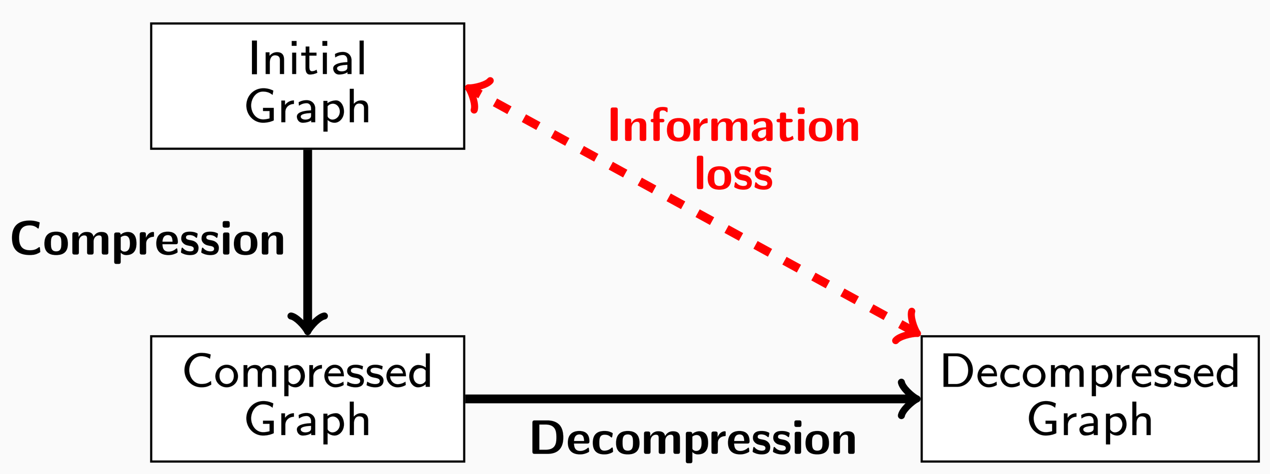
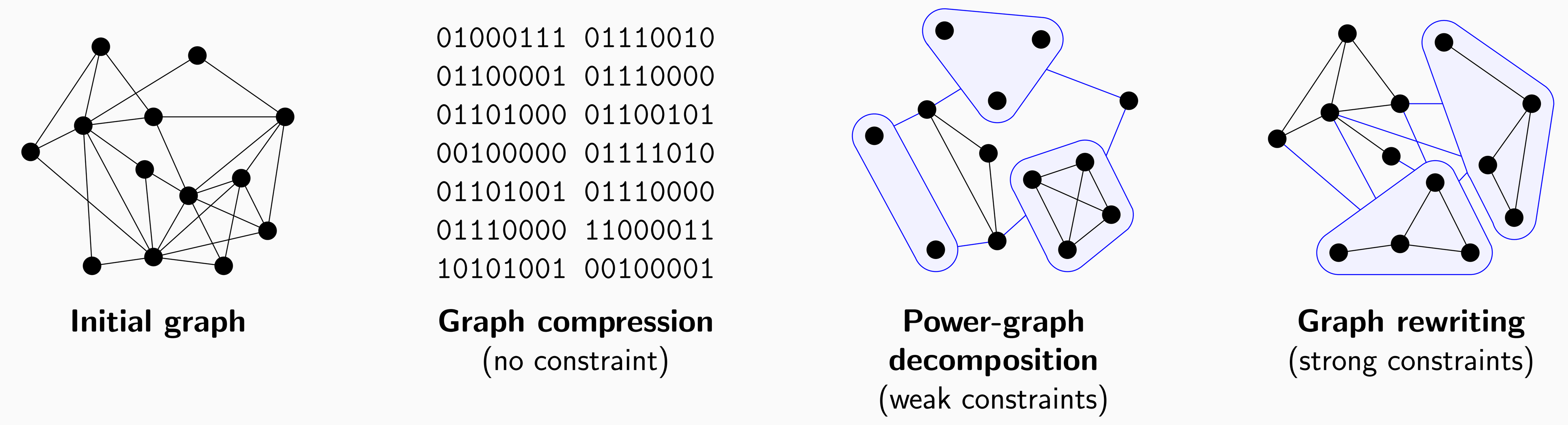
Information-theoretic Compression of Weighted Graphs

Robin Lamarche-Perrin, Lionel Tabourier, and Fabien Tarissan
{FirstName.LastName}@lip6.fr

