Big Data Analytics for Societal Event Forecasting

Liang Zhao (George Mason University), Feng Chen (University at Albany, SUNY)

IEEE Conference on Big Data 2018

Dec 12, 2018
Introduction and motivation
Part 1: Temporal Event Forecasting
Part 2: Spatio-temporal Event Forecasting
Conclusion and Future Directions
What are societal events?

- Terrorism events in Africa
- Epidemics outbreak on Week 47 ending Nov 22, 2014 in southern region
- Traffic congestions
What are societal events?

Civil unrest events on Mar 17, 2013 in Brazil

Economics crisis

Earthquake events
Societal Events

- Riots
- Terrorism
- Epidemics
- Traffic
- Congestions
- Earthquake
- Boycotts
- Pandemics
- Floods
- Crimes
- Protests
- Crisis
- Strikes
- Economic
- Storm
Civil unrest

Complaining, July 1, 2012~July 3, 2012
- Fair Election? @on #EPN seem to keep tricking to FRAUD the election...or?
- Those votes will not be counted only way to ensure a fair and honest election
- This is just beginning...this has not been a democratic election

- Never let anyone stop you from speaking out against injustice. It is okay if others disagree with you. Stand up for what you believe in.
- Today #auro do not hesitate to raise your voice in protest against injustice. Who comes to march?

Dissemination, July 5, 2012~July 6, 2012
- Let us go to the Mega March on July 7 to Zócalo Angel on 3:00 pm. Spread the word!
Societal Events are Forecastable

- Transportation congestion

### Practitioners want much more than just prediction results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Trigger event type?</td>
<td>detoured traffic</td>
<td>detoured traffic</td>
<td>fire on road</td>
<td>roadblock ahead</td>
</tr>
<tr>
<td>2. Trigger event?</td>
<td>Event 2</td>
<td>Event 1</td>
<td>fires accident</td>
<td>Event 1</td>
</tr>
</tbody>
</table>

[Map of Atlanta with highlighted areas and roads]
Societal Events are Forecastable

Epidemics

Social Media Space

I feel like crap! I'm getting the flu

Just got my flu shot from @rxjude ouch!!

Maybe I will feel better tomorrow and can go back to school Thursday! :D hate the flu

I think I got flu

get the flu, in bed for 3 days.

I just realized that I lost my ability to taste. This flu is just trolling me now

4th day with flu

Real-world Space

: infected for n days

: isolated

: vaccinated

: uninfected
• Given some indicators, the task of societal event forecasting is to predict the time, location, and topic of a thing occurring in the future with significant social impact.

• Underlying mechanism of societal events
  – Complex
  – Hard to comprehensively model
  – Largely unknown

Data-driven model as surrogate, thanks to Big Data!

Build the forecaster driven by large historical data
Lead Time

- \( t_1 \) Forecast Date
- \( t_2 \) Event Date
- \( t_3 \) Predicted Event Date
- \( t_4 \) Reported Date

Date Quality

Lead Time
Examples of Social Indicators

The GDELT Project
Characteristics of Social Indicators in Big data Era

• Ubiquitousness
  – Every user/agent of social media/web/forum is a social sensor.
  – They are everywhere monitoring the world all the time.

• Timeliness
  – 6,000 tweets every second.
  – 500 million tweets per day.
  – Usually beat the earliest official reports.

• Indicative and predictive signals

Complaints toward Trump on Change.org
Challenges in Societal Event Forecasting

1. Dynamics
new #hashtags, abbreviations, new words

2. Multiple resolution
many messages with country info, few with coordinates

3. Data incompleteness
Reddits enable geo-info this year

4. Big Data Paradox
many data in total, few data for each user

5. Noisy
typos, chit-chat, rumors
Challenges in Societal Event Forecasting

7. Multilingual, multi-modal

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Tweets</th>
<th>SPA (%)</th>
<th>ENG (%)</th>
<th>POR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>160,564,890</td>
<td>91.6</td>
<td>7.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Brazil</td>
<td>185,286,958</td>
<td>10.1</td>
<td>16.0</td>
<td>73.9</td>
</tr>
<tr>
<td>Chile</td>
<td>97,781,414</td>
<td>82.8</td>
<td>16.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Colombia</td>
<td>158,332,002</td>
<td>89.8</td>
<td>9.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Ecuador</td>
<td>50,289,195</td>
<td>91.1</td>
<td>8.1</td>
<td>0.8</td>
</tr>
<tr>
<td>El Salvador</td>
<td>21,992,962</td>
<td>91.5</td>
<td>7.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Mexico</td>
<td>197,550,208</td>
<td>83.7</td>
<td>15.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Paraguay</td>
<td>30,891,602</td>
<td>92.2</td>
<td>6.4</td>
<td>1.4</td>
</tr>
<tr>
<td>Uruguay</td>
<td>10,310,514</td>
<td>89.7</td>
<td>8.8</td>
<td>1.4</td>
</tr>
<tr>
<td>Venezuela</td>
<td>167,411,358</td>
<td>92.3</td>
<td>6.9</td>
<td>0.8</td>
</tr>
</tbody>
</table>

8. Heterogeneous network

9. Sparsity in high-dimensional features
Numerous features of vocabulary and profile few are of interest for the research task
Other challenges

• Dependencies among events, e.g., spatial dependencies
• Uncertainty
• Lack of label data, Cannot afford to label massive data

• Model interpretability – societal event is influential
• Lack Mechanism Models
Comparisons with Event Detection

Event detection

– Historical or Ongoing events
– Discover anomaly
– Model types
  – Unsupervised learning
– Relevant techniques
  – Anomaly detection
  – Outlier detection
  – Change detection
  – Motif discovery

Event forecasting

– Future events
– Discover the mapping
– Model types
  – Supervised learning
  – Self-supervised learning
  – Semi-supervised learning
– Relevant techniques
  – Autoregressive
  – Markov chain
  – Classification
  – Causal inference
Comparisons with Spatial Prediction

Prediction v.s. Forecasting:
- “Forecasting”: Must be variable in the future.
- “Prediction”: Not necessarily variable in the future.

• Spatial Prediction
  – Dependent variable
    – No need be in the future
    – Usually continuous values – “index”
  – Must have spatial dimension

• Event Forecasting
  – Dependent variable
    – Must be in the future
    – Usually discrete values – “event”
  – No need be in spatial dimension
**Taxonomy**

Societal event forecasting

Temporal event forecasting

Spatiotemporal event forecasting

Discriminative Models

- Multi-task
- Multi-modal
- Multi-level

Generative Models

- Bayesian Networks
- Deep Generative

Ensemble models

- Computational
- Data-driven
Big Data Analytics for Societal Event Forecasting

Temporal event forecasting

Spatio-temporal event forecasting

Casual dependency mining

Temporal dependency mining

Anomaly mining

Predefined causality

Optimized causality

Markov decision processes

Deep Neural Networks

Scan Statistics based

Distance based
Probabilistic real-time alert system

- Historical News Archives
- Chains Extractor
- Chains
- Feature Enhancer
- Feature Generalization
- Examples Labeled Pairs
- Learner

World Knowledge Ontologies

- Probability of target event at time $t' > t$
- Event Generalization
- Predictor
- Events $\leq$ time $t$
- Target Event
Probabilistic real-time alert system

Probability of target event at time $t' > t$
# Event Chains

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 16, 1992</td>
<td>Jury in Shooting by Officer Hears Conflicting Accounts</td>
</tr>
<tr>
<td>Feb 11, 1992</td>
<td>Closing Arguments Conflict on Killing by Teaneck Officer</td>
</tr>
<tr>
<td>Feb 12, 1992</td>
<td>Officer Acquitted in Teaneck Killing</td>
</tr>
<tr>
<td>Feb 13, 1992</td>
<td>Acquitted Officer Expresses Only Relief, Not Joy</td>
</tr>
<tr>
<td>Feb 16, 1992</td>
<td>250 March in Rain to Protest Teaneck Verdict</td>
</tr>
</tbody>
</table>
Event Chains

Cluster documents with similar text (using bag of words similarity)

Improve Precision:
Greedily optimize Story Entropy (entropy in its entities) to grow “slowly”
Probabilistic real-time alert system

- Historical News Archives
  - Chains Extractor
  - Chains
  - Feature Enhancer
  - Feature Generalization
  - World Knowledge Ontologies
  - Event Generalization
  - Probability of target event at time $t' > t$
  - Examples Labeled Pairs
  - Learner
  - Predictor
  - Events $\leq$ time $t$
  - Target Event
Cuba faces its worst drought for 50 years

Cuba is facing its worst drought in half a century, with tens of thousands of families almost entirely reliant on water trucks for essential supplies.

The drought started two years ago, and reservoirs are now down to a fifth of their normal levels.

The government is providing road deliveries of water to more than 100,000 people in the worst affected areas of the capital, Havana.

The BBC's Michael Vos is asked people in Havana how they were coping.

Tropical Storm Isaac drenches Haiti, swipes Cuba

(Reuters) - Tropical Storm Isaac dumped torrential rains on Haiti and flattened tent camps housing survivors of a devastating earthquake, then began an assault on eastern Cuba on Saturday.

