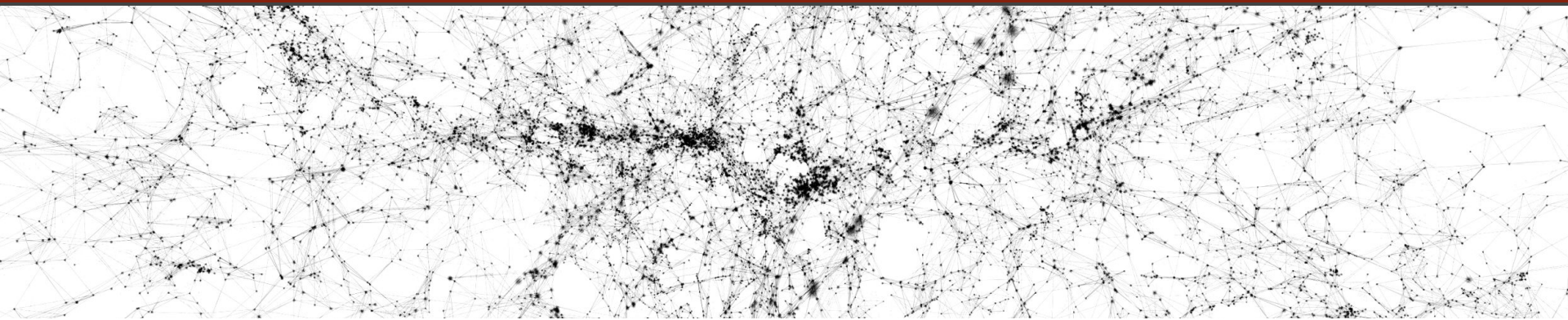
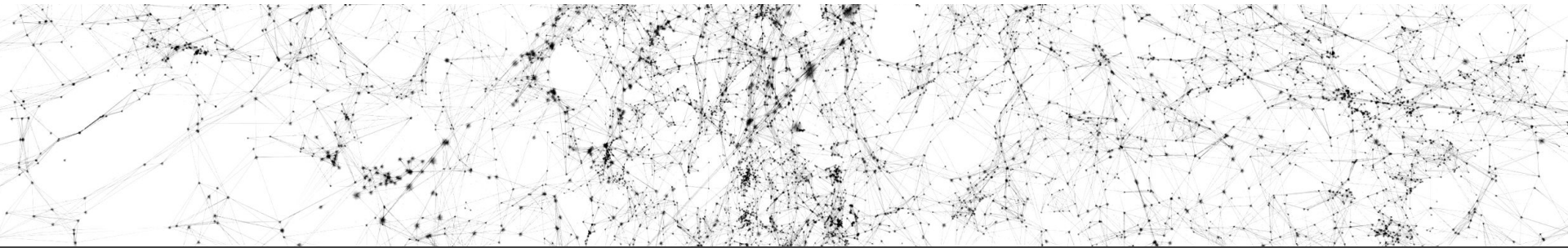


Node Classification

Analysis of complex interconnected data



Quick Notes

- Last assignment is out and is due on Oct 18th
  - http://www.reirab.com/Teaching/NS20/Assignment_3.pdf
  - Submit 2 files (report.pdf, code.zip) as a Group (pairs or two or individual) in Mycourses
- Please select your seminars
  - Due date for selection: Tuesday night (let me know if there are any issues)
  - Please put your name for **two** slots, do not change the entries that are already booked
  - Find the link to editable spreadsheet in slack or Mycourses
- Any questions?

Deadlines

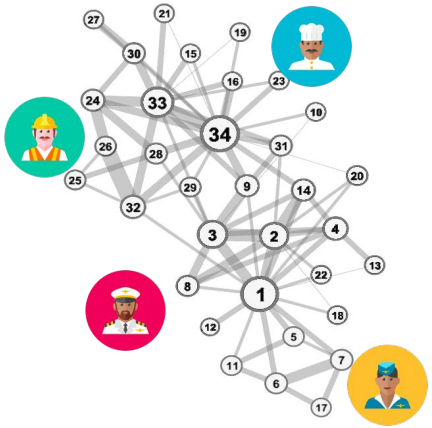
- assignment 1 due on Sep. 20th
- assignment 2 due on Oct. 4th
- assignment 3 due on Oct. 18th
- project proposal slides due on Oct. 25th
- project proposal due on Nov. 1th
- Reviews (first round) due on Nov. 8th
- project progress report due on Nov. 22nd
- Reviews (second round) due on Nov. 29th
- project final report slides due on Dec. 1st
- project final report due on Dec. 6th
- Reviews (third round) due on Dec. 13th
- project revised report and rebuttal due on Dec. 20th
- note: dates are tentative, please check them for the updated deadlines

