How Random are Online Social Interactions?

Chunyan Wang[1] & Bernardo A. Huberman[2]

[1]Department of Applied Physics, Stanford University, California, USA, [2]Social Computing Group, HP Labs, California, USA.

The massive amounts of data that social media generates has facilitated the study of online human behavior on a scale unimaginable a few years ago. At the same time, the much discussed apparent randomness with which people interact online makes it appear as if these studies cannot reveal predictive social behaviors that could be used for developing better platforms and services. We use two large social databases to measure the mutual information entropy that both individual and group actions generate as they evolve over time. We show that user's interaction sequences have strong deterministic components, in contrast with existing assumptions and models. In addition, we show that individual interactions are more predictable when users act on their own rather than when attending group activities.

Recent developments in digital technology have made possible the collection and analysis of massive amount of human social data and the ensuing discovery of a number of strong online behavioral patterns[1–11]. These patterns are important for two reasons. First, they yield predictions about future behavior that can be incorporated into the design of useful social media and services, and second, they provide an empirical test of the many social theoretical models that have been proposed in the literature. As an example, the assumption that events in web traffic data are described by a series of Poisson process[12] was shown to be contradicted by measurements of the the waiting time between two consecutive events, which display power law scaling. These power laws are ubiquitous and appear in the analysis of email exchanges[13–15] and web browsing[16–18]. On the other hand, regular behavioral patterns in real life are a well known phenomenon, as exemplified by vehicular traffic patterns, daily routines, work schedules and the seasonality of economic transactions. At the aggregate level, these regularities are often induced by spatial and temporal constraints, such as the disposition of roads and streets in urban settings or the timing of daily routines. Other examples are provided by the existence of deterministic patterns in human daily communication[13,19] and phone call location sequences[20].

When it comes to human online activities many theoretical studies curiously assume uncorrelated random events on the part of the users[12,21–23] which makes their behavior rather unpredictable. Moreover, that literature assumes that a user's future partners in comments and reviews, or how web pages are visited are independent of the history of the process or at best on the previous time step. While these assumptions work well for page ranking in web searching[21], online recommendation systems[22], link prediction[24], and advertising[23], it is not clear that they apply to more interactive processes such as contacting friends within online social networks, participating in online discourse and exchanges of email and text messages. Even in cases where a Markovian assumption seems to yield good results, the discovery of deterministic components to online browsing and searching can improve existing algorithms[25].

In this paper we study the predictability of online interactions both at the group and individual levels. To this end, we measure the predictability of online user behavior by using information-theoretic methods applied to real time data of online user activities. This is in the same spirit as a recent study of offline conversations within an organization[19]. Using ideas first articulated in studies of gene expressions[26], predictability is here defined as the degree to which one can forecast a user's interacions based on observations of his previous activity. The main focus of this study is to be contrasted to existing studies of online social behavior, such as recommender systems[22] and link prediction[24], which use statistical learning models to improve the prediction accuracy of novel links and recommendations. By examining datasets from user commenting activities and place visiting logs, we found that the observed activity sequences deviate from a random walk model with deterministic components. Furthermore, we also compared the predictability of activities when individuals act alone as opposed to as members of a group. In contrast to many model assumptions in studies of online communites and group behavior[27–29], we observed that individuals are less predictable when attending group or social activities than when acting on their own.

Results

We examined the predictability of online user behavior using data-sets from two different websites: Epinions and Whrrl. Epinions is a who-trust-who consumer review site, where users write their personal reviews of a wide variety of products, ranging from automobiles to media (including music, books and movies). Each user can comment on other users' reviews or comments. The thread of comments forms a conversation of two or more users. To trace the predictability of commenting partners, we collected 88,859 unique users' comments from the website. For each user, we used the website's API to collect all of their comments with a time stamp for each comment. In total, we gathered 286,317 threads of comments from different categories containing 722,475 user comments. The other dataset that we used is from Whrrl.com. Whrrl is a popular LBSN (Location Based Social Network) that people use to explore, rate and share points-of-interest. It also allows users to declare friendships with each other and to interact through visits and check-ins at physical places. Users can check in by using a mobile application on a GPS-equipped smart phone. The types of places that are often visited include restaurants, hotels and bookstores. A distinctive feature of this dataset is that a user can check-in by herself or with a group of other people, thus providing a forum for social activities. Users of the site are identified by unique user-ids. In our study, we crawled a friendship network consisting of 24,002 users and 145,228 social ties and collected the check-in records of these users' activities from January 2009 to January 2011. The resulting undirected graph had an average degree of 12.101 and an average shortest-path length of 4.718, which is typical of a small-world social network. In our observational period of 2 years, there were 357,393 check-in records over 120,726 different places associated with these users. For each check-in record, we also collected information such as the exact location (i.e., longitude and latitude), time of check-in and the users involved (i.e., there may be more than one user-id for group check-ins). We were thus able to obtain a series of places the users visited in chronological order.

