Matchmaker, matchmaker, make me a match: migration of populations via marriages in the past

Sang Hoon Lee (이상훈),[1,*] Robyn Ffrancon,[2,*] Daniel M. Abrams,[3] Beom Jun Kim (김범준),[4] and Mason A. Porter[1,5]

[1]*Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, Oxford, OX2 6GG, United Kingdom*
[2]*Department of Physics, University of Gothenburg, Origovägen 6B, 412 96 Gothenburg, Sweden*
[3]*Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois 60208, USA*
[4]*Department of Physics, Sungkyunkwan University, Suwon 440-746, Korea*
[5]*CABDyN Complexity Centre, University of Oxford, Oxford, United Kingdom*

The study of human mobility is both of fundamental importance and of great potential value. For example, it can be leveraged to facilitate efficient city planning and improve prevention strategies when faced with epidemics. The newfound wealth of rich sources of data—including banknote flows, mobile phone records, and transportation data—have led to an explosion of attempts to characterize modern human mobility. Unfortunately, the dearth of comparable historical data makes it much more difficult to study human mobility patterns from the past. In this paper, we present such an analysis: we demonstrate that the data record from Korean family books (called "*jokbo*") can be used to estimate migration patterns via marriages from the past 750 years. We apply two generative models of long-term human mobility to quantify the relevance of geographical information to human marriage records in the data. Using the different geographic distributions of clans, we quantify the "ergodicity" of clans in terms of how widely and uniformly they have spread across Korea, and we compare these results to those obtained using surname data from the Czech Republic. Based on the correlation between ergodicity and migration patterns in Korea, we identify two different types of migration patterns: diffusive and convective.

Keywords: complex systems; sociophysics; human mobility; human migration

Since Quetelet's advocacy of "social physics" in the 1830s [1] and Ravenstein's seminal work later in the nineteenth century [2], quantitative studies of human mobility have suggested that our movements follow statistically predictable patterns [3–10]. Research has focused predominantly on modern mobility—rather than historical mobility and migration—due to the disproportionate availability of large, rich data sets from modern life [11–15]. By contrast, historical data tends to be sparse, incomplete, and noisy; these constraints limit the scope of conclusions that one can draw about how humans mingled, mixed, and migrated over long time scales [16, 17].

In this paper, we investigate *historical* human mobility, and associated human migration, by studying the matchmaking process for traditional marriages in Korea combined with modern census data in South Korea. We obtain our data from Korean "family books" called *jokbo* (족보 in Korean). Such a confluence of historical and modern data is rare, and it allows a novel test of generative models for human mobility.

According to Korean tradition, family names are subdivided into clans called *bon-gwan* (본관), which are identified by a unique place of origin. For example, the two Korean authors of this paper belong to the clans "Kim from Gimhae (김해 김)" and "Lee from Hakseong (학성 이)," and the clan "Lee from Hakseong" is distinct from the clan "Lee from Jeonju (전주 이)" [the royal clan of the Joseon dynasty and the Great Korean Empire (1392–1910)]. When two Koreans marry, the bride's clan is customarily recorded in the *jokbo* owned by the groom's family. These *jokbo* are kept in the groom's family and passed down through the generations, serving primarily as a record of the names and birth years of all male descendants [18]. In previous work, researchers used the marriage data contained in these books to estimate the population sizes and distributions of clans in Korea as far as 750 years in the past [19–21]. Such distributions are useful for understanding quantitative aspects of human culture, and we proceed even further by conducting a systematic investigation of the geographical information embedded in *jokbo*.

We examine a set of ten *jokbo* to try to understand how geographical separation affected human interaction in the past in Korea. Specifically, we examine how inter-clan marriage rates can be predicted by physical distance and how clans themselves have spread across the country during the past several hundred years. To do this, we apply two generative models for describing human mobility patterns to *jokbo* records of past marriages between two clans. Note that the identification of clans with specific geographical origins is not unique to Korea: for example, the origins of British and Czech surnames were also the subject of recent investigations [22, 23].

I. JOKBO DATA SETS

We use the same ten *jokbo* data sets that were employed in Ref. [19–21]. An individual book contains between 1 873 and 104 356 marriage entries, and there are a total of 221 598 entries across all books [see Table S1 and Figs. S3 and S4 in the Supporting Information (SI) [24] for details]. Each entry contains the bride's clan and year of birth [25]; the oldest book has entries that date back to the 13th century.

Previous studies of this data set [19–21] did not use any of the information implicitly encoded in the geographical origins of each clan. This, together with the modern geographical distribution of clans, comprises a key ingredient of our

* These authors contributed equally to this work.

analysis. We convert location names to geographical coordinates using the Google Maps Application Programming Interface (GMAPI) [26]. Because of the much sparser coverage of North Korean regions by Google Maps (see Fig. S7 in the SI), this geolocation data is a biased sample of the full data. However, data for the southern half of the Korean peninsula is rich, and it is sufficient to draw interesting and robust conclusions. For example, the effect of a change in the legality of intraclan marriage in 1997 is clearly observable in the data.

II. MODERN NAME DISTRIBUTIONS

In addition to the *jokbo* data sets, we also use data from two Korean census reports—from 1985 and 2000—to evaluate the current distribution of clans in Korea [27, 28]. As illustrated in Fig. 1, some clans have dispersed rather broadly but others are localized (usually near their place of origin). Drawing on ideas from statistical mechanics [29, 30], we use the term "ergodic" as an analogy to describe clans that have spread broadly throughout Korea. We suppose that such clans have reached a dynamic equilibrium: an ergodic clan is "spread equally" throughout Korea in the sense that one expects it to have roughly the same geographical distribution as the population as a whole. Note that we do not expect an ergodic clan to reach a spatially uniform state for the same reasons that the full population is not spatially uniform (e.g., inhomogeneities in natural resources, advantages to congregating in cities, etc.).

Non-ergodic clans should have rather different distributions from those we dub ergodic, because their distribution must be significantly different from that of the full population. One can construe the notion of ergodicity as a natural extension of other physical analogies that were used in previous quantitative studies (including the original ones) on human migration [1–10]. As we discuss later, we can quantify the extent of clan ergodicity.

III. GENERATIVE MODELS FOR MARRIAGE FLUX

We compute a "marriage flux"—the rate of marriage of women from clan i into clan j—for all clan pairs (i, j) in our data [31]. Historically, professional matchmakers were employed to travel between families to arrange marriages [32], so we posit that physical distance plays a significant role in determining this flux. We examine this hypothesis using two generative models: a conventional gravity model [4, 5] and a recently-developed radiation model [33, 34].

The gravity model has been used to explain phenomena such as commuting patterns and disease spread [35–38]. In this model, the flux of population G_{ij} from a site i to a site j is

$$G_{ij} = \frac{m_i^{\alpha} m_j^{\beta}}{r_{ij}^{\gamma}}\,, \tag{1}$$

where α, β, and γ are adjustable exponents. For our purposes, G_{ij} is proportional to the flux of women from clan i to clan j through marriage. The total population of clan i is given by m_i, and the variable r_{ij} is the distance between the centroids of clans i and j. (We calculate centroids using the spatial distribution for each clan from the census data in 2000 [27].)

Determining the centroid locations of clans from modern census data is more accurate than attempting to locate the origins' names themselves [39]. for two reasons. First, many clans already existed much earlier than the advent of *jokbo*, so the modern centroid locations may be closer to the real centroids at the time of creation of the *jokbo*. Second, the origin-place names for many clans have become outdated and cannot be located accurately via the names of modern administrative regions. For instance, the clan origin 'Hakseong' of the first author is an old name for the city Ulsan in South Korea, but the name 'Hakseong' is currently only used to describe the small administrative region 'Hakseong-dong' in Ulsan. However, as we demonstrate in Fig. 1, using the centroid location of 'Lee from Hakseong' correctly gives the modern city Ulsan. This works in part because Lee from Hakseong is a non-ergodic clan; for ergodic clans such as Kim from Gimhae, the spatial precision is much worse. This is an important observation that we will discuss in detail later.

The radiation model that we examine takes finite-size effects into account [34]. The population flux R_{ij} from clan i to clan j is

$$R_{ij} = \frac{R_i}{1 - m_i/N} \times \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}\,, \tag{2}$$

where $R_i = \sum_j R_{ij}$ is proportional to the total population marrying from clan i into any other clan, N is the total population, and s_{ij} is the exclusive population within a circle of radius r_{ij} centered on the centroid of clan i. Note that members of clans i and j are *not* included in computing s_{ij} [33]. As before, m_i is the population of clan i, members of clan i marry into clan j, and clan j keeps the marriage records.

For both models, we use census data from the year 2000 [27] as a proxy for past population to compute the quantities r_{ij}, m_i, and s_{ij}. This approximation is supported by previously reported estimates of stability in Korean society: historically, clans have grown in parallel with the total population, so the relative sizes of clans have remained nearly constant [21]. In both Eqs. (1) and (2), only the relative sizes m_i/N and s_{ij}/N matter for calculating the flux (up to a constant of proportionality).

IV. HUMAN DIFFUSION AND ERGODICITY

One way to quantify the notion of clan ergodicity is to examine what we call the "clan density anomaly," which describes the local deviation in density of members of a given clan. The clan density anomaly is the dimensionless variable $\phi_i(x, y, t) = c_i - (m_i/N)\rho$ at position (x, y) and time t, where $c_i = c_i(x, y, t)$ is the (spatially and temporally varying) local clan concentration (i.e., the clan population density), $m_i(t)$ is the total clan population, $\rho = \rho(x, y, t)$ is the local population density (i.e., the total population of all clans at point (x, y) and

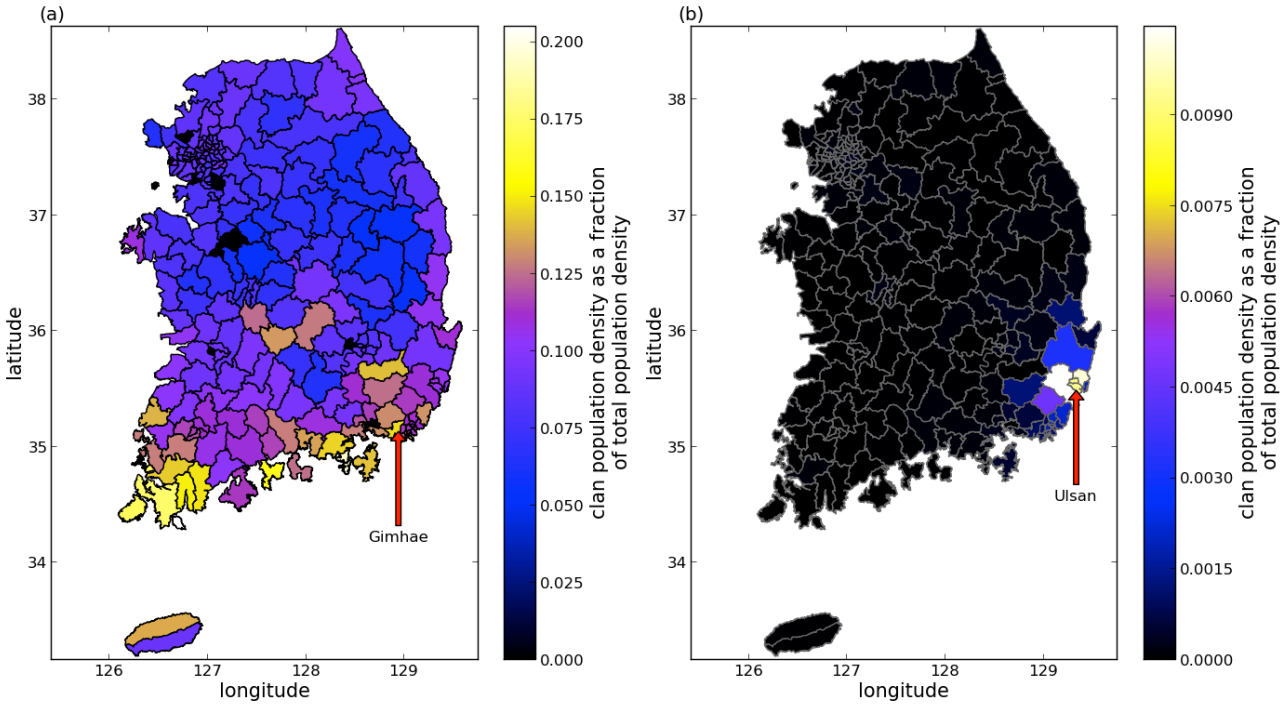

FIG. 1. Examples of (a) ergodic and (b) non-ergodic clans. We color the regions of South Korea based on the fraction of the population composed of members of the clan in the year 2000. We use arrows to indicate the origins of the two clans: Gimhae on the left and Ulsan ("Hakseong" is the old name of the city) on the right. In this map, we use the 2010 administrative boundaries [28]. (See the SI for discussions of data sets and data cleaning.)

