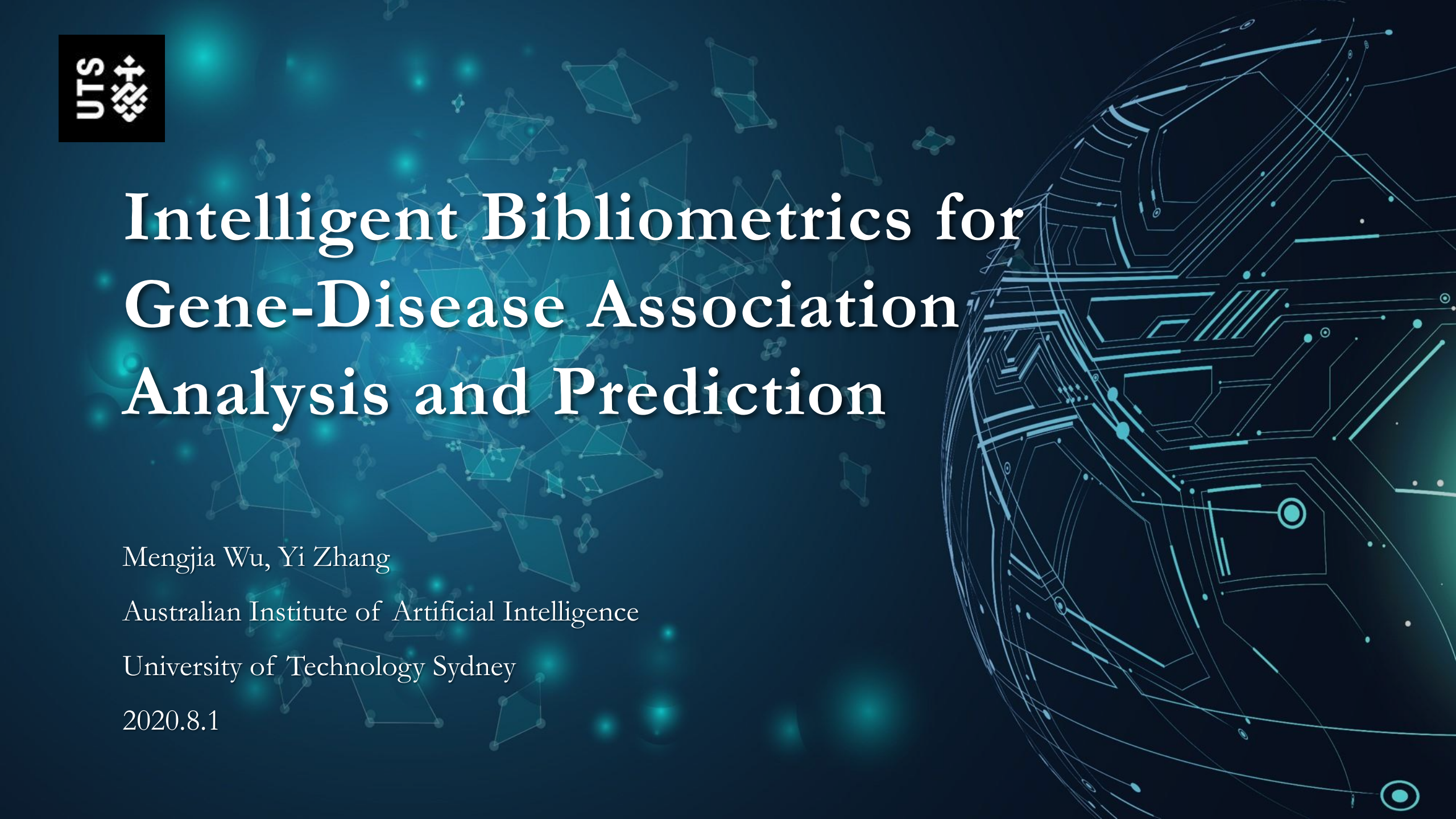


Intelligent Bibliometrics for Gene-Disease Association Analysis and Prediction

Mengjia Wu, Yi Zhang

Australian Institute of Artificial Intelligence

University of Technology Sydney

2020.8.1

1. Research Motivation

Genetic Analysis for Disease: occurrence, diagnosis and treatment

Data-driven Disease-Gene Association Prediction:

- Curated Databases – limited knowledge within established frameworks
- Literature Based Discovery (LBD) – the requirement of expert knowledge

- Propose an adaptable and automatic LBD approach for the following tasks:

