On Compression of Uniform Random Intersection Graphs

Zbigniew Golebiewski Marcin Kardas Jakub Lemiesz
Krzysztof Majcher

Wroclaw University of Science and Technology, Poland

AofA 2017

Structural Information

What:

- to develop compression algorithms for various structured data
- we focus on compression of unlabelled graphs

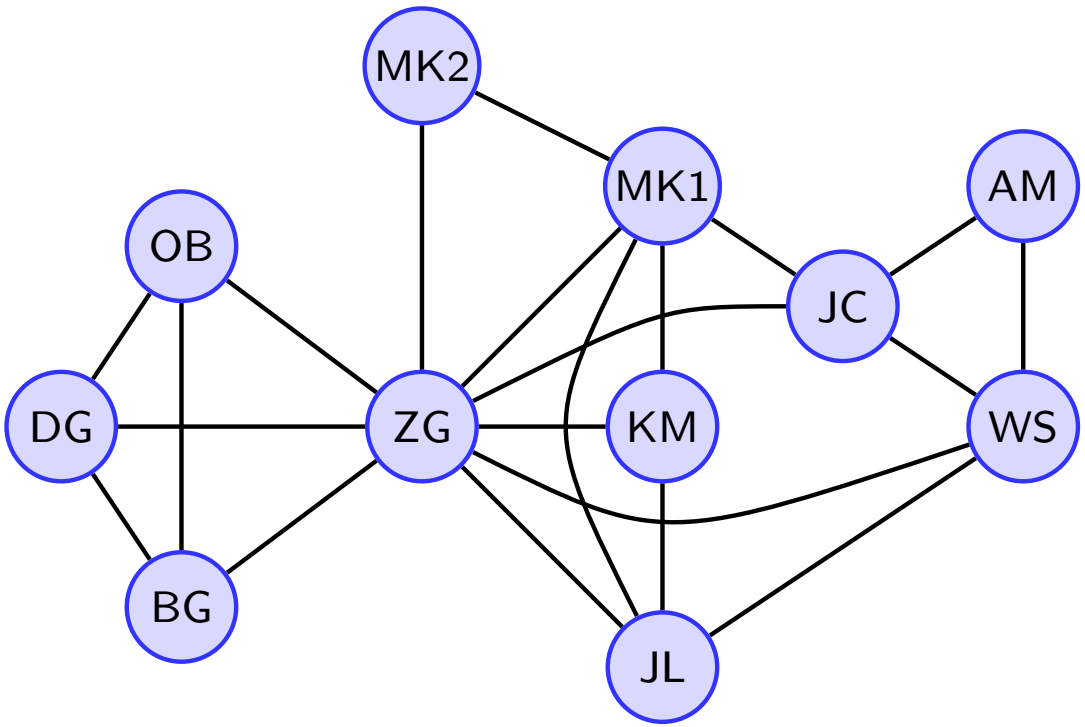
Why:

- many interesting combinatorial objects have structure (*web graph, protein-protein interactions, collaboration networks . . .*)
- they can be abstracted by (unlabelled) graphs

How:

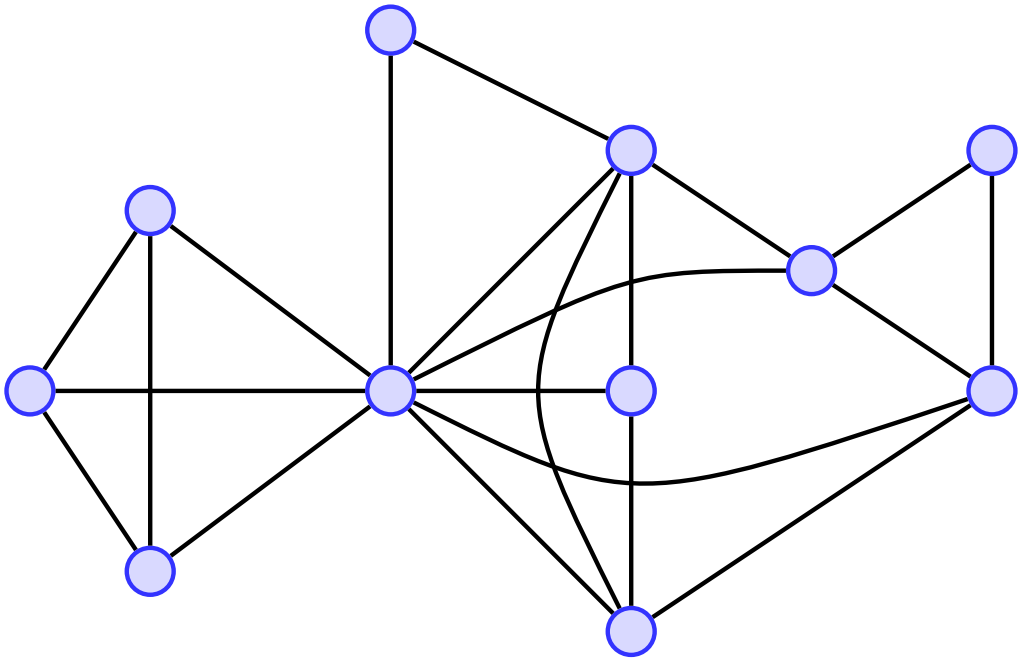
- using the structural entropy metric in order to measure the amount of information embodied in a graph structure

Graphs and Structural Entropy



ZG - Zbigniew Golebiewski
MK1 - Marcin Kardas
MK2 - Marek Klonowski
KM - Krzysztof Majcher
JL - Jakub Lemiesz
JC - Jacek Cichon
WS - Wojtek Szpankowski
AM - Abram Magner
OB - Olivier Bodini
DG - Daniele Gardy
BG - Bernhard Gittenberger

Graphs and Structural Entropy



&&

ZG - Zbigniew Golebiewski
MK1 - Marcin Kardas
MK2 - Marek Klonowski
KM - Krzysztof Majcher
JL - Jakub Lemiesz
JC - Jacek Cichon
WS - Wojtek Szpankowski
AM - Abram Magner
OB - Olivier Bodini
DG - Daniele Gardy
BG - Bernhard Gittenberger

{ZG, MK1, MK2, KM, JL, JC, WS, AM, OB, DG, BG}

Q: How many bits are required to describe the structure of a graph?

Graphs and Structural Entropy

- let $\mathcal{G}$ be a memory-less source producing graph according to some random graph model
- the classic entropy of $\mathcal{G}$ is defined as

$$H(\mathcal{G}) = -\sum_{G \in \mathcal{G}} \mathbb{P}_{\mathcal{G}}(G) \lg \mathbb{P}_{\mathcal{G}}(G)$$

- let $\mathcal{S}$ be a random *structure model* for the random graph model $\mathcal{G}$
- the probability of generating a given structure $S \in \mathcal{S}$ is

$$\mathbb{P}_{\mathcal{S}}(S) = \sum_{G \cong S, G \in \mathcal{G}} \mathbb{P}_{\mathcal{G}}(G)$$

- the structural entropy for model $\mathcal{G}$ is defined as

$$H_{\mathcal{S}}(\mathcal{G}) = -\sum_{S \in \mathcal{S}} \mathbb{P}_{\mathcal{S}}(S) \lg \mathbb{P}_{\mathcal{S}}(S)$$

Graphs and Structural Entropy

- let $\mathcal{G}$ be a memory-less source producing graph according to some random graph model
- the classic entropy of $\mathcal{G}$ is defined as

$$H(\mathcal{G}) = -\sum_{G \in \mathcal{G}} \mathbb{P}_{\mathcal{G}}(G) \lg \mathbb{P}_{\mathcal{G}}(G)$$