Link Prediction- Quick recap

- Heuristic predictors
  - Rank the pairs of nodes using a similarity score
  - Number of common neighbours, Jaccard, Adamic/Adar and many more
- Measuring performance
  - Turn into binary classification using a cutoff threshold for links v.s. do not link score
  - Use AUC for threshold independent measure
- Learn the predictors based on a set of features instead of a single score
- Learn the features as well
  - Directly learn the features for link prediction
  - Use any node embedding technique and use proximity in the embedded space
- Use clusters for link prediction
  - Links that optimize a clustering objective, e.g. modularity
  - Fit (with maximum likelihood) a SBM variation to the data to get the probability linking based on block assignments of the nodes

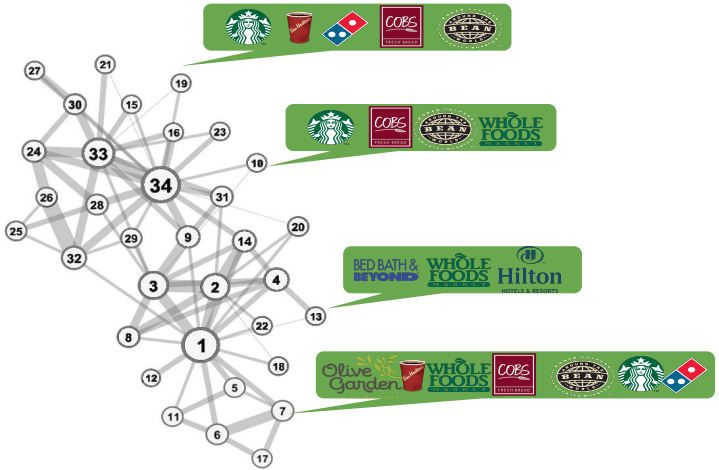
Attributed Graphs

Individual characteristics or activity (attributes) & relations (graph)



characteristics

age, occupation, salary, sex, etc.



activity & interest

check-ins, page-likes, group memberships, movies

Attributed Graphs

Interplay between attributes and 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

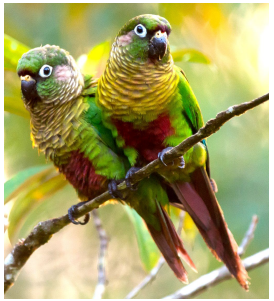
  ⇒ Your neighbours' attributes can reveal yours

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?

birds of the same feather flock together



Assortativity & Homophily

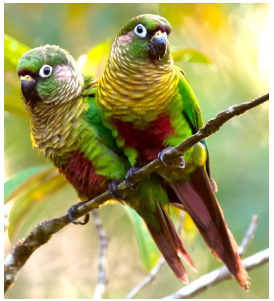
Similar nodes tend to link to each other

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?

categorical attribute instead of a numeric one

birds of the same feather flock together



Assortativity & Homophily

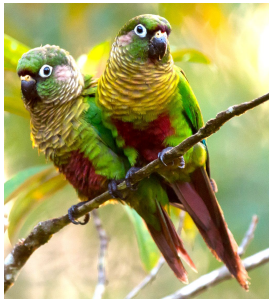
Similar nodes tend to link to each other

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?

categorical attribute instead of a numeric one

birds of the same feather flock together

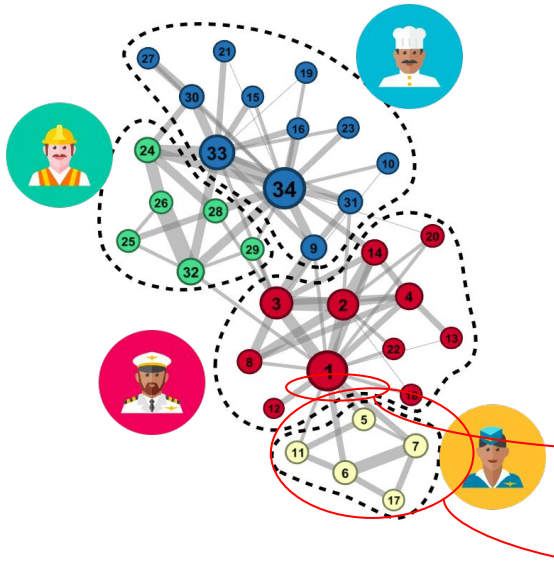


Assortativity & Homophily

Look at the mixing patterns

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

What indicates homophily in this matrix?
How does homophily look like?



