Lepri et al. *EPJ Data Science* (2015) 4:16
DOI 10.1140/epjds/s13688-015-0050-4

EPJ Data Science
a SpringerOpen Journal

EDITORIAL Open Access

Making big data work: smart, sustainable, and safe cities

Bruno Lepri[1*], Fabrizio Antonelli[2], Fabio Pianesi[1] and Alex Pentland[3]

*Correspondence: lepri@fbk.eu
[1]Foundation Bruno Kessler, via Sommarive, 18, Trento, 24105, Italy
Full list of author information is available at the end of the article

Abstract

The goal of the present thematic series is to showcase some of the most relevant contributions submitted to the 'Telecom Italia Big Data Challenge 2014' and to provide a discussion venue about recent advances in the appplication of mobile phone and social media data to the study of individual and collective behaviors. Particular attention is devoted to data-driven studies aimed at understanding city dynamics. These studies include: modeling individual and collective traffic patterns and automatically identifying areas with traffic congestion, creating high-resolution population estimates for Milan inhabitants, clustering urban dynamics of migrants and visitors traveling to a city for business or tourism, and investigating the relationship between urban communication and urban happiness.

Keywords: mobile phone data; social media data; human behavior; city dynamics

1 Introduction

We live in a world of data. Nowadays, there are 6.8 billion of mobile phone subscribers worldwide, with millions of new subscribers every day [1]. More importantly, the almost universal adoption of mobile phones and the exponential increase in the use of social media and other Internet services is generating an enormous amount of data about human behaviors with a breadth and depth that was previously inconceivable. As recently reviewed by Blondel et al. [2], the Call Detail Records (CDRs), needed by the mobile phone operators for billing purposes, can be exploited to extract mobility patterns [3–5], to model social interactions [6, 7], city's structures [8], and epidemic spreading [9, 10], to estimate population densities [11], and to predict socio-economic indicators and outcomes of territories [12, 13]. Similarly, the emergence of social media (e.g. Twitter, Foursquare, Facebook) provides further opportunities to researchers to study different aspects of human behavior such as people's mobility [14] and social well-being of individuals and communities [15].

In this context, research challenges that provide access to a large number of research teams to the same dataset are becoming a valuable framework to advance the state of the art in the field and to sustain the process of reproducibility needed by the scientific community. An example is the Orange's 'Data for Development' (D4D) initiative [16, 17]. Last year, Telecom Italia with support from MIT Media Lab, Northeastern University, Fondazione Bruno Kessler, Polytechnic University of Milan, University of Trento, EIT ICT Labs, Trento Rise, and Spazio Dati organized the 'Telecom Italia Big Data Challenge' [18],

providing a multi-source geo-referenced and anonymized dataset composed by telecommunications, weather, news, Twitter and electricity data from two Italian areas: the city of Milan and the Trentino province [19].

More than 650 teams from more than 100 universities have participated to the 'Telecom Italia Big Data Challenge'. The projects ranged from predicting energy consumption to exploring the impact on mobility of some specific events and comparing mobile phone calling patterns with economic, demographic, and well-being indicators.

The goal of the present thematic series is to showcase some of the most outstanding contributions submitted to the 'Telecom Italia Big Data Challenge 2014' and to provide a discussion venue about recent advances in the application of CDRs and social media data to the study of individual and collective behaviors, with a particular attention devoted to the city dynamics.

2 Contributions

The first contribution, by De Domenico et al. [20], investigates route assignments in smart multimodal systems [21, 22], where individual daily trips follow recommendations based on personal and community constraints. The proposed approach is of special interest for designing efficient cities, where inhabitants could be automatically routed in order to reduce traffic and pollution. A person might want to avoid routes with high traffic or areas with high criminality, or to favorite routes across shopping and touristic areas. However, the individual choices of certain routes, without accounting for the state of the whole urban system, may lead to traffic congestion, increasing pollution, etc. [23]. In their paper, the authors proposed to model the trips in an urban system as interacting particles with data-driven origin-destination pairs. The route choices of the interacting particles are based on a time-varying potential energy landscape that seeks to simultaneously satisfy individual's (e.g. avoiding specific areas of the city) and community's (e.g. traffic and pollution reduction in specific city areas) constraints. Specifically, the proposed framework integrates multiple layers of constraints to favor certain routes and to study the effects of the proposed recommendations. The obtained results showed that the synergy among the individual choices plays a fundamental role in designing an efficient and smart city: only when all the individuals move according to the recommended routes, the city traffic is closer to the most ideal mobility scenario. Interestingly, the proposed method allows to monitor the traffic state of the city in real time, automatically identifying areas that are experiencing a congestion and hence supporting urban authorities and policy makers in planning interventions.

