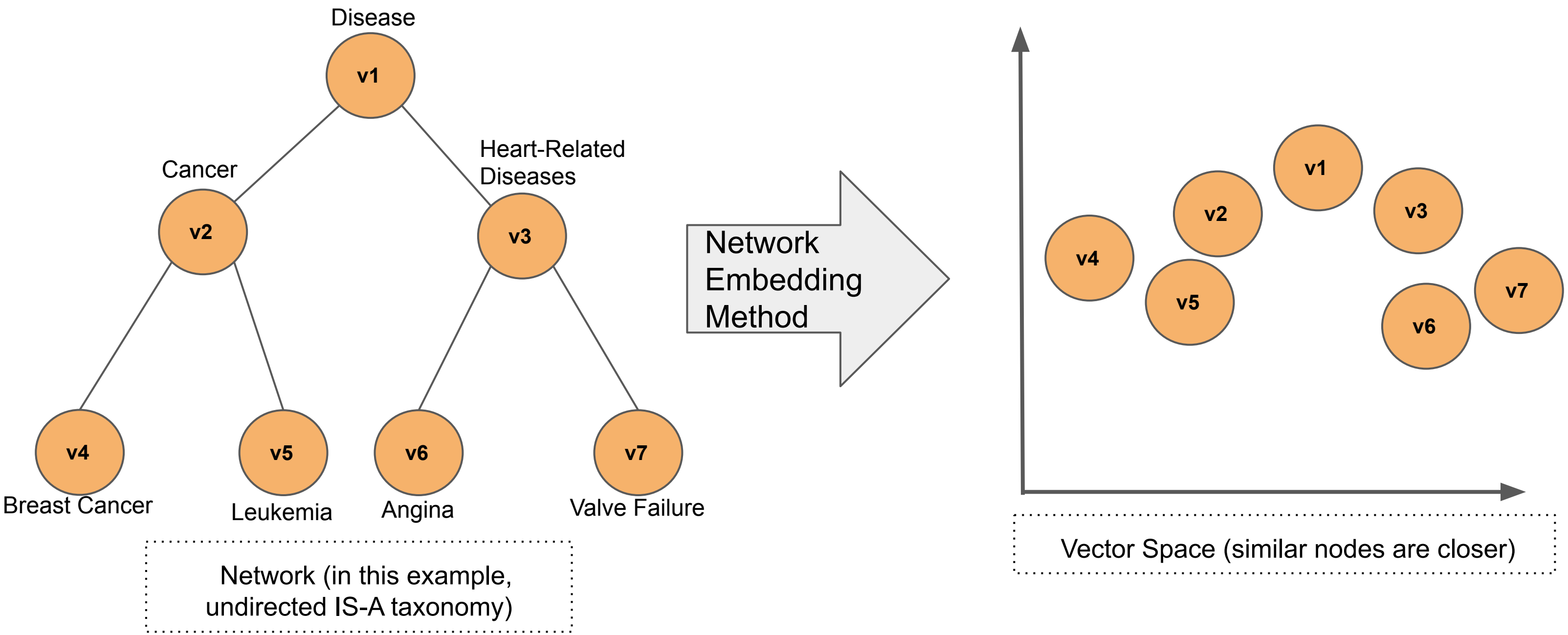
Embedding Biomedical Ontologies by Jointly Encoding Network Structure and Textual Node Descriptors

Sotiris Kotitsas, Dimitris Pappas, **Ion Androutsopoulos**,
Ryan McDonald, Marianna Apidianaki

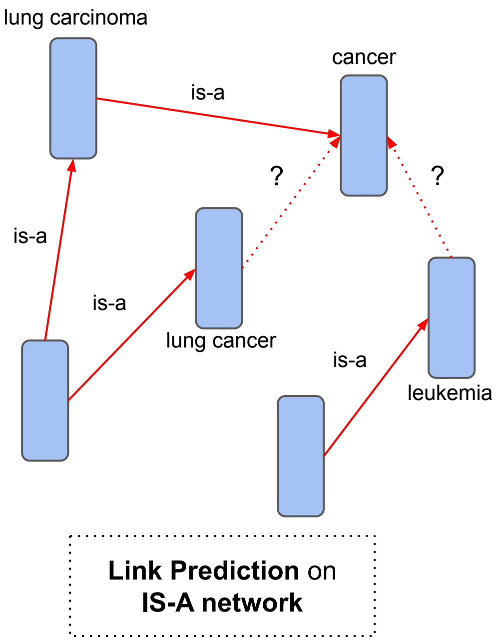
Network Embedding Methods

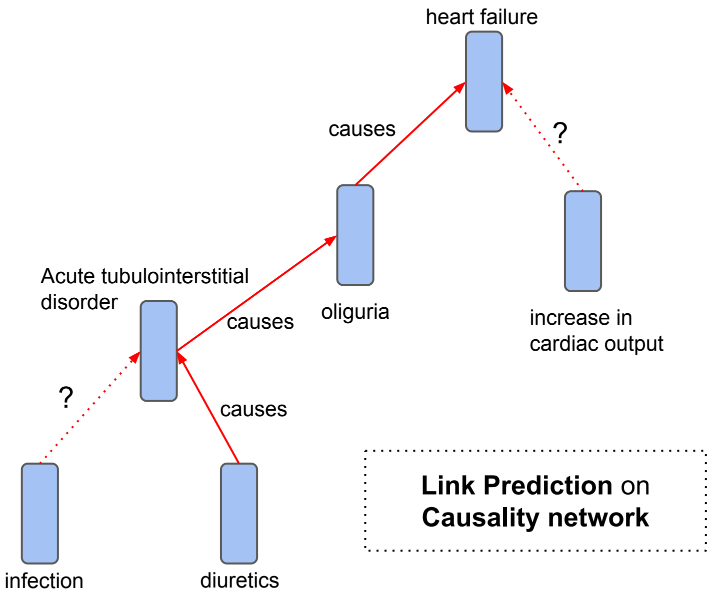
- Map each **node** of a network to a low-dimensional **feature vector ("node embedding")**, so that more **"similar" nodes** will be **closer in the vector space**.



Applications of Network Embedding Methods

- **Link prediction** (discover missing interactions): **classify pairs** of node embeddings.



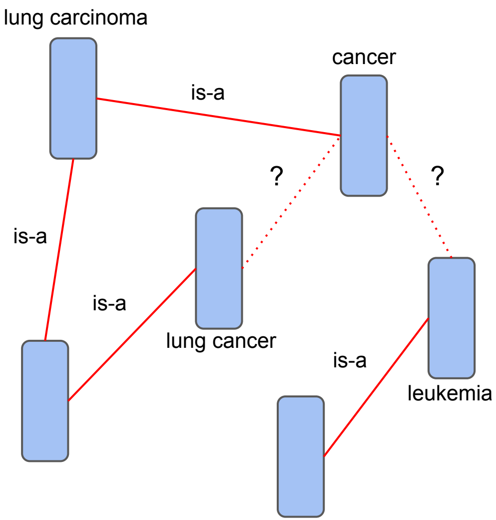


Applications of Network Embedding Methods

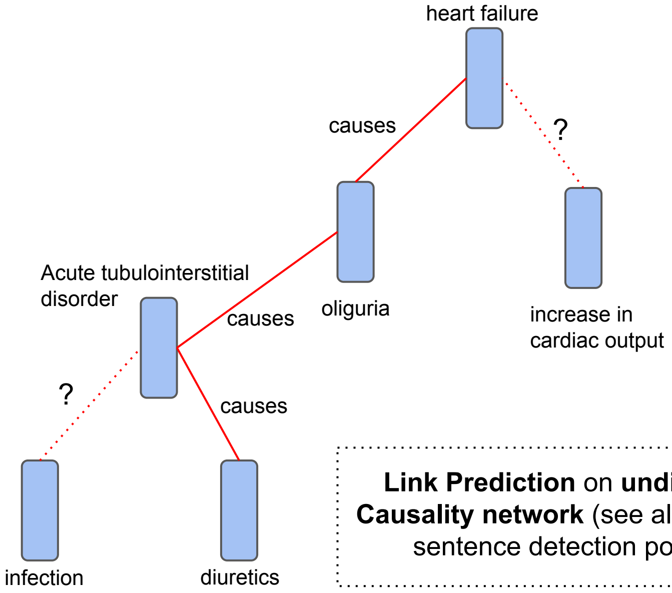
- **Link prediction** (discover missing interactions): classify **pairs** of node embeddings.
- **Node classification** (e.g., identify proteins linked with diseases): classify **single** node embeddings.

Applications of Network Embedding Methods

- **Link prediction** (discover missing interactions): classify **pairs** of node embeddings.
- **Node classification** (e.g., identify proteins linked with diseases): classify **single** node embeddings.
- We experimented with **link prediction** on **undirected IS-A and PART-OF** networks.

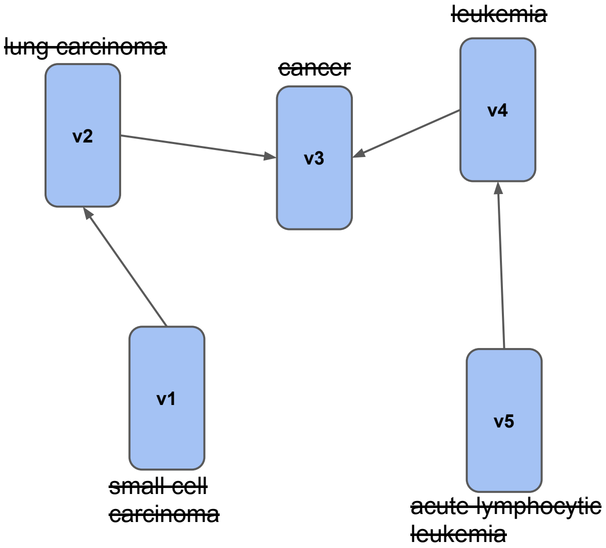


Link Prediction on **undirected IS-A network**



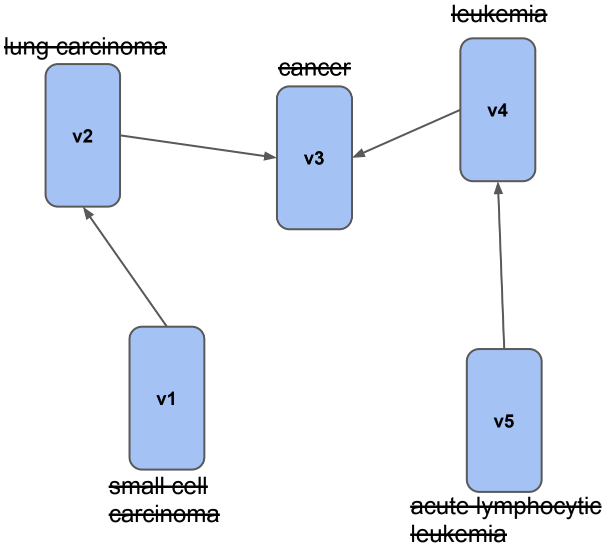
Link Prediction on **undirected Causality network** (see also causal sentence detection poster)

Structure-Oriented and Content-Oriented Methods



Structure-Oriented methods consider only the **network structure**. **Similar nodes** are nodes with **similar neighbors**.

