The Novelty "Sweet Spot" of Invention

Yuejun He, Jianxi Luo

Appendix

Multivariable Regression Analyses

Robustness Checks

Novelty Profiles of Countries and Regions

Figures S1-S11

Table S1

1. Multivariable Regression Analyses

We ran a multivariable regression analysis to confirm the pattern found in Fig. 9. Table S1 presents the results. The variables *Central_10levels* and *Extreme_10levels* are 10 quantiles of the patents' central and extreme novelty. The fixed effects of the total number of references (backward citations) and the total number of co-inventors of a patent are included, whose impact on the value of invention has been implied by the literature. The variable *Ref_10levels* is 10 quantiles of the patents' reference counts. The variable *Team_type* equals to 1 when the patent has one single inventor, 2 when the patent has two inventors or 3 when the patent has more than two inventors.

When invention value (measured as normalized forward citation) is the dependent variable, we use OLS regression. When "*hit invention"* (whose value is either 0 or 1) is used as the dependent variable, logistic regression is used. In Model 1 and Model 3, the estimated coefficients of $Central_10levels^2$ are negative and statistically significant, which implies that the effects of central novelty on invention value follow a parabola that opens downwards. These results are consistent with the patterns in Fig. 9A and Fig. 9C. In Model 2 and Model 4, the estimated coefficients of $Extreme_10levels^3$ are positive and statistically significant, which suggests that the effects of extreme novelty on invention value follow a cubic curve going upwards. Such results are consistent with the patterns shown in Fig. 9B and Fig. 9D.

Table S1. Fixed effects regressions (heteroskedasticity-robust standard errors in parentheses).

Fixed effects: *Reference_10levels*, *Team_type*

	Invention Value (Normalized Forward Citation) OLS Regression		**Hit Invention (Top 5% Invention Value)** Logistic Regression	
Repressors	Model 1	Model 2	Model 3	Model 4
Central_10levels	0.0405** (0.0027)		0.1150** (0.0097)	
$Central_10levels^2$	-0.0026** (0.0002)		-0.0080** (0.0008)	
Extreme_10levels		0.1107** (0.0076)		0.2434** (0.0293)
$Extreme_10levels^2$		-0.0227** (0.0017)		-0.0510** (0.0058)

Extreme_10levels^3		0.0015** (0.0001)		0.0034** (0.0003)
Constant	0.6217** (0.0071)	0.5864** (0.0100)	-3.9368** (0.0298)	-3.9706** (0.0444)
Summary Statistics				
$\bar{R}^2$	0.0263	0.0267		
Pseudo R^2			0.0293	0.0309
Log likelihood			-115951.28	-115756.22
n	601715	601715	601715	601715

Legend: * p<0.05; ** p<0.01

2. Robustness Checks

Results using data subsets with different ranges of reference IPC4 pairs

In the main text, we focused on all the US patents granted in the 1990s with 5 or more references to earlier patents (i.e., backward citations) in the USPTO database. As a result, in the references of these patents, there are no fewer than 10 IPC4 class pairs. We also ran analyses on samples of patents with no fewer than 20, 30, or 50 IPC4 class pairs in their reference lists. Relevant results are reported in Figs. S2-S5. General patterns hold, despite subtle variations.

Results using alternative definitions of extreme novelty

In the main text, we used the additive inverse of the *z*-score at the minimum of the *z*-score distribution as the measure of extreme novelty of a patented invention. We also ran alternative analyses that use the 3rd, 5th, 8th, and 10th percentiles to define extreme novelty. Relevant results are reported in Figs. S2-S4, and show similar patterns in the central-extreme novelty matrices for both mean invention values and hit invention rates.

Results using alternative definitions of hit inventions

In the main text, we defined the patents with top 5% invention value as hit invention. We also ran alternative analyses that define hit inventions as those with top 10%, 5%, 3%, and 1% invention values. Results are reported in Fig. S5 and show a similar "sweet spot" in the central-extreme novelty matrices of different scenarios.

Results using alternative definitions of *z*-scores

In the main text, the *z*-scores in the distribution for a patent were calculated based on the historical data of co-occurrences of patent class pairs until the grant year of the focal patent, because central and extreme novelty of inventions are subjective to the past and present, and should vary temporarily as newer and newer technologies are developed over time. We alternatively calculated the *z*-scores based on the data of co-occurrences of patent class pairs in the single year when the focal patent was granted, i.e., only focusing on the technology combinations of the patents granted in the same year as the focal patent. Fig. S1 and Figs. S6-S10 display the results using the alternative calculations of *z*-scores. Compared with the results shown in Figs. S2-S5, the sweet spots in Figs. S6-S9 are similar despite blurry boundaries.

The generally consistent patterns found by alternative data sampling and analyzing settings suggest the qualitative findings reported in the main text hold.

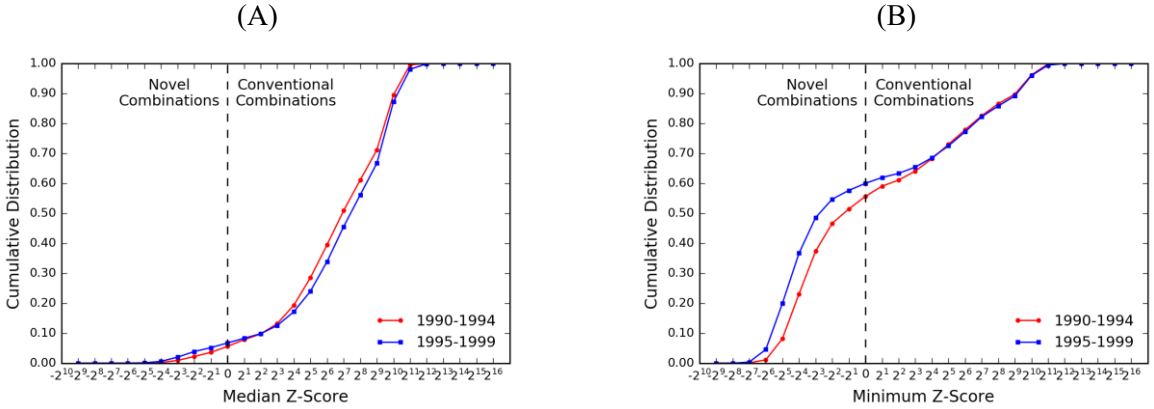


Fig. S1. Central and extreme novelty in invention. (A) Median *z*-score (i.e., additive inverse of central novelty) cumulative distribution of patents when *z*-scores are calculated based on the data in the granting year; (B) Minimum *z*-score (i.e., additive inverse of extreme novelty) cumulative distribution of patents when *z*-scores are calculated based on the data in the granting year.

