IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Toward two-dimensional search engines

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2012 J. Phys. A: Math. Theor. 45 275101

(http://iopscience.iop.org/1751-8121/45/27/275101)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 130.120.231.163
The article was downloaded on 18/06/2012 at 12:01

Please note that terms and conditions apply.

J. Phys. A: Math. Theor. **45** (2012) 275101 (20pp) doi:10.1088/1751-8113/45/27/275101

Toward two-dimensional search engines

L Ermann[1], **A D Chepelianskii**[2] **and D L Shepelyansky**[1]

[1] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France
[2] Department of Physics, Cavendish Laboratory, University of Cambridge, CB3 0HE, UK

E-mail: dima@irsamc.ups-tlse.fr

Received 27 September 2011, in final form 17 May 2012
Published 18 June 2012
Online at stacks.iop.org/JPhysA/45/275101

Abstract
We study the statistical properties of various directed networks using ranking of their nodes based on the dominant vectors of the Google matrix known as PageRank and CheiRank. On average PageRank orders nodes proportionally to a number of ingoing links, while CheiRank orders nodes proportionally to a number of outgoing links. In this way, the ranking of nodes becomes two dimensional which paves the way for the development of two-dimensional search engines of a new type. Statistical properties of information flow on the PageRank–CheiRank plane are analyzed for networks of British, French and Italian universities, Wikipedia, Linux Kernel, gene regulation and other networks. A special emphasis is done for British universities networks using the large database publicly available in the UK. Methods of spam links control are also analyzed.

PACS numbers: 89.75.Fb, 89.75.Hc, 89.20.Hh

(Some figures may appear in colour only in the online journal)

1. Introduction

During the past decade, modern society has developed enormously large communication networks. The well-known example is the World Wide Web (WWW) which has started approaching 10^{11} webpages [1]. The sizes of social networks like Facebook [2] and VKONTAKTE [3] have also become enormously large, reaching 600 and 100 millions user pages, respectively. The information retrieval from such huge databases becomes the foundation and main challenge for search engines [4, 5]. The fundamental basis of the Google search engine is the PageRank algorithm [6]. This algorithm ranks all websites in a decreasing order of components of the PageRank vector (see e.g. detailed description at [7], historical surveys of PageRank are given at [8, 9]). This vector is a right eigenvector of the Google matrix at the unit eigenvalue, it is constructed on the basis of the adjacency matrix of the directed network, its components give a probability of finding a random surfer on a given node.

The Google matrix G of a directed network with N nodes is given by

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

where the matrix S is obtained by normalizing to unity all columns of the adjacency matrix $A_{i,j}$, and replacing columns with zero elements by $1/N$. An element A_{ij} of the adjacency matrix is equal to unity, if a node j points to node i and zero otherwise. The damping parameter α in the WWW context describes the probability $(1 - \alpha)$ for a random surfer to jump to any node. The value $\alpha = 0.85$ gives a good classification for the WWW [7] and thus we also use this value here. A few examples of Google matrix for various directed networks are shown in figure 1. The matrix G belongs to the class of Perron–Frobenius operators [7], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leqslant \alpha$. The right eigenvector at $\lambda = 1$ gives the probability $P(i)$ to find a random surfer at site i and is called the PageRank. Once the PageRank is found, all nodes can be sorted by decreasing probabilities $P(i)$. The node rank is then given by index $K(i)$ which reflects the relevance of the node i. The PageRank dependence on K is well described by a power law $P(K) \propto 1/K^{\beta_{\rm in}}$ with $\beta_{\rm in} \approx 0.9$. This is consistent with the relation $\beta_{\rm in} = 1/(\mu_{\rm in} - 1)$ corresponding to the average proportionality of PageRank probability $P(i)$ to its in-degree distribution $w_{in}(k) \propto 1/k^{\mu_{\rm in}}$, where $k(i)$ is a number of ingoing links for a node i [7, 10]. For the WWW, it is established that for the ingoing links $\mu_{\rm in} \approx 2.1$ (with $\beta_{\rm in} \approx 0.9$) while for the out-degree distribution $w_{\rm out}$ of outgoing links a power law has the exponent $\mu_{\rm out} \approx 2.7$ [11, 12]. Similar values of these exponents are found for the WWW British university networks [13], the procedure call network (PCN) of Linux Kernel software introduced in [14] and for Wikipedia hyperlink citation network of English articles (see e.g. [15]).

The PageRank gives at the top the most known and popular nodes. However, an example of the Linux PCN studied in [14] shows that in this case the PageRank puts at the top certain procedures which are not very important from the software view point (e.g. *printk*). As a result it was proposed [14] to use in addition another ranking taking the network with inverse link directions in the adjacency matrix corresponding to $A_{ij} \to A^T = A_{ji}$ and constructing from it an additional Google matrix G^* according to relation (1) at the same α. The eigenvector of G^* with eigenvalue $\lambda = 1$ then gives a new inverse PageRank $P^*(i)$ with ranking index $K^*(i)$. This ranking was named CheiRank [15] to mark that it allows us to *chercher l'information* in a new way (which in English means *search the information* in a new way). Indeed, for the Linux PCN the CheiRank gives at the top more interesting and important procedures compared to the PageRank [14] (e.g. *start_kernel*). While the PageRank ranks the network nodes in average proportionally to a number of ingoing links, the CheiRank ranks nodes in average proportionally to a number of outgoing links. The physical meaning of PageRank vector components is that they give the probability to find a random surfer on a given node when a surfer follows the given directions of network links. In a similar way, the CheiRank vector components give the probability to find a random surfer on a given node when a surfer follows the inverted directions of network links. The inversion of links is a mathematical way to give a weight to outgoing links. We note that each directed network has both outgoing and ingoing links, and thus it is important to characterize these two complementary properties of information flow on directed networks. Since each node belongs both to CheiRank and PageRank vectors, the ranking of information flow on a directed network becomes two dimensional. We note that there have been earlier studies of PageRank of the Google matrix with inverted directions of links [16, 17], but no systematic analysis of statistical properties of 2DRanking was presented there.

An example of variation of PageRank probability $P(K)$ with K and CheiRank probability $P^*(K^*)$ with K^* is shown in figure 2, for the WWW network of University of Cambridge in

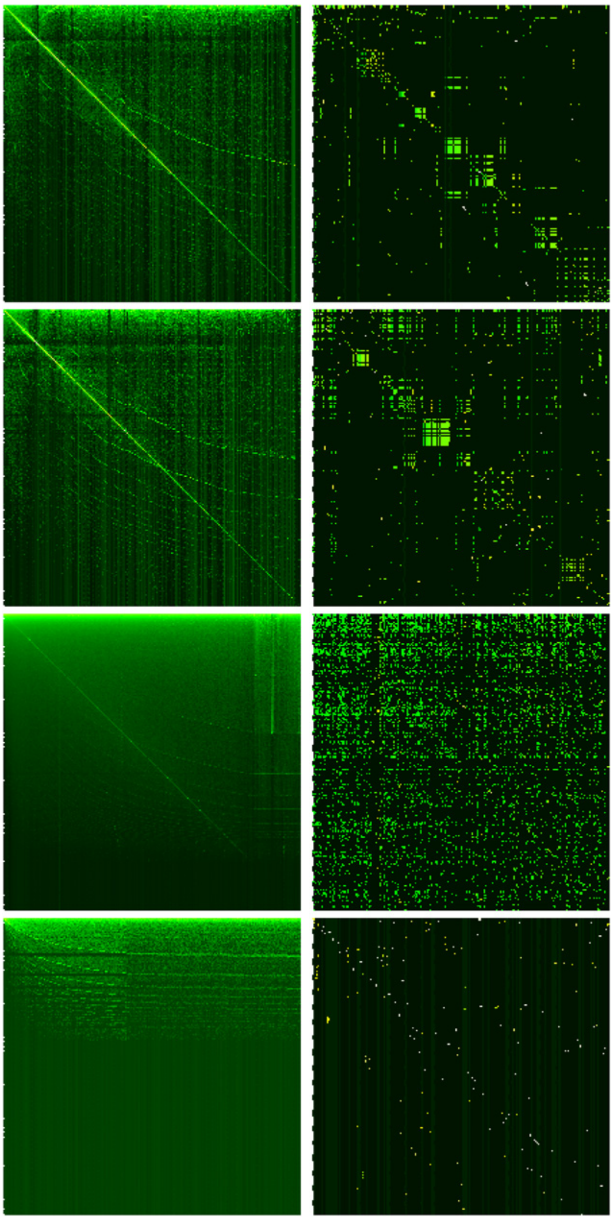

Figure 1. Google matrix gallery: all matrices are shown in the basis of PageRank index K (and K') of matrix $G_{KK'}$, which corresponds to x (and y) axis with $1 \leqslant K, K' \leqslant N$ (left column) and $1 \leqslant K, K' \leqslant 200$ (right column); all nodes are ordered by PageRank index K of matrix G and thus we have two matrix indexes K, K' for matrix elements in this basis. Left column: coarse-grained density of Google matrix elements $G_{K,K'}$ written in the PageRank basis $K(i)$ with indexes $j \to K(i)$ (in x-axis) and $i \to K'(i)$ (in a usual matrix representation with $K = K' = 1$ on the top-left corner); the coarse graining is done on 500×500 square cells for the networks of University of Cambridge 2006, University of Oxford 2006, Wikipedia English articles, PCN of Linux Kernel V2.6 (from top to bottom). Right column shows the first 200×200 matrix elements of G matrix at $\alpha = 0.85$ without coarse graining with the same order of panels as in the left column. Color shows the density of matrix elements changing from black for minimum value $((1 - \alpha)/N$ to white for maximum value via green and yellow (density is coarse grained in the left column).

