bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Matchmaker, Matchmaker, Make Me a Match: Migration of Populations via Marriages in the Past

Sang Hoon Lee (이상훈),[1,2,*] Robyn Ffrancon,[3,*] Daniel M. Abrams,[4] Beom Jun Kim (김범준),[5] and Mason A. Porter[2,6]

[1]*Integrated Energy Center for Fostering Global Creative Researcher (BK 21 plus) and Department of Energy Science, Sungkyunkwan University, Suwon 440–746, Korea*
[2]*Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, Oxford, OX2 6GG, United Kingdom*
[3]*Department of Physics, University of Gothenburg, 412 96 Gothenburg, Sweden*
[4]*Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois, 60208, USA*
[5]*Department of Physics, Sungkyunkwan University, Suwon, 440-746, Korea*
[6]*CABDyN Complexity Centre, University of Oxford, Oxford, OX1 1HP, 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, and we find that the wide variety in the geographical distributions of the clans poses interesting challenges for the direct application of these models. Using the different geographical 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. To examine population flow in more detail, we also construct and examine a population-flow network between regions. Based on the correlation between ergodicity and migration patterns in Korea, we identify two different types of migration patterns: diffusive and convective. We expect the analysis of diffusive versus convective effects in population flows to be widely applicable to the study of mobility and migration patterns across different cultures.

I. INTRODUCTION

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 human movements follow statistically predictable patterns [3–10]. Such systems-level studies are an important complement to individual-based approaches, as they can reveal population-level phenomena that are difficult to deduce by focusing on the characteristics of isolated members [11].

Research that takes a physics-based approach 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 [12–16]. 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 [17, 18]. 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; they serve primarily as a record of the names and birth years of all male descendants [19, 20]. 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 [21–23]. 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

[*] These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were also the subject of recent investigations [24, 25].

Our analysis consists of two parallel approaches: we use marriages recorded in *jokbo* to obtain snapshots of migrations (mainly of individual women) for a "marriage-flux analysis." We apply two generative models for population flow, discuss the results of applying these models, and explain the limitations that arise from the wide variety in the geographical-distribution patterns of the clans. To consider the geographical spread of clans in more detail, we conduct an "ergodicity analysis." We use the modern geographical distribution of clans from census data to infer "ergodicity" of clans (mainly caused by past movement of male descent lines). To provide an additional perspective, we also use this data to construct a network model of population flows. To the best of our knowledge, the notion of diffusive versus convective population flow is new for data-driven studies of human mobility and migration, and we believe that this kind of approach can provide valuable insights for many problems in population mobility and migration. In the present paper, we focus on long-term migration, which has significant effects on many processes over a variety spatial and temporal scales. Such processes include urban population growth and the demographic structure of cities [26]; city infrastructure and planning [27]; unemployment [28]; and the spread of culture, religion, and other ideas [29]. Most early studies of migration emphasized so-called "internal migration" (i.e., movement within a country) [2, 3, 30, 31] like the phenomena that we study, though international migration is also a prominent field of study [32, 33]. International migration has been a more popular field of study than internal migration during the past few decades, but present-day urbanization processes in Asia, Africa, and Latin America have led to renewed interest in internal migration [26, 31]. We hope that our work provides useful ideas to help solve some of the fundamental questions in the migration literature: who migrates, why people migrate, and the consequences of migration (e.g., rural depopulation).

The remainder of our paper is organized as follows. In Sec. II, we introduce the *jokbo* and census data that we use in our investigation. In Sec. III, we present our primary methodology for data analysis: the gravity and radiation models for marriage-flux analysis, a special case of the gravity model that we call the population-product model, and a diffusion model for ergodicity analysis. We present our main results in Sec. IV, and we conclude in Sec. V. We include detailed information on the data sets, data cleaning, additional results and practical considerations for our analysis, an investigation of a network model for population flow, and various other results in Appendices A–I.

II. DATA SETS

A. *Jokbo* Data Sets

For our marriage-flux analysis, we use the same ten *jokbo* data sets that were employed in [21–23]. 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 I and Figs. 8, 9 in Appendix A for details.) Each entry contains the bride's clan and year of birth [34]. The oldest book has entries that date back to the 13th century.

Previous studies of this data set [21–23] did not use any of the information that is encoded implicitly in the geographical origins of each clan. Such information, together with the modern geographical distribution of clans, comprises a key ingredient of our analysis. We convert location names to geographical coordinates using the Google Maps Application Programming Interface (GMAPI) [35]. Because of the much sparser coverage of North Korean regions by Google Maps (see Fig. 12 in Appendix D), this geolocation data is a biased sample of the full data. However, data for the southern half of the Korean peninsula is rich [36], and it is sufficient to draw interesting and robust conclusions. For example, the effect of a change in the legality of intra-clan marriage in 1997 is clearly observable in the data.

B. Modern Name Distributions

In addition to the *jokbo* data sets that we employ for marriage-flux analysis, we also use data from two Korean census reports (1985 and 2000) to evaluate the current spatial distribution of clans in Korea [37, 38]. As illustrated in Fig. 1, some clans have dispersed rather broadly but others remain localized (usually near their place of origin). Drawing on ideas from statistical mechanics [39, 40], 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 that we dub ergodic, because their distribution must differ significantly 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. METHODS

A. Generative Models for Marriage-Flux Analysis

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 [41]. Historically, professional matchmakers were employed to travel between families to arrange marriages [42], so we posit that physical distance plays a significant role in

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

determining marriage flux. We examine this hypothesis using two generative models: a conventional gravity model with adjustable parameters that incorporates the distance between regions and the effects (or lack thereof) of each regions's population [4, 5]; and a recently developed, parameter-free radiation model [43, 44].

The gravity model has been used to explain phenomena such as commuting patterns and disease spread [45–48]. 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 employ census data from 2000 to calculate centroids using the spatial population distribution for each clan [37]. Importantly, note that choosing $\gamma = 0$ in the gravity model yields a special case in which flux is independent of distance. As we will see in Sec. IV A, this situation arises when large uncertainties in geographical locations (due to clan ergodicity) hinder the accuracy of estimations of distances.

Determining the centroid locations of clans from modern census data is more accurate than attempting to locate the origins' names themselves [49] for two reasons. First, for many clans, origin-place names have differed from geographical clan centres since the beginning of recorded Korean history, in particular, during the period spanning our *jokbo* data sets [20, 50]. 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 procedure 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.

We use a version of the radiation model that takes finite-size effects into account [44]. 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 that marries 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} [43]. 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. In contrast to the gravity model, the radiation model does not include any external parameters. Importantly, this renders it unable to describe the geographically-independent situation that we need to consider in our study (and which we can obtain by setting $\gamma = 0$ in the gravity model).

For both the gravity and radiation models, we use census data from the year 2000 [37] as a proxy for past populations. This allows us compute the quantities r_{ij}, m_i, and s_{ij}. Our approximation is supported by previously reported estimates of stability in Korean society: historically, most clans have grown in parallel with the total population, so we assume that the relative sizes of clans have remained roughly constant [23]. 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).

B. Human Diffusion and the Ergodicity Analysis

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 $\phi_i(\mathbf{r}, t) = c_i(\mathbf{r}, t) - [m_i(t)/N(t)]\rho(\mathbf{r}, t)$ at position $\mathbf{r} = (x, y)$ and time t, where $c_i(\mathbf{r}, t)$ is the (spatially and temporally varying) local clan concentration (i.e., the clan population density), $m_i(t)$ is the total clan population, $\rho(\mathbf{r}, t)$ is the local population density (i.e., the total population of all clans at point $\mathbf{r}$ and time t, divided by the differential area), and $N(t)$ is the total population of all of 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 corresponds to perfect ergodicity.) The range of typical values for the clan density anomaly depends a clan's aggregate concentration in the country. Examining the anomaly relative to clan concentration, the year-2000 numbers for $\phi_i/(m_i\rho/N)$ 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. 17 in Appendix I).

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 [51] for human transport on long time scales: we propose that the flux of clan members is $\mathbf{J}_i \propto \nabla\phi_i$, so individuals move preferentially away from high concentrations of their clans. This implies that $\partial c_i/\partial t = \nabla \cdot \mathbf{J}_i \propto \nabla^2\phi_i$ (where we have assumed that the constant of proportionality is independent of space), which 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 an average 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 [52] and demic [53] diffusion). An important distinction is that

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

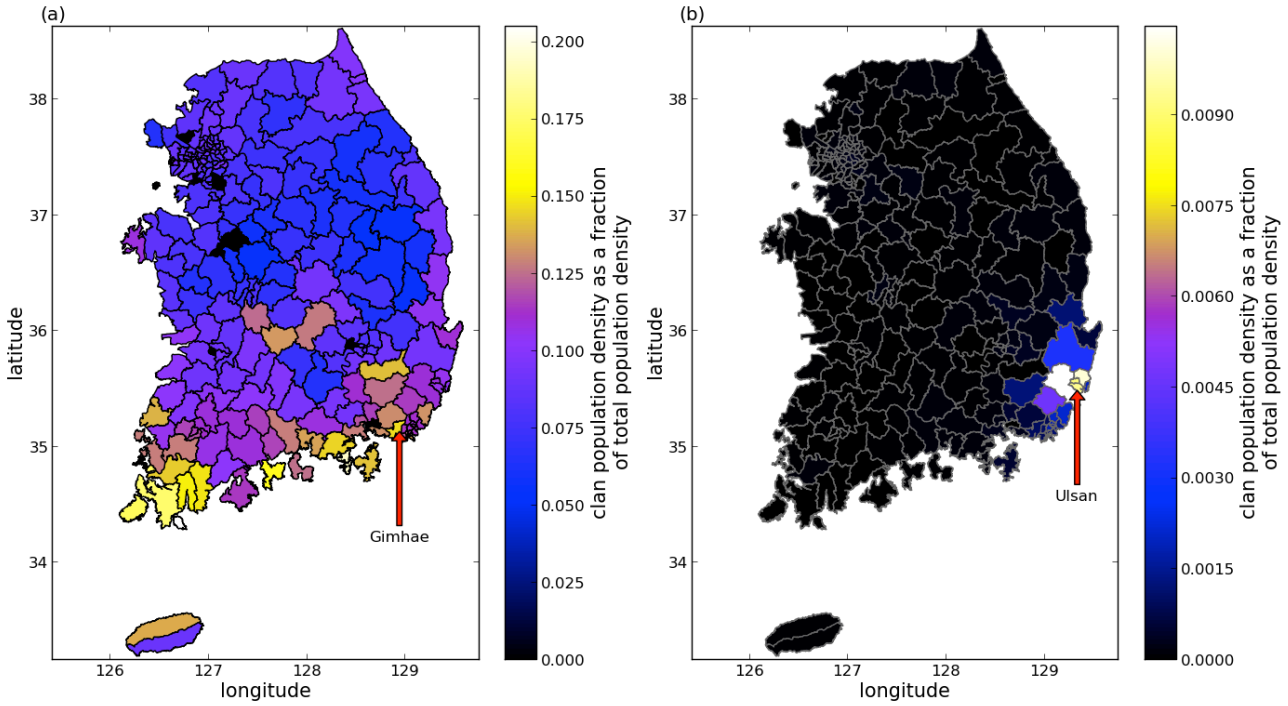

FIG. 1. Examples of (a) ergodic and (b) non-ergodic clans. We color the regions of South Korea based on the fraction of the population composed of members of the clan in the year 2000. We use arrows to indicate the origins of the two clans: Gimhae on the left and Ulsan ("Hakseong" is the old name of the city) on the right. In this map, we use the 2010 administrative boundaries [38]. See the Appendices for discussions of data sets and data cleaning.

