INFX 574
Applied Machine Learning

Spring 2018

How would you build a search engine?

Most Referenced Big Data Contributions

Stalzer & Mentzel (2015, arXiv

Count	Year	Citation
63	2008	MapReduce[1]
51	2009	*Fourth Paradigm*[2]
43	2009	*Elements of Statistical Learning*[3]
30	2001	Initial sequencing of the human genome[4]
24	1948	A mathematical theory of communication[5]
23	2000	Sloan Digital Sky Survey[6]
20	1990	BLAST[7]
19	1996	Lasso[8]
19	2003	Latent Dirichlet allocation[9]
17	1977	EM algorithm[10]
17	1995	Support vector networks[11]
15	2001	Random forests[12]
14	2006	*Pattern Recognition*[13]
14	1998	Anatomy of web search engine[14]
13	2007	*Numerical Recipes*[15]
11	1979	Bootstrap methods[16]
11	1953	Equation of state calculations[17]
11	1977	Exploratory data analysis[18]
11	1988	*Probabilistic reasoning*[19]
10	1999	PageRank[20]
10	2013	*Bayesian Data Analysis*[21]
10	2009	Unreasonable effectiveness of data[22]

TABLE I

WORKS THAT WERE CITED AT LEAST TEN TIMES, WITH COUNT, YEAR, AND CITATION.

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1 Introduction and Motivation

The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous. Current estimates are that there are over 150 million web pages with a doubling life of less than one year. More importantly, the web pages are extremely diverse, ranging from "What is Joe having for lunch today?" to journals about information retrieval. In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions.

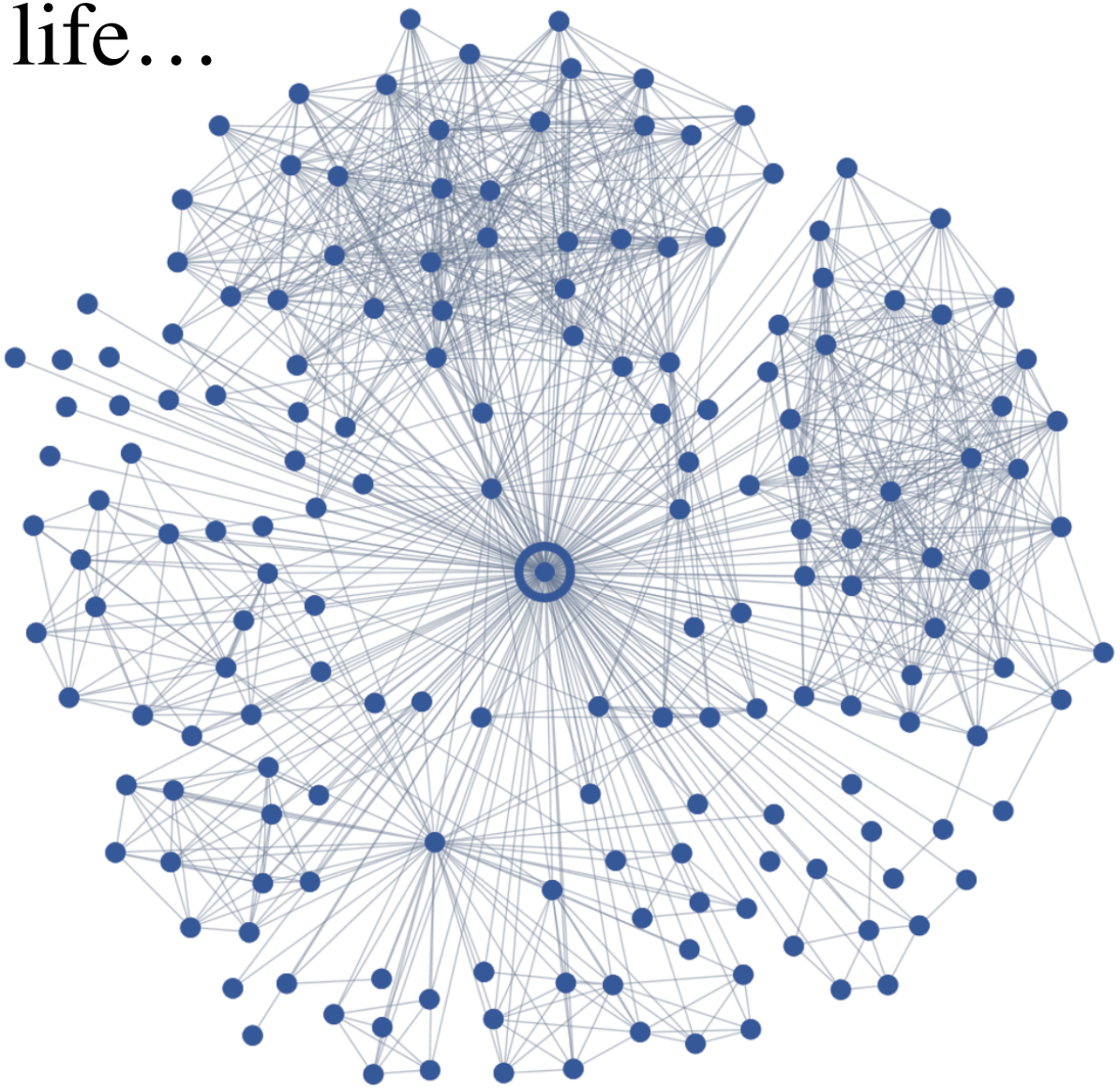
However, unlike "flat" document collections, the World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages, such as link structure and link text. In this paper, we take advantage of the link structure of the Web to produce a global "importance" ranking of every web page. This ranking, called PageRank, helps search engines and users quickly make sense of the vast heterogeneity of the World Wide Web.

