MATHEMATICAL FRONTIERS

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/MathFrontiers

Board on Mathematical Sciences & Analytics

MATHEMATICAL FRONTIERS
2018 Monthly Webinar Series, 2-3pm ET

February 13: ***Recording posted***
Mathematics of the Electric Grid

March 13:
Probability for People and Places

April 10:
Social and Biological Networks

May 8:
Mathematics of Redistricting

June 12:
Number Theory: The Riemann Hypothesis

July 10: *Topology*

August 14:
Algorithms for Threat Detection

September 11:
Mathematical Analysis

October 9: *Combinatorics*

November 13:
Why Machine Learning Works

December 11:
Mathematics of Epidemics

View webinar videos and learn more about BMSA at www.nas.edu/MathFrontiers

MATHEMATICAL FRONTIERS
Probability for People and Places

Gregory F. Lawler,
University of Chicago

Kenneth L. Lange,
UCLA

Elizabeth A. Thompson,
University of Washington

View webinar videos and learn more about BMSA at www.nas.edu/MathFrontiers

MATHEMATICAL FRONTIERS
Probability for People and Places

Gregory F. Lawler,
University of Chicago

George Wells Beadle Distinguished Professor in Chicago in departments of mathematics and statistics

Random fractal curves

Applying mathematics to the "real world"

- Observe some phenomenon
- Make a mathematical model. This should be precise but generally includes some simplification.
- Make conclusion from the mathematical model using
  - Rigorous mathematical theorems
  - Heuristic (nonrigorous) arguments
  - Computer calculation and in case of random models, Monte Carlo simulations
- Check mathematical predictions with real world
- If wrong, either model is wrong or mathematical deductions were incorrect
- Users of mathematics should always understand the assumptions in mathematical models even if they do not understand the mathematical derivations.

Probability

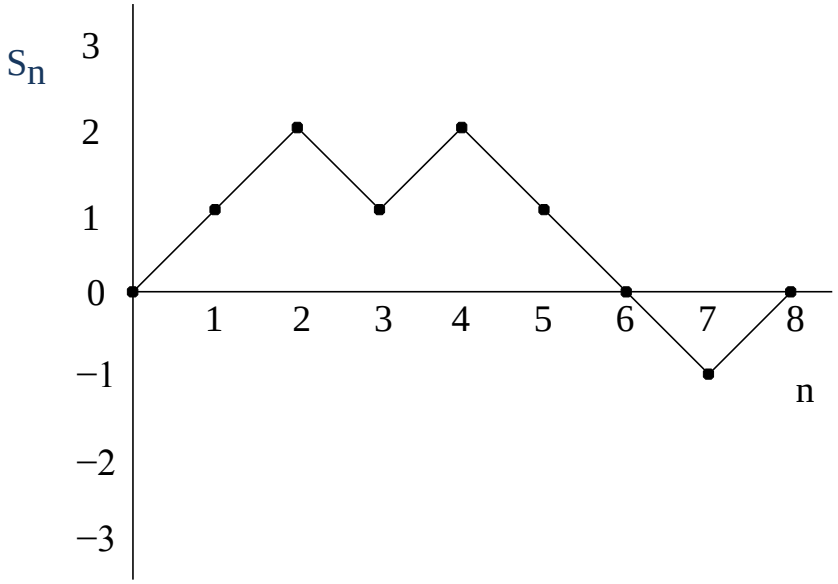
- Probability is the area of mathematics that incorporates randomness in the mathematical model.
- This randomness might indicate some "real" randomness or just lack of total information about the behavior
- Statistics is the inverse of probability — view the real word and try to determine the appropriate mathematical model and the parameters.
- Flip a fair coin, what is the probability that one gets 18 heads out of 20? (probability problem)
- Flip a coin, get 18 heads out of 20, and then ask: is the coin fair? (statistics problem)
- Cannot do statistics without good background in probability.

Random walk / Brownian motion
(Completely) random continuous motion

Basic model for random motion: a random walker takes independent random steps at each time unit.

