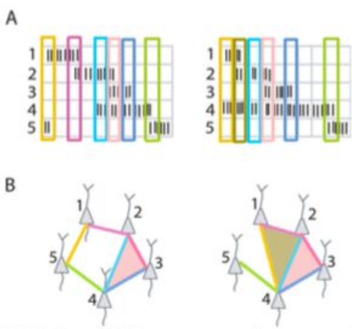
INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

Computational tools for handling simplicial complexes in real datasets

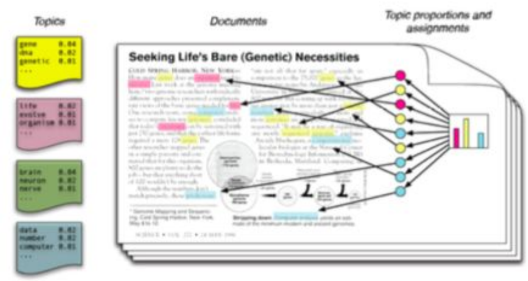
Alice Patania
Network Science Institute (IUNI), Indiana University

Why?



Curto, Carina, and Vladimir Itskov. "Cell groups reveal structure of stimulus space." PLoS Comput Biol 4, no. 10 (2008): e1000205.

brain networks

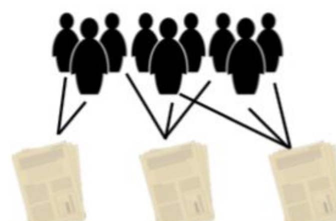


Rhody, Lisa. "Topic modeling and figurative language." Journal of Digital Humanities 2, no. 1 (2012): 19-35.

topic modelling

Wikipedia

protein interaction networks

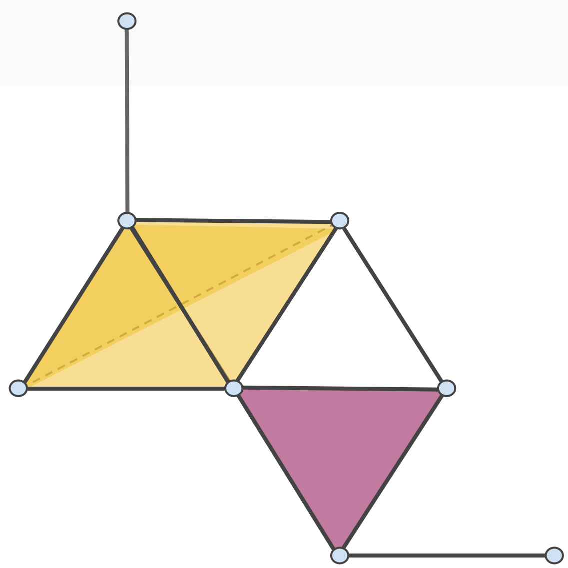


collaboration networks

Simplicial Complexes

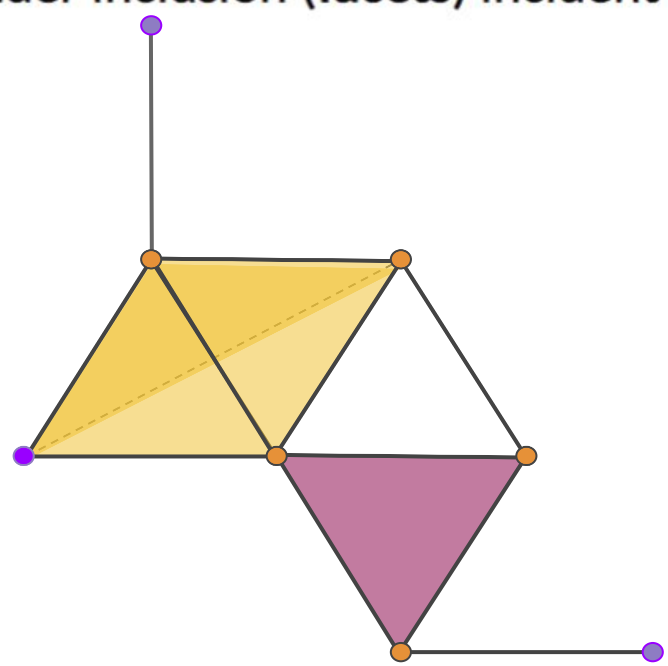
A **simplicial complex** X is a collection of simplices such that:

- $\forall \sigma \in X$ its faces are still in X,
- $\forall \sigma, \tau \in X$, $\sigma \cap \tau$ is either the empty set or a face of both σ and τ.



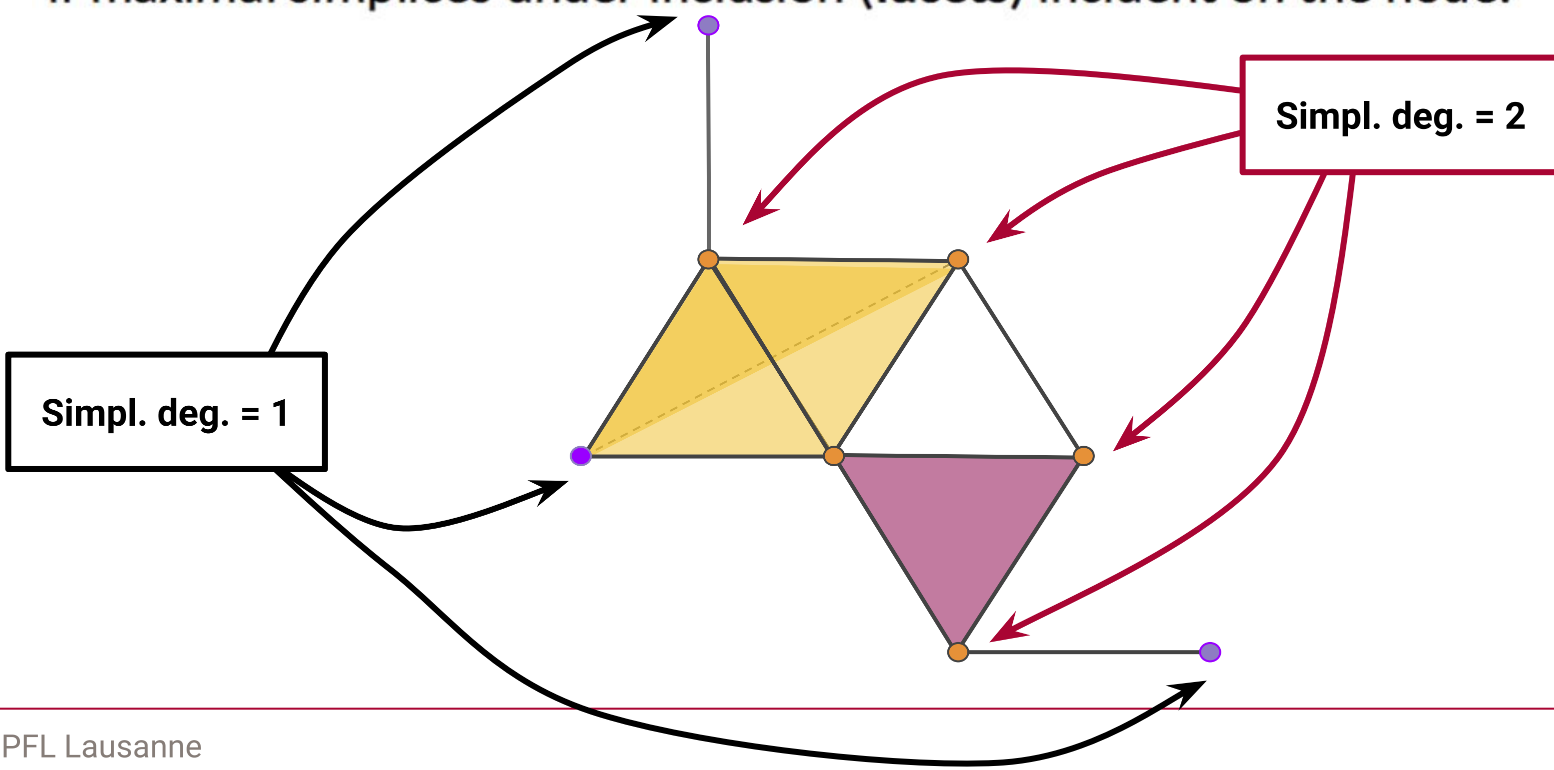
Simplicial Complexes

The **simplicial degree** of a node in a simplicial complex is the number if maximal simplices under inclusion (**facets**) incident on the node.



Simplicial Complexes 101

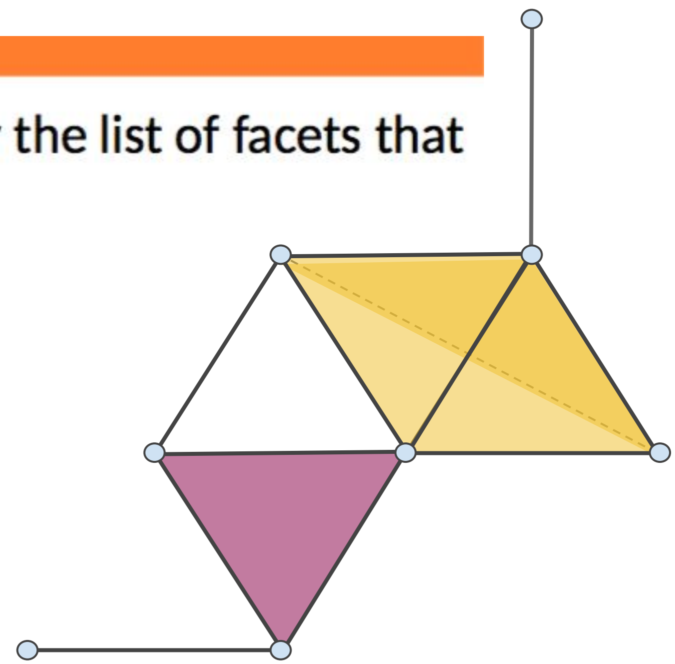
The **simplicial degree** of a node in a simplicial complex is the number if maximal simplices under inclusion (**facets**) incident on the node.



Simplicial Complexes

The **simplicial degree** of a node in a simplicial complex is the number if maximal simplices under inclusion (**facets**) incident on the node.

A simplicial complex is completely described by the list of facets that belong to it.



Simplicial Complexes

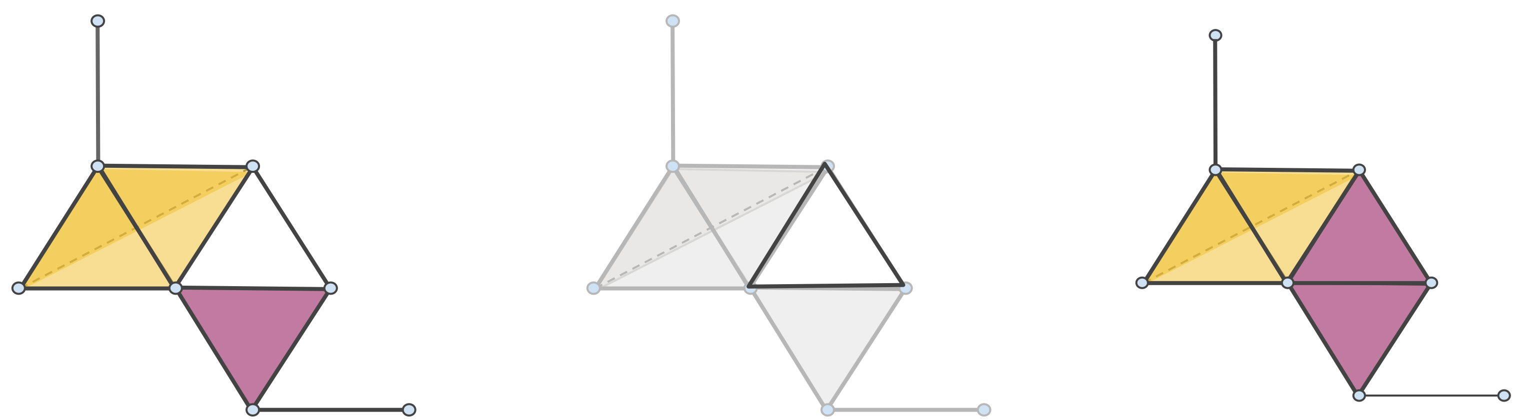
Simplicial homology

What hinders application of homology to data?

Representation: it's difficult to find an optimal representative

Memory and efficiency: the algorithm for computing homology grows with the number of simplices in the complex

Null model: There is a lack of samplers that could easily be used in practice.



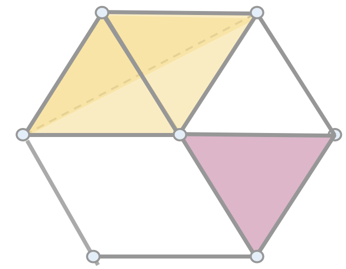
Structure

- **Random Simplicial Complexes**
  - Simplicial Configuration Model
- **Reducing the complexity of homology computation**
- **1D-Homology and network communities**
  - arXiv case study

Random Simplicial Complexes

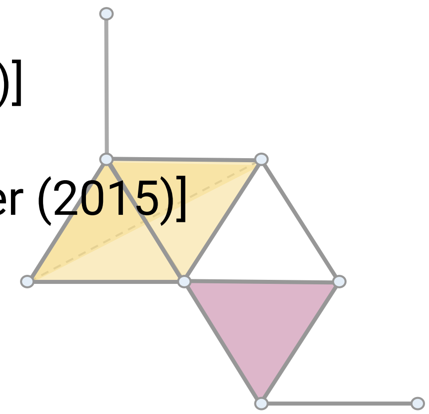
Simplicial Complexes

Sampling



Erdos-Renyi inspired:

- **Random pure simplicial complexes** [Linial-Meshulam (2006)]
- **Random simplicial complexes** [Kahle (2009)]
- **Multi-parameter random simplicial complexes** [Costa-Farber (2015)]

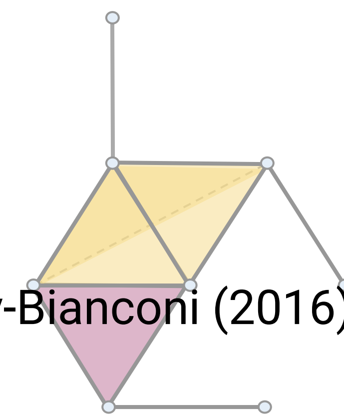


Exponential random graphs inspired:

- **Exponential random simplicial complex** [Zuev et al. (2015)]

Preferential attachment for simplicial complexes:

- **Network geometry with flavor** [Bianconi-Rahmede (2016)]



Configuration model:

- **Configuration model for pure simplicial complexes** [Courtney-Bianconi (2016)]

Configuration model

The **configuration model** is a generative model that creates a random graph with a fixed degree sequence.

It implies the following are fixed?

the number of nodes n

the number of edges in the network $m = \frac{1}{2}\sum_i k_i$

Configuration model

The **configuration model** is a generative model that creates a random graph with a fixed degree sequence.

Suppose to have n vertices with fixed degrees k_i for $i = 1, \ldots, n$, the random graph is constructed in the following way.

1. Each vertex i is provided with k_i edge 'stubs', there are therefore $\sum_i k_i = 2\,m$ stubs.
2. Uniformly at random two stubs are chosen and an edge is created connecting the two of them, until no free stubs are left in the graph.

Bipartite graphs and simplicial complexes

Theorem

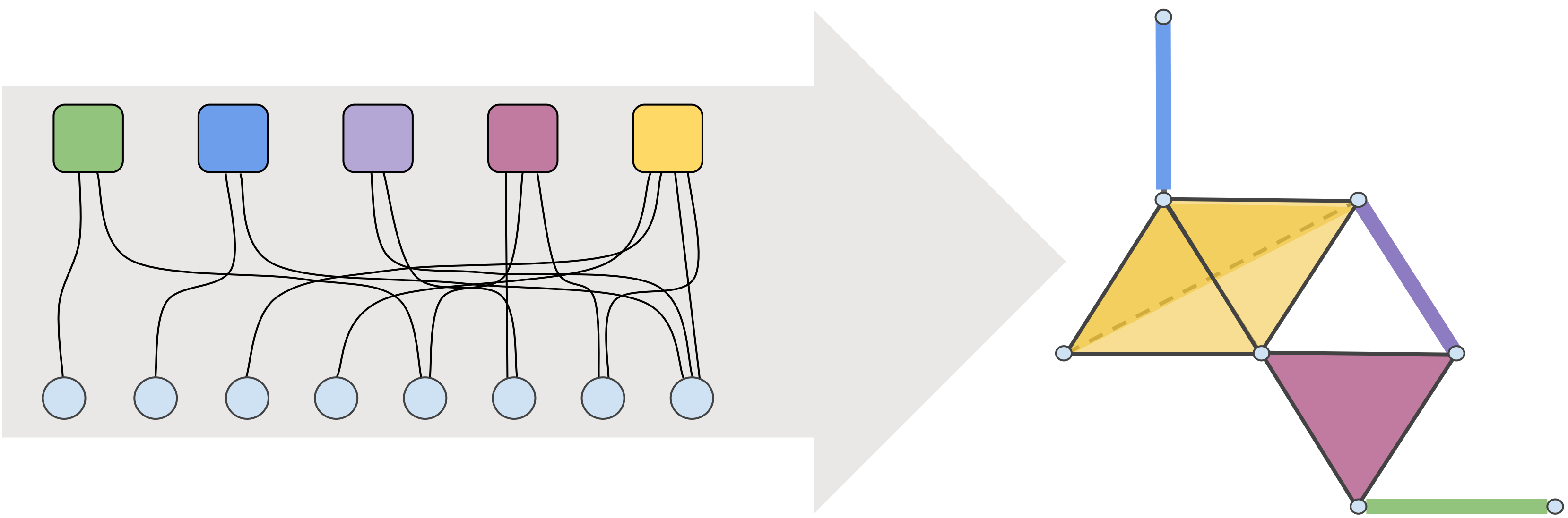
Let G be a bipartite graph with vertex sets $\{F, V\}$, G_V its one-mode projections onto the vertex set V.
Then it exists a simplicial complex Σ whose underlying graph is G_V.

