ECONOMICS

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1]* Gabriel Cadamuro,[2] Robert On[3]

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

Reliable, quantitative data on the economic characteristics of a country's population are essential for sound economic policy and research. The geographic distribution of poverty and wealth is used to make decisions about resource allocation and provides a foundation for the study of inequality and the determinants of economic growth (*1*, *2*). In developing countries, however, the scarcity of reliable quantitative data represents a major challenge to policy-makers and researchers. In much of Africa, for instance, national statistics on economic production may be off by as much as 50% (*3*). Spatially disaggregated data, which are necessary for small-area statistics and which are used by both the private and public sector, often do not exist (*4*, *5*).

In wealthy nations, novel sources of passively collected data are enabling new approaches to demographic modeling and measurement (*6*–*8*). Data from social media and the "Internet of Things," for instance, have been used to measure unemployment (*9*), electoral outcomes (*10*), and economic development (*8*). Although most comparable sources of big data are scarce in the world's poorest nations, mobile phones are a notable exception: They are used by 3.4 billion individuals worldwide and are becoming increasingly ubiquitous in developing regions (*11*).

Here we examine the extent to which anonymized data from mobile phone networks can be used to predict the poverty and wealth of individual subscribers, as well as to create high-resolution maps of the geographic distribution of wealth. That this may prove fruitful is motivated by the fact that mobile phone data capture rich information, not only on the frequency and timing of communication events (*12*) but also reflecting the intricate structure of an individual's social network (*13*, *14*), patterns of travel and location choice (*15*–*17*), and histories of consumption and expenditure. Regionally aggregated measures of phone penetration and use have also been shown to correlate with regionally aggregated population statistics from censuses and household surveys (*8*, *18*, *19*).

Our approach is different from prior work that has examined the relation between regional wealth and regional phone use, as we focus on understanding how the digital footprints of a single individual can be used to accurately predict that same individual's socioeconomic characteristics. This distinction is a scientific one, which also has several important implications: First, it allows for the method to be used in contexts for which recent census or household survey data are unavailable. Second, when an authoritative source of data does exist, it can be used to more objectively validate or refute the model's predictions. This limits the likelihood that the model is overfit on data from a single source, which is otherwise difficult to control, even with careful cross-validation (*20*). Third, our approach allows for a broad class of potential applications that require inferences about specific individuals instead of census tracts. As we discuss in the supplementary materials (section 6), future iterations of this approach could help to improve the targeting of humanitarian aid and social welfare, disseminate information to vulnerable populations, and measure the effects of policy interventions.

For this study, we used an anonymized database containing records of billions of interactions on Rwanda's largest mobile phone network and supplemented this with follow-up phone surveys of a geographically stratified random sample of 856 individual subscribers. Upon contacting and surveying each of these individuals, we received informed consent to merge their survey responses with the mobile phone transaction database. The surveys solicited no personally identifying information but contained questions on asset ownership, housing characteristics, and several other basic welfare indicators. From these data, we constructed a composite wealth index using the first principal component of several survey responses related to wealth (*21*, *22*) (supplementary materials section 1D). For each of the 856 respondents, we thus have ~75 survey responses, as well as the historical records of thousands of phone-based interactions such as calls and text messages (Table 1).

We use the merged data from this sample of 856 phone survey respondents to show that a mobile phone subscriber's wealth can be predicted from his or her historical patterns of phone use (Fig. 1A) (cross-validated correlation coefficient $r = 0.68$). Our approach to modeling combines feature engineering with feature selection by first transforming each person's mobile phone transaction logs into a large set of quantitative metrics and then winnowing out metrics

[1]Information School, University of Washington, Seattle, WA 98195, USA. [2]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. [3]School of Information, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. E-mail: joshblum@uw.edu

Table 1. Summary statistics for primary data sets. Phone survey data were collected by the authors in Kigali, in collaboration with the Kigali Institute of Science and Technology. Call detail records were collected by the primary mobile phone operator in Rwanda at the time of the phone survey. Demographic and Health Survey (DHS) data were collected by the Rwandan National Institute of Statistics. N/A, not applicable.