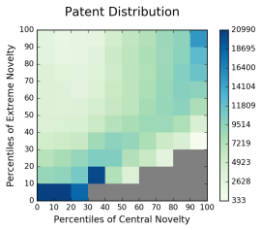
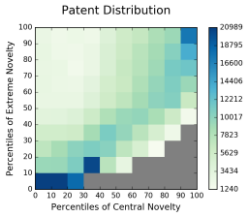
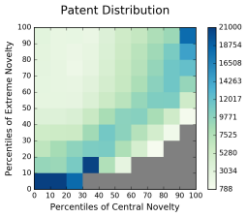
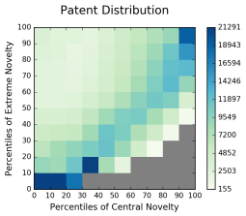
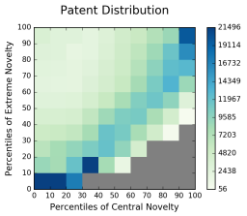
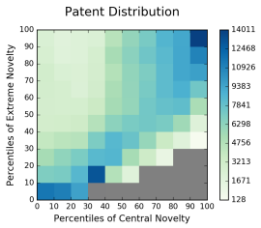
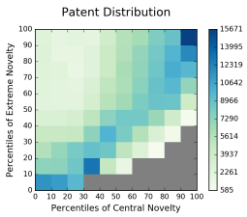
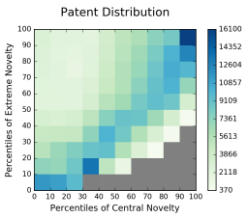
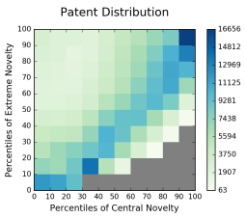
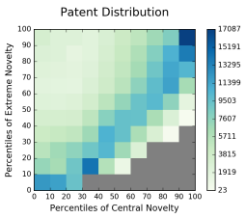
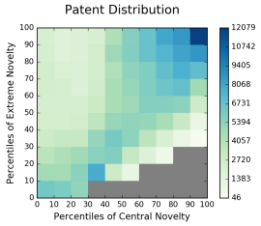
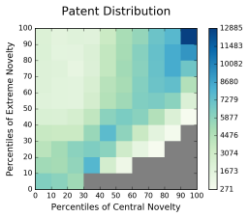
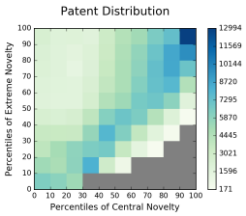
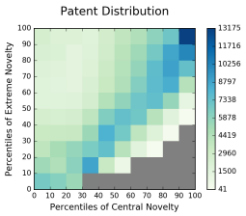
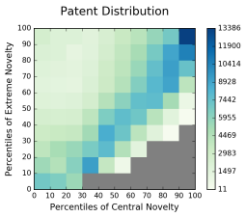
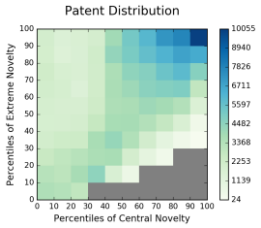
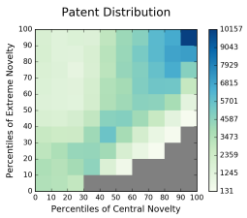
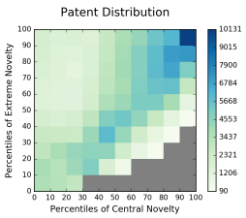
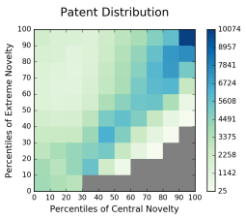
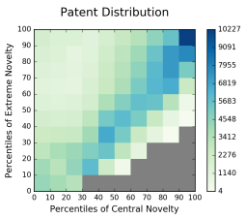
Reference IPC4 pairs	Patent Distribution by Alternative Extreme Novelty Definitions				
	minimum	3rd percentile	5th percentile	8th percentile	10th percentile
≥10					
≥20					
≥30					
≥50					

Fig. S2. Patent distribution in the central-extreme novelty matrices, based on alternative definitions of extreme novelty (Gray indicates no data).

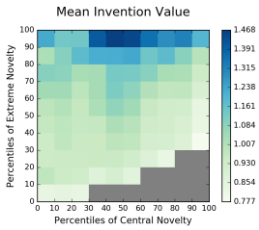
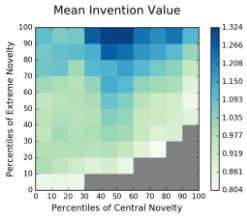
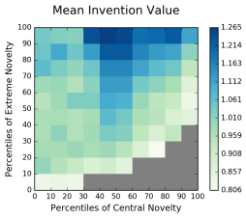
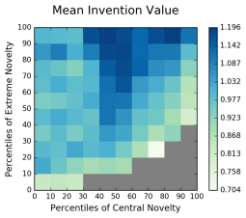
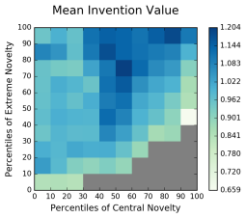
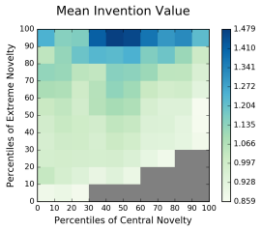
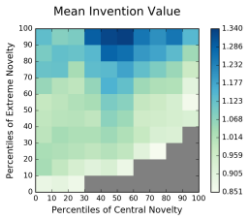
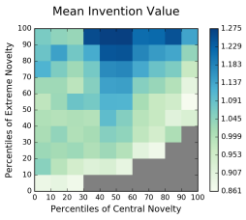
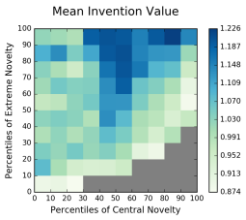
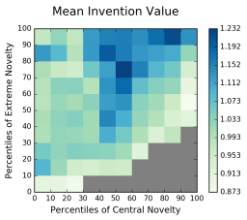
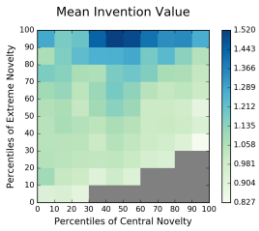
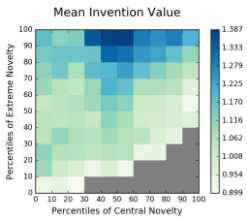
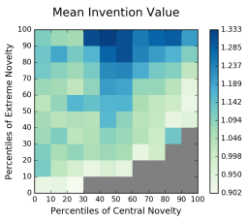
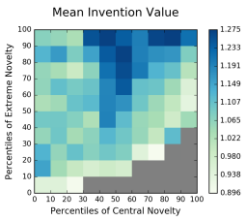
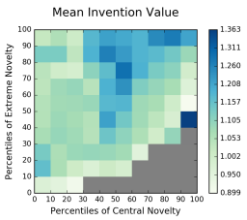
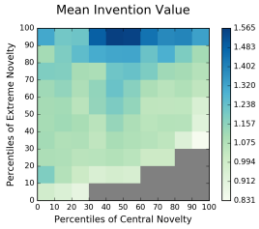
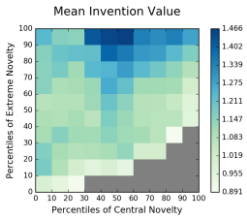
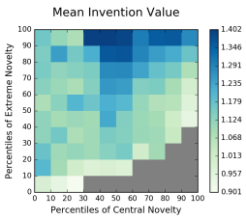
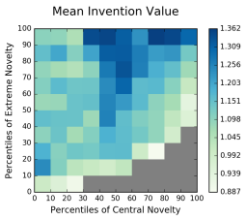
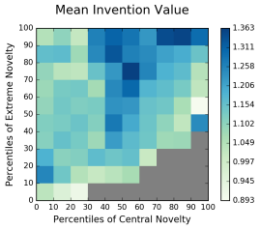
Reference IPC4 pairs	Mean Invention Value by Alternative Extreme Novelty Definitions				
	minimum	3rd percentile	5th percentile	8th percentile	10th percentile
≥10					
≥20					
≥30					
≥50					

