Spatiotemporal Analysis of Location-based Social Media Data

Guofeng Cao

Department of Geosciences
Texas Tech University
guofeng.cao@ttu.edu

Fall 2015

Location-based social media

- Social media has been experiencing a spectacular rise and popularity in the past several years
- With the widespread of mobile devices, location has become a critical component of user-generated social media

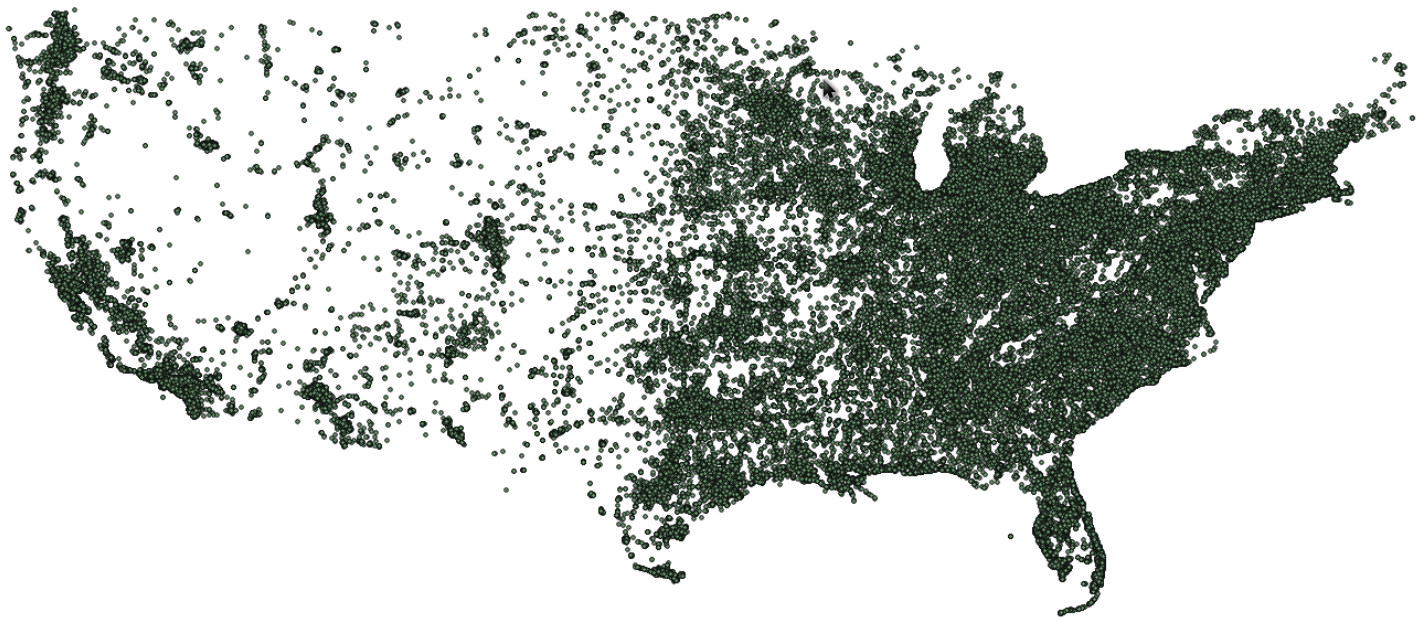
Opportunities of location-based social media

Location-based social media

- A particular type of volunteered geographic information (VGI)
- Location-based social media (LBSM) data offer a unique opportunity to investigate complex social dynamics at multiple spatiotemporal and social scales
  - individual and continuous
  - group of population (population vs subpopulation)
  - spatiotemporal aggregation (modified areal unit problem)
- Extensive studies with significant societal impacts have been recently reported based on location-based social media data, e.g.,
  - public health surveillance, digital epidemiology
  - human mobility
  - political survey
  - ...

Challenges

- Characteristics of LBSM data pose fundamental representation, modeling and computational challenges to GIScience, spatiotemporal databases and analysis
  - generated by a massive number of social media users, often "big" and "coming" continuously
  - dynamic and real-time
  - unstructured forms of media (e.g., text, photos and media)

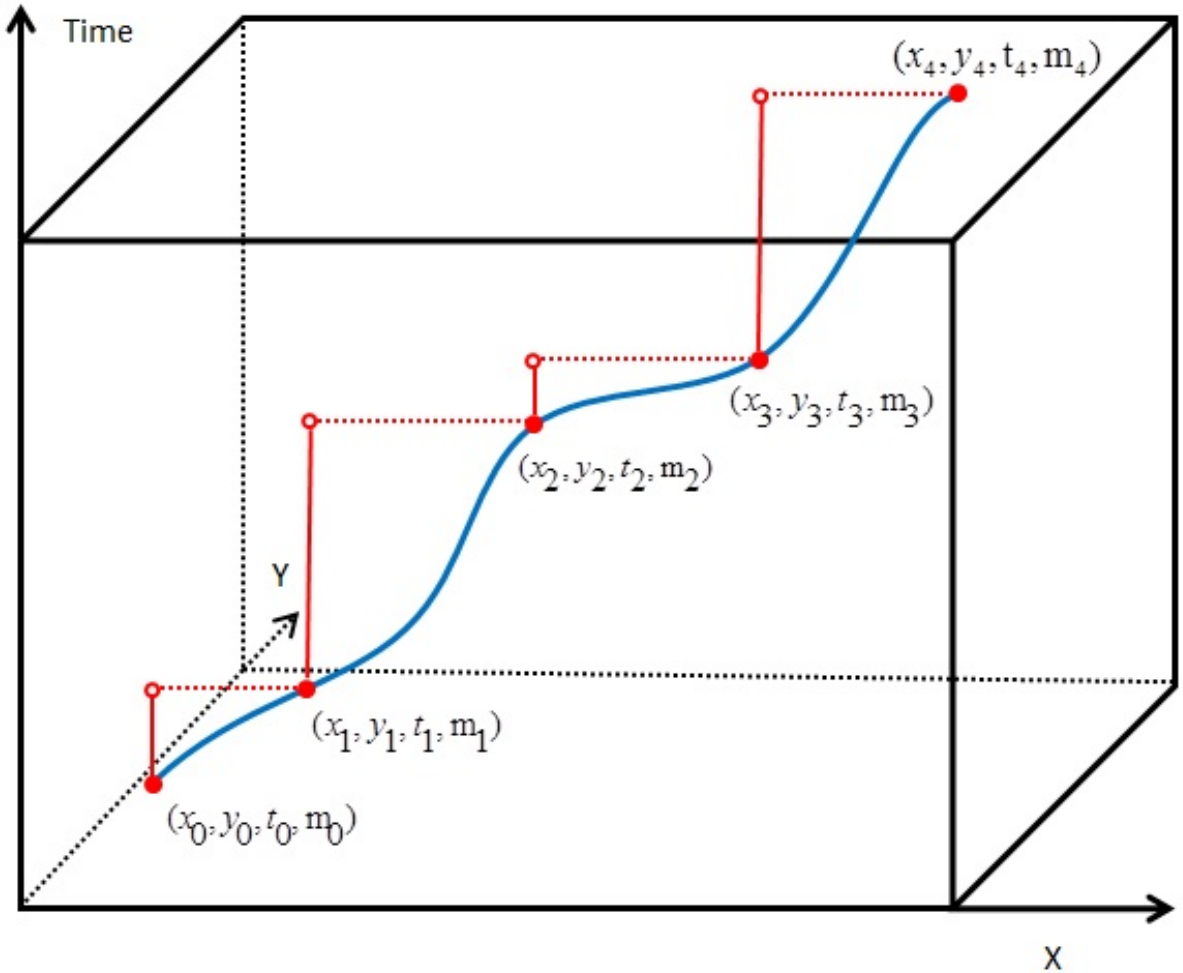


In this presentation

To address these challenges, we develop a scalable computational framework to harness the massive location-based social media data, specifically Twitter posts, to support systematic and efficient analysis of spatiotemporal dynamics.

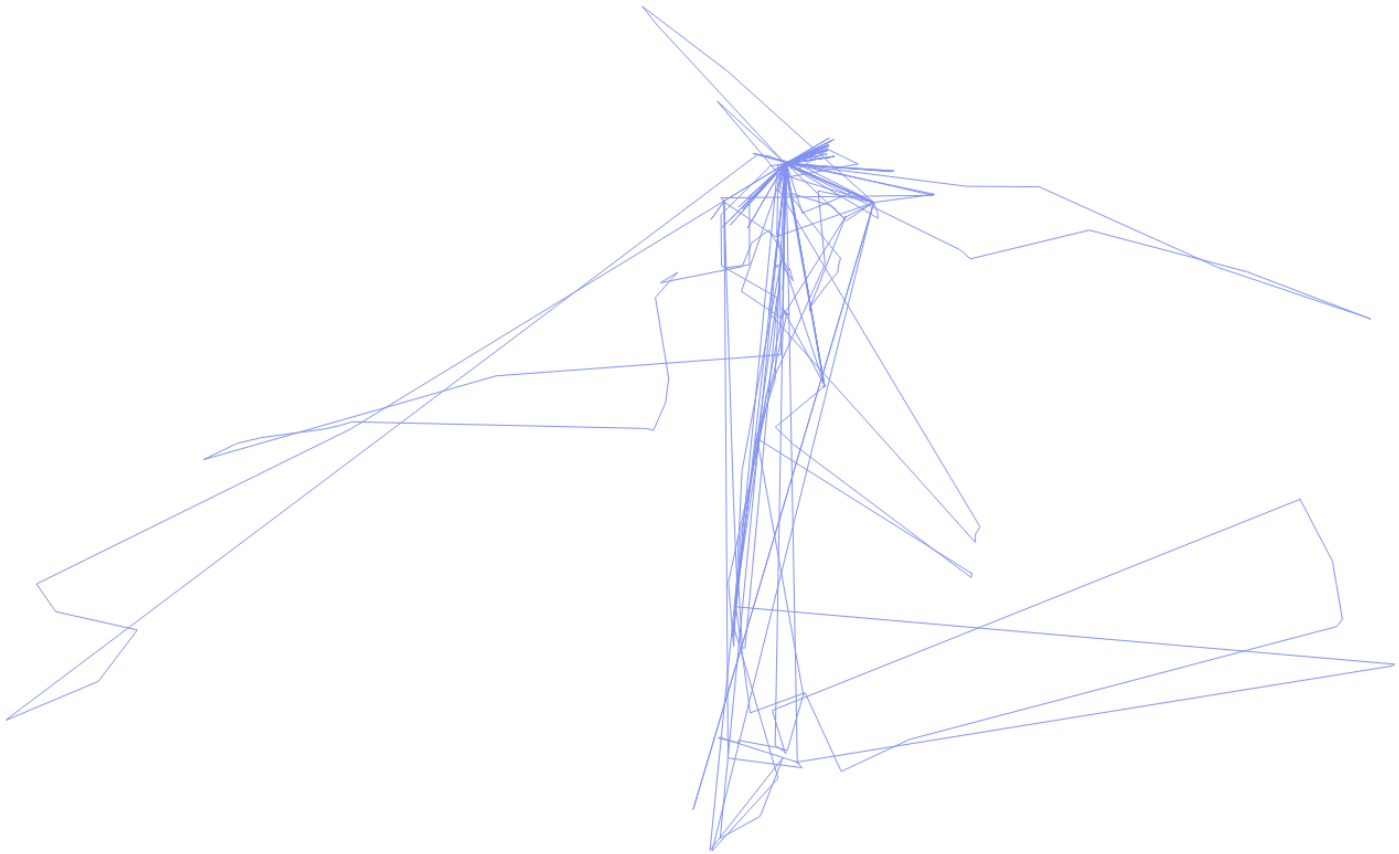
- Space-time trajectories
- Text mining (e.g., health status, demographic information, sentimental analysis)
- A hierarchical multi-scale spatiotemporal data cube
- Analysis of spatiotemporal data cube

