Precept 9: LDA, graph/network properties and analysis

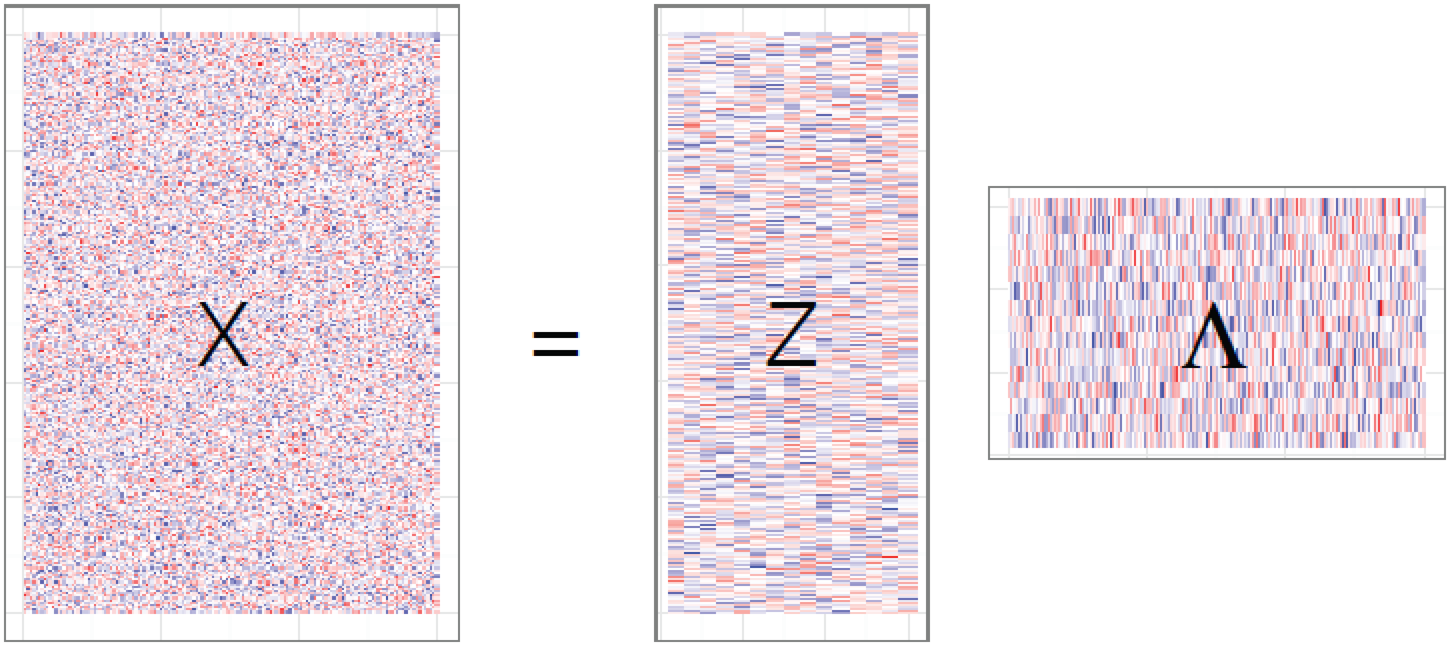
COS424/524/SML302 Spring 2021

Xiaoyan Li

Topics:

- Latent Dirichlet Allocation (LDA)
- Graph Properties and Measurements
- Final Course Project
  - Final project proposal (4/16)
  - Poster session (5/3)
  - Final project report (5/5 by 5pm)

Dimension Reduction—basic idea



Compute a reduced representation Z of data X from p dimensional to k dimensional. X is a nxp matrix and Z is a nxk matrix where p >> k.

Q: What are the two main benefits of representing X with Z?

A: data compression, de-noising

Can reconstruct the p-dimensional data from the k-dimensional data.

Q: why?