Fig. S3. Mean invention values in different regions of the central-extreme novelty matrices, based on alternative definitions of extreme novelty (Gray indicates no data).

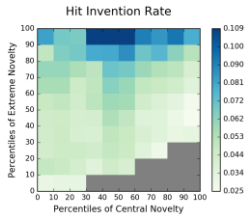
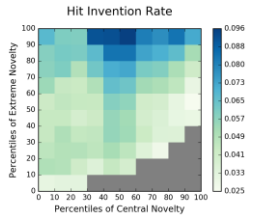
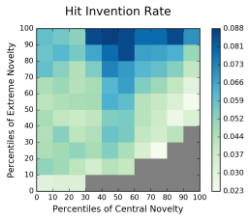
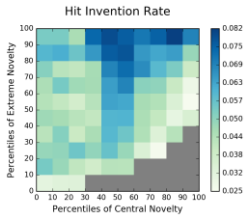
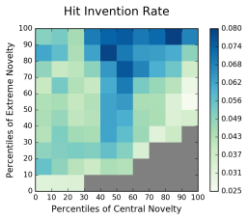
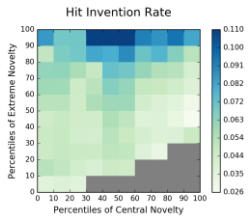
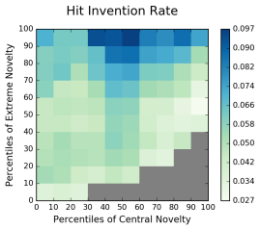
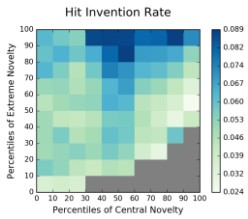
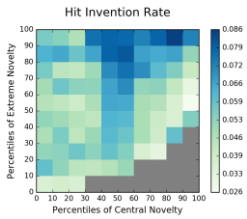
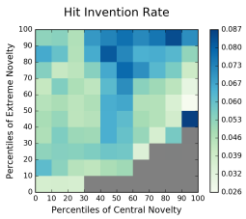
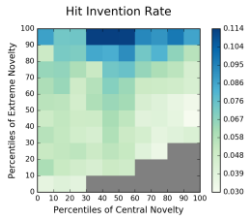
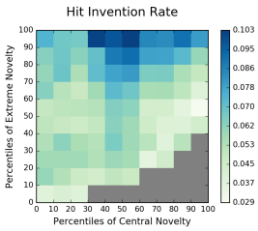
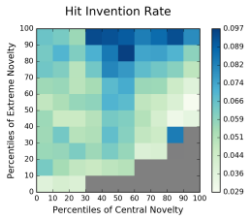
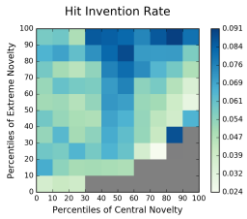
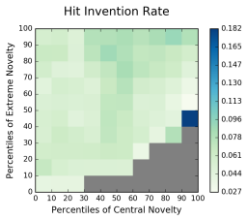
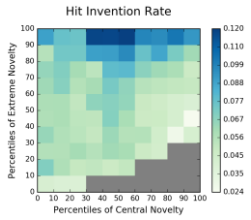
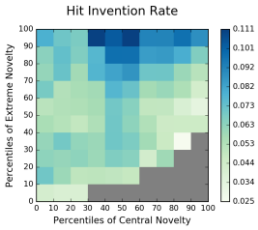
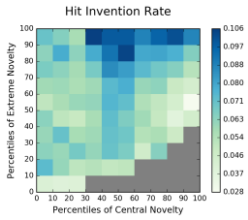
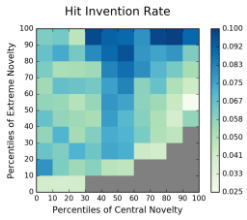
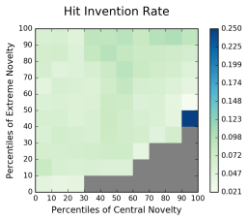
Reference IPC4 pairs	Hit Invention Rates by Alternative Extreme Novelty Definitions (A hit invention is defined as a top 5% patent in terms of normalized forward citation count)				
	minimum	3rd percentile	5th percentile	8th percentile	10th percentile
≥10					
≥20					
≥30					
≥50					

Fig. S4. Hit Invention rates in different regions of central-extreme novelty matrices, based on alternative definitions of extreme novelty (Gray indicates no data).

