




Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia

Lorenzo Lucchini^{1,2*} , Sara Tonelli¹ and Bruno Lepri¹

*Correspondence: llucchini@fbk.eu

¹Fondazione Bruno Kessler, Trento, Italy

²Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Abstract

The steady growth of digitized historical information is continuously stimulating new different approaches to the fields of Digital Humanities and Computational Social Science. In this work we use Natural Language Processing techniques to retrieve large amounts of historical information from Wikipedia. In particular, the pages of a set of historically notable individuals are processed to catch the locations and the date of people's movements. This information is then structured in a geographical network of mobility patterns.

We analyze the mobility of historically notable individuals from different perspectives to better understand the role of migrations and international collaborations in the context of innovation and cultural development. In this work, we first present some general characteristics of the dataset from a social and geographical perspective. Then, we build a spatial network of cities, and we model and quantify the tendency to explore of a set of people that can be considered as historically and culturally notable. In this framework, we show that by using a multilevel radiation model for human mobility, we are able to catch important features of migration's behavior. Results show that the choice of the target migration place for historically and culturally relevant people is limited to a small number of locations and that it depends on the discipline a notable is interested in and on the number of opportunities she/he can find there.

Keywords: Human behavior understanding; Natural language processing; Computational social science; Network theory; Human mobility

1 Introduction

Ever since the first villages were built by primitive people, humankind has moved from one community to another, in search of better life conditions or new opportunities [1]. These conditions can be represented by different factors, such as the opportunity of a job, better living standards, or the distance from the home country [2, 3]. The set of all these factors is difficult to define a-priori.

For example, according to the dominant neo-classical theory, people tend to make choices in order to maximize their income or level of well-being [4, 5]. Thus, the search for

© The Author(s) 2019. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

better economic conditions is one of the most important factors in the decision of moving from one location to a more attractive one. However, there are other factors that could play an important role in the decision-making process of a specific group of people. The attractiveness of an opportunity can also depend on cultural and linguistic barriers, and on the presence of particular communities at the destination [4, 6, 7]. Hence, in order to define a complete migration framework it is important to consider all the linked aspects that intervene in modifying the attractiveness of a location (e.g. economical, environmental, cultural, and political aspects) [4, 7, 8]. As a consequence, we believe that, in the same way as the economic conditions play a central role for those seeking employment, it is very relevant to investigate if there are other factors playing a role for specific kinds of migration, for example the migration of notable people and intellectuals in the course of history [9].

For the general problem of human migration different mathematical models have been built both in a descriptive [10–13] and predictive [14] framework. However, these models have never been applied to the specific scenario of modeling the historical movements and migrations of intellectual figures. Instead, mobility and migrations historically played, and still play, an important role in the process of cultural evolution introducing seeds of change in different places around the world [16]. Thus, understanding the patterns that historically notable individuals followed during their lives and how these affected the cultural evolution and the human history is an intriguing and still open research question [17]. This approach introduces new challenges because of the specificity of the problem and the relatively small number of people that had an impact on cultural evolution and human history. New perspectives were opened by a recent approach proposed by Schich *et al.* [18]. By analyzing birth and death locations historically notable individuals they captured, from a network perspective, the key characteristics of their exploratory behavior.

In our paper, we propose a way of using Natural Language Processing (NLP) and Network Science techniques to further characterize and model the mobility of historically and culturally notable individuals and to investigate the factors playing a role in their migration patterns.

In the last decade, the NLP community has developed technologies for extracting information from unstructured texts, thus enabling their application also to interdisciplinary research areas. Understanding and modeling historical migration phenomena require specific historical data. To this end, we propose to use NLP techniques to process the digital biographies contained in Wikipedia and to extract migratory events from its encyclopedic information. In particular, a subset of the biographies available in the English version of Wikipedia is used as raw data source. From this source, for each notable person we search for her/his footsteps hidden in her/his Wikipedia page and collect the following information: (i) the place and date of birth, (ii) the place and date of death, and (iii) the place and date of the various in-life migrations (e.g. moving from one city to a different one). This results in a more complete set of data with respect to [18], enriching the global picture of the mobility of notable individuals with a finer temporal granularity, while enabling, to the best of our knowledge, to model this process from a historical point of view for the first time.

As proposed in [19], we use the term *culture* to focus our attention on the set of notable contributions to the development of human history in its broadest sense: from poetry to sports, from music to physics and mathematics. In particular, we model the mobility

dynamics of notable people, namely those people whose cultural production is known at a global level. To this end, we introduce a modification to the *radiation model* for human mobility [13]. The assumption behind the *radiation model* resides in using the city size (i.e. the city population) as a proxy of the number of job opportunities. In our paper, we modify the *radiation model* to take into account, in addition to the role played by the city size, also the attractive role played by the different disciplines and by the number of notable people as proxies of cultural opportunities.

Then, we compare the predictive performances of our *cultural-based radiation model* and of the state-of-the art population-based radiation model on three main aspects of the migration processes of notable people: (i) the radius of gyration of each notable person, (ii) the number of different cities the notable people lived in during their lives, and (iii) the distances between the source of a migration and its destination (jump lengths). Interestingly, our results show that the radius of gyration and the jump lengths are best modeled considering three different factors: (i) the population of the city, as a proxy for the economical wealth and job opportunities, (ii) the number of notable individuals that spent some time of their lives in a given city, as a proxy for the role played by the city as a cultural attractor, and (iii) the specific discipline an historically and culturally notable person is working in, as a proxy both for the interests a city has in investing on a specific cultural area and for the tendency that people, interested or working on that discipline, have to follow notable figures from the same domain.

Our results pave the way for further investigations on the historical role played by places and cities (e.g. ancient Athens, Renaissance Florence, Song Dynasty Hangzhou, Vienna of 1900, Silicon Valley, etc.) in becoming cultural attractors for historically and culturally notable individuals and, thus, making them grow into flourishing places for novel artistic and literary movements, for new scientific and philosophical theories, for social, technological and political innovations [20].

2 Materials and methods

2.1 Data

In this section, we present and discuss how we build a dataset containing biographical information on thousands of historically relevant people in the context of cultural production and innovation. Together with information about their field of influence, we also extract information about their mobility patterns.

We start from the set of notable people identified by the Pantheon project [19]. This project collected the biographies of 11,341 historically and culturally relevant people that lived from 3500 B.C. to 2010 A.D. More precisely, the dataset was built by extracting information from the Wikipedia biography and info-boxes. Here, a person is defined as *notable* if the corresponding Wikipedia page is translated in 25 or more different languages, since the focus is on global historical and cultural contributions.