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 the clan density anomaly as a proxy for measuring ergodicity. Suppose that clan i's concentration $c_i(\mathbf{r}, t)$ is known on a set of discrete regions $\{S_k\}$ with areas $\{A_k\}$. We define the centroid coordinates for the k-th region as

$$\mathbf{r}(k) = \frac{1}{|S_k|} \sum_{\mathbf{r} \in S_k} \mathbf{r}\,, \tag{4}$$

where $|S_k|$ is the total number of coordinate points $\mathbf{r}$ in S_k for normalization, and we henceforth use $\phi_i(k, t)$ to indicate $\phi_i[\mathbf{r}(k), t]$. The centroid of the clan's anomaly has coordinates

$$\mathbf{r}_{i,\mathrm{C}}(t) = \frac{1}{\phi_{i,\mathrm{tot}}(t)} \sum_k \mathbf{r}(k)\phi_i(k, t)A_k\,, \tag{5}$$

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

$$\phi_{i,\mathrm{tot}}(t) = \sum_k \phi_i(k, t)A_k\,, \tag{6}$$

where $\phi_i(k, t)$ 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. (5) and (6) in which ϕ_i is replaced by the concentration c_i. The radius of gyration (i.e., the spatial second moment) $r_{g_i}(t)$ of clan i at time t is then defined by

$$r_{g_i}^2(t) = \frac{1}{\phi_{i,\mathrm{tot}}} \sum_k \|\mathbf{r}(t) - \mathbf{r}_{i,C}(t)\|^2 \phi_i(k, t)A_k\,, \tag{7}$$

where $\|\cdot\|$ is the Euclidean norm. We can use the set of radii of gyration $\{r_g(t)\}$ from Eq. (7) as a proxy for ergodicity, because (by construction) $r_{g_i}(t)$ quantifies how widely the clan density anomaly of clan i has spread across Korea [54].

We simulate 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 D_i 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 the clans.

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

IV. RESULTS

A. Marriage-Flux Analysis Based on *Jokbo* and Modern Census Data

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 geographical factors irrelevant. (In the context of population genetics, this corresponds to "full mixing" [55–58].) In other words, as we discussed in Sec. III A, this special case of gravity model (for which we use $\gamma = 0$ in our analysis) corresponds to having geographical independence. Consequently, we will henceforth use the term *population-product model* for the gravity model with $\gamma = 0$. For our analysis of other *jokbo* and additional details, see Appendix A (and Tables I and II therein).

With little loss of accuracy for the fit, we take $\gamma = 0$ (i.e., we use the population-product model) to avoid divergence in the rare cases in which a bride comes from the same clan as the groom (for which the distance is $r_{ij} = 0$). We also take $\alpha = 1$ with little loss of accuracy. Using $\gamma = 0$ allows us to include data from the approximately 22% of clans for which geographical 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 population-product-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 been strongly discouraged from marrying men listed in the *jokbo* (it is possible that they were even recorded under false names in the book), and it was illegal until 1997 [59]. For the other *jokbo*, see Fig. 6 in Appendix A. In the bottom panel of Fig. 2, we illustrate that the radiation model does not give a good fit to the data. Recall from our discussion in Sec. III A that the lack of parameters in the radiation model does not allow us to explicitly consider a geographically-independent special case when using it. We emphasize, however, that this does not imply that the gravity model is "better" than the radiation model, as a direct comparison between the two models is hampered by the ergodicity of clans. In other words, the current formulations of the gravity and radiation models do not provide a solution for how to estimate fluxes between the clan centroids. Consequently, to investigate population fluxes, we incorporate modern census data. See our discussions in the next subsection and in Appendix H.

B. Ergodicity Analysis Based on Modern Census Data and a Simple Diffusion Model

We use census data from the year 2000 [37] to examine the ergodicity 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. (7), 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 the dichotomy of ergodic versus non-ergodic clans with the bimodal distribution in Fig. 3(a). However, from the perspective of individual clan members [see Fig. 3(b)], such a 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. 17 in Appendix I, we illustrate the dichotomy for Kim from Gimhae and Lee from Hakseong.

As we indicate in Table I in Appendix A, 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 of their geographical precision, which is consistent both with $\gamma = 0$ (i.e., with using the population-product model) and with $\alpha = 0$. 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 diagnostics are only weakly correlated (see Fig. 18 in Appendix I).

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 [25] (see Fig. 19). 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) [20, 60]. 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 nearby years. In Fig. 20 in Appendix I, we show the correlations between the diffusion constants and other measures.

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

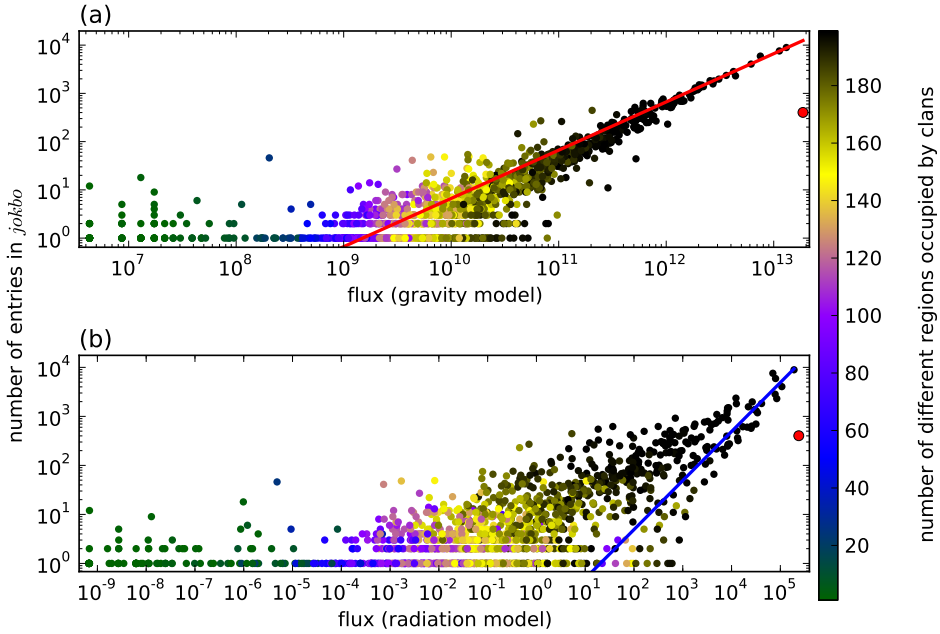

FIG. 2. Flux predictions from the population-product model (i.e., the special case of the gravity model with $\gamma = 0$) with $\alpha = 1$ and the radiation models for *jokbo* 1. (a) Scatter plot of the number of clan entries in *jokbo* 1 versus the corresponding centroid in 2000 using the population-product-model flux with $\alpha = 1$. 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 population-product-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 that are 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 (i.e., the case $i = j$).

C. Convection in Addition to Diffusion as Another Mechanism for Migration

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 [61]. This suggests the presence of a strong attractor or "sink" for the bulk flow of population into Seoul, as has been discussed in rural-urban labor migration studies [28]. The density-equalizing population cartogram [62] in Fig. 21 in Appendix I 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, 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 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). This further supports our claim that both convective and diffusive transport have occurred. To further examine clan ergodicity, we also compare each clan's radii of gyration r_g to the distances of their origin location to (1) Seoul and (2) its present-day centroid (see Fig. 22 in Appendix I). The latter shows the same general tendency as in Fig. 5. We speculate that the absence of statistical significance in the correlation between r_g and the distances from between clan origin locations and Seoul is a sampling issue, as we could not include many of the small clans in this calculation because we cannot estimate the locations of their centroids from our data (see Appendix B).

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

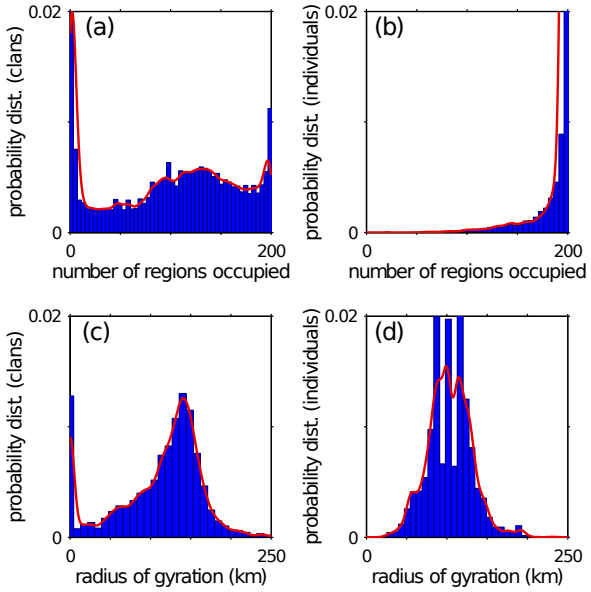

FIG. 3. Distribution of the number of different administrative regions occupied by clans. (a) Probability distribution of the number of different administrative regions occupied by a Korean clan in the year 2000. (b) Probability distribution of the number of different administrative regions occupied 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).

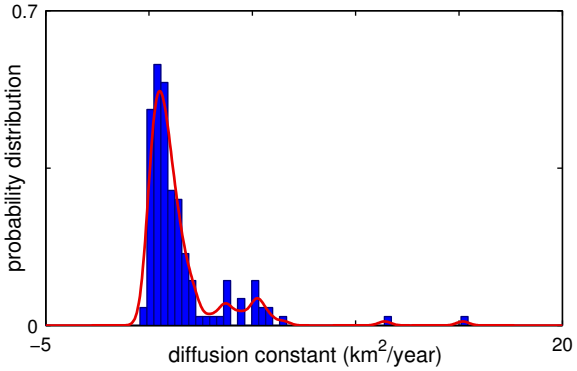

FIG. 4. Distribution of estimated diffusion constants (in km^2/year) computed using 1985 and 2000 census data and Eq. (3). The solid curve is a kernel density estimate (from MATLAB R2011a's `ksdensity` function with default smoothing). See the Appendix for details of the calculation of diffusion constants.

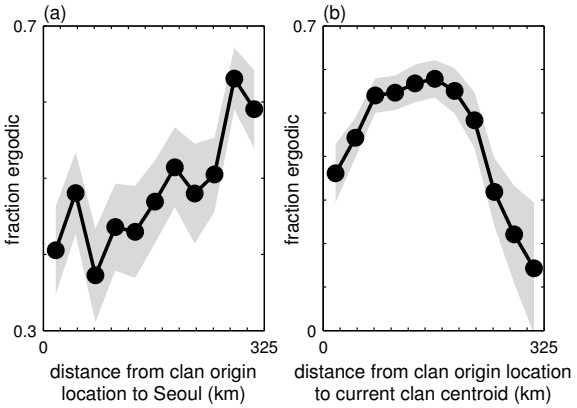

FIG. 5. Fraction of ergodic clans and distance scales of clans. (a) Fraction of ergodic clans versus distance to Seoul. The correlation between the variables is positive and statistically significant. (The Pearson correlation coefficient is $r \approx 0.83$, and the p-value is $p \approx 0.0017$.) For 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 for 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 positive and significant up to 150 km ($r \approx 0.94$, $p \approx 0.0098$) and is negative and significant for larger distances ($r \approx -0.98$, $p \approx 2.4 \times 10^{-4}$).