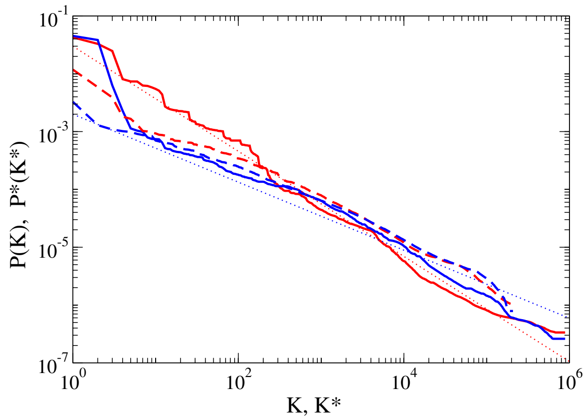

Figure 2. Dependence of probabilities of PageRank $P(K)$ (red/gray curve) and CheiRank $P^*(K^*)$ (blue/black curve) on corresponding ranks K and K^* for the network of University of Cambridge in 2006 (dashed curve) and in 2011 (full curve). The power-law dependences with the exponents $\beta \approx 0.91, 0.59$, corresponding to the relation $\beta = 1/(\mu - 1)$ with $\mu = 2.1, 2.7$, respectively, are shown by dotted straight lines.

years 2006 and 2011. Other examples for PCN Linux Kernel and Wikipedia can be found in [14, 15]. Detailed parameters of networks which we analyze in this paper and their sources are given in the appendix.

A detailed comparative analysis of PageRank and CheiRank two-dimensional classification was done in [15] for the example of the Wikipedia hyperlink citation network of English articles. It was shown that CheiRank highlights communicative property of nodes leading to a new way of two-dimensional ranking. While according to PageRank the top three countries are (1) USA, (2) UK and (3) France, CheiRank gives (1) India, (2) Singapore and (3) Pakistan as the most communicative Wikipedia country articles. The top 100 personalities of PageRank have the following percents in five main category activities: 58 (politics), 10 (religion), 17 (arts), 15 (science) and 0 (sport) [15]. Clearly, the significance of politicians is overestimated (many of them are USA presidents not broadly known to public). In contrast, CheiRank gives a more balanced distribution over these categories with 15, 1, 52, 16 and 16, respectively. It allows us to classify information in a new way finding composers, architects, botanists and astronomers who are not well known but who, for example, discovered a lot of Australian butterflies (*George Lyell*) or many asteroids (*Nikolai Chernykh*). These two people appear in the large listings of Australian butterflies and in the listing of asteroids (since they discovered many of them) and due to that they gain high CheiRank values. In a similar way, popular singers and musicians have long listings of their songs and music which increase their outgoing links and CheiRank. This shows that the information retrieval, which uses both PageRank and CheiRank, allows us to rank nodes not only by an amount of their popularity (how known is a given node) but also by an amount of their communicative property (how communicative is a given node). This 2DRanking was also applied to the brain model of the neuronal network [18] and the business process management network [19], and it was shown that it gives a new useful way of information treatment in these networks. The 2DRanking in the PageRank–CheiRank plane also naturally appears for the world trade network corresponding to import and export trade flows [20]. Thus, the 2DRanking based on PageRank and CheiRank paves the way for the development of 2D search engines which can become more intelligent than the present Google search based on the 1D PageRank algorithm.

In this work, we study the statistical properties of such a 2DRanking using examples of various real directed networks, including the WWW of British, French and Italian university networks [21], Wikipedia network [15], Linux Kernel networks [14, 22], gene regulation networks [23, 24] and other networks. The rest of the paper is organized as follows: in section 2, we study the properties of node density in the plane of PageRank and CheiRank; in section 3, the correlator properties between PageRank and CheiRank vectors are analyzed for various networks; information flow on the plane of PageRank and CheiRank is analyzed in section 4; the methods of control of SPAM outgoing links are discussed in section 5; 2DRanking applications for the gene regulation networks are considered in section 6 and the discussion of results is presented in section 7. The parameters of the networks and references on their sources are given in the appendix.

2. Node density of 2DRanking

A few examples of the Google matrix for four directed networks are shown in figure 1. There is a significant similarity in the global structure of G for the Universities of Cambridge and Oxford with well-visible hyperbolic curves (left column) even if at small scales the matrix elements are rather different (right column) in these two networks (see figure 1). Such hyperbolic curves are also visible in the Google matrix of Wikipedia (left column) even if here they are less pronounced due to much larger averaging inside the cells which contain about 15 times larger number of nodes (see network parameters in the appendix). We make a conjecture that the appearance of such curves is related to the existence of certain natural categories existing in the network, e.g., departments for universities or countries, cities, personalities etc for Wikipedia. We expect that there are relatively more links inside a given category compared to links between categories. However, this is only a statistical property, since on small scales at small K values the hyperbolic curves are not visible (right column in figure 1). Hence, more detailed studies are required to verify this conjecture. At small scale, the G matrix of Wikipedia is much more dense compared to the cases of Cambridge and Oxford (right column). We attribute such an increase of density of significant matrix elements to a stronger connectivity between nodes with large K in Wikipedia compared to the case of universities where the links have a more hierarchical structure. Partially this increase of density can be attributed to a larger number of links per node in the case of Wikipedia, but this increase by a factor 2.1 is not so strong and cannot explain all the differences of densities at small K scale. For Wikipedia, there are about 20% of nodes at the bottom of the matrix where there are almost no links. For PCN of Linux Kernel, this fraction becomes significantly larger with about 60% of nodes. The hyperbolic curves are still well visible for Linux PCN inside the remaining 40% of nodes. On a small scale, the density of matrix elements for Linux is rather small compared to the three previous cases. We attribute this to a much smaller number of links per node which is by factor 5 smaller for Linux compared to the university networks of figure 1 (see data in appendix).

The distributions of density of nodes $W(K, K^*) = dN_i/dKdK^*$ in the plane of PageRank and CheiRank in the logscale are shown for four networks of British universities in figure 3. Here, dN_i is a number of nodes in a cell of size $dKdK^*$ (see the detailed description in [15]). Even if the coarse-grained G matrices for Cambridge and Oxford look rather similar the density distributions in the (K, K^*) plane are rather different, at least at moderate values of K, K^*. The density distributions for all four universities clearly show that nodes with high PageRank have low CheiRank that corresponds to zero density at low K, K^* values. At large K, K^* values, there is a maximum line of density which is located not very far from the diagonal $K \approx K^*$. The presence of such a line should correspond to significant correlations between $P(K(i))$ and