For each notable person, several features were annotated and manually verified, among which we consider the following ones:

- birth place (geo-localized), state, and date;
- occupation, work area, and discipline.

While the process of selection of notable people in Yu *et al.* [19] resulted in 11,341 biographies selected from more than one million biographies available in Wikipedia, we further restrict our analyses on the notable people that were actively working and producing cultural outcomes in a specific time-window, the first fifty years of the 20th century.

Table 1 Composition of the selected subset of notable people, organized by the different work areas and disciplines present in the Pantheon dataset [19]. In particular, 15 work areas are presented, sorted by discipline

Work area	Discipline	Percentage (%)
Film and Theatre	Arts	10.35
Music	Arts	8.81
Fine Arts	Arts	5.61
Design	Arts	1.68
Natural Sciences	Science and Technology	13.74
Social Sciences	Science and Technology	3.92
Medicine	Science and Technology	2.25
Math	Science and Technology	2.04
Language	Humanities	17.39
Philosophy	Humanities	3.51
Government	Institutions	15.90
Military	Institutions	3.81
Activism	Public Figure	1.23
Individual Sports	Sports	1.28
Business	Business and Law	1.07

The focus on this specific time-window is justified by two combined needs. The first is to have a sufficiently short time span so that the properties of the global mobility do not change. The second is to have the highest possible number of notable people with complete and precise migratory information, i.e. with birth, in-life, and death locations. The first requirement reduces the width of the time-window, while the second one requires to consider relatively recent years. We therefore consider active the notable people that were at least 20 years old by the end of the considered time-window to ensure that their historical and cultural contribution was made during this time span, and that no bias in the type of migratory event (birth, in-life, and death) is introduced.

By applying this filtering procedure based on the time-window of interest, we reduce the number of Wikipedia biographies to be processed to 2407. We report in Table 1 the distribution of work areas and disciplines present in this subset of Wikipedia biographies, which will be further processed as described in the next Section.

2.1.1 Extracting migration footsteps

For the purposes of our study, we are interested in identifying the different locations that were visited by the selected set of notable people during their lives and the year of their visits, in order to build a trajectory (made of multiple footsteps) for each notable person. This kind of information is not present in the Pantheon dataset and we therefore need to extract it automatically. The approach we adopt follows the one recently proposed with Ramble-On [21], a text processing pipeline dealing with two main tasks: (i) the identification of predicates of migration and their arguments (i.e. the subject of the migration frame) in Wikipedia biography pages, and (ii) the recognition and classification of dates, places, and mentions. We focus our attention on migration processes because they are more likely to describe a motion action that resulted in a long time permanence in a new location. As a consequence, if the permanence in a specific location is long, it is more likely that the notable person had the time to provide his/her cultural contribution there. While there have been recent attempts to automatically extract people's trajectories and an associated time period from Wikipedia biographical pages [22], these rely on shallow NLP

approaches based on the presence of keywords and geo-links in the pages. Furthermore, trajectories are coupled with time spans and not time points like in our study, leading to a lower granularity of the extracted information. In our case, the use of semantic parsing associated with a selection of predicates describing possible trajectories enables a very precise analysis of the resulting data.

To identify the predicates related to migration events, the Ramble-On application^a calls PIKES [23], a suite for NLP that extracts information from English plain texts, which in turn automatically assigns to each predicate a semantic frame based on the FrameNet classification [24]. Also the arguments attached to each predicate are automatically labeled with semantic roles, relying on the frame-semantic parser Semafor [25]. To better distinguish migration predicates, we again follow the approach proposed in [21], thus removing 16 motion frames (e.g. *Escaping*, *Getting underway*, *Touring*) out of 45 because of the high number of false positives found during the identification. Hence, 29 motion frames (e.g. *Arriving*, *Being employed*, *Transfer*, *Travel*) were used for the identification of notable people's migration actions described in their biographies.

Once a migration frame was identified in a sentence extracted from a Wikipedia biography page, three elements are required to be present in order to extract a migration trajectory: (i) the time/date of the motion, (ii) the traveler, and (iii) the destination. Again, the Ramble-On application selects only sentences satisfying these constraints, where the date, the notable person (or a reference to him/her) and the destination have been identified. With this approach, precision is favoured over recall, requiring that these three elements are explicitly mentioned in the same sentence.

In Table 2 we show as an example the snippet of a sentence identified as a movement. In this case the predicate “moved” is assigned to the frame “Motion” and it is selected as a migration frame. Then, the different arguments of the sentence are identified and labelled depending on their role in the sentence (e.g. the time, the traveller and the destination/place).

Once the movements have been identified and the information about the date and the location have been extracted, Ramble-On geo-locates each word related to the identified destinations using OpenStreetMap *Nominatim*,^b a search engine for geo-referenced OpenStreetMap locations. Destinations that lack coordinates are discarded from the movements' list since they may be erroneously annotated as destinations. Besides, place and date of death are retrieved for each biography by Ramble-On using *DBpedia*,^c where structured data about each notable person are stored.

In Table 3 we present the information retrieved by processing the same biography as in Table 2. Together with the birth information retrieved from Pantheon, other important information such as date and place of death are extracted from *DBpedia* as discussed above. The *Migration 1* element, M_1 , is the first additional trajectory retrieved by running the Ramble-On tool on the biography page. Merging the information retrieved about the inventor and musician Les Paul (see Table 2), we identify two jumps, i.e. two migratory

Table 2 Example of identification and classification of a sentence using Semaphor and FrameNet

Snippet	Predicate	Frame	Place	Time
“Paul moved to Chicago in 1934, where he continued to perform on radio.”	moved	Motion	Chicago	in 1936

Table 3 Results obtained by processing Les Paul's biography using the Ramble-On pipeline

