A Cloud Computing Workflow for Scalable Integration of Remote Sensing and Social Media Data in Urban Studies

Aiman Soliman[1], Kiumars Soltani[1], Junjun Yin[1], Balaji Subramaniam[2], Pierre Riteau[2], Kate Keahey[2], Yan Liu[1], Anand Padmanabhan[1], Shaowen Wang[1]

1 CyberGIS Center for Advanced Digital and Spatial Studies
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign

2 Mathematics and Computer Science Division,
Argonne National Laboratory
University of Chicago

December 18, 2015

CyberGIS@ncsa.illinois.edu

Outline

- **Fusing Social and Sensor Data**
- **Fusion Conceptual Framework**
- **Scalable Spatial Synthesis Capabilities**
- **Demo : Urban Flow**

Fusing Social and Sensor Data

Fusing Social and Sensor Data

Location Based Social Networks (LBSN) is defined as ambient geographic information provided through different social channels (Twitter, Flicker, etc.)

Recent studies suggested the potential of fusing LBSN with physical sensor data

- LBSN provides direct observations about **human presence** over urban landscapes (social sensing).
- LBSN could complement sparse sensor data (e.g., disaster management).

Challenges associated with LBSN data

- Data reliability
- Spatiotemporal properties of LBSN
- Lack of conceptual integration models
- Big data management and synthesis

Spatiotemporal Characteristics of LBSN

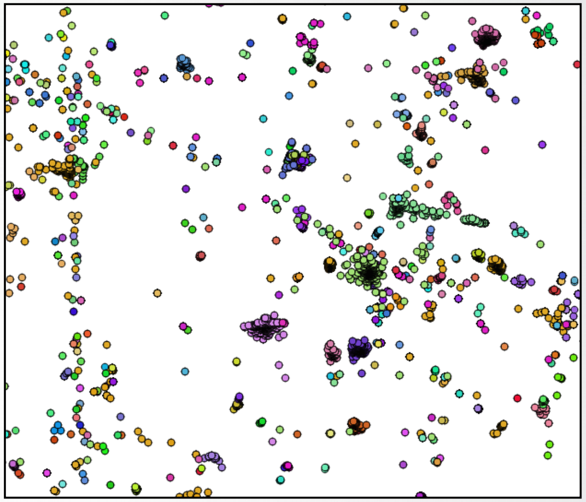
LBSN data has **peculiar spatiotemporal properties** that need to be considered when fused with sensor data

Spatial characteristics

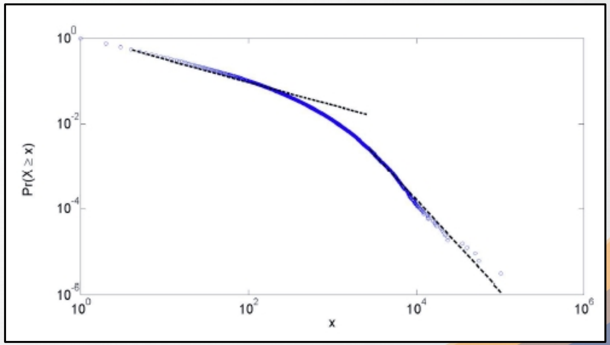
- Geographically sparse
- Distinct sub-populations (e.g., clusters, sporadic, linear)

Temporal characteristics

- Users' engagement follows a bursty behavior
- Observations exhibit a fat-tail distribution.
- Fewer users engage the most (frequent user bias)



Geo-located Twitter data spatial distribution in Detroit suburbs



Distribution of total number of tweets per user

Conceptual Framework

Urban Dynamics and Connectivity

Mapping **Urban Connectivity** is important in **urban planning**, **transportation** and design of **smart cities**.

Remote sensing has been used successfully to delineate urban cores (e.g., **landcover, landuse** new **urban development** and **sprawl)**

Revealing the connectivity network between urban units can not be completed using sensor data and usually obtained from low latency surveys

A cyberGIS based workflow was developed to study urban connectivity based on fusion of remotely sensed landuse maps and mobility patterns extracted from geo-located Twitter data

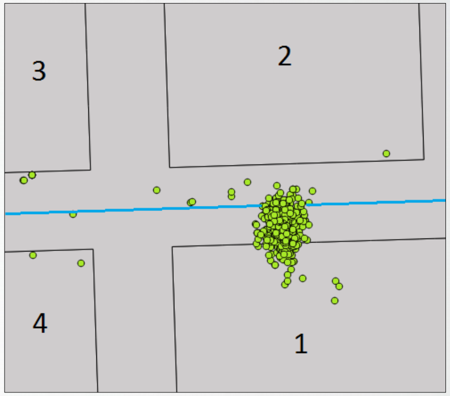
Conceptual Framework (I)

Each tweet contains information about the user, the geographic location, time stamp and 140 characters text.

