# Is the Whole Greater Than the Sum of Its Parts?

Liangyue Li
Arizona State University
liangyue@asu.edu

Hanghang Tong*
Arizona State University
hanghang.tong@asu.edu

Yong Wang
HKUST
ywangct@cse.ust.hk

Conglei Shi
IBM Research
shiconglei@gmail.com

Nan Cao
Tongji University
nan.cao@gmail.com

Norbou Buchler
US Army Research Laboratory
norbou.buchler.civ@mail.mil

## ABSTRACT

The PART-WHOLE relationship routinely finds itself in many disciplines, ranging from collaborative teams, crowdsourcing, autonomous systems to networked systems. From the algorithmic perspective, the existing work has primarily focused on predicting the outcomes of the whole and parts, by either *separate* models or *linear joint* models, which assume the outcome of the parts has a linear and independent effect on the outcome of the whole. In this paper, we propose a joint predictive method named PAROLE to simultaneously and mutually predict the part and whole outcomes. The proposed method offers two distinct advantages over the existing work. First (*Model Generality*), we formulate joint PART-WHOLE outcome prediction as a generic optimization problem, which is able to encode a variety of complex relationships between the outcome of the whole and parts, beyond the linear independence assumption. Second (*Algorithm Efficacy*), we propose an effective and efficient block coordinate descent algorithm, which is able to find the coordinate-wise optimum with a linear complexity in both time and space. Extensive empirical evaluations on real-world datasets demonstrate that the proposed PAROLE (1) leads to consistent prediction performance improvement by modeling the non-linear part-whole relationship as well as part-part interdependency, and (2) scales linearly in terms of the size of the training dataset.

## KEYWORDS

Joint predictive model; part-whole relationship

## 1 INTRODUCTION

The great Greek philosopher Aristotle articulated more than 2,000 years ago that "*the whole is greater than the sum of its parts*". This is probably most evident in *teams*, which, through appropriate synergy, promise a collective outcome (i.e., team performance) that is superior than the simple addition of what each individual team member could achieve (i.e., individual productivity). For example, in the scientific community, the new breakthrough is increasingly resulting from the teamwork, compared with individual researcher's sole endevour [20]; in professional sports (e.g., NBA), the peak performance of a grass-root team is often attributed to the harmonic teamwork between the team players rather than the individual player's capability. Beyond teams, the *part-whole* relationship also routinely finds itself in other disciplines, ranging from crowdsourcing (e.g., Community-based Question Answering (CQA) sites [23]), collective decision-making in autonomous system (e.g., a self-orchestrated swarm of drones[1]), to reliability assessment of a networked system of components [4, 24].

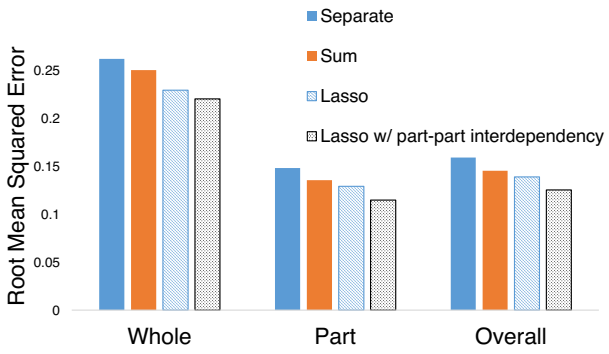



**Figure 1: Prediction error comparison on *Movie* dataset. Lower is better. Best Viewed in Color. The right two bars are the proposed methods, which encode the non-linear part-whole relationship and the non-linearity with part-part interdependency respectively.**

From the algorithmic perspective, an interesting problem is to predict the outcome of the whole and/or parts [7]. In organizational teams, it is critical to appraise the individual performance, its contribution to the team outcome as well as the team's overall performance [15]. In the emerging field of the "science of science", the dream of being able to predict breakthroughs, e.g. predicting the likelihood of a researcher making disruptive contributions and foreseeing the future impact of her research products (e.g., manuscripts, proposals, system prototypes) pervades almost all aspects of modern science [5]. In Community-based Question Answering (CQA) sites, predicting the long-term impact of a question (*whole*) and its associated answers (*parts*) enables users to spot valuable questions and answers at an early stage. Despite much progress has been

*To whom correspondence should be addressed.


*KDD '17, August 13–17, 2017, Halifax, NS, Canada*

DOI: 10.1145/3097983.3098006

[1]CBS 60 minutes report: http://www.cbsnews.com/news/60-minutes-autonomous-drones-set-to-revolutionize-military-technology/

made, the existing work either develop separate models for predicting the outcome of whole and parts without explicitly utilizing the part-whole relationship [10, 14], or implicitly assume the outcome of the whole is a *linear* sum of the outcome of the parts [23], which might oversimplify the complicated part-whole relationships (e.g., non-linearity).

The key to address these limitations largely lies in the answers to the following questions, i.e., to what extent does the outcome of parts (e.g., individual productivity) and that of the whole (e.g., team performance) correlated, beyond the existing linear, independency assumption? How can we leverage such potentially non-linear and interdependent 'coupling' effect to mutually improve the prediction of the outcome of the whole and parts collectively? This is exactly the focus of this paper, which is highly challenging for the following reasons. First (*Modeling Challenge*), the relationship between the parts outcome and whole outcome might be complicated, beyond the simple addition or linear combination. For example, the authors in [22] empirically identified a non-linear correlation between the impacts of questions and the associated answers, that is, the impact of a question is much more strongly correlated with that of the *best* answer it receives, compared with the *average* impact of its associated answers. However, how to leverage such non-linear relationship between the parts and whole outcome has largely remained open. For teams, the team performance might be mainly dominated by a few top-performing team members, and/or be hindered by one or more struggling team members (i.e., the classic Wooden Bucket Theory, which says that "A bucket (whole) can only fill with the volume of water the shortest plank (parts) allows"). Moreover, the composing parts of the whole might not be independent with each other. In a networked system, the composing parts are connected with each other via an underlying network. Such part-part interdependency could have a profound impact on both the part outcome correlation as well as each part's contribution to the whole outcome. How can we mathematically encode the non-linear part-whole relationship as well as part-part interdependency? Second (*Algorithmic Challenge*), the complicated part-whole relationship (i.e., non-linearity and interdependency) also poses an algorithmic challenge, as it will inevitably increase the complexity of the corresponding optimization problem. How can we develop scalable algorithms whose theoretic properties are well-understood (e.g., the convergence, the optimality, and the complexity)?

To address these challenges, in this paper, we propose a joint predictive model named PAROLE to simultaneously and mutually predict the part and whole outcomes. First, *model generality*, the proposed model is flexible in admitting a variety of linear as well as non-linear relationships between the parts and whole outcomes, including *maximum aggregation, linear aggregation, sparse aggregation, ordered sparse aggregation* and *robust aggregation*. Moreover, it is able to characterize part-part interdependency via a graph-based regularization, which encourages the tightly connected parts to share similar outcomes as well as have similar effect on the whole outcome. Second, *algorithm efficacy*, we propose an effective and efficient block coordinate descent optimization algorithm, which converges to the coordinate-wise optimum with a linear complexity.

The main contributions of the paper can be summarized as follows:

**Table 1: Table of symbols**

| Symbols | Definition |
|---|---|
| $\mathbf{F}^o$, $\mathbf{F}^p$ | feature matrices for whole and part entities |
| $\mathbf{y}^o$, $\mathbf{y}^p$ | impact vectors for whole and part entities |
| $\mathcal{O} = \{o_1, o_2, \ldots, o_{n_o}\}$ | set of whole entities |
| $\mathcal{P} = \{p_1, p_2, \ldots, p_{n_p}\}$ | set of part entities |
| $\phi(\cdot)$ | whole to parts mapping function |
| $G^p$ | the network connectivity among part entities |
| $a^i_j$ | the contribution of part $p_j$ to whole $o_i$ |
| $n_o/n_p$ | number of whole/part entities |
| $\mathrm{Agg}(\cdot)$ | the function that aggregates parts outcome |
| $e_i$ | predicted whole outcome using whole feature vs. predicted whole outcome using aggregated parts outcome |

- **Models.** We propose a joint predictive model (PAROLE) that is able to admit a variety of linear as well as non-linear part-whole relationships and encode the part-part interdependency.
- **Algorithms and Analysis.** We propose an effective and efficient block coordinate descent optimization algorithm that converges to the coordinate-wise optimum with a linear complexity in both time and space.
- **Empirical Evaluations.** We conduct extensive empirical studies on several real-world datasets and demonstrate that the proposed PAROLE achieves consistent prediction performance improvement and scales linearly. See Fig. 1 for some sampling results.

