Heterogeneous Information Learning in Large-Scale Network

Xiao Huang,[†] Jundong Li,[‡] Na Zou,[†] and Xia (Ben) Hu[†]

[†]Texas A&M University, College Station, TX, USA
[‡]Arizona State University, Tempe, AZ, USA
Emails: {xhuang, nzou1, xiahu}@tamu.edu, jundongl@asu.edu

Roadmap

- Network Embedding

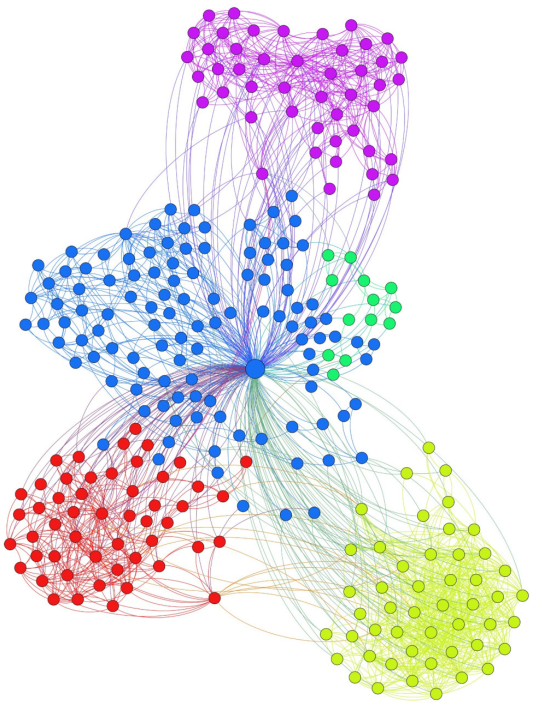
- Heterogeneous Information

- Challenges: Heterogeneity and Large Scale

- Proposed Framework *Heterogeneous Information Learning in Large-Scale Networks* (HILL)

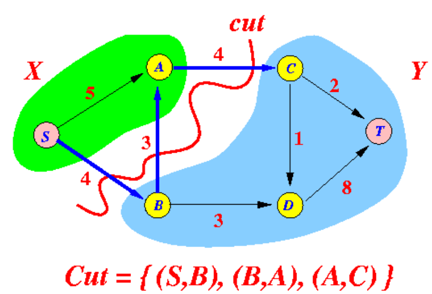
Traditional Network Analysis

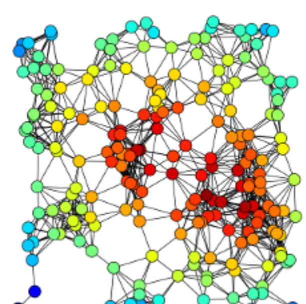
Network



Graph Theory

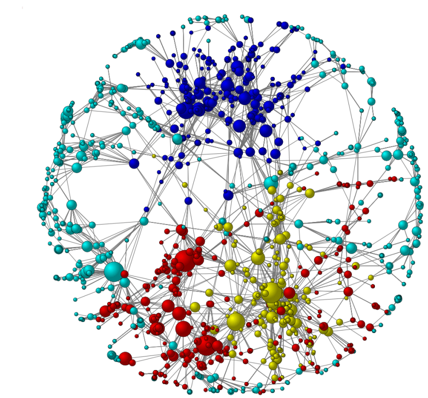
- Shortest path
- Maximum flow
- Graph partition
- Centrality
- …



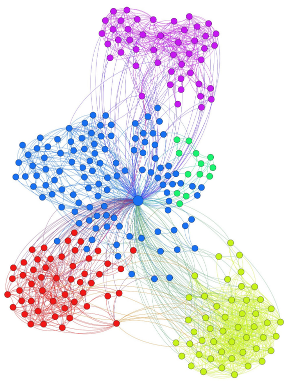


Tasks

- Clustering
- Link Prediction
- Classification
- Visualization
- …



Network Embedding



n-dimensional

Embedding Representation

$$\mathbf{H} = \begin{array}{c} \begin{bmatrix} 0.54 & 0.27 \\ 0.22 & 0.91 \\ 0.55 & 0.28 \\ 0.98 & 0.11 \\ 0.32 & 0.87 \\ 0.26 & 0.11 \end{bmatrix} \end{array} \begin{array}{c} n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{array}$$

$\mathrm{d} \ll n$

Machine Learning

Tasks

- Classification
- Clustering
- Link Prediction
- Visualization
- …

➢ To take advantage of machine learning, it learns a low-dimensional vector representation for each node, to preserve the geometrical structure **G**.

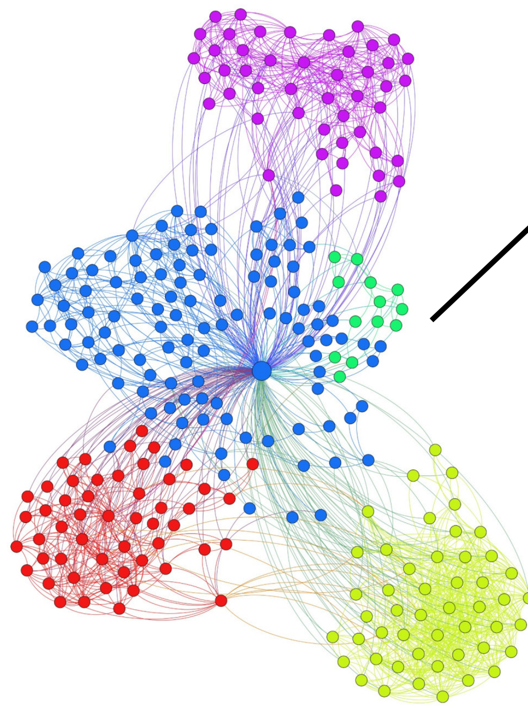
➢ Nodes with similar structure $\longrightarrow$ similar vectors.

➢ **H** benefits real-world applications.

Examples of Node Attributes

- Examples: user content in social media, reviews in co-purchasing networks, & paper abstracts in citation networks.
- Rich node attributes are available.

Attributed Networks



Nodes Have Different Attributes

- Nodes are not just vertices.
- Node attributes: a rich set of data that describes the unique features of each node.

Heterogeneous Information

- Nodes are accompanied with other types of meaningful information.
  - Node attributes
  - Second-order proximity
  - Link directionality
- Incorporating it into network embedding is potentially helpful in learning better vector representations.

