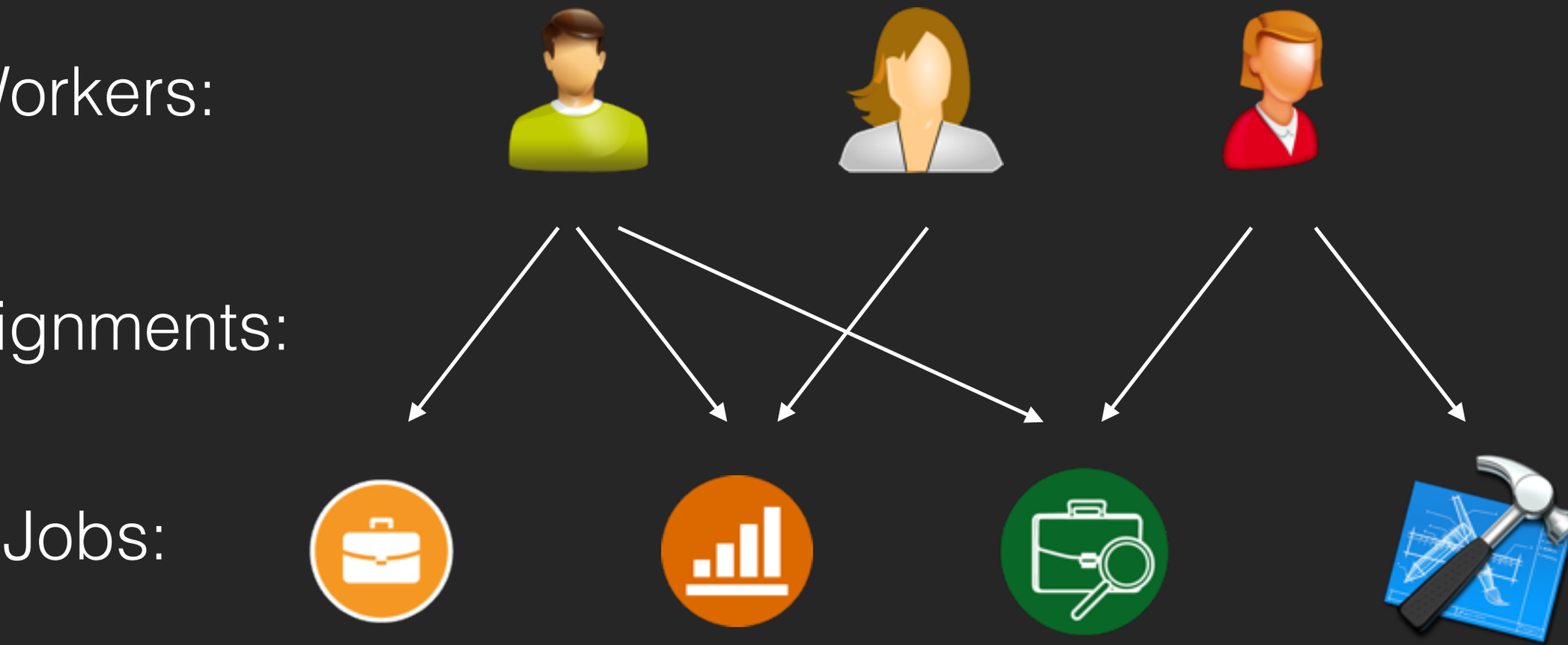
Effective Crowd Expertise Modeling via Cross Domain Sparsity and Uncertainty Reduction

Sihong Xie, Qingbo Hu, Weixiang Shao, Jingyuan Zhang,
Jing Gao, Wei Fan, Philip S. Yu



