Syndromic Surveillance Systems: Overview and the BioPortal System

Hsinchun Chen, Ph.D.
Artificial Intelligence Lab, U. of Arizona
NSF BioPortal Center

陳炘鈞, 亞利桑納大學人工智慧實驗室任
NCTU ➔ NYU ➔ Arizona

Digital Library ➔ Biomedical Informatics ➔ Intelligence and Security Informatics

COPLINK ➔ BorderSafe ➔ Dark Web ➔ BioPortal

NSF ➔ DOD ➔ DOJ ➔ DHS ➔ CIA ➔ NIH/NLM/NCI
Medical Informatics: The computational, algorithmic, database and information-centric approach to the study of medical and health care problems.

From Medical Informatics to Infectious Disease Informatics
Syndromic Surveillance

- A syndrome is a set of symptoms or conditions that occur together and suggest the presence of a certain disease or an increased chance of developing the disease (from NIH/NLM)

- Syndromic surveillance is based on health-related data that precede diagnosis and signals a sufficient probability of a case or an outbreak to warrant further public health response (from CDC)
  - Targeting investigation of potential cases
  - Detecting outbreaks associated with bioterrorism
Syndromic Surveillance Data Sources in Different Stages of Developing a Disease

Reproduced from Mandl et. al. (2004)
# Syndromic Surveillance System Survey

<table>
<thead>
<tr>
<th>Projects</th>
<th>User population</th>
<th>Stakeholders</th>
</tr>
</thead>
<tbody>
<tr>
<td>RODS</td>
<td>-Pennsylvania, Utah, Ohio, New Jersey, Michigan etc</td>
<td>RODS laboratory, U of Pittsburgh</td>
</tr>
<tr>
<td></td>
<td>-418 facilities connected to RODS</td>
<td></td>
</tr>
<tr>
<td>STEM</td>
<td>N/A</td>
<td>IBM</td>
</tr>
<tr>
<td>ESSENCE II</td>
<td>300 world wide DOD medical facilities</td>
<td>DoD</td>
</tr>
<tr>
<td>EARS</td>
<td>-Various city, county, and state public health officials in the United States and abroad of US</td>
<td>CDC</td>
</tr>
<tr>
<td>BioSense</td>
<td>Various city, county, and state public health officials in the United States and abroad of US</td>
<td>CDC</td>
</tr>
<tr>
<td>RSVP</td>
<td>Rapid Syndrome Validation Project; Kansas, NM</td>
<td>Sandia NL, NM</td>
</tr>
<tr>
<td>BioPortal</td>
<td>NY, CA, Kansas, AZ, Taiwan</td>
<td>U of Arizona</td>
</tr>
</tbody>
</table>
## Sample Systems and Data Sources Utilized

<table>
<thead>
<tr>
<th>Projects</th>
<th>Data sources/Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>RODS</td>
<td>- Chief complaints (CC); OTC medication sales</td>
</tr>
<tr>
<td></td>
<td>- Free-text Bayesian disease classification</td>
</tr>
<tr>
<td>STEM</td>
<td>- Simulated disease data</td>
</tr>
<tr>
<td></td>
<td>- Disease modeling and visualization, SIR</td>
</tr>
<tr>
<td>ESSENCE II</td>
<td>- Military ambulatory visits; CC; Absenteeism data</td>
</tr>
<tr>
<td>EARS</td>
<td>- 911 calls; CC; Absenteeism; OTC drug sales</td>
</tr>
<tr>
<td></td>
<td>- Human-developed CC classification rules</td>
</tr>
<tr>
<td>BioSense</td>
<td>- City/state generated geocoded clinical data</td>
</tr>
<tr>
<td></td>
<td>- Graphing/mapping displays</td>
</tr>
<tr>
<td>RSVP</td>
<td>- Clinical and demographic data</td>
</tr>
<tr>
<td></td>
<td>- PDA entry and access</td>
</tr>
<tr>
<td>BioPortal</td>
<td>- Geo-coded clinical data; Gemonic sequences; Multilingual CC</td>
</tr>
<tr>
<td></td>
<td>- Real-time access and visualization; Web based hotspot analysis; Sequence visualization; Multilingual ontology-based CC classification</td>
</tr>
</tbody>
</table>
BioPortal: Overview, WNV, BOT
Project Background

- In September, 2002, representatives of 18 different agencies, including DOD, DOE, DOJ, DHS, NIH/NLM, CDC, CIA, NSF, and NASA, are convened to discuss “disease surveillance”

- AI Lab was chosen to be the technical integrator to work with New York and California States to develop a prototype system targeting West Nile Virus and Botulism
BioPortal Project Goals

- Demonstrate and assess the technical feasibility and scalability of an infectious disease information sharing (across species and jurisdictions), alerting, and analysis framework.

- Develop and assess advanced data mining and visualization techniques for infectious disease data analysis and predictive modeling.

- Identify important technical and policy-related challenges in developing a national infectious disease information infrastructure.
Data Ingest Control Module
Cleansing / Normalization

Portal Data Store (MS SQL 2000)

Adaptor

PHINMS Network

NYSDOH

Adaptor

XML/HL7 Network

CADHS

Adaptor

New

SSL/RSA

SSL/RSA

SSL/RSA

Info-Sharing Infrastructure Design
Data Access Infrastructure Design

Browser (IE/Mozilla/…) → SSL connection

Web Server (Tomcat 4.21 / Struts 1.2)

Data Search and Query
Spatial-Temporal Visualization
Analysis / Prediction
HAN or Personal Alert Management
Dataset Privileges Management

User Access Control API (Java)

Data Store
(MS SQL 2000)

Public health professionals, researchers, policy makers, law enforcement agencies & other users

Access Privilege Def.