Space-time trajectories



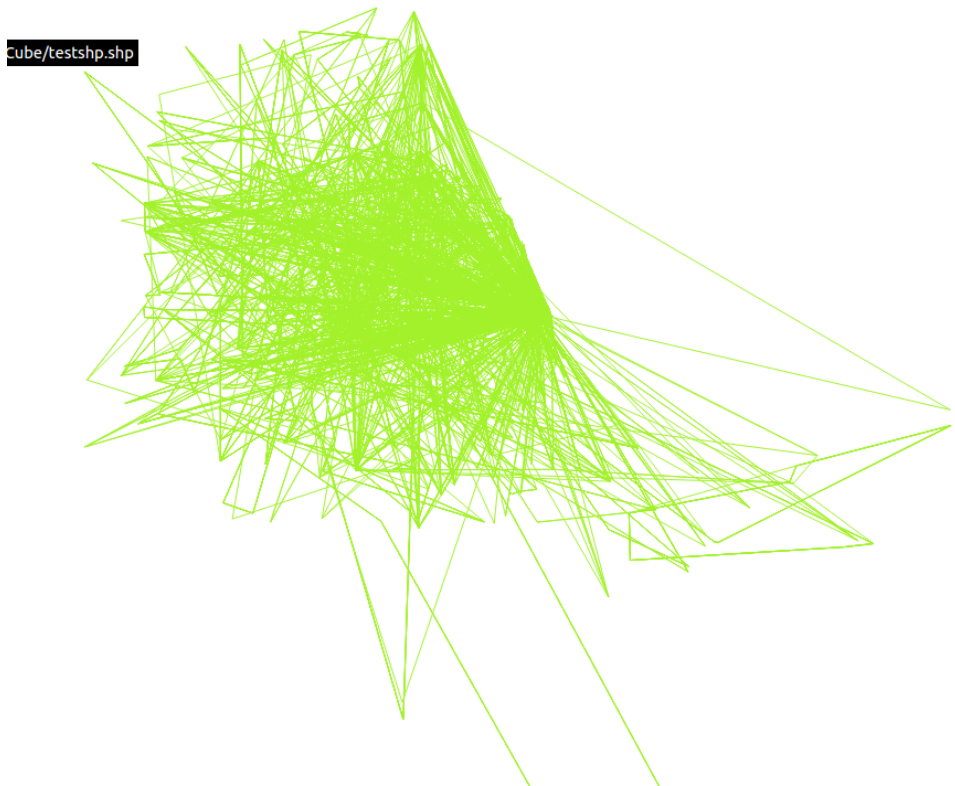
An illustration of space-time path constructed from location-based social media data in a space-time cube of Hägerstrand; a space-time trajectory is essentially a collection of step functions.

Trajectory examples from Twitter



A trajectory example from Twitter collections.

Trajectory examples from Twitter



A trajectory example from Twitter collections.

Trajectory examples from Twitter

A trajectory example from Twitter collections.

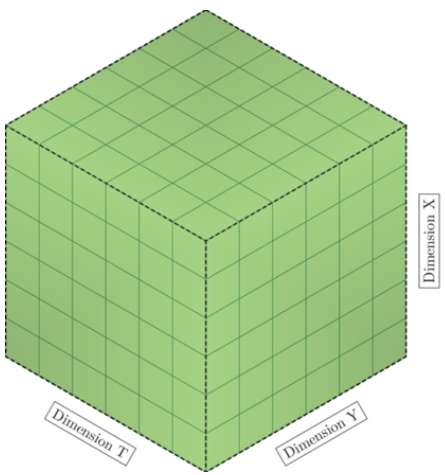
Elements of space-time trajectories

- Identifiers and spatiotemporal footprint
- Frequently visited locations or mode locations, e.g. *home*
- *Radius of gyration*
- Compared to the trajectories constructed from GPS logs or mobile phone records, space-time trajectories based on location-based social media provide access to the contents of messages and activities, which provide clues of latent attributes of social media users, e.g., *health statuses*, *socio-demography*, and *opinions* on specific subjects
- Here we focus on infection spread of *ILI*, and diagnose the chance that an individual is ILI affected by monitoring text contents of social media messages

A spatiotemporal data cube model

Spatiotemporal data cube is developed to support efficient management of aggregated statistics

- Data cube decomposes the spatiotemporal space into a lattice of multi-scale, hierarchical *cuboids*, with *base cuboids* representing primitive compartments in multidimensional spaces at the finest level
- Provides multiple scales of spatial index for efficient spatial query
- Given a region, statistics such as the number of social media users, and number of trajectories traveling out and in from this region are efficiently retrieved by data cube operations, e.g. merging/splitting cuboids



A spatiotemporal data cube model: elements

Given a spatiotemporal cuboid c and social media users u, we specifically defined the following measures/statistics:

1. $R(c)$ (residents): the number of distinct social media users in u whose homes locate within spatial boundary of c;
2. $V(c)$ (visitors): the number of distinct social media users in u that post activities in c;
3. $A(c)$ (activities): of social media activities by individuals in u occurring in c;
4. $O(c)$ (out): the number of moves made by u from c to other cells;
5. $I(c)$ (in): the number of moves made by u into c from other cells;
6. $S(c)$ (centroid): the expected location of social media activities by individuals in u occurring in c;
7. $V_{flu}(c)$ (occurrences): the number of distinct social media users in u that post activities in c diagnosed as ILI affected occurrences;

Notes:

- Apparently, $O(c) \leq V(c)$, $I(c) \leq V(c)$, $V_{flu}(c) \leq V(c)$ and $V(c) \leq A(c)$.

A spatiotemporal data cube model: elements

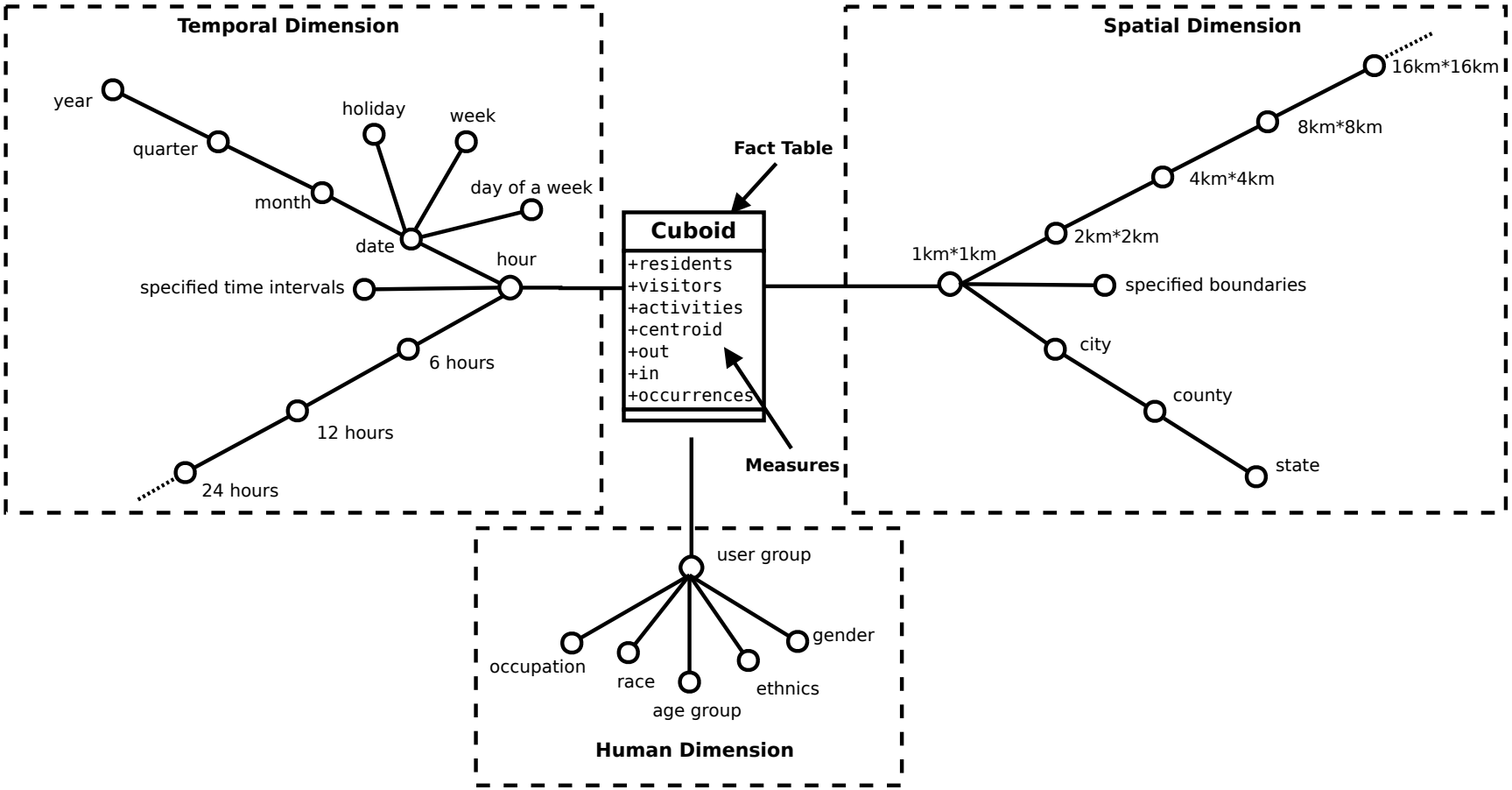
Similarly for a pair of spatiotemporal cuboid c_i and c_j, we specifically defined the following measures/statistics:

1. $F(c_i, c_j)$ (travel flows): the number of *moves* made by social media users u starting from cuboid c_i and ending in cuboid c_j;
2. $F_{flu}(c_i, c_j)$ (flu travel flows): the number of *moves* made by social media users u starting from cuboid c_i and ending in cuboid c_j made by ILI occurrences;
3. $F_{migration}(c_i, c_j)$ (migration flows): the number of social media users in u migrating home location from cuboid c_i to cuboid c_j;

Notes:

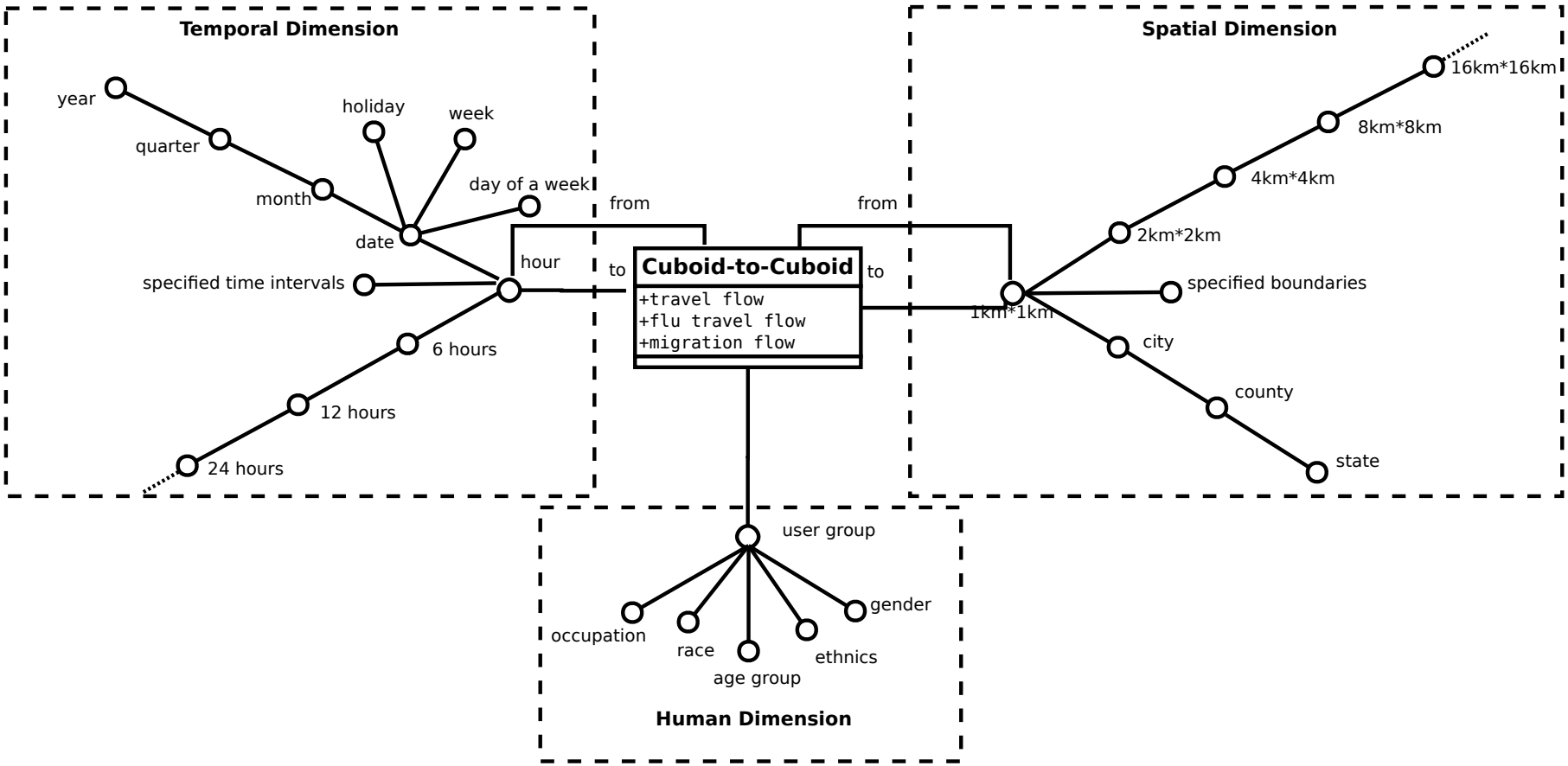
- Apparently, $F_{flu}(c_i, c_j) \leq F(c_i, c_j)$ and $F_{migration}(c_i, c_j) \leq F(c_i, c_j)$ and all of these three flow measures are asymmetry
- Flow is an aggregation of space-time trajectories based on points of trajectories instead of the segments in-between

A spatiotemporal data cube model: conceptual model



Fact schema of a spatiotemporal data cube (cuboid)

A spatiotemporal data cube model: conceptual model



Fact schema of a spatiotemporal data cube (cuboid-to-cuboid)

A spatiotemporal data cube model: change of scales

Suppose for two disjoint cuboids $\boldsymbol{c}_1 = \cup_{i=1}^{K} \boldsymbol{c}_{1,i}$, and $\boldsymbol{c}_2 = \cup_{j}^{K} \boldsymbol{c}_{2,j}$, the super-aggregates of A, F, F_{flu}, $F_{migration}$, can be written as recursive functions of sub-aggregates:

$$A(\boldsymbol{c}_1) = \sum_{i=1}^{K} A(\boldsymbol{c}_{1,i}) \tag{1}$$

$$F(\boldsymbol{c}_1, \boldsymbol{c}_2) = \sum_{i,j=1}^{i,j=K} F(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{2,j}) \tag{2}$$

$$F_{flu}(\boldsymbol{c}_1, \boldsymbol{c}_2) = \sum_{i,j=1}^{i,j=K} F_{flu}(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{2,j}) \tag{3}$$

$$F_{migration}(\boldsymbol{c}_1, \boldsymbol{c}_2) = \sum_{i,j=1}^{i,j=K} F_{migration}(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{2,j}) \tag{4}$$

A spatiotemporal data cube model: change of scales

For measure S, I and O, the super-aggregates $\boldsymbol{c}_1$ needs the support of other measures: Specifically for S:

$$S(c_1) = \frac{1}{A(c_1)} \sum_{i=1}^{K} A(c_{1,i}) S(c_{1,i}) = \frac{\sum_{i=1}^{K} A(c_{1,i}) S(c_{1,i})}{\sum_{i=1}^{K} A(c_{1,i})} \tag{5}$$

For measures I and O on $\boldsymbol{c}_1$, we need to remove the space-time trajectories that occurred within the boundaries of $\boldsymbol{c}_1$ according to the definition of I and O. Hence, we have:

$$O(\boldsymbol{c}_1) = \sum_{i=1}^{K} O(\boldsymbol{c}_{1,i}) - \sum_{i=1}^{K} \sum_{j=1}^{K} F(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{1,j}) \tag{6}$$

$$I(\boldsymbol{c}_1) = \sum_{i=1}^{K} I(\boldsymbol{c}_{1,i}) - \sum_{i=1}^{K} \sum_{j=1}^{K} F(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{1,j}) \tag{7}$$

A spatiotemporal data cube model: change of scales

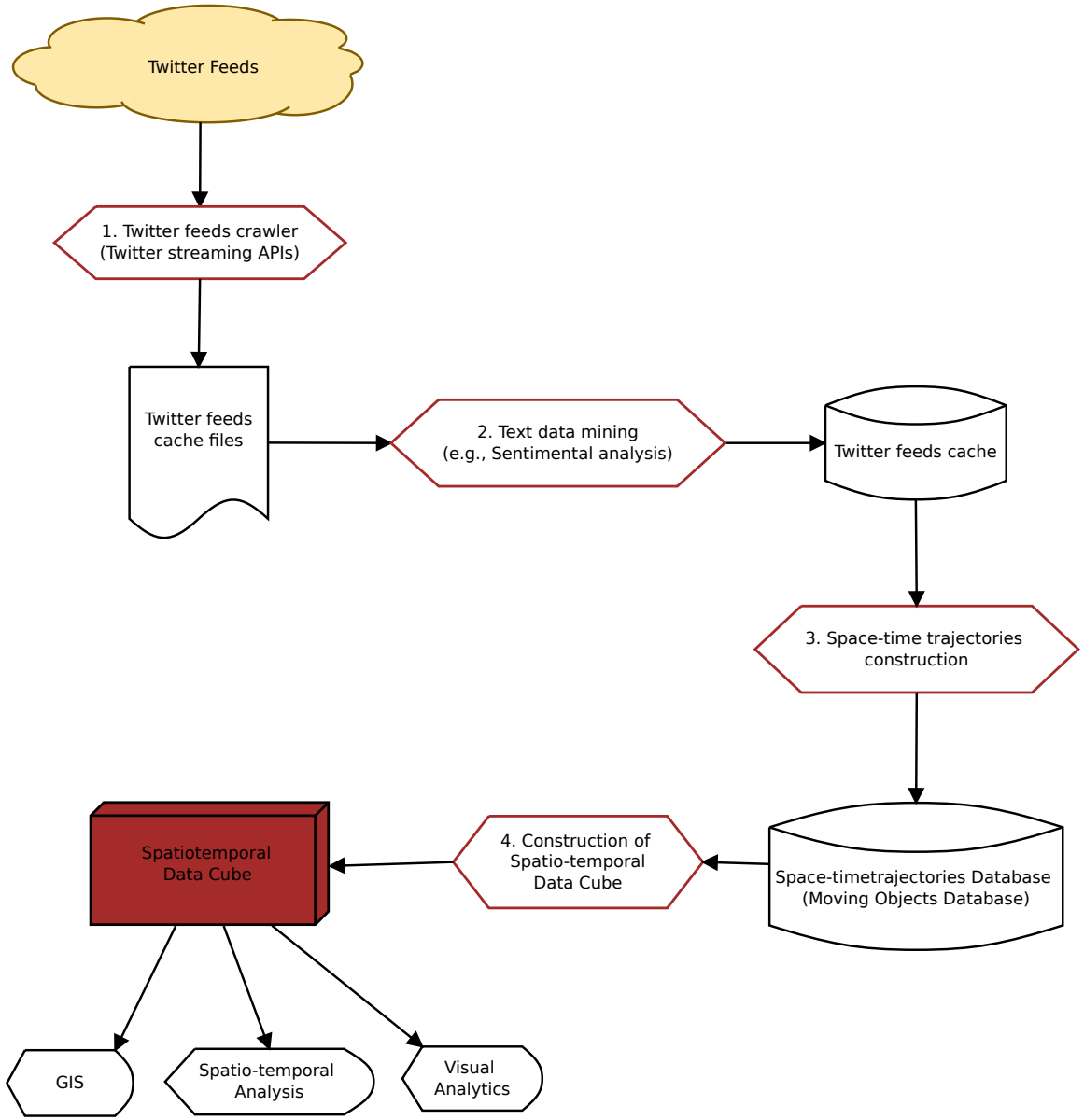
The super-aggregates of R, V and V_{flu} cannot be obtained as a recursive function of sub-aggregates, which amounts to the *distinct counting problem* and needs approximation:

$$V(\boldsymbol{c}_1) = \sum_{i=1}^{K} V(\boldsymbol{c}_{1,i}) - \sum_{i=1}^{K}\sum_{j=1}^{K} F(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{1,j}) \tag{8}$$

$$V_{flu}(\boldsymbol{c}_1) = \sum_{i=1}^{K} V_{flu}(\boldsymbol{c}_{1,i}) - \sum_{i=1}^{K}\sum_{j=1}^{K} F_{flu}(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{1,j}) \tag{9}$$

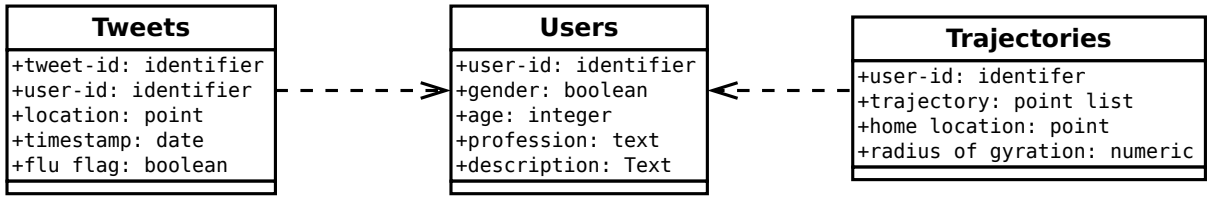
$$R(\boldsymbol{c}_1) = \sum_{i=1}^{K} R(\boldsymbol{c}_{1,i}) - \sum_{i=1}^{K}\sum_{j=1}^{K} F_{migration}(\boldsymbol{c}_{1,i}, \boldsymbol{c}_{1,j}) \tag{10}$$

