1}{2} \), which we think of as an array of capacity-one boxes that can fit at most one ball of any of the κ colors. Balls gradually move to the right, and the particular time evolution of BBS makes the system to be eventually decomposed into ‘solitons’ with various lengths. By arranging the solitons as columns, we can associate a time-invariant Young diagram to a given BBS configuration. An example of a 5-color BBS trajectory \( (X_t)_{t \geq 0} \) and its associated Young diagram \( \Lambda(X_0) \) is shown:

\[
\begin{align*}
& t = 0 : \\
& t = 1 : \\
& t = 2 : \\
& t = 3 : \\
& t = 4 : \\
& t = 5 : \\
& t = 6 :
\end{align*}
\]

\[
\Lambda(X_0) = \begin{array}{c}
\text{ } \text{ } \text{ } \\
\text{ } \text{ } \text{ } \\
\text{ } \text{ } \text{ } \\
\text{ } \text{ } \text{ } \\
\end{array}
\]

BBS can be obtained through different limiting procedures both from the quantum and classical inegrable systems, which makes the theory of BBS very rich [IKT12].

Recently BBS with random initial configuration is getting a considerable attention in the probability community [LLP17, CKST18, KL18, FG18, KL18, CS19a, CS19b]. There are roughly two central questions that the researchers are aiming to answer: 1) If the random initial configuration is one-sided, what is the limiting shape of the invariant random Young diagram as the system size tends to infinity? 2) If one considers the two-sided BBS (where the initial configuration is a bi-directional array of balls), what are the two-sided random initial configurations that are invariant under the BBS dynamics? In collaboration with a number of researchers in the field, I have worked extensively on the second problem in the past years, combining a number of techniques from different areas of mathematics and discovering interesting new phenomena.

<table>
<thead>
<tr>
<th>( i \geq 1, j \geq 2 ) fixed</th>
<th>( \rho_i(n) )</th>
<th>( \lambda_j(n) )</th>
<th>( \lambda_j(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcritical phase ( p^* &lt; p_0 )</td>
<td>( \Theta(n) )</td>
<td>( \Theta(\log n) )</td>
<td>( \Theta(\log n) )</td>
</tr>
<tr>
<td>Critical phase ( p^* = p_0 )</td>
<td>( \Theta(n) )</td>
<td>( \Theta(\sqrt{n}) )</td>
<td>( \Theta(\sqrt{n}) )</td>
</tr>
<tr>
<td>Supercritical phase ( p^* &gt; p_0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple ( (p^* = p_\ell \text{ for unique } \ell) )</td>
<td>( \Theta(n) )</td>
<td>( \Theta(n) )</td>
<td>( \Theta(\log n) )</td>
</tr>
<tr>
<td>Non-simple ( (p^* = p_\ell \text{ for multiple } \ell) )</td>
<td></td>
<td></td>
<td>( \Theta(\sqrt{n}) \cap \Omega(\sqrt{n}/\log n) )</td>
</tr>
</tbody>
</table>

*Table 1.* Asymptotic scaling of column \( \lambda_j \) and row \( \lambda_i \) lengths for the independence model with ball density \( p = (p_0, p_1, \cdots, p_\kappa) \) and \( p^* = \max(p_1, \cdots, p_\kappa) \). The asymptotic soliton lengths undergo a similar ‘double-jump’ phase transition depending on \( p^* - p_0 \), as in the \( k = 1 \) case established in [LLP17]. The existence of non-simple supercritical phase is unique to the multicolor \( (k \geq 2) \) case, where subsequent soliton lengths scales as \( \sqrt{n} \) instead of \( \log n \). Sharp asymptotics for the row lengths has been obtained in [KL18]. The constant factors depend \( p, i, \) and \( j \).