Isaac killed at least four people in Haiti and was expected to strengthen into a hurricane before hitting the Florida Keys on Sunday and crossing into the Gulf of Mexico.

Fueled by warm Gulf waters, it was forecast to strengthen into a Category 2 hurricane with 100-mph (160-kph) winds and hit the U.S. coast somewhere between the Florida Panhandle and New Orleans at midweek.
P(Cholera in Havana|Cuba) 

Never appeared in the news archive...
\[
P(Cholera \text{ in Havana} | \text{Cuba},
\begin{array}{l}
\text{AreaTotal} : 109884.0,
\text{PopulationDensity} : 102.3,
\text{GdpNominalPerCapita} : 5100.0,
\text{PercentWater} : \text{negligible}
\end{array})
\]
(Cholera in Havana
Cuba,
Social States,
Island Countries)
Abstraction Process

Time

06/17/2010 11:00AM + (2h)

First Event

Bangladesh

“Bangladesh floods: 1000s homeless”

07/27/2011 11:15AM + (3h)

Following Event

Dhaka

“Dhaka's Cholera wars”
Big Data Analytics for Societal Event Forecasting

Temporal event forecasting

- Casual dependency mining
  - Predefined causality
  - Optimized causality
  - Markov decision processes

- Temporal dependency mining
  - Deep Neural Networks

Spatio-temporal event forecasting

- Anomaly mining
  - Statistics
  - Distance based
Why Can We Detect & Forecast Events from Social Media?

- Event = Large-scale population behavior
- Social media is a real-time “sensor” of large population behavior
- Event Detection vs. Forecasting
  - Sense of public discussions about ongoing events vs. trigger events using social media
People are dying from hantavirus in Osorno hydroelectric government workers do not report Camila I beg help @ camila_vallejo

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue

Confirmed: Young man dies in Pencahue Hanta: This is a 26-year residence in the commune of http://t.co/5lkD0CZDmf

RT @ RADIOPALOMAFM: ISP confirmed case of hantavirus nvo rural sector in Linares. Woman, 38, who died May 11 at the UCI via @ SeremiSaludM
Elephant And The Blind Men

- It's a Fan!
- It's a Spear!
- It's a Wall!
- It's a Rope!
- It's a Snake!
- It's a Tree!
“#VIRUSHANTA” mentioned 100 times

Keyword “Hantavirus” Mentioned 90 times

Araucania State has 15 active users and 100 tweets

Influential User “SeremiSaludM” (1497 followers) posted 8 tweets

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue re-tweeted 50 times

Confirmed: Young man dies from hantavirus in Pencahue

http://t.co/5RD0CZBDm mentioned 10 times
Influential User “SeremiSaludM” (1497 followers) posted 8 tweets

Hantavirus Disease Outbreak

Keyword “Hantavirus” mentioned 90 times

Araucanía State has 15 active users and 100 tweets

Keyword “#VIRUSHANTA” mentioned 100 times

Hashtag

Tweet

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue

re-tweeted 50 times

Link

http://t.co/5lkD0CZDmf mentioned 10 times

User

Location

Keyword

“Hantavirus” Disease Outbreak

http://t.co/5lkD0CZDmf

Influential User “SeremiSaludM” (1497 followers) posted 8 tweets
Twitter Heterogeneous Network

Co-occurrence

Term

Geographic Neighbors

Location

Geographical Relationship

Tweet

Geographical Relationship

User Mention

Owner

Hashtag

Part of

Part of

Part of

Part of

Link

Followers
Twitter Heterogeneous Network (Example)

RT People are dying from hantavirus in Osorno hydroelectric government workers do not report Camila I beg help @ camila_vallejo

Hantavirus outbreak in Osorno? 1 dead 2 Serious and more, it involved a lawyer hydrogen q wants to help the family? #Valdiviocl

@ juanjosellanten @ meganoticiascl Virus kills Jorge Vasquez have now moved to Palm Jc evicted more contagious mutual

In observation ? N man suspected of having Hanta virus in Osorno is http://t.co/xey8hbJ4

RT @ rioenlinea [? WHAT LAST ] confirmed case of Hantavirus in children or Malalhue remains severe, life-threatening

http://t.co/xey8hbJ4
## Node Attributes

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td># tweets, # retweets, # followers, #followees, #mentioned_by, #replied_by, diffusion graph depth, diffusion graph size</td>
</tr>
<tr>
<td>Tweet</td>
<td>Klout, sentiment, replied_by_graph_size, reply_graph_size, retweet_graph_size, retweet_graph_depth</td>
</tr>
<tr>
<td>City, State, Country</td>
<td># tweets, # active users</td>
</tr>
<tr>
<td>Term</td>
<td># tweets</td>
</tr>
<tr>
<td>Link</td>
<td># tweets</td>
</tr>
<tr>
<td>Hashtag</td>
<td># tweets</td>
</tr>
</tbody>
</table>
A heterogeneous graph (HG) is composed of nodes, attributes, and relations that could be of multiple different types.

The following questions are answered:

Q1: How to define an appropriate scan statistic for a given connected subgraph (“window”) of HG?

Q2: How to efficiently find connected subgraphs in HG that have the largest scan statistic scores?
Summary of Our Major Contributions

• Q1: How to define an appropriate scan statistic for a given connected subgraph (“window”) of HG?
  – First, we propose a two-stage empirical calibration process to calculate an empirical p-value for each node of HG
  – Second, we present a nonparametric scan statistic for a given connected subgraph of HG based on node-level empirical p-values

• Q2 How to efficiently find connected subgraphs that have the largest scan statistic scores?
  – We design an efficient algorithm to approximately maximize the proposed nonparametric scan statistic over connected subgraphs with the time complexity (O(|V| log |V|)), where |V| refers to the total number of nodes in HG
Two Stage Empirical Calibration Process

\[ p_{value} = 0.02 \]

\[ p_{value} = 0.07 \]
Two Stage Empirical Calibration Process

• Question: Each node has multiple p-values. How can we aggregate multiple p-values into a single p-value?
  – Step 1: Calculate the minimal of multiple p-values
  – Step 2: Estimate the single p-value based on the empirical calibration of minimal p-values in historical data

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td># tweets, # retweets, # followers, #followees,</td>
</tr>
<tr>
<td></td>
<td>#mentioned_by, #replied_by,</td>
</tr>
<tr>
<td></td>
<td>diffusion graph depth, diffusion graph size</td>
</tr>
<tr>
<td>Tweet</td>
<td>Klout, sentiment, replied_by_graph_size, reply_graph_size,</td>
</tr>
<tr>
<td></td>
<td>retweet_graph_size, retweet_graph_depth</td>
</tr>
<tr>
<td>City, State, Country</td>
<td># tweets, # active users</td>
</tr>
<tr>
<td>Term</td>
<td># tweets</td>
</tr>
<tr>
<td>Link</td>
<td># tweets</td>
</tr>
<tr>
<td>Hashtag</td>
<td># tweets</td>
</tr>
</tbody>
</table>
• THEOREM: The empirical p-value \( p(v) \) calculated using two-stage empirical calibration process follows a uniform distribution on \([0, 1]\) under the assumption that the current multivariate observations for a single node are exchangeable within the reference set given the null hypothesis that no events of interest are occurring.
Nonparametric Scan Statistics

Sub-graph

\[ F(S) = \max_{\alpha \leq \alpha_{\text{max}}} \max_{\alpha \leq \alpha_{\text{max}}} \phi(\alpha, N_{\alpha}(S), N(S)) \]

Significance level

Berk-Jones (BJ) Statistic

Number of nodes in \( S \) with p values \( \leq \alpha \)

\[ \phi_{BJ}(\alpha, N_{\alpha}(S), N(S)) = N(S)K\left(\frac{N_{\alpha}}{N}, \alpha\right) \]

Kullback-Liebler Divergence

\[ K(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y}, \]
The BJ statistic can be described as the log-likelihood ratio statistic for testing whether the empirical p-values are uniformly distributed on [0, 1], where the alternative hypothesis assumes a piecewise constant distribution with probability density function

\[
f(x) = \begin{cases} 
  f_1 & \text{for } 0 \leq x \leq \alpha \\
  f_2 & \text{for } \alpha \leq x \text{ with } f_1 \geq f_2 
\end{cases}
\]

Berk and Jones (1979) demonstrated this test statistic fulfills several optimality properties and has greater power than any weighted Kolmogorov statistic.
Nonparametric Graph Scanning

We propose an approximate algorithm with time cost $O(|V_T| \log |V_T|)$.

$$S^* = \arg\max_{S \in V_T, S \text{ is connected}} F(S)$$
Nonparametric Graph Scanning

\[
\max_{S \subseteq V, S \text{ is connected}} \max_{\alpha \leq \alpha_{\text{max}}} \phi(\alpha, N_\alpha(S), N(S))
\]

