Predicting Nodes Attributes using Network Structure

Sarwan Ali

joint work with

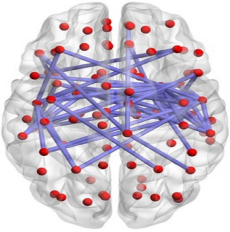
M H Shakeel, M A Khan, S. Faizullah, I Khan

appeared in ACM Transactions on Intelligent Systems and Technology

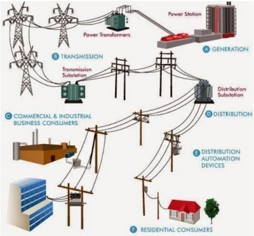
Graphs model many systems

Graphs: entities (**nodes**) interconnected (**through edges**)

World Wide Web



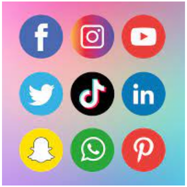
Biological Network



Power Network

Social Network

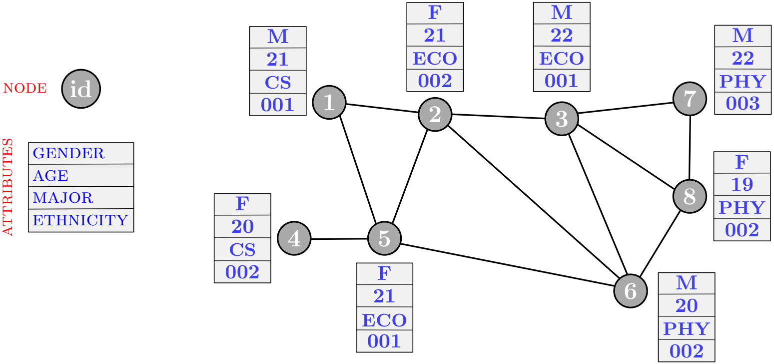
The Internet



Online Social Nework

Attributed Graph

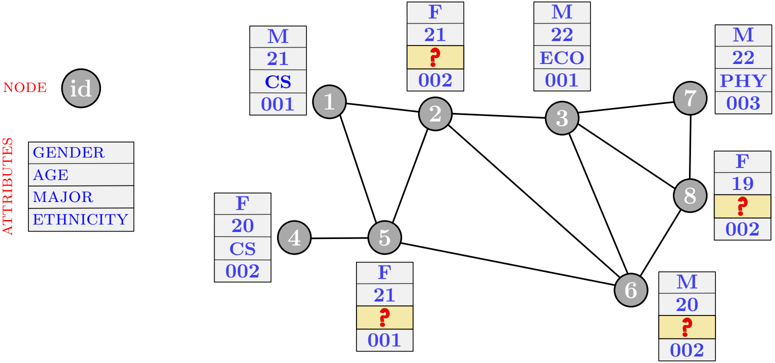
Nodes in attributed graphs have additional properties/attributes



Node Attributes Prediction

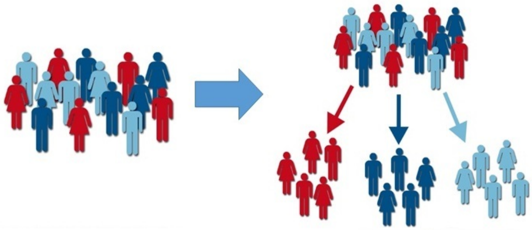
Attributes of some nodes can be missing

- Goal: predict missing attributes of nodes



Attribute Prediction Application in Targeted Advertisement

Determine characteristics of consumers from social network



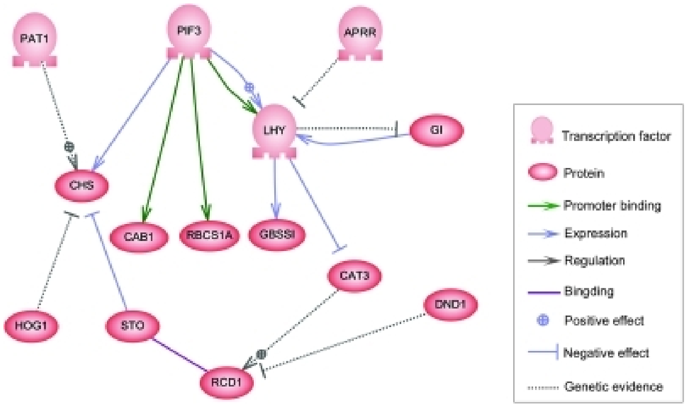
Attribute Prediction Application in Privacy and Security

