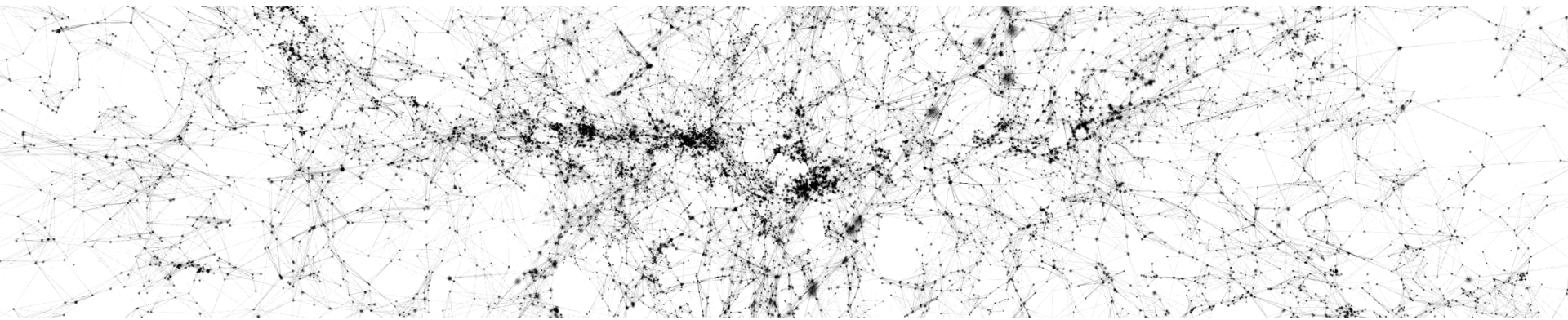
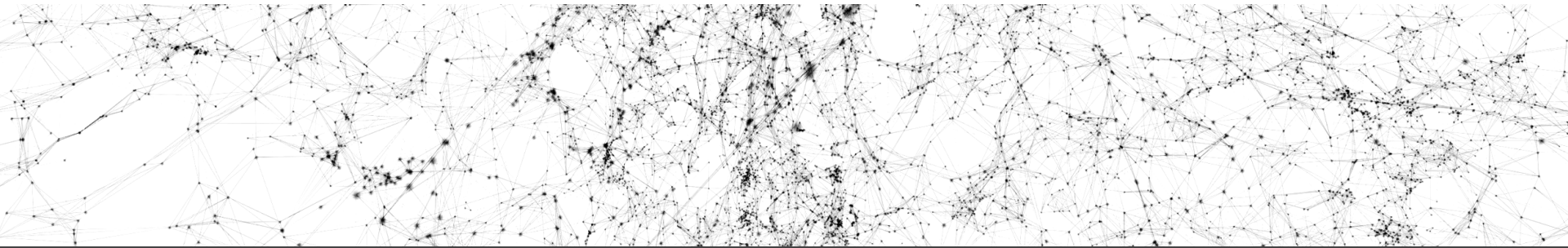


Prediction

Analysis of complex interconnected data



Common prediction tasks

- Link Prediction
- Node Classification
- Graph Classification

Examples:

https://paperswithcode.com/task/link-prediction

https://paperswithcode.com/task/node-classification

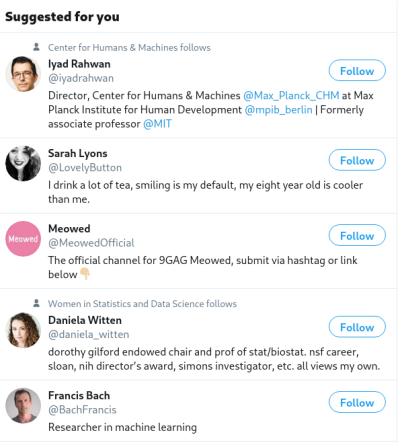
https://paperswithcode.com/task/graph-classification

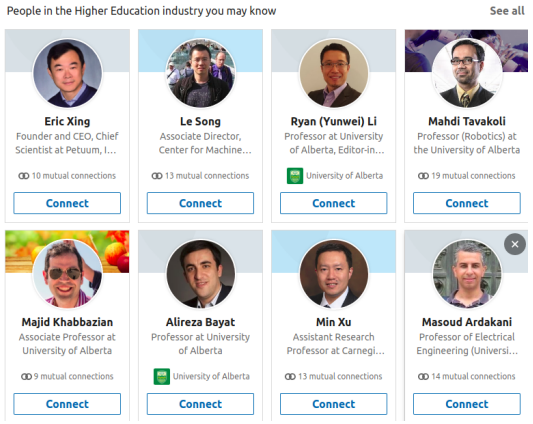
What is unsupervised node classification?

