

# Following Soccer Fans from Geotagged Tweets at FIFA World Cup 2014

Eugenio Cesario <sup>1</sup>, Chiara Congedo <sup>2</sup>, Fabrizio Marozzo <sup>3</sup>, Gianni Riotta <sup>4</sup>,  
Alessandra Spada <sup>2</sup>, Domenico Talia <sup>3</sup>, **Paolo Trunfio** <sup>3,\*</sup>, Carlo Turri <sup>2</sup>



<sup>1</sup> ICAR-CNR & DtoK Lab, Italy

<sup>2</sup> Alkemy Lab, Italy

<sup>3</sup> University of Calabria & DtoK Lab, Italy

<sup>4</sup> Princeton University, USA

\* [paolo.trunfio@unical.it](mailto:paolo.trunfio@unical.it)



ICSDM 2015  
IEEE International Conference on Spatial Data Mining  
and Geographical Knowledge Services  
July 8-10, 2015 – Fuzhou, P.R. China

## Motivations and goals (1/2)

- In the past, understanding people behavior in a large-scale event was extremely difficult to catch
- Today, using geo-localized services of social media, we can analyze the behavior of large groups of people attending popular events
- Example: geotagged tweets can be used to understand users' mobility behaviors that are useful in travel route discovery
- **Goal** of this work: monitoring the attendance of **Twitter** users during the **FIFA World Cup 2014** matches to discover the most frequent movements of fans

## Motivations and goal (2/2)

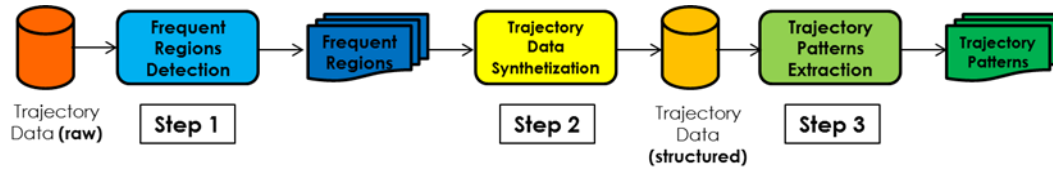
- **Data source:** more than half million geotagged tweets posted from inside the stadiums during the 64 matches of the World Cup from June 12 to July 13, 2014
- **Trajectory pattern mining** was carried out to identify the most frequent movement patterns of **Twitter** users attending the **World Cup matches**
- **Original results:**
  - **Number of matches attended** by fans
  - **Most frequent sequences of matches** attended by fans, either in the same stadium or to follow a given soccer team
  - **Most frequent movement patterns** obtained by grouping matches based on the phase in which they were played



# Outline

- Trajectory pattern mining
- Definitions
- Analysis process
  - *Data acquisition*
  - *Data pre-processing*
  - *Data mining*
  - *Results visualization*
- Results
  - *Number of Matches Attended*
  - *Frequent Sequences*
  - *Aggregate Analysis*
- *Conclusions*

# Trajectory pattern mining



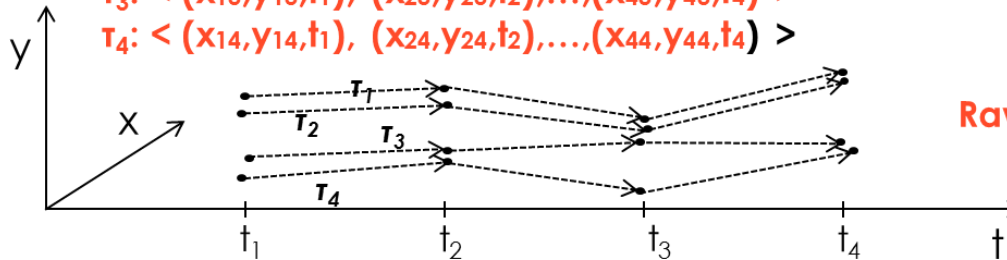
$$T_1: \langle (x_{11}, y_{11}, t_1), (x_{21}, y_{21}, t_2), \dots, (x_{41}, y_{41}, t_4) \rangle$$

$$T_2: \langle (x_{12}, y_{12}, t_1), (x_{22}, y_{22}, t_2), \dots, (x_{42}, y_{42}, t_4) \rangle$$