We assume that clans that have moved a larger 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 than from diffusive ones, 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 the clans that moved farther than 150 km 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 I in Appendix A for our ten *jokbo*) could have resulted from the invention of characters or printing devices rather than from the true birth of a clan [20].

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 "convection." Furthermore, our analysis suggests that *if* the hypothesis of pure diffusion is correct, then our estimated diffu-

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sion 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 $(100000\ \mathrm{km}^2)/(1.5\ \mathrm{km}^2/\mathrm{year}) \approx 67000$ 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. Despite the limitations imposed by our data, we try to estimate and quantify the centrality of Seoul using a network-flow model for population, and we find suggestive differences between the flow patterns of ergodic and non-ergodic clans. For details, see Appendix H.

V. 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 extent of "ergodicity" of Korean clans as reflected in time series of marriage snapshots. This underscores the utility of investigating the location distributions of individual clans. Additionally, by comparing our results from Korean clans to those from Czech families, we have also demonstrated that our approach can give insightful indications of different mobility and migration patterns in different cultures. Our ergodicity analysis using modern census data clearly illustrates that there are both ergodic and non-ergodic clans, and we have used these results to suggest two different mechanisms for human migration on long time scales. Many mobility processes involve a balance between diffusive spreading and attractiveness of a central location (and between more general diffusive and convective fluxes), so we believe that our approach in the present paper will be valuable for many situations.

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 [55–58].

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.

ACKNOWLEDGMENTS

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 and Erik Bollt, Valentin Danchev, Sandra González-Bailón, James Irish, Philip Kreager, Michael Murphy, and Tommy Murphy for helpful comments and discussions. We thank Marc Barthelemy and Richard Morris for details about their work on constructing flow networks [83], and we thank the anonymous referees for their helpful comments and suggestions. M.A.P. and S.H.L. acknowledge a grant (EP/J001759/1) from the EPSRC. B.J.K. was supported by an NRF grant funded by Korean government (2011-0015731). D.M.A. acknowledges grant #220020230 from the James S. McDonnell Foundation. S.H.L. did the majority of his work at University of Oxford.

Appendix A: Jokbo Data

In our investigation, we examine ten digitized *jokbo* that were first studied in Ref. [21]. In Table I, we give basic information about the ten *jokbo* and here we summarise the results of some of our computations.

First, we applied the same gravity-model fit that we used for *jokbo* 1 to all of the *jokbo* data, and the results do not deviate much from those for *jokbo*. That is, $\gamma \approx 0$ and $\alpha \approx 1$, so we can apply the population-product model with $\alpha = 1$. 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 [4, 5, 44–48], and it seems to be extremely rare to report any empirical values at all for gravity-model parameters. As one can see in Fig. 6, 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. 6. 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. 7.

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 surprising, 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 II, and we give results that use clan origin locations instead of population centroid in 2000 in Table III. (We also temporally divided the data from *jokbo* 6—because, as shown in Table I, it has the largest γ value among the 10 clans—and we found that it too does not exhibit systematic changes over time.) 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. 8 and 9, 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

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE I. 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 [21]. 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 use the latter set of clans in the gravity model when $\gamma = 0$ (i.e., for the population-product model, which is applicable without geographical information) and $\alpha = 1$. (See the discussion in Appendix B.) We also indicate the best values for the fitting parameters α and γ of the gravity model in Eq. (1). We apply this fit to the brides' side of a marriage, and we calculated these values by minimizing the sum of squared differences using the `scipy.optimize` package in Python [63] (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, 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	1752	3 500	390	4 364	1.0853	0.2000	199	199	132.7	151.5	Y	0.426
4	1698	15 445	915	4 524	0.9678	0.1210	199	199	132.7	151.5	Y	0.426
5	1439	17 911	923	4 551	0.9452	0.2346	198	199	101.2	97.4	Y	0.062
6	1476	16 379	727	4 462	1.1102	0.5377	130	196	144.6	128.8	Y	2.253
7	1802	1 873	289	4 359	1.4930	−0.0961	199	199	110.2	116.1	Y	−0.062
8	1254	15 006	958	4 570	0.9651	0.1285	198	198	114.1	109.6	Y	0.101
9	1458	6 463	548	4 376	1.1253	0.3650	196	195	118.6	121.5	Y	0.784
10	1475	11 526	736	4 463	0.9947	0.4502	198	196	117.7	127.7	Y	0.461

TABLE II. Gravity-model parameters α and γ in Eq. (1) calculated for temporally divided entries of *jokbo* 1 by minimizing the sum of squared differences using the `scipy.optimize` package in Python [63]. (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

the aggregate trend of population change throughout the last several hundred years of Korean history.

Appendix B: Census Data / Population and Number of Clans

Since 1925, the South Korean government has conducted a census every five years [37]. 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. [37].

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

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

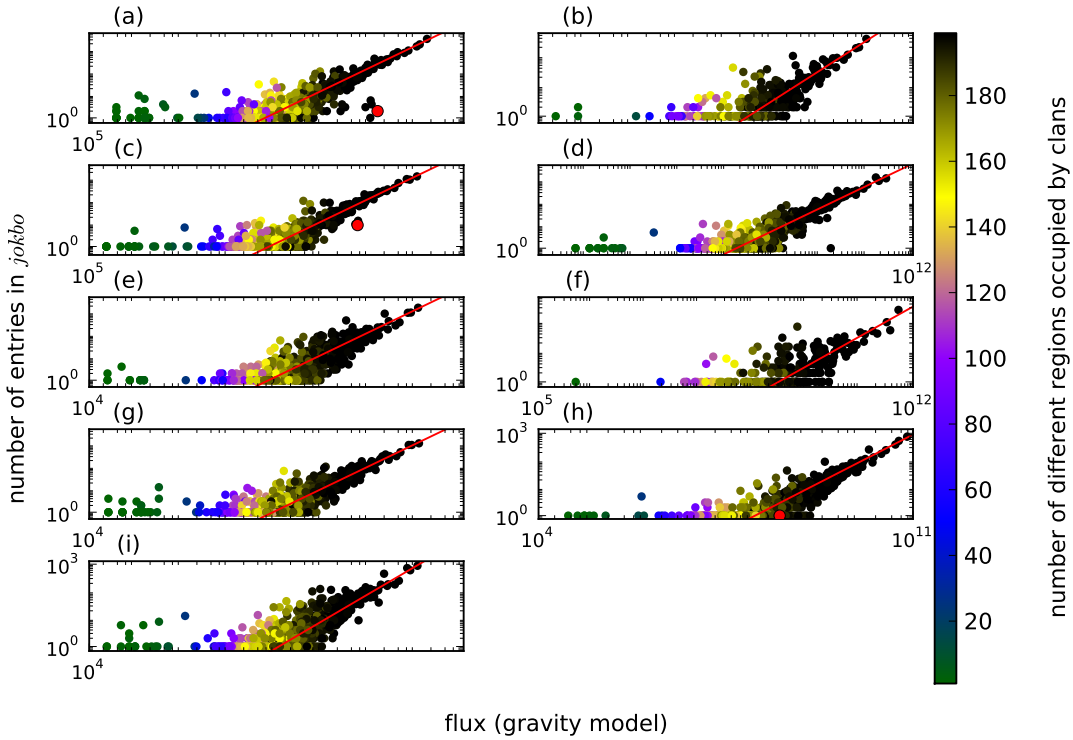

FIG. 6. Scatter plots of the number of clan entries in *jokbo* 2–10 versus the corresponding centroid in 2000 using the population-product-model flux with $\alpha = 1$. We show our results in numerical order of the *jokbo* in panels (a)–(i), so *jokbo* 2 is in panel (a), etc. In each panel, we calculate the line using linear regression to determine the fitting parameter a_G for $N_i = a_G G_{ij}$, where G_{ij} is the population-product-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.

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. 10, 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 10 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 I, we indicate the number of distinct clans in each of the ten *jokbo*. There are 162 clans that appear in all ten *jokbo*. For all calculations with the gravity and radiation models, we use the 4 303 clans listed in the 2000 census data. When we use the population-product model (for which $\gamma = 0$), we do not require geometrical information, so we also use the additional clans listed in each *jokbo*. In this case, we denote the number of clans by $N_{\gamma=0}$ (see Table I).

Appendix C: 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) [65]. 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)

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

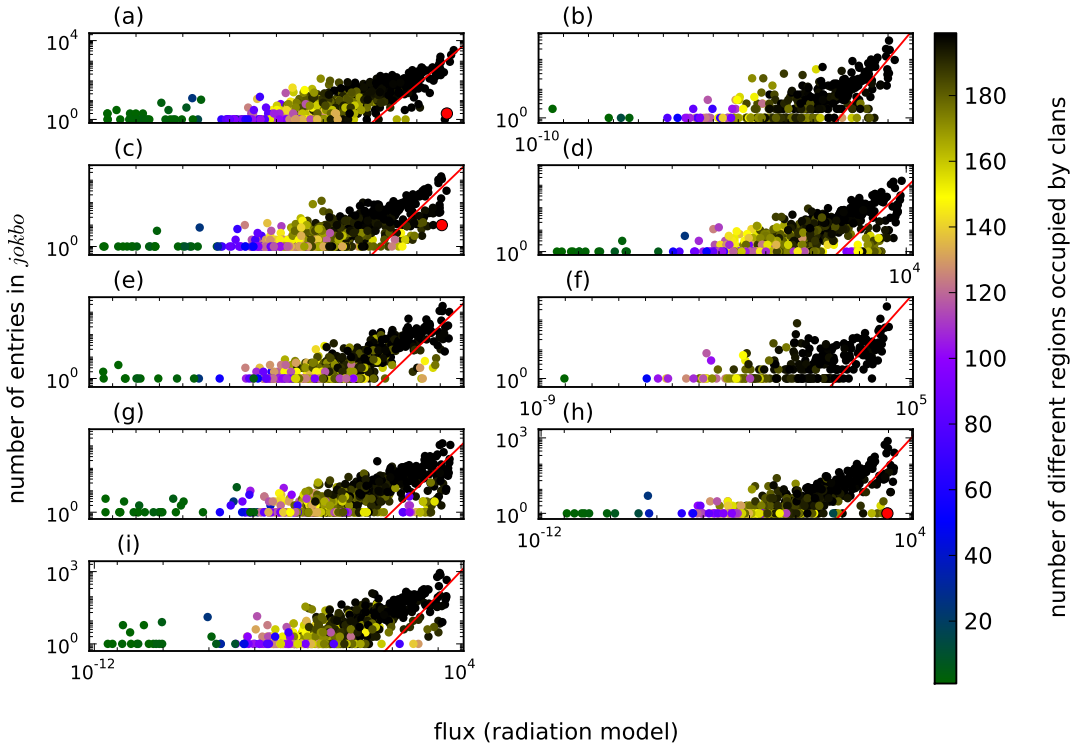

FIG. 7. Scatter plots of the number of clan entries in *jokbo* 2–10 versus the corresponding centroid in 2000 using the radiation-model flux. We show our results in numerical order of the *jokbo* in panels (a)–(i), so *jokbo* 2 is in panel (a), etc. In each panel, we calculate the line using a linear 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.