	B	G	R	Y
B	22	7	2	0
G	7	6	7	0
R	2	7	19	4
Y	0	0	4	6

Assortativity & Homophily

Look at the mixing patterns

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

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

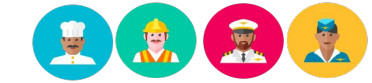
$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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

Is normalized Q-modularity assuming attributes partition the graph

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



	B	G	R	Y
B	22	7	2	0
G	7	6	7	0
R	2	7	19	4
Y	0	0	4	6

/sum (E)

Assortativity & Homophily

Look at the mixing patterns

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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

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

Is there other mixing patterns?

Assortativity & Homophily

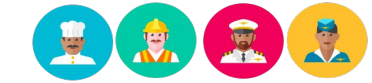
Look at the mixing patterns

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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



	B	G	R	Y
B	22	7	2	0
G	7	6	7	0
R	2	7	4	19
Y	0	0	19	6

/sum (E)

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

Is there other mixing patterns?

e.g. opposites attract

Structural Correlation

Look at the mixing patterns

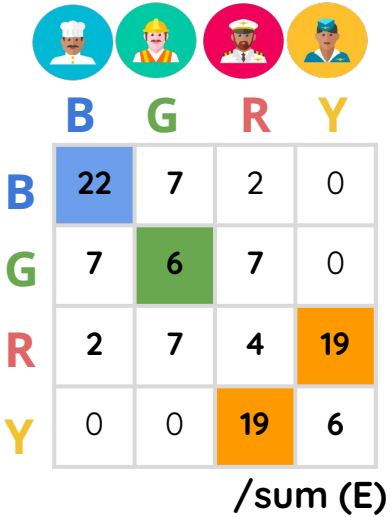
Mixing matrix shows the number of edges connecting each pair of attribute values

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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



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

Structural Correlation

Look at the mixing patterns

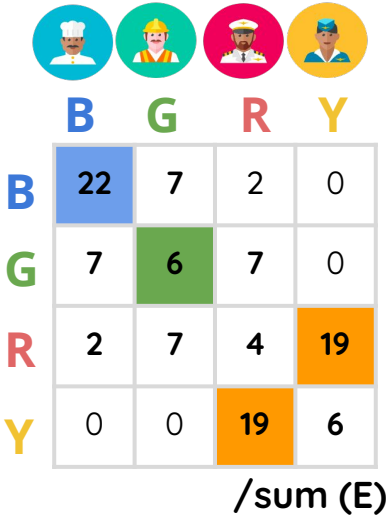
Mixing matrix shows the number of edges connecting each pair of attribute values

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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



How does heterophily look like in the mixing matrix?
How does **correlation** look like in the mixing matrix?
A **dominant cell** in each row and column

Structural Correlation

Look at the mixing patterns

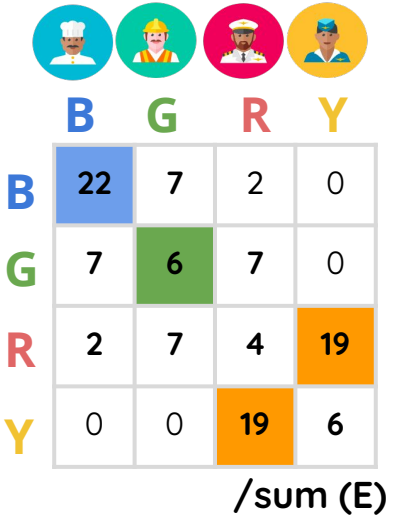
Mixing matrix shows the number of edges connecting each pair of attribute values

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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



How to quantify the overall correlation?

Does it resemble anything?

