

# NODE2VEC: EMBEDDING FOR GRAPH DATA

Kien Trinh

AVM Team - Weekly Meeting

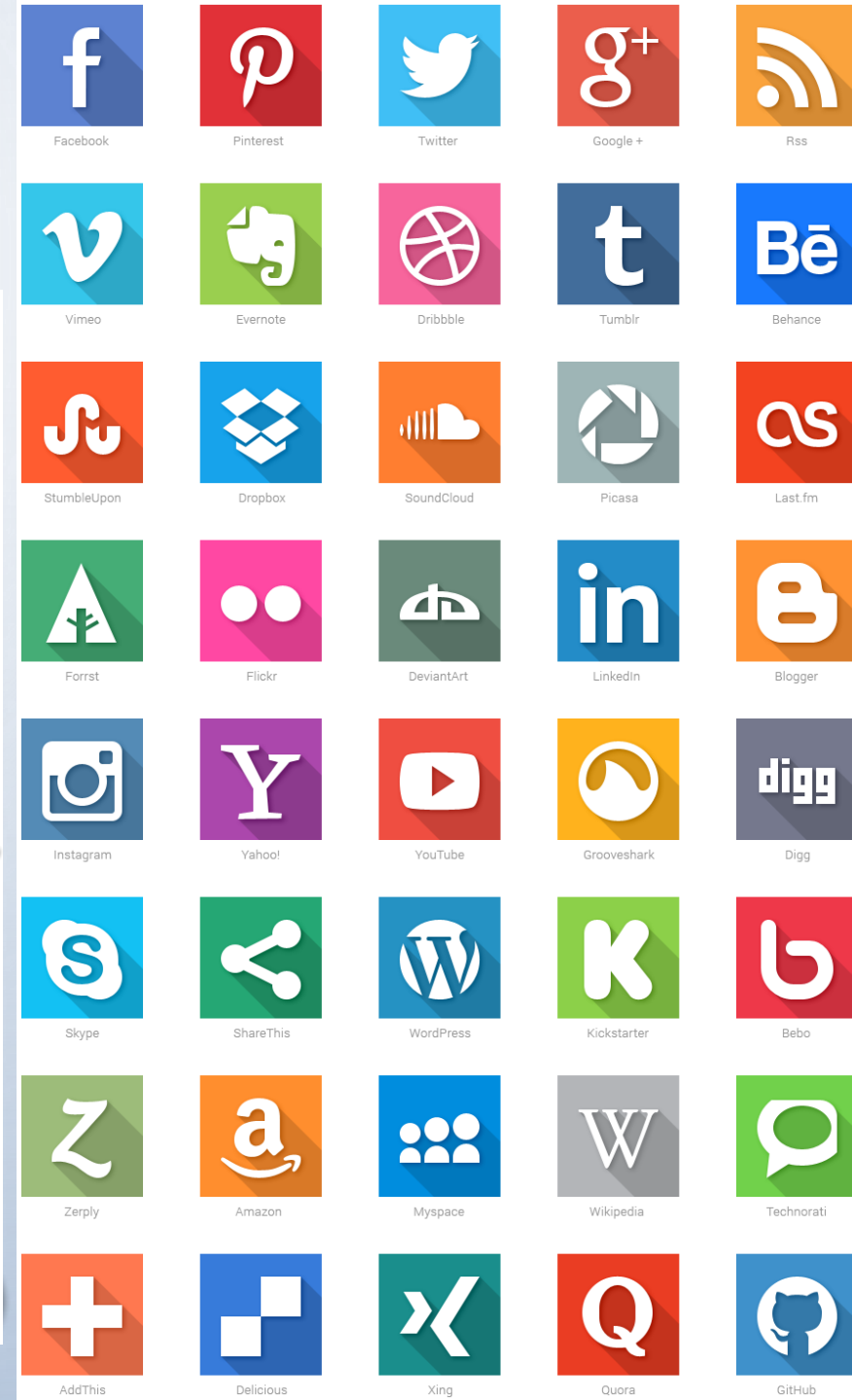
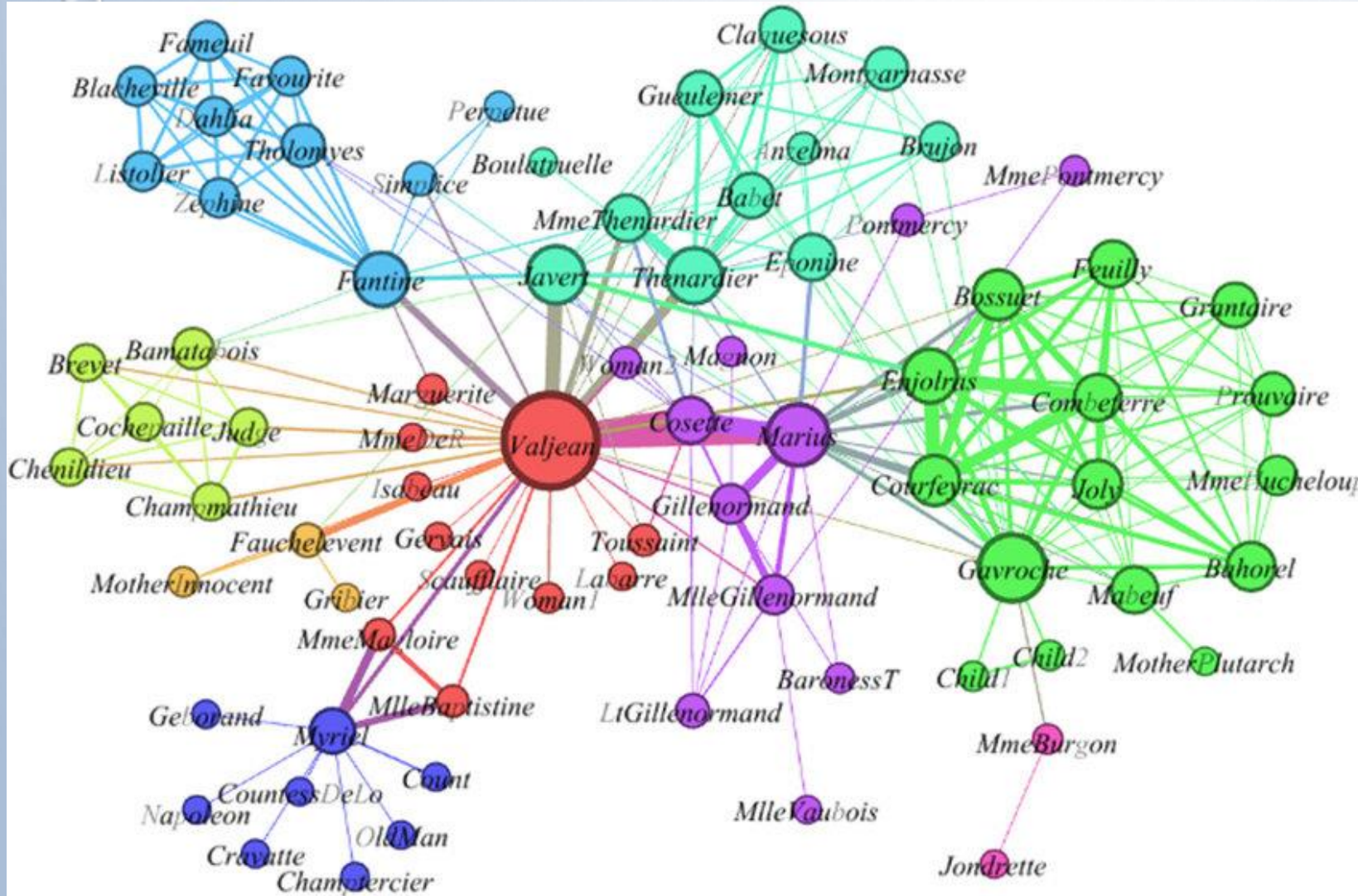
May 7<sup>th</sup>, 2020

