Grid Computing, Social Media, and Geospatial Cyberinfrastructure

Ming-Hsiang (Ming) Tsou mtsou@mail.sdsu.edu, Professor (Geography), San Diego State University,

PI of NSF-CDI “Mapping Ideas” project. Funded by NSF Cyber-Enabled Discovery and Innovation (CDI) program. Award # 1028177 (2010-2014)

Co-PIs:
Dr. Dipak K Gupta (Political Science), Dr. Jean Marc Gawron (Linguistic), Dr. Brian Spitzberg (Communication), Dr. Li An (Geography)

San Diego State University, USA
Three Links for Geospatial Cyberinfrastructure

Enabling Advanced/Complex GIS simulation Models for Various Applications (including Public Health)


Cloud Computing

Grid Computing

Executes data processing modules and HPC simulation modules

Provides input data processing results

Transfer simulation results

Executes HPC simulation modules

Send simulation results

Provides Web mapping APIs and databases

Users

Geospatial Web Server
Enabling Complex GIS Model (ABM)

- **Case Study**: Schelling Residential Segregation Model built by Thomas Schelling in 1971 to understand dynamic residential segregation patterns.

- Apply the complex GIS model to the real-world data (the County of San Diego) with **2.8 million agents** (the total number of residents in the County of San Diego in 2000).

**Tolerance Rates**: 40%

- Small circles: dissatisfied agents
  - Different races > tolerance rates

- Large circles: satisfied agents
  - Different races < tolerance rates

**Box 1**: Red 50%, Blue 50%
**Box 2**: Red 62.5%, Blue 37.5%

Domain Experts (Model Designers)

HPC/GIS CyberGIS Programmers

Collaborate???
Web-based Interface for Model Designers

Geospatial Web Portal Prototype

User Scenario Setup (Tolerance Rates)
Select Computing Resources
Analysis and Visualization Module
KML
Analysis Results

Grid Computing (XSEDE)

Resident Agents
Housing Units
Scenario
Main Simulation Parallel Processing Module
Simulation Results

Cloud Computing (Amazon EC2)

Resident Agents
Housing Units
Scenario
Main Simulation Parallel Processing Module
Simulation Results

GSISSH
GSISCP
SSH
SCP
VIDEO DEMO (2 MINUTES)

Web-based GIS simulation model (parallel Schelling model)
How to Utilize CyberGIS (Geospatial Cyberinfrastructure – including Grid computing and Cloud computing resources) to advance the health sciences and NIH research?

- **Web-based Interfaces** for Simulation Model builders (Domain Experts just focus on the models, not on the programming tasks).

- **Dynamic Integration** of grid computing and cloud computing resources for public health databases/analysis. Different tasks required different arrangements (ArcGIS online – Cloud, GIS processing portals – Grid, Spatial Analysis + Visualization -- Both).

- **CON:** Data Security/Privacy, Job Queue (long waiting time)

- Need more collaborations between the two research communities (GIScience and Public Health)!!!
Social Media (Tweets) – Semi-Private Communication:
Use Twitter APIs to retrieve tweets based on Keywords or #Hashtag and geolocations (self-defined home-towns or GPS locations).

Twitter
- An online social networking service
- A micro-blogging service
- Created in March, 2006
- Reached 300 million users as of 2011
- Over 300 million tweets per day
- Over 1.6 billion search queries per day

Twitter Application Programming Interface (API)
- For software to communicate with other programs
- Search API: query for Twitter content + spatial query function.
- Streaming API: real-time data trend mining + location info.
Twitter - Spatial Search API

Please enter the search Keyword: Mitt Romney

Please enter the type of this search
General (G) // Location-specific (L): L

Please input search center-location
(example format: 32.777393,-117.073727): 41.961295,-93.281859

Please input search radius (miles)(example: 10): 180

Enter current time in HHMM format (example: 0445 or 1945): 1806

Searching tweets from Twitter now with the keyword: “Mitt Romney”

Excel file successfully generated: Mitt Romney(41.961295,-93.281859,L80m1)-2012-01-05-1806.xls

Number of tweets with GPS-geocoded: 32 (in 1346 results)
Percentage of GPS-geocoded tweets in this search: 2.37741456166 %

Center: 41.961295, -93.281859
Radius: 45 miles

Limitations:
1. Spatial Search can only trace back up to seven days. (Regular search can trace back to 14 days.)
2. Each search results can not exceed 1500 tweets.
Whooping Cough Epidemic (WA State)
How does the frequency of tweets (search by keywords “whooping cough”) compare to Pertussis cases? in different cities?

