Recommending Tumblr Blogs to Follow

Donghyuk Shin
Dept of Computer Science
UT-Austin

Group Meeting
May 4, 2015

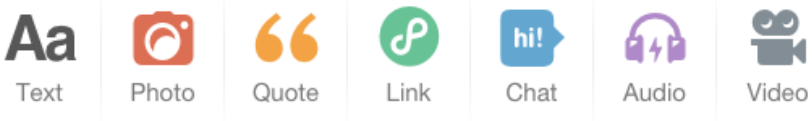
Joint work with S. Cetintas and K.C. Lee

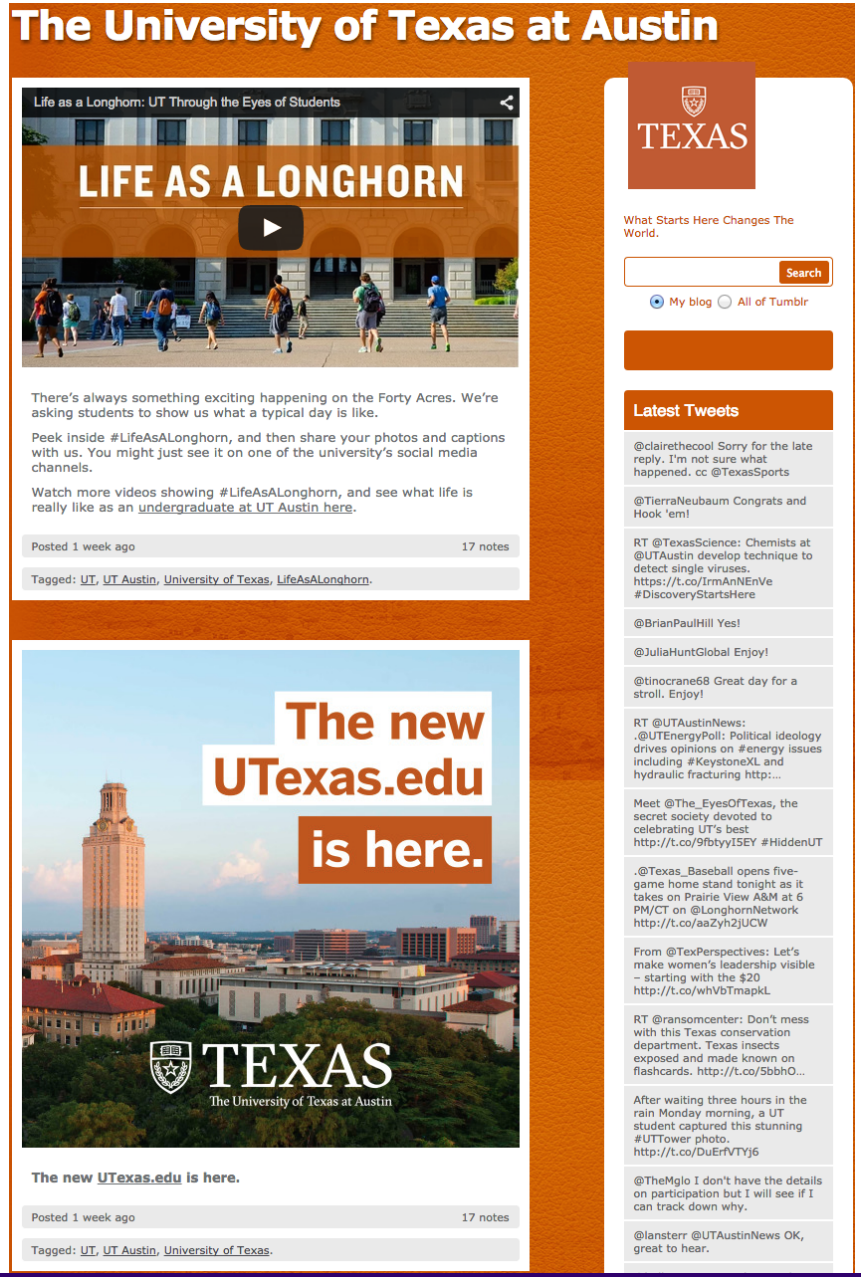
Tumblr Blog Recommendation

Tumblr

Microblogging service:

- No length limitation of posts.
- Supports various types of posts:

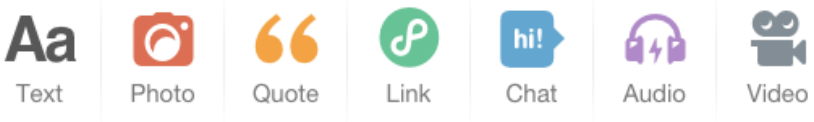




Tumblr

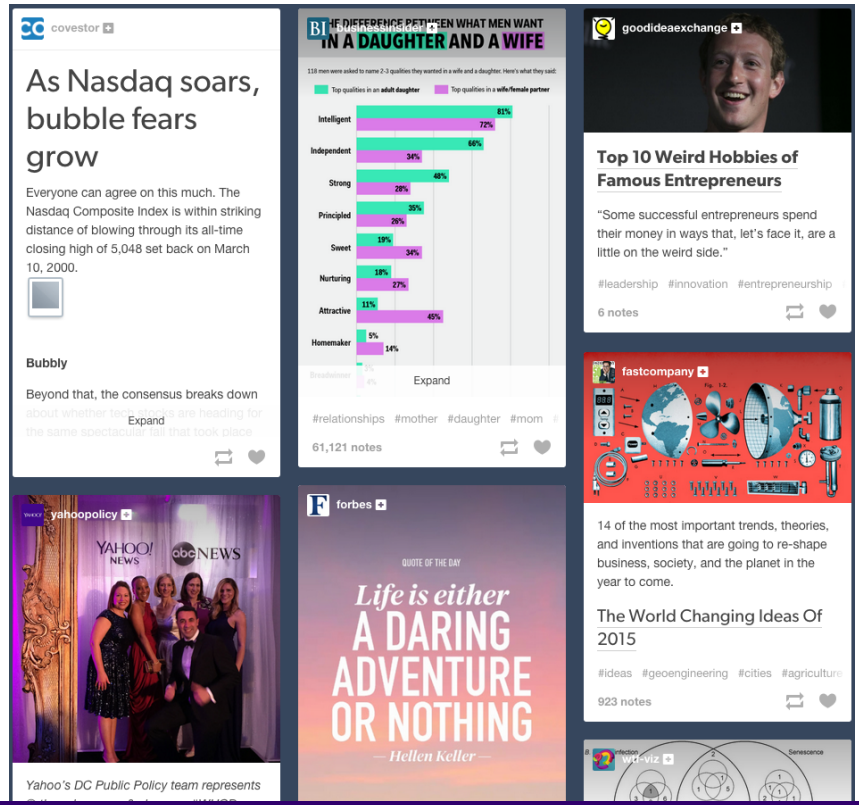
Microblogging service:

- No length limitation of posts.
- Supports various types of posts:



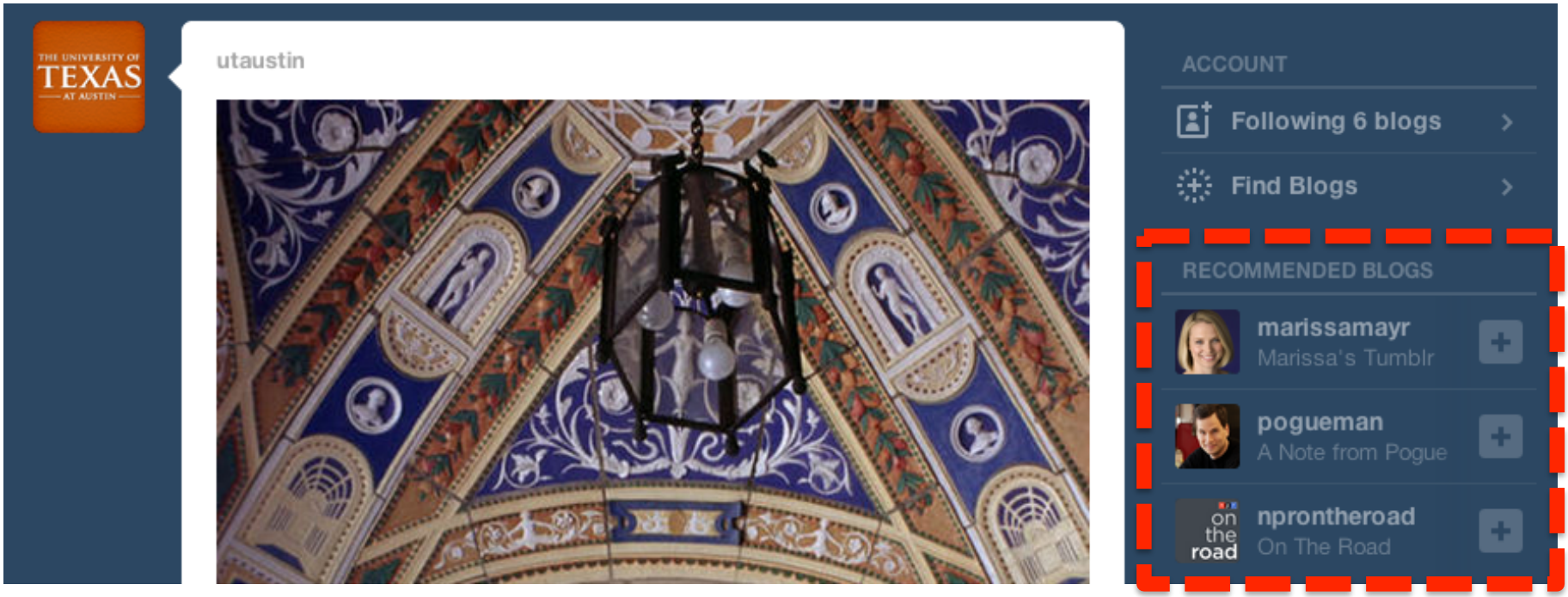
Social network service:

- *Follow* other blogs – 230M blogs
- Like or reblog other posts – 110B posts



Blog Recommendation

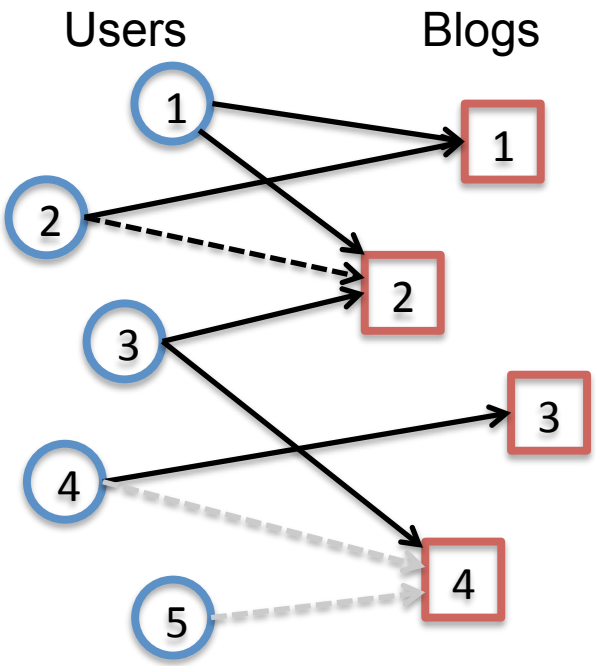
Who to follow?