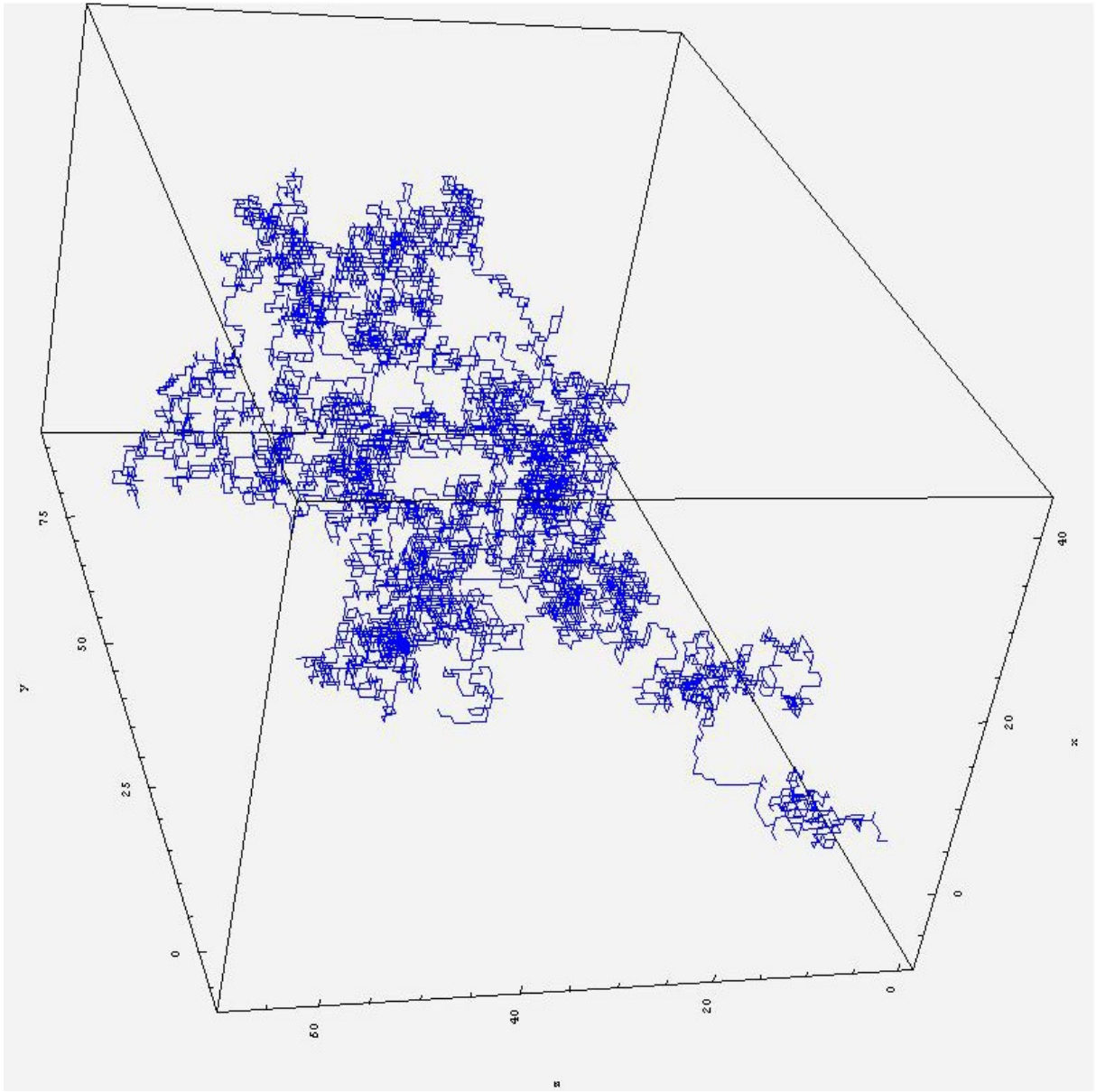
S_n = position after n steps

$$S_n = X_1 + \cdots + X_n$$



$$P\{X_j = +1\} = P\{X_j = -1\} = 1/2$$

- Brownian motion or Wiener process is the continuous analog of this.
- In a small time increment Δt one chooses moves a distance (either positive or negative) of size $(\Delta t)^{1/2}$.
- For small Δt, $(\Delta t)^{1/2} \gg \Delta t$. However, since there are both positive and negative jumps there is a lot of cancellation.
- From this we see that Brownian paths are nowhere differentiable.
- This can be done for random motion in any number of dimensions.
- The principle $\Delta W_t \approx (\Delta t)^{1/2}$ is always true for completely random, continuous motion.


y
75
50
25
0
z
60
40
20
0
x
40
20
0

Fractal dimension of random paths

- What is the fractal dimension α of a random walk or Brownian path?
- For random walk, we define this roughly by: in the ball of radius N the random walker visits N^{α} points.
- Equivalently, the number of steps needed to reach distance N is N^{α}.
- For random walk, $\alpha = 2$: this follows from the basic scaling rule for random walk or Brownian motion,

$$\Delta W_t \approx (\Delta t)^{1/2}.$$

Usual calculus and differential equations

$$\frac{dy(t)}{dt} = F(t, y(t)),$$

$$dy(t) = F(t, y(t))\, dt.$$

At time t, the process moves $F(t, y(t))\,\Delta t$ in time Δt.

Stochastic calculus

$$dX_t = m(X_t)\,dt + \sigma(X_t)\,dW_t$$

where W_t is a Brownian motion.

- From time t to time $t + \Delta t$, X_t moves $m(X_t)\Delta t$ plus a random jump with mean 0 and standard deviation $\sigma(X_t)(\Delta t)^{1/2}$
- This is basic mathematics of diffusion used in physics, chemistry, biological sciences
- This is also the main tool for mathematical finance such as the Black-Scholes formula for pricing options.
- The "average" value of process satisfies elliptic and parabolic partial differential equations (PDE).
- Paths look like Brownian motion paths with a little "drift". They have fractal dimension $\alpha = 2$

Harder problem: walks with self repulsion

- In statistical physics, there are many models of curves with very strong interactions.
- For example, polymer chains can be represented as random walks with the requirement that the chain does not cross itself. (Flory)
- Lattice model (self-avoiding walk) : give all random walk paths of the same number of steps the same probability

- Unlike usual random walk and Brownian motion, these processes are nonMarkovian. The future evolution of a curve depends on the entire past and not just on the current position.
- Self-avoiding walk is only one of many models arising in statistical physics "at criticality" where fractal curves arise.
- Another model is the loop-erased random walk obtained from a usual random walk and erasing the loops.
- Other curves arise as interfaces such as the percolation exploration process that we will see.

