The fundamental structures of dynamic social networks

Vedran Sekara[1], Arkadiusz Stopczynski[1,2] & Sune Lehmann[1,3]

[1]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark,*
[2]*Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA,*
[3]*The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.*

Networks provide a powerful mathematical framework for analyzing the structure and dynamics of complex systems[1–3]. The study of group behavior has deep roots in the social science literature[4,5] and community detection is a central part of modern network science. Network communities have been found to be highly overlapping and organized in a hierarchical structure[6–9]. Recent technological advances have provided a toolset for measuring the detailed social dynamics at scale[10,11]. In spite of great progress, a quantitative description of the complex temporal behavior of social groups—with dynamics spanning from minute-by-minute changes to patterns expressed on the timescale of years—is still absent. Here we uncover a class of fundamental structures embedded within highly dynamic social networks. On the shortest time-scale, we find that social gatherings are fluid, with members coming and going, but organized via a stable core of individuals. We show that cores represent social contexts[9], with recurring meetings across weeks and months, each with varying degrees of regularity. In this sense, cores provide a vocabulary which quantifies the patterns of social life. The simplification is so powerful that participation in social contexts is predictable with high precision. Our results offer new perspectives for modeling and understanding of processes in social systems, with applications including epidemiology, social contagion, urban planning, digital economy, and quantitative sociology.

Human societies, organizations, and communities are characterized by structure and regularity. For example, we have recently seen impressive progress in understanding the basic laws that govern human mobility[12,13]. Our understanding of the temporal behavior of social interactions is much less complete[14]. For this reason, recent models for dynamic complex networks[15,16] are based on complicated mathematical heuristics and report network structures that are difficult to interpret and act upon. Using data with minute-by-minute resolution, the apparent complexity underlying social interactions vanishes, and we are able to observe basic structural elements directly and without ambiguity. When time slices are shorter than the turnover rate, *gatherings* of individuals can be observed directly (Fig. 1A-B). From minute to minute, members join and leave such gatherings but the social structure itself is maintained via a stable *core* of individuals. A simple matching across time exposes the long-term temporal behavior of each core, revealing intricate patterns of recurring meetings over time (Fig. 1C and supporting material section S2). The approach proposed here dramatically simplifies the temporal dynamics of the social system, renders community detection unnecessary, and reveals components that are directly interpretable. Our results are based on high-resolution multi-channel social interaction data from the Copenhagen Networks Study[17], which includes face-to-face interactions[18] of a large densely-connected

population of approximately $1\,000$ freshman students.

Dynamically evolving gatherings represent a completely new object of quantitative study. Fig. 1C illustrates the dynamics. Each node is part of only a single gathering per time step, but may switch affiliations between co-existing gatherings. It is due to this gradual turnover that community detection has proven difficult in many other settings. Individuals participating in multiple groups create a highly overlapping structure, which is difficult to untangle[9] (Fig. 1A-B). The gatherings we discover, are broadly distributed in both size and duration, capturing meetings ranging from small cliques to large aggregations, and from short interactions on the order of minutes to prolonged encounters lasting many hours, covering the multiple temporal scales of social lives. We find that small groups tend to have shorter meetings while large groups have a typical duration of 1-2 hours. Partitioning gatherings according to location, there are 42% on-campus (work) and 58% off-campus (recreation) gatherings. Comparing work/recreation campus statistics, recreational meetings tend to be smaller but last considerably longer, illustrating that the context of meetings can influence their properties (see SI sections S3 and S4 for full statistics). Unlike the typical community detection assumption of binary assignment to social contexts[19], we find that real world gatherings have soft boundaries (Fig. 1D), with some members participating for the total duration of the gathering, while others participate only briefly. We quantify this tendency in Fig. 1E where we investigate the stability of gatherings as a function of their duration. In terms of local stability (black line), which measures average turnover of nodes between subsequent network slices, we see that gatherings tend to be highly stable between time slices. When we compare each time slice to the aggregated network (global stability, gray line), we find that $\approx 70\%$ of all nodes are present in each slice. Both trends are largely independent of meeting duration. Comparing global to local stability, we see that high similarity between consecutive slices combined with a fixed global stability for any meeting duration reveals the stable core generating each gathering.