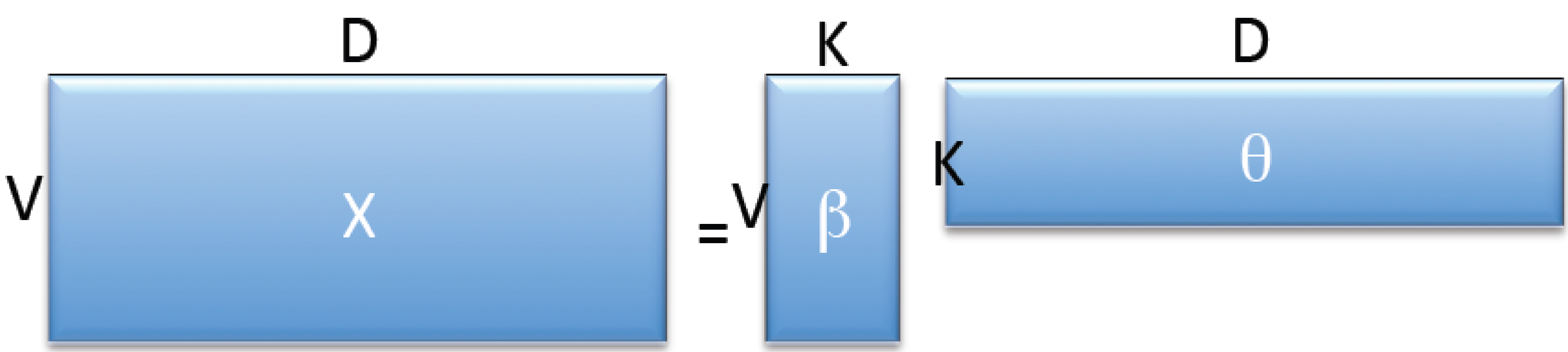
A: can fill in missing values.

LDA as dimension reduction

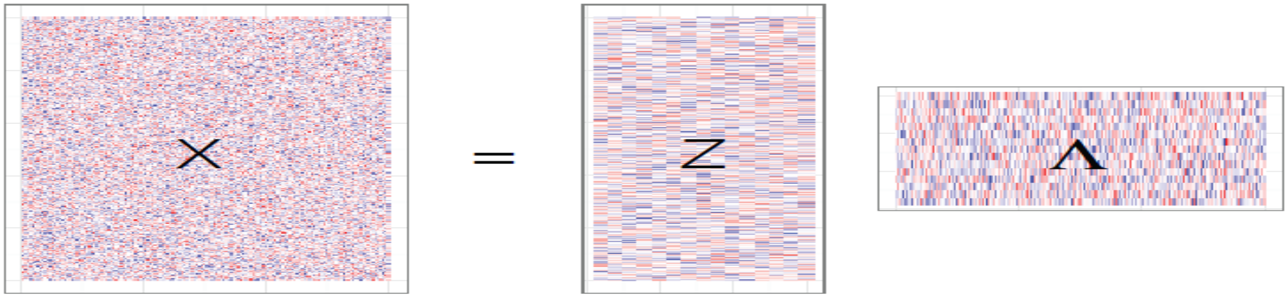
Marginal likelihood of data:

$$x_d \mid \theta, \beta, \alpha, \eta \quad \sim \quad Mult(\theta_d^T \beta)$$

- We can rewrite this likelihood in terms of a matrix factorization.
- Consider variables and parameters as matrices: $X \in \mathcal{Z}^{D \times V}$, $\theta \in (0,1)^{D \times K}$, $\beta \in (0,1)^{K \times V}$



PCA, SVD, NMF and LDA



- **Principal Component Analysis(PCA):**
  - row vectors in Λ: Λ1, Λ2, ... Λk. are orthogonal.
- **Singular Value Decomposition(SVD):**
  - $X \sim X_k = U_k S_k V^T_k$, $(Z = U_k S_k$, $V^T_k = \Lambda)$, both U and V are orthogonal, S is diagonal
- **Non-Negative Matrix Factorization(NMF):**
  - **V = WH** (W = Z , H = Λ); V, M, and H have no negative elements.
- **Latent Dirichlet Allocation(LDA):**
  - **X** = $\boldsymbol{\beta\theta}$, $\boldsymbol{\beta}$: word-to-topic matrix, $\boldsymbol{\theta}$: document-to-topic matrix
  - Each document in $\boldsymbol{\theta}$ is a distribution over topics, each column sums to 1
  - Each topic in $\boldsymbol{\beta}$ is a distribution over words, each column sums to 1.

