JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Šuvakov M, Mitrović M, Gligorijević V, Tadić B. 2012 How the online social networks are used: dialogues-based structure of MySpace. J R Soc Interface 10: 20120819.
http://dx.doi.org/10.1098/rsif.2012.0819

Received: 6 October 2012
Accepted: 6 November 2012

Subject Areas:
mathematical physics, medical physics

Keywords:
online social networks, user communities, patterns of emotion flow, self-organized dynamics

Author for correspondence:
Bosiljka Tadić
e-mail: bosiljka.tadic@ijs.si

How the online social networks are used: dialogues-based structure of MySpace

Milovan Šuvakov[1,2], Marija Mitrović[1], Vladimir Gligorijević[1] and Bosiljka Tadić[1]

[1]Department of Theoretical Physics, Jožef Stefan Institute, Box 3000, 1001 Ljubljana, Slovenia
[2]Institute of Physics, Belgrade University, Belgrade, Serbia

Quantitative study of collective dynamics in online social networks is a new challenge based on the abundance of empirical data. Conclusions, however, may depend on factors such as user's psychology profiles and their reasons to use the online contacts. In this study, we have compiled and analysed two datasets from MySpace. The data contain networked dialogues occurring within a specified time depth, high temporal resolution and texts of messages, in which the emotion valence is assessed by using the SentiStrength classifier. Performing a comprehensive analysis, we obtain three groups of results: dynamic topology of the dialogues-based networks have a characteristic structure with Zipf's distribution of communities, low link reciprocity and disassortative correlations. Overlaps supporting 'weak-ties' hypothesis are found to follow the laws recently conjectured for online games. Long-range temporal correlations and persistent fluctuations occur in the time series of messages carrying positive (negative) emotion; patterns of user communications have dominant positive emotion (attractiveness) and strong impact of circadian cycles and interactivity times longer than 1 day. Taken together, these results give a new insight into the functioning of online social networks and unveil the importance of the amount of information and emotion that is communicated along the social links. All data used in this study are fully anonymized.

1. Introduction

The Web as a new social space provides 'unbearable easiness of communication' that may lead to new social phenomena in the online world and affect the behaviour of users [1–3]. The emergent online social networks represent some of the largest social structures in the world [4].

Recently, data regarding user activity at various Web portals have been analysed across different science disciplines [5–12]. It has been recognized that the emotions, known to drive social behaviour in face-to-face contacts, also play an important role in online interactions [12–19]. However, full understanding of the mechanisms that drive online social behaviours still remains elusive. It is of primary importance to understand how individual (emotional) actions of users in the network may lead to dynamically robust collective behaviours. What are the appropriate quantitative measures of the emergent phenomena? Furthermore, in what respect do the dynamics of online social networks differ from those of blogs, forums, games and other forms of online communications? Here, we address these questions by compiling and analysing the empirical data of *user dialogues* from the social network MySpace.

In recent psychology research, quantitative study of emotion [20–22] is conducted based on Russell's multidimensional model of affect [23]. Specifically, each emotion known in common life can be represented by a set of numerical values in the corresponding multidimensional space [23]. Three important dimensions of emotion are recognized as valence, arousal and dominance. Valence measures the degree of attractiveness (positive valence) or aversiveness (negative valence) to a stimulus. Similarly, arousal and dominance can have a range of values corresponding to different degrees of reactivity to stimuli, and the power of a reaction, respectively. Moreover, normative emotional rating has been developed for a large number of words [24]. Recently, on the basis of the lexicon of emotional words and machine-learning approaches, methods have been developed [16,22,25] for the effective inference of the emotional content from text messages appearing in online communications.

How are online social networks used and who uses them? Social psychology research begins to recognize the relationship between the personal profile, social and emotional loneliness and attitudes of the Web users in general, and users of social networks, in particular [9,26–28]. The 'friendship' association is the framework for communications via online social networks. Analysis of the topology of the *friendship network* in Facebook finds 'mostly, but not entirely, agreement on common structural network characteristics' [4, p. 1,29]. A somewhat different structure was reported for the *friendship network* of MySpace [30]. However, it is not the mere structure of the network of declared 'friendship' links, but rather the *dynamics of message exchange* along these links that contain relevant information for the study of emergent social phenomena, in particular temporal bursts of (emotional) messages involving many users. Therefore, developing a methodology for systematic collection and analysis of the data that contain complete information about the dynamics of dialogues is of key importance for the diagnostics of bursting events, for detecting and characterizing collective behaviours, and identifying the users involved.

In this work, we study the dialogues-based social networks that represent dynamical structures situated on the underlying friendship network in MySpace. They involve only a certain number (or type?) of active users, and their structure can vary in time, depending on the events and time window of interest. We consider two sets of data with dialogues among the users in MySpace social network, and analyse them as complex dynamical systems in order to determine quantitative measures of the collective behaviour of users. We first develop a procedure to compile the data that have the required structure and high temporal resolution for this type of analysis. The datasets are then studied by methods of graph theory and statistical physics to determine the topology of the dialogues-based networks, and to define and compute several other quantitative measures of the collective behaviours, in particular, the temporal correlations and the patterns of user activity. Furthermore, using the emotion classifier SentiStrength [15,16], which has been developed for a graded estimate of the emotional content in short informal texts, we assess the emotion valence in the text of each message in our datasets. This enables us to further analyse how the flow of emotion adheres to the topology of the network and with the observed collective behaviours of users.

2 Empirical data from MySpace social network

For the research that we pursue in this study, high-resolution data are necessary, in particular, (i) a *connected network* of users identified by their IDs as nodes, and the exchanged messages as links, and (ii) *each message identified* by its source and target nodes (user IDs), the time the message was generated, and its textual content. Arguably, an ideal online social network for this analysis is MySpace, having the data of such structure systematically recorded until 2010.

2.1. Data crawling and structure

Communications between users in the online social network MySpace occur along friendship links: writing messages on their *own wall*, where they can be seen by linked friends, or writing to the linked friend's wall. Privacy policy varies from user to user, and hence some users do not allow access to their data. However, the messages that they sent to the users who allow access can be identified.

To obtain the networked dialogues in MySpace, we developed a parametrized Web crawler that visits mutually linked users and collects the dialogues (messages), which were posted by them *within a specified time window*. To start, we first specify the time window that we are interested in, and find an appropriate node, representing a *user* who was active within that time window. The crawler then proceeds to search in the neighbourhood of that node in a *breadth-first manner*, as schematically shown in figure 1, and collects all messages posted within that time window on the current user's wall and identifies their source and posting time. The links in figure 1 indicate that at least one message was exchanged between the linked users within the specified time window. Starting from the initial node marked as '1' the list with first layer of the connected nodes is explored, then the search is continued from each node on that list, thus making the second layer nodes. Then, the lists are swapped, and so on. The crawler is instructed to avoid the nodes whose data are marked as 'private' and also the nodes that contain over 1000 pages of messages (probably representing a non-human user). The crawler can identify the messages posted from these types of nodes that can be seen at the walls of their neighbours. Such nodes and their messages are not considered in the analysis. Full information about the exchanged messages along each discovered link is stored in a database, in particular, the identity of the source and the target node, the creation time and full text of each message. Given a parametrized search, the crawler can stop either when no new nodes are found that satisfy the parameter criteria (time depth) or when it reaches a given diameter (number of layers).

Our data (fully anonymized data available, doi:10.5061/dryad.5253n or from http://www-f1.ijs.si/~tadic/Art/SI/ (registration required)) were compiled in 2009 and contain messages for two time windows—two months and three months depth, but starting from the same initial node. One dataset for the two month period, January and February 2008, consists of $N_M = 80\,754$ messages exchanged between $N_U = 36\,462$ users. The larger dataset corresponds to a time depth of three months, from 1 June till 1 September 2008, contains $N_U = 64\,739$ users and $N_M = 172\,127$ their messages.