With Levein and Pike [LLP17], I have studied various soliton statistics of the basic 1-color BBS when the system is initialized according to a Bernoulli product measure with ball density \( p \) on the first \( n \) boxes. One of their main results is that the length of the longest soliton is of order \( \log n \) for \( p < 1/2 \), order \( \sqrt{n} \) for \( p = 1/2 \), and order \( n \) for \( p > 1/2 \). Additionally, there is a condensation toward the longest soliton in the supercritical \( p > 1/2 \) regime in the sense that, for each fixed \( j \geq 1 \), the top \( j \) soliton lengths have the same order as the longest for \( p \leq 1/2 \), whereas all but the longest have order \( \log n \) for \( p > 1/2 \). Their analysis is based on geometric mappings from the associated simple random walks to the invariant Young diagrams, which enable robust analysis of the scaling limit of the...
invariant Young diagram. However, this connection is not apparent in the general $\kappa \geq 1$ case. In fact, one of the main difficulties in analyzing the soliton lengths in the multicolor BBS is that within a single regime, there is a mixture of behaviors that we see from different regimes in the single-color case.

The row lengths in the multicolor BBS are well-understood due to my recent works with Kuniba and Okado [KLO18, KL18]. The central observation is that, when the initial configuration is given by a product measure, then the sum of row lengths can be computed via some additive functional (called ‘energy’) of carrier processes of various shapes, which are finite-state Markov chains whose time evolution is given by combinatorial $R$. In [KLO18], the ‘stationary shape’ of the Young diagram for the most general type of BBS is identified by the logarithmic derivative of a deformed character of the KR modules (or Schur polynomials in the basic case). In [KL18], I showed that the row lengths satisfy a large deviations principle and hence the Young diagram converges to the stationary shape at an exponential rate, in the sense of row scaling.

Recently, with Lewis, Pylyavskyy, and Sen [LLPS19], I have obtained scaling limits of the columns lengths in multicolor BBS using a modified version of Greene-Kleitman invariants for BBS and circular exclusion processes. Especially, the latter can be regarded as a circular version of the well-known Totally Asymmetric Simple Exclusion Process (TASEP) on a line (see, e.g., [F]). Starting from some finite number of points, at each time, a new point is added to $S^1$ independently from a fixed distribution, which then deletes the nearest counterclockwise point already on the circle. Equivalently, one can think of each point in the circle trying to jump to the clockwise direction. It turns out that this process is crucial in analyzing the permutation model, whereas for the independence model, the relevant circular exclusion process is defined on the integer ring $\mathbb{Z}$ that this process is crucial in analyzing the permutation model, whereas for the independence model, the relevant circular exclusion process is defined on the integer ring $\mathbb{Z}$.

4. Phase transition in random contingency tables with non-uniform margins

Denote by $\mathcal{M}(r,c)$ the set of all $(n \times m)$ contingency tables with row sums $r_i$ and column sums $c_j$ given by two nonnegative integer vectors $r = (r_1, \ldots, r_m) \in \mathbb{N}^m$, $c = (c_1, \ldots, c_n) \in \mathbb{N}^n$. Let $X = (X_{ij})$ be the contingency table chosen uniformly at random from $\mathcal{M}(r,c)$. When the margins are uniform, i.e. $r_1 = \ldots = r_m$ and $c_1 = \ldots = c_n$, the exact asymptotics for $|\mathcal{M}(r,c)|$ are known [CM10, GM08]. In fact, the distribution of individual entries $X_{ij}$ is asymptotically geometric and the dependence between the entries vanish as the size of the table goes to infinity [CDS10]. This agrees the asymptotic behavior of random contingency tables following the hypergeometric distribution. However, when the margins are far from being constant, Barvinok [Bar10] conjectured that the behavior of the uniform random contingency table becomes drastically different from that of hypergeometric table.

A natural model to test Barvinok’s conjecture on uniform contingency tables with non-uniform margins is to investigate properties of uniform contingency tables with only two distinct values for the margin, in relation to their ratio. In particular, fix a parameter $B \geq 1$, and take the margin vectors $r$ and $c$ so that their first entry is large with $|Bn|$ and the last $n$ entries are small with $n$ (see Figure 6 left with $\delta = 0$ and $C = 1$). In [DLP19a], we tested Barvinok’s conjecture empirically using a new MCMC algorithm introduced in [PD19], and we found that the corner entry $X_{11}$ undergoes phase transition as $B$ passes a critical value $B_c = 1 + \sqrt{2}$; it is uniformly bounded for $B < B_c$, but grows linearly in $n$ for $B > B_c$.