- We started collecting tweets at **week 20 (May 14, 2012)**
- (30 major U.S. cities – using **17 miles** radius search).
Tweets Search Results
(Python programming with Twitter API and save results to SQL databases)
Time, keywords, text-content, user-name, ID, message to user, location (self-defined city or region, or GPS coordinates – only 1% with GPS).
TWEETS in 30 Cities (search by “whooping cough”) (from 5-14-2012 to 7-7-2012 - two months, 2170 TWEETS)

Seattle, WA had the greatest number of tweets in absolute terms and relative to population size.

The raw counts of tweets

The adjusted number (divided by the population of cities)
Daily Frequency of Tweets (containing “whooping cough” keyword) - combined all 30 major U.S. Cities. (Temporal Change)

RED Line: The actual Pertussis Cases in Washington State. The weekly pertussis cases were plotted on top of the daily tweets. In general is seem that as the number of cases is decreasing so is the number of tweets.

If we look solely at the tweets produced in Seattle, WA we still see a cyclical trend.
May 16: lots of RT:
@gatesfoundation
@HarvardHSPH

http://t.co/tsEBP7My

(http://www.impatientoptimists.org/Posts/2012/05/Whooping-Cough-Epidemic-in-Washington-State)

Same situation on June 11
(An article from NPR was re-tweeted. It was created in the early morning of June 11th)
1745 different users who tweeted on whooping cough. One group stands out. AANaturalHealth which is the user name for Naturally Healthy
Top 3 tweeters on Whooping Cough

- **AANaturalHealth** (User Name)
  - Naturally Healthy
  - Worldwide
  - 12,521 tweets and **10,218 followers (since 2006)**
  - Focus on natural medicine with no mention of vaccines in tweets or website

- **Medical_Discuss** (User Name)
  - Medical Discussions
  - Worldwide
  - 214,544 tweets and 1,890 followers
  - Positive view on vaccines
  - Website: [http://medical-discussions.com](http://medical-discussions.com) discussion page with various health related topics

- **anushka_sen** (User Name)
  - Anushka Sen
  - New York
  - 205,231 tweets and 2,362 followers
  - Website: [http://www.fametimes.com](http://www.fametimes.com) (page suspended)
Pertussis (the causative agent of whooping cough) is the term used by most health care professionals. The term whooping cough is usually used by the lay person. Are the trends for these keywords different?

Total numbers: **639** tweets for “Pertussis” vs, **2170** tweets for “Whooping cough”

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cough</td>
<td>6762</td>
<td>cough</td>
<td>466</td>
</tr>
<tr>
<td>whooping</td>
<td>6758</td>
<td>whooping</td>
<td>465</td>
</tr>
<tr>
<td>worst</td>
<td>3201</td>
<td>outbreak</td>
<td>401</td>
</tr>
<tr>
<td>epidemic</td>
<td>3058</td>
<td>worst</td>
<td>313</td>
</tr>
<tr>
<td>years</td>
<td>2037</td>
<td>years</td>
<td>237</td>
</tr>
<tr>
<td>antivaccination</td>
<td>1254</td>
<td>cdc</td>
<td>230</td>
</tr>
<tr>
<td>antivaccine</td>
<td>1089</td>
<td>vaccine</td>
<td>147</td>
</tr>
<tr>
<td>specterm</td>
<td>1051</td>
<td>epidemic</td>
<td>146</td>
</tr>
<tr>
<td>propaganda</td>
<td>1048</td>
<td>abc</td>
<td>137</td>
</tr>
<tr>
<td>mims</td>
<td>1046</td>
<td>cases</td>
<td>135</td>
</tr>
<tr>
<td>//t.co/0hsfoq64</td>
<td>1019</td>
<td>baby</td>
<td>124</td>
</tr>
<tr>
<td>cases</td>
<td>996</td>
<td>lizszabo</td>
<td>119</td>
</tr>
<tr>
<td>jamesgunn</td>
<td>931</td>
<td>kids</td>
<td>112</td>
</tr>
<tr>
<td>jenny</td>
<td>871</td>
<td>largest</td>
<td>107</td>
</tr>
<tr>
<td>mccarthy</td>
<td>867</td>
<td>washington</td>
<td>93</td>
</tr>
<tr>
<td>dead</td>
<td>848</td>
<td>yrs</td>
<td>84</td>
</tr>
<tr>
<td>babies</td>
<td>813</td>
<td>current</td>
<td>81</td>
</tr>
<tr>
<td>yrs</td>
<td>807</td>
<td>deaths</td>
<td>78</td>
</tr>
<tr>
<td>achieving</td>
<td>730</td>
<td>dead</td>
<td>77</td>
</tr>
</tbody>
</table>


