Lossy Compression of Dynamic, Weighted Graphs

Wilko Henecka and Matthew Roughan

matthew.roughan@adelaide.edu.au

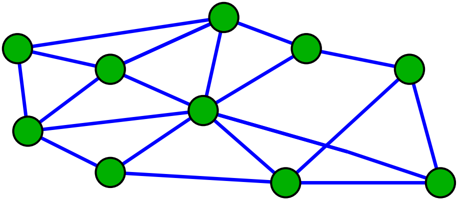
http://www.maths.adelaide.edu.au/matthew.roughan/

UoA

August 6, 2015

Graphs

- Graph: $G(N, E)$
  - N = set of nodes (vertices)
  - E = set of edges (links)



- Often we have additional information on links, e.g.,
  - link distance
  - link capacity
  - link strength

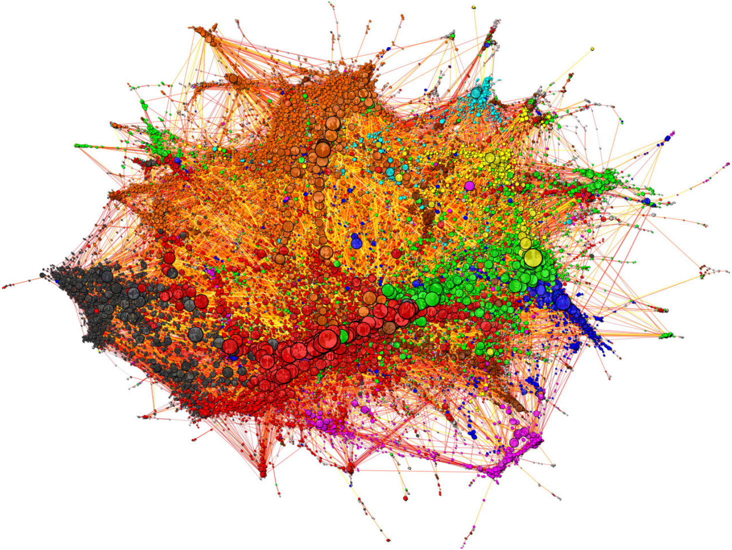
  we call these *weights*.
- Often graphs change over time
  - nodes, links, and weights can change

  we call these *dynamic* graphs

Why?

- To represent data where "connections" are 1st class objects in their own right
  - storing the data in the right format improves access, processing, ...
  - it's natural, elegant, and might be *efficient* if we do it properly
- Many examples
  - Telephone call records: how often does person A call B
    - AT&T use this to detect fraudsters (amongst other things)
  - Musicians – how alike are musicians A and B
    - last.fm use to make music recommendations

Network of Musicians (last.fm)



http://sixdegrees.hu/last.fm/

Compression

- Compression is almost ubiquitous now
  - lossless vs lossy (*e.g.,* GIF vs JPEG)
  - algorithm vs encoding (*e.g.,* DCT+quantisation vs Huffman Coding)
- Type of graph that is compressed

	lossless	lossy
static, unweighted	[1, 2, 3]	[4, 5]
static, weighted	[6]	[7, 8]
dynamic, unweighted	[9]	
dynamic, weighted		[10, 11, 12]

Tool

Weighted sum of weighted graphs

$$G = \alpha A \oplus \beta B$$

means

$$\begin{aligned} N(G) &= N(A) \cup N(B), \\ E(G) &= E(A) \cup E(B), \end{aligned}$$

and

$$w_G(e) = \alpha w_A(e) + \beta w_B(e), \text{ for all } e \in E(G),$$

where if an edge is not present, we treat it as if it has weight 0.

Measuring dynamic graphs

- Usually we can't see the graph itself
  - we see a proxy measurement
  - we have errors because network changes, and measurement errors
- Example: call records
  - underlying graph gives social connections
  - measure a links strength by number of calls
  - underlying graph evolves at the same time as we measure it
- Exponentially Weighted Moving Average (EWMA) Graph

$$G_t = \theta G_{t-1} \oplus (1 - \theta) g_t,$$

  - g_t is measured graph in current time interval t
  - G_t is updated estimate of graph
  - $1 - \theta$ is the "gain"

Approximation

- Lossy compression is approximation
  - on-line algorithms combine estimation and approximation
- Mathematical representation in this context

$$\hat{G}_t = A\Big(\theta \hat{G}_{t-1} \oplus (1-\theta) g_t\Big).$$

where $A(\cdot)$ is an approximation function

  - can prune edges
  - can approximate edge weights

Top-k approximation

- Idea is to model *Community of Interest (COI)* signature [11, 12]
  - approximation is just "take the top k edges"
  - also prune edges whose weight falls below ϵ
- Parameters (k, ϵ)
  - choose so that 95% of edges are kept
- Applied to detecting fraudsters
  - you are who you call
  - compare COI signature of new customers to database of "bad" accounts
- Problems:
  - non-trivial to choose parameters
  - doesn't work well as general approximation technique

