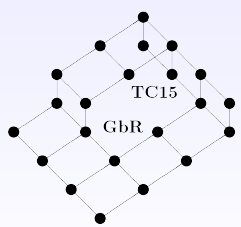
# Graph descriptors from B-matrix representation

Wojciech Czech

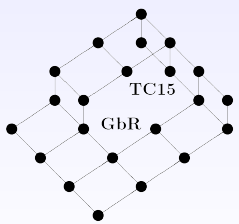
Institute of Computer Science
AGH University of Science and Technology, Kraków

May 17, 2011



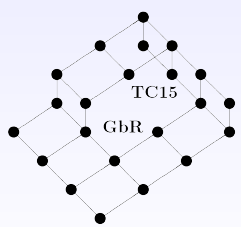
# Agenda

- Introduction
- Vertex B-matrix
- Edge B-matrix
- Graph descriptors
- Experiments
- Summary



# Motivation

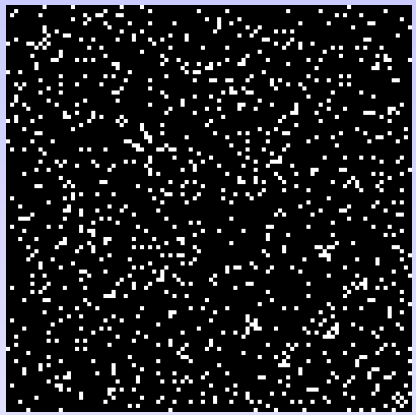
- Explore shortest-paths-based representations of a graph
- Investigate usability of B-matrices for graph feature vectors generation
- Test B-matrices-based pattern vectors using clustering and classification algorithms



# Graph representations

- Adjacency matrix

$$\mathbf{A}_{u,v} = \begin{cases} 1 & \text{if } (u,v) \in E \\ 0 & \text{if } (u,v) \notin E. \end{cases}$$



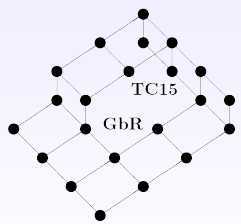
- Laplace matrix

$$\mathbf{L} = \mathbf{D} - \mathbf{A},$$

$\mathbf{D}$ is diagonal matrix of vertex degrees.

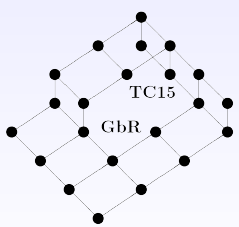
- Incidence matrix
- Neighbourhood lists

Vertex ordering needed



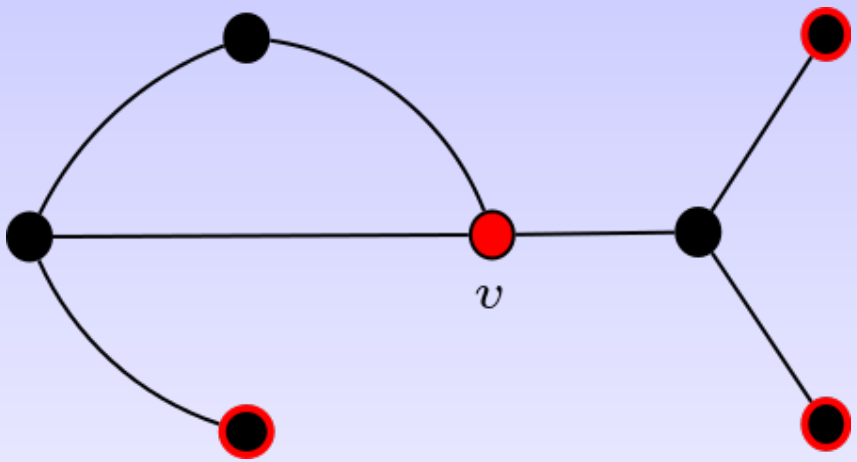
# Graph characteristics and isomorphism invariance

- Permutation invariant functions
- Aggregating histogram bins
- Using reference graph and inter-vertex correspondences

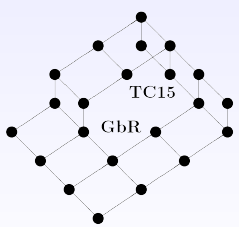


# Vertex $l$-shell

$l$-shell of vertex $v$ is a set of graph vertices at distance $l$ from vertex $v$