The neighbours $\mathcal{N}(f_i)$ of f_i are the vertices that form the maximal simplex f_i, for each i, or equivalently, the neighbours $\mathcal{N}(v_i)$ of vertex v_i are the facets in which node v_i appears.

Bipartite graphs and simplicial complexes

Theorem

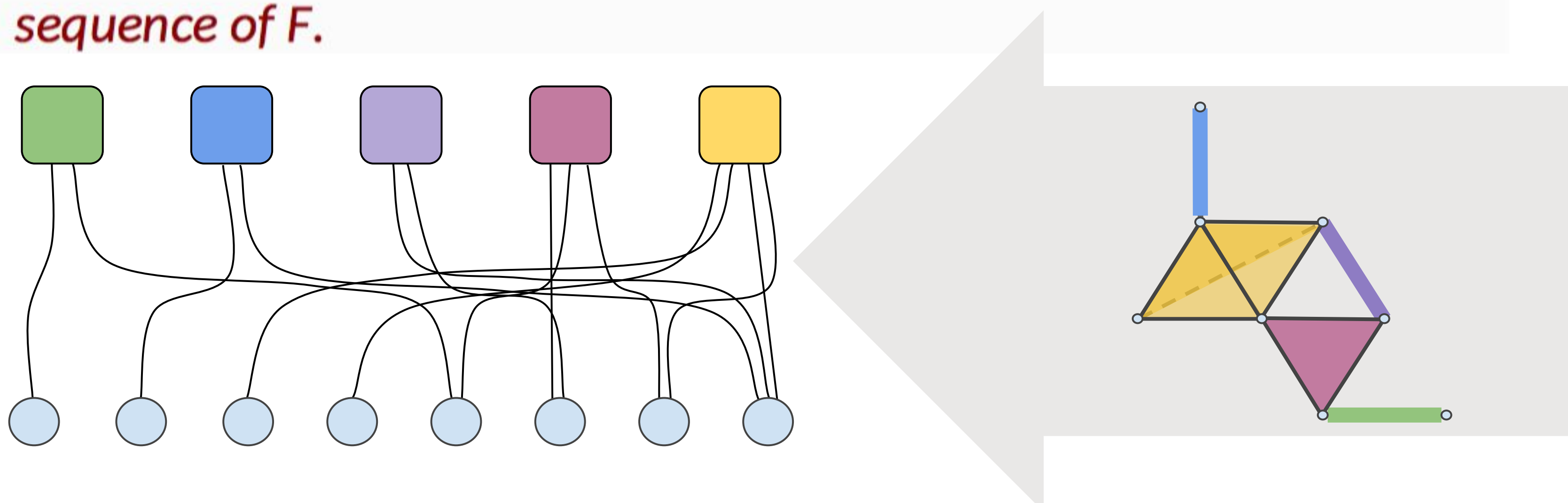
Let G be a bipartite graph with vertex sets $\{F, V\}$, G_V its one-mode projections onto the vertex set V.
Then it exists a simplicial complex Σ whose underlying graph is G_V.



Bipartite graphs and simplicial complexes

Theorem

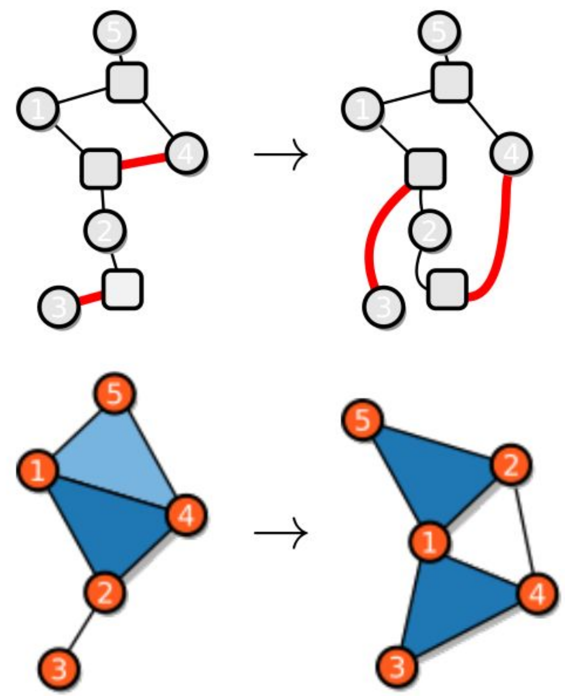
For every Σ, $\exists G$, bipartite graph, s.t. one of its two one-mode projections G_V is the underlying graph of Σ.
Moreover, the facet size sequence of Σ is equal to the degree sequence of F.



Bipartite graphs and simplicial complexes

Idea:

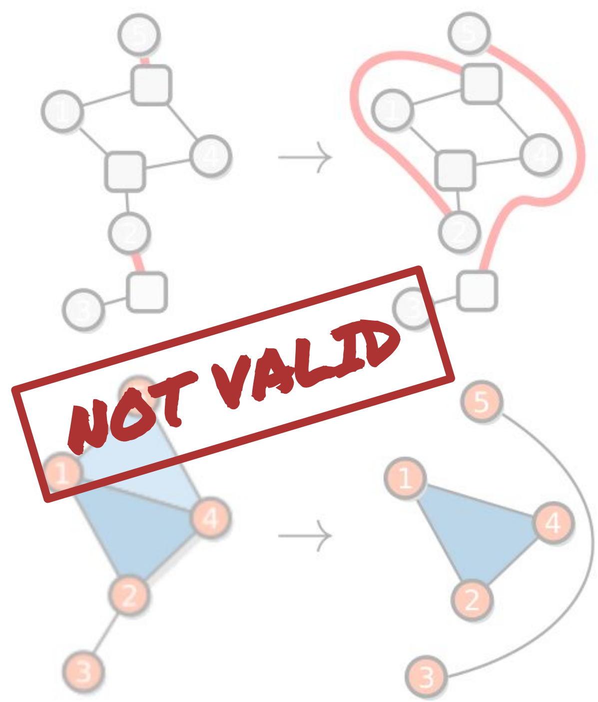
Use the configuration model for bipartite graphs and the maps to construct a sampling method for SCM.



Bipartite graphs and simplicial complexes

Idea:

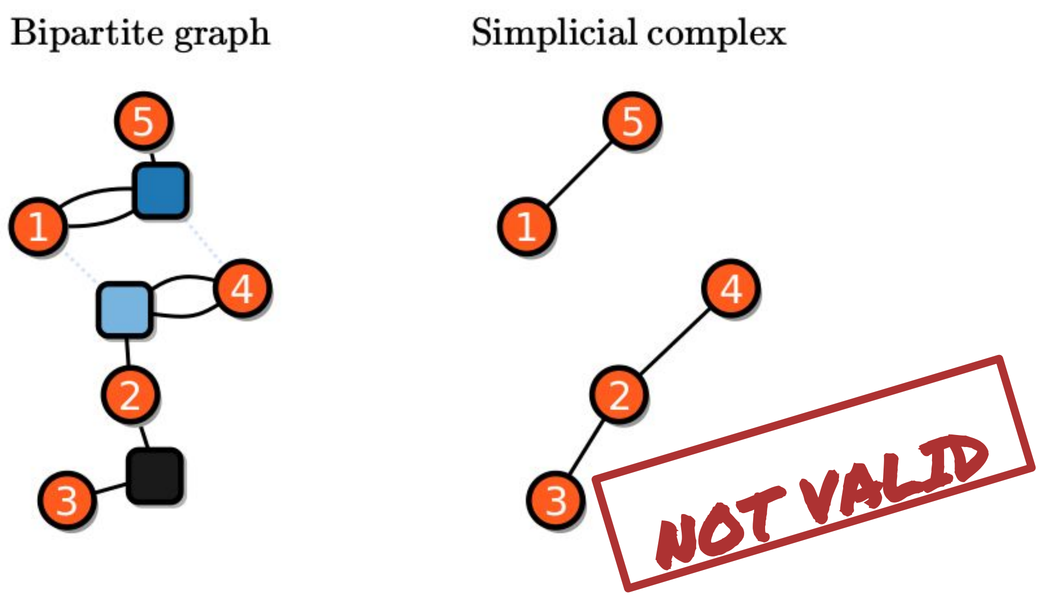
Use the configuration model for bipartite graphs and the maps to construct a sampling method for SCM.



Adding constraints

First constraint: **No multi-edges**

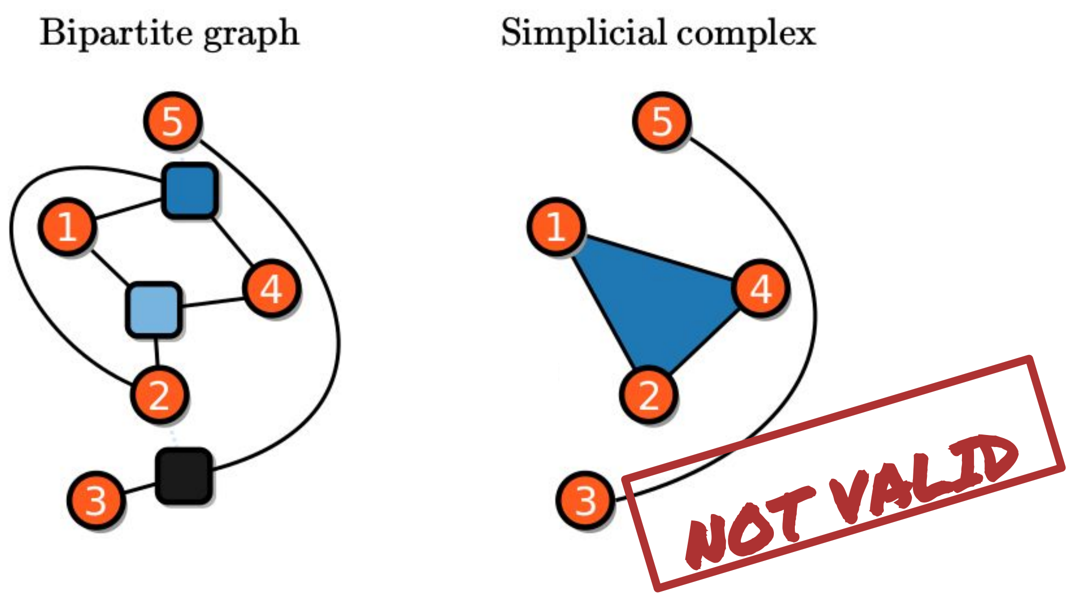
Multi-edges decrease the size of the maximal simplices.



Adding constraints

Second constraint: **No included neighborhoods**

Included neighborhoods violate the maximality assumption of the facets.



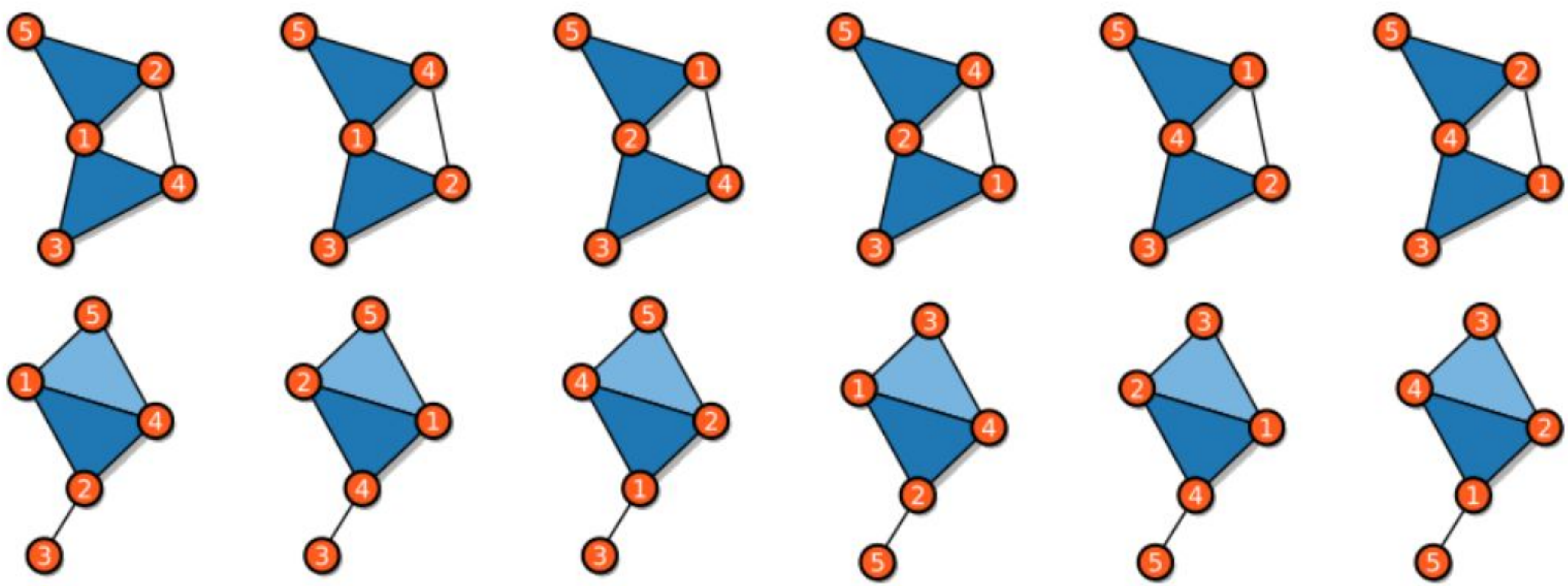
Adding constraints

Constraints:

No multi-edges

No included neighborhoods

Then the acceptable configurations for the toy example are the following:

