REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
12-07-2017	Final Report	15-Apr-2014 - 14-Apr-2017

4. TITLE AND SUBTITLE
Final Report: Scaling limits in stochastic interacting systems

5a. CONTRACT NUMBER
W911NF-14-1-0179

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Sunder Sethuraman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Arizona
P.O. Box 210158, Rm 510

Tucson, AZ 85721 -0158

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
65389-MA.26

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The proposed project considered the large scale structure of several stochastic models used to understand features of traffic, fluids, social networks, trapping phenomena, data clustering etc. By connecting microscopic behaviors, that is the interactions of individual agents at the `street level', to continuum descriptions, that is a `bird's eye' view of the system, the project identified high-level, macroscopic rules of behavior, and explained how they can be categorized in terms of types of microscopic interactions.

15. SUBJECT TERMS
scaling limits, interacting particle systems, random networks, degree distribution, modularity, clustering, random geometric graph, kelvin tiling, range of random walk, geodesic approximation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sunder Sethuraman
UU	UU	UU	UU		19b. TELEPHONE NUMBER 520-621-1774

Report Title

Final Report: Scaling limits in stochastic interacting systems

ABSTRACT

The proposed project considered the large scale structure of several stochastic models used to understand features of traffic, fluids, social networks, trapping phenomena, data clustering etc. By connecting microscopic behaviors, that is the interactions of individual agents at the `street level', to continuum descriptions, that is a `bird's eye' view of the system, the project identified high-level, macroscopic rules of behavior, and explained how they can be categorized in terms of types of microscopic interactions.

Problems studied included scaling limits of (1) the space-time distribution of mass in a system of interacting particles, (2) the degree sequence in random network growth models, (3) data clustering in random geometric graphs,
(4) collision times of particles under elastic and nonelastic rules, (5) the range of a random walk before exit from a domain, (6) approximations of geodesic paths via random points.

In terms of significance with respect to the Army Research Office (ARO) and broader effects, the project found certain links between areas, highlighted in the solicitation, of `stochastic partial differential equations', `measure-valued stochastic processes', and `interacting particle systems', as well as `random graph/network structures', which helped reveal basic phenomenology, salient to applications. Moreover, aspects of the work advanced interests in other fields, such as physics and data science.

In terms of education, the project has been a source of good questions for participation by PhD/MS graduate (3 PhD, 1 MS) and undergraduate students (1), and postdocs (1). Two PhD students graduated in 2016, 2017 and the undergraduate finished in 2017. 4 of the students won the top scholar award in the years 2016, 2017.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
04/17/2017	21	Joceline Lega, Sunder Sethuraman, Alex Young. On collision times of self-sorting interacting particles in one-dimension with random initial positions and velocities, J. Stat. Phys., ( ): . doi: 1,039,251.00
04/18/2017	22	Sunder Sethuraman, Shankar Venkataramani. On the growth of a superlinear preferential attachment scheme, Springer Proceedings in Mathematics and Statistics: Varadhan 75, ( ): . doi: 1,039,252.00
06/28/2017	10	Erik Davis, Sunder Sethuraman. Consistency of modularity clustering on random geometric graphs, Ann. Appl. Probab., ( ): . doi: 1,015,094.00
08/23/2016	7	Jihyeok Choi, Sunder Sethuraman, Shankar C. Venkataramani. A scaling limit for the degree distribution in sublinear preferential attachment schemes, Random Structures & Algorithms, ( ): 703. doi: 1,015,086.00
08/23/2016	8	Sunder Sethuraman. On Microscopic Derivation of a Fractional Stochastic Burgers Equation, Communications in Mathematical Physics, ( ): 625. doi: 1,015,089.00
08/23/2016	9	Cédric Bernardin, Patrícia Gonçalves, Sunder Sethuraman. Occupation times of long-range exclusion and connections to KPZ class exponents, Probability Theory and Related Fields, ( ): . doi: 1,015,092.00
09/21/2016	15	Sunder Sethuraman, Shankar Venkataramani. On the growth of superlinear preferential attachment schemes, combinatorics, probability and computing, ( ): . doi: 1,018,172.00
09/30/2016	16	Joceline Lega, Sunder Sethuraman, Alex Young. On collisions times of certain interacting particles in one-dimension with random initial positions and velocities, J. Stat. Phys., ( ): . doi: 1,018,173.00
TOTAL:	**8**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Dissemination: There were 7 outside seminars, 5 lectures in conferences, and 1 outside colloquia given on the work in the grant.

AMS Sectional Conference, East Lansing, March 16-20, 2015.
SIAM conference on PDE, Scottsdale Dec 7 - 10, 2015.
Conference in honor of Prof SRS Varadhan, TU Berlin, Aug 15-19, 2016
Asia Pacific Rim IMS conference, CUHK, Hong Kong, June 27-30, 2016
Discrete Geometry and Statistics, Chula. U., Bangkok, Jan 31-Feb 4, 2017.

Indian Statistical Institute Kolkata, India, July 11, 2014.
Indian Institute of Science Bangalore, India, July 15, 2014
U. Tokyo, July 26, 2014.
U. Michigan, Mar 18, 2015.
CIMAT, Guanajuato, MX, Jan. 18, 2016
Indian Statistical Institute Delhi, July 14, 2016
Bangalore Prob Seminar Indian Statistical Institute,July 25, 2016.

Iowa State University, April 21, 2017.