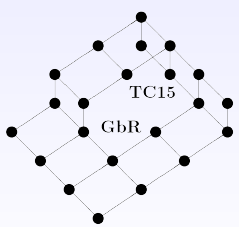

2-shell of vertex $v$



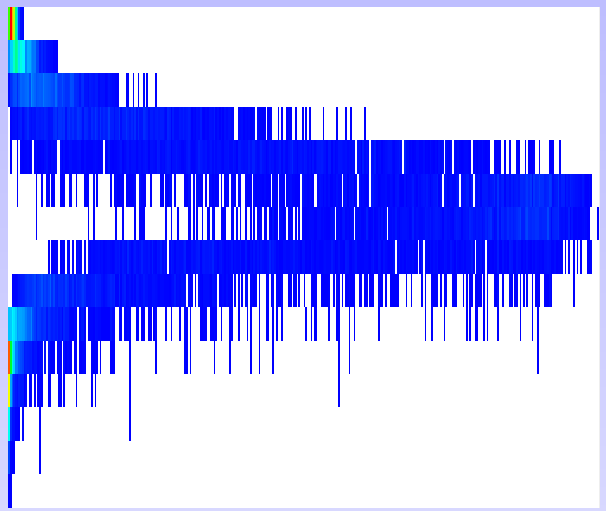
# Vertex B-matrix

$B^V_{l,k}$ = number of nodes that have $k$ members in their $l$-shells

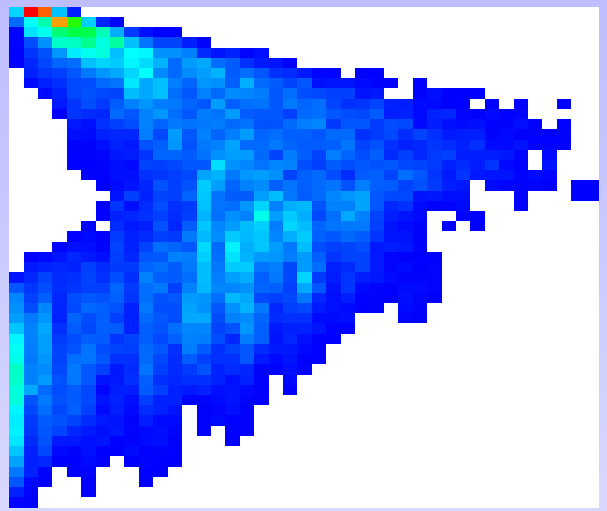
- $l \leq n$, $k \leq n$, where $n$ is number of graph vertices
- Sequence of histograms, $l$-level degree distributions
- Graph diameter determines number of non-empty rows
- *Breadth-First Search* used to enumerate shell members gives $\mathcal{O}(n^2)$ time-complexity for sparse graphs



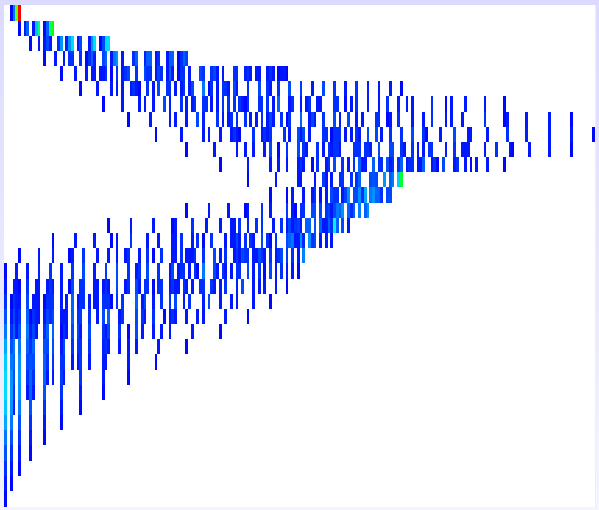
# Vertex B-matrices examples



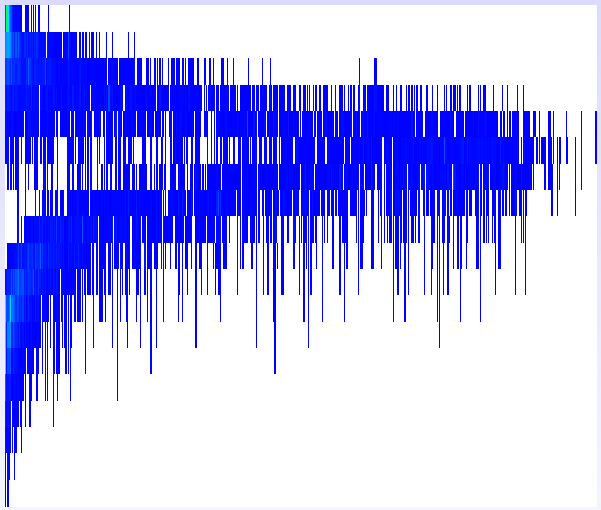
random graph (1028, 1582)



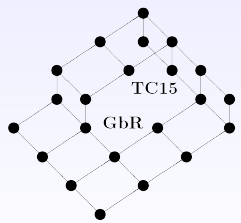
tumor vascular network (497, 622)



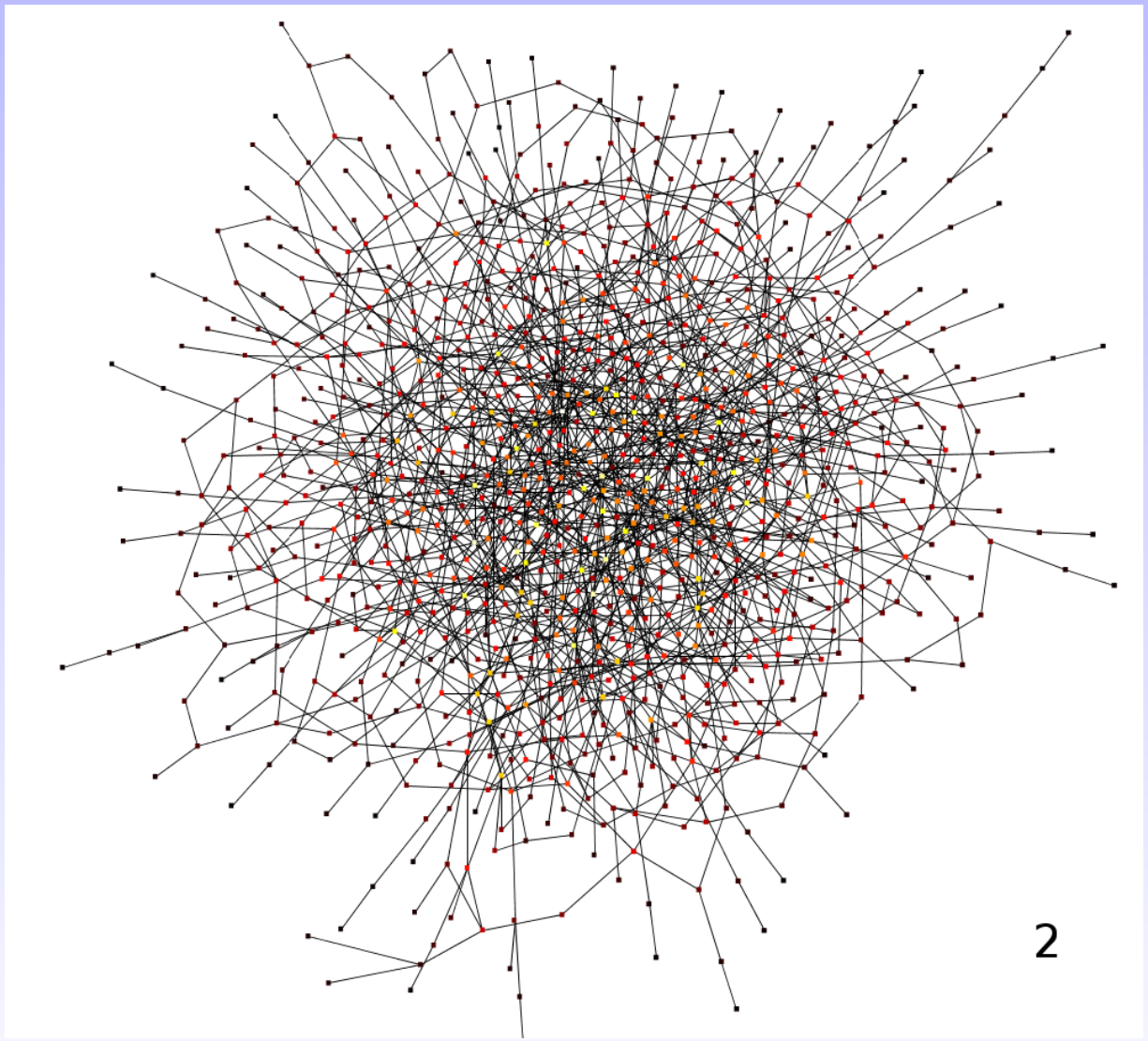
3D lattice (1728, 4752)



