Mathematical Embeddings of Complex Systems

Daryl DeFord

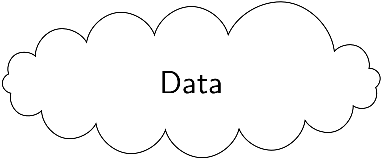
Dartmouth College
Department of Mathematics

Omidyar Fellowship Presentation
Santa Fe Institute
Santa Fe, NM
January 29, 2018

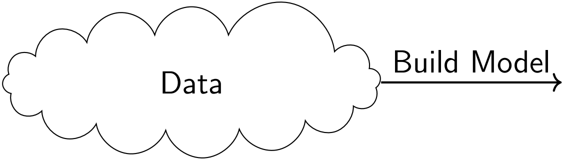
Dartmouth

Outline

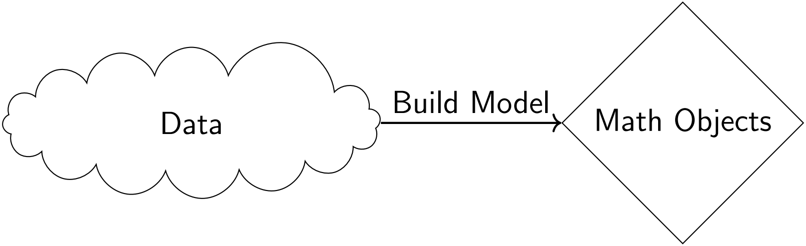
Philosophy



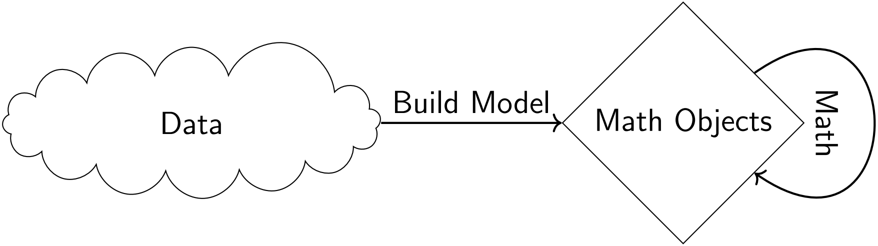
Philosophy



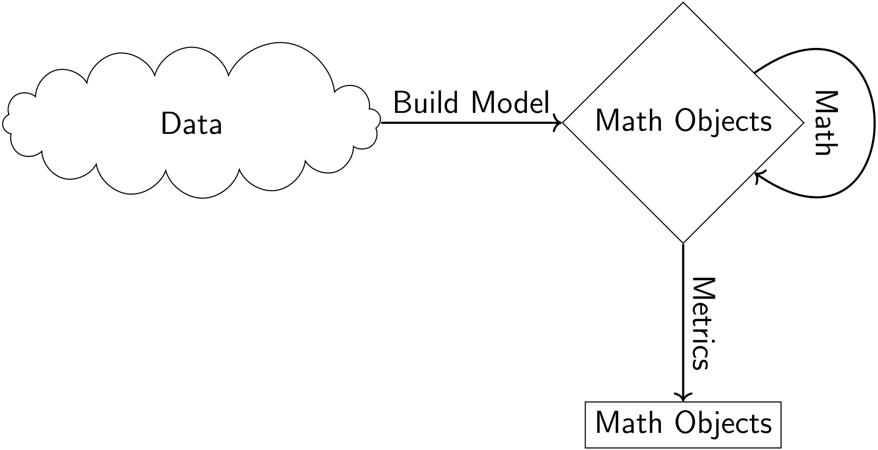
Philosophy



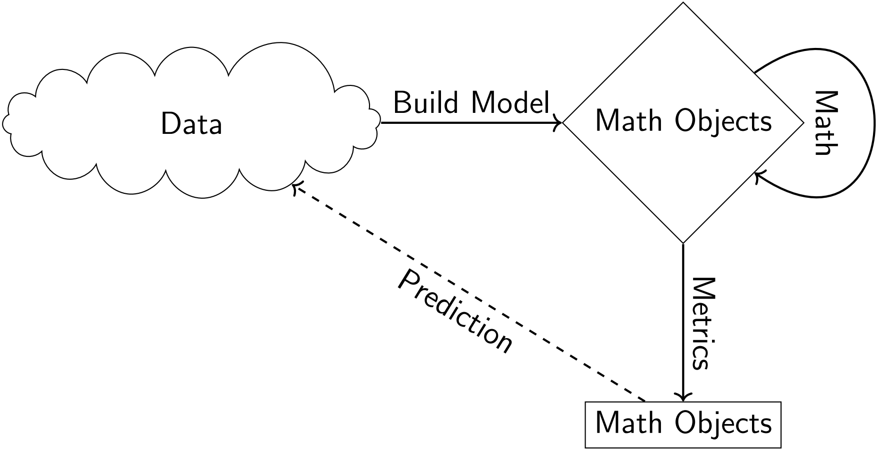
Philosophy



Philosophy



Philosophy


Data
Build Model
Math Objects
Math
Metrics
Prediction
Math Objects

Motivating Examples

(a) C.O.M.P.A.S.

Images from Wikipedia.

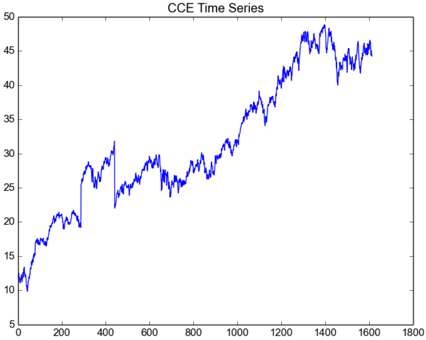
Motivating Examples

(a) C.O.M.P.A.S.

(b) Gerrymandering

Images from Wikipedia.

Time Series


CCE Time Series
50
45
40
35
30
25
20
15
10
5
0
200
400
600
800
1000
1200
1400
1600
1800

Iterated Maps

Given a function $f : [0, 1] \to [0, 1]$ and a point $x \in [0, 1]$, consider the behavior of $\{x, f(x), f(f(x)), f(f(f(x))), \ldots\}$.

