My graduate research has been mostly concentrated on problems in probability and theoretical computer science. More specifically, I have worked on problems in deterministic random walks, diffusions on graphs, stochastic block models. My research explores variations of randomness in complex systems, and more specifically, how drastically the dynamics and structure of a network change when a little bit of information is added to “chaos”. On one hand, I investigate how much determinism in diffusions (the random spread of various objects in media) de-randomizes the process, and on the other hand, I look at how superposing “planted” information on a random network changes its structure in such a way that the “planted” structure can be recovered. For these types of problems, a great deal of inspiration, motivation and intuition comes from statistical physics. Below I detail my current and past research projects and also outline further research directions arising out of these projects which I am planning to pursue. I wish to emphasize that I am also very interested in exploring new areas and working on new problems.

1 Deterministic random walks

The first main part of my dissertation is concerned with rotor-router walks [50], a deterministic counterpart to random walk, which is the mathematical model of a path consisting of a succession of random steps. Around the time of the initiation of the study of deterministic walks (published under the name ”Eulerian walkers as a model of self-organized criticality” by Priezzhev, Dhar, Dhar and Krishnamurthy [50]), there was, and still is, much interest in the study of complex systems exhibiting self-organized criticality [8]. Different models have been proposed for these type of systems such as sandpiles [8], earthquakes [55], forest-fires [18], and biological evolution [7]. These models involve a slowly driven system, in which an external disturbance propagates in the random medium following a random or deterministic rule. After some time, the disturbance produces long-range spatial correlations in the system [25].

In a rotor-router walk, each node in a network remembers its neighbors in a specific cyclical order, and each time the node is visited by a particle, it sends it to its neighbors in that order. This model has been employed in studying fundamental questions such as optimal transport [42], information spreading in networks [17], load balancing in distributed computing [27], condensed matter [49]. Deep connections with famous statistical mechanics also arise through the concept of self-organized criticality, which is a property of dynamical systems to stabilize without outside intervention [21].

More formally, let $G = (V, E)$ be a finite or infinite directed graph. For $v \in V$ let $E_v \subset E$ be the set of outbound edges from $v$, and let $C_v$ be the set of all cyclic permutations of $E_v$. A rotor configuration on $G$ is a choice of an
outbound edge $\rho(v) \in E_v$ for each $v \in V$. A rotor mechanism on $G$ is a choice of cyclic permutation $m(v) \in C_v$ for each $v \in V$. Given $\rho$ and $m$, the simple rotor walk started at $X_0$ is a sequence of vertices $X_0, X_1, \ldots \in \mathbb{Z}^d$ and rotor configurations $\rho = \rho_0, \rho_1, \ldots$ such that for all integer times $t \geq 0$

$$
\rho_{t+1}(v) = \begin{cases} 
m(v)(\rho_t(v)), & v = X_t \\
\rho_t(v), & v \neq X_t
\end{cases}
$$

and $X_{t+1} = \rho_{t+1}(X_t)^+$, where $e^+$ denotes the target of the directed edge $e$. In words, the rotor at $X_t$ “rotates” to point to a new neighbor of $X_t$ and then the walker steps to that neighbor.

### 1.1 Range

I study and show results on the volume (the range) of the territory explored by the random rotor-router model, confirming an old prediction of physicists that in two dimensions the range in $n$ steps is about $n^{2/3}$ [24]. This result is especially interesting because it shows that the territory covered by rotor-walk on the two dimensional lattice is significantly smaller than that of random walk on the same network. In recent work, we also show results on the range of rotor walk on other families of networks, such as the hypercube [23].

We show the following results, where $R_t$ denotes the range of the walk at time $t$:

**Theorem 1** For any Eulerian graph $G$ of bounded degree that is “at least d-dimensional”, the number of distinct sites visited by a rotor walk started at $o$ in $t$ steps satisfies

$$
\# R_t \geq ct^{d/(d+1)}.
$$

for a constant $c > 0$ depending only on $G$ (and not on $\rho$ or $m$).