The activity sequence is obtained by neglecting the absolute timing of events in the raw dataset. To generate the activity sequence of a certain user, we first sifted out all the events that are associated with the user and we then listed the chronologically ordered sequence of states identified by a unique number. For the Whrrl dataset, we labeled each activity as a group one if the user was checking in with others. To determine the extent to which user behavior is predictable we used standard information-theoretic methods similar to those used in the analysis of gene expression[26,30]. For instance, we consider a user A as having M_A possible states, where each state in the sequence can correspond to either an online conversation partner or a check-in location. An example of a user's activity sequence is shown in Figure 1, where two states, 1 and 2, form the sequence. We then used the observed sequenc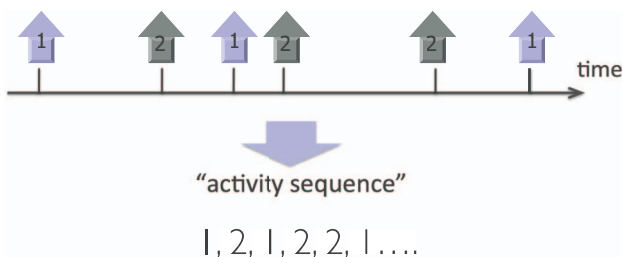es to examine the degree of second order dependences, which signal the extend to which activities depart from random interactions.

We used entropy to measure the randomness of a user A's activities. The estimated probabilities for all states $p_A(i)$ have the property that $\sum_{i=1}^{M_A} p_A(i) = 1$. If we assume that these probabilities do not change with time, the randomness of user A's possible states can be measured by the uncorrelated entropy, defined as

$$H_A^1 = -\sum_{i=1}^{M_A} p_A(i) \log p_A(i). \tag{1}$$

Notice that if each state is equally probable, this uncorrelated entropy is maximal and equal to

$$H_A{}^0 = \log M_A. \tag{2}$$

To measure the randomness of the sequence from knowledge of the previous states we introduce the conditional entropy, defined as

$$H_A^2(i|j) = -\sum_{j=1}^{M_A} p_A(j) \sum_{j=1}^{M_A} p_A(i|j) \log p_A(i|j). \tag{3}$$

And we quantify the predictability of the user's activity sequence by using the mutual information

$$I_A = H_A^1(i) - H_A^2(i|j). \tag{4}$$

For each user, the inequalities $0 \leq H_A^2 \leq H_A^1 \leq H_A^0$ are satisfied. I_A is equal to the amount of information one can gain about the next state by knowing the current state. If there is no second order correlation between state sequences, H_A^1 is equal to H_A^2, and I_A takes the minimum value of 0. If the next state is completely determined by the previous state, or in other words the user activity is completely predictable, I_A takes the maximum value of H_A^1.

