Lossy Compression of Dynamic, Weighted Graphs

Wilko Henecka and Matthew Roughan
matthew.roughan@adelaide.edu.au

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/
Compressions

- 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

<table>
<thead>
<tr>
<th></th>
<th>lossless</th>
<th>lossy</th>
</tr>
</thead>
<tbody>
<tr>
<td>static, unweighted</td>
<td>[1, 2, 3]</td>
<td>[4, 5]</td>
</tr>
<tr>
<td>static, weighted</td>
<td>[6]</td>
<td>[7, 8]</td>
</tr>
<tr>
<td>dynamic, unweighted</td>
<td>[9]</td>
<td></td>
</tr>
<tr>
<td>dynamic, weighted</td>
<td></td>
<td>[10, 11, 12]</td>
</tr>
</tbody>
</table>
Weighted sum of weighted graphs

\[ G = \alpha A \oplus \beta B \]

means

\[ N(G) = N(A) \cup N(B), \]
\[ E(G) = E(A) \cup E(B), \]

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”
Lossy compression is approximation
- on-line algorithms combine estimation and approximation

Mathematical representation in this context

\[ \hat{G}_t = A\left( \theta \hat{G}_{t-1} \oplus (1 - \theta)g_t \right). \]

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 $\epsilon$
- Parameters $(k, \epsilon)$
  - 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

$$w_G(e) = [w_G(e) - \lambda]^+,$$

- Only one parameter $\lambda$
- Draws on ideas from de-noising
Results: errors with compression approximately 10:1
Results: compressed degree distributions

Graph Compression
August 6, 2015 12 / 14
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)


Bonus frames
Results

Graph Compression

August 6, 2015

![Graph showing the relationship between shrinkage factor and error. The x-axis represents the shrinkage factor ranging from 0.04 to 0.11, and the y-axis represents the error ranging from 0.158 to 0.174. The line shows an increasing trend as the shrinkage factor increases.]