FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

Yasuko Matsubara[†], Yasushi Sakurai[†], Willem G. van Panhuis[§], Christos Faloutsos[‡]
[†] Dept. of Computer Science and Electrical Engineering, Kumamoto University,
[§] Dept. of Epidemiology, University of Pittsburgh, [‡] Dept. of Computer Science, Carnegie Mellon University
{yasuko,yasushi}@cs.kumamoto-u.ac.jp, wav10@pitt.edu, christos@cs.cmu.edu

ABSTRACT

Given a large collection of epidemiological data consisting of the count of d contagious diseases for l locations of duration n, how can we find patterns, rules and outliers? For example, the Project Tycho provides open access to the count infections for U.S. states from 1888 to 2013, for 56 contagious diseases (e.g., measles, influenza), which include missing values, possible recording errors, sudden spikes (or dives) of infections, etc. So how can we find a combined model, for all these diseases, locations, and time-ticks?

In this paper, we present FUNNEL, a unifying analytical model for large scale epidemiological data, as well as a novel fitting algorithm, FUNNELFIT, which solves the above problem. Our method has the following properties: (a) *Sense-making*: it detects important patterns of epidemics, such as periodicities, the appearance of vaccines, external shock events, and more; (b) *Parameter-free*: our modeling framework frees the user from providing parameter values; (c) *Scalable*: FUNNELFIT is carefully designed to be linear on the input size; (d) *General*: our model is general and practical, which can be applied to various types of epidemics, including computer-virus propagation, as well as human diseases.

Extensive experiments on real data demonstrate that FUNNELFIT does indeed discover important properties of epidemics: (P1) disease seasonality, e.g., influenza spikes in January, Lyme disease spikes in July and the absence of yearly periodicity for gonorrhea; (P2) disease reduction effect, e.g., the appearance of vaccines; (P3) local/state-level sensitivity, e.g., many measles cases in NY; (P4) external shock events, e.g., historical flu pandemics; (P5) detect incongruous values, i.e., data reporting errors.

Categories and Subject Descriptors: H.2.8 [**Database management**]: Database applications–*Data mining*

Keywords: Epidemics; Time-series; Automatic mining

1. INTRODUCTION

Given a huge collection of co-evolving epidemic time-series, such as measles and influenza, how can we find typical patterns or anomalies, and statistically summarize all the epidemic sequences? In this paper, we present a unifying model, namely FUNNEL, which provides a good description of large collections of epidemiological data. [1] Intuitively, the problem we wish to solve is as follows:

INFORMAL PROBLEM 1. *Given a large collection of epidemiological data, which consists of d diseases in l locations of duration n, with missing values and recording errors, we want to*

- *find basic patterns of diseases (e.g., seasonality)*
- *find extra patterns (e.g., outbreaks, sudden drops)*
- *detect anomalies (i.e., possible errors)*

Uncovering the mechanisms and patterns of contagious diseases is an important and challenging task for public health scientists and policy makers. In this paper, we study a publicly available resource of epidemiological data: *Tycho* [32], which contains the count of infections of 56 diseases in the U.S, covering over 125 years on a weekly basis. [2]

Preview of our results. Figure 1 (a) shows the number of measles cases in the United States from 1928 to 1982, as gray circles, and our fitted model, as a solid red line. The sequence has a clear yearly periodicity, but also characteristic bi- and triennial patterns resulting in alternating large (1941, 1958) and small (1940, 1947) epidemic years [22], which is known as "skip" phenomena [29]. It should also be noted that the number of cases suddenly dropped in 1965. This was achieved because of the vaccination program that started in 1963. Figure 1 (b) shows the potential population of susceptibles for measles for each state in the U.S. The top three states in this respect are, NY, PA, and CA.

Figure 1 (c) shows a scatter plot of the seasonality strength vs. the peak season for each disease. We determined four categories of diseases in terms of (1) the strength of annual periodicity (radius) and (2) their phase difference, i.e., the month in which they peak (angle). For example, we found previously characterized epidemic peaks for influenza in January-February and for respiratory childhood diseases (e.g., measles) in the spring [27], and for tick-borne diseases (e.g., Lyme disease) peaks in the summer [28]; There is no periodicity for sexually transmitted diseases, (e.g., gonorrhea).

We can capture these important patterns in epidemic data with our proposed model, simply by changing its parameters. More importantly, our method is *fully-automatic*, that is, it provides a good description of a large collection of epidemiological data, without user intervention, prior training, or parameter tuning.

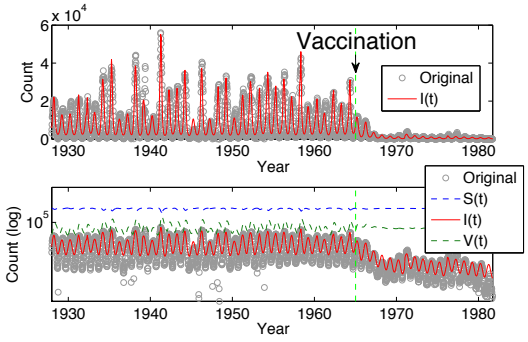
Contrast with competitors. Table 1 illustrates the relative advantages of our method. Only our approach has checks against all entries, while,

- The SI model (and SIR, SIRS, etc.) can compress the data into a fixed number of parameters, and capture the dynamics of epidemiological data, however, it cannot describe periodic patterns, and is incapable of forecasting.

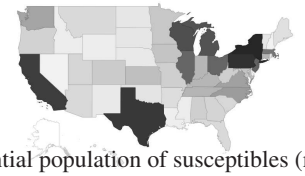
[1] Available at http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html
[2] Project Tycho at University of Pittsburgh: http://www.tycho.pitt.edu/

KDD'14, August 24–27, 2014, New York, NY, USA.

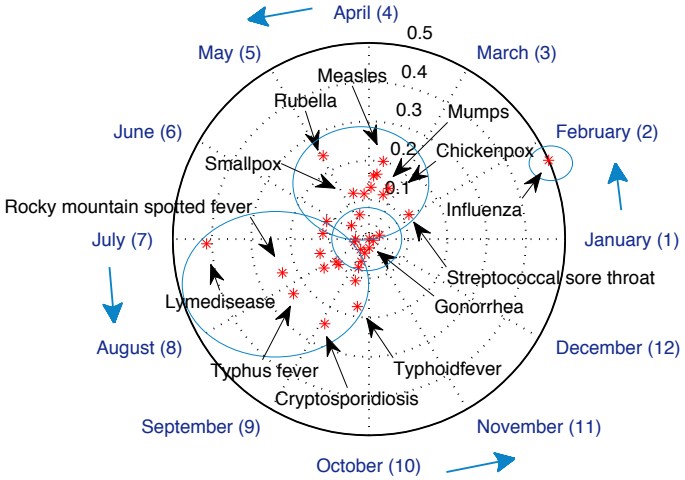
http://dx.doi.org/10.1145/2623330.2623624.



(a) Fitting result of FUNNELFIT (measles)



(b) Potential population of susceptibles (measles)



(c) Seasonality strength (radius) vs. peak season (angle)

Figure 1: Modeling power of FUNNELFIT: (a) the original number of measles cases (gray dots), and our model (red lines). It captures the yearly cycle, external spikes, and vaccination starting in 1963, as well as (b) the local sensitivity (e.g., many patients in NY, PA, CA, TX); (c) the scatter plot of the seasonality strength vs. the peak season - (angle): the peak month for each disease and (radius): the strength of the fluctuation, e.g., influenza peaks every winter, measles in the spring, and no periodicity for gonorrhea.

Table 1: Capabilities of approaches. Only our approach meets all specifications.

	SIRS	AR/*PLiF*	PARAFAC	FUNNELFIT
Compression	√	√	√	√
Domain knowledge	√			√
Missing values	√			√
Periodicity		√		√
Forecasting		√		√
Parameter free				√

- The auto regression (AR) model and *PLiF* [15] have the ability to compress and forecast sequences, but they are fundamentally unsuitable for epidemic data, and cannot capture the non-linear patterns of virus propagation.
- Our epidemic data can be turned into a tensor. PARAFAC is capable of compression, but it cannot handle missing values, periodicity, or forecasting.

Most importantly, none of above are parameter-free methods.

Contributions. Our method has the following desirable properties:

1. **Sense-making**: thanks to our modeling framework, our method can provide an intuitive explanation for epidemics, such as the seasonality of diseases, vaccination, and external shocks. It matches the behavior of various types of contagious diseases, such as measles, influenza, and smallpox.
2. **Automatic**: it is fully automatic, requiring no human intervention. Our algorithm is theoretically founded on the idea of minimizing the cost of the resulting modeling.
3. **Scalable**: it scales linearly with the input size.
4. **Generality**: it includes earlier patterns and models as special cases (e.g., SIRS), and it can be applied to various types of epidemic data including computer virus infections.

Outline. The rest of the paper is organized in the conventional way: Next we describe related work, followed by our proposed model and algorithms, experiments, discussion and conclusions.

2. RELATED WORK

We provide a survey of the related literature, which falls broadly into two categories: (1) epidemiology, and (2) pattern discovery in time series.

