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 [PARS14].

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 approches 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 $\mathbf{r} = (a_1, \cdots, a_m)$ and columns sums $\mathbf{c} = (b_1, \cdots, b_n)$ 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 $\mathscr{M}(\mathbf{r},\mathbf{c})$. 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 $X_t : V \to \mathbb{Z}_\kappa$ 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 $\mathbb{P}(X_t \text{ has property } P)$ 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 checmial reaction) and the Greenberg-Hastings model (neural network) [GLS16], the cyclic particle system (multicolor acyclic voter model) [FL17], the parking process and ballistic annihilation (annihilating particle systems) [DGJ+17, JL18, DLS20], and diffusion-limited annihilating systems [JJLS20]. 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 $\mathbb{Z}$ 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 [FYO00, 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 [KLO18, KL18], 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

[Bar10] Alexander Barvinok, *What does a random contingency table look like?*, Combinatorics, Probability and Computing **19** (2010), no. 4, 517–539.

[BH12] Alexander Barvinok and J Hartigan, *An asymptotic formula for the number of non-negative integer matrices with prescribed row and column sums*, Transactions of the American Mathematical Society **364** (2012), no. 8, 4323–4368.

[BL91] Maury Bramson and Joel L Lebowitz, *Asymptotic behavior of densities for two-particle annihilating random walks*, Journal of statistical physics **62** (1991), no. 1-2, 297–372.

[BLSY10] Alexander Barvinok, Zur Luria, Alex Samorodnitsky, and Alexander Yong, *An approximation algorithm for counting contingency tables*, Random Structures & Algorithms **37** (2010), no. 1, 25–66.

[CKST18] David A Croydon, Tsuyoshi Kato, Makiko Sasada, and Satoshi Tsujimoto, *Dynamics of the box-ball system with random initial conditions via Pitman's transformation*, arXiv preprint arXiv:1806.02147 (2018).

[CM10] E Rodney Canfield and Brendan D McKay, *Asymptotic enumeration of integer matrices with large equal row and column sums*, Combinatorica **30** (2010), no. 6, 655.

[CS19a] David A Croydon and Makiko Sasada, *Duality between box-ball systems of finite box and/or carrier capacity*, arXiv preprint arXiv:1905.00189 (2019).

[CS19b] ______, *Invariant measures for the box-ball system based on stationary Markov chains and periodic Gibbs measures*, arXiv preprint arXiv:1905.00186 (2019).

[DB12] Florian Dorfler and Francesco Bullo, *Synchronization and transient stability in power networks and nonuniform kuramoto oscillators*, SIAM Journal on Control and Optimization **50** (2012), no. 3, 1616–1642.

[DGJ+17] Micheal Damron, Janko Gravner, Matthew Junge, Hanbaek Lyu, and David Sivakoff, *Parking on transitive unimodular graphs*, arXiv.org/1710.10529 (2017).

[DJ06] EM De Jager, *On the origin of the korteweg-de vries equation*, arXiv preprint math/0602661 (2006).

[DLP20] Sam Dittmer, Hanbaek Lyu, and Igor Pak, *Phase transition in random contingency tables with non-uniform margins*, Transactions of the AMS. **373** (2020), 8313–8338.

[DLS20] Michael Damron, Hanbaek Lyu, and David Sivakoff, *Stretched exponential decay for subcritical parking times on* $\mathbb{Z}^d$, arXiv preprint arXiv:2008.05072 (2020).

[EA06] Michael Elad and Michal Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Transactions on Image processing **15** (2006), no. 12, 3736–3745.

[Eve92] Brian S Everitt, *The analysis of contingency tables*, Chapman and Hall/CRC, 1992.

[FG18] Pablo A Ferrari and Davide Gabrielli, *BBS invariant measures with independent soliton components*, arXiv preprint arXiv:1812.02437 (2018).

[FL17] Eric Foxall and Hanbaek Lyu, *Clustering of three and four color cyclic particle systems in one dimension*, arXiv.org/1711.04741 (2017).

[FLL17] Morten Fagerland, Stian Lydersen, and Petter Laake, *Statistical analysis of contingency tables*, Chapman and Hall/CRC, 2017.