1.1 Diversity of Web Pages

Although there is already a large literature on academic citation analysis, there are a number of significant differences between web pages and academic publications. Unlike academic papers which are scrupulously reviewed, web pages proliferate free of quality control or publishing costs. With a simple program, huge numbers of pages can be created easily, artificially inflating citation counts. Because the Web environment contains competing profit seeking ventures, attention getting strategies evolve in response to search engine algorithms. For this reason, any evaluation strategy which counts replicable features of web pages is prone to manipulation. Further, academic papers are well defined units of work, roughly similar in quality and number of citations, as well as in their purpose – to extend the body of knowledge. Web pages vary on a much wider scale than academic papers in quality, usage, citations, and length. A random archived message posting

What is a network?

What your Facebook network tells you about your love life…



Backstrom, L. and Kleinberg, J (2014) arXiv. 1310.6753v1

Immune signalling network

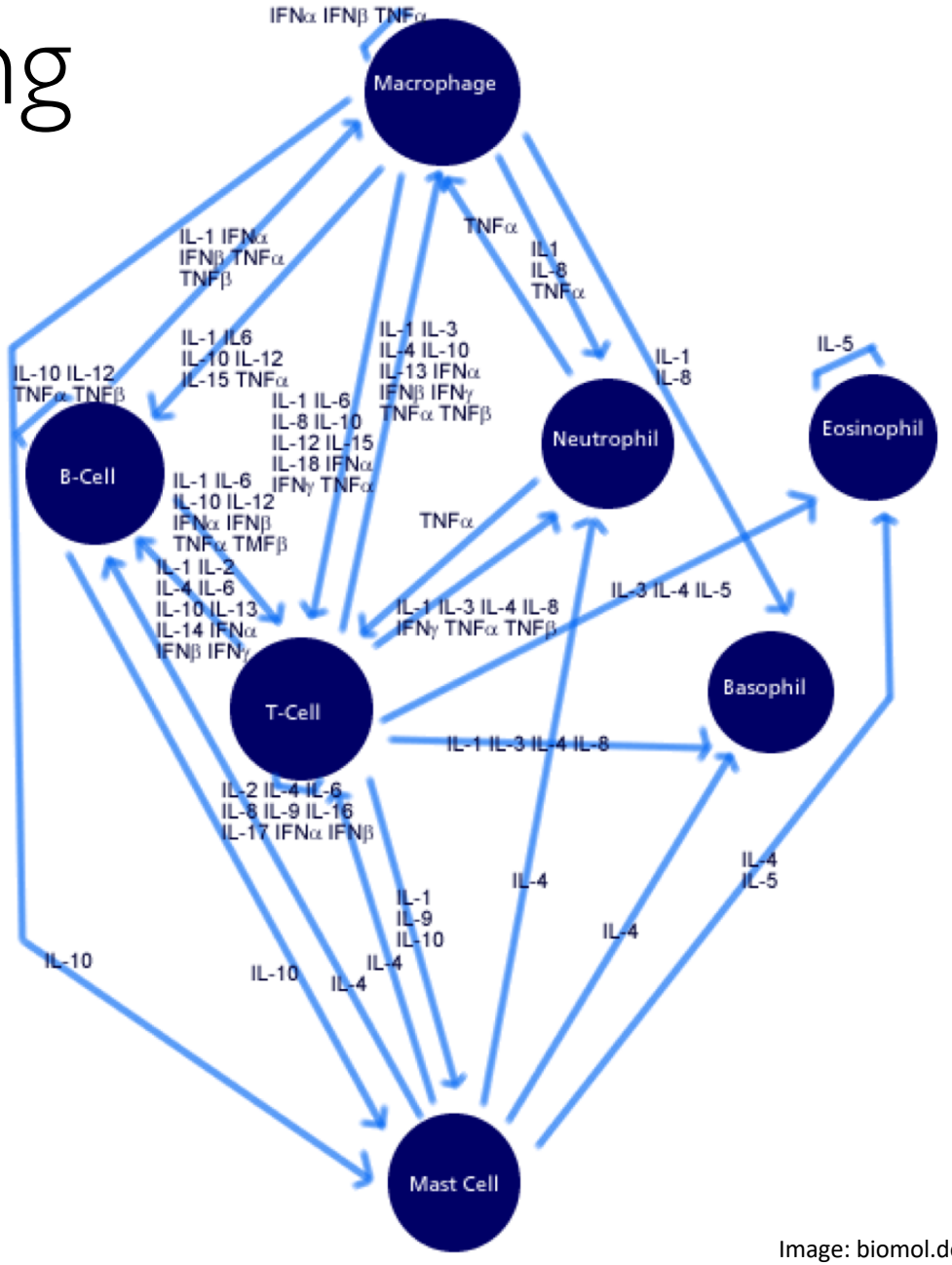
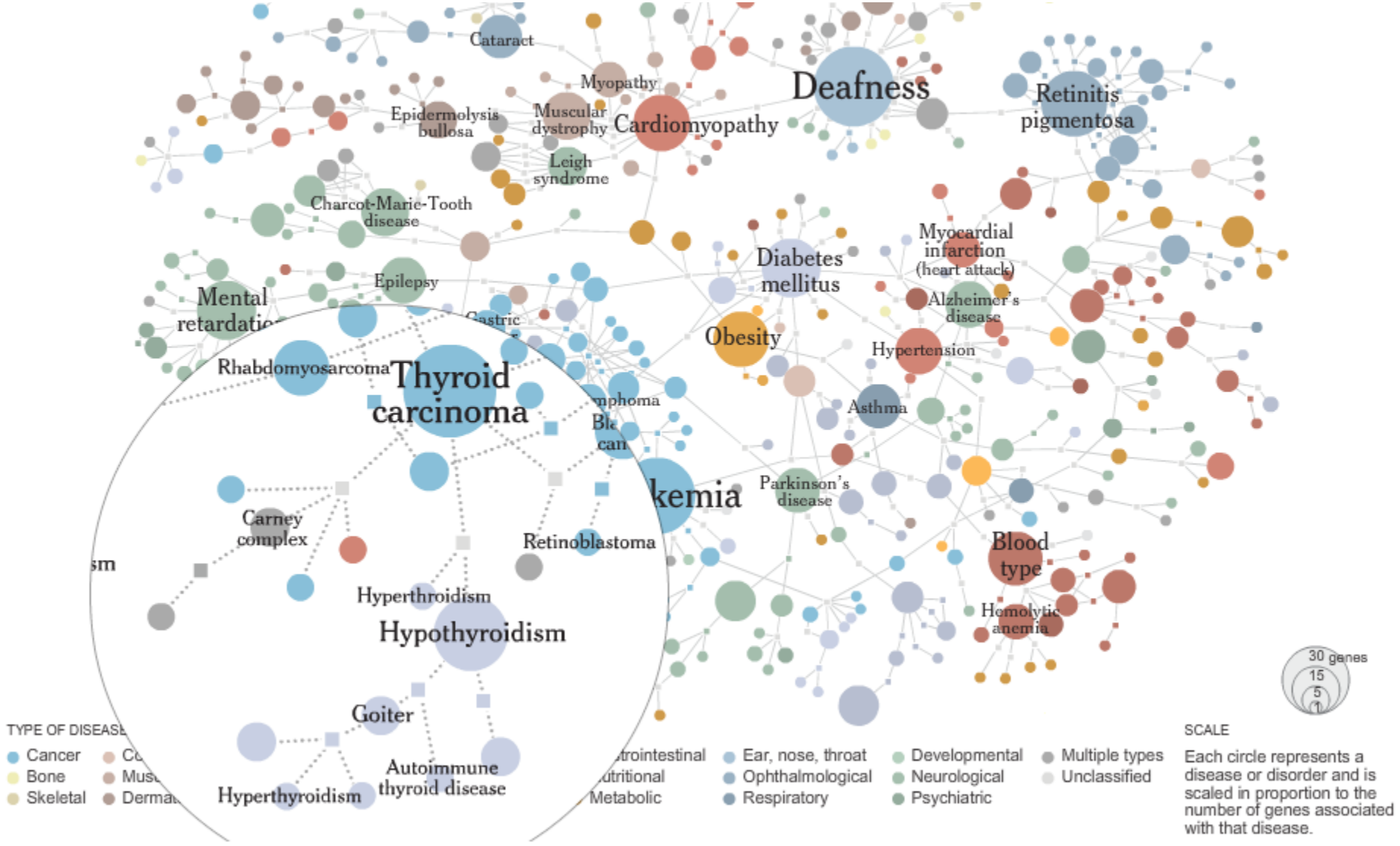


Image: biomol.de

Disease association network



Sources: Marc Vidal; Albert-Laszlo Barabasi; Michael Cusick; Proceedings of the National Academy of Sciences