Dataset Privileges Management
Spatial-Temporal Visualization

- Integrates four visualization techniques
  - GIS View
  - Periodic Pattern View
  - Timeline View
  - Central Time Slider
- Visualizes the events in multiple dimensions to identify hidden patterns
  - Spatial
  - Temporal
  - Hotspot analysis
  - Phylogenetic (planned)
BioPortal Prototype Systems

The BioPortal is a system for managing and visualizing biological data. It includes tools for searching and selecting data sets, as well as features for mapping and analyzing data. The BioPortal also provides access to various databases, such as West Nile Virus, Botulism, Foot & Mouth Disease, and BioWatch. The FMD BioPortal is specifically designed for the study of Foot and Mouth Disease, with tools for analyzing livestock data.
Outbreak Detection & Hotspot Analysis

- **Hotspot** is a condition indicating some form of clustering in a spatial and temporal distribution (Rogerson & Sun 2001; Theophilides et. al. 2003; Patil & Tailie 2004; Zeng et. al. 2004; Chang et. al. 2005)

For WNV, localized clusters of dead birds typically identify high-risk disease areas (Gotham et. al. 2001); *automatic* detection of dead bird clusters can help predict disease outbreaks and allocate prevention/control resources effectively
Retrospective Hotspot Analysis
Problem Statement
Risk-Adjusted Support Vector Clustering (RSVC)

- Estimate baseline density
- Minimum sphere
- Split into several clusters
- High baseline density makes two points far apart in feature space
- Feature space
- Case points
- Baseline points

High baseline density makes two points far apart in feature space.
Study II: NY WNV

- On May 26, 2002, the first dead bird with WNV was found in NY
  - Based on NY’s test dataset

140 records
March 5
baseline

224 records
May 26
new cases
July 2
Dead Bird Hotspots Identified

Clusters using SatScan, RNNH, RSVC
### Select Maps to add to Visualization

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>☑</td>
<td>Precipitations</td>
</tr>
<tr>
<td></td>
<td>Census 2000</td>
</tr>
<tr>
<td>☑</td>
<td>Census 2000 New York Detail</td>
</tr>
<tr>
<td>☑</td>
<td>1999 California Population</td>
</tr>
<tr>
<td></td>
<td>Labor</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
</tr>
<tr>
<td></td>
<td>Vegetation</td>
</tr>
<tr>
<td></td>
<td>Mexican Rivers</td>
</tr>
<tr>
<td>☑</td>
<td>Detailed Rivers</td>
</tr>
<tr>
<td>☑</td>
<td>Major Lakes</td>
</tr>
<tr>
<td>☑</td>
<td>Major Rivers</td>
</tr>
<tr>
<td></td>
<td>Airports</td>
</tr>
<tr>
<td></td>
<td>Cities and Towns</td>
</tr>
<tr>
<td></td>
<td>Urban Areas</td>
</tr>
</tbody>
</table>

Select NY / CA population, river and lakes

Start STV

Select background maps
NY dead bird temporal distribution pattern

1 year window in 3 year span

Concentrated in May / Jun

Similar time pattern

Year 2001 data

Spatial distribution pattern

2 weeks window
Enable population map

Overlay population map

Dead bird cases distribute along populated areas near Hudson river
BioPortal HotSpot Analysis: RSVC, SaTScan, and CrimeStat Integrated (first visual, real-time hotspot analysis system for disease surveillance)

RSVC is a clustering-based, spatio-temporal hotspot analysis algorithm developed at the Artificial Intelligence Laboratory of the University of Arizona. It combines the power of support vector machines (SVM) with the risk adjustment approach from CrimeStat. It clusters points with consideration for baseline information (data under normal conditions) to find the emerging at risk area.

Perform hotspot analysis using SaTScan Ver. 5.1

In this version of SaTScan, which uses the Bernoulli method, the distribution of baseline observations (or controls) and the distribution of new observations (or cases) are compared and circular clusters are identified where the proportion of new observations is significantly higher than the proportion of new observations outside the circle. Click here to find more information about SaTScan and the spatial scan statistic.

P-Value: 10 percent

Perform hotspot analysis using CrimeStat® III

CrimeStat® III offers several hotspot analysis algorithms. BioPortal uses RNNH.

The nearest neighbor hierarchical clustering (NNH) routine in CrimeStat identifies groups of incidents that are spatially close. It clusters points together and then proceeds to group the clusters together. The risk-adjusted nearest neighbor hierarchical clustering routine (RNNH) combines the hierarchical clustering capabilities with kernel density interpolation techniques.

Minimum number of points per cluster: 10 (within one standard deviation from the ellipse)
Hotspot Analysis-Enabled STV

Select hotspot to highlight case points.
BioPortal - FMD
International FMD BioPortal

- Real time web-based situational awareness of FMD outbreaks worldwide through the establishment of an international information technology system.
- FMDv characterization at the genomic level integrated with associated epidemiological information and modeling tools to forecast national, regional and/or international spread and the prospect of importation into the US and the rest of North America.
- Web-based crisis management of resources—facilities, personnel, diagnostics, and therapeutics.
Global foot-and-mouth disease surveillance

Dr. Mark Thurmond

FMD Lab, Center for Animal Disease Modeling and Surveillance, School of Veterinary Medicine, University of California, Davis, CA 95616
Preliminary Global FMD Dataset