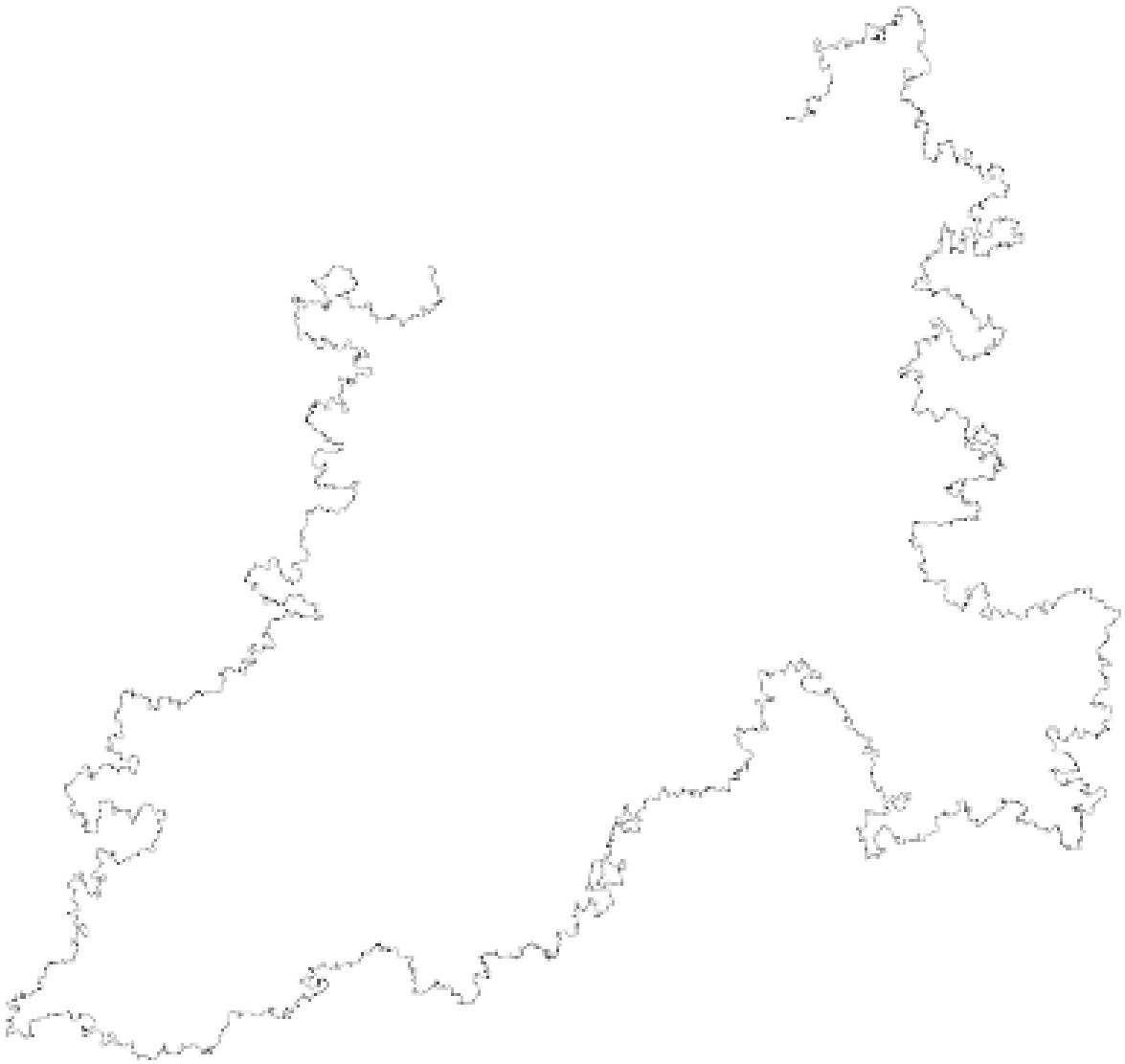


Figure: Self-avoiding walk (simulation by V. Beffara)