The main difficulty in proving upper bounds for $\# R_t$ lies in showing that the uniform rotor walk is recurrent. This seems to be a difficult problem in $\mathbb{Z}^2$, but we can show it for two different directed graphs obtained by orienting the edges of $\mathbb{Z}^2$: the Manhattan lattice and the F-lattice, pictured in Figure 1. The F-lattice has two outgoing horizontal edges at every odd node and two outgoing vertical edges at every even node (we call $(x,y)$ odd or even according to whether $x+y$ is odd or even). The Manhattan lattice is full of one-way streets: rows alternate pointing left and right, while columns alternate pointing up and down.

![Figure 1: Two different periodic orientations of the square grid with indegree and outdegree 2.](image)

**Theorem 2** Uniform rotor walk is recurrent on both the F-lattice and the Manhattan lattice.

The proof uses a connection to the mirror model and critical bond percolation on $\mathbb{Z}^2$.

We also show a range result on the comb graph:

**Theorem 3** Consider uniform rotor walk on the comb with any rotor mechanism. Let $n \geq 2$ and $t = \left\lfloor \frac{16}{3} n^3 \right\rfloor$. For any $a > 0$ there exist constants $c, C > 0$ such that

$$
P\{D_{n-\sqrt{cn \log n}} \subset R_t \subset D_{n+\sqrt{cn \log n}} \} > 1 - Cn^{-a},
$$

where $D_n$ represents the diamond of side length $n$.

In parallel to exploring the connections to self-organized criticality of such deterministic walks, in general, the main research directions are to compare the behavior of the deterministic walk with that of the random walk. In some networks, random walk explores a vaster portion than rotor walk. A striking result we show is that the rotor walk covers a much larger territory than its random counterpart on a different class of complex networks [24].
1.2 Escape rates

A different quantity of interest is the fraction of particles performing rotor-walk that escape to infinity. This is a way of quantifying recurrence/transience - the property of the walk to cover the territory infinitely many times or to escape toward infinity. We show asymptotic results for this quantity on the lattices in two and higher dimensions [21]. Throughout this research, I make use of various simulations that have helped in generating the ideas to prove the theorems, as well as those leading to bold conjectures; some of these can be found at [1]. An important impact of rigorously proving such theorems about whether rotor walk escapes to infinity or tight bounds on the volume of the territory typically covered by the walk is in making complex systems simulations faster, more reliable, predictable and informative.

We say that \( \rho \) is recurrent if the rotor walk with initial configuration \( \rho \) returns to the origin infinitely often \((x_n = o \text{ for infinitely many } n)\); otherwise, we say that \( \rho \) is transient. To quantify the degree of transience, consider the following experiment: let each of \( n \) particles in turn perform rotor walk starting from the origin until either returning to the origin or escaping to infinity. Denote by \( I(\rho,n) \) the number of walks that escape to infinity. (Importantly, we do not reset the rotors in between trials!)

Our first result gives a corresponding lower bound for the initial configuration \( \uparrow \) in which all rotors start pointing in the same direction: \( \uparrow(x) = e_d \) for all \( x \in \mathbb{Z}^d \).

**Theorem 1** For the rotor walk on \( \mathbb{Z}^d \) with \( d \geq 3 \) with all rotors initially aligned \( \uparrow \), a positive fraction of particles escape to infinity; that is,

\[
\lim\inf_{n \to \infty} \frac{I(\uparrow,n)}{n} > 0.
\]

Our next result concerns the fraction of particles that escape in dimension 2: for any rotor configuration \( \rho \) this fraction is at most \( \frac{\pi}{2} \log n \), and for the initial configuration \( \uparrow \) it is at least \( \frac{c}{\log n} \) for some \( c > 0 \).

