Limits of Predictability in Human Mobility

Chaoming Song *et al.*

Science **327**, 1018 (2010);
DOI: 10.1126/science.1177170

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

criterion: They have higher energies in the experimentally biased optimization.

For the proteins in our set in the ~30-kD molecular-weight range, the computed structures are not completely converged and have large disordered regions. This is clearly a sampling problem because the native structure has lower energy (Fig. 4C and fig. S3); even with the NMR data as a guide, Rosetta trajectories fail to sample very close to the native state. Increased convergence on the low-energy native state can be achieved either by collecting and using additional experimental data (1ilb_2 in fig. S3) or by improved sampling. Though at present the former is the more reliable solution, the latter will probably become increasingly competitive as the cost of computing decreases and conformational search algorithms improve.

We have shown that accurate structures can be computed for a wide range of proteins using backbone-only NMR data. These results suggest a change in the traditional NOE-constraint–based approach to NMR structure determination (fig. S4). In the new approach, the bottlenecks of side-chain chemical-shift assignment and NOESY assignment are eliminated, and instead, more backbone information is collected: RDCs in one or more media and a small number of unambiguous H^N-H^N constraints from three- or four-dimensional experiments, which restrict possible β-strand registers. Advantages of the approach are that ^1H,^{15}N-based NOE and RDC data quality is relatively unaffected in slower tumbling, larger proteins and that the analysis of resonance and NOESY peak assignments can be done in a largely automated fashion with fewer opportunities for error. The approach is compatible with deuteration necessary for proteins greater than 15 kD and, for larger proteins, can be extended to include methyl NOEs on selectively protonated samples. The method should also enable a more complete structural characterization of transiently populated states (*25*) for which the available data are generally quite sparse.

References and Notes

1. D. E. Zimmerman *et al.*, *J. Mol. Biol.* **269**, 592 (1997).
2. C. Bartels, P. Güntert, M. Billeter, K. Wüthrich, *J. Comput. Chem.* **18**, 139 (1998).
3. M. C. Baran, Y. J. Huang, H. N. Moseley, G. T. Montelione, *Chem. Rev.* **104**, 3541 (2004).
4. Y. S. Jung, M. Zweckstetter, *J. Biomol. NMR* **30**, 11 (2004).
5. W. Lee, W. M. Westler, A. Bahrami, H. R. Eghbalnia, J. L. Markley, *Bioinformatics* **25**, 2085 (2009).
6. M. Berjanskii *et al.*, *Nucleic Acids Res.* **37** (Web Server issue), W670 (2009).
7. Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J. Biomol. NMR* **44**, 213 (2009).
8. G. Kontaxis, F. Delaglio, A. Bax, *Methods Enzymol.* **394**, 42 (2005).
9. I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, *Dalton Trans.* **29**, 3782 (2008).
10. J. H. Prestegard, C. M. Bougault, A. I. Kishore, *Chem. Rev.* **104**, 3519 (2004).
11. S. Grzesiek, A. Bax, *J. Biomol. NMR* **3**, 627 (1993).
12. D. M. LeMaster, F. M. Richards, *Biochemistry* **27**, 142 (1988).
13. K. H. Gardner, M. K. Rosen, L. E. Kay, *Biochemistry* **36**, 1389 (1997).
14. G. Wagner, *J. Biomol. NMR* **3**, 375 (1993).
15. R. Das, D. Baker, *Ann. Rev. Biochem.* **77**, 363 (2008).
16. P. Bradley, K. M. S. Misura, D. Baker, *Science* **309**, 1868 (2005).
17. Materials and methods are available as supporting material on *Science* Online.
18. A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9615 (2007).
19. Y. Shen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4685 (2008).
20. J. A. Losonczi, M. Andrec, M. W. Fischer, J. H. Prestegard, *J. Magn. Reson.* **138**, 334 (1999).
21. C. A. Rohl, D. Baker, *J. Am. Chem. Soc.* **124**, 2723 (2002).
22. B. Qian *et al.*, *Nature* **450**, 259 (2007).
23. T. J. Brunette, O. Brock, *Proteins* **73**, 958 (2008).
24. X. Wang, T. Weldeghiorghis, G. Zhang, B. Imperiali, J. H. Prestegard, *Structure* **16**, 965 (2008).
25. H. van Ingen, D. M. Korzhnev, L. E. Kay, *J. Phys. Chem. B* **113**, 9968 (2009).
26. D. A. Snyder, G. T. Montelione, *Proteins* **59**, 673 (2005).
27. The FindCore algorithm is available at http://fps.nesg.org.
28. A. Zemla, *Nucleic Acids Res.* **31**, 3370 (2003).

