Impact of heterogeneity and socio-economic factors on individual behavior in decentralized sharing ecosystems

Supporting information

A. Gavaldà-Miralles, D. Choffnes, J. Otto, M. Sánchez, F. Bustamante, L.A.N. Amaral, J. Duch, R.Guimerà

Preprocessing of the Data. We obtain the decentralized content distribution data from a plug-in for the BitTorrent client Vuze. This plug-in, known as Ono, improves the download speed for the users while reduces the total overall network load. Ono was created by AquaLab at Northwestern University and, in May 2013, has been used by 1,469,187 worldwide users [1]. The plug-in has a built-in feature that, after user approval, assigns him or her a unique anonymous identifier and collects anonymous information about his/her activity in BitTorrent. This information is sent and stored in a centralized database.

The process of collecting information goes as follows: When the user starts the BitTorrent client, Ono collects the user identifier, IP address, geographical position and time. Every time a user starts a new download using BitTorrent, Ono collects the IP address, geographical position of this IP, time and file information (file size and file id). Since we are interested in assigning a download to a unique user identifier instead of the IP address behind it we built a method to map IP addresses to users ids. This method works as follows: (a) Using the information collected when the users start BitTorrent, we build a set of time series that tells for each IP address what user id was active at any time. (b) With these time series we can ask which user id that was active at the specified IP address at the time a download took place.

We apply this method to the more than 4 billion files downloaded by the plug-in users during March 2009. With the method we obtain a detailed list of the downloads made by 63,604 BitTorrent users. Since the user base of Ono grow continuously, in the analysis performed in this article we focus in what we define as the *active* users, that is, users for which we observe activity at least once the month before and at least once the month after. In March 2009 there were 9,783 of these active users which shared a total of 217,982 files and did 10,976,607 downloads during March 2009.

Naming the file size ranges with "The Pirate Bay" categories. We observe remarkable peaks in the file size distribution of the active users (Fig. 1A) which we use to delimit 7 file size ranges to study (Fig. 1B). To associate each file size range with its relevant content types so we can name it appropriately we collect the size and categories of torrents from the most popular torrent website [2], "The Pirate Bay". In particular we randomly sampled 456,949 torrents from this website. We observe that there are categories more popular than others, for example Video - Movies, Video - TV shows or Audio - Music are the most popular (Fig. S1).

For each file size range there are categories that appear more than expected by random chance (binomial, $p < 0.05$) and with a remarkable contribution to the total, larger than 10%. In Fig. 1B we color and name these categories and we represent in grey the ones that are not relevant, i.e. the ones which are smaller than 10% or which are expected by the null model. For example, between 401MB and 830MB, there are 3 categories with a contribution bigger than 10%. One of these categories is expected by the null model, therefore we color it in grey together with other categories that contribute less than 10%. Then we have two representative categories for file sizes between 401MB and 830MB: Videos (movies) and Porn (movies); Using this information, we summarize this file size range with the content type name "Movies low".

Comparing the distribution of file sizes of "The Pirate Bay" versus our dataset. Using the file size data of "The Pirate Bay" website described for the previous section, we also check if the distribution of file sizes that we observe in Fig. 1A is equivalent to the same distribution of the size of files available for downloading at this website. In (Fig. S2) we compare the histogram of the file sizes, and we observe that in both cases the peaks appear at very similar positions -14MB, 195MB, 400MB, 830MB, 1.65GB and 5.6GB-.

Alternative clusterings of the users keep our conclusions. In the manuscript we choose to cluster the users in 17 different user profiles according to the distribution of their downloads. If we select a number lower than 17 user profiles, it suddenly appears a large group that includes 3,500 users - more than 1/3 of all the active users (Fig. S3)-. As we increase the number of profiles, a lot of small user profiles that seem irrelevant appear. Therefore, based on the observation of Fig. S3, 17 user profiles seemed a fair number since is a balance between having a large user profile which takes more than one third of the data and too many small groups of user profiles.

Note that the conclusions of the paper remain the same for different number of user profiles. We repeated the analysis done in the main text but using 12 clusters (Fig. S4, S5) and using 22 clusters (Fig. S6, S7). The results presented in these figures show that the conclusions of the paper remain the same independently on the number of clusters used.

Effective number of contents. The effective number of contents E enables us to quantify the specialization of a user or group of users. The effective number of contents is based and equivalent to the inverse of the Herfindahl–Hirschman Index of Economy or the inverse of Simpson diversity index of Ecology [3]. To calculate E we use the frequency f_i with which the user or group of users download content of each type $i \in C$:

$$E = \frac{1}{\sum_{i \in C} f_i^2} \tag{1}$$

User profile prediction. To predict the profile of a user from their observed downloads, we consider two alternative approaches. The first approach is the complete Bayesian inference treatment for users that behave exactly according to one profile and have no correlations in their downloads. Under these assumptions, the predicted profile $\hat{u}$ is the one that maximizes the posterior over the existing user profiles U

$$\hat{u} = \arg\max_{\omega \in U} p(\omega \mid \mathbf{n}), \tag{2}$$

where $\mathbf{n}$ is a vector whose elements n_i represent the number of downloads of type i for the user under consideration, and

Reserved for Publication Footnotes

$p(\omega \mid \mathbf{n})$ is the probability that ω is the true profile of the user given $\mathbf{n}$.

Using Bayes theorem we have that

$$p(\omega \mid \mathbf{n}) \propto p(\mathbf{n}|\omega)p(\omega) = p(\omega)\prod_i (f_i^\omega)^{n_i} \tag{3}$$

where f_i^ω is the probability of downloading a file of type i for a user with profile ω. We set the priors $p(\omega)$ to be the overall abundances of each profile.

The second approach is heuristic and estimates the user profile $\hat{u}$ as the closest to the observed user behavior in terms of the cosine similarity

$$\hat{u} = \arg\min_{\omega \in U} \frac{\mathbf{n} \cdot \mathbf{f}^\omega}{\|\mathbf{n}\| \, \|\mathbf{f}^\omega\|} \tag{4}$$

where $\mathbf{f}^\omega$ is a vector whose elements f_i^ω are as defined above.

To compare the two methods for our data we plot their predictive accuracy as a function of the number of observed download (Fig. S6A). In this plot we see that the cosine method has a better performance for most of the specialist users. Also this method has a slightly lower performance for generalist users. More than 70% of the users are specialists so we choose the method that performs better on them, the cosine.

Next download prediction accuracy and model. We use a Bayesian model to predict the content type $\hat{c}$ of the next download of a user from the observed downloads $\mathbf{n}$ of that user:

$$\hat{c} = \arg\max_{\kappa \in C} \sum_{u \in U} p(\kappa|u)\, p(u|\mathbf{n}) \tag{5}$$

where we have that

$$\sum_{u \in U} p(\kappa|u)\, p(u|\mathbf{n}) \propto \sum_{u \in U} f_\kappa^u \, p(u) \prod_i (f_i^u)^{n_i} \ . \tag{6}$$

With our model, we can predict the next content type of 45% to 50% specialists. We can do that for 30% to 33% of the generalists.

Summary of the 9,783 active users data with absolute numbers. Our analysis is based on probabilities, therefore in Table S1 we give the absolute numbers of the users that we user in our analysis. Specifically, we show the number of users for each country for which we have more than 100 users and for each profile. We obtained the country of residence of 97,33% of the users. The most frequent countries are Spain with 1,060 users, USA with 1,035 users and France with 899. The most popular user profiles are Movies Low Definition (2,270 users), Generalist 2 (1,184 users) and Small (1,089 users).

Effective number of contents threshold between generalists and specialists. To chose the threshold between generalist and specialists, we compute the effective number of contents, E, for each profile (Table S2). Since we have 7 content types the maximum value of E is 7 (the user is focused on all the content types) and the minimum value is 1 (the user is focused on only content type).

Then define two different profiles classes based on the biggest separation of effective contents and the range of interests. We have generalists that have a broad range of interests, they are focused on at least 4 out of the 7 content types. And we have specialists that they are focused on at most 3 out of 7 content types. The separation between generalists and specialists is considerable, 0.73 units of effective contents. Specialists are more frequent and more diverse, whereas generalists represent an small part of the total users.

The main conclusions remain stable in almost 5 years of monthly data. The Ono plug-in has been collecting data for almost 5 years, between February 2009 to June 2013, with the exception of April 2012 to October 2012 when the collection servers had technical problems. During these 48 months, 543,577 active users shared 20,233,288 different files for which we could obtain the file sizes.

We study if the behavior of the users during March 2009 –the month used in the main text– is similar to 10 additional months of data, 5 randomly selected months amongst the 48 available (April 2009, December 2010, August 2011, February 2012 and May 2013) and 5 months selected periodically with a 6 month span period (October 2009, March 2010, October 2010, March 2011, October 2011).