**Theorem 2** For rotor walk in \( \mathbb{Z}^2 \) with any rotor configuration \( \rho \), we have

\[
\lim\sup_{n \to \infty} \frac{I(\rho,n)}{n/\log n} \leq \frac{\pi}{2}.
\]

Moreover, if all rotors are initially aligned \( \uparrow \), then

\[
\lim\inf_{n \to \infty} \frac{I(\uparrow,n)}{n/\log n} > 0.
\]

![Figure 2: The configuration of rotors in \( \mathbb{Z}^2 \) after \( n \) particles started at the origin have escaped to infinity, with initial configuration \( \uparrow \) (that is, all rotors send their first particle North). Top: \( n = 100 \); Bottom: \( n = 480 \). Each non-white pixel represents a point in \( \mathbb{Z}^2 \) that was visited at least once, and its color indicates the direction of its rotor.](image)

1.3 Future directions

A natural followup of the results in [24] is the following conjecture

**Conjecture 1 (Kapri-Dhar [?])** The set of sites \( R_t \) visited by the clockwise uniform rotor walk in \( \mathbb{Z}^2 \) is asymptotically a disk. There exists a constant \( c \) such that for any \( \epsilon > 0 \),

\[
\mathbb{P}\{\mathcal{D}_{(c-\epsilon)t^{1/3}} \subset R_t \subset \mathcal{D}_{(c+\epsilon)t^{1/3}}\} \to 1
\]

as \( t \to \infty \), where \( \mathcal{D}_r = \{(x,y) \in \mathbb{Z}^2 : x^2 + y^2 < r^2\} \).
Rumor-spreading is a simple stochastic process for dissemination of information across a network. In each round, each node chooses a neighbor and propagates its knowledge. When each node chooses the neighbor at random, the process is called randomized rumor spreading. This has turned out to be very efficient in terms of time and message complexity while keeping robustness to failures [37, 19]. It has also found many applications both in communication networks and social networks, such as updating a database replicated at many sites [15, 37], resource discovery [34], computation of aggregate information [39], multicast via network coding [14], membership services [33], the spread of influence and gossip in social networks [38, 10].

There is also, of course, a deterministic counterpart to this process, quasirandom rumor spreading. In [17], the authors show that it is faster than randomized rumor spreading on some networks. This is important because not only it allows for faster dissemination of information through the network, but it does so using just the local memory at each node (since each node only keeps a list of its neighbors). I want to explore other networks where this happens, as well as other similar algorithms which can benefit from the use of determinism.

2 Diffusions on graphs

I am also interested in various diffusion processes on graphs, and while the projects I have worked on during my graduate studies involve discrete models, I would also like to work on continuous processes further on.

2.1 Optimal control for diffusions

In a project with Yuval Peres, we investigate the following process. Starting from a unit mass on a vertex of a graph, we give bounds on the minimum number of “controlled diffusion” steps needed to transport a constant mass $p$ outside of the ball of radius $n$. In a step of a controlled diffusion process we may select any site with positive mass and topple its mass equally to its neighbors. Our initial motivation comes from the maximum overhang question in one dimension, but the more general case arises from optimal mass transport problems.

Our main result concerns the lattice $\mathbb{Z}^d$:

**Theorem 3** Start with initial unit mass $\delta_0$ at the origin 0 of $\mathbb{Z}^d$, $d \geq 2$, and let $p \in (0,1)$ be constant. The minimum number of toppling moves needed to transport mass $p$ to distance at least $n$ from the origin is

$$N_p (\mathbb{Z}^d, B_n, \delta_0) = \Theta (n^{d+2}) ,$$

where the implied constants depend only on $d$ and $p$.