Test privacy preservation of anonymization schemes

Find the likelihood of users sharing fake news

image source: https://hotspotshield.com/

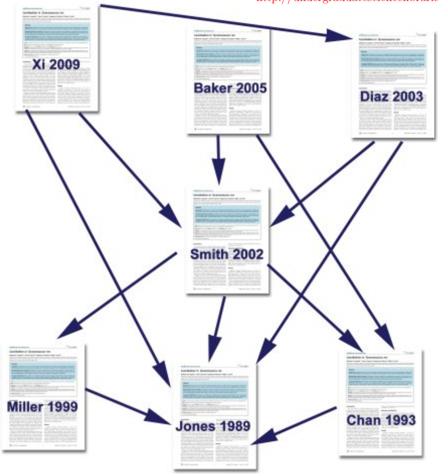
Determine structural and functional properties of proteins in Protein-Protein Interaction networks



Attribute Prediction Application in Research Informatics

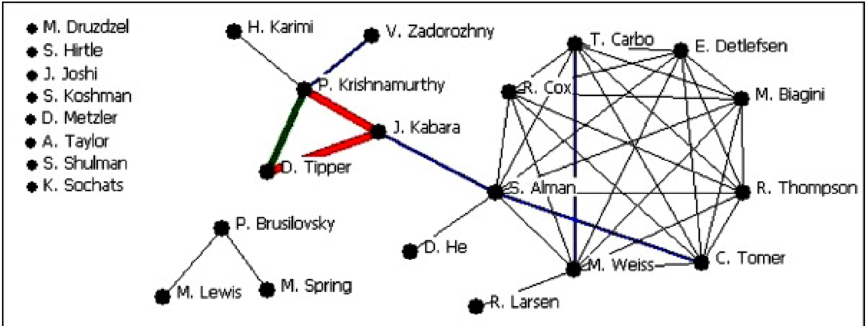
Determine subject areas of research papers in citation networks

http://undergraduatesciencelibrarian.org/



Attribute Prediction Application in Research Informatics

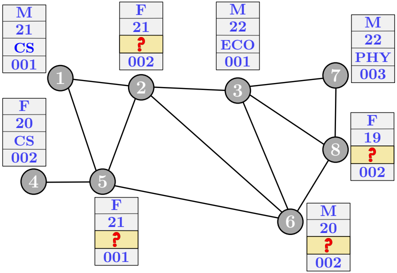
Determine research areas of scholars in coauthorship networks

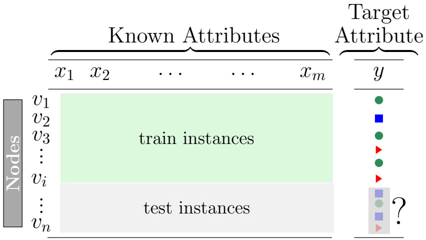


Abbasi et.al (2011) Journal of Informetrics

Attribute Prediction as a classification problem

- Nodes as feature vectors 'Desired attribute' as class label





Issue with Attribute Prediction as classification

- Does not take into account ‘network structure’

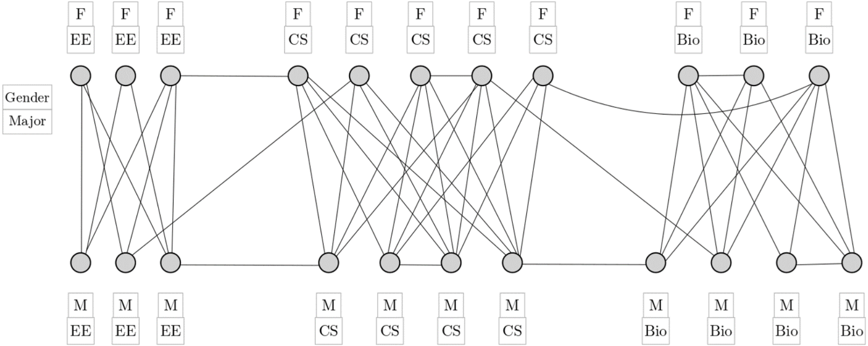
Attribute values and network structure are highly inter-dependent

Two important phenomena in Sociology

- Social Selection: Individual’s attributes drive the interaction with others
- Social Influence: Interactions between people shape their attributes

Attribute values and network structure are highly inter-dependent

- Homophily: Connections among nodes having same attribute values
- Heterophily: Connections among nodes having different attribute values



'MAJOR' attribute is homophilic

'GENDER' attribute heterophilic

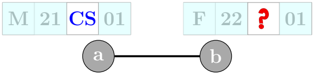
From Sociology to Data Science

Social Selection, Social Influence, Homophily, and Heterophily are abstract concepts, not quantified measures

▷ We give a metric to measure dependency between attributes

Dependency of node attributes on interconnections is limited to the same attribute of ‘friends’

If a’s MAJOR is ‘CS’, what is MAJOR of his friend b?