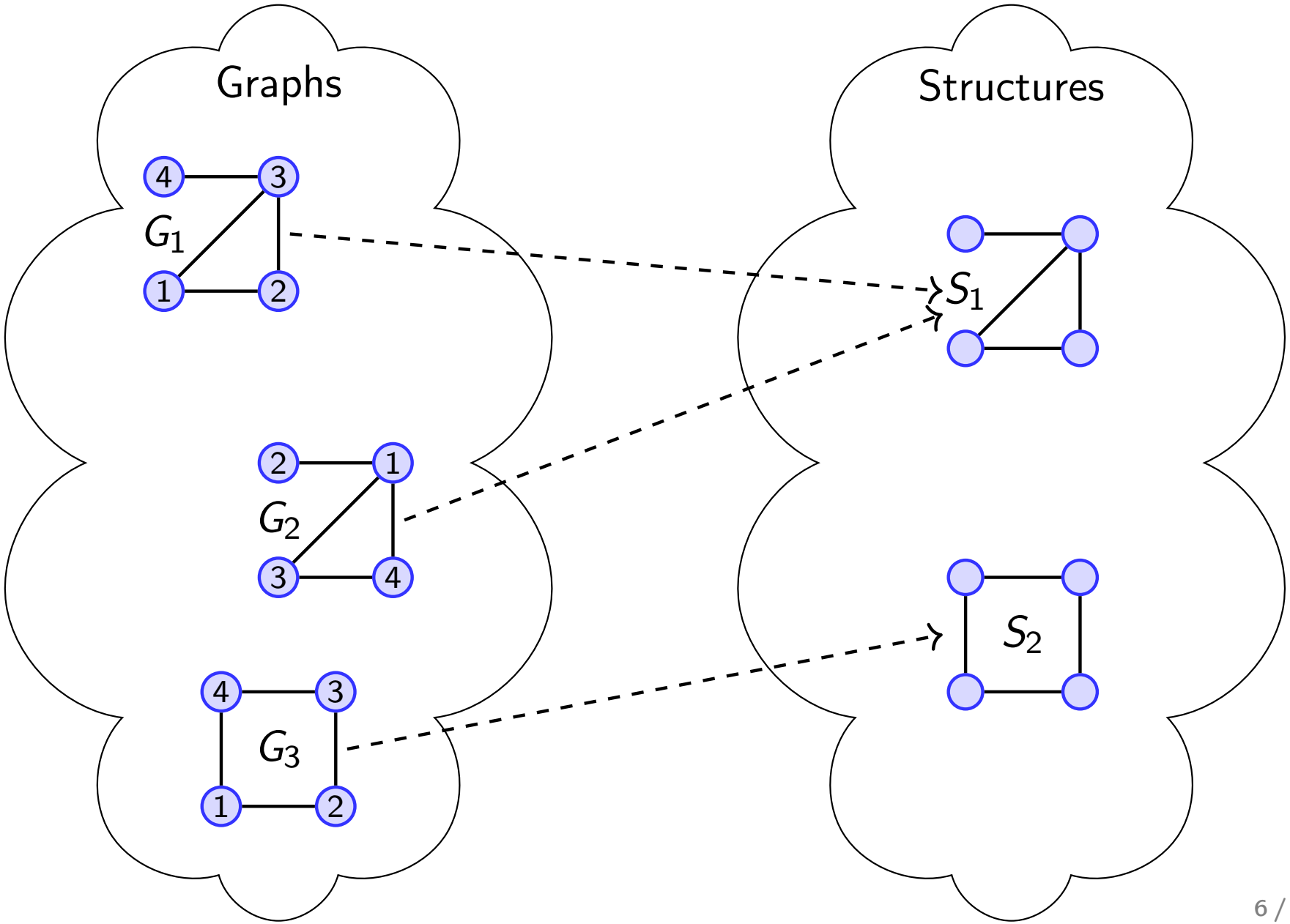
- let $\mathcal{S}$ be a random *structure model* for the random graph model $\mathcal{G}$
- the probability of generating a given structure $S \in \mathcal{S}$ is

$$\mathbb{P}_{\mathcal{S}}(S) = \sum_{G \cong S, G \in \mathcal{G}} \mathbb{P}_{\mathcal{G}}(G)$$