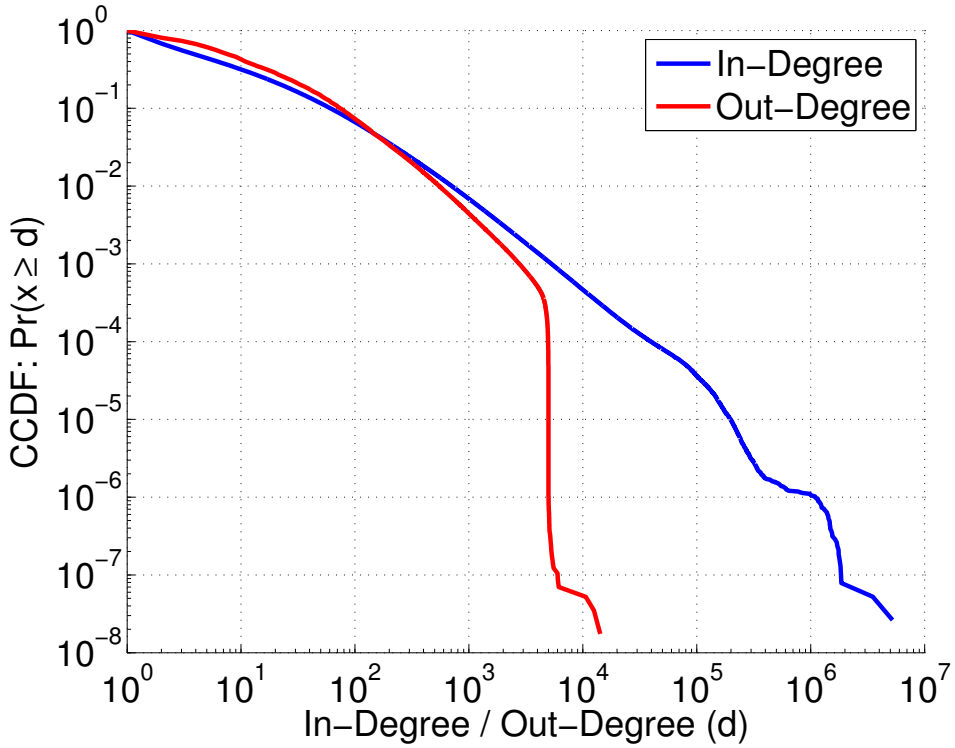


- Help users find content they want $\Rightarrow$ enhance user experience and engagement
- Help increase followers for sponsored or advertiser blogs

Follower Graph

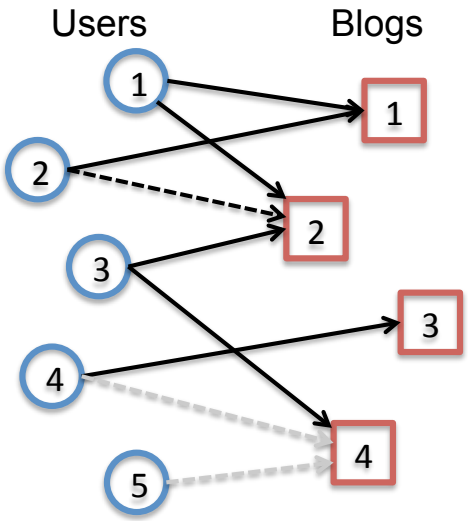
- Snapshot from Jun. 2014
- Directed graph: 76.86M nodes (users/blogs) and 2.27B edges (follows) with timestamps
- 50% have 0 in-degree, 25% have 0 out-degree

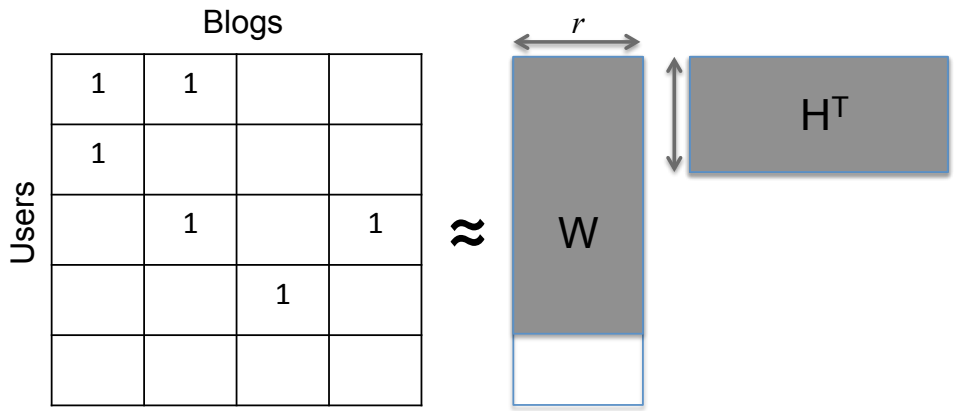




Existing Methods

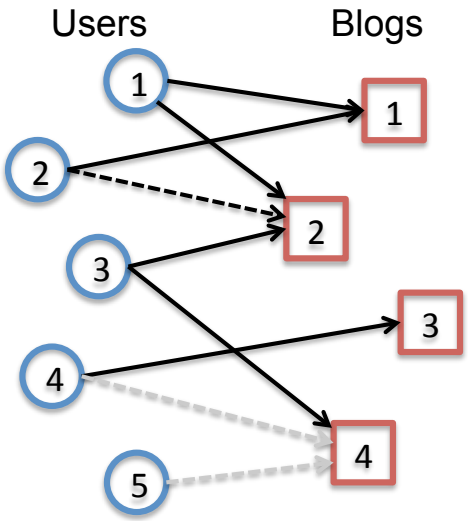
- Global ranking: recommend most popular blogs
- Graph proximity: Common Neighbors, Katz, etc
- Standard matrix completion

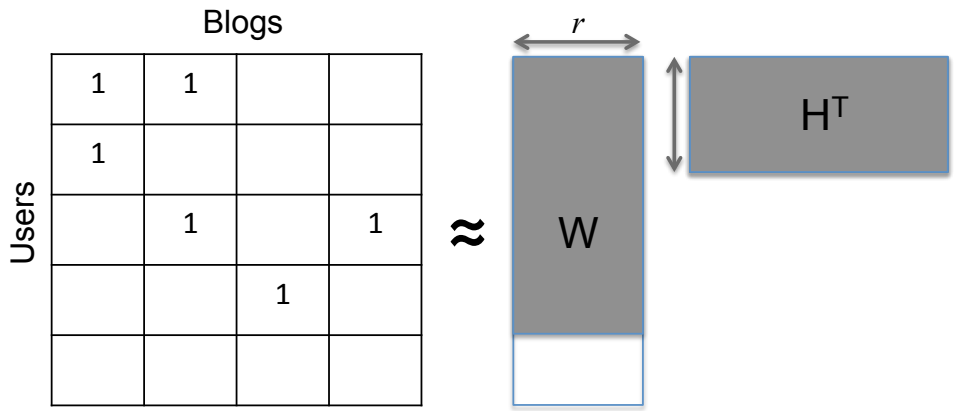




Existing Methods

- Global ranking: recommend most popular blogs
- Graph proximity: Common Neighbors, Katz, etc
- Standard matrix completion





Can NOT make meaningful recommendations for users or blogs with no follow information (e.g., user 5)

Additional Information from Posts

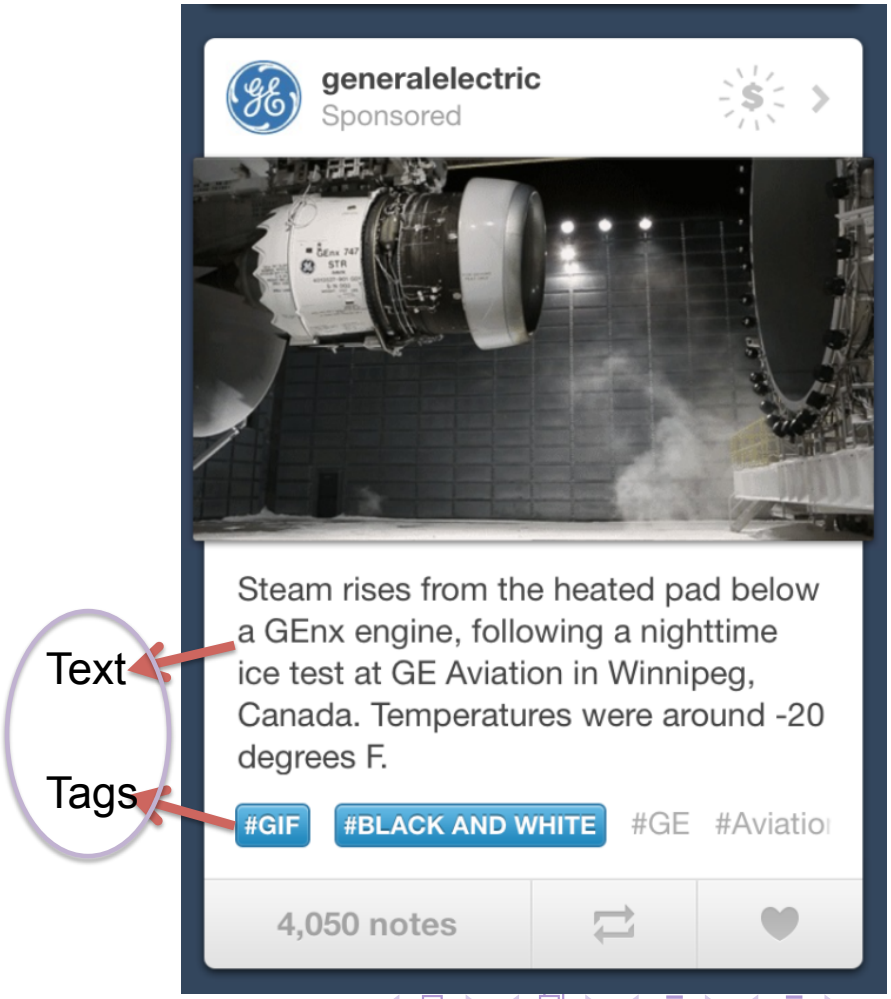
Tumblr Posts

Various sources of additional information:

Tumblr Posts

Various sources of additional information:

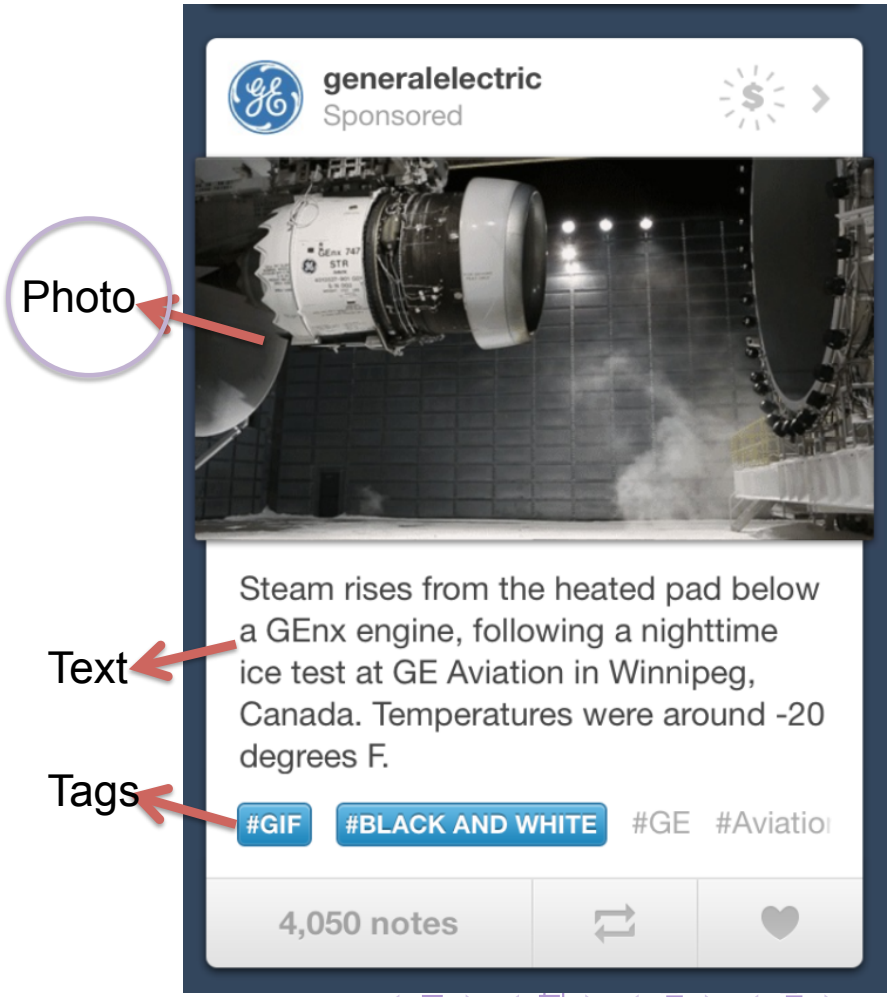
- Textual: text and tags



Tumblr Posts

Various sources of additional information:

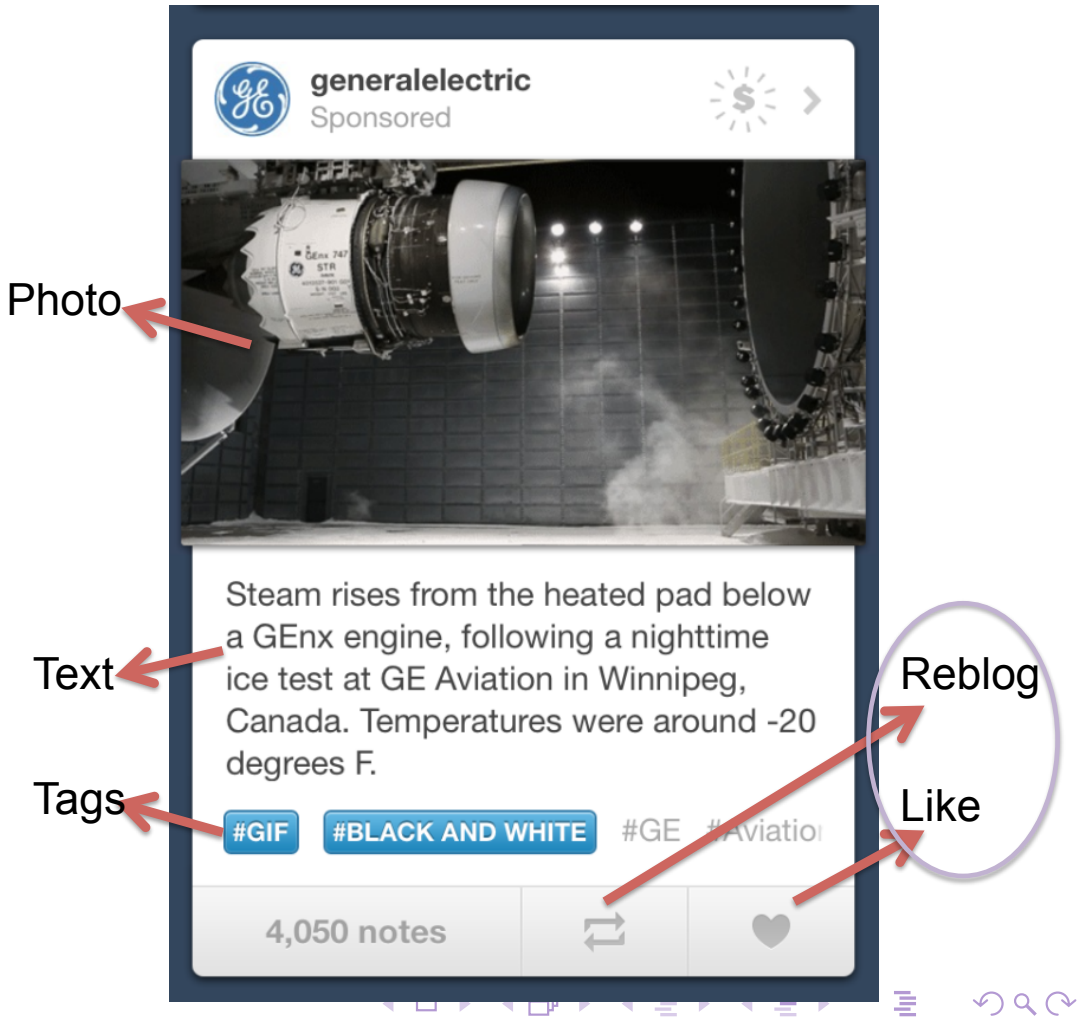
- Textual: text and tags
- Visual: photos (and videos)



Tumblr Posts

Various sources of additional information:

- Textual: text and tags
- Visual: photos (and videos)
- Activity: reblog and like



Tumblr Posts

Various sources of additional information:

- Textual: text and tags
- Visual: photos (and videos)
- Activity: reblog and like

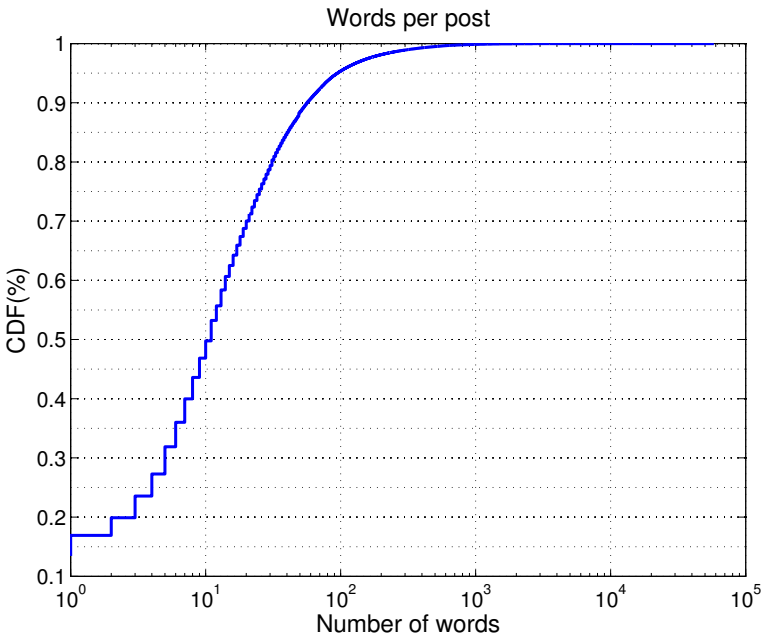
User and blog features are generated from each source.

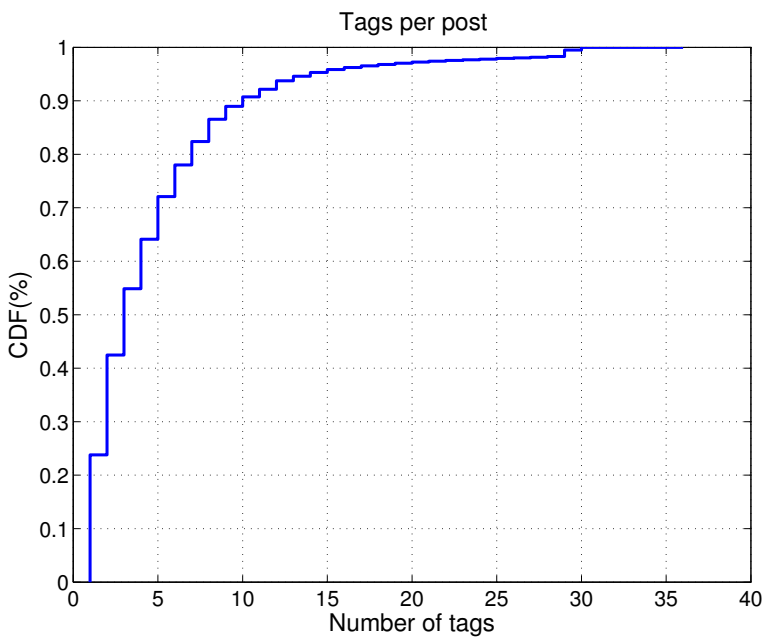
- 150M new posts per month
- Collected over a 5 month period: more than 7.5 TB of data

User and Blog Features – Textual

Text and Tags:

- Text data is extremely sparse and noisy – 28.7 words and 4.8 tags per post on average





Problematic for existing bag-of-words models (LSA, LDA, etc).

User and Blog Features – Textual

Text and Tags: extract features using word2vec [Mikolov, et.al. 2013]

User and Blog Features – Textual

Text and Tags: extract features using word2vec [Mikolov, et.al. 2013]

- Learns a vector representation of each word such that words in similar context are close to each other

User and Blog Features – Textual

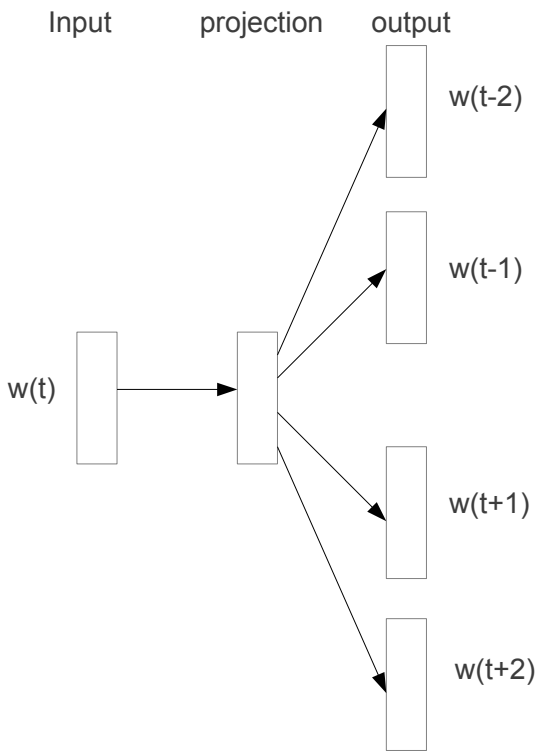
Text and Tags: extract features using word2vec [Mikolov, et.al. 2013]

- Learns a vector representation of each word such that words in similar context are close to each other
- Trains a neural network maximizing the conditional probability of context given a word: $\arg\max_\theta \prod_{(w,c)\in D} p(c|w;\theta)$

User and Blog Features – Textual

Text and Tags: extract features using word2vec [Mikolov, et.al. 2013]