The rest of the paper is organized as follows. Section 2 formally defines the Part-Whole Outcome Prediction problem. Section 3 introduces the proposed PAROLE model and section 4 presents the optimization algorithm with analysis. The empirical evaluation results are given in Section 5. After reviewing related work in Section 6, we conclude the paper in Section 7.

## 2 PROBLEM DEFINITION

The main symbols are summarized in Table 1. We use bold capital letters (e.g., $\mathbf{A}$) for matrices and bold lowercase letters (e.g, $\mathbf{w}$) for vectors. We index the elements in a matrix using a convention similar to Matlab, e.g., $\mathbf{A}(:, j)$ is the $j^{th}$ column of $\mathbf{A}$, etc. The vector obtained by sorting the components in non-increasing order of $\mathbf{x}$ is denoted by $\mathbf{x}_\downarrow$. Such sorting operation can be defined by a permutation matrix $\mathbf{P}_\mathbf{x}$, i.e., $\mathbf{P}_\mathbf{x}\mathbf{x} = \mathbf{x}_\downarrow$. We use $\mathcal{K}_{m+}$ to denote the *monotone non-negative cone*, i.e., $\mathcal{K}_{m+} = \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq x_2 \geq \ldots x_n \geq 0\} \subset \mathbb{R}^n_+$. Similarly, we use $\mathcal{K}_m$ for the monotone cone.

We consider predicting the outcome for both the whole and their composing parts. Fig. 2 presents an illustrative example, which aims to predict the popularity (e.g., Facebook likes) of a particular movie (*whole*) and the popularities of the participating actors/actresses (*parts*). We denote the set of whole entities by $\mathcal{O} = \{o_1, o_2, \ldots, o_{n_o}\}$, and denote the set of part entities by $\mathcal{P} = \{p_1, p_2, \ldots, p_{n_p}\}$, where $n_o$ and $n_p$ are the number of the whole and parts, respectively. To specify the part-whole associations, we also define a mapping function $\phi$ that maps a whole entity to the set of its composing parts, e.g., $\phi(o_i) = \{p_{i_1}, p_{i_2}, \ldots, p_{i_{n_i}}\}$ (i.e., the edges between a movie and actors/actresses in Fig. 2). Note that the two sets $\phi(o_i)$ and $\phi(o_j)$ might have overlap. In the example of movies as whole

entities, one actor could participate in multiple movies. Let $\mathbf{F}^o$ be the feature matrix for the whole entities, where the $i^{th}$ row $\mathbf{F}^o(i,:)$ is the feature vector for the $i^{th}$ whole entity. Similarly, let $\mathbf{F}^p$ be the feature matrix for the part entities, where the $j^{th}$ row $\mathbf{F}^p(j,:)$ is the feature vector for the $j^{th}$ part entity. The outcome vector of the whole entities is denoted as $\mathbf{y}^o$ and the outcome vector of the part entities is denoted as $\mathbf{y}^p$. In addition, we might also observe a network connectivity among the part entities, denoted as $G^p$. In the movie example, the network $G^p$ could be the collaboration network among the actors/actresses (the connections among the actors/actresses in Fig. 2).

With the above notations, we formally define our Part-Whole Outcome Prediction problem as follows:

Problem 1. *Part-Whole Outcome Prediction*

**Given:** *the feature matrix for the whole/part entities $\mathbf{F}^o$/$\mathbf{F}^p$, the outcome vector for the whole/part entities $\mathbf{y}^o$/$\mathbf{y}^p$, the whole to part mapping function $\phi$, and the parts' network $G^p$ (optional);*

**Predict:** *the outcome of new whole and parts' entities.*

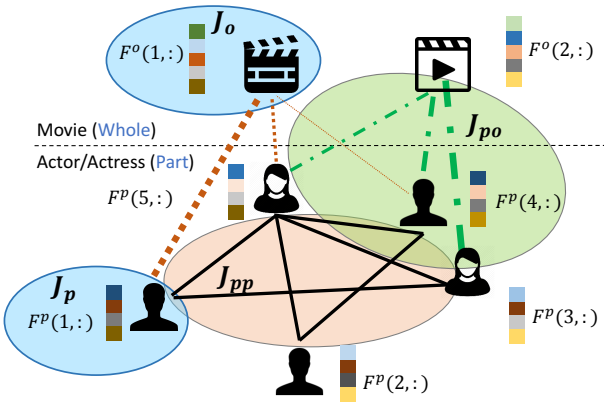



**Figure 2: An illustrative example of part-whole outcome prediction where movies are the whole entities and the actors/actresses are the part entities. The four shadowed ellipses correspond to the key sub-objectives in our proposed PAROLE model (Eq. (1), Sec. 3.1.).**

## 3 PROPOSED MODEL – PAROLE

In this section, we present our joint predictive model PAROLE to simultaneously and mutually predict the outcome of the whole and parts. We first formulate it as a generic optimization problem, and then present the details on how to instantiate the part-whole relationship and part-part interdependency, respectively.

### 3.1 A Generic Joint Prediction Framework

In order to fully embrace the complexity of the part-whole and part-part relationship, our joint predictive model should meet the following desiderata.

First (*part-whole relationship*), the outcome of the whole and that of the parts might be strongly correlated with each other. For example, the team outcome is usually a collective effort of the team members. Consequently, the team performance is likely to be correlated/coupled with each individual's productivity, which might be beyond a simple linear correlation. This is because a few top-performing team members might dominate the overall team performance, or reversely, a few struggling team members might drag down the performance of the entire team. Likewise, in scientific community, a scientist's reputation is generally built by one or a few of her highest-impact work. Our joint predictive model should have the capability to encode such non-linear part-whole relationships, so that the prediction of the parts outcome and that of the whole can mutually benefit from each other.

Second (*part-part interdependency*), the composing parts of a whole entity might be interdependent/interconnected via an underlying network, e.g., the collaboration network among the actors/actresses. The part-part interdependency could have a profound impact on the part-whole outcome prediction performance. That is, not only might the closely connected parts have similar effect on the whole outcome, but also these parts are very likely to share similar outcomes between themselves. Therefore, it is desirable to encode the part-part interdependency in the joint model to boost the prediction performance.

