MAYO CLINIC

Constructing Co-occurrence Network Embeddings to Assist Association Extraction for COVID-19 and Other Coronavirus Infectious Diseases

Feichen Shen, Ph.D.
Assistant Professor of Biomedical Informatics
Division of Digital Health Sciences

David Oniani
R&D Intern
Kern Center for the Science of Health Care Delivery

BACKGROUND
-COVID-19

- An infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

- As of today, more than 8 million cases across 188 countries and territories, resulting in more than 435,000 deaths.

MAYO CLINIC

https://en.wikipedia.org/wiki/Coronavirus_disease_2019

BACKGROUND
-CORD-19

- The White House and leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19).
- Over 138,000 scholarly articles, including over 69,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses.

MAYO CLINIC

https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

BACKGROUND
-CORD-19-ON-FHIR

- A linked data version of CORD-19.
- Incorporate annotations from different sources, such as PubTator and LitCovid.
- Easy to retrieve co-occurrences among entities by using SPARQL.

MAYO CLINIC

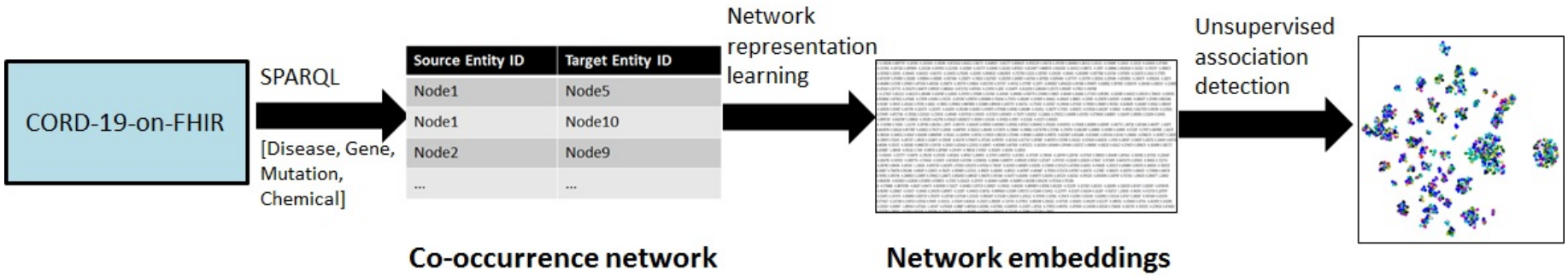
https://github.com/fhircat/CORD-19-on-FHIR

OBJECTIVE

- Co-occurrence network
- Network embeddings
- Accelerate knowledge discovery for COVID-19

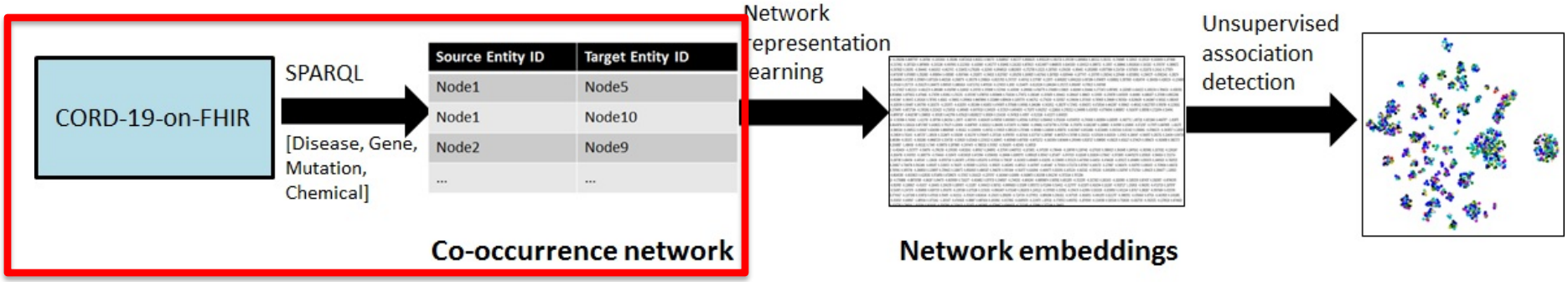
MAYO CLINIC

METHODS
-OVERVIEW



MAYO CLINIC

METHODS



MAYO CLINIC

METHODS

-CO-OCCURRENCE NETWORK

```
pmc:annotations [
        pmc:id "1" ;
                pmc:infons [ pmc:identifier "MESH:D003371" ; pmc:type "Disease" ] ;
                pmc:locations [ pmc:length "5"^^xsd:int ; pmc:offset "20312"^^xsd:int ] ;
                pmc:text "cough" ],
        pmc:id "2" ;
                pmc:infons [ pmc:identifier "MESH:C000657245" ; pmc:type "Disease" ] ;
                pmc:locations [ pmc:length "19"^^xsd:int ; pmc:offset "14766"^^xsd:int ] ;
                pmc:text "2019-nCoV infection" ],],
pmc:annotations [
        pmc:id "5" ;
                pmc:infons [ pmc:identifier "59272" ; pmc:ncbi_homologene "41448" ; pmc:type "Gene" ] ;
                pmc:locations [ pmc:length "31"^^xsd:int ; pmc:offset "1986"^^xsd:int ] ;
                pmc:text "angiotensin-converting enzyme 2" ],
        pmc:id "7" ;
                pmc:infons [ pmc:identifier "MESH:C000657245" ; pmc:type "Disease" ] ;
                pmc:locations [ pmc:length "19"^^xsd:int ; pmc:offset "14766"^^xsd:int ] ;
                pmc:text "2019-nCoV infection" ],],
```