time t, divided by the differential area), and $N = N(t)$ is the total population of all the clans at time t. If a clan were to occupy a constant fraction of the population everywhere in the country, then $\phi_i = 0$ everywhere because its local concentration would be $c_i = (m_i/N)\rho$. (This situation would be perfectly ergodic.) The range of typical values for the anomaly depends a clan's aggregate concentration in the country. Examining the anomaly relative to clan concentration, the year-2000 numbers for $\phi_i/(m_i/N)$ range from -1700 to 7400 for Kim from Gimhae and from -19000 to 87000 for Lee from Hakseong. Clearly, the distribution of the latter is much more heterogeneous (see Fig. S8 in the SI).

Combining the notion of clan density anomaly with traditional arguments—flow ideas based on Ohm's law and "molecular weights for population" are mentioned explicitly in [6, 10]—about migration from population gradients [2–10] suggests a simple Fickian law [40] for human transport on long time scales: we propose that the flux of clan members is $\vec{J}_i \propto \vec{\nabla}\phi_i$, so individuals move preferentially away from high concentrations of their clans. This implies that $\frac{\partial c_i}{\partial t} = \vec{\nabla} \cdot \vec{J}_i \propto \nabla^2 \phi_i$ (where we have assumed that the constant of proportionality is independent of space), which in turn yields the diffusion equation

$$\frac{\partial \phi_i}{\partial t} = D_i \nabla^2 \phi_i \,. \tag{3}$$

We thereby identify the constant of proportionality as a diffusion constant D_i with dimensions [length^2/time]. This prediction of diffusion of clan members is consistent with past theories that posited human diffusion (e.g., cultural [41] and demic [42] diffusion). An important distinction is that we are proposing a process of diffusive mixing of clans rather than diffusive expansion of an idea or group. If this theory is correct, then one should expect clan density anomalies to simply diffuse over time. One should also be able to estimate diffusion constants by comparing the spatial variance at two points in time.

One can gain insights into the above diffusion process by calculating the radius of gyration (a second moment) of clan density anomaly as a proxy for measuring ergodicity. Suppose that clan i's concentration $c_i(x, y, t)$ is known on a set of discrete regions of areas $\{A_k\}$. The centroid of the clan's anomaly has coordinates

$$\begin{aligned} x_{i,\mathrm{C}} &= \frac{1}{\phi_{i,\mathrm{tot}}} \sum_k x(k)\phi_i(k)A_k \,, \\ y_{i,\mathrm{C}} &= \frac{1}{\phi_{i,\mathrm{tot}}} \sum_k y(k)\phi_i(k)A_k \,, \end{aligned} \tag{4}$$

where $\vec{r}(k) = (x(k), y(k))$ gives the coordinates of the centroid of region k and the normalization constant is

$$\phi_{i,\mathrm{tot}}(t) = \sum_k \phi_i(k)A_k \,, \tag{5}$$

where $\phi_i(k)$ is the anomaly of clan i in region k at time t. Note that we calculate the centroid of population for the ith clan (as opposed to the centroid of its anomaly) using analogous formulas to Eqs. (4) and (5), but with ϕ_i replaced by the concentration c_i. The radius of gyration (i.e., the spatial second moment) r_{g_i} of clan i at time t is then defined by

$$r_{g_i}^2 = \frac{1}{\phi_{i,\mathrm{tot}}} \sum_k \left[(x(k) - x_{i,\mathrm{C}})^2 + (y(k) - y_{i,\mathrm{C}})^2 \right] \phi_i(k) A_k \,. \quad (6)$$

We can use the set of radii of gyration $\{r_g\}$ in Eq. (6) as a proxy for ergodicity, because (by construction) r_{g_i} quantifies how widely the clan density anomaly of clan i has spread across Korea [43].

We simulate the diffusion Eq. (3) between the known anomaly distributions from census data at $t_1 = 1985$ and $t_2 = 2000$ to estimate a best-fit diffusion constant for each clan. We compare our results to a null model in which movement is diffusive but driven by the aggregate population density in each region rather than by clan population anomaly. Our clan-based diffusive model performs better than the null model for approximately 84% of all clans.

V. RESULTS

We apply a least-squares fit on a doubly logarithmic scale to determine the coefficients α and γ from Eq. (1) (along with the proportional coefficient a_G, which is essentially a normalization constant, for the total number of marriages). The parameter β is irrelevant for the aggregated entries in a single *jokbo* because m_j is constant (and is equal to the total number of grooms in that *jokbo*). The strongest correlation between the gravity-model flux and the number of entries for each clan appearing in *jokbo* 1 occurs for $\alpha \approx 1.0749$ and $\gamma \approx -0.0349$, which suggests that the frequency of marriage between two families is proportional to the product of the populations of the two clans and, in particular, that there is little or no geographical dependence. The likely explanation is that the clan in *jokbo* 1 is ergodic, so the grooms could have been almost anywhere in the country, which would indeed make geographic factors irrelevant. (In the context of population genetics, this corresponds to "full mixing" [44–47].) For analysis of other *jokbo* and further details, see Sec. S1 and Tables S1–S2 of the SI.

With little loss of accuracy for the fit, we take $\alpha = 1$ and $\gamma = 0$ to avoid divergence in the rare cases in which a bride comes from the same clan as the groom (for which the distance $r_{ij} = 0$). This approximation allows us to use data from the approximately 22% of clans for which geographic origin information is not available. In Fig. 2, we show the fit for *jokbo* 1, where we have used linear regression to quantify the correlation between the gravity-model flux and the number of entries for each clan in the *jokbo*. The noticeably lower outlier to the right of the line is the data point that corresponds to the clan of *jokbo* 1, and we remark that this deviation results from a cultural taboo against marrying into one's own clan. Women from the same clan as the owners of a *jokbo* have traditionally

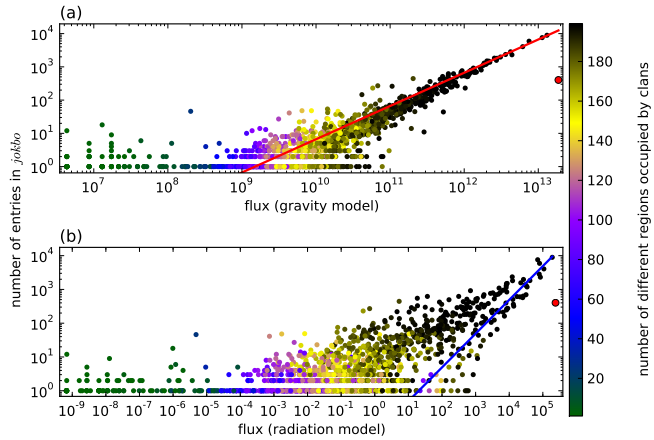

FIG. 2. (a) Scatter plot of the number of clan entries in *jokbo* 1 versus the corresponding centroid in 2000 using the gravity-model flux with parameters $\alpha = 1$ and $\gamma = 0$. We compute the line using a linear regression to find the fitting parameter, $a_G \approx 6.55(4) \times 10^{-11}$ with a 95% confidence interval, to satisfy the expression $N_i = a_G G_{ij}$, where G_{ij} is the gravity-model flux and N_i is the total number of entries from clan i in the *jokbo*. (b) The same clan entries compared to the radiation model. We compute the line using a linear regression to find the fitting parameter, $a_R \approx 0.049(2)$, to satisfy the expression $N_i = a_R R_{ij}$, where R_{ij} is the radiation-model flux and N_i is the total number of entries from clan i in the *jokbo*. In both panels, we color the points using the number of administrative regions occupied by the corresponding clans [see Figs. 3(a) and (b)]. The red markers (outliers) in both panels correspond to the clan of *jokbo* 1 (the case $i = j$).

been strongly discouraged from marrying men listed in the *jokbo* (it's possible that they were even recorded under false names in the book), and it was illegal until 1997 [48]. For the other *jokbo*, see Fig. S1 in the SI. In the bottom panel of Fig. 2, we illustrate that the radiation model gives a worse fit to the data than the gravity model.

We use census data from the year 2000 [27] to estimate the ergodicity level of clans in three different ways: (1) the number of administrative regions quantifies how "widely" each clan is distributed; (2) the radius of gyration, which we calculate from the clan density anomaly using Eq. (6), quantifies how "uniformly" each clan is distributed; and (3) the standard deviation of anomaly values, which measures how much the anomaly varies across regions. For instance, using data from the 2000 census and considering all of the clans and the 199 standardized regions, we find that 3.04% of the clans have a member in every region but that 22.1% of the clans have members in 10 or fewer regions.

We illustrate this dichotomy among the clans with the bimodal distribution in Fig. 3(a). However, from the perspective of individual clan members (see Fig. 3(b)), this dichotomy is not apparent. We show the radii of gyration that we calculate from the 2000 census data in Figs. 3(c) and (d). We can again see the bimodality in Fig. 3(c). In Fig. S8 in the SI, we illustrate the dichotomy for Kim from Gimhae and Lee from

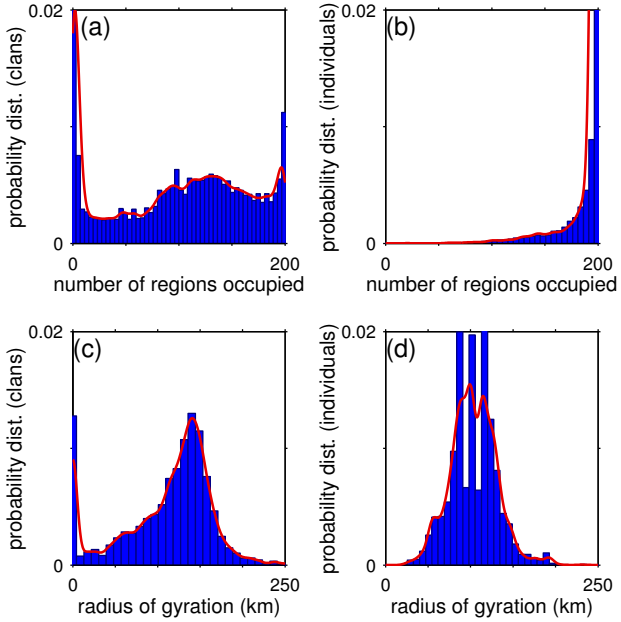

FIG. 3. (a) Probability distribution of the number of different administrative regions occupied by a Korean clan in the year 2000. (b) Probability distribution of the number of different administrative regions occupied by the clan of a Korean individual selected uniformly at random in the year 2000. The difference between this panel and the previous one arises from the fact that clans with larger populations tend to occupy more administrative regions. Note that the rightmost bar has a height of 0.17 but has been truncated for visual presentation. (c) Probability distribution of radii of gyration (in km) for clans in 2000. (d) Probability distribution of radii of gyration (in km) for clans of a Korean individual selected uniformly at random in 2000. The difference between this panel and the previous one arises from the fact that clans with larger populations tend to occupy more administrative regions. Solid curves are kernel density estimates (from MATLAB R2011a's `ksdensity` function with a Gaussian smoothing kernel of width 5).