The union of \(\{\alpha_{\text{max}}\}\) and the set of distinct p-values less than \(\alpha_{\text{max}}\) in \(V\)
The union of \( \{ \alpha_{\text{max}} \} \) and the set of distinct p-values less than \( \alpha_{\text{max}} \) in \( V \)
Consider each node as a candidate cluster center (or start point)

In this example, we start from the seed set $S^{(0)} = \{v_1\}$. 
Expand $S^{(0)}$ by adding positive neighbor nodes:

$\mathcal{N}(S^{(0)}) = \{v_2, v_3, v_4, v_5, v_6\}$

$S^{(1)} = S^{(0)} \cup \arg \max_{\alpha \in U(S^{(0)} \cup S, \alpha_{max})} \left\{ \max_{S \in \mathcal{N}(S^{(0)})} NK \left( \frac{N\alpha(S^{(0)} \cup S)}{N(S^{(0)} \cup S)}, \alpha \right) \right\}$
Expand $S^{(0)}$ by adding positive neighbor nodes:

\[ \mathcal{N}(S^{(0)}) = \{v_2, v_3, v_4, v_5, v_6\} \]

\[ S^{(1)} = S^{(0)} \cup \arg\max_{\alpha \in U(S^{(0)} \cup S, \alpha_{\text{max}})} \left\{ \max_{S \in \mathcal{N}(S^{(0)})} NK \left( \frac{N\alpha(S^{(0)} \cup S)}{N(S^{(0)} \cup S)} , \alpha \right) \right\} \]

\[ = \{v_1, v_2, v_3, v_4, v_6\} \]
Nonparametric Graph Scanning Algorithm

Expand $S^{(1)}$ by adding positive neighbor nodes:

$\mathcal{N}(S^{(1)}) = \{v_5, v_7, v_8, v_{11}, v_{16}\}$

$S^{(2)} = S^{(1)} \cup \arg \max_{\alpha \in U(S^{(1)} \cup S, \alpha_{max})} \left\{ \max_{S \in \mathcal{N}(S^{(1)})} NK \left( \frac{N_{\alpha}(S^{(1)} \cup S)}{N(S^{(1)} \cup S)}, \alpha \right) \right\}$
Expand $S^{(1)}$ by adding positive neighbor nodes:

$$\mathcal{N}(S^{(1)}) = \{v_5, v_7, v_8, v_{11}, v_{16}\}$$

$$S^{(2)} = S^{(1)} \cup \arg\max_{\alpha \in U(S^{(1)} \cup S, \alpha_{\text{max}})} \left\{ \max_{S \in \mathcal{N}(S^{(1)})} NK \left( \frac{N\alpha(S^{(1)} \cup S)}{N(S^{(1)} \cup S)}, \alpha \right) \right\}$$

$$= \{v_1, v_2, v_3, v_4, v_6, v_8, v_{11}\}$$
Consider each node as a candidate cluster center (or start point).

In this example, we start from the seed set $\hat{S} = \{v_1\}$, and after four expansions, we obtain the local optimum solution:

$$S_{v_1}^* = \{v_1, v_2, v_3, v_4, v_6, v_8, v_9, v_{11}, v_{12}, v_{17}\}$$
In this example, we start from the seed set $S = \{ (v_1, 0.09) \}$, and after four expansions, we obtain the local optimum solution:

$$S^*_{v_1} = \{ v_1, v_2, v_3, v_4, v_6, v_8, v_9, v_{11}, v_{12}, v_{17} \}$$
Experimental Evaluations

• Detection and Forecasting of Hantavirus Disease Outbreaks
• Detection and Forecasting of Civil Unrest Events
Experiment Settings

• Twitter Dataset
  – 10% random sample of public twitter data from 2012-June to 2013-April in four countries, including Argentina, Chile, Colombia, and Ecuador.

• Event Labels
  – 17 hantavirus outbreaks in Chile
  – 918 civil unrest events in the four countries

• Performance Metrics
  – Forecasting: Have an alert in the same state 1-7 days before the event
  – Detection: Did not have an alert in that state 1-7 days before the event but did have an alert in the event 07 days after the event
# Twitter Dataset for Hantavirus Outbreaks

<table>
<thead>
<tr>
<th>Country</th>
<th># of tweets</th>
<th>News source*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chile</td>
<td>14,000,000</td>
<td>La Tercera; Las Últimas Noticias; El Mercurio</td>
</tr>
</tbody>
</table>

**Time Period:** From *2013 Jan.* to *2013 Jun.* Totally **17** Hantavirus outbreaks

Example of an event label: (PROVINCE = “La Araucanía”, COUNTRY = “Chile”, DATE = “2013-01-19”, News Title = “A 11 year old boy who was admitted to a clinic in Temuco down suspected hantavirus died during the day on Saturday.”, News Link = “http://www.biobiochile.cl/2013/01/19/muere-menor-sospechoso-de-hanta-en-temuco.shtml”).
Twitter Dataset for Civil Unrests

<table>
<thead>
<tr>
<th>Country</th>
<th># of tweets</th>
<th>News source*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>29,000,000</td>
<td>Clarín; La Nación; Infobae</td>
</tr>
<tr>
<td>Chile</td>
<td>14,000,000</td>
<td>La Tercera; Las Últimas Noticias; El Mercurio</td>
</tr>
<tr>
<td>Colombia</td>
<td>22,000,000</td>
<td>El Espectador; El Tiempo; El Colombiano</td>
</tr>
<tr>
<td>Ecuador</td>
<td>6,900,000</td>
<td>El Universo; El Comercio; Hoy</td>
</tr>
</tbody>
</table>

Time Period: From **2012 Jul.** to **2013 Apr.** Totally **918** civil unrest events
### Comparison with Baseline Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>FPR (FP/Day)</th>
<th>TPR (Forecasting)</th>
<th>TPR (Forecasting &amp; Detection)</th>
<th>Lead Time (Days)</th>
<th>Lag Time (Days)</th>
<th>Run Time (Hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST Burst Detection</td>
<td>0.65</td>
<td>0.07</td>
<td>0.42</td>
<td>1.10</td>
<td>4.57</td>
<td>30.1</td>
</tr>
<tr>
<td>Graph Partition</td>
<td>0.29</td>
<td>0.03</td>
<td>0.15</td>
<td>0.59</td>
<td>6.13</td>
<td>18.9</td>
</tr>
<tr>
<td>Earthquake</td>
<td>0.04</td>
<td>0.06</td>
<td>0.17</td>
<td>0.49</td>
<td>5.95</td>
<td>18.9</td>
</tr>
<tr>
<td>RW Event</td>
<td>0.10</td>
<td>0.22</td>
<td>0.25</td>
<td>0.93</td>
<td>5.83</td>
<td>16.3</td>
</tr>
<tr>
<td>Geo Topic Modeling</td>
<td>0.09</td>
<td>0.06</td>
<td>0.08</td>
<td>0.01</td>
<td>6.94</td>
<td>9.7</td>
</tr>
<tr>
<td>NPHGS (FPR=.05)</td>
<td>0.05</td>
<td>0.15</td>
<td>0.23</td>
<td>0.65</td>
<td>5.65</td>
<td>38.4</td>
</tr>
<tr>
<td>NPHGS (FPR=.10)</td>
<td>0.10</td>
<td>0.31</td>
<td>0.38</td>
<td>1.94</td>
<td>4.49</td>
<td>38.4</td>
</tr>
<tr>
<td>NPHGS (FPR=.15)</td>
<td>0.15</td>
<td>0.37</td>
<td>0.42</td>
<td>2.28</td>
<td>4.17</td>
<td>38.4</td>
</tr>
<tr>
<td>NPHGS (FPR=.20)</td>
<td>0.20</td>
<td>0.39</td>
<td>0.46</td>
<td>2.36</td>
<td>3.98</td>
<td>38.4</td>
</tr>
</tbody>
</table>

Table 3: Comparison between NPHGS and Existing Methods on the civil unrest datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>FPR (FP/Day)</th>
<th>TPR (Forecasting)</th>
<th>TPR (Forecasting &amp; Detection)</th>
<th>Lead Time (Days)</th>
<th>Lag Time (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST Burst Detection</td>
<td>0.57</td>
<td>0.25</td>
<td>0.63</td>
<td>1.13</td>
<td>3.81</td>
</tr>
<tr>
<td>Graph Partition</td>
<td>0.57</td>
<td>0.06</td>
<td>0.19</td>
<td>0.19</td>
<td>6.10</td>
</tr>
<tr>
<td>Earthquake</td>
<td>0.92</td>
<td>0.13</td>
<td>0.19</td>
<td>0.75</td>
<td>5.69</td>
</tr>
<tr>
<td>RW Event</td>
<td>0.40</td>
<td>0.19</td>
<td>0.41</td>
<td>0.43</td>
<td>4.91</td>
</tr>
<tr>
<td>Geo Topic Modeling</td>
<td>0.43</td>
<td>0.19</td>
<td>0.50</td>
<td>0.62</td>
<td>4.31</td>
</tr>
<tr>
<td>NPHGS (FPR=.05)</td>
<td>0.05</td>
<td>0.20</td>
<td>0.78</td>
<td>0.71</td>
<td>2.44</td>
</tr>
<tr>
<td>NPHGS (FPR=.10)</td>
<td>0.10</td>
<td>0.22</td>
<td>0.85</td>
<td>0.76</td>
<td>1.90</td>
</tr>
<tr>
<td>NPHGS (FPR=.15)</td>
<td>0.15</td>
<td>0.25</td>
<td>0.93</td>
<td>0.80</td>
<td>1.36</td>
</tr>
<tr>
<td>NPHGS (FPR=.20)</td>
<td>0.20</td>
<td>0.29</td>
<td>0.94</td>
<td>0.82</td>
<td>1.24</td>
</tr>
</tbody>
</table>