▷ We quantify dependency of an attribute on all other attributes of friends

Only considers direct connections-not mult-hop connections

▷ We consider remote connections

No work on using interconnections to predict attributes

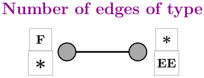
▷ We give efficient and explainable attribute prediction algorithm

Mixing Matrix: Summary of Interconnections two attributes **a** and **b** is

- A row/column corresponding to each possible value of attribute **a**/**b**
- $(i, j)^{th}$ entry of $M_{(\mathbf{a},\mathbf{b})}$ is the number of edges connecting nodes with attribute value a_i of **a** to nodes with attribute value b_j of **b**

$$M_{(\mathbf{a},\mathbf{b})}(i,j) = \left|\left\{(u,v) \in E : \mathbf{a}(u) = a_i \;\text{ AND }\; \mathbf{b}(v) = b_j\right\}\right|$$

	CS	EE	Bio
M	23	9	13
F	16	10	14



Mixing Matrix of
GENDER & MAJOR

Divergence of Mixing Matrix: The spread of values in $M_{(\mathbf{a},\mathbf{b})}$

- The divergence D_f of a matrix M with respect to a function f is

$$D_f = \frac{\sum_i [f(e_{i\cdot}) - \sum_j f(e_{ij})] + \sum_j [f(e_{\cdot j}) - \sum_i f(e_{ij})]}{\sum_i f(e_{i\cdot}) + \sum_j f(e_{\cdot j}) - 2\sum_i \sum_j f\left(\frac{e_{\cdot j} e_{i\cdot}}{e_{\cdot\cdot}}\right)}$$

  f can be $= x^2, x^3$, or $x \log x$
  - $e_{i\cdot}$: sum of values in the i^{th} row
  - $e_{\cdot j}$: sum of values in the j^{th} column
  - $e_{\cdot\cdot}$: sum of all entries of the matrix M
- The numerator aggregates the per-row and per-column divergences of this matrix, while the denominator normalizes this quantity using the maximum divergence value when the marginals are fixed

Essentially a measure of non-uniformity of the matrix

Proclivity Value

- The proclivity value "PRONE" (self/cross) between a pair of attributes, also called correlation or agreement between two attributes, (based on $M_{(\mathbf{a},\mathbf{b})}$) is inversely proportional to the divergence of $M_{(\mathbf{a},\mathbf{b})}$
- PRONE value between two attributes **a** and **b** is defined as:

$$\text{PRONE}_{(\mathbf{a},\mathbf{b})} := \rho_{(\mathbf{a},\mathbf{b})} := 1 - D_f$$

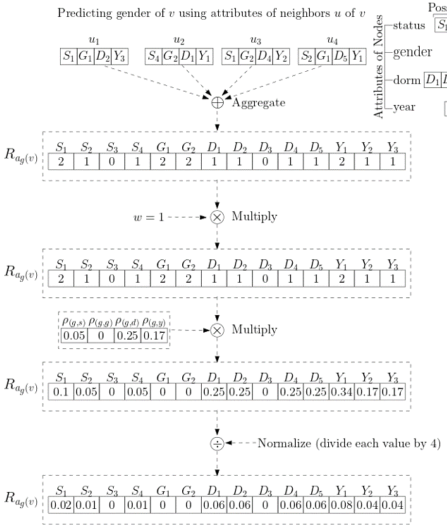
Proposed Approach

ALGORITHM 1: N-FVR (Graph $G = (V, E)$, $A(v), \forall\, v \in V$, hop length h, attribute weight ρ, hop weight w, attribute to be predicted a_t)

for $v \in V$ **do**
    $R_v^0 \leftarrow A(v)$
    for $i \leftarrow 1 : h$ **do** ▷ search depth
        $h^i \leftarrow []$
        for $j \leftarrow 1 : |A|$ **do**
            $vec \leftarrow w_i \times (\rho_{(a_t, a_j)}) \times (A_j(N^i(v)))$ ▷ from Equations (2), (3), and (6)
            $y^i \leftarrow \textsc{Concat}(y^i, vec))$
    y $\leftarrow \textsc{Aggregate}(y^1, y^2, \ldots, y^h)$
    $\textsc{Normalize}(y)$ ▷ divide by $\deg(v)$
    $R1_{a_t}(v) \leftarrow y$ ▷ N-FVR
    $R2_{a_t}(v) \leftarrow \textsc{Concat}(R_v^0, y)$ ▷ NN-FVR
return $R1, R2$

