Impact of heterogeneity and socioeconomic factors on individual behavior in decentralized sharing ecosystems

Arnau Gavaldà-Miralles[a,b,c], David R. Choffnes[d], John S. Otto[e], Mario A. Sánchez[e], Fabián E. Bustamante[e], Luís A. N. Amaral[c,f,g,h], Jordi Duch[a,1], and Roger Guimerà[b,i]

Departaments [a]d'Enginyeria Informàtica i Matemàtiques and [b]d'Enginyeria Química, Universitat Rovira i Virgili, 43007 Tarragona, Spain; Departments of [c]Chemical and Biological Engineering, [e]Electrical Engineering and Computer Science, and [f]Physics and Astronomy, [g]Northwestern Institute on Complex Systems, and [h]Howard Hughes Medical Institute, Northwestern University, Evanston, IL 60208; [d]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350; and [i]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

Edited by Alessandro Vespignani, Northeastern University, Boston, MA, and accepted by the Editorial Board September 10, 2014 (received for review May 31, 2013)

Tens of millions of individuals around the world use decentralized content distribution systems, a fact of growing social, economic, and technological importance. These sharing systems are poorly understood because, unlike in other technosocial systems, it is difficult to gather large-scale data about user behavior. Here, we investigate user activity patterns and the socioeconomic factors that could explain the behavior. Our analysis reveals that (*i*) the ecosystem is heterogeneous at several levels: content types are heterogeneous, users specialize in a few content types, and countries are heterogeneous in user profiles; and (*ii*) there is a strong correlation between socioeconomic indicators of a country and users behavior. Our findings open a research area on the dynamics of decentralized sharing ecosystems and the socioeconomic factors affecting them, and may have implications for the design of algorithms and for policymaking.

human activity | Internet | content sharing | privacy | BitTorrent

Every month, ~150 million users worldwide share files over the Internet using BitTorrent (1), the most widely used decentralized peer-to-peer (P2P) communication protocol. Eleven years after its inception, file sharing through BitTorrent is one of the top three major contributors to the overall Internet traffic, accounting for 9–27% of the total traffic, depending on the continent (2, 3).

The expansion in scale and breadth of decentralized file-sharing has highlighted the conflicts between the interests of creators (musicians and writers, e.g.) and those of P2P users. Creators and creative industries argue that they are being deprived of fair compensation for their work (4), which is being widely distributed for free in violation of copyright laws. Users, however, argue that P2P can be (and is) used for sharing nonproprietary contents, and warn that widespread monitoring of online activity by corporations and law enforcement violates P2P users' right to privacy. Proof of the complexity of the situation includes the rejection of the Anti-Counterfeiting Trade Agreement by the European Parliament and the controversy with the Stop Online Piracy Act in the United States.

Despite the growing social, economic, and technological importance of BitTorrent (4), there is currently little understanding of how users behave in this complex technosocial (5, 6) ecosystem. Due to the decentralized structure of P2P ecosystems, it is very difficult to gather large-scale data about interactions and behavioral patterns of the users without their explicit consent; this is in contrast to other forms of online exchange where all of the information is stored in a central system, be it publicly accessible as in Wikipedia (7), partially accessible through a public interface as in Twitter (8, 9) or Google [through its search logs (10) or its public services (11, 12)], or restricted as in Facebook (13, 14) or in email communications within organizations (15–18).

Because of the difficulty to collect complete user-level data of large and representative samples of users (3), studies of user behavior in P2P networks have so far been based on (*i*) small datasets; (*ii*) aggregate data collected from "trackers" or from individual Internet service providers (ISPs); and (*iii*) incomplete user data collected using a single crawler client connected to the network (19–23).

Here, we investigate the complete activity patterns of a large and representative pool of BitTorrent users. Our analysis reveals that P2P sharing is highly heterogeneous, that users are specialized, giving rise to well-defined user profiles, and that the abundance of certain user profiles in a country is highly correlated with socioeconomic factors. Our findings open a research area on the dynamics of decentralized sharing ecosystems, and may have implications for the understanding and design of algorithms and for policymaking.

