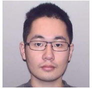
From Fiedler Cuts to Community Detection across Resolution Variation

Duke CS Nov 19, 2024

Dimitris Floros[2]



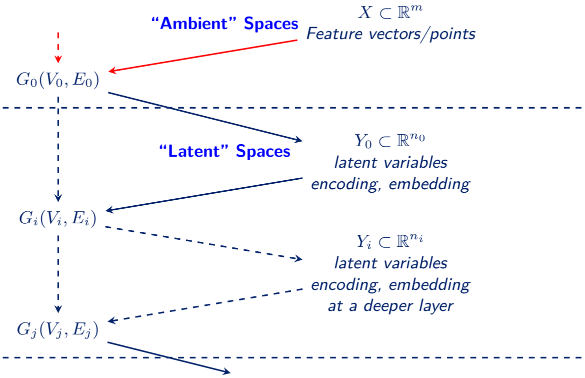
Tiancheng Liu[3]

Nikos Pitsianis [1,3]

Xiaobai Sun[3]

[1]Electrical & Computer Engineering, Aristotle Univ., Thessaloniki, Greece
[2]Nicholas School of the Environment, Duke Univ., Durham, NC, USA
[3]Computer Science, Duke Univ., Durham, NC, USA

Graph Analysis: goals & means

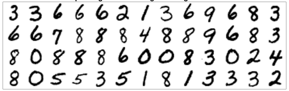


Objectives: $\begin{cases} \text{Differentiation} \\ \text{Identification} \\ \text{Recognition} \end{cases}$ of $\begin{cases} \text{Patterns} \\ \text{Communities} \\ \text{Functional units} \end{cases}$

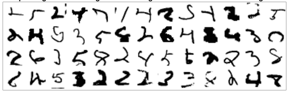
- ubiquitous in data analysis
- understanding
- underlying, unifying

Outline

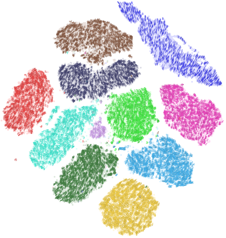
Top-52 digits with the largest in-degree values



Top-52 digits with 0 in-degree and the largest distance to their nearest neighbor



`mnist-handwritten-digit` images (28×28-pixels)



2D embedding of **70000** `mnist-digit` images

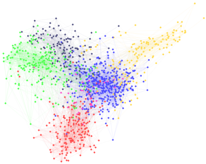
- ▷ Feature points ↦ graphs (discrete manifolds)
- ▷ Community detection a.k.a. graph clustering

  From Fiedler cuts to Modularity

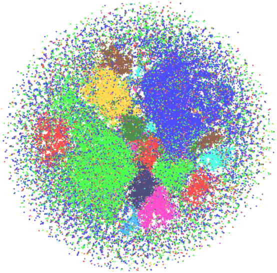
  Resolution limit: with all previous models
- ▷ **BlueRed**: across the resolution variation
- ▷ Needs, opportunities for extensions and applications

☐ Analytical, experimental results throughout the talk

Context-specific networks & abstract graph representation



URV-EMAIL-2003: (n, m) 1133, 5451)



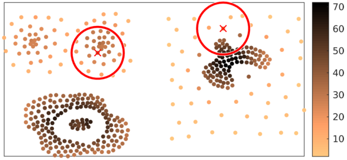
Drosophila connectome[2024] $(n, m) = (110\text{K}, 13.5\text{M})$

Graph $G(V, E)$ Matrix $A(V, V)$

- $n = |V|$ vertices: entities
- $m = |E|$ edges: pairwise entity relationships
- $A(u, v) > 0 \iff (u, v) \in E$
- directed, undirected, weighted, unweighted
- Context-specific networks:
  social, biological, ecological, technological, epidemiological
- Types of graph data sources
  - obtained directly from observation
  - derived from feature vectors/points
  - generated by graph/feature models

From feature-point set to near-neighbor graph

- ▷ X: a feature point cloud in a metric space $\mathbb{R}^D$ with $d(x, x')$
- ▷ $X \to G(X, E)$, a near-neighbor graph: $(x, y) \in E \iff y \in \mathcal{N}(x)$
- ▷ Two types of near-neighborhoods : rNN, kNN

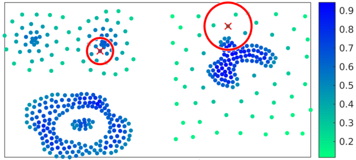


Local density $\boldsymbol{\rho}$ with $\boldsymbol{r} = 3.1$

- ◦ by range within distance r

$$\mathcal{N}(x \mid r) = \{x' \mid d(x, x') < r\}$$

- ◦ parameter r: real-valued, elusive (scale)
- ◦ high-dimension curse:
  overcrowded or vanishing
  highly sensitive to change of r



Dual local density $\boldsymbol{\rho}^*$ with $\boldsymbol{k} = 15$

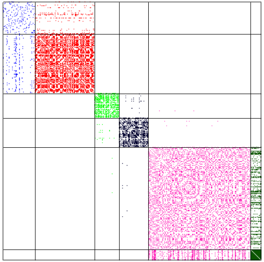
- ◦ by distance ranking to k

$$\mathcal{N}(x \mid k) = \{x_i \mid d(x, x_i) \le d(x, x_{i+1}), i = 1 : k\}$$

- ◦ parameter k: integer, perceptive
- ◦ non-empty, sparse, maintain density information
- ◦ relative robust to change in D (dimension) and k

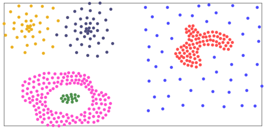
Feature points to rNN and kNN graphs

r**NN**

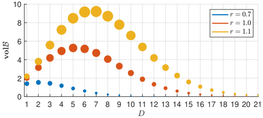


$\mathbf{G}_r$: rNN matrix, symmetric
rows/columns ordered by labeled classes

Compound in $\mathbb{R}^2$

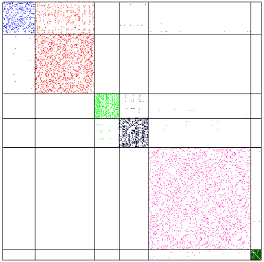


399 feature points in 6 classes
(color-coded)



vanishing volume of spherical balls

k**NN**



$\mathbf{G}_k$: kNN matrix, non-symmetric
rows/columns ordered by labeled classes

Problem description: community detection a.k.a. graph clustering

Community detection in $G(V, E)$ governed by an inference model

$$\Omega_*(h, G) \triangleq \arg \max_{\Omega \in \mathcal{L}(G)} h(\Omega, \Theta)$$

- $\boldsymbol{h}$: clustering function (Hamiltonian-like)
- Θ: hyper-parameter set
- Ω: cluster configuration, $\Omega = \{C_i = G(V_i, E_i) \mid i = 1 : q\}$
- $\mathcal{L}(G)$: the lattice of all feasible cluster configurations, the search space