The New York Times

Yeast protein interaction network

Bader and Hogue (2002) *Nature*

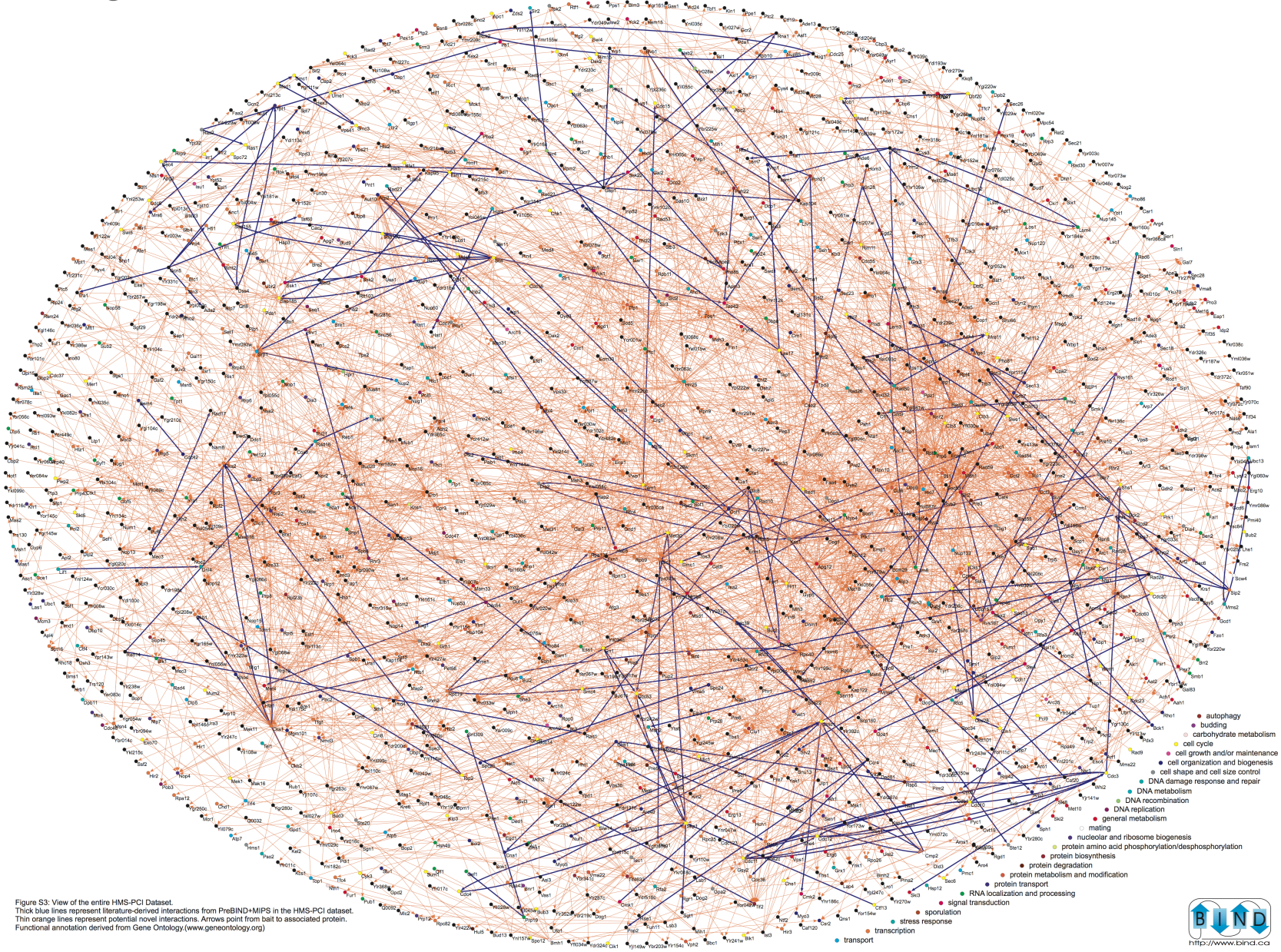


Figure S3: View of the entire HMS-PCI Dataset.
Thick blue lines represent literature-derived interactions from PreBIND+MIPS in the HMS-PCI dataset.
Thin orange lines represent potential novel interactions. Arrows point from bait to associated protein.
Functional annotation derived from Gene Ontology.(www.geneontology.org)

BIND
http://www.bind.ca

US Banks


citigroup
GOLDEN WEST FINANCIAL CORPORATION
FIRST TENNESSEE
BB&T
JPMorganChase
REGIONS
Comerica
charles SCHWAB
M&T Bank
MetLife
mbna
HSBC
HIBERNIA
usbank
WACHOVIA
Citizens Bank
Not your typical bank?
Mellon
RBC Centura
Bank of America
Higher Standards
WELLS FARGO
Huntington
AmSouth BANK
THE RELATIONSHIP PEOPLE
North Fork Bank
Capital One
LaSalle Bank
ABN AMRO
TD
Banknorth
National City
Sovereign Bank
Fifth Third Bank
SYNOVUS
Commerce Bank
Merrill Lynch
Compass Bank
STATE STREET
PNC
M&I
The BANK of NEW YORK
UNION BANK OF CALIFORNIA
SUNTRUST
The Family of Community Banks
DORAL FINANCIAL CORPORATION
Countrywide Financial
POPULAR
Washington Mutual
Commerce Bank
BANCWEST
A BNP PARIBAS company

