2011 International Symposium on Nonlinear Theory and its Applications
NOLTA2011, Kobe, Japan, September 4-7, 2011

SNS Analysis Based on Statistical-Thermodynamical Formalism

Syuji Miyazaki† and Taro Takaguchi‡

†Department of Applied Analysis and Complex Dynamical Systems, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
‡Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan
Email: syuji@acs.i.kyoto-u.ac.jp

Abstract—Random walk on a real social networking service consisting of 2271 nodes is analyzed on the basis of the statistical-thermodynamics formalism to find phase transitions in network structure. Each phase can be related to a characteristic local structure of the network such as a cluster or a hub. For this purpose, the generalized transition matrix is introduced, whose largest eigenvalue yields statistical structure functions. The weighted visiting frequency related to the Gibbs probability measure, which is useful for extracting characteristic local structures, is obtained from the products of the right and left eigenvectors corresponding to the largest eigenvalue. An algorithm to extract the characteristic local structure of each phase is also suggested on the basis of this weighted visiting frequency.

1. Introduction

The statistical-thermodynamical formalism has been successfully applied to temporal fluctuations caused by chaotic or stochastic dynamics. In chaotic dynamical systems, local expansion rates which evaluate an orbital instability fluctuate largely in time, reflecting a complex structure in the phase space. Its average is called the Lyapunov exponent, whose positive sign is a practical criterion of chaos. There exist numerous investigations based on large deviation statistics in which one considers distributions of coarse-grained expansion rates (finite-time Lyapunov exponent) in order to extract large deviations caused by non-hyperbolicities or long correlations in the vicinity of bifurcation points[1]. In general, statistical structure functions consisting of weighted averages, variances, and these partition functions as well as fluctuation spectra of coarse-grained dynamic variables can be obtained by processing the time series numerically. In some cases, we can obtain these structure functions from matrix calculations. We herein try to apply to network analyses an approach based on an weighted visiting frequency corresponding to the Gibbs probability measure and large deviation statistics in the research field of chaotic dynamical systems. Along this line, graphs and networks can be related to chaotic dynamics[2].

In the q-phase transitions of the chaotic dynamics at the band crisis or the band merging, an chaotic attractor collides with another attractor or repeller[3]. One of our motivations is to establish an analogy between the characteristic local structure of the graph and the former attractor or repeller of the whole attractor at the band crisis.

2. Statistical-thermodynamics formalism for temporal fluctuations

For stationary discrete-time signals $\tilde{u}_j$ $(j = 1, 2, \cdots)$, we consider the following local average over n steps $\overline{u}_n = \frac{1}{n}\sum_{j=1}^{n}\tilde{u}_j$. For $n \to \infty$, $\overline{u}_n$ coincides with the long-time average $\langle u\rangle$. For a large but finite n, $\overline{u}_n$ fluctuates and distributes. Let the distribution function be $P_n(u)$. Even for random or chaotic time series, there exists a characteristic time scale n_c of correlation decay. For $n \gg n_c$, the following scaling holds: $P_n(u) \propto \exp(-nS(u))$, where $S(u)$ is called fluctuation spectrum or rate function. Note that the following limit holds: $P_\infty(u) = \delta(u - \overline{u}_\infty)$, $\quad \overline{u}_\infty = \langle u\rangle$. For a real parameter q, we define the following generating function $M_q(T)$: $M_q(n) \equiv \langle e^{qn\overline{u}_n}\rangle = \int_{-\infty}^{\infty} P_n(u)e^{qnu}du$. For $n \gg n_c$, the following scaling holds: $M_q(n) \propto \exp\left(n\phi(q)\right)$, where the characteristic function $\phi(q)$ is introduced in the limit of $n \to \infty$. Thus, we have $M_q(n) \propto \int_{-\infty}^{\infty} e^{-[S(u)-qu]n}du$ for large n. Assuming the concavity of $S(u)$ $(S''(u) > 0)$, we can apply the saddle-point method to the integral and we have the following Legendre transformation between $\phi(q)$ and $S(u)$: $\phi(q) = -\min_{u'}[S(u') - qu']$ for large n. Since the integrand $S(u') - qu'$ takes minimum at $u' = u(q)$, we have $\frac{dS(u(q))}{du(q)} = \phi(q) = -S(u(q)) + qu(q)$, where $\phi(q)$ is convex downward and $\phi(q)/q$ is monotonically increasing with respect to q. Differentiating $\phi(q)$ with respect to q, we have $u(q) = \frac{d\phi(q)}{dq} = \lim_{n\to\infty}\frac{\langle\overline{u}_n e^{qn\overline{u}_n}\rangle}{M_q(n)} = \lim_{n\to\infty}\langle\overline{u}_n; q\rangle_n$, where the weighted average $\langle ...; q\rangle_n = \langle ...e^{qn\overline{u}_n}\rangle / M_q(n)$ is defined. Using this weighted average, we can extract larger (smaller) local averages than the long-time average for $q > 0$ $(q < 0)$ from among various local averages. The long-time average corresponds to $q = 0$. The weighted variance $\chi(q) = \frac{du(q)}{dq} = \lim_{n\to\infty} n\left\langle(\overline{u}_n - u(q))^2; q\right\rangle_n = \frac{1}{S''(u(q))}$ corresponds to a fluctuation intensity as a function of q. The functions $\phi(q)$, $u(q)$, $\chi(q)$ and $S(u)$ are called statistical structure functions characterizing temporal fluctuations. The relationship among the parameter q and the statistical

