Is the Whole Greater Than the Sum of Its Parts?

Presenter: **Liangyue Li**

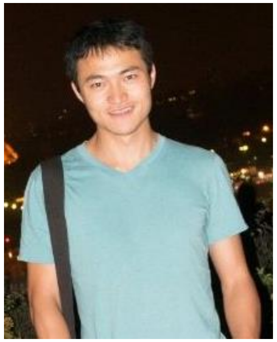
Joint work with

Liangyue Li
(ASU)

Hanghang Tong
(ASU)

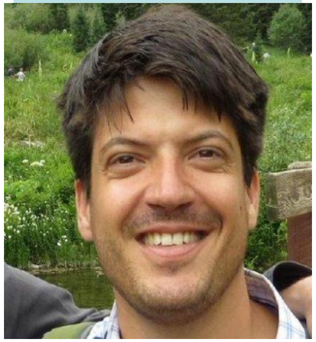
Yong Wang
(HKUST)

Conglei Shi
(IBM->Airbnb)

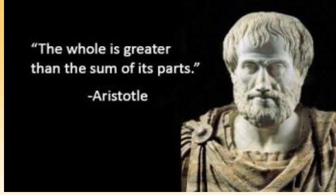


Nan Cao
(Tongji)

Norbou Buchler
(US ARL)



From the Ancient Philosophy



The whole is greater than the sum of its parts. -- **Aristotle**

- **Whole**: a collection of parts
- **Parts**: individual elements
- **Aristotle's hypothesis**:
  - whole > sum of parts

Part-Whole in Team Science

Research Team

Sports Team

Film Crew

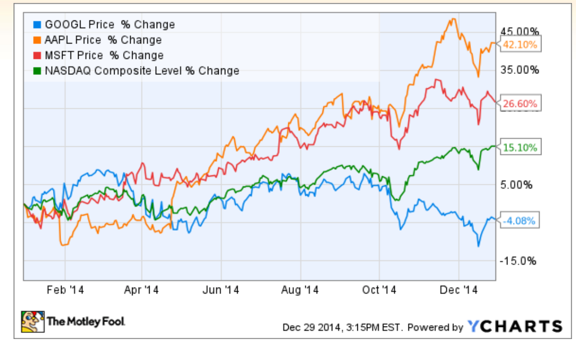
Sales Team

Whole – Team

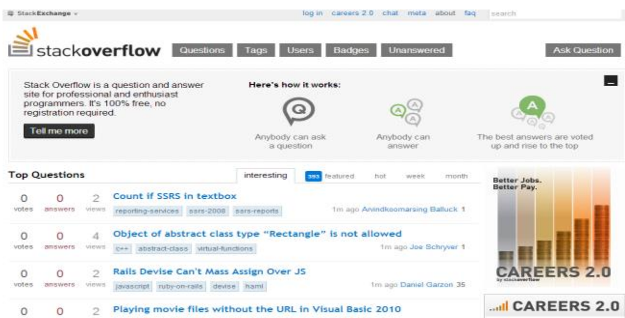
Parts – Team members

Part-Whole Beyond Teams

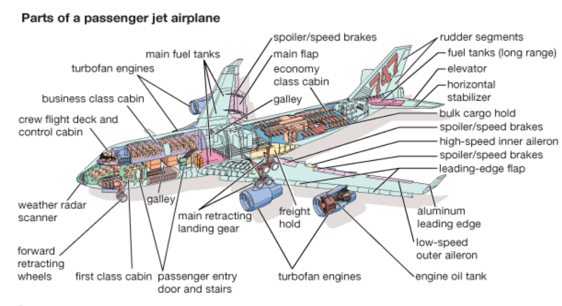
Autonomous System
Whole: system
Parts: individual drones



Stock Market
Whole: DJIA
Parts: individual stock



Community Question Answering
Whole: question
Parts: individual answers

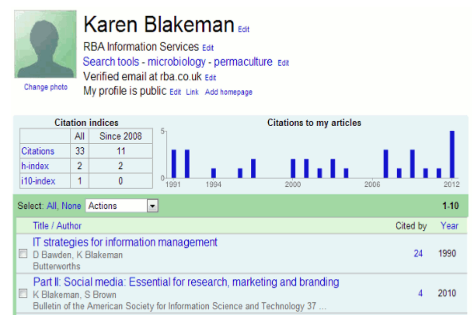


System Reliability
Whole: system
Parts: individual component

Outcome of Part-Whole

Whole: Team
Part: Members

Whole outcome: Team's performance
Part outcome: each member's performance



Whole: Researcher
Part: Publications

Whole outcome: h-index
Part outcome: #citations of publications

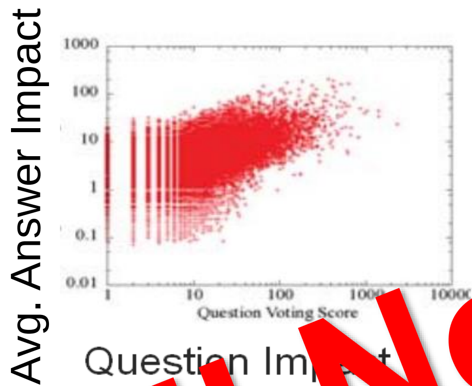
Question: how can we predict the outcome of whole/parts?

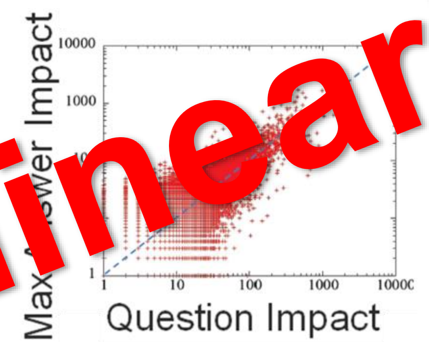
Predict the Part-Whole Outcomes

- Existing Algorithmic Work
  - Separate models for parts and whole
  - Joint linear models
- Aristotle's hypothesis: whole>sum(parts)
- Question: how to jointly predict part/whole
  - by leveraging the part-whole relationship *beyond* the linear models?

Challenges -- Modeling

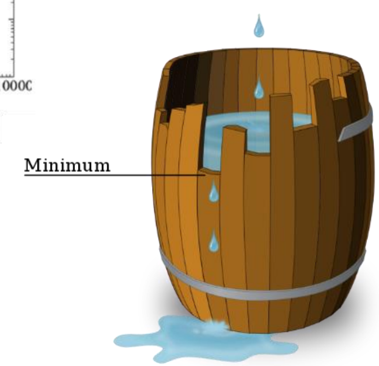
- Nonlinear Part-whole Relationship
  - **Max**: impact of a question is strongly correlated with that of the *best* answer





All Non-linear!

