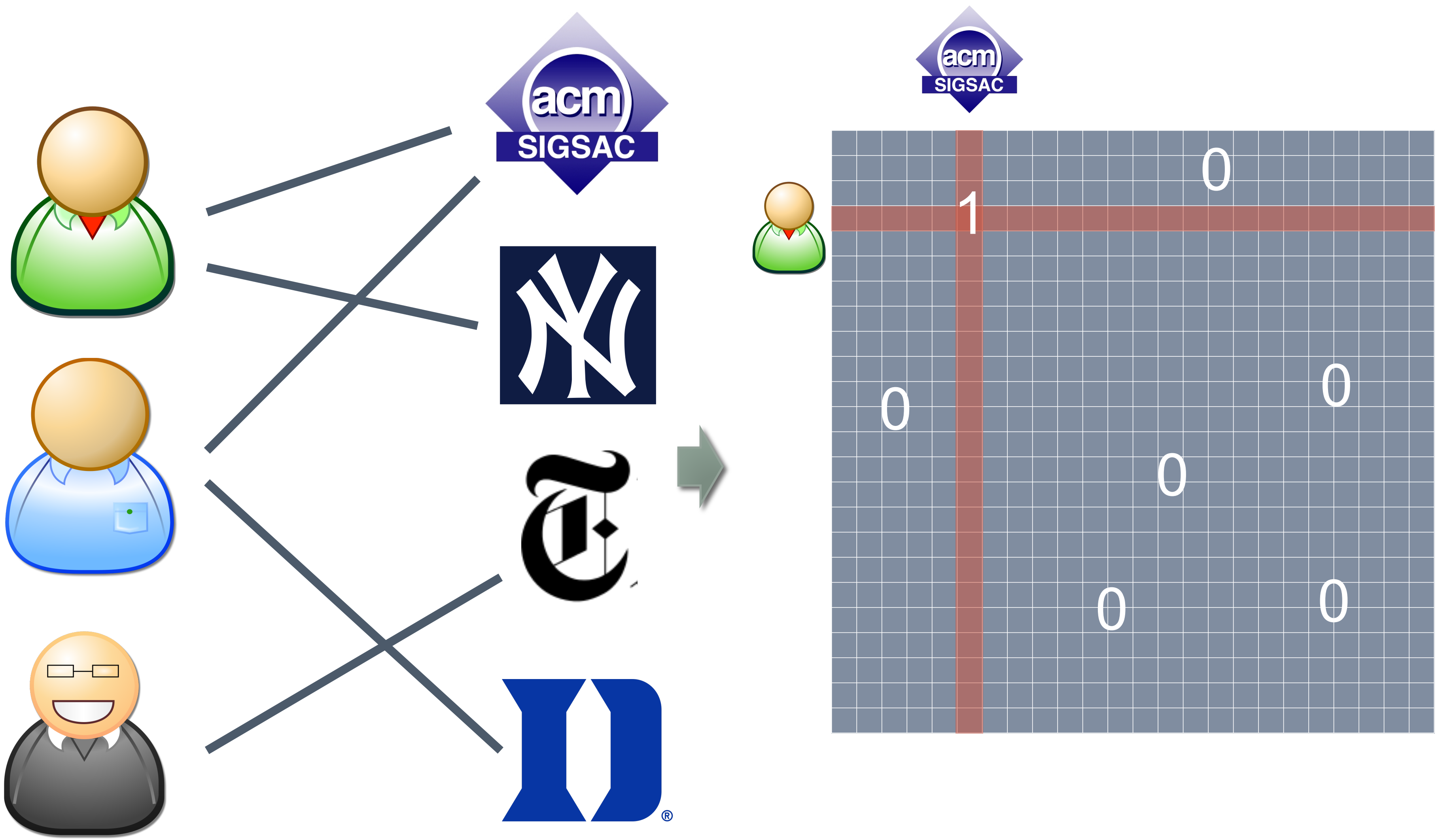
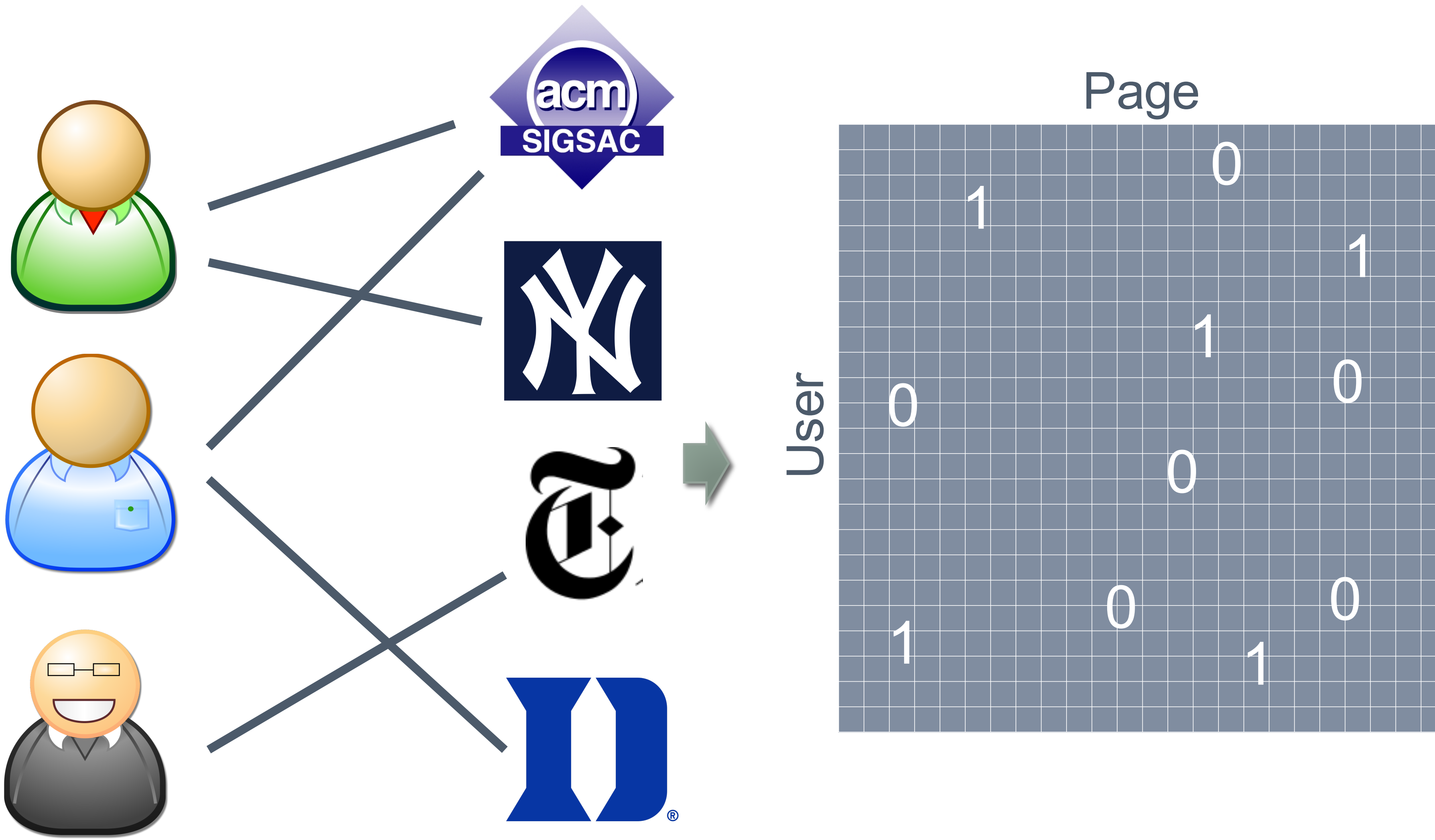
1. Subgraph Analysis
2. Propagation Methods
3. **Latent Factor Models**
   a) Background
   b) Normal Behavior
   c) Abnormal Behavior

Matrix Modeling



Matrix Modeling

Matrix M



Matrix Modeling

HITS

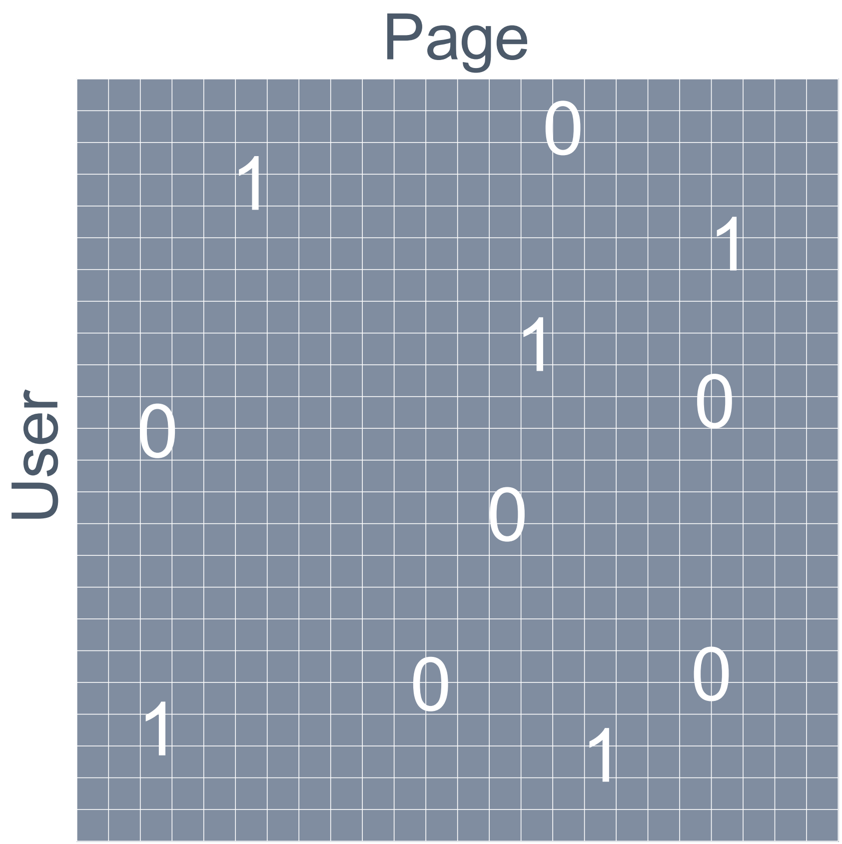
Authoritativeness $\vec{v}$ is first eigenvector of $M^{\mathrm{T}}M$

$$\vec{v} = cM^{\mathrm{T}}M\vec{v}$$

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

$$\vec{u} = cMM^{\mathrm{T}}\vec{u}$$

Matrix M



Matrix Modeling

Matrix M

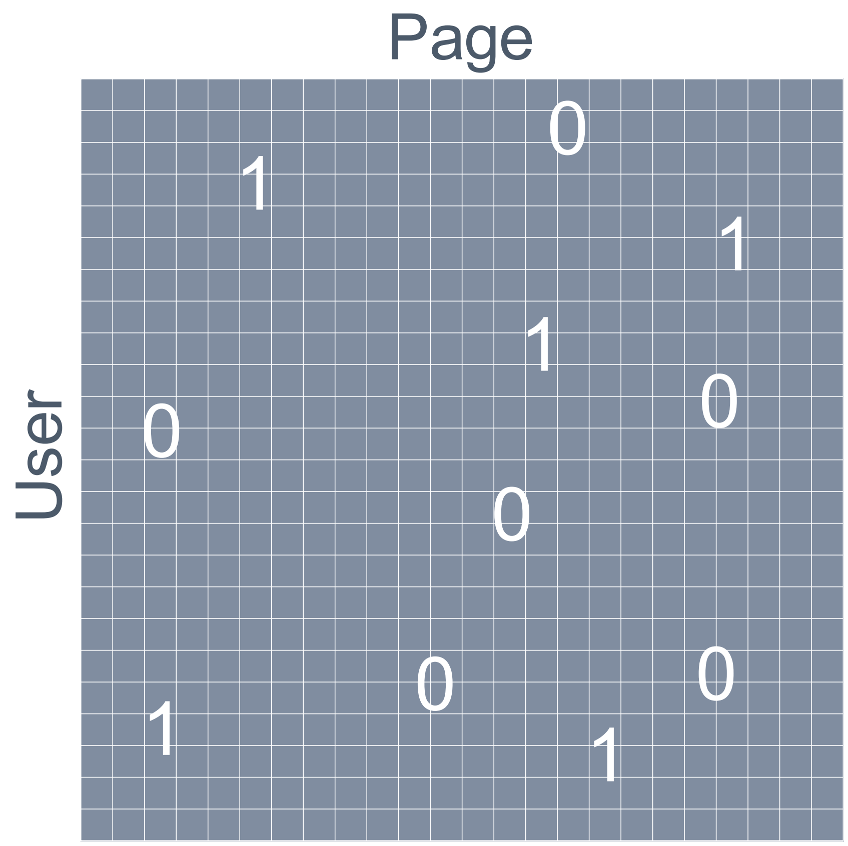
HITS

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

$$\vec{v} = cM^{\mathrm{T}}M\vec{v}$$

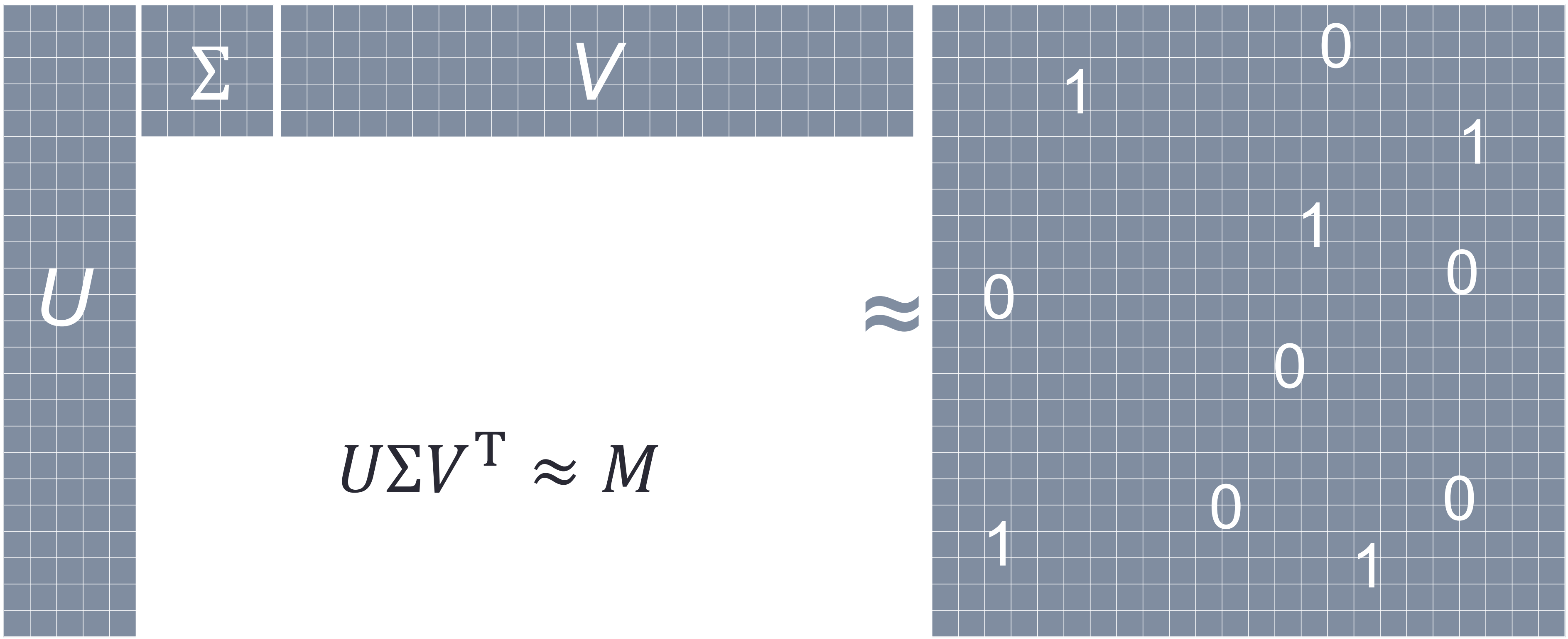
Hubness $\vec{u}$ is first eigenvector of MM^{T}

$$\vec{u} = cMM^{\mathrm{T}}\vec{u}$$



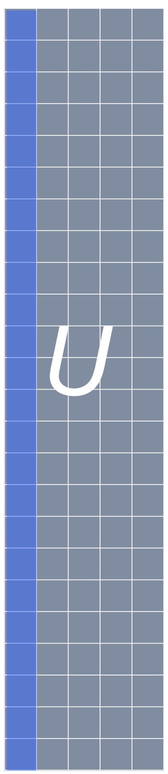
What about the other eigenvectors?

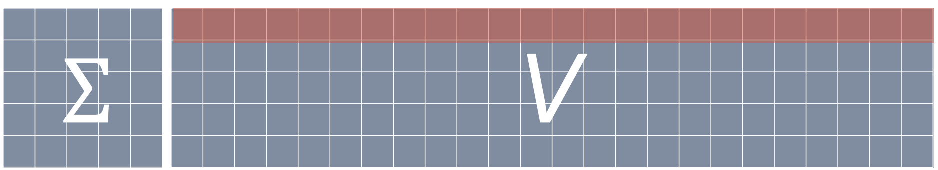
Matrix Modeling
Singular Value Decomposition

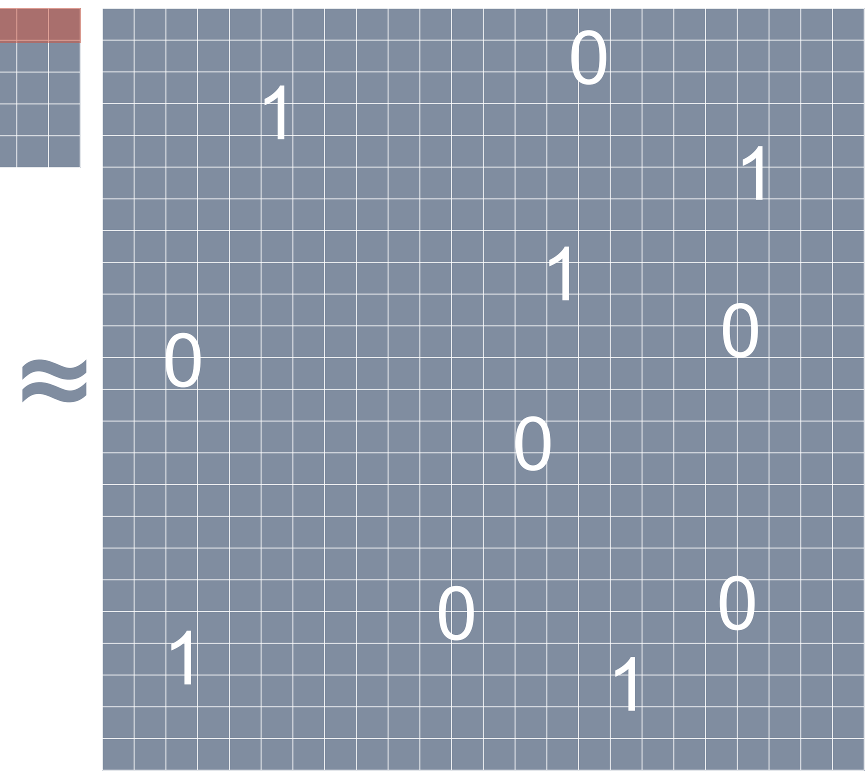


$$U \Sigma V^{\mathrm{T}} \approx M$$

Matrix Modeling
Singular Value Decomposition





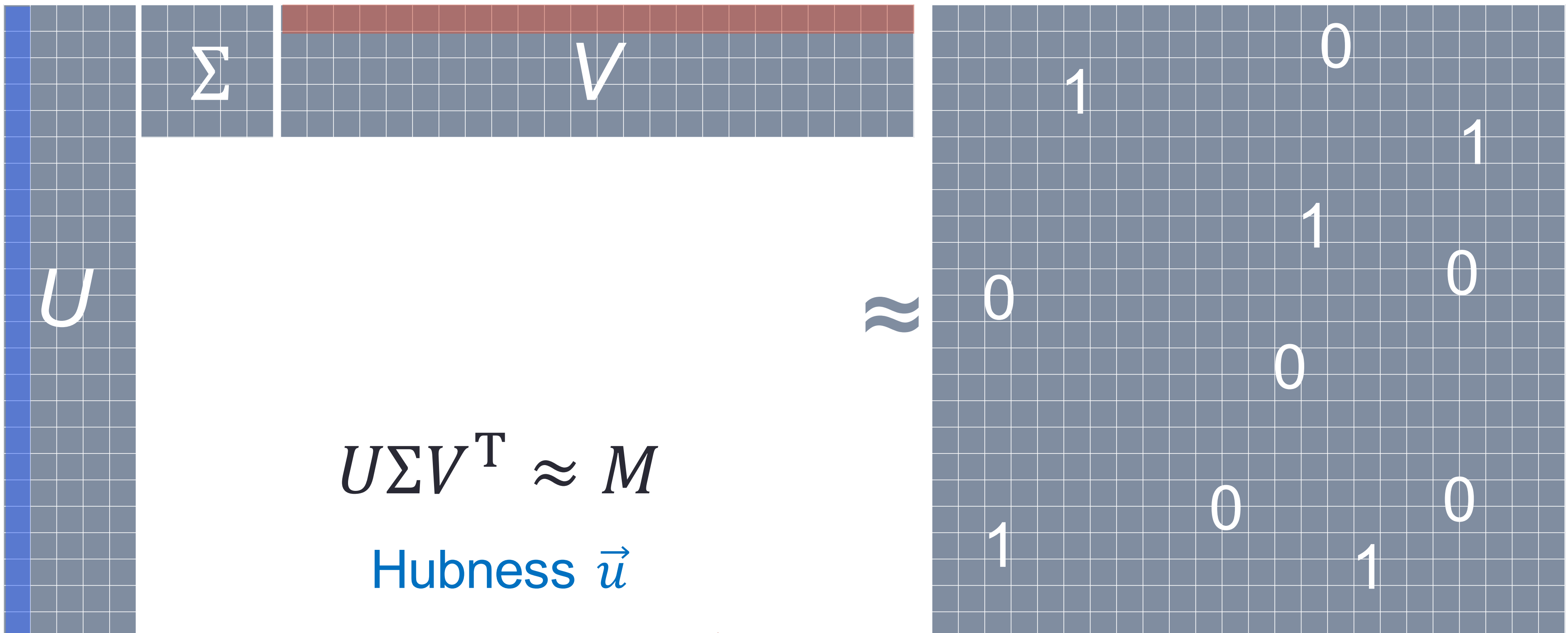


$$U\Sigma V^{\mathrm{T}} \approx M$$

Hubness $\vec{u}$

Authoritativeness $\vec{v}$

Matrix Modeling
Singular Value Decomposition



$$U \Sigma V^{\mathrm{T}} \approx M$$

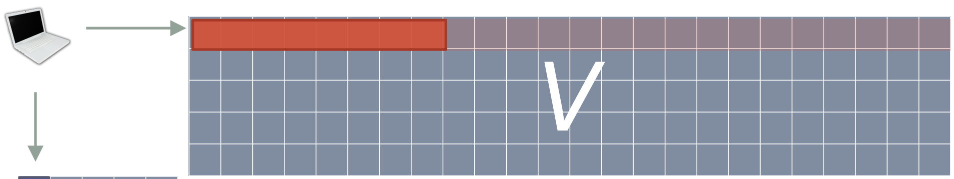
Hubness $\vec{u}$

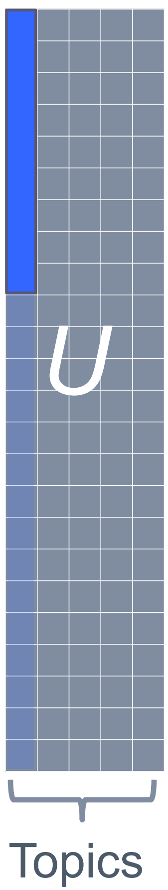
Authoritativeness $\vec{v}$

Σ contains normalization for $\vec{u}$ and $\vec{v}$

Matrix Factorization

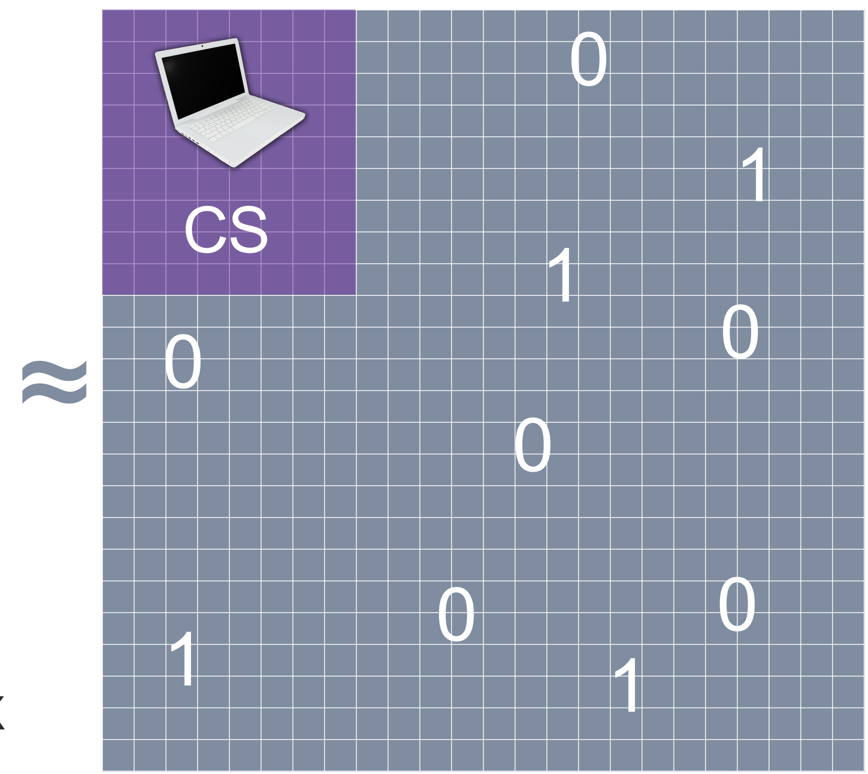
What does each eigenvector capture?





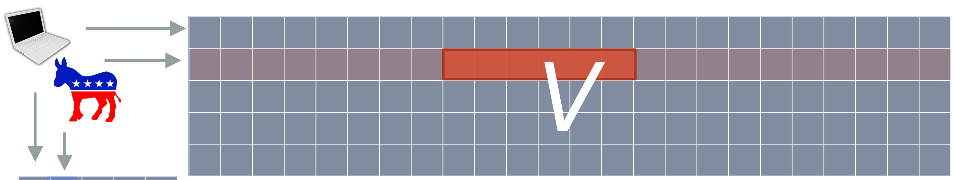
$$UV^{\mathrm{T}} \approx M$$

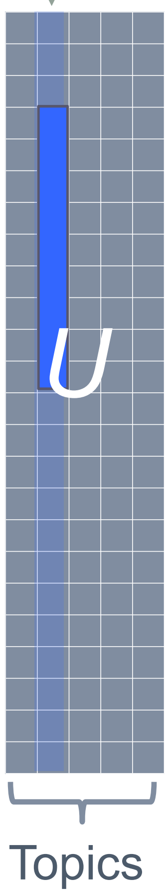
Each factor captures a dense block in the matrix



Matrix Factorization

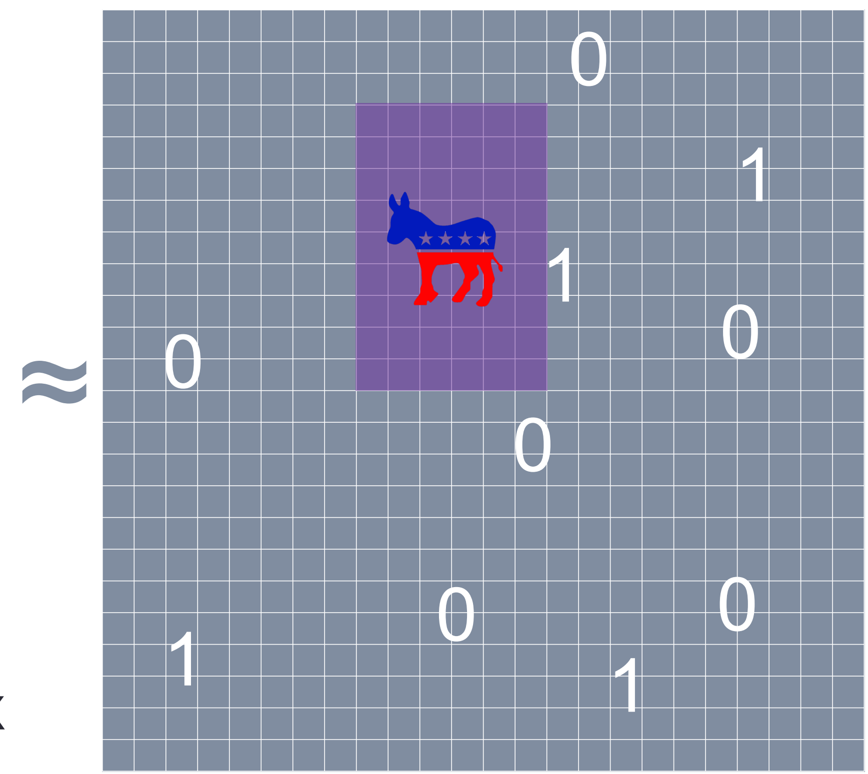
What does each eigenvector capture?