Structure-Oriented and Content-Oriented Methods

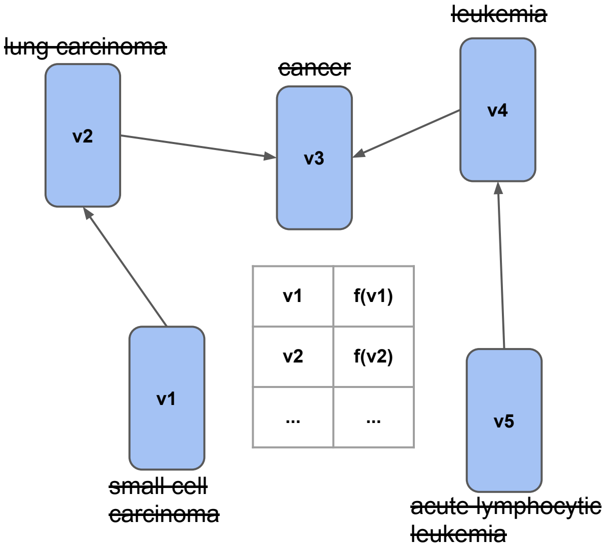


Structure-Oriented methods consider only the **network structure**. **Similar nodes** are nodes with **similar neighbors**.

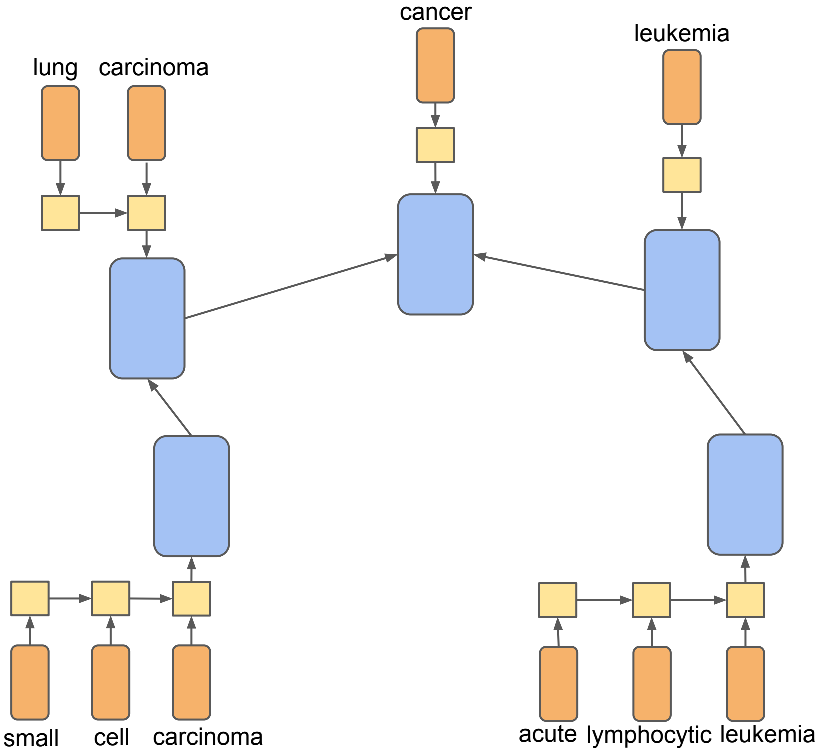
Node	Embedding
v1	f(v1)
v2	f(v2)
v3	f(v3)
v4	f(v4)
v5	f(v5)
v6	f(v6)

Lookup table directly links **nodes to embeddings**, ignoring texts.

Structure-Oriented and Content-Oriented Methods

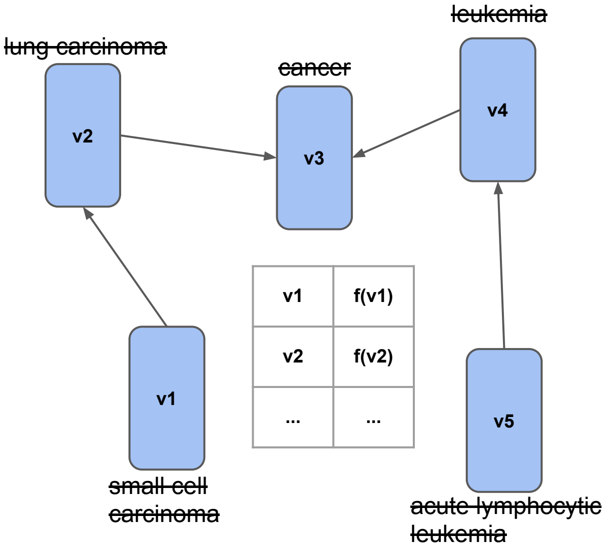


Structure-Oriented methods consider only the **network structure**. **Similar nodes** are nodes with **similar neighbors**.



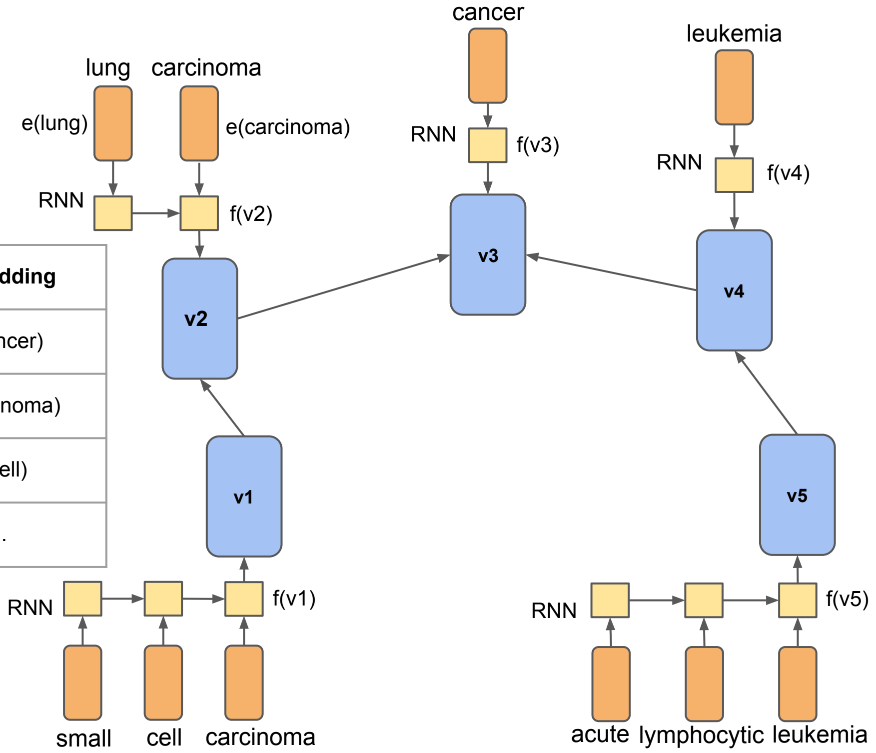
Content-Oriented methods also consider the **textual descriptors** (or longer texts) of the nodes. **Similar nodes** are nodes with **similar neighbors and similar texts**.

Structure-Oriented and Content-Oriented Methods



v1	f(v1)
v2	f(v2)
...	...

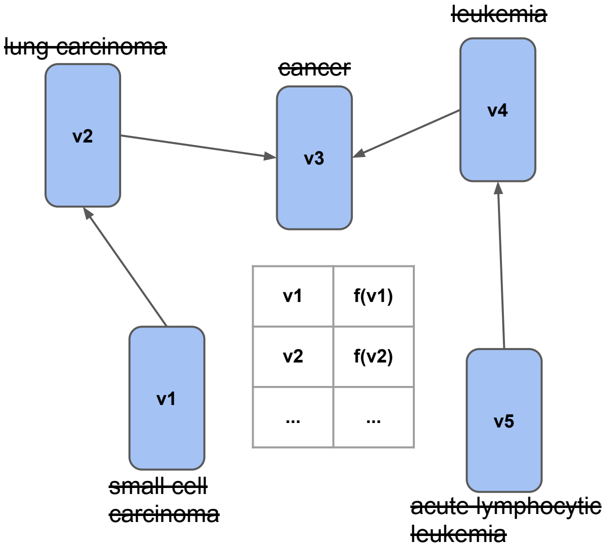
Word	Embedding
cancer	e(cancer)
carcinoma	e(carcinoma)
cell	e(cell)
...	...



Structure-Oriented methods consider only the **network structure**. **Similar nodes** are nodes with **similar neighbors**.

Content-Oriented methods also consider the **textual descriptors** (or longer texts) of the nodes. **Similar nodes** are nodes with **similar neighbors and similar texts**.

Structure-Oriented and Content-Oriented Methods

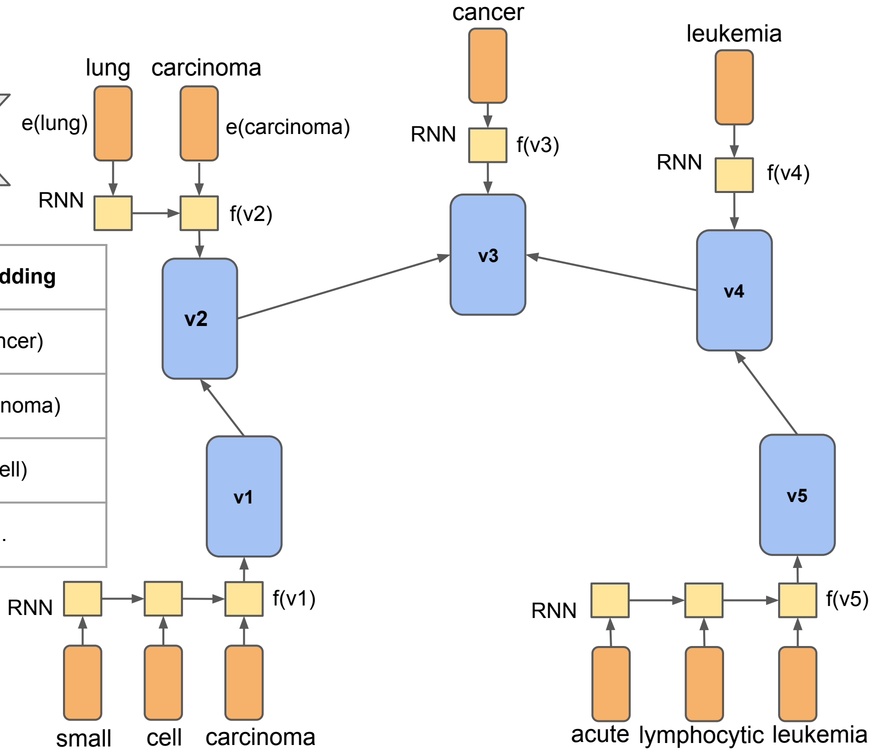


v1	**f(v1)**
v2	**f(v2)**
...	...

Structure-Oriented methods consider only the **network structure**. **Similar nodes** are nodes with **similar neighbors**.

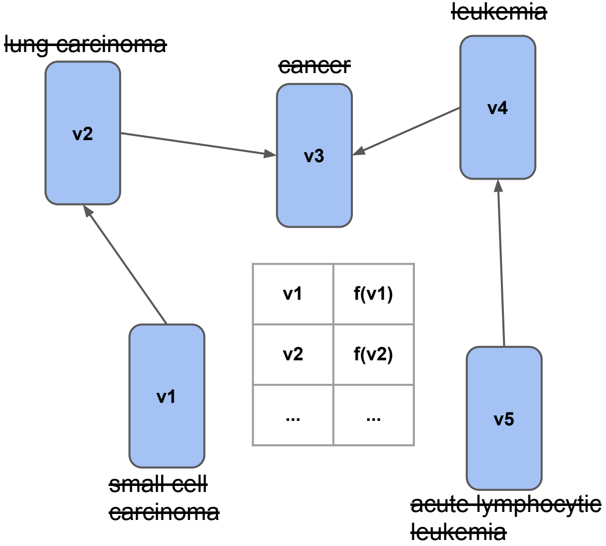
By-product: word embeddings

Word	Embedding
cancer	e(cancer)
carcinoma	e(carcinoma)
cell	e(cell)
...	...