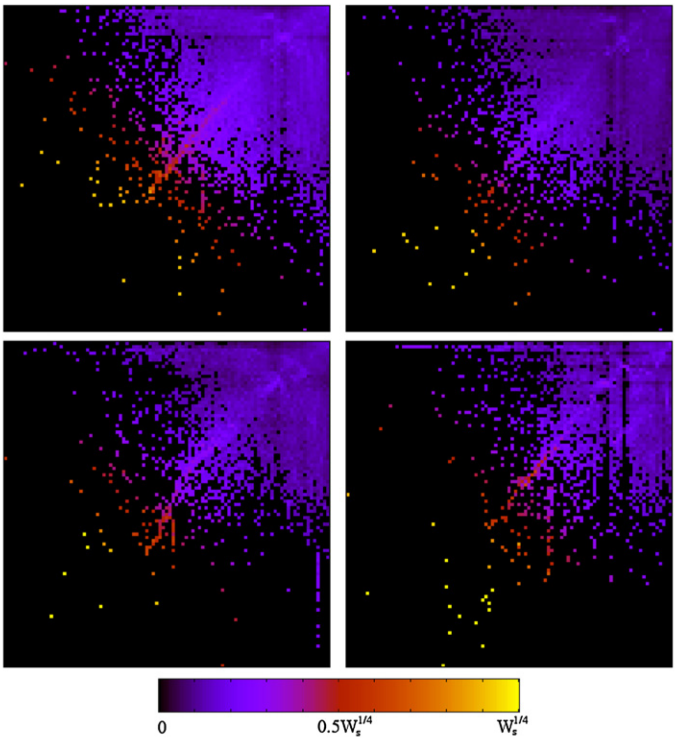

Figure 3. Density distribution $W(K, K^*) = dN_i/dKdK^*$ for networks of four British universities in the plane of PageRank K and CheiRank K^* indexes in the logscale $(\log_N K, \log_N K^*)$. The density is shown for 100×100 equidistant grid in $\log_N K, \log_N K^* \in [0, 1]$, the density is averaged over all nodes inside each cell of the grid, the normalization condition is $\sum_{K,K^*} W(K, K^*) = 1$. Color varies from black for zero to yellow for maximum density value W_M with a saturation value of $W_s^{1/4} = 0.5W_M^{1/4}$ so that the same color is fixed for $0.5W_M^{1/4} \leqslant W^{1/4} \leqslant W_M^{1/4}$ to show low densities in a better way. The panels show networks of University of Cambridge (2006) with $N = 212\,710$ (top left); University of Oxford with $N = 200\,823$ (top right); University of Bath with $N = 73\,491$ (bottom left); University of East Anglia with $N = 33\,623$ (bottom right). The axes show $\log_N K$ in the x-axis and $\log_N K^*$ in the y-axis, in both axes the variation range is $(0, 1)$.

$P^*(K^*(i))$ vectors that will be discussed in more detail in the next section. The presence of correlations between $P(K(i))$ and $P^*(K^*(i))$ leads to a probability distribution with one main maximum along a diagonal at $K - K^* = \text{const}$. This is similar to the properties of density distribution for the Wikipedia network discussed in [15] (see also the bottom-right panel in figure 13).

The density of nodes for Linux networks is shown in figure 4. In these networks, the density is homogeneous along lines $K + K^* = \text{const}$ that correspond to absence of correlations between $P(K(i))$ and $P^*(K^*(i))$. Indeed, in the absence of such correlations the distribution of nodes in the K, K^* plane is given by the product of independent probabilities. In the log-scale format used in figure 4, this leads to a homogeneous density of nodes in the top-right corner of the $(\log_N K, \log_N K^*)$ plane as it was discussed in [15, see right panel in figure 4]. Indeed, the distributions in figure 4 are very homogeneous inside the top-right triangle. We note that, a part from fluctuations, the distributions remain rather stable even if the size of the network is changed by factor 20 from the V2.0 to V2.6 version. The physical reasons for the absence of correlations between $P(i)$ and $P^*(i)$ have been explained in [14] on the basis of the concept of 'separation of concerns' used in software architecture. As discussed in [14], a good code should

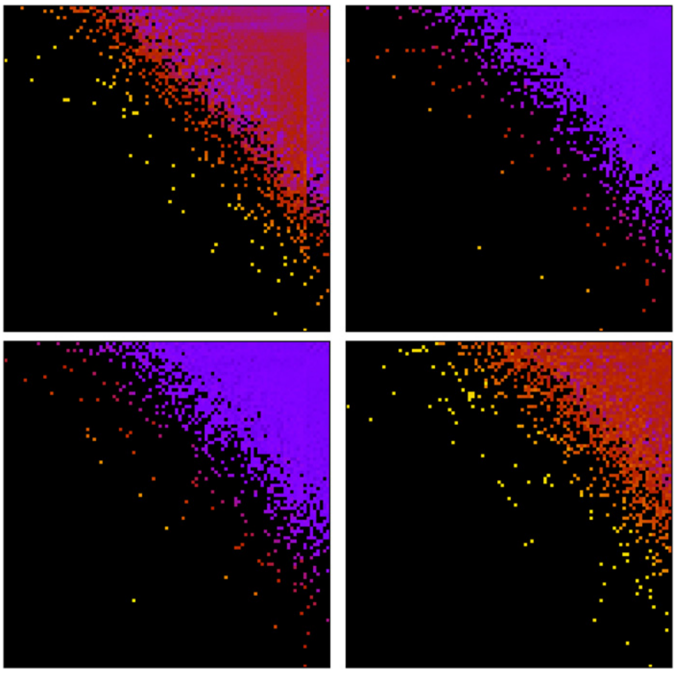

Figure 4. Density distribution $W(K, K^*) = dN_i/dKdK^*$ of four Linux Kernel networks shown in the same frame as in figure 3. The panels show networks for Linux versions V2.0 with $N = 14\,080$ (top left); V2.3 with $N = 41\,117$ (top right); V2.4 with $N = 85\,757$ (bottom left); V2.6 with $N = 285\,510$ (bottom right). Color panel is the same as in figure 3 with a saturation value of $W_s^{1/4} = 0.2W_M^{1/4}$ so that the same color is fixed for $0.2W_M^{1/4} \leqslant W^{1/4} \leqslant W_M^{1/4}$ to show low densities in a better way. The axes show $\log_N K$ in the x-axis and $\log_N K^*$ in the y-axis, in both axes the variation range is $(0, 1)$.

decrease a number of procedures that have high values of both PageRank and CheiRank; such procedures will play a critical role in error propagation since they are both popular and highly communicative at the same time. For example in the Linux Kernel, *do_fork()*, that creates new processes, belongs to this class. These critical procedures may introduce subtle errors because they entangle otherwise independent segments of code. The above observations suggest that the independence between popular procedures, which have high $P(K_i)$ and fulfil important but well-defined tasks, and communicative procedures, which have high $P^*(K_i^*)$ and organize and assign tasks in the code, is an important ingredient of well-structured software. We discuss the properties of PageRank–CheiRank correlations in the next section.

3. Correlations between PageRank and CheiRank

The correlations between PageRank and CheiRank can be quantitatively characterized by the correlator

$$\kappa(\tau) = N \sum_{i=1}^{N} P(K(i) + \tau)P^*(K^*(i)) - 1. \qquad (2)$$

Such a correlator was introduced in [14] for $\tau = 0$ and we will use the same notation $\kappa = \kappa(\tau = 0)$. This correlator at $\tau = 0$ shows if there are correlations and dependences between PageRank and CheiRank vectors. Indeed, for homogeneous vectors $P(K) = P^*(K^*) = 1/N$ we have $\kappa = 0$ corresponding to absence of correlations. We will see below that the values

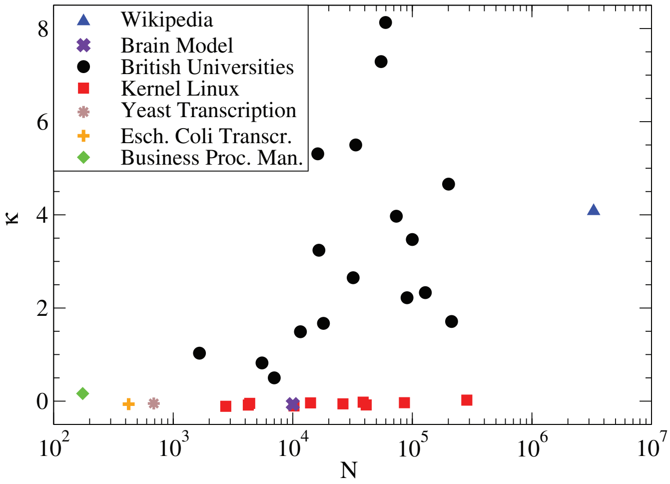

Figure 5. Correlator κ as a function of the number of nodes N for different networks: Wikipedia network, 17 British universities, ten versions of Kernel Linux Kernel PCN, *Escherichia Coli* and Yeast transcription gene networks, brain model network and business process management network. The parameters of networks are given in the appendix.

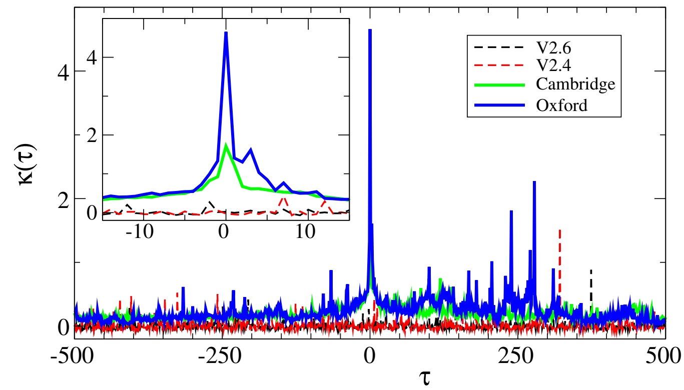

Figure 6. Correlator $\kappa(\tau)$ for two different long and short range of τ in the main and inset panel, respectively. The Kernel Linux PCN V2.6 and V2.4 are shown by dashed curves while universities networks of Cambridge and Oxford are shown by full curves.

of κ are very different for various directed networks. Hence, this new characteristic is able to distinguish various types of networks even if they have rather similar algebraic decay of PageRank and CheiRank vectors.