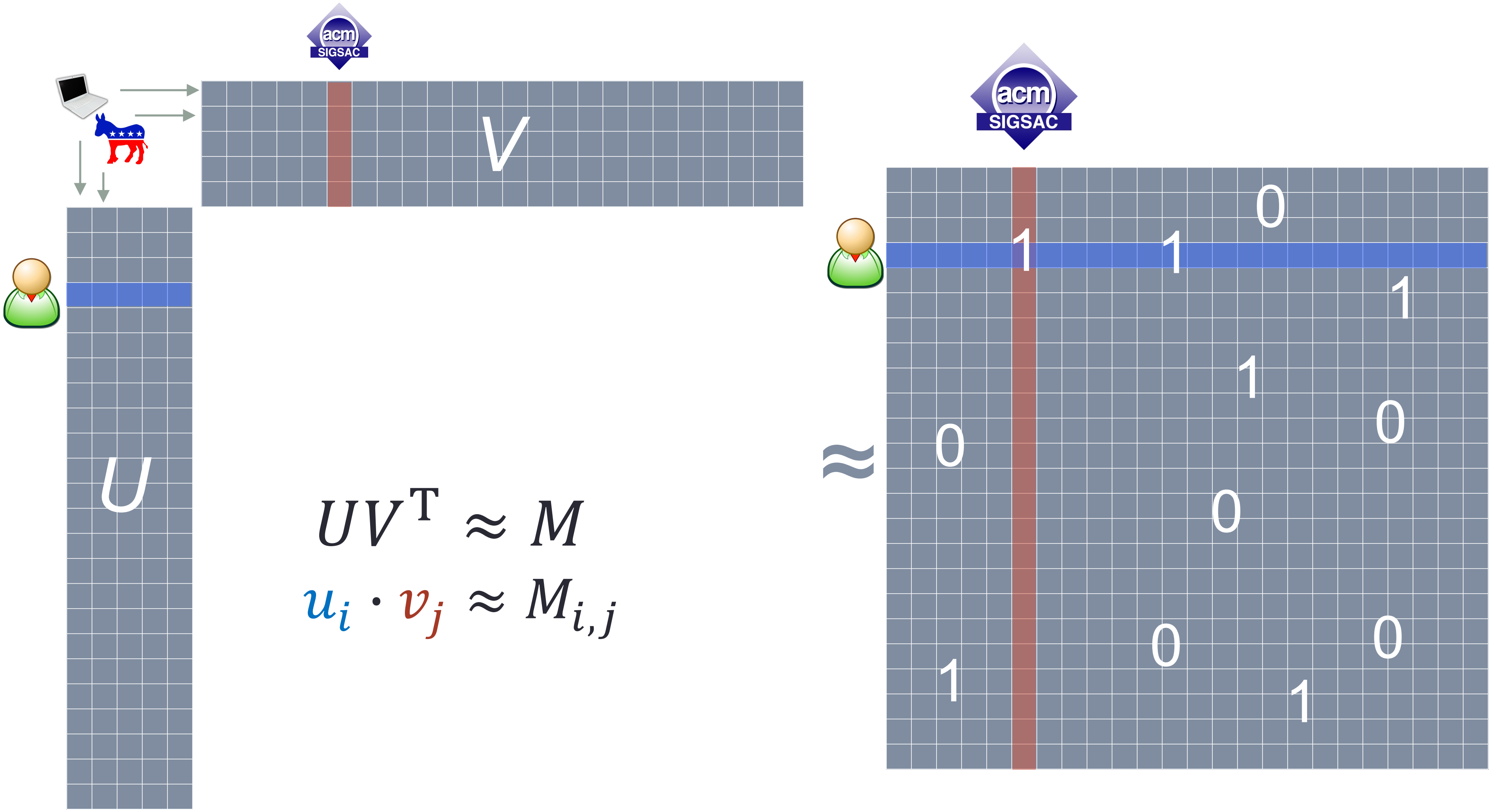


$$UV^{\mathrm{T}} \approx M$$

Each factor captures a dense block in the matrix



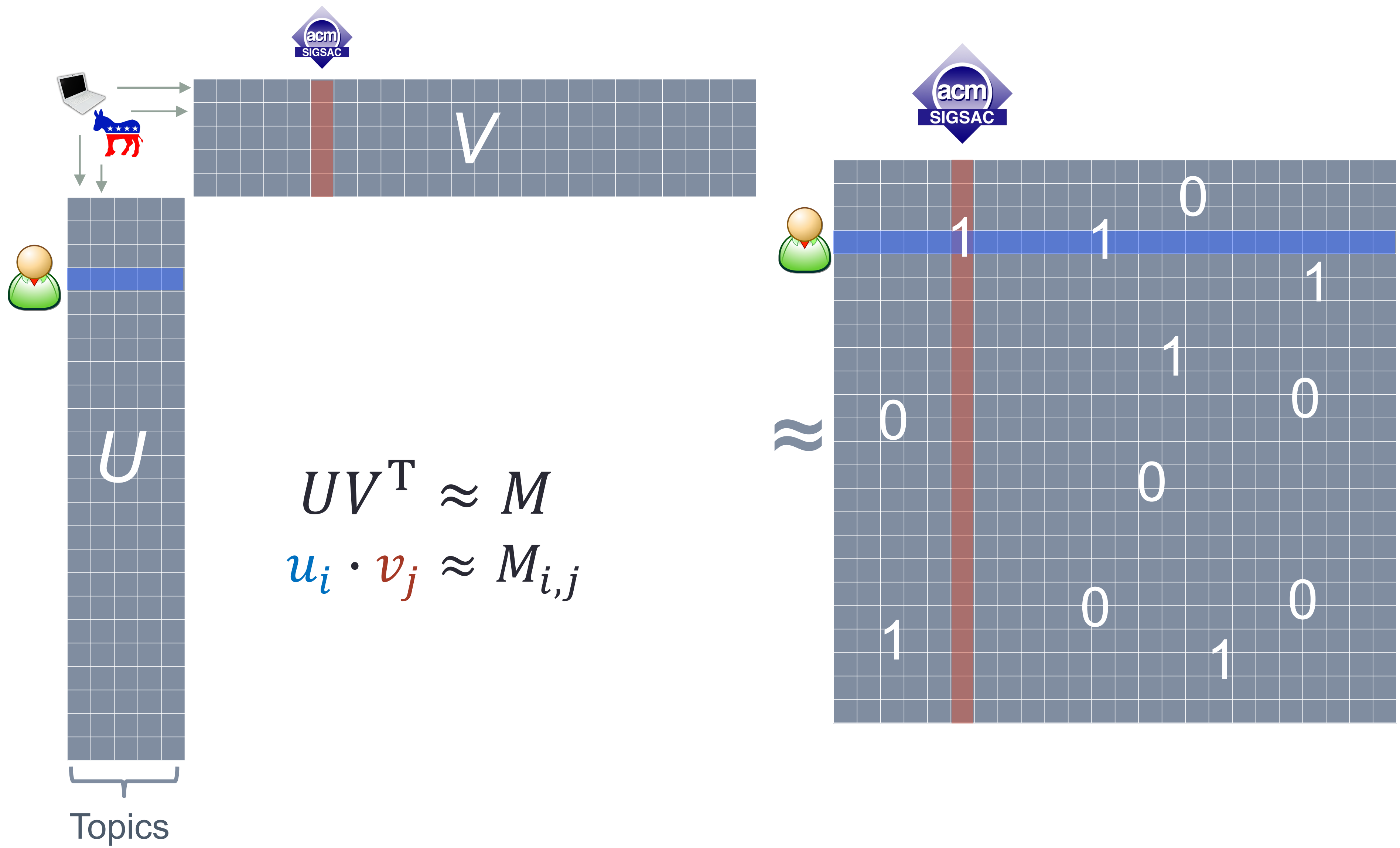
Matrix Factorization



$$UV^{\mathrm{T}} \approx M$$

$$u_i \cdot v_j \approx M_{i,j}$$

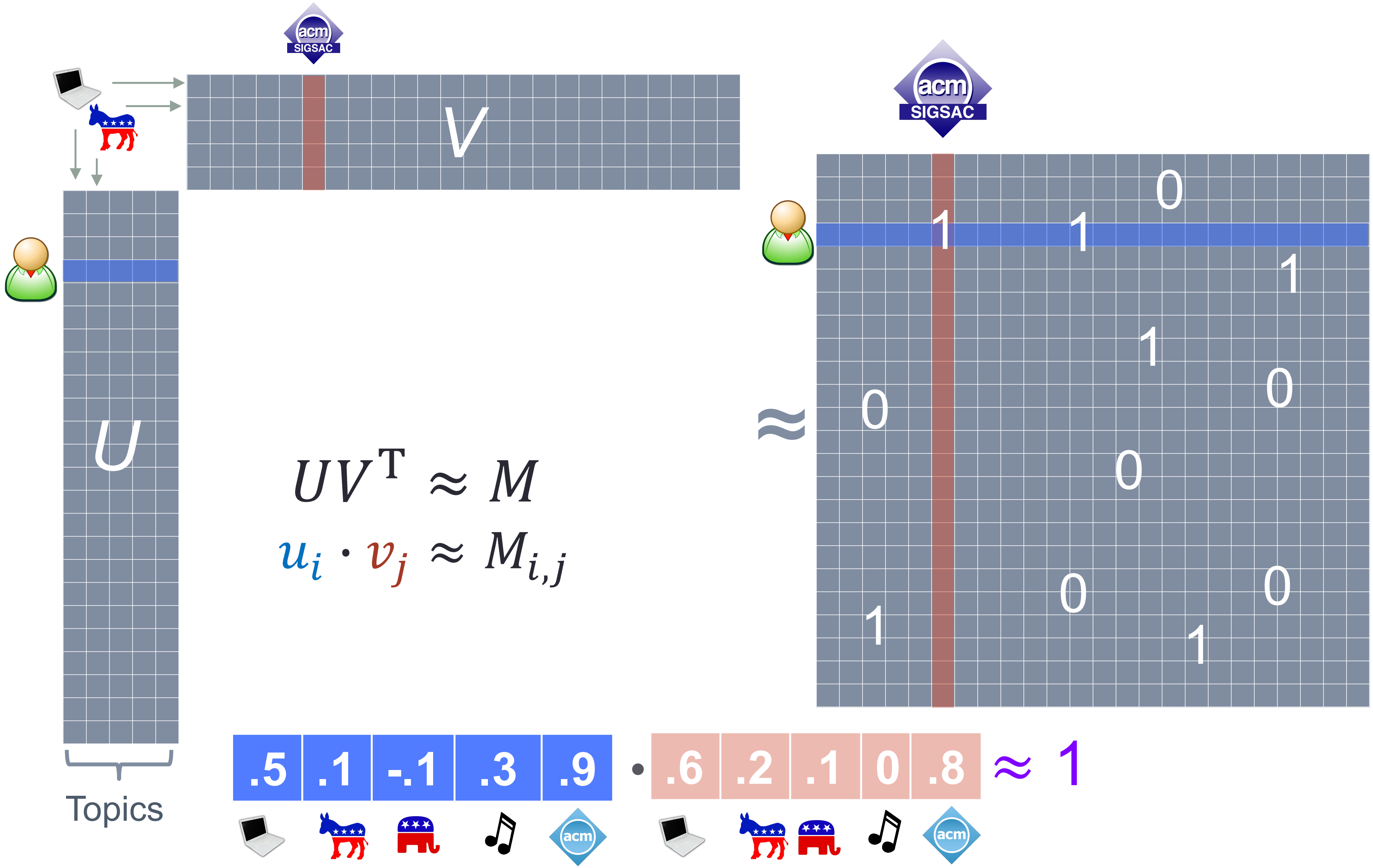
Matrix Factorization



$$UV^{\mathrm{T}} \approx M$$

$$u_i \cdot v_j \approx M_{i,j}$$

Matrix Factorization

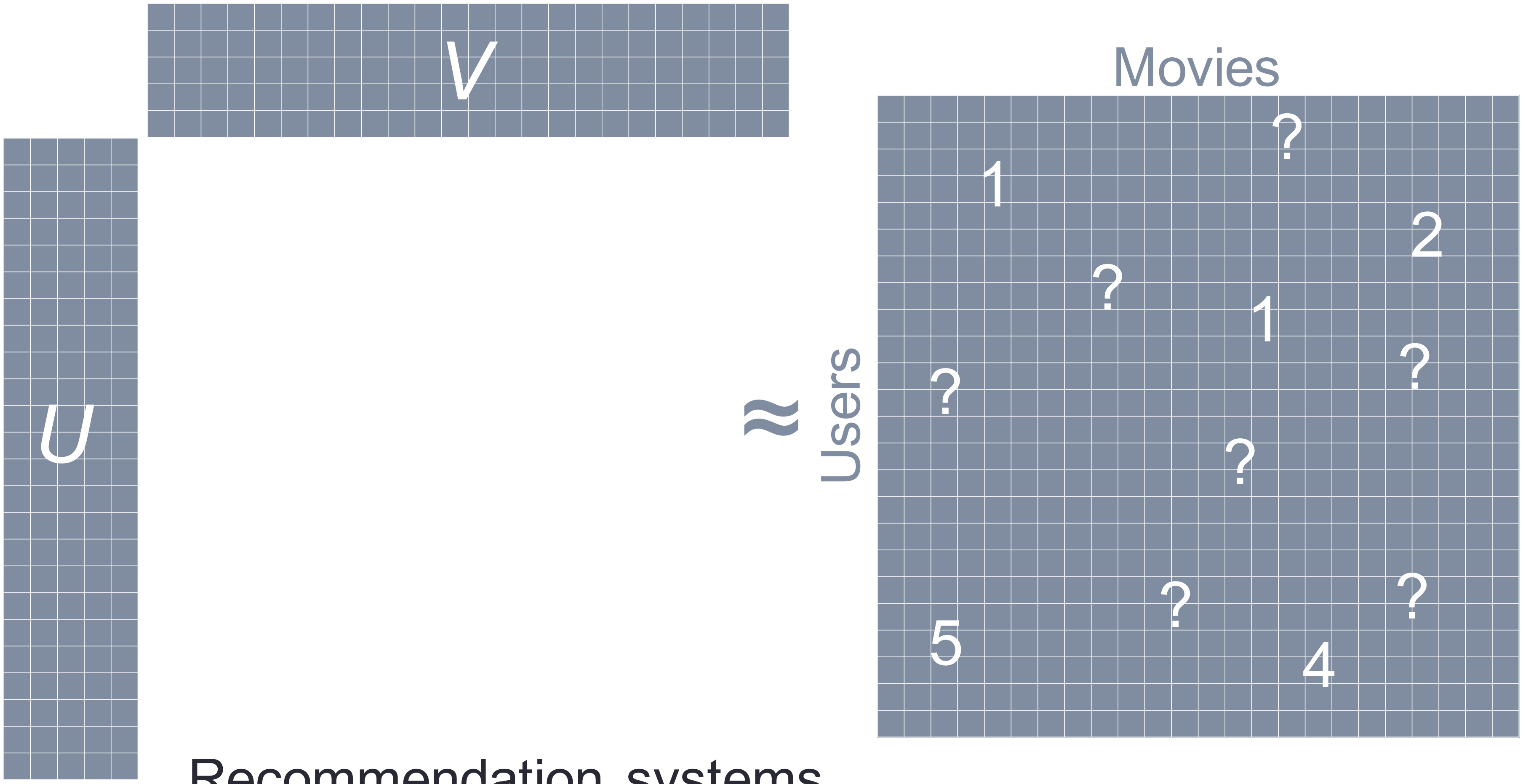


$$UV^{\mathrm{T}} \approx M$$

$$u_i \cdot v_j \approx M_{i,j}$$

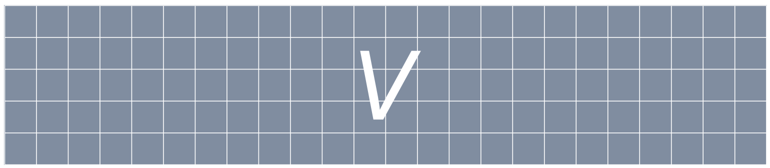
1. Subgraph Analysis
2. Propagation Methods
3. **Latent Factor Models**
   a) Background
   b) Normal Behavior
   c) Abnormal Behavior

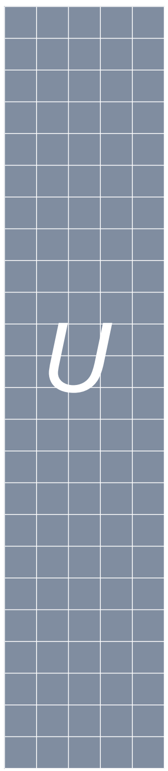
Matrix Completion



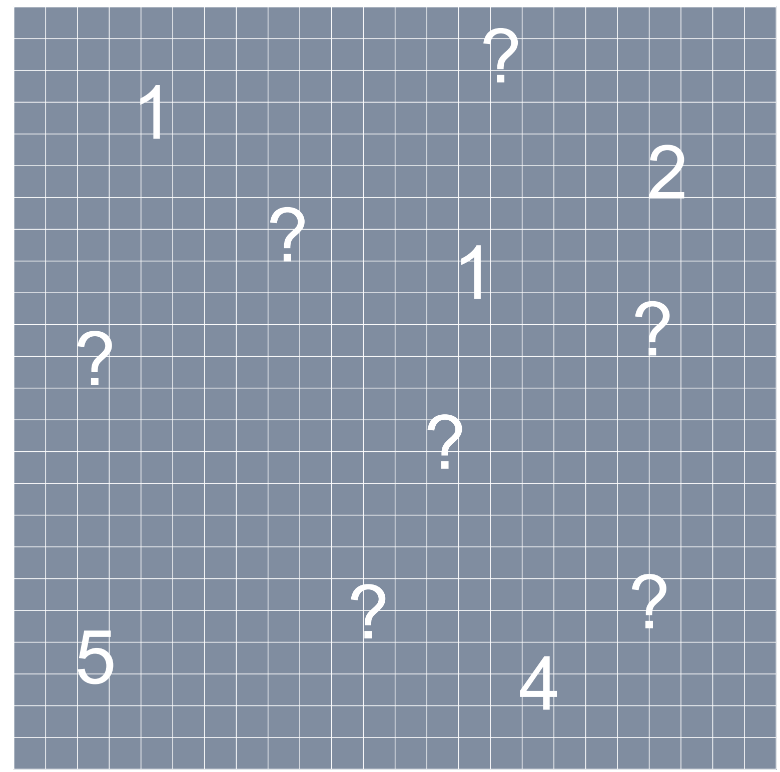
Recommendation systems predict missing entries

Matrix Completion





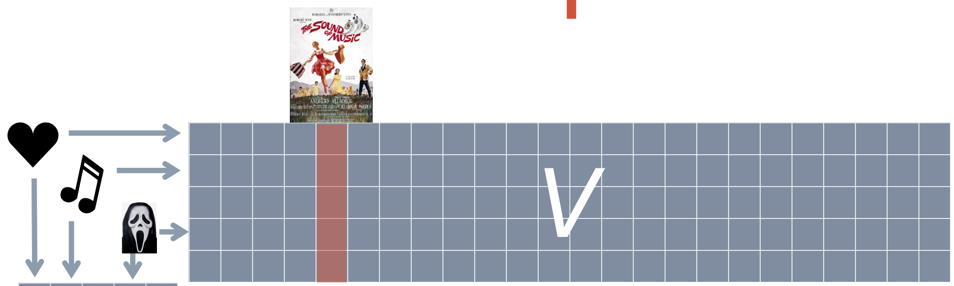
$\approx$

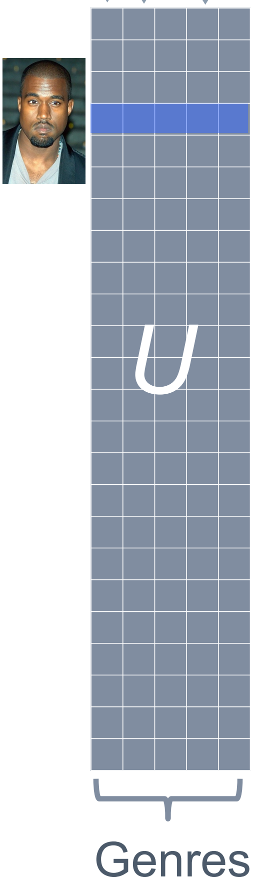


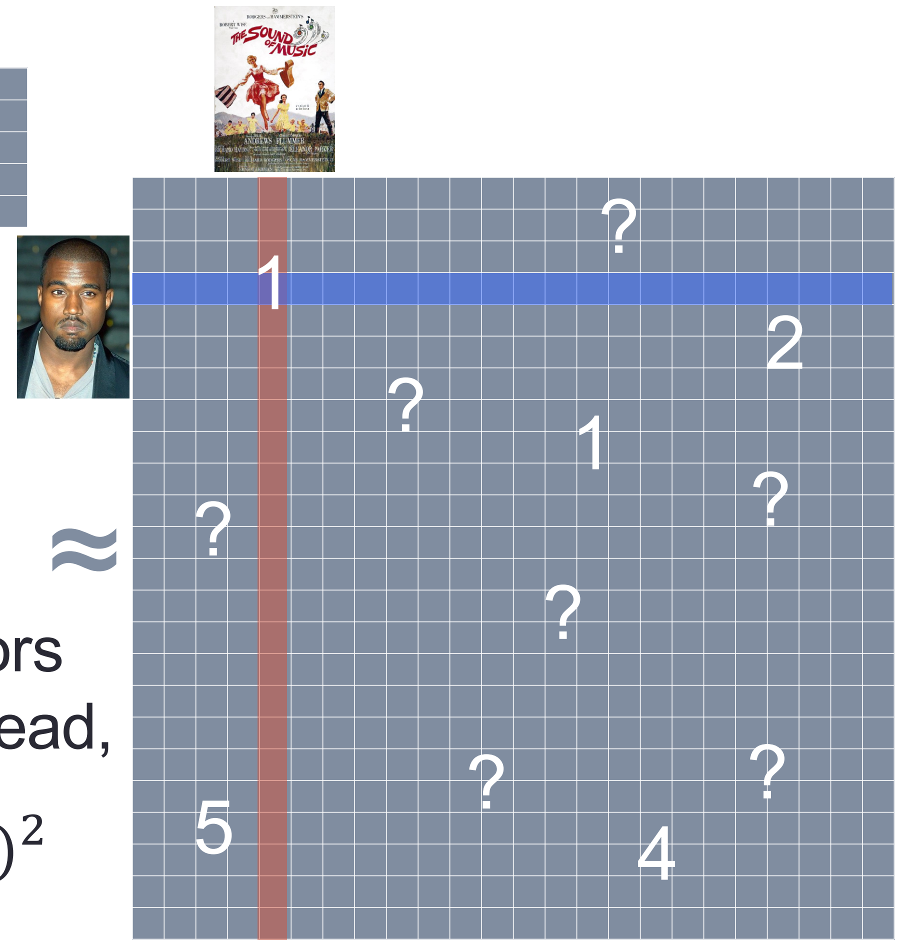
Can't find singular vectors with missing entries. Instead,

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

Matrix Completion



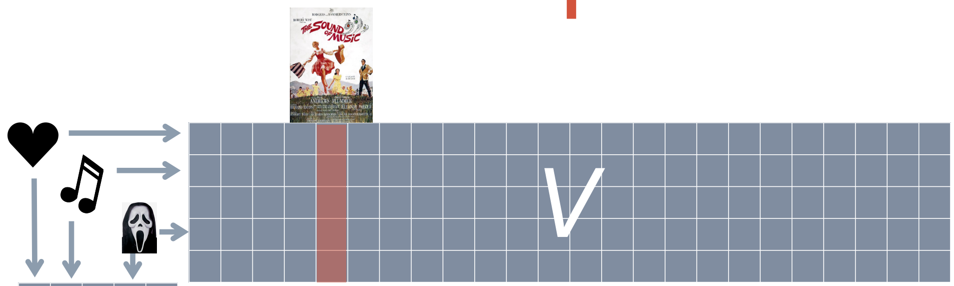


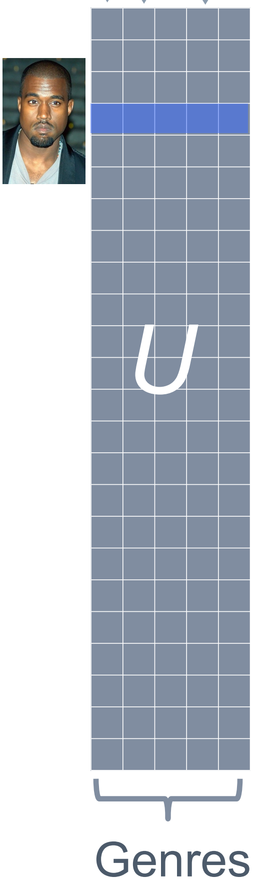


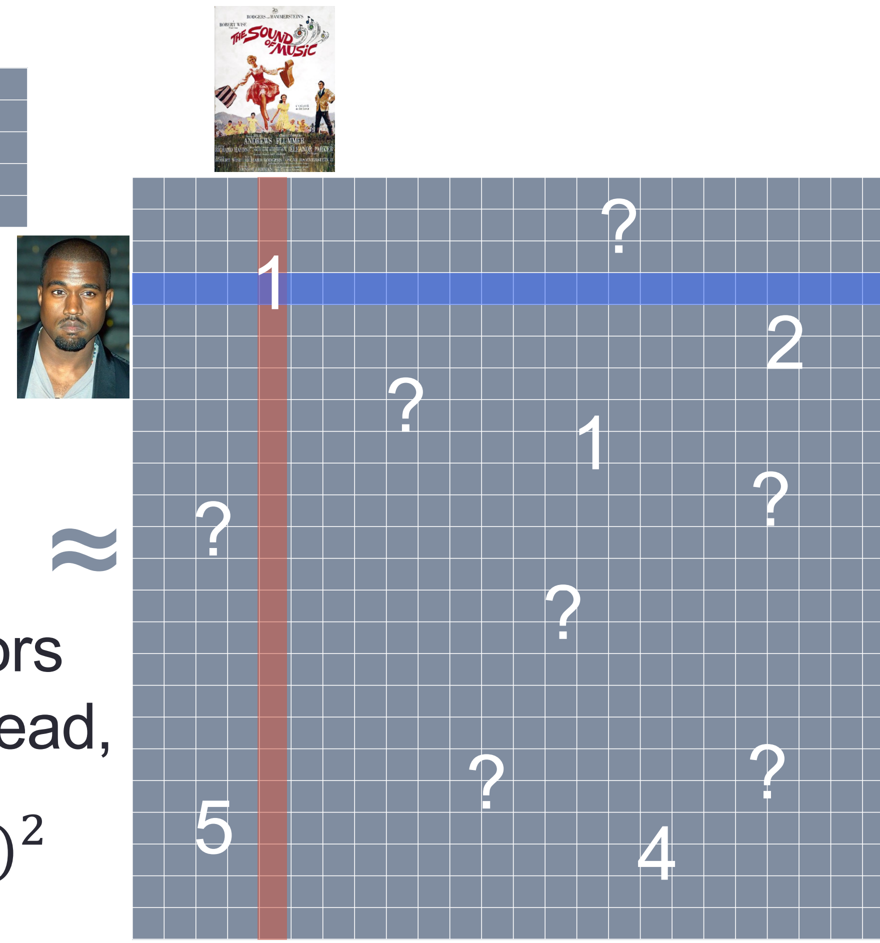
Can’t find singular vectors with missing entries. Instead,

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

Matrix Completion





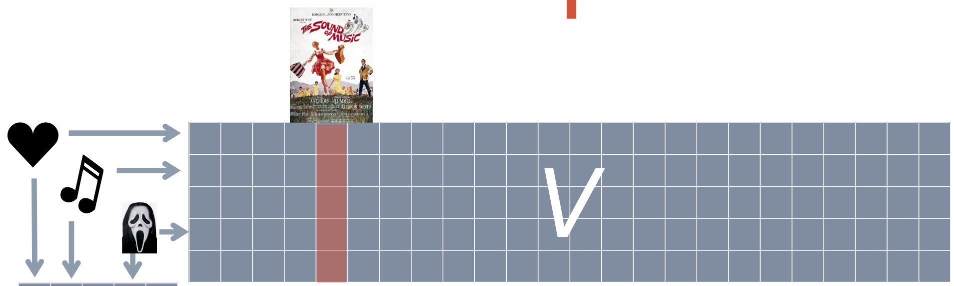


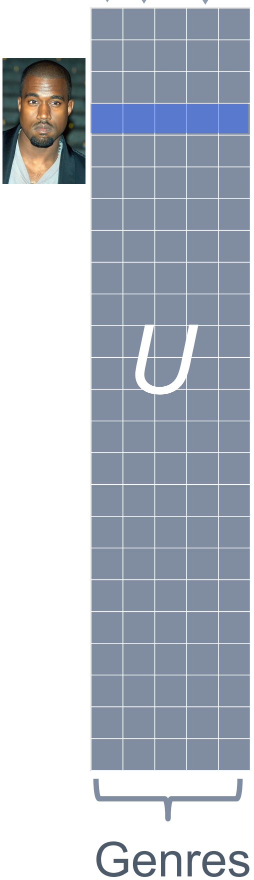
Can't find singular vectors with missing entries. Instead,