Iterated Maps

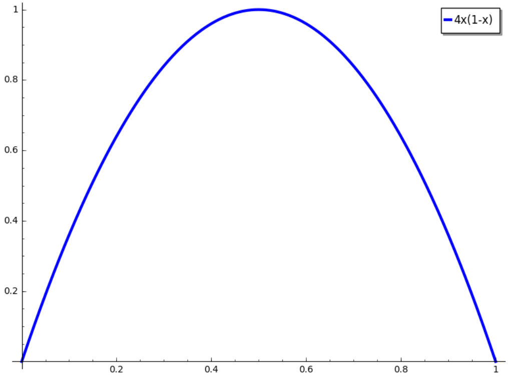
Given a function $f : [0,1] \to [0,1]$ and a point $x \in [0,1]$, consider the behavior of $\{x, f(x), f(f(x)), f(f(f(x))), \ldots\}$.

Example

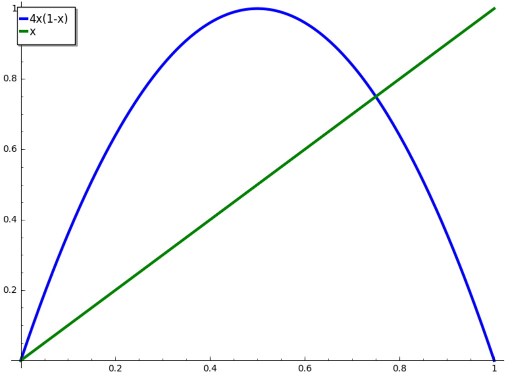
Let $f(x) = 4x(1-x)$ *and* $x_0 = .2$. *Then, the list of values is:*

$$[0.20, 0.64, 0.92, 0.28, 0.82, 0.58, 0.97, 0.11, 0.40, \ldots].$$

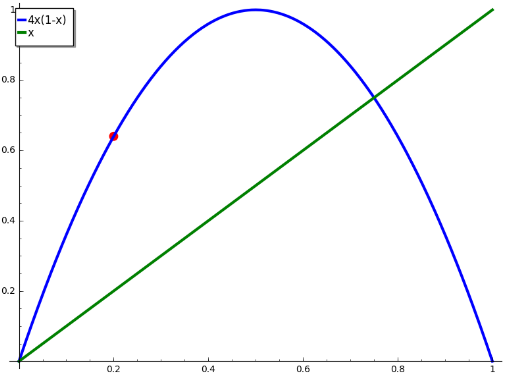
Iterated Example



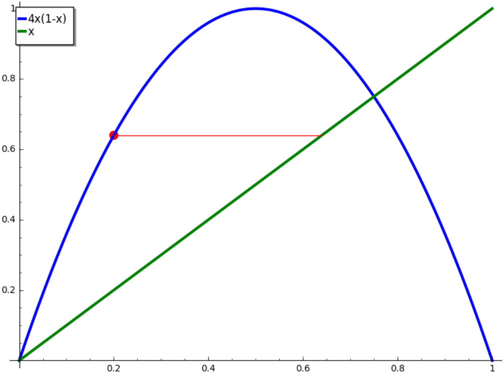
Iterated Example



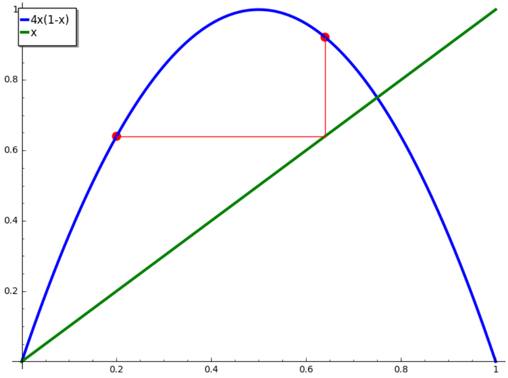
Iterated Example



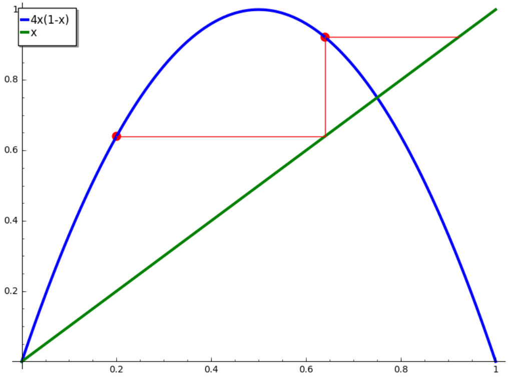
Iterated Example



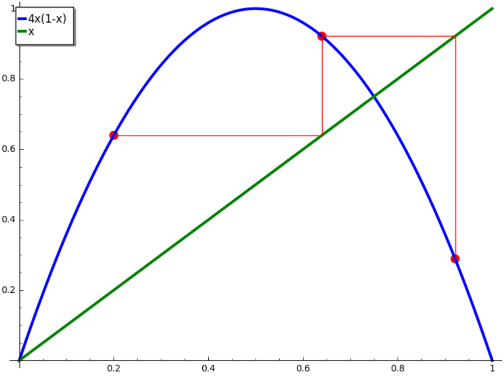
Iterated Example (12)



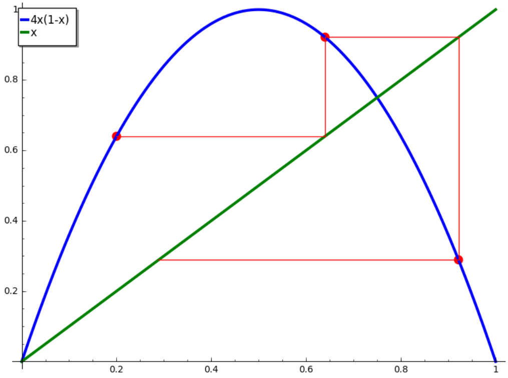
Iterated Example (12)