Furthermore, we also give results on graphs with positive speed/entropy and satisfying Shannon’s theorem, as well as on Galton-Watson trees, the product of trees $T_d \times T_k$, the comb graph, and graphs with bounded degree and exponential decay of Green’s function, an example of which is the lamplighter group.
2.2 Heat diffusion with frozen boundary

I also enjoy problems arising from other fields, like physics or chemistry. Consider “Frozen Random Walk” on $\mathbb{Z}$: $n$ particles start at the origin. At any discrete time, the leftmost and rightmost $\lfloor \frac{n}{4} \rfloor$ particles are “frozen” and do not move. The rest of the particles in the “bulk” independently jump to the left and right uniformly. The goal of this note is to understand the limit of this process under scaling of mass and time. To this end we study the following deterministic mass splitting process: start with mass 1 at the origin. At each step the extreme quarter mass on each side is “frozen”. The remaining “free” mass in the center evolves according to the discrete heat equation. We establish diffusive behavior of this mass evolution and identify the scaling limit under the assumption of its existence. It is natural to expect the limit to be a truncated Gaussian. A naive guess for the truncation point might be the 1/4 quantile points on either side of the origin. We show that this is not the case and it is in fact determined by the evolution of the second moment of the mass distribution.

More specifically, we define Frozen-Boundary Diffusion with parameter $\alpha \in (0,1)$ (or FBD-$\alpha$) as follows. Informally it is a sequence $\mu_t$ of symmetric probability distributions on $\mathbb{Z}$. The sequence has the following recursive definition: given $\mu_t$, the leftmost and rightmost $\lfloor \frac{n}{4} \rfloor$ masses are constrained to not move, and the remaining $1-\alpha$ mass diffuses according to one step of the discrete heat equation to yield $\mu_{t+1}$. In other words, we split the mass at site $x$ equally to its two neighbors. Formal descriptions appear later. We briefly remark that this process is similar to Stefan type problems, which have been studied for example in [?].

Denoting $\tilde{\beta}_t$ as the boundary of the process at time $t$, and $\tilde{\mu}_t$ is the scaled distribution, the main result of this work is in precisely describing the mass distribution and the boundary evolution:

**Theorem 4** Assuming that $\lim_{t \to \infty} \frac{\tilde{\beta}_t}{\sqrt{t}}$ is a constant, the following is true:

$$\tilde{\mu}_t \xrightarrow{\text{weak}} \mu_{\infty}(\alpha),$$

where,

$$\mu_{\infty}(\alpha) = \frac{\alpha}{2} \delta(-q_\alpha) + (1-\alpha)\Phi_{q_\alpha} + \frac{\alpha}{2} \delta(q_\alpha),$$

and $q_\alpha$ is the unique positive number such that:

$$\frac{\alpha}{2} q_\alpha = \frac{(1-\alpha)e^{-q_\alpha^2/2}}{\sqrt{2\pi}\Phi(\{-q_\alpha, q_\alpha\})}. $$

While there are many theoretical interesting questions remaining from the projects above, I am also interested in pursuing models of diffusions on networks arising from the real-world. Other types of diffusions on graphs that I’m interested in exploring more are models of disease spread, contagion on networks, traffic flow, dissemination of information, influence estimations.

3 Stochastic Block Models

Another main part of my dissertation is detecting communities in networks, or more generally, clustering networks. This is a fundamental problem in mathematics, machine learning, biology and economics [6], both for its theoretical foundations as well as for its practical implications. The problem of finding a highly connected subset of vertices in a large complex network arises in a number of applications across science and engineering. Within social network analysis, a highly connected subset of nodes is interpreted as a community [26].

Detecting communities (or clusters) in graphs is a fundamental and long studied problem in computer science and machine learning. The techniques apply to a large variety of complex networks, such as social or biological networks, or to data sets engineered as networks via graphs of similarity. Extracting communities in networks has been used in particular to find like-minded people in social networks [31, 48], to perform recommendations [43], to segment or classify images [52, 53], to detect protein complexes [12, 20], to find genetically related sub-populations [51, 36], to discover new tumor subclasses [54], among other applications.

