Data Analytics for Big Data

José M. F. Moura
Phillip L. and Marsha Dowd University Professor, CMU
moura@nyu.edu, www.ece.cmu.edu/~moura

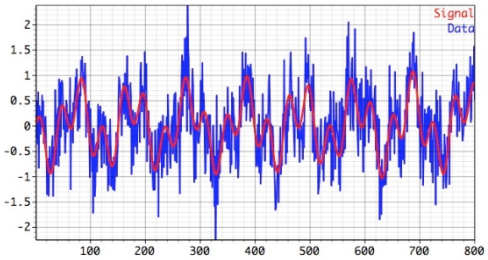
Electrical & Computer ENGINEERING

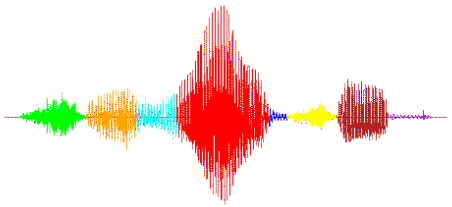
Data Data Data Data Data Data Data Data Data Data Data

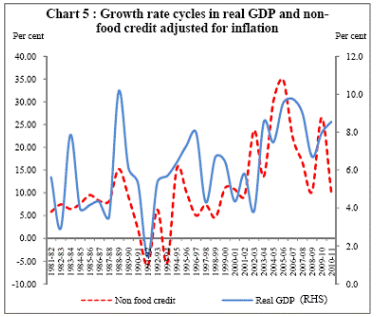
- **Big Data:**
  - Variety
  - Volume
  - Velocity
  - Veracity, Variability, Value, Visualization
  - Unstructured
  - Distributed
  - •••
  - By 2020, all digital data created, replicated, consumed, in a year (IDC, Dec 2012):
    - 40 ZB ≈ 170 M US LoC

Traditional Signals

- Time signals

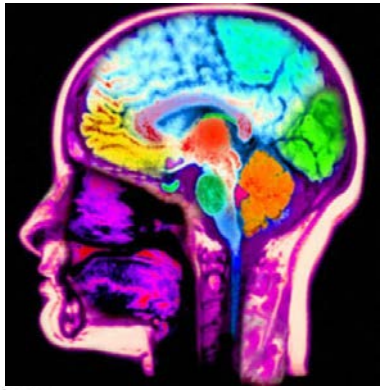






- Images, video

Forbes, 03/05/ 2013



KU Band SAR Image
Sandia Nat Lab

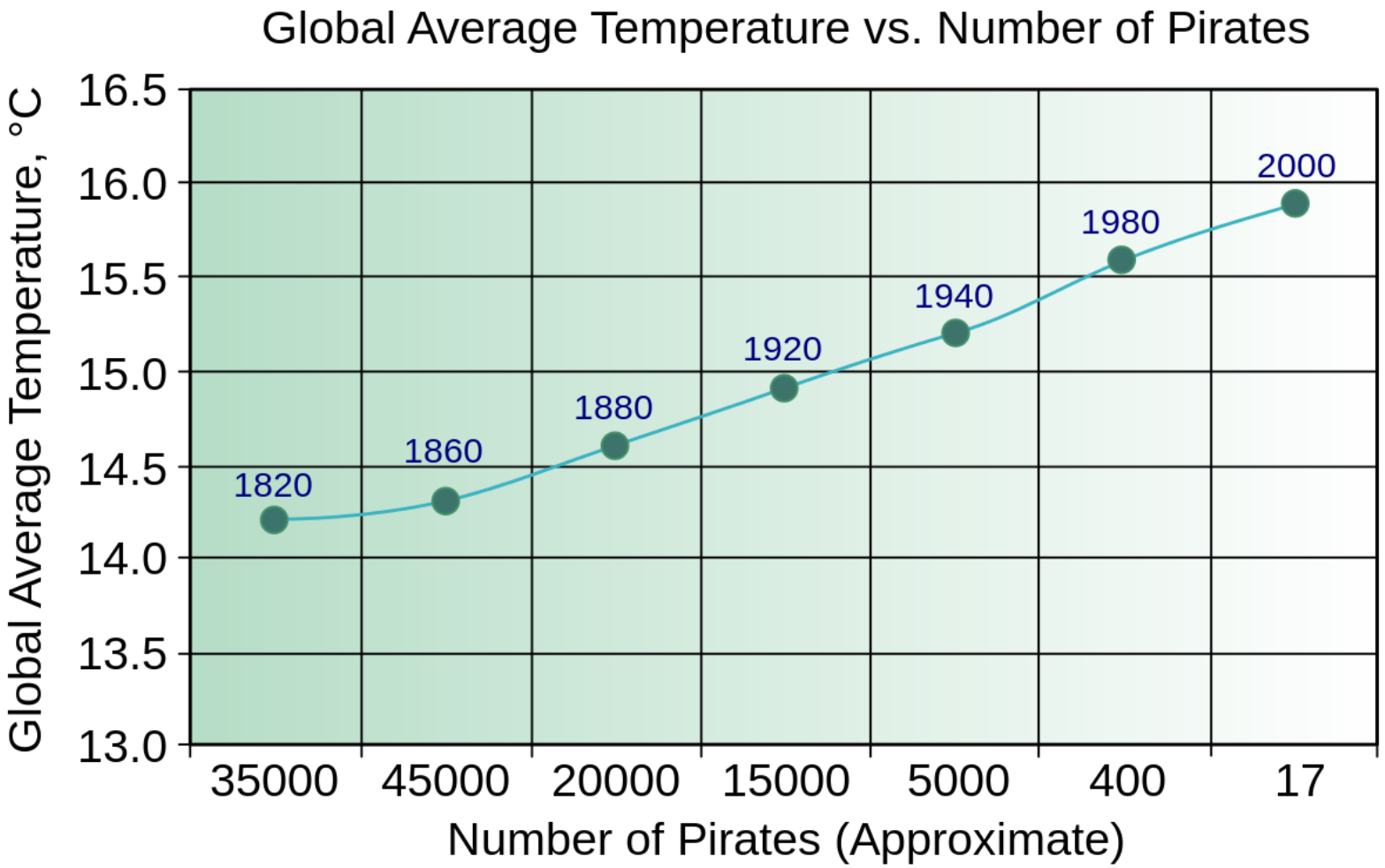
Mendrelic: Time Lapse
http://vimeo.com/18554749

Electrical & Computer ENGINEERING

Data: Volume & Variety

- 190 million records of taxi rides in NYC
  - Time of the day, date, geolocation of start and end, duration, fare paid, tax, tip, # riders, car, driver info, …
- Political blogs:
  - 2004 presidential run Bush vs Kerry, political opiniated blogs, self described as liberal (blue) vs conservative (red), hyperlinked
- 112 million Verizon cell phone subscribers:
  - Every cell phone user, who calls whom for how long, cell phone tower used, location, date and time of day, …
- Health care data:
  - Patient, providers, visit, exams, tests, results, costs, insurance, hospital visits, procedures, …
- Data analytics
  - *Un*structured, can we define a DSP methodology, what is a signal, can we filter the signal, can we Fourier transform it, frequency response of signal and of filter, lowpass, bandpass, highpass signal, …