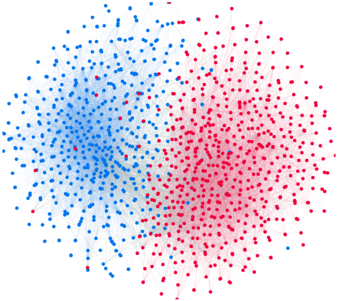
- $\boldsymbol{G}(V, E)$ is connected
- $\{V_i \mid i = 1 : q\}$ is a partition of V, q is part of the solution
- feasibility: cluster $C_i(V_i, E_i)$ is connected

  two primal configurations: $\Omega_\vee$ (all in one), $\Omega_\wedge$ (all singletons)
- encoding: $E \mapsto \Omega$, C_i: community membership (label)

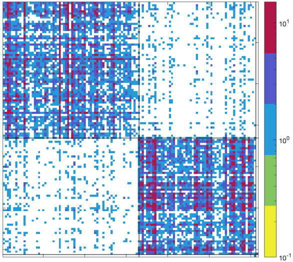
A real-world network: two parties

POLBLOGS-2004

- 2004 US election blogsite: V: 1,224 blog sites, E: 19,025 citation links
- determine: #communities, individual community membership
- our experimental results are consistent with the ground-truth labels $\{\text{DSC}, \text{ARI}\} = \{0.95, 0.81\}$



2D spatial embedding



Adjacency matrix by clusters

Connection between the Fiedler cut & Modularity

Basic case: $\boldsymbol{q \leq 2}$ (cut or not), $A^{\mathrm{T}} = A$ (undirected graph)

The **Fiedler cut:** 1973, 1998

$$\lambda_2(L) = \min_{\substack{v^{\mathrm{T}} v_1 = 0 \\ v^{\mathrm{T}} v = 1}} v^{\mathrm{T}} L v \quad \text{(Fielder value)}$$

$$v_2(L) = \arg \min_{\substack{v^{\mathrm{T}} v_1 = 0 \\ v^{\mathrm{T}} v = 1}} v^{\mathrm{T}} L v \quad \text{(Fielder vector)}$$

- Laplacian: $L = D - A$, $D = \mathrm{diag}(d)$, $Lv_1 = 0$

  or, $L = D^{-1/2}(D - A)D^{-1/2} \sim (I - AD^{-1})$ (normalized)

  AD^{-1}: the random-walk transition matrix
- $\boldsymbol{\lambda_2}$: known as the algebraic connectivity
- $\boldsymbol{v_2}$: gives the sign-parity cut $\&$ vertex ordering

Modularity for community detection, 2004

$$\Omega_* = \arg\min_{\Omega} Q(\Omega) = \sum_{\substack{C \in \Omega \\ i,j \in C}} \frac{A(i,j)}{2m} - \frac{d(i)d(j)}{(2m)^2}$$

where $d = A \times \mathbf{1}$, $2m = d^{\mathrm{T}} \mathbf{1}$

- probabilistic, combinatorial
- not limited to $q \leq 2$, nor to $A^{\mathrm{T}} = A$

  when $q \leq 2$, decide a cut or no cut
- first connection to the algebraic cut, 2013
- first rigorous analysis of the connection, 2023

The Fiedler vector renders the minimal normalized cut

The normalized cut $[C_*, \bar{C}_*]$ with $\bar{C} = V - C$ becomes widely used since 1999

$$C_* = \min_{C \subset V} \frac{\alpha(C,\bar{C})}{\alpha(C,C)} + \frac{\alpha(C,\bar{C})}{\alpha(\bar{C},\bar{C})} = \frac{\alpha(C,\bar{C})}{\text{harmonic-mean}(\alpha(C,C), \alpha(\bar{C},\bar{C}))}$$

where $\alpha(C,C') = \mathbf{1}^{\mathrm{T}} A(C,C')\mathbf{1}$, $C, C' \subset V$

- The minimal cut: normalized by the harmonic mean of the intra-volumes of clusters C and $\bar{C}$
- Search: expensive combinatorially, efficient algebraically
- The Fiedler vector, sorted in as-/descending order, makes both a vertex ordering (1D encoding) *and* the cut
  - the vertex ordering: neighbors are closer
  - the cut: at the sign change
  - numerical location of the cut can be stabilized
- Left plot: the first cut at 50 among 150 flowers in dataset IRIS with 4 feature attributes

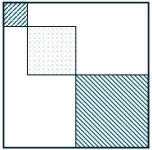
Modularity: reformulation & reinterpretation

Modularity Q (clustering function), simple, significant, widely used since 2004

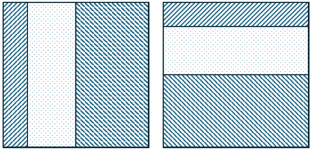
$$\max Q(\Omega) = \sum_{C \in \Omega} \alpha(C,C) - \alpha(C,V)\alpha(V,C) = Q_{\text{attraction}} + Q_{\text{repulsion}}$$

where

$$\alpha(C,S) = \mathbf{1}^{\mathrm{T}} A(C,S)\mathbf{1}, \qquad C,S \subset V, \quad \alpha(V,V) = 1$$



Attraction term $\boldsymbol{Q}_{\textbf{attraction}}$
the sum of intra-volumes
increases with **merges**



Repulsion term $\boldsymbol{Q}_{\textbf{repulsion}}$
the sum of inter-volume-products
increases with **splits**

Model evolution: the tale of modularity $h = Q$

- Modularity parameterized, 2006

$$Q(\Omega\,|\,\boldsymbol{\gamma}) = \sum_{C\in\Omega} \alpha(C,C) - \boldsymbol{\gamma}\, \alpha(C,V)\alpha(V,C)$$

$\alpha(C,C') = \mathbf{1}^{\mathrm{T}} A(C,C')\mathbf{1},\ C,C' \subset V,\ \alpha(V,V) = 1$

$\boldsymbol{\gamma} \in [\mathbf{0},\boldsymbol{\infty})$: **resolution** parameter(external)

$Q(\Omega \mid \gamma = 1)$: the original modularity

- $Q(\gamma,\Omega)$: γ as an internal variable, 2021

no longer external, free of γ-tuning

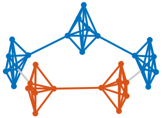
- $Q_{\mathrm{stoch}}(\gamma,\Omega)$: stochastic model, 2024

unsupervised attention to robust configurations

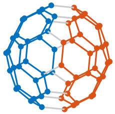
Progressive understanding of the $\boldsymbol{\gamma}$-dynamics:

- the resolution limit at $\gamma = 1$, 2007
- the resolution limit at any particular γ-value, 2023
  * the critical value γ_c in cut $(q \leq 2)$ is $\boldsymbol{\lambda}_2(L)$
  * the Fiedler value under stochastic fluctuations $\lambda_2 \Longrightarrow \Lambda_2$, the Fiedler pseudo-set (FPS)
- **BlueRed** across the γ-spectrum, 2021-2023
  * no single resolution γ-value is universal
  * universal: $\exists$ a unique set of γ-bands by p critical values $\{\gamma_i < \gamma_{i+1}, i = 1:p\}$
- **BlueRed** admitting stochastic fluctuations, 2024
  * recognizing persistent and steady γ-bands

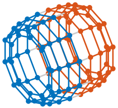
Split transition analysis: the Fiedler connection to modularity Q



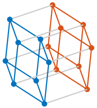
$\mathbf{RoK}_{5,6}$, [2007]
Ring-5 of K_6-cliques



Buckyball, C_{60}



Cylindrical mesh 10×10



Cube_{16}, 4D hypercube

Splits by Fiedler Cuts

- Two-cluster configuration $\Omega_{\pm 1} = \{C_1, C_{-1}\}$
- Q_γ: at a fixed γ-value
- Connection to the (plain) Laplacian

$$Q_\gamma(\Omega_{\pm 1}) = \frac{s^{\mathrm{T}}\,(A - \gamma\, dd^{\mathrm{T}})\, s}{2s^{\mathrm{T}}s}, \qquad s \in \{1, -1\}^n,$$

Connection to the normalized Laplacian

$$Q_\gamma(\Omega_{\pm 1}) = \frac{y(s)^{\mathrm{T}}\,\left(\hat{A} - \gamma\, \hat{d}\hat{d}^{\mathrm{T}}\right)\, y(s)}{y(s)^{\mathrm{T}} y(s)},$$

$$\hat{d} = d^{1/2}, \quad y(s) = \hat{d} \odot s, \quad \hat{A} = \hat{D}A\hat{D}, \quad \hat{D} = \mathrm{diag}\left(1./\hat{d}\right)$$

- Our findings [2023]:
  - reveal the critical value γ_c between cut and no-cut
  - show the resolution limit at any particular value γ_o
  - explain the erroneous split or merge when $\gamma_c \neq \gamma_o$

Analysis of the critical transition

Theorem

- G: undirected, connected graph
- $\gamma > 0$: resolution parameter value
- Q_γ: parametrized modularity
- λ_2: the Fiedler value of $L = I - \hat{A}$

Then,

◇ $\gamma < \lambda_2$: no split in $Q_{*,\gamma}$

◇ $\gamma > \lambda_2$: split in $Q_{*,\gamma}$

◇ $\gamma = \lambda_2$: $Q_{*,\gamma}$ nonunique

`Implications:`

- ▷ $\boldsymbol{\lambda_2(G)}$ marks the critical transition point $\boldsymbol{\gamma_c}$ between cut and no-cut by Q_γ
- ▷ Ω_γ at any $\gamma \neq 1$ is as resolution-limited as Ω_1

  i.e., no single magic γ-value serves for all graphs

  while γ is fixed in Q, λ_2 varies with G

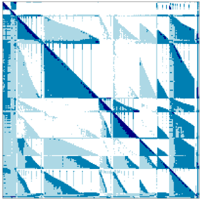
  `f.g.` $\lambda_2(\texttt{RoK}_{17,6}) = 0.003$, $\lambda_2(\texttt{Cube}_8) = 0.25$
- ▶ It is **necessary** to free Q from γ-fixing/tuning

Additional issues with the Fiedler pair $(\boldsymbol{\lambda_2}, \boldsymbol{v_2})$:

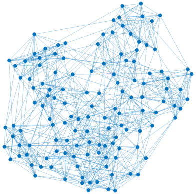
- ◦ $\boldsymbol{\lambda}_2$: overfitting to data graph G
- ◦ $\boldsymbol{v_2}$: non-unique when λ_2 is not simple

  `f.g.` isomorphic split patterns in `cubes`, `torus`

BlueRed: across resolution variation



APS-2020 citation: matrix in DOI order



FBS-2023 network: 132 American college football teams (nodes) of Football Bowl Subdivision (FBS) play 738 games (edges) in the regular season of 2023

- A broad family of clustering functions: $\boldsymbol{h}$ existing or novel
- Clustering analysis of $\min \boldsymbol{h}(\boldsymbol{\gamma})$ across $\gamma \in [0, \infty)$
  (recall: necessary to be free of γ-setting or turning)
- A universal γ-spectral structure with p transition points
  $\{(\Omega_j, \Gamma_j),\ j = 1 : p\}$, each Ω_j on γ-band $\Gamma_j = (\gamma_{j-1}, \gamma_j)$
  ($\exists$ a unique set of $p \geq 1$ critical transition points $\{\gamma_j\}$)

  A master optimization model
  (parallel to the Courant-Fischer theorem with matrix spectral)
- Theoretical γ-spectral analysis of KSC graphs (sketched)
- Descending triangulation (DT) algorithms (sketched)
- Computational γ-spectral analysis of graphs of diverse types

BlueRed: clustering function family

$$h(\gamma, \Omega) \quad = h_{\textbf{attraction}}(\Omega) \quad + \gamma \; h_{\textbf{repulsion}}(\Omega),$$

$$\Omega \in \mathcal{L}(G) \qquad \text{\% lattice of feasible configurations}$$
$$\gamma \in (0, \infty) \qquad \text{\% internal resolution variable}$$

- `attraction` & `repulsion` terms: *non-colinear* and *dialectical*

$$\Omega_{\vee} = \arg \min_{\Omega \in \mathcal{L}(\mathcal{G})} h_{\text{attraction}}(\Omega), \qquad \Omega_{\wedge} = \arg \min_{\Omega \in \mathcal{L}(\mathcal{G})} h_{\text{repulsion}}(\Omega)$$

- including/unifying most existing models (w./w.o. modification) and open to novel models
- inspecting the community structure in graph G through the lenses of h
- encoding, mapping (2D): $\Omega \mapsto (h_{\text{a}}(\Omega), h_{\text{r}}(\Omega))$, i.e., $\boldsymbol{\mathcal{L}(G)} \mapsto$ **har(*G*)-plane**

