FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

By Yasuo Matsubara, Yasushi Sakurai, Willem G. van Panhuis, and Christos Faloutsos
SIGKDD 2014

Presented by Sarunya Pumma
This presentation has been modified from a presentation slide from SIGKDD 2014

Link to the presentation: www.cs.kumamoto-u.ac.jp/~yasushi/publications/funnel.pptx
Outline

• Introduction & motivation
• Problem definition
• Epidemic modeling using FUNNEL
  • With a single epidemic
  • With multi-evolving epidemics
• Optimization algorithm
• Experiments and results
• Conclusion
Introduction & Motivation
Introduction & Motivation

Given a large set of epidemiological data

Measles cases in the U.S. (number of infectious per week)

How to find a model that can represent it efficiently?
Introduction & Motivation

Considering some existing models...

The SI model

• Pros
  • Use a few number of parameters
  • Can capture the dynamics of epidemiological data

• Cons
  • Cannot represent periodic patterns
  • Cannot use for predicting a future trend

Recall a general SIR model

Introduction & Motivation

Considering some existing models…

The auto regression (AR) model
• Pros
  • Use a few number of parameters
• Cons
  • Cannot model the non-linear patterns (i.e. virus propagation)

Non-linear pattern

Introduction & Motivation

Considering some existing models...

The PARAFAC model (a generalization of PCA)

- Pros
  - Use a few number of parameters
- Cons
  - Cannot handle missing values

A graphical representation of a two-component PARAFAC model of the data array X

Introduction & Motivation

They proposed **FUNNEL**
- A model for describing non-linear dynamics in coevolving epidemics

Properties
- **Sense-making**: it can describe the seasonality of diseases, vaccination, and external shocks
- **Fully automatic**: it is parameter-free and be able to handle missing/mistake values. No human interaction required
- **Scalable**: it scales linearly with the input size
- **Generality**: it can be applied to many types of epidemic data
Problem Definition
Project Tycho provides infectious diseases data in the U.S.

\[ x \] is number of cases (infectious)  
\[ \text{e.g. ('measles', 'NY', 'April 1-7, 1931', '4000')} \]
Problem Definition

**Given:** Tensor $\mathbf{X}$ (disease x state x time)

**Find:** Compact representation of $\mathbf{X}$

$\mathbf{X} = \text{FUNNEL} = \{P_1, P_2, P_3, P_4, P_5\}$

$P_i$ is property $i$
Problem Definition

Important properties of epidemics

• P1 - yearly periodicity
• P2 - disease reduction effects
• P3 - area specificity and sensitivity
• P4 - external shock events
• P5 - mistakes, incorrect values
Problem Definition

P1 - Yearly periodicity
Radius: seasonality strength
Angle: peak season
Problem Definition

P1 - Yearly periodicity
Radius: seasonality strength
Angle: peak season

Influenza (peak in cold season)
Problem Definition

**P1 - Yearly periodicity**

Radius: seasonality strength
Angle: peak season

Measles (peak in spring)
Problem Definition

**P1 - Yearly periodicity**

Radius: seasonality strength  
Angle: peak season

Lyme-disease (peak in summer)
Problem Definition

P1 - Yearly periodicity

Radius: seasonality strength
Angle: peak season

Gonorrhea (STD) (no peak season)
Problem Definition

P2 - Disease reduction effects

Vaccine licensure in 1963
Problem Definition

**P3 - Area specificity and sensitivity**

Potential population of susceptible for measles
Problem Definition

P4 - External shock events

Large outbreaks of measles
Problem Definition

P5 - Mistakes, incorrect values
Epidemic Modeling using FUNNEL
FUNNEL with a Single Epidemic

FUNNEL-BASE

\[
S(t + 1) = S(t) - \beta(t)S(t)I(t) + \gamma V(t)
\]
\[
I(t + 1) = I(t) + \beta(t)S(t)I(t) - \delta I(t)
\]
\[
V(t + 1) = V(t) + \delta I(t) - \gamma V(t)
\]

where

\[
\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot \cos\left(\frac{2\pi}{P_p}(t + P_s)\right)\right)
\]

Amplitude of the fluctuation
Phase shift of the seasonal cycle
Average of rate $\beta$ over the year
Number of weeks in a year

\[
b = \{N, \beta_0, \delta, \gamma, P_a, P_s\}
\]
FUNNEL with a Single Epidemic

**FUNNEL-R** (with disease reduction)

\[
S(t+1) = S(t) - \beta(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)
\]

\[
I(t+1) = I(t) + \beta(t)S(t)I(t) - \delta I(t)
\]

\[
V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)
\]

where

\[
\theta(t) = \begin{cases} 
0 & (t < t_\theta) \\
\theta_0 & (t \geq t_\theta) 
\end{cases}
\]

\[
r = \{t_\theta, \theta_0\}
\]
FUNNEL with a Single Epidemic

**FUNNEL-RE** (with disease reduction and external shocks)

\[
S(t + 1) = S(t) - \beta(t) \varepsilon(t) S(t) I(t) + \gamma V(t) - \theta(t) S(t)
\]
\[
I(t + 1) = I(t) + \beta(t) \varepsilon(t) S(t) I(t) - \delta I(t)
\]
\[
V(t + 1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t) S(t)
\]

where

- **Total number of events**
- **Strength of the external shock**
- **Duration of the event**
- **Central time point of the external shock event**

