***Data Analysis Paper***

# GB-R: A Fast and Effective Gray-Box Reconstruction of Cascade Time-Series

*Hyun Ah Song* *
Machine Learning Deparment
Carnegie Mellon University
hyunahs@cs.cmu.edu

*DAP advisor*
Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

*DAP committee*
Vladimir Zadorozhny
University of Pittsburgh
vladimir@sis.pitt.edu

April 15, 2018

**Abstract**

Given some (but not all) monthly totals of people with measles (or counts of product-units sold, or counts of retweets), how can we recover the weekly counts? Requiring smoothness between successive weeks is reasonable - but can we do better, if we have some domain knowledge? For example, we know that measles (flu, count-of-retweets, etc) follow a specific cascade model, like the so-called 'SIS'.

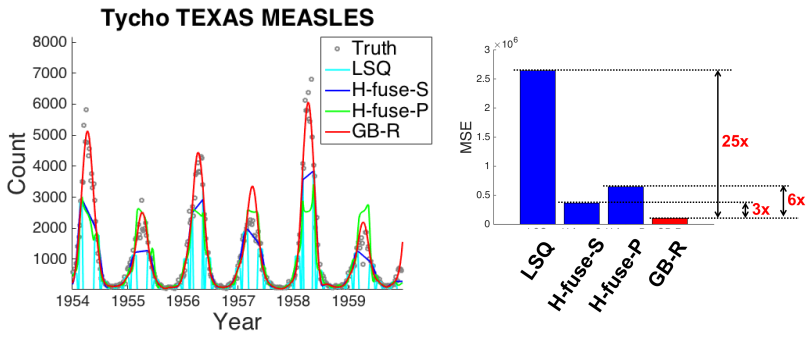
The answer is 'yes'. With our proposed GB-R we show how to inject domain knowledge, creating a *gray-box* model; we show how to set up and efficiently solve the appropriate optimization problem. The desirable properties of our GB-R are: (a) **Effectiveness**, outperforming the best competitors on real, epidemiology data, often by **3x - 25x** in reconstruction error; (b) **Scalability**, being linear on the sequence length and (c) **Interpretability**, accurately estimating the parameters of the gray-box model.

## 1 Introduction

Given reports from multiple data sources, in several granularities, (like *monthly* counts of sales from one source, and *weekly* counts of sales from another), with some of the reports missing, how can we reconstruct the *daily* sales counts? The problem is usually under-determined - how can we find a realistic solution, if experts advice us that the real sequence obeys a well-defined model, like the susceptible-infected-susceptible ('SIS') cascade model?

More formally, the problem is as follows: Let $\mathbf{x} = (x_1, \ldots, x_T)$ be the unknown time sequence (say, weekly counts of flu patients); let $\mathbf{y} = (y_1, \ldots, y_N)$ be the aggregated information (say, reports of monthly sums, with some of them missing); and suppose that domain experts advise that the unknown sequence $\mathbf{x}$ obeys a differential equation, like the SIS model (see equation (2)), or some other model; such models have a few parameters, like the patient-recovery rate $\delta$, the spreading-rate $\beta$, etc. Our goal is to reconstruct the time

*This is co-work with Fan Yang (University of Pittsburgh), Zongge Liu (Carnegie Mellon University), Wilbert van Panhuis (University of Pittsburgh), Nicholas Sidiropoulos (University of Minnesota), Christos Faloutsos (Carnegie Mellon), Vladimir Zadorozhny (University of Pittsburgh).




(a) (b)

Figure 1: GB-R outperforms competitors: (a) GB-R (*red*) is closer to truth (*gray dot*), in measles patient-counts over time (1954-1960); (b) it gives **3x** to **25x** better MSE in reconstruction.

sequence $\mathbf{x}$ (and the model parameters $\beta$ etc), so that $\mathbf{x}$ satisfies the aggregated information ($\mathbf{y}$), and also matches the gray-box model as best as it can.

**Informal Problem 1 (Gray-Box Reconstruction)**

- Given: *the reports* $\mathbf{y}$*; and a differential equation (gray box model, with some unknown parameters)*
- Reconstruct: *the true time series* $\mathbf{x}$*, and the gray-box-model parameters*

One especially challenging setting is when the peak values are wrong, or missing (*"flooding effect"*). For example, when there is an outburst of an epidemic in an developing regions (e.g. Ebola epidemic in West Africa[1]), clinical practitioners may prioritize clinical care over case reporting. Similarly, in computer traffic monitoring, during peak times, routers drop packets, resulting in incorrect count of packets.
We propose the "GB-R" method to solve the gray-box reconstruction problem. What sets GB-R apart from all other reconstruction methods is that it carefully infuses domain knowledge (like the SIS cascade model), with dual benefits: (a) *accuracy*: it better reconstructs the unknown time series $\mathbf{x}$; and (b) *interpretability*: it also estimates the gray-box model parameters (infection rate, etc), thus allowing a better understanding of the cascade.
Figure 1 compares GB-R against top contenders. The real data were the Project Tycho level1 measles weekly counts in the state of Texas, during 1954-1960 [10] (*gray dots* in Figure 1-(a)). We generated the 5-week sums, omitting the peak times ("flooding effect" scenario). From those *5-week sums*, we reconstructed the true data to *weekly* granularity, using GB-R in *red*, and the competitors, 'LSQ' ([12]), 'h-fuse-s' ([7]), 'h-fuse-p' ([7]) in *cyan*, *blue*, and *green*, respectively. The proposed GB-R method reconstructs the true data very closely, including the peaks, while other methods fail to reconstruct. Figure 1(b) shows GB-R achieves **3x- 25x** better reconstruction than the competitors.
The contribution of our work is as follows:

1. **Effectiveness**: GB-R outperforms the competitors, often by **3x-25x** in reconstruction error.
2. **Scalability**: GB-R scales linearly with respect to the length of the time sequence.
3. **Interpretability**: GB-R also estimates the gray-box-model parameters (spreading-rate, patient-recovery rate, timing of peak, etc), providing a better understanding of the time sequence.