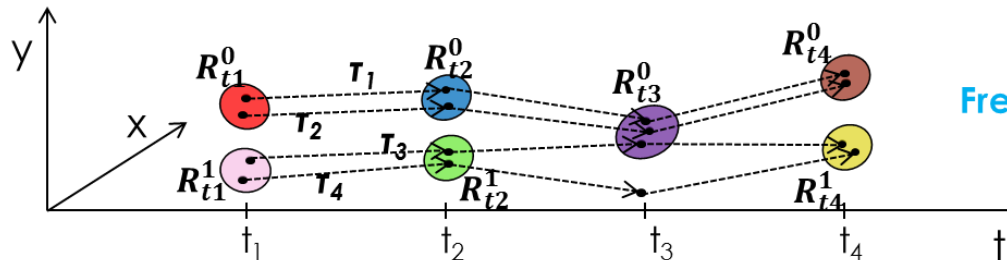
$$T_3: \langle (x_{13}, y_{13}, t_1), (x_{23}, y_{23}, t_2), \dots, (x_{43}, y_{43}, t_4) \rangle$$

$$T_4: \langle (x_{14}, y_{14}, t_1), (x_{24}, y_{24}, t_2), \dots, (x_{44}, y_{44}, t_4) \rangle$$

Raw Trajectory Data



Raw Trajectories



Frequent Region Extraction

$$T_1: \langle R^0_{t1}, R^0_{t2}, R^0_{t3}, R^0_{t4} \rangle$$

$$T_2: \langle R^0_{t1}, R^0_{t2}, R^0_{t3}, R^0_{t4} \rangle$$

$$T_3: \langle R^1_{t1}, R^1_{t2}, R^0_{t3}, R^0_{t4} \rangle$$

$$T_4: \langle R^1_{t1}, R^1_{t2}, R^1_{t4} \rangle$$

$$R^0_{t1} \wedge R^0_{t2} \wedge R^0_{t3} \rightarrow R^0_{t4}$$

$$R^0_{t1} \wedge R^0_{t2} \rightarrow R^0_{t3}$$

...

Structured Trajectory Data

Trajectory Patterns

# Definitions

- $S=\{s_1, \dots, s_{12}\}$ : set of **stadiums**, where for each stadium  $s_i$  are known the four corner coordinates of the rectangle containing it
- $TW=\{tw_1, \dots, tw_N\}$ : set of **geotagged tweets**, where each tweet  $tw_i$  is described by the following properties:
  - user who posted  $tw_i$
  - *latitude* and *longitude* (of the place from where  $tw_i$  was sent)
  - *source* (device or application used to generate  $tw_i$ )
  - *date* and *text*
- $M=\{m_1, \dots, m_{64}\}$ : the 64 **matches**, where each match  $m_i$  is described by the following properties:
  - *stadium*
  - *date*
  - $team_1$  and  $team_2$  (the two teams playing the match)

# Analysis process

- The analysis process is composed of four steps:
  - **Data acquisition**, collecting the geotagged Twitter data
  - **Data pre-processing**, cleaning, selection and transformation of data to make it suitable for analysis
  - **Data mining**, analyzing pre-processed data to infer trajectory patterns
  - **Results visualization**, making results readable and usable

# Data acquisition

- **Twitter REST APIs** used to collect all the geotagged tweets posted during the World Cup matches
- Only **tweets** whose coordinates fall **within the area of stadiums during the matches**