Number of Presentations: 13.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
01/08/2015	3.00	Sunder Sethuraman. On microscopic derivation of a fractional stochastic Burgers equation, COMM. Math. Phys. (09 2014)
08/23/2016	1.00	Jihyeok Choi, Sunder Sethuraman, Shankar Venkataramani. A scaling limit for the degree distribution in sublinear preferential attachment schemes, Random Structures And Algorithms (02 2014)
08/23/2016	5.00	Jihyeok Choi, Sunder Sethuraman, Shankar Venkataramani. A scaling limit for the degree distribution in sublinear preferential attachment schemes, Random Structures Algorithms (07 2015)
08/28/2015	6.00	Cedric Bernardin, Patricia Goncalves, Sunder Sethuraman. Occupation times of long-range exclusion and connections to KPZ class exponents, Probability Theory and Related Fields (08 2015)
09/03/2014	2.00	Cedric Bernardin, Patricia Goncalves, Sunder Sethuraman. Occupation times of long-range exclusion and connections to KPZ class exponents, Probability Theory and Related Fields (07 2014)
TOTAL:		**5**

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Students working on problems in the grant received the following awards:

Erik Davis (PhD in Math, graduated 2016, postdoc in 2017). Bartlett Prize 2016 for top Math student. Next: Conversant Data Science in Chicago.

Alex Young (PhD in Applied Math, graduated 2017). Al Scott Prize 2015 for top AMath student. Next: Postdoc at Duke U.

Doron Shahar (PhD expected 2017). Bartlett Prize 2017 for top Math student.

Thomas Doehrman (UG graduated 2016). Top senior in Math award 2017, Harvill Fellowship 2017. Next: Grad student at U. Arizona

Graduate Students

NAME	PERCENT_SUPPORTED	DISCIPLINE
Doron Shahar	0	
Derick Bishop	0	
Alex Young	0	Mathematics
FTE Equivalent:	**0.00**	
Total Number:	**3**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Erik Davis	1.00
FTE Equivalent:	**1.00**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	DISCIPLINE
Thomas Doehrman	50	
FTE Equivalent:	**0.50**	
Total Number:	**1**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	1.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	1.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	1.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):.....	1.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:.....	1.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME
Alex Young
Erik Davis
Total Number: 2

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Please see the final report attached as a pdf.

Technology Transfer

FINAL REPORT (W911NF1410179): SCALING LIMITS IN STOCHASTIC INTERACTING SYSTEMS

SUNDER SETHURAMAN

1. Overview

The proposed project considered the large scale structure of several stochastic models used to understand features of traffic, fluids, social networks, trapping phenomena, data clustering etc. By connecting microscopic behaviors, that is the interactions of individual agents at the 'street level', to continuum descriptions, that is a 'bird's eye' view of the system, the project identified high-level, macroscopic rules of behavior, and explained how they can be categorized in terms of types of microscopic interactions.

Problems studied included scaling behaviors of the following objects, fundamental in diverse real-world settings.

(1) The space-time distribution of mass in systems of interacting particles.
Understanding how 'traffic' in an interacting system evolves is crucial in applications, for instance.

(2) The degree sequence, or counts of nodes with varying numbers of connections in random network growth models.
Capturing the structure of social media, modeled by random network growth processes, is of interest.

(3) Data clustering in random geometric graphs.
It is of interest to establish benchmarks on naturally generated data sets, such as random point clouds, for popular clustering methods on which there is little theoretical foundation.

(4) Final times of collision in a system of particles.
In systems of particles moving by Newtonian laws, for instance, it is of interest to understand, in terms of the system size, the scales of the final times of collision.

(5) The range or number of locations visited by a random walk up to the time of exit from a domain,
It is a basic concern to understand the extent of visitation of an individual before exit from a region.

(6) Polygonal approximations of geodesic or shortest paths in spatially varying fields via random points.
For instance, in moving optimally between two points through a 'minefield' restricted to certain roads, it of interest to understand how close such a path is to the one without restriction.

Key words and phrases. scaling limits, interacting particle systems, random networks, degree distribution, modularity, clustering, random geometric graph, kelvin tiling, range of random walk, geodesic approximation.

In terms of significance with respect to the Army Research Office (ARO) and broader effects, the project found several innovative connections between the different mathematical subareas, highlighted in the solicitation, of 'stochastic partial differential equations', 'measure-valued stochastic processes', and 'interacting particle systems', as well as 'random graph/network structures', which helped reveal basic phenomenology, salient to applications. Moreover, aspects of the work advanced interests in other fields, such as physics and data science.

In terms of education, the project has been a source of good questions for participation and training of PhD/MS graduate (3 PhD, 1 MS) and undergraduate students (1), and postdocs (1), who look forward to entering the STEM workforce after graduating. Two PhD students graduated in 2016, 2017 and the undergraduate finished in 2017. 4 of the students won the top scholar award in their respective programs in the years 2016, 2017.

From a more mathematical view, the challenge in the proposal has been to understand how 'averaging' occurs in different contexts, that is the 'coarsening' process which allows to approximate noisy/rough 'on the street' behaviors by 'bird's eye' large scale limit laws, which govern the essential features. Since many of the problems studied, although stochastic in nature, could be recast in terms of other disciplines, such as dynamical systems and geometry, there was the opportunity for cross-fertilization of ideas, which allowed to develop new techniques. Part of the significance of the results is that they often make conclusions in other disciplines in unexpected ways.

The report is organized as follows. In Section 2, in 6 subsections, the problems above are discussed in more detail. In Section 3, the education aspect of the project is discussed. In Section 4, dissemination in terms of talks is listed. In Section 5, the papers/manuscripts produced is listed. References are collected at the end.

2. Specific projects

We discuss now specific problems, based on 9 papers/manuscripts listed in Section 5.

