Link Prediction

Supervised vs unsupervised

The supervised learner doesn't know much

The unsupervised learner knows what it is doing

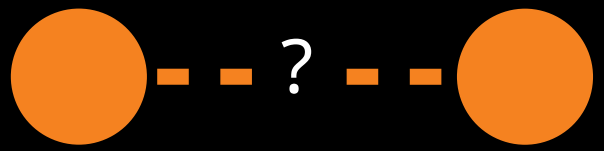
Supervised learning tasks with networks

- Node classification
  - Given a network (e.g. friendship network) and some labels (e.g. political party). Can we predict the labels of a node from the labels of their neighbours?

- Graph classification
  - Given many networks (e.g. ego-networks, brain networks) and outcomes (e.g. political party, mental disorders). Can we predict the outcomes from the topology of the network?

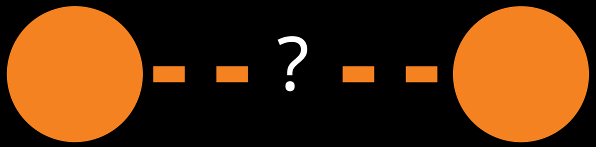
- Link prediction
  - Given a network (e.g. friendship network) and optionally some metadata (e.g. political party). Can we predict which links we are missing (or will be created)?

Link prediction



Does this link exist?

Link prediction



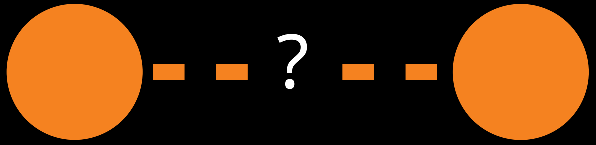
Does this link exist?

Many tasks

<u>1. Model Validation</u>

- Observe part of the adjacency matrix (fit model)
- Predict held out entries (cross validation)

Link prediction



Does this link exist?

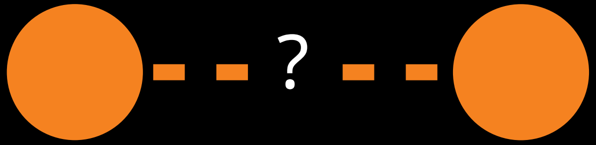
Many tasks

1. Model Validation

2. De-noising / network reconstruction

- Real-world data are noisy / contain errors

Link prediction



Does this link exist?

Many tasks

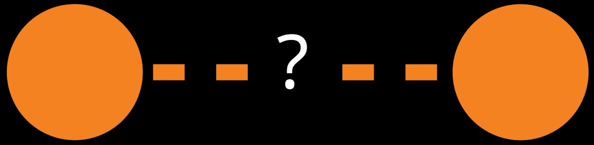
1. Model Validation

2. De-noising / network reconstruction

3. Predict missing links

- Observed edges are assumed correct
- Predict which unobserved edges exist

Link prediction



Does this link exist?

Many tasks

1. Model Validation

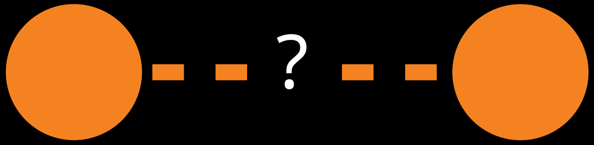
2. De-noising / network reconstruction

3. Predict missing links

4. Predict future links

- Observe the adjacency matrix at time (t)
- Predict edges in time (t+1)

Link prediction



Does this link exist?