With these design objectives in mind, we propose a generic framework for the joint predictive model as follows:

$$\begin{aligned}\min_{\mathbf{w}^o,\mathbf{w}^p} \mathcal{J} = & \underbrace{\frac{1}{n_o}\sum_{i=1}^{n_o}\mathcal{L}[f(\mathbf{F}^o(i,:),\mathbf{w}^o),\mathbf{y}^o(i))]}_{\mathcal{J}_o\text{: predictive model for whole entities}} \\ & + \underbrace{\frac{1}{n_p}\sum_{i=1}^{n_p}\mathcal{L}[f(\mathbf{F}^p(i,:),\mathbf{w}^p),\mathbf{y}^p(i))]}_{\mathcal{J}_p\text{: predictive model for part entities}} \\ & + \underbrace{\frac{\alpha}{n_o}\sum_{i=1}^{n_o}h(f(\mathbf{F}^o(i,:),\mathbf{w}^o),\text{Agg}(\phi(o_i)))}_{\mathcal{J}_{po}\text{: part-whole relationship}} \\ & + \underbrace{\frac{\beta}{n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{n_p}G^p_{ij}g(f(\mathbf{F}^p(i,:),\mathbf{w}^p),f(\mathbf{F}^p(j,:),\mathbf{w}^p))}_{\mathcal{J}_{pp}\text{: part-part interdependency}} \\ & + \underbrace{\gamma(\Omega(\mathbf{w}^o)+\Omega(\mathbf{w}^p))}_{\mathcal{J}_r\text{: parameter regularizer}}\end{aligned} \tag{1}$$

where the objective function is a sum of five sub-objective functions. The first two sub-objectives $\mathcal{J}_o$ and $\mathcal{J}_p$ (the two blue shadowed ellipses in Fig. 2) minimize the training loss for whole and parts outcome predictions, where $f(\cdot,\cdot)$ is the prediction function parameterized by $\mathbf{w}^o$ and $\mathbf{w}^p$. The prediction function could be either linear or non-linear; and $\mathcal{L}(\cdot)$ is a loss function, e.g., squared loss for regression or logistic loss for classification. The core of the objective function is the third term $\mathcal{J}_{po}$ (the green shadowed ellipse in Fig. 2) and the fourth term $\mathcal{J}_{pp}$ (the pink shadowed ellipse in Fig. 2). $\mathcal{J}_{po}$ characterizes the part-whole relationship, where $\text{Agg}(\cdot)$ is a function that aggregates the predicted outcomes of all

the composing parts for the whole to a single outcome, e.g., maximum, summation/mean or more complicated aggregations; and $h(\cdot)$ function measures the correlation between the predicted whole outcome and the aggregated predicted parts outcome. In $\mathcal{J}_{pp}$, the function $g(\cdot)$ characterizes the relationship of the predicted outcomes of parts $i$ and $j$ based on their connectivity $G^p_{ij}$, such that tightly connected parts would share similar outcomes. Lastly, $\mathcal{J}_r$ regularizes $\mathbf{w}^o$ and $\mathbf{w}^p$ to prevent overfitting. The regularization parameters $\alpha$, $\beta$ and $\gamma$ are used to balance the relative importance of each aspect.

*Remarks:* Depending on the specific choices of the aggregation function $\text{Agg}(\cdot)$ and the $h(\cdot)$ function, the proposed model in Eq. (1) is able to admit a variety of part-whole relationships, which we elaborate below.

## 3.2 Modeling Part-Whole Relationships

**Overview.** In this subsection, we give the instantiations for a variety of part-whole relationships. For each whole entity $o_i$, define $e_i$ as follows:

$$e_i = \mathbf{F}^o(i,:)\mathbf{w}^o - \text{Agg}(o_i) \tag{2}$$

which measures the difference between the predicted whole outcome using whole features (i.e., $\mathbf{F}^o(i,:)\mathbf{w}^o$) and predicted whole outcome using aggregated parts outcome (i.e., $\text{Agg}(o_i)$). Our proposed model will be able to characterize a variety of part-whole relationship, by using (a) different aggregation functions $\text{Agg}(\cdot)$ with augmented regularizations; and (b) different loss functions on $e_i$ (e.g., squared loss or robust estimator).

**Maximum aggregation.** Let us first consider using maximum as the aggregation function, which can model the correlation between the whole outcome and the maximum parts outcome. Given that the max function is not differentiable, we propose to approximate it with a differentiable function that will largely facilitate the optimization process. In details, we propose to use the smooth "soft" maximum function, which was first used in economic literature for consumer choice [17]: $\max(x_1, x_2, \ldots, x_n) \approx \ln(\exp(x_1)+\exp(x_2)+\ldots+\exp(x_n))$, where the maximum is approximated by summing up the exponential of each item followed by a logarithm. With this, we define the maximum aggregation function as follows:

$$\text{Agg}(o_i) = \ln\Big(\sum_{j\in\phi(o_i)} \exp(\mathbf{F}^p(j,:)\mathbf{w}^p)\Big) \tag{3}$$

which approximates the maximum predicted parts outcome. The part-whole relationship with maximum aggregation can be formulated as follows:

$$\mathcal{J}_{po} = \frac{\alpha}{2n_o}\sum_{i=1}^{n_o} e_i^2 \tag{4}$$

where we use the squared loss to measure the difference between the predicted whole outcome and the predicted approximated maximum parts outcome.

For the remaining part-whole relationships, we instantiate $\text{Agg}(o_i)$ using a linear function as follows:

$$\text{Agg}(o_i) = \sum_{j\in\phi(o_i)} a^i_j \mathbf{F}^p(j,:)\mathbf{w}^p \tag{5}$$

where each $a^i_j$ is the weight of a particular part $j$'s contribution to the whole $o_i$'s outcome. Defining $\mathbf{a}_i$ as the vector whose components are $a^i_j$, $j \in \phi(o_i)$ and by imposing (i) different loss functions on $e_i$, and/or (ii) different norms on $\mathbf{a}_i$, we can model either linear or nonlinear part-whole relationships.

**Linear aggregation.** In this scenario, the whole outcome is a weighted linear combination of the parts outcome, where the weights determine each individual part's contribution to the whole outcome. The intuition of linear aggregation is that , in contributing to the final whole outcome, some parts play more important roles than the others. This part-whole relationship can be formulated as follows:

$$\mathcal{J}_{po} = \frac{\alpha}{2n_o}\sum_{i=1}^{n_o} e_i^2 \tag{6}$$

where we use the squared loss to measure the difference between the whole outcome and the aggregated parts outcome.

*Remark:* this formulation generalizes several special part-whole relationships. The expression that "*the whole is the sum of its parts*" is a special case of Eq. (6) where various $a^i_j$ is 1, which we refer to as *Sum* in the empirical study. The average coupling formulated in [23] is also its special case with $a^i_j = \frac{1}{|o_i|}$. Instead of fixing the weights, Eq. (6) allows the model to learn to what extent each part contributes to the prediction of the whole outcome. Nonetheless, in all these variants, we have assumed that the part outcomes always have a *linear* effect on the whole outcome.

**Sparse aggregation.** The above linear aggregation assumes that each part would contribute to the whole outcome, which might not be the case as some parts have little or no effect on the whole outcome. This scenario can be seen in large teams, where the team performance could be primarily determined by a few members, who could either make or break the team performance. To encourage such a sparse selection among the composing parts of a whole entity, a natural choice is to introduce the $l_1$ norm on the vector $\mathbf{a}_i$ [18]:

$$\mathcal{J}_{po} = \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\big(\frac{1}{2}e_i^2 + \lambda|\mathbf{a}_i|_1\big) \tag{7}$$

where the $l_1$ norm can shrink some part contributions to exactly zero and the parameter $\lambda$ controls the degree of sparsity.

**Ordered sparse aggregation.** In some cases, the team performance (i.e., the whole outcome) is determined by not only a few key members, but also the structural hierarchy between such key members within the organization. To model such parts performance ranking in addition to the sparse selection, we adopt the ordered weighted $l_1$ norm (OWL) [25] that is able to give more weights to those parts with bigger effect on the whole outcome. Such part-whole relationship can be formulated as follows:

$$\mathcal{J}_{po} = \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\big(\frac{1}{2}e_i^2 + \lambda\Omega_{\mathbf{w}}(\mathbf{a}_i)\big) \tag{8}$$

where $\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^n |x|_{[i]} w_i = \mathbf{w}^T|\mathbf{x}|_\downarrow$ is the ordered weighted $l_1$ norm, where $|x|_{[i]}$ is the $i$-th largest component of the vector $|\mathbf{x}|$ and $\mathbf{w} \in \mathcal{K}_{m+}$ is a vector of non-increasing non-negative weights.

**Robust aggregation.** In all the above formulations, we model the difference between the whole outcome and the aggregated parts outcome using squared loss, which is prone to outlying parts/wholes.