yeast PPI network (1870, 4406)

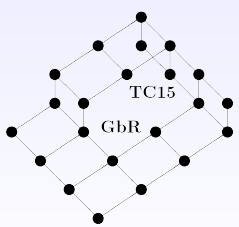


# Visualization of $l$-shell sizes (*random graph*)

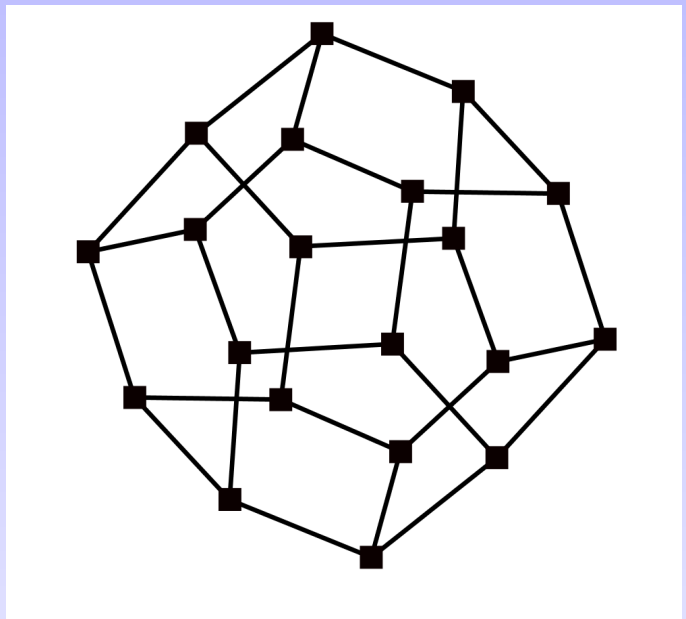


# Vertex B-matrix and graph structure

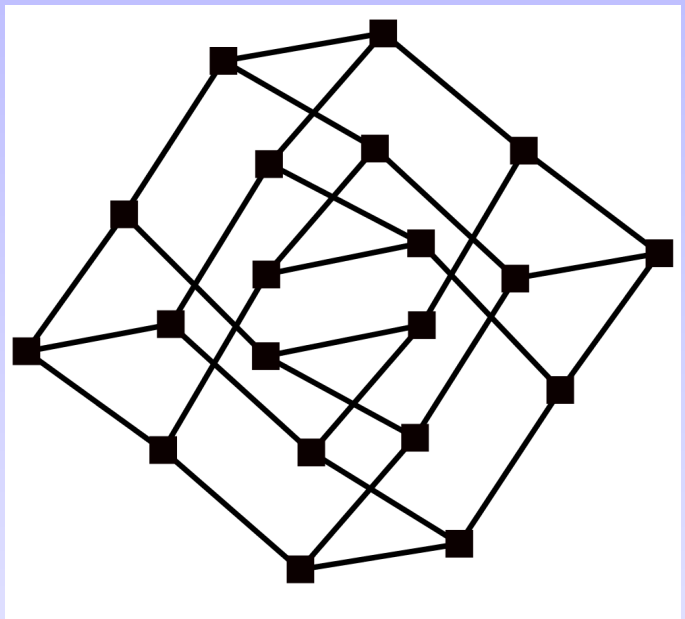
- *small-worldliness*
- *dimensionality* and *regularity*
- *assortativity* and *disassortativity*
- visualization of percolation



