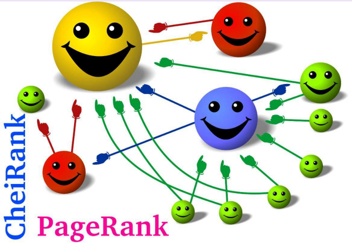
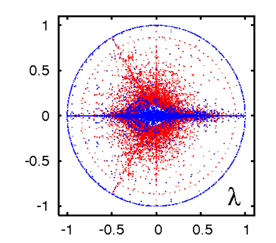
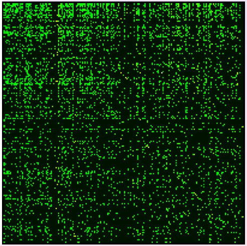
New tools and algorithms for directed network analysis

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/







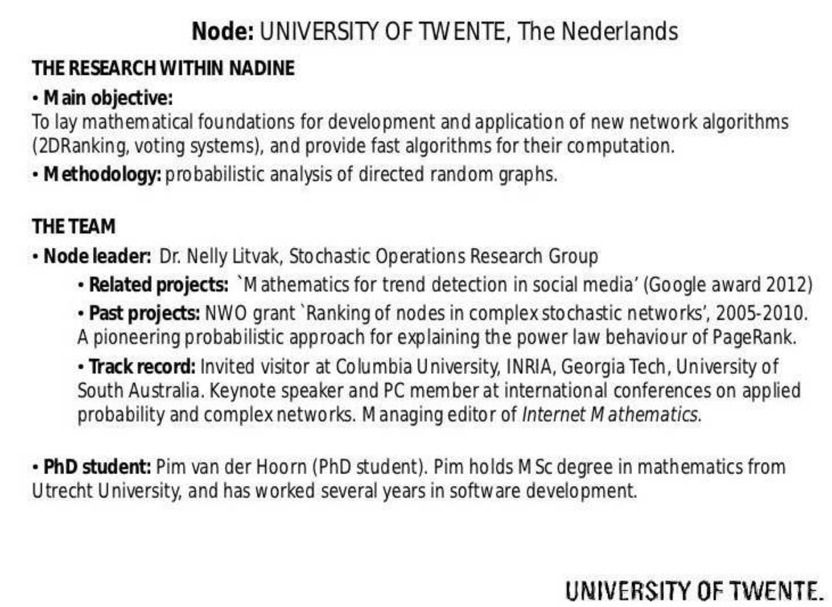
1945: Nuclear physics → Wigner (1955)→ Random Matrix Theory

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)

Partner 2 Univ Twente



Node: UNIVERSITY OF TWENTE, The Nederlands

THE RESEARCH WITHIN NADINE

- **Main objective:**

To lay mathematical foundations for development and application of new network algorithms (2DRanking, voting systems), and provide fast algorithms for their computation.

- **Methodology:** probabilistic analysis of directed random graphs.

