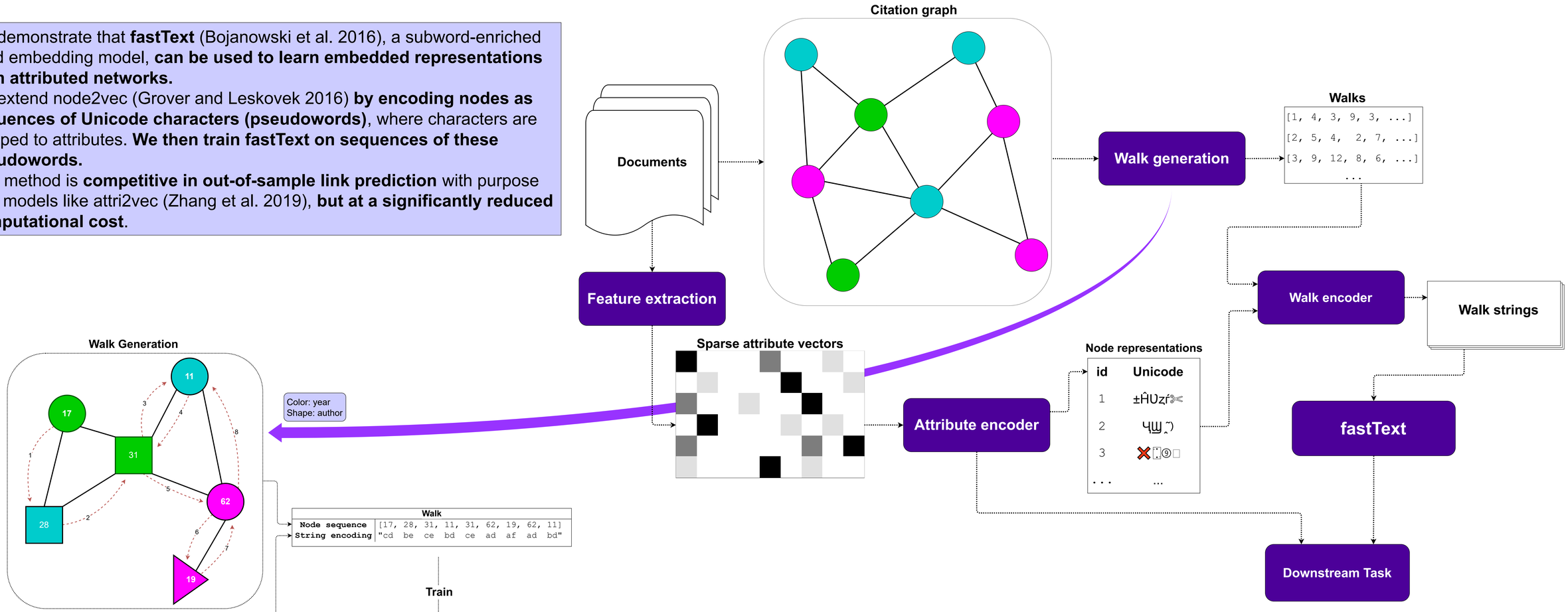
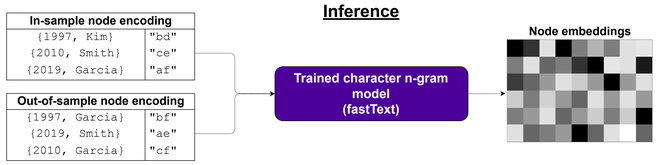
An embarrassingly simple method for attributed network embedding

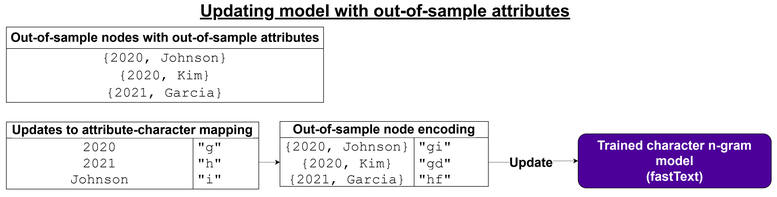
Jacob A. Matthews, Imane Terhmina, Laurent Dubreuil and Marten van Schijndel

jam963@cornell.edu

- We demonstrate that **fastText** (Bojanowski et al. 2016), a subword-enriched word embedding model, **can be used to learn embedded representations from attributed networks.**
- We extend node2vec (Grover and Leskovek 2016) **by encoding nodes as sequences of Unicode characters (pseudowords)**, where characters are mapped to attributes. **We then train fastText on sequences of these pseudowords.**
- This method is **competitive in out-of-sample link prediction** with purpose built models like attri2vec (Zhang et al. 2019), **but at a significantly reduced computational cost**.



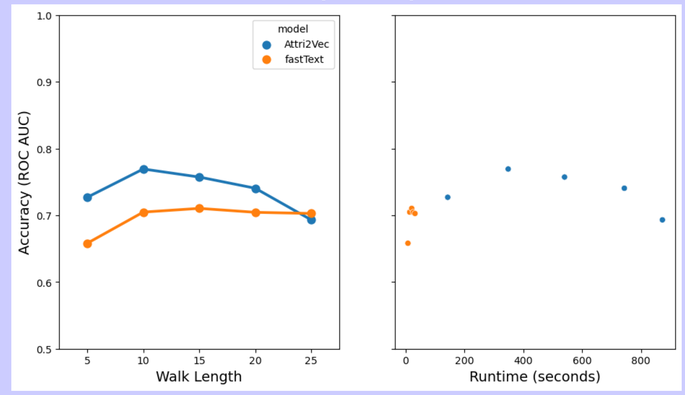




Advantages

- **Capable of representing any out-of-sample node** while retaining the relative simplicity of node2vec.
- **Can easily be updated with out-of-sample attributes** (e.g. new publication year, author, or keyword). Not possible with other methods.
- **Interpretable**: vector representations of single attributes can be compared to full node representations with straightforward vector arithmetic.
- **Does not require familiarity with low-level ML libraries**

Out-of-sample link prediction



- **Our approach is competitive with attri2vec** (Zhang et al. 2019) **in out-of-sample link prediction** when tested on the DBLP dataset (18k nodes, 45k edges) (left).
- However, **our approach is many times faster than attri2vec** at all walk lengths (right).