Supplementary Information for "Socioeconomic factors of national representation in the global film festival circuit: skewed toward the large and wealthy, but small countries can beat the odds" by Karjus and Zemaityte

This appendix contains details on the simulation and a few more graphs, expanding the content that was explored in the main text.

Simulation procedure implementation details

This section supplements the Methods subsection concerning the simulation. The parameters of the simulation were as follows, with only the population and GDP per capita values being manipulated (across a uniform 14x14 grid of value combinations on the log scale).

- Grid resolution: $14 \times 14 = 196$ target combinations across the population and GDP per capita axes; see graph below;
- Population mean targets μ_{pop}: log-scale target mean of country population (varied across simulation grid from 1,000,000 to 330,000,000)
- GDP per capita mean targets μ_{gdp}: log-scale target mean of country GDP per capita (varied across simulation grid from 1,200 to 60,000)
- Standard deviation $\sigma_{\text{pop}} = \sigma_{\text{gdp}} = 0.3$: fixed standard deviation for both truncated log-normal distributions (log-scale)
- Truncation bounds $[a_{\text{pop}}, b_{\text{pop}}]$, $[a_{\text{gdp}}, b_{\text{gdp}}]$: minimum and maximum observed values (on log scale) for population and GDP per capita, inferred from the dataset
- Replicates per combination: 100 bootstrap replicates per parameter combination
- Initial subsample limit per country $n_{\max} = 20$: maximum number of films drawn per country before weighting
- Target sample size $n = 500$: intended number of films in the final sample (adjusted dynamically via density ratio r)

In each simulation replicate, a two-stage sampling procedure was applied. First, a near-uniform stratified subsample was drawn by selecting up to a maximum of $n_{\max}$ films per country, in order to mitigate the overrepresentation of high-output countries. For each unit in the resulting subsample, truncated normal weights were computed for population and GDP per capita using fixed simulation parameters. The population weight was calculated as:

$$w^{\text{pop}} = \frac{\phi\left(\frac{x-\mu_{\text{pop}}}{\sigma_{\text{pop}}}\right)}{\sigma_{\text{pop}}\left[\Phi\left(\frac{b_{\text{pop}}-\mu_{\text{pop}}}{\sigma_{\text{pop}}}\right) - \Phi\left(\frac{a_{\text{pop}}-\mu_{\text{pop}}}{\sigma_{\text{pop}}}\right)\right]}$$

where $x = \log_{10}(\text{Population})$, $\mu_{\text{pop}} = \log_{10}(\text{TargetPopulation})$, and σ_{pop} is the fixed standard deviation parameter. The GDP weights were calculated analogously using log-transformed GDP per capita values and corresponding parameters. ϕ and Φ denote the standard normal probability density and cumulative distribution functions, respectively.
Joint sampling weights were obtained by normalizing and multiplying the marginal weights:

$$w^{\text{joint}} = \frac{w^{\text{pop}}}{\sum w^{\text{pop}}} \cdot \frac{w^{\text{gdp}}}{\sum w^{\text{gdp}}}$$

To adjust for representativeness with respect to the target population and GDP distributions, a scaling factor r was computed as:

$$r = \frac{\sqrt{\left(\frac{\sum w^{\text{pop}}}{n} \cdot \frac{\sum w^{\text{gdp}}}{n}\right)}}{\sqrt{\bar{w}^{\text{ref}}_{\text{pop}} \cdot \bar{w}^{\text{ref}}_{\text{gdp}}}}$$

The reference densities used in the denominator were computed by evaluating truncated normal densities centered at the empirical means of the subsample. For population:

$$\bar{w}^{\text{ref}}_{\text{pop}} = \frac{1}{n}\sum_{j=1}^{n} \frac{\phi\left(\frac{x_j-\bar{x}}{\sigma_{\text{pop}}}\right)}{\sigma_{\text{pop}}\left[\Phi\left(\frac{b_{\text{pop}}-\bar{x}}{\sigma_{\text{pop}}}\right) - \Phi\left(\frac{a_{\text{pop}}-\bar{x}}{\sigma_{\text{pop}}}\right)\right]}$$

with $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$. The reference weights for GDP were computed in the same way using $y_j = \log_{10}(\mathrm{GDP}_j)$.
The final sample S, to be used in diversity calculations, was drawn with replacement using the joint weights w^{joint}, and the total sample size was scaled by the factor r to avoid drawing large samples from ranges of the parameter space where there are actually no or only a few countries:

$$S \sim \mathrm{sampler}\left(\{x_j\},\ \mathrm{size} = n \cdot r,\ \mathrm{weights} = \{w_j^{\mathrm{joint}}\}\right)$$

Additional simulation graphs

This section provides a few extra graphs to provide insights into the simulations, starting with the parameter grid.