# Agenda

- What is graph embedding and motivation from word2vec.
- Node2vec (Grover and Leskovec, 2016):
  - Sampling strategy technique
  - Applications of graph embedding
- Graph embedding with ModelX

# 1. Les Misérables

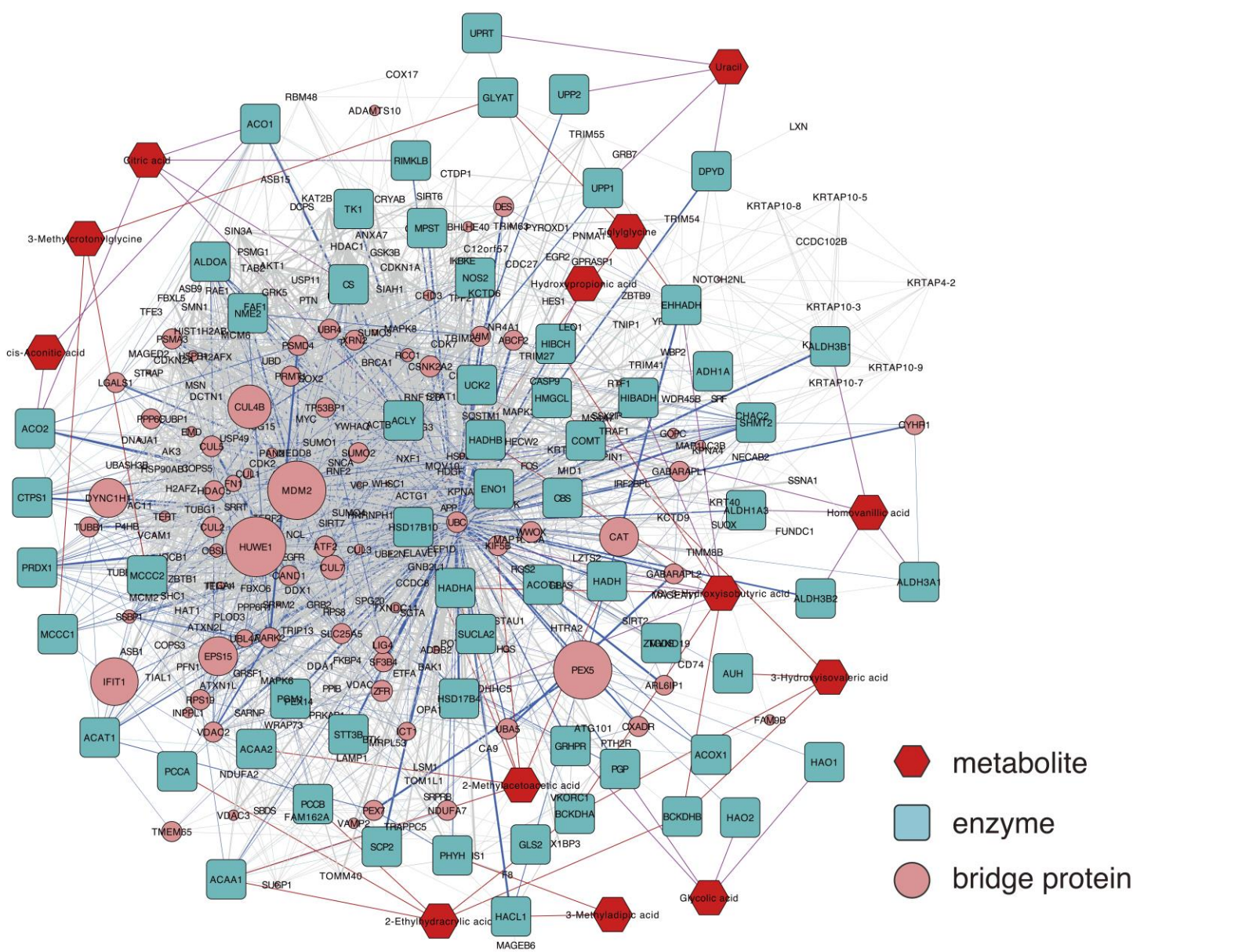
A “Prehistoric” Social Network





# 1. Protein-Protein Interaction

We would like to predict the functional labels of the proteins



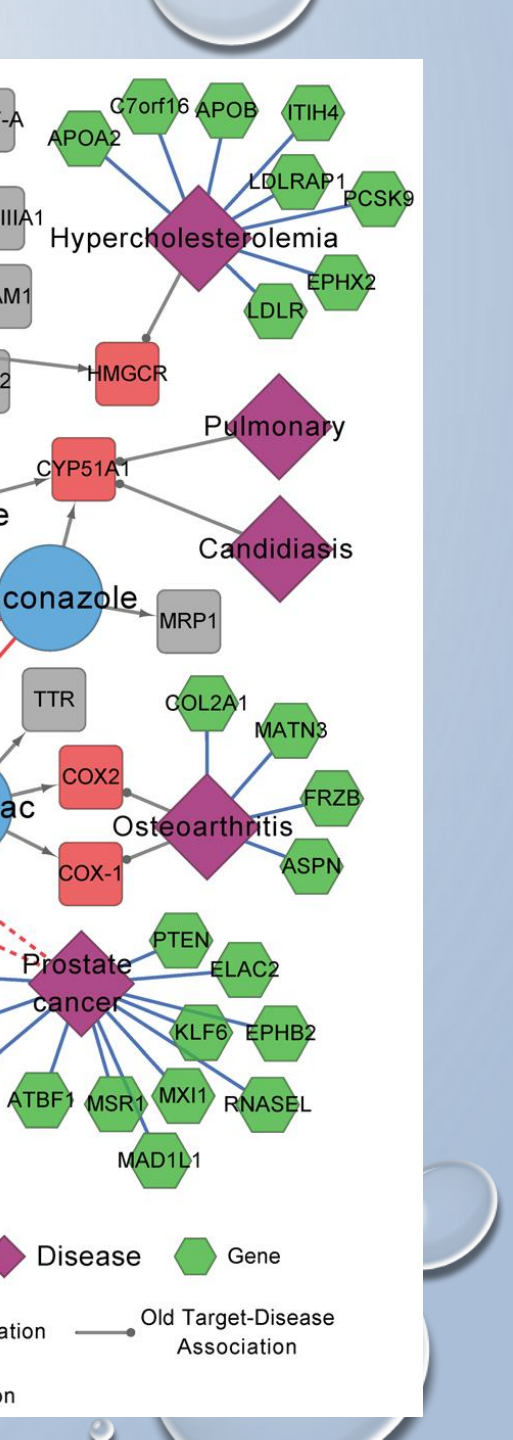
BioGRID-based PPI

— between significant bridge proteins — between enzymes — other

Association of metabolite with enzyme

— KEGG-based — Internally curated










