RESEARCH STATEMENT (Short version)

Hanbaek Lyu

My research is broadly driven by questions related to understanding emergent or latent properties of complex systems and data sets. Such questions and techniques that I develop to address them span across the fields of probability, combinatorics, dynamical systems, optimization, and machine learning.

1. **Online Matrix/Tensor factorization + MCMC sampling + Networks.** A central step in modern data analysis is to find a low-dimensional representation to better understand, compress, or convey the key phenomena captured in the data. Matrix/Tensor factorization algorithms are machine learning techniques that learn interpretable latent structures of complex data sets and are applied regularly in text and image data analysis [EA06, MES07, Pey09, KB09]. Modern data are not only large in their size, but they may also have intricate structure. A prime example is in the form of *networks*, which are formal representation of the architecture of interactions between entities in many complex systems in nature. As most real-world networks are sparse, independently choosing a set of *k* nodes from a network does not return any meaningful information with high probability. Therefore, *in order to learn efficiently and correctly from large and structured data, we need to develop a comprehensive theory for simultaneous Markov Chain Monte Carlo (MCMC) sampling and matrix/tensor factorization* [LNB20, SLN20]. Our results opens up a wide variety of applications by combining classical OMF algorithms with MCMC sampling in diverse domains.

*Network dictionary learning* (NDL) is a novel framework that we develop in [LKVP20] to learn latent mesoscale structures of networks. Combining OMF for Markovian data in [LNB20] and MCMC motif sampling algorithms for networks that we develop in [LMS19], NDL shows that various social, collaboration, and PPI networks possess a few ‘latent motifs’ that together can well-approximate most subnetworks at a fixed mesoscale. The ability to encode a network using a set of latent motifs opens up a wide variety of network-analysis tasks, such as comparison, denoising, and edge inference. For instance, NDL applied for network denoising and edge inference tasks achieves state-of-the-art results even compared to other well-known supervised graph neural network algorithms such as node2vec [GL16] and DeepWalk [PAR14].

2. **Combinatorial and probabilistic approaches to oscillator and clock synchronization (NSF: DMS-2010035).** If a group of people is given local clocks with arbitrarily set times, and there is no global reference (for example GPS), is it possible for the group to synchronize all clocks by only communicating with nearby members? In order for a distributed system to be able to perform high-level tasks that may go beyond the capability of an individual agent, the system must first solve a "clock synchronization" problem to establish a shared notion of time. The study of clock synchronization (or coupled oscillators) has been an important subject of research in mathematics and various areas of science for decades [Str00], with fruitful applications in many areas including wildfire monitoring, electric power networks, robotic vehicle networks, large-scale information fusion, and wireless sensor networks [DB12, NL07, PS11].

However, there has been a gap between our theoretical understanding of systems of coupled oscillators and practical requirements for clock synchronization algorithms in modern application contexts such as robustness in arbitrary perturbation, bounded memory consumption, and energy efficiency in communication [MS90, KB12, WND13]. In a series of solo papers [Lyu15, Lyu16, Lyu17], I have developed a systematic approaches using discrete pulse-coupled oscillators to not only break the notorious “half-circle barrier” in the literature, but also to meet with the minimal resource and energy constraint in clock synchronization algorithms. The novelty and significance of my contribution and my vision in the field of oscillator and clock synchronization have been recognized by the National Science Foundation and was awarded by the NSF Grant DMS-2010035 in May 2020. Supported by NSF, I have already successfully mentored a group of REU students in summer 2020 on the project “Machine learning approaches to oscillator and clock synchronization”.

