Eur. Phys. J. B **53**, 381–385 (2006)
DOI: 10.1140/epjb/e2006-00377-4

THE EUROPEAN
PHYSICAL JOURNAL B

Condition numbers and scale free graphs

G. Acosta[1], M. Graña[2], and J.P. Pinasco[1,a]

[1] Instituto de Ciencias, Univ. Nac. de Gral. Sarmiento, J.M. Gutierrez 1150, (1613) Los Polvorines, Buenos Aires, Argentina
[2] Depto de Matematicas, FCEyN, Universidad de Buenos Aires, Ciudad Universitaria, Pab. 1 (1428) Ciudad de Buenos Aires, Argentina

Received 3 February 2006 / Received in final form 16 August 2006
Published online 18 October 2006 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2006

Abstract. In this work we study the condition number of the least square matrix corresponding to scale free networks. We compute a theoretical lower bound of the condition number which proves that they are ill conditioned. Also, we analyze several matrices from networks generated with Linear Preferential Attachment, Edge Redirection and Attach to Edges models, showing that it is very difficult to compute the power law exponent by the least square method due to the severe lost of accuracy expected from the corresponding condition numbers.

PACS. 02.60.Dc Numerical linear algebra – 05.10Ln Monte Carlo methods – 89.75.-k Complex systems

1 Introduction

The problem of fitting data by means of an underlying model is a common issue in science. Among several methods, the least square method (LSM) is widely used in linear and non linear cases. It is well known that even in the former (and simpler) case some stability analysis is required in order to evaluate in advance how the errors in the empirical data could be propagated to the model's coefficients.

For linear systems of equations (like those arising from linear models in the LSM) the condition number of the involved matrix [1] measures precisely how relative errors in the data can be magnified during the computation of the solution. In some way, "large" condition numbers imply a loss of accuracy in the solutions of the system. However, since relative errors are considered, the same condition number which can be regarded as harmless when the error source comes from floating point rounding matters (an important issue in numerical analysis) could be unacceptable when empirical measurements are taking into account.

Ill conditioned matrices are usually found in large linear systems. Nevertheless we will show that this is also the case for 2×2 matrices corresponding to LSM whenever the assumed independent variable ranges across several scales.

In statistics practice the cutoff value for the condition number is 225, since the effects of collinearity become strong when it exceeds this value. Condition numbers greater than 900 seems to be large indeed, causing substantial variance inflation and degrade regression estimates (see [2–4]). Experiment design could compensate the numerical error due to ill conditioned matrices by performing several measurements for a fixed value of the independent variable, which gives an error in the measured data decreasing as the inverse of the number of measurements. However, in the case of networks, it is not always clear how to reduce the measurement errors. Let us note that even the process of data mining could change severely the exponents (see for example [5,6]). For instance, in econometrics practice, where the data is collected and there is no hope to perform more experiments, condition numbers between 225 and 900 are again the borderline case suggesting strong collinearity [7].

In recent years many real problems have arisen in the literature with data ranging across several scales, like the distribution of populations or wealth, [8,9], name frequencies, and web hits, among several others. We refer to Figure 4 in [10] for several cumulative distributions of ranks/frequency plots.

Also, the analysis of small world networks offers a challenge to LSM matrices, since the size of the networks — the number of nodes — could attain very huge values. For example several networks related to Internet routers and domains, the World Wide Web, citations and coauthors in different fields and databases, words, and phone calls, range from $\sim 10^4$ to $\sim 10^8$ nodes (see tables 1 and 2 in [11] and references therein).

The cumulative and node degree distributions of these networks seem to follow a power law. Also, several models of graph growth were presented in order to explain the emergence of the supposed power law distribution (see [12–14] among others) although harsh criticism has

[a] e-mail: jpinasco@ungs.edu.ar

appeared. Sometimes, the accuracy of the data fitting for different power law distributions was questioned, as reported in [10] (where different methods were considered to avoid the linear fit on the log log scale, such as logarithmic binning or maximum likelihood estimators), or in [15] (where several biological networks were reconsidered). Also, in [16], the power law character of node distributions of human sexual contacts networks and email traffic was challenged; as well as the power law character of time intervals for e-mail communications (see [17]).

A different problem was studied in [5,6,18–22], mainly focusing on sampling bias. We may think that the large x range could be shortened by selecting small samples of a given network. However, the results in those papers show that this is not the case. For example, the reliability of the data was questioned [18]; even random networks can be mistaken as scale free networks selecting random nodes [5]; and different search algorithms give different exponents for a same scale free network [6].