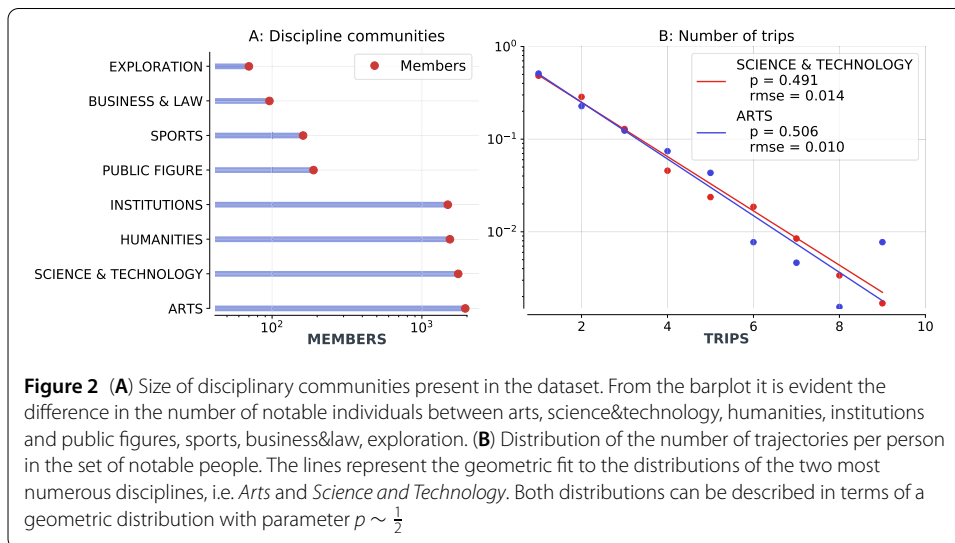
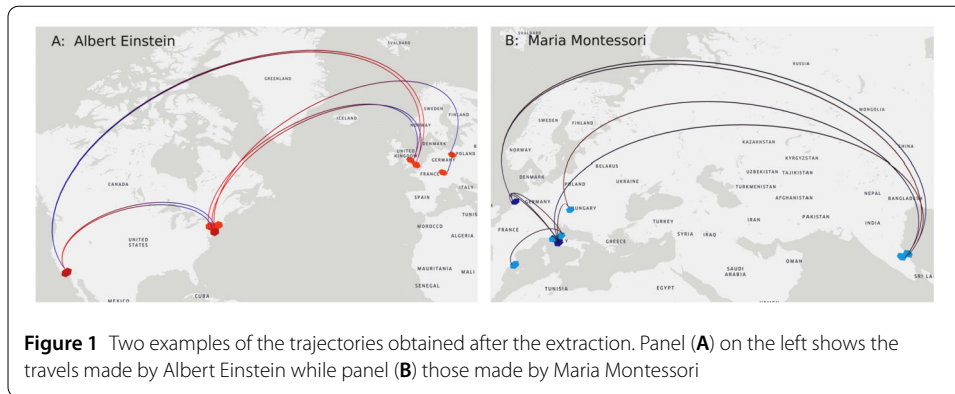
	Birth	Movement 1	Death
Date	19150000	19340000	20091231
Place	Waukesha	Chicago	White Plains
Latitude	43.0117	41.8369	41.0400
Longitude	-88.2317	-87.6847	-73.7786
Predicate	null	moved	null
Resource	dbpedia	FrameNet	dbpedia
Place frame	null	@Goal	null
Resource frame	Birth	Motion	Death

events: (i) the first one from the birth location to *Chicago*, and (ii) the second one from *Chicago* to the death location.

We refer to [21] for a detailed discussion on the performance of the extraction process. In brief, Ramble-On has a precision of 0.86 in correctly identifying migration frames. As mentioned before, however, the strategy adopted to identify trajectories may penalize recall, failing to extract movements whose date or destination are mentioned implicitly or in two different sentences. In order to estimate the amount of locations visited (birth and death locations included) that we are not able to capture in our study, we manually annotate the trajectories in the biographies of 50 notable people randomly sampled stratifying over the number of locations found. In this way, we have estimated that the recall of Ramble-on approach is equal to 0.59.

2.1.2 Dataset composition

The processing of the 2407 biographies results in a set of 7240 locations connected with notable persons' trajectories. Among these, we consider the 4028 movements taking place in the 1900–1950 time-window. Each movement with the associated date and destination was then manually checked by comparing the extracted information with the source Wikipedia sentence, and corrected if necessary. Also the coordinates associated with the extracted locations by *Nominatim* were manually checked, since the geographical information associated with trajectories is at the core of our migration model and possible errors must be minimized. These are then collapsed to the nearest *great city*, where we adopted as a definition of *great cities* the list proposed in [26]. In their work, Reba *et al.* [26] collected also precious historical demographic data for most of these cities, that we used to test our baseline for the migration model. More specifically, geo-localized locations are merged based on a Voronoi tessellation of the Earth. In this framework each cell is built from a list of cities for which historical population data was available [26]. The space is built using the great-circle approximation to associate each identified location with the corresponding Voronoi cell. The distribution of the clustering process is reported in Fig. SM7 of the Supplementary Materials (SM) (Additional file 1). In Fig. 1 we present two examples of trajectories for Albert Einstein, the famous physicist, and Maria Montessori, the renowned physician and educator. The arcs connects different locations where these two notable figures spent a part of their lives. The blue coloured side of an arc indicates the origin of the migration while the red one its destination. For example, the extraction well captures Einstein's movements from Zurich to Berlin and from Berlin to US. His first movement from Ulm, his home town, is missing since it happened before the beginning of 20th century. We also notice that his short period as visiting professor to Caltech is detected by Ramble-On. Similarly, Maria Montessori's experiences around Europe (i.e.

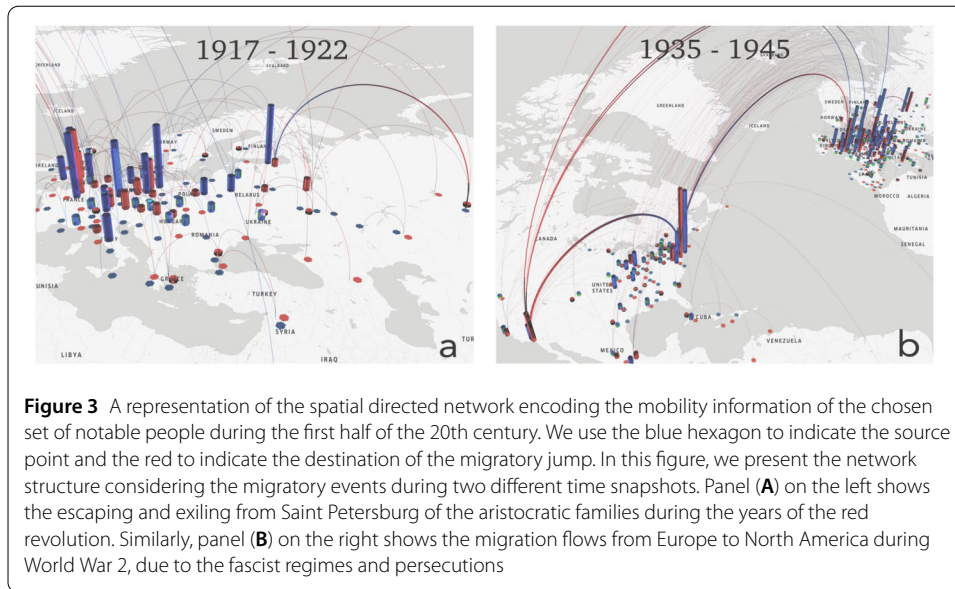


Barcelona, Amsterdam, Vienna, Rome) are correctly identified. In contrast, we stress that, due to the lack of population data, her trips to Sri-Lanka are collapsed to cities in India. This shows how the collapsing process might impact the actual migration distribution.