2.2. Inference of the emotional valence from the text of messages

For further analysis, we performed automatic screening of the text of dialogues in our dataset to extract the emotional valence of each message. We apply the emotion classifier, which has been developed by Thelwall and co-workers [15,16,25] for short text messages in MySpace dialogues. The classifier considers each message as a single document

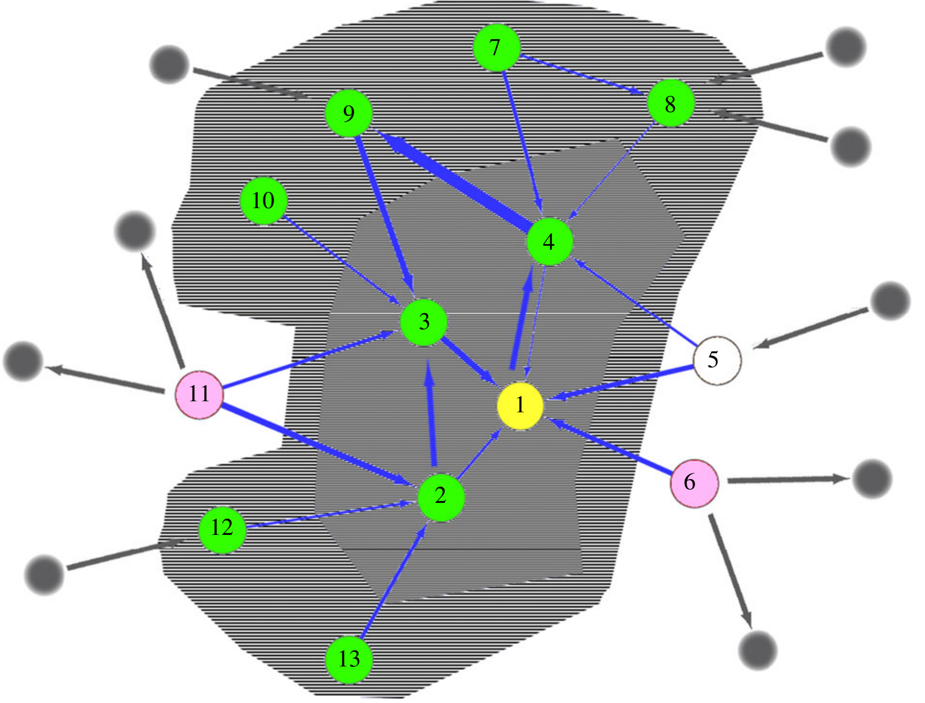

Figure 1. Schematic of the parametrized breadth-first search of the dialogues occurring within a specified time window, starting from a given user in `MySpace`, marked as node '1'. Nodes '6' and '11' indicate users who do not allow public access to their 'walls'; however, their messages left on the neighbouring 'walls' can be identified. These nodes and their links are not included in the data. The presence of bots, companies or another non-ordinary users is identified, an example depicted by node '5', and dropped from further search. (Online version in colour.)

and can extract graded emotion valence as two numbers (e_-, e_+), representing the intensity of the negative and positive emotion content of the same message.

According to references [15,16,25], the classifier algorithm is based on two emotional dictionaries *the general inquirer* and *linguistic inquiry and word count*. In each message, the algorithm detects all words that belong to the emotional dictionary and extracts their polarity and intensity. The obtained scores are then modified with additional linguistic-based rules if special terms, such as negators (*good* versus *bad*), intensifiers (*liked* versus *liked very much*), diminishers (*excellent* versus *rather excellent*) are found in the neighbourhood of that word in an area of *five* words before and after the emotional term or the beginning or end of the sentence. Other special terms such as capitalization (*bad* versus *BAD*), exclamation and emotion detection (*happy!* or *:-)*) are searched and treated similarly as intensifiers. If an intensifier (or diminisher) word is found, then the absolute original emotional value of the term is increased (or decreased) by one. For example, if *bad* is given (-3) then *very bad* is (-4). Similarly, if *good* is judged as ($+3$) then *somewhat good* is ($+2$), whereas *very good* is ($+4$). The scores e_+ and e_- that the classifier returns represent the maximum positive and the maximum negative number for a given message. Accuracy and other relevant details can be found in the original references [15,16,25].

The classifier returns two independent ratings for every message: one for the positive $e_+ \in \{+1, +2, +3, +4, +5\}$ and one for the negative $e_- \in \{-1, -2, -3, -4, -5\}$ dimension. On this scale, $e_- = -5$ corresponds to very negative and $e_+ = +5$ to very positive emotion valence. While $e_- = -1$ and $e_+ = +1$ indicate the absence of negative and positive emotion, respectively. On the basis of these automated ratings, we can construct the overall valence polarity of a particular message. Specifically, all messages for which the scores are in the range ($e_- \leq -3$ and $e_+ = +1$ or $e_+ = +2$) are considered as carrying *negative emotion valence*, and symmetrically the messages with ($e_+ \geq +3$ and $e_- = -1$ or $e_- = -2$) are classified as carrying *positive emotion valence*. Notice that this excludes a certain number of messages for which the two scores are exactly balanced, ($e_- = e_+$), although they may contain 'emotional' words. Also a small fraction of messages for which the negative and positive scores are simultaneously very high ($e_+ \geq 3$ and $e_- \leq -3$) are disregarded as possible artefacts of the graded-strength classifier. In total, a fraction 0.15 of all messages from the two month sample, and 0.16 from the three month sample, remain unclassified.

2.3. Methodology for data analysis

The dataset for a given time window is mapped onto the dialogues-based network in the following way: each user is represented by a node and a link is inserted between the pair of nodes if a dialogue (at least one message) was detected between them in that time window. The direction of the link indicates the message from source-to-target, and the link multiplicity (weight) represents the number of messages. In some figures, colours are used to suggest net emotion balance along the link, which is derived form the original scores of all messages along that link.

Our high-resolution datasets contain valuable information, which enables us to study other features, in particular the temporal sequences of user activities and the patterns of emotion flow via message exchange. To study such diverse aspects of the online social networks, appropriate methodologies and mathematical approaches are applied. In particular:

— *the network topology* is analysed using the graph theory methods [31]. Specifically, we determine several topology measures at global network level (distributions of degree, strength, weight, betweenness) and at local

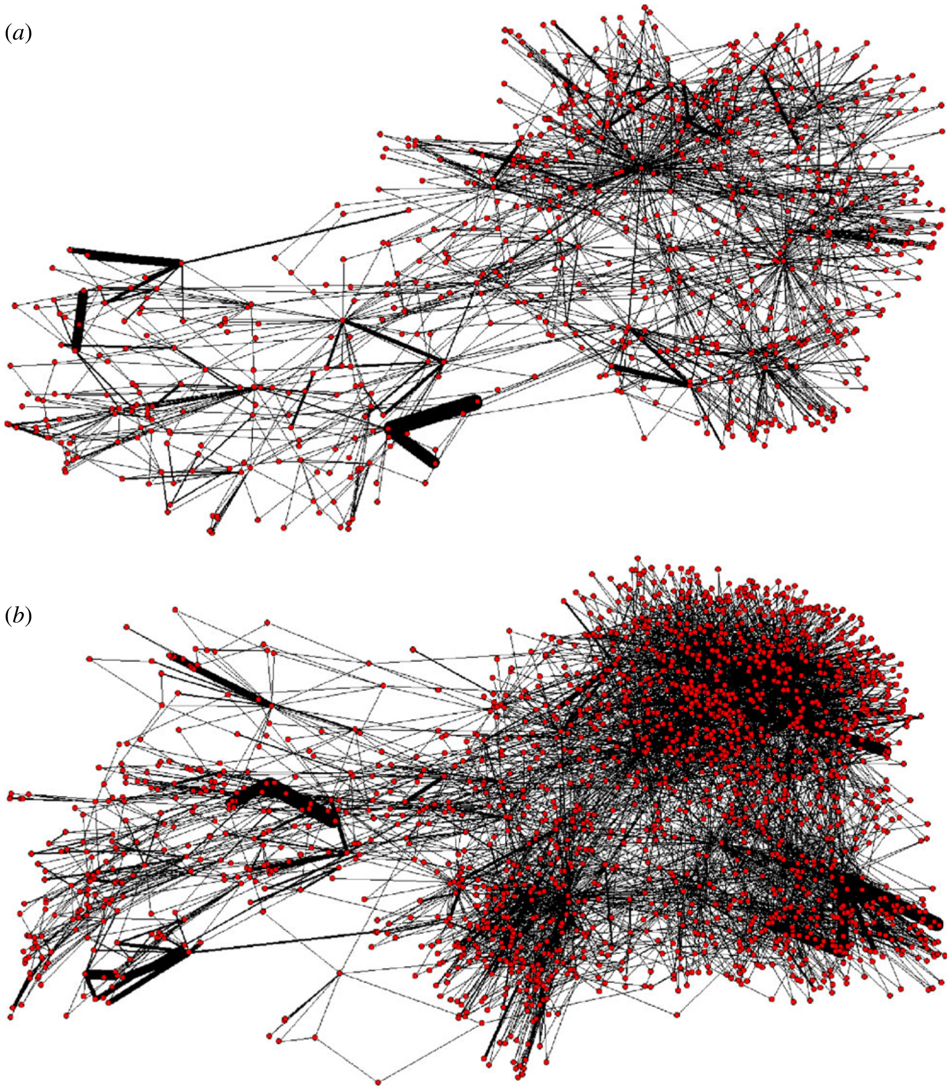

Figure 2. Initial part of the network of users connected by dialogues in `MySpace`, as compiled by our crawler for the time window of (*a*) two months and (*b*) three months starting from the same user-node. (Online version in colour.)

node-neighbourhood level (assortativity measures and tests of social weak-tie hypothesis), as well as the mesoscopic level (community structure). For the community structure analysis in compact and relatively small networks, we apply the eigenvalue spectral analysis of the Laplacian operator related to the *weighted symmetrical* adjacency matrix [32]. In the case of large graphs, the communities are detected by `Gephi` software, which uses a maximum modularity approach [33]. The greedy algorithm is used to determine maximum-flow spanning trees of the networks;

— *temporal correlations* are studied by the power spectrum and fractal detrended fluctuations analysis of the time series. Three types of time series are constructed, i.e. the time series of the number of messages, and the number of messages carrying either positive or negative emotion, per small time bin $t_{\rm bin} = 5$ min; and

— *patterns of user activity* are mapped directly from the dataset in the original temporal resolution. Each action (message) of a given user at a given time is represented by a point on the temporal pattern. The interactivity time is identified as the distance between two consecutive points along the time axis from the activity pattern of each user.