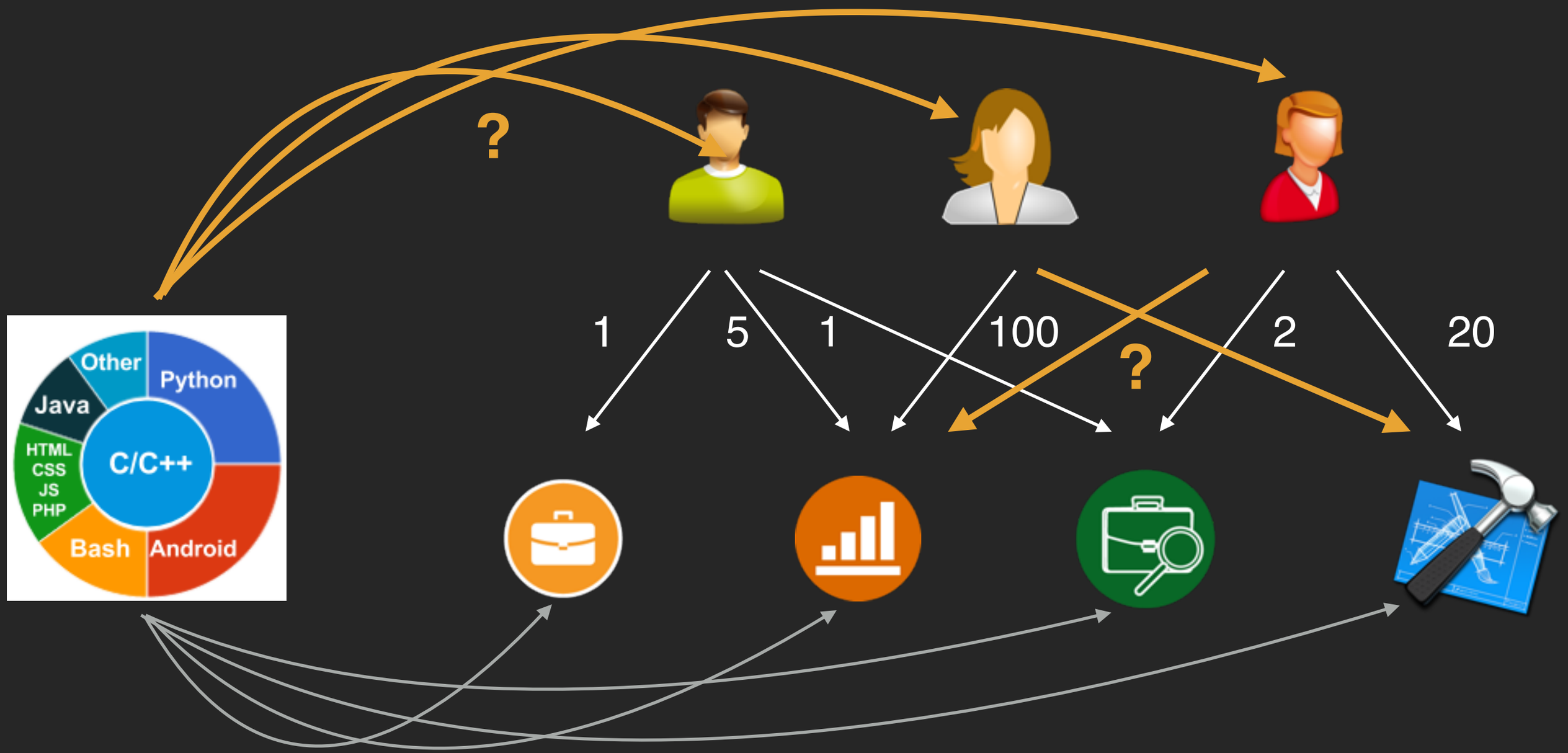

Motivations

Online job and crowdsourcing markets:

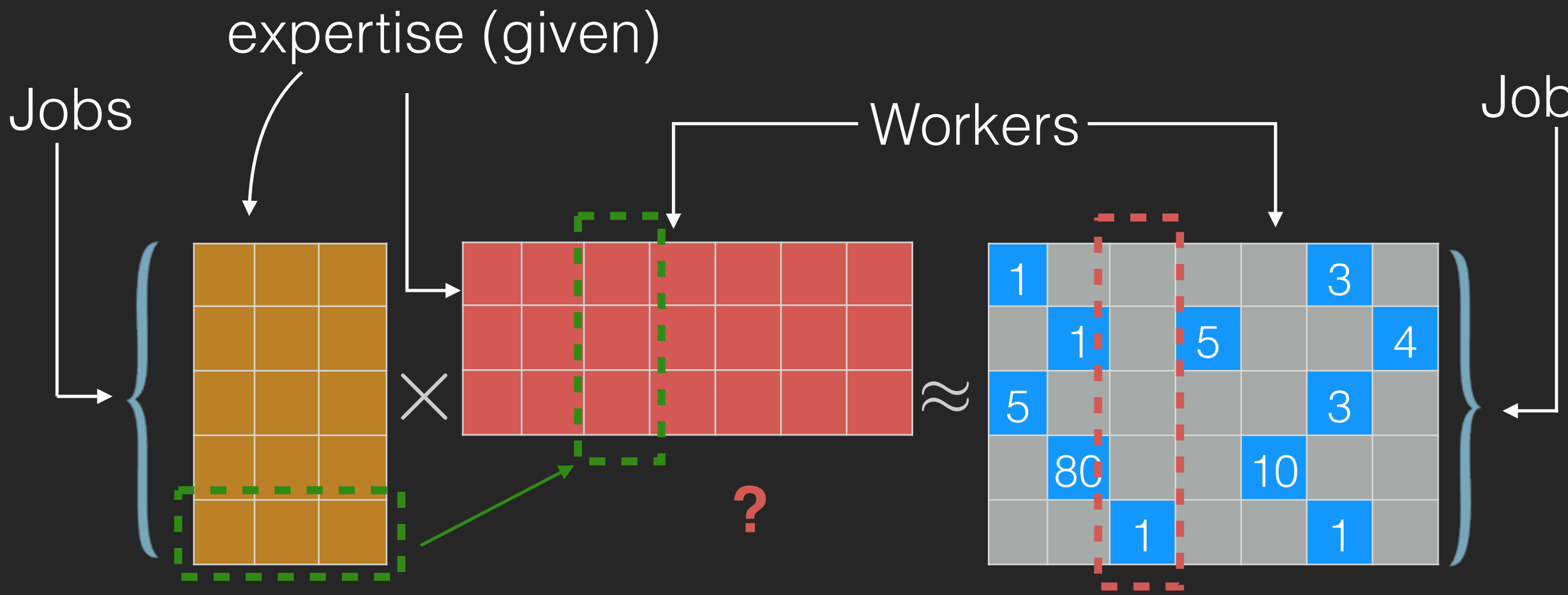


Motivations

Matching workers to jobs using expertise



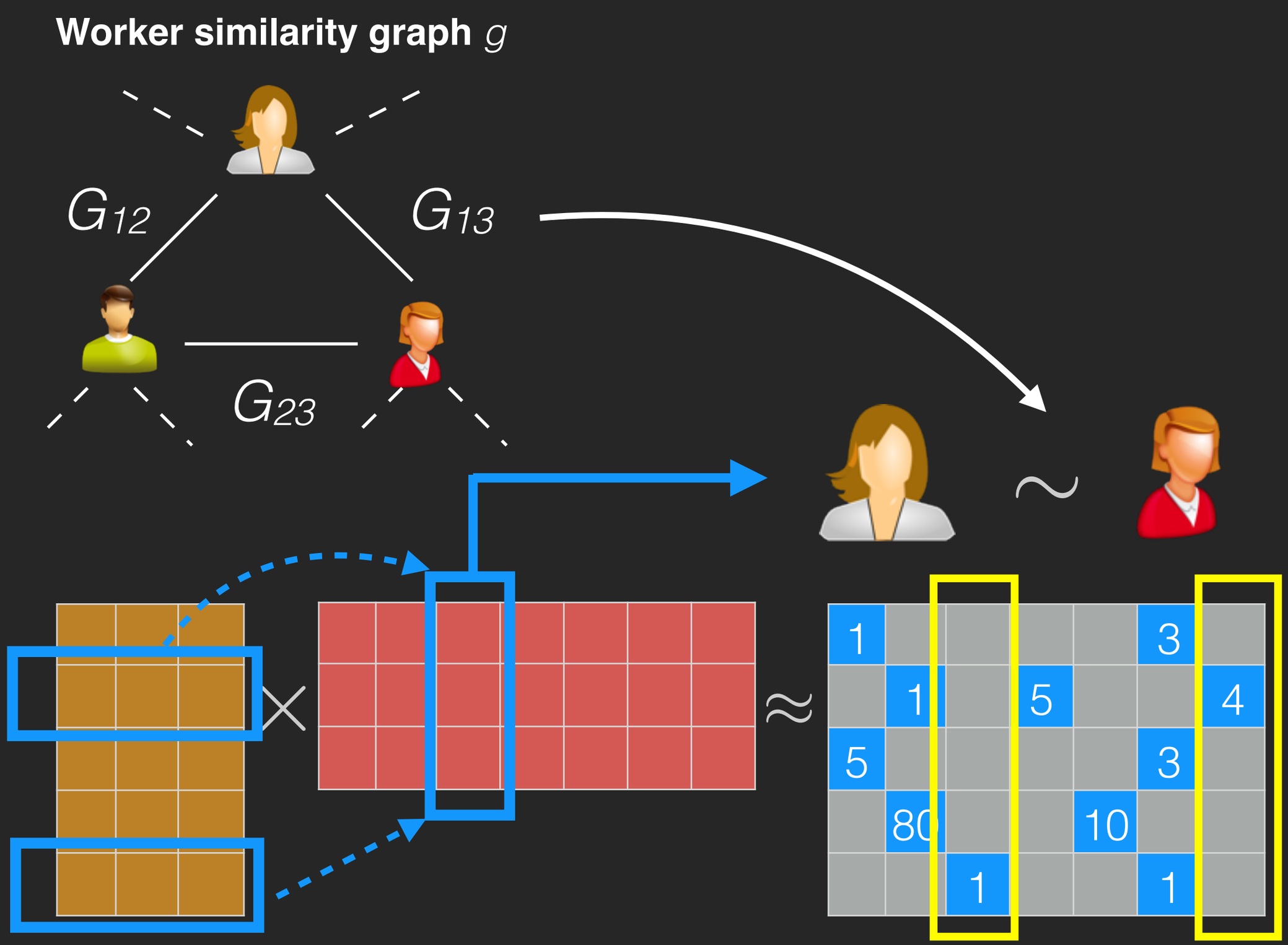
Problem formulation



suffer from sparsity? Yes

$$X \times B \approx Y$$

Graph-fused multi-task regression

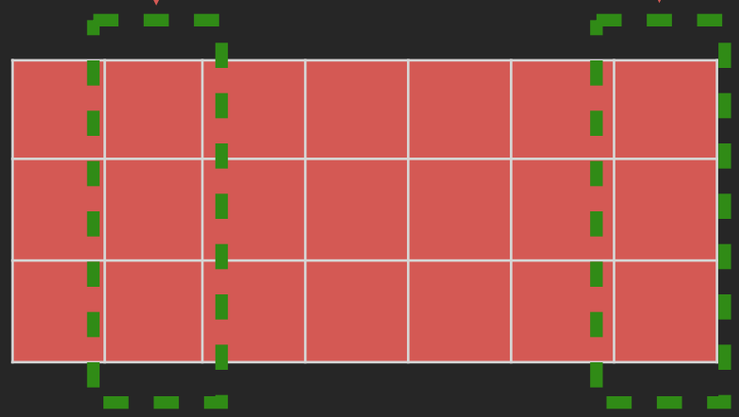


Graph-fused multi-task regression

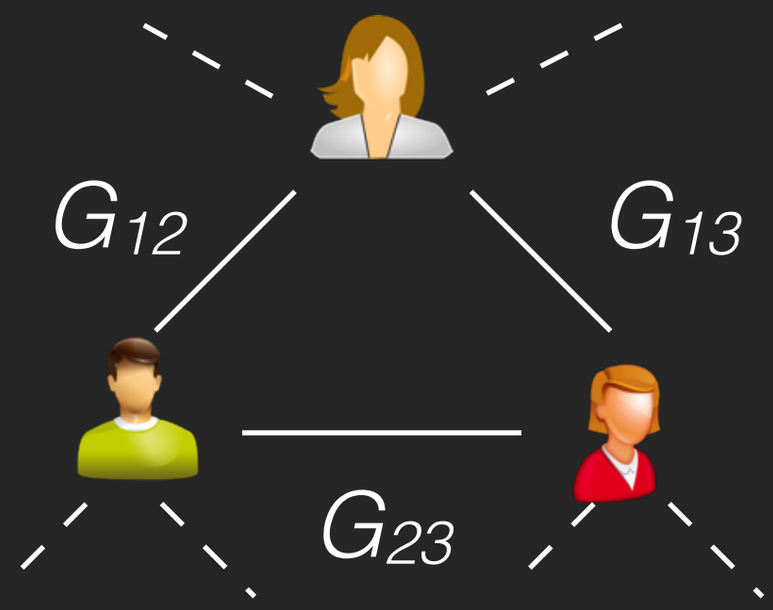
$$\min_{B \in \mathbb{R}^{K \times M}} \frac{1}{2}\|Y - XB\|_F^2 + \lambda \Omega_g(B)$$

$$\Omega_g(B) = \sum_{(i,j) \in G} |G_{ij}| \|\beta_i - \mathrm{sign}(G_{ij})\beta_j\|_1.$$

worker expertise matrix B:

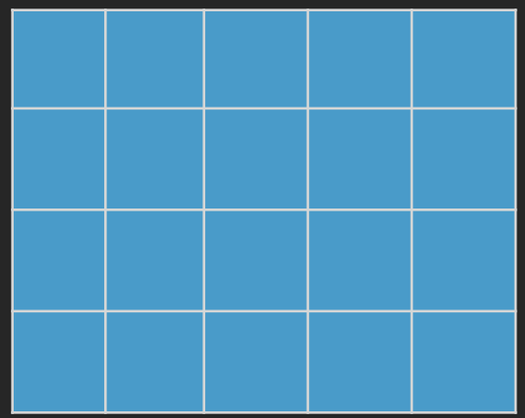


Worker similarity graph g

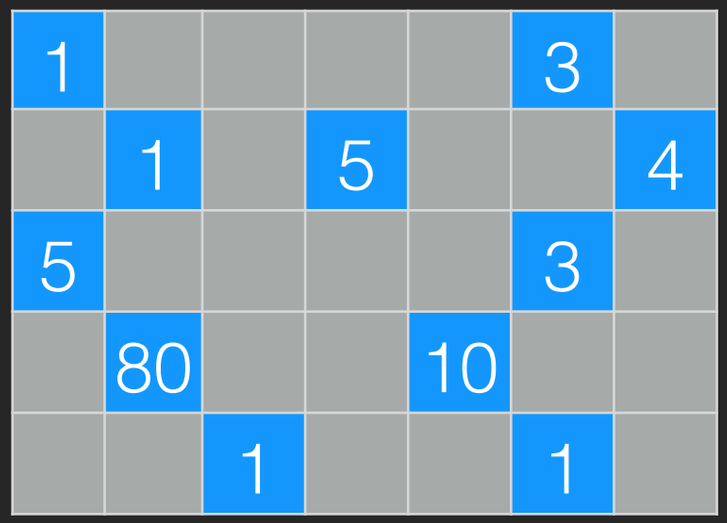


Building the worker similarity graph

Job profiles

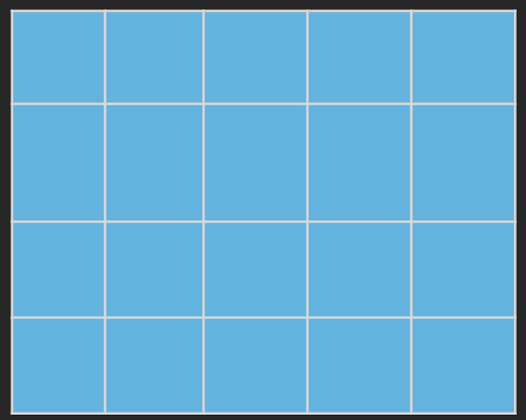


X



=

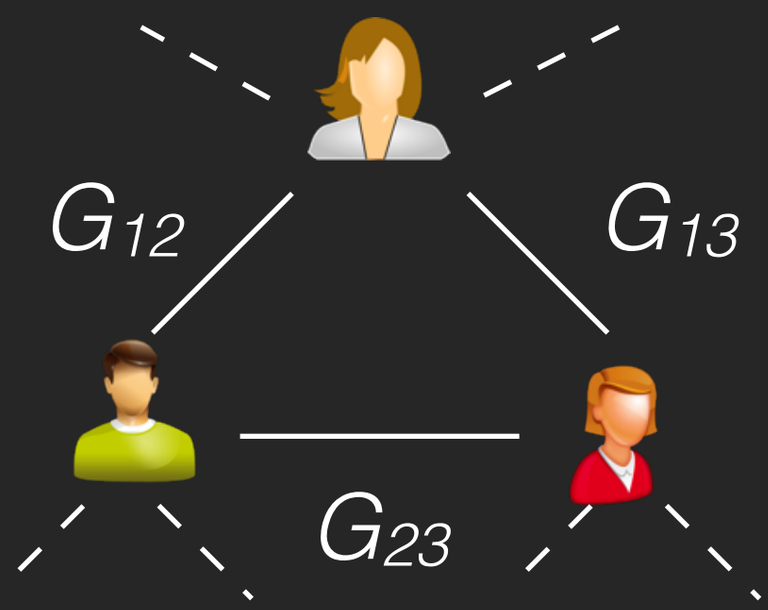
Worker profiles



suffer from sparsity? Yes

$$\min_{B \in \mathbb{R}^{K \times M}} \frac{1}{2} \|Y - XB\|_F^2 + \lambda \Omega_g(B)$$

