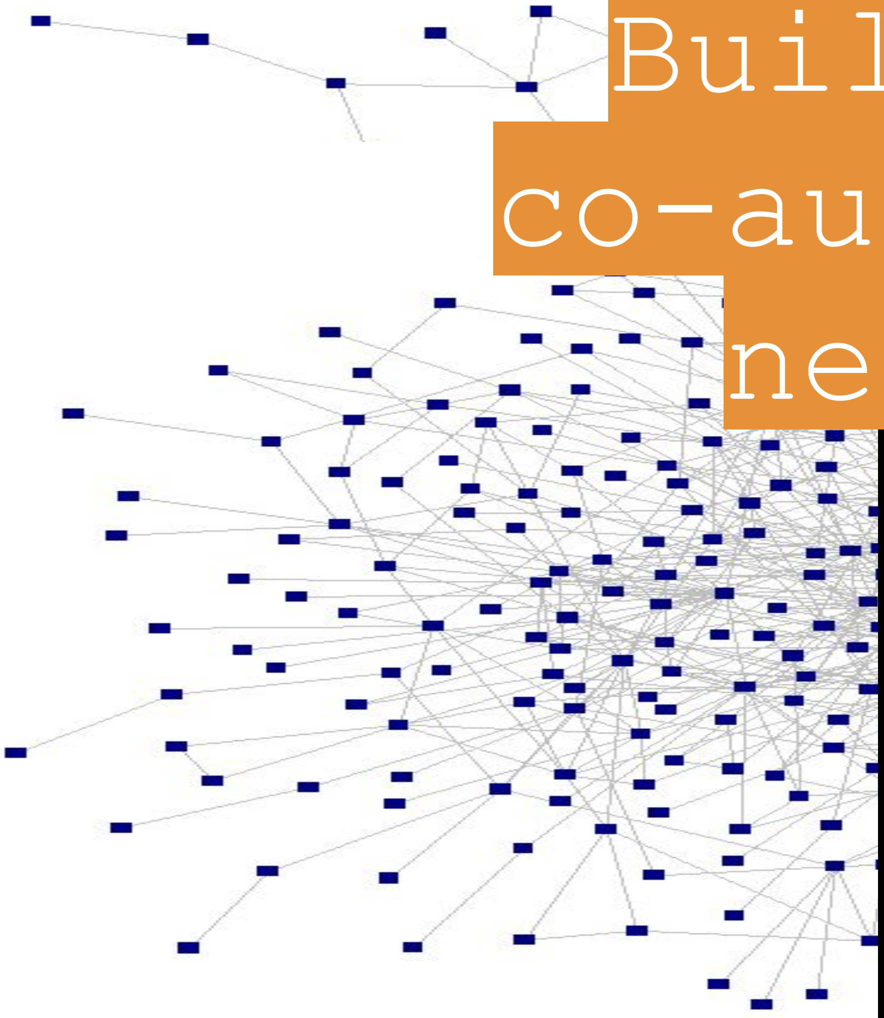


Building a co-authorship network

Project made by
Vladimir Vlasov
& Anastasia Khorosheva
HSE 2019

_strategy

Step 0: transform data into graph embeddings

Step1: build a co-authorship graph

Step2: train NN to predict future possible collaborations between authors

_dataset

arxivData.json

45000-line dict containing info about papers


```
{
    "author": "[{'name': 'Kenji Kawaguchi'}, {'name': 'Leslie Pack Kaelbling'}, {'name': 'Yoshua Bengio'}]",
    "day": 16,
    "id": "1710.05468v3",
    "link": "[{'rel': 'alternate', 'href': 'http://arxiv.org/abs/1710.05468v3', 'type': 'text/html'}, {'rel':
        'related', 'href': 'http://arxiv.org/pdf/1710.05468v3', 'type': 'application/pdf', 'title': 'pdf'}]",
    "month": 10,
    "summary": "With a direct analysis of neural networks, this paper presents a\nmathematically tight
        generalization theory to partially address an open problem\nregarding the generalization of deep
        learning. Unlike previous bound-based\ntheory, our main theory is quantitatively as tight as possible
        for every\ndataset individually, while producing qualitative insights competitively. Our\nresults give
        insight into why and how deep learning can generalize well,\ndespite its large capacity, complexity,
        possible algorithmic instability,\nnonrobustness, and sharp minima, answering to an open question in the
        \nliterature. We also discuss limitations of our results and propose additional\nopen problems.",
    "tag": "[{'term': 'stat.ML', 'scheme': 'http://arxiv.org/schemas/atom', 'label': None}, {'term': 'cs.AI',
        'scheme': 'http://arxiv.org/schemas/atom', 'label': None}, {'term': 'cs.LG', 'scheme': 'http://arxiv.org
        /schemas/atom', 'label': None}, {'term': 'cs.NE', 'scheme': 'http://arxiv.org/schemas/atom', 'label':
        None}]",
    "title": "Generalization in Deep Learning",
    "year": 2017
},
```