**Reproducibility**: Our code for GB-R is open-sourced at: `http://www.cs.cmu.edu/~hyunahs/code/GB_R.zip`. All the data we used are publicly available at `https://www.tycho.pitt.edu/`
The outline of the paper is the typical one: Background and related work; method description; intuition behind our approach; experiments; and conclusions.

[1] `http://www.nejm.org/doi/full/10.1056/NEJMsr1600236?af=R&rss=currentIssue#t=article`

## 2 Background and Related Work

In this section, we provide a brief background on the basic epidemic model, and the information fusion problem. Then we introduce some of the previous works related to our work.

### 2.1 SIS epidemic model basics

Among the numerous epidemic models (SI, SIR, SIRS, etc [1]), we choose here the SIS model, because it is the simplest one and still gives good fits. The SIS model is suitable for, say, the flu, where we assume no immunity: a person is either susceptible ('S', i.e., healthy but possible to get infected), or infected ('I', i.e., sick). An infected person heals with probability $\delta$, and becomes susceptible again (no immunity); an infected person may infect a healthy person with probability $\beta$.
Let $x_t$ be the count of infected people at time $t$. Then, the SIS model uses the following differential equation:

$$\frac{dx_t}{dt} = \beta x_t(P - x_t) - \delta x_t \tag{1}$$

The equivalent difference equation is: $x_t = x_{t-1} + \beta x_{t-1}(P - x_{t-1}) - \delta x_{t-1}$ where $\beta$, $\delta$ are the infection rate and recovery rate of the disease, and $P$ is the population size.
In [8], the authors modified the SIS model to allow for periodicity (flu propagates faster during the colder months, etc). and they modeled the infection rate $\beta$ as periodic function of time, $\beta_t$, as follows:

$$x_t = x_{t-1} + \beta_{t-1} x_{t-1}(P - x_{t-1}) - \delta x_{t-1} \tag{2}$$

where $\beta_t$ is defined as

$$\beta_t = \beta_0 \Big(1 + P_a \cos\big(\frac{2\pi}{P_p}(t + P_s)\big)\Big) \tag{3}$$

and $\beta_0$, $P_a$, $P_s$, $P_p$ are the average infection rate of the disease, the amplitude, the phase shift, and the period of the infection rate, respectively.

### 2.2 Information fusion basics

As mentioned, the problem is to recover the *true data* $\mathbf{x}$=$(x_1, \ldots, x_T)$ (e.g., daily counts of infected people) given some aggregated counts or reports $\mathbf{y}$=$(y_1, \ldots, y_N)$, say, monthly sums. Next, we show that this becomes an under-specified, linear problem.
Figure 2 illustrates the concepts that we need, which are defined below:

- *true data* $\mathbf{x}$: the finer-scale data that we want to reconstruct, in length of $T$.
- *a report* ($y$): a sum of true data over a longer period (in Figure 2, report 1 ($y_{i=1}$) is a *sum of 30 weeks*) The *report coverage* of report $i$ is the time interval that it spans ($t^i_{start}, t^i_{end}$), and the *report duration* is obviously the time length ($t^i_{end} - t^i_{start} + 1$).

The given $N$ reports $\mathbf{y} = (y_1, \ldots, y_N)$ are the hard constraints that we need to satisfy.
The information fusion task can be formulated as a system of linear equations, the *characteristic linear system* [7]. We need the concept of mixing matrix $\mathbf{A}$: this is an $N \times T$ matrix, where $N$ and $T$ are the total number of reports and total number of timeticks in the true data, respectively; with all entries as zero, except $A_{it} = 1$ when the $i$-th report includes the $t$-th timetick (i.e., $t^i_{start} \leq t \leq t^i_{end}$, for each $i^{th}$ report as shown in Figure 3).
With the above definition of the mixing matrix $\mathbf{A}$, we have

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{4}$$

which is under-determined, since $N < T$.

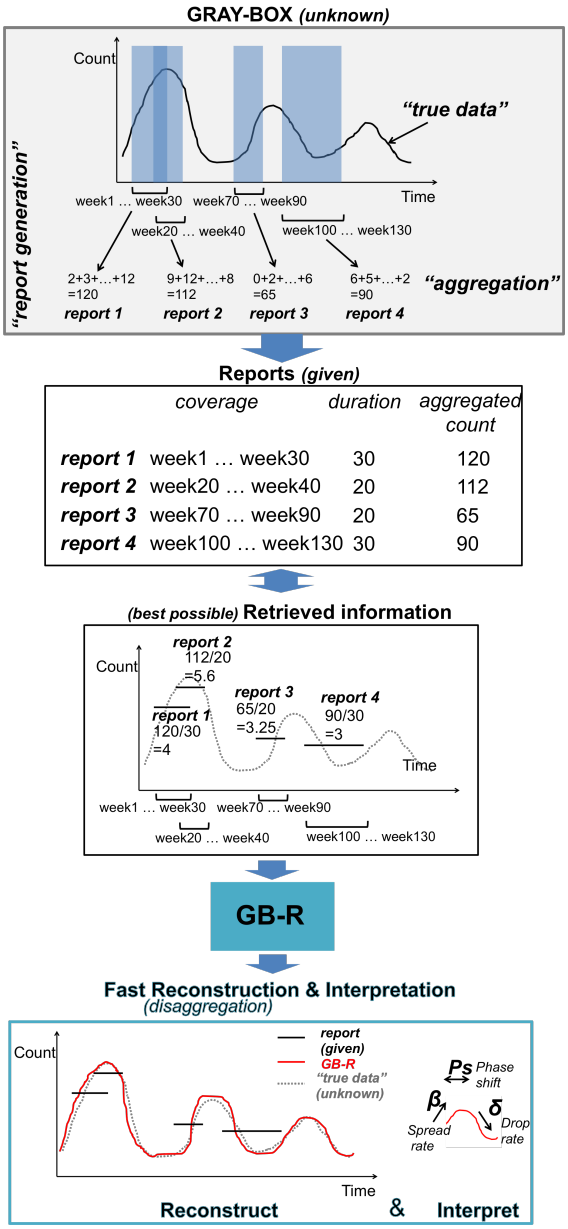



Figure 2: Information fusion basics.