1 How to identify the crucial genetic entities for a specific disease.

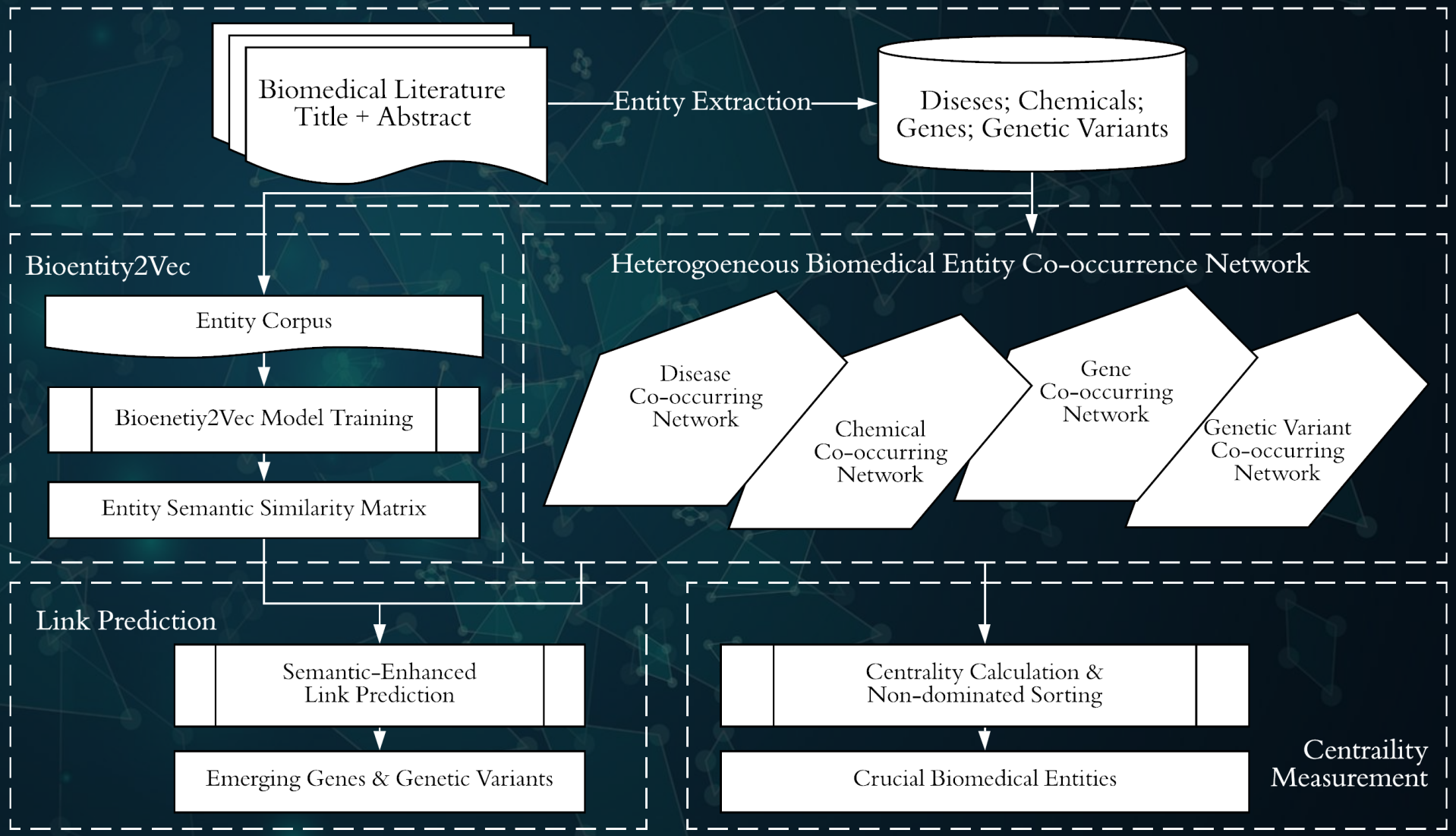
2 How to predict emerging genetic factors for the target disease.

2. Methodology Framework

Stage 1
Data Collection and Pre-processing

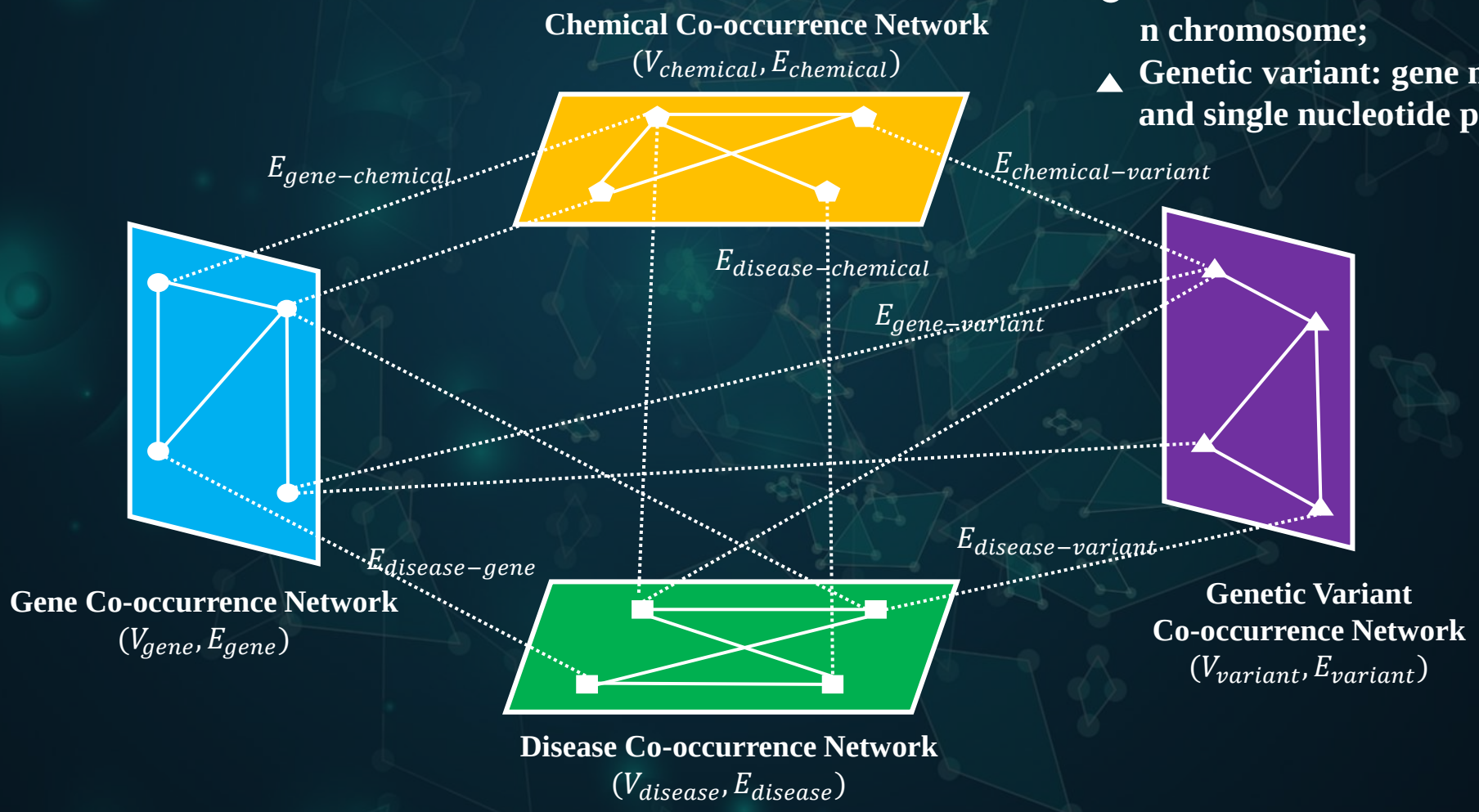
Stage 2
Bioentity2Vec Training and Network Construction