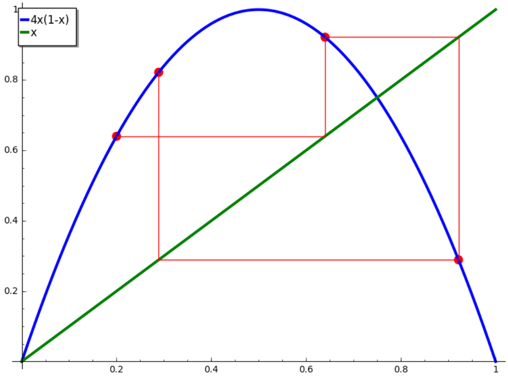
Iterated Example (231)



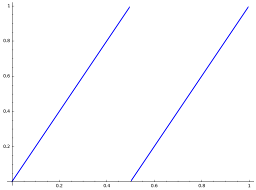
Iterated Example (231)

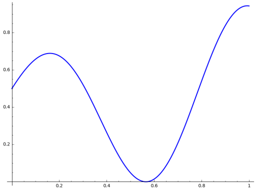


Iterated Example (2413)

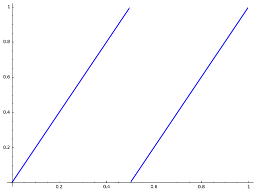


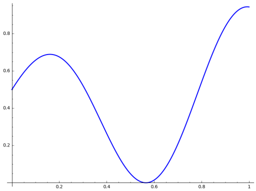
Forbidden Patterns





Forbidden Patterns

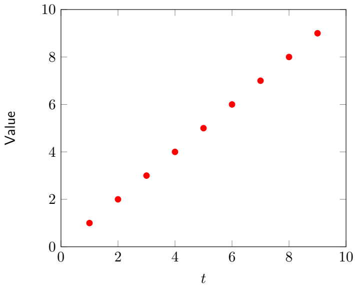


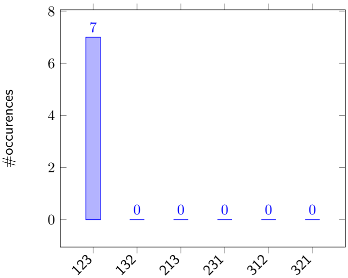


Definition (Topological Entropy)

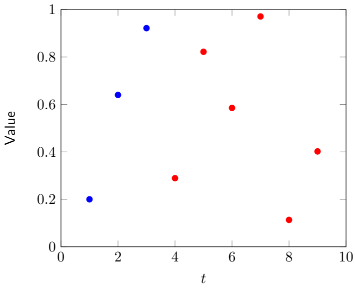
$$TE = \lim_{n \to \infty} \frac{\log(|\text{Allow}(f)|)}{n-1}$$

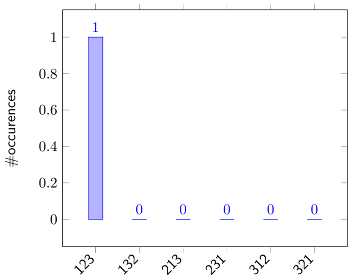
Simple Time Series



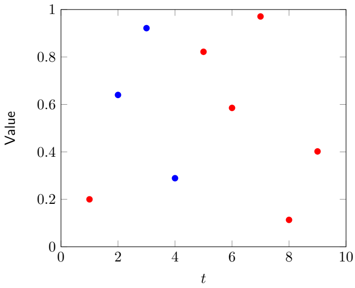


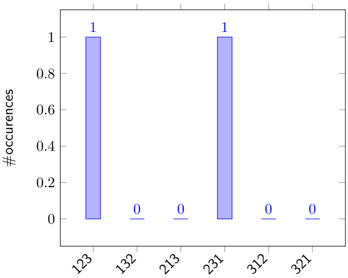
Complex Time Series



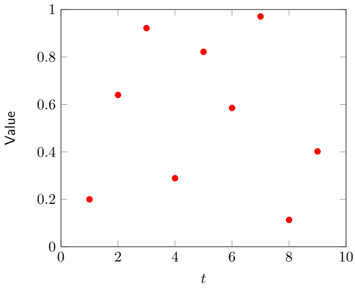


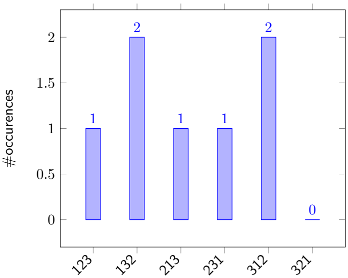
Complex Time Series





Complex Time Series





Complexity Measures

Definition (Normalized Permutation Entropy)

$$NPE(\{X_i\}) = -\frac{1}{\log(N!)} \sum_{\pi \in S_n} p_\pi \log(p_\pi)$$

Complexity Measures

Definition (Normalized Permutation Entropy)

$$NPE(\{X_i\}) = -\frac{1}{\log(N!)} \sum_{\pi \in S_n} p_\pi \log(p_\pi)$$

Definition (Uniform KL Divergence)

$$D_{KL}(\{X_i\}||\text{UNIFORM}) = \sum_{\pi \in S_n} p_\pi \log\left(\frac{p_\pi}{\frac{1}{n!}}\right)$$

Complexity Measures

Definition (Normalized Permutation Entropy)

$$NPE(\{X_i\}) = -\frac{1}{\log(N!)} \sum_{\pi \in S_n} p_\pi \log(p_\pi)$$

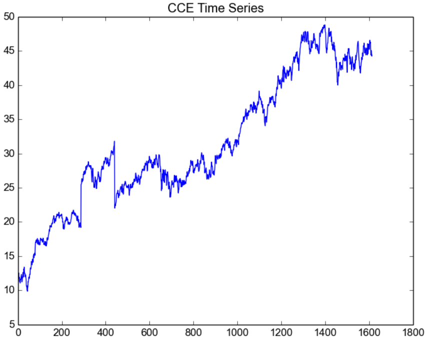
Definition (Uniform KL Divergence)

$$D_{KL}(\{X_i\}||\text{UNIFORM}) = \sum_{\pi \in S_n} p_\pi \log\left(\frac{p_\pi}{\frac{1}{n!}}\right)$$