General Setting: lossy compression



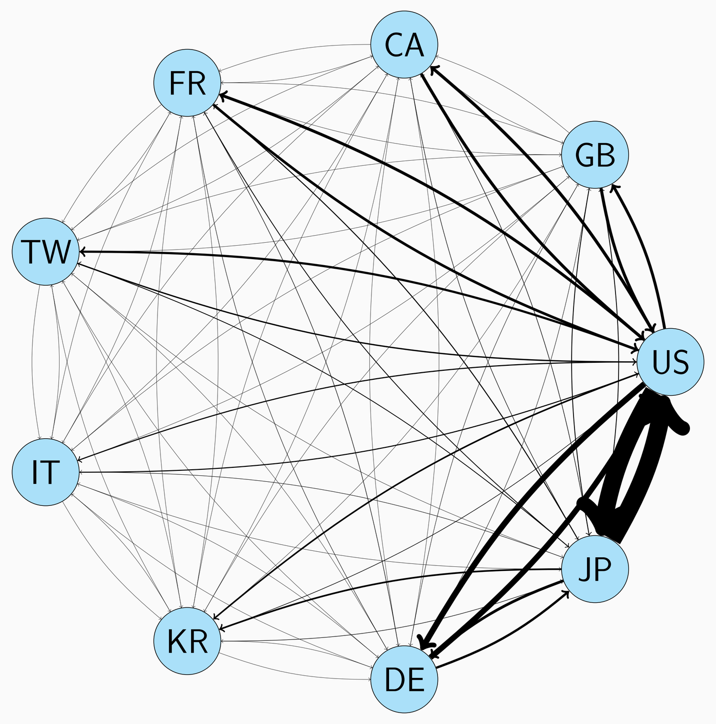
Related Work: lossless compression



1. Initial Graph

- **Weighted graph:** $G = (V_G, E_G, w_G)$
- **Set of vertices:** $V_G = \{v_1, \dots, v_n\}$
- **Set of arcs:** $E_G \subseteq V_G \times V_G$ with $(v_1, v_2) \in E_G \Rightarrow v_1 \neq v_2$
- **Weight function:** $w_G : V_G \times V_G \to \mathbb{R}^+$ with $(v_1, v_2) \notin E_G \Rightarrow w_G(v_1, v_2) = 0$
- **Empirical distribution:**

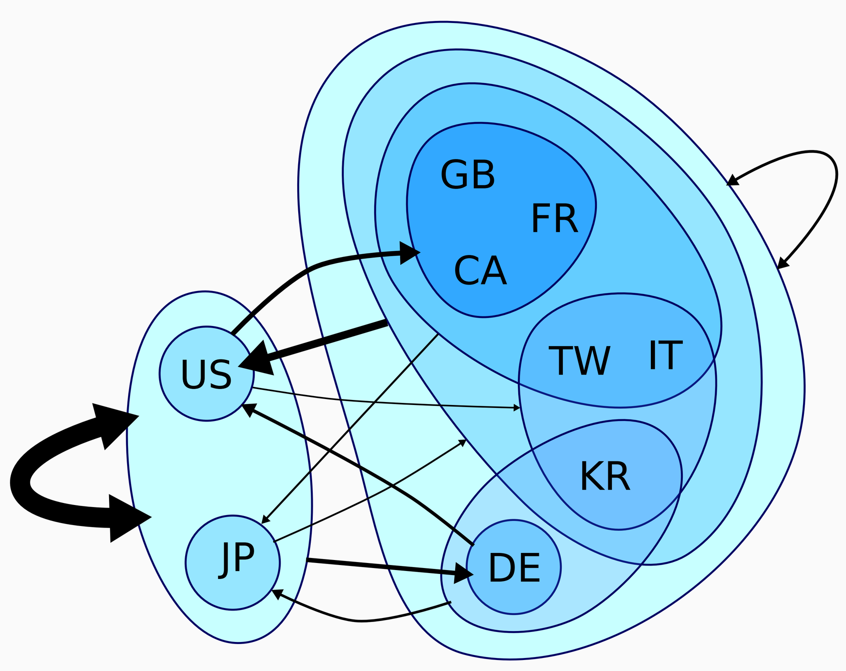
$$p_G(v_1, v_2) = \frac{w_G(v_1, v_2)}{\sum_{(v_1', v_2') \in E_G} w_G(v_1', v_2')}$$

