

# Predicting Missing Links in Infection Networks: Accelerate contact tracing investigations using network theory.

**Pavlos Alexandros Dimitriou<sup>1,2</sup>, V. Silvestros<sup>3</sup>, C.Pitris<sup>1,2</sup>, P. Kolios<sup>1,4</sup>**

<sup>1</sup>University of Cyprus, KIOS Research and Innovation Center of Excellence, Nicosia, Cyprus.

<sup>2</sup>University of Cyprus, Dept. of Electrical & Computer Engineering, Nicosia, Cyprus.

<sup>3</sup>Unit for Surveillance and Control of Communicable Diseases, Ministry of Health, Nicosia, Cyprus.

<sup>4</sup>University of Cyprus, Dept. of Computer Science, Nicosia, Cyprus.

**24-27 September, World Congress of Epidemiology**

# Building networks from contact tracing data

Data store in excel tables

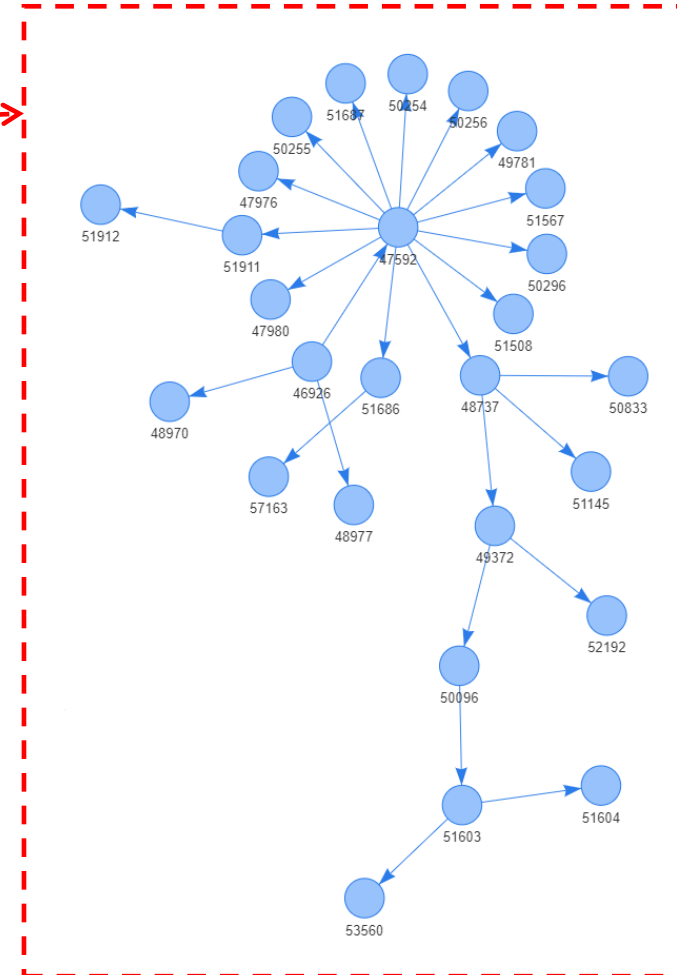
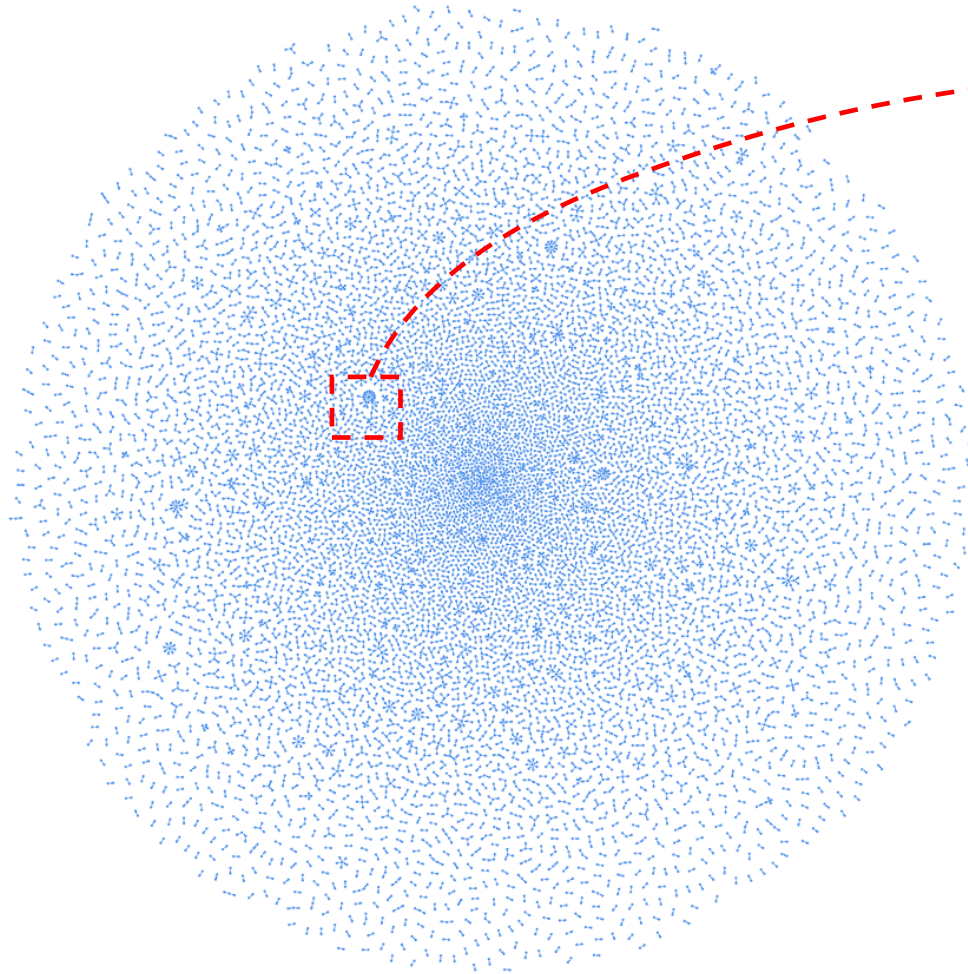