2.1. **Scaling limits in interacting particle systems, which model traffic, queues, fluids etc.**

Given the interest in traffic of various sorts, queues, fluids, etc., it is a basic concern in applications to understand how the the mass in a system of interacting components (particles) evolves in space and time. For instance, one may like to know what fraction of the system would be in a certain region at a certain time.

The first-order behavior of the mass, that is the 'law of large numbers' (or 'hydrodynamics' by another name), and the 'fluctuations, that is the second-order behavior or error made in the first-order estimate, in models where the particles interactions are local in short time-windows is well known since the '90s. These 'hydrodynamic limits' and 'fluctuations show that the space-time mass density solves a partial differential equation with parameters reflecting the local individual particle interactions, and that associated errors are understood in terms of a stochastic partial differential equation. See [1] for more discussion.

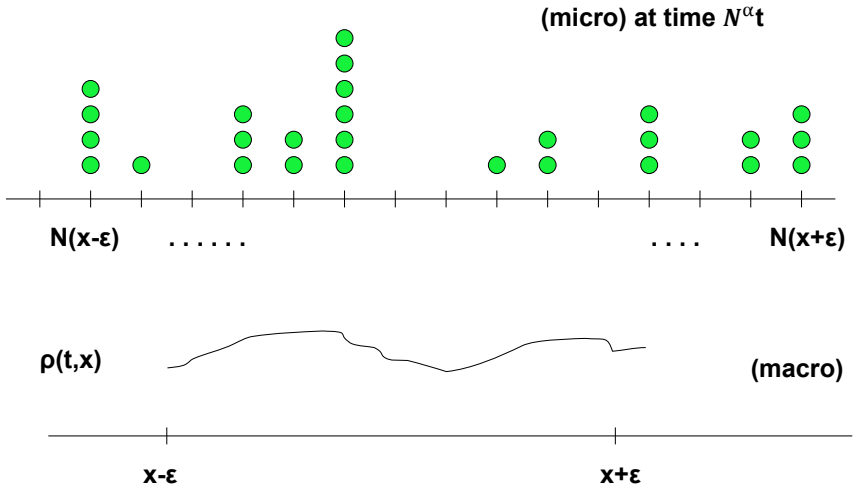
However, hydrodynamics and fluctuations are not well understood when individual particles may interact with others very far away in short time-windows, that is when there are 'long-range interactions'. Such systems however find use in the

modeling of various applied problems, such as with respect to communication, temperature distribution in nuclear reactors for instance; see [2] and references therein. Here, information or particles may displace in one jump long distances at some rate, depending on the configuration of the other particles.

Our results, described below, find novel equations and behaviors depending on the scope of the long-range interactions. Such knowledge is crucial in applied modeling, and was hitherto unknown before.

More technically speaking, we consider systems acting on the d-dimensional grid $\mathbb{Z}^d$, where a particle displaces by x with rate of the order of $c(\eta)|x|^{-d+\alpha}$; here, $\alpha > 0$ is a parameter reflecting the strength of the jump, and $c(\eta)$ reflects a particle interaction term. That is, when α is small, long jumps happen more often than when α is large.

Examples of interactions studied include 'exclusion', 'zero-range' and 'misanthrope' processes, which follow particles on grids $\mathbb{Z}^d$. In the exclusion process, the interaction is spatially given in that particles can only jump to unoccupied locations. In zero-range processes, the interaction is temporal in that the time of jump of a particle depends on the number of particles nearby it. Misanthrope processes combine exclusion and zero-range features in that the interaction is both spatial and temporal. See [3], [4] and [5] for more discussion of these models.



The figure depicts a cartoon, for $1D$ interacting particle systems, of the relation between microscopic and macroscopic scales with respect to the space-time mass density $\rho(t, x)$. Microscopically, one is looking at particles in the system moving between grid points $N(x - \epsilon)$ and $N(x + \epsilon)$ in the rapid time scale $N^\alpha t$. In the macroscopic view, a bird would see the system smeared out in a coarse-grained sense where the flow of mass density between locations $x - \epsilon$ and $x + \epsilon$ is on a slower time scale t. The function $\rho(t, x)$ represents the smoothed/averaged density of particles in this view.

(A) In [Paper 8], for systems where the particles jump asymmetrically, that is when jumps are allowed only in certain positive directions, we derive, depending

on the strength of the long-range interactions, two types of equations for the law of large numbers or hydrodynamics. In the 'weak regime, that is when $\alpha \geq 1$, a nonlinear equation for the mass density is recovered similar to that known in the local interactions setting. However in the 'strong case, when $\alpha < 1$, a new nonlinear fractional equation, under an anomalous scaling, is found. This is part of the PhD work of Doron Shahar.

(B) In [Paper 1], with respect to certain initial configurations and jump rate asymmetries, we derived new stochastic partial differential equations governing the macroscopic fluctuation field, associated to second-order errors made in the law of large numbers approximation of the mass density by its mean. Again, depending on the strength of the long-range interactions, the type of equation found differs. In particular, interestingly, this dichotomy is not the same as for the hydrodynamics.

When the interactions are strong, that is $\alpha < 3/2$, the equation is a stochastic fractional heat equation. However, when the interactions are weak, that is $\alpha \geq 3/2$, the equation is a type of Kardar-Parisi-Zhang Burgers stochastic partial differential equation, on which there has been much recent activity, and which has scaling behaviors different than the equation when the interactions are strong. Moreover, we note that this dichotomy was unexpected as the role of '3/2' was not understood even in the physics literature.