The values of κ for networks of various size N are shown in figure 5. The two types of networks are well visible according to these data. The human created university and Wikipedia networks have typical values of κ in the range $1 < \kappa < 8$. Other networks like Linux PCN, gene transcription networks, brain model and business process management networks have $\kappa \approx 0$.

The dependence of $\kappa(\tau)$ on the correlation 'time' τ is shown in figure 6. For the PCN of Linux there are no correlations at any τ, while for the university networks we find that the correlator drops to small values with increase of $|\tau|$ (e.g. $|\tau| > 5$) even if at certain rather large values of $|\tau|$ significant values of correlator κ can reappear.

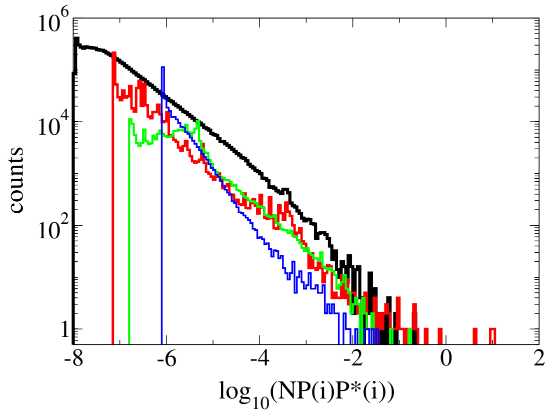

Figure 7. Histogram of frequency appearance of correlator components $\kappa_i = NP(K(i))P^*(K^*(i))$ for networks of Wikipedia (black), University of Cambridge in 2006 (green) and in 2011 (red), and PCN of Linux Kernel V2.6 (blue). For the histogram the whole interval $10^{-8} \leqslant \kappa_i \leqslant 10^2$ is divided into 200 cells of equal size in the logarithmic scale. Curve colors are black, red, green and blue from left to right at the bottom of the vertical axis.

It is interesting to see what are typical values $\kappa_i = NP(K(i))P^*(K^*(i))$ of contributions in the correlator sum (2) at $\tau = 0$. The distribution of κ_i values for a few networks are shown in figure 7. All of them follow a power law with an exponent a =1.23 for PCN Linux, 0.70 for Wikipedia and 0.76 (2006) and 0.66 (2011) for University of Cambridge. We note that further studies are required to analytically obtain the values of the exponent a. In the latter two cases the exponent and the distribution shape remains stable in time; however, in 2011 there appear few nodes with very large κ_i values which give a significant increase of the correlator from $\kappa = 1.71$ (in 2006) up to $\kappa = 30.0$ (in 2011). It is possible that such a situation can appear if it is imposed that practically any page points to the main university page, which may have a rather high CheiRank due to many outgoing links to other departments and university divisions. We suppose that these are also the reasons why we have the appearance of large values of $\kappa(\tau)$ in university networks. At the same time more detailed studies are required to clarify the correlation properties on directed networks of a deeper level. We will return in section 7 to a discussion of university networks collected in 2011.

Another way to analyze the correlations between PageRank and CheiRank is simply to count the number of nodes $\Delta(n)$ inside a square $1 \leqslant K(i), K^*(i) \leqslant n$. For a totally correlated distribution with $K(i) = K^*(i)$ we have $\Delta(n)/N = n/N$, while in absence of correlations we should have points homogeneously distributed inside a square $n \times n$ that gives $\Delta(n)/N = (n/N)^2$. The dependence of such point-count correlator $\Delta(n)$ on size n is displayed in figure 8 for various networks. These data clearly show that the Linux PCN is uncorrelated being close to the limiting uncorrelated dependence, while Wikipedia and British university networks show intermediate strength of correlations being between the two limiting functions of $\Delta(n)$.

4. Information flow of 2DRanking

According to 2DRanking, all network nodes are distributed on a two-dimensional plane (K, K^*). The directed links of the network create an information flow in this plane. To visualize this flow, we use the following procedure:

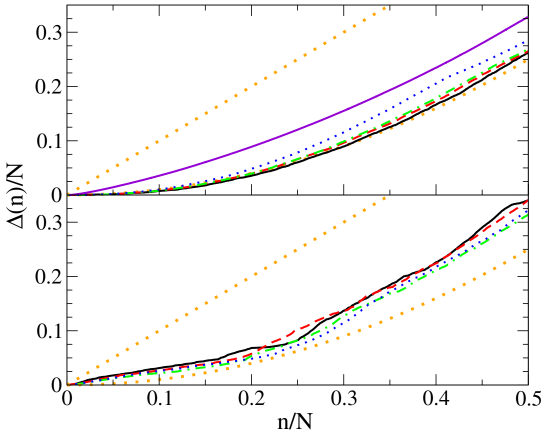

Figure 8. Dependence of the point-count correlation function $\Delta(n)/N$ on n/N for networks of Wikipedia, British universities and Kernel Linux PCN. The curves in the top panel show the cases of Wikipedia (solid violet/gray) and four versions of PCN of Linux Kernel with V2.0 (solid black), V2.3 (dashed red), V2.4 (dot-dashed green) and V2.6 (dotted blue). The curves in the bottom panel show the cases of British universities with East Anglia (solid black), Bath (dashed red), Oxford (dot-dashed green) and Cambridge 2006 (dotted blue). Dotted orange curves represent the totally correlated case with $\Delta(n)/N = n/N$ and the totally uncorrelated one with $\Delta(n)/N = (n/N)^2$.

(a) each node is represented by one point in the (K, K^*) plane;
(b) the whole space is divided into equal size cells with indexes (i, i^*) with the number of nodes inside each cell being n_{i,i^*}, in figure 9 we use cells of equal size in usual (left column) and logarithmic (right column) scales;
(c) for each node inside the cell (i, i^*), pointing to any other cell $(i', i^{*\prime})$, we compute the vector $(i' - i, i^{*\prime} - i^*)$ and average it over all nodes n_{i,i^*} inside the cell (the weight of links is not taken into account);
(d) we put an arrow centered at (i, i^*) with the modulus and direction given by the average vector computed in (c).

Examples of such average flows for the networks of figure 1 are shown in figure 9. All flows have a fixed point attractor. The fixed point is located at rather large values $K, K^* \sim N/4$, that is, due to the fact that in average nodes with maximal values $K, K^* \sim N$ point to lower values. At the same time nodes with very small $K, K^* \sim 1$ still point to some nodes which have larger values of K, K^* that places the fixed point at certain intermediate K, K^* values. We note that the analyzed directed networks have dangling nodes which have no outgoing links, the fraction of such nodes is especially large for the Linux network. Due to the absence of outgoing links, we obtain an empty white region in the information flow shown in figure 9. A more detailed analysis of statistical properties of information flows on the PageRank–CheiRank plane requires further study.

5. Control of spam links

For many networks, ingoing and outgoing links have their own importance and thus should be treated on equal grounds by PageRank and CheiRank as described above. However, for the WWW it is more easy to manipulate outgoing links which are handled by an owner of a