To address this issue, we employ robust regression models [9] to reduce the effect of outliers as follows:

$$\mathcal{J}_{po} = \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\rho(e_i) \tag{9}$$

where $\rho(\cdot)$ is a nonnegative and symmetric function that gives the contribution of each residual $e_i$ to the objective function. In this paper, we consider two robust estimators, namely Huber and Bisquare estimators as follows:

| 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 $\rho_B(e)$ | $\frac{t^2}{6}\left\{1-[1-(\frac{e}{t})^2]^3\right\}$ | $\frac{t^2}{6}$ |

where the value $t$ is a tuning constant. Smaller $t$ values have more resistance to outliers.

### 3.3 Modeling Part-Part Interdependency

As mentioned in Sec. 3.1, the part-part interdependency, if exists, can play two roles in the part-whole outcome predictions, i.e., closely connected parts would (A) have similar effect on the whole outcome and (B) share similar part outcomes between themselves.

**A - The effect on the whole outcome:** the closely connected parts might have similar impact on the whole outcome. It turns out we can use the same method to model such a part-part effect for various aggregation methods in Sec. 3.2. Let us take *sparse aggregation* as an example and instantiate the term $\mathcal{J}_{po}$ in Eq. (1) as follows:

$$\mathcal{J}_{po} = \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\left[\frac{1}{2}e_i^2 + \lambda|\mathbf{a}_i|_1 + \frac{1}{2}\sum_{k,l\in\phi(o_i)} G^p_{kl}(a^i_k - a^i_l)^2\right] \tag{10}$$

where if the two parts $k$ and $l$ of $o_i$ are tightly connected, i.e., $G^p_{k,l}$ is large, then the difference between their impacts on the whole outcome, $a^i_k$ and $a^i_l$, is small.

**B - The effect on the parts' outcomes:** the tightly connected parts might share similar outcomes themselves. Such parts outcome similarity can be instantiated by a graph regularization as follows:

$$\mathcal{J}_{pp} = \frac{\beta}{2n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{n_p} G^p_{ij}(\mathbf{F}^p(i,:)\mathbf{w}^p - \mathbf{F}^p(j,:)\mathbf{w}^p)^2 \tag{11}$$

where tightly connected two parts $i$ and $j$ with large $G^p_{k,l}$ is encouraged to be closer to each other in the output space, i.e., with similar predicted outcomes.

## 4 OPTIMIZATION ALGORITHM

In this section, we propose an effective and efficient block coordinate descent optimization algorithm to solve the joint prediction framework in Eq. (1), followed by the convergence and complexity analysis.

### 4.1 Block Coordinate Descent Algorithm

The proposed Eq. (1) is general, being able to admit a variety of different separate models ($\mathcal{J}_o$ and $\mathcal{J}_p$) as well as part-whole relationship ($\mathcal{J}_{po}$). Let us first present our algorithm to solve a specific instance of Eq. (1) by instantiating it using linear predictive functions, squared loss and sparse aggregation as follows:

$$\begin{aligned}\min_{\mathbf{w}^o,\mathbf{w}^p}\ & \frac{1}{2n_o}\sum_{i=1}^{n_o}(\mathbf{F}^o(i,:)\mathbf{w}^o - \mathbf{y}^o(i))^2 + \frac{1}{2n_p}\sum_{i=1}^{n_p}((\mathbf{F}^p(i,:)\mathbf{w}^p - \mathbf{y}^p(i))^2\\ &+ \frac{\beta}{2n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{n_p} G^p_{ij}(\mathbf{F}^p(i,:)\mathbf{w}^p - \mathbf{F}^p(j,:)\mathbf{w}^p)^2 + \frac{\gamma}{2}(\|\mathbf{w}^o\|_2^2 + \|\mathbf{w}^p\|_2^2)\\ &+ \frac{\alpha}{n_o}\sum_{i=1}^{n_o}\left[\frac{1}{2}e_i^2 + \lambda|\mathbf{a}_i|_1 + \frac{1}{2}\sum_{k,l\in\phi(o_i)} G^p_{kl}(a^i_k - a^i_l)^2\right]\end{aligned} \tag{12}$$

In the formulation, we identify three coordinate blocks, namely $\mathbf{w}^o$, $\mathbf{w}^p$ and various $a^i_j$. We propose a block coordinate descent (BCD) algorithm to optimize Eq. (12) by updating one coordinate block while fixing the other two.

**1. Updating $\mathbf{w}^o$ while fixing others:** Observing that only $\mathcal{J}_o$, $\mathcal{J}_{po}$ and $\mathcal{J}_r$ are functions of $\mathbf{w}^o$, we have

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial\mathbf{w}^o} &= \frac{\partial\mathcal{J}_o}{\partial\mathbf{w}^o} + \frac{\partial\mathcal{J}_{po}}{\partial\mathbf{w}^o} + \frac{\partial\mathcal{J}_r}{\partial\mathbf{w}^o}\\ &= \frac{1}{n_o}(\mathbf{F}^o)'(\mathbf{F}^o\mathbf{w}^o - \mathbf{y}^o) + \gamma\mathbf{w}^o + \frac{\alpha}{n_o}(\mathbf{F}^o)'(\mathbf{F}^o\mathbf{w}^o - \mathbf{M}\mathbf{F}^p\mathbf{w}^p)\end{aligned} \tag{13}$$

where $\mathbf{M}$ is a $n_o$ by $n_p$ sparse matrix with $\mathbf{M}(i,j) = a^i_j$, for $j \in \phi(o_i)$. We then update $\mathbf{w}^o$ as $\mathbf{w}^o \leftarrow \mathbf{w}^o - \tau\frac{\partial\mathcal{J}}{\partial\mathbf{w}^o}$, where $\tau$ is the step size.

**2. Updating $\mathbf{w}^p$ while fixing others:** The sub-objective functions that are related to $\mathbf{w}^p$ are $\mathcal{J}_p$, $\mathcal{J}_{pp}$, $\mathcal{J}_{po}$ and $\mathcal{J}_r$. Therefore,

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial\mathbf{w}^p} &= \frac{\partial\mathcal{J}_p}{\partial\mathbf{w}^p} + \frac{\partial\mathcal{J}_{pp}}{\partial\mathbf{w}^p} + \frac{\partial\mathcal{J}_{po}}{\partial\mathbf{w}^p} + \frac{\partial\mathcal{J}_r}{\partial\mathbf{w}^p}\\ &= \frac{1}{n_p}(\mathbf{F}^p)'(\mathbf{F}^p\mathbf{w}^p - \mathbf{y}^p) + \frac{\beta}{n_p}(\mathbf{F}^p)'\mathcal{L}^p\mathbf{F}^p\mathbf{w}^p + \gamma\mathbf{w}^p\\ &\quad - \frac{\alpha}{n_o}(\mathbf{F}^p)'\mathbf{M}'(\mathbf{F}^o\mathbf{w}^o - \mathbf{M}\mathbf{F}^p\mathbf{w}^p)\end{aligned} \tag{14}$$

where $\mathcal{L}^p$ is the Laplacian of the graph $G^p$ [1]. Similarly, $\mathbf{w}^p$ can be updated by $\mathbf{w}^p \leftarrow \mathbf{w}^p - \tau\frac{\partial\mathcal{J}}{\partial\mathbf{w}^p}$.

**3. Updating $a^i_j$ while fixing others:** Let us fix a whole $o_i$ and the sub-problem with respect to $\mathbf{a}_i$ becomes:

$$\min_{\mathbf{a}_i}\frac{1}{2}e_i^2 + \lambda|\mathbf{a}_i|_1 + \frac{1}{2}\sum_{k,l\in\phi(o_i)} G^p_{kl}(a^i_k - a^i_l)^2 \tag{15}$$

Observing that the sub-problem is a composite of a non-smooth convex function ($\lambda|\mathbf{a}_i|_1$) and a differentiable convex function (the remaining terms), we update $\mathbf{a}_i$ using the proximal gradient descent method [2]. We first take a gradient step by moving $\mathbf{a}_i$ along the negative direction of the derivative of the smooth part w.r.t. $\mathbf{a}_i$, as follows:

$$\mathbf{z} = \mathbf{a}_i - \tau[e_i(-\mathbf{F}^p(\phi(o_i),:)\mathbf{w}^p) + \mathcal{L}^p_i\mathbf{a}_i] \tag{16}$$

where $\mathcal{L}^p_i$ is a shorthand notation for the Laplacian of the subgraph $G^p(\phi(o_i),\phi(o_i))$. Next, we compute the proximal-gradient update for the $l_1$ norm using soft-thresholding as $\mathbf{a}_i \leftarrow \mathcal{S}_{\tau\lambda}(\mathbf{z})$, where the soft-thresholding operator is defined as follows:

$$[\mathcal{S}_t(\mathbf{z})]_j = \text{sign}(\mathbf{z}_j)(|\mathbf{z}_j| - t)_+, \tag{17}$$

where we use $(x)_+$ as a shorthand for $\max\{x, 0\}$.