Structural Correlation

Look at the mixing patterns

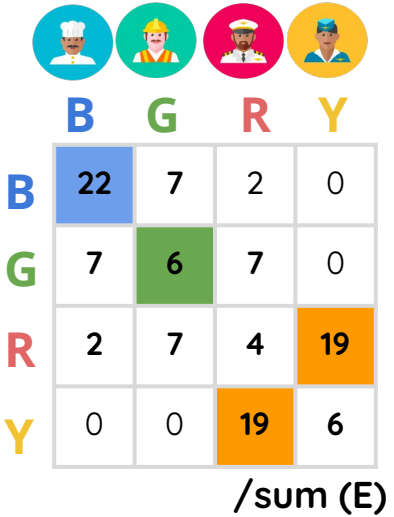
Mixing matrix shows the number of edges connecting each pair of attribute values

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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



How to quantify the overall correlation?

Does it resemble anything? confusion matrix where pairwise cluster overlaps are changed to edges between pair of values

Structural Correlation

Look at the mixing patterns

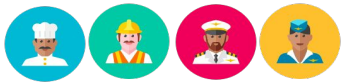
Mixing matrix shows the number of edges connecting each pair of attribute values

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

$\mathbf{e_{ij}}$: ratio of edges between each pair of values

Assortativity index

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



.2	.07	.02	0
.07	.06	.07	0
.02	.07	.04	.17
0	0	.17	.06

How to quantify the overall correlation?

measure the total dispersion similar to clustering agreement indexes

Beyond Assortativity: Proclivity Index for Attributed Networks, PAKDD (2017)

Structural Correlation of Attributes

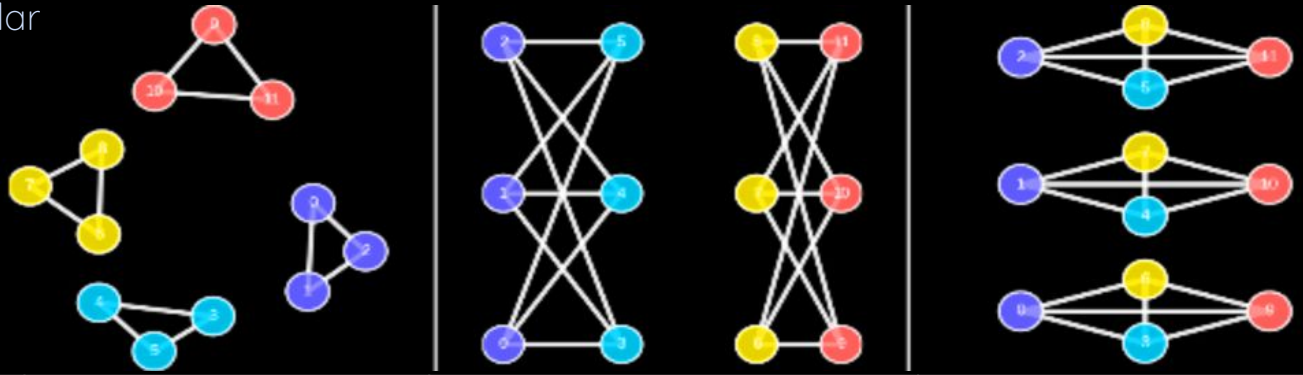
Proclivity

"inclination or predisposition toward a particular thing"

similar
homophily

different
heterophily

random



Assortativity	**1.0**	**-0.33**	**-0.33**
Prone	**1.0**	**1.0**	**0.11**

Beyond Assortativity: Proclivity Index for Attributed Networks. PAKDD (2017)