Many tasks

1. Model Validation

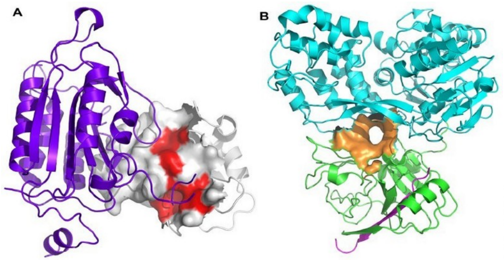
2. De-noising / network reconstruction

3. Predict missing links

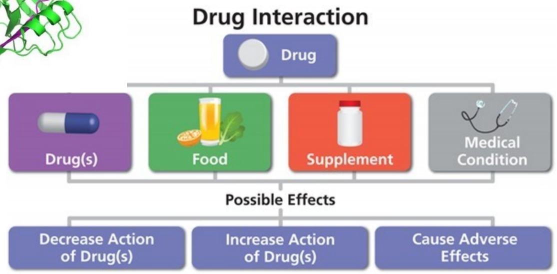
4. Predict future links

Applications

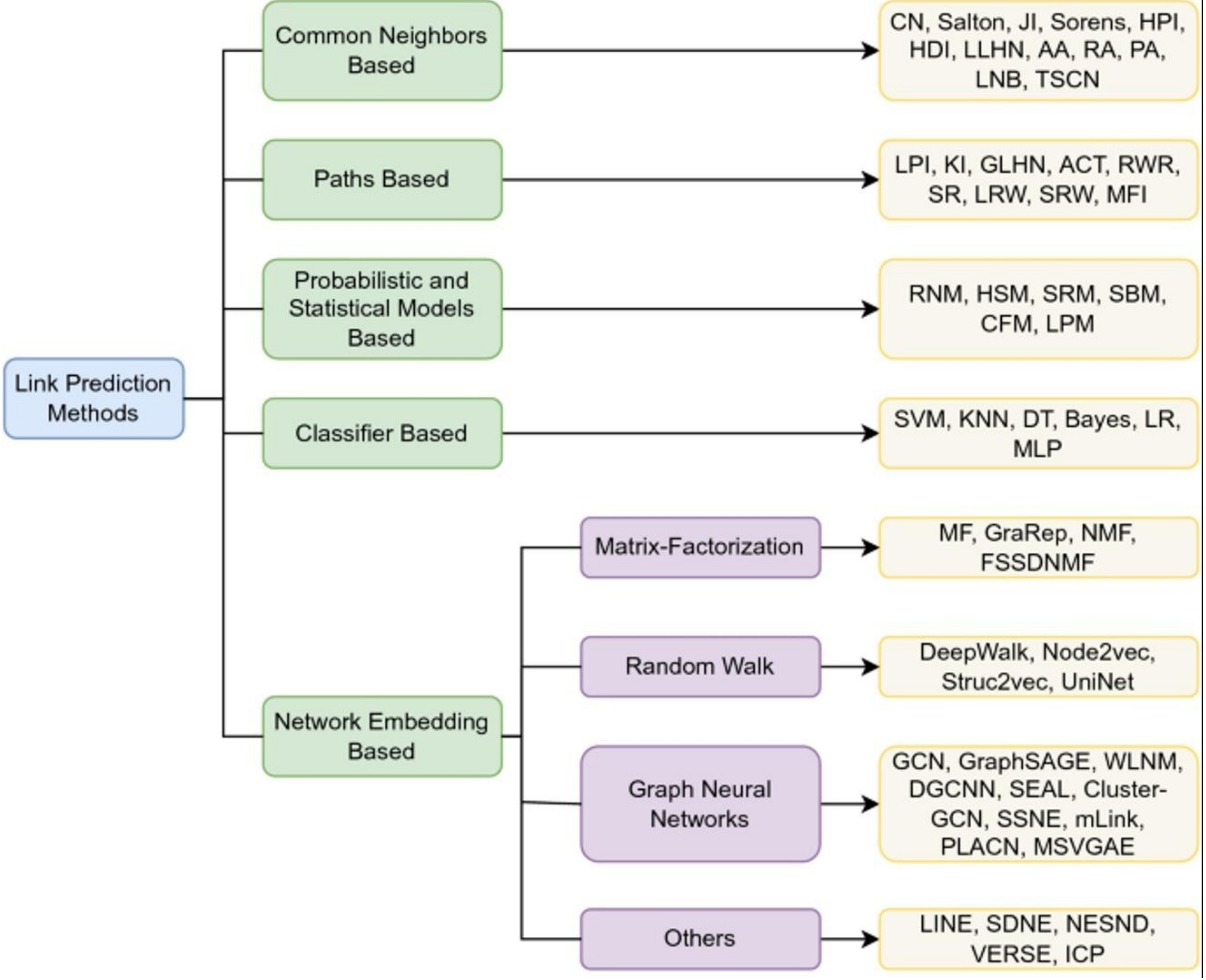
Suggesting social and professional connections