(C) Instead of the mass density, which corresponds to the 'bulk in the system, one is also interested in fluctuations of the time that a single location is occupied. For instance, what are the statistics of when a cafe is occupied? In [Paper 2], we derived the fluctuation scales and limits in models with all types of long range interactions. In fact, this was the first paper suggesting that anomalous behaviors were possible depending on the strength of the long range interactions, and was a precursor to the discussion in the previous paragraphs.

Technical Methods. To handle, in particular, the fluctuations work in [Paper 2], we developed an analytical machinery, dubbed 'H_{-1} norm analysis, which can estimate the level of volatility or mixing in the system, the main technical challenge. In [Paper 1], we formulated a 'martingale problem' to identify the nonlinear, nonlocal limit of the long-range fluctuation fields, which is of interest itself, given that such characterizations are in their infancy in comparision with martingale problems for linear, local evolutions. In [Paper 8], several estimates, reflecting the long-range character of the model, needed to be developed, which should be of use in other problems.

2.2. **Limits of the degree structure in random graphs which model social networks and other 'real world' systems.** In a social network, individuals with a large number of 'friends', since their reach is more, have a larger chance to gain more 'friends' than individuals with few 'friends'. One of the basic models of this phenomenon is as follows: Start with a small network, and grow the network by adding vertices and/or links, one at a time, to locations in the network selected with chance according to their connectivity. The form of the selection chance depends on the application, but in social networks it is proportional to an increasing function of the connectivity–so already well connected individuals tend to become even more well connected, the reinforcement phenomenon mentioned earlier. See [6] and [7] in this context.

In these models, sometimes called 'preferential attachment' models, it is of interest to characterize the growth of the count of nodes of degree $1, 2, \ldots$, or so to speak the 'degree distribution' of the network. For instance, how many individuals have only one 'friend', two 'friends', or 100 'friends'? From an applied perspective, given a social or a 'real world' network, one can compute the empirical degree distribution as a histogram–so many individuals have 1 connection, so many have 2 connections, etc. Then, if one knew the degree distribution for a class of models, one could fit one of these theoretical models to the data, based on the empirical degree distribution, justifying a virtual model of the applied setting for more detailed analysis.

Previously, degree distribution limits have been found when the selection is done linearly, that is when a node with k connections is chosen with chance proportional to k. However, there is little work on models with nonlinear selection functions, that is say when a node with k connections is chosen with chance proportional to a nonlinear function of k such as k^β where β is an arbitrary exponent. Such nonlinear models are quite natural, and important to characterize for applications.

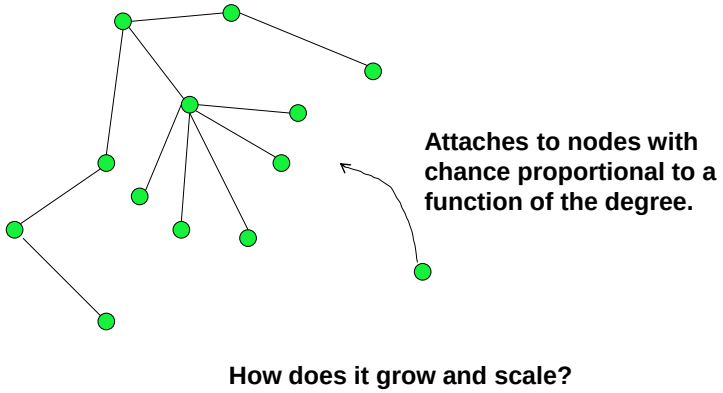
Below, our results describe in detail the law of large numbers and fluctuation behaviors of the degree distribution in a large class of nonlinear models. It turns out there is a dichotomy depending on when the nonlinearity is 'sublinear' or 'superlinear'. In sublinear models, the degree counts depend on each other in an interesting way and are all of the same proportionate size. While in superlinear models, there is a certain 'explosion', and growth behavior, different than in sublinear models, which shows the counts have varying asymptotics.

(D) For a general class of schemes, when the selection function is sublinear, that is say when a node with k connections is selected with chance proportional to k^β where $\beta < 1$, in [Paper 3], we derived a system of coupled ordinary differential equations, governing the limiting degree distribution. Here, interestingly, every count of nodes with k connections depends on every other count, unlike in the linear case where the dependency is more restricted. This work is robust and does not rely of specific features of the evolution.

(E) When the selection function is superlinear, that is when the selection chance of a node with k connections is proportional to k^β for $\beta > 1$, it is known that the preference given to highly connected nodes is so strong that the network condenses, in that after some time one (random) vertex gobbles up most of the incoming connections, obtaining an infinite number of links in the limit, whereas all other nodes have bounded degree. However, the specifics of how the graph converges to such a limit was not known.

In [Paper 5], we derived the precise orders of growth of the counts of nodes with degrees $1, 2, \ldots$, and also show how these quantities fluctuate. In particular, the count of the 'leaves', the count of nodes of degree 1, is of order N at time N, as might be expected, as they mostly connect to the condensing vertex. However, there are not so many nodes of degree 2 and higher, whose counts are sublinear in N, and depend on the power β.

The figure shows a cartoon of a growing preferential attachment graph where an incoming node seeks to attach to an existing node with probability proportional to a superlinear function of its degree. After some lead in time, one of the nodes will accumulate a large number of connections, and this node will more likely be picked at all later times, leading to its explosion.



Technical Methods. In [Paper 3] and [Paper 4], we introduced martingale and dynamical system analysis, allowing to treat robustly a large class of models. Most of the previous work on these problems has been combinatorial or from a branching process view which applies to a limited type of model. However, the martingale approach used seems flexible, allowing vistas into the growth process, capturing even lower order terms in growths and fluctuations, the first such results in nonlinear selection models.

