Towards a Digital Time Machine fueled by Big Data & Social Mining

Dino Pedreschi
KDD LAB
Università di Pisa & ISTI-CNR
http://kdd.isti.cnr.it

AI*IA Special Track – Pisa 12 dicembre 2014
SIAMO TUTTI
POLLICINI DIGITALI

La Vita Nova, e-magazine de Il Sole 24 Ore
Fosca Giannotti, Dino Pedreschi
Big data “proxies” of social life

Shopping patterns & lifestyle

Relationships & social ties

Desires, opinions, sentiments

 Movements
ERCIM White Paper on Big Data Analytics
A European way to the fair and open use of Big Data

- A problem-driven path towards fair use of Big data
  - Creation of an ecosystem for Big data Analytics-as-a-service
  - Based on Federated, Open (and trusted) Platforms for Knowledge Accelleration (FOAPKA)
Data&Knowledge&People infrastructure

ETHICAL VALUES: privacy, trust, transparency

EDUCATION: data literacy – data scientists
The long term vision

- Empower citizens, communities, business and institutions with a **Digital Time Machine** to:
  - Explore the past and present to gain better understanding and self-knowledge
  - Explore plausible future to reason on the consequences of decision making
PI’s

- ISTI-CNR
  - Fosca GIANNOTTI
  - Raffele PEREGO
  - Fabrizio SEBASTIANI

- IIT-CNR
  - Andrea PASSARELLA
  - Maurizio TESCONI

- UNIPI
  - Paolo FERRAGINA
  - Dino PEDRESCHI

- SNS
  - Fabrizio LILLO

- IMT-LU
  - Guido CALDARELLI

www.sobigdata.eu
Focus on country-wide CDR data
CITY USERS’ SOCIOMETER

Partner: Telecomunicazioni, PA, ISTAT
Mobile phone socio-meters

Analyze individual call habits to recognize profiles of city users

– Resident
– Commuters
– Visitors/Tourists
Dimmi come chiami ... ti dirò chi sei!

Residente

Pendolare

Visitatore
Call Habit Profiles

Week: working days & weekend

Time slots
0:00-7:59
8:00-18:59
19:00-23:59

Users' call habit profile
User profile quantification

- Resident profile
- Commuter profile
- Visitor profile

Classification outcome
- Residents: 26%
- Commuters: 20%
- Visitors: 9%
- Unclassified: 45%
City User Sociometer

**Chiamate**

<table>
<thead>
<tr>
<th>userID</th>
<th>CellStart ID</th>
<th>CellEnd ID</th>
<th>TimestampStart</th>
<th>TimestampEnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1092594</td>
<td>PD02U</td>
<td>PD02U</td>
<td>2012-01-19 18:46:33+01</td>
<td>2012-01-19 18:46:44+01</td>
</tr>
<tr>
<td>1092594</td>
<td>PD02U</td>
<td>PD02U</td>
<td>2012-01-19 18:53:10+01</td>
<td>2012-01-19 18:53:33+01</td>
</tr>
<tr>
<td>1092639</td>
<td>PD34U</td>
<td>PD34U</td>
<td>2012-01-18 13:59:20+01</td>
<td>2012-01-18 14:02:33+01</td>
</tr>
</tbody>
</table>

**Anonimizzazione**

**Profilo aggregato**

**Analisi di data mining**

**Mappa dei profili**

**Pendolari**

**Visitatori/Turisti**

**Residenti**
Urban Sociometer: Pisa

Classification outcome

- Residents: 26%
- Commuters: 20%
- Visitors: 45%
- Unclassified: 9%

Pisa, January 2012

Urban Sociometer: Cosenza

Quantification of the Categories
- Cosenza -

- Visitors: 35%
- Commuters: 22%
- Residents: 25%
- Unknown: 13%
- In transit: 5%
ISTAT – Commissione Big Data

- Measuring urban population and inter-city mobility using big data in an integrated approach
  - Barbara Furletti, Lorenzo Gabrielli, Giuseppe Garofalo, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio

- SIS 2014 – 47° Simposio Italiano di Statistica
Users Classification (Sociometer)