Epidemiology. The canonical textbook for epidemiological models including SI/SIR is Anderson and May [2]. Grenfell et al. [9] studied the recurrent travelling waves for measles, while the work in [8] explained the complex dynamical transitions in epidemics. Stone et al. [29] studied the seasonal dynamics of recurrent epidemics including measles, and identified a new threshold for predicting the occurrence of either a future epidemic, or a 'skip' (i.e., a year in which an epidemic fails to initiate). Van Panhuis et al. [32] digitized the entire history of weekly Nationally Notifiable Disease Surveillance Reports for the U.S. from 1888 to 2013.

Pattern discovery in time series. In recent years, there has been an explosion of interest in mining time series [4, 6, 21, 16]. Traditional approaches applied to data mining include auto-regression (AR), linear dynamical systems (LDS), Kalman filters (KF) and their variants [10, 15, 31]. Similarity search and pattern discovery in time sequences have also attracted huge interest [33, 13, 30, 26, 7]. Regarding large-scale time-series mining, TriMine [18] is a scalable method for forecasting co-evolving multiple (thousands of) sequences, while, [17] developed a fully-automatic mining algorithm for co-evolving time sequences. Rakthanmanon et al. [25] proposed a similarity search algorithm for "trillions of time series" under the DTW distance. Recently, analyses of epidemics, social media, propagation and the cascades they create have attracted much interest [24, 12, 14, 23, 19].

However, none of these methods specifically focused on automatic mining of non-linear dynamics in coevolving epidemics.

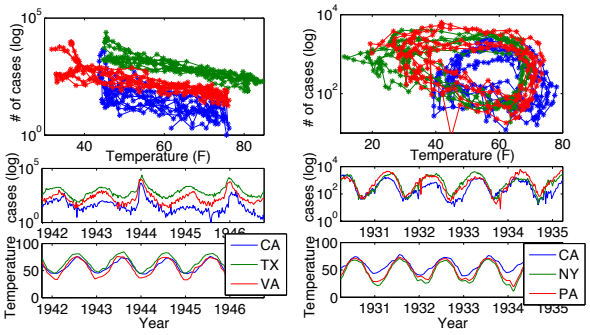
3. PROPOSED MODEL

In this section we present our proposed model.

3.1 Design philosophy of FUNNEL

Data description. The Project Tycho [32] covers more than a century of weekly surveillance reports of nationally notifiable diseases (56 diseases in total) for all 50 states in the U.S, from 1888 to the present, with 87,950,807 reported individual cases for diseases.

This dataset consists of tuples of the form: *(disease, location, timestamp)*. We then have a collection of entries with d unique diseases, and l states, with duration n (on a weekly basis). We can



(a) Influenza (in CA, TX, VA) (b) Measles (in CA, NY, PA)

Figure 2: The air temperature vs. # of cases: (a) influenza is completely anti-correlated with the air temperature (i.e., peaking in the winter), while, (b) measles also has strong periodicity, but it peaks in the spring (i.e., with a phase shift).

treat this set of $d \times l$ epidemic sequences as a 3rd-order tensor, i.e., $\mathcal{X} \in \mathbb{N}^{d \times l \times n}$, where the element $x_{ij}(t)$ of $\mathcal{X}$ shows the total number of entries of the i-th disease in the j-th state at time-tick t.

For example, ('measles', 'PA', 'April 1-7, 1931'; 4740), means that the number of cases due to 'measles' in 'PA' on 'April 1-7 in 1931' is '4740'.

We refer to each sequence of the i-th disease in the j-th state: $\boldsymbol{x}_{ij} = \{x_{ij}(t)\}_{t=1}^{n}$, as a "local/state"-level epidemic sequence. Similarly, we can turn these local sequences into "global/country"-level epidemics: $\bar{\boldsymbol{x}}_i = \{\bar{x}_i(t)\}_{t=1}^{n}$, where $\bar{x}_i(t)$ shows the total count of the i-th disease at time-tick t, i.e., $\bar{x}_i(t) = \sum_{j=1}^{l} x_{ij}(t)$.

Preliminary observations. Here, we provide the reader with several important observations. Figure 2 shows the scatter plots (top) and sequence plots (bottom) of the original local-level sequences of influenza and measles counts in three states, versus the average air temperature for five years. [3] In Figure 2 (a), influenza cases are strongly anti-correlated with the air temperature, corresponding to influenza epidemics in colder seasons. On the other hand, for measles (Figure 2 (b)), the scatter plot exhibits characteristic loop shapes, which indicates that there is a phase shift of measles vs. temperature - actually, measles peaks in the spring. As shown in Figure 1 (c), there are several groups of infectious diseases with specific seasonal patterns, including children's diseases (e.g., measles, mumps) in the spring, and tick-borne diseases (e.g., Lyme disease) in the summer. Consequently we have:

OBSERVATION 1 (DISEASE SEASONALITY). *Many diseases have yearly cycles with different phases, that is, they are correlated with air temperature and the seasons.*

The next observation refers to the abrupt decline of several diseases. Luckily, many diseases have been eradicated or significantly reduced over the last century, through various factors including vaccination, sanitation and antibiotics. For example, in Figure 1 (a), the number of measles cases has been decreasing since the vaccination program was introduced in 1963. We will collectively refer to such abrupt declines as *disease reduction* effects.

OBSERVATION 2 (DISEASE REDUCTION EFFECT). *Many infectious diseases have been reduced or eliminated through vaccination programs, antibiotics, sanitation, etc.*

Next, let us look at the topic from a local point of view. In Figure 2, three local sequences are correlated with each other, but with different fractions of patients, which correspond to the number of susceptible people in each state. For example, measles mainly affects children, and so, the more children there are, the more cases of measles there will be (see, NY, PA, CA, TX, in Figure 1 (b)).

[3] National climate data center: http://www.ncdc.noaa.gov/cag/

Table 2: Symbols and definitions

Symbol	Definition
d	Number of diseases
l	Number of states (i.e., locations)
n	Duration of sequences
$\mathcal{X}$	3rd-order tensor ($\mathcal{X} \in \mathbb{N}^{d \times l \times n}$)
$\boldsymbol{x}_{ij}$	Local-level epidemic sequence of disease i in state j
$\bar{\boldsymbol{x}}_i$	Global-level epidemic sequence of disease i
$S_{ij}(t)$	Count of susceptibles of disease i in state j at time t
$I_{ij}(t)$	Count of infectives of disease i in state j at time t
$V_{ij}(t)$	Count of vigilants of disease i in state j at time t
$\mathbf{B}$	Base matrix ($d \times 6$) i.e., $\mathbf{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}$
$\mathbf{R}$	Disease reduction matrix ($d \times 2$) i.e., $\mathbf{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_d\}$
$\mathbf{N}$	Geo-disease matrix ($d \times l$) i.e., $\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$
$\mathcal{E}$	External shock tensor i.e., $\mathcal{E} = \{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{E}^{(\mathbf{S})}\}$
$\mathcal{M}$	Mistake tensor i.e., $\mathcal{M} = \{m_{ij}(t)\}_{i,j,t=1}^{d,l,n}$
$\mathcal{F}$	Complete set of FUNNEL i.e., $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$

OBSERVATION 3 (AREA SPECIFICITY AND SENSITIVITY). *For each disease, neighbors are correlated with different sensitivity.*

The last two observations are the extra properties of epidemics. Figure 1 (a) shows large outbreaks of measles in 1941 and 1958, while Figure 2 (a) shows two large flu pandemics in 1944 and 1946.

OBSERVATION 4 (EXTERNAL SHOCK EVENTS). *There are some extreme spikes, representing major events such as historical flu pandemics.*

Basically, real-world datasets are subject to quality constraints such as typing errors and incorrect reports (we refer to them as "mistakes").

OBSERVATION 5 (MISTAKES). *There are some implausible spikes, which are completely independent of the dynamics of epidemic patterns.*

Summary. In this paper, we propose a new model, namely, FUNNEL, which tries to incorporate all the above important properties that we observed in the real epidemic data. Consequently, we would like to capture the following properties:

- (**P1**): yearly periodicity
- (**P2**): disease reduction effects
- (**P3**): area specificity and sensitivity
- (**P4**): external shock events
- (**P5**): mistakes, incorrect values

For simplicity, let's focus on a simple step first, where (a) we assume that we are given a single epidemic sequence, say, the number of measles cases in NY. We then (b) extend our model to multiple co-evolving epidemics, that is, to capture the individual patterns of d diseases in l states.

3.2 FUNNEL - with a single epidemic

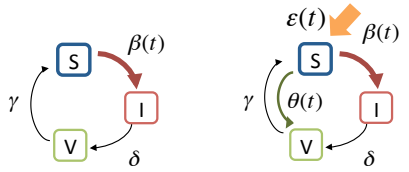
We begin with the simplest case, where we assume that we are given a single epidemic sequence.

3.2.1 *Base model* - FUNNEL-BASE

The model we propose has nodes (=people) of three classes:

- **S**usceptible: nodes in this class can get infected by any neighboring node who is infectious.
- **I**nfected: nodes who have been infected and are capable of transmitting the infection to those in the susceptible class.
- **V**igilant (i.e., recovered/immune): nodes in this class cannot get infected nor can they cause infections.