- About **526,000 tweets** collected from June 12 to July 13, 2014



# Data pre-processing

- A three-step task:
  1. **Cleaned** data by removing tweets with unreliable positions (e.g., tweets with coordinates manually set by users or applications)
  2. **Selected** only tweets written by users present at the matches, by removing *re-tweets* and *favorites* posted by other users
  3. **Transformed** data by keeping one tweet per user per match, as we were interested to know only if a user attended a match or not
- Final dataset **D** with about 10,000 transactions, each one containing the list of matches attended by a single user:

$$D = \{T_1, T_2, \dots, T_n\}$$

where  $T_i = \langle u_i, \{m_{i1}, m_{i2}, \dots, m_{ik}\} \rangle$  and  $m_{i1}, m_{i2}, \dots, m_{ik}$  are the matches attended by a Twitter user  $u_i$

## Data mining (1/2)

- Trajectory pattern mining to extract the most frequent movements of fans starting from  $D$
- **Trajectory pattern**: sequence of geographic regions that emerge as frequently visited in a given temporal order
- The **support** of a trajectory pattern  $p$  (# of transactions containing  $p$ ) is a measure of its reliability
- In our case, a **frequent pattern  $fp$**  with support  $s$ :

$$fp = \langle m_i, m_j, \dots, m_k \rangle (s)$$

is an ordered sequence of matches  $m_i, m_j, \dots, m_k$  where  $s$  is the percentage of transactions in  $D$  containing  $fp$

## Data mining (2/2)

- Pattern extraction algorithm:
  - Compute the support of each match in  $D$
  - *Iteratively*:
    - Generate new candidate  **$k$ -match-sets\*** and compute their support, using the frequent  **$(k-1)$ -match-sets** found in the previous iteration
    - Delete all the candidate match-sets whose support is lower than a given minimum support
  - Terminate when no more frequent match-sets are generated

\* **$k$ -match-set** = set of matches of cardinality  $k$

## Results visualization

- Creation of **Infographics** for presenting the mobility patterns
- Main design guidelines:
  - Visual representation of quantitative information
  - Minimising the efforts necessary to decoding symbols
- **Result:** a visualization model helping readers to easily catch the key meaning of extracted knowledge

# Results

- Three main categories:
  - **Number of matches attended** by fans during the competition
  - **Most frequent sequences of matches** attended by fans, either in the same stadium or to follow a given soccer team
  - **Most frequent movement patterns** obtained by grouping matches based on the phase in which they were played

## Results: Number of matches attended

No. of matches	Spectators
1	71.3%
2	16.0%
3	6.0%
4	3.0%
5 or more	3.7%

- 3.7% of the spectators attended **five or more** matches during the whole World Cup
- Twitter profiles of those who attended several matches, show that many of them were journalists

