The Networks of Science. Data-driven Understanding of Scientific Production

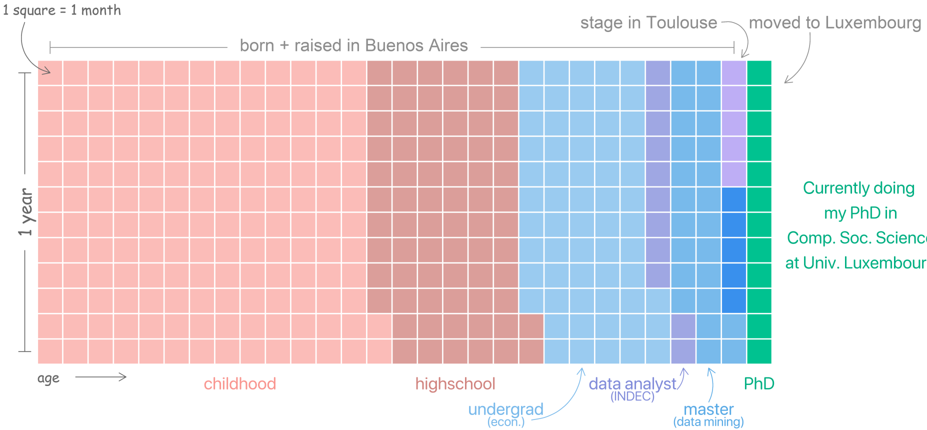
Diego Kozlowski

July-2020

DTU DRIVEN

uni.lu
UNIVERSITÉ DU LUXEMBOURG

about me



Currently doing
my PhD in
Comp. Soc. Science
at Univ. Luxembourg

Objectives

My PhD (cumulative) has three main projects:

1. Analysis of the *Science of Science* field using the citation network, metadata and text.
2. *Heterogeneous networks in science*: Combining types of entities (authors, articles) and relations (authorship, collaboration, references),
3. extend the concepts of *Globalized Science* and *Knowledge Economy* to the different roles countries play in the international production of science, and its relation with the role these play in global economy as a whole.

Methodology: Graph Neural Networks

For this objectives, I will use **Graph Neural Networks**(GNN):

- GNN extend the idea of convolutions in Deep Learning to non-euclidean spaces, like graphs.
- This models build a **dense-vector representation of a node, *embedding*, as an aggregation of the information of its neighbors**. The model leverages on both the graph structure and the nodes (and maybe the edges) attributes.
- This is a booming topic in the Deep Learning community, and surprisingly some of the most standardized benchmarks are on the task of node prediction in citation networks.

1. Science of science analysis

Data: Journal-based: using scopus API, I built a dataset with 15000 articles. The list of journals was based on the ISSI recommendations, extended with journals from social sciences.

- Journal of informetrics (900)
- Research Policy (3200)
- Science, Technology, & Human Values (750)
- Research Evaluation (650)
- Scientometrics (5000)
- Social Studies of Science (1050)
- Science and Public Policy (1700)
- Minerva (400)
- Science and Technology Studies (100)
- Public Understanding of Science (1000)

Doubt: Is this a comprehensive list? should I add journals that include also unrelated topics (e.g. Information Processing & Managment)?

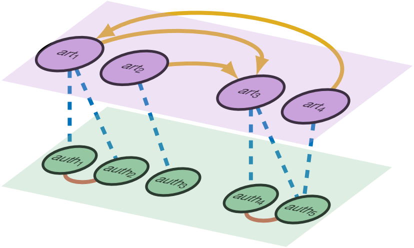
1. Science of science analysis

As a social-science-motivated computational-science-oriented project, this project has several goals:

1. work with a clear objective benchmark (link prediction) to compare the different possible approaches,
2. use the node embeddings to develop new insights of the field of Sciences of Sciences,
   - Do articles cluster by subfield, journal, year and/or country?
   - How are articles positioned in the embedding according to their collaboration status?
3. build a recommendation system with the link prediction model that also considers other factors, like semantic similarity and impact.

2. Multi-type network embedding

- Collaboration networks and citation networks represent different information,. so does textual information and metadata.
- the idea of this project is to check weather combining this different networks is useful for a more comprehensive understanding.
- Although working directly on heterogeneous networks might not give good results, GNN might be able to encode all this information in *embeddings*, easy to handle.



2. Multi-type network embedding

In order to see the potential benefits, we need an objective benchmark to compare this with simple networks.

- One option is **author-disambiguation**:
    - Gold standard: author-curated information of their own publications from Google Scholar, ORCID or universities websites.
- **doubt**: Is it a good task? In GNN the traditional benchmark is the subject of the article (inferred from journal), would that be a more interesting benchmark? or link prediction?

Data: Web Of Science raw data. Germany 2011-2019. From the *Q-Know project*.

3. International Division of Science.

- This project will be descriptive and it will focus on a characterization of global science:
  - Funding distribution along countries, and the internal distribution within fields.
  - Collaboration patterns of institutions, with focus on international collaboration.
  - Specialization Index across countries,
  - Relation of the economic role of countries in global economy and the specificity of their scientific production.
- **data:**
  - Web Of Science, only authors, institutions and publications (no references or text). Global coverage. 1900-2011, as part of the *Q-Know project*.
  - Input data from OCDE, Word Bank, etc.

