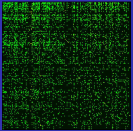
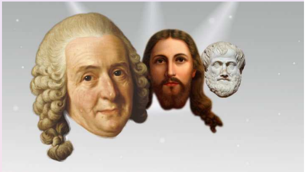
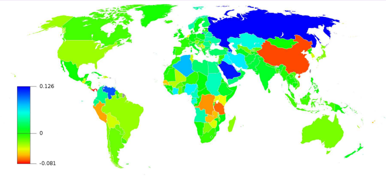
New tools and algorithms for directed network analysis

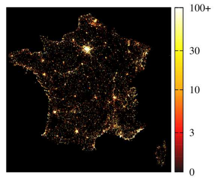
Dima Shepelyansky (CNRS, Toulouse)
www.quantware.ups-tlse.fr/dima



FET Open NADINE project No 288956 (4 partners) =>
http://www.quantware.ups-tlse.fr/FETNADINE/







1945: Nuclear physics → Wigner (1955)→ Random Matrix Theory

1991: WWW, small world social networks → Markov chains (1906) → Google matrix

Despite the importance of large-scale search engines on the web, very little academic research has been done on them.

S.Brin and L.Page, Comp. Networks ISDN Systems **30**, 107 (1998)

Workpackages and milestones

WP1: CheiRank versus PageRank, RTD centrality measures and network structure (UTWE)

WP2: Network analysis through Google matrix eigenspectrum and eigenstates (CNRS)

WP3: Applications to voting systems in social networks (UMIL)

WP4: Applications of new tools and algorithms to real-world network structures (MTA_SZTAKI)

WP5: Database development of real-world networks (UMIL)

RM1: Correlation properties of directed networks (WP1.1);

RM2: Statistical characterization of 2DRanking (WP1.2, WP2.1, WP4.3);

RM3: Eigenstate community detection (WP2.2, WP3.1);

RM4: Spam filter protocols (WP4.2);

RM5: Network-specific centrality measures (WP1.1, WP1.3, WP3.1, WP3.2)

Milestones period 2

DM6: Fractal Weyl law properties of networks;

DM7: Protocols for large-scale network processing;

DM8: Characterization of multi-product world trade network;

DM9: Webcrawler development and database collection;

DM10: Monte Carlo algorithms for centrality measures;

DM11: Delocalization conditions for Google matrix eigenstates;

DM12: New protocols for social voting and recommendation;

DM13: Characterization of ranking of Wikipedia and other networks;

DM14: Characterization of time-evolving Web structures

4 joint partner publications during 2nd period (5 in total)

Highlights

* - Highlights in the second reporting period include work on interactions of cultures and top 100 people of Wikipedia from ranking of 24 language editions [69] (P1+P4),

* - Google matrix analysis of the multiproduct world trade network [39] (P1 with UN COMTRADE database),

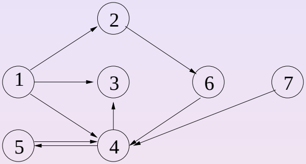
* - development of Monte Carlo algorithm for quick detection of high-degree entities in large directed networks [50] (P2),

* - results for RecSys Challenge 2014: an ensemble of binary classifiers and matrix factorization [53] (P3),

* - construction of a weighted correlation index for rankings with ties [70] (P4)

Organized:
3 NADINE workshops, 1 + 1 Luchon summer school (2014 + 2015)

Markov chains (1906) and Directed networks



Weighted adjacency matrix

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

For a directed network with N nodes the adjacency matrix $\mathbf{A}$ is defined as $A_{ij} = 1$ if there is a link from node j to node i and $A_{ij} = 0$ otherwise. The weighted adjacency matrix is

$$S_{ij} = A_{ij} / \sum_k A_{kj}$$

In addition the elements of columns with only zeros elements are replaced by $1/N$.

How Google works

Google Matrix and Computation of PageRank

$\mathbf{P} = \mathbf{SP} \Rightarrow \mathbf{P}$= stationary vector of $\mathbf{S}$; can be computed by iteration of $\mathbf{S}$.