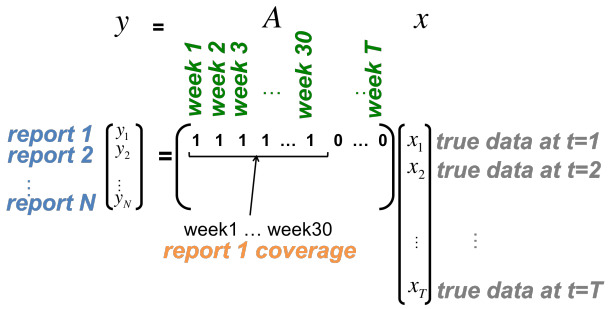



Figure 3: Information fusion formulation

### 2.3 Related works

Previous works fall in the following groups: information fusion, super resolution reconstruction, and epidemics modeling.

**Infofusion:** One of the most well-known information fusion projects in the historical data domain is the Tycho project [10, 11, 12]. Project Tycho is an open-access repository for disease data for which over 6,500 weekly disease reports were digitized and curated, spanning 125+ years and 100+ diseases. Data entry for Project Tycho took 200 million keystrokes, making it into a valuable real-world dataset and a testbed for the problem of information fusion.

A recent work appeared in [7], where the main idea is to assume that the unknown sequence $\mathbf{x}$ is smooth, and/or periodic. The resulting methods, h-fuse-s and h-fuse-p, give good results on Tycho data, and we use them as baselines here, together with the least squares method "LSQ". (throughout the paper, we use the term LSQ to refer to the minimum norm solution to the under-determined system, following the convention of [7] - while in signal processing, it usually means the least squares solution to the over-determined problem.)

**Super resolution reconstruction:** Information fusion and super resolution reconstruction share a similar goal of reconstructing higher resolution data from low. The most basic algorithm for super resolution reconstruction is the least squares (LSQ) method along with Tikhonov regularization [5, 4] that enforce additional constraints such as smoothness. Super resolution reconstruction has been studied in various fields like signal processing ("F.J.S. method" [2]), or brain imaging (BrainZoom [3]).

**Epidemics modeling:** With respect to the epidemic data modeling, cascade and epidemic analysis have attracted huge interest [6], [9], also in biological/computer virus propagation, word-of-mouth marketing campaigns, etc.

In contrast to GB-R, none of the algorithms mentioned allows for information fusion using the domain knowledge, lacking the interpretability of the result (e.g. phase shift of the sinusoidal pattern, healing rate, etc).

In Table 1, a we compare GB-R and other competitors on the basis of selected properties.

## 3 Proposed Method: GB-R

In this section, we provide the details of GB-R.

A table of symbols that is used throughout this paper is shown in Table 2.

Our goal is to reconstruct the unknown true time series data $\mathbf{x}$ given the multiple reports $\mathbf{y}$ of aggregated sums and the mixing matrix $\mathbf{A}$ that tells us the coverage of each report. Because this is an underdetermined optimization problem ($N \ll T$), we need to incorporate additional constraints and/or regularization, based on and reflecting our domain knowledge about the nature of the data.

Table 1: GB-R captures all of the listed properties.

| *Properties* | LSQ | F.J.S. method [2] | BrainZoom [3] | h-fuse [7] | GB-R |
|---|---|---|---|---|---|
| Scalability | ✓ | ✓ | ✓ | ✓ | ✔ |
| Smooth reconstruction | | ✓ | ✓ | ✓ | ✔ |
| Periodic reconstruction | | | | ✓ | ✔ |
| Nonlinear modeling | | | ✓ | | ✔ |
| Interpretability | | | | | ✔ |
| Infusion of domain knowledge | | | | | ✔ |

In our work of GB-R, we utilize one of the prominent epidemic model called "*SIS model*" (section 2.1) to fit the unknown time series $\mathbf{x}$. This expert-model explains how individuals in the population become infected, and how they recover and become susceptible to the disease again. The model uses parameters such as infection rate, peak phase, peak amplitude, etc that help interpret the evolving pattern of the time series data in terms of the model parameters, which capture how rapidly infection spreads, when it reaches its peak values, etc. Other models, like SIR (allowing immunity), SIRS (temporary immunity) etc, may reflect the nature of the disease even better, but the expert-model we propose to use here, gives very good accuracy and it is simpler to fit. Thus we decided to use the SIS model as our expert-model.
Details of our proposed algorithm GB-R are provided in the following section.

Table 2: Symbols and definitions

| Symbols | Definitions |
|---|---|
| $x_t$ | true (unknown) time series data in fine scale that we want to reconstruct. $t = 1, \cdots, T$ |
| $y_i$ | a report count (an aggregated sum of $\mathbf{x}$). $i = 1, \cdots, N$ |
| $\mathbf{A}$ | mixing matrix used to generate report from $\mathbf{x}$: $\mathbf{y} = \mathbf{A}\mathbf{x}$. $\mathbf{A} \in \mathbb{R}^{N \times T}$ |
| $T$ | total timeticks of $\mathbf{x}$ |
| $N$ | total number of reports $\mathbf{y}$ |
| $\beta_0$ | average infection rate in SIS model |
| $\delta$ | recovery rate of the disease in SIS model |
| $P_a$ | amplitude of the SIS model |
| $P_s$ | phase shift of the SIS model |
| $P_p$ | period of the SIS model |
| $\mathbf{\Theta}$ | a set of parameters to optimize. $\{\mathbf{x}, \mathbf{z}, P_a, P_s, \beta_0, \delta\}$ |
| $\mathcal{M}$ | a set of SIS model parameters. $\{P_a, P_s, \beta_0, \delta\}$ |
| $P$ | the size of population |

### 3.1 Proposed problem formulation

What is the simplest way to *reconstruct the true data from the aggregated reports*, i.e. to solve for infofusion problem? The most naive approach is to solve the following problem:

$$\min_{\mathbf{x}} \sum_{i=1}^{N} (y_i - \sum_{t=1}^{T} A_{it} x_t)^2 \tag{5}$$

where $\mathbf{x}$ is the true data that we want to reconstruct, $\mathbf{y}$ is a vector of $N$ given reports, and $\mathbf{A}$ is the mixing matrix that generates the reports. However, as we mentioned in the previous section, this is underdetermined problem ($N < T$), so we need to have proper regularizers that can reflect the nature of the true data.