- **Provider**: UC Davis FMD Lab
- **Information sources**: reference labs and OIE
- **Coverage**: 28 countries globally
- **Time span**: May, 1905 – March, 2005
- **Dataset size**: 30,000+ records of which 6789 records are complete
- **Host species**: Cattle, Caprine, Ovine, Bovine, Swine, NK, Elephant, Buffalo, Sheep, Camelidae, Goat

**Regionwise Distribution of FMD Data**
- South America: 66%
- Africa: 4%
- Middle East Asia: 4%
- Europe: 14%
- Central and South Asia: 15%

**Host Species Distribution**
- Ovine: 37%
- Bovine: 37%
- Caprine: 4%
- Cattle: 5%
- Sheep: 11%
- Swine: 3%
- Buffaloes: 0%
- Goats: 0%
- Elephant: 0%
- Camelidae: 0%
Global FMD Coverage in BioPortal
FMD BioPortal link to Google Earth
<table>
<thead>
<tr>
<th>Sample #</th>
<th>Country</th>
<th>Serotype</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>IND/1994/00011</td>
<td>INDIA</td>
<td>O</td>
<td>1995-12-04</td>
</tr>
<tr>
<td>IND/1994/00025</td>
<td>INDIA</td>
<td>ASIA-1</td>
<td>1995-12-04</td>
</tr>
<tr>
<td>IND/1990/00005</td>
<td>INDIA</td>
<td>O</td>
<td>1995-12-04</td>
</tr>
<tr>
<td>IND/1991/00002</td>
<td>INDIA</td>
<td>ASIA-1</td>
<td>1995-12-04</td>
</tr>
</tbody>
</table>

Directions: To here - From here
Hotspot analysis

Focus on Africa

Use ~1999 as baseline distribution; 2000+ as observing target
Hotspot analysis
Hotspot analysis

Hotspot

Mixed Area
International FMD News

- **Provider**: UC Davis FMD Lab
- **Information sources**: Google, Yahoo, and open Internet sources
- **Time span**: Oct 4, 2004 – present (real-time messaging under development)
- **Data size**: 460 events (6/21/05)
- **Coverage**: 51 countries (Africa:11, Asia:16, Europe:12, Americas:12)
Searching FMD News

- [ ] http://fmd.ucdavis.edu/
- [ ] Searchable by
  - Date range
  - Country
  - Keyword
Visualizing FMD News on BioPortal
FMD Genetic Visualization

- Goal: Extend STV to incorporate 3rd dimension, phylogenetic distance
  - Include a phylogenetic tree.
  - Identify phylogenetic groups and color-code the isolate points on the map.
  - Leverage available NCBI tools such as BLAST.

- Proof of concept: SAT 2 & 3 analysis
  - **Data**: 54 partial DNA sequence records in South Africa received from UC Davis FMD Lab (Bastos, A.D. *et al.* 2000, 2003)
  - **Date range**: 1978-1998
  - **Countries covered**: South Africa, Zimbabwe, Zambia, Namibia, Botswana
Textual View of Gene Sequence

2015/10/12
FMDV Genomics BioPortal (under development)
This is full view of the phylogenetic tree.

The RED ring is the threshold circle.

This value is the genetic distance between the threshold and the root.

Each label is an accession number (selectable via mouse).
As the threshold circle is pulled inwards, the leaves falling outside the threshold are grouped into the color of their parent in the tree.
When the circle is moved to the root (the distance is 0.00) position, all the nodes are grouped into one color, i.e., the color of the root.
The nodes on GIS map acquire the corresponding color from the phylogenetic tree.
Select any accession on the phylogenetic tree. The corresponding node(s) on phylogenetic tree and the GIS map are highlighted.
FMD BioPortal activity

Launched: January 5, 2007

65 users from >15 countries

- Belgium, Brazil, Canada, France, Germany, Italy, India, Iran, Netherlands, Pakistan, Paraguay, South Africa, Sweden, U.S., U.K.
- Research institutes, diagnostic labs, government and international agencies and organizations, universities (7)

Applications:

- Promed
- Bioinformatics support to DHS Plum Island
- Teaching veterinary students
- FMD status evaluations and risk assessments for USDA
- Research on FMD in southern Africa
- Teaching at US Army Command and General Staff College
BioPortal – Arizona Syndromic Surveillance
Chief Complaints As a Data Source

- Chief complaints (CCs) are short free-text phrases entered by triage practitioners describing reasons for patients’ ER visit
  - *Examples:* lt foot pain [left foot pain]; cp [chest pain]; sob [shortness of breath]; so [should be ‘sob’]; poss uti [possibly urinary tract infection]

- Advantages of using CCs for surveillance purposes
  - **Timeliness:** Diagnose results are on average 6 hours slower than CCs
  - **Availability and low-cost:** Most hospitals have free-text CCs available in electronic form
# Existing CC Classification Methods

<table>
<thead>
<tr>
<th>Classification Method</th>
<th>Systems</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword Match + Synonym List + Mapping Rules</td>
<td>DOHMH (NY City), EARS</td>
<td>Mikosz et. al. (2004)</td>
</tr>
<tr>
<td>Naïve Bayesian</td>
<td>RODS</td>
<td>Olszewski (2003), Ivanov et. al (2002)</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>N/A</td>
<td>Chapman et. al. (2004)</td>
</tr>
</tbody>
</table>
## Syndromic Categories in Different Systems