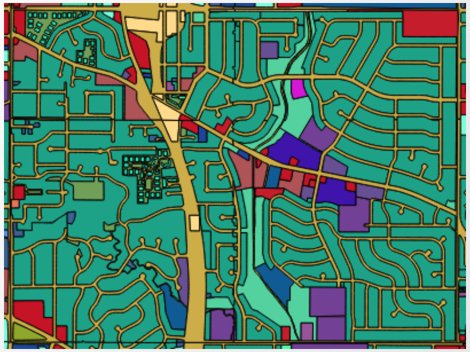
Individual twitter user trajectories are dominated by frequently visited locations

These clusters are explained by the high predictability of human movement (preferential return)

We examined the semantic composition of preferential return locations using remotely-sensed landuse maps



Example of frequently visited location

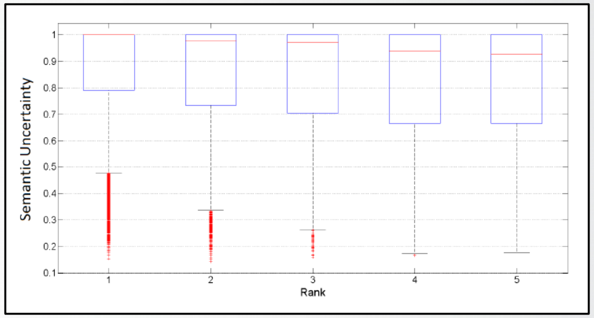


CMAP landuse map

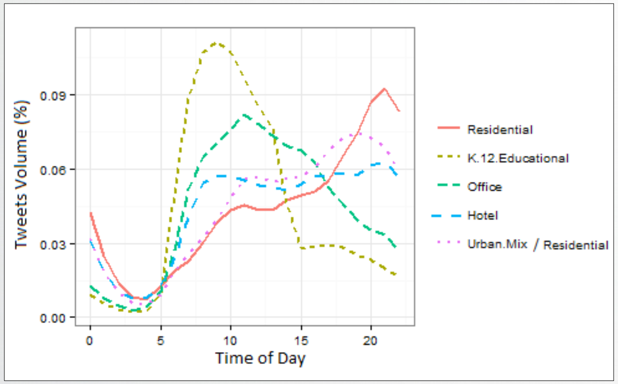
CMAP 2014. Chicago Metropolitan Agency for Planning's 2010 Land Use Inventory for Northeastern Illinois.

Conceptual Framework (II)

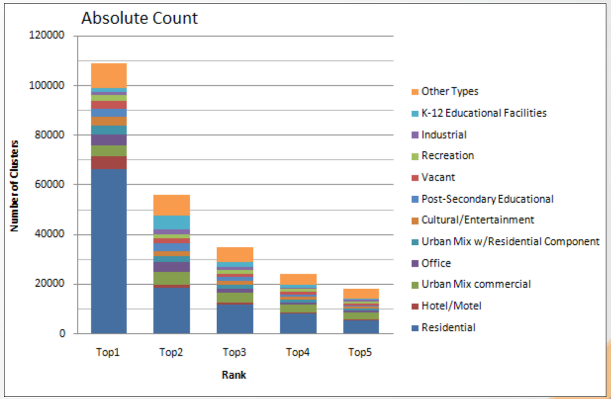
- Semantics of frequent visited locations are dominated by one land use type
- Clusters' semantics are partially dependent on visitation rank
- LBSN has distinct temporal signatures



Landuse purity of top visited locations



Volume of tweets per hour



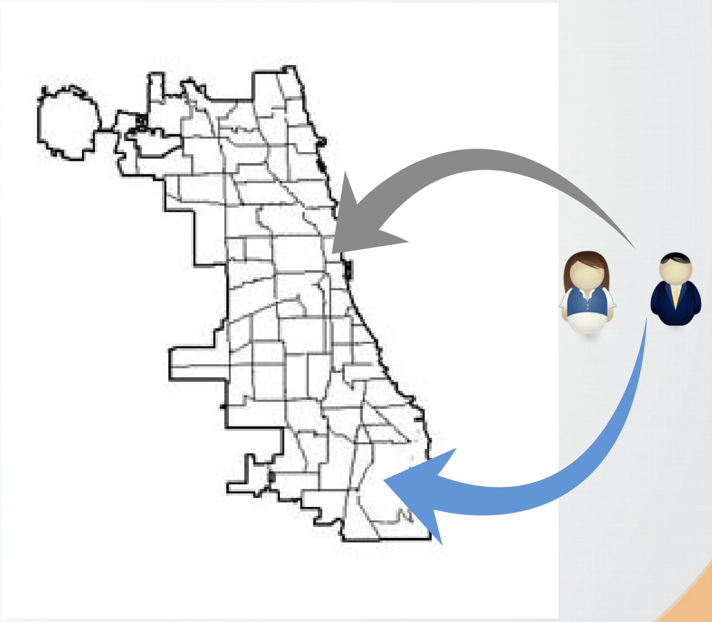
Dominant land use composition for top 5 visited locations

Soliman, A., Yin, J., Soltani, K., Padmanabhan, A., and Wang, S. 2015. "Where Chicagoans tweet the most: Semantic analysis of preferential return locations of Twitter users". Proceedings of the First ACM SIGSPATIAL International Workshop on Smart Cities and Urban Analytics (UrbanGIS'15)

