Regular decomposition of large graphs and other structures: scalability and robustness towards missing data

Hannu Reittu (VTT, Finland)

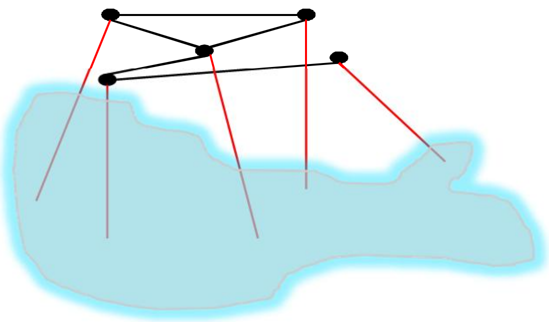
Joint work with

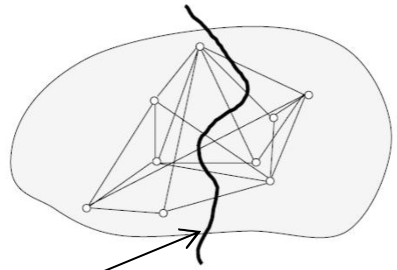
Ilkka Norros (VTT) and

Fülöp Bazsó (Wigner Research Centre, Hungary)

Huge networks are everywhere!

- Infer properties from small samples of large graphs
  - Property testing (Goldreich et al (1998)- Alon (2009)…)
  - Graph parameter testing





- Example; Lovasz: a dense cut in the large graph => dense cut in the sample graph

Noga Alon, Eldar Fischer, Ilan Newman, and Asaf Shapira

2009:

SIAM J. Comput., 39(1), 143–167. (25 pages)

A Combinatorial Characterization of the Testable Graph Properties: It's All About Regularity

DEFINITION 2.5. (REGULAR-REDUCIBLE) *A graph property $\mathcal{P}$ is* regular-reducible *if for any $\delta > 0$ there exists an $r = r(\delta)$ such that for any n there is a family $\mathcal{R}$ of at most r regularity-instances each of complexity at most r, such that the following holds for every n-vertex graph G:*

1. *If G satisfies $\mathcal{P}$ then for some $R \in \mathcal{R}$, G is δ-close to satisfying R.*
2. *If G is ϵ-far from satisfying $\mathcal{P}$, then for any $R \in \mathcal{R}$, G is $(\epsilon - \delta)$-far from satisfying R.*

THEOREM 2. (MAIN RESULT) *A graph property is testable* **if and only if** *it is regular-reducible.*

WIKIPEDIA

Szemerédi regularity lemma (SRL)

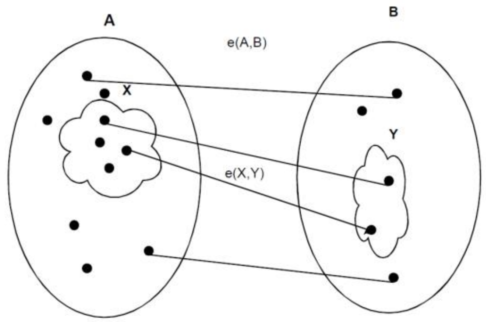
Definition 1. Let X, Y be disjoint subsets of V. The **density** of the pair (X, Y) is defined as:

$$d(X,Y) := \frac{|E(X,Y)|}{|X||Y|}$$

where $E(X, Y)$ denotes the set of edges having one end vertex in X and one in Y.

Definition 2. For $\varepsilon > 0$, a pair of vertex sets X and Y is called **ε-regular**, if for all subsets $A \subseteq X$, $B \subseteq Y$ satisfying $|A| \geq \varepsilon|X|$, $|B| \geq \varepsilon|Y|$, we have

$$|d(X,Y) - d(A,B)| \leq \varepsilon.$$



Definition 3. A partition of V into k sets: $V_1, \ldots, V_k$, is called an **ε-regular partition**, if:

- for all i, j we have: $||V_i| - |V_j|| \leq 1$;
- all except εk^2 of the pairs V_i, V_j, $i < j$, are ε-regular.

Regularity Lemma. For every $\varepsilon > 0$ and positive integer m there exists an integer M such that if G is a graph with at least M vertices, there exists an integer k in the range $m \leq k \leq M$ and an ε-regular partition of the vertex set of G into k sets whose sizes differ by at most 1.