<table>
<thead>
<tr>
<th>System</th>
<th># Sdms</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC (2002)</td>
<td>11</td>
<td>Botulism, Hemorrhagic, Lymphadenitis, Cutaneous Lesion, Gastrointestinal, Respiratory, Neurological, Rash, Specific Infection, Fever, Severe Illness or Death</td>
</tr>
<tr>
<td>EARS</td>
<td>41</td>
<td>Lower Resp., Upper Resp., Neuro, Febrile, Poison, Hemorrhage, Botulinic, Rash, Fever, etc. (41 categories)</td>
</tr>
<tr>
<td>RODS</td>
<td>8</td>
<td>Gastrointestinal, Constitutional, Respiratory, Rash, Hemorrhagic, Botulinic, Neurological, Other</td>
</tr>
<tr>
<td>ESSENCE</td>
<td>8</td>
<td>Death, Gastr, Neuro, Rash, Respi, Sepsi, Unspe, Other</td>
</tr>
</tbody>
</table>
Overall System Design

Stage 1

Chief Complaints

CC Standardization

EMT-P

UMLS Concepts
Synonym List

Stage 2

Symptom Grouping

Weighted Semantic Similarity Score

UMLS Ontology
Symptom Grouping Table

EMT-P

Stage 3

Syndrome Classification

JESS

EARS Syndrome Rules

EARS Symptom Table
A Stage 2 Example: CC Concepts ➔ Symptom Group Concepts

- **Symptom Group Concepts**
  - **Blood**
    - In urine
  - **UMLS**
    - **coagulopathy**
      - 4
    - **purpura**
      - 5
    - **ecchymosis**
      - 6
      - bleeding = $\frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 0.62$
    - **ureteral stone**
      - 5
    - **coma**
      - 5
      - other = $\frac{1}{5} = 0.2$
      - coma = $\frac{1}{5} = 0.2$
      - dead = $\frac{1}{5} = 0.2$
    - **out pass**
      - 5
      - altered_mental_status = $\frac{1}{5} = 0.2$
System Benchmarks

- Both RODS (Tsui et. al., 2003) and EARS (CDC, 2006; Hutwagner et. al., 2003) serve as the benchmarks
  - RODS uses supervised learning method
  - EARS uses rule-based method
  - Both system are available for test
- Performance criteria are calculated by comparing system outputs with the gold standard
### Syndromic Categories in Different Systems

<table>
<thead>
<tr>
<th></th>
<th>EARS</th>
<th>RODS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botulism-like</td>
<td>s_botulism</td>
<td>Botulism-like</td>
</tr>
<tr>
<td>Constitutional</td>
<td>N/A</td>
<td>Constitutional</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>s_gastrointestinal, s_gicat</td>
<td>Gastrointestinal</td>
</tr>
<tr>
<td>Hemorrhagic</td>
<td>s_hemorrhagic</td>
<td>Hemorrhagic</td>
</tr>
<tr>
<td>Neurological</td>
<td>s_neurons, s_neurological</td>
<td>Neurological</td>
</tr>
<tr>
<td>Rash</td>
<td>s_rashcat</td>
<td>Rash</td>
</tr>
<tr>
<td>Respiratory</td>
<td>Upper Respiratory, Lower Respiratory</td>
<td>Respiratory</td>
</tr>
<tr>
<td>Upper Respiratory</td>
<td>s_upperresp, s_sb_upper_respiratory</td>
<td>N/A</td>
</tr>
<tr>
<td>Lower Respiratory</td>
<td>s_lowerresp, s_sb_lower_respiratory</td>
<td>N/A</td>
</tr>
<tr>
<td>Fever</td>
<td>s_fever, s_febrile</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Research Test Bed

- Training Dataset
  - Chief Complaints from a large hospital in Phoenix from Aug. 22, 2005 to Sep. 1, 2005
  - Total 2256 records

- Testing Dataset
  - Random sample of 1000 records from the same hospital during July 2005 to Nov. 2005
  - No overlap with training dataset
  - Generate the gold standard
Generating Gold Standard

- Three experts (two physicians and one nurse) were given a description of syndrome definition and 1,000 chief complaints
- The experts worked independently to assign CCs into syndromic categories
- Majority vote was used to determine syndromic assignments. Another physician reviewed CCs with three-way tie
- One CC can be assigned to more than one syndromic category
**Expert Agreement by Syndromic Category**

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botulism-like</td>
<td>0.22</td>
</tr>
<tr>
<td>Constitutional</td>
<td>0.24</td>
</tr>
<tr>
<td>Lower Respiratory</td>
<td>0.38</td>
</tr>
<tr>
<td>Fever</td>
<td>0.46</td>
</tr>
<tr>
<td>Neurological</td>
<td>0.64</td>
</tr>
<tr>
<td>Other</td>
<td>0.74</td>
</tr>
<tr>
<td>Upper Respiratory</td>
<td>0.77</td>
</tr>
<tr>
<td>Respiratory</td>
<td>0.80</td>
</tr>
<tr>
<td>Hemorrhagic</td>
<td>0.81</td>
</tr>
<tr>
<td>Rash</td>
<td>0.82</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>0.85</td>
</tr>
<tr>
<td>Overall</td>
<td>0.71</td>
</tr>
</tbody>
</table>

 Syndromic categories with kappa lower than 0.7 and Other were both excluded in the evaluation.
Performance Criteria

- **Sensitivity** (recall) = TP/(TP+FN)
- **Specificity** (negative recall) = TN/(FP+TN)
- **Precision** = TP/(TP+FP)
- **F-measure** = 2 * Precision * Recall / (Precision + Recall)

In the context of syndromic surveillance, sensitivity is more important than precision and specificity (Chapman, 2005). Thus, the F2-measure is used:

- F2 measure weights recall twice as much as precision.
- **F2-measure** = (1+2)*Precision * Recall / (2*Recall + Precision)

- Note: TP=True Positive, TN=True Negative
  FP=False Positive, FN=False Negative
## Comparing BioPortal to RODS