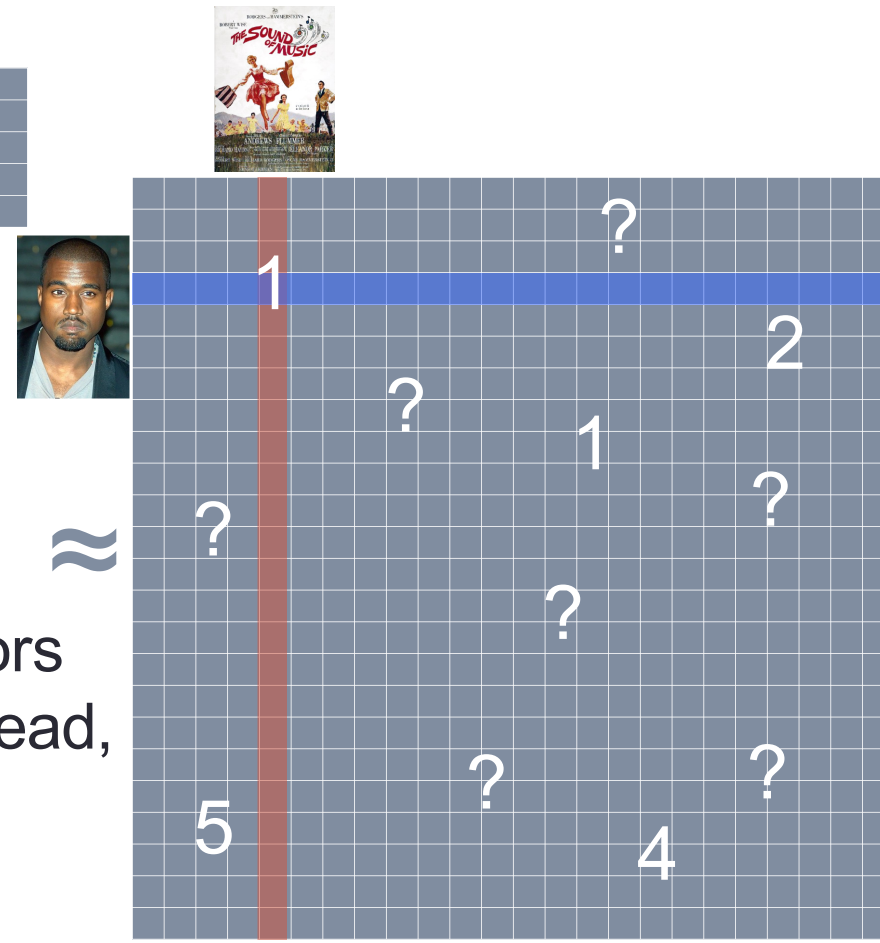
$$\min_{U,V} \sum_{(i,j)\in M} (M_{i,j} - \vec{u}_i \cdot \vec{v}_j)^2$$

$$[1.2 \quad -.1 \quad .5 \quad .8 \quad -.5] \cdot [.6 \quad .8 \quad 0 \quad 0 \quad .1] \approx 1$$

Matrix Completion







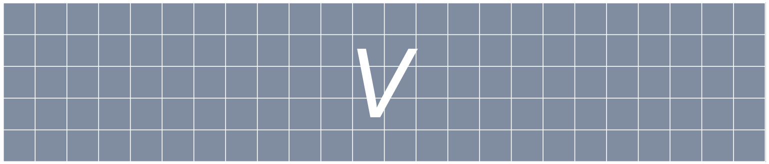
Can't find singular vectors with missing entries. Instead,

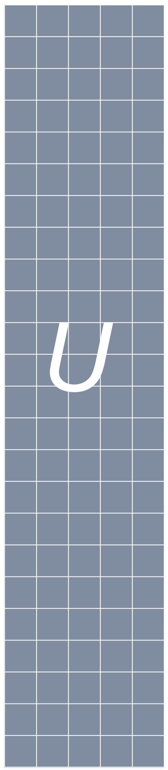
$$\min_{U,V} \sum_{(i,j)\in M} (M_{i,j} - \widehat{M}_{i,j})^2$$

$$\widehat{M}_{i,j} = \vec{u}_i \cdot \vec{v}_j$$

Adding Latent Factors

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

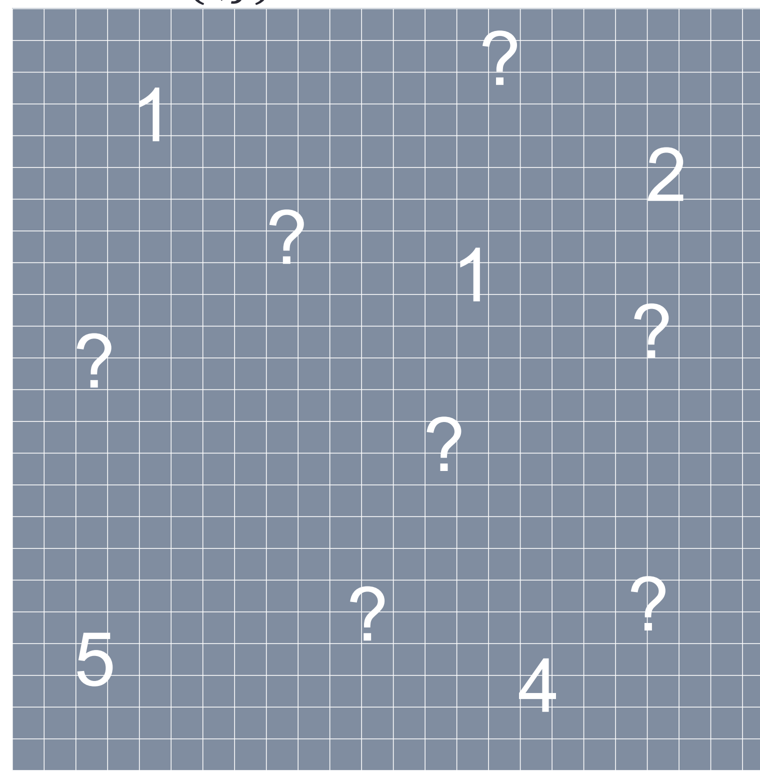




Consider additional factors:

- Dataset mean μ
- Row (user) baseline b_i
- Column (movie) baseline b_j

$\approx$



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

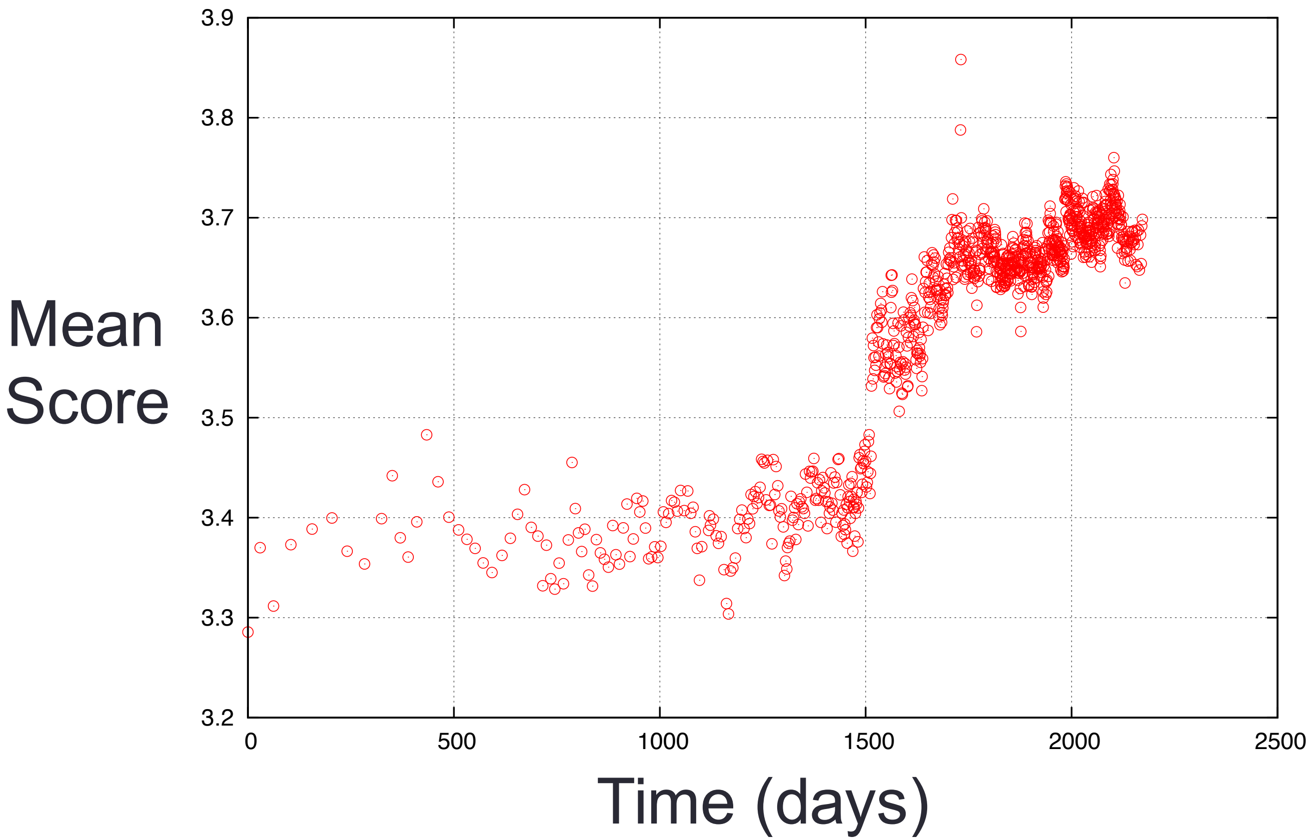
Collaborative Filtering with Temporal Dynamics
Yehuda Koren
KDD 2009

Adding Latent Factors

What if we know the time of the rating (time of the edge being created)?

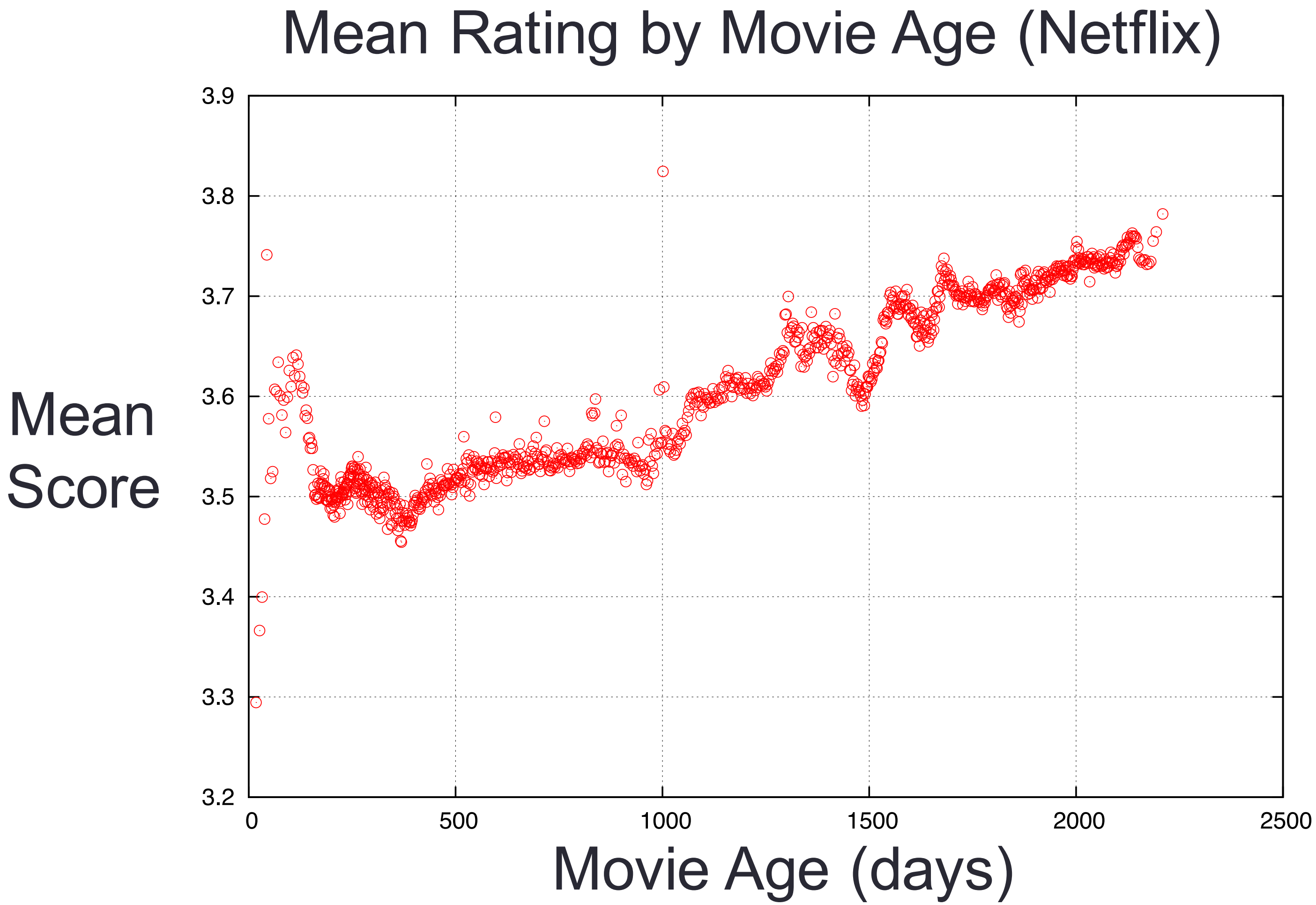
Adding Latent Factors

Mean Rating by Date (Netflix)



Collaborative Filtering with Temporal Dynamics
Yehuda Koren
KDD 2009

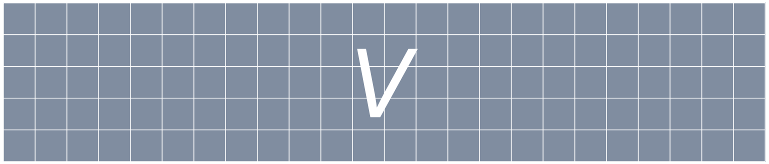
Adding Latent Factors

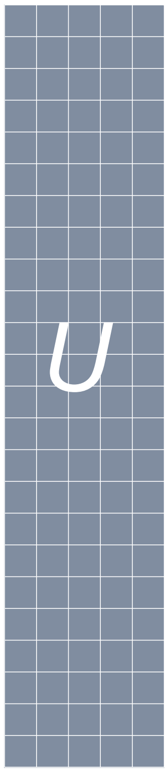


Collaborative Filtering with Temporal Dynamics
Yehuda Koren
KDD 2009

Adding Latent Factors

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

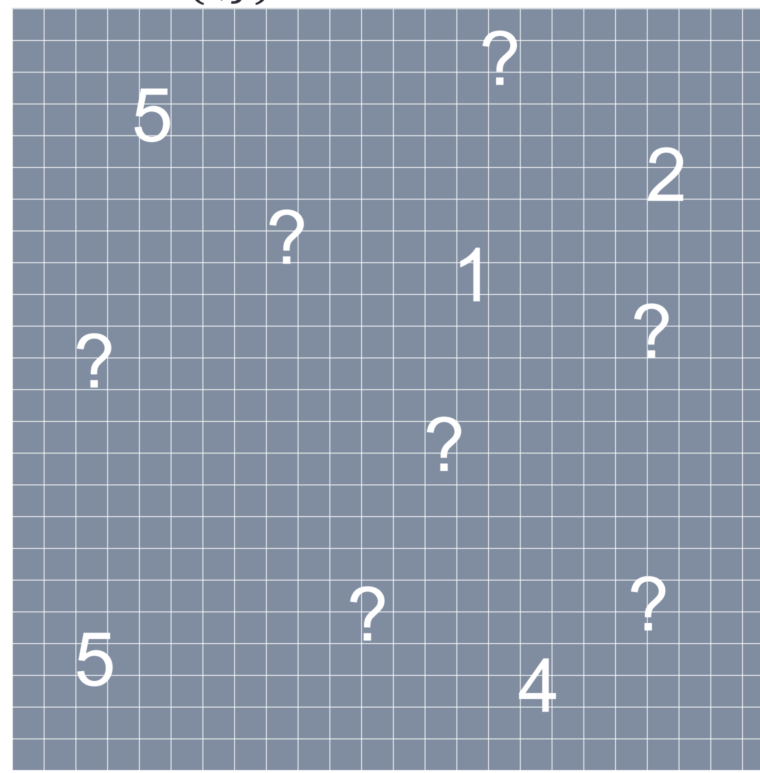




Time factors:
- Column (movie)- time baseline $b_{j,\mathrm{Bin}(t)}$
- Row (user)-time baseline *function* $b_i(t)$

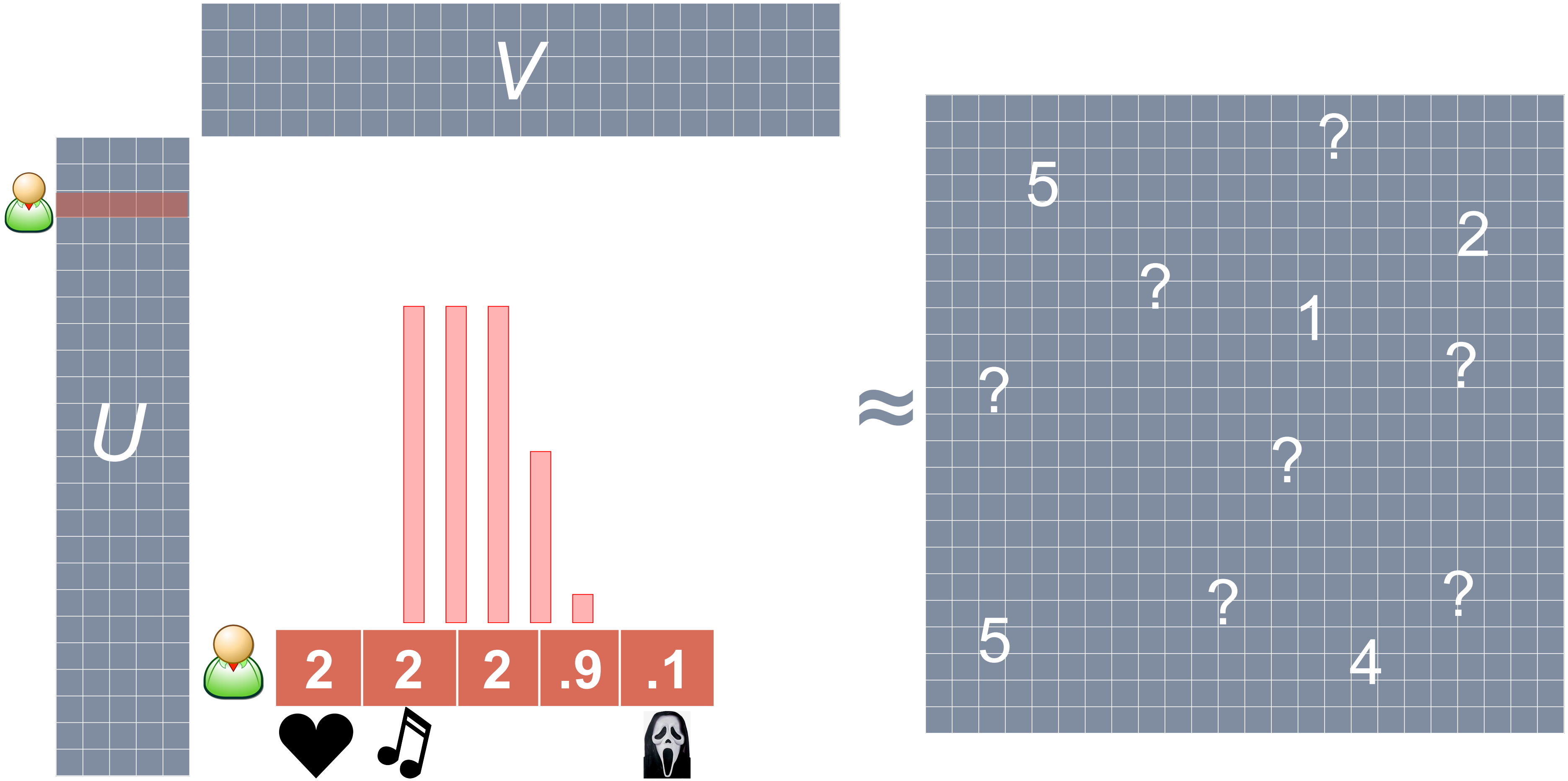
$$\widehat{M}_{i,j} = \mu + b_i + b_j + \vec{u}_i \cdot \vec{v}_j + b_{j,\mathrm{Bin}(t)} + b_i(t)$$

$\approx$



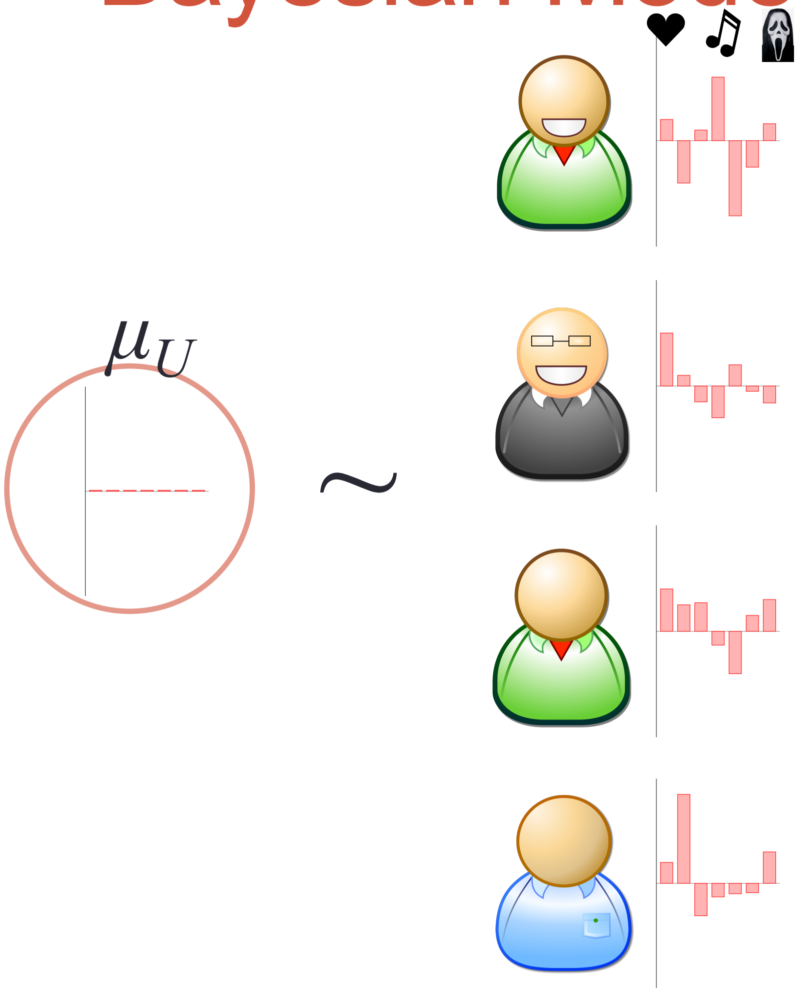
Collaborative Filtering with Temporal Dynamics
Yehuda Koren
KDD 2009

Bayesian Modeling



Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008

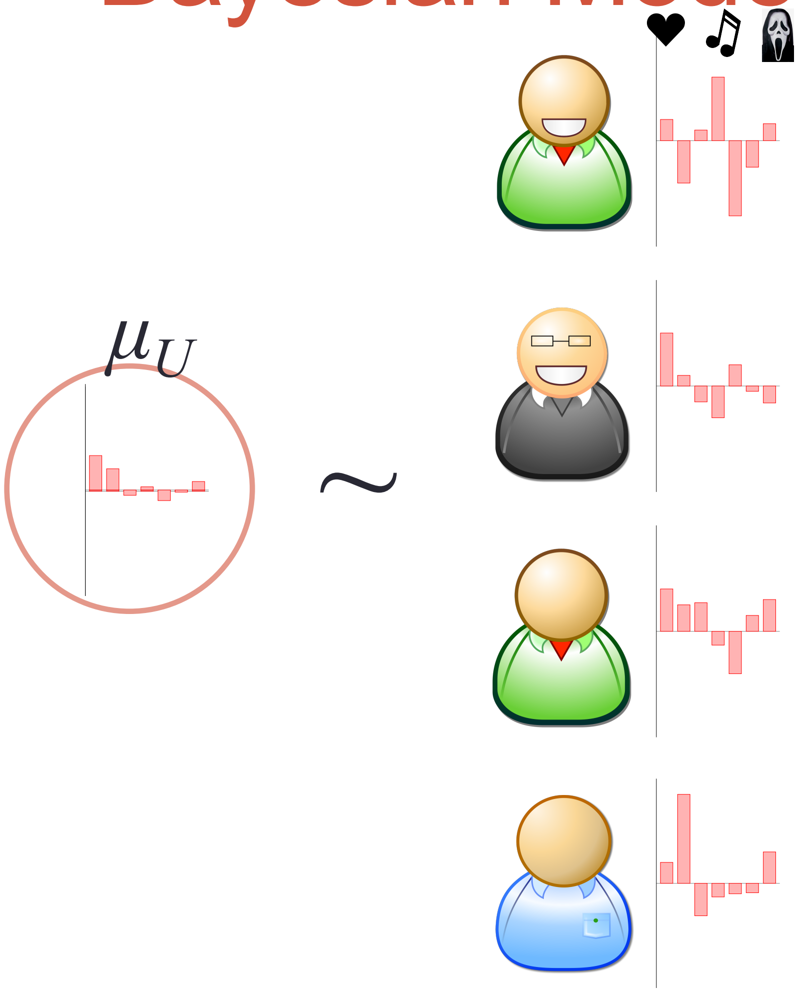
Bayesian Modeling



Sample user factors from Normal distribution

Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008

Bayesian Modeling

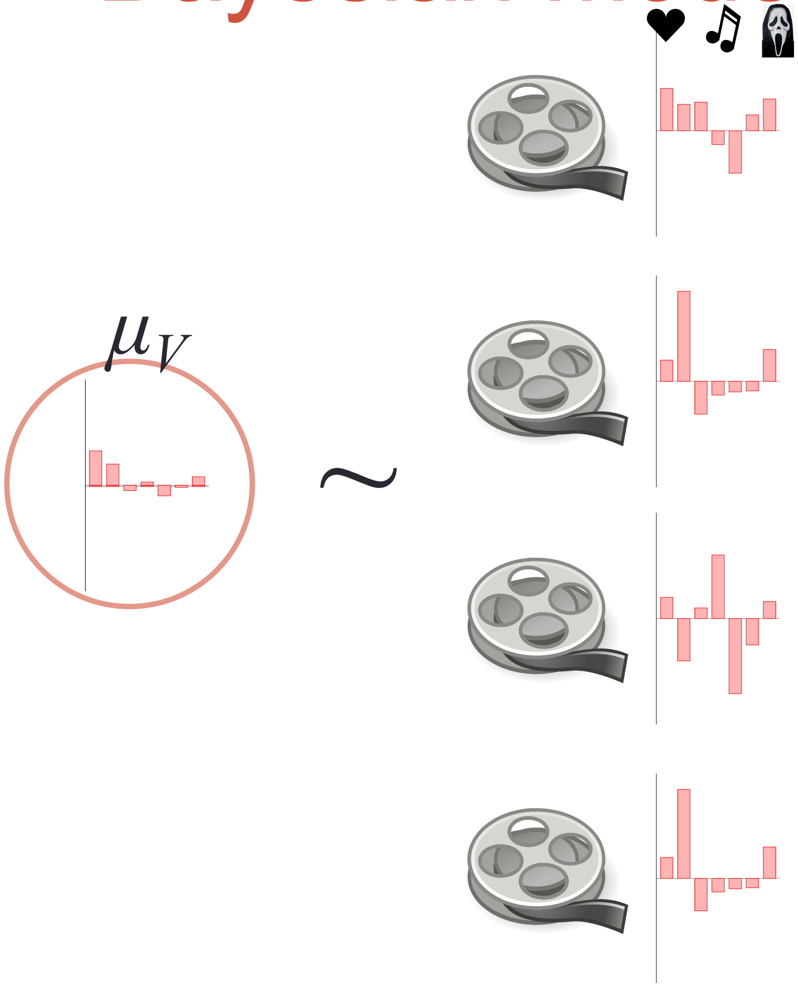


Sample user factors from Normal distribution

Update mean based on user factors

Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008

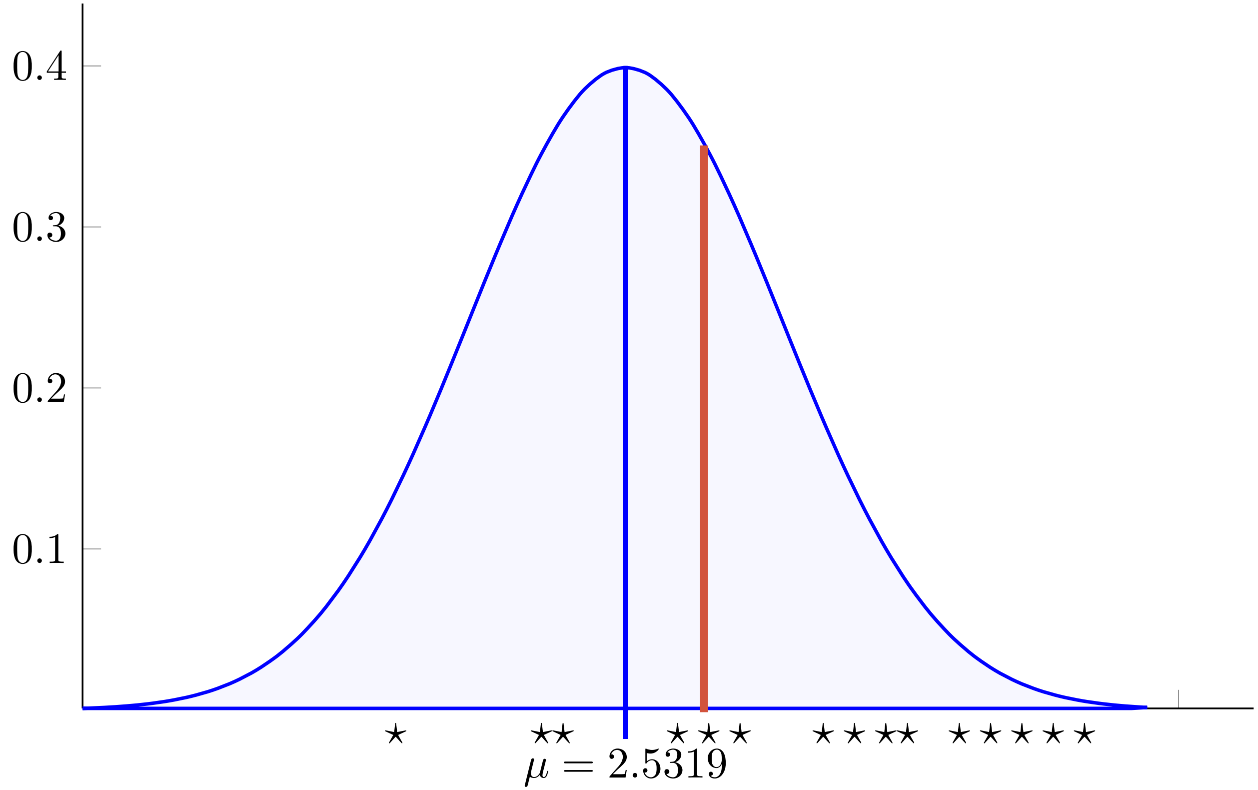
Bayesian Modeling



Similarly sample movie factors

Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008

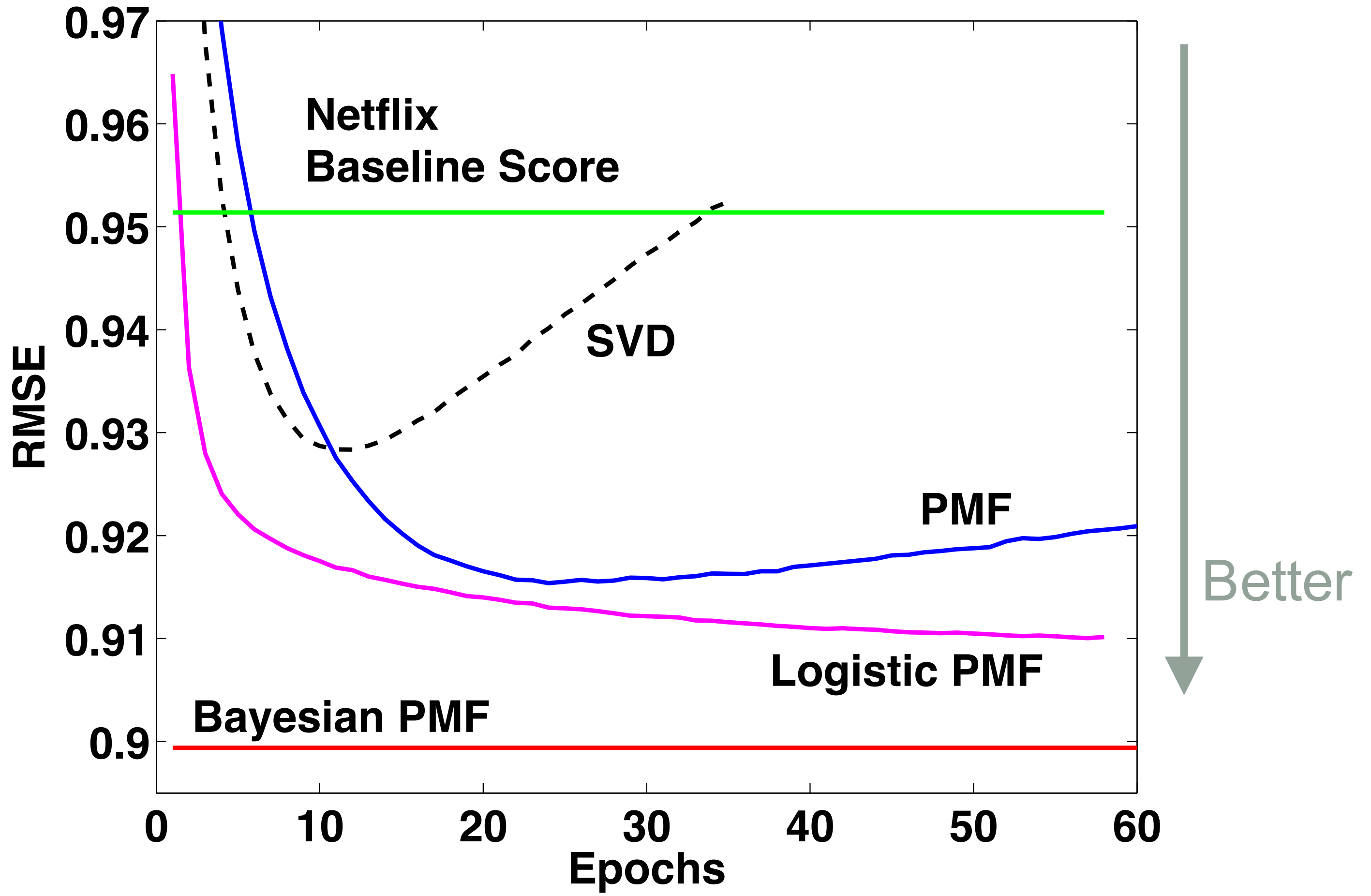
Bayesian Modeling



$$p(M_{i,j} | U, V) = \mathcal{N}(M_{i,j} | \vec{u}_i \cdot \vec{v}_j, \sigma^2)$$