structure functions is similar to that among several quantities and the thermodynamics functions of the ferromagnet below the Curie temperature where the magnet field, the magnetizaton and the susceptibility correspod respectively to q, $u(q)$, and $\chi(q)$. One may also relate the inverse temperature to q. The name, "statistical-thermodynamics formalism", comes from this analogy. Let us consider the discrete-time N-state Markovian process given by the evolution equation $\mathbf{P}(n+1) = H\mathbf{P}(n)$ $(n = 0, 1, 2, \cdots)$, where $\mathbf{P}(n) = (P_1(n), P_2(n), \cdots, P_N(n))^T$ consists of the probability $P_j(n)$ that the system is in the j-th state at time n, and H denotes transition matrix with jk element H_{jk} being equal to the transition probability from the k-th state to the j-th one. The transition probability satisfies the normalization $\sum_{j=1}^{N} H_{jk} = 1$. Let us consider the time series of $\tilde{u}_n$, which takes the value a_j if the system is in the j-th state. The generating function $M_q(n)$ for the time series $\{\tilde{u}_n\}$ is given by $M_q(n) = \left\langle \exp\left(q\sum_{s=0}^{n-1}\tilde{u}_s\right)\right\rangle = \sum_{j=1}^{N}\left(H_q^n\mathbf{P}_*\right)_j$, where $\mathbf{P}_*$ is the steady probability density, and is commercially valuable information in the field of the World Wide Web called PageRank (`http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf`). The generalized transition matrix H_q is defined by $H_q = H\,\mathrm{e}^{qU}$, where U is the diagonal matrix with the jk element being equal to $U_{jk} = a_j\delta_{jk}$. For large n, we have $M_q(n) \propto \exp(n\phi(q))$. Thus, we find that the characteristic function $\phi(q)$ is identical to the logarithm of the largest eigenvalue ν_q of H_q as $\phi(q) = \log\nu_q$. Note that $\nu_0 = 1$ holds. The other statistical structure functions can be obtained analytically from the relations described above.