2.3. **Data clustering and connections to geometric tilings.** Clustering data into groups of similar points is an old but still quite a relevant problem with myriad important applications. For instance, in a large collection of images of people, one may like to organize the collection into groups where each group consists of images of the same person. See [8] for a taste of this literature.

In the last 10 years, new methods have emerged which involve treating the data as nodes of a graph where links between vertices, any two images say, are assigned weights depending on how similar they are. A popular method is then to optimize a 'modularity' functional of a partition, or grouping, of the data. This functional favors partitions which are 'well-connectedness' inside sets in the partition, and 'lack of connectedness' across sets in the partition. The optimizer, which can be found automatically from a computer program, is declared the best 'clustering'. See [9], [10].

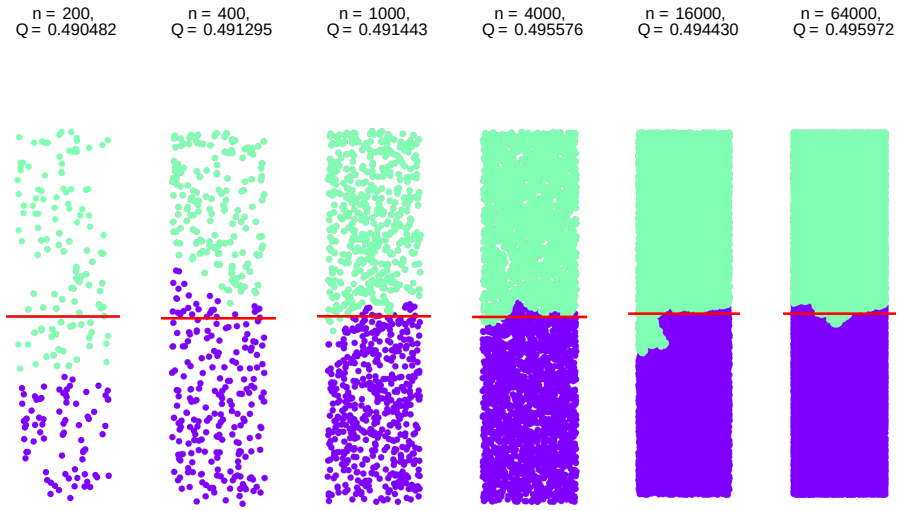
However, it is not known what to expect from the 'modularity' or other clustering procedures, even with respect to basic, known data sets. Our results, in a nutshell, establish as a useful benchmark what modularity clustering optimally selects when the data is given geometrically as a point cloud in the space $\mathbb{R}^d$. Interestingly, the answer connects to classical and still relevant questions in tiling geometry, which was unexpected.

(F) In [Paper 4], we show that modularity clustering is 'consistent', and derive a continuum functional associated to it, relating to a certain tiling optimization problem. That is, suppose the data is given as N independent samples, uniformly drawn from a region of space, and that two points are connected only if they are within a small distance ϵ of each other. One would like to know how the 'modularity' method behaves on such 'known' data to develop benchmarks. Colloquially

speaking, if the points are locations of people, what are the 'county lines' drawn by the modularity procedure?

In an appropriate scale $\epsilon = \epsilon_N \downarrow 0$, we show the best modularity clustering, as the number N of data points grows, converges to a partition of the underlying space where each partition set has the same volume, but that the boundary perimeter of the sets are minimized.

Such geometric tiling problems go back to Lord Kelvin in the 1800's (see [11] for a survey). The answer in dimension $d = 2$ is a type of 'honeycomb' tiling, but the answer is not explicit in $d \geq 3$! It was unexpected to make such a geometric connection in this statistical problem. Interestingly, from the viewpoint of geometry, this work allows to approximate such optimal partitions, which is difficult to do directly from the continuum problem numerically. This was part of the PhD work of Erik Davis.



The figure shows a run of the modularity clustering algorithm to separate the data into two clusters as the number of data points grows with $\epsilon_n = n^{-0.3}$. The modularity functional Q has a maximum of $1/2$ when used to separate the data into two clusters. The continuum partioning problem is to separate the domain into two pieces of equal area which minimizes the length of the boundary between them, which in this case is to cut the slender rectangle in in the middle of the long side. As seen, the data points do indeed segregate into two clusters approaching the optimal cut, with optimal Q values nearing $1/2$.

Technical Methods. One of the main vehicles used is 'Gamma convergence', a technique from analysis, however less known in the probability community, which allows to identify a limit optimization problem from a sequence of optimizations. Importantly, we introduce a probabilistic form of Gamma convergence, which may be of use in other problems.

2.4. **Final collision times in a 1D model of particle interactions.** By Newton's laws, the movement of a collection of particles on d-dimensional space $\mathbb{R}^d$, each with unit mass, can be modeled as follows. Between collisions, the particles move by free flight according to their positions and velocities. When they collide, some part of their velocities are exchanged. 'Kinetic theory' is devoted to the study of such systems, which has much application in material science and other domains (see [12]). For instance, one can think of the system as a group of messengers in space, exchanging information when they meet.

However, to our knowledge, if there are N particles in the system, the dependence on the system size N, of the times of final collision of a given particle t_N and that in the whole system T_N, has not been studied. In the messenger example, when is the last time of exchange of information? We remark this seems to be a new and important problem, given the relevance to know when the system enters a steady-state solution.

Our results capture the scale of t_N and T_N, in $d = 1$, when the particles begin with random initial positions and velocities (the only stochasticity in the system), and the collision rule is either elastic, that is when the particles exchange velocities, or nonelastic when the exchange is only partial. Beyond kinetic theory, this problem connects with 'sorting' of velocities, which after a moment's thought seems reasonable, but is interesting nonetheless.