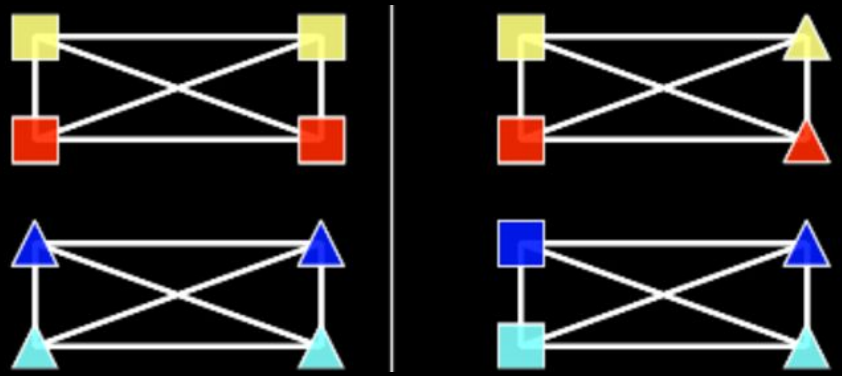
Structural Correlation of Attributes

Proclivity

Cross-proclivity in addition to self-proclivity

Shape and Color

correlation | no correlation

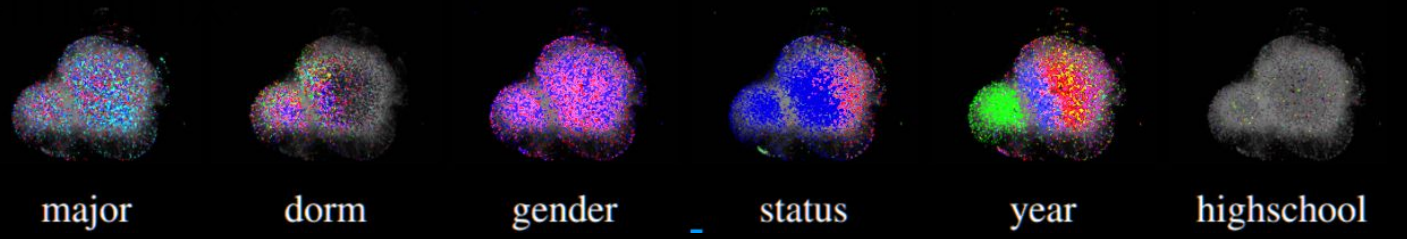


Assortativity	X	X
Prone	**0.5**	**0.0**

Correlation between your income and occupations of your friends

Structural Correlation of Attributes

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



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

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

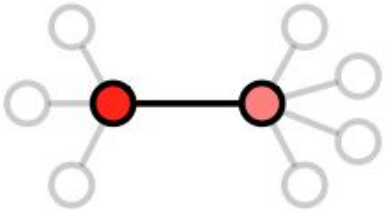
Even if you don't put your information online, that information can be inferred/predicted based on what your friends reveal about themselves

Predicting missing node attributes

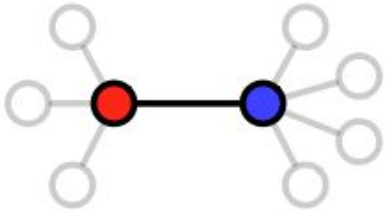
Most graphs are incomplete, and often attributes of some nodes are missing

We can use structural correlations to predict a missing attribute to be e.g. the average (scalar) or most common (categorical) value of its neighbors' non-missing attributes

What does this local smoothing assume about the mixing patterns?



assortative mixing



disassortative mixing

From Clauset's Slides

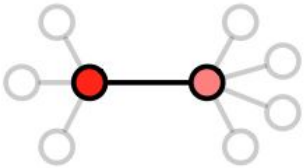
Predicting missing node attributes

Most graphs are incomplete, and often attributes of some nodes are missing

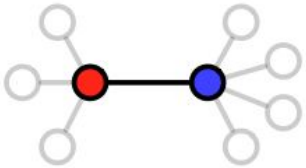
We can use structural correlations to predict a missing attribute to be e.g. the average (scalar) or most common (categorical) value of its neighbors' non-missing attributes

What does this local smoothing assume about the mixing patterns?

mean (scalar) & mode (categorical) ⇒ assortative



assortative mixing



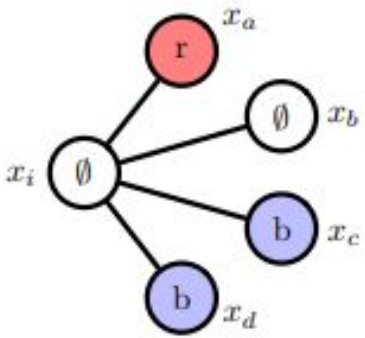
disassortative mixing

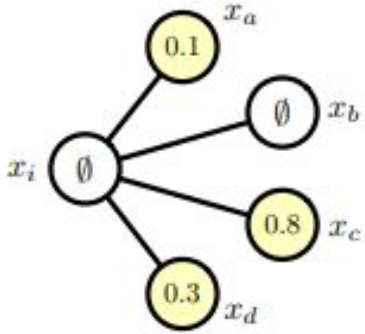
From Clauset's Slides

Predicting missing node attributes, example

missing = mean (scalar) & mode (categorical)

what is the prediction in these two cases for node i?



