

Multimedia and Web Databases

CSE – 515

Fall 2014

ARIZONA STATE UNIVERSITY

Professor K. Selcuk Candan

Spatial-Temporal Epidemic Simulation Datasets

Project Phase-1

Group 6:

Prerna Satija

1206312714

Masters in Computer Science

ABSTRACT

Spatial – Temporal Epidemic Simulation files for 51 states of USA are given. An epidemic is a situation wherein the number of new cases of a particular disease surpass the expected rate in given duration of time. The data of the simulation sets is normalized to get all values in the range of 0 to 1 for easy visualization. This data is quantized into a number of Gaussian Bands which is the resolution input by the user. The quantized data is then moved in windows each of length entered by the user. Each window is shifted h time units where h is input by the user. Based on the adjacency matrix of each state, average and difference files were created. Any file number input by the user is then viewed in the form of a heatmap and the states with lowest and highest strength (2-norm) are highlighted on the map along with their 1-hop neighbors in the epidemic word file, average file or difference file.

Keywords- STEM, epidemic simulation data sets, data normalization, time series, heatmap visualization, Gaussian distribution

INTRODUCTION

Terminologies:

- Epidemic:

An epidemic is a scenario when new cases of an infectious disease increase at a rate higher than usual. There could be various reasons for an epidemic: change of ecology, an unusual genetic condition, or the spread of a new parasite/virus.

- Spatial data:

Data that helps locate geographic location of features and boundaries on Earth and can be mapped.

- Temporal data:

Temporal data is represented by a timestamp. This data is different from non-temporal data in that it has a time period associated with all the records.

- Spatiotemporal Epidemiological Modeler:

STEM is an open source framework and a development tool designed to create spatial and temporal models of infectious disease. STEM allows to simulate an epidemic at state level, national level or county level.

Goal Description (Project Specification):

Sample files of epidemic simulations across 51 states of USA are given. This data needs to be normalized so that each value falls between 0 and 1. The normalized data should be quantized into r Gaussian bands where r is the resolution input by user. The length of each band will be computed using the given mean μ and standard deviation σ . The data should then be mapped to the center of the band depending on the range of each band computed. Window of window length, input by the user, needs to be moved on the time series of quantized data, with a shift of time units (input). The data represented in windows will be then written to a file. Based on the adjacency matrix of states, average and difference are computed using an input alpha value. Any simulation file input by user, will be viewed in the form of a heatmap and the highest and lowest strengths (norm-2) computed from any one of the epidemic word file, average file or difference file need to be highlighted on the heatmap.

Assumptions:

- The states in all simulation files are assumed to be in sorted order.
- The input file names and paths entered by the user are correct.
- Sample simulation files should be .csv files and locationMatrix should be .xls file.
- Each simulation file should have the same number of time stamps.

Description of proposed solution:

For an input directory path, each sample file in that directory is read and the data is normalized. Normalization is done for better visualization of data as the range of values after normalization will be in the range 0 to 1. The formula used for normalization is:

$$A_norm = (A - Amin)/(Amax - Amin)$$

The normalized data is quantized into r number of Gaussian bands, where r is the resolution input by user. The length of each band depends on mean and standard deviation. For this study the mean (μ) is zero and standard deviation (σ) is 0.25. The formulas used are:

$$length_i = \frac{\int_{(i-1)/r}^{i/r} \text{Gaussian}_{(\mu=0.0, \sigma=0.25)}(x) \delta x}{\int_0^1 \text{Gaussian}_{(\mu=0.0, \sigma=0.25)}(x) \delta x}$$

$$\text{Gaussian}(x) = f(x, \mu, \sigma) = (1/\sigma\sqrt{2\pi}) \exp(-(x-\mu)^2/2\sigma^2)$$