Node Attributes Benefit Embedding

- Node attributes are informative.
- Network and node attributes influence each other and are inherently correlated. (Homophily & social influence)
  - High correlation of user posts and following relationships
  - Strong association between paper topics and citations

Attributes & Network are Correlated

Dataset	Scenarios	CorrCoef	Intersect	p-value
BlogCatalog	Real-world	**3.69e-002**	**42**	**0.00e-016**
	RandomMean	3.14e-005	7.32	0.18
	RandomMax	1.40e-003	13	4.42e-016
Flickr	Real-world	**1.85e-002**	**25**	**0.00e-016**
	RandomMean	2.15e-005	3.56	0.49
	RandomMax	5.48e-004	9	3.37e-003

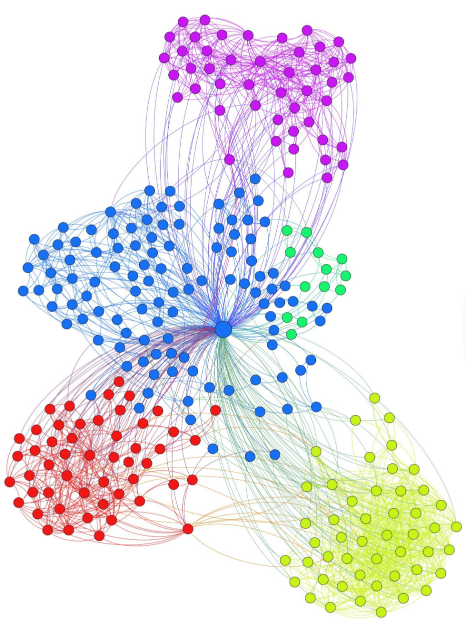
- Hypothesis: there is no correlation between network affinities and node attribute affinities.

- Real-world networks vs randomly-generated networks.
  Mean and max results on synthetic networks as baselines
  A significance level of 0.05

How to Incorporate the Heterogeneous Information?

Heterogeneous Information, e.g., Node Attributes

Embedding Representation

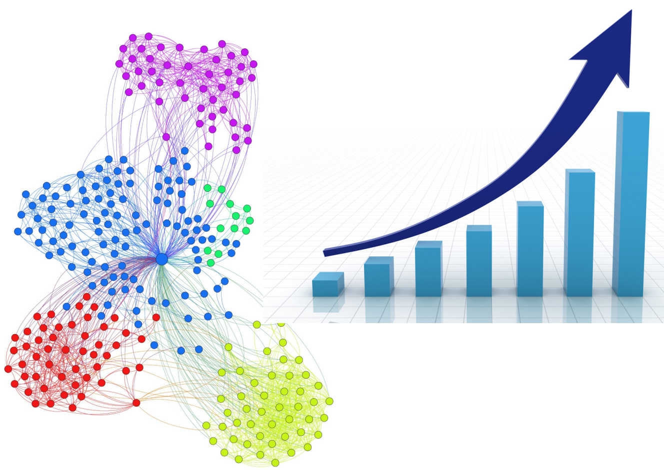


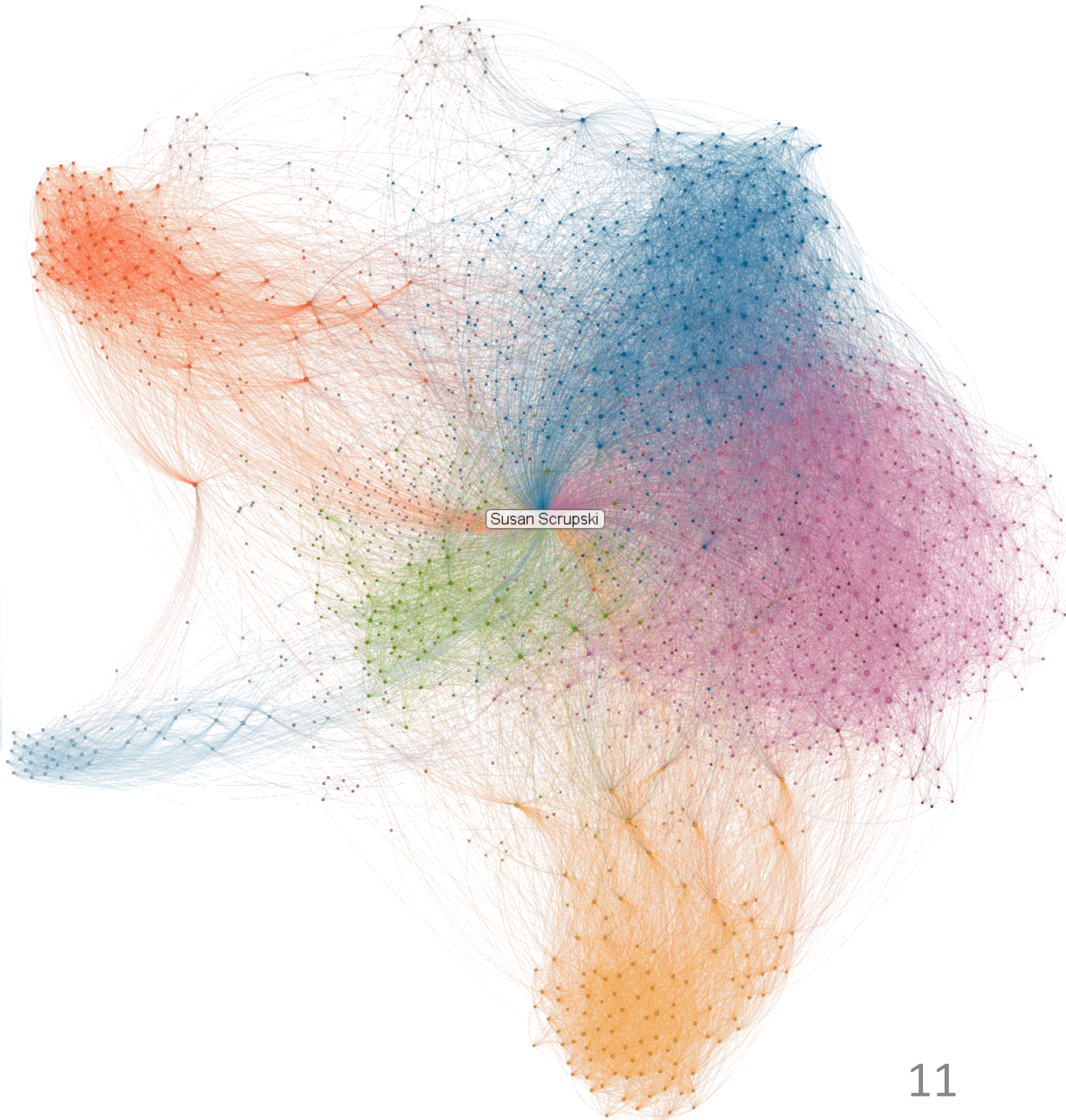
Network Embedding

$$\mathbf{H} = \begin{bmatrix} 0.54 & 0.27 \\ 0.22 & 0.91 \\ 0.55 & 0.28 \\ 0.98 & 0.11 \\ 0.32 & 0.87 \\ 0.26 & 0.11 \end{bmatrix} \begin{array}{c} n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{array}$$