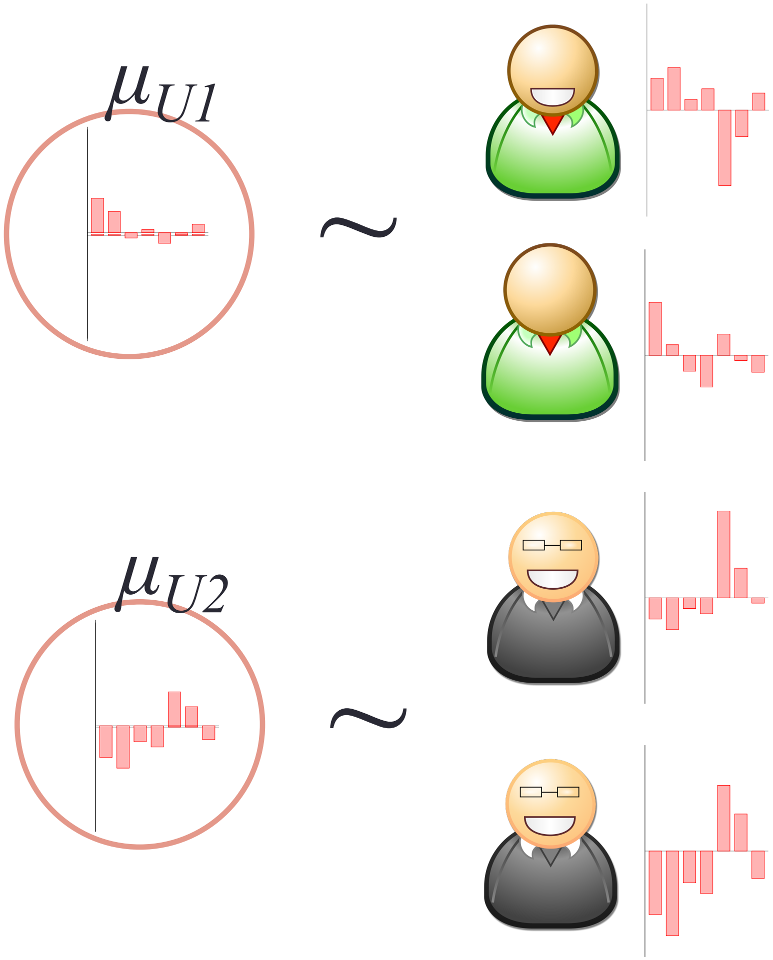
Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008

Bayesian Modeling



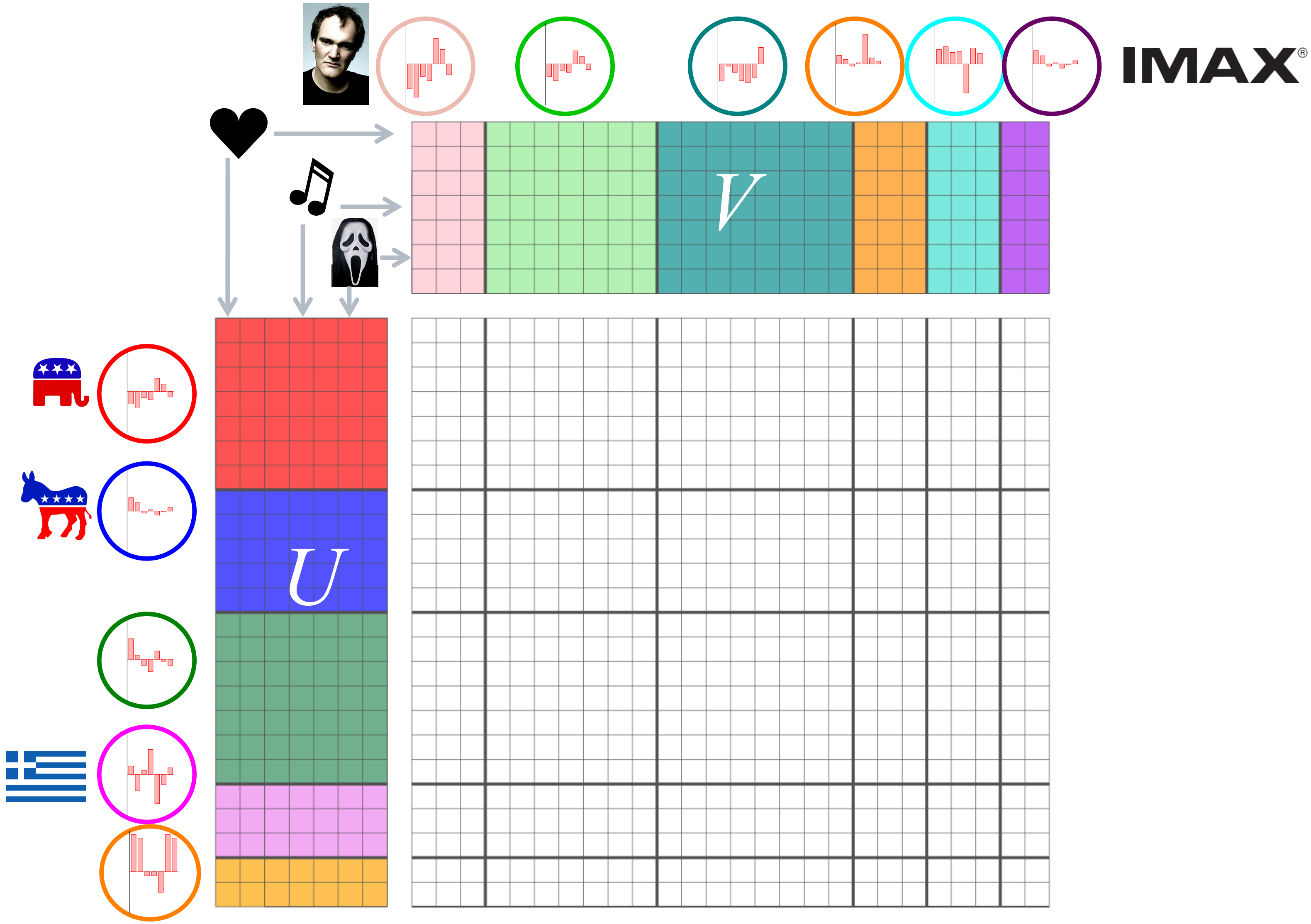
Bayesian Probabilistic Matrix Factorization
Ruslan Salakhutdinov and Andriy Mnih
ICML 2008

Bayesian Modeling with Co-Clustering



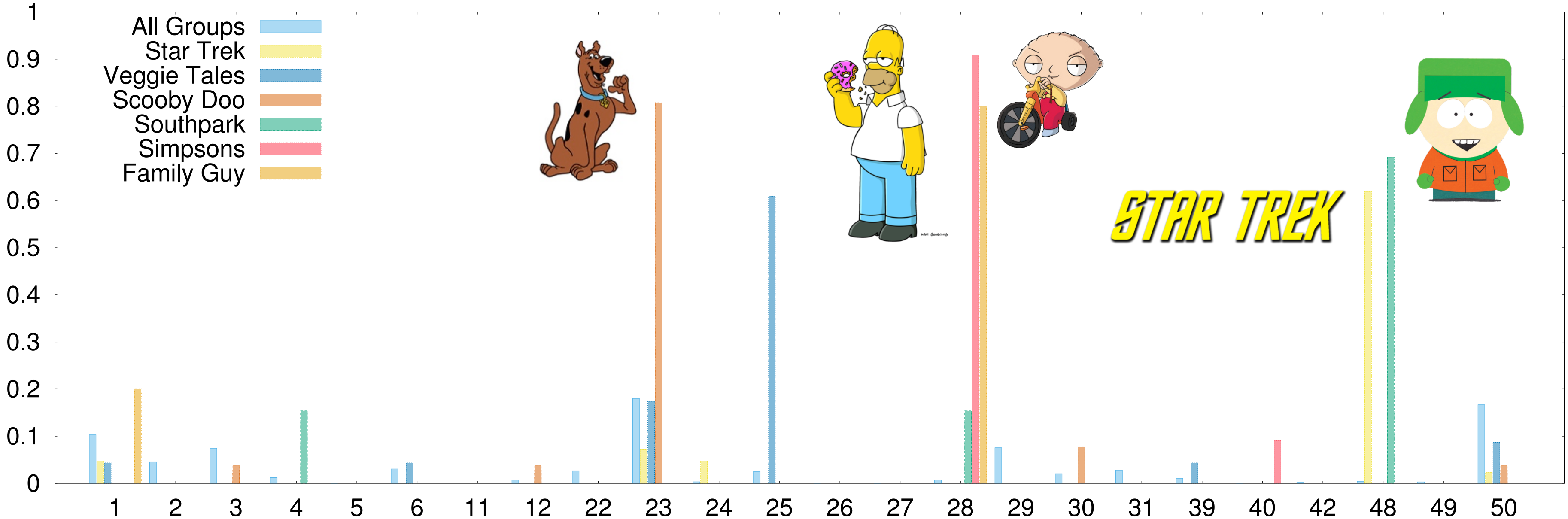
Cluster users
with similar factors

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



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

Bayesian Modeling with Co-Clustering

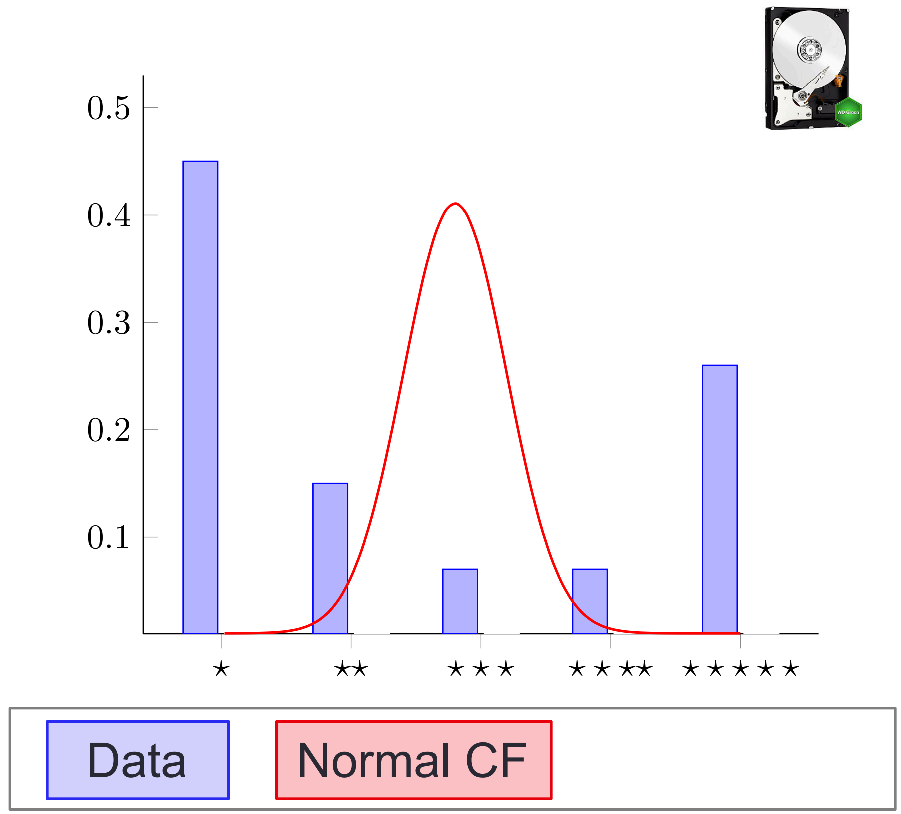


Cluster 28	Cluster 30	Cluster 48
Simpsons	Scooby Doo	Star Trek
Family Guy	Spy Kids	Back to the Future
Monty Python	Stuart Little	Southpark
Curb your Enthusiasm	Dr. Dolittle	Lord of the Rings
The Twilight Zone	Lion King	Harry Potter
Arrested Development	Agent Cody Banks	The X-Files

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

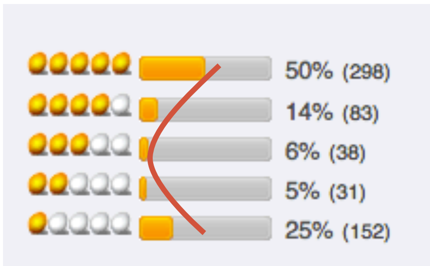
Online Rating Models

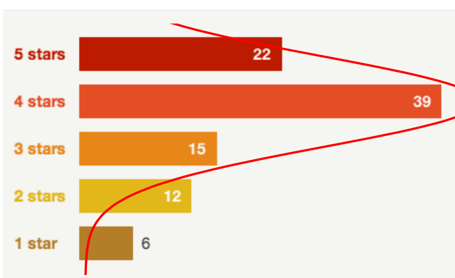
Typically fit a Gaussian - Minimize RMSE



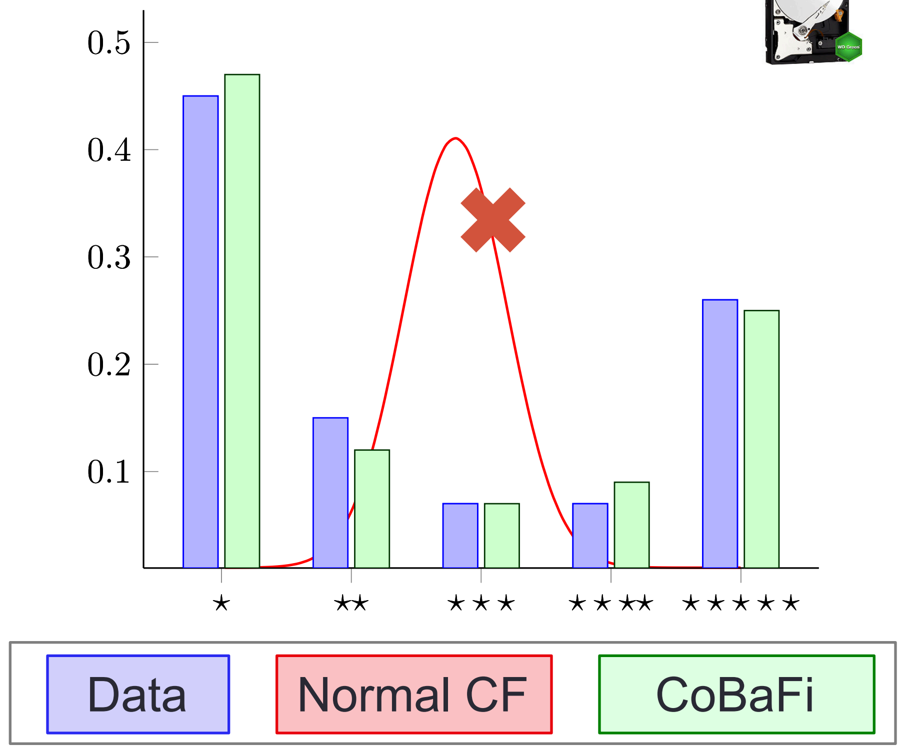
CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014

Online Rating Models



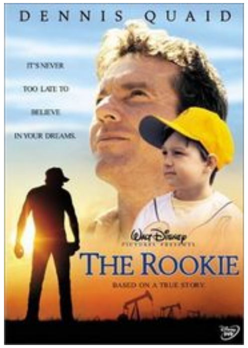


Typically fit a Gaussian - Minimize RMSE

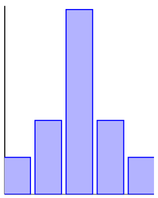


CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014

Shape of Netflix reviews



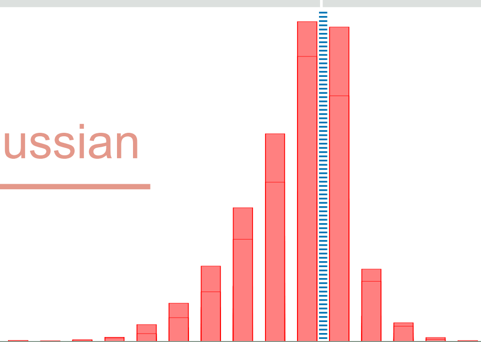
Most Gaussian	Most skewed
The Rookie	The O.C. Season 2
The Fan	Samurai X: Trust and Betrayal
Cadet Kelly	Aqua Teen Hunger Force: Vol. 2
Money Train	Sealab 2001: Season 1
Alice Doesn't Live Here	Aqua Teen Hunger Force: Vol. 2
Sea of Love	Gilmore Girls: Season 3
Boiling Point	Felicity: Season 4

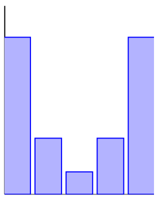


\# Stars

Movies

More Gaussian ← → More Skewed





\# Stars

TV Shows

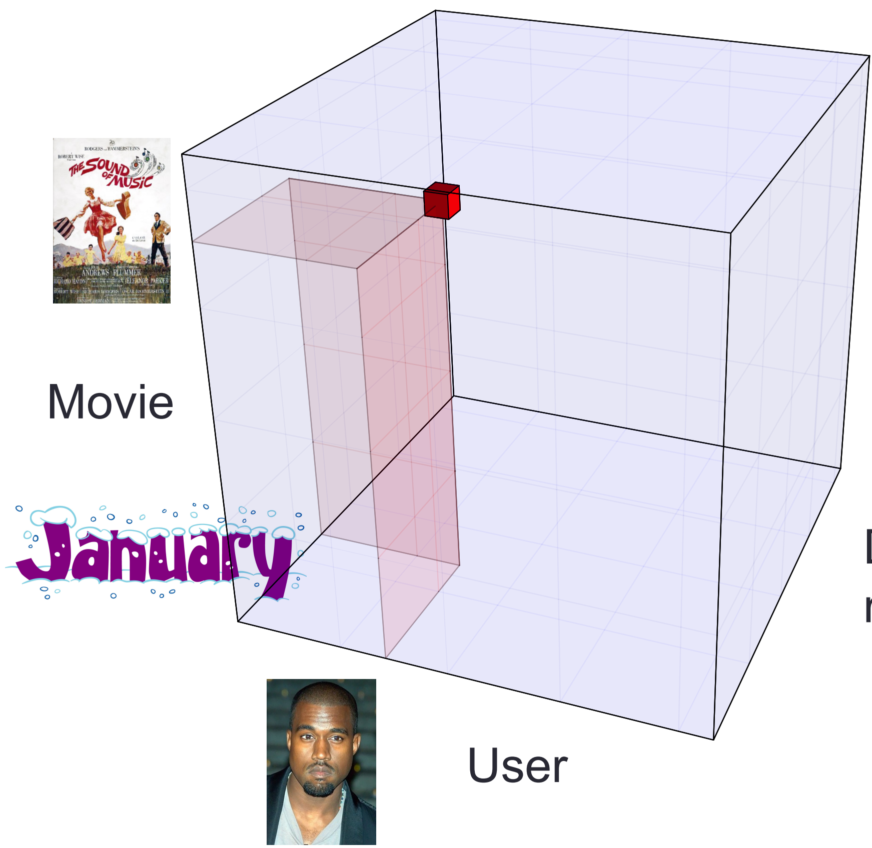
CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014

What is a tensor?

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

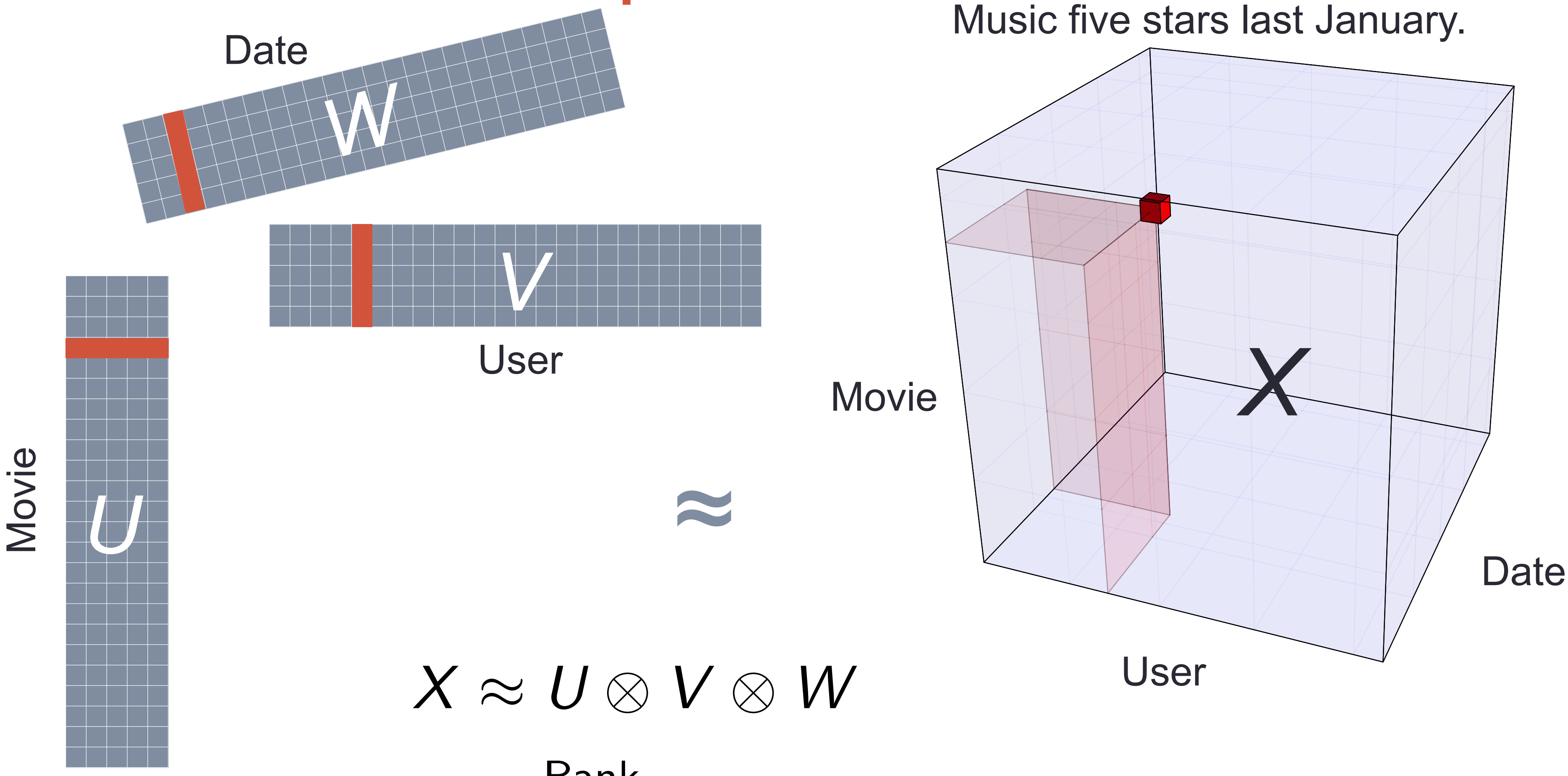
For example:

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



Tensor Decomposition

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



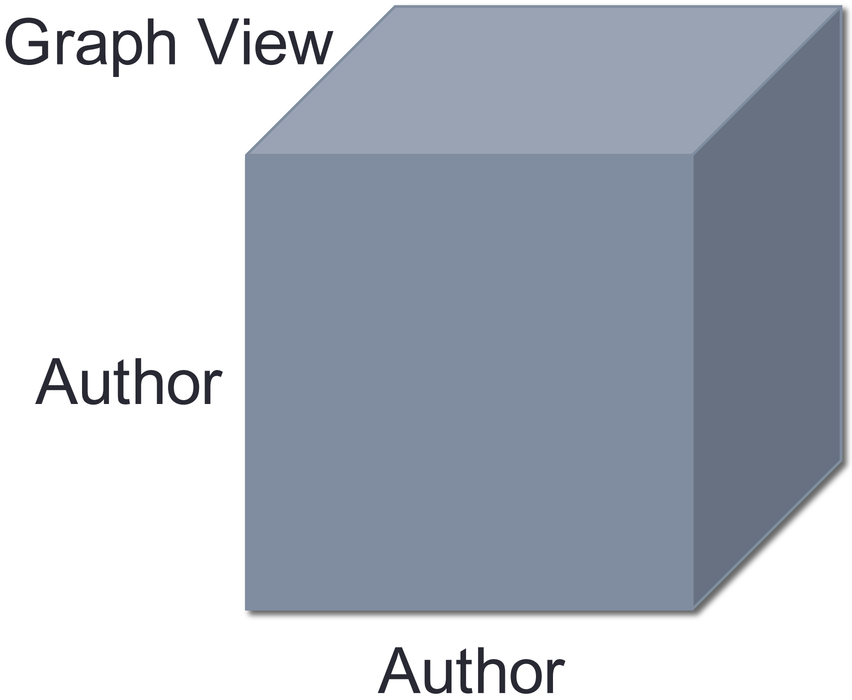
$$X \approx U \otimes V \otimes W$$

$$X_{i,j,k} \approx \sum_{r=1}^{\text{Rank}} U_{i,r} V_{j,r} W_{k,r}$$

Graph Clustering with Tensors

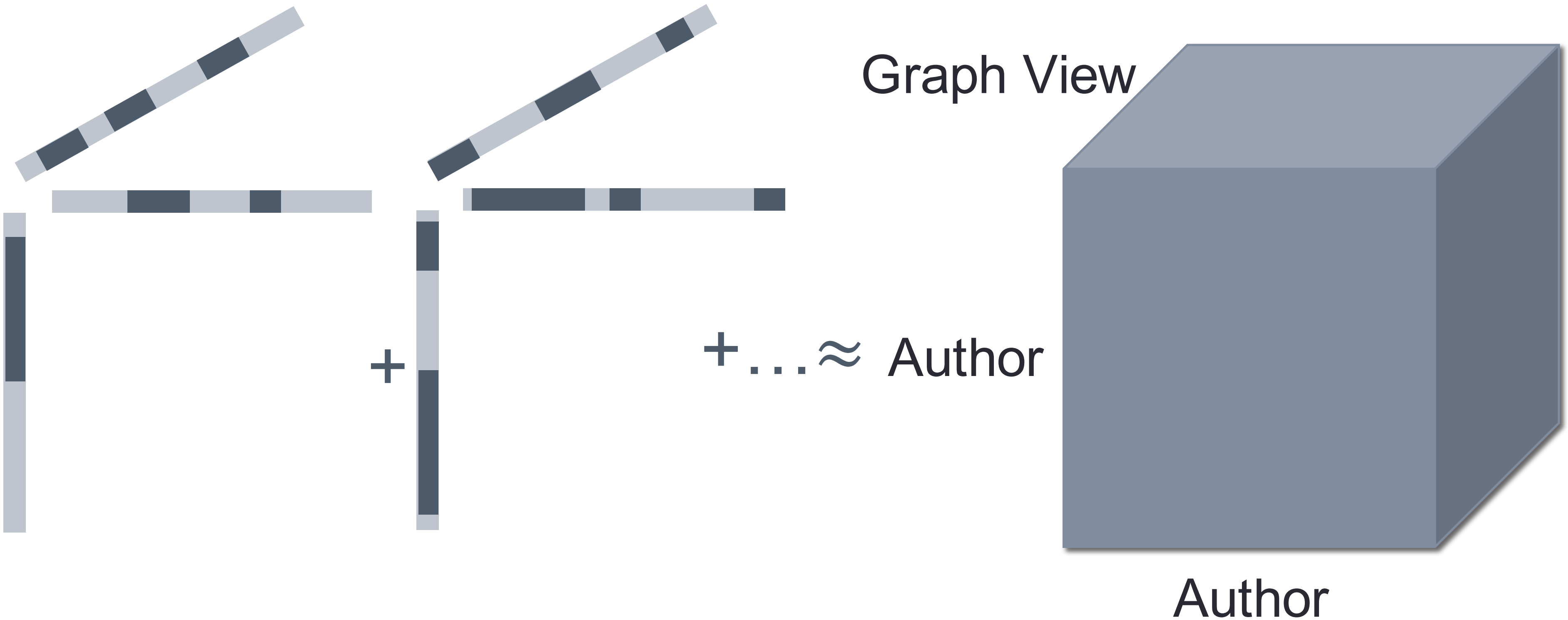
Multiple possible views of the DBLP network:

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

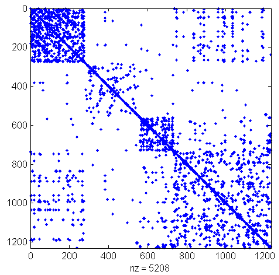


Sparse Tensor Factorization

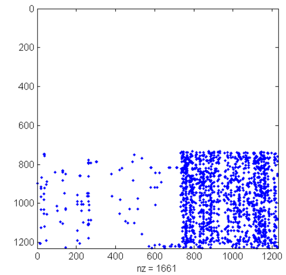
Do more Views of a Graph help? Community Detection and Clustering in Multi-Graphs
Evangelos E. Papalexakis, Leman Akoglu, Dino Ienco
FUSION 2013

Graph Clustering with Tensors

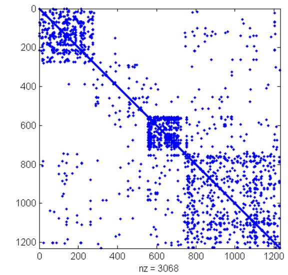
DBLP-1



(a) citation

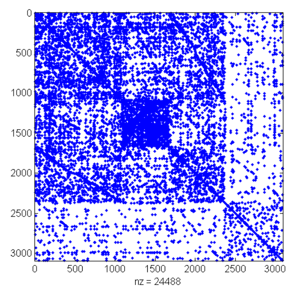


(b) co-auth.

