Leveraging Social Media Data in Agent-based Simulations

Presenter:
Hamdi Kavak, PhD Student, hkava001@odu.edu

Authors:
Jose J. Padilla, Saikou Y. Diallo, Hamdi Kavak,
Olcay Sahin, Brit Nicholson

Virginia Modeling, Analysis and Simulation Center
Old Dominion University

Agent-based Simulations

- Applied to various phenomena in
  - Social Sciences, Economics, Biology, …
- Typically, **agent properties** are derived from related theories.
- Usually, **property initialization** is based on random distributions.

Agent-based Simulations

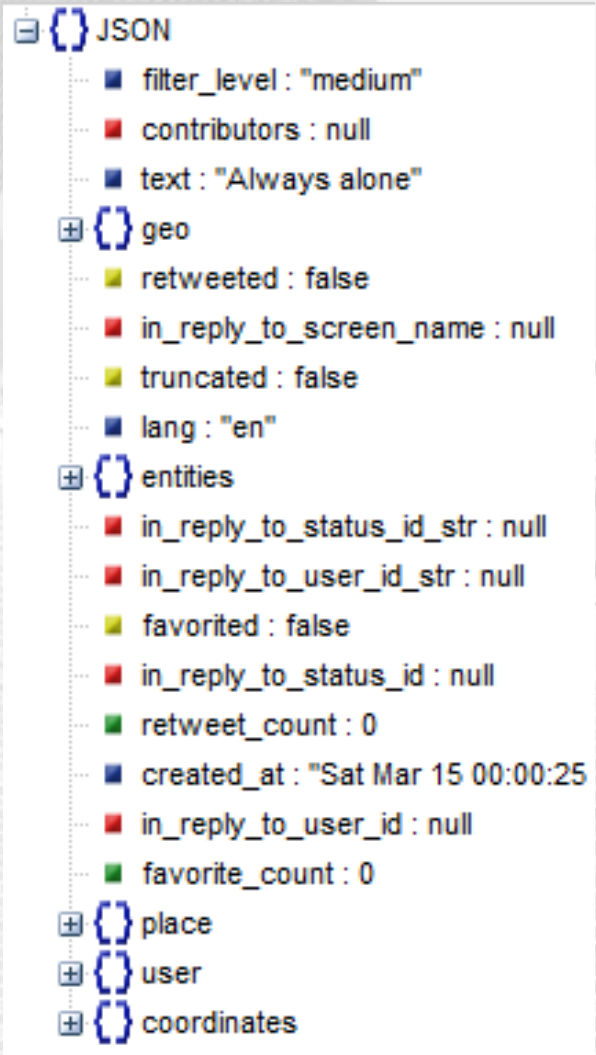
- Limitation:
  - Lack of data to make simulations more empirically grounded.
- Data of:
  - Individuals
  - Population
  - Individuals' interaction
  - Environment

Data Collection Alternatives

1. Surveys
   - Representativeness
   - Truthfulness
   - Quickly become outdated
   - Snapshot
2. Electronic devices
   - Sensors/badges *(Eagle and Pentland 2006)*
   - Cellphone trajectories *(Gonzalez, Hidalgo, and Barabasi 2008)*
   - Smartphone apps *(Funf Framework by Behavio)*
   - Longitudinal

Data Collection Alternatives

3. Social Media
   - Rich human behavior data
     - Preferences, Hobbies, Activities Opinions,…
   - Variable structure
   - Large
   - **Not** widely used in M&S studies
   - Repetitiveness
   - Near real-time & Longitudinal