Case Id	Contact case Id	Gender	District	Age group	NACE	Infection date
1	11	M	Pafos	20-29	P85.4	12/1/2021
2	12	M	Pafos	3-5	U6	12/1/2021
3	nan	M	Pafos	30-39	P85.4	12/1/2021
4	3	F	Pafos	40-49	P85.4	12/1/2021
5	12	M	Pafos	3-5	U6	12/1/2021
6	12,13	F	Pafos	0-2	U6	12/1/2021
7	12	F	Pafos	30-39	nan	12/1/2021
8	4	F	Pafos	40-49	nan	12/1/2021
9	3	F	Pafos	60-69	R65	12/1/2021
10	12	M	Pafos	20-29	T97	12/1/2021
11	3	F	Pafos	50-59	P85.4	12/1/2021
12	3	F	Pafos	60-69	R65	12/1/2021
13	3	M	Pafos	70-79	R65	12/1/2021

Challenging to analyse



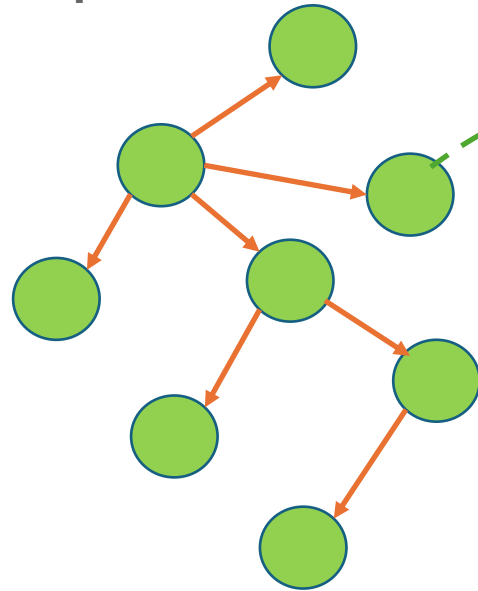
# Infection networks are sparse



# Can we predict missing links?

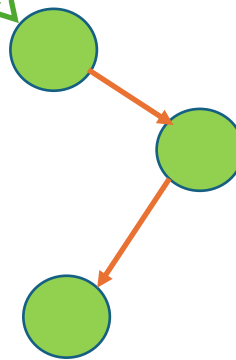
## Infection Network

Component 1



?