Schedule

- Week 2, Jan. 10: **Problem Set #1 DUE**
- Week 3, Jan. 17: Problem Set #2 DUE
- Week 4, Jan. 22: Linear Algebra Lab DUE
- Week 5, Jan. 29: Quiz #1

Agenda

- 5:30 – 5:40 Logistics
- 5:40 – 6:00 Networks?
- 6:00 – 6:20 PageRank
- 6:20 - 6:30 Break
- 6:30 – 7:20 Assignment #2

Homework

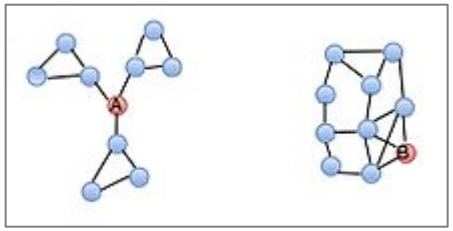
- Required Readings
  - Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999). [10k citations]

How would you build a search engine?

CULTURAL

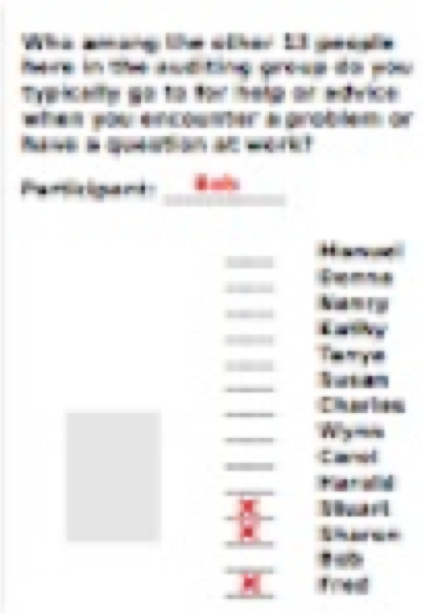
pollination, pollen, pollinators, flowers, flower, seed set, number flowers, pollinator, pollinated, nectar, stigmas, bees, anthers, outcrossing, floral, fruit set, pollen grains, breeding system, seed production, flowering, pollination ecology, selfing, ovules, number pollen, inbreeding depression, breeding systems, pollen transfer, flowers visited, flowering plants, stigma, pollen dispersal, inflorescence, inflorescences, nectar production, open flowers, flowers produced, inbreeding, bee, number seeds, individual flowers, floral display, flowers open, corolla, amount pollen, hermaphroditic, pollination biology, pollinator visitation, fruit seed, bumblebees, male function, flowers plants, floral traits, flowering season, visitation, experimental pollination, seeds produced, pollen production, flowers had, fruit production, pollinator behavior, pollinator visits, selfed, pollinations, pollen flow, outcrossing rate, fruit, visit flowers, pollen deposition, flower number, floral morphology, seeds, pollen limitation, outcrossing rates, female function, anthesis, seed, variation floral, plant reproductive, s c h, principles pollination, anther, variation pollen, grains, pollen deposited, flowers pollinated, flowering period, plants, reproductive biology, pollen donors, ipomopsis aggregata, pollen tubes, flowers plant, pollen load, flowers have, nectar pollen, plant populations, number ovules, floral biology, hummingbirds, flowering phenology, flower visitors, set fruit, pollen tube growth, visitors, handbook experimental, mating system, pollination success, pollen tube, visitation rates, visiting flowers, pollinator limitation, visited flowers, flower production, flower size, dioecy, hummingbird pollination, natural populations, journal botany, pollination systems, visited, flower fruit, seeds fruit, bumble bees, effects pollen, van der pijl, c h barrett, effect pollen, inflorescence size, pollinator activity, insect visitors, stamens, pollen removal, plant, b charlesworth, sex allocation, seed number, stigmatic, evolution dioecy, evolution floral, bagged, outcrossed, floral visitors, pollinator attraction, c e jones, fruits