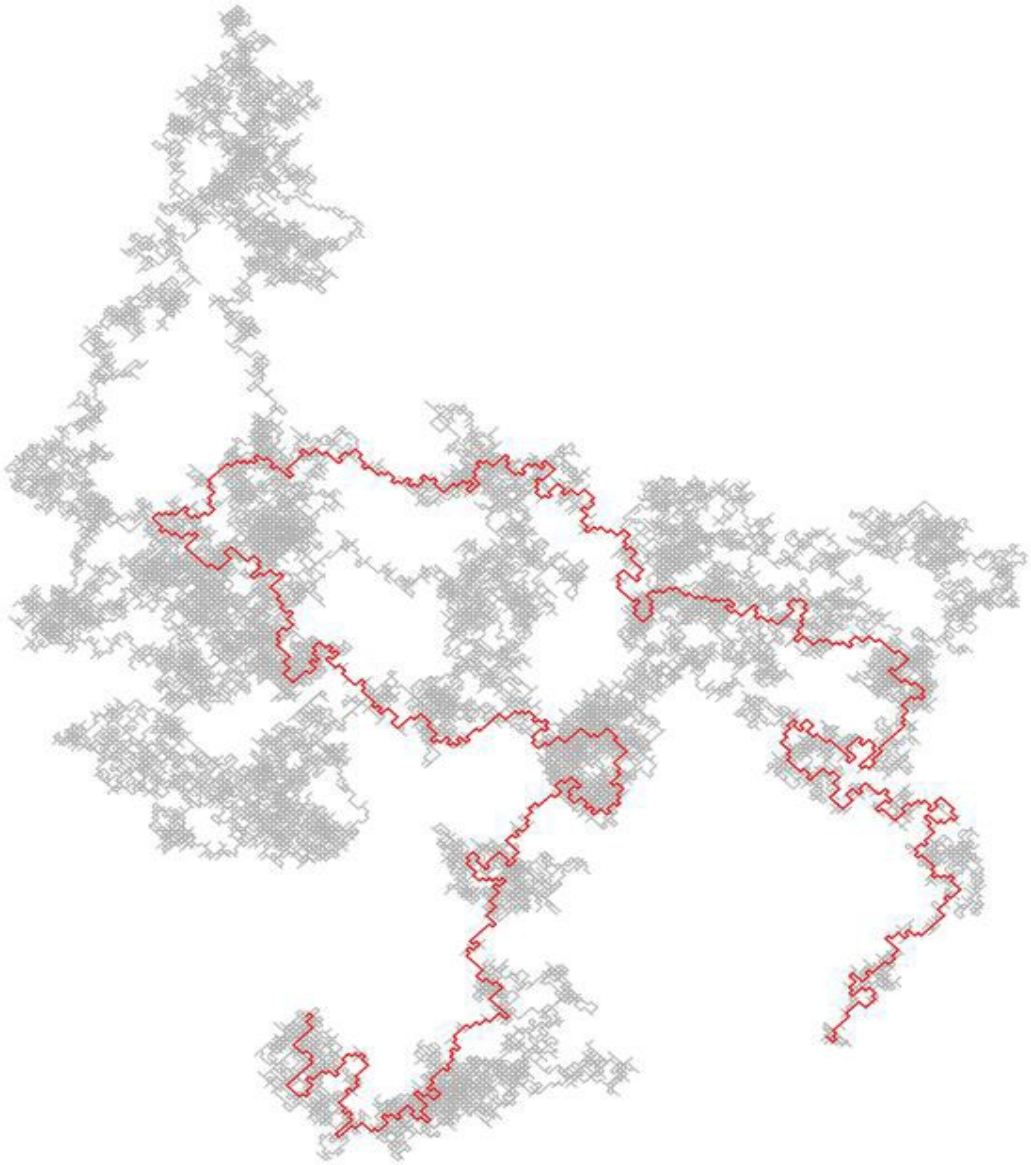
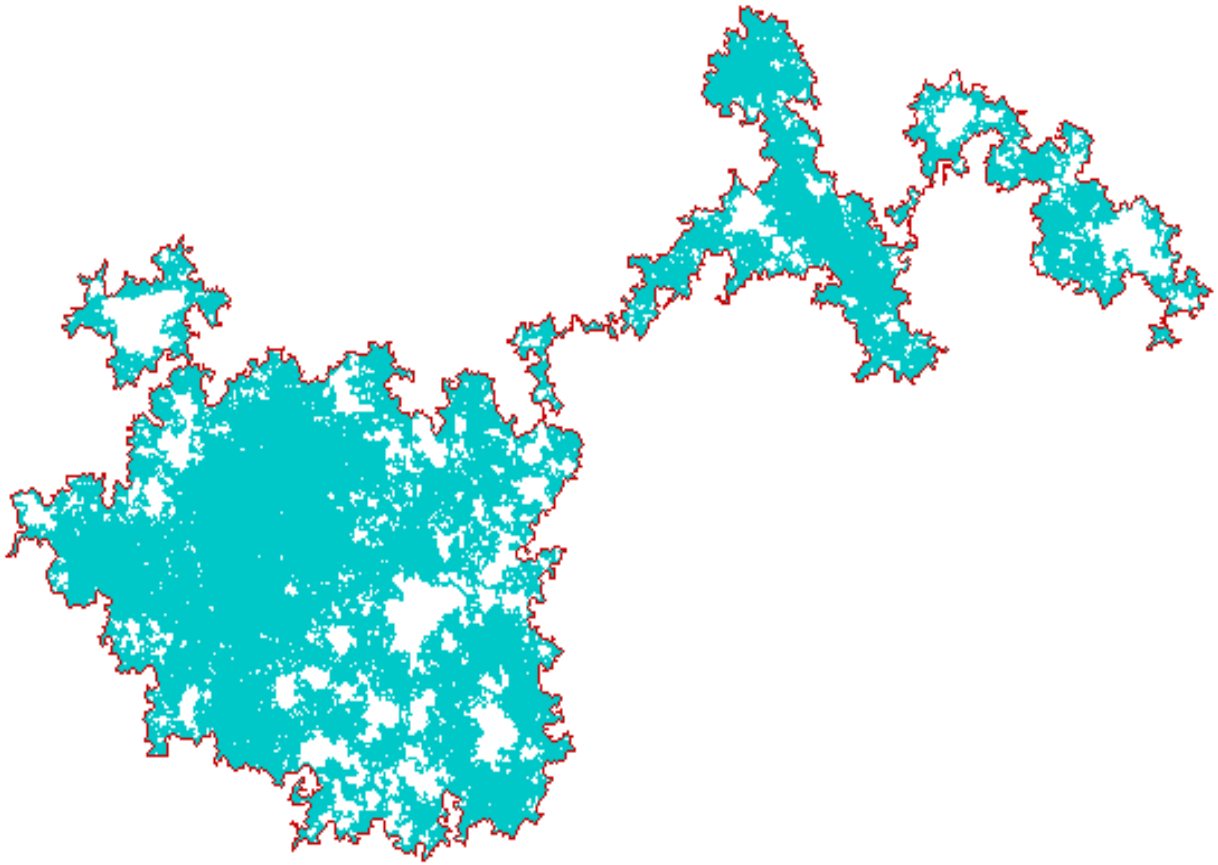