Correlation vs Causation

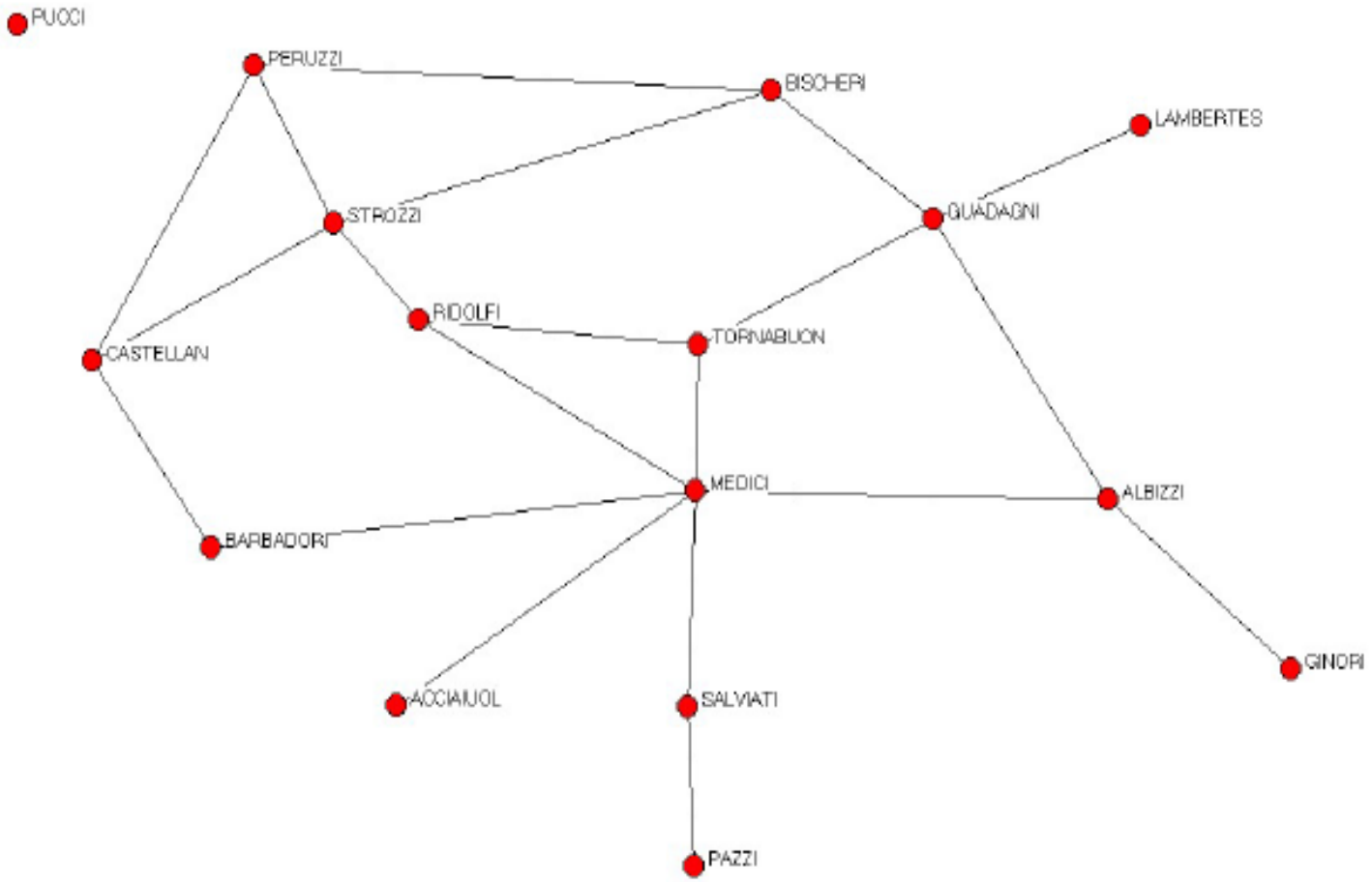


http://web.archive.org/web/20070407182624/http://www.venganza.org/about/open-letter/

Henderson, Bobby (2005). "Open Letter To Kansas School Board". Venganza.org.

Unstructured Data

- **Why were the Medicis so influential in the 15th Century in Florence**

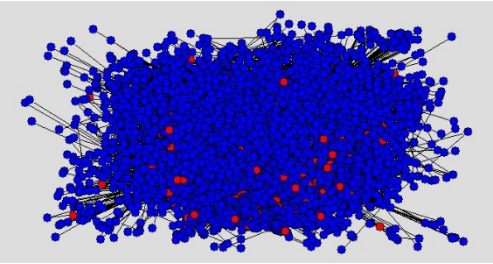


Electrical & Computer ENGINEERING

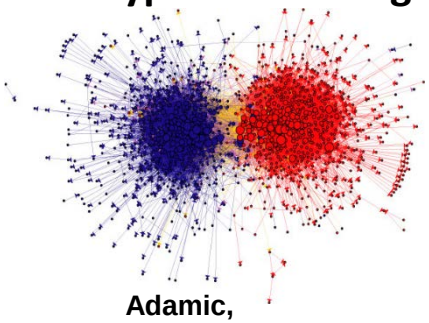
15th Century Florentine Marriages (Padgett and Ansell)

Currently: Social Nets, Web, Company Data, …

Wireless Service Providers

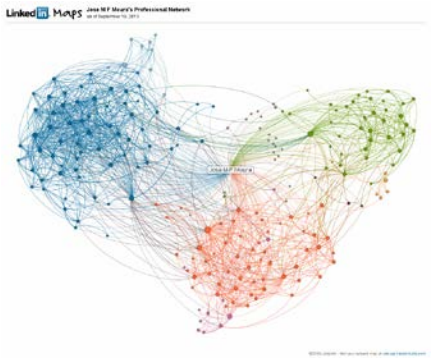


Web: hyperlinked blogs



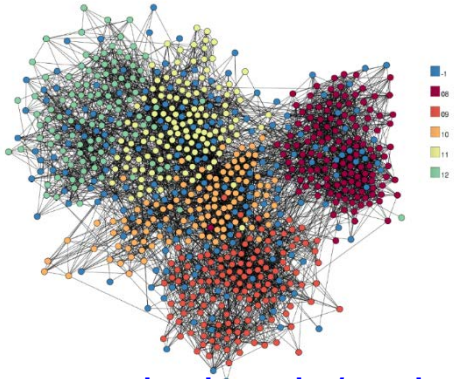
Adamic, Glance

Social networks
Linkedin Contacts

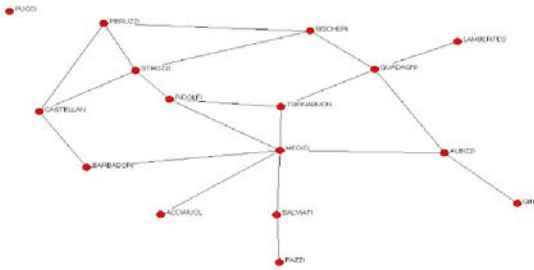


Sensor Networks

Friendship Networks



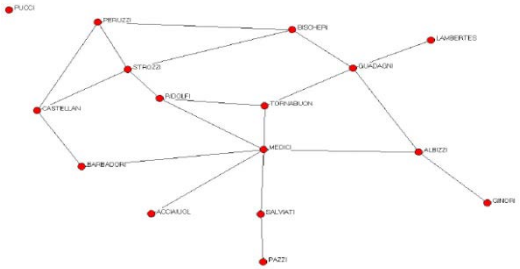
XV Century Florentine Families



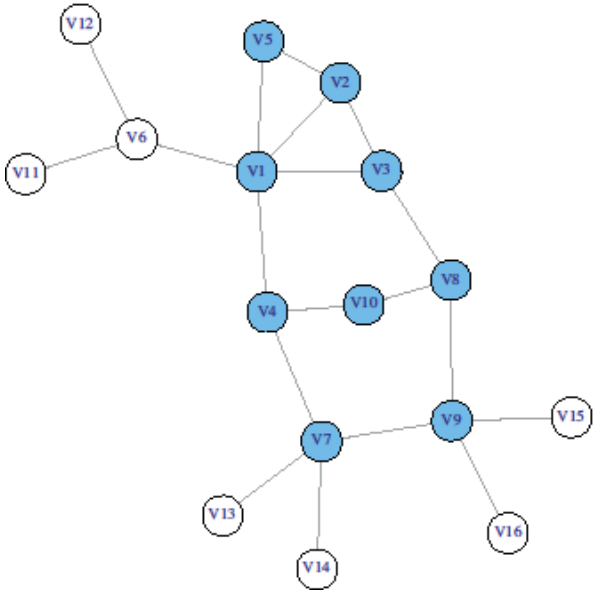
http://vidi.cs.ucdavis.edu/projects/AggressionNetworks/

Networks and Data

- Networks of agents, networks of sensors
  - *Structure* of the network
- Network *processes*
- *Distributed* inference:
  - consensus
  - consensus+innovations
- Data analytics

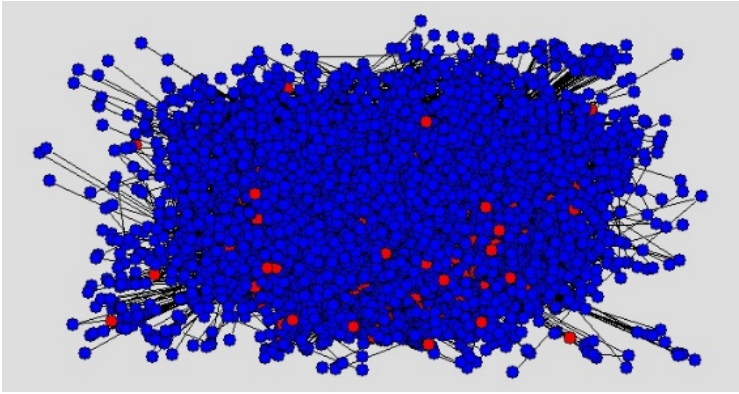


15th Century Florentine Marriages
(Padgett and Ansell)



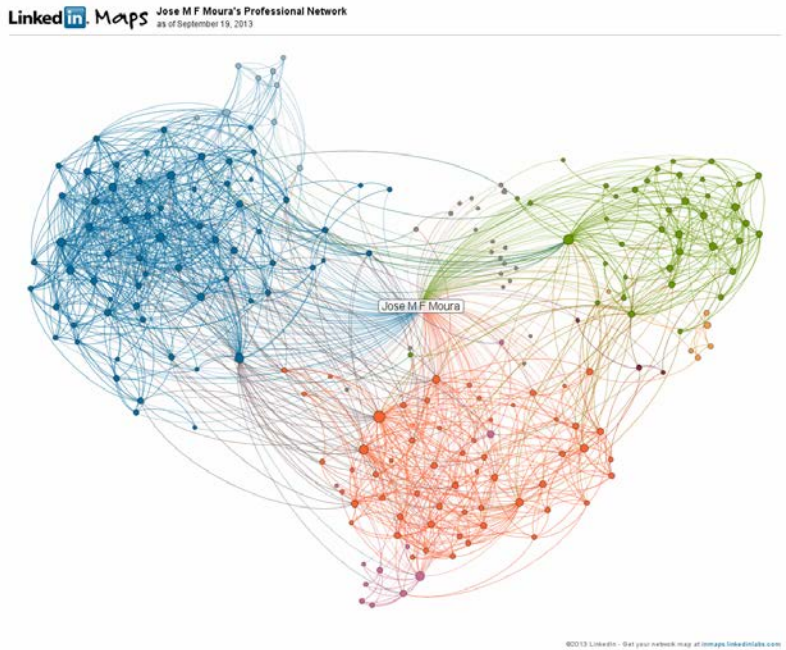
Data Science: Graph+Data

- Graphical models: Markov random fields
  - Machine Learning approaches (Jordan, Willsky, …)
- Data transforms:
  - Regression analysis, wavelets on irregular sensor network (Baraniuk), filter banks (Vandergheynst), diffusion wavelets (Coifman)
  - A eigen spectral analysis based on the graph Laplacian (assumed undirected, non-negative weights) (Vandergheynst, Barbarossa, Ortega …)
- Discrete Signal Processing on graphs:
  - Sandryhaila and Moura (T-SP 2013, May; 2014, April)



