A High-order Statistical Approach to Understand Urban Structures

Rui Zhu

STKO Lab
Department of Geography
University of California, Santa Barbara

LocBigData 2019: July 15, Tokyo, Japan

Location-based Big Data

From Spatial to Platial

- Location-based big data allow us to model human perceptions of places
- The **order** of platial analysis increases:
  - Spatial, temporal, thematic and emotional perspectives
  - **High-order interactions**

Beyond Pairs

Due to the **significant heterogeneity** and **complex dependence** of places, **high-order spatial interactions** have to be considered in platial analysis.

- **Geo-dipole** (Goodchild et al., 2007)

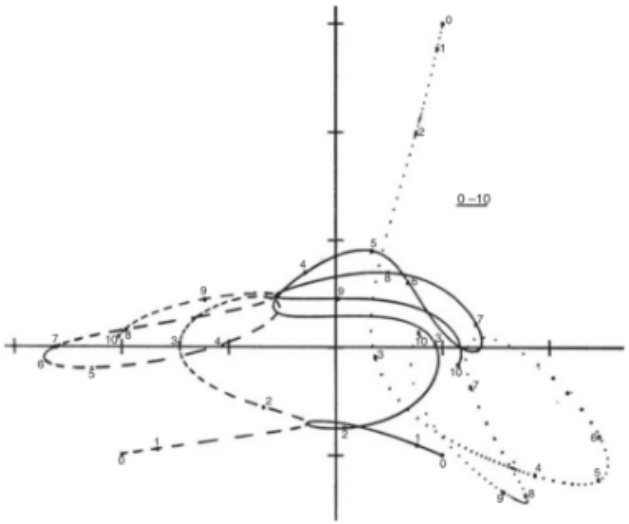
$$< x, x', Z, z(x, x') >$$

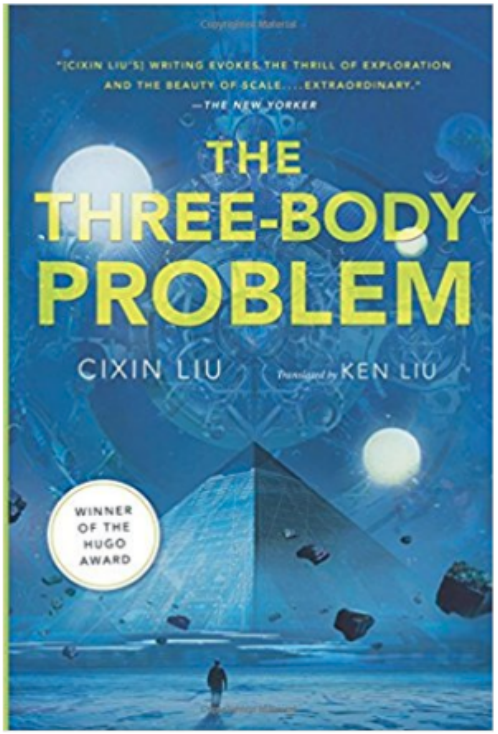
- **Geo-multipole** (Zhu et al., 2017):

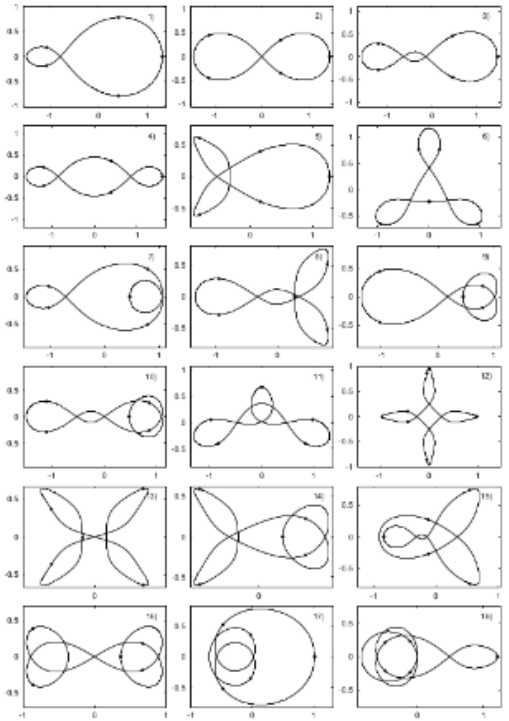
$$\langle x, t_N, Z, z(x, t_N) \rangle$$