We will cycle through the above three steps to update the three coordinate blocks until convergence. The algorithm is summarized in Algorithm 1.

**Algorithm 1** PAROLE - Part-Whole Outcome Predictions

**Input:** (1) the feature matrix for whole/part entities $\mathbf{F}^o/\mathbf{F}^p$,
(2) outcome vector for the whole/part entities $\mathbf{y}^o/\mathbf{y}^p$,
(3) the whole to parts mapping function $\phi$,
(4) the part-part network $G^p$ (optional),
(5) parameters $\alpha, \beta, \gamma, \lambda, \tau$.
**Output:** Model parameters $\mathbf{w}^o$ and $\mathbf{w}^p$.

1: Initialize $\mathbf{w}^o$ and $\mathbf{w}^p$ and $a_j, j \in \phi(o_i), i = 1, \ldots, n_o$
2: **while** Not converged **do**
3: Update $\mathbf{w}^o \leftarrow \mathbf{w}^o - \tau \frac{\partial \mathcal{J}}{\partial \mathbf{w}^o}$
4: Update $\mathbf{w}^p \leftarrow \mathbf{w}^p - \tau \frac{\partial \mathcal{J}}{\partial \mathbf{w}^p}$
5: Update $\mathbf{a}_i$ via proximal gradient descent for $i = 1, \ldots, n_o$

*Remarks:* we want to emphasize that Algorithm 1 provides a general optimization framework that not only works for the formulation with *sparse aggregation* in Eq. (12), but is also applicable to the other part-whole relationships introduced in Sec. 3.2. The only difference is that, since $\mathcal{J}_{po}$ varies for each part-whole relationship, its derivatives w.r.t. the coordinate blocks would also change. Next, for each of the other part-whole relationships, we give their derivative or proximal gradient w.r.t. the three coordinate blocks.

**1. Maximum aggregation:** the derivatives of $\mathcal{J}_{po}$ w.r.t. $\mathbf{w}^o$ and $\mathbf{w}^p$ are as follows:

$$\frac{\partial \mathcal{J}_{po}}{\partial \mathbf{w}^o} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} e_i (\mathbf{F}^o(i,:))'$$

$$\frac{\partial \mathcal{J}_{po}}{\partial \mathbf{w}^p} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} e_i \cdot \frac{\sum_{j \in \phi(o_i)} (\mathbf{F}^p(j,:))' \tilde{\mathbf{y}}_i^p}{\sum_{j \in \phi(o_i)} \tilde{\mathbf{y}}_i^p}$$

where we denote $\tilde{\mathbf{y}}_i^p = \exp(\mathbf{F}^p(j,:)\mathbf{w}^p)$.

**2. Linear aggregation:** the derivatives of $\mathcal{J}_{po}$ w.r.t. $\mathbf{w}^o$ and $\mathbf{w}^p$ are the same as in the *sparse aggregation* case. Its derivative w.r.t. $\mathbf{a}_i$ is same as in Eq. (16) without the following proximal-gradient update.

**3. Ordered sparse aggregation:** the only difference from the *sparse aggregation* lies in the proximal-gradient update for the OWL norm, which can be computed as follows [25]:

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{v}) = \text{sign}(\mathbf{v}) \odot \left( \mathbf{P}_{|\mathbf{v}|}^T \text{proj}_{\mathbb{R}_+^n}(\text{proj}_{\mathcal{K}_m}(|\mathbf{v}|_\downarrow - \mathbf{w})) \right) \quad (18)$$

In the above equation, to compute $\text{prox}_{\Omega_\mathbf{w}}(\mathbf{v})$, we first compute the Euclidean project of $(|\mathbf{v}|_\downarrow - \mathbf{w})$ onto $\mathcal{K}_m$ using the linear time *pool adjacent violators* (PAV) algorithm [16]. This is followed by a projection onto the first orthant by a clipping operation. The resulting vector is sorted back according to the permutation matrix $\mathbf{P}_{|\mathbf{v}|}$ and then element-wisely multiplied by the signs of $\mathbf{v}$.

**4. Robust aggregation:** we compute the gradient of $\mathcal{J}_{po}$ using chain rule as follows:

$$\frac{\partial \mathcal{J}_{po}}{\partial \mathbf{w}^o} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} \frac{\partial \rho(e_i)}{\partial e_i} \frac{\partial e_i}{\partial \mathbf{w}^o}, \quad \frac{\partial \mathcal{J}_{po}}{\partial \mathbf{w}^p} = \frac{\alpha}{n_o} \sum_{i=1}^{n_o} \frac{\partial \rho(e_i)}{\partial e_i} \frac{\partial e_i}{\partial \mathbf{w}^p}$$

$$\frac{\partial \mathcal{J}_{po}}{\partial \mathbf{a}_i} = \frac{\alpha}{n_o} [\frac{\partial \rho(e_i)}{\partial e_i} \frac{\partial e_i}{\partial \mathbf{a}_i} + \mathcal{L}_i^p \mathbf{a}_i]$$

where $\frac{\partial e_i}{\partial \mathbf{w}^o} = \mathbf{F}^o(i,:)'$, $\frac{\partial e_i}{\partial \mathbf{w}^p} = -\sum_{j \in \phi(o_i)} a_j \mathbf{F}^p(j,:)'$, and $\frac{\partial e_i}{\partial \mathbf{a}_i} = -\mathbf{F}^p(\phi(o_i),:)\mathbf{w}^p$; and the gradient of the Huber and Bisquare estimator can be computed as follows:

| Method \ Case | $\lvert e \rvert \le t$ | $\lvert e \rvert > t$ |
|---|---|---|
| Huber $\frac{\partial \rho_H(e)}{\partial e}$ | $e$ | $t \cdot \text{sign}(e)$ |
| Bisquare $\frac{\partial \rho_B(e)}{\partial e}$ | $e[1-(e/t)^2]^2$ | 0 |

## 4.2 Proofs and Analysis

In this subsection, we analyze the proposed PAROLE algorithm in terms of its convergence, optimality and complexity.

First, building upon the proposition from [19], we have the following theorem regarding the proposed Algorithm 1, which says that under a mild assumption, it converges to a local optimum (i.e., coordinate-wise minimum) of Eq. (12).

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

Proof. Omitted for brevity. □

Next, we analyze the complexity of Algorithm 1, which is summarized in Lemma 4.2.

Lemma 4.2. *(*Complexity of PAROLE). Algorithm 1 takes $O(T(n_o d_o + n_p d_p + m_{po} + m_{pp}))$ time for *linear aggregation*, *maximum aggregation*, *sparse aggregation*, and *robust aggregation*, and it takes $O(T(n_o d_o + n_p d_p + C n_p d_p + m_{pp}))$ for *ordered sparse aggregation*, where $d_o$ and $d_p$ are the dimensionality of the whole and part feature vectors, $m_{po} = \sum_i |\phi(o_i)|$ is the number of associations between the whole and parts and $m_{pp}$ is the number of edges in the part-part network, $T$ is the number of iterations, $C = \max_i \log(|\phi(o_i)|)$ is a constant. The space complexity for Algorithm 1 is $O(n_o d_o + n_p d_p + m_{po} + m_{pp})$ for all the part-whole relationships.