The merging step results in a set of 629 different cities visited by our notable people during their lifetime. Figure 2(A) shows the members of our set of notable individuals listed by discipline (as labelled by [19]); while Fig. 2(B) shows the distribution of different visited cities for the top two discipline communities, namely “Arts” and “Science and Technology”. The colored dots represent the data while the lines the geometrical fit to the data. Both the distributions can be described using a geometrical distribution with parameter p , representing the probability of successfully settling in a city, $p \sim \frac{1}{2}$.

2.2 Cultural network and migration modeling

In our framework, we assume that a culturally notable person living in a place for a certain period of time contributed in some way to make such place a cultural attractor for other people interested in cultural innovation and development. At the same time, when a culturally notable person was moving from one place to another, s/he linked the cultural relations s/he had and the work s/he did in the first place with the relations and work in the second place. In this way, each movement creates a cultural connection between two different places around the world. Depending on the number of notable people moving from



one place to another, we can add a weight to the links of this cultural network. Thus, the nodes correspond to the different cities visited by our set of notable people, while the edges are cultural links, built by interpersonal relations together with the cultural contamination a person brings with herself/himself while migrating from one place to another, weighted by the number of occurrences.

In Fig. 3 we show a representation of this weighted directed network, where the weight is the percentage of notable people migrating from a location to their destination. In the left panel and right panel we show a sector of the network considering different time frames. On the left, we restrict our focus on a 3-year time-window centering the map to highlight the connections between Eastern Europe and Asia (Union of Soviet Socialist Republics in particular) during the October revolution in 1917. It is interesting to notice that the structure of the network nicely catches the known phenomenon of Siberian exile of aristocratic families and important political personalities such as representatives of the previous tsarists' power and relevant persons not aligned with the current regime. In this example, we are capturing the movements of people that were forced to migrate to specific locations. Hence, our notion of cultural attractor both includes locations to which people moved voluntarily and locations to which culturally notable people were forced to move. In particular, the number of forced movements (i.e. people sent to concentration camps and imprisoned people) is 36 out of 3474 identified movements. Top right panel shows the strong migratory flux of intellectuals from Europe to the US during the Second World War.

As previously said, we are interested in modeling the mobility of culturally relevant figures and in investigating the factors playing a role in their migration patterns. To this end, we modify the radiation model to take into account the role played by the city size (i.e. proxy for job opportunities) as well as the ones played by the different disciplines and by the number of notable people (i.e. proxies for cultural opportunities). The radiation model [13] describes the mobility of people seeking job opportunities in terms of job openings per number of inhabitants. The model is developed in the framework of network theory since it treats cities as nodes of a completely connected weighted network. Specifically,

the radiation model describes the human mobility behavior at long distances, e.g. at the country or global scale, better than other often used models (e.g. gravity model) [27]. It is also important to notice that it undershoots the real flows. Moreover, its performances are dependent on the structure of the system even though it directly accounts for variations of the population between the source and the destination of a migration, i.e. the less population you have between two cities the more probable is to migrate from one to the other.

Simini *et al.* [13] show that, using this formulation of the problem, the flow of people between cities only depends on the population of the two cities (namely m_i and n_i), and the population living in the circle of radius r_{ij} is equal to the distance between the two cities (namely s_{ij}). The relation can be summarized in a simple and parameter-free equation. We report here the formula for the probability to move from city i to city j :

$$P_{ij} = P_i \frac{n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \quad (1)$$

where P_i is the normalization coefficient for city i that ensures that P_{ij} is the probability of moving from i to every city (i included): $P_i = \sum_{j \in N} \frac{(m_i + s_{ij})(m_i + n_j + s_{ij})}{n_j}$, where N is the set of all nodes present in the network. In our work we make use of these concepts to model in a similar way the mobility of culturally relevant people. In particular, inspired by multi-dimensional network theory and its recent applications in modeling human mobility [15, 31, 32], we propose a multilevel approach to cultural mobility. In this framework, every cultural discipline works as a separate system described by a cultural radiation model. Formally, a level is a fully connected weighted and directed network in which the nodes are the cities visited by all the notables of a specific discipline and the links represent the probability of migrating from a city to a different one. Each node has also a link pointing to itself, representing the probability of remaining in the same city instead of moving to a different place. The different levels, $l \in L$ where L is the set of all disciplines, do not interact with each other but their contribution to the overall migratory exploration sums up. Each level contributes to the global migration model with a factor proportional to the share of notable people the discipline has, NS_l . Thus, the probability of this multilevel migratory network can be described by

$$P_{ij} = \sum_{l \in L} NS_l P_{il} \frac{n_{jl}}{(m_{il} + s_{ijl})(m_{il} + n_{jl} + s_{ijl})}, \quad (2)$$

where NS_l is the notable share of discipline $l \in L$, m_{il} and n_{jl} are the population of locations i and j respectively in the discipline level l , and s_{ijl} is equivalent to s_{ij} for the specific level l . In a similar way, the generalized P_i normalizes P_{ij} to a probability following the idea of equation (1):

$$P_i = \sum_{j \in N} \frac{(m_{il} + s_{ijl})(m_{il} + n_{jl} + s_{ijl})}{n_{jl}}. \quad (3)$$

Starting from these two equations, we propose different implementations of the radiation model. In particular, we stress that the radiation model introduces the concept of attractiveness of a city based on the number of job opportunities that a city can provide.

The assumption that this number is directly proportional to the number of people living in a city [13] directly connects the concept of attractiveness with the population size of a city. Here, with similar assumptions we propose different formulations of the model based on different possible ways of modeling cultural attractiveness. In particular, we use, as possible alternatives to the standard formulation of the model, the number of notable people that visited a city and its combination with the population of the city. In the case of notable people that visited a city, we count this quantity considering all the visits during the whole time-window. As a consequence, we model cultural dynamics of individuals within this period of observation considering the effects of the notable people distribution as a constant feature of our model, as we do for general population, which is not updated after every step of the dynamics. This is equivalent to assume that a single step of the dynamics has a latency larger than the size of the considered time-window in affecting the importance of cultural attractors. Moreover, counting notable people in this way also relies on the simplifying assumption that they all equally contributed to the importance of a city from a cultural perspective. More realistic modeling will require a *relevance score* based on the historical relevance of the notable people.