## Results: Frequent sequences (1/4)

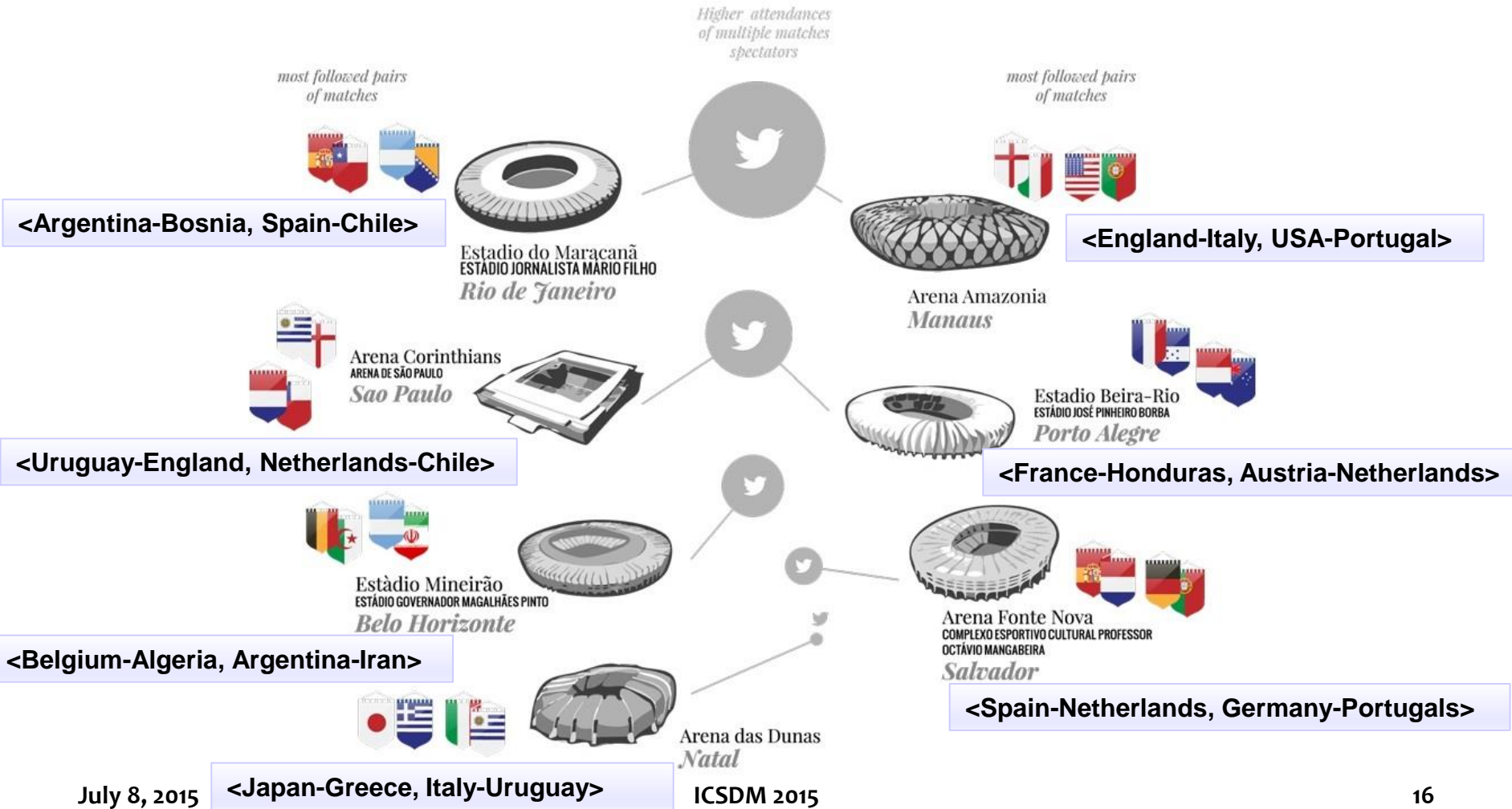
- General classification of the paths followed by fans who attended at least two matches:

No. of matches	Same stadium	Same team
2	62.9%	22.2%
3	48.8%	11.8%
4	41.0%	7.2%
5	37.0%	8.4%
6	33.7%	4.8%

- Results show that **most of who attended multiple matches did it staying in the same city**

# Results: Frequent sequences (2/4)

- Most frequent 2-match-sets observed during the group stage, from June 12 to June 26, 2014