Figure 3 (a) shows a diagram of our base model, where, $\beta(t)$ represents the rate of effective contacts between infected and susceptible individuals; δ is the rate at which infected individuals recovered; γ is the immunization loss probability for a recovered or



(a) FUNNEL-BASE (b) FUNNEL-RE

Figure 3: FUNNEL diagrams: there are three classes - susceptible (i.e., healthy, but can get infected), infected (i.e., capable of transmission), vigilant (i.e., healthy, and cannot get infected).

vigilant individual. [4] More importantly, to handle the first property of epidemics: (**P1**), we assume that the infection rate $\beta(t)$ is a periodic function of time t. We refer to it as FUNNEL-BASE.

MODEL 1 (FUNNEL-BASE). *Let $S(t)$, $I(t)$, $V(t)$ be the number of susceptible, infected, vigilant people at time-tick t. Our base model is governed by the following equations:*

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)S(t)I(t) + \gamma V(t) \\ I(t+1) &= I(t) + \beta(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) \end{aligned} \quad (1)$$

where $\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos\left(\frac{2\pi}{P_p}(t+P_s)\right)\right)$, $P_p = 52$, [5] and, we have the invariant $N = S(t) + I(t) + V(t)$, with initial conditions $S(1) = N-1$, $I(1) = 1$, $V(1) = 0$.

Consequently, FUNNEL-BASE consists of a set of the following parameters: $\boldsymbol{b} = \{N, \beta_0, \delta, \gamma, P_a, P_s\}$, specifically,

- N: Potential population of the disease. N is composed of susceptible, infected and vigilant individuals.
- β_0: Rate of effective contacts between infected and susceptible individuals averaged over the year.
- δ: Healing rate of the disease.
- γ: Forgetting rate of the diseases.
- P_a: Amplitude of the fluctuation, specifically, it gives the relative value of the peak/off-season.
- P_s: Phase shift of the seasonal cycle.

3.2.2 *With disease reduction* - FUNNEL-R

With respect to the second property: (**P2**), we also introduce an essential concept, namely, the "disease reduction" effect.

MODEL 2 (FUNNEL-R). *We add a disease reduction rate: $\theta(t)$, to capture the effect of the disease reduction program, that is,*

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\ I(t+1) &= I(t) + \beta(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \end{aligned} \quad (2)$$

where, the disease reduction program started at time t_θ and $\theta(t)$ is defined as: $\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}$

The model is identical to FUNNEL-BASE, with the addition of the disease reduction factor, $\theta(t)$, which corresponds to the direct immunization probability when susceptible (see Figure 3 (b)). Note that this effect is due to vaccination, antibiotics and any other anti-disease factors. Hereafter, we simply say the "disease reduction effect", unless otherwise specified.

In addition to the base parameters $\boldsymbol{b}$, FUNNEL-R requires a set of two parameters, $\boldsymbol{r} = \{t_\theta, \theta_0\}$, where,

- t_θ: Starting time of the disease reduction effect.
- θ_0: Diffusion rate of the disease reduction effect.

[4] This factor also incorporates the birth and mortality rate.
[5] We have 52 time-ticks (weeks) in one year.

3.2.3 *With external shocks* - FUNNEL-RE

Next, with respect to the property: (**P4**), we assume that there are external shock events, such as flu pandemics. So how do we go about capturing such unexpected patterns? Assume that there is a swine flu pandemic. In this situation, many more people in the susceptible class would become infected than in previous years.

An elementary concept we need to introduce is the *temporal susceptible rate*: $\epsilon(t)$. Figure 3 (b) describes how this is done. The idea is that the number of susceptibles $S(t)$ is the count of victims available for infection, and if there is an external shock event at time-tick t, the virus attacks are much stronger than usual, and, each victim-attack pair would lead to a new victim, and will eventually cause a major pandemic.

MODEL 3 (FUNNEL-RE). *Our full model can be described as the following equations:*

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\ I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \end{aligned} \quad (3)$$

In addition, we introduce the temporal susceptible rate, $\epsilon(t)$, which is defined as follows:

$$\epsilon(t) = 1 + \sum_{i=1}^{k} f(t; \boldsymbol{e}_i^{(\boldsymbol{T})}), f(t; \boldsymbol{e}^{(\boldsymbol{T})}) = \begin{cases} \epsilon_0 & (t_\mu - t_\sigma < t < t_\mu + t_\sigma) \\ 0 & (else) \end{cases}$$

where, k is the number of shocks, and if $k = 0$, then $\epsilon(t) = 1$.

Here, each external shock consists of $\boldsymbol{e}^{(\boldsymbol{T})} = \{t_\mu, t_\sigma, \epsilon_0\}$, i.e.,

- t_μ: Central time point of the external shock event.
- t_σ: Duration of the event.
- ϵ_0: Strength of the external shock effect.

3.3 FUNNEL - with multi-evolving epidemics

So far we have seen how FUNNEL captures the dynamics of a single epidemic sequence. The next question is, "how can we apply FUNNEL to multiple co-evolving epidemics in $\mathcal{X}$, and capture the individual behavior of d diseases in l states?"

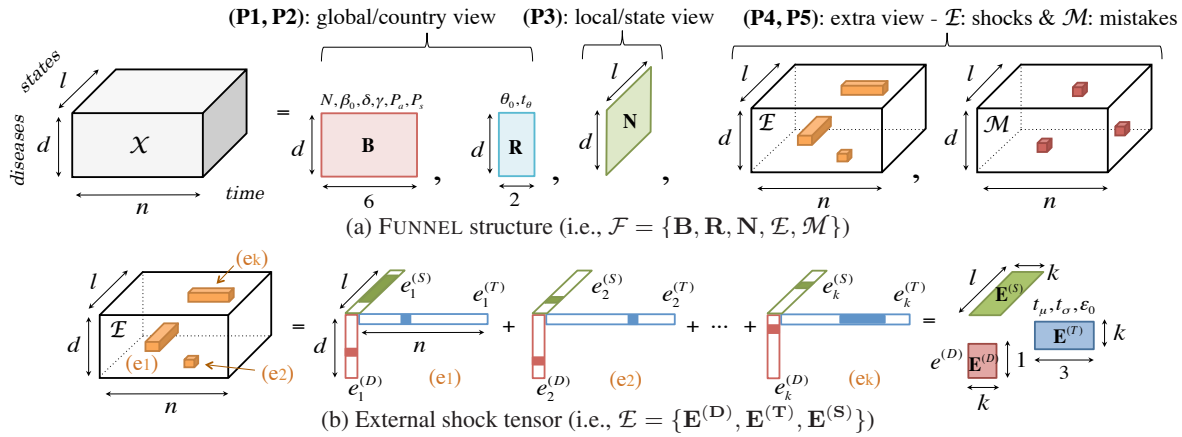
We want to estimate the parameter set of FUNNEL, for each individual epidemic sequence in $\mathcal{X}$. The straightforward solution would be that we consider a set of $(d \times l)$ sequences of length n generated from $\mathcal{X}$: $\{\boldsymbol{x}_{ij}\}_{i,j=1}^{d,l}$, (i.e, "local-level" epidemic sequences), and estimate parameter set: $\{\boldsymbol{b}, \boldsymbol{r}, \boldsymbol{e}^{(\boldsymbol{T})}\}$ for each sequence. However, some of the *(disease, state)* pairs have very sparse sequences (e.g., Lyme disease in Alaska), which derails the fitting result. Also, we are interested in capturing global/country-level patterns, as well as local/state-level trends. So how can we deal with this issue? We thus propose "sharing" the global-level parameters for all l states, to achieve much better modeling.

FUNNEL - full model parameter set. Our goal is to extract the main trends and external patterns of co-evolving epidemics $\mathcal{X} \in \mathbb{N}^{d \times l \times n}$, and make a good representation of $\mathcal{X}$. Figure 4 shows our modeling framework. Given epidemic data $\mathcal{X}$, we try to find important patterns with respect to the following five aspects, (**P1**) $\mathbf{B}$: base properties of diseases, (**P2**) $\mathbf{R}$: disease reduction effects, (**P3**) $\mathbf{N}$: locations vs. diseases, (**P4**) $\mathcal{E}$: external shock events, and (**P5**) $\mathcal{M}$: mistake values. The first two are global/country-level parameter sets, and the third is a local/state-level parameter set, and the last two are used for describing extra trends in $\mathcal{X}$.

DEFINITION 1 (COMPLETE SET OF FUNNEL). *Let $\mathcal{F}$ be a complete set of parameters (namely, $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$) that describe the global/local/extra patterns of epidemics in $\mathcal{X}$.*

Next, we will see each property in detail.

(P1), (P2) Global/country view. Basically, we assume that the following parameters are the same for all l states.



(a) FUNNEL structure (i.e., $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$)

(b) External shock tensor (i.e., $\mathcal{E} = \{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{E}^{(\mathbf{S})}\}$)

Figure 4: Illustration of FUNNEL structure: (a) we extract the important behavior of epidemics from an original tensor $\mathcal{X}$, (i.e., the base matrix B, disease reduction matrix R, geo-disease matrix N, external shock tensor $\mathcal{E}$, and mistake tensor $\mathcal{M}$); Also, (b) the external shock tensor $\mathcal{E}$ can be described as a set of matrices i.e., $\{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{E}^{(\mathbf{S})}\}$, for *disease, state, time*.