We now turn our attention to these stable cores and identify repeated appearances across the full duration of our dataset (section S4). The number of appearances per core is a heavy-tailed distribution; some cores appear only once, while the most active cores can appear multiple times per day over the full observation period (section S4.2). In the following, we focus on the temporal patterns of recurring gatherings, so we restrict our dataset to cores that, on average, are observed more than once per month. Fig. 2A shows a clear difference between how individuals engage and spend time with respect to varying social contexts (*work cores*—primarily observed on campus—and *recreational cores*—primarily observed elsewhere).

Cores leave traces in other data channels that emphasize the differences between work and recreation meetings. One such trace is coordination behavior, which we can explore by studying how call and text-message activity increases in the time leading up to a meeting. For each individual, we count the number of calls and texts within a hourly time-bin and compare to a null model based on typical hourly calling patterns for each participant. In this telecommunication network, we see clear evidence of coordination prior to meetings, which is accentuated during weekends when we expect meetings to be less schedule-driven (Fig. 2B). We also find that meetings re-

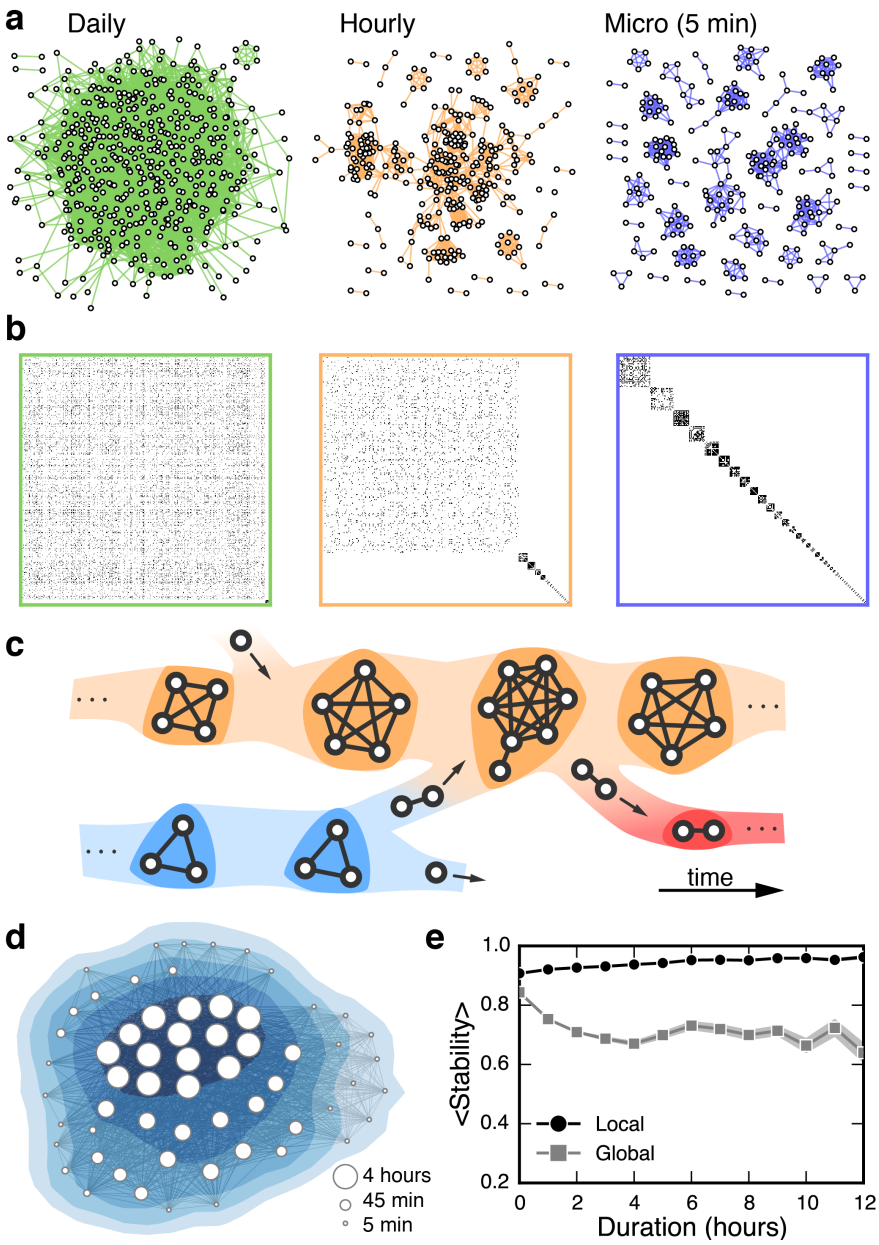

Figure 1: **Properties of gatherings.** **a**, The network formed by face-to-face meetings within one day (green), 60-minute (orange), and 5-minute temporal aggregation (blue). **b**, Corresponding adjacency matrices sorted according to connected components. Groups are directly observable for short time-slices, but become overlapping as more time is aggregated in each bin. **c**, Illustration of gathering dynamics. Gatherings change gradually with members flowing in and out of social contexts. **d**, Real world gatherings have soft boundaries, with nodes organized in a stable core and periphery nodes with varying degree of participation. Node-size corresponds to participation. **e**, The stability of gatherings as a function of duration. Global stability is defined as $\sum_{t_{\text{birth}}}^{t_{\text{death}}} J(g_t, G)/(t_{\text{death}} - t_{\text{birth}})$, where J denotes the Jaccard similarity and G is the aggregated network of slices ($G = g_{\text{birth}} \cup \cdots \cup g_{\text{death}}$) while local stability is defined as $\sum_{t_{\text{birth}}}^{t_{\text{death}}-1} J(g_t, g_{t+1})/(t_{\text{death}} - t_{\text{birth}} - 1)$.