29. We are thankful to the U.S. Department of Energy Innovative and Novel Computational Impact on Theory and Experiment Award for providing access to the Blue Gene/P supercomputer at the Argonne Leadership Computing Facility and to Rosetta@home participants for their generous contributions of computing power.
We thank Y. Shen and A. Bax for fruitful discussions; Y. J. Huang and Y. Tang for their contribution during preliminary studies using sparse NOE constraints with CS-Rosetta; S. Bansal, H.-w. Lee, and Y. Liu for collection of RDC data, A. Lemak for providing the Crystallography and NMR System RDC refinement protocol, and the NESG consortium for access to other unpublished NMR data that has facilitated methods development. S.R., O.F.L., P.R., G.T.M., and D.B. designed research; S.R. designed and tested the CS-RDC-Rosetta protocol; O.F.L. designed and tested the iterative CS-RDC-NOE-Rosetta protocol; M.T. developed the all-atom refinement protocol; S.R., O.F.L. and D.B. designed and performed research for energy based structure validation; X.W and J.P analyzed the ALG13 ensemble; J.A, G.L, T.R, A.E, M.K, and T.S provided blind NMR data sets; and S.R., O.F.L., P.R., G.T.M., and D.B. wrote the manuscript. This work was supported by the Human Frontiers of Science Program (O.F.L.), NIH grant GM76222 (D.B.), the HHMI, the National Institutes of General Medical Science Protein Structure Initiative program grant U54 GM074958 (G.T.M.), and the Research Resource grant RR005351 (J.P.). M.T. holds a Sir Henry Wellcome Postdoctoral Fellowship. RDC and Paramagnetic Relaxation Enhancement data as deposited in the Protein Data Bank (PDB) with accession code 2jzc.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1183649/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 to S3
References

21 October 2009; accepted 14 January 2010
Published online 4 February 2010;
10.1126/science.1183649
Include this information when citing this paper.

Limits of Predictability in Human Mobility

Chaoming Song,[1,2] Zehui Qu,[1,2,3] Nicholas Blumm,[1,2] Albert-László Barabási[1,2]*

A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to foresee the whereabouts and mobility of individuals, raising a fundamental question: To what degree is human behavior predictable? Here we explore the limits of predictability in human dynamics by studying the mobility patterns of anonymized mobile phone users. By measuring the entropy of each individual's trajectory, we find a 93% potential predictability in user mobility across the whole user base. Despite the significant differences in the travel patterns, we find a remarkable lack of variability in predictability, which is largely independent of the distance users cover on a regular basis.

When it comes to the emerging field of human dynamics, there is a fundamental gap between our intuition and the current modeling paradigms. Indeed, although we rarely perceive any of our actions to be random, from the perspective of an outside observer who is unaware of our motivations and schedule, our activity pattern can easily appear random and unpredictable. Therefore, current models of human activity are fundamentally stochastic (*1*) from Erlang's formula (*2*) used in telephony to Lévy-walk models describing human mobility (*3*–*7*) and their applications in viral dynamics (*8*–*10*), queuing models capturing human communication patterns (*11*–*13*), and models capturing body balancing (*14*) or panic (*15*). Yet the probabilistic nature of the existing modeling framework raises fundamental questions: What is the role of randomness in human behavior and to what degree are individual human actions predictable? Our goal here is to quantify

[1]Center for Complex Network Research, Departments of Physics, Biology, and Computer Science, Northeastern University, Boston, MA 02115, USA. [2]Department of Medicine, Harvard Medical School, and Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [3]School of Computer Science and Engineering, University of Electric Science and Technology of China, Chengdu 610054, China.