In a joint work with Dittmer and Pak [DLP19b], I was able to prove Barvinok’s conjecture in a more refined form (but for the ‘think bezel’ case $\delta \in (1/2, 1)$). The proof uses Barvinok’s typical table theory for uniform contingency tables [Bar09, Bar10] as well as ‘transference principle’ and concentration inequalities in probability literature. Below is a rough summary of the main results in [DLP19b].

Theorem 4.1 (Dittmer, L., and Pak ’19+). Fix constants $B, C > 0$ and $1/2 < \delta < 1$. Let $X = (X_{ij})$ be the uniform random contingency table with the first $\lfloor n^{\delta} \rfloor$ row and column margins $\lfloor BCn \rfloor$, and the last $n$ row and column margins $\lfloor Cn \rfloor$. Let $B_c = 1 + \sqrt{1 + 7/C}$. Then

$$
E[X_{11}] = \Theta(1) \quad \text{for } B < B_c, \quad \text{and} \quad E[X_{11}] = \Theta(n^{1-\delta}) \text{ for } B > B_c.
$$

Furthermore, $X$ marginally converges to the matrices given in Figure 6 in total variation distance.
5. MULTITYPE ANNIHILATING PARTICLE SYSTEMS

5.1. Two-type annihilating systems. A multitype particle system is a stochastic process defined on a d-dimensional integer lattice $\mathbb{Z}^d$ with following set of information: particle types with corresponding random walk kernel and jump rate, and their collision rule. Each site is assigned with one type initially independently, and all particles perform independent random walks and interact upon collision. The system is said to be annihilating if only particles with different types interact upon collision and they do it by annihilating with each other. Physicists have long been interested in this type of process as a model for irreversible reactions with mobile particles since the seminal works [OZ78, TW83]. In probability literature, one-type annihilating system is well-understood in all dimensions due to Arratia [Arr83], under the name of annihilating random walks.

Now consider two-type annihilating systems with particle types $A, B$ and $\lambda_A, \lambda_B$ for their corresponding jump rates. Let $p_A$ denote the density of initial $A$-particles. Bramson and Lebowitz gave a complete analysis of this system for equal jump rates $\lambda_A = \lambda_B$ in their celebrated paper [BL88]. Most notably, for critical density $p_A = 1/2$, they showed that the particle density $\rho_t$ at the origin has the following asymptotic

$$
\rho_t \approx \begin{cases} 
  t^{-d/4}, & d \leq 3 \\
  t^{-1}, & d \geq 4.
\end{cases}
$$

This is in agreement with mean field approximation in the physics literature for $d \geq 4$, but deviates from it for low dimension $d \leq 3$. The above asymptotics were conjectured to hold when $\lambda_A \neq \lambda_B$ [LC95, Koz96, KR84]. The case of asymmetric jump rate seems to be challenging, with the best known result due to Cabezas, Rolla, and Sidoravicius [CRS18]: $\rho_t \geq 1/t$ in all dimensions for $p_A = 1/2$, and that the system undergoes a phase transition from infinite visits from $A$-particles to the root (recurrence) when $p_A \geq 1/2$ to only finitely many visits (transience) for $p_A < 1/2$.

In joint work with Damron, Gravner, Junge, and Sivakoff [DGJS17], I studied the two-type annihilating system with extremely asymmetric jump rate with $\lambda_B = 0 < \lambda_A$ (a.k.a. parking process). It is helpful to visualize the $A$-particles as ‘cars’ and the $B$-particles as ‘parking spots’. Cars simultaneously perform independent simple symmetric random walks, and when one or more car encounters an available parking spot, uniformly chosen car parks there and the spot becomes unavailable (see Figure 7). Our main results show that for a large class of transitive and unimodular graphs in place of $\mathbb{Z}^d$, there exists a sharp phase transition at $p_A = 1/2$ some quantitative estimates that was lacking in the literature:

**Theorem 5.1.** Let $V_t$ be the number of visits of cars to the origin, and let $V := \lim_{t \to \infty} V_t$.

(i) If $p_A \geq 1/2$, then $V$ is infinite almost surely. Moreover, $\mathbb{E}_p V_t = (2p_A - 1)t + o(t)$.