Figure: Loop-erased walk (simulation by F. Viklund)

In Benoit Mandelbrot's book Fractal Geometry of Nature we see another walk with self-avoidance. This is the outer boundary or frontier of a two dimension random walk.



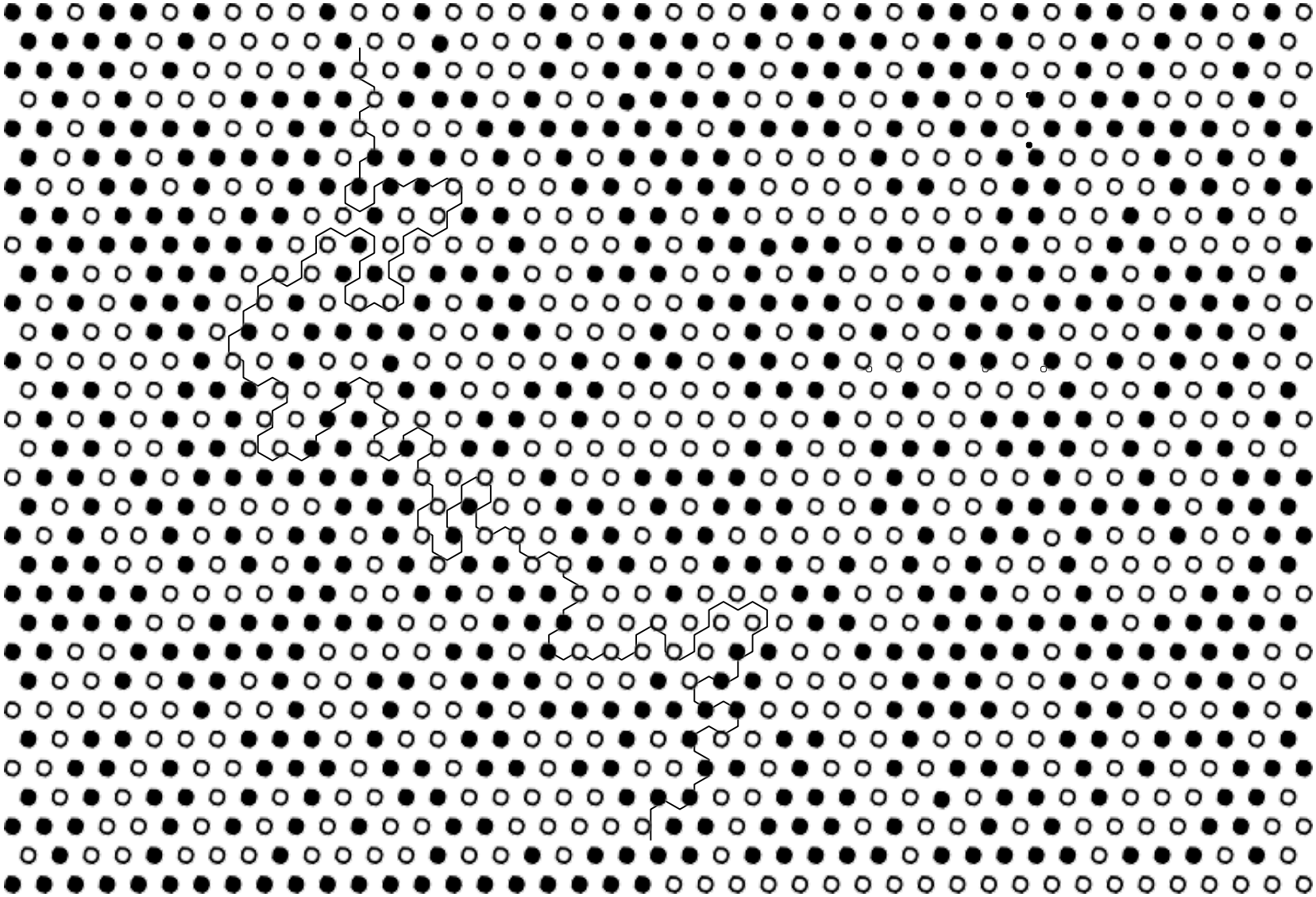


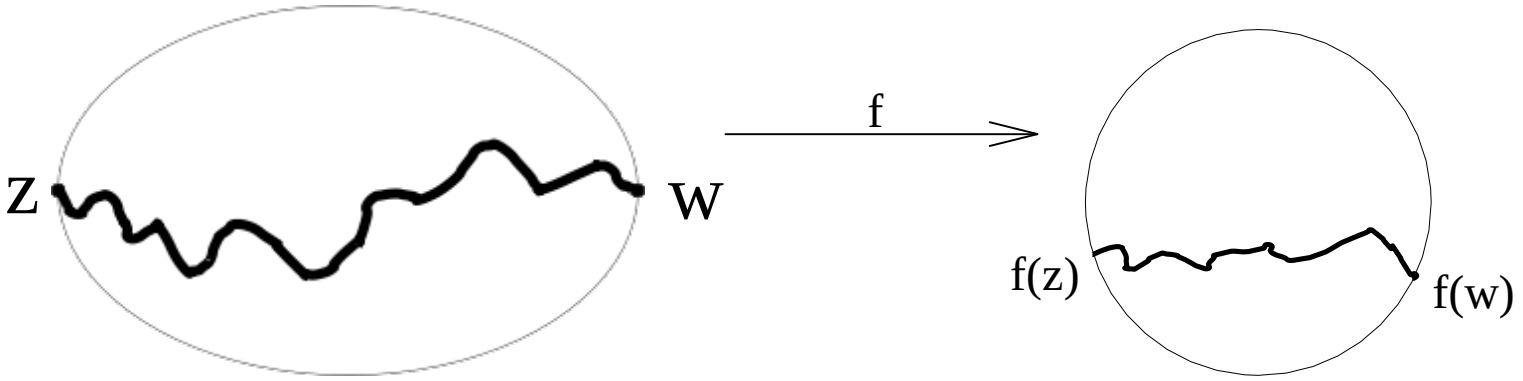
Figure: Percolation exploration process (simulation by G. Grimmett)

Critical dimension for paths with self repulsion

- If the spatial dimension is 4 or greater, than random paths do not tend to intersect (paths are two dimensional)
- The interesting dimensions are $d = 2, 3$.
- Focus on fractal dimension α: In the ball of radius N one visits about N^{α} points.
- The number of steps needed to reach distance N is N^{α}. Flory predicted $\alpha = 4/3$ for $d = 2$ and $\alpha = 5/3$ for $d = 3$. He was correct for $d = 2$ but (we expect) slightly wrong for
- $d = 3$, $\alpha = 1/.588\cdots$.

Conformal invariance for two dimensional systems

- It was first predicted nonrigorously by theoretical physicists that two dimensional systems exhibit conformal invariance in the limit.
- Conformal invariance implies that a function is locally a dilation and a rotation. For $d = 2$ this is the same as being complex differentiable and one-to-one.



- Brownian motion had already been shown to be conformally invariant. ”Random continuous motion” is rotationally invariant and has a kind of scale invariance.
- There have been incredible advance in the last twenty years on rigorous analysis of conformally invariant random planar fractals. The limit curves are called Schramm-Loewner evolutions.
  - Polymer (Self-avoiding walk) $\alpha = 4/3$. Loop-erased random walk $\alpha = 5/4$.
  - Percolation exploration process $\alpha = 7/4$.
  - The outer boundary of random walk is $\alpha = 4/3$.

Three dimensions: many open problems

- Can we define a process $\gamma(t)$ that is continuous, avoids the past, and is a potential limit for either the self-avoiding walk or the loop-erased random walk?
- It should have a fractal dimension α and satisfy the scaling rule

$$|\Delta\gamma(t)| \approx (\Delta t)^{1/\alpha}.$$

  For SAW, α a little bigger than 5/3; for loop-erased walk, a little smaller.

- Can α be determined exactly or is it just an unknown constant? There is no reason to expect nice rational numbers of values for the fractal dimensions.
- For percolation the separation between white and black sites would be a random fractal surface.

