REFEREE NETWORKS AND THEIR SPECTRAL PROPERTIES*

František Slanina

Institute of Physics, Academy of Sciences of the Czech Republic
Na Slovance 2, CZ-18221 Praha, Czech Republic

Yi-Cheng Zhang

Institut de Physique Théorique, Université de Fribourg
Pérolles, CH-1700 Fribourg, Switzerland

(Received July 1, 2005)

The bipartite graph connecting products and reviewers of that product is studied empirically in the case of amazon.com. We find that the network has power-law degree distribution on the side of reviewers, while on the side of products the distribution is better fitted by stretched exponential. The spectrum of normalised adjacency matrix shows power-law tail in the density of states. Establishing the community structures by finding localised eigenstates is not straightforward as the localised and delocalised states are mixed throughout the whole support of the spectrum.

PACS numbers: 89.20.–a, 89.75.–k, 89.90.+n

1. Introduction

It is often very difficult to extract information from a source even if it is plain and seemingly at hand. Nobody would deny that assessing a quality of a product without touching and trying it is a hard problem. In many cases somebody can do, and already did, that for you. There are customer groups, specialised magazines, web pages, all trying to shed some light in the jungle of goods. It seems natural to read reviews on several alternative choices for your desired thing and by comparing them select the proper one for your purpose. But you would soon find that reducing the level of uncertainty is not enough. The reviews are often still too many, while sometimes they

[*] Presented at the Conference on Applications of Random Matrices to Economy and Other Complex Systems, Kraków, Poland, May 25–28, 2005.

are missing, they focus on different features of the product, they are biased, very often they are superficial and bear little insight into the working of the products *etc.* Adding reviews in fact increases the level of complexity of the system. Instead of pure hiding the information now the information is entangled and mingled in a complicated way. However, the positive hope is, that smart enough algorithm may help solving the tangle and bring the information to light.

Extracting hidden information is in fact an old problem and significant progress was made quite recently within statistical physics [1–3]. The problem took a new spin with the boom of complex network studies [4–6]. We will concentrate only on a small segment of that area. Specifically, we will try to filter the information contained in reviews on products sold over the Internet. Many of the reviews are repeating what was already said, while some of them are so eccentric that we cannot take them seriously. As a first step, we want to group the reviewers together according to similar focus, so that many reviews can be replaced by a representative one. This amounts to looking for a community structure on the interaction network of the referees. To determine network communities is a hard problem and several algorithms are used [7–11]. We will apply the spectral approach, following the work of Refs. [10, 11]. Let us also note that communities on the on-line auction sites were identified recently [12].

2. Referee network on amazon.com

Studying activity on Internet commerce sites is relatively new branch within the physics community [12, 13]. One of the studies carried so far investigated the bipartite graph of auctions on eBay site. The nodes are of two types, first are the agents participation in the auctions, second group consists of the items to be sold. An edge is drawn connecting an agent with an item, if the agent participates in the auction concerning the particular item. It was found [13] that the degree distribution of the auction network is highly asymmetric. On the side of the agents, the power-law distribution is observed, while on the side if items the distribution is exponential.

We performed a similar study on the referee network on the site amazon.com. The first group of nodes are the books on sale (for practical reasons we neglect other types of goods sold on amazon.com), the second group are the people who wrote review(s) on any of that books. The connectivity matrix M has elements $M_{br} = 1$ if the person r reviewed book b; otherwise $M_{br} = 0$. At this stage we do not consider other information contained within the review, most notably the number of points attributed to the book. Such information is essential if we want to determine how much the reviewers agree or disagree among themselves, but for now we take into account only the bare existence or non-existence of a link between a reviewer and a book.

The empirical data were downloaded from the amazon.com site for $N_r = 1000$ reviewers with highest rank (the rank is attributed to them by amazon, based mainly on number of books reviewed and other auxiliary factors). On that small sample it was found that on the side of reviewers the node degree (*i.e.* the number of reviews written by one person) is power-law distributed, $P(k) \sim k^{-\gamma}$, with $\gamma \simeq 2$, as shown in Fig. 1. Such finding is no surprise, as the network grows by adding the books one by one and it is reasonable to suppose that the probability that a reviewer will read the book is proportional to her current productivity level, measured as the number of books reviewed so far. Thus, a mechanism analogous to preferential attachment is active here, leading naturally to power-law degree distribution [14].