The second paper, contributed by Douglass et al. [24], used telecommunications activity data to create high-resolution population estimates. The traditional local census estimates are expensive, contingent on participation, and often suffer from several logistical issues. As shown by [11], telecommunications data are a promising new source of real-time estimates of population. In their paper, Douglass et al. have shown that the correlation between call volume and population in a given area of Milan is scale invariant above a certain population size. Then, the authors by means of a Random Forest regression [25] provided a reliable estimate of population for populous areas. The obtained results suggest that the method could be extended also to estimate population in less dense areas and to create estimates by gender, age, and ethnicity. Finally, the authors evaluated models for predicting the percentage of foreign population.

In the third paper, Bajardi et al. [26] studied urban spaces through the analysis of mobile phone records of users with strong international links, e.g. migrants and visitors travelling to a city for tourism or for business. More precisely, the authors focused on mobile phone records collected in Milan and used an entropy function to measure the level of country codes' hetereogenity in the calling patterns of a city's neighborhood. Then, they proposed a topological classification based on persistent *homology* and clustered the nationalities associated to the calls' sources and destinations outside Italy into two main groups. The first group comprises low-income countries, whose topological spatial patterns show a strong cyclic spatial distribution. The second group is formed by high-income countries, whose spatial distribution is scattered in small areas over the city. These results indicate that migrant communities from low income countries tend to aggregate in cohesive spatial structures and to live in the city's residential areas, mainly around the city centre; while communities associated with higher income countries tend to represent movement patterns of tourists and/or highly specialized professionals in central and high-entropy urban areas. As pointed out by the authors, the findings are in line with the ones predicted by the *spatial assimilation theory* [27] and confirm the empirical observation that different socio-economic migrant conditions can show distinct spatial clustering patterns [28]. Moreover, the authors demonstrated how mobile phone data can provide very specific spatial and temporal trajectories of visitors from a given country during a mass gathering event (e.g. large sport events).

The fourth and last contribution, by Alshamsi et al. [29] focuses on the relationship between urban communication and urban happiness. Specifically, the authors analyzed geo-located tweets within Milan to produce a detailed spatial map of urban sentiments. Then, they used communication intensity data to build the directional network of urban areas where the weights of the edges represent the communication strength between the areas. Their results found that there is no correlation between the happiness level of urban areas and the amount of communication the areas receive or initiate. Instead, happy urban areas tend to interact with other happy areas more than they interact with unhappy areas and, similarly, unhappy areas tend to interact with other unhappy areas more than they interact with happy areas. Interestingly, the urban happiness homophily supports previous findings on individual happiness homophily [30]. The obtained results may be relevant to guide policy makers in setting strategies that increase urban happiness.

3 Conclusion

The fourth papers in this series are excellent demonstrations of how mobile phone and social media data can contribute to many discoveries on daily life of individuals, communities and cities.

Telecom Italia is currently running a second edition of the Challenge [31]. This year, the data are released on 7 Italian cities: Bari, Milan, Naples, Rome, Turin, Venice and Palermo. Datasets include CDRs, demographic data from Telecom Italia (e.g. gender, age-range and living area), Twitter data, energy consumption data, private mobility data (trips performed by customers of some car security and insurance companies), and detailed Italian companies' information (e.g employees, size and locations). Hence, there are good reasons to continue with a second edition of this thematic series as follow up of the Big Data Challenge 2015.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Author details
[1]Foundation Bruno Kessler, via Sommarive, 18, Trento, 24105, Italy. [2]Skil-Telecom Italia, via Sommarive 18, Trento, 24105, Italy. [3]MIT Media Lab, 20 Ames Street, Cambridge, MA 02139, USA.