Then how can we reconstruct the true data from the aggregated reports, in a way that it *obeys the domain knowledge (expert-model)*? The first attempt (the most naive approach) to include the expert-model is shown below:

$$\min_{\mathbf{\Theta}} \Big[ \sum_{i=1}^{N} (y_i - \sum_{t=1}^{T} A_{it} x_t)^2 + \mu \sum_{t=2}^{T} (x_t - x_{t-1} - \beta_{t-1} x_{t-1} (P - x_{t-1}) + \delta x_{t-1})^2 \Big] \tag{6}$$

where $\mathbf{\Theta}$ is a set of parameters we optimize $\mathbf{\Theta} = \{\mathbf{x}, \mathbf{z}, P_a, P_s, \beta_0, \delta\}$ and $\beta_t$ is defined as equation (3). The first term is for the reconstruction of the true data from the aggregated reports, and the second term is to fit the true data using our expert-model to incorporate our domain knowledge.

Then our next question is: *how can we optimize* the nonlinear equation given in equation (6)? In the second term (expert-model), we see that the key challenge is in the quadratic form of $x_{t-1}$, which makes the second term a fourth-order polynomial, which is very hard to optimize.

In GB-R, we reformulate our problem to turn it to a bilinear one, which can be handled using alternating optimization. This is the key step needed to derive a computationally tractable and effective algorithm. We use variable splitting method, where we introduce an auxiliary variable $\mathbf{z}$ that is a copy of $\mathbf{x}$, and replaced one of $\mathbf{x}$ with its copy $\mathbf{z}$ so that our problem becomes bilinear. Then we enforce that the copy $\mathbf{z}$ is same as $\mathbf{x}$. This reformulation of the problem reduces to the optimization problem below:

$$\begin{aligned} \min_{\mathbf{\Theta}} \Big[ & \sum_{i=1}^{N} (y_i - \sum_{t=1}^{T} A_{it} x_t)^2 \\ & + \mu \sum_{t=2}^{T} (x_t - x_{t-1} - \beta_{t-1} x_{t-1} (P - z_{t-1}) + \delta x_{t-1})^2 \\ & + \xi \sum_{t=1}^{T} (x_t - z_t)^2 \Big] \end{aligned} \tag{7}$$

From the experimental results, we can simply set $\mu$ and $\xi$ to 1.

Our optimization problem of equation (7) consists of three terms: T1: hard constraint (reconstruction), T2: soft constraint (expert-model), and T3: variable splitting. We will explain each term in more detail in the following.

#### 3.1.1 T1: hard constraint (reconstruction)

$$\sum_{i=1}^{N} (y_i - \sum_{t=1}^{T} A_{it} x_t)^2 \tag{8}$$

The first term in our optimization problem is shown above. It is to ensure that when we aggregate our reconstruction $\hat{\mathbf{x}}$ based on the mixing matrix $\mathbf{A}$, we are able to generate as good matches to the observed reports $\mathbf{y}$ as possible. This is a hard constraint that must be met.

#### 3.1.2 T2: soft constraint (expert-model)

$$\sum_{t=2}^{T}(x_t - x_{t-1} - \beta_{t-1}x_{t-1}(P - z_{t-1}) + \delta x_{t-1})^2 \tag{9}$$

The second term of the optimization problem shown above involves the domain knowledge modeling of $\mathbf{x}$. T2 nudges $\hat{\mathbf{x}}$ towards the solution that best obeys the domain knowledge. By this term, we enforce 1) smoothly but possibly peaky patterns, and 2) periodic patterns in $\hat{\mathbf{x}}$. T2 enables us to recover the underlying evolving characteristics of given data by learning basic parameters of $\beta_0$, $\delta$, $P_a$, and $P_s$.
As discussed briefly earlier in the section, in our reformulation, we introduce an auxiliary variable $\mathbf{z}$ that is a copy of $\mathbf{x}$ and replace one of $\mathbf{x}$ with its copy $\mathbf{z}$ (*variable splitting method*). This reformulated optimization problem becomes bilinear and becomes easier to solve using alternating optimization method. This is to replace the quadratic nonlinear term in the cost function by one that is conditionally linear in one of the two variables given the other. This greatly facilitates optimization, as we will see.

#### 3.1.3 T3: Variable splitting

$$\sum_{t=1}^{T}(x_t - z_t)^2 \tag{10}$$

With our formulation in GB-R, in T3 we enforce that $\mathbf{z}$, a copy of $\mathbf{x}$, matches $\mathbf{x}$.

### 3.2 Proposed optimization steps

In this section, we explain how we solve our optimization problem shown in equation (7). Given that our optimization problem is reformulated into bilinear form, we propose to use an alternating optimization method. The iterative update rule of GB-R algorithm is shown in Algorithm 1. We alternatively solve for each of the variables in the parameter set $\mathbf{\Theta} = \{\mathbf{x}, \mathbf{z}, P_a, P_s, \beta_0, \delta\}$ one at a time, while fixing the rest. The update equations are derived by setting the corresponding partial derivative of equation (7) to zero.
Notice that thanks to the variable splitting method (i.e. the introduction of an auxiliary variable $\mathbf{z}$), each alternating optimization step of our problem can be solved as a least-squares problem, using the Moore-Penrose pseudoinverse. We will explain the update equations in the following section.
**How to optimize for $\mathbf{x}$**: The solution is explained in Lemma 1 below.