*To whom correspondence should be addressed. E-mail: alb@neu.edu

the interplay between the regular and thus predictable and the random and thus unforeseeable, probing through human mobility the fundamental limits that characterize the predictability of human dynamics.

At present, the most detailed information on human mobility across a large segment of the population is collected by mobile phone carriers (*4*, *16*–*21*). Mobile carriers record the closest mobile tower each time the user uses his or her phone. Here we use a 3-month-long record, collected for billing purposes and anonymized by the data source, capturing the mobility patterns of 50,000 individuals chosen from ~10 million anonymous mobile phone users with the criteria that they visit more than two locations (tower vicinity) during the observational period and that their average call frequency f is ≥0.5 hour^{-1} [(*22*) sections S1 and S2].

The trajectories of two users with widely different mobility patterns are shown in Fig. 1A: The first user moves in the vicinity of $N = 22$ towers in a 30-km region, whereas the second visits as many as $N = 76$ towers spanning approximately a 90-km neighborhood. To understand the recurrent nature of individual mobility, we assigned to each user a mobility network (*23*) (Fig. 1B), in which nodes are the locations visited by the user (each location corresponding to a mobile phone tower, with about a 3-km^2 reception area on average, representing the uncertainty in our ability to determine the user's whereabouts), and links represent the observed movements between these. The uneven node sizes, corresponding to the percentage of time the user spent in the vicinity of the particular tower, indicate that individuals tend to spend most of their time in a few selected locations. Finally, each mobility network has an associated dynamical pattern (Fig. 1C), capturing the temporal sequence of towers visited by the user.

Entropy is probably the most fundamental quantity capturing the degree of predictability characterizing a time series (*24*). We assign three entropy measures to each individual's mobility pattern: (i) The random entropy $S_i^{\mathrm{rand}} \equiv \log_2 N_i$, where N_i is the number of distinct locations visited by user i, capturing the degree of predictability of the user's whereabouts if each location is visited with equal probability; (ii) the temporal-uncorrelated entropy $S_i^{\mathrm{unc}} \equiv -\sum_{j=1}^{N_i} p_i(j) \log_2 p_i(j)$, where $p_i(j)$ is the historical probability that location j was visited by the user i, characterizing the heterogeneity of visitation patterns; (iii) the actual entropy, S_i, which depends not only on the frequency of visitation, but also the order in which the nodes were visited and the time spent at each location, thus capturing the full spatiotemporal order present in a person's mobility pattern. To be specific, if $T_i = \{X_1, X_2, \cdots, X_L\}$ denotes the sequence of towers at which user i was observed at each consecutive hourly interval, the entropy S_i is given by $-\sum_{T_i' \subset T_i} P(T_i') \log_2 [P(T_i')]$, where $P(T_i')$ is the probability of finding a particular time-ordered subsequence T_i' in the trajectory T_i [(*22*) section *S4*]. Naturally, for each user, $S_i \leq S_i^{\mathrm{unc}} \leq S_i^{\mathrm{rand}}$.

To calculate the real entropy S_i, we need a continuous (e.g., hourly) record of a user's momentary location. Mobile phone records provide location information only when a person uses his or her phone. The users tend to place most of their calls in short bursts (*11*–*13*, *25*) (Fig. 1D), followed by long periods with no call activity, during which we have no information about the user's location (Fig. 1C). This incompleteness of the collected data is captured by the parameter q, representing the fraction of hour-long intervals when the user's location is unknown to us. As Fig. 1E shows, $P(q)$ across our user base peaked around $q = 0.7$, which indicated that, for a typical user, we have no location update for about 70% of the hourly intervals, which masks the user's real entropy S_i. We therefore studied the dependence of the entropy $S(q)$ on the incompleteness q, which