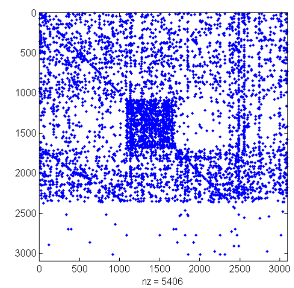


(c) co-term

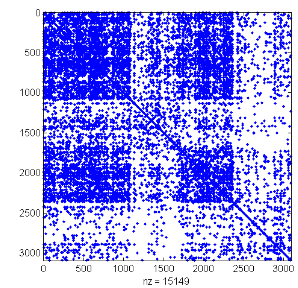
DBLP-2



(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

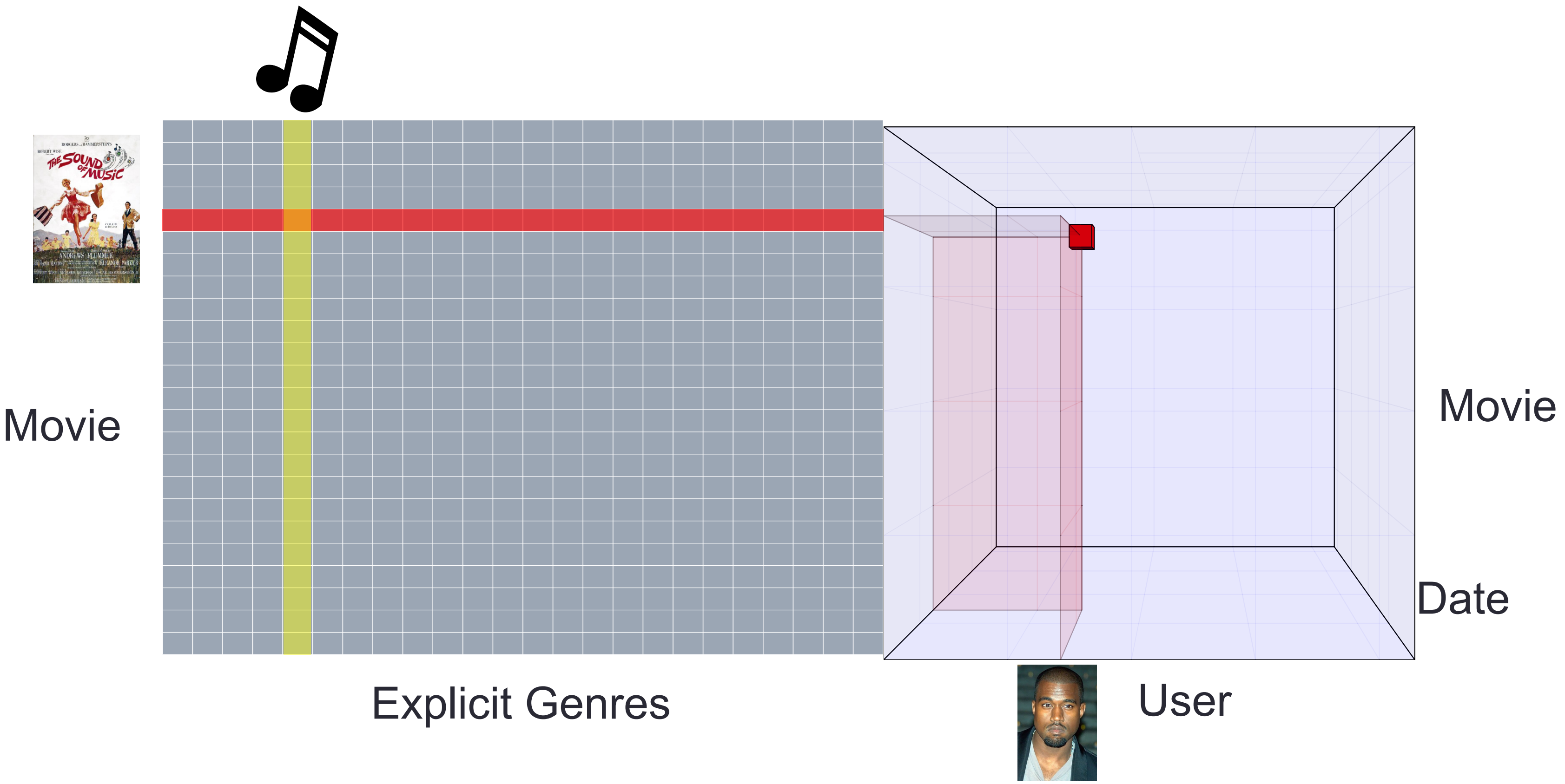
Graph Clustering with Tensors

Dataset	Baseline	GraphFuse
DBLP-1	0.12	**0.30**
DBLP-2	0.08	**0.12**

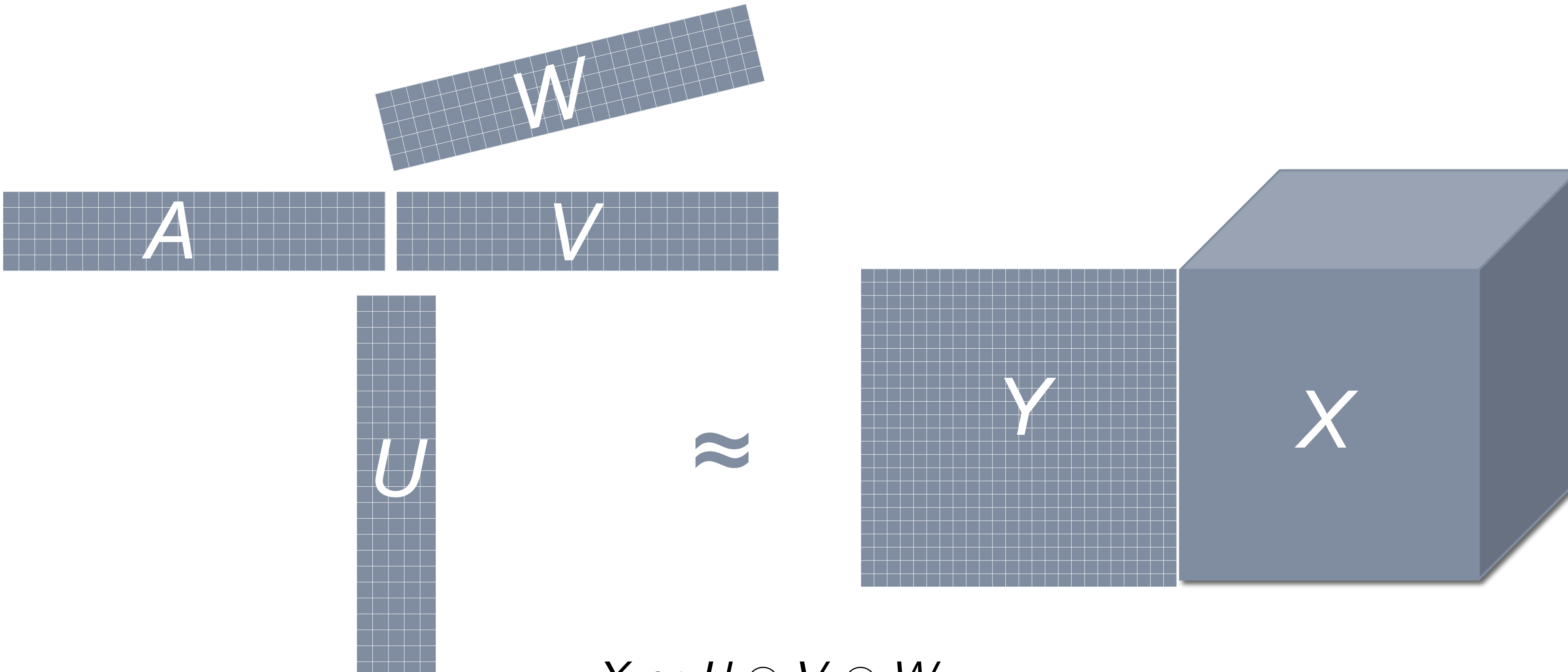
Modeling Accuracy

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

Coupled Matrix + Tensor Decomposition



Coupled Matrix + Tensor Decomposition

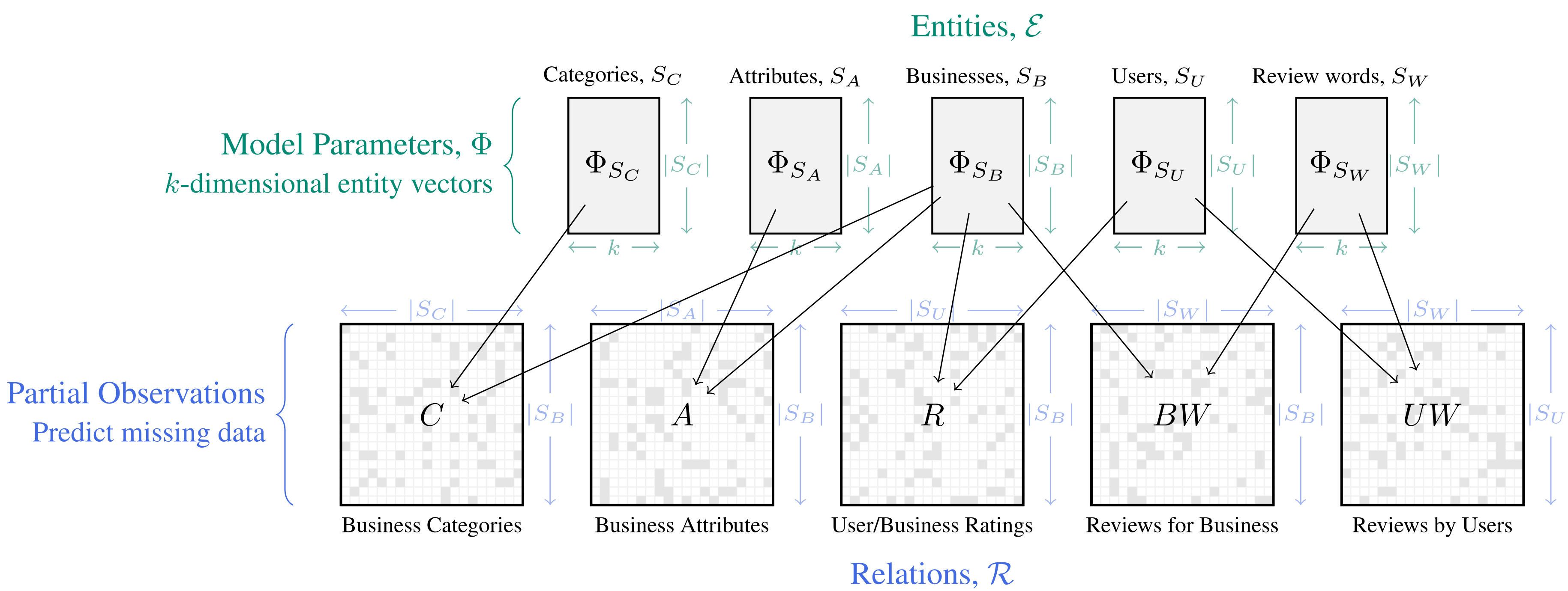


$$X \approx U \otimes V \otimes W$$

$$Y \approx UA^\top$$

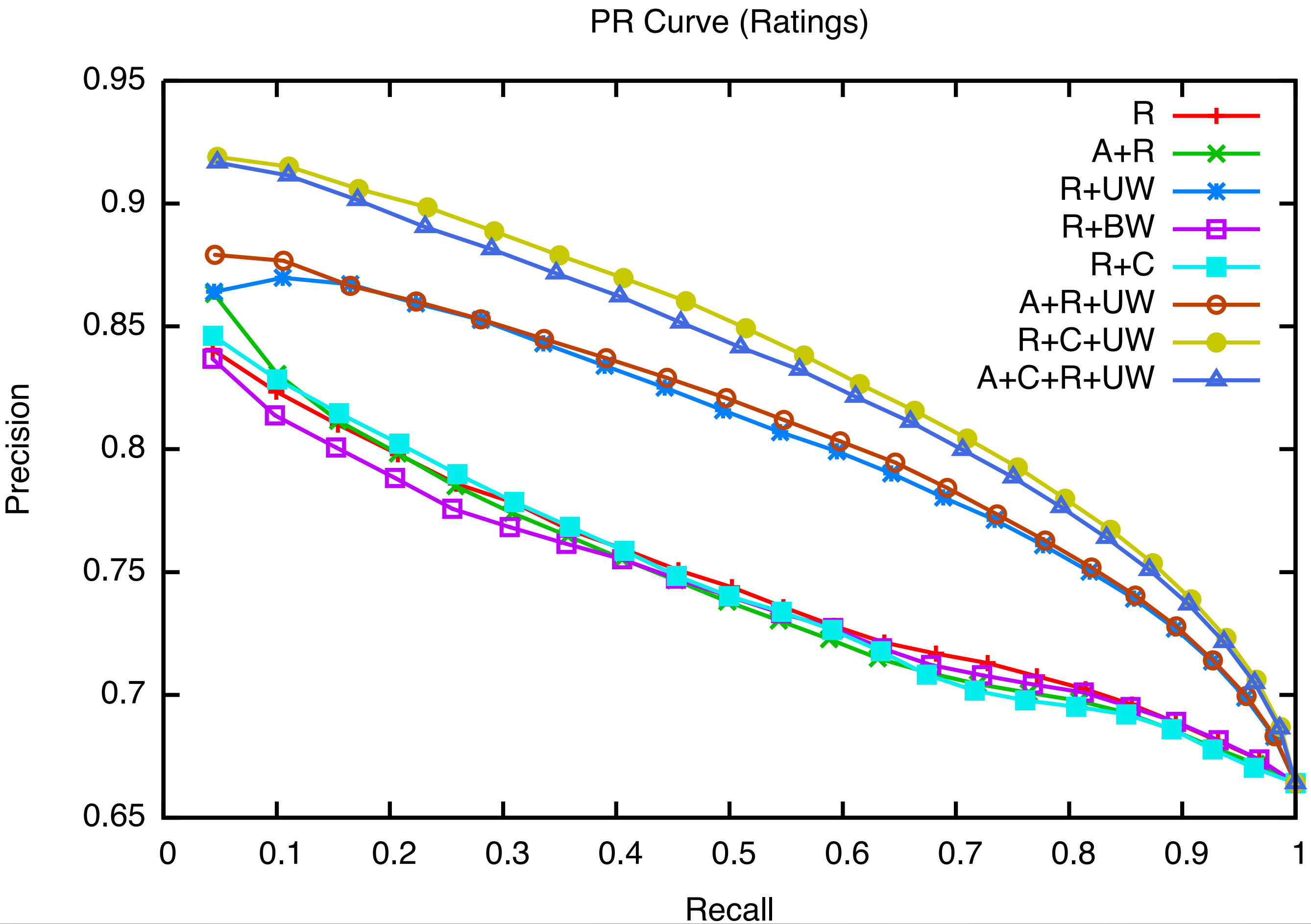
$$\min_{U,V,W,A} \|X - U \otimes V \otimes W\|_F^2 + \|Y - UV^\top\|_F^2$$

Joint Factorization



Collective Factorization for Relational Data:
An Evaluation on the Yelp Datasets
Nitish Gupta, Sameer Singh


Joint Factorization



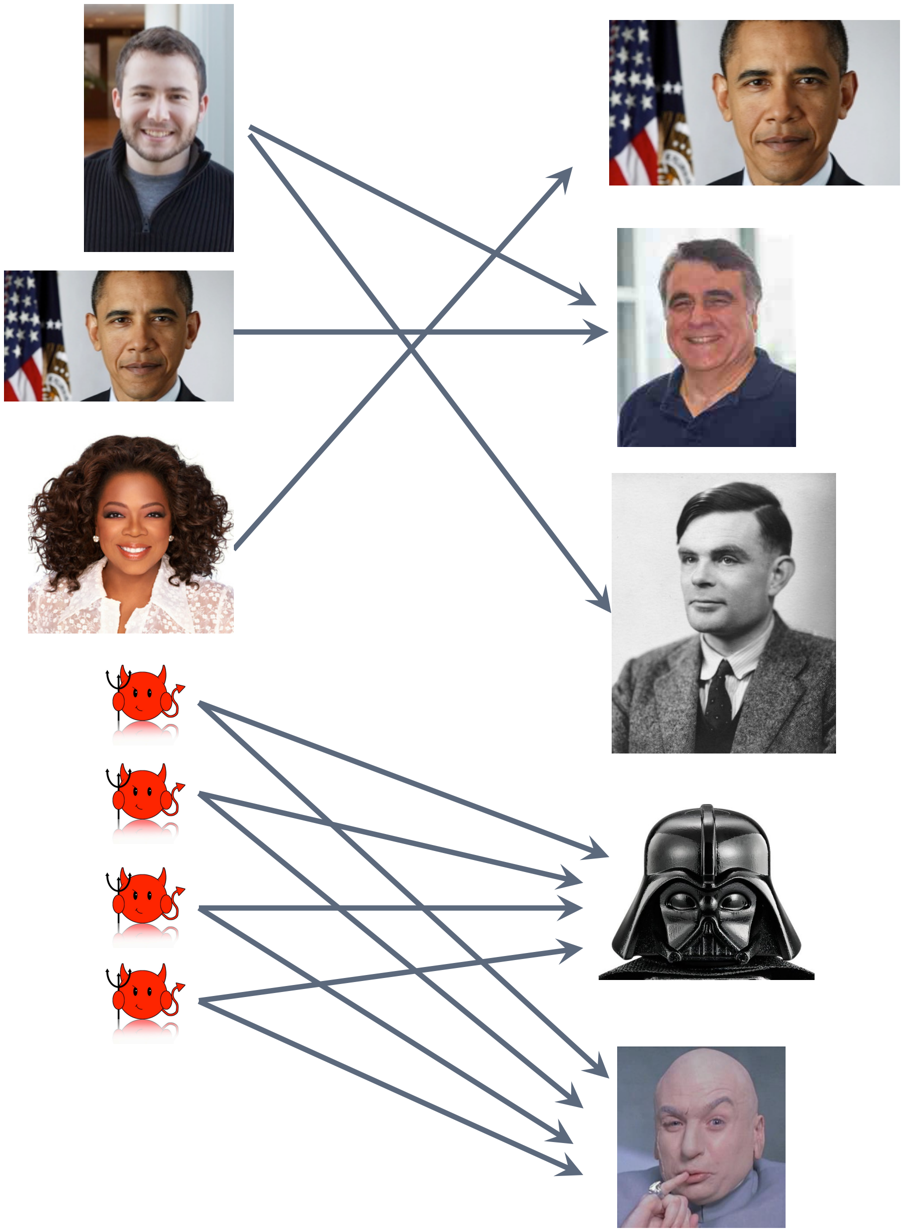
Most valuable:

1. Ratings
2. Review text
3. Business Categories

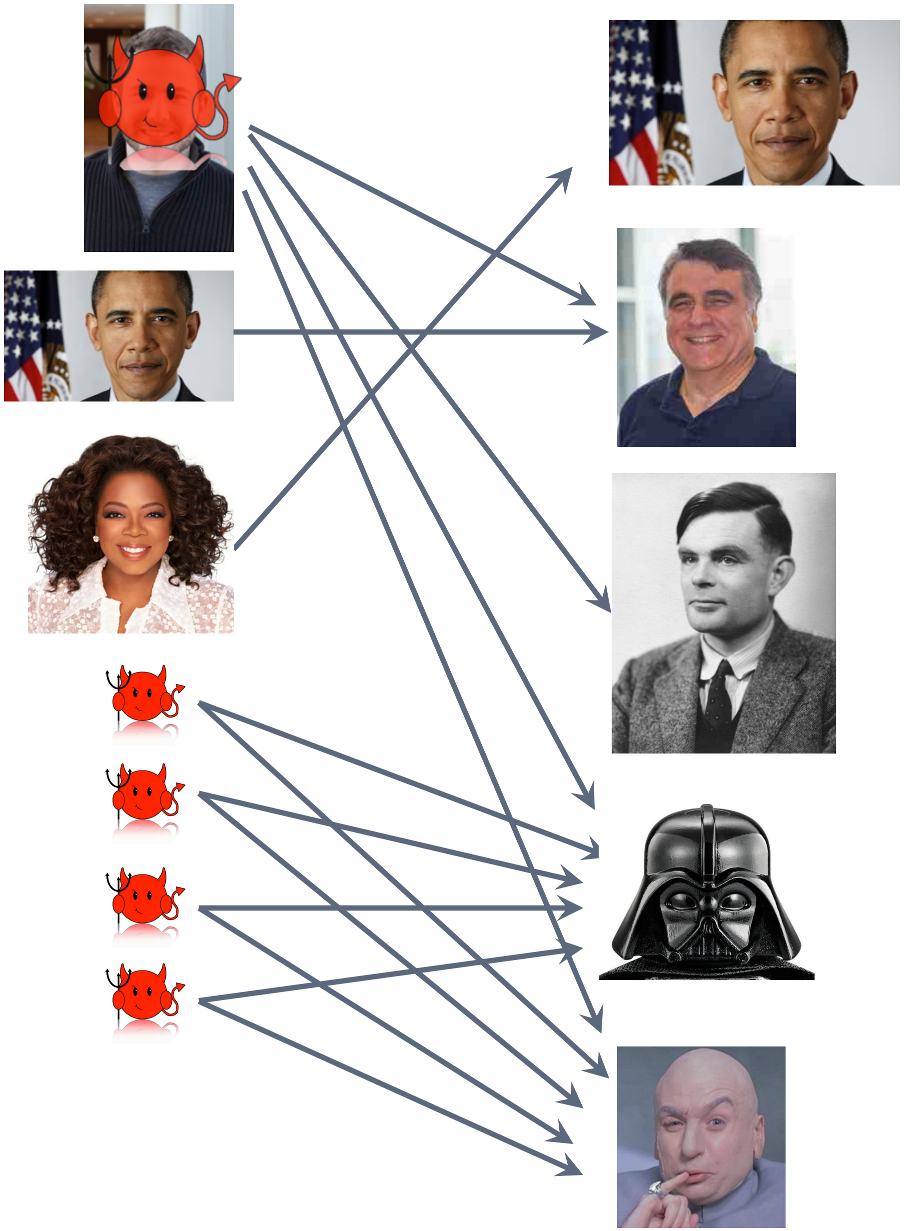
Collective Factorization for Relational Data: An Evaluation on the Yelp Datasets
Nitish Gupta, Sameer Singh

1. Subgraph Analysis
2. Propagation Methods
3. **Latent Factor Models**
   a) Background
   b) Normal Behavior
   c) Abnormal Behavior

