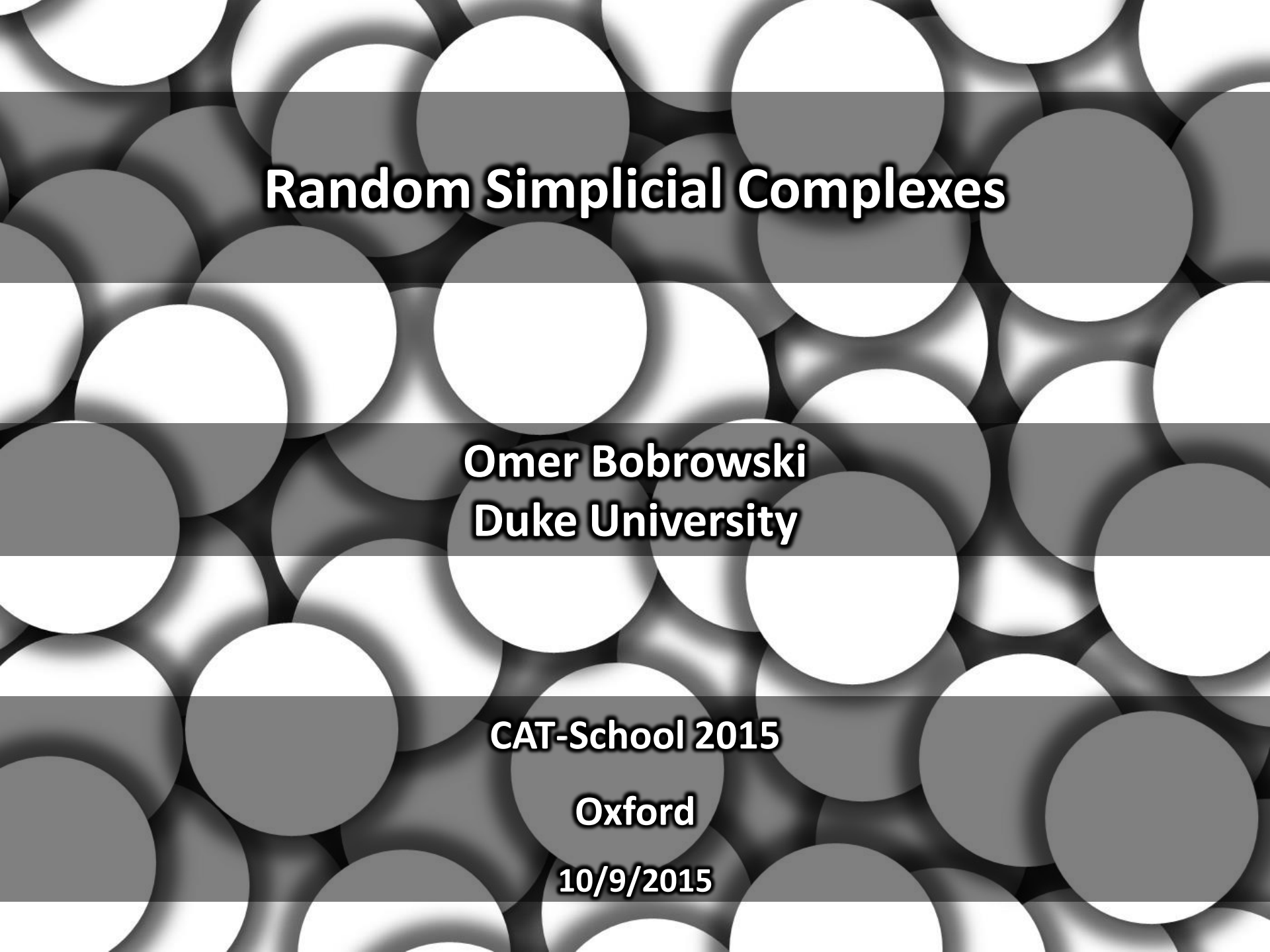


# Random Simplicial Complexes

Omer Bobrowski
Duke University

CAT-School 2015

Oxford

10/9/2015

# Part III
# Extensions & Applications

# Contents

# Contents

## Morse Theory for the Distance Function

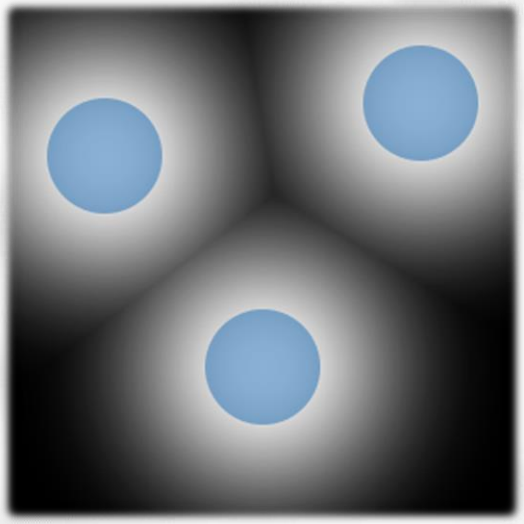
Joint work with Robert Adler

## Persistent Homology and Maximal Cycles

# The Distance Function

- Definition: For a finite set $\mathcal{P} \subset \mathbb{R}^d$

$$d_{\mathcal{P}}(x) = \min_{p \in \mathcal{P}} \|x - p\|_2 , \quad x \in \mathbb{R}^d$$



- Sublevel sets = $d_{\mathcal{P}}^{-1}((-\infty, r/2]) = \mathcal{U}(\mathcal{P}, r) \simeq \mathcal{C}(\mathcal{P}, r)$
- Morse Theory: homology $\leftrightarrow$ critical points
- Problem: $d_{\mathcal{P}}$ is not everywhere differentiable