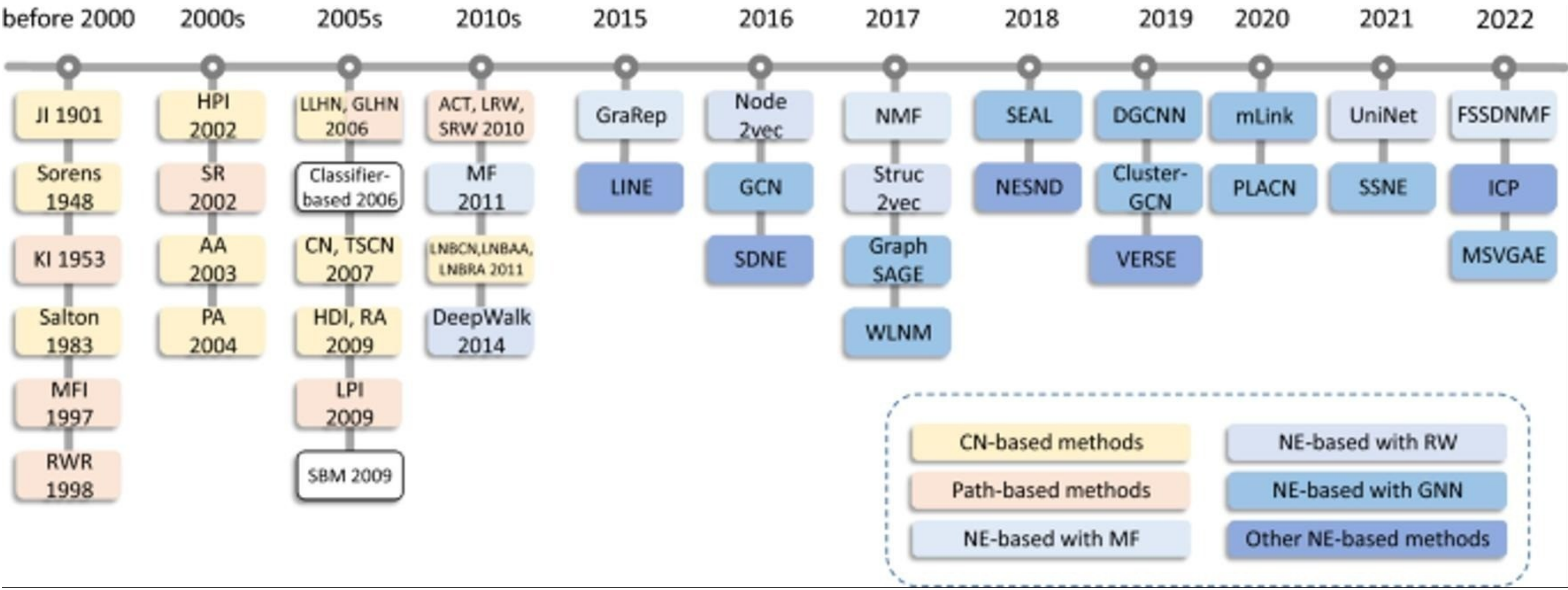


Predicting biological interactions



Recommending products and services


Link Prediction Methods
Common Neighbors Based
CN, Salton, JI, Sorens, HPI, HDI, LLHN, AA, RA, PA, LNB, TSCN
Paths Based
LPI, KI, GLHN, ACT, RWR, SR, LRW, SRW, MFI
Probabilistic and Statistical Models Based
RNM, HSM, SRM, SBM, CFM, LPM
Classifier Based
SVM, KNN, DT, Bayes, LR, MLP
Network Embedding Based
Matrix-Factorization
MF, GraRep, NMF, FSSDNMF
Random Walk
DeepWalk, Node2vec, Struc2vec, UniNet
Graph Neural Networks
GCN, GraphSAGE, WLNM, DGCNN, SEAL, Cluster-GCN, SSNE, mLink, PLACN, MSVGAE
Others
LINE, SDNE, NESND, VERSE, ICP



Link Prediction on Complex Networks: An Experimental Survey; Wu, Song, Ge, and Ge (2022)