Disease
Gene
Mutation
Chemical

METHODS

-CO-OCCURRENCE NETWORK

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX fhir: <http://hl7.org/fhir/>
PREFIX pmc: <https://www.ncbi.nlm.nih.gov/pmc/articles#>
SELECT distinct ?pmc_id0 ?text0 ?pmc_id1 ?text1 (count(?text1) as ?count) WHERE {
?pmc pmc:annotations
[ pmc:id ?id0 ; pmc:text ?text0 ; pmc:infons
[ pmc:type ?type0 ; pmc:identifier ?pmc_id0 ] ] .
    FILTER ((?type0 = 'Biomedical_Entity') ) .
    {select * where{
        ?pmc pmc:annotations
        [ pmc:id ?id1 ; pmc:text ?text1 ; pmc:infons
        [ pmc:type ?type1 ; pmc:identifier ?pmc_id1 ] ] .
            FILTER ((?type1='Disease') && (contains (lcase(str(?text1)), "coronavirus") || contains (lcase(str(?text1)), "sars") || contains (lcase(str(?
text1)), "covid-19") || contains (lcase(str(?text1)), "pneumonia") || contains (lcase(str(?text1)), "fever") || contains (lcase(str(?text1)), "fibrosis") ||
 contains (lcase(str(?text1)), "diarrhea") || contains (lcase(str(?text1)), "bronchitis") || contains (lcase(str(?text1)), "ebola") || contains (lcase(str(?
text1)), "influenze") || contains (lcase(str(?text1)), "zika"))).

