Joint Non-negative Matrix Factorization for Learning Ideological Leaning on Twitter

Preethi Lahoti\textsuperscript{1}, Kiran Garimella\textsuperscript{2}, Aristides Gionis\textsuperscript{2}

\textsuperscript{1}Max Planck Institute for Informatics, \textsuperscript{2}Aalto University
Access to Diverse Information Around World ...
Filtered and Cherry Picked Content ...
Twitter

(Retweet or Follow other users)

User – User social graph
Twitter

User – User social graph

(Retweet or Follow other users)

User - Content content graph

(Retweet a news article)
User and Content Ideology

Liberal

0.5

Conservative

Ideology Scale
User and Content Ideology

Liberal

Clinton puts Trump on defense at first debate

0

Ideology Scale

0.5

DONALD TRUMP WINS SECOND DEBATE; CNN SAYS IT DOESN’T MATTER

Conservative

1
User and Content Ideology

Clinton puts Trump on defense at first debate

DONALD TRUMP WINS SECOND DEBATE; CNN SAYS IT DOESN’T MATTER

Liberal

Ideology Scale

Conservative
Filter bubble...

User A

User B

Topic: Presidential Debate
Filter bubble ...

Clinton puts Trump on defense at first debate

User A

User B

Topic: Presidential Debate
Filter bubble...

Topic: Presidential Debate
Motivation
Ideological segregation, polarization, biased views
Bursting the filter bubble...

(1) Recommend content From other side

Liberal side

Conservative side
Bursting the filter bubble...

(1) Recommend content from other side

(2) Bridge the Gap
Bursting the filter bubble...

(1) Recommend content from other side

(2) Bridge the Gap

Ideology stance is not enough
Bursting the filter bubble...

(1) Recommend content from other side

(2) Bridge the Gap

Ideology stance is not enough

Influence (popularity) of users and content matters
Problem Statement

- Input:

![Social graph (A)](image1)

![Content graph (C)](image2)
Problem Statement

- Input:

- Learn the **shared latent space** between A and C
Problem Statement

- Input:

- Learn the **shared latent space** between A and C
- Discover **ideology-popularity** latent dimensions
Problem Statement

- Input:
  - Learn the **shared latent space** between A and C
  - Discover **ideology-popularity** latent dimensions
  - Estimate **ideology and popularity scores** for users and content
Proposed Methodology
Orthogonal Non-negative Matrix Factorization as Co-Clustering Model [Ding et al]

\[ X \approx [U][H][V^T] \]

- Input matrix (e.g., user – item)
- Clustering of rows of X
- Association between row and column clusters
- Clustering of columns of X
Combining Link and Content

\[ J = \| A - U H_u U^T \|_F^2 + \| C - U H_s V^T \|_F^2 \]

- **user-user matrix**
- **user-content matrix**

**Shared Latent Space:**
- A and C are related via users
- row datatype of matrix A is the same as rows of C
Combining Link and Content

\[ J = \| A - UH_u U^T \|_F^2 + \| C - UH_s V^T \|_F^2 \]  

Joint Matrix factorization

Shared Latent Space:
- A and C are related via users
- row datatype of matrix A is the same as rows of C
Learning Shared Latent Space

\[ J = \| A - UH_uU^T \|^2_F + \| C - UH_sV^T \|^2_F \]

- Joint Matrix factorization
- User-user matrix
- User-content matrix
- Shared latent factors (U and V)

Shared Latent Space:
- A and C are related via users
- Row datatype of matrix A is the same as rows of C
Learning Hidden Manifolds in The Data

\[ J = \|A - UH_u U^T\|_F^2 + \|C - UH_s V^T\|_F^2 + \lambda Tr(U^T L_u U) + \lambda Tr(V^T L_s V) \]

User Manifold and content Manifold are tied together

- Users connected in social graph tend to be ideologically similar
- Ideologically similar users share similar content (and ideologically similar content is shared by similar users)
Latent factors have a **probabilistic interpretation**

\[
J = \| A - UH_u U^T \|_F^2 + \| C - UH_s V^T \|_F^2
\]

Latent factors \( U \) and \( V \)
Latent factors have a probabilistic interpretation

\[ J = \| A - U H_u U^T \|_F^2 + \| C - U H_s V^T \|_F^2 \]

Latent factors U and V

degree to which user i belongs to ideology j

\[ U_{ij} \]
Latent factors have a probabilistic interpretation

\[ J = \| A - U H_u U^T \|_F^2 + \| C - U H_s V^T \|_F^2 \]

Latent factors U and V

degree to which user i belongs to ideology j

degree to which content i belongs to ideology j
Latent factors have a **probabilistic interpretation**

Each row of $U$ and $V$ can be represented as a two dimensional vector $\langle x, y \rangle$
Estimating Ideological Leaning

Each row of $U$ and $V$ can be represented as a vector $\langle x, y \rangle$ in the two dimensional space.
Evaluation
Dataset

Twitter Streaming API (2011 - 2016)
- 7000 users
- 19 million tweets

Three controversial topics
- Gun control
- Abortion
- Obamacare

Ground Truth
- 500 news media channels
- Bayesian point estimate using large annotated data set [Barbera et al]
Baselines

- **Link (user-user)**
  - Graph partitioning (Retweet/Follows)
  - NMF [Lee & Seung]