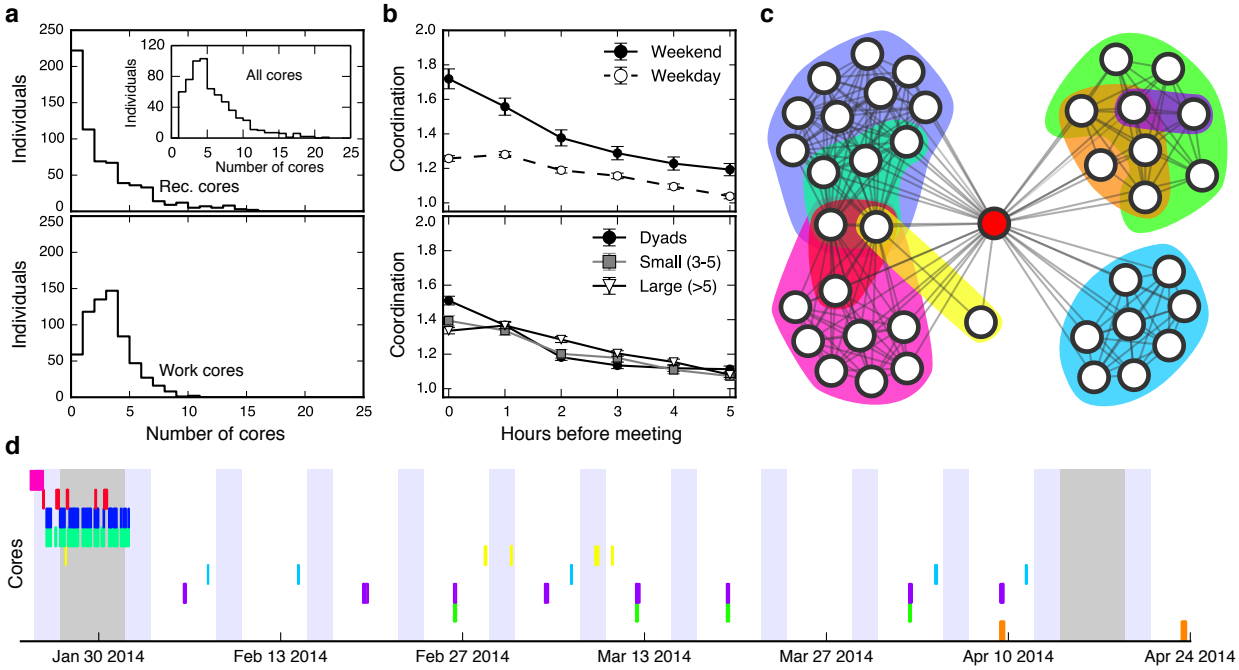

Figure 2: **Cores summarize social contexts for individuals.** **a**, The distributions of work and recreational core membership, inset shows participation across both categories. Participation in recreational cores reveals that individuals typically participate in only one or two recreational contexts, although the tail of the distribution show some individuals with more gregarious behavior. The distribution of work cores is localized, with an average of 2.74 ± 1.85 work cores per individual, mainly reflecting participation in classes or group work. **b**, Coordination prior to meetings, defined as $c_t = 1/N \sum_{n=1}^{N} a_t^n / \widetilde{a_t^n}$, where N is the number of participants, a_t^n is the individual activity of person n in time-bin t, compared to an individual baseline denoted by the average activity $\widetilde{a_t^n}$. More coordination is required to organize meetings during weekends than during weekdays. Larger meetings do not require additional coordination per participant. **c**, Ego view of communities; we observe overlapping and hierarchically stacked structures. **d**, The temporal complexity of participation for the cores in panel, we summarize this complexity using time-correlated entropy[13].

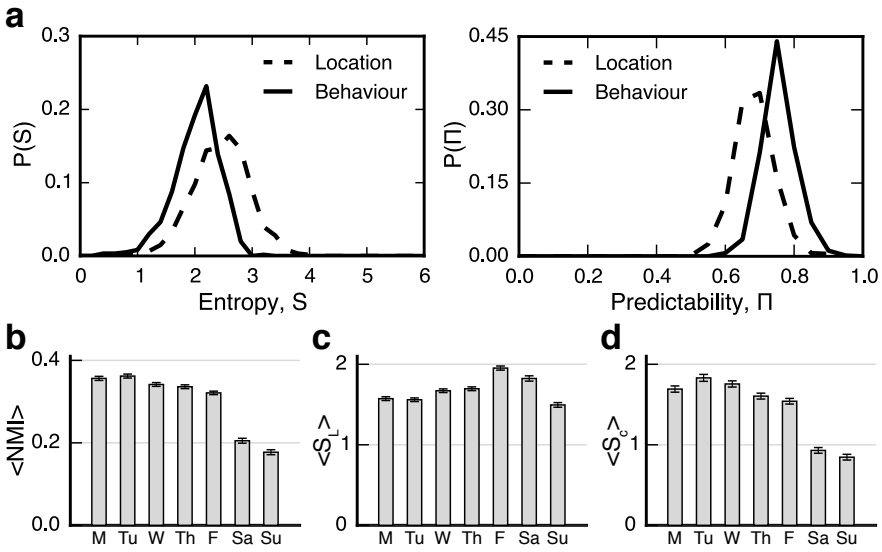