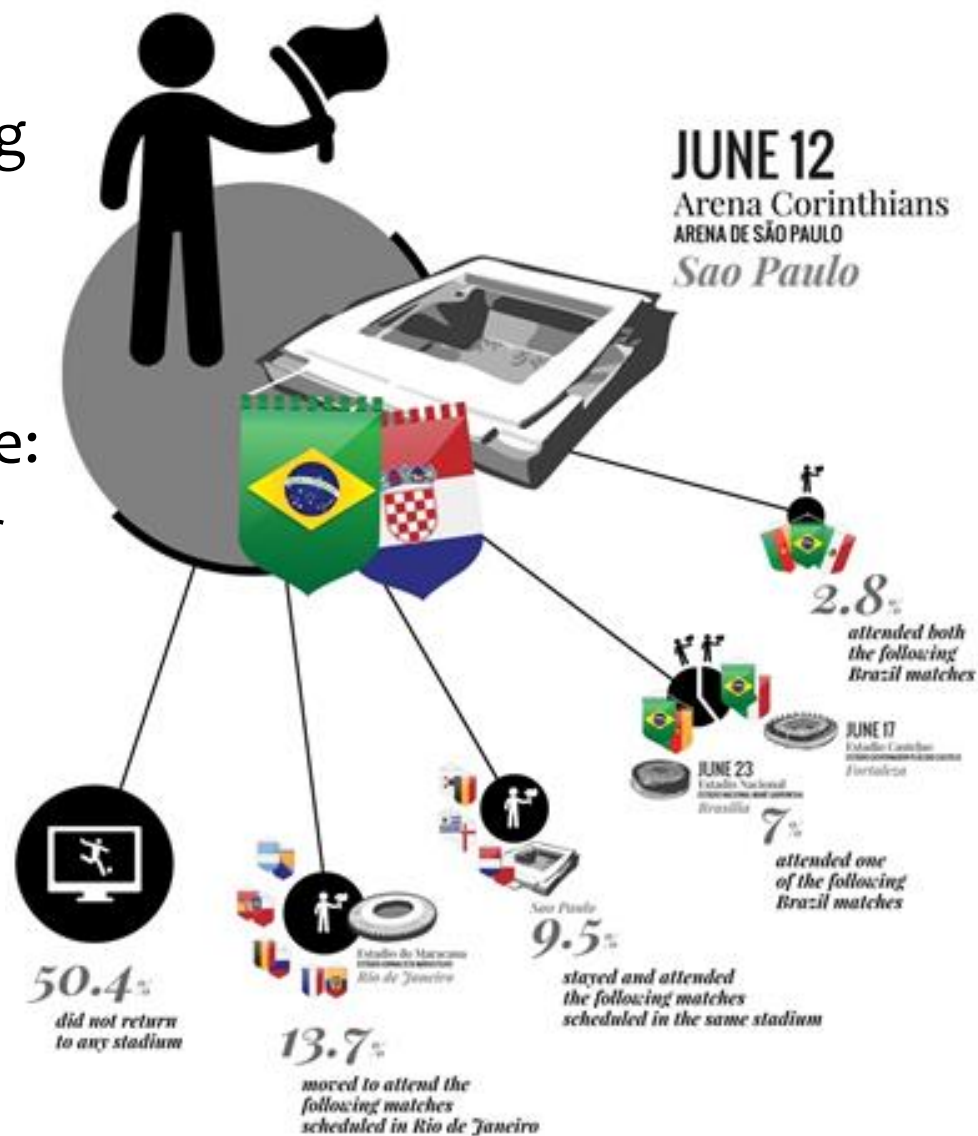






## Results: Frequent sequences (4/4)

- Specific analysis on the spectators of the opening match <Brazil-Croatia> played on in São Paulo
- At the end of group stage:
  - 50.4% did not attend other matches
  - 13.7% moved to Rio de Janeiro to attend other matches
  - 9.5% attended other matches in the same stadium