- $A + B$ (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 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 Appendix E) 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 [66] for the Korean characters.

The regional boundaries drawn in Fig. 1 are from the 2010 data downloaded from Ref. [38]. There is a slight difference between the regional boundaries in 2000 and 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.

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

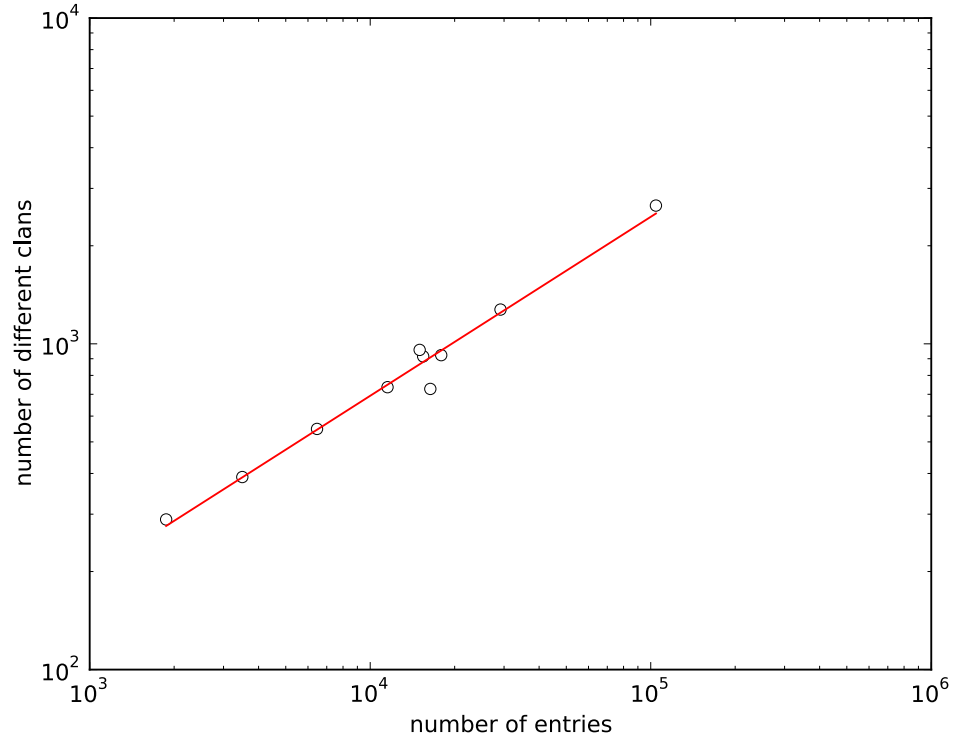

FIG. 8. For each *jokbo*, we plot the number of distinct clans N_c versus the total number of entries N_e on a doubly logarithmic scale. We calculate the red line via a linear regression to Heaps' law [64] using the expression $N_c = 10^{b_J} N_e^{a_J}$. This yields a slope of $a_J \approx 0.55(7)$ and an intercept of $b_J \approx 0.6(3)$ (with 95% confidence intervals).

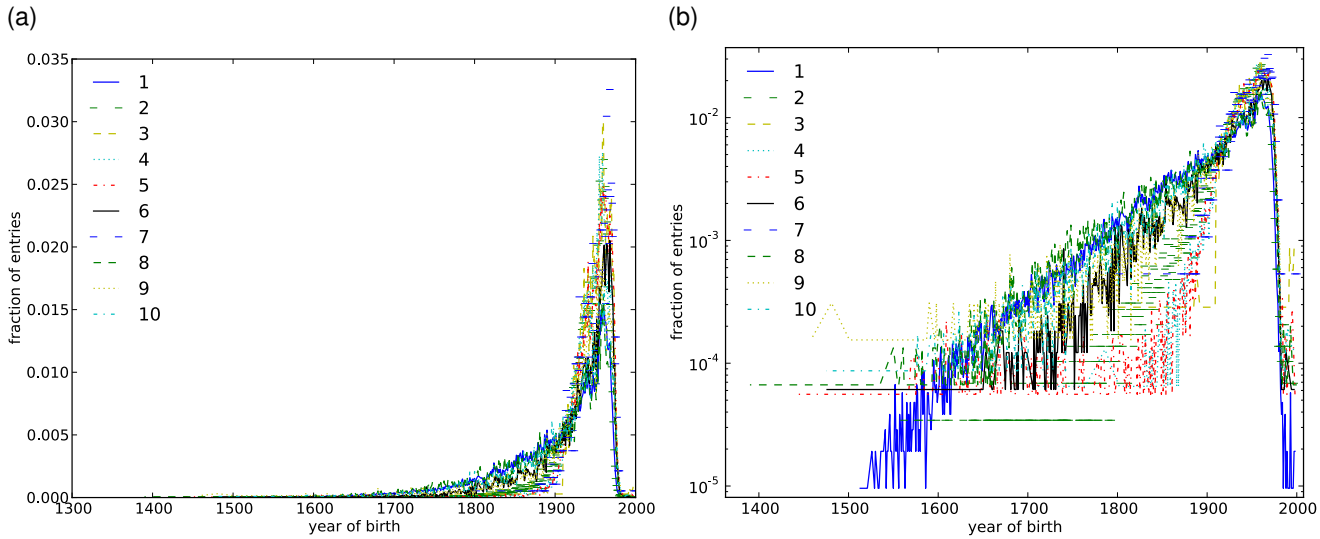

FIG. 9. Fraction of entries in each *jokbo* versus the birth year of brides using (a) linear and (b) semi-logarithmic scales. The sudden drop on the right of each panel simply reflects the fact that women who are too young are not married yet.

Appendix D: 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) [67–69]. 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. 11). We excluded the remaining 403 clan origin locations as erroneous because in each case there is a distance of more than

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE III. Gravity-model parameters α and γ in Eq. (1) 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 squared differences using with the `scipy.optimize` package in Python [63]. (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

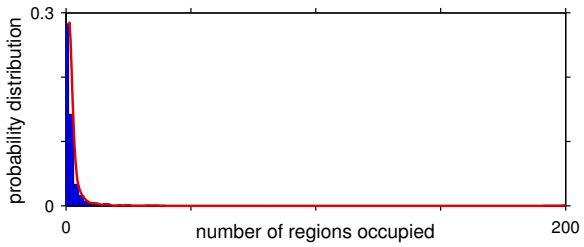

FIG. 10. Probability distribution for the number of different administrative regions occupied by the 822 "new" clans that are in the 2000 census data but are not in the 1985 census data. The solid curve is a kernel density estimate (from Matlab R2011a's `ksdensity` function with smoothing width 1.)

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. 12, 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 [71].

In Table IV, 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 [73]. The code returns coordinates in latitude and longitude, which can then be converted to metric units (see Appendix E).

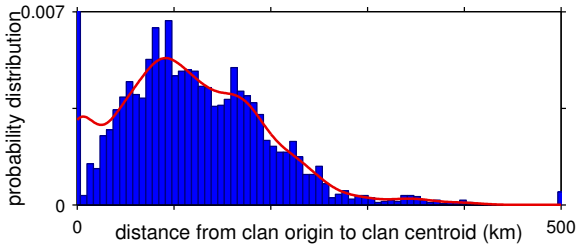

FIG. 11. Probability distribution for how far clans have moved in terms of the distance from the historical clan origin location to the clan centroid from 2000. We geographically identified the origin and centroid for 3 900 clans among the 4 303 clans in the 2000 census data. The rightmost bar corresponds to ≥ 500 km, and the solid curve is a kernel density estimate (from Matlab R2011a's `ksdensity` function with default smoothing).

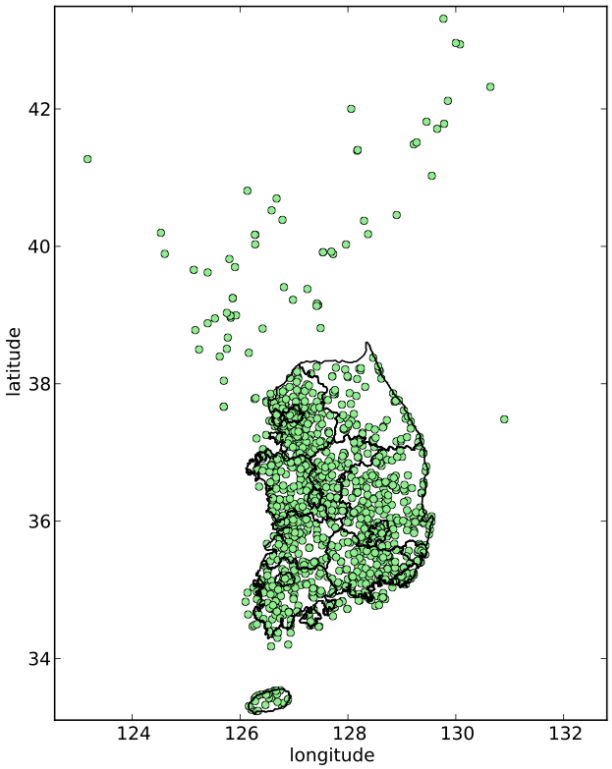

FIG. 12. Locations of clan origin names found using the Google Maps application programming interface (API) on top of the Korean map. We show administrative boundaries of South Korea corresponding to the upper-level local autonomy (광역자치단체) composed of provinces (도), special autonomous province (특별자치도), special city (특별시), and metropolitan cities (광역시) [70].

Appendix E: Universal Transverse Mercator (UTM) Coordinates

All of the distance measures that we employ use the Universal Transverse Mercator (UTM) geographical coordinate system, which assigns a local two-dimensional Cartesian coordinate system to a given zone on the surface of Earth [74].