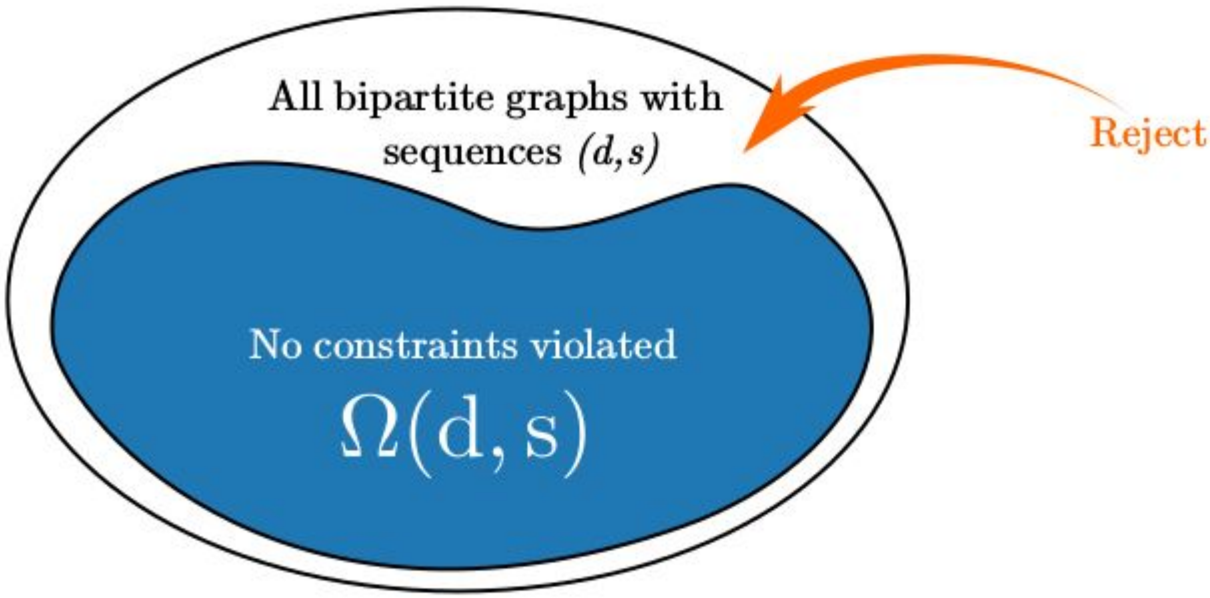


Adding constraints

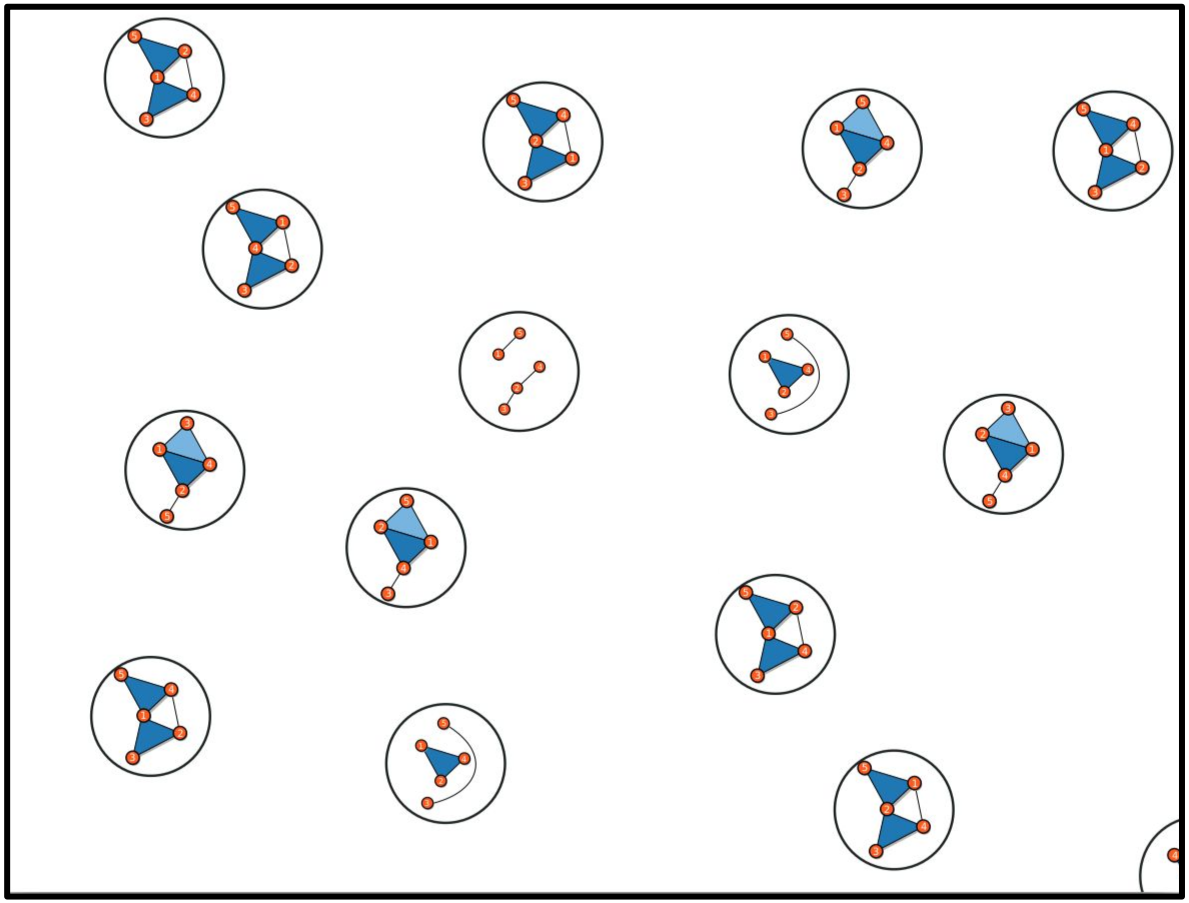
! **Problem with rejection sampling**: Far too many rejections!

Loose upper bound :

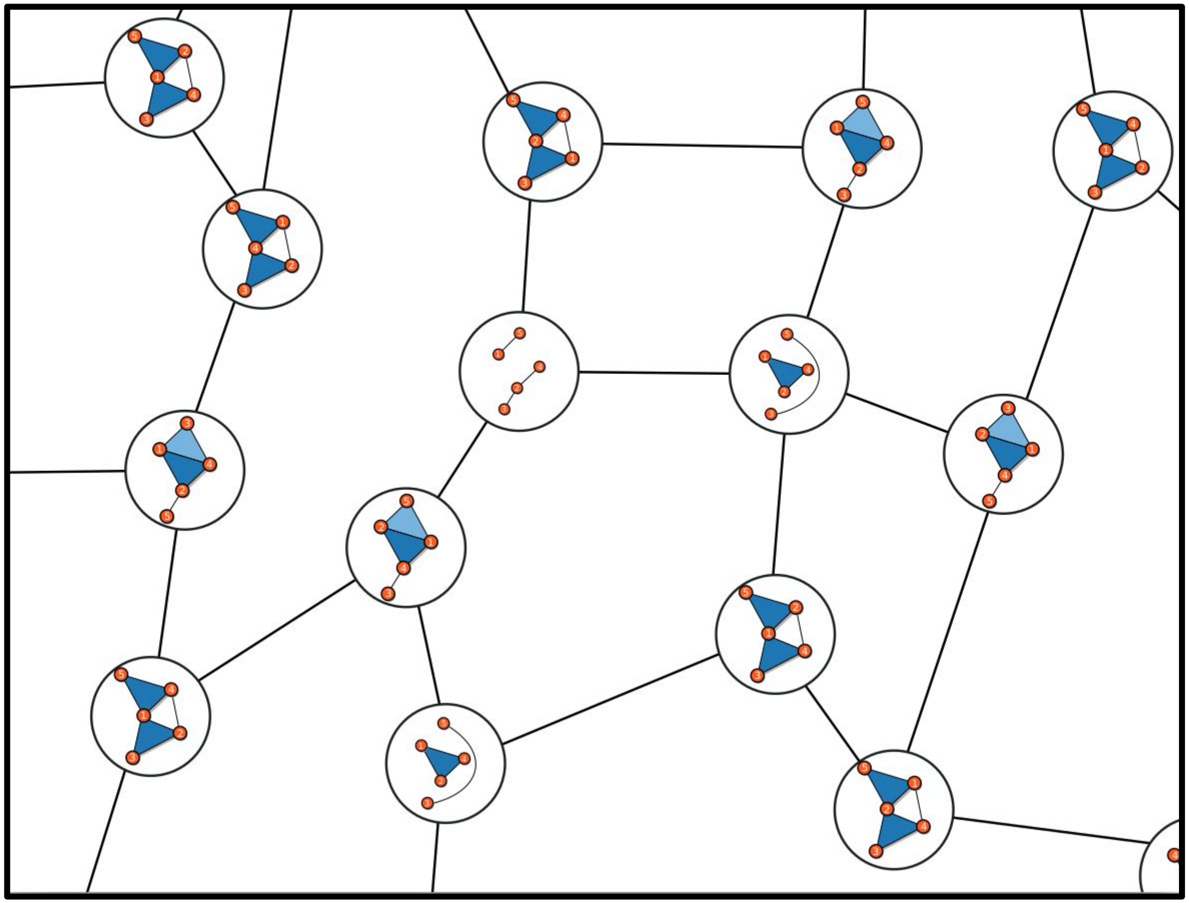
Pr[reject] > exp[-0.5($\langle d^2 \rangle / \langle d \rangle - 1$) ($\langle s^2 \rangle / \langle s \rangle - 1$)]



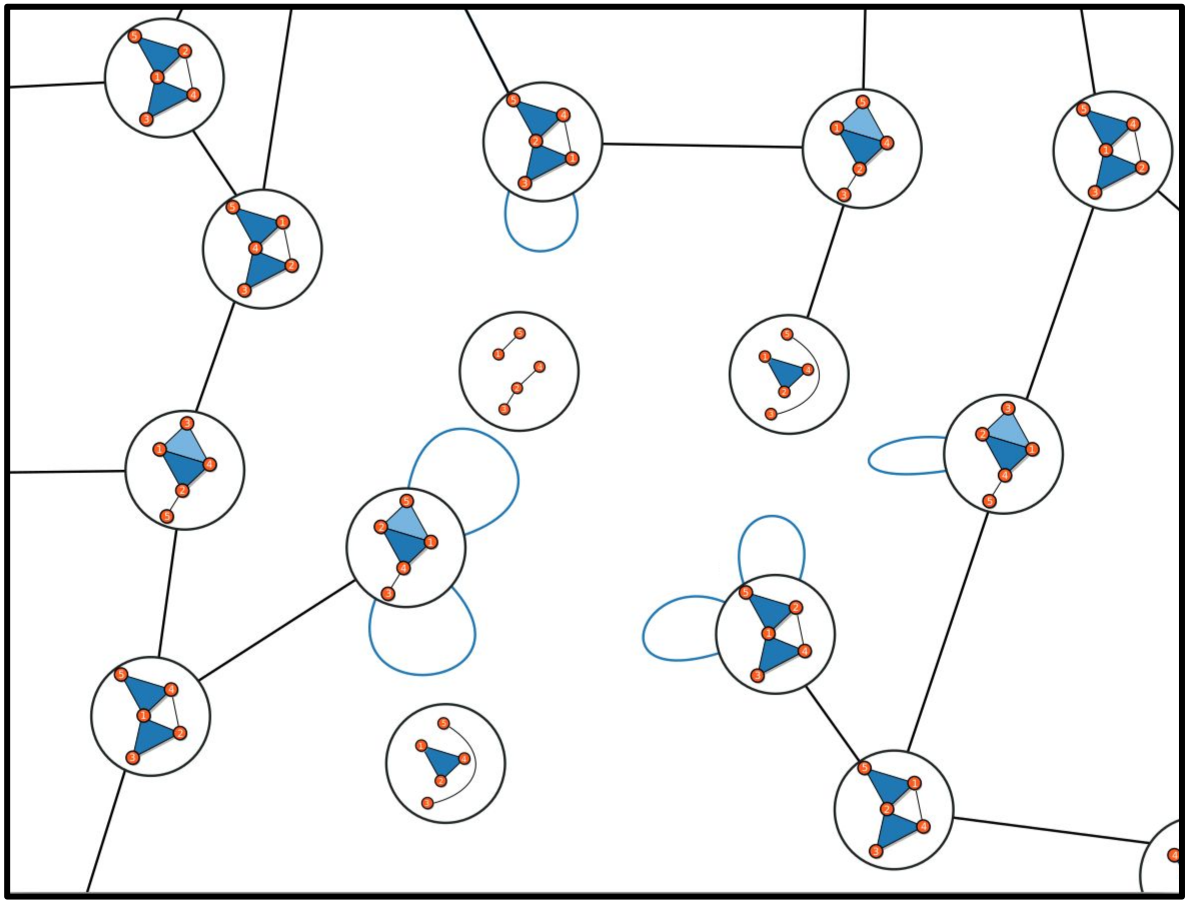
Markov Chain Monte Carlo sampling



Markov Chain Monte Carlo sampling



Markov Chain Monte Carlo sampling



MCMC sampling: The details

Move set

1. Pick L~P random edges in bipartite graph
   P can be arbitrary, we use $\Pr[L = l] = \exp[\lambda l]/Z$
2. Rewire edges. If multi-edge or included neighbors, reject.

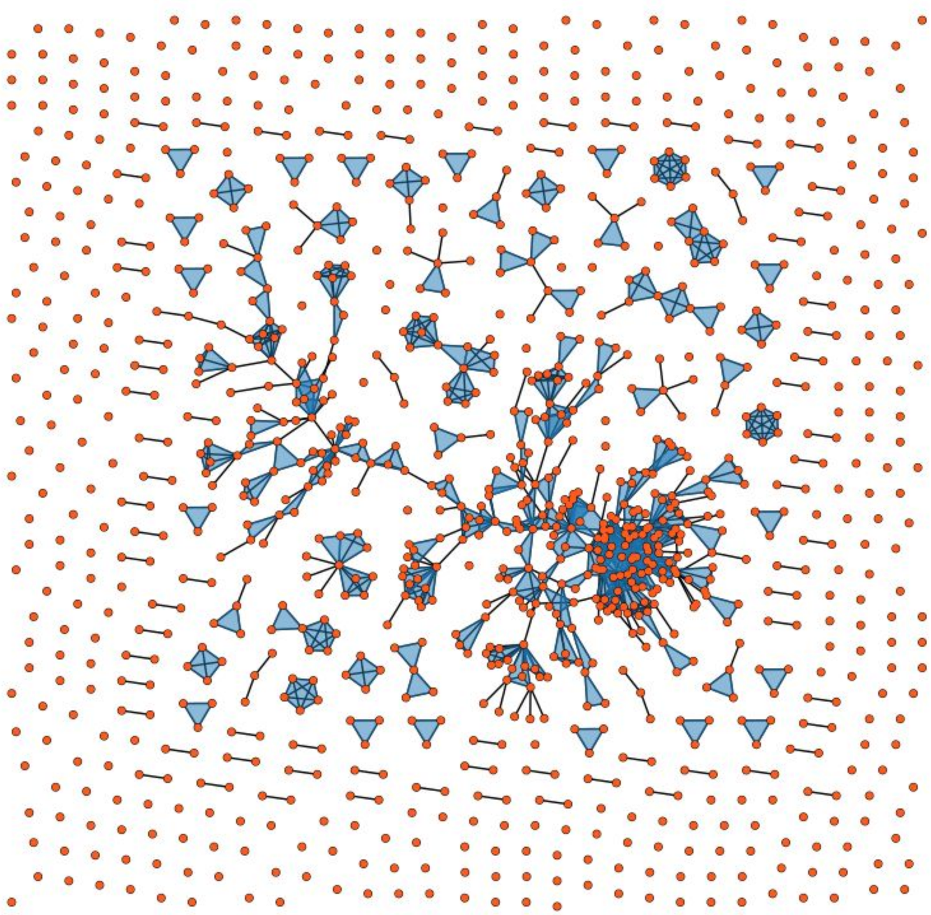
Similar to [Miklós–Erdős–Soukup, Electron. J. Combin., 20, (2013)]

- MCMC is uniform over $\Omega(d, s)$
- Move set yields aperiodic chain
- Move set connects the space

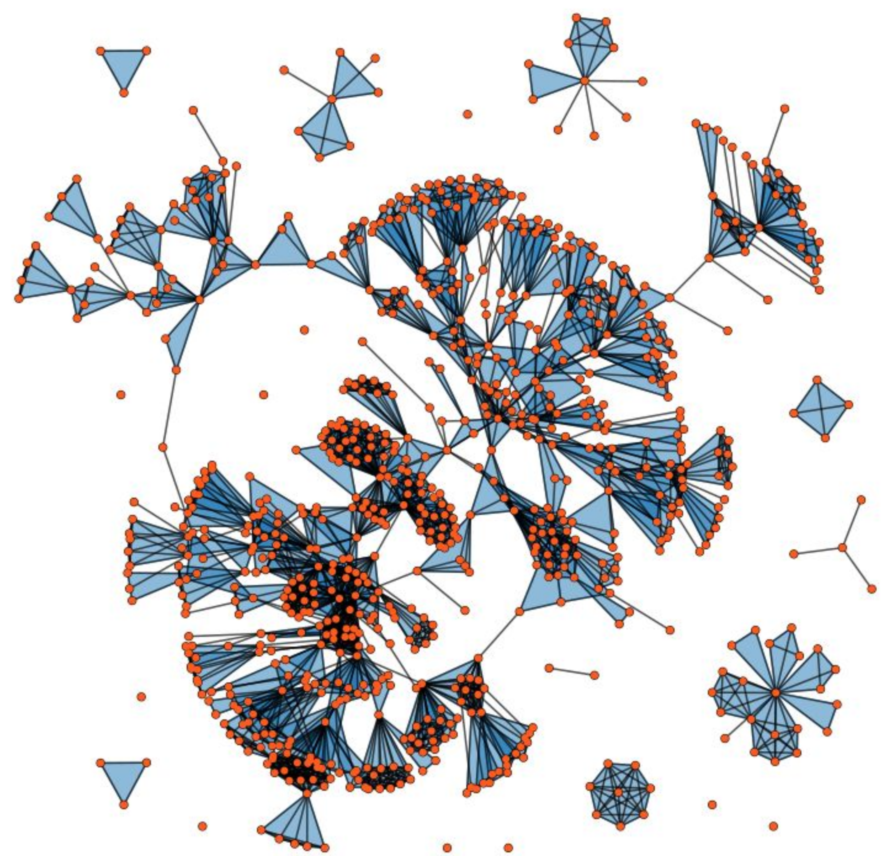
INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

Results - True sustems

Disease regulation dataset
(facets : *genes*, nodes : *human diseases*)
[Goh et al., PNAS, **104**, (2007)]