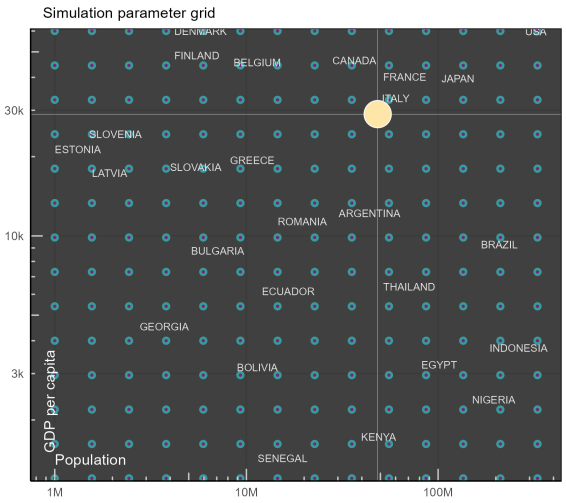

Figure S1: This graph complements the Method section and the simulation graphs, showing the parameter grid that is used to initiate the simulation. Each dot marks a parameter combination of population and GDP capita, used to create (log) normal distributions of sampling weights. Example countries are placed on the graph according to their average values on these axes in the dataset to illustrate the distributions, along with the position of the current dataset global average (larger yellow circle). The axes are bounded by the real ranges of the two variables in the dataset, with some padding to avoid sampling from the boundaries, but there are areas (e.g. bottom left) which are sparse or empty in reality, i.e. there are countries that are tiny and those which have low GDP per capita, but none which are both. That is accounted for in the simulation, which adjust sample size according to density, and the graphs, which show the empirical means of the resulting samples, not the initiating target values.

