Growing Better Graphs with Latent-Variable Probabilistic Graph

Xinyi Wang, Salvador Aguinaga, Tim Weninger, David Chiang

Background and Problems

- Hyperedge Replacement Grammar (HRG)
  - Generate graphs like CFG generating strings
  - Extract from the tree decomposition of a graph
  - Problem: Lack of context
- Graph Generator
  - Problem: Evaluate on training data

Solution: latent variable HRG

- Nonterminal Splitting [add CONTEXT to HRG]
  - Learning with Expectation Maximization
  - Objective: max P(training graphs)
  - Rule splits to differentiate contexts
  - Evaluation [robust and accurate]
    - Log likelihood of TEST graph

Experiments and Results

- Train/test graphs
  - Two synthesized graphs
  - Four real world graphs
  - Left: log likelihood is an effective metric
  - Right: latent variable HRG improves over HRG
  - Comparable with other graph generators in terms of GCD
  - On the test graph of similar structure with the training graph, the log likelihood is higher than the test graph of different structure
  - Log Likelihood always maximize at number of split n > 1
Watch Your Step: Learning Graph Embeddings through Attention
Sami Abu-El-Haija¹,², Bryan Perozzi², Rami Al-Rfou², Alex Alemi²
Information Sciences Institute¹

**Task: Node Embeddings**
- Goal: Learn Node Embeddings.
- Useful for various tasks (Link Prediction & Node Classification).
- Modern methods pass random walk sequences to word2vec [1], which samples context using uniform distribution:

```
<table>
<thead>
<tr>
<th>Surfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>v₁ → v₃ → v₉ → v₇ → ...</td>
</tr>
</tbody>
</table>
```

\[ C = 3 \text{ context choices} \]

**E[statistics]**
We derive analytical solution on (anchor, context) sampling:

\[
E \left[ D^{\text{DEEPWALK}[C]} \right] = \tilde{p}(0) \sum_{k=1}^{C} \left(1 - \frac{k - 1}{C} \right) (T)^k
\]

\[
E \left[ D^{\text{GloVe}[C]} \right] = \tilde{p}(0) \sum_{k=1}^{C} \frac{1}{k} (T)^k
\]

**Ours:**
We train the context distribution jointly with embeddings:

\[
E [D \mid Q₁, Q₂, \ldots, Qₖ] = \tilde{p}(0) \sum_{k=1}^{C} Qₖ (T)^k
\]

**Our Objective:** extends [2]

\[
\text{min } \beta \|q\|_2^2 + \|E[D \mid q] \circ \log \left(\sigma(L \times R^T)\right) - 1[A = 0] \circ \log \left(1 - \sigma(L \times R^T)\right)\|_1
\]

**t-SNE:** node2vec [3] VS ours

Learned Q: differs per net

**Results:** reduce errors (link prediction by 20%-40%; Node classification by up to 10%)

[1] Perozzi et al, DeepWalk, KDD’14
**t-PINE: Tensor-based Predictable and Interpretable Node Embeddings**

Saba Al-Sayouri, Ekta Gujral, Danai Koutra, Evangelos E. Papalexakis, and Sarah S. Lam

<table>
<thead>
<tr>
<th>Baselines</th>
<th>Present Gap</th>
<th>t-PINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsatisfactory accuracy</td>
<td>Accuracy</td>
<td>Better performance (Multi-view information graph)</td>
</tr>
<tr>
<td>Explicit representation learning</td>
<td>Shallow models</td>
<td>Explicit &amp; Implicit representation learning</td>
</tr>
<tr>
<td>Disjoint explicit &amp; implicit representation learning</td>
<td>Representations concatenation</td>
<td>Joint explicit &amp; implicit representation learning (CP decomposition)</td>
</tr>
<tr>
<td>Uninterpretable</td>
<td>Interpretability</td>
<td>Interpretable</td>
</tr>
</tbody>
</table>

**Tensor formation**

**Representation Learning**

SIGKDD MLG Workshop 2018 – London, United Kingdom, August 2018
Can exploiting links in relational data lead to greater accuracy in predicting elections as they unfold in real-time?

How to “bootstrap” initial predictions to provide a baseline for inference?
- Combine vote and region features

How to compute links so as to connect the regions into a useful graph?
- Leverage region-to-region correlations

How can we perform effective collective inference?
- Executes over 100x faster!
Reducing Network Incompleteness Through Online Learning

Timothy LaRock*  Timothy Sakharov*  Sahely Bhadra†  Tina Eliassi-Rad*
*Northeastern University  †IIT Palakkad

• Network data is often incomplete
• Acquiring more data can be expensive and/or hard
• Research question:
  • Given a network and limited resources to collect more data, how can we get the most bang for our buck?

Supported by NSF 1314603
What the HAK? Estimating Ranking Deviations in Incomplete Graphs

Helge Holzmann, Avishek Anand, Megha Khosla

- Graphs collected on the Web are typically incomplete
- Hypothesis: Incomplete graphs (e.g., crawls, Web archives, ...) cause deviations in random walk algorithms, such as PageRank
- Consequence: Rankings corresponding to PageRank differ from the (unavailable) complete / original graph

**RQ I:** Do incomplete real-world graphs show a deviation in their PageRank?

**RQ II:** How can we reliably measure the extent of such ranking deviations for incomplete graphs?
Most real graphs have a multi-scale structure.