# Critical Points of the Distance Function

- Gershkovich & Rubinstein 97 – *Morse theory for min-type functions*
- Example: $\mathcal{P} = \{p_1, p_2, p_3\} \subset \mathbb{R}^2$

$$d_{\mathcal{P}} : \mathbb{R}^2 \to \mathbb{R}$$

**Index 0 (minimum)**

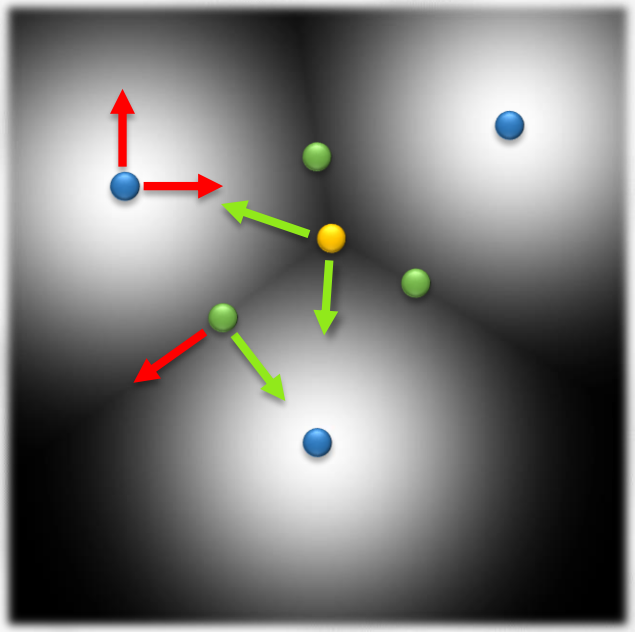
Generated by 1 point

**Index 1 (saddle)**

Generated by 2 points

**Index 2 (maximum)**

Generated by 3 points



increasing

decreasing

Index $k$ critical points are "generated" by subsets of $k+1$ points

# Morse Theory for the Distance Function

- Study how the homology of $d_{\mathcal{P}}^{-1}((-\infty, r])$ changes as $r : 0 \to \infty$
- Morse Theory: Homology is changed only at critical values
- Critical point of index $k$ → create $H_k$ (birth) – OR – destroy $H_{k-1}$ (death)

# Setup & Goals

## Setup:

- $\mathcal{X}_n = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f$ - a set of random points in $\mathbb{R}^d$
- $\beta_k = \beta_k(n, r)$ - the $k$-th Betti number of $\mathcal{U}(n, r) \simeq \mathcal{C}(n, r)$
- $N_k = N_k(n, r)$ - number of index-$k$ critical points $p$ of $d_{\mathcal{X}_n}$ with $p \in \mathcal{U}(n, r)$

## Goals:

- Limiting behavior of $N_k$ as $n \to \infty, r \to 0$
- Conclusions about $\beta_k$

# The Subcritical Regime ( $\Lambda \longrightarrow 0$ )

- **Reminder:** most $k$-cycles are generated by $k+2$ points $\Rightarrow \beta_k \approx n\Lambda^{k+1}\mu_k^C$
- **Here:** all $k$-critical points are generated by $k+1$ points
- Similar counting arguments yield

**Theorem [B. & Adler, 14]**

For $1 \leq k \leq d$

$$\mathbb{E}\{N_k\} \approx \mathrm{Var}(N_k) \approx n\Lambda^k \mu_k^M,$$

where

$$\mu_k^M := \frac{1}{(k+1)!}\int h_1(0,\mathbf{y})d\mathbf{y}$$

and

$$h_r(\mathbf{x}) := \mathbb{1}\{\mathbf{x} \text{ generates a crit. pt.}\}$$

- Other limit theorems can be proved as well

# The Critical Regime ( $\Lambda \to \lambda$ )

- **Reminder:** $k$-cycles are generated by at least $k+2$ points $\longrightarrow$ intractable
- **Here:** $k$-critical points are generated by exactly $k+1$ points

**Theorem [B. & Adler, 14]**

For $1 \leq k \leq d$, if $\Lambda \to \lambda \in (0, \infty)$ then

$$\mathbb{E}\{N_k\} \approx n\gamma_k(\lambda),$$

where

$$\gamma_k(\lambda) := \frac{\lambda^k}{(k+1)!} \int h_1(0, \mathbf{y}) e^{-\lambda R^d(0, \mathbf{y})} d\mathbf{y} dx.$$

- CLT and LLN are also proved

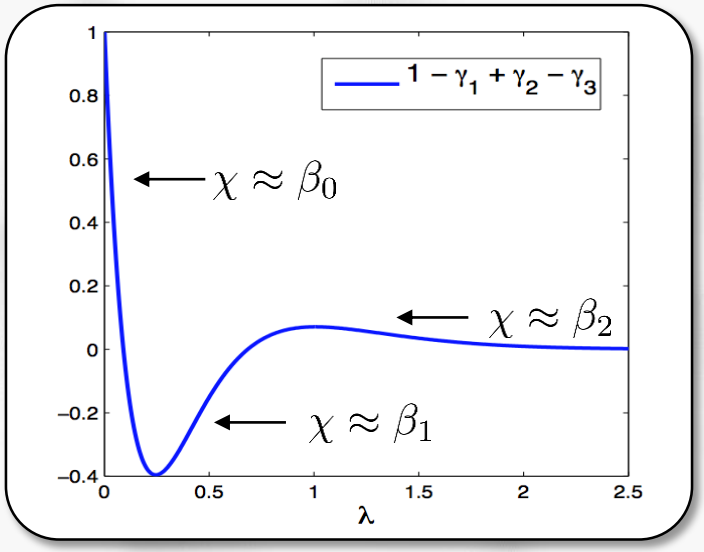
# The Euler Characteristic

- The Euler characteristic: $\chi = \sum_{k=0}^{d-1} (-1)^k \beta_k$
- Morse Theory: $\chi = \sum_{k=0}^{d} (-1)^k N_k$
- Conclusion:

$$\mathbb{E}\{\chi\} \approx n \left( 1 + \sum_{k=1}^{d} (-1)^k \gamma_k(\lambda) \right)$$