        }
    }
}Group by ?pmc_id0 ?text0 ?pmc_id1 ?text1 Order by DESC(?count)
```

Disease
Gene
Mutation
Chemical

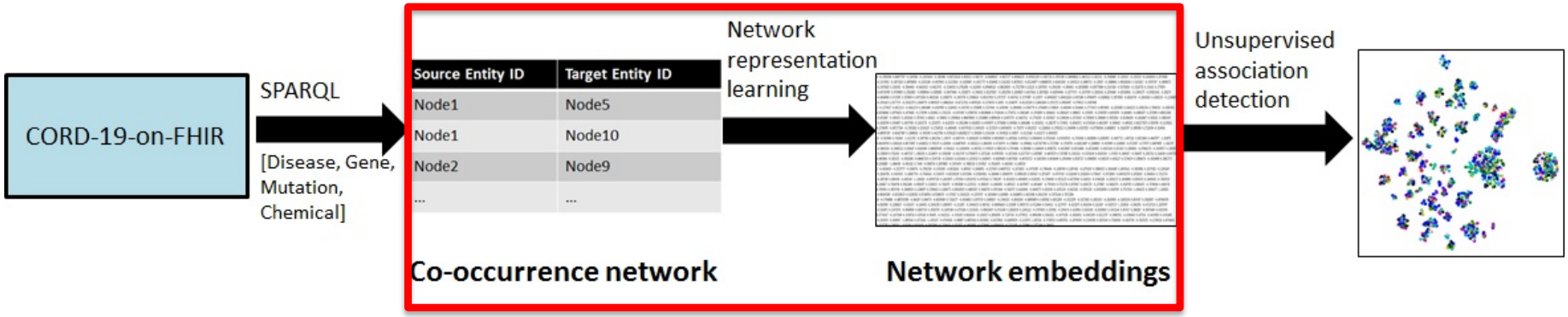
MAYO CLINIC

METHODS
-CO-OCCURRENCE NETWORK

	text0	text1
1	Hu B	pneumonia
2	albumin	fever
3	alanine aminotransferase	fever
4	aspartate aminotransferase	fever
5	SARS	pneumonia
6	SARS	2019-nCoV pneumonia
7	Nam	2019-nCoV pneumonia
8	Sri	2019-nCoV pneumonia
9	bat	coronavirus infection
10	bat	SARS
11	bat	pneumonia
12	ACE2	SARS
13	angiotensin-converting enzyme 2	SARS

MAYO CLINIC

METHODS



MAYO CLINIC

METHODS
-NETWORK EMBEDDINGS

- Similar to word embeddings
- Node→Word, Neighborhood→Context
- However, sliding window for text is not suitable for non-linear graph
- One solution is to use random walk to select "context" in graph

MAYO CLINIC

METHODS
-NETWORK EMBEDDINGS

Node2vec

- Scalable feature learning for networks.
- Grover A, Leskovec J.