Clustering functions in BlueRed expression

	BlueRed expression	
Existing clustering function	**Attraction h_a**	**Repulsion h_r**
Altered q-state Potts model $-J\sum_{(i,j)\in E}\delta_{\sigma_i\sigma_j}+\gamma\sum_{s=1}^{q}\frac{n_s(n_s-1)}{2}$	$-J\sum_{C_i\in\Omega}\alpha(C_i,C_i)$	$\sum_{C_i\in\Omega} n_i\,(n_i-1)$
Absolute Potts model $\sum_s(-w_s+\gamma u_s)$	$-\sum_{C_i\in\Omega}\alpha(C_i,C_i)$	$\sum_{C_i\in\Omega} n_i\,(n_i-1)-\alpha(C_i,C_i)$
Constant Potts model $-\sum_{ij}(A_{ij}w_{ij}-\gamma)\,\delta_{\sigma_i\sigma_j}$	$-\sum_{C_i\in\Omega}\alpha(C_i,C_i)$	$\sum_{C_i\in\Omega} n_i\,(n_i-1)$
Modularity $\frac{1}{2m}\sum_{ij}\left(A_{ij}-\gamma\frac{k_ik_j}{2m}\right)\delta_{\sigma_i\sigma_j}$	$-\sum_{C_i\in\Omega}\alpha(C_i,C_i)$	$\sum_{C_i\in\Omega}\alpha^2(C_i,V)$
Normalized cuts $\frac{\alpha(C_1,C_2)}{\alpha(C_1,V)}+\frac{\alpha(C_1,C_2)}{\alpha(C_2,V)}$	$-\sum_{C_i\in\Omega}\frac{\alpha(C_i)}{\alpha(C_i,V)}$	$\sum_{C_i\in\Omega}\frac{\alpha(C_i,\overline{C}_i)}{\alpha(C_i,V)}$
Generalized modularity density $\frac{1}{2m}\sum_c\left(2m_c-\frac{K_c^2}{2m}\right)\rho_c^\chi$	$-\sum_{C_i\in\Omega}\alpha(C_i)\cdot\frac{\rho^\chi(C_i)}{\sum_{c_j\in\Omega}\rho^\chi(C_j)}$	$\sum_{C_i\in\Omega}\alpha^2(C_i,V)\cdot\frac{\rho^\chi(C_i)}{\sum_{c_j\in\Omega}\rho^\chi(C_j)}$
Infomap $L(M)=q_\curvearrowright H(\mathcal{Q})+\sum_{i=1}^{m}p_\circlearrowright^i H(\mathcal{P}^i)$	$\sum_{C_i\in\Omega} f\left(\frac{\alpha(C_i,C_i)+2\alpha(C_i,\overline{C}_i)}{2e}\right)$	$f\left(\frac{1}{e}-\sum_{C_i\in\Omega}\frac{\alpha(C_i,C_i)}{e}\right)-2\sum_{C_i\in\Omega} f\left(\frac{\alpha(C_i,\overline{C}_i)}{e}\right)$

($*$) The functions in the bottom group are modified to introduce γ-variation while the specific encoding principle for each is maintained

Multiple critical transition points across γ-variation

Theorem

- G: connected, un-/directed, un-/weighted
- $h(\gamma)$: a BlueRed function, $\gamma \in [0, \infty)$

$\exists$ a unique set of p critical transition values

$$\{\gamma_j, j = 1, 2, \cdots, p\}, \quad 0 < p < \infty$$
$$\gamma_{j-1} < \gamma_j, \quad \gamma_0 = 0$$

and a finite set of configurations

$$\boldsymbol{\Omega_j} = \arg \min_{\substack{\Omega \in \mathcal{L}(G) \\ \gamma \in \Gamma_j}} h(\gamma, \Omega), \quad \boldsymbol{\Gamma_j} = (\gamma_{j-1}, \gamma_j)$$

BlueRed Front:

$$\mathcal{B}(h, G) \triangleq \{ (\Omega_j, \Gamma_j), 1 \le j \le p \}$$

- $\{\Gamma_j\}$: γ-spectral bands, non-overlapping
  - $*$ analogous to multispectral imaging: image/band
  - $*$ #bands, band locations, bandwidth: not prescribed
- $\{\Omega_j\}$: Ω_j optimal over band Γ_j, not a single value
- State transition at each and every critical value γ_j

  $$\Omega_j \to \Omega_{j-1} \text{ or } \Omega_j \to \Omega_{j+1}$$

  explaining the *sensitivity* at γ close to a critical value

  or the relative *robustness* at γ away from critical values

The master model & the Courant-Fischer theorem

- critical values: $\gamma_{j-1} < \gamma_j$, $j \le p$
- configurations: $\Omega_j \in \mathcal{L}(G)$, a lattice
- $\boldsymbol{h}$: nonlinear on a nonnegative matrix
- equivalent expressions: sup-inf or inf-sup
- probabilistic, combinatorial
- #bands unknown a priori

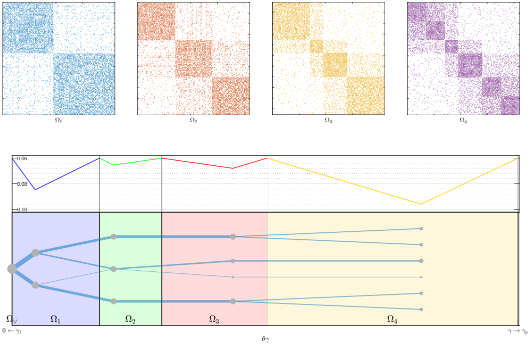
- eigenvalues: $\lambda_{j-1} \le \lambda_j$
- eigenvectors: $\boldsymbol{v}_j \in \mathbb{F}^n$, a vector field
- Rayleigh quotient on a Hermitian matrix
- equivalent expressions: max-min or min-max
- algebraic, geometric
- #distinct eigenvalues unknown a priori

$$\gamma_1 = \sup_{\gamma_1' \in (0,\infty)} \Big\{ \Omega_\vee = \arg\inf_{\substack{\Omega \in \mathcal{L}(G) \\ \gamma \in [0, \gamma_1')}} h(\gamma, \Omega) \Big\},$$

$$\gamma_j = \sup_{\gamma_j' \in (\gamma_{j-1},\infty)} \Big\{ \Omega_j = \arg\inf_{\substack{\Omega \in \mathcal{L}(G) \\ \gamma \in [\gamma_{j-1}, \gamma_j')}} h(\gamma, \Omega) \Big\}, \quad 2 \le j \le p$$

$$\Omega_\wedge = \arg\inf_{\substack{\Omega \in \mathcal{L}(G) \\ \gamma \in [\gamma_p, \infty)}} h(\Omega, \gamma).$$

Illustration of γ-resolution band structure in a graph

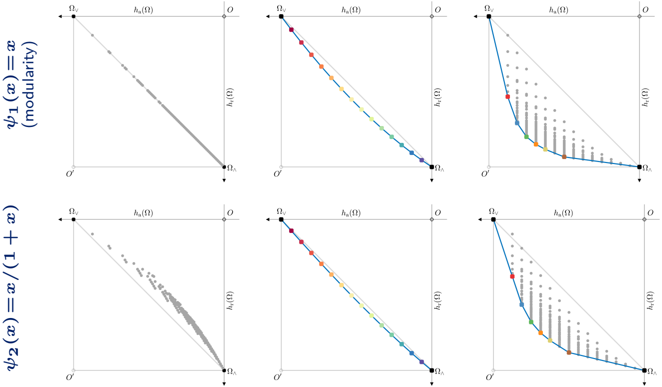


- ▷ Middle: the relative energy over the four γ bands. The *peak* locations coincide with $\{\gamma_j\}$. Configuration Ω_j at the *dip* location of Γ_j is more robust to changes
- ▷ Bottom: the lineage pyramid relates clusters in Ω_j across the γ-bands