For each of these months we identify the 'active users' (i.e. users that had activity the month before and the month after) and we repeat the analysis described in the main text. First, we analyze the distribution of the file sizes of the downloads and we observe that all of them have a similar distribution with peaks appearing at the same values – 14MB, 195MB, 400MB, 830MB, 1.65GB and 5.6GB– (Fig. S9). Second, we study the user profiles and we also observe in Figs. S10-S11 that the they remain fairly consistent over time, with users specializing mostly in one or two different content types. To confirm this similarity, in Fig. S12 we plot the distribution of the effective number of contents that the each user downloads, and we again observe that a large part of the users focus in a very small number of contents. Finally, we study the frequency of downloading each content type (Fig. S13) and we find again that it remains quite similar over time.

There is a significant correlation between user behavior and countries' socio-economic indicators. We quantitatively investigate whether there is a correlation between user behavior and five socio-economic indicators of the country where the user lives: GDP per capita (PPP in US Dollars 2011), number of Internet users per 100 people, number of broadband users per 100 people, payments per capita paid to other countries for the use of intellectual property and payments per capita received from other countries for the use of intellectual property (both in current US dollars).

Similarity between pairs of country profiles. We start by analyzing the similarity between pairs of country profiles. We present in Fig. 4A and Fig. S14 the country profile similarity, defined as the cosine similarity between the vectors of user profile z-scores, as a function of the absolute difference in each of the five socio-economic indicators, averaged over pairs of countries with a similar difference in the indicator. We calculate the Spearman ρ statistic for the observed pairs (S_{ij}, I_{ij}), where S_{ij} is the similarity between countries i and j, and I_{ij} is the absolute difference between countries i and j in socio-economic indicator I.

To establish the significance of the statistic (and given that not all points are independent, so that we cannot use directly the Spearman test) we bootstrap the values of the indicators for each country, and compute the p-value comparing the observed ρ to what one should expect from the bootstrapped samples ρ_{obs}. In Fig. S15 we show that the correlation is significant in all cases, the weakest being for number of broadband and internet users ($p < 10^{-2}$), and the strongest for payments made for intellectual property ($p < 10^{-4}$).

Correlation between fraction of user types and socio-economic indicators. We also study if the fraction of users in a country with a given profile is directly correlated with the socio-economic indicators. For each user profile we check using

Spearman's test if there is a correlation between the fraction of users of that profile in a country and each of the indicators. In Table S3, we present the p-values using Spearman's rank correlation test and we find that there are four types of user profiles that have a significant correlation with most of the indicators, "Small", "Small; Music", "Small; Movies LD" and "Movies LD". In Figs. 4B-D and S16 we present the correlations between three of these profiles versus the five indicators and the value of their significance. Interestingly, we find again that the number of internet users and broadband users are the least significant of all the studied socio-economic indicators.

Predictive models for the fraction of user types. The socio-economic indicators we consider (and most others that one may plausibly consider) are all highly correlated to each other, which poses difficulties to establish which variables are really responsible for the observed effects. The correlation analysis presented in the paragraph above (where each socio-economic indicator is considered separately) is the simplest and most agnostic and conservative approach, but of course has limitations in terms of the conclusions one can draw from it.

Therefore, we complete this analysis with a more in-depth model-selection analysis. In this analysis, we assume that the dependent variable (fraction of users of a given profile) is a linear combination of a subset S of all the socio-economic indicators. To identify the most predictive model, we try all possible subsets S of the socio-economic indicators and select the one with the smallest Bayesian information criterion (BIC) (Table S4).

This approach has the limitation of having to assume linear dependencies of the dependent variable on the socio-economic indicators, but it is easy to interpret in terms of the ability to predict the value of the independent variable—the model with the lowest BIC is (asymptotically) the one with the highest predictive power (that is, the one that would predict most accurately the fraction of users of a given profile in a country other than those we consider in our analysis).

Consistent with our previous analysis, we observe that the most predictive model always includes GDP, and only in one case adding other predictor variables improves the predictive power of GDP alone (Table S4).

1. Choffnes D, Bustamante F (2008) Taming the torrent: a practical approach to reducing cross-isp traffic in peer-to-peer systems. In ACM SIGCOMM Computer Communication Review (ACM), vol. 38, pp. 363–374.
2. Cuevas R, et al. (2010) Is content publishing in bittorrent altruistic or profit-driven? In Proceedings of the 6th International Conference (ACM), p. 11.
3. Simpson E (1949) Measurement of diversity. Nature 163:688.

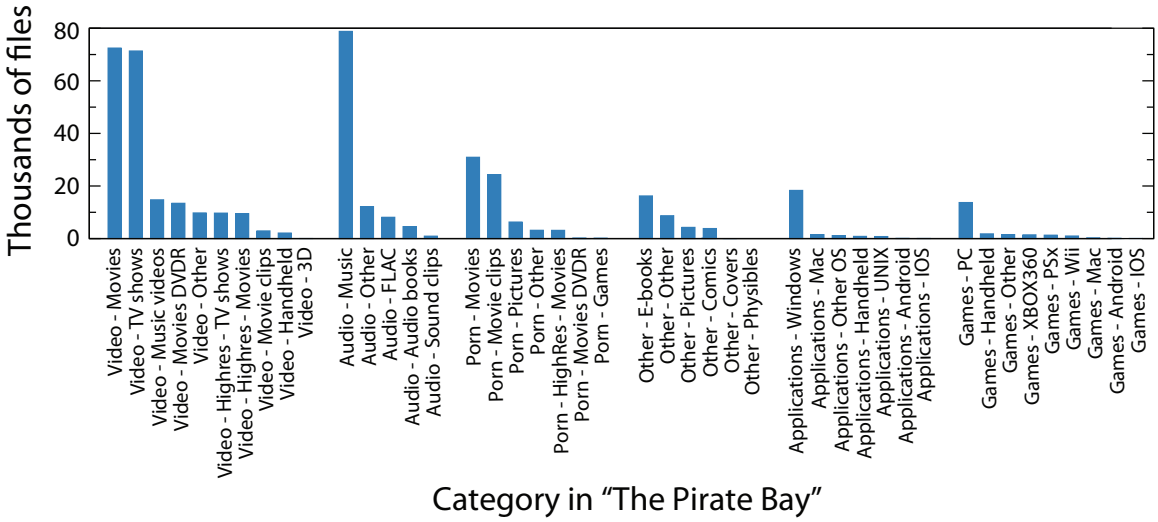

Fig. S1: Histogram of the categories from the most popular torrent site, "The Pirate Bay" [2]

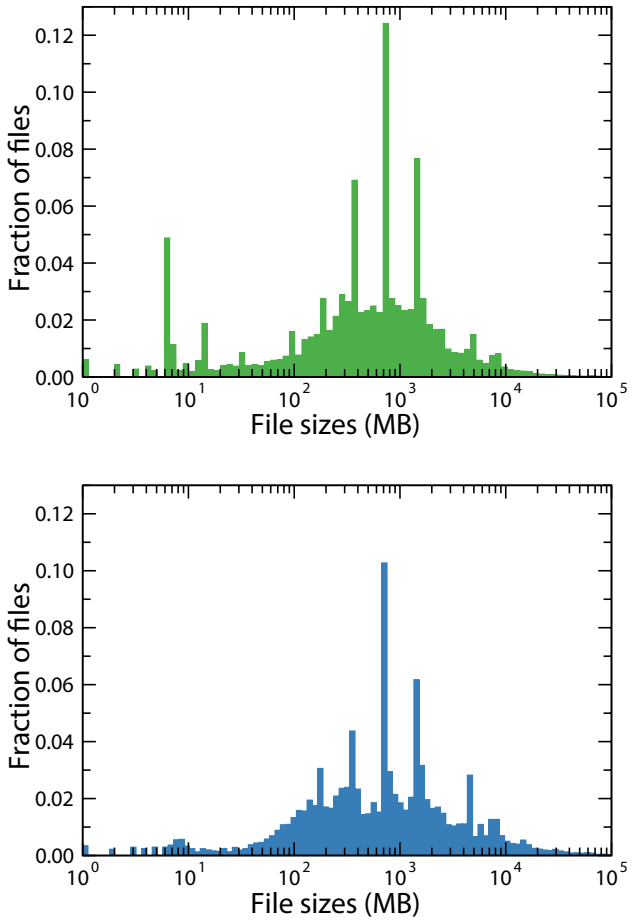

Fig. S2: The distribution of the file sizes in our dataset and the files available at the website "The Pirate Bay" are similar. In the top panel we show the histogram of the size of the files downloaded by the users in our dataset and in the bottom panel we show the "The Pirate Bay" file size histogram. Each file size is weighted by the total number of users that are uploading or downloading it. In both figures we observe the existence of peaks at or near 14MB, 195MB, 400MB, 830MB, 1.65GB and 5.6GB.