Stage 3
Network Analytics



2. Methodology Framework

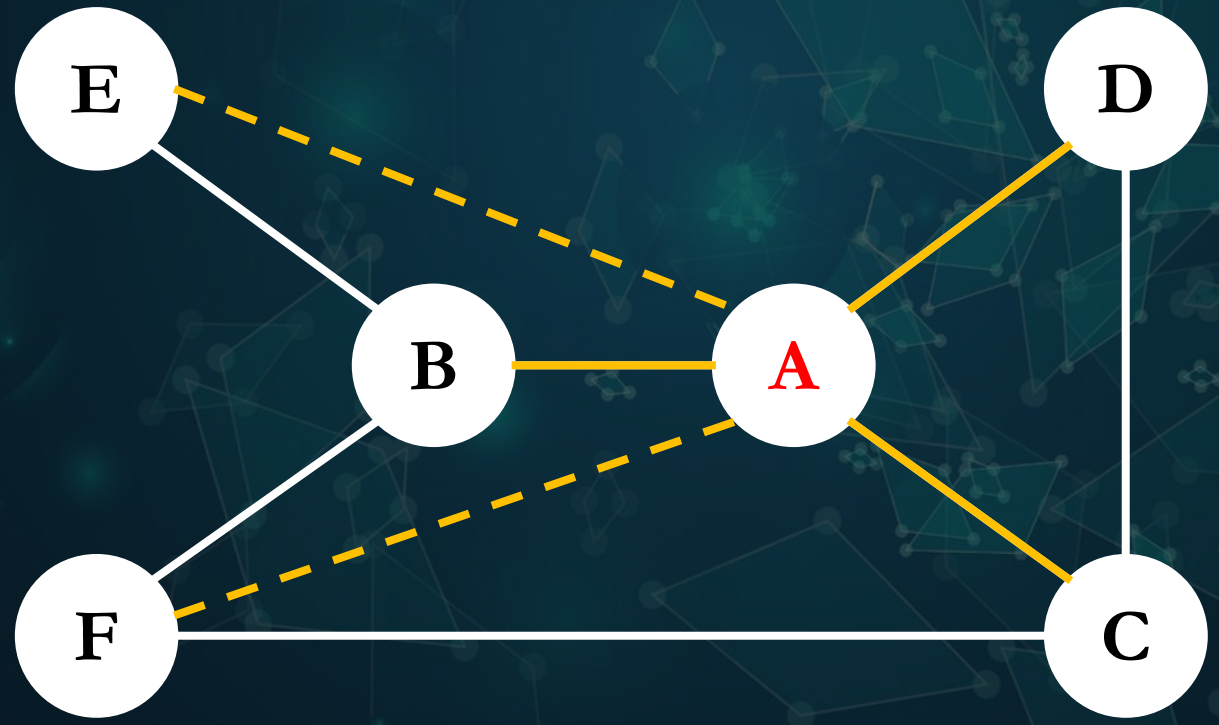
- Heterogenous Network Construction



- Disease: target disease, symptoms, risk factors, complications etc.
- Chemical: chemical elements, compounds, drugs etc.
- Gene: refers to a certain segment of nucleotides on chromosome;
- Genetic variant: gene mutation, protein mutation and single nucleotide polymorphism (SNP)

2. Methodology Framework

- Network Analytics – Centrality Measurement



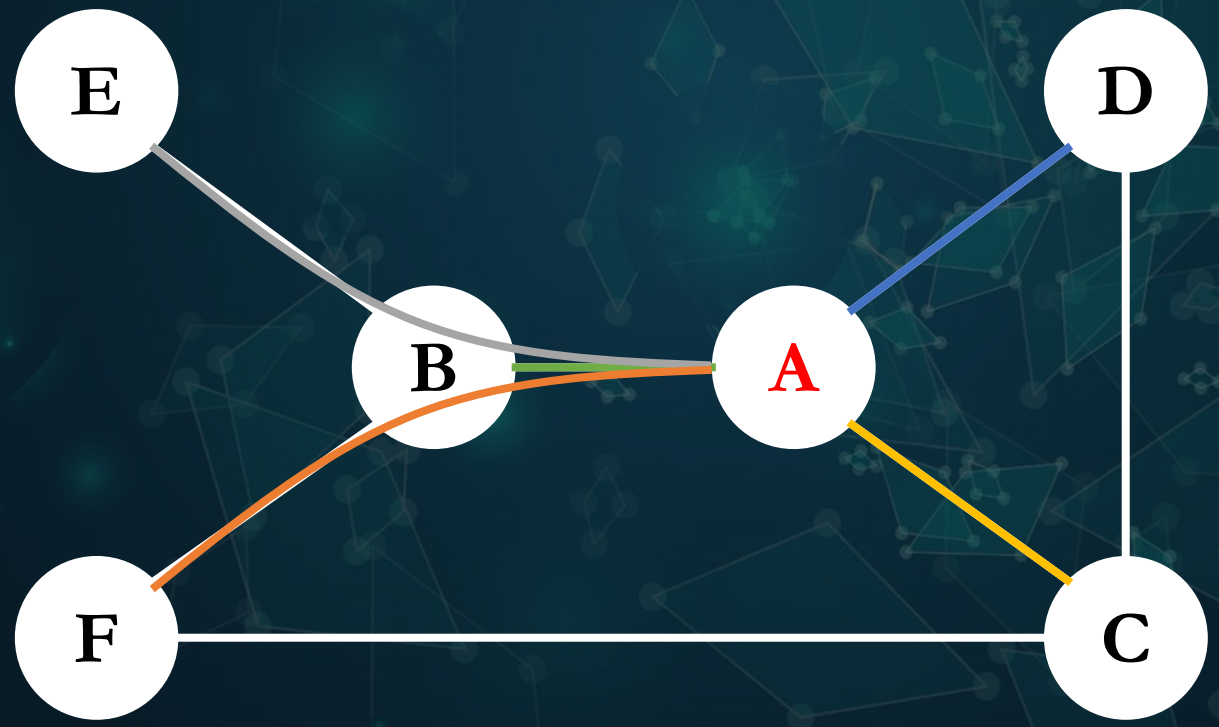
Degree Centrality (*DC*)