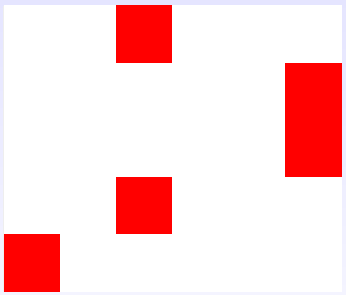
# Incompleteness of vertex B-matrix

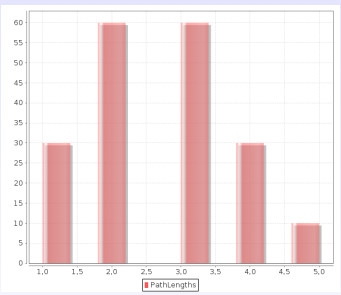


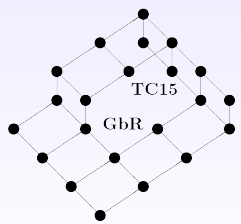
dodecahedral graph



Desargues graph

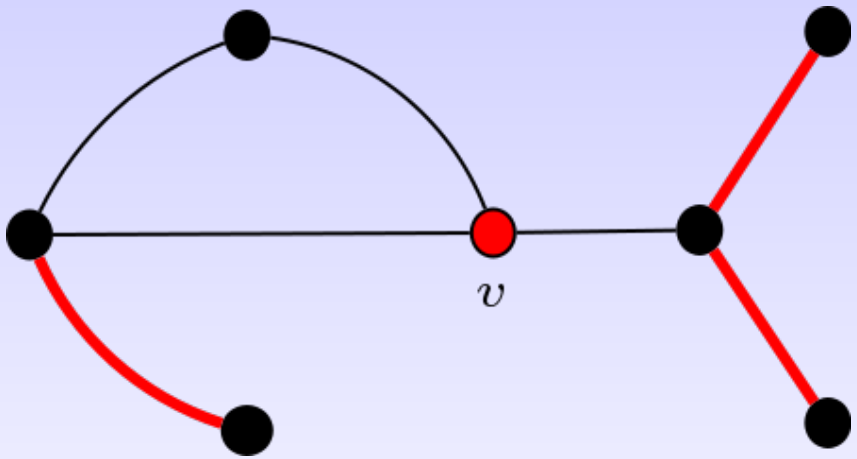




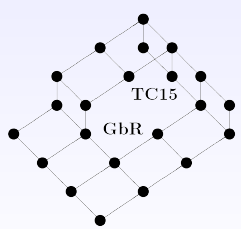



# Edge $l$-shell

The distance from vertex $v$ to an edge $(u, w)$ is the mean of distances $d(v, u)$ and $d(v, w)$. The $l$-edge-shell of vertex $v$ is a set of graph edges at distance $l$ from $v$ ($l$ can have half-integer values).



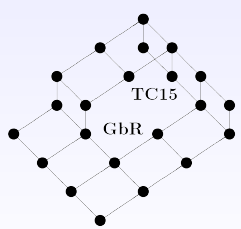

1.5-edge-shell of vertex $v$



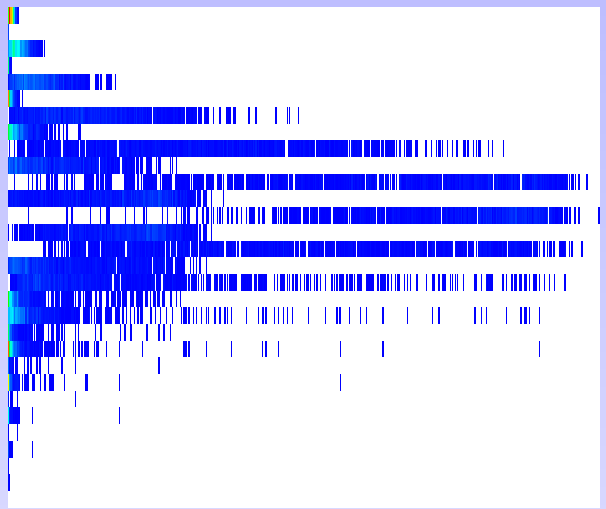
# Edge B-matrix

$B^E_{i,k}$ = number of nodes that have $k$ edges in their $(\frac{1}{2}i)$-edge-shells.

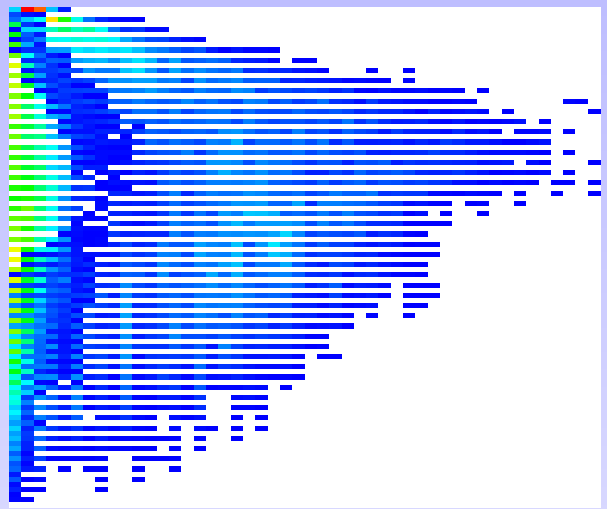
- Size $2n \times m$, $n$ - number of vertices, $m$ - number of edges
- Bipartite graphs (no odd cycles) have empty even rows
- For dense graphs $m$ is of order $\mathcal{O}(n^2)$