MATHEMATICAL FRONTIERS
Probability for People and Places

Kenneth L. Lange,
UCLA

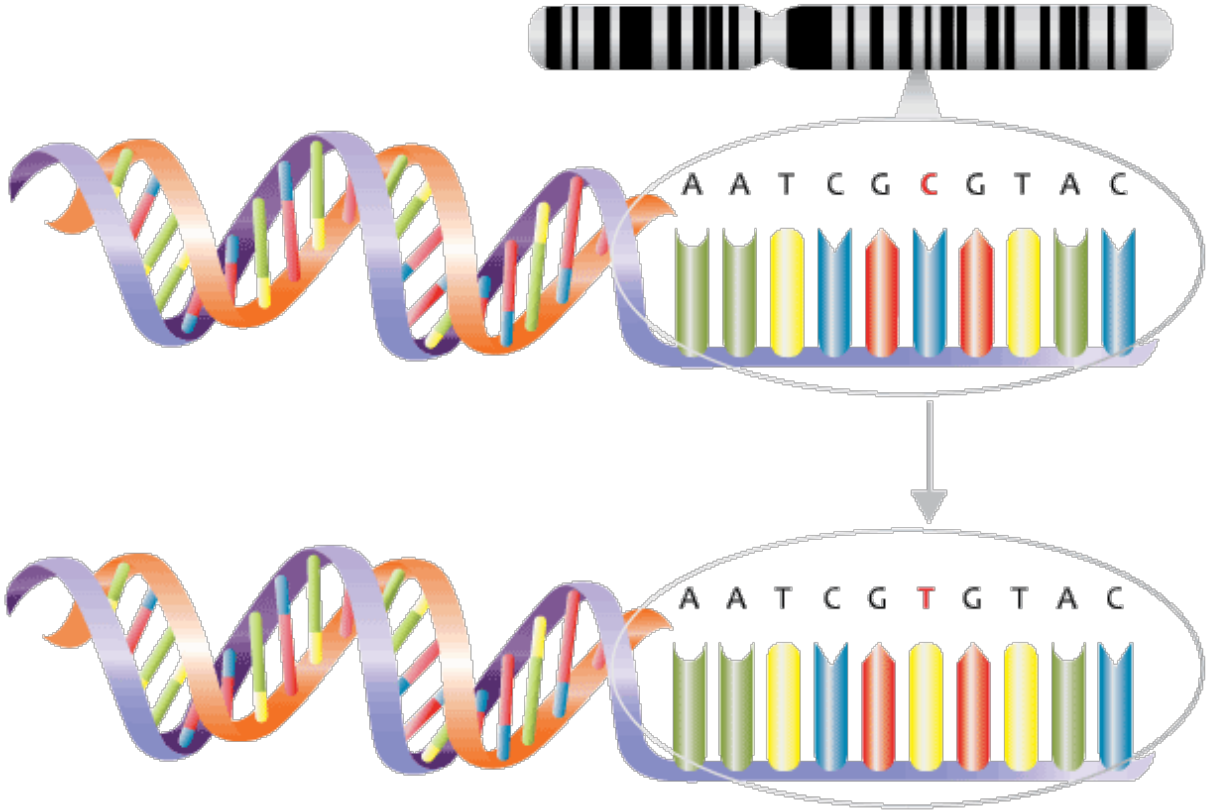
Rosenfeld Professor of
Computational Genetics
in the Department of Human Genetics