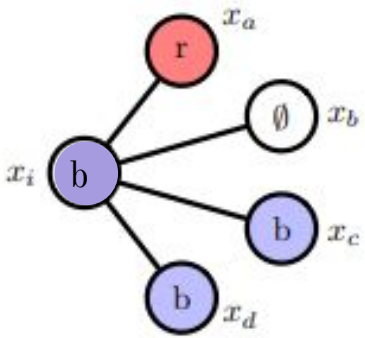


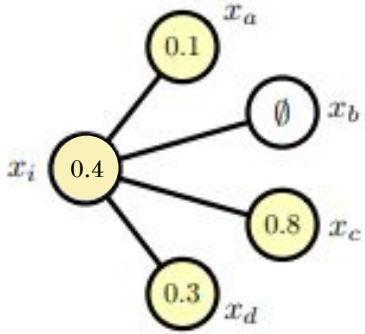
From Clauset's Slides

Predicting missing node attributes, example

missing = mean (scalar) & mode (categorical)

what is the prediction in these two cases for node i?



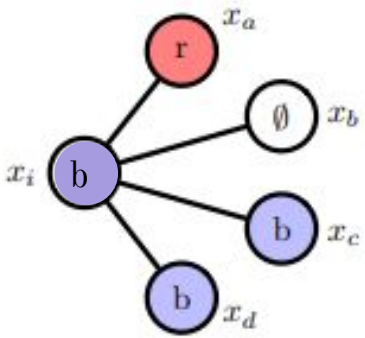


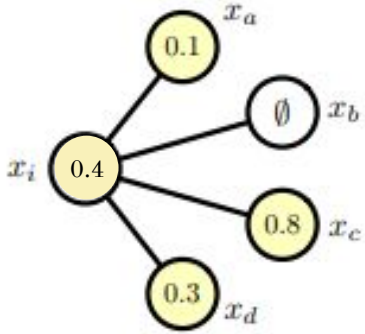
what is the predictions for node x_b ?

Predicting missing node attributes, example

missing = mean (scalar) & mode (categorical)

what is the prediction in these two cases for node i?