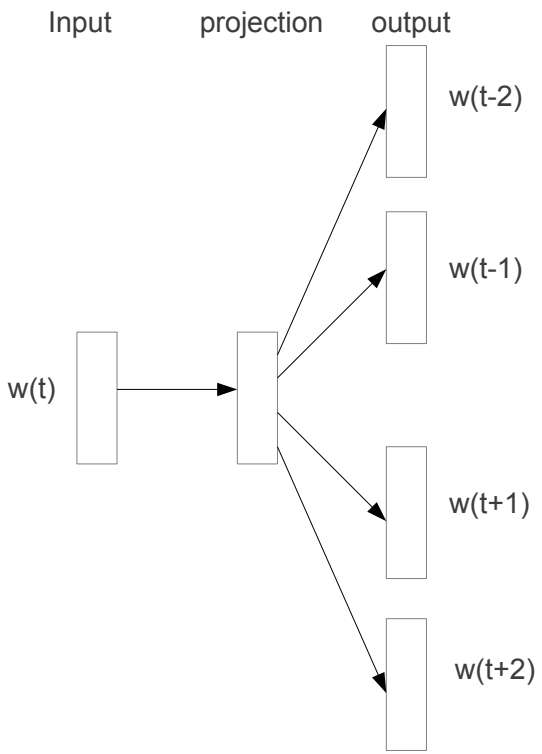
- Learns a vector representation of each word such that words in similar context are close to each other
- Trains a neural network maximizing the conditional probability of context given a word: $\arg\max_\theta \prod_{(w,c)\in D} p(c|w;\theta)$

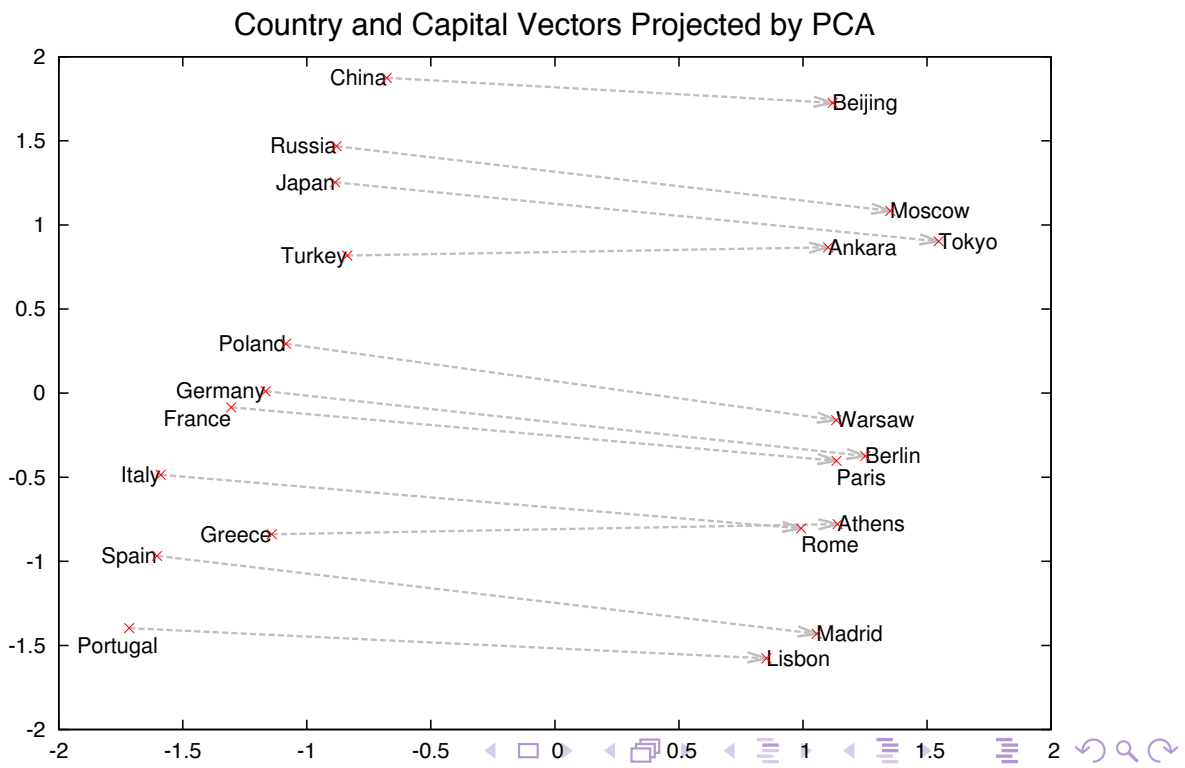


User and Blog Features – Textual

Text and Tags: extract features using word2vec [Mikolov, et.al. 2013]

- Learns a vector representation of each word such that words in similar context are close to each other
- Trains a neural network maximizing the conditional probability of context given a word: $\arg\max_\theta \prod_{(w,c)\in D} p(c|w;\theta)$





User and Blog Features – Textual

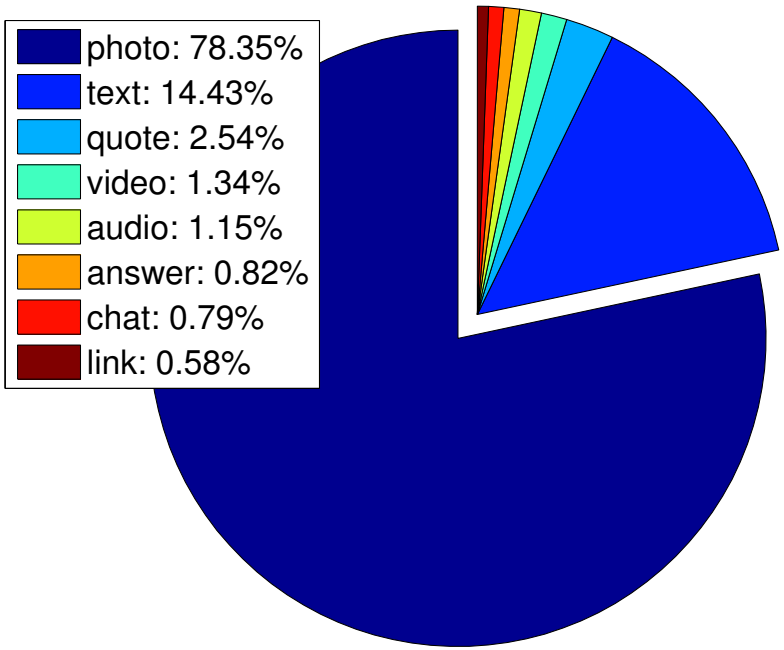
Text and Tags: extract features using word2vec [Mikolov, et.al. 2013]

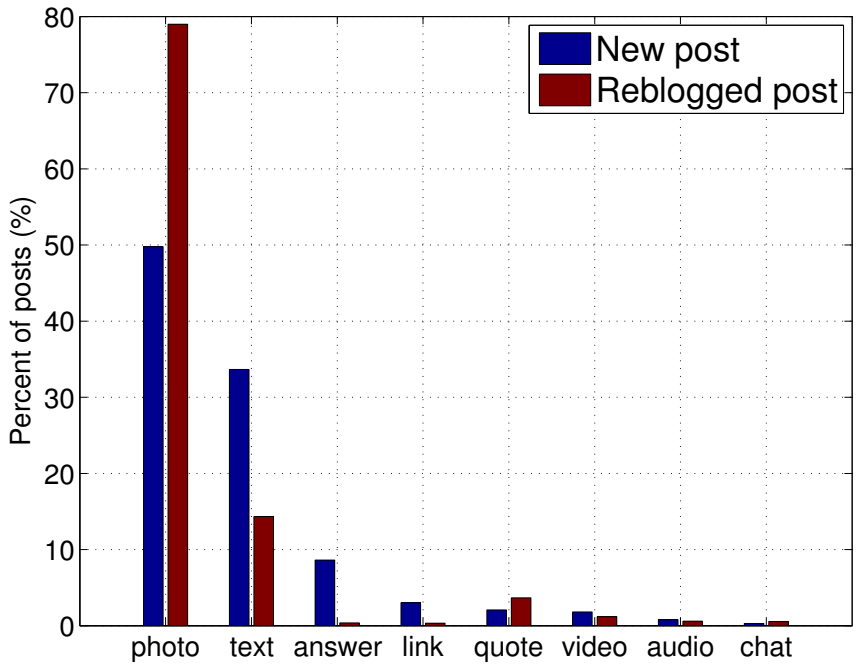
1. Learn d-dimensional vector representations of each word
2. Cluster the words into c clusters via k-means algorithm
3. Compute histogram of clusters for each post
4. Take average of histograms as features of users and blogs

User and Blog Features – Visual

Photos:

- About 78% of total posts are photo posts.
- Majority of reblogged posts are also photo posts.

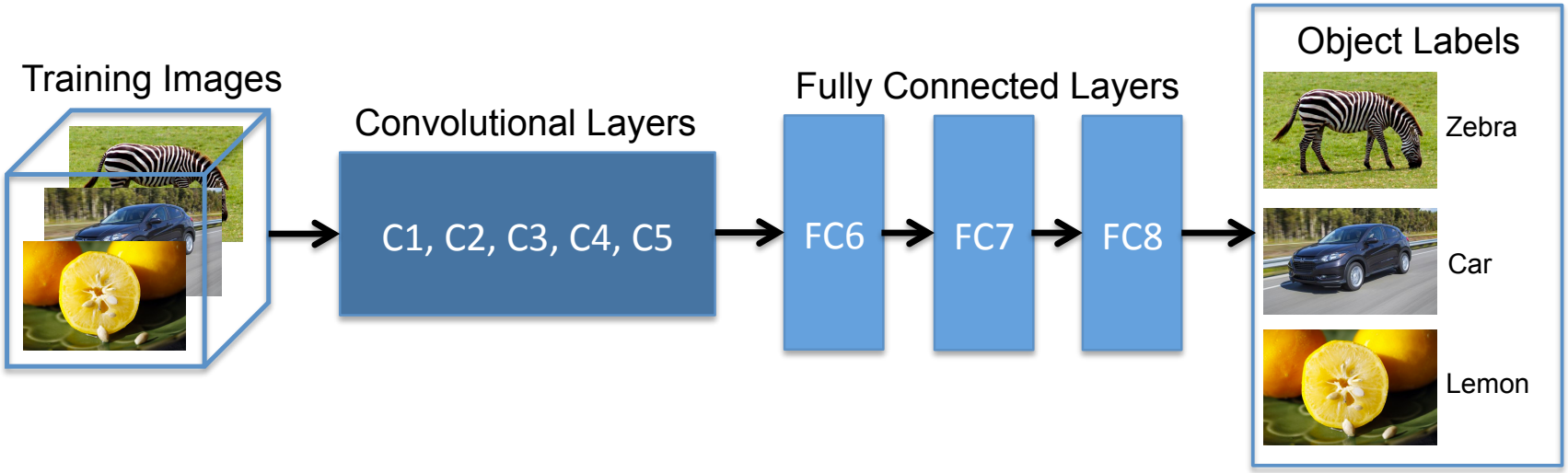




User and Blog Features – Visual

Photos: extract features using deep learning

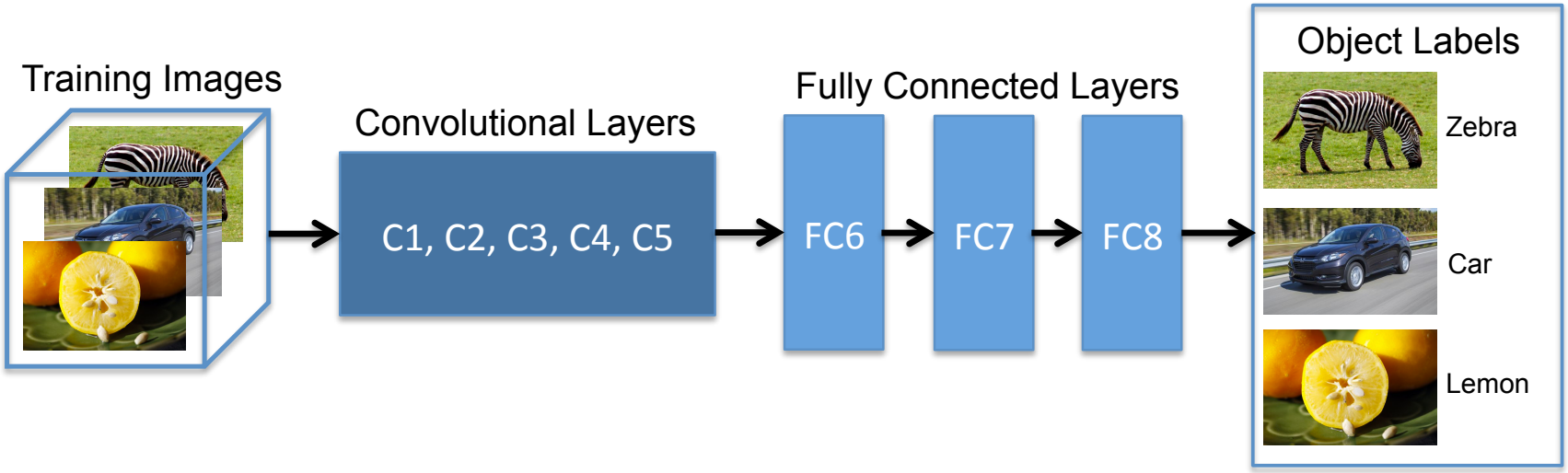
- Used a convolutional neural network with a total of 8 layers – 7 hidden layers and a soft-max layer [Krizhevsky, et.al. 2012, Donahue, et.al. 2014]
  - Trained on 1.5M Flickr images (labels unavailable for Tumblr images)