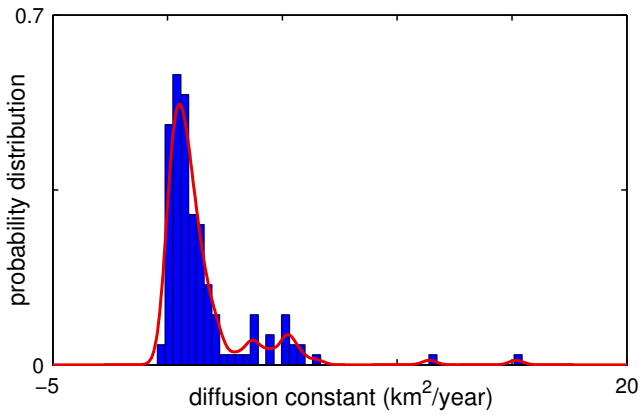

FIG. 4. Distribution of estimated diffusion constants (in km^2/year) computed using 1985 and 2000 census data and Eq. (3). The solid curve is a kernel density estimate (from MATLAB R2011a's `ksdensity` function with default smoothing). See the SI for details of the calculation of diffusion constants.

Hakseong.

As indicated in Table S1 in SI, all ten of the clans for which we have *jokbo* are fairly ergodic, so the variables associated with the j indices (i.e., the grooms) in Eqs. (1) and (2) have already lost much geographical precision, which is consistent both with the value $\gamma = 0$ in the gravity model and the poor fit to the radiation model. Again see the scatter plots in Fig. 2, in which we color each clan according to the number of different administrative regions that it occupies. Note that the three different ergodicity measures are only weakly correlated (see Fig. S9 in the SI).

Our observations of clan bimodality for Korea contrast sharply with our observations for family names in the Czech republic, where most family names appear to be non-ergodic [23] (see Fig. S10 in the SI). One possible explanation of the ubiquity of ergodic Korean names is the historical fact that many families from the lower social classes adopted (or even purchased) names of noble clans from the upper classes near the end of the Joseon dynasty (19th–20th centuries) [49]. At the time, Korean society was very unstable, and this process might have, in essence, introduced a preferential growth of ergodic names.

In Fig. 4, we show the distribution of the diffusion constants that we computed by fitting to Eq. (3). Some of the values are negative, which presumably arises from finite-size effects in ergodic clans as well as basic limitations in estimating diffusion constants using only a pair of recent years. In Fig. S11 in the SI, we show the correlations between the diffusion constants and other measures.

The assumption that human populations simply diffuse is a gross oversimplification of reality. We will thus consider the intriguing (but still grossly oversimplified) possibility of simultaneous diffusive and convective (bulk) transport. In the past century, a dramatic movement from rural to urban areas has caused Seoul's population to increase by a factor of more than 50, tremendously outpacing Korea's population growth as a whole [50]. This suggests the presence of a strong attractor or "sink" for the bulk flow of population into the city, as discussed in the rural-urban labor migration study of Ref. [51]. The density-equalizing population cartogram [52] in Fig. S12 in the SI clearly demonstrates the rapid growth of Seoul and its surroundings between 1970 and 2010.

If convection (i.e., bulk flow) directed towards Seoul has indeed occurred throughout Korea while clans were simultaneously diffusing from their points of origin, then one ought to be able to detect a signature of such a flow. In Fig. 5(a), we show what we believe is such a signature: we observe that the fraction of ergodic clans increases with the distance between Seoul and a clan's place of origin. This would be unexpected for a purely diffusive system or, indeed, in any other simple model that excludes convective transport. By allowing for bulk flow, however, we expect to observe that a clan's members preferentially occupy territory in the flow path that is located geographically between the clan's starting point and Seoul. For clans that start closer to Seoul, this path is short; for those that start farther away, the longer flow path ought to contribute to an increased number of administrative regions

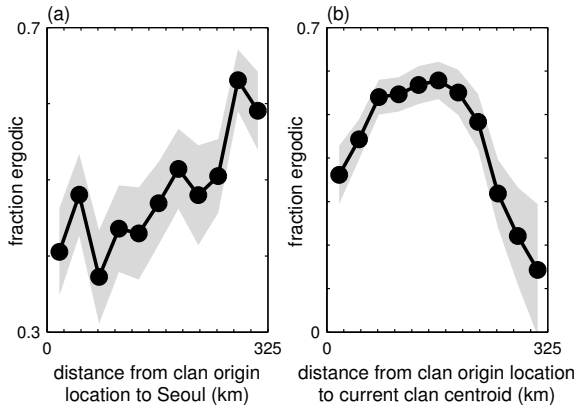

FIG. 5. (a) Fraction of ergodic clans versus distance to Seoul. The correlation between the variables is positive and statistically significant (the Pearson correlation coefficient is $r \approx 0.83$, and the p-value is $p \approx 0.0017$). For the purpose of this calculation, we call a clan "ergodic" if it is present in at least 150 administrative regions. We estimate this fraction separately in each of 11 equally-sized bins over the displayed range of distances. The gray regions give 95% confidence intervals. (b) Fraction of ergodic clans versus the distance between location of clan origin and the present-day centroid. We measure ergodicity as in the left panel, and we estimate the fraction separately for each range of binned distances. (We use the same bins as in the left panel.) The correlation between the variables is significantly positive up to 150 km ($r \approx 0.94$, $p \approx 0.0098$) and is significantly negative ($r \approx -0.98$, $p \approx 2.4 \times 10^{-4}$) for larger differences.

occupied and hence to a greater aggregate ergodicity. We plot the fraction of ergodic clans versus the distance a clan has moved (which we estimate by calculating distances between clan-origin locations and the corresponding modern clan centroids) in Fig. 5(b), and this further supports our conclusion that both convective and diffusive transport have occurred.

We assume that clans that have moved a greater distance have also existed for a longer time and hence have undergone diffusion longer; we thus also expect such clans to be more ergodic. This is consistent with our observations in Fig. 5(b) for distances less than about 150 km, but it is difficult to use the same logic to explain our observations for distances greater than 150 km. However, if one assumes that long-distance moves are more likely to arise from convective effects, then our observations for both short and long distances become understandable: the fraction of moves from bulk-flow effects like resettlement or transplantation is larger for long-distance moves, and they become increasingly dominant as the distance approaches 325 km (roughly the size of the Korean peninsula). We speculate that such clans are likely to be ones that originated in the most remote areas of Korea, or even outside of Korea, and that they have only relatively recently been transplanted to major Korean population centers, from which they have had little time to spread. This observation is necessarily speculative because the age of a clan is not easy to determine: the first entry in a *jokbo* (see Table S1 of the SI for our ten *jokbo*) could have resulted from the invention of characters or printing devices rather than from the true birth of a clan.

Ultimately, our data are insufficient to definitively accept or reject the hypothesis of human diffusion. However, as our analysis demonstrates, our data are consistent with the theory of simultaneous human "diffusion" and "convention." Furthermore, our analysis suggests that *if* the hypothesis of pure diffusion is correct, then our estimated diffusion constants indicate a possible time scale for relaxation to a dynamic equilibrium and thus for mixing in human societies. In mainland South Korea, it would take approximately $(100000\text{km}^2)/(1.5\text{km}^2/\text{yr}) \approx 67000$ years for purely diffusive mixing to produce a well-mixed society. A convective process thus appears to be playing the important role of promoting human interaction by accelerating mixing in the population.

VI. CONCLUSIONS

The long history of detailed record-keeping in Korean culture provides an unusual opportunity for quantitative research on historical human mobility and migration, and our investigation strongly suggests that both "diffusive" and "convective" patterns have played important roles in establishing the current distribution of clans in Korea. By studying the geographical locations of clan origins in *jokbo* (Korean family books), we have quantified the "ergodicity" of Korean clans. Additionally, by comparing our results from Korean clans to those that we obtained using data from Czech families, we have also demonstrated that these ideas can give robust indications of different mobility and migration patterns in different cultures.

A noteworthy feature of our analysis is that we used both data with high temporal resolution but low spatial resolution (*jokbo* data) and data with high spatial resolution but low temporal resolution (census data). This allowed us to consider both the patterns of human movement on short time scales (mobility via individual marriage processes) and its consequences for human locations on long time scales (human migration via clan ergodicity). An interesting further wrinkle would be to compare such mobility-derived time scales for human mixing patterns to genetically-derived patterns [44–47].

From a more general perspective, our research has allowed us to test the idea of using a physical analogy for modeling human *migration*—an idea put forth (but not quantified) as early as the 19th century [1–10]. Physics-inspired ideas have been very successful for the study of human *mobility*, which occurs on shorter time scales than human migration, and we propose that Ravenstein was correct when he posited that such ideas are also useful for human migration.

ACKNOWLEDGEMENTS

We thank Hawoong Jeong (정하웅) for providing data from Korean family books and Josef Novotný for providing data on surnames in the Czech Republic. We thank Tim Evans for introducing us to helpful references, Valentin Danchev for relevant references and thoughtful comments on an early version

of the draft, and Erik Bollt and Philip Kreager for useful discussions. MAP and SHL acknowledge a grant (EP/J001759/1) from the EPSRC. BJK was supported by an NRF grant funded by Korean government (2011-0015731). DMA acknowledges a grant from the James S. McDonnell Foundation.