Discrete Signal Processing on Graphs

- (Linear) DSP for social, biological and physical networks data

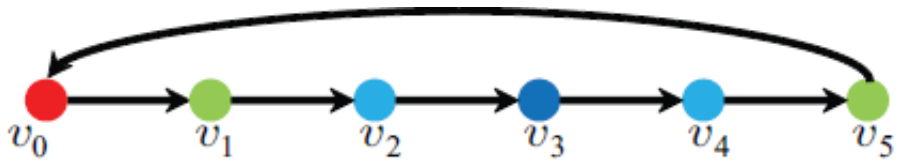


- Graph signal model, filters, filtering and convolution, impulse response, *z*- and Fourier transforms, spectrum, frequency response, …

Sandryhaila & Moura, "DSP on graphs," IEEE Tr-SP, 2013

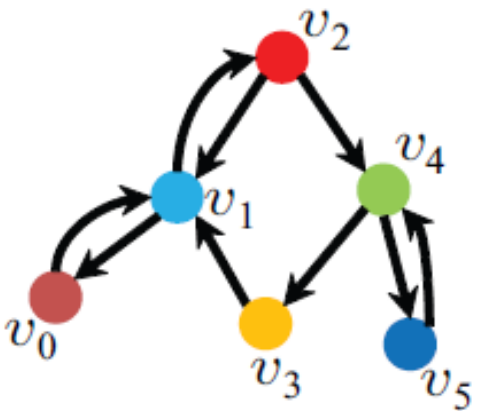
Graph Signals

- **Time signals:**
  - **Cosine signal** $\cos\left(\frac{2\pi}{6}k\right)$, **k=0, …, 5**

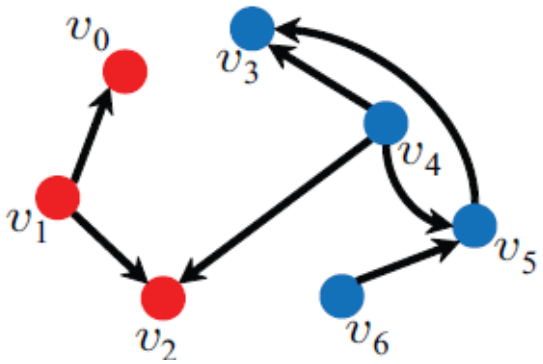


- **Average temperature in US cities**

- **Website topics in hyperkinked blogs**



- **Average # tweets**



Electrical & Computer ENGINEERING

DSP_G: Graph Fourier Transform

- For simplicity assume graph adjacency matrix is diagonalizable

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1} = [\mathbf{v}_0, \cdots, \mathbf{v}_{n-1}] \begin{bmatrix} \lambda_0 & & \\ & \ddots & \\ & & \lambda_{n-1} \end{bmatrix} = [\mathbf{v}_0, \cdots, \mathbf{v}_{n-1}]^{-1}$$

- Graph frequencies:

$$\lambda_0, \cdots, \lambda_{n-1}$$

- Graph frequency components:

$$\mathbf{v}_0, \cdots, \mathbf{v}_{n-1}$$

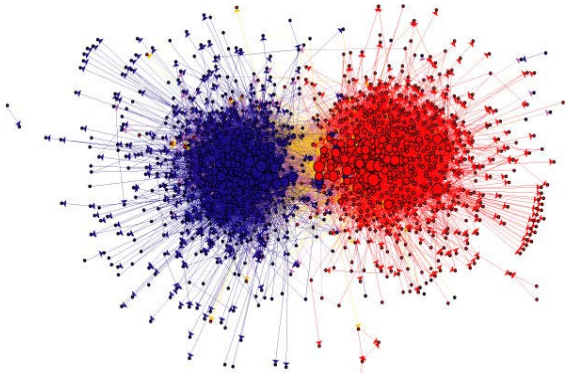
- Graph Fourier transform:

$$\widehat{\mathbf{S}} = \begin{bmatrix} \widehat{S}_0 \\ \vdots \\ \widehat{S}_{n-1} \end{bmatrix} = \mathbf{V}^{-1}\mathbf{s}$$

$$= \mathbf{F}\mathbf{s}$$

DSP_G: Political Blogs

- **Graph signal:** 1224 conservative & liberal political blogs
  - Graph: 1224 blogs are the nodes, hyperlinks are the edges

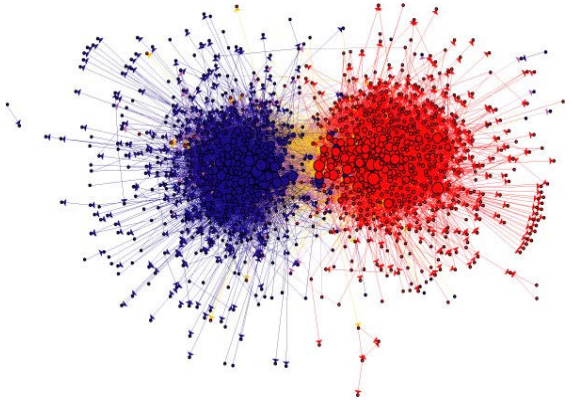


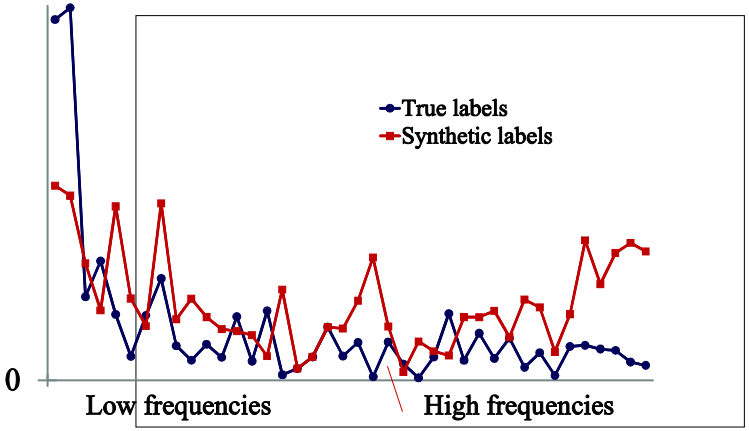
Adamic, Glance

  - Graph signal: blue ⟹ 1, red ⟹ - 1
  - Shift: Adjacency matrix $\mathbf{A}$ given by hyperlinks among blogs

DSP_G: Blogosphere Low- vs High Pass

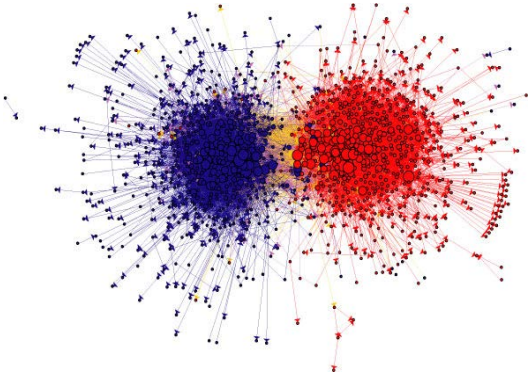
- 1224 political blogs, hyperlinked: conservative & liberal
  - Adjacency matrix A given by hyperlinks