Content-Oriented methods also consider the **textual descriptors** (or longer texts) of the nodes. **Similar nodes** are nodes with **similar neighbors and similar texts.**

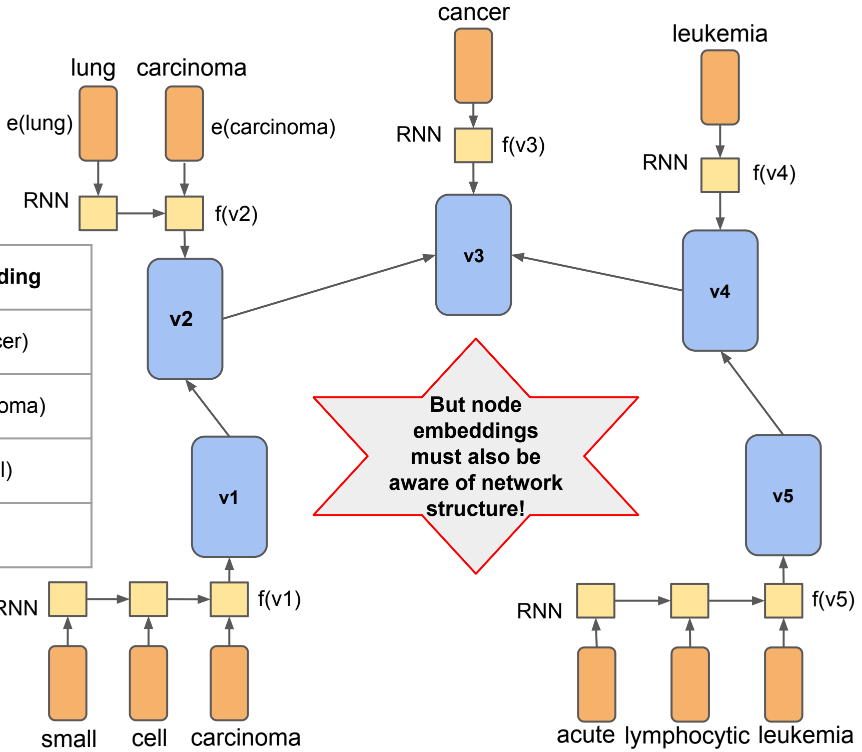
Structure-Oriented and Content-Oriented Methods



v1	f(v1)
v2	f(v2)
...	...

By-product: word embeddings

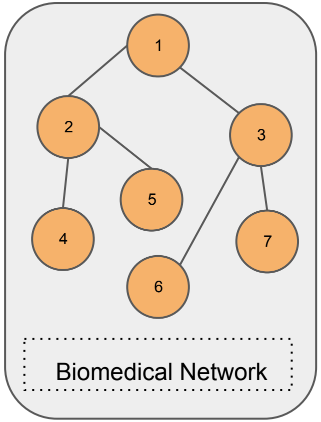
Word	Embedding
cancer	e(cancer)
carcinoma	e(carcinoma)
cell	e(cell)
...	...



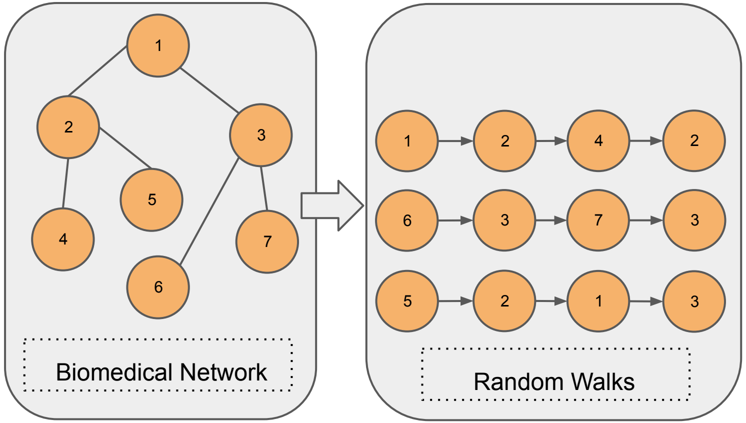
Structure-Oriented methods consider only the **network structure**. **Similar nodes** are nodes with **similar neighbors**.

Content-Oriented methods also consider the **textual descriptors** (or longer texts) of the nodes. **Similar nodes** are nodes with **similar neighbors and similar texts**.

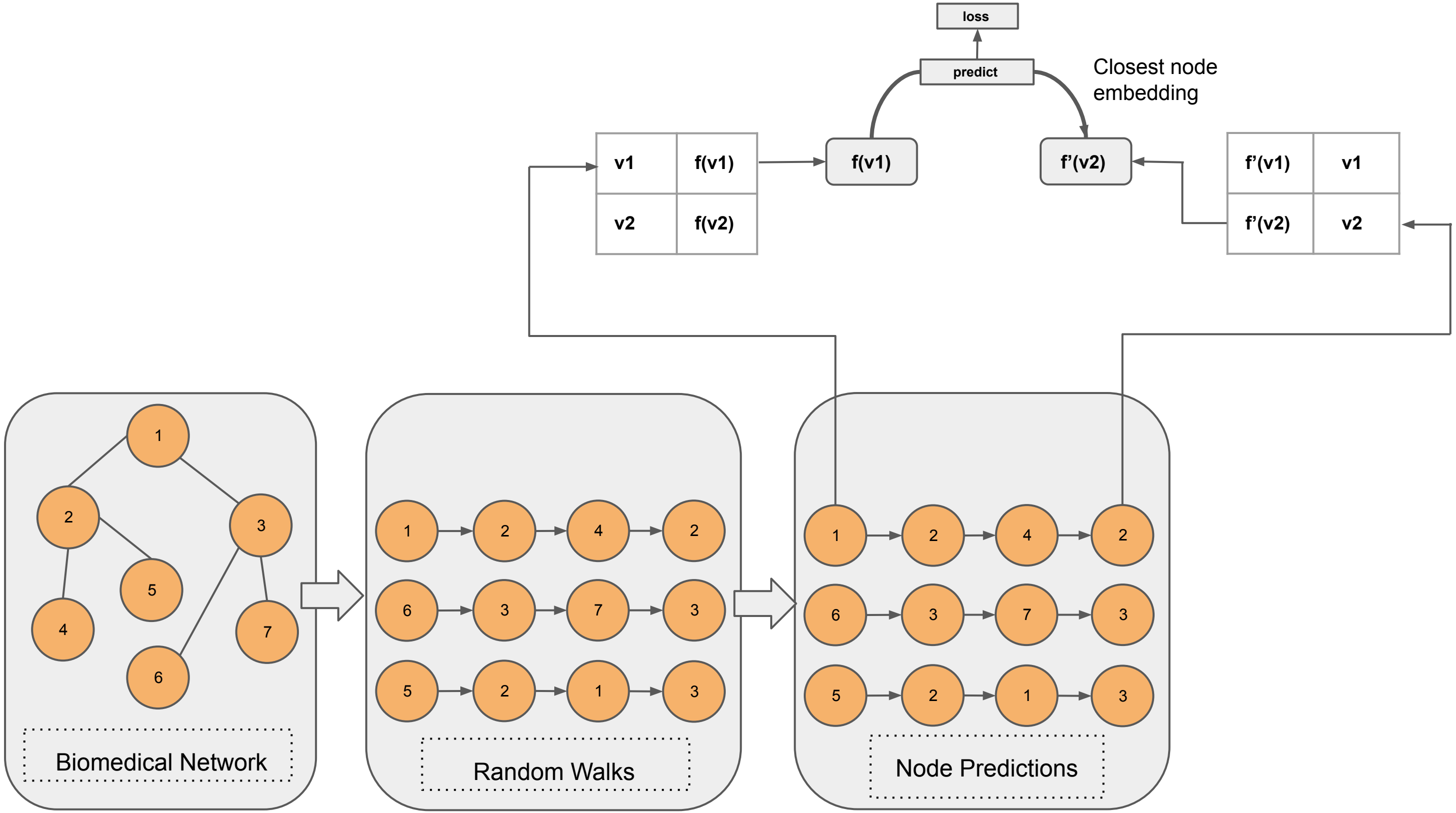
Original Node2Vec (Structure-Oriented)



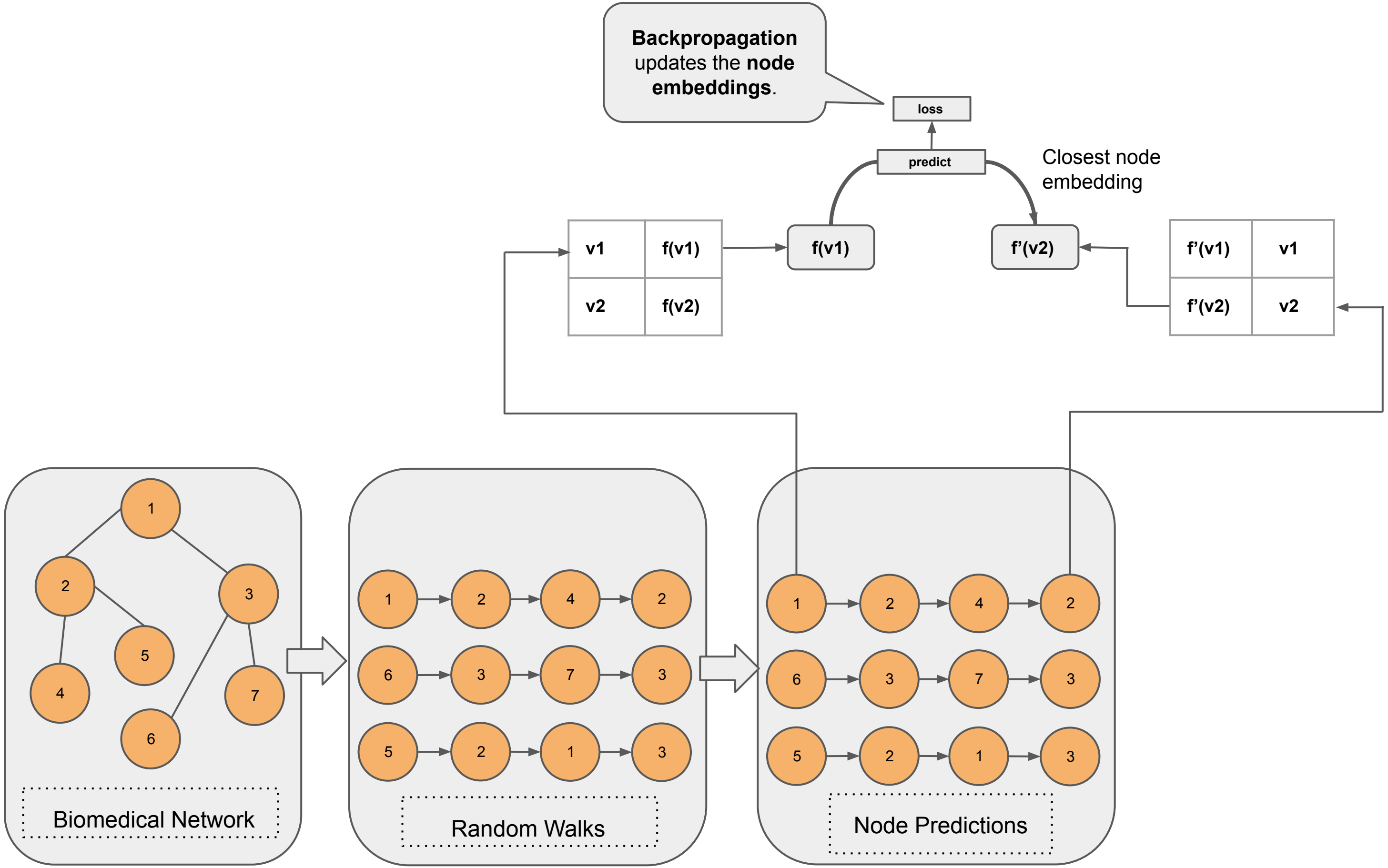
Original Node2Vec (Structure-Oriented)