Link Prediction

Given a graph $G(V, E)$ predict the missing (future) links between nodes

- Modelling of the network evolution
- Predict likely interactions, not explicitly observed (e.g. terrorist network monitoring)
- Link recommendation: “friend” suggestion in social networks





Predicting Missing Links

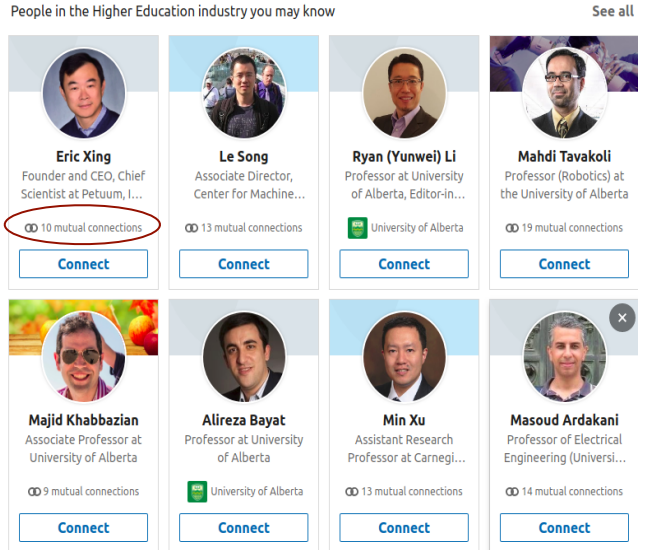
Only a subset of edges are observed, and graph is sparse ⇒ searching for O(n) needles in a $\Theta(n^2)$ haystack and a random baseline would have 1/n accuracy

Define a **Score(i , j)** over unconnected pairs

local topological predictors

- Number of common neighbours
- Number of shortest paths
- Product of degree
- Same cluster
- …

Which one corresponds to the common neighbours?



Slides based on lecture notes from Caluset

Local Neighbourhood Scores

- Common Neighbours

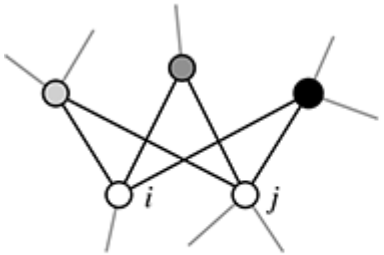
$$n_{ij} = \sum_k A_{ik} A_{kj} = |\mathcal{N}(i) \cap \mathcal{N}(j)|$$

Neighbours of node j

- Jaccard similarity

$$J_{ij} = \frac{\sum_k A_{ik} A_{kj}}{d_i + d_j - \sum_k A_{ik} A_{kj}} = \frac{n_{ij}}{d_i + d_j - n_{ij}} = \frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}$$

Example:



$n_{ij} = 3$

$J_{ij} = 3/6$

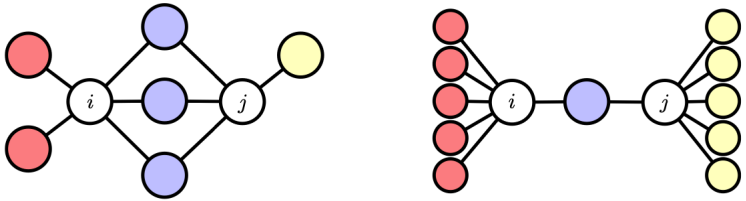
Link Prediction with Jaccard Similarity

$$J_{ij} = \frac{\sum_k A_{ik}A_{kj}}{d_i + d_j - \sum_k A_{ik}A_{kj}} = \frac{n_{ij}}{d_i + d_j - n_{ij}} = \frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}$$

We can add a small noise to break the ties randomly: $score(i, j) = J_{ij} + \epsilon$

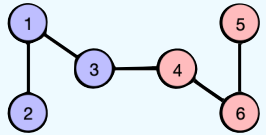
What happens to the network if we add edges based on this? Closes triangles

What is J_{ij} for this example ?