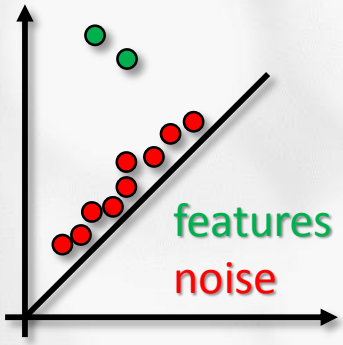
- Example ($d$=3):

$$\chi = \beta_0 - \beta_1 + \beta_2$$



# A Word on Persistent Homology

- Most “noisy” cycles are born and die in the critical regime ( $\sim n$ )
- Critical points analysis might tell us something…
- Example - critical points density:

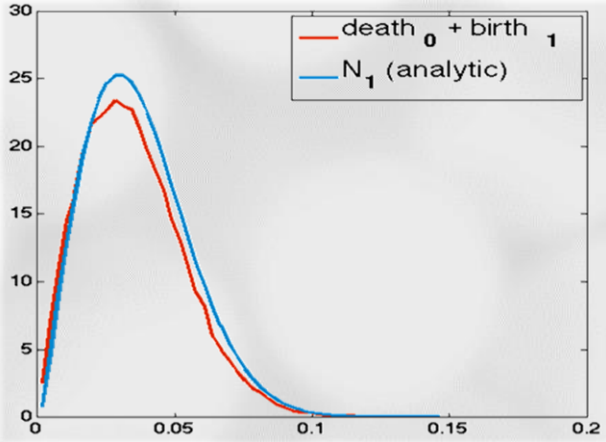


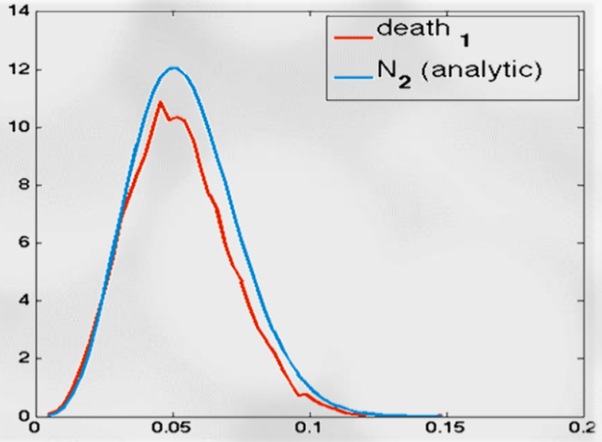

Take a random sample on the unit box in $\mathbb{R}^2$

Compute (analytically) the density of critical points: $\frac{d}{dr}\mathbb{E}\{N_k(r)\} \approx n\frac{d}{dr}\gamma_k(r)$

$N_1 = \#death_0 + \#birth_1$

$N_2 = \#death_1$




$n$=200

# The Supercritical Regime ( $\Lambda \to \infty$ )

- Recall: Uniform distribution on the torus $\mathbb{T}^d$
- Define: $\Delta N_k(r) = N_k(\infty) - N_k(r)$ (the critical points "we didn't use yet")
- Morse theory:

$$\beta_k(r) \leq \beta_k(\mathbb{T}^d) + \Delta N_{k+1}(r)$$

(all "small" cycles must get killed)

**Theorem [B. & Weinberger, 15]**

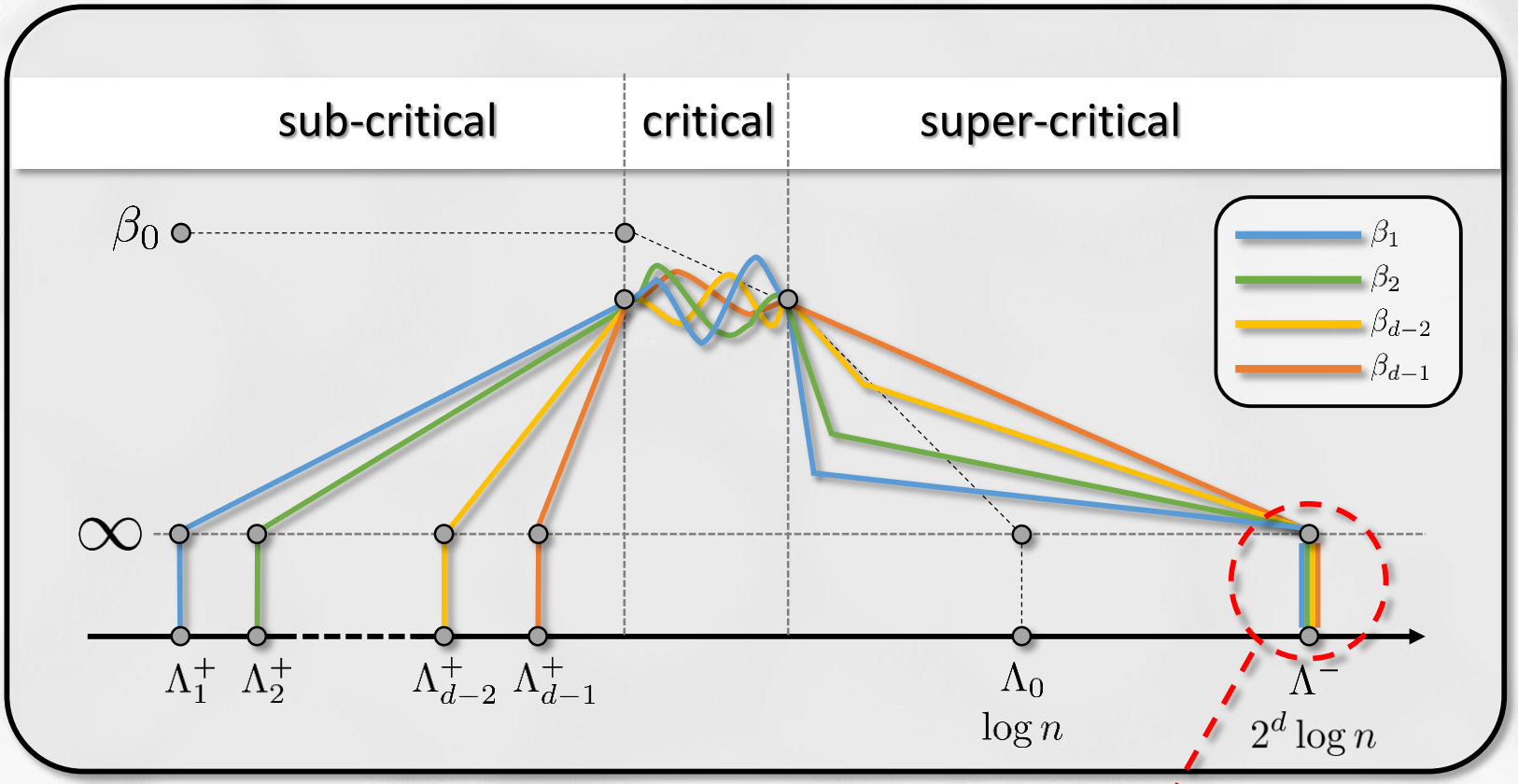
For $1 \leq k \leq d$, if $\Lambda \to \infty$ then

