The magnitude of graphs and finite metric spaces

Tom Leinster

Edinburgh

Plan

1. Background
2. The magnitude of a finite set of points
3. Diversity
4. The magnitude of a graph
5. The future: magnitude homology

1. Background

Size

For many types of mathematical object, there is a canonical notion of size.

- Sets have cardinality. It satisfies

$$|A \cup B| = |A| + |B| - |A \cap B|$$
$$|A \times B| = |A| \times |B| .$$

- Subsets of $\mathbb{R}^n$ have volume. It satisfies

$$\mathrm{vol}(A \cup B) = \mathrm{vol}(A) + \mathrm{vol}(B) - \mathrm{vol}(A \cap B)$$
$$\mathrm{vol}(A \times B) = \mathrm{vol}(A) \times \mathrm{vol}(B).$$

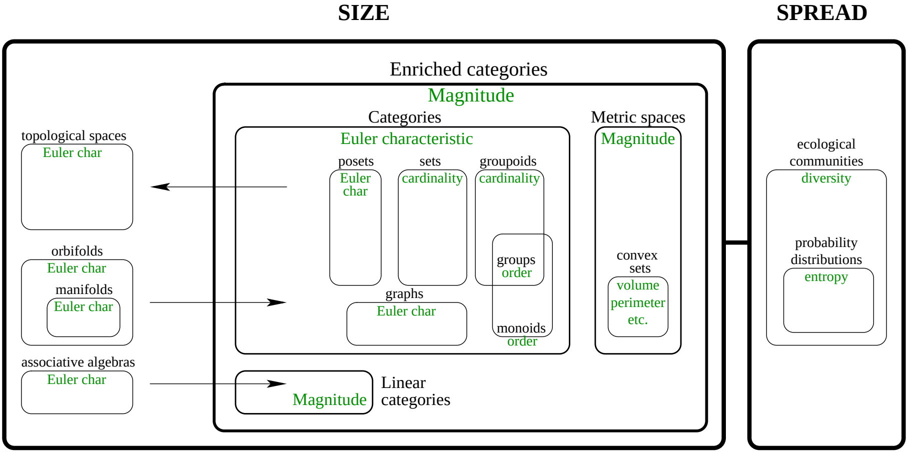
- Topological spaces have Euler characteristic. It satisfies

$$\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B) \quad \text{(under hypotheses)}$$
$$\chi(A \times B) = \chi(A) \times \chi(B).$$

Challenge Find a general definition of ‘size’, including these and other examples.

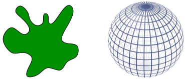
One answer The magnitude of an “enriched category”.

The wide world of magnitude



The magnitude of a compact metric space

Let A be a compact metric space, e.g. a closed bounded subset of $\mathbb{R}^n$.



The magnitude $|A|$ of A is a real number measuring the 'size' of A. (Definition later.)

Olaf, yesterday: 'There is no privileged scale!' So...

Given $t > 0$, write tA for A scaled up by a factor of t.

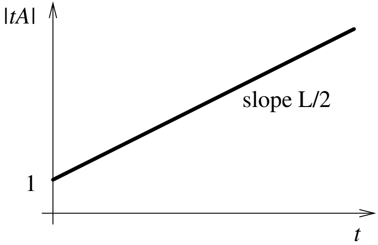
The magnitude function of A is the function $t \mapsto |tA|$.

Thus, magnitude assigns to each space not just a *number*, but a *function*.

The magnitude of a line segment

Example: Let A be a straight line of length L.

The magnitude function of A is



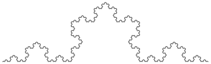
Euler characteristic, length, dimension

$$t \mapsto |tA| = \boxed{1} + \tfrac{1}{2}\boxed{L} \cdot t^{\boxed{1}}$$

The magnitude of a compact metric space: theorems

Let A be a compact subset of $\mathbb{R}^n$.

Theorem (Meckes) *The asymptotic growth rate of $|tA|$ as $t \to \infty$ is the Minkowski dimension of A.*

E.g. $\Big| t$  $\Big|$ grows like t^2 and $\Big| t$  $\Big|$ grows like $t^{1.261\ldots}$, for large t.

Theorem (Barceló and Carbery; Gimperlein and Goffeng) *Under technical hypotheses,*

$$|tA| = c_n \operatorname{vol}_n(A) \cdot t^n + c_{n-1} \operatorname{vol}_{n-1}(\partial A) \cdot t^{n-1} + O(t^{n-2})$$

as $t \to \infty$, where c_n and c_{n-1} are known constants.