User and Blog Features – Visual

Photos: extract features using deep learning

- Used a convolutional neural network with a total of 8 layers – 7 hidden layers and a soft-max layer [Krizhevsky, et.al. 2012, Donahue, et.al. 2014]
  - Trained on 1.5M Flickr images (labels unavailable for Tumblr images)



- Final output is a 958-dimensional vector of confidence scores over pre-defined categories
  - For users, averaged over photos that a user posted/liked/reblogged
  - For blogs, averaged over photos that a blog posted/reblogged

User and Blog Features – Activity

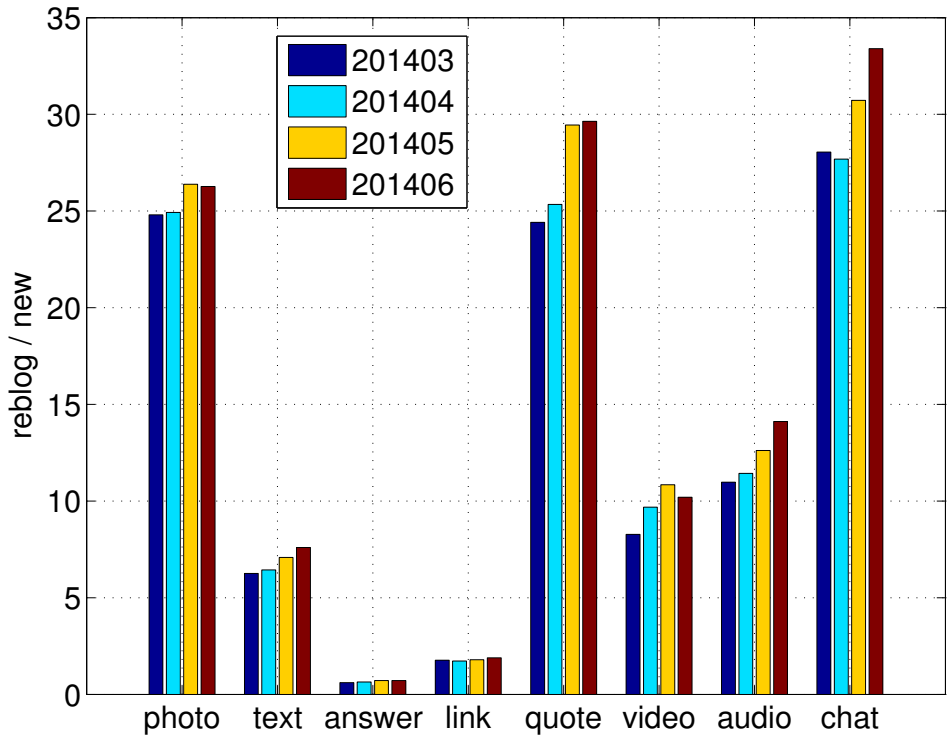
User actions:

- Reblog: clones the post to the user's blog (similar to Twitter's retweet)
- Like: marks what a user likes (similar to Facebook's like)

User and Blog Features – Activity

User actions:

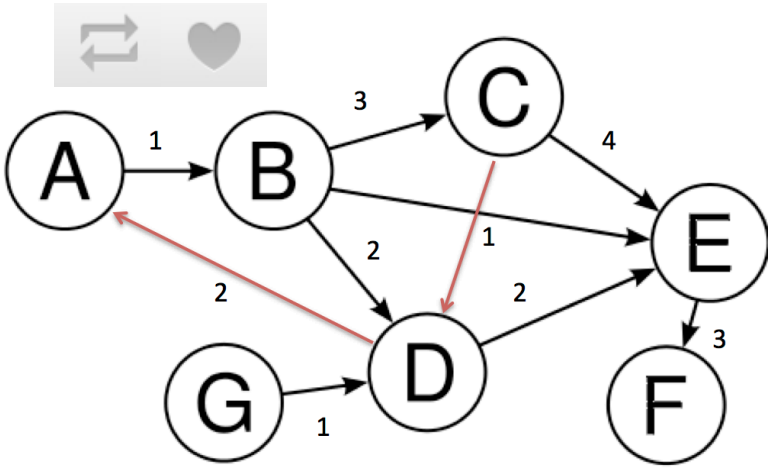
- Reblog: clones the post to the user's blog (similar to Twitter's retweet)
- Like: marks what a user likes (similar to Facebook's like)
- 2.5B reblogs and 2.0B likes per month
- Reblog posts consist of more than 90% of total posts



User and Blog Features – Activity

Both activity can be represented as a graph W:

- W_{ij}: number of reblogged/liked posts of blog j by user i.

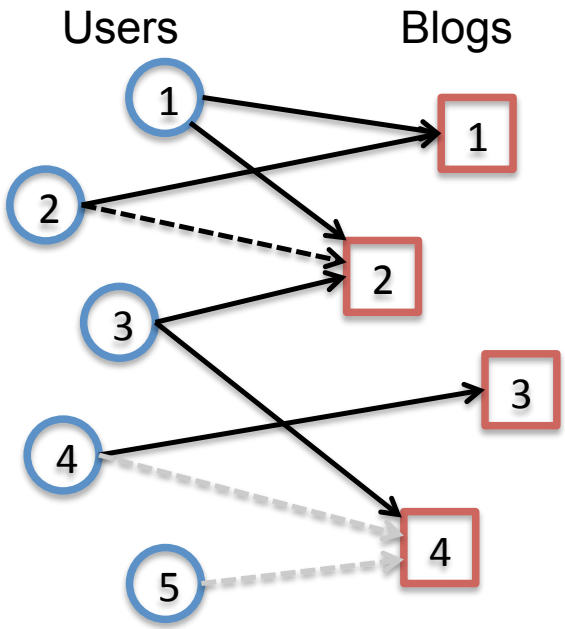


We extract principal components of the reblog/like graph.

Methods Incorporating Additional Information

Graph-based Method: Katz

Follower graph:

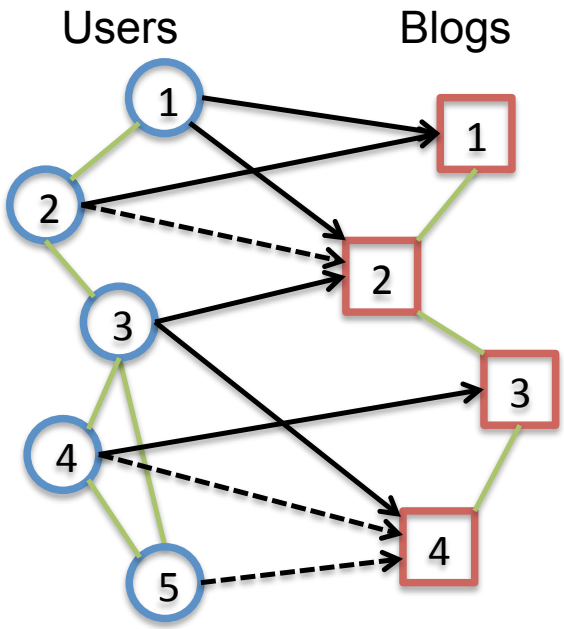


$$C = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

- Compute proximity measures (Katz) between users and blogs on C.

Graph-based Method: Katz

Follower graph + user and blog features:

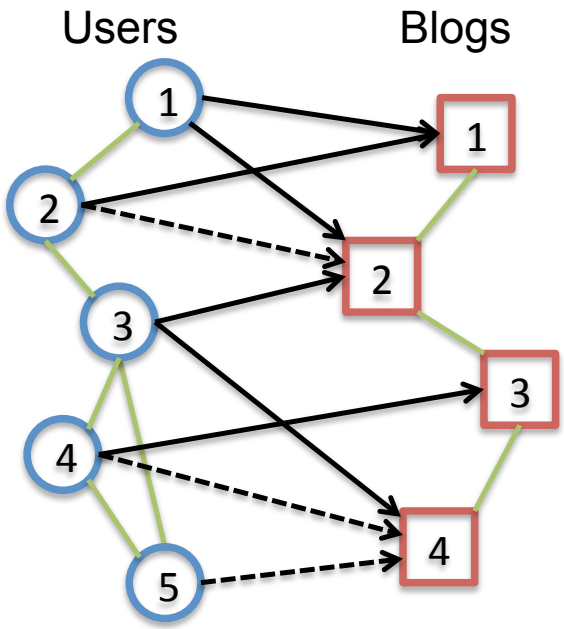


$$C = \begin{bmatrix} U & A \\ A^T & B \end{bmatrix}$$

- Compute proximity measures (Katz) between users and blogs on C.
- U and B are similarity matrices computed from user and blog features, respectively.