The normalized data is mapped to the center of the bands to get the quantized data. A window of length w , given by the user, is used to compute the window vector win_i for each state s . win_i is a vector all the values from the quantized matrix from a time stamp (index) t to $t + w$. The window shifts by user input shift length h . Each such entry consisting of file number (f), state (s), time stamp (t) and window vector (win_i) is stored in a file called, *epidemic_word_file*. Each entry in this file will be:

$$\begin{aligned} &\langle idx_i, win_i \rangle \\ &\text{where, } idx_i = \langle f, s, t \rangle \\ &\text{and, } win_i = \langle \text{window vector} \rangle. \end{aligned}$$

For the second task, using the connectivity graph G , found out the neighbors of all states. For each entry $\langle idx_i \rangle$ in *epidemic_word_file* average window vector ($win_{avg,i}$) is computed. This data matrix ($\langle idx_i, win_{avg,i} \rangle$) is then stored in file *epidemic_word_file_avg*. For computation, let win_i be the window vector corresponding to $\langle idx_i \rangle = \langle f_i, s_i, t_i \rangle$ and s_j be the array of all the 1-hop neighbors states of s_i obtained from connectivity graph G , then:

$$\begin{aligned} win_{avg,i} &= (\alpha * win_i) + (1 - \alpha) * AVG(\sum win_j \mathcal{E} \langle f_i, s_j, t_i \rangle) \\ &\text{where, } win_j \text{ is the window vectors for states } s_j \text{ with file number } f_i \text{ and time stamp } t_i \\ &\text{and, } \alpha \text{ is the weightage given by the user.} \\ &\langle idx_i, win_{avg,i} \rangle \text{ is stored in the file.} \end{aligned}$$

Similarly, $win_{diff,i}$ is computed and stored in file *epidemic_word_file_diff*. The method to compute the average window vector for each $\langle idx_i \rangle$ is the same as used to compute $win_{avg,i}$.

$$\begin{aligned} win_{diff,i} &= (win_i - AVG(\sum win_j \mathcal{E} \langle f_i, s_j, t_i \rangle)) / win_i \\ &\langle idx_i, win_{diff,i} \rangle \text{ is stored in the file.} \end{aligned}$$

For the third task, a user selected sample simulation file is viewed in the form of a heatmap, and the highest and the lowest strength states with their 1-hop neighbors are highlighted in the heatmap. The highest and lowest strength are computed according to the epidemic file chosen by the user. The strength is calculated as norm-2 length of the window vector (win_i) corresponding to a given word $\langle idx_i, win_i \rangle$. The norm-2 length is computed as:

$$\begin{aligned} ||x_p|| &= (\sum_{i=1}^n |x_i|^p)^{1/p} \\ &\text{where, } p = 2 \text{ for norm-2} \end{aligned}$$

The states with maximum and minimum 2-norm length are the highest and lowest strength states respectively. These states along with their 1-hop neighbors as computed from the connectivity graph G are highlighted on the

heatmap. The highest and lowest states are marked in red for identification.

Interface specifications

Implemented the proposed solution for Sample1.csv and Sample2.csv.

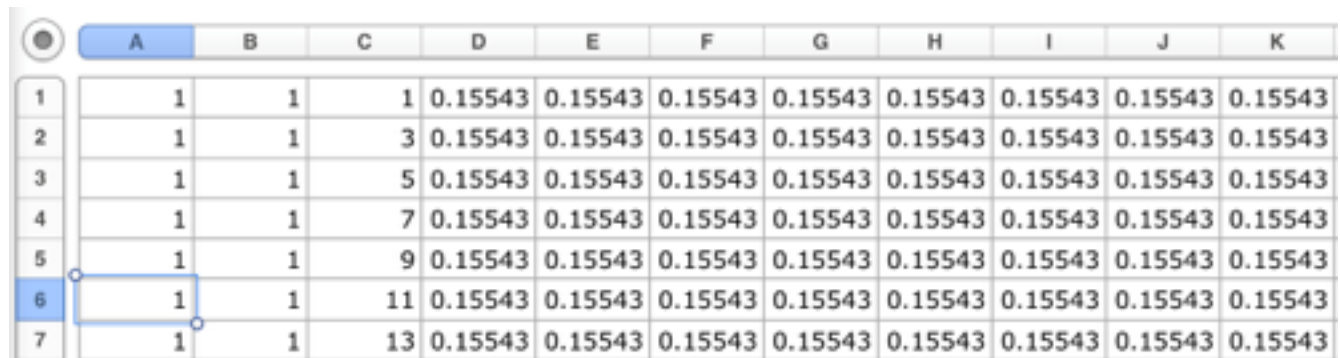
Expected Inputs:

- directoryPath: It is the path of the directory which has the sample simulation files.
- windowLength (w) : Size of window on quantized data.
- shiftLength (h): unit time the window should shift.
- resolution (r): number of bands.
- alpha (α): weightage (value between 0 and 1).
- locationFile: file path for LocationMatrix (connectivity graph).
- fileSelected: selected sample simulation file.
- filePath: choice between epidemic_word_file or epidemic_word_file_avg or epidemic_word_file_diff. This choice maps to the location of the respective file.

Outputs:

Task 1

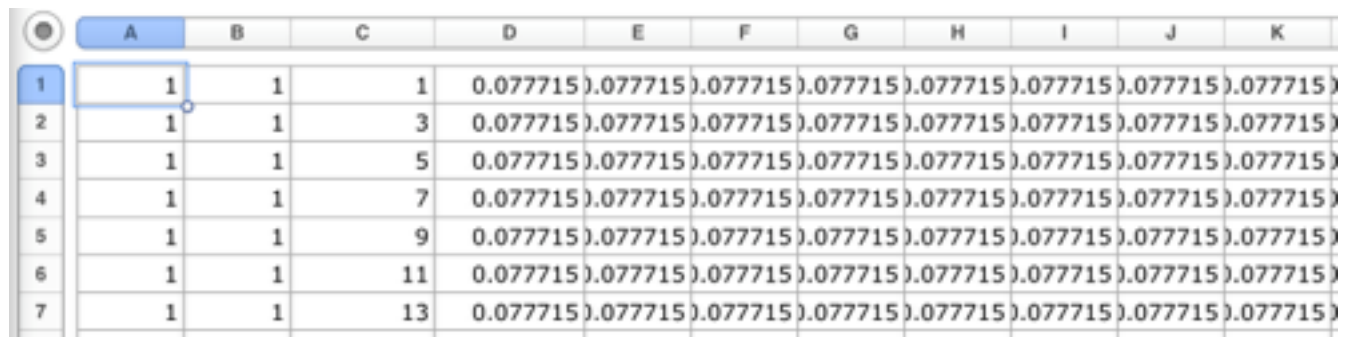
Screen shot of the epidemic_word_file. As can be observed from the screen shot also most values lie in the first band.



	A	B	C	D	E	F	G	H	I	J	K
1	1	1	1	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543
2	1	1	3	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543
3	1	1	5	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543
4	1	1	7	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543
5	1	1	9	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543
6	1	1	11	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543
7	1	1	13	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543	0.15543

Task 2

Screen shots of the epidemic_word_file_avg and epidemic_word_file_diff.



	A	B	C	D	E	F	G	H	I	J	K
1	1	1	1	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715
2	1	1	3	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715
3	1	1	5	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715
4	1	1	7	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715
5	1	1	9	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715
6	1	1	11	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715
7	1	1	13	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715	0.077715

