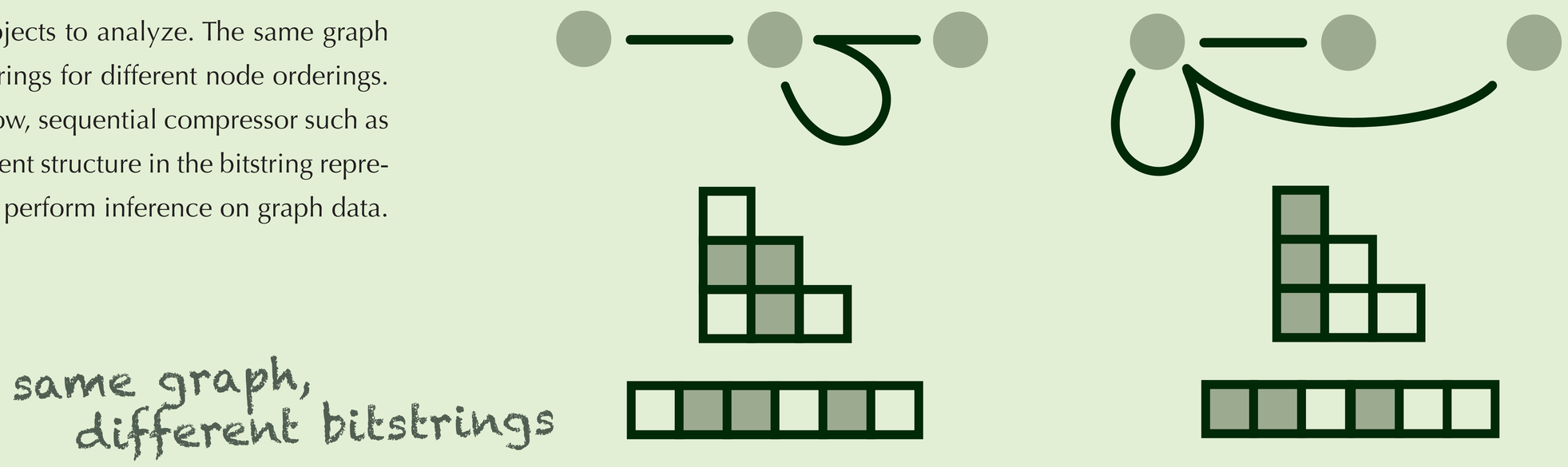
Compression-based inference on graphs

Peter Bloem, p@peterbloem.nl
System and Network Engineering Group, University of Amsterdam

THE PROBLEM OF ISOMORPHISM: SEQUENTIAL REPRESENTATIONS OF GRAPHS

Graphs are complex objects to analyze. The same graph generates different bitstrings for different node orderings. We ask whether a shallow, sequential compressor such as gzip can still find sufficient structure in the bitstring representation to succesfully perform inference on graph data.



To test whether the problem of isomorphism affects sequential compressors, we performed the following experiment: we generated data from four synthetic sources and computed the NCD between all pairs, clustering the full set with the k-medoids algorithm. As a compressor we used either the fast and sequential gzip, or the slower, graph-based Subdue[3] algorithm. The Subdue compressor does not suffer the problem of isomorphism, as it operates on a graph directly, rather than on a sequential representation of a graph.

The results show some performance gain by Subdue (at the cost of significantly higher running time), but more interestingly, they show that gzip—despite the problem of isomorphism—can still pick up sufficient patterns from the sequential representation to outperform the baseline.

random	0.17	0.083	0	0
pa	0.083	0.17	0	0
fractal	0.083	0	0.17	0
fractal sw	0	0	0	0.25
mean error	0.46 (0.11)			

	0	0.25	0	0
	0	0.25	0	0
	0	0	0.25	0
	0	0	0	0.25
	0.27 (0.12)			

	0.25	0	0	0
	0	0.25	0	0
	0	0	0.25	0
	0	0	0	0.25
	0.14 (0.14)			

cellular	0.11	0.11	0.11
neural	0.11	0.11	0.11
co-purchase	0.11	0	0.22
mean error	0.43 (0.11)		

random baseline

	0.33	0	0
	0.22	0.11	0
	0	0.22	0.11
	0.28 (0.17)		

gzip

	0.33	0	0
	0	0.33	0
	0.33	0	0
	0.34 (0.17)		

subdue

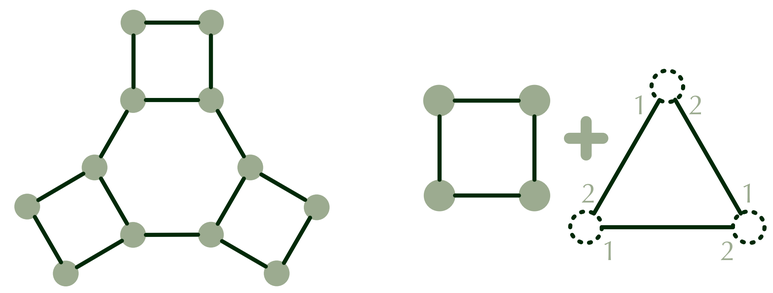
NORMALIZED COMPRESSION DISTANCE

The normalized compression distance[1,2] converts a general-purpose compressor into an inference tool. It measures the distance between two objects by compressing them separately and put together, and compares the two values.

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\left[C(x), C(y)\right]}{\max\left[C(x), C(y)\right]}$$

The distances this defines over a dataset of objects produce a clustering that often conforms to natural intuitions. The normalized compression distance has been used to successfully cluster such objects as books, genomic information and MIDI files of works by classical composers.

SUBDUE



data compressed

REFERENCES

1. Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. (2004). The similarity metric. *Information Theory, IEEE Transactions on*, 50, 3250–3264.
2. Cilibrasi, R., & Vitányi, P. M. (2005). Clustering by compression. *Information Theory, IEEE Transactions on*, 51, 1523–1545.
3. Ketkar, N. S., Holder, L. B., & Cook, D. J. (2005). Subdue: compression-based frequent pattern discovery in graph data. *Proceedings of the 1st int. workshop on open source data mining*. (pp. 71–76).

For code and data, see peterbloem.nl/benelearn2013

COMMIT/ Data2Semantics From Data to Semantics for Scientific Data Publishers University of Amsterdam System and Network Engineering