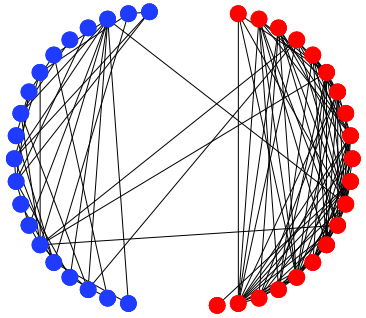


DSP_G: Classification–Political Blogsphere

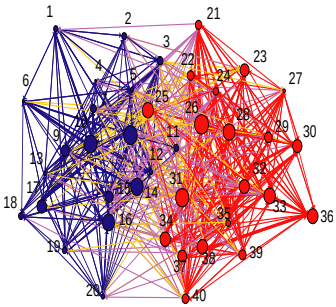
- **Political blogs:**



Adamic, Glance

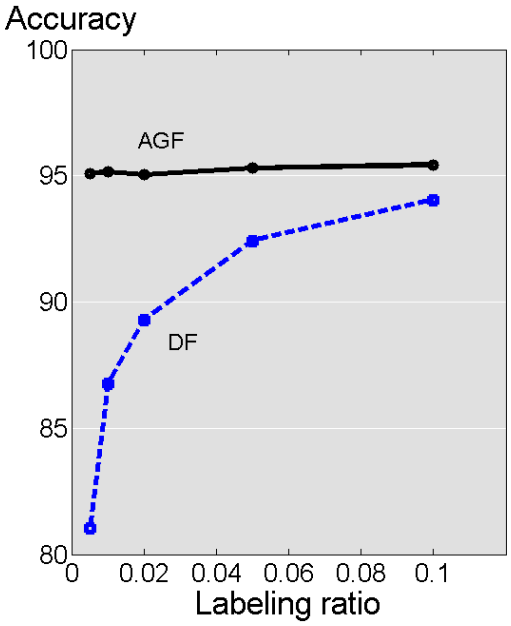


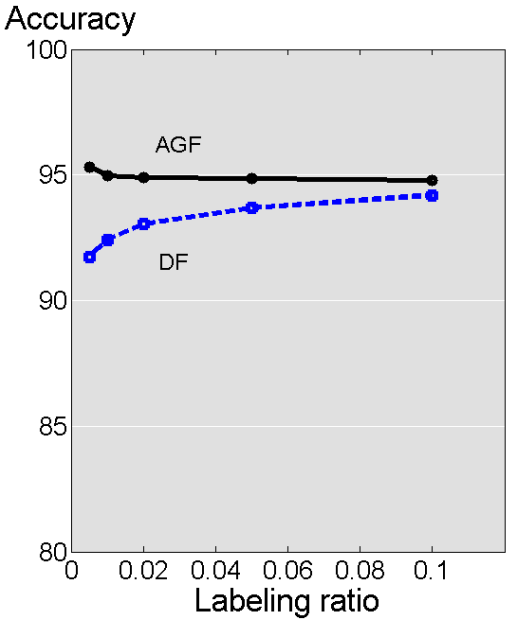
Random



Most connected

- **Classifier *P=10*:** $h(\mathbf{A}) = (\mathbf{I} + h_p\mathbf{A})(\mathbf{I} + h_{p-1}\mathbf{A}) \cdots (\mathbf{I} + h_1\mathbf{A})$





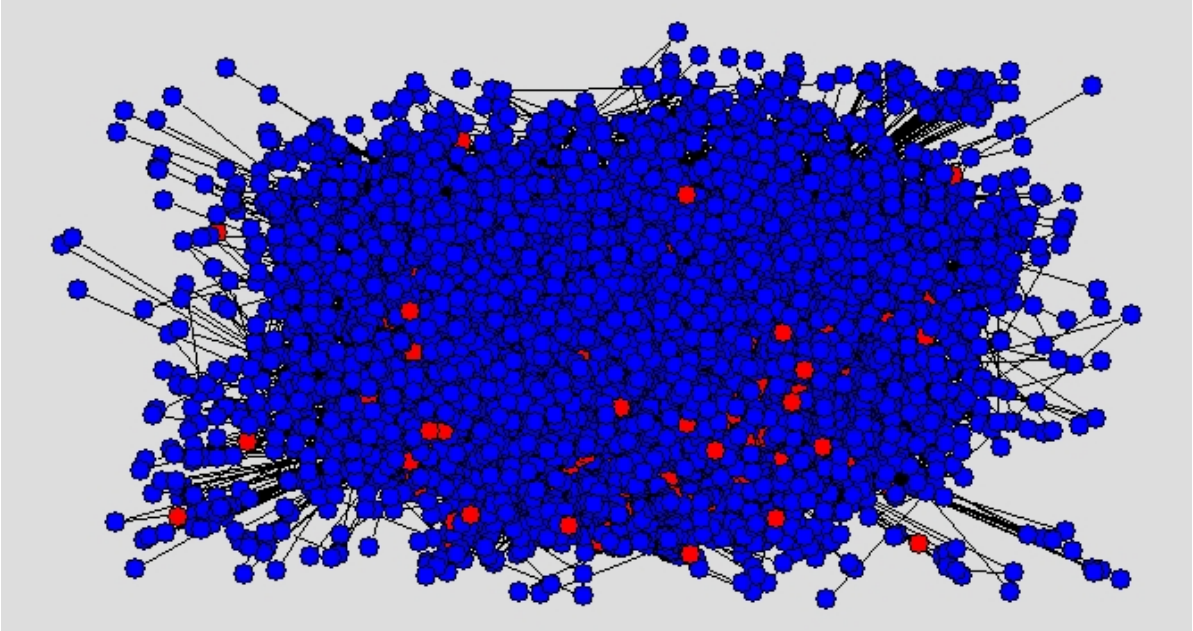
AGF: Globalsip, 2013: Chen. Sandryhaila, Moura, Kovacevic

DF: Diffusion Functions, 2008, J. Mach Learn. Res., Szlam, Coiffman, Maggioni

Electrical & Computer ENGINEERING

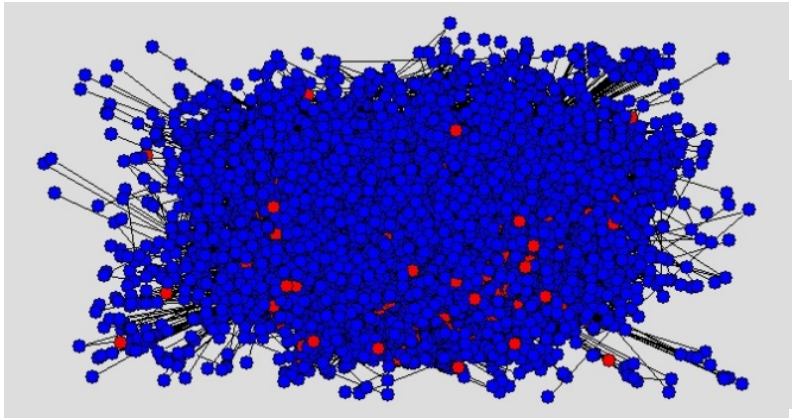
DSP_G: Service Provider–Predict Customer Behavior

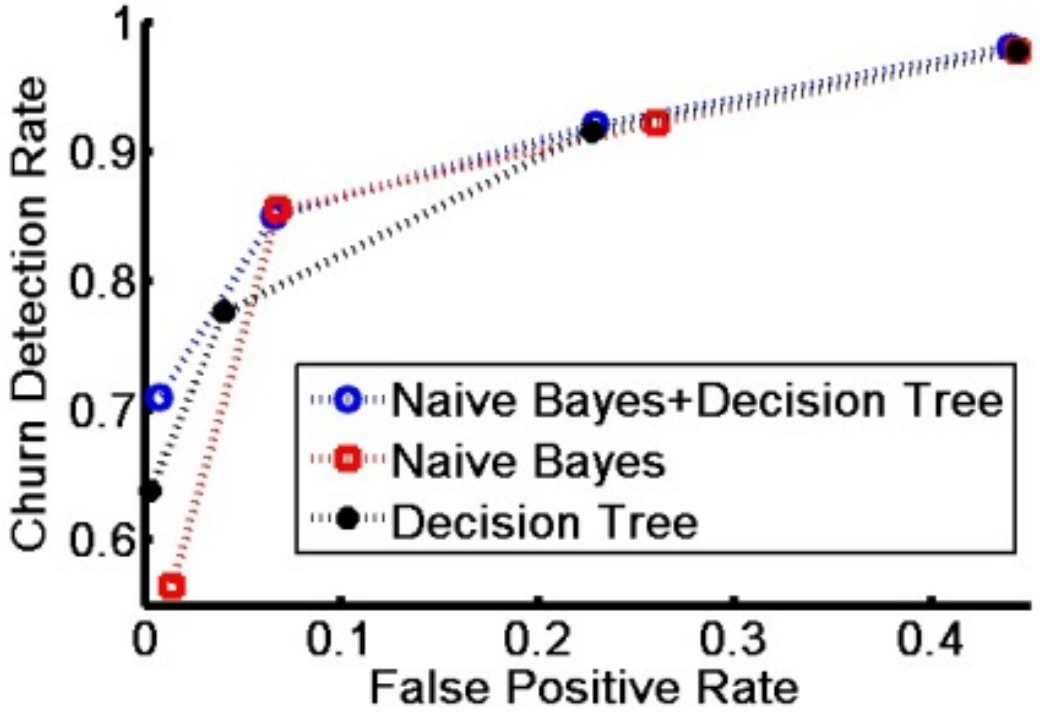
- **3.7 Million customers: 3.6 M non-churners, 100K churners**
- **Adjacency matrix:** $A_{n,m} = \frac{T_{n,m}}{\sum_{m \in \eta_n} T_{n,m}}$
- **10 months log: Learn from few churners in month A who will churn month A+1**



DSP_G: Service Provider–Predict Customer Behavior

- **Classifier**





Deri & Moura, ICASSP, May 2014

DSP_G: News Article Dataset–Classification

- **News articles dataset:** Belkin, Matveeva, Nyogi, "Regularization and semi-supervised learning on large graphs,"Conf. Learn. Th., 2004