- **Content (user-content)**
  - ONMTF [Ding et al]
  - DMCC [Gu et al]

- **Combined (link+ content)**
  - IFD / IFD-NGR [Proposed method]
  - BIAS WATCH [Lu et al]
  - KULSHRESTHA [Kulshrestha et al]
Baselines

- **Link (user-user)**
  - Graph partitioning (Retweet/Follows)
  - NMF [Lee & Seung]

- **Content (user-content)**
  - ONMTF [Ding et al]
  - DMCC [Gu et al]

- **Combined (link+ content)**
  - IFD / IFD-NGR [Proposed method]
  - BIAS WATCH [Lu et al]
  - KULSHRESTHA [Kulshrestha et al]

No ideology scores
Baselines

- **Link (user-user)**
  - Graph partitioning (Retweet/Follows)
  - NMF [Lee & Seung]

- **Content (user-content)**
  - ONMTF [Ding et al]
  - DMCC [Gu et al]

- **Combined (link+ content)**
  - IFD / IFD-NGR [Proposed method]
  - BIAS WATCH [Lu et al]
  - KULSHRESTHA [Kulshrestha et al]

Matrix Factorization based approaches

No ideology scores
Baselines

- **Link (user-user)**
  - Graph partitioning (Retweet/Follows)
  - NMF [Lee & Seung]

- **Content (user-content)**
  - ONMTF [Ding et al]
  - DMCC [Gu et al]

- **Combined (link+ content)**
  - IFD / IFD-NGR [Proposed method]
  - BIAS WATCH [Lu et al]
  - KULSHRESTHA [Kulshrestha et al]

Matrix Factorization based approaches
No ideology scores
No ideology scores for media channels
| Twitter Users (Ideology) | News Media (Ideology) |

### Purity

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIASWATCH</td>
<td></td>
</tr>
<tr>
<td>KULSHRESTHA</td>
<td></td>
</tr>
<tr>
<td>IFD</td>
<td></td>
</tr>
<tr>
<td>DMCC</td>
<td></td>
</tr>
<tr>
<td>ONMTF</td>
<td></td>
</tr>
<tr>
<td>GRAPH - FOLLOW</td>
<td></td>
</tr>
<tr>
<td>GRAPH - RETWEET</td>
<td></td>
</tr>
<tr>
<td>NMF</td>
<td></td>
</tr>
</tbody>
</table>

### Pearson Correlation Coefficient

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIASWATCH</td>
<td></td>
</tr>
<tr>
<td>KULSHRESTHA</td>
<td></td>
</tr>
<tr>
<td>IFD</td>
<td></td>
</tr>
<tr>
<td>DMCC</td>
<td></td>
</tr>
<tr>
<td>ONMTF</td>
<td></td>
</tr>
<tr>
<td>GRAPH - FOLLOW</td>
<td></td>
</tr>
<tr>
<td>GRAPH - RETWEET</td>
<td></td>
</tr>
<tr>
<td>NMF</td>
<td></td>
</tr>
</tbody>
</table>
Estimated Ideology scores of high quality across the ideology spectrum (including center)
Ideological Latent Space

Motivation
(Revisiting)
We now have access to ideological position of all the users and content
(1) Bridge the Gap

(2) User guided recommendations
Summary

IFD framework
- Combines Link and Content Graph
- Jointly Matrix Factorization
- Shared Multidimensional Latent Space

Compared to Baselines IFD Estimates
- Ideology and Popularity Scores
- Twitter Users and Media Channels
- High Quality
THANK YOU
Learning Hidden Manifolds in The Data

Manifold assumption

If two data points $x_i, x_j$ are close to each other in the input space then their projections in the new basis $u_i, u_j$ are also close.

\[
J = \|A - UH_u U^T\|_F^2 + \|C - UH_s V^T\|_F^2 + \lambda Tr(U^T L_u U) + \lambda Tr(V^T L_s V)
\]

Graph regularization constraints [Cai et al]

where

- $L_u, L_s$ are graph laplacians of the row and column affinity matrices of $X$
- $Tr(\cdot)$ is trace of the matrix
Optimization Problem

IFD (Ideology Factor Decomposition)

\[ J = \|A - UH_u U^T\|_F^2 + \|C - UH_s V^T\|_F^2 + \lambda \text{Tr}(U^T L_u U) + \lambda \text{Tr}(V^T L_s V) \]

Constraints
- Bi-orthogonality \((U^T U = I ; V^T V = I)\)
- Non-negativity \((U_+ ; H_{u_+} ; H_{s_+} ; V_+)\)

Solution
- Derive multiplicative update rules for \(U, V, H_u\) and \(H_s\)
- Iterative update algorithm
- Locally optimal solution
Estimating Ideological Leaning

The latent factors $U$ and $V$ have a probabilistic interpretation:

- $U_{ij}$: degree to which user $i$ belongs to ideology $j$
- $V_{ij}$: degree to which content $i$ belongs to ideology $j$

Each row of $U$ and $V$ can be represented as a two dimensional vector $\langle x, y \rangle$

Ideology: $i(x, y) = \frac{\theta}{\pi/2} = \frac{\arctan\left(\frac{y}{x}\right)}{\pi/2}$

Popularity: $\rho(x, y) = \sqrt{x^2 + y^2}$
Estimating **Ideological Leaning**

![Diagram showing the relationship between latent dimensions and ideological leaning.](image)