Proof. Omitted for brevity. □

*Remarks:* suppose we have a conceptual part-whole graph $G = \{\mathcal{O}, \mathcal{P}\}$, which has $n_o$ nodes for the whole entities and $n_p$ nodes for the part entities, $m_{po}$ links from whole nodes to their composing parts nodes and $m_{pp}$ links in the part-part networks. The Lemma 4.2 says that PAROLE scales linearly w.r.t. the size of this part-whole graph in both time and space.

**Table 2: Summary of Datasets.**

| Data | Whole | Part | # of whole | # of part |
|---|---|---|---|---|
| *Math* | Question | Answer | 16,638 | 32,876 |
| *SO* | Question | Answer | 1,966,272 | 4,282,570 |
| *DBLP* | Author | Paper | 234,681 | 129,756 |
| *Movie* | Movie | Actors/Actresses | 5,043 | 37,365 |

# 5 EXPERIMENTS

In this section, we present the empirical evaluation results. The experiments are designed to evaluate the following aspects:

- *Effectiveness:* how accurate is the proposed PAROLE algorithm for predicting the outcomes of parts and whole?
- *Efficiency:* how fast and scalable is the proposed PAROLE algorithm?

## 5.1 Datasets

The real-world datasets used for evaluations are as follows:

*CQA.* We use Mathematics Stack Exchange (*Math*) and Stack Overflow (*SO*) data from [23]. The questions are whole and answers are parts both with voting scores as outcome. For each question, we treat all the answers it receives as its composing parts. The extracted features are described in [23].

*DBLP.* DBLP dataset provides the bibliographic information of computer science research papers. We treat authors as whole with $h$-index as outcome and papers as parts with citation counts as outcome. For each author, his/her composing parts are the papers s/he has co-authored. Paper features include temporal attributes and author features include productivity and social attributes.

*Movie.* We crawl the metadata of 5,043 movies with budget information[2] from IMDb website. The meta information includes movie title, genres, cast, budget, etc. We treat movies as whole and the actors/actresses as parts both with the number of Facebook likes as the outcome. For each movie, we treat its cast as the composing parts. Movie features include contextual attributes and actors/actresses features include productivity and social attributes.

The statistics of these datasets are summarized in Table 2. For each dataset, we first sort the whole in chronological order, gather the first $x$ percent of whole and their corresponding parts as training examples and always test on the last 10% percent of whole and their corresponding parts. The percentage of training $x$ could vary. The root mean squared error (RMSE) between the actual outcomes and the predicted ones is adopted for effectiveness evaluation. The parameters are set for each method on each dataset via a grid search.

*Repeatability of experimental results:* all the datasets are publicly available. We will release the datasets and code of the proposed algorithms through authors' website after the paper is published. The experiments are performed on a Windows machine with four 3.5GHz Intel Cores and 256GB RAM.

## 5.2 Effectiveness Results

We compare the effectiveness of the following methods:

(1) *Separate:* train a linear regression model for parts and whole separately.
(2) *Sum:* a joint model with Sum part-whole relationship.
(3) *Linear:* our PAROLE with linear aggregation.
(4) *Max:* our PAROLE with maximum aggregation.
(5) *Huber:* our PAROLE with robust Huber estimator.
(6) *Bisquare:* our PAROLE with robust Bisquare estimator.
(7) *Lasso:* our PAROLE with sparse aggregation.
(8) *OWL:* our PAROLE with ordered sparse aggregation.

**A - Outcome prediction performance:** the RMSE results of all the comparison methods for predicting the outcomes of parts and whole on all the datasets are shown from Fig. 3 to Fig. 6. We draw several interesting observations from these results. First, all the joint prediction models outperform the separate model in most cases, which suggests that the part outcome indeed has a profound impact on the whole outcome, and vice versa. Second, among the joint prediction models, in general, the linear methods (*Sum* and *Linear*) are not as good as the non-linear counterparts (*Max*, *Huber*, *Bisqaure*, *Lasso* and *OWL*), and in some cases (Fig. 3b, Fig. 5b), the linear joint models are even worse than the separate method, which indicates that the part-whole relationship is indeed more complicated than the linear aggregation. Third, among the non-linear methods, a consistent observation across all the datasets is that *Lasso* and *OWL* are the best two methods in almost all the cases. This suggests that the whole outcome is mostly dominated by a few, often high-performing, parts.

**B - The effect of part-part interdependency:** in the proposed joint prediction model, we have hypothesized that the part-part interdependency might help boost the predictions in two ways, i.e., regularizing the parts' contribution to the whole outcome as well as part outcome correlation. Here, we verify and validate to what extent these two aspects contribute to the performance gain, when such part-part interdependency information is available. Fig. 7 shows the results of *Lasso* on the *Movie* dataset with 50% training data. The network among the parts, i.e., actors/actresses, is their collaboration network. The "PAROLE-Basic" does not use the network information. The "PAROLE-GraphForWhole" applies the graph regularization on the parts' contribution to the whole, which brings a 8% overall prediction error reduction. On top of that, "PAROLE-GraphForWhole&Parts" uses the graph regularization on the parts' outcome, which brings a 14.5% decrease in the overall prediction error.

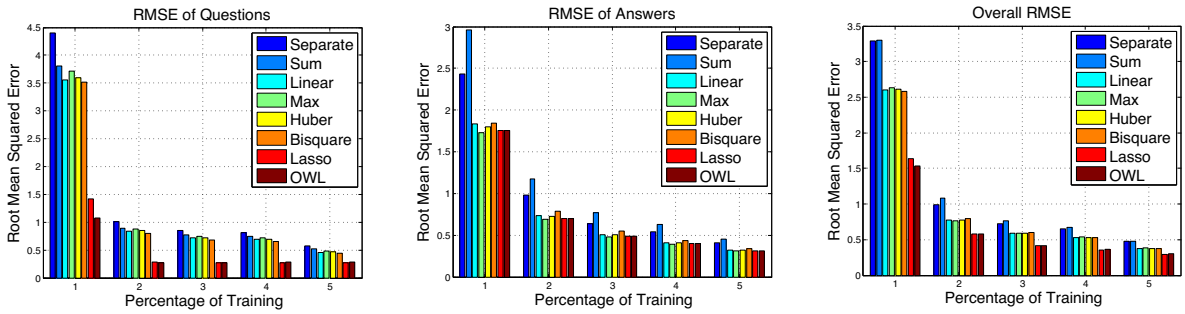
**C - Convergence analysis:** Fig. 8 shows the objective function value vs. the number of iterations on the *SO* dataset using *OWL* with 5% training data. As we can see, the proposed PAROLE algorithm coverges very fast, after 25-30 iterations.

**D - Sensitivity analysis:** to investigate the parameter sensitivity, we perform parametric studies with the two most important parameters in PAROLE, i.e., $\alpha$ that controls the importance of part-whole relationship and $\beta$ that controls the importance of part-part interdependency on the parts outcome. The bowl shaped surface in Fig. 9 suggests that the proposed model can achieve good performance in a large volumn of the parameter space.

## 5.3 Efficiency Results

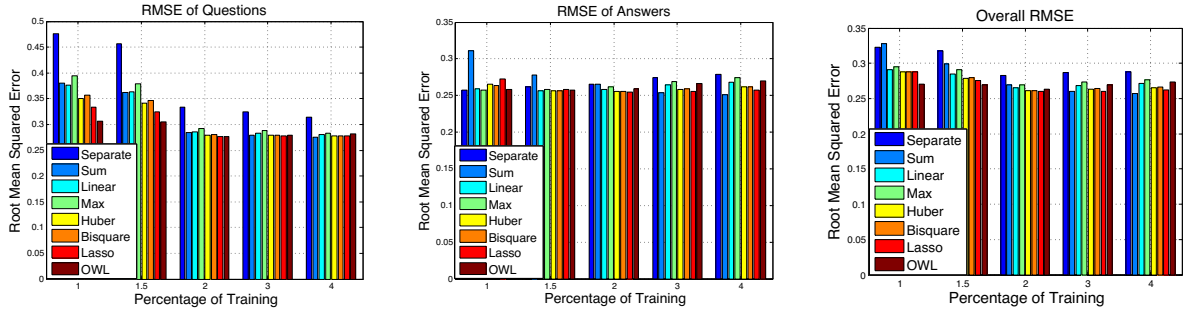
Fig. 10 shows the running time of all the proposed methods with varying size of training data ($n_o + n_p + m_{po}$). We can see that all the proposed methods scale linearly, which is consistent with Lemma 4.2. *OWL* takes the longest time due to the additional sorting operation in the proximal-gradient update for the OWL norm.