We test each of these three possible definitions, namely the standard one based on population size, the one based only on the number of notable people, and the combination of the two, using both a single level formulation of the radiation model and a multilevel formulation. Using this probabilistic model, our aim is to understand if the radiation model abstraction can be used to describe the level of exploration of the historically and culturally relevant figures (namely, the radius of gyration of each notable people, the number of different cities visited, and the distance distribution of the migration jumps) and which formulation better captures these properties. In the next Section, we discuss first the general information that can be obtained by analyzing the system in terms of network theory metrics and then the comparison between the different formulations proposed.

3 Results

3.1 Properties of the migration network

One of the most interesting characteristics of cultural migration patterns is the tendency of notable figures to explore different cities. To study this property, we can define $S(t)$ as the number of different cities and $N(t)$ as the number of notable people's birth locations, the number of their death locations, and the number of their jumps during the selected time-window, as displayed in Fig. 4(A). by the curves for *Birth*, *Death* and *In-life* respectively. The growth of $S(t)$ is modeled as a function of $N(t)$ using a Heap's law $S(t) = N(t)^\alpha$. Our result is consistent with the estimate of the parameter α for the *Birth* curve obtained by Schich *et al.* in [18]. A similar result is obtained also for the *Death* and *In-life* curves representing the growth of the location for which the exponent, $\alpha = 0.85$, suggests a tendency to migrate to a smaller number of cities with respect to the number of different cities where notable individuals were born. This finding may be interpreted as a general and global tendency of notable figures to migrate to a more culturally renowned subset of cities with respect to all the possible available locations.

Focusing on the migration jumps that notable people made during their life, we can study the most central cities both from a global and discipline-based perspective. Here, we use the Page-Rank centrality [28] to measure the importance in terms of the number of incoming links that point to a city and the relative importance of the cities from which

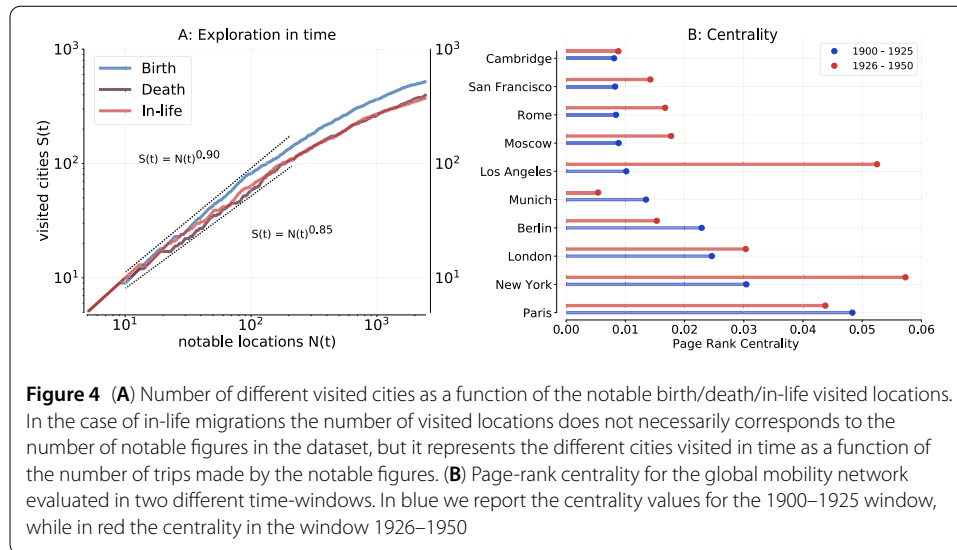


Figure 4 (A) Number of different visited cities as a function of the notable birth/death/in-life visited locations. In the case of in-life migrations the number of visited locations does not necessarily corresponds to the number of notable figures in the dataset, but it represents the different cities visited in time as a function of the number of trips made by the notable figures. (B) Page-rank centrality for the global mobility network evaluated in two different time-windows. In blue we report the centrality values for the 1900–1925 window, while in red the centrality in the window 1926–1950

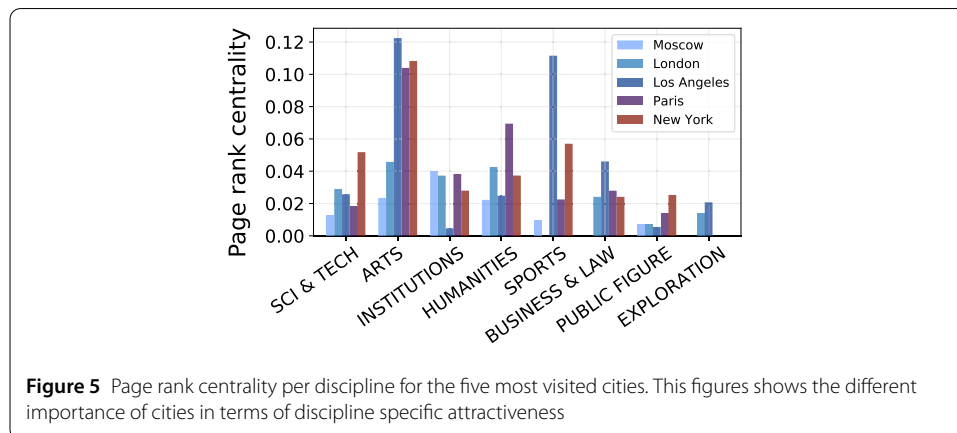


Figure 5 Page rank centrality per discipline for the five most visited cities. This figures shows the different importance of cities in terms of discipline specific attractiveness

these links are coming. In Fig. 4(B), we measure Page-Rank centrality for two different time-windows, namely 1900–1925 and 1926–1950, to show the structural changes of the network during the first half of the 20th century. It is interesting to notice how the development of the film industry in Los Angeles attracted several figures to the city. It is also worth noticing how, due to the Second World War (WW2), Berlin loses positions in the ranking of the *more central* cities. In Fig. SM4, we highlight the specific effect of WW2, evaluating Page-Rank centrality before and after the rise of the Nazism regime in Germany, showing the overall loss in cultural centrality for most of the European cities.