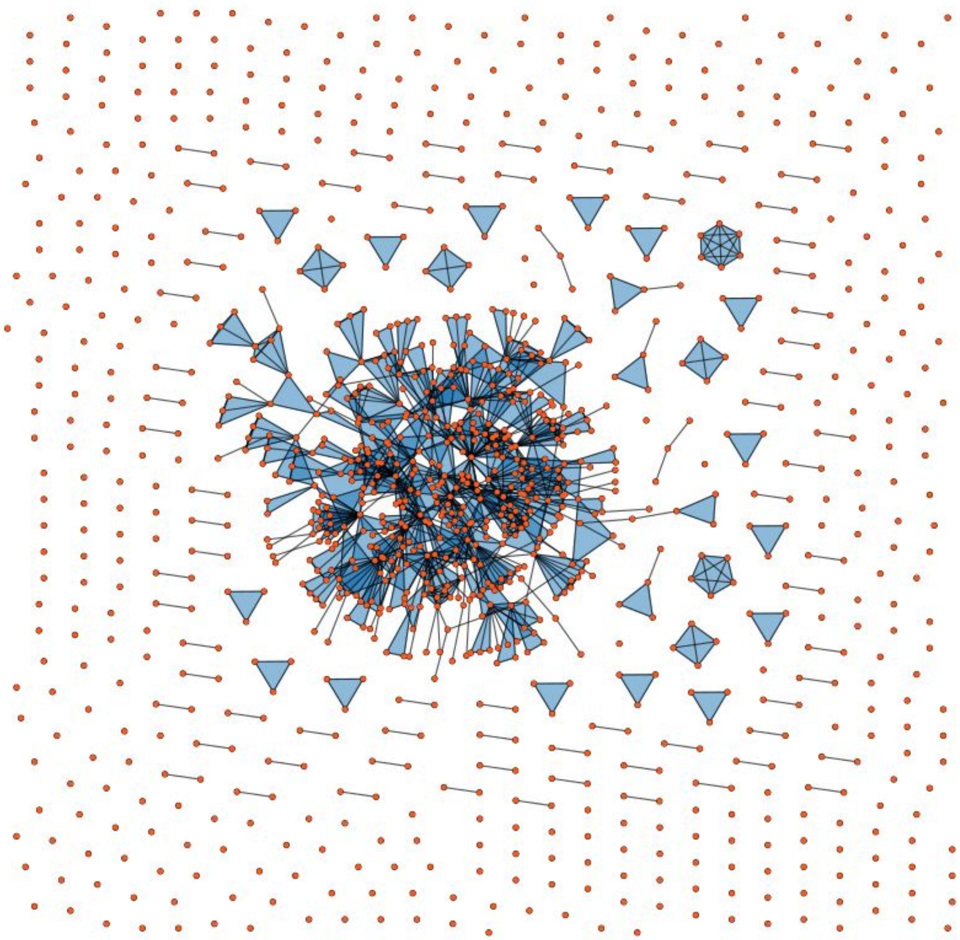


Crimes in St-Louis (true system)
(facets : *people*, nodes : *crimes*)
[Rosenfeld et al., (1991)]

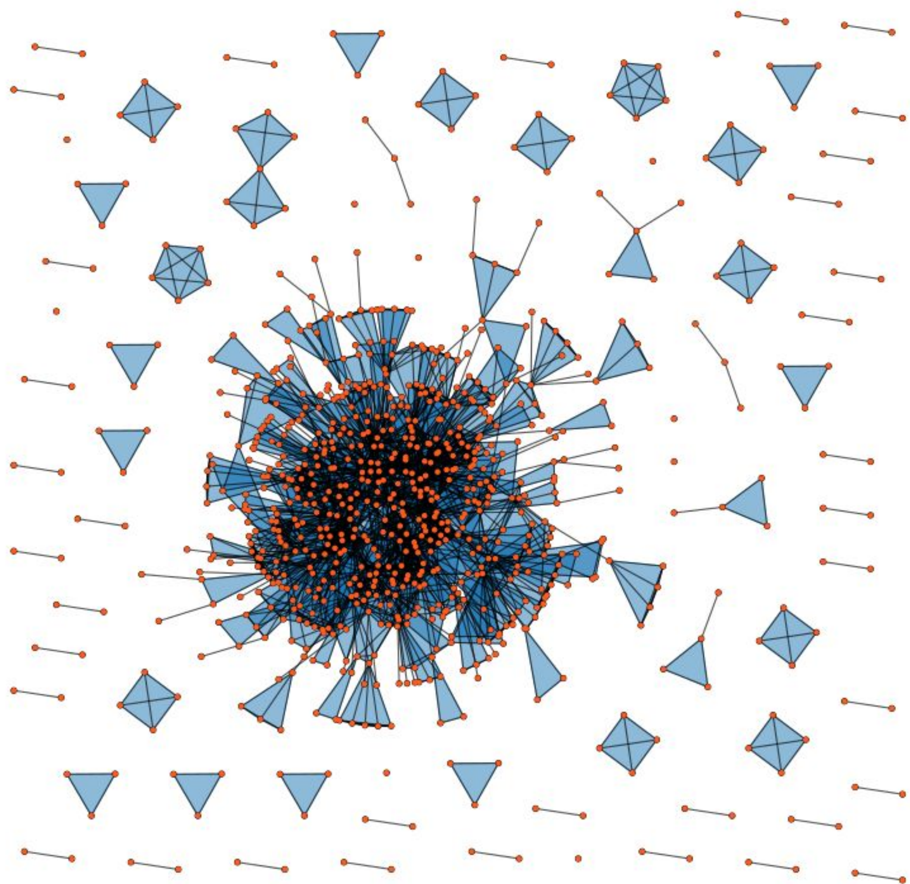


Results - Random instances

Disease regulation dataset (random instance)
(facets : *genes*, nodes : *human diseases*)
[Goh et al., PNAS, **104**, (2007)]



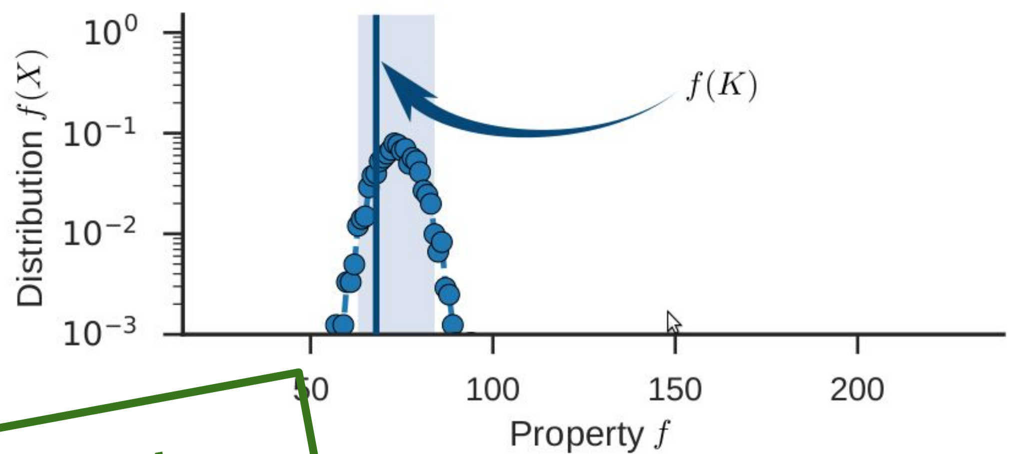
Crimes in St-Louis (random instance)
(facets : *people*, nodes : *crimes*)
[Rosenfeld et al., (1991)]



Concept for a null model

Null model

Is the quantity **f(X)** close to **f(K)** for random simplicial complexes **X~SCM[d(K), s(K)]** ?



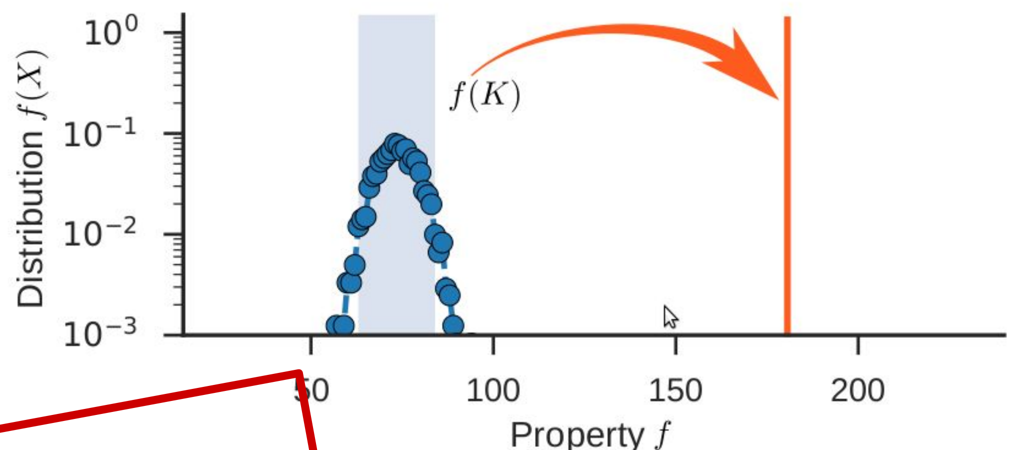
Pr[| f(K) – f(X)| <] ≈ 1

K is typical, the local quantities (***d,s***) explain **f**.

Concept for a null model

Null model

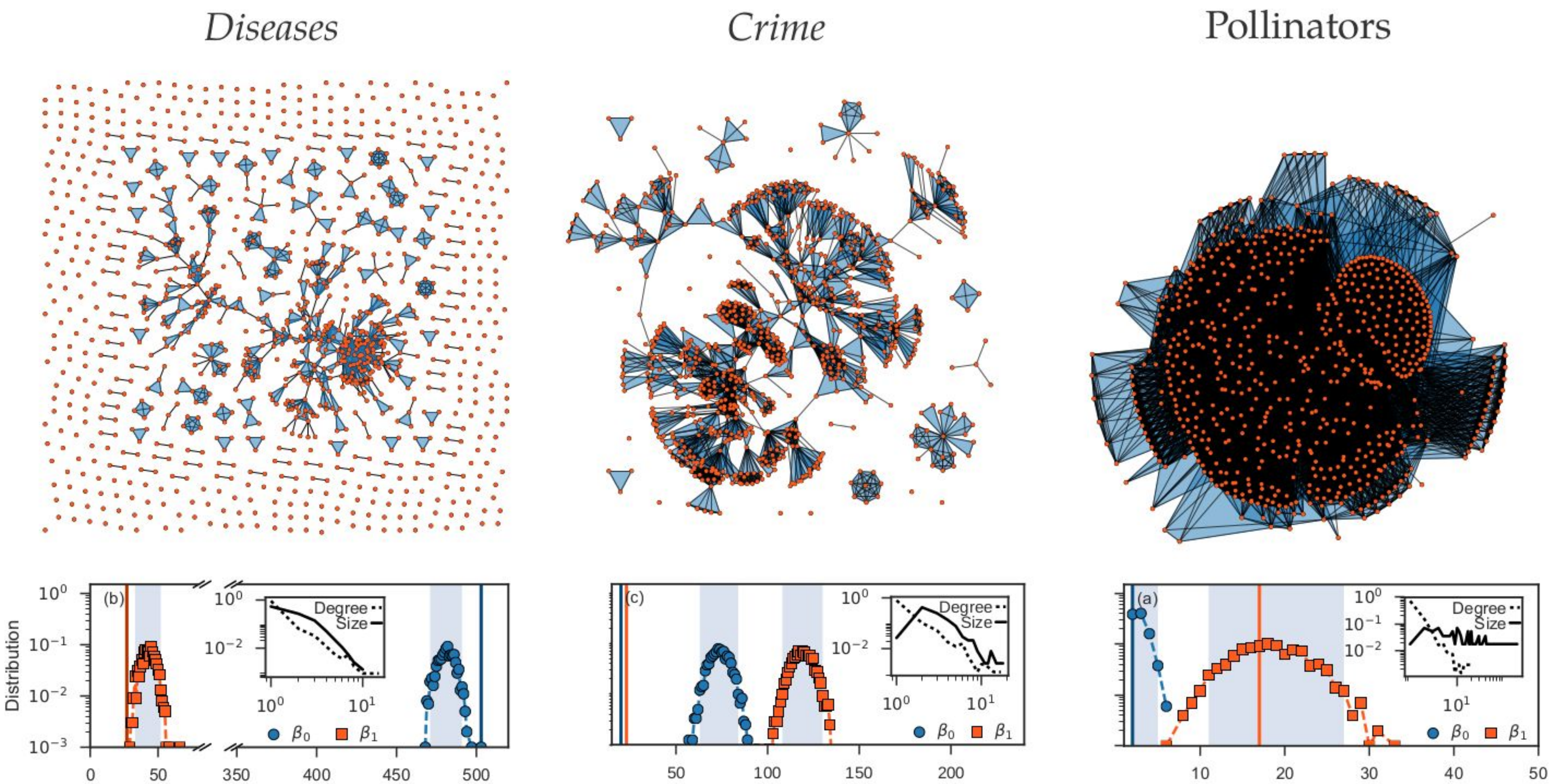
Is the quantity f(X) close to f(K) for random simplicial complexes X~SCM[d(K), s(K)] ?



$\Pr[\,|f(K) - f(X)| <] \ll 1$

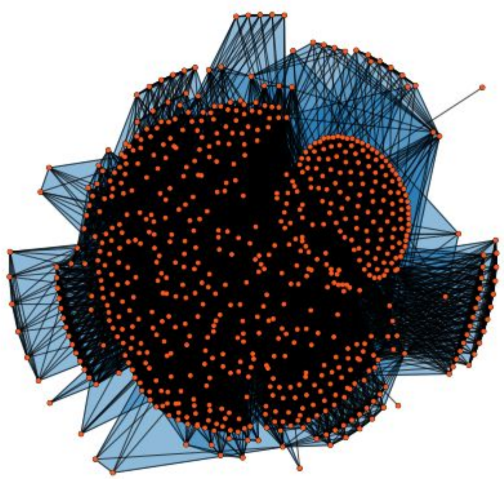
K is atypical, K is organized beyond the local scale.

Results on Betti numbers of real data sets

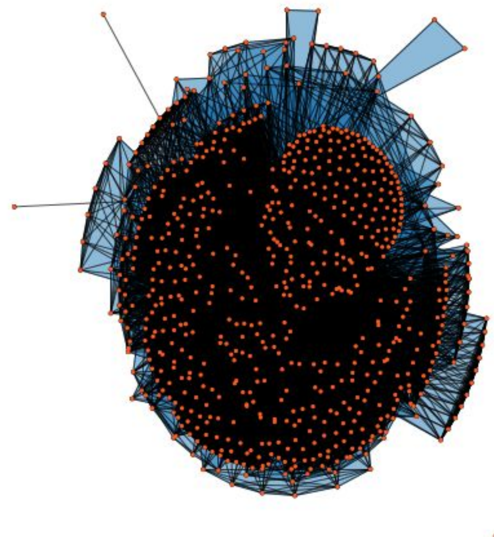


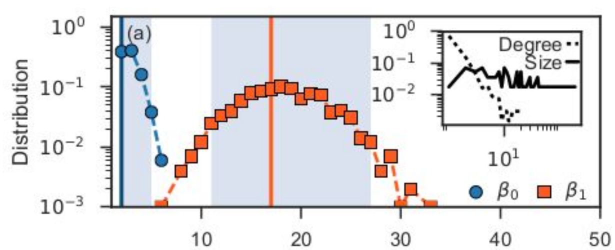
Results on Betti numbers of real data sets

Pollinators (*real*)



Pollinators (*random*)





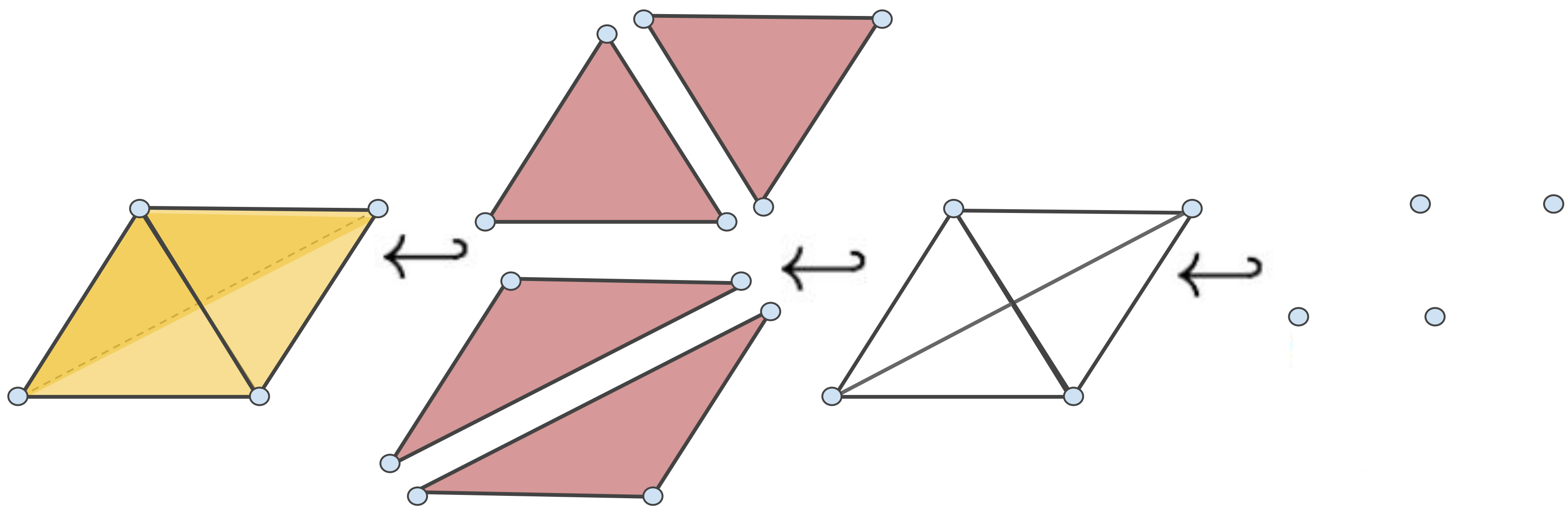
Reducing the algorithmic complexity

Simplicial Complexes

homological hiccups

The computational complexity of homology is $\mathbf{O(m^3) > O([2^{max(s)}]^3)}$

where m is the number of **ALL** simplices in the complex not only the maximal facets.