(ii) If $p_A < 1/2$, then $V$ is finite almost surely. Moreover, $\mathbb{E}_p V < \infty$ if $p_A < (256d^6e^2)^{-1}$.

![Figure 6](image6.png)

**Figure 6.** (Left) Contingency table with parameters $n, \delta, B$ and $C$. First $\lfloor n^\delta \rfloor$ rows and columns have margins $[Cn]$, the last $n$ rows and columns have margins $[BCn]$. (Right) Limiting distributions of the entries in the uniform contingency table $X$ in the subcritical $B < B_c$ (left) and supercritical $B > B_c$ (right) regimes for thick bezels $1/2 < \delta < 1$. $\text{Geom}(\lambda)$ denotes geometric distribution with mean $\lambda$.

![Figure 7](image7.png)

**Figure 7.** Simulation of parking process with critical density $p = 0.5$. Blue=cars and Red=spots. Time goes upwards.
In [DGJ+17], we conjectured that particle densities in the critical \((p_A = 1/2)\) parking process satisfies the Bramson-Lebowitz asymptotics (5) especially for low dimensions \(d \leq 3\). This was recently addressed for \(d = 1\) by Przykucki, Roberts, and Scott [PRS19]. Simultaneously, in a work in preparation with Johnson, Damron, Junge, and Sivakoff, we have obtained a matching lower bound for all \(d \leq 3\) and for general jump rates \(0 \leq \lambda_B \leq \lambda_A\) as well as matching upper bound and upper and lower bounds on the critical exponent when \(\lambda_B = 0\):

**Theorem 5.2.** The following hold.

(i) Let \(p_A = 1/2\) and \(0 \leq \lambda_B \leq 1\). On \(\mathbb{Z}^d\) with \(d \leq 3\), we have \(p_t \gtrsim (t \log t)^{-d/4}\).

(ii) Let \(p_A = 1/2\) and \(\lambda_B = 0\). On \(\mathbb{Z}\) it holds that \(E V_t \lesssim t^{3/4}\).

(iii) Let \(\lambda_B = 0\). On \(\mathbb{Z}\), there exists \(C > 0\) such that for all \(1/4 < p_A < 1/2\) we have

\[
E p V_\infty \leq C(1 - 2p)^{-3}.
\]

(iv) Let \(\lambda_B = 0\). On \(\mathbb{Z}^d\) for \(d \leq 3\), there exists \(C > 0\) such that for all \(1/4 < p_A < 1/2\) it holds that

\[
E p V_\infty \gtrsim (1 - 2p)^{-(4/d) + 1} / \log (1 - 2p)^{-1}\]

Our proof for the original parking process in [DGJ+17] relies on a recursive distributional equation and mass-transport principle coming from the unimodularity. In our following work [DJJ+19], we were able to obtain sharper results by using a path-swapping coupling that reconstructs the parallel process as a sequential abelian process that is easier to analyze. In a work in progress with Ahn, Richey, Wang, Junge, and Sivakoff [ARW+19], I am investigating particle densities in two-type annihilating systems when same-type particles coalesce upon collision.

### 5.2. Ballistic annihilation

Ballistic annihilation (BA) is a particular instance of the multitype annihilating system we described in the previous subsection where the random walk kernels are degenerate so that each particle type maintain constant velocity until collision. This relatively simple to define system has formidable long term dependence that is both interesting and challenging to understand rigorously.