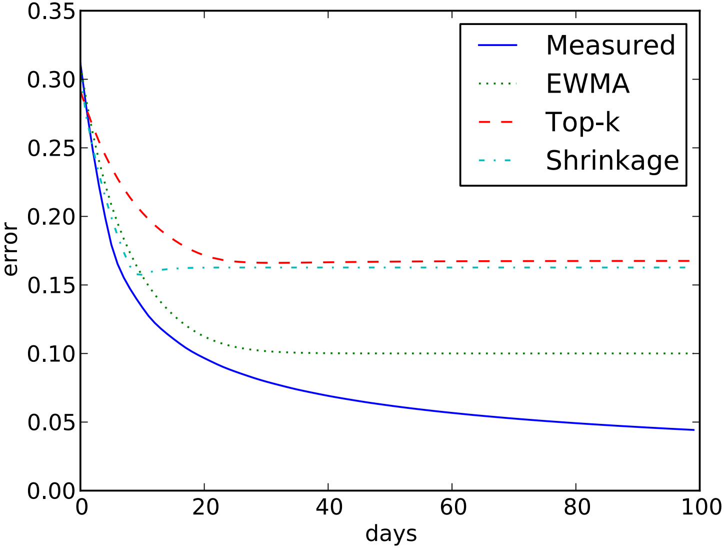
Shrinkage approximation

- Similar idea, but don't make a fixed k
- All weights are soft thresholded

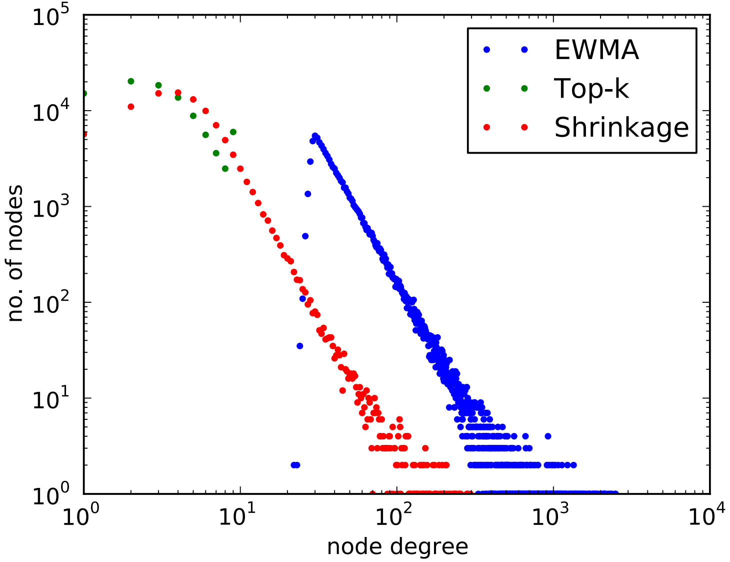
$$w_{\hat{G}}(e) = [w_G(e) - \lambda]^+ ,$$

- Only one parameter λ
- Draws on ideas from de-noising

Results: errors with compression approximately 10:1



Results: compressed degree distributions



Conclusion

- Graph Compression is Good
  - lossy compression can reduce size of data 10:1 with reasonable errors
- New method
  - shrinkage outperforms top-k in many respects
- Haven't talked about encoding at all
  - we don't know how this approximation interacts with encoding, but it should be good as we are de-noising
  - encoding works better on structured data (as opposed to noise)

S. Chen and J. Reif, "Efficient lossless compression of trees and graphs," in *Proceedings of the IEE Data Compression Conference (DCC '96)*, pp. 428–437, Mar 1996.

P. Boldi and S. Vigna, "The WebGraph framework I: Compression techniques," in *Thirteenth World-Wide Web Conference*, pp. 595–601, 2004.

F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan, "On compressing social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, (New York, NY, USA), pp. 219–228, ACM, 2009.

S. Navlakha, R. Rastogi, and N. Shrivastava, "Graph summarization with bounded error," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, (New York, NY, USA), pp. 419–432, ACM, 2008.

A. C. Gilbert and K. Levchenko, "Compressing network graphs," in *Proceedings of the LinkKDD workshop at the 10th ACM Conference on KDD*, August 2004.

J. Willcock and A. Lumsdaine, "Accelerating sparse matrix computations via data compression," in *Proceedings of the 20th Annual International Conference on Supercomputing*, ICS '06, (New York, NY, USA), pp. 307–316, ACM, 2006.

F. Zhou, S. Mahler, and H. Toivonen, "Simplification of networks by edge pruning," in *Bisociative Knowledge Discovery* (M. Berthold, ed.), vol. 7250 of *LNCS*, pp. 179–198, Springer, 2012.

H. Toivonen, F. Zhou, A. Hartikainen, and A. Hinkka, "Compression of weighted graphs," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, (New York, NY, USA), pp. 965–973, ACM, 2011.

C. H. You, L. Holder, and D. Cook, "Graph-based data mining in dynamic networks: Empirical comparison of compression-based and frequency-based subgraph mining," in *IEEE International Conference on Data Mining Workshops ICDMW '08*, pp. 929–938, Dec 2008.