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE IV. 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` [72] 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]
```

We use the UTM PYTHON module [75], 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 [76]. Conversion from (φ, λ) coordinates to UTM coordinates can be also done using the website [77].

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 [78], 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. [79]. In a given grid zone, these formulas are accurate to within less than a meter.

Appendix F: 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. [25]. The vast majority of the Czech Republic is within zone 33, although a small part of it is in zone 34 [74, 78]. 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 surname concentration (i.e., surname population density) to define an "anomaly" that indicates the difference in value from what would be observed for an "ergodic" for surname, whose members are well-mixed with the population. First, we obtain the centroid coordinates as in Eq. 4. The surname density anomaly, similar to what we defined for the Korean clans, is

$$\phi_i(k,t) = c_i(k,t) - [m_i(t)/N(t)]\rho(k,t)\,. \tag{F1}$$

where $c_i(k,t)$ is the population density of surname i in region k at time t, the quantity $m_i(t)$ is the total population of surname i in all regions at time t, the quantity $N(t)$ is the total population of all surnames in all regions at time t, and $\rho(k,t)$ is the population density of all surnames in region k. We use the same notational conventions as for Korean clans, so $\phi_i(k,t) = \phi_i[\mathbf{r}(k),t]$, $c_i(k,t) = c_i[\mathbf{r}(k),t]$, and $\rho(k,t) = \rho[\mathbf{r}(k),t]$. It is necessary to introduce the anomaly (F1) because the total population is not conserved. Otherwise, we could simply take $\mathbf{J} \propto \nabla c_i$ as the flux of people with surname i. Instead, we take the flux to be

$$\mathbf{J} \propto \nabla \phi_i\,. \tag{F2}$$

Appendix G: 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` [80], which implements a Nelder-Mead downhill simplex search.

1. 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 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 adap-

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tive 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.

2. Testing against a Null Hypothesis

Because our hypothesis of clan diffusion 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,

$$\frac{\partial \phi_i}{\partial t} = D_i \nabla^2 \rho \,. \tag{G1}$$

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.

3. 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 $\bar{D} = \langle D_i \rangle$ (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, 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 $\sigma_D \approx 2.1$.

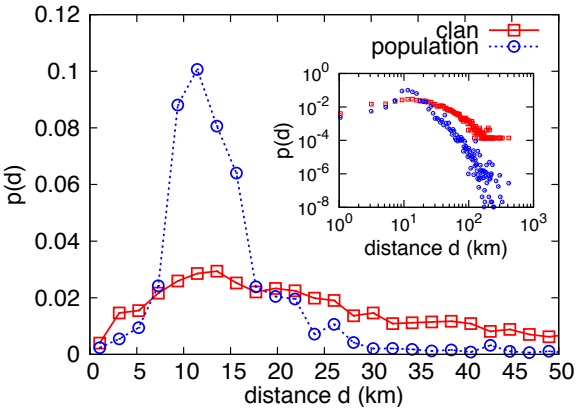

FIG. 13. Distribution of clan-centroid movement between 1985 and 2000, which we compute using $\|\mathbf{r}^{\rm pop}_{i,\rm C}(t = 1985) - \mathbf{r}^{\rm pop}_{i,\rm C}(t = 2000)\|$ from Eq. (H1). (The norm $\|\cdot\|$ is the Euclidean norm.) The curve marked by squares weighs each clan equally, and the curve marked by circles weighs each clan by its population. We indicate distance in units of populations in 2000. (The 1985 data gives a very similar distribution.) As one can see in the inset (for which we use a doubly logarithmic scale), the maximum movement distance is larger than 400 km. However, as the main panel illustrates, most clans moved considerably shorter distances.

Appendix H: Construction of Population-Flow Network between Regions

Although it is impossible to track the movement of individual people from the census data (because it does not include such information), it is possible to construct a rough estimate of the population flow between a pair of regions by considering the movement of clans (i.e., of the smallest demographic unit that it is possible to resolve with our data) between 1985 and 2000. For each clan i, we define its population centroid [note the contrast with the clan *anomaly* centroid from Eqs. (5) and (6)] as

$$\mathbf{r}^{\rm pop}_{i,\rm C}(t) = \frac{1}{c_{i,\rm tot}(t)} \sum_k \mathbf{r}(k) c_i(k,t) A_k \,, \tag{H1}$$

where the normalization constant is

$$c_{i,\rm tot}(t) = \sum_k c_i(k,t) A_k \equiv \sum_k N_i(k,t) \,. \tag{H2}$$

Recall that k indexes the region in Korea, $\mathbf{r}(k) = [x(k), y(k)]$ gives the coordinate of that region's centroid, A_k is its area, and $c_i(k,t)$ is the population density of clan i in region k at time t. Additionally, recall that $N_i(k,t) = c_i(k,t)A_k$ is the population of clan i in region k at time t (see Sec. III B)

We considered approximating the movement of each clan i by

$$\mathbf{r}^{\rm pop}_{i,\rm C}(t = 2000) - \mathbf{r}^{\rm pop}_{i,\rm C}(t = 1985) \,,$$

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

but this would entail treating an entire clan population as a "point mass," so it neglects valuable information from the spatial variation (as illustrated by our calculations of clan ergodicity). In addition, as we show in Fig. 13, the amount of movement for the majority of clans is too small to proceed further if we took such an approach. (One can also infer the small scale of clan movements from Fig. 4.)

As an alternative that avoids the undesirable point-mass approximation, we attempt to infer the flow of a clan between two regions from changes in population ratios. Let each of the 199 standardized administrative regions of South Korea (see Appendix C for details) be individual nodes of a population-flow network [81] between 1985 and 2000. To examine the central nature of Seoul in such a network, we merge the 17 regions that corresponding to different parts of Seoul into a single node that we call "Seoul." The resulting population-flow network has 183 nodes. The raw change in relative populations between regions k and k' in our census data is

$$\tilde{W}_{i,k\to k'} = \frac{N_i(k', t = 2000)}{N_i(k, t = 2000)} - \frac{N_i(k', t = 1985)}{N_i(k, t = 1985)} , \qquad \text{(H3)}$$

where we exclude the regions with $N_i(k, t = 1985) = 0$ or $N_i(k, t = 2000) = 0$ to avoid singularities. We then define the normalized relative population change as

$$U_{i,k\to k'} = \frac{\tilde{W}_{i,k\to k'}}{\max_{q\to q'} \left\{ \left| \tilde{W}_{i,q\to q'} \right| \right\}} , \qquad \text{(H4)}$$

so that $U_{i,k\to k'} \in [-1, 1]$.

The quantity $U_{i,k\to k'}$ is a proxy for a more finely-grained quantity, which we denote by $W_{i,k\to k'}$, that describes the real population flow of clan i from region k to region k' (where $k \to k'$ is a directed edge) and would be desirable to construct from data for flows of individuals. The smallest units in the census data are clans, so we need to estimate population flows from them as our basic units. We thus instead calculate $\tilde{W}_{i,k\to k'}$ and use the normalized relative population changes $U_{i,k\to k'}$ in Eq. (H4) as the adjacency-matrix elements of a weighted and directed population-flow network.

Our proxy is not guaranteed to be "correct," but it has several properties that are consistent with reasonable flow behavior: (1) If the ratio of clan populations N_1/N_2 (in regions $k' = 1, k = 2$) does not change from 1985 to 2000, then both the proxy flow $U_{i,2\to1}$ and the inferred flow $W_{i,2\to1}$ from region 2 to 1 are equal to 0; (2) if the ratio N_1/N_2 increases from 1985 to 2000, then the proxy and inferred flows from region 2 to region 1 are both positive; and (3) if the ratio N_1/N_2 decreases from 1985 to 2000, then the proxy and inferred flows from region 2 to region 1 are both negative. Naturally, both the proxy flow and the inferred flow are asymmetric (so $W_{i,2\to1} \neq -W_{i,1\to2}$).

As a downside, uniform population growth biases the proxy calculation slightly in favor of flow towards regions with smaller populations. Additionally, the proxy cannot capture circulating flows and is unlikely to do a good job when flow is strongly multipolar (i.e., if more than one area attracts a significant amount of flow). When flow is mostly between Seoul and other regions, we call it "unipolar."

