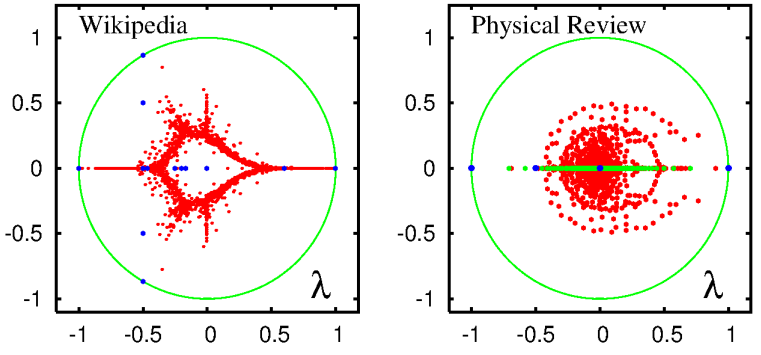



# Spectral analysis of Wikipedia and PhysRev networks

**Klaus Frahm**

***Quantware MIPS Center***

Université Paul Sabatier

Laboratoire de Physique Théorique, UMR 5152, IRSAMC, CNRS

supported by EC FET Open project NADINE

FET NADINE Workshop, Directed Networks Days 2013, Milano, 13 Juin 2013

# Google matrix for directed networks

Define the ***adjacency matrix*** $A$ by $A_{ij} = 1$ if there is a link from the node $j$ to $i$ in the network (of size $N$) and $A_{ij} = 0$ otherwise. Let $S_{ij} = A_{ij} / \sum_i A_{ij}$ and $S_{ij} = 1/N$ if $\sum_i A_{ij} = 0$ (dangling nodes). $S$ is of Perron-Frobenius type but for many networks the eigenvalue $\lambda_1 = 1$ is highly degenerate [$\Rightarrow$ convergence problem to arrive at the stationary limit of $p(t+1) = S\, p(t)$].

Therefore define the ***Google matrix***:

$$G(\alpha) = \alpha S + (1-\alpha)\,\frac{1}{N}\, e e^T$$

where $e = (1, \ldots, 1)^T$ and $\alpha = 0.85$ is a typical damping factor. Here there is a unique eigenvector for $\lambda_1 = 1$ called the ***PageRank*** $P$ and the convergence goes with $\alpha^t$.

(***CheiRank*** $P^*$ by replacing: $A \to A^* = A^T$).

# Arnoldi method

to (partly) diagonalize large sparse non-symmetric $d \times d$ matrices:

- choose an initial normalized vector $\xi_0$ (random or “otherwise”)
- determine the ***Krylov space*** of dimension $n_A$ (typically: $1 \ll n_A \ll d$) spanned by the vectors: $\xi_0,\ G\,\xi_0,\ \ldots,\ G^{n_A-1}\xi_0$
- determine by ***Gram-Schmidt*** orthogonalization an orthonormal basis $\{\xi_0,\ \ldots,\ \xi_{n_A-1}\}$ and the representation of $G$ in this basis:

$$G\,\xi_k = \sum_{j=0}^{k+1} H_{jk}\,\xi_j$$

- diagonalize the ***Arnoldi matrix*** $H$ which has ***Hessenberg*** form:

$$H = \left(\begin{array}{ccccc} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & \cdots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & * & * \\ \hline 0 & 0 & \cdots & 0 & * \end{array}\right)$$

which provides the ***Ritz eigenvalues*** that are very good aproximations to the “largest” eigenvalues of $A$.

# Invariant subspaces

In realistic WWW or other networks invariant subspaces of nodes create (possibly) large degeneracies of $\lambda_1$ (or $\lambda_2$ if $\alpha < 1$) which is very problematic for the Arnoldi method.

Therefore one needs to determine the ***invariant subspaces*** defined as ***subsets of nodes such that for any node in a subspace each outgoing link stays in the subspace***. One can efficiently find all subspaces of maximal size (or dimension) $N_c$ (with $N_c = bN$ a certain fraction of the network size $N$, e.g. $b = 0.1$) and then all subspaces with common members are merged resulting in a decomposition of the network in many ***separate subspaces*** with $N_s$ nodes and a “big” ***core space*** of the remaining $N - N_s$ nodes.

Note that ***dangling nodes*** are by construction ***core space nodes***.

Possible: core space node $\longrightarrow$ subspace node

Impossible: subspace node $\longrightarrow$ core space node

The decomposition in subspaces and a core space implies a block structure of the matrix $S$:

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix} \qquad , \qquad S_{ss} = \begin{pmatrix} S_1 & 0 & \dots \\ 0 & S_2 & \\ \vdots & & \ddots \end{pmatrix}$$

where $S_{ss}$ is block diagonal according to the subspaces. The subspace blocks of $S_{ss}$ are all matrices of PF type with at least one eigenvalue $\lambda_1 = 1$ explaining the high degeneracies.

To determine the spectrum of $S$ apply:

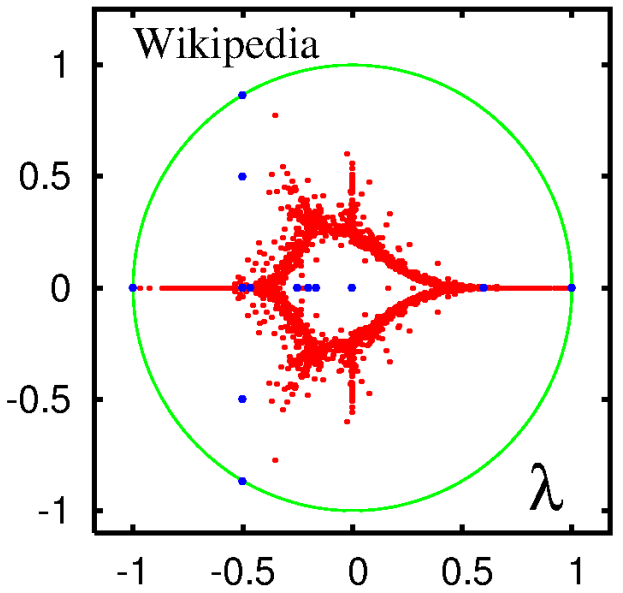
- Exact (or Arnoldi) diagonalization on each subspace.