Figure 3: **Geospatial and social predictability. a**, The distributions of entropy and predictability for social and location patterns. We find that overall social patterns tend towards lower entropy than geospatial traces, resulting in higher predictability. The fact that our location-predictability is lower than previously found[13] is connected to a number of factors. For example, our location data is based on GPS rather than cell towers and has significantly higher precision[20] (see SI for full details). **b**, The average daily normalized mutual information between social and location sequences. Notice a significant drop on weekends. **c**, The average daily entropy of social engagements. The entropy is reduced on weekends, indicating a simpler pattern of social engagements, in agreement with Fig. 2a. **d**, The average daily entropy of location sequences, which increases on Fridays and Saturdays, indicates an increased geographical exploration on those days. In panels b-d we use the time-uncorrelated entropy to quantify the behavioral complexity.

quire the same amount of prior communication per person regardless of the size of the gathering (Fig. 2B). In terms of network structure, we find that cores are highly overlapping, and that large cores may contain rich inner structure with hierarchically nested sub-cores (section S4.3). Here, however, we focus on cores from the perspective of individuals; Fig. 2C shows an ego-perspective. In Fig. 2D, we can observe the temporal patterns of core participation from late January to late April for the ego-network shown in Fig. 2C. The participation patterns are complex, displaying regularity mixed with randomness.