  - **Min**: classic Wooden Bucket Theory



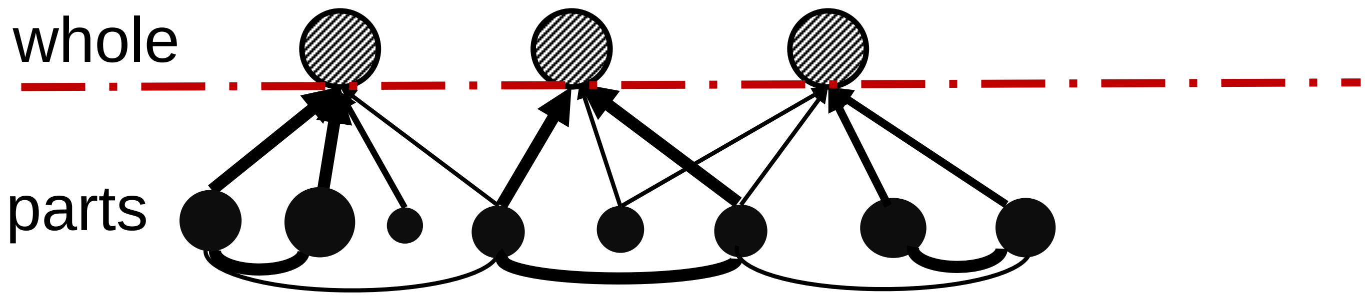
  - **Sparsity**: team performance often dominated by a few top-performing team members

DATA Lab

Challenges – Modeling (con't)

- Part-part Interdependency
  - Parts are connected via underlying network
  - Impact of such interdependency on outcomes

Hypothesis-1: similar parts -> similar contribution to whole
Hypothesis-2: similar parts -> similar parts outcome



Question: how can we utilize
1. nonlinear part-whole relationship
2. part-part interdependency

Challenges -- Algorithm

Non-linearity

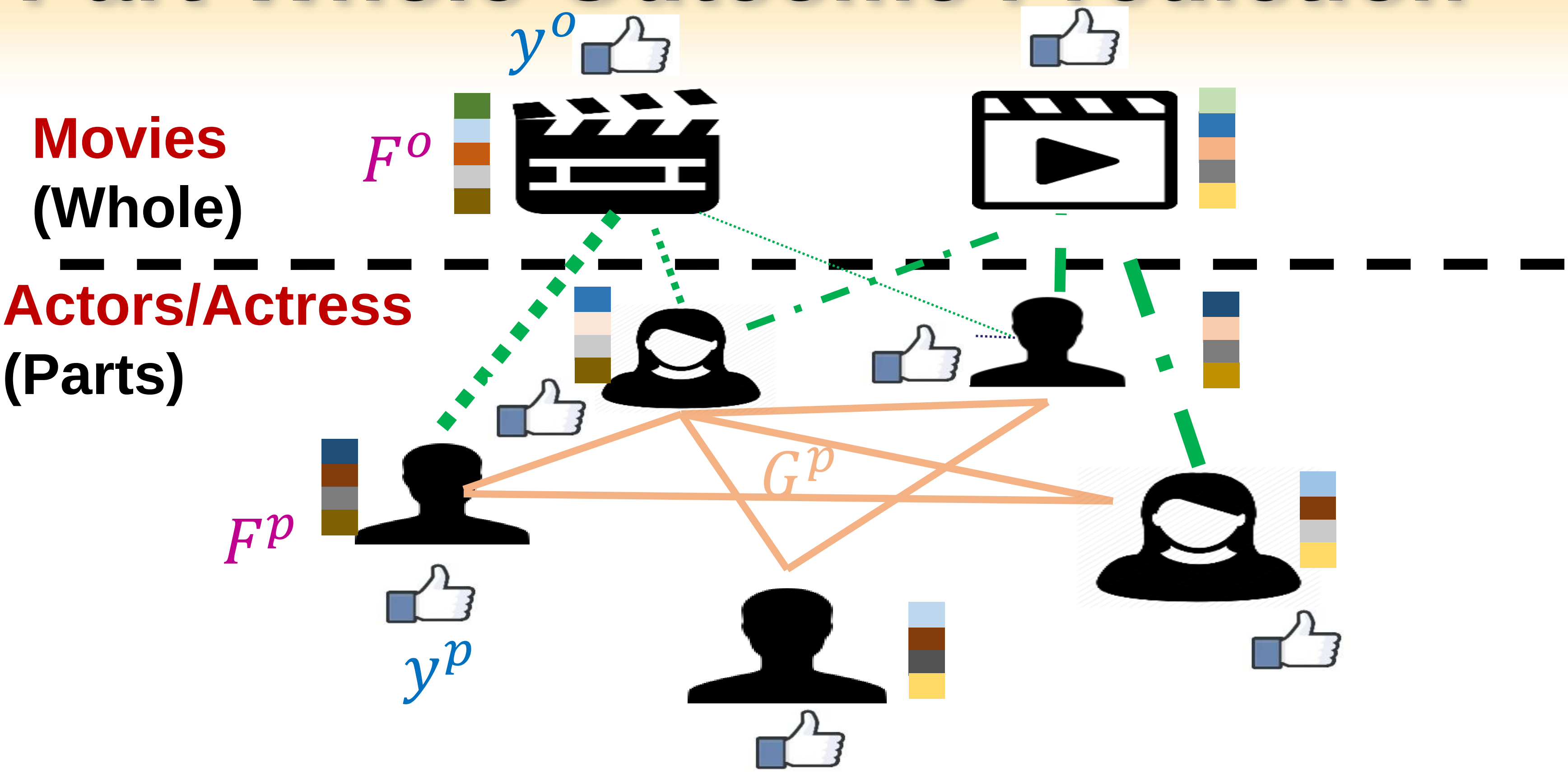
+

Interdependency

} **high complexity**

Question: how to scale up the computation?

Part-Whole Outcome Prediction



Given: 1. feature matrix for whole/part F^o/F^p

2. outcome vector for whole/part y^o/y^p

3. whole to part mapping ϕ

4. parts' network G^p (optional)

Predict: outcome of new whole/parts

Roadmap

- Motivations
- **PAROLE -- Modeling**
  - **Generic Framework**
  - Modeling Part-Whole Relationship
  - Modeling Part-Part Interdependency
- PAROLE -- Optimization
- Empirical Evaluations
- Conclusions

A Generic Joint Prediction Framework -- PAROLE

- Formulation

$$\min J = J_o + J_p + J_{po} + J_{pp} + J_r$$

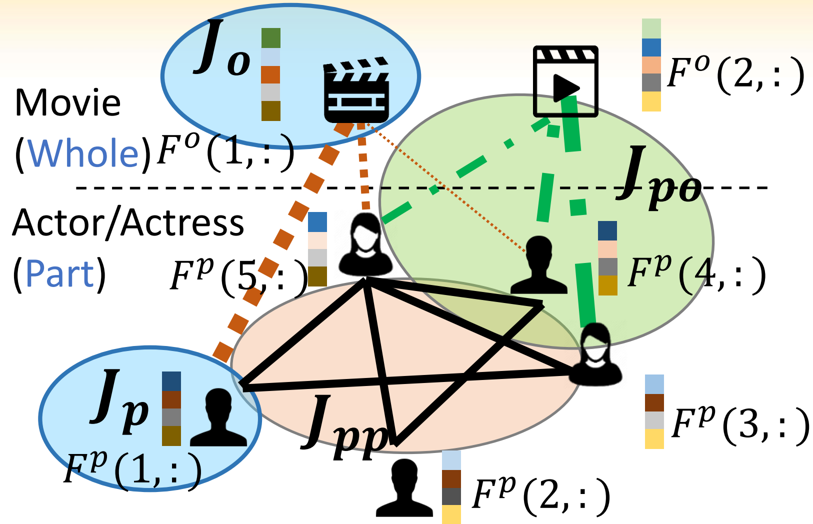
$$= \frac{1}{n_o}\sum_{i=1}^{n_o} L[f(F^o(i,:), w^o), y^o(i))]$$

$$+ \frac{1}{n_p}\sum_{i=1}^{n_p} L[f(F^p(i,:), w^p), y^p(i))]$$

$$+ \frac{\alpha}{n_o}\sum_{i=1}^{n_o} h(f(F^o(i,:), w^o), Agg(\phi(o_i)))$$

$$+ \frac{\beta}{n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{n_p} G_{ij}^p g(f(F^p(i,:), w^p), f(F^p(j,:), w^p))$$