0.50 (3/6) vs **0.091** (1/11)

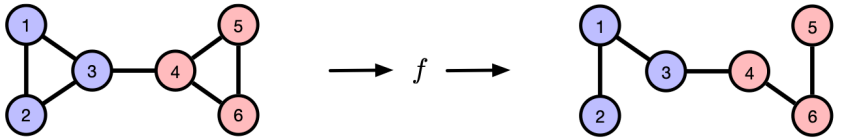
Example:



i	j	Jaccard score(i, j)
4	**5**	$1/2 + r$
2	**3**	$1/2 + r$
3	6	$1/3 + r$
1	4	$1/3 + r$
1	5	r
1	6	r
2	6	r
2	4	r
2	5	r
3	5	r

Link prediction: comparing different scores

Consider the example below and two predictors, Jaccard and Degree product



actual $G = (V, E)$ observed $G^\circ = (V, E^\circ)$

Which one is more accurate?

We can consider a threshold and convert this to a binary classification, and use any of the performance metrics for that, e.g. accuracy

i	j	Jaccard $score(i,j) = J_{ij}$
4	**5**	$1/2 + r$
2	**3**	$1/2 + r$
3	6	$1/3 + r$
1	4	$1/3 + r$
1	5	r
1	6	r
2	6	r
2	4	r
2	5	r
3	5	r

Predict edge

i	j	degree product $score(i,j) = d_i d_j$
1	4	$4 + r$
1	6	$4 + r$
3	6	$4 + r$
1	5	$2 + r$
2	**3**	$2 + r$
2	6	$2 + r$
2	4	$2 + r$
3	5	$2 + r$
4	**5**	$2 + r$
2	5	$1 + r$

Based on Preferential attachment

Measuring performance for binary classification

Reminder.

		Truth		Σ
		TP	FP	RP
Results		FN	TN	RN
Σ		P	N	

Defined based on four values: TP (positive in both results and ground-truth), FP (positive in results but not the ground truth), FN (negative in the results but positive in the ground-truth), and TN (negative in results and ground truth)

Accuracy = (TP+TN) / (P+N)
Precision = TP / RP
Recall = TP / P {also called sensitivity}
F1 score = 2. Precision x Recall / (Precision + Recall)
{Harmonic mean}

Miss rate = FN / P
Fallout = FP / N {also called false positive rate}
False discovery rate = FP / RP
Selectivity = TN / N {also called specificity}
False omission rate = FN / RN
Negative predictive value = TN / RN

These measures depend on the threshold

The tradeoff between precision and recall

Max threshold: everything negative
Perfect Fallout & Precision

Min threshold: everything positive
Perfect Recall

Threshold Invariant Metrics: ROC & AUC

Reminder.

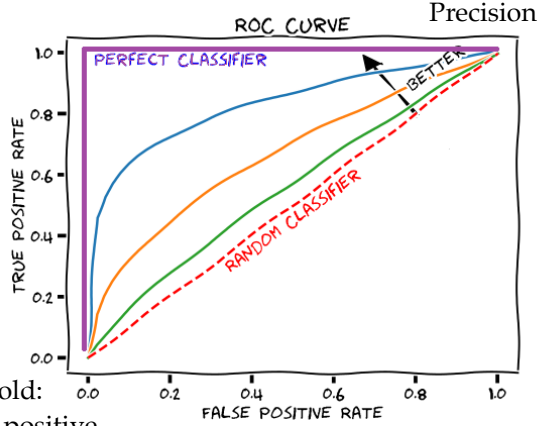
Receiver Operating Characteristic (ROC) as a function of prediction threshold

TPR(t) = TP(t)/P (recall, sensitivity at t)
FPR(t) = FP(t)/N (fallout, false alarm at t)

Area Under the Curve (AUC) of ROC
gives the probability of ranking a random positive edge higher than a random negative edge

For **unbalanced** data use the Precision Recall Curve instead
(when the number of negatives and positives differ significantly)

Max threshold:
everything negative
Perfect Fallout &
Precision



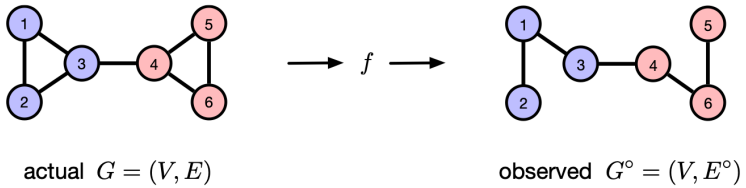
Min threshold:
everything positive
Perfect Recall