Summary statistic	Phone survey	Call detail records	DHS (2007)	DHS (2010)
Number of unique individuals	856	1.5 million	7377	12,792
Data collection period	July 2009	May 2008–May 2009	Dec. 2007–Apr. 2008	Sept. 2010–Mar. 2011
Number of questions in survey	75	N/A	1615	3396
Primary geographic units	30 districts	30 districts	30 districts	30 districts
Secondary geographic units	300 cell towers	300 cell towers	247 clusters	492 clusters

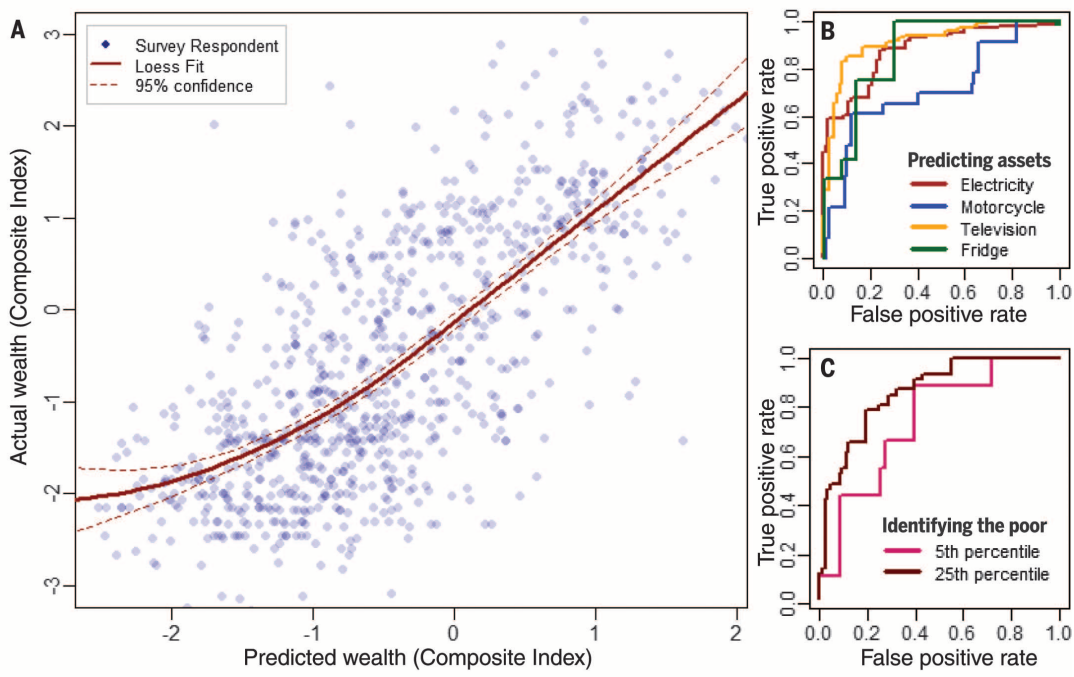

Fig. 1. Predicting survey responses with phone data. (**A**) Relation between actual wealth (as reported in a phone survey) and predicted wealth (as inferred from mobile phone data) for each of the 856 survey respondents. (**B**) Receiver operating characteristic (ROC) curve showing the model's ability to predict whether the respondent owns several different assets. AUC values for electricity, motorcycle, television, and fridge, respectively, are as follows: 0.85, 0.67, 0.84, and 0.88. (**C**) ROC curve illustrates the model's ability to correctly identify the poorest individuals. The poor are defined as those in the 5th percentile (AUC = 0.72) and the 25th percentile (AUC = 0.81) of the composite wealth index distribution.