epidemic_word_file_avg

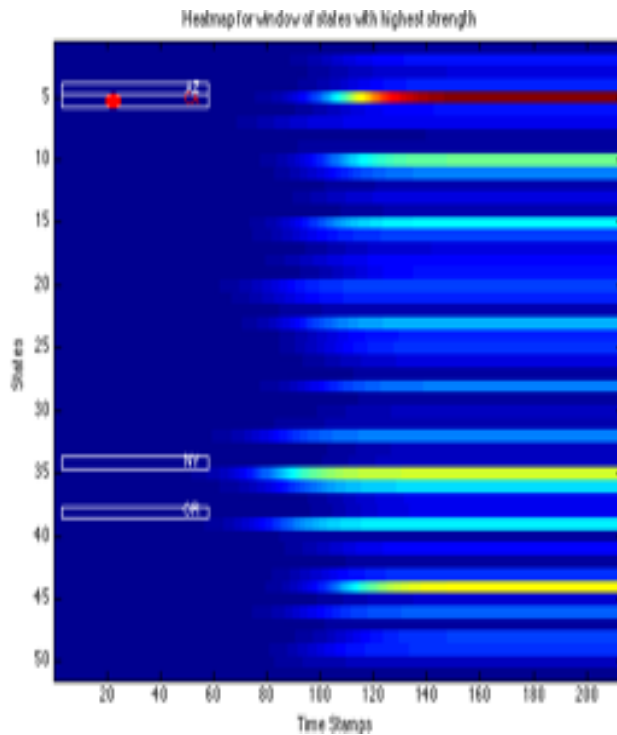
	A	B	C	D	E	F	G	H	I	J	K
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	3	1	1	1	1	1	1	1	1
3	1	1	5	1	1	1	1	1	1	1	1
4	1	1	7	1	1	1	1	1	1	1	1
5	1	1	9	1	1	1	1	1	1	1	1
6	1	1	11	1	1	1	1	1	1	1	1
7	1	1	13	1	1	1	1	1	1	1	1

epidemic_word_file_diff

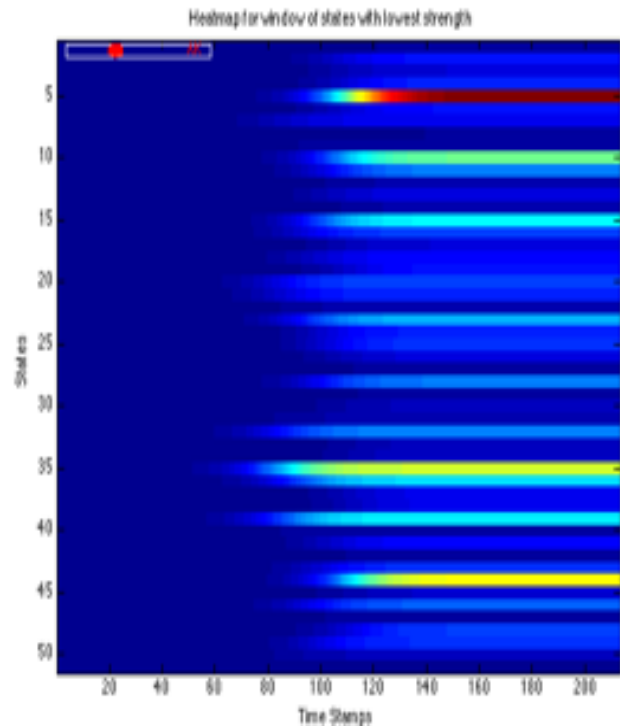
• Task 3

1. Heatmaps for Sample1.csv

Input: w=10, r=10, h=2, $\alpha=0.5$, fileSelected=1, filePath='MWDB_Phase1/Output/epidemic_word_file.csv'