Example

every candidate (i, j) is either a missing link or not

TPR = TP/P (**recall**, sensitivity)
Links above threshold / Total Positives

FPR = FP/N (**fallout**, false alarm)
Non-links above threshold / Total Negatives



actual $G = (V, E)$ observed $G^{\circ} = (V, E^{\circ})$

$\text{AUC} = \sum_t \text{TPR(t)}\ (\text{FPR (t)- FPR (t-1)})$

box-rule approximation

	i	j	Jaccard score(i,j)
link	**4**	**5**	$1/2 + r$
link	**2**	**3**	$1/2 + r$
link	3	6	$1/3 + r$
threshold →			
nonlink	1	4	$1/3 + r$
nonlink	1	5	r
nonlink	1	6	r
nonlink	2	6	r
nonlink	2	4	r
nonlink	2	5	r
nonlink	3	5	r

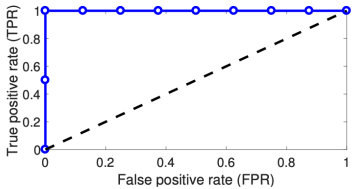
	i	j	degree product score(i,j)
	1	4	$4 + r$
	1	6	$4 + r$
	3	6	$4 + r$
	1	5	$2 + r$
	2	**3**	$2 + r$
	2	6	$2 + r$
	2	4	$2 + r$
	3	5	$2 + r$
	4	**5**	$2 + r$
	2	5	$1 + r$

	Truth		Σ
results	TP = 2	FP = 2	RP = 4
results	FN = 0	TN = 6	RN = 6
Σ	P = 2	N = 8	

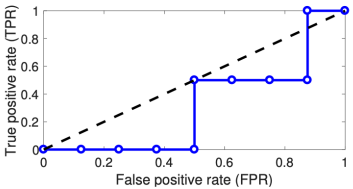
TPR = 2/2 = 1, FPR = 2/8=0.25

	Truth		Σ
results	TP = 0	FP = 4	RP = 4
results	FN = 2	TN= 4	RN = 6
Σ	P = 2	N = 8	

TPR = 0/2 = 0, FPR = 4/8=0.5



Jaccard coefficient, AUC = 1.00



degree product, AUC = 0.31

Topological Link Predictors

CN	common neighbors of i, j
SP	shortest path between i, j
LHN	Leicht-Holme-Newman index of neighbor sets of i, j
PPR	j-th entry of the personalized page rank of node i
PA	preferential attachment (degree product) of i, j
JC	Jaccard's coefficient of neighbor sets of i, j
AA	Adamic/Adar index of i, j
RA	resource allocation index of i, j
LRA	entry i, j in low rank approximation (LRA) via singular value decomposition (SVD)
dLRA	dot product of columns i and j in LRA via SVD for each pair of nodes i, j
mLRA	average of entries i and j's neighbors in low rank approximation
LRA-approx	an approximation of LRA
dLRA-approx	an approximation of dLRA
mLRA-approx	an approximation of mLRA
LCC_i, LCC_j	local clustering coefficients for i and j
AND_i, AND_j	average neighbor degrees for i and j
$SPBC_i$, $SPBC_j$	shortest-path betweenness centralities for i and j
CC_i, CC_j	closeness centralities for i and j
DC_i, DC_j	degree centralities for i and j
EC_i, EC_j	eigenvector centralities for i and j
KC_i, KC_j	Katz centralities for i and j
LNT_i, LNT_j	local number of triangles for i and j
PR_i, PR_j	Page rank values for i and j
LC_i, LC_j	load centralities for i and j

There are many **alternatives** to give a score to a given pair of nodes (i,j)

Many also used as similarity measures since homophily is a strong force in link creation

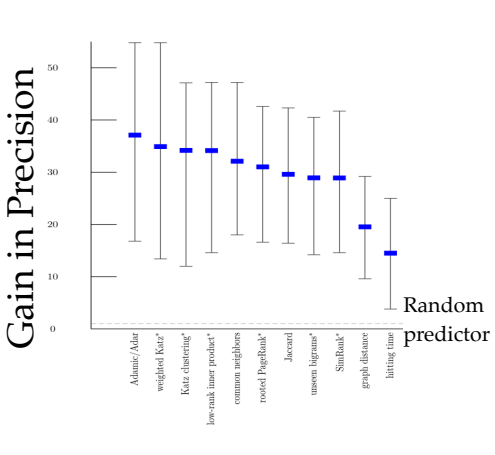
See A Survey of Link Prediction in Complex Networks, 2015 & Link prediction in complex networks: A survey, 2011