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>TP+FN</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained BioPortal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GI</td>
<td>124</td>
<td>91.41%</td>
<td>94.35%***</td>
<td>98.74%</td>
<td>0.93***</td>
<td>0.93***</td>
</tr>
<tr>
<td>HEMO</td>
<td>30</td>
<td>82.86%</td>
<td>96.67%***</td>
<td>99.38%</td>
<td>0.89**</td>
<td>0.92***</td>
</tr>
<tr>
<td>RASH</td>
<td>15</td>
<td>66.67%</td>
<td>66.67%**</td>
<td>99.49%</td>
<td>0.67*</td>
<td>0.67**</td>
</tr>
<tr>
<td>RESP</td>
<td>110</td>
<td>92.08%</td>
<td>84.55%****</td>
<td>99.10%</td>
<td>0.88***</td>
<td>0.87***</td>
</tr>
<tr>
<td>UPPER_RESP</td>
<td>43</td>
<td>80.43%</td>
<td>86.05%</td>
<td>99.06%</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>RODS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GI</td>
<td>124</td>
<td>89.89%</td>
<td>64.52%</td>
<td>98.97%</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>HEMO</td>
<td>30</td>
<td>90.91%</td>
<td>66.67%</td>
<td>99.79%*</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>RASH</td>
<td>15</td>
<td>58.33%</td>
<td>46.67%</td>
<td>99.49%</td>
<td>0.52</td>
<td>0.50</td>
</tr>
<tr>
<td>RESP</td>
<td>110</td>
<td>87.84%</td>
<td>59.09%</td>
<td>98.99%</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>UPPER_RESP</td>
<td>43</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

* p-value < 0.1 ** p-value < 0.05 *** p-value < 0.01

Statistical test is based on 2,500 bootstrapings.
Comparing BioPortal to EARS

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Trained BioPortal</th>
<th>EARS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP+FN</td>
<td>PPV</td>
</tr>
<tr>
<td>GI</td>
<td>124</td>
<td>91.41%</td>
</tr>
<tr>
<td>HEMO</td>
<td>30</td>
<td>82.86%</td>
</tr>
<tr>
<td>RASH</td>
<td>15</td>
<td>66.67%</td>
</tr>
<tr>
<td>RESP</td>
<td>110</td>
<td>92.08%</td>
</tr>
<tr>
<td>UPPER_RESP</td>
<td>43</td>
<td>80.4%***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>TP+FN</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI</td>
<td>124</td>
<td>93.75%*</td>
<td>72.58%</td>
<td>99.32%***</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>HEMO</td>
<td>30</td>
<td>100.00%***</td>
<td>33.33%</td>
<td>100.00%***</td>
<td>0.50</td>
<td>0.43</td>
</tr>
<tr>
<td>RASH</td>
<td>15</td>
<td>70.00%</td>
<td>46.67%</td>
<td>99.70%</td>
<td>0.56</td>
<td>0.53</td>
</tr>
<tr>
<td>RESP</td>
<td>110</td>
<td>90.36%</td>
<td>68.18%</td>
<td>99.10%</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>UPPER_RESP</td>
<td>43</td>
<td>58.70%</td>
<td>62.79%</td>
<td>98.01%</td>
<td>0.61</td>
<td>0.61</td>
</tr>
</tbody>
</table>

* p-value < 0.1 ** p-value < 0.05 *** p-value < 0.01

Statistical test is based on 2,500 bootstrapings.

2015/10/12
Conclusions

- Medical Ontology (UMLS) and Weighted Semantic Similarity Score can significantly help improve syndromic surveillance system performance.
- Rule-based approach can be easily adopted in different syndromic surveillance systems.
- Edit Distance can prove the handling of word variations in CCs.
BioPortal – Taiwan Syndromic Surveillance
Multi-lingual Chief Complaints: Chinese Example

- Data Characteristics:
  - Mixed expressions in both Chinese and English
    - 头痛; 头晕; FEVER; 腹痛; 噁心呕吐多次; 旅遊史 (無)
    - 車禍，導致左手背 A/W，疼痛不適，咳嗽有痰
    - 18% CC records from NTU Med. Center contain Chinese expressions.
    - Some hospitals have 100% CC records in Chinese (For example, 馬偕紀念醫院)

- Misspellings and typographic errors are not serious
Prevalence of Chinese Chief Complaints

<table>
<thead>
<tr>
<th></th>
<th># Records</th>
<th># Hospitals</th>
<th>% Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Center</td>
<td>222,893</td>
<td>17</td>
<td>52%</td>
</tr>
<tr>
<td>Regional Hospital</td>
<td>484,123</td>
<td>57</td>
<td>16%</td>
</tr>
<tr>
<td>Local Hospital</td>
<td>232,008</td>
<td>89</td>
<td>19%</td>
</tr>
<tr>
<td>Total</td>
<td>939,024</td>
<td>163</td>
<td>100%</td>
</tr>
</tbody>
</table>

- Medical Center: 馬偕紀念醫院 (100%), 台大醫院 (18%), 慈濟綜合醫院 (8%)
- Regional Hospital: 中興醫院 (99%), 國軍高雄總醫院 (87%), 嘉義榮民醫院 (72%), 光田綜合醫院 (50%), …, etc.
- Local Hospital: 縣立板橋醫院 (100%), 西園醫院 (93%), 灣橋榮民醫院 (88%), …, etc.
The Role of Chinese Chief Complaints in Syndromic Surveillance Systems