**Lemma 1** $\mathbf{x}$ *can be solved via Moore-Penrose pseudoinverse:*

$$\mathbf{x} = \begin{bmatrix} \mathbf{A} \\ - \\ \mathbf{M}_1 \\ - \\ \mathbf{I} \end{bmatrix}^{\dagger} \begin{bmatrix} \mathbf{y} \\ - \\ \mathbf{0} \\ - \\ \mathbf{z} \end{bmatrix} \tag{11}$$

*where* $\dagger$ *means Moore-Penrose pseudoinverse and* $\mathbf{M}_1$ *is a 2-band diagonal matrix of size* $(T-1 \times T)$*:*

$$\begin{aligned} \mathbf{M}_1(t,t) &= -1 - \beta_t(P - z_t) + \delta \\ \mathbf{M}_1(t,t+1) &= 1 \end{aligned} \tag{12}$$

*for* $\forall t = 1 \cdots, T-1$. $\mathbf{I}$ *is an identity matrix of size* $(T \times T)$.

**Proof 1** *To update* $\mathbf{x}$*, consider only the terms involving* $\mathbf{x}$*. Then we can reformulate the optimization problem as follows.*

$$\min_{\mathbf{x}} \left\| \begin{bmatrix} \mathbf{y} \\ - \\ \mathbf{0} \\ - \\ \mathbf{z} \end{bmatrix} - \begin{bmatrix} \mathbf{A} \\ - \\ \mathbf{M}_1 \\ - \\ \mathbf{I} \end{bmatrix} \mathbf{x} \right\|_2^2 . \tag{13}$$

*Solving for above optimization problem enables us to solve* $\mathbf{x}$ *via Moore-Penrose pseudoinverse as in equation (11).* ■

**How to optimize for z**: $\mathbf{z}$ can be solved in the similar manner. The solution is shown in Lemma 2.

**Lemma 2** $\mathbf{z}$ *can be solved via Moore-Penrose pseudoinverse:*

$$\mathbf{z} = \begin{bmatrix} \mathbf{I} \\ - \\ \mathbf{M}_2 \end{bmatrix}^{\dagger} \begin{bmatrix} \mathbf{x} \\ - \\ \mathbf{v}_1 \end{bmatrix} \tag{14}$$

*Here,* $\mathbf{v}_1$ *is a vector of size* $(T-1)$*:*

$$\mathbf{v}_1(t) = -x_{t+1} + x_t + \beta_t P x_t - \delta x_t, \tag{15}$$

$\forall t = 1, \cdots, T-1$. $\mathbf{M}_2$ *is a diagonal matrix of size* $(T-1 \times T)$*:*

$$\mathbf{M}_2(t,t) = \beta_t x_t, \tag{16}$$

$\forall t = 1, \cdots, T-1$, *and* $\mathbf{I}$ *is an identity matrix of size* $(T \times T)$*.*

**Proof 2** *Considering only the terms involving* $\mathbf{z}$*, we can reformulate optimization problem regarding* $\mathbf{z}$ *as follows.*

$$\min_{\mathbf{z}} \left\| \begin{bmatrix} \mathbf{x} \\ - \\ \mathbf{v}_1 \end{bmatrix} - \begin{bmatrix} \mathbf{I} \\ - \\ \mathbf{M}_2 \end{bmatrix} \mathbf{z} \right\|_2^2 . \tag{17}$$

*Solving for above optimization problem enables us to solve* $\mathbf{z}$ *via Moore-Penrose pseudoinverse as in equation (14).* ■

**How to optimize for expert-model parameters** $\mathcal{M}$: The rest of the variables $\mathcal{M} = \{P_a, P_s, \beta_0, \delta\}$ can be updated in a similar manner:
Update equation for $P_a$:

$$P_a = \sum_{t=2}^{T} \frac{x_t - x_{t-1} - \beta_0 x_{t-1}(P - z_{t-1}) + \delta x_{t-1}}{\beta_0 \cos(\frac{2\pi}{P_p}(t - 1 + P_s)) x_{t-1}(P - z_{t-1})} \tag{18}$$

Update equation for $P_s$: For better interpretability of $P_s$, we seek $P_s$ to be an integer (week). Thus we perform exhaustive search of 52 possible positions (weeks throughout a year) and pick the one with the smallest cost.

$$P_s = \arg \min_{P_s' = 1,2,\cdots,52} \sum_{t=2}^{T} x_t - x_{t-1} - \beta_0(1 + P_a cos(\frac{2\pi}{P_p}(t - 1 + P_s'))) x_{t-1}(P - z_{t-1}) + \delta x_{t-1} \tag{19}$$

Update equation for $\beta_0$:

$$\beta_0 = \sum_{t=2}^{T} \frac{x_t - x_{t-1} - \beta_{t-1} x_{t-1}(P - z_{t-1})}{(1 + P_a cos(\frac{2\pi}{P_p}(t - 1 + P_s)))} \tag{20}$$

Update equation for $\delta$:

$$\delta = \sum_{t=2}^{T} \frac{x_t - x_{t-1} - \beta_{t-1} x_{t-1}(P - z_{t-1})}{-x_{t-1}} \tag{21}$$