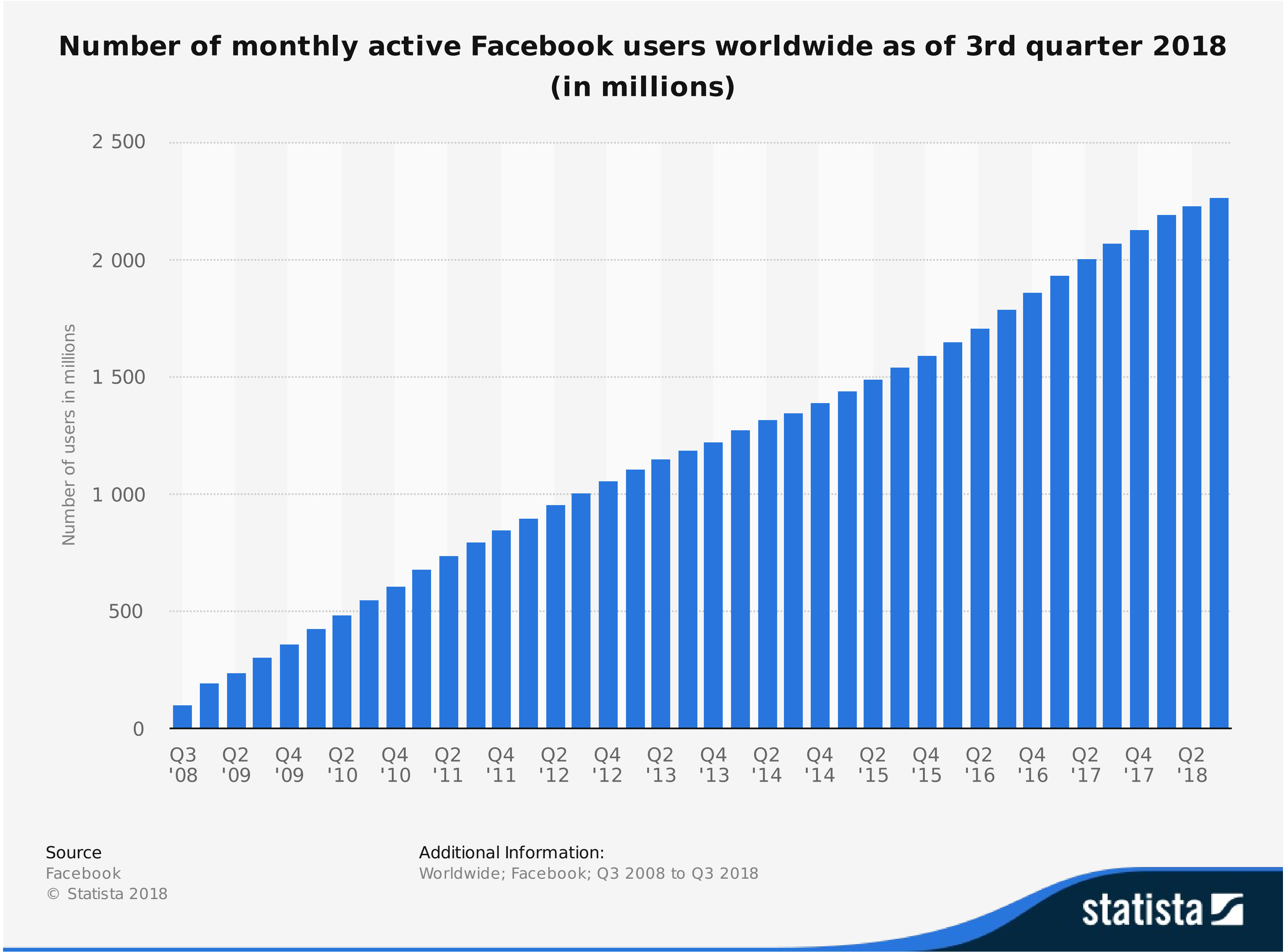
$d \ll n$

What if We Have a Large Network?