To remove convergence problems:

- Replace columns of 0 (dangling nodes) by $\frac{1}{N}$:

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 1 & 1 & 1 \\ 0 & 0 & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \end{pmatrix}; \mathbf{S}^* = \begin{pmatrix} \frac{1}{7} & 1 & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \frac{1}{7} \\ \frac{1}{7} & 0 & 0 & 0 & 0 & 1 & \frac{1}{7} \\ \frac{1}{7} & 0 & 0 & 0 & 0 & 0 & \frac{1}{7} \\ \frac{1}{7} & 0 & \frac{1}{2} & 0 & 1 & 0 & \frac{1}{7} \\ \frac{1}{7} & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{7} \\ \frac{1}{7} & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{7} \\ \frac{1}{7} & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{7} \end{pmatrix}.$$

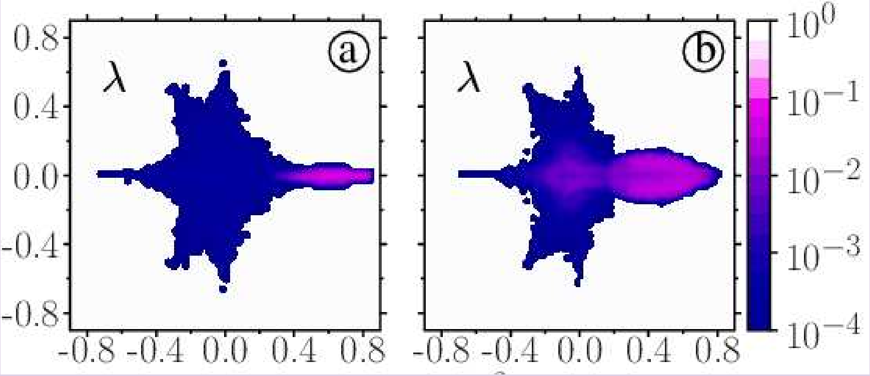
- To remove degeneracies of $\lambda = 1$, replace $\mathbf{S}$ by **Google matrix**

  $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\frac{\mathbf{E}}{N}$; $\quad GP = \lambda P$ => Perron-Frobenius operator

- α models a random surfer with a random jump after approximately 6 clicks (usually $\alpha = 0.85$); PageRank vector => P at $\lambda = 1$ ($\sum_j P_j = 1$).
- CheiRank vector P^*: $G^* = \alpha\mathbf{S}^* + (1-\alpha)\frac{\mathbf{E}}{N}$, $G^* P^* = P^*$
  ($\mathbf{S}^*$ with inverted link directions)
  Fogaras (2003) ... Chepelianskii arXiv:1003.5455 (2010) ...

Anderson transition on directed networks

Anderson (1958) metal-insulator transition for electron transport in disordered solids (Nobel prize 1977)



Panels show distribution of IPR values ξ (number of nodes contributing to an eigenstate) on the plane λ or eigenvalues of G matrix for two nodels of directed networks with disorder; color shows the ration ξ/N, $\alpha = 0.85$.

O.Zhirov, D.S [40] (2015)

Quantware group + partners

Toulouse (physics) Permanent:
Klaus Frahm (PR Univ. P. Sabatier)
Bertrand Georgeot (DR CNRS), DS (DR CNRS)

Post-doc: Young-Ho Eom (NADINE)

PhD: Vivek Kandiah (CNRS - Region Midi-Pyrenees) PhD Oct 2014

External collaborators:
Alexei Chepelianskii (U Cambridge - LPS Orsay)
Leonardo Ermann (CNEA Buenos Aires)
Oleg Zhirov (BINP Russian Academy Novosibirsk)
NOMAO.COM: Estelle Delpech, Samuel Phan

Twente => Nelly Litvak (mathematics)

Budapest => Andras Benczur (computer science)

Milano => Sebastiano Vigna (computer science)

Results 2nd period: 41 publications (73 in total), 30 conference presentations (67 in total), ...