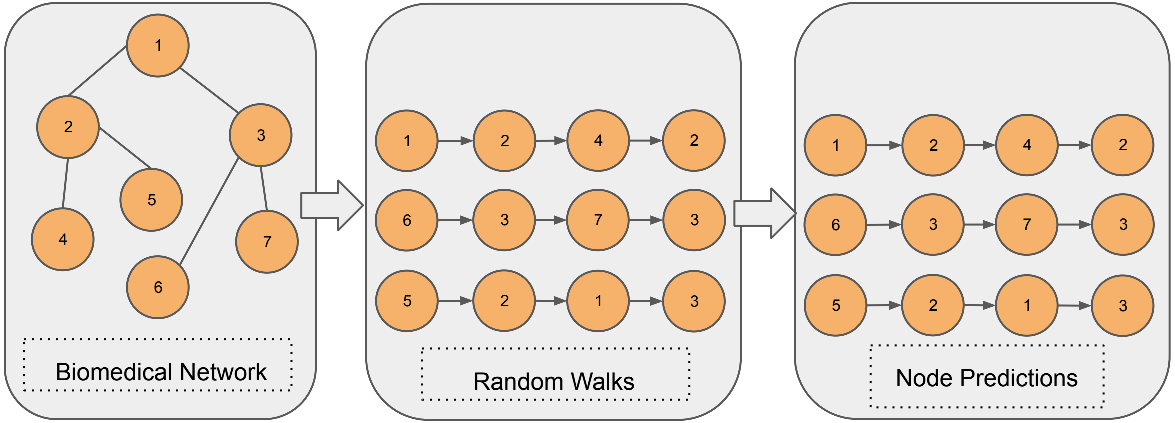
Original Node2Vec (Structure-Oriented)



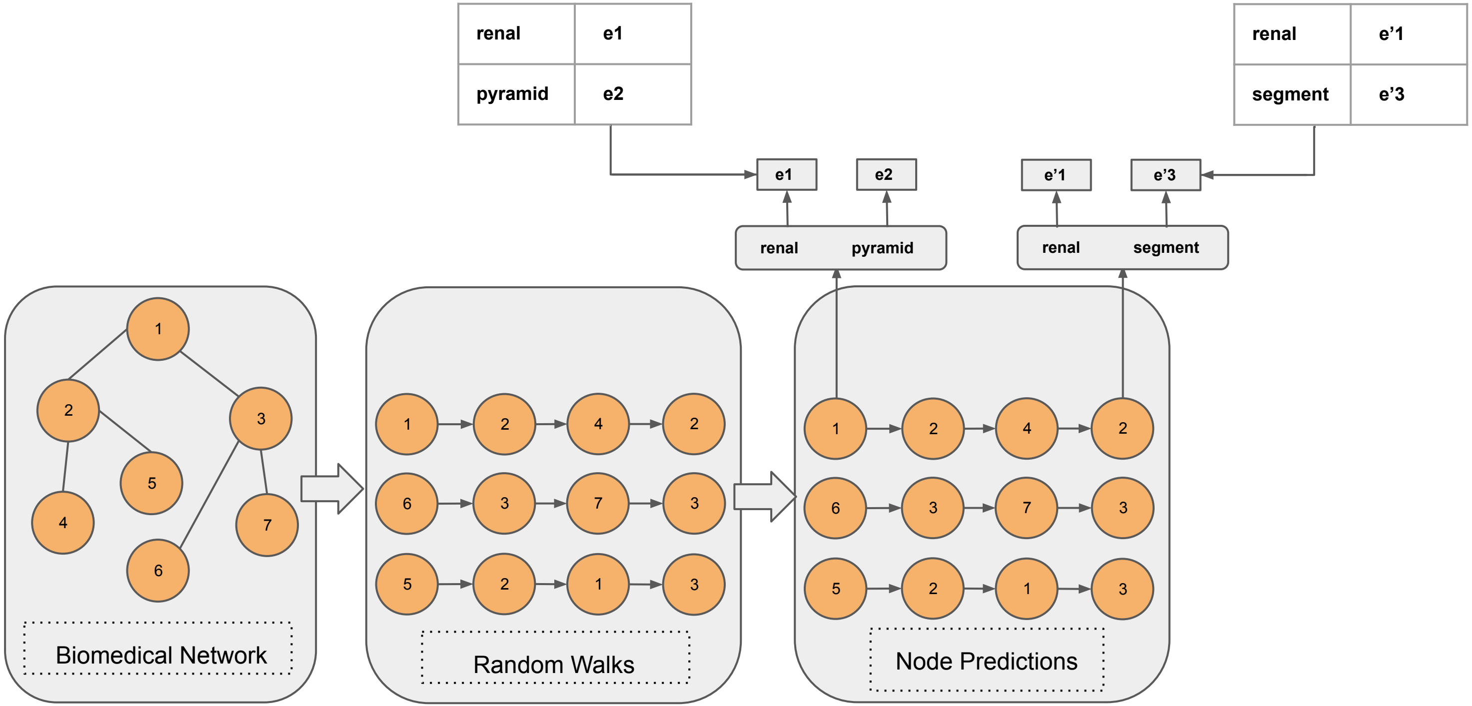
Original Node2Vec (Structure-Oriented)



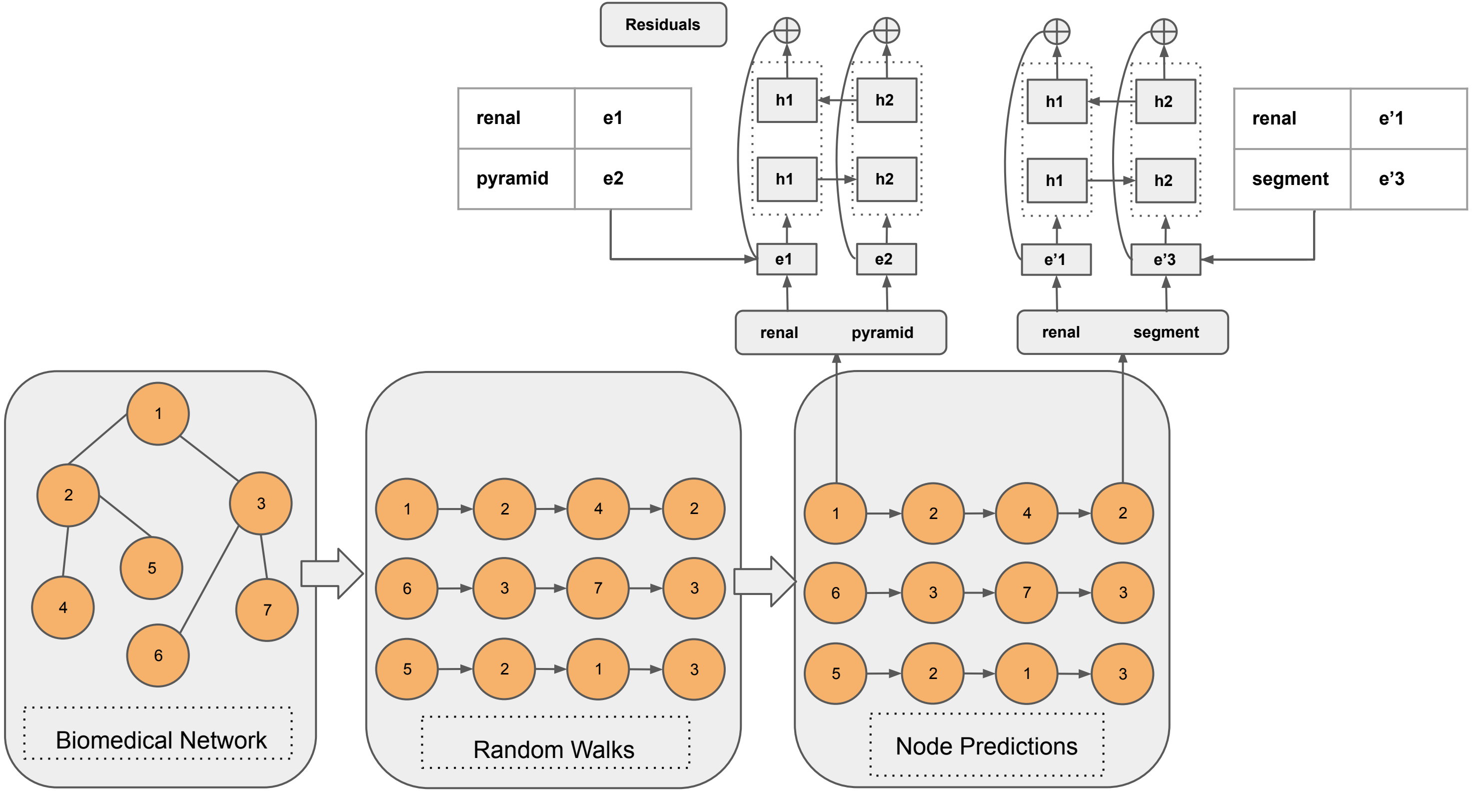
Proposed Content-Oriented Extension of Node2Vec