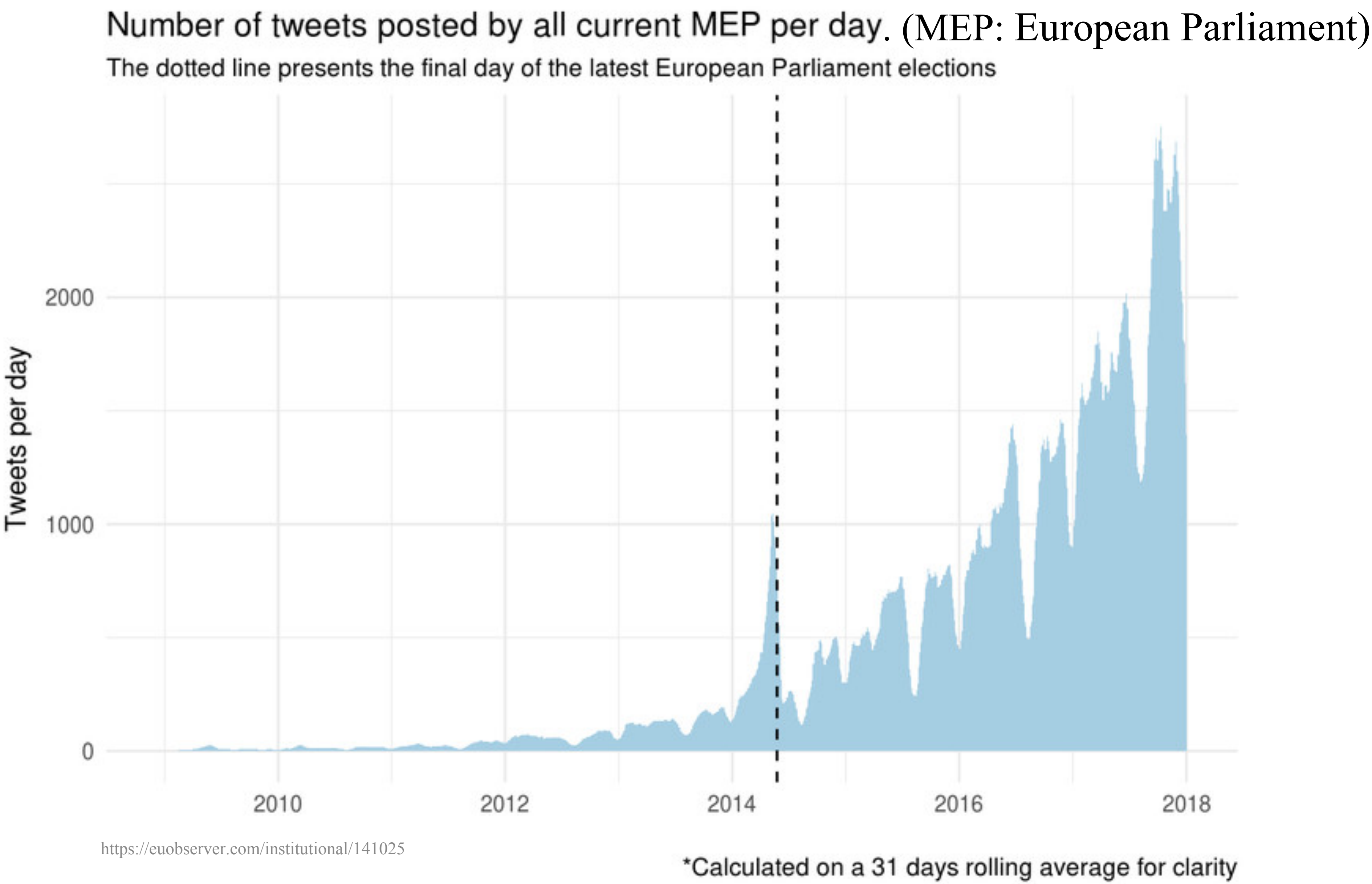




Real-world Attributed Network are Large



Real-world Node Attributes are High-dimensional

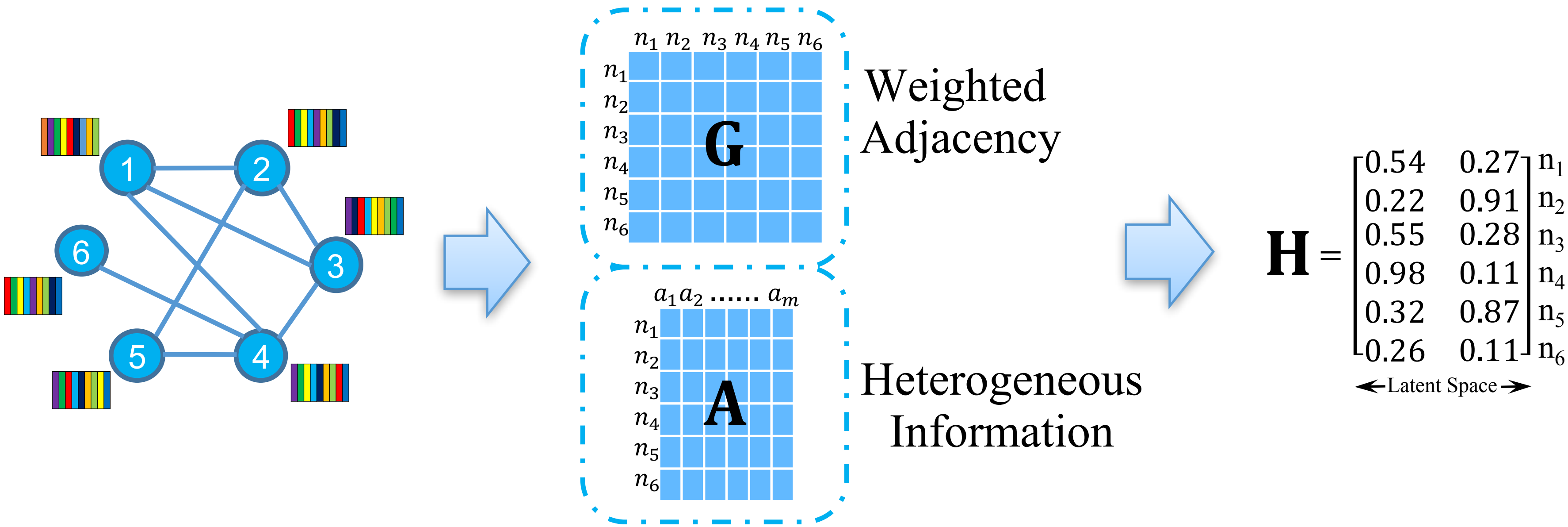


https://euobserver.com/institutional/141025

Challenges

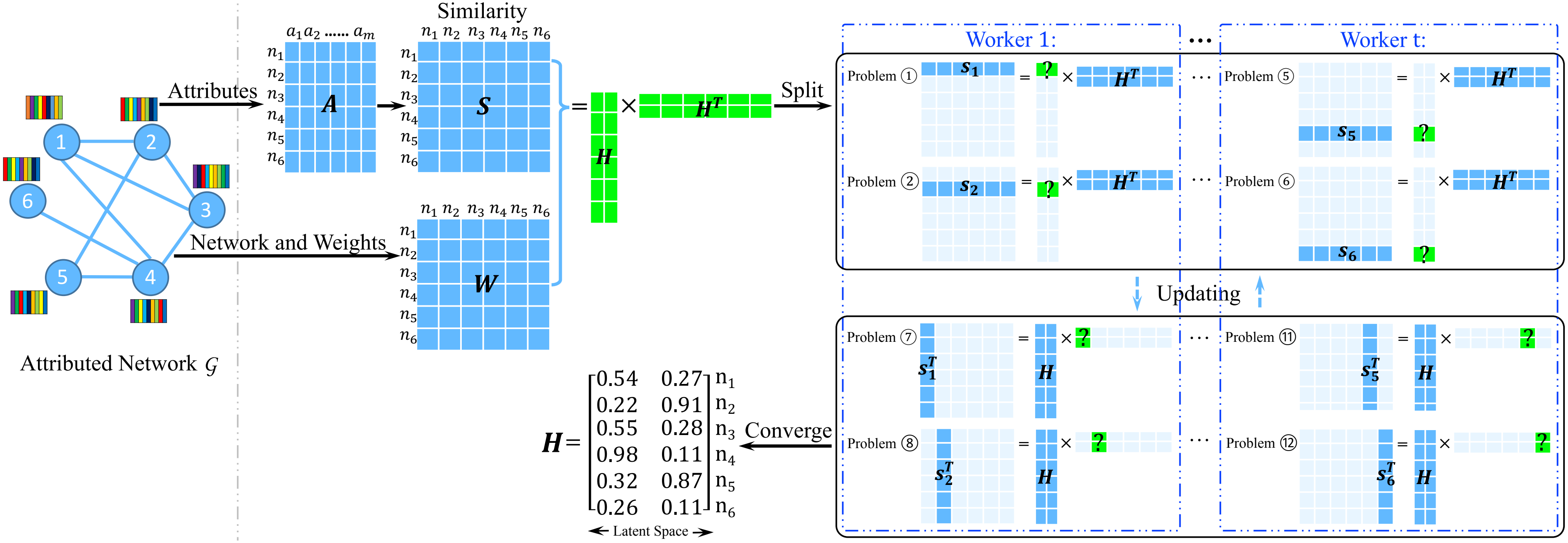
- Hard to jointly assessing node proximity from heterogeneous information.
  - Node attribute information such as paper abstracts and user posts is distinct from network topological structure
  - Data could be sparse, incomplete and noisy
- Number of nodes and dimension of attributes could be large.
  - Classical algorithms such as eigen-decomposition and gradient descent cannot be applied
  - It could be expensive to store or manipulate the high-dimensional matrices such as node attribute similarity