With only two velocities (types), a comparison to simple random walk ensures that the phase transition occurs when particle types are in balance. The 3-velocity case (velocities \(-v, 0, 1\)) becomes remarkably more complicated (See Figure 9 for some simulations). Ininitially each particle is assigned with one of the three velocities \(-v, 0, 1\) independently with probabilities \((1 - p)(1 - \lambda), p, (1 - p)\lambda\), respectively, for parameters \(p\) and \(\lambda\). The central question is to characterize the critical value \(p_c(\lambda, v)\) of density \(p\) of inert particles so that the origin is eventually visited (fluctuation) or not (fixation), for each fixed \(\lambda\) and \(v\). Due to lack of monotonicity, it is not even clear that there exists a single such critical value.

In the early 1990s, physicists were interested in proving for the totally symmetric case \((v = 1\) and \(\lambda = 1/2\)), confirming a simple heuristic that \(p_c(1/2, 1) = 1/4\) [BNRL93]. This was a well-known open problem in the mathematics community for decades, and only recently it has been addressed by Haslegrave, Sidoravicius, and Tournier [HST18]. However, their argument, which relies on the fact that inverting a configuration on a finite interval preserves both the measure and collision event in the totally symmetric case, does not generalize to the general asymmetric case.

In a joint work with Junge [JL18], I have obtained a nearly complete phase structure of BA in the most general setting, giving characterization of the fluctuation and fixation regimes that depending on certain conditional probabilities for collisions between three particles. We achieve this by invoking a mass transport principle in several places to relate key quantities in order to derive a master equation that describes the
system. The existence of sharp phase transition depends on whether these conditional probabilities are monotonic in $p$, which we are working to address in order to obtain the ‘critical curve’ in the phase plane.

6. Synchronization of Pulse-Coupled Oscillators

The fundamental difficulty in understanding collective behaviors of coupled oscillators is the lack of monotonicity due to cyclic hierarchy of phase space $S^1$. Three commonly used techniques bypass this difficulty but have their weaknesses. First, a total ordering on oscillator phases emerges if the initial phases are concentrated within a half of $S^1$ (e.g., from the most lagging to the most advanced). A broad class of couplings respects this ordering and contracts the phase configuration toward synchrony. Second, in order to achieve such concentration in the first place, heavy randomization could be used to turn the system into a Markov chain; then the strong Markov property ensures that concentration is achieved after exponential mount of time. On the other hand, by using an unbounded memory per node, one can ‘lift’ the cyclic phase space $S^1$ into totally ordered $\mathbb{R}$, and let the nodes tune toward local maxima. Then global maximum propagates and subsumes all the other nodes in $O(d)$ time. This classic idea in theoretical computer science dates back to Lamport [Lam78], which in fact also inspired our lifting technique for the 3-color CCA/GHM on arbitrary graphs. However, use of heavy randomization and unbounded memory are too expensive in terms of running time and memory efficiency. Hence we are left with the following question: how can we drive the system close to synchrony from an arbitrary initial configuration in polynomial time without using unbounded memory?

In a series of three solo papers, I addressed the above question for certain class of PCOs on finite trees. In the first two papers [Lyu15, Lyu16], I introduced the $\kappa$-color FCA as a discrete model for inhibitory PCOs and proved the following theorem on finite trees.

**Theorem 5.1.** Let $(X_t)_{t \geq 0}$ be the random $\kappa$-color FCA trajectory on a finite tree $T$ with maximum degree $\Delta$.

(i) If $\Delta \geq \kappa$, then $X_t$ does not synchronize with positive probability.

(ii) For $\kappa \leq 6$, $X_t$ synchronizes with probability 1 iff $\Delta < \kappa$.

(iii) For $\kappa \geq 7$, $X_t$ does not synchronize with positive probability on some $T$ with $\Delta \leq \kappa/2 + 1$.

In fact, the maximum degree condition (i) in the above theorem is a fundamental issue in inhibitory PCOs on finite trees. Namely, nodes with large degree may receive input pulses so often that its phase is constantly inhibited and it may never send pulses to its neighbors. This divides the tree into non-communicating components so global synchrony may not emerge.

