

# The statistical mechanics of Twitter

Gavin Hall<sup>a,b</sup> and William Bialek<sup>b,c</sup>

<sup>a</sup>*Feinberg School of Medicine, Northwestern University, Chicago, IL 60611*

<sup>b</sup>*Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544*

<sup>c</sup>*Initiative for the Theoretical Sciences, The Graduate Center,  
City University of New York, 365 Fifth Ave., New York, NY 10016*

(Dated: December 19, 2018)

We build models for the distribution of social states in Twitter communities. States can be defined by the participation vs 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.

## I. 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 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 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.

## II. 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.

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, so we focus further, breaking these networks into sub-communities using the Clauset–Newman–

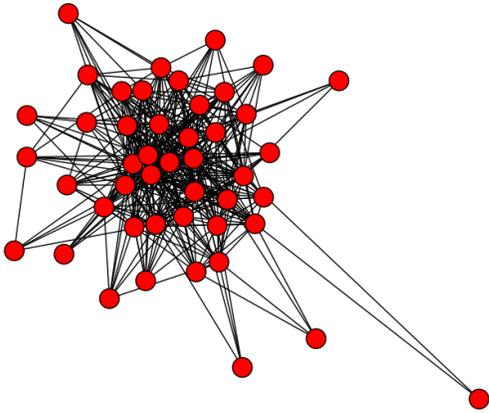


FIG. 1: **Example Social Network.** An example social network used for building a pairwise max-ent model. Same sub-community as used in Fig 4. We include more information on the topology of the social networks examined in Appendix A.

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 networks that contain between 8 and 80 people, and there don't appear to be any simple trends of topology vs size (see Appendix A for these and other details). An example of the networks that we identify is shown in Fig 1.

Having defined a network of individuals  $i = 1, 2, \dots, N$ , what are the states  $\sigma_i$  taken on by these individuals? In examining the raw data, we find prominent words that 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 J.M. Kosterlitz in a very short period of time after he shared the 2016 Nobel Prize.

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.

### III. 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't make any progress without making more than  $2^N$  measurements.

The maximum entropy method, as first emphasized by Jaynes [25, 26], gives us a way of searching systematically for simplified descriptions. Without constraints, the distribution that maximizes the entropy, and hence has the minimal structure, is the uniform distribution, which is a clearly unrealistic model. We therefore introduce constraints to ensure that the model is capable of reproducing features of the empirical data, and then maximize the entropy to remove any structures which are not absolutely necessary to meet these constraints.

In the present context, it is natural to constrain one-body marginals, which means that our model will reproduce the observed average activity of each user in the community,

$$\langle \sigma_i \rangle_P \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \sigma_i = \langle \sigma_i \rangle_{\text{obs}}. \quad (1)$$

To capture the interactions among individuals, we insist that our model also match the correlations between pairs of users, so that

$$\langle \sigma_i \sigma_j \rangle_P \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \sigma_i \sigma_j = \langle \sigma_i \sigma_j \rangle_{\text{obs}}; \quad (2)$$

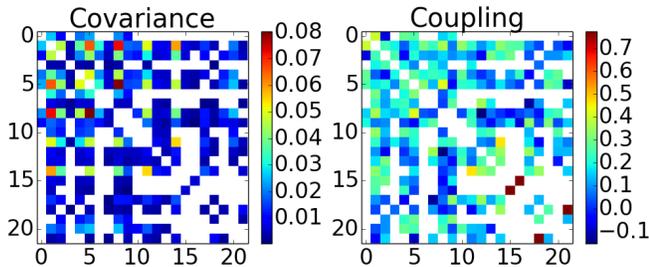


FIG. 2: **Covariance and coupling matrices.** An example of the covariance matrix [Eq (5)] and the corresponding coupling matrix  $J_{ij}$  [Eq (4)] for a Twitter community. Same community as used in Figures 3 and 4. Blank elements correspond to pairs of users without a direct social connection.

to respect the known connectivity of the network, and to limit the complexity of our models, we enforce this constraint only among pairs that have a direct social connection. Given these constraints, we look for the distribution  $P(\boldsymbol{\sigma})$  with the largest possible entropy, and the form of this distribution (Appendix C) is