3. Application to an SNS network

We apply our analysis based on the statistical-thermodynamics formalism to a real social networking service (SNS). Our analysis object is in such a way constructed that we choose all users within second-neighbor distance from a specific user belonging to the largest SNS in Japan called *mixi* (`http://mixi.jp/`). Let us regard user as node, *my-mixi* relation indicating a friendship on the SNS as undirected link, so that we have an undirected graph with 2271 nodes, among which 11559 undirected links exist as shown in Fig. 1 by use of the program called *Pajek* for analysis and visualization of large networks (`http://pajek.imfm.si/`). The *mixi* users specify some *keywords* such as *fashion*, *cooking* as their matters of concern. For a fixed *keyword*, we assign the node-dependent quantity $a = 1$, when the node (user) chooses the *keyword*, and $a = 0$ otherwise. Random walk on the object graph yields random sequence of 0 and 1 denoted by $\{\tilde{u}\}$. The statistical structure functions are obtained from the largest eigenvalue of the 2271×2271 matrix H_q. There are some remarkable non-analytical behaviors, which implies the presence of q-phase transitions. Stepwise discontinuous leaps are observed in the weighted average $u(q)$ of the analysis object, which separate five *phases*. Four eminent sharp peaks are also observed in the q-dependence of the weighted variance $\chi(q)$ at the q-phase transition points. The whole graph is not uniform and separated into some local structure which can be characterized by the same fluctuation property of the node-dependent quantity a, so that such a local structure appears as a *phase* in the q-phase transition. Although the link structure of the graph is identical, different choice of the node-dependent quantity yields different q-phase transitions, which implies that our method characterizes simultaneously both the link structure and the distribution of the node-dependent quantity. The weighted visiting frequency $v_i(q)h_i(q)$ $(i = 1, 2, \cdots, 2271)$ given by the left and right eigenvector corresponding to the largest eigenvalue ν_q of the generalized transition matrix H_q also reflects the phase. The top hundred nodes of the weighted visiting frequency $v_i(q)h_i(q)$ differ from phase to phase as shown Fig. 2. We extract community structure as a phase of the statistical structure functions. We hereafter regulate ourselves to the case of *cooking*. We find five phases in Fig. 2, whose transition points are given by $-\infty = q_0 < q_1 < q_2 < q_3 < q_4 < q_5 = \infty$. We call the phase corresponding to $q \in [q_{\alpha-1}, q_\alpha)$ $(\alpha = 1, 2, 3, 4, 5)$ phase α, which is extracted by the following procedures: (1) [Calculate the phase-averaged weighted visiting frequencies of each node $\bar{p}_\alpha^{(i)} \equiv \frac{1}{q_\alpha - q_{\alpha-1}}\int_{q_{\alpha-1}}^{q_\alpha} v_i(q)h_i(q)dq$, where $q_0 = -\infty$ and $q_5 = \infty$ are replaced by suitable finite cutoff values in numerical estimations.] (2) [Sort $\bar{p}_\alpha^{(i)}$ in descending order and choose m nodes, such that m is the minimum number satisfying $\sum_m \bar{p}_\alpha^{(i_m)} \geq P$, where P $(0 \leq P \leq 1)$ is a ratio of the chosen nodes to the total nodes contained in phase α called contribution rate in the following.] When P is equal to unity, all nodes are chosen. Note that an identical node may have a large value of the weighted visiting frequency and may be chosen in different phases according to our procedures. Although many known methods divide a network into subnetworks, our method does not make a complete division. In the case of the network of Fig. 1, community structure as a phase is obtained for $P = 0.7$ and shown in Fig. 3 (a)-(e). When q is nearly equal to zero, the node-dependent quantities consist of zeros and ones, and the networks between such nodes have a few hubs and many satellites, as shown in Fig. 3 (b) and (c). For large (small) q, all the node-dependent quantities are equal to one (zero), and the networks are tightly clustered as shown in Fig. 3 (a), (d) and (e). We observed in Fig. 2 eminent large values of $v_i(q)h_i(q)$ of one node for the phase $-0.5 \lesssim q \lesssim 0$, one for $0 \lesssim q \lesssim 0.8$ and two for $q \gtrsim 2.2$, which might be regarded as *hubs* yielding the corresponding value of the local average of the node-dependent quantity $\tilde{u}$ or equivalently of the weighted average $u(q)$. It should be noted that we have q-phase transition points at $q = 0$ in many cases, so that the PageRank, the unweighted visiting frequency $v_i(0)h_i(0)$, is a special case in our formalism. The weighted visiting frequencies $v_i(q)h_i(q)$ just before and after the transition point $q = 0$ are quite different, as shown in Fig. 2.

4. Level dynamics as a future problem

In a sense, the unweighted visiting frequency $v_i(0)h_i(0)$ is degenerated. Regarding the standard and generalized transition matrices H and H_q as unperturbed and perturbed Hamiltonian, respectively, we can break the degeneracy by use of the perturbation. Such a quantum dynamics like analog of our statistical thermodynamical formalism called level dynamics is developed. Fujisaka and Yamada derived a system of differential equations (equations of motion) for eigenvalues and eigenvectors of H_q, regarding q as a virtual time[4]. For the initial conditions of eigenvalues and eigenvectors of the conventional transition matrix or the Frobenius-Perron matrix at $q = 0$, eigenvalues and eigenvectors of the generalized transition matrix or the Frobenius-Perron matrix are determined by solving the above-mentioned equations of motion.

Even for a simple discrete-time two-state Markov chain, the equations of motion turn out to be a system of strongly nonlinear differential ones. For many states, it seems impossible to solve. In the case of numerical analyses, we have to numerically solve an initial-value problem of a system of N^2 nonlinear differential equations, where N denotes a number of states. It would be rather faster to solve an eigenvalue problem of the generalized transition matrix or Frobenius-Perron matrix H_q than to solve the level dynamics.

At this moment, we find no advantage of level dynamics approaches to large deviation statistics both analytically and numerically. However, we are still interested in hidden conserved quantities and other theoretical aspects. The relationship between nearest-neighbor-spacing distributions or other spectral statistics and level dynamics for stochastic matrices is an interesting and open problem.