# Classification or link prediction requires representation learning

- In machine learning, representation learning (or feature learning) is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data.
- Two most prominent representations:

## word2vec

### Distributed representations of words and phrases and their compositionality

**Authors:**  [Tomas Mikolov](#),  [Ilya Sutskever](#),  [Kai Chen](#),  [Greg Corrado](#),  [Jeffrey Dean](#)

[Authors Info & Affiliations](#)

**Publication:** NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2 • December 2013 • Pages 3111–3119

## doc2vec

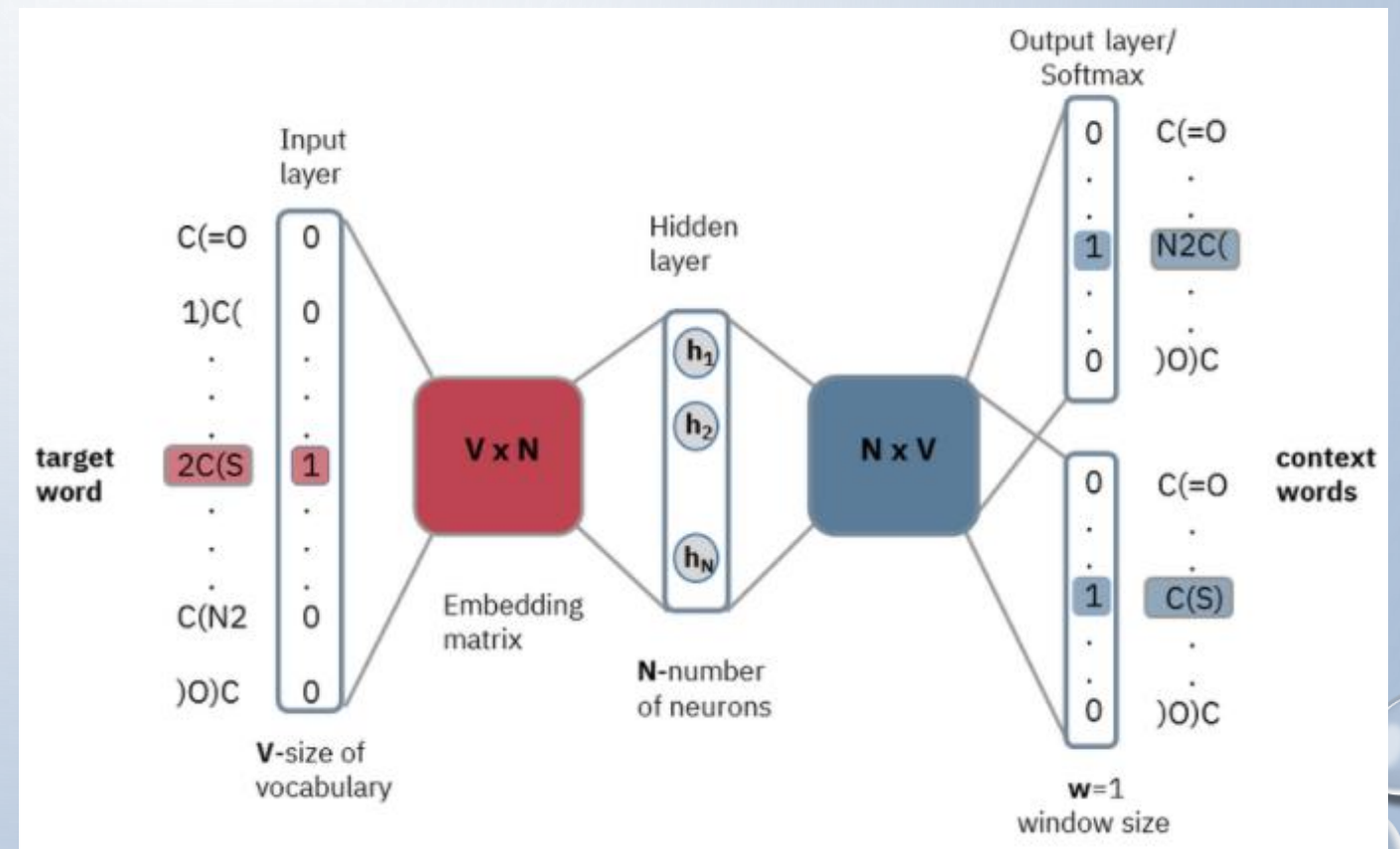
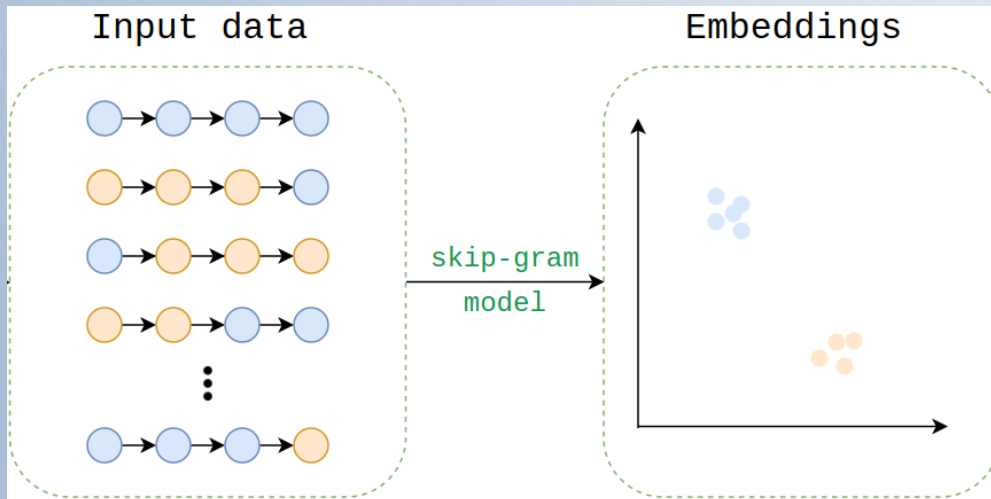
### Distributed Representations of Sentences and Documents