**Algorithm 1:** GB-R.

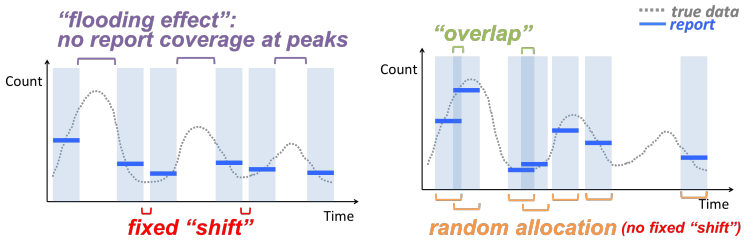
**Data**: mixing matrix $\mathbf{A}$, reports $\mathbf{y}$, the population $P$, the periodicity $P_p$
**Result**: A set of parameters $\mathbf{\Theta}$: unknown true data $\{\mathbf{x}\}$ of finer scale, and expert-model parameters $\{P_a, P_s, \beta_0, \delta\}$
**for** *iter* $= 1, \cdots,$ *until convergence* **do**
  (block coordinate descent);
  Update each variable from $\mathbf{\Theta} = \{\mathbf{x}, \mathbf{z}, P_a, P_s, \beta_0, \delta\}$, while fixing the rest;

# 4 Experiments

In this section, we demonstrate experimental results of GB-R to address three questions: *Q1*: how effective is the method; *Q2*: how scalable; and *Q3*: how easy it is to interpret its answer.
**Data description** Tycho data [10, 11, 12] is an epidemic database that is integrated from more than $6,500$ weekly reports of the United States epidemic data spanning over 100 years from multiple sources in different granularities. There are 3 types of dataset: Level 1, 2, and 3 based on the level of standardization of the data. Level 1 dataset is in the most tailored and standardized, thus the most complete form, but with limitations in the types of diseases and the location of the data. There are 7 state-level diseases (hepatitis A, measles, mumps, pertussis, polio, rubella, and smallpox) and 1 city-level disease (diphtheria) for 50 states and 122 cities, respectively. The total time coverage of the counts varies for different types of diseases, some of which starts as early as 1916 and ends as recent as 2009. The counts are weekly-basis, and there exist missing values.
We applied GB-R on Tycho Level 1 dataset, and 7 state-level diseases.
**Experimental setup** We incorporate our knowledge on the yearly periodicity of the epidemic data, and set $P_p = 52$ (since the Tycho data is weekly basis, and there are 52 weeks in one year). We fixed the total population $P = 10$ million, following the population of the New York State. (the result is not sensitive to the population as long as it is set to sufficiently large value, so we fixed $P$ for all of the states)
**Scenarios** In Figure 4, different types of report generation processes are illustrated: 1) "flooding effect" scenario, and 2) random reporting.

### 4.0.1 "flooding effect" scenario

As mentioned in section 1, "flooding effect" refers to the case where the reported counts during the peak season may not be accurate due to prioritization of clinical care.
Illustration on the "flooding effect" scenario is shown in Figure 4 (a). The report duration and the intervals between each report are fixed, except that during the peak seasons, there is no report covering the time period.




(a) "flooding effect" (b) random reporting

Figure 4: Report generation scenarios.

#### 4.0.2 Random reporting

Random reporting is illustrated in Figure 4 (b). In this setting, report duration is still fixed, but the allocation of the report coverage is randomly assigned. This results in overlaps between reports, and long space between reports (long time range with no report coverage). This is a plausible scenario when there are multiple data sources, each of them having different "shifts".

### 4.1 Q1: Effectiveness

In this section, we demonstrate and prove the effectiveness of GB-R on the real-world data.

#### 4.1.1 Case of "flooding effect"

The effectiveness of GB-R on the real-world Tycho data under "flooding effect" scenario was demonstrated earlier in Figure 1: We chose Texas state because the time series pattern in Texas state was more spiky compared to that of New York state, thus more suitable to demonstrate the "flooding effect" scenario. We fixed the report duration to be 5-week, and the "shift" to be 2-week. We left out the peak seasons with no report.
In Figure 1 (a) and (b), we have demonstrated that GB-R successfully reconstructs the *spiky peak* patterns by utilizing its domain expert knowledge ensuring the data follows the natural cascade pattern, resulting in reduction of the reconstruction error compared to the competitors.

#### 4.1.2 Case of random reporting

Random reporting scenario is the most likely scenario in the real world, so we focus more on the experiments under this setting.
**Generalization over diseases and states** In Figure 5, bar plots of MSE of the reconstructions by GB-R and the competitor algorithms on different types of diseases (measles, pertussis, hepatitis A, polio, rubella, mumps, and smallpox, all 7 state-level diseases available in level 1 Tycho dataset) in three largest states (California (CA), New York (NY), and Texas (TX)) are shown.
We used first $T = 1,000$ time points of the weekly counts as our unknown true dataset and generated reports from it using random reporting process as illustrated in Figure 4 (b). Each disease starts in different time of the year. We fixed the report duration to be 20-week and the total number of reports as $N = 100$.
GB-R wins over competitors in reconstructing all of the diseases except for Hepatitis A, and Smallpox. Even for the losing cases, the amount of loss is small.
**Intuition behind GB-R** In this section, we share our intuition on why GB-R outperforms the competitors. In Figure 6, a reconstruction of weekly measles count of New York state by h-fuse-s, and h-fuse-p is shown in top and bottom, respectively.