$$\mathbb{E}\{\beta_k(r)\} = \beta_k(\mathbb{T}^d) + O(n\Lambda^k e^{-\Lambda/2^d})$$

- Threshold:

$$\Lambda_k^- := 2^d(\log n + k \log\log n) \Rightarrow H_k(\mathcal{C}(n,r)) \cong H_k(\mathbb{T}^d)$$

# Summary




$$\Lambda_k^- = 2^d(\log n + k \log\log n)$$

$$H_k(\mathcal{C}(n,r)) \cong H_k(\mathbb{T}^d)$$

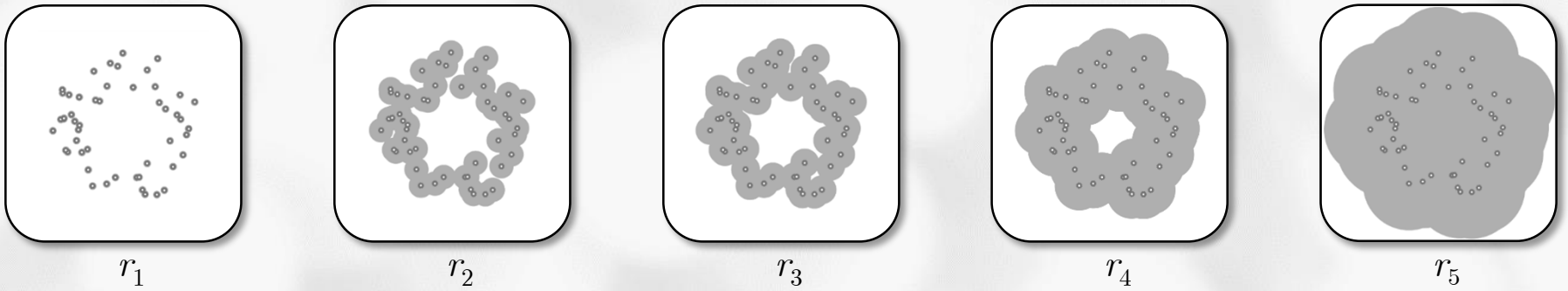
# Contents

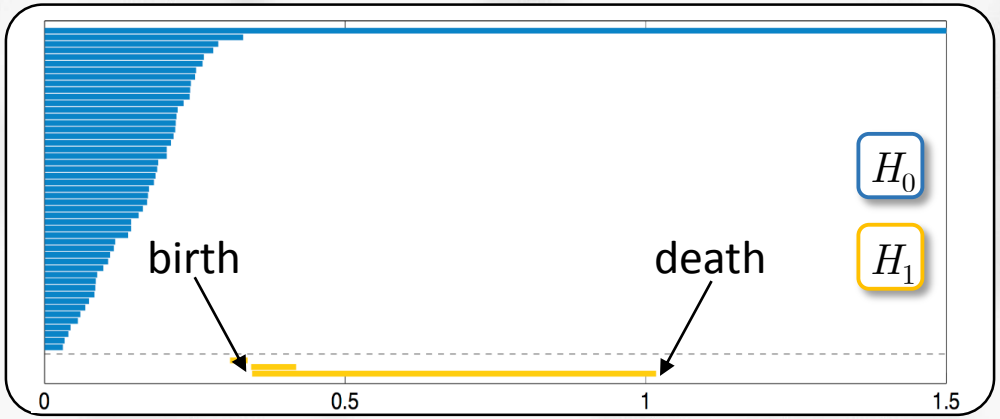
## Morse Theory for the Distance Function