- ▷ Graph G_{sbm3}: generated by a SBM model with 3 block modes
- ▷ Four cluster configurations Ω_j are shown in the top with the adjacency matrices in corresponding orders, color-coded
- ▷ The four supporting γ-bands Γ_j are shown in the bottom
- ▷ Ω_{blue}, Ω_{red} and Ω_{purple} recover the three generation modes, Ω_{yellow} is induced

- $\mathbf{\Omega_{blue}}$ and $\mathbf{\Omega_{red}}$ are anti-chain on lattice $\mathcal{L}(G_{\text{sbm3}})$ Their coexistence on $\mathbf{\Gamma_{blue}}$ and $\mathbf{\Gamma_{red}}$ cannot be detected accurately at a single γ-value
- This also explains the fuzziness observed at the critical value between $\mathbf{\Gamma_{blue}}$ and $\mathbf{\Gamma_{red}}$

BlueRed Fronts (BRF) & the critical values in the HAR plane



clique K_{16} star S_{16} cycle C_{16}

◇ Theoretical BRFs for KSC graphs:
clique K_n, star S_n, cycle C_n
(closed-form expressions omitted)

◇ HAR-images of $\mathcal{L}(G)$ by two functions

$$\mathrm{har}(\mathcal{L}(G)) \triangleq \{(h_a(\Omega), h_r(\Omega)) \mid \Omega \in \mathcal{L}(G)\}$$

$$h_a(\Omega) = \sum_{C\in\Omega} \alpha(C, C)$$

$$h_r(\Omega) = \sum_{C\in\Omega} \alpha(C, V)\psi_i(\alpha(C, V)), i = 1, 2$$

- $\mathrm{BRF}(h, G)$ encloses $\mathrm{har}(\mathcal{L}(G))$ in the smallest convex hull
- Critical values: $\gamma_j = \mathrm{slope}\left(\overline{\Omega_{j-1}\Omega_j}^{\perp}\right)$

These BRFs also serve for alg. testing

BRF construction: Descending Triangulation I

Descending Triangulation (I)

Input: h // a **BlueRed** clustering function

G // a connected graph

Output: $\mathrm{BRF}(h,G) = \{(\Omega_j, \Theta_j) \mid j = 0, 1, \ldots, p\}$

Initialization: $\mathcal{B}_\omega \leftarrow \{\Omega_\wedge, \Omega_\vee\}$ // the primal configurations

$\mathcal{B}_\theta \leftarrow \{0, \pi/4, \pi/2\}$

$L \leftarrow \{\, \overline{\Omega_\wedge \Omega_\vee} \,\}$ // active line segments

while (L is not empty) **do**

  $\ell_{\mathrm{ab}} \triangleq \overline{\Omega_{\mathrm{a}}\Omega_{\mathrm{b}}} \leftarrow \mathrm{nextSegment}(\,L\,)$; $L \leftarrow L - \ell_{\mathrm{ab}}$

  $\theta_{\mathrm{ab}} \triangleq \mathrm{atan}(s^{\perp}(\ell_{\mathrm{ab}}))$

  $\Omega_{\mathrm{c}} \leftarrow \arg\min_{\Omega} h(\,\Omega, \theta_{\mathrm{ab}}\,)$ // θ-specific minimization

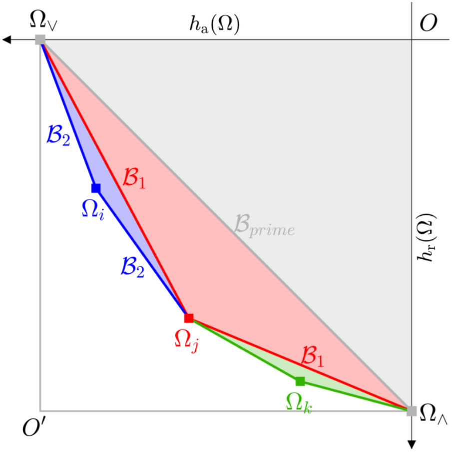
  if (Ω_c is below ℓ_{ab}) **then**

    $\mathcal{B}_\omega \leftarrow \mathcal{B}_\omega \cup \{\Omega_c\}$; $\mathcal{B}_\theta \leftarrow \{\theta_{\mathrm{ac}}, \theta_{\mathrm{cb}}\} \cup \mathcal{B}_\theta / \{\theta_{\mathrm{ab}}\}$

    $L \leftarrow L \cup \{\,\ell_{ac}, \ell_{cb}\,\}$ // update the active set

$\mathrm{BRF}(h,G) \leftarrow \mathrm{combine}\,(\mathcal{B}_\omega, \mathcal{B}_\theta)$

* locate BRF configurations with simple, successive triangulation steps
* arbitrary initial $\gamma(\theta)$
* autonomous termination in p steps

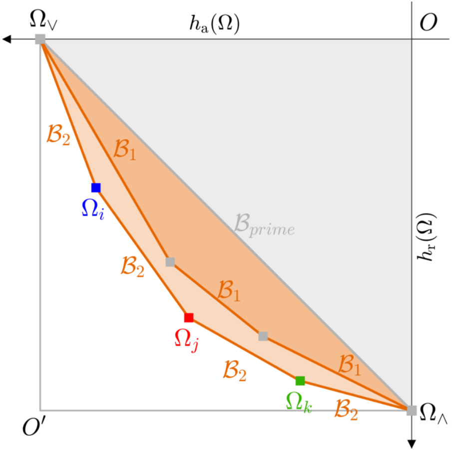


BRF construction: Descending Triangulation II

Descending Triangulation (II)

Input: h // a **BlueRed** clustering function
G // a connected graph
$\mathcal{B}_0$ // a valid bisection front
τ // a threshold on angular precision
Output: $\mathrm{BRF}_{\mathrm{sa}}(h, G) = \{(\Omega_j, \Theta_j) \mid j = 0, 1, \ldots, p\}$
Initialization: $\mathcal{B}_j \leftarrow \mathcal{B}_0$ // current bisection front
while ($\mathcal{B}_j$ has not converged) **do**
  $\ell_{\mathrm{ab}} \triangleq \overline{\Omega_{\mathrm{a}}\Omega_{\mathrm{b}}} \leftarrow \mathrm{selectSegment}(\mathcal{B}_j, \tau)$
  $\theta_{\mathrm{ab}} \triangleq \mathrm{atan}(s^{\perp}(\ell_{\mathrm{ab}}))$
  $\Omega_{\mathrm{c}} \leftarrow \arg \operatorname*{saDescend}_{\Omega} h(\Omega, \theta_{\mathrm{ab}})$ // stochastic descending
  if (Ω_c is below $\mathcal{B}_j$) **then**
    $\mathcal{B}_j \leftarrow \mathrm{updateBisectionFront}(\mathcal{B}_j \cup \{\Omega_c\}, \tau)$
$\mathrm{BRF}_{\mathrm{sa}}(h, G) \leftarrow \mathcal{B}_j$