STRUCTURAL

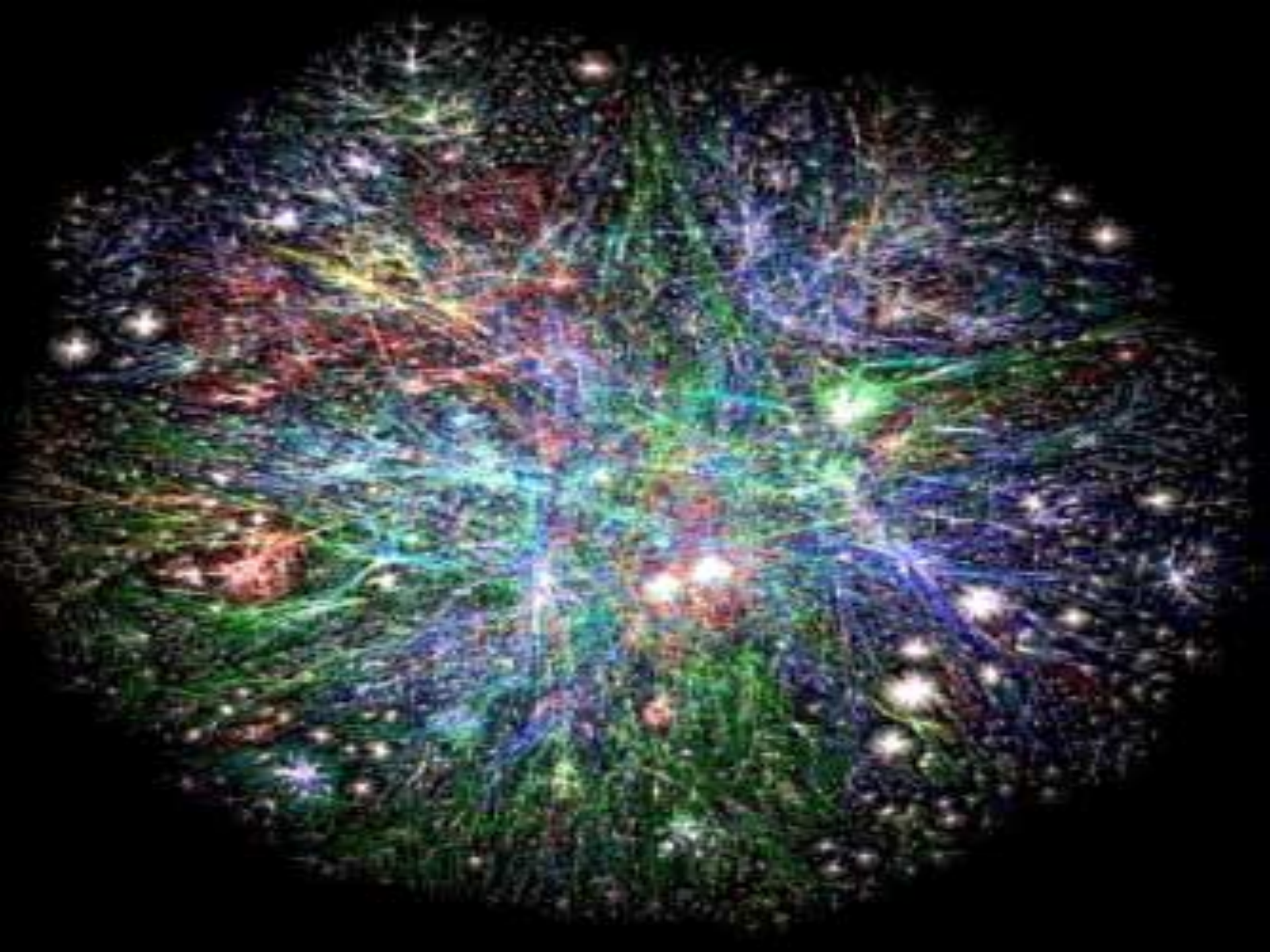


Adjacency Matrix

- Network data can be represented by adjacency matrices
  - Vertices on rows and columns
  - $A_{ij} = 1$ if i sends a tie to j



	Manuel	Donna	Nancy	Kathy	Tanya	Susan	Charles	Wynn	Carol	Harold	Stuart	Sharon	Bob	Fred
Manuel	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Donna	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Nancy	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Kathy	0	1	1	0	1	0	0	0	0	0	0	0	0	0
Tanya	0	1	1	1	0	0	0	0	0	0	0	0	0	0
Susan	0	1	0	1	1	0	0	0	0	0	0	0	0	0
Charles	1	0	1	0	0	0	0	0	0	0	1	0	0	0
Wynn	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Carol	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Harold	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Stuart	1	0	1	0	0	0	1	0	0	0	0	0	0	0
Sharon	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Bob	0	0	0	0	0	0	0	0	0	0	1	1	0	1
Fred	0	0	0	0	0	0	0	0	0	0	1	0	0	0



Which is the most 'influential' node?