Component 2



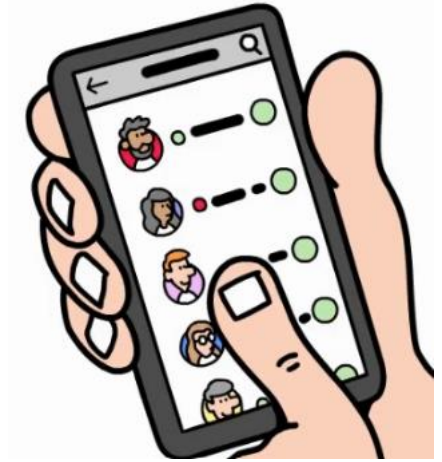
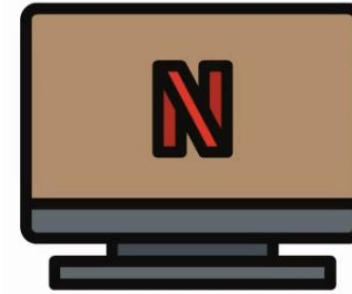
Link prediction

RECONSTRUCTION OF  
THE TRANSMISSION  
CHAINS



# Link Prediction and Its Application:

- Widely used in different domains
  - Recommender system
    - friendships in social networks
    - e-commerce websites



Feature vector ( $x$ )

$$x = [f_{0,0}, \dots, f_{i,j}, \dots, f_{n,n}]$$

Class label ( $y$ )