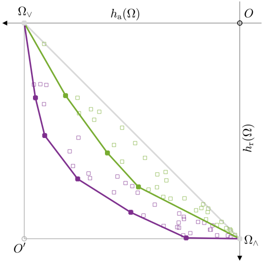
* based on an additional optimality bisection front $\searrow$ BRF
* addressing computational issues including stochastic approaches for optimization



Descending triangulation in parallel (parallel-DT)

* LEIDEN[1]: γ-specific, with stochastic searches, sequential
* known limitations in parallelizing γ-specific clustering[2]
* in need of parallel clustering across resolution variation

Novelty: integrating stochastic searches across resolution variation
even accelerating γ-specific searches

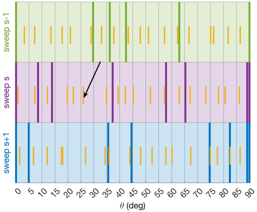


Simultaneous construction & descending of bisection fronts

* parallel stochastic searches in *sweeps*
* at each sweep s:

$$\Omega_{ij}^{[s]} = \text{LEIDEN}\left(\phi_i^{[s]}, r_{ij}^{[s]}, \Omega_{ij}^{[s,o]}\right),\ 1 \le i \le |\Phi|,\ 1 \le j \le t$$

  - Φ: randomized finite partition of θ interval
  - t: number of stochastic search trails for each θ-search
  - r: random seed
* bisection front at the end of sweep s: $\mathcal{B}^{[s]} = \left\{\text{HAR}\left(\Omega_{ij}^{[s]}\right)\right\}$



[1] traag2019 [2] floros2022

Experiments: setup

Clustering function & algorithms

- Function: stochastic modularity Q_{s}, 2024
- Algorithm: Descending Triangulation (DT-II), 2021

  internally deploys LEIDEN,**traag2019**

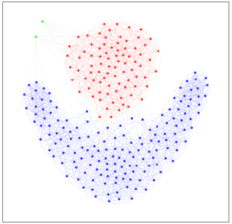
  unsupervised attention to robust configurations, 2024

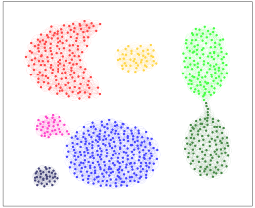
  (robustness indices: persistence λ, steadiness μ)
- Spatial embedding by SG-t-SNE, **pitsianis2019**

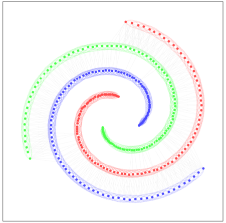
Diverse graph types

- real-world and synthetic
- data networks by direct observation
- knn graphs derived from feature point data
- topology-determinate graphs
- graphs with different degree distributions
- directed, undirected, weighted, unweighted

Experiment: random geometric graphs (2D RGG)





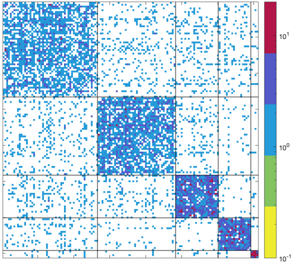


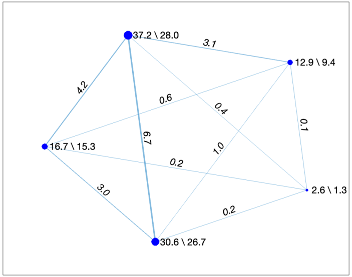
Dataset	$\|V\|$	k	$\|\Omega_*\|$	ARI	Γ_*	λ_*	μ_*
FLAME	240	12	3	**0.96**	(0.01, 0.08)	1.0	0.86
AGGREGATION	788	12	7	**0.99**	(0.03, 0.17)	1.0	0.88
SPIRALS	312	12	3	**1.00**	(0.0001, 0.08)	1.0	0.90

- data for benchmarking point-clustering methods
- different in topology, geometry, sample distribution

- ▷ knn graphs: richer tests for graph clustering
- ▷ **BlueRed** achieved high consistency scores

Experiment: URV-EMAIL-2003





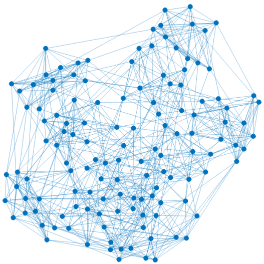
- Email network at Univ. Rovira i Virgili (URV), Tarragona, Spain, 2003
- 1133 users, 5451 email exchange links
- big difference in community size, intra-link density

▷ Graph minor by cluster contraction

- weighted node: 2-tuple (% population, % links)
- weighted link: % links

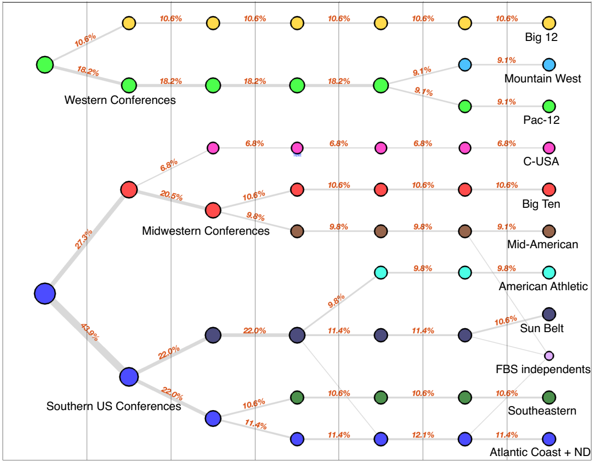
▷ A small dense group is detected besides much larger communities

FBS-2023: multi-level league structure



FBS-2023: 132 college football teams (nodes) of Football Bowl Subdivision (FBS) play 738 games (edges) in the regular season of 2023

BlueRed detects the multi-resolution structure and uncovers all the FBS Conferences with $\text{ari} = \mathbf{0.98}$



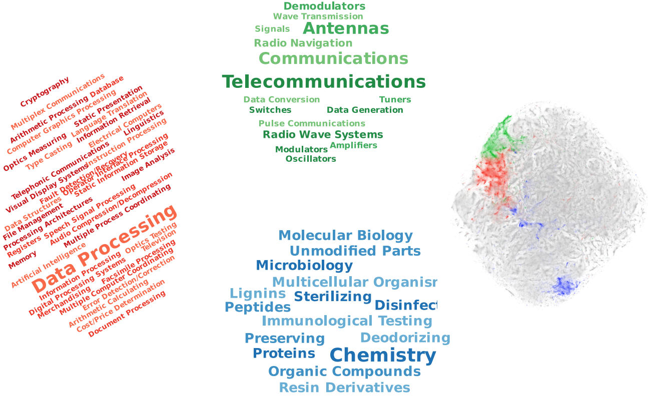
League configurations of FBS-2023 on multiple γ-resolution bands shown in the lineage pyramid. There are 3 non-tree links.

