Journal of Statistical Mechanics:
Theory and Experiment

SISSA

PAPER: INTERDISCIPLINARY STATISTICAL MECHANICS

The statistical mechanics of Twitter communities

To cite this article: Gavin Hall and William Bialek *J. Stat. Mech.* (2019) 093406

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

The statistical mechanics of Twitter communities

Gavin Hall[1,2] **and William Bialek**[2,3]

[1] Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, United States of America
[2] Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544, United States of America
[3] Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, 365 Fifth Ave., New York, NY 10016, United States of America
E-mail: gthall5816@gmail.com

Received 4 February 2019
Accepted for publication 31 July 2019
Published 27 September 2019

Online at stacks.iop.org/JSTAT/2019/093406
https://doi.org/10.1088/1742-5468/ab3af0

Abstract. We build models for the distribution of social states in Twitter communities. States can be defined by the participation versus silence of individuals in conversations that surround key words, and we approximate the joint distribution of these binary variables using the maximum entropy principle, finding the least structured models that match the mean probability of individuals tweeting and their pairwise correlations. These models provide very accurate, quantitative descriptions of higher order structure in these social networks. The parameters of these models seem poised close to critical surfaces in the space of possible models, and we observe scaling behavior of the data under coarse-graining. These results suggest that simple models, grounded in statistical physics, may provide a useful point of view on the larger data sets now emerging from complex social systems.

Keywords: critical phenomena of socio-economic systems, inference in socio-economic system, socio-economic networks

1742-5468/19/093406+29$33.00

Contents

1. Introduction

Social systems exhibit rich collective behaviors. Many large-scale social processes, from cultural fads [1] to residential segregation [2] to the polarization of political opinions [3], depend on the interactions of many individuals. Indeed, many social phenomena are emergent almost by definition. Sociologists have long explored the relationship between individual actions, interactions among individuals, and macroscopic social outcomes [4, 5].

While there is general agreement on the qualitative idea that social phenomena are emergent, there has been much less progress toward quantitative theories. For inanimate systems, especially near thermal equilibrium, statistical mechanics provides a framework for building quantitative theories of how macroscopic behaviors emerge from microscopic interactions. Importantly, successful theories in statistical mechanics often are simpler than the underlying microscopic reality, and the renormalization group (RG) allows us to understand how this simplification is possible [6, 7]. Inspired by these examples, there have been efforts to examine social phenomena using ideas and methods borrowed from statistical physics [8]. Examples include the dynamics of a strike [9], the emergence of group consensus [10, 11], the behavior of dancers at heavy metal concerts [12], and temporal patterns of activity and inactivity on Twitter [13].

In much previous work, methods from statistical physics were used to construct mathematically precise versions of existing sociological theories [14], but the resulting

J. Stat. Mech. (2019) 093406

models might or might not engage with quantitative data on real social systems. Here, inspired by a stream of work on biological systems ranging from families of proteins [15–18] to networks of neurons [19–22] to flocks of birds [23, 24], we take a different approach, using the maximum entropy method [25, 26] to build a statistical mechanics description of a social system directly from real data, independent of traditional sociological hypotheses. Previous efforts in this direction include analyses of voting patterns on the US Supreme Court [27, 28] and patterns of conflict in troops of macaques [29].

Here we adopt the strategy of building models directly from data, and explore the emergence of collective behaviors in Twitter communities. In order to carry out this program we need to identify communities and to define behavioral states for all the individuals in those communities. As a first step, we take these states to be participating or not participating in conversations that involve particular keywords. Then, states are binary and the maximum entropy models consistent with the pairwise correlations among these variables are equivalent to Ising spin glasses [30]. These relatively simple models successfully predict higher order structure in the data. Analysis of these models, as well as a direct coarse-graining of the data, suggests that these systems are close to a critical point or critical surface in their parameter space. We explore what this might mean for social functioning.

2. Networks and states

The full set of Twitter users is vast, beyond our ability to build explicit models. As a start, we want to focus on smaller networks of users who are well connected with one another. We start by choosing a single Twitter user and then find the people whom this user follows, and the people whom those neighbors follow. The result is a social network with known connectivity and relatively short path lengths. Twitter provides public access to the last 3200 tweets for each user, so each node in our network is associated with a stream of timestamped text. For an average user in our datasets, this covers two to three years of tweets.

We note that the initial choice of root users is arbitrary, and it is difficult to ask in what sense our results are representative (except by trying many examples). Happily, none of the individuals identified in this way were public figures, or otherwise strong outliers in terms of their social media presence.

Even the networks at depth two from a random user are quite large (see table A1 in appendix A), so we focus further, breaking these networks into sub-communities using the Clauset–Newman–Moore algorithm [31]. This algorithm builds sub-communities such that the proportion of edges within sub-communities is maximized while minimizing the number of edges between sub-communities. The resulting sub-communities are the networks that we use for further analysis. Among 106 examples we analyze sub-communities that contain between 8 and 80 people, and there do not appear to be any simple trends of topology versus size (see appendix A for these and other details). An example of the networks that we identify is in figure 1.

Having defined a network of individuals $i = 1, 2, \cdots, N$, what are the states σ_i taken on by these individuals? In examining the raw data, we find prominent words that

J. Stat. Mech. (2019) 093406

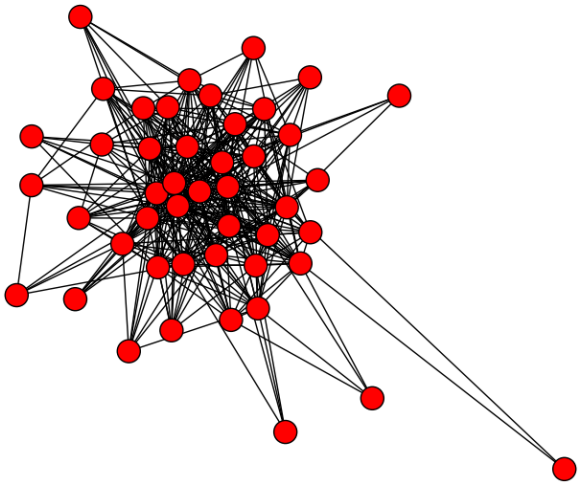

Figure 1. Example social network. An example social network used for building a pairwise max-ent model. Same sub-community as used in figure 4. We include more information on the topology of the social networks examined in appendix A.

are used many times within a short period of time (appendix B). For the remainder of this paper, we will call these prominent words 'keywords', and they can intuitively be thought of as something akin to a topic of conversation in the community being studied. For example, in a community with many physicists, one prominent keyword identified was 'Kosterlitz', as many people talked about Kosterlitz in a very short period of time after he shared the 2016 Nobel Prize. We have provided an example list of keywords in a supplementary dataset.

These keywords suggest a simple and intuitive way to binarize data from Twitter: either a given individual has used a given keyword, or they have not. So, in the physicists' example, we can find who talked about Kosterlitz and who did not, and assign everyone who did talk about Kosterlitz a state $\sigma_i = 1$ and everyone who did not talk about Kosterlitz a variable $\sigma_i = -1$. We can then conglomerate these variables into a vector $\boldsymbol{\sigma}$ representing whether or not everyone in the dataset used the given keyword. In this sense, the variable $\boldsymbol{\sigma}$ represents a social state of the community—there is an event happening in the world (represented by the keyword), and members of the community can either participate in this event or remain silent.

The succession of keywords, which by definition are each well localized in time, provide a series of snapshots of the social state $\boldsymbol{\sigma}$. These snapshots are drawn out of some distribution $P(\boldsymbol{\sigma})$ which characterizes the collective states taken on by the network as a whole. Our goal is to characterize this distribution.

There obviously are many ways to simplify or binarize data from Twitter. Even with the use of keywords as a tool for simplification, these keywords themselves could be chosen in different ways. While we cannot claim uniqueness, we do feel that our choice of simplification is intuitive and easy to implement (appendix B). Importantly, we will see that the states defined in this way have orderly behavior.

J. Stat. Mech. (2019) 093406

3. Maximum entropy models

The social states $\boldsymbol{\sigma}$ are the 'microscopic' states of our system, describing what each individual user is doing during a single conversation. In the spirit of statistical mechanics, we would like to write down the analog of the Boltzmann distribution, $P(\boldsymbol{\sigma})$, which tells us which social states are favored in a community and which states are disfavored. As usual, the number of possible states $\boldsymbol{\sigma}$ is so large (2^N) that we cannot directly 'measure' $P(\boldsymbol{\sigma})$ from any reasonable amount of data once we are looking at networks of reasonable size ($N \gg 10$). More precisely, the distribution $P(\boldsymbol{\sigma})$ is a list of length 2^N, constrained only by normalization, and so if there is no simpler underlying structure then we can not make any progress without making more than 2^N measurements.

How can we search, systematically, for simplifications of the distribution $P(\boldsymbol{\sigma})$? One idea is to take seriously some mean properties of the system that we can estimate, reliably, from our data, and start our search by insisting that our models match these empirical facts. It seems natural to ask, for example, that any reasonable model of a Twitter community match the mean probability $\langle\sigma_i\rangle$ that each user is active. We also suspect that correlations between pairs of users encode important information about collective behavior, so we propose that reasonable models should match these correlations $\langle\sigma_i\sigma_j\rangle$. It is not at all clear whether these low-order correlations are sufficient to describe the system, or whether we need more. Notice that if we continue down this path, asking that our model match higher and higher-order correlations, eventually we will have added so many constraints that we have specified the distribution. But this does not work, because a finite data set is not sufficient to determine arbitrarily high-order correlations with reasonable accuracy.

Even if we think that pairwise correlations are sufficient to characterize the collective behavior of the system, there are of course infinitely many distributions $P(\boldsymbol{\sigma})$ that are consistent with the measured $\langle\sigma_i\sigma_j\rangle$. Following a recent stream of work on the description of biological systems [15–24], we will choose the distribution that has the maximum entropy consistent with the measured correlations. This approach has its roots in the work of Jaynes [25, 26], but it is important to emphasize that searching for maximum entropy distributions does not require us to adopt any of the ideological positions articulated by Jaynes and his followers, positions which have been repeatedly criticized by the statistical physics community; for a recent example see [32]. We emphasize also that we do not know the details of the underlying dynamics—it seems unlikely, for example, that the states we have defined have a Markovian evolution—and hence we are not assuming that the maximum entropy model is the stationary distribution of some known stochastic process. We are interested in the maximum entropy model consistent with the means and pairwise correlations of user activity because this is well-motivated approximate model, one which matches some plausibly crucial features of the data exactly while discarding all other structure. The question of whether this model is a 'good model' of the underlying distribution is an empirical question, not a theoretical question, because we have no theoretical ground truth.