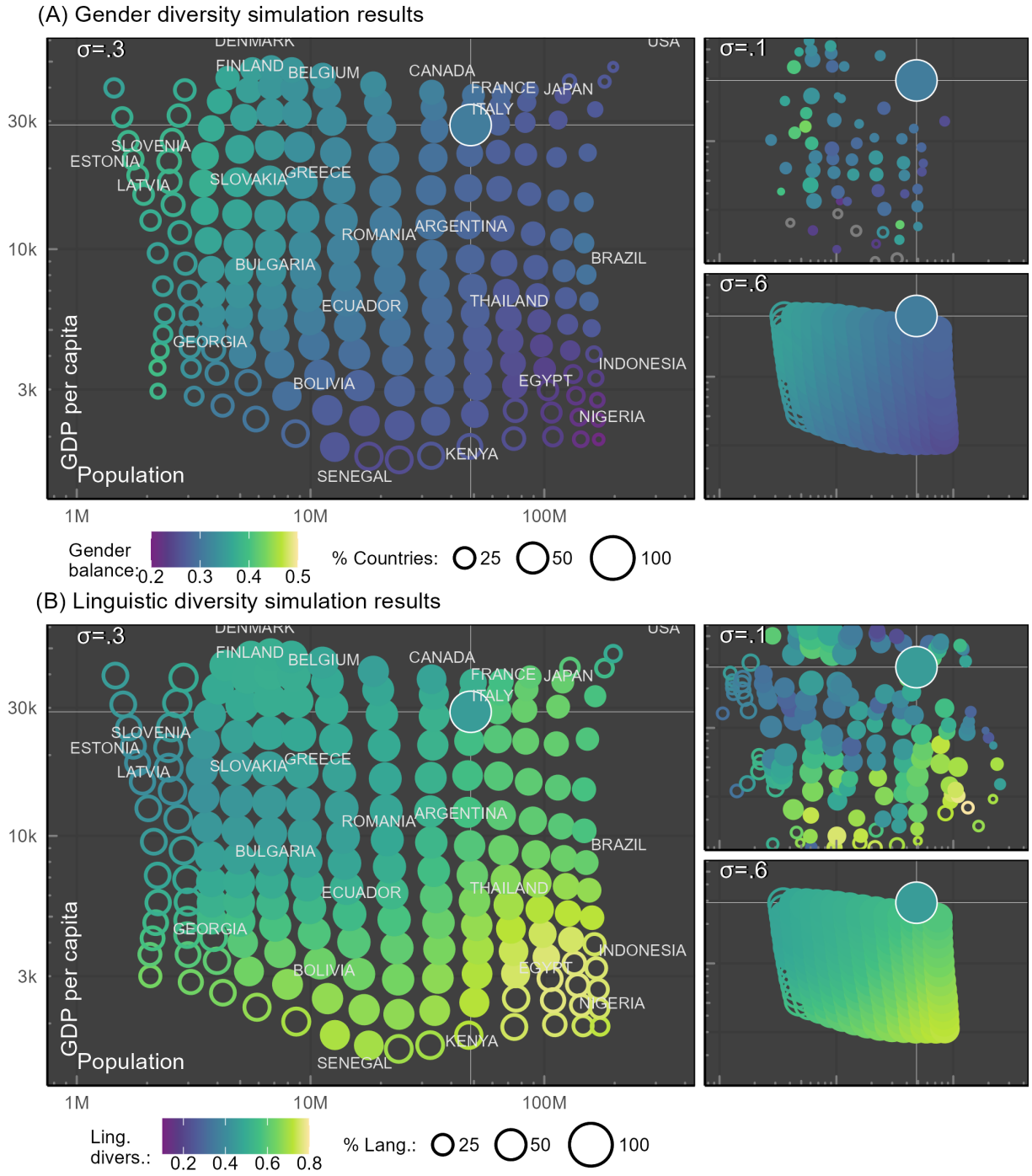

Figure S2: This graph complements the simulation section in the main text, showing results with two more values for the sampling distribution standard deviation parameter, 0.1 and 0.6 (note the log scales). The interpretation is otherwise exactly the same as the simulation figure in the main text, and the leftmost larger figure (for 0.3 value) reproduces the latter for comparison here. The color scales have been extended slightly to accommodate the wider range of values in the additional graphs and keep them comparable; hence the somewhat different colors in the central graph compared to the one in the main text. As expected, low SD leads to an even more distorted parameter grid, as the samples are drawn from more narrow ranges on the population and GDP axes, where some ranges have low or no support. High SD leads to the parameter grid constricting towards global average, as films from larger ranges of production country population and relative GDP are sampled. Compare this to the initial parameter grid above.