Algorithmic completion of semantic attribute annotation

USPTO PATENTS 1963-1999:

- Citation graph: V: 3.7 million US patents, E: 16.5 million citations
- USPTO codes (in 3-digits) categorize the patents with semantic attributes (manual)
- 1.1 million patents without USPTO class codes

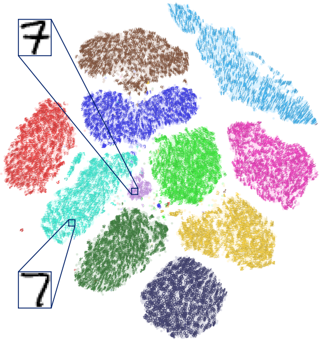
▷ Completion of semantic attribute annotation by community detection combined with language processing



Three of the differentiated 54 classes (RGB-color-coded):
Cluster in red: signal processing in digital devices and formats
Cluster in green: telecommunication techniques
Cluster in blue: chemical and biological techniques

Unsupervised recognition of handwritten digit images

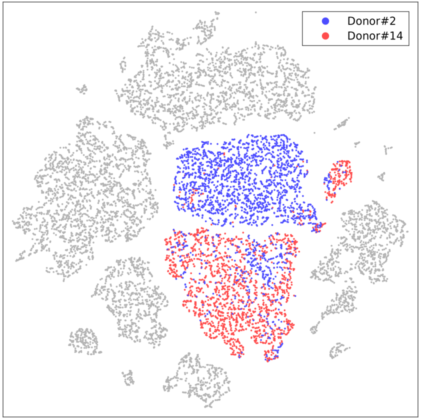
Estimated class \ True class	0	1	2	3	4	5	6	7	8	9	-	
0	**6882** 9.8%	**1** 0.0%	**8** 0.0%	**5** 0.0%	**0** 0.0%	**5** 0.0%	**9** 0.0%	**1** 0.0%	**20** 0.0%	**12** 0.0%	**0** 0.0%	99.1% 0.9%
1	**8** 0.0%	**7659** 10.9%	**3** 0.0%	**0** 0.0%	**7** 0.0%	**0** 0.0%	**8** 0.0%	**9** 0.0%	**25** 0.0%	**7** 0.0%	**0** 0.0%	99.1% 0.9%
2	**0** 0.0%	**162** 0.2%	**6913** 9.9%	**16** 0.0%	**0** 0.0%	**0** 0.0%	**3** 0.0%	**31** 0.0%	**1** 0.0%	**1** 0.0%	**0** 0.0%	97.0% 3.0%
3	**2** 0.0%	**0** 0.0%	**18** 0.0%	**7040** 10.1%	**0** 0.0%	**37** 0.1%	**1** 0.0%	**1** 0.0%	**15** 0.0%	**145** 0.2%	**0** 0.0%	97.0% 3.0%
4	**0** 0.0%	**13** 0.0%	**6** 0.0%	**0** 0.0%	**6683** 9.5%	**3** 0.0%	**4** 0.0%	**10** 0.0%	**5** 0.0%	**23** 0.0%	**0** 0.0%	99.1% 0.9%
5	**1** 0.0%	**2** 0.0%	**0** 0.0%	**14** 0.0%	**0** 0.0%	**6183** 8.8%	**6** 0.0%	**0** 0.0%	**5** 0.0%	**1** 0.0%	**0** 0.0%	99.5% 0.5%
6	**8** 0.0%	**4** 0.0%	**1** 0.0%	**0** 0.0%	**12** 0.0%	**24** 0.0%	**6834** 9.8%	**0** 0.0%	**17** 0.0%	**2** 0.0%	**0** 0.0%	99.0% 1.0%
7	**0** 0.0%	**21** 0.0%	**22** 0.0%	**18** 0.0%	**13** 0.0%	**1** 0.0%	**0** 0.0%	**7146** 10.2%	**4** 0.0%	**30** 0.0%	**0** 0.0%	98.5% 1.5%
8	**1** 0.0%	**6** 0.0%	**15** 0.0%	**38** 0.1%	**6** 0.0%	**36** 0.1%	**11** 0.0%	**4** 0.0%	**6697** 9.6%	**17** 0.0%	**0** 0.0%	98.0% 2.0%
9	**1** 0.0%	**6** 0.0%	**4** 0.0%	**10** 0.0%	**96** 0.1%	**9** 0.0%	**0** 0.0%	**80** 0.1%	**32** 0.0%	**6719** 9.6%	**0** 0.0%	96.6% 3.4%
-	**0** 0.0%	**3** 0.0%	**0** 0.0%	**0** 0.0%	**7** 0.0%	**15** 0.0%	**0** 0.0%	**11** 0.0%	**4** 0.0%	**1** 0.0%	**0** 0.0%	0.0% 100%
	99.7% 0.3%	97.2% 2.8%	98.9% 1.1%	98.6% 1.4%	97.9% 2.1%	97.9% 2.1%	99.4% 0.6%	98.0% 2.0%	98.1% 1.9%	96.6% 3.4%	NaN% NaN%	**98.2% 1.8%**



- The confusion matrix against the truth labels
  total accuracy $\mathbf{98.2\%}$
  recall scores in the last row
  precision scores in the last column
- ARI score $\mathbf{0.96}$

- 70.000 MNIST Digit-images in 11 clusters
- two distinct subclusters of digit-7 images (90% **cyan**, 10% **purple** $\approx 1\%$ total)
- with zero training, with higher differentiation