Finally, a simple and striking experiment was recently presented in [23]. The authors obtained a computationally generated dataset of 10.000 samples using a random deviate generator to produce a zeta (power law) distribution with exponent $\gamma = 2.500$, and they tried to recover the exponent with different methods from the data. For the linear fit on the log log scale, a severe bias error was reported of 36%, since the mean exponent estimated was 1.590; for logarithmic bins, the estimated exponent was 1.777, that is, an error of 29%. Although the work [23] shows that the broadly used fitting methods tend to provide biased estimates for the power law exponent, there are no hints or explanations about how this happens. In contradistinction to the works mentioned in the previous paragraph, the errors in [23] cannot be related to sampling bias or data reliability.

In this work we present an underlying problem which explains those errors: as mentioned before, the matrix in the least square method is ill-conditioned. More precisely, calling n the maximum degree of the network, we show that the condition number grows at least as a power of the logarithm of n.

We also introduce a parameter $c \in [0, 1]$ and we consider only the node degree distribution on $[cn, n]$. In fact, this is a common procedure, since the decay of node degree distribution usually begins at a certain point x_{min}, see [10]. Thus, we may consider c as x_{min}/n and we took $c = 0$, $c = 0.05$ and $c = 0.1$ for computations. Numerical computations show that the situation becomes worse when we focus on the tail of the distribution in this way.

Our results complement those in [15], where biological networks were considered and a different statistical problem arose, since on that work the power law fit was performed with the maximum likelihood method. On the other hand, our results explain the situation in [23], where the causes of the severe error were not shown.

Also, we compute the matrix condition for scale free graphs generated with three models of graph generation: the Linear Preferential Attachment model (LPA) introduced by Barabasi and Albert [12], where one target node is selected at random, with a probability proportional to the degree of each node; the Edge Redirection method (ER) of Krapivsky and Redner [13], where the target node is selected at random with uniform probability, and then with probability $1 - r$ is changed by the node it points to; and the Attach to Edges model (ATE) of Dorogovtsev, Mendes and Samukhin [14], where one link is selected at random with uniform probability (among the links) and the new node connects to both ends of the chosen link. We show that the matrix condition grows as expected when the network size increases in all of them.

2 Main results

2.1 Condition number

The results on this subsection are well known and could be found in any textbook of numerical analysis; we refer the interested reader to [1,2] for an advanced exposition.

For a given matrix $A \in R^{m\times m}$, and a matrix norm $\|.\|$, the condition number is defined as

$$\operatorname{cond}(A) = \|A\|\|A^{-1}\|, \qquad \operatorname{cond}(A) = \infty \text{ if } \det(A) = 0$$

Usually, for the 2-norm the condition is denoted $\operatorname{cond}(A)_2$. The 2-norm is an operator type norm, i.e. for $v \in R^m$, taking the vectorial Euclidean norm

$$\|v\|_2 := \left(\sum_{i=1}^{m} |v_i|^2\right)^{\frac{1}{2}}$$

we have

$$\|A\|_2 = \sup\{\|Av\|_2 \quad : \quad \|v\|_2 = 1\}.$$

Concerning the condition number, the following results are well known for symmetric matrices:

$$\operatorname{cond}(A)_2 = \frac{\lambda_{max}}{\lambda_{min}}. \tag{1}$$

where λ_{min} and λ_{max} are respectively the minimum and maximum eigenvalues (in absolute value), and

$$\frac{1}{\operatorname{cond}(A)_2} = \inf\left\{\frac{\|A - S\|_2}{\|A\|_2} \ : \ S \text{ singular}\right\} \tag{2}$$

which says that $\operatorname{cond}(A)_2$ is the reciprocal of the relative distance of A to the set of singular matrices.

The interest in the condition number for matrices is related to the accuracy of computations, since it gives a bound for the propagation of the relative error in the data when a linear system is solved. If $\operatorname{cond}(A) \sim 10^k$, then k is roughly the number of significant figures we can expect to lose in computations.

More precisely, for a general system $Ax = b$, if we consider a perturbation on the right hand side $\tilde{b}$, then calling $\tilde{x}$ to the exact solution of $A\tilde{x} = \tilde{b}$ it can be shown that

$$\frac{\|x - \tilde{x}\|_2}{\|x\|_2} \le \operatorname{cond}(A)_2 \frac{\|b - \tilde{b}\|_2}{\|b\|_2}.$$

Let us note that in our case A is not the matrix of connections of the underlying graph or network, but the symmetric matrix corresponding to the least square fit.