$$y = [y_{0,0}, \dots, y_{i,j}, \dots, y_{n,n}]$$

# Creating Feature Vectors:

## 1. Node2vec:

- Features are calculated solely on network characteristics



**FEATURE VECTORS BASED ON THE CONNECTIONS AND TOPOLOGY OF THE INFECTION NETWORK**

## 2. Shallow embeddings with handcrafted features:

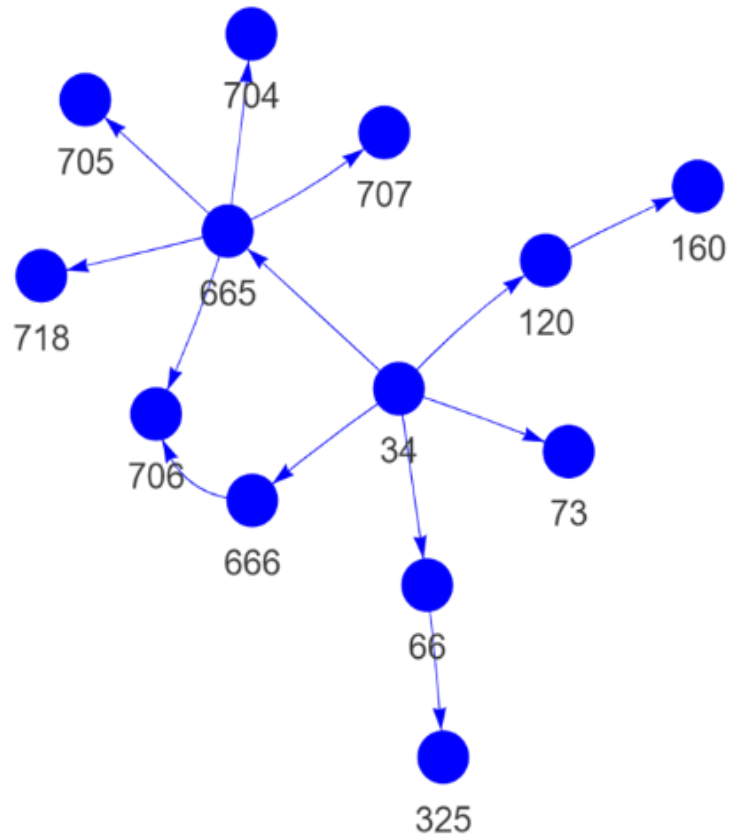
**EPIDEMIOLOGICAL FEATURES**  
+  
**NETWORK STRUCTURE**

- THE TIME DIFFERENCE BETWEEN INFECTIONS
- THE PHYSICAL DISTANCE BETWEEN INDIVIDUALS BASED ON THEIR RESIDENCE
- THE AGE DIFFERENCE BETWEEN INFECTED INDIVIDUALS
- THE OVERLAP OF OCCUPATIONS (NACE CODES) AMONG CHAINS OF INFECTED INDIVIDUALS
- THE OVERLAP OF POSTCODES AMONG INFECTION CHAINS