(G) In [Paper 6], we derive the scaling limits for these times in terms of N, for a class of interactions. Comprehensive results are found when the collisions are elastic, that is when the particles exchange velocities. Detailed numerics are given in nonelastic scenarios. In both situations, the results connect with 'sorting: Eventually, for each fixed N, after the final collision time T_N, the particles are sorted according to their velocities, the particles at the end with largest velocities being to the far right, and those with the least velocities to the far left.

In the elastic case, the limit laws, in scale N for t_N and scale N^2 for T_N, are mixtures of Frechet type distributions, which arise in order statistics theory, depending on features of the initial position and velocity distributions. The scales reflect that an individual particle crosses order N other particles, and that there are order N^2 collisions in the system.

On the other hand, in the nonelastic case, velocities exchanged are amped or damped, that is when two particles meet with velocities v_1 and v_2, and particle one now takes on velocity $(1-\epsilon)v_2$ and particle two takes on velocity $(1-\epsilon)v_1$, for $\epsilon < 1$. We show the associated scalings of the final times depend on the strength of this amping/damping ϵ and involve exponential (!) corrections to the elastic scales, something further to explore. This was part of the PhD work of Alex Young.

The figure below shows how two particles in one dimension interact by elastic collision. When they meet, they simply exchange their velocities, moving however on the lines depicted. One can infer that all collision times correspond to the times the N lines intersect.

Technical Methods. As mentioned, in the case of elastic collisions, the collision times of the particles can be understood in terms of the intersection times of lines with random y-intercepts (initial positions) and random slopes (initial velocities). Then, the problem interestingly may be put in the framework of order statistics of exchangeable random variables, which allowed some calculation. The numerical

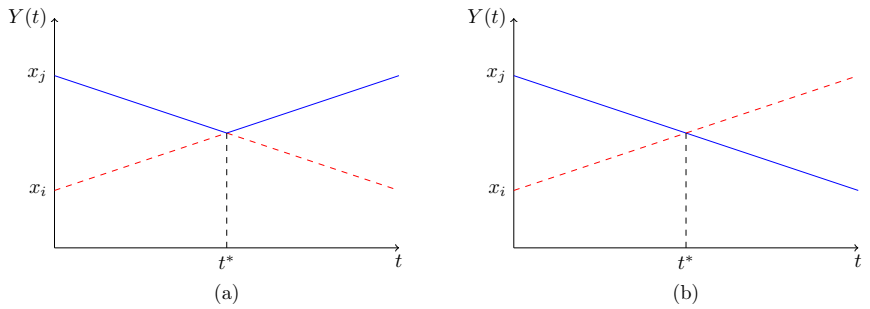


Figure 1: (a) Trajectories of particles i (in red) and j (in blue) before and after their collision. The elastic collision results in an exchange of velocities at t^*, the time of the collision. (b) Paths of particles i (in red) and j (in blue) if they do not interact. Paths intersect at time t^*.

work, for the nonelastic collisions, was a full molecular simulation, as in this case no memory is lost in the system.

2.5. **The range of random walk in a domain.** In applications, say with chemical traces, or tracking of individuals, the number of locations visited up to a certain time, the 'range' of the trace or individual, is an important variable. One can abstract in the following way: Consider a random walk on the lattice $\mathbb{Z}^d$ which moves to a nearest-neighbor grid point at each step with equal chance. Denote the number of nodes that the random walk visits in the first N steps, the range of random walk, by R_N.

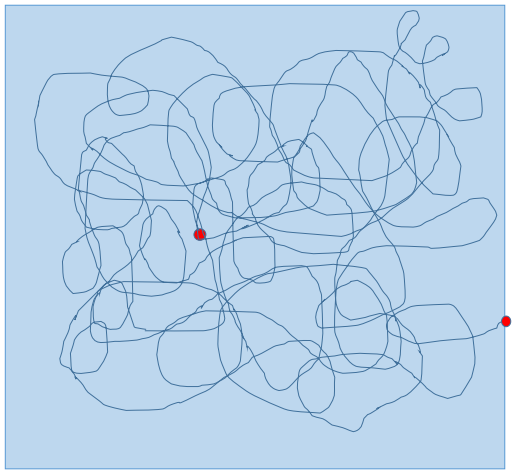
Although the behavior of the range, for unrestricted walks, is a well-studied problem [13], less is known when the walk is restricted in a certain sense, although such walks would seem relevant for applied trapping concerns. For instance, what is the range of the walk before it exits a cube in $\mathbb{Z}^d$ with side length N?

The answers depend on the dimension d, and also importantly the starting point of the random walk. In dimension $d = 1$, it is known that R_N scales as N, and special properties help solve this problem [14]. However, in dimension $d = 2$, as the walker can loop around a point, the problem is harder, and importantly of a different character, and as we see in our results connects with Brownian motion in an interesting way.

(H) In [Paper 7], we show that the mean behavior of R_N, starting from a point near (aN, bN), where $0 < a, b < 1$, scales as $c(a, b)N^2/\log(N)$ where $c(a, b) = const.E_{a,b}[T]$, $E_{a,b}[T]$ is the expected time of exit of a Brownian motion from a cube of side length 1 starting from (a, b), and *const.* is an explicit constant. The form of the coefficient $c(a, b)$ was unexpected! The next step in this problem, which we are pursuing presently, is push to higher dimensions, and an almost sure result. This was part of the senior thesis of Thomas Doehrmann.