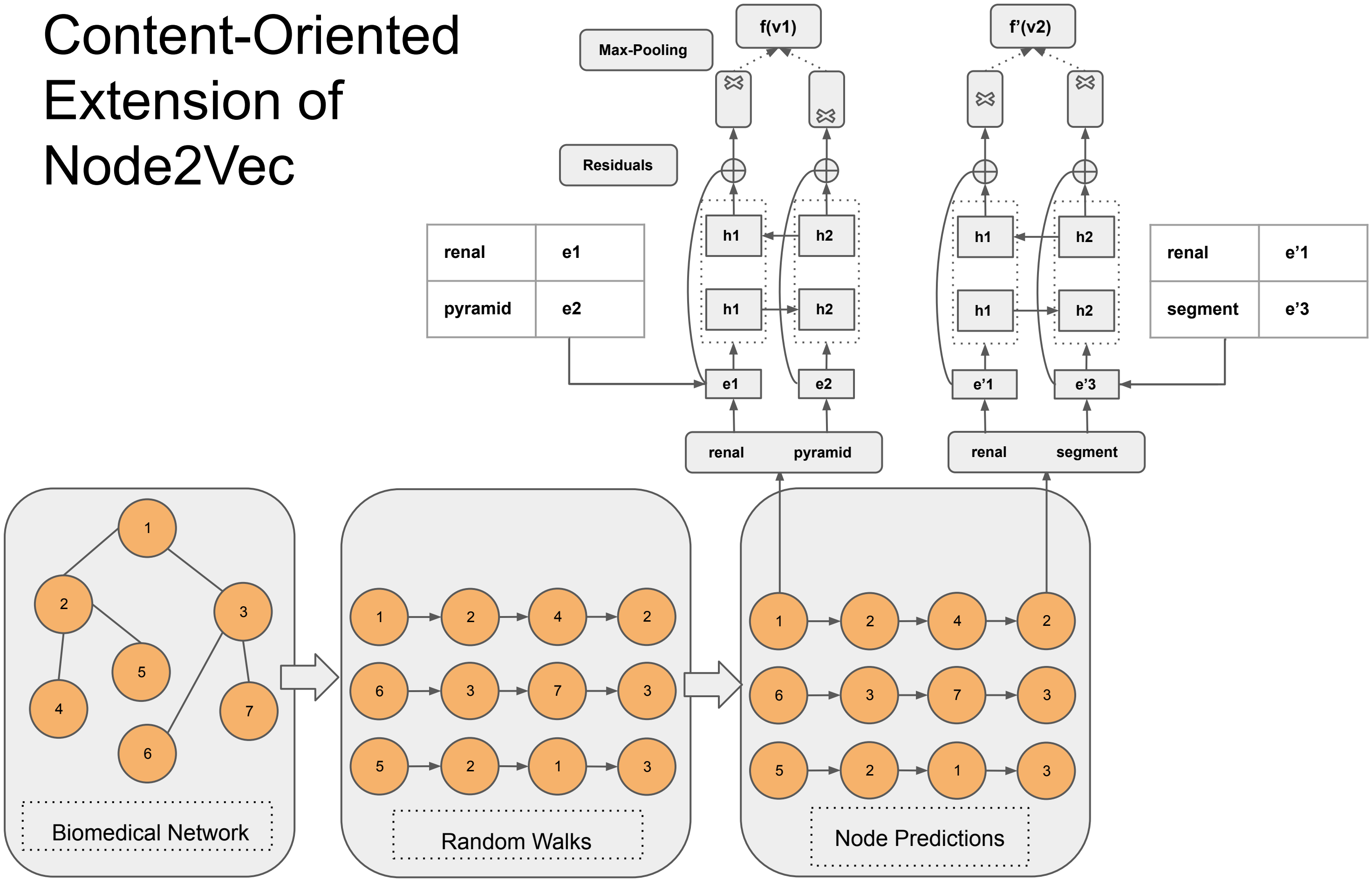
Proposed Content-Oriented Extension of Node2Vec



Proposed Content-Oriented Extension of Node2Vec

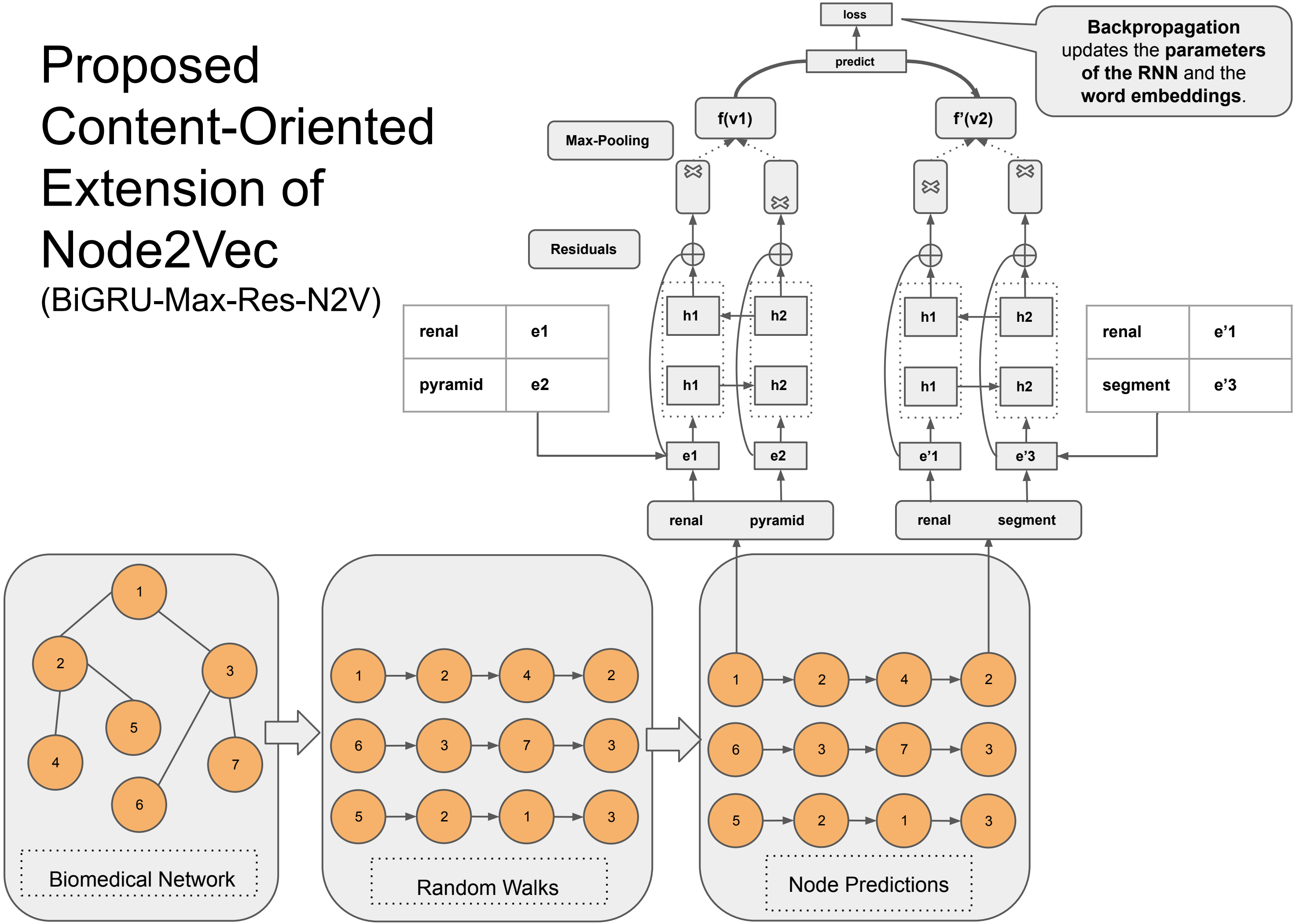


Proposed Content-Oriented Extension of Node2Vec



Proposed Content-Oriented Extension of Node2Vec

(BiGRU-Max-Res-N2V)

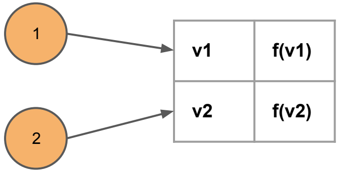


Link Prediction (to evaluate node embeddings)

- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.

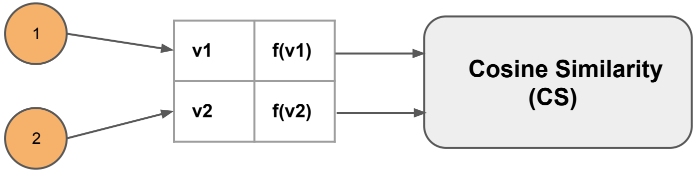
Link Prediction (to evaluate node embeddings)

- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.



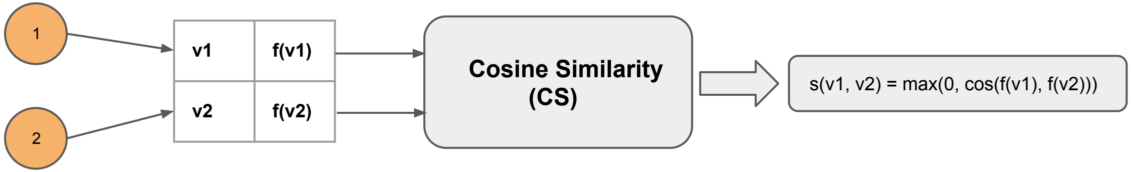
Link Prediction (to evaluate node embeddings)

- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.



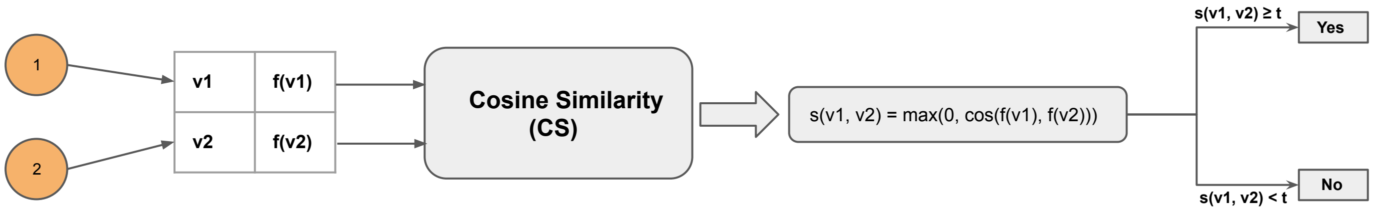
Link Prediction (to evaluate node embeddings)

- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.



Link Prediction (to evaluate node embeddings)

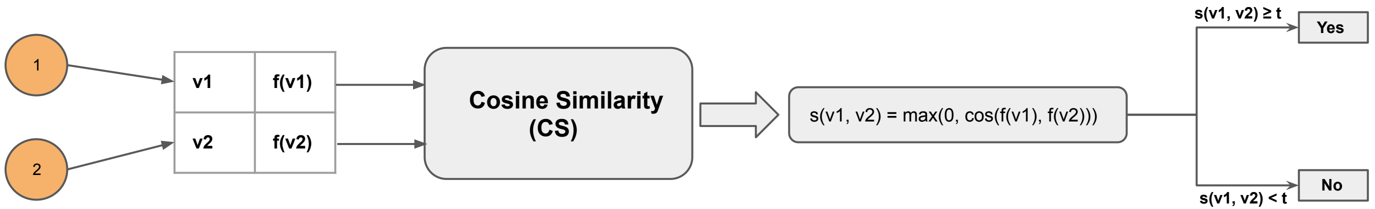
- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.



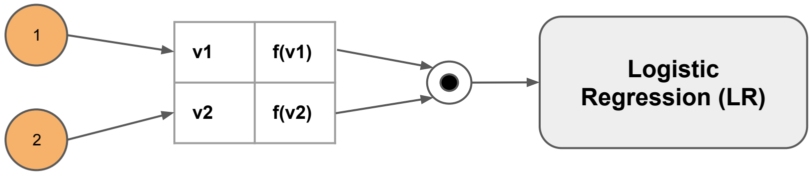
AUC scores computed for **evaluation**

Link Prediction (to evaluate node embeddings)

- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.