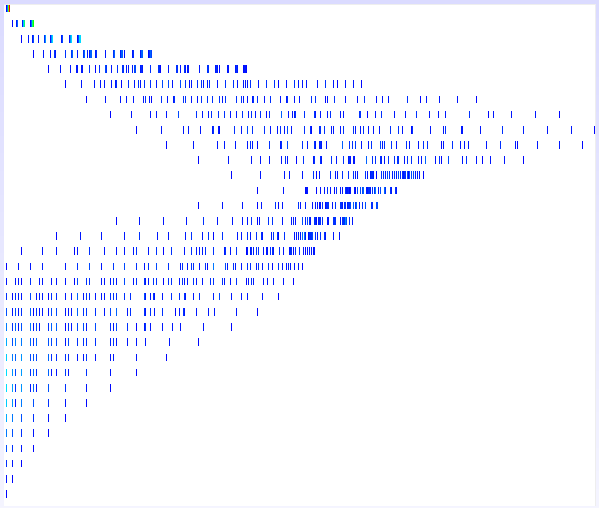
# Edge B-matrices examples



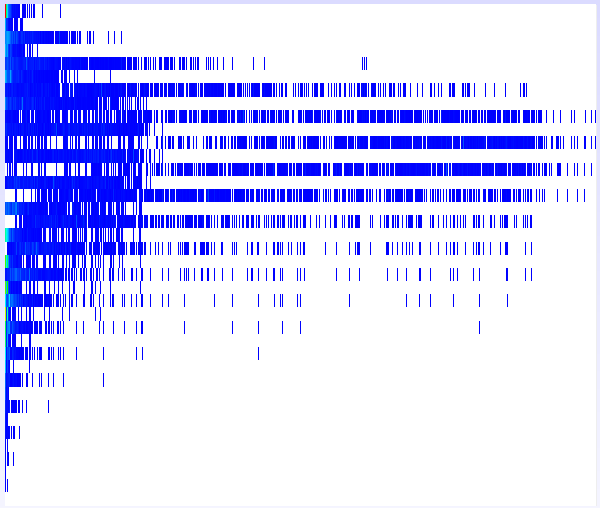
random graph (1028, 1582)



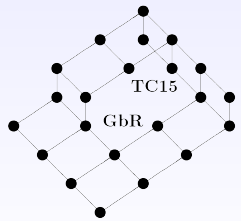
tumor vascular network (497, 622)



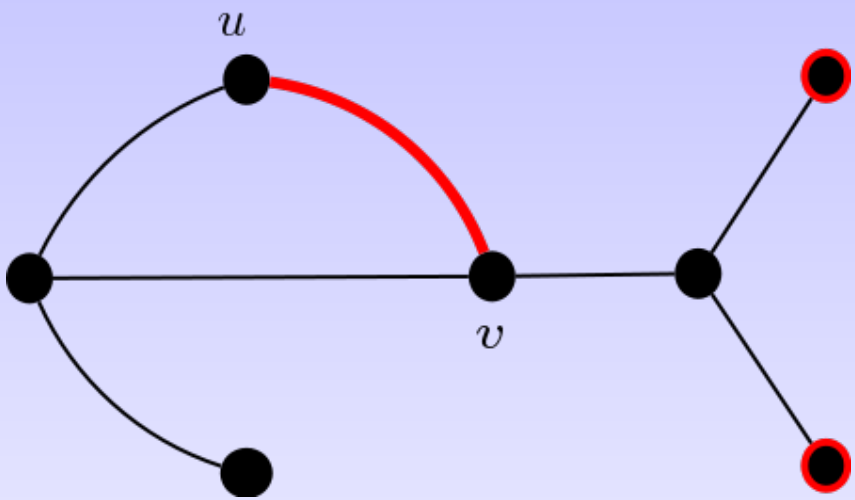
3D lattice (1728, 4752)



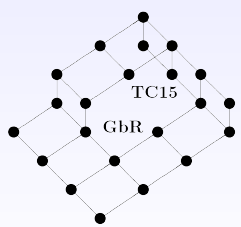
yeast PPI network (1870, 4406)



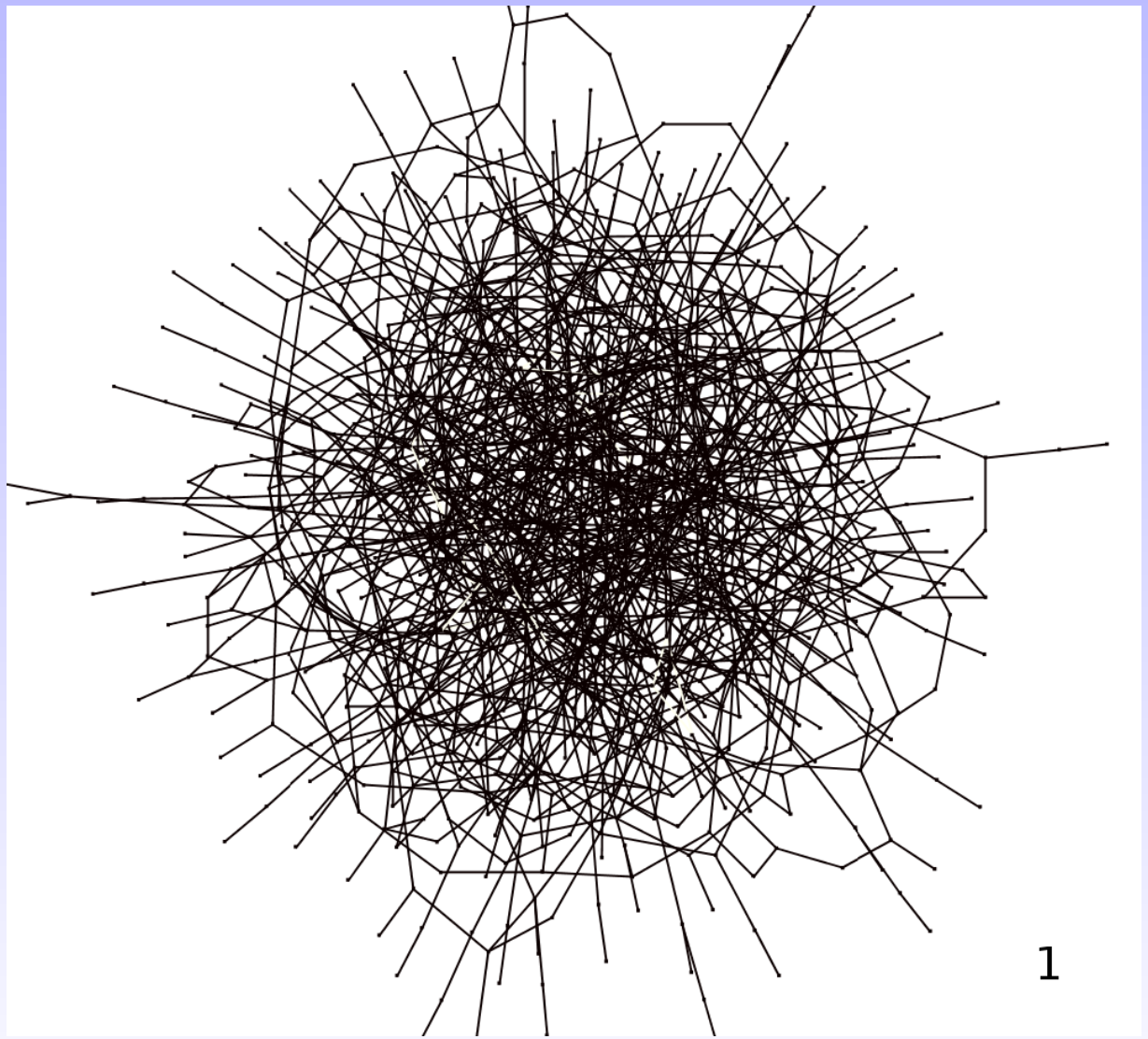
# Edge $l$-shell - alternative definition