- the structural entropy for model $\mathcal{G}$ is defined as

$$H_{\mathcal{S}}(\mathcal{G}) = -\sum_{S \in \mathcal{S}} \mathbb{P}_{\mathcal{S}}(S) \lg \mathbb{P}_{\mathcal{S}}(S)$$

Graphs and Structural Entropy



Uniform Random Intersection Graph (URIG) Model

- $n \in \mathbb{N}_+$ – a number of nodes
- $m \in \mathbb{N}_+$ – a number of colors
- $k \in \{1, \ldots, m\}$ – a number of colors sampled (without replacement) independently by each node

By $\mathcal{U}_{n,m,k}$ we understand a random memory-less source producing undirected graphs with n vertices according to the following process:

- each vertex chooses uniformly at random a set of k colors out of m possible
- two vertices u and v are connected if and only if both sampled at least one common color.

Uniform Random Intersection Graph (URIG) Model

- $n \in \mathbb{N}_+$ – a number of nodes
- $m \in \mathbb{N}_+$ – a number of colors
- $k \in \{1, \ldots, m\}$ – a number of colors sampled (without replacement) independently by each node

By $\mathcal{U}_{n,m,k}$ we understand a random memory-less source producing undirected graphs with n vertices according to the following process:

- each vertex chooses uniformly at random a set of k colors out of m possible
- two vertices u and v are connected if and only if both sampled at least one common color.

Underlying Intersection Graph $G_{m,k}$

- has $m' = \binom{m}{k}$ vertices that correspond to all distinct k-element subsets of a set of colors $\{1, \ldots, m\}$
- to vertices v and u are connected iff they share at least one color, i.e. $v \cap u \neq \emptyset$

Lemma

The number of automorphisms of the underlying intersection graph $G_{m,k}$ is

$$|Aut(G_{m,k})| = \begin{cases} \binom{m}{k}! & \text{when } m < 2k, \\ \binom{m}{k}!! & \text{when } m = 2k, \\ m! & \text{when } m > 2k. \end{cases}$$

Underlying Intersection Graph $G_{m,k}$

- has $m' = \binom{m}{k}$ vertices that correspond to all distinct k-element subsets of a set of colors $\{1, \ldots, m\}$
- to vertices v and u are connected iff they share at least one color, i.e. $v \cap u \neq \emptyset$

Lemma

The number of automorphisms of the underlying intersection graph $G_{m,k}$ is

$$|Aut\,(G_{m,k})| = \begin{cases} \binom{m}{k}! & \text{when } m < 2k, \\ \binom{m}{k}!! & \text{when } m = 2k, \\ m! & \text{when } m > 2k. \end{cases}$$