Data

We collected anonymized user activity data during the period March 2009 to October 2013 from more than 1.4 million users of the Ono plugin who gave informed consent for the use of their sharing behavior for research purposes (24) (*SI Appendix*). To protect the privacy of these users, we restricted our data collection

Significance

The emergence of the Internet as the primary medium for information exchange has led to the development of many decentralized sharing systems. The most popular among them, BitTorrent, is used by tens of millions of people monthly and is responsible for more than one-third of the total Internet traffic. Despite its growing social, economic, and technological importance, there is little understanding of how users behave in this ecosystem. Because of the decentralized structure of peer-to-peer services, it is very difficult to gather data on users behaviors, and it is in this sense that peer-to-peer file-sharing has been called the "dark matter" of the Internet. Here, we investigate users activity patterns and uncover socioeconomic factors that could explain their behavior.

Author contributions: A.G.-M., F.E.B., L.A.N.A., J.D., and R.G. designed research; A.G.-M., L.A.N.A., J.D., and R.G. performed research; A.G.-M., D.R.C., J.S.O., M.A.S., F.E.B., L.A.N.A., J.D., and R.G. analyzed data; and A.G.-M., D.R.C., F.E.B., L.A.N.A., J.D., and R.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.V. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: jordi.duch@urv.cat.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309389111/-/DCSupplemental.

to country of residency of the user, time of initiation of file sharing, and size of the shared file. We did not collect the name of the file or its content type classification. Although users of the Ono plugin constitute only ~1% of estimated BitTorrent users, we found that they are a representative sample of the BitTorrent ecosystem both in terms of country representation (3) and the sizes of the files they share (*SI Appendix*, Fig. S2).

We define active users for a given month of interest as those individuals that reported sharing activity during that month, and also during the prior and subsequent months (*SI Appendix*). From the complete log files for each month, which once compressed are ~100 gigabytes each, we extracted the complete set of sharing interactions of active users for 11 distinct months (*SI Appendix*). We report here results for the 9,783 active users during March 2009, who shared 217,982 different files for a total of 10,976,607 downloads. As we show in *SI Appendix*, Figs. S9–S13, the findings we report for March 2009 hold for all other months considered.

Results

File Sizes Are Informative of Content Types. As we show in Fig. 1, file size is informative of content types. The file size distribution has six major peaks corresponding to file sizes preferred by users (Fig. 1*A*), in agreement with results derived from aggregate data (22). Some of these peaks are clearly related to physical support [e.g., the peak around 830 megabytes (MB) reflects that many files are likely stored in compact disks]; other peaks are likely related to content types [e.g., a 40-min television (TV) show requires a file size of 200–400 MB].

To establish a relationship between file size and content type, we randomly sampled 456,949 torrents from a widely used BitTorrent repository. The metadata for these torrents includes both file size and content category. We determine the most common content classes for the file size classes suggested by the peaks (*SI Appendix* and Fig. 1*B*). We find that, for all size classes, a small number of content categories accounts for a disproportionately large fraction of the files. For example, high-resolution movies and pornographic movies account for over 60% of all files with sizes between 831 MB and 1,650 MB. Based on this observation, we define seven content types as follows (Fig. 1*B*): Small (accounts for 17% of all downloads in our database), Music (18%), TV Shows (12%), Movies Low Definition (LD; 26%), Movies Standard Definition (14%), Movies High Definition (HD; 9%), and Large (4%).

User Behavior Is Remarkably Predictable. We use the ability to infer the content type to investigate whether users participate in the ecosystem as generalists (i.e., sharing according to the average proportions observed for all content types) or as specialists (i.e., focusing on a small number of types). As we show in Fig. 2*A*, most users have a strong tendency toward sharing only one or two content types. In particular, for 96% of the users, their two most downloaded content types account for more than 50% of their downloads. Therefore, most users behave as specialists, at least within our current classification of content types (we cannot establish to what extent users are specialists/generalists at a finer scale, e.g., if they download movies of a single genre or several genres).

Because most users behave as specialists we surmise that they can be clustered into groups with common sharing behaviors. We use hierarchical clustering to group the 9,783 active users and define 17 different user profiles (alternative clusterings do not change the conclusions of the paper; *SI Appendix*). In Fig. 2*B* we display the average sharing behavior of users in each of the groups.