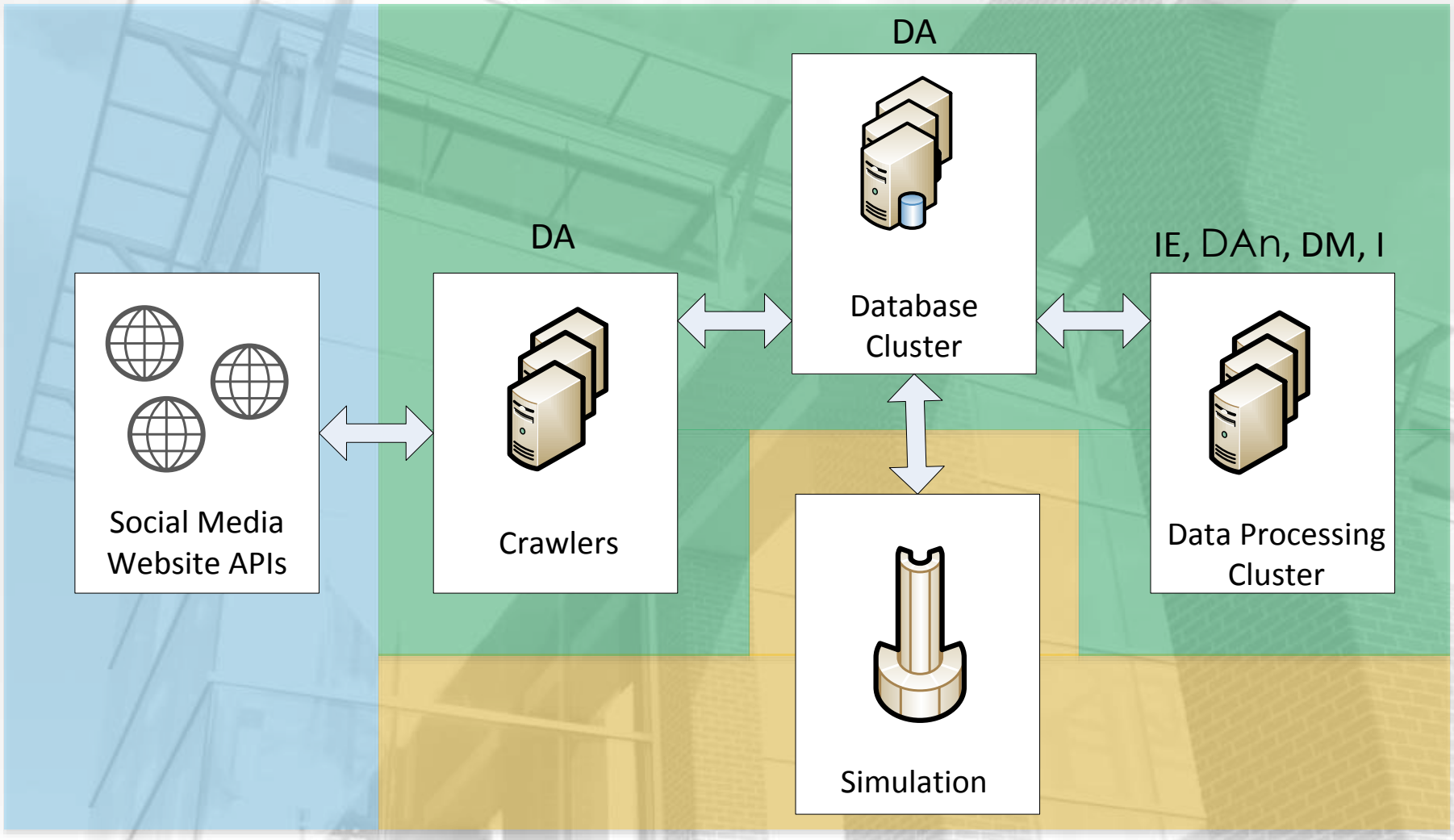
Sample Social Media Studies

- General
  - National suicide numbers prediction *(Won et al. 2013)*
  - Stock market change prediction *(Zhang et al. 2011)*
  - Prediction of the spread of diseases *(Sadilek et al. 2012)*
- Modeling and Simulation
  - As Simulation Input Data *(Yang, Liu, and Mo 2013; van Maanen and der Vecht 2013)*
  - As Validation/Calibration reference *(Yang, Liu, and Mo 2013; van Maanen and der Vecht 2013; Malleson and Birkin 2012)*

Research Approach

1. Data Source Identification
   - Single or multiple
2. Data Processing *(Labrinidis and Jagadish 2012)*
   - Data acquisition (DA)
   - Information extraction (IE)
   - Data analysis (DAn)
   - Data mining (DM)
   - Interpretation (I)
3. What-if Analysis

Tool Framework



Use Cases

1. **Population Preference Identification**
2. Individual Mobility Patterns
3. People's Communication Based on Common Interests