[1] Quetelet A (1835) *Sur l'homme et le développement de ses facultés: ou Essai de physique sociale* (Paris, Bachelier, imprimeur-libraire, quai des Augustins, no 55).
[2] Ravenstein E (1885) The Laws of Migration. *J Stat Soc* 46:167–235.
[3] Stouffer SA (1940) Intervening Opportunities: A Theory Relating Mobility and Distance. *Am Sociol Rev* 5(6):845–867.
[4] Stewart JQ, Warntz W (1958) Physics of Population Distribution. *J Regional Sci* 1:99–121.
[5] Wilson AG (1967) A statistical theory of spatial distribution models. *Transpn Res* 1(3):253–269.
[6] Thornthwaite C (1934) *Internal Migration in the United States* (University of Pennsylvania Press, Philadelphia).
[7] Lucas R (1981) in *Global Trends in Migration*, eds Kritz M, Keely C, Tomasai S (Center for Migration Studies, New York), p 85.
[8] Dorigo G, Tobler W (1983) Push-Pull Migration Laws. *Annals Assn Am Geographers* 73(1):1–17.
[9] Greenwood MJ, Hunt GL (2003) The Early History of Migration Research. *Int Reg Sci Rev* 26:3–37.
[10] Stewart JQ (1950) The development of social physics. *Am J Phys* 18:239–253.
[11] González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453:779–782.
[12] Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. *Nature* 439:462–465.
[13] Szell M, Sinatra R, Petri G, Thurner S, Latora V (2012) Understanding mobility in a social petri dish. *Sci Rep* 2:457.
[14] Davies TP, Fry HM, Wilson AG, Bishop SR (2013) A mathematical model of the London riots and their policing. *Sci Rep* 3:1303.
[15] Asgari F, Gauthier V, Becker M (2013) A survey on Human Mobility and its applications. *e-print arXiv*:1307.0814.
[16] Park RE (1928) Human Migration and the Marginal Man. *Am J Sociol* 33:881–893.
[17] Sjaastad LA (1962) The Costs and Returns of Human Migration. *J Polit Econ* 70:80–93.
[18] Jokbo. `http://en.wikipedia.org/wiki/Genealogy_book#Korea`, accessed on October 11, 2013.
[19] Kiet HAT, Baek SK, Jeong H, Kim BJ (2007) Korean family name distribution in the past. *J Korean Phys Soc* 51:1812–1816.
[20] Baek SK, Kiet HAT, Kim BJ (2007) Family name distributions: master equation approach. *Phys Rev E* 76:046113.
[21] Baek SK, Minnhagen P, Kim BJ (2011) The ten thousand Kims. *New J Phys* 13:073036.
[22] Longley PA, Cheshire JA, Mateos P (2011) Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum* 42:506–516.
[23] Novotný J, Cheshire JA (2012) The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-Occurrence of Surnames. *PLOS ONE* 7:e48568.
[24] See Supporting Information for additional text, tables, and figures.
[25] An original *jokbo* usually includes more details about clans, but we only use information that we mention explicitly.
[26] Google Maps, `https://developers.google.com/maps/`, accessed on October 11, 2013.
[27] Korean Statistical Information Service, `http://kosis.kr/` (Korean version) and `http://kosis.kr/eng/` (English version), accessed on October 11, 2013.
[28] Statistical Geographic Information Service (통계지리정보서비스), `http://sgis.kostat.go.kr/` (Korean version; no English version available), accessed on October 11, 2013.
[29] Sethna JP (2006) *Statistical Mechanics: Entropy, Order Parameters and Complexity* (Oxford University Press, Oxford, United Kingdom)
[30] van Lith J (2001) Ergodic theory, interpretations of probability and the foundations of statistical mechanics. *Stud Hist Philos Sci* 32:581–594.
[31] Because we only have ten family books, our data allows only ten distinct possiblities for "clan j" (the family into which a woman marries).
[32] Cho S (1996) *Chapter 8. Traditional Korean Culture* in *Cultural and Ethnic Diversity: A Guide for Genetics Professionals* eds Fisher NL (The Johns Hopkins University Press, Maryland, United States)
[33] Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. *Nature* 484:96–100.
[34] 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* 88:022812.
[35] Young EC (1924) *The movement of farm population*. (Cornell Agricultural Experiment Station Bulletin No. 426, Ithaca, New York).
[36] Jung WS, Wang F, Stanley HE (2008) Gravity model in the Korean highway. *EPL* 81:48005.
[37] Goh S, Lee K, Park JS, Choi MY (2012) Modification of the gravity model and application to the metropolitan Seoul subway system. *Phys Rev E* 86:026102.
[38] Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2008) Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci USA* 106:21484–21489.
[39] We did attempt to locate clan origin locations; see Fig. S6 in the SI.
[40] One might call this "Thornthwaite's law" of human diffusion.
[41] Ammerman AJ, Cavalli-Sforza LL (1971) Measuring the Rate of Spread of Early Farming in Europe. *Man* 6(4):674–688.
[42] Fort J (2012) Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proc Natl Acad Sci USA* 109:18669–18673.
[43] In practice, we manipulate the anomaly data before computing radii of gyration. See the SI for details.
[44] Nei M, Imaizumi J (1966) Genetic structure of human population. I. Local differentiation of blood group gene frequencies in Japan. *Heredity* 21:9–36; Genetic structure of human population. II. Differentiation of blood group gene frequencies among isolated populations. *Heredity* 21:183–190.
[45] Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I (2007) Surnames in Western Europe: A comparison of the

subcontinental populations through isonymy. *Theor Popul Biol* 71:37–48.

[46] Rossi P (2013) Surname distribution in population genetics and in statistical physics. *Phys Life Rev* (in press) http://www.sciencedirect.com/science/article/pii/S1571064513000845, accessed on October 11, 2013.

[47] Ralph P, Coop G (2013) The Geography of Recent Genetic Ancestry across Europe. *PLOS Biol* 11(5):e1001555.

[48] Article 809 of the Korean Civil Code in *The First Ten Years of the Korean Constitutional Court* (Constitutional Court of Korea, 2001), p. 242 http://www.ccourt.go.kr/home/english/download/decision_10years.pdf, accessed on October 11, 2013.

[49] Ko DH (2005) *How people lived in the Joseon Era* (in Korean; the original Korean title: '조선시대 사람들은 어떻게 살았을까') (Cheongnyeonsa, Paju, South Korea).

[50] Seoul: City Population History, http://www.demographia.com/db-seoul-pop.htm, accessed on October 11, 2013.

[51] Todaro MP (1969) A Model of Labor Migration and Urban Unemployment in Less Developed Countries. *Am Econ Rev* 59:138–148.

[52] Gastner MT, Newman MEJ (2004) Diffusion-based method for producing density equalizing maps. *Proc Natl Acad Sci USA* 101:7499–7504.

Supporting Information

S1. JOKBO DATA

In our investigation, we examine ten digitized *jokbo* that were first studied in Ref. [S1]. In Table S1, we give basic information about the ten *jokbo* and here we summarize the results of some of our computations.

First of all, we applied the same gravity-model fit that we used for *jokbo* 1 (see discussion in the main text) to all of the *jokbo* data, and the results do not deviate much from those for *jokbo* 1 (for which $\alpha \approx 1$ and $\gamma \approx 0$). The largest deviations in the two parameter values are $\alpha \approx 1.4930$ (for *jokbo* 7) and $\gamma \approx 0.5377$ (for *jokbo* 6). Interestingly, we could not find *any* empirical value of $\gamma < 0.6$ reported in literature [S3–S9], and it seems to be extremely rare to report any empirical values at all for gravity-model parameters. As one can see in Fig. S1, the choice of $\alpha = 1$ and $\gamma = 0$ fits the data reasonably well. Note that the suppressed case of bride and groom being from the same clan is apparent in Fig. S1. This is indicated by the red markers, which are significantly below the other points in some of the panels and do not exist at all in other panels. We show the radiation-model results for the other *jokbo* in Fig. S2. As we observed for *jokbo* 1, the gravity model with $\alpha = 1$ and $\gamma = 0$ fits better than the radiation model for all *jokbo*.

Additionally, we can see that all of the clans in the *jokbo* data that we study (i.e., the grooms' side of marriages) are "ergodic" in the sense that they are widespread across the nation in 2000. This is not unexpected, as the availability of digitized *jokbo* data itself reflects clan popularity. We present the gravity-model fitting results for temporally divided *jokbo* 1 data in Table S2, and we give results that use clan-origin locations instead of population centroid in 2000 in Table S3. With these calculations, we again find that $\alpha \approx 1$ and $\gamma \approx 0$ appear to be reasonable. The general trend of population change in Korea is also reflected in the *jokbo* data. In Figs. S3 and S4, we show the time series of entries for each *jokbo* normalized by the total entries. These plots suggest that *jokbo* of the different sizes at different times tend to follow the aggregate trend of population change throughout the last several hundred years of Korean history.

S2. CENSUS DATA / POPULATION AND NUMBER OF CLANS

Since 1925, the South Korean government has conducted a census every five years [S11]. The only years in which the populations of different clans were recorded separately for different administrative regions were 1985 and 2000. This data makes it possible to estimate distribution statistics (e.g., centroid and radius of gyration) for each clan. All of the data are publicly available to download at Ref. [S11].

The total population reported in the 1985 South Korean census was 40 419 647, and clan information is available for 40 315 813 individuals. In the 2000 South Korean census, a population of 45 985 289 was reported, and a clan is indicated for every individual. The number of different clans identified in the 1985 (respectively, 2000) census was 3 520 (respectively, 4 303). There are 3 481 clans in common in the two censuses: 39 clans disappeared and 822 appeared.

In Fig. S5, we indicate how many administrative regions the 822 "new" clans occupy. New clans might correspond to foreigners who obtained South Korean citizenship during the fifteen-year period 1985–2000, or these clans might simply have been missing from the 1985 census due to error. Figure S5 supports the hypothesis that these are genuinely new clans, because their members have not spread to a large number of administrative regions. This gives a total of 6 687 distinct clans after we also incorporate the 2 384 additional clans that are listed only in the *jokbo*. In Table S1, we indicate the number of distinct clans in each of the ten *jokbo*. There are 162 clans that appear in all ten *jokbo*. For our calculations with the gravity and radiation models, we use the 4 303 clans listed in the 2000 census data, while we use the union of the 2000 census data and the additional clans listed in each *jokbo* (the number of clans in that case denoted as $N_{\gamma=0}$ is different for each clan as shown in Table S1) for the gravity model with $\gamma = 0$ that does not require the geometric information.

S3. STANDARDIZING ADMINISTRATIVE REGIONS IN 1985 AND 2000

For the administrative regions, we use municipal divisions that are composed of city (시 in Korean), county (군 in Korean), and district (구 in Korean) [S12]. In South Korea, there were 232 (respectively, 246) such administrative regions in 1985 (respectively, 2000). The difference in the number of regions between the two years reflects a slight restructuring of the political units.

For our computations, we need to unify the two different partitionings to be able to systematically compare results from 1985 and 2000 and to compute diffusion constants. To do this, we manually extract 199 "standardized" regions that we use for all computations involving administrative regions. Our construction necessitates many instances of operations like the following:

- $A + B$ (1985) $\rightarrow C$ (2000) $\Rightarrow C$ (standardized region)
- A (1985) $\rightarrow B + C$ (2000) $\Rightarrow A$ (standardized region)
- $A + B$ (1985) $\rightarrow C + D + E$ (2000) $\Rightarrow F$ (renamed standardized region)

For each operation, the region on the right is the standardized one that we use in our computations. In a given line, each different region is represented by a different letter. Thus, in the first line, two distinct regions from the 1985 census have merged into one region (and correspond exactly to that region) from the 2000 census, and we use this last region as one of our 199 standardized regions. In other examples, such as in the third line above, the standardized region does not correspond exactly to a single region from either census. Finally, we remark that the above operations are examples of what we

TABLE S1. Number of entries and other information available in each *jokbo*, values that we determined by using additional data that we obtained from other sources, and a summary of some of our computational results for the clan corresponding to each *jokbo*. For each *jokbo*, we indicate the ID (1–10), the year t_0 of its earliest entry, its number of entries N_e, and the number of distinct clans (including at least one bride for each clan) N_c among those entries [S1]. The quantity $N_{\gamma=0}$ gives the number of clans from the 2000 census (which is 4 303) plus the number of clans in each *jokbo* that are not already in the census. We can only use the latter set of clans when $\gamma = 0$ (applicable without geographic information) in the gravity model. (See the discussion in Sec. S2 in Supporting Text.) We also indicate the best values for the fitting parameters α and γ of the gravity model in Eq. (1) of the main text. We apply this fit to the brides' side of a marriage (see the main text), and we calculated these values by minimizing the sum of differences squared using the `scipy.optimize` package in `Python` [S2] (with initial values of $\alpha = \gamma = a_G = 1.0$ in our computations). For the clan that *corresponds to* each *jokbo* (i.e., the grooms' side), we compute the number of administrative regions N_{admin} in which it appears based on census data from 1985 and 2000. We use the census data to compute a radius of gyration r_g (km) for both 1985 and 2000 and to estimate a diffusion constant D (km^2/year) for diffusion of clans between those two years. As in Fig. 5 of the main text, the clans with $N_{\text{admin}}^{2000} \geq 150$ is considered to be ergodic. Based on this definition, all ten clans in the *jokbo* data are ergodic.