A celebrated result:

Abel Prize Laureate 2012
Endre Szemerédi

Szemerédi`s Regularity lemma

A main ingredient in Szemerédis theorem about arithmetic progressions in sets of positive density is the Regularity lemma. Szemerédi used a weak form of this lemma, for bipartite graphs, to prove the theorem. Later he also proved a strong version, for more general graphs.

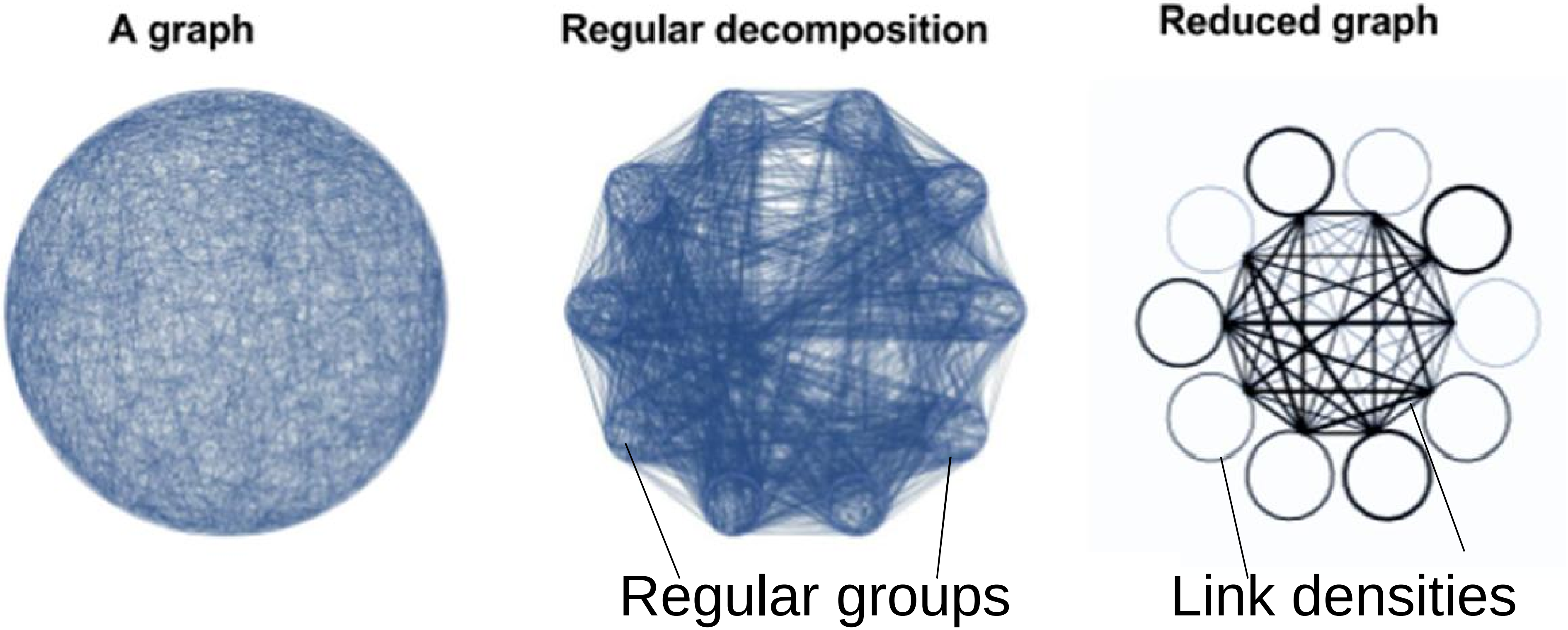
Szemerédi's Regularity Lemma and big data?