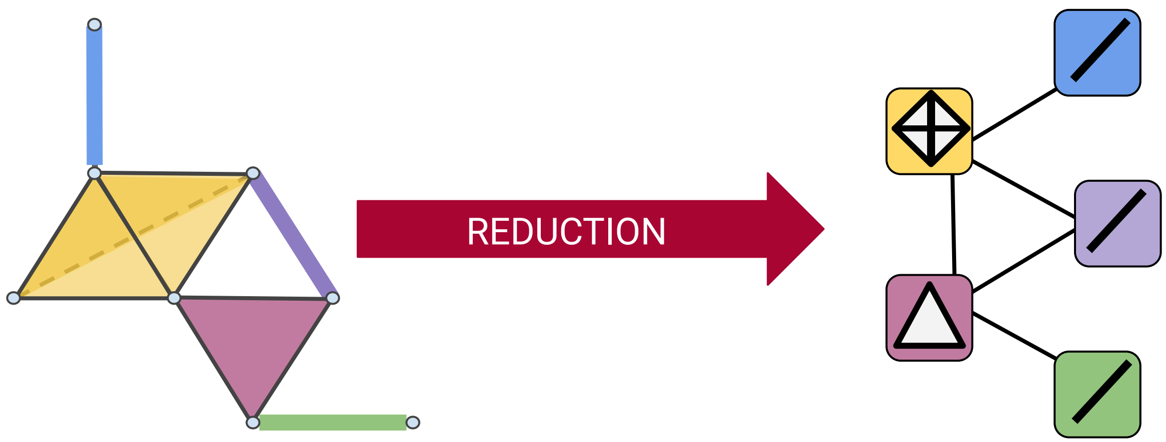


Simplicial Complexes

homological hiccups

The computational complexity of homology is $\mathbf{O(m^3) > O([2^{max(s)}]^3)}$

where m is the number of **ALL** simplices in the complex not only the maximal facets.

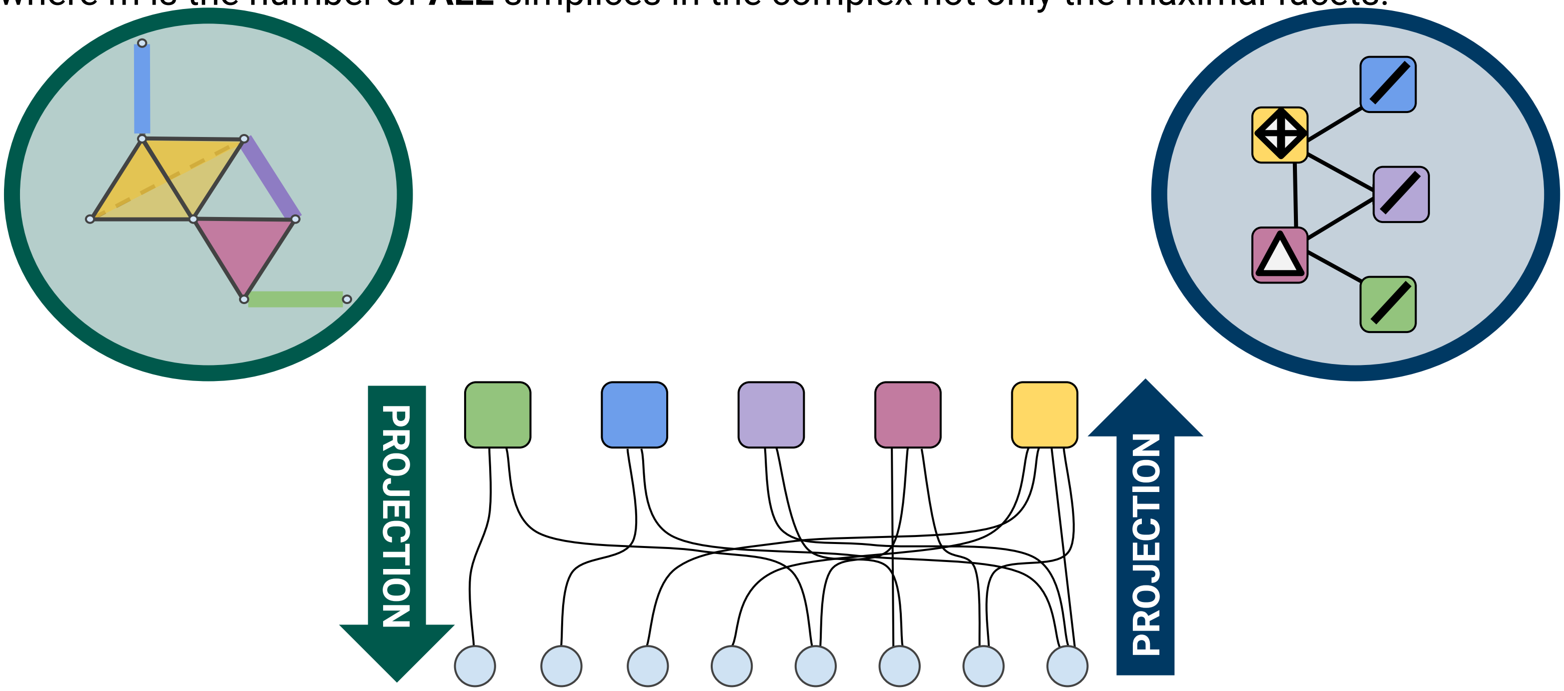


Simplicial Complexes

homological hiccups

The computational complexity of homology is $\mathbf{O(m^3) > O([2^{max(s)}]^3)}$

where m is the number of **ALL** simplices in the complex not only the maximal facets.

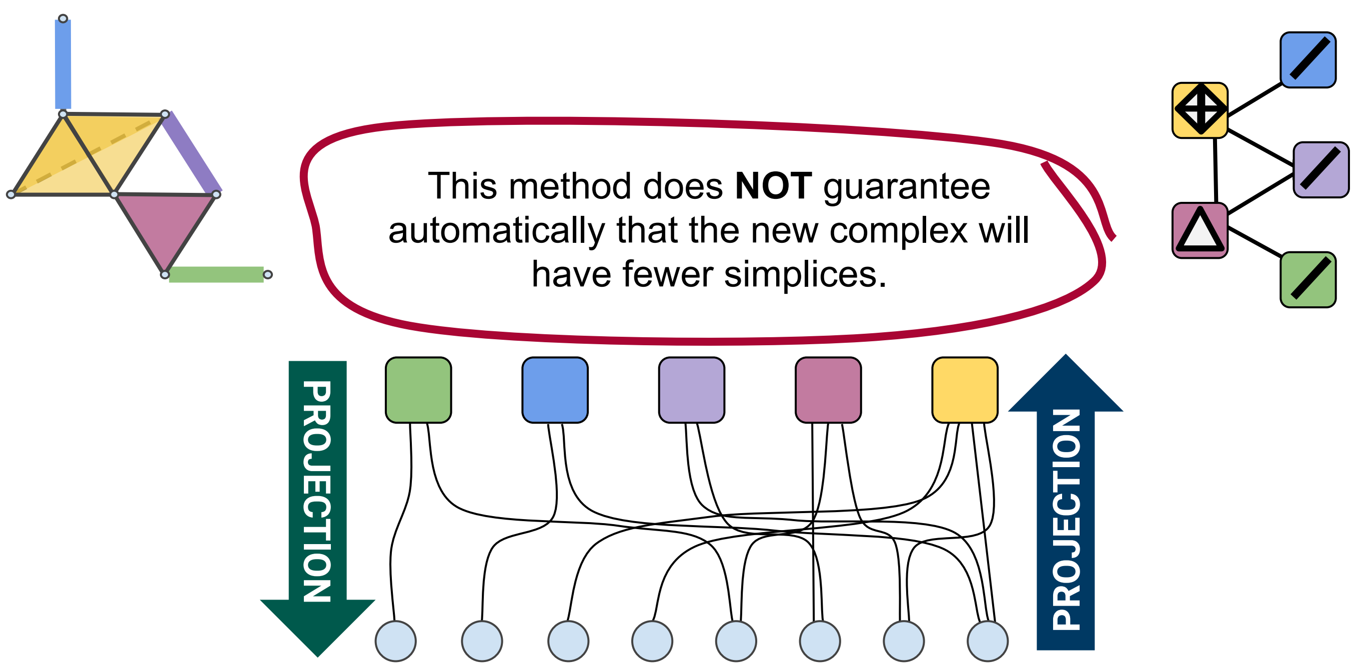


Simplicial Complexes

homological hiccups

The computational complexity of homology is $\mathbf{O(m^3) > O([2^{max(s)}]^3)}$

where m is the number of **ALL** simplices in the complex not only the maximal facets.

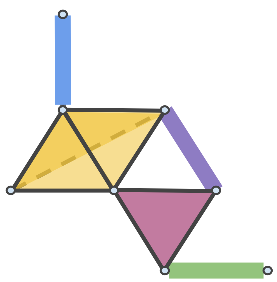


Simplicial Complexes

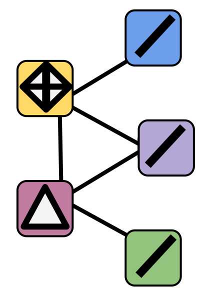
homological hiccups

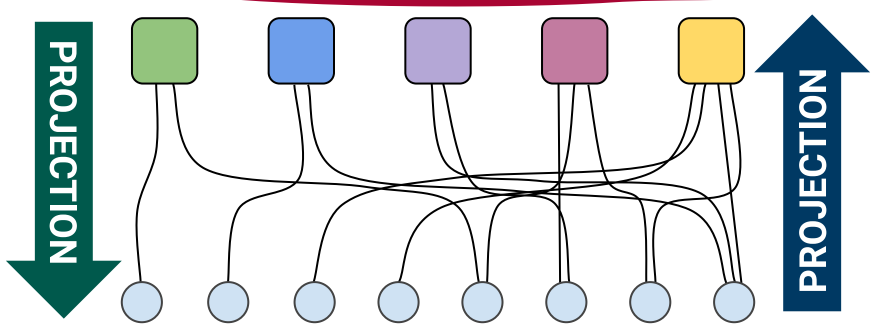
The computational complexity of homology is $\mathbf{O(m^3) > O([2^{max(d)}]^3)}$

where m is the number of **ALL** simplices in the complex not only the maximal facets.



This method does **NOT** guarantee automatically that the new complex will have fewer simplices.



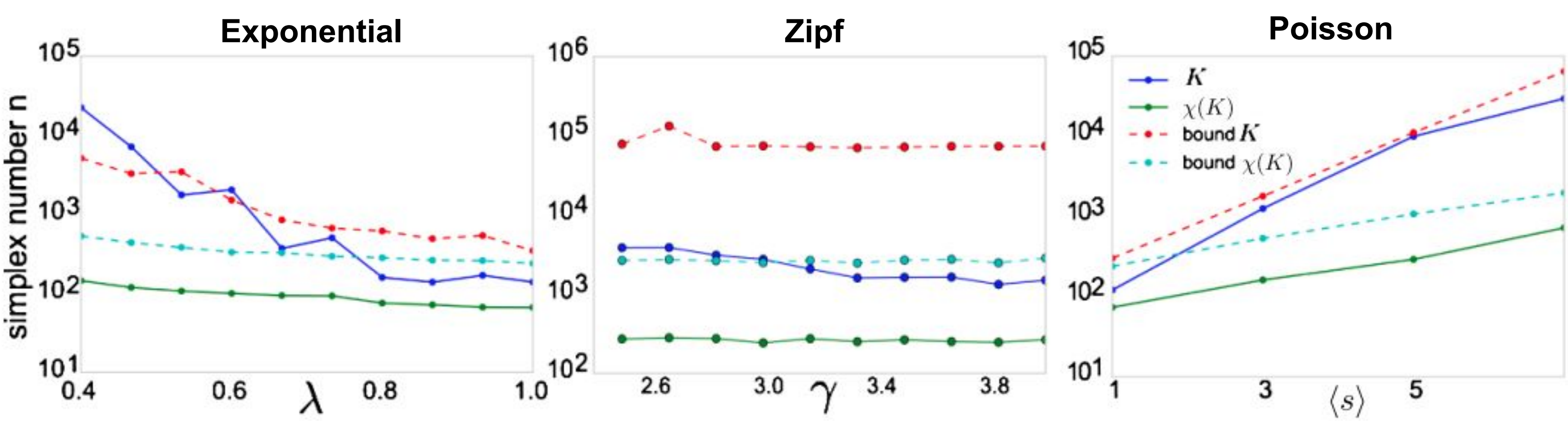


Simplicial Complexes

homological hiccups

The computational complexity of homology is $\mathbf{O(m^3) > O([2^{max(d)}]^3)}$

where m is the number of **ALL** simplices in the complex not only the maximal facets.



Simplicial Configuration Model, J-G Young et. al. 2017 PRE

1D Homology and communities

Empirical proof of concept

The data set

The data span 9 years, from 2007 to 2016, and are split according to the 18 major categories of arXiv.

This major categories correspond to different thematic areas and thus can be used as rough representative of different scientific fields.

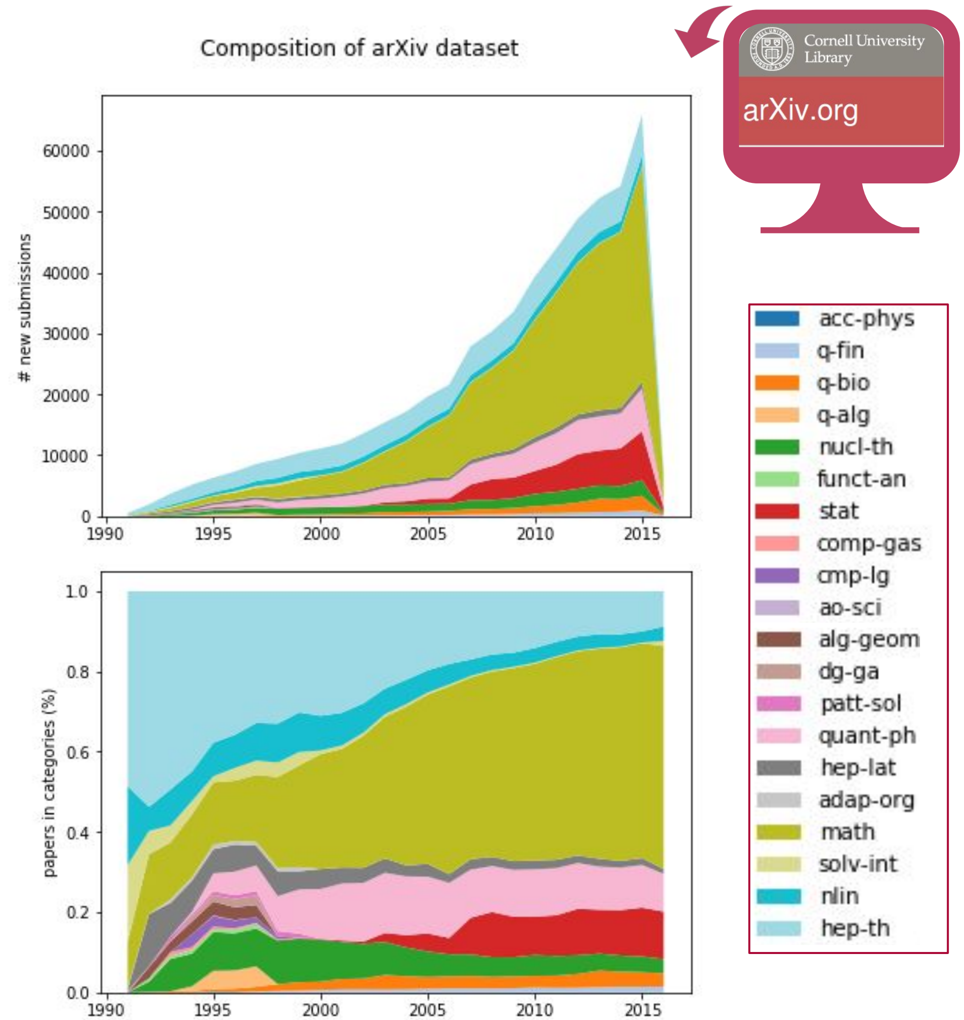
Notice: Due to arXiv's history, there is a bias toward mathematical and physical topics.

The data set

The data span 9 years, from 2007 to 2016, and are split according to the 18 major categories of arXiv.

This major categories correspond to different thematic areas and thus can be used as rough representative of different scientific fields.

Notice: Due to arXiv's history, there is a bias toward mathematical and physical topics.