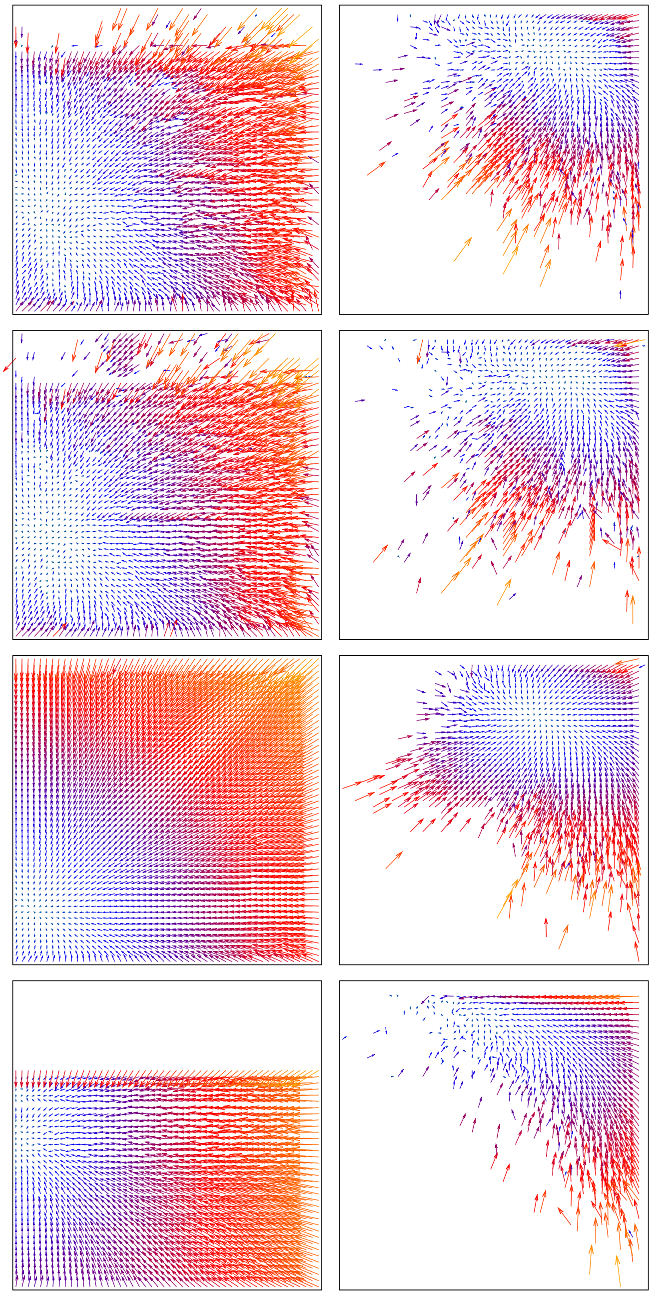

Figure 9. Information flow on the PageRank–CheiRank plane (K, K^*) generated by directed links of the networks of figure 1. Outgoing links flow is shown in the linear scale (K, K^*) with $K, K^* \in [1, N]$ on left panels, and in the logarithmic scale $(\log_N K, \log_N K^*)$ for $\log_N K, \log_N K^* \in [0, 1]$ on right panels. The flow is shown by arrows whose size is proportional to the vector amplitude, which is also indicated by color [from yellow for large to blue for small amplitudes]. The rows corresponds to University of Cambridge (2006); University of Oxford (2006), Wikipedia English articles, PCN of Linux Kernel V2.6 (from top to bottom). The axes show: on the left column K/N in the x-axis, K^*/N in the y-axis; on the right column $\log_N K$ in the x-axis, $\log_N K^*$ in the y-axis; in all axes the variation range is $(0, 1)$.

given webpage, while ingoing links are handled by other users. This requires the introduction of some level of control on the outgoing links which should be taken into account for the ratings. Since it is very easy to create links to highly popular sites, we will call 'spam links' links for which the destination site is much more popular than the source. A quantitative measure of popularity can be provided by the PageRank of the sites. We do not think that spam links are frequent in networks such as procedure calls in the Linux kernel, Wikipedia and gene regulation. Even for university networks we think that there is not much reason to put spam links inside the university domain. However, for a large-scale WWW an excessive number of such spam links can become harmful for the network performance. However, for WWW networks spam links are probably more widespread. Some websites may try to improve their rating by carefully choosing their outgoing links. Also it is a common policy to have links back to a website's root pages to facilitate navigation. Naturally, a good rating should not be sensitive to the presence of such links. Thus it is important to treat spam links appropriately in order to construct a two-dimensional web-search engine. Below we propose a method for spam links control and test it on an example of the Wikipedia network which has the largest size among networks analyzed in this paper. We stress that this is done as a test example and not because we think that there are spam links between Wikipedia articles.

With this aim, we propose the following filter procedure for computation of CheiRank. The standard procedure described above is to invert the directions of all links of the network and then to compute the CheiRank. The filter procedure inverts a link from j to i only if $\eta P(K(j)) > P(K(i))$, where η is some positive filter parameter. After such an inversion of certain links, while other links remain unchanged, the matrix S^* and G^* are computed and the CheiRank vector $P^*(K^*(i))$ of G^* is determined in a usual way. From the definition it is clear that for $\eta = 0$ there are no inverted links, and thus after filtering P^* is the same as the PageRank vector P. In the opposite limit $\eta = \infty$ all links are inverted and P^* is then the usual CheiRank discussed in previous sections. Thus intermediate values of η allow us to handle the properties of CheiRank depending on a wanted strength of filtering. We note that the proposed filtering procedure is rather generic and can be applied to various types of directed networks.

The dependence of the fraction f of inverted links (defined as a ratio between the number of inverted links to the total number of links) on the filter parameter η is shown for various networks in figure 10. There is a significant jump of f at $\eta \approx 1$ for British university networks. In fact the condition $\eta \approx 1$ corresponds approximately to the border relation $P(K) \approx P(K')$ with $K \approx K'$ that marks the diagonal of the G matrix shown in figure 1, which has a significant density of matrix elements. As a result for $\eta > 1$, we have a significant increase of inversion of links leading to a jump of f present in figure 10. The diagonal density is most pronounced for university networks so that for them the jump of f is mostly sharp.

It is also convenient to consider another condition for link inversion defined not for $P(K_i)$ but directly in the plane (K, K') defined by the condition: links are inverted only if $K(j) < \eta_K K(i)$ (where node j points to node i, $j \to i$). In a first approximation, we can assume that the links are homogeneously distributed in the plane of transitions from K to K'. This density is similar to the density distribution of Google matrix elements $G_{K'K}$ shown in figure 1. For the homogeneous distribution, the fraction f of inverted links is given by an area $\eta_K/2$ of a triangle, whose height is 1 and the basis is η_K, for $\eta_K \leqslant 1$. In a similar way, we have $f = 1 - 1/2\eta_K$ for $\eta_K \geqslant 1$. We can generalize this distribution assuming that there are only links with $1 \leqslant K' \leqslant aN$, that is, approximately the case for Linux network where $a = 0.4$ (see figure 1 bottom row), and that inside this interval the density of links decreases as $1/(K')^{\nu}$. Then after computing the area we obtain the expression for the fraction of inverted links valid

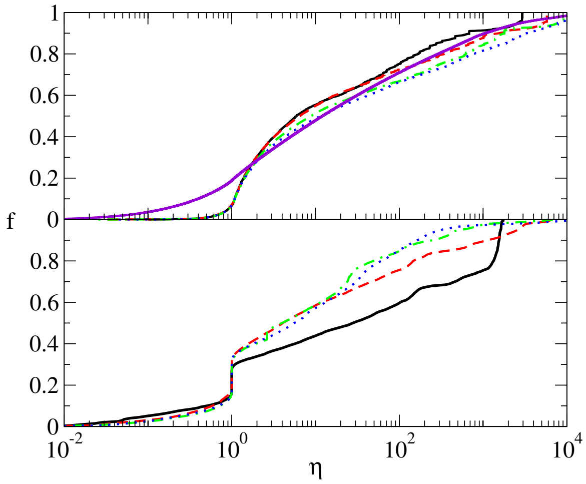

Figure 10. Fraction f of inverted links as a function of filter parameter η for various studied networks. Top panel: Wikipedia (violet/gray curve) and four versions of Kernel Linux PCN with V2.0 (solid black curve), V2.3 (dashed red curve), V2.4 (dot-dashed green curve) V2.6 (dotted blue curve). Bottom panel shows data for British university networks with East Anglia (solid black curve), Bath (dashed red curve), Oxford (dot-dashed green curve) and Cambridge 2006 (dotted blue curve).

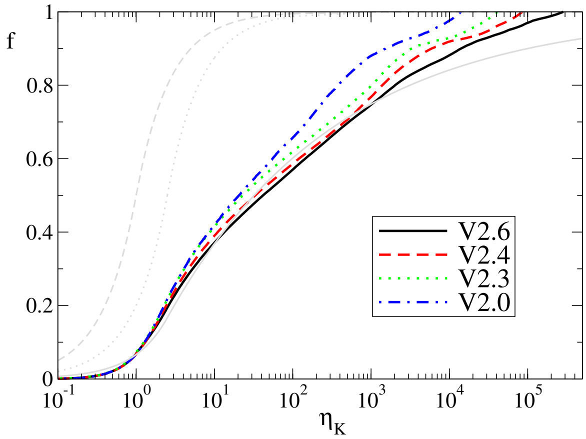

Figure 11. Fraction f of inverted links in the (K, K') plane with the condition $K(j) < \eta_K K(i)$ shown as a function of filter parameter η_K for Linux networks versions shown by different curves. Gray curves from left to right are the theory curves with $a = 1$, $\nu = 0$ (dashed); $a = 0.4$, $\nu = 0$ (dotted) and $a = 0.4$, $\nu = 0.8$ (full) (see text).

for $0 \leqslant \nu < 1$:

$$f(\eta_K) = \begin{cases} \frac{1-\nu}{2-\nu}(a\eta_K) & \eta_K \leqslant 1/a \\ 1 + \left(\frac{1-\nu}{2-\nu} - 1\right)(a\eta_K)^{\nu-1} & \eta_K > 1/a \end{cases}. \quad (3)$$