Table 4: Comparison between NPHGS and Existing Methods on the Hantavirus dataset
(a) FPR vs. TPR (Forecasting)

(b) FPR vs. TPR (Forecasting and Detection)

(c) FPR vs. Leadtime (Forecasting)

(d) TPR (Forecasting and Detection) vs. Precision
Part I: References


Big Data Analytics for Societal Event Forecasting

(Part II)

Speaker: Liang Zhao (George Mason University)

Preparation: Liang Zhao (George Mason University)

IEEE Conference on Big Data 2018

Dec 12, 2018
Taxonomy

Societal event forecasting

Temporal event forecasting

Discriminative Learning
- All-to-all
- All-to-one
- One-to-one

Generative / Mechanistic Learning
- Data-driven
- Data-driven+Mechanistic

Ensemble Learning
- Mechanistic
Discriminative Learning-based

Predictive mapping

Event indicators

Future events

features

locations

Forecasting model:
Spatial dependency:
• Which locations are correlated
Feature selection:
• Which features are discriminative
Model selection:
• Which predictive model to choose
Societal event forecasting

Temporal event forecasting

Discriminative Learning

- All-to-all
- All-to-one
- One-to-one

Generative / Mechanistic Learning

- Data-driven
- Data-driven+Mechanistic

Spatiotemporal event forecasting

Ensemble Learning

- Mechanistic
- Generative
**Categorization**

**All-to-all models**

- **Features**
- **Locations**

**Advantages:**
- No need to define spatial autocorrelation
- Simple model, easy to train
- Small data is needed
- More efficient: Complexity $\geq K$

**Disadvantages:**
- Cannot consider spatial dependency of inputs
- Ignore potential correlation among the events

**S**: number of locations, **K**: number of features

**All-to-one models**

- **Features**
- **Locations**

**Advantages:**
- Consider spatial dependency of inputs
- No need to define spatial autocorrelation
- More efficient: Complexity $\geq S \cdot K$

**Disadvantages:**
- Complex model
- Large data required
- Ignore potential correlation among the events

**One-to-one models**

- **Features**
- **Locations**

**Advantages:**
- Consider spatial dependency of inputs
- Consider spatial autocorrelation of outputs

**Disadvantages:**
- Time&memory consuming, complexity $\geq S^2 \cdot K$
- Complex model
- Large data required
- Tricky to define how locations are auto-correlated.
One-to-One models

Use individual location to forecast for each corresponding individual location

\[ f_W(\cdot) \rightarrow f(W^T \cdot X_{t,s}) \]

\( f \): predictive model
- Logistic regression [Gerber, DSS’15]
- Support vector machine [Gerber, TCSS’18]
- So on so forth…
Category 1: All locations share a single model

Pro: sufficient data to train model

Con: ignore the individual city’s exclusive characteristics (size, population, etc.)

City: Mexico City
Population: 8M
Size: 573 mi²

City: Taxco
Population: 39K
Size: 134 mi²

1K protest tweets have different meanings to these two locations
Category 2: Each model for each location

Pro: consider the individual location’s exclusive characteristics

Con: 1. Insufficient data for small cities.
   2. Ignore the relatedness among different locations

Many cities have little data
Few cities have huge data

Similar expressions
Same languages
Shared keywords
Relevant events
Similar topics
Multi-task learning for Spatial Event Forecasting [Zhao et al., KDD’15]

All locations share a single model

Pro: consider the exclusive characteristics

Con: 1. Ignore the relatedness among different locations
   2. Insufficient data for small cities.

Each model for each location

Pro: Sufficient training data

Con: Ignore the individual city’s characteristics

Regularize all the models
Enforce knowledge sharing

Combine
Multi-task learning for Spatial Event Forecasting [Zhao et al., KDD’15]

\[
\min_W \sum_{i=1}^{S} \mathcal{L}(W_i^T X_i, Y_i) + \lambda \cdot \mathcal{R}(W)
\]

\[
\mathcal{R}(W) = \|W\|_{2,1} = \sum_i \|W_i\|_2
\]

Minimizing \(l_{2,1}\) norm will make the matrix row sparse

- A feature important for a location will also tend to be also important
- Their weights value can be different.

- Keywords “basketball” and “music” is unimportant for “influenza outbreaks” for various locations;
- Keywords “cold” and “cough” is important to forecast “influenza outbreaks” for various locations;
- However, their weights are different in different locations due to different population size in each location.
More constraints

\[
\min_{W} \sum_{i=1}^{L} \mathcal{L}(W_i^T X_i) + \lambda \cdot \mathcal{R}(W)
\]

s.t. \( \sum_j^m I(\|W_j\| > 0) \leq r \)

Sometimes, the users have preference on how many features to select

Model optimization algorithm: Solved by projected gradient descent.
Experiments: Event Forecasting Performance

- Combining static features and dynamic features is BETTER
- The proposed CMTFL II is generally the BEST
- CMTFL II is able to ensure the inclusion of both type of features
- Multitask models outperform the traditional LASSO models

Training set: Twitter data from July 1, 2012 to December 31, 2012
Testing set: Twitter data from January 1, 2013 to May 31, 2013
Label set: Authoritative news reports on civil unrest events

<table>
<thead>
<tr>
<th>method</th>
<th>Mexico</th>
<th>Paraguay</th>
<th>Brazil</th>
<th>All Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQEF</td>
<td>0.56, 0.40, 0.47</td>
<td>0.90, 0.15, 0.26</td>
<td>0.37, 0.34, 0.35</td>
<td>0.54, 0.38, 0.45</td>
</tr>
<tr>
<td>LASSO-K</td>
<td>0.68, 0.32, 0.44</td>
<td>1.00, 0.17, 0.29</td>
<td>0.62, 0.44, 0.51</td>
<td>0.72, 0.28, 0.40</td>
</tr>
<tr>
<td>DQEF+LASSO</td>
<td>0.57, 0.49, 0.53</td>
<td>1.00, 0.11, 0.20</td>
<td>0.42, 0.49, 0.45</td>
<td>0.55, 0.44, 0.49</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.70, 0.36, 0.48</td>
<td>1.00, 0.17, 0.29</td>
<td>0.63, 0.43, 0.51</td>
<td>0.73, 0.30, 0.43</td>
</tr>
<tr>
<td>rMTFL-D</td>
<td>0.96, 0.12, 0.21</td>
<td>1.00, 0.02, 0.04</td>
<td>1.00, 0.07, 0.13</td>
<td>0.77, 0.15, 0.25</td>
</tr>
<tr>
<td>rMTFL-K</td>
<td>0.78, 0.45, 0.57</td>
<td>0.93, 0.43, 0.59</td>
<td>0.79, 0.55, 0.65</td>
<td>0.71, 0.51, 0.59</td>
</tr>
<tr>
<td>rMTFL</td>
<td>0.70, 0.70, 0.70</td>
<td>0.96, 0.32, 0.48</td>
<td>0.71, 0.52, 0.60</td>
<td>0.68, 0.57, 0.62</td>
</tr>
<tr>
<td>CMTFL-I</td>
<td>0.59, 0.87, 0.70</td>
<td>0.95, 0.39, 0.55</td>
<td>0.72, 0.60, 0.66</td>
<td>0.62, 0.68, 0.65</td>
</tr>
<tr>
<td>CMTFL-II</td>
<td>0.71, 0.79, 0.75</td>
<td>0.78, 0.81, 0.79</td>
<td>0.76, 0.57, 0.65</td>
<td>0.69, 0.71, 0.70</td>
</tr>
</tbody>
</table>
Few and not relevant keywords, due to the sparsity of the training data for small state

Does not ensure to include the dynamic features

Does not ensure to include the dynamic features
Multi-task Event Scale Forecasting
[Gao, and Zhao, AAAI’18]

Event Scale Forecasting (Gao et al., AAAI’18)

Influenza outbreaks in Week 13 of 2016

Generalize the output to ordinal!

Ordinal regression
Multi-task Event Subtype Forecasting
[Gao, et al. AAAI’19]

Event Subtype Forecasting (Gao et al., AAAI’19)

Primary Pollutant in one day in Shenzhen, China, 2013.

Multi-class classification

Generalize the output to multi-class!