- The Arnoldi method to $S_{cc}$ to determine the largest core space eigenvalues $\lambda_j$ (note: $|\lambda_j| < 1$). The largest eigenvalues of $S_{cc}$ are no longer degenerate but other degeneracies are possible (e.g. $\lambda_j = 0.9$ for Wikipedia).

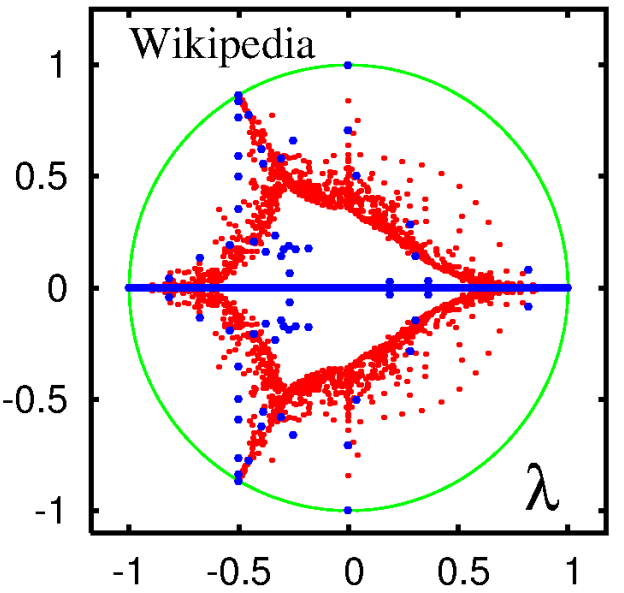
# Spectrum of Wikipedia

L. Ermann, KMF and D.L. Shepelyansky, Eur. Phys. J. B **86**, 193 (2013)

Wikipedia 2009 : $N = 3282257$ nodes, $N_\ell = 71012307$ network links.



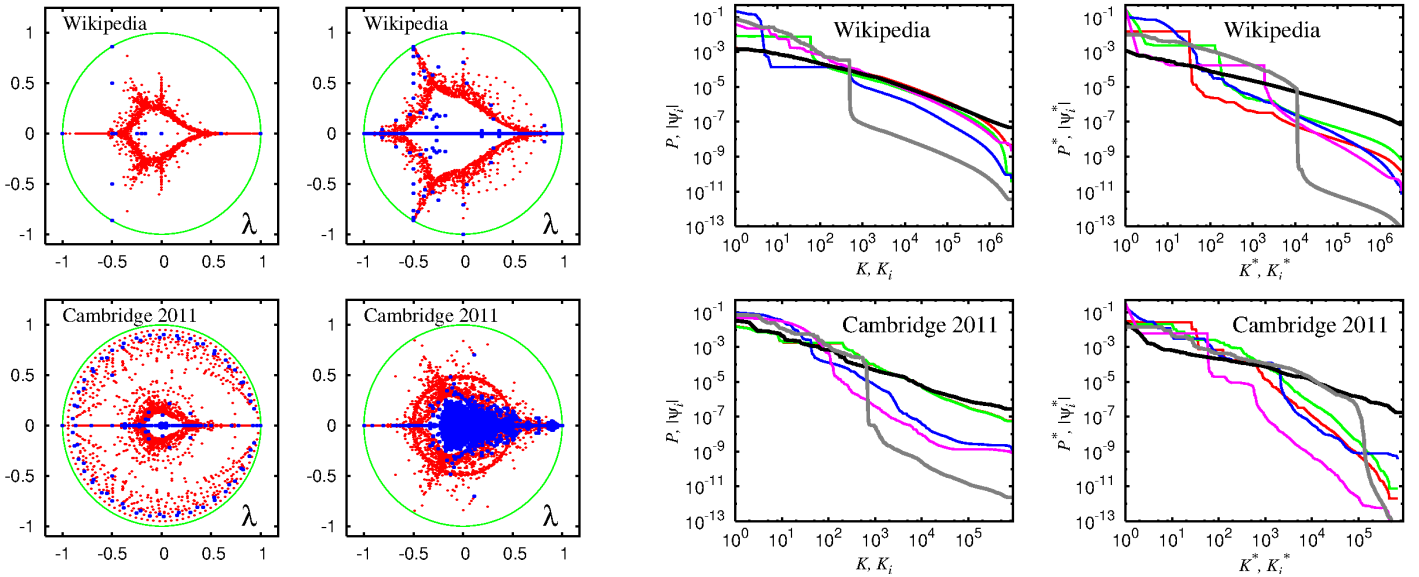

spectrum of $S$, $N_s = 515$




spectrum of $S^*$, $N_s = 21198$

$n_A = 6000$ for both cases

# Some Eigenvectors:




left (right): PageRank (CheiRank)