Unsupervised labeling of vasculature cells

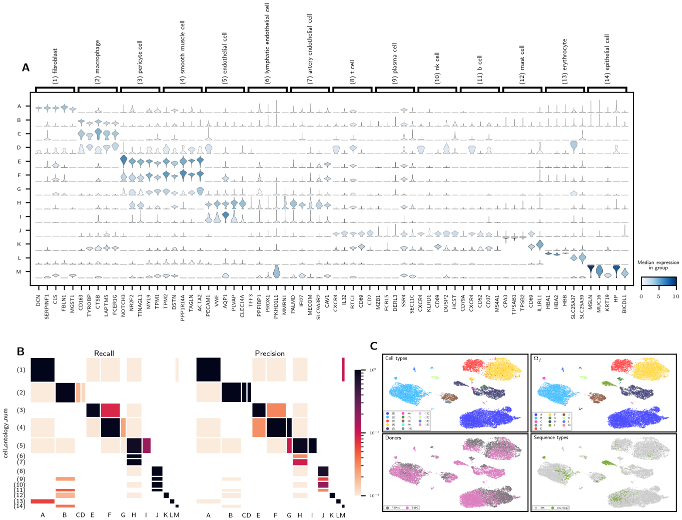


Two fibroblast cell clusters in Vasculature of the Tabula Sapiens, 2002

- 16 035 cells in the vasculature organ
  from the Tabula Sapiens, Science 2022
- cell type annotation assisted by tissue experts
- knn ($k \approx 15$) provided by a gene expression pipeline at the atlas data site

- higher consistency score: $\mathbf{0.95}$ in ARI against the ground-truth cell labels; all 16 035 cells are labeled without learning or training
- higher differentiating capability:
  two distinctive subclusters in fibroblast cells
  reflecting two different sequencing techniques

Unsupervised labeling of vasculature cells analysis

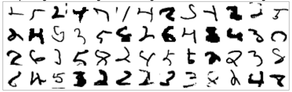


Extension in theory and application

More to invest in theory:

- clustering function bank: interpretable, composible, evaluated, supporting diverse tasks
- interplay between transformation of graph G and change of clustering function h
  f.g. $Q(G)$ vs. $h(G) = Q(f(G))$
- evolution: emergence, growth, decline, disappearance of sub-communities
- divide-and-conquer graph algorithms
- random models for graph data augmentation
- $\cdots$

Top-52 digits with 0 in-degree and the largest distance to their nearest neighbor



Abundant opportunities in applications:

- data $\rightarrow$ information $\rightarrow$ insight
- identification of influential players, outliers, strong, weak links, pattern features
- context-specific transformation of data graphs
- recognition, prediction of propagation patterns
- feature-guided graph partition or alignment
  `f.g.` super-pixel segmentation
- graph compression, decompression for query
- $\cdots$

Top-52 digits with the largest in-degree values

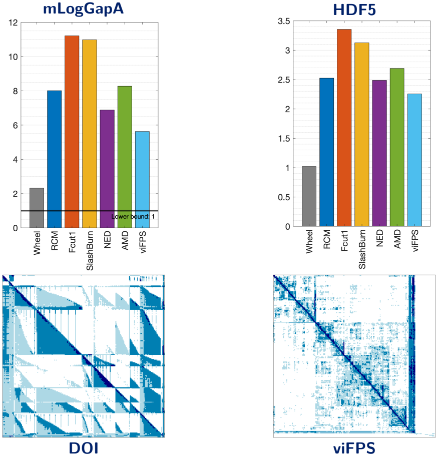
Graph compression $\rightleftharpoons$ Vertex ordering

To expedite responses to frequent, scattered queries

- accommodate large (portions) of knowledge networks in local memories
- use narrow (de)compression windows
- be compatible with storage formats (e.g. HDF5)

- ☐ exploit & encode similarities in neighborhood subgraphs (block cols/rows of adjacency matrix)
- ☐ π: order vertices to minimize the total variation in neighborhood subgraph patterns (NP-hard)

$$\text{mLogGapA}(G,\pi) = \frac{|V|}{|E|} + \sum_{\substack{u_i \in \mathcal{N}[v] \\ i=2:d(v)}} \log_2(1+\pi(u_i)-\pi(u_{i-1}))$$



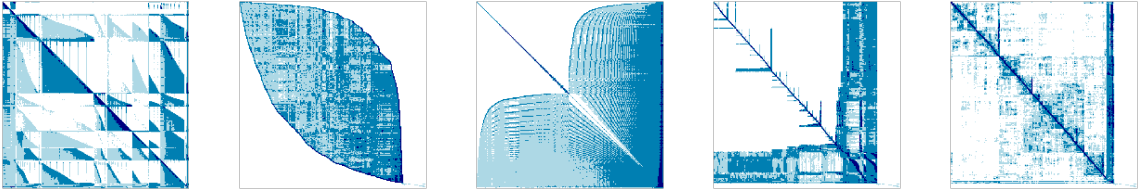
Algebraic graph compression

Connections & constraints

- ◦ Vertex ordering $\rightleftharpoons$ Graph compression
- ◦ Fast query responses: large capacity & narrow (de)compression window

- ▷ Theoretical analysis
- ▷ Highly efficient algorithm **viFPS**:
  - ◦ Higher compression ratios over diverse graphs
  - ◦ Faster sparse matrix-vector products

Citation graph APS2020: in 5 different vertex orderings (667 K articles & 8.850 M citation links)



DOI RCM SlashBurn AMD **viFPS**

cuthill1969reducingbandwidth; **george1981computersolution**; **amestoy2004algorithm837**; **amestoy1996approximateminimum**; **lim2014slashburngraph**

Superpixel segmentation

For embedded vision systems

- ▷ scene parsing
- ▷ image matching & registration
- ▷ motion tracking or estimation
- ▷ autonomous navigation
- ▷ saliency recognition
- ▷ feature and filter learning
- ▷ time-critical in many circumstances

Superpixel[1]

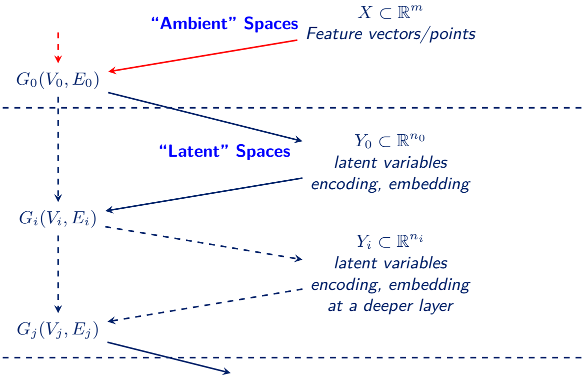
- ◇ a set of pixels spatially connected and chromatically homogeneous
- ◇ regulation on size & number (k)
- ◇ a semantic element encoded by
  [size, shape, shade]

Airplane [10081][2]

Church of Saint Anna, Mykonos [118035][2]

[1] Ren and Malik 2003 [2] BSDS-500 (**martin2001**)

Graph Analysis: to advance deeper and broader



Objectives: $\begin{cases} \text{Differentiation} \\ \text{Identification} \\ \text{Recognition} \end{cases}$ of $\begin{cases} \text{Patterns} \\ \text{Communities} \\ \text{Functional units} \end{cases}$

- ubiquitous in data analysis
- understanding
- underlying, unifying

References