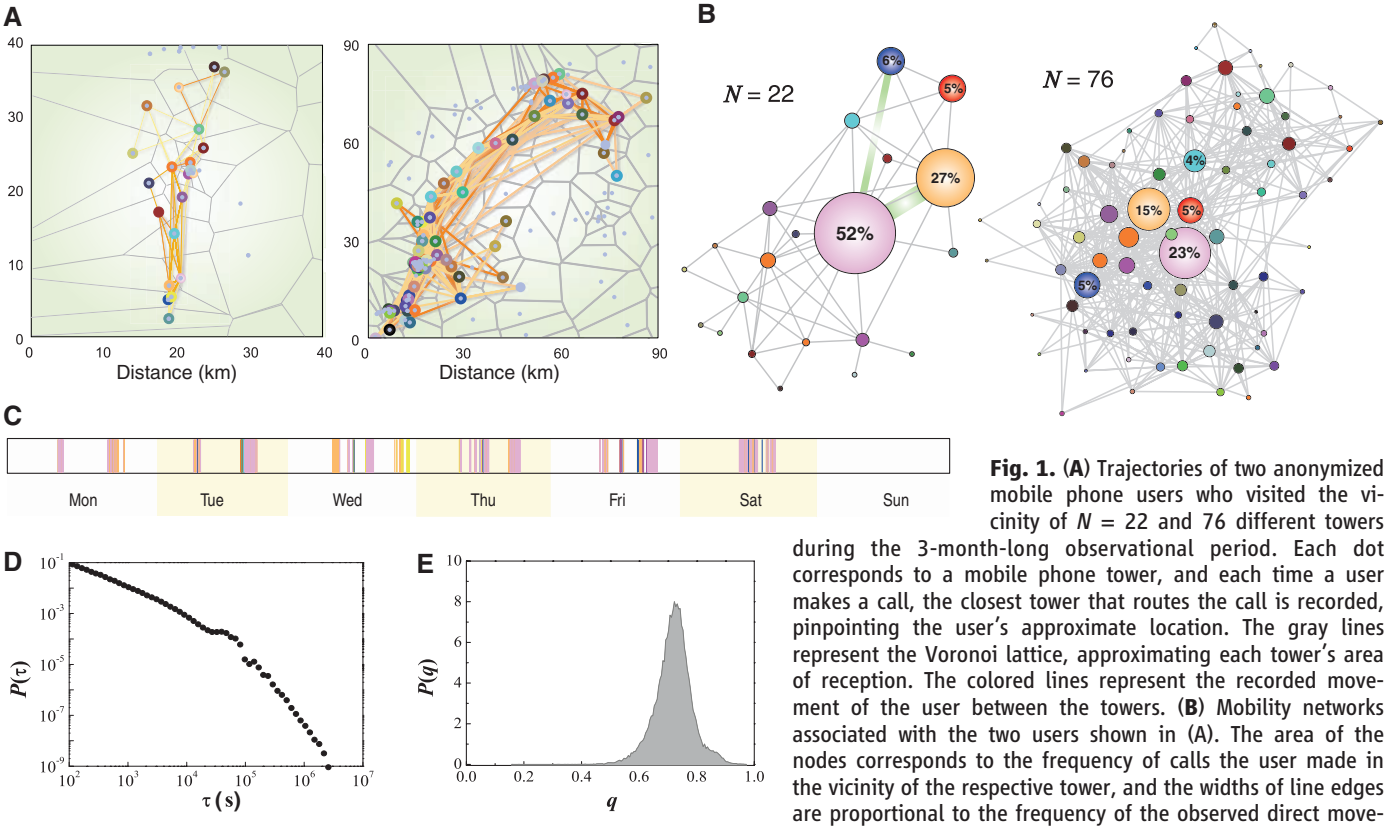

Fig. 1. (**A**) Trajectories of two anonymized mobile phone users who visited the vicinity of $N = 22$ and 76 different towers during the 3-month-long observational period. Each dot corresponds to a mobile phone tower, and each time a user makes a call, the closest tower that routes the call is recorded, pinpointing the user's approximate location. The gray lines represent the Voronoi lattice, approximating each tower's area of reception. The colored lines represent the recorded movement of the user between the towers. (**B**) Mobility networks associated with the two users shown in (A). The area of the nodes corresponds to the frequency of calls the user made in the vicinity of the respective tower, and the widths of line edges are proportional to the frequency of the observed direct movement between two towers. (**C**) A week-long call pattern that captures the time-dependent location of the user with $N = 22$. Each vertical line corresponds to a call, and its color matches the tower from where the call was placed. This sequence of locations serves as the basis of our mobility prediction. (**D**) The distribution of the time intervals between consecutive calls, τ, across the whole user population, documenting the nature of the call pattern as coming in bursts (*11*). (**E**) The distribution of the fraction of unknown locations, q, representing the hourly intervals when the user did not make a call, and thus his or her location remains unknown to us.