- About big graphs (testability, graph limits,…)
- Algorithmic versions: Regular structure can be found efficiently (deterministic: $O(n^2)$ time, randomized: $O(n)$ time)
- Rigorous algorithms have huge constants like: $O\left(k^2 2^{2^{(k/\alpha\epsilon)^{O(1)}}} n^2\right)$, where $k, 1/\epsilon\alpha$ are bounded yet possibly very large numbers
- => impossible to use in practice
- Needs some approximating scheme to find regular structure

Mimic Regularity Lemma in ‘practical’ way:

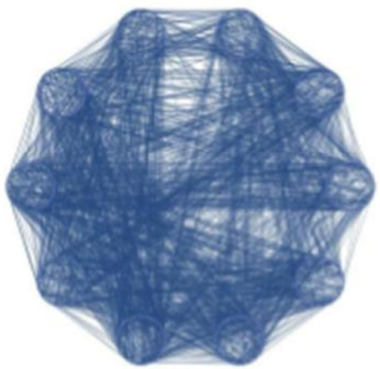
- VTT -> regular decomposition algorithm for ‘Big Data’ and machine learning
- See also:
- Marcello Pelillo, Ismail Elezi, Marco Fiorucci: Revealing Structure in Large Graphs: Szemerédi's Regularity Lemma and its Use in Pattern Recognition, Pattern Recog. Letters, 2017
- Hannu Reittu, Fülöp Bazsó, Ilkka Norros: Regular Decomposition: an information and graph theoretic approach to stochastic block models, ArXiv, 2017

Regular decomposition



Reduced graph

Regular decomposition



$\rightarrow$ $(P)_{i,j}$, $k \times k$, symmetric, elements $0 \leq p_{i,j} \leq 1$, are link densities between – and inside regular groups

Partition ξ of nodes into k regular groups

Minimum description length principle (MDL) for finding regular decomposition:

- Coding length of a graph given a regular decomposition:

(1) $L_k(G|P) := -\log P(G|P) = \sum_{1 \le i \le j \le k} n_{i,j} h(p_{i,j})$,

- $h(p) := -p \log p - (1-p)\log(1-p), 0 \le p \le 1$, $n_{i,j}$ is # node pairs inside ($i = j$) and between ($i \neq j$) groups
- Coding length of a partition ξ

(2) $L_k(\xi = \{V_1, V_2, \dots, V_k\}) = -\sum_{1 \le i \le k} n_i \log r_i$

- r_i is relative size of set V_i in the partition and $n_i = |V_i|$

(3) $L_k(P) = \sum_{1 \le i \le k} \log(e_{i,j})$, $e_{i,j}$, number of links between groups or inside groups.

- Regular decomposition (MDL) min((1)+(2)+(3))

$$(V_1, V_2, \dots, V_{k^*}) = \underset{k}{\text{argmin}} \underset{\xi=\{V_1,V_2,\dots,V_k\}}{\text{argmin}} (L_k(G|P) + L_k(\xi) + L_k(P))$$

Greedy regular decomposition algorithm

- For a given k make a random k-partition ξ_0,
- Compute link densities and get link density matrix P_0
- Apply mapping $P_{i+1} = \Phi(P_i)$, i = 0, 1, ... , until fixed point $P_{i+1} = P_i = P^*$ is reached on corresponding partition ξ^*
- Find coding length of the graph corresponding to $\xi^*, L(\xi^*)$
- Repeat above procedure several times and find the partition that correspond to $\min L(\xi^*)$ over all repetition
- Search above optimization in a range of k,
- Result an approximate MDL optimal regular decomposition

Other related works

- Spectral approach to regular decomposition:

Bolla, M.: Spectral clustering and biclustering, Wiley, 2013

- Stochastic block modeling and MDL, see e.g.

Peixoto, T.P.:   Parsimonious Model Inference in Large Networks, Phys. Rev. Lett. 110, 2013

- Algorithmic version of reg. lemma

A Sperotto, M Pelillo: Szemerédi's regularity lemma and its applications to pairwise clustering and segmentation, in proc. Energy minimization methods in computer vision and pattern recognition, 13-27, 2007

