Supplementary Materials for

A Network Framework of Cultural History

Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky,
Mauro Martino, Albert-László Barabási, Dirk Helbing

*Corresponding author. E-mail: maximilian.schich@utdallas.edu

Published 1 August 2014, *Science* **345**, 558 (2014)
DOI: 10.1126/science.1240064

This PDF file includes

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/345/6196/558/suppl/DC1)

Movies S1 and S2
Databases S1 to S4 as zipped archives:
SchichDataS1_FB.xlsx (Freebase.com)
SchichDataS2_AKL.xlsx (Allgemeines Künstlerlexikon)
SchichDataS3_ULAN.xlsx (Getty Union List of Artist Names)
SchichDataS4_WCEN.xlsx (Winckelmann corpus)

A Network Framework of Cultural History

Contents

Data Collection and Preprocessing

We collected notable individuals including dates and locations for birth and death. In addition we also acquired related information on individual genres or professional roles. Locations in two of our datasets are geocoded, providing good quality latitude and longitude information. Supplementary table S1 summarizes the basic statistics. The nature, acquisition, and pre-processing of the datasets is explained further below.

Table S1. Basic statistics for the FB, AKL, ULAN, and WCEN datasets. 'Type' characterizes individuals as defined by dataset publishers; we further indicate the total number of 'individuals'; the number of individuals with complete information on birth and death that act as 'links', including 'percentage' of total; the number of 'locations' or nodes connected by the 'links'; and the availability of 'geocodes', i.e. latitude and longitude information for 'locations' in a dataset.

Source	Type	#individuals	#links	#locations	geocoded
Freebase.com (FB)	deceased people	447.000	120.211 (~28%)	37.062	yes
Allgmeines Künstlerlexikon (AKL)	artists	1.1m	153.794 (~14%)	57.169	no
Getty Union List of Artist Names® (ULAN)	artists	202.719	20.173 (~12%)	7.285	yes
Winckelmann Corpus (WCEN)	antiquarians	720	233 (~32%)	128	yes

A. Freebase.com (FB)

Freebase.com (*3*) is a large Google-owned knowledge base that is publicly editable and available under a Creative Commons Attribution (CC-BY) license, which allows for both sharing and remixing of the data (see http://creativecommons.org/licenses/by/2.5/). Freebase stores information in structured graph format, integrating and disambiguating data from a variety of sources, including Wikipedia, the Internet Movie Database (IMDb), and others. Freebase data was obtained with a Python script using the proprietary MQL query language via the public Google API. Beyond a large set of individuals that may be dead or not, Freebase contains about 447.000 'deceased persons' with different professions. Taking only individuals into account that include full birth and death date as well as location information, Freebase returns 120.211 birth-death records connecting over 37.062 locations, where lat/long information is included and covers a timeframe of 3500 years.

During FB data preparation, we reduced dates, such as 01.01.2012, to the year, 2012; we removed individuals with ages above 130 and below 0; we cleaned out latitude

and longitude coordinates that were out of range; and we queried and attached FB 'profession' and 'sex' (meaning gender) information to our birth-death records (see Table S2). For known bias in FB see the respective section below.

B. Allgemeines Künstlerlexikon (AKL)

Allgemeines Künstlerlexikon (1969-2011) *(4)* is the most comprehensive scholarly artist lexicon in existence. It is based on the Thieme-Becker and Vollmer lexica (1907-1962/1980) and is widely used as a reference in art research (*5, 6*). The publisher Verlag Walter DeGruyter, has provided us with a “controlled vocabulary extract” that is usually catered to librarians as a cataloging aid to disambiguate individuals. The originally provided extract contained about 1.1 million XML files, i.e. one file per individual, including an internal ID, a label name, variant names, professional specializations, birth and death information, as well as (dissociated from the latter) countries of activity. Taking only individuals into account that include full birth and death date as well as place information, the AKL data, as published in the SOM with permission of the provider to allow for replication of our conclusions, contains 153.794 birth-death records connecting 57.169 locations. The AKL covers a timeframe of 2500 years, and unfortunately does not include latitude and longitude information for birth and death locations.

During AKL data preparation we combined the granular XML files to a single list; we reduced dates to the year, with '-' indicating dates before common era (BCE); we systematically chose the first year in date strings such as '1485' in '1485/1490'; we disregarded fuzzy qualifiers such as 'gegen (towards)', 'nach (after)', 'um (around)', and 'vor (before)'; we assigned location label strings with a unique ID (without disambiguating or merging entries any further, such as 'Tiflis' (Russian) and 'Tiblisi' (Georgian), or 'London' and 'London?'. We did not fully geocode the AKL locations, as AKL goes beyond the coverage of existing web services both temporally and spatially. Further major hurdles for geocoding location-label-strings in AKL are their notation in a mixture of German and native language transliteration, plus a disconnect of town and country information. We took into account AKL professional roles (see Table S2). For known bias in AKL see the respective section below.

C. The Getty Union List of Artist Names® (ULAN)

The Getty Union List of Artist Names® (*7*) is a controlled vocabulary that, similar to the AKL, is used as a cataloging and retrieval aid in art research. The Getty Research Institute has provided the authors with a full ULAN dump of relational tables. Part of the Getty Open Content program, the ULAN dataset is now made available as Linked Open Data, with no restrictions in terms of further use. The data contains about 202.718 'preferred names', of which we take only individuals into account that include full birth and death date as well as location information. After filtering, ULAN data contains 20.173 birth-death records connecting 7.286 locations. Like AKL, the ULAN covers a timeframe of 2500 years. While the ULAN dataset is substantially smaller than AKL, it contains expert curated latitude and longitude information for locations, linking to the additionally provided Getty Thesaurus of Geographic Names® (TGN), which is also now available in the Getty Open Content program (*31*). ULAN includes professional specializations and roles (see Table S2). For known bias in ULAN see the respective section below.

D. Winckelmann Corpus (WCen)

The Winckelmann Corpus (*11*) is a scholarly reference database that brings together ancient Greek and Roman sculptures and monuments as known by 18th century pioneer classical archaeologist Johann Joachim Winckelmann and others. The publisher of the 2001 DVD version, Verlag Biering & Brinkmann, provided us with the data, allowing us to map and analyze the full dataset as a network of complex networks (cf. *32*).

Beyond its main intended focus on monument-documentation, we can use the Winckelmann Corpus to construct a network, whose Giant Connected Component (GCC) connects 128 locations all over Europe across birth and death of 233 individuals. The strikingly heterogeneous structure of this small sample network, is visible in a clean way in Figure 1D. The raw version of this figure, based on data included in the SOM with permission of the provider to allow for replication, gave the initial hint for the viability of our macroscopic perspective on cultural history using birth-death data.

E. Google NGrams (Ngrams)

The Google Ngram datasets record the frequency of words and word-combinations in an estimated 5% of all books ever published, and were originally used and provided to establish the field of culturomics (*29*). For our purpose, we downloaded the 'Google Ngram English dataset version 20090715' (*28*).

From the data, we first extracted 3-grams and 4-grams by using regular expressions "^[A-Z].+ in [0-9]+" and "^[A-Z].+ .+ in [0-9]+", where the expression matches strings such as 'Paris in 1789' or 'Los Angeles in 1984' respectively. In the resulting preprocessed file, we looked up the Ngram count for each FB and AKL location-label-string in each year from 1400 until 2009. For instance, for 'Los Angeles', we identified all Ngram counts for the 4-grams 'Los Angeles in 1400', 'Los Angeles in 1401', etc. to 'Los Angeles in 2009'. For known bias in the Google Ngrams English dataset see the respective section below.

Table S2. FB, AKL, and ULAN include hundreds of professional roles and specializations by genre, in addition to plain birth and death information. We manually curated these roles and genres into a small set of categories: For FB we summarize performing arts, creative, governance (including politicians, lawyers, military, activists, and religion), academic (including education and health), sports, and business (including industry and travel). For AKL and ULAN we aggregate applied arts, architecture, fine arts, performing arts, and print & graphics. ULAN also includes academics and others. The table indicates our aggregate 'category' names, their 'total' frequency across individuals in a given dataset, and a list of most frequent 'example roles' or genres (with frequency). In all datasets, individuals can belong to more than one category.

AKL category	**Total**	**Example AKL roles (including frequency)**
Applied Arts	23633	goldsmith (3466), caramicist (2147), restorer (1379), designer (1324), silversmith (1148), decoration painter (1134), decorator (1007), etc.
Architecture	19171	architect (16494), master-builder (1417), stone-mason (374), urban planner (325), mason (286), building engineer (192), carpenter (155), etc.
Fine Arts	112276	painter (79174), sculptor (19922), landscape painter (10490), portrait painter (4305), genre painter (3977), watercolorist (1896), woodcarver (724), etc.
Performing Arts	3289	stage designer (2154), scenographer (530), costume designer (396), film maker (99), stage painter (87), stage decoration painter (86), film artist (77), etc.
Print&Graphics	58735	draughtsman (25127), graphic artist (14381), illustrator (9162), etcher (7063), lithographer (5693), photographer (4186), engraver (4046), etc.
ULAN category	**Total**	**Example ULAN roles (including frequency)**
Applied Arts	1803	designer (687), medalist (232), decorative artist (146), restorer (96), goldsmith (93), furniture designer (79), interior designer (66), etc.
Architecture	2488	architect (2331), engineer (168), urban planner (106), landscape architect (61), mason (27), garden designer (26), military engineer (21), etc.
Fine Arts	16331	painter (14309), sculptor (2348), portraitist (908), landscapist (822), watercolorist (750), genre subject artist (376), history subject artist (273), etc.
Performing Arts	507	scenographer (249), filmmaker (74), performance artist (37), musician (36), playwright (28), film director (26), composer (22), etc.
Print&Graphics	8465	printmaker (4221), draftsman (3635), engraver (1939), illustrator (1509), etcher (1045), lithographer (771), photographer (608), etc.
Academic	1339	teacher (593), theorist (142), critic (113), art historian (105), professor (101), archaeologist (56), curator (51), etc.
Other	1907	writer (713), author (426), collector (239), poet (188), patron (184), dealer (70), administrator (37), etc.
FB category	**Total**	**Example Freebase professions (including frequency)**
Performing Arts	24215	Actor (14176), Screenwriter (3608), Film Director (3308), Singer (2205), Film Producer (1798), Musician (1125), Playwright (759), etc.
Creative	13009	Writer (5651), Novelist (1908), Journalist (1625), Composer (1414), Author (1134), Artist (1102), Architect (1098), etc.
Governance	10643	Politician (7123), Lawyer (2915), Judge (662), Diplomat (378), Soldier (274), Priest (102), Social activist (80), etc.
Academic	8816	Mathematician (1320), Physician (1241), Philosopher (1027), Scientist (991), Physicist (961), Chemist (764), Historian (458), etc.
Sports	7128	Baseball player (4937), American football player (763), Racecar driver (315), Coach (249), Professional boxer (160), Golfer (158), Tennis player (146), etc.
Business/Industry/Travel	4196	Engineer (1036), Businessperson (969), Sailor (356), Manager (327), Entrepreneur (275), Aviator (265), Civil engineer (196), etc.

Known Dataset Biases

No historical dataset is free of bias. Here we provide an overview of the biases we encounter in FB, AKL, ULAN, and the Google Ngram English dataset. None of these biases fundamentally questions our results, but it puts them into perspective.