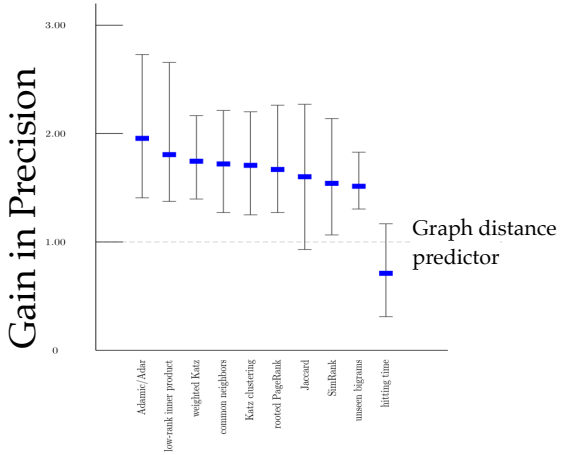
Which one to use?

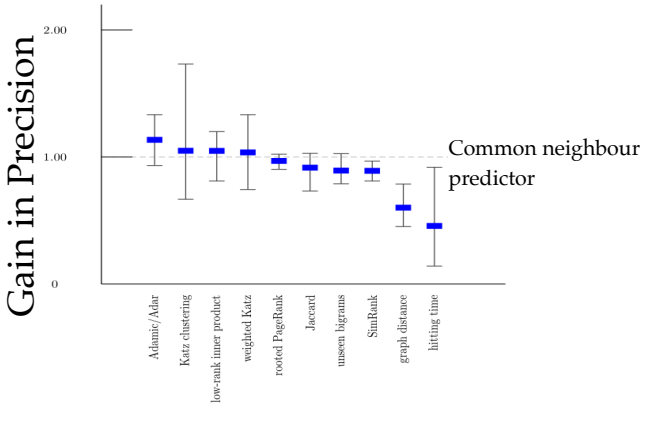
no one predictor is best, or worst, across all realistic inputs

See: Ghasemian A, Hosseinmardi H, Galstyan A, Airoldi EM, Clauset A. Stacking Models for Nearly Optimal Link Prediction in Complex Networks. PNAS 2020.,

Common Neighbour is a Strong Predictor







Adamic/Adar: a notable measure

Sums over size inverse of log of degree of common neighbours

$$AA(i,j) = \sum_{k \in \mathcal{N}(i) \cap \mathcal{N}(j)} \frac{1}{|\mathcal{N}(k)|} = \sum_{k} \frac{A_{ik} A_{kj}}{d_k}$$

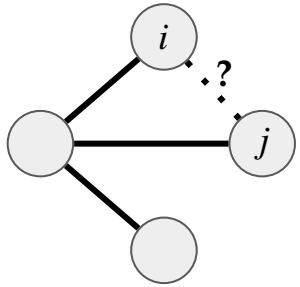
Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American society for information science and technology. 2007

Common prediction tasks

- Link Prediction

  Classic example:

  $$z(i,j) = |\mathcal{N}(i) \cap \mathcal{N}(j)|$$



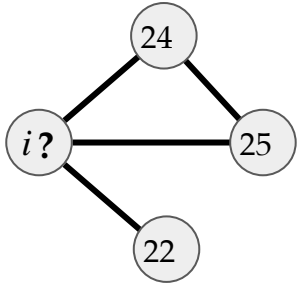
$z(i,j)$

e.g. how likely it is for them to become friend?

- Node Classification

  What can be a simple predictor?

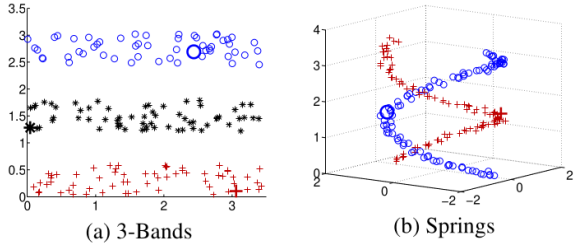
  $$z(i) = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} z(j)$$



$z(i)$

e.g. what is the age of a user based on his friends ages?