Cores provide a powerful simplification of the complexity of dynamic networks. A core represents a social context, and for an individual, the full set of cores provides a vocabulary for quantifying social life. With access to detailed mobility data Song *et al.* [13] made the highly surprising discovery that human mobility patterns contain great potential for predicting future locations based on past behavior. Below we demonstrate that, when encoded through the core representation, our social interactions feature even higher levels of predictability than our mobility behavior. Given a sequence of social contexts, we use the time-correlated entropy[13] to construct a measure

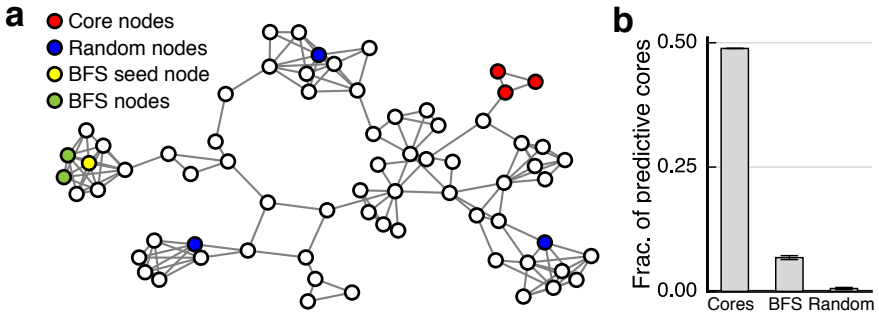

Figure 4: **Cores are predictive.** **a**, Illustration of null models. Starting from co-located nodes in the daily graph, we select nodes (1) chosen randomly and (2) found using breadth first search (BFS) in a daily graph. **b**, Social prediction using incomplete cores to predict arrival of remaining core-members. The strong increase compared to the BFS model emphasizes that full cores are needed for prediction, not just pairwise friendships. The ability to predict is tested on a month where cores have not been inferred. Error bars are calculated over $n = 100$ independent trials.

of the complexity for each person in our dataset. In order to incorporate the full complexity of social encounters in the calculation, we include cores with any number of appearances as well as dyadic cores. Time-correlated entropy quantifies the amount of uncertainty within a data sequence, accounting for both for frequency and ordering of states and simultaneously provides an upper bound of the predictability based on social routine. Fig. 3A shows the distribution of entropy and predictability. Social activity in our population is characterized by low temporal entropy, resulting in an average routine-based predictability limit of approximately 80%.

The core-representation allows us to examine existing results on predictability based on routine in location data[13] in the light of social patterns. Fig. 3A shows the distributions of entropy and predictability, calculated separately for the sequence of social states and spatial locations, respectively. Comparing social and location traces leads to a number of interesting findings. Firstly, we find that the social behavior tends to be more predictable and routine-driven than geospatial behavior. Secondly, the overall level of social and geospatial predictability is not correlated for individuals (p-value$= 0.85$ and section S5.3). Thus, highly routine driven location sequences do not imply predictable social behavior or *vice versa*. Thirdly, while the overall element of routine in a social or geospatial trajectory is not correlated for an individual, predictability in both contexts is closely related to daily and weekly schedules[21]. During the week, our social and location behavior is correlated; we tend to meet the same people in the same places. This correlation between our social and location behavior is reduced on weekends. In Fig. 3B, we use the average (uncorrelated) normalized mutual information between daily social and geolocation traces to illustrate this behavior. The mutual information is a measure of how much knowing one variable reduces uncertainty about the other. Interestingly, we find that location-entropy increases during weekends, indicating a more exploratory behavior (Fig 3C). During that same time-period, social traces become simpler and more predictable (Fig 3D), consistent with Fig. 3A and previous work[22]. Thus, in our population, periods of geospatial exploration are associated with social consolidation.