allowed us to extrapolate the entropy to $q = 0$. We tested the method's accuracy on the trajectory of 100 users whose whereabouts were recorded every hour [(*22*) section *S4*] and found that it performed well for $q < 0.8$, which represented 92% of the users in our data set. We therefore removed 5000 users with the highest q from our data set, which ensured that all remaining 45,000 users satisfied $q < 0.8$.

To characterize the inherent predictability across the user population, we determined S_i, S_i^{unc}, and S_i^{rand} for each user i; the obtained $P(S)$, $P(S^{unc})$, and $P(S^{rand})$ distributions are shown in Fig. 2A. The most striking result is the prominent shift of $P(S)$ compared with $P(S^{rand})$. Indeed, $P(S^{rand})$ peaks at $S^{rand} \approx 6$, which indicates that, on average, each update of the user's location represents six bits per hour of new information; that is, a user who chooses randomly his or her next location could be found on average in any of $2^{S^{rand}} \approx 64$ locations. In contrast, the fact that $P(S)$ peaks at $S = 0.8$ indicates that the real uncertainty in a typical user's whereabouts is not 64 but $2^{0.8} = 1.74$, i.e., fewer than two locations.

The typical distances covered by individuals during their daily mobility pattern, as captured by each user's radius of gyration, r_g, follows a fat-tailed distribution (*4*), which indicates that, although most individuals' daily activity is confined to a limited neighborhood of 1 to 10 km, a few users regularly cover hundreds of kilometers (fig. S2). These differences suggest that predictability should also follow a fat-tailed distribution. In other words, we expect that individuals who travel less should be easy to predict (small entropy), whereas those with large r_g should be much less predictable (high entropy).

An important measure of predictability is the probability Π that an appropriate predictive algorithm can predict correctly the user's future whereabouts. This quantity is subject to Fano's inequality (*24*, *26*). That is, if a user with entropy S moves between N locations, then her or his predictability $\Pi \leq \Pi^{max}(S,N)$, where Π^{max} is given by $S = H(\Pi^{max}) + (1 - \Pi^{max}) \log_2(N - 1)$ with the binary entropy function $H(\Pi^{max}) = -\Pi^{max} \log_2(\Pi^{max}) - (1 - \Pi^{max}) \log_2(1 - \Pi^{max})$. For a user with $\Pi^{max} = 0.2$, this means that at least 80% of the time the individual chooses his location in a manner that appears to be random, and only in the remaining 20% of the time can we hope to predict his or her whereabouts. In other terms, no matter how good our predictive algorithm, we cannot predict with better than 20% accuracy the future whereabouts of a user with $\Pi^{max} = 0.2$. Therefore, Π^{max} represents the fundamental limit for each individual's predictability.

We determined Π^{max} separately for each user in the database. To our surprise, we found that $P(\Pi^{max})$ does not follow the fat-tailed distribution suggested by the travel distances, but it is narrowly peaked near $\Pi^{max} \approx 0.93$ (Fig. 2B). This highly bounded distribution indicates that, despite the apparent randomness of the individuals' trajectories, a historical record of the daily mobility pattern of the users hides an unexpectedly high degree of potential predictability. We have also determined the maximal predictability Π^{unc} and the random predictability Π^{rand} extracted from S^{unc} and S^{rand}. As Fig. 2B shows, the result is strikingly different—$P(\Pi^{unc})$ is extremely widely distributed and peaked at $\Pi^{unc} \sim 0.3$, which indicates that, if we rely only on the heterogeneous spatial distribution, the predictability across the whole population is insignificant and varies widely from person to person. Similarly, $P(\Pi^{rand})$ has a peak at $\Pi^{rand} = 0$, which suggests not only that Π^{rand} and Π^{unc} are ineffective as predictive tools, but also that a significant share of predictability is encoded in the temporal order of the visitation pattern.

