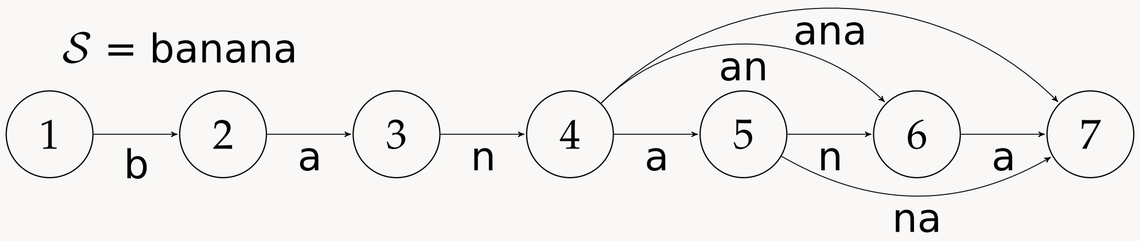
Bicriteria LZ77 Compression

The Bicriteria LZ77 Parsing Problem

- **More than just compression ratio** The advent of massive datasets and high-performing storage systems have reignited the interest towards the design of lossless data compressors which achieve effective compression ratio and very efficient decompression speed.
- **LZ77** Lempel-Ziv's LZ77 algorithm is the *de facto* choice due to its decompression performance and its algorithmic flexibility, which allow to trade decompression speed for compression ratio.
- **Picking between different trade-offs** Each existing implementation offers a *single* trade-off between space occupancy and decompression speed, so software engineers have to content themselves by picking the one which comes closer to the requirements of the application in their hands.
- **The Bicriteria LZ77 Parsing Problem** Find a parsing which minimize the consumption of one resource (decompression time, compressed size) given a bound on the consumption of the other one.

Modeling as a WCSPP

- The set of LZ77 parsings of a string $\mathcal{S}$ of length n may be expressed as the source-destination paths over a graph $\mathcal{G}(\mathcal{S})$ with $O(n^2)$ edges, such that:
  - there is a vertex v_i for each character $\mathcal{S}[i]$;
  - there is an edge (v_i, v_j) for each substring $\mathcal{S}[i, j-1]$ in the dictionary.



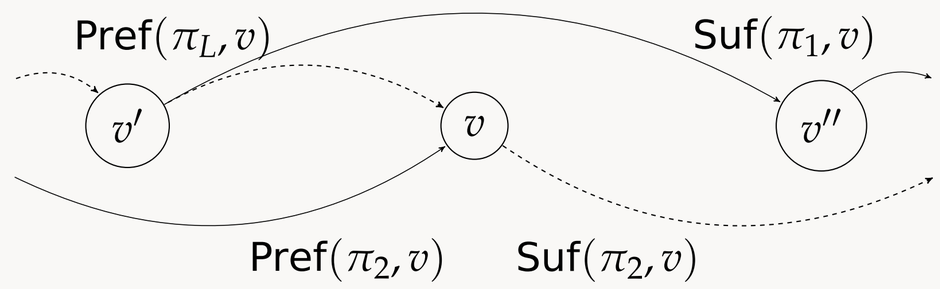
- each edge, which correspond to a phrase, is weighted with a *time* and *space* weight.
  - the space weight is the codeword length in bits, while its time weight is given by an experimental, scan-based time model.
- The Bicriteria LZ77 Parsing Problem is thus reduced to a **Weight-Constrained Shortest Path Problem** over $\mathcal{G}(\mathcal{S})$.

Solving the WCSPP on $\mathcal{G}(\mathcal{S})$

- General-purpose WCSPP resolution algorithms are not appropriate in this context.
  - the graph may be very huge: the number of edges of the graph induced by a one-gigabyte file can be up to $2^{32 \cdot 2} = 2^{64}$ edges, which make storing it unfeasible.
  - state-of-the-art algorithms for WCSPP, when applied to the Bicriteria LZ77 Parsing problem, have a complexity of at least $\Omega(n^2)$, which is unacceptable in practice.
- Our algorithm exploits some peculiar structural properties of $\mathcal{G}(\mathcal{S})$ to achieve $O(n \log^2 n)$ time and $O(n)$ auxiliary space complexity.
- The algorithm is an **additive** $(O(\log n), O(\log n))$-approximation algorithm.
  - Assuming that the optimal solution has compressed size s and the decompression time bound is T, the algorithm finds a solution with compressed space and decompression time bounded by $s + O(\log n)$ and $T + O(\log n)$.
  - Very close to the optimum, even on small files.

Our solution

- Our solution can be decomposed in four steps.
- **Pruning** Under some broad assumptions about the encoding functions and the memory hierarchy, the number of edges may be reduced from $O(n^2)$ to just $O(n \log n)$ in an implicit fashion.
- **Forward Star Generation** Each edge is dynamically generated when needed in $O(1)$ amortized time, in order to achieve $O(n)$ space complexity.
- **Lagrangian Relaxation** We solve the Lagrangian Dual relaxation of the WCSPP in $O(n \log^2 n)$ time through the Cutting Plane algorithm. This phase yields a lower and upper-bound on the cost of the optimal solution, plus a pair of paths (π_L, π_R) which constitute an *optimal basis* of the dual problem.
- **Approximate Gap-closing** We obtain an additive $(O(\log n), O(\log n))$-approximation by combining together the paths π_L and π_R in $O(n)$ time and space. The resulting path is composed by a prefix of π_L and a suffix of π_R starting from a carefully-picked vertex v, plus a *swapping bridge* connecting the two sub-paths.



Experimental Results (DBLP, 1GB)

Parsing	Compressed size (MB)	Decompression time (seconds)
BC-ZIP - 1	129.8	2.95
BC-ZIP - 0.8	131.4	2.77
BC-ZIP - 0.6	134.6	2.56
BC-ZIP - 0.4	139.3	2.32
BC-ZIP - 0.2	**148.5**	**1.96**
Snappy	323.4	2.13
LZ4	214.7	1.98
zlib	190.5	11.65
bzip2	121.4	48.98

Experimental results show that our approach allows to effectively control the time-space trade-off in a practical yet principled manner. Moreover, it leads to parsings which are faster to decode and more space-succinct than those generated by highly tuned and engineered parsing heuristics, like those of Google Snappy and LZ4

References

- Andrea Farruggia, Paolo Ferragina, Antonio Frangioni, and Rossano Venturini.
  Bicriteria data compression.
  ArXiv e-prints, July 2013.
- Paolo Ferragina, Igor Nitto, and Rossano Venturini.
  On the bit-complexity of lempel-ziv compression.
  In *SODA*, pages 768–777, 2009.

Andrea Farruggia, Paolo Ferragina, Antonio Frangioni, Rossano Venturini
{farruggi, ferragina, frangio, rossano}@di.unipi.it
Dipartimento di Informatica, University of Pisa
UNIVERSITÀ DI PISA