- In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016 Aug 13 (pp. 855-864). ACM.

MAYO CLINIC

NODE2VEC

2 steps

- Sampling strategy

Apply random walk on graph to prepare input data

- Node embeddings

Apply word2vec on prepared input data to generate embeddings for node

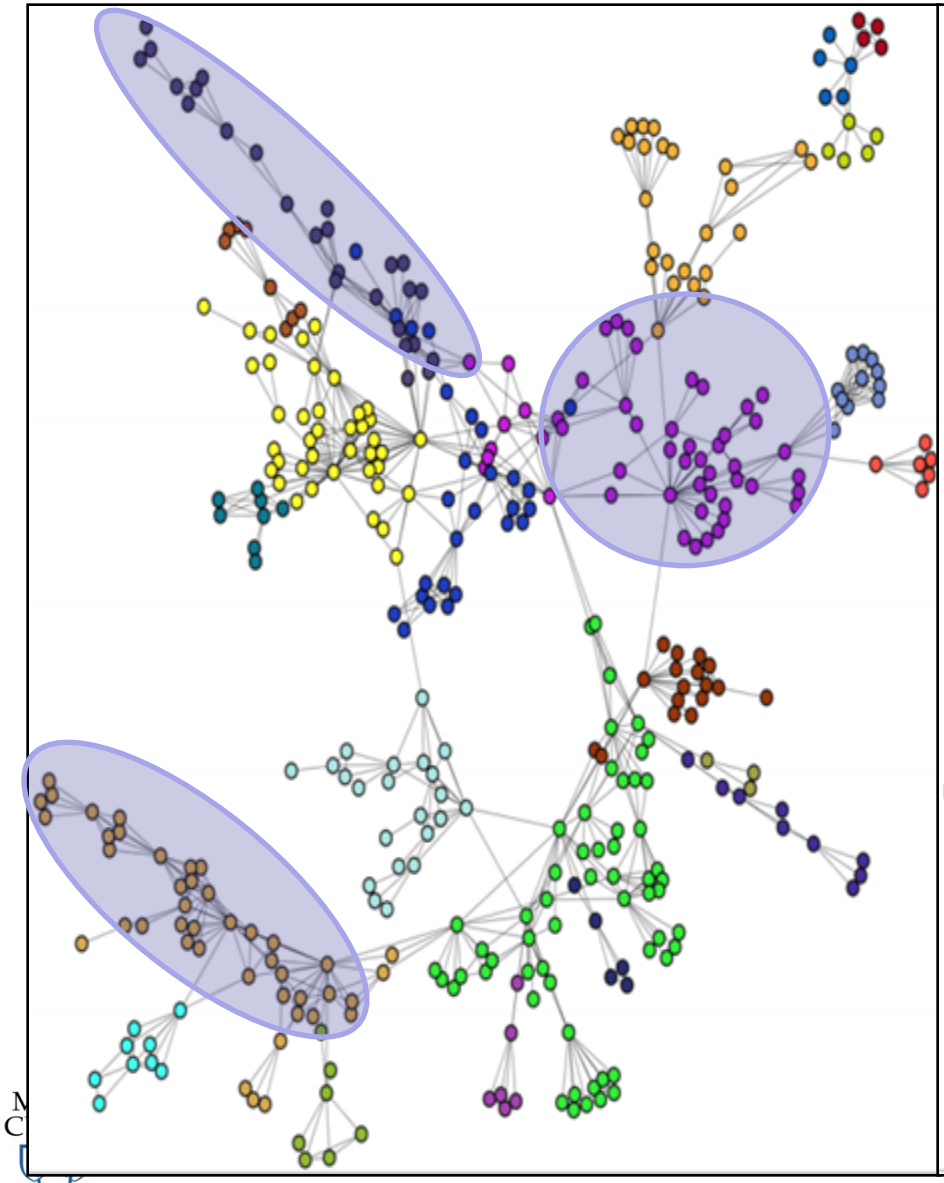
MAYO CLINIC

1. Sampling strategy

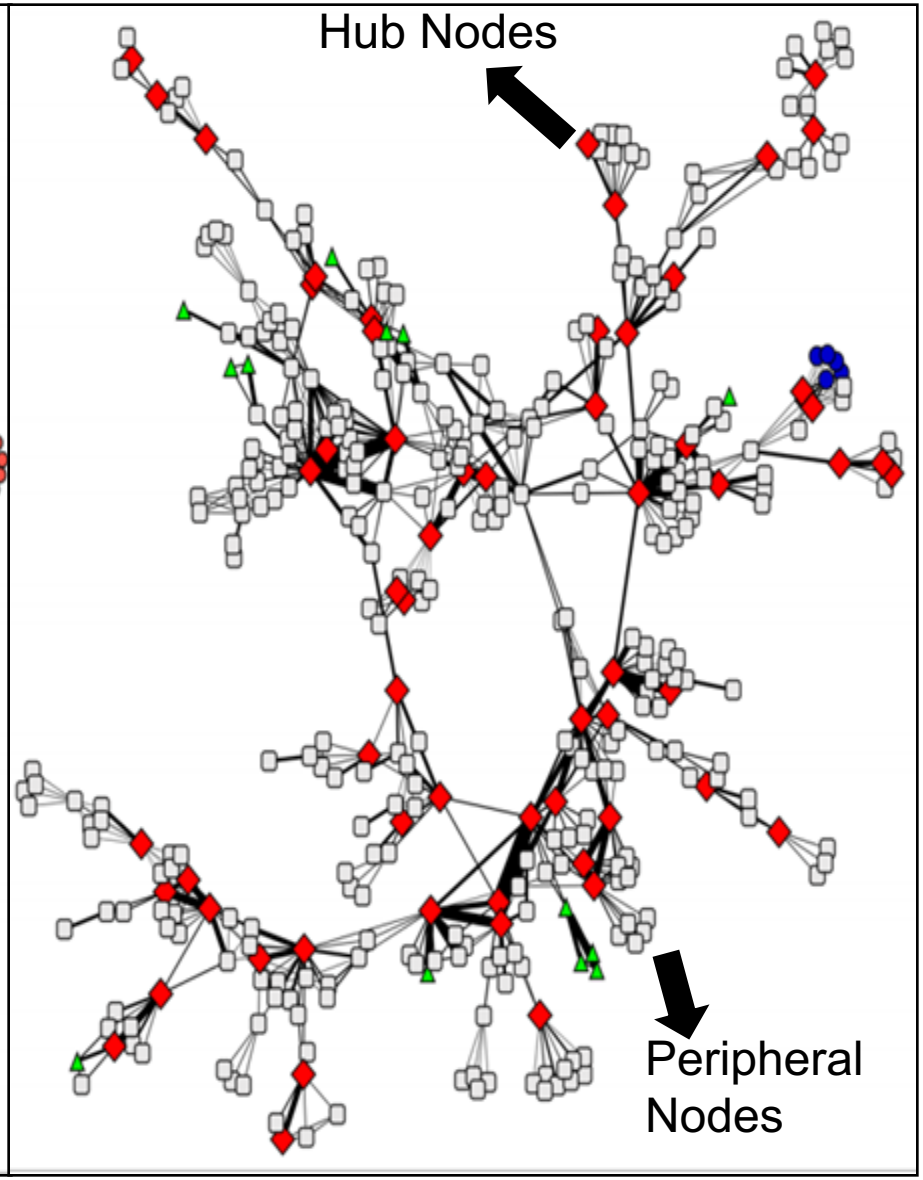
- 1. **Homophily equivalence**: embed nodes from the same network community closely together
- 2. **Structural equivalence**: nodes share similar roles have similar embeddings

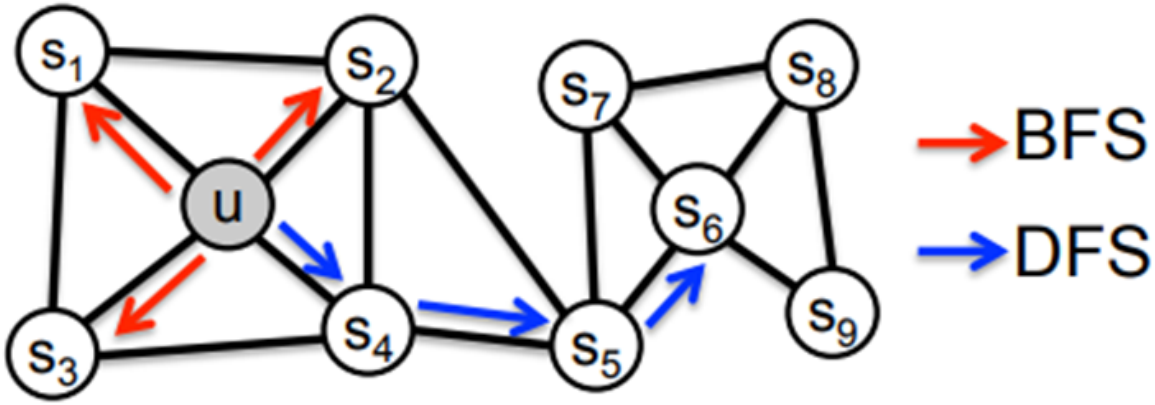
MAYO CLINIC

Homophily equivalence



Structural equivalence

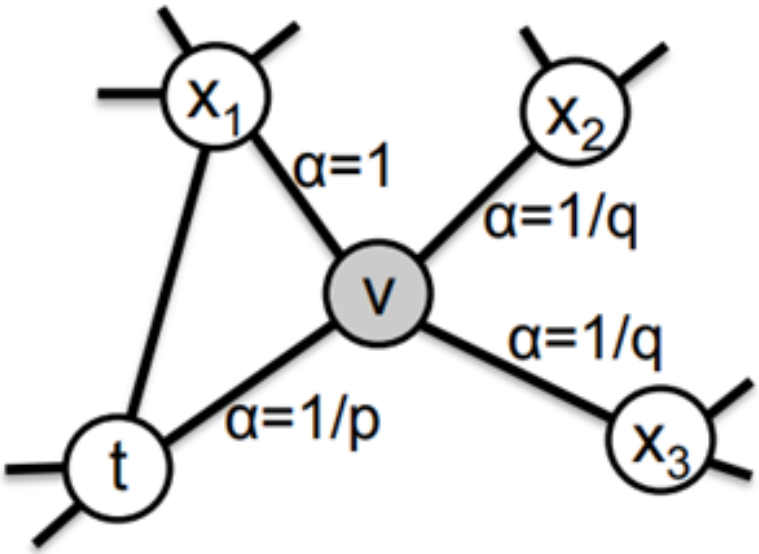




- **Breadth-first search (BFS): structural analysis, microscopic view** of neighborhood of every node
- **Depth-first search (DFS): homophily analysis, macroscopic view** of the neighborhood among different communities