Worker similarity graph *g*



An iterative sparsity reduction method

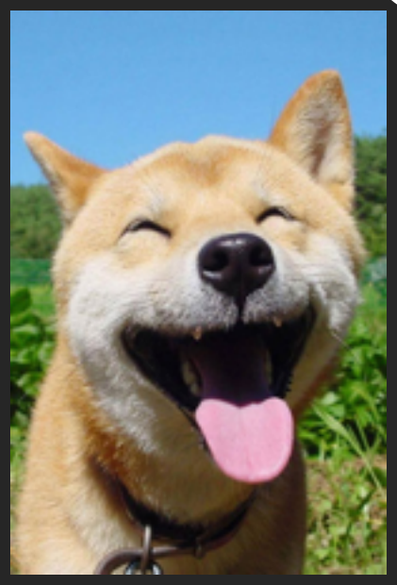
Worker profiles

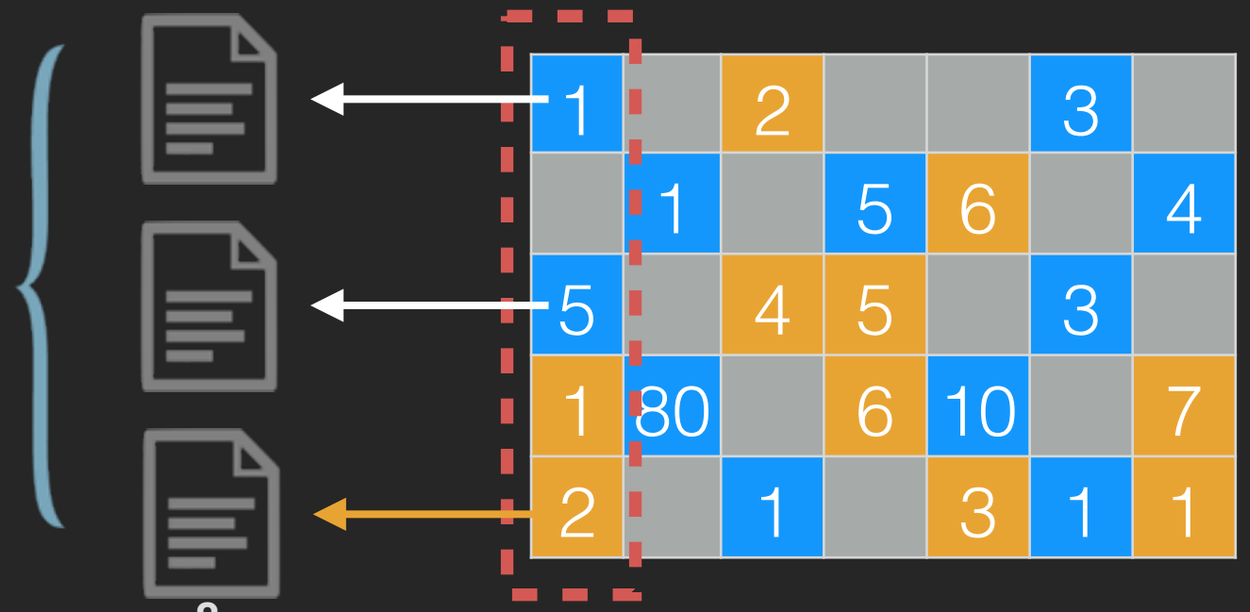
Worker similarity graph

Multi-task regression

$$\min_{B \in \mathbb{R}^{K \times M}} \frac{1}{2}\|Y - XB\|_F^2 + \lambda \Omega_g(B)$$

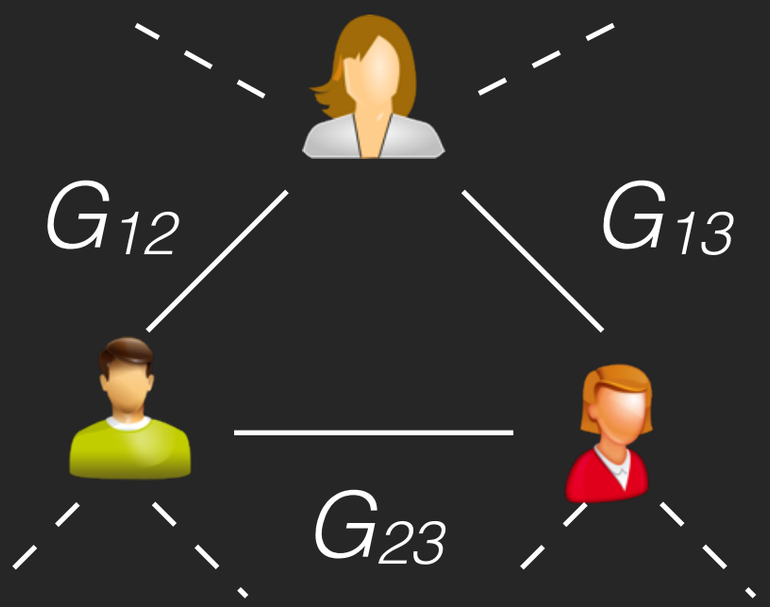
impute top-k missing entries





Scalability

Worker similarity graph g



$$\min_{B \in \mathbb{R}^{K \times M}} \frac{1}{2}\|Y - XB\|_F^2 + \lambda \Omega_g(B)$$