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} e^{-E_{\boldsymbol{\sigma}}} \quad (3)$$

$$E_{\boldsymbol{\sigma}} = - \sum_i h_i \sigma_i - \frac{1}{2} \sum_{i \neq j} A_{ij} J_{ij} \sigma_i \sigma_j, \quad (4)$$

where  $h_i$  are parameters corresponding to constraints on one body marginals [Eq (1)] and  $J_{ij}$  are parameters corresponding to constraints on two body marginals [Eq (2)]. 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. We find the values of  $\{h_i, J_{ij}\}$  by solving Eqs (1, 2). This is in general a difficult computational problem [32], but we have solved it numerically for communities of up to 80 people using Monte Carlo methods [33] (see Appendix C for details). An example of a covariance matrix,

$$C_{ij} = \langle \sigma_i \sigma_j \rangle_{\text{obs}} - \langle \sigma_i \rangle_{\text{obs}} \langle \sigma_j \rangle_{\text{obs}} \quad (5)$$

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

Since real communities exhibit correlations with both signs, the model in Eqs (3, 4) corresponds to an Ising spin glass [30, 34]. Importantly, the couplings  $J_{ij}$  are not completely random, but are determined by the observed correlations. Generally, the couplings  $J_{ij}$  can take on both positive and negative values, and frustration is common (Appendix D).

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

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

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 Fig 3 we show the comparison of predicted vs 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 Fig 13 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,

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

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 Fig 4 that the observed distribution

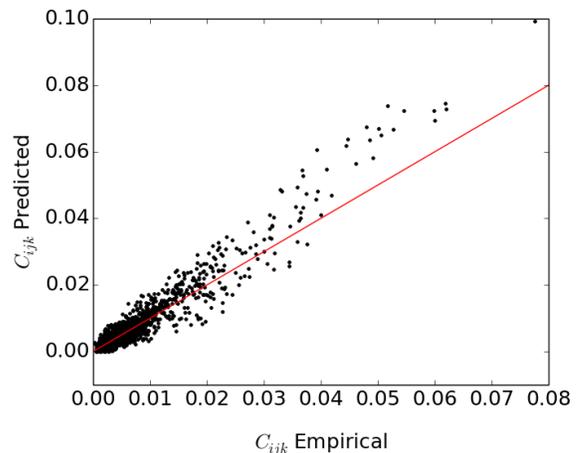


FIG. 3: **Triplet correlations.** Predicted three point correlations from our model vs the empirical value of the three point correlations estimated from data, for one sub-community. Error bars typically are  $\sim 10\%$  of the measured values; see Appendix E.

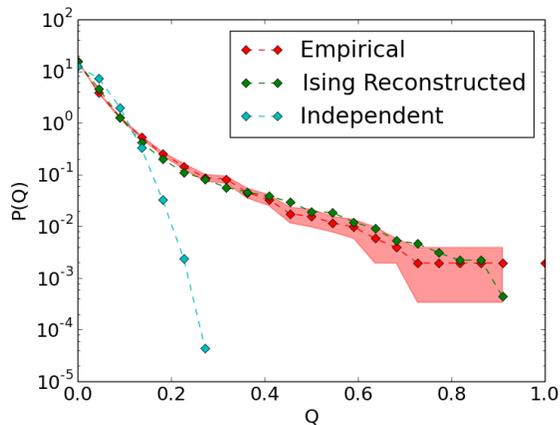


FIG. 4: **Distribution of simultaneous activity.** Probability that a fraction  $Q$  of users tweet about a particular keyword. Empirical data (red), maximum entropy model (green), and a maximum entropy model including only one-body constraints (cyan).

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 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 Figs 3 and 4 is that we don’t 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.

#### IV. 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  $N \rightarrow \infty$ , 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.