- Real-world network has a lot of such mixture of BFS and DFS

MAYO CLINIC

Grover A, Leskovec J. node2vec: Scalable feature learning for networks. InProceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016 Aug 13 (pp. 855-864)



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

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

- p and q are controller to balance between BFS and DFS
- p controls the return step, and q controls the step of walk to outside world

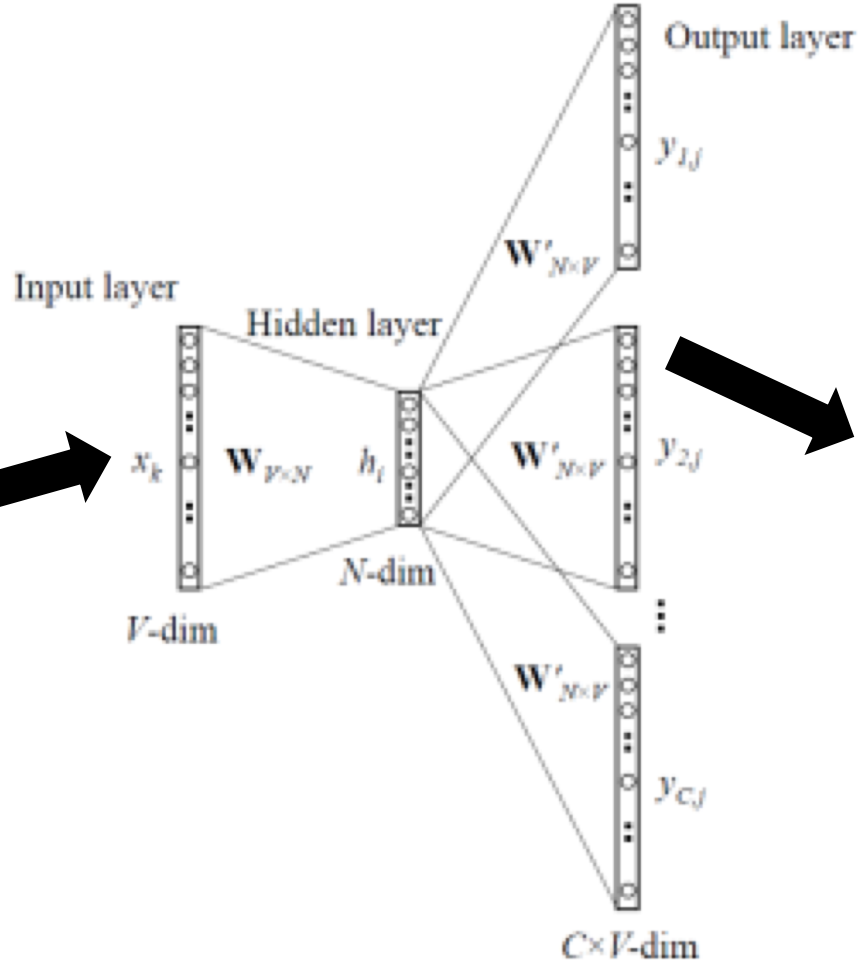
MAYO CLINIC

Grover A, Leskovec J. node2vec: Scalable feature learning for networks. InProceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016 Aug 13 (pp. 855-864)

2. WORD2VEC

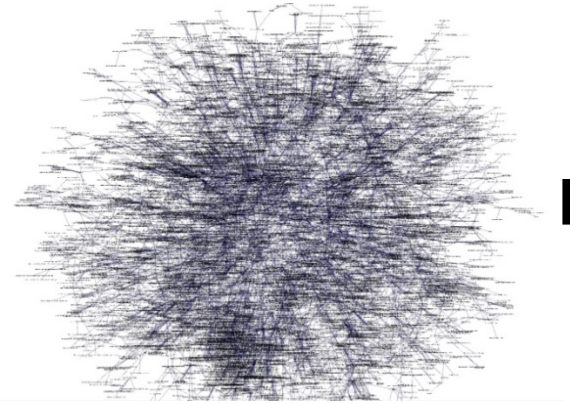
Skip-gram

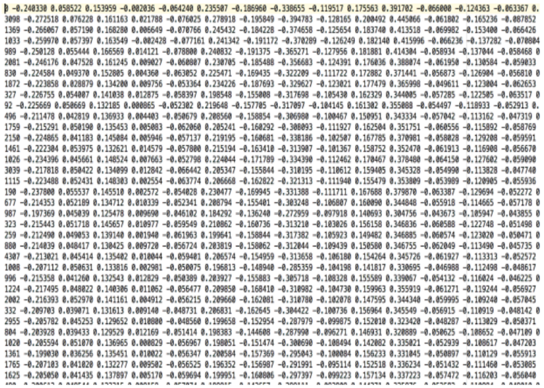
Node u



Neighbor Nodes Ns(u)

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111-3119).



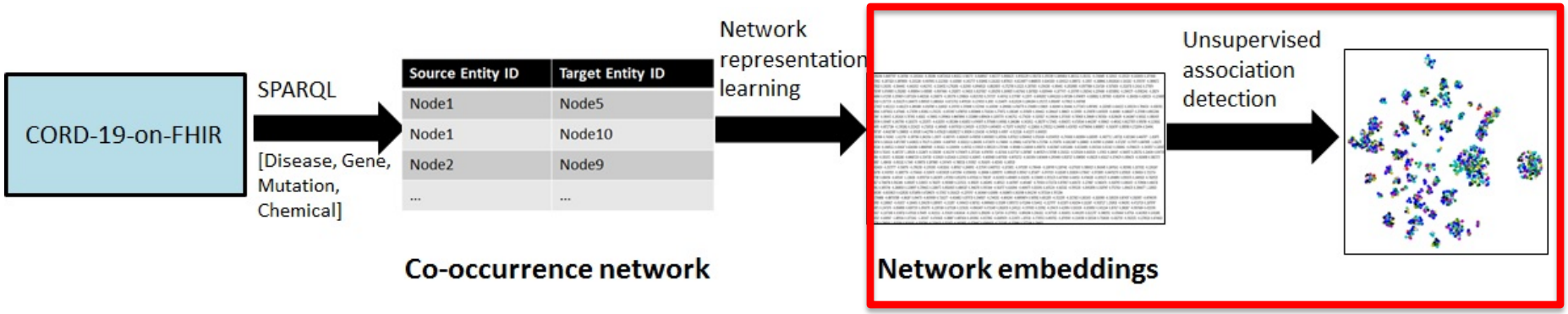