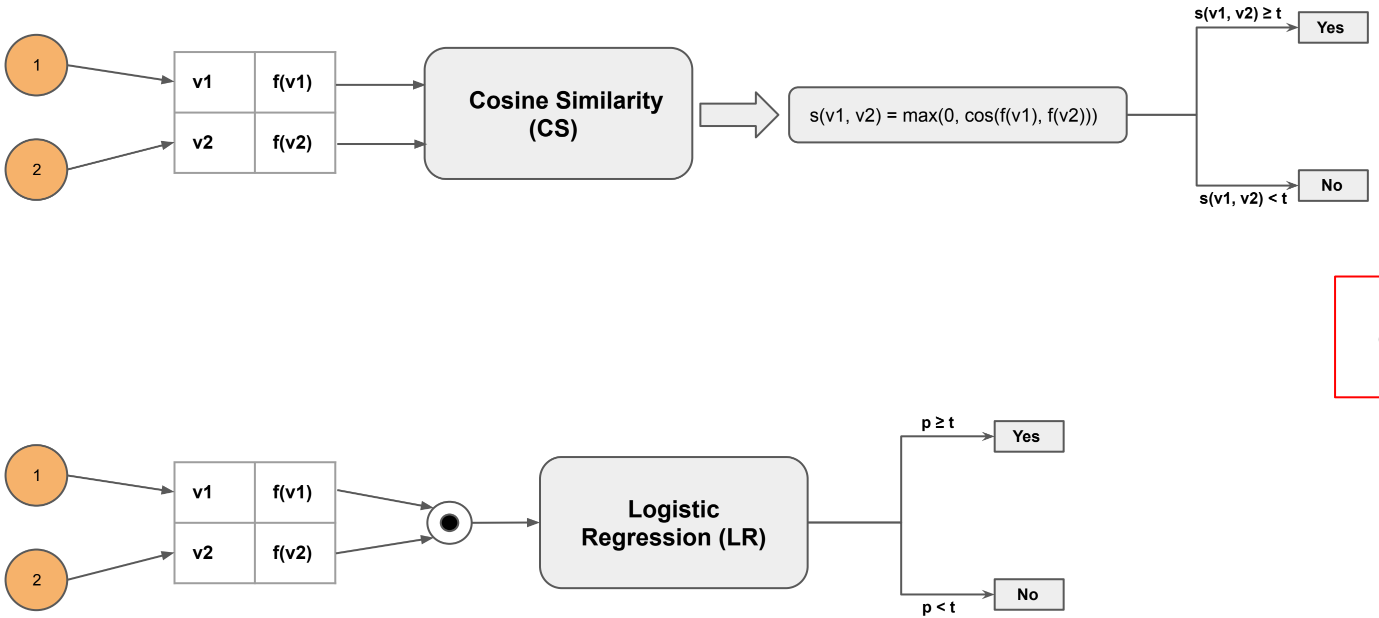


AUC scores computed for **evaluation**



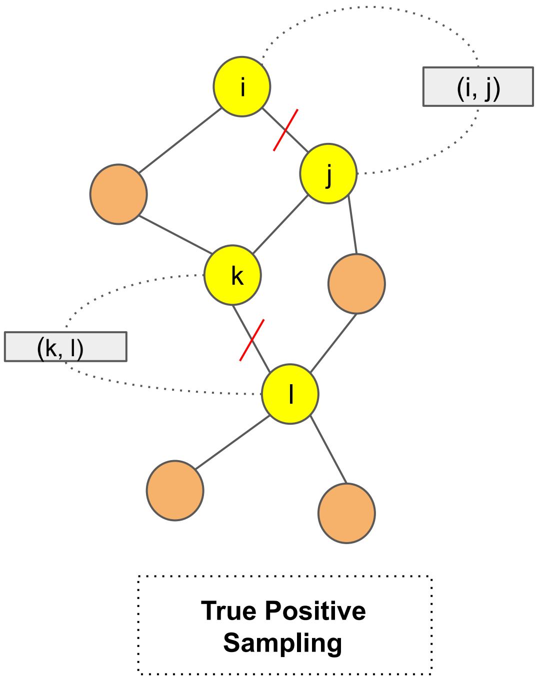
Link Prediction (to evaluate node embeddings)

- We experiment with: **Cosine Similarity Predictor** (**CS**), **Logistic Regression** (**LR**).
- **Node embeddings** have already been **generated**.



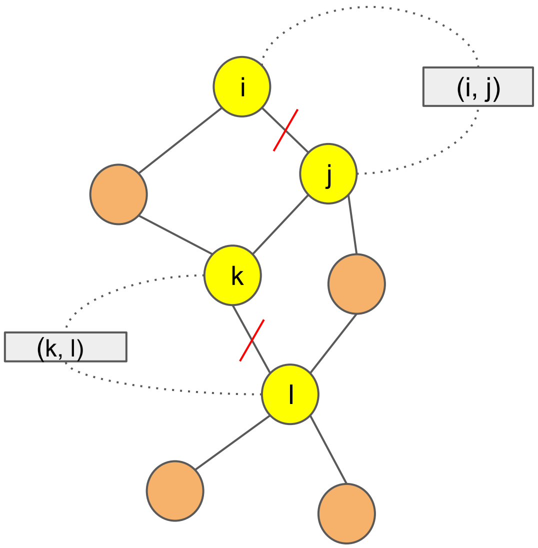
AUC scores computed for **evaluation**

Link Prediction: True Positive & True Negative Edges for Testing



Link Prediction: True Positive & True Negative Edges for Testing

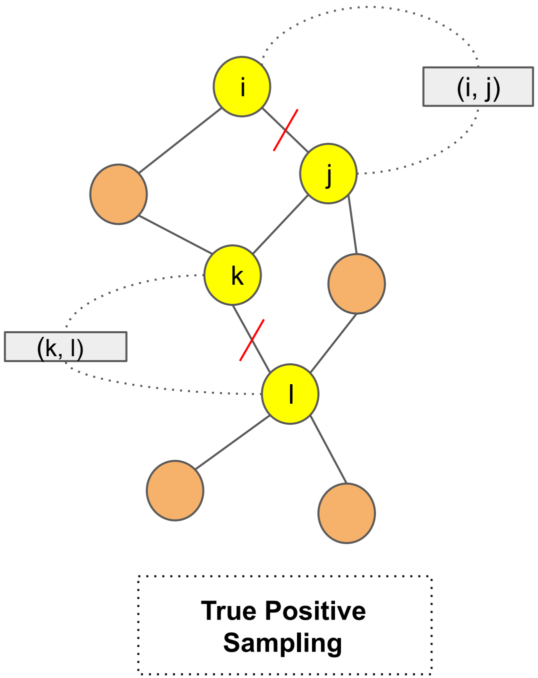
- We take care **when removing edges** to ensure that the **network remains connected**.



True Positive Sampling

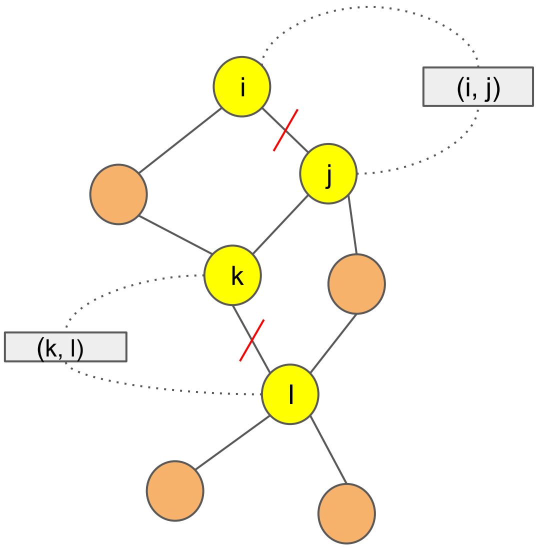
Link Prediction: True Positive & True Negative Edges for Testing

- We take care **when removing edges** to ensure that the **network remains connected**.
- The **network embedding method** is applied **after removing the true positive edges**.

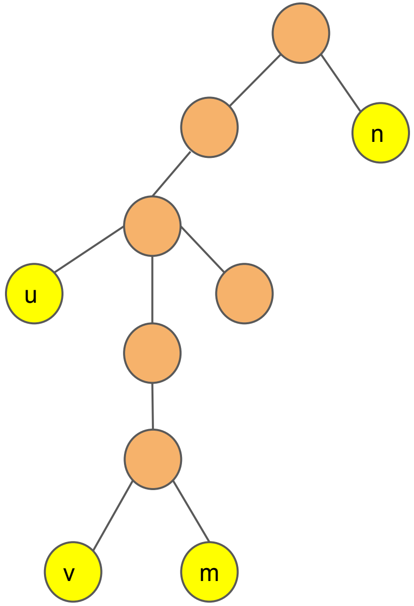


Link Prediction: True Positive & True Negative Edges for Testing

- We take care **when removing edges** to ensure that the **network remains connected**.
- The **network embedding method** is applied **after removing the true positive edges**.



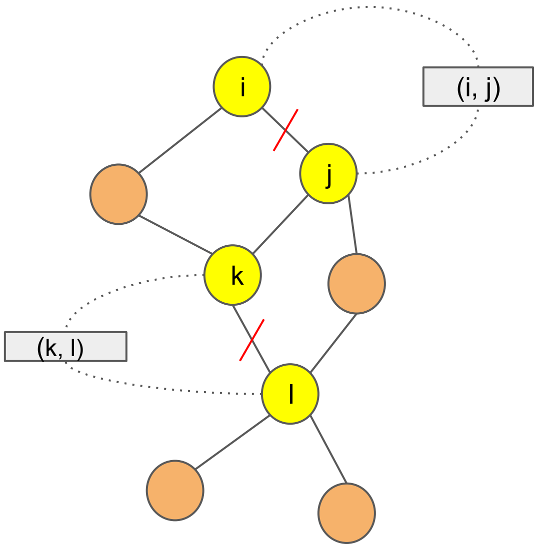
True Positive Sampling



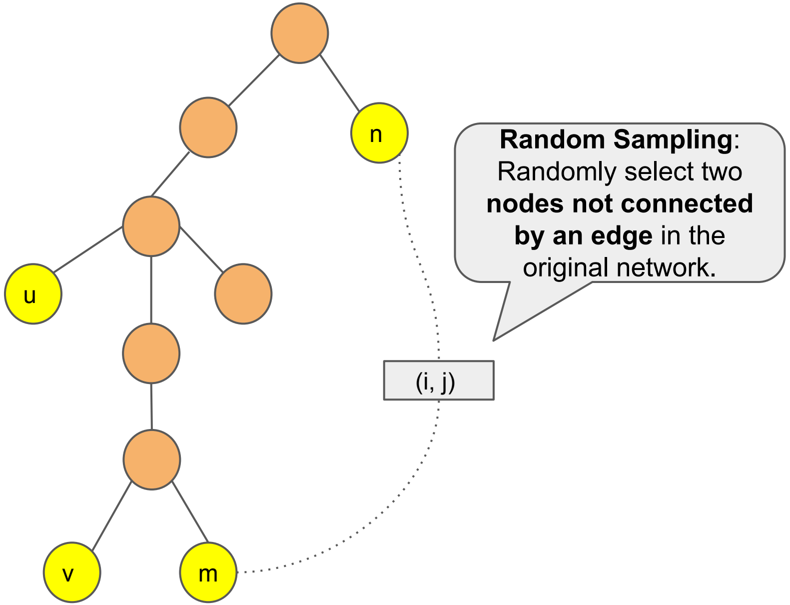
True Negative Sampling with **Random Sampling** or **Close Proximity Sampling**

Link Prediction: True Positive & True Negative Edges for Testing

- We take care **when removing edges** to ensure that the **network remains connected**.
- The **network embedding method** is applied **after removing the true positive edges**.