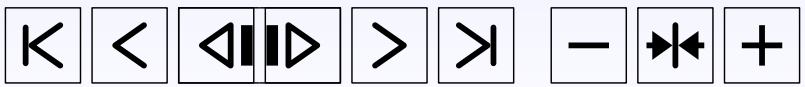


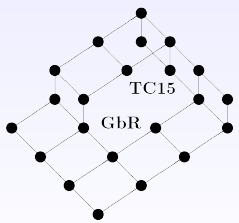

2.5-shell of edge $(u, v)$



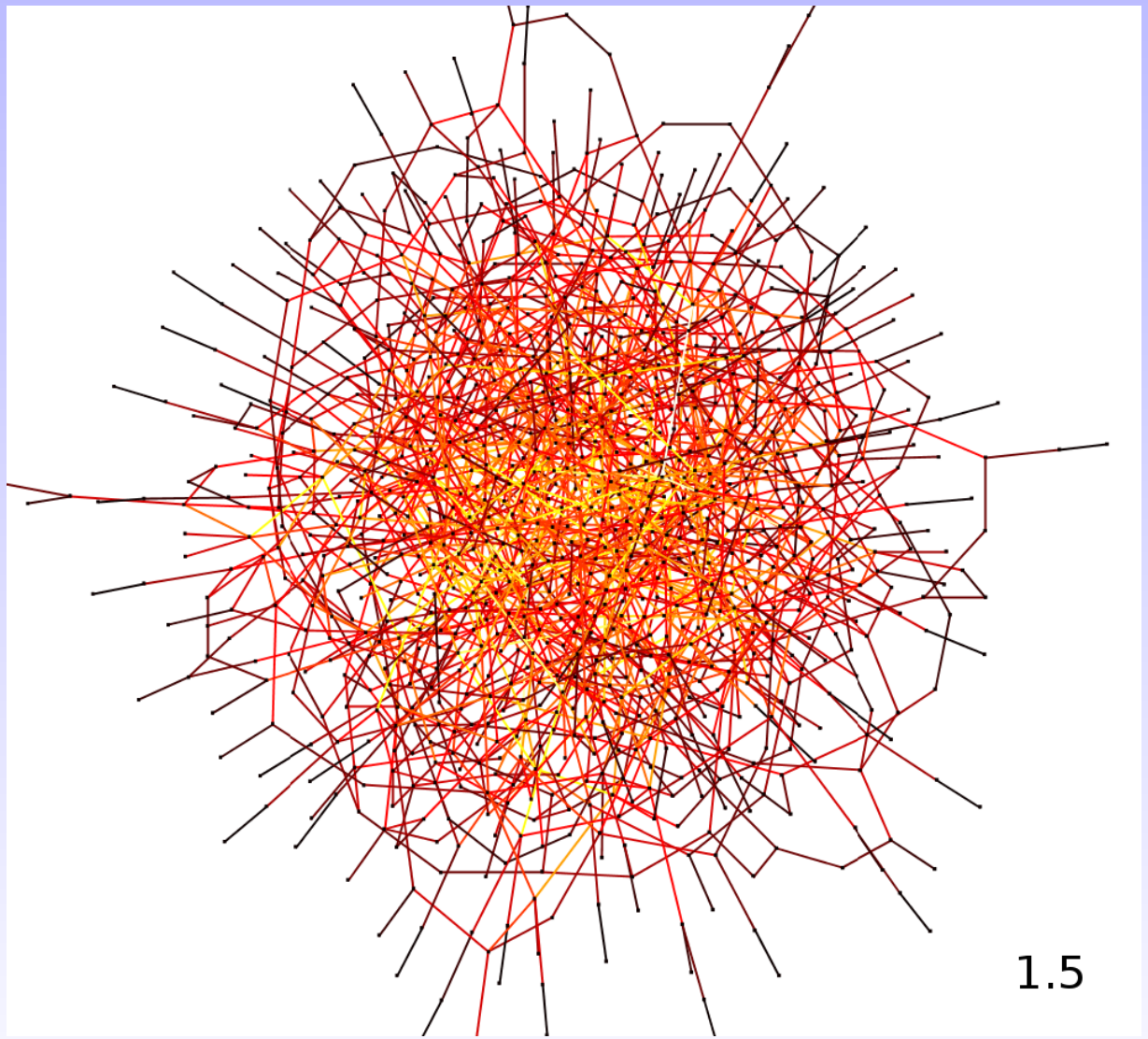
# Visualization of $l$-edge-shell sizes - integer $l$'s

