1. Subgraph Analysis

2. Propagation Methods

3. Latent Factor Models
   a) Background
   b) Normal Behavior
   c) Abnormal Behavior
Matrix Modeling
Matrix Modeling

Matrix $M$

User

Page

\[
\begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix}
\]
Matrix Modeling

**HITS**

Authoritativeness $\vec{v}$ is first eigenvector of $M^TM$

$$\vec{v} = cM^TM\vec{v}$$

Hubness $\vec{u}$ is first eigenvector of $MM^T$

$$\vec{u} = cMM^T\vec{u}$$
Matrix Modeling

HITS
Authoritativness $\vec{v}$ is first eigenvector of $M^T M$
$$\vec{v} = c M^T M \vec{v}$$

Hubness $\vec{u}$ is first eigenvector of $MM^T$
$$\vec{u} = c M M^T \vec{u}$$

What about the other eigenvectors?
Matrix Modeling
Singular Value Decomposition

\[ U \Sigma V^T \approx M \]
Matrix Modeling
Singular Value Decomposition

\[ U \Sigma V^T \approx M \]

- Hubness \( \vec{u} \)
- Authoritativeness \( \vec{v} \)
Matrix Modeling
Singular Value Decomposition

\[ U \Sigma V^T \approx M \]

Hubness \( \vec{u} \)

Authoritativeness \( \vec{v} \)

\( \Sigma \) contains normalization for \( \vec{u} \) and \( \vec{v} \)
Matrix Factorization

What does each eigenvector capture?

Each factor captures a dense block in the matrix

$$UV^T \approx M$$
Matrix Factorization

What does each eigenvector capture?

Each factor captures a dense block in the matrix

\[ UV^T \approx M \]
Matrix Factorization

\[ UV^T \approx M \]

\[ u_i \cdot v_j \approx M_{i,j} \]
Matrix Factorization

\[ UV^T \approx M \]

\[ u_i \cdot v_j \approx M_{i,j} \]
Matrix Factorization

\[ U V^T \approx M \]
\[ u_i \cdot v_j \approx M_{i,j} \]
1. Subgraph Analysis

2. Propagation Methods

3. Latent Factor Models
   a) Background
   b) Normal Behavior
   c) Abnormal Behavior
Matrix Completion

Recommendation systems predict missing entries
Matrix Completion

Can’t find singular vectors with missing entries. Instead,

$$\min_{U,V} \sum_{(i,j) \in M} (M_{i,j} - \vec{u}_i \cdot \vec{v}_j)^2$$
Matrix Completion

Can’t find singular vectors with missing entries. Instead,

$$\min_{U,V} \sum_{(i,j) \in M} (M_{i,j} - \tilde{u}_i \cdot \tilde{v}_j)^2$$
Matrix Completion

Can’t find singular vectors with missing entries. Instead,

\[
\min_{U,V} \sum_{(i,j) \in M} (M_{i,j} - \hat{u}_i \cdot \hat{v}_j)^2
\]

Genres

\[
\begin{array}{c|c|c|c|c}
1.2 & -.1 & .5 & .8 & -.5 \\
\end{array}
\]

≈ 1
Matrix Completion

Can’t find singular vectors with missing entries. Instead,

\[
\min_{U, V} \sum_{(i, j) \in M} (M_{i, j} - \hat{M}_{i, j})^2
\]

\[
\hat{M}_{i, j} = \hat{u}_i \cdot \hat{v}_j
\]
Adding Latent Factors

Consider additional factors:
- Dataset mean $\mu$
- Row (user) baseline $b_i$
- Column (movie) baseline $b_j$

$$\hat{M}_{i,j} = \mu + b_i + b_j + \vec{u}_i \cdot \vec{v}_j$$

$$\min_{U,V} \sum_{(i,j) \in M} (M_{i,j} - \hat{M}_{i,j})^2$$
Adding Latent Factors

What if we know the \textit{time} of the rating (time of the edge being created)?
Adding Latent Factors

Mean Rating by Date (Netflix)

Collaborative Filtering with Temporal Dynamics
Yehuda Koren
KDD 2009
Adding Latent Factors

Mean Rating by Movie Age (Netflix)

Collaborative Filtering with Temporal Dynamics
Yehuda Koren
KDD 2009
Collaborative Filtering with Temporal Dynamics

