TenGAN: A Tensor-Factorization-based GAN for Multi-view Graph Generation

William Shiao [1] Benjamin A. Miller [2] Kevin Chan [3] Paul Yu [3] Tina Eliassi-Rad [2] Evangelos E. Papalexakis [1]

[1]University of California, Riverside [2]Northeastern University [3]U.S. Army Research Laboratory

Problem

Given a set of multi-view graphs G, **generate** a set of graphs G' that are not identical to G, but possess similar graph attributes to and are indistinguishable (to a classifier) from them.

There are three major challenges we have to address:

- *Sampling from multi-view graphs*: Most datasets consist of a single graph with multiple views. However, since we are emulating a set of multi-view graphs, we need many smaller multi-view graphs to form a distribution. Therefore, we need a method that samples smaller sub-multi-view-graphs from a larger multi-view graph.
- *Inadequate Evaluation Criteria*: Before we can decide on an appropriate method to accomplish the aforementioned sampling task, we must determine our evaluation criteria. This is challenging because we need to consider not only the graph attributes of each view but also the relationships between each view. This means that many of the graph evaluation metrics commonly used in graph generation are insufficient for multi-view graph generation.
- *Large Number of Parameters*: If we naively attempt to generate a multi-view graph, the number of parameters required will explode. This is because we need $\mathcal{O}(k \times n^2)$ parameters to generate an adjacency tensor for a multi-view graph with k views and n nodes.

Proposed Method

Sampling: Many generative models require multiple input samples, rather than a single example. We perform random-walk sampling across each view. We then use the induced sub-graph on the remainder of the views.

However, one issue with many large multi-view datasets is that most of the nodes may be disconnected in any given view. In extreme examples, like in some knowledge base graphs, almost all nodes will be disconnected in each view. Oftentimes, even taking the union of all edges across all views will still result in a disconnected graph. In order to produce better quality samples, we use the BiasedSample method of ParCube [4].

Model: We chose to use a Generative Adversarial Network (GAN) due to its proven effectiveness for image and graph generation tasks.

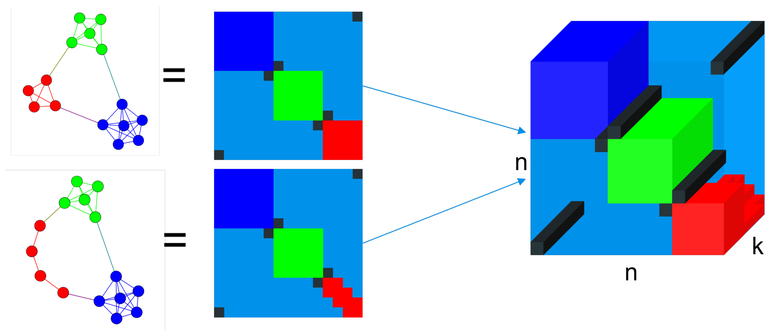


Figure 1. A visual representation of an adjacency tensor

Parameter Complexity: If we attempted to generate an adjacency tensor for a multi-view graph with n nodes and k views directly, we would have to use $\mathcal{O}(kn^2)$ parameters in the final layer. However, if we generate the CPD factors first, we only need $\mathcal{O}(r(n+k))$ parameters in the final layer, where r is a hyperparameter that increases the quality of the fit at the cost of more parameters. This offers significant savings for $r \ll n^2$, and we show that our models work well for this case.

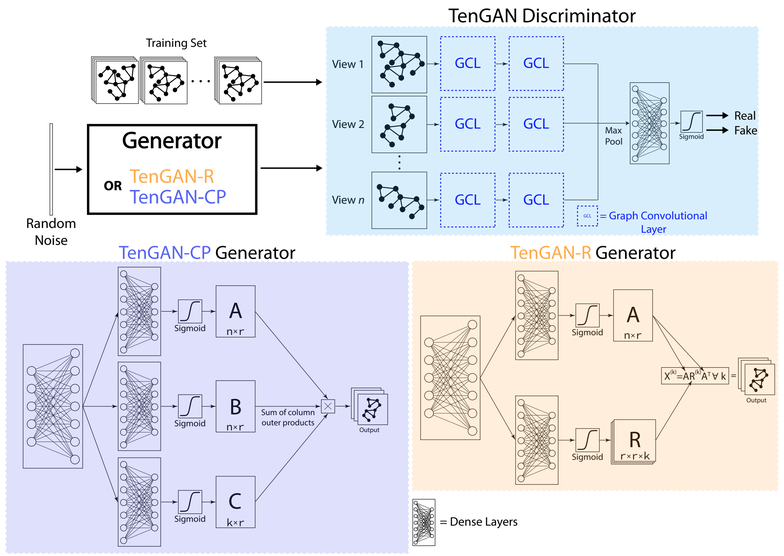
For the RESCAL-based formulation, we need $\mathcal{O}(nr+kr^2)$ parameters in the final layer. In this case, we save parameters in the case where $r < n$.

Architecture

The TenGAN-CP architecture uses a shared feature extractor layer and splits into separate networks, each of which generates a different factor in the CPD. This can be considered a higher-order extension of the BRGAN-B [5] architecture and uses the tensor CPD instead of the matrix SVD.

After generating the factors, we calculate the sum of the outer products of vectors from our factor matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$: $\sum_{i=1}^{r} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$. This helps us reduce the number of parameters needed to generate a given multi-view graph.