Heterogeneous Information Learning with Joint Network Embedding



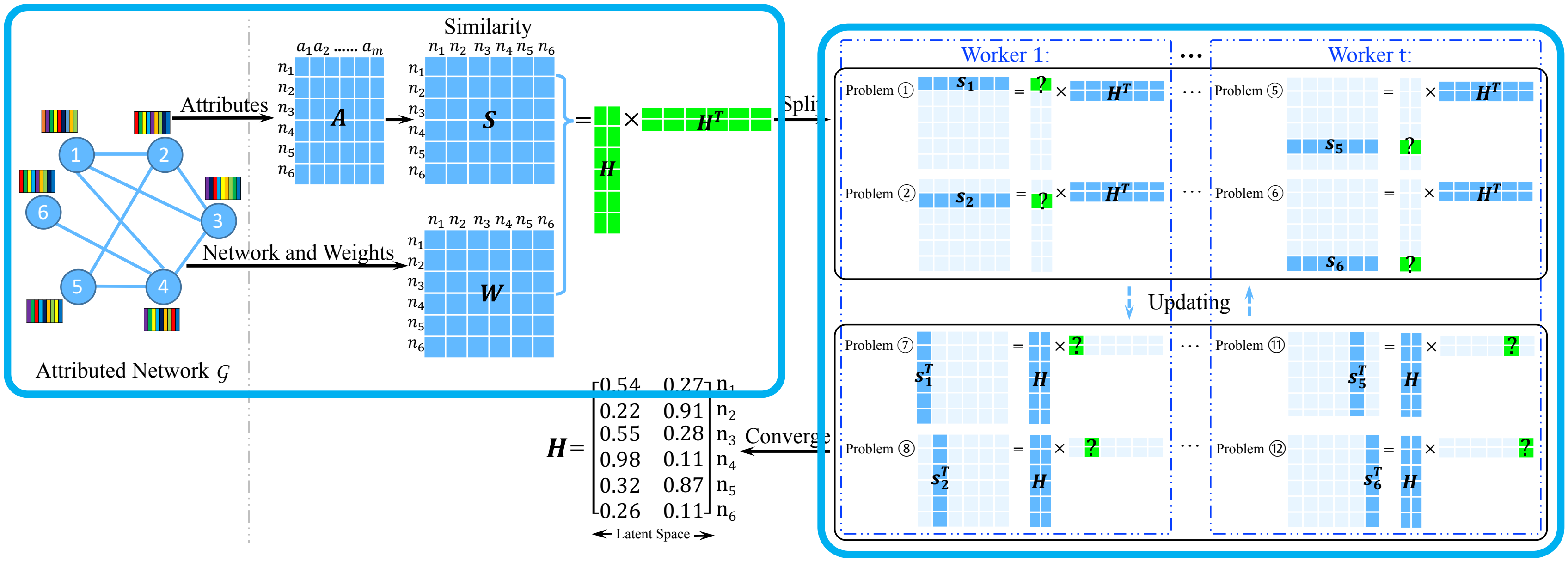
- Given **G** and **A**, we aim to represent each node as a d-dimensional row $\mathbf{h}_i$, such that **H** can preserve node proximity both in network and the heterogeneous information.
- Examples of **A**: node attributes, second-order proximity, link directionality.

Framework HILL



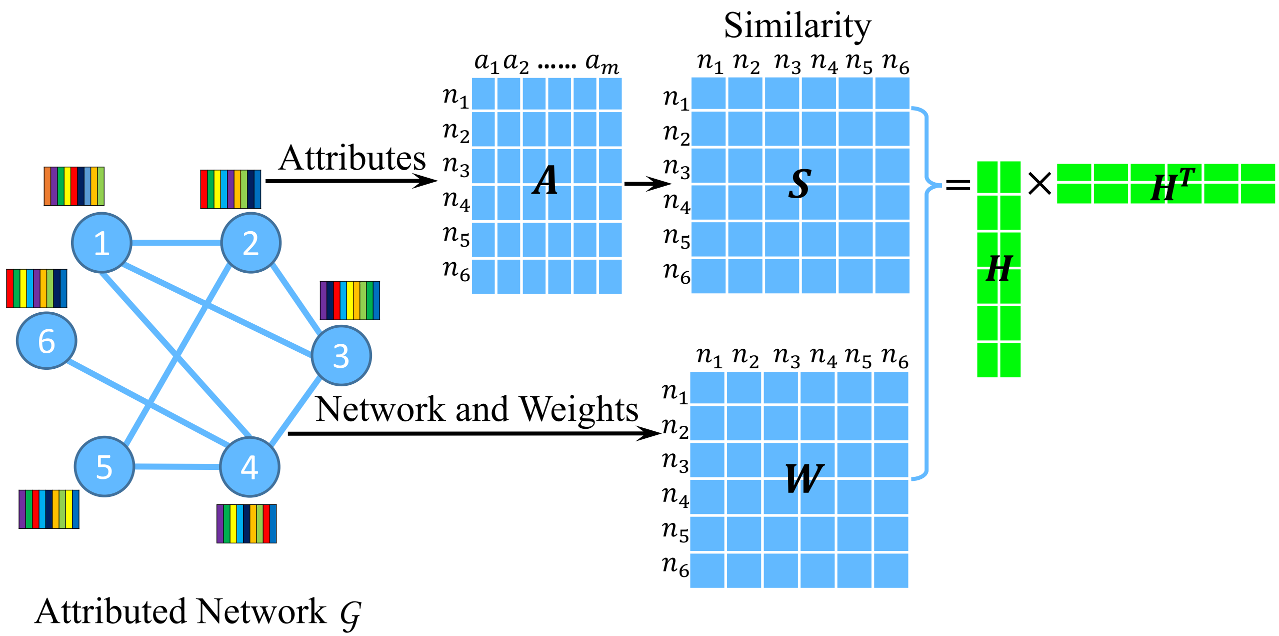
- A General Embedding Framework for Heterogeneous Information Learning in Large-Scale Networks, TKDD 2018.
- HILL accelerates the optimization by decomposing it into low complexity sub-problems.

Strategies of HILL



1) Assimilate the two info in the similarity space to tackle heterogeneity, but without calculating network similarity matrix.
2) Avoid high-dimensional matrix manipulation.
3) Make sub-problems independent to each other to allow parallel computation.