## Persistent Homology and Maximal Cycles

Joint work with Matthew Kahle and Primoz Skraba

# Reminder - Persistent Homology



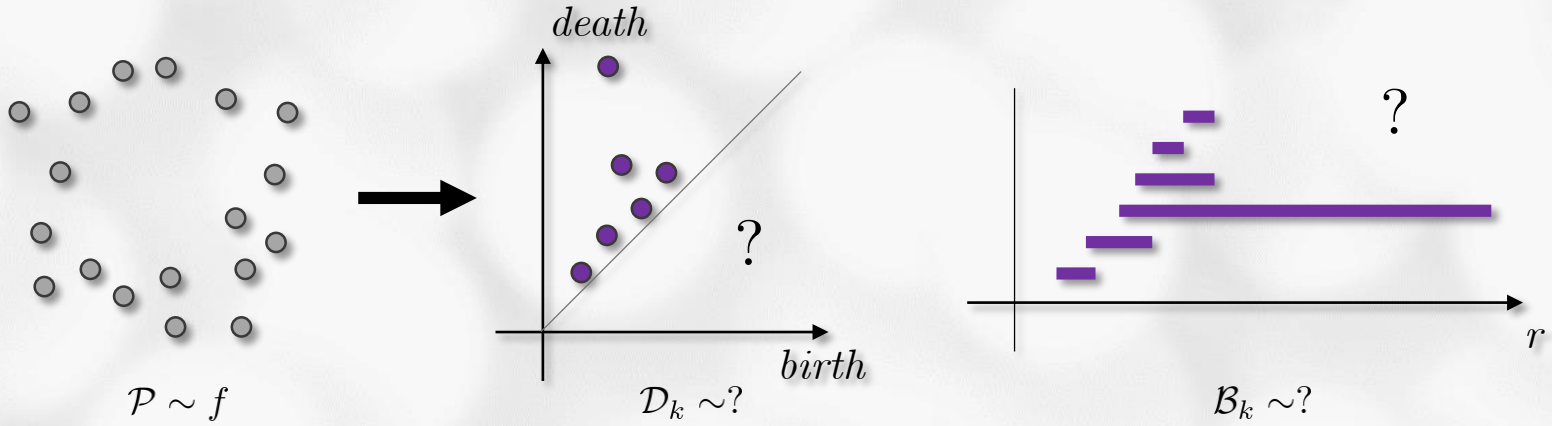




- Persistent (long bar) = significant

# The Ultimate Goal

- $\mathcal{P}$ = a random point process, generated by a probability density $f$
- **Goal:**

  Find the distribution of barcodes / persistence diagrams

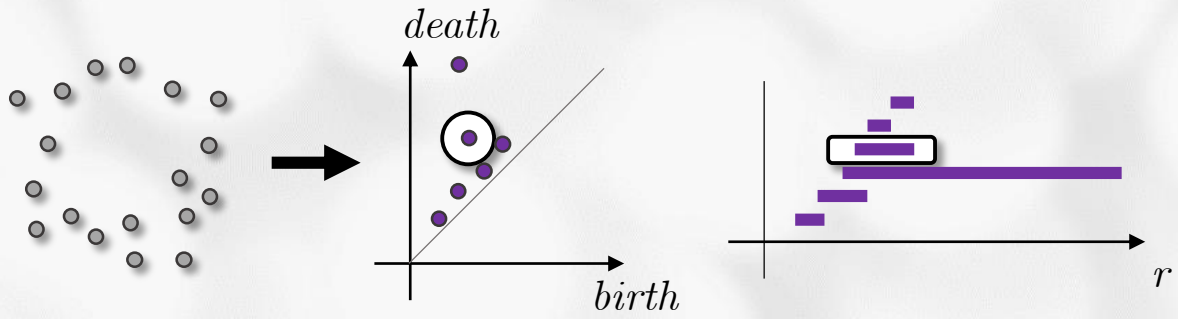



- Problem:

  The mappings $\mathcal{P} \to \mathcal{B}_k$, $\mathcal{P} \to \mathcal{D}_k$ are extremely difficult to analyze ☹

  (points in metric space $\to$ abstract algebraic structure)

# A Simpler Problem




- Barcodes / persistence diagram are believed to be of the form

**signal**

cycles in the “true” underlying space

longest bars
points away from the diagonal

+

**noise**

artefacts of the sampling

short bars
points near the diagonal

- What can we say about the maximal persistence of cycles in the “noise”?
- $\Rightarrow$ threshold for filtering the noise

## Setup

- $\mathcal{X}_n$ - uniformly distributed in $\mathcal{Q}^d = [0,1]^d$ (noise only)
- $\mathrm{PH}_k(n)$ - the $k$-th persistent homology of $\{\mathcal{C}(n,r)\}_{r=0}^{\infty}$
- For every cycle $\gamma \in \mathrm{PH}_k(n)$ we define: $\gamma_{birth}$, $\gamma_{death}$
- "Traditional" persistence: $\gamma_{death} - \gamma_{birth}$ (= length of the bar)
- Our definition:

$$\pi(\gamma) = \frac{\gamma_{death}}{\gamma_{birth}}$$

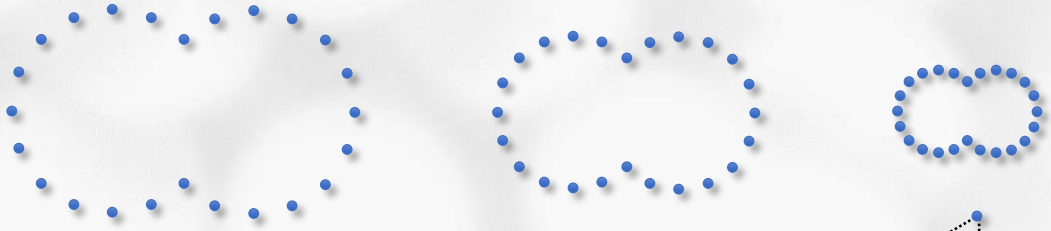
- Maximal persistence:

$$\Pi_k(n) := \max_{\gamma \in \mathrm{PH}_k(n)} \pi(\gamma).$$

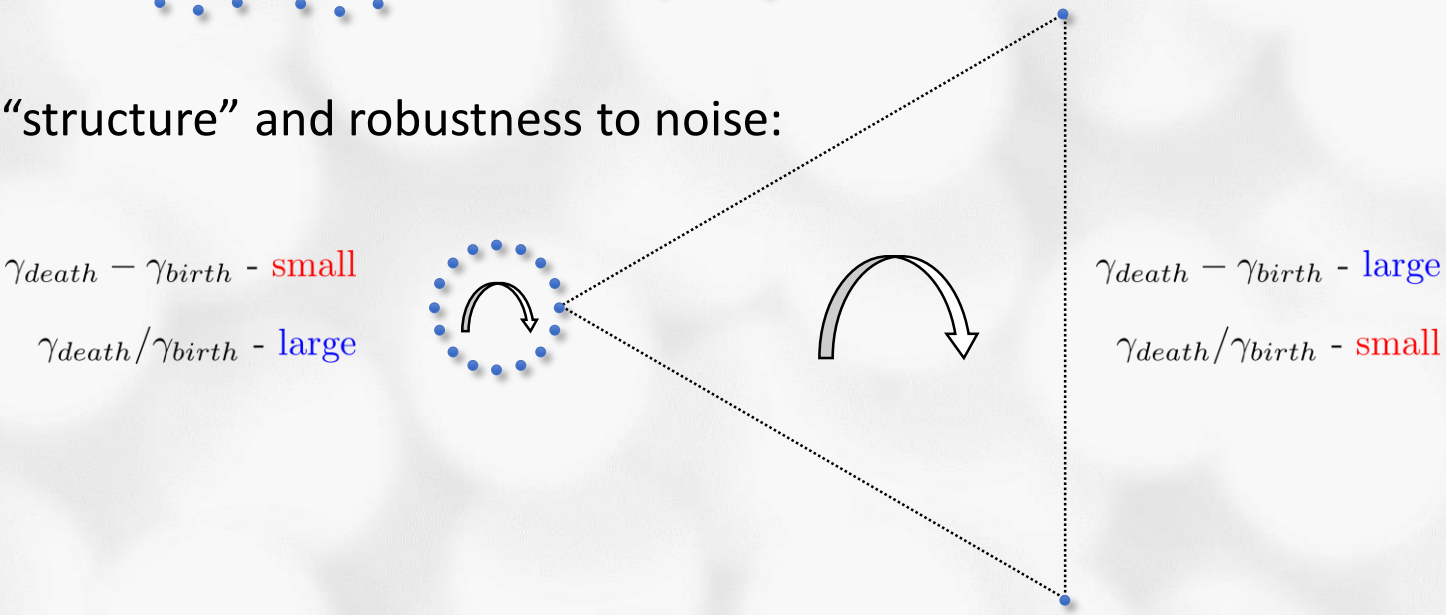
# Multiplicative Persistence

$$\pi(\gamma) = \frac{\gamma_{death}}{\gamma_{birth}}$$

- Scale invariance:



- Preference for “structure” and robustness to noise:



# Main Result

**Theorem [B, Kahle, Skraba]**

Let $\mathrm{PH}_k(n)$ be the $k$-th persistent homology of $\mathcal{C}(n,r)$.

Define

$$\Pi_k(n) := \max_{\gamma \in \mathrm{PH}_k(n)} \pi(\gamma),$$

and

$$\Delta_k(n) = \left(\frac{\log n}{\log\log n}\right)^{1/k}.$$

Then $\exists A_k, B_k > 0$, such that

$$A_k \Delta_k(n) \leq \Pi_k(n) \leq B_k \Delta_k(n) \quad w.h.p.$$

(w.h.p. = with high probability = $\mathbb{P} \longrightarrow 1$)

# Upper Bound – Main Idea

- **Main idea:**

  Show that large cycles require (too) large connected components.

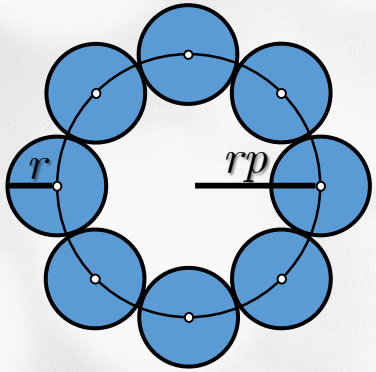
**Lemma**

Let $\gamma \in \mathrm{PH}_k(n)$ with $\gamma_{birth} = r$ and $\pi(\gamma) = p$.

Then $\exists C_k > 0$ s.t. $\mathcal{C}(n,r)$ has a component with at least $C_k p^k$ vertices.

- **Some intuition:**

  - $k$=1: Given $\gamma_{birth} = r$, $\gamma_{death} = r \cdot p$, the cycle with the fewest vertices is




$$\Rightarrow \quad \#\text{vertices} \approx \frac{2\pi rp}{2r} = \pi p$$

  - Higher $k$ :

    Cover a $k$-sphere of radius $rp$ with balls of radius $r$ $\Rightarrow$ $\#\text{vertices} \sim \frac{(rp)^k}{r^k} = p^k$

# Upper Bound – Key Steps

$$\pi(\gamma) = \gamma_{death}/\gamma_{birth}, \quad \gamma_{birth} = r$$

$$\gamma_{death} \leq C\left(\frac{\log n}{n}\right)^{1/d} \quad \text{(coverage)}$$