On the books' side the situation is less clear. Neither power nor exponential fit are satisfactory. We found that the data are best represented by the stretched exponential $P(k) \sim \exp(-9.3\, k^{0.2})$. Currently we do not have any explanation for such specific dependence. It may well be that the observed dependence is due to some cutoff, discouraging the reviewers to examine books already read by too many people.

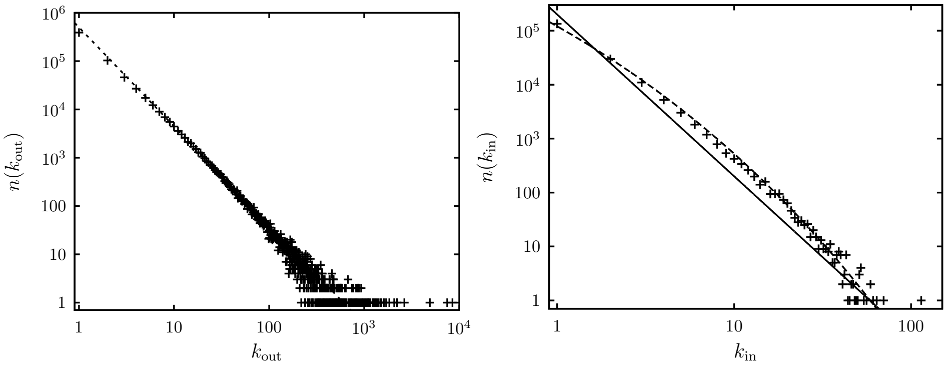

Fig. 1. Degree distributions of the referee network. In the left panel, number of referees connected to k_{out} books. The dashed line is the power $\sim k_{\mathrm{out}}^{-2.05}$. In the right panel, number of books connected to k_{in} reviewers. The solid line is the power $\sim k_{\mathrm{in}}^{-3}$ and the dashed line the stretched exponential $\sim \exp(-9.3\, k_{\mathrm{in}}^{0.2})$.

3. Extracting the community structure

Among several approaches to establish the community structures on networks the following is perhaps the most straightforward. Imagine a random walker moving along the edges of the network. If the graph were composed of disconnected clusters, the probability to find the walker will be large inside its original cluster and zero elsewhere. The average time spent at node i of the network is given by the stationary probability distribution P_i for the random walk on the graph. For isolated clusters, there will be eigenmodes of the diffusion operator which will be localised on that clusters. Identifying the localised states means finding the clusters, *i.e.* the network communities.

In realistic case the clusters, or communities, are not isolated, but weakly bound to other clusters and the localisation of the eigenstates is not perfect. However, the states which are localised in the sense that the probability is concentrated on a specific subgraph and vanishingly small elsewhere, can be well taken as signatures of the network communities also in the general case. Of course, the distinction between well localised and extended state is not sharp, and the definition of communities through localisation is fuzzy as well. For practical purposes, though, it can be very useful, at least for finding the most evident communities.

Usually the distinction between extended and localised states on the network represented by a matrix A can be made in the "energy" domain, or on the axis of eigenvalues. Denote N number of nodes in the network. The degree of localisation of a normalised eigenvector $v_{i\lambda}$ corresponding to the eigenvalue λ is measured by inverse participation number (IPN)

$$m(\lambda) = \sum_{i}^{N} v_{i\lambda}^4 \,. \tag{1}$$

The localised states have $m(\lambda)$ finite in the thermodynamic limit $N \to \infty$, while for the extended states IPN decreases as $m(\lambda) \sim 1/N$. Usually it is supposed that a thresholds $\lambda_l < \lambda_u$, often called mobility edges, exist, such that the eigenstates corresponding to eigenvalues within the interval $\lambda \in (\lambda_l, \lambda_u)$ are extended, while for $\lambda < \lambda_l$ or $\lambda > \lambda_u$ we have localised states.

For the diffusion operator this should be modified by the fact that the largest eigenvalue corresponds to the stationary state populated proportionally to the degree of the nodes and therefore the state is extended, except for extremely heterogeneous networks. Such rationale lies behind the attempts to determine the industrial sectors from the eigenvectors of the matrix of stock price correlations. Large eigenvalues except the single largest one, or rather their corresponding eigenmodes, are attributed to areas like banking, oil industry and the like [15, 16].