Yehuda Koren

KDD 2009

Adding Latent Factors

\[
\min_{U, V} \sum_{(i,j) \in M} (M_{i,j} - \hat{M}_{i,j})^2
\]

Time factors:
- Column (movie)- time baseline \(b_{j, \text{Bin}(t)}\)
- Row (user)-time baseline function \(b_i(t)\)

\[
\hat{M}_{i,j} = \mu + b_i + b_j + \vec{u}_i \cdot \vec{v}_j + b_{j, \text{Bin}(t)} + b_i(t)
\]
Bayesian Probabilistic Matrix Factorization

Ruslan Salakhutdinov and Andriy Mnih
ICML 2008
Bayesian Modeling

Sample user factors from Normal distribution

Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008
Bayesian Modeling

Sample user factors from Normal distribution

Update mean based on user factors
Bayesian Modeling

Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008
Bayesian Modeling

\[ p(M_{i,j} | U, V) = \mathcal{N}(M_{i,j} | \bar{u}_i \cdot \bar{v}_j, \sigma^2) \]
Bayesian Modeling

![Graph showing RMSE vs. Epochs for different models: SVD, PMF, Logistic PMF, and Bayesian PMF.](image)

Left panel: Performance of SVD, PMF, logistic PMF, and Bayesian PMF using 30D feature vectors. The y-axis displays RMSE (root mean squared error), and the x-axis shows the number of epochs. Right panel: RMS error for the models with 30D and 60D feature vectors. Note that this model can be seen as a PMF model without regularization applied to the feature vectors.
Bayesian Modeling with Co-Clustering

Cluster users with similar factors
CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014
Bayesian Modeling with Co-Clustering

CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray, Christos Faloutsos Alex Smola
WWW 2014
Online Rating Models

Typically fit a Gaussian - Minimize RMSE

Data Normal CF

CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray, Christos Faloutsos Alex Smola
WWW 2014
Online Rating Models

Typically fit a Gaussian - Minimize RMSE

Data  Normal CF  CoBaFi

CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray, Christos Faloutsos Alex Smola
WWW 2014
### Shape of Netflix reviews

<table>
<thead>
<tr>
<th>Most Gaussian</th>
<th>Most skewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Rookie</td>
<td>The O.C. Season 2</td>
</tr>
<tr>
<td>The Fan</td>
<td>Samurai X: Trust and Betrayal</td>
</tr>
<tr>
<td>Cadet Kelly</td>
<td>Aqua Teen Hunger Force: Vol. 2</td>
</tr>
<tr>
<td>Money Train</td>
<td>Sealab 2001: Season 1</td>
</tr>
<tr>
<td>Alice Doesn’t Live Here</td>
<td>Aqua Teen Hunger Force: Vol. 2</td>
</tr>
<tr>
<td>Sea of Love</td>
<td>Gilmore Girls: Season 3</td>
</tr>
<tr>
<td>Boiling Point</td>
<td>Felicity: Season 4</td>
</tr>
</tbody>
</table>

- **More Gaussian**: Movies
- **More Skewed**: TV Shows

---

**CoBaFi**: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray, Christos Faloutsos Alex Smola
WWW 2014
What is a tensor?

- Tensors are used for structured data > 2 dimensions
- Think of as a 3D-matrix

For example:

Kanye West rated The Sound of Music five stars last January.
Tensor Decomposition

Kanye West rated The Sound of Music five stars last January.

\[ X \approx U \otimes V \otimes W \]

\[ X_{i,j,k} \approx \sum_{r=1}^{\text{Rank}} U_{i,r} V_{j,r} W_{k,r} \]
Graph Clustering with Tensors

Multiple possible views of the DBLP network:
1. Who-cites-whom
2. Co-authorship
3. Using same words in title

Do more Views of a Graph help? Community Detection and Clustering in Multi-Graphs
Evangelos E. Papalexakis, Leman Akoglu, Dino Ienco
FUSION 2013
Graph Clustering with Tensors

Do more Views of a Graph help? Community Detection and Clustering in Multi-Graphs
Evangelos E. Papalexakis, Leman Akoglu, Dino Ienco
FUSION 2013
Graph Clustering with Tensors

DBLP-1

(a) citation

(b) co-auth.

(c) co-term

DBLP-2

(a) citation

(b) co-auth.