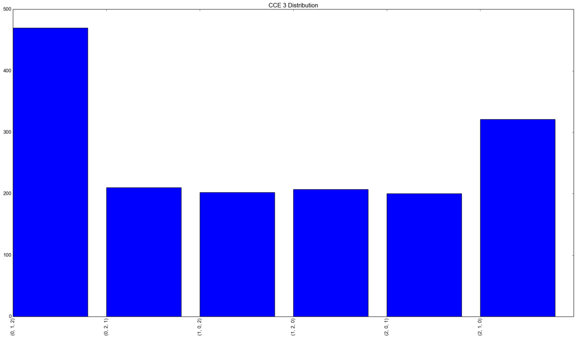
Observation

$$1 - NPE(\{X_i\}) = \frac{1}{\log(N!)} D_{KL}(\{X_i\}||\text{UNIFORM})$$

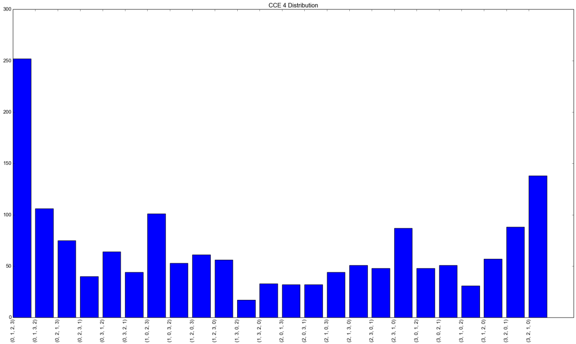
Stock Data (Closing Prices)



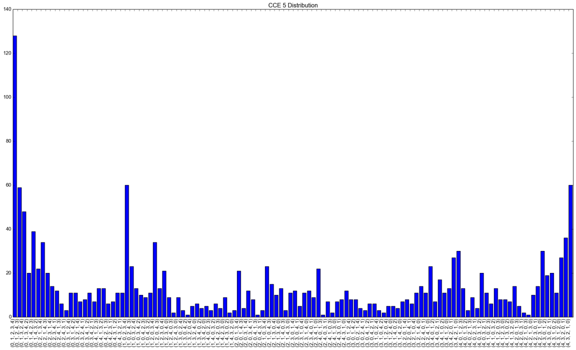
Stock Data (n=3)



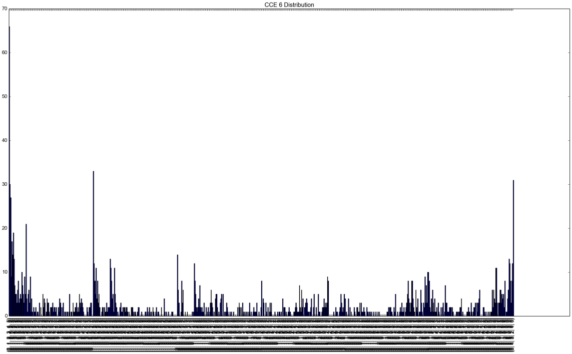
Stock Data (n=4)



Stock Data (n=5)



Stock Data (n=6)



Random Walk Null Models

Definition (Random Walk)

Let $\{X_i\}$ be a set of I.I.D. continuous random variables and define $\{Z_i\}$ by $Z_j = \sum_{i=0}^{j} X_j$. Usually the steps $\{X_i\}$ will be uniformly or normally distributed.

Random Walk Null Models

Definition (Random Walk)

Let $\{X_i\}$ be a set of I.I.D. continuous random variables and define $\{Z_i\}$ by $Z_j = \sum_{i=0}^{j} X_j$. Usually the steps $\{X_i\}$ will be uniformly or normally distributed.

Proposition (No Forbidden Patterns)

If $\{Z_i\}$ are defined as above then every permutation occurs with some positive probability.

Random Walk Null Models

Definition (Random Walk)

Let $\{X_i\}$ be a set of I.I.D. continuous random variables and define $\{Z_i\}$ by $Z_j = \sum_{i=0}^{j} X_j$. Usually the steps $\{X_i\}$ will be uniformly or normally distributed.

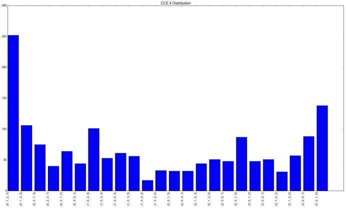
Proposition (No Forbidden Patterns)

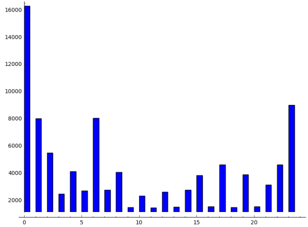
If $\{Z_i\}$ are defined as above then every permutation occurs with some positive probability.

Proposition (No Uniform Distribution)

If $\{Z_i\}$ are defined as above and $n \geq 3$ then the expected distribution of permutations is not uniform.

Uniform Steps CCE





New Complexity Measure

Definition (Null Model KL Divergence)

$$\mathsf{D_{KL}}_n(X) := \mathsf{D_{KL}}_n(X||Z) = \sum_{\pi \in \mathcal{S}_n} p_\pi \log\left(\frac{p_\pi}{q_\pi}\right),$$

where p_π is the relative frequency of π in X and q_π is the relative frequency of π in Z.

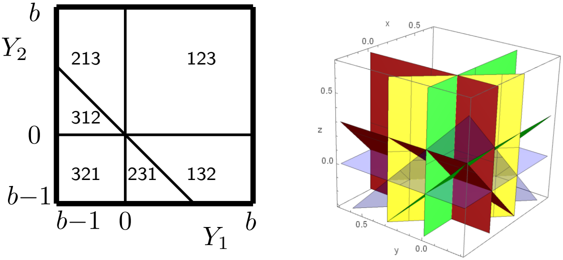
Hyperplanes

Example (Uniformly distributed steps)

In order for the pattern 1342 *to appear in the random walk time series we need the following inequalities to hold:*

- $X_1 > 0$
- $X_2 > 0$
- $X_3 < 0$
- $X_3 > X_2$
- $X_3 < X_1 + X_2$

Integration Regions



(a) (b)

Figure: The regions of integration for patterns in uniform random walks for (**a**) $n = 3$ and (**b**) $n = 4$, sketched here for $b = 0.65$.