Graph: 18,000 news articles, 20 different topics, construct graph from randomly selected 500 from each class, each article a vector of 6000 most common keywords, use cosine distance between keywords:

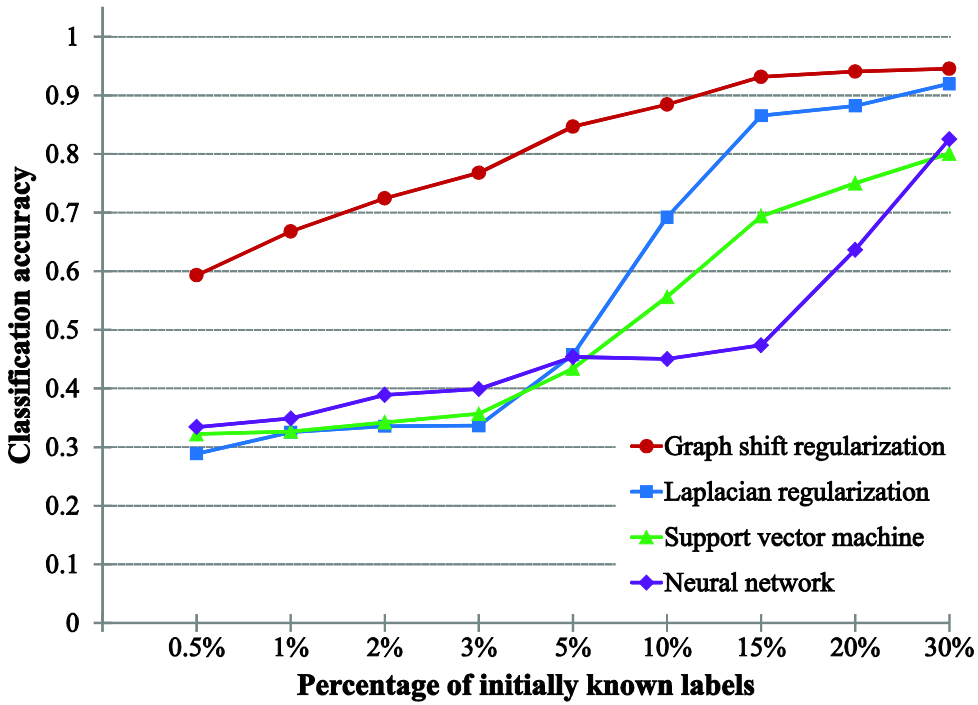
$$d_{n,m} = \frac{\langle \mathbf{v}_n, \mathbf{v}_m \rangle}{||\mathbf{v}_n||_2 \; ||\mathbf{v}_m||_2}.$$

$$\mathbf{A}_{n,m} = 1 - d_{n,m}.$$

$$\mathrm{TV}_G(\mathbf{s}) = \frac{1}{||\mathbf{s}||_2^2} \left|\left| \mathbf{s} - \frac{1}{|\lambda_{max}|} \mathbf{A}\mathbf{s} \right|\right|_2^2$$

$$\mathbf{s}^{(\text{predicted})} = \underset{\mathbf{s} \in \mathbb{R}^N}{\operatorname{argmin}} \; \mathrm{TV}_G(\mathbf{s}).$$

$$|| \mathbf{C}\,\mathbf{s}^{(\text{known})} - \mathbf{C}\,\mathbf{s}||_2^2 < \epsilon$$

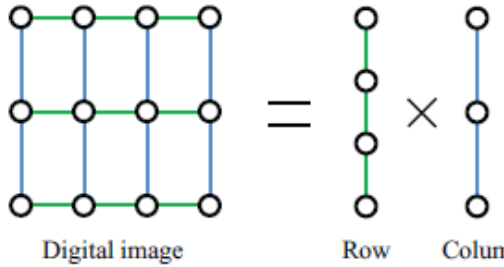


Electrical & Computer ENGINEERING

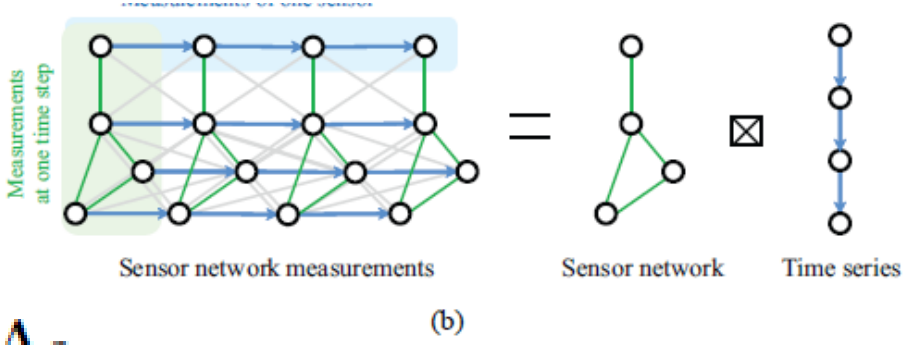
Globalsip, 2013: Sandryhaila, Moura

Big Data

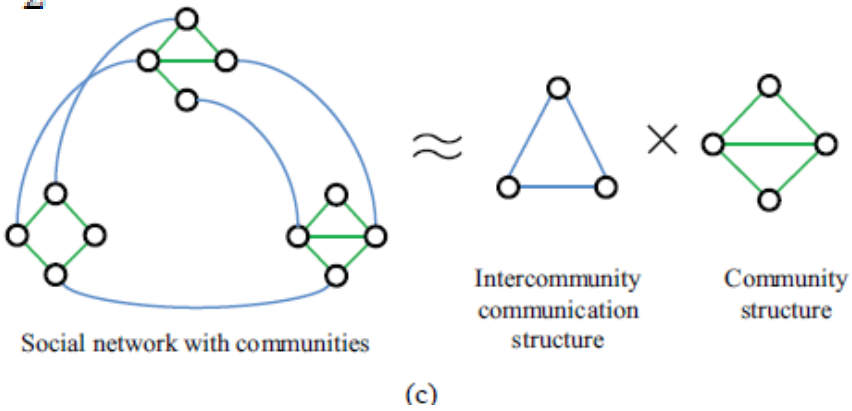
- $\mathbf{A}_{\times} = \mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2.$



- $\mathbf{A}_{\boxtimes} = \mathbf{A}_1 \otimes \mathbf{A}_2 + \mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2.$



- $\mathbf{A}_{\times} = \mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2.$



Big Data

- **Parallel constructs:**

$$\Big(\mathbf{I}_3 \otimes \mathbf{A}\Big)\mathbf{s} = \begin{bmatrix} \mathbf{A} & & \\ & \mathbf{A} & \\ & & \mathbf{A} \end{bmatrix} \mathbf{s} = \begin{bmatrix} \mathbf{A}\begin{bmatrix} s_0 \\ s_1 \end{bmatrix} \\ \mathbf{A}\begin{bmatrix} s_2 \\ s_3 \end{bmatrix} \\ \mathbf{A}\begin{bmatrix} s_4 \\ s_5 \end{bmatrix} \end{bmatrix}.$$

- **Vectorizable constructs:**

$$\Big(\mathbf{A} \otimes \mathbf{I}_3\Big)\mathbf{s} = \begin{bmatrix} a_{00}\begin{bmatrix} s_0 \\ s_1 \\ s_2 \end{bmatrix} + a_{01}\begin{bmatrix} s_3 \\ s_4 \\ s_5 \end{bmatrix} \\ a_{10}\begin{bmatrix} s_0 \\ s_1 \\ s_2 \end{bmatrix} + a_{11}\begin{bmatrix} s_3 \\ s_4 \\ s_5 \end{bmatrix} \end{bmatrix}.$$

Big Data: Fourier Transform

- **Cartesian graphs:** $\mathbf{V} = \mathbf{V}_1 \otimes \mathbf{V}_2,$

$$\mathbf{A}_\times = \mathbf{V}(\mathbf{\Lambda}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{\Lambda}_2)\,\mathbf{V}^{-1}.$$

$$\mathbf{F}_\times = (\mathbf{V}_1 \otimes \mathbf{V}_2)^{-1} = \mathbf{V}_1^{-1} \otimes \mathbf{V}_2^{-1} = \mathbf{F}_1 \otimes \mathbf{F}_2,$$