A. Biographical coverage

As visible in Table S1, only 12% to 28% of individuals in FB, AKL, and ULAN have a death location in addition to a birth location. Moreover, birth and death locations are usually the two most frequent locations associated with historical individuals, with the number of locations per individual decaying as a fat tail. In other words, building full life-paths for a large number of individuals remains a persistent challenge, which is why we focus exclusively on birth and death locations.

B. Temporal coverage

In terms of coverage over time, FB fluctuates in waves between 0 and 900 CE (see Fig. 1A). These fluctuations are a result of changes in geographic focus, where Roman, Chinese, and Arab individuals alternate in prominence. As more and more individuals come in, more and more geographic foci become concurrent and eventually wash out such fluctuations. AKL mostly covers artists, with little data before 1200 CE, when artists in Western Europe start to claim authorship with signatures (Nicholas of Verdun in 1189 CE being a famous example). The emancipation of artists leads to increasing documentation of birth and death information. The void of data between 300 and 600 CE, which becomes visible in Fig. 1A, is a classical bias of scholarship in Western Art History, where the early Roman Empire, certain periods of Byzantium, Carolingian art, and Ottonic art get more attention than other periods due to the nature of source material. ULAN mostly covers artists, and is subject to very similar biases as AKL.

C. Spatial coverage

In terms of coverage in space, it turns out that our geographic bias is similar to that in the Human Mortality Database (HMD) (*33*), a standard demographic source that also focuses on Western countries and Japan. In Figure S1 the relative spatial bias of FB,

AKL, and ULAN becomes visible, comparing our data with the total amount of migration within and between 15 world regions in 2005 (*9*).

D. Curated versus crowd sourced data

In figures 1A and S3 we find that the partially crowd sourced FB covers more individuals up to the present, while the professionally curated AKL and ULAN cover relatively more individuals in the past. This indicates the complementary value of crowd sourced and curated data.

E. Still alive bias

FB, AKL, and ULAN are subject to a perceivable bias due to an increasing number of notable individuals still alive. We have marked the onset of this bias, currently approximately in 1930 CE, in figures 1A, S13A-F, and S17. Due to this bias, some locations appear of relatively low weight, as in the case of present-day Istanbul, which grew from a population of 680,000 in 1927 to 14 million in 2014.

F. Place-aggregation, location name changes, and spelling variants

Practitioners in demography and urban scaling, in particular, may wonder why we do not aggregate or reconcile locations to places. There are three reasons: Aggregating locations to places would (a) hide interesting ambiguities in location labels (Tiflis vs. Tbilisi), (b) introduce bias by sticking to an aggregation level that reflects a historical state, disregarding dynamics for which data is hard to obtain for many different countries (for e.g., even in case of wildly changing county boundaries in the US), and (c) hamper further research as the datasets continue to grow and aggregation would have to be done again and again on a growing number of locations.

Reconciling location name changes (Constantinople vs. Istanbul) and spelling variants (Tiflis vs. Tbilisi) would certainly introduce a bias by changing the frequency distribution for large, well-known, and therefore easy-to-reconcile places, versus small, less-known, and therefore hard-to-reconcile places (cf. fig. S10). By avoiding reconciliation, we embrace the "gradient of uncertainty" in historical sources as opposed to threshold against it. As a consequence, we enable further investigation into the relative importance of *toponyms* over time, using our macroscope approach.

AKL contains "uncertain variant names" ("Moskau?"), whose reconciliation would not change the general frequency distribution of "main names" ("Moskau") (r = 0.99). Even though the presence of these "uncertain variant names" in AKL affects our analysis in a minor way (cf. fig. S10), we decided to keep them explicit, as they are hard won assets of historical research that allow for further investigation using our macroscope approach.

Beyond that, we hardly encounter bias due to a lack of aggregation to places or reconciliation of location names. Besides exceptionally interesting historical name changes, most variant names for locations are way less frequent than the main name. For birth-death distance there is no major bias, as the results for FB and ULAN are essentially the same, even though country *toponyms* are much more prominent in FB (cf. Fig. S13D/F). On the upside, we can look at several levels of resolution without re-aggregating. Zooming into large places such as Paris or London, we find that the birth-death-attractor pattern remains valid down to the level of city-quarters, while the main location point for these places will indicate the general attractiveness. This level of granularity is visible for London and Paris in Figs. S5/6, indicating a potential for further applications using our macroscope approach.

G. Dataset language bias

The *Google Ngrams English* dataset is based on English literature. It is therefore no surprise that the Treaty of Paris in 1763 appears as the strongest signal in Fig. 4A. Using the Google Ngrams French or German datasets, the French Revolution in 1789 would emerge as the strongest signal, which makes sense from a French or German perspective. This difference indicates that Ngram trajectories in different languages can be used to measure mutual attention or ignorance towards historical events. A similar bias can be expected in FB and ULAN, both curated in English, and AKL, which is edited in German. A logic next step would be to engage in a comparative analysis of further datasets in different languages, striving for a world-wide coverage of the birth-death macroscope.

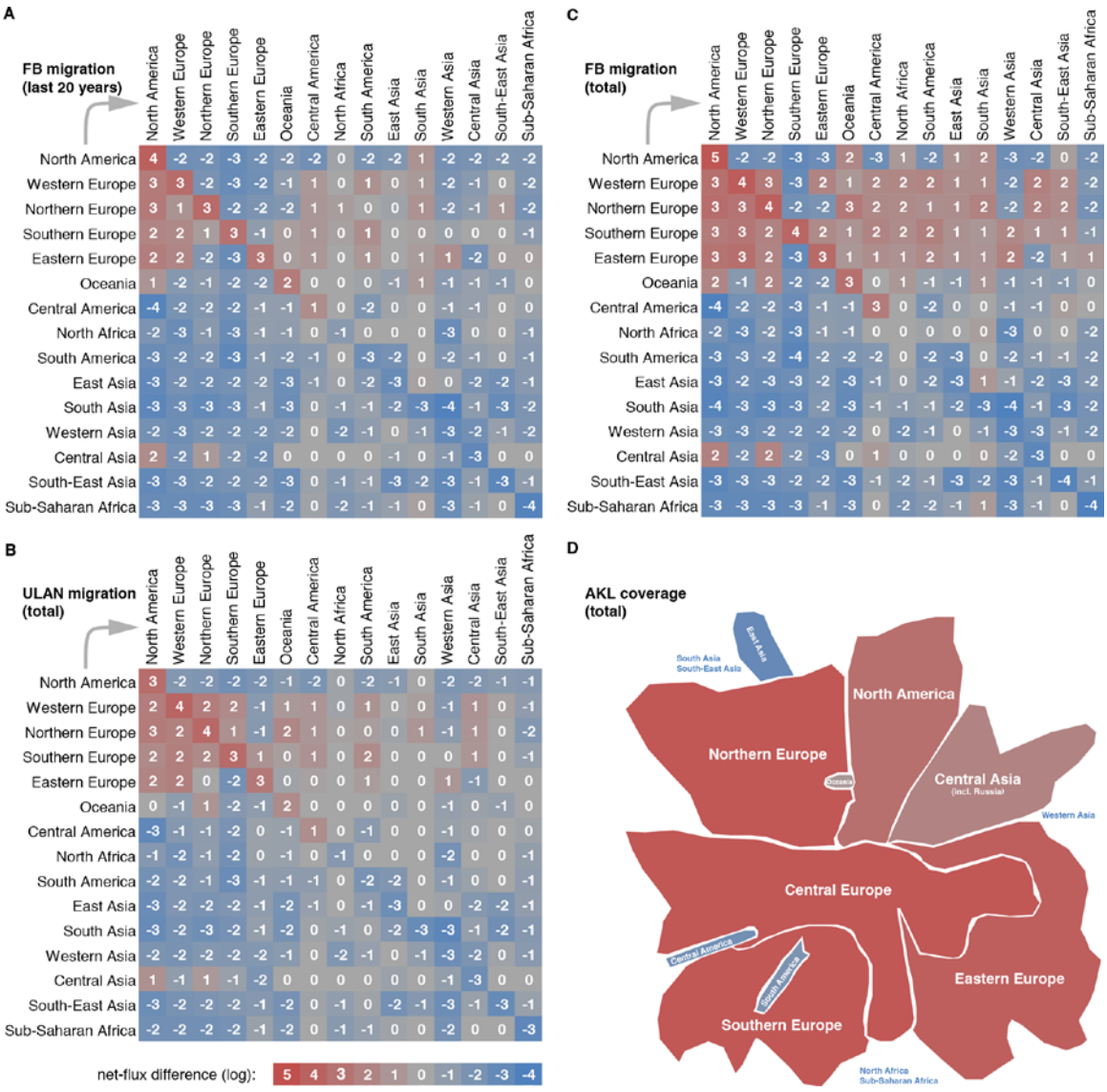

Fig. S1. Migration bias in FB, ULAN, and coverage bias in AKL. To map spatial bias we normalize the net migration flux from birth to death in our datasets against migration within and between 15 world regions in 2005 (*9*), for (**A**) deaths in *FB (last 20 years)*, (**B**) *FB total*, and (**C**) *ULAN total*. We map the net flux difference, given in orders of magnitude to indicate over-, neutral, and under-representation of migration flux in our data (red-grey-blue). For (**D**) AKL, where we lack the geocodes necessary to assign world regions to all locations, we map the coverage of regions, based on the visible clusters in Fig. S8, including information on the disconnected components. As expected, our datasets focus on migration within and out of Europe and North America. Underrepresented migration in our data includes notable individuals and artists moving from Central to North America, South America to Southern Europe, and in particular within Asia and Africa. This mapping could be used to suggest funding efforts to cover art and cultural research in underrepresented areas.

Historical World Population

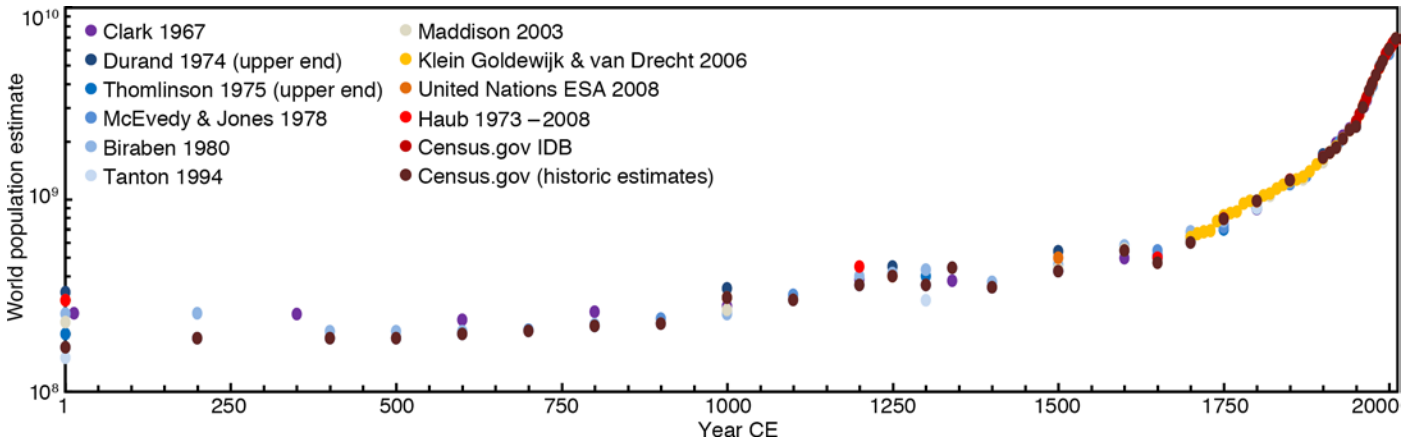

Fig. S2. World population estimates from 1 to 2012 CE. The commonly used sources of historical world population (*34-51*) largely agree on a super-exponential growth curve (*52*), while data points are sparse before 1700 CE.