Additional explorative graphs

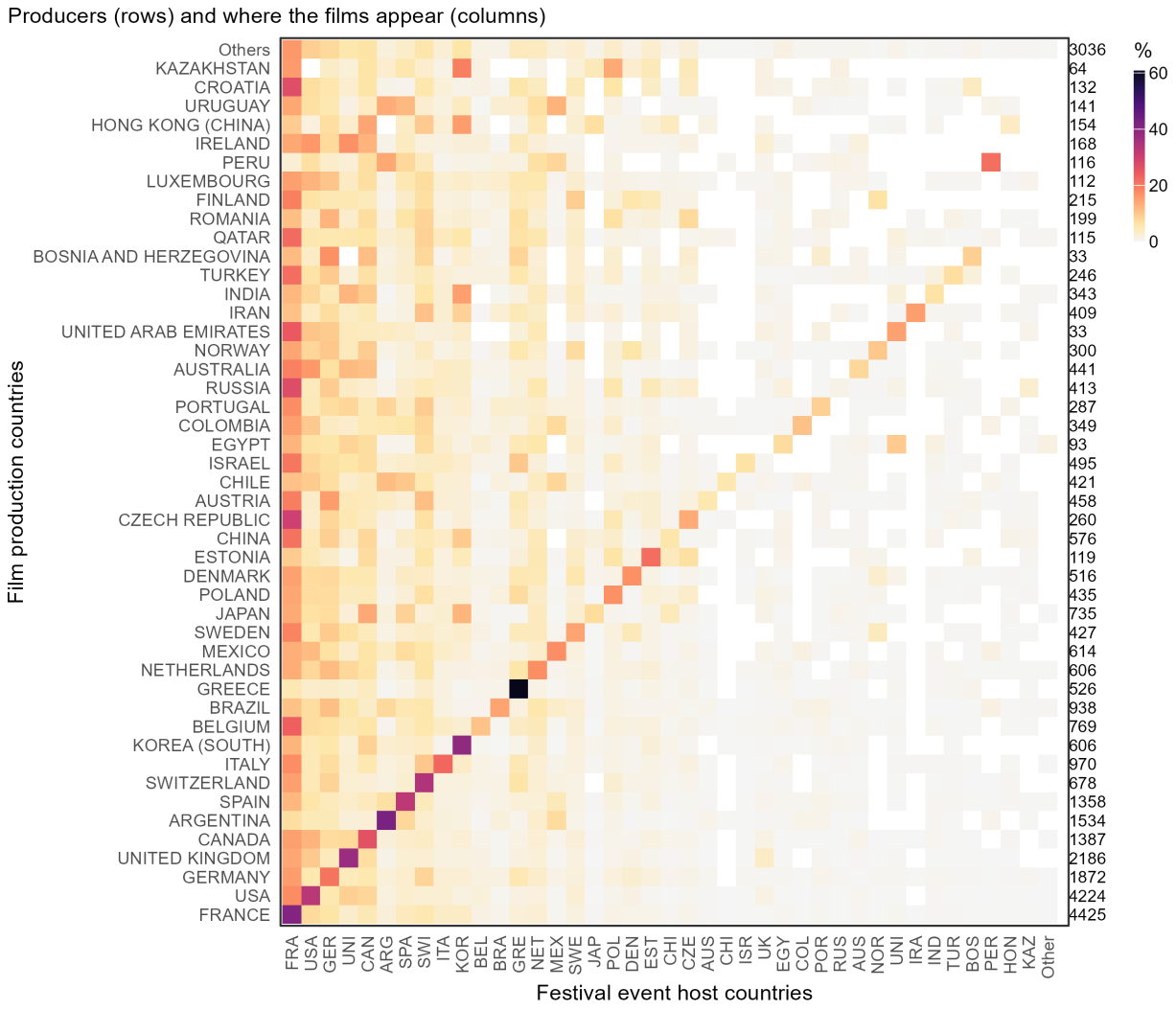

Figure S3: This graph complements the flows matrix in the main text, showing more countries, and the total (rounded) sums of production by the production countries (right side numbers column). The values differ slightly from the absolute values of database entry sums discussed in the text, being based on the co-production weighted values, as in the flow analysis. As in the main figure in the text, rows are producers and columns are host countries, and the cells, normalized row-wise, show the percentage of film-festival pairs in each host country.

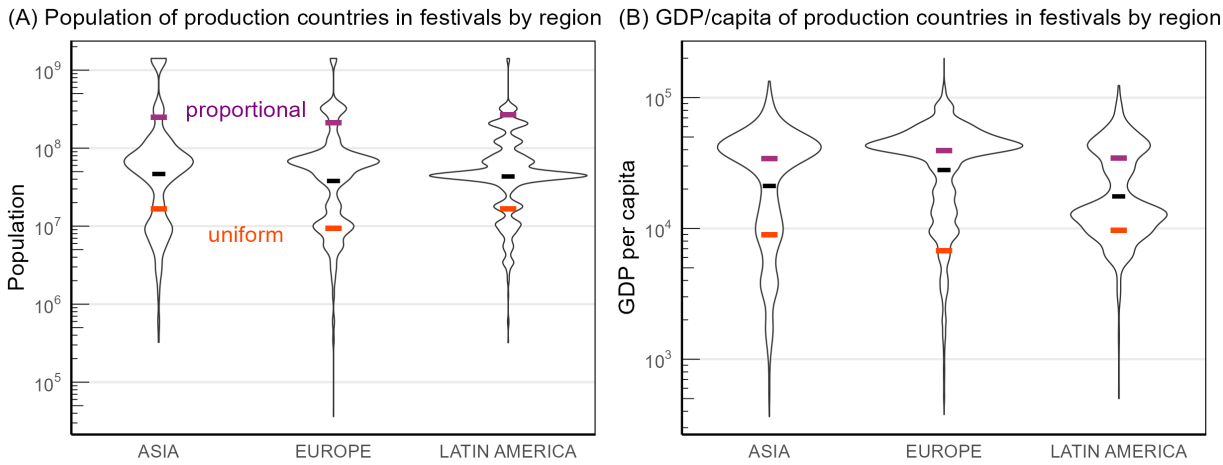

Figure S4: This graph complements the global distributions graph in the main text, showing the distributions and means across three regions, showing the average population (A) and GDP per capita (B) of films in festivals hosted in these regions (films coming from other regions are included).