To get a "large" k-cycle

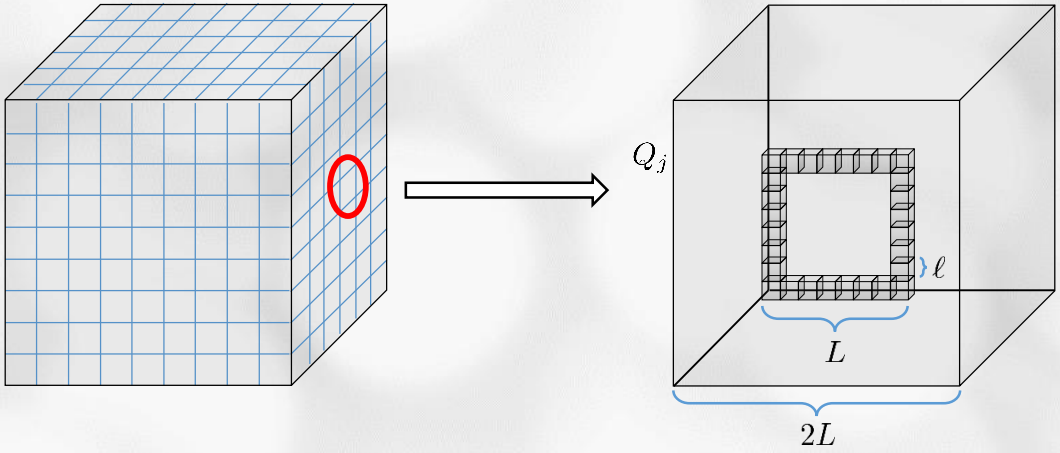
- a large component in $\mathcal{C}(n,r)$
- $\gamma_{birth}$ - small
  - $\mathcal{C}(n,r)$ - sparse

$$\pi(\gamma) \leq B_k \Delta_k(n)$$

# Lower Bound – Main Idea

- **Goal:** Show that there exists a cycle with $\pi(\gamma) \geq A_k \Delta_k(n)$ (for some $A_k > 0$)

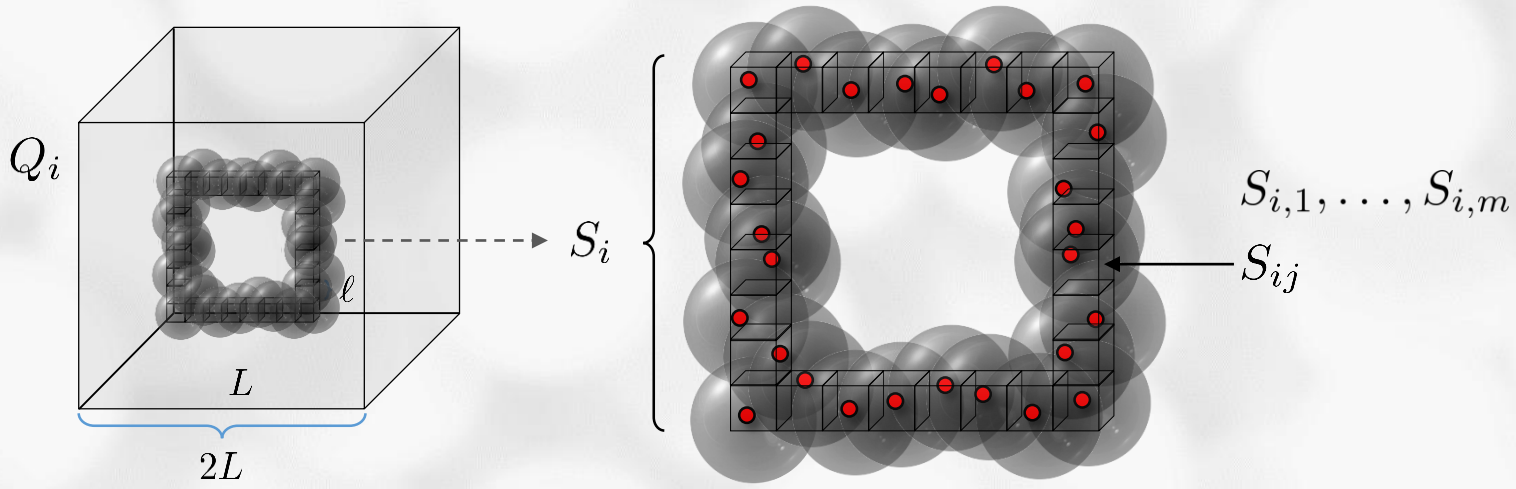
- **Main Idea:** Split the cube $[0,1]^d$ into small cubes $Q_1, \ldots, Q_M$




In each small cube - look for a "floating" $k$-cycle that

1. covers the boundary of a $(k+1)$-box
2. is disconnected from the rest of the complex
3. its persistence $\pi(\gamma) \sim L/\ell$ is large

# Lower Bound – Floating Cycles




- **Claim:** A single point in each $S_{ij}$ + nothing else in $Q_i \Rightarrow$ a k-cycle $\gamma$ with

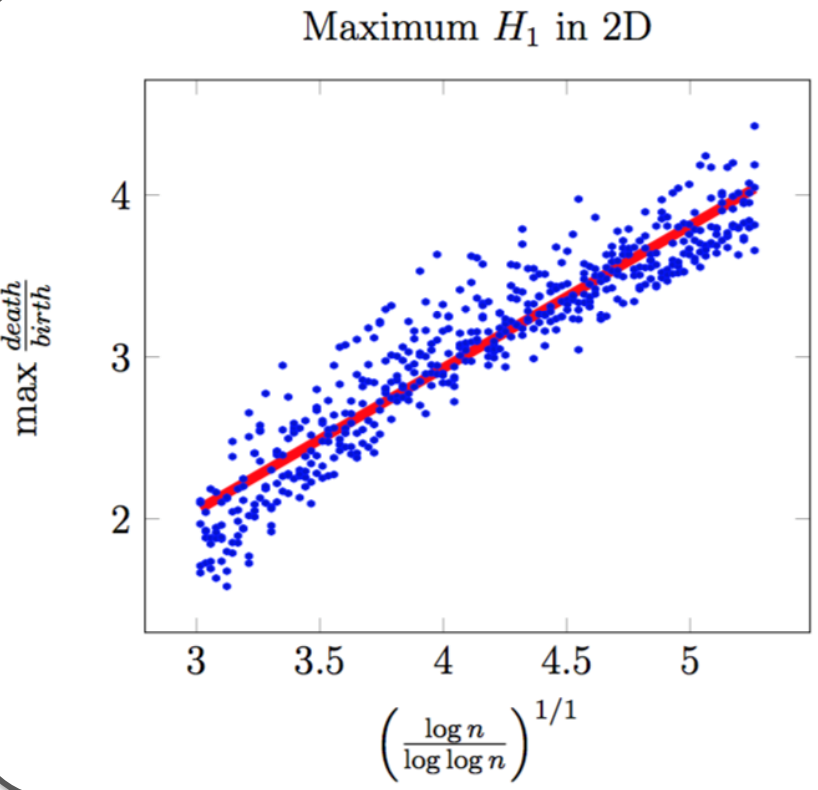
$$\pi(\gamma) \geq \frac{1}{4\sqrt{d}} \times \frac{L}{\ell}$$