Demographic Heat Maps

Demographic heat maps based on birth-death data expand modern demography. While standard demographic sources, such as the Human Mortality Database (HMD) (*33*), cover a time frame of several decades for most countries, our birth-death datasets reproduce essential patterns, while adding up to five centuries of information. Our demographic heat maps in Figure S4 are simplified versions of Lexis surfaces (*10*) or shaded contour maps (*53*).

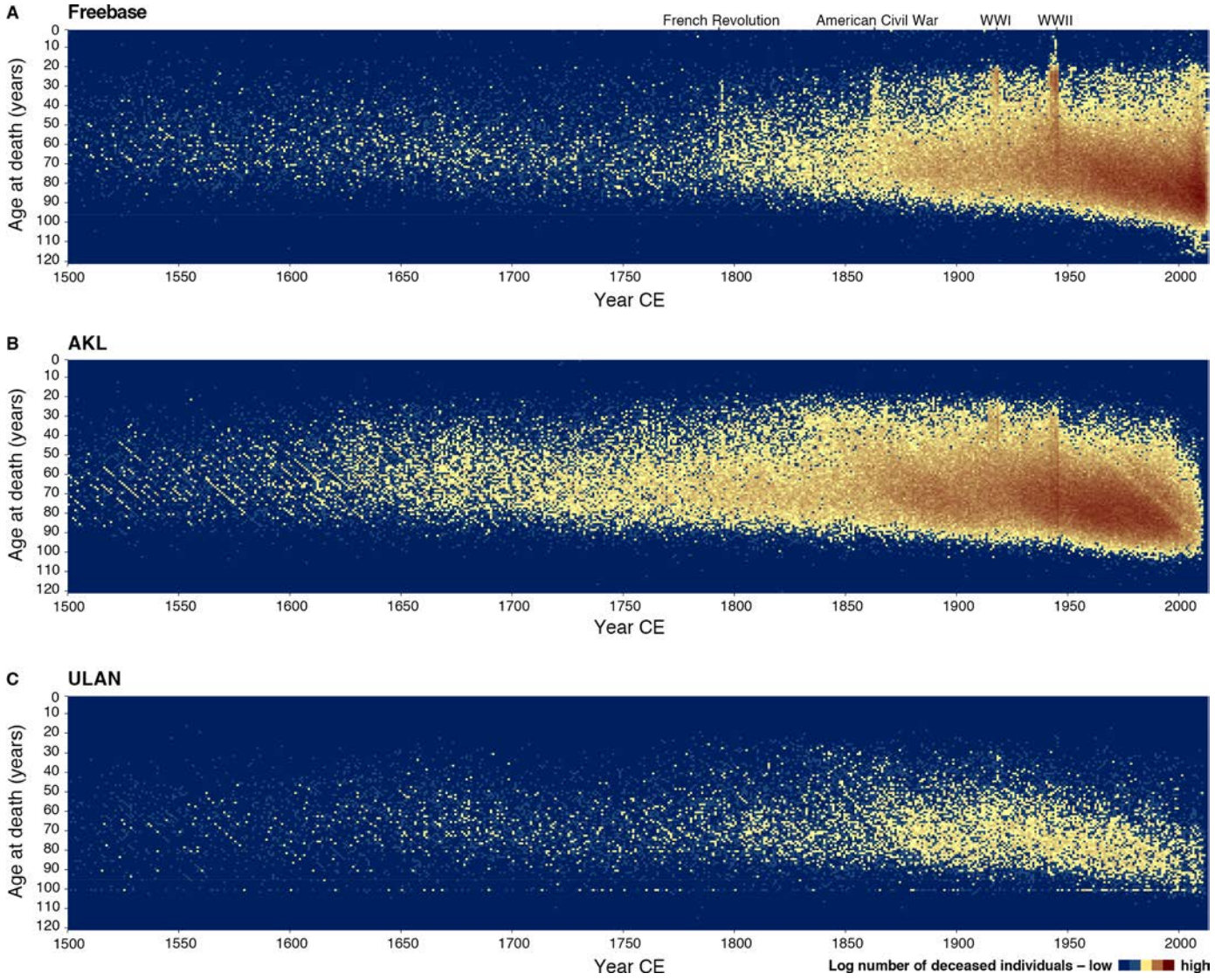

Fig. S3 Demographic heat maps for FB, AKL, and ULAN from 1500 to 2012 CE. Time on the x-axis is plotted against death age on the y-axis, with the log number of deceased individuals given in color from low (blue) to high (red). **(A)**, FB data reveals spikes of premature deaths as a result of crisis and war, such as the French Revolution, the American Civil War, World War I, or World War II. FB has exceptional coverage until present time, indicating a difference between crowd-sourced and expert data curation. **(B)**, AKL data has more coverage in the past, while fading off in the last decades of the 20th century. Light-colored diagonals, best visible during the 20th century, indicate "lost generations" due to potential parents dying in crisis or war. **(C)**, ULAN has substantially less data but is consistent with the general pattern. The over-expression at age 100 in ULAN is an artifact due to a data entry guideline that improves local queries.

Birth-Death Imbalance

In Figures 1C and S4 we plot the number of births versus deaths for locations in FB, AKL, and ULAN, differentiating death-attractors (red) from birth-sources (blue). Deviations from the balanced diagonal are measured as multiples m_e of statistical error e, which is calculated using the equation:

$$e = \begin{cases} -\left(\dfrac{(b-d)}{\sqrt{b}}\right) & for \quad b > d \\ 0 & for \quad b = d \\ \left(\dfrac{(d-b)}{\sqrt{d}}\right) & for \quad b < d \end{cases} \qquad (1)$$

where b is the number of births and d is the number of deaths in a location.

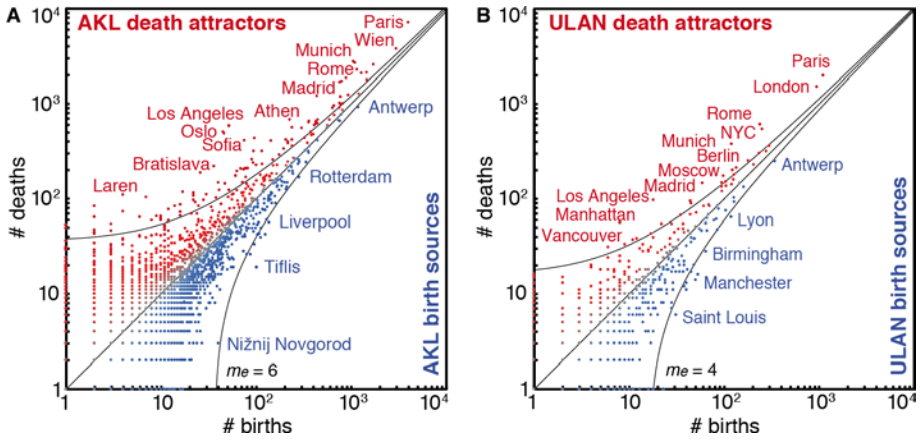

Fig. S4. Birth-death scatter plots with birth sources (blue) and death attractors (red) for locations in **(A)** AKL and **(B)** ULAN cumulated over all time. As in FB (cf. Fig. 1C), deviations from the balanced diagonal, measured as multiples m_e of statistical error e, span up to four orders of magnitude; in AKL and ULAN too, significant outliers host multiple times more deaths than births. Note that outliers are highly significant, even if the numbers of birth and death are low: 'Laren', for example, is a small but highly affluent location in the Netherlands; strong bias towards births in 'Tiflis' on the other hand points to a Russian exodus from 'Tbilisi', Georgia (both location labels are present in AKL, with the latter being located closer to the diagonal).

See Fig. S13 below, for dynamics and further quantification of significance.

Birth to Death Migration

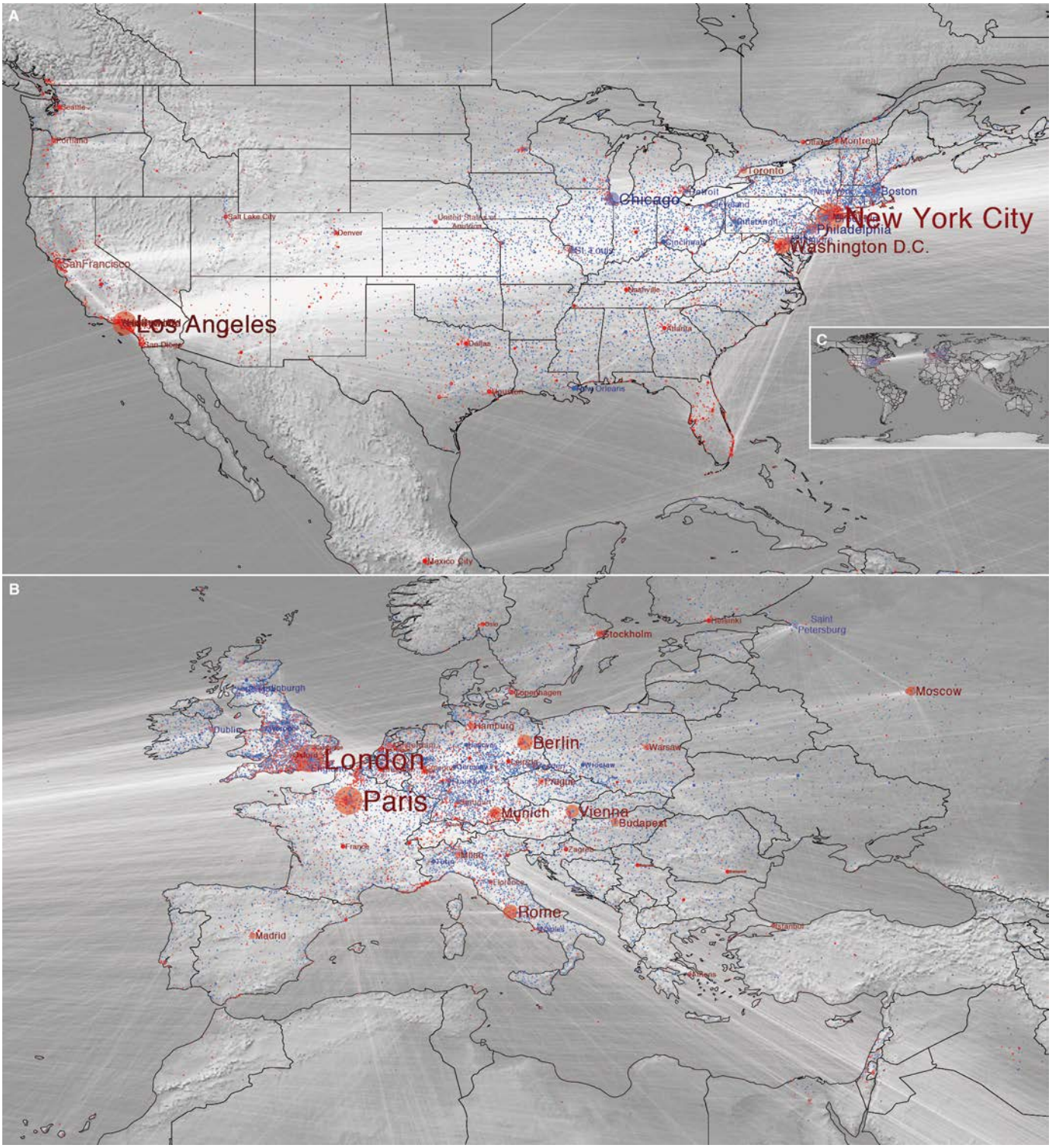

Fig. S5. Birth to death migration in FB, cumulated over all time to 2012 CE: (A), North America; and **(B)**, Europe, and **(C)** worldwide (cf. Figs. 1E and S6/7).