The figure below shows a cartoon of a random walker, starting at the red dot in the square with side length N, moving inside until exiting at the right. The range R_N would be the number of grid points visited before exit, that is the number of pixels shaded by the path of the walk, counting each pixel crossed not more than once, even if it is crossed several times.

Technical Methods. As mentioned, in dimensions larger than 1, one cannot use ordering of the underlying space to help estimates. However, in [Paper 7], we have



used discrete harmonic analysis to obtain the scaling of the mean of R_N, and the dominant prefactor. The techniques border on potential theory, and we feel there is much more to do on this and related problems.

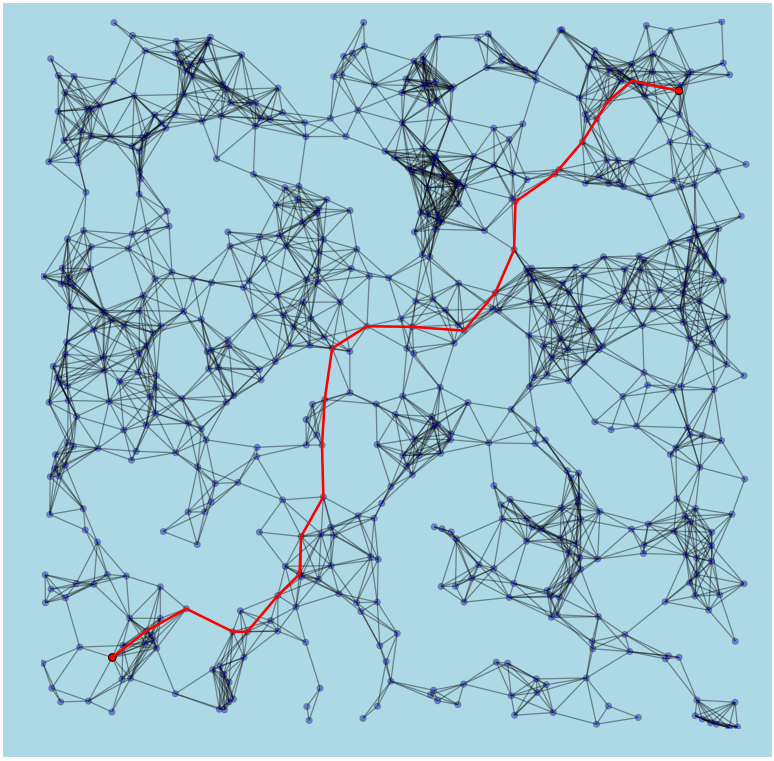
2.6. **Approximating geodesics.** Given a distance function on a space with an underlying metric, one can define the shortest path between points a and b, namely the geodesic path (see [15] for a technical definition). It is of interest to study approximations of this distance and the associated geodesics. For instance, one may like to compute the optimal way to navigate a 'minefield' using existing trails, and to know if the trails are dense enough so that this path is close to the geodesic path computed without restrictions.

We now abstract the problem in the setting of a random point cloud. In a space $\mathbb{R}^d$, consider N points drawn at random. We will say that two points are connected if they are within ϵ Euclidean distance of each other. If we insist that paths from a to b must be formed by linear interpolations through connected points, the question above is rephrased as follows: How close are these shortest paths to the underlying geodesic? Although such questions are natural, there is little work on them.

Previously, for some type of underlying metrics, the distance of the interpolating path has been shown to converge to the distance of the geodesic as N grows [16]. However, the convergence of the actual interpolating paths to the geodesic path has not been treated, although it would seem to be a natural concern.

We note part of our results is to treat for the first time convergence of these paths in various senses. Interestingly, some of the methods involve results on Hilbert's 4th problem, stated famously in 1900, which asks for what underlying metrics is the geodesic between points a and b a straight line [17].

(I) In [Paper 9], we show for a large class of underlying metric, as $\epsilon = \epsilon_N$ vanishes, that shortest interpolations converge in various senses, both in uniform and Hausdorff norms, to the geodesic with probability 1. Convergence of the associated distances is also immediate from this result. This was part of the postdoc project of Erik Davis.



The figure shows the shortest path, through $N = 400$ points sprinkled on a square with side length 2 with $\epsilon_n = n^{-0.3}$, with the usual underlying Euclidean metric, which approximates the straight line continuum geodesic path between the extreme red points.

Technical Methods. We use a form of probabilistic Gamma convergence, mentioned earlier to show convergence of the optimal interpolation to the geodesic. Several detailed geometric and probabilistic estimates on the structure of the point cloud, including as mentioned, results on Hilbert's 4th problem, may be of interest themselves.

3. Education

The project has been a good source of problems for students who I have been fortunate to work with (3 PhD, 1 MS, 1 Undergrad) and postdocs (1 Postdoc). Two PhD students graduated in 2016, 2017, and the undergraduate finished in 2017. 4 of the students have won the top scholar award in the years 2016, 2017.

- Erik Davis (PhD in Math, graduated 2016, postdoc in 2017). Bartlett Prize 2016 for top Math student. Next: Conversant Data Science in Chicago.
- Alex Young (PhD in Applied Math, graduated 2017). Al Scott Prize 2015 for top AMath student. Next: Postdoc at Duke U.
- Doron Shahar (PhD expected 2017). Bartlett Prize 2017 for top Math student.
- Thomas Doehrman (UG graduated 2016). Top senior in Math award 2017, Harvill Fellowship 2017. Next: Grad student at U. Arizona
- Derick Bishop (MS expected 2018).

4. Dissemination

In total there were 13 invited talks given at various conferences and universities on the work in the grant.