Label Propagation Algorithm

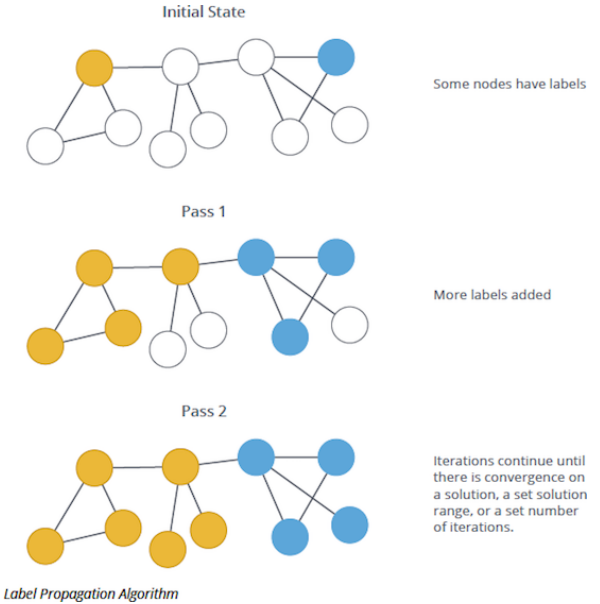


proposed originally for semi-supervised classification of iid data by defining a fully connected distance graph

Transition matrix $M = AD^{-1}, M_{ij} = \frac{A_{ij}}{\sum_k A_{kj}}$

row normalize $\hat{M}_{ij} = \frac{M_{ij}}{\sum_k M_{ik}}$

Repeat: Set the labeled data in Y and propagate labels with $Y = \hat{M}Y$



What is the main assumption here?

Homophily, similar nodes are connected

Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation.

Feedback Loop of Relations & Characteristics

There is an interplay between characteristics of nodes and their relations, a positive feedback loop derived by two social theories:

- Social Selection
  - Similarity of individuals' characteristics motivates them to form relations

⇒ Similarity of node's attributes is a link predictors in addition to structure proximity

- Social Influence
  - Characteristics of individuals may be affected by the characteristics of their relations

⇒ Your neighbours' attributes can reveal yours

Assortativity & Homophily

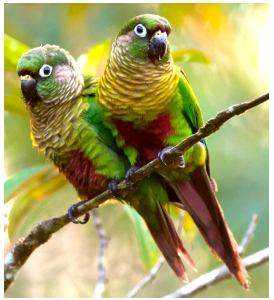
Similar nodes tend to link to each other

How to measure age homophily in a given friendship graph when you know age of every node?

Similar to degree assortativity, measure the correlation of age across all edges

How to measure occupation homophily?

birds of the same feather flock together



Assortativity & Homophily: Mixing Matrix

Mixing matrix measures the **ratio** of edges connecting each pair of attribute values

What indicates homophily in this matrix? How does homophily look like? **dominant diagonal**

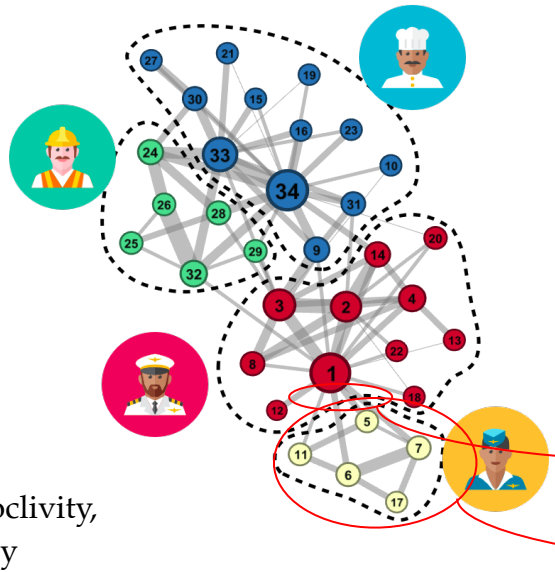
Assortativity index

$$r = \frac{\sum_i e_{ii} - \sum_i e_{i.}e_{.i}}{1 - \sum_i e_{i.}e_{.i}} = \frac{Tr[E] - \|E^2\|}{1 - \|E^2\|}$$

This is a normalized Q-modularity assuming the characteristic partitions the graph

$$Q = \sum_i e_{ii} - \sum_i e_{i.}e_{.i} = Tr[E] - \|E^2\|$$

$$E_{kl} = \sum_{ij} \frac{A_{ij}}{2m}\delta(C(i),k)\delta(C(j),l) = \sum_{i\in c_k,\ j\in c_l} \frac{A_{ij}}{2m}$$



	B	G	R	Y
B	44	7	2	0
G	7	12	7	0
R	2	7	38	4
Y	0	0	4	12

/2m

Is there other mixing patterns?