Using Eq. (H4), we define a population-flow network for each clan i. We include all clans by using the entire population density of region k in year t. That is, we calculate $N_{\rm tot}(k, t) = \sum_i N_i(k, t)$ and obtain a raw total population-flow network. It has corresponding adjacency-matrix elements

$$\tilde{W}_{{\rm tot},k\to k'} = \frac{N_{\rm tot}(k', t = 2000)}{N_{\rm tot}(k, t = 2000)} - \frac{N_{\rm tot}(k', t = 1985)}{N_{\rm tot}(k, t = 1985)} , \qquad \text{(H5)}$$

and the adjacency-matrix elements for the associated normalized relative total population change are

$$U_{{\rm tot},k\to k'} = \frac{\tilde{W}_{{\rm tot},k\to k'}}{\max_{q\to q'} \left\{ \left| \tilde{W}_{{\rm tot},q\to q'} \right| \right\}} . \qquad \text{(H6)}$$

Our normalization guarantees that $U_{{\rm tot},k\to k'} \in [-1, 1]$.

In Fig. 14(a), we show box plots of the distribution of $U_{{\rm tot},k\to k'}$ using all pairs k, k'. We also show the distributions of $U_{i,k\to k'}$ for the clans Kim from Gimhae and Lee from Hakseong. We show flows to Seoul separately from flows to other regions. Note that the values of $U_{{\rm tot},k\to k'}$ are distributed much more broadly for flows to Seoul than for flows to other regions, even though there are many more adjacency-matrix elements for the latter (182 for flow to Seoul and 33 124 for flow to other regions). One can also observe this feature in the shapes of distribution themselves [see Figs. 14(d)–(f)].

One simple but intuitive way to check the centrality of Seoul is to extract a maximum relatedness subnetwork (MRS) [82] from each population-flow network. We construct an MRS as follows. For each node, we examine the weight of each of its edges and keep only the single directed edge with maximum weight. (When there are ties, we keep all edges that have the maximum weight.) We exclude out-edges from Seoul for the MRS in order to focus on the movement from other regions to Seoul. We will later compare the MRS to a null-model network which also disallow out-edges from the central node. In Figs. 15(a,b), we show the MRS from the adjacency matrix with elements $U_{{\rm tot},k\to k'}$. The central role of Seoul is apparent. As we indicate in Table V, Seoul's in-degree in this MRS is 109. This constitutes nearly 60% of the MRS edges and is consistent with the rapid growth of the Seoul area that we illustrate in Fig. 21.

We model the population flow using a simple rewired-network model inspired by the model in Ref. [83]. We start with a "monocentric" network, with Seoul as the central node, in which all directed edges start in some region (aside from the center) and terminate at Seoul. We then rewire each edge with independent, uniform probability p. Each network that we construct in this way has one directed edge for each node aside from the central one, so we can use an ensemble of such networks as a null model for our MRSs.

As we indicate in Table V, the edges in the MRSs are distributed rather heterogeneously among the regions. For example, the region in Gyeonggi Province (which has the second largest in-degree) has about 39% of the edges for the MRS that we constructed using all clans. When constructing null-model networks, we use a rewiring probability of $p = 0.4$

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

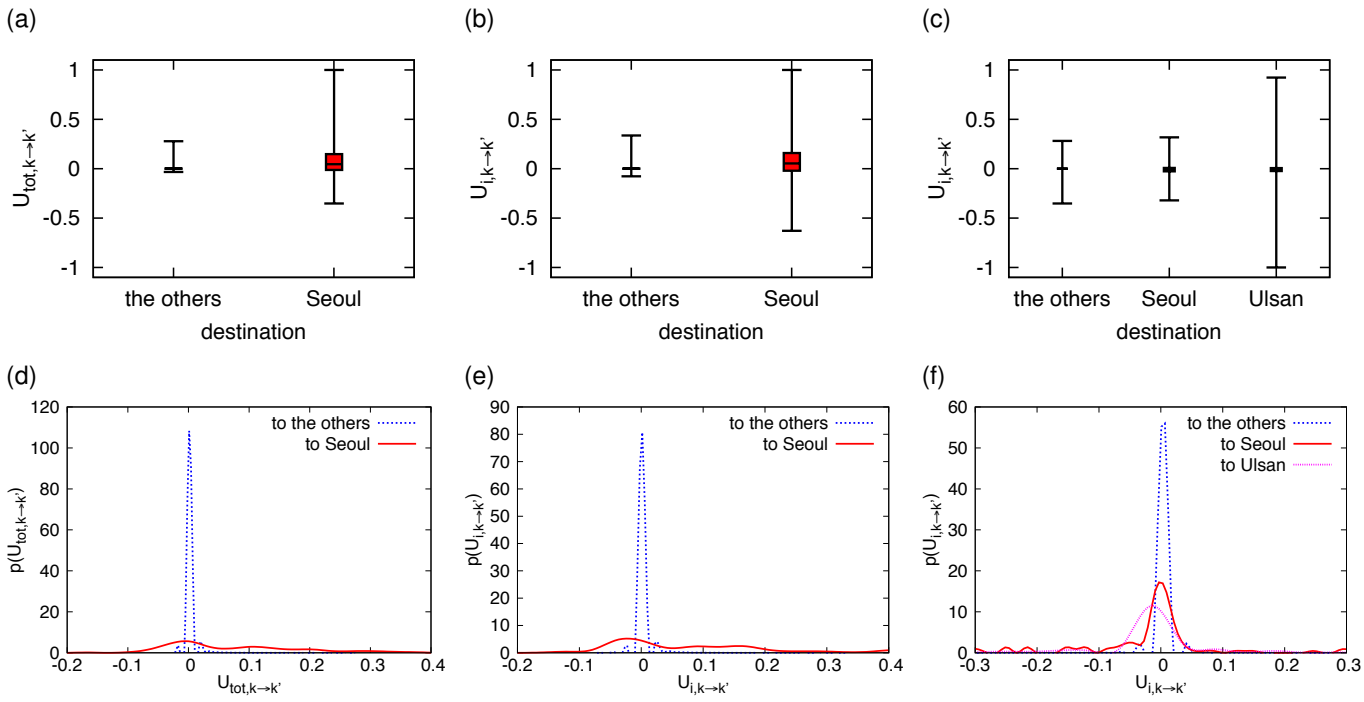

FIG. 14. We show distributions of (a) $U_{\mathrm{tot},k\to k'}$ for total population, (b) $U_{i,k\to k'}$ for Kim from Gimhae and (c) $U_{i,k\to k'}$ for Lee from Hakseong with box plots, where the boxes indicate quartile values and the whiskers give the minimum and maximum values. We also show the distributions for (d) $U_{\mathrm{tot},k\to k'}$ for total population, (e) $U_{i,k\to k'}$ for Kim from Gimhae, and (f) $U_{i,k\to k'}$ for Lee from Hakseong. In all cases, we apply `cspline` smoothing of `gnuplot` to the histograms using 50 equally-sized bins.

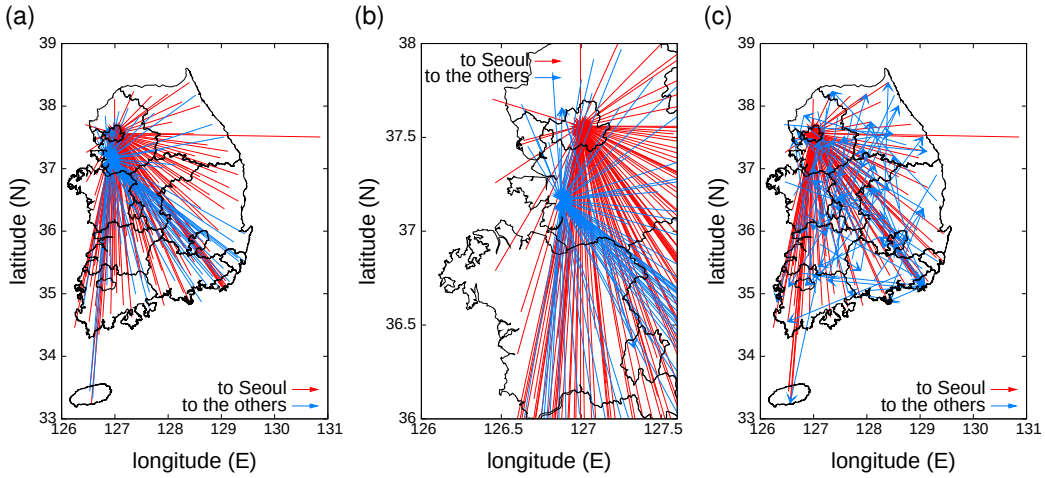

FIG. 15. (a) Maximum-relatedness subnetwork (MRS) [82] of the combined population-flow network for all clans, (b) a magnified portion of this MRS that includes all regions with nonzero in-degrees, and (c) an instance of a model (inspired by the one in Ref. [83]) of a rewired version of a monocentric network (with Seoul as the center) with a rewiring probability of $p = 0.4$. (See the discussion in the text.) We color the directed edges toward Seoul in blue, and we color the directed edges towards other regions in blue.

to ensure that about 60% of the directed edges terminate in Seoul on average (as suggested by the data when considering all clans). The null-model network ensemble generated from the rewiring process has a binomial (or approximately Poisson, as the MRS is rather sparse) in-degree distribution as a result of the given fraction p of edges that are redirected uniformly at random except for the central node or Seoul [81]. Therefore, the emergence of a second-largest hub comparable in size to the largest hub (Seoul) is extremely unlikely. We illustrate one instance of such a rewired network in Fig. 15(c), and the MRS for all clans that we constructed from empirical data differs significantly from the null-model network. (See

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE V. List of regions (which we name based on the current administrative regions) with nonzero values of in-degree in our MRSs.

clan: figure	region	in-degree	percentage of total in-degree
all clans	Seoul (서울)	109	59.9%
Figs. 15(a) and (b)	Hwaseong (화성) + Ansan (안산) + Osan (오산) + Siheung (시흥) + Gwacheon (과천) + Gunpo (군포) + Euiwang (의왕) of Gyeonggi Province (경기도)	70	38.7%
	Goyang (고양) of Gyeonggi Province (경기도)	2	1.1%
	Yuseong-gu (유성구) of Daejeon (대전)	1	0.5%
Kim from Gimhae (김해 김): Fig. 16(a)	Seoul (서울)	106	58.2%
	Hwaseong (화성) + Ansan (안산) + Osan (오산) + Siheung (시흥) + Gwacheon (과천) + Gunpo (군포) + Euiwang (의왕) of Gyeonggi Province (경기도)	74	40.7%
	Yuseong-gu (유성구) of Daejeon (대전)	1	0.5%
	Daedeok-gu (대덕구) of Daejeon (대전)	1	0.5%
Lee from Hakseong (학성 이): Fig. 16(b)	Nam-gu (남구) of Ulsan (울산)	52	31.0%
	Yongin (용인) of Gyeonggi Province (경기도)	41	24.4%
	Goyang (고양) of Gyeonggi Province (경기도)	37	22.0%
	Jung-gu (중구) + Buk-gu (북구) + Ulju-gun (울주군) of Ulsan (울산)	28	16.7%
	Buk-gu (북구) + Gangseo-gu (강서구) + Sasang-gu (사상구) of Busan (부산)	8	4.8%
	Yeongi (연기) of Chungcheongnamdo (충청남도)	2	1.2%

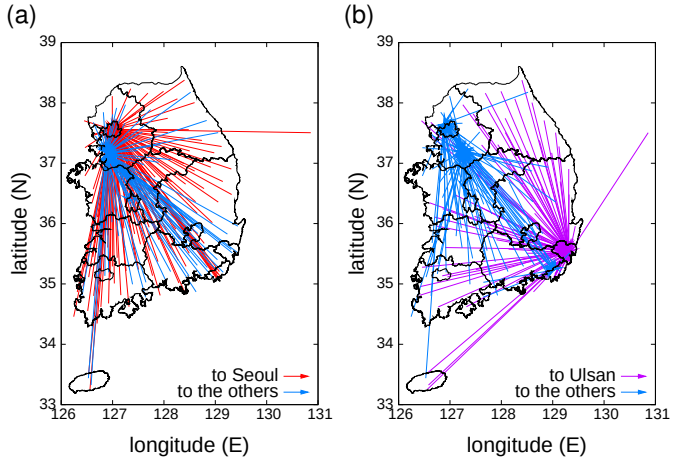

FIG. 16. (a) MRS [82] of the population-flow network for (a) Kim from Gimhae and (b) Lee from Hakseong. In panel (a), we color directed edges towards Seoul in red and directed edges towards other regions in blue. In panel (b), we color directed edges towards Ulsan in purple and directed edges towards other regions in blue.

Table V as well).

It is also instructive to examine the population-flow networks for individual clans. As with prior discussions, we will use Kim from Gimhae as an example of an ergodic clam and non-ergodic Lee from Hakseong as an example of a non-ergodic clan (see Fig. 1).

When we consider the population-flow network for the clan Kim from Gimhae [by using $N_i(k,t)$ with i corresponding to Kim from Gimhae in Eq. (H3)], we obtain a qualitatively similar result — namely, an abundance of edges terminating in Seoul — as what we obtained when using all clans. See Fig. 14(b), Fig. 16(a), and Table V. By contrast, we find that two different locations "attract" the population for Lee from Hakseong. Following the general trend in the population, one area is the Gyeonggi Province in the northwestern part of South Korea surrounding the Seoul area. (The name means "the area surrounding capital" in Korean and it is often construed to be essentially an "extended Seoul.") The other area is Ulsan/Busan in the southeastern part of South Korea (where the clan origin is located). See Fig. 14(c), Fig. 16(b), and Table V As we can see from Fig. 14(c), the Seoul region is not special for this clan. Therefore, we see that this young, non-ergodic clan has a different mobility pattern from the stabilized, ergodic clans that follow the general trend in population flow.

Appendix I: Other Results