Various histograms obtained in this analysis are fitted to the corresponding theoretical expressions using either χ^2 or the maximum-likelihood estimator method. All fits passed the χ^2-test of goodness. Logarithmic binning with base $b = 1.1$ is often applied to obtain smooth curves.

3. Topology of the dialogues-based networks in `MySpace`

Using the earlier-described procedure of data mapping, we obtain networks of `MySpace` users as nodes, connected by directed weighted links, representing the messages sent from one user to another. In figure 2, two examples of such dialogue-based networks are shown. In figure 2*a*, a part of the network of dialogues observed within the time window $T_{\rm WIN}$= two months is shown, whereas figure 2*b* shows the corresponding network for the situation when the searched time depth is $T_{\rm WIN}$= three months, starting from the same initial node. Only a small initial part of the corresponding dataset is shown for the illustration. The increased time depth is manifested in that more nodes are connected to the network, the link density is increased, as well as the widths of some already existing links. Moreover, the community structure—visually identified as groups of nodes—is evolving.

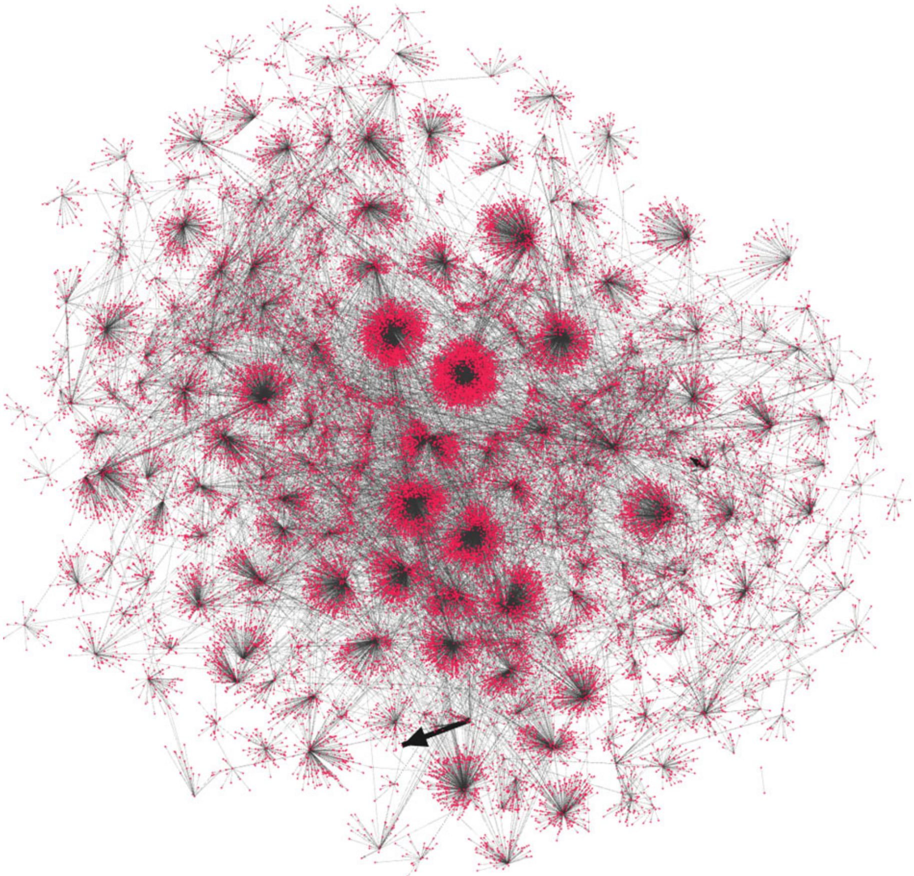

Figure 3. A view of the network Net2M from the dataset of dialogues with two-months depth in MySpace. The 33 649 users are organized in 87 communities, visualized as blobs of different sizes and densities.

3.1. Density and reciprocity of links

Here, we consider the topology of two networks representing all *emotion-classified dialogues* in a two-month time window (Net2M), and in a three-month time window (Net3M). They contain $N = 33\,649$ and 58 957 users, respectively. These networks appear to be very sparse (cf. figure 3). The average link density, defined as $\rho \equiv L/N(N-1)$, where (L) is the number of all directed links, is found as $\rho = 3.345 \times 10^{-5}$ for Net2M, and $\rho = 2.08 \times 10^{-5}$ for Net3M network. Furthermore, we compute the *link reciprocity*, which is defined [34] as $r \equiv (L^{\leftrightarrow}/L - \rho)/(1-\rho)$, where $L^{\leftrightarrow}$ is the number of links occurring in both directions $i \to j$ and $j \to i$ disregarding the weights. We find $r = 0.0227$ for Net2M, and $r = 0.0214$ for Net3M network, i.e. the reciprocity is barely positive. According to Garlaschelli & Loffredo [34], in social networks, larger positive reciprocity is expected. The clustering coefficients $C_c = 0.013$ and 0.014 are found for these two networks when the directedness of the links is ignored. We also consider a reduced network, termed Net3321, which is extracted from the two-months dataset. In this network, the users who sent and received less than four messages within two months were excluded. The reduced network contains $N_U = 3321$ nodes. It is more compact, i.e. $\rho = 7.19 \times 10^{-4}$ and has larger link reciprocity $r = 0.118$ and a $C_c = 0.084$ for directed, and 0.115 for undirected links.

3.2. Community structure

In figure 3, the entire network of Net2M is shown. The network is organized as a large number of *communities*. Specifically, 87 communities shown in figure 3 are obtained using the weighted maximum modularity algorithm [33] with Gephi software. The largest community contains 2543 users. It is interesting to note that the size distribution of these communities obeys Zipf's law. In figure 4(*a*), two curves in the inset are the ranking distributions of the community sizes, which are detected in the networks of the two-month dialogues and three-month dialogues, respectively. It should be stressed that the number of communities that one observes depends on the resolution of the algorithm. By increasing the resolution, some communities can further split into two or more groups, and conversely, join together to make a larger community when the resolution is decreased. However, the scale-free organization of communities up to a certain size persists (with changed parameters), as shown in the main figure 4(*a*). All curves are fitted by the discrete generalized beta distribution [35], $y = Ax^{-b}\,(N+1-x)^c$ with the parameters (*b*,*c*) listed in the figure caption.

In the Net3321, nodes of low strength $\ell_{\text{in}} + \ell_{\text{out}} \leq 4$ are excluded, as mentioned earlier. In figure 4(*b*), we confirm that this more compact network also exhibits a community structure. We use the eigenvalues spectral analysis [32] of the normalized Laplacian $\mathscr{L} = \delta_{ij} - W^S_{ij}/\sqrt{\ell_i \ell_j}$, where ℓ_i and ℓ_j are the total strengths and $W^S_{ij} \equiv W_{ij} + W_{ji}$ is the symmetrical weighted adjacency matrix of the network. The property that the eigenvectors belonging to the lowest non-zero eigenvalues of the Laplacian tend to localize on the subgraphs [32] is used to identify four communities of this network.