Graph-based Method: Katz

Follower graph + user and blog features:

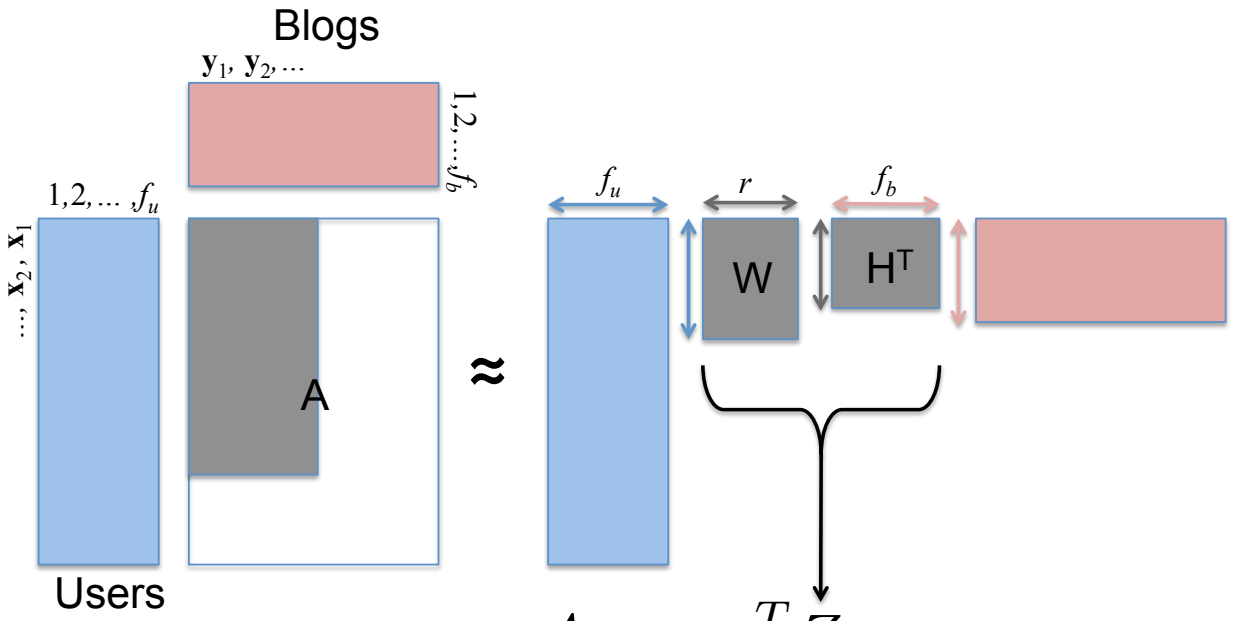


$$C = \begin{bmatrix} U & A \\ A^T & B \end{bmatrix}$$

- Compute proximity measures (Katz) between users and blogs on C.
- U and B are similarity matrices computed from user and blog features, respectively.

High computational cost to compute U and B, and Katz scores

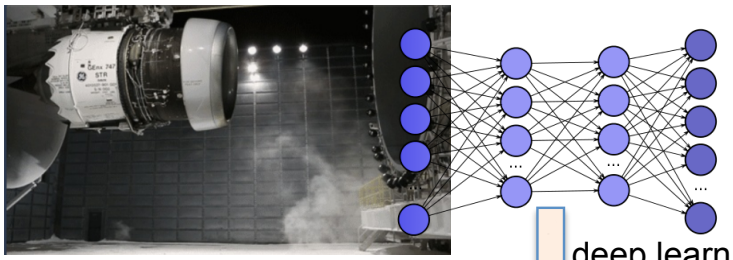
Inductive Matrix Completion



$$\min_{W,H} \sum_{(i,j)\in\Omega} (A_{ij} - \mathbf{x}_i^T W H^T \mathbf{y}_j)^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2)$$

[Jain and Dhillon 2013, Natarajan and Dhillon 2014]

Inductive Matrix Completion



Photo

deep learning

Text and Tags

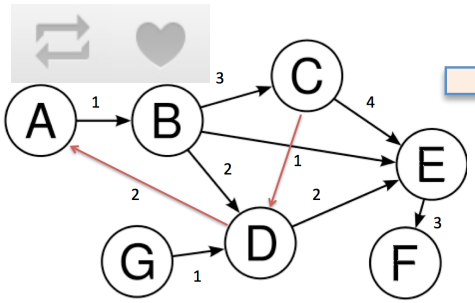
Steam rises from the heated pad below a GEnx engine, following a nighttime ice test at GE Aviation in Winnipeg, Canada. Temperatures were around -20 degrees F.

#GIF #BLACK AND WHITE #GE

word vec

latent

Like and Reblog Graphs



Blogs

Users

A

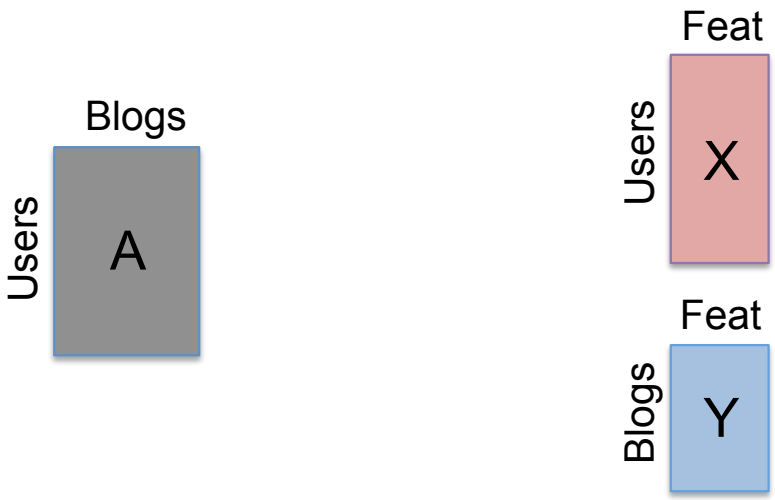
$$\approx$$

f_u r f_b

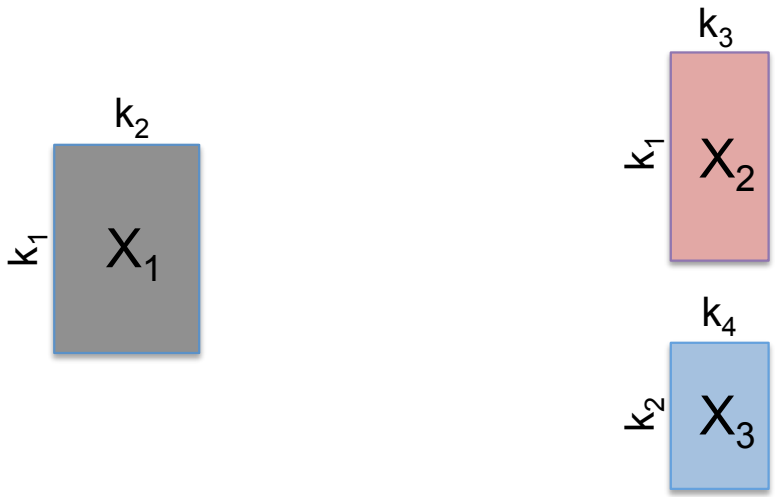
W H^T

$$A_{ij} = \mathbf{x}_i^T Z \mathbf{y}_j$$

Collective Matrix Completion

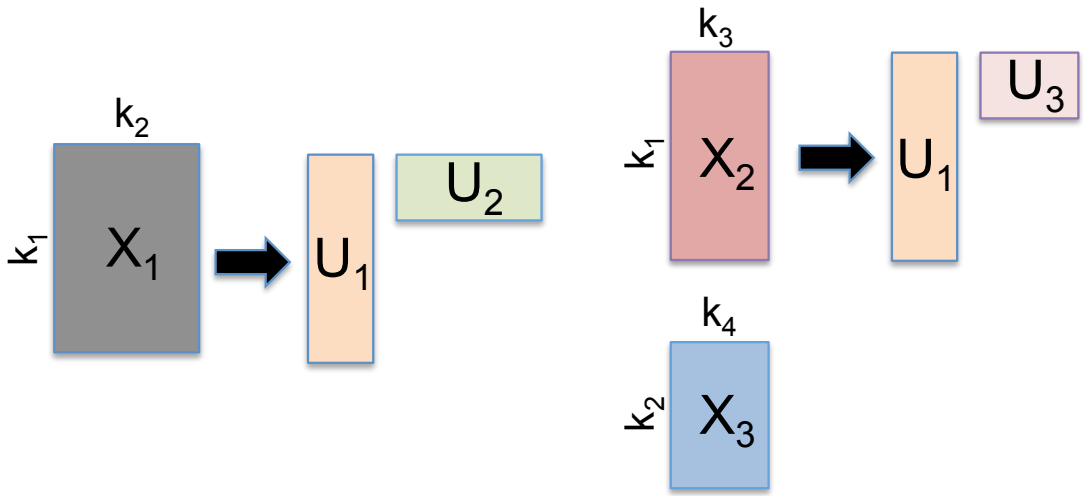


Collective Matrix Completion



- Each matrix (view) X_v represents relations between entities r_v and c_v

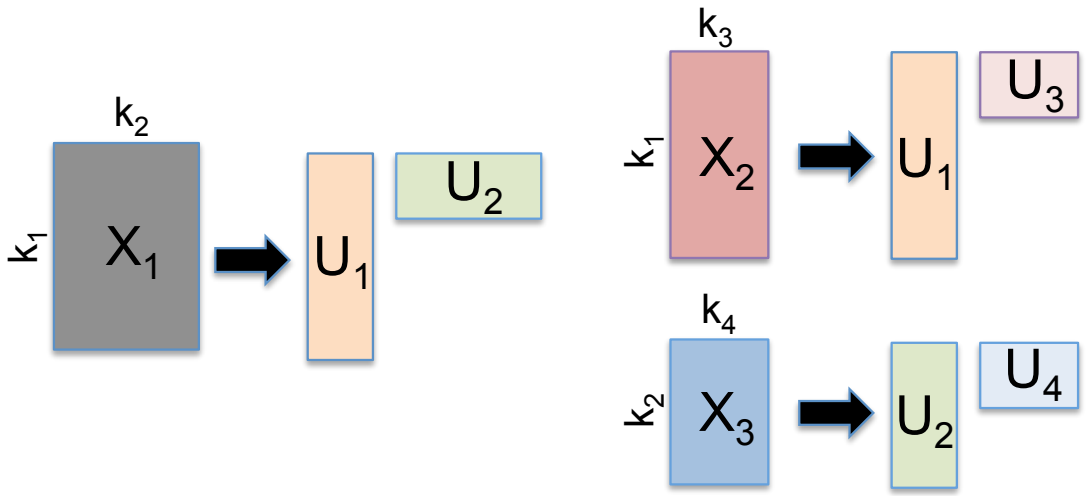
Collective Matrix Completion



- Each matrix (view) X_v represents relations between entities r_v and c_v
- Goal: jointly recover a collection of matrices with shared low rank structure

$$X_v = U_{r_v} U_{c_v}^T$$

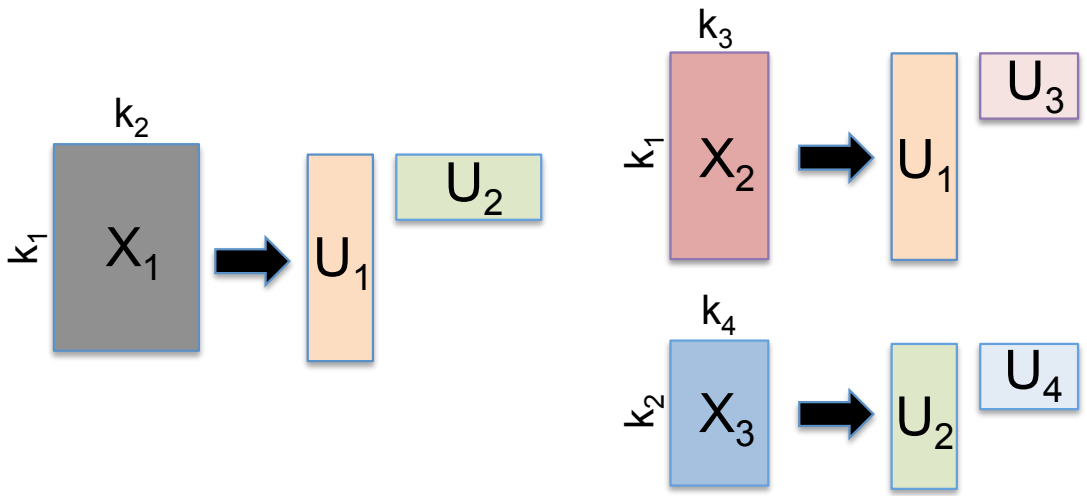
Collective Matrix Completion



- Each matrix (view) X_v represents relations between entities r_v and c_v
- Goal: jointly recover a collection of matrices with shared low rank structure

$$X_v = U_{r_v} U_{c_v}^T$$

Collective Matrix Completion



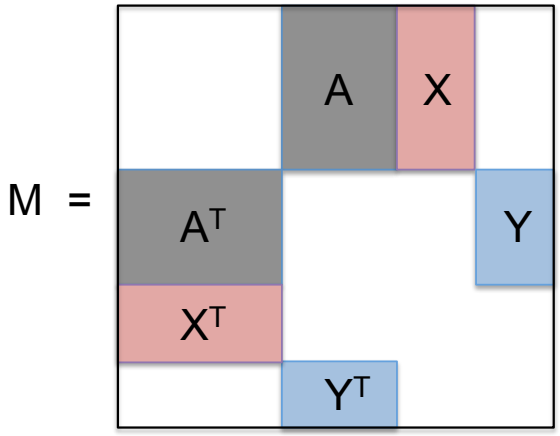
- Each matrix (view) X_v represents relations between entities r_v and c_v
- Goal: jointly recover a collection of matrices with shared low rank structure

$$X_v = U_{r_v} U_{c_v}^T$$

$$\min_{\{U_k \in \mathbb{R}^{n_k \times r}\}_k} \sum_{v=1}^{V} \ell(X_v, U_{r_v} U_{c_v}^T) + \lambda \sum_{k=1}^{K} \|U_k\|_F^2$$