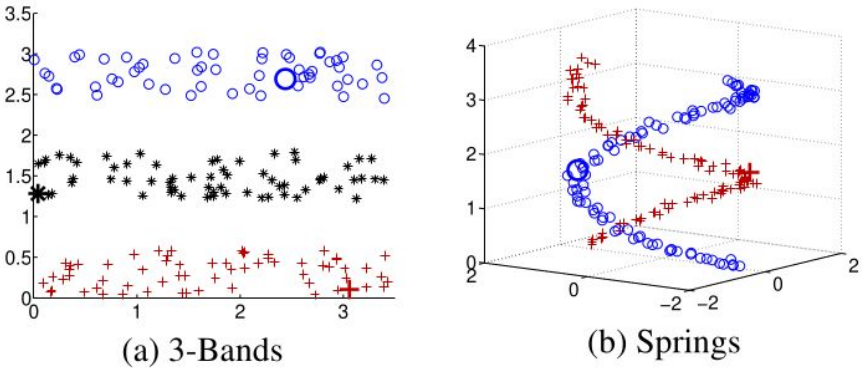


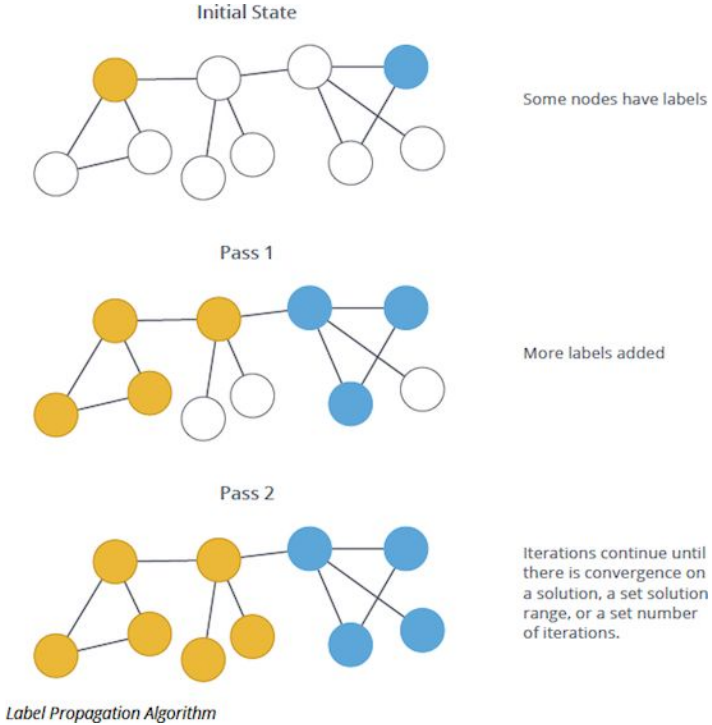
what is the predictions for node x_b ? repeat given current predictions

Label Propagation Algorithm

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

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

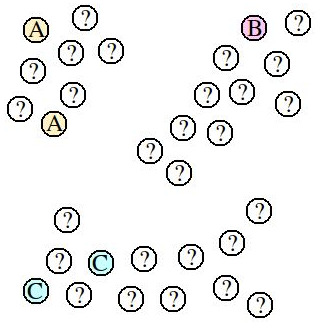




Label Smoothing

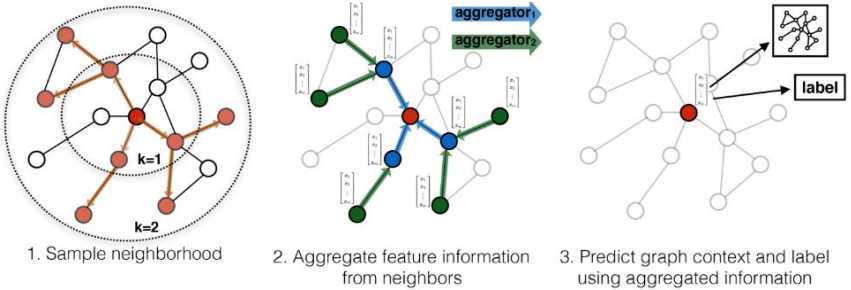
Node classification

- Unsupervised learning
  - clustering, only graph is given, classes/clusters are not predefined
- Supervised learning
  - classifying, input is graph and labels on all nodes
    - You mask some nodes (labels and their connections) for training [inductive]
    - You mask some nodes (only labels) for training [transductive]
- Semi-supervised learning
  - input is graph and labels on some nodes
  - You mask some node labels for training (seeing the whole graph: transductive)
- Active learning
  - Input is graph and a budget that determines how many nodes you can query for labels
  - labels come in sequence and can be queried based on the current set



Semi-Supervised Node classification

- Traditional
  - label propagation & belief propagation
- Recent end-to-end methods (Feature Smoothing)
  - GCN and variants, which use a classification loss
- Embedding based
  - Unsupervised embedding extraction (e.g. node2vec) then apply a classifier



1. Sample neighborhood
2. Aggregate feature information from neighbors
3. Predict graph context and label using aggregated information

Unifying Graph Convolutional Neural Networks and Label Propagation, 2020

Measuring performance of attribute prediction

Scalar values ⇒ correlation of predicted & actual values (r^2 correlation)

Categorical values ⇒ confusion matrix & average accuray

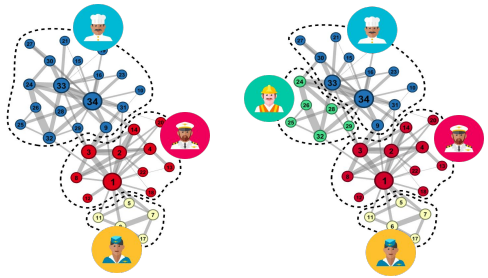
C_{ij} = the number of nodes with predicted label i and actual label j

	G	B	R	Y	Σ
G	0	0	0	0	**0**
B	**12**	**6**	0	0	18
R	0	0	**11**	0	11
Y	0	0	0	**5**	5
Σ	12	6	11	5	34

Accuracy = 1/34 Tr(C)

truth \ results		
	90	0
	10	0

What is the accuracy?