$$DC(A) = \frac{The\ degree\ of\ A}{Num\ of\ nodes - 1}$$

For node A, $DC = 3/5 = 0.6$

2. Methodology Framework

- Network Analytics – Centrality Measurement



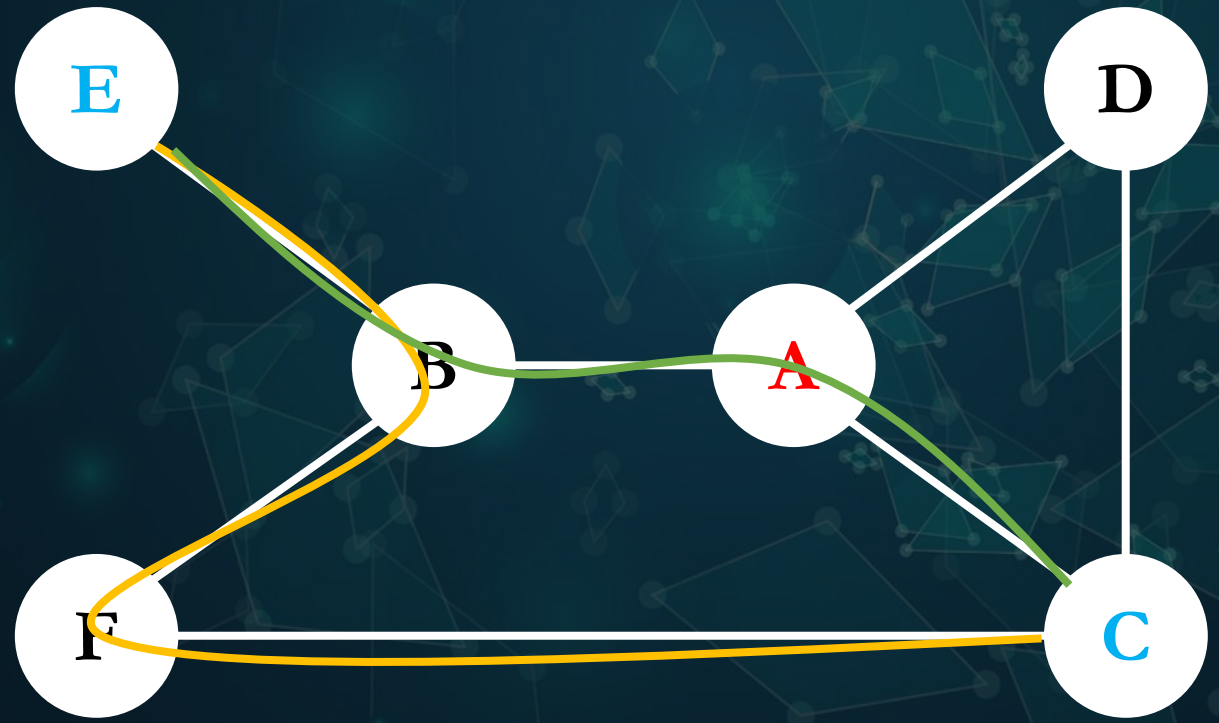
Closeness Centrality (CC)

$$CC(A) = \frac{Num\ of\ nodes - 1}{the\ sum\ of\ topological\ distances\ of\ A\ to\ other\ nodes}$$

For node A, $CC = \frac{5}{1+1+1+2+2} = 0.714$

2. Methodology Framework

- Network Analytics – Centrality Measurement



Betweenness Centrality (BC)

$$BC\left(V_i^m\right) = \frac{\sum_{all\ pairs} \frac{num\ of\ the\ shortest\ paths\ pass\ A}{Total\ num\ of\ the\ shortest\ paths}}{the\ num\ of\ node\ pairs}$$

For node A, $BC = \frac{\frac{1}{2}+\cdots+\cdots}{(5*4)/2}$

2. Methodology Framework

- Centrality Integration: Non-dominating sorting[2]