To be concrete we consider the maximum entropy model consistent with the mean activity of each user in the community, and with the correlations between pairs of users who are connected in the network. The resulting probability distribution is

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\mathrm{e}^{-E_{\boldsymbol{\sigma}}} \tag{1}$$

$$E_{\boldsymbol{\sigma}} = -\sum_i h_i\sigma_i - \frac{1}{2}\sum_{i\neq j} A_{ij}J_{ij}\sigma_i\sigma_j, \tag{2}$$

where h_i are parameters corresponding to constraints on one body marginals (equation (3)) and J_{ij} are parameters corresponding to constraints on two body marginals (equation (4)). We introduce the adjacency matrix A for the community to remind us that we have constraints only among connected pairs; $A_{ij} = 1$ if there is a social tie between individuals i and j and 0 otherwise. Thus, A represents a symmetrization of the underlying directed graph of following relationships on Twitter.

We find the values of $\{h_i, J_{ij}\}$ by solving

$$\langle\sigma_i\rangle_P \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\sigma_i = \langle\sigma_i\rangle_{\mathrm{obs}}. \tag{3}$$

$$\langle\sigma_i\sigma_j\rangle_P \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\sigma_i\sigma_j = \langle\sigma_i\sigma_j\rangle_{\mathrm{obs}}. \tag{4}$$

This is in general a difficult computational problem [33], but we have solved it numerically for communities of up to 80 people using Monte Carlo methods [34] (see appendix C for details). An example of a covariance matrix,

$$C_{ij} = \langle\sigma_i\sigma_j\rangle_{\mathrm{obs}} - \langle\sigma_i\rangle_{\mathrm{obs}}\langle\sigma_j\rangle_{\mathrm{obs}} \tag{5}$$

and the corresponding coupling matrix J_{ij} for a community is shown in figure 2.

A positive J_{ij} corresponds to a tendency for two users to use the same keywords and a negative J_{ij} corresponds to a tendency for users to avoid keywords used by another user. We find couplings of both signs (appendix D), as in spin glasses [30, 35], but it is important that couplings are not completely random, since they are determined by the observed correlations.

Maximum entropy models are appealing both because of their simplicity and because of their connections to statistical physics. But these are not arguments for their correctness. We could easily imagine, for example, that there are important multibody interactions among individuals, and in this case we could not give an accurate description of the joint distribution by matching pairwise correlations alone. To test these models we can compute higher order statistical quantities, and ask if these agree with the data. Importantly, once we have matched the one-body and two-body marginals, there are no free parameters left to adjust, so we are not 'fitting' these higher order structures—either we get them right or we get them wrong. We focus here on two such structures, the triplet correlations and the distribution of how many people tweet about each keyword.

For every distinct group of three users in a network, we can define the triplet correlation

$$C_{ijk} = \langle(\sigma_i - \langle\sigma_i\rangle)(\sigma_j - \langle\sigma_j\rangle)(\sigma_k - \langle\sigma_k\rangle)\rangle. \tag{6}$$

These correlations typically are quite small ($C \sim 0.01$) but can be estimated with fractional errors $\sim$10% given the sizes of our data sets (appendix E). In figure 3 we show

J. Stat. Mech. (2019) 093406

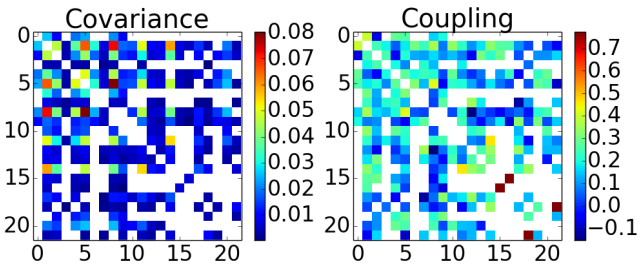

Figure 2. Covariance and coupling matrices. An example of the covariance matrix (equation (5)) and the corresponding coupling matrix J_{ij} (equation (2)) for a Twitter community. Same community as used in figures 3 and 4. Blank elements correspond to pairs of users without a direct social connection.

the comparison of predicted versus observed triplet correlations, in one sub-community of 22 users. Results for many other sub-communities are similar, with prediction errors on the same scale as our measurement errors, although in certain communities the pairwise model fails to capture some aspects of three point correlation structure (see figure E1 in appendix E for examples).

A different way of assessing higher order structure in the network is to ask about the fraction of users that participate in a conversation,

$$Q = \frac{1}{2} + \frac{1}{2N}\sum_{i=1}^{N}\sigma_i. \tag{7}$$

If the users tweet independently, then for large N we would see a Gaussian distribution of Q, and even for smaller communities the tail at large Q would be very restricted. We see in figure 4 that the observed distribution of Q is quite broad, with an extended tail, and that this is captured within error bars by our model. Thus, although we match only correlations among pairs of users, we can predict, quantitatively, the probability that many users will be active in the same conversation.

There are many reasons why a pairwise maximum entropy model should not work. In contrast to Jaynes, we do not view the maximum entropy principle as a way of making statistical inferences about the laws of this system, as these inferences are not guaranteed to be correct in nonequilibrium systems where the choice of canonical coordinates is unclear [32]. Thus in our system, there is no guarantee a maximum entropy model will produce accurate predictions. In particular, even in the absence of explicit combinatorial effects, averaging over many unseen factors that affect all the users, or different subsets of users, will generate effective multibody interactions in the joint distribution. These effects may be present, but what we see from figures 3 and 4 is that we do not need to make explicit models of these effects in order to generate quantitative predictions for the joint distribution of social behaviors. These models, while simple, are sufficiently precise that it makes sense to take them seriously as statistical physics problems and ask what we can learn about the collective behavior of the network.

J. Stat. Mech. (2019) 093406

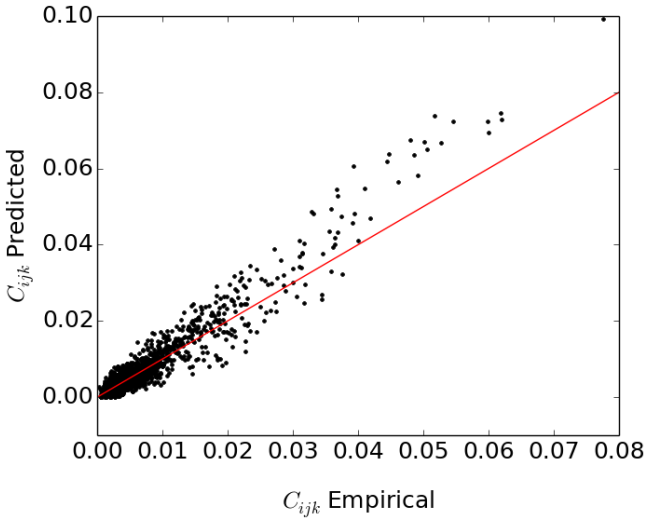

Figure 3. Triplet correlations. Predicted three point correlations from our model versus the empirical value of the three point correlations estimated from data, for one sub-community. Error bars typically are ∼10% of the measured values; see appendix E.

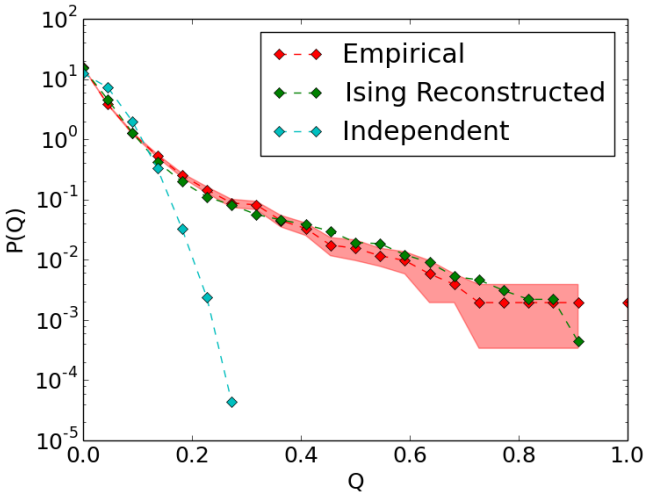

Figure 4. Distribution of simultaneous activity. Probability that a fraction Q of users tweet about a particular keyword. Empirical data (red), maximum entropy model (green), and a maximum entropy model including only one-body constraints (cyan).

4. Toward a phase diagram

A crucial lesson of statistical physics is that the parameter space of models for systems with many interacting degrees of freedom breaks up, at large N, into distinct phases with qualitatively different behaviors. The boundaries between these phases become sharp as $N \to \infty$, and on the boundaries the behavior of the system is a singular function of its parameters. Here we try to locate real networks of Twitter users in relation to these critical surfaces in parameter space.

https://doi.org/10.1088/1742-5468/ab3af0

J. Stat. Mech. (2019) 093406

Maximum entropy models have the form of a Boltzmann distribution, and so we can think about an 'energy landscape' as a function of the social state $\boldsymbol{\sigma}$; energy minima correspond to probability maxima, identifying states that are favored by the network. Every sub-community we have examined has the same dominant energy basin that contains the vast majority of the data. This basin is defined by silence in the sub-community ($\sigma_i = -1$ for all i). This in turn allows for an enormous simplification of our data, as we can define an order parameter by the overlap with the silent state [30]. The overlap with silence is just the negative of the usual magnetization, but it is important that we do not choose the magnetization arbitrarily.

Once we have an order parameter, we can define, in the usual way, a conjugate field and a susceptibility of the order parameter to this field; the susceptibility will be equal to the variance of the order parameter. Again this example is relatively simple—the conjugate field is a uniform 'magnetic field' Δh that biases each user to tweet or remain silent, and the susceptibility χ is proportional to the variance of the fractional activity Q defined above (equation (7)). In figure 5, we track χ versus Δh in a sub-community of 46 people (figure 1).