where $t_N = \{x_1, .., x_N\}$ are the N neighbors of x

The Three-body Problem

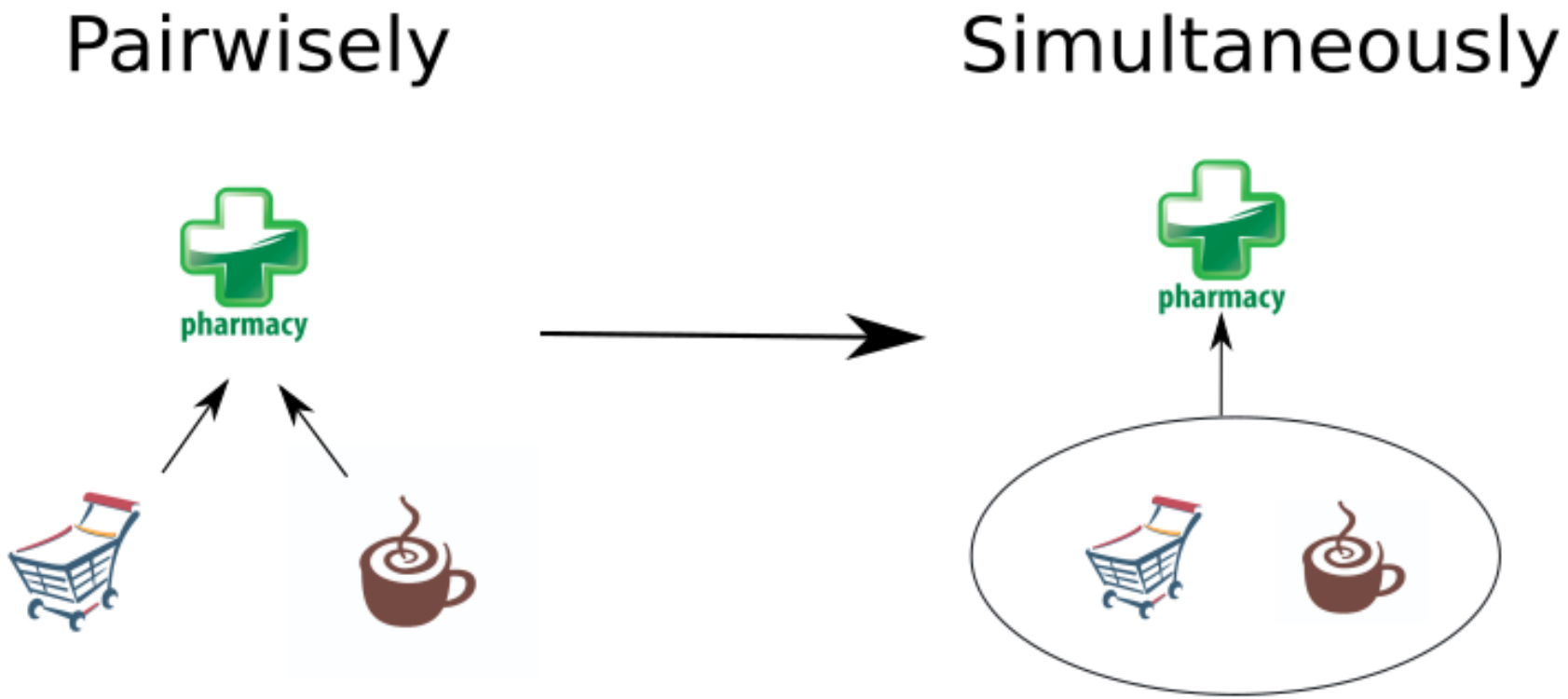






Simó, C. (2001)

The Three-body Problem in Geography



Pairwise Independence $\neq$ Global Independence

- Two independent random indicators: x_1 and x_2, each could be either 1 or 0 with probability p
- The third indicator $x_3 = x_1 x_2$
- x_3 is determined by both x_1 and x_2 but not by either of them individually
- Even worse, x_1 and x_2 are often not independent as well

High-order Spatial Interaction

- Therefore, the interaction between x_3 and (x_1, x_2) should be modeled through a **trivariate** statistics
- However, most traditional spatial statistics are bivariate