Strategy 1. Incorporating Node Similarities



➢ Based on the decomposition of attribute similarity and penalty of embedding difference between connected nodes.

$$\min_{\mathbf{H}} \ \mathcal{J} = \|\mathbf{S} - \mathbf{H}\mathbf{H}^\top\|_{\mathrm{F}}^2 + \lambda \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2$$

- ℓ_2 norm alleviates the impacts from outliers and missing data.
- Fused lasso clusters the network without similarity matrix.
- λ adjusts the size of clustering group.

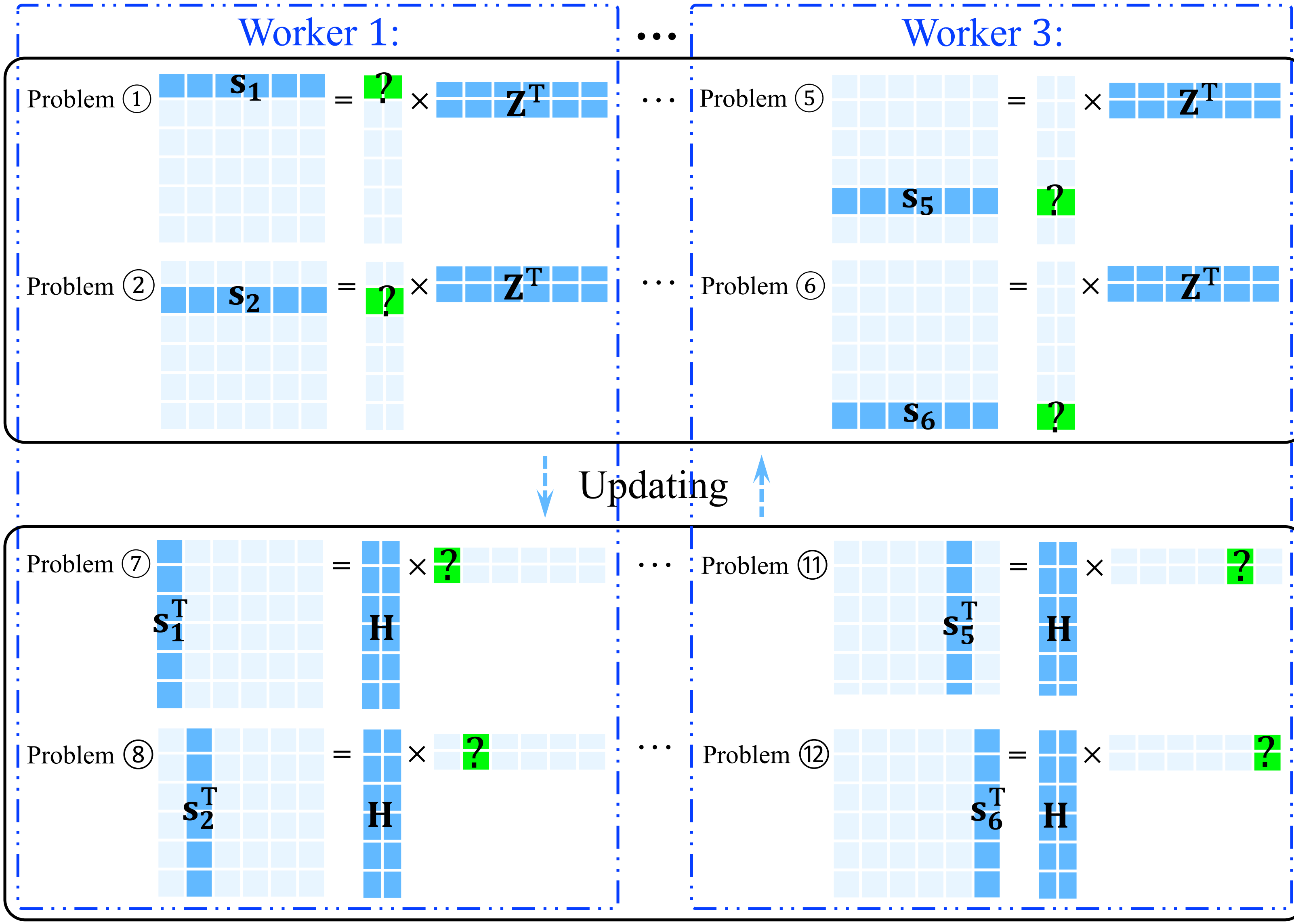
Strategy 2. Avoid High-dimensional Matrix Manipulation

- Make a copy of **H** and reformulate into a linearly constrained problem.

$$
\begin{aligned}
\min_{\mathbf{H}} \quad & \sum_{i=1}^{n} \|\mathbf{s}_i - \mathbf{h}_i \mathbf{Z}^\top\|_2^2 + \lambda \sum_{(i,j)\in\mathcal{E}} w_{ij} \|\mathbf{h}_i - \mathbf{z}_j\|_2, \\
\text{subject to} \quad & \mathbf{h}_i = \mathbf{z}_i, \; i = 1, \ldots, n.
\end{aligned}
$$

- Given fixed **H**, all the row $\mathbf{z}_i$ could be calculated independently.
- Each sub-problem only needs row $\mathbf{s}_i$, not the entire **S**.
- Time complexity of updating $\mathbf{h}_i$ is $\mathcal{O}(d^3 + dn + d|N(i)|)$, with space complexity $\mathcal{O}(n)$.
- Alternating Direction Method of Multipliers (ADMM) converges to a modest accuracy in a few iterations.

Strategy 3. Enabling Parallel Computation



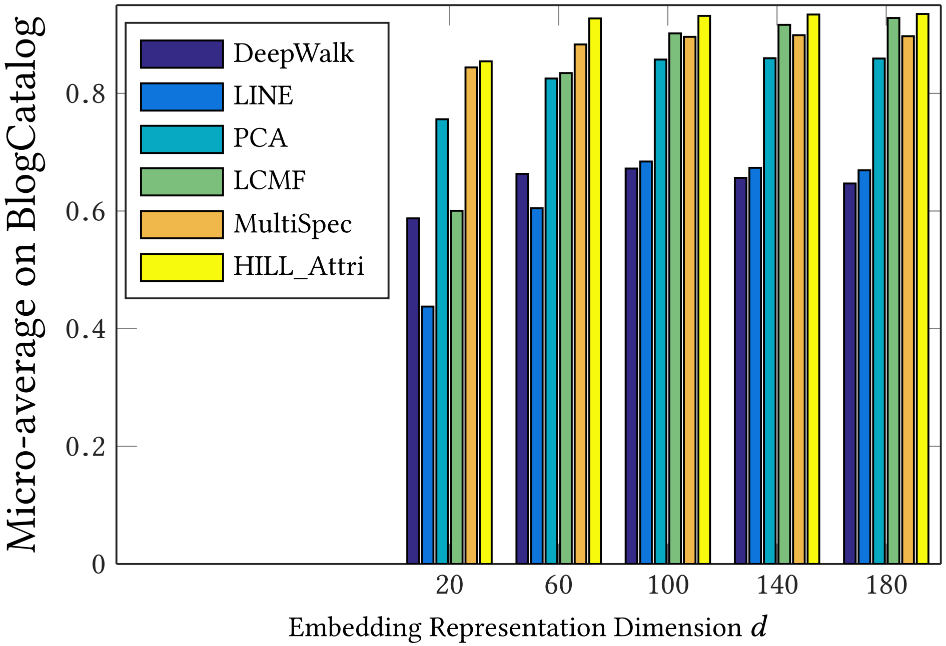
Experimental Settings

- Classification on three real-world network.
  - BlogCatalog (5,196 nodes)
  - Flickr (7,564 nodes)
  - Yelp (249,012 nodes, 1,779,803 edges, 20,000 attribute categories, 47,216,356 entities)

