Modeling Spread of Disease from Social Interactions

Adam Sadilek
Henry Kautz
Vincent Silenzio
The Data

• New York City
The Data

- New York City
- 16M tweets over a month
The Data

- New York City
- 16M tweets over a month
- 0.6M unique users
The Data

- 6K geo-active users
The Data

- 6K **geo-active** users
- 2.5M **tweets** by geo-active users
The Data

- 6K geo-active users
- 2.5M tweets by geo-active users
- 103K “follows” relationships
The Data

• 6K geo-active users
• 2.5M tweets by geo-active users
• 103K “follows” relationships
• 32K “friends” relationships
“feeling horrible.. this flu had me up till 3am..... Jesus!”
Spread of Disease

- **Identify** people with **symptoms**
- **Quantify** the impact of:
  - Co-location
  - Social ties
- **Model contagion** with fine granularity
“feeling sick”
“feeling sick”
“feeling sick”
“feeling sick”
“feeling sick”

Diagram showing nodes and connections labeled as follows:
- Feeling
- Sick
- House

The diagram indicates relationships and positions in a network or graph representation.
Cascade SVM
Cascade SVM

Corpus of 5,128 tweets labeled by human workers

Corpus of 1.6 million machine-labeled tweets

Random sample of 200 million tweets

Corpus of "other" tweets

Corpus of "sick" tweets

Final corpus

Training
### SVM Features

<table>
<thead>
<tr>
<th>Positive Features</th>
<th>Weight</th>
<th>Negative Features</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>sick</td>
<td>0.9579</td>
<td>sick of</td>
<td>−0.4005</td>
</tr>
<tr>
<td>headache</td>
<td>0.5249</td>
<td>you</td>
<td>−0.3662</td>
</tr>
<tr>
<td>flu</td>
<td>0.5051</td>
<td>lol</td>
<td>−0.3017</td>
</tr>
<tr>
<td>fever</td>
<td>0.3879</td>
<td>love</td>
<td>−0.1753</td>
</tr>
<tr>
<td>feel</td>
<td>0.3451</td>
<td>i feel your</td>
<td>−0.1416</td>
</tr>
<tr>
<td>coughing</td>
<td>0.2917</td>
<td>so sick of</td>
<td>−0.0887</td>
</tr>
<tr>
<td>being sick</td>
<td>0.1919</td>
<td>bieber fever</td>
<td>−0.1026</td>
</tr>
<tr>
<td>better</td>
<td>0.1988</td>
<td>smoking</td>
<td>−0.0980</td>
</tr>
<tr>
<td>being</td>
<td>0.1943</td>
<td>i’m sick of</td>
<td>−0.0894</td>
</tr>
<tr>
<td>stomach</td>
<td>0.1703</td>
<td>pressure</td>
<td>−0.0837</td>
</tr>
<tr>
<td>and my</td>
<td>0.1687</td>
<td>massage</td>
<td>−0.0726</td>
</tr>
<tr>
<td>infection</td>
<td>0.1686</td>
<td>i love</td>
<td>−0.0719</td>
</tr>
<tr>
<td>morning</td>
<td>0.1647</td>
<td>pregnant</td>
<td>−0.0639</td>
</tr>
</tbody>
</table>
SVM Robustness

- Held-out set of 700,000 tweets
- 0.98 precision
- 0.97 recall
CherishmySTYLE [1.37667832]
On my way to work! I feel soooooo sick!!!
I better get it together b4 tonight!!!
lafemmefatalexx [0.74721562]
My feet hurt soooooo baddd I need to be carried home waaaaahhhhhhhhhhhhhhhhhhh never wearing heels again.
Impact of Co-Location

Conditional probability of getting sick at \( t + 1 \)

- 1 hour time window (\( T=1h \))
- 4 hour time window (\( T=4h \))
- 12 hour time window (\( T=12h \))

Prior probability of being sick

\[ f(x) = 0.013e^{(0.055x)} \]
\[ f(x) = 0.002e^{(0.054x)} \]
\[ f(x) = 0.001e^{(0.055x)} \]