Implementation

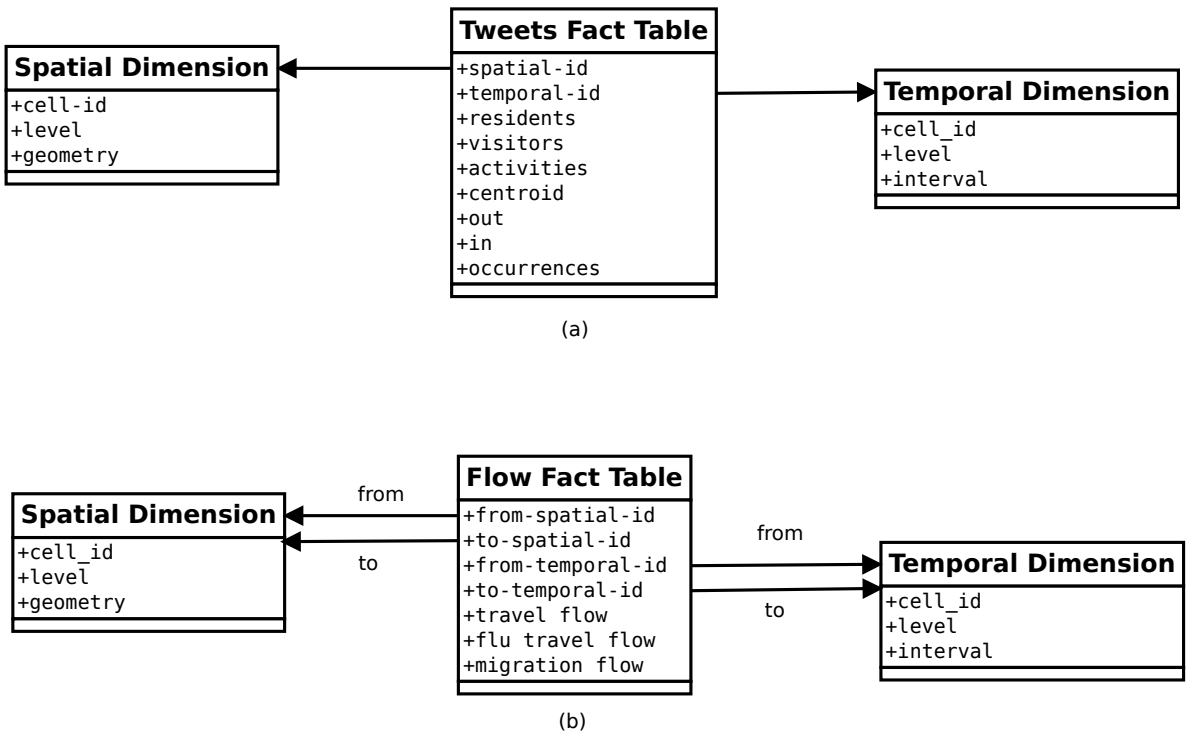


Framework architecture

Implementation

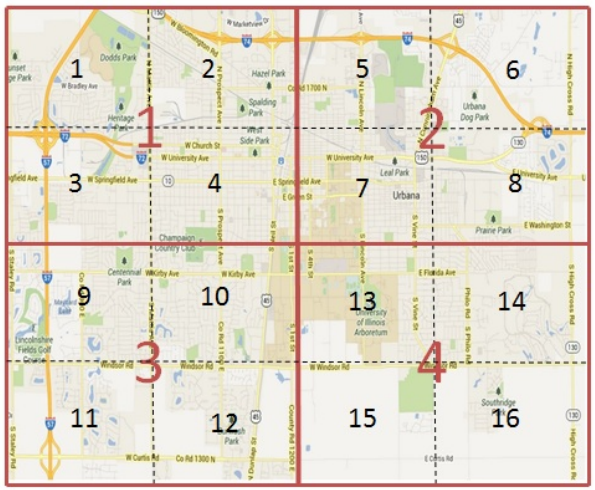


Schema of a space-time trajectories database



Schema of a spatiotemporal data cube for Twitter feeds. There are two fact tables, one is for the facts of a single cuboid (a), and the other for facts of flows between cuboids (b).

Implementation



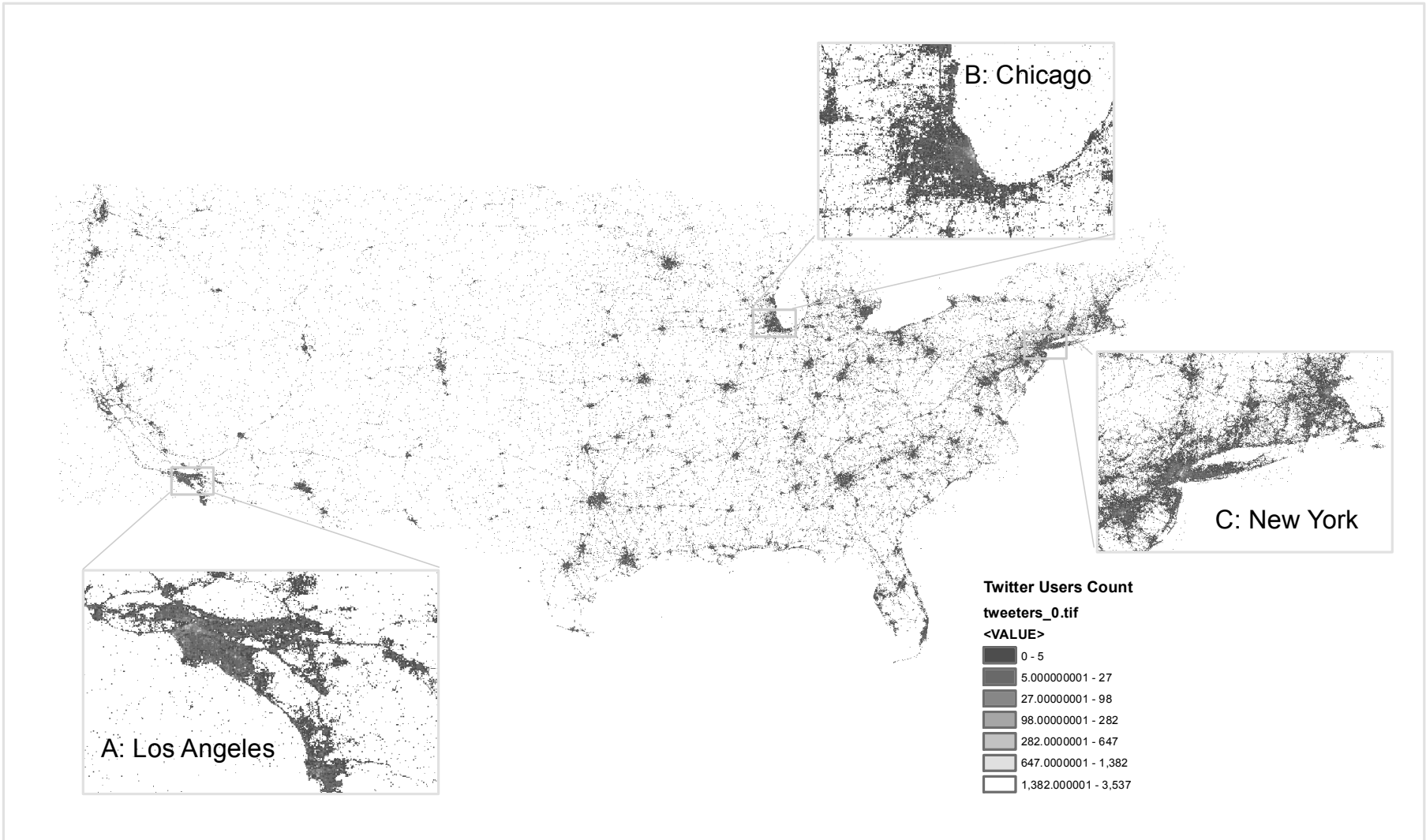
(a)

spatial: level	spatial: cell-id	temporal: level	temporal: cell-id	visitors	activities	residents	...
1	1	1	1	5	10	8	
1	2	1	1	3	5	6	
1	3	1	1	2	4	3	
1	4	1	1	5	8	7	
2	1	1	1	10	22	24	
1	5	1	1	5	6	5	
1	6	1	1	4	9	7	
...	...	1	1	...	...	...	
2	2	1	1	12	20	18	

(b)

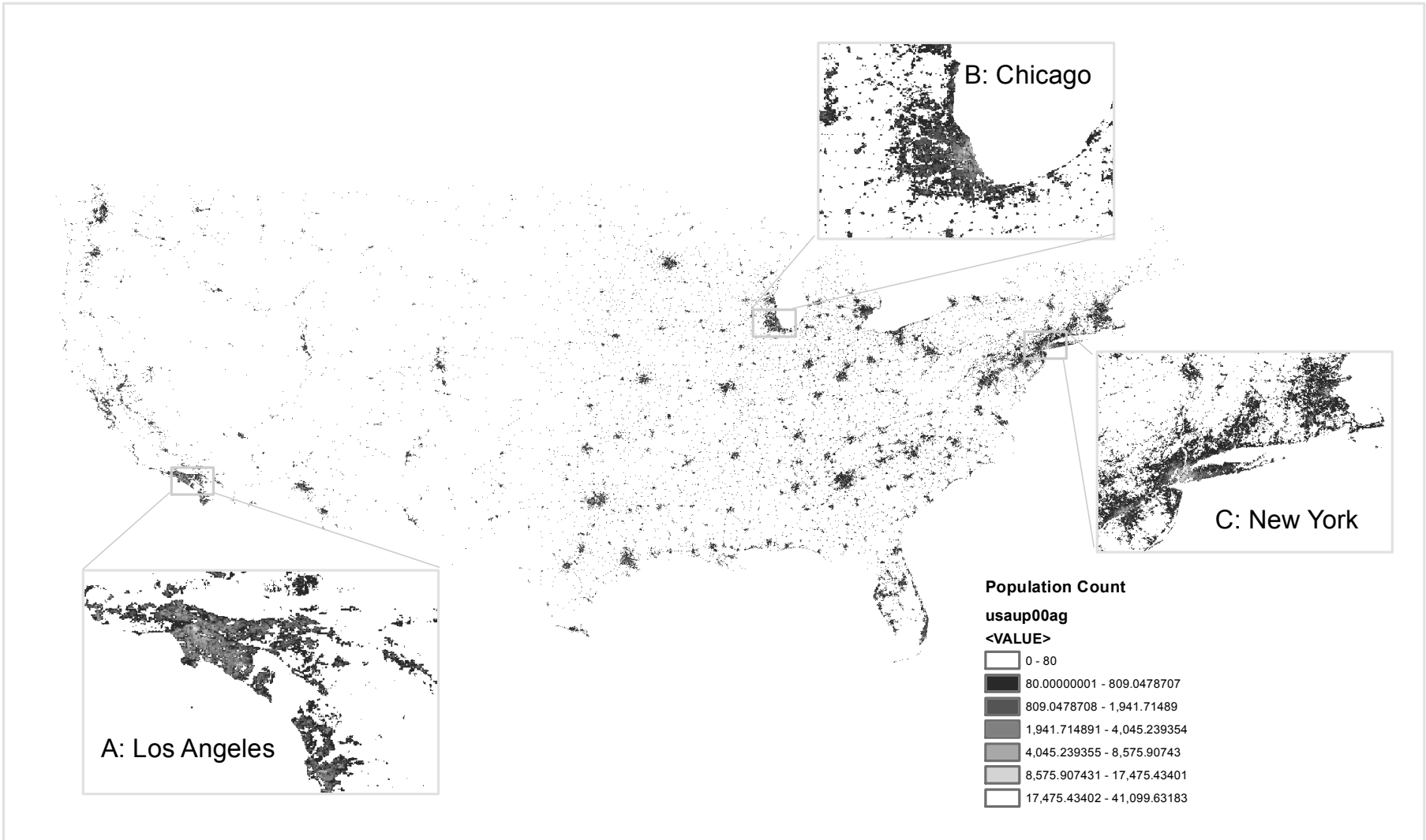
An example table of a data cube with two levels of hierarchy on spatial dimension. (a) shows an example map and the boundaries of two levels of spatial hierarchy. Black dashed lines indicate the spatial boundaries of the cuboids at (fine) level 1, and red solid lines indicate the spatial boundaries of cuboids at (coarse) level 2; (b) is the fact table associated with each cuboid defined in (a).