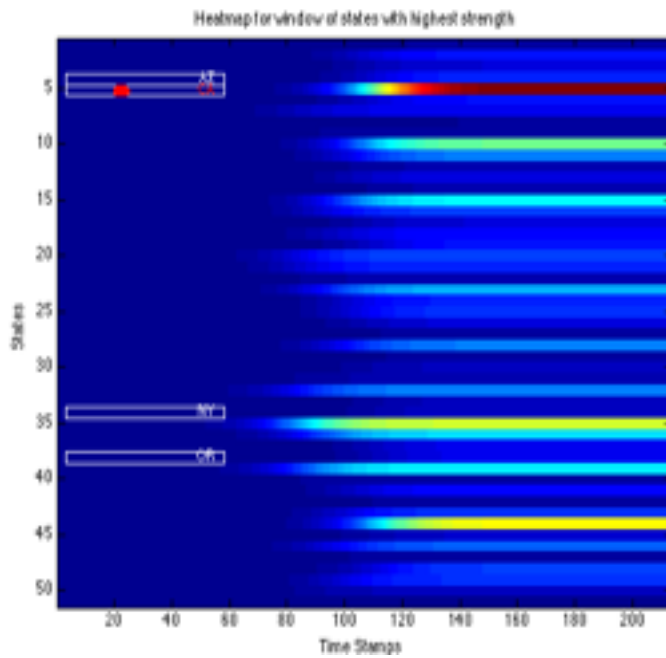


State with highest Strength is CA
1 Hop Neighbors are AZ, NY, and OR

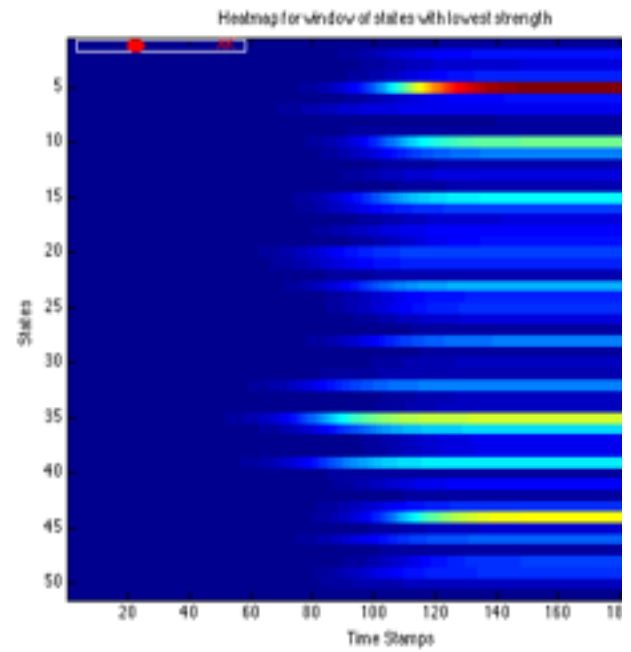


State with Lowest strength is AK
AK has no 1 Hop Neighbors

Input: w=10, r=10, h=2, $\alpha=0.5$, fileSelected=1, filePath='MWDB_Phase1/Output/epidemic_word_file_avg.csv'