Multi-class Classification

Spatial Incomplete Multi-task Deep Learning (SIMDA)
- Enforce shared hidden representations across tasks
- Enforce similar subtype patterns for adjacent tasks
Taxonomy

Societal event forecasting

Temporal event forecasting
- Discriminative Learning
  - All-to-all
  - All-to-one
  - One-to-one

Spatiotemporal event forecasting
- Generative / Mechanistic Learning
  - Data-driven
  - Data-driven + Mechanistic

- Ensemble Learning
  - Mechanistic

- Generative
All-to-one models

Use Multiple locations to forecast for each individual location

All the locations (spatial dependency among indicators)

When the inputs have strong:
  Spatial hierarchy
  Spatial dependency
  Spatial multi-resolution

Each individual location (spatial independency among events)
Spatial dependency among inputs

[Zhao et al., KDD’16]

Different feature in different spatial levels

City-level features

State-level features

Country-level features

A city
Applications: Multi-source Event Forecasting

Why multiple data sources?
• Spatiotemporal events are often influenced by different aspects of the society.
• Different data sources complement each other.
• One single source cannot cover all aspects of an event.

News reports

Social Media

Economic Index

A wave of massive protest events in Brazil in 2013, caused by economic issues, organized by social media, and reported by news reports.
Spatial hierarchy among inputs

Challenge 1: Hierarchical topology
- E.g., country-level, state-level, city-level
- Higher-level features can influence lower-level ones

Challenge 2: Interactive missing values
- Different data sources, different spans
- Need to consider the interactions among different sources.
Hierarchical Incomplete Multisource Feature Learning

Given the multi-source data for a location \( l \) at time \( t \), predict whether the event will happen at time \( \tau \)

\[
f : \{X_{t, l_1}, \ldots, X_{t, l_N}\} \rightarrow Y_{\tau, l}
\]

- Each location has features at multiple levels \( l=(l_1, l_2, \ldots, l_N) \)
  E.g., (San Francisco, CA, USA)

Variables are dependent on the variables in their parent level

\[
\begin{align*}
(l_{level} - 1) & \quad Y_{\tau, l} = \alpha_0 + \sum_{i=1}^{F_1} \alpha_i^T \cdot [X_{t, l_1}]_i + \varepsilon \\
(l_{level} - 2) & \quad \alpha_i = \beta_{i,0} + \sum_{j=1}^{F_2} \beta_{i,j}^T \cdot [X_{t, l_2}]_j + \varepsilon_i \\
(l_{level} - 3) & \quad \beta_{i,j} = W_{i,j,0} + \sum_{k=1}^{F_3} W_{i,j,k}^T \cdot [X_{t, l_3}]_k + \varepsilon_{i,j}
\end{align*}
\]

\[
Y_{\tau, l} = \sum_{i=0}^{F_1} \sum_{j=0}^{F_2} \sum_{k=0}^{F_3} W_{i,j,k} \cdot [X_{t, l_1}]_i \cdot [X_{t, l_2}]_j \cdot [X_{t, l_3}]_k + \varepsilon
\]

Encode hierarchical feature correlation by nth-order strong hierarchy

Tensor form:

\[
Y_{\tau, l} = W \odot Z_{t, l} + \varepsilon
\]
## Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Label sources ¹</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>CU</td>
<td>Clarín; La Nación; Infobae</td>
<td>1306</td>
</tr>
<tr>
<td>Brazil</td>
<td>CU</td>
<td>O Globo; O Estado de São Paulo; Jornal do Brasil</td>
<td>3226</td>
</tr>
<tr>
<td>Chile</td>
<td>CU</td>
<td>La Tercera; Las Últimas Noticias; El Mercurio</td>
<td>706</td>
</tr>
<tr>
<td>Colombia</td>
<td>CU</td>
<td>El Espectador; El Tiempo; El Colombiano</td>
<td>1196</td>
</tr>
<tr>
<td>El Salvador</td>
<td>CU</td>
<td>El Diario de Hoy; La Prensa Gráfica; El Mundo</td>
<td>657</td>
</tr>
<tr>
<td>Mexico</td>
<td>CU</td>
<td>La Jornada; Reforma; Milenio</td>
<td>5465</td>
</tr>
<tr>
<td>Paraguay</td>
<td>CU</td>
<td>ABC Color; Ultima Hora; La Nación</td>
<td>1932</td>
</tr>
<tr>
<td>Uruguay</td>
<td>CU</td>
<td>El País; El Observador</td>
<td>624</td>
</tr>
<tr>
<td>Venezuela</td>
<td>CU</td>
<td>El Universal; El Nacional; Ultimas Noticias</td>
<td>3105</td>
</tr>
<tr>
<td>U.S.</td>
<td>FLU</td>
<td>CDC Flu Activity Map</td>
<td>1027</td>
</tr>
</tbody>
</table>

¹ Label sources include news reports, disease surveillance reports, and specific news outlets for each country.
Hierarchical features and missing values

Multi-level sources

Block-wise missing values

Multi-source features

Multi-source features
AUC for different missing ratios

(AUC: area under ROC curve)

<table>
<thead>
<tr>
<th>Method</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Chile</th>
<th>Colombia</th>
<th>El Salvador</th>
<th>Mexico</th>
<th>Paraguay</th>
<th>Uruguay</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>0.5287</td>
<td>0.7476</td>
<td>0.5624</td>
<td>0.8032</td>
<td>0.3148</td>
<td>0.7823</td>
<td>0.5572</td>
<td>0.4693</td>
<td>0.8973</td>
</tr>
<tr>
<td>LASSO-INT</td>
<td>0.5268</td>
<td>0.7191</td>
<td>0.5935</td>
<td>0.7861</td>
<td>0.5269</td>
<td>0.777</td>
<td>0.4887</td>
<td>0.5069</td>
<td>0.7543</td>
</tr>
<tr>
<td>iMSF</td>
<td>0.4795</td>
<td>0.4611</td>
<td>0.6033</td>
<td>0.7213</td>
<td>0.5569</td>
<td>0.4486</td>
<td>0.4904</td>
<td>0.5</td>
<td>0.5501</td>
</tr>
<tr>
<td>MTL</td>
<td>0.3885</td>
<td>0.5017</td>
<td>0.5011</td>
<td>0.4334</td>
<td>0.3452</td>
<td>0.4674</td>
<td>0.4313</td>
<td>0.3507</td>
<td>0.5501</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.5085</td>
<td>0.7317</td>
<td>0.6148</td>
<td>0.8084</td>
<td>0.777</td>
<td>0.8037</td>
<td>0.7339</td>
<td>0.7264</td>
<td>0.7846</td>
</tr>
<tr>
<td>HIML</td>
<td>0.5873</td>
<td>0.8353</td>
<td>0.5705</td>
<td>0.8160</td>
<td>0.7101</td>
<td>0.7478</td>
<td>0.8573</td>
<td>0.7488</td>
<td>0.7488</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Chile</th>
<th>Colombia</th>
<th>El Salvador</th>
<th>Mexico</th>
<th>Paraguay</th>
<th>Uruguay</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>0.5035</td>
<td>0.7362</td>
<td>0.588</td>
<td>0.8412</td>
<td>0.3785</td>
<td>0.7896</td>
<td>0.478</td>
<td>0.6749</td>
<td>0.681</td>
</tr>
<tr>
<td>LASSO-INT</td>
<td>0.4976</td>
<td>0.6361</td>
<td>0.5912</td>
<td>0.8151</td>
<td>0.3852</td>
<td>0.7622</td>
<td>0.426</td>
<td>0.7177</td>
<td>0.6428</td>
</tr>
<tr>
<td>iMSF</td>
<td>0.4797</td>
<td>0.4611</td>
<td>0.4950</td>
<td>0.6845</td>
<td>0.5</td>
<td>0.5569</td>
<td>0.4811</td>
<td>0.4898</td>
<td>0.5</td>
</tr>
<tr>
<td>MTL</td>
<td>0.4207</td>
<td>0.5156</td>
<td>0.5023</td>
<td>0.5978</td>
<td>0.3413</td>
<td>0.4666</td>
<td>0.4318</td>
<td>0.347</td>
<td>0.4397</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.5012</td>
<td>0.7724</td>
<td>0.6245</td>
<td>0.8032</td>
<td>0.7628</td>
<td>0.7598</td>
<td>0.738</td>
<td>0.8205</td>
<td>0.7621</td>
</tr>
<tr>
<td>HIML</td>
<td>0.5854</td>
<td>0.8497</td>
<td>0.6072</td>
<td>0.8440</td>
<td>0.726</td>
<td>0.7907</td>
<td>0.7471</td>
<td>0.8576</td>
<td>0.7378</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Chile</th>
<th>Colombia</th>
<th>El Salvador</th>
<th>Mexico</th>
<th>Paraguay</th>
<th>Uruguay</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>0.5128</td>
<td>0.7461</td>
<td>0.5301</td>
<td>0.8167</td>
<td>0.3139</td>
<td>0.7552</td>
<td>0.5285</td>
<td>0.5992</td>
<td>0.6678</td>
</tr>
<tr>
<td>LASSO-INT</td>
<td>0.504</td>
<td>0.6145</td>
<td>0.5537</td>
<td>0.7339</td>
<td>0.4283</td>
<td>0.7309</td>
<td>0.4745</td>
<td>0.5396</td>
<td>0.6155</td>
</tr>
<tr>
<td>iMSF</td>
<td>0.4796</td>
<td>0.4611</td>
<td>0.4962</td>
<td>0.7467</td>
<td>0.4899</td>
<td>0.5488</td>
<td>0.4804</td>
<td>0.487</td>
<td>0.5</td>
</tr>
<tr>
<td>MTL</td>
<td>0.5104</td>
<td>0.4818</td>
<td>0.4715</td>
<td>0.65</td>
<td>0.3375</td>
<td>0.4744</td>
<td>0.436</td>
<td>0.3578</td>
<td>0.3839</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.5101</td>
<td>0.7717</td>
<td>0.6599</td>
<td>0.842</td>
<td>0.7685</td>
<td>0.8079</td>
<td>0.7324</td>
<td>0.8112</td>
<td>0.7759</td>
</tr>
<tr>
<td>HIML</td>
<td>0.5795</td>
<td>0.8463</td>
<td>0.548</td>
<td>0.8432</td>
<td>0.7126</td>
<td>0.7892</td>
<td>0.7477</td>
<td>0.856</td>
<td>0.7176</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Chile</th>
<th>Colombia</th>
<th>El Salvador</th>
<th>Mexico</th>
<th>Paraguay</th>
<th>Uruguay</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>0.5162</td>
<td>0.6874</td>
<td>0.5947</td>
<td>0.8344</td>
<td>0.2507</td>
<td>0.7485</td>
<td>0.4075</td>
<td>0.3652</td>
<td>0.6999</td>
</tr>
<tr>
<td>LASSO-INT</td>
<td>0.4691</td>
<td>0.5557</td>
<td>0.5169</td>
<td>0.7167</td>
<td>0.2116</td>
<td>0.7086</td>
<td>0.4075</td>
<td>0.3652</td>
<td>0.6999</td>
</tr>
<tr>
<td>iMSF</td>
<td>0.4796</td>
<td>0.4611</td>
<td>0.5503</td>
<td>0.7855</td>
<td>0.5567</td>
<td>0.4795</td>
<td>0.5221</td>
<td>0.5</td>
<td>0.5501</td>
</tr>
<tr>
<td>MTL</td>
<td>0.4128</td>
<td>0.5023</td>
<td>0.5069</td>
<td>0.6195</td>
<td>0.3323</td>
<td>0.4702</td>
<td>0.4283</td>
<td>0.3569</td>
<td>0.6164</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.5188</td>
<td>0.7741</td>
<td>0.6659</td>
<td>0.8121</td>
<td>0.7557</td>
<td>0.8097</td>
<td>0.7136</td>
<td>0.72</td>
<td>0.6993</td>
</tr>
<tr>
<td>HIML</td>
<td>0.5484</td>
<td>0.7812</td>
<td>0.3887</td>
<td>0.8416</td>
<td>0.7181</td>
<td>0.8001</td>
<td>0.7146</td>
<td>0.8453</td>
<td>0.716</td>
</tr>
</tbody>
</table>