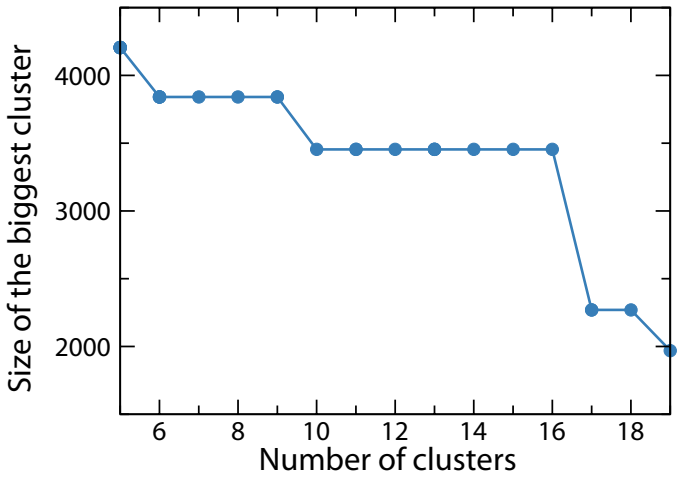

Fig. S3: Size of the largest user profile group for each number of user profiles. With more than 17 user profiles, we obtain small user profiles that are irrelevant. With less than 17 user profiles we obtain a large cluster which contains almost 3500 users, one third of the total data.

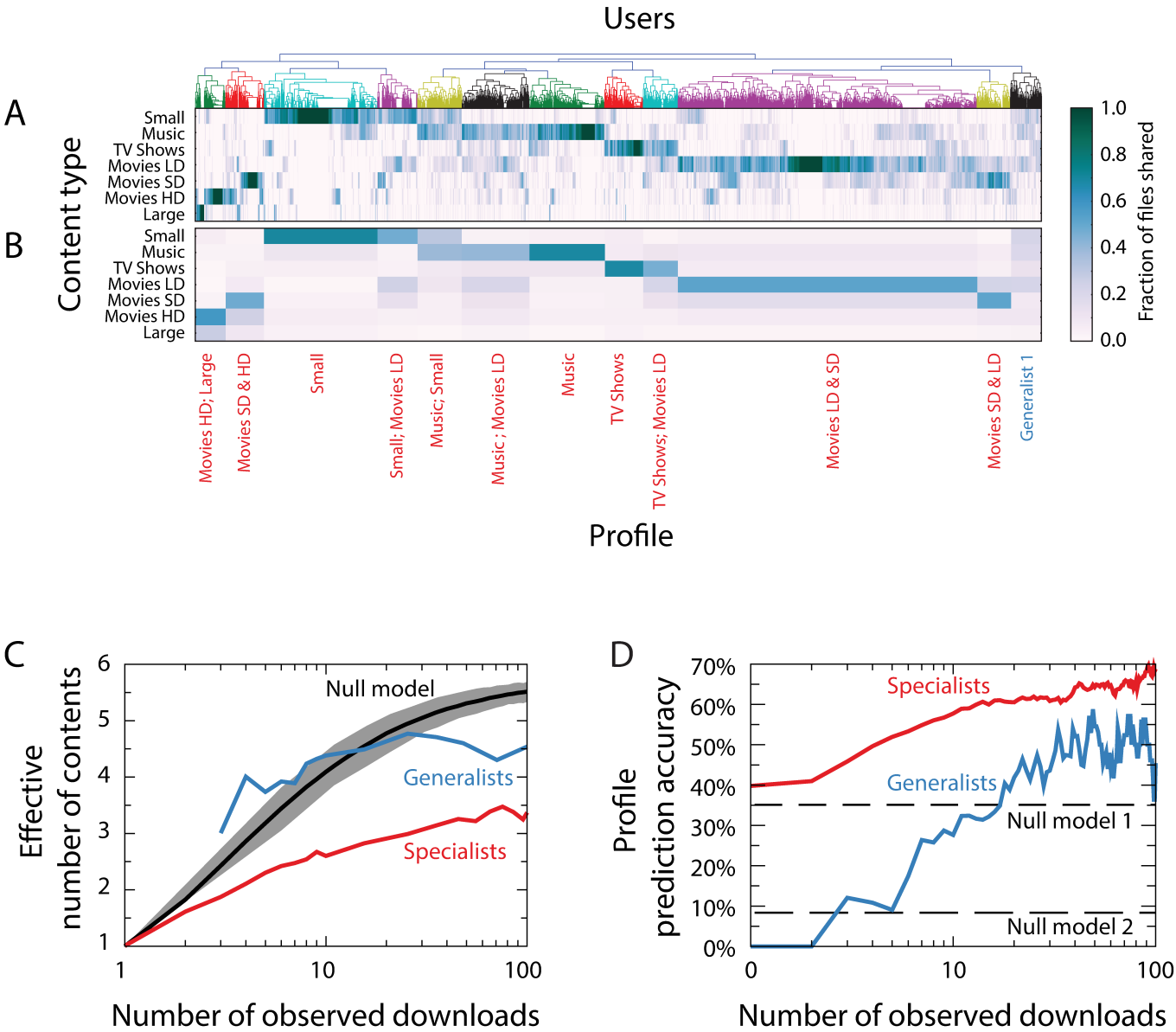

Fig. S4: Same as Fig. 2 of the main text with 12 user profiles instead of 17. (A) With less clusters generalists are blurred inside close specialists. (B) Then specialists become more generalists and the remaining generalist are quite close to the hypothetical average user. (C) This causes an increase of approximately 0.5 in the effective number of contents of both user types. (D) The user profile prediction accuracy is quite similar to the one with 17 clusters. With less generalists the user prediction accuracy fluctuates more, specially, with 20 downloads or more.

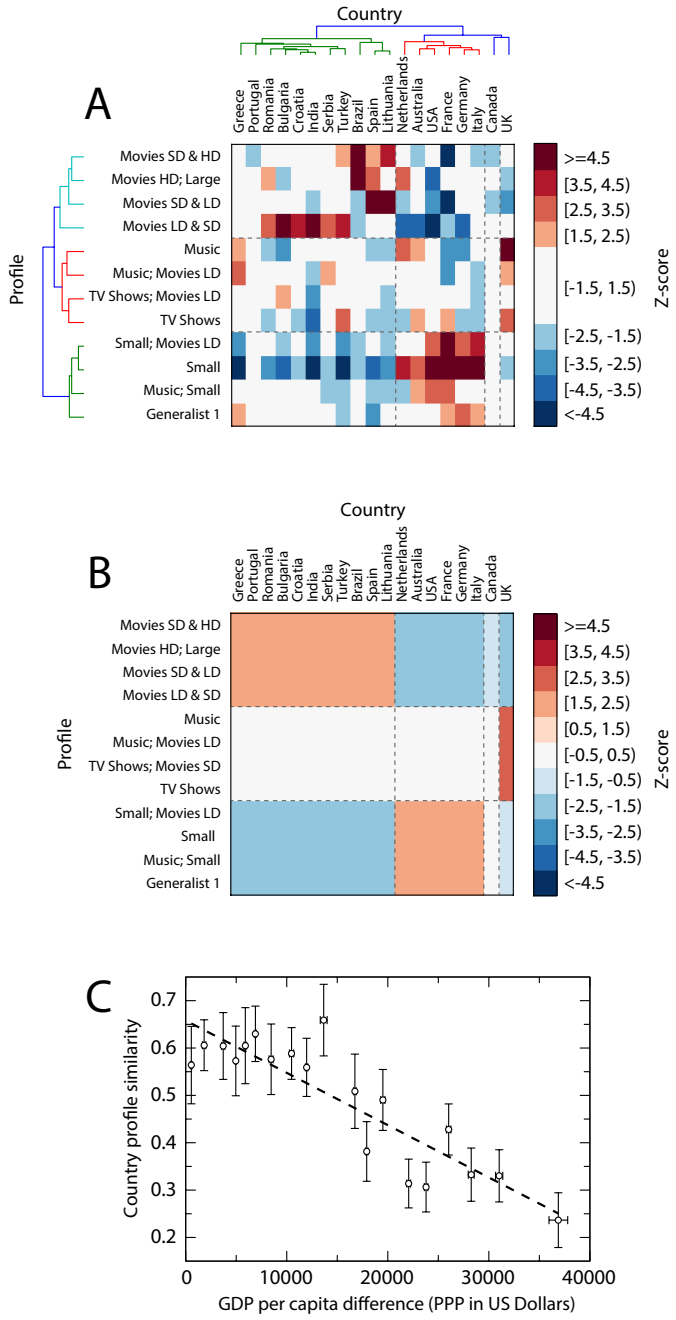

Fig. S5: Same as Figs. 3 and 4A (for GDP) of the main text with 12 user profiles instead of 17. (A) The movies categories are now condensed into fewer ones. Brazil, Greece, Lithuania, and Spain join the group of mostly low & middle economic level, which tend to share more movies than the expected and less small files. The cluster of the rich ones remains the same. Canada and UK users keep themselves apart by displaying an almost random uniform sharing behavior or by focusing mostly on music and TV shows respectively. (B) When we average the profiles by clusters, we see even more clearly the pattern that users living in countries with low GDP per capita tend to share more movies than the expected and less small files. Users in countries with higher GDP per capita do the opposite. (C) The country profile similarity vs GDP per capita trend is even steeper than with 17 clusters.