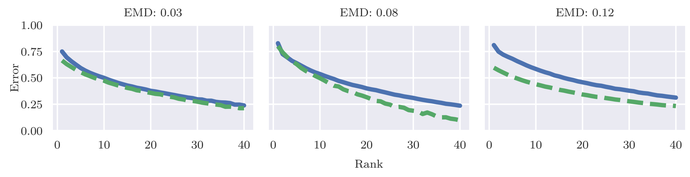
We also propose the TenGAN-R architecture, which is initially similar to the TenGAN-CP architecture, with the distinction that we use the RESCAL decomposition instead of the CPD. This results in more parameters for the same value of r, but performs better on certain datasets.



Evaluation

We propose the following methods to evaluation the quality of the generated graphs:

- *MMD-Based Evaluation*: We measure the Mean MMD (M-MMD) score between the distributions of different graph attributes. More concretely, for each graph attribute, we take the mean of the MMD between the i-th view of a generated graph G' and a graph G. We can use the clustering coefficient, degree distribution, and the orbit of the graphs similar to You *et al.* [7].
- *Classifier-Based Evaluation*: We train a classifier on generated and original data; then check to see if it correctly predicts the origin of an example. We calculate accuracy and F1 score of the resulting model (the closer to 0.5 or 50%, the better).
- *Tensor-Based Evaluation*: We calculate the sum of the Wasserstein metric between all n^2 pairs. We then normalize the sum of generated-real distances by the sum of pairwise real-real distances. The lower this score, the more similar pairs are (on average).



$$\text{TenScore} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \text{EMD}(\mathbf{E}_i, \mathbf{E'}_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \text{EMD}(\mathbf{E}_i, \mathbf{E}_j)} \quad (1)$$

Datasets

1. **Football** [2]: 248 English Premier League football players and clubs on Twitter, where each of the 6 views corresponds to a different interaction between the accounts (follows, followed-by, mentions, mentioned-by, retweets, retweeted-by). Note that 3 of the views are essentially transposes of the other 3.
2. **NELL-2** [1]: A sampled version of the NELL-2 dataset (from [6]) that consists of (*entity*, *relation*, *entity*) tuples. The original size is 12,092 x 9,184 x 28,818, but we resample it to a 1,000 x 4 x 1,000 tensor (where 4 is the number of views) for the purpose of evaluation.
3. **Comm**: An enterprise communication network dataset of 1,558,594 computers. Each view corresponds to communications between nodes on one of five ports (22, 23, 80, 443, and 445), with one view for each port. In the view associated with port p, a directed edge from u to v exists if u initiates a connection to v over port p.
4. **Enron** [3]: A multi-view graph of emails sent between Enron employees, where each view represents a two-month (60-day) time interval and edges represent emails. The original tensor [6] is 6,066 senders x 5,699 recipients x 244,268 words x 1,176 days. We collapse the words dimension and simply add an unweighted edge for each email sent in a given time interval. We also aggregate the slices so that each view represents a 60-day period to reduce the number of views. Finally, we sample 1,000 senders and 1,000 recipients.

Results

	Football						NELL-2					
Model Name	Deg	Clust	Orbit	F1	Acc	TenScore	Deg	Clust	Orbit	F1	Acc	TenScore
TenGAN-CP	**0.10**	**0.45**	**0.10**	**0.57**	**0.70**	0.92	**0.48**	**0.70**	**0.31**	0.94	0.94	**0.89**
TenGAN-R	1.09	1.12	0.75	0.94	0.94	**0.77**	0.93	0.77	0.91	**0.49**	**0.66**	2.64
HGEN	1.03	1.45	0.84	1.00	1.00	5.00	0.98	0.74	0.65	1.00	1.00	5.67

	Enron						Comm					
Model Name	Deg	Clust	Orbit	F1	Acc	TenScore	Deg	Clust	Orbit	F1	Acc	TenScore
TenGAN-CP	1.34	1.00	1.19	**0.87**	**0.89**	0.86	0.37	0.39	0.38	0.72	0.61	1.50
TenGAN-R	1.77	1.97	1.74	0.99	0.99	4.32	**0.34**	**0.40**	**0.35**	**0.69**	**0.56**	**1.31**
HGEN	**0.59**	**0.02**	**0.06**	1.00	1.00	**0.90**	-	-	-	-	-	-

References

[1] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1306–1313, Atlanta, Georgia, 2010. AAAI Press.

[2] Derek Greene and Pádraig Cunningham. Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 118–121, New York, NY, USA, 2013. Association for Computing Machinery.

[3] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, ECML'04, page 217–226, Berlin, Heidelberg, 2004. Springer-Verlag.

[4] Evangelos E. Papalexakis, Christos Faloutsos, and Nicholas D. Sidiropoulos. Parcube: Sparse parallelizable tensor decompositions. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 521–536, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[5] William Shiao and Evangelos E. Papalexakis. Adversarially generating rank-constrained graphs. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–8, Porto, Portugal, 2021. IEEE.

[6] Shaden Smith, Jee W. Choi, Jiajia Li, Richard Vuduc, Jongsoo Park, Xing Liu, and George Karypis. FROSTT: The formidable repository of open sparse tensors and tools, 2017.

[7] Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5694–5703, Stockholm, Sweden, 2018. PMLR.

Acknowledgements

We would like to thank Ananthram Swami for his valuable feedback and discussions. Research was supported by the National Science Foundation under CAREER grant no. IIS 2046086. This research was also sponsored by the Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). BAM and TER were also sponsored in part by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory, the United States Air Force, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here on.

wshia002@ucr.edu