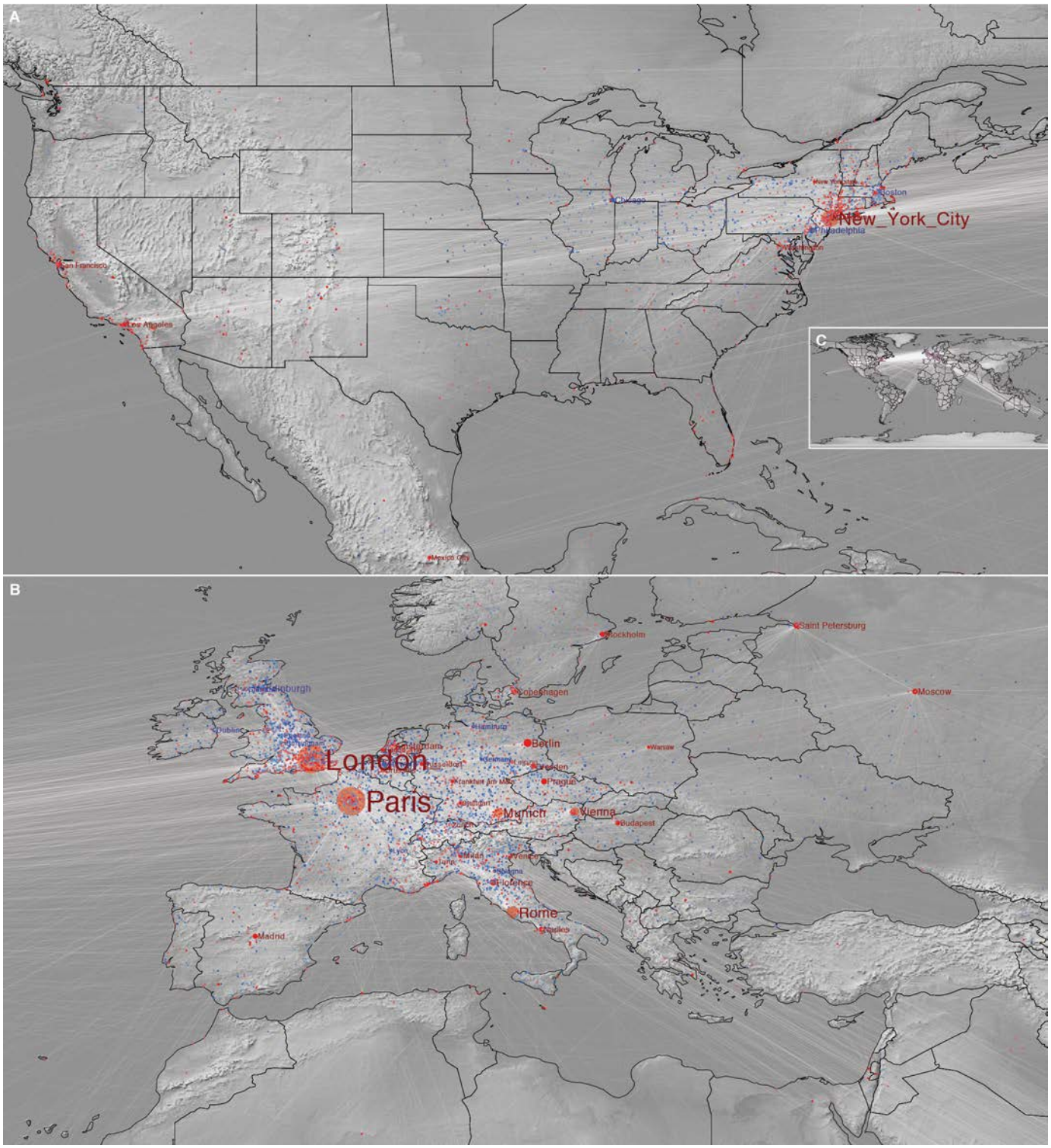

Fig. S6. Birth to death migration in ULAN, cumulated over all time to 2012 CE: (A), North America; **(B)**, Europe, and **(C)** worldwide (cf. Figs. 1E and S5/7).

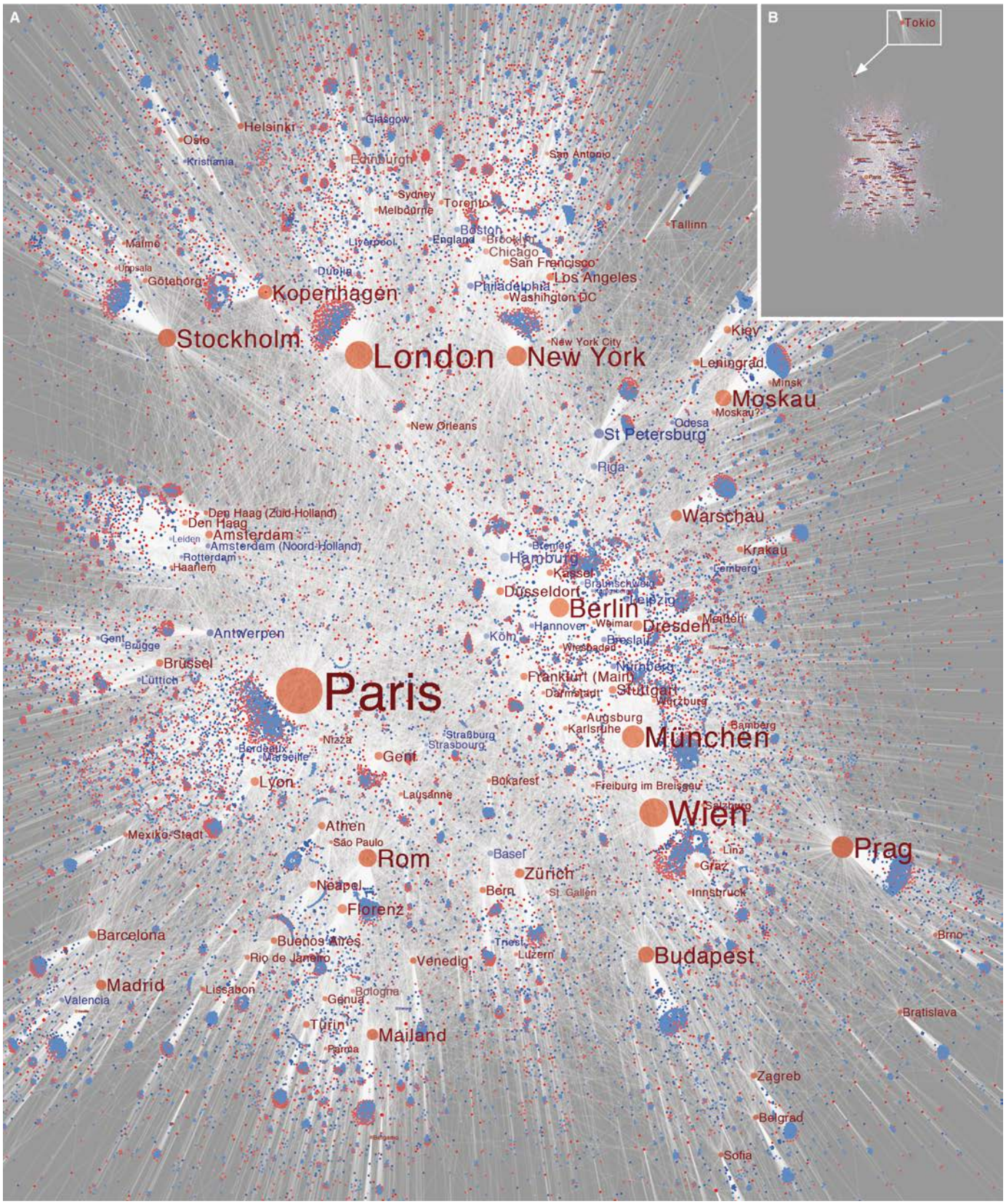

Fig. S7. Birth to death migration in AKL, cumulated over all time to 2012 CE. Without geocodes, cultural proximity emerges using a spring-embedded graph layout (*54*). **(A)**, Europe and North America dominate the global cluster; **(B)**, Japan attaches to the cluster as a tendril (cf. Figs. 1E and S5/6).

PageRank versus Eigenvector Centrality

We chose to use PageRank in our static maps in Figs. S5-S7 over other eigenvalue-based centralities, as PageRank is (1) well-defined in directed networks, (2) not sensitive to the outgoing edges of hubs, and (3) does not have the localization problem that can be found in other eigenvector-based centralities.

Directionality: Eigenvector centrality has problems when applied to directed graphs. For instance, it is well known that all nodes in the in-component of a directed graph will have the value of zero. More precisely, non-zero eigenvector centrality can be assigned only to the nodes "in a strongly connected component of two or more vertices, or the out-component of such a component" (*55* p. 172). Katz centrality (*56*) solves this problem by assigning a small centrality value to every node. However, it is still less useful than PageRank, as it fully transfers node importance to each of its out-neighbors. For instance, if we adopt Katz centrality (or eigenvector centrality), each out-neighbor of Paris will get a huge boost in their centrality and a small town near a hub may become more important, according to the centrality measure, relative to other cultural centers (*55* p. 175). PageRank addresses this problem by 'diluting' the influence of a hub into its outgoing edges.

Strong bias towards hubs (localization transition): Another reason why we have used PageRank is the existence of localization in the eigenvector centrality. As shown in a recent paper (*57*), even under a common condition, eigenvector centrality may fail to assign appropriate centrality to all nodes, as the hubs and their neighbors consume a non-vanishing fraction of the eigenvector centrality while that of the other nodes converges to zero (~ O(1/n)). In other words, the eigenvector centrality may ignore a large fraction of a network when there are large hubs. This indeed happens if we apply Eigenvector Centrality to FB (34%), AKL (41%), and ULAN (42%).

Modeling Statistical Regularities

Heaps' Law

We expect the total number of notable individuals $N(t)$ in our birth-death data to grow exponentially

$$N(t) \sim \exp(rt), \tag{2}$$

with a given growth rate r. In Figure 1A we observe that r is usually stable over several centuries. The number of birth or death locations for notable individuals $S(t)$ also grows exponentially

$$S(t) \sim \exp(st), \tag{3}$$

which implies an underlying Heap's law $S(t) \sim N(t)^{\alpha}$, where $\alpha = s/r \approx 0.9$ (see Figs. 2A and S8).