2.2 Theoretical results

Let us consider a graph G with k nodes $x_1, \cdots, x_k$, and let $d(x_i)$ be the degree of node x_i, that is, the number of links emanating from x_i. Let us define

$$n = \max\{d(x_i) \,:\, 1 \le i \le k\}.$$

As proved in [24] for the ER model with parameter r, n grows as $k^{(1-r)}$, where N is the total number of nodes.

For each j, $1 \le j \le n$, let $h(j)$ be the number of nodes with degree j. The existence of a power law dependence $h(d) = ad^\gamma$ is usually observed in a log-log plot, and its parameters are computed with the least square method after a logarithmic change of variables.

We first assume that the degrees span the full integer interval $[1, n]$. In this case the matrix A_n corresponding to the least square fit, regardless of the measured data, is given by

$$A_n = \begin{pmatrix} n & \sum_{j=1}^n \ln(j) \\ \sum_{j=1}^n \ln(j) & \sum_{j=1}^n \ln^2(j) \end{pmatrix}. \tag{3}$$

In a certain sense, this corresponds to the best situation, where the data span the full range of variables. The following result estimates the condition number of A_n, when $n \to \infty$:

Theorem 1 *For n large, it holds*

$$\mathrm{cond}(A_n)_2 \sim \ln^4(n)$$

Proof : We use here (1). A straightforward computation of the eigenvalues of A_n gives

$$2\lambda_{max} = \Big(n + \sum_{j=1}^n \ln^2(j)\Big) + \sqrt{\Delta} \tag{4}$$

$$2\lambda_{min} = \Big(n + \sum_{j=1}^n \ln^2(j)\Big) - \sqrt{\Delta}, \tag{5}$$

where

$$\Delta = \Big(n - \sum_{j=1}^n \ln^2(j)\Big)^2 + 4\Big(\sum_{j=1}^n \ln(j)\Big)^2.$$

For n large we can write

$$\sum_{j=1}^n \ln(j) \sim n(\ln(n) - 1)) + O(\ln(n))$$

and

$$\sum_{j=1}^n \ln^2(j) \sim n(\ln^2(n) - 2\ln(n) + 2) + O(\ln^2(n)).$$

Replacing this expressions in (4) and (5), we get by taking limit

$$\lim_{n\to\infty} \frac{\frac{\lambda_{max}}{\lambda_{min}}}{\ln^4(n)} = 1.$$

Since in practice logarithmic bin is preferred (see for example [10]), due to the sparsity of measurements at the tail of the distribution, our next result shows that also the corresponding matrix is ill conditioned. We suppose that the selected degrees for the computation are of the form e^j with $1 \le j \le n$. Calling A_{e^n} the corresponding least square matrix, we can write

$$A_{e^n} = \begin{pmatrix} n & \sum_{j=1}^n j \\ \sum_{j=1}^n j & \sum_{j=1}^n j^2 \end{pmatrix} = \begin{pmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} \end{pmatrix}.$$

And the following holds

Theorem 2 *For n large*

$$\mathrm{cond}(A_{e^n})_2 \sim \frac{4}{3}n^2.$$

Proof : Using again (1), and computing explicitly the eigenvalues of A_{e^n}, we have

$$\frac{\lambda_{max}}{\lambda_{min}} = \frac{2n^2 + 3n + 7 + \sqrt{4n^4 + 12n^3 + 25n^2 + 42n + 61}}{2n^2 + 3n + 7 - \sqrt{4n^4 + 12n^3 + 25n^2 + 42n + 61}}$$

Hence, for n large

$$\mathrm{cond}(A_{e^n})_2 = \frac{\lambda_{max}}{\lambda_{min}} \sim \frac{4}{3}n^2.$$

Taking bins of the form ab^j, $1 \le j \le n$ for values of (a, b) other than $(1, e)$ does not help either: it can be shown that the condition number does not depend on a nor b in the limit $n \to \infty$. We show condition numbers for logarithmic bins in Table 1; the effect of discarding a few values (which is usually necessary) is also shown.

2.3 Numerical computations

In this section we present several numerical computations of matrix conditions.

We computed the condition number of matrix A_n numerically. We also computed the condition number of the truncated matrix A_n: for each n we took the matrix obtained with degree values between cn and n. The results are shown in Figure 1 for $n \le 1000$ with $c = 0$ and $c = 0.1$.

We show the dependence on c in Figure 2. Also, in Table 1 we show the condition number of the matrix A_{e^n} with logarithmic bins.