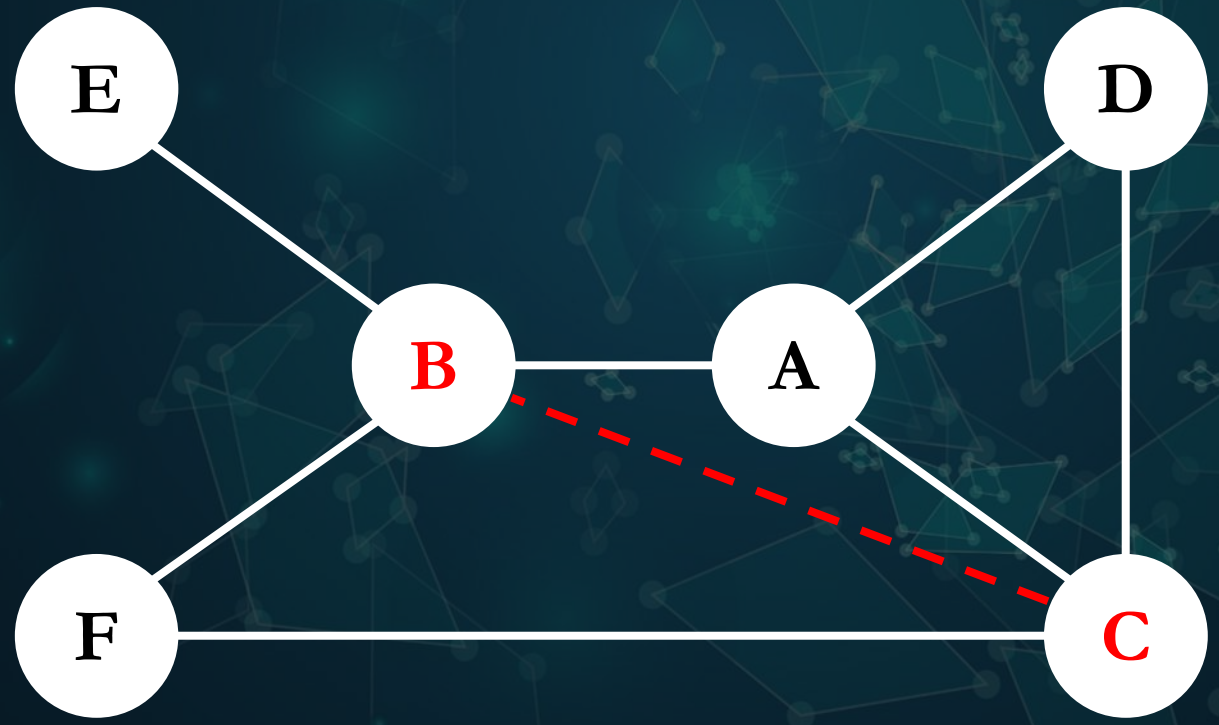
	Degree Centrality	Closeness Centrality	Betweenness Centrality
Node A	0.8	0.5	0.7
Node B	0.1	0.3	0.5
Node C	0.3	0.2	0.5
Node D	0.2	0.1	0.2
Node E	0.4	0.5	0.6

- Objective: Comprehensively identify dominant nodes with 3 prior values for all the centralities

[2] Y. Yuan, H. Xu, and B. Wang, "An improved NSGA-III procedure for evolutionary many-objective optimization," in Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, 2014, pp. 661-668.

2. Methodology Framework

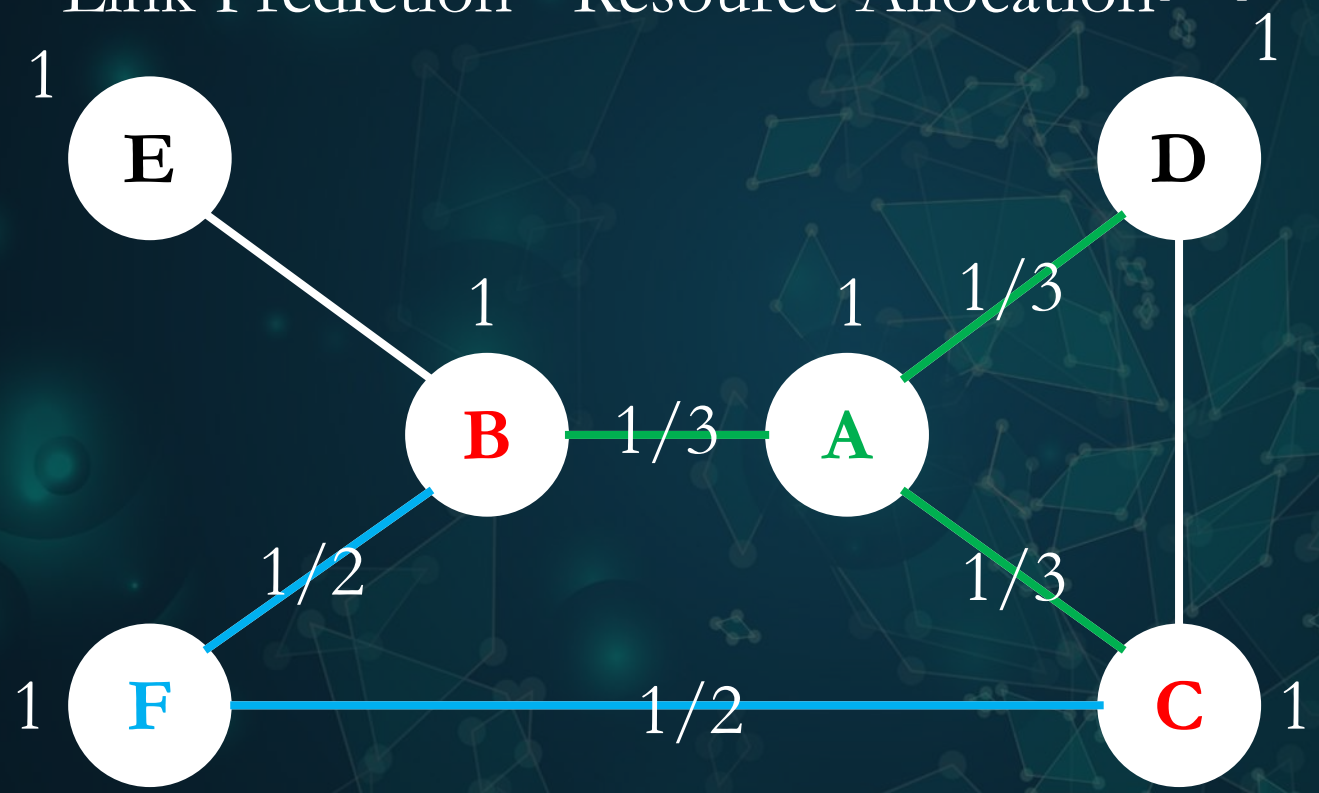
- Network Analytics – Link Prediction



- Common neighbor-based Assumption: If two unconnected nodes share common neighbor(s), there is possibility that an edge will emerge between them.