As we can see in figure 5, the system exhibits a large peak in susceptibility at a small forcing field. This is a collective effect, and would not be present if the users all tweeted independently (as shown in cyan). This peak in susceptibility is reminiscent of what we see at a critical point, where incremental changes in the control parameter lead to disproportionately large changes in observable behavior. Relative to the width of the peak, the system seems to be poised quite close to this near-critical point.

One may object that the language of 'applied fields' involves taking the mapping between maximum entropy models and their physical counterparts a bit too seriously. As an alternative, we can bias the system by choosing individuals out of the community and conditioning the distribution of all other users on these individuals being in the state $\sigma_\mu = 1$ (tweeting). Mathematically, if we hold $\sigma_\mu = 1$ for $\mu = \mu_1, \mu_2, \cdots$, then for all the remaining users the joint distribution still is given by equations (1) and (2), but with $h_i \to h_i + \sum_\mu J_{i\mu}$ [29]. An example susceptibility as a function of number of individuals forced is shown in figure 6 for the same sub-community as figure 5.

The fact that forcing some users to be active will bias the mean activity of other users is not surprising. More interesting is that the variance of total activity in the other users changes nonmonotonically as a function of the number of users that we force, echoing the dependence of susceptibility on applied field. In figure 6 we can see this peak occurs at 2–3 people being forced, out of a community of 46 people. In all the sub-communities that we have examined, the peak in variance occurs upon forcing just a handful of users, often just one; there is no indication that this depends systematically on N. We conclude that many Twitter communities are within one to three users of being near maximal variance in activity [36]. This is a direct but perhaps more intuitive analog of the peak in susceptibility for very small applied fields (figure 5).

If we have a statistical mechanics, then it should be possible to construct a thermodynamics. Much of thermodynamics is about the tradeoff between energy and entropy, and it might be unclear what this has to do with tweeting. But in the Boltzmann distribution, energy is just the (negative) log probability, and (microcanonical) entropy counts the number of states that have this probability. Intuitively, social states in

J. Stat. Mech. (2019) 093406

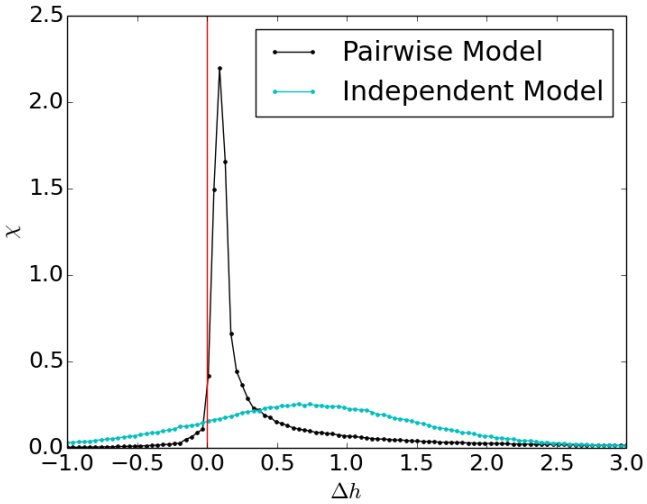

Figure 5. Susceptibility. Predicted susceptibility against a forcing field for the community of 46 people shown in figure 1. Pairwise maximum entropy model (black) and independent model (cyan); red line indicates $\Delta h = 0$, corresponding to the parameters inferred for the real network.

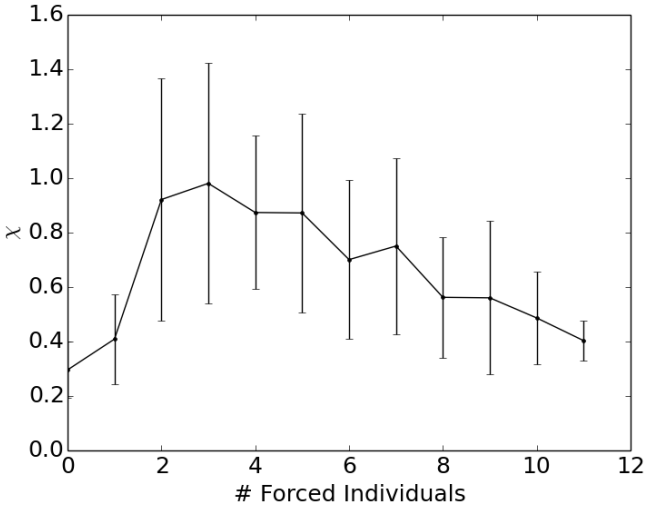

Figure 6. Forcing individuals. Susceptibility as a function of the number of forced individuals for the community of 46 people shown in figure 1. Error bars represent standard deviation over different configurations of individuals being forced.

which more users are active have lower probability (figure 4), but until fully half the users are active there are more distinct states available at larger Q (equation (7)). Thus, less probable (higher energy) states are more numerous (higher entropy). We explore this tradeoff between probability and numerosity following [21, 37, 38]; for details see appendix F.

The maximum entropy model assigns to each state $\boldsymbol{\sigma}$ an energy $E_{\boldsymbol{\sigma}}$, through equation (2). We would like to count the number of states that have a particular energy, or range of energies, but this involves making bins along the energy axis. A simple alternative is to count the number of states with energy less than E, so we define the entropy

J. Stat. Mech. (2019) 093406

$$S(E) = \ln\left(\sum_{\sigma} \Theta(E - E_{\sigma})\right) \tag{8}$$

where $\Theta(x)$ is the step function: $\Theta(x > 0) = 1$, $\Theta(x < 0) = 0$. We recall that the temperature is the derivative of the entropy with respect to energy. In our case we have $T = 1$, from equation (1), and so the condition

$$\frac{\mathrm{d}S(E)}{\mathrm{d}E} = 1 \tag{9}$$

picks out the typical energy of the system. The fluctuations around the typical energy are related to the heat capacity C,

$$\langle(\delta E)^2\rangle = \left(-\frac{\mathrm{d}^2 S}{\mathrm{d}E^2}\right)^{-1} = C \tag{10}$$

again with $T = 1$. We expect that energies and entropies both are extensive, that is proportional to system size N for large N, so that C itself also is of order N. Then the fractional fluctuations $\langle(\delta E)^2\rangle/E^2 \sim 1/N$ vanish rapidly for larger systems. At many critical points, $\mathrm{d}^2S/\mathrm{d}E^2$ vanishes, the specific heat C/N diverges with N, and the variance of energy fluctuations are similarly large.

Starting with our maximum entropy model, we can find entropy as a function of energy numerically using Wang–Landau sampling [39]. An example plot of S/N versus E/N is shown in figure 7 for a sub-community of 52 people. We see that the entropy is very nearly a linear function of energy across a wide range of energies near the typical value. It is not merely that $\mathrm{d}^2S/\mathrm{d}E^2$ vanishes at a single critical point, but it is very nearly zero all together. This unusual form of critical behavior was seen previously in the analysis of activity in networks of neurons [37].

5. Coarse-graining social data

The idea that social networks might be poised near a critical point is intriguing. Related notions of criticality have emerged from the analysis of neural networks, but this has also generated controversy. It is in principle possible that inference from finite data sets, using the maximum entropy framework, is biased toward finding models near criticality, or that some of the phenomenology which seems to be a signature of criticality could have more mundane explanations [40–43]. One response to these concerns is to look very closely and ask whether the alternatives to criticality really explain the data in detail, as discussed for a population of neurons in [21]. But the approach to a critical point seems so dramatic that we should be able to give a more direct argument.

In our modern view, a critical point can be defined as a nontrivial fixed point of the RG [6, 44]. In the standard formulation, microscopic variables live in real space, and have their dominant interactions with near neighbors. The RG involves averaging over spatial neighborhoods [45], and then tracking the distribution of these coarse-grained

J. Stat. Mech. (2019) 093406

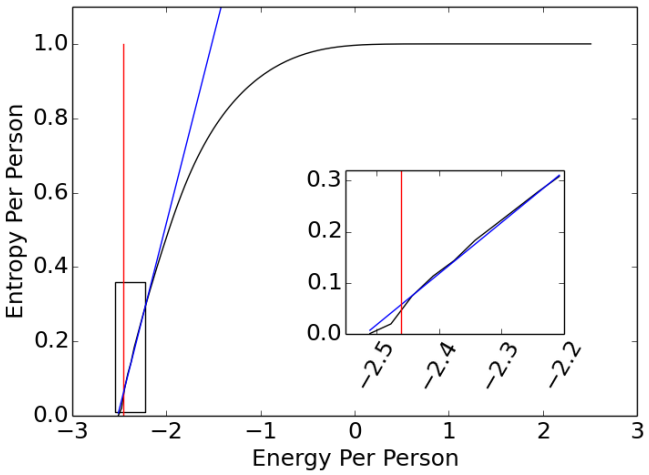

Figure 7. Entropy versus energy. $S(E)/N$ from equation (8), plotted versus E/N, for a sub-community of $N = 52$ users. Red line indicates average energy of system, blue line indicates line of slope 1 fit around the actual energy.

variables as a function of the averaging scale. A crucial result is that the joint distributions of coarse-grained variables become simpler as the scale becomes larger, so that most interactions are 'irrelevant'; models of macroscopic behavior thus are simpler and more universal than the underlying microscopic details. More subtly, there are special parameter settings such that the joint distribution of coarse-grained variables is invariant to the scale of averaging. This is a fixed point of the RG transformation, and these fixed points correspond to critical points.

In order to explore RG ideas in more complex systems we have to find coarse-graining strategies that do not lean on the locality of interactions. Recent work on large populations of neurons suggests that a natural analog of averaging with spatial neighbors is averaging with maximally correlated partners [46], and we follow this approach here. In brief, we walk through the network, identifying maximally correlated pairs $i,j_*(i)$, and then add the corresponding variables together,

$$\sigma_i^{(2)} = \sigma_i^{(1)} + \sigma_{j_*(i)}^{(1)}, \tag{11}$$

where superscripts refer to the level of coarse-graining. We then repeat this for the next most correlated pair of people and so on, until the original N variables have become $N/2$. If we iterate this full procedure k times, then we turn our data on N people into data on N/K clusters, each with $K = 2^k$ people in them. Though the underlying interactions have an irregular topology, we choose to coarse-grain with regular cluster sizes in order to yield a reproducible coarse-graining operation that can then be iterated multiple times.