Use Case 1: Population Preference Identification

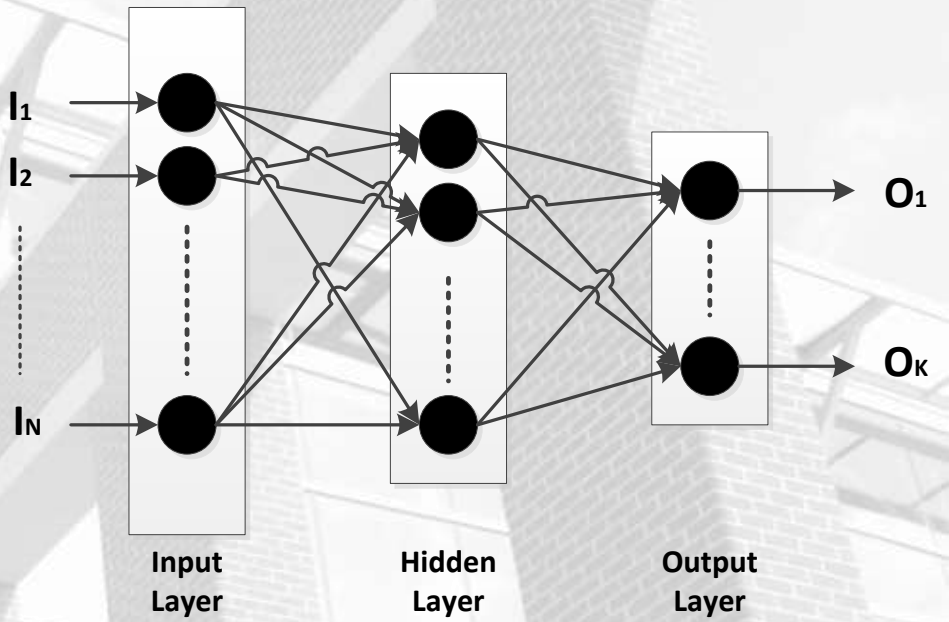
Factor	Percentage
Telecommunication	6.67
Chemical Hazardous Material	4.41
Healthcare and Public Health	5.94
Agriculture and Food	5.82
Government Facilities	5.34
Water	7.05
Information Technology	5.56
Nuclear	4.13
Transportation	5.74
National Monuments and Icons	7.29
Dams	4.84
Energy	5.40
Defense	5.62
Emergency Services	6.13
Postal and Shipping	5.45
Commercial Facilities	8.11
Banking and Finance	6.50

Use Cases

1. Population Preference Identification
2. **Individual Mobility Patterns**
3. People's Communication Based on Common Interests

Use Case 2: Individual Mobility Patterns

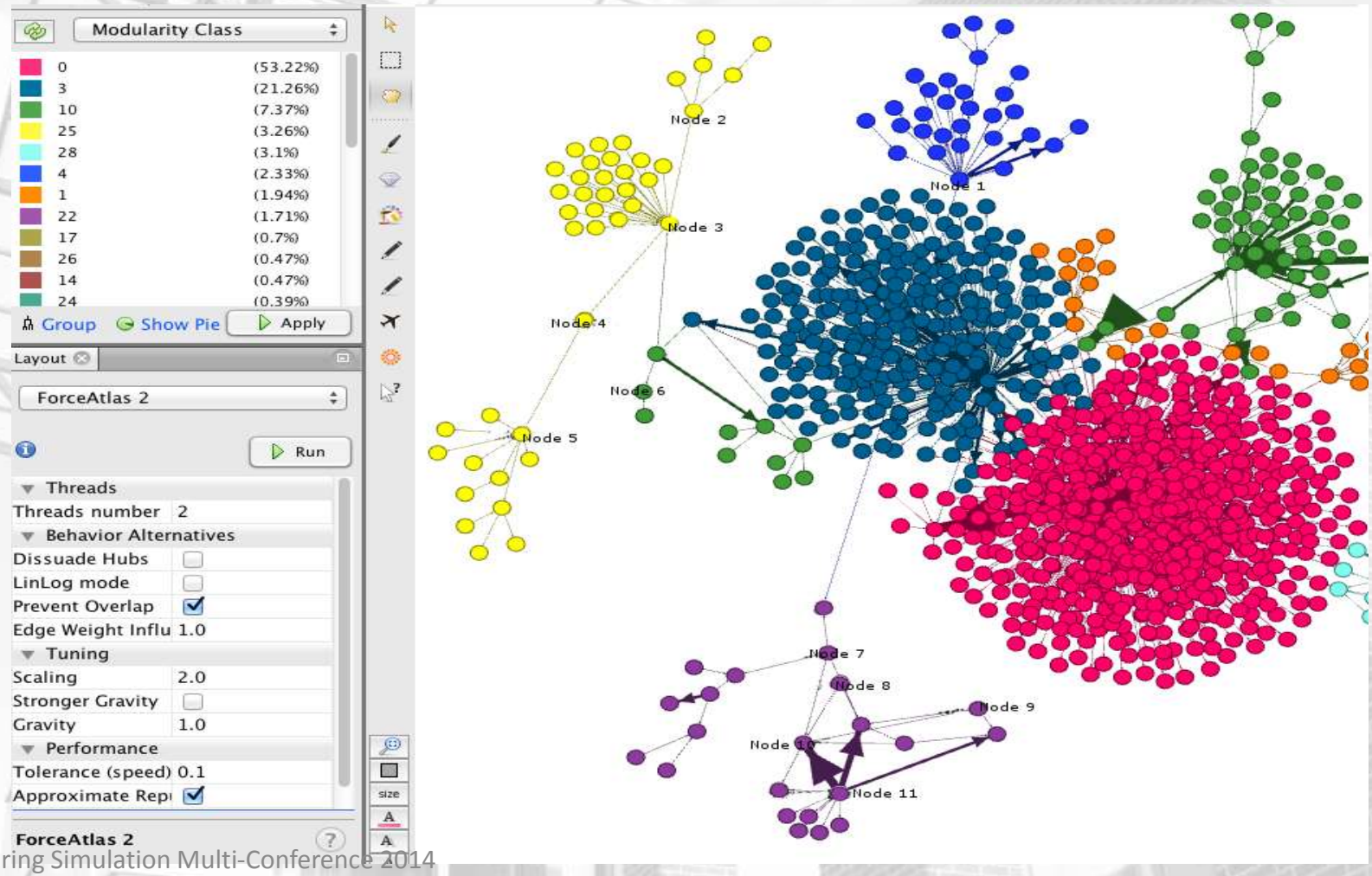
- Input:
  - Last location Id
  - Day of the week
  - Hour of the day
- Output:
  - Predicted location Id