High-order Conditional Probability

- Derived from the **extended normal equation:**

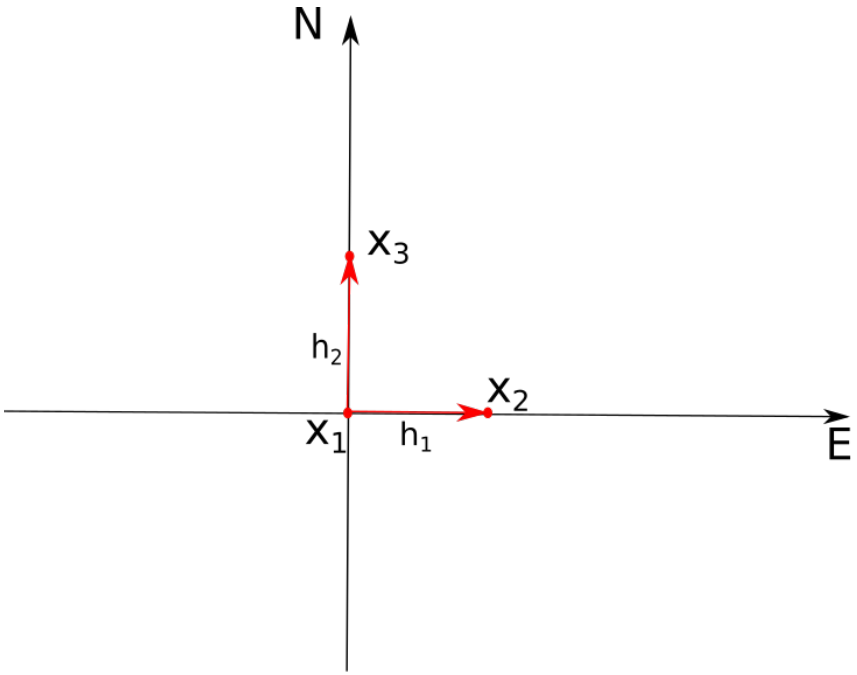
$$p(x_0 = 1|x_i, i = 1, \cdots, n) = \lambda_0 + \sum_{i_1=1}^{n} \lambda_{i_1}^{(1)} p(x_{i_1}) + \sum_{i_1=1}^{n} \sum_{i_2>i_1}^{n} \lambda_{i_1 i_2}^{(2)} p(x_{i_1} x_{i_2}) + \sum_{i_1=1}^{n} \sum_{i_2>i_1}^{n} \sum_{i_3>i_2}^{n} \lambda_{i_1 i_2 i_3}^{(3)} p(x_{i_1} x_{i_2} x_{i_3}) + \cdots + \lambda^{(n)} p(\prod_{i=1}^{n} x_i)$$

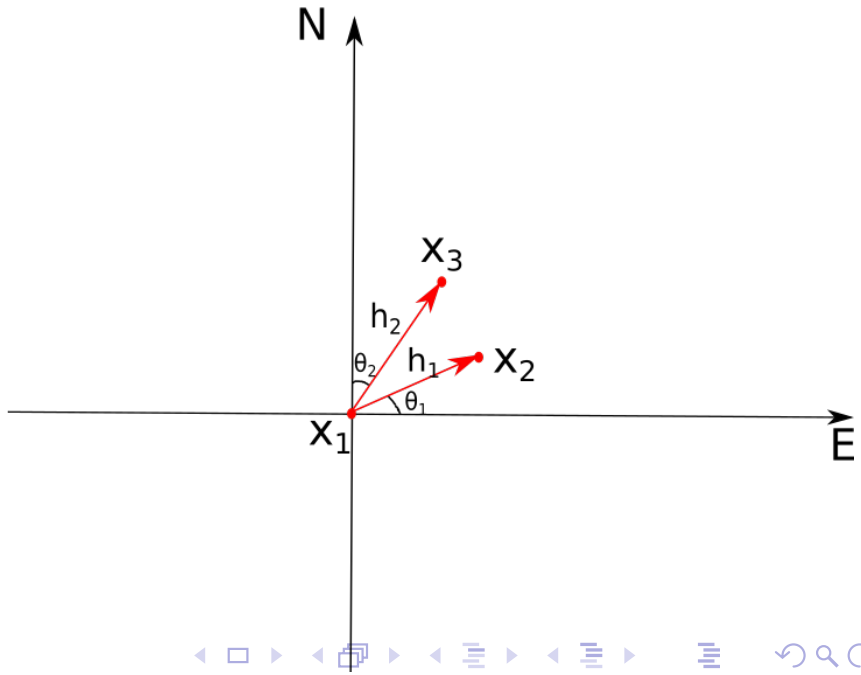
where the indicator x_i is defined as:

$$x_i = \begin{cases} 1, \text{ if location } i \text{ has feature type } t \\ 0, \text{ otherwise} \end{cases}$$