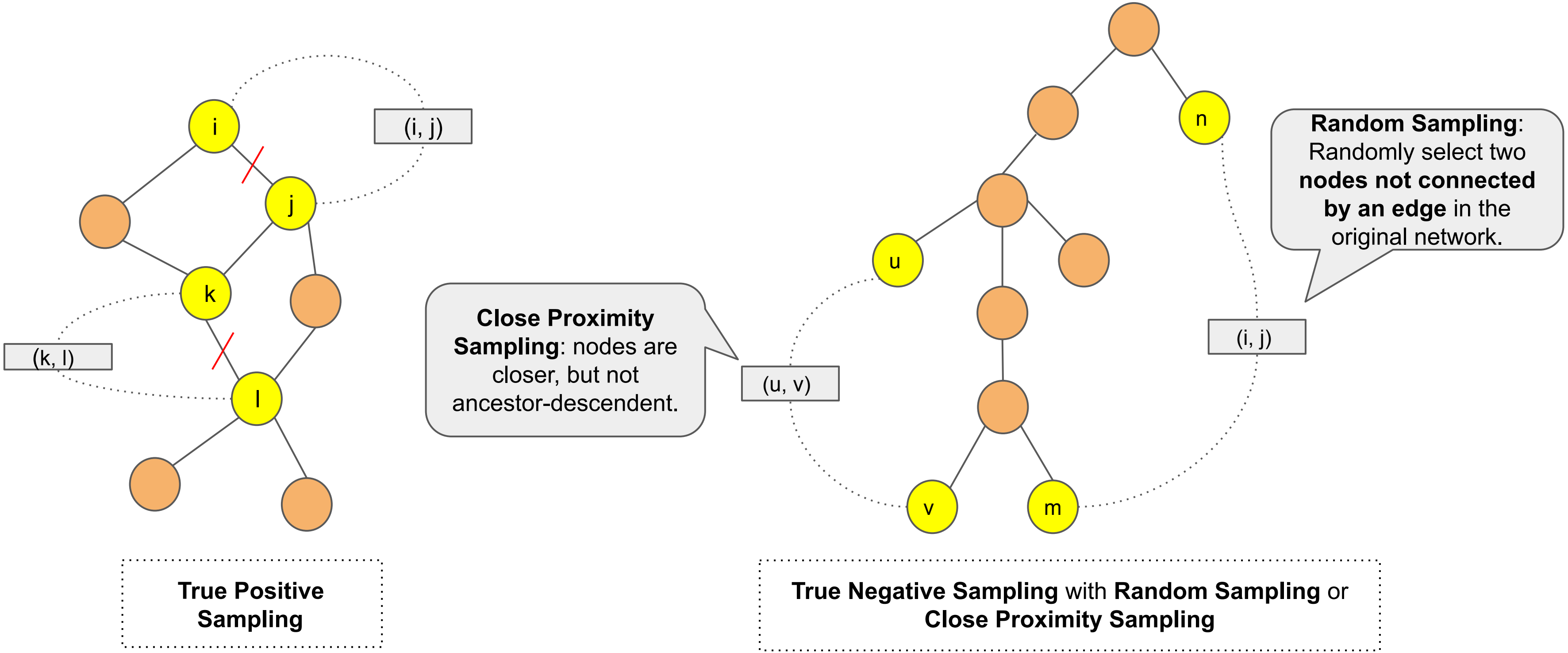
True Positive Sampling



True Negative Sampling with **Random Sampling** or **Close Proximity Sampling**

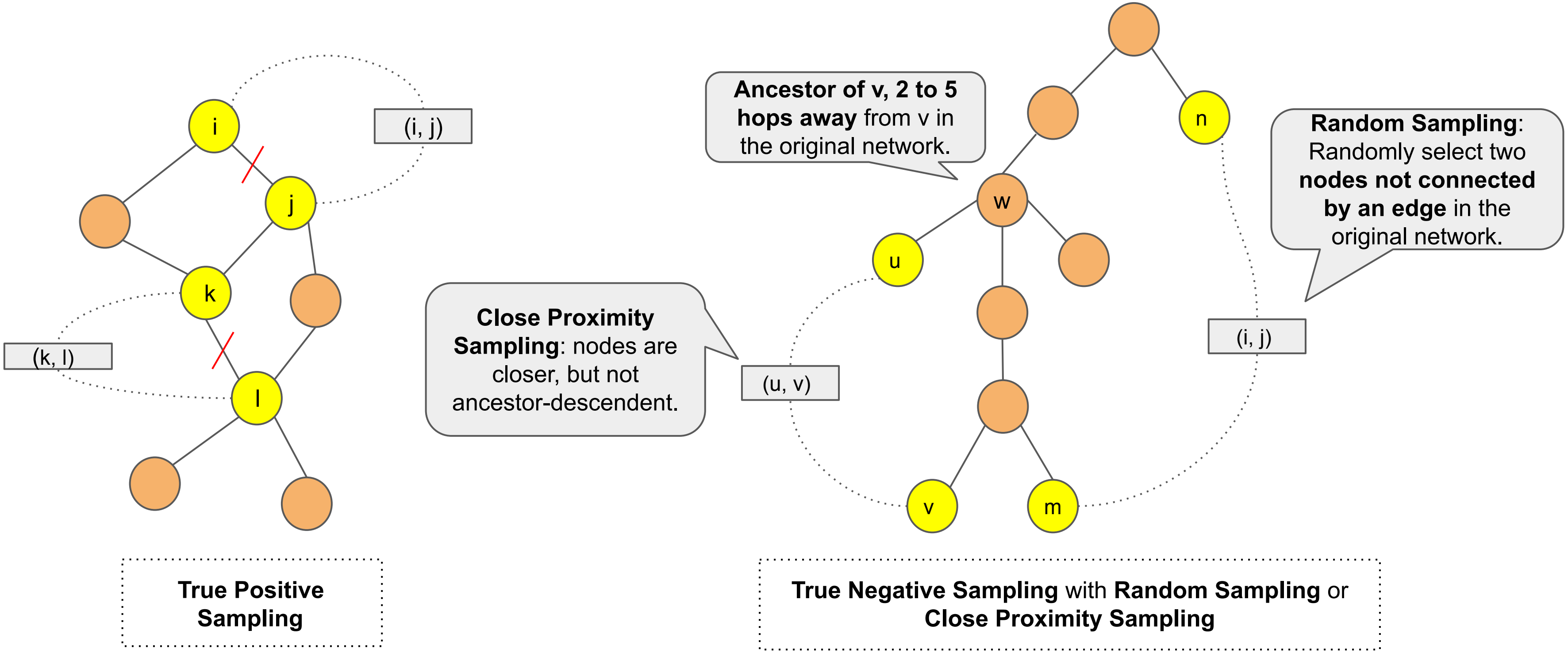
Link Prediction: True Positive & True Negative Edges for Testing

- We take care **when removing edges** to ensure that the **network remains connected**.
- The **network embedding method** is applied **after removing the true positive edges**.



Link Prediction: True Positive & True Negative Edges for Testing

- We take care **when removing edges** to ensure that the **network remains connected**.
- The **network embedding method** is applied **after removing the true positive edges**.

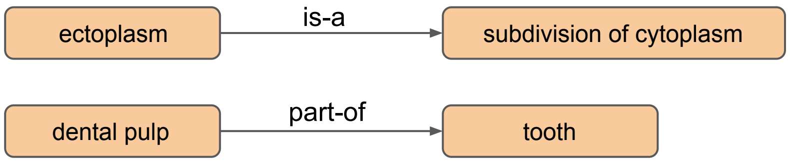


Datasets

- **Two semantic networks** (undirected) extracted from **UMLS**: **PART-OF** and **IS-A.**

Dataset	**IS-A** (large)	**PART-OF** (small)
Nodes	294,693	16,894
Edges	356,541	19,436
Avg. Descriptor Length	5 words	6 words
Max. Descriptor Length	31 words	14 words

- **Examples**:



- **Other biomedical ontologies** could also be used (e.g., Gene Ontology, Disease Ontology).

Link Prediction: True Positive & True Negative Edges for Training/Test

- To **test link prediction**, we obtain **true positive and true negative edges** from the network.

Statistics	**IS-A**	**PART-OF**
Nodes	294,693	16,894
Edges	356,541	19,436
Training true positive edges	294,692	16,893
Training true negative edges	294,692	16,893
Test true positive edges	61,849	2,543
Test true negative edges	61,849	2,543
Avg. descriptor length	5 words	6 words
Max. descriptor length	31 words	14 words

Link Prediction: True Positive & True Negative Edges for Training/Test

- To **test link prediction**, we obtain **true positive and true negative edges** from the network.
- For the **Logistic Regression** predictor, where we also need training instances, we also create **training true positive and true negative edges** in the **same manner**.

Statistics	**IS-A**	**PART-OF**
Nodes	294,693	16,894
Edges	356,541	19,436
Training true positive edges	294,692	16,893
Training true negative edges	294,692	16,893
Test true positive edges	61,849	2,543
Test true negative edges	61,849	2,543
Avg. descriptor length	5 words	6 words
Max. descriptor length	31 words	14 words

Link Prediction: True Positive & True Negative Edges for Training/Test

- To **test link prediction**, we obtain **true positive and true negative edges** from the network.
- For the **Logistic Regression** predictor, where we also need training instances, we also create **training true positive and true negative edges** in the **same manner**.
- In all cases, we use an **equal number** of **true positive** and **true negative edges**.

Statistics	**IS-A**	**PART-OF**
Nodes	294,693	16,894
Edges	356,541	19,436
Training true positive edges	294,692	16,893
Training true negative edges	294,692	16,893
Test true positive edges	61,849	2,543
Test true negative edges	61,849	2,543
Avg. descriptor length	5 words	6 words
Max. descriptor length	31 words	14 words

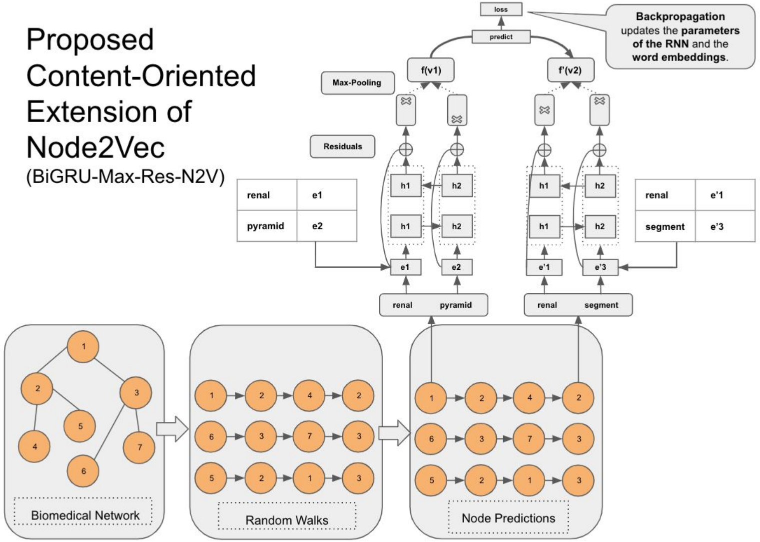
Methods Compared

- **Previous** network embedding methods:
  - **Structure-Oriented: Node2Vec** (Grover and Leskovec, 2016)
  - **Content-Oriented: CANE** (Tu et al., 2017) learns **separate text-based and network-based embeddings of nodes**, considers **only direct neighbors** (unlike Node2Vec).

Methods Compared

- **Previous** network embedding methods:
    - **Structure-Oriented: Node2Vec** (Grover and Leskovec, 2016)
    - **Content-Oriented: CANE** (Tu et al., 2017) learns **separate text-based and network-based embeddings of nodes**, considers **only direct neighbors** (unlike Node2Vec).
- **Our** methods:
    - **BiGRU-Max-Res-N2V**: Our **full content-oriented extension of Node2Vec**, as discussed.
    - **GRU-N2V**: Simpler form of our method, with a **unidirectional GRU** encoder, **no max-pooling**, **no residuals** (node embedding is the GRU's **last state**).
    - **AVG-N2V**: Even simpler, **no RNN encoder**, node embedding is **average of the word embeddings**.