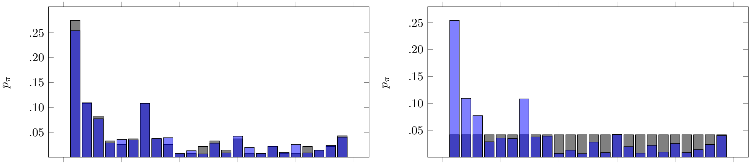
Null Distributions ($n = 3$)

Pattern	**Normal: $\mu = 0$**	**Uniform: $\mu = 0$**	**Uniform: $\mathbb{P}(Y > 0) = b$**
{123}	$1/4$	$1/4$	b^2
{132, 213}	$1/8$	$1/8$	$(1/2)(1-b)^2$
{231, 312}	$1/8$	$1/8$	$(1/2)(b^2+2b-1)$
{321}	$1/4$	$1/4$	$(1-b)^2$

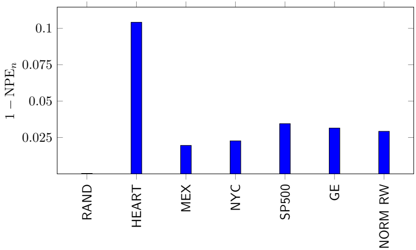
Null Distributions ($n = 4$)

Pattern	Normal: $\mu = 0$	Uniform: $\mu = 0$	Uniform: $\mathbb{P}(Y > 0) = b$
{1234}	0.1250	1/8	b^3
{1243, 2134}	0.0625	1/16	$(1/2)b(1-b)(3b-1)$
{1324}	0.0417	1/24	$(1/3)(1-b)(7b^2-5b+1)$
{1342, 3124}	0.0208	1/24	$(1/6)(1-b)^2(4b-1)$
{1423, 2314}	0.0355	1/48	$(1/6)(1-b)^2(5b-2)$
{1432, 2143, 3214}	0.0270	1/48	$\begin{cases}(1/6)(2-24b+48b^2-15b^3) & \text{if } b \leq 2/3 \\ (b-1)^2(2b-1) & \text{if } b > 2/3\end{cases}$
{2341, 3412, 4123}	0.0270	1/48	$(1/6)(1-b)^3$
{2413}	0.0146	1/48	$(1/6)(1-b)^3$
{2431, 4213}	0.0208	1/24	$\begin{cases}(1/6)(24b^3-45b^2+27b-5) & \text{if } b \leq 2/3 \\ (1/2)(1-b)^3 & \text{if } b > 2/3\end{cases}$
{3142}	0.0146	1/48	$\begin{cases}(1/6)(25b^3-48b^2+30b-6) & \text{if } b \leq 2/3 \\ (1/3)(1-b)^3 & \text{if } b > 2/3\end{cases}$
{3241, 4132}	0.0355	1/48	$(1/6)(1-b)^3$
{3421, 4312}	0.0625	1/16	$(1/2)(1-b)^3$
{4231}	0.0417	1/24	$(1/3)(1-b)^3$
{4321}	0.1250	1/8	$(1-b)^3$

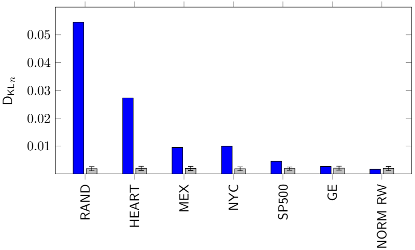
Uniform Steps S&P 500



Data Comparisons

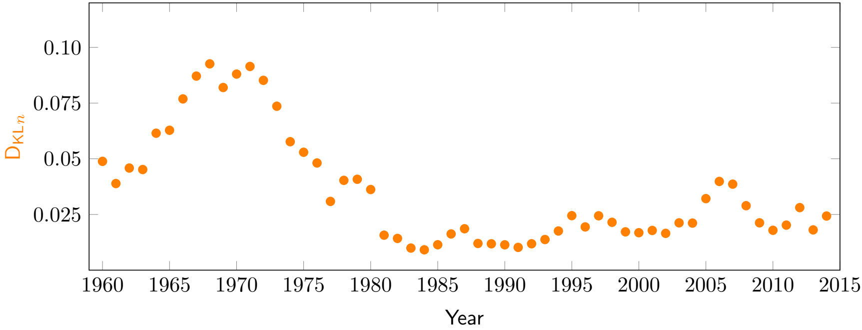


(a)

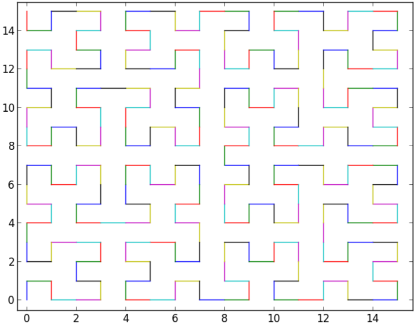


(b)

Stock Market Example



Hilbert Curve



Parallel Computing

- Latency in communication overhead has become a limiting factor in designing parallel algorithms
- Designing ways to efficiently embed problems to minimize communication is an important problem

Parallel Computing

- Latency in communication overhead has become a limiting factor in designing parallel algorithms
- Designing ways to efficiently embed problems to minimize communication is an important problem
- Factors:
  - Processor Topology
  - System Size
  - Distribution of Data
  - ...

Parallel Computing

- Latency in communication overhead has become a limiting factor in designing parallel algorithms
- Designing ways to efficiently embed problems to minimize communication is an important problem
- Factors:
    - Processor Topology
    - System Size
    - Distribution of Data
    - ...
- Previous Approaches
    - Database Methods
    - Average Nearest Neighbor Stretch

Discrete Space Filling Curves

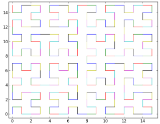
Definition

A Discrete Space Filling Curve is a mapping from multi–dimensional space to a linear order that allows for a unique indexing of the points.

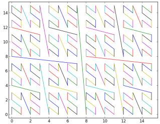
Discrete Space Filling Curves

Definition

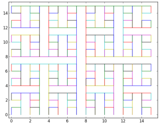
A Discrete Space Filling Curve is a mapping from multi–dimensional space to a linear order that allows for a unique indexing of the points.