Q1: How do you determine k=? (# of components, dimension of the reduced space)

Q2: Do they have unique solutions? Will they converge?

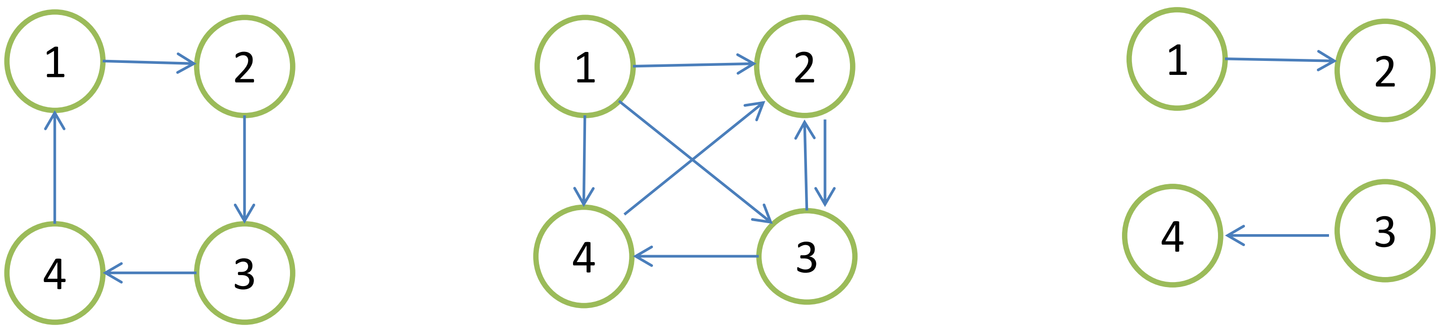
Q3: What do you do if you do not have enough memory to reconstruct the original data matrix?

Graph Properties and Measurements: Connectivity of a graph

- Strongly connected
  - A directed graph is called strongly connected if there is a path in each direction between each pair of vertices of the graph
  - Can generate strongly connected subgraphs if the graph itself Is not strongly connected.
- Weakly connected
  - A directed graph is weakly connected if, and only if, the graph is connected when the direction of the edge between nodes is ignored.

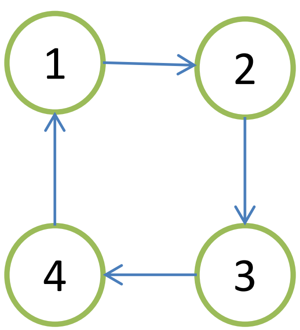
Graph Properties and Measurements: Connectivity of a graph

- Q: Strongly connected, weakly connected, or not connected?

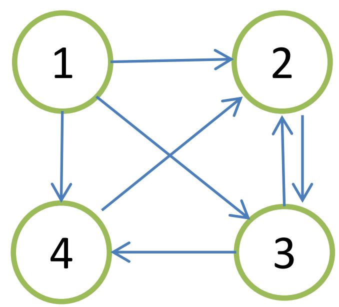


Graph Properties and Measurements: Connectivity of a graph

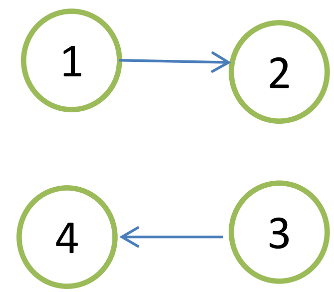
- Q: Strongly connected, weakly connected, or not connected?



Strong connected



Weak connected



Not connected

Graph Properties and Measurements: Centrality of a node

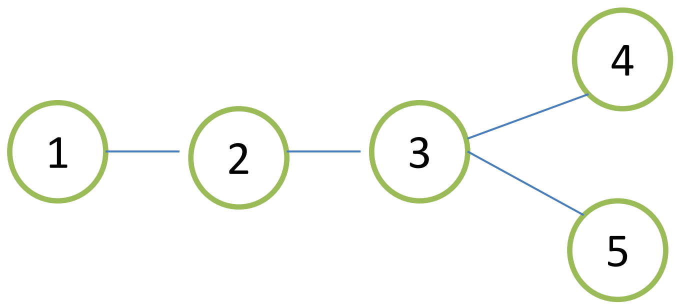
- degree_centrality (v) , in_degree_centrality(v) , out_degree_centrality(v)