[2] http://www.the-numbers.com/movie/budgets/all



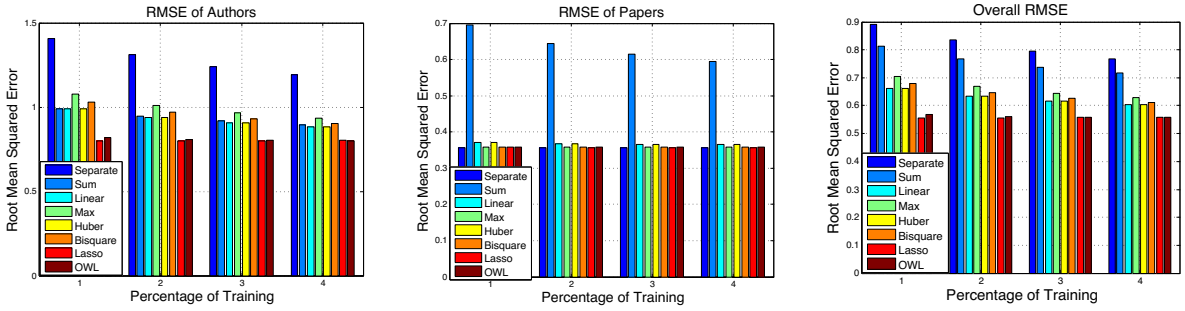

(a) RMSE of question outcome prediction. (b) RMSE of answer outcome prediction. (c) Overall RMSE.

**Figure 3: RMSE comparisons on *Math*. Best viewed in color. From left to right: *Separate*, *Sum*, *Linear*, *Max*, *Huber*, *Bisquare*, *Lasso* and *OWL*.**



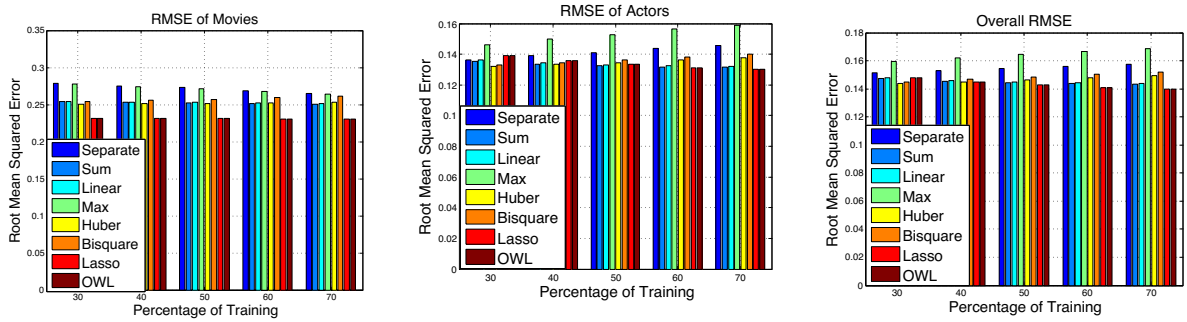

(a) RMSE of question outcome prediction. (b) RMSE of answer outcome prediction. (c) Overall RMSE.

**Figure 4: RMSE comparisons on *SO*. Best viewed in color. From left to right: *Separate*, *Sum*, *Linear*, *Max*, *Huber*, *Bisquare*, *Lasso* and *OWL*.**




(a) RMSE of author outcome prediction. (b) RMSE of paper outcome prediction. (c) Overall RMSE.

**Figure 5: RMSE comparisons on *DBLP*. Best viewed in color. From left to right: *Separate*, *Sum*, *Linear*, *Max*, *Huber*, *Bisquare*, *Lasso* and *OWL*.**




(a) RMSE of movie outcome prediction. (b) RMSE of actors/actress outcome prediction. (c) Overall RMSE.

**Figure 6: RMSE comparisons on *Moive*. Best viewed in color. From left to right: *Separate*, *Sum*, *Linear*, *Max*, *Huber*, *Bisquare*, *Lasso* and *OWL*.**