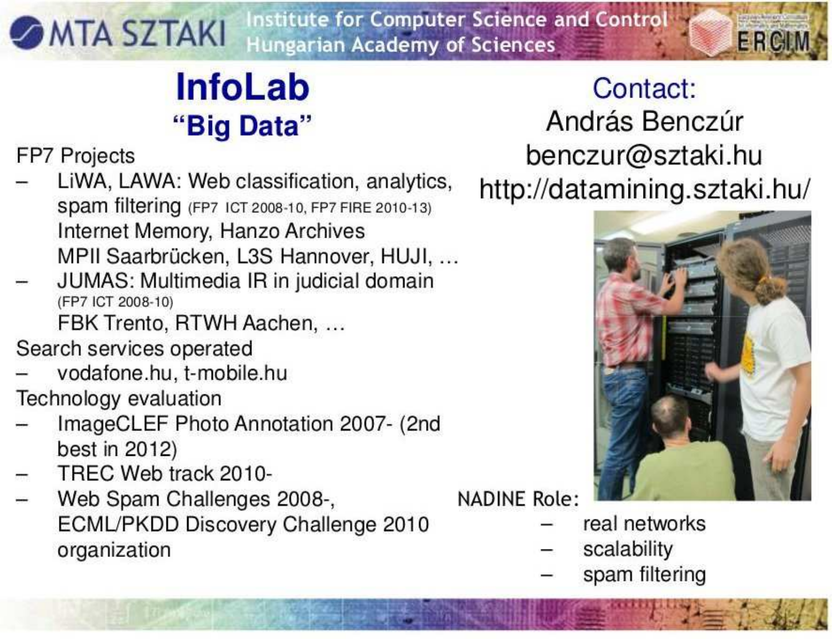
THE TEAM

- **Node leader:** Dr. Nelly Litvak, Stochastic Operations Research Group
  - **Related projects:** `Mathematics for trend detection in social media' (Google award 2012)
  - **Past projects:** NWO grant `Ranking of nodes in complex stochastic networks', 2005-2010. A pioneering probabilistic approach for explaining the power law behaviour of PageRank.
  - **Track record:** Invited visitor at Columbia University, INRIA, Georgia Tech, University of South Australia. Keynote speaker and PC member at international conferences on applied probability and complex networks. Managing editor of *Internet Mathematics*.

- **PhD student:** Pim van der Hoorn (PhD student). Pim holds MSc degree in mathematics from Utrecht University, and has worked several years in software development.

UNIVERSITY OF TWENTE.

Partner 3 MTA_SZTAKI Budapest



MTA SZTAKI Institute for Computer Science and Control
Hungarian Academy of Sciences
ERCIM

InfoLab
“Big Data”

FP7 Projects

- LiWA, LAWA: Web classification, analytics, spam filtering (FP7 ICT 2008-10, FP7 FIRE 2010-13)
  Internet Memory, Hanzo Archives
  MPII Saarbrücken, L3S Hannover, HUJI, …
- JUMAS: Multimedia IR in judicial domain (FP7 ICT 2008-10)
  FBK Trento, RTWH Aachen, …

Search services operated

- vodafone.hu, t-mobile.hu

Technology evaluation

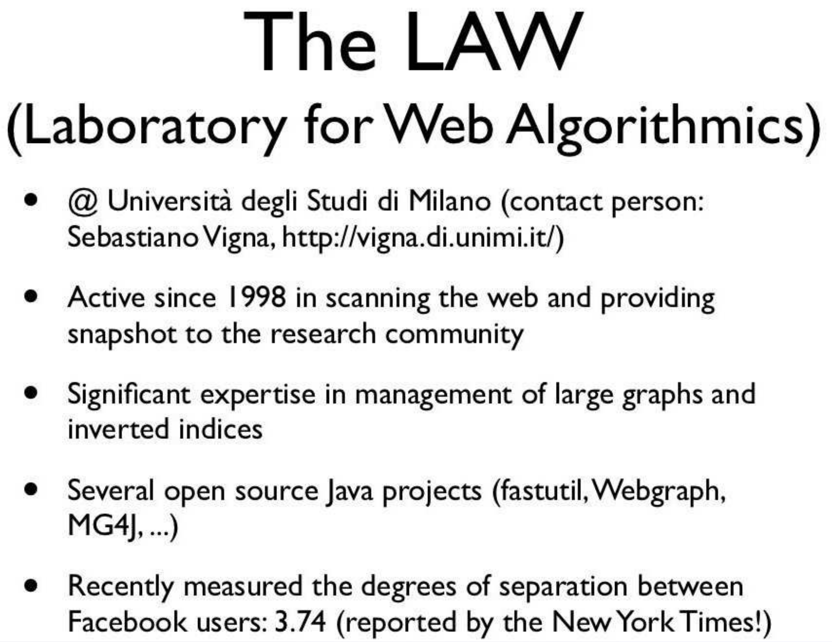
- ImageCLEF Photo Annotation 2007- (2nd best in 2012)
- TREC Web track 2010-
- Web Spam Challenges 2008-, ECML/PKDD Discovery Challenge 2010 organization

Contact:
András Benczúr
benczur@sztaki.hu
http://datamining.sztaki.hu/

NADINE Role:

- real networks
- scalability
- spam filtering

Partner 4 Univ Milano



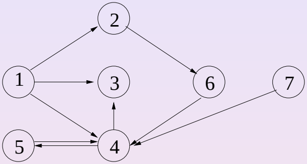
The LAW

(Laboratory for Web Algorithmics)

- @ Università degli Studi di Milano (contact person: Sebastiano Vigna, http://vigna.di.unimi.it/)
- Active since 1998 in scanning the web and providing snapshot to the research community
- Significant expertise in management of large graphs and inverted indices
- Several open source Java projects (fastutil, Webgraph, MG4J, ...)
- Recently measured the degrees of separation between Facebook users: 3.74 (reported by the New York Times!)

How Google works

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

$\mathbf{P} = \mathbf{S}\mathbf{P} \Rightarrow$ **P**= stationary vector of **S**; can be computed by iteration of **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 **S** by **Google matrix**

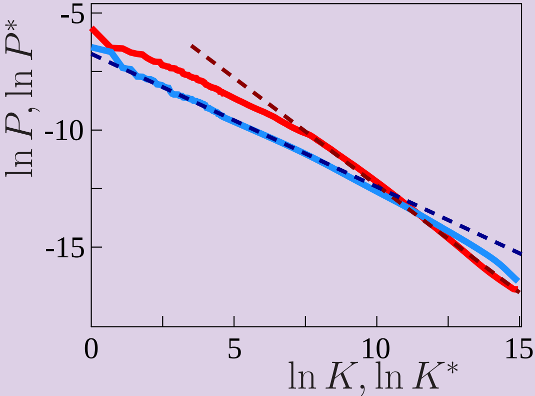
  $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\frac{\mathbf{E}}{N}$; $\quad GP = \lambda P \quad$ => Perron-Frobenius operator

- 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) ...

Wikipedia ranking of human knowledge

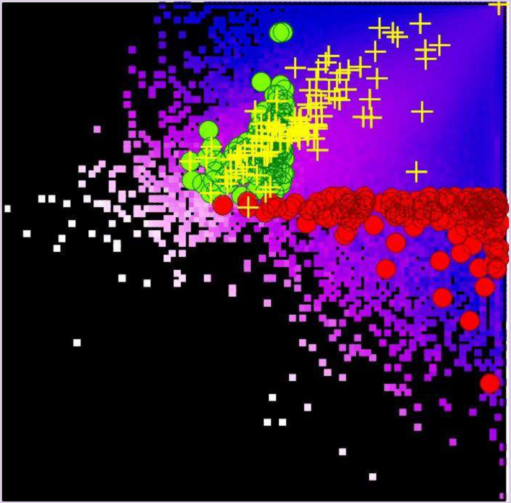
Wikipedia English articles $N = 3282257$ dated Aug 18, 2009



Dependence of probability of PagRank P (red) and CheiRank P^* (blue) on corresponding rank indexes K, K^*; lines show slopes $\beta = 1/(\nu - 1)$ with $\beta = 0.92; 0.57$ respectively for $\nu = 2.09; 2.76$.

[Zhirov, Zhirov, DS EPJB **77**, 523 (2010)]

Two-dimensional ranking of Wikipedia articles



Density distribution in plane of PageRank and CheiRank indexes ($\ln K, \ln K^*$): 100 top personalities from PageRank (green), CheiRank (red) and Hart book (yellow)

Wikipedia ranking of universities, personalities

Universities:

PageRank: 1. Harvard, 2. Oxford, 3. Cambridge, 4. Columbia, 5. Yale, 6. MIT, 7. Stanford, 8. Berkeley, 9. Princeton, 10. Cornell.

2DRank: 1. Columbia, 2. U. of Florida, 3. Florida State U., 4. Berkeley, 5. Northwestern U., 6. Brown, 7. U. Southern CA, 8. Carnegie Mellon, 9. MIT, 10. U. Michigan.

CheiRank: 1. Columbia, 2. U. of Florida, 3. Florida State U., 4. Brooklyn College, 5. Amherst College, 6. U. of Western Ontario, 7. U. Sheffield, 8. Berkeley, 9. Northwestern U., 10. Northeastern U.