```
0 -0.240330 0.058522 0.153959 -0.002036 -0.064240 0.235507 -0.186960 -0.338655 -0.119517 0.175563 0.391702 -0.066000 -0.124363 -0.063367 0.
3098 -0.272518 0.076228 0.161163 0.021788 -0.076025 0.278918 -0.195849 -0.394783 -0.128165 0.200492 0.445066 -0.061802 -0.165236 -0.087852
1369 -0.266067 0.057190 0.168280 0.006649 -0.070766 0.245432 -0.184228 -0.374658 -0.125654 0.183740 0.413518 -0.069982 -0.153400 -0.066426
1033 -0.259970 0.057397 0.163549 -0.002428 -0.077161 0.241342 -0.191172 -0.370289 -0.126249 0.182140 0.415996 -0.066236 -0.137282 -0.070804
989 -0.250128 0.055444 0.166569 0.014121 -0.078800 0.240832 -0.191375 -0.365271 -0.127956 0.181881 0.414304 -0.058934 -0.137044 -0.058468 0
2081 -0.246176 0.047528 0.161245 0.009027 -0.060807 0.230705 -0.185488 -0.356683 -0.124391 0.176036 0.388074 -0.061950 -0.130584 -0.059033
830 -0.224584 0.049370 0.152805 0.004360 -0.063052 0.225471 -0.169435 -0.322209 -0.111722 0.172882 0.371441 -0.056873 -0.126904 -0.056810 0
1872 -0.223858 0.028879 0.134200 0.009756 -0.053364 0.234226 -0.187693 -0.329627 -0.123021 0.177479 0.365998 -0.049611 -0.123004 -0.062653
327 -0.226755 0.054007 0.141038 0.012875 -0.058397 0.198548 -0.155008 -0.317698 -0.105430 0.162329 0.344005 -0.057285 -0.122505 -0.063517 0
92 -0.225669 0.050669 0.132185 0.000065 -0.052302 0.219648 -0.157705 -0.317097 -0.104145 0.161302 0.355088 -0.054497 -0.118933 -0.052913 0.
496 -0.211478 0.042819 0.136933 0.004403 -0.050679 0.208560 -0.158854 -0.306980 -0.100467 0.150951 0.343334 -0.057042 -0.113162 -0.047319 0
1759 -0.215291 0.050190 0.135453 0.005083 -0.062060 0.205241 -0.160292 -0.308093 -0.111927 0.162504 0.351751 -0.060556 -0.115892 -0.058769
2150 -0.224865 0.041183 0.145084 0.005946 -0.057137 0.219195 -0.160681 -0.338186 -0.102507 0.167785 0.370981 -0.058028 -0.129208 -0.059591
1461 -0.222304 0.053975 0.132621 0.014579 -0.057800 0.215194 -0.163410 -0.313907 -0.101367 0.158752 0.352470 -0.061913 -0.116908 -0.056670
1026 -0.234396 0.045661 0.148524 0.007663 -0.052798 0.224044 -0.171789 -0.334390 -0.112462 0.170467 0.378480 -0.064150 -0.127602 -0.059090
3039 -0.217818 0.050422 0.134099 0.012842 -0.066442 0.205347 -0.155844 -0.310195 -0.110612 0.159405 0.345328 -0.054990 -0.113828 -0.047740
1115 -0.223488 0.052431 0.148303 0.002554 -0.063774 0.206668 -0.162822 -0.321313 -0.111940 0.155479 0.353809 -0.053989 -0.120905 -0.055936
190 -0.237800 0.055537 0.145510 0.002572 -0.054028 0.230477 -0.169945 -0.331388 -0.111711 0.167688 0.379870 -0.063387 -0.129694 -0.052272 0
677 -0.214353 0.052189 0.134712 0.010339 -0.052341 0.208794 -0.155401 -0.303248 -0.106807 0.160090 0.344848 -0.055918 -0.114665 -0.057178 0
987 -0.197369 0.045039 0.125478 0.009690 -0.046102 0.184292 -0.136240 -0.272959 -0.097918 0.140693 0.304756 -0.043673 -0.105947 -0.043855 0