black: PageRank (CheiRank) at $\alpha = 0.85$

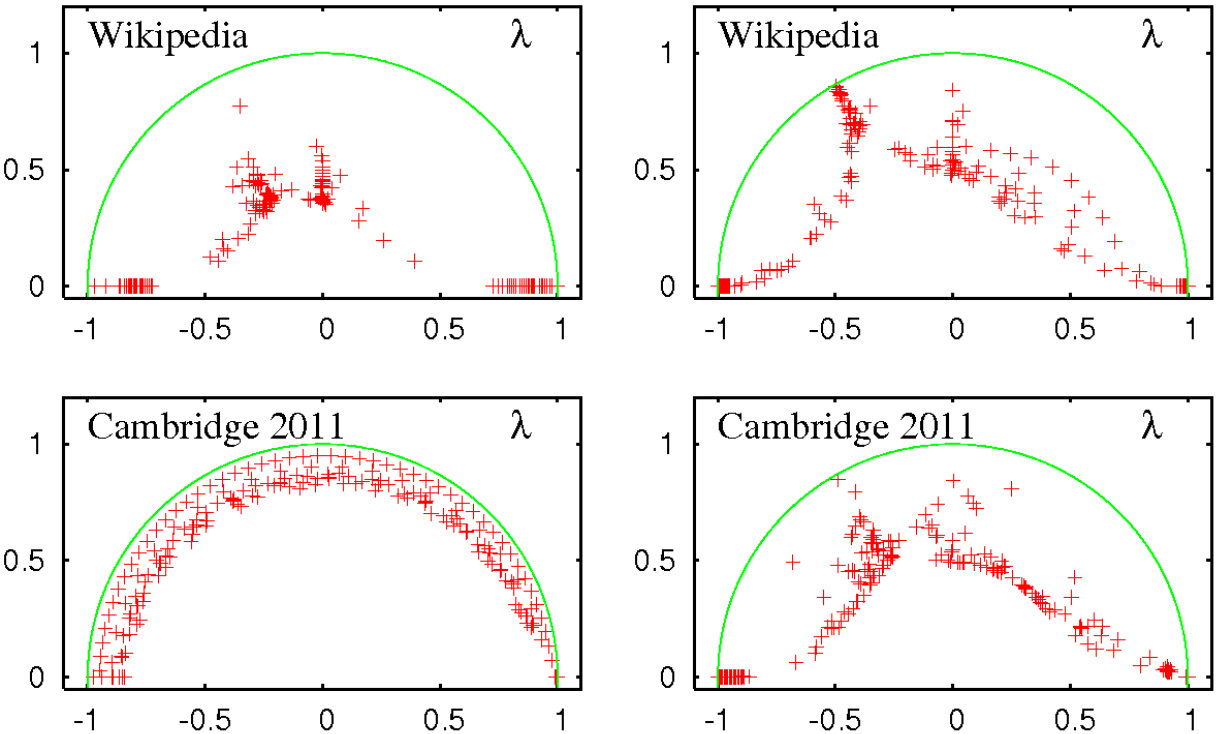
grey: PageRank (CheiRank) at $\alpha = 1 - 10^{-8}$

red and green: first two core space eigenvectors

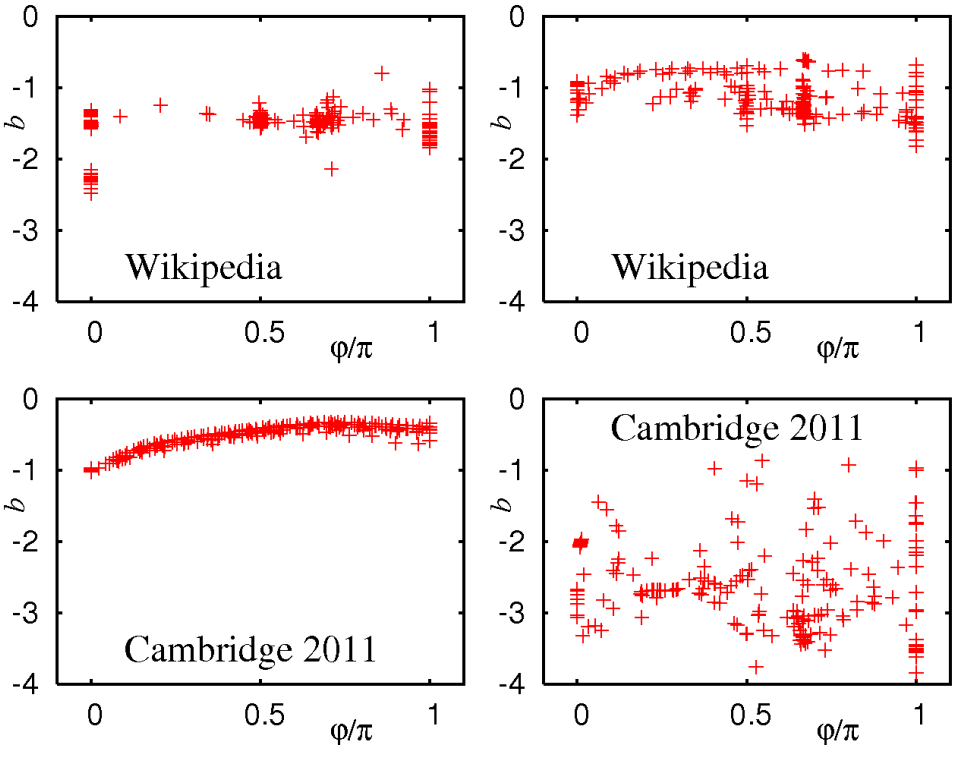
blue and pink: two eigenvectors with large imaginary part in the eigenvalue

Detail study of 200 selected eigenvectors

with eigenvalues “close” to the unit circle:



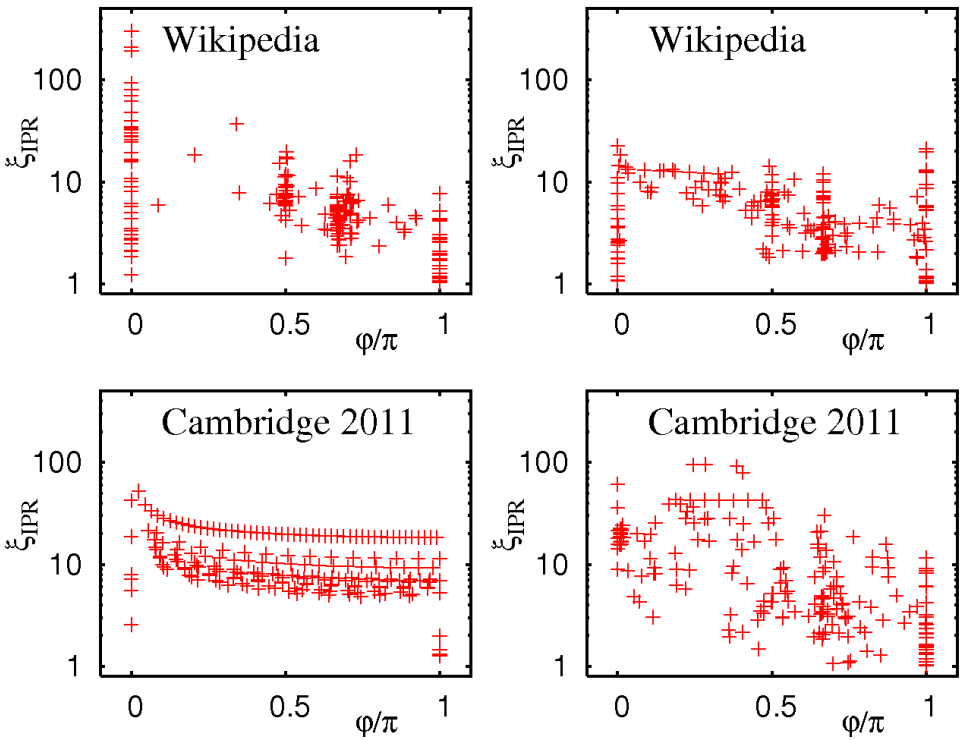
# Power law decay of eigenvectors:




$$|\psi_i(K_i)| \sim K_i^b \quad \text{for} \quad K_i \geq 10^4$$

$$\varphi = \arg(\lambda_i)$$

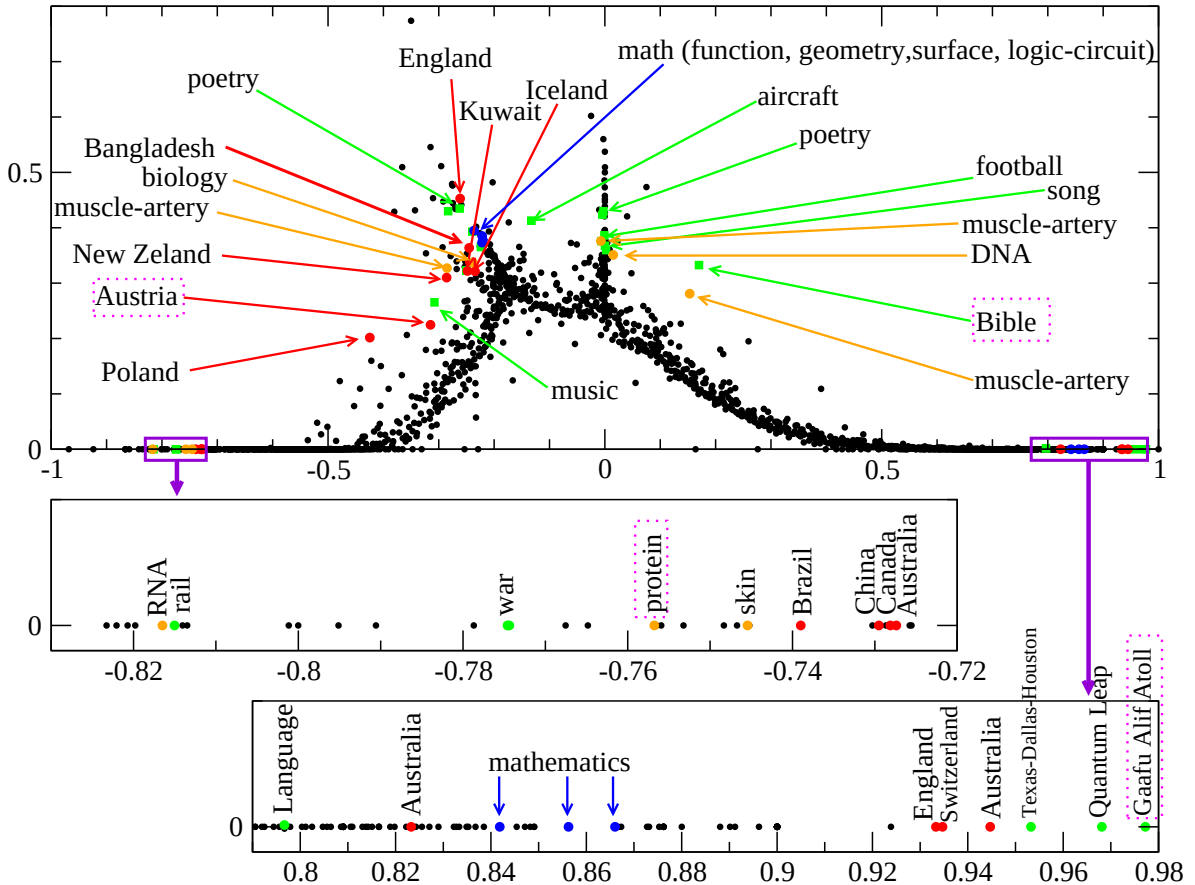
# Inverse participation ratio of eigenvectors:



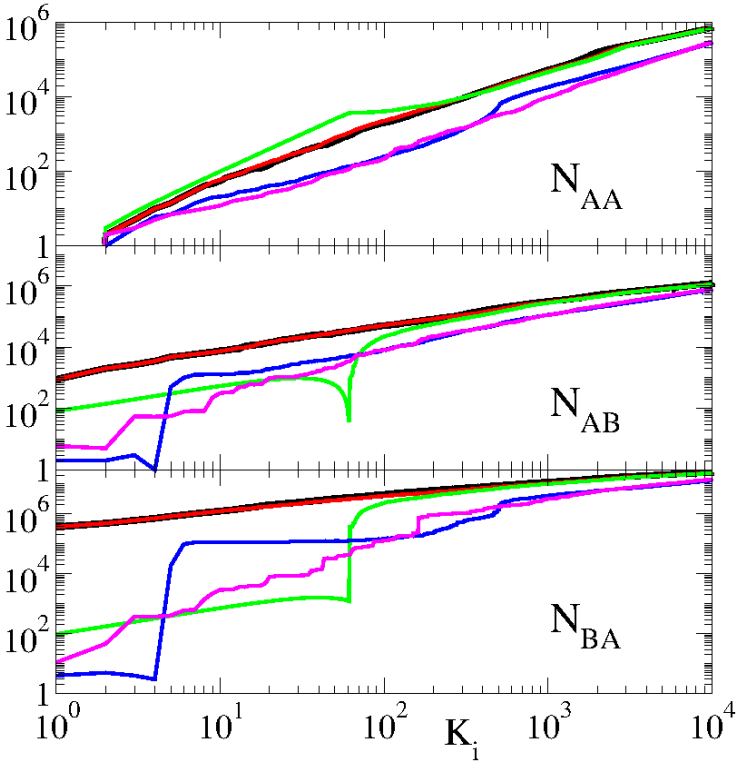

$$\xi_{\mathrm{IPR}} = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$$

$$\varphi = \arg(\lambda_i)$$

# “Themes” of certain eigenvectors:



Number of links between or inside sets $A$ and $B$ defined by the index $K_i$ ordered by decreasing absolute value of Wikipedia eigenstates:



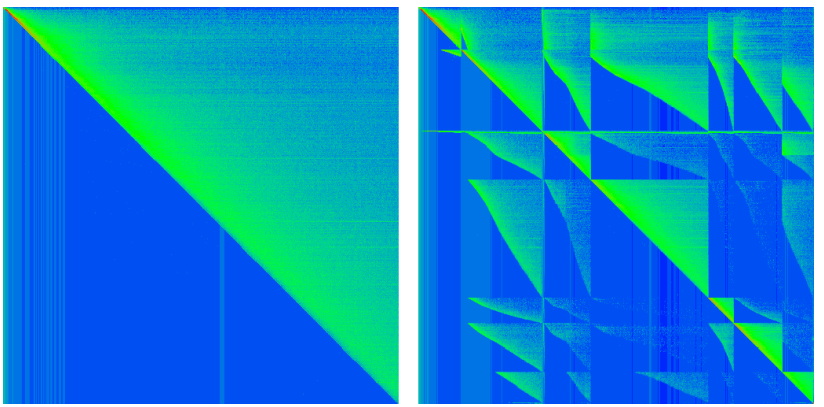

$$A = \{1, \ldots, K_i\}$$
$$B = \{K_i + 1, \ldots, N\}$$

# Physical Review network

(work in progress: KMF, Young-Ho Eom, D. Shepelyansky)

$N = 463347$ nodes and $N_\ell = 4691015$ links.

Coarse-grained matrix structure ($500 \times 500$ cells):



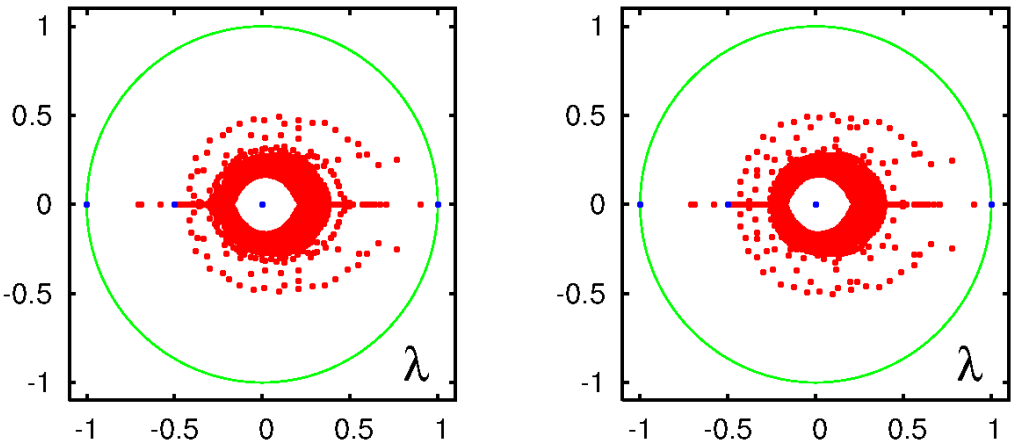
left: time ordered

right: journal and then time ordered

"11" Journals of Physical Review: (Phys. Rev. Series I), Phys. Rev., Phys. Rev. Lett., (Rev. Mod. Phys.), Phys. Rev. A, B, C, D, E, (Phys. Rev. STAB and Phys. Rev. STPER).

$\Rightarrow$ nearly triangular matrix structure of adjancy matrix: most citations links $t \longrightarrow t'$ are for $t > t'$ ("past citations") but there is small number ($12126 = 2.6 \times 10^{-3} N_\ell$) of links $t \longrightarrow t'$ with $t \leq t'$ corresponding to ***future citations***.

Spectrum by "double-precision" Arnoldi method with $n_A = 8000$:




Numerical problem: eigenvalues with $|\lambda| < 0.3 - 0.4$ are not reliable!

Reason: large Jordan subspaces associated to the eigenvalue $\lambda = 0$.

"very bad" Jordan perturbation theory:

Consider a "perturbed" Jordan block of size $D$:

$$\begin{pmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \varepsilon & 0 & \cdots & 0 & 0 \end{pmatrix}$$

characteristic polynomial: $\lambda^D - (-1)^D \varepsilon$

$\varepsilon = 0 \quad \Rightarrow \quad \lambda = 0$

$\varepsilon \neq 0 \quad \Rightarrow \quad \lambda_j = -\varepsilon^{1/D} \exp(2\pi i j / D)$

for $D \approx 10^2$ and $\varepsilon = 10^{-16} \quad \Rightarrow$ "Jordan-cloud" of artifical eigenvalues due to rounding errors in the region $|\lambda| < 0.3 - 0.4$.

# Triangular approximation

Remove the small number of links due to “future citations”.

***Semi-analytical diagonalization*** is possible:

$$\boxed{S = S_0 + e\, d^T/N}$$

where $e_n = 1$ for all nodes $n$, $d_n = 1$ for dangling nodes $n$ and $d_n = 0$ otherwise. $S_0$ is the pure link matrix which is ***nil-potent***:

$$\boxed{S_0^l = 0} \text{ with } l = 352.$$

Let $\psi$ be an eigenvector of $S$ with eigenvalue $\lambda$ and $C = d^T\psi$.

- If $C = 0 \;\Rightarrow\; \psi$ eigenvector of $S_0 \;\Rightarrow\; \lambda = 0$ since $S_0$ nil-potent.

  These eigenvectors belong to large Jordan blocks and are responsible for the numerical problems.

  Note: Similar situation as in ***network of integer numbers*** where $l = [\log_2(N)]$ and numerical instability for $|\lambda| < 0.01$.

- If $C \neq 0 \;\Rightarrow\; \lambda \neq 0$ since the equation $S_0\psi = -C\, e/N$ does not have a solution $\Rightarrow\; \lambda\mathbf{1} - S_0$ invertible.

$$\Rightarrow\; \psi = C\,(\lambda\mathbf{1} - S_0)^{-1}\, e/N = \frac{C}{\lambda} \sum_{j=0}^{l-1} \left(\frac{S_0}{\lambda}\right)^j e/N \quad .$$

$$\text{From } \lambda^l = (d^T\psi/C)\lambda^l \;\Rightarrow\; \boxed{\mathcal{P}_r(\lambda) = 0}$$

with the ***reduced polynomial*** of degree $l = 352$ :

$$\mathcal{P}_r(\lambda) = \lambda^l - \sum_{j=0}^{l-1} \lambda^{l-1-j}\, c_j = 0 \quad , \quad c_j = d^T\, S_0^j\, e/N \; .$$

$\Rightarrow$ at most $l = 352$ eigenvalues $\lambda \neq 0$ which can be numerically determined as the zeros of $\mathcal{P}_r(\lambda)$.