- closeness_centrality(v)

- betweenness_centrality(v)

Graph Properties and Measurements: Centrality of a node

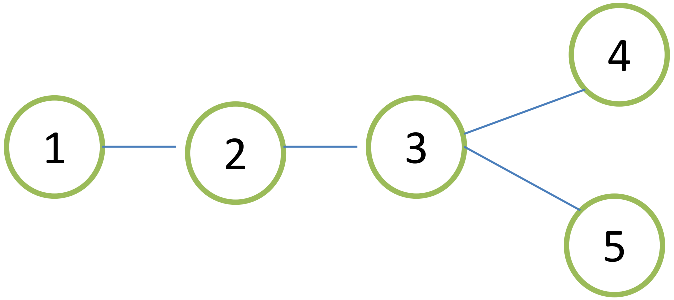
- degree_centrality (v) , in_degree_centrality(v) , out_degree_centrality(v)
  - The degree centrality for a node v is the fraction of nodes it is connected to.
  - The in-degree centrality for a node v is the fraction of nodes its incoming edges are connected to.
  - The out-degree centrality for a node v is the fraction of nodes its outgoing edges are connected to.



Node	Degree_centrality
1	¼
2	½
3	¾
4	¼
5	¼

Graph Properties and Measurements: Centrality of a node

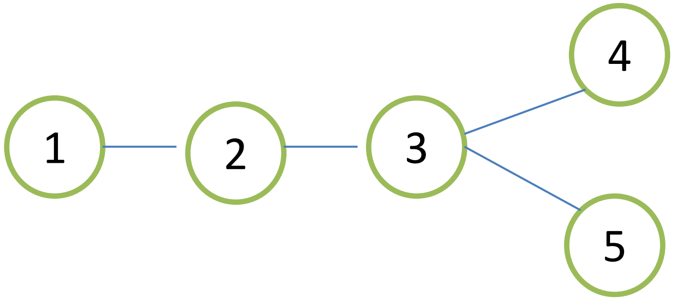
- closeness_centrality(v)
  - the reciprocal of the sum of the shortest path distances from v to all n-1 other nodes.
  - Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum possible distances n-1.
  - 1/(average distance to all other nodes)



Node	closeness_centrality
1	4/9
2	2/3
3	4/5
4	1/2
5	1/2

Graph Properties and Measurements: Centrality of a node

- betweenness_centrality(v)
  - the sum of the fraction of all-pairs shortest paths that pass through v
  - Since it scales with the number of pairs of nodes in the graph, betweenness is normalized with the number of pairs of nodes in the graph.
  - Divided by (n-1)(n-2)/2 for undirected graphs. (6 in the following graph)



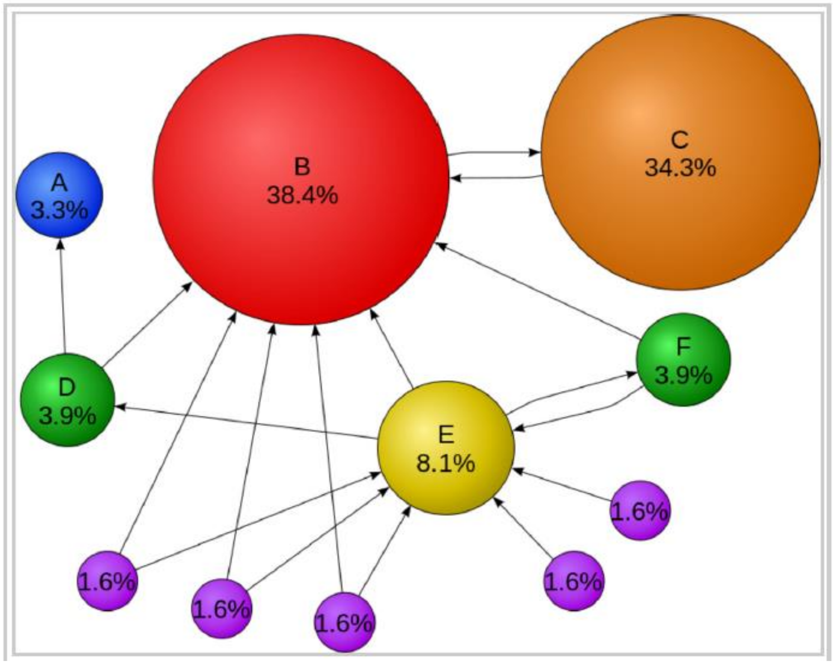
Node	Betweeness_centrality
1	0
2	1/2
3	5/6
4	0
5	0