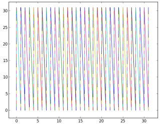
(a) Hilbert Curve $\mathcal{H}_4$



(b) Z–Curve $\mathcal{Z}_4$

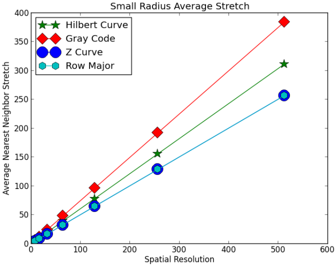


(c) Gray Order $\mathcal{G}_4$

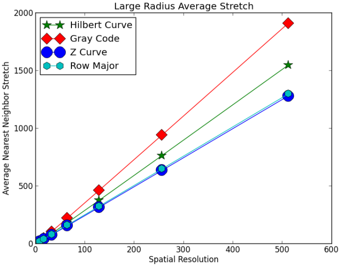


(d) Column Order

ANNS Result



(a) Standard ANNS



(b) Large Radius ANNS

Average Communicated Distance

Definition (ACD)

Given a particular problem instance, the *Average Communicated Distance (ACD)* is defined as the average distance for every pairwise communication made over the course of the entire application. The communication distance between any two communicating processors is given by the length of the shortest path (measured in the number of hops) between the two processors along the network intraconnect.

Average Communicated Distance

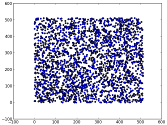
Definition (ACD)

Given a particular problem instance, the *Average Communicated Distance (ACD)* is defined as the average distance for every pairwise communication made over the course of the entire application. The communication distance between any two communicating processors is given by the length of the shortest path (measured in the number of hops) between the two processors along the network intraconnect.

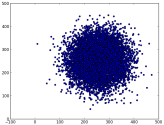
Definition (FMM)

The Fast Multipole Method (FMM) is an algorithm for approximating the interactions in an n body problem. The computations can be separated into two components: Near Field Interactions (NFI) and Far Field Interactions (FFI), with very different communications profiles.

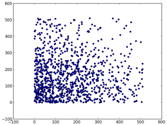
Point Distributions



(a) Uniform

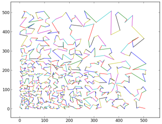


(b) Normal

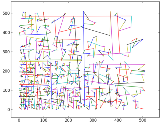


(c) Exponential

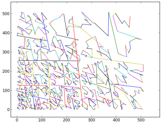
Ordered Points



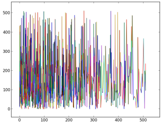
(a) Hilbert Ordering



(b) Gray Ordering



(c) Z Ordering



(d) Row Major Ordering

Main Results (NFI)

Table: A comparison of different particle/processor-order SFC combinations for NFI under various distributions. The lowest ACD value within each row is displayed in **boldface blue**, while the lowest ACD value within each column is displayed in *red italics*. The best option for each distribution is displayed in ***bold green italics***.

	Particle Order			
Processor Order ↓	Hilbert Curve	Z–Curve	Gray Code	Row Major
Hilbert Curve	**4.008**	*4.308*	*4.939*	*13.117*
Z–Curve	**5.486**	5.758	6.573	18.127
Gray Code	**5.802**	6.010	6.970	19.220
Row Major	**9.126**	9.763	11.713	70.353

Table: Uniform Distribution

Main Results (NFI) Continued

	Particle Order			
Processor Order ↓	Hilbert Curve	Z–Curve	Gray Code	Row Major
Hilbert Curve	**8.561**	*9.297*	*10.123*	*20.340*
Z–Curve	**11.003**	11.551	12.984	26.842
Gray Code	**11.881**	12.595	13.249	28.188
Row Major	**20.143**	22.221	24.053	66.719

(a) Normal Distribution

	Particle Order			
Processor Order ↓	Hilbert Curve	Z–Curve	Gray Code	Row Major
Hilbert Curve	**5.238**	*5.654*	*6.271*	*14.943*
Z–Curve	**6.943**	7.070	8.235	20.851
Gray Code	**7.276**	7.663	8.760	22.269
Row Major	**12.483**	13.017	15.289	61.227

(b) Exponential Distribution

Main Results (FFI)

Table: A comparison of different particle/processor-order SFC combinations for FFI under various distributions. The lowest ACD value within each row is displayed in **blue boldface**, while the lowest ACD value within each column is displayed in *red italics*. The best option for each distribution is displayed in ***bold green italics***.

	Particle Order			
Processor Order ↓	Hilbert Curve	Z–Curve	Gray Code	Row Major
Hilbert Curve	***19.494***	*20.841*	*22.572*	*31.124*
Z–Curve	**24.217**	24.793	27.787	37.709
Gray Code	**24.622**	25.446	27.997	39.282
Row Major	**44.513**	48.762	50.118	57.880

Table: Uniform Distribution

Main Results (NFI) Continued

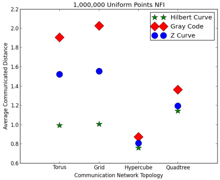
	Particle Order			
Processor Order ↓	Hilbert Curve	Z–Curve	Gray Code	Row Major
Hilbert Curve	**26.336**	*26.824*	31.963	*32.542*
Z–Curve	29.160	**28.036**	34.241	36.663
Gray Code	29.449	**27.981**	*31.909*	37.291
Row Major	**43.639**	44.636	49.133	45.475

(a) Normal Distribution

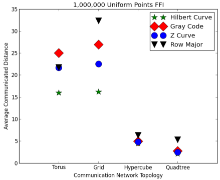
	Particle Order			
Processor Order ↓	Hilbert Curve	Z–Curve	Gray Code	Row Major
Hilbert Curve	**18.960**	*19.841*	*23.007*	*31.368*
Z–Curve	24.672	**23.316**	26.315	37.576
Gray Code	**23.762**	24.076	27.973	37.863
Row Major	**42.447**	44.067	46.872	50.963

(b) Exponential Distribution

Topology Comparison



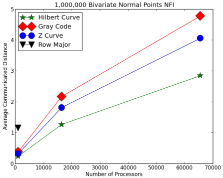
(e) Near–Field Interactions



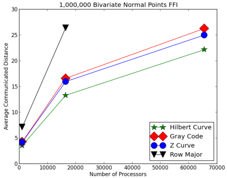
(f) Far–Field Interactions

Figure: The charts show the results of comparing different network topologies for a) NFI and b) FFI, respectively. All experiments were performed using $1,000,000$ uniformly distributed particles on a 4096×4096 spatial resolution. This plot is representative of all the experiments we performed to evaluate the topologies. It is important to note that quadtree structures have disproportionately large issues with contention in high volume communications.

