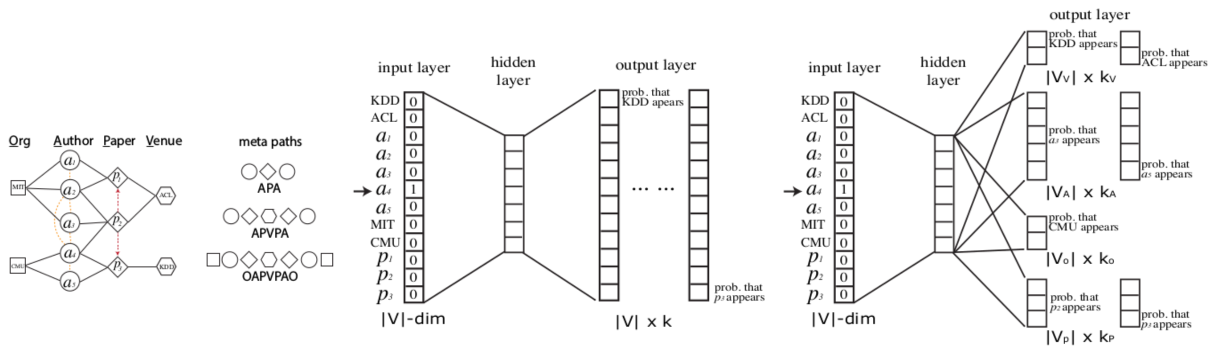
# Paper Summary

David Miller
CIS 5930: Social Network Mining

February 5, 2018



(a) An academic network (b) Skip-gram in *metapath2vec*, node2vec, & DeepWalk (c) Skip-gram in *metapath2vec++*

Figure 1

A Heterogeneous Network is defined as a graph $G = (V, E, T)$ in which each node $v$ and each link $e$ are associated with their mapping functions $\phi(v) : V \to T_V$ and $\phi(e) : E \to T_E$ , respectively. $T_V$ and $T_E$ denote the sets of object and relation types, where $|T_V| + |T_E| > 2$ [1]. Given this definition, we can state the problem as follows:

**Problem.** Given a heterogeneous network $G$, the task is to learn the $d$-dimensional latent representations $X \in \mathbb{R}^{|V|d}, d << |V|$ that are able to capture the structural and semantic relations among them

What the problem is essentially stating given some input $I$ we try to find the best (lowest) dimension $d$ that accurately represent and captures the relationship amongst elements in $I$.

The problem generates similar neighborhoods based on random meta-paths $\mathcal{P} : V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \ldots V_t \xrightarrow{R_t} V_{t+1} \ldots \xrightarrow{R_{l-1}} V_l$, where $R_i$ is the relationship between nod $V_i$ and $V_{i+1}$ and where probability of transition is given by

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) = t + 1) \\ 0 & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) \neq t + 1 \\ 0 & \notin E \end{cases}$$

which essentially states that the walker is likely to go to a node within its neighborhood. Once all nodes on the neighborhood have been visited, there is no place for the random walker to go and thus this creates a neighborhood of some shared context. This is essentially what *metapath2vec* does. What *metapath2vec++* does it just add probability distributions to context types $c_1, \ldots, c_t$ to increases neighborhood classification. Future work is discussed very well in the paper. It includes different improvements and optimizations that can be done: optimizing sampling space, machine learning to uncover meaningful meta-paths, allow the model to work with dynamic data, and generalizing the model for all types of heterogeneous networks.

Three strengths I found with the paper are

1. *metapath2vec* and *metapath2vec++* models are efficient and scalable for large-scale heterogeneous networks with millions of nodes [1].

2. The area of application lends itself well to data visualization, allowing for easy communication with the public and possible commercial consumers if sold as a product.

3. Parameter sensitivity allows contexts to be weighted respective to their significance.

Three weaknesses I found with the paper are

1. A lack of applications; the paper (results including) focused too much on the author and venue networks.

2. There was no evidence suggesting that this can be applied to GPU computing, which I believe to be an appropriate computing venue for *metapath2vec* and *metapath2vec*.

3. Lack of analysis. Topics such as statistics and probability deserve a thorough analysis to provide evidence that algorithms based on them are valid and reliable. There is a lack of this sort of analysis in the paper.

Questions for the reader

1. What exactly does *metapath2vec++* offer over *metapath2vec*?

2. Is there a way to embed the data in one dimension while keeping context of data?

3. If machine learning is used for metapath finding, how would that work?

# References

[1] Yuxiao Dong, Nitesh V. Chawla, Ananthram Swami, *metapath2vec: Scalable Representation Learning for Heterogeneous Networks*, KDD17, August 13-17, 2017, Halifax, NS, Canada.