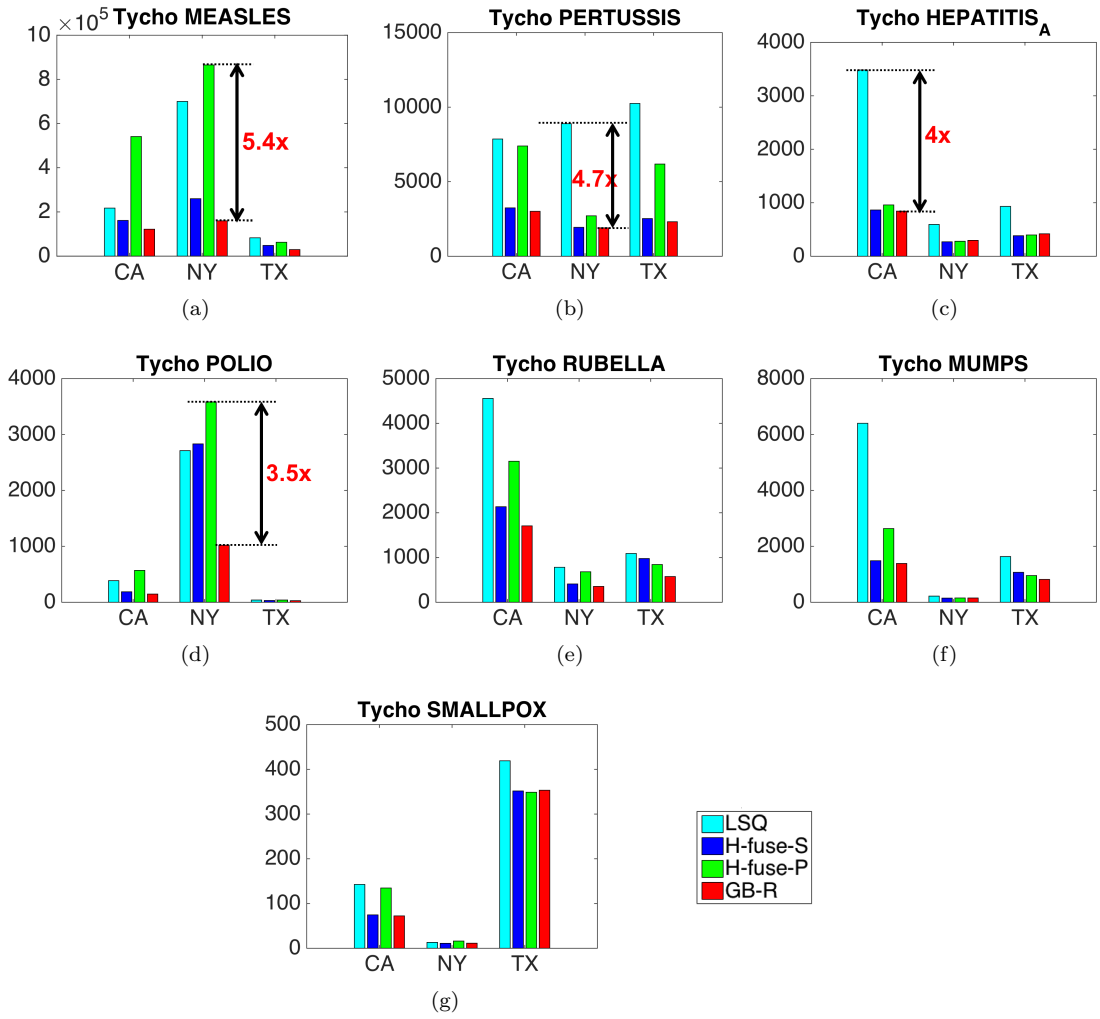



Figure 5: **GB-R almost always wins**, often by large margin. Bar plots of MSE of reconstruction of (a) measles, (b) pertussis, (c) hepatitis A, (d) polio, (e) rubella, (f) mumps, and (g) smallpox in three largest states California (CA), New York (NY), and Texas (TX) are shown.

In the top plot, h-fuse-s fails to reconstruct the first and the eighth peaks (no report coverage). For the time region where there is too scarce reports as reference to make a guess on the true data, h-fuse-s interpolates, or *"smooths out"* the counts of the adjacent time points with report coverage.
In the bottom plot, h-fuse-p *overestimates* the peak values for years 1940-1946 due to the strong assumption of periodic pattern that expects repetitive heights for the recurring peaks.
On the other hand, in both plots, GB-R demonstrates good reconstructions; no smoothing out or overestimation problem as h-fuse-s or h-fuse-p, thanks to the expert-model, that forces the reconstruction to follow more natural evolving patterns with spiky peaks, spread rate, and recovery rate.

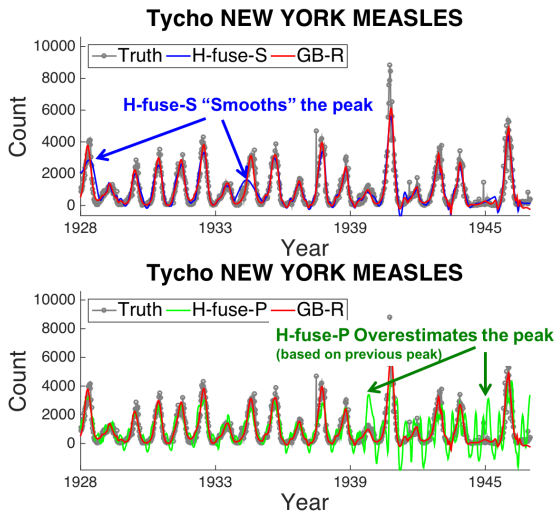



Figure 6: **Superiority of GB-R**. Competitors enforcing smoothness or periodicity may miss / over represent the spikes

## 4.2 Q2: Scalability

In Figure 7, the wall clock time (s) required for GB-R to run on different length of timeticks ($T$) is plotted along with a linear line for better understanding. It shows that GB-R scales linearly with respect to the length of the timeticks $T$ in the given data.

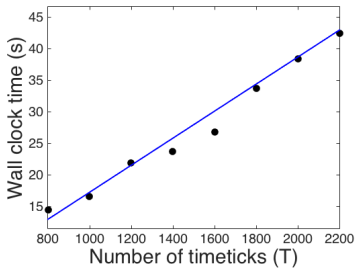



Figure 7: **GB-R scales linearly** with respect to the signal length.

Table 3: expert-model parameters

| | Measles | Polio | Pertussis |
|---|---|---|---|
| $P_s$ | 52 weeks (peak: Jan) | 27 weeks (peak: Jul) | 47 weeks (peak: Nov) |
| $\delta$ | 0.31 | 0.31 | 0.92 |

### 4.3 Q3: Interpretability

Some of the expert-model parameters values $(P_s, \delta)$ learned by GB-R on the weekly counts of measles, polio, and pertussis in California state are shown in Table 3.