[\[edit\]](#)

**Quoc Le, Tomas Mikolov ; Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, 2014.**

# 1. Skip-gram Architecture of the Word2vec Algorithm

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

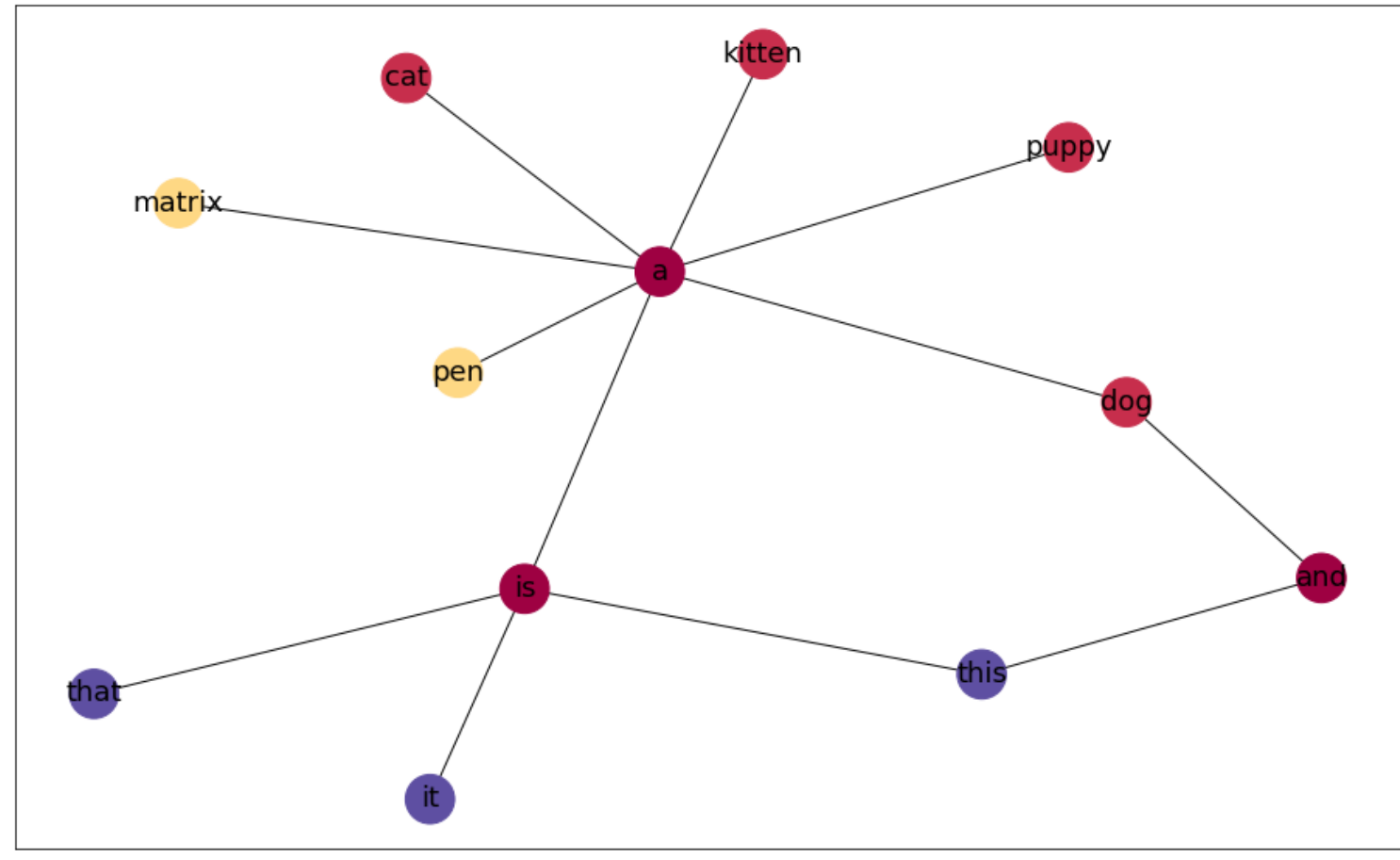




# 1. Text is a special case of graph

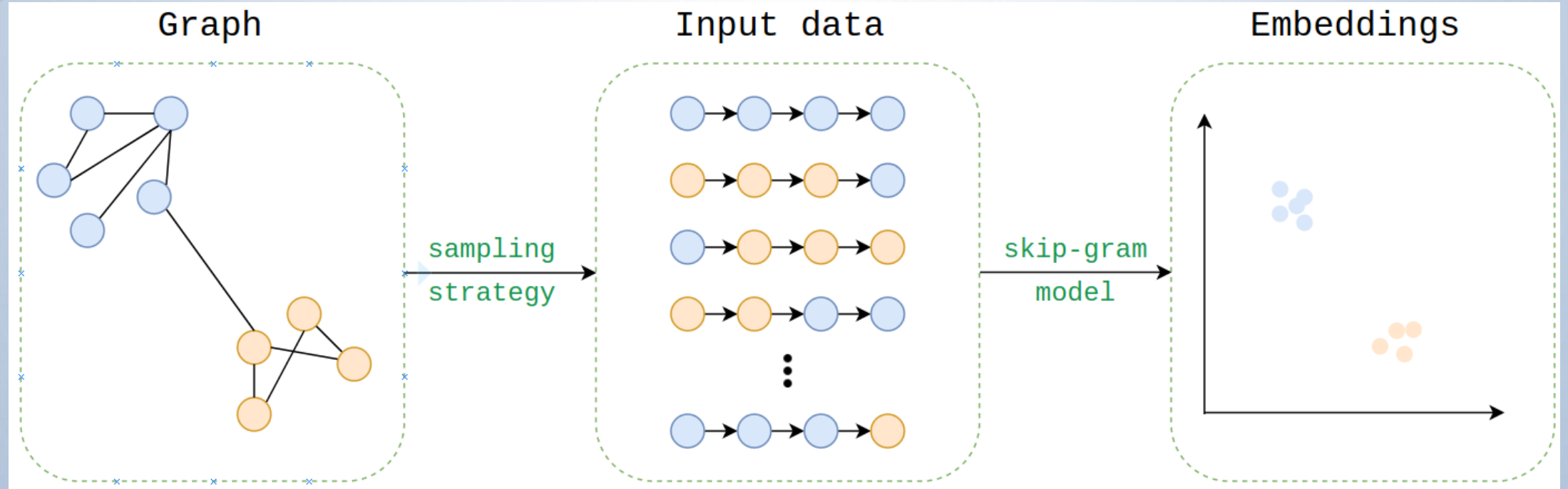
	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