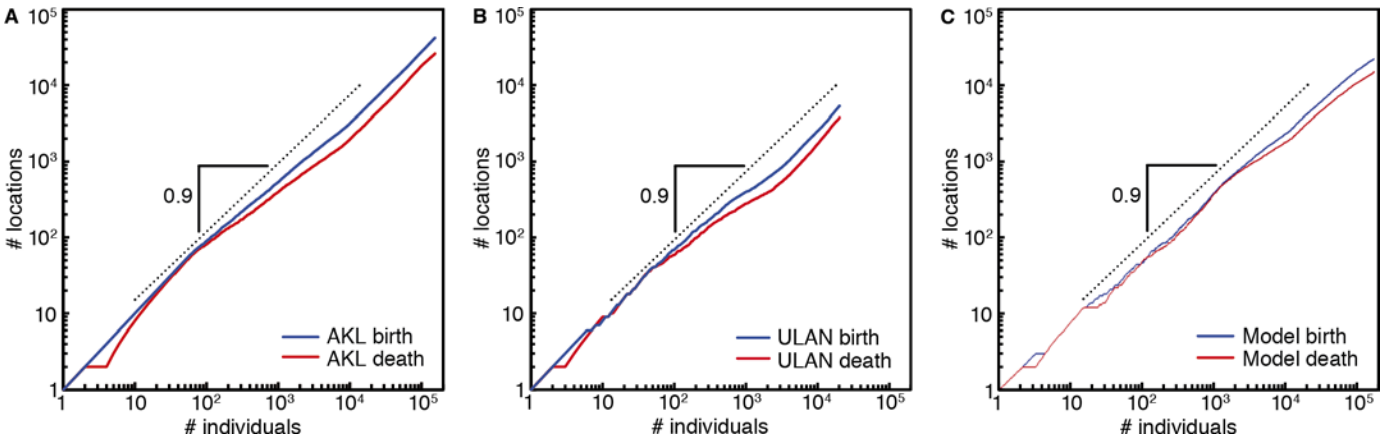

Fig. S8. The number of birth and death locations *S(t)* over the number of individuals *N(t)* for (**A**) AKL and (**B**) ULAN cumulating over all time, and (**C**) our generalized mobility model (see below) (cf. Fig. 2A).

Zipf's Law

The probabilities f^B and f^D of birth and death of an individual in the k^{th} most frequent location follows Zipf's law

$$f_k^B \sim k^{-\zeta_B} \text{ and } f_k^D \sim k^{-\zeta_D}, \tag{4}$$

where $\zeta_B \approx 0.9 \pm 0.1$ and $\zeta_D \approx 1.1 \pm 0.1$. Fig. S11 presents the respective Zipf plots for FB, AKL, ULAN, and our generalized mobility model (see below). The cumulative plots in Figs. S12 are equivalent and include precise quantifications of the slopes (Fig. 2C and S12G-J). Note that the CDF slopes correspond to the Zipf slopes plus 1.

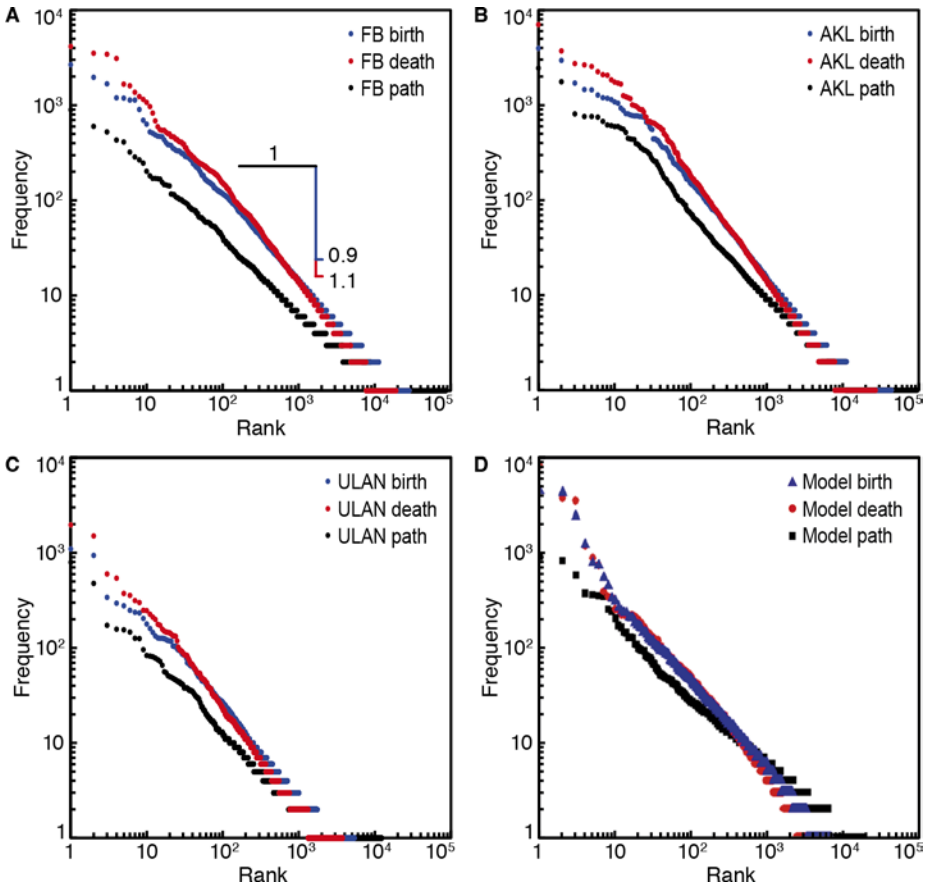

Fig. S9. Scaling patterns for birth, death, and birth-death paths. (**A-D**) Zipf plots for FB, AKL, ULAN, and our generalized mobility model, equivalent to Fig. S12, here cumulated over all time. The plots show the respective rank frequencies f_k of the k-th most frequent birth location (blue), death location (red), and birth to death link (black).

Growth of Individual Locations

To model birth and death population growth for individual locations, we denote $B_i(t)$ and $D_i(t)$ as the total number of respective births and deaths in a location i. Inspired from the fact that the total population grows exponentially (Eq. 2), we may write

$$\begin{aligned} B_i(t) &\sim \exp(\beta_i(t)) \\ D_i(t) &\sim \exp(\delta_i(t)) \end{aligned}, \qquad (5)$$

where $\beta_i(t)$ and $\delta_i(t)$ represents the growth rate of the respective birth and death processes. In Fig. S17, we plot $B(t)$ and $D(t)$ for seven exemplary locations in FB, finding that $\beta_i(t)$ and $\delta_i(t)$ are approximately linear with time t over stretches of several centuries.

Network Model

We model the dynamics of birth to death migration as a network, where locations are the nodes, and notable individuals contribute to the link weight of their birth and death locations over time. Following the work in (*58*), our model includes two main mechanisms governing the growth of the birth–death network: (a) nodes join the network with rate dS/dt and (b) new internal links are generated with rate $d(N\text{-}S)/dt$. More precisely, we have following differential equations

$$\begin{aligned} \frac{dB_i}{dt} &= \Pi_i^B \frac{dS}{dt} + \sum_j \Pi_{i\to j}^I \frac{d(N-S)}{dt} \\ \frac{dD_i}{dt} &= \Pi_i^D \frac{dS}{dt} + \sum_j \Pi_{j\to i}^I \frac{d(N-S)}{dt}, \\ \frac{dW_{i\to j}}{dt} &= \Pi_{i\to j}^I \frac{d(N-S)}{dt} \end{aligned} \tag{6}$$

where the probability Π_i^B, Π_i^D and $\Pi_{i\to j}^I$ satisfy the preferential attachment principle

$$\begin{aligned} &\Pi_i^B(t) \propto B_i(t), \Pi_i^D(t) \propto D_i(t) \\ &\Pi_{i\to j}^I(t) \propto B_i(t) D_j(t) \end{aligned}. \tag{7}$$

Denoting $\sum_j \Pi_{i\to j}^I(t) = \eta_i^B(t) B_i(t) / \sum_i \eta_i^B(t) B_i(t)$ and $\sum_j \Pi_{j\to i}^I(t) = \eta_i^D(t) D_i(t) / \sum_i \eta_i^D(t) D_i(t)$, where $\eta_i^B(t)$ and $\eta_i^D(t)$ are a time-dependent fitness of the place *i*, our model (6-7) predicts the following asymptotical solutions:

$$\begin{aligned} \beta_i(t) &\propto \int_{t_i}^t \eta_i^B(t')dt' \\ \delta_i(t) &\propto \int_{t_i}^t \eta_i^D(t')dt' \end{aligned}, \tag{8}$$

revealing a surprising connection between the growth rate and the fitness function. The fact that $\beta_i(t)$ and $\delta_i(t)$ grow linearly with time indicates that $\eta_i^B(t)$ and $\eta_i^D(t)$ are

nearly constant under time evolution. The Zipf's law of f^B and f^D, or equivalently the fat tail nature of the population distribution $P(B)$ and $P(D)$ can be calculated from the fitness distribution $\rho(\eta)$, which can be measured directly from the empirical data.

For Heaps and Zipf distributions resulting from our model see Figs. S8C and S9D. The fitness measure is used in the death rate trajectories in Figs. 4 and S16.