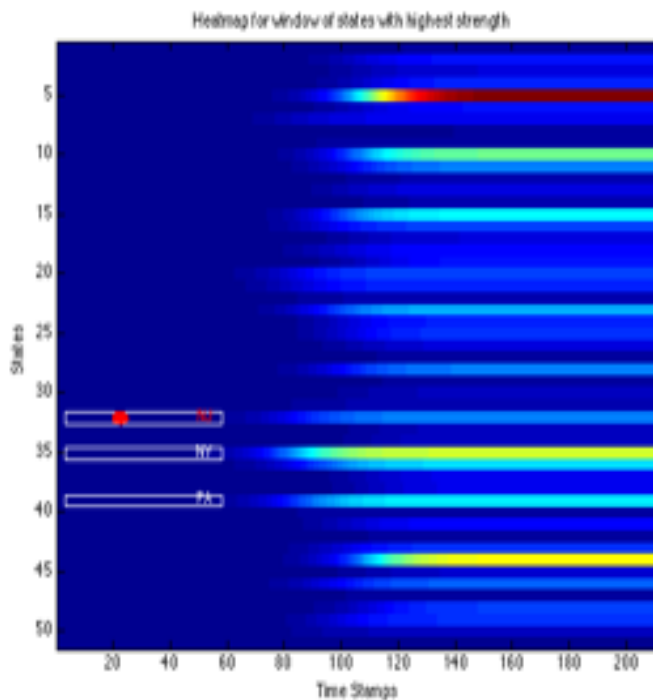


State with highest Strength is CA
1 Hop Neighbors are AZ, NY, and OR

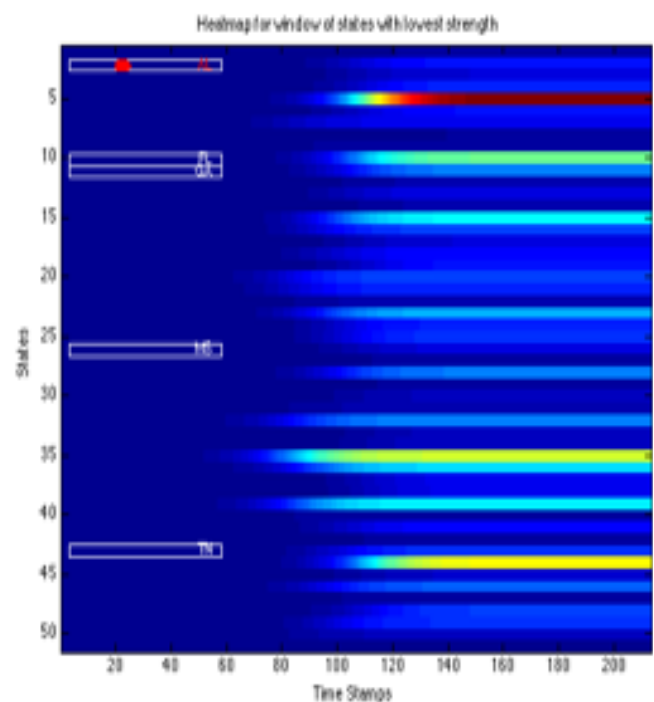


State with Lowest strength is AK
AK has no 1 Hop Neighbors

Input: $w=10$, $r=10$, $h=2$, $\alpha=0.5$, fileSelected=1, filePath = 'MWDB_Phase1/Output/epidemic_word_file_diff.csv'



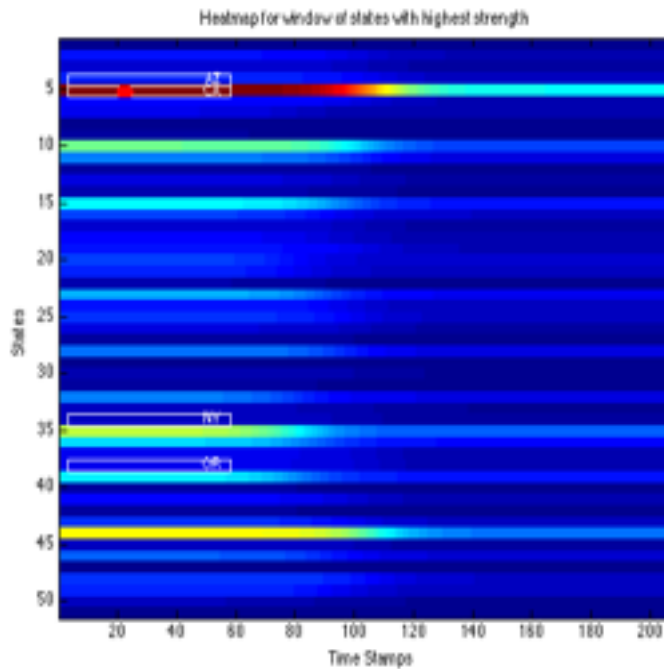
State with highest Strength is NJ
1 Hop Neighbors are NY and PA



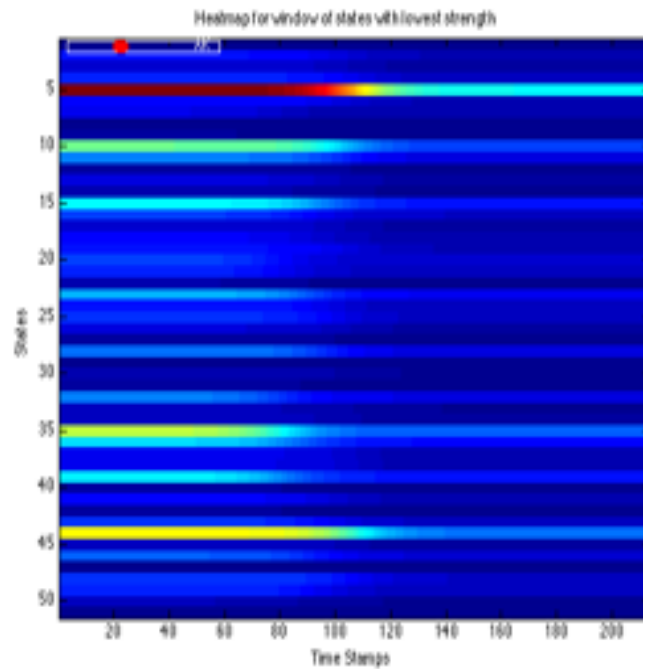
State with lowest Strength is AL
1 Hop Neighbors are FL, GA, MS, TN