How can we reconcile the wide variability in the observed travel distances, as captured by the fat-tailed $P(r_g)$, with the highly bounded predictability observed across the user population? To answer this, we measured the dependency of Π^{max} on r_g, and found that, for $r_g \geq 10$ km, predictability becomes largely independent of r_g, saturating at $\Pi^{max} \approx 0.93$ (Fig. 2C). Therefore, Fig. 2C explains the failure of our earlier hypothesis: Individuals with $r_g \geq 100$ km, covering hundreds of kilometers on a regular basis, are just as predictable as those whose life is constrained to a $r_g \approx 10$-km neighborhood, a saturation that lies behind the high predictability observed across the whole user base.

To determine how much of our predictability is really rooted in the visitation patterns of the top locations, we calculated the probability $\tilde{\Pi}$ that, in a given moment, the user is in one of the top n most visited locations, where $n = 2$ typically captures home and work. Thus, $\tilde{\Pi}$ represents an upper bound for Π^{max}, as, even if our predictive algorithm is 100% accurate, it can foresee the future location only when the user is found in one of the top n locations monitored by the algorithm. As Fig. 2D shows, the top two locations ($n = 2$)

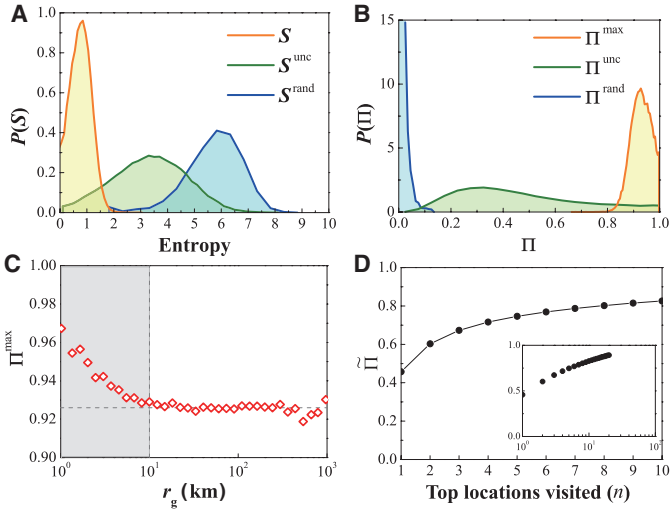

Fig. 2. (**A**) The distribution of the entropy S, the random entropy S^{rand}, and the uncorrelated entropy S^{unc} across 45,000 users. (**B**) The distribution of the Π^{max}, Π^{rand}, and Π^{unc} across all users. (**C**) The dependence of predictability Π^{max} on the user's radius of gyration r_g, capturing the distances regularly covered by each user. Note that, for $r_g >$ 10 km, Π^{max} is largely independent of r_g, saturating at $\Pi^{max} \approx 0.93$. (**D**) The fraction of time a user spends in the top n most visited locations, the resulting measure $\tilde{\Pi}$ representing an upper bound of predictability Π^{max}. Thus for $n = 1$ we can predict correctly the user's location when he or she is found at the most likely location ("home"), whereas for $n = 2$, we can predict correctly the user's whereabouts when he is in one of his top two locations ("home" or "office"). (Inset) $\tilde{\Pi}$ appears to grow approximately logarithmically with n.