$$+ \gamma(\Omega(w^o) + \Omega(w^p))$$



J_o: **Predictive Model for Whole**

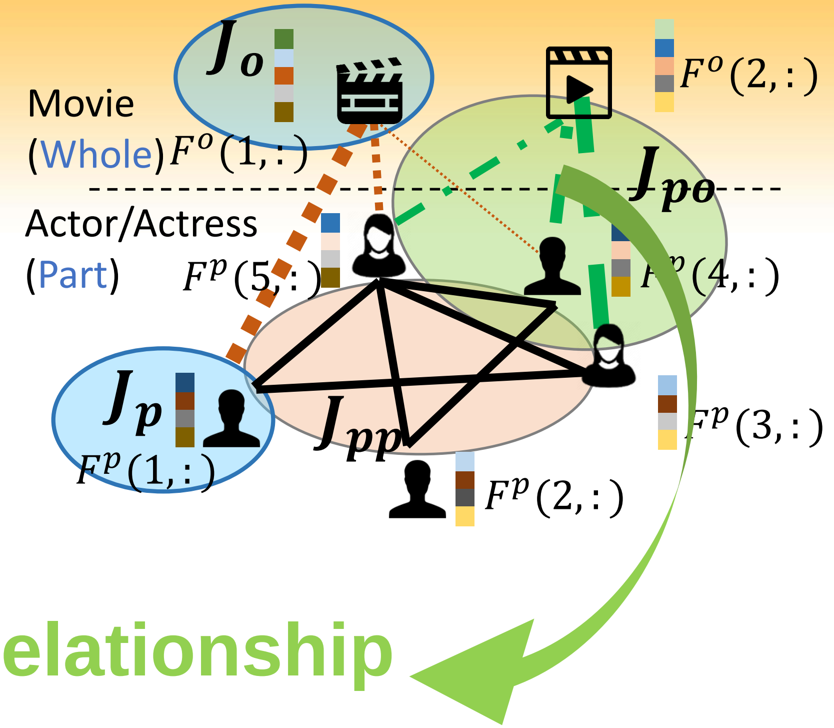
J_p: **Predictive Model for Part**

J_{po}: **Part-whole Relationship**

J_{pp}: **Part-part Interdependency**

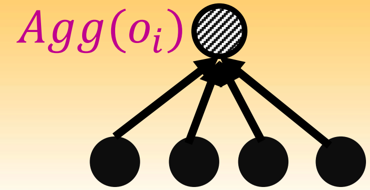
J_r: **parameter regularizer**

Roadmap



- Motivations
- **PAROLE -- Modeling**
  - Generic Framework
  - **Modeling Part-Whole Relationship**
  - Modeling Part-Part Interdependency
- PAROLE -- Optimization
- Empirical Evaluations
- Conclusions

Modeling Part-Whole Relationship



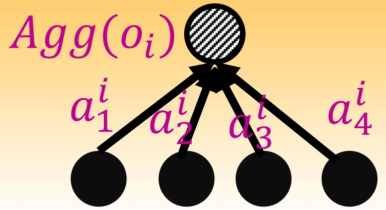
- **Overview**: for each whole entity $\mathrm{o_i}$, define

$$e_i = \mathbf{F}^{\mathrm{o}}(\mathrm{i},:)\mathbf{w}^{\mathrm{o}} - \mathrm{Agg}(\mathrm{o_i})$$

  – e_i: Measure the difference between
    - predicted whole outcome using whole feature
    - predicted whole outcome using aggregated parts outcome

- **Key idea**: model part-whole relations by
    - Different loss functions on e_i
    - Different aggregation functions $Agg(\cdot)$

Overview



- **Intuition:** whole ← (weighted) sum of parts
- **Details**:

$$e_i = F^o(i,:)w^o - Agg(o_i)$$

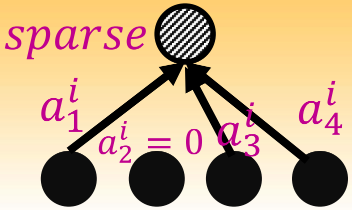
$$Agg(o_i) = \sum_{j \in \phi(o_i)} a_j^i F^p(j,:)w^p$$

  – a_j^i: weight of part j's contribution to the whole o_i's outcome
- **Remark**:
  – Characterize part-whole relationships
    • Use different loss functions on e_i
    • Use different norms on a_i

Linear Part-Whole Relation

- **Intuition**: Whole ← linear combination of parts
  - some parts play more important roles than the others in contributing to the whole outcome
- **Details**: $J_{po} = \frac{\alpha}{2n_o}\sum_{i=1}^{n_o} e_i^2$
- **Remark**:
  - $a_j^i = 1$: the whole is the sum of its parts
  - $a_j^i = \frac{1}{|o_i|}$: average coupling

Sparse Part-Whole Relation



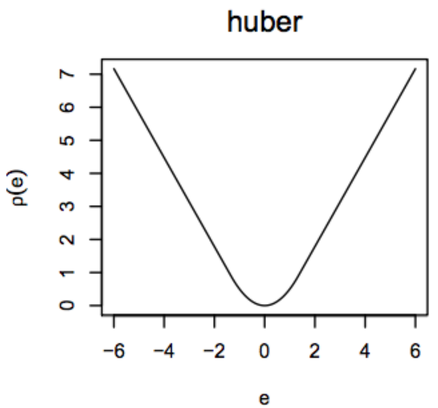
- **Intuition**: Whole ← a few parts
  - some parts have little or no effect on the whole outcome
- **Details**: $J_{po} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} (\frac{1}{2} e_i^2 + \lambda |\mathbf{a}_i|_1)$
- **Remark**:
  - The l_1 norm can shrink some part contributions a_j^i to exactly zero
  - **Nonlinear** part-whole relation

Ordered Sparse Part-Whole Relation

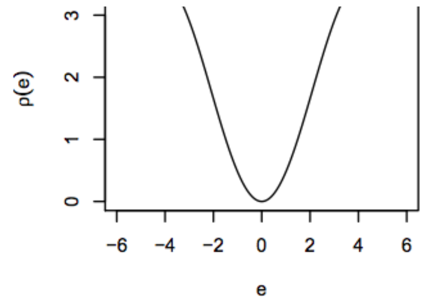
- **Intuition**: Whole ← a few top parts
  - team performance is determined by not only a few key members, but also the structural hierarchy between them
- **Details:** $J_{po} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} (\frac{1}{2} e_i^2 + \lambda \Omega_w(\mathbf{a}_i))$
  - $\Omega_w(x) = \sum_{i=1}^{n} |x|_{[i]} w_i = \boldsymbol{w}^T |\boldsymbol{x}|_{\downarrow}$: ordered weighted l_1 norm
  - $w \in \mathcal{K}_{m+}$: vector of non-increasing non-negative weights

Robust Part-Whole Relation

- **Intuition**: Whole ← parts that are not outliers
  – squared loss is sensitive to outliers.
- **Solution**: robust regression model
- **Details**: $J_{po} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} \rho(e_i)$
  – $\rho(\cdot)$ is robust estimator



Method \ Case	$\lvert e\rvert \le t$	$\lvert e\rvert > t$
Huber $\rho_H(e)$	$\frac{1}{2}e^2$	$t\lvert e\rvert - \frac{1}{2}t^2$
Bisquare $\rho_B(e)$	$\frac{t^2}{6}\left\{1 - [1 - (\frac{e}{t})^2]^3\right\}$	$\frac{t^2}{6}$



