Poisson statistics of PageRank probabilities of Twitter and Wikipedia networks

K.M. Frahm and D.L. Shepelyansky

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received: February 25, 2014

Abstract. We use the methods of quantum chaos and Random Matrix Theory for analysis of statistical fluctuations of PageRank probabilities in directed networks. In this approach the effective energy levels are given by a logarithm of PageRank probability at a given node. After the standard energy level unfolding procedure we establish that the nearest spacing distribution of PageRank probabilities is described by the Poisson law typical for integrable quantum systems. Our studies are done for the Twitter network and three networks of Wikipedia editions in English, French and German. We argue that due to absence of level repulsion the PageRank order of nearby nodes can be easily interchanged. The obtained Poisson law implies that the nearby PageRank probabilities fluctuate as random independent variables.

PACS. 89.75.Hc Networks and genealogical trees – 05.45.Mt Quantum chaos; semiclassical methods – 89.75.Fb Structures and organization in complex systems – 89.20.Hh World Wide Web, Internet

1 Introduction

The PageRank vector $P(K)$ of the Google matrix $G_{ij}$ had been proposed by Brin and Page for ranking of nodes of the World Wide Web (WWW) in 1998 [1]. At present the PageRank algorithm became a fundamental element of various search engines including Google search [2]. This ranking works reliably also for other networks like the Physical Review citation network [3], Wikipedia [5,6,7] and other networks including even the world trade network [8]. Thus it is important to understand the statistical properties of the PageRank vector.

To study the properties of PageRank probabilities we use the standard approach [1,2] following the notation used in [3]. The directed network is constructed in a usual way: a directed link is formed from a node $j$ to a node $i$ when $j$ quotes $i$ and an element $A_{ij}$ of the adjacency matrix is taken to be unity when there is such a link and zero in absence of link. Then the matrix $S_{ij}$ of Markov transitions is constructed by normalizing elements of each column to unity ($\sum_j S_{ij} = 1$) and replacing columns with only zero elements (dangling nodes) by $1/N$, with $N$ being the matrix size. Then the Google matrix of the network takes the form [1,2]:

$$G_{ij} = \alpha S_{ij} + (1 - \alpha)/N .$$

The damping parameter $\alpha$ in the WWW context describes the probability $(1 - \alpha)$ to jump to any node for a random surfer. For WWW the Google search engine uses $\alpha \approx 0.85$ [2]. The matrix $G$ belongs to the class of Perron-Frobenius operators [2], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leq \alpha$. The right eigenvector at $\lambda = 1$, which is called the PageRank, has real non-negative elements $P(i)$ and gives a probability $P(i)$ to find a random surfer at site $i$. Thus we can rank all nodes in a decreasing order of PageRank probability $P(K(i))$ so that the PageRank index $K(i)$ counts all $N$ nodes $i$ according to their ranking, placing the most popular nodes at the top values $K = 1, 2, 3, ...$. In numerical simulations the vector $P(K_i)$ can be obtained by the power iteration method [2]. The Arnoldi method allows to compute efficiently a significant number of eigenvalues and eigenvectors corresponding to large values of $|\lambda|$ (see e.g. [9,10,11]).

From a physical viewpoint we can make a conjecture that the PageRank probabilities are described by a steady-state quantum Gibbs distribution [12] over certain quantum levels with energies $E_i$. In the frame of this conjecture the PageRank probabilities on nodes $i$ are given by

$$P(i) = \exp(-E_i/T)/Z , \quad Z = \sum_i \exp(-E_i/T)$$

and inversely the effective energies $E_i$ are given by

$$E_i = -T \ln P(i) = T \ln Z .$$

Here $Z$ is the statistical sum and $T$ is a certain effective temperature. In some sense the above conjecture assumes that the operator matrix $G$ can be represented as a sum of two operators $G_H$ and $G_{NH}$ where $G_H$ describes a hermitian system while $G_{NH}$ represents a non-hermitian operator which creates a system thermalization at a certain effective temperature $T$ with the quantum Gibbs distribution over energy levels $E_i$ of operator $G_H$. The last term

$$\sum_i \exp(-E_i/T)$$

summarizes the quantum transitions.
in [8] is independent of \(i\) and gives a global energy shift which is not important.

The statistical properties of fluctuations of levels have been extensively studied in the fields of Random Matrix Theory (RMT) [13] and quantum chaos [14]. The most direct characteristics is the probability distribution \(p(s)\) of level spacings \(s\) statistics. Here \(s = (E_{i+1} - E_i)/\Delta E\) is a spacing between nearest levels measured in the units of average local energy spacing \(\Delta E\). Thus the probability distribution \(p(s)\) is obtained via the unfolding procedure which takes into account the variation of energy level density with energy \(E\) [14]. We note that the value of \(T\) in [8] does not influence the statistics \(p(s)\) due to spectrum unfolding and definition of \(s\) in units of local level spacing.