**Temporal susceptible rate**

\[
\varepsilon(t) = \sum_{k} \epsilon(t_k) \delta(T)
\]

\[
f(t; \varepsilon(T)) = \begin{cases} 
\epsilon_0 & (t_\mu - t_\sigma < t < t_\mu + t_\sigma) \\
0 & (\text{else})
\end{cases}
\]

\[
\varepsilon(T) = \{t_\mu, t_\sigma, \epsilon_0\}
\]
FUNNEL with Multi-Evolving Epidemics

**Find:** Compact representation of $X$

**FUNNEL** $F = \{B, R, N, E, M\}$

---

**Seasonality**

$P_1$  
$b_i = \{N_i, \beta_0i, \delta_i, \gamma_i, P_{ai}, P_{si}\}, b_i \in B$

**Reduction**

$P_2$  
$r_i = \{t_{\theta i}, \theta_{0i}\}, r_i \in R$

$i$ is a disease ID
FUNNEL with Multi-Evolving Epidemics

**Find:** Compact representation of $\mathcal{X}$

**FUNNEL** $F = \{B, R, N, E, M\}$

$$N_{ij} = S_{ij}(t) + I_{ij}(t) + V_{ij}(t)$$

$i$ is a disease ID

$j$ is a state ID
FUNNEL with Multi-Evolving Epidemics

**Find:** Compact representation of $\mathbf{X}$

\[
\text{FUNNEL } \mathbf{F} = \{ \mathbf{B}, \mathbf{R}, \mathbf{N}, \mathbf{E}, \mathbf{M} \}
\]

Recall: temporal susceptible rate

\[
\epsilon(t) = 1 + \sum_{i=1}^{k} f(t; e_{i}^{(T)})
\]

We replace $e^{(T)}$ in the full model with $e^{(S)}$ to consider $\ell$ individual location
**Find:** Compact representation of $\mathbf{X}$

**FUNNEL $\mathbf{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathbf{E}, \mathbf{M}\}$**

Mistake values

$\mathbf{M}$ is a 3rd-order tensor of mistake data points
(a very sparse tensor)
Optimization Algorithm
Optimization Algorithm

- The optimization follows the minimum description length (MDL) principle
- **Objective**: minimize the size of data by compressing it
- **Objective function**: cost of model description (spacewise)

\[Cost_T(\mathcal{X}; \mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n) + Cost_M(\mathcal{B}) + Cost_M(\mathcal{R}) + Cost_M(\mathcal{N}) + Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X} | \mathcal{F})\]

- **Data compression**: Huffman coding with a Gaussian distribution is adopted
Optimization Algorithm

- Multi-layer optimization
  - Global-fit is $F_G = \{B, R, E^D, E^T, M^D, M^T\}$
  - Local-fit is $F_L = \{N, E^S, M^S\}$

- Levenberg-Marquardt (LM) is adopted to minimize the cost function
Experiments and Results
Mumps (Seasonality Disease)

(c) Mumps (P1), (P2), (P4)
Local-Level Fittings for Measles

(a) New York State (NY)  
(b) Pennsylvania (PA)
External-Shock Effects

World War II (1945)

(a) Diphtheria (P1), (P2), (P4)

(b) Scarlet fever (P1), (P2), (P4)
Missing Data and Incorrect Data Point Effects

(i) Typhoid fever (P1), (P2), (P4), (P5)
Model Accuracy

(a) Global fitting

(b) Local fitting
Model Scalability

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

Train data: 2/3
Forecast: 1/3
Apply to Other Applications

Computer virus outbreaks!
Conclusion
Conclusion

• FUNNEL is an efficient model for representing epidemic dynamics
  • It can capture yearly-periodicity, epidemic reduction, and local sensitivity efficiently
  • It is fully automatic and highly scalable
  • It can handle missing/wrong data effectively
  • It is general which can be applied to various domains
Thank you for your attention

http://www.sltrib.com/opinion/2117762-155/bagley-cartoon-measles-outbreak
Dynamic Poisson Autoregression for Influenza-Like-Illness Case Count Prediction

By Zheng Wang, Prithwish Chakrabortty, Sumiko R. Mekaru, John S. Brownstein, Jieping Ye, Naren Ramakrishnan
KDD 2015

Presented by Sarunya Pumma
Outline

• Introduction & motivation
• Case count forecasting
  • ARX model for time series prediction
  • Dynamic ARX model
  • Dynamic Poisson ARX model
• Experiments and results
• Conclusion
Introduction & Motivation
Introduction & Motivation

- Influenza-like-illness (ILI) is the most common disease

Goal: predicting characteristics of ILI

- Seasonal characteristics
  - Overall shape of ILI counts for a specific season (e.g., peak value, peak size)
- Short-term characteristics
  - Absolute values of next data points

Example of Influenza-like-illness case count

Picture reference: http://publichealth.jmir.org/2015/1/e5/
Introduction & Motivation