Until now, prediction has implied understanding future behavior based on past routine. We now demonstrate how the recreational cores effectively summarize information in the underlying network by proposing a completely new kind of prediction, based on the cohesion of the social fabric itself. It is a well established fact that there is a correlation of spatial patterns between individuals that share a social tie[23–26]. Pairs of traces, however, do not contain information that reveal at which times two location traces overlap, making it non-trivial to use this information for prediction. Cores provide such a temporal signature—an incomplete set of core members implies that the remaining members will arrive shortly. We illustrate this phenomenon on cores of size three. Given that two members of a core are observed, we measure the probability that the remaining member will arrive within one hour. To avoid testing on scheduled meetings, we only consider weekends and weekday evenings and nights (6pm-8am), where meetings are not driven by an academic schedule. Furthermore, we test on a month of data that has not been used for identifying cores. We now compare social prediction using cores to two null models (Fig. 4A). In the first, random, null model we create reference groups by randomly choosing groups of nodes from a daily graph. For the second null model, we form reference groups by performing a breadth first search (BFS) on the daily graph of interactions. While Fig. 4B shows that $\sim 50\%$ of cores are predictive, both the random and the BFS reference groups fare poorly. By requiring that nodes share social connections, as well as a spatial location, the BFS null model demonstrates that that it is not simply pairwise friendships that are predictive. The predictions of the arrival of the final group member is possible because the social context requires all core members to be present. We have no reason to expect this social behavior to be exclusive to our population of students, but expect that social prediction based on cores can be applied broadly.

Within the existing literature, incorporating a temporal dimension dramatically complicates the mathematical description of complex networks[14]. Starting from fundamental structures, we find the opposite. By observing social gatherings at the right time scale, when the temporal granularity is higher than the turnover rate, a simple matching across time slices reveals dynamically changing gatherings with stable cores that can be matched across time, providing a strong simplification of the social dynamics. These cores manifest in other data channels, such as through coordination behavior, and provide a finite vocabulary, which dramatically simplifies individuals' social activity. As a demonstration of the saliency of the description, we use the cores to (1) quantify predictability within the social realm and (2) allow for a new kind of non-routine prediction, based solely on the signal encoded in the core representation. Our work provides a first quantitative look at the rich patterns encoded in the micro-dynamics of a large system of closely interacting individuals, characterized by a high degree of order and predictability. The work presented here provides a new framework for describing human behavior and hints at the promise of our approach. We have focused on predictability, but we expect our work will support better modeling of processes in social systems, from epidemiology and social contagion to urban planning, as well quantitative sociology, and public health.

Methods

Inferring gatherings In each temporal slice we identify connected components as groups. A gathering is defined by matching these groups across time. Matching is done using single-linkage agglomerative hierarchical clustering that iteratively merges the two clusters with smallest pairwise distance. This merge criterion is strictly local and will agglomerate groups into chains, a preferable effect when considering time-series. Distance is calculated using $d(c_t, c_{t'}) = 1 - |c_t \cap c_{t'}|/|c_t \cup c_{t'}| f(\Delta t, \gamma))$, where $f(\Delta t, \gamma)$ denotes coupling between temporal slices (may assume any functional form) and $\Delta t = t' - t$ denotes the distance between two bins. We model the decay using an exponential form $f = \exp(-\gamma(\Delta t - 1)))$. By definition $f = 1$ (zero decay) between two consecutive temporal slices. The optimal threshold for partitioning the dendrogram is discussed in the Supplementary Information.

Dynamical communities Gatherings only contain information about their local appearance, to gain a dynamic picture we match them across time. Counting the fraction of times each nodes has been present in a gathering we construct a participation profile and match gatherings according to them. Average linkage agglomerative hierarchical clustering is applied and distance is calculated as:

$$D\left(G_i, G_j\right) = 1 - \sum_{n=1}^{N} \min\left(G_i, G_j\right) / \sum_{n=1}^{N} \max\left(G_i, G_j\right). \tag{1}$$

G_i is a vector that denotes participation values for nodes belonging to gathering i, N is the total number of nodes for $G_i \cup G_j$, and the two functions $\max$ and $\min$ act piecewise on the vectors. Further, $D(G_i, G_j)$ is defined as 1 between two gatherings that have zero overlap. Iteratively this method builds a dendrogram with gatherings as leafs and thresholding the tree clusters similar gatherings. The optimal threshold is set using the gap measure (see SI).