Number of estimated encounters with sick individuals at time \( t \)
Impact of Friendships

\[ f(x) = 0.003 \exp(0.413x) \]

- Red: Probability of getting sick at \( t+1 \) given \( n \) friends are sick at \( t \)
- Blue: Probability of getting sick given having \( n \) friends (any)
- Green: Probability of getting sick at \( t+1 \) given \( n \) unencountred friends are sick at \( t \)
- Dashed: Prior probability of being sick

Number of friends (\( n \))

Conditional probability of getting sick
Twitter Health

• Aggregate accuracy comparable with:
  • Google Flu Trends (R = 0.73)
  • CDC statistics

+ we can model fine-grained interactions between specific individuals
Health Insights

[Snow, 1855]
CON EDISON - HUDSON AVE STATION:
1 HUDSON AVE BROOKLYN NY 11201
NAICS: 22133
Steam and Air-Conditioning Supply

Annual Air Emissions

<table>
<thead>
<tr>
<th>Emissions Type</th>
<th>2008 Emissions Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon Monoxide</td>
<td>47</td>
</tr>
<tr>
<td>Lead</td>
<td>0</td>
</tr>
<tr>
<td>Nitrogen Oxides</td>
<td>318</td>
</tr>
<tr>
<td>Particulate Matter 10</td>
<td>27</td>
</tr>
<tr>
<td>Particulate Matter 2.5</td>
<td>27</td>
</tr>
<tr>
<td>Sulfur Dioxide</td>
<td>416</td>
</tr>
<tr>
<td>Volatile Organic Compounds</td>
<td>7</td>
</tr>
</tbody>
</table>
Summary
Summary

- **Quantify** impact of physical *encounters* and *social ties* on public *health*
  - Real world observed via Twitter
  - Fine granularity
  - Real-time
  - No active user participation
  - High accuracy
Current Work
Refining Epidemiological Models
Health Prediction

\[ h_t - 1 \quad \rightarrow \quad h_t \quad \rightarrow \quad h_t + 1 \quad \rightarrow \quad \ldots \]
Health Prediction

... $h_{t-1}$ $h_t$ $h_{t+1}$ ...

$O_{t-1}$ $O_t$ $O_{t+1}$...
Health Prediction

- Conditional probability of getting sick at time $t+1$ given $h_t$
  - 1 hour time window ($T=1h$)
  - 4 hour time window ($T=4h$)
  - 12 hour time window ($T=12h$)

Prior probability of being sick

$f(x) = 0.013e^{0.055x}$

Number of estimated encounters with sick individuals at time $t$
Health Prediction

\[
h_t = h_{t-1} + O_t
\]

Number of estimated encounters with sick individuals at time \(t\):

\[
f(x) = 0.002e^{(0.054x)}
\]

Conditional probability of getting sick at \(t+1\) given \(n\) unencountred friends are sick at \(t\):

\[
f(x) = 0.003e^{(0.443x)}
\]

Prior probability of being sick:

\[
f(x) = 0.013e^{(0.055x)}
\]

Number of friends (\(n\)):

\[
f(x) = 0.001e^{(0.055x)}
\]

Prob. of getting sick at \(t+1\) given \(n\) friends are sick at \(t\):

\[
f(x) = 0.002e^{(0.054x)}
\]

Prob. of getting sick given having \(n\) friends (any):

\[
f(x) = 0.003e^{(0.443x)}
\]

Prob. of getting sick at \(t+1\) given \(n\) unencountred friends are sick at \(t\):

\[
f(x) = 0.001e^{(0.055x)}
\]

Prior probability of being sick:

\[
f(x) = 0.013e^{(0.055x)}
\]
Adam Sadilek, Henry Kautz & Vincent Silenzio: Predicting Disease Transmission from Geo-Tagged Micro-Blog Data.

AAAI 2012 in Toronto
Demo

bit.ly/twitterHealth
Modeling Spread of Disease from Social Interactions

bit.ly/twitterHealth

Adam Sadilek
Henry Kautz
Vincent Silenzio
Modeling “Hidden” Population
Modeling “Hidden” Population
Modeling “Hidden” Population

Terrible headache!