High-order Stationarity

- **Approximate** the equation to the **order of three**
- Second-order stationarity: based on the **displacement** $\overrightarrow{h}$
- High-order stationarity: based on the **shape** comprised of **more than two locations**
- Examples of third-order shape:



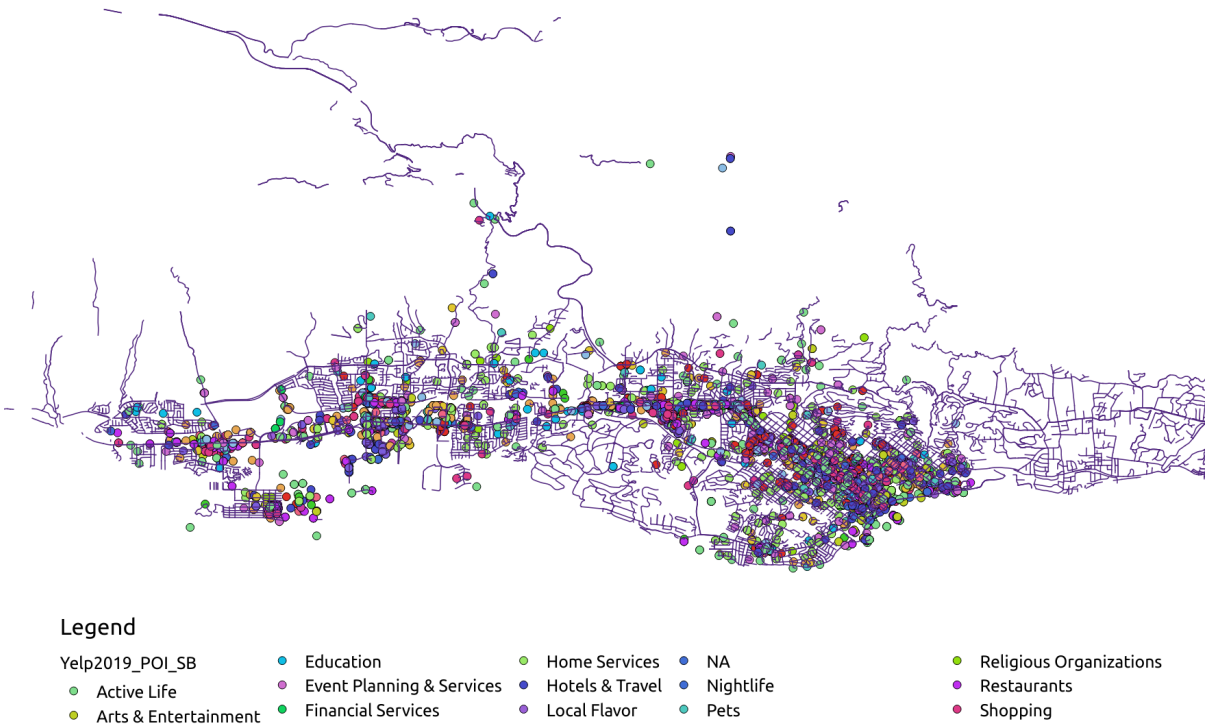


Workflow

1. Use the shape to scan the data
2. Find data events that fit the shape
3. The parameters (i.e., λ) are estimated using either least square estimation or Legendre polynomial approximation
4. The high order conditional probability is inferred using the extended normal equation

Experimental Data

- We focus on modeling interaction of **place types**
- **Points of Interests** (22 root place types) from Foursquare are collected in three cities:
  - Santa Barbara
  - Washington D.C.
  - Las Vegas



Hypothesis and Expected Result

- Places in urban cities are too complex to be modeled through the traditional second-order statistics, and by applying our proposed approach, rather complicated platial patterns could be extracted and characterized
- The proposed high-order spatial statistics are capable of addressing multiple challenges in location-based big data:
  - to compare spatial patterns between cities
  - to predict/suggest place types for a new data entry
  - to align typing schema for different data sources
  - to clean crowdsourced data