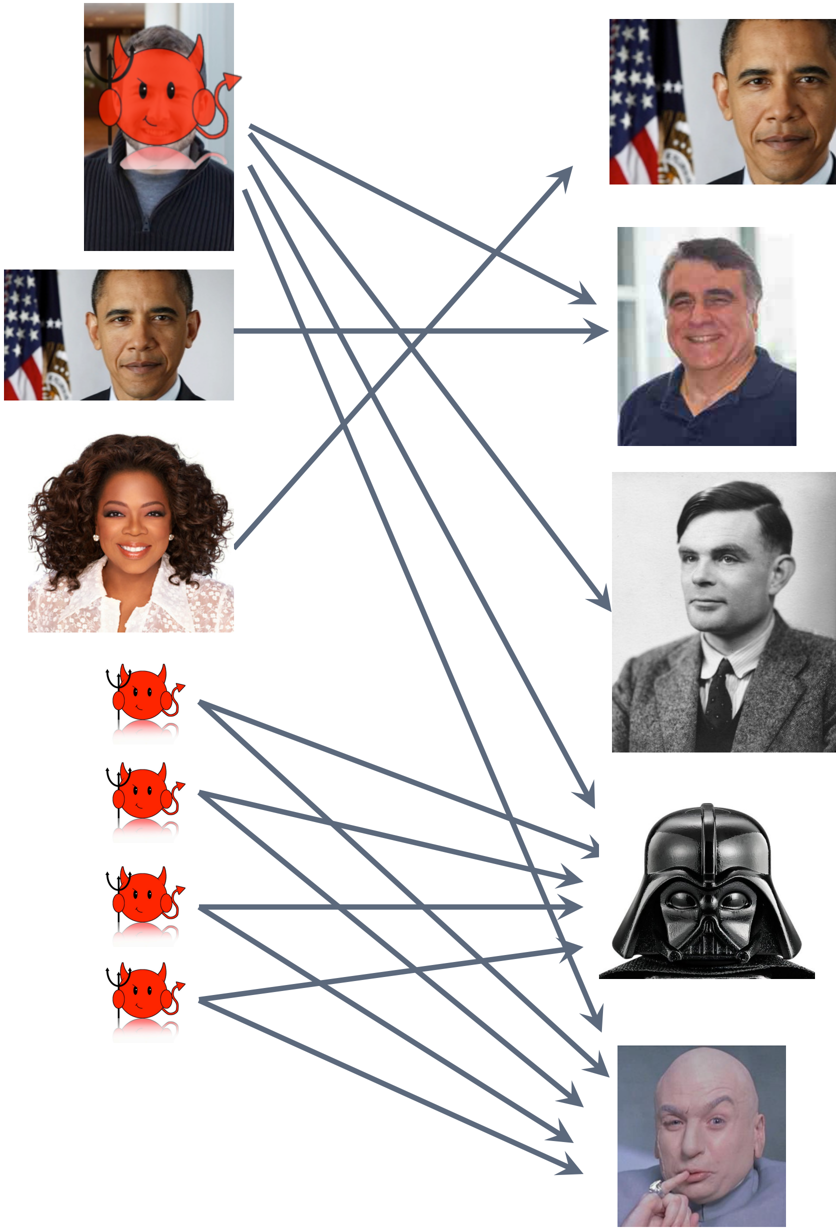
Fraud Detection

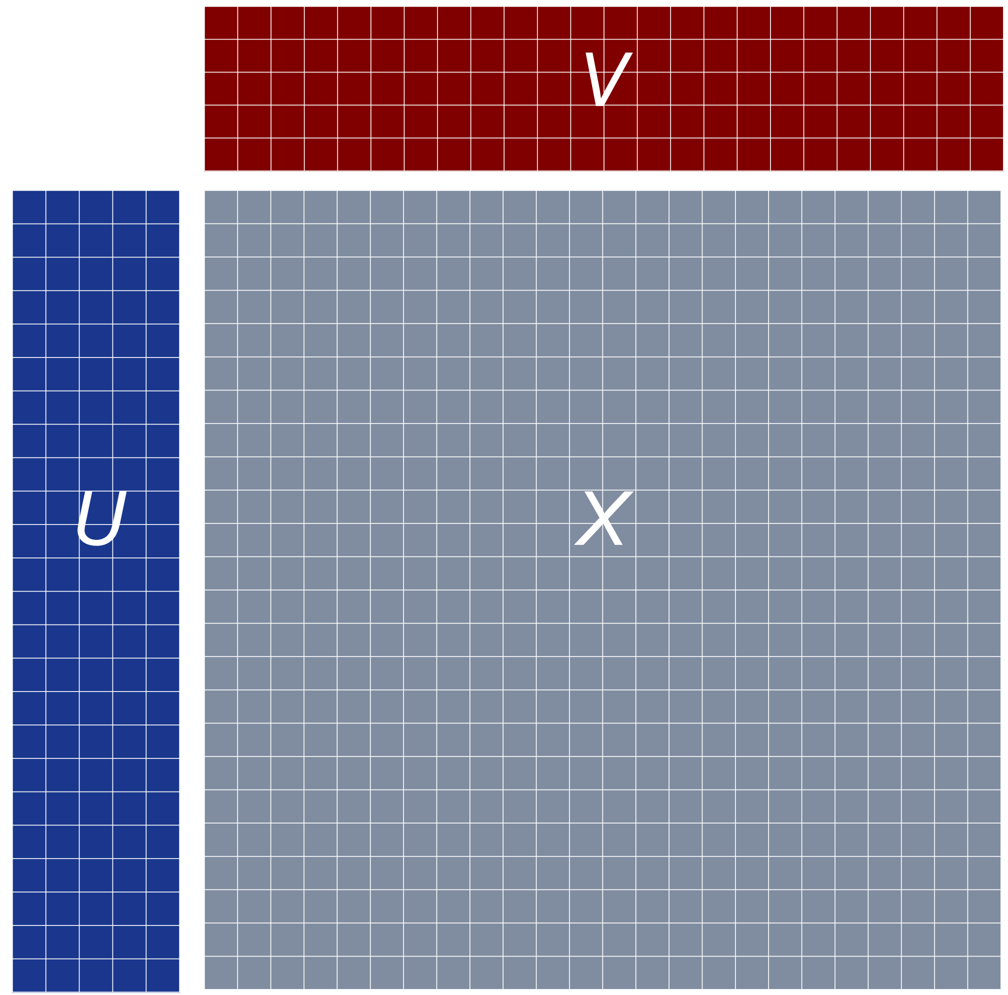


Fraud Detection

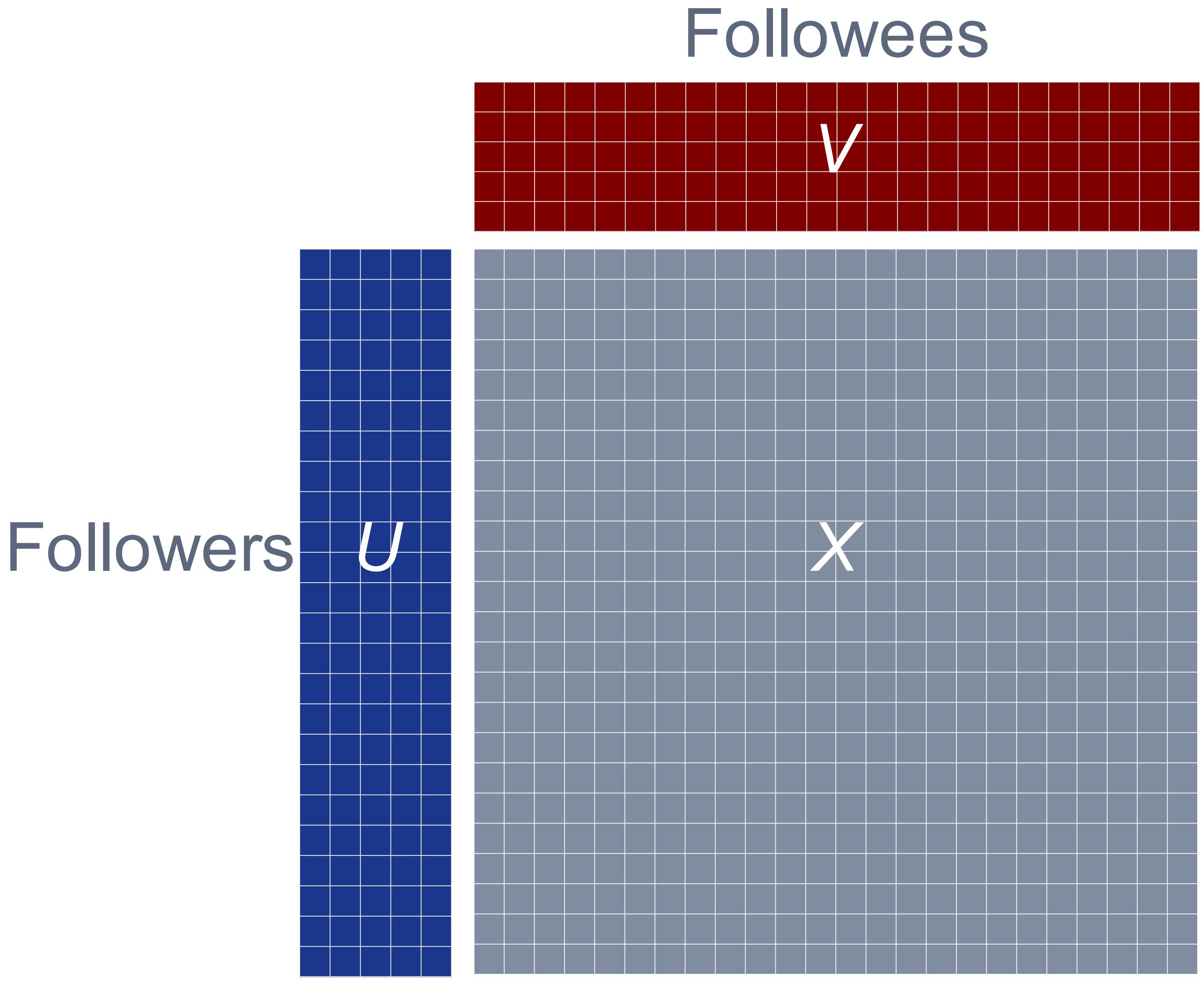


Fraud within a factorization

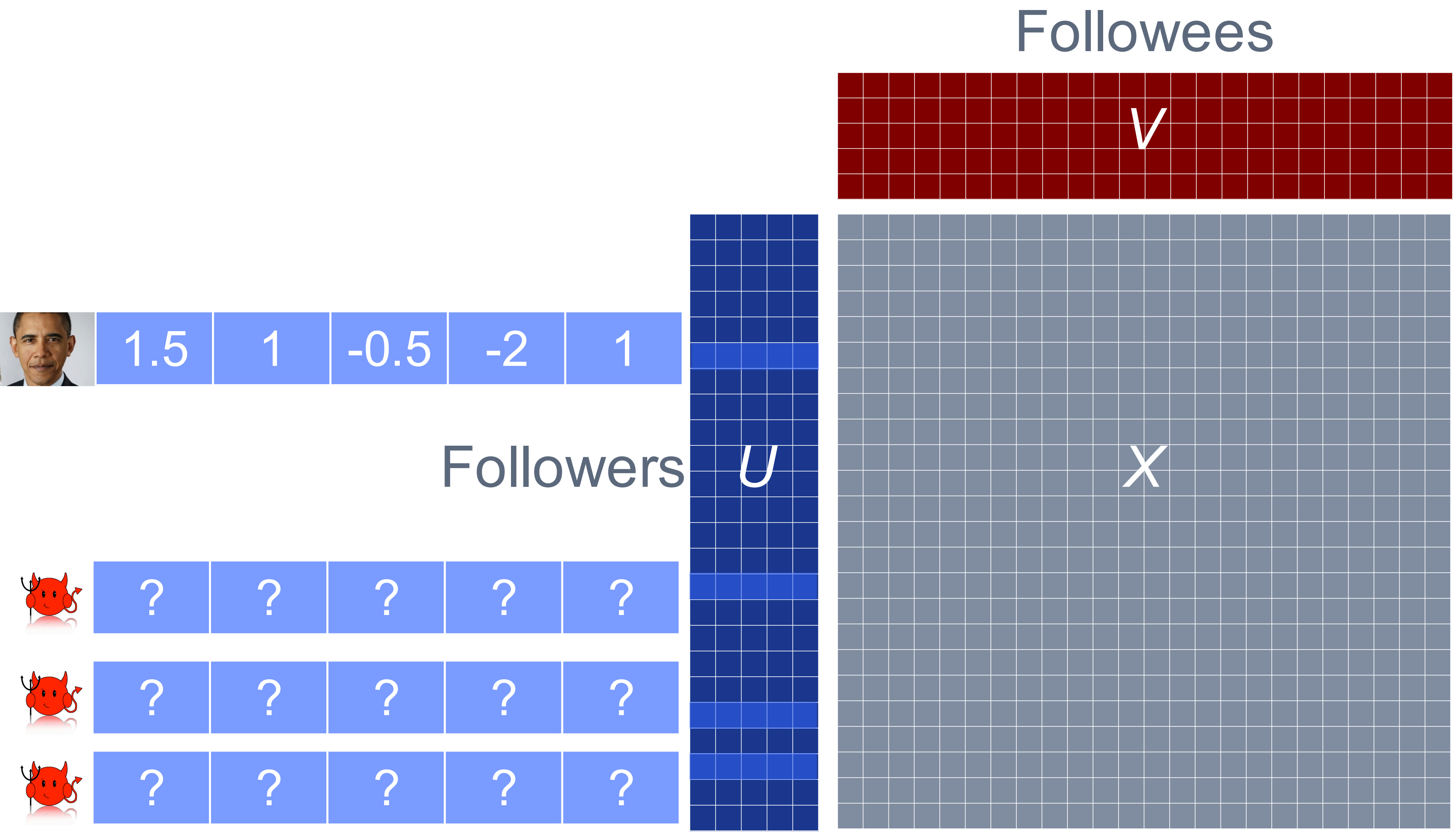




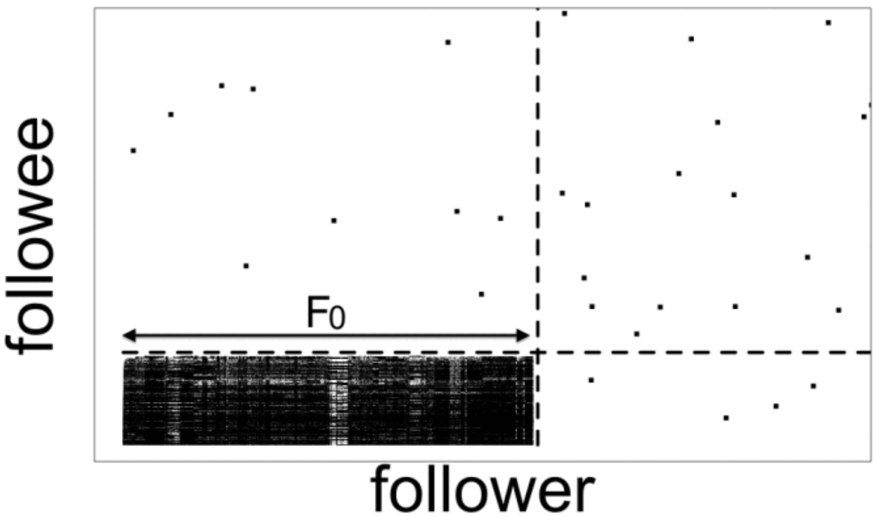
Fraud within a factorization



Fraud within a factorization

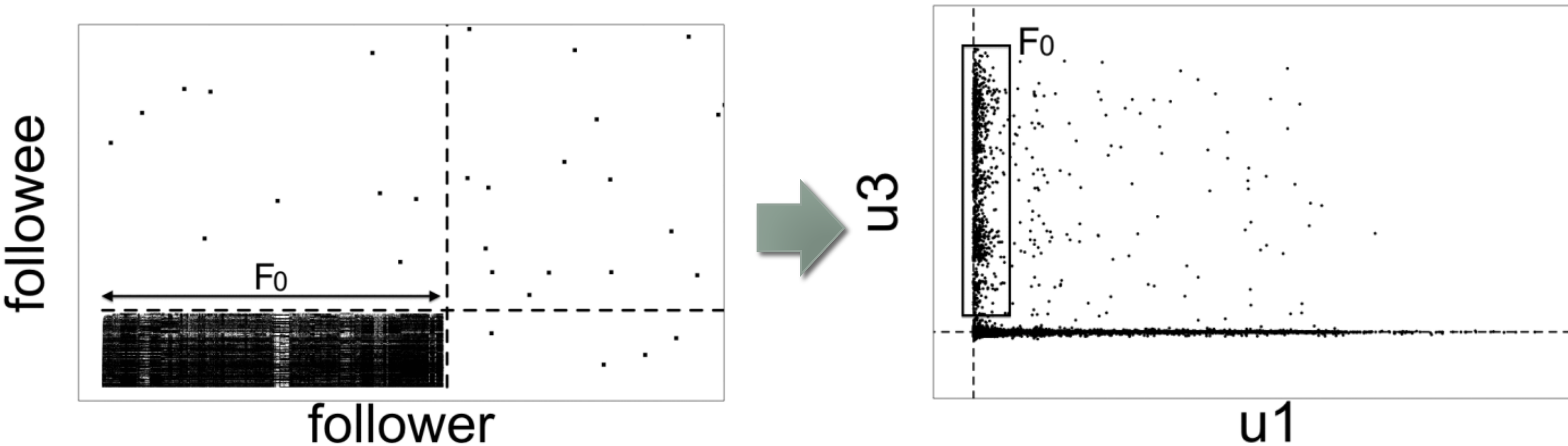


Fraud within a factorization



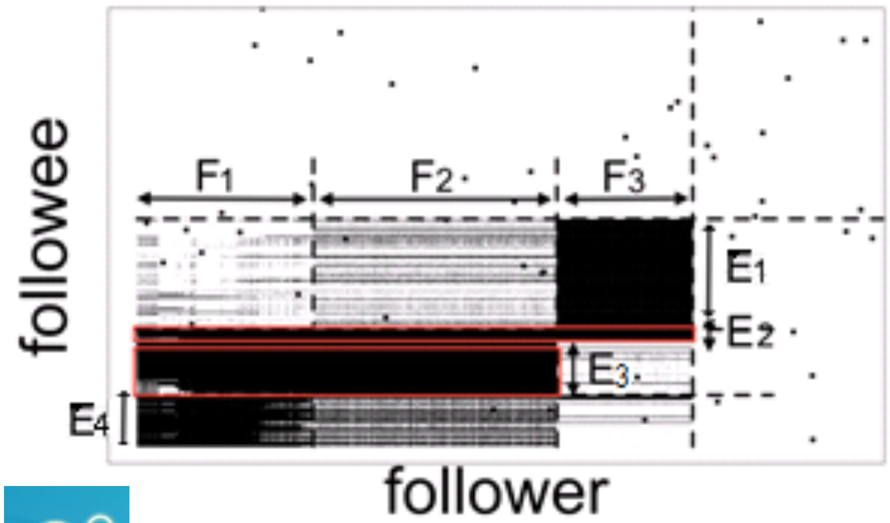
EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs
B. Aditya Prakash, Ashwin Sridharan, Mukund Seshadri,
Sridhar Machiraju, Christos Faloutsos
PAKDD, 2010

Fraud within a factorization

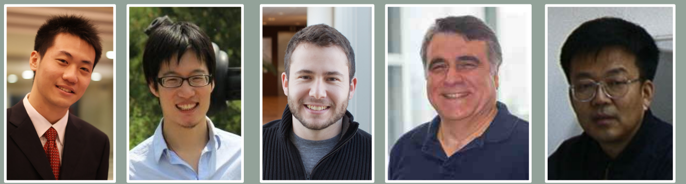


EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs
B. Aditya Prakash, Ashwin Sridharan, Mukund Seshadri,
Sridhar Machiraju, Christos Faloutsos
PAKDD, 2010

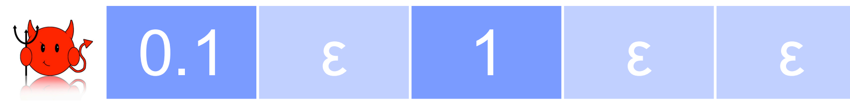
Fraud within a factorization

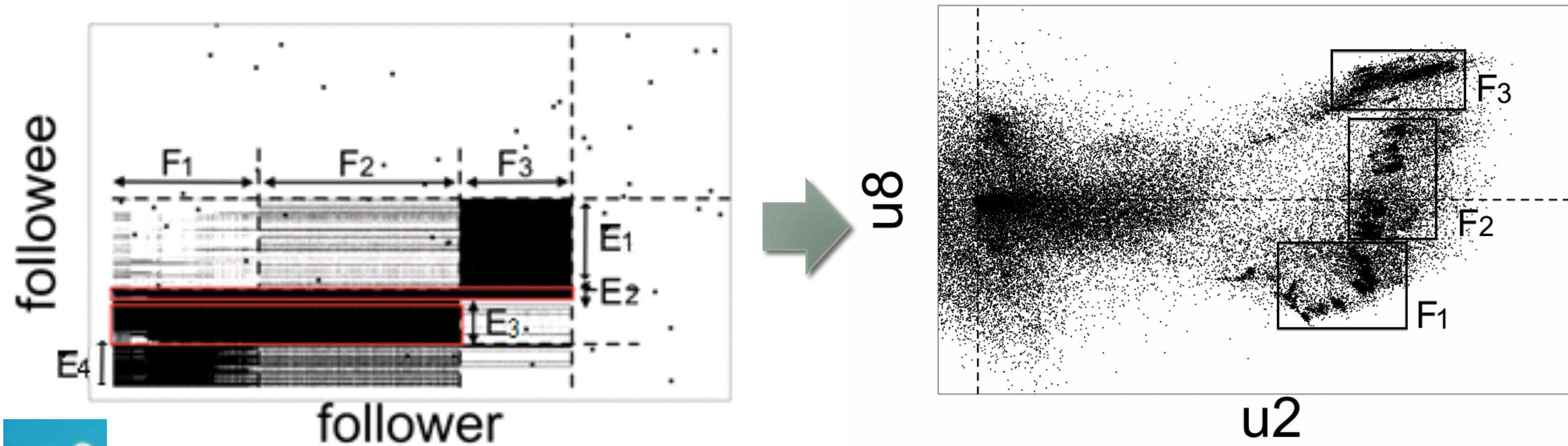


Inferring Strange Behavior from Connectivity Pattern in Social Networks
Meng Jiang, Peng Cui, Alex Beutel,
Christos Faloutsos, Shiqiang Yang.
PAKDD, 2014



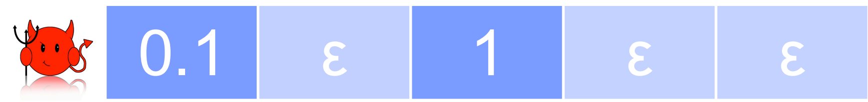
Fraud within a factorization

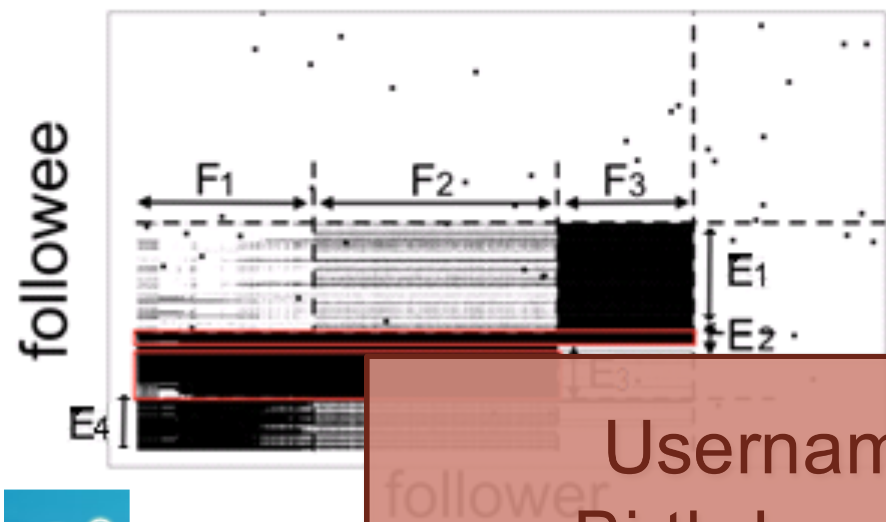


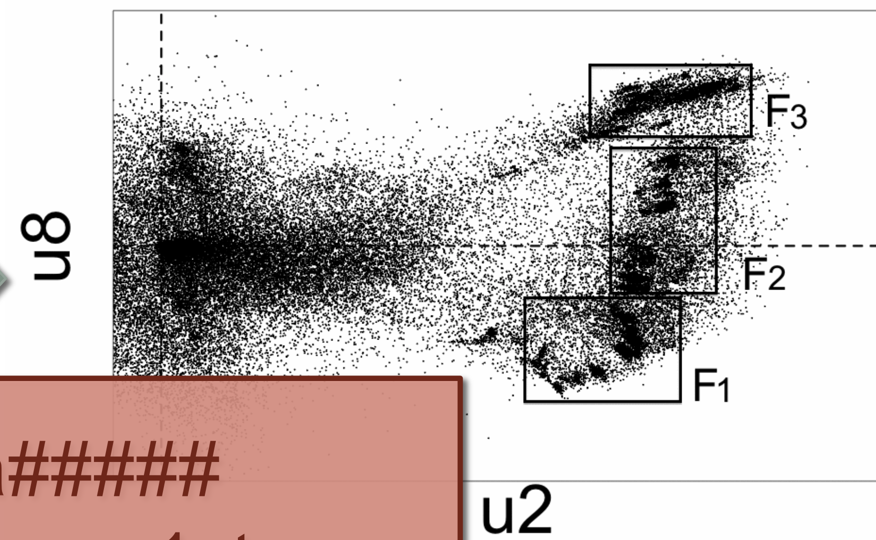


Inferring Strange Behavior from Connectivity Pattern in Social Networks
Meng Jiang, Peng Cui, Alex Beutel,
Christos Faloutsos, Shiqiang Yang.
PAKDD, 2014

Fraud within a factorization



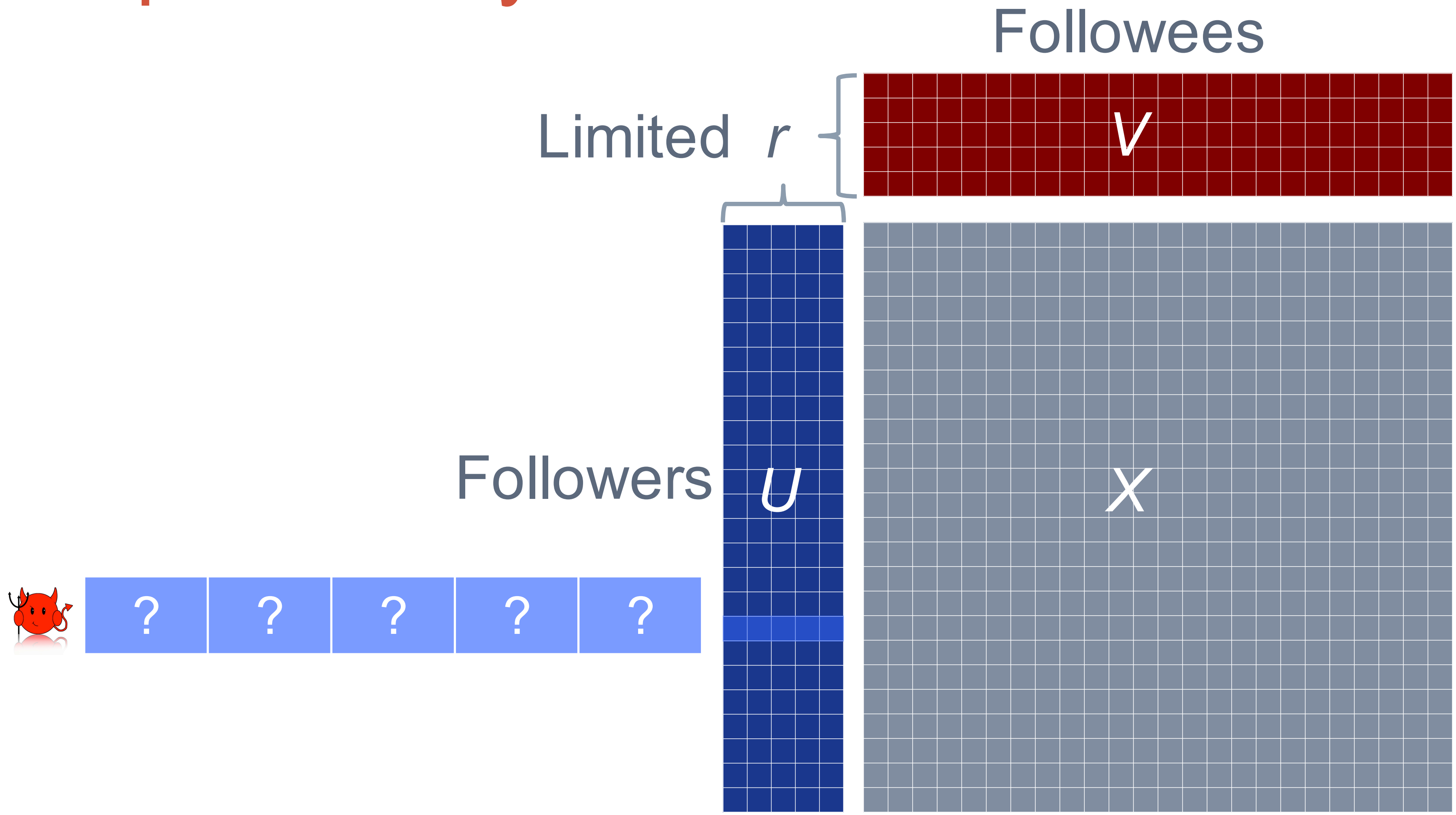




Username: a######
Birthday: January 1st

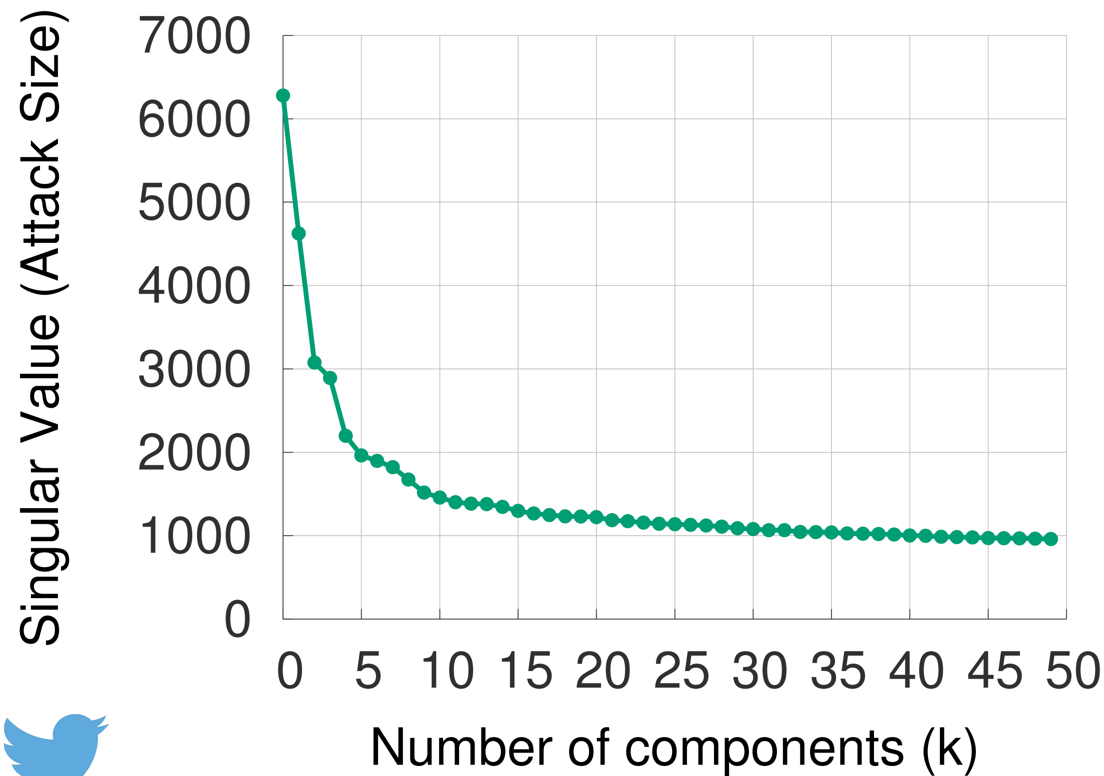
Inferring Strange Behavior from Connectivity Pattern in Social Networks
Meng Jiang, Peng Cui, Alex Beutel,
Christos Faloutsos, Shiqiang Yang.
PAKDD, 2014

Complementary Fraud Detection

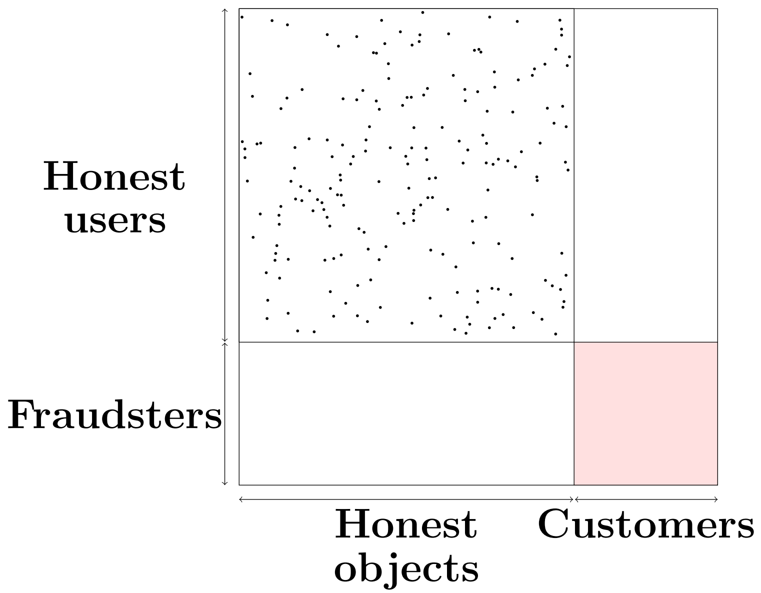


Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher,
Christos Faloutsos
ICDM, 2014.