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  $\sigma$ ; energy minima correspond to probability maxima, identifying states that are favored by the network. Remarkably, 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’s important that we don’t 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’  $\Delta h$  that biases each user to tweet or remain silent, and the susceptibility  $\chi$  is proportional to the variance of the fractional activity  $Q$  defined above [Eq (7)]. In Fig 5, we track  $\chi$  vs  $\Delta h$  in a sub-community of 46 people (Fig 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 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

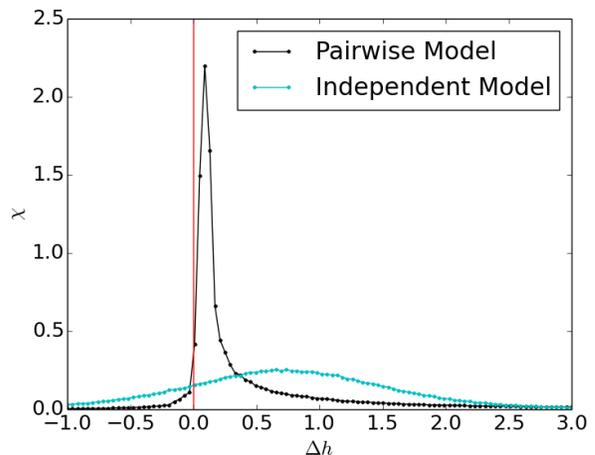


FIG. 5: **Susceptibility.** Predicted susceptibility against a forcing field for the community of 46 people shown in Fig 1. Pairwise maximum entropy model (black) and independent model (cyan); red line indicates  $\Delta h = 0$ , corresponding to the parameters inferred for the real network.

in the state  $\sigma_\mu = 1$  (tweeting). Mathematically, if we hold  $\sigma_\mu = 1$  for  $\mu = \mu_1, \mu_2, \dots$ , then for all the remaining users the joint distribution still is given by Eqs (3, 4), but with  $h_i \rightarrow h_i + \sum_\mu J_{i\mu}$  [29].

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 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 user of being near maximal variance in activity [35]. This is a direct but perhaps more intuitive analog of the peak in susceptibility for very small applied fields (Fig 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 which more users are active have lower probability (Fig 4), but until fully half the users are active there are more distinct states available at larger  $Q$  [Eq (7)]. Thus, less probable (higher energy) states are more numerous (higher entropy). We explore this tradeoff between probability and numerosity following Refs [21, 36, 37]; for details see Appendix F.

The maximum entropy model assigns to each state  $\sigma$  an energy  $E_\sigma$ , through Eq (4). 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

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

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 Eq (3), and so the condition

$$\frac{dS(E)}{dE} = 1 \quad (9)$$

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

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

again with  $T = 1$ . We expect that energies and entropies both are extensive, that is proportional to system size  $N$

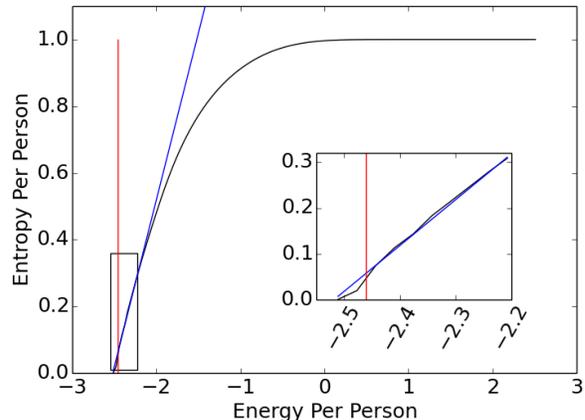


FIG. 6: **Entropy vs energy.**  $S(E)/N$  from Eq (8), plotted vs  $E/N$ , for a sub-community of  $N = 52$  users. Red line indicates average energy of system, blue line indicates line of slope 1 fit around the actual energy.

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,  $d^2 S / dE^2$  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 [38]. An example plot of  $S/N$  vs  $E/N$  is shown in Fig 6 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  $d^2 S / dE^2$  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 [36].

## V. 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 [39–42]. 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 Ref [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 renormalization group

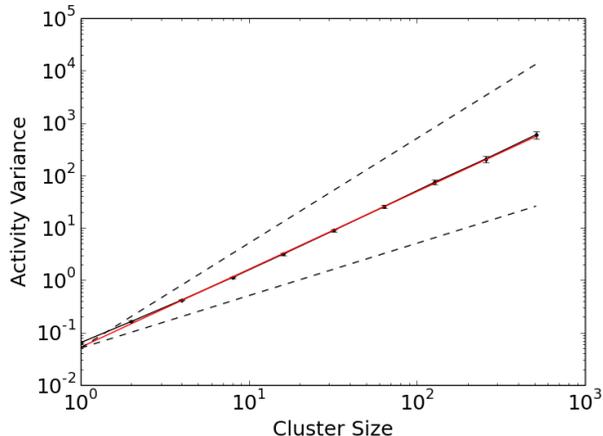


FIG. 7: **Variance Scaling.** The variance of  $\sigma_i^{(k)}$  against cluster size  $K = 2^k$ . Dotted lines indicate linear and quadratic scaling; red line indicates a power-law fit with exponent  $1.49 \pm .02$ . Error bars are the standard deviation over random halves of the data. Fit obtained by nonlinear least squares with reduced  $\chi^2 = 0.26$ .