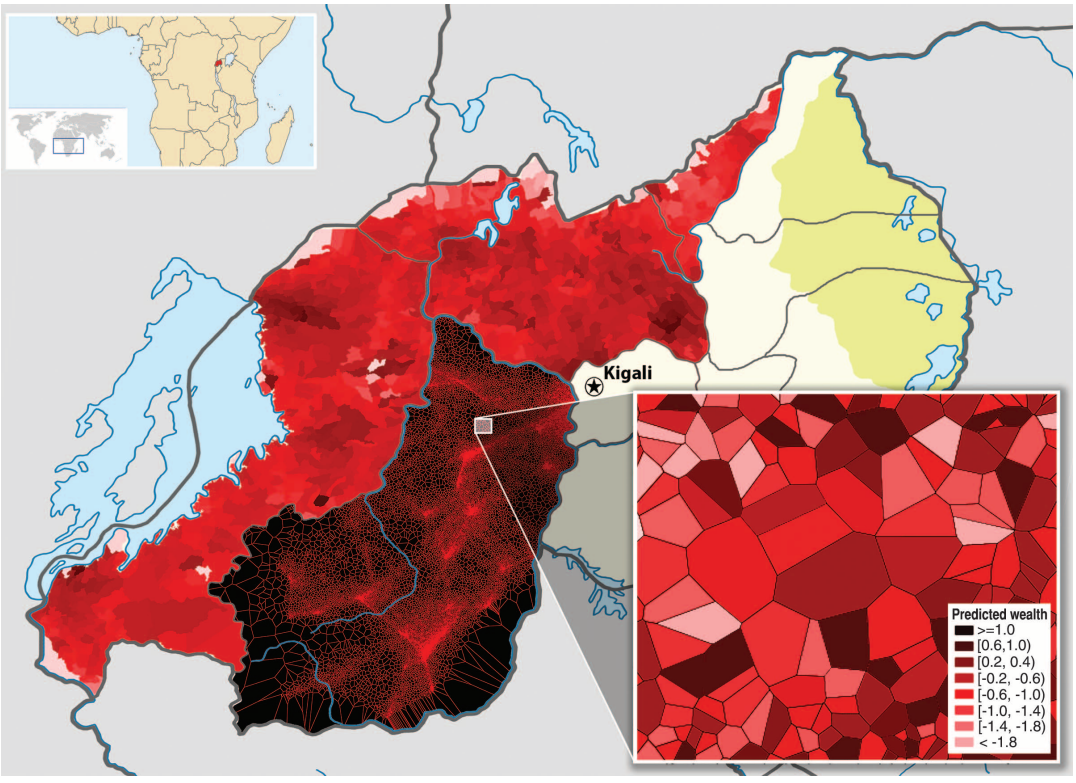

Fig. 2. Construction of high-resolution maps of poverty and wealth from call records. Information derived from the call records of 1.5 million subscribers is overlaid on a map of Rwanda. The northern and western provinces are divided into cells (the smallest administrative unit of the country), and the cell is shaded according to the average (predicted) wealth of all mobile subscribers in that cell. The southern province is overlaid with a Voronoi division that uses geographic identifiers in the call data to segment the region into several hundred thousand small partitions. (**Bottom right inset**) Enlargement of a 1-km^2 region near Kiyonza, with Voronoi cells shaded by the predicted wealth of small groups (5 to 15 subscribers) who live in each region.

that are not predictive of wealth. The first step employs a structured, combinatorial method to automatically generate several thousand metrics from the phone logs that quantify factors such as the total volume, intensity, timing, and directionality of communication; the structure of the individual's contact network; patterns of mobility and migration based on geospatial markers in the data; and so forth. The second step uses "elastic net" regularization to eliminate irrelevant phone metrics and select a parsimonious model that is more likely to generalize (*23*). We use cross-validation to limit the possibility that the model is overfit on the small sample on which it is trained. In the supplementary materials (section 3B), we provide details on these methods and show that comparable results are obtained under a variety of alternative supervised-learning models, including tree-based ensemble regressors and classifiers (*24*). We also show that this two-step approach to feature engineering and model selection performs significantly better than a more intuitive approach based on a small number of hand-crafted metrics (table S1).

In addition to predicting composite wealth, this same approach can be used to estimate, with varying degrees of accuracy, how a phone survey participant will respond to any question, such as whether the respondent owns a motorcycle or has electricity in the household (Fig. 1B and table S1). Cross-validated area-under-the-curve (AUC) scores—which indicate the probability that the model will rank a randomly chosen positive response higher than a randomly chosen negative one—range from 0.50 (no better than random) to 0.88 (quite effective). An analogous method can be used to accurately identify the individuals in the sample who are living below a relative poverty threshold (AUC = 0.72 to 0.81) (Fig. 1C). With further refinement, such methods could prove useful to policy-makers and organizations that target resources to the extreme poor (*25*) (supplementary materials section 6).