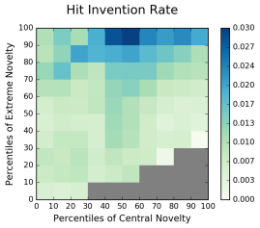
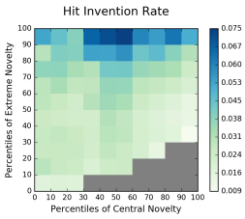
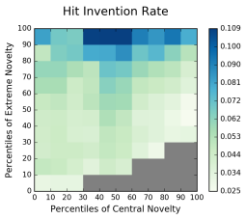
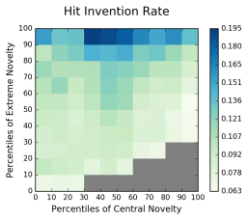
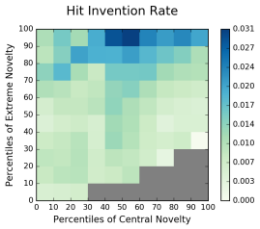
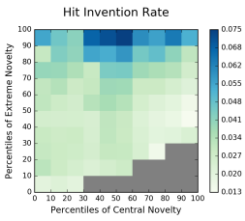
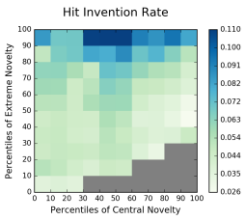
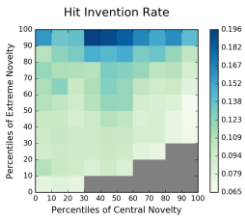
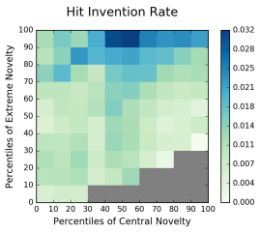
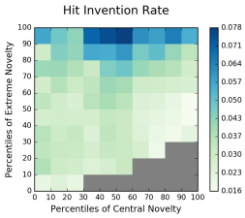
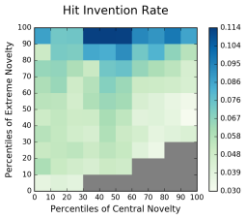
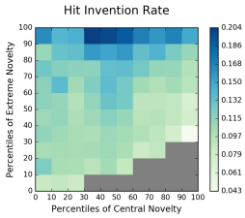
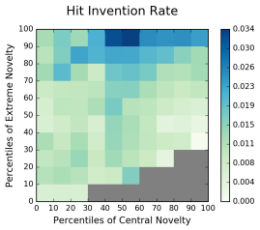
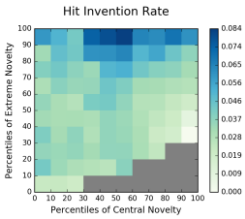
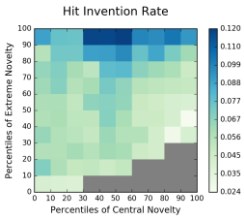
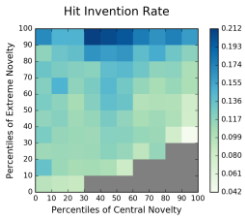
Reference IPC4 pairs	Hit Invention Rates by Alternative Definitions of Hit Inventions (Hit invention: in terms of the percentile of the normalized forward citation count; Extreme novelty: the additive inverse of the minimum of the *z*-score distribution)			
	Top 1%	Top 3%	Top 5%	Top 10%
≥10				
≥20				
≥30				
≥50				

Fig. S5. Hit invention rates in different regions of the central-extreme novelty matrices, based on alternative definitions of a hit invention (Gray indicates no data).

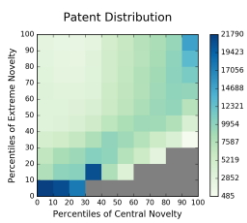
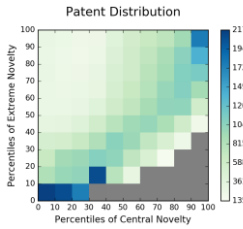
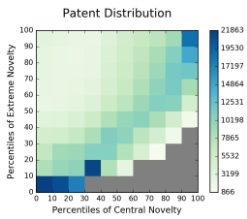
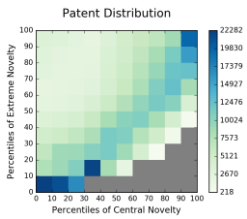
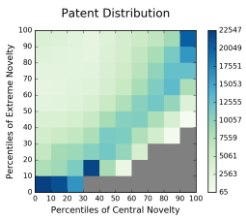
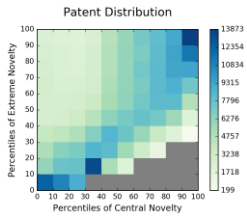
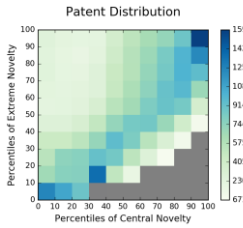
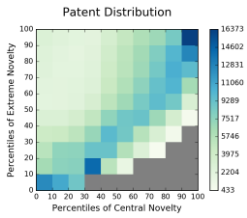
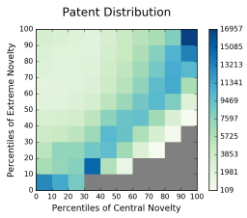
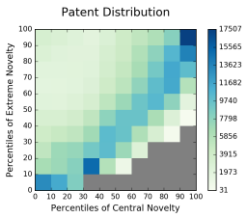
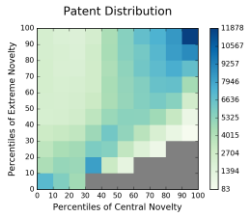
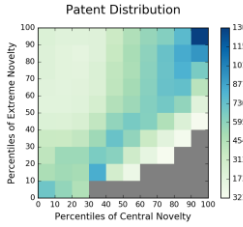
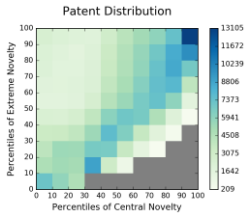
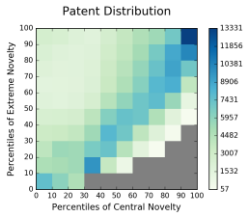
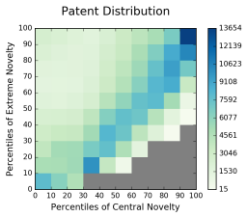
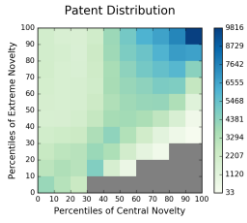
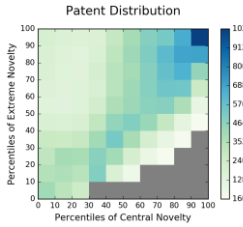
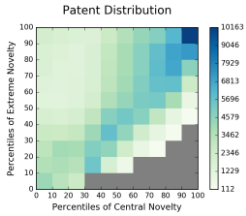
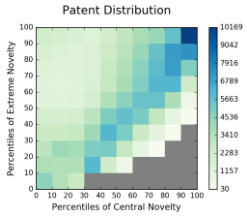
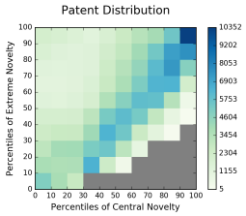
Reference IPC4 pairs	Patent Distribution by Alternative Extreme Novelty Definitions				
	minimum	3rd percentile	5th percentile	8th percentile	10th percentile
≥10					
≥20					
≥30					
≥50					

Fig. S6. Patent distribution in the central-extreme novelty matrices, when *z*-scores are calculated based on the data in the granting year of the focal patent (Gray indicates no data).