E.g. If $n = 3$ then $\operatorname{vol}_n(A)$ and $\operatorname{vol}_{n-1}(\partial A)$ are the volume and surface area of A.

So: if you know the magnitude function of a space, you know its dimension, volume and surface area.

The magnitude of a compact metric space: theorems

Theorem (Willerton) *Let A be a homogeneous Riemannian n-manifold. Then*

$$|tA| = C_n \,\mathrm{vol}_n(A) \cdot t^n + C_{n-2}\mathrm{TotalScalarCurvature}(A) \cdot t^{n-2} + O(t^{n-4})$$

as $t \to \infty$, where C_n and C_{n-2} are known constants. In particular, when $n = 2$,

$$|tA| = \tfrac{1}{2\pi}\mathrm{area}(A) \cdot t^2 + \chi(A) + O(t^{-2}).$$

Theorem (Barceló and Carbery) *The magnitude of an odd-dimensional Euclidean ball is a rational function of its radius. Specifically:*

$$\left|tB^1\right| = 1 + t,$$
$$\left|tB^3\right| = \frac{1}{3!}(6 + 12t + 6t^2 + t^3),$$
$$\left|tB^5\right| = \frac{1}{5!}\,\frac{360 + 1080t + 525t^2 + 135t^4 + 18t^5 + t^6}{3 + t}.$$

The moral

For geometrically interesting subsets of $\mathbb{R}^n$, the magnitude function conveys geometrically interesting information.

(Despite—or because of?—its very general, abstract categorical origins.)

What information does the magnitude function contain for *finite* sets of points?

2. The magnitude of a finite set of points

The definition

Let A be a finite metric space with points $1, \ldots, n$ and distance d_{ij} from point i to point j.

Write $Z_A = Z$ for the $n \times n$ matrix with entries

$$Z_{ij} = e^{-d_{ij}}.$$

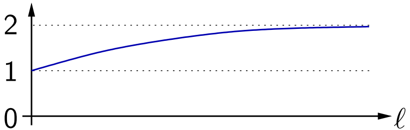
Assuming Z is invertible (which it usually is), the magnitude of A is

$$|A| = \sum_{i,j} (Z^{-1})_{ij}$$

—the sum of all n^2 entries of Z^{-1}.

First examples

- $|\emptyset| = 0$.
- $|\bullet| = 1$.
- $\left|\bullet\!\overset{\leftarrow\ \ell\ \rightarrow}{}\!\bullet\right| = \text{sum of entries of } \begin{pmatrix} e^{-0} & e^{-\ell} \\ e^{-\ell} & e^{-0} \end{pmatrix}^{-1} = \dfrac{2}{1+e^{-\ell}}$.



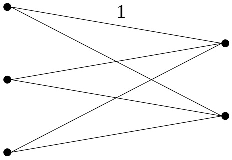
- If $d(i,j) = \infty$ for all $i \neq j$ then $|A| = n$ (number of points).

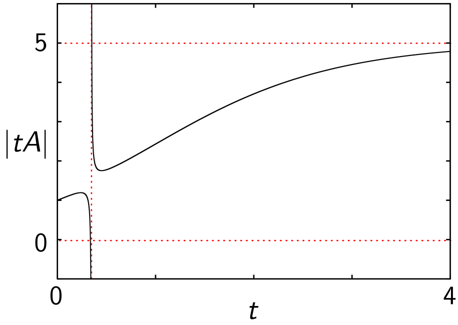
Slogan: Magnitude is the 'effective number of points'
(or clusters, or modules, . . .)

Magnitude functions

Let A be a finite metric space. The magnitude function of A is the (partially-defined) function $t \mapsto |tA|$ $(t > 0)$.

Example:





Properties:

- The magnitude function has only finitely many singularities (none if $A \subseteq \mathbb{R}^n$)
- $\lim_{t\to\infty} |tA|$ is equal to n, the number of points
- $|tA|$ is increasing in t for $t \gg 0$.

Detecting clusters at different scales (Willerton)

Magnitude: 1

Detecting clusters at different scales (Willerton)

Magnitude: 1.01

Detecting clusters at different scales (Willerton)

Magnitude: 1.2

Detecting clusters at different scales (Willerton)

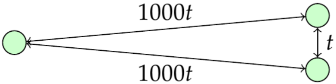
Magnitude: 1.6

Detecting clusters at different scales (Willerton)

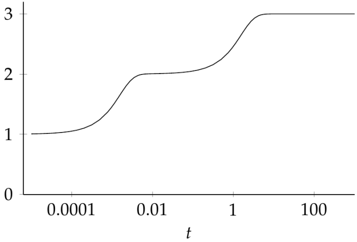
Magnitude: 2.3

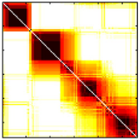
As the points get further apart, the magnitude gets closer to 3.