In the third paper [Lyu17], I extended the 4-color FCA to what I call the adaptive 4-coupling, which is a continuous-time and continuous-state pulse-coupling for period 1 phase oscillators. The key innovation to surpass the degree constraint is that each oscillator has an auxiliary state variable, which may throttle the input pulse. This effectively breaks symmetries in local configurations on finite trees and reduces the relevant diameter to consider in constant time. The main result is the following:

**Theorem 5.2.** Let $(\Sigma_\tau(t))_{t \geq 0}$ be an adaptive 4-coupling trajectory on a finite tree $T$ with diameter $d$. For arbitrary $\Sigma_\tau(0)$, the trajectory synchronizes in time $87d$.

Moreover, in order to overcome the restriction on tree topology, we precompose the adaptive 4-coupling with distance $\leq 2$ coloring and spanning tree algorithms (both of which use randomization). Denote the resulting multi-layer distributed algorithm by $\mathcal{A}$. We have the following corollary.

**Corollary 5.3.** Consider a distributed communication network of anonymous processors on a connected simple graph $G = (V, E)$ of maximum degree $\Delta$ and diameter $d$. Suppose each processor has a local clock of identical frequency.

![Figure 10. Illustration of Theorem 5.2.](image-url)
(i) The adaptive 4-coupling can be implemented with $O(\log \Delta)$ memory per node and uses $O(\Delta)$ bits of binary pulse communication per unit time.

(ii) Let $\tau_G$ be the worst case time until synchronizing a given configuration. Then

$$E(\tau_G) = O(|V| + (\Delta^2 + \Delta) \log |V|).$$

To my knowledge, the derivation of global convergence of PCOs on trees, especially with an explicit bound which is optimal up to a constant factor, was for the first time. Moreover, the clock synchronization algorithm is scalable and can be applied for dynamic and growing networks, as long as the maximum degree is bounded. Hence it is especially suitable in modern wireless sensor networks.

The largest bottleneck for the upper bound (6) comes from the unknown diameter of random spanning tree constructed by the algorithm of Itkis and Levin [IL94]. This is essentially due to the unknown output of a leader-election subroutine. We can omit the $O(|V|)$ term in the bound (6) if we could answer the following question:

**Question 5.4.** Is there a (randomized) distributed spanning tree algorithm $\mathcal{T}$, which computes a spanning tree $T$ of a given connected graph $G = (V,E)$ with maximum degree $\Delta$ and diameter $d$, with the following properties?

(i) $\mathcal{T}$ can be implemented on $G$ with $O(\log \Delta)$ memory per node.

(ii) $\text{diam}(T) = \text{diam}(G)^{O(1)}\log |V|$.

(iii) $E(\text{worst case running time}) = \text{diam}(G)^{O(1)}\log |V|$.

Another open question that I am working on, which is important both for theory and application aspects, is to extend the above results for oscillators with non-identical frequencies.

**Question 5.5.** Find a coupling under which arbitrary (or almost all) initial phase configuration on every undirected finite tree synchronizes, where the oscillators may have different natural frequency or communication between oscillators may have propagation delay. Furthermore, show that the worst-case time to convergence is a polynomial in the diameter of the underlying tree.

**References**


Correction for ambiguous solutions in factor analysis using a penalized dictionary learning approach.


Sampling random graph homomorphisms and applications to network data analysis.


Sampling random graph homomorphisms and applications to network data analysis.


Hanbaek Lyu, Deana Needell, and Laura Balzano, Online dictionary learning.


Hanbaek Lyu, Deana Needell, and Laura Balzano, Online matrix factorization for markovian data and applications to network dictionary learning.


Parking on the integers.

Double jump phase transition in a random soliton cellular automaton.


Sampling random graph homomorphisms and applications to network data analysis.


Parking on the integers.


HANBAEK LYU, DEPARTMENT OF MATHEMATICS, THE OHIO STATE UNIVERSITY, COLUMBUS, OH 43210.

*Email address: colourgraph@gmail.com*