Exploration of spatiotemporal data cube



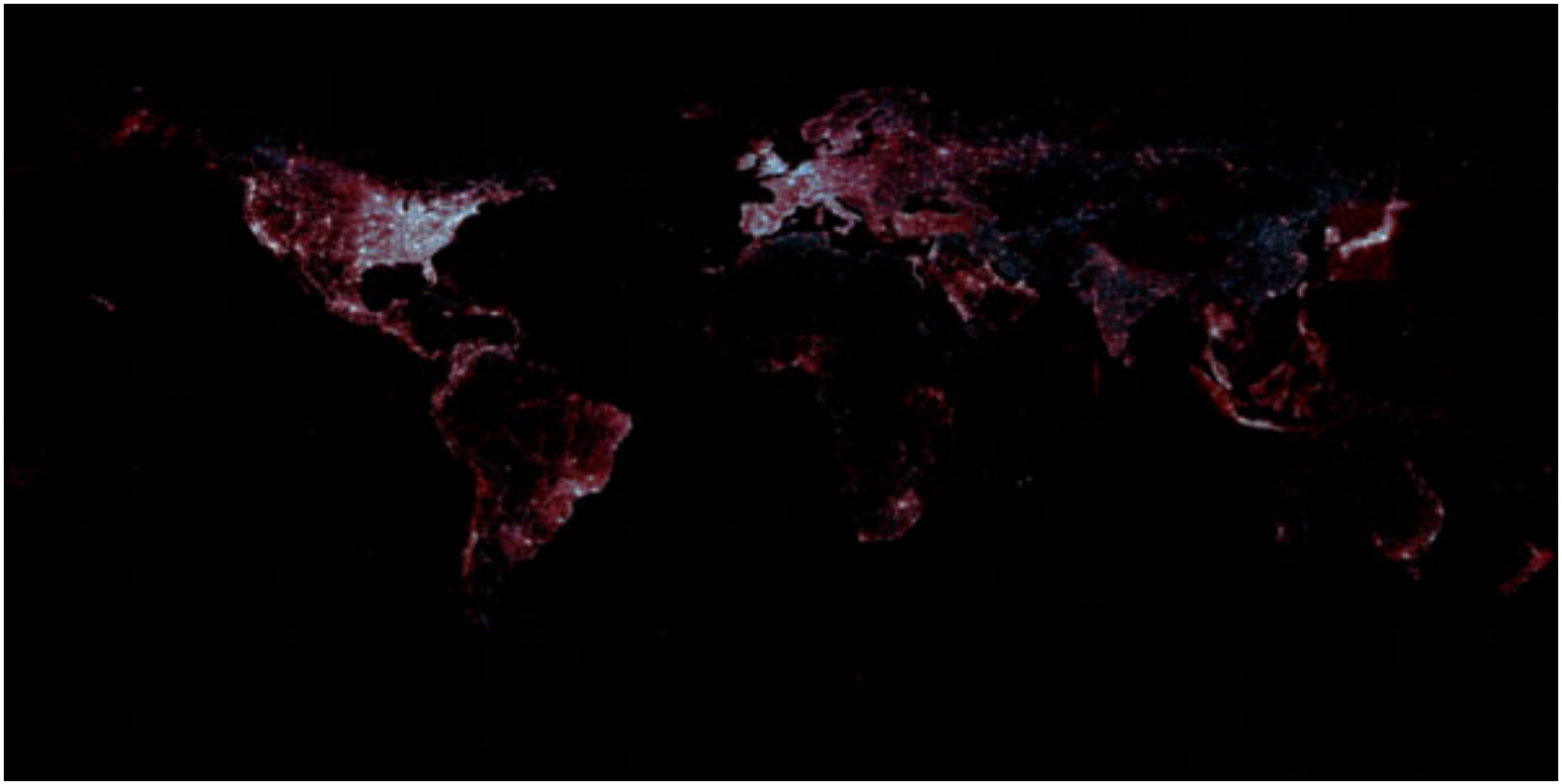
Number of distinct Twitter users (during 17:00 June 22 2013 to 16:59 June 29 2013) in the finest level of spatio-temporal data cube ($1km \times 1km$).

Exploration of spatiotemporal data cube



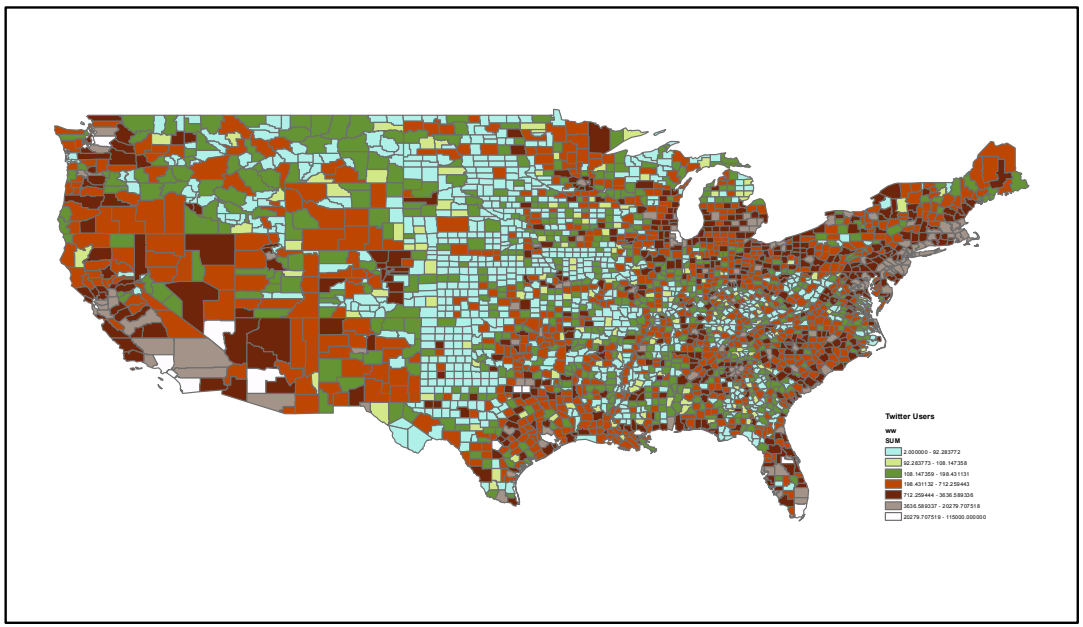
Population distribution of Global Rural-Urban Mapping Project (GRUMP) with more than 80 people per cell ($1km \times 1km$).

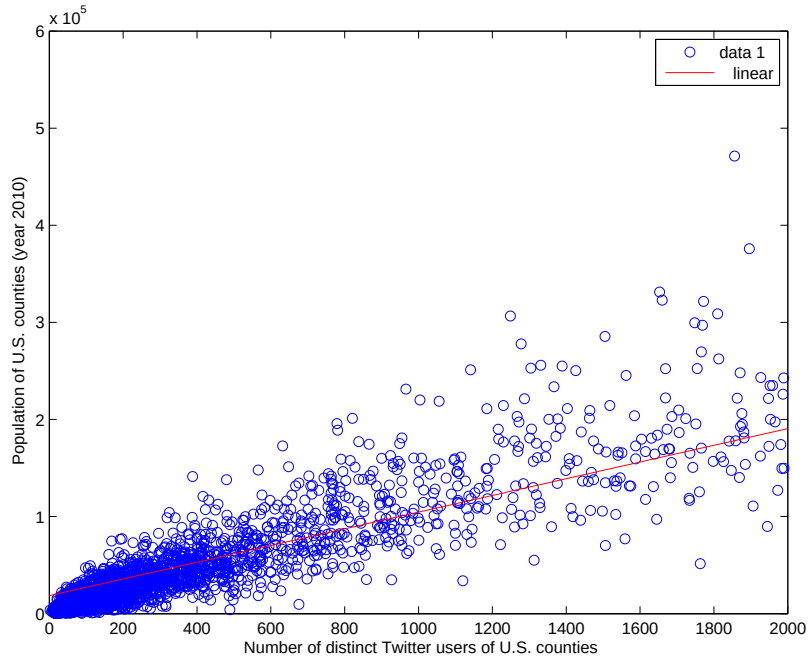
Exploration of spatiotemporal data cube



Comparison of georeferenced tweets from the Twitter Decahose 23 October 2012 to 30 November 2012 with NASA Visible Earth imagery (red areas overemphasize tweeting/blue underemphasize), correlation coefficient $r = 0.79$.

Exploration of spatiotemporal data cube





Left: Number of distinct Twitter users (during 17:00 June 22 2013 to 16:49 June 29 2013) aggregated by U.S. counties. Right: Correlation between the number of distinct Twitter users and population of U.S. counties.

Visual analytics of spatiotemporal data cube

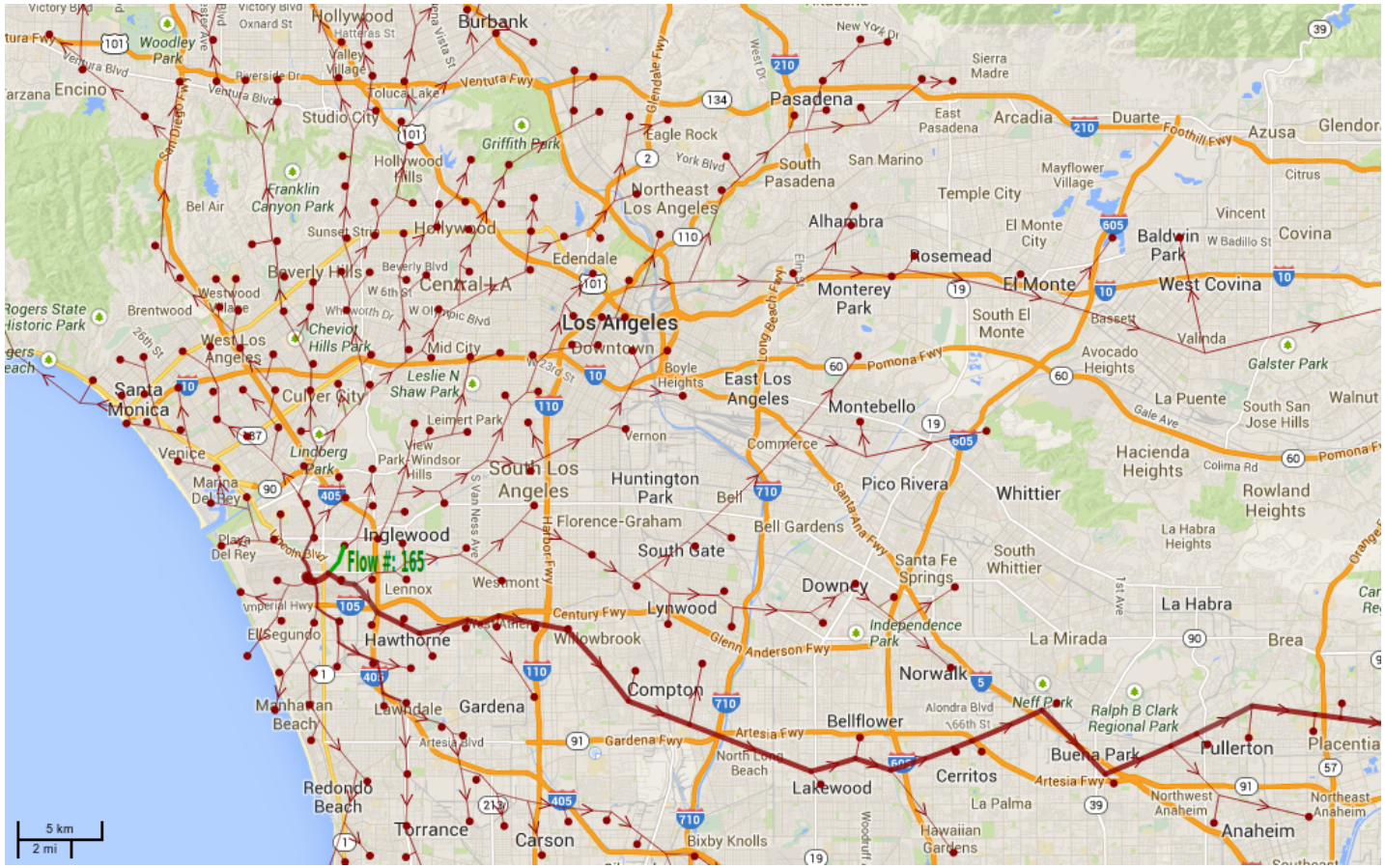
Flow mapping is a widely used visual analytical method to depict and represent geographical dynamics of movement