We can have any proclivity, including heterophily

Mixing Patterns & Structural Correlation

Mixing matrix shows the number of edges connecting each pair of attribute values

Other mixing patterns, such as **heterophily** can also be reflected in the mixing matrix

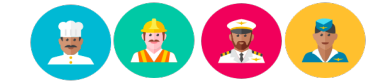
How does heterophily look like in the mixing matrix? How does correlation look like in the mixing matrix?

Low dispersion, a **dominant cell** in each row and column

How to quantify the overall correlation Does it resemble anything?

confusion matrix for pairwise cluster overlaps but with edges between pair of values

measure the total dispersion similar to clustering agreement indexes



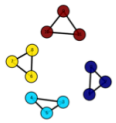
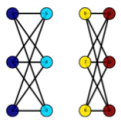
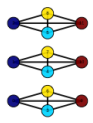
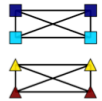
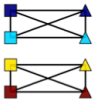
	B	G	R	Y
B	44	7	2	0
G	7	12	7	0
R	2	7	4	19
Y	0	0	19	12

/2m

Proclivity "inclination or predisposition toward a particular thing"

Can also be used to measure cross-correlation: correlation between your income and occupations of your friends

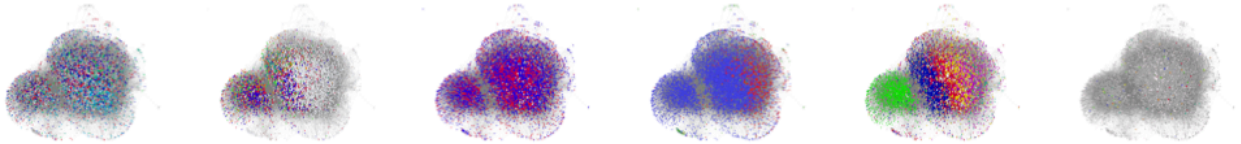
Toy Examples:

	(i) self-proclivity: Given: node's color; Guess: neighbor's colors			(ii) cross-proclivity: Given: node's color; Guess: neighbor's shape	
	i.1	i.2	i.3	ii.1	ii.2
					
	homophily	heterophily	random	correlation	no correlation
Q [16]	0.75	-0.25	-0.25	✗	✗
r [17]	**1.0**	-0.33	-0.33	✗	✗
ProNE$_l$	**1.0**	**1.0**	0.21	0.67	**0.0**
ProNE$_2$	**1.0**	**1.0**	0.11	0.5	**0.0**
ProNE$_3$	**1.0**	**1.0**	0.05	0.33	**0.0**

Example:
Facebook friendship network of a US college
Color: distinct attribute values
Placement : connections

	major	gender	year	status	dorm	highschool	minor
major	0.01	0.00	0.01	0.00	0.01	**0.05**	0.01
gender	0.00	0.00	0.00	0.00	0.00	0.00	0.00
year	0.01	0.00	**0.25**	**0.07**	**0.07**	**0.07**	0.01
status	0.00	0.00	**0.07**	**0.09**	0.02	0.02	0.00
dorm	0.01	0.00	**0.07**	0.02	**0.16**	**0.10**	0.02
highschool	**0.05**	0.00	**0.07**	0.02	**0.10**	**0.31**	**0.07**
minor	0.01	0.00	0.01	0.00	0.02	**0.07**	0.01

correlation between the entrance year of a student and the high school his friends are from, the dormitory they are in and most strongly their entrance year



major
62(76) values
9.94% missing

dorm
23(25) values
48.2% missing

gender
2(2) values
5.87% missing

status
5(6) values
0.03% missing

year
9(20) values
12% missing

highschool
198(2881) values
13.7% missing

From Beyond Assortativity: Proclivity Index for Attributed Networks, 2017

Even if you don't put your information online, that information can be inferred/predicted based on what your friends reveal about themselves