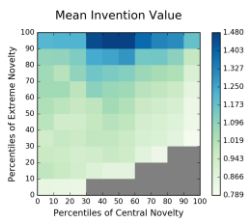
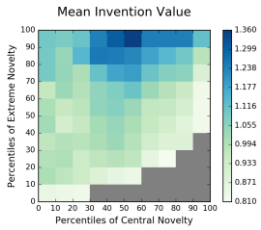
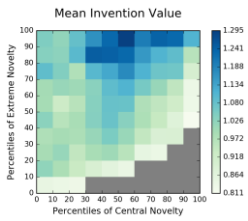
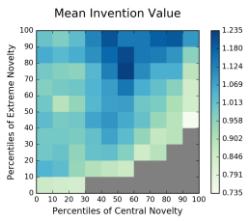
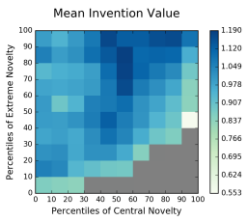
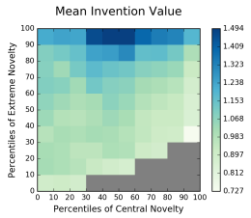
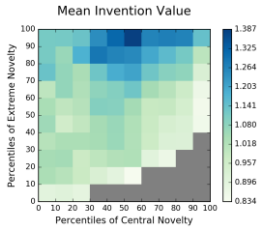
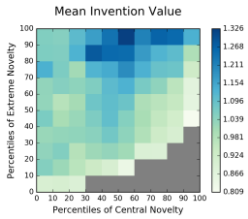
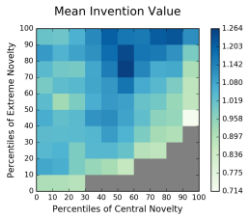
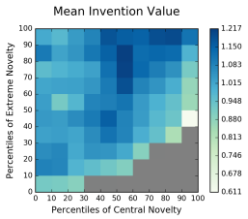
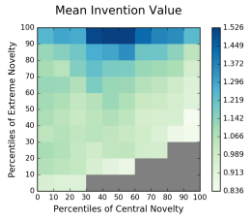
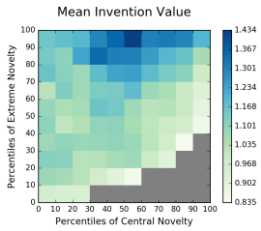
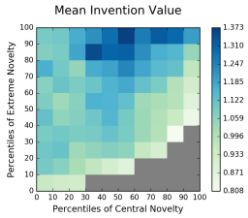
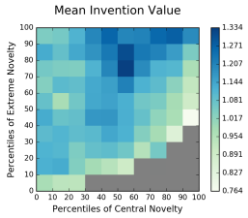
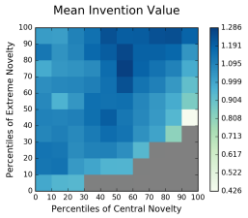
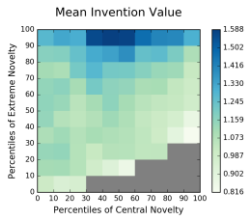
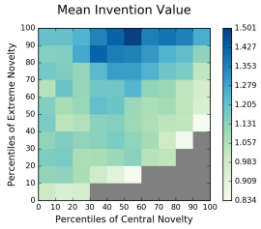
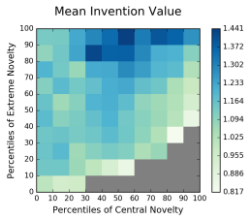
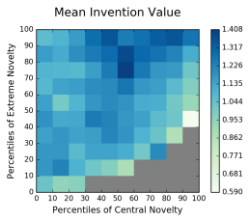
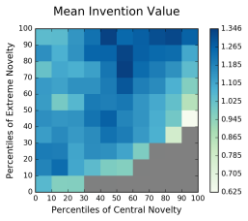
Reference IPC4 pairs	Mean Invention Value by Alternative Extreme Novelty Definitions				
	minimum	3rd percentile	5th percentile	8th percentile	10th percentile
≥10					
≥20					
≥30					
≥50					

Fig. S7. Mean invention values in different regions of central-extreme novelty matrices, when *z*-scores are calculated based on the data in the granting year of the focal patent (Gray indicates no data).

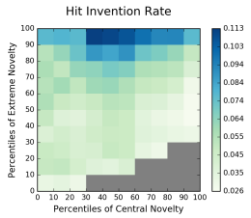
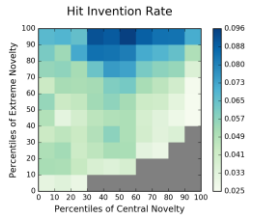
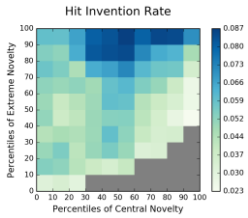
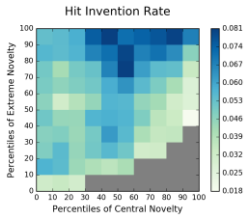
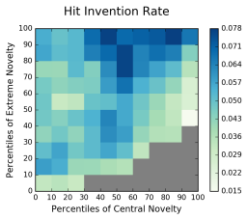
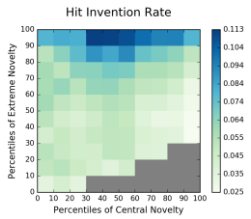
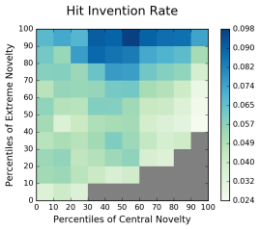
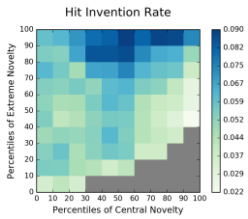
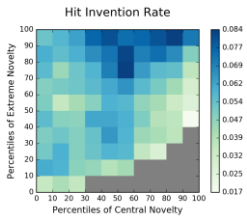
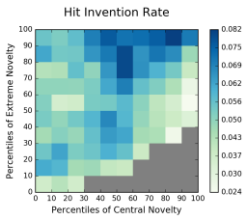
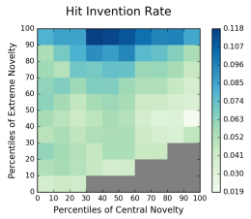
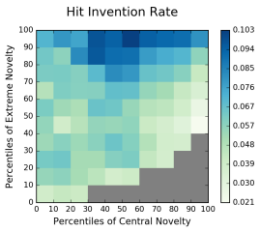
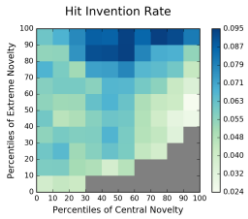
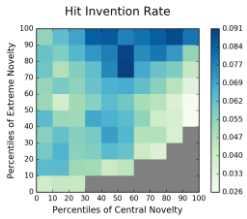
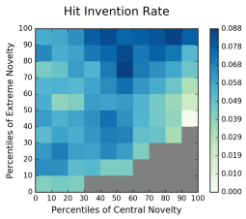
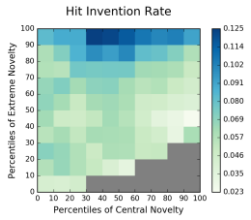
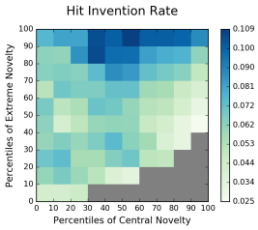
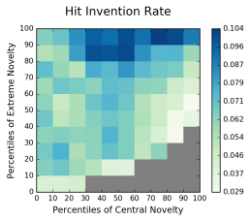
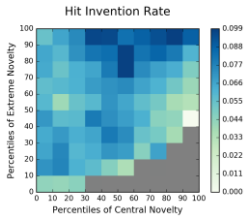
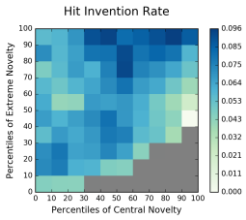
Reference IPC4 pairs	Extreme Value Probabilities by Alternative Extreme Novelty Definitions (A hit invention is defined as a top 5% patent in terms of normalized forward citation count)				
	minimum	3rd percentile	5th percentile	8th percentile	10th percentile
≥10					
≥20					
≥30					
≥50					