If correlations are weak, the central limit theorem drives the activity of the clustered variables toward the normal distribution as the clustering scale increases. Near criticality, the self-similar structure of correlations under our coarse-graining operation should evade the central limit theorem, driving the distribution toward a non-Gaussian fixed point. In the same way that the central limit theorem predicts a linear scaling (for example) of the variance with the number of variables that are being summed, the

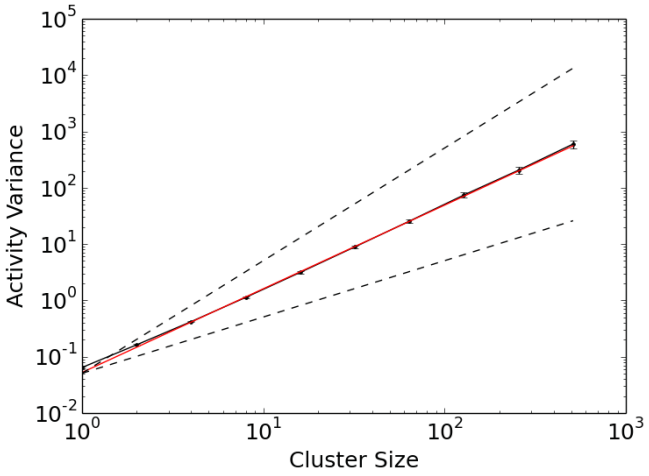

Figure 8. Variance scaling. The variance of $\sigma_i^{(k)}$ against cluster size $K = 2^k$. Dotted lines indicate linear and quadratic scaling; red line indicates a power-law fit with exponent $1.49 \pm .02$. Error bars are the standard deviation over random halves of the data. Fit obtained by nonlinear least squares with reduced $\chi^2 = 0.26$.

approach to nontrivial fixed points typically is associated with different scaling behavior. It is this scaling behavior that we are looking for in the data.

The means of the coarse-grained variables scale linearly with the cluster size K, so the first interesting question concerns the variance. For independent variables the variance will be linear in the cluster size K, while for perfectly correlated variables the growth would be quadratic. In figure 8 we show the behavior of the variance versus cluster size in a community of 583 people, and we see that the variance has near perfect scaling at an intermediate exponent $\sim$1.5. This intermediate scaling indicates that there is nontrivial structure to the correlations in the dataset, independent of the level of coarse-graining, and we note that this scaling is very precise across nearly three decades. When analyzing different communities, we find a range of scaling exponents between 1.29 and 1.69. We could not discern any clear pattern or clustering in the scaling of different communities. Although there is significant variability across communities, the precision of scaling within communities is surprising.

We can also look at the structure of correlations within clusters. In the analogy with physical critical points, we expect correlations to have the same structure at each stage of coarse-graining. In translation invariant systems with spatially local interactions, this corresponds to a scale-free correlation function, or equivalently a power-law dependence of the spectrum on momentum. In our systems, the closest analogue to this correlation function is the spectrum of the correlation matrix within clusters [46, 47]. In figure 9 we plot the averaged spectra of cluster correlation matrices, for different clusters sizes K, as a function of the fractional rank of the spectrum for the same community as used in figure 8. We stop at $K = 128$ to avoid contamination of the spectrum by finite sample effects.

As we can see in figure 9, the clusters seem to have a correlation spectrum that is independent of the size of the cluster. That is, plotted as a function of fractional rank, the spectra collapse onto a single curve that is independent of the degree of coarse-graining, with the exception of the largest eigenvalue of each cluster. Furthermore,

https://doi.org/10.1088/1742-5468/ab3af0

J. Stat. Mech. (2019) 093406

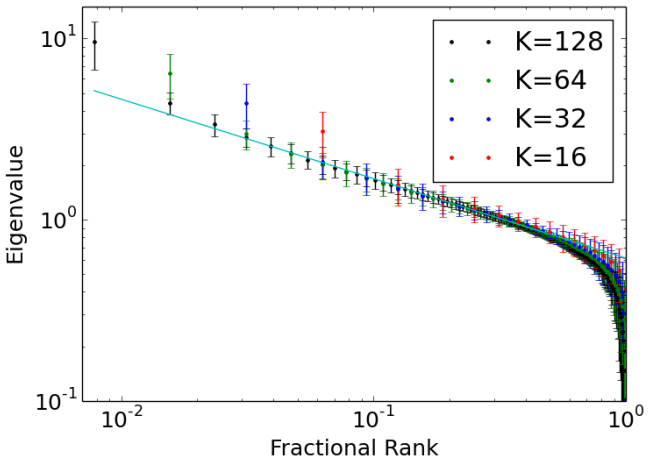

Figure 9. Correlation spectra. Correlation spectra of various cluster sizes plotted as a function of the fractional rank. Error bars are standard deviations across different clusters and random halves of the data. Cyan line is a power-law with exponent -0.438 ± 0.015.

while the limited dynamic range in figure 9 precludes identification of a power law, the data are consistent with scale-free behavior. Although noisy, even this largest eigenvalue seems to have a regular behavior as a function of cluster size. The correlations are scale free both in the sense that they are independent of the degree of coarse-graining and in the sense that they are consistent with a power law form.

6. Discussion

The dynamics of human behavior on social media are complex, and what we have done here is a first try. Nonetheless, we find it striking that the social states of these networks have relatively simple, orderly behavior that can be captured in the language of statistical physics. The joint distribution of activity in a community is described quite accurately by models that match only pairwise correlations, and are equivalent to familiar Ising models. More deeply, the parameters of these models seem not to be arbitrary, but rather are poised near critical surfaces, and we see independent evidence of this near-criticality in the scaling behavior of the system under coarse-graining.

It is an old idea that complex systems, far from equilibrium, might organize themselves to states that are analogous to critical states in equilibrium statistical physics [48]. One can think of many reasons why such an organization might be advantageous: the system becomes infinitely sensitive to (some) small signals, distant parts of the system can exchange information, the system grows long time scales, and more, although in different contexts the same features might be disadvantageous. Importantly, all of these features arise together at the critical point, and so it is necessarily difficult to disentangle which ones are actually functional.

There is an intuitive if non-rigorous connection between criticality and some familiar properties of social systems. Specifically, the high susceptibility and information transfer in critical networks evoke the tendency of online content to ‘go viral’, or to

J. Stat. Mech. (2019) 093406

very quickly become very prominent on a social network. Just as small perturbations become amplified in critical networks, a social phenomenon initiated by a small group of people can quickly become amplified in online social networks.

While it is tempting to suggest that the proximity of a critical point is the 'mechanism' by which things go viral on a social network, it is difficult to imagine a mechanism for social systems to tune themselves to this kind of critical point. Generically, any system capable of producing such complex behavior will likely be controlled by many different parameters, and critical dynamics will only hold in a relatively small section of this high-dimensional space. It is unclear how an online social system would naturally tune itself to this area of its parameter space. Nonetheless, this is what we see.

It is important to interpret these potential signs of criticality cautiously. In physical systems far from equilibrium, large fluctuations in the density can extend over a much longer range than is generally possible in equilibrium systems, by mechanisms which are not connected with criticality [49–51]. Although the system we study here cannot be described by the specific models where these effects arise, they provide cautionary examples. The Twitter community surely is not in equilibrium, and as such, we are unable to say whether the behavior we observe comes from true criticality or from some other mechanism. What we can say is that the data are represented faithfully by a model that is mathematically equivalent to an equilibrium statistical mechanics problem close to a critical point, and that aspects of the system's behavior exhibit scaling and an approach to a fixed distribution under coarse-graining.

As far as we know, neither the maximum entropy method nor the RG has been used previously in thinking about social networks. As the social science community accumulates more 'big data', more such tools will be needed. We hope to have made clear that these relatively simple ideas, grounded in statistical physics, are quite successful in revealing interesting regularities of human behavior in these social systems.

Acknowledgments

We thank Alex Aparicio, Curtis Callan, and Ned Wingreen for helpful comments and discussions. This work was supported in part by the National Science Foundation through the Center for the Physics of Biological Function (PHY–1734030), the Center for the Science of Information (CCF–0939370), and PHY–1607612.

Appendix A. Acquiring data

Raw data were scraped from publicly available tweets using the Twitter API (https://developer.twitter.com/). As described in the main text, for each dataset a root user was chosen and a social network was built out to second degree from that root. That is, we find who the root user follows, and then who those people follow; our examples of social networks are built from those connections. Characteristics of the datasets that we analyze are summarized in table A1, including information on the number of keywords or topics in each dataset, discussed further in appendix B.

J. Stat. Mech. (2019) 093406

Table A1. Dataset characteristics. Basic graph characteristics about the datasets analyzed here. Keywords are defined as those words used at least ten times with a standard deviation in time of less than 130 d.

Dataset name	# of people	# of edges	# of topics
A	927	69 280	7849
B	583	45 670	7030
C	646	72 870	9703
D	1184	205 834	8190
E	688	31 245	4159
F	261	5274	7940
G	575	17 422	3356
H	851	54 654	9298
I	498	15 798	6900
J	551	29 202	3374
K	99	2550	531
L	269	8788	1440
M	1128	47 762	9841
N	787	37 912	6731
O	94	582	382
P	651	24 000	11 329
Q	625	41 092	4923
Total	10 417	709 935	102 926

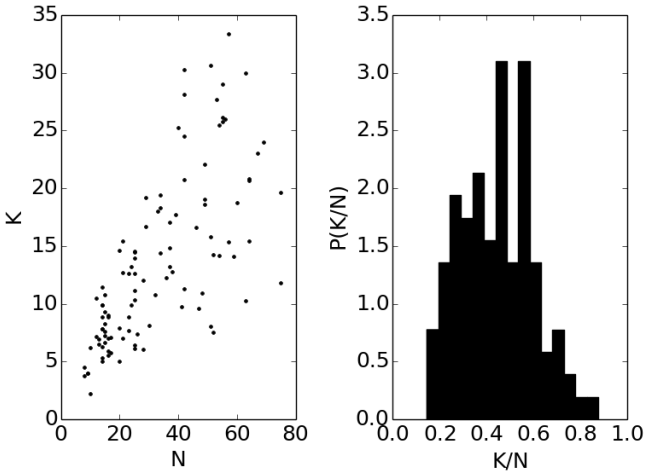

Figure A1. Topological characteristics. For 106 sub-communities with inferred pairwise max-ent models. (Left) Mean degree of social network K against community size N for all communities. (Right) Distribution of K/N.