Proposed Approach

Neighborhood-based Feature Vector Representation (N-FVR) of nodes



Baselines

- NNS: No Network structure. The feature vector is generated using only the attributes of nodes.
- Line: It defines a loss function based on one-step and two-step relational information between nodes and combines them to get the final feature vector.
- SLR: It is an integrative probabilistic model, which is used to capture the statistical correlations (homophily effect) among attributes. It uses the triangular motif representation of the network for improved scalability and predictive performance.
- Majority: The majority approach takes the most frequently occurring attribute value from the neighboring nodes.
- MNE: It captures multiple structures (facets) of the network by learning multiple embeddings simultaneously. Uses the Hilbert Schmidt Independence Criterion (HSIC) as a diversity constraint.

Baselines

- LMMG: Based upon the idea of Multiplicative Attribute Graph (MAG) Mode. In this approach, each node can belong to multiple groups, and the occurrence of each node feature is determined by a logistic model based on the group memberships of the given node.
- WVRN: It is a weighted relational neighbor classifier that estimates attribute value a_i of a node v using the weighted mean of the same attribute of v's neighbors.
- DeepWalk: It uses random walk method to translate graph structure into linear sequences. The skip-gram model with hierarchical softmax is used as the loss function.
- GraRep: It incorporates both local and global structural information of the graph to learn the feature vector representations.
- Node2Vec: It generalizes the DeepWalk method with the combination of BFS and DFS random walks. This method considers both network structure and graph homophily.

Datasets

Dataset	Name	No. of Nodes	No. of Edges	No. of Attributes	Train (%)
Facebook100	Caltech	769	33,312	4	70
	Haverford	1,446	59,589	4	1,5,9
	Rice	4,088	369,657	4	70
	American	6,387	435,325	4	70
	UChicago	6,591	208,103	4	1,5,9
	Mississippi	10,521	610,911	4	1,5,9
	Temple	13,686	360,795	4	1,5,9
	UNC	18,163	766,800	4	80
Slovak Social Network	Pokec	1,000	6,303	3	70
Bibliography Network	4area	26,144	217,100	4	70

Results and Comparisons

Prone values for the attributes of each dataset (used as attribute weights)

(a) Caltech (b) UNC (c) American (d) Haverford (e) Rice

(f) 4area (g) Temple (h) Pokec (i) Mississippi (j) UChicago

Results and Comparisons

Accuracy Comparison of N-FVR and NN-FVR with WVRN, MAJORITY, and NNS Approaches on American and Rice Datasets

Method	American				Rice			
	Status	Gender	Dormitory	Year	Status	Gender	Dormitory	Year
WVRN	85.49	56.67	67.06	71.71	86.12	54.80	84.46	74.72
MAJORITY	85.43	56.84	67.11	70.99	85.71	55	83.50	73.64
NNS	79.59	59.70	16.93	38.96	70.33	57.22	10.65	29.79
N-FVR KNN	88.56	62.86	70.66	**83.45**	89.07	62.02	**94.29**	**84.91**
N-FVR NB	80.16	62.86	36.79	48.63	76.28	52.70	78.59	50.27
N-FVR DT	88.51	62.86	54.23	81.55	87.85	52.70	92.24	81.71
N-FVR SVM	80.84	62.86	43.44	81.25	84.59	52.70	**94.29**	77.33
NN-FVR KNN	80.01	59.92	21.57	61.80	75.55	57.22	45.70	47.80
NN-FVR NB	76.36	61.23	79.06	50.54	80.16	60.80	36.79	48.87
NN-FVR DT	**91.60**	**64.89**	92.24	82.54	**90.24**	**67.68**	55.14	83.51
NN-FVR SVM	85	52.52	**93.55**	78.88	81.41	62.86	43.85	81.43
Improvement from NNS to N-FVR (%)	8.97	3.16	53.73	44.49	18.74	4.8	83.64	55.12
Improvement from NNS to NN-FVR (%)	10.65	5.19	76.62	43.58	19.91	10.46	44.49	53.72