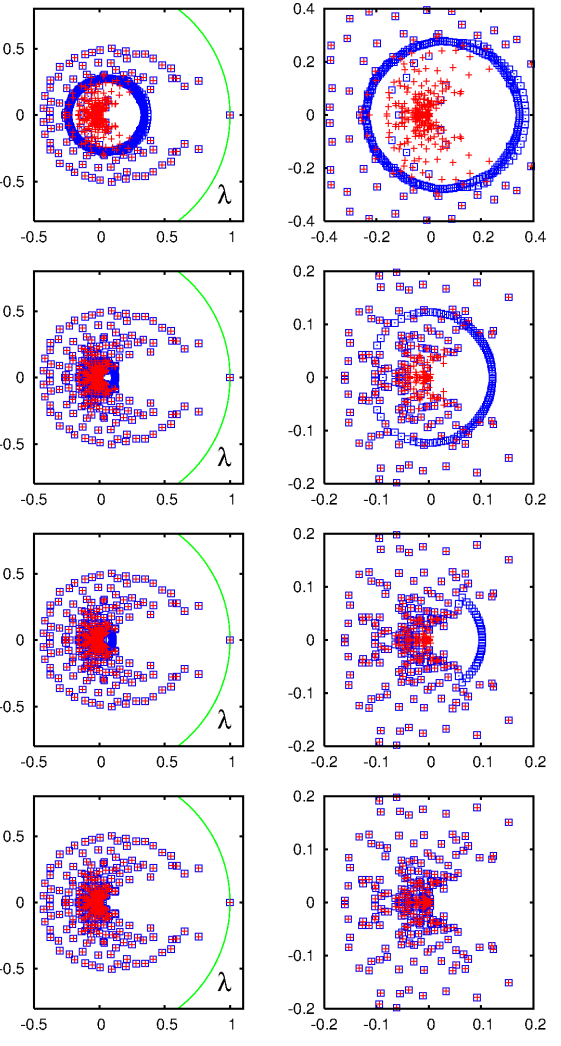
However: still numerical problems:

- $c_{l-1} \approx 3.6 \times 10^{-352}$
- alternate sign problem with a strong loss of significance.
- big sensitivity of eigenvalues on $c_j$

# Solution:

Using the multi precision library GMP with 256 binary digits the zeros of $\mathcal{P}_r(\lambda)$ can be determined with accuracy $\sim 10^{-18}$.

Furthermore the Arnoldi method can also be implemented with higher precision.

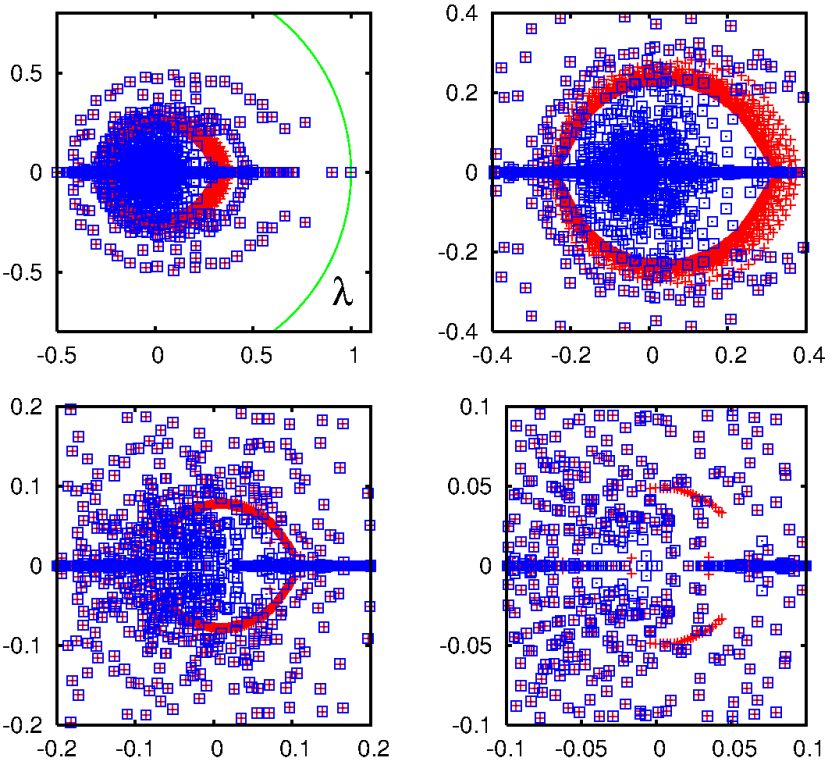



red crosses: zeros of $\mathcal{P}_r(\lambda)$ from 256 binary digits calculation

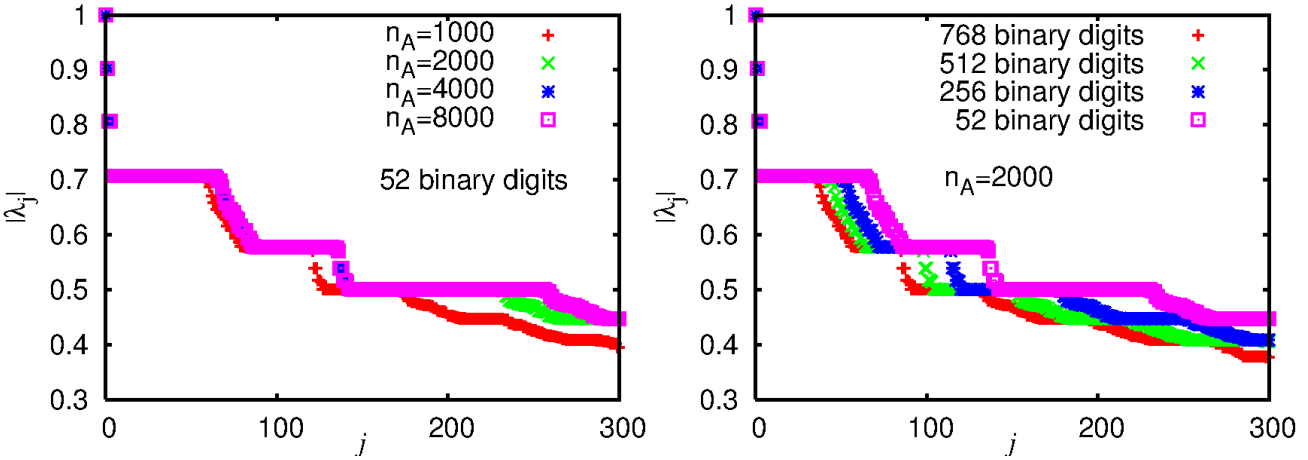
blue squares: eigenvalues from Arnoldi method with 52, 256, 512, 1280 binary digits. In the last case: $\Rightarrow$ break off at $n_A = 352$ with vanishing coupling element.