The calculations of these quantities require an accurate estimation of the probabilities $P_A(j)$ and $P(j|i)$. However, in the absence of unlimited data, estimating these probabilities with finite sampling renders a biased estimation of the entropy, since the finite sampling makes the user activity less variable than it is, resulting in a downward bias of the entropy, and a upward bias of the the mutual information[30]. The problems associated with estimating entropies for sparse data have been extensively explored in the literature and a variety of remedies proposed[31]. The most common solution is to restrict the measurements to situations where one has an adequate amount of user activity data[32]. In what follows we filter out users who are below a certain activity level, 1000 in our observational period. Since both H^1 and H^2 generally decrease by different amounts when taking into account finite size effects, we also performed a through bootstrap test to confirm that the empirical values of mutual information are significantly different from zero. Another widely accepted method is to estimate the magnitude of the systematic bias that originates from finite size effects and then subtract this bias from the estimated entropy. To do so, we used the Panzeri-Treves bias correction method[31] in our calculations. The lead terms in the bias are, respectively

$$BIAS[H_A(i)] = -\frac{1}{2N\ln(2)}\left[\overline{M} - 1\right], \tag{5}$$

and

$$BIAS[H_A(i|j)] = -\frac{1}{2N\ln(2)}\sum_j\left[\overline{M_j} - 1\right], \tag{6}$$

where $\overline{M}$ denotes the estimated number of outcome states, $\overline{M_j}$ denotes the number of different states i with nonzero probability of being observed given that the previous state is j, and N is the total number of observations. Thus, the leading term of the mutual information bias equals
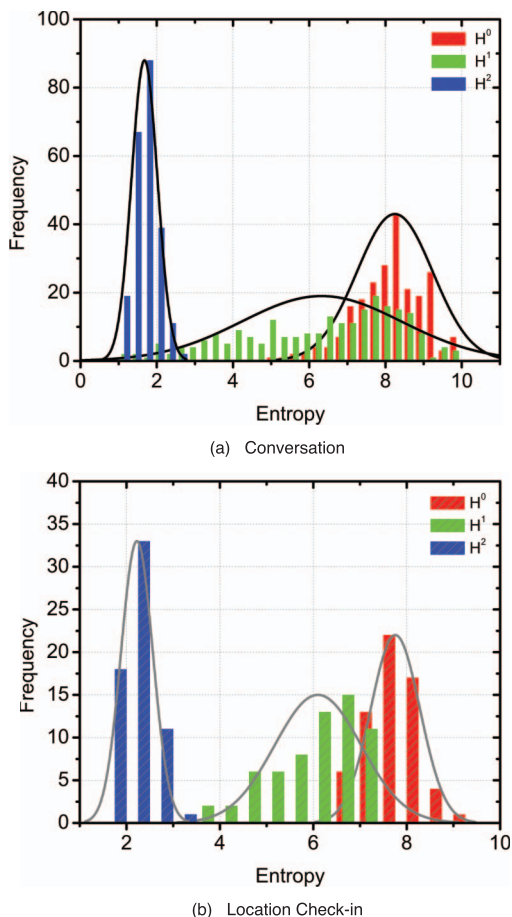

Figure 1 | Online activity sequence of a sample user. Every short vertical line in the figure represents the time of a user activity. There are two observed states for the sampled user's activity sequence, state 1 and state 2. The second order correlation, or predictability, of this sequence is measured through the conditional entropy.

(a) Conversation

(b) Location Check-in

Figure 2 | Frequency count of estimated H^0, H^1 and H^2 from users in (a) online conversation partner sequence and (b) online location check-in place sequence.

$$BIAS[I(i;j)] = \frac{1}{2N\ln(2)}\left\{\sum_j \left[\overline{M_j} - 1\right] - \left[\overline{M} - 1\right]\right\}. \quad (7)$$

In what follows, we include the above adjustments to eliminate the impact of the finite size amount of data.

We start by looking at the predictability of individual activities as measured by both the entropy and the mutual information extracted from sequences in the Whrrl and Epinion datasets, respectively. The histograms of H^0, H^1 and H^2 calculated from user activity sequences

are shown in Figure 2. The gray solid line in the plot shows a normal fit to the frequency count. The gap between H^1 and H^0 suggests a preference for certain activities, while the difference between the values of H^1 and H^2 in the figure indicates the existence of second order correlations between states. Values of H^1 and H^2 for each individual in the online conversation network and the location check-in one are shown in Figure 3. The straight line corresponds to H^1 equal to H^2. One interesting fact is that all the dots are below the straight line, which confirms that there is a positive difference between H^1 and H^2 for all individuals. This difference, which is the mutual information conditioned on previous states of user activity, is plotted in increasing order as the red line in Figure 4 for (a) conversations and (b) location check-ins. The positive values of the mutual information indicate information gain, or predictability, conditioned on historical states.