For each of these prediction tasks, we use the two-step procedure to select a different model with different metrics and parameters. Although not the focus of our analysis, we note discernible patterns in the set of features identified as the best joint predictors of these different response variables. For instance, features related to an individual's patterns of mobility are generally predictive of motorcycle ownership, whereas factors related to an individual's position within his or her social network are more useful in predicting

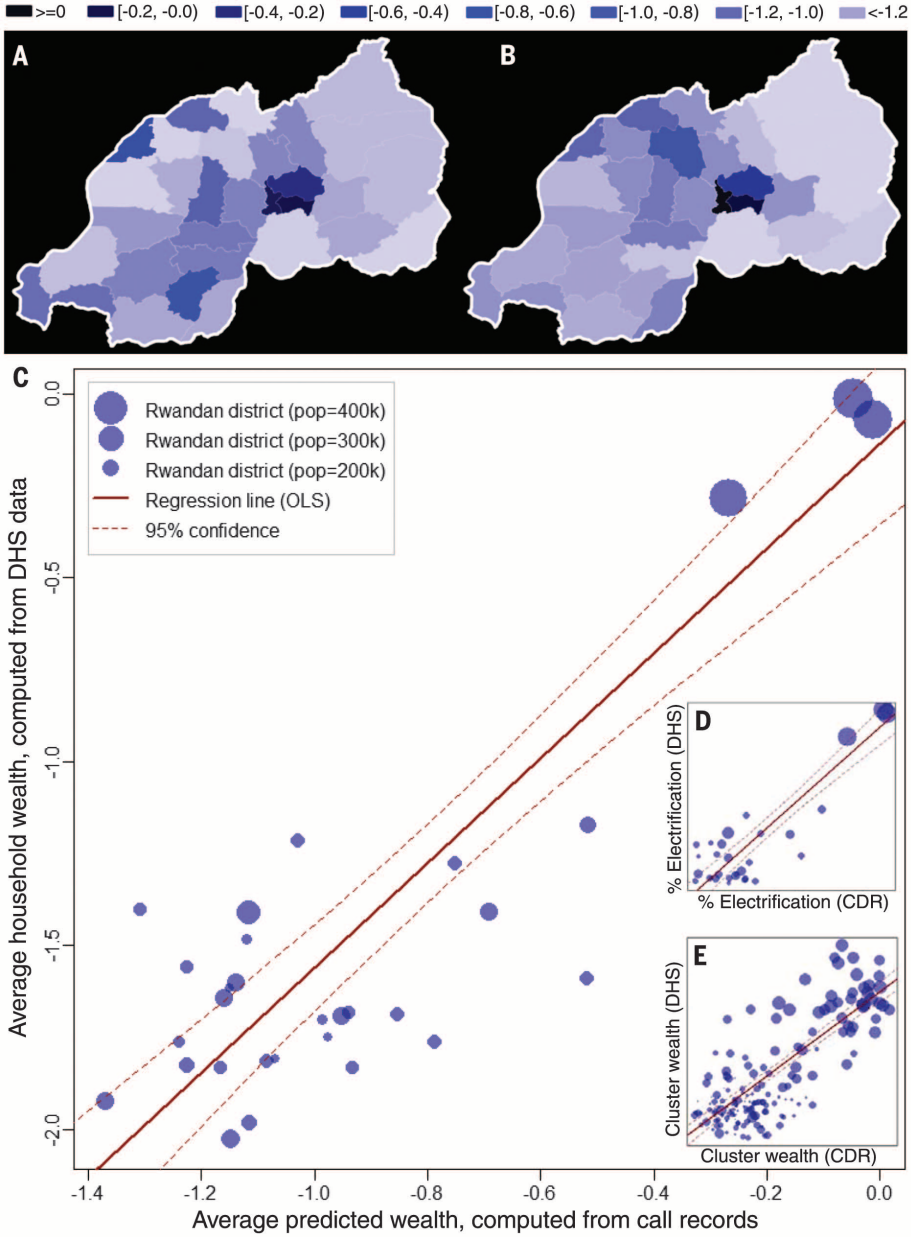

Fig. 3. Comparison of wealth predictions to government survey data. (**A**) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (**B**) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households. (**C**) Comparison of actual and predicted district wealth, for each of the 30 districts, with dots sized by population. (**D**) Comparison of actual and predicted rates of electrification, for each of the 30 districts. (**E**) Comparison of actual and predicted cluster wealth, for each of the 492 DHS clusters. CDR, call detail records.

poverty and wealth (fig. S3). These results suggest that our approach might be generalized to predict a broader class of survey responses, such as the subjective opinions and perceptions of mobile subscribers.

