Understanding Human Mobility from Twitter

Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice Abou Jaoude, Mark Cameron, David Newth

PLoS ONE 10(7): e0131469. doi:10.1371/journal.pone.0131469
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0131469
Understanding Human Mobility

• Significance
  – Urban planning
  – Transport
  – Disease spread

• Issues with current data sources
  – Census: coarse-grained
  – GPS/Call data records/Wifi: proprietary/private

• Geo-tagged tweets as a proxy
  – 500 Million users
  – 340 Million tweets/day
  – Up to 10m accuracy
Open Issues with Geo-Tagged Tweets as Mobility Proxy

• Potential sampling bias
  – Demographic
  – Geographic

• Content limit on tweets
  – 140 characters/tweet
  – Unknown effect of this limit on tweet locations

• Tweet location preference
  – Unclear how it can affect observed movement patterns
Overview of this work

• Determine how representative are Twitter-based mobility patterns of population and individual-level movement

• Analyse a large dataset with 7,811,004 tweets from 156,607 Twitter users

• Compare the mobility patterns observed through Twitter with the patterns observed through other technologies, such as call data records
Displacement Distribution

Displacement distribution, namely spatial dispersal kernel \( P(d) \), where \( d \) is the distance between a user’s two consecutive reported locations.

Previously observed distributions:
- Power-law (banknotes)
- Truncated power-law (mobile phones & travel surveys)
- Exponential and log-normal (GPS from cars/taxis)

\[
P(d) \sim q\lambda_1 e^{-\lambda_1(d-d_{\text{min}})} + (1-q)\beta\lambda_2 d^{\beta-1}e^{-\lambda_2(d^\beta-d_{\text{min}}^\beta)},
\]
What could be different here?

• May stem from multiplicative processes, i.e. the displacement $d$ is determined by the product of $k$ random variables.

• These random variables can be transportation cost, lifestyle aspects such as the preference on commute distance, or socio-economic status such as personal income.

• The number of these variables $k$, namely the number of levels in the multiplicative cascade, is indicated by the exponent $\beta$ in the above equation.

• When $k$ is small, $P(d)$ converges to a stretched-exponential asymptotically, and $k \rightarrow +\infty$ leads to the classic log-normal distribution. In particular, if these random variables are Gaussian distributed, we have $k \approx 2/\beta$, and the value of $k$ is around 3 or 4 for our data ($\beta \approx 0.55$).
Another fitting possibility

- 2 separate power laws
- Differences between short and long distance travel patterns

\[ P(d) \sim \begin{cases} 
  d^{-\gamma_1} & d_{\text{min}} \leq d < d_m \\
  d^{-\gamma_2} & d_m \leq d < d_c 
\end{cases} \]
Radius of Gyration

• Quantifies the spatial stretch of an individual trajectory or the traveling scale of an individual

\[ r_g = \sqrt{\frac{1}{n} \sum_i (\vec{r}_i - \vec{r}_c)^2} \]

where \( \vec{r}_i \) the individual’s i-th location,
\( \vec{r}_c = \frac{1}{n} \sum_i \vec{r}_i \) is the geometric center of the trajectory and \( n \) is the number of locations in the trajectory.
Rg Distribution

\[ P(r_g) \]

- **Data**: \( y = y_1 + y_2 \)
- \( y_1 \sim e^{-0.12x} \)
- \( y_2 \sim x^{-0.23}e^{-0.0015x^{0.77}} \)
- \( y_3 \sim x^{-1.11} \)

\[ y_1 \sim x^{-0.40} \]
\[ y_2 \sim x^{-1.60} \]
First passage time

• $F_{pt}(t)$, i.e. the probability of finding a user at the same location after a period of $t$

• Similar to CDR trends
• Periodic behavior with daily cycles
• Content limit does not appear to affect $F_{pt}$
Preferential return to visited locations

• The probability function $P(L)$ of finding an individual at his/her $L$-th most visited location.
• Sort visited locations and perform spatial clustering (250m)
• Generally follow Zipf law of preferential return
• People are 50% likely to tweet from most popular location, higher than other data sources
Predictability of Tweet Locations

• Study the randomness (entropy) and predictability of the sequence of tweeting locations for each user
• Bi-modal distribution for users with >20 locations suggests 2 types of users
Probabilistic Modeling of Human Movement

Rg: 1-10 km

Rg: 10-100 km

Rg: 100-500 km

Rg: 500-1000 km
Isotropy of Motion Patterns

- Isotropy ratio $\sigma = \frac{\delta_y}{\delta_x}$, where $\delta_y$ is the standard deviation of $P(x, y)$ along the y-axis and $\delta_x$ is the standard deviation of $P(x, y)$ along the x-axis, to characterise the orbit of each $r_g$ group.
- Second peak at $\sim$1000 km differentiates results from previous studies.
Preferential return decreases with larger orbits

- Likelihood to tweet from home location drops with increasing $r_g$
- Exponent $\alpha$ also decreases with increasing $r_g$
Twitter-based Mobility Patterns

Rg: 1-10 km

Rg: 10-100 km

Rg: 100-500 km

Rg: 500-1000 km
Discussion

• Three observed modes of mobility
  – Intrasisite
  – Metropolitan
  – Intercity

• Two apparent groups of tweeters
  – Highly predictable group where geo-tags are not highly useful for mobility prediction
  – Less predictable group where geo-tagged tweets can be representative of movement patterns
Discussion

• Long distance movers more diffusive in their movement than intermediate distance movers, most likely as a reflection of a switch in transportation mode towards air travel and local circulation around destination cities.

• Preferential return strongly dependent on a person’s orbit of movement, with long distance movers less likely to return to previously visited locations.

• Population-level mobility patterns are well-represented by geo-tagged tweets, while individual-level patterns are more sensitive to contextual factors
Implications and Future Work

• Develop agent-based model for disease spread based on rg group features
• Use tweets to better understand drivers for movement
• Use location inference algorithms on tweet content to increase data sample with higher uncertainty per location