Underlying Intersection Graph $G_{m,k}$

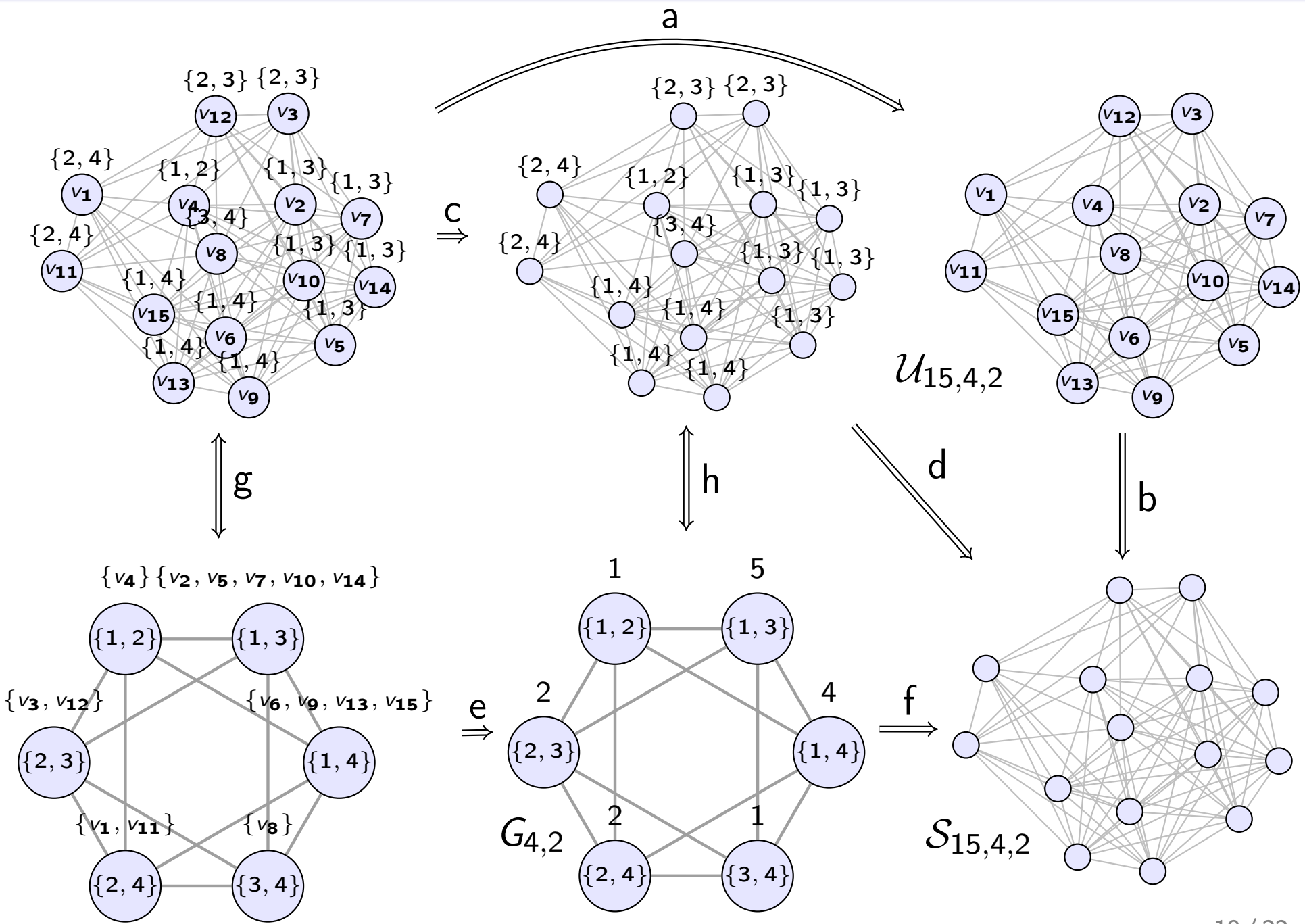
proof idea:

$$|Aut\,(G_{m,k})| = \begin{cases} \binom{m}{k}! & \text{when } m < 2k, \\ \binom{m}{k}!! & \text{when } m = 2k, \\ m! & \text{when } m > 2k, \end{cases}$$

complete graph

the complement graph contains only separated cliques of size 2

least symmetric case, known result for a Kneser graph, use of Erdős-Ko-Rado theorem


a
c
g
h
d
b
e
f
$\mathcal{U}_{15,4,2}$
$G_{4,2}$
$\mathcal{S}_{15,4,2}$
$\{v_4\}$ $\{v_2, v_5, v_7, v_{10}, v_{14}\}$
$\{v_3, v_{12}\}$
$\{v_6, v_9, v_{13}, v_{15}\}$
$\{v_1, v_{11}\}$
$\{v_8\}$
{1, 2}
{1, 3}
{2, 3}
{1, 4}
{2, 4}
{3, 4}
1
2
4
5
2
1

Random Composition Source

- let $\mathcal{K}_{n,G_{m,k}}$ be a set of all m'-element compositions of n $\left(m' = \binom{m}{k}\right)$, where the elements are indexed by vertices of $G_{m,k}$

$$\mathcal{K}_{n,G_{m,k}} = \left\{ K \in \mathbb{N}^V : \sum_{v \in V} K(v) = n \right\}$$

- observe that

$$\mathbb{P}\left(K = (k_1, \ldots, k_{m'})\right) = \binom{n}{k_1, \ldots, k_{m'}} \frac{1}{m'^n}$$