ID	t_0	N_e	N_c	$N_{\gamma=0}$	α	γ	N_{admin}^{1985}	N_{admin}^{2000}	r_g (1985)	r_g (2000)	Ergodic?	D
1	1513	104 356	2 657	5 510	1.0749	−0.0349	199	199	115.5	113.5	Y	0.062
2	1562	29 139	1 274	4 796	1.0145	0.2305	199	199	124.4	128.7	Y	0.737
3	1439	3 500	390	4 364	1.0853	0.2000	199	199	132.7	151.5	Y	0.426
4	1476	15 445	915	4 524	0.9678	0.1210	199	199	132.7	151.5	Y	0.426
5	1698	17 911	923	4 551	0.9452	0.2346	198	199	101.2	97.4	Y	0.062
6	1254	16 379	727	4 462	1.1102	0.5377	130	196	144.6	128.8	Y	2.253
7	1475	1 873	289	4 359	1.4930	−0.0961	199	199	110.2	116.1	Y	−0.062
8	1457	15 006	958	4 570	0.9651	0.1285	198	198	114.1	109.6	Y	0.101
9	1752	6 463	548	4 376	1.1253	0.3650	196	195	118.6	121.5	Y	0.784
10	1802	11 526	736	4 463	0.9947	0.4502	198	196	117.7	127.7	Y	0.461

TABLE S2. Gravity-model parameters α and γ in Eq. (1) of the main text calculated for temporally divided entries of *jokbo* 1 by minimizing the sum of differences squared using the `scipy.optimize` package in `Python` [S2]. (We again use initial values of $\alpha = \gamma = a_G = 1.0$ in these computations.) We sort the list of brides according to birth year, (temporally) partition the data such that each time window (except for the last one) has 10 001 entries, and indicate the mean and median birth year in each window.

window	year (mean)	year (median)	α	γ
1–10 001	1739.72	1756	1.0943	−0.1019
10 002–20 002	1828.51	1829	1.1130	−0.0396
20 003–30 003	1865.08	1865	1.1186	−0.0776
30 004–40 004	1890.72	1891	1.1277	−0.0272
40 005–50 005	1910.91	1911	1.0802	0.0209
50 006–60 006	1926.80	1927	1.0463	0.0270
60 007–70 007	1938.99	1939	1.0886	−0.0146
70 008–80 008	1949.64	1950	1.0405	0.0027
80 009–90 009	1958.01	1958	1.0443	−0.0807
90 010–100 010	1964.90	1965	1.0030	−0.0247
100 011–104 356	1971.78	1971	1.0240	−0.1077

needed to do to reconcile the 1985 and 2000 administrative regions. This is not an exhaustive list (e.g., four regions in 1985 corresponding to six regions in 2000), and we treat these other cases similarly.

For each standardized region, we sum the associated areas and populations of the constituent regions to obtain the area and population values that we use in our computations. The list of standardized regions is available online at `https://drive.google.com/file/d/0B6cEJA5TN6vfNFU3OWhjcmJkS3c`. For each standardized region, this data includes the component region names (in Korean) in 1985 and 2000, the latitudes and longitudes (and UTM easting and northing coordinates; see Sec. S5) of the component region administrative centers, the geographical areas of the component regions, and the populations of component regions in 1985 and 2000. The data are in a tab-delimited text file, for which we have used the UTF-16 (16-bit Unicode Transformation Format) encoding scheme [S13] for the Korean characters.

The regional boundaries drawn in Fig. 1 of the main text are from the 2010 data downloaded from Ref. [S14]. There is a slight difference between the regional boundaries in 2000 and

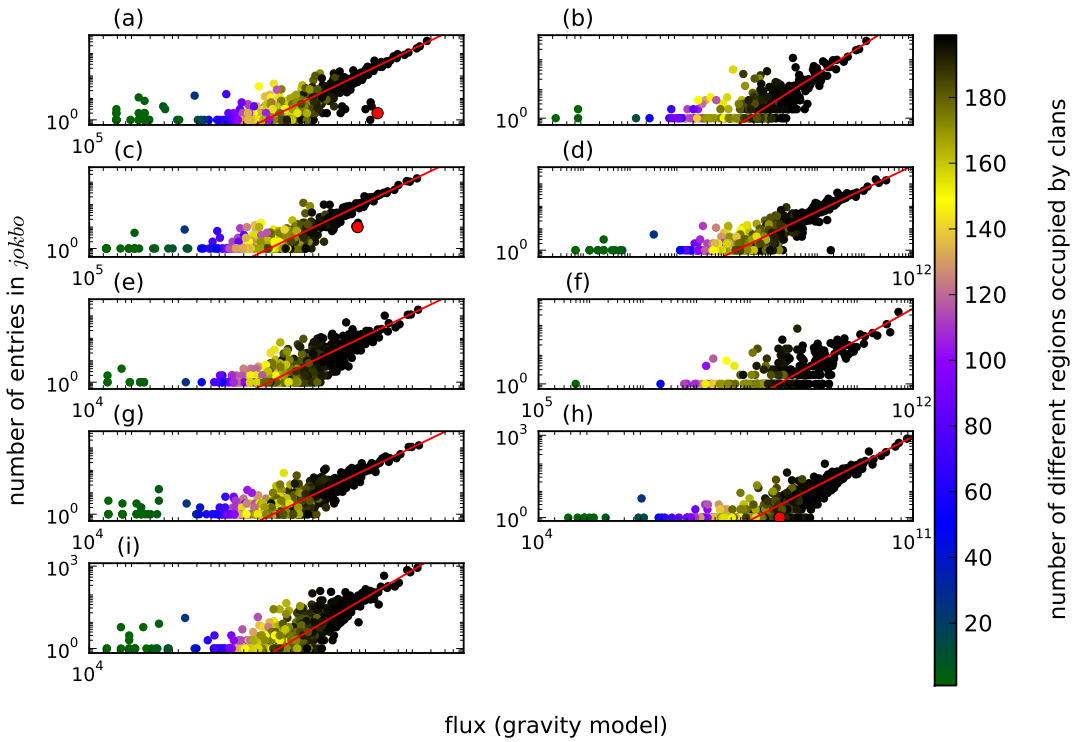

FIG. S1. Scatter plots of the number of clan entries in *jokbo* 2–10 versus the corresponding centroid in 2000 using the gravity-model flux with parameters $\alpha = 1$ and $\gamma = 0$. We show our results in numerical order of the *jokbo* in panels (a)–(i), so *jokbo* 2 is in panel (a), etc. In each panel, we calculate the line using linear regression to determine the fitting parameter a_G for $N_i = a_G G_{ij}$, where G_{ij} is the gravity-model flux and N_i is the total number of entries from clan i in the given *jokbo*. The parameter values are (a) $a_G \approx 2.36(1) \times 10^{-9}$ [*jokbo* 2], (b) $a_G \approx 6.6(1) \times 10^{-11}$ [*jokbo* 3], (c) $a_G \approx 5.15(5) \times 10^{-9}$ [*jokbo* 4], (d) $a_G \approx 5.15(5) \times 10^{-8}$ [*jokbo* 5], (e) $a_G \approx 5.8(1) \times 10^{-9}$ [*jokbo* 6], (f) $a_G \approx 5.1(2) \times 10^{-16}$ [*jokbo* 7], (g) $a_G \approx 4.25(5) \times 10^{-8}$ [*jokbo* 8], (h) $a_G \approx 1.44(1) \times 10^{-9}$ [*jokbo* 9], and (i) $a_G \approx 4.71(8) \times 10^{-8}$ [*jokbo* 10]. The red markers in panels (a), (c), and (h) correspond to the clans of the depicted *jokbo*, and $N_i|_{i=j=\text{own clan}} = 0$ for all of the other *jokbo*. In each case, we use a 95% confidence interval and color the points according to the number of administrative regions occupied by the corresponding clans.

2010, so we map the coordinates of administrative regions in 2000 to those in 2010 by checking which "polygon" in 2010 encloses the coordinates of administrative regions from 2000.

S4. OBTAINING GEOGRAPHICAL INFORMATION FROM GOOGLE MAPS

To obtain the coordinates of the clans' origins and the administrative regions, we wrote a `Python` script that returns the latitude and longitude given a clan origin location's name. We used a `Python` module for geocoding via Google Maps Application Programming Interface (API) [S15–S17]. We were able to successfully retrieve 3 900 clan origin locations out of the total of 4 303 clans present in the 2000 census data (see Fig. S6). We excluded the remaining 403 clan origin locations as erroneous because in each case there is a distance of more than 1 000 km between the identified origin location and the modern clan centroid. (Such distances are much larger than the scale of the Korean peninsula).

We confirmed by exhaustive checking that the modern administrative regions of South Korea are accurate. (The first author, who is South Korean, manually checked all of the locations.) However, as shown in Fig. S7, the clan origin locations are severely undersampled in the northern part of the Korean peninsula due to Google Maps' lack of information about North Korea. We hope to include more North Korean regions in future studies, and this might be possible because Google is adding details of North Korea to their mapping service [S19].

In Table S4, we present our `Python` code using Google Maps API. It requires pygeocoder (we used version 1.1.4), which is available as of July 16, 2013 [S21]. The code returns coordinates in latitude and longitude, which can then be converted to metric units (see Sec. S5).

S5. UNIVERSAL TRANSVERSE MERCATOR (UTM) COORDINATES

All of the distance measures that we employ use the Universal Transverse Mercator (UTM) geographical coordinate