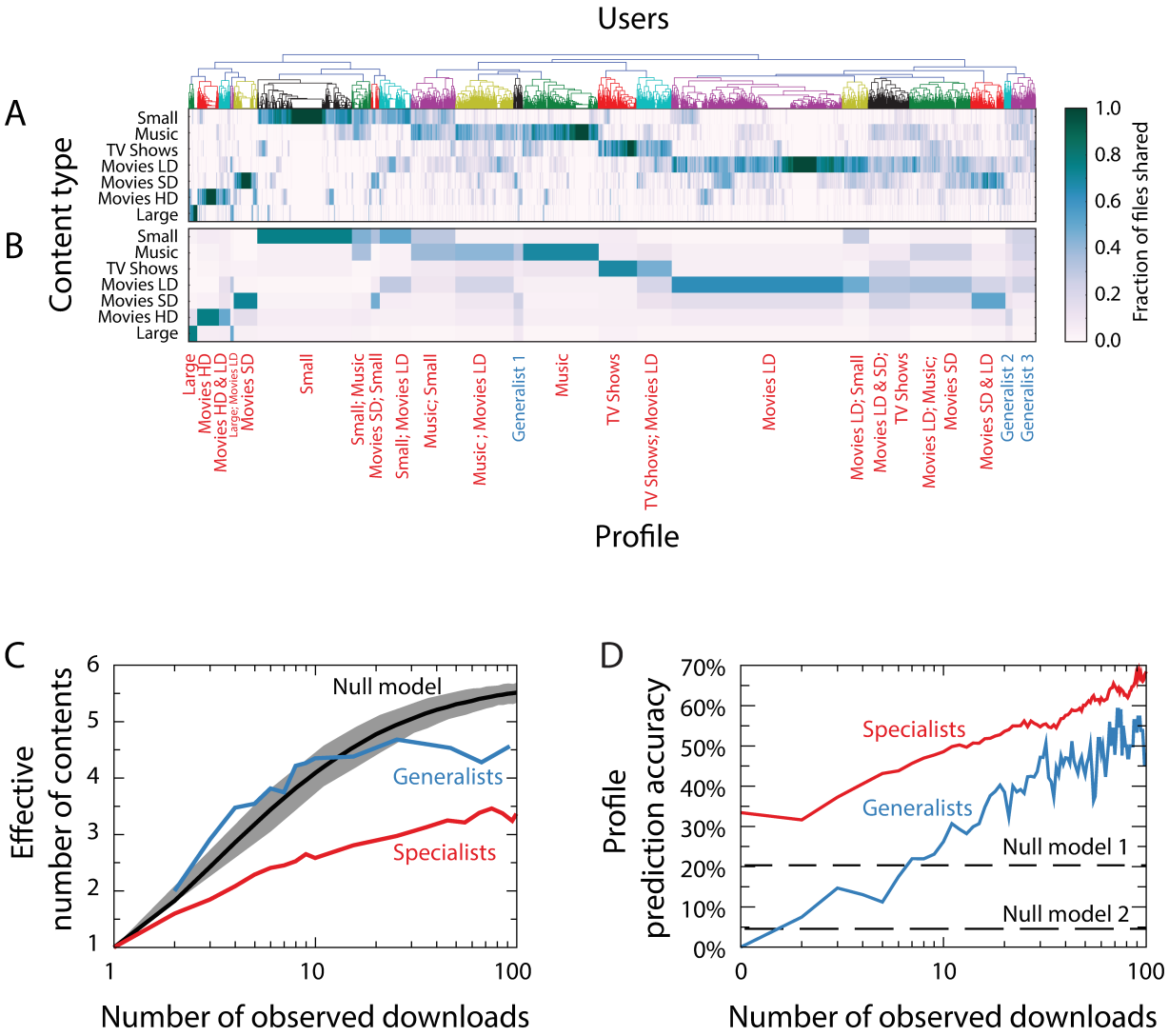

Fig. S6: Same as Fig. 2 of the main text with 22 user profiles instead of 17. (A) With more clusters the specialists get divided to even more specialists and we discover some "pseudo-specialists" inside previous generalists. (B) Then generalists are less common and very generalists (they are closer the null model) while the new specialists come mostly from previous generalists or some specialists. (C) The effective number of contents increases approximately by 0.5 for both types. (D) With more profiles to choose from it, is harder for our simplistic model to predict it. This fact makes specialist profiles 5% on average harder to predict, specially with small number of downloads. The prediction accuracy fluctuates because we have less generalist users too.

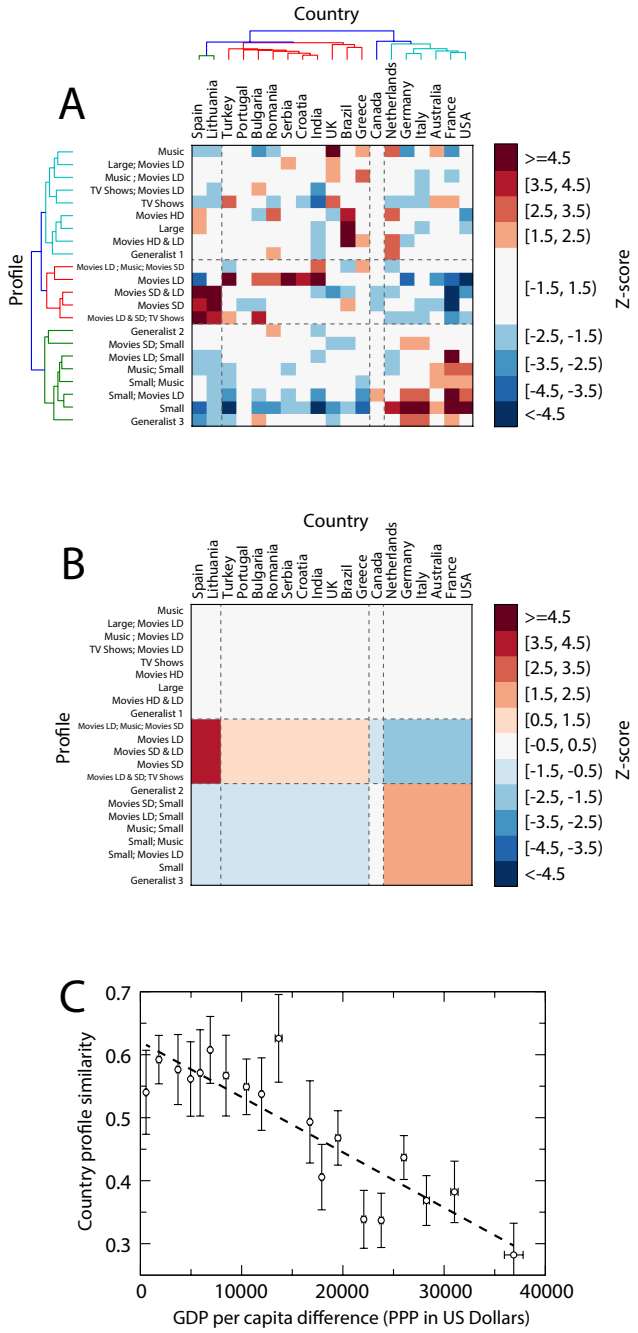

Fig. S7: Same as Figs. 3 and 4A (for GDP) of the main text with 22 user profiles instead of 17. (A) The 22 clusters provide higher detail. Then we can see that, Movies low, Movies medium and small content types play a key role to distinguish the country profiles. (B) When we average the boxes we see that users with low-middle GDP per capita share more Movies low and/or Movies medium than expected and less small files. Also, we see that Brazil, Greece and UK join the group of mostly low-middle class countries. (C) The trend of the country profile similarity vs GDP per capita difference plot is similar to the one of 17 clusters.

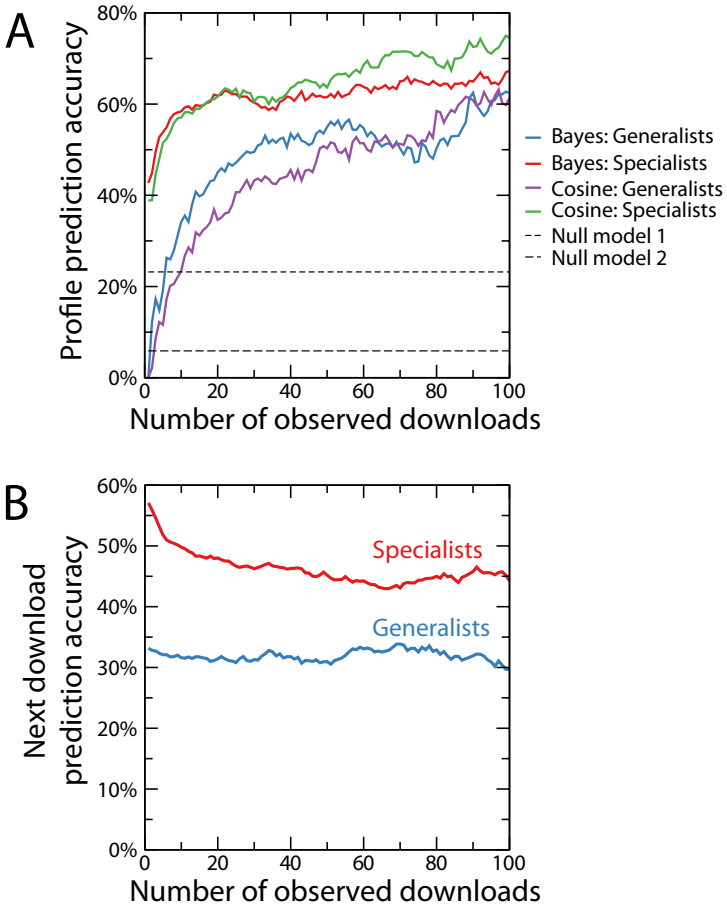

Fig. S8: (A) As the observed downloads increase, the cosine model gets better than the Bayesian. With a small number of downloads, the Bayesian model has better predictive accuracy. Also the cosine method has a slightly better performance with specialists. Since more than 70% of the users are specialists we conclude that the cosine model is more accurate in general and use it in the manuscript. (B) The specialists next download is fairly predictable. With our simplistic model we can predict exactly the next download of between 45% to 50% of the specialists and for up to 30% to 33% of the generalists.