Maximum Part-Whole Relation

- **Intuition**: Whole ← max(parts)
  – team performance dominated by the best team member/leader
- **Details**:
  – $Agg(o_i) = \max(parts' outcome)$ [not differentiable]
  – Soft max function: $\max(x_1, x_2, \ldots, x_n) \approx \ln(\exp(x_1) + \exp(x_2) + \cdots + \exp(x_n))$
  – Aggregation: $Agg(o_i) = \ln(\sum_{j \in \phi(o_i)} \exp(F^p(j,:)w^p))$

$$J_{po} = \frac{\alpha}{2n_o} \sum_{i=1}^{n_o} e_i^2$$

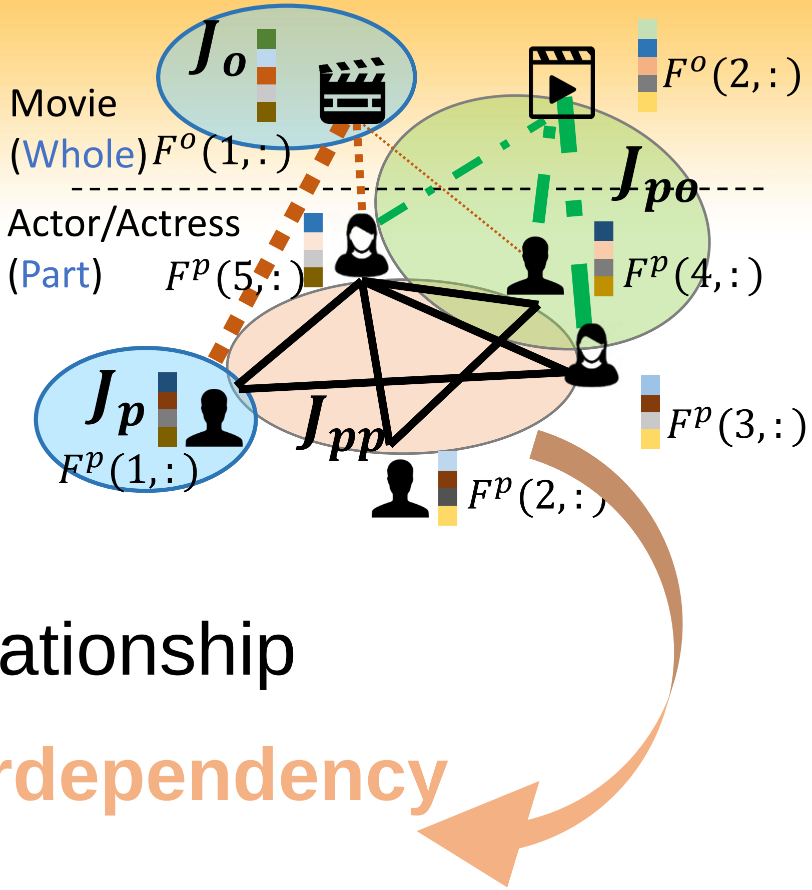
DATA Lab

Summarize Part-Whole Relations

Name	$Agg(o_i)$ Aggregation of parts	J_{po} Sub-objective	Remark
Maximum	$\ln(\sum \exp(F^p(j,:)w^p))$	$\frac{\alpha}{2n_o}\sum e_i^2$	Nonlinear Whole ← max(parts)
Linear	$\sum a_j^i F^p(j,:)w^p$	$\frac{\alpha}{2n_o}\sum e_i^2$	Linear Whole ← linear combination of parts
Sparse	$\sum a_j^i F^p(j,:)w^p$	$\frac{\alpha}{n_o}\sum(\frac{1}{2}e_i^2 + \lambda\lvert a_i\rvert_1)$	Nonlinear Whole ← a few parts
Ordered Sparse	$\sum a_j^i F^p(j,:)w^p$	$\frac{\alpha}{n_o}\sum(\frac{1}{2}e_i^2 + \lambda\Omega_w(a_i))$	Nonlinear Whole ← a few top parts
Robust	$\sum a_j^i F^p(j,:)w^p$	$\frac{\alpha}{n_o}\sum \rho(e_i)$	Nonlinear Whole ← parts that are not outliers

Roadmap

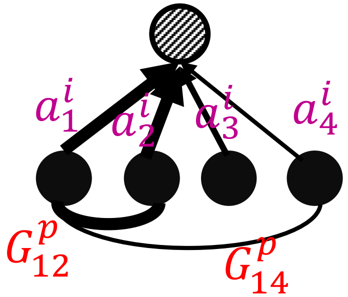
- Motivations
- **PAROLE -- Modeling**
  - Generic Framework
  - Modeling Part-Whole Relationship
  - **Modeling Part-Part Interdependency**
- PAROLE -- Optimization
- Empirical Evaluations
- Conclusions



Modeling Part-Part Interdependency

- **Effect on the whole outcome**
  - **Intuition**: closely connected parts might have similar contribution to the whole outcome
  - **Details**:

$$\mathcal{J}_{po} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} \left[\frac{1}{2} e_i^2 + \lambda |\mathbf{a}_i|_1 + \frac{1}{2} \sum_{k,l \in \phi(o_i)} G_{kl}^p (a_k^i - a_l^i)^2 \right]$$



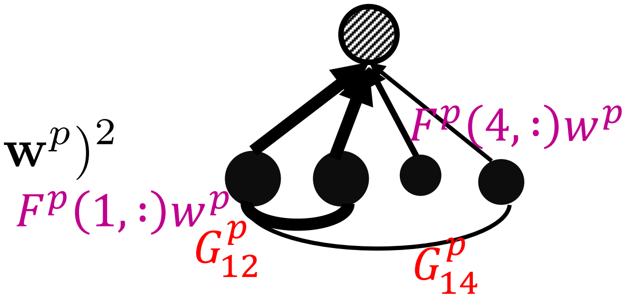
- Similar parts (large G_{kl}^p)

→ similar contributions ($a_k^i \approx a_l^i$)

Modeling Part-Part Interdependency

- **Effect on the parts outcome**
  - **Intuition**: closely connected parts might share similar outcomes themselves
  - **Details**:

$$\mathcal{J}_{pp} = \frac{\beta}{2n_p} \sum_{i=1}^{n_p} \sum_{j=1}^{n_p} G_{ij}^p (\mathbf{F}^p(i,:)\mathbf{w}^p - \mathbf{F}^p(j,:)\mathbf{w}^p)^2$$



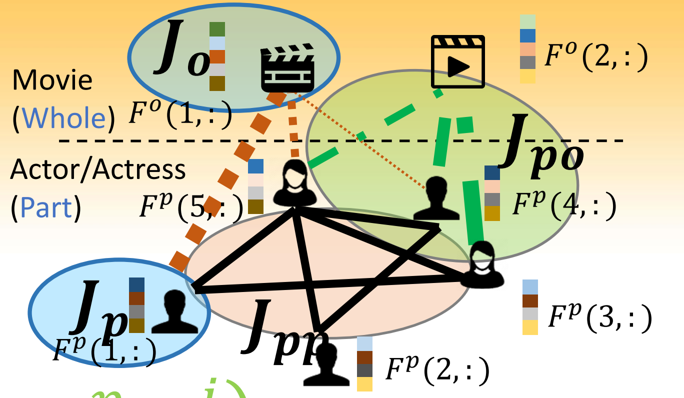
    - Similar parts (large G_{ij}^p)

    → similar predicted outcomes ($F^p(i,:)w^p \approx F^p(j,:)w^p$)

Roadmap

- Motivations
- PAROLE -- Modeling
- **PAROLE -- Optimization**
- Empirical Evaluations
- Conclusions

Optimization Solution



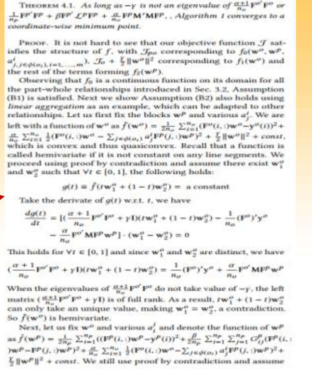
- **Formulation**:
  - $J = J_o(w^o) + J_p(w^p) + J_{po}(w^o, w^p, a_j^i) + J_{pp}(w^p) + J_r(w^o, w^p)$
- **Observation**:
  - not jointly convex w.r.t. w^o, w^p, a_i^j
  - Convex w.r.t. to one block while fixing others
- **Solution**: block coordinate descent

Block Coordinate Descent

- Three coordinate blocks: w^o, w^p, a_j^i
- Update one block while fixing others
- Update each block
  - (proximal) gradient descent