To further quantify the degree of specialization, we measure the effective number E of contents downloaded by a user or a group of users, which we define as $E = 1/\sum_i f_i^2$, with f_i being the fraction of all downloads that are of content type i (*SI Appendix*) (25). For example, if a user downloaded three content types, each amounting to one-third of the user's downloads, then $E = 3$. A user sharing content according to the overall probabilities would have $E = 5.7$. Whereas 13 of the 17 average user profiles have $1.7 \le E \le 3.8$, four have $4.5 \le E \le 6.3$. Based on this observation, we define specialist user profiles (if the average user profile has $E < 4.5$) and generalist user profiles (if the average user profile has $E \ge 4.5$). Even users that we classify as generalist are on average more specialized that a hypothetical perfect generalist that downloads contents types with the average proportions of each content type (Fig. 2*C*).

An important consequence of the fact that most users behave as specialists is that even a few downloads from a user are highly informative of the user's profile and, therefore, of their future sharing behavior. Just five downloads enable us to correctly identify the profile of more than 50% of the specialists (Fig. 2*D*). The assignment accuracy increases to 75% for 100 observed downloads. Similarly, one can accurately predict the next content type that a specialist user will download (*SI Appendix*). Significantly, the high predictability of user behaviors raises the concern of threats to privacy and guilt-by-association attacks (26, 27).

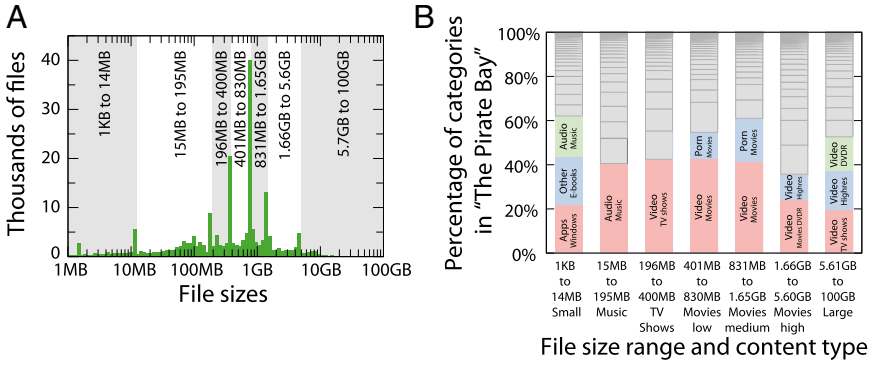

Fig. 1. File sizes define distinct content types. (*A*) For each user, we collect the size of the files they downloaded. We plot the distribution of all those file sizes, with sizes binned logarithmically. This distribution has pronounced peaks at 14 MB, 195 MB, 400 MB, 830 MB, 1.65 GB, and 5.6 GB. Based on these peaks, we define seven file size ranges (alternating white and gray bands). (*B*) File size ranges can be associated to distinct content types. We randomly sample half a million torrents at "The Pirate Bay" and analyze their content categories as a function of their sizes. For each file size range, 1–3 categories account for most of the observed files. For example, for file sizes in the range 196–400 MB, which we denote as Videos of TV Shows, accounts for 40% of all files. For each size range, we color and name all categories that account for more than 10% of the files in the range and that are significantly overrepresented, $P < 0.05$, with respect to a null model in which categories are uniformly distributed among file size ranges (*SI Appendix*).