There were 7 outside seminars:

- AMS Sectional Conference, East Lansing, March 16-20, 2015.
- SIAM conference on PDE, Scottsdale Dec 7 - 10, 2015.
- Conference in honor of Prof SRS Varadhan, TU Berlin, August 15-19, 2016
- Asia Pacific Rim IMS conference, CUHK, Hong Kong, June 27-30, 2016
- Discrete Geometry and Statistics, Chulalongkorn U., Bangkok, Jan 31- Feb 4, 2017.

There were 5 lectures in conferences:

- Indian Statistical Institute Kolkata, India, July 11, 2014.
- Indian Institute of Science (two talks) Bangalore, India, July 15, 2014
- U. Tokyo, July 26, 2014.
- U. Michigan, Mar 18, 2015.
- CIMAT, Guanajuato, MX, Jan. 18, 2016
- Indian Statistical Institute Delhi, July 14, 2016
- Bangalore Probability Seminar Indian Statistical Institute,July 25, 2016.

There was 1 outside colloquium:

- Iowa State University, April 21, 2017.

5. Papers/Manuscripts

There were 9 papers/manuscripts written. 4 were published/accepted (40, 64, 28, 66 pages) in the top journals in probability, combinatorics, and mathematical physics. 2 papers are submitted, and available on the arXiv. 2 other papers are close to final form. 1 more, although a part of it is in final form as a thesis, is also nearing completion as a more general paper.

[Paper 1] S. Sethuraman, On microscopic derivation of a fractional stochastic Burgers equation (2016) Commun. Math. Phys. 341, 625–665.

[Paper 2] C. Bernardin, P. Goncalves, S. Sethuraman, Occupation times of long-range exclusion and connections to KPZ class exponents. (2016) Prob. Theory Rel. Fields. 166, 365–428.

[Paper 3] J. Choi, S. Sethuraman, S. Venkataramani, A scaling limit for the degree distribution in sublinear preferential attachment schemes. (2016) Random Structures and Algorithms, 48, 703731.

[Paper 4] E. Davis, S. Sethuraman, Consistency of modularity clustering on random geometric graphs. (2016) To appear in Ann. Appl. Probab. (66 pgs). Available at arXiv: 1604.03993

[Paper 5] S. Sethuraman, S. Venkataramani, On the asymptotic growth of superlinear preferential attachment random graphs. (2017) Submitted to Springer volume in honor of Prof. SRS Varadhan. Available at arXiv: 1704.05568

[Paper 6] J. Lega, S. Sethuraman, A. Young, On collisions times of self-sorting interacting particles in one-dimension with random initial positions and velocities. (2017) Submitted to J. Stat. Phys. Available at arXiv: 1704.01251

[Paper 7] T. Doehrman, S. Sethuraman, S. Venkataramani, The range of random walk up to the time of exit from a domain. (2017) Part is the senior honors thesis of T. Doehrman. We are working on a more involved version. Available at the ARO technical reports site.

[Paper 8] S. Sethuraman, D. Shahar, Hydrodynamics for long-range asymmetric particle systems. (2017) Manuscript, all results there, to be polished/finished Summer 2017. See ARO technical reports site.

[Paper 9] Approximating geodesics via random points. (2017) Manuscript, all results there, to be polished/finished Summer 2017.

References

[1] Kipnis, C., Landim, C. (1999). *Scaling Limits of Interacting Particle Systems* Springer-Verlag, Berlin.

[2] Sornete, D. (2006). *Critical Phenomena in Natural Sciences* Springer Series in Synergetics, Springer

[3] Andjel, E. (1982). Invariant measures for the zero range process. *Ann. Prob.* **10**, 525–547.

[4] Cocozza, C.T. (1985). Processus des misanthropes. *Z. Wahr. Verw. Gebiete* **70**, 509–523.

[5] Liggett, T. M. (1985). *Interacting particle systems* Springer-Verlag, New York.

[6] Albert, R., Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74**, 47–97.

[7] Durrett, R. (2007). *Random Graph Dynamics* Campridge U. Press

[8] He, X., Yan, S., Niyogi, P., Zhang, H. (2005). Face Recognition Using Laplacianfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 328–340.

[9] Porter, M., Onnela, J.P., Mucha, P. (2009). Communities in networks. *Notices of the Amer. Math. Soc.* **56**, 1082–1097.

[10] Newman, M., Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113

[11] Morgan, F. (2008). *Geometric Measure Theory: A Beginner's Guide* Academic Press, New York

[12] Bellomo, N., Pulvirenti, M. (2000). *Modeling in Applied Sciences: A Kinetic Theory Approach* Birkhauser

[13] (1972). Jain, N., Pruitt, W. The range of random walk. in *Proc. Sixth Berkeley Symposium on Math. Stat. and Probab.* **3**, 31–50. University of California Press, Berkeley, Calif.

[14] Athreya, S., Sethuraman, S., Toth, B. (2011). On the range, local times, and periodicity of a random walk in an interval. *ALEA Lat. Am. J. Prob. Stat.* **8**, 269–284.

[15] (2000). Bao, D., Chern, S.S., and Shen, Z. *An Introduction to RiemannFinsler Geometry* Springer-Verlag

[16] Alamgir, M., von Luxburg, U. (2012). Shortest path distance in random k-nearest neighbor graphs. in *Proc. 29th Int. Conf. on Machine Learning*; arXiv:1206.6381

[17] Papadopoulos, A. (2013). On Hilbert's fourth problem. in *Handbook of Hilbert geometry* Ed. Papadopoulos, A. and Troyanov, M., European Mathematical Society. arXiv:1312.3172

Department of Mathematics, University of Arizona, Tucson, AZ 85721
e-mail: sethuram@math.arizona.edu