DEFINITION 2 (BASE MATRIX $\mathbf{B}$ $(d \times 6)$). *Let $\mathbf{B}$ be a set of base parameters of d diseases, i.e., $\mathbf{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}$ where $\boldsymbol{b}_i$ is the parameter set of the i-th disease.*

For example, the infection/healing rate of measles should be the same for NY and FL. Similarly, once the measles vaccine has been introduced, (i.e., the disease reduction effect), it could be immediately spread all over the country, that is, the starting time of the disease reduction effect would be the same for all locations.

DEFINITION 3 (DISEASE REDUCTION MATRIX $\mathbf{R}$ $(d \times 2)$). *Let $\mathbf{R}$ be a parameter set of the reduction of d diseases, i.e., $\mathbf{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_d\}$ where $\boldsymbol{r}_i$ is the parameter set of the i-th disease.*

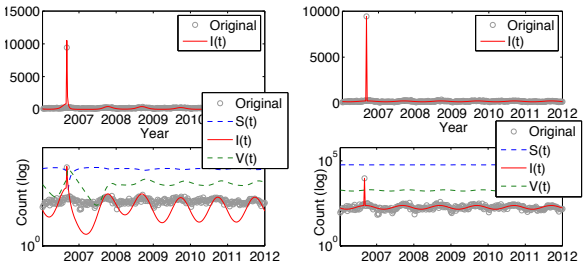
(P3) Local/state view. We also want to analyze and explain local-specific patterns and trends in $\mathcal{X}$. So, what is the difference between measles *in NY* and *in FL*? Our answer is: they are exactly the same, except for the "local sensitivity" of the disease. The idea is that we share the parameters of the global-level matrices for all l states with but one exception, local sensitivity, N_{ij}, which describes the potential population of the disease i in the j-th state. Specifically, we set the invariant, $N_{ij} = S_{ij}(t) + I_{ij}(t) + V_{ij}(t)$ in Model 3. This parameter corresponds to the fraction of individuals who are likely to be infected by the disease. For example, NY has more measles patients than FL, because it mainly affects children (i.e., there were more children in NY than FL, in the last century).

DEFINITION 4 (GEO-DISEASE MATRIX $\mathbf{N}$ $(d \times l)$). *Let $\mathbf{N}$ be a parameter set of the potential population of d diseases and l states, i.e., $\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$, where N_{ij} is the potential population of susceptibles of the i-th disease in the j-th state.*

(P4) Extra view - external shocks. Consider that in 1946 a serious flu pandemic spread throughout the country. We want to describe this external shock event in terms of three aspects, *(disease, state, time)*, e.g., (e1) "influenza, country-wide, 1946". Similarly, there was a community-wide outbreak of cryptosporidiosis in Utah, in 2007. i.e., (e2) "cryptosporidiosis, Utah, 2007". To describe these external shock events, we create a new parameter set, namely external shock tensor $\mathcal{E}$, which consists of a set of k external shock events, as described in Figure 4 (b).

The external shock tensor $\mathcal{E}$ can be also decomposed into three-aspect matrices, $\{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{S})}, \mathbf{E}^{(\mathbf{T})}\}$, each of which shows the patterns in terms of *disease, state, time*. A single external shock event can be described as triplet vectors $\{\boldsymbol{e}^{(D)}, \boldsymbol{e}^{(S)}, \boldsymbol{e}^{(T)}\}$, where,

- The shock (disease) vector $\boldsymbol{e}^{(D)}$ shows the assignment of the external shock to the disease ID (i.e., $1 \leq \boldsymbol{e}^{(D)} \leq d$).



(a) External shock fitting (b) Mistake fitting

Figure 5: External shock vs. mistake for giardiasis in 2007: (a) the model (i.e., red line) is greatly influenced by the large distance of the outlier from the original sequence, while (b) it filters out the mistake point, and fits the sequence very well.

- The shock (state) vector $\boldsymbol{e}^{(S)}$ describes the participation strength of each state for each external shock event.
- The shock (time) vector $\boldsymbol{e}^{(T)}$ shows the temporal pattern of the external shock event.

Specifically, $\boldsymbol{e}^{(T)}$ is the global-level parameters, as described in subsubsection 3.2.3, and $\boldsymbol{e}^{(S)}$ is the local-level parameters, i.e., $\boldsymbol{e}^{(S)} = \{e_j^{(S)}\}_{j=1}^{l}$, where, we change ϵ_0 in Model 3 to describe the strength of the external shock for l individual locations. That is, the strength of the shock effect in the j-th state is, $\epsilon_0 \cdot e_j^{(S)}$. Consequently, we have the following:

DEFINITION 5 (EXTERNAL SHOCK TENSOR $\mathcal{E}$). *Let $\mathcal{E}$ be a 3rd-order tensor of k external shock events, i.e., $\mathcal{E} = \{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{S})}, \mathbf{E}^{(\mathbf{T})}\}$, where triplet matrices show the parameters in terms of three aspects, namely, "disease", "state", and "time".*

(P5) Extra view - Mistakes. Basically, real datasets contain many errors such as incorrect reports. FUNNEL should detect and filter them out as outliers. We thus introduce an additional concept.

DEFINITION 6 (MISTAKE TENSOR $\mathcal{M}$). *Let $\mathcal{M}$ be a 3rd-order tensor of mistake data points, where, the element $m_{ij}(t)$ of $\mathcal{M}$ shows the entry of the i-th disease in the j-th state at time-tick t.*

Note that $\mathcal{M}$ is very sparse, and very often $m_{ij}(t) = 0$.

Figure 5 compares the fitting results of the external shock vs. mistake for the giardiasis cases, which contains an incongruous point in 2006 (approximately, $10,000$). In this case, the point should be treated as (b) a mistake value, instead of (a) an external shock event; in figure (a), the model (red line) is strongly influenced by the extreme point, while in (b), it successfully captures the real patterns of the original sequence.

4. OPTIMIZATION ALGORITHM

In this section, we describe our fitting algorithm, FUNNELFIT. Our goal is to extract the important patterns of epidemics from $\mathcal{X}$. More specifically, the problem that we want to solve is as follows:

PROBLEM 1. **Given** *a tensor $\mathcal{X}$ of $(disease, state, time)$ triplets,* **Find** *a compact description that best summarizes $\mathcal{X}$, that is, $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$.*

We want to find a good representation $\mathcal{F}$ to solve the problem. The essential questions are: (a) How can we estimate the parameter set that best captures the dynamics and patterns in $\mathcal{X}$? (b) How should we decide the number of external shocks k? (c) How can we ignore mistake (i.e., outlier) values in $\mathcal{X}$?

4.1 Model quality and data compression

We provide a new intuitive coding scheme, which is based on the minimum description length (MDL) principle. In short, it follows the assumption that the more we can compress the data, the more we can learn about its underlying patterns.

Model description cost. The description complexity of model parameter set consists of the following terms,

- The number of diseases d, states l, and time-ticks n require $\log^*(d) + \log^*(l) + \log^*(n)$ bits. [6]
- The model parameter set of the base ($\mathbf{B}$), reduction ($\mathbf{R}$), geo-disease ($\mathbf{N}$) matrices require $d \times 6$, $d \times 2$, $d \times l$ parameters, respectively, i.e., $Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N}) = c_F \cdot d(6 + 2 + l)$, where c_F is the floating point cost[7].

Similarly, the model description cost of the external shock tensor $\mathcal{E} = \{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{S})}, \mathbf{E}^{(\mathbf{T})}\}$) consists of the following:

- The number of external shocks k requires $\log^*(k)$ bits.
- The shock-disease matrix $\mathbf{E}^{(\mathbf{D})}$ requires $k \log(d)$.
- The shock-time parameter set $\boldsymbol{e}^{(\boldsymbol{T})} = \{t_\mu, t_\sigma, \epsilon_0\}$ in $\mathbf{E}^{(\mathbf{T})}$ requires $\log(n)$, $\log(n)$, c_F, respectively.
- The shock-state matrix $\mathbf{E}^{(\mathbf{S})}$ requires $c_F \cdot kl$.

Consequently, the model cost of the external shock tensor $\mathcal{E}$ is $Cost_M(\mathcal{E}) = \log^*(k) + k\big(\log(d) + 2\log(n) + c_F \cdot (1 + l)\big)$. The model cost of mistake tensor $\mathcal{M}$ consists of

- The number of non-zero elements in $\mathcal{M}$ requires $\log^*(|\mathcal{M}|)$
- The location of each non-zero element and its value, $m_{ij}(t)$ require $\log(d)$,$\log(l)$,$\log(n)$, $\log^*(m_{ij}(t))$, respectively.

Thus, $Cost_M(\mathcal{M}) = \log^*(|\mathcal{M}|) + \sum_{m_{ij}(t)>0}^{|\mathcal{M}|}(\log(d) + \log(l) + \log(n) + \log^*(m_{ij}(t)))$, where, $|\mathcal{M}|$ is the number of non-zero elements in $\mathcal{M}$.