In the field of quantum chaos it is well established that \(p(s)\) is a powerful tool to characterize the spectral properties of quantum systems. For quantum systems, which have a chaotic dynamics in the classical limit (e.g. Sinai or Bunimovich billiards [15]), it is known that in generic cases the statistics \(p(s)\) is the same as for the RMT, invented by Wigner to describe the spectra of complex nuclei [13,16,17]. This statement is known as the Bohigas-Giannoni-Schnit conjecture [16]. In such cases the distribution is well described by the so-called Wigner surmise \(p(s) = (\pi s^2/2)e^{-\pi s^2/4}\) [13,14,17]. For integrable quantum systems (e.g. circular of elliptic billiards) one finds a Poisson distribution \(p(s) = e^{-s}\) corresponding to the fluctuations of random independent variables. Such a Poisson distribution is drastically different from the RMT results characterized by the level repulsion at small \(s\) values.

The strong feature of \(p(s)\) statistics is that it describes the universal statistical fluctuations. Thus its use for description of PageRank fluctuations is very relevant, it provides a new statistical information about PageRank properties. We describe the results obtained within such an approach in next Sections.

2 Statistical properties of PageRank probabilities

For our studies we use the network of entire Twitter 2009 studied in [11] with number of nodes \(N = 41652230\) and number of links \(N_L = 1468365182\); network of English Wikipedia (Aug 2009; noted below as Wikipedia) articles from [15] with \(N = 3282257\), \(N_L = 71012307\); German Wikipedia (dated November 2013, noted below as Wikipedia-DE) with \(N = 1532077\), \(N_L = 36781077\) and French Wikipedia (dated November 2013; noted below as Wikipedia-FR) with \(N = 1352825\), \(N_L = 34431943\). For the last two cases we use the network data collected by S. Vigna [18]. For a given network the PageRank is computed as usually by the power or iteration method for a typical value of the damping factor \(\alpha = 0.85\). The probabilities \(P_i\) are computed with a relative precision better than \(10^{-12}\). For each node \(i\) its PageRank value \(P_i\) is associated to a pseudo-energy \(E_i\) by the relation \(E_i = -\ln(P_i)\). Obviously the energy spectrum is ordered if the index is given in the rank index \(K\), i.e. \(E_{K+1} \geq E_K\). Therefore the number \(n\) of levels below a given pseudo-energy \(E\) is given by \(n = K\) if \(E_K < E < E_{K+1}\) (we also use index \(i\) for \(E_i\)).

The evolution of energy levels \(E_i\) with the variation of the damping factor \(\alpha\) are shown in Fig. 1 for Twitter and Wikipedia networks. The results show many level crossings which are typical of Poisson statistics. We note that here each level has its own index so that it is rather easy to see if there is a real or avoided level crossing. In this respect the situation is simpler compared to energy levels in quantum systems.

In the following we fix the damping factor to the standard value \(\alpha = 0.85\). To obtain the unfolded spectrum with an average uniform level spacing of unity (see e.g. [14]) one has to replace the function \(E_i\) by a smooth func-
As shown in Fig. 3, one can very well approximate $E_K$ by a polynomial $Q(x)$ of modest degree in the variable $x = \ln(K)$. In this procedure it is better to exclude the first ten nodes with $K \leq 10$ which do not affect the global statistics. For a fit range $10 < K \leq 10^4$ a polynomial of degree 2 is already sufficient. However, for larger intervals, e.g. $10 < K \leq 10^7$ for Twitter or $10 < K \leq 10^6$ for Wikipedia it is better to increase the polynomial degree up to 20. Once the polynomial fit is known one obtains the unfolded energy eigenvalues $S_i$ by solving the equation $E_i = Q(\ln(S_i))$ using the Newton method. For each energy the obtained value of $S_i \approx i$ is rather close to $K = i$ index with an average spacing of unity. In certain cases this equation does not provide a solution for energies close to the boundary of the fit range. In these cases the unfolded spectrum is slightly reduced with respect to the initial fit range.

In Fig. 4 only a polynomial of degree 2 is used since the fit range $10 < K \leq 10^4$ is rather small and the histogram fluctuations, compared with the Poisson distribution, are still quite considerable due to the limited number of $N_s \approx 10^4$ data points. The obtained data show a good agreement of results with the Poisson statistics.

In Fig. 4 we show the integrated probability to find a level spacing larger than $s$:

$$I_p(s) = \int_s^{\infty} d\tilde{s} \tilde{p}(\tilde{s}). \tag{4}$$

This quantity is numerically more stable since no histogram is required. One simply orders the spacings $s_i = S_{i+1} - S_i$ and draws the ratio $1 - i/N_s$ versus $s_i$ where $i$ is the ordering index of the spacings and $N_s$ is the number of spacings in the numerical data.

The data shown in Fig. 4 clearly demonstrate that $I_p(s)$ follows the Poisson expression $I_p(s) = \exp(-s)$ for a quite large range of level spacings. Of course, for the largest values of $s$ there are deviations which are either due to the lack of statistics (especially for modest values of $K_{\text{max}}$) or due to the fact that the number of levels is close to the total network size.