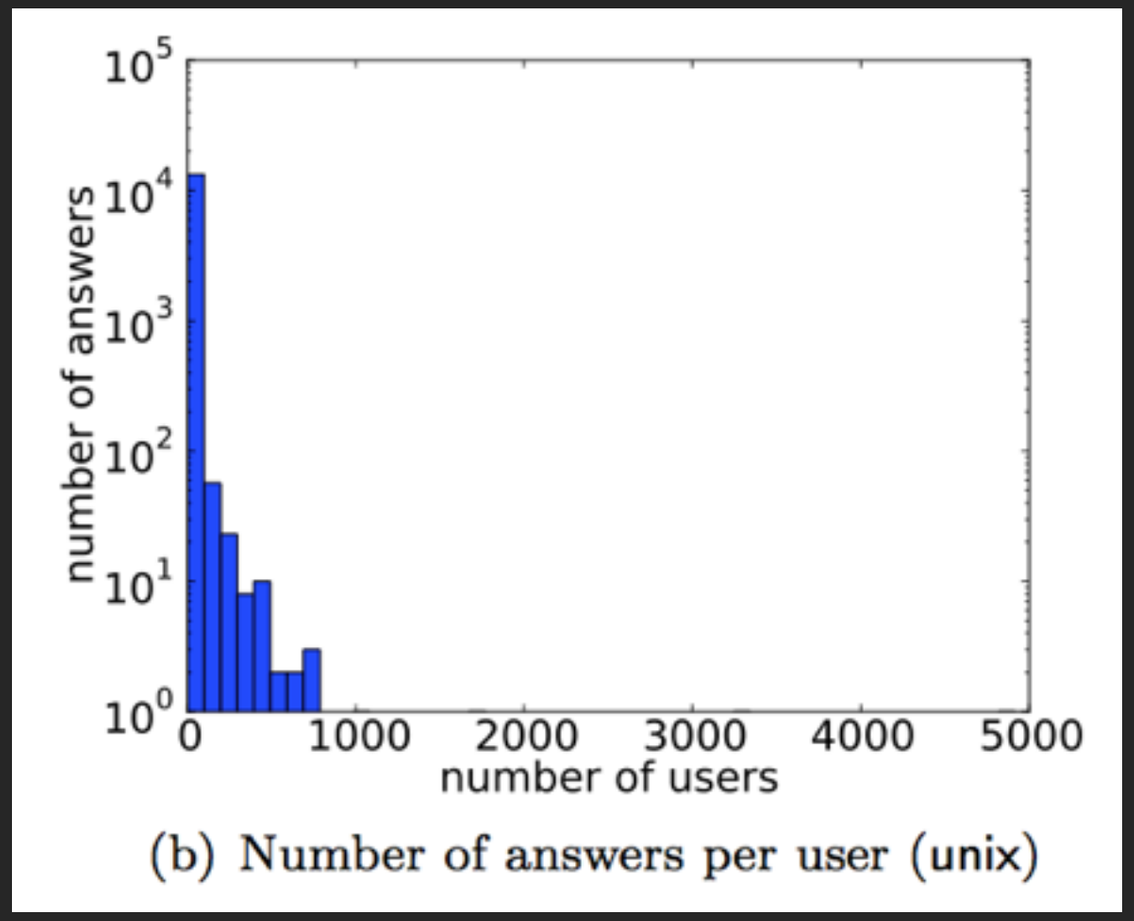
ADMM distributed computing

Graph sparsification

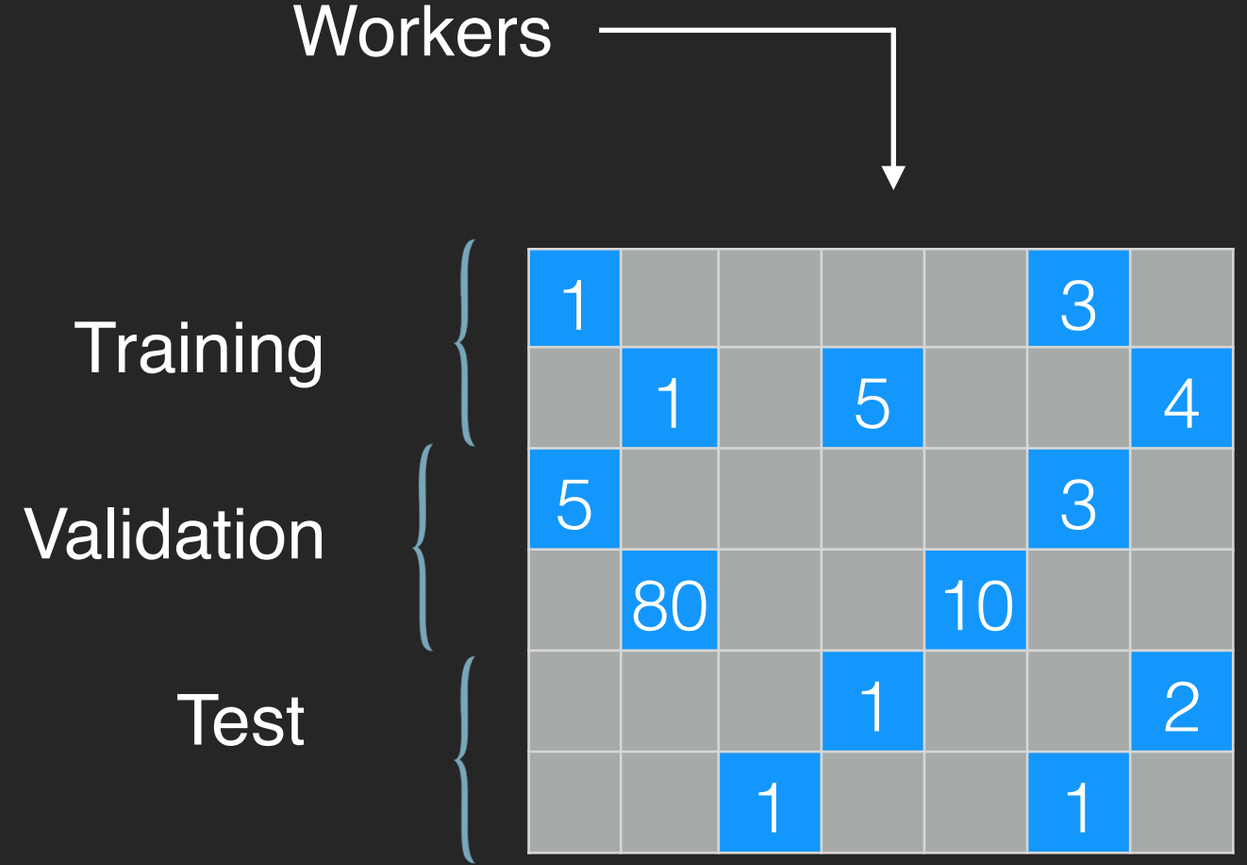
Experiments - datasets

Stackexchange: cstheory, unix and english.

1. thousands of workers; tens of thousands of jobs;
2. bag-of-words repr for the questions (jobs) -> LDA to expertise;
3. answerers (workers) profiles -> SVD -> worker similarity graph.



(b) Number of answers per user (unix)