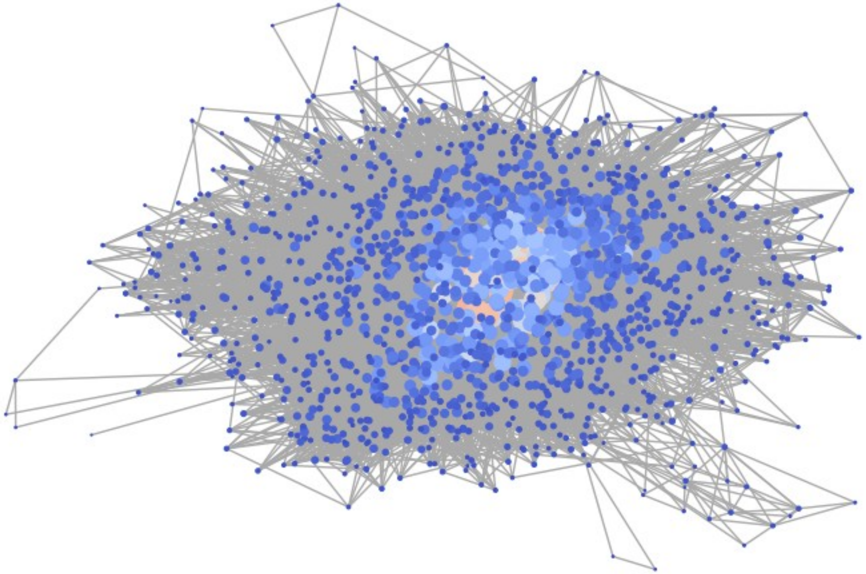
Predicting missing links

Goal: Rank all non-edges according to how likely they are to exist

Assessed using measures such as accuracy, F1, AUC…

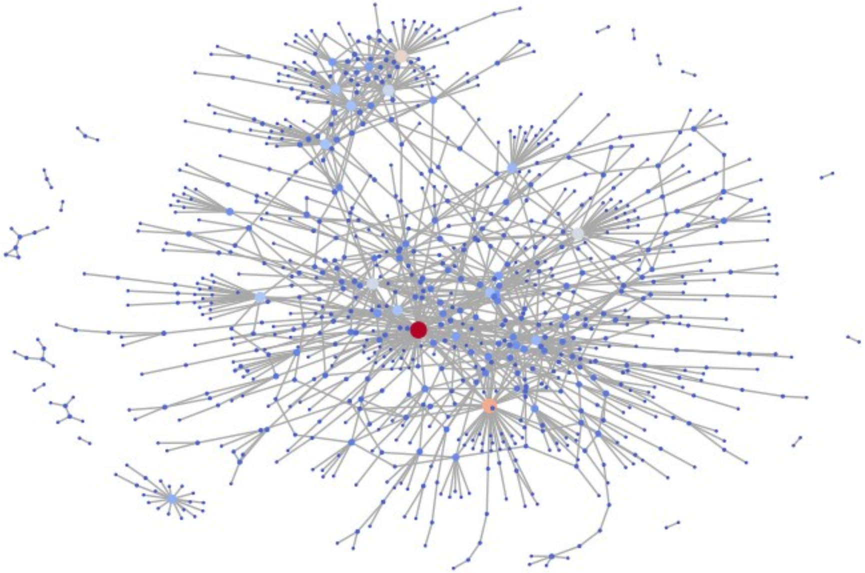
Data Challenge: Link prediction

Twitter network



Global clustering: 0.172
Degree assortativity: -0.03.3

PPI network



Global clustering: 0.014
Degree assortativity: -0.157