- The proposed HIML performs the best
- Methods considers hierarchical features performs better
- Performance decreases when missing ratio increases
- Methods that can handle incomplete data decreases slower in performance
Spatial multi-resolution among inputs

Same features but in different spatial resolutions
Spatial big data typically is with multiple geo-resolutions.

In Fine-grained Spatial Event Forecasting tasks, Most of the data discarded.

[Zhao et al., ICDM’16]
Geographically hierarchical relationship

Model optimization

\[
\min_w \sum_m \mathcal{L}(Y_m, f(W_m)) + \gamma_1 \sum_{N \in N} \left\| \left\{ W_m^{(1)} \right\}_m^N \right\|_{2,1} + \gamma_2 \sum_{R \in R} \left\| \left\{ W_m^{(2)} \right\}_m^R \right\|_{2,1}
\]

s.t. \( W_m = W_m^{(1)} + W_m^{(2)} \)

Empirical loss

Logical constraints:
If event in Los Angeles, then event in California;

Spatial neighborhood relation

Spatial resolution relation

Parent-child relation
Multi-resolution event forecasting

(Precision, Recall, F-measure)

<table>
<thead>
<tr>
<th>Method</th>
<th>State-level</th>
<th>Region-level</th>
<th>Country-level</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARX</td>
<td>0.53,0.60,0.56</td>
<td>0.33,0.57,0.42</td>
<td>0.60,0.90,0.72</td>
<td>0.49,0.70,0.58</td>
</tr>
<tr>
<td>LR</td>
<td>0.48,0.47,0.48</td>
<td>0.49,0.45,0.47</td>
<td>0.63,0.93,0.75</td>
<td>0.53,0.62,0.57</td>
</tr>
<tr>
<td>KDE-LR</td>
<td>0.13,0.93,0.22</td>
<td>0.23,0.20,0.21</td>
<td>0.97,0.69,0.80</td>
<td>0.44,0.61,0.51</td>
</tr>
<tr>
<td>LDA-LR</td>
<td>0.14,0.93,0.24</td>
<td>0.19,0.21,0.20</td>
<td>0.94,0.70,0.80</td>
<td>0.42,0.61,0.50</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.61,0.46,0.52</td>
<td>0.52,0.49,0.50</td>
<td>0.54,0.99,0.70</td>
<td>0.56,0.65,0.60</td>
</tr>
<tr>
<td>MTL</td>
<td>0.60,0.54,0.56</td>
<td>0.58,0.55,0.56</td>
<td>0.68,0.75,0.71</td>
<td>0.62,0.61,0.61</td>
</tr>
<tr>
<td>TMTL</td>
<td>0.60,0.64,0.62</td>
<td>0.52,0.52,0.52</td>
<td>0.67,0.72,0.70</td>
<td>0.60,0.63,0.61</td>
</tr>
<tr>
<td>MREF</td>
<td>0.66,0.60,0.63</td>
<td>0.63,0.58,0.61</td>
<td>0.69,0.97,0.81</td>
<td>0.66,0.72,0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>ARX</th>
<th>LR</th>
<th>KDE-LR</th>
<th>LDA-LR</th>
<th>LASSO</th>
<th>MTL</th>
<th>TMTL</th>
<th>MREF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runtime (sec)</td>
<td>21</td>
<td>37</td>
<td>2026</td>
<td>296</td>
<td>118</td>
<td>45</td>
<td>656</td>
<td>923</td>
</tr>
</tbody>
</table>

- The proposed MREF performs the best
- Multi-task Methods perform better
- Accuracy increases when geo-resolution becomes coarser
Precision-Recall curves

The proposed MREF performs best in general. Multi-task Methods perform better.
Taxonomy

Societal event forecasting

Temporal event forecasting
Discriminative Learning
  All-to-all
  All-to-one
  One-to-one

Generative / Mechanistic Learning
  Data-driven
  Data-driven + Mechanistic
  Generative
  Mechanistic

Spatiotemporal event forecasting
Ensemble Learning
All-to-all models

Use all the locations to forecast for each location individually

All the locations (spatial dependency among indicators)

In some domain, events can trigger other events

Practitioners want much more than just prediction results

Additional Questions

1. Trigger event type?
   - Event 4 (on I85/17ST) detoured traffic
   - Event 3 (on Peachtree RD) detoured traffic
2. Trigger event?
   - Event 4
   - Event 1
3. Indicative messages/signals?
   - posts/images
   - posts/images
   - posts/images

(a) current crowdsourced observations
(b) future congestion events
Challenges:

- Existing methods fail to sufficiently utilize all different sources.
- Spatial heterogeneity
  - e.g., rural vs urban
- Class imbalance
  - a.k.a., accidents are rare
The structure of the regional ConvLSTM model

Regional Accident Risk Model

Feed Data Through Time

Input Feature Maps

Predicting Crash Map

Regional Accident Risk Model

Input Feature Maps

Data Source

DOT

NWS

Google Earth

$7 \times 64 \times 128$
Using heterogeneous data sources is advantageous!

