Over-squashing and bottlenecks

# Over-squashing and bottlenecks

2022-04-29 15:51  
**Tags**:

@topping2022UNDERSTANDING

Message passing paradigm has "over-squashing" problem where messages from distant nodes have their information distorted

## Message passing:

* learnable non-linear functions diffuse info in graph
* GCN and GAT (popular GNN frameworks) are posed as flavors of this scheme and considered instances of more general geom DL framework
* Drawbacks have been formalized, like oversmoothing and limits of expressive power

## Oversquashing

* Distortion of messages from distant notes is less understood
* One proposed way to fix is reducing the "bottleneck"  
  Screen Shot 2022-04-29 at 3.59.33 PM.png
* this topological framing is still not so well understood, so we look at general message passing NNs (MPNNs)
* Defined in terms of an adjacency mat A, update function phi, and message function family psi, the l+1th hidden layer output h is  
  Screen Shot 2022-04-29 at 5.22.39 PM.png
* *Long-range dependencies* in MPNNs exist when the output depends on represenations from distant nodes interacting.
  + If they exist, they need to be propagated across the network without distortion
  + The problem is that the size of the receptive field of a node grows exponentially with layer count (r), forcing exponentially more information to be compressed in a fixed code size
* Authors propose using the Jacobian of hidden layers wrt x to formally assess over-squashing  
  Screen Shot 2022-04-29 at 5.26.18 PM.png
* If the MP function and update functions have bounded derivatives, then the propagation of messages is controlled by a suitable power of Ahat.  
  The Jacobian (RHS) that measures over-squashing is related to graph topology **via powers of the augmented normalized adjacency matrix** (I do not understand why lmao)

## Graph Curvature

They define the **Balanced Forman curvature**, a metric that captures the local curvature of a region in a graph (from the geometric notion, hyperbolic=>parallel lines pull apart=>information lost=>negative curvature on graph)  
In the neighborhood of an edge i~j:, they count the triangles containing that edge, the vertex neighbors forming a 4-cycle based at ij without diagonals inside, and the maximal number of 4-cycles based at i~j traversing a common node  
Screen Shot 2022-04-29 at 5.44.48 PM.png  
Screen Shot 2022-04-29 at 5.44.57 PM.png  
If i~j is a "bridge" between the 1-neighbors of i and j, curvature on that edge is negative, else it is positive, and the edges stay connected even if i and j are removed.

They then show that if this Ric(i,j) is lower-bounded at every point by a positive number, the curvature is positive everywhere, and the receptive field of each node will be polynomial in a hop-distance, there fore the bottleneck effect will not play a "crucial role"

They use this to show that **negatively-curved edges are the source of the over-squashing problem**, using an epsilon-delta proof to demonstrate bounds on the gradients of the transition and update functions, Ric(i,j) upper bounded by -2+delta, then there exists a Q that fixes a bound on the layerwise jacobian , which

implies that if we have a negatively curved edge as in (ii), then there exist a large number of nodes *k* such that GNNs---on average---struggle to propagate messages from *i* to *k* in two layers despite these nodes *k* being at distance 2 from *i*. In this case the over-squashing occurs as measured by the Jacobian in equation 4 and hence the propagation of information suffers.

### Connection to Cheeger Constant

They further connect their curvature idea directly to spectral graph theory by showing the Cheeger constant (spectral gap) \lambda\_1 is bounded by Ric  
Screen Shot 2022-04-29 at 7.13.09 PM.png

## Rewiring with Curvature

The authors finally propose using their technique to augment the input graph with additional network passing edges (which has been previously proposed) in a novel way---reduing bottlenecks by adding edges between nodels *k* and *l* that maximize gain in Ric(i,j) on minimal edges, and remove high Ricci curvature edges to convergence

Screen Shot 2022-04-29 at 7.18.34 PM.png

* Adds edges to "support" the negatively-curved edge, taking them away from places where they aren't needed in the message passing graph
* Compare efficacy against random-walk-based rewiring
  + Authors suggest a weakness of these methods is they mostly smooth across short diffusion distances
  + This might not help with larger structural problems in bottlenecks
  + They show this with more math I'm frankly not very interested in
  + They also demonstrate that structure is better-preserved under the bottleneck-targeting rewiring approach due to its surgical nature, something I'm more inclined to believe

### Results

Across a variety of tasks they find that *unsupervised node classification models* perform better under the graphs rewired with SDRF, particularly on low-homophily datasets, whereas more heterophilic datasets (adjacent nodes have different labels) performance is worse, as noise actually gets *injected*  
Screen Shot 2022-04-29 at 7.37.28 PM.png  
Screen Shot 2022-04-29 at 7.37.38 PM.png