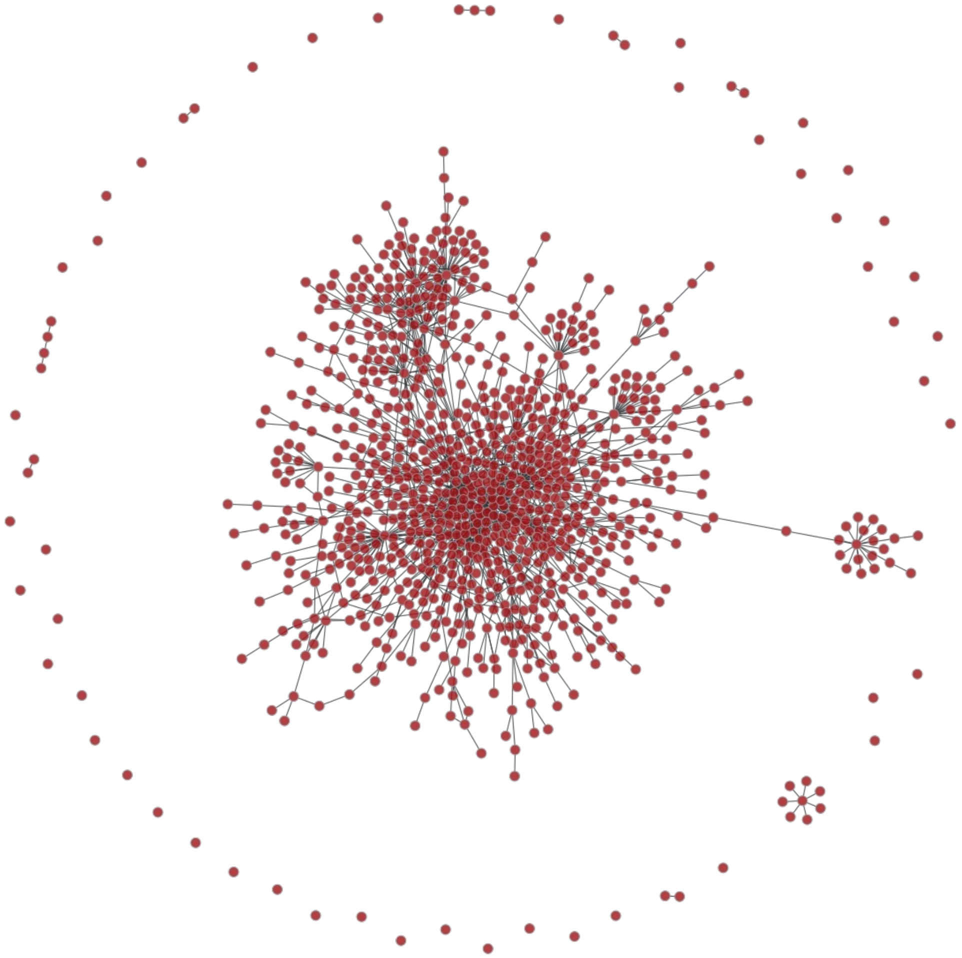
We have removed some edges, your objective is to predict those accurately.

We give you:

- Graph: Used for training
- Test dataset (a series of node pairs, some with a link associated)

How:

- Methods based on common neighbors
- Methods based on paths
- Methods based on embeddings
    - Spectral methods
    - Matrix factorization
    - Node2vec
    - GraphSAGE