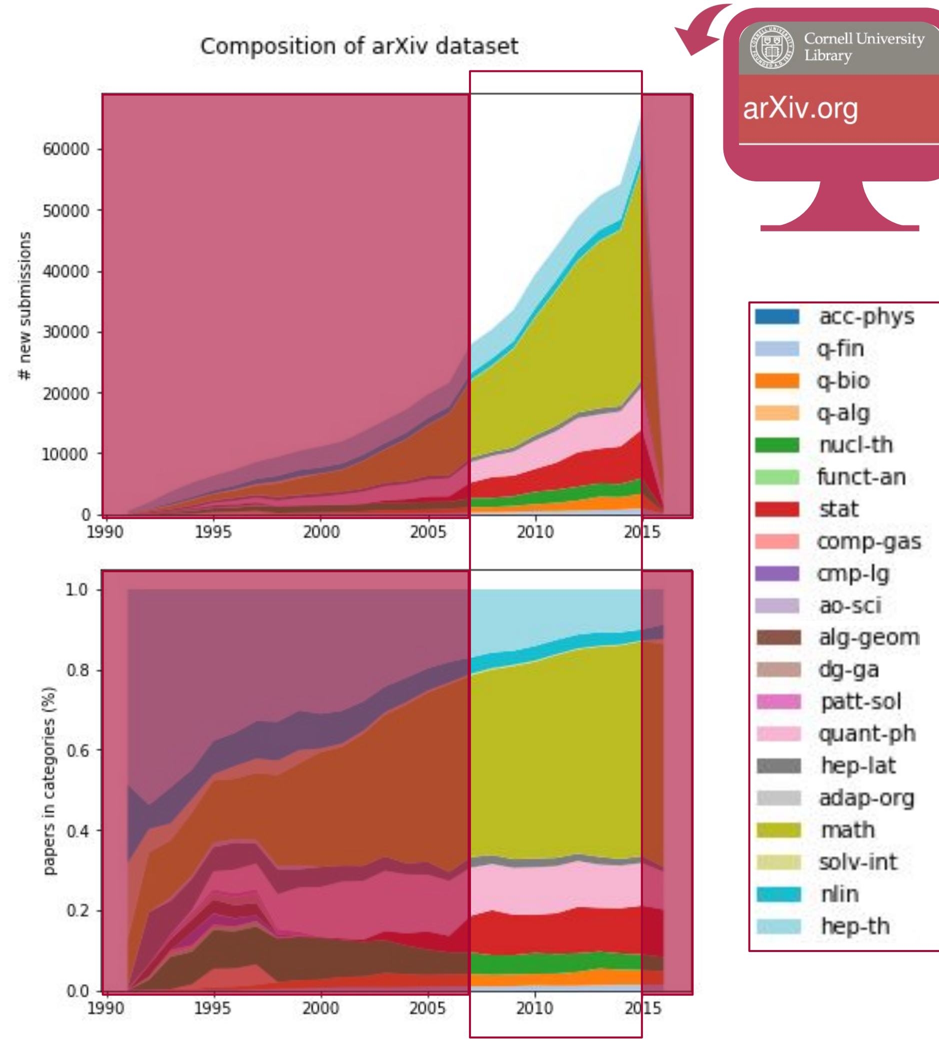


The data set

The data span 9 years, from 2007 to 2016, and are split according to the 18 major categories of arXiv.

This major categories correspond to different thematic areas and thus can be used as rough representative of different scientific fields.

Notice: Due to arXiv's history, there is a bias toward mathematical and physical topics.



Facets size and simplicial degree

Assess commonalities in the statistical properties of the different categories

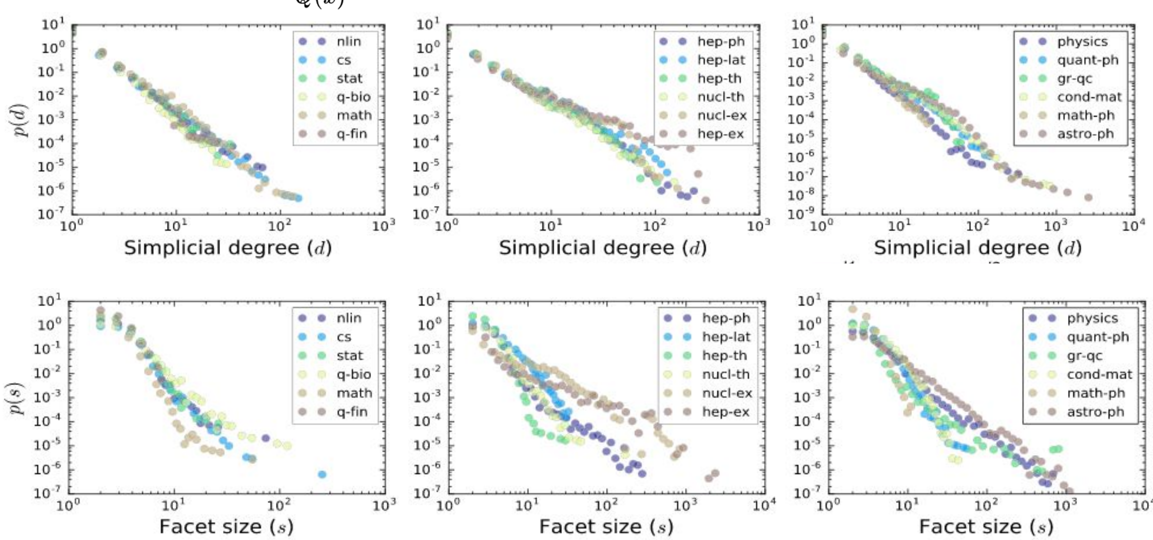
Jensen-Shannon Divergence

$$JSD(P,Q) = \frac{1}{2} D_{KL}(P \mid M) + \frac{1}{2} D_{KL}(Q \mid M)$$

where:

$$M = P + Q$$

$$D_{KL}(P \mid Q) = -\sum_x P(x) \log \frac{P(x)}{Q(x)}$$



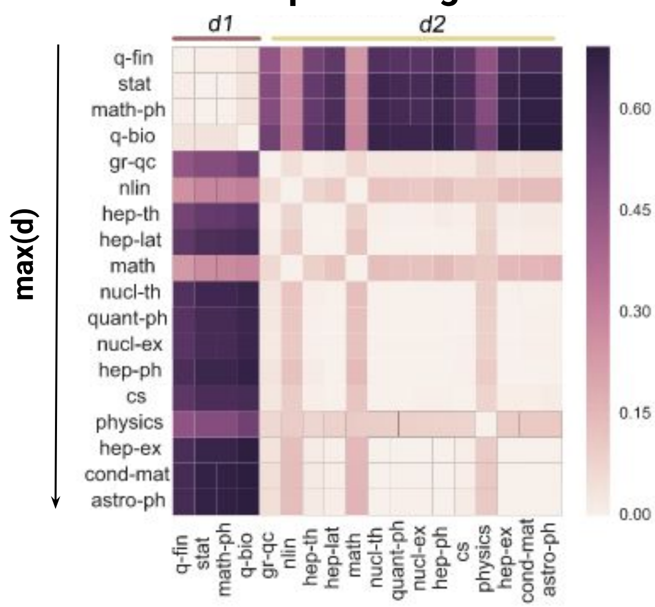
Facets size and simplicial degree

Assess commonalities in the statistical properties of the different categories

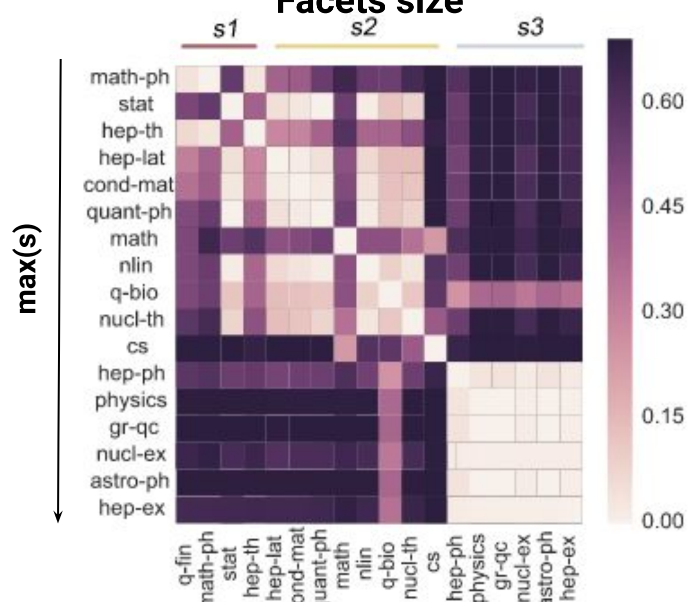
Jensen-Shannon Divergence

$$JSD(P,Q) = \frac{1}{2}D_{KL}(P \mid M) + \frac{1}{2}D_{KL}(Q \mid M)$$

Simplicial degree



Facets size

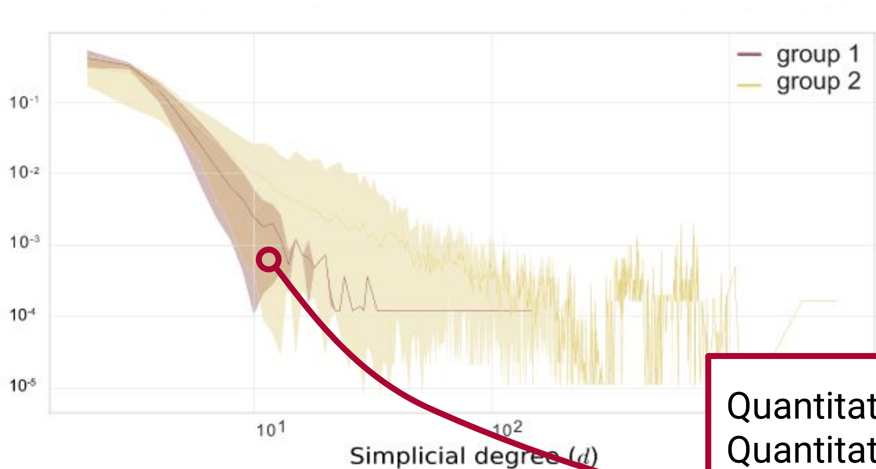


Facets size and simplicial degree

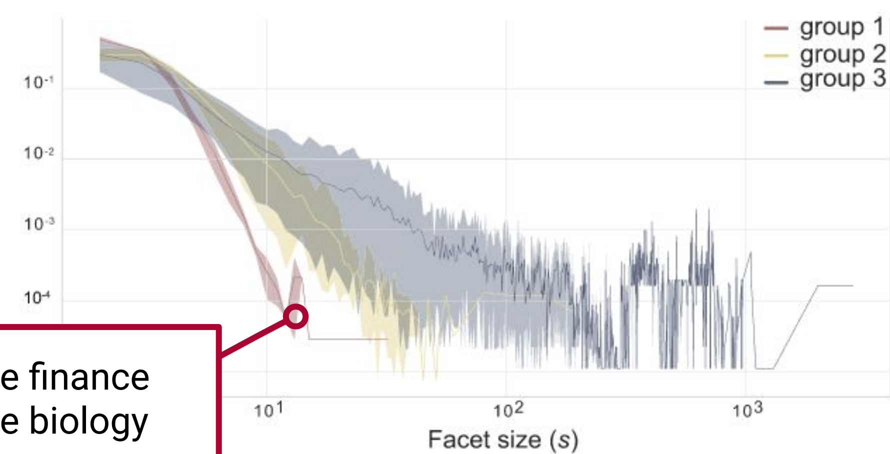
Examples for the biggest connected component for each group.

Facets size and simplicial degree

Simplicial degree



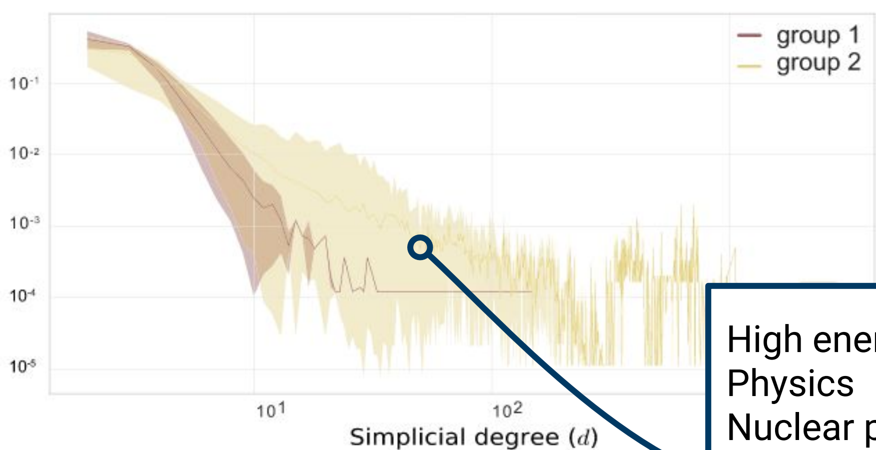
Facets size



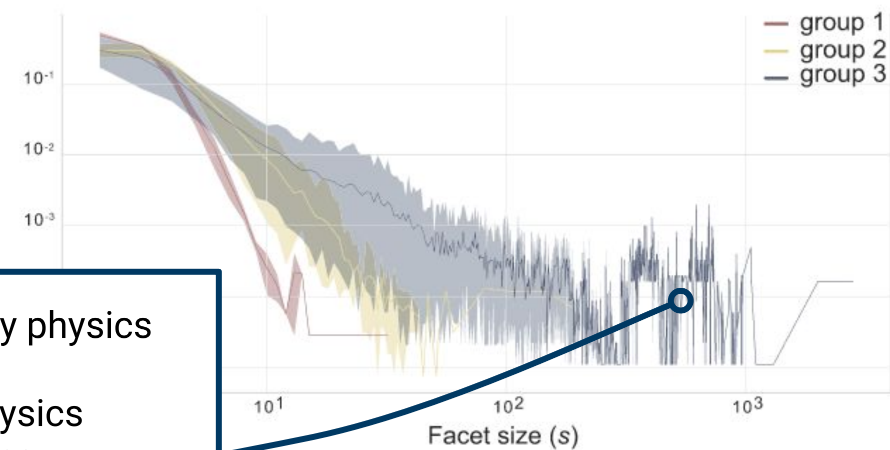
Quantitative finance
Quantitative biology
Statistics
Mathematical physics