Fig. S8. Hit invention rates in different regions of central-extreme novelty matrices, when *z*-scores are calculated based on the data in the granting year of the focal patent (Gray indicates no data).

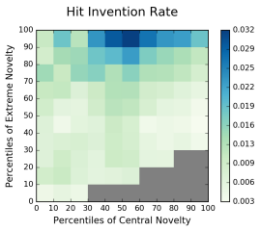
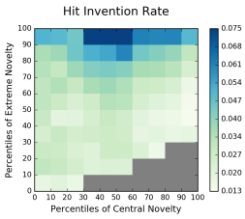
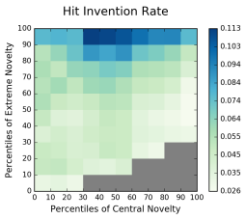
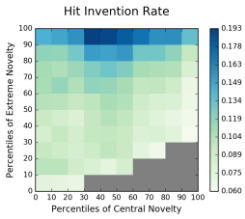
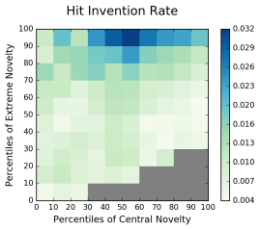
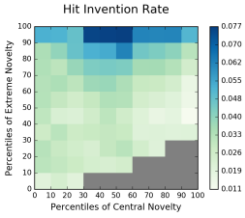
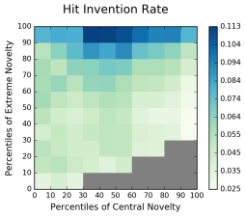
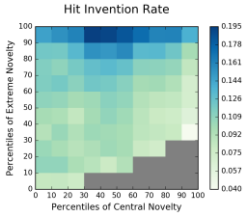
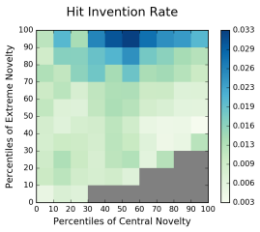
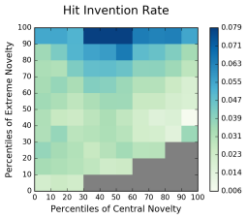
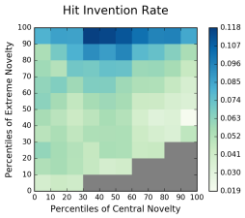
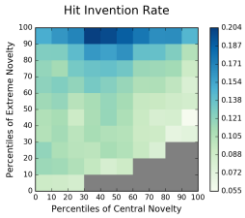
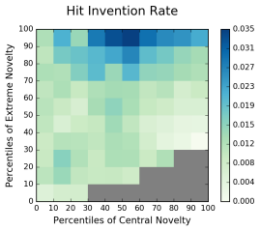
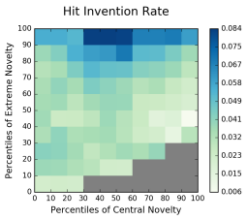
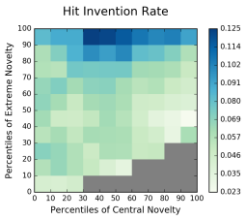
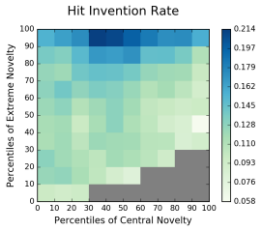
Reference IPC4 pairs	Hit Invention Rates by Alternative Definitions of Hit Inventions (Hit invention: in terms of the percentile of the normalized forward citation count; Extreme novelty: the additive inverse of the minimum of the *z*-score distribution)			
	Top 1%	Top 3%	Top 5%	Top 10%
≥10				
≥20				
≥30				
≥50				

Fig. S9. Hit invention rates in different regions of central-extreme novelty matrices, when *z*-scores are calculated based on the data in the granting year of the focal patent (Gray indicates no data).