3.3. Nodes inhomogeneity and mixing patterns

The cumulative distributions of node's degree and strength, and the mixing patterns for the networks Net2M and Net3M are computed, and the results are shown in figures 5 and 6. The out-degree and the out-strength distributions,

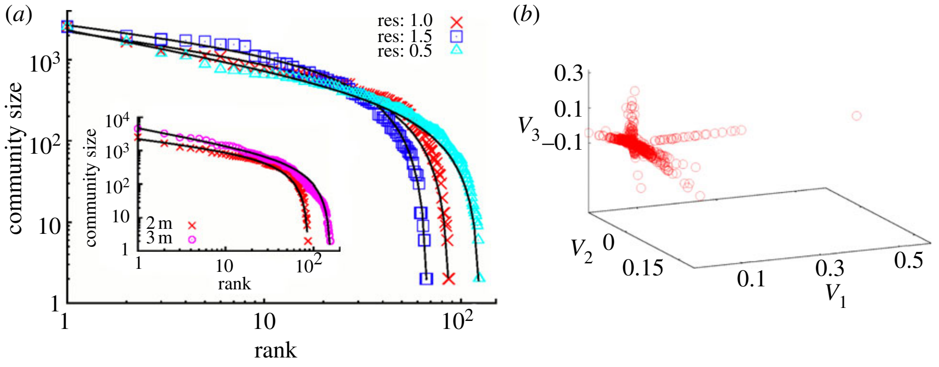

Figure 4. (*a*) Ranking distributions of the size of communities in the networks of two-months dialogues in `MySpace` computed in standard, lower and higher resolution. The solid curves are the best-fit discrete generalized beta distributions with parameters $(b;c) = (0.34 \pm 0.01;\ 1.45 \pm 0.05)$, $(0.31 \pm 0.01;\ 1.49 \pm 0.005)$ and $(0.46 \pm 0.01;\ 1.249 \pm 0.003)$, respectively. Inset: comparison of the distributions for two-months and three-months dialogues networks in standard resolution. The three-months data fit: $(0.505 \pm 0.005;\ 1.579 \pm 0.002)$. (*b*) The communities in `Net3321` are indicated by different branches in the scatter-plot of the eigenvectors (V_1, V_2, V_3) of three lowest non-zero eigenvalues of the Laplacian. (Online version in colour.)

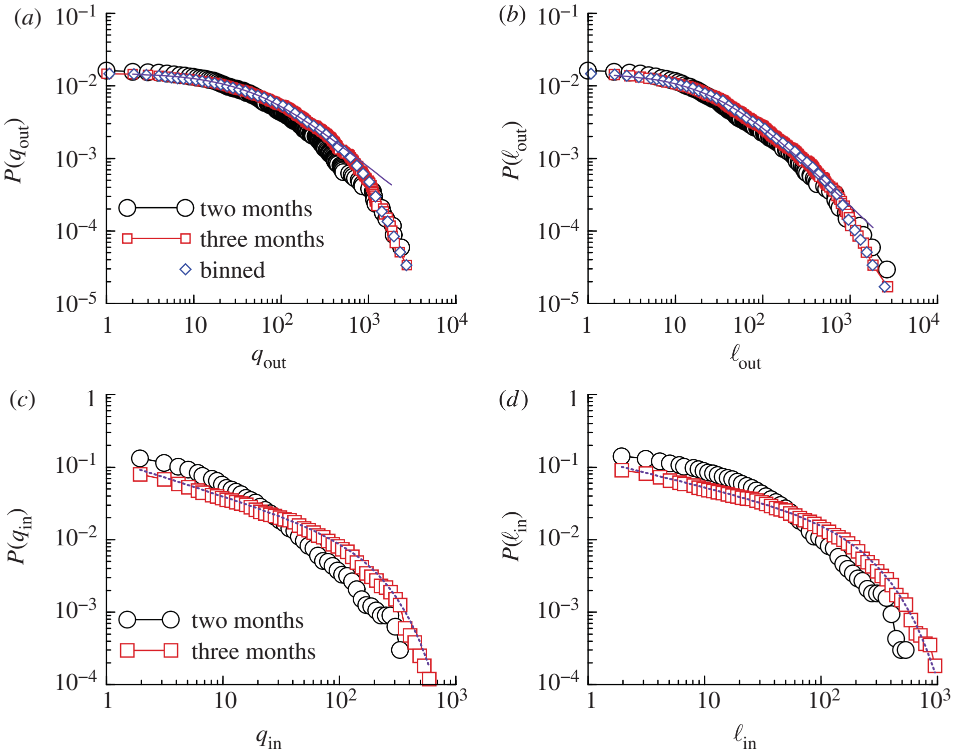

Figure 5. Cumulative distributions of (*a*) out-degree, (*b*) out-strength, (*c*) in-degree and (*d*) in-strengths of nodes in the networks of `MySpace` dialogues within two-months and three-months time windows. Log-binned data. Fits are described in the text. (Online version in colour.)

which are controlled by the user represented by that node, appear to be stable with the increased time depth (cf. figure 5*a*,*b*). Apart from very large degrees where data are missing, these distributions are fitted by the expression $P_{\rm out}(X) = A(1 + X/X_0)^{-\tau_{\rm out}}$. The power-law section with the exponent $\tau_{\rm out} = 1.18 \pm 0.01$ after the cut-off $X_0 = 20.3 \pm 0.2$ is found for the degree, and $\tau_{\rm out} = 0.99 \pm 0.01$ after the cut-off $X_0 = 53.1 \pm 0.5$ for the strength distribution, respectively. By contrast, the in-degree and in-strength of a node are given by incoming messages to the wall of the user, thus depending on the cumulative activity of its neighbours. Consequently, we find that the corresponding distributions evolve with the searched time depth and eventually obtain a power-law decay with a stretched exponential cut-off $P_{\rm in}(X) = BX^{-\tau_{\rm in}} {\rm e}^{-(X/X_0)^{\sigma}}$, as shown in figure 5*c*,*d*. The fit parameters are $\tau_{\rm in} = 0.53 \pm 0.07$, $X_0 = 226 \pm 32$, $\sigma = 1.25$ for the in-degree, and $\tau_{\rm in} = 0.88 \pm 0.04$, $X_0 = 248 \pm 17$, $\sigma = 1.06$ for the in-strength distribution, respectively.

The assortativity measures are another characteristic of the network topology at the level of node's neighbourhood. Specifically, situations when nodes with large degree are linked to each other (assortativity), or conversely, nodes with large degree have a large number of nodes with small degree (disassortativity), will be expressed by increasing (decreasing) slopes of the degree–correlations plots, such as those in figure 6(*a*). In view of the *weighted directed* links of the dialogues network, several such measures can be determined. The results are shown in figure 6*a*,*b* for the degree and for the strength mixing, respectively. Specifically, the *in–out* correlation is determined by considering a node with a given in-degree $q_\kappa = {\rm in}$, and computing the average out-degree $\langle q_{\mu={\rm out}}\rangle_{\rm nn}$ over the neighbour nodes, sources of the incoming links. We find a disassortative pattern, i.e. $\langle q_{\mu={\rm out}}\rangle_{\rm nn} \sim q_{\rm in}^{-\mu}$ with the decreasing slope $\mu \sim 1$. The tendency of flattening at large-degree nodes is compatible with the presence of communities. A similar finding applies for

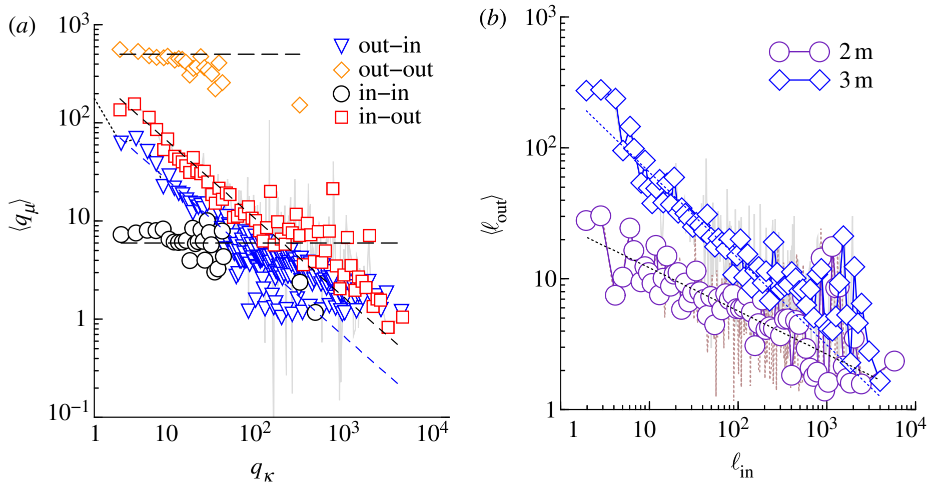

Figure 6. Mixing patterns in MySpace dialogues network: (*a*) correlations between in- and out-degrees (indicated) for the three-months dialogues; (*b*) in–out-strengths for the networks of three-months and two-months dialogues. Log-binned data. The dashed lines indicate the slope $\mu \sim 1$, and the dotted lines $\mu = 0.33$ and $\mu = 0.86$. (Online version in colour.)