We also evaluated Page-Rank centrality for the sub-network built only by considering the migration jumps of the four top disciplines in those years. Figure 5 shows how notable people from different disciplines migrated to different cities, suggesting that the cultural centrality of a city depends on its cultural characteristics, e.g. Los Angeles and the film industry. This, indeed, results in Los Angeles being a central node for Arts (in particular for *actors*) and Sports, but a more peripheral one for Institutions, having their more central nodes in capital cities such as London, Paris and Moscow. In Fig. SM3 we performed a discipline Page-Rank analysis for two time windows as in Fig. 4, showing that an important change is present also at the discipline level. An example is given by the dramatic change

(i.e. a decrease) in the centrality of Berlin for the scientific community before and after 1933.

3.2 Cultural attractiveness in a multilevel radiation model

With the results obtained so far we aim at modeling notable figures' migratory patterns to better understand what is the process driving the choice of the location where to migrate. The radiation model proposed in [13] finds the motivating factor of mobility for those seeking a job in the number of opportunities a city can provide. Following this idea and using the results obtained by Simini *et al.* [13], we propose a similar approach to understand if such a model can catch the main factors of cultural mobility. We assume that the number of opportunities that are available in the selected time-window is directly proportional to the number of notable figures that lived in a city during the same time interval. In particular, we explore the following different configurations:

- cultural opportunities are uniformly distributed among cities;
- cultural opportunities are proportional to the population of a city;
- cultural opportunities of a city are directly proportional to the number of notable figures that lived in that city;
- cultural opportunities are directly proportional both to the population of a city and to the number of notable figures that lived in that city.

In addition, we also want to check whether cultural opportunities depend on the discipline a notable individual is part of. We study all these possibilities using the formulation proposed in Sect. 2.2.

To find which model better describes the historical mobility of the first half of 20th century, we simulate mobility using a set of walkers that can move following a radiation model over the cultural network, based on different equations depending on the model we are simulating. Thus, depending on the selected configuration, we are using respectively (i) a random walker dynamics' model over the network, (ii) a standard (population-related) radiation model, and (iii) an implementation of the radiation model that specifically considers the cultural opportunities as a subset of job opportunities not necessarily driven by the same factors. The starting location of each simulated notable individual is chosen based on a random choice weighted by the population size of each city.

To compare the models we analyze their impact on predicting (i) the number of different cities visited, as a proxy of the availability to explore different and new destinations, (ii) the radius of gyration of each notable figure simulated, and (iii) the distribution of the length of the migration jumps. We report in Table 4 the results obtained for five representative models: (i) the random walker model, (ii) the notable-based jump probability on a single-level structure, (iii) the notable-based jump probability on a multilevel structure, (iv) the population-based jump probability on a multilevel structure, and (v) the mixed population-notable-based jump probability on a multilevel structure. Results for the other models are reported in the SM (Additional file 1), specifically in Table SM1, Table SM2, and Table SM3.

Among the different possibilities tested, we found that a modeling approach considering cultural attractors as a product of both job-opportunities and cultural interests, by means of the population number and the effective number of notable people migrated in the time-window under investigation, better captures key features of notables' mobility. Moreover, we stress that the model treating different disciplines as different dynamics outperforms

Table 4 Models' performances. Results obtained for the five models on predicting the number of destinations, the radius of gyration, and the distribution of the length of the migration jumps. The metrics used are the adjusted- R^2 , the Pearson correlation coefficient, ρ , between models and data, the Kullback–Leibler distance (K-L dist), and the first Wasserstein distance (Wass. dist)

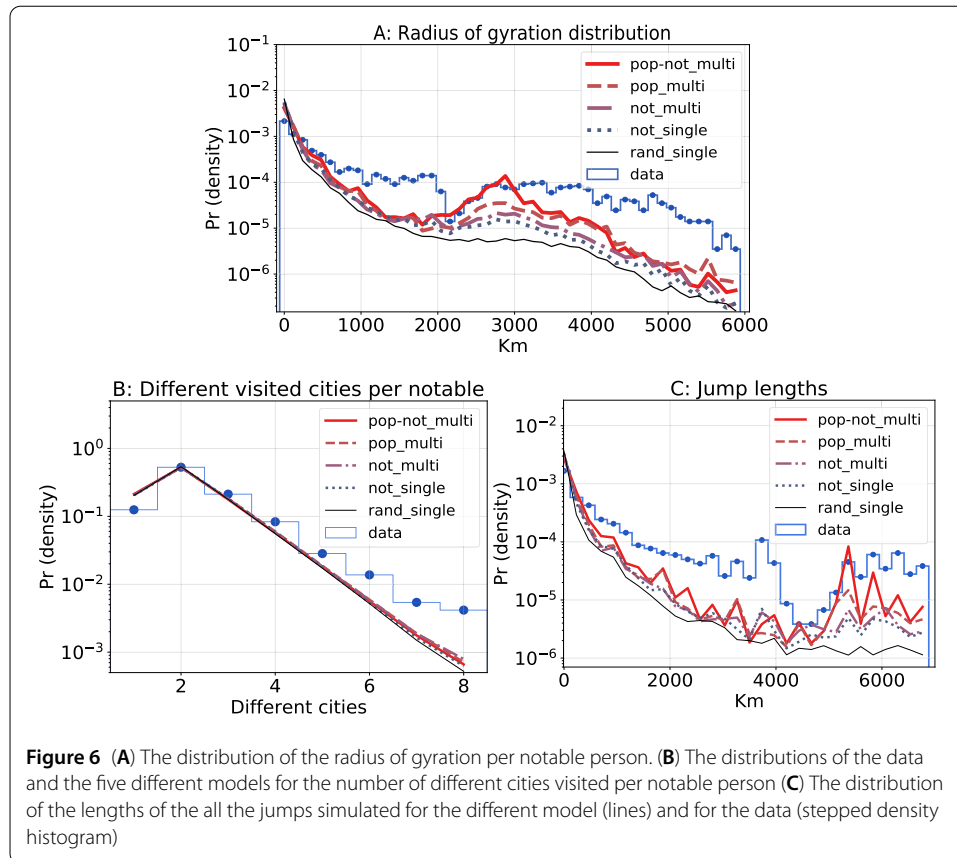
Model	adj- R^2	Pearson ρ	K-L dist	Wass. dist
<i>Radius of gyration</i>				
Pop-notable-multilevel	0.2414 \pm 0.0027	0.962***	0.00554 \pm 0.00004	0.000100 \pm 2e-7
Pop-multilevel	-0.2004 \pm 0.0034	0.953***	0.00655 \pm 0.00005	0.000125 \pm 1.e-7
Notable-multilevel	-0.6849 \pm 0.0041	0.947***	0.00836 \pm 0.00005	0.000139 \pm 1.e-7
Notable-singlelevel	-1.0249 \pm 0.0048	0.923***	0.01006 \pm 0.00006	0.000143 \pm 1.e-7
Random-singlelevel	-2.2673 \pm 0.0054	0.886***	0.01559 \pm 0.00009	0.000173 \pm 1.e-7
<i>Different destinations</i>				
Pop-notable-multilevel	0.9547 \pm 0.0004	0.978***	0.0649 \pm 0.002	0.0150 \pm 0.0001
Pop-multilevel	0.9612 \pm 0.0004	0.981***	0.0561 \pm 0.001	0.0154 \pm 0.0001
Notable-multilevel	0.9619 \pm 0.0003	0.982***	0.0570 \pm 0.001	0.0155 \pm 0.0001
Notable-singlelevel	0.9624 \pm 0.0003	0.982***	0.0623 \pm 0.002	0.0159 \pm 0.0001
Random-singlelevel	0.9606 \pm 0.0004	0.982***	0.0724 \pm 0.002	0.0163 \pm 0.0001
<i>Length of migration jumps</i>				
Pop-notable-multilevel	0.5104 \pm 0.0019	0.982***	0.00533 \pm 0.00005	0.000080 \pm 1e-7
Pop-multilevel	0.2249 \pm 0.0023	0.974***	0.00686 \pm 0.00005	0.000099 \pm 1e-7
Notable-multilevel	-0.0640 \pm 0.0029	0.967***	0.00795 \pm 0.00005	0.000109 \pm 1e-7
Notable-singlelevel	-0.2192 \pm 0.0029	0.962***	0.00790 \pm 0.00006	0.000112 \pm 1e-7
Random-singlelevel	-0.8313 \pm 0.0034	0.947***	0.01265 \pm 0.00006	0.000131 \pm 1e-7