(RG) [6, 43]. 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 [44], and then tracking the distribution of these coarse-grained 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 [45], 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,

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

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.

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 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 approach to non-trivial 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 Fig 7 we show the behavior of the variance vs 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. When comparing across different communities, we see 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 [45, 46]. In Figure 8 we plot the averaged spectra of cluster correlation matrices, for different cluster sizes  $K$ , as a function of the fractional rank of the spectrum for the same community as used in Fig 7. We stop at  $K = 128$  to avoid contamination of the spectrum by finite sample effects.

As we can see in Figure 8, 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, these spectra seem to exhibit a power law dependence as a function of their fractional rank for a large range. 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 have a power law form.

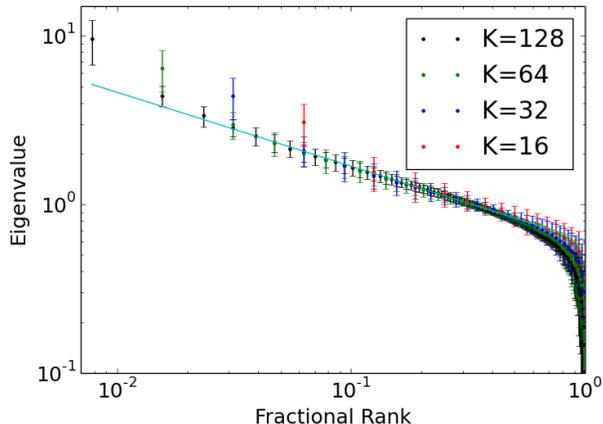


FIG. 8: **Correlation Spectra.** Correlation spectra of various cluster sizes plotted as a function of the fractional rank. Error bars are standard deviations across different clusters and random halves of the data. Cyan line is a power-law with exponent  $-0.438 \pm 0.015$ .

## VI. 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 [47]. 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 very quickly spread over a long distance in 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.

As far as we know, neither the maximum entropy method nor the renormalization group 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. 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 Fig 9.

Mean degree (at left in Fig 9) 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. For communities larger than 40 people, there is no discernible relationship between community size and the number of social connections. Interestingly, while one might imagine that these two types of social communities have different behaviors, in the communities that we have examined

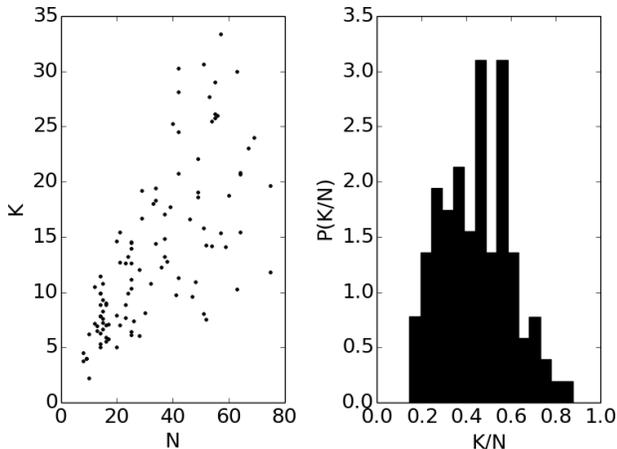


FIG. 9: **Topological Characteristics.** For 106 sub-communities with inferred pairwise max-ent models. (Left) Mean degree of social network  $K$  against community size  $N$  for all communities. (Right) Distribution of  $K/N$ .

max-ent models are capable of describing both types of systems.

### 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 [48], 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 Fig 10. As we can see in Fig 10, this distribution has two peaks, one corresponding to words that are used in a very short amount of time (far left) and words that are used with a standard deviation in time of around 500 days. The second peak begins with a kink in the distribution at a standard deviation of around 200 days. 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 key-

words 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 Fig 10.