### Table 1: Model Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Type-1 Urban</th>
<th>Type-2 Rural</th>
<th>Type-3 Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>RMSE</td>
<td>CE</td>
</tr>
<tr>
<td>LR(C=0.1)</td>
<td>0.146</td>
<td>0.382</td>
<td>0.051</td>
</tr>
<tr>
<td>DTR(depth=30)</td>
<td>0.172</td>
<td>0.415</td>
<td>0.243</td>
</tr>
<tr>
<td>DNN(2048x2048)</td>
<td>0.140</td>
<td>0.374</td>
<td>0.033</td>
</tr>
<tr>
<td>FC-LSTM(2048x2048)</td>
<td>0.187</td>
<td>0.434</td>
<td>0.419</td>
</tr>
<tr>
<td>ConvLSTM (128x128x128x128)</td>
<td>0.117</td>
<td>0.343</td>
<td>0.074</td>
</tr>
<tr>
<td>Historical Average (7 years)</td>
<td>0.050</td>
<td>0.224</td>
<td>0.340</td>
</tr>
<tr>
<td>Hetero-ConvLSTM (128x128x128x128)</td>
<td><strong>0.021</strong></td>
<td><strong>0.144</strong></td>
<td><strong>0.014</strong></td>
</tr>
</tbody>
</table>

### Table 2: Impact of Feature Groups

<table>
<thead>
<tr>
<th>Model</th>
<th>Type-1 Urban</th>
<th>Type-2 Rural</th>
<th>Type-3 Mixed</th>
<th>All Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>RMSE</td>
<td>CE</td>
<td>MSE</td>
</tr>
<tr>
<td>N</td>
<td>0.120</td>
<td>0.346</td>
<td>0.089</td>
<td>0.063</td>
</tr>
<tr>
<td>N+RW+RA</td>
<td>0.126</td>
<td>0.356</td>
<td>0.073</td>
<td>0.038</td>
</tr>
<tr>
<td>N+RW+RA+V+RC</td>
<td>0.123</td>
<td>0.351</td>
<td>0.127</td>
<td>0.039</td>
</tr>
<tr>
<td>N+RW+RA+V+RC+G</td>
<td>0.148</td>
<td>0.384</td>
<td>0.247</td>
<td>0.038</td>
</tr>
<tr>
<td>N+RW+RA+V+RC+G+CL</td>
<td>0.118</td>
<td>0.344</td>
<td>0.075</td>
<td>0.046</td>
</tr>
<tr>
<td>N+RW+RA+V+RC+G+CL+E</td>
<td><strong>0.117</strong></td>
<td><strong>0.343</strong></td>
<td><strong>0.074</strong></td>
<td><strong>0.037</strong></td>
</tr>
</tbody>
</table>
Case study of traffic accidents

Ground Truth at Dec. 8th 2013

Predictions at Dec. 8th 2013

Multi-vehicle crash at Des Moines (I35)
Traffic Accidents at Cedar Rapids
Three Vehicles Crashed at Coralville, Iowa (I80)
References


• Zhuoning Yuan, Xun Zhou, Tianbao Yang. Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data. In 24th ACM SIGKDD International Conference on Knowledge Discovery from Data (KDD), 2018 (Accepted).


Taxonomy

Societal event forecasting

Temporal event forecasting

Discriminative Learning

- All-to-all
- All-to-one
- One-to-one

Generative / Mechanistic Learning

- Data-driven
- Data-driven+Mechanistic

Spatiotemporal event forecasting

Ensemble Learning

- Generative
- Mechanistic
Generative/ Mechanistic Models

**Event indicators**
**Future events**

Forecaster

Indicators

Features

locations

Advantages:
- Embed prior knowledge.
- Good at longer-term prediction.
- Better explainability

Disadvantages:
- Low efficiency.
- Labor intensive.
- Maybe biased and incomplete
Taxonomy

Societal event forecasting

Temporal event forecasting

Discriminative Learning
- All-to-all
- All-to-one
- One-to-one

Generative / Mechanistic Learning
- Data-driven
- Data-driven+Mechanistic

Spatiotemporal event forecasting

Ensemble Learning

- Mechanistic
Spatiotemporal Event Forecasting in Social Media
(Zhao, et al., SDM 2015; Zhao, et al., TSAS)

Inside: low
Outside: low

Inside: high
Outside: low

Inside: high
Outside: medium

Background words

This is just beginning ... this has not been a democratic election

Event specific words

Today #tauro do not hesitate to raise your voice in protest against injustice. Who comes to march?

Topic 1 words

Let us go to the MegaMarch on July 7 to Zócalo Angel on 3:00 pm. spread the word

Topic 2 words

Complaining
July 1, 2012

Planning
July 3, 2012

Dissemination
July 5, 2012

Protest event
July 7, 2012

tweet volume>500

50<tweet volume<500

0<tweet volume<50

★ Location of event
Utilizing prior geographical knowledge

Crime events happen more often in regions with bad security

Earthquakes occur more frequently on continent plate border

Civil unrest events occur more frequently on large cities
Generative process 1: Gaussian-based

Count of neighbors  Count of current location  Latent status

Spatial burstiness model

Structural context model

Event temporal progression

Whether background?
Which topic?
What words?

Trendy words
Generative process 2: Poisson-based

Count of neighbors

Count of current location

Spatial burstiness model

Structural context model

Event temporal progression

Whether background?

Which topic?

What words?

Trendy words

- Poisson-based Spatial burstiness model
- Event temporal progression
- Whether background?
- Which topic?
- What words?
- Trendy words
Experiments: Datasets

• Civil unrest:
  – #raw tweets: 32,459,668
  – #processed tweets: 57,856
  – #events: 726

• Flu outbreaks:
  – Time duration: 2011-01-01 - 2013-12-31
  – #raw tweets: 8,627,664,399
  – #processed tweets: 2,252,436
  – #events: 102
Performance comparisons

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Baseline</th>
<th>ARX</th>
<th>LR</th>
<th>LDA-LR</th>
<th>KDE-LDA-LR</th>
<th>Proposed algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>civil unrest data</td>
<td>precision</td>
<td>0.44</td>
<td>0.26</td>
<td>0.7</td>
<td>0.31</td>
<td>0.42</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>0.59</td>
<td>0.43</td>
<td>0.18</td>
<td>0.7</td>
<td>0.69</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>f1-score</td>
<td>0.5</td>
<td>0.32</td>
<td>0.29</td>
<td>0.43</td>
<td>0.52</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>runtime per day (sec)</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>0.001</td>
<td>0.005</td>
<td>0.005</td>
<td>0.32</td>
</tr>
<tr>
<td>flu data</td>
<td>precision</td>
<td>0.28</td>
<td>0.14</td>
<td>0.64</td>
<td>0.27</td>
<td>0.78</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>0.39</td>
<td>0.66</td>
<td>0.31</td>
<td>0.55</td>
<td>0.32</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>f1-score</td>
<td>0.33</td>
<td>0.23</td>
<td>0.41</td>
<td>0.36</td>
<td>0.46</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>runtime per day (sec)</td>
<td>$10^{-3}$</td>
<td>0.001</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>2.1</td>
</tr>
</tbody>
</table>

(a) Civil unrest dataset

(b) Flu dataset
Hidden topics

(a) topic 1 (describe protest)
(b) topic 2 (background)
(c) topic 3 (describe protest)

(d) topic 4 (descriptive)
(e) topic 5 (background)
(f) topic 6 (descriptive)
(g) topic 7 (descriptive)

(h) topic 8 (call for protest)
(i) topic 9 (advertise protest)
(j) topic 10 (call for protest)
Latent States

(a) State 1 (advertise protest)

(b) State 2 (describe protest)

(c) State 3 (advertise protest)

(d) State 4 (describe protest)

(e) State 5 (call for protest)
Multivariate Gaussian of Burstiness

(a) State 1 (advertise protest)  (b) State 2 (describe protest)  (c) State 3 (advertise protest)

(d) State 4 (describe protest)  (e) State 5 (call for protest)
Modeling Mass Protest Adoption in Social Network Communities using Geometric Brownian Motion [Fang et al., 2014]

Goal:
• Model the growth of protest participants within a social network.
• Understand the underlying social network and structural dynamics.
Bi-space: propagation via mention network and other latent network

Latent Space

Mentions network

(SEED QUERY)
Protest, march, demonstration ...

We consider the mentions network to be stable
Twitter mentioning propagation model

Trust threshold by Brownian distance

Brownian distance defined by mentioning frequency:

\[ d_{ij} = \frac{1}{(\omega_{ij} + 1)(\omega_{ji} + 1)^{\gamma}(\eta_{ij} + 1)^{\gamma}}, \gamma \geq 1 \]

\[ eg : d_{14} = \frac{1}{(1+1)(6+1)(2+1)}, \gamma = 1 \]

Trust threshold \( d_{ij} \) for each pair of nodes i and j.

Trust Function

If close friend, trusting his information need less time;
If not close, then more time needed.

The curve is defined by Geometric Brownian Motion process:

\[ S_t = S_0 \exp \left( \left( \mu - \frac{\sigma^2}{2} \right) t + \sigma W_t \right) \]

Trust the information \( \rightarrow \) Forward the tweet

Simulate the tweet forwarding process
We assume the probability of the number of newly infected users, $X$, in a given time interval satisfies the Poisson distribution:

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$
Simulation results

Bispace simulation value = Poisson value + GBM value
Simulation results: More cases

- **Mexican Yosoy132 protest**
- **Colombian anti-government protest**
- **Mexican teachers’ protest**
- **Accuracy for 7 protest events**
### GBM simulation results for teacher protest events on Sep 2, 2013

<table>
<thead>
<tr>
<th></th>
<th>Average degree</th>
<th>Diameter</th>
<th>Graph density</th>
<th>Connected components</th>
<th>Average clustering coefficient</th>
<th>Average path length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>1.791</td>
<td>11</td>
<td>0.002</td>
<td>183</td>
<td>0.083</td>
<td>4.786</td>
</tr>
<tr>
<td>Ground truth</td>
<td>1.726</td>
<td>18</td>
<td>0.002</td>
<td>204</td>
<td>0.008</td>
<td>6.261</td>
</tr>
</tbody>
</table>
Taxonomy

- Societal event forecasting
  - Temporal event forecasting
    - Discriminative Learning
      - All-to-all
      - All-to-one
      - One-to-one
    - Generative / Mechanistic Learning
      - Data-driven
      - Data-driven + Mechanistic
  - Spatiotemporal event forecasting
    - Ensemble Learning
      - Generative
      - Mechanistic
Due to the complexity of the societal phenomena.