These networks were then reduced, identifying sub-communities using the Clauset–Newman–Moore algorithm [31]. The resulting sub-communities contain between 8 and 80 people, and vary considerably in topology, as summarized in figure A1.

Mean degree (at left in figure A1) represents the typical number of social connections that a given individual in a sub-community has within that sub-community. Below a sub-community size $N \sim 40$ people, the mean degree K grows roughly linearly with the system size (Pearson correlation .74). For communities larger than 40 people, there is no discernible relationship between community size and the number of social

connections (Pearson correlation −.001). Interestingly, while one might imagine that these two types of social communities have different behaviors, in the communities that we have examined max-ent models are capable of describing both types of systems.

Our method of collecting data from Twitter differs from previous approaches, potentially with implications for our conclusions. Broadly speaking, there are two approaches to handling Twitter data—one that attempts to use a sample of global Twitter activity and one that focuses on local subgraphs. Global analyses often use some version of the 'firehose', or a live stream that contains some fraction of all tweets, either provided by Twitter directly or by a third party vendor [52, 53]. This leads to a very large amount of data, but is computationally intensive and is poorly suited for examining behavior in specific, small communities.

For our purposes we need to use a local approach to be able to model small sub-communities, but this necessitates an arbitrary choice of where to look in the larger Twitter network. There are many ways of defining such local sub-communities, but our method is most closely aligned with [54], which attempts to find communities with shared interests by examining the Twitter users that follow all of a set of celebrities identified with that interest and then examining the topology of social links among those users. Like our method, the authors in [54] find a group of users centered around a pre-defined central user or set of users, but our method differs in that (1) our central user need not be a celebrity or anyone of import in the community and (2) going to follower depth two gives us larger datasets and communities. This allows us to both explore a wider variety of Twitter communities while maintaining community sizes capable of interesting collective behavior.

Given the relatively small size of the communities we examine we cannot claim that we are meaningfully sampling the entire Twitter network. As such, it is important to interpret our results in the context of the behavior of local social dynamics in small communities. It is also unclear to us how our particular method of selecting communities influences the results presented here, and it is possible that other ways of defining communities might yield different results.

Appendix B. Defining keywords

We define keywords to be words that are used many times while staying localized in a short period of time. Of course, we then must define how many times a word must be used and how localized a word must be to be considered a keyword. We use the standard deviation in the times that a word is used to quantify the degree to which a word is localized in time. These two criteria (number of times used, standard deviation in time) define a two dimensional space in which we can place each word that is used in a dataset. Our task is to find cutoffs in this space to define a clear set of keywords. Unfortunately, this is hard. All datasets examined are approximately Zipfian [55], which means that there is no clear scale for usage, making a non-arbitrary cutoff for the number of times a word is used quite difficult.

The distribution of standard deviation in time for words is more interesting. We show the distribution of the standard deviation in time for language used in a dataset of 651 people in figure B1. As we can see in figure B1, this distribution has two peaks, one corresponding to words that are used in a very short amount of time (far left) and

J. Stat. Mech. (2019) 093406

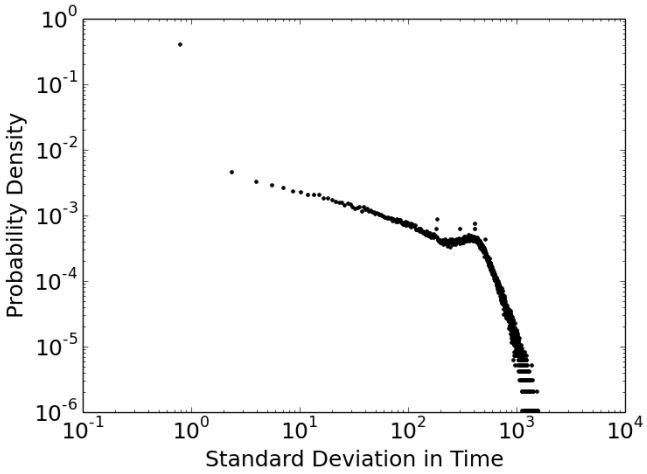

Figure B1. Distribution of standard deviations in time of word usage. For a dataset of 651 people. X-axis is in units of days.

words that are used with a standard deviation in time of around 500 d. The second peak begins with a kink in the distribution at a standard deviation of around 200 d. This second peak corresponds to words that are used independent of context, such as the staples of standard English vocabulary ('the', 'she', etc). Obviously, we do not want to include such words as keywords, as they are independent of context, and what makes keywords interesting is that they are highly contextual. We therefore bound the cutoff in standard deviation in time to be well clear of this second peak in figure B1.

The statistics of word usage do not provide more detailed guidance for how to define keywords. As such, we examined the data at a range of different definitions for keywords [36], and generally found that the qualitative nature of results did not change with different definitions of keywords. For the purposes of presenting this data, we choose a definition for keywords that attempts to maximize the number of clearly meaningful data points. The exact parameters used for the datasets presented here are a cutoff in standard deviation in time of 130 d and a requirement that the word must be used at least 10 times in the data. In our datasets, this generally is sufficient to obtain on the order of 10 times the number of keywords as there are people in the dataset, while still yielding generally comprehensible keywords. The number of keywords generated by this procedure for each dataset is shown in the right column of table A1. We have provided a text file with example keywords from community P in table A1.

Our method of defining keywords differs from how past work on Twitter has analyzed the function of online communities, but we believe our approach offers unique conveniences for our methods of data analysis. Much of past literature on collective behavior on Twitter can be classified by the type of social interaction that the authors analyze. Twitter offers many different ways that users can interact, and these different types of interactions presumably lead to different social properties. Past quantitative work on social interaction on Twitter has analyzed directed tweets (when one user tweets at another specific user) [56], retweets (when one user amplifies a tweet from another user) [57, 58], shared use of hashtags (a tool to tag tweets in a given category) [59–62] or spread of topics as inferred from the text of tweets [63–65]. Analysis of hashtags and more traditional topic models from natural language processing are the

J. Stat. Mech. (2019) 093406

techniques most closely aligned with our goals in this paper, and therefore it is worth contrasting our approach with these alternatives.

We chose to focus on individual words instead of hashtags in our analysis of Twitter data as this gave us access to a much larger pool of data. The large majority of tweets have no hashtags [66], which would mean that focusing only on hashtag usage would discard a large amount of information about the communities we are studying.

Beyond embodying intuition about the topics of Twitter conversations, our approach to identifying keywords provides a natural binary state for each user, who either tweets that keyword or does not. Many other topic models, such as Latent Dirichlet Allocation, model texts as probabilistic mixtures of various topics rather than binary classifications of whether or not a topic is relevant [67], making it difficult to define binary states of the users.

It is also worth emphasizing what our definition of keywords neglects to capture about Twitter activity. Perhaps most importantly, we do not take into account the temporal order in which tweets are sent. This means our data does not distinguish whether person A tweets word X before or after person B. Temporal order of tweets is clearly important, for example when considering causality or influence between users, and techniques in network theory have been developed to contextualize the temporal interactions between nodes [68]. Additionally, there is compelling empirical work on Twitter communities indicating that temporal patterns of social interactions on Twitter have meaningful effects. Repeated exposure to hashtags has been shown to increase their probability of adoption [59], and popular hashtags have different dynamics than less popular hashtags [69]. In short, it is clear that temporal patterns matter on Twitter.

We choose not to include temporal ordering in our analysis not because it is uninteresting, but because we wanted to start with something simpler. We have the ambitious goal of describing the full joint distribution of user activity in a community. If we wanted to achieve the same goal in a dynamic setting, we would need to construct the full joint distribution of events in time, which is vastly more complex, and it is likely that this complexity could be tamed only by strong modeling assumptions, e.g. that some aspect of time evolution is Markovian. By focusing on states of the community in single epochs, rather than their sequence, we avoid all these assumptions and can be driven by the data.

Appendix C. Maximum entropy models

The maximum entropy method has a long history, and has received new attention in efforts to build statistical mechanics descriptions of biological networks directly from data. Much of what we need thus is well known in some communities. To make the discussion accessible to a wider community, we provide some review of these ideas here. We start with general ideas and proceed to the specifics of our problem.

We recall that entropy, in addition to its thermodynamic meaning, provides the unique measure of available information consistent with simple and plausible conditions [70]. Distributions with larger entropy thus describe variables about which we know less, *a priori*. Maximizing the entropy is then a strategy for building models that

J. Stat. Mech. (2019) 093406

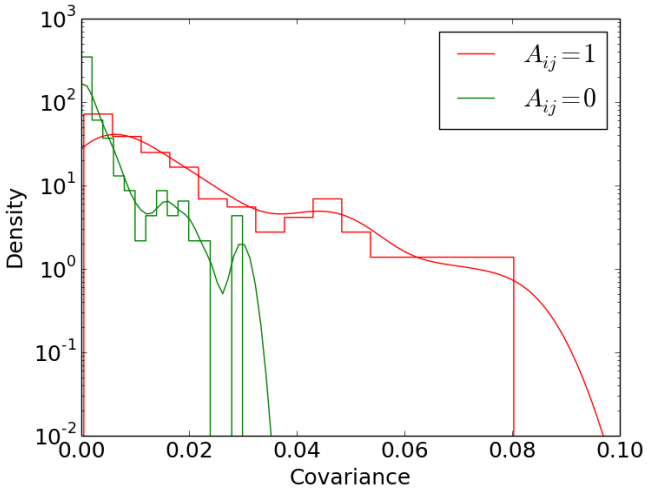

Figure C1. Distribution of covariance matrix elements for connected and disconnected pairs. (red) Distribution of C_{ij} across connected pairs ($A_{ij} = 1$). (green) Distribution of C_{ij} across disconnected pairs ($A_{ij} = 0$). Square lines are histograms, smooth lines are Gaussian Kernel density estimates of the underlying continuous distribution with bandwidth determined by Scott's rule [74]. Data are from the same community as in figure 3.