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 [35], 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 days and a requirement that the word must be used at least 10 times in the data. In our datasets, this is generally 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.

### 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 [49]. 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 inject as little structure or knowledge as possible, as first emphasized by Jaynes [25, 26]. More specifically, we

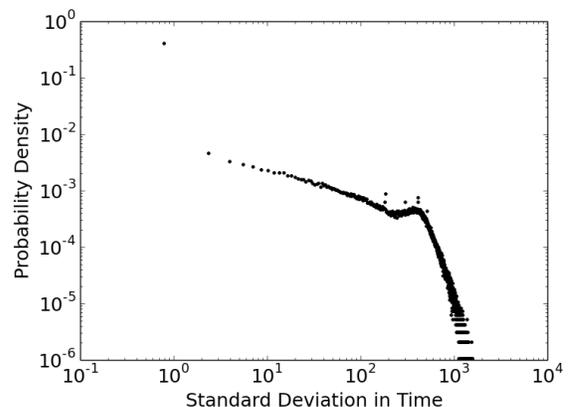


FIG. 10: **Distribution of Standard Deviations in Time of Word Usage.** For a dataset of 651 people.

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

$$\langle f_\mu(\sigma) \rangle_{\text{expt}} = \sum_{\sigma} P(\sigma) f_\mu(\sigma), \quad (\text{C1})$$

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

$$S[P(\sigma)] = - \sum_{\sigma} P(\sigma) \log P(\sigma). \quad (\text{C2})$$

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

$$\tilde{S}[P(\sigma)] = - \sum_{\sigma} P(\sigma) \log P(\sigma) + \sum_{\mu} \lambda_{\mu} \left( \langle f_{\mu}(\sigma) \rangle_{\text{expt}} - \sum_{\sigma} f_{\mu}(\sigma) P(\sigma) \right) + \lambda_0 \left( 1 - \sum_{\sigma} P(\sigma) \right), \quad (\text{C3})$$

where the Lagrange multipliers  $\lambda_{\mu}$  correspond to each constrained feature and the term proportional to  $\lambda_0$  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  $\tilde{S}[P(\sigma)]$ , as usual we take the derivative and set it to zero,

$$\frac{\partial \tilde{S}[P(\sigma)]}{\partial P(\sigma)} = 0; \quad (\text{C4})$$

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

$$P(\sigma) = \frac{1}{Z} \exp \left( - \sum_{\mu=1}^N \lambda_{\mu} f_{\mu}(\sigma) \right), \quad (\text{C5})$$

where the partition function  $Z$  absorbs  $\lambda_0$  and thus depends implicitly on the values of all the other Lagrange multipliers,

$$Z = \sum_{\sigma} \exp \left( - \sum_{\mu=1}^N \lambda_{\mu} f_{\mu}(\sigma) \right). \quad (\text{C6})$$

While Equation (C5) gives the correct form of the maximum model, it glosses over a major sticking point, which is that the correct values of the Lagrange multipliers  $\lambda_{\mu}$  must be determined. This is a difficult computational problem [32].

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  $f_{\mu}$ , 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

$$\lambda_{\mu}(t+1) = \lambda_{\mu}(t) - \eta \left[ \langle f_{\mu} \rangle_t - \langle f_{\mu} \rangle_{\text{expt}} \right], \quad (\text{C7})$$

where  $\eta$  is a learning rate. For discussions about convergence see Refs [50, 51]. 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  $\lambda$  to estimate features of the distribution parameterized by a different set of parameters  $\lambda'$  [33, 52].

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 Eqs (3, 4) 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  $\langle \sigma_i \sigma_j \rangle$ .