	$\frac{\partial J_{po}}{\partial w^o}$	$\frac{\partial J_{po}}{\partial w^p}$	$\frac{\partial J_{po}}{\partial a_i}$ or proximal gradient update
Maximum Agg	$\frac{\alpha}{n_o}\sum_{i=1}^{n_o} e_i(F^o(i,:))'$	$\frac{\alpha}{n_o}\sum_{i=1}^{n_o} e_i \frac{\sum_{j\in\phi(o_i)}(F^p(j,:))'\tilde{y}_i^p}{\sum_{j\in\phi(o_i)}\tilde{y}_i^p}$	N/A
Linear Agg	$\frac{\alpha}{n_o}(\mathrm{F^o})'(\mathrm{F^o w^o} - \mathrm{MF^p w^p})$	$-\frac{\alpha}{n_o}(\mathrm{F^p})'\mathrm{M}'(\mathrm{F^o w^o} - \mathrm{MF^p w^p})$	$e_i(-F^p(\phi(o_i),:)w^p) + L_i^p a_i$
Sparse Agg	$\frac{\alpha}{n_o}(\mathrm{F^o})'(\mathrm{F^o w^o} - \mathrm{MF^p w^p})$	$-\frac{\alpha}{n_o}(\mathrm{F^p})'\mathrm{M}'(\mathrm{F^o w^o} - \mathrm{MF^p w^p})$	$z = a_i - \tau\left[e_i(-F^p(\phi(o_i),:)w^p) + L_i^p a_i\right]$ $a_i \leftarrow prox_{\lambda\tau l_1}(z)$
Order Sparse Agg	$\frac{\alpha}{n_o}(\mathrm{F^o})'(\mathrm{F^o w^o} - \mathrm{MF^p w^p})$	$-\frac{\alpha}{n_o}(\mathrm{F^p})'\mathrm{M}'(\mathrm{F^o w^o} - \mathrm{MF^p w^p})$	$z = a_i - \tau\left[e_i(-F^p(\phi(o_i),:)w^p) + L_i^p a_i\right]$ $a_i \leftarrow prox_{\lambda\tau\Omega_w}(z)$
Robust Agg	$\frac{\alpha}{n_o}\sum_{i=1}^{n_o}\frac{\partial\rho(e_i)}{\partial e_i}F^o(i,:)'$	$\frac{\alpha}{n_o}\sum_{i=1}^{n_o}\frac{\partial\rho(e_i)}{\partial e_i}(-\sum_{j\in\phi(o_i)} a_j F^p(j,:)')$	$\frac{\alpha}{n_o}\left[\frac{\partial\rho(e_i)}{\partial e_i}(-F^p(\phi(o_i),:)w^p) + L_i^p a_i\right]$

Optimization Properties



Theorem 4.1. *As long as $-\gamma$ is not an eigenvalue of $\frac{\alpha+1}{n_o}\mathbf{F}^{o\prime}\mathbf{F}^o$ or $\frac{1}{n_p}\mathbf{F}^{p\prime}\mathbf{F}^p + \beta\mathbf{F}^{p\prime}\mathcal{L}^p\mathbf{F}^p + \frac{\alpha}{n_o}\mathbf{F}^p\mathbf{M}'\mathbf{M}\mathbf{F}^p$, , Algorithm 1 converges to a coordinate-wise minimum point.*

Proof. It is not hard to see that our objective function $\mathcal{J}$ satisfies the structure of f, with $\mathcal{J}_{po}$ corresponding to $f_0(\mathbf{w}^o, \mathbf{w}^p, a^i_{j, j\in\phi(o_i), i=1,\ldots,m})$, $\mathcal{J}_o + \frac{\gamma}{2}\|\mathbf{w}^o\|^2$ corresponding to $f_1(\mathbf{w}^o)$ and the rest of the terms forming $f_2(\mathbf{w}^p)$.

Observing that f_0 is a continuous function on its domain for all the part-whole relationships introduced in Sec. 3.2, Assumption (B1) is satisfied. Next we show Assumption (B2) also holds using *linear aggregation* as an example, which can be adapted to other relationships. Let us first fix the blocks $\mathbf{w}^p$ and various a^i_j. We are left with a function of $\mathbf{w}^o$ as $\bar{f}(\mathbf{w}^o) = \frac{1}{2n_o}\sum_{i=1}^{n_o}(\mathbf{F}^o(i,:)\mathbf{w}^o - \mathbf{y}^o(i))^2 + \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\frac{1}{2}(\mathbf{F}^o(i,:)\mathbf{w}^o - \sum_{j\in\phi(o_i)} a^i_j\mathbf{F}^p(j,:)\mathbf{w}^p)^2 + \frac{\gamma}{2}\|\mathbf{w}^o\|^2 + const$, which is convex and thus quasiconvex. Recall that a function is called hemivariate if it is not constant on any line segments. We proceed using proof by contradiction and assume there exist $\mathbf{w}^o_1$ and $\mathbf{w}^o_2$ such that $\forall t \in [0, 1]$, the following holds:

$$g(t) = \bar{f}(t\mathbf{w}^o_1 + (1-t)\mathbf{w}^o_2) = \text{ a constant}$$

Take the derivate of $g(t)$ w.r.t. t, we have

$$\frac{dg(t)}{dt} = [(\frac{\alpha+1}{n_o}\mathbf{F}^{o\prime}\mathbf{F}^o + \gamma\mathbf{I})(t\mathbf{w}^o_1 + (1-t)\mathbf{w}^o_2) - \frac{1}{n_o}(\mathbf{F}^o)'\mathbf{y}^o - \frac{\alpha}{n_o}\mathbf{F}^{o\prime}\mathbf{M}\mathbf{F}^p\mathbf{w}^p]\cdot(\mathbf{w}^o_1 - \mathbf{w}^o_2) = 0$$

This holds for $\forall t \in [0, 1]$ and since $\mathbf{w}^o_1$ and $\mathbf{w}^o_2$ are distinct, we have

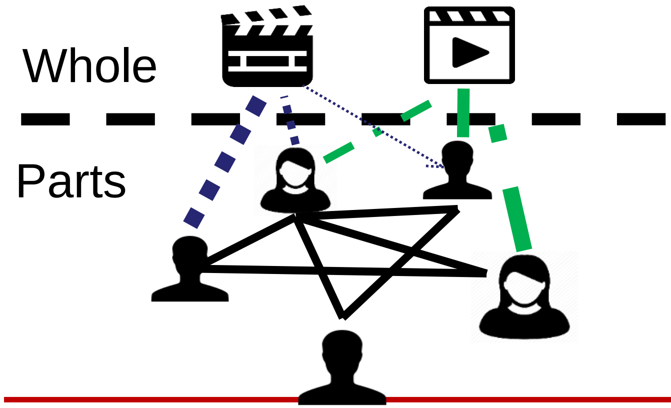
$$(\frac{\alpha+1}{n_o}\mathbf{F}^{o\prime}\mathbf{F}^o + \gamma\mathbf{I})(t\mathbf{w}^o_1 + (1-t)\mathbf{w}^o_2) = \frac{1}{n_o}(\mathbf{F}^o)'\mathbf{y}^o + \frac{\alpha}{n_o}\mathbf{F}^{o\prime}\mathbf{M}\mathbf{F}^p\mathbf{w}^p$$

When the eigenvalues of $\frac{\alpha+1}{n_o}\mathbf{F}^{o\prime}\mathbf{F}^o$ do not take value of $-\gamma$, the left matrix $(\frac{\alpha+1}{n_o}\mathbf{F}^{o\prime}\mathbf{F}^o + \gamma\mathbf{I})$ is of full rank. As a result, $t\mathbf{w}^o_1 + (1-t)\mathbf{w}^o_2$ can only take an unique value, making $\mathbf{w}^o_1 = \mathbf{w}^o_2$, a contradiction. So $\bar{f}(\mathbf{w}^o)$ is hemivariate.