Results and Comparisons

Accuracy Comparison of N-FVR and NN-FVR with WVRN, MAJORITY, and NNS Approaches on Pokec and 4area Datasets

Method	Pokec			4area			
	Public	Gender	Age	DB	DM	IR	ML
WVRN	46.2	42.2	**25.6**	90.40	88.94	88.97	89.95
MAJORITY	49.1	40.5	25.2	90.17	88.84	88.57	89.68
NNS	52.33	61.33	16.74	97.60	97.50	97.30	97.90
N-FVR KNN	87	**66**	23.78	92.83	92.26	92.01	92.40
N-FVR NB	86.33	60.66	18.50	88.71	87.21	87.65	88.05
N-FVR DT	**87.66**	61	21.58	92.45	92.59	91.48	92.93
N-FVR SVM	81	57	14.09	92.64	92.47	91.80	93.11
NN-FVR KNN	80.66	64.33	25.11	96.30	95.52	95.46	96.18
NN-FVR NB	86.33	60.66	18.94	90.04	89.31	89.02	90.88
NN-FVR DT	**87.66**	65.33	18.50	95.80	95.49	95.06	95.69
NN-FVR SVM	82.66	57	16.74	**97.61**	**97.59**	**97.46**	**97.92**
Improvement from NNS to N-FVR (%)	35.33	4.67	7.04	−4.77	−4.91	−5.29	−4.79
Improvement from NNS to NN-FVR (%)	35.33	4	8.37	0.01	0.09	0.16	0.02

Results and Comparisons

Accuracy Comparison of N-FVR and NN-FVR Using KNN Classifier with DeepWalk, Line, GraRep, Node2Vec, and MNE.

Techniques	UChicago									Temple								
	Gender			Year			Dormitory			Gender			Year			Dormitory		
	1%	5%	9%	1%	5%	9%	1%	5%	9%	1%	5%	9%	1%	5%	9%	1%	5%	9%
DeepWalk	50.1	52.3	55.9	55.6	59.1	63.8	20.2	35.7	47.4	50.1	55.5	58.2	51.1	55.7	60.3	21.4	31.8	36.1
LINE	52.1	54.1	56.9	61	61.9	65.2	21.1	43.5	50.1	52.9	57.9	58.5	56.3	66.9	69.6	25.4	32.7	38.2
GraRep	47.7	48.5	50.1	50.5	55.3	59.9	18.6	30.3	40	45.6	49	55	50.3	57.2	65.1	21.7	29.6	31.5
node2vec	51.3	53.5	55.2	60.2	61.2	64.1	22.1	39.8	49.7	51	54.8	57.9	52.8	55.3	64.2	20.2	29.8	38.1
MNE	54.5	**57.7**	**59.7**	58.1	65.9	67.7	24.8	**48.2**	**54.4**	55.9	**61.4**	**62.9**	61.5	69.9	**72.7**	30.1	36.1	41.9
N-FVR	**55.8**	56.6	56.7	**70.7**	**74.3**	**75**	**25.8**	46.1	54.1	**57.3**	57.5	58.3	**69.4**	**70.3**	70.6	**37.8**	**45.4**	**48.4**
NN-FVR	52.3	52.4	52.4	52.5	71.1	71	8	16.6	22.1	57	57.1	57.1	57.3	67.8	68	21.6	29.2	29.4
Value of k	12	97	3	4	4	17	1	1	1	89	45	24	26	18	25	6	9	10
Improvement from MNE (%)	1.3	−1.1	−3	12.6	8.4	7.3	1	−2.1	−0.3	1.4	−3.9	−4.6	7.9	0.4	−2.1	7.7	9.3	6.5