- Single-source flow mapping
- Multiple-source flow mapping

Flow mapping for movement dynamics represented in spatiotemporal data cube

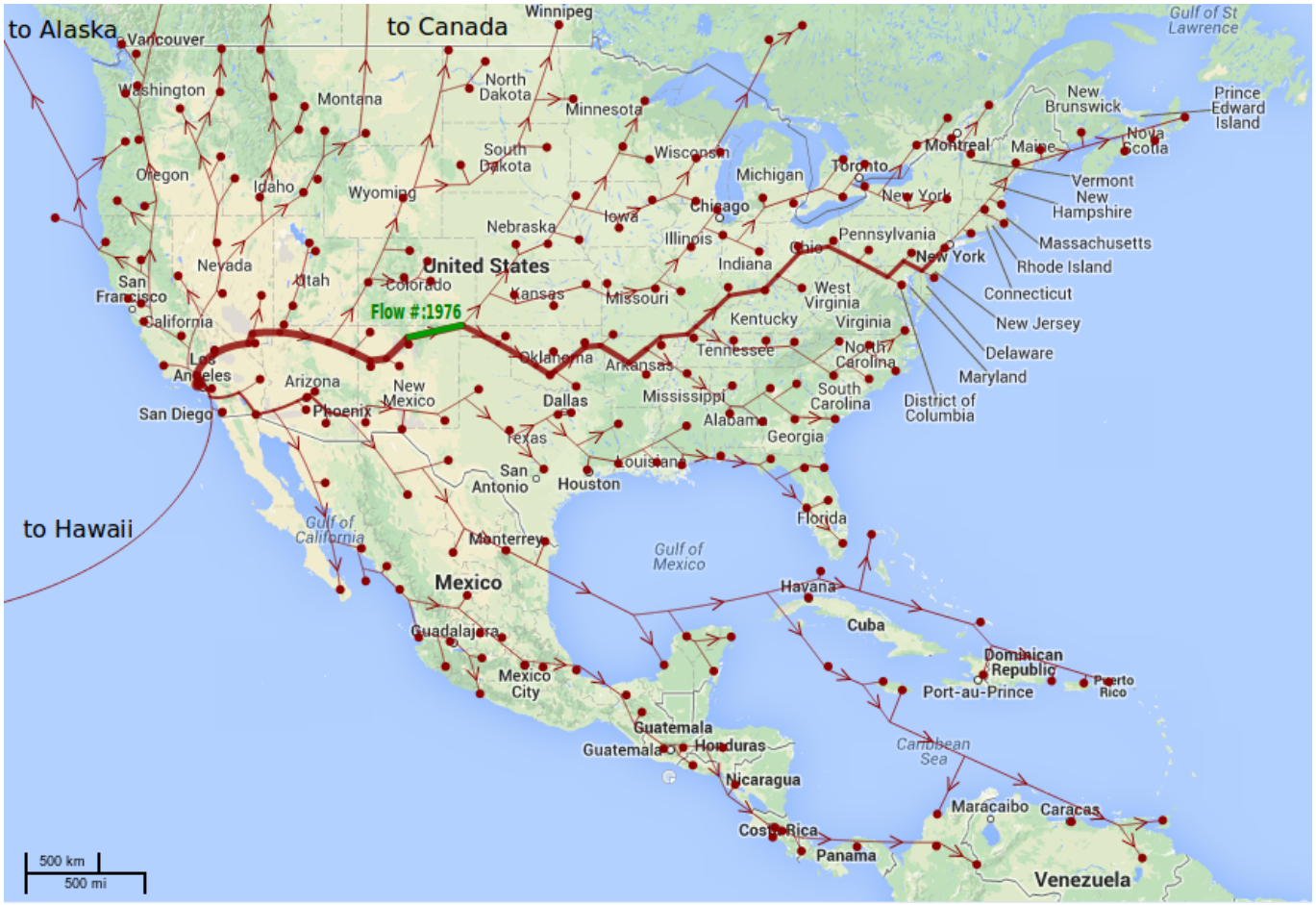
- Based on location-based Twitter feeds posted in the continent of North America
- Visual exploration of movement dynamics across multiple spatiotemporal scales, from macro migration trends across the globe to characteristics of individual daily activity
- Interactive, near real-time on-line interface

Travel flow from LAX at city levels:



A flow map of number of travels during seven days (January 29th to February 5th, 2014) from the LAX neighborhood to the rest area of Los Angeles.

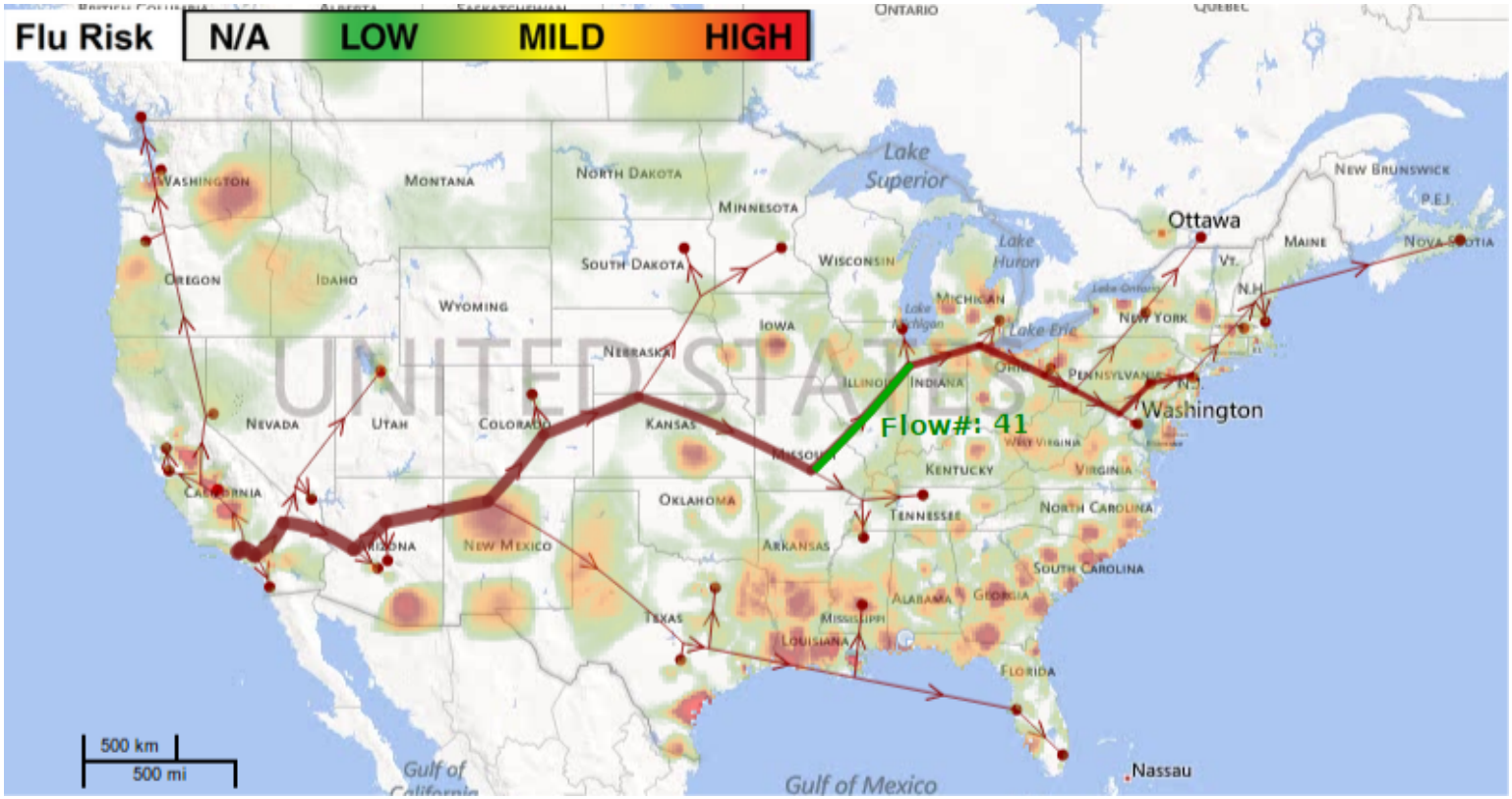
Travel flow from Los Angeles at continent levels:



A flow map of number of travels during seven days (January 29th to February 5th, 2014) from the Los Angeles to the other areas of North America.

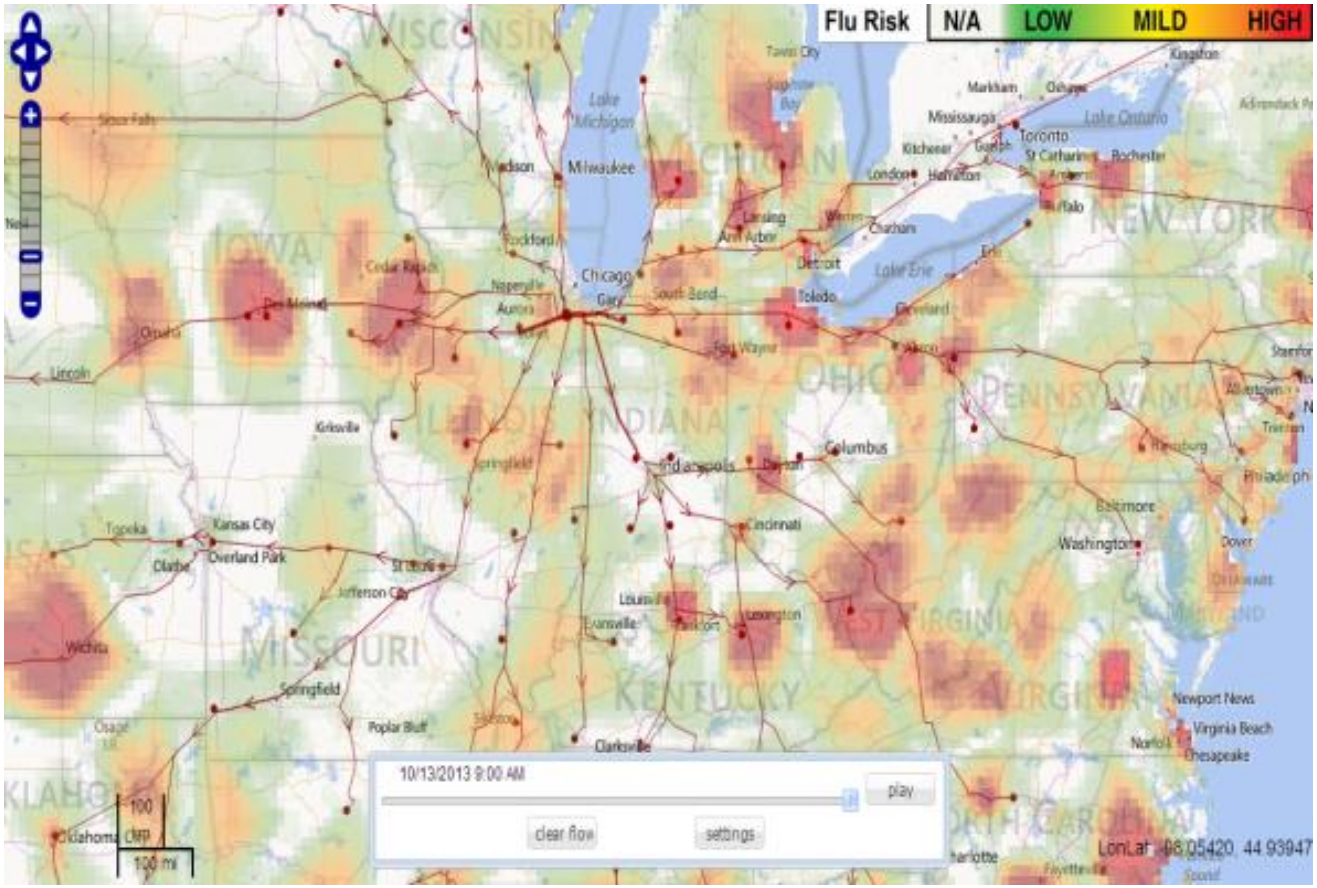
Single-source flow mapping III

Travel flows with flu from Los Angeles at continent scales:



A flow map of number of travels made by potential flu-affected Twitter users during seven days (January 29th to February 5th, 2013) from the Los Angeles to the other areas of North America.