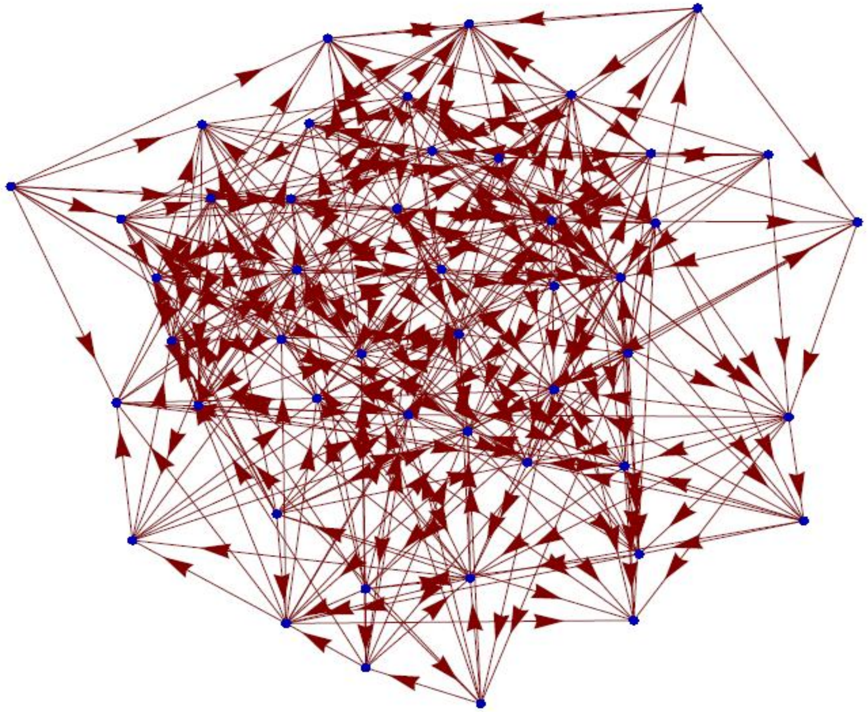
Gábor N. Sárközy, Fei Song, Endre Szemerédi, Shubhendu Trivedi:

A Practical Regularity Partitioning Algorithm and its Applications in Clustering, Arxiv

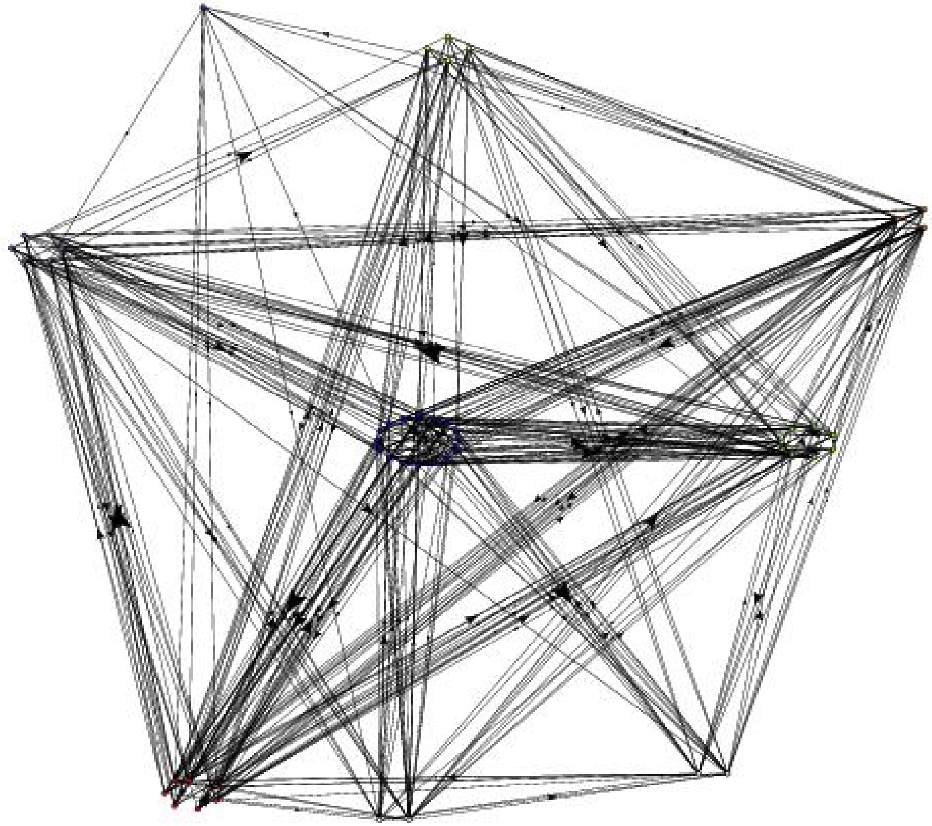
- Testability, graph limits, regularity, see e.g.