Measuring performance of attribute prediction

Scalar values ⇒ correlation of predicted & actual values (r^2 correlation)

Categorical values ⇒ confusion matrix & average accuray

C_{ij} = the number of nodes with predicted label i and actual label j

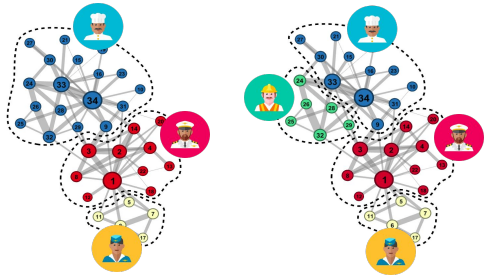
	G	B	R	Y	Σ
G	0	0	0	0	0
B	12	6	0	0	18
R	0	0	11	0	11
Y	0	0	0	5	5
Σ	12	6	11	5	34

Accuracy = 1/34 Tr(C)

truth \ results		
	90	0
	10	0

What is the accuracy? 90% but is always guessing the majority class and never getting the minority class correct

Class imbalance problem

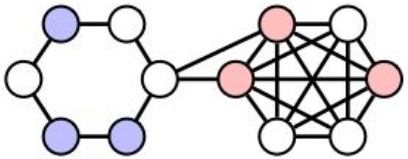


Measuring performance of attribute prediction

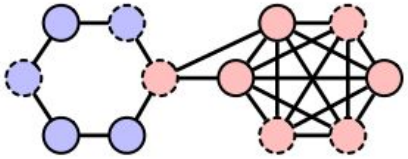
Example

Truth: left circle all blue, right circle all red

Observed: 6 missing values



observed



predicted

What is the accuracy?

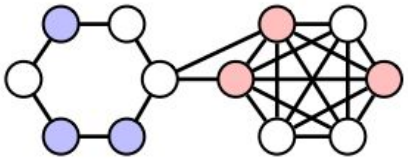
	C	actual r	actual b
pred.	r	3	1
pred.	b	0	2

Measuring performance of attribute prediction

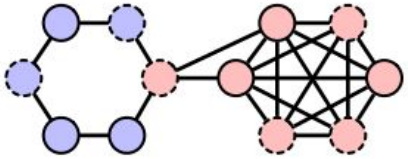
Example

Truth: left circle all blue, right circle all red

Observed: 6 missing values



observed



predicted

What is the accuracy?

		actual	
	C	r	b
pred.	r	3	1
	b	0	2

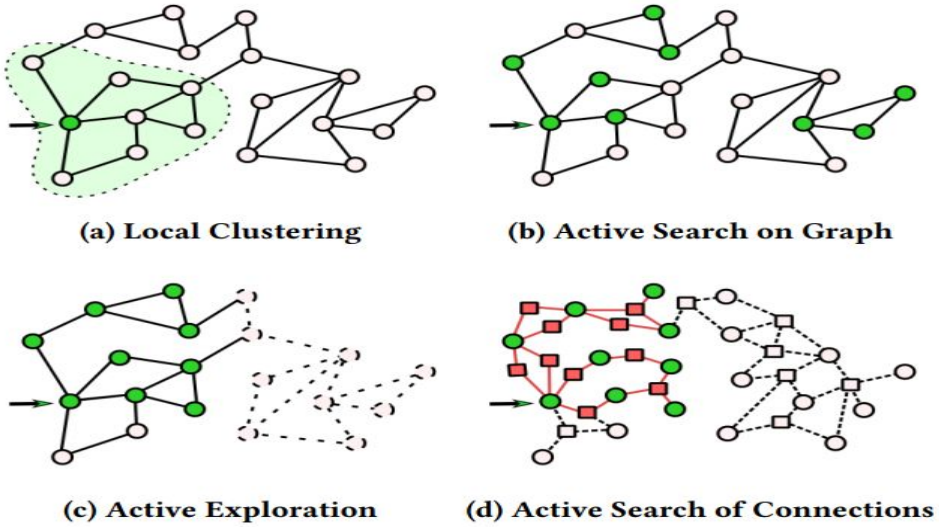
Accuracy = ⅚ = 0.83

Active Search of Connections

What if you can query with some cost or given a budget?

Semi-supervised ⇒ Active learning

labels are local and depend on the given seed



Active Search of Connections for Case Building and Combating Human Trafficking, 2018