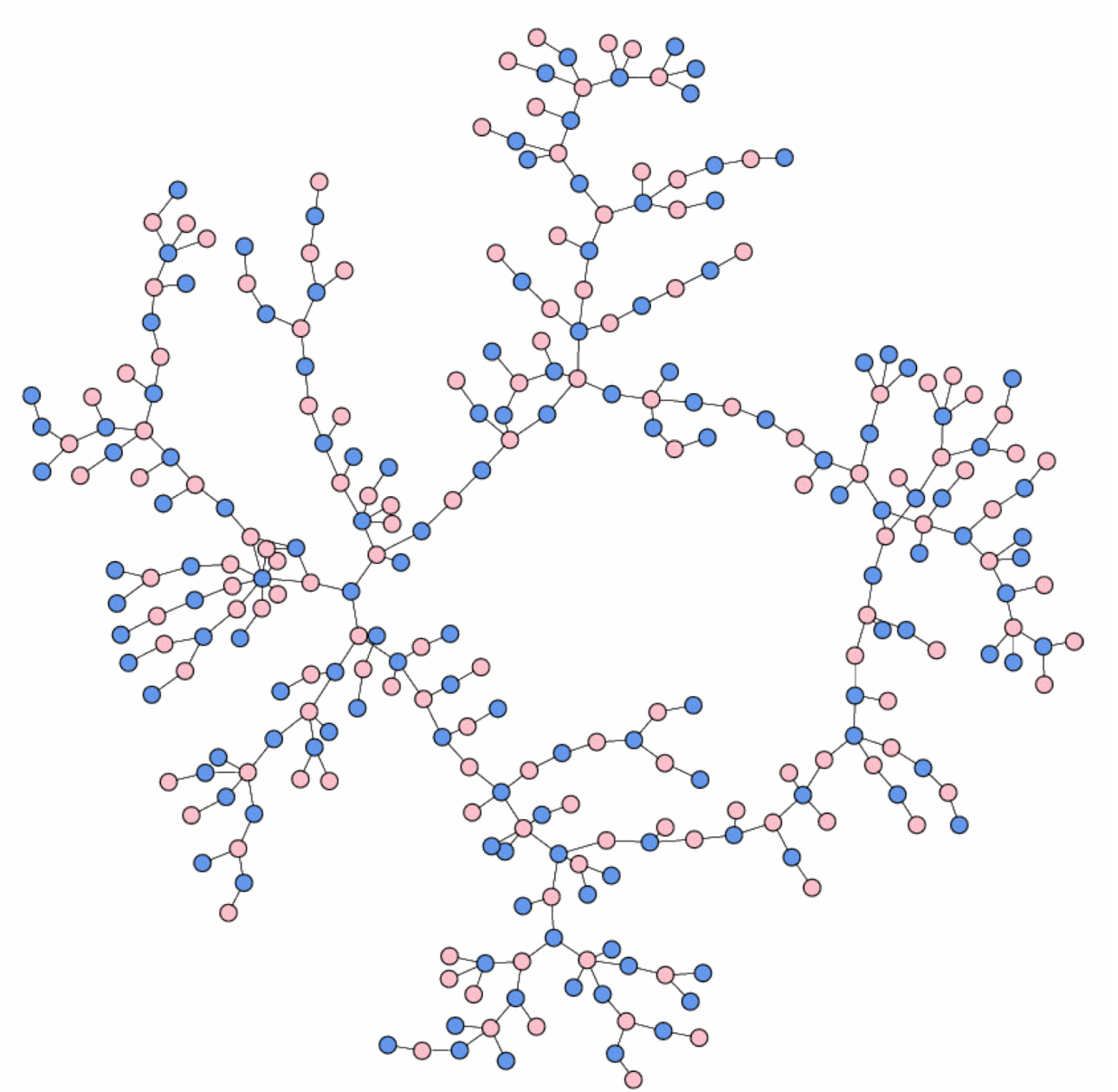


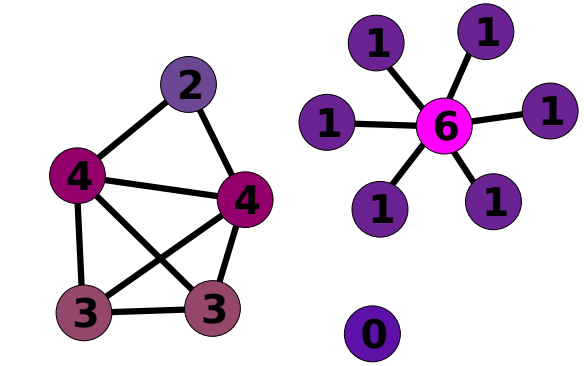
Image Courtesy of Mark Newman

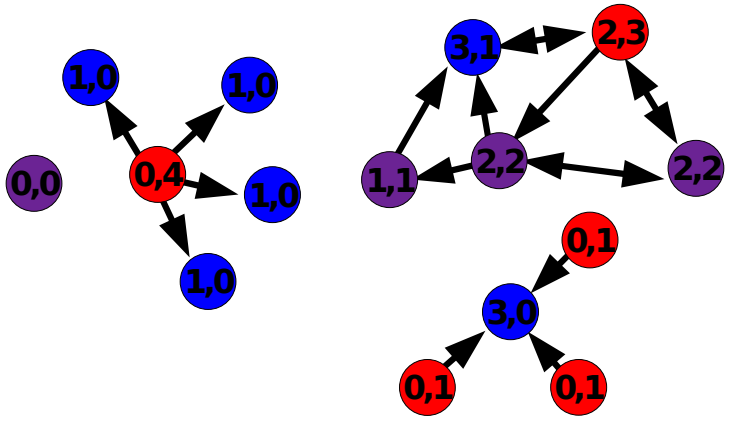
Degree

- **Degree:** number of direct ties
  - Measure of overall activity or extent of involvement
  - High degree positions are influential, but also may be subject to a great deal of influence from others