Precise version: the magnitude function of the 3-point space



is




Rubinov & Sporns

Observation

Magnitude can be seen as ‘effective number of clusters’, but it’s not always an integer! Awkward reality: ‘clusters’ are ill-defined.

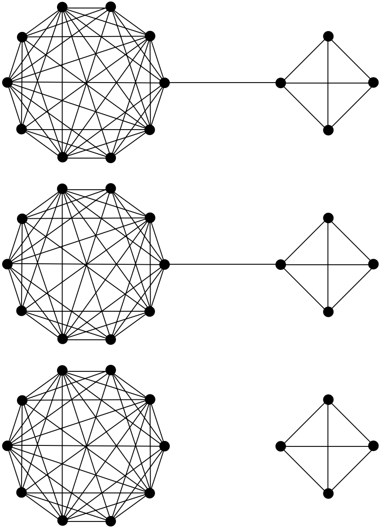
Detecting the critical nodes and edges

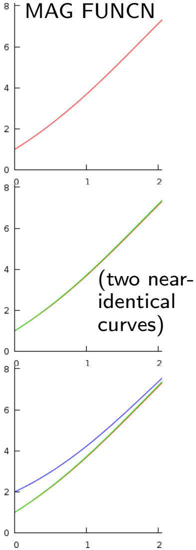
View a graph as a metric space: the points are the nodes, and the distance between nodes is the length of a shortest path between them.

whole graph

one edge on left deleted

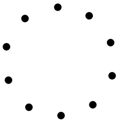
bridge deleted

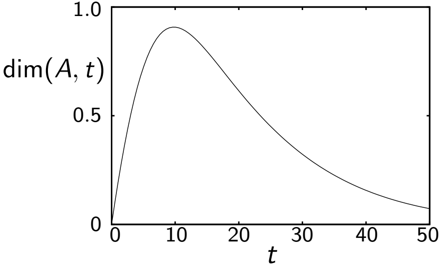


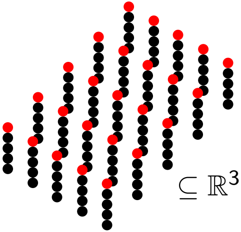


Detecting dimension at multiple scales

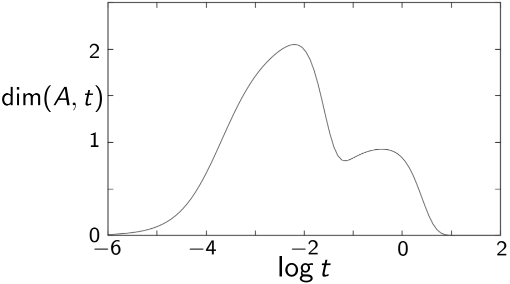
Definition (Willerton) The dimension of A at scale t is $\dim(A, t) = \frac{d(\log|tA|)}{d(\log t)}$.

E.g. $A =$ 



E.g. $A \approx$  $\subseteq \mathbb{R}^3$

grid is $100 \times 100 \times 100$,
with 100:1 ratio of
horizontal:vertical spacing.



An appeal

The only ‘results’ we have on the ability of magnitude functions to detect features of data-sets are a handful of specific examples. We need:

1. More empirical exploration
2. Theorems. . . or at least, conjectures!

Back to compact spaces

There are several equivalent ways to define the magnitude of a compact metric space X (assuming a technical hypothesis).

The simplest:

$$|X| = \sup\{|A| \ : \ \text{finite } A \subseteq X\}.$$

With this definition, and lots of analysis, we get all the results on geometric invariants stated earlier.

3. Diversity

Joint with Christina Cobbold (*Ecology*, 2012)

A brief history of diversity measurement

Challenge Given a biological community, derive a single real number measuring its 'diversity' (whatever *that* means).

There are practical problems . . . which we'll ignore.

There are statistical problems . . . which we'll ignore.

There are conceptual problems . . . which we'll focus on.

Some conceptual questions:

- How much importance to attach to rare species?
  E.g. there are 8 species of great ape, but 99.99% are humans.
- How to incorporate the varying similarities between species?
  E.g. 10 species of pine vs. 10 very different tree species.