Global Scaling Dynamics

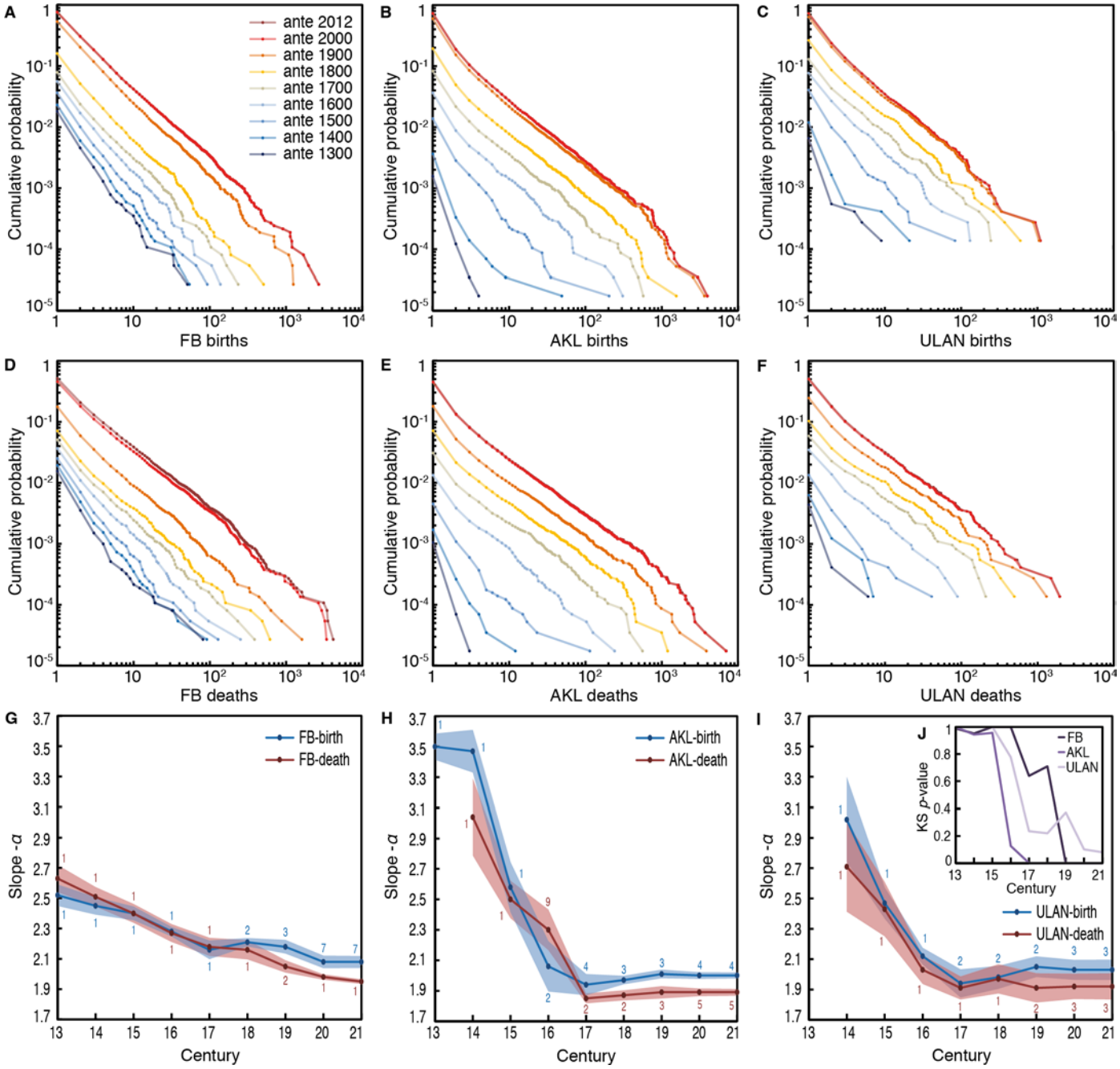

Fig. S10. Cumulative probability plots tracking global scaling dynamics. (A-C) Births and **(D-F)** deaths in FB, AKL, and ULAN *ante* (meaning up to) 1300 to 2012 CE**.** The plots are normalized over final number of locations in the system, indicating that the nature of the distributions for birth and death is consistent over many centuries. (**G-I**) Evolution of power-law slopes over centuries for births and deaths in FB, AKL, and ULAN, fitted using the method in (*18*). The uncertainty interval indicates goodness-of-fit, while the numbers associated to the data points provide the respective x-min value. Note that birth and death scale significantly different in FB from the 19th and in AKL from the

17th/18th century onwards, while ULAN has less data but confirms the trend. (**J**) A pairwise Kolmogorov-Smirnov KS test, comparing birth and death distributions directly, confirms this result, as the *p*-value falls below 0.01 for FB in the 19th century, for AKL in the 17th century and to 12% for ULAN in the 20th.

Note that we do not fit *q*-exponentials, as found for small city sizes and used for tracking dynamics in urban scaling literature (*25, 59, 60*), because small and large locations in our datasets do not scale along exponential curves with power-law tails (as in *25* based on data from *59*). In fact, the *q*-exponential nature of cities in the urban scaling literature is likely to be an effect of the aggregation of *locations* to *places*, where small locations are more likely to be integrated into mid-size and larger locations. Indeed, the AKL dataset (Fig. S10B/E) shows the opposite effect for small locations with one or two births or deaths due to the presence of uncertain or ambiguous location labels, such as "Moskau(?)" in addition to "Moskau". This systematic effect of *unknown* or *uncertain* nodes changing distribution slopes systematically is a known phenomenon in cultural databases (*32*). Trying to clean such systematic properties in large datasets is likely to create more bias due to the inherent gradient of uncertainty; not cleaning such uncertainties exposes the existing bias as in the case of AKL in comparison to FB (see small numbers of births in Fig. S10A/B). It also serves to note that such uncertainties in cultural data collections are hard-won assets, subject to editorial guidelines, that should not be removed before analysis. During data preparation, we made sure that the resulting biases are systematic and do not influence our results in a negative way (see the section on known biases above).

Birth to Death Distances

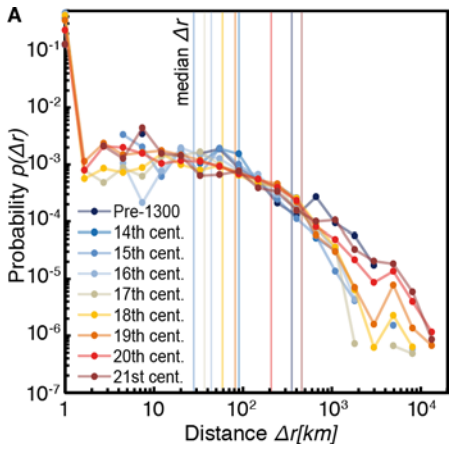

Fig. S11. The probability of birth to death distances *Δr* over several centuries in ULAN from before 1300 to 2012 CE. As in the case of FB (Fig. 2C), the distribution changes very little over time and reveals a clear preference for shorter-range migration on the order of zero to hundreds of kilometers. As expected for a substantially smaller dataset, the median distance (vertical lines) in ULAN fluctuates over a wider range than in FB. For AKL we did not calculate distances due to missing geocodes.

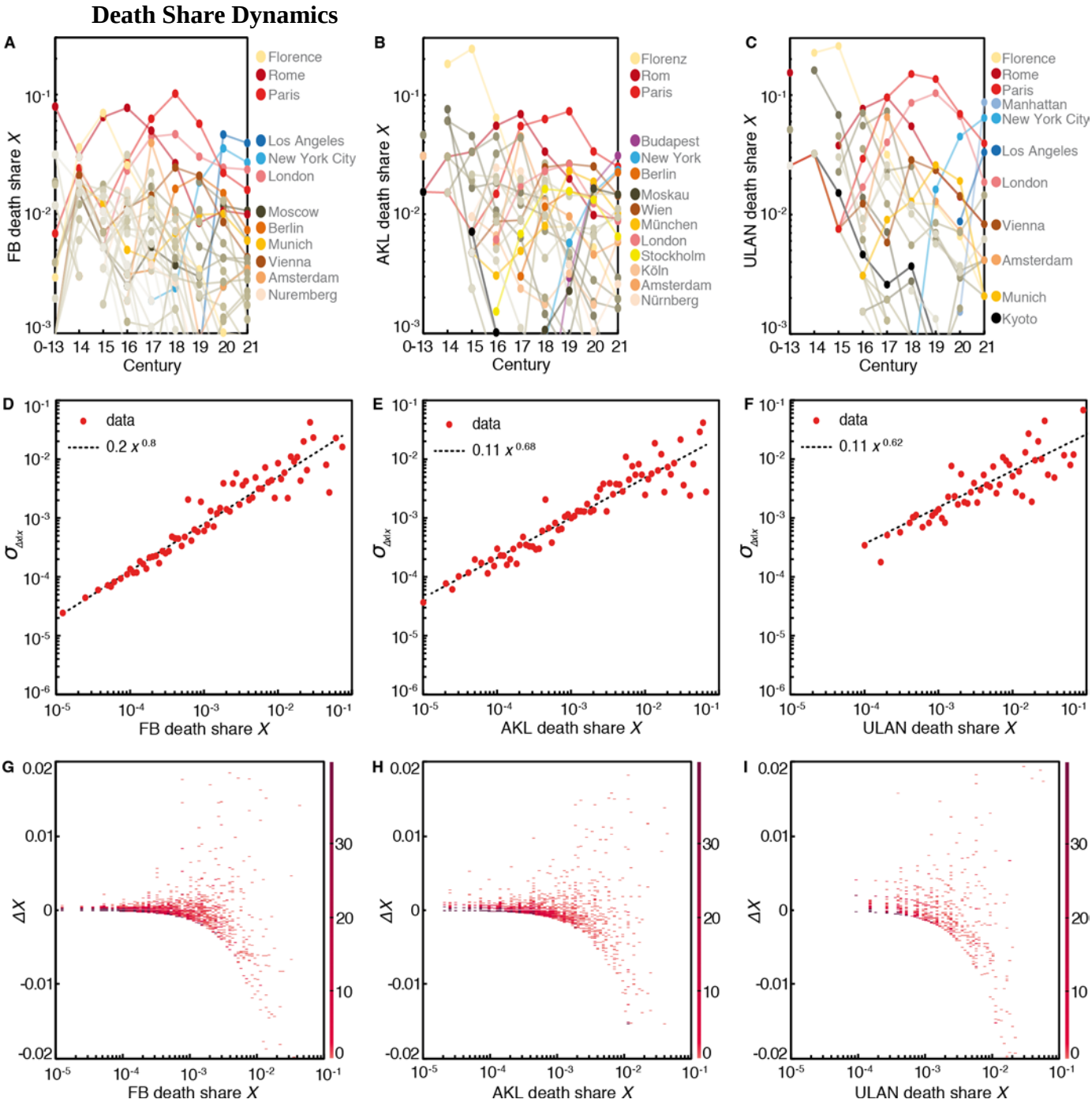

Fig. S12. Death-share dynamics for locations in FB, AKL, and ULAN indicate substantial instabilities of ranking in the system due to noise (cf. *26*). (**A-C**) Relative death-share and consequently rank of major FB, AKL, and ULAN locations fluctuates over centuries. Cross-dataset correlations, such as visible for Florence, Rome, and Paris point to stable local specifics that need to be addressed by qualitative scholarship. (**D-F**) The corrected sample standard deviation of death-share change $\sigma_{\Delta x|x}$ in a function of death-share x for deaths in FB, AKL, and ULAN locations. The linear regression exponents 0.62 to 0.9 < 1 indicate that locations with higher market share tend to be more stable, while the coefficients 0.11 to 0.2 point to strong general instability in both death-

share and rank. (**G-I**) Surface plots of death-share-change Δx in a function of death-share x, indicating that Δx fluctuates asymmetrically for high x, as typical for complex systems with substantial rank instabilities due to noise (cf. *26* for other systems).