- Each data source may only cover one part
- Each model may only explain a portion of the truth
- Some truth are unobservable.

Ensemble learning:

- Leverage the complementary strength of different models
- Sufficiently utilize different data sources in modeling different phenomena
References


Taxonomy

Societal event forecasting

Temporal event forecasting

Discriminative Learning
- All-to-all
- All-to-one
- One-to-one

Generative / Mechanistic Learning
- Data-driven
- Data-driven+Mechanistic

Spatiotemporal event forecasting

Ensemble Learning

Generative
Mechanistic
SimNest: Social Media Nested Epidemic Simulation via Online Semi-supervised Deep Learning

[Zhao, et al., ICDM’15]

• **Goal**: Utilize social media data and disease mechanism to model the underlying influenza epidemics progression.

• **Model characteristics**:
  – Ensembles of Data-driven and Mechanistic Models
  – Online Learning
1. Model the following mechanisms
   a. Demographics and social contact network
   b. Disease progression: SEIR model
   c. Interventions

   - School Closure
   - Vaccination
   - Isolation

2. Tune parameters against surveillance data

3. Run simulation model
Epidemics Modeling (Category 1): Computational Epidemiology

- Challenges
  - Challenge 1: Coarse-grained surveillance data
  - Challenge 2: Dynamics of contact networks
  - Challenge 3: Poor timeliness
  - Surveillance data is at least one week behind.

This year much more people get flu shot

Peter moved out to another city because he lost job.

Jim is suddenly on vacation.
Epidemics Modeling (Category 2): Data-driven Techniques on Social Media

- Fast monitoring real-time epidemics
  - Temporally fine-grained
  - No delay

- Individual-wise health condition mining
  1. Identify the response to flu
     - Avoid crowds
     - Get flu shot
     - ... in flu season, What Peter will do?
  2. Identify the individual’s disease progression
     - Feel I’m getting flu
     - 3rd day in the bed
     - Maybe it’s time back to work
Epidemics Modeling (Category 2): Data-driven Techniques on Social Media

Have No Idea of the Underlying Mechanism

Challenge: Real Mechanism is hidden to social media

What is the real disease contact network?
What is diffusion process of epidemics?
What is the consequence if someone took vaccine?
Any influence on infectivity if someone has summer holiday?
Motivations

Computational Epidemiology

• Advantages:
  – Mechanism on disease progression
  – Mechanism on disease diffusion
  – Consideration on interventions

• Drawbacks:
  – Temporally coarse-grained
  – Spatially coarse-grained
  – Poor dynamics in social contact network
  – One week delay

Social Media Mining

• Advantages:
  – Temporally fine-grained
  – Spatially fine-grained
  – Change in social contact network is observable in real time
  – No time delay

• Drawbacks:
  – No mechanism on disease progression
  – No mechanism on disease diffusion
  – No consideration on interventions

Combine
Idea

Timely and fine-grained observations

Mechanisms of epidemics diffusion

Combine
Model: Overview

- B. Bispace Inconsistency Loss
  - Date t, t+1, t+2, t+3, t+4, t+5

- C. Infectious Period Loss

- D. Temporal Pattern Loss

- A. Supervised Loss
  - Health State Layer
  - Abstract Semantics Layer
  - Text Layer
  - a keyword

Legend:
- : Healthy
- : Infectious
Model: Overview

Our objective: Minimize loss

$$\min \mathcal{L} = \min \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_a$$
The Proposed Model

• Learn a mapping: \( f_W(X_{u,t}) : X_{u,t} \rightarrow Y_{u,t} \)

• Minimize supervised Loss: \( \mathcal{L}_A = \min_W \sum_u \sum_{t} \left\| f_W(X_{u,t}) - Y_{u,t} \right\|^2 \)

• Minimize bi-space inconsistency: \( \mathcal{L}_B = \min_{\Theta, W, \lambda_1} \sum_{l,t} \left\| \lambda_1 \sum_v Q_{v,t} - \sum_u f_W(X_{u,t}) \right\|^2 \)

• Maximize the likelihood of infectious period distribution: Online training by alternating optimization

• Health stage should be consecutive:

\[
\mathcal{L}_D = \min_W \sum_u \sum_t \left\| f_W(X_{u,t}) - f_W(X_{u,t+1}) \right\|^2
\]
Online Training Algorithm

• Objective function:

\[ \mathcal{L} = \mathcal{L}_A(Y_1, X_1, W) + \mathcal{L}_B(X_2, G, p_E, p_I, W) + \mathcal{L}_C(X_2, p_I, W) + \mathcal{L}_D(X_2, W) \]

• Alternating optimization:
  – Solving for \( W \), fix others.
    • Stochastic Gradient Descent
  – Solving for \( \Theta \), fix others.
    • Nelder-Mead method.
  – Solving for \( p_I, \lambda_1 \).
Model Extensions

1. Consider dynamics of contact network

Dynamically adjust the transmissibility:
\[ p(W(v_1, v_2)) \]

Figure 3: Counts of Twitter users in Virginia who got flu shot

2. Consider heterogeneous surveillance

- Loss:
\[ \mathcal{L}_E = \min_{W, \lambda_2} \sum_{i} \left\| \lambda_2 (a_e - a_s + 1) \right\|_2 \sum_{l, t = a_s}^{L, a_e} \sum_{u} f_W(X_{u, t}) - C(i) \right\|^2 \]

- Scaling down time frame:
\[ T' = \lceil T \cdot \tau' / \tau \rceil, \quad a_s = \lceil i \cdot \tau' / \tau \rceil, \quad a_e = \lceil (i + 1) \cdot \tau' / \tau \rceil - 1 \]
Experiments: Dataset

• Dataset:
  – Twitter: Year 2011 ~ Year 2015 in the US.

<table>
<thead>
<tr>
<th></th>
<th>Demographics</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>population size</td>
<td>#connections</td>
</tr>
<tr>
<td>CT</td>
<td>3,518,288</td>
<td>175,866,264</td>
</tr>
<tr>
<td>DC</td>
<td>599,657</td>
<td>19,984,180</td>
</tr>
<tr>
<td>MA</td>
<td>6,593,587</td>
<td>332,194,314</td>
</tr>
<tr>
<td>MD</td>
<td>5,699,478</td>
<td>285,159,648</td>
</tr>
<tr>
<td>VA</td>
<td>7,882,590</td>
<td>407,976,012</td>
</tr>
</tbody>
</table>

Connecticut (CT), Massachusetts (MA), Maryland (MD), and Virginia (VA), and the District of Columbia (DC)
Experiments: Label and Metrics

• Label:
  – influenza statistics reported by the Centers for Disease Control and Prevention (CDC).
  – The CDC weekly publishes the percentage of the number of physician visits related to influenza-like illness (ILI) within each major region in the United States.

• Metrics:
  – Lead time: How much time the output is ahead of the input.
  – Mean squared error (MSE)
  – Pearson correlation
  – P-value
  – Peak time error: Error of the predicted time of peak value
Experiments: Comparison Methods

- **social media mining methods:**
  - Linear Autoregressive Exogenous model (LinARX)
  - Logistic Autoregressive Exogenous model (LogARX)
  - Simple Linear Regression model (simpleLinReg)
  - Multi-variable linear regression model (multiLinReg)

- **computational epidemiology methods:**
  - SEIR
  - EpiFast

- **Detailed parameter settings:**
Influenza Epidemic Forecasting Performance

Training set: Tweets in Aug 2011 ~ Jul 2012 in the US.
Test set: Tweets Aug 2012 ~ Jul 2014 in the US.

Label set: CDC surveillance data

P-value: likelihood that the null hypothesis is true.
Pearson correlation: Strength of linear relation
Lead time: How much time the output is ahead of the input.
Taxonomy

Societal event forecasting

Temporal event forecasting

Discriminative Learning
- All-to-all
- All-to-one
- One-to-one

Generative / Mechanistic Learning
- Data-driven
- Data-driven+Mechanistic

Spatiotemporal event forecasting

Ensemble Learning
- Mechanistic
Ensembles of Complementary Models
(Ramakrishnan, et al., KDD’14, Muthiah, et al., KDD’16)

Data Sources
Tor, RSS, Twitter, YouTube, WJCEWS, The GDELT Project

Planned protest detection
Trabajadores de Southern Perú amenazan con huelga indefinida desde enero

Causal-based prediction

Cascade regression
(tracks online recruitment and viral spread)

Propagation-Modeling based

Dynamic query expansion
(automatically detects emerging keyword groups)

Volume-based model
(LASSO approach)

Discriminative learning-based

Baseline model
(GSR-based)

Historical prior

Each model has high precision
Different models are complementary

high precision & recall


Thank you