Received: 13 September 2015 Accepted: 21 September 2015 Published online: 16 October 2015

References

1. The world in 2014: ICT facts and figures. International Telecommunication Union. http://www.itu.int/
2. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4:10
3. Gonzalez M, Hidalgo C, Barabasi A (2008) Understanding individual human mobility patterns. Nature 435(7196):779-82
4. Song C, Qu Z, Blumm N, Barabasi A (2010) Limits of predictability in human mobility. Science 327(5968):1018-21
5. Kung K, Greco K, Sobolevsky S, Ratti C (2014) Exploring universal patterns in human home-work commuting from mobile phone data. PLoS ONE 9(6):e96180
6. Miritello G, Rubén L, Cebrian M, Moro E (2013) Limited communication capacity unveils strategies for human interaction. Sci Rep 3:1950
7. Schläpfer M, Bettencourt LMA, Grauwin S, Raschke M, Claxton R, Smoreda Z, West GB, Ratti C (2014) The scaling of human interactions with city size. J R Soc Interface 11(98):20130789
8. Louail T, Lenormand M, Cantú O, Picornell M, Herranz R, Frias-Martinez E (2014) From mobile phone data to the spatial structure of cities. Sci Rep 4:5276
9. Wesolowski A, Eagle N, Tatem A, Smith D, Noor A, Snow R, Buckee C (2012) Quantifying the impact of human mobility on malaria. Science 338(6104):267-70
10. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider C, Blondel V, Smoreda Z, González M, Colizza V (2014) On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol 10(7):e1003716
11. Deville P, Linard C, Martin S, Gilbert M, Stevens F, Gaughan A (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888-93
12. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029-31
13. Bogomolov A, Lepri B, Staiano J, Oliver N, Pianesi F, Pentland A (2014) Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th international conference on multimodal interaction (ICMI). ACM, New York, pp 427-34
14. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260-71
15. Quercia D, Ellis J, Capra L, Crowcroft J (2012) Tracking 'gross community happiness' from tweets. In: Proceedings of the ACM 2012 conference on computer supported cooperative work. ACM, New York, pp 965-8
16. Blondel VD, Esch M, Chan C, Clerot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C (2012) Data for development: the D4D challenge on mobile phone data. arXiv:1210.0137
17. de Montjoye Y, Smoreda Z, Trinquart R, Ziemlicki C, Blondel V (2014) D4D-Senegal: the second mobile phone data for development challenge. arXiv:1407.4885
18. Telecom Italia big data challenge 2014. http://www.telecomitalia.com/tit/en/bigdatachallenge/contest.html
19. Barlacchi G, De Nadai M, Larcher R, Casella A, Chitic C, Torrisi G, Antonelli F, Vespignani A, Pentland A, Lepri B (2015) A multi-source dataset of urban life in the city of Milan and Trentino province (in press)
20. De Domenico M, Lima A, González M, Arenas A (2015) Personalized routing for multitudes in smart cities. EPJ Data Sci 4:1
21. De Domenico M, Solé-Ribalta A, Gomez S, Arenas A (2013) Navigability of interconnected networks under random failures. Proc Natl Acad Sci USA 111(23):8351-6
22. Gallotti R, Barthelemy M (2014) Anatomy and efficiency of urban multimodal mobility. Sci Rep 4:6911
23. Wang P, Hunter T, Bayen AM, Schechtner K, Gonzaález M (2012) Understanding road usage patterns in urban areas. Sci Rep 2:1001
24. Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2015) High resolution population estimates from telecommunications data. EPJ Data Sci 4:4
25. Breiman L (2001) Random forests. Mach Learn 45(1):5-32
26. Bajardi P, Delfino M, Panisson A, Petri G, Tizzoni M (2015) Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Sci 4:3
27. Massey DS (1985) Ethnic residential segregation: a theoretical synthesis and empirical review. Sociol Soc Res 69(3):315-50
28. Pamuk A (2004) Geography of immigrant clusters in global cities: a case study of San Francisco. Int J Urban Reg Res 28(2):287-307
29. Alshamsi A, Awad E, Almehrezi M, Babushkin V, Chang P-J, Shoroye Z, Toth A-P, Rahwan I (2015) Misery loves company: happines and communication in cities. EPJ Data Sci 4:7
30. Bollen J, Goncalves B, Ruan G, Mao H (2011) Happiness is assortative in online social networks. Artif Life 17(3):237-51
31. TIM big data challenge 2015. http://www.telecomitalia.com/tit/en/innovazione/big-data-challenge-2015.html