_approach

- apply node2vec to data
- feed embeddings to NN
  - 3 hidden layers
  - activation: relu, tanh
  - loss: categorial cross-entropy
- LogReg to predict future co-authorship

metrics&results

_allocation_index

```
nx.resource_allocation_index
              precision    recall  f1-score   support
           0       0.83      1.00      0.91     69435
           1       0.76      0.00      0.00     13887

   micro avg       0.83      0.83      0.83     83322
   macro avg       0.80      0.50      0.46     83322
weighted avg       0.82      0.83      0.76     83322
[[69429     6]
 [13868    19]]
0.5990594303456509
```

_jaccard_coefficient

```
nx.jaccard_coefficient
              precision    recall  f1-score   support

           0       0.83      1.00      0.91     69435
           1       1.00      0.00      0.01     13887

   micro avg       0.83      0.83      0.83     83322
   macro avg       0.92      0.50      0.46     83322
weighted avg       0.86      0.83      0.76     83322
[[69435     0]
 [13832    55]]
0.5990585638635837
```

_LogReg

```
1*1 Logreg
              precision    recall  f1-score   support

           0       0.86      0.99      0.92     69353
           1       0.83      0.19      0.31     13887

   micro avg       0.86      0.86      0.86     83240
   macro avg       0.84      0.59      0.62     83240
weighted avg       0.85      0.86      0.82     83240
[[68812   541]
 [11241  2646]]
0.6395078283412827
```

_all_pairs

```
5*1 All_pairs
              precision    recall  f1-score   support

           0       0.92      0.98      0.95    135084
           1       0.84      0.59      0.69     27216

   micro avg       0.91      0.91      0.91    162300
   macro avg       0.88      0.78      0.82    162300
weighted avg       0.91      0.91      0.91    162300
[[132044   3040]
 [ 11130  16086]]
```

_new_pairs

```
5*1 New pairs
              precision    recall  f1-score   support

           0       0.87      0.98      0.92     69367
           1       0.68      0.24      0.35     13887

   micro avg       0.85      0.85      0.85     83254
   macro avg       0.77      0.61      0.64     83254
weighted avg       0.83      0.85      0.82     83254
[[67779  1588]
 [10569  3318]]
0.6553634735001194
```

_new_pairs ver2.0

```
1*1 New pairs
              precision    recall  f1-score   support

           0       0.87      0.95      0.91     69353
           1       0.55      0.29      0.38     13887

   micro avg       0.84      0.84      0.84     83240
   macro avg       0.71      0.62      0.65     83240
weighted avg       0.82      0.84      0.82     83240
[[66055  3298]
 [ 9823  4064]]
0.6657878721401573
```

_new_pairs ver3.0

```
1*1 New pairs Stack more layers 40e
              precision    recall  f1-score   support

           0       0.87      0.96      0.91     69341
           1       0.59      0.30      0.39     13887

   micro avg       0.85      0.85      0.85     83228
   macro avg       0.73      0.63      0.65     83228
weighted avg       0.82      0.85      0.83     83228
[[66511  2830]
 [ 9789  4098]]
0.7027983388267736
```

_new_pairs ver3.0 -updated

```
1*1 New pairs Stack MORE layers 40e
              precision    recall  f1-score   support

           0       0.87      0.97      0.92     69341
           1       0.64      0.27      0.38     13887

   micro avg       0.85      0.85      0.85     83228
   macro avg       0.75      0.62      0.65     83228
weighted avg       0.83      0.85      0.83     83228
[[67246  2095]
 [10151  3736]]
0.6888330186522706
```