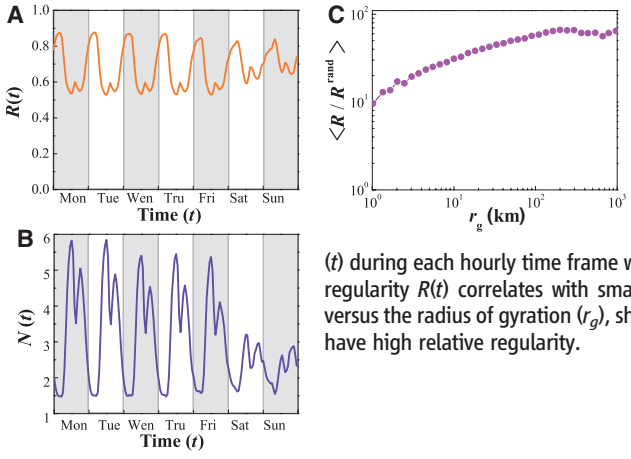

Fig. 3. (**A**) The hourly regularity $R(t)$ over a week-long time period, measuring the fraction of instances when the user is found in his or her most visited location during the corresponding hour-long period. (**B**) The average number of visited locations $N(t)$ during each hourly time frame within a week, revealing that high regularity $R(t)$ correlates with small $N(t)$. (**C**) The averaged R/R^{rand} versus the radius of gyration (r_g), showing that the users with large r_g have high relative regularity.

offer only a 60% overall predictability. Gradually adding more locations increases $\tilde{\Pi}(n)$, but we need several dozen distinct locations to converge to $\tilde{\Pi} = 1$ (Fig. 2D, inset).

To understand the origin of the observed high potential predictability, we segmented each week into $24 \times 7 = 168$ hourly intervals, and within each hour, we identified for each user the most visited location (Fig. 3, A and B). For example, if, between 8 and 9 a.m. on Monday, a user was found 10 times at tower 1, twice at tower 2, and once at tower 3, we assumed that her most likely location during this hour will be tower 1. Next, we measured each user's regularity, R, defined as the probability of finding the user in his most visited location during that hour. R represents a lower bound for predictability Π, as it ignores the temporal correlations in user mobility. We found that across the whole user base, $R \approx 0.7$, which meant that, on average, 70% of the time the most visited location coincides with the user's actual location. The pattern is time dependent: During the night, when most people tend to be reliably at home, R peaks at ≈ 0.9, but between noon and 1 p.m. and between 6 and 7 p.m., R has clear minima, corresponding to transition periods (travel to lunch or home). Indeed, if we measure the total number of distinct locations $N(t)$ a user visited each hour (Fig. 3B), we find that moments of low regularity R correspond to significant increase in $N(t)$, a signature of high mobility, and when R peaks there is a drop in $N(t)$.

If the users were to move randomly between their N locations, then $R^{\text{rand}} = 1/N$, which is $1/2^{S^{\text{rand}}} \approx 0.016$, an order of magnitude smaller than the observed $R \approx 0.7$. This gap once again indicates that the high regularity characterizing each user's mobility represents a significant departure from the expectation that they will be random. In Fig. 3C, we plot the relative regularity R/R^{rand} as a function of r_g, observing a clearly increasing tendency. That is, counterintuitively, the relative regularity of users who travel the most (i.e., have high r_g) is higher than the relative regularity of the more homebound individuals.

To explore whether demographic factors influence the users' regularity and predictability, we measured R and $\Pi^{\max}$ for different age and gender groups (fig. S10). It was surprising that we did not observe gender- or age-based differences in $\Pi^{\max}$, but only a systematic, but statistically insignificant, gender-based difference emerged in regularity. We also explored the impact of home, language groups, population density, and rural versus urban environment on predictability and found only insignificant variations (figs. S8, S12, and S13). Finally, we did not find significant changes in user regularity over the weekends compared with their weekday mobility (fig. S8), which suggested that regularity is not imposed by the work schedule, but potentially is intrinsic to human activities.

In summary, the combination of the empirically determined user entropy and Fano's inequality indicates that there is a potential 93% average predictability in user mobility, an exceptionally high value rooted in the inherent regularity of human behavior. Yet it is not the 93% predictability that we find the most surprising. Rather, it is the lack of variability in predictability across the population. Indeed, given the fat-tailed distribution of the distances over which users travel on a regular basis (see fig. S2), most individuals are well localized in a finite neighborhood, but a few travel widely. Furthermore, a number of demographic and external parameters, from age to population density and the number of towers visited, vary widely from user to user. It is not unreasonable to expect, therefore, that predictability should also vary widely: For people who travel little, it should be easier to foresee their location, whereas those who regularly cover hundreds of kilometers should have a low predictability. Despite this inherent population heterogeneity, the maximal predictability varies very little—indeed $P(\Pi^{\max})$ is narrowly peaked at 93%, and we see no users whose predictability would be under 80%.