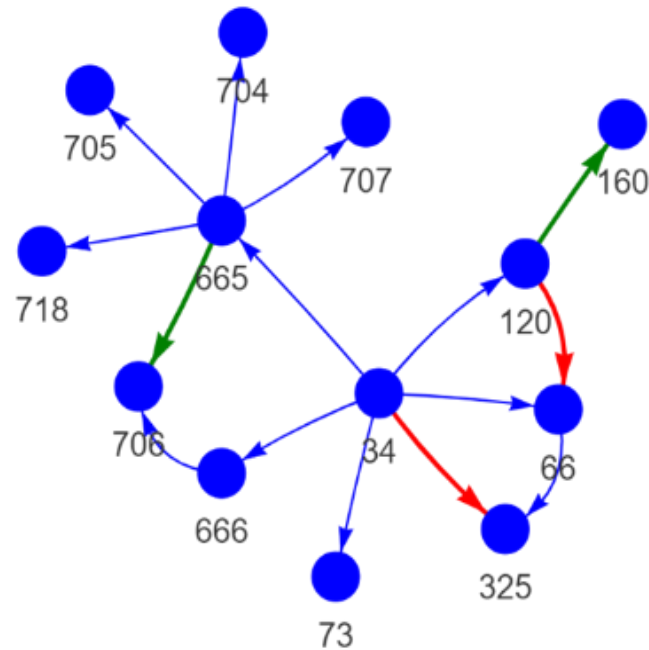
# Splitting the Dataset

- Sampling positive and negative edges to create training and test set

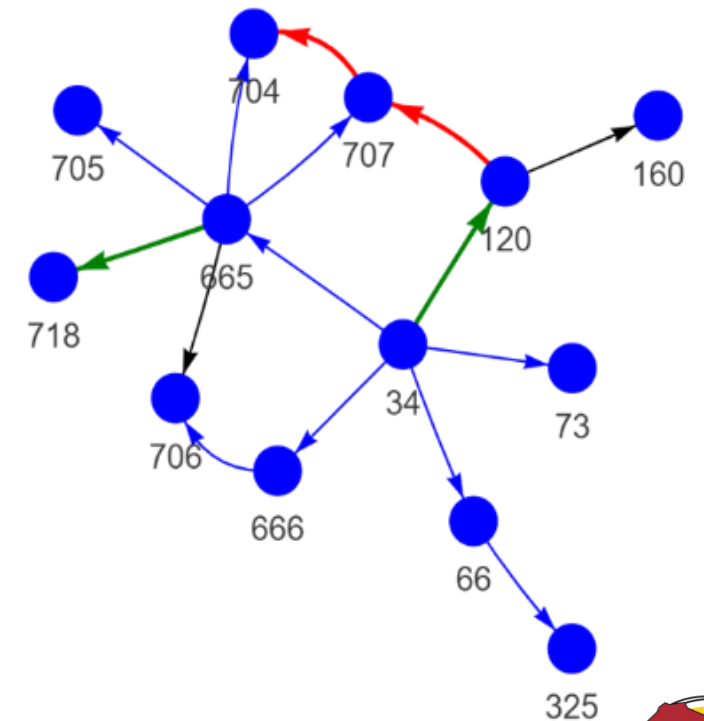
Full network



Training network

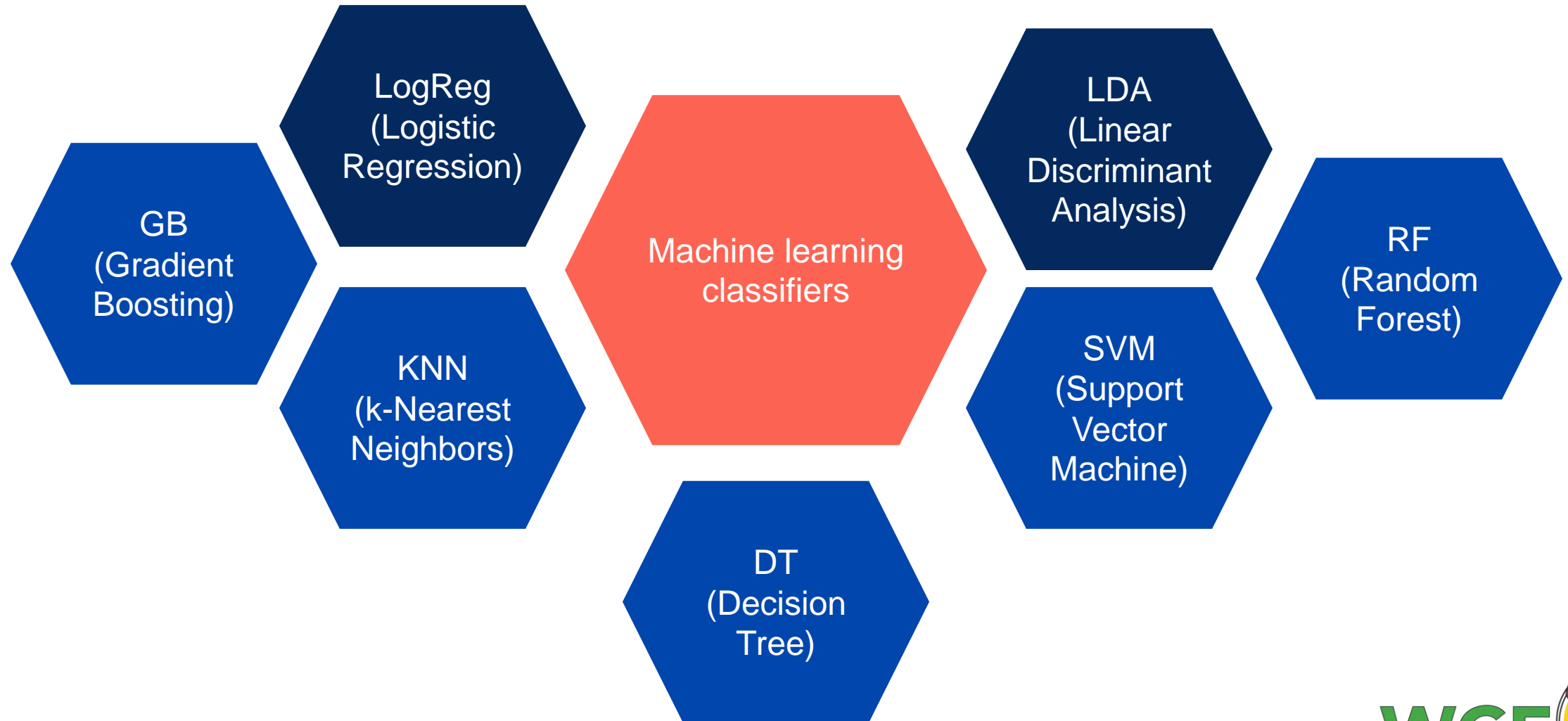


Test network





# Machine Learning: Classification algorithms



# Performance Metrics: Node2vec Link prediction

	Wave 1		Wave 2		Wave 3		Wave 4	
	<u>AUC</u>	<u>F1-Score</u>	<u>AUC</u>	<u>F1-Score</u>	<u>AUC</u>	<u>F1-Score</u>	<u>AUC</u>	<u>F1-Score</u>
RF	0.54	0.49	0.54	0.46	0.57	0.53	0.67	0.65
LDA	0.50	0.51	0.48	0.47	0.55	0.57	0.65	0.71
<b>SVM</b>	<b>0.54</b>	<b>0.53</b>	<b>0.50</b>	<b>0.51</b>	<b>0.61</b>	<b>0.64</b>	<b>0.70</b>	<b>0.75</b>
LOGREG	0.53	0.52	0.49	0.49	0.58	0.60	0.68	0.75
KNN	0.56	0.61	0.49	0.50	0.58	0.65	0.64	0.62