We now examine the validity of the positive mutual information values in greater detail. There are usually two limitations when performing mutual information measurements. The first one is the potential bias resulting from the finite data size. The second one is the possibility of missing data points in the observation process. To make sure that our results are significant and are not impacted by these two limitations, we performed the following analysis. To establish that the observed positive value of the mutual information is not due to the finite size of our data sets, we performed a bootstrap test similar to that used in human conversation studies[19]. The null hypothesis of this test is that the mutual information has a positive value because of the finite size of the dataset. For this test we set the significance level to 5%. We first shuffled the true activity sequence and constructed a new sequence by drawing elements randomly one by one from the original sequence without replacement. If there is a second order correlation in the original sequence the shuffled sequence breaks the order and will thus have a higher H_A^2 value, while the value of H_A^1 would be the same before and after the bootstrap. This would result in a mutual information I_A value smaller than that of the true sequence. The test checks if the value of I_A obtained from the original activity sequence is significantly different from the shuffled one. To obtain an estimate of the distribution for the shuffled sequence we performed the shuffling procedure a 1000 times for each user and calculated each individual's shuffled mutual information. The value of the simulated sequence ranging from 2.5% to 97.5% is shown by the blue column in Figure 4. As can be seen, the red line (mutual information for true activity sequence) lies well above the upper end of the 97.5% error bar, which suggests that the value of the original sequence is significantly different from that generated by the simulated sequences. We can then reject the null hypothesis at the 5% significance level and conclude that the positive mutual information we obtained is not due to the limited size of the
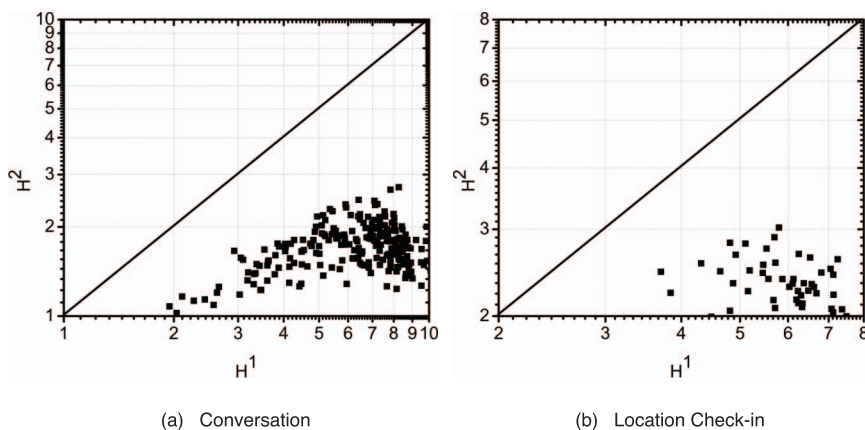

(a) Conversation

(b) Location Check-in

Figure 3 | Relationship between the measured H^1 and H^2 in (a) online conversations and (b) location check-ins. The solid line in the plot represents the line where H^1 and H^2 are equal. Black dots in the plot correspond to individual activity sequences.
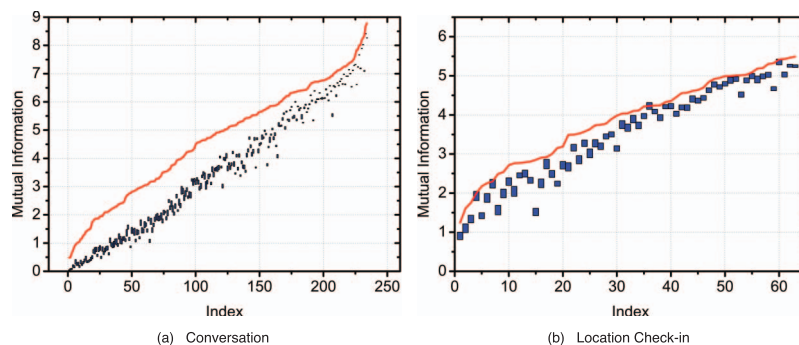
(a)   Conversation