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 Eqs (3, 4), but with all  $J_{ij} = 0$  and

$$h_i = \text{arctanh}(\langle \sigma_i \rangle). \quad (\text{C8})$$

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

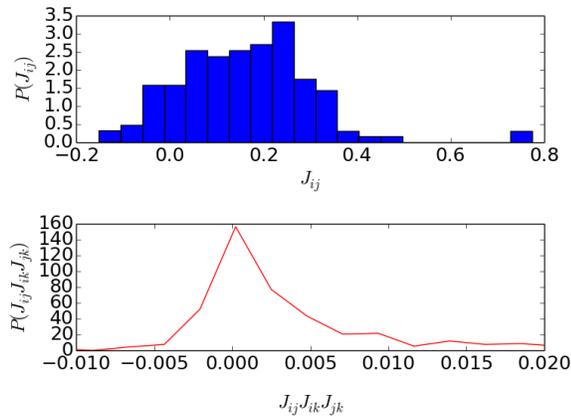


FIG. 11: **Characteristics of Coupling Matrix.** (Top) Distribution of couplings  $J_{ij}$  for the community in Fig 2. (Bottom) Distribution of the product of all interactions from closed social triangles in the same community. Approximately 30% of all possible triangles are frustrated.

#### Appendix D: Energy landscapes

The model laid out in Eqs (3) and (4) is similar to canonical models of spin glasses. Spin glasses tend to have many minima in their energy landscape due to frustration [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 Fig 11. 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 Fig 12 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 ther-

modynamic limit it would put the energy landscape of these systems into the very complex class identified for the mean-field spin glass [30].

#### 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 sheer 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 13 we show two examples of these errors. Data from the community on the left is shown in Fig 3 of the main text.

For the community on the left of Fig 13, 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 Fig 13 shows the cumulative distribution of three point correlations. As we can see, the accuracy in the top of figure 13 covers the bulk of this distribution. However, for the community shown on the right of Fig 13, 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

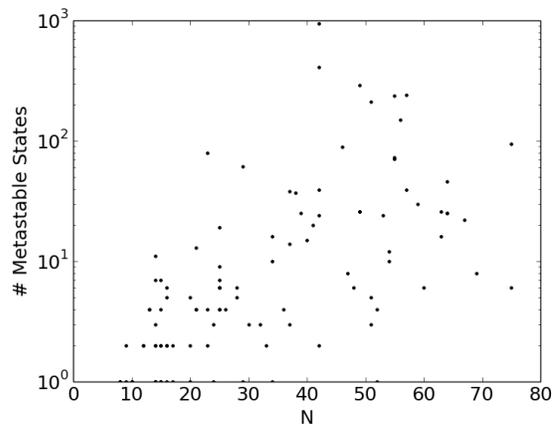


FIG. 12: **Number of Metastable States.** Number of metastable states, estimated as described in the text, for each of the communities that we analyzed.

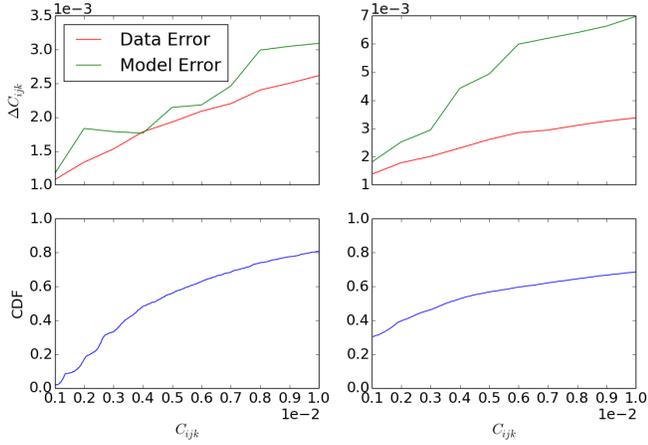


FIG. 13:  $C_{ijk}$  **Prediction Error**. Top shows the root mean square error for predictions of three point correlations from the pairwise max-ent model (green) as well as the root mean square uncertainty for empirical three point correlations (red). Data were binned to compute root mean square errors. Bottom shows cumulative distribution of empirical three point correlation values. Left and right represent data from two different communities (data from left shown in Fig 3).

a success, even if we do not reach the maximum possible accuracy.