We also note that for large values of $K \geq 10^6$ there are $N_d$ degenerate nodes with identical $P(i)$ values with at least one more another node or a few nodes. Such an effect has been pointed in [11]. These artificial degeneracies provide an additional delta function contribution $w_0 \delta(s)$ in the Poisson statistics $p(s)$ where $w_0$ is the probability to find such a degeneracy. There are about $N_d \approx 10^2$ ($N_d \approx 10^4$) degeneracies for Twitter nodes for $K < 10^6$ ($K < 10^7$) which gives $w_0 \approx 10^{-4}$ ($w_0 \approx 10^{-2}$). In a histogram of bin-width $\Delta s = 0.1$ this gives a relative change of the height of the first bin at $s = 0$ of $10 w_0 \approx 10^{-3}$ ($\approx 10^{-1}$) and unless we use too large $K$ value the statistical contribution of such degenerate nodes is indeed very small.

We note that if we use all nodes of Twitter up to $K < 4.2 \times 10^7$ we have $N_d \approx 1.1 \times 10^7$ with $w_0 \approx 0.26$ which is indeed considerable. In this particular case also the distribution of close degeneracies $(0 < s \ll 1)$ is quite different from the (rescaled) Poisson distribution $(1-w_0) \exp[-(1-w_0)s]$ for the non-degenerate levels. Apparently a particular network structure, which is responsible for the degeneracies, also enhances the number of close degeneracies. We attribute the appearance of such degeneracies to a weak interconnections between nodes at the tail of PageRank probability where the fluctuations are not stabilized being sensible to the finite network size.

Our data show that the Poisson statistics gives a good description of fluctuations of PageRank probabilities. It may be interesting to determine what are the nodes which have very large spacings $s$ from nearest levels on both sides. It is natural to expect that those nodes will be rather stable in respect to modifications of network or damping factor variations. Such nodes for Wikipedia network are shown in Table 1 for $s > 4$ and $K < 10^4$. Such a selection...
captures two important figures of English history but the
deviations of spacings values can provide a new
interest in a correctly weighted dimensionless representa-
validity of the Poisson statistics means that the
validity of nearby PageRank index $K_i$.
We also analyzed the statistics of PageRank proba-
dependence for a random triangular matrix model (triangu-
statistics. We also consider CheiRank probability vector
linked nodes are not really so important since after that we
difference is not really so important since after that we
dependences so that nearby PageRank probabilities behave
can be easily interchanged between nodes
with large spacings with nearby nodes in $K$ can provide a
and also find here the Poisson distribution.

### 3 Discussion

We use the methods of quantum chaos to study the sta-
the next. We think that a further study of nodes with large
interest in a correctly weighted dimensionless representa-
right of nodes of a given net-
validity of the Poisson statistics means that the
validity of nearby PageRank index $K_i$.
We also analyzed the statistics of PageRank proba-
dependence for a random triangular matrix model (triangu-
statistics. We also consider CheiRank probability vector
linked nodes are not really so important since after that we
dependences so that nearby PageRank probabilities behave

dependences for a random triangular matrix model (triangu-
1. B. Brin and L. Page, Computer Networks and ISDN Sys-
3. S. Redner, Phys. Today 58(6), 49 (2005)
4. F. Radicchi, S. Fortunato, B. Markines and A. Vespignani,
5. A.O.Zhirov, O.V.Zhirov and D.L.Shepelyansky, Eur. Phys. J. B 77, 523 (2010);
6. Y.-H.Eom, K.M.Frahm, A.Benczur and D.L.Shepelyansky, Eur. Phys. J. B 86, 492 (2013).
7. Y.-H. Eom and D.L. Shepelyansky, PLoS ONE 8(10), e74554 (2013).
8. L. Ermann and D.L. Shepelyansky, Acta Phys. Polonica A 120(6A), A158 (2011);
9. G.W. Stewart, Matrix Algorithms Volume II: Eigensystems, SIAM (2001).
10. K.M. Frahm and D.L. Shepelyansky, Eur. Phys. J. B 76, 57 (2010).
11. K.M. Frahm and D.L. Shepelyansky, Eur. Phys. J. B 85, 355 (2012);
12. L.D. Landau and E.M. Lifshitz, Statistical Mechanics, Vol.5, Nauka, Moskva (1976) (in Russian)
13. M.L.Mehta, Random matrices, Elsevier-Academic Press, Amsterdam (2004).
14. F. Haake, Quantum Signatures of Chaos, Springer, Berlin (2010).
15. L. Buninovich, Scholarpedia 2(8), 1813 (2007).
16. O. Bohigas, M.-J. Giannoni and C. Schmit, Phys. Rev. Lett. 52, 1 (1984).
17. H.-J. Stöckmann, Scholarpedia 5(10), 10243 (2010).
18. S.Vigna, web page http://vigna.di.unimi.it/}
19. K.M. Frahm, Y.-H. Eom and D.L. Shepelyansky, preprint arXiv:1310.5624 [physics.soc-ph] (2013).