In this section, we discuss several figures that illustrate additional results. Figures 17–20 explore clan ergodicity in more detail, and Fig. 21 illustrates the "convective" effect of movement into the Seoul metropolitan area.

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

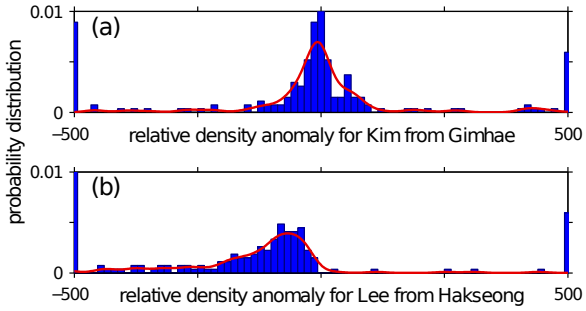

FIG. 17. Distributions of clan density anomalies $\{\phi_i\}$ in 2000 over the 199 standardized administrative regions for (a) Kim from Gimhae and (b) Lee from Hakseong. The leftmost and rightmost peaks correspond to ≤ -500 and ≥ 500, respectively. Solid curves are kernel density estimates (from MATLAB R2011a's `ksdensity` function with a Gaussian smoothing kernel of width 20).

In Fig. 17, 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, Kim from Gimhae appears to be ergodic, whereas Lee from Hakseong appears to be more localized.

In Fig. 18, 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. Some clans do not exist in the 2000 census data, and other clans only exist in one administrative region in 2000. We thus obtain radii of gyration of $r_g = 0$ for these clans. There are 3 120 clans with $r_g > 0$, and our calculations involving radius of gyration only use these clans. In Fig. 19, 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, in which we showed this information for Korea. In Fig. 20, 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. 21, we show two "cartograms" [62] 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.

To examine an alternative characterization of ergodicity as the fraction of ergodic clans (see Fig. 5 in the main text), we examine radii of gyration r_g versus distance to Seoul and versus distance between clan origin location and present-day centroid. We show our results in Fig. 22, which we see are qualitatively similar to those in Fig. 5. As in Figs. 18 and 20, we use the 3 120 clans instead of 3 900. Consequently, we repeat the computation from Fig. 5 using this smaller set of clans. As one can see in Fig. 23, we obtain the same qualitative result.

[1] A. Quetelet, *Sur l'homme et le développement de ses facultés : Ou essai de physique sociale* (Paris, Bachelier, imprimeur-libraire, quai des Augustins, no 55, 1835).

[2] E. Ravenstein, *The Laws of Migration*, J. Stat. Soc. **46**, 167 (1885).

[3] S. A. Stouffer, *Intervening Opportunities: A Theory Relating Mobility and Distance*, Am. Sociol. Rev. **5**, 845 (1940).

[4] J. Q. Stewart and W. Warntz, *Physics of Population Distribution*, J. Regional Sci. **1**, 99 (1958).

[5] A. G. Wilson, *A Statistical Theory of Spatial Distribution Models*, Transpn. Res. **1**, 253 (1967).

[6] C. Thornthwaite, *Internal Migration in the United States* (University of Pennsylvania Press, Philadelphia, 1934).

[7] R. Lucas, in *Global Trends in Migration*, eds M. Kritz, C. Keely, and S. Tomasai (Center for Migration Studies, New York, 1981), p 85.

[8] G. Dorigo and W. Tobler, *Push-Pull Migration Laws*, Annals Assn. Am. Geographers **73**, 1 (1983).

[9] M. J. Greenwood and G. L. Hunt, *The Early History of Migration Research*, Int. Reg. Sci. Rev. **26**, 3 (2003).

[10] J. Q. Stewart, *The Development of Social Physics*, Am. J. Phys. **18**, 239 (1950).

[11] J. E. Cohen, M. Roig, D. C. Reuman, and C. GoGwilt, *International Migration beyond Gravity: A Statistical Model for Use in Population Projections*, Proc. Natl. Acad. Sci. USA **105**, 15269 (2008).

[12] M. C. González, C. A. Hidalgo, and A.-L. Barabási, *Understanding Individual Human Mobility Patterns*, Nature (London) **453**, 779 (2008).

[13] D. Brockmann, L. Hufnagel, and T. Geisel, *The Scaling Laws of Human Travel*, Nature (London) **439**, 462 (2006).

[14] M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora, *Understanding Mobility in a Social Petri Dish*, Sci. Rep. **2**, 457 (2012).

[15] T. P. Davies, H. M. Fry, A. G. Wilson, and S. R. Bishop, *A Mathematical Model of the London Riots and Their Policing*, Sci. Rep. **3**, 1303 (2013).

[16] F. Asgari, V. Gauthier, and M. Becker, *A Survey on Human Mobility and Its Applications*, arXiv:1307.0814.

[17] R. E. Park, *Human Migration and the Marginal Man*, Am. J. Sociol. **33**, 881 (1928).

[18] L. A. Sjaastad, *The Costs and Returns of Human Migration*, J. Polit. Econ. **70**, 80 (1962).

[19] Jokbo. `http://en.wikipedia.org/wiki/Genealogy_book#Korea`.

[20] S.-G. Lee, *Korean Family Names and Genealogies* (in Korean; the original Korean title: '한국의 성씨와 족보') (Seoul National University Press, Seoul, South Korea, 2003).

[21] H. A. T. Kiet, S. K. Baek, H. Jeong, and B. J. Kim, *Korean Family Name Distribution in the Past*, J. Korean Phys. Soc. **51**, 1812 (2007).

[22] S. K. Baek, H. A. T. Kiet, and B. J. Kim, *Family Name Distributions: Master Equation Approach*, Phys. Rev. E **76**, 046113 (2007).

[23] S. K. Baek, P. Minnhagen, and B. J. Kim, *The Ten Thousand Kims*, New J. Phys. **13**, 073036 (2011).

[24] P. A. Longley, J. A. Cheshire, and P. Mateos, *Creating a Regional Geography of Britain through the Spatial Analysis of*

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Surnames, Geoforum **42**, 506 (2011).

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

[26] World Urbanization Prospects: The 2014 Revision (Department of Economic and Social Affairs, United Nation), `http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf` (2014).

[27] M. Batty, *The Size, Scale, and Shape of Cities*, Science **319**, 769 (2008).

[28] M. P. Todaro, *A Model of Labor Migration and Urban Unemployment in Less Developed Countries*, Am. Econ. Rev. **59**, 138 (1969).

[29] D. Hoerder, *Cultures in Contact: World Migrations in the Second Millennium (Comparative and International Working-Class History)* (Duke University Press Books, Durham, North Carolina, 2002).

[30] M. J. Greenwood, *Internal Migration in Developed Countries* in *Handbook of Population and Family Economics*, edited by M. R. Rosenzweig and O. Stark (North Holland, Amsterdam, the Netherlands, 1997), pp. 647–720.

[31] R. Lucas, *Internal Migration in Developing Countries*, In: *Handbook of Population and Family Economics*, edited by M. R. Rosenzweig and O. Stark (North Holland, Amsterdam, the Netherlands, 1997), pp. 721–798.

[32] E. S. Lee, *A Theory of Migration*, Demography **3**, 47 (1966).

[33] B. Cushing and J. Poot, *Crossing Boundaries and Borders: Regional Science Advances in Migration Modelling*, Pap. Reg. Sci. **83**, 317 (2003).

[34] An original *jokbo* usually includes more details about clans, but we only use information that we mention explicitly.

[35] Google Maps, `https://developers.google.com/maps/`.

[36] In fact, it is known that clans that originated in the southern part of Korea have been more abundant in recent Korean history

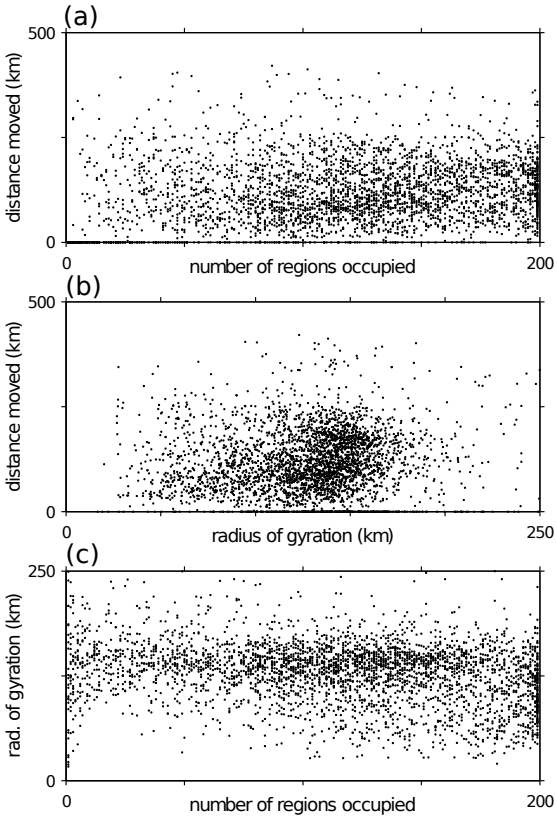

FIG. 18. Scatter plots of (a) distance between clan origin location and population centroid versus number of administrative regions, (b) distance between clan origin location and population centroid versus radius of gyration, and (c) radius of gyration versus number of administrative regions. The corresponding 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 is $p \approx 9.4 \times 10^{-5}$), 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 ($r \approx 0.39$, $p \approx 2.3 \times 10^{-4}$).

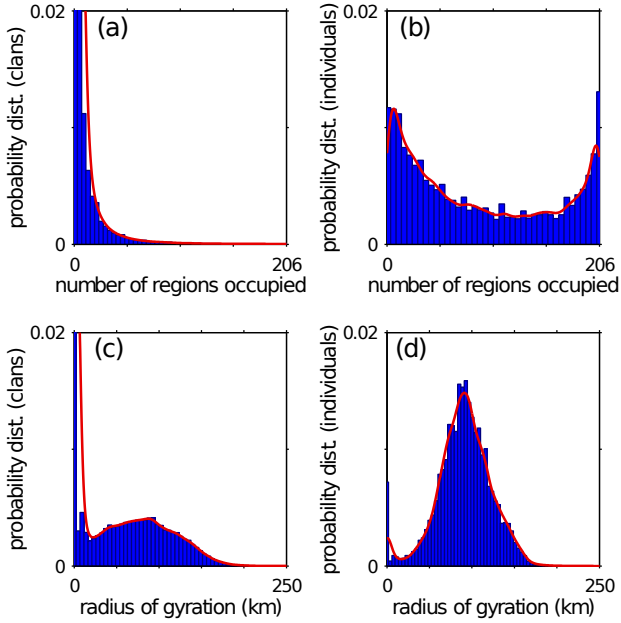

FIG. 19. (a) Probability distribution of the number of different administrative regions occupied a Czech family name in 2009. Note that the leftmost two bars have heights of 0.17 and 0.03 but have been truncated for visual presentation. (This data was initially analysed in Ref. [25].) (b) Probability 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. Solid curves are kernel density estimates (from MATLAB R2011a's `ksdensity` function with a Gaussian smoothing kernel of width 5).

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

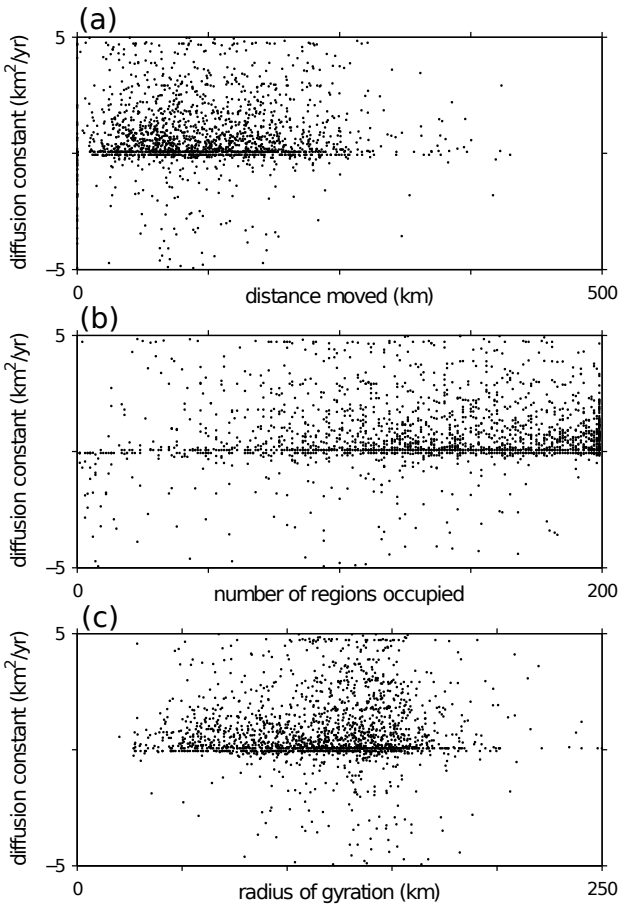

FIG. 20. Scatter plots of diffusion constants from the difference between the 1985 and 2000 census data versus (a) distance between origin and population centroid, (b) number of administrative regions, and (c) radius of gyration of clans in 2000. The corresponding Pearson correlation values are (a) $r \approx 0.13$ (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 $p \approx 1.0 \times 10^{-8}$).