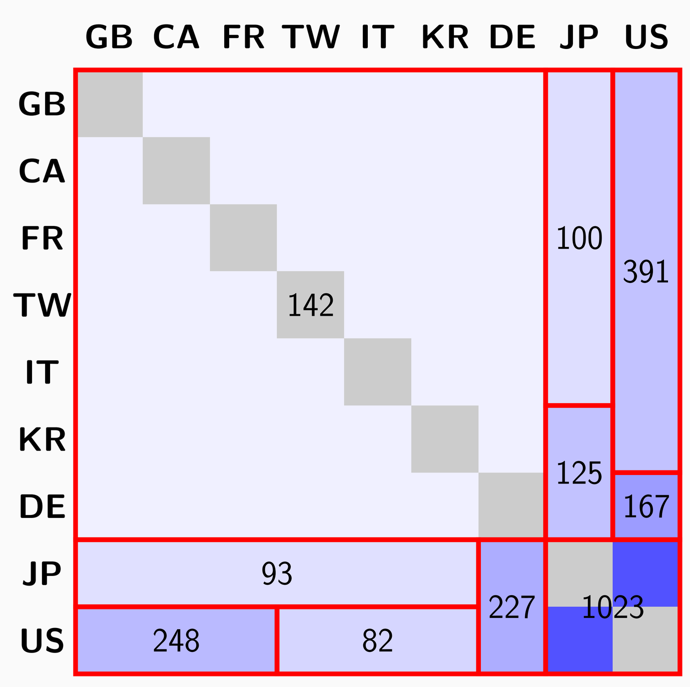


	GB	CA	FR	TW	IT	KR	DE	JP	US
GB		3	5	1	2	0	11	23	82
CA	3		3	2	1	0	6	15	89
FR	5	3		1	3	1	14	28	83
TW	2	3	2		1	3	4	22	62
IT	2	1	3	1		0	7	12	31
KR	2	1	2	2	1		3	47	44
DE	11	6	12	2	6	1		78	167
JP	24	14	23	9	9	14	66		504
US	86	87	75	37	29	16	161	519	

2. Compressed Graph

- **Super-vertex:** $V \subseteq V_G$
- **Super-edge:** $V_1 \times V_2 \subseteq V_G \times V_G$ with $V_1 \subseteq V_G$ and $V_2 \subseteq V_G$
- **Super-edge partition:** $\mathcal{E} = \{V_1^1 \times V_2^1, \dots, V_1^m \times V_2^m\}$ with $V_1^i \subseteq V_G$ and $V_2^i \subseteq V_G$, $\cup_i (V_1^i \times V_2^i) = V_G \times V_G$, $(V_1^i \times V_2^i) \cap (V_1^j \times V_2^j) = \emptyset$
- **Compressed weight function:** $w_G(V_1, V_2) = \sum_{(v_1, v_2) \in V_1 \times V_2} w_G(v_1, v_2)$





3. Decompressed Graph

- **Decompressed weight function:** $w_{\mathcal{E}} : V_G \times V_G \to \mathbb{R}^+$ such that $\forall V_1 \times V_2 \in \mathcal{E}$, $\forall (v_1, v_2) \in V_1 \times V_2$,

$$v_1 = v_2 \Rightarrow w_{\mathcal{E}}(v_1, v_2) = 0$$

$$v_1 \neq v_2 \Rightarrow w_{\mathcal{E}}(v_1, v_2) = \frac{w_G(V_1, V_2)}{(|V_1 \times V_2| - |V_1 \cup V_2|)}$$

- **Decompressed empirical distribution:**

$$p_{\mathcal{E}}(v_1, v_2) = \frac{w_{\mathcal{E}}(v_1, v_2)}{\sum_{(v_1', v_2') \in E_G} w_G(v_1', v_2')}$$

	GB	CA	FR	TW	IT	KR	DE	JP	US
GB		3.4	3.4	3.4	3.4	3.4	3.4	20.0	65.2
CA	3.4		3.4	3.4	3.4	3.4	3.4	20.0	65.2
FR	3.4	3.4		3.4	3.4	3.4	3.4	20.0	65.2
TW	3.4	3.4	3.4		3.4	3.4	3.4	20.0	65.2
IT	3.4	3.4	3.4	3.4		3.4	3.4	20.0	65.2
KR	3.4	3.4	3.4	3.4	3.4		3.4	62.5	65.2
DE	3.4	3.4	3.4	3.4	3.4	3.4		62.5	167
JP	16.5	16.5	16.5	16.5	16.5	16.5	114		512
US	82.7	82.7	82.7	27.3	27.3	27.3	114	512	

Optimisation Problem

Set Partitioning Problem (SPP):

- given a ground set $\Omega = \{x_1, \dots, x_n\}$,
- a collection of admissible subsets $\mathcal{P} = \{X_1, \dots, X_m\} \subseteq 2^\Omega$,
- a cost function $c : \mathcal{P} \to \mathbb{R}$,
- find a partition $\mathcal{X}$ of Ω using subsets in $\mathcal{P}$ and minimising the sum of the costs $\min_{\mathcal{X}} \sum_{X \in \mathcal{X}} c(X)$.