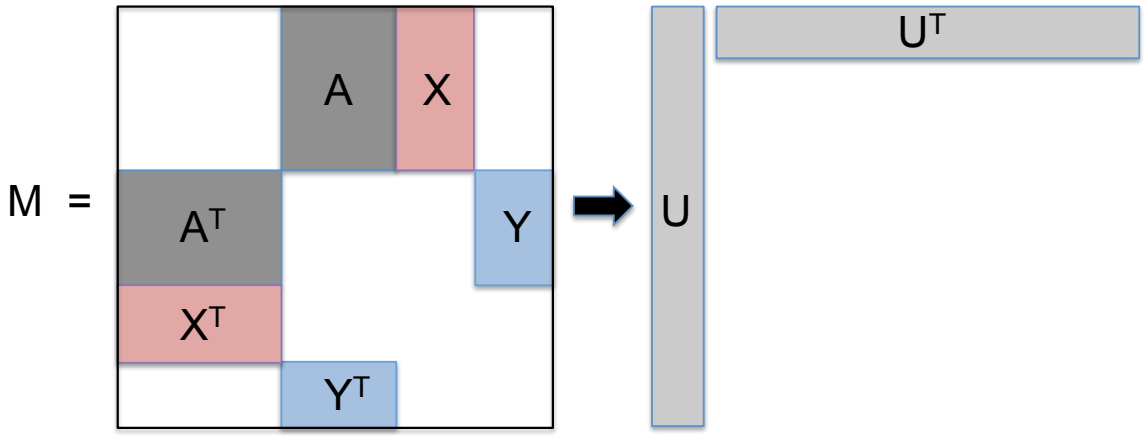
Collective Matrix Completion

- Equivalently, can be represented as a block symmetric matrix M



Collective Matrix Completion

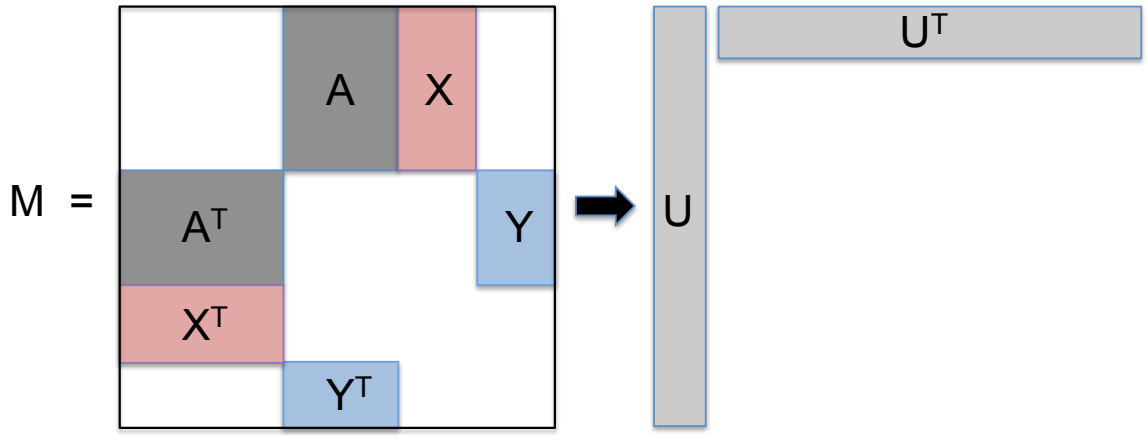
- Equivalently, can be represented as a block symmetric matrix M



$$M = UU^T$$

Collective Matrix Completion

- Equivalently, can be represented as a block symmetric matrix M



$$M = UU^T$$

- Recent work:
  - Original formulation: [Singh and Gordon. 2008]
  - Convex formulation: [Bouchard, et.al. 2013]
  - Consistency guarantees: [Gunasekar, et.al. 2015]

Experimental Results

- 1M randomly sampled users and blogs

Method	PRC@10	RCL@10	AUC
Global	1.03%	4.80%	0.8687
SVD	1.28%	5.10%	0.8530
MC	1.28%	5.07%	0.8515
Katz	1.90%	8.15%	0.9209
CMC	0.49%	2.41%	0.8996
IMC	2.93%	11.33%	0.9075

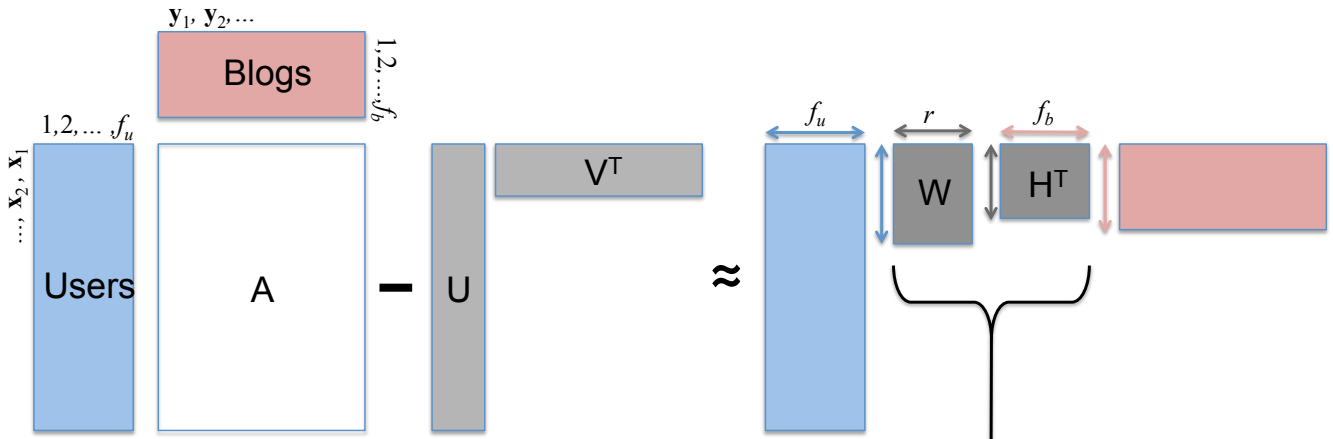
- Both IMC and CMC can be restrictive in the way they model A
- Especially, when features X and Y are noisy or does not support A
  - e.g., text data in Tumblr

Inductive Matrix Completion on the Residual

Combine MC and IMC – Boosted IMC (BIMC):

$$A = UV^T + XWH^T Y^T$$

1. Learn U and V using standard MC
2. Train IMC on the residual $R = A - UV^T$



$$A_{ij} - (UV^T)_{ij} = \mathbf{x}_i^T Z \mathbf{y}_j$$

- First try to capture links of A with U and V, and then focus on the part it can not accurately model with IMC

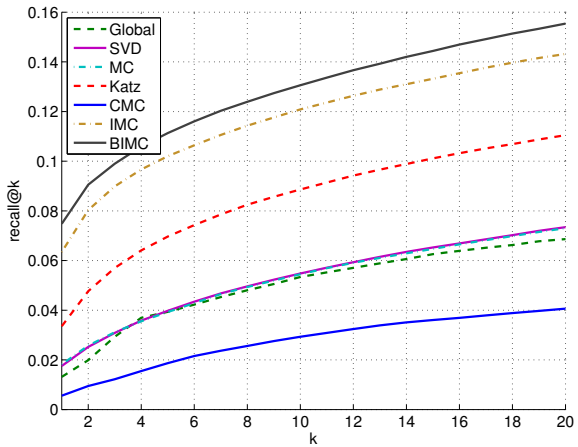
Experimental Results

- 1M randomly sampled users and blogs

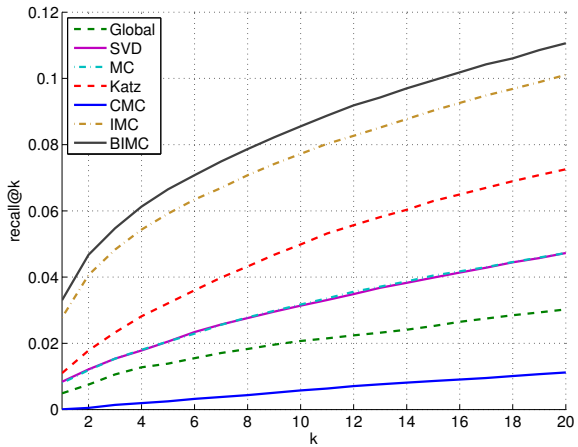
Method	PRC@10	RCL@10	AUC
Global	1.03%	4.80%	0.8687
SVD	1.28%	5.10%	0.8530
MC	1.28%	5.07%	0.8515
Katz	1.90%	8.15%	0.9209
CMC	0.49%	2.41%	0.8996
IMC	2.93%	11.33%	0.9075
BIMC	**3.21%**	**12.28%**	**0.9221**

User and Blog Groups

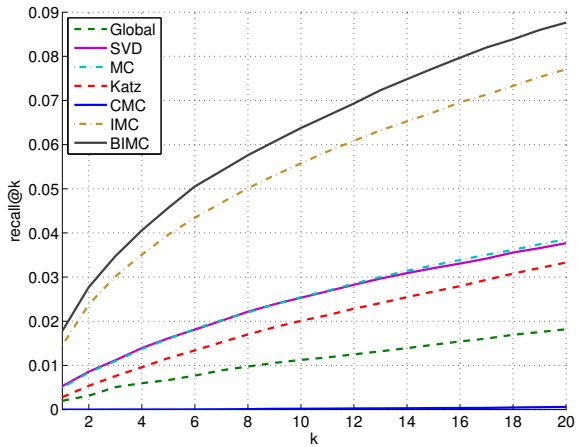
User groups with different number of followees ($n_f \leq 40$, $40 < n_f \leq 100$, $100 < n_f$)



(a) Low (89.4%)