(including the period that is spanned by our data set) [20].

[37] Korean Statistical Information Service, `http://kosis.kr/` (Korean version) and `http://kosis.kr/eng/` (English version).

[38] Statistical Geographic Information Service (통계지리정보서비스 in Korean), `http://sgis.kostat.go.kr/` (Korean version; no English version available).

[39] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters and Complexity* (Oxford University Press, Oxford, United Kingdom, 2006)

[40] J. van Lith, *Ergodic Theory, Interpretations of Probability and the Foundations of Statistical Mechanics*, Stud. Hist. Philos. Sci. **32**, 581 (2001).

[41] Because we only have ten family books, our data allows only ten distinct possibilities for "clan *j*" (the family into which a woman marries).

[42] S. Cho, *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, 1996).

[43] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, *A Universal Model for Mobility and Migration Patterns*, Nature (London) **484**, 96 (2012).

[44] A. P. Masucci, J. Serras, A. Johansson, and M. Batty, *Gravity versus Radiation Models: On the Importance of Scale and Heterogeneity in Commuting Flows*, Phys. Rev. E **88**, 022812 (2013).

[45] E. C. Young, *The Movement of Farm Population* (Cornell Agricultural Experiment Station Bulletin No. 426, Ithaca, New York, 1924).

[46] W.-S. Jung, F. Wang, and H. E. Stanley, *Gravity Model in the Korean Highway*, EPL **81**, 48005 (2008).

[47] S. Goh, K. Lee, J. S. Park, and M. Y. Choi, *Modification of the Gravity Model and Application to the Metropolitan Seoul Subway System*, Phys. Rev. E **86**, 026102 (2012).

[48] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, *Multiscale Mobility Networks and the Spatial Spreading of Infectious Diseases*, Proc. Natl. Acad. Sci. USA **106**, 21484 (2008).

[49] We did attempt to locate clan origin locations; see Fig. 11.

[50] S. Noh, S. Yang, and H. Kang, *Donggukyeojiseungram* (동국여지승람 in Korean) (Joseon Dynasty, 1481).

[51] One might call this "Thornthwaite's law" of human diffusion.

[52] A. J. Ammerman and L. L. Cavalli-Sforza, *Measuring the Rate of Spread of Early Farming in Europe*, Man. **6**, 674 (1971).

[53] J. Fort, *Synthesis between Demic and Cultural Diffusion in the Neolithic Transition in Europe*, Proc. Natl. Acad. Sci. USA **109**, 18669 (2012).

[54] In practice, we manipulate the anomaly data before computing radii of gyration. See the Appendix for details.

[55] M. Nei and J. Imaizumi, *Genetic Structure of Human Population. I. Local Differentiation of Blood Group Gene Frequencies in Japan*, Heredity **21**, 9 (1966); *Genetic Structure of Human Population. II. Differentiation of Blood Group Gene Frequencies among Isolated Populations*, *ibid* **21**, 183 (1966).

[56] C. Scapoli, E. Mamolini, A. Carrieri, A. Rodriguez-Larralde, and I. Barrai, *Surnames in Western Europe: A Comparison of the Subcontinental Populations through Isonymy*, Theor. Popul. Biol. **71**, 37 (2007).

[57] P. Rossi, *Surname Distribution in Population Genetics and in Statistical Physics*, Phys. Life Rev. **10**, 395 (2013).

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

[59] 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`.

[60] D. H. Ko *et al.*, *How People Lived in the Joseon Era* (in Korean; the original Korean title: '조선시대 사람들은 어떻게 살았을까') (Cheongnyeonsa, Paju, South Korea, 2005).

[61] Seoul: City Population History, `http://www.demographia.com/db-seoul-pop.htm`.

[62] M. T. Gastner and M. E. J. Newman, *Diffusion-Based Method for Producing Density Equalizing Maps*, Proc. Natl. Acad. Sci. USA **101**, 7499 (2004).

[63] Optimization and root finding (`scipy.optimize`). `http://docs.scipy.org/doc/scipy/reference/optimize.html`.

[64] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, New York, 1978), pp 206–208.

[65] Municipal level divisions in Administrative divisions of South Korea. `http://en.wikipedia.org/wiki/`

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Administrative_divisions_of_South_Korea#Municipal_level_divisions.

[66] UTF-16 (16-bit Unicode Transformation Format). http://en.wikipedia.org/wiki/UTF-16.

[67] Google Maps API licensing. https://developers.google.com/maps/licensing.

[68] The Google Geocoding API. https://developers.google.com/maps/documentation/geocoding/.

[69] 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.

[70] Administrative divisions of South Korea. http://en.wikipedia.org/wiki/Administrative_divisions_of_South_Korea.

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

[72] Usage Limits for Google Maps API Web Services. https://developers.google.com/maps/documentation/business/articles/usage_limits.

[73] pygeocoder. https://bitbucket.org/xster/pygeocoder/wiki/Home.

[74] Universal Transverse Mercator coordinate system. http://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system.

[75] utm 0.3.0: Bidirectional UTM-WGS84 converter for python. https://pypi.python.org/pypi/utm.

[76] How WGS 84 defines Earth. http://home.online.no/~sigurdhu/WGS84_Eng.html.

[77] S. Dutch, *Converting UTM to Latitude and Longitude (or Vice Versa)*, http://www.uwgb.edu/dutchs/UsefulData/UTMFormulas.htm.

[78] A. Morton, *UTM Grid Zones of the World*, http://www.dmap.co.uk/utmworld.htm.

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

[80] fminbnd: Find Minimum of Single-Variable Function on Fixed Interval. http://www.mathworks.co.uk/help/matlab/ref/fminbnd.html.

[81] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, United Kingdom, 2010).

[82] S. H. Lee, P.-J. Kim, Y.-Y. Ahn, and H. Jeong, *Googling Social Interactions: Web Search Engine Based Social Network Construction*, PLOS ONE **5**, e11233 (2010).

[83] R. G. Morris and M. Barthelemy, *Transport on Coupled Spatial Networks*, Phys. Rev. Lett. **109**, 128703 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

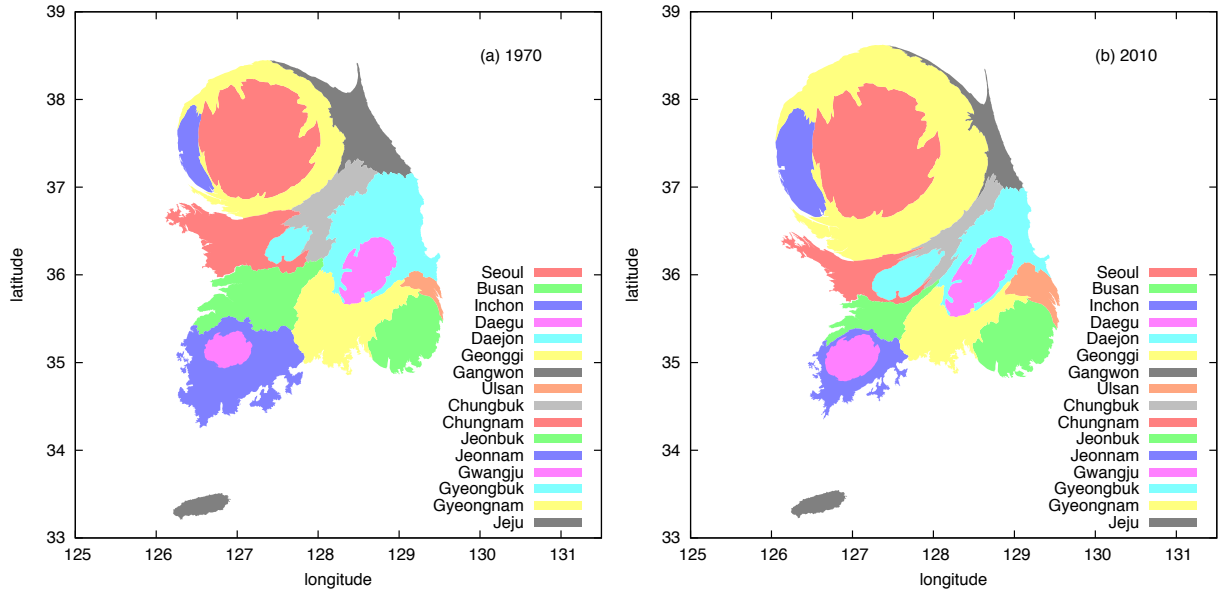

FIG. 21. Density-equalizing population cartograms [62] for South Korea using population data from (a) 1970 and (b) 2010 censuses [37]. The coordinates are longitude on the horizontal axis and latitude on the vertical axis. The growth of the Seoul metropolitan area over the past 40 years is clearly visible (compare this to a regular map of South Korea, such as the one in Fig. 1).

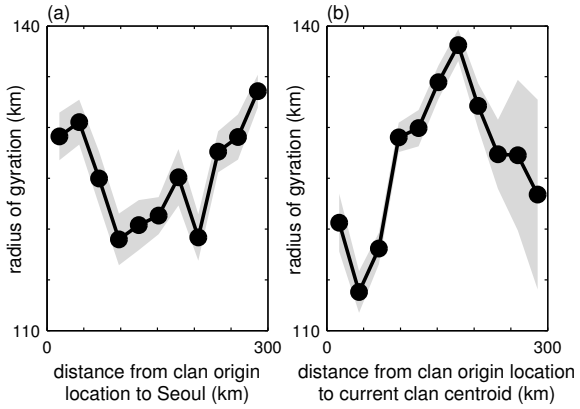

FIG. 22. Radii of gyration and distance scales of clans. (a) Radius of gyration r_g versus distance to Seoul. The Pearson correlation between the variables is not statistically significant ($r \approx 0.18$; the p-value is $p \approx 0.6$). (b) Radius of gyration r_g versus distance between clan origin location and present-day centroid. The Pearson correlation between the diagnostics is positive and statistically significant up to 170 km ($r \approx 0.86$, $p \approx 0.01$) and is negative and significant for larger distances ($r \approx -0.96$, $p \approx 0.005$). For all of the panels, we estimate r_g separately in each of 11 equally-sized bins for the displayed range of distances. The gray regions give 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/000257; this version posted August 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

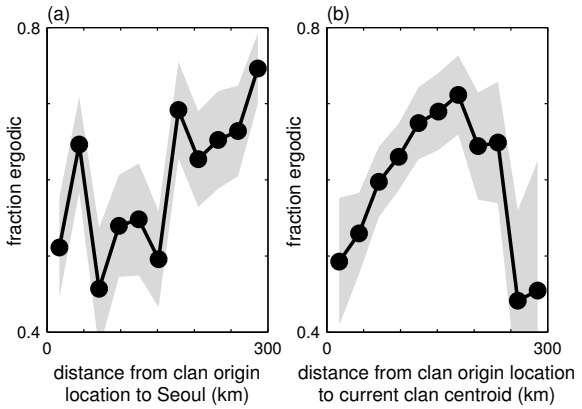

FIG. 23. Fraction of ergodic clans and distance scales of clans using only the clans that we used for the calculations in Fig. 22. We obtain the same qualitative result as in Fig. 5 in the main text. (a) Fraction of ergodic clans versus distance to Seoul. The correlation between the variables is positive and statistically significant. (The Pearson correlation coefficient is $r \approx 0.70$, and the p-value is $p \approx 0.02$.) For the purpose of this calculation, we call a clan "ergodic" if it is present in at least 150 administrative regions. (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 positive and significant up to 170 km ($r \approx 0.99$, $p \approx 0.0001$) and is negative and significantly for larger distances ($r \approx -0.92$, $p \approx 0.01$). For all of the panels, we estimate the fraction of ergodic clans in each of 11 equally-sized bins for the displayed range of distances. The gray regions give 95% confidence intervals.