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 Refs [21, 36, 37] 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,

$$Z = \sum_{\sigma} e^{-E} = \int dE e^{-E} \rho(E), \quad (\text{F1})$$

where

$$\rho(E) = \sum_{\sigma} \delta(E - E_{\sigma}). \quad (\text{F2})$$

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

$$Z = \int dE e^{S(E) - E}. \quad (\text{F3})$$

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

$$Z = N \int d\epsilon e^{N(s(\epsilon) - \epsilon)} = \int dE e^{-Nf(\epsilon)}, \quad (\text{F4})$$

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

$$\lim_{N \rightarrow \infty} S(E)/N = s(\epsilon) \quad (\text{F5})$$

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  $N \rightarrow \infty$ , 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  $\epsilon^*$  such that:

$$\left. \frac{df}{d\epsilon} \right|_{\epsilon^*} = 0 \implies \left. \frac{ds}{d\epsilon} \right|_{\epsilon^*} = 1 \quad (\text{F6})$$

This is true for all systems, and is another way of defining temperature[36, 53], 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:

$$Z \approx N e^{-Nf(\epsilon^*)} \int d\epsilon \exp\left(\frac{N}{2}(\epsilon - \epsilon^*)^2 \frac{d^2 s(\epsilon)}{d\epsilon^2} \Big|_{\epsilon^*}\right) \quad (\text{F7})$$

In this equation, it seems that the term outside the integral provides a contribution from the typical energy  $\epsilon^*$ , 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.

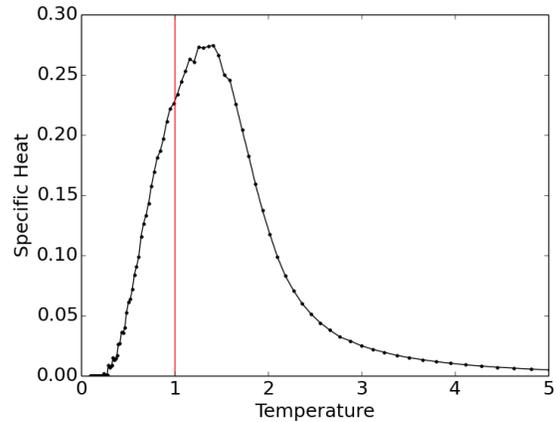


FIG. 14: **Heat Capacity** Heat Capacity (equation F8) as a function of a temperature coupled to the inferred max-ent Hamiltonian. Red Line indicates actual temperature ( $T = 1$ ).

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

$$C_T = N \left( -\frac{d^2 S}{dE^2} \right)^{-1} = \sigma_E^2 \quad (\text{F8})$$

Where  $\sigma_E^2$  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 Fig 14, where we simulate the system shown in Fig 5 with a Hamiltonian scaled by various fictitious temperatures  $H \rightarrow H/T$ . In Fig 14, 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.