<table>
<thead>
<tr>
<th>Chinese Function</th>
<th>Symptom Related</th>
<th>Chinese Punctuation</th>
<th>Name Entity</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Chinese CCs</td>
<td>40.79%</td>
<td>20.32%</td>
<td>13.97%</td>
<td>24.92%</td>
</tr>
</tbody>
</table>

- The most important role of Chinese words/phrases is for describing symptom related information
  - Example: 今早開始腹痛; 剛吃藥後始雙眼腫 現呼吸不適 心悸
- Chinese Punctuation
- Name Entity
  - Example: Diarrhea SINCE THIS MORNING. Group poisoning. Having dinner at 喜滿客 restaurant.
Chinese CC Preprocessing: System Design

Stage 0.1: Separate Chinese and English Expressions

Stage 0.2: Chinese Phrase Segmentation

Stage 0.3: Chinese Phrase Translation

English Expressions

Chinese Expressions

Segmented Chinese Phrases

Translated Chinese Phrases

Chinese Medical Phrases

Common Chinese Phrases

Chinese to English Dictionary

Raw Chinese CCs

Mutual Info.

Chinese Chief Complaints
Chinese Phrases Segmentation

- Technology Used:
  - MI (Mutual Information)

- Test bed:
  - 1978 records from hospital A
  - 18% records have Chinese expression

- Results:
  - 726 phrases extracted
  - 370 (51%) are medical related

- Example:
  - Input: 與人打架, 用鍋鏟打到頭部,流鼻血
  - Output: 與-人-打架, 用-鍋鏟-打到頭-部, 流鼻血
Chinese Phrases Translation

- Recruited 3 physicians to help translating 370 extracted Chinese terms
- 280 (76%) terms have consistent translation

Example:

- Input:
  與-人-打架，用-鍋鏟-打到頭-部，流鼻血
- Intermediate output:
  N/A-N/A-fighting, N/A-N/A-head injury-N/A, epistaxis
- Final result:
  fighting, head injury, epistaxis

<table>
<thead>
<tr>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>小便量少</td>
<td>oliguria</td>
</tr>
<tr>
<td>心臟痛</td>
<td>chest pain</td>
</tr>
<tr>
<td>叫喚不醒</td>
<td>coma</td>
</tr>
<tr>
<td>打倒頭頂</td>
<td>head injury</td>
</tr>
<tr>
<td>肝指數高</td>
<td>abnormal liver function</td>
</tr>
<tr>
<td>食道吞嚥痛</td>
<td>dysphagia</td>
</tr>
<tr>
<td>排尿困難</td>
<td>dysuria</td>
</tr>
</tbody>
</table>
Result: Self Validation

- Use the 280 translations against 1978 chief complaints from hospital A
  - 1610 (82%) records are in English
  - 368 (18%) records contain Chinese
    - 36% contains trivial info. (Eg. “r/o septic shock 外院轉入”)
    - 64% contains non-trivial info. (Eg. “poor intake and 味覺喪失”)
  - 67% has complete translation
  - 2% has partial translation
  - 20% does not have translation
Taiwan Surveillance Data Visualization

- 2.2M+ scrubbed chief complaints records
General Grouping
Group by Hospital
Group by Syndrome Classification
Incorporating Geographical Contacts into Social Network Analysis for Contact Tracing in Epidemiology: A Study of Taiwan SARS Data

Hsinchun Chen; Yida Chen; Cathy Larson; Chunju Tseng; The BioPortal Team, Artificial Intelligence Lab, University of Arizona; Chwan-Chuen King, Tsung-Shu Joseph Wu, National Taiwan University

Acknowledgements: NSF ITR Program
Social Network Analysis in Epidemiology

- **Conceptualizing** a population as a set of individuals linked together to form a large social network provides a fruitful perspective for better **understanding** the spread of some infectious diseases. (Klovdahl, 1985)

- Social Network Analysis in epidemiology has two major activities:
  - **Network Construction**
    - Link the whole set of persons in a particular population with relationships or types of contacts
  - **Network Analysis**
    - Measure and make inferences about structural properties of the social networks through which infectious agent spread
# A Taxonomy of Network Construction

<table>
<thead>
<tr>
<th>Disease</th>
<th>Linking Relationship</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexually Transmitted Disease (STD)</td>
<td>Sexual Contact</td>
<td>AIDS (Klovdahl, 1985)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gonorrhea (Ghani et al., 1997)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Syphilis (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td>Sexually Transmitted Disease (STD)</td>
<td>Drug Use</td>
<td>AIDS (Klovdahl et al., 1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td>Sexually Transmitted Disease (STD)</td>
<td>Needle Sharing</td>
<td>AIDS (Klovdahl et al., 1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td>Tuberculosis (TB)</td>
<td>Social Contact</td>
<td>AIDS (Klovdahl et al., 1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td>Severe Acute Respiratory Syndrome (SARS)</td>
<td>Personal Contact</td>
<td>(Klovdahl et al., 2001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(McElroy et al, 2003)</td>
</tr>
<tr>
<td>Severe Acute Respiratory Syndrome (SARS)</td>
<td>Geographical Contact</td>
<td>(Klovdahl et al., 2001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(McElroy et al, 2003)</td>
</tr>
<tr>
<td>Severe Acute Respiratory Syndrome (SARS)</td>
<td>The Source of Infection</td>
<td>(CDC*, 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Shen et al., 2004)</td>
</tr>
<tr>
<td></td>
<td>Personal Contact</td>
<td>(Meyers et al., 2005)</td>
</tr>
</tbody>
</table>

*CDC: Centers for Disease Control and Prevention*
## A Taxonomy of Network Analysis

