Information-theoretic Compression of Weighted Graphs

Robin Lamarche-Perrin[1], Lionel Tabourier[1], and Fabien Tarissan[2]
{Firstname.Lastname}@lip6.fr

Keywords: Graph Compression, Multilevel Analysis, Information Theory, Power Graph Analysis, Combinatorial Optimisation, Set Partitioning Problem.

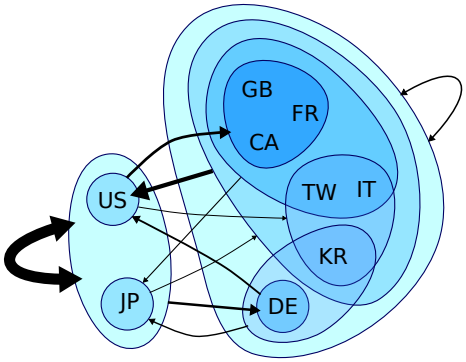
In order to handle large-scale and real-world networks that are nowadays available for analysis, one has to develop proper computational tools for the understanding of such complex datasets. In this poster, we present a *data compression framework* providing domain experts with some tools for *multilevel description of weighted directed graphs*. In this context, data compression is modelled as a combinatorial optimisation problem for which one has to make explicit (1) the set of formal operators that can be used for data processing (*e.g.*, wavelet transform or tiling in the case of image compression) and (2) a quantitative objective function to be optimised (*e.g.*, reducing the size of data in memory while maintaining its information content).

Regarding formal operators, the literature on graph compression spans from strongly-constrained rewriting approaches (where only very specific topological patterns, such as cliques, are subject to compression) to fully unconstrained compression schemes (where any operation is allowed, hence hugely modifying the original graph structure). We decide here to focus on *power graph decomposition*, that is an intermediate approach which stays quite generic while preserving a graph-like language: nodes are aggregated into *supernodes* and edges into *superedges* [1]. We show that this approach is quite close to *block model analysis*, that is a compression scheme for equivalence relations consisting in partitioning a symmetric adjacency matrix with "guillotine cuts". We generalise this approach by allowing any partition of the adjacency matrix into "rectangular tiles", thus allowing the compression of edges wrt overlapping supernodes (see figure on next page). We identify the resulting optimisation problem as a special version of the *Set Partitioning Problem* for which the solution space is defined from the Cartesian product of two unconstrained sets (CSPP$\times$CSPP, see [3]).

Regarding objective functions, previous work addressing weighted graph compression uses *distance-based* objectives to compare a graph with its compressed versions [5]. However, when the weights represent a *counting process* over edges, we instead recommend to use *information-based* measures. Indeed, in this case, we propose to interpret the graph as a probability distribution over the edge set (that is the probability of selecting a given edge according to the counting process) and the compressed graph as an approximation of this distribution. The information that is lost in the process can then be quantified by the Kullback-Leibler divergence, that is the expected number of supplementary bits required to encode the counts while using the approximate distribution instead

[1]Sorbonne Universités, UPMC Univ. Paris 6, CNRS, LIP6 UMR 7606, 75005 Paris, France.
[2]Université Paris-Saclay, ISP, ENS Cachan, CNRS, 94235 Cachan, France.

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



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

Compressing the network of patent citations between countries: On the left, the adjacency matrix gives the number of patents (in hundreds) granted during the 90s in a given country and that cite a patent from another country. Our approach allows to partition this matrix in multiscale rectangular tiles, each representing on the right a superedge between two supernodes. In this example, the information is summarised by 10 superedges, resulting in only a 3 % information loss (according to Kullback-Leibler divergence).

of the original distribution [4]. We propose a measure that linearly combines this information loss and the gain in terms of description size (*i.e.*, the number of superedges) by introducing a *scale parameter* in order to provide a cursor going from the complete microscopic description (minimal information loss) to the most compressed description (minimal size).

The resulting optimisation problem is NP-complete [3]. However, because the information-based objectives we define are *additively decomposable* (*i.e.*, the quality of a partition can be defined as the sum of the quality of its parts), dynamic programming helps us to exploit the algebraic structure of the solution space to efficiently compute an optimal partition (see [3] for details). We implemented such an exponential algorithm [2] and applied it on a very small real-world network to illustrate our approach (see figure). The multilevel descriptions resulting from this work are quite encouraging, and we will now work on the design of heuristics to scale our approach to much larger graphs.

[1] T. Dwyer, N. H. Riche, K. Marriott, and C. Mears. Edge Compression Techniques for Visualization of Dense Directed Graphs. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2596–2605, 2013.

[2] R. Lamarche-Perrin. Optimal Partition: A Toolbox to Solve Structured Versions of the Set Partitioning Problem. `https://github.com/Lamarche-Perrin/optimal_partition/`, 2015.

[3] 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 ICTAI'14*, pages 891–897. IEEE Computer Society, 2014.

[4] 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.

[5] H. Toivonen, F. Zhou, A. Hartikainen, and A. Hinkka. Compression of Weighted Graphs. In *Proceedings of KDD'11*, pages 965–973. ACM, 2011.