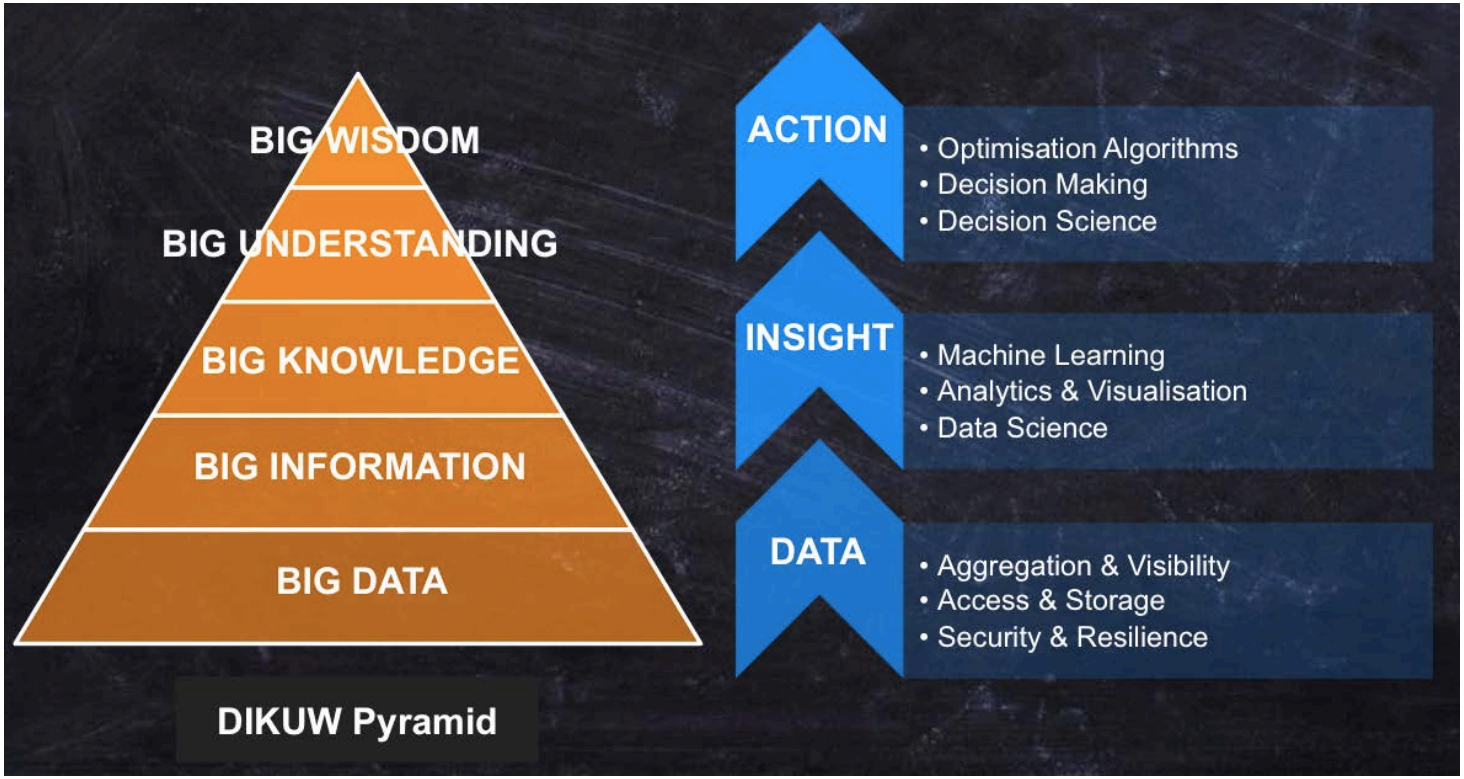
$$\mathbf{F}_1 \otimes \mathbf{F}_2 = (\mathbf{F}_1 \otimes \mathbf{I}_{N_2})(\mathbf{I}_{N_1} \otimes \mathbf{F}_2)$$

- **Kronecker graphs and Strong graphs:**

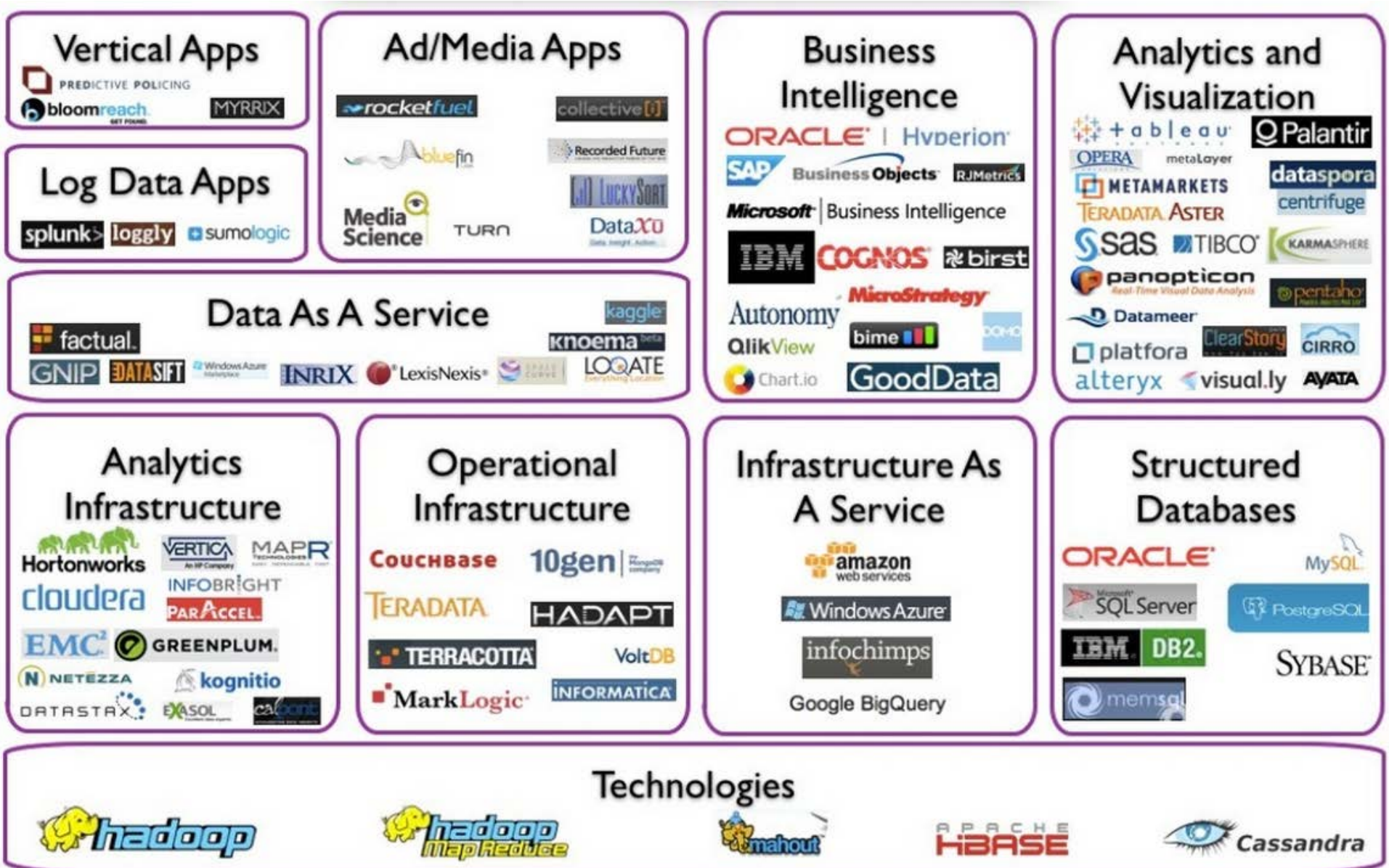
$$\mathbf{A}_\otimes = \mathbf{V}(\mathbf{\Lambda}_1 \otimes \mathbf{\Lambda}_2)\,\mathbf{V}^{-1},$$

$$\mathbf{A}_\boxtimes = \mathbf{V}(\mathbf{\Lambda}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{\Lambda}_2 + \mathbf{\Lambda}_1 \otimes \mathbf{\Lambda}_2)\,\mathbf{V}^{-1}.$$

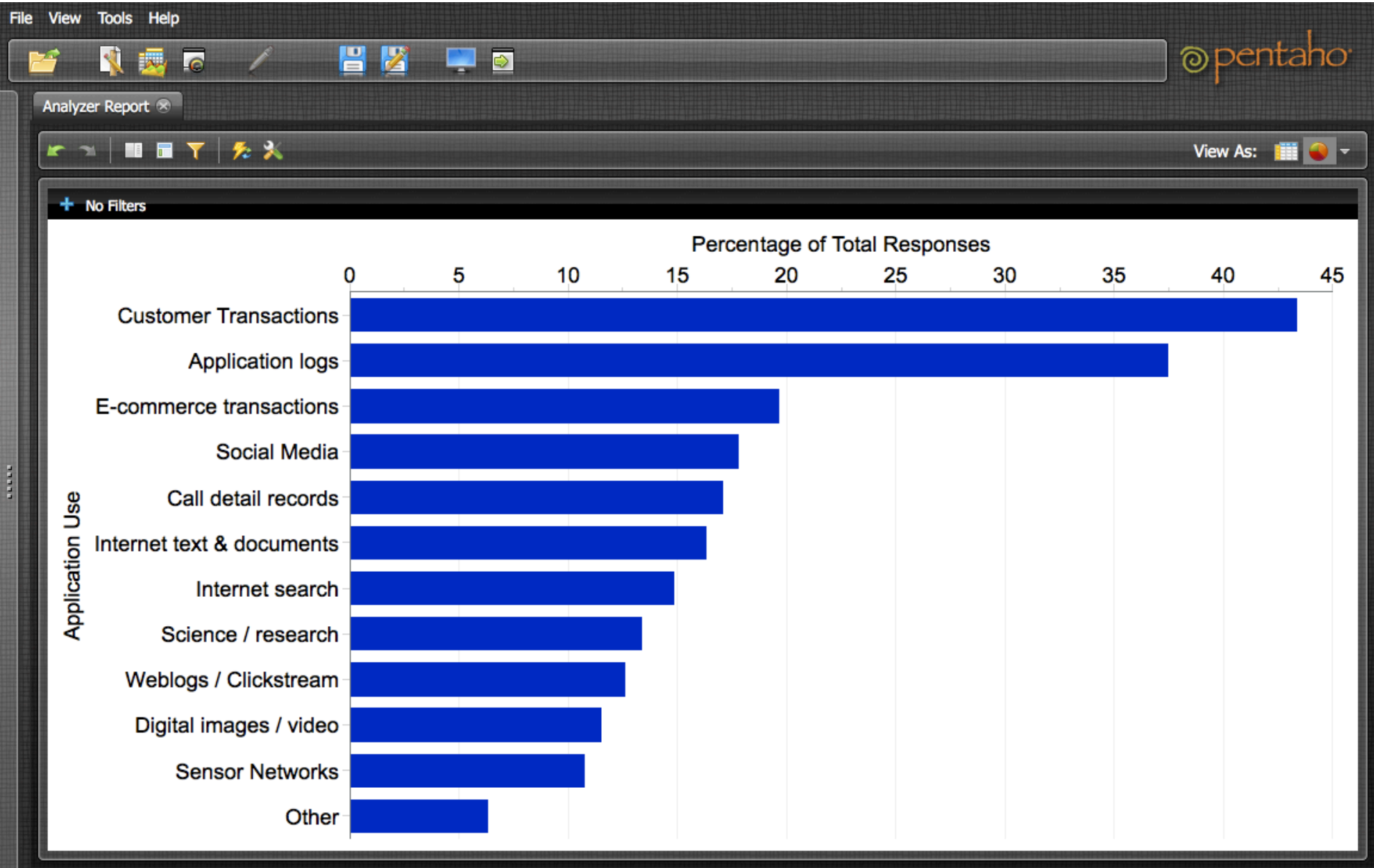
Big Data Pyramid (DIKUW)