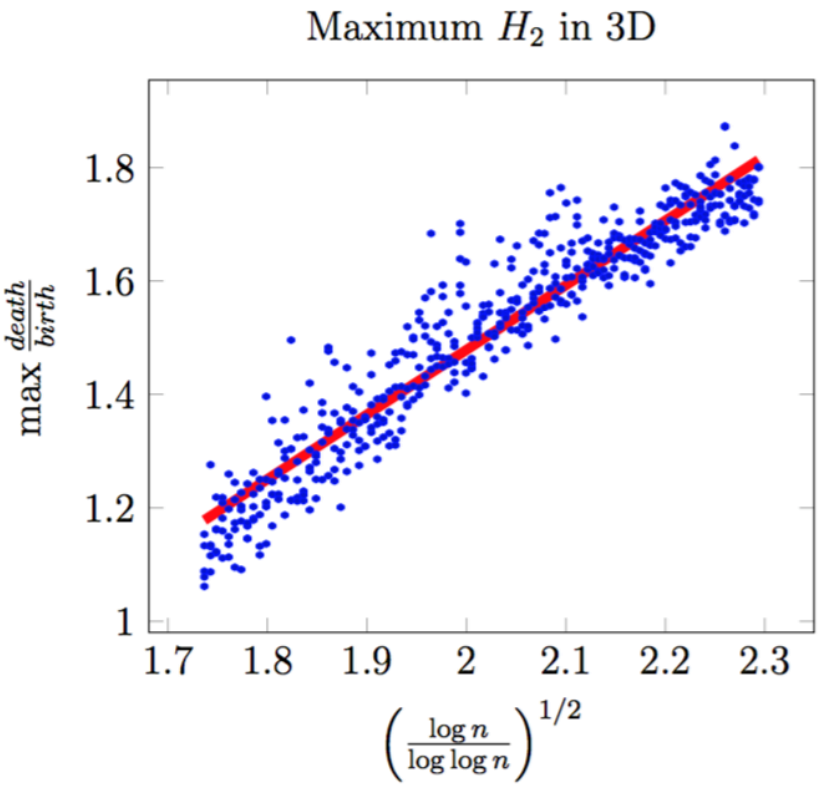
- **Why?**
  - The balls of radius $r = \sqrt{d}\ell$ cover $S_i$ completely $\Rightarrow$ we see the k-cycle
  - The balls of radius $r = L/4$ still do not fill in the hole, and do not connect outside $Q_i$

$$\Rightarrow \gamma_{birth} \leq \sqrt{d}\ell, \quad \gamma_{death} \geq L/4$$

# Simulations

$$\Pi_k(n) \sim \left(\frac{\log n}{\log\log n}\right)^{1/k}$$




$$n = 1,000,000 \quad \Rightarrow \quad \frac{\log n}{\log\log n} \approx 5.3 \quad \left(\frac{\log n}{\log\log n}\right)^{1/2} \approx 2.3$$

# Summary

- The maximal noisy cycle:

$$\Pi_k(n) \sim \Delta_k(n) = \left(\frac{\log n}{\log\log n}\right)^{1/k} \to \infty$$

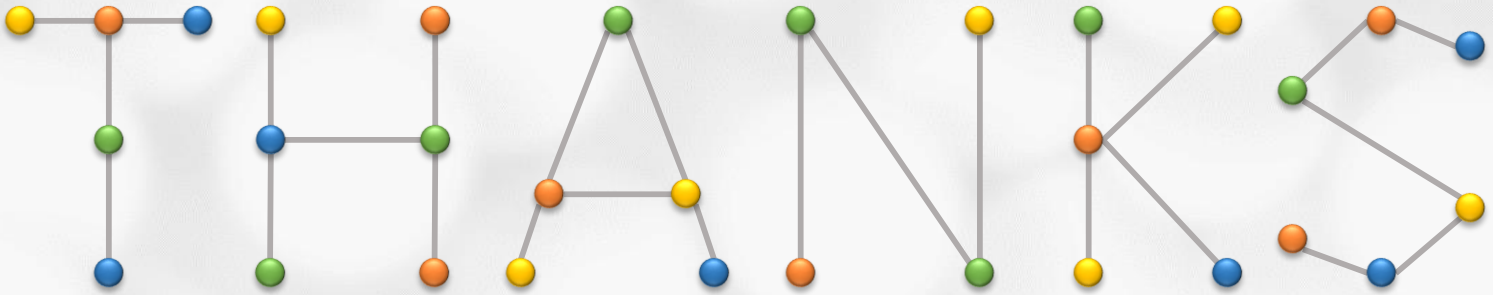
- "True" cycles:

$$\gamma_{death} = const$$

$$\gamma_{birth} \leq const \left(\frac{\log n}{n}\right)^{1/d} \quad \text{(coverage)} \qquad \Rightarrow \quad \pi(\gamma) \geq \left(\frac{n}{\log n}\right)^{1/d}$$

- Conclusion ("SNR"):

$$\frac{\pi(\gamma)}{\Pi_k(n)} \geq C n^{\frac{1}{d}-\epsilon} \to \infty$$


THANKS