Flow on ground road network:



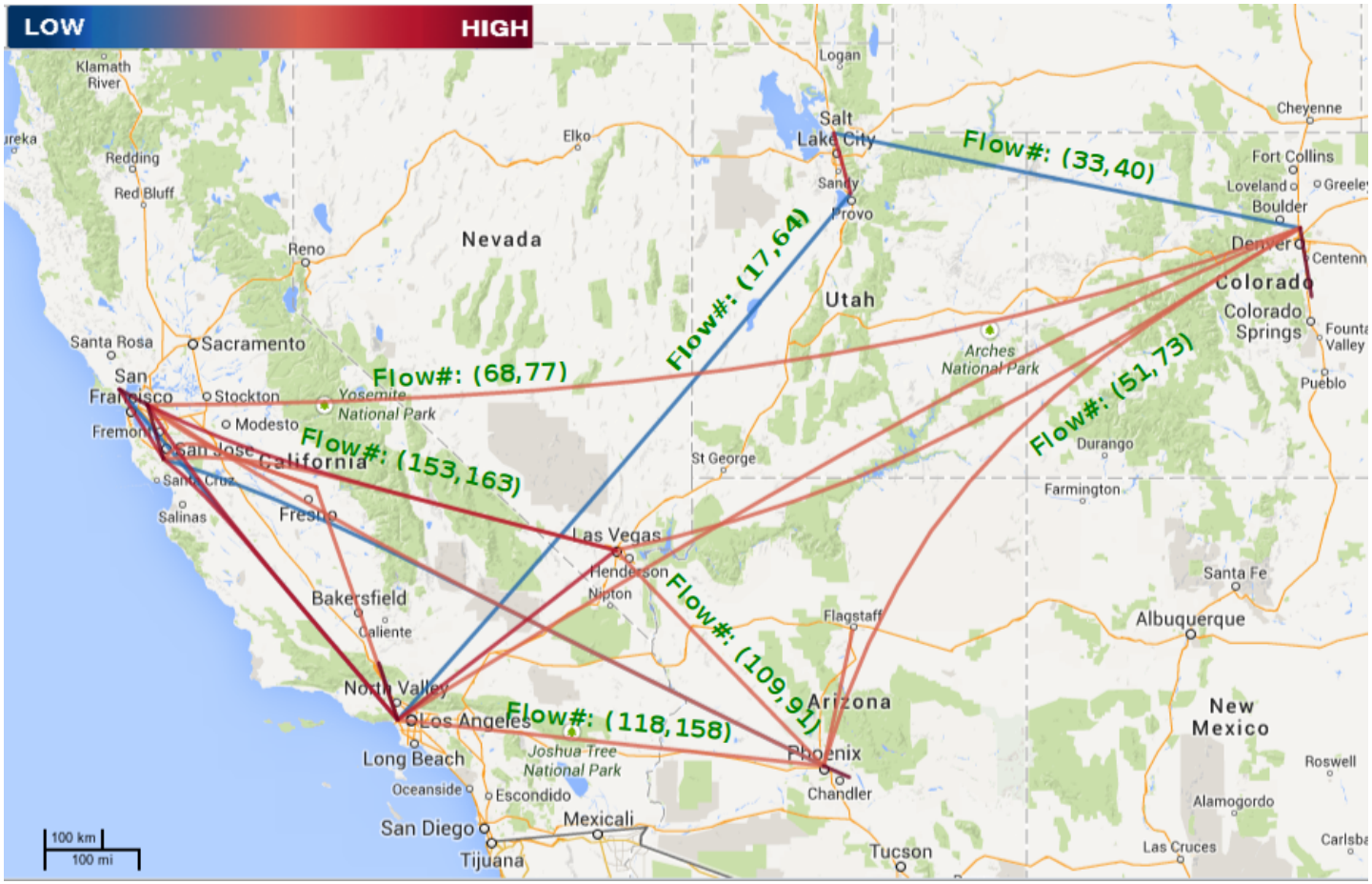
Visualization of Twitter users on ground movements between 2013/10/7 9:00am to 2013/10/13 9:00am

Flow on ground road network:

Visualization of Twitter users on ground movements between 2013/10/7 9:00am to 2013/10/13 9:00am

Multiple-source flow mapping I

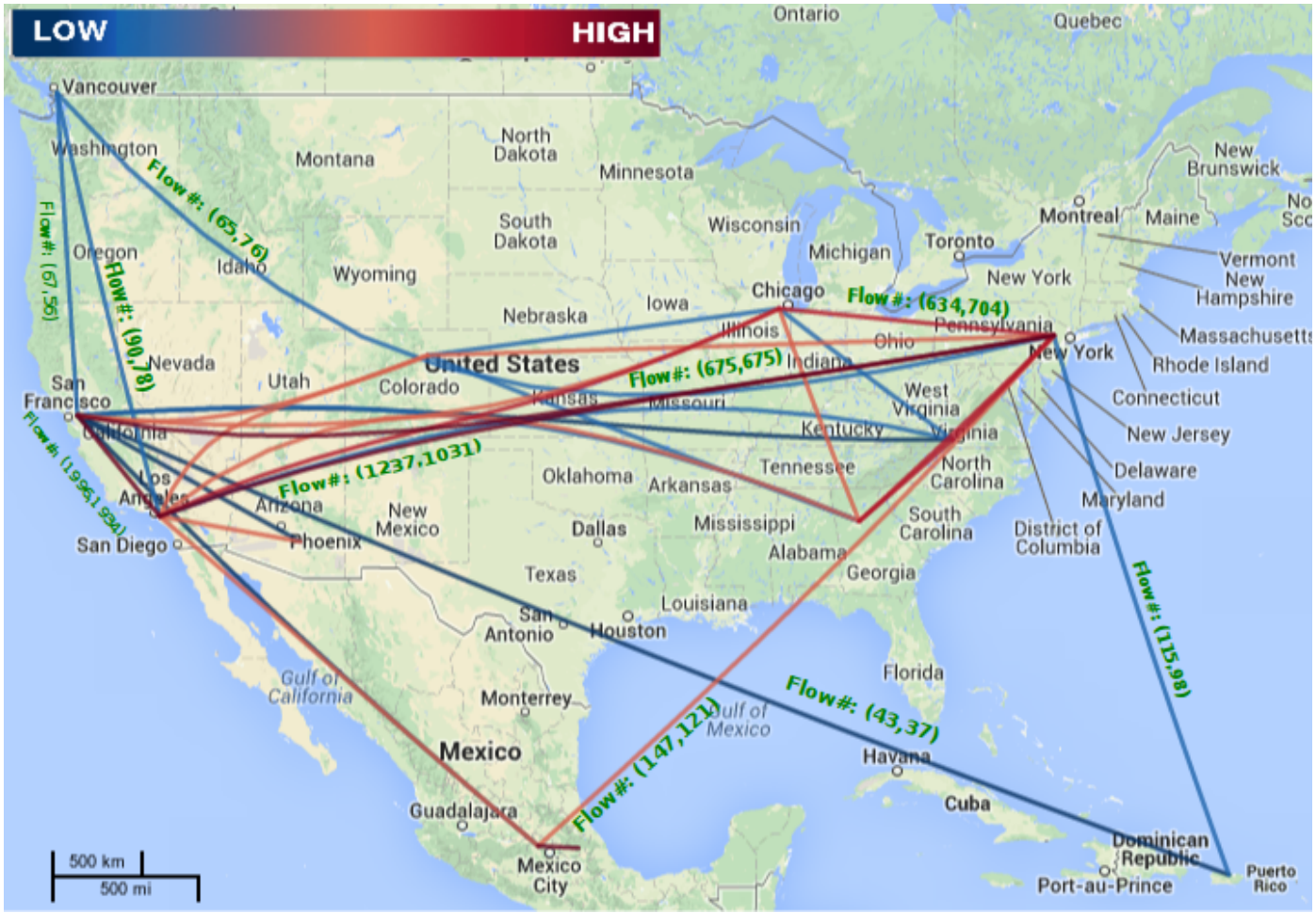
Regional scales:



Multiple-source flow maps of the travel flows between major cities in the southwest of the United States during the 22:00 of January 31st, 2014 to the 21:59 of February 7th, 2014.

Multiple-source flow mapping II

Continent scales:



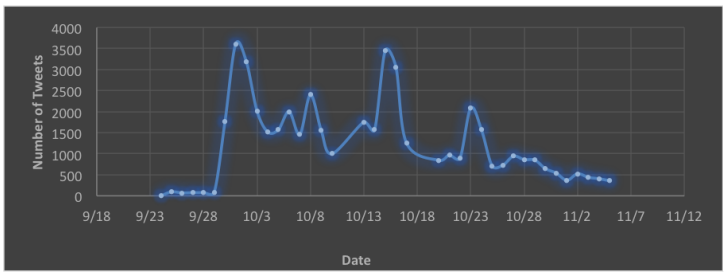
Multiple-source flow maps of the travel flows of the major cities in North America during the 22:00 of January 31st, 2014 to the 21:59 of February 7th, 2014.

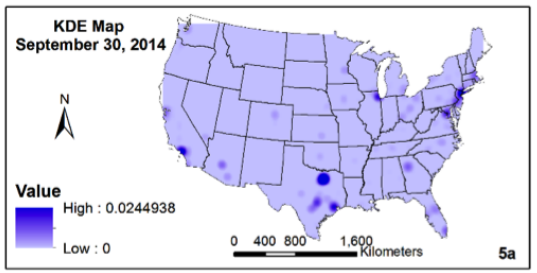
Conclusion and ongoing work

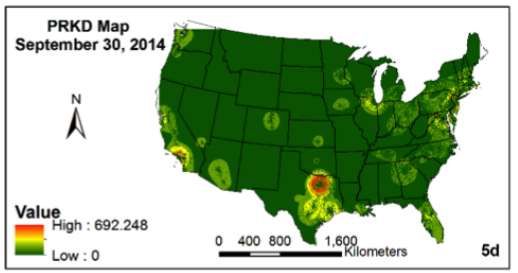
- A general framework to harness the massive location-based social media data for scalable and efficient spatiotemporal analysis of massive location-based social media data.
  - space-time trajectories
  - Text mining (health status, demographic information, sentimental analysis)
  - a hierarchical multi-scale spatiotemporal data cube
  - visual analytics of spatiotemporal data
- The framework transforms unstructured location-based social media to be easily integrated with conventional GIS data sources and spatiotemporal analysis