Big Data Space



Big Data: Applications



Conclusion

- Big Data Analytics: complex relations among data captured by a graph
- Shift: adjacency matrix
- Filters: polynomials in the adjacency matrix
- Graph Fourier transform: eigen decomposition of adjacency matrix
- Range of applications
- IEEE Signal Processing Society activity in Big Data:
- New Transactions on Networked Systems and Data
- Special issues: SP Magazine, May 2013, Jun 2014; Transactions on Selected Topics on SP

Special Issues IEEE Transactions on Special Topics in SP

- Signal and Information Processing for Privacy
- Anomalous Pattern Discovery for Spatial, Temporal, Networked, and High-Dimensional Signals (Feb 2013)
- Adaptation and Learning over Complex Networks (Apr 2013
- Signal Processing for Social Networks (Aug 2014))
- Signal Processing for Situational Awareness from Networked Sensors and Social Media (Mar 2015)
- Signal Processing for Big Data (Jun 2015)

Special Issues IEEE SP Magazine

- Signal Processing for Financial Applications (Sep 2011)
- Special Issue on Genomic and Proteomic Signal Processing in Biomolecular Pathways (Jan 2012)
- Adaptation and Learning over Complex Networks (May 2013)
- Signal Processing for Cyber-Security and Privacy (Sep 2013)
- Signal Processing for Big Data (Sep 2014)

IEEE Transactions on Information Forensics and Security

- Privacy and Trust Management in Cloud and Distributed Systems (Jun 2013)

GlobalSip: 2013

- Advancing Neural Engineering Through Big Data
- Bioinformatics and Systems Biology
- Controlled Sensing For Inference: Applications, Theory and Algorithms
- Cyber-Security and Privacy
- Emerging Challenges in Network Sensing, Inference, and Communication
- Energy Harvesting and Green Wireless Communications
- Graph Signal Processing
- Information Processing in the Smart Grid
- Information Processing over Networks
- Network Theory
- Optimization in Machine Learning and Signal Processing
- Signal and Information Processing in Finance and Economics

GlobalSip: 2014

- Information Processing for Big Data
- Signal Processing Applications Related to Animal Environments
- Signal and Information Processing for Energy Exchange and Intelligent Trading
- Game Theory for Signal Processing and Communications
- Network Theory

Big Data SiG

THANKS

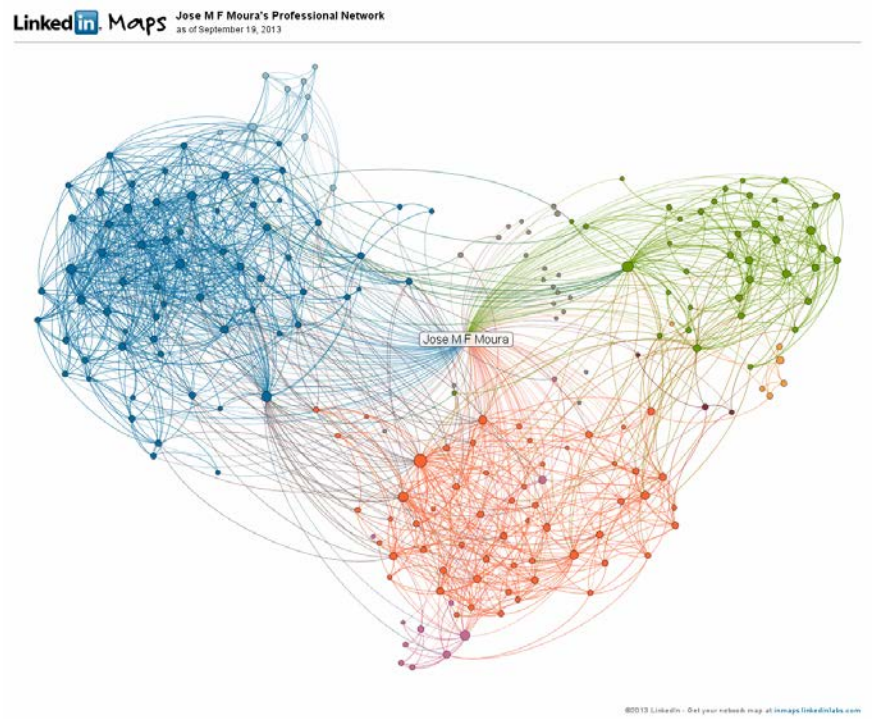
Urban Science: What Is in a Name?

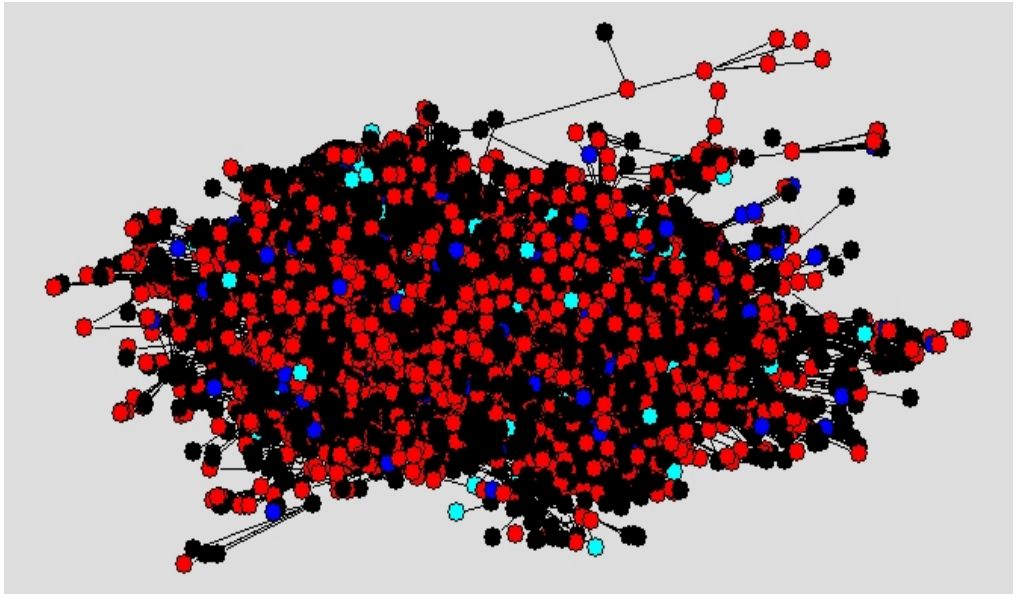
- Google search leads to *interesting* obvious and nonobvious links on “urban science”
- **CUSP:** Center for Urban Science and Progress – does come on on top
- **Urban Science**
  - Founded in 1977, Urban Science is a business-solutions company focused on supporting the needs of the sales and marketing function of the global automotive industry. Leveraging a scientific methodology, Urban Science helps its client partners sell more vehicles, improve profitability and increase customer loyalty. With headquarters in Detroit, Urban Science serves its global clientele from offices in Beijing, Frankfurt, London, Los Angeles, Madrid, Melbourne, Mexico City, Moscow, Delhi NCR, Munich, Nashville, New York, Paris, Rome, Sao Paulo, Shanghai, Tokyo and Washington, D.C.

    For more information on Urban Science, visit http://www.urbanscience.com
- **Urban Science Academy**: middle school serving the Concourse, Melrose, and Morrisania neighborhoods of the South Bronx with instruction in grades 6 through 8.