We propose a novel hierarchical graph clustering algorithm.

The algorithm is agglomerative, with a distance between clusters induced by node pair sampling.
Generalized Embedding Model for Knowledge Graph Mining

◆ Contribution

a) Propose GEN, an efficient embedding learning framework for generalized KGs

b) Consider “multi-shot” information for embedding learning simultaneously
   • (Subject, Predicate) ⇒ Object
   • (Object, Predicate) ⇒ Subject
   • (Subject, Object) ⇒ Predicate

c) We show that GEN can works on graphs in different domains

◆ Task

- Learning reasonable and accurate distributed representations for knowledge graph.
- Flexible enough to adapt to variations networks
Network Signatures from Image Representation of Adjacency Matrices: Deep/Transfer Learning for Subgraph Classification
Kshiteesh Hegde, Malik Magdon-Ismail, Ram Ramanathan and Bishal Thapa
Testing Alignment of Node Attributes with Network Structure through Label Propagation

Natalie Stanley (Stanford), Marc Niethammer (UNC-CH), Peter Mucha (UNC-CH)

In this work, we developed test to measure the extent to which node attributes and network connectivity align. This relationship is reflected through an empirical p-value in a label propagation task.

Application: Single Cell Mass Cytometry

CD8, $p=0$

TNFa, $p=0$

IL4, $p=.47$

CD14, $p=.97$
Our Goal: Extend spectral clustering to the case where different kind of interactions are present.

\[ G = (\text{Work}, \text{Coauthors}, \text{Sports}, \text{Lunch}) \]

Power Mean Laplacian: 
\[ L_p = \left( \frac{1}{T} \sum_{i=1}^{T} \left( L_{\text{sym}}^{(i)} \right)^p \right)^{1/p} \]
An iterative node sampling method that

- Achieves better community diversity than state-of-the-art
- Has linear time complexity
- Is a better seeding strategy for PPR-based community detection

**Spread Sampling for Graphs: Theory and Application**

Yu Wang, Bortik Bandyopadhyay, Aniket Chakrabarti, David Sivakoff and Srinivasan Parthasarathy
Temporal Walk Based Centrality Metric for Graph Streams

Temporal Katz Centrality

- Centrality measure for dynamic graphs
- Online updateable from the edge stream
- $\varphi$ is an arbitrary time decay function

$$r_u(t) := \sum_v \sum_{\text{temporal paths } z \text{ from } v \text{ to } u} \Phi(z, t)$$

$$\Phi(z, t) = \varphi(t_2 - t_1) \cdot ... \cdot \varphi(t_j - t_{j-1}) \cdot \varphi(t - t_j)$$

Supervised Evaluation

- Roland-Garros, USOpen 2017 Twitter data
- Daily tennis players are considered relevant
- Predict relevant nodes of the mention network with graph centrality

$t_1 < t_2 < ... < t_j \leq t$
A Method for Learning Representations of Signed Networks

- Signed networks comprise +ve and -ve edges.
- Representation learning useful for downstream tasks.
- Methods for unsigned networks don’t work well for signed networks.
- Present a method for learning representations using maximum likelihood estimation
- Opposing communities separated in representation space
Model the problem as thresholded tropical matrix factorization.

Solve using stochastic gradient descent.
Traffic prediction at individual level is hard

Can we solve the problem at aggregate level instead?

Key questions:
- How do we represent traffic congestion for a region?
- Which inputs help in predict traffic congestion?
- Can we use the underlying road network graph?
- Can we use prior knowledge of choices made by individuals?
- Can we identify the likely cause of future congestion?
Temporal Graph Generation Based on a Distribution of Temporal Motifs
Sumit Purohit, Lawrence Holder, George Chin

Motivations
• Generate High fidelity synthetic temporal graph
• Privacy Preservation
• Benchmarking

Approach
• Non-overlapping temporal motif
• Generate distribution of temporal motifs
• Up to 3-edges, 3-vertices motifs
• No self-loop, Non-overlapping
• Model motif formation time
• Distributed algorithms using:
  • Apache Spark, GraphFrame, Python

Result

Next Step
• Scalability Analysis
• Define Temporal Metrics to measure fidelity
• Deep Autoregressive models to generate graphs
• Code availability
  • Generator code: https://github.com/lbholder/graphstream-generator
Relevance Measurements in Online Signed Social Networks

Tyler Derr¹, Chenxing Wang¹, Suhang Wang², and Jiliang Tang¹

Recently accepted papers on signed network modeling and applications!

Please see my homepage for details!

Thank you to the following:

1: Data Science and Engineering Lab, Michigan State University
2: Data Mining and Machine Learning Lab, Arizona State University

KDD2018

Aug. 20, 2018
A Marketing Game:  
a rigorous model for strategic resource allocation  

Matthew G. Reyes

Features / Contributions:

- stochastic choice updates rather than best-response
- including marketers in the model
- optimize allocation based on expected market share
GeniePath adaptively selects “neighbors” to aggregate

Existing Graph Neural Networks (GNN) study how to aggregate “neighbors” but ignore which “neighbor” to aggregate.

Why GeniePath?
1. Graph is noisy.
2. Different nodes, different roles.
3. Performance & interpretability

GeniePath, learns to explore the breadth and depth of neighborhood.