**phase shift** $P_s$ represents the *peak seasons* of the data. In Table 3, GB-R learned different peak seasons for each disease. Our interpretation of the peak seasons of measles, polio, and pertussis matches with the epidemiology:

1) *"Measles increases during the late winter and early spring"* [2]
2) *"Polio typically peaks in the summer months"* [3]
3) *"Peak season for Pertussis is late summer and early fall."* [4]

**recovery rate** $\delta$ corresponds to how sharp is the drop in the spike. Figure 8 (top) shows the synthetically generated data with increasing values of $\delta$: larger the $\delta$, sharper the drop.

Figure 8 bottom left shows real Tycho data (*gray*) and the fitting by GB-R (*red*). Figure 8 bottom right shows the fittings of measles, polio, and pertussis on top of each other for the comparison of the sharpness of the drop. GB-R successfully learned small $\delta$ for smooth drop, and large $\delta$ for sharp drop that matches with the actual pattern of the disease.

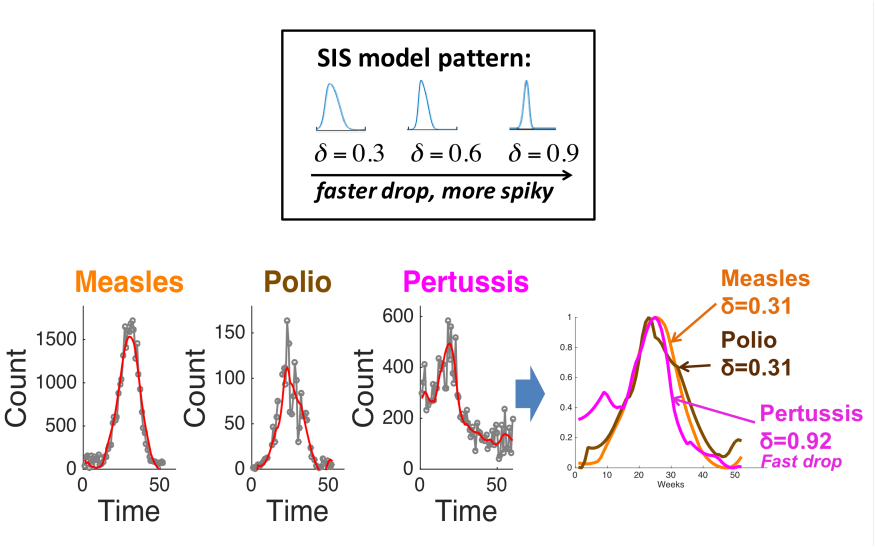



Figure 8: **GB-R is interpretable**. GB-R learns the healing rate parameter. Pertussis with larger parameter value heals faster.

[2] http://www.who.int/ith/diseases/measles/en/

[3] https://www.cdc.gov/vaccines/pubs/pinkbook/polio.html

[4] https://consumer.healthday.com/infectious-disease-information-21/misc-infections-news-411/spread-of-whooping-cough-raises-concern-641758.html

# 5 Conclusions

In this paper, we proposed GB-R, an algorithm that reconstructs time series counts from aggregated reports by careful infusion of domain knowledge.

1. **Effectiveness**: GB-R outperforms top competitors by **3x-25x** in reconstructing real-world data thanks to its principled infusion of domain knowledge.
2. **Scalability**: GB-R scales linearly on the duration of the output sequence.
3. **Interpretability**: GB-R provides understanding of the time sequence, by estimating the gray-box-model parameters, like the average infection rate ($\beta_0$), the peak timing ($P_s$), the peak size ($P_a$), etc.

**Reproducibility**: We provide our source code for at: `http://www.cs.cmu.edu/~hyunahs/code/GB_R.zip`. All our data are publicly available at `https://www.tycho.pitt.edu/`

# References


[1] Roy M. Anderson and Robert M. May. *Infectious diseases of humans: Dynamics and control*. Oxford Press, 2002.

[2] Christos Faloutsos, H. V. Jagadish, and Nikolaos Sidiropoulos. Recovering information from summary data. In *VLDB'97, August 25-29, 1997, Athens, Greece*, pages 36–45, 1997.

[3] Xiao Fu, Kejun Huang, Otilia Stretcu, Hyun Ah Song, Evangelos Papalexakis, Partha Tulakdar, Tom Mitchell, Nicholas Sidiropoulos, and Christos Faloutsos. Brainzoom: High resolution reconstruction from multi-modal brain signals. In *SIAM Data Mining*, 2017.

[4] Silvia Gazzola and James G. Nagy. Generalized arnoldi–tikhonov method for sparse reconstruction. *SIAM Journal on Scientific Computing*, 36(2):B225–B247, 2014.

[5] Gene H. Golub, Per Christian Hansen, and Dianne P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.

[6] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 137–146, 2003.

[7] Zongge Liu, Hyun Ah Song, Vladimir Zadorozhny, Christos Faloutsos, and Nicholas Sidiropoulos. H-fuse: Efficient fuse of aggregated historical data. In *SIAM Data Mining*, 2017.

[8] Yasuko Matsubara, Yasushi Sakurai, Willem G Van Panhuis, and Christos Faloutsos. Funnel: automatic mining of spatially coevolving epidemics. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–114. ACM, 2014.

[9] B Aditya Prakash, Deepayan Chakrabarti, Nicholas C Valler, Michalis Faloutsos, and Christos Faloutsos. Threshold conditions for arbitrary cascade models on arbitrary networks. *Knowledge and information systems*, 33(3):549–575, 2012.

[10] Tycho Project:. https://www.tycho.pitt.edu.

[11] W. van Panhuis, J. Grefenstette, S. Jung, N. Chok, A. Cross, H. Eng, B. Lee, V. Zadorozhny, S. Brown, D. Cummings, and D. Burke. Contagious diseases in the united states from 1888 to the present. *The New England Journal of Medicine*, 369(22), 2013.

[12] V. Zadorozhny and Y.-F. Hsu. Conflict-aware fusion of historical data. In *Proc. of the 5th International Conference on Scalable Uncertainty Management (SUM11)*, 2013.