Having fit and cross-validated the model on the phone survey sample—a sample drawn to be representative of all active mobile phone users—we next generate out-of-sample predictions for the characteristics of the remaining 1.5 million Rwandan mobile phone users who did not participate in the survey. Combined with the rich geospatial markers in the phone data, the predicted attributes of millions of individual subscribers enable us to study the geographic distribution of subscriber wealth at an extremely fine degree of spatial granularity (Fig. 2). Whereas public data from Rwanda are only accurate at the level of the district (of which there are 30), the phone data can be used to infer characteristics of each of Rwanda's 2148 cells, as well as small microregions of just a few mobile subscribers (Fig. 2, bottom right inset).

The accuracy of these microregional wealth estimates cannot be directly verified, because no other data set provides wealth information with sufficient geographic resolution. However, when further aggregated to the district level, we can compare the distribution of wealth predicted from the call records of mobile subscribers (Fig. 3A) to the distribution of wealth measured with "ground truth" data collected by the Rwandan government (Fig. 3B). The former estimates are computed by averaging predicted wealth across the thousands of individual mobile phone–based predictions in each of Rwanda's 30 districts; the latter estimates are calculated using data from a nationally representative Demographic and Health Survey (DHS) of 12,792 households, conducted in person by the National Institute of Statistics of Rwanda (*26*). The strong correlation between these two predictions is evident in Fig. 3C and exists whether the ground truth is estimated from only those DHS households that report owning a mobile phone (r = 0.917) or from all households in the survey (r = 0.916). As we discuss in the supplementary materials (section 5A), the first correlation shows that the model's out-of-sample predictions are representative of the population of Rwandan mobile phone owners. The second correlation indicates that in countries like Rwanda, where patterns of mobile phone adoption are similar across regions, this method can provide a close approximation of the distribution of wealth of the full national population. Similar results are obtained when the analysis is disaggregated to the level of the DHS "cluster" (r = 0.79) (Fig. 3E), a geographic unit designed to be comparable to a village. These strong correlations are partially driven by the stark differences between urban and rural areas in Rwanda, but the correlations persist even when comparing clusters within urban or rural areas (fig. S6).

This same approach can be used to predict more than just the average wealth of a district. For instance, rates of district electrification estimated from phone records are comparable to those reported in the DHS survey (r = 0.93) (Fig. 3D). In the urban capital of Kigali, we also find a correlation (r = 0.58) between satellite estimates of night light intensity in 0.55-km^2 grid cells (fig. S7B) and the predicted distribution—based on phone data and the methods described earlier—of responses to the question "Does your household have electricity?" (fig. S7C).

How might such methods be used in practice? In addition to small-area estimation, one promising application is as a source of low-cost, interim national statistics. In many developing economies, long lag times typically occur between successive national surveys. In Angola, for instance, the most recent census before 2014 was conducted in 1970. In that 44-year period, the official population grew by more than 400%. Rwanda has better resources for data collection, and the DHS preceding the 2010 DHS was conducted in 2007. However, even in that relatively short period, the distribution of wealth in Rwanda shifted slightly. Thus, we find that the 2010 distribution of wealth is more accurately reflected in projections based on our analysis of phone data from 2009 than in estimates based on the 2007 DHS (fig. S8). This implies that a policy-maker tasked with targeting the poorest districts in Rwanda would obtain more accurate information from estimates based on mobile phone data than from estimates based on 2007 DHS data (supplementary materials section 6A).

In developing economies, where traditional sources of population data are scarce but mobile phones are increasingly common, these methods may provide a cost-effective option for measuring population characteristics. Whereas a typical national household survey costs more than $1 million and requires 12 to 18 months to complete (*27*), the phone survey we conducted cost only $12,000 and took 4 weeks to administer. Looking forward, the greatest challenge to such work lies in identifying protocols that enable analysis of similar data while respecting the privacy of individual subscribers and the commercial concerns of mobile operators (*28*, *29*). With careful consideration, however, many compelling (and some speculative) applications are within reach, including population monitoring in remote and inaccessible regions, real-time policy evaluation, and the targeting of resources to those with the greatest need.