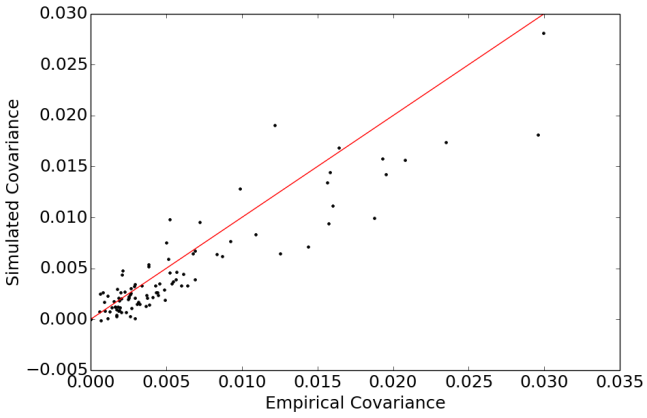

Figure C2. Covariance prediction for $A_{ij} = 0$. For the same community as shown in figure 3. Red line indicates equality.

inject as little structure or knowledge as possible, as first emphasized by Jaynes [25, 26]. Specifically, we want to insist that our models match certain features of the data, but otherwise have as little structure as possible.

In a system with states $\boldsymbol{\sigma}$, we can construct features $f_\mu(\boldsymbol{\sigma})$, for $\mu = 1, 2, \cdots, K$. Then if we are trying to make a model of the probability distribution $P(\boldsymbol{\sigma})$, we insist that the average of these features in the model matches those seen experimentally,

$$\langle f_\mu(\boldsymbol{\sigma})\rangle_{\text{expt}} = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\, f_\mu(\boldsymbol{\sigma}), \tag{C.1}$$

J. Stat. Mech. (2019) 093406

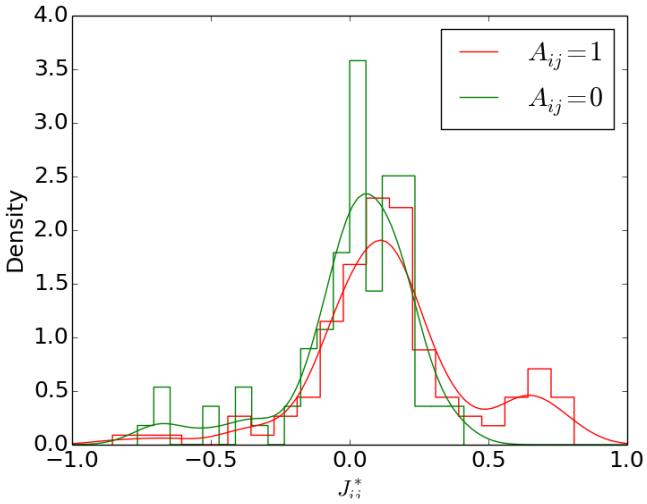

Figure C3. Couplings for $A_{ij}=0$ and $A_{ij}=1$ for a fully connected model. For the same community as shown in figure 3. Maximum entropy model inferred with $A_{ij}=1$ for all pairs, then sorted by actual A_{ij} in the community. Square lines are histograms, smooth lines are Gaussian kernel density estimates of the underlying continuous distribution with bandwidth determined by Scott's rule [74]. Data are from the same community as in figure 3.

for each μ. Among the distributions that obey this matching condition, we want to choose the one that maximizes the entropy

$$S[P(\boldsymbol{\sigma})] = -\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log P(\boldsymbol{\sigma}). \tag{C.2}$$

To solve the constrained maximization problem we introduce Lagrange multipliers and define

$$\tilde{S}[P(\boldsymbol{\sigma})] = -\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log P(\boldsymbol{\sigma}) + \sum_{\mu} \lambda_\mu \left(\langle f_\mu(\boldsymbol{\sigma}) \rangle_{\text{expt}} - \sum_{\boldsymbol{\sigma}} f_\mu(\boldsymbol{\sigma}) P(\boldsymbol{\sigma}) \right) + \lambda_0 \left(1 - \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \right), \tag{C.3}$$

where the Lagrange multipliers λ_μ correspond to each constrained feature and the term proportional to λ_0 constrains the distribution to be normalized. Now we can search over all distributions, and adjust the values of the Lagrange multipliers at the end to be sure that the constraints are satisfied.

To maximize $\tilde{S}[P(\boldsymbol{\sigma})]$, as usual we take the derivative and set it to zero,

$$\frac{\partial \tilde{S}[P(\boldsymbol{\sigma})]}{\partial P(\boldsymbol{\sigma})} = 0; \tag{C.4}$$

we can verify that the second derivatives are negative so that we really are finding a maximum of the entropy. The solution is

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(-\sum_{\mu=1}^{N} \lambda_\mu f_\mu(\boldsymbol{\sigma}) \right), \tag{C.5}$$

J. Stat. Mech. (2019) 093406

where the partition function Z absorbs λ_0 and thus depends implicitly on the values of all the other Lagrange multipliers,

$$Z = \sum_{\boldsymbol{\sigma}} \exp\left(-\sum_{\mu=1}^{N} \lambda_\mu f_\mu(\boldsymbol{\sigma})\right). \tag{C.6}$$

While equation (C.5) gives the correct form of the max-ent model, it glosses over a major sticking point, which is that the correct values of the Lagrange multipliers λ_μ must be determined. In general this is a difficult computational problem [33], and we use a method based on Monte Carlo sampling. Briefly, we simulate the model with some set of parameters $\{\lambda_\mu\}$ and then examine the expectation values of the observables f_μ, computed as averages over the Monte Carlo samples in the simulation epoch labelled by t. We then adjust the parameters by a factor proportional to the error [20, 22], so that the basic learning step is

$$\lambda_\mu(t+1) = \lambda_\mu(t) - \eta\left[\langle f_\mu\rangle_t - \langle f_\mu\rangle_{\text{expt}}\right], \tag{C.7}$$

where η is a learning rate. For discussions about convergence see [71, 72]. We iterate until the differences between model and observed expectation values are comparable to the errors in the observed expectation values.

The expensive part of this procedure is generating new Monte Carlo samples for each set of parameters. To speed up this process we use the histogram Monte Carlo method, which allows us to recycle Monte Carlo samples generated with parameters $\boldsymbol{\lambda}$ to estimate features of the distribution parameterized by a different set of parameters $\boldsymbol{\lambda}'$ [34, 73].

In our particular case, the features that we choose are $f_\mu = \sigma_i$ and $f_\mu = \sigma_i\sigma_j$, for all values of i and j that share a social tie. Constraining the expectation values of these features corresponds to fixing the probability that individual i participates in a Twitter conversation, and the correlations between participation by individuals i and j. With these choices, it is convenient to think of the Lagrange multipliers as 'effective fields' h_i and 'couplings' J_{ij}, and we arrive at the form of the model shown in equations (1) and (2) of the main text. As indicated above, we need to follow the Monte Carlo procedure to arrive at values of these parameters given our measurements of $\langle\sigma_i\rangle$ and $\langle\sigma_i\sigma_j\rangle$.

Although we are not trying to describe the dynamics by which communities move through sequences of states, these dynamics generate correlations that limit the number of independent samples and hence the statistical reliability of our analysis. We can estimate the effective number of independent data points by examining how the errors in the averages of our binary variables σ_i scale with the total number of included data points [22]. Briefly, for a variable σ_i we expect the squared error on the estimate of the average of σ_i to scale as $\delta^2_{\langle\sigma_i\rangle} = \text{Var}[\sigma_i]/n$ if we have n independent samples. This number of independent samples should be proportional to the raw number of keywords included, so that by looking at different subsets of the data we can find the proportionality constant. Perhaps surprisingly, we find that we have approximately as many independent samples as nominal samples, indicating that correlations from one Twitter conversation to the next are weak.

J. Stat. Mech. (2019) 093406

In building the maximum entropy model, we chose to constrain pairwise correlations only between users that have a social connection. In figure C1 we check that this makes sense, comparing the distribution of covariance matrix elements $C_{ij} = \langle \sigma_i \sigma_j \rangle$ for connected ($A_{ij} = 1$) and unconnected ($A_{ij} = 0$) pairs. We see, as expected, that correlations between people who share a social tie (red) are much stronger than correlations between people who do not share a social tie (green), and that highly correlated pairs of individuals are always socially connected in this community. This simple pattern would not appear if the keywords that we identify did not have some social content.

A more subtle question is whether knowing the correlations between connected individuals is sufficient to predict the correlations between unconnected individuals. Our maximum entropy model, which matches the correlations for pairs with $A_{ij} = 1$, makes predictions for *all* elements of the covariance matrix C_{ij}. In figure C2 we test these predictions, and find good though not perfect agreement. The Pearson correlation between predicted and observed values is 0.95, which is quite high; again we emphasize that there are no free parameters that can be adjusted to improve this prediction.

Once we have inferred the maximum entropy model, it is reasonable to ask whether the parameters of this model can be related back to measurable features of the system. Specifically, can we correlate strong couplings with some measurable feature of the systems? A similar question arises in studies that attempt to use a fully connected max-ent model in order to infer an underlying interaction topology based on which couplings are large. For example, in applications of max-ent modeling to protein structure determination, residues with large couplings are predicted to be in direct contact in the protein's three dimensional structure [16]. We can ask an analogous question in our system: if we infer a fully connected model (that is, assume that $A_{ij} = 1$ for all pairs $i \neq j$), can we recover the known social network topology?

In figure C3 we see the distribution of couplings J^*_{ij} inferred from a fully connected model. The distributions of couplings for people who are in fact socially connected (red) and those who are not (green) overlap significantly, indicating that coupling strength alone would not be a reliable indicator to recover the social network topology. However, we can also see in figure C3 that all strongly positive couplings correspond to pairs of users that in fact share a social tie. That is, a fully connected model is not capable of identifying every following relationship in the community, but it is capable of picking out strong relationships that can only exist among people who follow one another.