\textbf{"Flu vaccine" Tweets Sentiment Analysis by City}

Manually conducted by Anna Nagel (our graduate student)

The raw counts of tweets

The adjusted number (divided by the population of cities)
1) Social media might provide an effective way to monitor the public opinions to outbreaks, epidemics, and public health issues – in near “real time”. (Social Radar).

2) Spatial Variation/Analysis is the key for understanding the contexts of public health issues.

3) Many “distortions” and “noises” in collected data. How to clean up these messages?

4) Comparing different types of social networks and social media (such as Twitters, Facebook, personal blogs, etc.).

5) Privacy Concerns, Ethical Implication.

Image source: http://www.censusscope.org/us/chart_age.html
Thank You  Q & A

Ming-Hsiang Tsou  mtsou@mail.sdsu.edu

http://mappingideas.sdsu.edu/

Social Media Research is funded by NSF Cyber-Enabled Discovery and Innovation (CDI) program. Award # 1028177. (four years 2010-2014)
Back-up Slides
Tweets Geolocation Analysis (based on the 2012 GOP primary keywords with GPS enabled tweets (1177 tweets) – Within the same State in the user-defined profile 60% (or higher). Outside the state: 20%, N/A 19%.

<table>
<thead>
<tr>
<th>State</th>
<th>Total Geo-Tweets</th>
<th>Consistent</th>
<th>Consistent (%)</th>
<th>Non-Consistent</th>
<th>Non-Consistent (%)</th>
<th>N/A</th>
<th>N/A (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>116</td>
<td>36</td>
<td>31.03</td>
<td>64</td>
<td>55.17</td>
<td>16</td>
<td>13.79</td>
</tr>
<tr>
<td>NH</td>
<td>21</td>
<td>2</td>
<td>9.52</td>
<td>18</td>
<td>85.71</td>
<td>1</td>
<td>4.76</td>
</tr>
<tr>
<td>SC</td>
<td>21</td>
<td>13</td>
<td>61.90</td>
<td>8</td>
<td>38.10</td>
<td>2</td>
<td>9.52</td>
</tr>
<tr>
<td>FL</td>
<td>143</td>
<td>69</td>
<td>48.25</td>
<td>50</td>
<td>34.97</td>
<td>24</td>
<td>16.78</td>
</tr>
<tr>
<td>NV</td>
<td>42</td>
<td>19</td>
<td>45.24</td>
<td>18</td>
<td>42.86</td>
<td>5</td>
<td>11.90</td>
</tr>
<tr>
<td>CO</td>
<td>22</td>
<td>18</td>
<td>81.82</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>18.18</td>
</tr>
<tr>
<td>MN</td>
<td>48</td>
<td>38</td>
<td>79.17</td>
<td>3</td>
<td>6.25</td>
<td>7</td>
<td>14.58</td>
</tr>
<tr>
<td>ME</td>
<td>4</td>
<td>2</td>
<td>50.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AZ</td>
<td>30</td>
<td>21</td>
<td>70.00</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>30.00</td>
</tr>
<tr>
<td>MI</td>
<td>113</td>
<td>78</td>
<td>68.35</td>
<td>12</td>
<td>16.62</td>
<td>23</td>