The comparison of this theoretical expression with the numerical data for Linux PCN is shown in figure 11. It shows that the data for Linux are well described by the theory (3) with $a = 0.4$

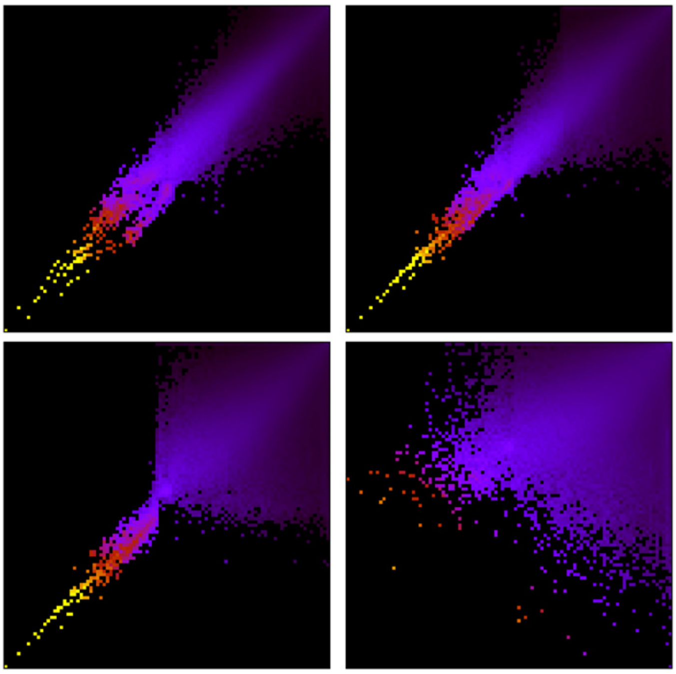

Figure 12. Density distribution $W(K, K^*) = dN_i/dKdK^*$ for Wikipedia in the plane of PageRank and filtered CheiRank indexes, $(\log_N K, \log_N K^*)$, in a equidistant 100×100 lattice with $\log_N K, \log_N K^* \in [0, 1]$. The filter parameter is $\eta = 10$ (left-top panel), 100 (right-top panel), 1000 (left-bottom panel), 10^5 where all links are inverted (right-bottom panel). The color panel is the same as in figure 3 with the saturation value $W_s^{1/4} = 0.5W_M^{1/4}$. The axes show: $\log_N K$ in the x-axis, $\log_N K^*$ in the y-axis, in both axes the variation range is $(0, 1)$.

and $\nu = 0.8$. The last value takes into account the fact that the density of links decreases with PageRank index K' as it is well visible in figure 1.

The variation of nodes density in the plane of PageRank and filtered CheiRank (K, K^*) for the Wikipedia network is shown in figure 12 with the filtering by η for $P(K)$ and $P(K')$ values. At moderate values $\eta = 10$ the density is concentrated near the diagonal, with further increase of $\eta = 100, 1000$ a broader density distribution appears at large K values which goes to smaller and smaller K until the limiting distribution without filtering is established at very large η. The top 100 Wikipedia articles obtained with filtered CheiRank at the above values of η are given at [25]. We also give there top articles in 2DRank which gives articles in order of their appearance on the borders of a square of increasing size in (K, K^*) plane (see the detailed description in [15]). These data clearly show that filtering eliminates articles with many outgoing links and gives a significant modification of top CheiRank articles. Thus the described method can be efficiently used for control of spam links present in the WWW.

6. 2DRanking of gene regulation networks

The method of 2DRanking described above is rather generic and can be applied to various types of directed networks. Here, we apply it to gene regulation networks of *Escherichia Coli* and Yeast with the network links taken from [24]. Such transcription regulation networks control the expression of genes and have important biological functions [23].

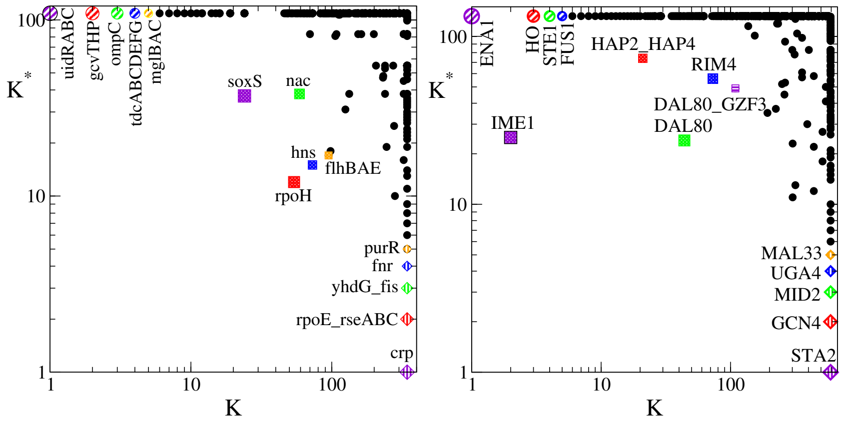

Figure 13. Distribution of nodes in the plane of PageRank K and CheiRank K^* for *Escherichia Coli* and Yeast transcription networks on left and right panels, respectively (network data are taken from [24]). The nodes with the five top probability values of PageRank, CheiRank and 2DRank are labeled by their corresponding node names; they correspond to five lowest index values.

The distribution of nodes in PageRank–CheiRank plane is shown in figure 13. The top five nodes in CheiRank probability value (lowest CheiRank indexes) are those which send many outgoing orders, the top five in PageRank probability value are those which obtain many incoming signals and the top five indexes in 2DRank (with five lowest 2DRank index values) combine these two functions. For these networks the correlator κ is close to zero (even slightly negative), which indicates the statistical independence between outgoing and ingoing links quite similarly to the case of the PCN for the Linux Kernel. This may indicate that a slightly negative correlator κ is a generic property for the data flow network of control and regulation systems. We use these networks here to show that the general methods proposed above can be applied to these directed networks as well. Whether the obtained ratings can bring deep insights into the functioning of gene regulation can only be assessed by experts in the field. However, we hope that such an analysis will prove to be useful for a better understanding of gene regulation networks.

7. Discussion

Above we presented extensive studies of statistical properties of 2DRanking based on PageRank and CheiRank for various types of directed networks. All studied networks are of a free-scale type with an algebraic distribution of ingoing and outgoing links with a usual value of exponents. In spite of that their statistical characteristics related to PageRank and CheiRank are rather different. Some networks have high correlators between PageRank and CheiRank (e.g. Wikipedia, British universities), while others have practically zero correlators (PCN of Linux Kernel, gene regulation networks). The distribution of nodes in PageRank–CheiRank plane also varies significantly between different types of networks. Thus 2DRanking discussed here gives more detailed classification of information flows on directed networks.

We think that 2DRanking gives new possibilities for information retrieval from large databases which are growing rapidly with time. Indeed, for example the size of the Cambridge network increased by a factor 4 from 2006 to 2011 (see appendix and figure 2). At present, web robots start automatically generating new webpages. These features can be responsible for the appearance of gaps in the density distribution in the (K, K^*) plane at large $K, K^* \sim N$ values visible for large-scale university networks of Cambridge and ENS Paris in 2011 (see

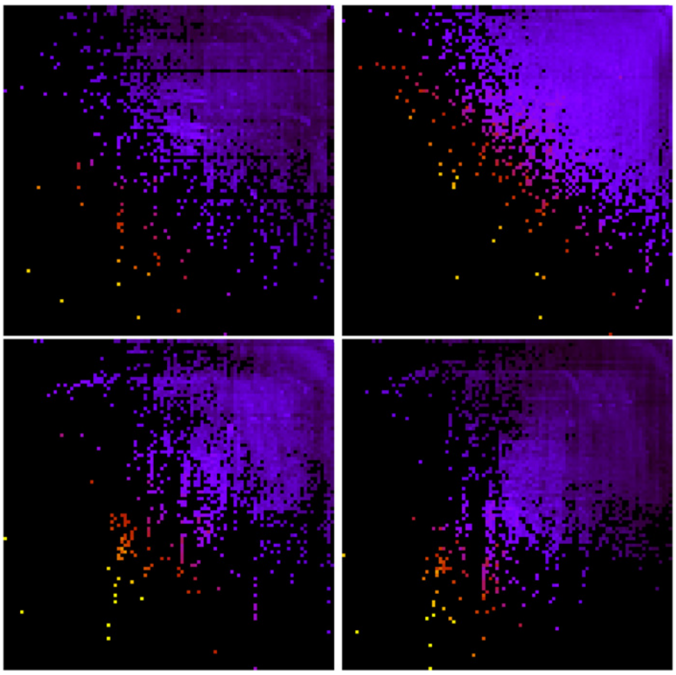

Figure 14. Density distribution $W(K, K^*) = dN_i/dKdK^*$ shown in the same frame as in figure 3 for networks collected in 2011: University of Cambridge (top left), University of Bologna (top right), ENS Paris for crawling level 5 (bottom left) and 7 (bottom right). The color panel is the same as in figure 3 with the saturation value $W_s^{1/4} = 0.5W_M^{1/4}$. The axes show: $\log_N K$ in the x-axis, $\log_N K^*$ in the y-axis, is both axes the variation range is $(0, 1)$.