2. Heatmaps for Sample2.csv

Input: $w=10$, $r=4$, $h=2$, $\alpha=0.5$, fileSelected =2, filePath = 'MWDB_Phase1/Output/epidemic_word_file.csv'

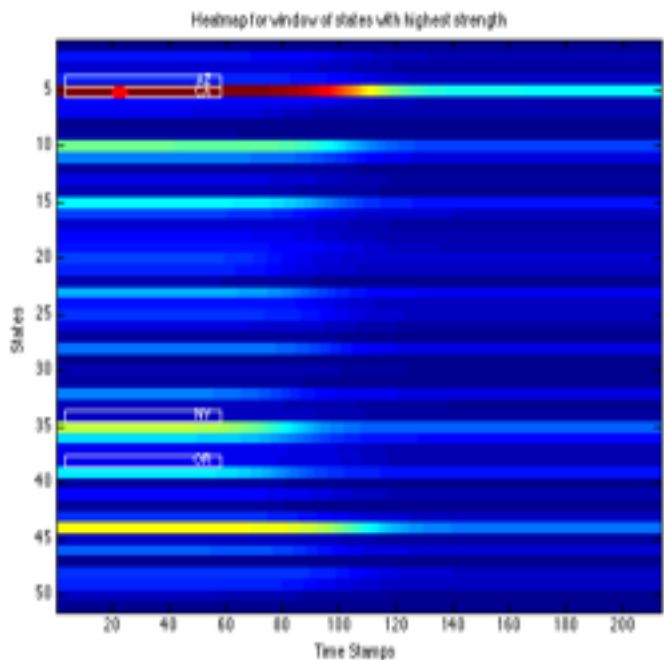


State with highest Strength is CA
1 Hop Neighbors are AZ, NY and OR

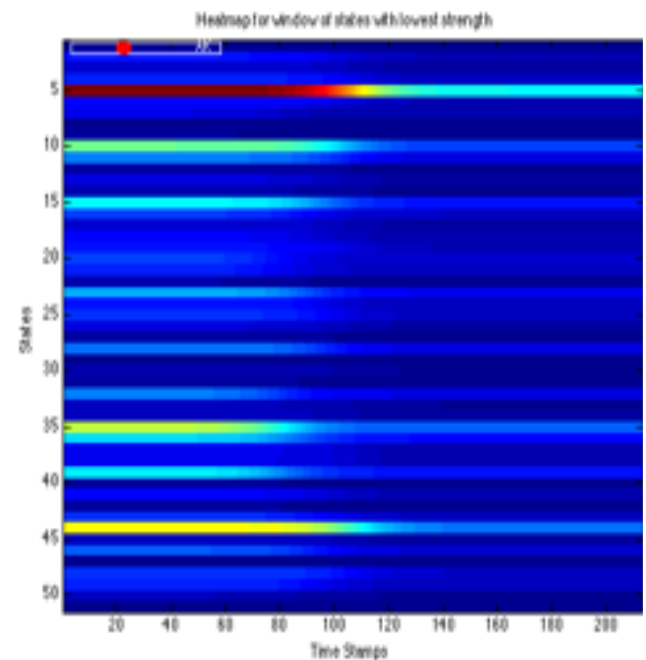


State with lowest Strength is AK
No 1 Hop Neighbors

Input: $w=10$, $r=4$, $h=2$, $\alpha=0.5$, fileSelected =2, filePath = 'MWDB_Phase1/Output/epidemic_word_file_avg.csv'