Complete SPP (CSPP):

- All subsets are admissible: $\mathcal{P} = 2^\Omega$

Bidimensional Complete SPP (CSPP × CSPP):

- The ground set is a Cartesian product: $\Omega = \Omega_1 \times \Omega_2$
- Admissible subsets are all Cartesian products: $\mathcal{P} = 2^{\Omega_1} \times 2^{\Omega_2}$
- In our case: $\Omega_1 = \Omega_2 = V_G$, $\mathcal{X} = \mathcal{E}$, and $c(V_1, V_2) = \mathrm{comp}(V_1, V_2) + \beta \,\mathrm{loss}(V_1, V_2)$ (see on the right)

Optimisation Algorithm: Dynamic programming algorithm (branching, recursion, memoization, non-redundancy)
https://github.com/Lamarche-Perrin/optimal_partition

Dataset

National Patent Citations: $w_G(v_1, v_2)$ is the number of patents granted in country v_1 and citing a patent granted in country v_2 (unit: hundred of patents, only for the 9 most cited countries over the period 1990-1999)

Source: NBER U.S. Patent Citations Data File
http://www.nber.org/patents/

Objective Function

- **Information loss:** Kullback-Leibler divergence between the initial and the resulting distribution

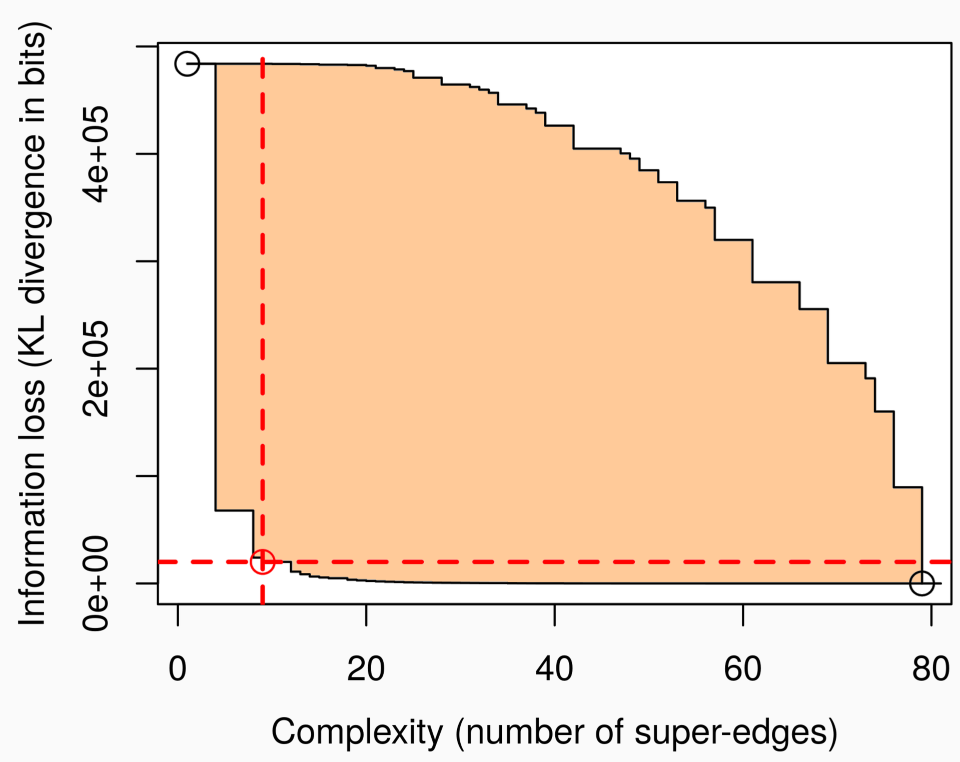
$$\mathrm{loss}(\mathcal{E}) = \sum_{(v_1, v_2) \in V_G \times V_G} p_G(v_1, v_2) \log_2 \left(\frac{p_G(v_1, v_2)}{p_{\mathcal{E}}(v_1, v_2)} \right)$$

- **Complexity:** number of super-edges

$$\mathrm{comp}(\mathcal{E}) = |\mathcal{E}|$$

- **Variational:**

$$\min_{\mathcal{E}} \ \mathrm{comp}(\mathcal{E}) + \beta \,\mathrm{loss}(\mathcal{E}) \text{ with } \beta \in \mathbb{R}^+$$



References

- R. Lamarche-Perrin.
Optimal Partition: A Toolbox to Solve Structured Versions of the Set Partitioning Problem with Decomposable Objectives.
GitHub, https://github.com/Lamarche-Perrin/optimal_partition/, 2015.
- R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent.
A Generic Algorithmic Framework to Solve Special Versions of the Set Partitioning Problem.
In *Proceedings of the 26th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'14)*, pages 891–897. IEEE Computer Society, 2014.
- R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent.
Building Optimal Macroscopic Representations of Complex Multi-agent Systems.
In *Transactions on Computational Collective Intelligence*, volume XV of *LNCS 8670*, pages 1–27. Springer-Verlag, 2014.