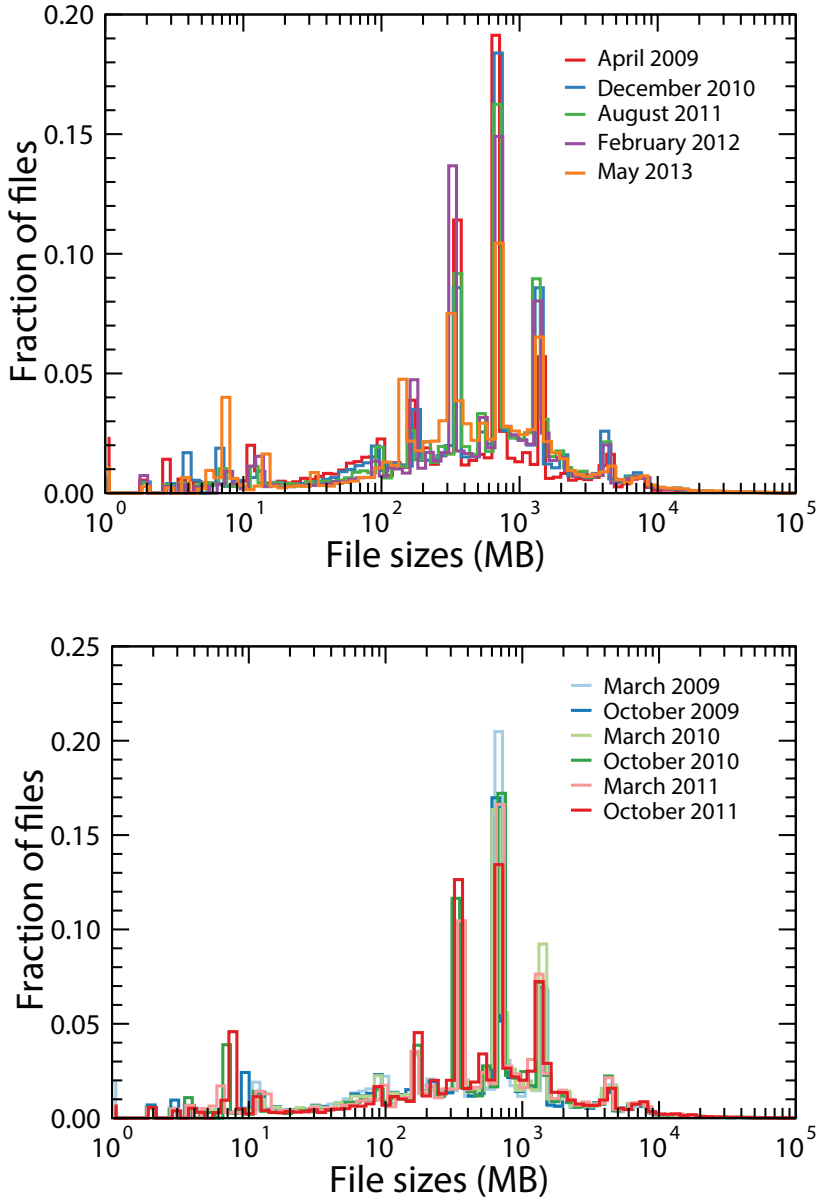

Fig. S9: The distribution of file sizes of the downloads remains stable for the 5 years of analyzed data, with the existence of peaks at the same file sizes in all of them. In top left panel we compare 5 randomly selected months amongst the 48 available, while in the bottom panel we compare 6 equally spaced months.

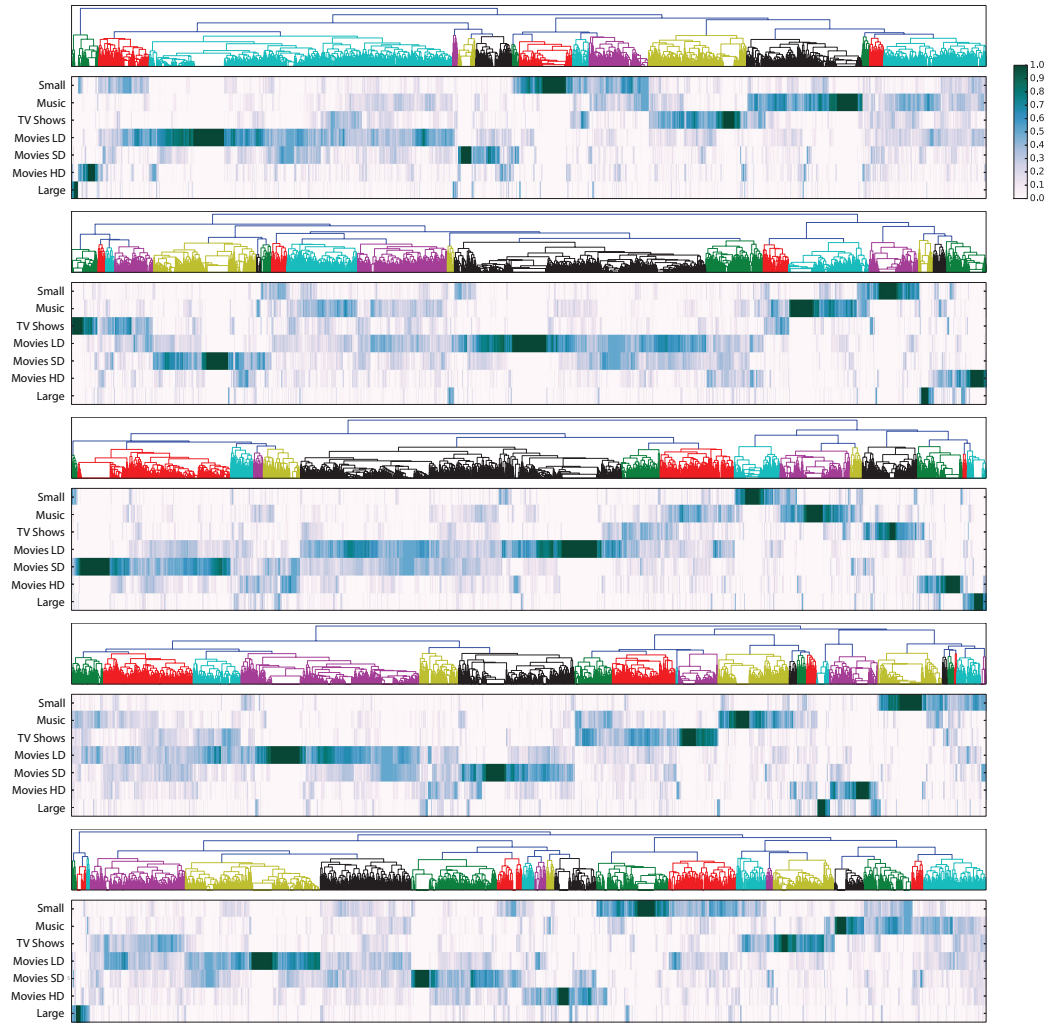

Fig. S10: User profile identification for 5 months selected randomly amongst the 48 available months (from top to bottom: April 2009, December 2010, August 2011, February 2012, May 2013). We observe that most users have a strong tendency towards sharing only one or two content types, and the resulting clusters are approximately of the same size.