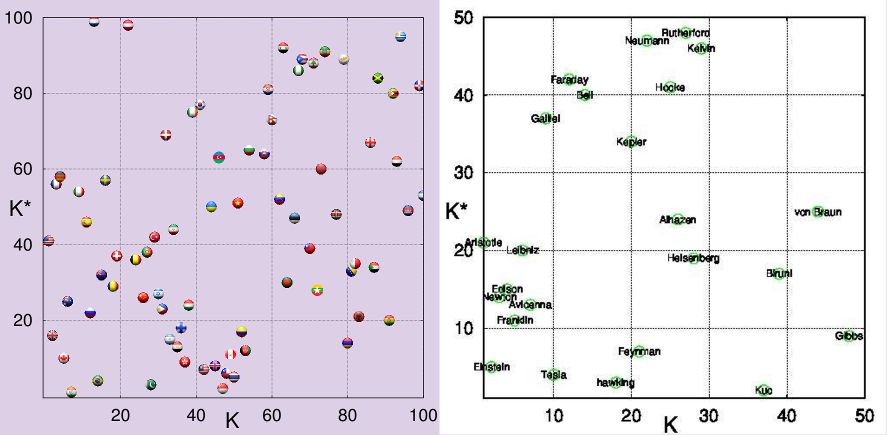
Personalities:

PageRank: 1. Napoleon I of France, 2. George W. Bush, 3. Elizabeth II of the United Kingdom, 4. William Shakespeare, 5. Carl Linnaeus, 6. Adolf Hitler, 7. Aristotle, 8. Bill Clinton, 9. Franklin D. Roosevelt, 10. Ronald Reagan.

2DRank: 1. Michael Jackson, 2. Frank Lloyd Wright, 3. David Bowie, 4. Hillary Rodham Clinton, 5. Charles Darwin, 6. Stephen King, 7. Richard Nixon, 8. Isaac Asimov, 9. Frank Sinatra, 10. Elvis Presley.

CheiRank: 1. Kasey S. Pipes, 2. Roger Calmel, 3. Yury G. Chernavsky, 4. Josh Billings (pitcher), 5. George Lyell, 6. Landon Donovan, 7. Marilyn C. Solvay, 8. Matt Kelley, 9. Johann Georg Hagen, 10. Chikage Oogi.