Professor in the Departments of
Biomathematics and Statistics

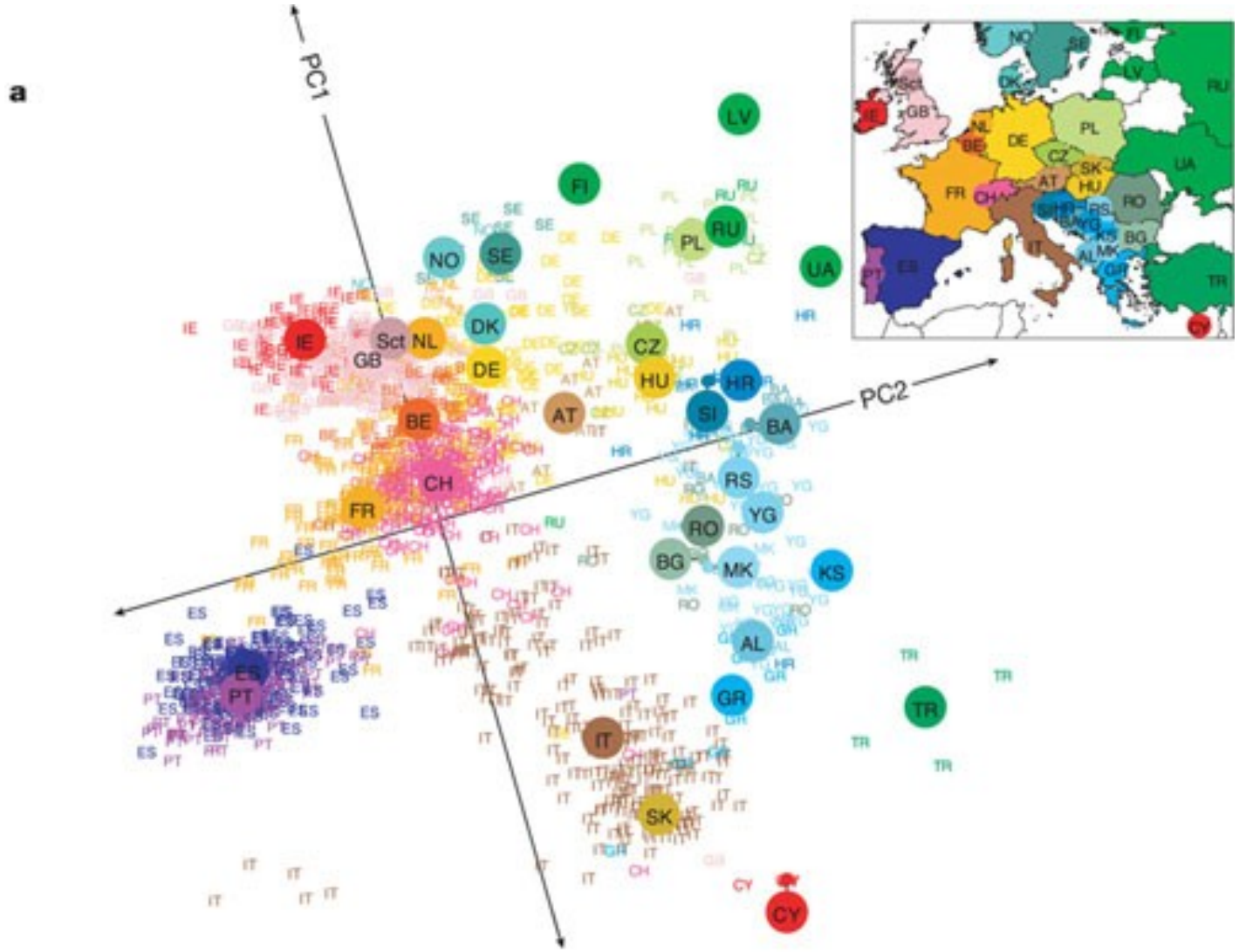
An Overview of Ancestry Estimation

Estimation of Ethnic Ancestry

- Major companies such as Ancestry.com exist to help people understand their origins.
- Population stratification is a potential confounding factor in genetic association studies. Estimated ancestries, derived from multi-locus genotype data, can serve as covariates (predictors) to correct for population stratification.
- Competing software: Structure uses Bayesian MCMC, Eigenstrat uses principal components, and Frappe uses an EM algorithm.
- Critique: Structure and Frappe are much too slow, and Eigenstrat does not deliver admixture fractions.
- Our program Admixture is orders of magnitude faster than Structure, which dominated the field for years.



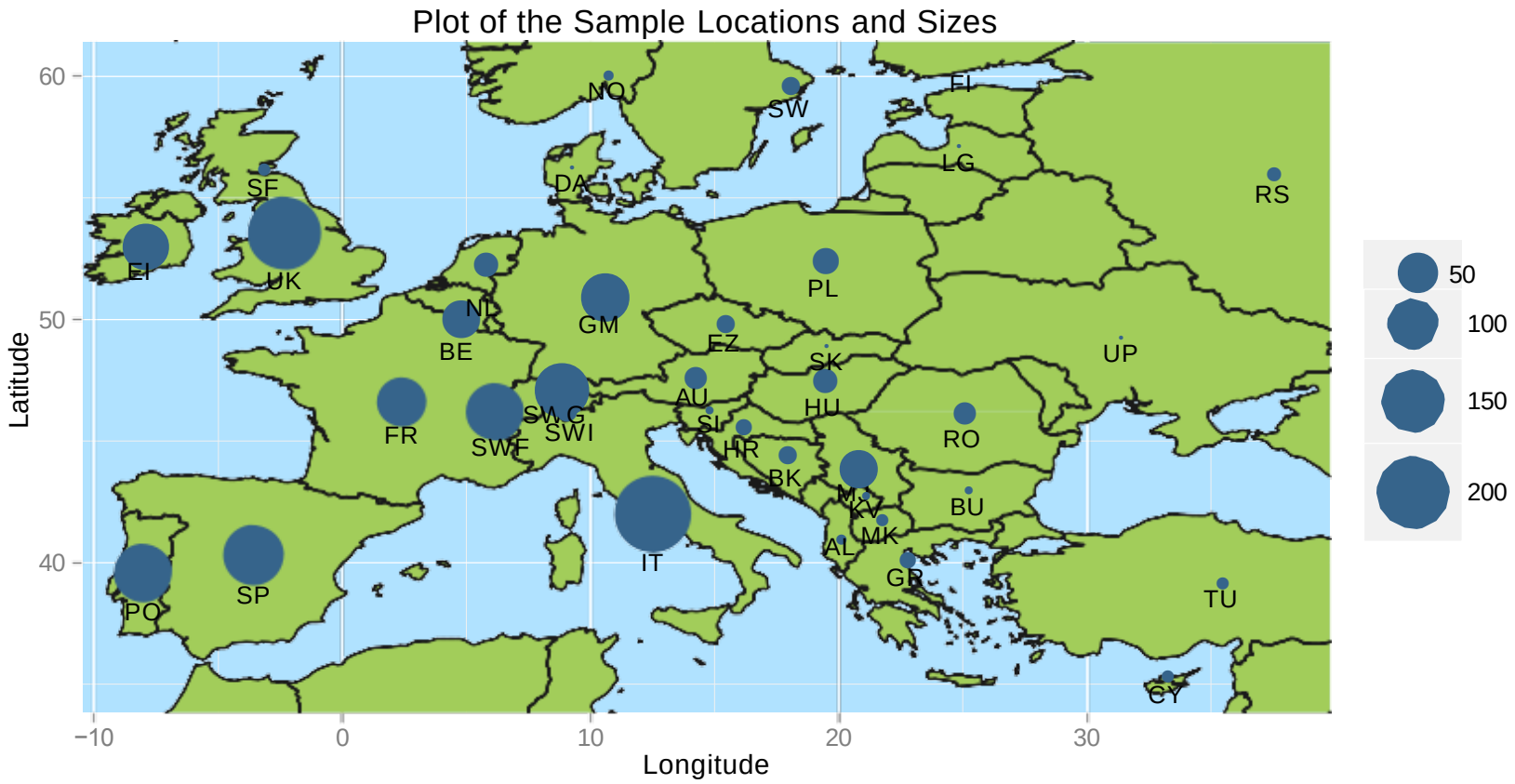
SNP: single nucleotide polymorphism (2 alleles per SNP)



Novembre, John et al (2008) Genes mirror geography within Europe. *Nature* 456:98–101

POPRES Data

1. 1,387 Europeans typed at 197,146 SNPs on an Affymetrix Genotyping Chip
2. All four grandparents of each subject came from the same region.
3. Tallies from each country appear on the next slide. Counts per county range from 219 (Italy) to 1 (Denmark, Finland, Latvia, Slovakia, and Ukraine). Swiss French, German, and Italians are counted separately.