3. **Phase transition in contingency tables with non-uniform margins.** Contingency tables are $n \times m$ matrices of nonnegative integer entries with prescribed row sums $\mathbf{r} = (a_1, \ldots, a_m)$ and columns sums $\mathbf{c} = (b_1, \ldots, b_n)$ called *margins*, where by $\mathcal{M}(\mathbf{r}, \mathbf{c})$ we denote the set of all such tables. They are fundamental objects in statistics for studying dependence structure between two or more variables, see e.g. [Eve92, FLL17, Kat14]. For the fundamental problems of *Counting* their number $|\mathcal{M}(\mathbf{r}, \mathbf{c})|$ and *sampling* an element from $\mathcal{M}(\mathbf{r}, \mathbf{c})$, the historic guiding principle has been the *independent heuristic*, which was introduced by I. J. Good as far back as in 1950 [Goo50] – It asserts that the constraints for the rows and columns of the table are asymptotically independent as the size of the table grows to infinity. This yields a simple yet surprisingly accurate formula that approximates the count $|\mathcal{M}(\mathbf{r}, \mathbf{c})|$. The independence heuristic and also implies the hypergeometric (or Fisher-Yates) should approximate the uniform distribution.
on $\mathcal{M}(r,c)$. Both of these implications of the independent heuristic have been proved and disproved in some extreme cases [CM10, GM08, BLSY10, Bar10, BH12], and understanding the what is really happening has been an open problem for a decade.

My contributions [DLP20, LP20] provide the first complete answer to this puzzle; Contingency tables exhibit a sharp phase transition when the heterogeneity of margins exceeds a certain critical threshold. This is quite surprising since the independence heuristic does not “notice” the phase transition. Our results show that the historic independent heuristic in 1950 captures the structure of contingency tables remarkably well for near-homogeneous margins, but in general positive correlation between rows and columns may emerge and the heuristic fails dramatically.

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, 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 $X_t : V \to Z_\kappa$ 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) [LS17b, LS17a], the cyclic cellular automata (BZ chemical reaction) and the Greenberg-Hastings model (neural network) [GLS16], the cyclic particle system (multicolor acrylic voter model) [FL17], the parking process and ballistic annihilation (annihilating particle systems) [DG1*17, JL18, DLS20], and diffusion-limited annihilating systems [JL18]. In the aforementioned works, I was able to settle a conjecture of Bramson and Griffeath in 1989 that the 3- and 4-color system clusters on $Z$ and also Bramson-Lebowitz asymptotics [BL91] in the asymmetric two-type particle setting.

5. Solitons, box-ball systems, and integrable probability. Integrable systems roughly refer to nonlinear systems where one can explicitly write down the solutions in terms of a set of more elementary functions, just like we can superimpose solutions to linear differential equations. Understanding such systems can greatly advance our understanding on more general nonlinear dynamical systems. One of the most significant integrable systems is given by the Korteweg-de Veris (KdV) equation, which was proposed by Korteweg and de Veris in 1895 in order to model particle-like waves (solitons) on shallow water surfaces [New85, Woy06, MQR16]. It has been a central topic in statistical and mathematical physics over the past seventy years [DJ06]. A line of my research program considers a discrete counterpart of the KdV equation, known as the box-ball systems (BBS) [TS90, TTMS96]. They are known to arise both from the quantum and classical integrable systems and enjoy deep connections to quantum groups, crystal base theory, solvable lattice models, the Bethe ansatz and so forth [FY00, HHI*01, KOS*06, IKT12].

An important but notoriously difficult question for KdV equations is the following: If the system is randomly initialized, what is the limiting statistics of the solitons emerging from the system, as the system size tends to infinity? My research on BBS aims to address this question in the discrete setting of BBS, and it leads a body of work in the literature of randomized BBS [LLP17, CKST18, KL18, FG18, KL18, CS19a, CS19b, LLPS]. In 2017, we addressed this question for the basic 1-color BBS with i.i.d. initial configuration [LLP17]. This work is considered as the foundational paper in the literature of randomized BBS and has cited by most papers in the literature. In [KL18, KLO18], we 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. One of the contribution there is to merge two entirely different approach – large deviations and Markov chains in probability and Thermodynamic Bethe Ansatz in statistical physics. Furthermore, in [LLPS], we 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. I am continuing to investigate on this topic and hope that we may get a better understanding on KdV from investigating its discrete counterpart of BBS.

References


Hanbaek Lyu, Department of Mathematics, The Ohio State University, Columbus, OH 43210.

Email address: colourgraph@gmail.com