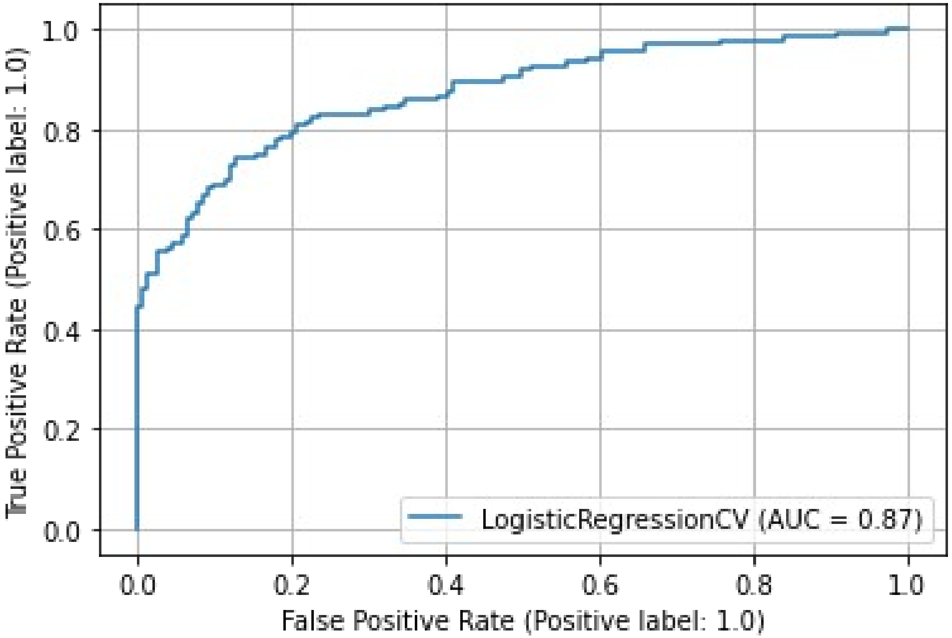


Protein-protein interaction network in S. cerevisiae

Clustering ~ 0; Assortativity ~ -0.2

ROC curve

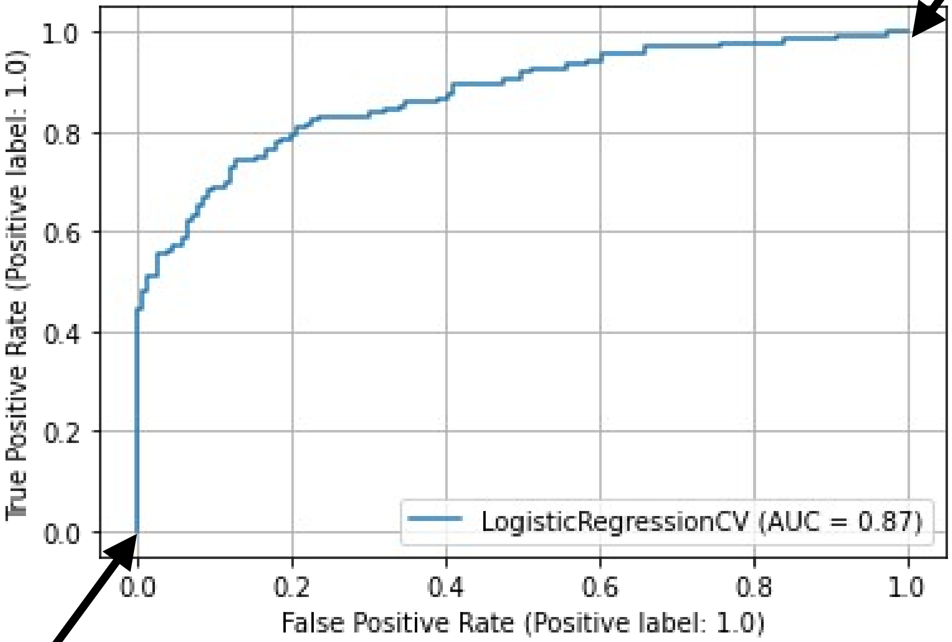
$$\frac{\text{Edges predicted correctly}}{\text{No. of edges}}$$



$$\frac{\text{Non-edges predicted as edges}}{\text{No. of non-edges}}$$

ROC curve

(everything predicted as an edge)



$$\frac{\text{Edges predicted correctly}}{\text{No. of edges}}$$

(nothing predicted as an edge)

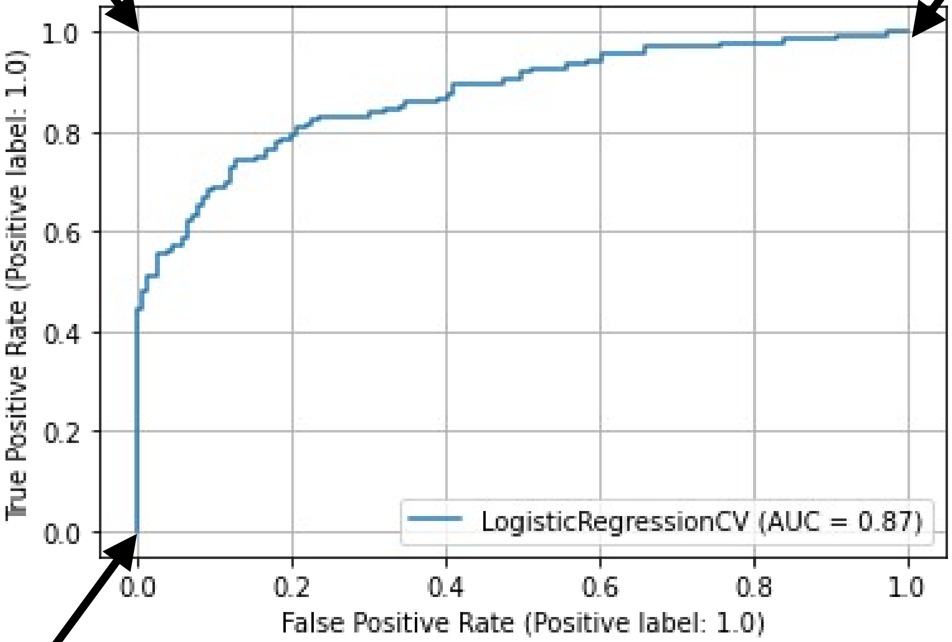
$$\frac{\text{Non-edges predicted as edges}}{\text{No. of non-edges}}$$