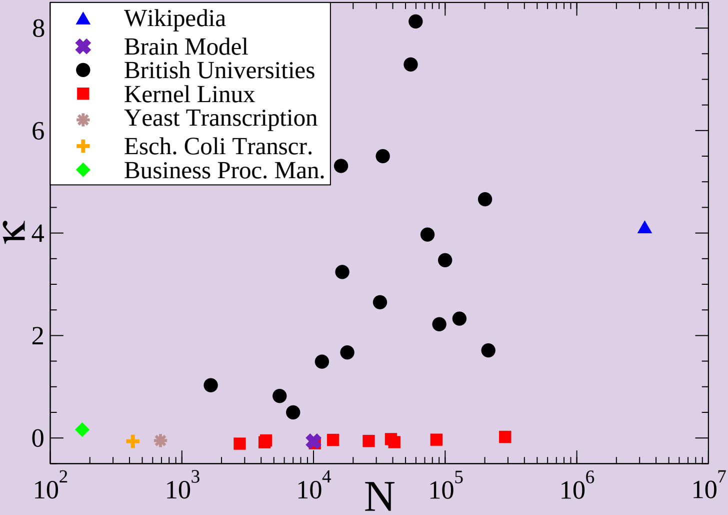
Toward two-dimensional search engines



local CheiRank vs PageRank of countries, physicists via Wikipedia

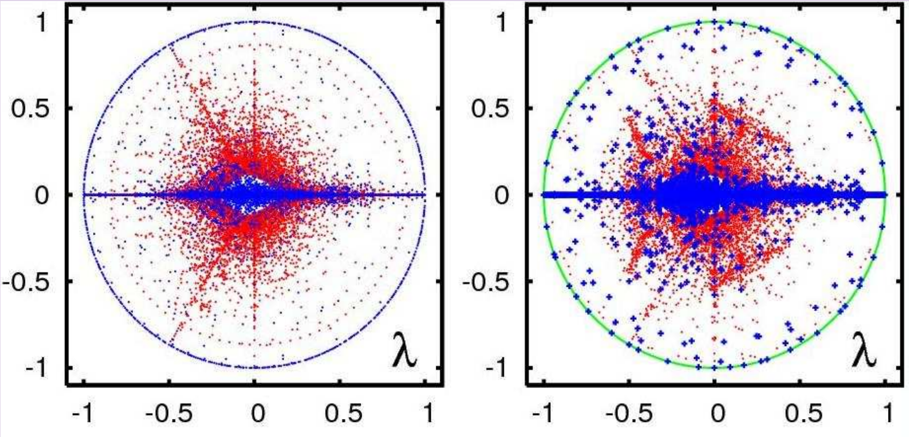
Ermann, Chepelianskii, DS J. Phys. A Math. Theor. **45**, 275101 (2012)

Correlator of PageRank and CheiRank



$\kappa = N \sum_i P(K(i))P^*(K^*(i)) - 1$

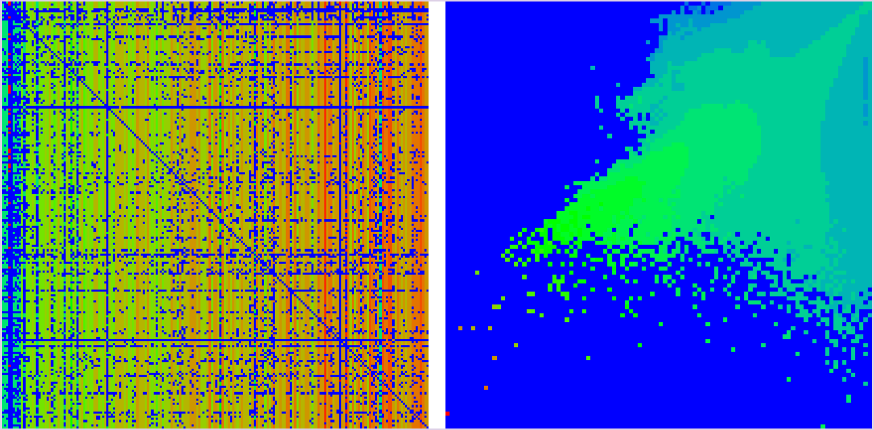
Spectrum of UK University networks



Arnoldi method: Spectrum of Google matrix for Univ. of Cambridge (left) and Oxford (right) in 2006; 20% at $\lambda = 1$ ($N \approx 200000$, $\alpha = 1$). [Frahm, Georgeot, DS arxiv:1105.1062 (2011)]

Google matrix of Twitter

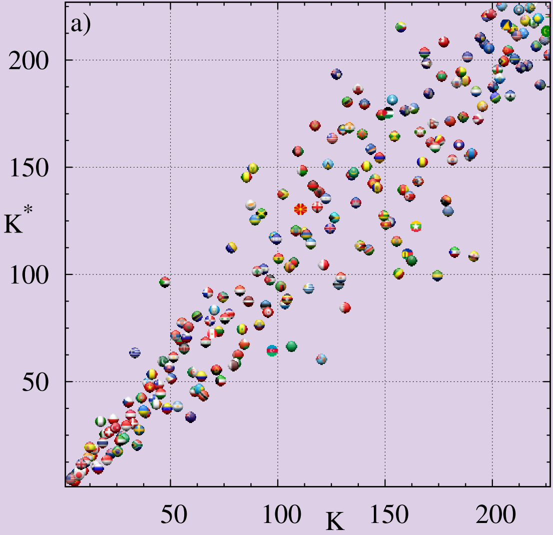
entier Twitter network 2009 => 41 million users



K.Frahm, DS arXiv:1207.3414[cs.SI] (2012)

Ranking of World Trade

UN COMTRADE database 2008: All commodities



Ermann, DS arxiv:1103.5027 (2011)

Google Matrix Applications

practically to everything ....

more data at

http://www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/ .../tradecheirank/