Data coding cost. Once we have decided the full parameter set $\mathcal{F}$, we can encode the data $\mathcal{X}$ using Huffman coding [3], i.e., a number of bits is assigned to each value in $\mathcal{X}$, which is the logarithm of the inverse of the probability of the values (here, we use a Gaussian distribution). The encoding cost of $\mathcal{X}$ given $\mathcal{F}$ is: $Cost_C(\mathcal{X}|\mathcal{F}) = \sum_{i,j,t=1}^{d,l,n} \log_2 p^{-1}_{Gauss(\mu,\sigma)}(x_{ij}(t) - m_{ij}(t) - I_{ij}(t))$, where, $x_{ij}(t)$, $m_{ij}(t)$ are the elements in $\mathcal{X}$ and the mistake tensor $\mathcal{M}$, respectively, and $I_{ij}(t)$ is the estimated count of infections (i.e., Model 3). Also, μ and σ are the mean and variance of the distance between the original and estimated values. [8]

Putting it all together. Consequently, the total code length for $\mathcal{X}$ with respect to a given parameter set $\mathcal{F}$ can be described as follows:

$$\begin{aligned} Cost_T(\mathcal{X}; \mathcal{F}) &= \log^*(d) + \log^*(l) + \log^*(n) \\ &+ Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N}) \\ &+ Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F}) \end{aligned} \quad (4)$$

Thus our next goal is to minimize the above function.

4.2 Multi-layer optimization

Until now, we have seen how we can measure the goodness of the representation of $\mathcal{X}$, if we are given a candidate parameter set $\mathcal{F}$. The next question is, how to find an optimal solution of the full parameter set: $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$.

As described in subsection 3.3, our FUNNEL model consists of multiple parameter sets, each of which explains either the local or global pattern of epidemics in $\mathcal{X}$. For example, the base and reduction matrices $\mathbf{B}$, $\mathbf{R}$ explain the global-level behavior of each disease, while the geo-disease matrix $\mathbf{N}$ describes the local-level trends. Also, the extra tensors $\mathcal{E}$, $\mathcal{M}$ consist of both the global and local-level parameters. More specifically, the external shocks consists of $\mathcal{E} = \{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{S})}, \mathbf{E}^{(\mathbf{T})}\}$), where, the first two are the global-level, and the last one is the local-level. Similarly, the mistake tensor can also be describes by the triplet matrix $\mathcal{M} = \{\mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{T})}\}$), each of which describes the location of the mistake values in terms of *disease, state, time*. So, how can we efficiently estimate these model parameters?

We propose a multi-layer optimization algorithm, to search for the optimal solution in terms of both the global and local-level parameters. The idea is that we split parameter set $\mathcal{F}$ into two subsets, i.e., $\mathcal{F}_\mathcal{G}$ and $\mathcal{F}_\mathcal{L}$, each of which corresponds to a global/local-level parameter set, and try to fit the parameter sets separately. Our algorithm consists of the following two phases:

- GLOBALFIT: find good global-level parameters for $\{\bar{\boldsymbol{x}}_i\}_{i=1}^{d}$, i.e., $\mathcal{F}_\mathcal{G} = \{\mathbf{B}, \mathbf{R}, \mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{T})}\}$
- LOCALFIT: find good local-level parameters: for $\{\boldsymbol{x}_{ij}\}_{i,j=1}^{d,l}$, i.e., $\mathcal{F}_\mathcal{L} = \{\mathbf{N}, \mathbf{E}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{S})}\}$

Here, the global epidemic sequence of the i-th disease: $\bar{\boldsymbol{x}}_i$ can be described as the sum of the l local sequences, i.e., $\bar{\boldsymbol{x}}_i(t) = \sum_{j=1}^{l} \boldsymbol{x}_{ij}(t)$. Algorithm 1 shows an overview of FUNNELFIT. Given a tensor $\mathcal{X}$, it finds the full set of FUNNEL parameters.

Algorithm 1 FUNNELFIT ($\mathcal{X}$)

```
1: Input: Tensor X (d × l × n)
2: Output: Complete set of parameters, i.e., F = {B, R, N, E, M}
3: /* Parameter fitting for global-level sequences */
4: {F_G} =GLOBALFIT (X);
5: /* Parameter fitting for local-level sequences */
6: {F_L} =LOCALFIT (X, F_G);
7: return F = {F_G, F_L};
```

4.2.1 Global-level parameter fitting

Given a tensor $\mathcal{X}$, our sub-goal is to find the optimal global-level parameter set: $\mathcal{F}_\mathcal{G}$, to minimize the cost function (i.e., Equation 4). We want to fit the basic parameters of each disease (i.e., the base and reduction matrices), and estimate the appropriate number of external shocks and mistake values, simultaneously. Finding the appropriate number of external-shocks/mistakes is a particular issue here, because the parameter fittings are very sensitive to outliers, as described in Figure 5 (a). To find a good basic parameter set for $\mathcal{X}$, we have to filter out the external shocks and mistakes appropriately. Simultaneously, a good external-shock/mistake filter requires a well estimated base model. We escape this circular dependency by applying an iterative method that employs external-shocks/mistakes detection and filtering, and basic model fitting in an alternating way until the cost function reaches a minimum value.

[6] Here, $\log^*$ is the universal code length for integers.

[7] We used 4×8 bits in our setting.

[8] Here, μ, σ need $2c_F$ bits, but we can eliminate them because they are constant values and independent of our modeling.

Algorithm 2 GLOBALFIT ($\mathcal{X}$)

Input: Tensor $\mathcal{X}$
Output: Set of global-level parameters $\mathcal{F}_\mathcal{G}$
for $i = 1 : d$ **do**
  Create $\bar{\boldsymbol{x}}_i$ from $\mathcal{X}$; /* Global sequence $\bar{\boldsymbol{x}}_i$ of i-th disease */
  /* Initialize external shocks and mistake values for disease i */
  $\mathbf{E}_i^{(\mathbf{D})} = \mathbf{E}_i^{(\mathbf{T})} = \mathbf{M}_i^{(\mathbf{D})} = \mathbf{M}_i^{(\mathbf{T})} = \emptyset$;
  while improving the cost **do**
    $\boldsymbol{b}_i = \arg\min_{\boldsymbol{b}'_i} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}'_i, \boldsymbol{r}_i, \mathbf{E}_i^{(\mathbf{T})}, \mathbf{M}_i^{(\mathbf{T})})$; /* Base */
    $\boldsymbol{r}_i = \arg\min_{\boldsymbol{r}'_i} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}_i, \boldsymbol{r}'_i, \mathbf{E}_i^{(\mathbf{T})}, \mathbf{M}_i^{(\mathbf{T})})$; /* Reduction */
    $\mathbf{E}_i^{(\mathbf{D})} = \mathbf{E}_i^{(\mathbf{T})} = \mathbf{M}_i^{(\mathbf{D})} = \mathbf{M}_i^{(\mathbf{T})} = \emptyset$; /* Initialize values */
    /* Find external shocks and mistakes for disease i */
    while improving the cost **do**
      $\boldsymbol{e}^{(\boldsymbol{T})} = \arg\min_{\boldsymbol{e}'^{(\boldsymbol{T})}} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}_i, \boldsymbol{r}_i, \{\mathbf{E}_i^{(\mathbf{T})} \cup \boldsymbol{e}'^{(\boldsymbol{T})}\}, \mathbf{M}_i^{(\mathbf{T})})$;
      $\boldsymbol{m}^{(\boldsymbol{T})} = \arg\min_{\boldsymbol{m}'^{(\boldsymbol{T})}} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}_i, \boldsymbol{r}_i, \mathbf{E}_i^{(\mathbf{T})}, \{\mathbf{M}_i^{(\mathbf{T})} \cup \boldsymbol{m}'^{(\boldsymbol{T})}\})$;
      /* Compare external shock vs. mistake */
      if $Cost_T(\bar{\boldsymbol{x}}_i; \boldsymbol{e}^{(\boldsymbol{T})}) < Cost_T(\bar{\boldsymbol{x}}_i; \boldsymbol{m}^{(\boldsymbol{T})})$ **then**
        /* External shock wins - treat as an external shock */
        $\mathbf{E}_i^{(\mathbf{D})} = \{\mathbf{E}_i^{(\mathbf{D})} \cup i\}$; $\mathbf{E}_i^{(\mathbf{T})} = \{\mathbf{E}_i^{(\mathbf{T})} \cup \boldsymbol{e}^{(\boldsymbol{T})}\}$;
      else
        /* Mistake wins - treat as a mistake value */
        $\mathbf{M}_i^{(\mathbf{D})} = \{\mathbf{M}_i^{(\mathbf{D})} \cup i\}$; $\mathbf{M}_i^{(\mathbf{T})} = \{\mathbf{M}_i^{(\mathbf{T})} \cup \boldsymbol{m}^{(\boldsymbol{T})}\}$;
      end if
    end while
  end while
  /* Update parameter set of i-th disease */
  $\mathbf{B} = \mathbf{B} \cup \boldsymbol{b}_i$; $\mathbf{R} = \mathbf{R} \cup \boldsymbol{r}_i$;
  $\mathbf{E}^{(\mathbf{D})} = \mathbf{E}^{(\mathbf{D})} \cup \mathbf{E}_i^{(\mathbf{D})}$; $\mathbf{E}^{(\mathbf{T})} = \mathbf{E}^{(\mathbf{T})} \cup \mathbf{E}_i^{(\mathbf{T})}$;
  $\mathbf{M}^{(\mathbf{D})} = \mathbf{M}^{(\mathbf{D})} \cup \mathbf{M}_i^{(\mathbf{D})}$; $\mathbf{M}^{(\mathbf{T})} = \mathbf{M}^{(\mathbf{T})} \cup \mathbf{M}_i^{(\mathbf{T})}$;
end for
return $\mathcal{F}_\mathcal{G} = \{\mathbf{B}, \mathbf{R}, \mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{T})}\}$;

External shock vs. mistake. There is also an important issue regarding the external shock vs. the mistake value. We want to distinguish automatically between an external shock event and a typing error. For example, in Figure 5, there is a clear "typo", rather than an external shock event. Our coding scheme enables us to provide the answer. The idea is that we try to fit the parameters by treating the data as both an external shock event and a mistake value, and then compare the cost of the two alternatives. For Figure 5, the cost of (b) is less than (a), thus the algorithm determines that there is a mistake value in 2007.