The study of planted problems in complex systems is furthermore motivated by cryptographic applications: in several works, including one of my own [22], there is a clear connection made between block models and one-way functions or pseudorandom number generators. Showing the impossibility of recovery of the code information-theoretically implies the discovery and provable efficiency of good one-way functions if the constraints are hard to solve [32]. Most importantly, it provides an excellent real-life model of a complex network, since in a network modeling real-life data, the constraints are not random, but rather correlated.
3.1 Bipartite SBM

More formally, the stochastic block model represents a random network where nodes belong to various communities, and the edges between the nodes are drawn independently at random, with different probabilities depending on whether they are inside the same community or not. Presented with only the edges and not the nodes labels, the goal is to recover the communities [35]. For such a model with two communities, I show both information theoretic thresholds when it is impossible to recover the communities based on the density of the edges planted between the communities, as well as thresholds for when it is computationally possible (and efficient, in the sense of the existence of algorithms running in polynomial time) to recover the communities [22]. I also implement these algorithms and show their efficiency for fundamental problems in computer science, such as random planted K-SAT [1].

Fix parameters $\delta \in [0, 2]$, $n_1 \leq n_2$, and $p \in [0, 1/2]$. Then we define the bipartite stochastic block model as follows:

- Take two vertex sets $V_1, V_2$, with $|V_1| = n_1$, $|V_2| = n_2$.
- Assign labels ‘+’ and ‘-’ independently with probability $1/2$ to each vertex in $V_1$ and $V_2$. Let $\sigma \in \{\pm 1\}^{n_1}$ denote the labels of the vertices in $V_1$ and $\tau \in \{\pm 1\}^{n_2}$ denote the labels of $V_2$.
- Add edges independently at random between $V_1$ and $V_2$ as follows: for $u \in V_1, v \in V_2$ with $\sigma(u) = \tau(v)$, add the edge $(u, v)$ with probability $\delta p$; for $\sigma(u) \neq \tau(v)$, add $(u, v)$ with probability $(2 - \delta)p$.

Algorithmic task: Determine the labels of the vertices given the bipartite graph, and do so with an efficient algorithm at the smallest possible edge density $p$.

Our results are

**Theorem 4** Let $\delta \in [0, 2 \setminus \{1\}$ be fixed and $n_2 = \omega(n_1)$. Then there is a polynomial-time algorithm that detects the partition $V_1 = A_1 \cup B_1$ whp if

$$p > \frac{1 + \epsilon}{(\delta - 1)^2 \sqrt{n_1 n_2}}$$

for any fixed $\epsilon > 0$.

**Theorem 5** On the other hand, if $n_2 \geq n_1$ and

$$p \leq \frac{1}{(\delta - 1)^2 \sqrt{n_1 n_2}}$$

then no algorithm can detect the partition whp.

We also investigate the limits and successes of the simple Singular Value Decomposition algorithm, as well as a modification of it, Diagonal Deletion SVD, which performs SVD on the adjacency matrix with diagonal entries removed. Our results locate two different thresholds for spectral algorithms for the bipartite block model: while the usual SVD is only effective with $p = \tilde{\Omega}(n_1^{-2/3} n_2^{-1/3})$, the modified diagonal deletion algorithm is effective already at $p = \tilde{\Omega}(n_1^{-1/2} n_2^{-1/2})$, which is optimal up to logarithmic factors. In particular, when $n_1 = n, n_2 = n^{k - 1}$ for some $k \geq 3$, as in the application above, these thresholds are separated by a polynomial factor in $n$. 
Theorem 6 Let \( n_2 \geq n_1 \log^4 n_1 \), with \( n_1 \to \infty \). Let \( \delta \in [0, 2] \setminus \{1\} \) be fixed with respect to \( n_1, n_2 \). Then there exists a universal constant \( C > 0 \) so that

1. If \( p = C(n_1 n_2)\frac{1}{2} \log n_1 \), then whp the diagonal deletion SVD algorithm recovers the partition \( V_1 = A_1 \cup B_1 \).
2. If \( p = Cn_1^{-2/3} n_2^{-1/3} \log n_1 \), then whp the unmodified SVD algorithm recovers the partition.