NB	0.52	0.42	0.51	0.47	0.58	0.55	0.62	0.71
GB	0.57	0.57	0.52	0.52	0.58	0.61	0.69	0.71
DT	0.53	0.52	0.50	0.51	0.54	0.54	0.62	0.62

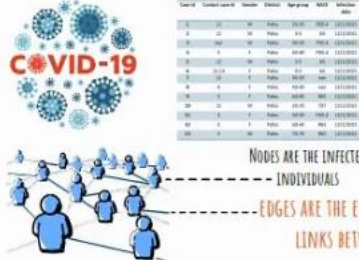
# Performance Metrics: Shallow embeddings with handcrafted features

	Wave 1		Wave 2		Wave 3		Wave 4	
	<u>AUC</u>	<u>F1-Score</u>	<u>AUC</u>	<u>F1-Score</u>	<u>AUC</u>	<u>F1-Score</u>	<u>AUC</u>	<u>F1-Score</u>
RF	0.87	0.85	0.82	0.82	0.93	0.93	0.94	0.94
LDA	0.85	0.86	<b>0.82</b>	<b>0.83</b>	0.90	0.90	0.91	0.91
SVM	0.56	0.57	0.55	0.61	0.54	0.59	0.54	0.53
LOGREG	0.54	0.17	0.53	0.24	0.53	0.10	0.53	0.11
KNN	0.78	0.75	0.74	0.75	0.67	0.67	0.73	0.73
NB	0.56	0.57	0.55	0.58	0.54	0.61	0.53	0.63
<b>GB</b>	<b>0.90</b>	<b>0.88</b>	<b>0.82</b>	<b>0.82</b>	<b>0.94</b>	<b>0.94</b>	<b>0.95</b>	<b>0.95</b>
DT	0.81	0.85	0.79	0.80	0.91	0.90	0.92	0.92