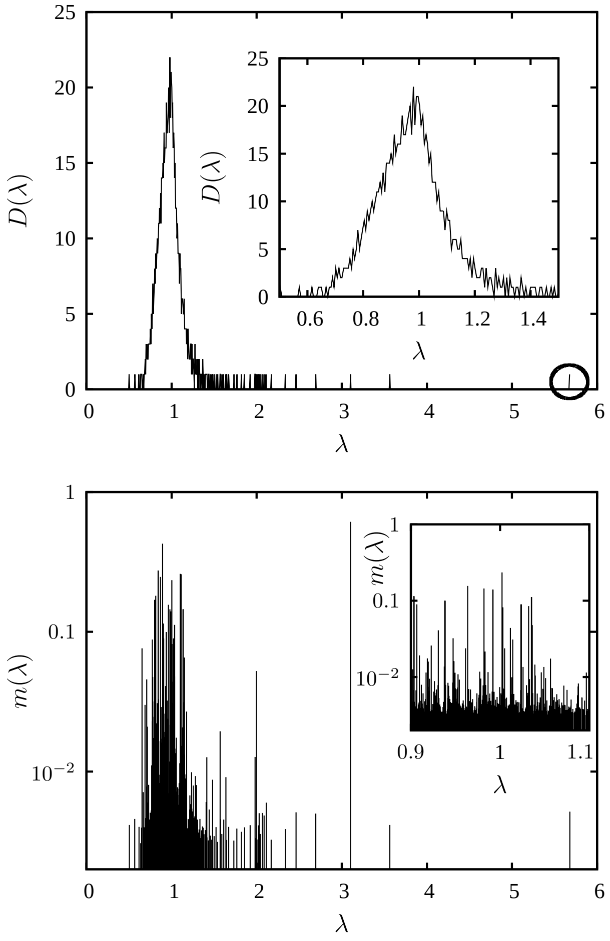

Fig. 2. Results of diagonalisation of the empirical matrix A_{rs} of 1000 highest rank amazon.com reviewers. Upper panel: the density of eigenvalues. In the inset we show detail of the same plot. The highest eigenvalue is highlighted by the circle. Lower panel: inverse participation number. Note that the most localised eigenmode is just above $\lambda = 3$, it corresponds to 3rd highest eigenvalue and the value $m(\lambda) \simeq 0.6$ means that it is localised on about 1.6 nodes. In the inset detail of the central part. We can clearly see how the localised and delocalised states are mixed on the λ axis.

Let us see how this strategy can be adapted for our referee network. We shall consider connections between pairs of referees mediated by books both of them have reviewed. So, we define the matrix

$$A_{rs} = \frac{\sum_b M_{br} M_{bs}}{\sqrt{\sum_b M_{br} M_{br} \sum_{b'} M_{b's} M_{b's}}} \tag{2}$$

which quantifies the relations between referees. $A_{rs} = 0$ implies no overlap between sets of books treated by referees r and s, while $A_{rs} = 1$ means that both of them reviewed exactly the same set. We stress again that we do not care about the correlations in the content of the reviews, but only examine the size of the intersection of the referees' interests.

The spectrum of the matrix A is shown in Fig. 2. We can observe a typical tent-shaped form of the density of states in the central region, which agrees with similar finding for the scale-free networks [17]. We can also see that the maximum eigenvalue (look at the circle in Fig. 2) lies very far from the bulk of the spectrum. This is a marked demonstration of another property of the spectrum, which is the power-law tail of the density of states,as can be seen in Fig. 3. This feature is also found in the scale-free networks [17–19] and makes connection to special class of random matrices, called Lévy matrices [20–22].

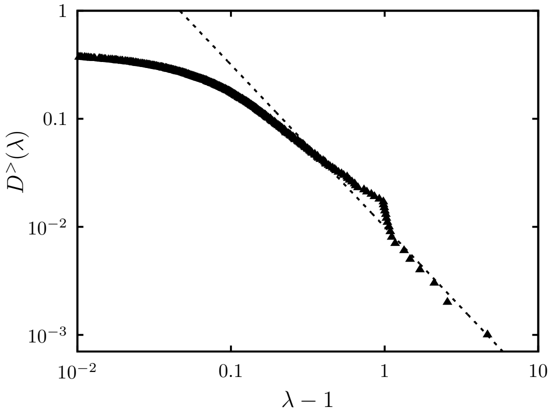

Fig. 3. Integrated density of eigenvalues for the referee–referee matrix A_{rs}. The dashed line is the power law $\sim (\lambda - 1)^{-1.5}$.