- L. Lovász and B. Szegedy: Szemerédi's Lemma for the analyst, *J. Geom. and Func. Anal.* **17** (2007), 252-270

A directed weighted graph:

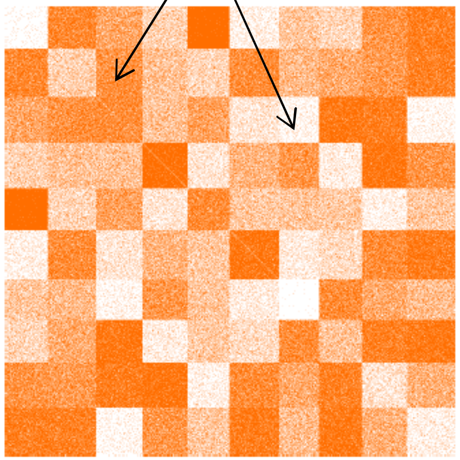


=>



In regular decomposition the mapping $\Phi(\cdot)$ involves matrix multiplication of adjacency matrix

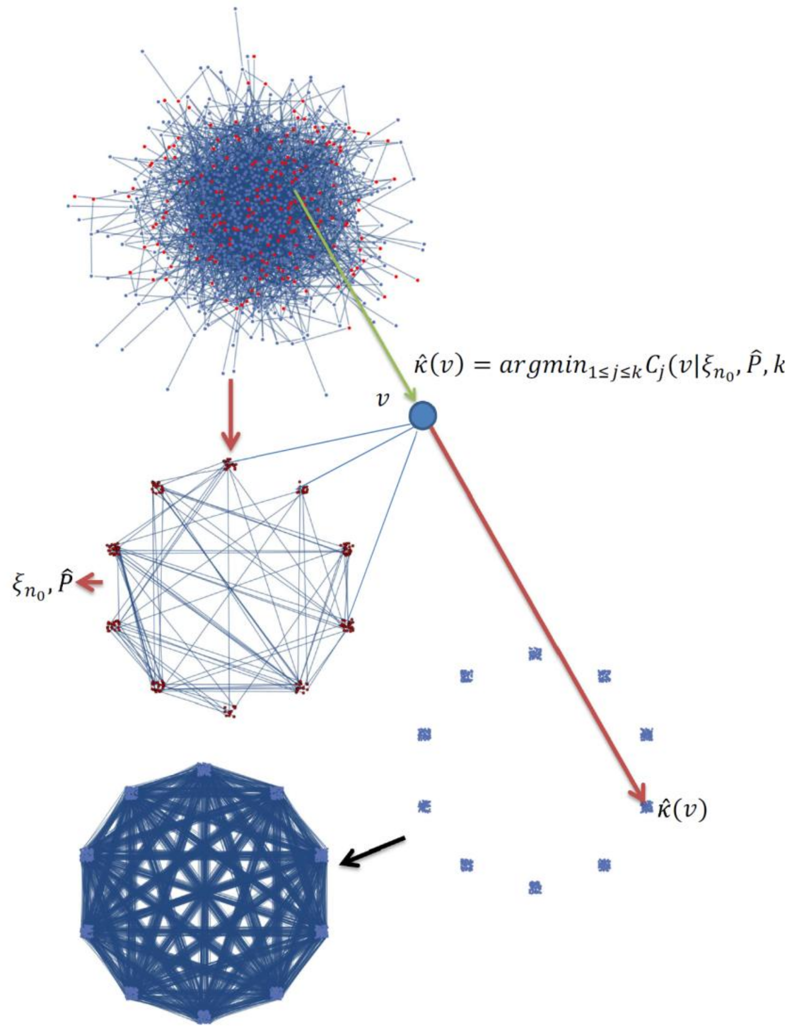
- => Too heavy for very large graphs
- Claim: if a regular structure with moderate k exists for a graph, then small sample is sufficient to find regular decomposition
- => regular decomposition is computationally feasible for big graphs
- Needs only to estimate link densities in every block
- => scales and tolerates missing link data



Sampling:
assume we have a large regular graph – k groups with regular link densities