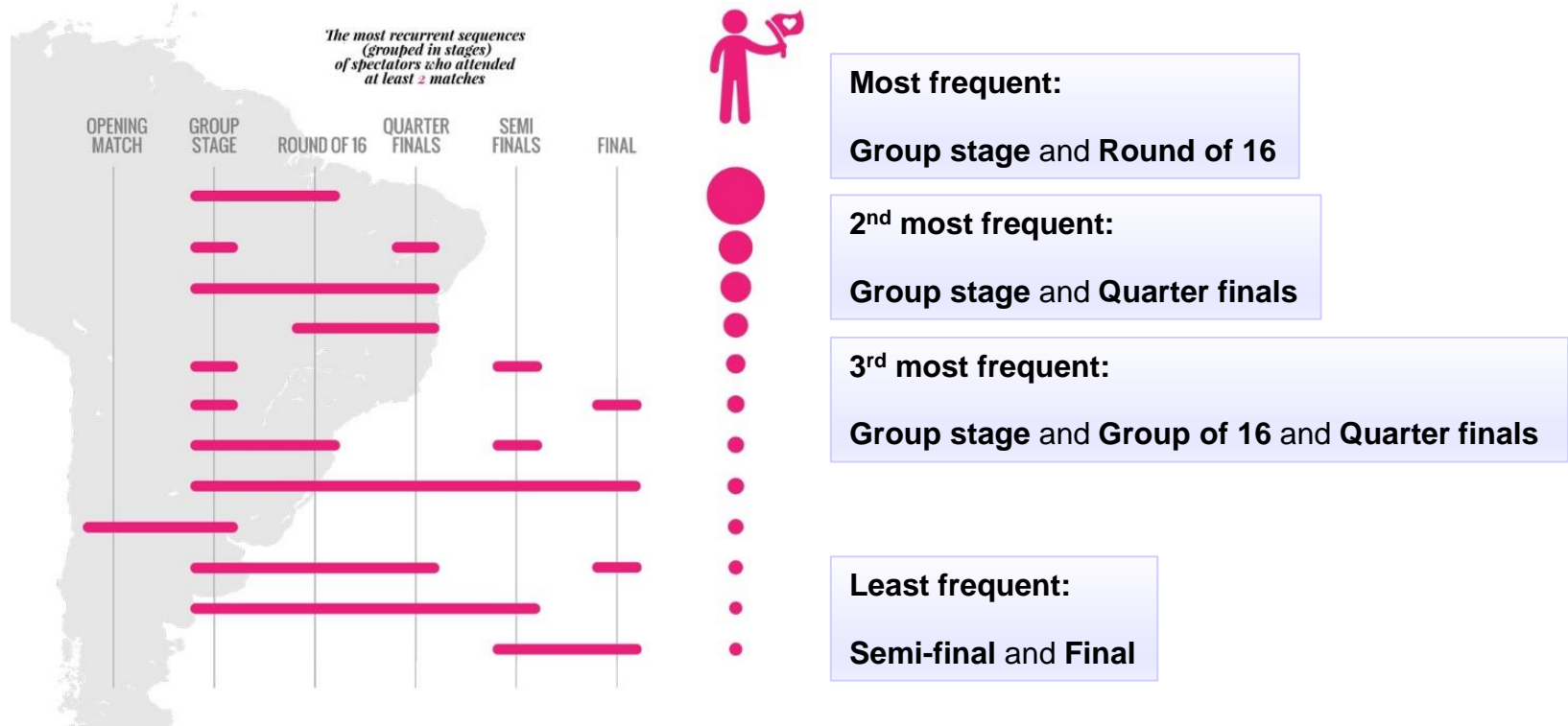


## Results: Aggregate analysis (1/2)

- *Goal*: studying the movements of fans during the different phases of the competition
- Matches were grouped into the following phases:
  - *Opening match* (match no. 1)
  - *Group stage* (matches no. 2-48)
  - *Round of 16* (matches no. 49-56)
  - *Quarter finals* (matches no. 57-60)
  - *Semi-finals* (matches no. 61-62)
  - *Final* (match no. 64)

## Results: Aggregate analysis (2/2)

- Patterns of movements based on the grouping above, and the relative frequency (support) of these patterns



- The relative frequency of each pattern is represented by a circle: the larger the circle, the higher the frequency

## Conclusions

- Analysis of fans' movements during the FIFA World Cup 2014: An example of how social data analysis can be used to know how people behave in big events
- Social data applications can help the organization of future events, e.g. monitoring and management of key services like transports, security, logistics, and others
- This methodology can be re-used in similar scenarios to understand collective behaviours that are very hard to discover with traditional social analysis techniques

# Questions?



*Thank you!*