(b)   Location Check-in

Figure 4 | Estimated mutual information and statistics of bootstrap samples. The red line is the mutual information estimated from observed online activity sequences. The upper and lower end of the blue columns represent the 2.5% and 97.5% percentile of 1000 shuffled sequences for (a) online conversation partner sequences and (b) online location check-in place sequences.

data. Furthermore, the fact that the mutual information is significantly different from zero suggests that a user's current online activity predicts his next interactions. Next, we assessed the impact of potential loss of data points in the observation period by marking off a percentage of data points from the observed location check-in sequence from Whrrl dataset. In real applications of predicting user behavior, a key question to apply maximum likelihood estimation depends on the size of observations and the ratio of missing points. To examine the impact of ratio, we perform a mark-off on the bootstrap test of mutual information. We hide data points randomly from the true sequence while keeping the chronological order in the remaining sequence. For example if we have a mark-off rate of 0.5, then 50% of the states from the true sequence is marked off. The result of the bootstrap test after mark-off is shown in Figure 5. In the plot, the red dot shows the mutual information of the true sequence after performing mark-off procedure. The thick blue bar in the plot demonstrates the mutual information of the exact same sequence with shuffling. The mutual information is significantly different from that of the random shuffled sequence until the mark-off rate reaches 60%. For values of the mark-off rate larger than 60%, the difference between the two is broken when we fail to reject the null hypothesis that the sequence is significantly different from randomly shuffled. It is thus a confirmation that the deterministic pattern we observed is a robust one. This test also suggests the existence of a higher order correlations, larger than two, in human social online behavior. Thus, the deterministic pattern discussed in this study is a robust
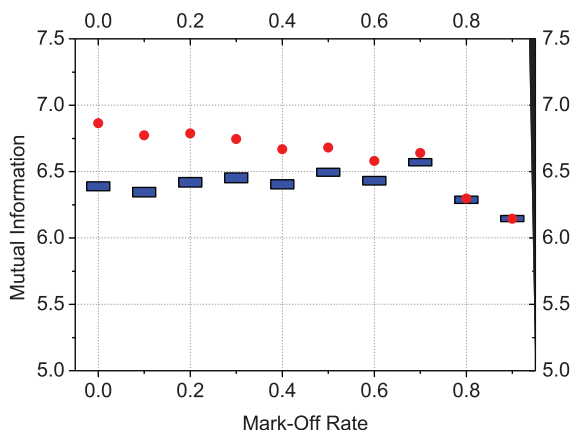
phenomenon which can be applied to the general situations with missing or incomplete observations.

As mentioned earlier, we also explored whether individuals acting alone are less predictable than when becoming members of a group. Specifically, we investigated how predictable each user's is when engaged in group activities as compared with the predictability of individual ones. In the Whrrl dataset users can expose their position with a group of other users thus providing a sequence of group attendances by users and filtering out the places that were checked in by the user alone. We then calculated the information entropies and performed the same bootstrap test as before. The calculated mutual information of the activity sequences and shuffled sequences are shown in Figure 6. Interestingly, the gap between the red line of true observation and the upper end of the error bar is is smaller than the one we obtained for the individual activities. In contrast with Figure 4(b), the differences between the randomly shuffled sequences and the true observations are smaller. To quantify the observed difference, we calculated the gap between the mutual information from the true activity sequence and the 97.5% percentile value of the shuffled sequences, defined by $G_{Individual} = I_{Individual} - I_{Individual}^{0.975}$ and $G_{Group} = I_{Group} - I_{Group}^{0.975}$. This allows for a comparison of sequences with different lengths. The relative frequency plot of this $G_{Individual}$ and G_{Group} is plotted in Figure 7. The upper plot in Figure 7 shows the density plot of the gap for individual activity sequences while the lower plot shows the gap for group activities. As can be seen, the gap for individual activities has a larger of the mode compared with that of the group activities. Under the assumption that both populations from $G_{Individual}$ and G_{Group} are random,