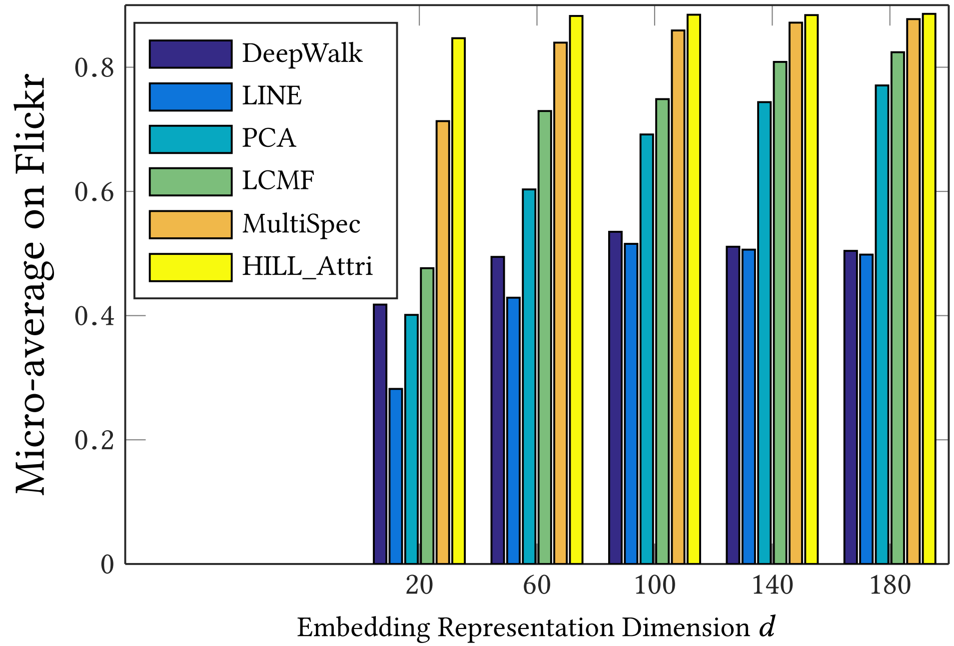
- Three types of baselines.
  - Scalable network embedding: DeepWalk & LINE.
  - Node attribute modeling based on PCA.
  - Attributed network representation learning: MultiSpec & LCMF.

Codes and datasets are available at http://people.tamu.edu/~xhuang/Code.html

Effectiveness Evaluation



(a) BlogCatalog



(b) Flickr

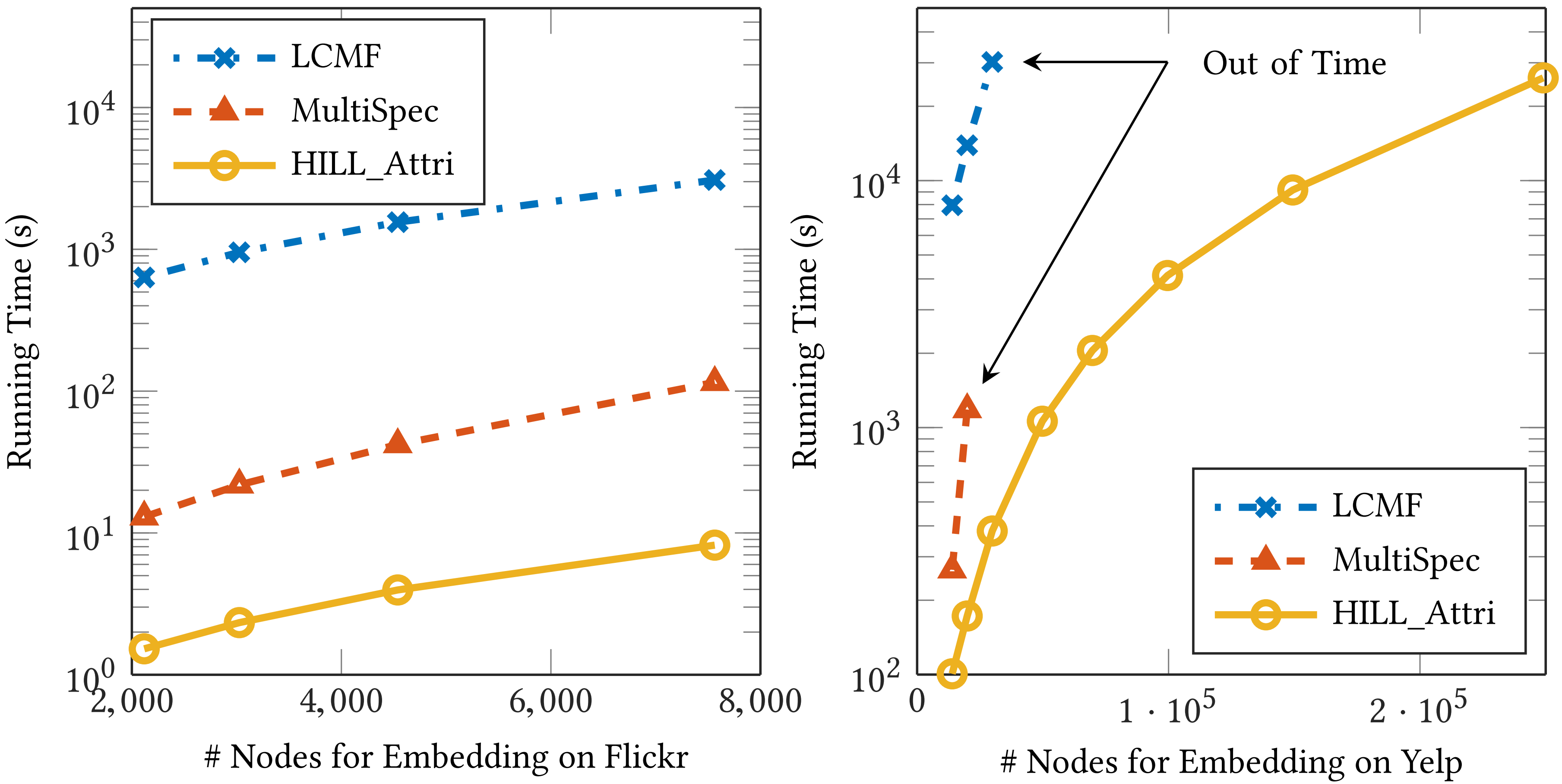
- HILL outperforms the state-of-the-art embedding algorithms with different latent dimension d.