Graph Properties and Measurements: Centrality of a node

- degree_centrality (v) , in_degree_centrality(v) , out_degree_centrality(v)
  - The degree centrality for a node v is the fraction of nodes it is connected to.
  - The in-degree centrality for a node v is the fraction of nodes its incoming edges are connected to.
  - The out-degree centrality for a node v is the fraction of nodes its outgoing edges are connected to.
- closeness_centrality(v)
  - the reciprocal of the sum of the shortest path distances from v to all n-1 other nodes.
  - Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum possible distances n-1.
- betweenness_centrality(v)
  - the sum of the fraction of all-pairs shortest paths that pass through v:

Graph Properties and Measurements: Centrality of a node

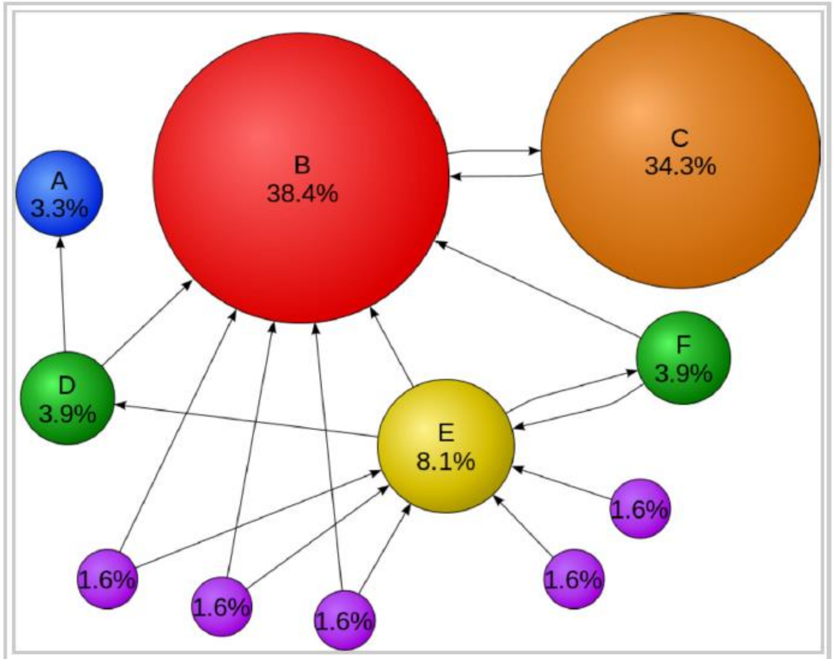
- PageRank:
  - PageRank of a node/webpage is the probability of that webpage being visited on a particular random walk.
  - Q: What kinds of nodes are likely to have a higher PageRank



PageRanks for a simple network, expressed as percentages.
https://en.wikipedia.org/wiki/PageRank

Graph Properties and Measurements: Centrality of a node

- PageRank:
  - PageRank of a node/webpage is the probability of that webpage being visited on a particular random walk.
  - A node pointed by many nodes is likely to have a higher PageRank. E.g. Node B, and node E
  - A node pointed by a node with a higher PageRank is likely to have a higher PageRank.



PageRanks for a simple network, expressed as percentages. https://en.wikipedia.org/wiki/PageRank

Graph Properties and Measurements: Centrality of a node

- PageRank
- Hubs and Authorities
  - also called Hyperlink-Induced Topic Search(HITS) algorithm, a link analysis algorithm that ranks web pages.
  - a hub is a node pointing to many other nodes (high out-degree)
  - An authority node is pointed to by many hubs.