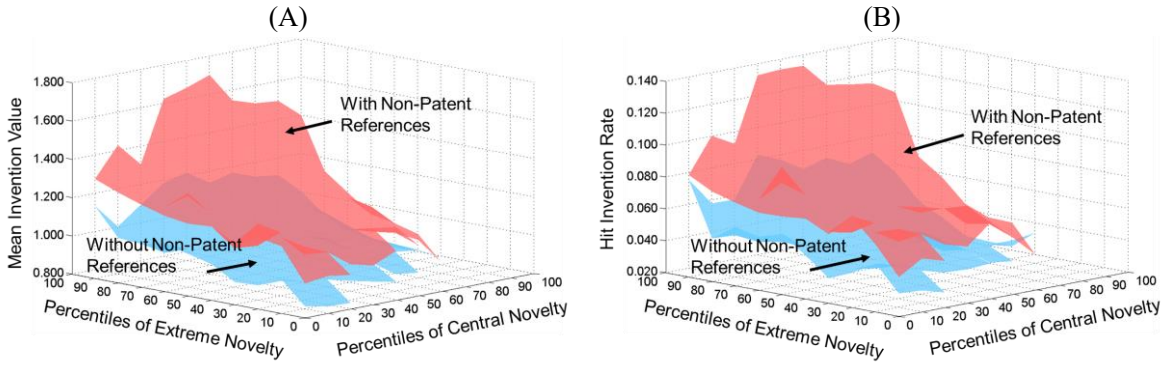


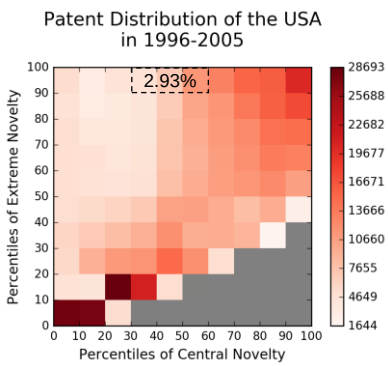
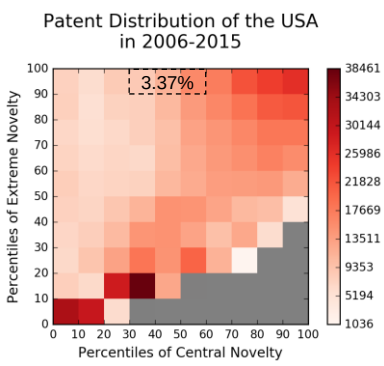
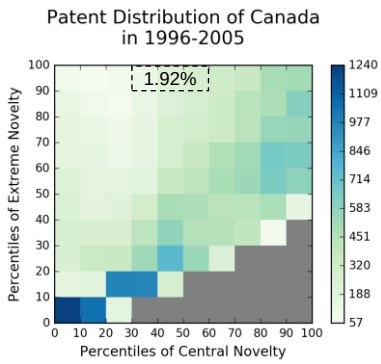
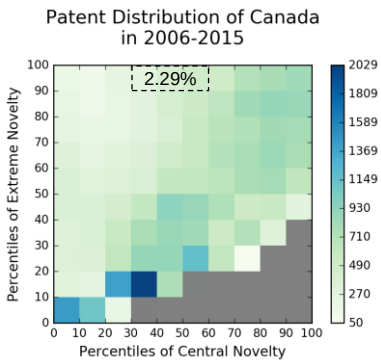
Fig. S10. Novelty sweet spot with and without non-patent references (e.g., scientific papers, technical reports, and books, etc.) by central and extreme novelty percentiles, when *z*-scores are calculated based on the data in the granting year of the focal patent. Each cell in the base matrix represents a category of patents. (A) Mean invention value, i.e., average normalized forward citation; (B) Hit invention rate, i.e., probability of top 5% invention value.

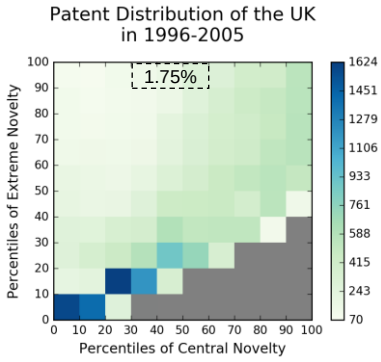
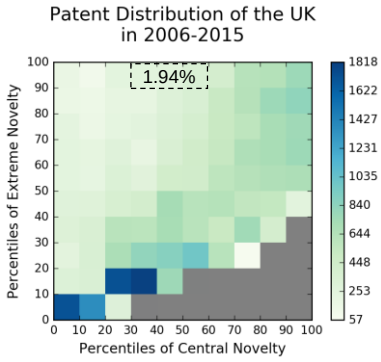
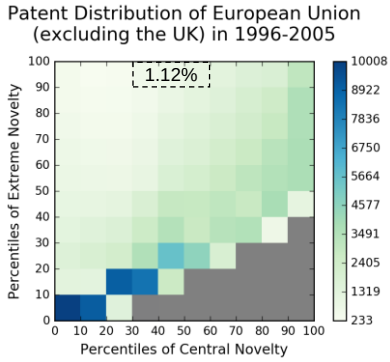
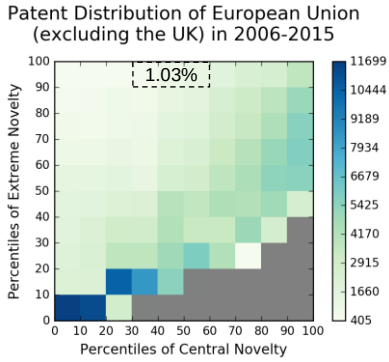
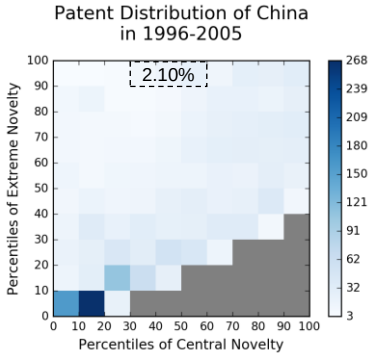
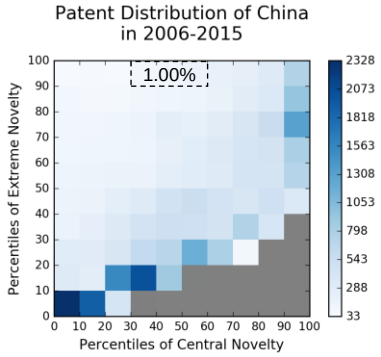
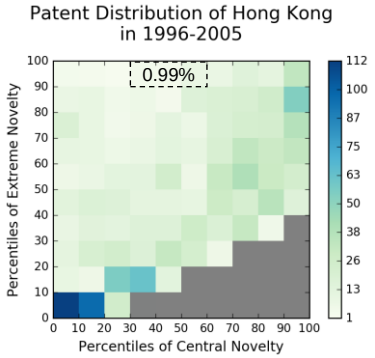
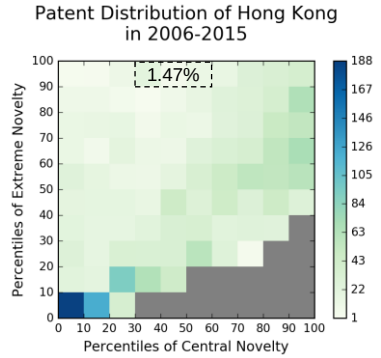
3. Novelty Profiles of Countries and Regions