Figure 1: The analysis object constructed from a social networking service with 2271 nodes are drawn. Here the *keyword* is fixed to *cooking*. The node-dependent quantity 0 (1) is indicated by a white (black) circle.

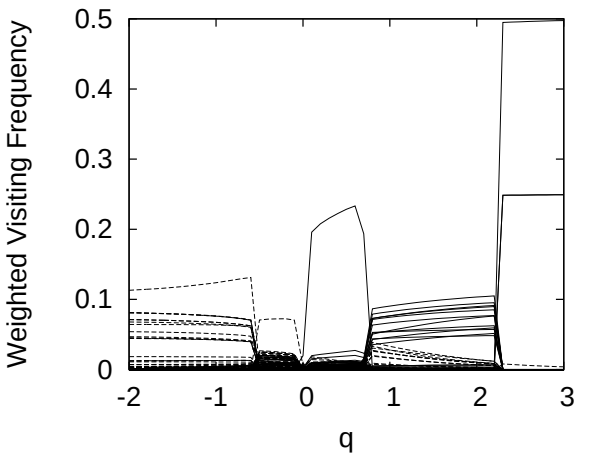

Figure 2: The weighted visiting frequencies $v_i(q)h_i(q)$ plotted against q for $i = 1, 2, \cdots, 2271$ in the case of the network shown in Fig.1 The weighted visiting frequencies of the identical node are connected with a line. In total 2271 lines are drawn. For a fixed value of q, $v_i(q)h_i(q)$ are nearly equal to zero for most nodes, and only a few nodes have finite values, which are distinguishable from the horizontal axis ($v_i(q)h_i(q) = 0$). The node-dependent quantity 0 (1) is indicated by a dashed (solid) line. Multiple nodes may degenerate into a single line. For $q > 2.2$, e. g., the lower line corresponds to two nodes.

Acknowledgments

We thank the *mixi* administrators for accepting our data acquisition. This work was supported by Grant-in-Aid for Scientific Research (c) (No.20540376).

References

[1] H. Mori and Y. Kuramoto, Dissipative Structures and Chaos, Springer, Berlin (1998).

[2] T. Takaguchi, K. Ejima and S. Miyazaki , Prog. Theor. Phys. **124**, 27 (2010).

[3] S. Miyazaki, N. Mori, T. Yoshida, H. Mori, H. Hata and T. Horita, Prog. Theor. Phys. **82**, 863 (1989) and references cited therein.

[4] H. Fujisaka and T. Yamada, Phys. Rev. E, **75**, 031116 (2007).

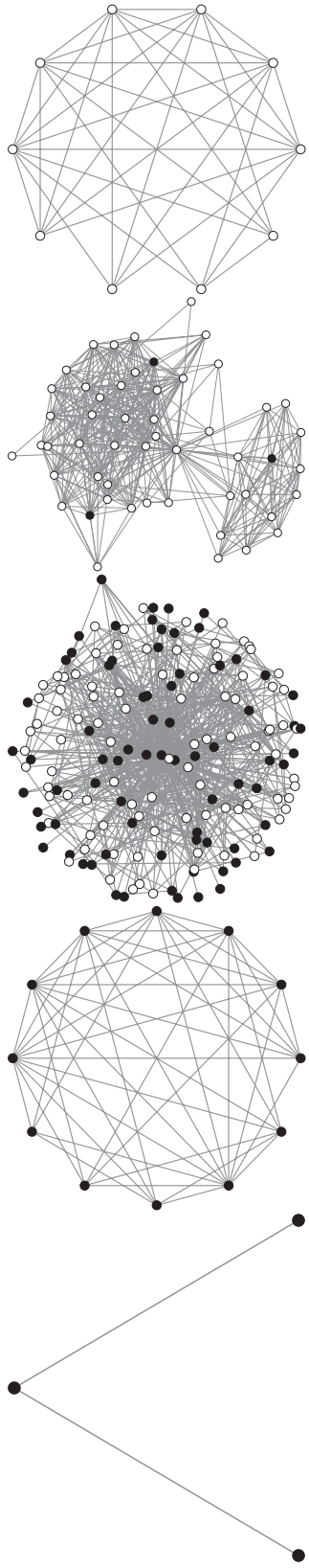

Figure 3: (a)-(e) from top down. Community structure of the network shown in Fig. 1 with $P = 0.7$. Each of the five subnetworks is extracted as a phase according to our procedures. The node-dependent quantity 0 (1) is indicated by a white (black) circle.