the single-level models in terms of Kullback–Leibler divergence [29] and first Wasserstein distance [30].

These quantities are measured after simulating the mobility for 2000 notable people, whose number of migration jumps was randomly sampled from a geometric distribution with parameter $p \sim 0.5$. We repeated the simulation for 500 times to estimate the stability and the standard error of the different metrics. Figure 6(A) shows the distribution of the mean radius of gyration of the 500 simulations against the data (blue-stepped line). Figure 6(B) shows the distribution of different cities visited by a notable person during her/his lifetime, while the distribution presented in Fig. 6(C) shows the probability of jumping to a destination that is at a specific distance from the origin. These distributions are highly dependent on the geographical distance. Short distance migrations are largely privileged while at ~ 2500 km and ~ 5500 km respectively a second smaller peak appears capturing the overseas migrations (across the Atlantic Ocean) mainly from Europe to US. We stress that the effect of slightly underestimating the log-distance trips, which also affects the radius-of-gyration distribution, has been proven to be a structural feature of the radiation model under irregular geographical configurations such as those imposed by oceans [27].

4 Discussion and conclusions

A complex question has been posed in [18], i.e. whether it is possible to describe the dynamical properties of the cultural migration phenomenon. Starting from their idea of using network theory to tackle the problem, we make some further steps in understanding this issue. First of all, we use NLP tools to capture a more detailed representation of the lives of historically notable people that can be considered as cultural developers or important actors in the evolutionary process of culture. Our approach gathers information not only from the birth and death events but also from in-life migratory events, enabling us to study in a more detailed way the cultural migration processes and to include in our model the years in which a person is professionally more active. Indeed, there is a differ-



ence between the birth location of a person and the locations s/he migrates to or where s/he decide to spend her/his last years. The first birth location is not determined by a decision of the born individual, while the in-life migrations and the death places are more likely to be chosen following some precise interests and motivations. Using our data we are able to capture this difference and quantify the contribution of the exploration level of notable people during these phases of life.

Moreover, we focus our attention on understanding the main features that drive this kind of mobility. Our results provide evidence that the mobility of historically and culturally notable individuals is best described by simultaneously considering three different factors: (i) the population of a city, as a proxy of economic wealth and generic job opportunities; (ii) the number of culturally notable people that spent some time of their lives there, as a proxy of the attractive role played by this city as a cultural hub and of the proneness of this city to invest in culture; and (iii) the discipline a culturally notable person is working in, as a proxy both of the interest a city has in investing on a specific cultural area and the tendency of people, interested or working on a given discipline, to follow notable ones from the same discipline. The solution proposed in our work represent a functional integration, in a quantitative theoretical model, of these components.

It is also worth highlighting some limitations of our work. First of all, we rely on Wikipedia as data source, which shows a clear bias towards the Western culture and male figures. This limitation is even more relevant since we focus only on pages written in English. Then, we consider a specific time-window in the cultural history, i.e. the first half of the 20th century. So, our results may be dependent on the time chosen (for example, we

may observe different behaviors during wider time-windows) and on the small available dataset of historically and culturally notable personalities. Besides, while our data on trajectories were extracted automatically but manually revised, we estimate a recall of our information extraction pipeline of 0.59, as pointed out in Sect. 2.1.1. This implies that some migration destinations that are mentioned in the Wikipedia biographies should be added (e.g. by improving the extraction performances of Ramble-On) in order to make our set of data even richer. A richer dataset will also help in stabilizing, constructing and precisely characterizing the structure of the network. However, we also stress that the extension to the radiation model proposed here only uses the visited locations and not the migration timelines of notables. Thus, the present recall level in extracting complete timelines does not directly affect the dynamical structure of our model. Thanks to these considerations, it is also interesting to discuss the specific limitations of the mobility model proposed. In particular, while the fit of the number of trips might depend on the recall limitations discussed above and thus affect the number of different visited locations, this does not explain the systematical underestimation of the probability of *high-distance jumps* and *high-radius of gyration*. We expect the geographical constraints (as discussed in Sect. 2.2) and the specific time-window we selected (e.g. WW2 forced to migrate many notable people whose choice was biased towards US) to be two determinants of these discrepancies.

Overall, our results open interesting possibilities on further investigating the historical role played by places and cities in attracting culturally relevant figures as well as on better analyzing the level of contribution of each of the factors identified by our approach (i.e. city's population, number of intellectuals living in the city, and strength of a specific cultural discipline in the city). Similarly, changing the perspective, it will become possible to quantify the impact of cultural communities on local well-being, helping our understanding on how individuals from similar or different disciplines combine and collaborate to seed the vital growth of cities' economies.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-019-0215-7>.

Additional file 1. Supplementary information (PDF 1.3 MB)

Funding

No funding supported our research.

Abbreviations

NLP, Natural Language Processing; WW2, World War 2; US, United States; SM, Supplementary Material.