It is useful to think about a null model in which we constrain only the probabilities that individuals participate in a Twitter conversation, but ignore correlations. Then the maximum entropy model has the same form as in equations (1) and (2), but with all $J_{ij} = 0$ and

$$h_i = \operatorname{arctanh}(\langle \sigma_i \rangle). \tag{C.8}$$

This model is also equivalent to the assumption that each individual makes independent decisions about whether to tweet.

Appendix D. Energy landscapes

The model laid out in equations (1) and (2) is similar to canonical models of spin glasses. Spin glasses tend to have many minima in their energy landscape due to frustration

J. Stat. Mech. (2019) 093406

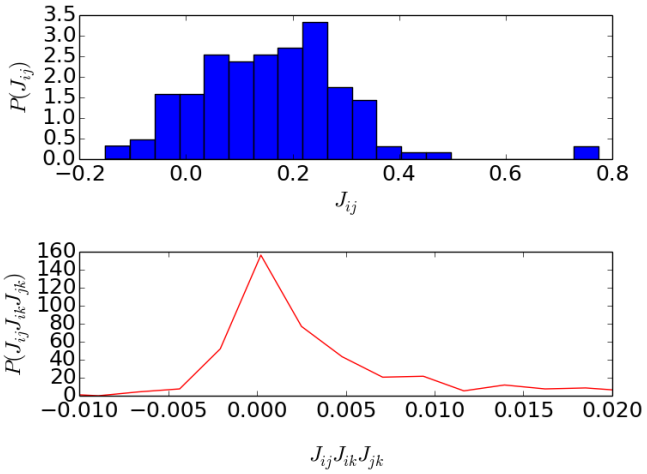

Figure D1. Characteristics of coupling matrix. (Top) Distribution of couplings J_{ij} for the community in figure 2. (Bottom) Distribution of the product of all interactions from closed social triangles in the same community. Approximately 30% of all possible triangles are frustrated.

[30], which occurs when there are couplings of mixed sign in the Hamiltonian, competing with one another. This in turn leads to many metastable states in the energy landscape.

We can assess the complexity of the energy landscape by measuring how frequently frustration occurs. We examine this by looking at the distribution of the product of all coupling terms representing a social triangle (where three people all have social ties to one another) in the communities. When the product of these interaction terms is positive, there exists a social configuration that can minimize all relevant terms in the Hamiltonian. When this product is negative, then no state can minimize all relevant terms and the system is frustrated. We show an example of the distribution of couplings J_{ij} and the distribution of the product of couplings from social triangles in figure D1. As we can see, there are a significant number of frustrated triangles. This is true for all the communities that we examined.

We can also evaluate the complexity of the energy landscape by directly estimating the number of metastable states in the energy landscape, which we do by moving 'downhill' in energy from each of 10^5 Monte Carlo samples. We can then see how the number of metastable states scales with the size of the system. We show this relationship in figure D2 for all 106 communities we examined. Over the range that we can observe, the number of metastable states increases roughly exponentially with system size, albeit with considerable variation from instance to instance. If this pattern persists into the thermodynamic limit it would put the energy landscape of these systems into the very complex class identified for the mean-field spin glass [30].

J. Stat. Mech. (2019) 093406

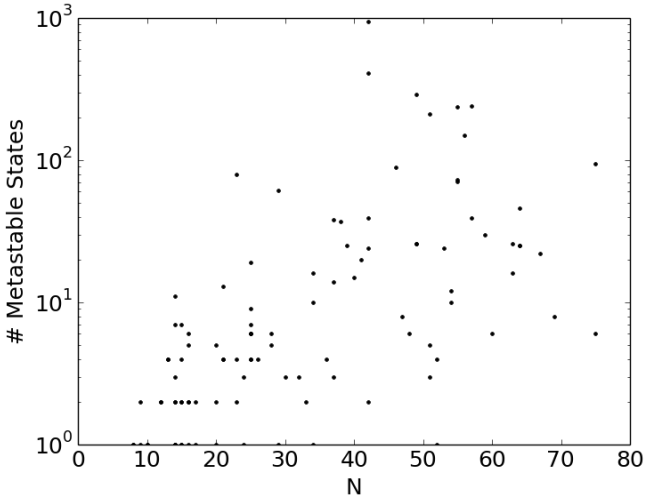

Figure D2. Number of metastable states, estimated as described in the text, for each of the communities that we analyzed.

Appendix E. Accuracy of three point correlations

In figure 3 of the main text, each three point correlation C_{ijk} carries an empirical uncertainty that can be estimated by bootstrap. The shear quantity of possible three point correlations in a heterogenous system makes it difficult to evaluate the accuracy of our predictions, so here we bin the three point correlations by their empirical values and then compare the root mean square error in prediction by the pairwise max-ent model to the root mean square uncertainty in the data. In figure E1 we show two examples of these errors. Data from the community on the left is shown in figure 3 of the main text.

For the community on the left of figure E1, the prediction error is of the same order of magnitude as the measurement error, indicating that the pairwise max-ent model is able to capture three point correlations almost as well as the data allow. The bottom of figure E1 shows the cumulative distribution of three point correlations. As we can see, the accuracy in the top of figure E1 covers the bulk of this distribution. However, for the community shown on the right of figure E1, the prediction error from the maximum entropy model is significantly larger than the intrinsic error in the data. In short, for the community on the right, the pairwise maximum entropy model is not capable of accurately reproducing 3 point correlations to the precision that the data allows. This does not mean that the maximum entropy model is incapable of making useful predictions on the three point correlations. Indeed, for the community on the right, the Pearson correlation coefficient between the empirical and predicted values of C_{ijk} is 0.93, which would normally be viewed as a success, even if we do not reach the maximum possible accuracy.

In the bottom subfigures of figure E1, we can see that essentially all three point correlations are positive. We have noted this to be the case across many communities, but we lack a convincing explanation for why this is the case.

J. Stat. Mech. (2019) 093406

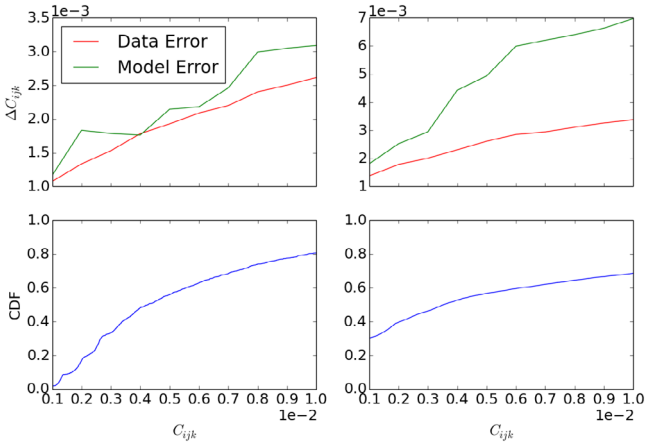

Figure E1. C_{ijk} prediction error. Top shows the root mean square error for predictions of three point correlations from the pairwise max-ent model (green) as well as the root mean square uncertainty for empirical three point correlations (red). Data were binned to compute root mean square errors. Bottom shows cumulative distribution of empirical three point correlation values. Left and right represent data from two different communities (data from left shown in figure 3).

It is unclear what determines how well a pairwise model is capable of fitting data from a given community, and we suspect that variation between communities could be a fruitful area of study.

Appendix F. Thermodynamics redux

If we can take our models seriously, then as the networks we study become larger the description in terms of statistical mechanics should imply an analog of thermodynamics. We follow [21, 37, 38] in this construction, and for completeness we recall the arguments presented there.

The essential step is to write the partition function not as a sum over states but as an integral over the density of states,

$$Z = \sum_{\boldsymbol{\sigma}} \mathrm{e}^{-E} = \int \mathrm{d}E \mathrm{e}^{-E} \rho(E), \tag{F.1}$$

where

$$\rho(E) = \sum_{\boldsymbol{\sigma}} \delta(E - E_{\boldsymbol{\sigma}}). \tag{F.2}$$

We can integrate by parts to yield an expression in terms of the cumulative density of states, and from there the micro-canonical entropy (equation (8)),

$$Z = \int \mathrm{d}E \mathrm{e}^{S(E)-E}. \tag{F.3}$$

We then scale the energy per node $\epsilon = E/N$ and the entropy per node $s = S(E)/N$, which gives us

J. Stat. Mech. (2019) 093406

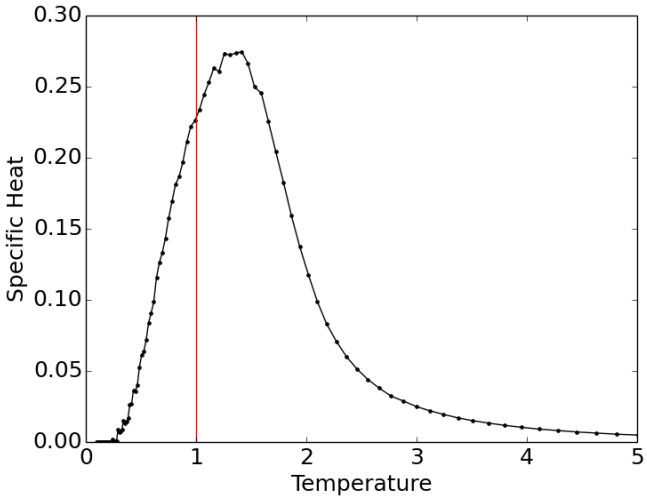

Figure F1. Heat capacity. Heat capacity (equation (F.8)) as a function of a temperature coupled to the inferred max-ent Hamiltonian. Red line indicates actual temperature ($T = 1$).

$$Z = N \int \mathrm{d}\epsilon \mathrm{e}^{N(s(\epsilon)-\epsilon)} = \int \mathrm{d}E \mathrm{e}^{-Nf(\epsilon)}, \tag{F.4}$$

where $f(\epsilon) = \epsilon - s(\epsilon)$, is the free energy per particle. The claim that

$$\lim_{N\to\infty} S(E)/N = s(\epsilon) \tag{F.5}$$

is the claim that a thermodynamic limit exists, which is far from obvious. But if it does exist, we can continue.