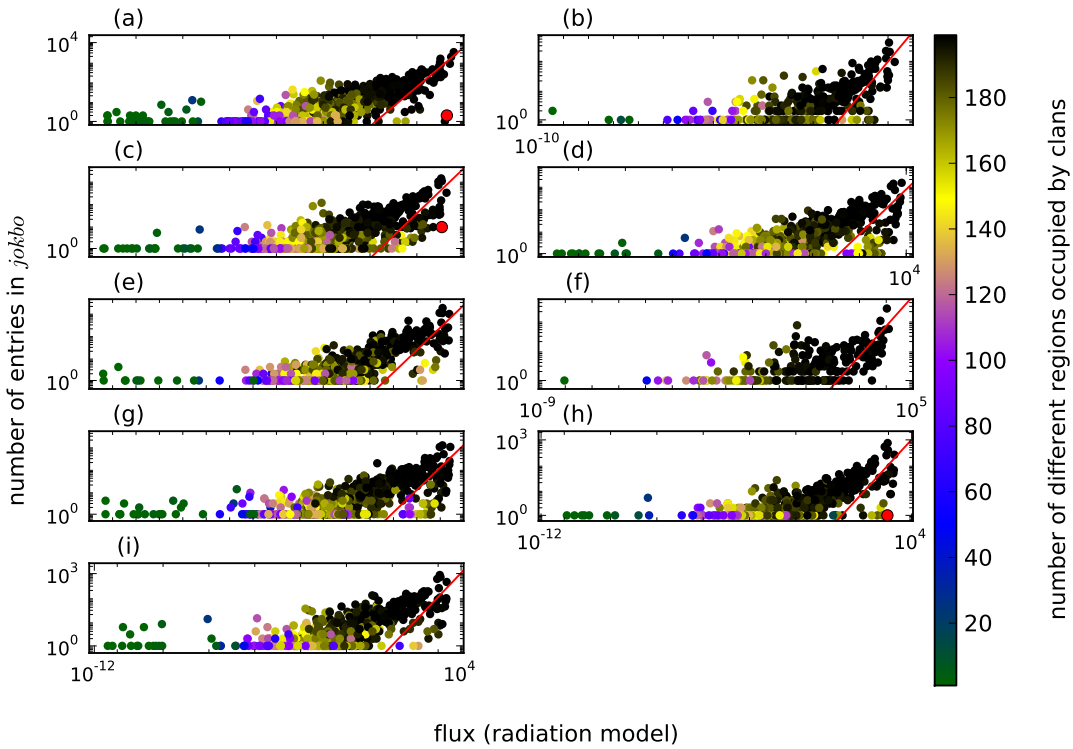

FIG. S2. Scatter plots of the number of clan entries in *jokbo* 2–10 versus the corresponding centroid in 2000 using the radiation-model flux with $\alpha = 1$ and $\gamma = 0$. We show our results in numerical order of the *jokbo* in panels (a)–(i), so *jokbo* 2 is in panel (a), etc. In each panel, we calculate the line using a linear regression to determine the fitting parameter a_R for $N_i = a_R R_{ij}$, where R_{ij} is the radiation-model flux and N_i is the total number of entries from clan i in the *jokbo*. The parameter values are (a) $a_R \approx 0.062(2)$ [*jokbo* 2], (b) $a_R \approx 0.0098(7)$ [*jokbo* 3], (c) $a_R \approx 0.040(3)$ [*jokbo* 4], (d) $a_R \approx 0.075(7)$ [*jokbo* 5], (e) $a_R \approx 0.23(2)$ [*jokbo* 6], (f) $a_R \approx 0.0069(5)$ [*jokbo* 7], (g) $a_R \approx 0.12(1)$ [*jokbo* 8], (h) $a_R \approx 0.11(1)$ [*jokbo* 9], and (i) $a_R \approx 0.11(1)$ [*jokbo* 10]. The red markers in panels (a), (c), and (h) correspond to the clans of the depicted *jokbo*, and $N_i|_{i=j=\text{own clan}} = 0$ for all of the other *jokbo*. In each case, we use a 95% confidence interval and color the points according to the number of administrative regions occupied by the corresponding clans.

system, which assigns a local two-dimensional Cartesian coordinate system to a given zone on the surface of Earth [S22]. We use the UTM `Python` module [S23], which converts (φ, λ) coordinates for latitude (φ) and longitude (λ) to UTM coordinates (and vice versa), where the UTM standard revision used by this module is WGS84 [S24]. Conversion from (φ, λ) coordinates to UTM coordinates can be also done using the website [S25].

A point (i_E, i_N) defined by UTM coordinates has two components: easting i_E and northing i_N. For example, the mean UTM coordinates of our standardized regions are $i_E \approx 381.3$ and $i_N \approx 4\,017.7$, where the numbers are in units of kilometers from a reference point. The UTM scheme splits Earth into 60 zones. Calculating distances between two points in different zones is complicated in general, but the Korean peninsula lies entirely in zone 52 [S26], which simplifies the calculation considerably. For example, Seoul's (latitude, longitude) coordinates are $(\varphi, \lambda) \approx (37.58, 127.00)$, and its UTM coordinates are $(i_E, i_N) \approx (323.4, 4\,161.5)$. For our computations, we use formulas from Ref. [S27]. In a given grid zone, these formulas are accurate to within less than a meter.

S6. CZECH REPUBLIC SURNAME DATA

The census data for the Czech republic was derived from the 2009 Central Population Register (produced by the Czech Ministry of the Interior) by the authors of Ref. [S28]. The vast majority of the Czech Republic is within zone 33, although a small part of it is in zone 34 [S22, S26]. As with the Korean clan origin locations, we used Google's API to roughly geolocate each of 206 Czech administrative regions by searching for the name of the administrative region followed by the string ", Czech Republic". We then converted all of the resulting latitude and longitude coordinates to UTM, forcing zone 33 for consistency.

We use the clan concentration (i.e., clan population density) to define an "anomaly" that indicates the difference in value from what would be observed for an "ergodic" clan, whose members are well-mixed with the population. The clan density anomaly, which we defined in the main text, is

$$\phi_i(r_k, t) = \frac{1}{A(r_k)} \left[p_i(r_k, t) - m_i(t) p_{\text{tot}}(r_k, t)/N(t) \right] . \qquad \text{(S1)}$$

TABLE S3. Gravity-model parameters α and γ in Eq. (1) of the main text calculated using the clan-origin locations (instead of the population centroids from the census data) for all of the *jokbo* by minimizing the sum of differences squared using with the `scipy.optimize` package in `Python` [S2]. (We again use initial values of $\alpha = \gamma = a_G = 1.0$ for *jokbo* 1–2, 4–6, and 8–10. For *jokbo* 3 and 7, we instead use $\alpha = a_G = 1.0$ and $\gamma = 0.01$ because the procedure did not converge when we used $\alpha = \gamma = a_G = 1.0$.)

jokbo ID	α	γ
1	1.1188	−0.0716
2	1.1261	0.1253
3	1.0983	0.0205
4	0.9310	−0.1441
5	0.8922	0.0540
6	0.9785	0.2776
7	1.5820	0.0183
8	0.9651	0.1285
9	1.1683	0.3600
10	0.9244	0.0868

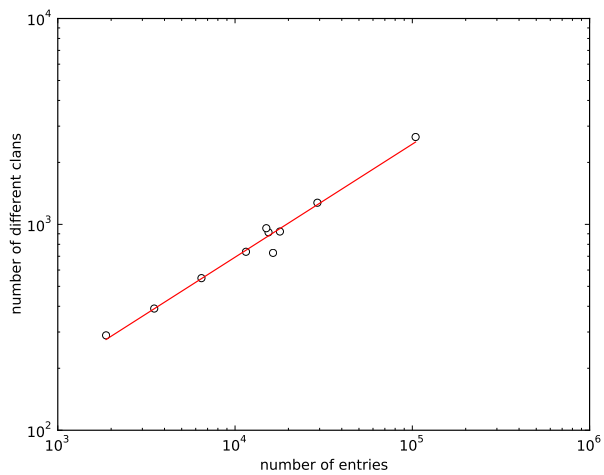

FIG. S3. For each *jokbo*, we plot the number of distinct clans N_c versus the total number of entries N_e on a doubly logarithmic scale. We calculate the red line via a linear regression to Heaps' law [S10] using the expression $N_c = 10^{b_J} N_e^{a_J}$. This yields a slope of $a_J \approx 0.55(7)$ and an intercept of $b_J \approx 0.6(3)$ (with 95% confidence intervals).

It is necessary to introduce this anomaly only because the total population is not conserved. Otherwise, we could simply take $\vec{J} \propto \vec{\nabla} c_i$ as the flux of clan members. Instead, we take the flux to be

$$\vec{J} \propto \vec{\nabla} \phi_i \,. \tag{S2}$$

S7. ESTIMATING DIFFUSION CONSTANTS

To estimate a diffusion constant for each clan, we start with the known anomaly distribution based on 1985 census data. Using an initial guess for the diffusion constant D_i (where i indexes the clan), we integrate forward in time until 2000. We then compare the numerical prediction to the known anomaly distribution based on 2000 census data using a single number: the relative error, which we define as the sum of square deviations divided by the sum of square anomalies from 2000.

After the relative error is known, we can adaptively change the "guess" for D_i and repeat the above process until we find the optimal D_i. In practice, we use MATLAB's built-in numerical optimization routine `fminbnd` [S29], which implements a Nelder-Mead downhill simplex search.

A. Some Subtleties

Because the census data is irregular, we first interpolate it to a regular grid before numerical integration of the diffusion equation. The grid that we use covers the UTM zone 52 rectangular region from 245 to 545 km easting and from 3800 to 4250 km northing with a uniform 2.5 km spacing between grid

TABLE S4. `Python` code to obtain location coordinates in latitude and longitude from the Google Maps API. We set the delay of two seconds not to exceed Google Maps API's `OVER QUERY LIMIT` [S20] in case of a large set of locations.

```
from time import sleep
import sys

## https://bitbucket.org/xster/pygeocoder/wiki/Home
from pygeocoder import Geocoder

## get latlng from address
#TODO: edit this as required
address = 'Seoul, Korea'

try:
    sleep(2)

    results = Geocoder.geocode(address)
    (lat, lng) = results[0].coordinates

except:
    print 'error (addr2coord): ', sys.exc_info()[0]
    lat = -1
    lng = -1

print 'lat/lng : ', [lat, lng]

## retrieve accosiated address
try:
    sleep(2)

    results = Geocoder.reverse_geocode(lat, lng)
    retri_addr = str(results[0])

    print 'retri_addr: ', retri_addr

except:
    print 'error (coord2addr): ', sys.exc_info()[0]
```

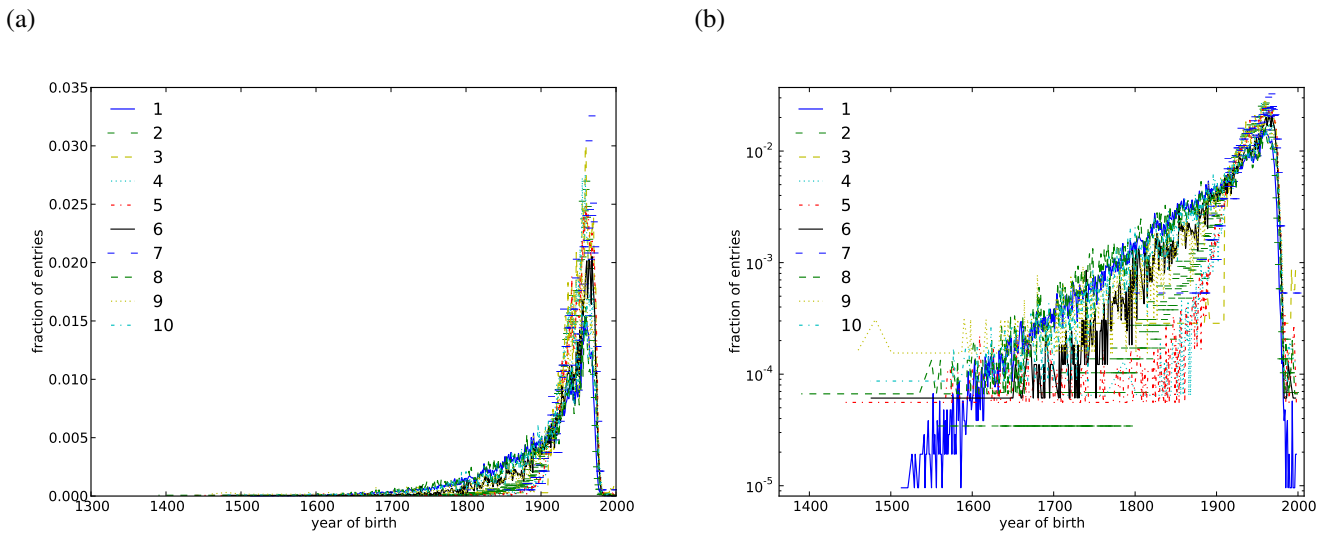

FIG. S4. Fraction of entries in each *jokbo* versus the birth year of brides using (a) linear and (b) semi-logarithmic scales. The sudden drop on the right of each panel simply reflects the fact that women who are too young are not married yet.

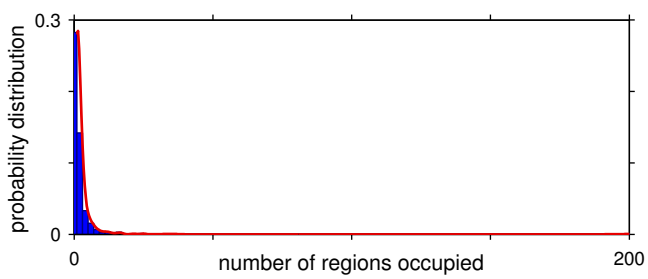

FIG. S5. Probability distribution for the number of different administrative regions occupied by the 822 "new" clans that are in the 2000 census data but are not in the 1985 census data. The solid curve is a kernel density estimate (from Matlab R2011a's `ksdensity` function with smoothing width 1.)