Figure 5 | Mutual information and statistics of bootstrap as a function of mark-off rate. Red dots in the plot shows mutual information of sequence after mark-off. Blue bar in the plot shows the mutual information of that marked-off sequence with random shuffling. Up to 60% of hidden data from the true sequence, there is deterministic pattern in the sequence after mark-off.
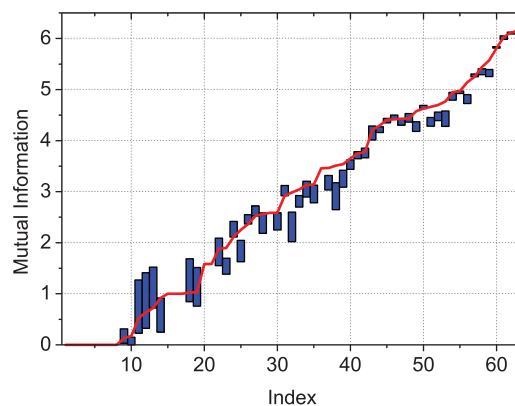

Figure 6 | Estimated mutual information and statistics of bootstrap samples for group activities from Whrrl dataset. The red line is the mutual information estimated from observed online activity sequences. The upper and lower end of the blue column represent the 2.5% and 97.5% percentile of the shuffled sequences.
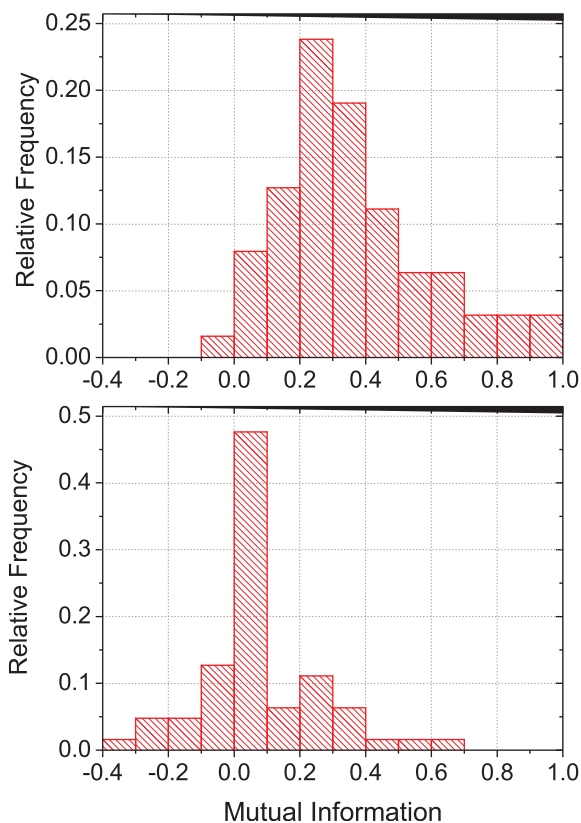
Figure 7 | The upper plot shows the density of $G_{Individual}$. The lower plot shows density of G_{Group}. The gap for individual activities has a larger mean compared with that of attending group activities.

independent, and arising from a normally distributed population with equal variances, the two sample t-test rejects the null hypothesis of an equal mean with a p-value of 4.88×10^{-12} under 5% significance level. This implies that it is harder to predict the a user's group activities than his individual ones. The values of $G_{Individual}$ versus G_{Group} for each individual are plotted in Figure 7. The mean of $G_{Individual}$ is larger than G_{Group}. One possible explanation for this observation is that when individuals attend group activities, the decision as to what to do next is not usually made by the individual himself. Thus, the tendency to follow others in their decisions tends to break one's regular patterns. This extra randomness would result in a larger value of H_A^2 and thus become less predictable.

Discussion

In summary, we have shown that sequences of user online activities have deterministic components that can be used for predicting future activities. Using methods from information theory, we experimentally measured how much additional information can be gained from knowledge of previous states within a users' activity sequences. While the degree of predictability varies from person to person, we also established that it is different when individuals join a group. Besides the intrinsic interest of these findings, the fact that one can predict online social interactions should be helpful in improving the design of algorithms and applications for online social sites.