Birth-Death Imbalance Dynamics

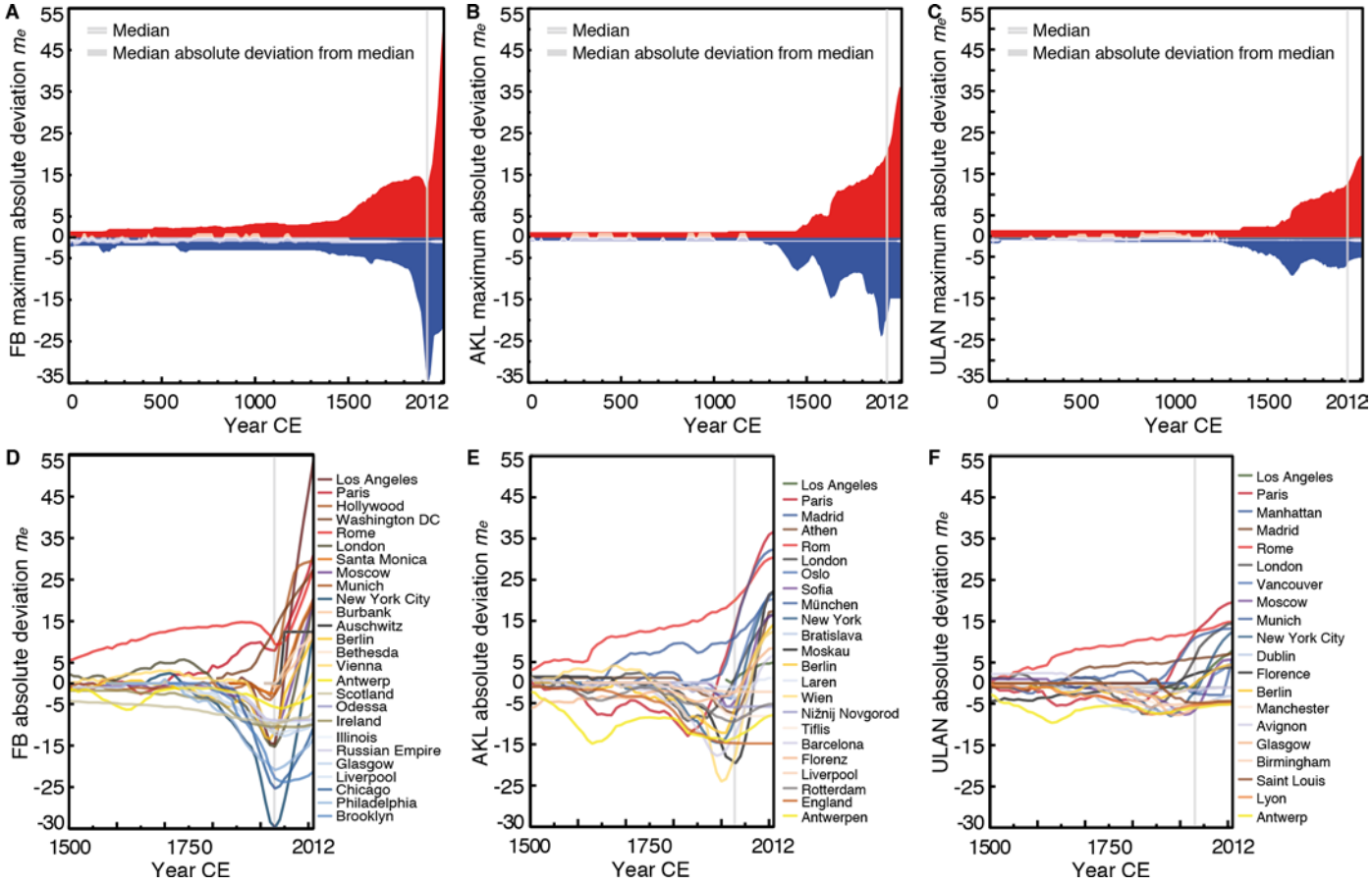

Fig. S13. Birth-death deviation dynamics over two millennia in FB, AKL, and ULAN. (A-C) The maximum absolute deviation, measured as multiples m_e of the statistical error e (cf. the birth death imbalance section above), grows with dataset size from 0 to 2012 CE. The median deviation fluctuates around 0 to -1 while the median absolute deviation (MAD) from the median, a robust measure of significance, fluctuates closely (+/–1) around the median. During best dataset coverage the MAD collapses with the median. **(D-F)** Fluctuating birth-death imbalance in FB, AKL, and ULAN locations from 1500 to 2012 CE. The vertical grey line at 1930 indicates the onset of a systematic bias where all locations dip towards birth and then deviate towards death (cf. known biases section above). This bias is due to notable individuals *not dead yet* (with their births not yet recorded) as well as an increase in life expectancy over the 20th century (delaying their birth record even further). As a result of this systematic bias our original Figures 1C and S4 are slightly shifted off the diagonal towards death. The general picture however is not influenced by the systematic bias. We can still observe major deviations towards birth and death during the bias phase in form of consistent differences in slopes between individual locations.

Frame Sequence for North America

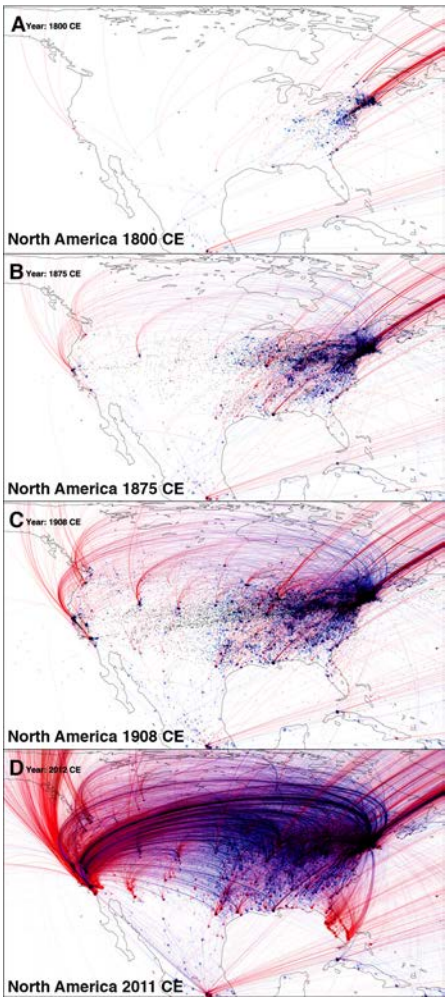

Fig. S14. Sequence of frames, based on movie S2, exemplifying the FB narrative for North America 1630 to 2011 CE. As in movie S1 and Fig. 3A, the dynamically applied color scheme (with black and white inverted in print) denotes birth-death imbalance (blue to red). Here too, individuals appear as particles in the movie, indicating collective directions of flow as they gravitate towards their death locations.

Centralization and Federal Competition

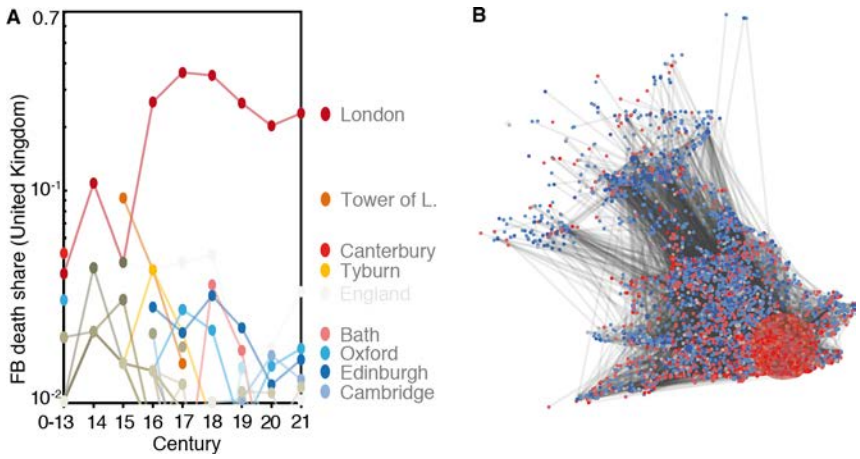

Fig. S15. Centralization and federal competition in the United Kingdom. (**A**) Death-share dynamics in FB locations from before 1300 to 2012 CE, and (**B**) the final network state for locations in the United Kingdom. London clearly emerges as the center of a *winner-takes-all*, while we still find exceptionally high regional cohesion. This special situation can be explained with the social season where individuals move seasonally between the city and the country-side. Note that during the 16th and 17th century, the Tower of London, a prison, and Tyburn, a village now part of London, known for capital punishment, are among the top ranked death locations for notable individuals in the United Kingdom.

Cultural Trajectories

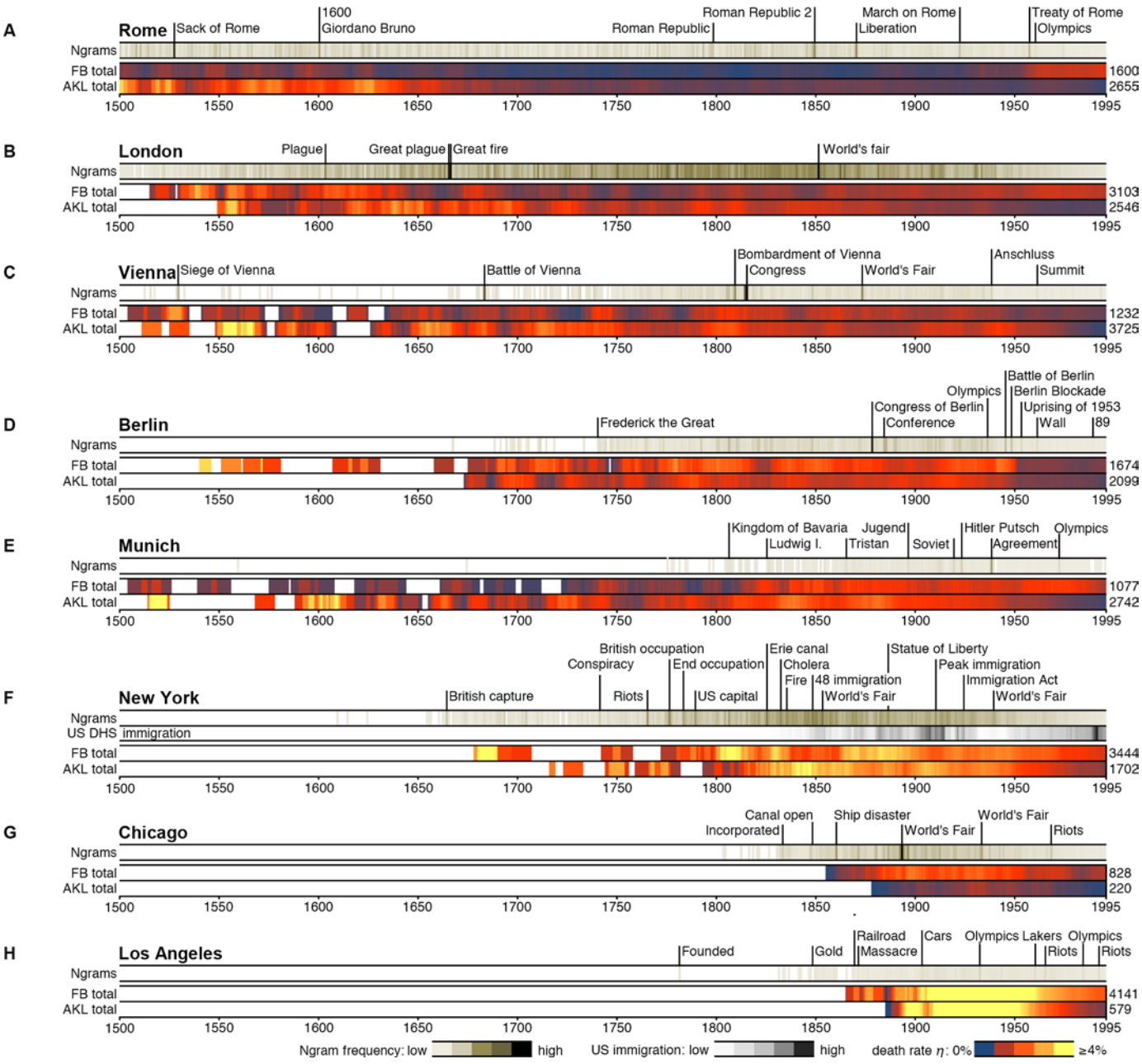

Fig. S16. Cultural trajectories for eight selected cultural centers indicate successive periods of cultural leadership, providing detail for further micro-readings (cf. Fig. S12A-C). For each city we show trajectories based on the *English Google Ngrams*, the *FB total*, and *AKL total* datasets, in correspondence to the Paris trajectories in Figs. 4A-B. Here, we show trajectories for **(A)**, Rome, **(B)**, London, **(C)**, Vienna, **(D)**, Berlin, **(E)**, Munich, **(F)**, New York, **(G)**, Chicago, and **(H)**, Los Angeles. For New York we include US immigration data for comparison (*61*).