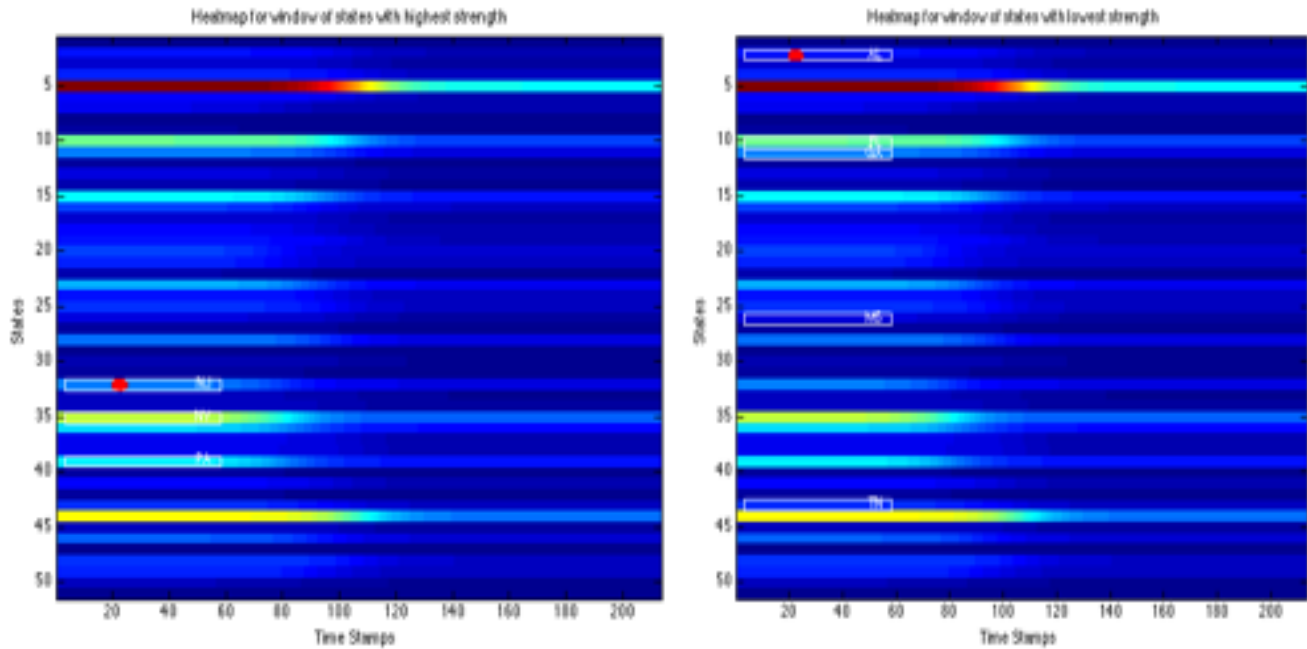


State with highest Strength is CA
1 Hop Neighbors are AZ, NY and OR



State with lowest Strength is AK
No 1 Hop Neighbors

Input: $w=10$, $r=4$, $h=2$, $\alpha=0.5$, $\text{fileSelected}=2$, $\text{filePath} = \text{'MWDB_Phase1/Output/epidemic_word_file_diff.csv'}$



State with highest Strength is NJ
1 Hop Neighbors are NY and PA

State with lowest Strength is AL
1 Hop Neighbors are GA, MS, TN

System requirements and execution instructions

MATLAB 2012 or 2013 should be installed on the system.

To run the program:

- Unzip the folder and add the path of the directory to Matlab.
- Make sure “*sort_nat.m*” is in the same folder.
- Open the script “*phase1_main.m*” and run.
- Input desired input when prompted.
- The output files will be in the “*Output*” folder.

Conclusion

Based on the input simulation files, I created 3 epidemic files: *epidemic_word_file*, *epidemic_word_file_avg* and *epidemic_word_file_diff*. For any input simulation file selected by the user, I generated its heatmap and highlighted highest and lowest strength states on it. The heatmaps for the word and average files are identical but the heatmaps for *epidemic_word_file_diff* is different.

References

1. www.wikipedia.com
1. http://www.webopedia.com/TERM/S/spatial_data.html

Appendix

It was an individual project. The whole team contributed in open discussions though.