# Full Physical Review network

High precision Arnoldi method for full Physical Review network (including the “future citations”) for 52, 256, 512, 768 binary digits and $n_A = 2000$:



# Degeneracies




High precision in Arnoldi method is “bad” to count the degeneracy of certain degenerate eigenvalues.

In theory the Arnoldi method cannot find several eigenvectors for degenerate eigenvalues, a shortcoming which is (partly) “repaired” by rounding errors.

**Q:** How are highly degenerate core space eigenvalues possible ?

# Semi-analytical argument for the full PR network:

$$\boxed{S = S_0 + e\, d^T/N}$$

There are ***two groups of eigenvectors*** $\psi$ with: $S\psi = \lambda\psi$

1. Those with $d^T\psi = 0 \quad \Rightarrow \quad \psi$ is also an eigenvector of $S_0$. Generically an arbitrary eigenvector of $S_0$ is **not** an eigenvector of $S$ **unless** the eigenvalue is degenerate with degeneracy $m > 1$. Using linear combinations of different eigenvectors for the same eigenvalue one can construct $m - 1$ eigenvectors $\psi$ respecting $d^T\psi = 0$ which are therefore eigenvectors of $S$.

   Pratically: determine degenerate subspace eigenvalues of $S_0$ (and also of $S_0^T$) which are of the form: $\lambda = \pm 1/\sqrt{n}$ with $n = 1,\ 2,\ 3,\ \ldots$ due to $2 \times 2$-blocks:

$$\begin{pmatrix} 0 & 1/n_1 \\ 1/n_2 & 0 \end{pmatrix} \quad \Rightarrow \quad \lambda = \pm\frac{1}{\sqrt{n_1\, n_2}}\,.$$

2. Those with $d^T\psi \neq 0 \quad \Rightarrow \quad \mathcal{R}(\lambda) = 0$ with the rational function:

$$\mathcal{R}(\lambda) = 1 - d^T \frac{\mathbf{1}}{\lambda\,\mathbf{1} - S_0} e/N = 1 - \sum_{j,q} \frac{C_{jq}}{(\lambda - \rho_j)^q}$$

Here $C_{jq}$ and $\rho_j$ are unknown, except for
$\rho_1 = 2\,\mathrm{Re}\,[(9 + i\sqrt{119})^{1/3}]/(135)^{1/3} \approx 0.9024$ and
$\rho_{2,3} = \pm 1/\sqrt{2} \approx \pm 0.7071$.

Idea: Expand the geometric matrix series $\Rightarrow$

$$\mathcal{R}(\lambda) = 1 - \sum_{j=0}^{\infty} c_j \lambda^{-1-j} \quad , \quad c_j = d^T\, S_0^j\, e/N$$

which converges for $|\lambda| > \rho_1 \approx 0.9024$ since $c_j \sim \rho_1^j$ for $j \to \infty$.

Problem: How to determine the zeros of $\mathcal{R}(\lambda)$ with $|\lambda| < \rho_1$ ?

## *Analytic continuation by rational interpolation:*

Use the series to evaluate $\mathcal{R}(z)$ at $n_S$ support points $z_j = \exp(2\pi ij/n_S)$ with a given precision of $p$ binary digits and determine the rational function $R_I(z)$ which interpolates $\mathcal{R}(z)$ at these support points. Two cases:

$$n_S = 2n_R + 1 \quad \Rightarrow \quad R_I(z) = \frac{P_{n_R}(z)}{Q_{n_R}(z)}$$

$$n_S = 2n_R + 2 \quad \Rightarrow \quad R_I(z) = \frac{P_{n_R}(z)}{Q_{n_R+1}(z)}$$

The $n_R$ zeros of $P_{n_R}(z)$ are approximations of the eigenvalues of $S$ (of the 2nd group).

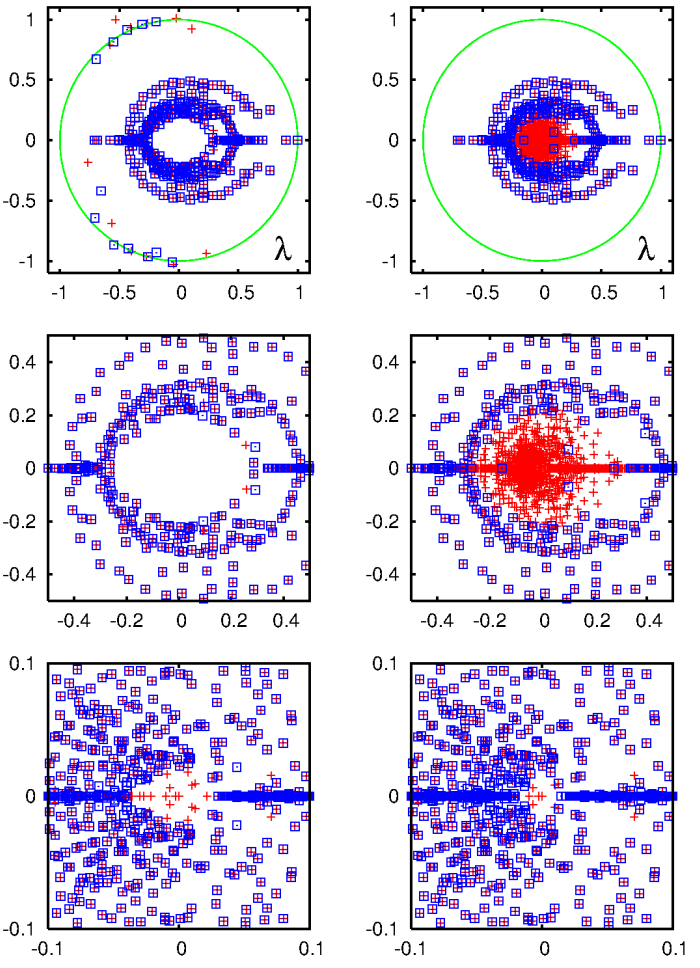
For a given precision, e. g. $p = 1024$ binary digits one can obtain a certain number of reliable eigenvalues, e. g. $n_R = 300$. The method can be pushed up to $p = 16384$ and $n_R = 2500$ which is better than the high precision Arnoldi method with $n_A = 2000$.