Algorithm. Algorithm 2 is a detailed algorithm of the global-level fitting. Given a tensor $\mathcal{X}$, it creates a set of d global sequences: $\{\bar{\boldsymbol{x}}_i\}_{i=1}^d$. It tries to fit the global-level parameter set, as well as find the appropriate number of external-shocks/mistakes. We use the *Levenberg-Marquardt (LM)* algorithm to minimize the cost function. Note that the extra tensors $\mathcal{E}$ and $\mathcal{M}$ consist of an entry $(disease, state, time)$, but this algorithm can find only the global-level entry, which consists of $(disease, time)$. The local-level entries $\mathbf{E}^{(\mathbf{S})}$ and $\mathbf{M}^{(\mathbf{S})}$ can be computed by local-level parameter fitting, as shown next in Algorithm 3. Also, the cost function (Equation 4) includes the cost of local-level parameters such as $\mathbf{N}$, but these terms are independent of the global model fitting. Hence, we can simply consider them to be constant.

4.2.2 Local-level parameter fitting

Given a set of $d \times l$ local-level sequences, $\{\boldsymbol{x}_{ij}\}_{i,j=1}^{d,l} \in \mathcal{X}$, and a set of global-level parameters, $\mathcal{F}_\mathcal{G}$, our next goal is to fit the individual parameters of each disease in each state, that is, $\mathcal{F}_\mathcal{L} = \{\mathbf{N}, \mathbf{E}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{S})}\}$. We propose an iterative optimization algorithm (see Algorithm 3). Our algorithm searches for the optimal solution with respect to (a) the geo-disease matrix $\mathbf{N}$, (b) the local-level external shocks $\mathbf{E}^{(\mathbf{S})}$, and (c) the local-level mistake values $\mathbf{M}^{(\mathbf{S})}$, so that the total coding cost is minimized.

Algorithm 3 LOCALFIT ($\mathcal{X}, \mathbf{B}, \mathbf{R}, \mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{T})}$)

Input: (a) Tensor $\mathcal{X}$, (b) global-level parameter set $\mathcal{F}_\mathcal{G}$
Output: Set of local-level parameters, i.e., $\mathcal{F}_\mathcal{L}$
while improving the cost **do**
  /* For each local sequence $\boldsymbol{x}_{ij}$ of i-th disease in j-th state */
  for $i = 1 : d$ **do**
    for $j = 1 : l$ **do**
      $N_{ij} = \arg\min_{N'_{ij}} Cost_C(\boldsymbol{x}_{ij} | \mathbf{B}, \mathbf{R}, N'_{ij}, \mathcal{E}, \mathcal{M})$;
    end for
  end for
  for each external shock $(\boldsymbol{e}^{(\boldsymbol{D})}, \boldsymbol{e}^{(\boldsymbol{S})}, \boldsymbol{e}^{(\boldsymbol{T})}) \in \mathcal{E}$ **do**
    Update $\boldsymbol{e}^{(\boldsymbol{S})}$ to minimize the cost /* Local participation rate */
  end for
  for each mistake $(\boldsymbol{m}^{(\boldsymbol{D})}, \boldsymbol{m}^{(\boldsymbol{S})}, \boldsymbol{m}^{(\boldsymbol{T})}) \in \mathcal{M}$ **do**
    Update $\boldsymbol{m}^{(\boldsymbol{S})}$ to minimize the cost /* Mistake value */
  end for
end while
return $\mathcal{F}_\mathcal{L} = \{\mathbf{N}, \mathbf{E}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{S})}\}$;

LEMMA 1. *The computation time of* FUNNELFIT *is* $O(dln)$.

PROOF. To create the global-level sequences from $\mathcal{X}$, the algorithm requires $O(dln)$ time. For global-level parameter fitting, it needs $O(\#iter \cdot (k + |\mathcal{M}|) \cdot dn)$ time, where $\#iter$ is the number of iterations, k and $|\mathcal{M}|$ show the number of external shocks and non-zero values in $\mathcal{M}$, respectively. Similarly, for the local-level parameter fitting, it needs $O(\#iter \cdot (k + |\mathcal{M}|) \cdot dln)$ time to fit the parameters. Note that $\#iter$, k and $|\mathcal{M}|$ are small constant values that are negligible. Thus, the complexity is $O(dln)$. □

5. EXPERIMENTS

In this section we demonstrate the effectiveness of FUNNEL with real epidemic data. The experiments were designed to answer the following questions:

- Q1 *Sense-making*: Can our method help us understand the given input epidemic data?
- Q2 *Accuracy*: How well does our method match the data?
- Q3 *Scalability*: How does our method scale in terms of computational time?

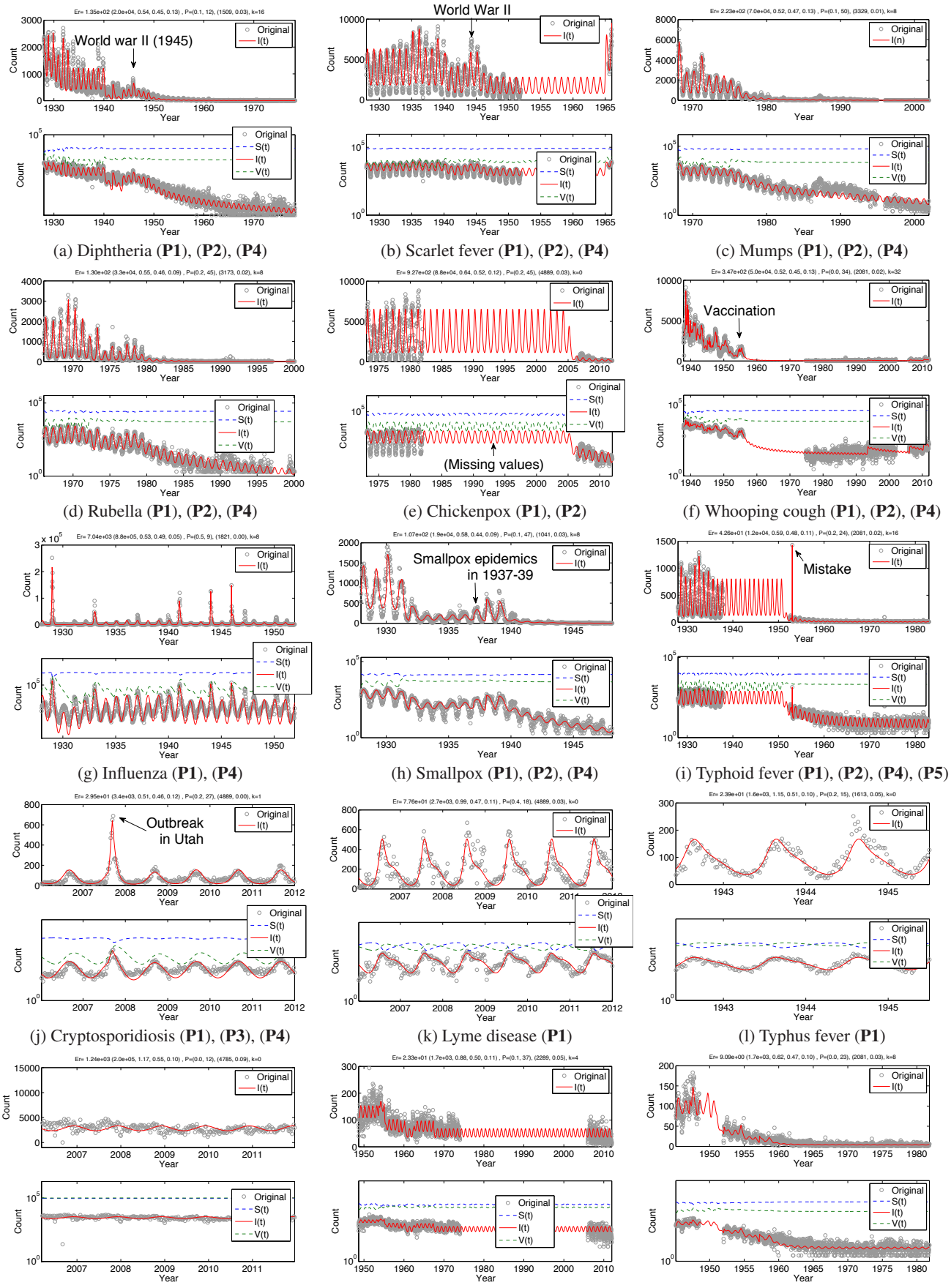
5.1 Matching co-evolving epidemic patterns

We demonstrate how effectively FUNNEL can learn important patterns given a large collection of epidemics. Figure 6 shows the results of model fitting on 15 typical diseases. We show the original sequences (i.e., black dots) and estimated sequences: $I(t)$ (i.e., red line) in linear-linear (top) and linear-log (bottom) scales. In the log-log scale, we also show the susceptible $S(t)$ and vigilant $V(t)$ counts. We made several important observations, which correspond to the five properties of the epidemic sequences.