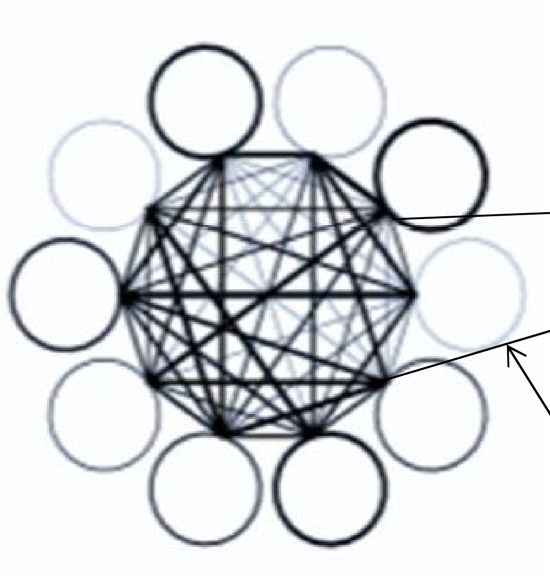
- Make a small uniformly random sample of nodes
- Retrieve links of induced small graph
- Find regular structure of the small sample graph
- Define a classifier based on sample graph
- Classify all nodes of the large graph (in linear time)
- => Compact representation of a graph => use in further analysis

Graphically:



Classifier:

A fixed sample graph with k regular groups ξ and a $k \times k$ link density matrix $\hat{d}_{i,j}$ and sizes of groups n_i



node v

Count number of links $e_j(v)$ to every regular group $1 \leq j \leq k$

choose the best class $1 \leq \alpha^* \leq k$

$$C_\alpha(v|\hat{\xi}, \hat{d}, k) := \sum_{j=1}^{k} \big[- e_j(v) \log \hat{d}_{j,\alpha} - (n_j - e_j(v)) \log(1 - \hat{d}_{j,\alpha}) \big],$$

$$\alpha^* = \arg\min_\alpha (C_\alpha(v|\hat{\xi}, \hat{d}, k)),$$

First experiments supporting conjectures of testability:

- 10 × 10 regular groups with uniformly random link densities $U(0,1)$
- 200 nodes is enough, 50 is too little; adjacency matrix

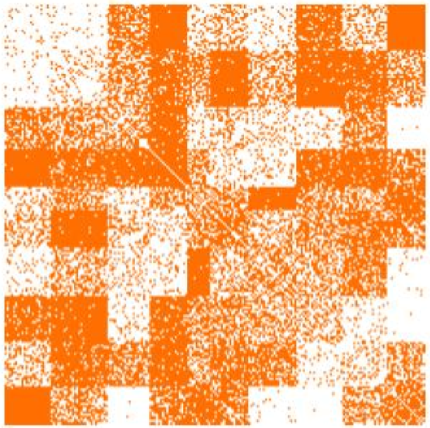


Fig. 6. 200 node sample, from the same model as above, that generates almost a perfect classifier - no errors detected in experiments.

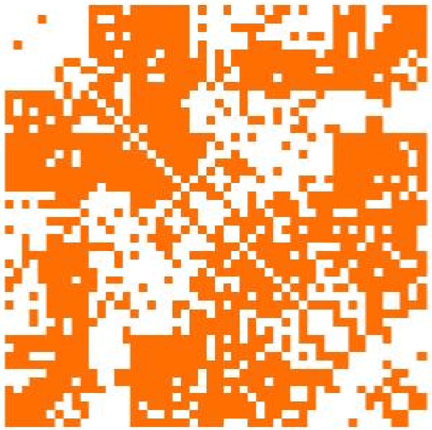


Fig. 5. A sample graph with 50 nodes that is insufficient to create a successful classifier - the result is similar to completely random classification.

Remarks:

- Error probability as a function of sample size?
  - 4 sources of classification errors (link densities, group sizes, misclassifications of sample, missing links)
  - Conjecture: exponentially small error probabilities
- Proof of existence (testability of graph sampling à la Lovasz)?
- Suggested sampling makes sense for <u>dense graphs</u>
  - How to extend to sparse case (different sampling style, sparse regularity…?)
- Similar approach should work also for real matrices, multi level graphs, tensors, hypergraphs (partly tested on data)

Thank You!