Complementary Fraud Detection

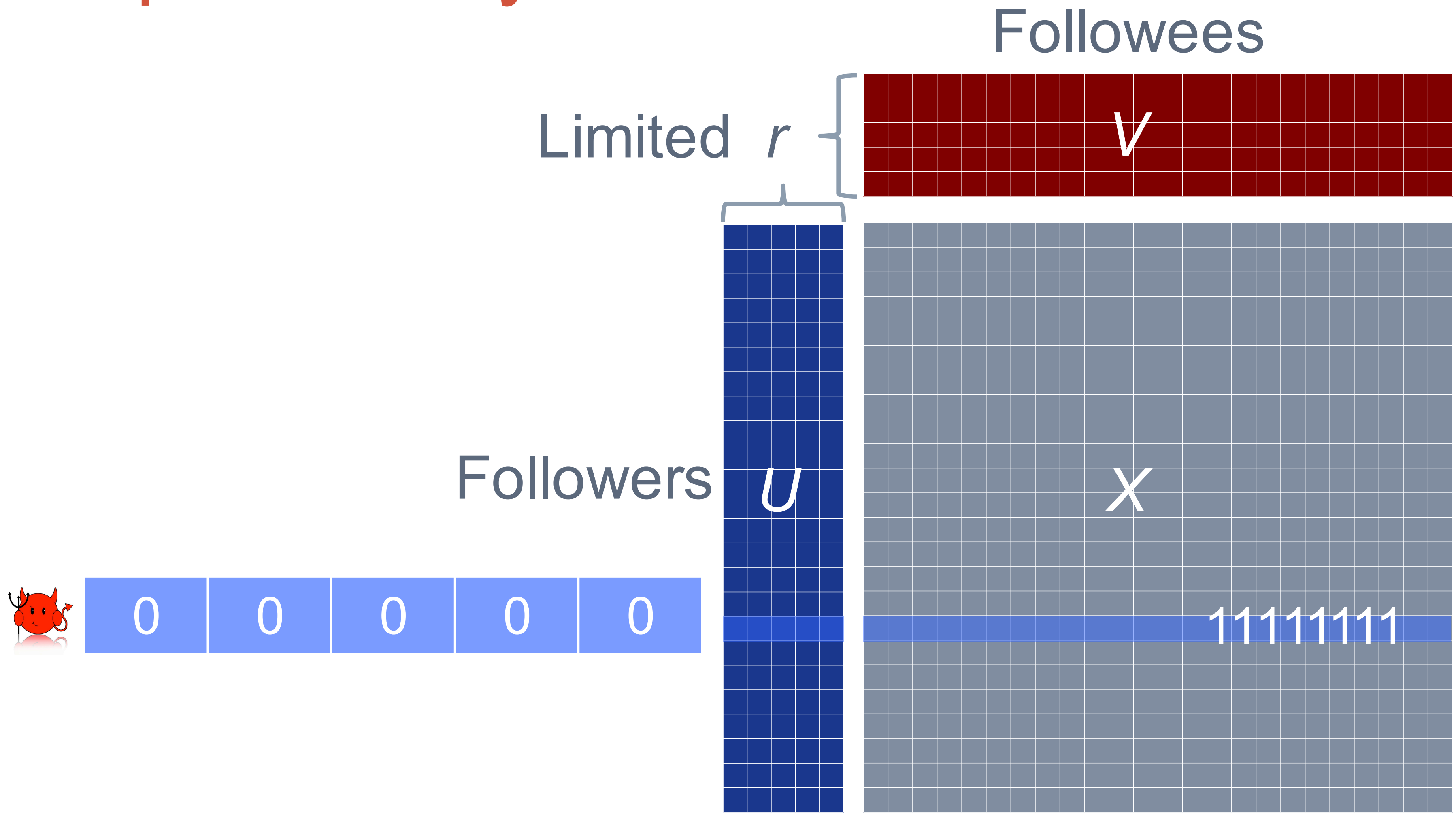


960 fraudsters safely following 960 customers



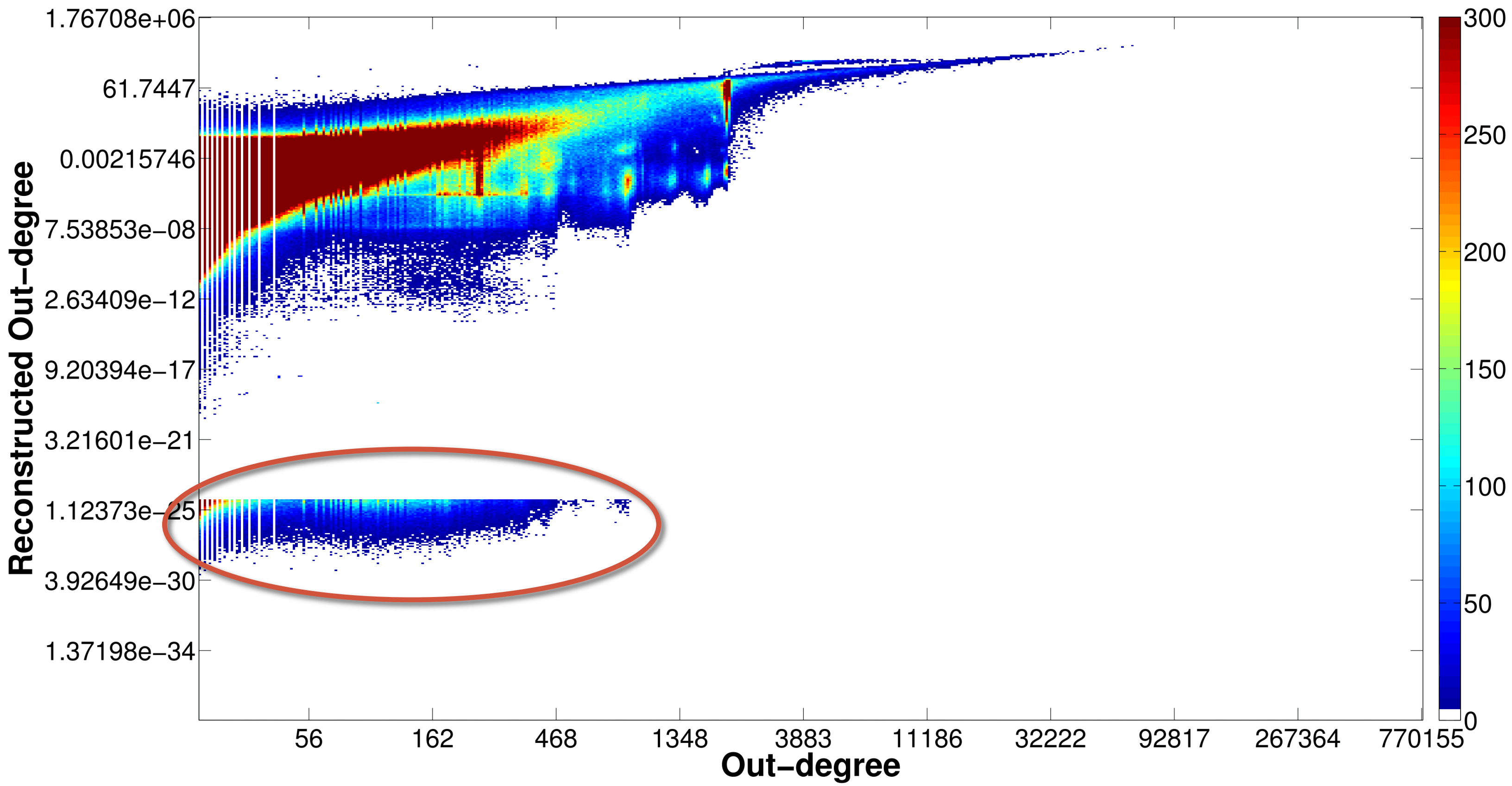
Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher,
Christos Faloutsos
ICDM, 2014.

Complementary Fraud Detection



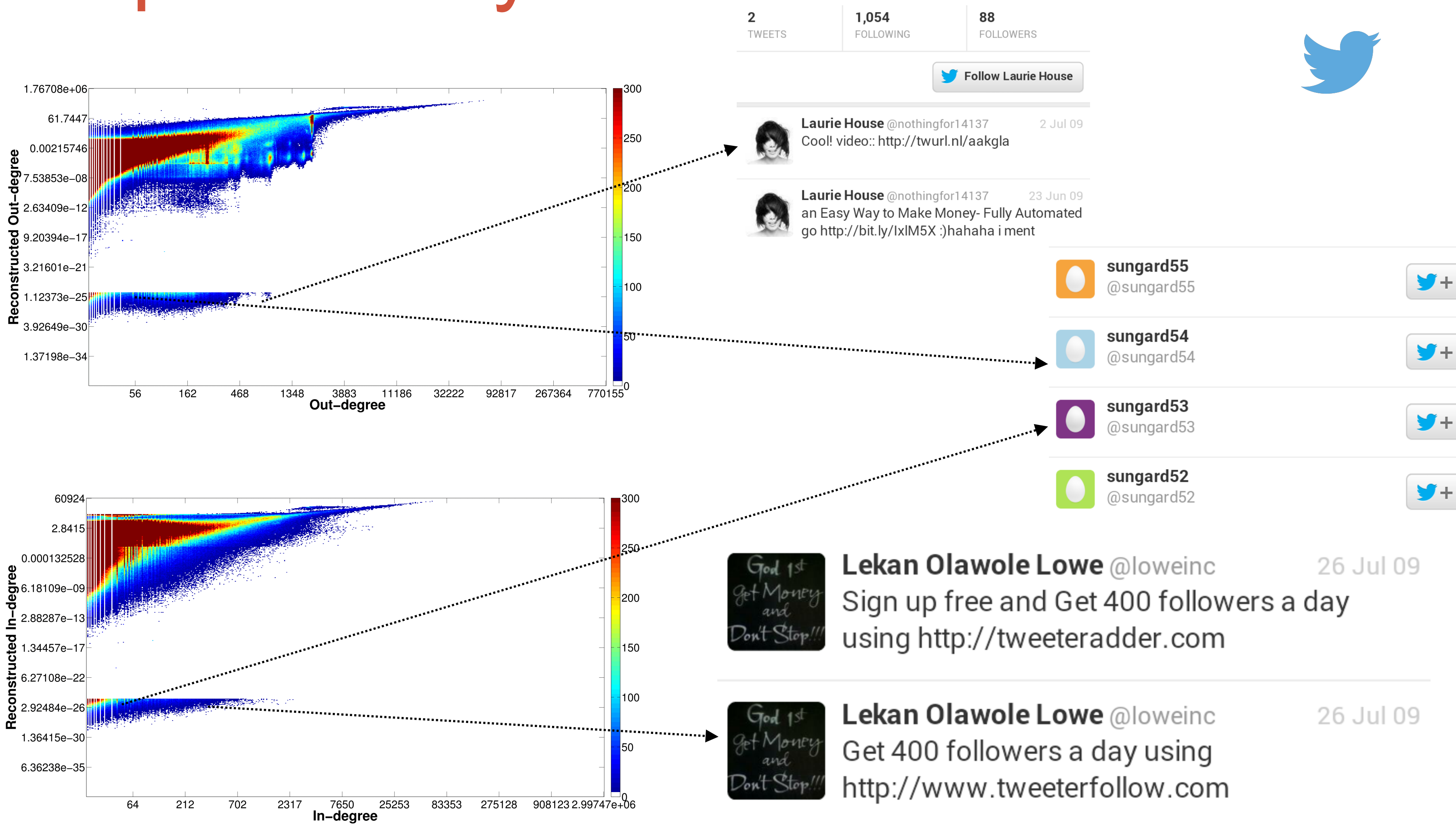
Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher,
Christos Faloutsos
ICDM, 2014.

Complementary Fraud Detection



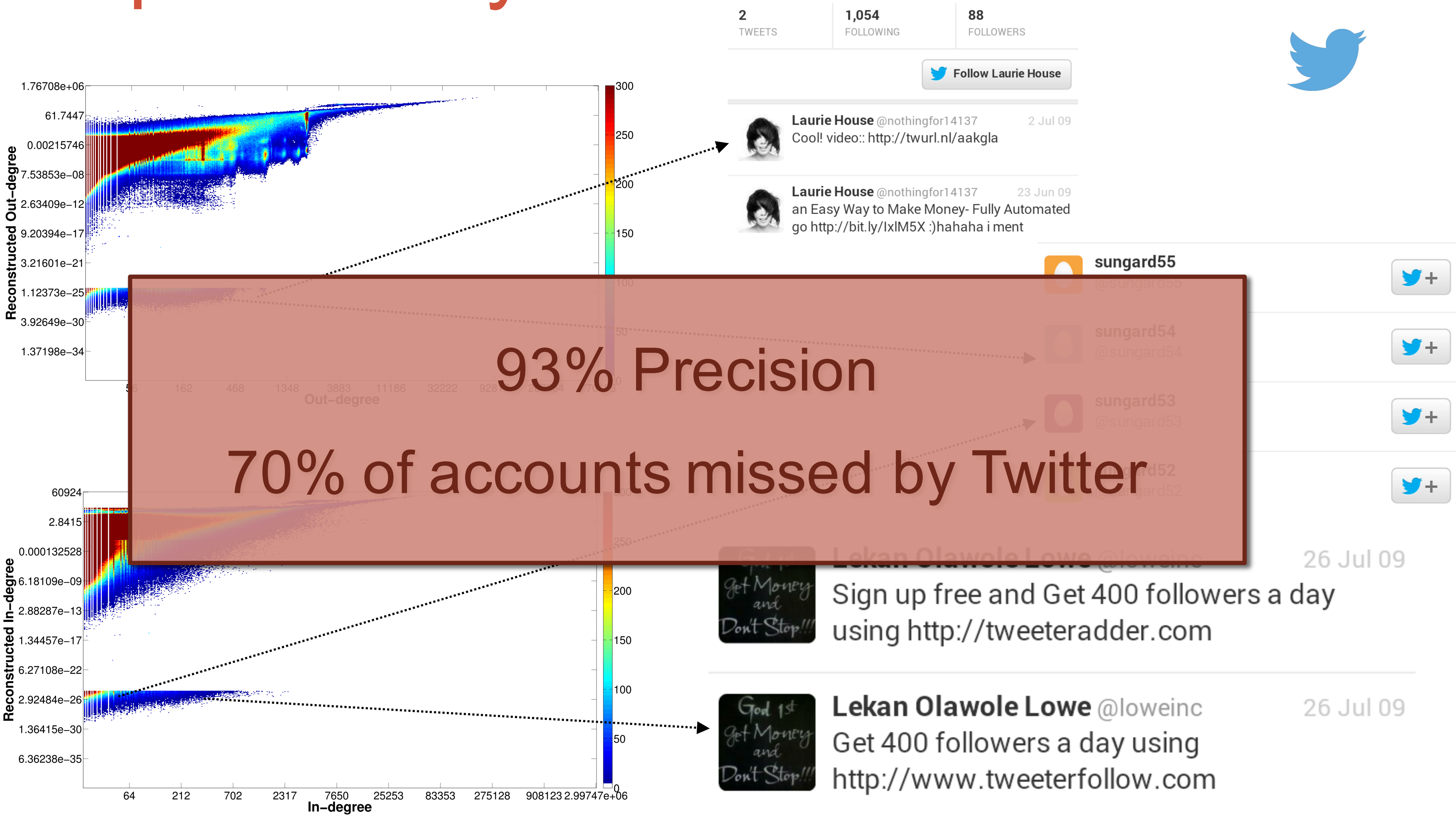
Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective
Neil Shah, Alex Beutel, Brian Gallagher,
Christos Faloutsos
ICDM, 2014.

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

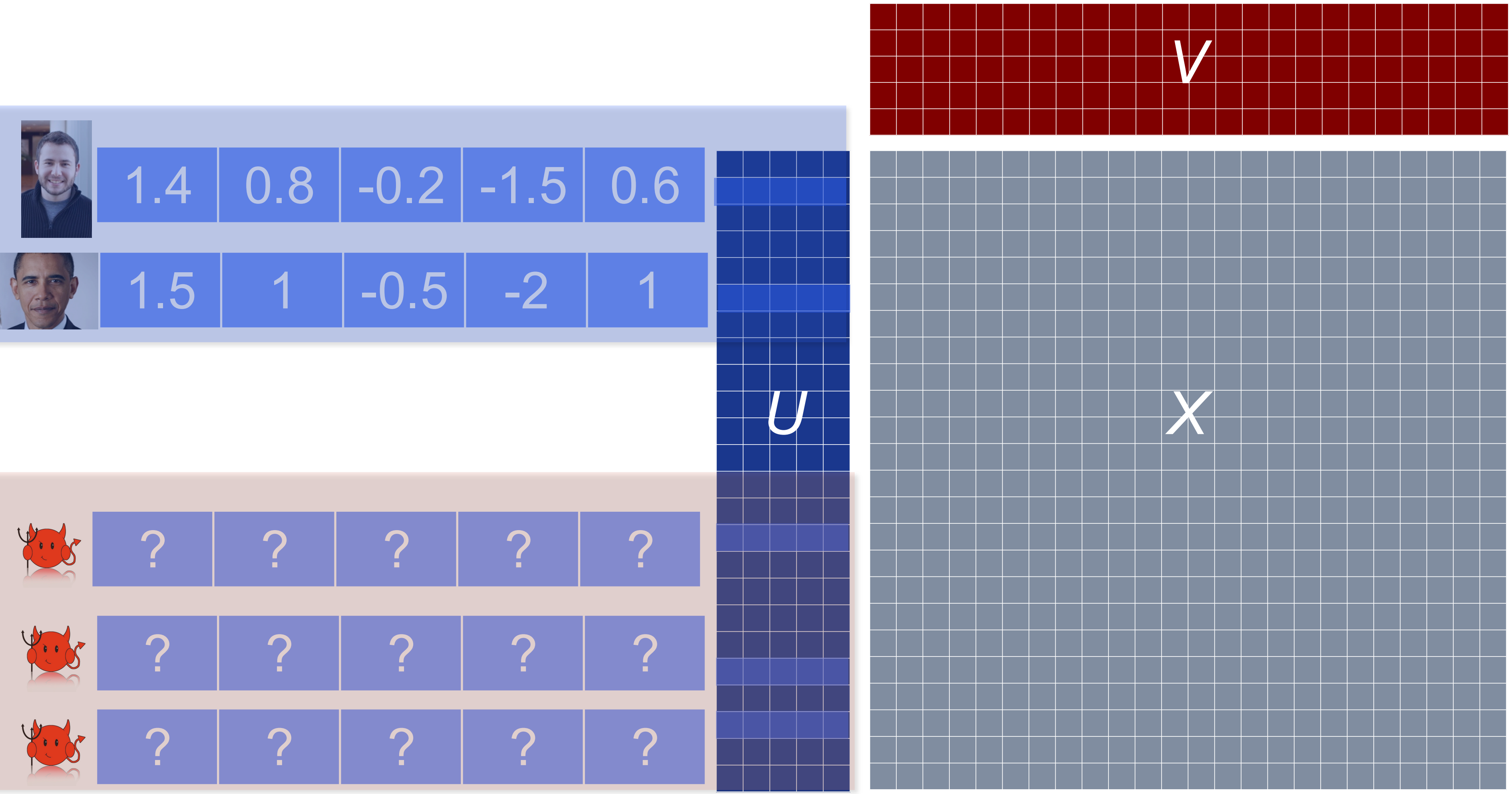
70% of accounts missed by Twitter

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

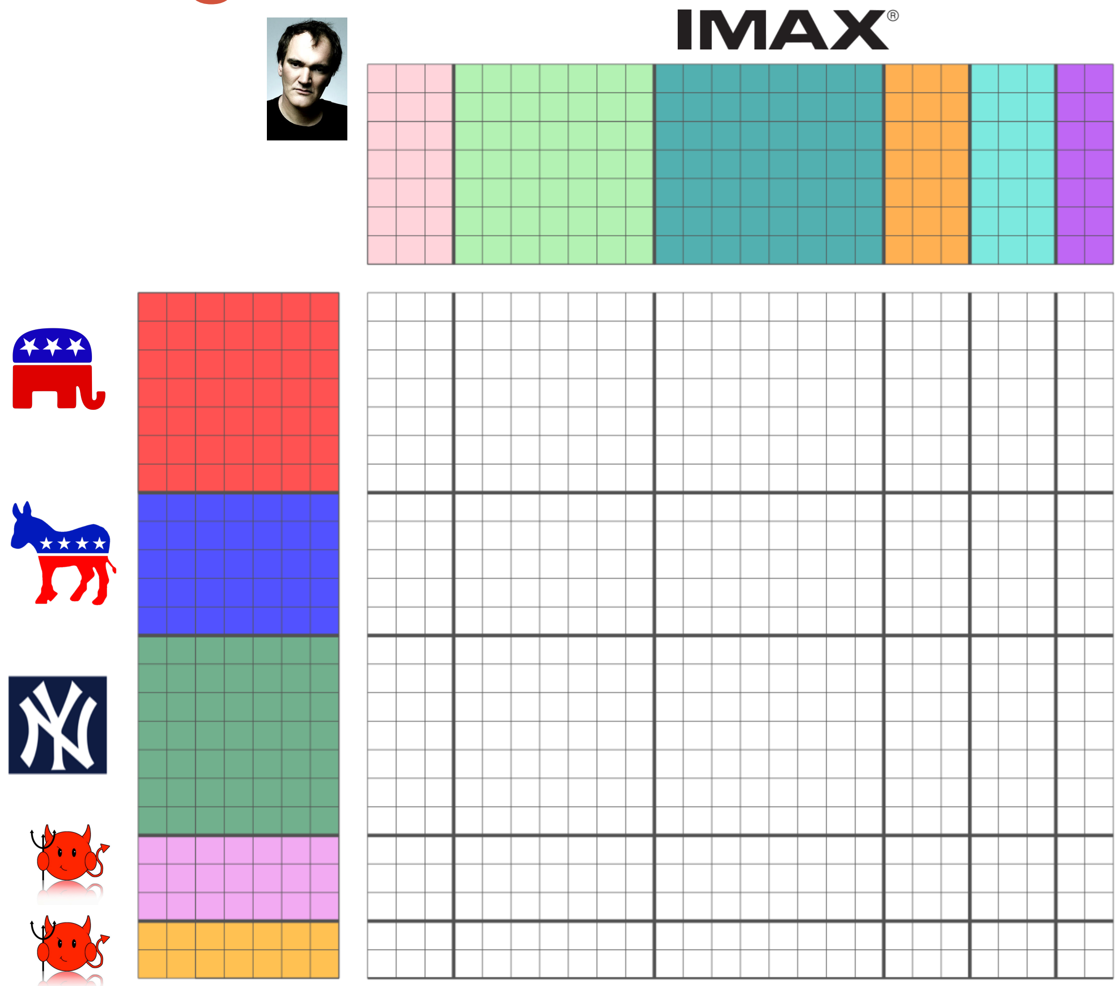
Practitioner's Guide

Method	Graph Type	Node Attributes	Edge Attributes	Seed Labels
EigenSpokes	Directed+			
Get-the-Scoop	Directed+			
fBox	Directed+			
CoBaFi	Bipartite+		✓	
CDOutliers	Undirected	✓		

Detecting Fraud within Recommendation

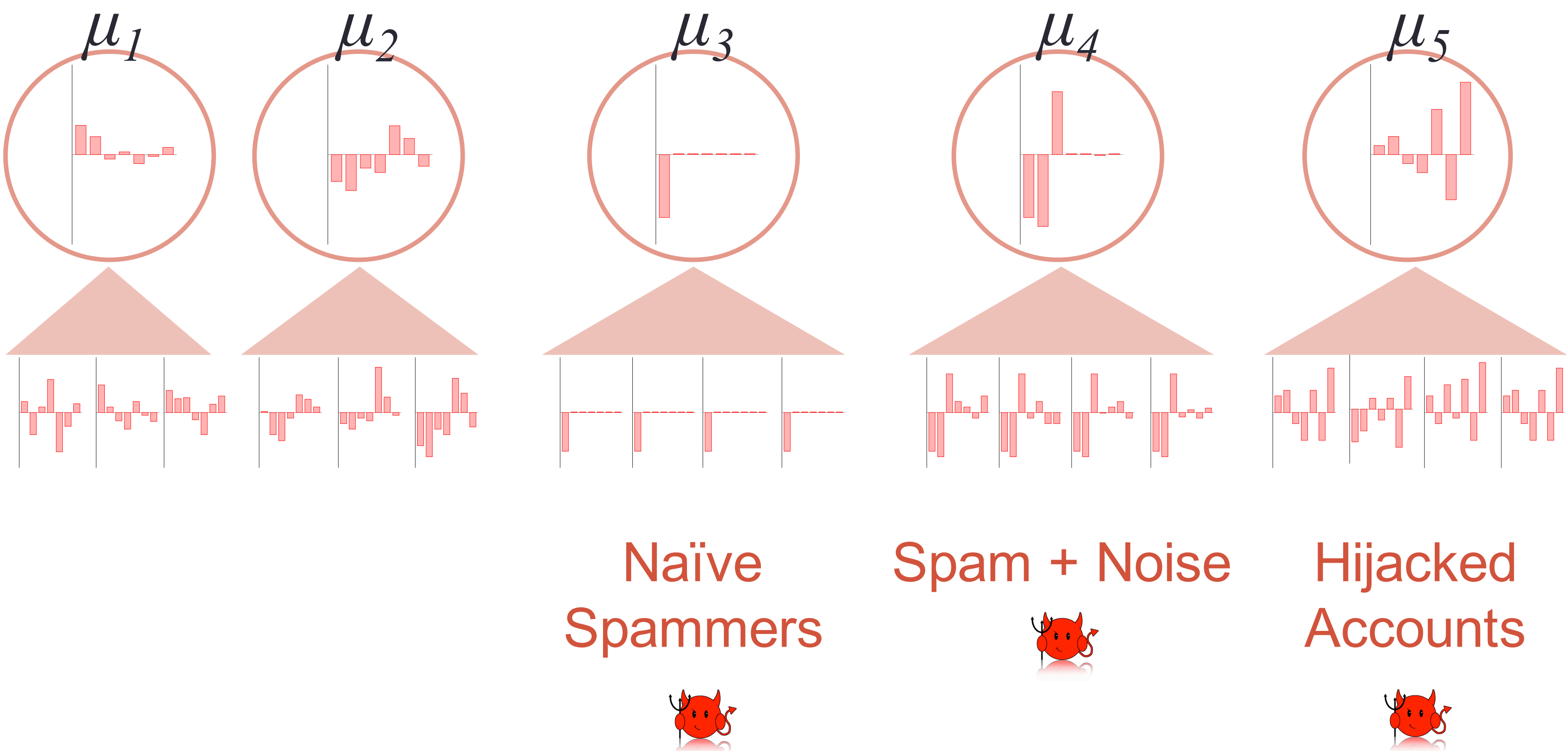


Detecting Fraud within Recommendation



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

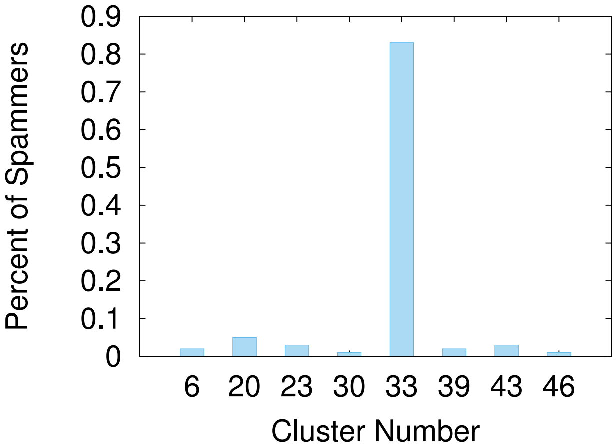
Clustering Fraudsters



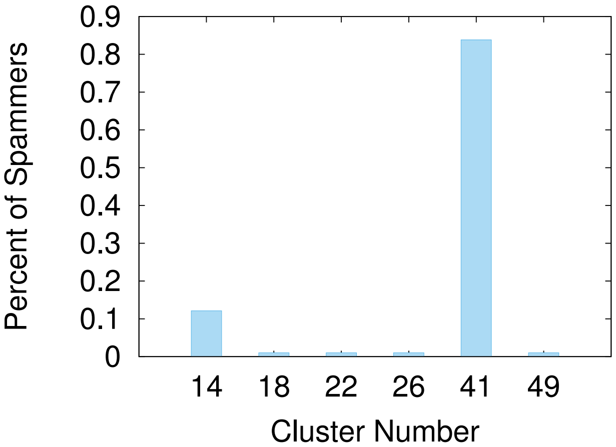
CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014