(P1) Disease seasonality. As we have already seen in the introduction section (Figure 1(c)), we identified four categories i.e.,

- Influenza has very strong periodic spikes, in January-February.
- Children's diseases (e.g., measles, mumps, chickenpox) also have strong periodicity, but they peak in spring [27].
- Tick-borne diseases (e.g., Lyme disease), and cryptosporidiosis (i.e., water-borne disease) have strong periodicity, peaking in the summer, related to vector and human behavior and climate factors [28].
- Gonorrhea, i.e., sexually transmitted disease (STD) has no periodicity.

(P2) Disease reduction effects. FUNNEL is capable of automatically detecting the disease reduction impact. For example, in Figure 6

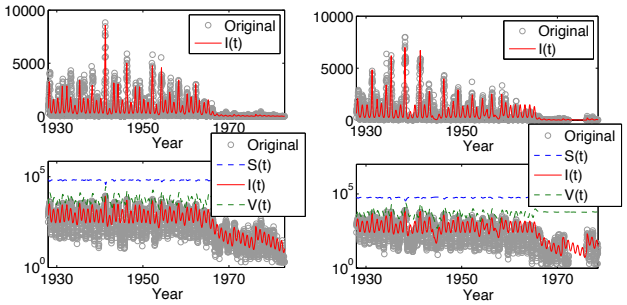


(a) Diphtheria (**P1**), (**P2**), (**P4**)
(b) Scarlet fever (**P1**), (**P2**), (**P4**)
(c) Mumps (**P1**), (**P2**), (**P4**)
(d) Rubella (**P1**), (**P2**), (**P4**)
(e) Chickenpox (**P1**), (**P2**)
(f) Whooping cough (**P1**), (**P2**), (**P4**)
(g) Influenza (**P1**), (**P4**)
(h) Smallpox (**P1**), (**P2**), (**P4**)
(i) Typhoid fever (**P1**), (**P2**), (**P4**), (**P5**)
(j) Cryptosporidiosis (**P1**), (**P3**), (**P4**)
(k) Lyme disease (**P1**)
(l) Typhus fever (**P1**)
(m) Gonorrhea
(n) Rabies in animals (**P1**), (**P2**), (**P4**)
(o) Brucellosis (**P1**), (**P2**), (**P4**)

Figure 6: Fitting results of FUNNEL for 15 diseases (global-level counts), shown in 'lin-lin'(top) and 'lin-log'(bottom) scales.

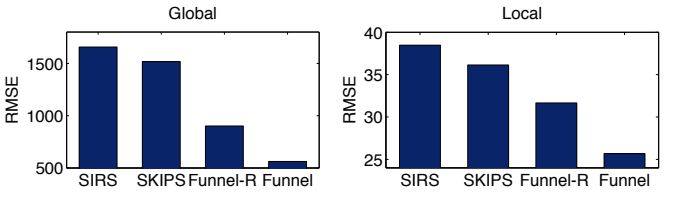
Figure 7: The year of vaccine licensure [32] vs. detection.

Disease	licensure	detected
Measles	1963	**1965**
Mumps	1967	**1975**
Whooping cough (pertussis)	1948	**1951**
Rubella	1969	**1972**



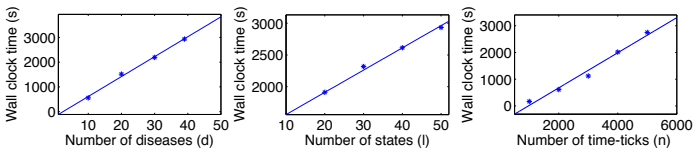
(a) New York State (NY) (b) Pennsylvania (PA)

Figure 8: Local-level fittings for measles: as with the global fitting shown in Figure 1, FUNNEL fits very well.



(a) Global fitting (b) Local fitting

Figure 9: Fitting accuracy for the (a) global sequences: $\{\bar{\boldsymbol{x}}_i(t)\}_{i,t}^{d,n}$ and (b) local sequences: $\{\boldsymbol{x}_{ij}(t)\}_{i,j,t}^{d,l,n}$. FUNNEL consistently outperforms the previous models w.r.t. ther RMSE between real and estimated values (lower is better).



(a) Diseases (d) (b) States (l) (c) Duration (n)

Figure 10: FUNNELFIT scales linearly: wall clock time vs. dataset size ($d \times l \times n$). Our method is linear on the data size.

(a-f), most children's diseases include the reduction patterns. Figure 7 shows the year the first vaccine was licensed vs. the year detected as the starting point of the disease reduction effect. Basically, there is a lag of 2-3 years between the licensure and the detected point, which could be related to the diffusion rate of a vaccination program. Note that FUNNEL does not detect any disease reduction effect for influenza; This is because influenza epidemics continue to occur due to high mutation rates in influenza virus that limit protective immunity.

(P3) Area specificity. FUNNEL can find the local dynamics of each disease, as well as the global-level pattern of epidemics. For example, as described in Figure 1 (b), there are many measles patients in NY and PA. Specifically, Figure 8 shows the original local-level epidemic sequences of measles in (a) NY and (b) PA, and the results of our fittings. FUNNEL successfully captures yearly-periodic patterns, the disease reduction effects, and, *also*, the local spikes of each location (i.e., notice that the strengths of the external shock effects differ state to state).

(P4) External shock events. Figure 6 shows that FUNNELFIT automatically detects some important external shock events, e.g.,

- (a) Diphtheria and (b) scarlet fever (i.e., children's diseases) have multiple external shocks, e.g., during World War II.
- (g) Flu has several major pandemics in 1929, 1941 etc., due to immune dynamics and antigenic changes in the virus [20].
- (h) Epidemics of a milder form of smallpox occurred in Northwestern states in 1937-39 due to low vaccination rates [5].
- (j) A major outbreak of Cryptosporidiosis, (i.e., water-borne disease), was detected in Utah due to contaminated public pools in 2007 [1].

(P5) Mistakes. One of the strong points of FUNNELFIT is its robustness against noise. It can handle "mistake" points as well as missing values (e.g., Figure 6 (i) typhoid fever, which contains missing values in the 1940s, and a mistake value in 1953).

5.2 Model quality and scalability

Next, we discuss the quality of FUNNEL in terms of fitting accuracy. We compared FUNNEL with the standard *SIRS* model and *SKIPS* [29]. To evaluate the effect of the disease reduction parameters, we also compared with FUNNEL-R, (i.e., removing external shocks and mistakes). Figure 9 (a) shows the root mean square error between the original and predicted counts of the global sequences $\{\bar{\boldsymbol{x}}_i(t)\}_{i,t}^{d,n}$. Similarly, Figure 9 (b) shows the results of the local counts $\{\boldsymbol{x}_{ij}(t)\}_{i,j,t}^{d,l,n}$, A lower value indicates a better fitting accuracy. Note that the *SIRS* model cannot capture seasonal dynamics, while *SKIPS* has the ability to capture periodic patterns. Moreover, they are not intended to detect disease reductions, or external shocks. As shown in the figures, the *SIRS* model and *SKIPS* failed to capture the complicated patterns of epidemics, while our method achieved high fitting accuracy.

We also evaluated the scalability of FUNNELFIT, and verified the complexity of our method, which we discussed in Lemma 1, in section 4. Figure 10 shows the computational cost of FUNNELFIT in terms of the dataset size. We varied the dataset size with respect to (a) diseases d, (b) states l, and (c) duration n. As shown in Figure 10, FUNNELFIT is linear with respect to data size.

6. DISCUSSION

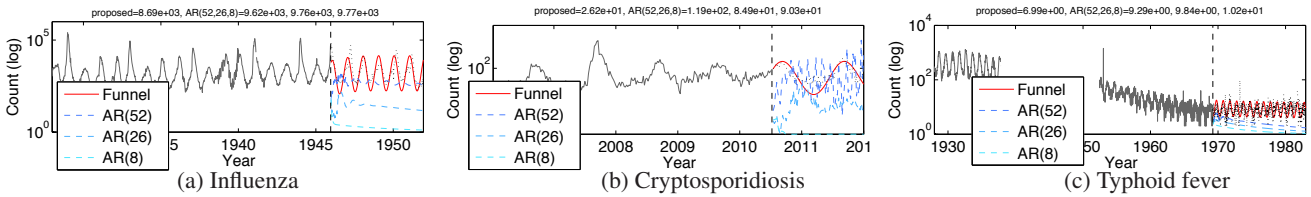
Here, we describe important applications and potential directions for our method.