Processor Scaling



(a) NFI



(b) FFI

Figure: These plots show ACD values for a) NFI, and b) FFI, as a function of the number of processors and the SFC used. The input used was fixed at 1,000,000 uniformly distributed particles. This demonstrates the effect scale on processor ranking SFCs. Some of the row–major data has been excluded from these plots because for this SFC, the ACD values at larger processor numbers were significantly higher than the other data–points.

What is a multiplex?

What is a multiplex?

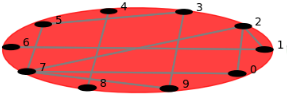
Definition

A *multiplex* is a collection of graphs all defined on the same node set.

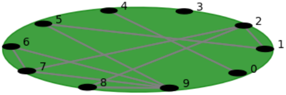
What is a multiplex?

Definition

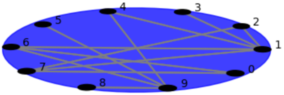
A *multiplex* is a collection of graphs all defined on the same node set.



(a) Family



(b) Colleagues



(c) Facebook

World Trade Web[1]

Figure: World trade networks

[1] R. Feenstra, R. Lipsey, H. Deng, A.C. Ma, and H. Mo. World Trade Flows: 1962-2000. NBER Working Paper 11040, (2005).

WTW Layers

Layer	Description	Volume	% Total	Transitivity
0	Food and live animals	291554437	5.1	.82
1	Beverages and tobacco	48046852	0.9	.67
2	Crude materials	188946835	3.3	.79
3	Mineral fuels	565811660	10.0	.62
4	Animal and vegetable oils	14578671	0.3	.64
5	Chemicals	535703156	9.5	.83
6	Manufactured Goods	790582194	13.9	.87
7	Machinery	2387828874	42.1	.85
8	Miscellaneous manufacturing	736642890	13.0	.83
9	Other commodities	107685024	1.9	.56
All	Aggregate Trade	5667380593	100	.93

Table: Layer information for the 2000 World Trade Web.

Disjoint Layers

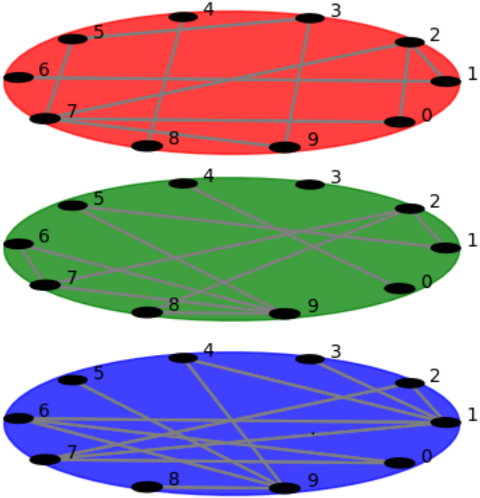
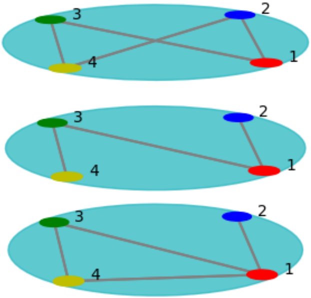
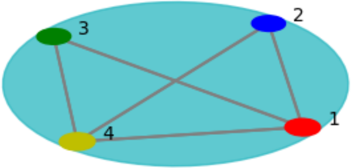


Figure: Disjoint Layers

Aggregate Models

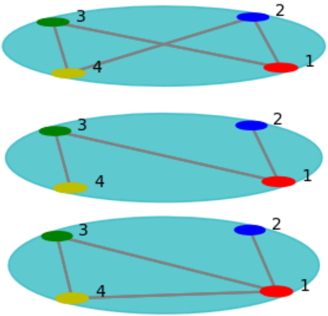


(a) Disjoint Layers

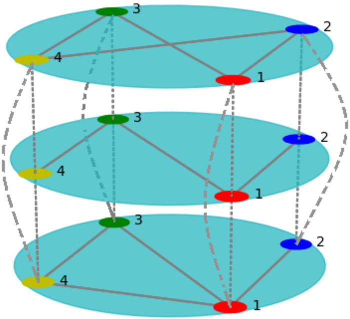


(b) Aggregate

Matched Sum



(a) Disjoint Layers



(b) Matched Sum

Strucutral Asymptotics

As the number of layers grows, what happens to the:

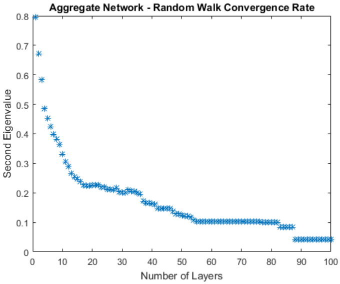
- Density?
- Degree Distribution?
- Transitivity?
- Average Path Length?
- Diameter?
- Clique Number?
- ...