Paradigm of Big Data: Networked Data

- **Big Data:**
  - Time series, speech, audio, images, video, visual …
  - Physical, sensor networks, large infrastructures, cyberphysical, …
  - Media, social media, biology, advertising, service ind., financial, …

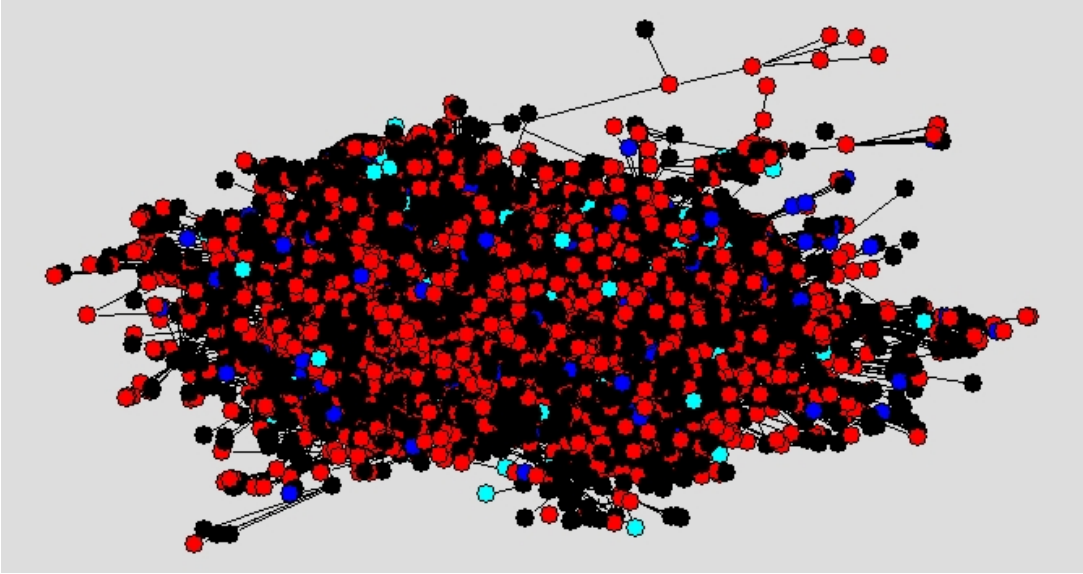




red = "nokia", blue="motorola", cyan="samsung", black = "other"

Networked Data: Graph+Data

- Cell phone users: wireless service provider

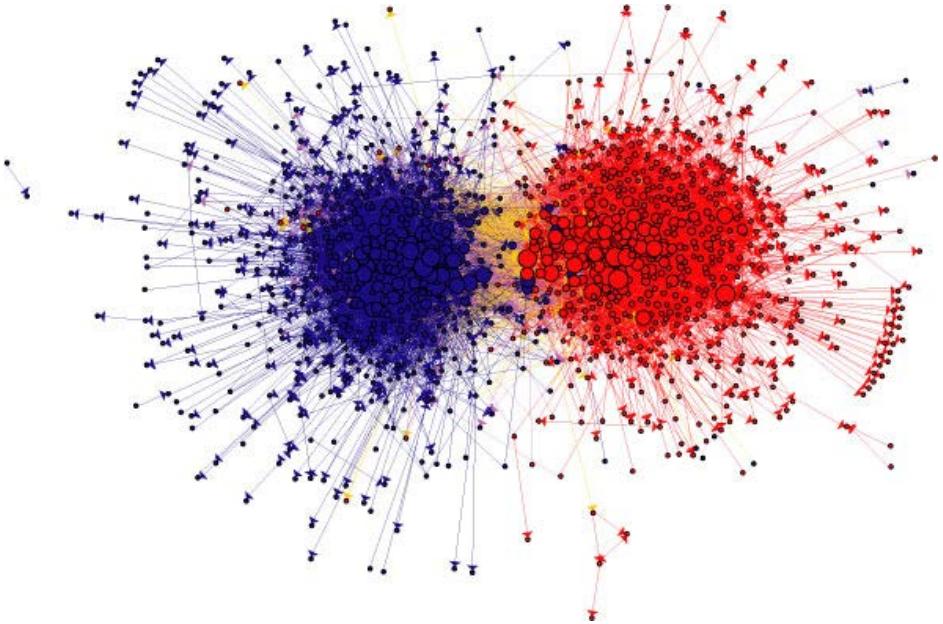


red = "nokia", blue="motorola", cyan="samsung", black = "other"

- Churning: who drops service at a given month

Social Media

- Political blogosphere (US 2004 election) (1224 blogs, L & R)

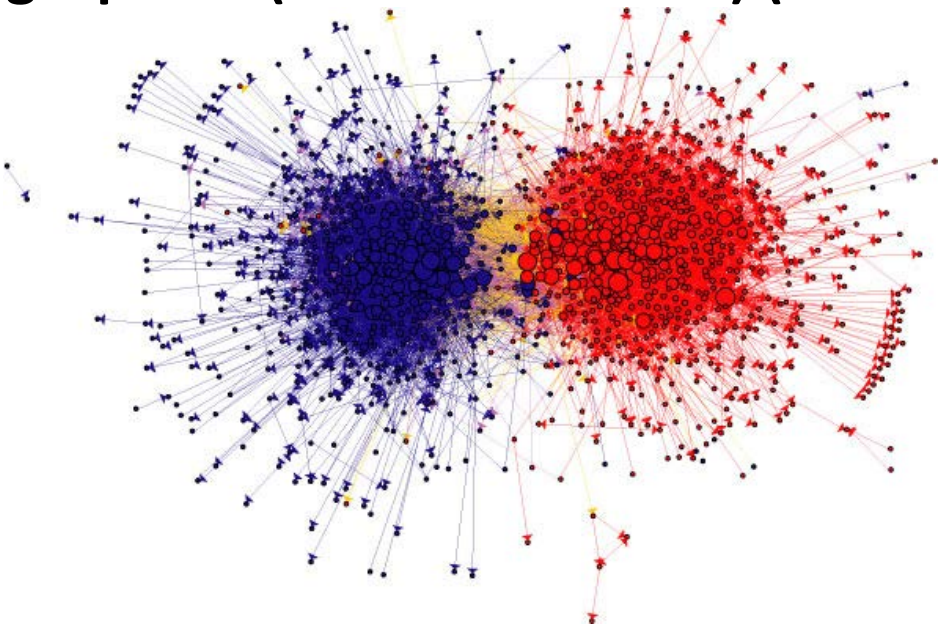


Lada Adamic, U Michigan, Natalie Glance @ Intelliseek

- Determine label of blogs from known labels of a few bloggers

Social Media: Graph Signal

- Political blogosphere (US 2004 election) (1224 blogs, L & R)

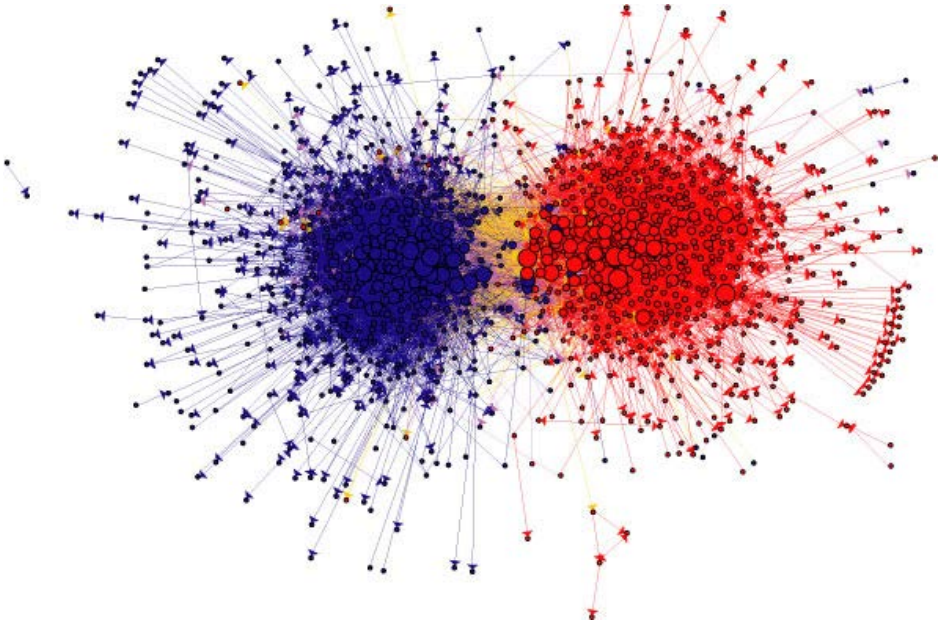


Lada Adamic, U Michigan, Natalie Glance @ Intelliseek

- Structure: relations among data, captured by a graph $\mathcal{G} = (V,A)$
- Graph signal (single snapshot): s: $V \Longrightarrow C, L$

Social Media: Graph Signal

- Political blogosphere (US 2004 election) (1224 blogs, L & R)



Lada Adamic, U Michigan, Natalie Glance @ Intelliseek

- Structure: relations among data, captured by a graph $\mathcal{G} = (V,A)$
- Graph signal (single snapshot): s: $V \Longrightarrow C, L$