Users' Call Profiles

Resident

Dynamic Resident

Commuter

Visitor

Residents

Dynamic Residents

Commuters

Visitors

Classification Algorithm
Validation

- Compare estimation with GSM data and administrative data:
  - Is the model a good proxy for estimating city users’ presence on each municipality?
  - GSM data rescaled using a penetration coefficient provided by the telecom operator
Commuters (incoming)

\[ y = -2009.25 + 1.48x \quad r = 0.977 \]
Dynamic residents (outgoing)

\[ y = 998.72 + 0.87x \quad r = 0.830 \]
Flow B → A is given by the users that are resident in B and work in A
OD Matrix
(Fluxes to Pisa)

\[ y = -22.84 + 1.95x \quad r = .989 \]
Main issues

- Linear vs. superlinear models to approximate the number of city users and fluxes
- How size of the municipalities influence the accuracy of the model?
- Alternative classification models for city users
Results

- Semi-automatic methodology for estimating presence and fluxes
- High experimental accuracy with ground-truth administrative data
- First steps towards exploiting big data in official statistics
- Basis for nowcasting and short-term prediction models
MP4-A Project: Mobility Planning For Africa

A joint work of

kdd.isti.cnr.it  www.goudappel.nl

Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Peter Van Der Mede, Joost De Bruijn, Erik de Romph, Gerard Bruil: Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast.
CitiSens 2013: Citizen in Sensor Networks – 2° Int. Workshop, pp. 15-25
The approach

- Analyze raw GSM data to
  - infer systematic mobility of individuals
- Build origin-destination matrices
  - Describe (expected) flows between areas
- Build a transportation model
  - Assigns O/D matrix to OSM road network through OmniTRANS system
Systematic mobility

- A single trace of an individual can be poorly informative about his/her movements
Systematic mobility

- Yet, several daily traces of the same individual might allow to identify regular places
Systematic mobility

- Yet, several daily traces of the same individual might allow to identify regular places
Systematic mobility

- Yet, several daily traces of the same individual might allow to identify regular places and trips
Systematic mobility

- The whole individual mobility is then summarized by its systematic movements

- They will be used as typical daily schedule of the individual
Systematic O/D matrix

- Combine the ten 2-weeks datasets into one
- For each user, extract significant L1 $\rightarrow$ L2
- Aggregate (individual) systematic movements into (collective) systematic flows
- Examples:

  ![Outgoing traffic](image1)
  ![Incoming traffic](image2)
Figure 12: Mobile phone movements in Ivory Coast and Abidjan.

Building the transport model

- Traffic assignment
  - Based on OmniTRANS V6 software

- Simulation assumptions
  - Assign each phone tower to the closest road
  - Use OSM information on speed limits
  - Adopt an all-or-nothing assignment
BIG DATA ANALYTICS & ECONOMIC DEVELOPMENT

Partner: Octo, WIND, Orange Telecom
Economic Development

social diversity

depprivation
What about mobility?
Economic Measures
Economic Measures

7,000 French cities

- Income
- Deprivation Index
- Education level
20 million users
200 million calls

user filtering

6 million users
mobility trajectories
social network
Four individual measures

- Radius of gyration
- Social degree
- Mobility entropy
- Social diversity

volume

diversity
mobility

\[ \rho = -0.43 \]

sociality

\[ \rho = -0.2 \]

\[ \rho = 0.07 \]
Entropy stable with age/gender

Song et al. 2010
mobility

Sociality

![Scatter plots showing correlations between European Deprivation Index and other variables.](image-url)
mobility

sociality
...also in Tuscany

HCR
Head Count Ratio

corr = -0.74
Discussion

1. **Mobility** diversity is linked to wellbeing
2. **Entropy** is stable across age/gender but varies with wellbeing
3. **Geography** matters
4. Big Data may become a pillar for real-time official statistics and nowcasting
Human Mobility, Social Networks and Economic Development
L. Pappalardo, M. Vanhoof, D. Pedreschi, Z. Smoreda, F. Giannotti

Small area model-based estimators using Big Data sources

To learn more: join Luca Pappalardo’s PhD Defense soon after this talk at 17:00 in Aula Seminari Ovest
Big Data Analytics & Social Mining

The Social Microscope
a tool to measure, understand, and possibly predict human behavior
EUROPEAN WAY TO BIG DATA