<table>
<thead>
<tr>
<th>Levels of Analysis</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network Visualization</strong></td>
<td>Show the spread of an infectious agent transmitted from one person to another</td>
<td>AIDS (Klovdahl, 1985)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Syphilis (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SARS (CDC*, 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SARS (Shen et al., 2004)</td>
</tr>
<tr>
<td><strong>Network Measurement</strong></td>
<td>Study the structure of a population through which an infectious agent is transmitted during close personal contact</td>
<td>Syphilis (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS (Klovdahl et al., 1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS (Rethenberg et al., 1998)</td>
</tr>
<tr>
<td></td>
<td>Develop disease containment strategies or programs</td>
<td></td>
</tr>
<tr>
<td><strong>Network Simulation</strong></td>
<td>Evaluate the spread of an infectious agent within a population with different network parameters</td>
<td>Gonorrhea (Ghani et al., 1997)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SARS (Meyers et al., 2005)</td>
</tr>
</tbody>
</table>

*CDC: Centers for Disease Control and Prevention*
Network Visualization

- Focus on the identification of
  - Subgroups within the population
  - Characteristics of each subgroup
  - Bridges between subgroups which transmit a disease from a subgroup to another

Clusters in Singapore Source (CDC, 2003)

Syphilis Transmission (Rothenberg et al., 1998)
Research Questions

- What are the differences in connectivity between personal and geographical contacts in the construction of contact networks?

- What are the differences in network topology between one-mode networks with only patients and multi-mode networks with patients and geographical locations?

- Whether SNA with geographical nodes can be used to identify epidemic phases of infectious diseases with multiple transmission modes?
SARS in Taiwan

- The first SARS case in Taiwan was a Taiwanese businessman who traveled to Guangdong Province via Hong Kong in the early February 2003.
  - Had onset of symptoms on February 26, 2003
  - Infected two family members and one healthcare worker

- Eighty percent of probable SARS cases were infected in hospital setting.
  - The first outbreak began at a municipal hospital in April 23, 2003.
  - Total seven hospital outbreaks were reported.
  - Hospital shopping and transfer were suspected to trigger such sequential hospital outbreaks.
Taiwan SARS Data

- Taiwan SARS data was collected by the Graduate Institute of Epidemiology at National Taiwan University during the SARS period.
- In this dataset, there are 961 patients, including 638 suspected SARS patients and 323 confirmed SARS patients.
- The contact-tracing data of patients in this dataset has two main categories, personal and geographical contacts, and nine types of contacts.
  - Personal contacts: family member, roommate, colleague/classmate, and close contact
  - Geographical contacts: foreign-country travel, hospital visit, high risk area visit, hospital admission history, and workplace
Taiwan SARS Data (Cont.)

- Hospital admission history is the category with largest number of records (43%).
- Personal contacts are primarily comprised of family member records.

<table>
<thead>
<tr>
<th>Category</th>
<th>Type of Contacts</th>
<th>Records</th>
<th>Suspected Patients</th>
<th>Confirmed Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal</td>
<td>Family Member</td>
<td>177</td>
<td>48</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Roommate</td>
<td>18</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Colleague/Classmate</td>
<td>40</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Close Contact</td>
<td>11</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Geographical</td>
<td>Foreign-Country Travel</td>
<td>162</td>
<td>100</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td><strong>Hospital Visit</strong></td>
<td>215</td>
<td>110</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>High Risk Area Visit</td>
<td>38</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td><strong>Hospital Admission History</strong></td>
<td>622</td>
<td>401</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>Workplace</td>
<td>142</td>
<td>22</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>1425</td>
<td>638</td>
<td>323</td>
</tr>
</tbody>
</table>
Phase Analysis

- In the phase analysis, we want to examine whether epidemic phases of an infectious disease with multiple transmission modes, such as SARS, could be identified through SNA with geographical nodes.

- SARS transmission in Taiwan has two main phases:
  - **Importation (February to the middle of April 2003)**
    - Small clusters of local transmission were initiated by the imported cases of SARS.
    - Patients were primarily infected through:
      - Travels in the mainland China and Hong Kong (Geographical contacts)
      - Family Transmission
  - **Hospital Outbreaks (The middle of April to July 2003)**
    - Patients were primarily infected through:
      - Hospital related contacts (Geographical contacts)
      - Close personal contacts
Phase Analysis (Cont.)

Network Partition
- We partition each contact network on a weekly basis with linkage accumulation.
- From 2/24 to 5/4, there are 10 weeks in total.

Personal Contact Network

- Week 1
- Week 2
- Week 3
- …………
- Week 10

2/24 3/3 3/10 3/17 5/4
Phase Analysis (Cont.)

- Network Measurement
  - We investigate two factors that contribute to the transmission of disease in macro-structure:
    - Density: the degree of intensity to which people are linked together
      - Density
      - Average degree of nodes
    - Transferability: the degree to which people can infect others
      - Betweenness
      - Number of components

Lower density

Higher density

Lower Transferability

Higher Transferability
Phase Analysis (Cont.)

- Measuring weekly changes:

\[
\Delta A_i = \frac{|A_i - A_{i-1}|}{A_n}
\]

for \(i = 2\) to \(n\)

where

- \(A_i\): a network measure of Week \(i\) partition
- \(A_n\): a network measure of the last week partition
Connectivity Analysis

- Geographical contacts provide much higher connectivity than personal contacts in the network construction.
  - Decrease the number of components from 961 to 82
  - Increase the average degree from 0.31 to 108.62

<table>
<thead>
<tr>
<th>Applied Contacts in the network construction</th>
<th>Average Degree (Patient Nodes)</th>
<th>Maximum Degree (Patient Nodes)</th>
<th>Number of Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Contacts</td>
<td>0.31</td>
<td>4</td>
<td>847</td>
</tr>
<tr>
<td><strong>Geographical Contacts</strong></td>
<td><strong>108.62</strong></td>
<td><strong>474</strong></td>
<td><strong>82</strong></td>
</tr>
<tr>
<td>Personal + Geographical Contacts</td>
<td>108.85</td>
<td>474</td>
<td>10</td>
</tr>
</tbody>
</table>
Connectivity Analysis (Cont.)