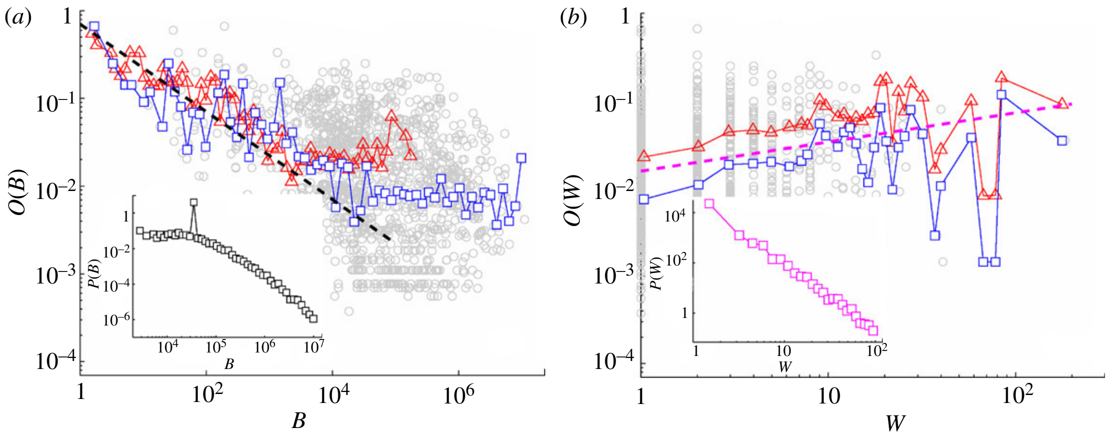

Figure 7. The averaged overlap *O* plotted against (*a*) betweenness centrality *B* and (*b*) weight *W*, for the symmetrical links on MySpace dialogue network of two-months window (squares) and Net3321 (triangles). The dashed lines indicate the slopes $-\frac{1}{2}$ and $+\frac{1}{3}$, conjectured in Szell & Thurner [14]. Insets: the distributions *P*(*B*) and *P*(*W*) for the network of two-months window. Data are logarithmically binned. (Online version in colour.)

the *out–in* mixing. On the other hand, no assortativity measures can be observed in out–out and in–in patterns, represented by the flat curves in figure 6(*a*). The results are from the three-months dialogues dataset. The disassortative mixing is already present in two-months dialogues, although with a smaller slope. In figure 6*b*, the results for the in–out strength mixing are compared for two-months and three-months dialogues networks. These findings suggest that a particular pattern, where a large number of small-degree nodes communicate with one large-degree node, occurs quite often. Note that the static structure of the friendship links in MySpace itself shows a tendency towards disassortative mixing [30]. The observed striking disassortativity of the dialogues-based networks in MySpace stands in a clear contrast to the assortativity of common social structures [36] and of the static friendship connections in Facebook [4].

3.4. Confirmation of the weak-ties hypothesis

As quantitative measures of the social dynamics weak-ties hypothesis, it is common practice to determine the nature of correlations between the *overlap* O_{ij} of two neighbouring nodes and properties of the link (*ij*) connecting them [17,37]. Specifically, considered are the weight W_{ij} and the betweenness centrality B_{ij} of the link. $B_{ij} = \sum_{s,t \neq s} (\sigma_{st}(ij)/\sigma_{st})$ is given as the fraction $\sigma_{st}(ij)$ passing through that link of the shortest paths σ_{st} between all pairs of nodes (*s*,*t*) on the network [31]. Note that for this purpose the network is considered as undirected. The overlap of two adjacent nodes *i* and *j* is $O_{ij} = m_{ij}/(q_i + q_j - 2 - m_{ij})$, where q_i and q_j is the total degree of node *i* and *j*, respectively, and m_{ij} is the number of their common neighbours. In traditional social dynamics, it is expected that the overlap increases with bond strength, i.e. $O_{ij}(W) \sim W^{\eta_1}$. Moreover, owing to the ubiquitous community structure, the bonds with large betweenness, e.g. connecting different communities, should not have a large overlap. Consequently, $O_{ij}(B) \sim B^{-\eta_2}$. In the case of online social contacts, the weak-ties hypothesis was confirmed for networks of e-mails [37] and online games [14]. It was conjectured in Szell & Thurner [14] that the universal exponents $\eta_1 = \frac{1}{3}$ and $\eta_2 = \frac{1}{2}$ should apply. Our numerical results for the MySpace two-months dialogues network and for more compact Net3321, shown in figure 7*a*,*b*, confirm this conjecture. The histograms of the betweenness $P(B)$ and of the weight $P(W)$ computed for all links in two-months dialogues window are shown in the insets. The power-law tails of these distributions suggest a large diversity both in the organization of the communities and in the intensity of communications within them.

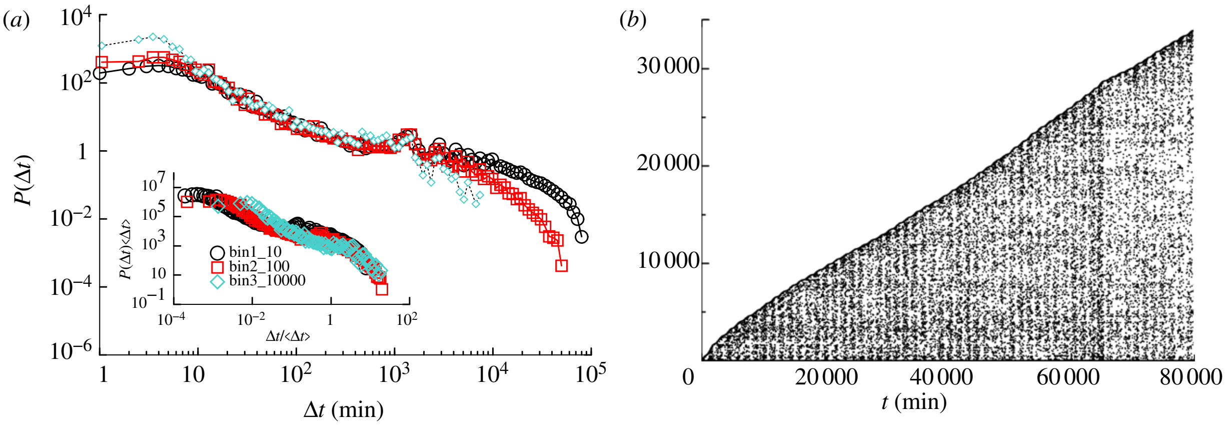

Figure 8. (*a*) Distributions $P(\Delta t,b)$ of the user delay time Δt for three activity bins. Inset: scaling plot of the distributions. Logarithmically binned data. (*b*) User-activity pattern in MySpace dialogues from two-months window. (Online version in colour.)

4. Activity patterns and emotion flow

4.1. Temporal patterns of user activity

In our high-resolution data, information about the activity of every user over time is presented as a pattern in figure 8(*b*). Time in the original resolution is plotted along the *x*-axis, and each integer number along the *y*-axis stands for an user index. The indexes are ordered by the user's first appearance in the dataset. Two characteristic features of these temporal patterns are (cf. figure 8):

— *arrival of new users*, depicted with the top boundary of the pattern, follows daily cycles. With the appearance of new users (relative to the beginning of the dataset), the system experiences increased activity, manifested as larger density of points below each 'wave' of new users. Note also the stripes inclined upwards, which indicate possible correlated actions of the users involved in later times (in §5 the temporal correlations are studied in detail). Both the arrival of new users and the increased activity of all other users obey periodicity, compatible with the *circadian cycles*, which is related to offline behaviour. Similar features are observed in Blogs & Diggs [17] and other social systems [7], confirming the importance of circadian cycles in online dynamics; and

— *delay (interactivity) times of user actions* Δt, defined as the time between two consecutive actions of a user, is a quantitative measure of fractality of the temporal activity pattern. Namely, following the line of a given user, we find no characteristic distance between two consecutive activity points. The distribution $P(\Delta t,b)$ of the distances Δt between subsequent points for a given user, then averaged over all users in a data bin, leads to a separate curve in figure 8(*a*). Three data bins are introduced to take into account large differences in user activity as follows: in the first bin are the users who posted between (1–10) messages for the entire period of two months. The majority of users belongs to this group. Consequently, their delay times can get very large (cf. figure 8*a*). Users in the second bin posted between (11–100) messages, having moderately shorter average delay time. The bin with (101–10000) messages contains few but very active users, whose delay times are much shorter. Compared with some other online systems [38], the observed inhomogeneity of MySpace users is stronger. Nevertheless, the distributions $P(\Delta t,b)$ for different bins share some common features that can be quantified by the universal scaling curve. The attempted scaling plot according to $P(\Delta t,b) = b^{-1}\mathcal{P}(\Delta t/b)$ is shown in the inset of figure 8(*a*), where the scale *b* of each data bin is chosen, following Radicchi [38], as the average delay time $b = \langle\Delta t\rangle$ in the corresponding data bin.

The occurrence of a broad distribution of the delay time was recently observed in different online communication systems [5,11,39]. The present analysis suggests that the activity in the MySpace social network is somewhat specific, exhibiting a different regime for the short delays (5–85) minutes, compared with the delays longer than 1 day.