Lots of measures of diversity have been proposed. . .

688 NATURE April 30, 1949 Vol. 163

Measurement of Diversity

THE 'characteristic' defined by Yule[1] and the 'index of diversity' defined by Fisher[2] are two measures of the degree of concentration or diversity

The third and fourth cumulants of the distribution of l have also been calculated exactly. They indicate that as N increases, the distribution tends to normality except when $\lambda = 1/Z$; in that case the distribution of

VEGETATION OF THE SISKIYOU MOUNTAINS, OREGON AND CALIFORNIA[1]

R. H. WHITTAKER

Biology Department, Brooklyn College, Brooklyn 10, N. Y.

Lots of measures of diversity have been proposed. . .

688 NATURE April 30, 1949 Vol. 163

THE NONCONCEPT OF SPECIES DIVERSITY: A CRITIQUE AND ALTERNATIVE PARAMETERS[1]

STUART H. HURLBERT[2]

Division of Biological Control, Department of Entomology, University of California, Riverside

Abstract. The recent literature on species diversity contains many semantic, conceptual, and technical problems. It is suggested that, as a result of these problems, species diversity has become a meaningless concept, that the term be abandoned, and that ecologists take a

THROUGH THE JUNGLE OF BIOLOGICAL DIVERSITY

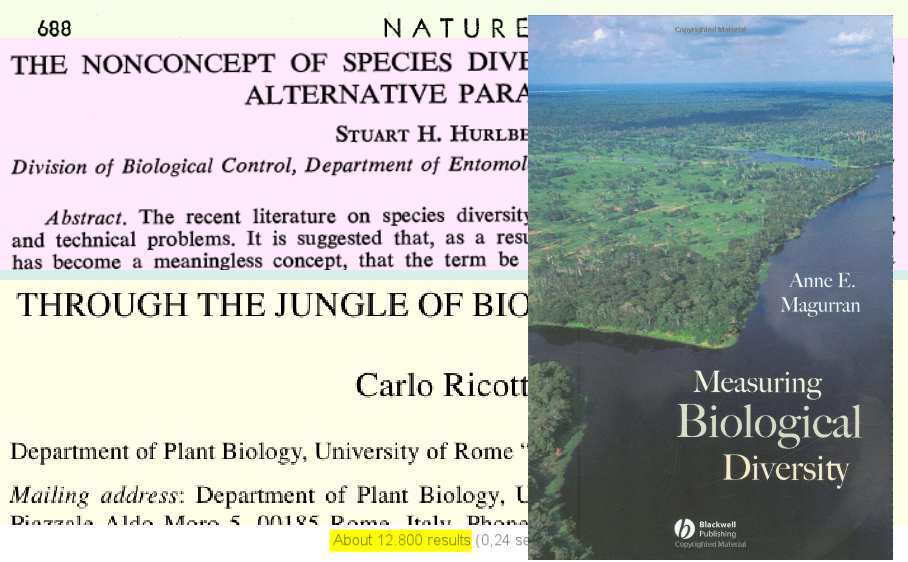
Carlo Ricotta

Department of Plant Biology, University of Rome "La Sapienza", Rome, Italy.

Mailing address: Department of Plant Biology, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy. Phone: +39 06 49912408. Fax: +39 06

About 12.800 results (0,24 seconds)

Lots of measures of diversity have been proposed. . .



We'll meet a family of measures that encompasses many of them.

Modelling a community

Model a biological community mathematically as follows:

- The organisms are classified into n species.
- They have relative abundances $\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$ (with $\sum p_i = 1$).
- The similarity between species i and j is $Z_{ij} \in [0, 1]$.
  Here 0 means totally dissimilar and 1 means identical.
  Assume $Z_{ii} = 1$ and $Z_{ij} = Z_{ji}$, so have symmetric $n \times n$ matrix Z.

Similarity can be measured in many ways, including:

- Naive model $Z = I$: distinct species have nothing in common.
- Percentage genetic similarity.
- Taxonomically, e.g. $Z_{ij} = \begin{cases} 1 & \text{if same species} \\ 0.7 & \text{if different species but same genus} \\ 0 & \text{otherwise.} \end{cases}$
- $Z_{ij} = e^{-d_{ij}}$ if (d_{ij}) is a metric on species.

A unifying family of diversity measures

Recap: We model a community by a relative abundance vector $\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$ and an $n \times n$ similarity matrix Z.