ROC curve

(all edges correctly predicted)

(everything predicted as an edge)

$$\frac{\text{Edges predicted correctly}}{\text{No. of edges}}$$



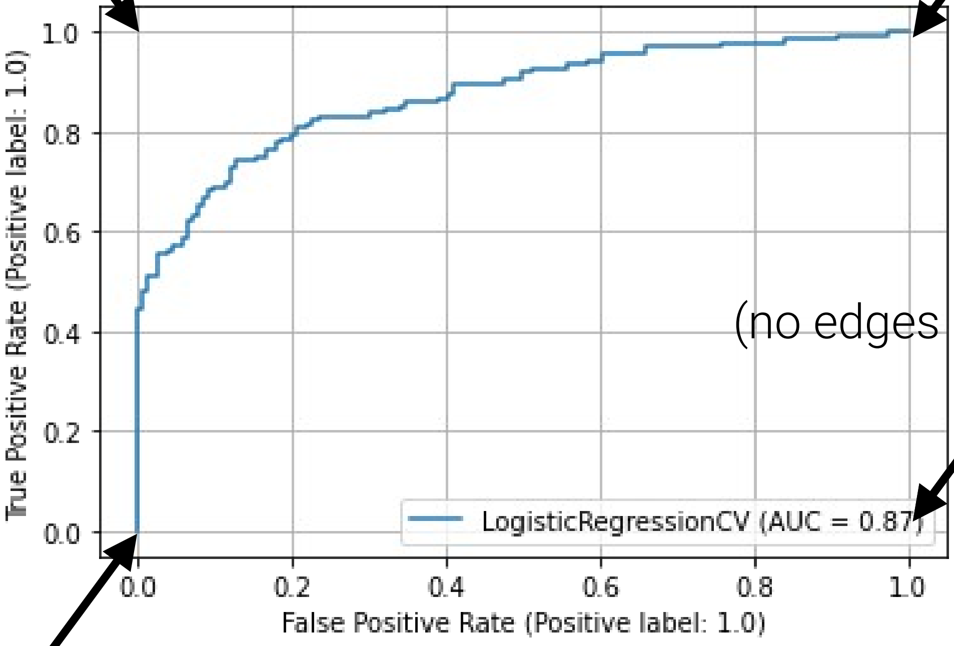
(nothing predicted as an edge)

$$\frac{\text{Non-edges predicted as edges}}{\text{No. of non-edges}}$$

ROC curve

(all edges correctly predicted)

(everything predicted as an edge)



(no edges predicted correctly)

$$\frac{\text{Edges predicted correctly}}{\text{No. of edges}}$$

(nothing predicted as an edge)

$$\frac{\text{Non-edges predicted as edges}}{\text{No. of non-edges}}$$

Local heuristics (common neighbours approach)

Based on similarity of node connections

$\Gamma(x)$ ← neighbours of x

k_x ← degree of x

Common neighbours

$$s_{xy}^{\mathrm{CN}} = |\Gamma(x) \cap \Gamma(y)|,$$

Jaccard similarity

$$s_{xy}^{\mathrm{Jaccard}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Cosine similarity

$$s_{xy}^{\mathrm{Salton}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}}$$

k_x ← degree of x

$\Gamma(x)$ ← neighbours of x

Common neighbour heuristics

Γ(x)	Γ(y)	CN	Jaccard	Cosine
ABC	BC	2	0.66	0.81
ABC	BCD	2	0.5	0.66
ABC	C	1	0.33	0.57
ABC	CD	1	0.25	0.41
ABC	CDE	1	0.2	0.33

Common Neighbours ignores degrees

Jaccard and Cosine provide similar rankings

Other local heuristics

Adamic-Adar

$$s_{xy}^{\mathrm{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$$

Resource Allocation

$$s_{xy}^{\mathrm{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$$

Preferential Attachment

$$s_{xy}^{\mathrm{PA}} = k_x \times k_y$$

Γ(x)	Γ(y)	CN	Jaccard	Cosine
ABCDEF	DEFGH	3	0.38	0.55
ABCDEF	DE	2	0.33	0.58

Jaccard and Cosine do not always provide the same ranking!

Jaccard is biased towards nodes with similar degree