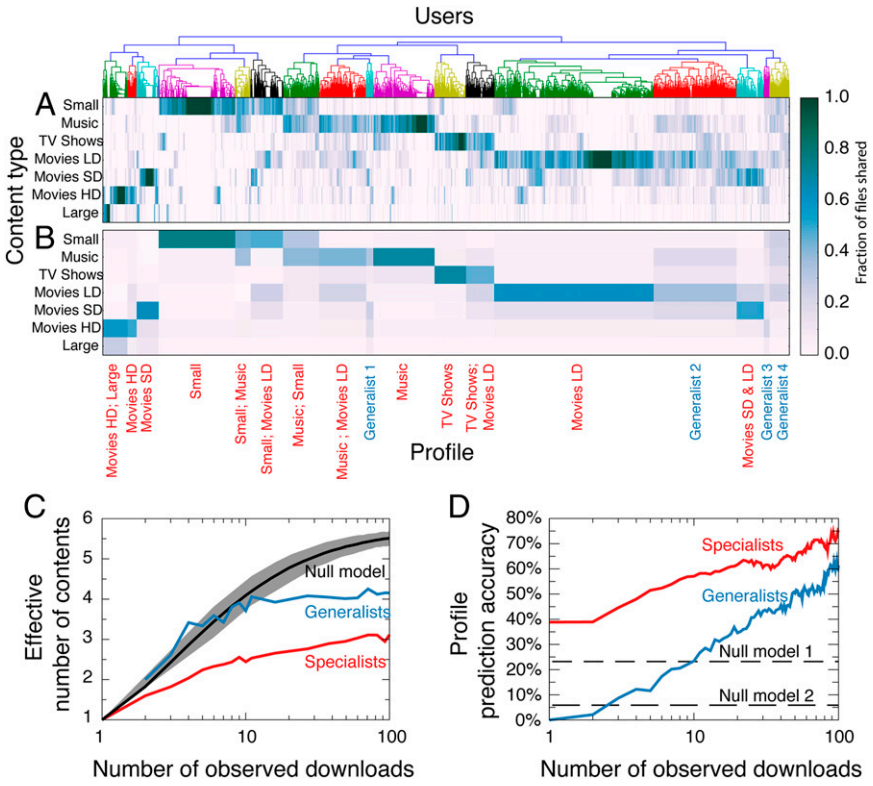

Fig. 2. Users are heterogeneous, mostly specialized, and predictable. (*A*) We calculate the frequency with which each user downloads files from each content type. We hierarchically cluster users according to these frequencies and identify 17 user profiles (see *SI Appendix* for alternative partitions of the users into groups, which support the conclusions of the manuscript). (*B*) For each group, we depict the average download frequencies, which provide a stylized profile of the users in the group. We label each profile according to the most prevalent content types in the profile. For instance, users with a Music profile download, on average, Small files (4% of the times), Music (70%), TV Shows (11%), Movies Low Definition (5%), Movies Standard Definition (3%), Movies High Definition (4%), and Large (2%). Users are often highly specialized in few content types. Indeed, for 8 of the 17 user profiles, one content type alone accounts for more than 50% of the downloads, and for 10 of the 17 two content types account for more than 70% of the downloads. We classify as generalists the users that download contents proportionally to their availability and as specialists the users that focus primarily on one or two content types. (*C*) The effective number of contents *E* is indicative of how the downloads of a user or a group of users are concentrated in a small number of content types (*SI Appendix*). We plot the effective number of contents as a function of the number of observed downloads, for specialists (red), generalists (blue), and a hypothetical average users that download files randomly chosen from all observed downloads (black; the gray region corresponds to the 95% confidence interval). (*D*) To evaluate the potential implications for privacy of user specialization, we use a simple model (*SI Appendix*) to infer the profile of users from their downloads alone. We find that specialists can be profiled quite easily with this simple model. Indeed, after having only five downloads we can correctly identify the profile of more than 50% of them. After 100 downloads, our accuracy goes up to 75%. In contrast, generalist users are more difficult to profile; around 50 downloads are necessary to achieve 50% accuracy. For comparison, random guessing of the user profile yields an accuracy of 6% (null model 2) and assigning all users the most frequent profile (Movies low) has 22% accuracy (null model 1).

Socioeconomic Characteristics of a Country Correlate with the Sharing Behavior of Its Users. The user profiles we identify are universal; e.g., a Japanese user that specializes in TV Shows and a Brazilian user with the same profile are indistinguishable in terms of the file types they download. A question prompted by the existence of such profiles is what motivates users to behave in a certain way. One possibility, which has not been quantitatively investigated to date for lack of data, is that different technological and economic conditions, as well as political priorities, will lead users to adopt one profile or another in different countries. Such country dependencies have in fact been observed at an aggregate level in P2P networks (21, 23) and at an individual level in other online behaviors [e.g., a recent study has been able to establish a correlation between the country's gross domestic product (GDP) and the tendency of its inhabitants to search for information about the future, rather than the past] (28).

The question of motivation is complex and difficult to address, because there are several factors that may drive user's behavior: content availability and accessibility through alternative channels, legislation, industry pressures or technological infrastructures. For instance, one may hypothesize that better infrastructure in wealthier countries will lead to widespread use of P2P for larger downloads, e.g., HD movies. However, one may hypothesize the opposite—namely, that widespread access to cable television and video streaming services in wealthier countries eliminates the need for P2P downloading of HD movies.

To investigate the role of economic factors in determining user profiles, we analyze the distribution by country of user profiles. We find that the number of users belonging to a certain profile in a given country significantly deviates from the null expectation that user profiles are randomly and uniformly distributed among countries (Fig. 3*A*). Indeed, we find that most countries have strong overrepresentation of certain user profiles and underrepresentation of others. Using hierarchical clustering, we identify five country profiles (Fig. 3*B*).

