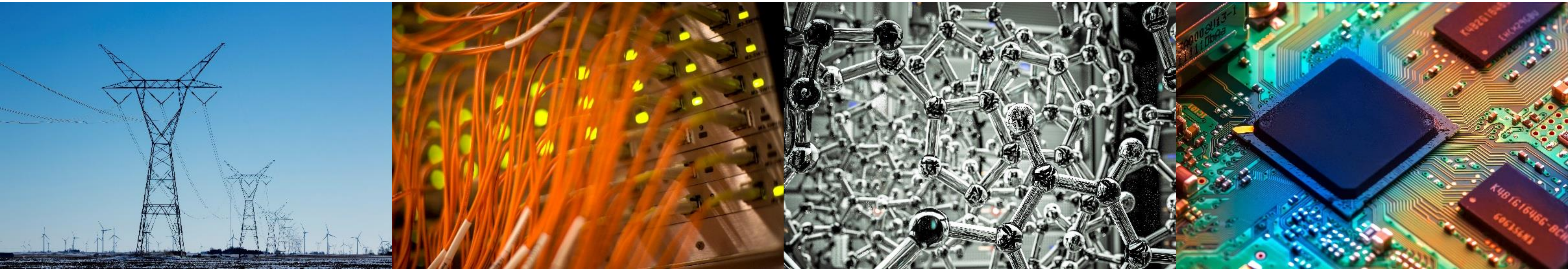
# The Twelvefold Way of Non-Sequential Lossless Compression

Illinois Information Theory Students (FA20), Sourya Basu, and Lav R. Varshney



ILLINOIS
Electrical & Computer Engineering
GRAINGER COLLEGE OF ENGINEERING

# Massively collaborative mathematical projects and consortium authors

polymathprojects.org

## The polymath blog

February 20, 2021

### Polymath projects 2021

Filed under: polymath proposals — Gil Kalai @ 4:11 pm

After the success of Polymath1 and the launching of Polymath3 and Polymath4, Tim Gowers wrote a blog post "Possible future Polymath projects" for planning the next polymath project on his blog. The post mentioned 9 possible projects. (Four of them later turned  to polymath projects.) Following the post and separate posts describing some of the proposed projects, a few polls were taken and a problem – the Erdős discrepancy problem, was selected for the next project polymath5. In *Combinatorics and more* I reviewed some of the proposed projects from 2009, and in the same post I briefly and sometimes vaguely discussed the 2021 list, that I plan to present and discuss in detail in the next couple of months.

Comments (1)

Recent Comments

Thomas Vu on Polymath projects 2021

Hyun seok on A sort of Polymath on a famous…

Polymath16, seventee… on Polymath proposal: finding sim…

Possible future Poly… on Two polymath (of a sort) propo…

Gil Kalai on A sort of Polymath on a fam…

Dan … A sort of

Follow

# Massively collaborative mathematical projects and consortium authors

Illinois Information Theory Students (FA20)

Taha Ameen ur Rahman, Alton S. Barbehenn, Xinan Chen, Hassan Dbouk, James A. Douglas, Yuncong Geng, Ian George, John B. Harvill, Sung Woo Jeon, Kartik K. Kansal, Kiwook Lee, Kelly A. Levick, Bochao Li, Ziyue Li, Yashaswini Murthy, Adarsh Muthuveeru-Subramaniam, S. Yagiz Olmez, Matthew J. Tomei, Tanya Veeravalli, Xuechao Wang, Eric A. Wayman, Fan Wu, Peng Xu, Shen Yan, Heling Zhang, Yibo Zhang, Yifan Zhang, Yibo Zhao

# Order and disorder in data compression

- Several important applications produce huge amounts of data that are not simple sequences of distinguishable symbols, but have certain equivalence classes for patterns of symbols

- Often the ordering of sequences of scientific data records are irrelevant and so only the histogram or type class must be represented [Varshney and Goyal, 2006; Reznik, 2011; Gripon, et al., 2012; Steinruecken, 2015; Orlitsky, et al., 2006]
- Biological data [Ginart, et al., 2018], social graph structure [Choi and Szpankowski, 2012], neural network architecture [Basu and Varshney, 2017]
- Some non-sequential sources also arise in computer science via trace theory, which can be studied explicitly using interchange entropy [Savari, 2004]

- General approach to considering non-sequentiality in data compression is via group theory
- Alternative approach to get explicit bounds/computations for limits of lossless data compression is via enumerative combinatorics

# Rota's Twelvefold Way in enumerative combinatorics

**The twelve combinatorial objects and their enumeration formulas.**

| *f*-class | Any *f* | Injective *f* | Surjective *f* |
|---|---|---|---|
| Distinct<br>$f$ | $n$-sequence in $X$<br>$x^n$ | $n$-permutation of $X$<br>$x^{\underline{n}}$ | composition of $N$ with $x$ subsets<br>$x!\left\{ {n \atop x} \right\}$ |
| $\mathbf{S}_n$ orbits<br>$f \circ S_n$ | $n$-multisubset of $X$<br>$\binom{x+n-1}{n}$ | $n$-subset of $X$<br>$\binom{x}{n}$ | composition of $n$ with $x$ terms<br>$\binom{n-1}{n-x}$ |
| $\mathbf{S}_x$ orbits<br>$S_x \circ f$ | partition of $N$ into $\leq x$ subsets<br>$\sum_{k=0}^{x} \left\{ {n \atop k} \right\}$ | partition of $N$ into $\leq x$ elements<br>$[n \leq x]$ | partition of $N$ into $x$ subsets<br>$\left\{ {n \atop x} \right\}$ |
| $\mathbf{S}_n \times \mathbf{S}_x$ orbits<br>$S_x \circ f \circ S_n$ | partition of $n$ into $\leq x$ parts<br>$p_x(n+x)$ | partition of $n$ into $\leq x$ parts 1<br>$[n \leq x]$ | partition of $n$ into $x$ parts<br>$p_x(n)$ |

# Entropy computations for the twelvefold way

- We characterized the information-theoretic lossless compression limits for a wide range of non-sequentialities governed by the twelvefold way in combinatorics
- Explicit computations for all twelve settings were carried out for i.i.d. uniform and Bernoulli distributions
- Comparisons among settings provide quantitative insight into which equivalences reduce entropy the most
- See the paper for details [https://arxiv.org/abs/2011.04069]