2. Methodology Framework

- Link Prediction - Resource Allocation[3, 4]



Resource Allocation Index (B, C)

$$= \sum_{w \in \Gamma(B) \cap \Gamma(C)} \frac{1}{|\Gamma(w)|}$$

$$= \frac{1}{2} + \frac{1}{3} = 0.833$$

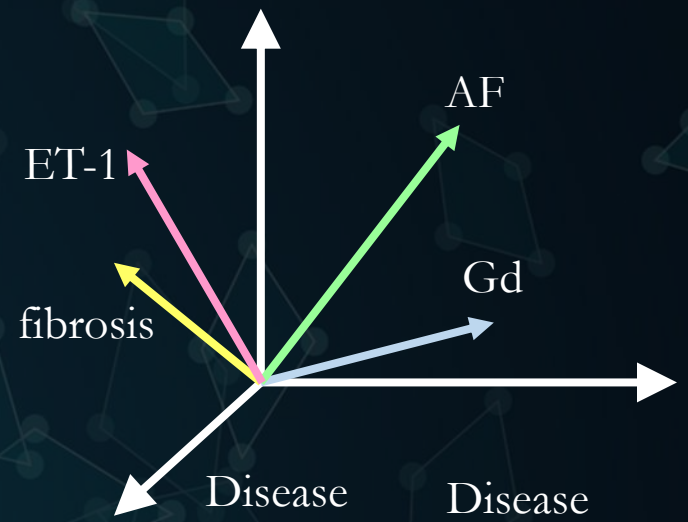
Resource Allocation Index (B, C) (weighted version)

$$= \sum_{w \in \Gamma(B) \cap \Gamma(C)} \frac{E(w,B)+E(w,C)}{\sum_{v \in \Gamma(w)} E(w,v)}$$

[3] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," The European Physical Journal B, vol. 71, no. 4, pp. 623-630, 2009.

[4] Zhang, Y., Wu, M., Zhu, Y., Huang, L., & Lu, J. (2020b). Characterizing the potential of being emerging generic technologies: A prediction method incorporating with bi-layer network analytics. Journal of Informetrics, under review.

2. Methodology Framework



- Bioentity2Vec Model Training

Gene Disease Chemical Disease Disease

…Plasma big endothelin-1 predicts atrial fibrillation … late gadolinium enhancement…of AF and fibrosis….

Skip-Gram Algorithm[1]

…	E(t-2) ET-1	E(t-1) AF	E(t) Gd	E(t+1) AF	E(t+2) fibrosis	…

Entity Window size = 5

E(t) Gd

- Semantic Similarity (“AF”, “ET-1”) = Cosine Similarity ($\overrightarrow{AF}, \overrightarrow{ET-1}$)

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

2. Methodology Framework

- Bioentity2Vec & Resource Allocation Incorporation

Proposed Semantic-Enhanced Resource Allocation Index:

$$R_{(B,C)} = \sum_{w \in \Gamma(B) \cap \Gamma(C)} \frac{CF(B,w)\left|S_{B,w}\right| + CF(w,C)\left|S_{w,C}\right|}{\sum_{v \in \Gamma(w)} CF(v,w) S\left(S_{v,w}\right)}$$

$CF(B,w)$ is the co-occurring frequency of entity B and entity w, $S_{B,w}$ represents the semantic similarity between entities B and w.

Output: a ranking list of genetic factors

3. Case Study

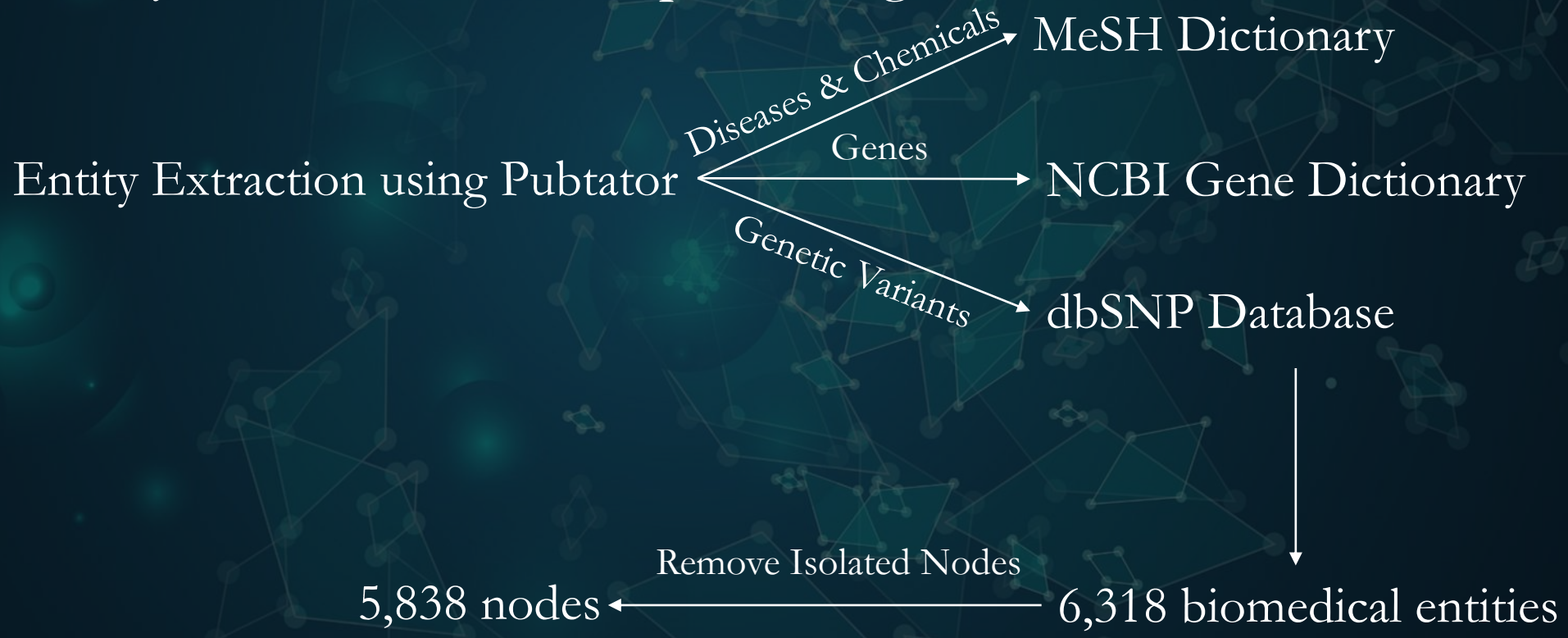
- Data Collection and Entity Extraction
- PubMed database

“("Atrial Fibrillation"[Mesh] AND Humans[Mesh])”
Search Date: 2020/04/28

Record Num: 54,219

3. Case Study

- Entity Extraction and Pre-processing



3. Case Study

PubTator Central

31092201

TUTORIAL API FTP

group type sort freq

Search...