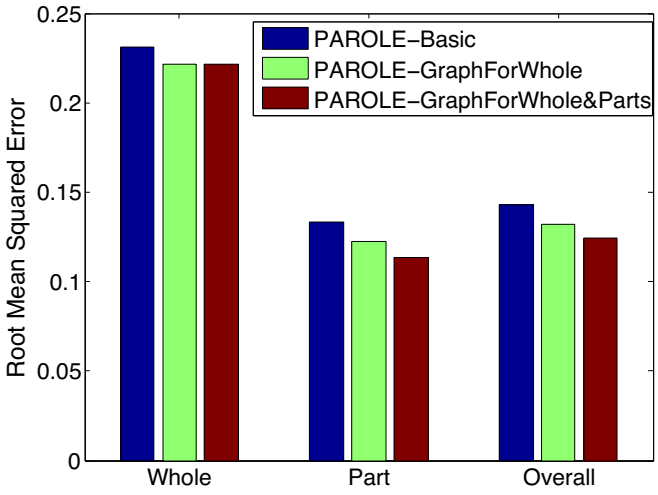



**Figure 7: Performance gain analysis on *Movie*.**

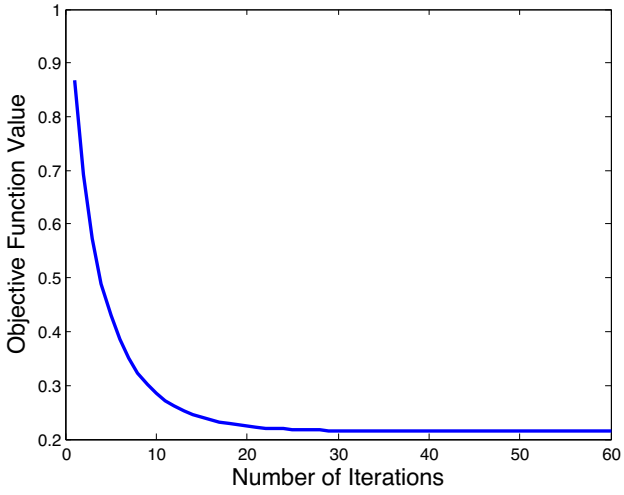



**Figure 8: Convergence analysis on *SO*.**

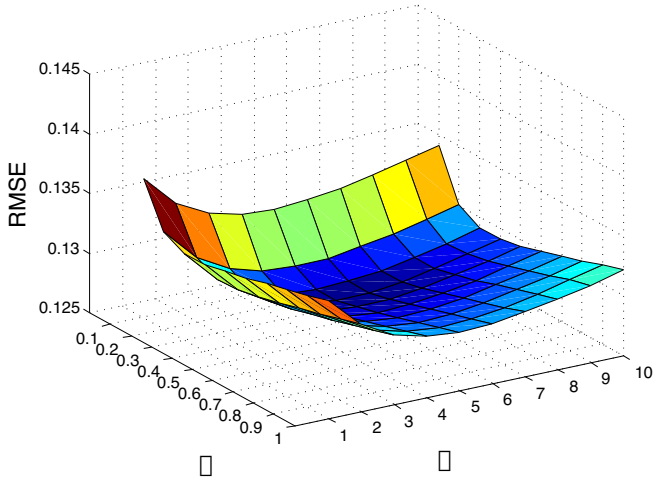



**Figure 9: RMSE with varying $\alpha$ and $\beta$ of *Lasso* on *Movie*.**

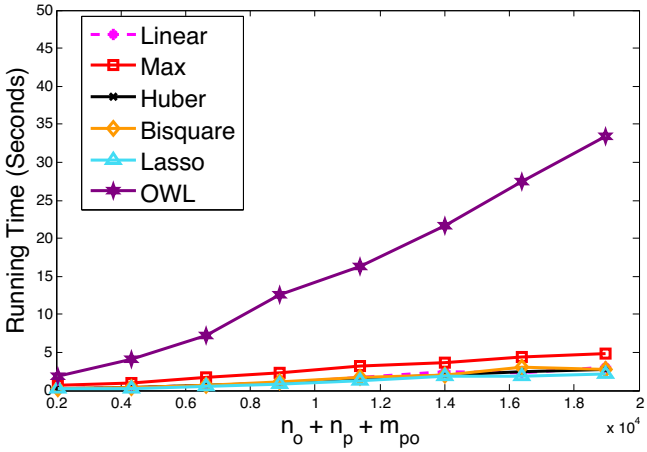



**Figure 10: Scalability plot on *SO*.**

## 6 RELATED WORK

Very few studies have quantitatively examined the part-whole relationships for the purpose of prediction. In CQA sites, empirical studies have shown a strong correlation between the question voting score and the average/maximum answer voting score [22]. Based on this observation, a joint predictive model that leverages question/answer coupling is proposed that is also able to capture the dynamics of the community posts [23]. Related to the part-whole relationship, there is a large body of literature on the collaborative teams. Along the line of team outcome prediction, the iBall [10] model focuses on the long-term impact forecasting of scholarly entities that exploits domain heterogeneity; going beyond the point prediction, the iPath [14] model aims to predict the pathway to impact. Though all the above predictive models can admit non-linear prediction functions for each separate model, they still assume a *linear* relationship between different models (e.g., part vs. whole).

On team formation and outcome optimization, the seminal work in [8] aims to form a team that can cover the required skill sets with strong team cohesion. Some recent work has been focusing on finding a good candidate to replace a team member [11] and enhance the team performance by allowing several enhancement operations, including refinement and expansion [13]. A visual interactive system is developed allowing users to explore and optimize teams [3, 12]. From the multi-task learning [6, 21] perspective, our method explicitly models two types of potentially non-linear task relatedness, i.e., part-whole relationship and part-part interdependency.

## 7 CONCLUSION

In this paper, we propose a joint predictive model PAROLE to simultaneously and mutually predict the parts and whole outcomes. First, *model generality*, the proposed model is able to (i) admit a

variety of linear as well as non-linear relationship between the parts and whole outcome and (ii) characterize part-part interdependency. Second, *algorithm efficacy*, we propose an effective and efficient block coordinate descent optimization algorithm that converges to the coordinate-wise optimum with a linear complexity in both time and space. The empirical evaluations on real-world datasets demonstrate that (i) by modeling the non-linear part-whole relationship and part-part interdependency, the proposed method leads to consistent prediction performance improvement, and (ii) the proposed algorithm scales linearly w.r.t. the size of the training data. In the future, we would like to explore the dynamics of the proposed model as well as the hierarchy in the parts (i.e., the parts of the parts).

## 8 ACKNOWLEDGMENTS

This work is supported by National Science Foundation under Grant No. IIS-1651203, DTRA under the grant number HDTRA1-16-0017, Army Research Office under the contract number W911NF-16-1-0168, National Institutes of Health under the grant number R01LM011986, Region II University Transportation Center under the project number 49997-33 25, NSFC grant no. 61602306 and a Baidu gift.

## REFERENCES

[1] Rie K Ando and Tong Zhang. 2006. Learning on Graph with Laplacian Regularization. In *NIPS*. 25–32.
[2] Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2, 1 (2009), 183–202.
[3] Nan Cao, Yu-Ru Lin, Liangyue Li, and Hanghang Tong. 2015. g-Miner: Interactive Visual Group Mining on Multivariate Graphs. In *CHI*. 279–288.
[4] Chen Chen, Hanghang Tong, Lei Xie, Lei Ying, and Qing He. 2016. FASCINATE: Fast Cross-Layer Dependency Inference on Multi-layered Networks. In *KDD (KDD '16)*. 765–774.
[5] Aaron Clauset, Daniel B. Larremore, and Roberta Sinatra. 2017. Data-driven predictions in the science of science. *Science* 355, 6324 (2017), 477–480.
[6] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized Multi–task Learning. In *KDD*. 109–117.
[7] Barbara R. Jasny and Richard Stone. 2017. Prediction and its limits. *Science* 355, 6324 (2017), 468–469. DOI: http://dx.doi.org/10.1126/science.355.6324.468
[8] Theodoros Lappas, Kun Liu, and Evimaria Terzi. 2009. Finding a team of experts in social networks. In *KDD*. ACM, 467–476.
[9] Kenneth D. Lawrence and Jeffrey L. Arthur. 1990. *Robust regression: analysis and applications*. Marcel Dekker Inc, New York.
[10] Liangyue Li and Hanghang Tong. 2015. The Child is Father of the Man: Foresee the Success at the Early Stage. In *KDD*. 655–664.
[11] Liangyue Li, Hanghang Tong, Nan Cao, Kate Ehrlich, Yu-Ru Lin, and Norbou Buchler. 2015. Replacing the Irreplaceable: Fast Algorithms for Team Member Recommendation. In *WWW*. 636–646.
[12] Liangyue Li, Hanghang Tong, Nan Cao, Kate Ehrlich, Yu-Ru Lin, and Norbou Buchler. 2016. TEAMOPT: Interactive Team Optimization in Big Networks. In *CIKM*. 2485–2487.
[13] Liangyue Li, Hanghang Tong, Nan Cao, Kate Ehrlich, Yu-Ru Lin, and Norbou Buchler. 2016. Enhancing Team Composition in Professional Networks: Problem Definitions and Fast Solutions. *TKDE* (2016).
[14] Liangyue Li, Hanghang Tong, Jie Tang, and Wei Fan. 2016. *iPath:* Forecasting the Pathway to Impact. In *SDM*. 468–476.
[15] Liwei Liu and Erdong Zhao. 2011. Team Performance and Individual Performance: Example from Engineering Consultancy Company in China. In *2011 International Conference on Management and Service Science*. 1–4.
[16] Patrick Mair, Kurt Hornik, and Jan de Leeuw. 2009. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of statistical software* 32, 5 (2009), 1–24.
[17] David Schmeidler. 1989. Subjective probability and expected utility without additivity. *Econometrica: Journal of the Econometric Society* (1989), 571–587.
[18] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.
[19] Paul Tseng. 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications* 109, 3 (2001), 475–494.
[20] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science* 316, 5827 (2007), 1036–1039.
[21] Jianpeng Xu, Pang-Ning Tan, Jiayu Zhou, and Lifeng Luo. 2017. Online Multi-task Learning Framework for Ensemble Forecasting. *TKDE* (2017).
[22] Yuan Yao, Hanghang Tong, Tao Xie, Leman Akoglu, Feng Xu, and Jian Lu. 2014. Joint voting prediction for questions and answers in CQA. In *ASONAM*. IEEE, 340–343.
[23] Yuan Yao, Hanghang Tong, Feng Xu, and Jian Lu. 2014. Predicting long-term impact of CQA posts: a comprehensive viewpoint. In *KDD*. ACM, 1496–1505.
[24] Petek Yontay and Rong Pan. 2016. A computational Bayesian approach to dependency assessment in system reliability. *Reliability Engineering & System Safety* 152 (2016), 104–114.
[25] Xiangrong Zeng and Mário A. T. Figueiredo. 2014. The Ordered Weighted $l_1$ Norm: Atomic Formulation, Dual Norm, and Projections. *CoRR* abs/1409.4271 (2014). http://arxiv.org/abs/1409.4271