Next, let us fix $\mathbf{w}^o$ and various a^i_j and denote the function of $\mathbf{w}^p$ as $\tilde{f}(\mathbf{w}^p) = \frac{1}{2n_p}\sum_{i=1}^{n_p}((\mathbf{F}^p(i,:)\mathbf{w}^p - \mathbf{y}^p(i))^2 + \frac{\beta}{2n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{n_p} G^p_{ij}(\mathbf{F}^p(i,:)\mathbf{w}^p - \mathbf{F}^p(j,:)\mathbf{w}^p)^2 + \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\frac{1}{2}(\mathbf{F}^o(i,:)\mathbf{w}^o - \sum_{j\in\phi(o_i)} a^i_j\mathbf{F}^p(j,:)\mathbf{w}^p)^2 + \frac{\gamma}{2}\|\mathbf{w}^p\|^2 + const$. We still use proof by contradiction and assume

details

- **Convergence and Optimality**
  - Under mild conditions, the optimization alg converges to a coordinate-wise minimum point
- **Complexity**
  - The alg scales linearly w.r.t. the size of part-whole graph in both time and space



Complexity: $O(n_o d_o + n_p d_p + m_{po} + m_{pp})$

n_o: #whole entities

n_p: #part entities

m_{po}: #links from whole to parts

m_{pp}: #links in part-part network

d_o, d_p: feature dimension of whole, parts

DATA Lab

Roadmap

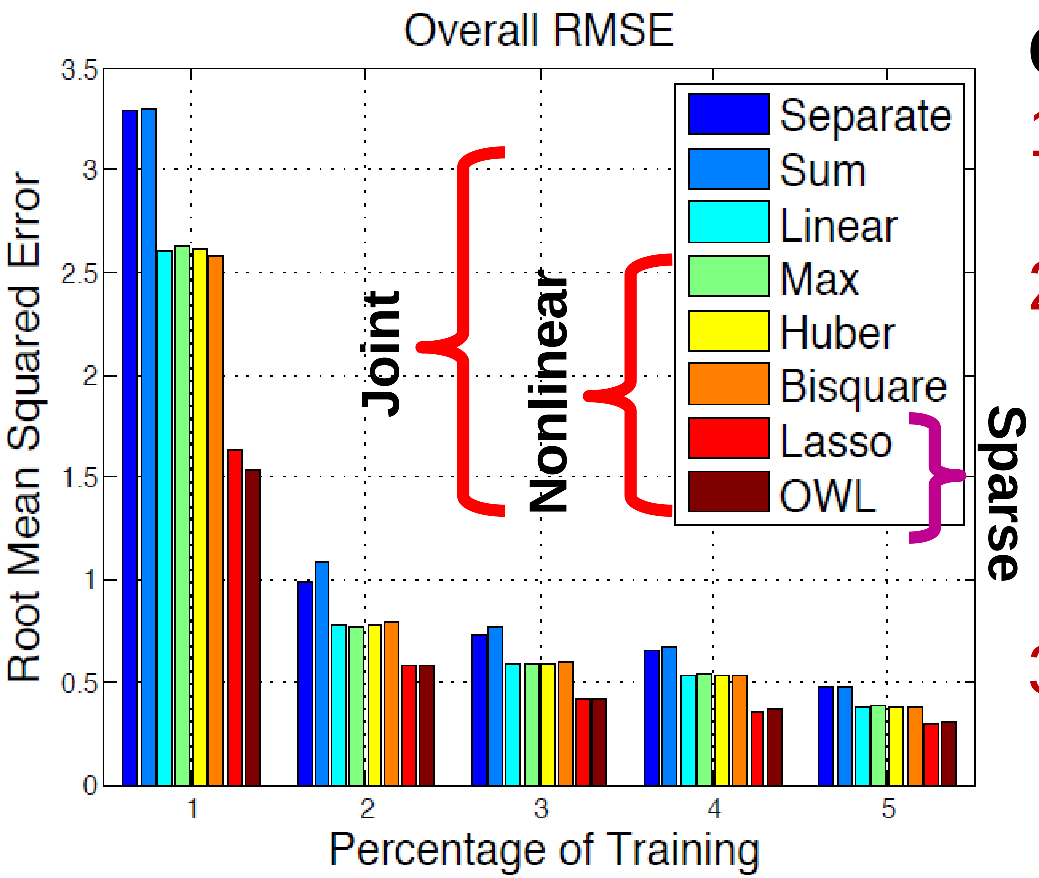
- Motivations
- PAROLE -- Modeling
- PAROLE -- Optimization
- **Empirical Evaluations**
- Conclusions

Datasets

Data	Whole	Part	#Whole	#Part
Math	Question (#votes)	Answer (#votes)	16,638	32,876
SO	Question (#votes)	Answer (#votes)	1,966,272	4,282,570
DBLP	Author (h-index)	Paper (#citation)	234,681	129,756
Movie	Movie (# )	Actors/Actress (# )	5,043	37,365

- **Setup**: sort whole in chronological order, gather first x percent and corresponding parts as training, test on last 10%
- **Metric**: root mean squared error (RMSE)

Outcome Prediction Performance



Math

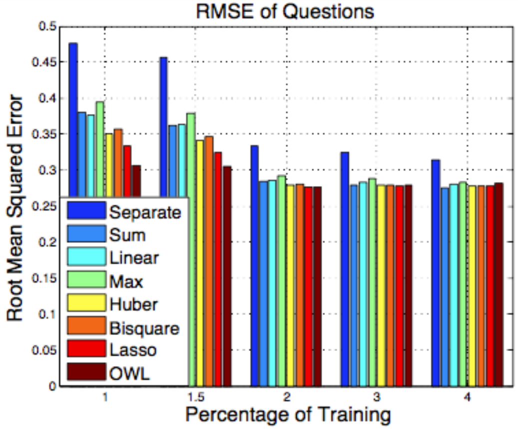
Observations

1. Joint prediction models > separate models
2. Non-linear part-whole relationships (max, Huber, Bisquare, Lasso, OWL) > linear relationships (Sum, Linear)
3. Lasso and OWL are the best methods in most cases

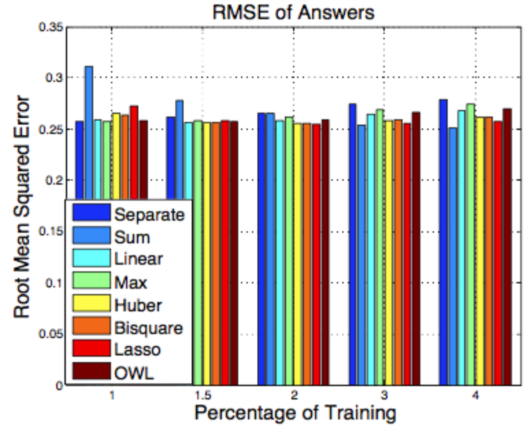
Outcome Prediction Performance

Whole | Parts | Overall

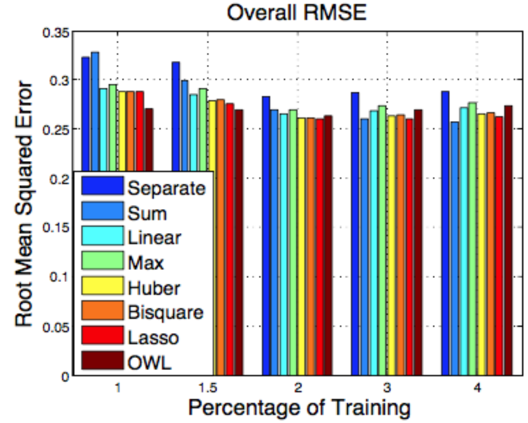
SO



(a) RMSE of question outcome prediction.