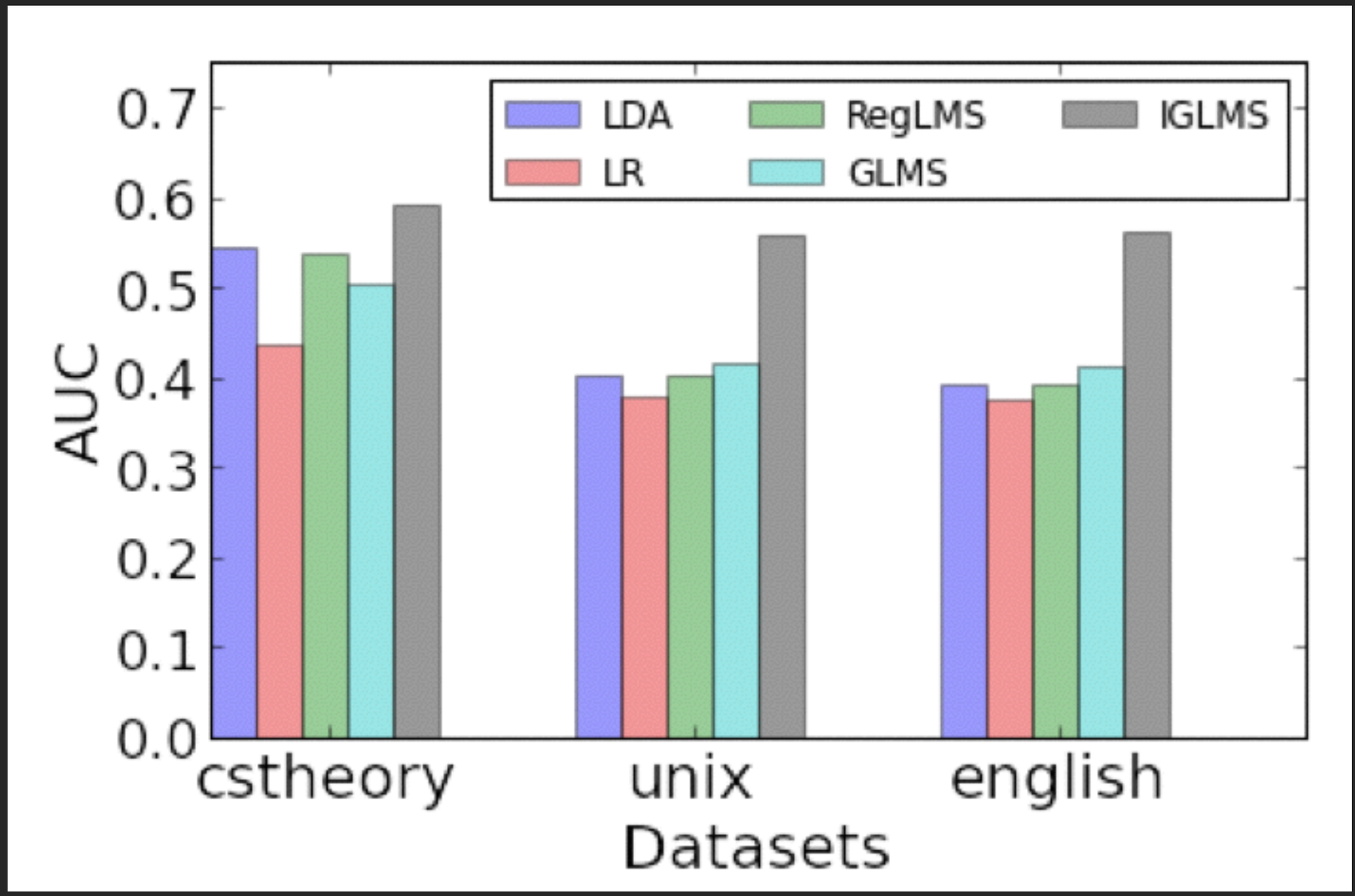
Experiments - baselines

1. LDA topics as expertise, *ignoring the relative quality*.
2. LR and RegLMS: Regression on expertise required by jobs (Logistic or Least Square), *ignoring inter-worker relations*.
3. GLMS: Graph-fused multi-task regression, *may suffer from sparsity of the auxiliary information.*

The proposed IGLMS considers:

1. the relative quality of the workers accomplishing their jobs and
2. side information while addressing the sparsity in both data sources.