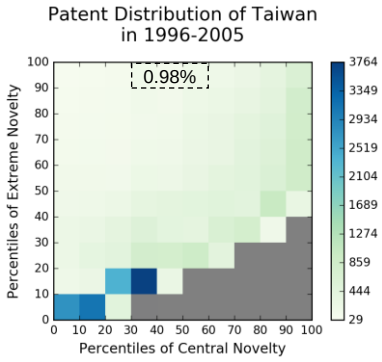
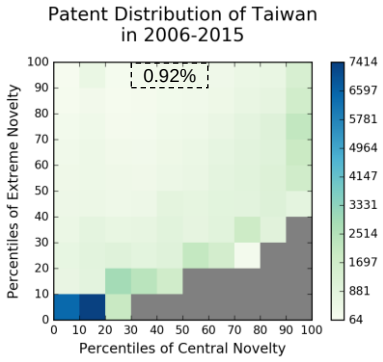
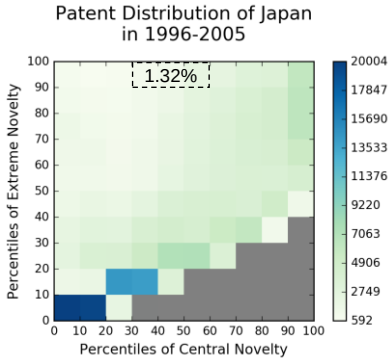
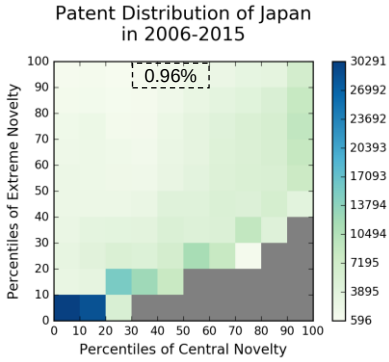
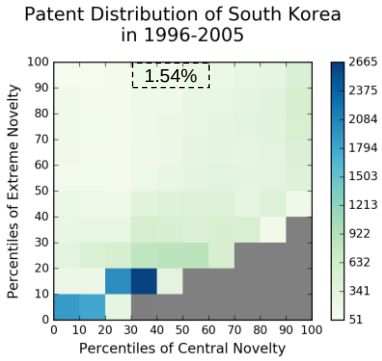
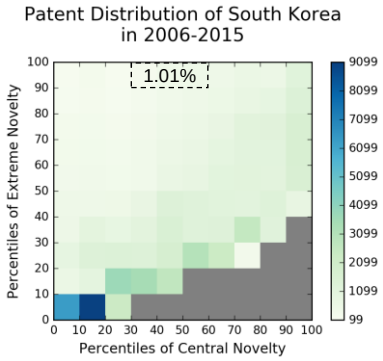
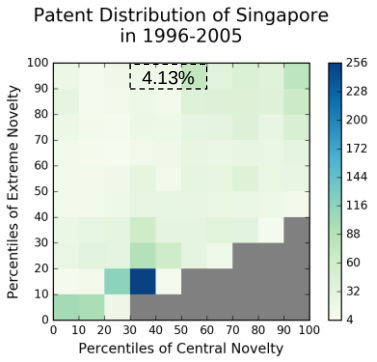
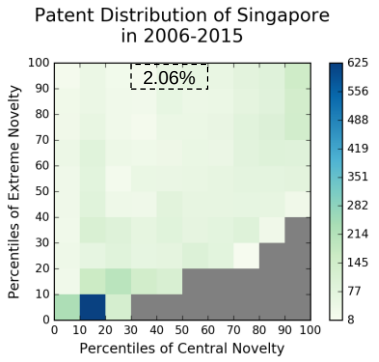
The central-extreme novelty matrix can be used to visually profile the two-dimensional novelty structures of patent portfolios for inventors at different aggregation levels, including persons, organizations, regions, etc. The desirable portfolio would have most patents concentrated in the sweet spot of the central-extreme novelty space. Fig. S11 shows more patent portfolio profiles for different countries and regions than the USA and China analyzed in the main text.

Several observations are noteworthy. First, in most cases, the highest concentrations of patents appeared at the bottom left corner in the matrix, where both central novelty and extreme novelty are low. However, the USA had relatively higher patent concentration at the upper right corner in the matrix, where both central novelty and extreme novelty are high. Second, the novelty profiles of most countries did not change obviously in the past two decades. Distinctively, China increased its concentration towards the upper right corner of the matrix, Singapore shifted its concentration towards the bottom left corner with low novelty in both dimensions, and ASEAN (excluding Singapore) decreased its concentration at the upper right corner. Third, the USA, Canada, the UK, Hong Kong, and ASEAN (excluding Singapore) increased their portions of patents in the sweet spot over time, while China, Taiwan, Japan, South Korea, Singapore, as well as the countries in Latina America and European Union (excluding the UK) dropped their portions of patents in the sweet spot.

The examples here show the utility of the central-extreme novelty matrix as a tool for assessing the novelty and estimating the value of the patent portfolios for countries and regions. This application may provide implications for national innovation policy. Such visual assessment can be conducted at different aggregation levels of patent portfolios for companies or individual inventors.

Geographical Region	Period	
	1996-2005	2006-2015
USA		
Canada		

UK	 
European Union (excluding the UK)	 
China	 
Hong Kong	 

Taiwan	 
Japan	 
South Korea	 
Singapore	 

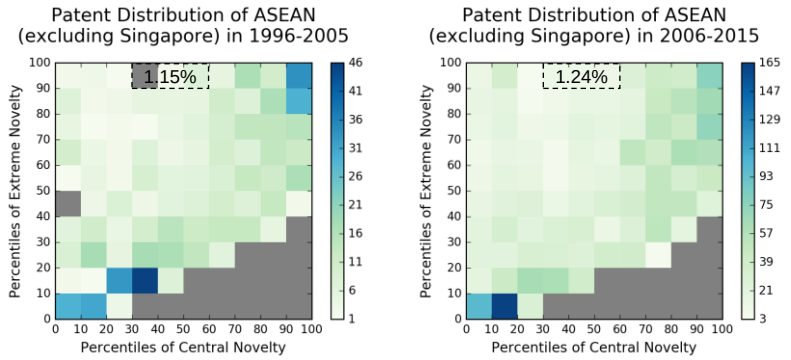
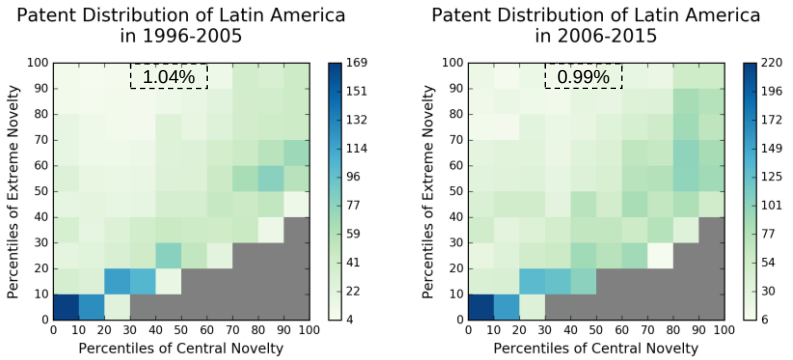
ASEAN (excluding Singapore)	
Latin America	

Fig. S11. Visual profiles of the patent portfolios of different countries and regions. Gray indicates no data. The numbers in the dashed boxes are the shares of the patents in the sweet spot, i.e., the region with 30th to 60th percentile of central novelty and 90th to 100th percentile of extreme novelty.