FUNNEL at work - forecasting. Since FUNNEL has a very high fitting accuracy on real epidemic data, the most practical application would be forecasting. Figure 11 shows results of our forecasting in relation to three different diseases. We trained the model parameters by using the $2/3$ values for each sequence (solid black lines in the figure), and then forecasted the following years (solid red lines). Note that the vertical axis uses a logarithmic scale.

We compared FUNNEL with the auto regressive (AR) model, where we used the regression coefficients: $r = 52$ (i.e., one year), 26 (i.e., half year), and 8 (i.e., the same size as our base (6) + reduction (2) parameters). Also, since the original sequences are bursty we took their logarithm for the AR forecast.

Our method achieves high forecasting accuracy while AR failed. Specifically, for (a) influenza, and (b) cryptosporidiosis, FUNNEL captured the future trend correctly, while AR was strongly affected by multiple extreme spikes (e.g., in figure (b), there is a spike in 2007). Similarly, for (c) typhoid fever, which includes the disease reduction pattern, missing/mistake values, FUNNEL successfully forecasted the periodic patterns after the disease reduction effect.

Generality - epidemics on computer networks. Another promising step for FUNNEL would be its generalization to other domains such as modeling computer viruses. Computer viruses have similar characteristics to biological viruses [11]. For example, Figure 12 shows the fitting result of FUNNEL on publicly available reports



(a) Influenza (b) Cryptosporidiosis (c) Typhoid fever

Figure 11: Forecasting result: we train the model parameters using $2/3$ of each sequence (i.e., solid black lines). We then start forecasting (at the vertical dotted line). Note that the vertical axis uses a logarithmic scale.

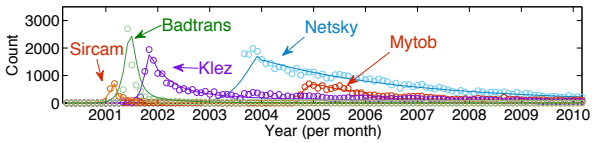


Figure 12: FUNNEL is general: our model (solid lines) fits computer virus data (in circles) very well. It captures rising spikes and viruse reduction effects (i.e., anti-virus software). Note that it shows the reported case count for each virus.

published by IPA, [9] which consists of annual reports on computer virus infections in Japan (e.g., private companies and public schools) covering more than ten years. The figure shows the original counts for the top five viruses (circles) and our fittings (solid lines) from 2000 to 2010. Here are some interesting observations: (a) "Badtrans" and "Klez" spread in 2001-2002 by exploiting a security hole in Microsoft Outlook. It spread very quickly due to the strong infection effect, but it also decayed very quick thanks to anti-virus software (i.e., the virus reduction effect); (b) "Netsky", which spreads via email attachment: there were huge infections in 2004, and it gradually decreased over the next 10 years, but still remains. Also, there is a weekly periodicity (i.e., less infection at weekends); (c) "Mytob", which spreads through corporate networks: there are some local-level (i.e., intra-office) infections. Very recently, there have been several worms (e.g., "Koobface", "Fbphotofake") that spread quickly through social networking sites including Facebook and Twitter, where they have a high potential population (# of users).

7. CONCLUSIONS

Our proposed method has the following appealing advantages:

1. It is **sense-making**: FUNNEL captures all essential aspects, i.e., yearly periodicity, discontinuities, local sensitivities. It can lead to disease clustering, find disease reduction effects (e.g., vaccines) and external shocks, and perform forecasting.
2. It is **automatic**: FUNNELFIT requires no training set and no hint regarding the number of parameters. Thanks to our coding scheme, it determines all of the above automatically.
3. It is **scalable**: FUNNELFIT scales very well, being linear on the database size, (i.e., $O(dln)$).
4. It is **general**: We demonstrated the generality of FUNNEL, by applying it to real epidemic datasets, including computer virus infections, as well as human diseases.

Acknowledgement. We thank Dr. Donald S. Burke, Dean of the Graduate School of Public Health, University of Pittsburgh for his support and expert opinion during this study. This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research Number 24500138, 26730060, 26280112, 25·7946. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1314632. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] Promotion of healthy swimming after a statewide outbreak of cryptosporidiosis associated with recreational water venues–utah, 2008-2009. *MMWR Morb Mortal Wkly Rep*, 61(19):348–52, 2012.

[9] IPA - IT security center: https://www.ipa.go.jp/security/english/index.html

[2] R. M. Anderson and R. M. May. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, 1992.
[3] C. Böhm, C. Faloutsos, J.-Y. Pan, and C. Plant. Ric: Parameter-free noise-robust clustering. *TKDD*, 1(3), 2007.
[4] G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1994.
[5] D. CC. Smallpox in the united states: It's decline and geographic distribution. *Public Health Reports*, 55(50):2303–2312, 1940.
[6] L. Chen and R. T. Ng. On the marriage of lp-norms and edit distance. In *VLDB*, pages 792–803, 2004.
[7] I. N. Davidson, S. Gilpin, O. T. Carmichael, and P. B. Walker. Network discovery via constrained tensor analysis of fmri data. In *KDD*, pages 194–202, 2013.
[8] D. J. Earn, P. Rohani, B. M. Bolker, and B. T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667–70, 2000.
[9] B. T. Grenfell, O. N. Bjornstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414:716, 2001.
[10] A. Jain, E. Y. Chang, and Y.-F. Wang. Adaptive stream resource management using kalman filters. In *SIGMOD*, pages 11–22, 2004.
[11] J. Kephart and S. White. Directed-graph epidemiological models of computer viruses. In *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*, pages 343–359, May 1991.
[12] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *KDD*, pages 553–562, 2010.
[13] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *SIGMOD*, pages 593–604, 2007.
[14] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, pages 462–470, 2008.
[15] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious linear fingerprinting for time series. *PVLDB*, 3(1):385–396, 2010.
[16] Y. Matsubara, L. Li, E. E. Papalexakis, D. Lo, Y. Sakurai, and C. Faloutsos. F-trail: Finding patterns in taxi trajectories. In *PAKDD (1)*, pages 86–98, 2013.
[17] Y. Matsubara, Y. Sakurai, and C. Faloutsos. Autoplait: Automatic mining of co-evolving time sequences. In *SIGMOD*, 2014.
[18] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa. Fast mining and forecasting of complex time-stamped events. In *KDD*, pages 271–279, 2012.
[19] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *KDD*, pages 6–14, 2012.
[20] F. NM, G. AP, and B. RM. Ecological and immunological determinants of influenza evolution. *Nature*, 422(6930):428–33, 2003.
[21] S. Papadimitriou and P. S. Yu. Optimal multi-scale patterns in time series streams. In *SIGMOD Conference*, pages 647–658, 2006.
[22] F. PE and C. JA. Measles in england and wales–i: An analysis of factors underlying seasonal patterns. *Epidemiol*, 11(1):5–14, 1982.
[23] B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos. Winner takes all: competing viruses or ideas on fair-play networks. In *WWW*, pages 1037–1046, 2012.
[24] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos. Threshold conditions for arbitrary cascade models on arbitrary networks. In *ICDM*, pages 537–546, 2011.
[25] T. Rakthanmanon, B. J. L. Campana, A. Mueen, G. E. A. P. A. Batista, M. B. Westover, Q. Zhu, J. Zakaria, and E. J. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *KDD*, pages 262–270, 2012.
[26] Y. Sakurai, S. Papadimitriou, and C. Faloutsos. Braid: Stream mining through group lag correlations. In *SIGMOD*, pages 599–610, 2005.
[27] D. SF. Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerg Infect Dis.*, 7(3):369–74, 2001.
[28] M. SM, E. RJ, M. A, and M. P. Seasonality in six enterically transmitted diseases and ambient temperature. *Am J Trop Med Hyg.*, 2014.
[29] L. Stone, R. Olinky, and A. Huppert. Seasonal dynamics of recurrent epidemics. *Nature*, 446:533–536, March 2007.
[30] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *KDD*, pages 374–383, 2006.
[31] Y. Tao, C. Faloutsos, D. Papadias, and B. Liu. Prediction and indexing of moving objects with unknown motion patterns. In *SIGMOD*, pages 611–622, 2004.
[32] W. G. van Panhuis, J. Grefenstette, S. Y. Jung, N. S. Chok, A. Cross, H. Eng, B. Y. Lee, V. Zadorozhny, S. Brown, D. Cummings, and D. S. Burke. Contagious diseases in the united states from 1888 to the present. *NEJM*, 369(22):2152–2158, 2013.
[33] M. Vlachos, D. Gunopulos, and G. Kollios. Discovering similar multidimensional trajectories. In *ICDE*, pages 673–684, 2002.