# Examples:

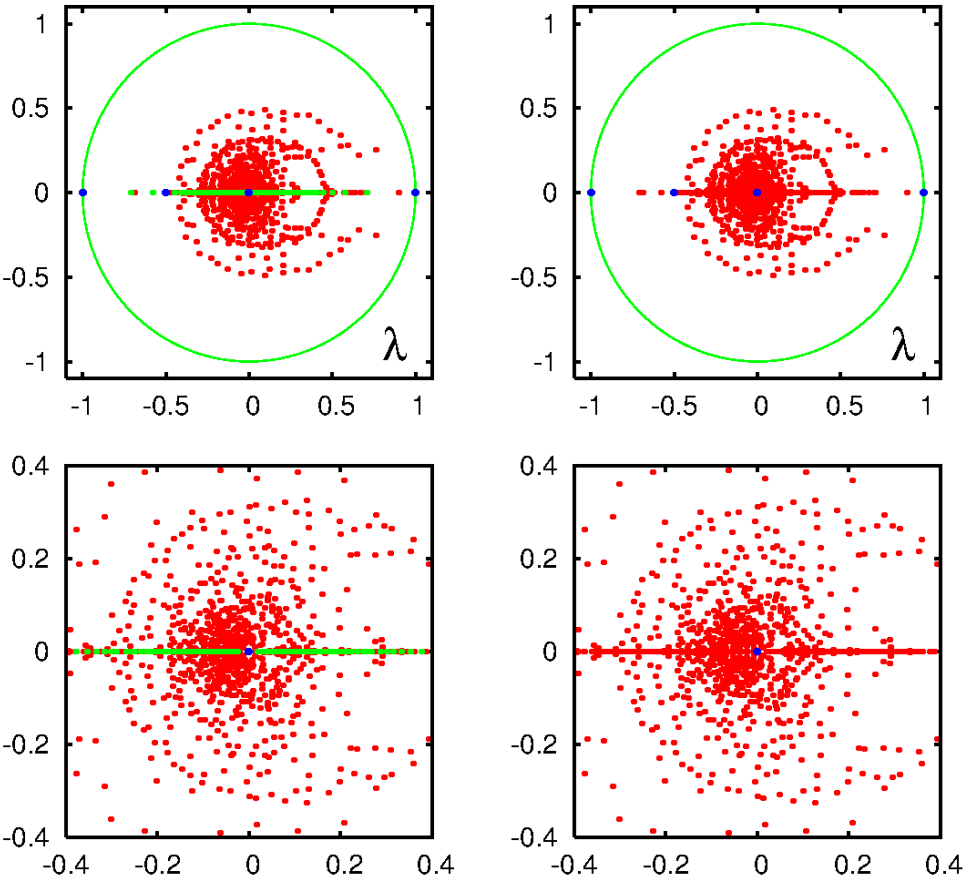
Some “artificial zeros” for $n_R = 340$ and $p = 1024$ (*left top and middle panels*) where both variants of the method differ.

For $n_R = 300$ and $p = 1024$ most zeros coincide with HP Arnoldi method (*right top and middle panels*) and both variants of the method coincide.

*Lower panels:* comparison for $n_R = 2000$, $p = 12288$ (left) or for $n_R = 2500$, $p = 16384$ with HP Arnoldi method.



Accurate eigenvalue spectrum for the full Physical Review network by the rational interpolation method (left) and the HP Arnoldi method (right):



# Conclusion

- Detailed eigenvector study for the ***Wikipedia network***.
- Identification of certain ***themes*** or ***communities*** with the help of eigenvectors.
- Subtle numerical problems for the eigenvalue problem of the ***Physical Review network*** which can be solved by a semi-analytical method and a high precision implementation of the Arnoldi method.
- Understanding of the ***degeneracies of core space eigenvalues*** and a decompostion of the core space eigenvalues in two groups. Important role of subspaces of $S_0$ (very different from the subspaces of $S$ !).

- New ***rational interpolation method*** to determine accurately the eigenvalues of a network matrix. Well suited for nearly triangular matrices but works in principle also for other case (e. g. Wikipedia but less efficient here).

- Drastic effect of the ***triangular approximation*** on the eigenvalue spectrum. Strong reduction of non-vanishing eigenvalues, from about $\sim 8000 - 10000$ to $352$ and only very few eigenvalues on the real axis. This implies a very strong effect of the few ***future citations*** on the spectrum.

- Very useful applications of the ***GNU high precision library GMP: http://gmplib.org/*** for different numerical methods: determination of zeros of the reduced polynomial, rational interpolation method, Arnoldi method.

# Appendix:

The subspace of $\lambda \neq 0$ is represented by the vectors $v^{(j)} = S_0^{j-1} e/N$ for $j = 1, \ldots, l$

$$\Rightarrow \quad S\, v^{(j)} = c_{j-1}\, v^{(1)} + v^{(j+1)} = \sum_{k=0}^{l-1} \bar{S}_{k,j}\, v^{(k)}$$

"Small" $l \times l$-representation matrix :

$$\bar{S} = \begin{pmatrix} c_0 & c_1 & \cdots & c_{l-2} & c_{l-1} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \quad , \quad \bar{P} = C \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

with $P = \sum_j \bar{P}_j\, v^{(j)} = C \sum_j v^{(j)}$ and due to sum rule: $\sum_j c_j = 1$.