The fact that countries in the same group tend to be similarly wealthy suggests that socioeconomic factors may indeed correlate with user behavior. To investigate this in more detail, we analyze whether countries with similar GDP also have users with similar profiles, and we find that that there is indeed a significant correlation [$P = 0.0004$, $\rho = -0.44$, $N = 171$ pairs of countries (Fig. 4*A*)]. We take Spearman's ρ as our statistic, but use bootstrapping

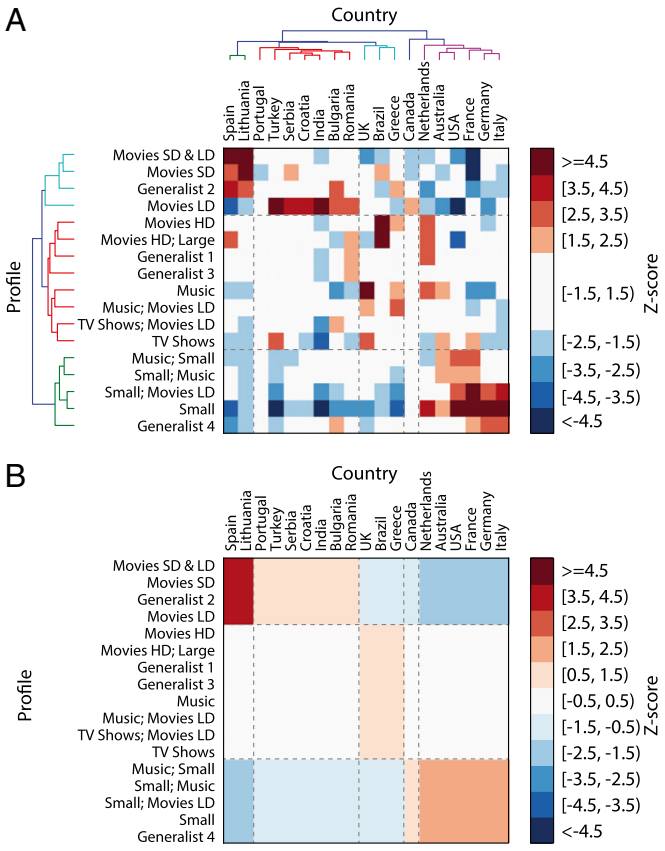

Fig. 3. The distribution of user profiles by countries is heterogeneous. (*A*) Over- and underrepresentation of the 17 identified user profiles in each country. The graph shows the *z*-score of the number of users with a certain profile with respect to the null expectation of uniform distribution of user profiles among countries (only countries with more than 100 users are shown). Countries and user profiles are hierarchically clustered. (*B*) Average of the *z*-scores for the blocks defined in *A*. The blocks reveal that wealthier countries (e.g., the Netherlands, Australia, and United States) have an overrepresentation of small files and an underrepresentation of users that download movies and other large files; countries in the other blocks have the opposite tendency (*SI Appendix*).

to establish the significance, as discussed in *SI Appendix* and *SI Appendix*, Fig. S15).

Of course, GDP is also correlated with other factors, such as Internet infrastructure, which may be relevant to explain users' behaviors. With our data, it is not possible to establish which factors causally and directly determine user behavior, but one can analyze whether these other factors also correlate to behavior, and to which extent. Therefore, we study other socioeconomic indicators of countries, in particular Internet users per 100 people (Fig. 4), as well as broadband availability, payments per capita made to other countries for the use of intellectual property, and payments per capita received from other countries for the use of intellectual property (*SI Appendix*, Figs. S14 and S15). We find that although all these factors significantly correlate with behavior, broadband availability and Internet use have the weakest correlations ($\rho = -0.24$, $P = 0.008$; and $\rho = -0.27$, $P = 0.005$, respectively), whereas intellectual property payments have the strongest ($\rho = -0.48$, $P = 0.00009$).