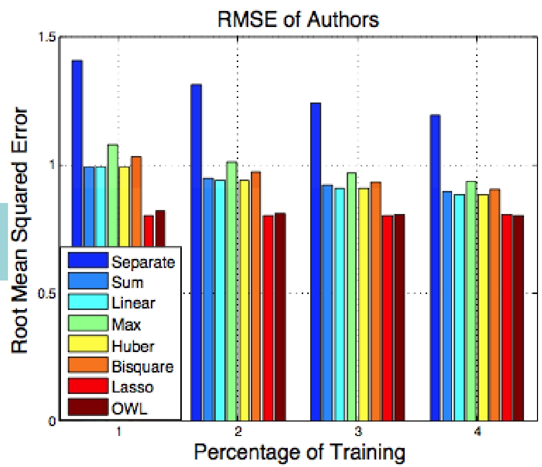


(b) RMSE of answer outcome prediction.

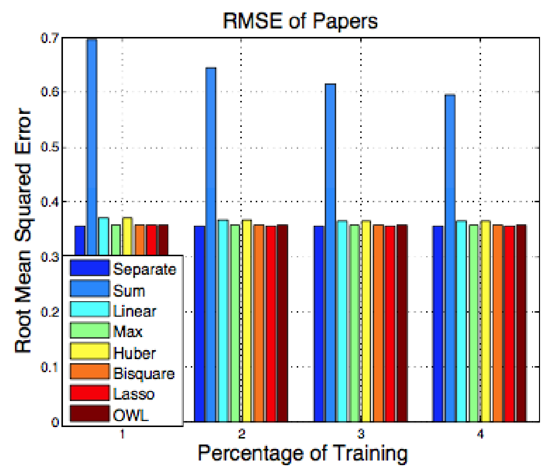


(c) Overall RMSE.

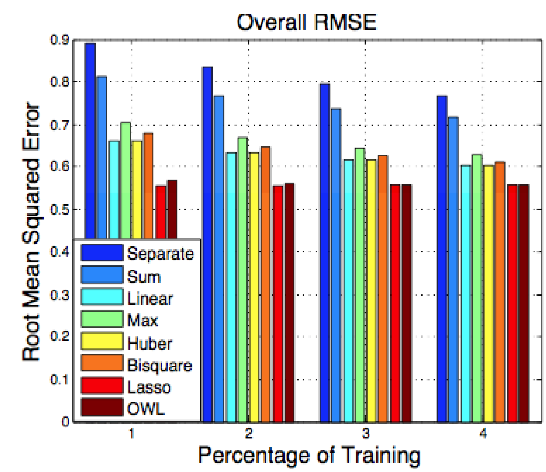
DBLP



(a) RMSE of author outcome prediction.



(b) RMSE of paper outcome prediction.



(c) Overall RMSE.

Effect of part-part interdependency

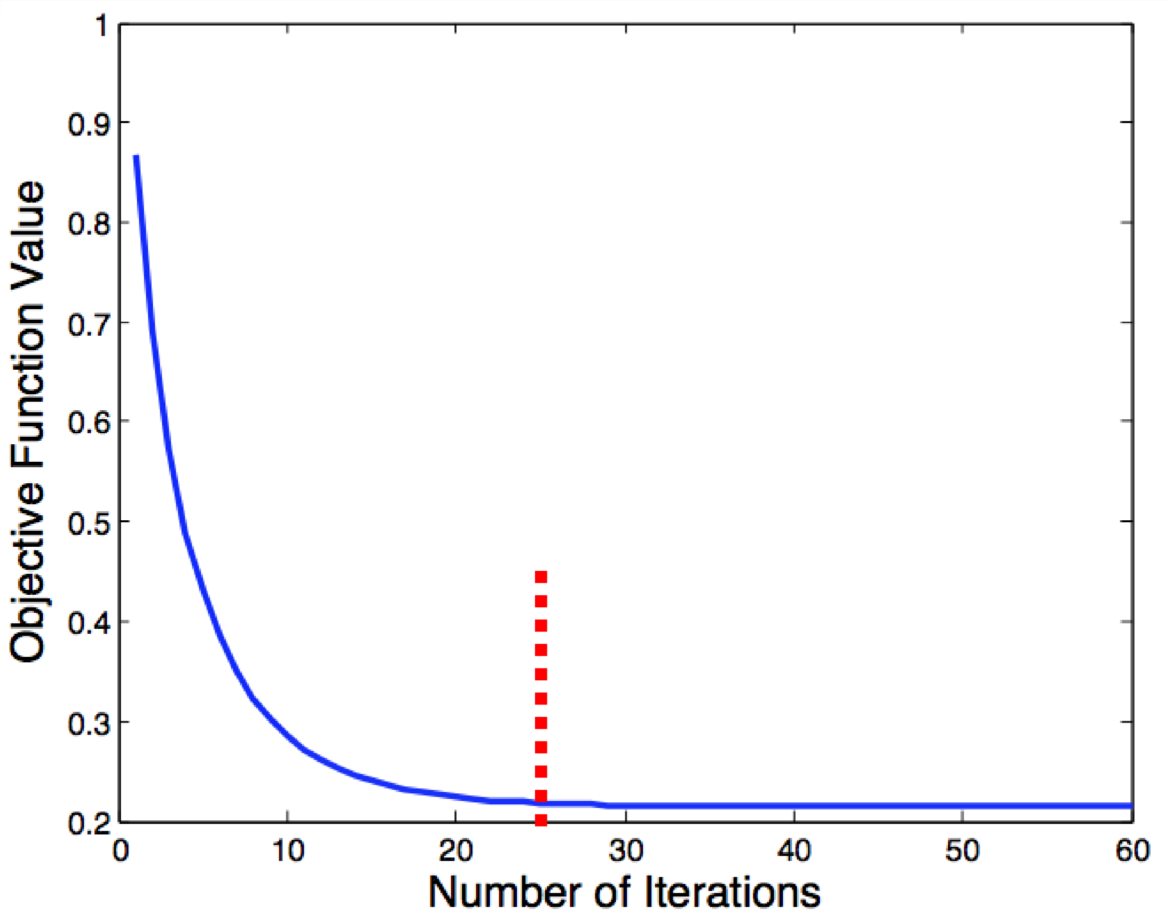
Movie

- PAROLE-Basic – no network information
- Part-part interdependency on whole outcome and parts outcome both boost the performance