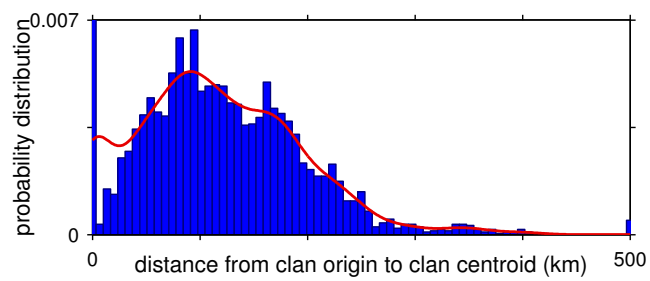

FIG. S6. Probability distribution for how far clans have moved in terms of the distance from the historical clan origin location to the clan centroid from 2000. We geographically identified the origin and centroid for 3 900 clans among the 4 303 clans in the 2000 census data. The rightmost bar corresponds to $\geq$ 500 km, and the solid curve is a kernel density estimate (from Matlab R2011a's `ksdensity` function with default smoothing).

points. (We exclude Jeju Island, which is distant from mainland Korea and is located south of the mainland.) We employ a standard five-point stencil to approximate the Laplacian operator in space and integrate in time with a 4th/5th order adaptive Runge-Kutta scheme. We impose Neumann conditions at the boundaries.

Because the numerical integration is unstable for negative values of D_i, we restrict D_i to be positive. We test for the possibility of "negative diffusion" by repeating the entire optimization procedure after interchanging the 1985 and 2000 data sets; that is, we start from 2000 and integrate backwards in time with a positive diffusion constant.

B. Testing against a Null Hypothesis

Because our hypothesis of clan diffusion (see the main text) is somewhat speculative, it is important to test it against a basic null hypothesis. We take the null hypothesis to be a model in which clans do indeed diffuse, but in which clan affiliation plays no role: members simply diffuse in accordance with the local population density ρ. Therefore,

$$\frac{\partial \phi_i}{\partial t} = D_i \nabla^2 \rho \,. \tag{S3}$$

We accept the null model as preferable to the clan diffusion model whenever it yields a lower relative error using its best-fit D_i.

C. Numerical Results

The results of our computational examination of diffusion are as follows. When we include all 3 481 clans for which data was available, we obtain a mean diffusion constant $\bar{D}$ =

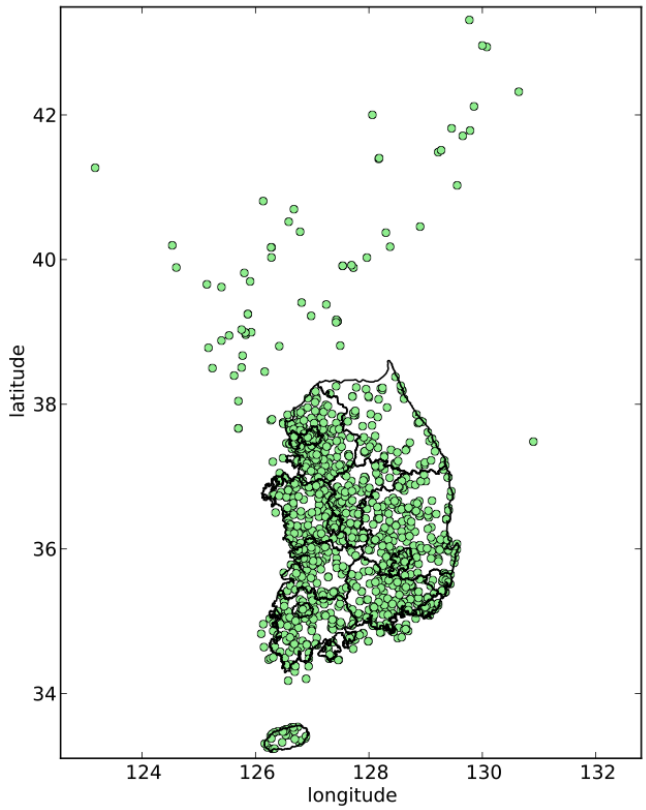

FIG. S7. Locations of clan origin names found using the Google Maps application programming interface (API) on top of the Korean map. We show administrative boundaries of South Korea corresponding to the upper-level local autonomy (광역자치단체) composed of provinces (도), special autonomous province (특별자치도), special city (특별시), and metropolitan cities (광역시) [S18].

$\langle D_i \rangle$ (where we average over the clans) of $\bar{D} \approx 2.91$ and a standard deviation of $\sigma_D \approx 10.4$. When we remove "small" clans (i.e., those with fewer than 2 000 members in the year 2000), the 707 remaining clans have a mean diffusion constant of $\bar{D} \approx 0.79$ and a standard deviation of $\sigma_D \approx 4.8$. When we also remove "ergodic clans" (by eliminating the 75% with the largest year-2000 radii of gyration), the 182 remaining clans have a mean diffusion constant of $\bar{D} \approx 1.3$ and a standard deviation of $\sigma_D \approx 2.0$.

When we use all 3 481 clans, our computations favor the clan-diffusion model over the null model in about 84% of the cases. Additionally, about 64% of all clans had both positive best-fit diffusion coefficients *and* have mobility patterns that are explained better by the clan-diffusion model.

When we exclude both small and ergodic clans, our computations favor the clan-diffusion model over the null model in about 81% of the cases (i.e., for 148 clans). Moreover, 78% of all clans had both positive best-fit diffusion constants *and* have mobility patterns that are explained better by the clan-diffusion model than by the null model.

In Fig. 4 of the main text, we show a histogram of diffusion constants for the subset of clans for which the clan-diffusion model appears to be valid. These 148 clans are non-ergodic, have a positive best-fit diffusion constant, and are fit better by the clan-diffusion model than by the null model. They have a mean diffusion constant of $\bar{D} \approx 1.6$ and a standard deviation of $\sigma_D \approx 2.1$.

S8. OTHER RESULTS

In this section, we discuss several figures shown after the references that illustrate additional results. Figures S8–S11 explore clan ergodicity in more detail, and Fig. S12 illustrates the "convective" effect of movement into the Seoul metropolitan area.

In Fig. S8, we show the distribution of clan density anomalies for the clans of the two Korean authors of this publication. As we illustrated in Fig. 1 of the main text, Kim from Gimhae appears to be ergodic, whereas Lee from Hakseong appears to be more localized.

In Fig. S9, we examine the correlation between the distance that a clan has moved and its current ergodicity. We consider two measures of ergodicity—radius of gyration and number of regions occupied—and we also show the correlation between these two diagnostics for all clans. In Fig. S10, we show the distribution of clan ergodicities—using both number of regions occupied and radius of gyration—for the Czech Republic. This is like Fig. 3 of the main text, in which we showed this information for Korea. In Fig. S11, we use scatter plots to examine the possible correlation between the calculated diffusion constants and distance a clan has moved. We similarly illustrate the connection between the diffusion constants and the two measures of ergodicity.

In Fig. S12, we show two "cartograms" [S30] of South Korea. In these images, we distort the administrative regions in proportion to the population of people who live there. The growth of the Seoul metropolitan area over the past 40 years is clearly visible.

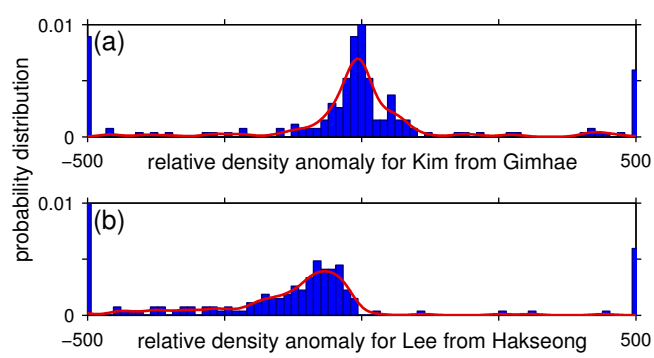

FIG. S8. Distributions of clan density anomalies $\{\phi_i\}$ (see the definition in the main text) in 2000 over the 199 standardized administrative regions for (a) Kim from Gimhae and (b) Lee from Hakseong. The leftmost and rightmost peaks correspond to ≤ -500 and ≥ 500, respectively. Solid curves are kernel density estimates (from Matlab R2011a's `ksdensity` function with a Gaussian smoothing kernel of width 20).

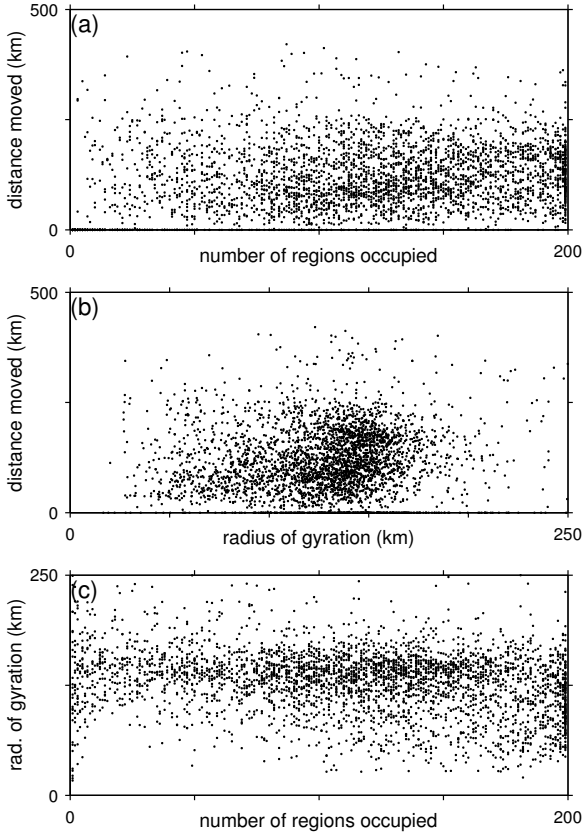

FIG. S9. Scatter plots of (a) the distance between the clan-origin location and the population centroid versus the number of administrative regions, (b) the distance between the clan-origin location and the population centroid versus the radius of gyration, and (c) the radius of gyration versus the number of administrative regions. The corresponding Pearson correlation values are (a) $r \approx 0.20$ (from 3 481 clans that include all of the required information; the p-value is $p \approx 5.7 \times 10^{-34}$), (b) $r \approx 0.066$ (from 3 481 clans that include all of the required information; the p-value $p \approx 9.4 \times 10^{-5}$), and (c) $r \approx -0.26$ (from 3 481 clans; the p-value is $p \approx 1.8 \times 10^{-53}$). Note that correlations over limited ranges may be different and significant: for example, in panel (c), the two metrics for ergodicity are significantly positively correlated when $r_g < 50$ km ($r \approx 0.39$, $p \approx 2.3 \times 10^{-4}$).