Nelson MR, et al (2008) The Population Reference Sample (POPRES)
Amer J Hum Genet 83:347–358

Same Person, Different Admixture Estimates

Company 1	Companies 1 & 2	Company 3	Companies 3 & 4	Company 5
29.1% British, Irish	67.4% English	51% British	67.3% English	72.8% Scandinavia
18.8% French, German	11% Irish, Scottish, Welsh	30% W Europe	11.4% Irish	23.3 % SW Europe
15.9% Scandinavia	16.9% Italian	8% Irish, Scottish, Welsh	18.4% Italian	3.9% Central Asian
26.0% NW Europe	3.5% Baltic	4% E Europe	2.9% Baltic	
3.8% E Europe	1.2% Jewish	4% Scandinavia		
2.1% Jewish		1% Jewish		
0.5% S Europe		1% Iberian Peninsula		
0.3% Balkan		< 1% Caucasus		
3.7% Europe		< 1% S Europe		
0.2% Middle East, North Africa				

The Standard Admixture Model

- **Unknowns**: 1) the number of populations, 2) the fraction w_{ik} of individual i's genome attributable to population k, and 3) the frequency f_{kj} of allele 1 of SNP j in population k.
- In **unsupervised learning**, both the matrices $\boldsymbol{W} = (w_{ik})$ and $\boldsymbol{F} = (f_{kj})$ are unknown. In **supervised learning**, the underlying populations and the frequency matrix $\boldsymbol{F}$ are known.
- **Model assumptions**: random union of gametes (eggs and sperm) and the independent inheritance of all SNPs. The latter assumption (linkage equilibrium) is violated for positionally close SNPs.
- **Observed data**: the observed number y_{ij} of copies of allele 1 at SNP j of person i. Thus, y_{ij} equals 0, 1, or 2.
- **Distributional assumption**: the y_{ij} are binomially distributed random variables with 2 trials and success probability $\sum_k w_{ik} f_{kj}$.

The Likelihood Equations

- The loglikelihood of the data is

$$L(\boldsymbol{W},\boldsymbol{F}) = \sum_i \sum_j \left\{ y_{ij} \ln \left[\sum_k w_{ik} f_{kj} \right] + (2 - y_{ij}) \ln \left[\sum_k w_{ik} (1 - f_{kj}) \right] \right\}.$$

  This function is maximized in estimating ***F*** and ***W***.

- At the maximum point, the following stationarity equations should hold:

$$\frac{\partial}{\partial w_{ik}} L(\boldsymbol{W},\boldsymbol{F}) = \sum_j \left[\frac{y_{ij} f_{kj}}{\sum_l w_{il} f_{lj}} + \frac{(2 - y_{ij})(1 - f_{kj})}{\sum_l w_{il} (1 - f_{lj})} \right] = 0$$

$$\frac{\partial}{\partial f_{kj}} L(\boldsymbol{W},\boldsymbol{F}) = \sum_i \left[\frac{y_{ij} w_{ik}}{\sum_l w_{il} f_{lj}} - \frac{(2 - y_{ij}) w_{ik}}{\sum_l w_{il} (1 - f_{lj})} \right] = 0.$$

- Unfortunately, there is no obvious solution to this system of equations.

Hindrances to Numerical Maximization

- Assume I unrelated sample people, J SNPs, and K ancestral populations. Recall that w_{ik} is an ethnic fraction and f_{kj} is an allele frequency.
- The parameter matrices $W = \{w_{ik}\}$ and $F = \{f_{kj}\}$ have dimensions
- $I \times K$ and $K \times J$, for a total of $N = IK + KJ$ parameters. For the modest choices $I = 1000$, $J = 10,000$, $K = 3$, there are $N = 33,000$ parameters to estimate.
- The sheer number of parameters makes Newton's method infeasible. The storage required for the $N \times N$ Hessian matrix is prohibitively large, and the required matrix inversion is intractable.
- The loglikelihood has at least $K!$ equivalent global maxima and is subject to the constraints by $0 \leq f_{kj} \leq 1$, $w_{ik} \geq 0$, and $\sum_k w_{ik} = 1$.

Block Ascent Maximization

- Block ascent maximizes ***W*** holding ***F*** fixed and then maximizes ***F*** holding ***W*** fixed. These two steps are alternated until convergence.
- In the ***W*** updates, the admixture proportions for each individual *i* are optimized separately. In the ***F*** updates, the allele frequencies for each SNP are optimized separately.
- The block updates are found iteratively by sequential quadratic programming that maximizes the second-order Taylor's expansion of $L(\boldsymbol{W}, \boldsymbol{F})$ around the current parameter vector. Without constraints sequential quadratic programming coincides with Newton's method.
- Block ascent is accelerated by a generic secant method: Zhou H, Alexander DH, Lange K (2011) *Statistics and Computing* 21:261-273
- Standard errors are calculated via the parametric bootstrap or by inverting the expected information matrix when ***F*** is known.

Recent Enhancements to Admixture

- Estimation of the number of underlying populations through cross-validation.
- Exploitation of individuals of known ancestry from reference databases.
- Encouragement of admixture parsimony through penalization of admixture coefficients. This eliminates low admixture fractions.
- Parallel and GPU processing.
- X chromosome data and sex-specific admixture analysis.
- Use of all SNPs, not just the ”ancestry informative” SNPs. Instead of 10,000 SNPs, use all 1,000,000 SNPs on a chip.

References

1. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19:1655–1664
2. Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246
3. Price AL et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38:904–909
4. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959
5. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology* 28:289–301

MATHEMATICAL FRONTIERS
Probability for People and Places – Q&A

Gregory F. Lawler,
University of Chicago

Kenneth L. Lange,
UCLA

Elizabeth A. Thompson,
University of Washington

View webinar videos and learn more about BMSA at www.nas.edu/MathFrontiers

MATHEMATICAL FRONTIERS
2018 Monthly Webinar Series, 2-3pm ET

February 13: ***Recording posted***
Mathematics of the Electric Grid

March 13:
Probability for People and Places

April 10:
Social and Biological Networks

May 8:
Mathematics of Redistricting

June 12:
Number Theory: The Riemann Hypothesis

July 10: *Topology*

August 14:
Algorithms for Threat Detection

September 11:
Mathematical Analysis

October 9: *Combinatorics*

November 13:
Why Machine Learning Works

December 11:
Mathematics of Epidemics

View webinar videos and learn more about BMSA at www.nas.edu/MathFrontiers