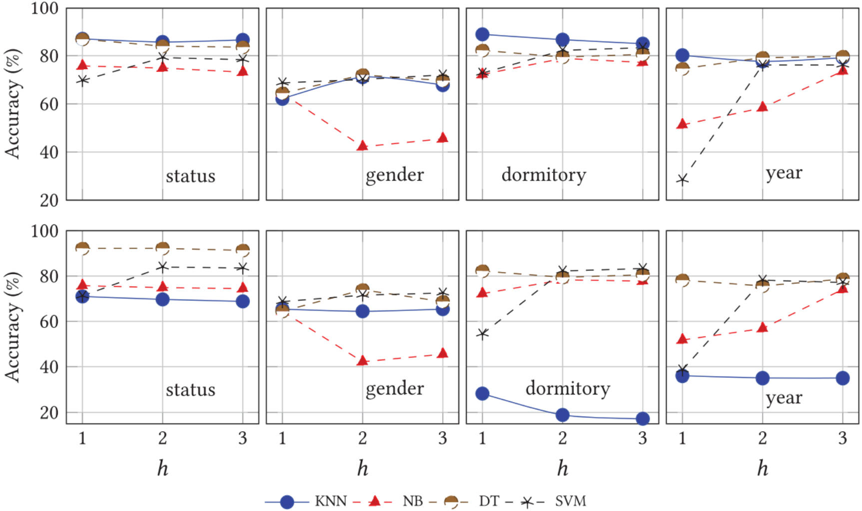
Results and Comparisons

Accuracy Comparison of N-FVR and NN-FVR Using KNN Classifier with DeepWalk, Line, GraRep, Node2Vec, and MNE.

Techniques	Haverford									Mississippi								
	Gender			Year			Dormitory			Gender			Year			Dormitory		
	1%	5%	9%	1%	5%	9%	1%	5%	9%	1%	5%	9%	1%	5%	9%	1%	5%	9%
DeepWalk	50.6	53.5	57.3	61.4	76.7	81.1	29	37.4	43.9	53.1	60.4	60.9	46.5	55.3	61.6	32.5	44.1	48.3
LINE	50.1	51.6	52.9	59.1	76.1	80.5	27.9	36.6	41.5	55.3	62.7	64.7	48.6	58.9	63.2	34.2	48.9	53.4
GraRep	48.8	51.1	51.9	57.4	72.1	77.5	29	39.8	42.9	44.6	48	52.9	42.7	48.3	49.2	32.5	45.9	52.1
node2vec	51.3	57.1	57.1	57.6	75.6	79.1	29.2	41.4	43.8	52.6	59.8	59.8	47.2	56.8	60.1	31.3	39.5	44.1
MNE	54.2	59.6	62.0	66.9	81.3	**84.4**	33	**45.7**	**47.6**	58.9	65.9	68	53.3	59.4	63.8	38.7	53.7	56.7
N-FVR	**63.8**	**64.2**	**63.9**	**78.9**	**81.9**	83.4	**38.3**	41.6	47.5	**63.2**	**67.1**	**68.7**	**68.7**	**68.4**	**68.8**	**43.9**	**56.6**	**60.8**
NN-FVR	54.3	55.9	57.3	34.3	54	63.9	37.1	38.3	39.4	55	58	60.4	56.3	67.9	67.7	15.2	27.9	30
Value of k	7	29	48	1	3	6	10	9	7	1	4	6	12	38	42	4	5	14
Improvement from MNE (%)	9.6	4.6	1.9	12	0.6	−1	5.3	−4.1	−0.1	4.3	1.2	0.7	15.4	9	5	5.2	2.9	4.1

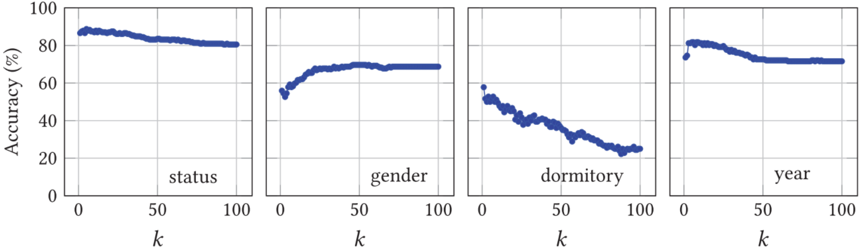
Results and Comparisons

Effect of h on N-FVR (top) and NN-FVR (bottom) methods using different classifiers for different attributes of Caltech dataset.



Results and Comparisons

Effect of k on accuracy using KNN algorithm on different attributes of Caltech dataset utilizing N-FVR



Limitations

- We observe that as the number of unique values in attributes increases, the accuracy of underlying classifiers tends to decrease. This behavior is observed for all methods

Conclusion And Future Work

- We propose a method to generate feature vectors for the nodes based on other attribute values of that node and its neighbors.
- These feature vectors then input to standard machine learning algorithms to predict attributes.
- One possible extension is to use the proposed method to design feature vectors for the nodes or graphs in general that can then be used for node or graph classification.
- Using Deep Learning to design embeddings based on the weighted neighborhood information is another potential future work.

Questions!!