- $(Z\mathbf{p})_i = \sum_j Z_{ij} p_j$ is the expected similarity between a random organism and one of species i. It measures the ordinariness of species i.
- So $1/(Z\mathbf{p})_i$ is the distinctiveness of species i.

A community is diverse if it contains many distinctive individuals.

So one measure of diversity is the average distinctiveness:

$$\sum_i p_i \frac{1}{(Z\mathbf{p})_i}.$$

More generally, it's worth considering the power mean

$$\left(\sum_i p_i \left(\frac{1}{(Z\mathbf{p})_i} \right)^{1-q} \right)^{1/(1-q)}$$

for every real q (but we'll stick to $q \geq 0$).

A unifying family of diversity measures

Definition

The diversity of the community, of order $q \geq 0$, is

$$D_q^Z(\mathbf{p}) = \left(\sum_i p_i (Z\mathbf{p})_i^{q-1} \right)^{1/(1-q)}.$$

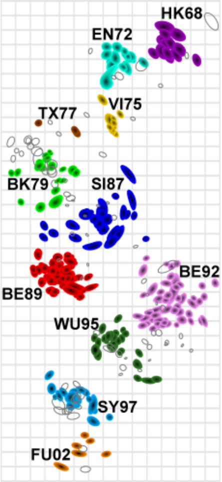
When $q = 1$, this doesn't make sense. Instead, define

$$D_1^Z(\mathbf{p}) = \lim_{q \to 1} D_q^Z(\mathbf{p}) = \exp\Big(- \sum p_i \log(Z\mathbf{p})_i \Big).$$

E.g. Naive model $Z = I$: then $D_1^Z(\mathbf{p}) = \exp(\text{Shannon entropy of } \mathbf{p})$.

Properties of these diversity measures

- This family of diversity measures encompasses many of the measures already defined and used by ecologists, geneticists, etc.
- They behave sensibly when species are reclassified (don't jump suddenly).
- They behave smoothly under change of resolution (e.g. if we go down to the subspecies level).
- They can be used in situations where we *don't have species classifications at all* (often the case for microbial communities).

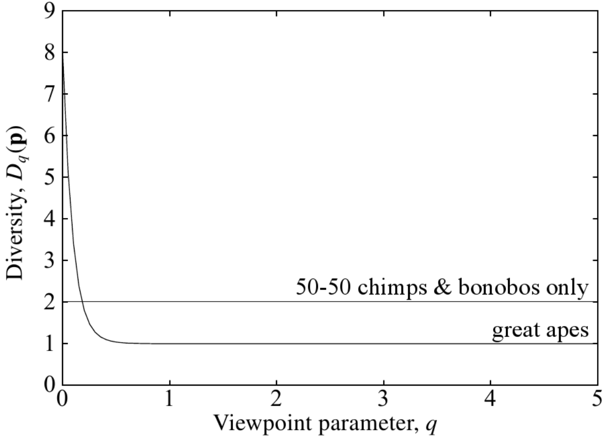


Influenza strains, 1968–2002

Comparing communities

The diversity profile of a community is the graph of $D_q^Z(\mathbf{p})$ against q.

E.g. Great apes worldwide, with naive similarity matrix ($Z = I$):



The parameter q controls the relative emphasis on rare or common species.

Maximizing diversity

Problem: Fix a list of species (i.e. a similarity matrix Z), and suppose we are free to choose their relative abundances $\mathbf{p}$.

- Which distribution $\mathbf{p}$ maximizes the diversity $D_q^Z(\mathbf{p})$?
- What is the value of the maximum diversity, $\sup_{\mathbf{p}} D_q^Z(\mathbf{p})$?

Diversity profiles can cross, so in principle, both answers depend on q.

Theorem (with Mark Meckes, *Entropy*, 2016)

Neither does. That is:

- *There is a single distribution* $\mathbf{p}_{max}$ *maximizing diversity of all orders* q *simultaneously (a 'best of all possible worlds').*
- *The maximum diversity* $D_q^Z(\mathbf{p}_{max})$ *is the same for all* q*. Call it* $D_{max}(Z)$.

Consequences of the maximum diversity theorem

Any $n \times n$ similarity matrix Z (e.g. coming from a metric via $Z_{ij} = e^{-d_{ij}}$) gives rise *canonically* to:

- a probability distribution $\mathbf{p}_{\mathrm{max}}$ on $\{1, \ldots, n\}$ (usually unique)
- a real number $D_{\mathrm{max}}(Z)$.

Often, $D_{\mathrm{max}}(Z)$ is equal to the magnitude $|Z| = \sum_{i,j} (Z^{-1})_{ij}$.