GENE
MIN/1 (8)

DISEASE
BLEEDING (29)
STROKE (21)
ATRIAL FIBRILLATION (18)
THROMBI (12)
THROMBOEMBOLIC (10)
more

CHEMICAL
OAC (12)
ASPIRIN (4)
VITAMIN-K (2)
CLOPIDOGREL (2)
SALINE (1)

Cerebrovascular events, bleeding complications and device related thrombi in atrial fibrillation patients with chronic kidney disease and left atrial appendage closure with the WATCHMAN device

PMID31092201

LUANI B, GENZ C ... RAUWOLF T • BMC CARDIOVASC DISORD. 2019 MAY 15 • 2019

FULL-TEXT

Background

Impaired renal function increases the bleeding risk, leading to a conservative prescription and frequent discontinuation of oral anticoagulation in atrial fibrillation patients with chronic kidney disease (CKD). Interventional left atrial appendage closure (LAAC) might be an alternative therapeutic strategy for these patients.

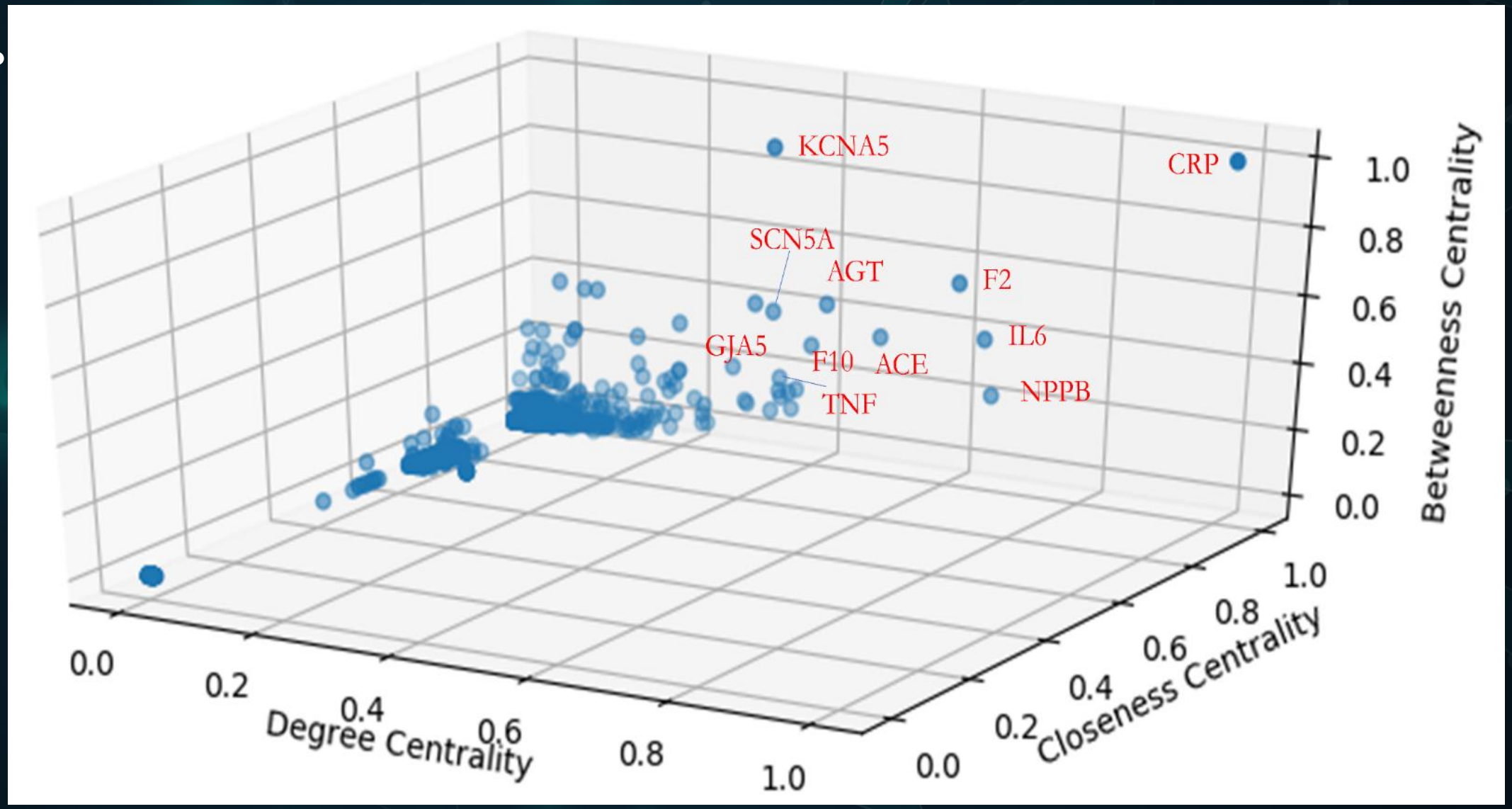
BioConcepts

- GENE
- DISEASE
- CHEMICAL
- MUTATION
- SPECIES
- CELLLINE

Navigation

TITLE
BACKGROUND
METHODS
RESULTS
DISCUSSION
CONCLUSIONS
ABBREVIATIONS

3. Case Study


KCNA5
CRP
SCN5A
AGT
F2
GJA5
F10
ACE
IL6
TNF
NPPB
Betweenness Centrality
Degree Centrality
Closeness Centrality
0.0
0.2
0.4
0.6
0.8
1.0