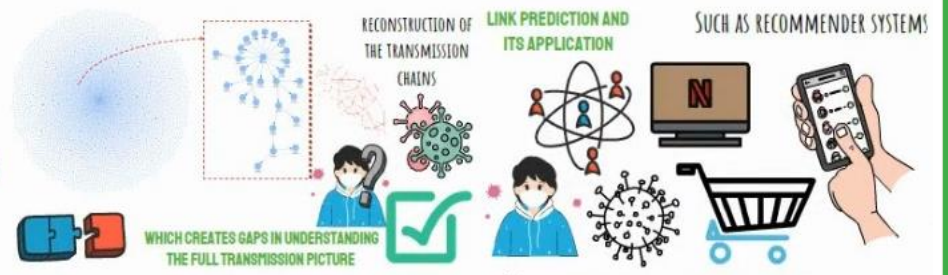


**PREDICTING MISSING LINKS IN INFECTION NETWORKS:  
ACCELERATE CONTACT TRACING INVESTIGATIONS  
USING NETWORK THEORY.**

24-27 SEPTEMBER, WORLD CONGRESS OF EPIDEMIOLOGY



**WHILE ALSO HELPING  
EPIDEMIOLOGISTS  
ACCELERATE CONTACT TRACING  
INVESTIGATIONS**

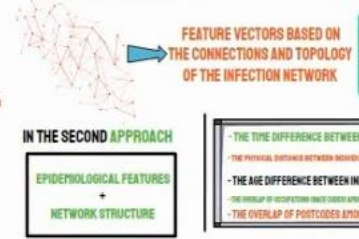


**SPLITTING THE DATASET**

ONCE THE FEATURE VECTORS ARE CREATED



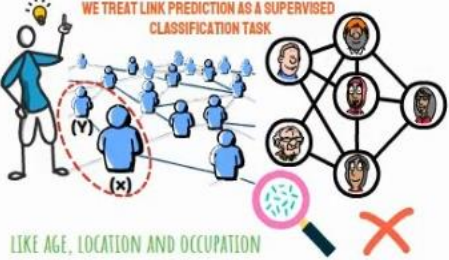
**THE FIRST APPROACH UTILIZES NODE2VEC**



**CREATING FEATURE VECTORS FOR LINK PREDICTION**



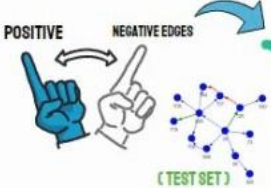
**USING THESE FEATURE VECTORS AND LABELS**



**POSITIVE EDGES**



**NEGATIVE EDGES**



**MACHINE LEARNING CLASSIFICATION ALGORITHMS**



PERFORMANCE ON PREDICTING MISSING LINKS

**APPROACH 1:**

	Wave 1		Wave 2		Wave 3		Wave 4	
	AUC	F1-Score	AUC	F1-Score	AUC	F1-Score	AUC	F1-Score
RF	0.54	0.49	0.54	0.46	0.57	0.53	0.67	0.65
LDA	0.50	0.51	0.48	0.47	0.55	0.57	0.65	0.71
SVM	0.54	0.58	0.50	0.51	0.61	0.64	0.70	0.75
LOGREG	0.53	0.52	0.49	0.49	0.58	0.60	0.68	0.75
KNN	0.56	0.61	0.49	0.50	0.58	0.65	0.64	0.62
NB	0.52	0.42	0.51	0.47	0.58	0.55	0.62	0.71
GB	0.57	0.57	0.52	0.52	0.56	0.61	0.69	0.71
DT	0.53	0.52	0.50	0.51	0.54	0.54	0.62	0.62

WITH THE NODE2VEC ALGORITHM

IDENTIFYING MISSING LINKS WITHIN THE INFECTION NETWORKS

**APPROACH 2:**

	Wave 1		Wave 2		Wave 3		Wave 4	
	AUC	F1-Score	AUC	F1-Score	AUC	F1-Score	AUC	F1-Score
RF	0.57	0.45	0.52	0.40	0.50	0.50	0.54	0.54
LDA	0.50	0.50	0.48	0.48	0.50	0.50	0.50	0.50
SVM	0.56	0.57	0.51	0.51	0.54	0.54	0.54	0.54
LOGREG	0.54	0.51	0.51	0.51	0.54	0.54	0.54	0.54
KNN	0.56	0.75	0.74	0.75	0.67	0.67	0.79	0.79
NB	0.50	0.52	0.51	0.51	0.54	0.54	0.54	0.54
GB	0.57	0.57	0.52	0.52	0.56	0.61	0.69	0.71
DT	0.53	0.52	0.50	0.51	0.54	0.54	0.62	0.62

ACHIEVING AN AUC AND F1 SCORE OF 95%

OCCUPATION OVERLAPS, AND POSTCODE OVERLAPS

EPIDEMIOLOGICAL & NETWORK FEATURES



**THANK YOU!**

Grant agreement No. 739551 (KIOS CoE).

CIPHS project (Cyprus Innovative Public Health ICT System), C1.1I2, of the NextGenerationEU-Recovery and Resilience