323 -0.215443 0.051718 0.145657 0.010977 -0.059549 0.210062 -0.160736 -0.313210 -0.103026 0.156158 0.346836 -0.060588 -0.122748 -0.051498 0
259 -0.212490 0.049053 0.139140 0.001940 -0.061963 0.199641 -0.158844 -0.317382 -0.105923 0.149482 0.346885 -0.060574 -0.123020 -0.050471 0
880 -0.214039 0.048417 0.130425 0.009720 -0.056724 0.203819 -0.158062 -0.312044 -0.109439 0.150580 0.346755 -0.062049 -0.113490 -0.045735 0
4307 -0.213021 0.045414 0.135402 0.010044 -0.059401 0.206574 -0.154959 -0.313658 -0.106180 0.154264 0.345726 -0.061927 -0.113313 -0.052572
1008 -0.207112 0.050631 0.133816 0.002981 -0.050075 0.196813 -0.148940 -0.285359 -0.104198 0.141817 0.330695 -0.046988 -0.112498 -0.048617
996 -0.215358 0.041260 0.132543 0.012829 -0.050389 0.203927 -0.155883 -0.305718 -0.108328 0.155589 0.339067 -0.054132 -0.116024 -0.046225 0
1224 -0.217495 0.048022 0.140306 0.011062 -0.056477 0.209850 -0.168410 -0.310982 -0.104730 0.159963 0.355919 -0.061271 -0.119244 -0.056927
2002 -0.216393 0.052970 0.141161 0.004912 -0.056215 0.209660 -0.162081 -0.310780 -0.102078 0.147595 0.344340 -0.059995 -0.109240 -0.057045
332 -0.209703 0.039071 0.131613 0.009140 -0.048731 0.206831 -0.162645 -0.304422 -0.100736 0.156964 0.345549 -0.056915 -0.110919 -0.048142 0
2955 -0.205782 0.045253 0.129652 0.010800 -0.048560 0.199658 -0.152954 -0.287979 -0.099875 0.152010 0.323420 -0.048287 -0.113029 -0.050371
804 -0.203928 0.039433 0.129529 0.012169 -0.051414 0.198383 -0.144608 -0.287990 -0.096271 0.146931 0.320889 -0.050625 -0.108652 -0.047109 0
1020 -0.205594 0.051070 0.136965 0.000829 -0.056967 0.198051 -0.151474 -0.300690 -0.108494 0.142082 0.335021 -0.052939 -0.108617 -0.047203
1361 -0.199030 0.036256 0.135451 0.010022 -0.056347 0.200584 -0.157369 -0.295043 -0.100084 0.156233 0.331045 -0.050897 -0.110129 -0.055913
1765 -0.207103 0.041020 0.132277 0.009502 -0.056525 0.196352 -0.156987 -0.291991 -0.095114 0.152518 0.336234 -0.051432 -0.111460 -0.053085
1625 -0.205050 0.041435 0.137897 0.005170 -0.059694 0.199951 -0.160806 -0.297397 -0.099223 0.157134 0.337223 -0.057472 -0.116203 -0.056040
```

Graph space

Embeddings space

METHODS



MAYO CLINIC

METHODS

-NODE CLUSTERING

- t-distributed stochastic neighbor embedding (t-SNE)
- Density-based spatial clustering of applications with noise (DBSCAN)

MAYO CLINIC

RESULTS

- 3,626 diseases
- 5,741 genes
- 524 mutations
- 6,878 chemicals
- 16,769 nodes and 49,696 edges

MAYO CLINIC

RESULTS

Link prediction

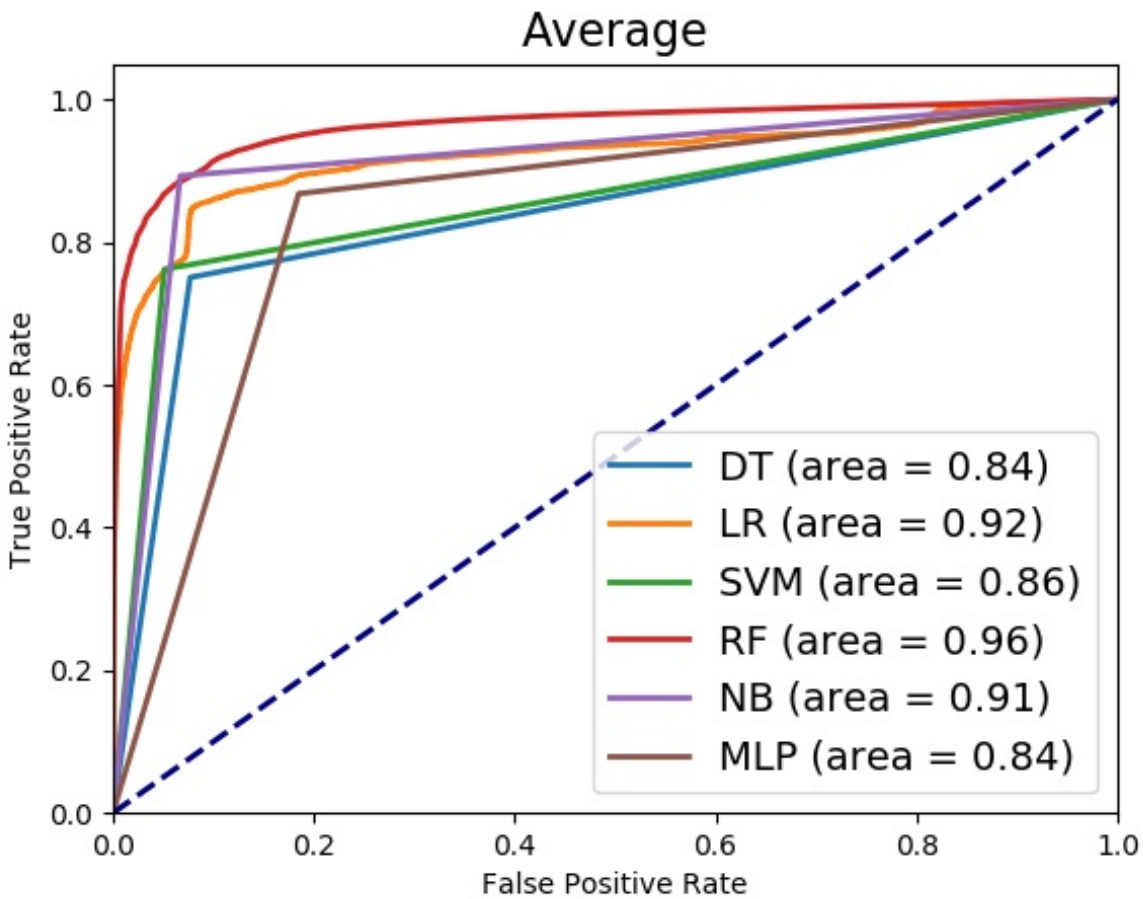
- Generate edge embedding based on node embedding
- Given nodes u and v, f(u) and f(v) denotes their feature representations

Operator	Definition
Average	$\frac{f(u)+f(v)}{2}$
Hadamard	$f(u) * f(v)$
Weighted-L1	$\lvert f(u)-f(v)\rvert$
Weighted-L2	$\lvert f(u)-f(v)\rvert^{2}$

MAYO CLINIC

RESULTS

Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Multi-layer Perceptron (MLP)



MAYO CLINIC

RESULTS

Algorithms	Precision	Recall	F1 score
DT	0.85	0.84	0.84
LR	0.87	0.85	0.85
SVM	0.87	0.86	0.85
RF	**0.91**	**0.91**	0.90
NB	**0.91**	**0.91**	**0.91**
MLP	0.84	0.84	0.84

MAYO CLINIC

RESULTS

COVID-19 (cluster #6)

Top 10 most relevant entities
VP35 (Gene)
HD11 (Gene)
Coronavirus infection process (Disease)
Fibroblast growth factor (FGF)-2 (Gene)
Acute respiratory infection illness (Disease)
PIGS (Gene)
TGF alpha (Gene)
SFPQ (Gene)
Tumour necrosis factor (TNF) (Gene)
Praziquantel (Chemical)

MAYO CLINIC

RESULTS

Pulmonary coronavirus infection (Cluster #1)

Top 10 most relevant entities
PTP (Gene)
SARS-CoV-infected human airway epithelia cell cultures (Disease)
"5'-tgg gat tca aca" (Chemical)
Tracheanasal respiratory epithelial cells nd llamas (lama glama) in (Disease)
Suppressor of cytokine signaling 3 (Gene)
KAT (Gene)
CD32 (Gene)
Maternal SARS infection (Disease)
Respiratory syndrome coronavirus (MERS-CoV) infections (Disease)
S27 (Gene)

MAYO CLINIC

RESULTS

Sars-cov infection damages lung (Cluster #2)

Top 10 most relevant entities
IL-1-alpha (Gene)
Sucralfate prn (Chemical)
Acute respiratory syndrome-cov infection (Disease)
IL-5- and IL-13-producing ilc-iis (Gene)
HAP1 (Gene)
FSK (Chemical)
Low fever (Disease)
HIV and Ebola virus infection (Disease)
YKL-40 (Gene)
ETF (Gene)

MAYO CLINIC

RESULTS

Coronavirus upper respiratory infection (Cluster #23)

Top 10 most relevant entities
Viruses actinobacillus pleuropneumoniae (Disease)
Plasmin (Gene)
JAM-1 (Gene)
TNF receptor-associated factor 6 (Gene)
GPC3 (Gene)
Renin (Gene)
ZO-1 (Gene)
Cathepsin G (Gene)
rs5743313 (Mutation)
Alpha1 antitrypsin (Gene)

MAYO CLINIC

RESULTS

Coronavirus-infected pneumonia (Cluster #10)

Top 10 most relevant entities
Respiratory syncytial viral infection (Disease)
Pegylated interferon-alpha (Chemical)
IFITM6 (Gene)
Feline b (Chemical)
E119V (Mutation)
Epac2 (Gene)
GFTP2 (Gene)
Hepatitis coronavirus infection (Disease)
Ouabain (Chemical)
LY6G (Gene)

MAYO CLINIC

CONCLUSION

- The construction of co-occurrence network embeddings for COVID-19 and related coronavirus infectious diseases.
- Published in the Journal of the American Medical Informatics Association. 2020 May 27.
- https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocaa117/5847598
- https://github.com/shenfc/COVID-19-network-embeddings
- https://www.davidoniani.com/covid-19-network

MAYO CLINIC

FUTURE WORK

- Adding more keywords to SPARQL query
- Including weights over edge for training purpose
- Manual evaluation by clinical investigators

MAYO CLINIC

ACKNOWLEDGEMENT

- CORD-19, CORD-19-on-FHIR
- David Oniani, Guoqian Jiang, Hongfang Liu
- National Institute of Health (NIH) U01TR0062-1.

MAYO CLINIC