- therefore

$$H_{\mathcal{S}}\left(\mathcal{U}_{n,m,k}\right) = H_{\mathcal{S}}\left(\mathcal{K}_{n,G_{m,k}}\right)$$

Structural Entropy of Uniform Random Intersection Graph Source

Theorem (G., Kardas, Lemiesz, Majcher)

The structural entropy of a source generating uniform intersection graphs $\mathcal{U}_{n,m,k}$, for $m \geq 2k$, is

$$H_S\left(\mathcal{U}_{n,m,k}\right) = H\left(\mathcal{K}_{n,G_{m,k}}\right) - \lg\left|Aut\left(G_{m,k}\right)\right| + \mathbb{E}\left(\lg\left|stab(K)\right|\right) + o(1),$$

assuming that $n, \binom{m}{k} \to \infty$ in such a way that $\frac{n}{\binom{m}{k}} = \Theta\left(n^{\tau}\right)$ for any $0 < \tau \leq 1$.

Proof idea

- for $m \geq 2k$, let $K \in \mathcal{K}_{n,G_{m,k}}$ be a positive composition, then

$$[K]_{\approx} = orb(K) \stackrel{df}{=} \{K \circ \pi : \pi \in Aut\left(G_{m,k}\right)\}$$

  and all compositions of $[K]_{\approx}$ are equiprobable

- $$H_{\mathcal{S}}\left(\mathcal{U}_{n,m,k}\right) = -\mathbb{E}\left(\lg \mathbb{P}\left([K]_{\approx}\right)\right)$$

- by orbit-stabilizer theorem:

$$|[K]_{\approx}| = |orb(K)| = \frac{|Aut\left(G_{m,k}\right)|}{|stab(K)|},$$

  where $stab(K) = \{\pi \in Aut\left(G_{m,k}\right) : \pi \circ K = K\}$

Structural Entropy of Uniform Random Intersection Graph Source

Theorem (G., Kardas, Lemiesz, Majcher)

The structural entropy of a source generating uniform intersection graphs $\mathcal{U}_{n,m,k}$, for $m > 2k$, is

$$H_S(\mathcal{U}_{n,m,k}) = \binom{m}{k} \lg \sqrt{2\pi\alpha} - \lg \sqrt{2\pi n} + \frac{\binom{m}{k} - 1}{2\ln(2)} - lg(m!) + \frac{\binom{m}{k}}{\alpha \ln(2)} \sum_{l=0}^{\lfloor \frac{1-2\tau}{\tau} \rfloor} (-1)^l l! \mathtt{G}_{l+2} \alpha^{-l} + o(1),$$

assuming that $n, \binom{m}{k} \to \infty$ in such a way that $\alpha = \frac{n}{\binom{m}{k}} = \Theta(n^\tau)$ for any $0 < \tau \leq 1$ and $\mathtt{G}_l$ is l'th Gregory coefficient.

Remarks

- the structural entropy of the source $\mathcal{U}_{n,m,k}$ is relatively low compared, for example, to the source $\mathcal{PA}_{n,m}$ producing preferential attachment graphs or to the source $\mathcal{G}_{n,p}$ producing Erdős-Rényi graphs
- for the same expected number of edges in a graph we get:

$$H_{\mathcal{S}}\left(\mathcal{U}_{n,m,k}\right) = \Theta\left(\lg n\right)$$
$$H_{\mathcal{S}}\left(\mathcal{PA}_{n,m}\right) = \Theta\left(n \lg n\right)$$
$$H_{\mathcal{S}}\left(\mathcal{G}_{n,p}\right) = \Theta\left(n^2\right)$$

- it can be justified by the symmetries present in the graphs generated by $\mathcal{U}_{n,m,k}$