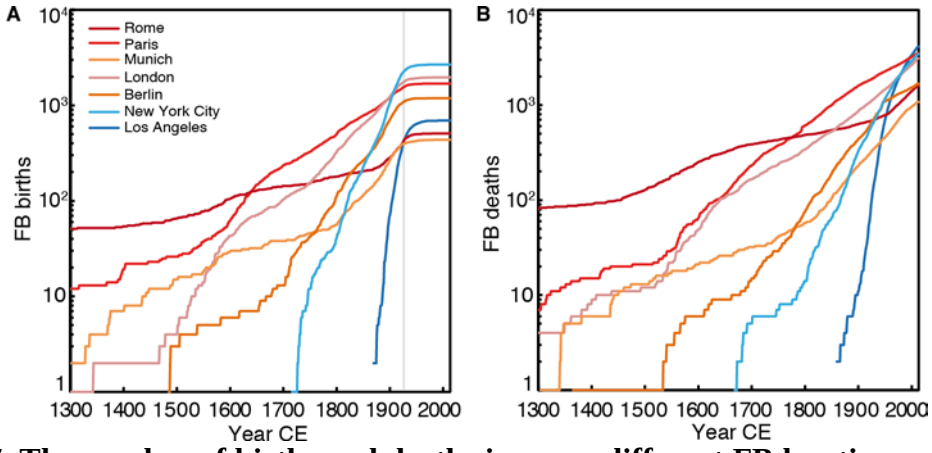

Fig. S17. The number of births and deaths in seven different FB locations over time from 1300 to 2012 CE. Changes in slope, such as for Munich deaths shortly after 1806, indicate a substantial change in fitness, such as becoming the capital of the Bavarian kingdom. The cutoff for births after the grey line at 1920 is due to notable people that are still alive and therefore not recorded yet.

Supplementary Movies

Movie S1. European birth-death network dynamics, offering a meta-narrative of cultural history based on the FB dataset, 0 to 2012 CE (67 seconds). The dynamically applied color scheme indicates birth sources (blue) and death attractors (red) in correspondence to Fig. 1C. Individuals in the videos appear as particles gravitating towards their death locations, indicating collective directions of flow. The video is rendered with one frame per year at 30 frames per second. Further characterization of the movie content is given in the text and Fig. 3A.

Movie S2. North American birth-death network dynamics, offering a meta-narrative of cultural history based on the FB dataset, 1620 to 2012 CE (13 seconds). As in Movie S1, the dynamically applied color scheme indicates birth sources (blue) and death attractors (red) in correspondence to Fig. 1C. Individuals in the videos appear as particles gravitating towards their death locations, indicating collective directions of flow. The video is rendered with one frame per year at 30 frames per second.

Additional Data Tables

Additional Data table S1 (separate file)
SchichDataS1_FB.xlsx

Additional Data table S2 (separate file)
SchichDataS2_AKL.xlsx

Additional Data table S3 (separate file)
SchichDataS3_ULAN.xlsx

Additional Data table S4 (separate file)
SchichDataS4_WCEN.xlsx

References and Notes

1. R. L. Carneiro, *The Muse of History and the Science of Culture* (Springer, New York, 2000).

2. L. Spinney, Human cycles: History as science. *Nature* **488**, 24–26 (2012). Medline doi:10.1038/488024a

3. Freebase.com: A community–curated database of well-known people, places, and things (Google, Mountain View, CA, 2011); http://www.freebase.com.

4. A. Beyer, S. Bénédicte, W. Tegethoff, Eds., *Allgemeines Künstlerlexikon (AKL). Die Bildenden Künstler aller Zeiten und Völker* (De Gruyter, Berlin, 1991, rev. ed. 2010).

5. U. Thieme, F. Becker, Eds., *Allgemeines Lexikon der bildenden Künstler von der Antike bis zur Gegenwart* (Seemann, Leipzig, 1907, rev. ed. 1950).

6. H. Vollmer, Ed., *Allgemeines Lexikon der bildenden Künstler des XX. Jahrhunderts* (Seemann, Leipzig, 1953, rev. eds. 1962 and 1980).

7. Getty Vocabulary Program, *Union List of Artist Names* (The J. Paul Getty Trust, Los Angeles, 2010); www.getty.edu/research/tools/vocabularies/ulan/.

8. Materials and methods are available as supplementary materials on *Science* Online.

9. G. J. Abel, N. Sander, Quantifying global international migration flows. *Science* **343**, 1520–1522 (2014). Medline doi:10.1126/science.1248676

10. N. Keiding, Statistical inference in the Lexis diagram. *Philos. Trans. R. Soc.* **332**, 487–509 (1990). doi:10.1098/rsta.1990.0128

11. Winckelmann-Gesellschaft Stendal, Eds., Corpus der antiken Denkmäler, die J.J. Winckelmann und seine Zeit kannten [database] (Biering & Brinkmann, München, 2000).

12. S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**, 107–117 (1998). doi:10.1016/S0169-7552(98)00110-X

13. M. Batty, Rank clocks. *Nature* **444**, 592–596 (2006). Medline doi:10.1038/nature05302

14. H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, Waltham, MA, 1978).

15. G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Boston, 1949).

16. L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G. B. West, Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7301–7306 (2007). Medline doi:10.1073/pnas.0610172104

17. L. M. A. Bettencourt, J. Lobo, D. Strumsky, G. B. West, Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PLOS ONE* **5**, e13541 (2010). Medline doi:10.1371/journal.pone.0013541

18. A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009). doi:10.1137/070710111

19. E. G. Ravenstein, The laws of migration. *J. Stat. Soc. Lond.* **48**, 167–235 (1885). doi:10.2307/2979181

20. E. G. Ravenstein, The laws of migration: Second paper. *J. R. Stat. Soc.* **52**, 241–305 (1889). doi:10.2307/2979333

21. G. K. Zipf, The $P_1 P_2/D$ hypothesis: On the intercity movement of persons. *Am. Sociol. Rev.* **11**, 677–686 (1946). doi:10.2307/2087063

22. P. Ren, *Lifetime Mobility in the United States: 2010* (U.S. Census Bureau, U.S. Department of Commerce, Washington, DC, 2011).

23. D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel. *Nature* **439**, 462–465 (2006). Medline doi:10.1038/nature04292

24. C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823 (2010). doi:10.1038/nphys1760

25. D. R. White, T. Laurent, N. Kejzar, in *Globalization as Evolutionary Process: Modeling Global Change*, G. Modelski, T. Devezas, W. R. Thompson, Eds. (Routledge, Milton Park, UK, 2007), pp. 190–225.

26. N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J. P. Bouchaud, A.-L. Barabási, Dynamics of ranking processes in complex systems. *Phys. Rev. Lett.* **109**, 128701 (2012). Medline doi:10.1103/PhysRevLett.109.128701

27. G. Bianconi, A.-L. Barabási, Bose-Einstein condensation in complex networks. *Phys. Rev. Lett.* **86**, 5632–5635 (2001). Medline doi:10.1103/PhysRevLett.86.5632

28. Google Ngram English data set, version 20090715 (Google, Mountain View, 2009); http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

29. J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, E. L. Aiden; Google Books Team, Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182 (2011). Medline doi:10.1126/science.1199644

30. M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. D. Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. Grant, A. R. Jones, An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012). Medline doi:10.1038/nature11405

31. *Getty Thesaurus of Geographic Names* (The J. Paul Getty Trust, Los Angeles, 2000); www.getty.edu/research/tools/vocabularies/tgn/.

32. M. Schich, in *Beautiful Visualization*, J. Steele and N. Iliinsky, Eds. (O'Reilly, Sebastopol, 2010), pp. 227–254.

33. The Human Mortality Database (Univ. of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock, 2000–2012); http://www.mortality.org.

34. C. Clark, *Population Growth and Land Use* (Macmillan, New York, 1967).

35. J. D. Durand, "Historical estimates of world population: An evaluation" (PSC Analytical and Tech. Rep. 10, Population and Study Center, Univ. of Pennsylvania, Philadelphia, 1974); http://repository.upenn.edu/psc_penn_papers/9/.

36. R. Thomlinson, *Demographic Problems: Controversy over Population Control* (Dickenson Publishing Company, Belmont, CA, 1975).

37. C. McEvedy, R. Jones, *Atlas of World Population History* (Penguin, Harmondsworth, UK, 1978).

38. J.-N. Biraben, *An Essay Concerning Mankind's Evolution* (Institut national d'études démographiques, Paris, 1980).

39. J. H. Tanton, End of the migration epoch? Time for a new paradigm. *The Social Contract* **4**, 162–173 (1994).

40. A. Maddison, *The World Economy: Historical Statistics* (OECD Publishing, Paris, 2004).

41. K. Klein Goldewijk, G. van Drecht, in *Integrated Modelling of Global Environmental Change*, A. F. Bouwman, T. Kram, K. Klein Goldewijk, Eds. [Netherlands Environmental Assessment Agency (MNP), Bilthoven, 2006], pp. 93–112.

42. United Nations Department of Economic and Social Affairs Estimates (ESA), *World Population Prospects: The 2008 Revision* (United Nations, New York, 2009).

43. United Nations Department of Economic and Social Affairs Estimates (ESA), "The World at Six Billion" (United Nations, New York, 1999).

44. C. Haub, How many people have ever lived on earth? *Popul. Today* **23**(2), 4–5 (1995). Medline

45. C. Haub, "2005 World population data sheet" (Population Reference Bureau, Washington, DC, 2005).

46. C. Haub, "2006 World population data sheet" (Population Reference Bureau, Washington, DC, 2006).

47. C. Haub, "2007 World population data sheet" (Population Reference Bureau, Washington, DC, 2007).

48. C. Haub, "2008 World population data sheet" (Population Reference Bureau, Washington, DC, 2008).

49. International Data Base (IDB) (U.S. Census Bureau, Washington, DC, 2012); www.census.gov/population/international/data/idb/informationGateway.php.

50. "World population: Historical estimates of world population" (U.S. Census Bureau, Washington, DC, 2012); www.census.gov/population/international/data/worldpop/table_history.php.

51. "World population estimates." Wikipedia (17 March 2013); http://en.wikipedia.org/w/index.php?title=World_population_estimates&oldid=544834734.

52. A. Johansen, D. Sornette, Finite-time singularity in the dynamics of the world population, economic and financial indices. *Physica A* **294**, 465–502 (2001). doi:10.1016/S0378-4371(01)00105-4

53. J. W. Vaupel, W. Zhenglian, K. F. Andreev, A. I. Yashin, *Population Data at a Glance: Shaded Contour Maps of Demographic Surfaces over Age and Time* (Univ. Press of Southern Denmark, Odense, 1998).

54. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE* **9**, e98679 (2014). Medline doi:10.1371/journal.pone.0098679

55. M. E. J. Newman, *Networks: An Introduction* (Oxford Univ. Press, Oxford, 2010).

56. P. Bonacich, Some unique properties of eigenvector centrality. *Soc. Networks* **29**, 555–564 (2007). doi:10.1016/j.socnet.2007.04.002

57. T. Martin, X. Zhang, M. E. J. Newman, "Localization and centrality in networks," arXiv 1401.5093 (2014).

58. A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations. *Physica A* **311**, 590–614 (2002). doi:10.1016/S0378-4371(02)00736-7

59. T. Chandler, *Four Thousand Years of Urban Growth: An Historical Census* (Edwin Mellon Press, Lewiston, NY, 1987).

60. D. R. White, N. Kejzar, C. Tsallis, C. Rozenblat, "Generative historical model of city-size hierarchies: 430BCE-2005" (ISCOM Working Paper, Institute of Mathematical Behavioral Sciences, Univ. of California, Irvine, CA, 2005).

61. *Yearbook of Immigration Statistics* (U.S. Department of Homeland Security, Washington, DC, 2012); www.dhs.gov/yearbook-immigration-statistics.