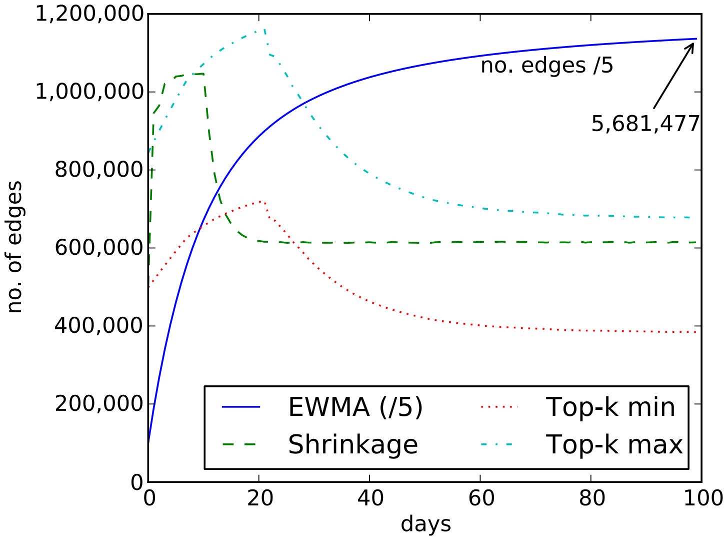
W. Liu, A. Kan, J. Chan, J. Bailey, C. Leckie, J. Pei, and R. Kotagiri, "On compressing weighted time-evolving graphs," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, (New York, NY, USA), pp. 2319–2322, ACM, 2012.

S. B. Hill, D. K. Agarwal, R. Bell, and C. Volinsky, "Building an effective representation for dynamic networks," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 584–608, 2006.

C. Cortes, D. Pregibon, and C. Volinsky, "Communities of interest," in *Advances in Intelligent Data Analysis*, pp. 105–114, Springer, 2001.

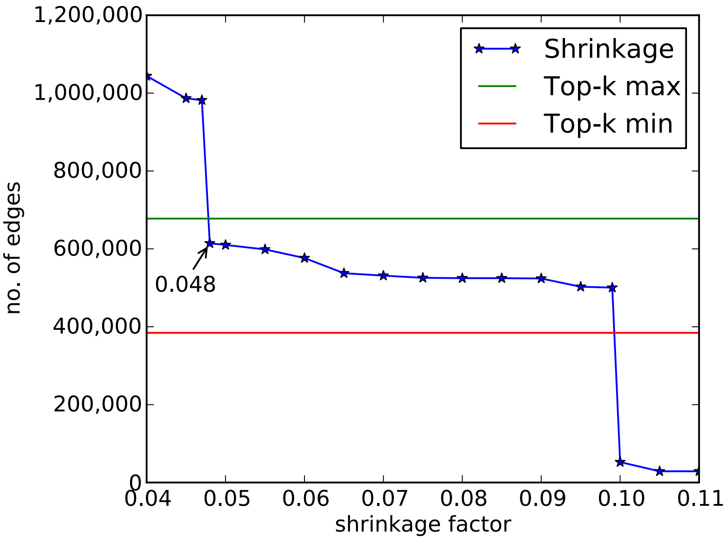
Bonus frames

Results



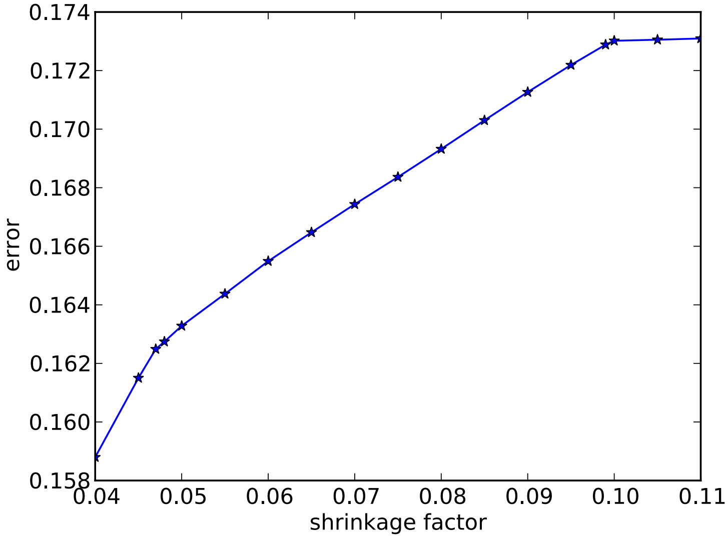
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
no. of edges
0
20
40
60
80
100
days
no. edges /5
5,681,477
EWMA (/5)
Shrinkage
Top-k min
Top-k max

Results



Shrinkage
Top-k max
Top-k min
0.048
no. of edges
shrinkage factor
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11

Results


0.174
0.172
0.170
0.168
0.166
0.164
0.162
0.160
0.158
error
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
shrinkage factor