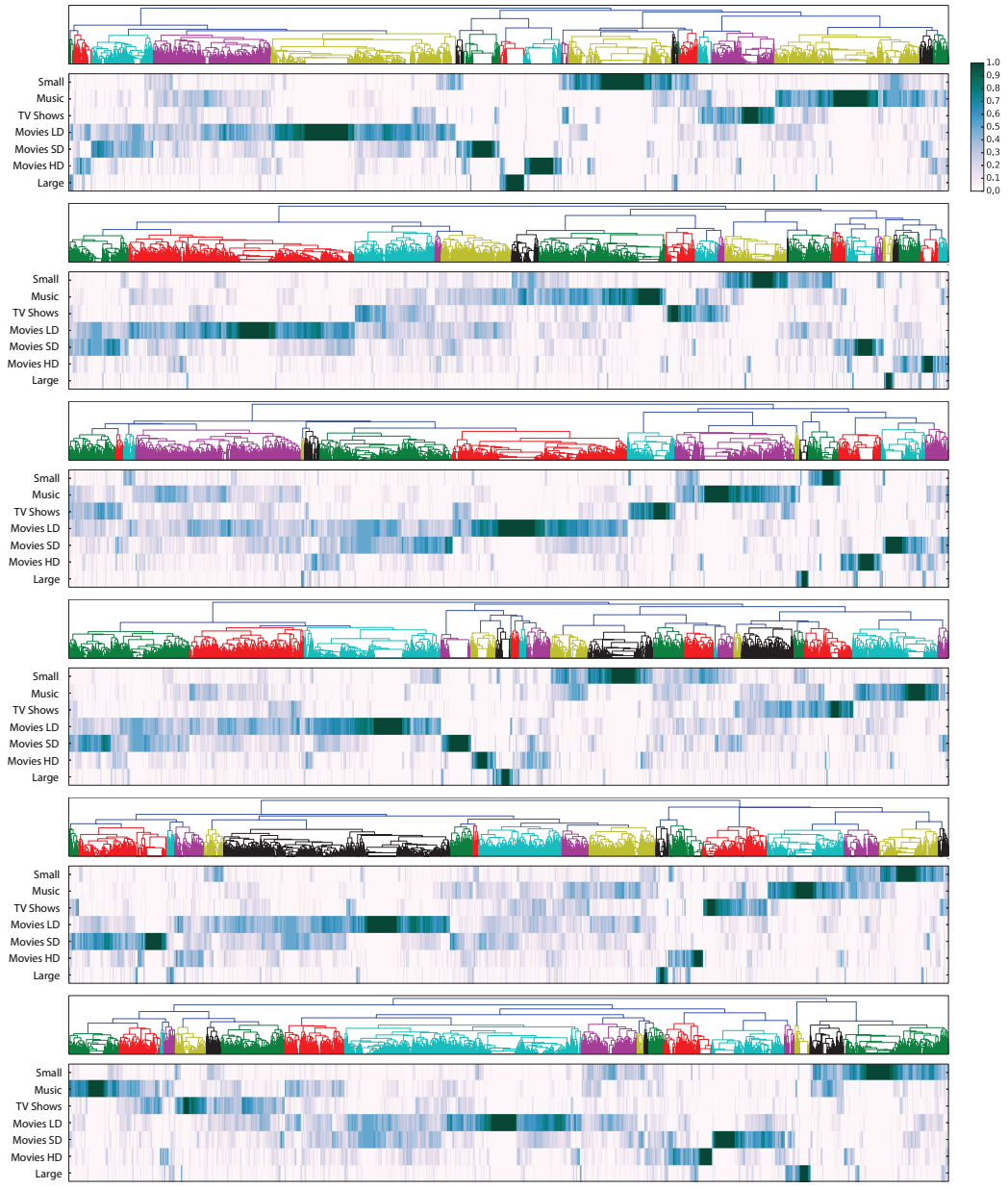

Fig. S11: User profile identification for 6 months selected periodically (from top to bottom: March 2009, October 2009, March 2010, October 2010, March 2011, October 2011). As in Fig. S10, we again observe that most users have a strong tendency towards sharing only one or two content types, and the resulting clusters are approximately of the same size.

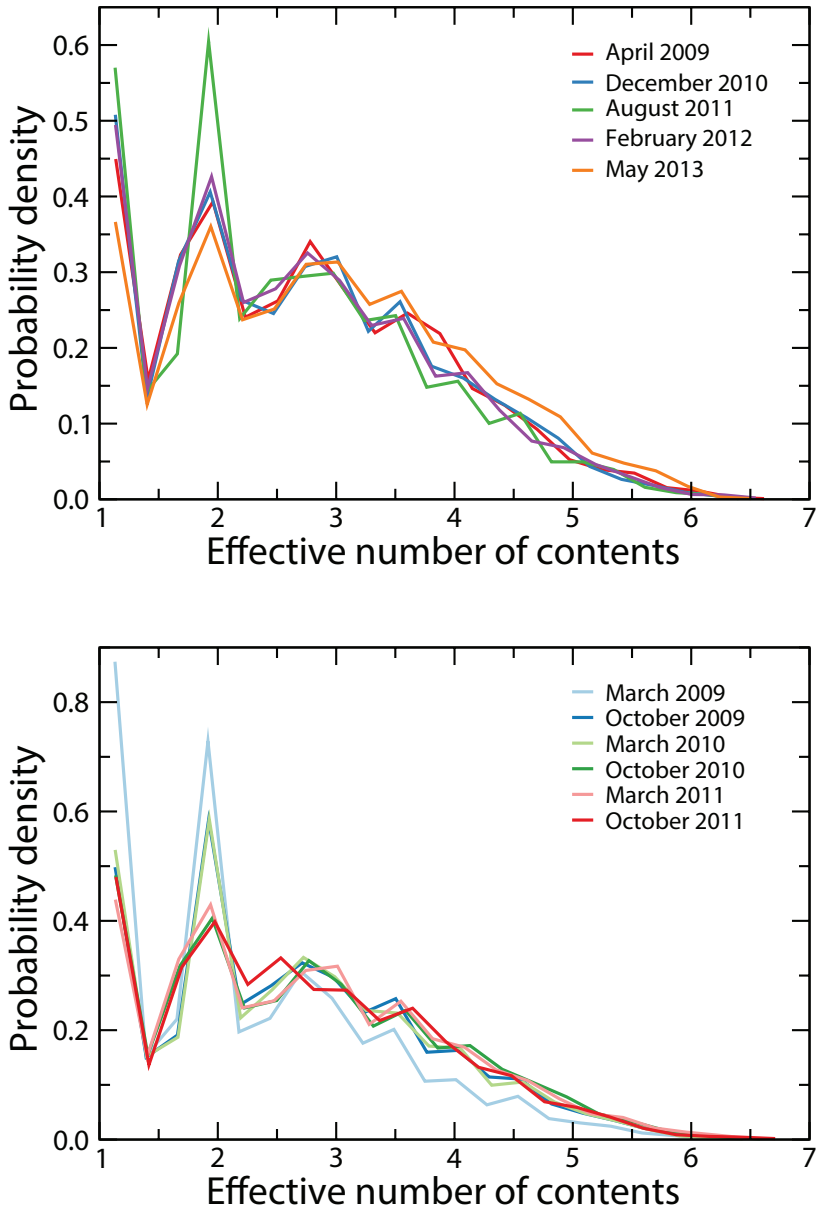

Fig. S12: The distribution of the effective number of contents is also stable for the 11 months that we have studied. In the top panel we show the effective number of contents of users belonging to 5 randomly selected months, while in the bottom panel we selected 5 months distributed periodically. In both figures we observe that the distributions are similar, with most of the users focusing in a very small number of content types.