1. Huberman, B. A. *The Laws of the Web: Patterns in the Ecology of Information* (The MIT Press, 2001).
2. Vázquez, A., Gama Oliveira, J., Dezsó, Z., Goh, K.-I., Kondor, I. & Barabási, A.-L. Modeling bursts and heavy tails in human dynamics. *Physical Review E* **73**(3), 036127 (2006).
3. Golder, S. A., Wilkinson, D. M. & Huberman, B. A. Rhythms of social interaction: messaging within a massive online network. In *International Conference on Communities and Technologies*, (2007).
4. Wu, F. & Huberman, B. A. Novelty and collective attention. *Proc. Natl. Acad.* Sci. **105** 17599 (2007).
5. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* (London) **453**, 779 (2008).
6. Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F. & Vespignani, A. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.* **105**(15), 158701 (2010).
7. Wu, Y., Zhou, C., Xiao, J., Kurths, J. & Schellnhuber, H. J. Evidence for a bimodal distribution in human communication. *Proc. Natl. Acad. Sci.* **107**, 18803 (2010).
8. Golder, S. A. & Macy, M. W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* **333** 6051 (2011).
9. Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. Strong regularities in world wide web surfing. *Science* **280** 95 (1998).
10. Tyler, J. R. & Tang, J. C. When can i expect an email response? a study of rhythms in email usage. In *ECSCW*, 239–258. Springer, (2003).
11. Gonalves, B., Perra, N. & Vespignani, A. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS ONE* **6** 8 (2011).
12. Scott, S. L. & Smyth, P. The markov modulated poisson process and markov poisson cascade with applications to web traffic data. *Bayesian Statistics* **7** (2003).
13. Eckmann, J., Moses, E. & Sergi, D. Entropy of dialogues creates coherent structures in e-mail traffic. *Proc. Natl. Acad. Sci.* **101** 14333 (2004).
14. Barabasi, A. The origin of bursts and heavy tails in human dynamics. *Nature* **435** (2005).
15. Rybski, D., Buldyrev, S., Havlin, S., Liljeros, F. & Makse, H. Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci.* **106** 12640 (2009).
16. Crane, R. & Sornette, D. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.* **105** 15649 (2008).
17. Adar, E., Teevan, J. & Dumais, S. Resonance on the web: Web dynamics and revisitation patterns. In *CHI*, (2009).
18. Chmiel, A., Kowalska, K. & Holyst, J. A. Scaling of human behavior during portal browsing. *Phys. Rev. E* **80** 066122 (2010).
19. Takaguchi, T., Nakamura, M., Sato, N., Yano, K. & Masuda, N. Predictability of conversation partners. *Phys. Rev. X* **1** 011008 (2011).
20. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**(5968), 1018–1021 (2010).
21. Page, L., Brin, S., Motwani, R. & Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, (1998).
22. Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D. & Aly, M. Video suggestion and discovery for youtube: taking random walks through the view graph. In *WWW*, 895–904. ACM, (2008).
23. Archak, N., Mirrokni, V. S. & Muthukrishnan, S. Mining advertiser-specific user behavior using adfactors. In *WWW*, 31–40. ACM, (2010).
24. Lu, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* **390**(6), 1150–1170 (2011).
25. Chierichetti, F., Kumar, R., Raghavan, P. & Sarlós, T. Are web users really markovian? In *WWW*, 609–618. ACM, (2012).
26. Steuer, R. E., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. In *ECCB*, 231–240 (2002).
27. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**(5827), 1036–1039 (2007).
28. Palla, G., Barabasi, A. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).
29. Choudhury, M. D. Modeling and predicting group activity over time in on-line social media. In *Hypertext* 349–350 (2009).
30. Panzeri, S., Senatore, R., Montemurro, M. A. & Petersen, R. S. Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol* **3**(98), 1064–1072 (2007).
31. Panzeri, S. & Treves, A. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems* **7**, 87107 (1996).
32. Steeg, G. V. & Galstyan, A. Information transfer in social media. In *WWW*, 509–518. ACM, (2012).

Author contributions

C.W. and B.H. conceived and designed the study. C.W. collected the data. C.W. and B.H. discussed the results and wrote the manuscript.

Additional information

How to cite this article: Wang, C. & Huberman, B.A. How Random are Online Social Interactions? *Sci. Rep.* **2**, 633; DOI:10.1038/srep00633 (2012).