Strucutral Asymptotics

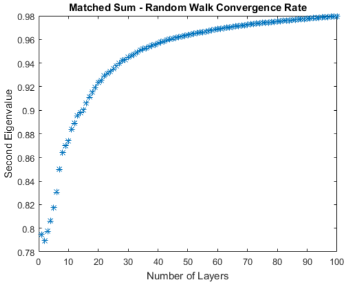
As the number of layers grows, what happens to the:

- Density?
- Degree Distribution?
- Transitivity?
- Average Path Length?
- Diameter?
- Clique Number?
- ...
- **Dynamics?!?**

Random Walk Convergence



(a) Aggregate



(b) Matched Sum

Dynamics on Multiplex Networks

- Two types of interactions
  - Within the individual layers
  - Between the layers
- Effects should “pass through” nodes
- Two step iterative model

Dynamics on Multiplex Networks

- Two types of interactions
  - Within the individual layers
  - Between the layers
- Effects should "pass through" nodes
- Two step iterative model
- Symbolically:

$$v' = \mathscr{D}v$$

$$(v')_i^\alpha = \sum_{\beta=1}^{k} m_i^{\alpha,\beta} c_i^{\alpha,\beta} (Dv)_i^\beta$$

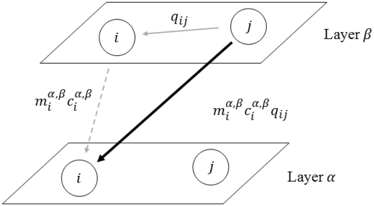
Dartmouth

Dynamics on Multiplex Networks

- Two types of interactions
  - Within the individual layers
  - Between the layers
- Effects should "pass through" nodes
- Two step iterative model
- Symbolically:

$$v' = \mathscr{D}v$$

$$(v')_i^\alpha = \sum_{\beta=1}^{k} m_i^{\alpha,\beta} c_i^{\alpha,\beta} (Dv)_i^\beta$$



Multiplex Random Walks

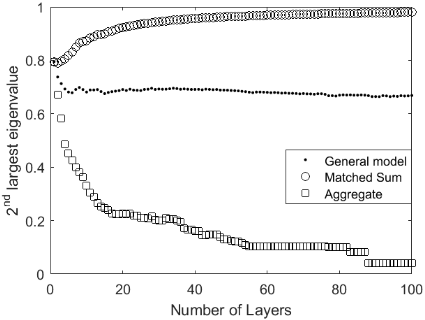
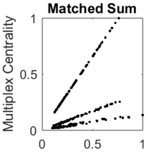
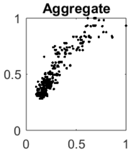


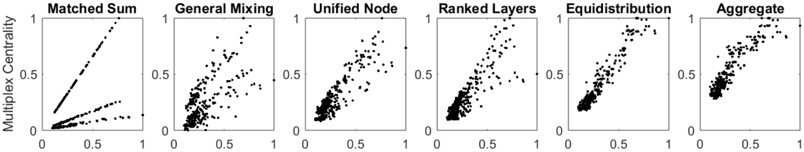
Figure: Comparison of random walk convergence for multiplex models.

Centrality Comparison





Centrality Comparison



Global Aggregate Rankings

Year	1970	1980	1990	2000
1	US	US	US	US
2	Germany	Germany	Germany	Germany
3	Canada	Japan	Japan	Japan
4	UK	UK	France	China
5	Japan	France	UK	UK
6	France	Saudi Arabia	Italy	France

Table: RWBC values for the aggregate WTW.

Full Multiplex RWBC

Ranking	Country	Layer
1	US	7
2	Germany	7
3	China	7
4	UK	7
5	Japan	7
6	US	8
7	Canada	7
8	France	7
9	Japan	3
10	US	6
12	US	3
13	Netherlands	7
14	Germany	6
15	Italy	7

Table: Multiplex RWBC values for the 2000 WTW.

Commodity Appearance

Layer	Ranking	Country
0	22	Japan
1	199	Germany
2	47	China
3	9	Japan
4	184	Australia
5	23	Germany
6	10	US
7	1	US
8	6	US
9	39	US

Table: First appearance of each layer in the rankings.

Ranking Movement

Layer 7 Ranking	Country	Multiplex Ranking
1	USA	1
2	Japan	5
3	Germany	2
4	China	3
5	France	8
6	UK	4
7	South Korea	18
8	Canada	7
9	Malaysia	16
10	Mexico	20

Table: Comparison of the relative rankings of the RWBC on Layer 7 versus the multiplex RWBC.

Multiplex References

- M. Kivela, A. Arenas, M. Barthelemy, J. Gleeson, Y. Moreno, and M. Porter: *Multilayer Networks*, Journal of Complex Networks, 1–69, (2014).
- D. DeFord and S. Pauls: *A New Framework for Dynamical Models on Multiplex Networks*, Journal of Complex Networks, to appear, (2017), 29 pages.
- D. DeFord and S. Pauls: *Spectral Clustering Methods for Multiplex Networks*, submitted, arXiv:1703.05355, (2017), 22 pages.
- D. DeFord: *Multiplex Dynamics on the World Trade Web*, Proc. 6th International Conference on Complex Networks and Applications, Studies in Computational Intelligence, Springer, 1111–1123, (2018).

SFC References

- S. Aluru and F. Sevilgen: *Parallel Domain Decomposition and Load Balancing Using Space–Filling Curves*, Fourth International Conference on High Performance Computing, (1997) 230–235.
- D. DeFord and A. Kalyanaraman: *Empirical Analysis of Space-Filling Curves for Scientific Computing Applications*, Proc. 42nd International Conference on Parallel Processing, 170-179, 2013.
- B. Moon, H. Jagadish, C. Faloutsos, and J. Saltz: *Analysis of the Clustering Properties of Hilbert Space–Filling Curve*, IEEE Transactions on Knowledge and Data Engineering, **13 (1)** (2001), 124–141.
- P. Xu and S. Tirthapura: *A Lower Bound on Proximity Preservation by Space Filling Curves*, Proc. International Parallel and Distributed Processing Symposium 2012, (2012), 1295 – 1305.
- P. Xu and S. Tirthapura: *On the Optimality of Clustering Properties of Space Filling Curves*, Proc. Principles of Database Systems 2012, (2012), 215–224.

Time Series References

- C. Bandt: *Permutation Entropy and Order Patterns in Long Time Series*, Time Series Analysis and Forecsting, Springer, 2016.
- D. DeFord and K. Moore: *Random Walk Null Models for Time Series Data*, Entropy, 19(11), 615, 2017.
- M. Zanin: *Forbidden patterns in financial time series*, Chaos 18 (2008) 013119.
- M. Zanin, L. Zunino, O.A. Rosso, and D. Papo : *Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review*, Entropy 2012, 14, 1553-1577.

That's all..

Thank You!