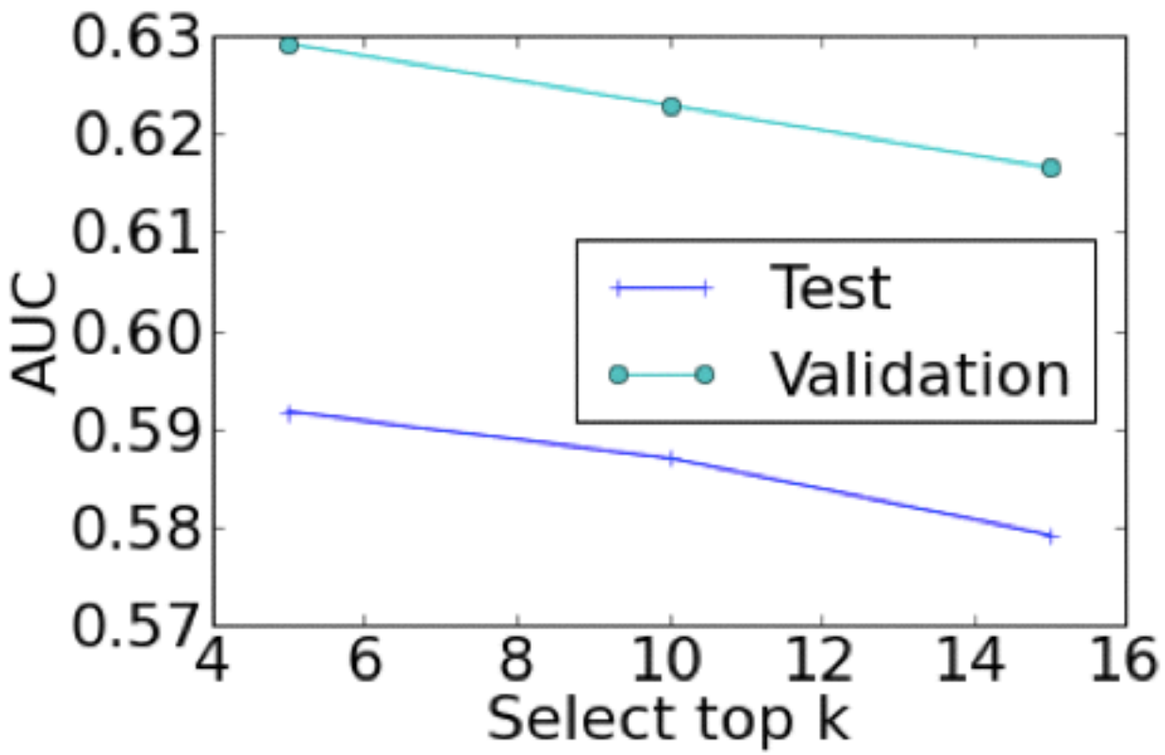
Experiments - overall results

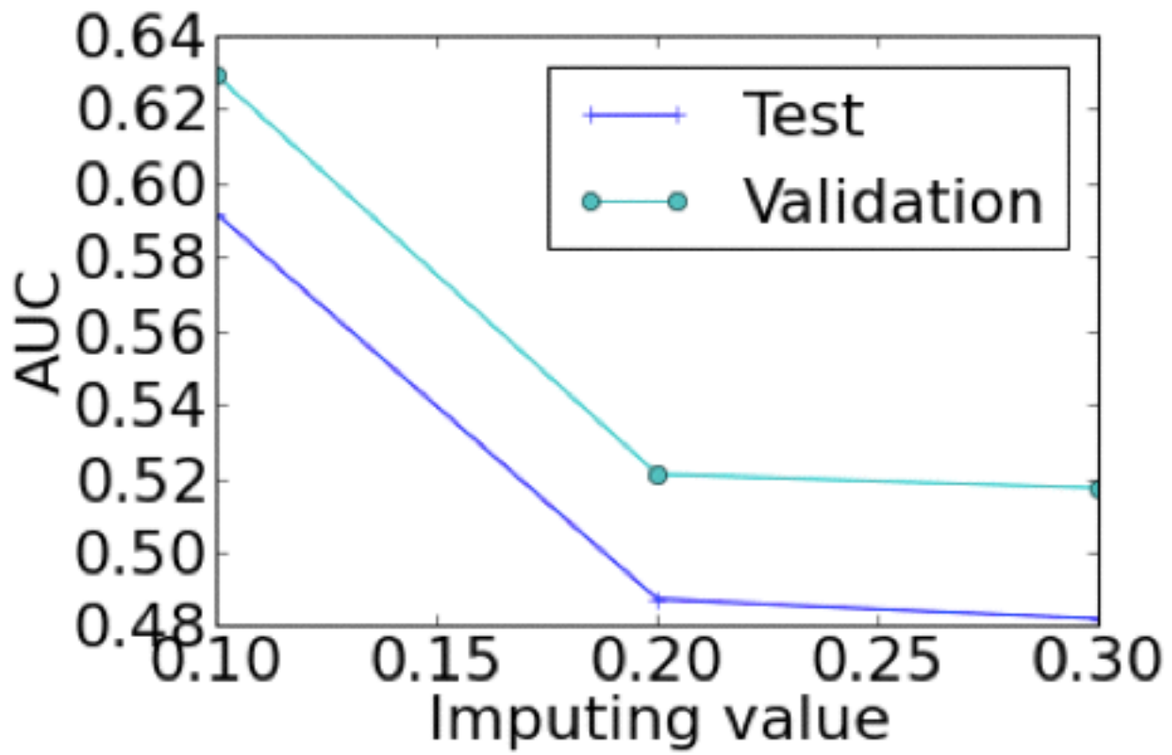


Experiments - sensitivity

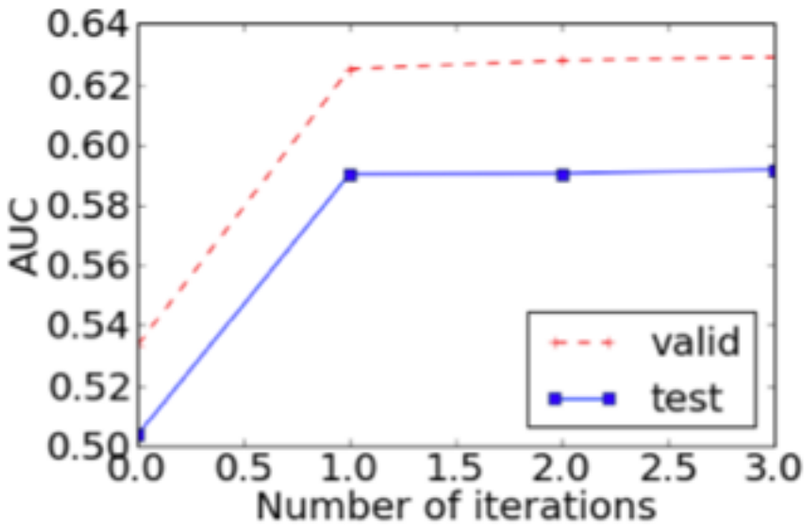
1. top-k entries?
2. what value to fill up

1		2			3	
	1		5	6		4
5		4	5		3	
1	80		6	10		7
2		1		3	1	1

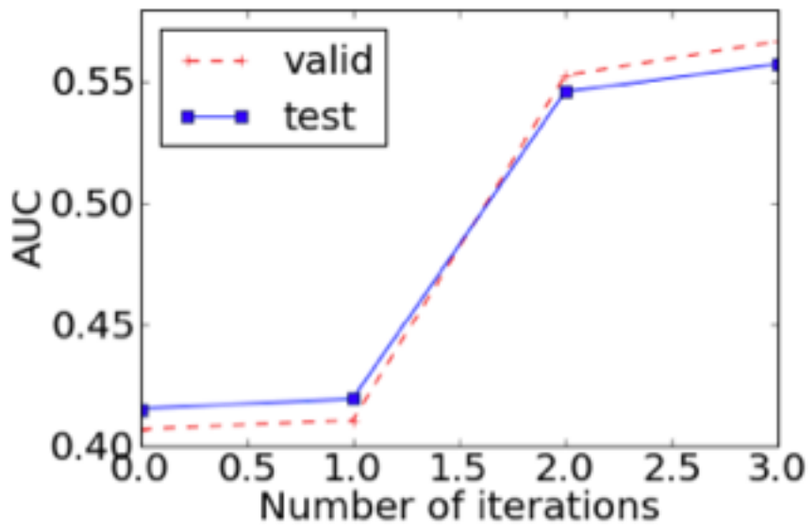




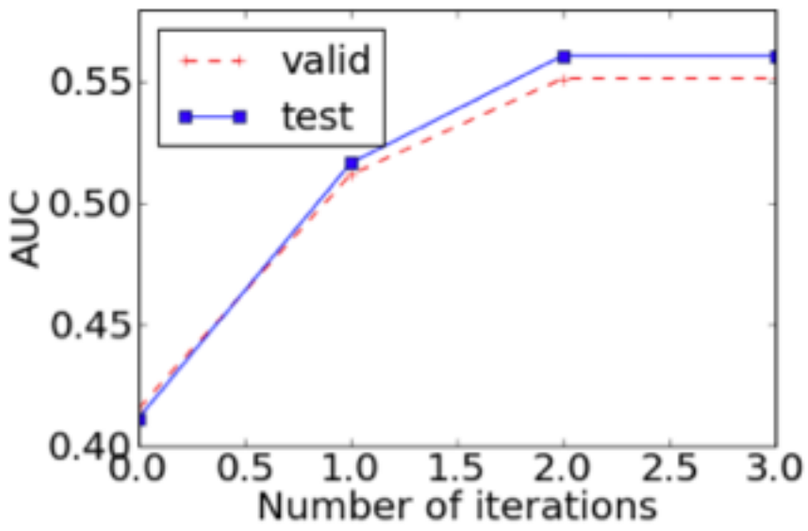
Experiments - convergence



(a) cstheory



(b) unix



(c) english

Conclusions

1. Worker expertise modeling is critical for many websites and services.
2. It is necessary and effective to find out the missing entries in the worker-job interactions to resolve the sparsity issue.