3.2 Planted hypergraph colorings

In another work [13], we investigate planted partitions in random hypergraphs, previously studied, for example, in [30]. This model is a prototypical example of a constraint satisfaction problem. Given a set of discrete variables and a set of constraints between these variables, a constraint satisfaction problem consists in deciding if there exists an assignment of the variables satisfying all the constraints. This is a setting that allows the tackling of diverse complex problems such as error-correcting codes [28, 29], register allocation in compilers, genetic regulatory networks [51], financial clusters [4], understanding the origin of average computational hardness, as well as testing new algorithmic ideas [11, 44, 45, 57]. More generally, the class of NP-complete problems, for which no algorithm is known that guarantees to decide satisfiability in a time polynomial in the number of variables, is of immense interest. Well studied examples of such problems are the satisfiability of boolean formulas and the coloring problem: given a graph and a fixed number of colors, can we color the vertices such that no two connected vertices have the same color? This can be translated into many fundamental complex systems problems. There are natural questions one can ask: Can the value of the colorable/uncolorable phase transition be computed? Can the number of all possible colorings be also computed? Are there other interesting phase transitions? Can these transitions explain the fact that solutions are sometimes very hard to find? Can this knowledge help us in designing new algorithms? These questions, and their answers, are at the roots of the interest of the statistical physics community in optimization problems [46, 44].

We consider the problem of coloring the vertices of a large sparse random hypergraph with a given number of colors so that no hyper edge is monochromatic. Using the cavity method, we present a detailed and systematic analytical study of the space of proper colorings (solutions). We show that for a fixed number of colors and as the average vertex degree (number of constraints) increases, the set of solutions undergoes several phase transitions similar to those observed in the mean field theory of glasses.

![Figure 5: The geometry of the set of solutions of CSPs](image)

Finally, we discuss the algorithmic consequences of our findings, and implement belief propagation (the code can be found at [1]). We argue that the onset of computational hardness is not associated with the clustering transition and we suggest instead that the freezing transition might be the relevant phenomenon, in a similar vein as [56]. This corresponds to an onset of hardness in local search algorithms, and is of great importance for complex systems research as it indicates precisely when a problem becomes intractable.

3.3 Future directions

Community detection naturally connects with information theory as it can be viewed as a decoding problem on a noisy channel: the community labels are the input to a black-box channel providing local and noisy interactions of the inputs. This view point was further developed in [3] with the notion of graphical channels, which allow to capture many extensions of the SBM, such as non-overlapping communities, edge-labeled or non-pairwise interactions. This can be further extended to problems such as topic modeling or ranking, with new notions of recovery. Ubiquitous to all these models are two quantities: a measure on how rich the observation graph is (e.g., the node degrees in the stochastic block model), and a measure on how noisy the connectivity channel is (the difference between the various
edge probabilities). These quantities are what make the problems novel and interesting. Through exploring various instances of such models, the hope is to build a general theory of fundamental limits in machine learning and data science problems, inspired by information theory and statistical physics.

There a few questions that I would like to answer further down the line. Most of the theoretical framework already developed for the stochastic block model can be much better translated and extended in order to adapt the models to real-world complex systems. There are a few directions I’d like to further this research in order to bridge gaps between information and computational thresholds for exact recovery for sublinear-sized non-balanced communities. Much is done when the communities are of equal linear-size but a lot more remains to be done beyond that. Similarly, there is hope to partially recover communities, for example in [16]. Other graphical channels, beyond the SBM, are also of interest [2, 3]. While the field of community detection has been expanding greatly since the 80’s, with impressive progress both on the algorithmic and application side, a large part of it has still remained more of an art than a science. In particular, understanding which structures can be discovered, how accurately can they be extracted, are far from being resolved.

As my research and interests are truly interdisciplinary, the ETH Institute for Theoretical Studies would be a perfect place for me both to further my own research agenda, as well as interact and start new projects with the fellows and other researchers at ETH. I thrive and am excited to work in environments in which I can explore new areas and start collaborations. At ETH, I am interested in working with Prof. Andreas Krause in diffusions on networks.

References