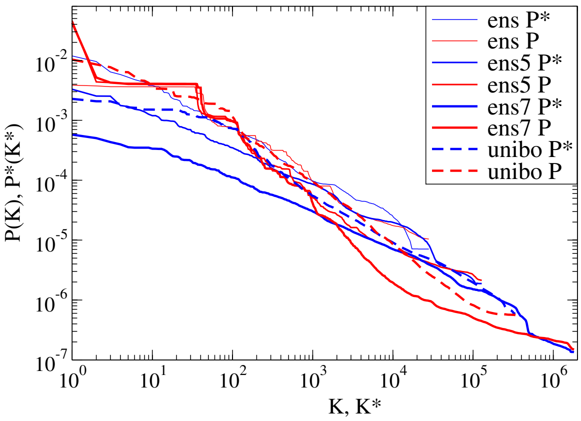

Figure 15. Dependence of probabilities of PageRank $P(K)$ (red/gray curve) and CheiRank $P^*(K^*)$ (blue/black curve) on corresponding ranks K and K^* for the networks of ENS Paris (crawling levels 3,5,7) and the University of Bologna.

figure 14). Such an automatic generation of links can change the scale-free properties of networks. Indeed, for ENS Paris we observe the appearance of a large step in the PageRank distribution $P(K)$ shown in figure 15. This step for $P(K)$ remains not sensitive to the deepness of crawling which goes on a level of 3, 5 and 7 links. However, the CheiRank distribution changes with the deepness level becoming more and more flat (see figure 15). Such a tendency in a modification of network statistical properties is visible in 2011 for large-size university networks, while networks of moderate size, like the University of Bologna 2011 (see data in figures 14 and 15), are not yet affected. A sign of ongoing changes is a significant growth of the correlator value κ which increases up to a very large value (30 for Cambridge 2011 and 63 for ENS Paris). There is a danger that automatic generation of links can lead to a delocalization transition of PageRank that can destroy efficiency of information retrieval from the WWW. We note that it is known that PageRank delocalization can appear in certain models of Markov chains and Ulam networks [26, 27] (see e.g. in [26] figure 1 (right-top panel) and figure 6 directly showing the delocalization of PageRank vector). Such a delocalization of PageRank would make the ranking of nodes inefficient due to high sensitivity of ranking to fluctuations that would create a very dangerous situation for the WWW information retrieval and ranking. We also note that the spectrum of the Google matrix of British universities networks has been recently analyzed in [28]. The spectrum and eigenstates analysis can be a sensitive tool for location of precursors of a delocalization transition.

Our studies of 2DRanking pave the way to the development of two-dimensional search engines which will use the advantages of both PageRank and CheiRank. Indeed, the Google search engine uses as the fundamental mathematical basis the one-dimension ranking related to PageRank [7]. Of course, there are various other important elements used by the Google search which remain the company secret, and not only PageRank order matters for the Google ranking. However, the mathematical aspects of these additional elements are not really known (e.g. they are not described in [7]). At the same time, the size of databases generated by the modern society continues its enormous growth. Due to that, the information retrieval and ordering of such datasets becomes of primary importance and new mathematical tools should be developed to operate and characterize efficiently their information flows and ranking. Here we proposed and analyzed the properties of the new two-dimensional search engine, which we call Dvvadi from Russian ‘dva (two)’ and ‘dimension’ that will use the complementary ranking abilities of both PageRank and CheiRank. Now the procedure of ordering of all network nodes uses not one but two vectors of the Google matrix of a network. The computational efforts are twice as expensive but for that we obtain a new quality, since now the nodes are ranked in the 2D plane not only by their degree of popularity but also by their degree of communicability. Thus for the Wikipedia network the top three articles in PageRank probability are three countries (most popular), while the top three articles in CheiRank probability are three listings of knowledge, state leaders and geographical places (most communicative). Hence, we can rank the nodes of the network in a new two-dimensional manner which highlight complementary properties of node popularity and communicability. Thus, the Dvvadi search can present nodes not in a line but on a 2D plane characterizing these two complementary properties of nodes. Examples of such 2D representation of nodes selected from Wikipedia articles by a specific subject are shown in figure 16: we determine global K and K^* indexes of all articles, select a specific subject (e.g. *countries*) and then represent countries in the local index K and K^* corresponding to their appearance in the global order via PageRank and CheiRank. For countries, we see a clear tendency that the countries on the top of PageRank probability (low K) have relatively high CheiRank index (high K^*) (e.g. US, UK, France) while small countries in the region $K \approx 50$, $K^* \approx 10$ have another tendency (e.g. Singapore). We attribute this to specific routes of cultural and industrial development of the world: e.g. Singapore was a colony of UK and became a strong trade country and due to that has historically many links pointing to the UK and other developed countries. For universities we also see that those at the top of PageRank (Harvard, Oxford and Cambridge) are not very communicative having high K^* values, while Columbia and Berkeley are more balanced, and Florida and FSU are very communicative probably due to the initial location of the Wikimedia Foundation

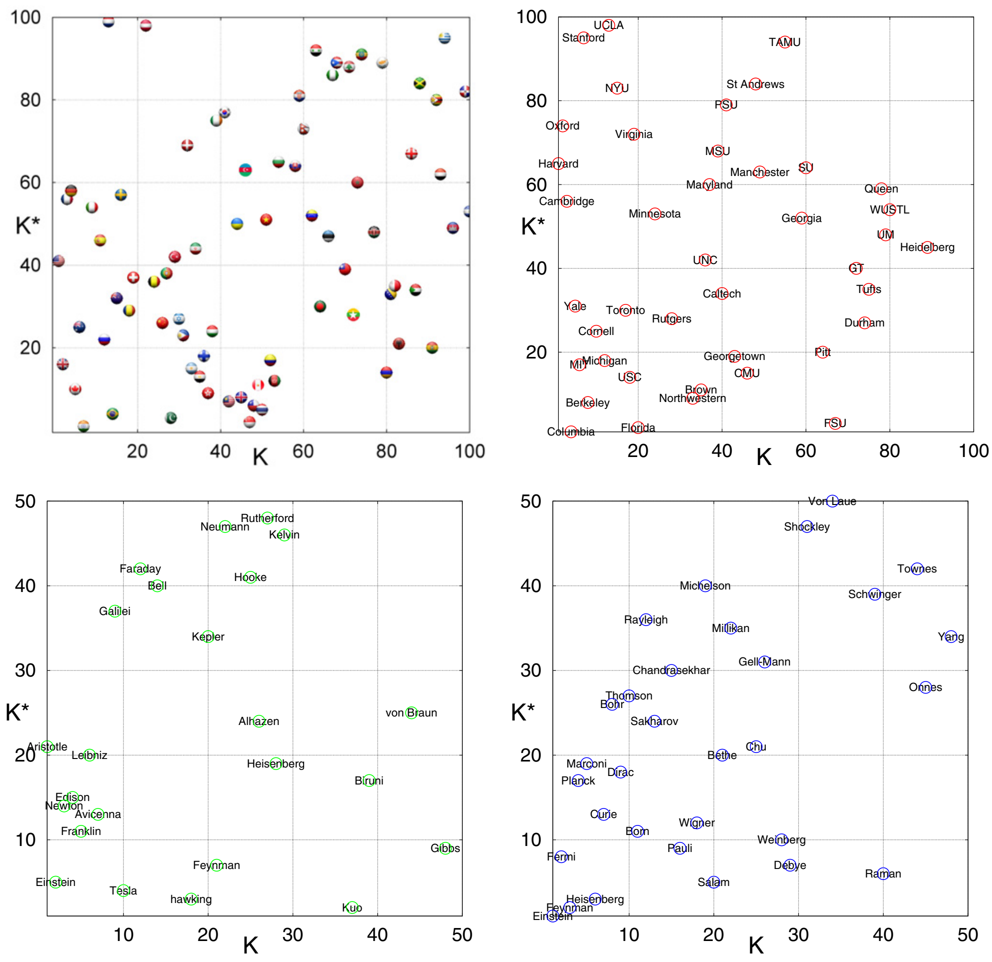

Figure 16. Examples of Dvvadi search analysis of Wikipedia articles shown on the 2D plane of PageRank K and CheiRank K^* local indexes for specific subjects (articles): countries marked by their flag (top left), universities (top right), physicists (bottom left), Nobel laureates in physics (bottom right), circles mark the node location; high resolution figures and listings of names with local (K, K^*) values in 100×100 square are available at [25] (listings with global ranking are available at [15]).

at Florida. For physicists, we see that links to many scientific fields (like Shen Kuo) or popularization of science (like Hawking and Feynman) place those people at the top positions of CheiRank. In a similar way, for the Nobel laureates in physics we see that CheiRank stresses the communicative aspects: e.g. Feynman, due to his popularization of physics; Salam, due to the institute with his name at Trieste, with a broad international activity; Raman, due to the Raman effect.