<td>19.05</td>
</tr>
<tr>
<td>WA</td>
<td>45</td>
<td>43</td>
<td>95.56</td>
<td>0</td>
<td>4.44</td>
<td>2</td>
<td>4.44</td>
</tr>
<tr>
<td>AK</td>
<td>2</td>
<td>2</td>
<td>100.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GA</td>
<td>71</td>
<td>40</td>
<td>56.34</td>
<td>7</td>
<td>4.44</td>
<td>21</td>
<td>33.80</td>
</tr>
<tr>
<td>ID</td>
<td>7</td>
<td>5</td>
<td>71.43</td>
<td>1</td>
<td>14.29</td>
<td>1</td>
<td>14.29</td>
</tr>
<tr>
<td>MA</td>
<td>67</td>
<td>45</td>
<td>67.16</td>
<td>8</td>
<td>11.94</td>
<td>14</td>
<td>20.90</td>
</tr>
<tr>
<td>ND</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100.00</td>
</tr>
<tr>
<td>OH</td>
<td>91</td>
<td>50</td>
<td>54.95</td>
<td>19</td>
<td>16.99</td>
<td>31</td>
<td>34.07</td>
</tr>
<tr>
<td>OK</td>
<td>20</td>
<td>13</td>
<td>75.00</td>
<td>1</td>
<td>5.00</td>
<td>4</td>
<td>20.00</td>
</tr>
<tr>
<td>TN</td>
<td>80</td>
<td>60</td>
<td>75.00</td>
<td>7</td>
<td>8.75</td>
<td>13</td>
<td>16.25</td>
</tr>
<tr>
<td>VT</td>
<td>3</td>
<td>0</td>
<td>100.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VA</td>
<td>42</td>
<td>28</td>
<td>66.67</td>
<td>3</td>
<td>7.14</td>
<td>11</td>
<td>26.19</td>
</tr>
<tr>
<td>KS</td>
<td>20</td>
<td>17</td>
<td>85.00</td>
<td>0</td>
<td>0.00</td>
<td>3</td>
<td>15.00</td>
</tr>
<tr>
<td>AL</td>
<td>33</td>
<td>16</td>
<td>48.48</td>
<td>5</td>
<td>15.15</td>
<td>12</td>
<td>36.36</td>
</tr>
<tr>
<td>MS</td>
<td>13</td>
<td>9</td>
<td>69.23</td>
<td>1</td>
<td>7.69</td>
<td>3</td>
<td>23.08</td>
</tr>
<tr>
<td>IL</td>
<td>46</td>
<td>36</td>
<td>78.26</td>
<td>5</td>
<td>10.87</td>
<td>5</td>
<td>10.87</td>
</tr>
<tr>
<td>LA</td>
<td>5</td>
<td>5</td>
<td>100.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DC</td>
<td>32</td>
<td>18</td>
<td>56.25</td>
<td>8</td>
<td>25.00</td>
<td>6</td>
<td>18.75</td>
</tr>
<tr>
<td>MD</td>
<td>14</td>
<td>6</td>
<td>42.86</td>
<td>1</td>
<td>7.14</td>
<td>7</td>
<td>50.00</td>
</tr>
<tr>
<td>WI</td>
<td>25</td>
<td>18</td>
<td>72.00</td>
<td>5</td>
<td>20.00</td>
<td>2</td>
<td>8.00</td>
</tr>
</tbody>
</table>

Location Consistency on Primary Tweets (Geo-tagged Location vs. Self-report Location)

- Consistent: 60%
- Non-Consistent: 20%
- N/A: 20%

Locations are aggregated by
All cyberspace maps are “distorted from the real world due to the “projection methods” -- communication methods or media characteristics.

- Explanation
- Space-Time Analysis
- Prediction?

Correct the distortion effects

Different Projection Methods (Distortion)
What kinds of “distortion”?