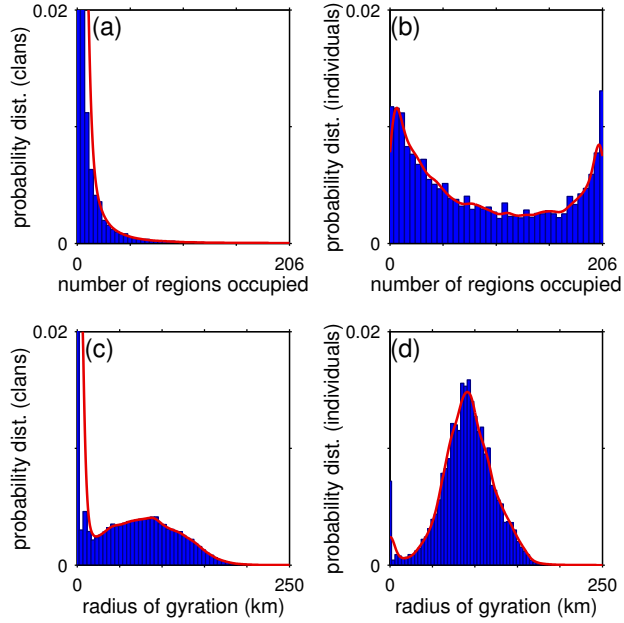

FIG. S10. (a) Probability distribution of the number of different administrative regions occupied a Czech family name in 2009. Note that the leftmost two bars have heights of 0.17 and 0.03 but have been truncated for visual presentation. (This data was initially analyzed in Ref. [S28].) (b) Probability distribution of the number of different administrative regions occupied by the clan of a Czech individual selected uniformly at random in 2009. The difference between this panel and the previous one arises from the fact that clans with larger populations tend to occupy more administrative regions (c) Probability distribution of radii of gyration (in km) of Czech family names in 2009. Note that the leftmost bar has a height of 0.11 but has been truncated for visual presentation. (d) Probability distribution of radii of gyration (in km) of Czech family names of a Czech individual selected uniformly at random in 2009. The difference between this panel and the previous one arises from the fact that clans with larger populations tend to occupy more administrative regions. Observe that the distributions in panels (a) and (b) are starkly different from the distributions in panels (a) and (b) from Fig. 3 in the main text. Solid curves are kernel density estimates (from MATLAB R2011a's `ksdensity` function with a Gaussian smoothing kernel of width 5).

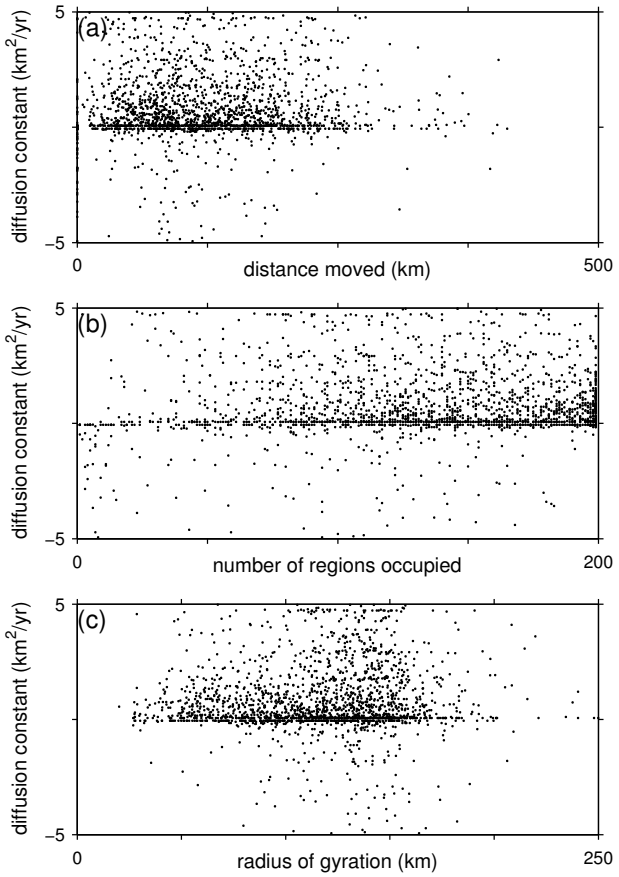

FIG. S11. Scatter plots of diffusion constant from the difference between the 1985 and 2000 census data versus (a) the distance between the origin to the population centroid, (b) the number of administrative regions, and (c) the radius of gyration of clans in 2000. The corresponding Pearson correlation values are (a) $r \approx 0.13$ (from 3 481 clans that include all of the required information; the p-value is $p \approx 1.9 \times 10^{-15}$), (b) $r \approx 0.088$ (from 3 481 clans that include all of the required information; the p-value is $p \approx 1.7 \times 10^{-7}$), and (c) $r \approx 0.097$ (from all 3 481 clans in the 1985 census data; the p-value is $p \approx 1.0 \times 10^{-8}$).

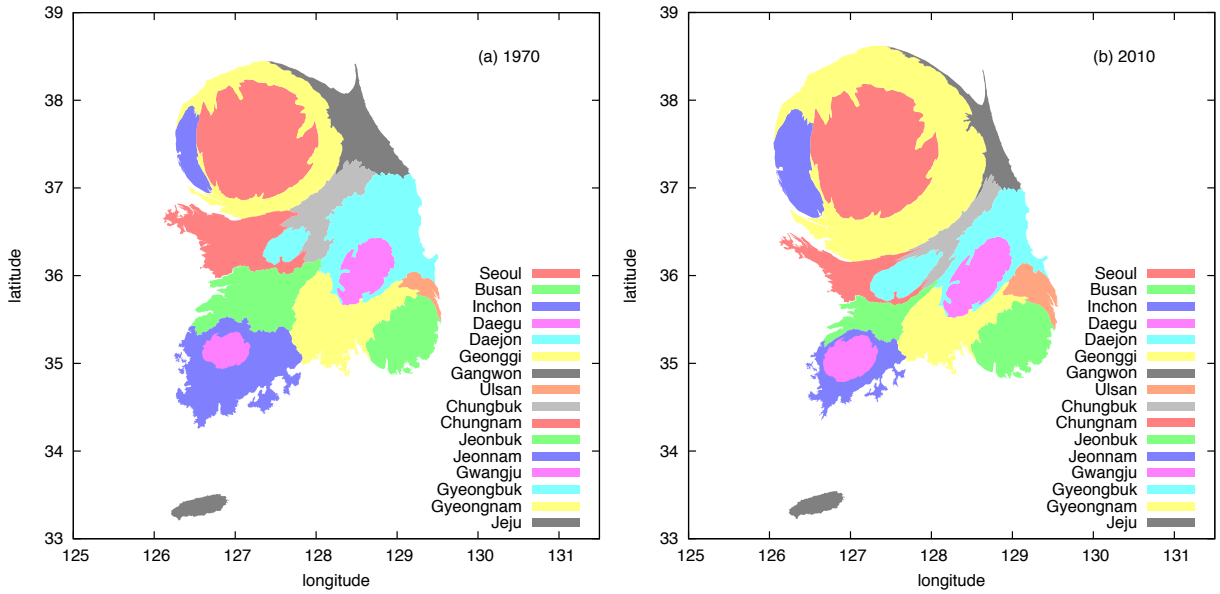

FIG. S12. Density-equalizing population cartograms [S30] for South Korea using population data from (a) 1970 and (b) 2010 censuses [S11]. The coordinates are longitude on the horizontal axis and latitude on the vertical axis. The growth of the Seoul metropolitan area over the past 40 years is clearly visible (compare this to a regular map of South Korea, e.g., Fig. 1 in the main text).

[S1] Kiet HAT, Baek SK, Jeong H, Kim BJ (2007) Korean family name distribution in the past. *J Korean Phys Soc* 51:1812–1816.

[S2] Optimization and root finding (`scipy.optimize`). `http://docs.scipy.org/doc/scipy/reference/optimize.html`, accessed on October 11, 2013.

[S3] Stewart JQ, Warntz W (1958) Physics of Population Distribution. *J Regional Sci* 1:99–121.

[S4] Wilson AG (1967) A statistical theory of spatial distribution models. *Transpn Res* 1(3):253–269.

[S5] 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* 88:022812.

[S6] Young EC (1924) *The movement of farm population*. (Cornell Agricultural Experiment Station Bulletin No. 426, Ithaca, New York).

[S7] Jung WS, Wang F, Stanley HE (2008) Gravity model in the Korean highway. *EPL* 81:48005.

[S8] Goh S, Lee K, Park JS, Choi MY (2012) Modification of the gravity model and application to the metropolitan Seoul subway system. *Phys Rev E* 86:026102.

[S9] Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2008) Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci USA* 106:21484–21489.

[S10] Heaps HS (1978) *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, New York), pp 206-208.

[S11] Korean Statistical Information Service, `http://kosis.kr/` (Korean version) and `http://kosis.kr/eng/` (English version), accessed on October 11, 2013.

[S12] Municipal level divisions in Administrative divisions of South Korea. `http://en.wikipedia.org/wiki/Administrative_divisions_of_South_Korea#Municipal_level_divisions`, accessed on October 11, 2013.

[S13] UTF-16 (16-bit Unicode Transformation Format). `http://en.wikipedia.org/wiki/UTF-16`, accessed on October 11, 2013.

[S14] Statistical Geographic Information Service, `http://sgis.kostat.go.kr/` (Korean version; no English version available), accessed on October 11, 2013.

[S15] Google Maps API licensing. `https://developers.google.com/maps/licensing`, accessed on October 11, 2013).

[S16] The Google Geocoding API. `https://developers.google.com/maps/documentation/geocoding/`, accessed on October 11, 2013.

[S17] `pygeocoder 1.2.0.3`: Python interface for Google Geocoding API V3. Can be used to easily geocode, reverse geocode, validate and format addresses. `https://pypi.python.org/pypi/pygeocoder`, accessed on October 11, 2013.

[S18] Administrative divisions of South Korea. `http://en.wikipedia.org/wiki/Administrative_divisions_of_South_Korea`, accessed on October 11, 2013.

[S19] McCurry J (2013) Google adds detail to North Korea map. *The Guardian* `http://www.guardian.co.uk/technology/2013/jan/29/google-detail-north-korea-map`, January 29, 2013.

[S20] Usage Limits for Google Maps API Web Services. `https://developers.google.com/maps/documentation/business/articles/usage_limits`, accessed on October 27, 2013.

[S21] `pygeocoder`. `https://bitbucket.org/xster/pygeocoder/wiki/Home`, accessed on October 11, 2013.

[S22] Universal Transverse Mercator coordinate system. `http://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system`, accessed on October 11, 2013.

[S23] `utm 0.3.0`: Bidirectional UTM-WGS84 converter for python. `https://pypi.python.org/pypi/utm`, accessed on October 11, 2013.

[S24] How WGS 84 defines Earth. `http://home.online.no/~sigurdhu/WGS84_Eng.html`, accessed on October 11, 2013.

[S25] Dutch S (2012) Converting UTM to Latitude and Longitude (Or Vice Versa). `http://www.uwgb.edu/dutchs/UsefulData/UTMFormulas.htm`, accessed on October 11, 2013.

[S26] Morton A, UTM Grid Zones of the World. `http://www.dmap.co.uk/utmworld.htm`, accessed on October 11, 2013.

[S27] Department of Army (1973) Universal Transverse Mercator Grid. *U.S. Army Technical Manual* 5-241-8, p. 64.

[S28] Novotný J, Cheshire JA (2012) The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-Occurrence of Surnames. *PLOS ONE* 7:e48568.

[S29] `fminbnd`: Find minimum of single-variable function on fixed interval. `http://www.mathworks.co.uk/help/matlab/ref/fminbnd.html`, accessed on October 16, 2013.

[S30] Gastner MT, Newman MEJ (2004) Diffusion-based method for producing density equalizing maps. *Proc Natl Acad Sci USA* 101:7499–7504; The code is available at `http://www-personal.umich.edu/~mejn/cart/`, accessed on October 11, 2013.