- 
- [1] R. Meyerson and E. Katz, *Am J Sociol* **62**, 594 (1957).  
 [2] T. Schelling, *J of Math Sociol* **1**, 143 (1971).  
 [3] N. McCarty, K. Poole, and H. Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches* (MIT Press, 2012), 2nd ed.  
 [4] M. Brodbeck, *Philos Sci* **25**, 1 (1958).  
 [5] R. K. Sawyer, *Am J Sociol* **107**, 551 (2001).  
 [6] J. Sethna, *Statistical Mechanics: Entropy, Order Parameters and Complexity* (Oxford University Press, 2006).  
 [7] K. Wilson, *Sci Am* **241**, 158 (1979).  
 [8] C. Castellano, S. Fortunato, and V. Loreto, *Rev Mod Phys* **81**, 591 (2009).  
 [9] S. Galam, Y. Gefen, and Y. Shapir, *Mathematical Sociology* **9**, 1 (1982).  
 [10] K. S. Weron, arXiv:physics/0503239 [physics.soc-ph] (2005).  
 [11] I. Dornic, H. Chate, J. Chave, and H. Hinrichsen, *Phys Rev Lett* **87**, 045701 (2001).  
 [12] J. Silverberg, M. Bierbaum, J. Sethna, and I. Cohen, *Phys Rev Lett* **110**, 228701 (2013).  
 [13] A. Duh, M. S. Rupnik, and D. Korosak, *Big Data* **6** (2018).  
 [14] E. Bruch and J. Atwell, *Sociol Methods Res* **44**, 186 (2015).  
 [15] W. Bialek and R. Ranganathan, arXiv.org:0712.4397 [q-bio.QM] (2007).  
 [16] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa, *Proc Natl Acad Sci (USA)* **106**, 67 (2009).  
 [17] T. Mora, A. Walczak, W. Bialek, and C. Callan, *Proc Natl Acad Sci (USA)* **107**, 5405 (2010).  
 [18] D. Marks, L. Colwell, R. Sheridan, T. Hopf, A. Pagnani, R. Zecchina, and C. Sander, *PLoS One* **6**, 28766 (2011).  
 [19] E. Schneidman, M. Berry, R. Segev, and W. Bialek, *Nature* **440**, 1007 (2006).  
 [20] G. Tkačik, E. Schneidman, M. Berry, and W. Bialek, arXiv:q-bio.NC/0611072 (2006).  
 [21] G. Tkačik, T. Mora, O. Marre, D. Amodei, S. Palmer, M. Berry, and W. Bialek, *Proc Natl Acad Sci (USA)* **112**, 11508 (2015).  
 [22] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. Berry, *PLoS Comput Biol* **10**, 1003408 (2014).  
 [23] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, *Proc Natl Acad Sci (USA)* **109**, 4786 (2012).  
 [24] W. Bialek, A. Cavagna, I. Giardina, T. Mora, O. Pohl, E. Silvestri, M. Viale, and A. M. Walczak, *Proc Natl Acad Sci (USA)* **111**, 7212 (2014).  
 [25] E. Jaynes, *Phys Rev* **106**, 620 (1957).  
 [26] E. Jaynes, *Phys Rev* **108**, 171 (1957).  
 [27] E. D. Lee, C. Broedersz, and W. Bialek, *J Stat Phys* **169**, 275 (2015).  
 [28] E. D. Lee, *J Stat Phys* **173**, 1722 (2018).  
 [29] B. Daniels, D. Krakauer, and J. Flack, *Nature Comm* **8**, 14301 (2017).  
 [30] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, 1987).  
 [31] A. Clauset, M. Newman, and C. Moore, *Phys Rev E* **70**, 066111 (2004).  
 [32] H. C. Nguyen, R. Zecchina, and J. Berg, *Adv Phys* **66**, 197 (2017).  
 [33] T. Broderick, M. Dudik, G. Tkačik, R. Schapire, and W. Bialek, arXiv:0712.2437 [q-bio.QM] (2007).  
 [34] S. F. Edwards and P. W. Anderson, *J Phys F* **5**, 965 (1975).  
 [35] G. Hall, Princeton University Department of Physics (2018).  
 [36] T. Mora and W. Bialek, *J Stat Phys* **144**, 268 (2011).  
 [37] G. Stephens, T. Mora, G. Tkačik, and W. Bialek, *Phys Rev Lett* **110**, 018701 (2013).  
 [38] F. Wang and D. Landau, *Phys Rev Lett* **86**, 2050 (2001).  
 [39] I. Mastromatteo and M. Marsili, *J Stat Mech* p. 10012 (2011).  
 [40] D. Schwab, I. Nemenman, and P. Mehta, *Phys Rev Lett* **113**, 068102 (2014).  
 [41] L. Aitchison, N. Corradi, and P. Latham, *PLoS Comput Biol* **12**, 1005110 (2016).  
 [42] M. Nonnenmacher, C. Behrens, P. Berens, M. Bethge, and J. Macke, arXiv:1603.0097v1 [q-bio.NC] (2016).  
 [43] M. E. Fisher, *Rev Mod Phys* **70**, 653 (1998).  
 [44] L. Kadanoff, *Physics* **2**, 263 (1966).  
 [45] L. Meshulam, J. Gauthier, C. Brody, D. Tank, and W. Bialek, arXiv:1809.08461 [q-bio.NC] (2018).  
 [46] S. Bradde and W. Bialek, *J Stat Phys* **167**, 462 (2017).  
 [47] P. Bak, *How Nature Works* (Copernicus, 1996).  
 [48] G. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).  
 [49] C. Shannon, *Bell Sys Tech J* **27**, 379 (1948).  
 [50] U. Ferrari, *Phys Rev E* **94**, 023301 (2016).  
 [51] M. Dudik, S. Phillips, and R. Schapire, in *Proceedings of the 17th Annual Conference on Computational Learning Theory* (Springer Verlag, 2004), pp. 472–486.  
 [52] A. Ferrenberg and R. Swendsen, *Phys Rev Lett* **61**, 2635 (1988).  
 [53] C. Kittel and H. Kroemer, *Thermal Physics* (W.H. Freeman, 1980).