• Existing works
  • SEIR and SIRS
  • Curve simulations with network-based models
  • Filtering-based methods

• Existing indicator sources
  • Online indicator (Google Flu Trend, Wikipedia, Twitter, etc.)
  • Weather

• Existing works used the **same model** to predict influenza for the entire season
  • Characteristics of influenza are different in the peak-season and off-season
Introduction & Motivation

• Proposed work
  • A time series model for predicting short-term characteristics of ILI by using external factors from indicator sources

• Overview of the model
  • Base model is autoregressive exogenous model (ARX)
  • Models are built separately for each time point
    • To model activities of the disease that evolve over time
    • Models share common characteristics
  • Block coordinate descent is used to fit the model
Case Count Forecasting
ARX Model for Time Series Prediction

• General ARX model

\[ y_t = F(y_{t-1}, y_{t-2}, y_{t-3}, \ldots, x_{t-1}, x_{t-2}, x_{t-3}) + \epsilon_t \]
ARX Model for Time Series Prediction

• General ARX model

\[ y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{b} \beta_i^T x_{t-d-i} + \varepsilon_t + c, \]

where \( p \) and \( b \) are previous time steps.
ARX Model for Time Series Prediction

Compact form

\[ z_t = [x_{t-d}, \cdots, x_{t-d-b}, y_{t-1}, \cdots, y_{t-p}, 1] \]
\[ y_t = wz_t^T + \varepsilon_t \]

where

\[ w = [\beta_{t-d}, \cdots, \beta_{t-d-b}, \alpha_{t-1}, \cdots, \alpha_{t-p}, c] \]

We would like to minimize the least square loss

\[ \min_w \sum_{t} l(z_t, y_t) = \sum_{t} (y_t - wz_t^T)^2, \]
Dynamic ARX Model

We build a model for each time point.

Model similarity graph

Node is the model \( (w_t) \) at each time point

Edge is a similarity between 2 models
Dynamic ARX Model

Model similarity graph with different prior knowledge

- Fully connected
- nearest neighbors
- seasonal nearest neighbors
Dynamic ARX Model

Adding a similarity measure (Euclidean distance) to the ARX model

$$\min_{\mathbf{w}} l(z_t, y_t) + \eta \sum_{i,j} S_{i,j} \| \mathbf{w}_i - \mathbf{w}_j \|_2^2$$

where, \( \mathbf{w}_0 = 0 \)

Solve a regression problem by using the block coordinate descent optimization
Dynamic Poisson ARX Model

Since we predict the infected people at different time points, we formulate the problem as Poisson regression

\[ Pr(y) = \frac{\lambda^y e^{-\lambda}}{y!}, \]

Likelihood function

\[ Pr(\{y_i\}|\{z_i\}) = \prod_{i=1}^{N} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \]

\[ \Rightarrow \log(Pr(\{y_i\}|\{z_i\})) = \sum_{i=1}^{N} (y_i \log(\lambda) - \lambda - \log(y_i!)) \]

From the link function

\[ \log(\mathbb{E}[y|z]) = \log(\lambda) = wz^T \]

we get

\[ \lambda = e^{zw^T} \]

\[ \sum_{i=1}^{N} \left( y_i (wz_i^T) - e^{(wz_i^T)} - \log(y_i!) \right) \]
Dynamic Poisson ARX Model

We maximize the log-likelihood to train weights

$$\max_w \sum_{i=1}^N \left( y_i (wz_i^T) - e^{(wz_i^T)} \right)$$

We can get \( \lambda = (zw^T) \) from the identity link function

Then we get our final model

$$l(X, y, W) = - \sum_{i=1}^N \left( y_i \log(wz_i^T) - (wz_i^T) - \log(y_i!) \right)$$
Experiment and Results
Datasets

• Reference source: PAHO/CDC
  • US and 14 Latin American countries
  • Weekly count
  • From Jan 1, 2010 to Oct 8, 2014

• Physical indicators
  • 5 weather attributes (absolute/relative/specific humidity, precipitation, and temperature)

• Social indicators
  • Google Flu Trends
  • Google Search Trends
  • HealthMap
Evaluation Measures

- Prediction error

$$error = \frac{4}{N} \sum_{t=t_s}^{t_e} \frac{|y_t - \hat{y}_t|}{\max(y_t, \hat{y}_t, 10)},$$

- Prediction Score is 4 - error
  (a perfect score is 4)
Model Similarity
Model Similarity

- Fully connected
- nearest neighbors
- seasonal nearest neighbors
Forecasting Results

• Training data
  • First 50 time points

• Time step settings
  • \( p = 1 \) (target variables) and \( b = 15 \) (exogenous params)

• Result
  • DPARX can yield better prediction accuracy than others
Seasonal Analysis

- They calculated the short-term predictions in multiple steps and fitted the curve to get seasonal prediction.
Conclusion
Conclusion

- DPARX can predict a number ILI cases at a desire time point
- DPARX extends the original ARX by having a dynamic term based on the similarity between models
- DPARX can also efficiently reflect the seasonal characteristics of ILI
Thank you for your attention