Proposed Content-Oriented Extension of Node2Vec
(BiGRU-Max-Res-N2V)



Link Prediction Results

NE Method + Link Predictor Is-a Dataset	Random Negative Sampling	Close Proximity Sampling (More Difficult)
Node2Vec + CS	66.6	54.3
CANE + CS	94.1	69.6
AVG-N2V + CS	95.0	78.6
GRU-N2V + CS	**98.7**	**79.2**
BiGRU-Max-Res-N2V + CS	98.5	79.0
Node2Vec + LR	77.2	56.3
CANE + LR	95.3	70.0
AVG-N2V + LR	97.6	73.9
GRU-N2V + LR	99.0	79.6
BiGRU-Max-Res-N2V + LR	**99.3**	**82.1**

NE Method + Link Predictor Part-of Dataset	Random Negative Sampling	Close Proximity Sampling (More Difficult)
Node2Vec + CS	76.8	61.8
CANE + CS	93.9	75.3
AVG-N2V + CS	95.9	81.8
GRU-N2V + CS	98.0	83.1
BiGRU-Max-Res-N2V + CS	**98.5**	**83.3**
Node2Vec + LR	85.2	66.5
CANE + LR	94.4	76.3
AVG-N2V + LR	97.6	79.4
GRU-N2V + LR	99.0	85.6
BiGRU-Max-Res-N2V + LR	**99.5**	**88.6**

Link Prediction Results

Content-oriented methods **outperform** the **structure-oriented Node2Vec**. Modeling **textual descriptors** along with network structure **improves results**.

NE Method + Link Predictor Is-a Dataset	Random Negative Sampling	Close Proximity Sampling (More Difficult)
Node2Vec + CS	66.6	54.3
CANE + CS	94.1	69.6
AVG-N2V + CS	95.0	78.6
GRU-N2V + CS	**98.7**	**79.2**
BiGRU-Max-Res-N2V + CS	98.5	79.0
Node2Vec + LR	77.2	56.3
CANE + LR	95.3	70.0
AVG-N2V + LR	97.6	73.9
GRU-N2V + LR	99.0	79.6
BiGRU-Max-Res-N2V + LR	**99.3**	**82.1**

NE Method + Link Predictor Part-of Dataset	Random Negative Sampling	Close Proximity Sampling (More Difficult)
Node2Vec + CS	76.8	61.8
CANE + CS	93.9	75.3
AVG-N2V + CS	95.9	81.8
GRU-N2V + CS	98.0	83.1
BiGRU-Max-Res-N2V + CS	**98.5**	**83.3**
Node2Vec + LR	85.2	66.5
CANE + LR	94.4	76.3
AVG-N2V + LR	97.6	79.4
GRU-N2V + LR	99.0	85.6
BiGRU-Max-Res-N2V + LR	**99.5**	**88.6**

Link Prediction Results

Close Proximity Negative Sampling produces more **difficult link prediction datasets** than Random Negative Sampling.

Content-oriented methods **outperform** the **structure-oriented Node2Vec**. Modeling **textual descriptors** along with network structure **improves results**.

NE Method + Link Predictor **Is-a Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	66.6	54.3
CANE + CS	94.1	69.6
AVG-N2V + CS	95.0	78.6
GRU-N2V + CS	**98.7**	**79.2**
BiGRU-Max-Res-N2V + CS	98.5	79.0
Node2Vec + LR	77.2	56.3
CANE + LR	95.3	70.0
AVG-N2V + LR	97.6	73.9
GRU-N2V + LR	99.0	79.6
BiGRU-Max-Res-N2V + LR	**99.3**	**82.1**

NE Method + Link Predictor **Part-of Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	76.8	61.8
CANE + CS	93.9	75.3
AVG-N2V + CS	95.9	81.8
GRU-N2V + CS	98.0	83.1
BiGRU-Max-Res-N2V + CS	**98.5**	**83.3**
Node2Vec + LR	85.2	66.5
CANE + LR	94.4	76.3
AVG-N2V + LR	97.6	79.4
GRU-N2V + LR	99.0	85.6
BiGRU-Max-Res-N2V + LR	**99.5**	**88.6**

Link Prediction Results

Content-oriented methods **outperform** the **structure-oriented Node2Vec**. Modeling **textual descriptors** along with network structure **improves results**.

Our content-oriented extensions of Node2Vec outperform CANE in all settings. Important to **model text and as much non-local structure as possible**.

NE Method + Link Predictor **Is-a Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	66.6	54.3
CANE + CS	94.1	69.6
AVG-N2V + CS	95.0	78.6
GRU-N2V + CS	**98.7**	**79.2**
BiGRU-Max-Res-N2V + CS	98.5	79.0
Node2Vec + LR	77.2	56.3
CANE + LR	95.3	70.0
AVG-N2V + LR	97.6	73.9
GRU-N2V + LR	99.0	79.6
BiGRU-Max-Res-N2V + LR	**99.3**	**82.1**

NE Method + Link Predictor **Part-of Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	76.8	61.8
CANE + CS	93.9	75.3
AVG-N2V + CS	95.9	81.8
GRU-N2V + CS	98.0	83.1
BiGRU-Max-Res-N2V + CS	**98.5**	**83.3**
Node2Vec + LR	85.2	66.5
CANE + LR	94.4	76.3
AVG-N2V + LR	97.6	79.4
GRU-N2V + LR	99.0	85.6
BiGRU-Max-Res-N2V + LR	**99.5**	**88.6**

Link Prediction Results

Content-oriented methods **outperform** the **structure-oriented Node2Vec**. Modeling **textual descriptors** along with network structure **improves results**.

Our content-oriented extensions of Node2Vec outperform CANE in all settings. Important to **model text and as much non-local structure as possible**.

BiGRU-Max-Res-N2V is **best overall**, but **small differences from GRU-N2V**. **Even AVG-N2V outperforms CANE**.

NE Method + Link Predictor **Is-a Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	66.6	54.3
CANE + CS	94.1	69.6
AVG-N2V + CS	95.0	78.6
GRU-N2V + CS	**98.7**	**79.2**
BiGRU-Max-Res-N2V + CS	98.5	79.0
Node2Vec + LR	77.2	56.3
CANE + LR	95.3	70.0
AVG-N2V + LR	97.6	73.9
GRU-N2V + LR	99.0	79.6
BiGRU-Max-Res-N2V + LR	**99.3**	**82.1**

NE Method + Link Predictor **Part-of Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	76.8	61.8
CANE + CS	93.9	75.3
AVG-N2V + CS	95.9	81.8
GRU-N2V + CS	98.0	83.1
BiGRU-Max-Res-N2V + CS	**98.5**	**83.3**
Node2Vec + LR	85.2	66.5
CANE + LR	94.4	76.3
AVG-N2V + LR	97.6	79.4
GRU-N2V + LR	99.0	85.6
BiGRU-Max-Res-N2V + LR	**99.5**	**88.6**

Link Prediction Results

Best results of LR predictor better than best results of CS predictor. LR can assign a **different weight to each dimension** of the node embeddings.

Content-oriented methods **outperform** the **structure-oriented Node2Vec**. Modeling **textual descriptors** along with network structure **improves results**.

Our content-oriented extensions of Node2Vec outperform CANE in all settings. Important to **model text and as much non-local structure as possible**.

BiGRU-Max-Res-N2V is **best overall,** but **small differences from GRU-N2V**. **Even AVG-N2V outperforms CANE**.

NE Method + Link Predictor **Is-a Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	66.6	54.3
CANE + CS	94.1	69.6
AVG-N2V + CS	95.0	78.6
GRU-N2V + CS	**98.7**	**79.2**
BiGRU-Max-Res-N2V + CS	98.5	79.0
Node2Vec + LR	77.2	56.3
CANE + LR	95.3	70.0
AVG-N2V + LR	97.6	73.9
GRU-N2V + LR	99.0	79.6
BiGRU-Max-Res-N2V + LR	**99.3**	**82.1**

NE Method + Link Predictor **Part-of Dataset**	**Random Negative Sampling**	**Close Proximity Sampling (More Difficult)**
Node2Vec + CS	76.8	61.8
CANE + CS	93.9	75.3
AVG-N2V + CS	95.9	81.8
GRU-N2V + CS	98.0	83.1
BiGRU-Max-Res-N2V + CS	**98.5**	**83.3**
Node2Vec + LR	85.2	66.5
CANE + LR	94.4	76.3
AVG-N2V + LR	97.6	79.4
GRU-N2V + LR	99.0	85.6
BiGRU-Max-Res-N2V + LR	**99.5**	**88.6**

Qualitative Analysis (More examples/discussion in the paper)

Nearest neighbors using our best NE method

Target Node: Left Eyeball (PART-OF)		
Most Similar Embeddings	**Cos**	**Hops**
equator of left eyeball	99.3	1
episcleral layer of left eyeball	99.2	4
cavity of left eyeball	99.1	1
wall of left eyeball	99.0	1
vascular layer of left eyeball	98.9	1

Target Node: Lung Carcinoma (IS-A)		
Most Similar Embeddings	**Cos**	**Hops**
recurrent lung carcinoma	97.6	1
papillary carcinoma	97.1	2
lung pleomorphic carcinoma	97.0	3
ureter carcinoma	96.6	2
lymphoepithelioma-like lung carcinoma	96.6	3

Qualitative Analysis (More examples/discussion in the paper)

Nearest neighbors using our best NE method

Target Node: Left Eyeball (PART-OF)		
Most Similar Embeddings	**Cos**	**Hops**
equator of left eyeball	99.3	1
episcleral layer of left eyeball	99.2	4
cavity of left eyeball	99.1	1
wall of left eyeball	99.0	1
vascular layer of left eyeball	98.9	1

Target Node: Lung Carcinoma (IS-A)		
Most Similar Embeddings	**Cos**	**Hops**
recurrent lung carcinoma	97.6	1
papillary carcinoma	97.1	2
lung pleomorphic carcinoma	97.0	3
ureter carcinoma	96.6	2
lymphoepithelioma-like lung carcinoma	96.6	3

Cosine similarities of true positive edges (IS-A, IS-A, PART-OF)

Edges/Descriptors	BN2V	CANE	N2V	Hops
(a) bariatric surgery (b) bypass gastrojejunostomy	82.7	38.0	56.2	11
(a) anatomical line (b) anterior malleolar fold	82.3	29.0	50.0	22
(a) zone of biceps brachii (b) short head of biceps brachii	93.0	70.0	61.6	13

Conclusions and Future Work

Conclusions

- **Extension of Node2Vec** that apart from **network structure** also considers the **texts of the nodes**.
- **Link prediction experiments** show the proposed method significantly **outperforms Node2Vec** (network structure only) and **CANE** (node texts and network structure, but only direct neighbors).
- **Important to model** both the **texts of the nodes and wide neighborhoods** (not just direct edges).

Conclusions and Future Work

Conclusions

- **Extension of Node2Vec** that apart from **network structure** also considers the **texts of the nodes**.
- **Link prediction experiments** show the proposed method significantly **outperforms Node2Vec** (network structure only) and **CANE** (node texts and network structure, but only direct neighbors).
- **Important to model** both the **texts of the nodes and wide neighborhoods** (not just direct edges).

Future Work

- Experiments with networks extracted from **other biomedical ontologies and knowledge bases** (e.g., **causal networks** for knowledge discovery).
- Explore if the **word embeddings** that our methods generate f**rom semantic networks** improve **biomedical question answering**, when combined with **word embeddings from text corpora**.

Thanks!

Source Code and Datasets

https://github.com/SotirisKot/Content-Aware-Node2Vec

For more papers

http://nlp.cs.aueb.gr/publications.html