(c) co-term

Do more Views of a Graph help? Community Detection and Clustering in Multi-Graphs

Evangelos E. Papalexakis, Leman Akoglu, Dino Ienco

FUSION 2013
Graph Clustering with Tensors

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>GraphFuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP-1</td>
<td>0.12</td>
<td>0.30</td>
</tr>
<tr>
<td>DBLP-2</td>
<td>0.08</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Modeling Accuracy
Coupled Matrix + Tensor Decomposition
Coupled Matrix + Tensor Decomposition

\[ X \approx U \circledast V \circledast W \]
\[ Y \approx UA^T \]

\[
\min_{U,V,W,A} \|X - U \circledast V \circledast W\|_F^2 + \|Y - UV^T\|_F^2
\]
Joint Factorization

Model Parameters, $\Phi$

- $k$-dimensional entity vectors

Partial Observations

- Predict missing data

Key Components:
- **Entities, $\mathcal{E}$**
  - Categories, $S_C$
  - Attributes, $S_A$
  - Businesses, $S_B$
  - Users, $S_U$
  - Review words, $S_W$

- **Relations, $\mathcal{R}$**
  - Business Categories, $C$
  - Business Attributes, $A$
  - User/Business Ratings, $R$
  - Reviews for Business, $BW$
  - Reviews by Users, $UW$

Collective Factorization for Relational Data: An Evaluation on the Yelp Datasets

Nitish Gupta, Sameer Singh
The most important problem in database completion is to be able to predict unobserved relations for entities that already exist in the database. For example, predicting ratings for businesses to which explicit relations are not observed.

To show how our model improves significantly on previous versions, Figure 2 presents the PR Curve (Ratings) for predicting held-out ratings. From this, it is clear that the user reviews are quite valuable, with an increase of 15% recall when incorporating information about the businesses in terms of their attributes, with an increase of 3% or review words used for them.

Table 3: Performance of different models by varying the business and user relations available during training for rating and attribute predictions. Results for collective factorization of reviews, categories, and attributes should improve predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.70</td>
</tr>
<tr>
<td>A+R</td>
<td>0.80</td>
</tr>
<tr>
<td>R+UW</td>
<td>0.90</td>
</tr>
<tr>
<td>R+BW</td>
<td>0.75</td>
</tr>
<tr>
<td>R+C</td>
<td>0.85</td>
</tr>
<tr>
<td>A+R+UW</td>
<td>0.95</td>
</tr>
<tr>
<td>R+C+UW</td>
<td>0.95</td>
</tr>
<tr>
<td>A+C+R+UW</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Most valuable:
1. Ratings
2. Review text
3. Business Categories

Collective Factorization for Relational Data: An Evaluation on the Yelp Datasets

Nitish Gupta, Sameer Singh
1. Subgraph Analysis

2. Propagation Methods

3. Latent Factor Models
   a) Background
   b) Normal Behavior
   c) Abnormal Behavior
Fraud Detection
Fraud Detection
Fraud within a factorization
Fraud within a factorization

Followees

Followers

U

V

X
Fraud within a factorization

Followees

Followers

1.5  1  -0.5  -2  1

V

U

X
Fraud within a factorization

EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs
B. Aditya Prakash, Ashwin Sridharan, Mukund Seshadri, Sridhar Machiraju, Christos Faloutsos
PAKDD, 2010
Fraud within a factorization

EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs
B. Aditya Prakash, Ashwin Sridharan, Mukund Seshadri, Sridhar Machiraju, Christos Faloutsos
PAKDD, 2010
Fraud within a factorization

Inferring Strange Behavior from Connectivity Pattern in Social Networks
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang.
PAKDD, 2014
Fraud within a factorization

Inferring Strange Behavior from Connectivity Pattern in Social Networks
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang.
PAKDD, 2014
Fraud within a factorization

Username: a####
Birthday: January 1st

Inferring Strange Behavior from Connectivity Pattern in Social Networks
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang.
PAKDD, 2014
Complementary Fraud Detection

Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher, Christos Faloutsos
ICDM, 2014.
Complementary Fraud Detection

960 fraudsters safely following 960 customers

Singular Value (Attack Size)

Number of components (k)

Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher,
Christos Faloutsos
ICDM, 2014.
Complementary Fraud Detection

Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher, Christos Faloutsos
ICDM, 2014.
Complementary Fraud Detection

Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher,
Christos Faloutsos
ICDM, 2014.
Complementary Fraud Detection

Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher, Christos Faloutsos
ICDM, 2014.
Complementary Fraud Detection

93% Precision
70% of accounts missed by Twitter

Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher, Christos Faloutsos
ICDM, 2014.
### Practitioner’s Guide

<table>
<thead>
<tr>
<th>Method</th>
<th>Graph Type</th>
<th>Node Attributes</th>
<th>Edge Attributes</th>
<th>Seed Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>EigenSpokes</td>
<td>Directed+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get-the-Scoop</td>
<td>Directed+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fBox</td>
<td>Directed+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoBaFi</td>
<td>Bipartite+</td>
<td></td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td>CDOoutliers</td>
<td>Undirected</td>
<td>✔️</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Detecting Fraud within Recommendation
Detecting Fraud within Recommendation

CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray, Christos Faloutsos Alex Smola
WWW 2014
Clustering Fraudsters

$\mu_1$  $\mu_2$  $\mu_3$  $\mu_4$  $\mu_5$

Naïve Spammers  Spam + Noise  Hijacked Accounts

CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray, Christos Faloutsos Alex Smola
WWW 2014
Clustered Fraudsters

Clustered naïve spammers
Clustered hijacked accounts
Clustered “attacked” movies

83% are clustered together

CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014
Outliers in Joint Factorization

\[ V_1 \]

\[ V_2 \]

\[ U_1 \]

\[ U_2 \]

\[ X_1 \]

\[ X_2 \]

Enforce \( U_1 \approx U_2 \) and \( U_1, U_2, V_1, V_2 \geq 0 \)
Outliers in Joint Factorization

Interesting design of $X_1$ and $X_2$; see paper for details

Enforce $U_1 \approx U_2$ and $U_1, U_2, V_1, V_2 \geq 0$
Outliers in Joint Factorization

Rows of $V_2$ represent common patterns in $X_2$ (cluster centroids)

Enforce $U_1 \approx U_2$ and $U_1, U_2, V_1, V_2 \geq 0$
An anomaly is a row of \( X_i \) that is \textit{not} similar to any row in \( V_i \).

**Rows of \( V_2 \) represent common patterns in \( X_2 \) (cluster centroids)**
## Practitioner’s Guide

<table>
<thead>
<tr>
<th>Method</th>
<th>Graph Type</th>
<th>Node Attributes</th>
<th>Edge Attributes</th>
<th>Seed Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>EigenSpokes</td>
<td>Directed+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get-the-Scoop</td>
<td>Directed+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fBox</td>
<td>Directed+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoBaFi</td>
<td>Bipartite+</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>CDOOutliers</td>
<td>Undirected</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
</tr>
</tbody>
</table>
Recap

• SVD captures communities of interest
• Bayesian methods can:
  • Handle missing values
  • Give factorization models (-> patterns, & anomalies)
• Group-outliers: spotted by CoBaFi, Get-the-Scoop, etc.
CONCLUSION
Open Problems / Opportunities

P1. Complex data: How should we integrate data from multiple data sources?

- Undirected
- Directed
- Bipartite
- Node Attributes
- Edge Attributes
- Unsupervised
  - Semi-Supervised
Open Problems / Opportunities

P2. Adversarial analysis: Can we offer provable guarantees on detecting fraud and spam?
Open Problems / Opportunities

P3. Early detection: Can we detect fraudsters before they cause significant damage?
Summary

Local Subgraph Analysis: Patterns and Features
e.g. using ego-nets

http://www.sizemore.co.uk/2005/08/i-feel-some-movies-coming-on.html

\[ 1.1094x + (-0.21414) = y \]
\[ 1.1054x + (-0.21432) = y \]
\[ 2.1054x + (-0.51535) = y \]
Summary

Propagation Methods
“Guilt-by-association”
“Importance-by-association” = PageRank

Given a graph and labels for some nodes, can we learn the labels for the other nodes? Generally, learn labels $X$ so neighbors have the same label.
Summary

Latent Factor Models

Find multiple communities, patterns and anomalies.
Take Away

User Modeling and Fraud Detection are two sides of the same coin.
Thanks again to

NSF Grant No. IIS-1408924, IIS-1408287, CAREER 1452425, DGE-1252522, ...
Questions?

References and resources available at cs.cmu.edu/~abeutel/ccs2015_tutorial