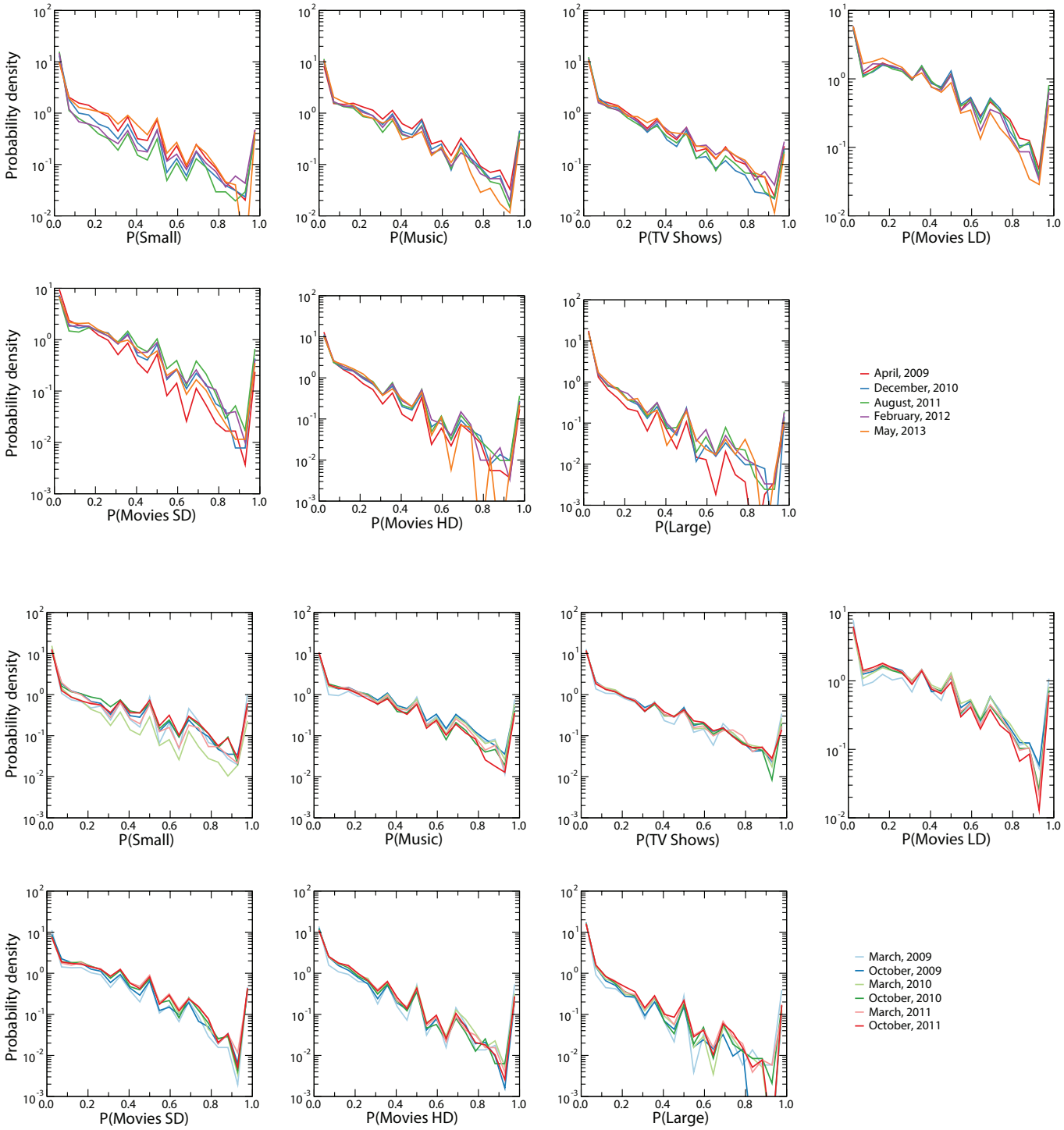

Fig. S13: The distribution of the fraction of each content type that users download remain stable over time. A value $f(\text{MoviesLD})=0.5$ corresponds to a user that downloads half of her files of content type "Movies low", and the corresponding probability density is the probability that a randomly selected user downloads half of her contents of that type. Here we depict the probability density of each content type for the users of a set of randomly selected months between 2009 to 2013 and for a set of periodically selected months between 2009 and 2011. In all the cases we observe that the distributions are fairly similar over time, i.e. the users select their downloads from the categories that we have defined equally independently of the month chosen.

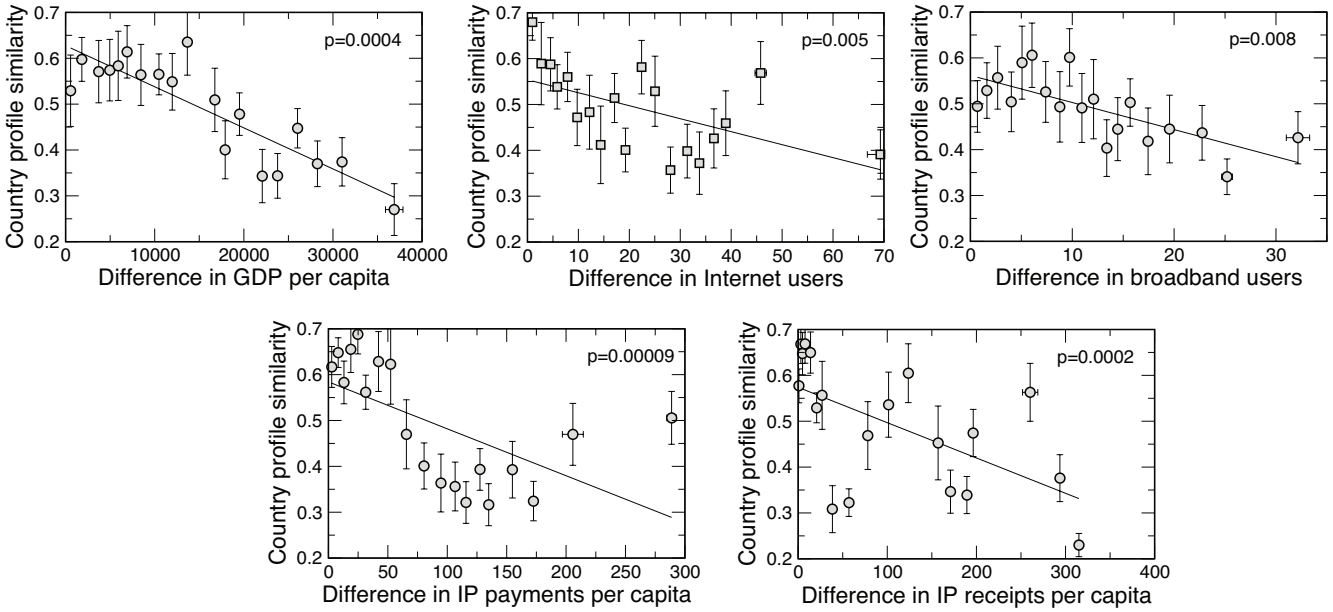

Fig. S14: We quantitatively investigate whether there is a correlation between user behavior and five socioeconomic indicators of where the user lives: GDP per capita (PPP in US Dollars 2011), number of Internet users per 100 people, number of broadband users per 100 people, payments per capita paid to other countries for the use of intellectual property and payments per capita received from other countries for the use of intellectual property (both in current US dollars). We analyze the similarity between pairs of country profiles, defined as the cosine similarity between the vectors of user profile z-scores; as a function of the difference of each indicator. We plot the country profile similarity as a function of the absolute difference in each of the five socio-economic indicators, and average over pairs of countries with a similar value in the indicator. In all cases the observed correlation is significant (see Fig. S15)

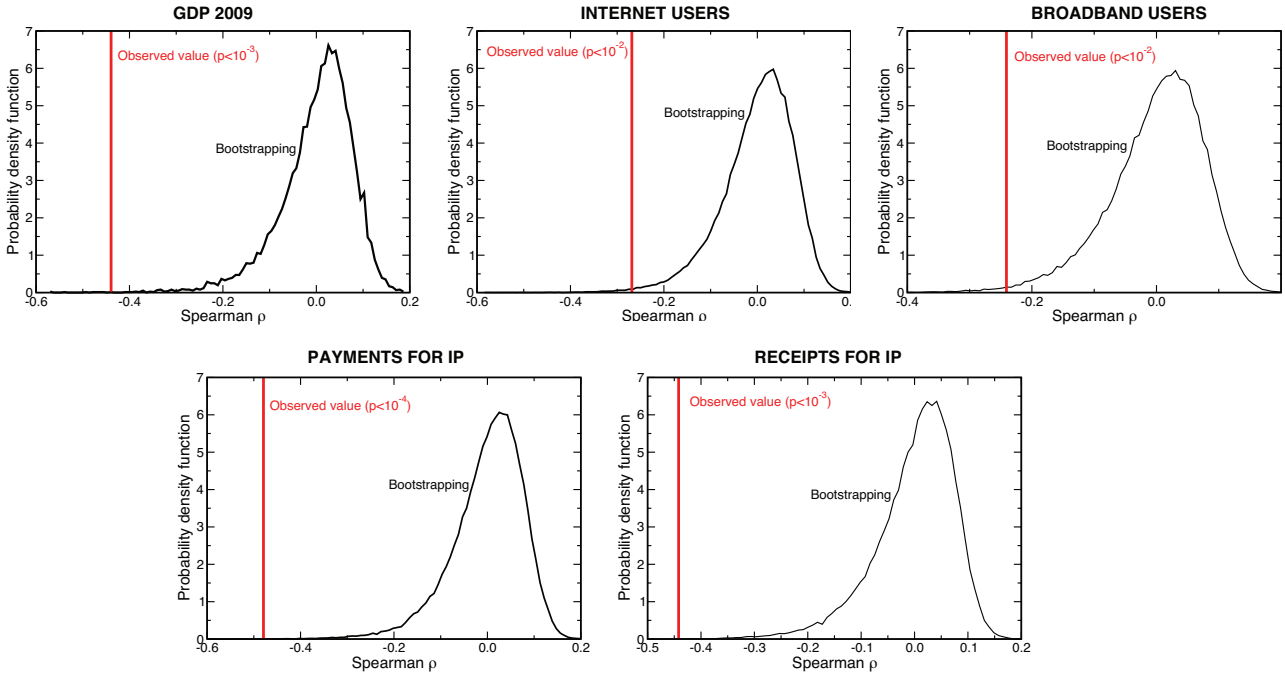

Fig. S15: We calculate the Spearman ρ statistic for the observed pairs (S_{ij}, I_{ij}), where S_{ij} is the similarity between countries i and j, and I_{ij} is the absolute difference between countries i and j in socio-economic indicator I. To establish the significance of the statistic (and given that not all points are independent, so that we cannot use directly the Spearman test) we bootstrap the values of the indicators for each country, and compute the p-value comparing the observed ρ to what one should expect from the bootstrapped samples. The correlation is significant in all cases, the weakest being for Internet users ($p < 10^{-2}$), and the strongest for payments made for intellectual property ($p < 10^{-4}$).