On the basis of the above results, we think that PageRank–CheiRank classification of network nodes on 2D plane will allow us to analyze the information flows on directed networks in a better way. It is also important to note that 2DRanking is very natural for financial and trade networks. Indeed, the world trade usually uses the import and export ranking which is analogous to PageRank and CheiRank, as it is shown in [20]. We think that such Dvvadi

Table A1. Linux Kernel network parameters

Version	N	N_{links}	κ
V1.0	2 752	5 933	$\kappa = -0.11$
V1.1	4 247	9 710	$\kappa = -0.083$
V1.2	4 359	10 215	$\kappa = -0.048$
V1.3	10 233	24 343	$\kappa = -0.102$
V2.0	14 080	34 551	$\kappa = -0.037$
V2.1	26 268	59 230	$\kappa = -0.058$
V2.2	38 767	87 480	$\kappa = -0.022$
V2.3	41 117	89 355	$\kappa = -0.081$
V2.4	85 757	195 106	$\kappa = -0.034$
V2.6	285 510	588 861	$\kappa = 0.022$

Table A2. British universities network parameters

University	N	N_{links}	κ
RGU (Abardeen)	1 658	15 295	$\kappa = 1.03$
Uwic (Wales)	5 524	111 733	$\kappa = 0.82$
NTU (Nottingham)	6 999	143 358	$\kappa = 0.50$
Liverpool	11 590	141 447	$\kappa = 1.49$
Hull	16 176	236 525	$\kappa = 5.31$
Keele	16 530	117 944	$\kappa = 3.24$
UCE (Birmingham)	18 055	351 227	$\kappa = 1.67$
Kent	31 972	277 044	$\kappa = 2.65$
East Anglia	33 623	325 967	$\kappa = 5.50$
Sussex	54 759	804 246	$\kappa = 7.29$
York	59 689	414 200	$\kappa = 8.13$
Bath	73 491	541 351	$\kappa = 3.97$
Glasgow	90 218	544 774	$\kappa = 2.22$
Manchester	99 930	1254 939	$\kappa = 3.47$
UCL (London)	128 450	1397 261	$\kappa = 2.33$
Oxford	200 823	1831 542	$\kappa = 4.66$
Cambridge (2006)	212 710	2015 265	$\kappa = 1.71$

engine/motor [25] will find useful applications for the treatment of enormously large databases created by modern society.

Acknowledgments

We thank K M Frahm and B Georgeot for useful discussions of properties of British university networks. This work is done in the frame of the EC FET Open project 'New tools and algorithms for directed network analysis' (NADINE No 288 956).

Appendix

We list below the directed networks used in this work giving for them number of nodes N, number of links N_{links} and correlator between PageRank and CheiRank κ. Additional data can be find at [25].

Linux Kernel PCNs are taken from [14] (see also [22]) with the parameters for various kernel versions shown in table A1.

Web networks of British universities dated by year 2006 are taken from [21] and are shown in table A2.

We also developed a special code with which we performed crawling of university web networks in January—March 2011 with the parameters given below: University of Cambridge (2011) with $N = 898\,262$, $N_{\text{links}} = 15\,027\,630$, $\kappa = 30.0$; École Normale Supérieure, Paris (ENS 2011) with $N = 28\,144$, $N_{\text{links}} = 971\,856$, $\kappa = 1.67$ (crawling deepness level of three links), $N = 129\,910$, $N_{\text{links}} = 2111\,944$, $\kappa = 16.2$ (crawling deepness level of five links), $N = 1820\,015$, $N_{\text{links}} = 25\,706\,373$, $\kappa = 63.6$ (crawling deepness level of seven links); University of Bologna with $N = 339\,872$, $N_{\text{links}} = 16\,345\,488$, $\kappa = 2.63$.

The data for the hyperlink network of Wikipedia English articles (2009) are taken from [15] with $N = 3282\,257$, $N_{\text{links}} = 71\,012\,307$, $\kappa = 4.08$.

Transcription gene networks are taken from [24]. We have for them: *Escherichia Coli* with $N = 423$, $N_{\text{links}} = 519$, $\kappa = -0.0645$; Yeast with $N = 690$, $N_{\text{links}} = 1079$, $\kappa = -0.0497$; for all links the weight is taken to be the same.

Business process management network is taken from [19] with $N = 175$, $N_{\text{links}} = 240$, $\kappa = 0.164$.

Brain model network is taken from [18] with $N = 10\,000$, $N_{\text{links}} = 1960\,108$, $\kappa = -0.054$ (unweighted), $\kappa = -0.065$ (weighted).

References

[1] See e.g. The Size of the World Wide Web www.worldwidewebsize.com/
[2] Wikipedia (The Free Encyclopedia) 2011 Facebook http://en.wikipedia.org/wiki/Facebook
[3] Wikipedia (The Free Encyclopedia) 2011 Vkontakte http://en.wikipedia.org/wiki/VK_(social_network)
[4] Wikipedia (The Free Encyclopedia) 2011 Web search engine http://en.wikipedia.org/wiki/Web_search_engine
[5] Büttcher S, Clarke C L A and Cormack G V 2010 *Information Retrieval: Implementing and Evaluating Search Engines* (Cambridge: MA: MIT Press)
[6] Brin S and Page L 1998 *Comput. Netw. ISDN Syst.* **30** 107
[7] Langville A M and Meyer C D 2006 *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton, NJ: Princeton University Press)
[8] Franceschet M 2010 PageRank: Standing on the shoulders of giants arXiv:1002.2858v3 [cs.IR]
[9] Vigna S 2011 Spectral Ranking (available at http://vigna.dsi.unimi.it/ftp/papers/SpectralRanking.pdf)
[10] Litvak N, Scheinhardt W R W and Volkovich Y 2008 *Lecture Notes Comput. Sci.* **4936** 72
[11] Donato D, Laura L, Leonardi S and Millozzi S 2004 *Eur. Phys. J.* B **38** 239
[12] Pandurangan G, Raghavan P and Upfal E 2005 *Internet Math.* **3** 1
[13] Georgeot B, Giraud O and Shepelyansky D L 2010 *Phys. Rev.* E **81** 056109
[14] Chepelianskii A D 2010 arXiv:1003.5455 [cs.SE] (www.quantware.ups-tlse.fr/QWLIB/linuxnetwork/)
[15] Zhirov A O, Zhirov O V and Shepelyansky D L 2010 *Eur. Phys. J.* B **77** 523 (www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/)
[16] Fogaras D 2003 Where to start browsing the Web? (*Lecture Notes in Computer Science* vol 2877) p 65
[17] Hrisitidis V, Hwang H and Papakonstantinou Y 2008 *ACM Trans. Database Syst.* **33** 1
[18] Shepelyansky D L and Zhirov O V 2010 *Phys. Lett.* A **374** 3206
[19] Abel M and Shepelyansky D L 2011 *Eur. Phys. J.* B **84** 493 (www.quantware.ups-tlse.fr/QWLIB/cheirankbusiness/)
[20] Ermann L and Shepelyansky D L 2011 *Acta Phys. Pol.* **120** A158 (arXiv:1103.5027 [q-fin.GN]) (www.quantware.ups-tlse.fr/QWLIB/tradecheirank/)
[21] Academic Web Link Database Project http://cybermetrics.wlv.ac.uk/database/
[22] Ermann L, Chepelianskii A D and Shepelyansky D L 2011 *Eur. Phys. J.* B **79** 115
[23] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 *Science* **298** 824
[24] Uri Alon Lab www.weizmann.ac.il/mcb/UriAlon/ (Complex networks section)
[25] Ermann L, Chepelianskii A D and Shepelyansky D L 2012 arXiv:1106.6215 [cs.IR] (www.quantware.ups-tlse.fr/QWLIB/dvvadi/)
[26] Shepelyansky D L and Zhirov O V 2010 *Phys. Rev.* E **81** 036213
[27] Ermann L and Shepelyansky D L 2010 *Phys. Rev.* E **81** 036221
[28] Frahm K M, Georgeot B and Shepelyansky D L 2011 *J. Phys. A: Math. Theor.* **44** 465101