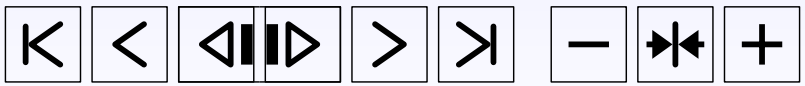


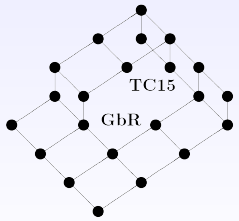





# Visualization of $l$-edge-shell sizes - fractional $l$'s

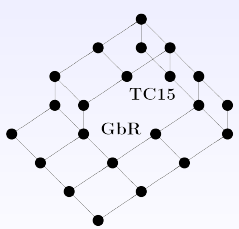




# Feature vectors based on B-matrices

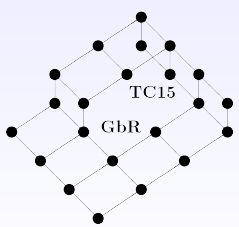
- Long vectors constructed on the basis of packed rows (columns)
- Aggregated statistics computed per row
- Assessing inter-row diversity



# Packed rows

$$D^{\star}_{long}(l_{min}, l_{max}, k_{min}, k_{max}) = [B^{\star}_{l,k}] \\ l_{min} \leq l \leq l_{max},\ k_{min} \leq k \leq k_{max} \tag{1}$$

- All information present in B-matrix can be retained
- Navigation between local and global features using $l$



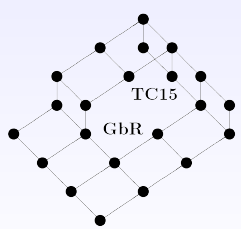
# Row statistics

$$D^{\star}_{rstd}(l) = \frac{\sigma^{\star}(l)}{\mu^{\star}(l)} \tag{2}$$

$$p(l,k) = \frac{B^{\star}_{l,k}}{\sum_k B^{\star}_{l,k}} \tag{3}$$

$$D^{\star}_{ent}(l) = -\sum_k p(l,k)\log(p(l,k)) \tag{4}$$

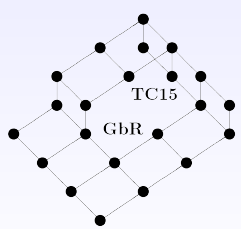
- Relative deviation captures $l$-shell regularity
- Shannon entropy measures unpredictability of $l$-shell size
- Only non-zero entries taken into account



# Difference of row statistics

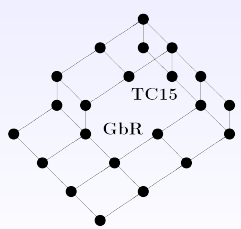
$$D^{\star}_{avgd}(l) = \mu^{\star}(l-1) - \mu^{\star}(l) \tag{5}$$

- Average offset between consecutive histograms
- Indicates branching and density of a graph



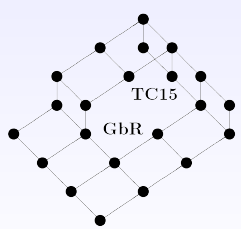
# Feature vector embedding

- Artificial dataset: 4 groups of graphs of similar density (100 vertices, 100 instances in a group)
- Comparison with heat-kernel-based descriptors
- Embedding into 2D or 3D space using PCA or LPMIP (Locality-Preserved Maximum Information Projection)
- Results evaluated using clustering validation indices: C index, Davies-Bouldin index and Rand index



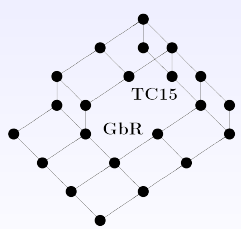
# Unsupervised learning results

| | Vector | Dim | Dim Red Method | 2D | 3D |
|---|---|---|---|---|---|
| 1 | $D^E_{long}(1,4,1,25)$ | 100 | PCA | 0.07, 1.00, 0.93 | 0.05, 1.13, 0.96 |
| 2 | $D^E_{long}(1,4,1,25)$ | 100 | LPMIP(20, 20, 0.1) | **0.04, 0.69, 0.98** | **0.03, 0.74, 0.99** |
| 4 | $D^V_{long}(1,4,1,20)$ | 80 | LPMIP(15, 20, 0.1) | 0.08, 2.11, 0.93 | 0.08, 2.13, 0.93 |
| 6 | $D^E_{avgd}$, $1 \leq l \leq 20$ | 20 | LPMIP(20, 20, 0.1) | 0.12, 1.67, 0.81 | 0.15, 1.96, 0.81 |
| 7 | $\mu^E, \mu^V, \sigma^E, \sigma^V$, $1 \leq l \leq 4$ | 16 | PCA | **0.07, 1.16, 0.97** | **0.11, 1.50, 0.97** |
| 8 | $D^V_{rstd}$, $1 \leq l \leq 10$ | 10 | LPMIP(20, 20, 0.1) | 0.09, 1.2, 0.86 | 0.15, 1.69, 0.87 |
| 9 | $D^E_{rstd}$, $1 \leq l \leq 20$ | 20 | LPMIP(20, 20, 0.1) | 0.09, 1.37, 0.81 | 0.16, 1.92, 0.84 |
| 10 | vectors from row 2 and 7 together | 116 | LPMIP(20, 20, 0.1) | **0.01, 0.47, 1.0** | **0.04, 0.54, 1.0** |
| 11 | $D_{hkc}$, $1 \leq t \leq 10$ | 10 | PCA | 0.05, 0.73, 0.89 | 0.05, 0.73, 0.98 |
| 12 | $D_{hkc}$, $1 \leq t \leq 20$ | 20 | PCA | 0.05, 0.86, 0.86 | 0.05, 0.86, 0.97 |
| 13 | $D_{hkcc}$, $1 \leq m \leq 10$ | 10 | PCA | 0.07, 1.46, 0.79 | 0.07, 1.47, 0.78 |
| 14 | $D_{hkcc}$, $1 \leq m \leq 20$ | 20 | PCA | 0.09, 1.16, 0.80 | 0.09, 1.17, 0.80 |
| 15 | $D_{con}$ | 7 | PCA | 0.11, 2.01, 0.79 | 0.11, 2.14, 0.86 |