REFERENCES AND NOTES

1. S. Kuznets, *Am. Econ. Rev.* **45**, 1–28 (1955).
2. G. S. Fields, *World Bank Res. Obs.* **4**, 167–185 (1989).
3. M. Jerven, *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It* (Cornell Univ. Press, Ithaca, NY, 2013).
4. C. Elbers, J. O. Lanjouw, P. Lanjouw, *Econometrica* **71**, 355–364 (2003).
5. M. Ghosh, J. N. K. Rao, *Stat. Sci.* **9**, 55–76 (1994).
6. D. Lazer *et al.*, *Science* **323**, 721–723 (2009).
7. G. King, *Science* **331**, 719–721 (2011).
8. N. Eagle, M. Macy, R. Claxton, *Science* **328**, 1029–1031 (2010).
9. H. Choi, H. Varian, *Econ. Rec.* **88**, 2–9 (2012).
10. W. Wang, D. Rothschild, S. Goel, A. Gelman, *Int. J. Forecast.* **31**, 980–991 (2015).
11. "The mobile economy 2014" (GSMA Intelligence, 2014); www.gsmamobileeconomy.com/GSMA_ME_Report_2014_R2_WEB.pdf.
12. J. Candia *et al.*, *J. Phys. A* **41**, 224015 (2008).
13. J.-P. Onnela *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332–7336 (2007).
14. G. Palla, A. L. Barabási, T. Vicsek, *Nature* **446**, 664–667 (2007).
15. M. C. González, C. A. Hidalgo, A.-L. Barabási, *Nature* **453**, 779–782 (2008).
16. X. Lu, E. Wetter, N. Bharti, A. J. Tatem, L. Bengtsson, *Sci. Rep.* **3**, 2923 (2013).
17. J. E. Blumenstock, *Inf. Technol. Dev.* **18**, 107–125 (2012).
18. V. Frias-Martinez, J. Virseda, in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (Association for Computing Machinery, New York, 2012), pp. 76–84; http://doi.acm.org/10.1145/2160673.2160684.
19. P. Deville *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15888–15893 (2014).
20. G. C. Cawley, N. L. C. Talbot, *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
21. D. Filmer, L. H. Pritchett, *Demography* **38**, 115–132 (2001).
22. J. Blumenstock, N. Eagle, *Inf. Technol. Int. Dev.* **8**, 1–16 (2012).
23. H. Zou, T. Hastie, *J. R. Stat. Soc. Ser. B* **67**, 301–320 (2005).
24. L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and Regression Trees* (Chapman and Hall/CRC Press, New York, ed. 1, 1984).
25. B. Abelson, K. R. Varshney, J. Sun, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2014), pp. 1563–1572; http://doi.acm.org/10.1145/2623330.2623335.
26. National Institute of Statistics of Rwanda (NISR) [Rwanda], Ministry of Health (MOH) [Rwanda], ICF International, "Rwanda Demographic and Health Survey 2010," *DHS Final Reports* (publication ID FR259, NISR, MOH, and ICF International, Calverton, MD, 2012).
27. M. Jerven, "Benefits and costs of the data for development targets for the post-2015 development agenda," in *Data for Development Assessment Paper* (Copenhagen Consensus Center, 2014).
28. Y.-A. de Montjoye, L. Radaelli, V. K. Singh, A. S. Pentland, *Science* **347**, 536–539 (2015).
29. A. Wesolowski *et al.*, *PLOS Curr.* 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e (2014).

ACKNOWLEDGMENTS

We received approval for this study from the University of Washington Human Subjects Division (protocol 44933) and the University of California Committee for Protection of Human Subjects (protocol 200949). We thank N. Eagle for providing access to the mobile phone records, Y. Yao for research assistance, S. Kumaran and the faculty and students at the Kigali Institute of Science and Technology for help in coordinating the phone survey, and N. Musaninkindi and the Rwandan National Institute of Statistics for assistance with DHS data. J.B. is supported by the NSF (Doctoral Dissertation Award 1025103); the Institute for Money, Technology, and Financial Inclusion (grant 2010-2366); and the Gates Foundation (grant OPP1106936). We do not have any real or apparent conflicts of interest. Mobile phone data were supplied by an anonymous service provider in Rwanda and are not available for distribution. All other data and code, including all intermediate data needed to replicate these results and apply these methods in other contexts, are available through the Inter-university Consortium for Political and Social Research (http://doi.org/10.3886/E50592V2).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/350/6264/1073/suppl/DC1
Materials and Methods
Figs. S1 to S8
Table S1
References (*30*–*51*)

29 April 2015; accepted 19 October 2015
10.1126/science.aac4420