These results suggest that the opportunity provided by good infrastructure is less of a driving factor than one may have thought, whereas other factors related to overall wealth and to how intellectual property is valued may be more relevant; this is confirmed by the analysis of the abundance of each user profile in different countries. We find that profiles focused in relatively small files (Small, Small; Music, Small; Movies LD, and Movies LD) are monotonically correlated with our socioeconomic indicators (Fig. 4 *B–D* and *SI Appendix*, Fig. S16 and Table S3); as before, the weakest correlation always occurs for broadband availability and Internet use. To parse out the interactions between the (highly correlated) factors we consider, we also carried a model-selection analysis in which we compared all possible linear models of the factors in terms of the Bayesian information criterion (29) (*SI Appendix* and *SI Appendix*, Table S4). We find that GDP is always in the most predictive model, and only in one case adding other factors improves the predictive power of GDP alone.

Interestingly, we observe that the abundance of users focused mostly in Small files correlates positively with all socioeconomic indicators, whereas abundance of users focused almost exclusively on Movies LD correlates negatively. Although the latter correlation may be explained in terms of accessibility to infrastructure (users in poorer countries download more LD movies because they cannot afford downloading larger files), the former cannot (users in richer countries download more Small files than poorer countries). Moreover, an abundance of users that focus on large files, such as Movies HD, is not significantly correlated with any of our socioeconomic indicators so, again, opportunity does not seem to be the main driving factor for use.

Discussion

Our work demonstrates that despite the decentralized nature and privacy safeguarding intrinsic to peer-to-peer ecosystems, they provide researchers with an extraordinary opportunity for investigating social and economic transactions on a large scale and to a level of detail not typically found for such large systems. For example, when studying financial transactions, one is not able to link a transaction to the user that initiated it, whereas in our study we were able to assign every transaction occurring during the March 2009 for the users involved; this opens the door for the use of P2P ecosystems to study economic and social transactions on a large scale and in a real-world context.

Our study also provides important insights concerning the ongoing disputes between creative industries and P2P users (30–34). First, opportunity to download does not in itself seem to lead to an increase in the amount of P2P exchanges. Specifically, HD movies and TV shows are not exchanged as much as one would expect in the United States and other wealthy countries, places where good internet infrastructure would allow for fast downloads of these content types. In contrast, in countries where streaming is not widely available because of poor infrastructure or their cost being out of reach for large portions of the population, we see high levels of P2P exchange of movies and TV shows, despite the exchange relying on poorer Internet infrastructure. Second, copyright laws have unequal impacts on inhibiting P2P exchange. Indeed, even though copyright law enforcement is stronger in the United States and other wealthy countries than in most other countries in the world, one finds a great deal more P2P exchanges of music and small files in wealthy countries than in poorer countries. We speculate that this unexpected high-level of P2P exchange may be related to the lack of convenient (and appropriately priced) distribution channels for music and electronic books.

Finally, our work illuminates some important aspects of the functioning of P2P networks. We have shown that most users in the network are specialists rather than generalists. As in natural ecosystems (35, 36), the specialist/generalist makeup of the sharing ecosystem may have important implications. In particular, specialization implies that the P2P network is compartmentalized, and that most users never interact but with those with a similar profile; this may explain why peer-selection algorithms are highly efficient (despite the very large number of peers connected to the network at any time), and conceivably help improve the algorithms. Moreover, the fact that each country has some user profiles