Proof idea

Lemma

If $n, \binom{m}{k} \to \infty$ *in such a way that* $\alpha = \frac{n}{\binom{m}{k}} = \Theta\left(n^{\tau}\right)$ *for any* $0 < \tau \leq 1$, *then*

$$H\left(\mathcal{K}_{n,G_{m,k}}\right) = \binom{m}{k} \lg\sqrt{2\pi\alpha} - \lg\sqrt{2\pi n} + \frac{\binom{m}{k} - 1}{2\ln 2} + \frac{\binom{m}{k}}{\alpha \ln 2} \sum_{l=0}^{\left\lfloor \frac{1-2\tau}{\tau} \right\rfloor} (-1)^l l! \mathtt{G}_{l+2} \alpha^{-l} + o(1),$$

where $\mathtt{G}_l$ *is* l*'th Gregory coefficient (known also as* l*'th logarithmic number) defined by a Maclaurin series expansion of*

$$\frac{y}{\log(1+y)} = 1 + \sum_{l=1}^{\infty} \mathtt{G}_l y^l.$$

Proof idea

- the entropy of the random composition source $\mathcal{K}_{n,G_{m,k}}$ is

$$H\left(\mathcal{K}_{n,G_{m,k}}\right) = -\sum_{K\in\mathcal{K}_{n,G_{m,k}}} \mathbb{P}(K)\lg\left(\mathbb{P}(K)\right)$$

- let us recall that $\mathbb{P}\left(K=(k_1,\ldots,k_{m'})\right) = \binom{n}{k_1,\ldots,k_{m'}}\frac{1}{m'^n}$
- after few steps we obtain a formula with Bernoulli sum

$$H\left(\mathcal{K}_{n,G_{m,k}}\right) = n\lg(m') - \lg(n!) + m'\sum_{t=0}^{n}\lg\left(t!\right)\binom{n}{t}\left(\frac{1}{m'}\right)^t\left(1-\frac{1}{m'}\right)^{n-t}$$

- choosing the approach due to Knessl, i.e. using:

$$\ln A = \lim_{\epsilon\to 0}\int_{\epsilon}^{\infty}\frac{e^{-x}-e^{-Ax}}{x}dx, \qquad A>0,$$

seems to give us the biggest freedom of choosing $\alpha = n/\binom{m}{k}$

Proof idea

Lemma

Let $m > 2k$, $k \geq 2$ and $n/\binom{m}{k} \to \infty$, then

$$\mathbb{E}\left(\lg |stab(K)|\right) = o(1) \quad \text{as } n \to \infty.$$

proof idea:

- for $n = \omega\left(\binom{m}{k}^5\right)$ we have different bins loads whp
- otherwise
  - let $\pi \in Aut\left(G_{m,k}\right)$ be a non-trivial automorphism of $G_{m,k}$ that is also a stabilizer of some random composition K
  - we can show that π has to move at least $2\binom{m-2}{k-1}$ vertices of $G_{m,k}$
  - let $\pi = C_1 \circ \ldots \circ C_\ell$ then for all i: $K = K \circ C_i$ and therefore all elements moved by the cycle C_i has to be equal
  - using poissonization technique we can bound the probability of such event

Optimal Compression Algorithm – work in progress

A lossless compression algorithm of a structure of URIG follows directions given by the analysis of the structural entropy.

Compress($U \in \mathcal{U}_{n,m,k}$) :

- ▷ contract vertices of U with the same colors subsets
- ▷ associate colors with the vertices of the underlying intersection graph

to this point we have reconstructed the underlying intersection graph $G_{m,k}$ and the composition K that corresponds to the graph U

- ▷ use arithmetic encoding to compress the composition K
- ▷ output: $(m, k, \mathit{compressed}(K))$

Future work

- quantitative result for the case $m = 2k$: stabilizers counts!
- case when n is smaller or comparable to $\binom{m}{k}$: contracted graph is a subgraph of the underlying intersection graph $G_{m,k}$
- structural entropy of the binomial intersection graphs source
- other symmetric graphs sources

Lessons learned

- asymmetric graphs $\rightarrow$ the difficulty can come from a source probability distribution
- with growth of the symmetry in the graphs generated by a source $\rightarrow$ the difficulty can come from both: a graph symmetries and a source probability distribution
- the most symmetric case is a clique $\rightarrow$ the analysis is trivial

Thank you!