DATA Lab

Convergence Analysis

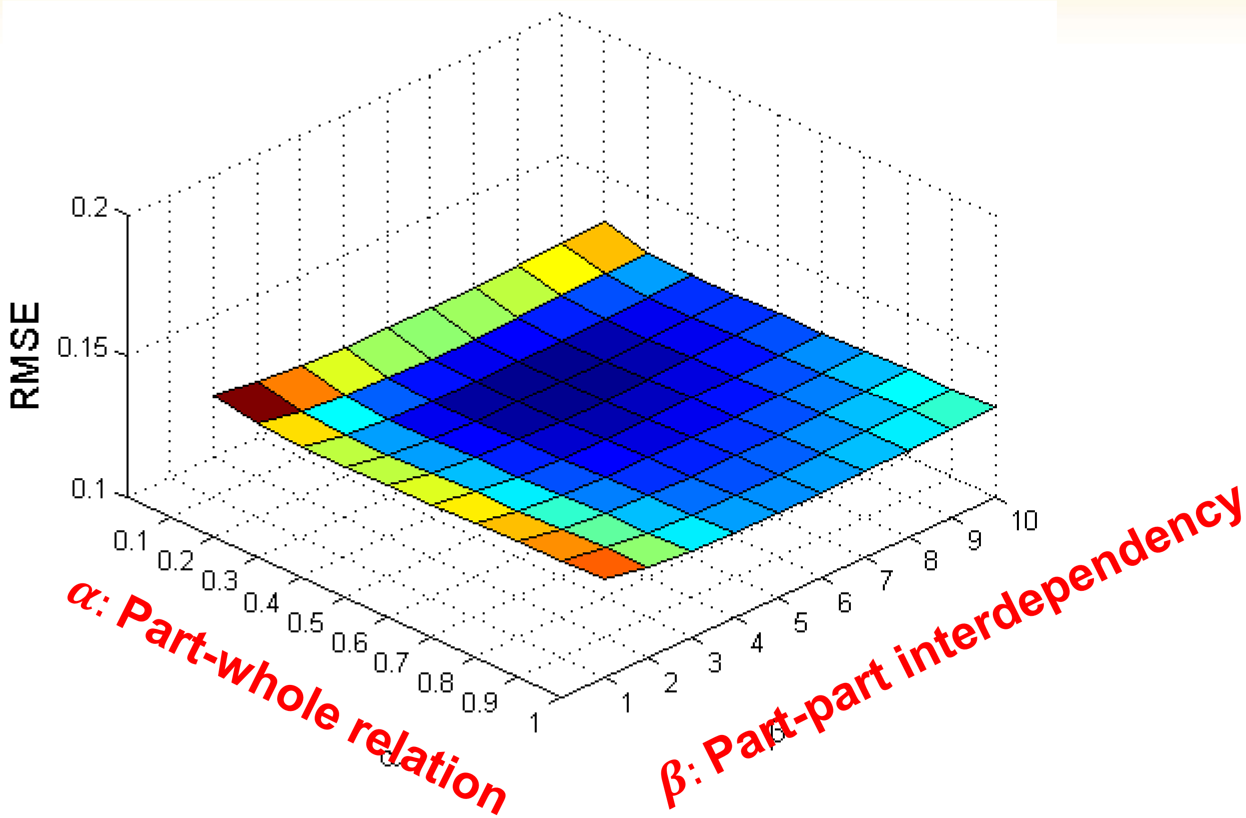
SO



- PAROLE converges fast (25-30 iterations)

Parameter Sensitivity

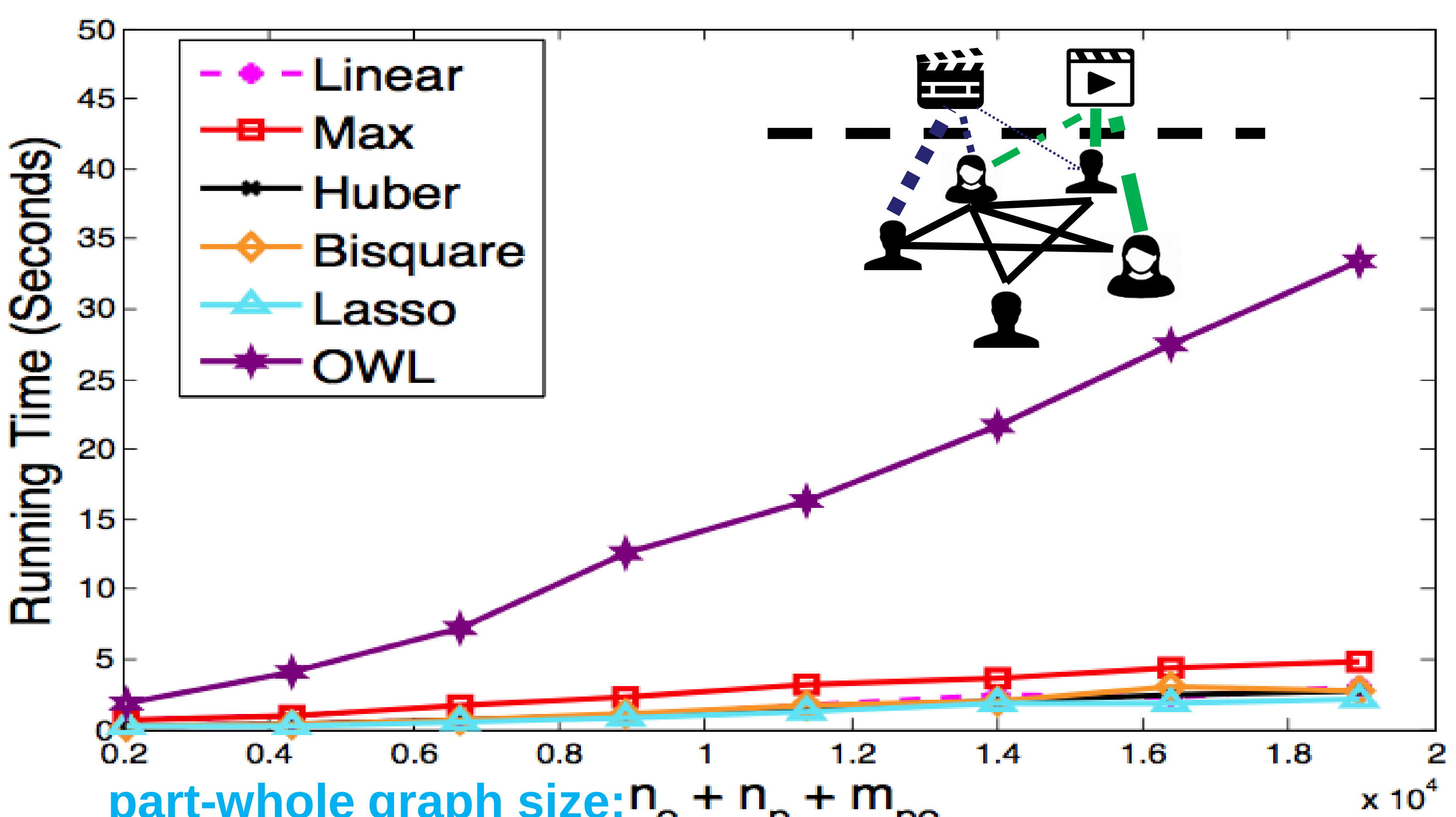
Movie



- α controls importance of part-whole relation
- β controls importance of part-part interdependency
- Stable in a relatively large parameter space

Scalability of PAROLE

SO



part-whole graph size: $n_o + n_p + m_{po}$

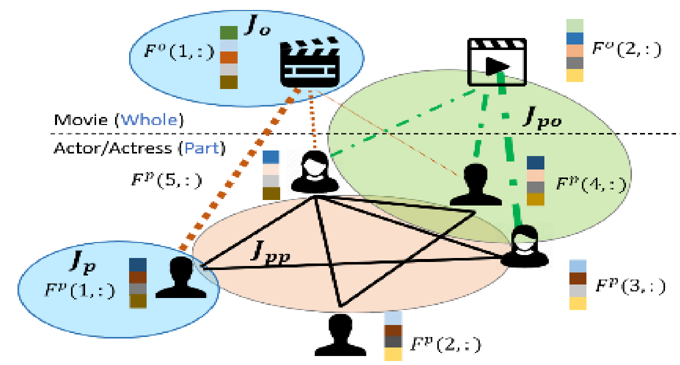
- PAROLE scales linearly w.r.t. part-whole graph size

Roadmap

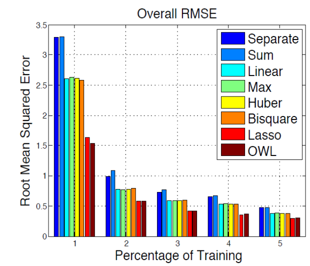
- Motivations
- PAROLE -- Modeling
- PAROLE -- Optimization
- Empirical Evaluations
- **Conclusions**

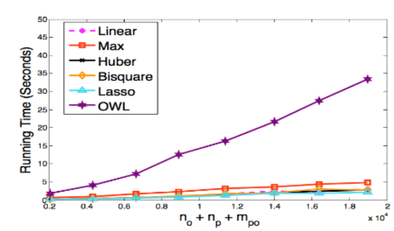
Conclusions -- PAROLE

- **Goals**: leverage potentially non-linear part-whole relationships for outcome prediction



- **Solutions**: PAROLE
  - **Modeling**
    - Part-whole relationship
    - Part-part interdependency





  - **Optimization**
    - Block coordinate descent
    - Converges to a coordinate-wise minimum point
    - Scales linearly w.r.t. the part-whole graph size

DATA Lab