Finally, we considered three models used to generate scale free graphs: Linear Preferential Attachment (LPA), Edge Redirection (ER) and Attach to Edges (ATE). See the introduction for short explanations and references. We averaged condition numbers for graphs generated with each of these models. We present the results in Tables 2–4. Actually, LPA is a special case of ER (for $r = 0.5$), but these models are highly sensitive to initial conditions, which explains small differences between them.

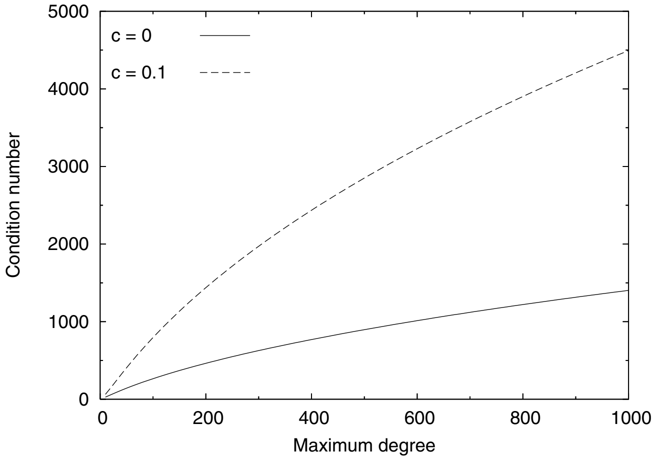

Fig. 1. Condition number of the least square matrix A_n, see equation (3), for $n \leq 1000$. The critical value 225 is reached for low values of n.

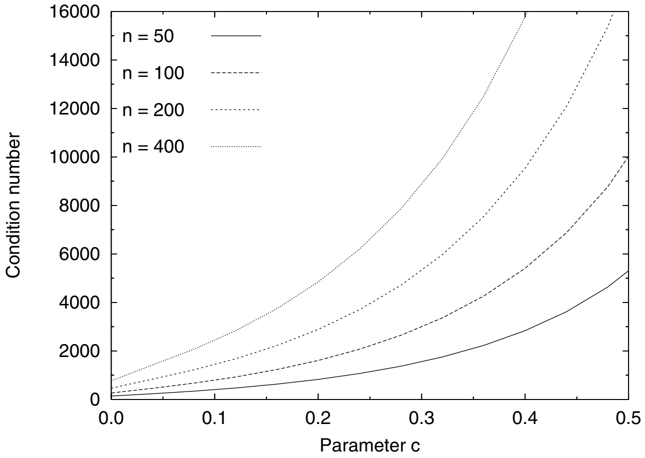

Fig. 2. Condition number of A_n as a function of c, $0 \leq c \leq 0.5$, where c measures the fraction of neglected degrees. Condition number increases when focusing on the tail.

Table 1. Condition number with logarithmic bins

e^j, $r \leq j \leq n$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$n = 5$	69.99	166.19	466.16	1847.00
$n = 10$	187.12	284.01	446.05	727.00
$n = 15$	373.00	490.94	656.28	891.15
$n = 20$	625.98	768.13	951.06	1188.04
$n = 25$	945.77	1113.23	1318.02	1569.48
$n = 30$	1332.30	1525.56	1754.05	2024.93

3 Conclusions

We have studied the condition number of the least square matrix corresponding to scale free networks. We computed theoretical lower bounds of the condition numbers showing that it behaves roughly as a power of the logarithm of the maximum degree of the network, and numerical simulations support this fact. We also showed that neglecting the less connected nodes of the network (a usual practice in fact, since the interest is on the tail) things become even worse. Similar conclusions can be drawn for logarithmic bins.

Table 2. Mean value of condition numbers for LPA graphs with different values of c.

Nodes	Graphs	$c = 0$	$c = 0.05$	$c = 0.1$
10^4	5×10^4	113.7	379.7	703.7
10^5	2.5×10^4	223.5	1058.4	1928.8
10^6	10^4	409.0	2648.5	4560.0
10^7	10^4	703.8	5897.6	9369.5

Table 3. Mean value of condition numbers for ER graphs, 1000 graphs for each line.