If we take the limit $N \to \infty$, we enter into the domain of Laplace's approximation, where Z will be dominated by the minima of f. These minima are given by the solutions ϵ^* such that:

$$\left.\frac{\mathrm{d}f}{\mathrm{d}\epsilon}\right|_{\epsilon^*} = 0 \Longrightarrow \left.\frac{\mathrm{d}s}{\mathrm{d}\epsilon}\right|_{\epsilon^*} = 1. \tag{F.6}$$

This is true for all systems, and is another way of defining temperature [37, 75], which we have set to be 1 in our discussion. Expanding to second order (as the first derivative disappears at the minima), we have that:

$$Z \approx N\mathrm{e}^{-Nf(\epsilon^*)} \int \mathrm{d}\epsilon \exp\left(\frac{N}{2}(\epsilon-\epsilon^*)^2 \left.\frac{\mathrm{d}^2 s(\epsilon)}{\mathrm{d}\epsilon^2}\right|_{\epsilon^*}\right). \tag{F.7}$$

In this equation, it seems that the term outside the integral provides a contribution from the typical energy ϵ^*, while the term inside the integral bounds the deviations from that typical energy. Crucially, the size of these deviations is controlled by the second derivative of the entropy with respect to the energy. When that second derivative is small, deviations from the typical energy will be large. This is a critical point.

The connection between the microcanonical entropy and the deviations from a system's typical energy is realized in the heat capacity, which can be expressed both in

J. Stat. Mech. (2019) 093406

terms of the second derivative of the microcanonical entropy or in terms of the variance of the energy:

$$C_T = N\left(-\frac{\mathrm{d}^2 S}{\mathrm{d}E^2}\right)^{-1} = \sigma_E^2. \tag{F.8}$$

Where σ_E^2 is the variance of the energy.

A nearly linear entropy should correspond to a large value of the heat capacity. We can see this in figure F1, where we simulate the system shown in figure 5 with a Hamiltonian scaled by various fictitious temperatures $H \to H/T$. In figure F1, the real system (at $T = 1$) is slightly on the low temperature side of the peak in the heat capacity. A peak in the heat capacity is another typical sign of criticality in physical systems, and it should increase our confidence that the systems examined here are near a critical point.

References

[1] Meyerson R and Katz E 1957 *Am. J. Sociol.* **62** 594
[2] Schelling T 1971 *J. Math. Sociol.* **1** 143
[3] McCarty N, Poole K and Rosenthal H 2012 *Polarized America: the Dance of Ideology and Unequal Riches* 2nd edn (Cambridge, MA: MIT Press)
[4] Brodbeck M 1958 *Phil. Sci.* **25** 1
[5] Sawyer R K 2001 *Am. J. Sociol.* **107** 551
[6] Sethna J 2006 *Statistical Mechanics: Entropy, Order Parameters and Complexity* (Oxford: Oxford University Press)
[7] Wilson K 1979 *Sci. Am.* **241** 158
[8] Castellano C, Fortunato S and Loreto V 2009 *Rev. Mod. Phys.* **81** 591
[9] Galam S, Gefen Y and Shapir Y 1982 *Math. Sociol.* **9** 1
[10] Weron K S 2005 (arXiv:physics/0503239 [physics.soc-ph])
[11] Dornic I, Chate H, Chave J and Hinrichsen H 2001 *Phys. Rev. Lett.* **87** 045701
[12] Silverberg J, Bierbaum M, Sethna J and Cohen I 2013 *Phys. Rev. Lett.* **110** 228701
[13] Duh A, Rupnik M S and Korosak D 2018 *Big Data* **6** 112
[14] Bruch E and Atwell J 2015 *Sociol. Methods Res.* **44** 186
[15] Bialek W and Ranganathan R 2007 (arXiv.org.org:0712.4397 [q–bio.QM])
[16] Weigt M, White R, Szurmant H, Hoch J and Hwa T 2009 *Proc. Natl Acad. Sci. USA* **106** 67
[17] Mora T, Walczak A, Bialek W and Callan C 2010 *Proc. Natl Acad. Sci. USA* **107** 5405
[18] Marks D, Colwell L, Sheridan R, Hopf T, Pagnani A, Zecchina R and Sander C 2011 *PLoS One* **6** 28766
[19] Schneidman E, Berry M, Segev R and Bialek W 2006 *Nature* **440** 1007
[20] Tkačik G, Schneidman E, Berry M and Bialek W 2006 (arXiv:q–bio.NC/0611072)
[21] Tkačik G, Mora T, Marre O, Amodei D, Palmer S, Berry M and Bialek W 2015 *Proc. Natl Acad. Sci. USA* **112** 11508
[22] Tkačik G, Marre O, Amodei D, Schneidman E, Bialek W and Berry M 2014 *PLoS Comput. Biol.* **10** 1003408
[23] Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M and Walczak A M 2012 *Proc. Natl Acad. Sci. USA* **109** 4786
[24] Bialek W, Cavagna A, Giardina I, Mora T, Pohl O, Silvestri E, Viale M and Walczak A M 2014 *Proc. Natl Acad. Sci. USA* **111** 7212
[25] Jaynes E 1957 *Phys. Rev.* **106** 620
[26] Jaynes E 1957 *Phys. Rev.* **108** 171
[27] Lee E D, Broedersz C and Bialek W 2015 *J. Stat. Phys.* **169** 275
[28] Lee E D 2018 *J. Stat. Phys.* **173** 1722
[29] Daniels B, Krakauer D and Flack J 2017 *Nat. Commun.* **8** 14301
[30] Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[31] Clauset A, Newman M and Moore C 2004 *Phys. Rev.* E **70** 066111

J. Stat. Mech. (2019) 093406

[32] Auletta G, Rondoni L and Vulpiani A 2017 *Eur. Phys.* J **226** 2327
[33] Nguyen H C, Zecchina R and Berg J 2017 *Adv. Phys.* **66** 197
[34] Broderick T, Dudik M, Tkačik G, Schapire R and Bialek W 2007 (arXiv:0712.2437 [q–bio.QM])
[35] Edwards S F and Anderson P W 1975 *J. Phys. F: Met. Phys.* **5** 965
[36] Hall G 2018 *Senior thesis* Princeton University Department of Physics
[37] Mora T and Bialek W 2011 *J. Stat. Phys.* **144** 268
[38] Stephens G, Mora T, Tkačik G and Bialek W 2013 *Phys. Rev. Lett.* **110** 018701
[39] Wang F and Landau D 2001 *Phys. Rev. Lett.* **86** 2050
[40] Mastromatteo I and Marsili M 2011 *J. Stat. Mech.* 10012
[41] Schwab D, Nemenman I and Mehta P 2014 *Phys. Rev. Lett.* **113** 068102
[42] Aitchison L, Corradi N and Latham P 2016 *PLoS Comput. Biol.* **12** 1005110
[43] Nonnenmacher M, Behrens C, Berens P, Bethge M and Macke J 2016 (arXiv:1603.0097v1 [q-bio.NC])
[44] Fisher M E 1998 *Rev. Mod. Phys.* **70** 653
[45] Kadanoff L 1966 *Physics* **2** 263
[46] Meshulam L, Gauthier J, Brody C, Tank D and Bialek W 2018 (arXiv:1809.08461 [q-bio.NC])
[47] Bradde S and Bialek W 2017 *J. Stat. Phys.* **167** 462
[48] Bak P 1996 *How Nature Works* (New York: Copernicus)
[49] Kirkpatrick T, Cohen E and Dorfman J 1982 *Phys. Rev.* A **26** 950
[50] Derrida B 2007 *J. Stat. Mech.* P07023
[51] Bertini L, Sole A D, Gabrielli D, Jona-Lasinio G and Landim C 2002 *J. Stat. Phys.* **107**
[52] Huang J, Kornfield R, Szczypka G and Emery S 2014 *Tobacco Control* **23** iii26
[53] Carley K, Malik M, Landwehr P, Pfeffer J and Kowalchuck M 2016 *Saf. Sci.* **90** 48
[54] Lim K and Datta A 2012 *Proc. 3rd Int. Workshop Modeling Social Media* pp 23–43
[55] Zipf G 1949 *Human Behavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
[56] Huberman B, Romero D and Wu F 2009 *First Monday* **14** 1
[57] Starbird K and Palen L 2012 *Proc. ACM 2012 Conf. Computer Supported Cooperative Work*
[58] Suh B, Hong L, Pirolli P and Chi E 2010 *IEEE Int. Conf. Social Computing*
[59] Romero D, Meeder B and Kleinberg J 2011 *Proc. 20th Int. Conf. World Wide Web*
[60] Tsur O and Rappoport A 2012 *Proceedings of the 5th ACM International Conference on Web Search and Data Mining* (New York: ACM Press) pp 643–52
[61] Gallagher R, Reagan A, Danforth C and Dodds P 2018 *PLoS One* **13** e0195644
[62] Lehmann J, Goncalves B, Ramasco J and Cattuto C 2012 *Proc. 21st Int. Conf. World Wide Web*
[63] Ramage D, Dumais S and Liebling D 2010 *Proc. 4th Int. AIII Conf. Weblogs and Social Media*
[64] Lansley G and Longley P 2016 *Comput. Environ. Urban Syst.* **58** 58
[65] Surian D, Nguyen D Q, Kennedy G, Johnson M, Coiera E and Dunn A 2016 *J. Med. Internet Res.* **18** e232
[66] An J and Weber I 2016 *Proc. of the 10th Int. AAAI Conf. on Web and Social Media*
[67] Blei D, Ng A and Jordan M 2003 *J. Mach. Learn. Res.* **3** 993
[68] Holme P 2015 *Eur. Phys. J.* B **88** 234
[69] Sanli C and Lambiotte R 2015 *PLoS One* **10** e0131704
[70] Shannon C 1948 *Bell Syst. Tech. J.* **27** 379
[71] Ferrari U 2016 *Phys. Rev.* E **94** 023301
[72] Dudik M, Phillips S and Schapire R 2004 *Proc. 17th Annual Conf. Computational Learning Theory* (Springer) pp 472–86
[73] Ferrenberg A and Swendsen R 1988 *Phys. Rev. Lett.* **61** 2635
[74] Scott D 2004 *Papers/Humboldt-Universität Berlin, Center for Applied Statistics and Economics* vol **16**
[75] Kittel C and Kroemer H 1980 *Thermal Physics* (San Francisco, CA: Freeman)

J. Stat. Mech. (2019) 093406