Acknowledgement

The Doctoral Training Unit **Data-driven computational modelling and applications** (DRIVEN) is funded by the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781).

`https://driven.uni.lu`

DTU DRIVEN

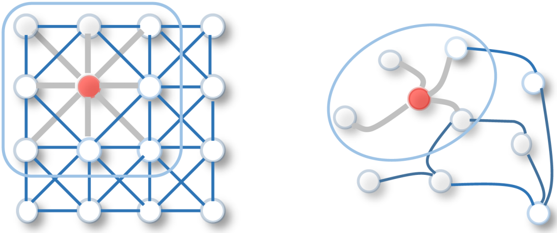
Fonds National de la Recherche Luxembourg

bibliography

[1] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* XX.Xx (2020), pp. 1–21. ISSN: 2162-237X. DOI: 10.1109/tnnls.2020.2978386. arXiv: 1901.00596.

[2] William L. Hamilton, Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs". In: *Advances in Neural Information Processing Systems* 2017-Decem.Nips (2017), pp. 1025–1035. ISSN: 10495258. arXiv: 1706.02216.

Apendix. Deep Learning on Graphs



(a) 2D Convolution. Analogous to a graph, each pixel in an image is taken as a node where neighbors are determined by the filter size. The 2D convolution takes the weighted average of pixel values of the red node along with its neighbors. The neighbors of a node are ordered and have a fixed size.

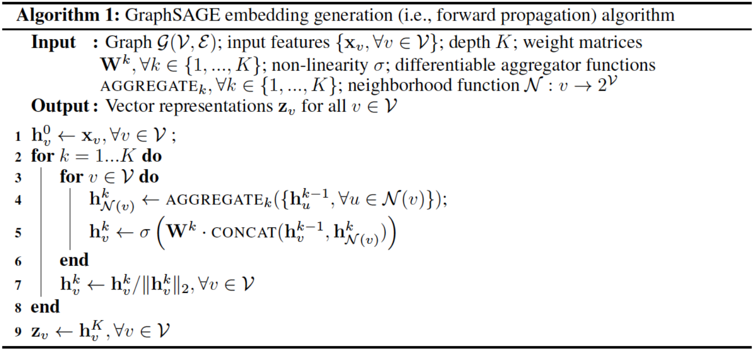
(b) Graph Convolution. To get a hidden representation of the red node, one simple solution of the graph convolutional operation is to take the average value of the node features of the red node along with its neighbors. Different from image data, the neighbors of a node are unordered and variable in size.

Fig. 1: 2D Convolution vs. Graph Convolution.

from Wu et al. 2020

Apendix. Deep Learning on Graphs

One possible implementation: GraphSAGE:



Algorithm 1: GraphSAGE embedding generation (i.e., forward propagation) algorithm

Input : Graph $\mathcal{G}(\mathcal{V},\mathcal{E})$; input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$; depth K; weight matrices $\mathbf{W}^k, \forall k \in \{1,...,K\}$; non-linearity σ; differentiable aggregator functions $\text{AGGREGATE}_k, \forall k \in \{1,...,K\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

Output : Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

```
$\mathbf{h}^0_v \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$ ;
for $k = 1...K$ do
    for $v \in \mathcal{V}$ do
        $\mathbf{h}^k_{\mathcal{N}(v)} \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}^{k-1}_u, \forall u \in \mathcal{N}(v)\})$;
        $\mathbf{h}^k_v \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}^{k-1}_v, \mathbf{h}^k_{\mathcal{N}(v)})\right)$
    end
    $\mathbf{h}^k_v \leftarrow \mathbf{h}^k_v / \|\mathbf{h}^k_v\|_2, \forall v \in \mathcal{V}$
end
$\mathbf{z}_v \leftarrow \mathbf{h}^K_v, \forall v \in \mathcal{V}$
```

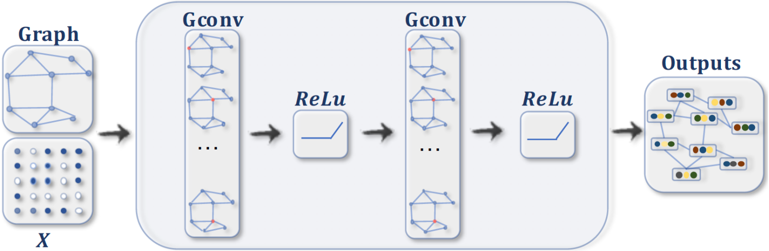
from Hamilton, Ying, and Leskovec 2017

One possibility would be the max aggregator

$$max(\{\sigma(W_{pool}\{h^k_{u_i} + b), \forall u_i \in \mathcal{N}(v)\})$$

Apendix. Deep Learning on Graphs

We can stack multiple layers and combine the outputs. The GNN receives the network structure and a matrix of features and generates a node embedding (or graph embedding) that can be used in different tasks.



from Wu et al. 2020