- The hospital admission history provides the highest connectivity of nodes in the network construction.
- The hospital visit provides the second highest connectivity.
- This result is consistent with the fact that most of patients got infected in the hospital outbreaks during the SARS period.

<table>
<thead>
<tr>
<th>Applied Contacts in the network construction</th>
<th>Average Degree</th>
<th>Maximum Degree</th>
<th>Number of Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Personal Contacts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family Member</td>
<td>0.204</td>
<td>4</td>
<td>893</td>
</tr>
<tr>
<td>Roommate</td>
<td>0.031</td>
<td>2</td>
<td>946</td>
</tr>
<tr>
<td>Colleague/Classmate</td>
<td>0.06</td>
<td>3</td>
<td>934</td>
</tr>
<tr>
<td>Close Contact</td>
<td>0.023</td>
<td>1</td>
<td>949</td>
</tr>
<tr>
<td><strong>Geographical Contacts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign-Country Travel</td>
<td>2.727</td>
<td>41</td>
<td>848</td>
</tr>
<tr>
<td><strong>Hospital Visits</strong></td>
<td>10.077</td>
<td>105</td>
<td>753</td>
</tr>
<tr>
<td>High Risk Area Visit</td>
<td>1.388</td>
<td>36</td>
<td>924</td>
</tr>
<tr>
<td><strong>Hospital Admission History</strong></td>
<td>50.479</td>
<td>289</td>
<td>409</td>
</tr>
<tr>
<td>Workplace</td>
<td>4.694</td>
<td>61</td>
<td>823</td>
</tr>
</tbody>
</table>
One-Mode Network with Only Patient Nodes

○ : Suspected
● : Confirmed
Contact Network with **Geographical Nodes**

[Diagram showing a network of hospitals and geographical areas with nodes and connections.]

- :Area
- :Hospital
- :Suspected
- :Confirmed
**Potential Bridges Among Geographical Nodes**

- Including geographical nodes helps to reveal some potential people who play the role as a bridge to transfer disease from one subgroup to another.
Network Visualization (Cont.)

- For a hospital outbreak, including geographical nodes and contacts in the network is also useful to see the possible disease transmission scenario within the hospital.

Background of the Example

- Mr. L, a laundry worker in Heping Hospital, had a fever on 2003/4/16 and was reported as a suspected SARS patient.
- Nurse C took care of Mr. Liu on 4/16 and 4/17.
- Nurse C and Ms. N, another laundry worker in Heping Hospital, began to have symptoms on 4/21.
- Heping Hospital was reported to have an SARS outbreak on 4/24.
- Nurse C’s daughter had a fever on 5/1.
Phase Analysis – Density

- Normalized density and average degree show similar patterns:
  - In the importation phase, foreign-country contact network increases dramatically in Week 4 (3/17-3/23), followed by personal contact network.
  - In the hospital outbreak phase, both personal and hospital networks increase dramatically. But in Week 10, personal network still increases while hospital network decreases.
Phase Analysis – Transferability

☐ From betweenness, we can see that personal network doesn’t have enough transferability until Week 9.

☐ Personal network just forms several small fragments without big groups in the importation phase.

☐ From the number of components, hospital network is the only one which can consistently link patients together.
Phase Analysis – Hospital Outbreak

- We further partition hospital network by patients and healthcare workers (HCW).
- From density and betweenness, we can see that before Week 9 hospital network is mainly affected by patients' hospital contacts. However, after Week 9, healthcare worker contacts lead the trend.
Data Selection – Select a Dataset

Select TAIWAN_SARS dataset for network visualization

<table>
<thead>
<tr>
<th>Selected</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TAIWAN_SARS</td>
<td>Taiwan SARS patient records. Data (997 records) was received on 09/10/2006, covering February to July 2003.</td>
</tr>
</tbody>
</table>

Advanced...  Next-->
Data Selection – Specify a Period of Time

Specify a period of time for data selection
Data Selection – Select Actor Types

Select the types of actors in network

<table>
<thead>
<tr>
<th>Default Actor Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
</tr>
<tr>
<td>Geographical Locations</td>
</tr>
</tbody>
</table>

You must have Java WebStart installed to visualize the results, go [here](#) and click "Get It Now" to install necessary software. This link will open in a new window.
Network Visualization (Cont.)

Social network visualization with patients and geographical locations

Scroll bar on time dimension to see the evolution of a network
The index patient of Heping Hospital began to have symptoms.
The SARS infection within the hospital started on 4/16.
The hospital outbreak started on 4/20.
The hospital outbreak was reported by the press on 4/24.
The outbreak spread to other hospitals.
The outbreak spread to other hospitals.
Conclusions

- Geographical contacts provide much higher connectivity in network construction than personal contacts.

- Introducing geographical locations in SNA provides a good way not only to see the role that those locations play in the disease transmission but also to identify potential bridges between those locations.

- SNA with geographical nodes can demonstrate the underlying context of transmission for the infectious diseases with multiple modes.
BioPortal Information

- Hsinchun Chen,
  hchen@eller.arizona.edu

- AI Lab, http://ai.arizona.edu

- BioPortal Demo and Information:
  http://bioportal.org