Graph Properties and Measurements: proximity measures of node pairs

- Common neighbors: the number of neighbors shared by node u and v.
  - Score(u,v) = $|N(u) \cap N(v)|$, where N(u) and N(v) are the number of neighbors for u and v respectively.
  - Used in social network analysis(SNA) for link prediction
    - Given a snapshot of a social network, predict future interactions between members.
  - Assumption: if two people have more common friends, then the probability of future interaction between them is higher.

Graph Properties and Measurements: proximity measures of node pairs

- Jaccard coefficient: percentage of neighbors shared by two nodes.
  - Score(u,v) = $\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$
- adamic_adar_index: frequency-weighted common neighbors, more weights for nodes with less neighbors
  - Score(u,v) = $\sum_{z \in N(u) \cap N(v)} \frac{1}{\log |N(z)|}$
- Preferential attachment score: product of the number of neighbors of two nodes
  - Score(u,v) = |N(u)|*|N(v)|
- Other proximity measures …

Apply graph-based methods

- 1. Generate more graph-based features for classification
- 2. Use proximity measures of node pairs directly to predict their future interactions
  - Common neighbors, Jaccard coefficient, adamic_adar_index, Preferential attachment score, etc.
- 3. Identify some structures in the graph
  - Find strongly connected components, and identify nodes that are in the same strongly connected subgraphs
  - Identify important nodes, or nodes with some strange behaviors, etc.
- 4. Come up with your own ideas …
- Q: Can I generate a graph with the HW3 data?
  - Nodes? Edges?

Final Project:

- Group project,
  - Work with 1-3 partners
- One-page proposal
  - Due 4/16, not graded
  - Feedback returned in about a week.
- Poster session
  - May 3rd. Online (one morning session and two afternoon sessions.)
- Final project report
  - 8-page
  - May 5th by 5pm

Final Project:

- Datasets
  - Existing, Cleaned, Processed…
  - Okay to use the data for HW1, HW2, or HW3
  - Describe size of the data, features. etc.
- Questions/tasks
  - Answer questions, answer them with the data
- Methods
  - Classifiers
  - Regressors
  - Unsupervised learning
- Evaluations
  - Accuracy/error, confusion matrix, ROC, precision, recall, F1 score,
  - R Squared, RMSE, MAE, …
  - Intra-cluster distance, inter-cluster distance
  - Visualization of clusters, label the topics/clusters

Final Project:

- Cross-validation
  - Quantify generalization error,
  - hyperparameter tuning
- Bootstrapping
  - Confidence intervals
- Feature selection
  - Why and how
- Dimension reduction
  - PCA, truncated SVD, NMF, LDA…
  - How to decide the number of components/dimensions?
- Graph analysis
  - What are the nodes? Edges? directed or undirected?
- …

Paper:

- Group discussion for COS524 in breakout rooms (~15 minutes)

- Share your opinions in the shared google doc:
  - Will send you the links in chat
  - You can focus on some of the questions

- Come back for precept wrap up

Wrap up for Precept 9:

- Latent Dirichlet Allocation (LDA)
- Graph Properties and Measurements
- Final Course Project
  - Final project proposal (4/16)
  - Poster session (5/3)
  - Final project report (5/5 by 5pm)

Some Packages in Python

- NetworkX
  - Creation and manipulation of graphs and network analysis, pure-python implementation.
  - https://networkx.github.io/documentation/networkx-1.10/reference/algorithms.html
- Graph-tool
  - implemented in C, https://pypi.python.org/pypi/graph-tool
- Igraph
  - implemented in C, https://igraph.org/python
- sklearn.decomposition.LatentDirichletAllocation
  - Latent Dirichlet Allocation with online variational Bayes algorithm
  - http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

Resources: (Some materials are taken from the following resources and lecture slides for cos424.)

- Latent Dirichlet Allocation from wikipedia
  - https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- Link Analysis
  - https://www.csc2.ncsu.edu/faculty/nfsamato/practical-graph-mining-with-R/sample/chapter_5_LinkAnalysis.pdf
- Graph-tool performance comparison
  - https://graph-tool.skewed.de/performance
- Graph Properties & Measurements
  - https://reference.wolfram.com/language/guide/GraphPropertiesAndMeasurements.html