CBOW is one of word representations



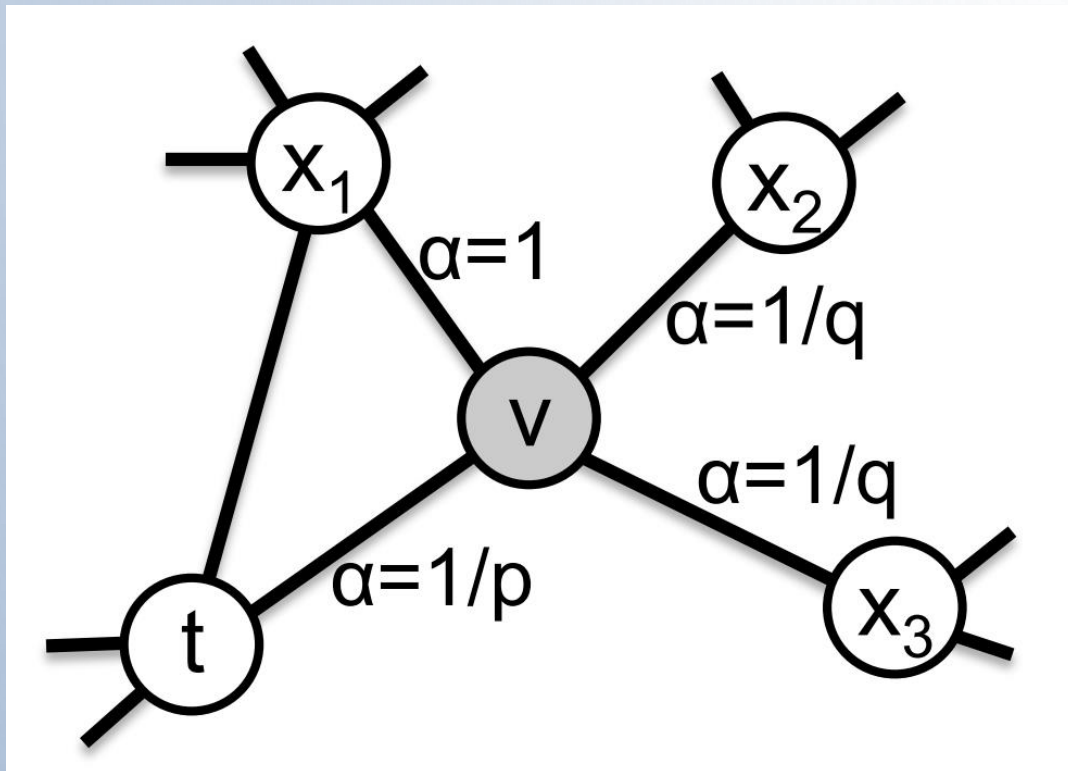


## 2. node2vec: Embeddings for graphs



$$\textit{node2vec}(G(V, E, W)) \rightarrow \mathbb{R}^n$$

## 2. Sampling strategy



$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & \text{if } (v, x) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx}$$

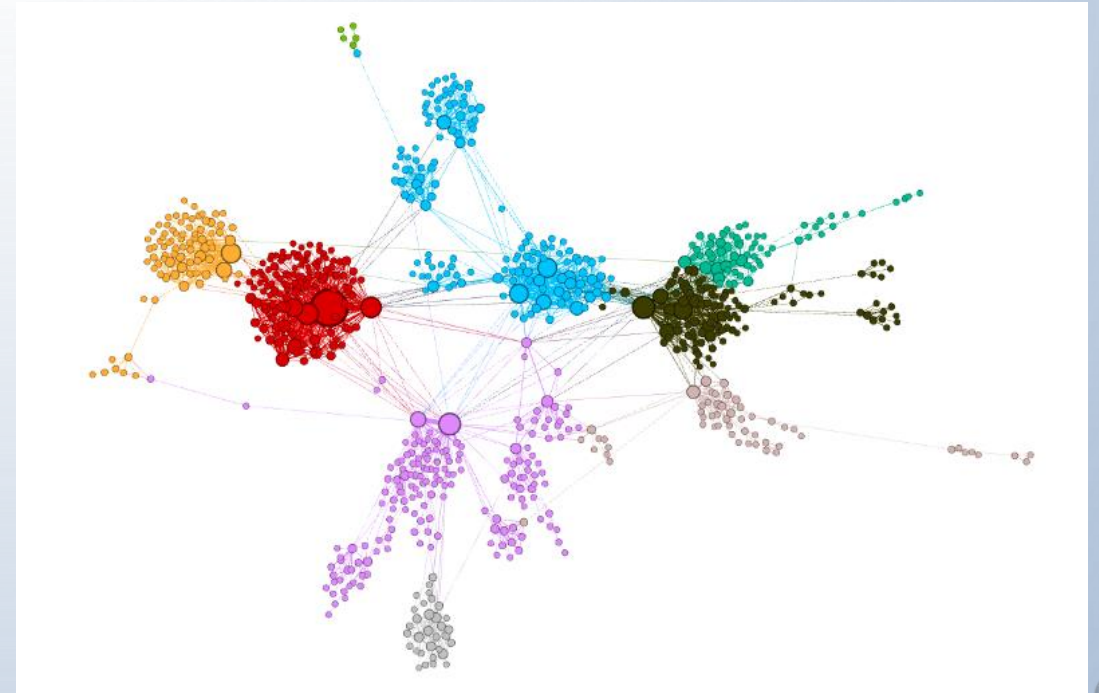
$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

- Node2vec's sampling strategy, accepts 4 arguments:
  - **number of walks**: number of random walks to be generated from each node in the graph
  - **walk length**: how many nodes are in each random walk
  - **p**: return hyperparameter
  - **q**: inout hyperaprameter

## 2. node2vec

Once we train a model and find embeddings for each node we can do:

- Clustering using node embeddings
- Node classification
- Link prediction

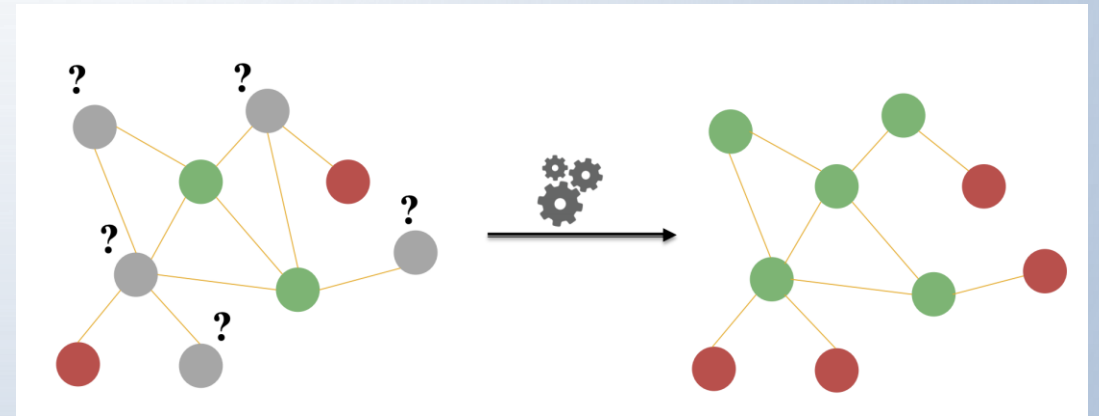




## 2. node2vec

Once we train a model and find embeddings for each node we can do:

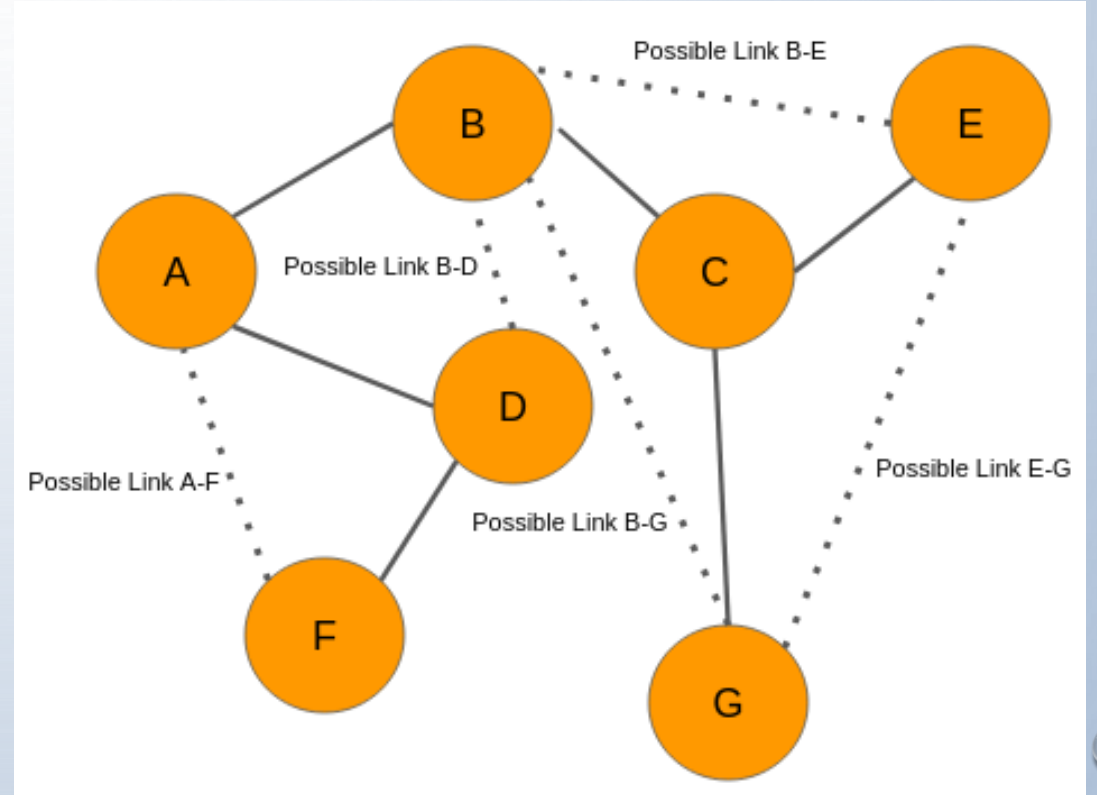
- Clustering using node embeddings
- Node classification
- Link prediction



## 2. node2vec

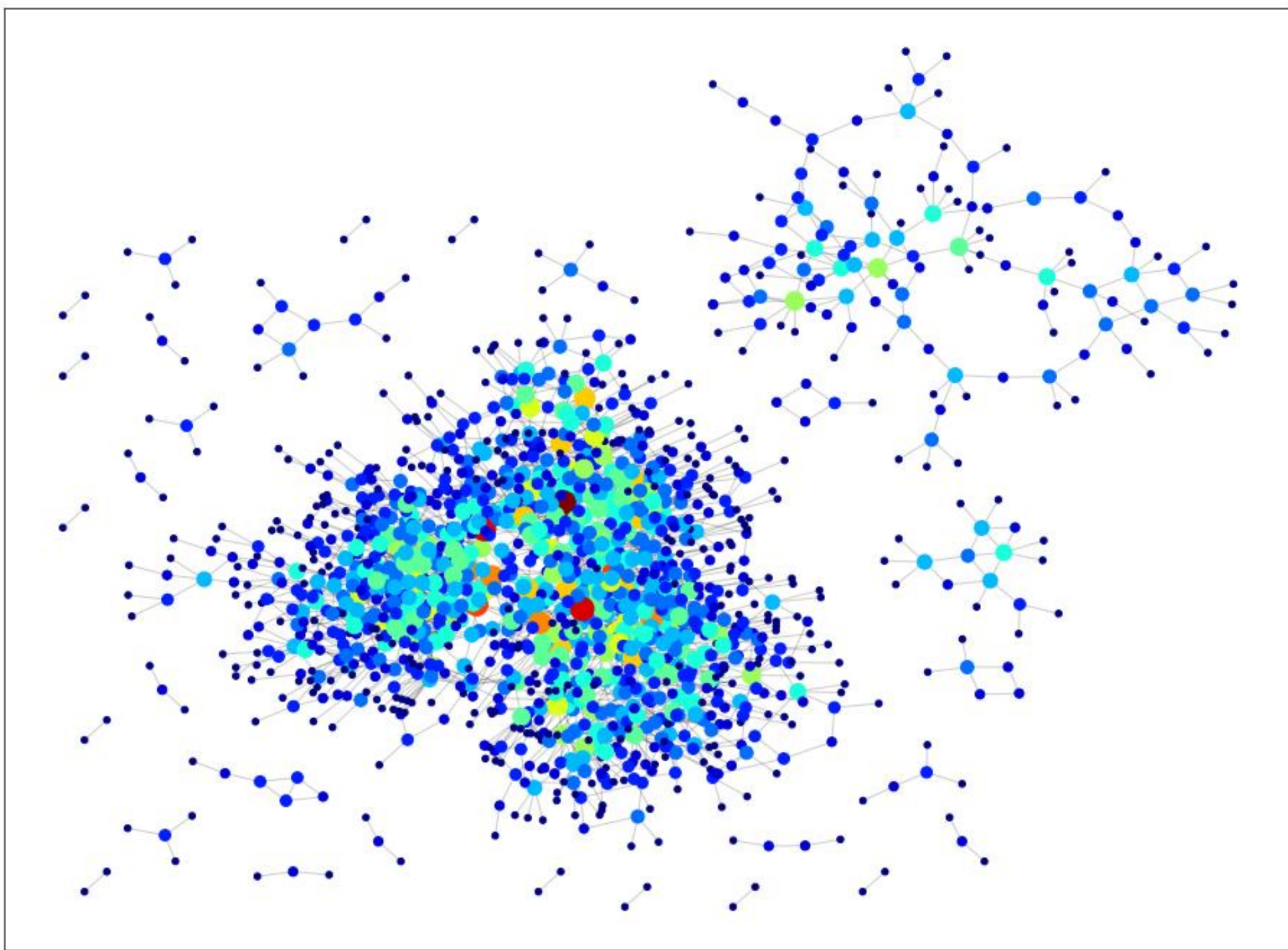
Once we train a model and find embeddings for each node we can do:

- Clustering using node embeddings
- Node classification
- Link prediction



### 3. Appraisal-comp graph

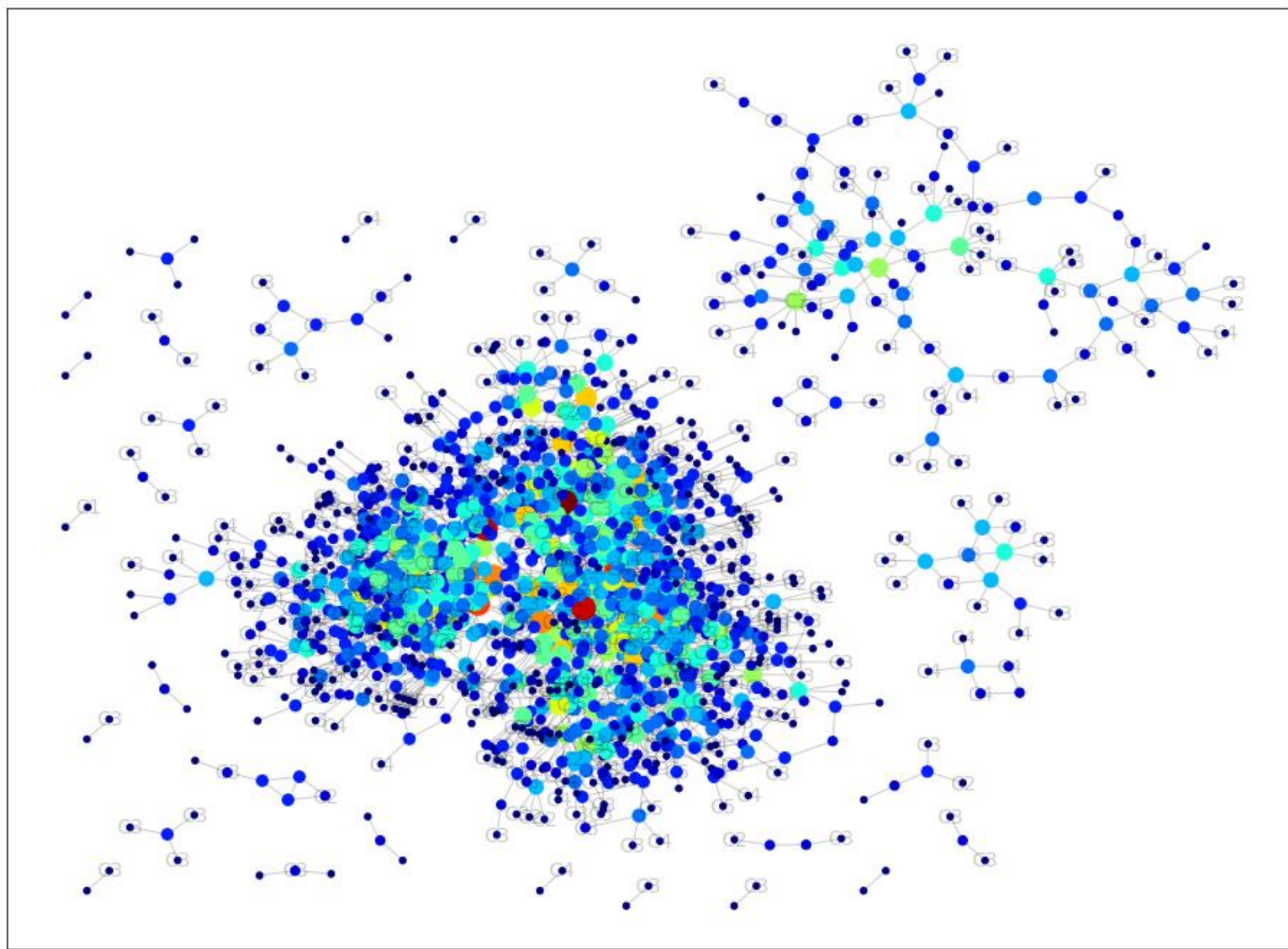
Appraisal embedding will help to find the best comps in the network (link prediction)