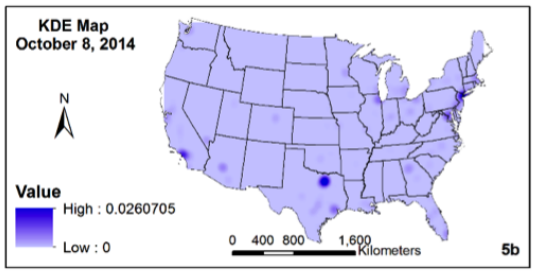
Ongoing related work I

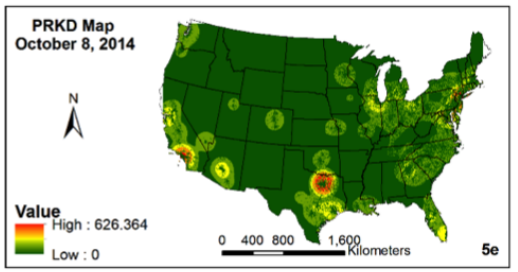
- Spatiotemporal event detection, e.g., Ebloa outbreak

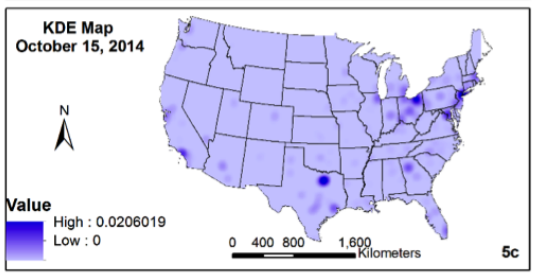


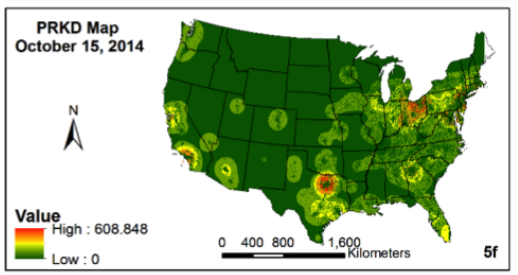






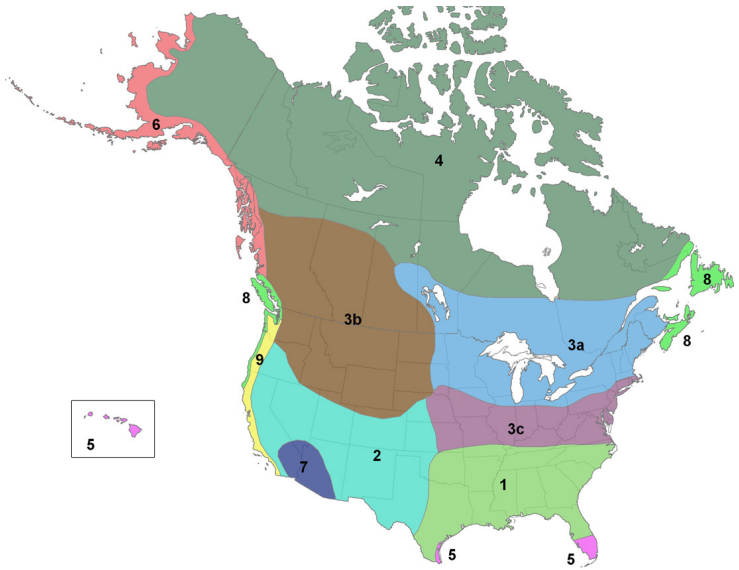






Ongoing related work II

- Incorporating environmental and climatic observations into digital epidemiology



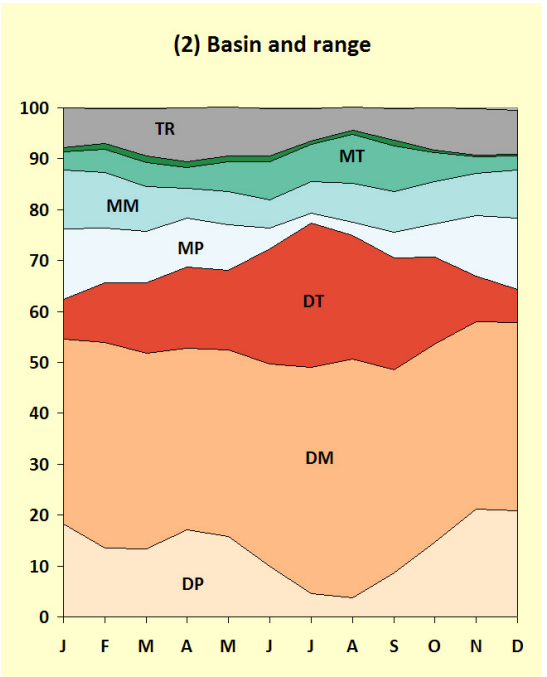
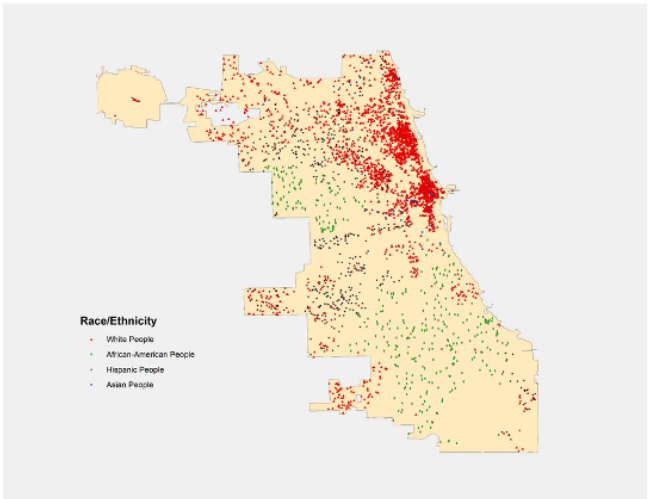


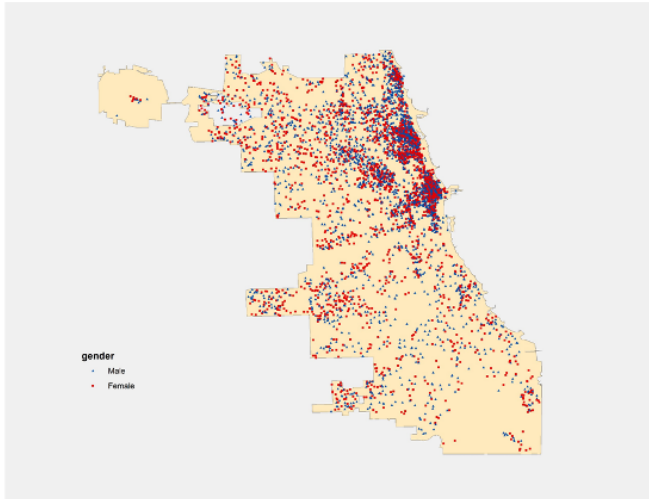
Image courtesy of Scott Sheridan (Kent State)

Ongoing related work III

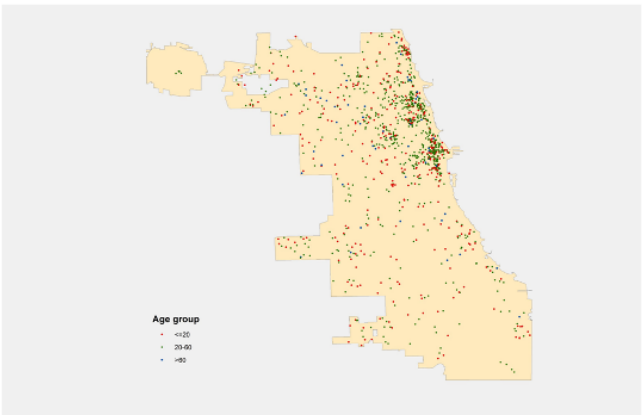
- Assessing bias and uncertainty of location-based social media data
  - how representative social media are?



(a)



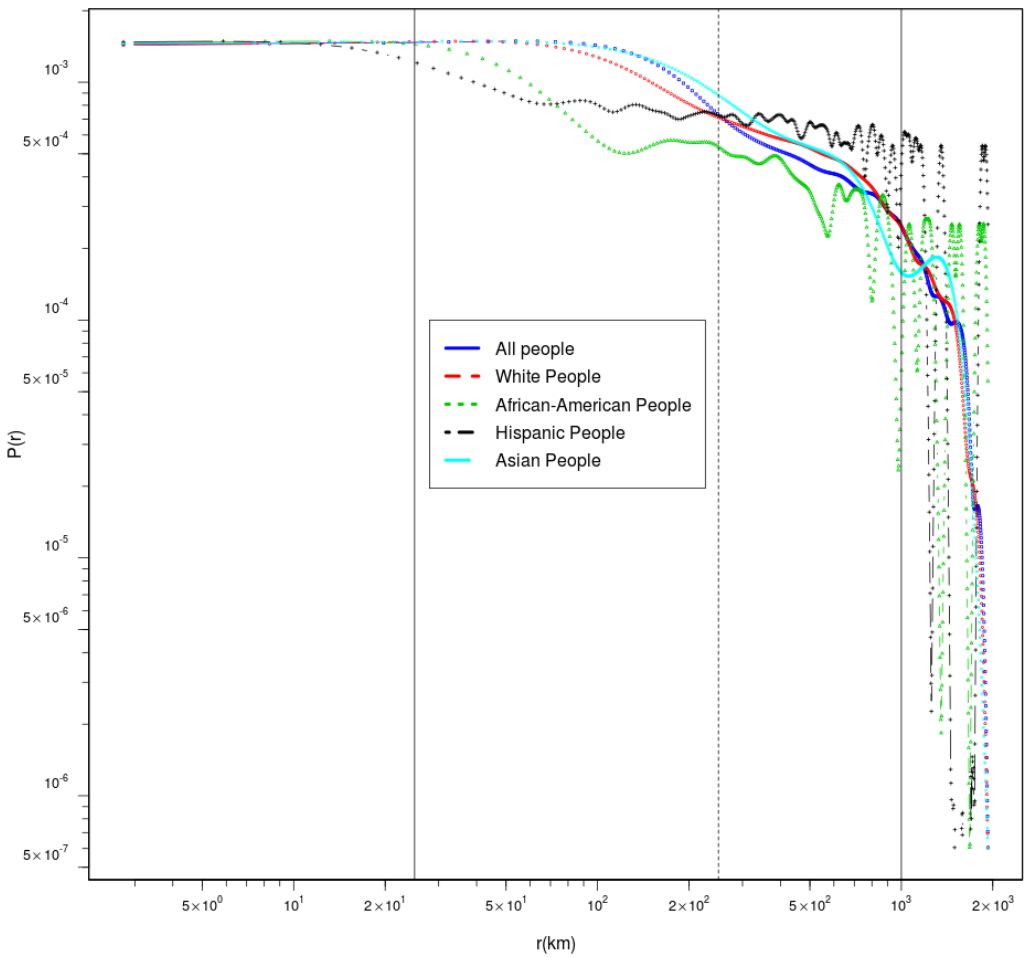
(b)



(c)

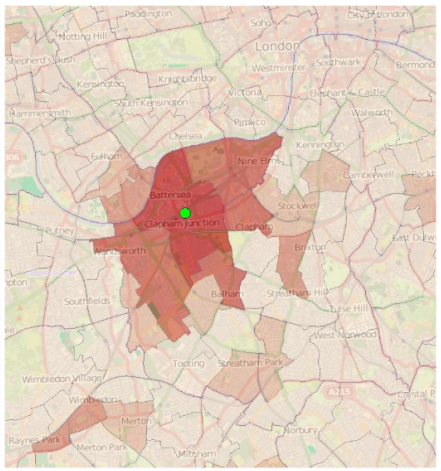
Ongoing related work IV

- Characteristics of human mobility, business locations and the catchment area



Ongoing related work V

- Characteristics of human mobility, business locations and the catchment area



Courtesy of Krzysztof Janowisz (UCSB)

Reading of this week

- Leetaru, K.,Wang, S., Cao, G., Padmananabhan, A., Shook, E. (2013): Mapping The Global Twitter HeartBeat: The Geography of Twitter. *First Monday*.
- Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*.

Thanks

Thank you, any questions?

Follow us on Twitter: @ttugis, @guofengcao