3. Case Study

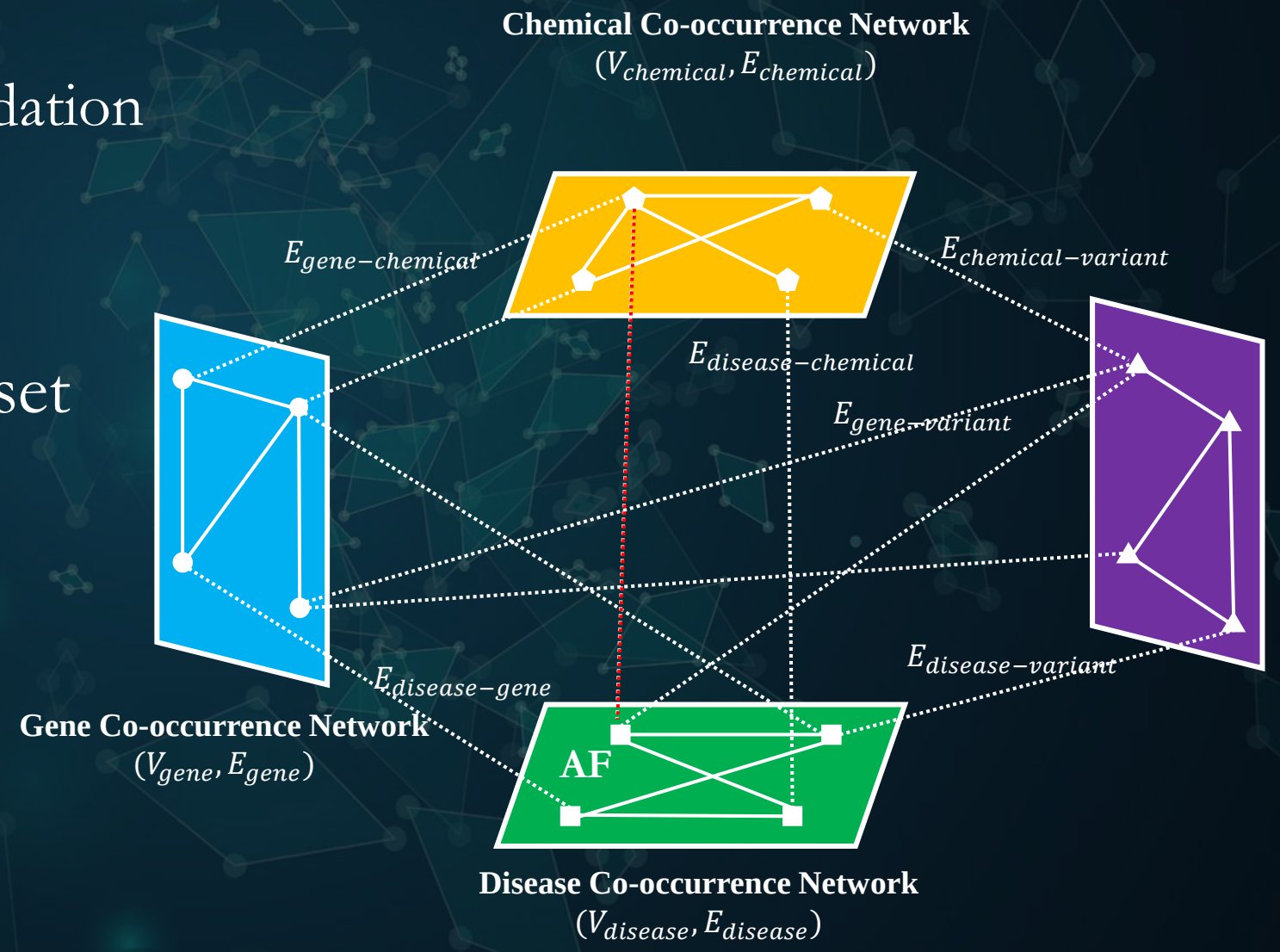
- Centrality Measurement - Gene

	Top 20 Results by Non-dominating Sorting
Disease	Atrial Fibrillation; Stroke; Heart Failure; Hypertension; Hemorrhage; Diabetes Mellitus; Fibrosis; Myocardial Infarction; Cerebral Infarction; Ischemia; Thromboembolism; Death; Thrombosis; Inflammation; Coronary Artery Disease; Tachycardia; Ventricular Fibrillation; Tachycardia, Supraventricular; Neoplasms; Atrioventricular Block
Chemical	Warfarin; Calcium; Amiodarone; Potassium; Digoxin; Ethanol; Verapamil; Sodium; Oxygen; Quinidine; Aspirin; Vitamin K; Glucose; Cholesterol; apixaban; Sotalol; Nitrogen; Magnesium; Heparin; Propafenone
Gene	CRP; F2; ACE; IL6; AGT; F10; SCN5A; NPPB; KCNA5; PITX2; FGB; GJA5; TNNI3; INS; TNF; TGFB1; VWF; KCNQ1; SERPINE1; AGTR1
SNP	rs2200733; rs6795970; rs2106261; rs2108622; rs3789678; rs13376333; rs17042171; rs1805127; rs7539020; rs11568023; rs10033464; rs3807989; rs7193343; rs3918242; rs3825214; rs16899974; rs699; rs7164883; rs6584555; rs10824026

3. Case Study

- Link Prediction Validation

Roll Back the dataset by 5 years



3. Case Study

- Validation Results

	Resource Allocation	Weighted Resource Allocation	Modified Resource Allocation (Purposed)
Top *k* Recall	0.245	0.208	**0.283**
Top 100 Recall	0.434	0.396	**0.472**
Top 200 Recall	0.604	0.642	**0.736**

k refers to the number of edges that were removed for node AF, in this experiment $k = 53$.

4. Limitations and Future Directions

Limitations:

- Negative associations collected when using co-occurrence
- The genetic research of AF is still at an early stage, some associations between AF and genes haven't been revealed yet

Future Study:

- Employ Sentiment analysis to exclude those negative associations
- Modify the entity extraction rules
- Involve the identified crucial genetic factors to improve predicting performance

Thank you!

Email address: Mengjia.wu@student.uts.edu.au