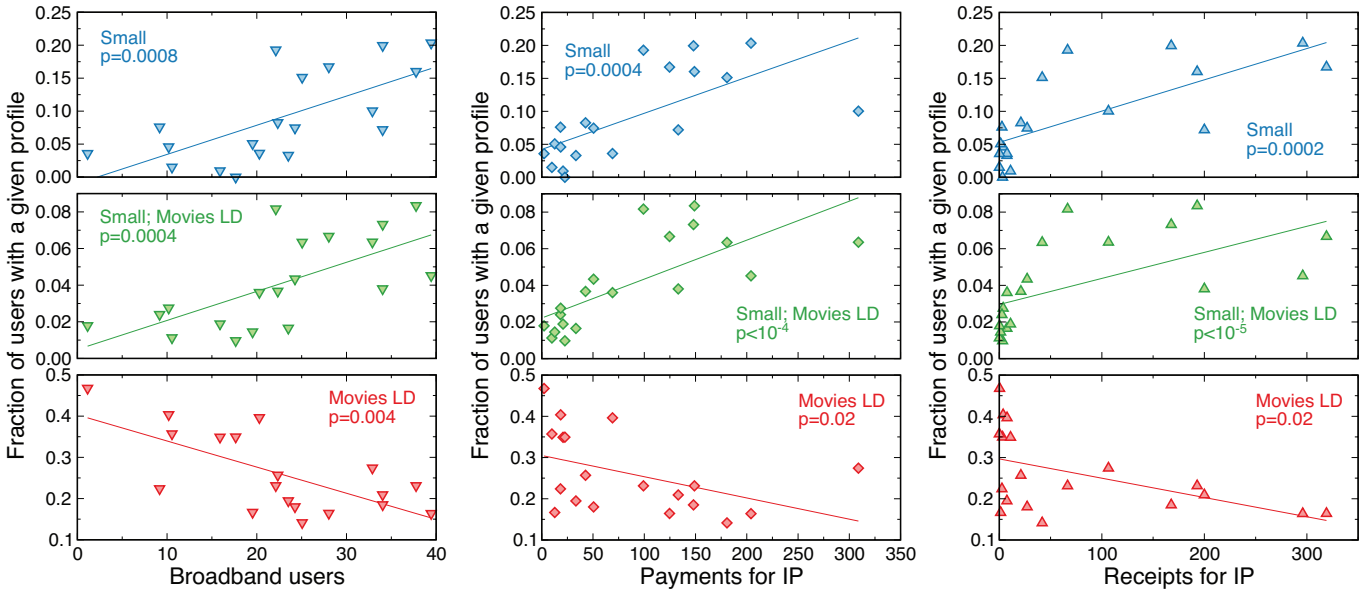

Fig. S16: Same as Figure 4B-D of the main text but for: (left) number of broadband users per 100 people, (center) payments per capita paid to other countries for the use of intellectual property and (right) payments per capita received from other countries for the use of intellectual property (both in current US dollars).

User profile	Effective contents
Generalist 3	6.27
Generalist 4	5.03
Generalist 1	4.64
Generalist 2	4.51
Music; Movies LD	3.78
Music; Small	3.64
TV Shows; Movies LD	3.37
Movies HD	3.33
Small; Movies LD	3.16
Movies SD & LD	2.91
Small; Music	2.91
Movies LD	2.38
Movies HD; Large	2.37
Movies SD	2.26
TV Shows	2.00
Music	1.97
Small	1.71

Table S1: Effective number of contents, E, for each user profile. The table is sorted from high to low effective number of contents. The difference between generalists and specialists is of at least 0.73 effective contents.

	User profiles																	
Country	Movies HD; Large	Movies HD	Movies SD	Small	Small; Music	Small; Movies LD	Music; Small	Music; Movies LD	Generalist 1	Music	TV Shows	TV Shows; Movies LD	Movies LD	Generalist 2	Movies SD & LD	Generalist 3	Generalist 4	Total
Australia	7	1	1	31	9	13	16	15	2	27	15	10	29	19	5	0	5	205
Bulgaria	0	2	1	0	0	1	4	9	2	1	3	8	36	24	6	0	6	103
Brazil	26	13	14	19	5	6	9	16	2	28	13	11	56	19	5	2	6	250
Canada	9	3	4	30	9	19	17	20	1	25	11	16	82	35	6	2	10	299
Croatia	3	1	1	4	3	4	4	10	0	11	1	3	44	13	4	2	3	111
Germany	19	4	23	139	13	51	30	52	10	42	24	27	129	70	24	6	34	697
France	24	9	2	144	30	75	67	44	7	55	52	46	208	86	5	10	35	899
Greece	11	9	10	10	1	5	16	36	3	37	14	15	59	50	10	4	13	303
India	10	2	13	16	7	8	23	25	1	34	3	7	209	64	9	0	16	447
Italy	11	4	7	85	9	36	29	19	4	32	12	9	102	42	14	4	22	441
Lithuania	6	1	14	7	0	2	1	9	2	7	0	2	23	27	34	2	1	138
Netherlands	13	7	4	36	4	8	3	11	6	28	3	5	29	11	1	2	6	177
Portugal	5	1	0	9	3	4	6	11	1	9	3	6	28	16	4	1	2	109
Romania	8	2	1	1	2	2	6	8	3	4	1	5	37	16	5	3	2	106
Serbia	3	0	7	5	1	3	1	11	2	8	3	5	44	11	3	0	2	109
Spain	56	11	53	79	17	46	41	69	14	77	34	50	191	172	129	7	14	1060
Turkey	8	5	12	4	1	3	6	17	2	29	22	15	95	35	9	0	3	266
UK	7	6	13	34	14	18	23	44	6	73	35	24	99	56	8	5	8	473
USA	13	15	29	173	34	69	76	73	11	95	53	39	170	113	25	12	35	1035
Others	90	33	91	207	47	76	120	160	19	211	134	94	561	297	78	20	56	2294
Unknown	21	4	13	56	13	11	17	12	5	38	11	6	39	8	3	1	3	261
Total	350	133	313	1089	222	460	515	671	103	871	447	403	2270	1184	387	83	282	9783

Table S2: Number of users for each user profiles and country with more than 100 users for the month March 2009. Countries with less than 100 users are into "Others". Users without country are labeled as "Unknown". The order of user profiles is the same as in Fig. 2B.

User Profile	**GDP 2009**	**Internet Users**	**Broadband Users**	**IP Payments**	**IP Receipts**
Movies HD; Large	0.626648	0.937498	0.737180	0.673252	0.490672
Movies HD	0.689053	0.920498	0.721024	0.988626	0.753443
Movies SD	0.177665	0.206516	0.140775	0.023791*†	0.085190
Small	0.000429**	0.000847**	0.000847**	0.000429**	0.000170**†
Small; Music	0.002146**	0.006098**	0.008223**	0.000245**†	0.000350**
Small; Movies LD	0.000450**	0.001337**	0.000409**	2.22e-05**	4.26e-06**†
Music; Small	0.053731	0.286430	0.120645	0.033415*	0.031282*†
Music; Movies LD	0.788924	0.911971	0.883736	0.869666	0.847234
Generalist 1	0.683772	0.445906	0.523260	0.909182	0.441550
Music	0.130437	0.325255	0.328931	0.256883	0.399223
TV Shows	0.119521	0.357061	0.181227	0.226184	0.293049
TV Shows, Movies LD	0.786171	0.866856	0.583674	0.601157	0.982932
Movies LD	0.000409**†	0.002120**	0.003983**	0.029921*	0.016082*
Generalist 2	0.109600	0.170297	0.160799	0.023237*†	0.038829*
Movies SD & LD	0.147245	0.177665	0.130437	0.061469	0.053731
Generalist 3	0.378622	0.230965	0.230965	0.519527	0.194192
Generalist 4	0.405215	0.728974	0.312360	0.158289	0.253507

Table S3: For each type of user profile in the month of March 2009 we study if there is a correlation between the fraction of users of each profile versus the following socio-economic factors: The GDP of the country in 2009, the number of Internet users per 100 people, the number of broadband users per 100 people, the payments per capita paid to other countries for the use of intellectual property and the payments per capita received from other countries for the use of intellectual property (both in current US dollars). For each column of this table we compute the Spearman rank correlation between the indicator and the type of user profile, and we present the significance of the value obtained. We indicate what values are significant using * if $p < 0.05$ and ** if $p < 0.01$. We also indicate with † the most significant correlation for each profile. We find that there are three types of user profiles that have a significant correlation with most of the indicators, "Small", "Small; Movies LD" and "Movies LD". We also find that the number of internet users and broadband users are the less significant of all the studied socio-economic indicators.

User Profile	Best linear model	$\mathbf{BIC}_1$	Second best linear model	$\mathbf{BIC}_2$
Small	$y = a\,\mathrm{GDP}$	-59.8	$y = a\,\mathrm{GDP} + b\,\mathrm{IPR}$	-59.3
Small; Movies LD	$y = a\,\mathrm{GDP}$	-97.6	$y = a\,\mathrm{GDP} + b\,\mathrm{IPP}$	-95.6
Movies LD	$y = a\,\mathrm{GDP} + b\,\mathrm{IPP} + c\,\mathrm{IU} + d$	-46.3	$y = a\,\mathrm{GDP} + b\,\mathrm{IPP} + c$	-45.5

Table S4: Most predictive linear models (according to BIC) for the fraction of users with each profile. GDP = Gross domestic product per capita; IPP = Intellectual property payments; IPR = Intellectual property receipts; IU = Internet users.