Cores Ranking nodes in a community according to their participation, we extract a core when there is a significant gap in the participation levels. This is done by comparing the participation profiles to a reference model, where participation is drawn from a representative null distribution. If the maximal gap in the real distribution exceeds the average plus variation of N reference distribution then we denote the gap as significant, i.e. if $\max_{gap}^{\text{real}} > \mu(\max_{gap}^{\text{reference}}) + \sigma(\max_{gap}^{\text{reference}})$. Nodes with participation levels above the gap are included in the core.

Definitions of entropy and predictability For an individual i given a sequence of states we define entropy, or uncertainty, in two ways, (1) uncorrelated entropy $S_i^{\text{unc}} = -\sum_j^{N_i} p_j \log_2 p_j$, where p_j is the probability of observing state j, captures the uncertainty of the behavioural history without taking the order of visits into account; (2) temporal (or time-correlated) entropy $S_i^{\text{temp}} = -\sum_{T_i' \subset T_i} p(T_i') \log_2[p(T_i')]$, where $p(T_i')$ is the probability of finding a subsequence T_i' in the trajectory T_i, takes both frequency and order of states into account. From the entropy one can estimate the upper bound of predictability by applying a limiting case of Fano's inequality[13,27]: $S_i = H(\Pi_i) + (1 - \Pi_i)\log_2(N - 1)$, where $H(\Pi_i) = -\Pi_i \log_2(\Pi_i) - (1 - \Pi_i)\log_2(1 - \Pi_i)$ and N is the number of states observed by person i.

1. Easley, D. & Kleinberg, J. *Networks, crowds, and markets: Reasoning about a highly connected world* (Cambridge University Press, 2010).

2. Newman, M. *Networks: An Introduction* (Oxford University Press, 2010).

3. Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications* (Cambridge university press, 1994).

4. Simmel, G. Quantitative aspects of the group. *The Sociology of Georg Simmel* 87–177 (1950).

5. Goffman, E. *Interaction ritual: Essays in Face to Face Behavior* (AldineTransaction, 2005).

6. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).

7. Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).

8. Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).

9. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).

10. Lazer, D. *et al.* Computational social science. *Science* **323**, 721–723 (2009).

11. Pentland, A. S. *Honest signals* (MIT press, 2010).

12. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).

13. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).

14. Holme, P. & Saramäki, J. Temporal networks. *Physics Reports* **519**, 97–125 (2012).

15. Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. & Onnela, J.-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328**, 876–878 (2010).

16. Gauvin, L., Panisson, A. & Cattuto, C. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PLoS ONE* **9**, e86028 (2014).

17. Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PLoS ONE* **9**, e95978 (2014).

18. Sekara, V. & Lehmann, S. The strength of friendship ties in proximity sensor data. *PLoS One* **9**, e100915 (2014).

19. Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

20. Lin, M., Hsu, W.-J. & Lee, Z. Q. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 381–390 (ACM, 2012).

21. McInerney, J., Stein, S., Rogers, A. & Jennings, N. R. Exploring periods of low predictability in daily life mobility. In *Nokia Mobile Data Challenge Workshop* (2012).

22. Eagle, N. & Pentland, A. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* **10**, 255–268 (2006).

23. Crandall, D. J. *et al.* Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences* **107**, 22436–22441 (2010).

24. Wang, D., Pedreschi, D., Song, C., Giannotti, F. & Barabási, A.-L. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1100–1108 (ACM, 2011).

25. Cho, E., Myers, S. A. & Leskovec, J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1082–1090 (ACM, 2011).

26. De Domenico, M., Lima, A. & Musolesi, M. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* **9**, 798–807 (2013).

27. Fano, R. M. *Transmission of information: a statistical theory of communications* (MIT Press, 1961).

Acknowledgements We thank L. K. Hansen, P. Sapiezynski, A. Cuttone, D. Wind, J. E. Larsen, B. S. Jensen, D. D. Lassen, M. A. Pedersen, A. Blok, T. B. Jørgensen, and Y. Y. Ahn for invaluable discussions and comments on the manuscript and R. Gatej for technical assistance. This work was supported a Young Investigator Grant from the Villum Foundation (High Resolution Networks, awarded to S.L.), and interdisciplinary UCPH 2016 grant (Social Fabric). Due to privacy implications we cannot share data but researchers are welcome to visit and work under our supervision.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to S.L. (email: sune.lehmann@gmail.com).