Conceptual Framework (III)

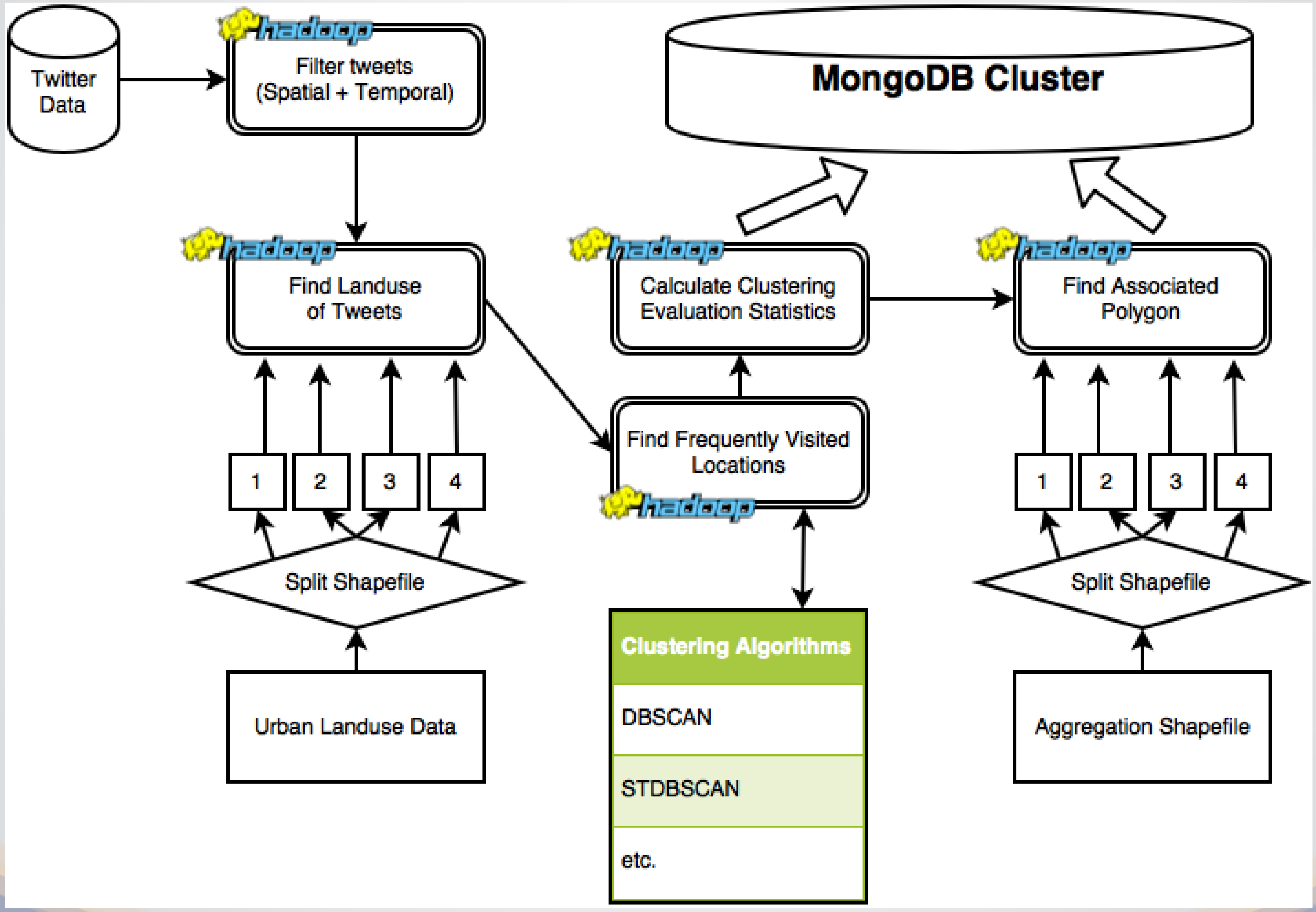
Our connectivity model is based on identifying shared frequent visitors between two polygons in terms of

- **Strength** : number of shared frequent visitors
- **Purpose** : semantic at origin and destination
- **Demographic** : dominant user language



Scalable Spatial Synthesis Capabilities

Architecture (Backend)



Distributed Point in Polygon

- Time-consuming to load the shapefile for each mapper (existing approach).
- Splits the large shapefile into smaller shapefiles.
  - Recursive bi-section method.
- Mapper decides which small shapefile the point lie into based on the geographical bounds.
  - Using R-tree.
- Reducer loads the small shapefile and finds the polygon the point lies into.
  - Using quad-tree.
- To find the closest polygon to a point, we split the original shapefile into overlapping shapefiles.

Database Cluster

- Multiple instances of MongoDB.
- Shared (partitioned).
- Directly connected to Hadoop to avoid storing data in HDFS first.
- Connected to NodeJS query service.
- Gateway application contacts the database through a NodeJS query service.

Urban Flow Demo

Acknowledgements

This research is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 1047916, 1443080, and 1354329.

Insightful comments were received from members of the CyberGIS Center for Advanced Digital and Spatial Studies.

Thank You