Although making explicit predictions on user whereabouts is beyond our goals here, appropriate data-mining algorithms (*19*, *20*, *27*) could turn the predictability identified in our study into actual mobility predictions. Most important, our results indicate that when it comes to processes driven by human mobility, from epidemic modeling to urban planning and traffic engineering, the development of accurate predictive models is a scientifically grounded possibility, with potential impact on our well-being and public health. At a more fundamental level, they also indicate that, despite our deep-rooted desire for change and spontaneity, our daily mobility is, in fact, characterized by a deep-rooted regularity.

References and Notes

1. C. Castellano, S. Fortunato, V. Loreto, *Rev. Mod. Phys.* **81**, 591 (2009).
2. A. K. Erlang, *Nyt Tidsskrift for Matematik B* **20**, 33 (1909).
3. D. Brockmann, L. Hufnagel, T. Geisel, *Nature* **439**, 462 (2006).
4. M. C. González, C. A. Hidalgo, A.-L. Barabási, *Nature* **453**, 779 (2008).
5. S. Havlin, D. Ben-Avraham, *Adv. Phys.* **51**, 187 (2002).
6. R. N. Mantegna, H. E. Stanley, *Phys. Rev. Lett.* **73**, 2946 (1994).
7. R. Metzler, A. V. Chechkin, V. Y. Gonchar, J. Klafter, *Chaos Solitons Fractals* **34**, 129 (2007).
8. V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, A. Vespignani, *PLoS Med.* **4**, e13 (2007).
9. L. Hufnagel, D. Brockmann, T. Geisel, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15124 (2004).
10. P. Wang, M. C. González, C. A. Hidalgo, A.-L. Barabási, *Science* **324**, 1071 (2009).
11. A.-L. Barabási, *Nature* **435**, 207 (2005).
12. G. Grinstein, R. Linsker, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **77**, 012101 (2008).
13. A. Gabrielli, G. Caldarelli, *Phys. Rev. Lett.* **98**, 208701 (2007).
14. J. D. Harry, J. B. Niemi, A. A. Priplata, J. J. Collins, *IEEE Spectr.* **42**, 36 (2005).
15. D. Helbing, I. Farkas, T. Vicsek, *Nature* **407**, 487 (2000).
16. W.-S. Sho, H. S. Kim, *IEEE Commun. Mag.* **41**, 86 (2002).
17. C. A. Hidalgo, C. Rodriguez-Sickert, *Physica A* **387**, 3017 (2008).
18. M. C. González, P. G. Lind, H. J. Herrmann, *Phys. Rev. Lett.* **96**, 088702 (2006).
19. N. Eagle, A. Pentland, *Pers. Ubiquitous Comput.* **10**, 255 (2006).
20. N. Eagle, A. Pentland, *Behav. Ecol. Sociobiol.* **63**, 1057 (2009).
21. R. Lambiotte *et al.*, *Physica A* **387**, 5317 (2008).
22. Materials and methods are available as supporting material on *Science* Online.
23. J. R. Banavar, A. Maritan, A. Rinaldo, *Nature* **399**, 130 (1999).
24. N. Navet, S.-H. Chen, *Natural Computing in Computational Finance* (Springer, Berlin, 2008).
25. A. Vázquez *et al.*, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **73**, 036127 (2006).
26. R. M. Fano, *Transmission of Information* (the MIT Press and Wiley, New York and London, 1961).
27. L. Song, D. Kotz, R. Jain, X. He, *Proc. 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)* **2**, 1414 (2004).

28. We thank M. Gonzales, P. Wang, J. Bagrow, and J. A. Aslam for discussions and comments on the manuscript. This work was supported by the James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems, NSF within the Information Technology Research (DMR-0426737) and IIS-0513650 programs, the Defense Threat Reduction Agency award HDTRA1-08-1-0027, and the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under agreement number W911NF-09-2-0053. Z.Q. was supported by a fellowship from the China Scholarship Council.

Supporting Online Material

www.sciencemag.org/cgi/content/full/327/5968/1018/DC1
Materials and Methods
SOM Text
Figs. S1 to S13
References

2 June 2009; accepted 28 December 2009
10.1126/science.1177170