Use Cases

1. Population Preference Identification
2. Individual Mobility Patterns
3. **People's Communication Based on Common Interests**

Use Case 3: People's Communication Based on Common Interests



Challenges

- Obtaining useful data
  - Introducing sample bias
  - Filtering out noise
- Data size & processing
  - 1% of tweets
  - Requires non-conventional data processing tools
- Designing simulations that make meaningful use of social media data

Thank you

Hamdi Kavak - hkava001@odu.edu

Details of the use cases

Use Case 1: Population Preference Identification

- Goal:
  Rank the preferences of individuals towards a predetermined set of factors.
- Method:
  - Determine keywords
  - Public Twitter data collection
  - Sentiment analysis
  - Initialize people's preferences based on sentiment ranks.

Use Case 2: Individual Mobility Patterns

- **Goal:**
  Modeling human mobility patterns using social media data
- **Method:**
  - Foursquare accounts were created and managed.
  - Individuals checked-in using Foursquare smartphone app.
  - A Foursquare developer application was created.

Use Case 2: Individual Mobility Patterns

- **Method(cont.):**
  - A history of user check-ins were obtained using Foursquare's API endpoints.
  - Two datasets were created using check-in history
    - **Place dataset:**

Place Id	Name	Location(lat, long)	Place category

    - **Check-in dataset**

Person Id	Place Id	Date	Time	Day of the week

Example 2: Human Dynamics Simulation (Cont.)

- Following conversions applied for NN training (per individual)

Raw Check-in History

Id	Place Id	Date	Time	Day
.				
.				
.				
13	5	3/1/2013	16:00	Fri.
14	2	3/1/2013	19:00	Fri.
15	1	3/1/2013	21:00	Fri.
16	3	3/2/2013	12:00	Sat.
.				
.				
.				

Interpolated Check-in History

Place Id	Time	Day
.		
.		
5	16:00	Fri.
5	17:00	Fri.
5	18:00	Fri.
2	19:00	Fri.
2	20:00	Fri.
1	21:00	Fri.
1	22:00	Fri.
.		
.		
1	11:00	Sat.
3	12:00	Sat.
.		
.		

Binary Codified Check-in History

Place Id	Time	Day
.		
.		
000101	01111	100
000101	10000	100
000101	10001	100
000010	10010	100
000010	10011	100
000001	10100	100
000001	10101	100
.		
.		
000001	01010	101
000011	01011	101
.		
.		

Use Case 3: People's Communication Based on Common Interests

- **Goal:**
  Simulate communication and information propagation among individuals
- **Method:**
  - Public tweets containing a specific keyword are obtained (June 05 – July 16 2013). Aprx: 1,3 million
  - A network of retweets is created based on imperative words used.
  - Sentiment analysis was conducted to verify network formation