_further_steps

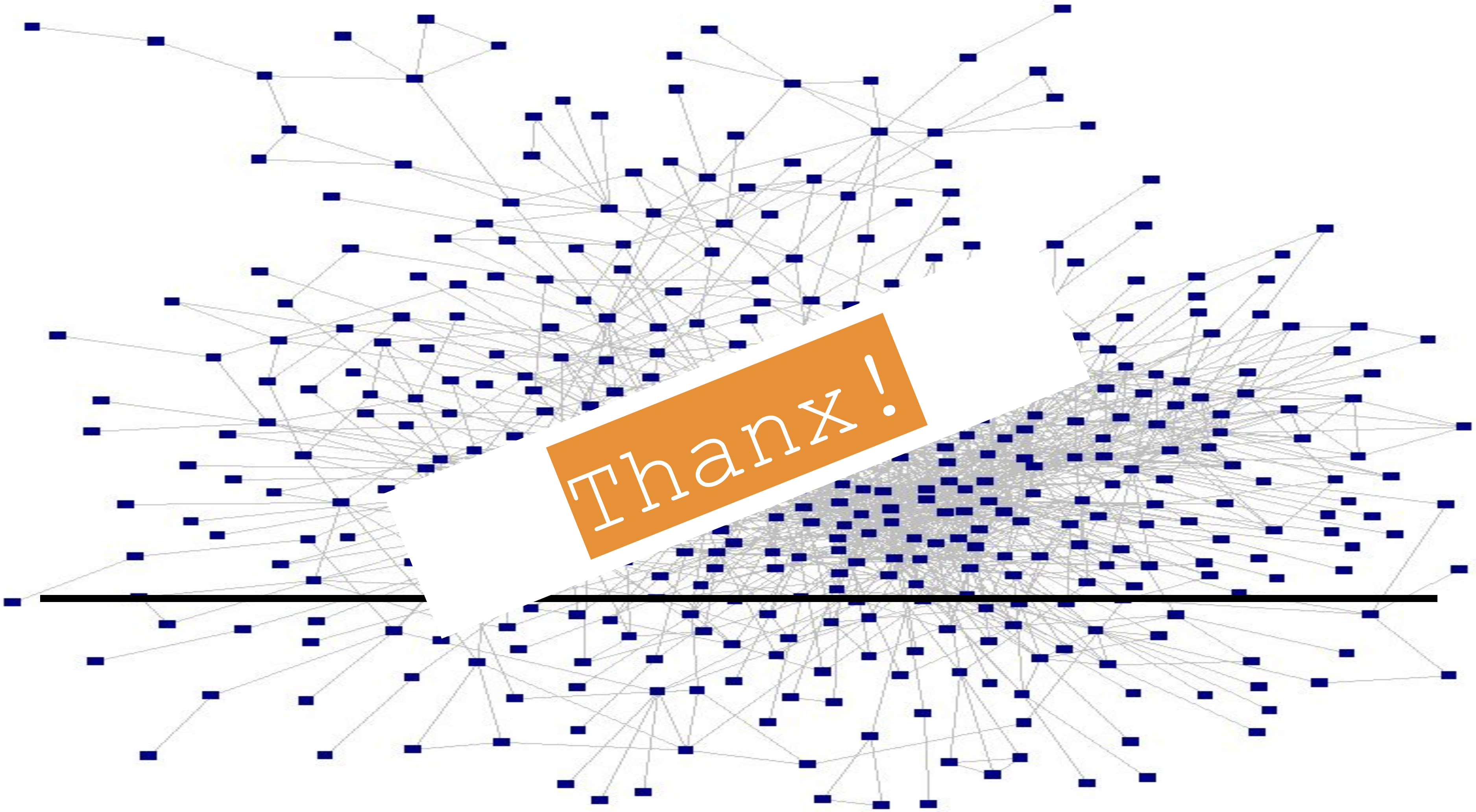
- no rush while learning (10+ random walks)
- add info about the authors (uni/degree/field/etc) and feed to NN as features
- add info about the papers
- make more interesting graphs

_conclusion

Our 1st attempt to build a network of co-authors for scientific papers

Not only build existing relations but also predict possible collaborations

There's a place for further improvement


Thanx!