For our purposes, it is the inverse participation number which is most important. The result of numerical diagonalisation of our empirical matrix is shown in the lower panel of Fig. 2. Even though it is difficult to state exactly which eigenvectors are localised, as this property becomes well defined only in the limit $N \to \infty$, we cannot be too wrong saying that the localised states correspond to IPD larger than about 0.1. On our network with $N = 1000$ it means that they cover less than 1 per cent of the network. The surprising fact is that we cannot determine any "mobility edge" on either side of the spectrum. The localised and delocalised states come very close in the spectrum. It is interesting to note that the Lévy matrices also exhibit a mixed behaviour in certain eigenvalue range, with both extended and localised characteristics [20].

What are the consequences of such a structure of localised states for determination of the community structures? First, one must note the absence of the mobility edge. Therefore, it is not possible to claim that the communities correspond to a few highest eigenvalues, except the maximum one. It is necessary to look through entire spectrum and determine the states with highest IPN. However, such procedure is never exact, so we should rather observe which communities emerge at given level of IPN. Of course, at too low IPN level the notion of community loses its sense, as the eigenstate is effectively delocalised.

4. Conclusions

We performed an empirical study of the referee network on the amazon.com site. The bipartite graph with books on one side and referees on the other side has power-law node degree distribution on the side of reviewers, while on the books' side the distribution is better fitted by a stretched exponential.

The matrix connecting the pairs of referees through books both of them reviewed is a starting point for determining the communities among the reviewers. We found that the matrix has several non-canonical properties. As for the spectrum, it exhibits a power-law upper tail, similarly to spectra of scale-free networks and Lévy matrices, studied earlier.

The localisation properties are rather complex and we observe both localised and extended states in all parts of the spectrum. The "mobility edge" separating the localised and extended states on the axis of eigenvalues, is missing. This means that the communities cannot be directly ascribed to the largest eigenvalues, but it is necessary to parametrise the communities by the level of inverse participation number, at which the eigenmodes are picked. As there is no sharp threshold in the value of IPN, there is also smooth transition from well-defined communities at high IPN to ill-defined ones at low IPN. The methodology therefore needs developing a refined criteria, which is the topic we devote our further study.

We want to thank for many useful discussions with Paolo Laureti and Lionel Moret. This work was supported by the Ministry of Education of the Czech Republic, under the COST P10 programme, grant No. 1P04OCP10.001.

REFERENCES

[1] S. Maslov, Y.-C. Zhang, *Phys. Rev. Lett.* **87**, 248701 (2001).

[2] A. Capocci, F. Slanina, Y.-C. Zhang, *Physica A* **317**, 259 (2003).

[3] P. Laureti, L. Moret, Y.-C. Zhang, *Physica A* **345**, 705 (2005).

[4] R. Albert, A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).

[5] S.N. Dorogovtsev, J.F.F. Mendes, *Adv. Phys.* **51**, 1079 (2002).

[6] S.H. Strogatz, *Nature* **410**, 268 (2001).

[7] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Proceedings of the WWW8 Conference (1999), `http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html`

[8] M. Girvan, M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 7821 (2002) [`cond-mat/0112110`].

[9] L. Danon, J. Duch, A. Arenas, A. Diaz-Guilera, `cond-mat/0505245`.

[10] K.A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen, *Phys. Rev. Lett.* **90**, 148701 (2003).

[11] I. Simonsen, K.A. Eriksen, S. Maslov, K. Sneppen, *Physica A* **336**, 163 (2004).

[12] J. Reichardt, S. Bornholdt, `physics/0503138`.

[13] I. Yang, H. Jeong, B. Kahng, A.-L. Barabási, *Phys. Rev.* **E68**, 016102 (2003).

[14] A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).

[15] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, *Phys. Rev. Lett.* **83**, 1467 (1999).

[16] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, *Phys. Rev. Lett.* **83**, 1471 (1999).

[17] I.J. Farkas, I. Derényi, A.-L. Barabási, T. Vicsek, *Phys. Rev.* **E64**, 026704 (2001).

[18] K.-I. Goh, B. Kahng, D. Kim, *Phys. Rev.* **E64**, 051903 (2001).

[19] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, A.N. Samukhin, *Phys. Rev.* **E68**, 046109 (2003).

[20] P. Cizeau, J.P. Bouchaud, *Phys. Rev.* **E50**, 1810 (1994).

[21] Z. Burda, J. Jurkiewicz, M.A. Nowak, G. Papp, I. Zahed, *Acta Phys. Pol. B* **34**, 4747 (2003).

[22] Z. Burda, J. Jurkiewicz, M.A. Nowak, G. Papp, I. Zahed, *Physica A* **343**, 694 (2004).