Facets size and simplicial degree

Simplicial degree



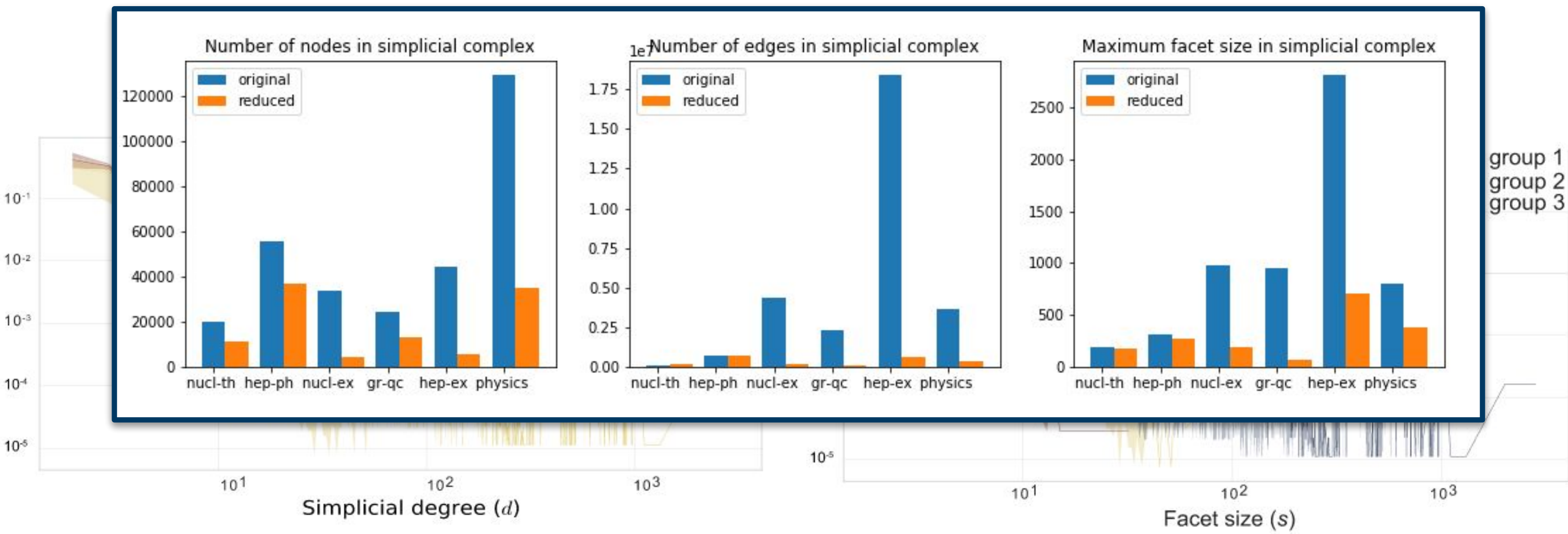
Facets size

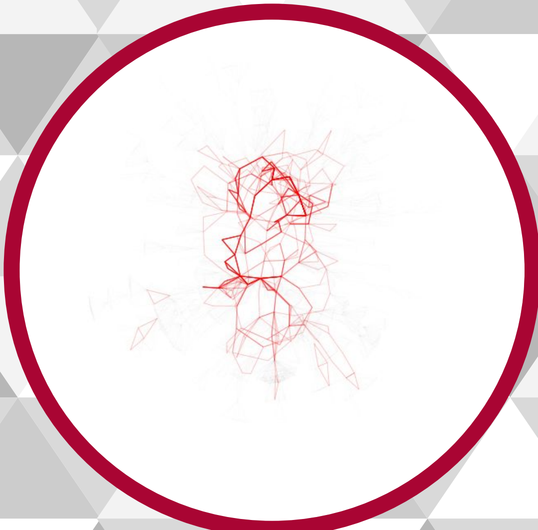


High energy physics
Physics
Nuclear physics
Astrophysics
General relativity

Facets size and simplicial degree

Homology equivalent simplicial complex

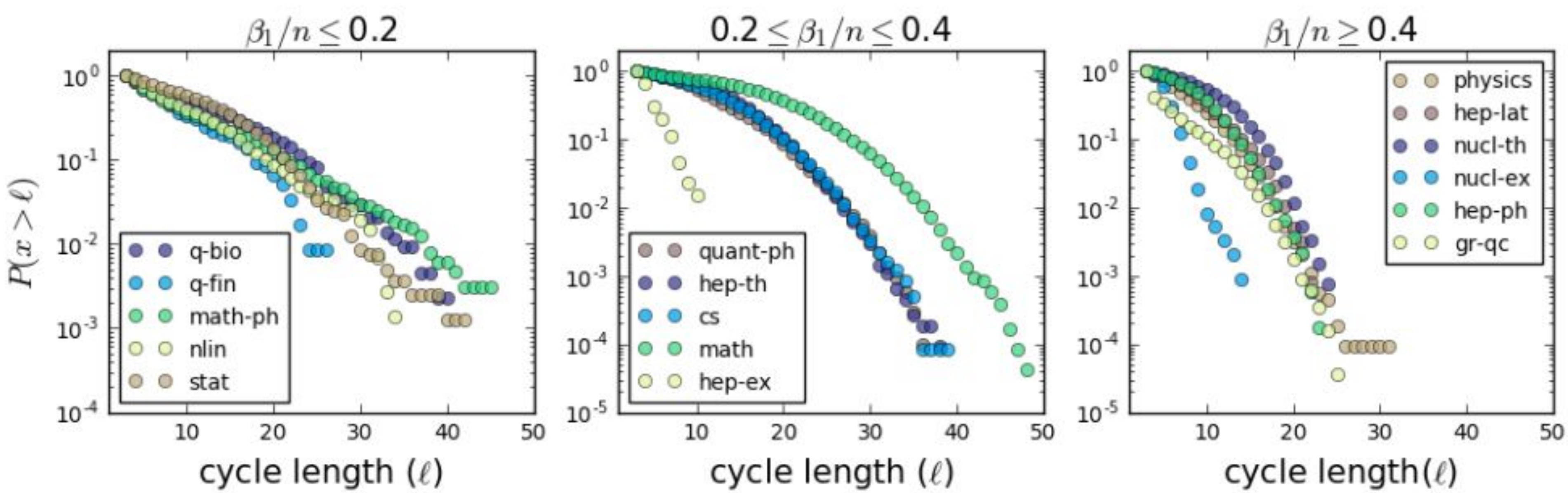




Homological Results

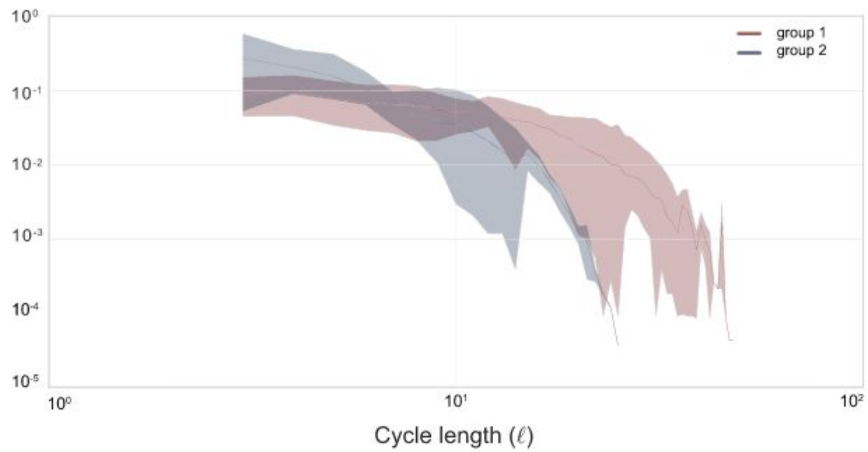
Homology

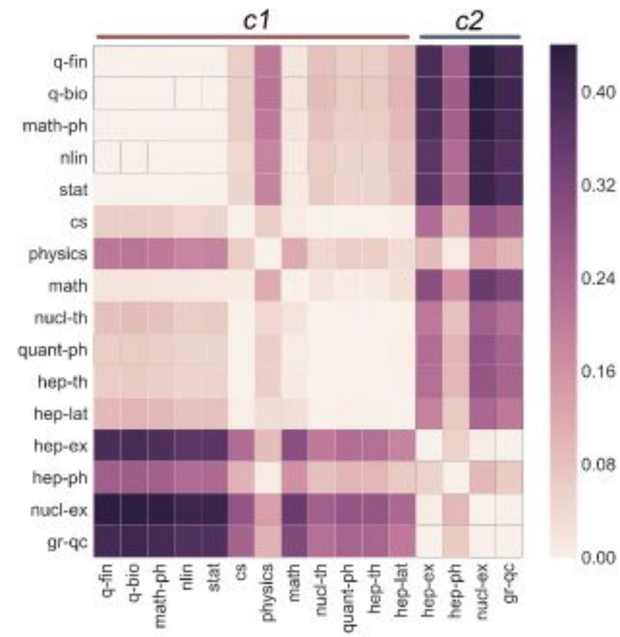
We introduce a new quantity $\boldsymbol{\beta_1/n}$ the ratio of the number of cycles over the number of nodes in the simplicial complex



Homology

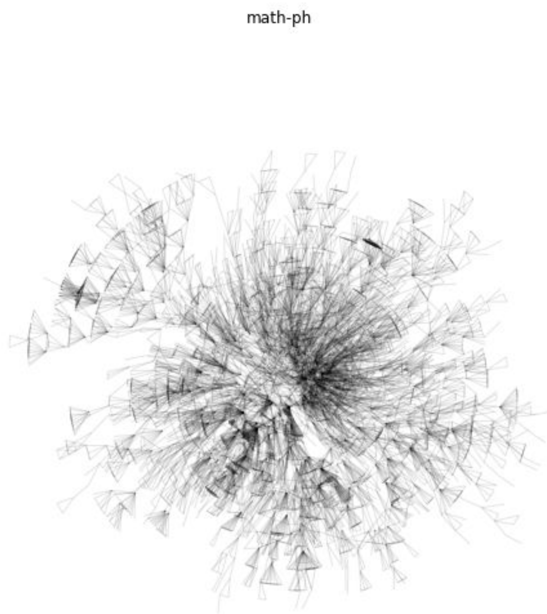
Cycle length distribution

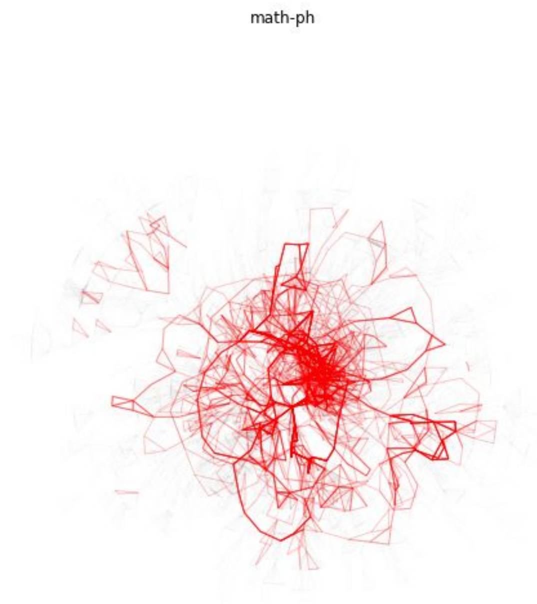




Homology

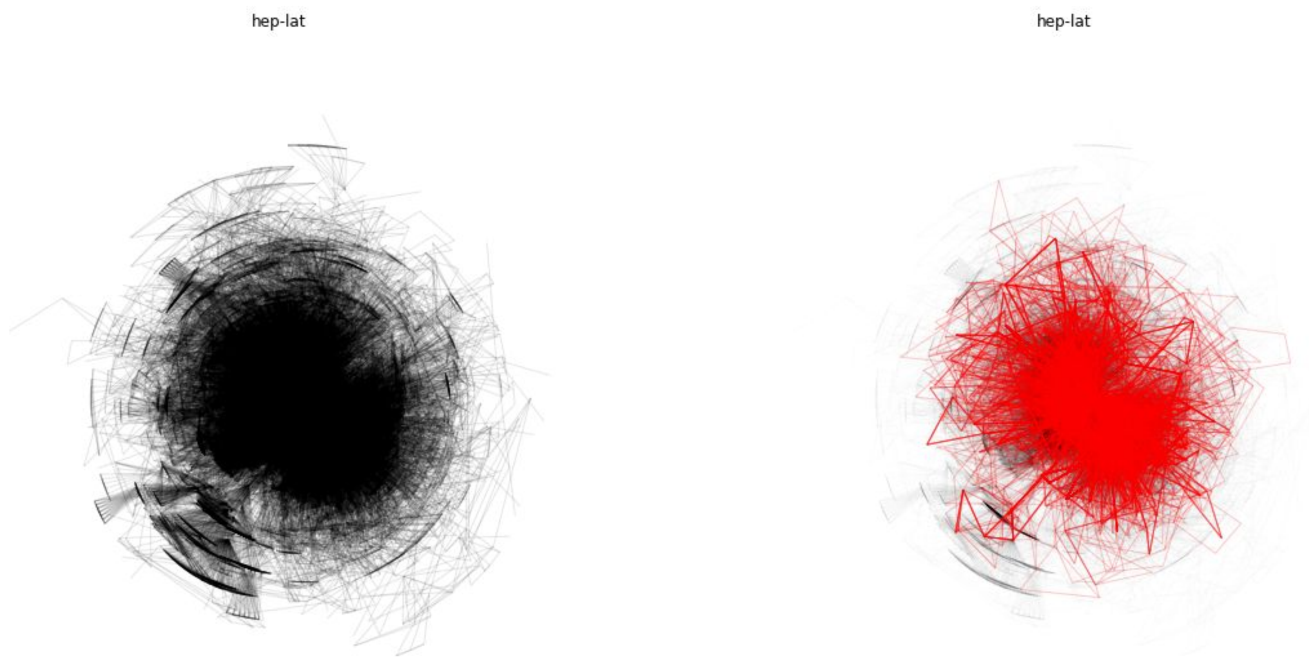
Cycle length distribution





Group 1

Cycles length distribution

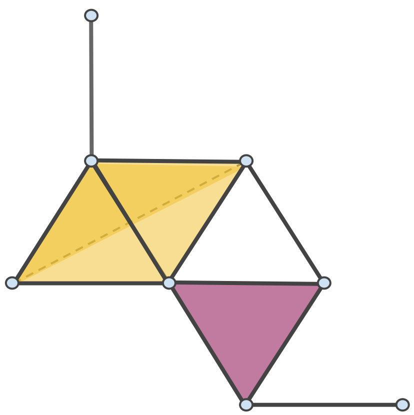


Group 2

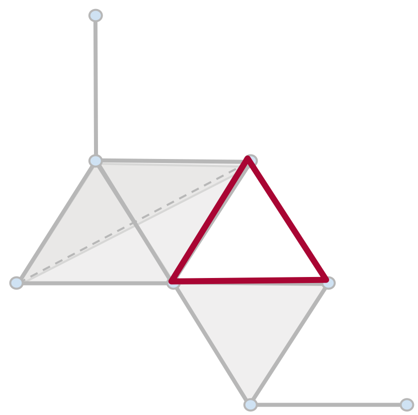
Community detection

Assumption:

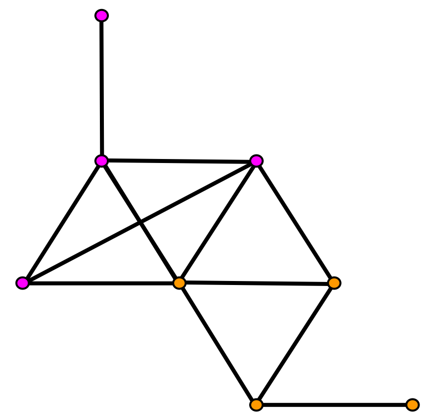
Homological cycles act as bridges between communities of the underlying graph



SIMPLICIAL COMPLEX

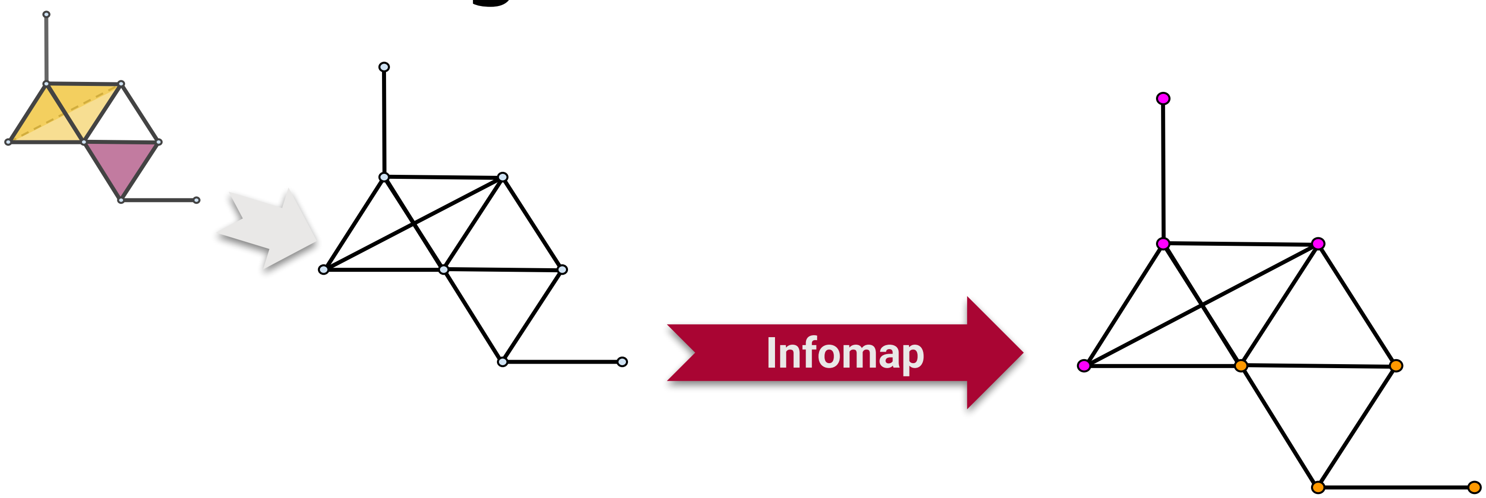


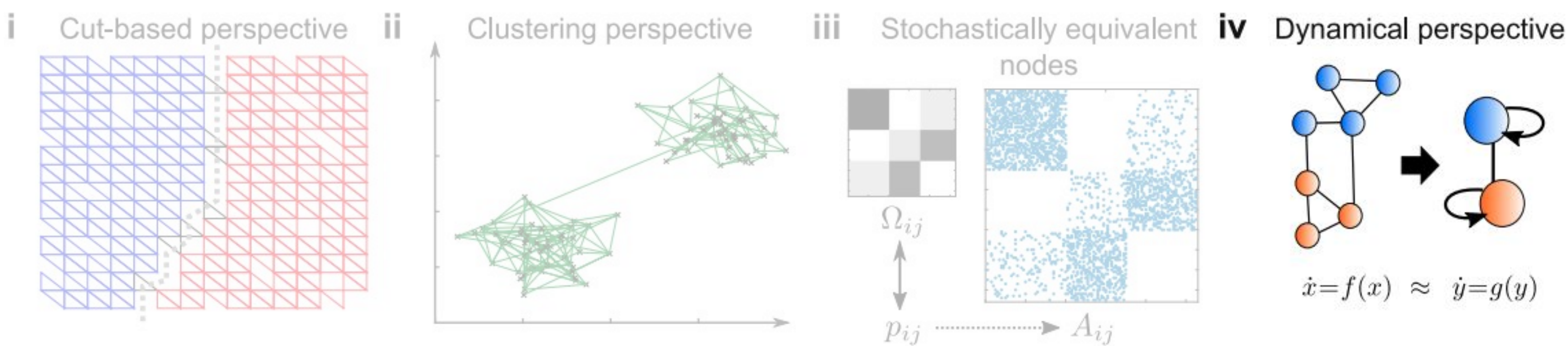
Homological cycles



COMMUNITY detection
in underlying graph

Community detection



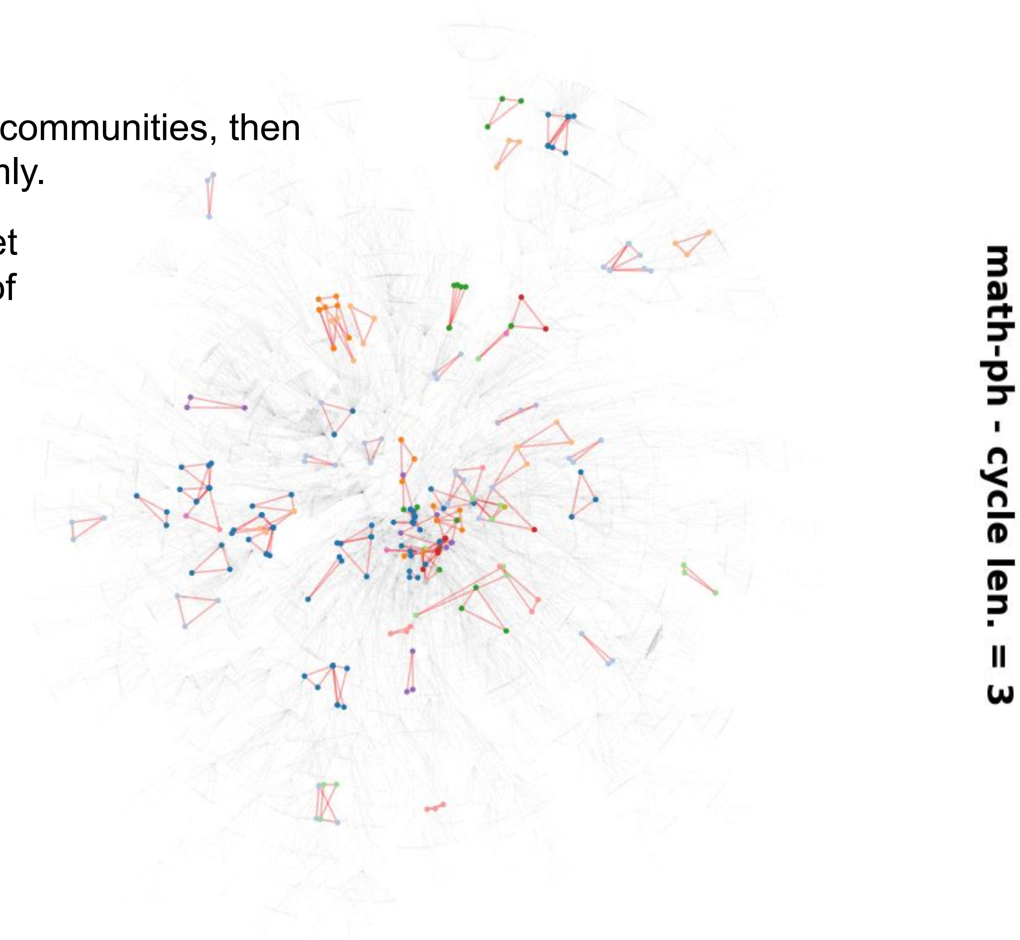


SCHAUB, Michael T., et al. The many facets of community detection in complex networks. *App. Net. Sc*, 2017.