4.2. Structure of the emotional dialogues

We consider in detail the emotion content transmitted along the links in the network Net3321. As stated earlier, the directed weighted link W_{ij} on this network represents the number of messages sent from the node *i* to the node *j* within the time window of two months. Here, we also include the emotion valence of these messages. By adding the emotion contained in each message along the link, we obtain the overall valence, i.e. as a positive, negative or neutral link. The network is shown in figure 9(*a*) with the links coloured according to their emotion content—red (positive), black (negative) and blue (neutral). The size of the nodes indicates their degree. Each node carries a label—unique user ID index derived from the original data. The enlarged part of the network, which is displayed in figure 9(*b*), illustrates a typical structure of the emotion-carrying links between the hubs and the number of small-centrality nodes surrounding them.

Considering the emotion content of the messages along the links, we find that positive emotion (attractiveness) dominates the connections in MySpace dialogues, whereas the links carrying negative emotion (aversiveness) occur rather sporadically. The temporal correlations of the positive (negative) emotion messages are further studied in §5. Here, we analyse the topology of the emotional connections on the network. We extract the subnetwork of negative links, whose fragments are found in different parts of the Net3321. The largest connected component of the negatively linked network is shown in figure 10(*a*). The enlarged part shown in

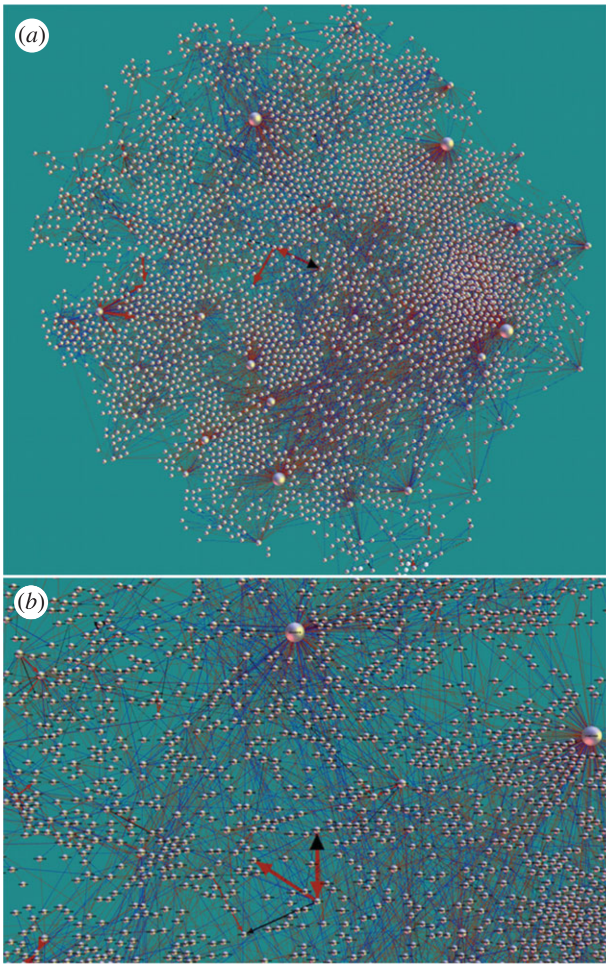

Figure 9. (*a*) A view of the dialogues network `Net3321` and (*b*) a close-up part. The sizes of nodes correspond to their degree. The widths of links represent the cumulative number of messages within two-months window, whereas the colours indicate their emotional valence—positive (red), neutral (blue) and negative (black).

figure 10(*b*) demonstrates typical flow patterns of the negative-emotion messages. Specifically, a node may act as a source or a sink of the negative links, can transmit or disseminate the emotion, or be involved in multiple reply-to events with the same emotion valence. Note that these two subnetworks with positive and negative emotion links are integrated into each other, and also that some messages (links) can be considered as neutral, i.e. carrying information, but no emotion. Therefore, the nodes that change the valence of the emotion messages have a special role as pinning centres for the propagation of the emotional bursts on the entire network.

For most of the links, the computed overlap is related to the widths of the links (cf. figure 7) according to the social tie hypothesis. However, occasionally a strong link appears, i.e. carrying a large number of messages (visible on the network in figure 9) disproportionally to the topological centrality of the adjacent nodes. To find out how most of the messages (and emotions) flow on the entire network, we analyse the maximum-flow spanning trees. It is a suitable tree-representation of the network, where each node is connected to the tree by its *strongest link*. In figure 11, the maximum-flow spanning three of the `Net3321` is shown. The tree is constructed using a variant of the greedy algorithm by ordering the *total weights* of the links $W_{ij} + W_{ji}$. Considerable side branching suggests heterogeneity in the intensity of the dialogues inside the existing communities. Moreover, it often occurs that a small-degree node interpolates in the branching process and transmits the flow of messages between the hubs. This is in agreement both with the observed community structure and the disassortativity of the dialogues-based network, discussed in §3. These findings further support the conclusion that the dynamical use of the links in `MySpace` social network, captured by the dialogues-based networks, is different from the conventional social structures.

5. Correlations in time series with emotional messages

The stochastic processes governing the communication with emotional messages among the users in `MySpace` can be studied from the point of view of time series analysis. From the datasets of the networked dialogues in a given time window, various time series have been constructed here—for instance, the series that contain the *number of all identified messages per small time bin*. Similarly, we construct the time series of the number of messages carrying either positive or negative emotion valence. The time bin $t_{\text{bin}} = 5$ min is used as the natural resolution in these data. The resulting time series are shown in figure 12*a*,*b*.

These time series exhibit fluctuations with strong daily periodicity (circadian cycles), as also observed in the activity patterns in figure 8(*b*). This periodicity is manifested as a pronounced peak in the power spectrum at the corresponding frequency, i.e. at the index value $\nu \approx 56$ in this case. Apart from the peak, the power spectra in figure 12(*c*) are of the coloured-noise type (fractal time series). Specifically, the spectrum can be expressed as $S(\nu) \sim 1/\nu^{\phi}$ for a range of frequency indexes below approximately 3000 (corresponding to the time scale longer than 2 h) in the case of all messages, and similarly in the time series of messages with positive emotion for $\nu \lesssim 1000$ (or $t \gtrsim 5$ h). These features of the time series suggest the occurrence of long-range correlations in the fluctuations in number of messages of all types and in the messages with positive-emotion valence. The spectrum of the negative-valence messages appears to be close to the white noise signal ($\phi \gtrsim 0$). The corresponding exponents are: $\phi^a = 0.59 \pm 0.08$, $\phi^+ = 0.55 \pm 0.08$ and $\phi^- = 0.15 \pm 0.06$, where the symbols $a, +, -$ stand for all messages, and messages with positive and negative-emotion valence, respectively.

The fractal analysis of these time series is completed by determining the Hurst exponent H. For each time series $n(k)$, $k = 1,2, \ldots T$, the profile $Y(i) = \sum_{k=1}^{i} (n(k) - \langle n \rangle)$ is divided into N_s segments of length s. The scaling behaviour of the fluctuation with the varied segment length defines the exponent via $F_2(s) = ((1/N_s) \sum_{\mu=1}^{N_s} F^2(\mu, s))^{1/2} \sim s^H$. The fluctuation at the μth segment $F^2(\mu, s) = (1/s) \sum_{i=1}^{s} [Y((\mu - 1)s + 1) - y_\mu(i)]^2$ is the standard deviation from a local trend $y_\mu(i)$ on the segment μ. Note that these time series are stationary. To remove the local trends related to the daily cycles (and potentially longer cycles), we adapted the *detrended fractal analysis* with polynomial interpolation in the intervals $s = 2m + 1$ overlapping over $m + 1$ points, described in references [40,41]. The trend at the scale of 144 time bins, corresponding to 12 h, is plotted over the original signal in figure 12(*b*). By removing the daily cycle, the fluctuations $F_2(s)$ of the detrended signals of

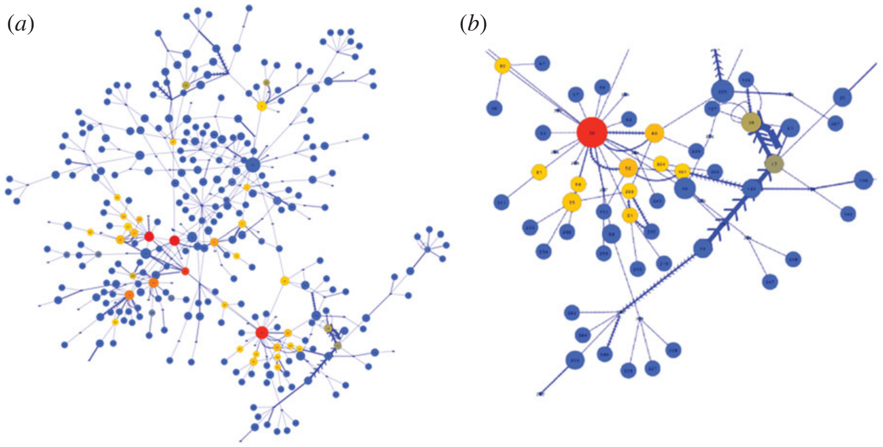

Figure 10. (*a*) The largest connected component of the subnetwork with negative dialogues on `Net3321`. (*b*) In the close-up part, the link directions and weights show a typical pattern of negative valence dialogues. The size and colour of nodes indicate their out-degree and betweenness-centrality. (Online version in colour.)