[FYO00] Kaori Fukuda, Yasuhiko Yamada, and Masato Okado, *Energy functions in box ball systems*, International Journal of Modern Physics A **15** (2000), no. 09, 1379–1392.

[GL16] Aditya Grover and Jure Leskovec, *node2vec: Scalable feature learning for networks*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.

[GLS16] Janko Gravner, Hanbaek Lyu, and David Sivakoff, *Limiting behavior of 3-color excitable media on arbitrary graphs*, Annals of Applied Probability (to appear) (2016).

[GM08] Catherine Greenhill and Brendan D McKay, *Asymptotic enumeration of sparse nonnegative integer matrices with specified row and column sums*, Advances in Applied Mathematics **41** (2008), no. 4, 459–481.

[Goo50] Isidore Jacob Good, *Probability and the weighing of evidence*.

[HHI+01] Goro Hatayama, Kazuhiro Hikami, Rei Inoue, Atsuo Kuniba, Taichiro Takagi, and Tetsuji Tokihiro, *The* $A_M^{(1)}$ *automata related to crystals of symmetric tensors*, Journal of Mathematical Physics **42** (2001), no. 1, 274–308.

[IKT12] Rei Inoue, Atsuo Kuniba, and Taichiro Takagi, *Integrable structure of box–ball systems: crystal, Bethe ansatz, ultradiscretization and tropical geometry*, Journal of Physics A: Mathematical and Theoretical **45** (2012), no. 7, 073001.

[JJLS20] Tobias Johnson, Matthew Junge, Hanbaek Lyu, and David Sivakoff, *Particle density in diffusion-limited annihilating systems*, arXiv preprint arXiv:2005.06018 (2020).

[JL18] Matthew Junge and Hanbaek Lyu, *The phase structure of asymmetric ballistic annihilation*, arXiv preprint arXiv:1811.08378 (2018).

[Kat14] Maria Kateri, *Contingency table analysis*, Statistics for Industry and Technology **525** (2014).

[KB09] Tamara G Kolda and Brett W Bader, *Tensor decompositions and applications*, SIAM Review **51** (2009), no. 3, 455–500.

[KB12] Johannes Klinglmayr and Christian Bettstetter, *Self-organizing synchronization with inhibitory-coupled oscillators: Convergence and robustness*, ACM Transactions on Autonomous and Adaptive Systems (TAAS) **7** (2012), no. 3, 30.

[KL18] Atsuo Kuniba and Hanbaek Lyu, *Large deviations and one-sided scaling limit of multicolor box-ball system*, arXiv preprint arXiv:1808.08074 (2018).

[KLO18] Atsuo Kuniba, Hanbaek Lyu, and Masato Okado, *Randomized box-ball systems, limit shape of rigged configurations and Thermodynamic Bethe ansatz*, arXiv preprint arXiv:1808.02626v4 (2018).

[KOS+06] Atsuo Kuniba, Masato Okado, Reiho Sakamoto, Taichiro Takagi, and Yasuhiko Yamada, *Crystal interpretation of Kerov–Kirillov–Reshetikhin bijection*, Nuclear Physics B **740** (2006), no. 3, 299–327.

[LKVP20] Hanbaek Lyu, Yacoub Kureh, Joshua Vendrow, and Mason A. Porter, *Learning low-rank latent mesoscale structures in networks*, Draft available: https://hanbaeklyudotcom.files.wordpress.com/2020/10/ndl-1.pdf (2020).

[LLP17] Lionel Levine, Hanbaek Lyu, and John Pike, *Double jump phase transition in a random soliton cellular automaton*, arXiv preprint arXiv:1706.05621 (2017).

[LLPS] Joel Lewis, Hanbaek Lyu, Pasha Pylyavskyy, and Arnab Sen, *Scaling limit of soliton lengths in a multicolor box-ball system*.

[LMS19] Hanbaek Lyu, Facundo Memoli, and David Sivakoff, *Sampling random graph homomorphisms and applications to network data analysis*, arXiv:1910.09483 (2019).