Homology and Communities

If cycles do not act as bridges between communities, then we expect them to go in and out randomly.

But we can clearly see that as cycles get longer the go through a larger number of communities.



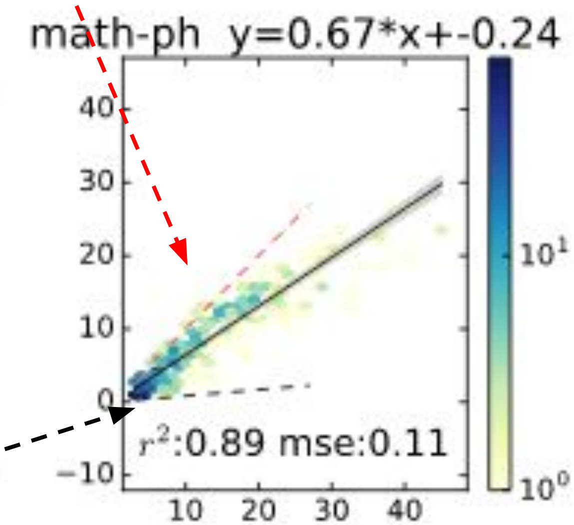
Homology bridges communities?

If cycles do not act as bridges between communities, then we expect them to go in and out randomly.

But we can clearly see that as cycles get longer the go through a larger number of communities.

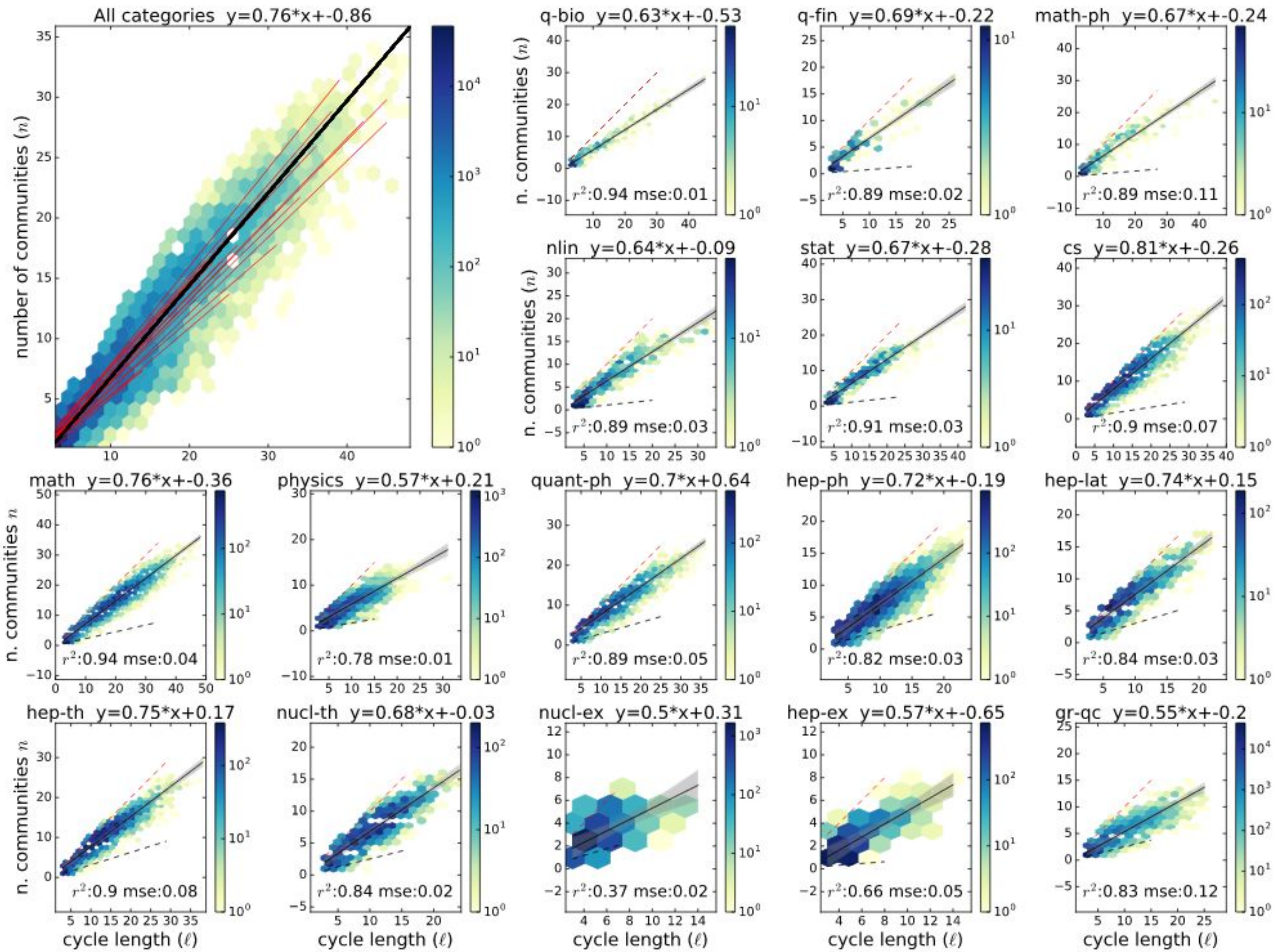
We decide to this as lower bound to assess if a cycle act as bridge in a category, and the length of the cycles as upper bound.

UPPER BOUND
y = x



y=m*x

Where **m** is the fraction of total edges between communities

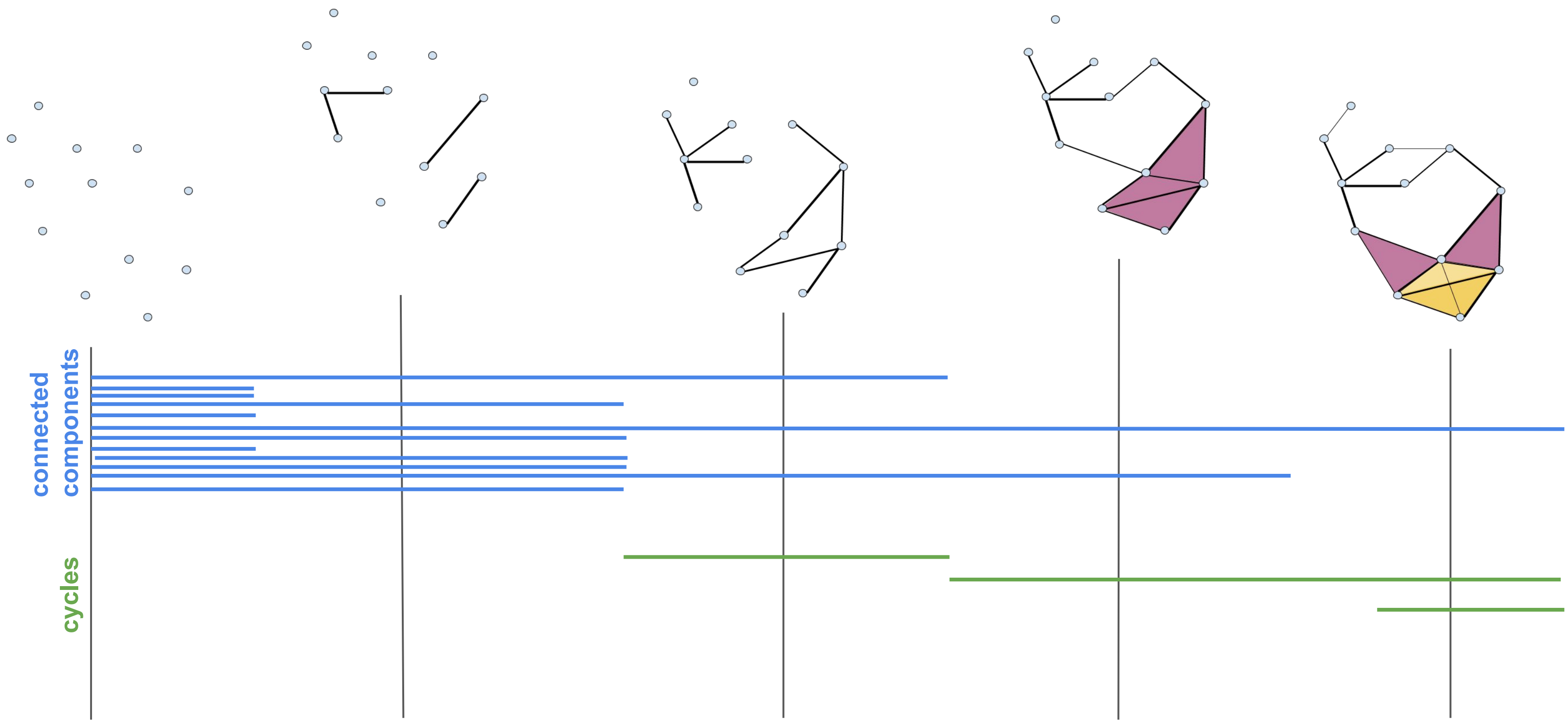


All categories y=0.76*x+-0.86
number of communities (n)
q-bio y=0.63*x+-0.53
r²:0.94 mse:0.01
q-fin y=0.69*x+-0.22
r²:0.89 mse:0.02
math-ph y=0.67*x+-0.24
r²:0.89 mse:0.11
nlin y=0.64*x+-0.09
r²:0.89 mse:0.03
stat y=0.67*x+-0.28
r²:0.91 mse:0.03
cs y=0.81*x+-0.26
r²:0.9 mse:0.07
math y=0.76*x+-0.36
r²:0.94 mse:0.04
physics y=0.57*x+0.21
r²:0.78 mse:0.01
quant-ph y=0.7*x+0.64
r²:0.89 mse:0.05
hep-ph y=0.72*x+-0.19
r²:0.82 mse:0.03
hep-lat y=0.74*x+0.15
r²:0.84 mse:0.03
hep-th y=0.75*x+0.17
r²:0.9 mse:0.08
nucl-th y=0.68*x+-0.03
r²:0.84 mse:0.02
nucl-ex y=0.5*x+0.31
r²:0.37 mse:0.02
hep-ex y=0.57*x+-0.65
r²:0.66 mse:0.05
gr-qc y=0.55*x+-0.2
r²:0.83 mse:0.12
n. communities (n)
n. communities n
cycle length (ℓ)

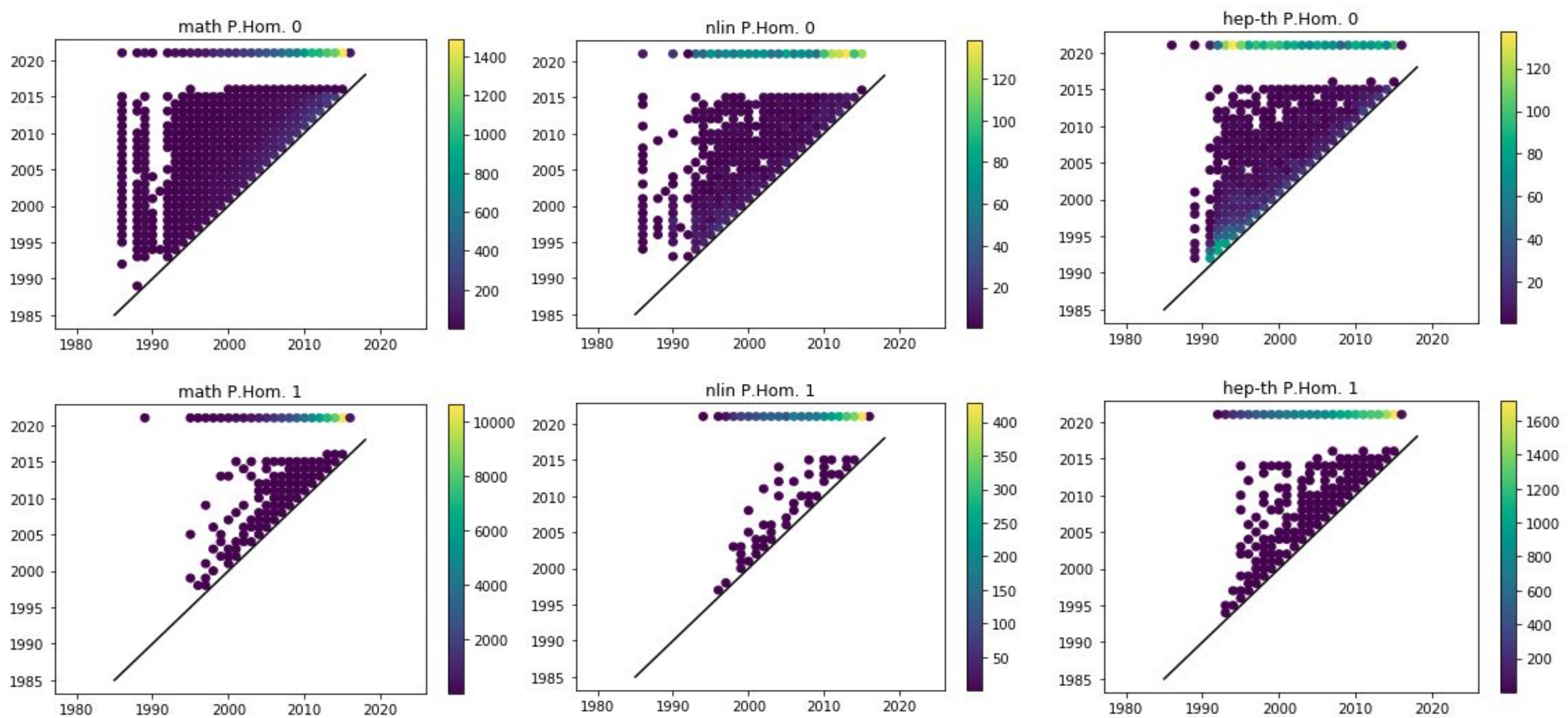
Future work

Future work

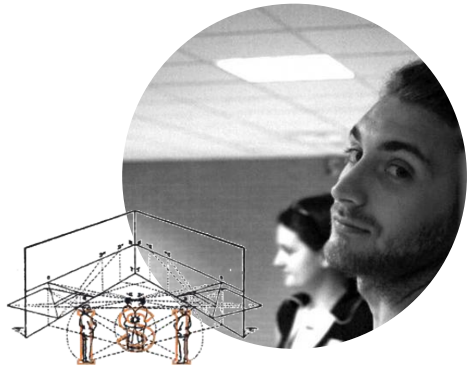
Extending to Persistent Homology



Future work


math P.Hom. 0
nlin P.Hom. 0
hep-th P.Hom. 0
math P.Hom. 1
nlin P.Hom. 1
hep-th P.Hom. 1

Prof. F. Vaccarino



Dott. G. Petri

J.-G. Young

Thank you for the attention

Patania A., Petri G., and Vaccarino F.
"The shape of collaborations." EPJ Data Science 6.1 (2017): 18.

Young J.-G., Petri G., Vaccarino F. and Patania A.
"Construction of and efficient sampling from the simplicial configuration model"
PRE 96 (3), 032312 (2017)

Indiana University
Network Science Institute