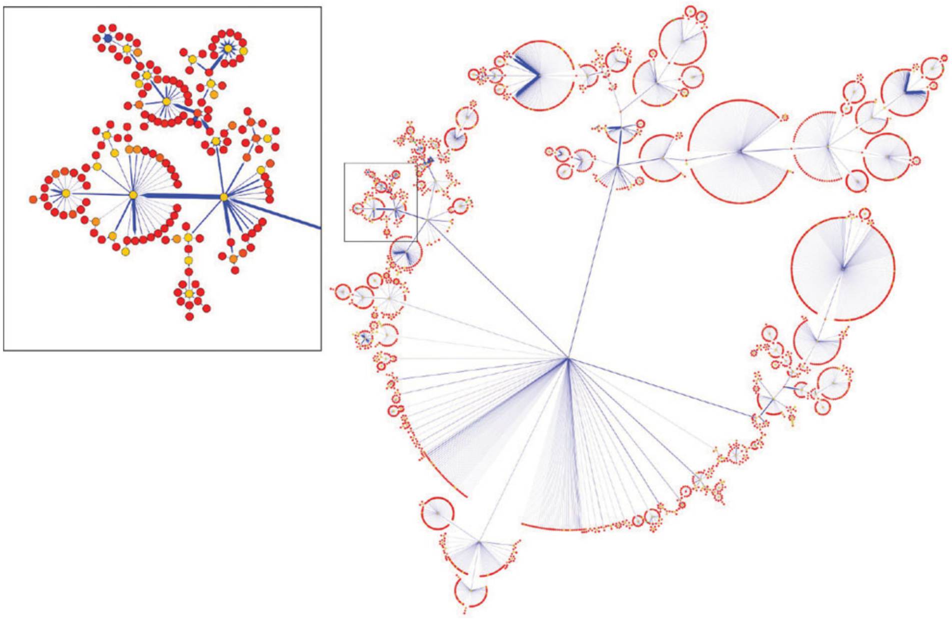

Figure 11. The maximum-flow spanning tree of `Net3321`. The enlarged part shows a typical structure of the tree away from the root. (Online version in colour.)

all three time series have been computed and are shown in figure 12(*d*) by the respective symbols. The corresponding Hurst exponents in the scaling region (4− − 288) time bins are listed in the legend. For comparison, also shown are the fluctuations of the trend (dotted lines) and of the whole signal (full lines). The obtained values of the Hurst exponents in the range $H \in (\frac{1}{2}, 1)$ confirm that the long-range correlations with *persistent fluctuations* occur in user activity and in the emotional messages of both polarities.

6. Conclusions

We have performed a comprehensive analysis of the empirical data of user dialogues from the social network `MySpace` focusing on the quantitative analysis of users collective behaviour. Our methodology that can be used across a large class of online systems with user-to-user interactions and high-resolution data includes: compiling the data of suitable structure, extracting emotion content from texts of messages by automated methods tailored for this type of textual documents, and analysing the data with graph theory and statistical physics by properly accounting for the nonlinear dynamic effects. Our main findings can be summarized as follows:

— *dialogues-based networks* in `MySpace` are dynamical structures built upon 'friendship' connections. They organize in a large number of communities of various sizes and the 'weak-tie' hypothesis holds in a manner similar to online games and e-mail networks. The actual use of links (within a given time window) reveals unbalanced message flow, yielding different organization of incoming and outgoing links; several hubs emerge to whom communication is directed from a large number of 'small' nodes, manifested in strongly disassortative mixing; furthermore, the emotion content of the messages passed along these links is dominated by positive emotion (attractiveness), while the links with negative emotion (aversiveness) appear less often, but make a specific local pattern on the network;

— *self-organized processes* of message exchange among the users have long-range temporal correlations of various degrees and persistent fluctuations, which clearly depend on the emotion content of the messages; and

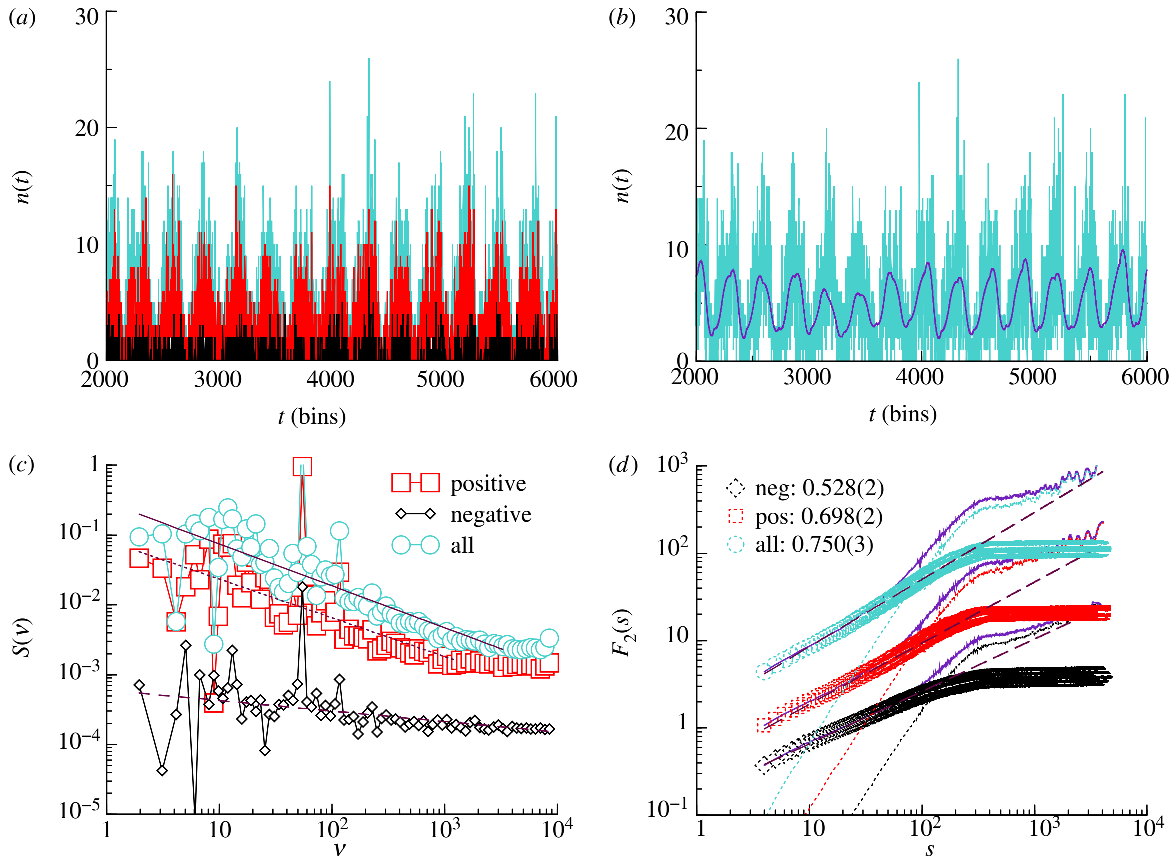

Figure 12. (*a*) A part of the time series of the number of messages from MySpace dialogues for the two-months window: all messages (cyan), and the messages classified as carrying positive (red) and negative (black) emotion valence. The length of each time series is $T = 16384$ time bins (one bin corresponds to 5 min). (*b*) The local trend with the daily cycle (thick line) plotted over the original signal. (*c,d*) The power spectrum $S(\nu)$ of the time series and their dispersion $F_2(s)$: the original (full lines) and the corresponding detrended series (symbols) and the trends (dotted lines).

— *robust patterns of user behaviours* are observed, which are linked with circadian cycles. Broad distributions of delay times obey scale invariance over different activity intervals.

In conclusion, the presented multi-analysis approach sheds new light on actual problems of the functional structure in online social networks. The studied patterns in MySpace, considered as one of the largest 'social sites', reveals the dynamical structure that by many measures disagrees with common social networks. The self-organized dynamics with message exchange leads to specific organization of the users, where a large diversity of groups occur as well as the importance of individuals within these groups. Moreover, the emerging collective behaviours depend both on the intensity of communications and the amount of emotion passed with the messages. Disproportional dominance of positive emotions (attractiveness) may also suggest the presence of euphoria ('24 h party'-like behaviour). We hope that our quantitative analysis with the results presented in §§3–5 can help in ongoing research regarding 'who' uses social networks and 'how' online social networks are used and may serve as a basis for further data analysis and theoretical modelling.