Clustered Fraudsters

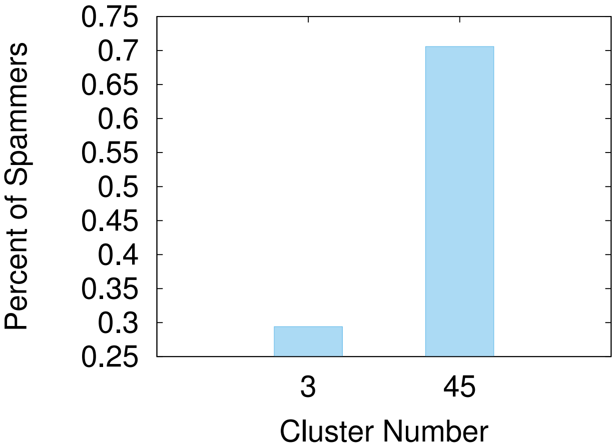
Clustered naïve spammers



Clustered hijacked accounts



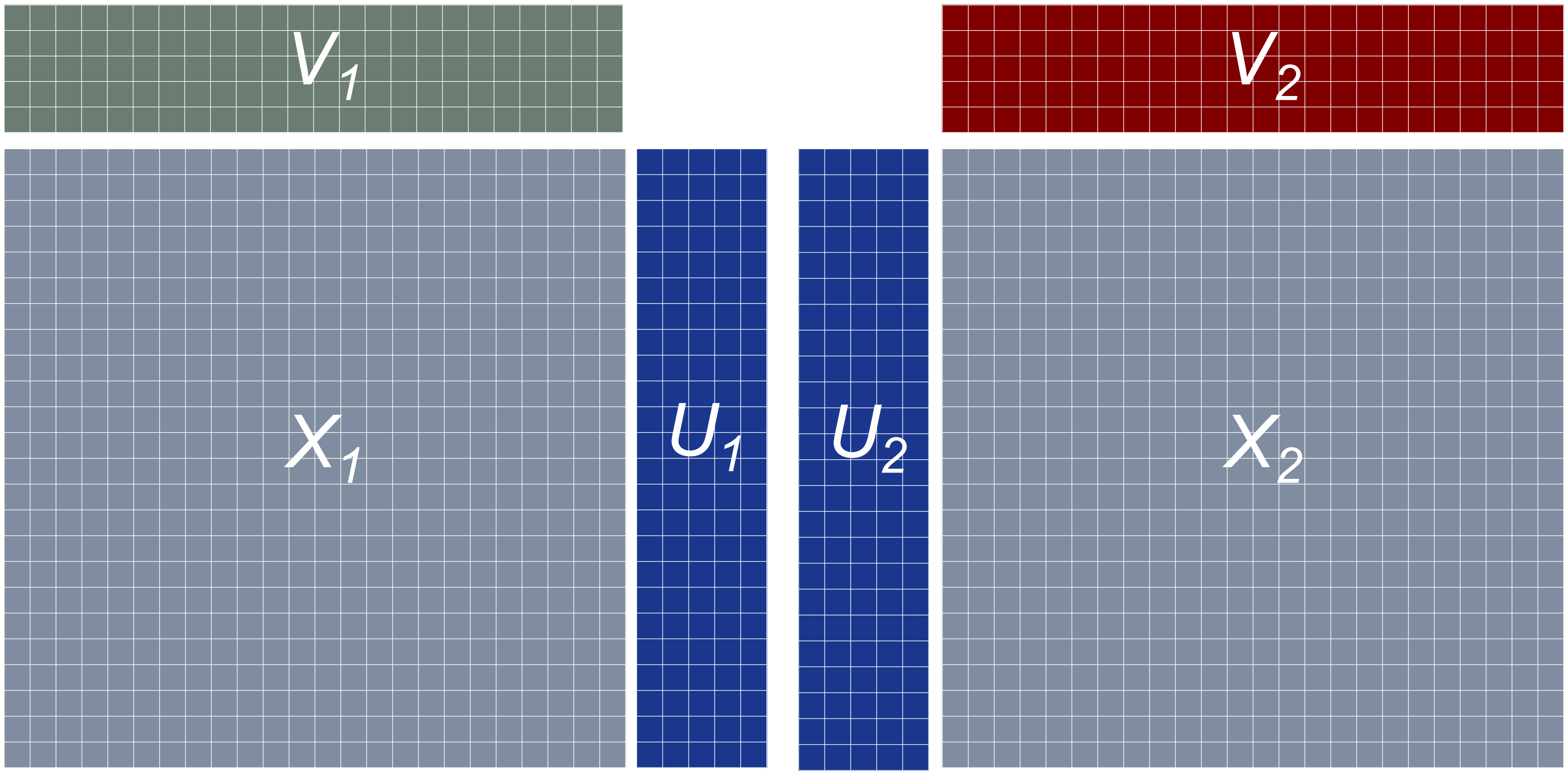
Clustered “attacked” movies



83% are clustered together

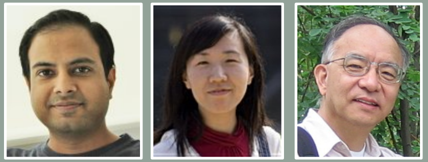
CoBaFi: Collaborative Bayesian Filtering
Alex Beutel, Kenton Murray,
Christos Faloutsos Alex Smola
WWW 2014

Outliers in Joint Factorization



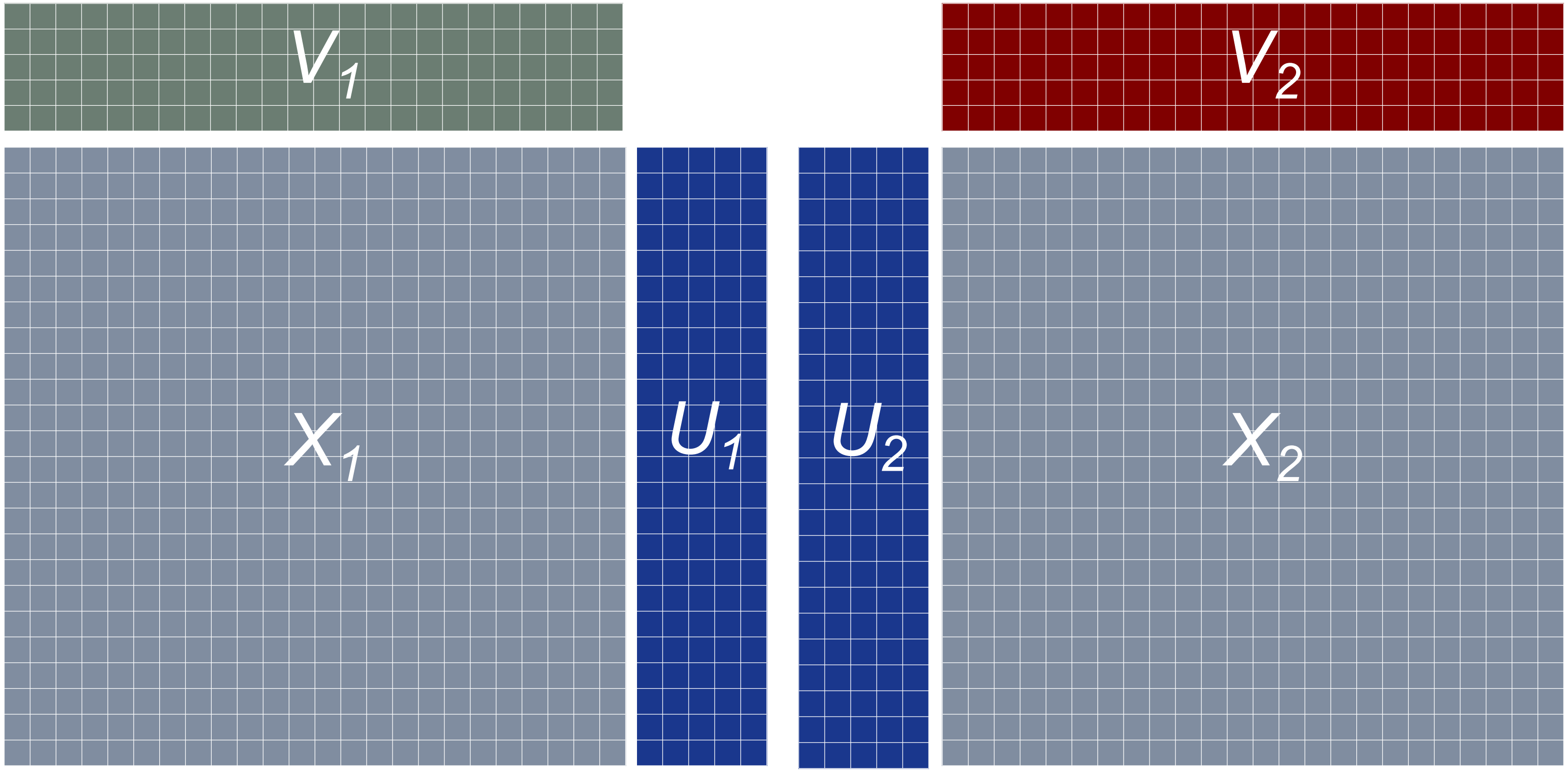
Enforce $U_1 \approx U_2$ and $U_1, U_2, V_1, V_2 \geq 0$

Community Distribution Outlier Detection in Heterogeneous Information Networks
Manish Gupta, Jing Gao, and Jiawei Han
ECML/PKDD 2013



Outliers in Joint Factorization

Interesting design of X_1 and X_2; see paper for details

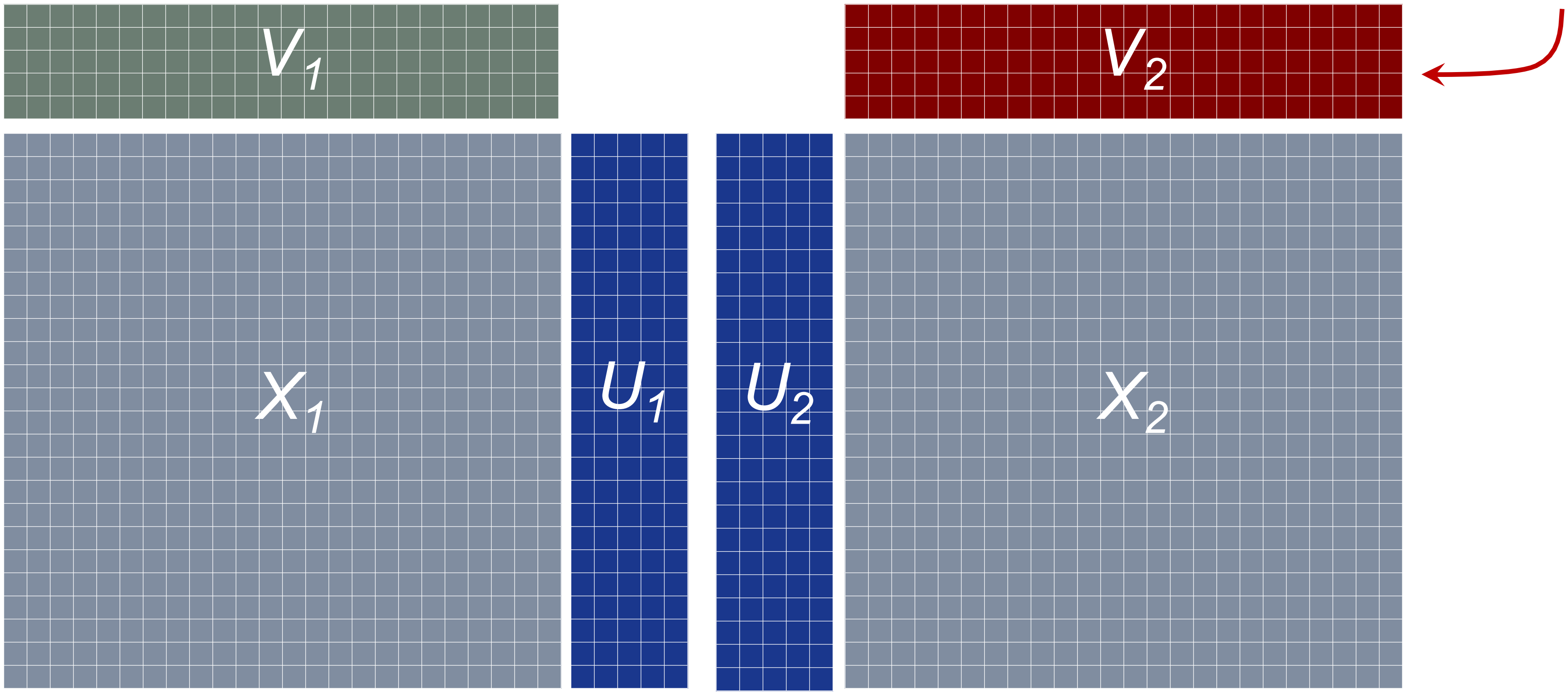


Enforce $U_1 \approx U_2$ and $U_1, U_2, V_1, V_2 \geq 0$

Community Distribution Outlier Detection in Heterogeneous Information Networks
Manish Gupta, Jing Gao, and Jiawei Han
ECML/PKDD 2013

Outliers in Joint Factorization

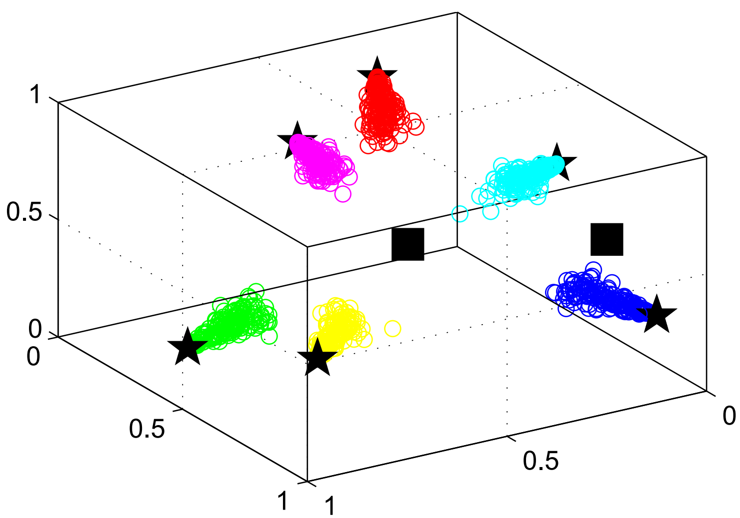
Rows of V_2 represent common patterns in X_2 (cluster centroids)



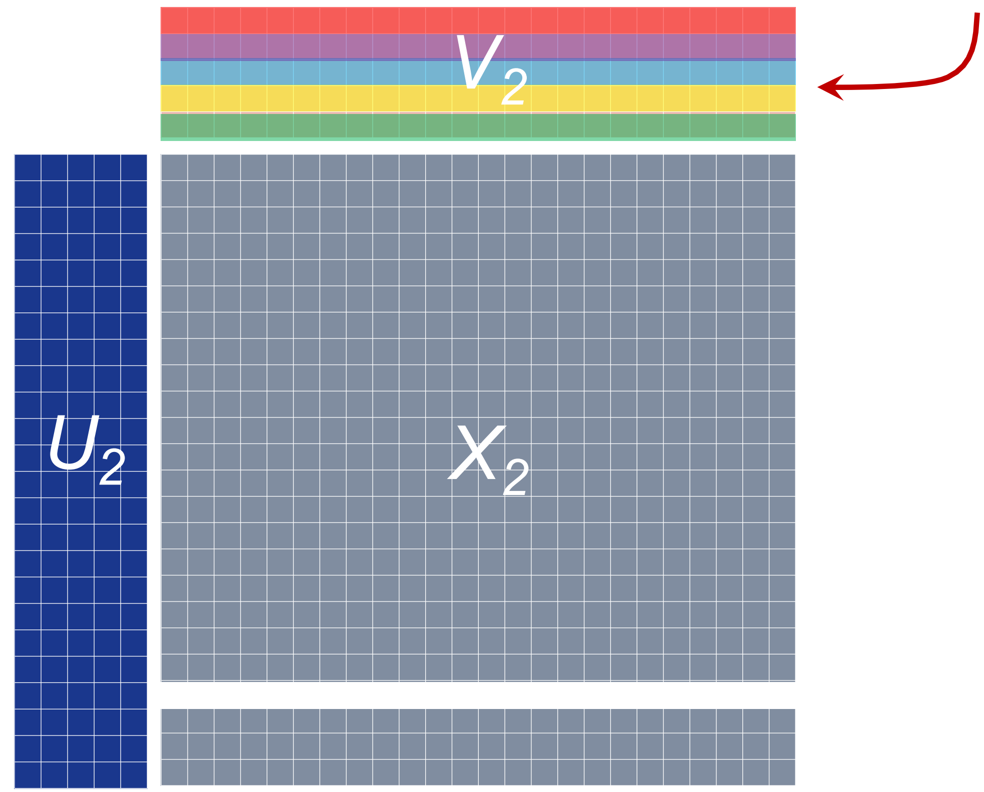
Enforce $U_1 \approx U_2$ and $U_1, U_2, V_1, V_2 \geq 0$

Outliers in Joint Factorization

An anomaly is a row of X_i that is *not* similar to any row in V_i



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



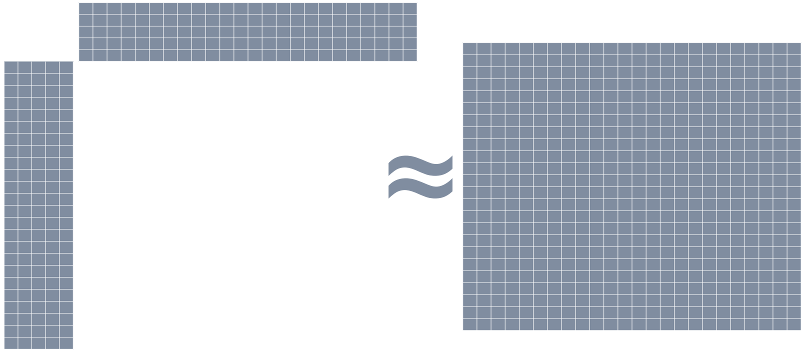
Community Distribution Outlier Detection in Heterogeneous Information Networks
Manish Gupta, Jing Gao, and Jiawei Han
ECML/PKDD 2013

Practitioner's Guide

Method	Graph Type	Node Attributes	Edge Attributes	Seed Labels
EigenSpokes	Directed+			
Get-the-Scoop	Directed+			
fBox	Directed+			
CoBaFi	Bipartite+		✓	
CDOutliers	Undirected	✓		

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

$$\left\{\begin{matrix}\text{Undirected}\\\text{Directed}\\\text{Bipartite}\end{matrix}\right\} \times \left\{\text{Node Attributes}\right\} \times \left\{\text{Edge Attributes}\right\} \times \left\{\begin{matrix}\text{Unsupervised}\\\text{Semi-Supervised}\end{matrix}\right\}$$

Open Problems / Opportunities

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

{ Undirected, Directed, Bipartite }

×

{ Node Attributes }

×

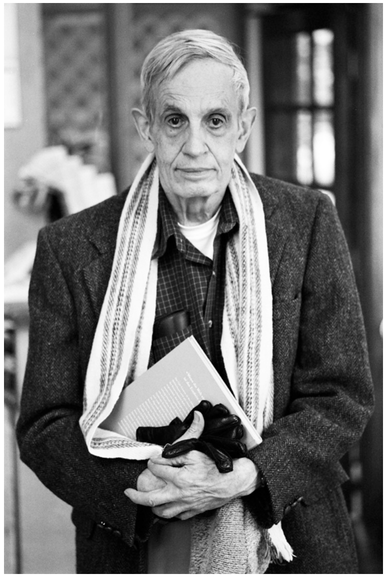
{ Edge Attributes }

×

{ Unsupervised, Semi-Supervised }

Open Problems / Opportunities

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



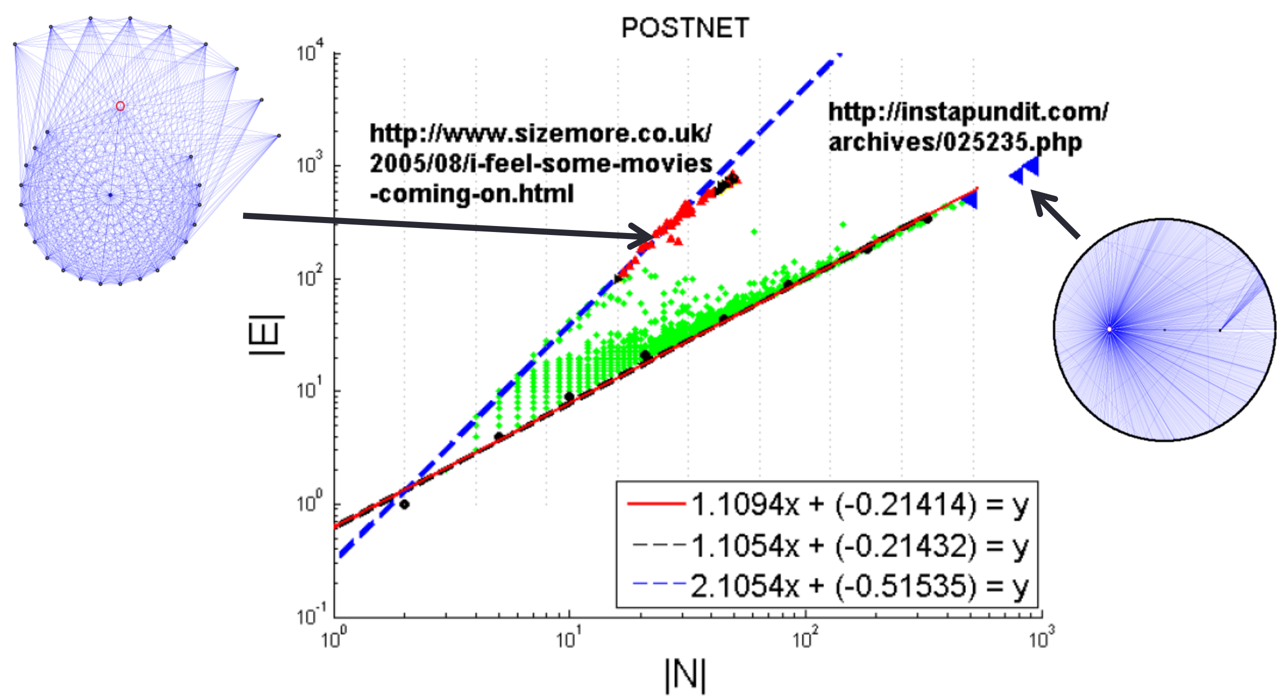
Open Problems / Opportunities

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

Summary

Local Subgraph Analysis: Patterns and Features

e.g. using ego-nets

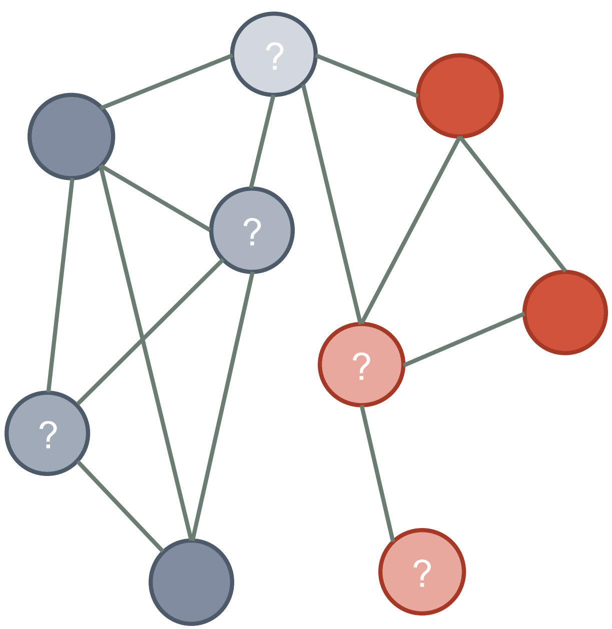


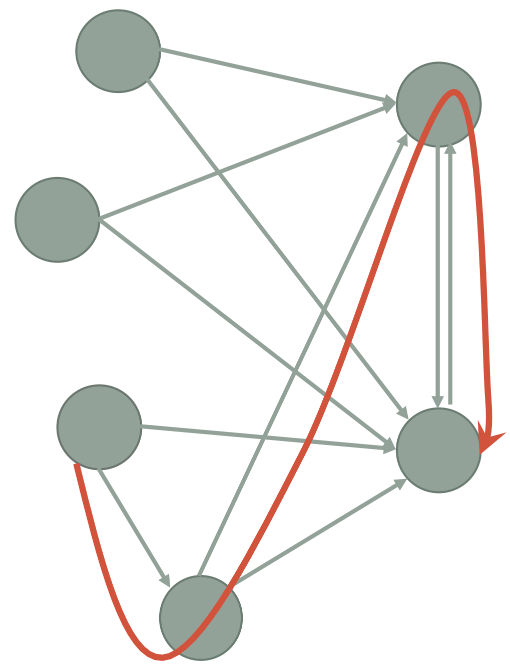
Summary

Propagation Methods

"Guilt-by-association"

"Importance-by-association" = PageRank

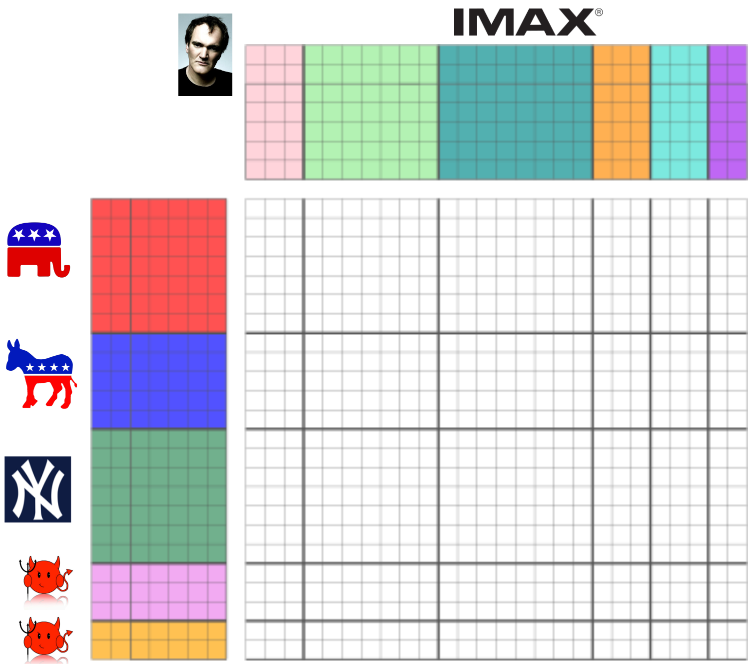


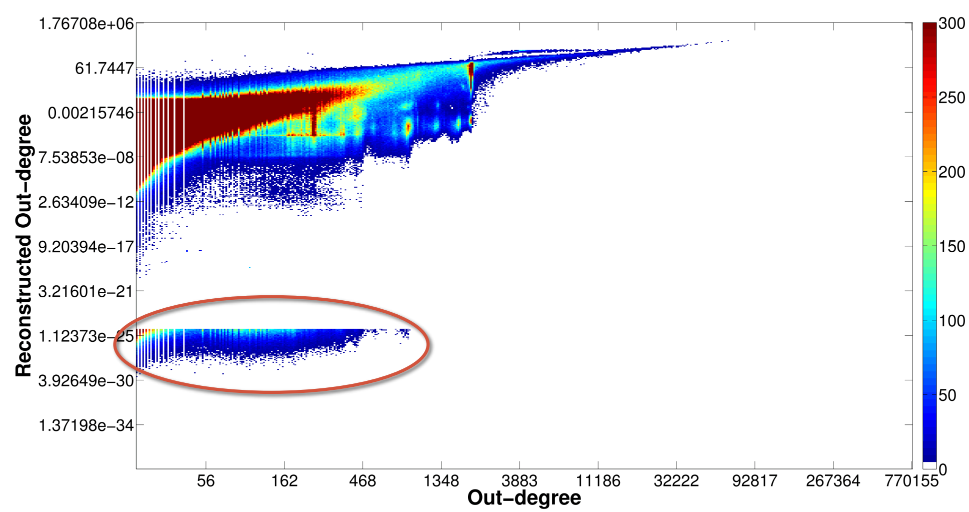


Summary

Latent Factor Models

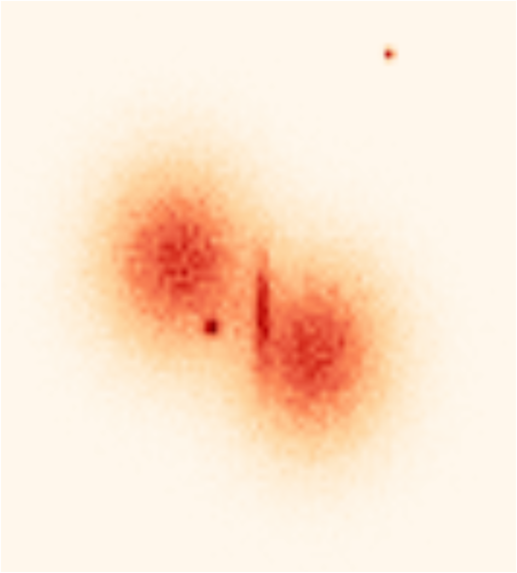
Find multiple communities, patterns and anomalies.

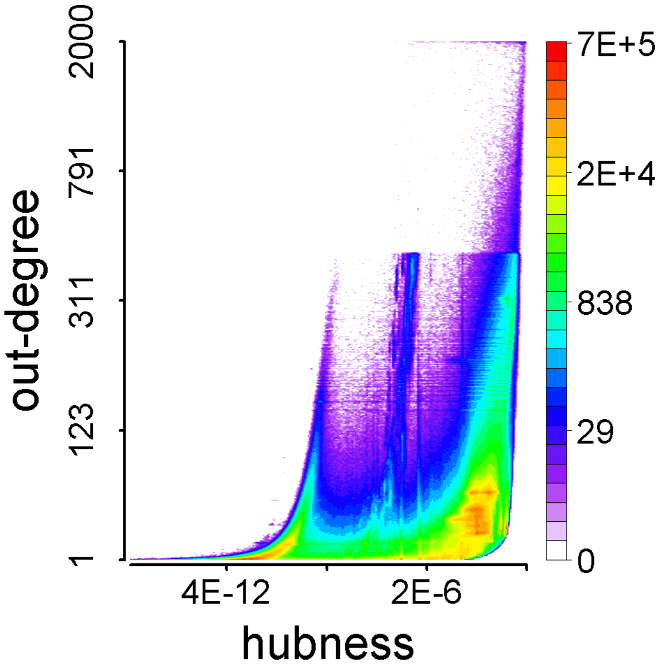




Take Away

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



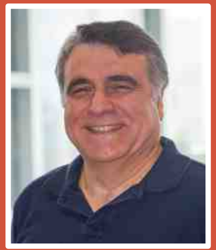


Thanks again to

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

Questions?





Carnegie Mellon University

Stony Brook University

References and resources available at

cs.cmu.edu/~abeutel/ccs2015_tutorial