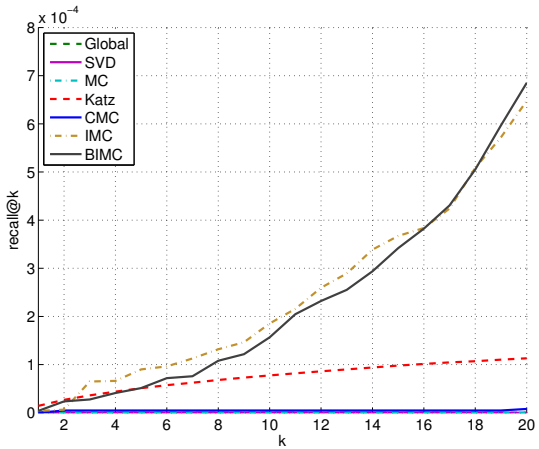


(b) Medium (7.8%)

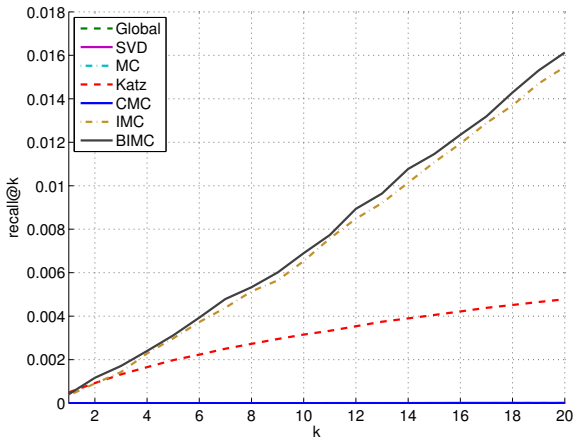


(c) High (2.8%)

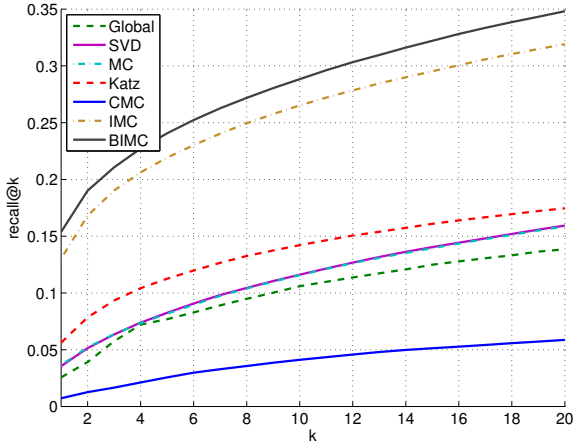
Blog groups with different number of followers ($n_f \leq 40$, $40 < n_f \leq 100$, $100 < n_f$)



(d) Low (95.1%)

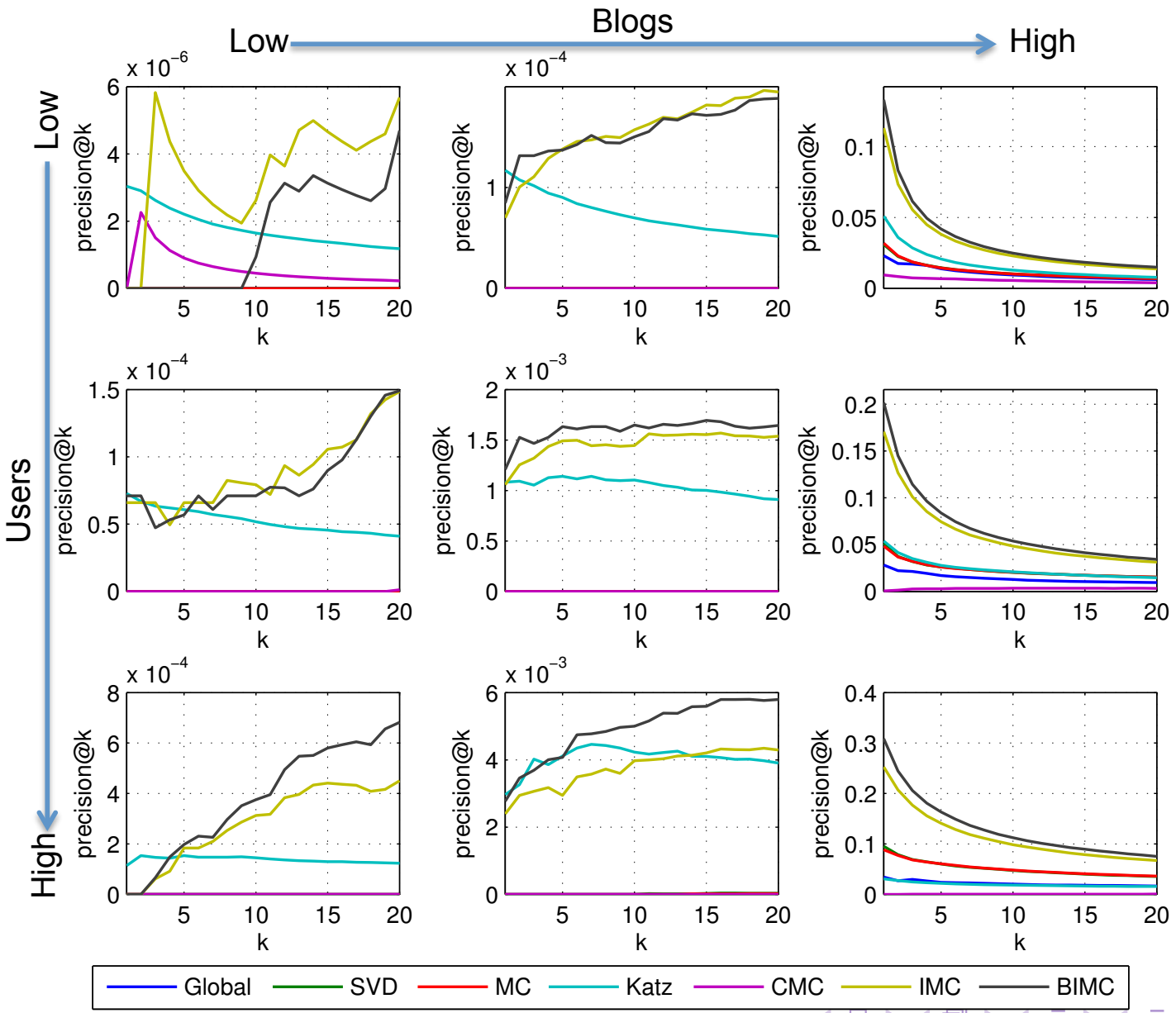


(e) Medium (3.3%)

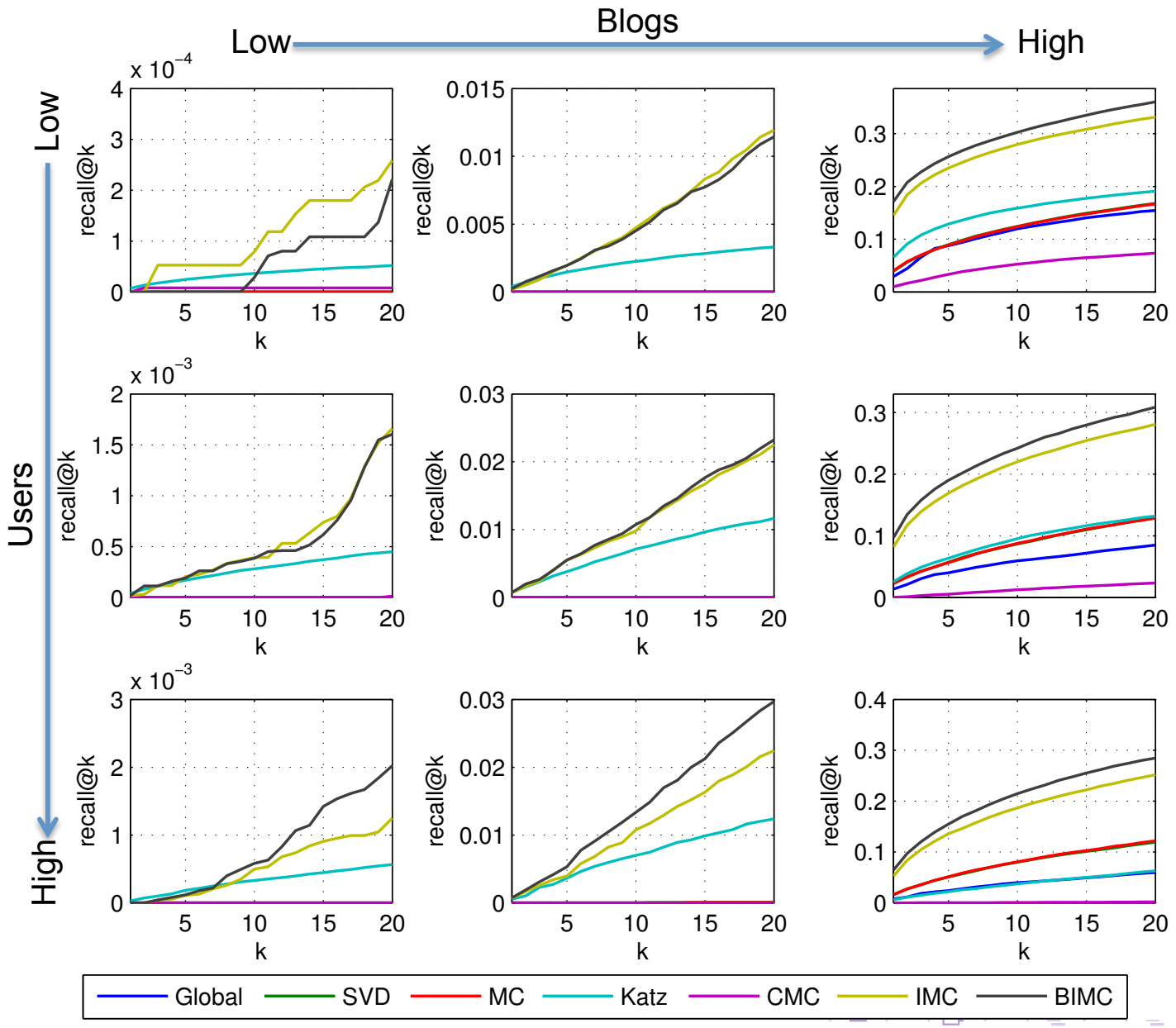


(f) High (1.6%)

Precision@k


Blogs
Low
High
Users
Low
High
precision@k
k
x 10^-6
x 10^-4
x 10^-3
Global
SVD
MC
Katz
CMC
IMC
BIMC

Recall@k


Blogs
Low
High
Users
Low
High
recall@k
k
x 10^-4
x 10^-3
Global
SVD
MC
Katz
CMC
IMC
BIMC

Conclusion

Tumblr Blog Recommendation

- Incorporating deep learned features extracted from various side information (text, image and activity).
- Combining both MC and IMC yields the best recommendation performance.
  - Especially, when users/blogs have only a few or no followees/followers.
- Relation between IMC and CMC?
- Multi-scale extensions?

Thanks!

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. *Distributed Representations of Words and Phrases and their Compositionality*. NIPS 2013.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks.* NIPS 2012.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.* ICML 2014.

[4] A. P. Singh and G. J. Gordon. *Relational Learning via Collective Matrix Factorization.* SIGKDD 2008.

[5] G. Bouchard, D. Yin, and S. Guo. *Convex Collective Matrix Factorization.* AISTATS 2013.

[6] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang. *Consistent Collective Matrix Completion under Joint Low Rank Structure.* AISTATS 2015.

[7] P. Jain and I. S. Dhillon. *Provable Inductive Matrix Completion.* CoRR 2013.

[8] N. Natarajan and I. S. Dhillon. *Inductive Matrix Completion for Predicting Gene-disease Associations.* Bioinformatics 2014.