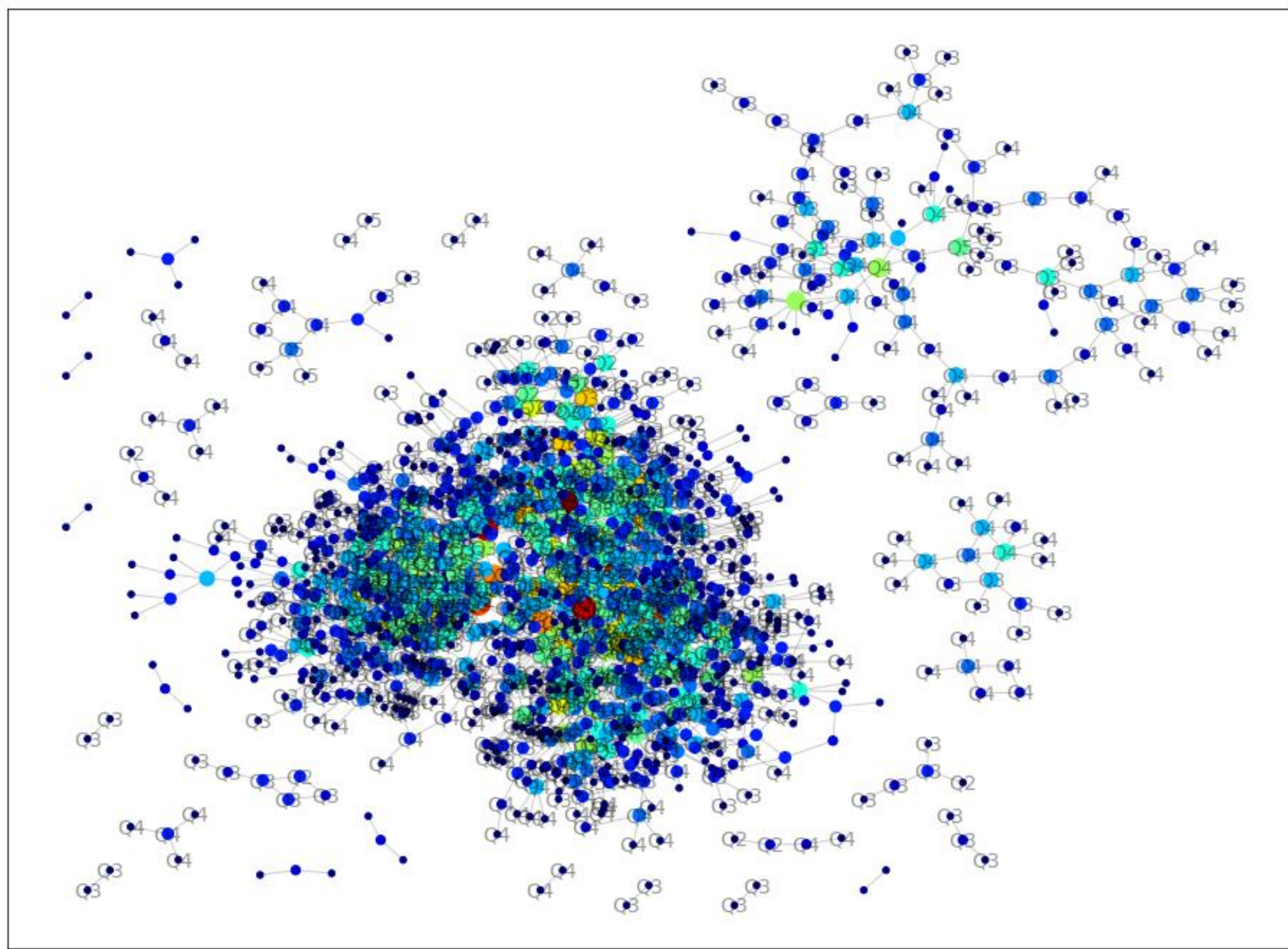
# Property's Condition

Node classification



# Property's Qualiy

Node classification





# 3. Potential Applications In ModelX

A graph of appraisal-comp data has:


- Link between subjects and comps
  - Condition and quality as labels
  - Note: subject-comp link embedded the relationship between subject and comp. Key attributes (beds/baths/age) are not necessary available.
- Appraisal embedding can be used as features in ModelX similarly to embedding of images or text.
  - For ReFi model, it could find best comps without actually looking at property attributes.
  - Back fill missing condition and quality values. Maybe open to other attributes?
  - Appraisal-comp collection is missing quite a lot of data.
  - Should work in non-disclosure states.



ARTICLE

Open Access

# Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2

Yadi Zhou<sup>1</sup>, Yuan Hou<sup>1</sup>, Jiayu Shen<sup>1</sup>, Yin Huang<sup>1</sup>, William Martin<sup>1</sup>  and Feixiong Cheng<sup>1,2,3</sup>

## Abstract

Human coronaviruses (HCoVs), including severe acute respiratory syndrome coronavirus (SARS-CoV) and 2019 novel coronavirus (2019-nCoV, also known as SARS-CoV-2), lead global epidemics with high morbidity and mortality. However, there are currently no effective drugs targeting 2019-nCoV/SARS-CoV-2. Drug repurposing, representing as an effective drug discovery strategy from existing drugs, could shorten the time and reduce the cost compared to de novo drug discovery. In this study, we present an integrative, antiviral drug repurposing methodology implementing a systems pharmacology-based network medicine platform, quantifying the interplay between the HCoV–host interactome and drug targets in the human protein–protein interaction network. Phylogenetic analyses of 15 HCoV whole genomes reveal that 2019-nCoV/SARS-CoV-2 shares the highest nucleotide sequence identity with SARS-CoV (79.7%). Specifically, the envelope and nucleocapsid proteins of 2019-nCoV/SARS-CoV-2 are two evolutionarily conserved regions, having the sequence identities of 96% and 89.6%, respectively, compared to SARS-CoV. Using network proximity analyses of drug targets and HCoV–host interactions in the human interactome, we prioritize 16 potential anti-HCoV repurposable drugs (e.g., melatonin, mercaptopurine, and sirolimus) that are further validated by enrichment analyses of drug-gene signatures and HCoV-induced transcriptomics data in human cell lines. We further identify three potential drug combinations (e.g., sirolimus plus dactinomycin, mercaptopurine plus melatonin, and toremifene plus emodin) captured by the “*Complementary Exposure*” pattern: the targets of the drugs both hit the HCoV–host subnetwork, but target separate neighborhoods in the human interactome network. In summary, this study offers powerful network-based methodologies for rapid identification of candidate repurposable drugs and potential drug combinations targeting 2019-nCoV/SARS-CoV-2.

# Link Prediction In Biomedical Networks

Method Category		Embedding Methods	Link Prediction Tasks			Node Classification Task
			drug-disease association prediction	drug-drug interaction prediction	protein-protein interaction prediction	medical term type classification
Traditional	Matrix Factorization-based	Laplacian	(Zhang <i>et al.</i> , 2018d)	(Zhang <i>et al.</i> , 2018b)	(Zhu <i>et al.</i> , 2013)	X
		SVD	(Dai <i>et al.</i> , 2015)	X	(You <i>et al.</i> , 2017)	X
		GF	(Yang <i>et al.</i> , 2014)	(Zhang <i>et al.</i> , 2018b)	X	X
		HOPE	X	X	X	X
Recently Proposed	Random Walk-based	GraRep	X	X	X	X
		DeepWalk	X	X	X	X
		node2vec	X	X	X	X
		struc2vec	X	X	X	X
		LINE	X	X	X	X
		SDNE	X	X	(Wang <i>et al.</i> , 2017b)	X
	Neural Network-based	GAE	X	(Zitnik <i>et al.</i> , 2018) (Ma <i>et al.</i> , 2018)	X	X

Yue, *et al.*, 2020

# References

- [Word2vec tutorial - the skip-gram model](#)
- [The Premise of Deep Learning](#)
- [Graph Embedding on Biomedical Networks: Methods, Applications, and Evaluations](#)
- [node2vec: Embeddings for Graph Data](#)
- [Graph Embeddings — The Summary](#)