PRIVACY-BY-DESIGN

NEW DEAL ON DATA
Privacy-by-design in big data analytics and social mining

Anna Monreale\textsuperscript{1,2*}, Salvatore Rinzivillo\textsuperscript{2}, Francesca Pratesi\textsuperscript{1,2}, Fosca Giannotti\textsuperscript{2} and Dino Pedreschi\textsuperscript{1}
City User Sociometer with Privacy-by-design

Chiamate

<table>
<thead>
<tr>
<th>userID</th>
<th>CellStar ID</th>
<th>CellEnd ID</th>
<th>TimestampStart</th>
<th>TimestampEnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1092594</td>
<td>P102U</td>
<td>P102U</td>
<td>2012-01-19</td>
<td>2012-01-19</td>
</tr>
<tr>
<td>1092594</td>
<td>P102U</td>
<td>P102U</td>
<td>2012-01-19</td>
<td>2012-01-19</td>
</tr>
<tr>
<td>1092594</td>
<td>P102U</td>
<td>P102U</td>
<td>2012-01-19</td>
<td>2012-01-19</td>
</tr>
<tr>
<td>1092639</td>
<td>P302U</td>
<td>P302U</td>
<td>2012-01-18</td>
<td>2012-01-18</td>
</tr>
<tr>
<td>1092639</td>
<td>P302U</td>
<td>P302U</td>
<td>2012-01-18</td>
<td>2012-01-18</td>
</tr>
<tr>
<td>1092639</td>
<td>P302U</td>
<td>P302U</td>
<td>2012-01-18</td>
<td>2012-01-18</td>
</tr>
</tbody>
</table>

Anonimizzazione

Profilo aggregato

Analisi di data mining

Mappa dei profili

(a)

(b)

(c)

(d)

Pendolari

Visitatori/Turisti

Residenti
A NEW DEAL ON PERSONAL DATA

DEMOCRATIZING BIG DATA
A user-centric ecosystem for Big Data

- **Engage people** in the creation and use of big data and knowledge, by empowering individuals with self-knowledge
- Incentivize individual to **participate**, to align own self-interest with broader societal goals
- Based on **transparency, trust & privacy**
Conoscenza individuale e conoscenza collettiva

Collective Knowledge

Collective patterns

Personal data store for self knowledge

Self-awareness

Social mining of many individuals

Social mining of individual histories

Novel Indicators

Individual profiles

Traditional DB sources
Companies who want to access data about individuals can request it through data agents.

Several data stores are now up and running allowing individuals to exercise control over how data about them is used.

Several governments are working with the private sector to give individuals access to a copy of data about them in a usable format which can then be stored in their locker and shared with other providers.
Toscana Mobile Territorial Lab

- HII project Trusted Cloud of EIT-ICT LAB
- A living lab of volunteer users who will create a Personal Data Store ecosystem
- Focus on mobility, telecom services and supermarket transactions
- Joint project of
Key publications

- F Giannotti, M Nanni, F Pinelli, D Pedreschi. Trajectory pattern mining. ACM SIGKDD 2007
- F Giannotti, D Pedreschi. Mobility, data mining and privacy: Geographic knowledge discovery. Springer, 2008
- D Wang, D Pedreschi, C Song, F Giannotti, AL Barabasi. Human mobility, social ties, and link prediction. ACM SIGKDD 2011
- F Giannotti, M Nanni, D Pedreschi, F Pinelli, C Renso, S Rinzivillo, R Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB 20(5) 2011
- R Trasarti, F Pinelli, M Nanni, F Giannotti. Mining mobility user profiles for car pooling. ACM SIGKDD 2011
Key publications

- S Rinzivillo, S Mainardi, F Pezzoni, M Coscia, D Pedreschi, F Giannotti. Discovering the geographical borders of human mobility. Künstliche Intelligenz 26 (3) 2012
- D Pennacchioli, M Coscia, S Rinzivillo, D Pedreschi, F Giannotti. Explaining the Product Range Effect in Purchase Data. IEEE BIGDATA 2013
Vision papers


This is the work of many people for a long time

- Fosca Giannotti, co-lead of KDD LAB
- Salvo Rinzivillo, Mirco Nanni, Roberto Trasarti, Salvatore Ruggieri, Chiara Renso, Anna Monreale, Franco Turini
- all the fantastic folks at KDD LAB Pisa
- many international collaborators