Nodes	r	$c = 0$	$c = 0.05$	$c = 0.1$
10^4	1.00	37.5	37.7	73.7
10^4	0.75	74.8	143.2	252.0
10^4	0.50	111.0	422.6	776.5
10^4	0.25	121.1	1068.5	2296.2
10^5	1.00	47.6	52.1	91.0
10^5	0.75	124.7	283.7	514.6
10^5	0.50	221.3	1200.8	2105.9
10^5	0.25	273.7	3194.1	8331.8
10^6	1.00	58.0	90.6	138.1
10^6	0.75	195.5	520.4	970.8
10^6	0.50	407.9	2996.9	4822.7
10^6	0.25	557.4	7762.2	14980.9
10^7	1.00	67.9	108.1	157.2
10^7	0.75	294.1	903.7	1749.4
10^7	0.50	703.8	6624.2	9823.2
10^7	0.25	1036.4	15835.3	25021.9

Table 4. Mean value of condition numbers for ATE graphs.

Nodes	Graphs	$c = 0$	$c = 0.05$	$c = 0.1$
10^4	10^5	107.8	194.2	353.1
10^5	10^5	192.6	559.5	1043.2
10^6	10^5	337.1	1448.5	2683.9
10^7	10^4	570.4	3433.8	6035.1

Finally, for random networks generated with Linear Preferential Attachment, Edge Redirection and Attach to Edges models, numerical computations of the condition numbers showed a severe ill condition of the least square matrices, even for small sized networks (10^4 nodes). Also, we confirmed the theoretical prediction of the condition number becoming worse when attention is paid to the tail. Clearly, in this context it is very difficult to compute the power law exponent by the least square method due to the lost of accuracy expected from the corresponding condition numbers.

GA and JPP are partially supported by Fundación Antorchas and ANPCyT. MG is partially supported by Fundación Antorchas and CONICET.

References

1. G.H. Golub, C.F. Van Loan, *Matrix Computations*, Johns Hopkins Series in Mathematical Sciences, 3rd edn. (The Johns Hopkins University Press, Baltimore and London, 1996)
2. D.A. Belsley, *Conditioning Diagnostics, Collinearity and Weak Data in Regression* (John Wiley & Sons, New York, 1991)
3. S. Chatterjee, A.S. Hadi, B. Price, *Regression Analysis by Example*, 3rd edn., Wiley Series in Probability and Statistics (John Wiley & Sons, 1991)
4. G.W. Stewart, Statistical Science **2**, 68 (1987)
5. D. Achlioptas, A. Clauset, D. Kempe, C. Moore *On the Bias of Traceroute Sampling*; or, *Power-law Degree Distributions in Regular Graphs* Proc. STOC (2005)
6. S.H. Lee, P.-J. Kim, H. Jeong, Phys. Rev. E **73**, 016102 (2006)
7. D.A. Belsley, *Multicollinearity: Diagnosing its Presence and Assessing the Potential Damage It Causes Least Squares Estimation*, National Bureau of Economic Research Working Paper No. 154 (1976)
8. M. Faloutsos, P. Faloutsos, C. Faloutsos, Computer Commun. Rev. **29**, 251 (1999)
9. H. Jeong, B. Tombor, B. Albert, Z.N. Oltvai, A.L. Barabasi, Nature **407**, 651 (2000)
10. M.E.J. Newman, Contemporary Physics **46**, 323 (2005)
11. R. Albert, A-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
12. A-L. Barabási, R. Albert, Science, **286**, 509 (1999)
13. P.L. Krapivsky, S. Redner, Phys. Rev. E **63**, 066123 (2001)
14. S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Phys. Rev. E **63**, 062101 (2001)
15. R. Khanin, E. Wit, J. Comput. Biology **13**, 810-8 (2006)
16. M.S. Handcock, J.H. Jones, Nature **423**, 605 (2003)
17. D.B. Stouffer, R.D. Malmgren, L.A.N. Amaral, Nature **435**, 207 (2005), `arXiv:physics/0510216`
18. T. Petermann, P. De Los Rios, Eur. Phys. J. B **38**, 201 (2004)
19. Q. Chen, H. Chang, R. Govindan, S. Jamin, S.J. Shenker, W. Willinger, *The Origin of Power Laws in Internet Topologies Revisited*, Proc. of IEEE Infocom (2002)
20. A. Clauset, C. Moore, Phys. Rev. Lett. **94**, 18701 (2005)
21. L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vazquez, A. Vespignani, Phys. Rev. E **71**, 036135 (2005)
22. A. Lakhina, J. Byers, M. Crovella, P. Xie *Sampling Biases in IP Topology Measurements*, *Proc. of IEEE INFOCOM '03* (2003)
23. M.L. Goldstein, S.A. Morris, G.G. Yen, Eur. Phys. J. B **41**, 255 (2004)
24. P.L. Krapivsky, S. Redner, J. Phys. A **35**, 9517 (2002)