[LNB20] Hanbaek Lyu, Deanna Needell, and Laura Balzano, *Online matrix factorization for Markovian data and applications to network dictionary learning*, To appear in Journal of Machine Learning Research (arXiv:1911.01931) (2020).

[LP20] Hanbaek Lyu and Igor Pak, *On the number of contingency tables and the independence heuristic*, arXiv preprint arXiv:2009.10810 (2020).

[LS17a] Hanbaek Lyu and David Sivakoff, *Persistence of sums of correlated increments and clustering in cellular automata*, Submitted. Preprint available at arXiv.org/1706.08117 (2017).

[LS17b] ———, *Synchronization of finite-state pulse-coupled oscillators on $\mathbb{Z}$*, arXiv.org:1701.00319 (2017).

[Lyu15] Hanbaek Lyu, *Synchronization of finite-state pulse-coupled oscillators*, Physica D: Nonlinear Phenomena **303** (2015), 28–38.

[Lyu16] ———, *Phase transition in firefly cellular automata on finite trees*, arXiv preprint arXiv:1610.00837 (2016).

[Lyu17] ———, *Global synchronization of pulse-coupled oscillators on trees*, To appear in SIAM Journal on Applied Dynamical Systems. Preprint available at arXiv:1604.08381 (2017).

[MES07] Julien Mairal, Michael Elad, and Guillermo Sapiro, *Sparse representation for color image restoration*, IEEE Transactions on Image Processing **17** (2007), no. 1, 53–69.

[MQR16] Konstantin Matetski, Jeremy Quastel, and Daniel Remenik, *The kpz fixed point*, arXiv preprint arXiv:1701.00018 (2016).

[MS90] Renato E Mirollo and Steven H Strogatz, *Synchronization of pulse-coupled biological oscillators*, SIAM Journal on Applied Mathematics **50** (1990), no. 6, 1645–1662.

[New85] Alan C Newell, *Solitons in mathematics and physics*, SIAM, 1985.

[NL07] Sujit Nair and Naomi Ehrich Leonard, *Stable synchronization of rigid body networks*, Networks and Heterogeneous Media **2** (2007), no. 4, 597.

[PARS14] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena, *DeepWalk: Online learning of social representations*, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 701–710.

[Pey09] Gabriel Peyré, *Sparse modeling of textures*, Journal of Mathematical Imaging and Vision **34** (2009), no. 1, 17–31.

[PS11] Roberto Pagliari and Anna Scaglione, *Scalable network synchronization with pulse-coupled oscillators*, IEEE Transactions on Mobile Computing **10** (2011), no. 3, 392–405.

[SLN20] Christopher Strohmeier, Hanbaek Lyu, and Deanna Needell, *Online nonnegative tensor factorization and cp-dictionary learning for markovian data*, arXiv preprint arXiv:2009.07612 (2020).

[Str00] Steven H Strogatz, *From kuramoto to crawford: exploring the onset of synchronization in populations of coupled oscillators*, Physica D: Nonlinear Phenomena **143** (2000), no. 1, 1–20.

[TS90] Daisuke Takahashi and Junkichi Satsuma, *A soliton cellular automaton*, J. Phys. Soc. Japan **59** (1990), no. 10, 3514–3519.

[TTMS96] Tetsuji Tokihiro, Daisuke Takahashi, Junta Matsukidaira, and Junkichi Satsuma, *From soliton equations to integrable cellular automata through a limiting procedure*, Physical Review Letters **76** (1996), no. 18, 3247.

[WND13] Yongqiang Wang, Felipe Nunez, and Francis J Doyle, *Increasing sync rate of pulse-coupled oscillators via phase response function design: theory and application to wireless networks*, Control Systems Technology, IEEE Transactions on **21** (2013), no. 4, 1455–1462.

[Woy06] Wojbor A Woyczynski, *Burgers-kpz turbulence: Göttingen lectures*, Springer, 2006.

HANBAEK LYU, DEPARTMENT OF MATHEMATICS, THE OHIO STATE UNIVERSITY, COLUMBUS, OH 43210.
Email address: colourgraph@gmail.com