Availability of data and materials

The data used in this work are available at: <https://doi.org/10.7910/DVN/PJS21L> or https://figshare.com/articles/Following_the_footsteps_of_giants/7352987. The code used to extract the destinations from the Wikipedia biographies is publicly released at <https://github.com/dhfbk/rambleon>.

Competing interests

The authors declare no competing financial or non-financial interests.

Authors' contributions

All authors conceptualized the project. LL acquired and cleaned the data, performed the investigation, the statistical analyses and drafted the original manuscript. All authors contributed revising the manuscript and gave final approval for publication.

Endnotes

^a Code available here: <https://github.com/dhfbk/rambleon>

^b <https://nominatim.openstreetmap.org/>

^c <http://wiki.dbpedia.org/>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 4 June 2019 Accepted: 3 December 2019 Published online: 12 December 2019

References

1. Manning P (2005) Migration in world history. Routledge, London
2. Lee ES (1966) A theory of migration. *Demography* 3:47–57
3. Arango J (2000) Explaining migration: a critical view. *Int Soc Sci J* 52:283–296
4. Castles S (2000) International migration at the beginning of the twenty-first century: global trends and issues. *Int Soc Sci J* 52:269–281. <https://doi.org/10.1111/1468-2451.00258>
5. Todaro MP, Smith S (2000) Economic development. Addison Wesley, Boston
6. Guiso L, Sapienza P, Zingales L (2009) Cultural biases in economic exchange? *Q J Econ* 124:1095–1131
7. Belot M, Ederveen S (2012) Cultural barriers in migration between OECD countries. *J Popul Econ* 25:1077–1105
8. Beine M, Parsons C (2015) Climatic factors as determinants of international migration. *Scand J Econ* 117:723–767
9. Khoshkish A (1966) Intellectual migration: a sociological approach to 'brain drain'. *J World Hist* 10:1095–1131
10. Zipf GK (1946) The P1 P2/D hypothesis: on the intercity movement of persons. *Am Sociol Rev* 11(6):677–686. <http://www.jstor.org/stable/2087063>
11. Stouffer SA (1940) Intervening opportunities: a theory relating mobility and distance. *Am Sociol Rev* 5(6):845–867. <http://www.jstor.org/stable/2084520>
12. Block HD (1974) Random orderings and stochastic theories of responses (1960). In: *Economic information, decision, and prediction*. Springer, Dordrecht, pp 172–217. ISBN 978-94-010-9276-0. https://doi.org/10.1007/978-94-010-9276-0_8
13. Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. *Nature* 484:96–100. <https://doi.org/10.1038/nature10856>
14. Ahmed M, Barlacchi G, Braghin S, Calabrese F, Ferretti M, Lonij V, Nair R, Novack R, Paraszczak J, Toor A (2016) A multi-scale approach to data-driven mass migration analysis. In: *CEUR workshop proceedings*, vol 1831.
15. Ferretti M, Barlacchi G, Pappalardo L, Lucchini L, Lepri B (2018) Weak nodes detection in urban transport systems: planning for resilience in singapore. In: *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)*, pp 472–480. <https://doi.org/10.1109/DSAA.2018.00061>
16. Cavalli-Sforza LL, Feldman MW (1981) Cultural transmission and evolution: a quantitative approach. Princeton University Press, Princeton. <http://groups.lis.illinois.edu/amag/langev/paper/cavallisforza81book.html>
17. Brewer J, Gelfand M, Jackson JC, MacDonald IF, Peregrine PN, Richerson PJ, Turchin P, Whitehouse H, Wilson DS (2017) Grand challenges for the study of cultural evolution. *Nat Ecol Evol* 1:0070. <https://doi.org/10.1038/s41559-017-0070>
18. Schich M, Song C, Ahn Y-Y, Mirsky A, Martino M, Barabási A-L, Helbing D (2014) A network framework of cultural history. *Science* 345(6196):558–562. <https://doi.org/10.1126/science.1240064>. <http://science.sciencemag.org/content/345/6196/558>
19. Yu AZ, Ronen S, Hu K, Lu T, Hidalgo CA (2016) Pantheon 1.0: a manually verified dataset of globally famous biographies. *Sci Data* 3:150075. <https://doi.org/10.1038/sdata.2015.75>
20. Weiner E (2016) *The geography of the genius: lessons from the world's most creative places*. Simon & Schuster, New York
21. Menini S, Sprugnoli R, Moretti G, Bignotti E, Tonelli S, Lepri B (2017) RAMBLE ON: Tracing movements of popular historical figures. In: *Proceedings of EACL 2017*
22. Gergaud O, Laouénan M, Wasmer E et al (2016) A brief history of human time: exploring a database of 'notable people'. Technical report 63, Sciences Po Department of Economics
23. Corcoglioniti F, Rospocher M, Aproso AP (2015) Extracting knowledge from text with PIKES. In: *ISWC 2015 posters & demonstrations track*, within the 14th international semantic web conference (ISWC 2015), Bethlehem, USA, October 11–15, 2015. http://ceur-ws.org/Vol-1486/paper_66.pdf
24. Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet project. In: *Proceedings of the 17th international conference on computational linguistics—volume 1. COLING '98*. Association for Computational Linguistics, Stroudsburg, pp 86–90. <https://doi.org/10.3115/980451.980860>
25. Kshirsagar M, Thomson S, Schneider N, Carbonell J, Smith N, Dyer C (2015) Frame-semantic role labeling with heterogeneous annotations. In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing*. <https://doi.org/10.3115/v1/P15-2036>
26. Reba M, Reitsma F, Seto KC (2016) Spatializing 6,000 years of global urbanization from 3700 BC to AD 2000. *Nat Sci Data* 3:160034. <https://doi.org/10.1038/sdata.2016.34>
27. Masucci AP, Serras J, Johansson A, Batty M (2013) Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. *Phys Rev E* 88(2):022812. <https://doi.org/10.1103/PhysRevE.88.022812>
28. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: *Proceedings of WWW 1998*
29. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):257–263
30. Olkin I, Pukelsheim F (1982) The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl* 48:257–263
31. Belyi A, Bojic I, Sobolevsky S, Sitko I, Hawelka B, Rudikova L, Kurbatski A, Ratti C (2016) Global multi-layer network of human mobility. *CoRR*. [arXiv:1601.05532](https://arxiv.org/abs/1601.05532)
32. De Domenico M, Solé-Ribalta A, Cozzo E, Kivelä M, Moreno Y, Porter MA, Gómez S, Arenas A (2013) Mathematical formulation of multilayer networks. *Phys Rev X* 3:041022. <https://doi.org/10.1103/PhysRevX.3.041022>