Effectiveness Evaluation

		BlogCatalog				Flickr			
Training Set Percentage		10%	25%	50%	100%	10%	25%	50%	100%
# nodes for embedding		1,455	2,079	3,118	5,196	2,118	3,026	4,538	7,564
Micro-average	DeepWalk	0.491	0.551	0.611	0.672	0.312	0.373	0.465	0.535
	LINE	0.433	0.545	0.624	0.684	0.259	0.332	0.421	0.516
	HILL_Net	0.556	0.628	0.690	0.747	0.315	0.397	0.496	0.626
	PCA	0.695	0.782	0.823	0.857	0.508	0.606	0.666	0.692
	Spectral	0.717	0.791	0.841	0.869	0.698	0.771	0.813	0.846
	LCMF	0.778	0.849	0.888	0.902	0.576	0.676	0.725	0.749
	MultiSpec	0.678	0.788	0.849	0.896	0.589	0.720	0.800	0.859
	HILL_Attri	**0.841**	**0.878**	**0.913**	**0.932**	**0.740**	**0.811**	**0.854**	**0.885**
	HILL_Stream	0.770	0.822	0.887	0.914	0.568	0.726	0.816	0.859

- HILL_Net uses network only. It employs the second-order proximity of network as the heterogeneous information.
- HILL_Attri embeds attributed network.
- For HILL_Stream, test nodes come one by one.

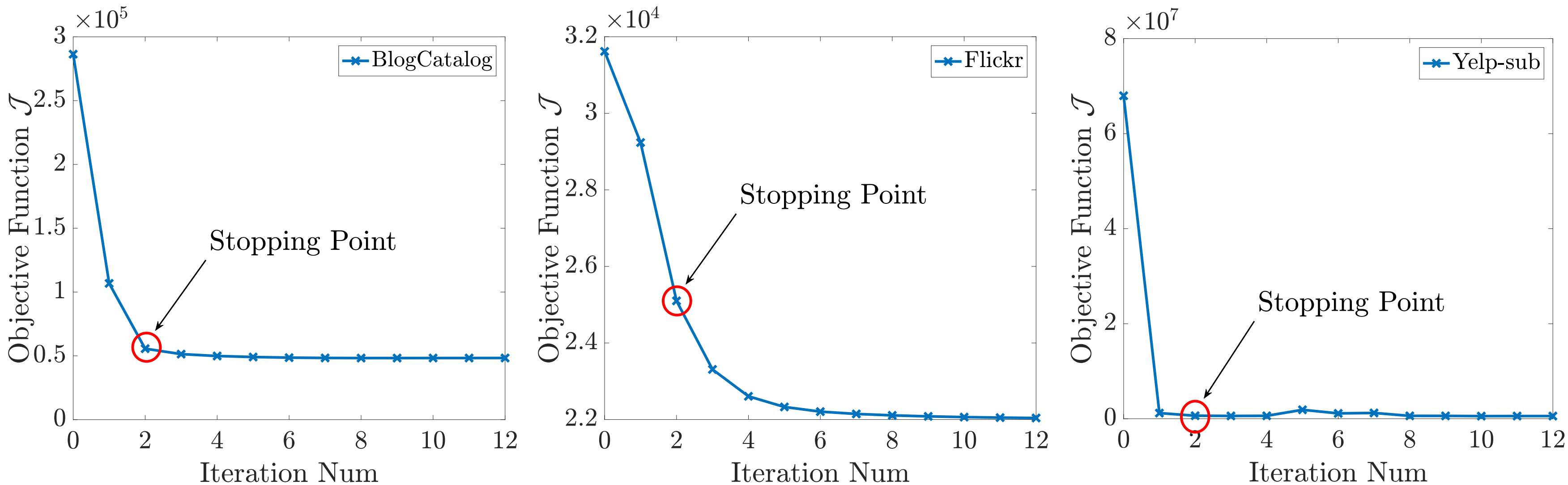
Efficiency Evaluation



- ➢ HILL takes much less running time than the attributed network representation learning methods even with single-thread.

Efficiency Evaluation

	BlogCatalog (sec)	Flickr (sec)	Yelp-sub (sec)
$c = 1$	26.301	33.751	1065.033
$c = 2$	14.233 (−45.9%)	17.510 (−48.1%)	581.544 (−45.4%)



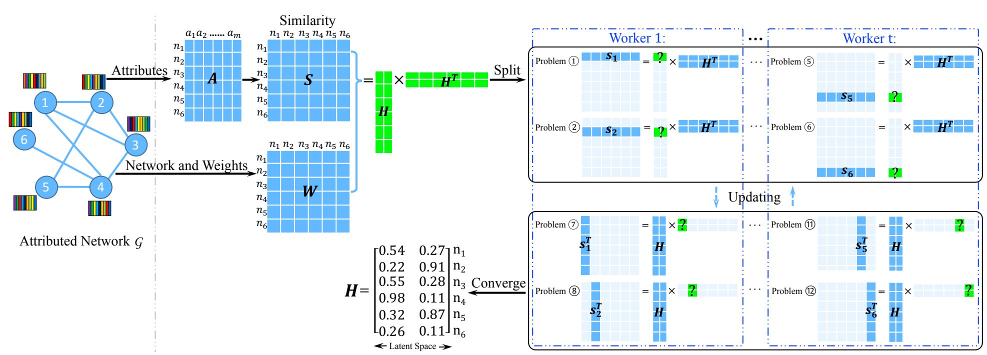
- ➢ Running time of HILL w.r.t. the number of workers c on a dual-core processor.
- ➢ One of the reasons HILL is efficient: it converges rapidly.

Conclusions

➢ Nodes are accompanied with other types of meaningful information.
  - Node attributes
  - Second-order proximity
  - Link directionality

➢ Challenges: Heterogeneity and Large Scale.

➢ HILL learns low-dimensional vectors to represent all nodes, such that the original network structure and the meaningful heterogeneous information are well preserved.



Acknowledgement

- DATA Lab and collaborators

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Data Analytics at Texas A&M (DATA Lab)

- Funding agencies
  - National Science Foundation
  - Defense Advanced Research Projects Agency

- Everyone attending the talk