It's *always* equal to the magnitude of Z restricted to some subset of $\{1, \ldots, n\}$.

In a slogan:

Magnitude $\approx$ maximum diversity

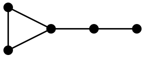
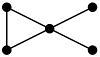
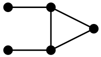
4. Graphs, revisited (briefly)

The magnitude of a graph

We've seen that any graph G can be understood as a special metric space, where all distances are integers.

(For simplicity, I'll stick to undirected graphs—but can do directed.)

Special property of graphs *Write $x = e^{-t}$. Then the magnitude function $t \mapsto |tG|$ is a rational function of x.*

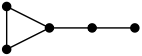
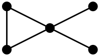
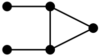
E.g.  ,  and  all have magnitude function

$$\frac{5 + 5x - 4x^2}{(1+x)(1+2x)}.$$

The magnitude of a graph can be studied as a graph invariant, and shares some invariance properties with the Tutte polynomial.

The magnitude of a graph

Lemma *The magnitude function of a graph can also be expressed as a power series with integer coefficients.*

E.g.  ,  and  have magnitude function

$$5 - 10x + 16x^2 - 28x^3 + 52x^4 - 100x^5 + \cdots .$$

In general, the magnitude function of a graph G is

$$c_0 + c_1 x + c_2 x^2 + \cdots$$

where

$$\begin{aligned}
c_0 &= \text{number of nodes} \\
c_1 &= -2 \cdot (\text{number of edges}) \\
c_2 &= \sum_{i,j : d_{ij} = 2} \underbrace{\Big((\text{num of configurations } i \!-\!\bullet\!-\! j) - 1 \Big)}_{\text{redundancy}} \\
&\quad + 6 \cdot (\text{num of triangles}) + 2 \cdot (\text{num of edges}).
\end{aligned}$$

5. The future: magnitude homology

Two points of view on Euler characteristic

So far: Euler characteristic has been treated as an analogue of cardinality.

Alternatively: Given any homology theory H_* of any kind of object X, can define

$$\chi(X) = \sum_{n=0}^{\infty} (-1)^n \operatorname{rank} H_n(X).$$

Note:

- $\chi(X)$ is a *number*
- $H_*(X)$ is an *algebraic structure*, and functorial in X.

In this sense, homology improves on ("categorifies") Euler characteristic.

The magnitude homology of a graph (Hepworth–Willerton)

There is a definition (omitted) of the magnitude homology of a graph.

It is a *graded* homology theory. That is, for each graph G and integer $k \geq 0$, it gives a *sequence*

$$H_{k,0}(G), H_{k,1}(G), H_{k,2}(G), \ldots$$

of abelian groups.

So for each $k \geq 0$, we have a power series

$$\text{rank}(H_{k,0}(G)) + \text{rank}(H_{k,1}(G))x + \text{rank}(H_{k,2}(G))x^2 + \cdots$$

The Euler characteristic $\chi(G)$ for this homology theory is (inevitably) defined as the alternating sum of these power series over $k = 0, 1, \ldots$.

Theorem (Hepworth, Willerton) *$\chi(G)$ is exactly the magnitude function of G.*

So: magnitude is the Euler characteristic of magnitude homology.

The magnitude homology of a metric space

The definition of magnitude homology can be generalized from graphs to enriched categories (Shulman).

In particular, there is a magnitude homology of metric spaces.

Sample Theorem *For a closed set $A \subseteq \mathbb{R}^n$,*

$$A \text{ is convex} \iff H_1(A) = 0.$$

Otter (2018) has done a comparison of magnitude homology with persistent homology:

- She proves a relationship between persistent homology and a 'blurred version' of magnitude homology...
- but concludes that 'morally, these are very different homology theories', conveying different information.

Summary

Summary

Magnitude is a numerical invariant of metric spaces (e.g. data sets, networks, and the kinds of space that geometers like thinking about).

By considering rescalings, magnitude assigns a function to each space.

- For geometrically interesting spaces, the magnitude function carries geometrically interesting information (volume, dimension, etc).
- For finite spaces, it seems—empirically—to carry multiscale information on number of clusters and dimensionality.

We can also measure the diversity ($\sim$ entropy) of any probability distribution on a finite metric space...

...and magnitude is closely related to maximum diversity.

There is a theory of magnitude homology for metric spaces.

It is related to persistent homology, but expresses different information about the space... Lots to explore here!

References and further reading: www.maths.ed.ac.uk/∼tl/magbib