Funding for this research was received in part from the programme P1-0044 of the Research agency of the Republic of Slovenia (project no. P1-0044-3). M.S. also thanks the research projects (OI171037 and III41011) of the Republic of Serbia and the project from FP7-ICT-2008-3 under grant agreement no. 231323. We thank George Paltoglou for providing a user-friendly version of the emotion classifier described in Paltoglou *et al.* [15]. Designed the research: B.T.; contributed new software and compiled the data: M.S.; performed emotion classification of the data: M.M.; analysed the results and produced figures: B.T., V.G., M.S., M.M.; wrote the manuscript: B.T.

References

1. Giles J. 2011 Social science lines up its biggest challenges. *Nature* **470**, 18–19. (doi:10.1038/470018a)
2. Kleinberg J. 2008 The convergence of social and technological networks. *Commun. ACM* **51**, 66–72. (doi:10.1145/1400214.1400232)
3. Cho A. 2009 Ourselves and our interactions: the ultimate physics problem? *Science* **325**, 406–408. (doi:10.1126/science.325_406)
4. Ugander J, Karrer B, Backstrom L, Marlow C. 2011 The anatomy of the Facebook social graph. (http://arxiv.org/abs/1111.4503)
5. Castellano C, Fortunato S, Loreto V. 2009 Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646. (doi:10.1103/RevModPhys.81.591).
6. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. 2003 Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103. (doi:10.1103/PhysRevE.68.065103)
7. Malmgren RD, Stouffer DB, Campanharo ASLO, Amaral LA. 2009 On universality in human correspondence activity. *Science* **325**, 1696–1700. (doi:10.1126/science.1174562)
8. Johnson NF *et al.* 2009 Human group formation in online guilds and offline gangs driven by a common

team dynamic. *Phys. Rev. E* **79**, 066117. (doi:10.1103/PhysRevE.79.066117)

9. Amichai-Hamburger Y, Vinitzky G. 2010 Social network use and personality. *Comp. Hum. Behav.* **26**, 1289–1295. (doi:10.1016/j.chb.2010.03.018)
10. Panzarasa P, Opsahl T, Carley KM. 2009 Patterns and dynamics of users' behavior and interactions: network analysis off and online community. *J. Am. Soc. Inf. Sci. Technol.* **60**, 911–932. (doi:10.1002/asi.v60:5)
11. Mitrović M, Tadić B. 2010 Bloggers behavior and emergent communities in blog space. *Eur. Phys. J. B* **73**, 293–301. (doi:10.1140/epjb/e2009-00431-9)
12. Szell M, Lambiotte R, Thurner S. 2010 Multirelational organization of large-scale social networks. *Proc. Natl Acad. Sci. USA* **107**, 13 636–13 641. (doi:10.1073/pnas.1004008107).
13. Mitrović M, Paltoglou G, Tadić B. 2010 Networks and emotion-driven user communities at popular blogs. *Eur. Phys. J. B* **77**, 597–609. (doi:10.1140/epjb/e2010-00279-x)
14. Szell M, Thurner S. 2010 Measuring social dynamics in a massive multiplayer online game. *Soc. Networks* **39**, 313–329. (doi:10.1016/j.socnet.2010.06.001)
15. Paltoglou G, Thelwall M, Buckely K. 2010 Online textual communication annotated with grades of emotion strength. In *Proc. Third Int. Workshop on EMOTION* (*satellite of LREC*)*: Corpora for research on emotion and affect*, pp. 25–30. 1–21 May 2010, Valletta, Malta. Paris: European Language Resources Association (ELRA). See http://www.lrec-conf.org/proceedings/lrec2010/workshops/W24.pdf#page=33.
16. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. 2010 Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **61**, 2544–2558. (doi:10.1002/asi.v61:12).
17. Mitrović M, Paltoglou G, Tadić B. 2011 Quantitative analysis of bloggers' collective behavior powered by emotions. *J. Stat. Mech. Theory Exp.* P02005. (doi:10.1088/1742-5468/2011/02/P02005).
18. Dodds PS, Harris KD, Koloumann IM, Bliss CA, Danforth CM. 2011 Temporal patterns of happiness and information in a global social network: Hedonometric and Twitter. (http://arxiv.org/abs/1101.5120v3)
19. Chmiel A, Sobkowicz P, Sienkiewicz J, Paltoglou G, Buckley K, Thelwall M, Holyst JA. 2011 Negative emotions boost user activity at BBC forum. *Physica A* **390**, 2936–2944. (doi:10.1016/j.physa.2011.03.040)
20. JA Coan, JJB Allen (eds) 2007 *The handbook of emotion elicitation and assessment*. Series in Affective Science. Oxford, UK: Oxford University Press.
21. Scherer K. 2005 What are emotions? And how can they be measured? *Soc. Sci. Inf.* **44**, 695–729. (doi:10.1177/0539018405058216)
22. Calvo RA, D'Mello S. 2010 Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* **1**, 18–37. (doi:10.1109/T-AFFC.2010.1)
23. Russell JA. 1980 A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178. (doi:10.1037/h0077714)
24. Bradley MM, Lang PJ. 1999 *Affective norms for English words* (*ANEW*)*: instruction manual and affective ratings*. Gainesville, FL: University of Florida. The Center for Research in Psychophysiology. See http://dionysus.psych.wisc.edu/methods/Stim/ANEW/ANEW.pdf.
25. Paltoglou G, Gobron S, Skowron M, Thelwall M, Thalmann D. 2010 Sentiment analysis of informal textual communication in cyberspace. In *Proc. ENGAGE 2010* (*Springer LNCS State-of-the-Art Survey*), pp. 13–23. Heidelberg, Germany: Springer.
26. Ryan T, Xsenos S. 2011 Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Comp. Hum. Behav.* **27**, 1658–1664. (doi:10.1016/j.chb.2011.02.004)
27. Yarkoni T. 2010 Personality in 100.000 words: a large-scale analysis of personality and word use among bloggers. *J. Res. Pers.* **44**, 363–373. (doi:10.1016/j.jrp.2010.04.001)
28. Cheung CMK, Chiu PY, Lee MKO. 2011 Online social networks: why do students use Facebook? *Comp. Hum. Behav.* **27**, 1337–1343. (doi:10.1016/j.chb.2010.07.028)
29. Ferrara E, Meo PD, Fiumara G, Provetti A. 2012 The role of strong and weak ties in Facebook: a community structure perspective. *Proc. Comp. Sci. Int. Conf. Comput. Sci. ICCS 2012*, pp. 1–10. See http://www.emilio.ferrara.name/wp-content/uploads/2011/06/ferrara-chasm-2012.pdf.
30. Ahn YY, Han S, Kwak H, Moon S, Jeong H. 2007 *Analysis of topological characteristics of huge online social networking services*. In *Proc. the 16th Int. Conf. World Wide Web*. pp. 835–844. Banff, Alberta, Canada. New York, NY: ACM. See http://doi.acm.org/10.1145/1242572.1242685. (doi:10.1145/1242572.1242685)
31. Bollobas B. 1998 *Modern graph theory*, 2nd edn. Heidelberg, Germany: Springer.
32. Mitrović M, Tadić B. 2009 Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Phys. Rev. E* **80**, 026123. (doi:10.1103/PhysRevE.80.026123)
33. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. 2008 Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* P10008. See http://stacks.iop.org/1742-5468/2008/P10008.
34. Garlaschelli D, Loffredo M. 2004 Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.* **93**, 268701. (doi:10.1103/PhysRevLett.93.268701)
35. Mart-nez-Mekler G *et al*. 2009 Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE* **4**, e4791. (doi:10.1371/journal.pone.0004791)
36. Newman MEJ. 2003 Mixing patterns in networks. *Phys. Rev. E* **67**, 026126. (doi:10.1103/PhysRevE.67.026126)
37. Onela J, Saramaki J, Hyvönen J, Szabo G, de Menezes M, Kaski K, Barabási A-L, Kertész J. 2007 Analysis of large-scale weighted networks of one-to-one human communications. *New J. Phys.* **9**, **179**. (doi:10.1088/1367-2630/9/6/179)
38. Radicchi F. 2009 Human activity in the web. *Phys. Rev. E* **80**, 026118. (doi:10.1103/PhysRevE.80.026118)
39. Vázquez A, Oliveira JG, Dezsö Z, Goh KI, Kondor I, Barabási AL. 2006 Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E* **73**, 036127. (doi:10.1103/PhysRevE.73.036127)
40. Gao J, Hu J, Mao X, Perc M. 2012 Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. *J. R. Soc. Interface* **9**, 1956–1964. (doi:10.1098/rsif.2011.0846)
41. Hu J, Gao J, Wang X. 2009 Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation. *J. Stat. Mech. Theory Exp.* P02066. See http://iopscience.iop.org/1742-5468/2009/02/P02066/.