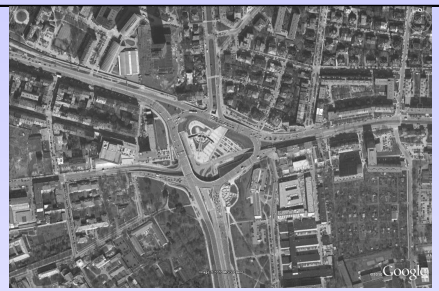


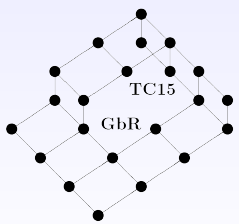
# Experiments on classification

- Satellite photos from *Google Earth* tansformed into graphs using corner detection (100 corners for each photo) and Delaunay triangulation
- Three groups of photos, 90 instances in each group
- Random selection of training and testing samples (75, 15)
- PCA, MMC (Maximum Margin Criterion) and LDA for dimensionality reduction
- Nearest centroid classifier
- Target dimensionality selected experimentally
- Average and maximum accuracy reported



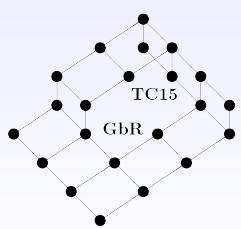
# Samples from dataset

| | | | |
|---|---|---|---|
| Photo |  |  |  |
| Instances | 90 | 90 | 90 |



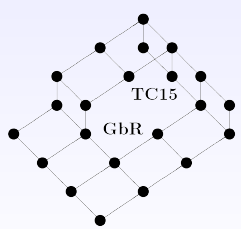
# Classification results

| Vector | Dim | Dim Red | Target Dim | Avg Accuracy | Max Accuracy |
|---|---|---|---|---|---|
| $D^E_{avgd}$, $1 \leq l \leq 15$ $D^E_{rstd}$, $1 \leq l \leq 15$ | 30 | PCA | 10 | 0.73 | 0.86 |
| | | PCA | 15 | 0.75 | 0.89 |
| | | LDA | 2 | **0.77** | **0.91** |
| | | MMC | 2 | 0.61 | 0.73 |
| | | MMC | 5 | 0.65 | 0.80 |
| $D^V_{long}(1,8,1,30)$ | 240 | MMC | 2 | **0.76** | **0.91** |
| $D^E_{long}(1,20,1,100)$ | 2000 | MMC | 100 | **0.78** | **0.89** |
| $D_{hkc}$ and $D_{hkcc}$, $1 \leq t \leq 15$ | 30 | PCA | 10 | 0.65 | 0.80 |
| | | PCA | 15 | 0.66 | 0.80 |
| | | LDA | 2 | 0.64 | 0.80 |
| | | MMC | 2 | 0.67 | 0.84 |
| $D_{con}$ | 7 | PCA | 5 | 0.60 | 0.71 |
| | | LDA | 2 | 0.69 | 0.84 |
| | | MMC | 2 | 0.68 | 0.80 |



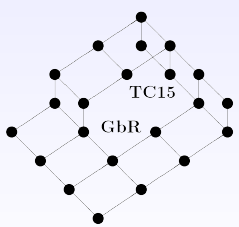
# B-matrices on GPU

- CUDA implementation of BFS
- CUDA implementation of all-shortest-paths using R-Kleene algorithm
- For graph with 8000 vertices computing ASPS takes 3s (Nvidia Tesla C2070)



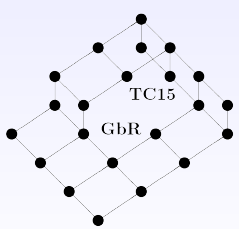
# B-matrices Pros

- Rich source of information about graph structure
- Invariance under graph isomorphism
- Computational efficiency
- Descriptors perform well on sample test cases
- Can be also used as a method of graph visualization



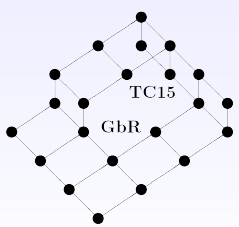
# B-matrices Cons

- Best descriptors are high-dimensional (parametrized dimensionality reduction needed)
- Vertex B-matrix is non-complete graph invariant
- Edge-B-matrix can be infeasible for large edge sets
- Scaling needed to compare graphs of different size



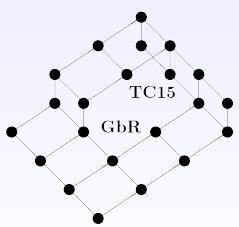
# Future work

- Generation of more elaborate features
- Feature selection for long vectors
- Testing on different real-world datasets such as metabolic networks



# Core references


- J.P. Bagrow, E.M. Bollt, J.D. Skufca,*Portraits of complex networks*, Europhysics Letters, vol. 81, 68004, 2008.
- B. Xiao, E.R. Hancock, R.C. Wilson, *Graph characteristics from the heat kernel trace*, Pattern Recognition, vol. 42, no. 11, 2589–2606, 2009.
- W. Czech, S. Goryczka, T. Arodz, W. Dzwinel, A. Dudek, *Exploring complex networks with Graph Investigator research application*, Computing and Informatics, vol. 30, no. 2, 2011.

Thank you for your attention...