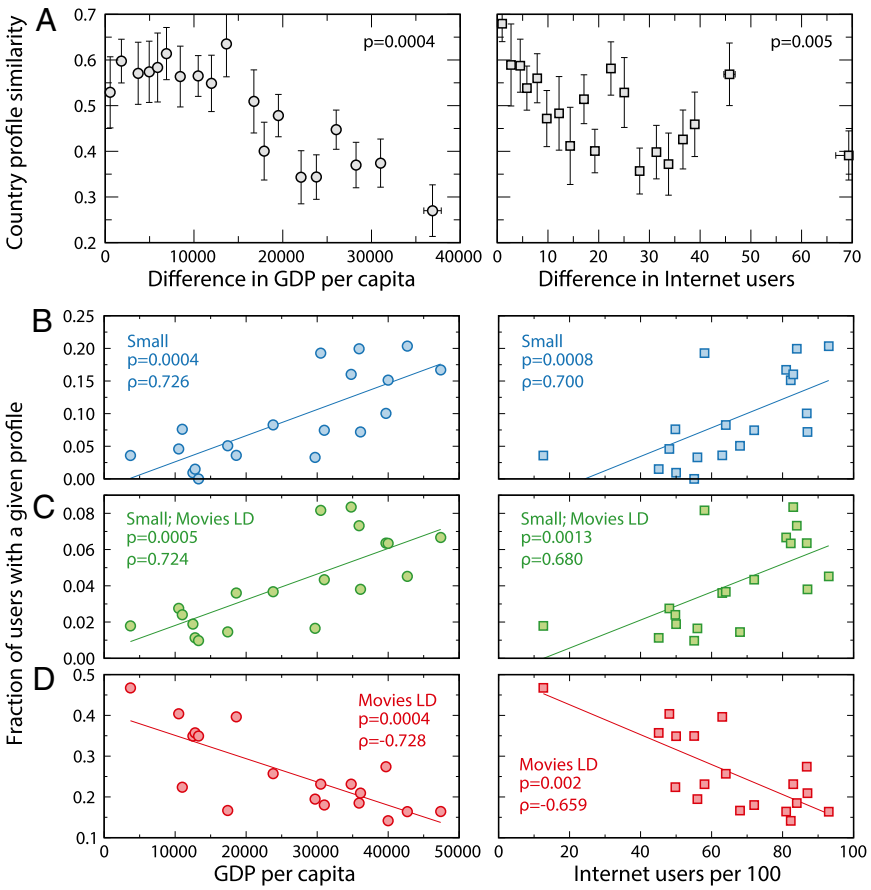

Fig. 4. Correlation between users' behavior and country socioeconomic indicators. We show here our results for two independent variables, GDP per capita in 2009 purchasing power parity US dollars (*Left*); Internet users (*Right*) per 100 people. (*A*) To investigate whether there is a correlation between user behavior and socioeconomic indicators of the country of residency, we analyze the similarity between pairs of country profiles (defined as the cosine similarity, scaled from 0 to 1, between the vectors of user profile *z*-scores; *SI Appendix*) as a function of their absolute difference in GDP per capita, and number of Internet users per 100 people. To establish the significance of these correlations, 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 socioeconomic indicator *I*. We bootstrap the values of the indicators for each country, and compute the *P* value comparing the observed ρ to the expected value from bootstrapped samples (*SI Appendix*). (*B–D*) Fraction of users in a country with a given profile [(*B*) profile Small; (*C*) profile Small; Movies LD; (*D*) profile Movies LD] as a function of the GDP per capita and number of Internet users per 100 people. We obtained the *P* values using Spearman's rank correlation test (see *SI Appendix* and *SI Appendix*, Table S4 for a model selection analysis). The other user profiles (with the exception of profile Small; Music, which behaves similar to profile Small) do not show significant correlations (*SI Appendix*, Table S3).

overrepresented means that the behavior of the network is more efficient in terms of cross-ISP traffic than one would expect from a homogeneous system. Our results also hint at how socioeconomic factors may alter this situation.

ACKNOWLEDGMENTS. We thank A. Aguilar-Mogas, A. Godoy-Lorite, F. A. Massucci, N. Rovira-Asenjo, M. Sales-Pardo, and T. Vallès-Català for useful comments and suggestions. We are especially grateful to the users/adopters of the Ono software for their invaluable data, and to Paul Gardner for his help with distributing the software. This work was supported by a James S. McDonnell Foundation Research Award (to R.G. and A.G.-M.), European Union Grant PIRG-GA-2010-277166 (to R.G.), Spanish Ministerio de Economía y Competitividad Grant FIS2010-18639 (to R.G. and J.D.), EC FET-Proactive Project MULTIPLEX Grant 317532 (to R.G. and J.D.), National Science Foundation (NSF) Grants CNS 0644062 and CNS 0917233 (to D.R.C. and F.E.B.), and NSF/Computing Research Association Computing Innovation Fellowship (to D.R.C.).

1. Cohen B (2003) Incentives build robustness in BitTorrent. *Proc First International Workshop on Economics of Peer-to-Peer Systems*, Vol 6, pp. 68–72. Available at www2.sims.berkeley.edu/research/conferences/p2pecon/program.html. Accessed September 29, 2014.
2. Schulze H, Mochalski K (2009) Internet study 2008/2009. *IPOQUE Rep* 37:351–362.
3. Otto J, Sánchez M, Choffnes D, Bustamante F, Siganos G (2011) On blind mice and the elephant: Understanding the network impact of a large distributed system. *Proc ACM SIGCOMM 2011* (Assoc Computing Machinery, New York), pp. 110–121.
4. Tera Consultants (2010) *Building a Digital Economy: The Importance of Saving Jobs in the EU's Creative Industries* (Intl Chamber of Commerce/BASCAP, Paris).
5. Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* 325(5939):425–428.
6. Lazer D, et al. (2009) Social science. Computational social science. *Science* 323(5915): 721–723.
7. Moat HS, et al. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Sci Rep*, 10.1038/srep01801.
8. Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.
9. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE* 6(12):e26752.
10. Ginsberg J, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.
11. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using google trends. *Sci Rep*, 10.1038/srep01684.
12. Michel JB, et al.; Google Books Team (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.
13. Bond RM, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.
14. Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N (2008) Tastes, ties, and time: A new social network dataset using facebook.com. *Soc Networks* 30(4): 330–342.

15. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. *Phys Rev E Stat Nonlin Soft Matter Phys* 68(6 Pt 2):065103.
16. Eckmann JP, Moses E, Sergi D (2004) Entropy of dialogues creates coherent structures in e-mail traffic. *Proc Natl Acad Sci USA* 101(40):14333–14337.
17. Malmgren RD, Stouffer DB, Motter AE, Amaral LAN (2008) A Poissonian explanation for heavy tails in e-mail communication. *Proc Natl Acad Sci USA* 105(47):18153–18158.
18. Malmgren RD, Stouffer DB, Campanharo ASLO, Amaral LAN (2009) On universality in human correspondence activity. *Science* 325(5948):1696–1700.
19. Saroiu S, Gummadi KP, Gribble SD (2003) Measuring and analyzing the characteristics of napster and gnutella hosts. *Multimedia Syst* 9(2):170–184.
20. Pouwelse JA, Garbacki P, Epema DHJ, Sips HJ (2005) The BitTorrent p2p file-sharing system: Measurements and analysis. *Lecture Notes in Computer Science* 3640:205–216.
21. Iosup A, Garbacki P, Pouwelse JA, Epema D (2005) Analyzing BitTorrent: Three lessons from one peer-level view. *Proc 11th ASCI Conference*, pp 96–104. Available at www.pds.ewi.tudelft.nl/~iosup/aiosup05asci.pdf. Accessed September 29, 2014.
22. Mazurczyk W, Kopiczko P (2013) Understanding BitTorrent through real measurements. *China Commun* 10(11):107–118.
23. Kigerl AC (2013) Infringing nations: Predicting software piracy rates, bittorrent tracker hosting, and p2p file sharing client downloads between countries. *Int J Cyber Criminol* 7(1):62–80.
24. Choffnes D, Bustamante F (2008) Taming the torrent: A practical approach to reducing cross-isp traffic in peer-to-peer systems. *ACM SIGCOMM Computer Communication Review* (Assoc Computing Machinery, New York), Vol 38, pp 363–374.
25. Simpson E (1949) Measurement of diversity. *Nature* 163:688.
26. Choffnes D, et al. (2010) Strange bedfellows: Community identification in BitTorrent. *Proc 9th International Conference on Peer-to-Peer Systems* (USENIX Assoc, Berkeley, CA), p 13.
27. Wetherall D, et al. (2011) Privacy revelations for web and mobile apps. *Proc 13th USENIX Conference on Hot Topics in Operating Systems* (USENIX Assoc, Berkeley, CA), p 21.
28. Preis T, Moat H, Stanley H, Bishop S (2012) Quantifying the advantage of looking forward. *Sci Rep* 2:350.
29. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464.
30. Gopal RD, Sanders GL (1998) International software piracy: Analysis of key issues and impacts. *Inf Syst Res* 9(4):380–397.
31. Yar M (2005) The global epidemic of movie piracy: Crime-wave or social construction? *Media Cult Soc* 27:677–696.
32. Chellappa RK, Shivendu S (2003) Economic implications of variable technology standards for movie piracy in a global context. *J Manage Inf Syst* 20(2):137–168.
33. Khouja M, Rajagopalan H (2008) Can piracy lead to higher prices in the music and motion picture industries? *J Oper Res Soc* 60(3):372–383.
34. Goles T, et al. (2008) Softlifting: Exploring determinants of attitude. *J Bus Ethics* 77(4): 481–499.
35. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. *Proc Natl Acad Sci USA* 100(16):9383–9387.
36. Dunne JA, Williams RJ, Martinez ND (2002) Food-web structure and network theory: The role of connectance and size. *Proc Natl Acad Sci USA* 99(20):12917–12922.