- Formulas:
  - Degree (undirected): $d(i, Y) = \sum_{j=1}^{N} Y_{ij}$
  - Indegree: $d_i(i, Y) = \sum_{j=1}^{N} Y_{ji}$
  - Outdegree: $d_o(i, Y) = \sum_{j=1}^{N} Y_{ij}$

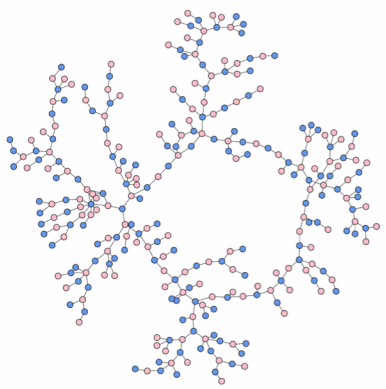
  where Y is the adjacency matrix.



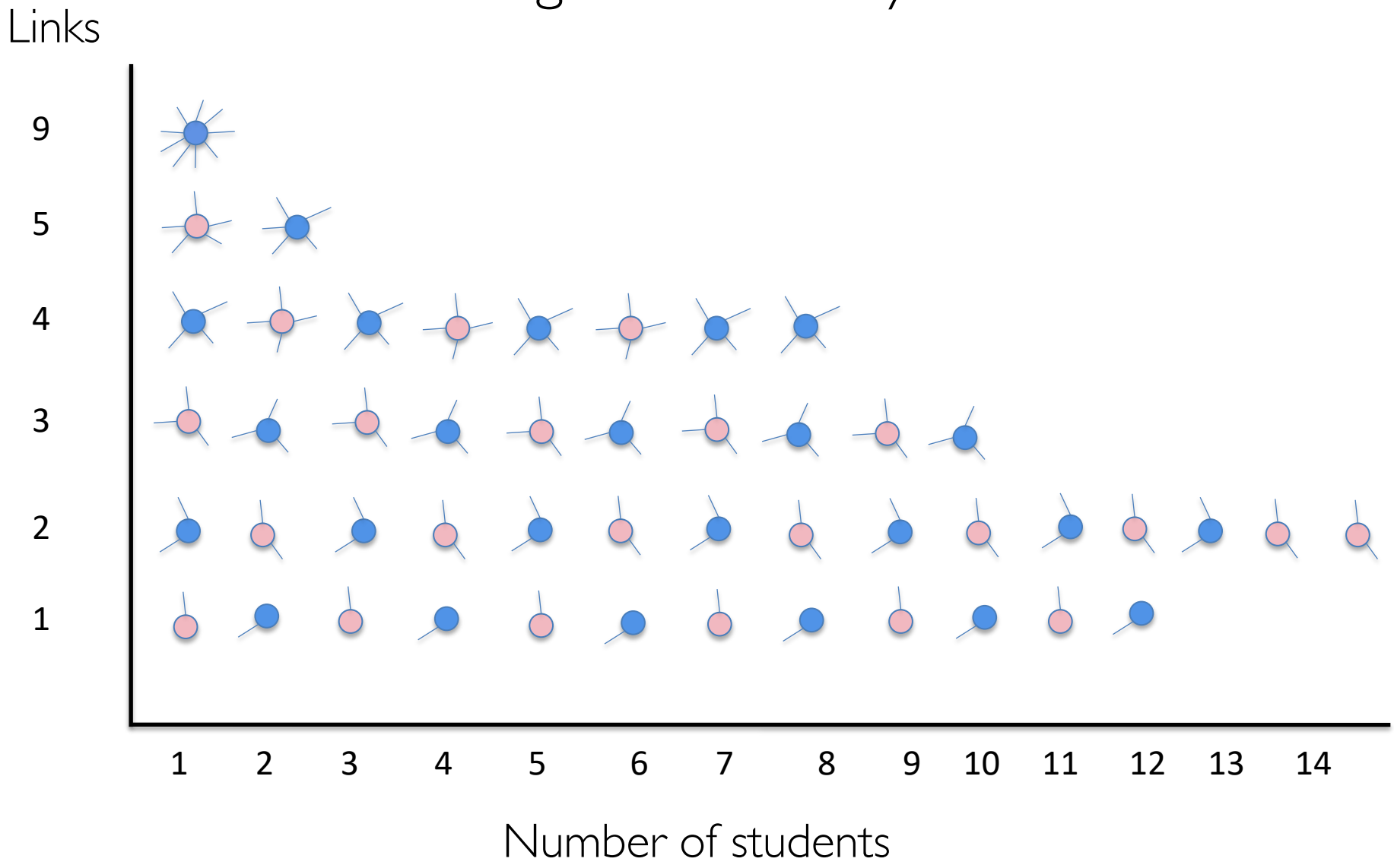


Degree Centrality

Eigenvector Centrality



Degree Centrality



Eigenvector Centrality

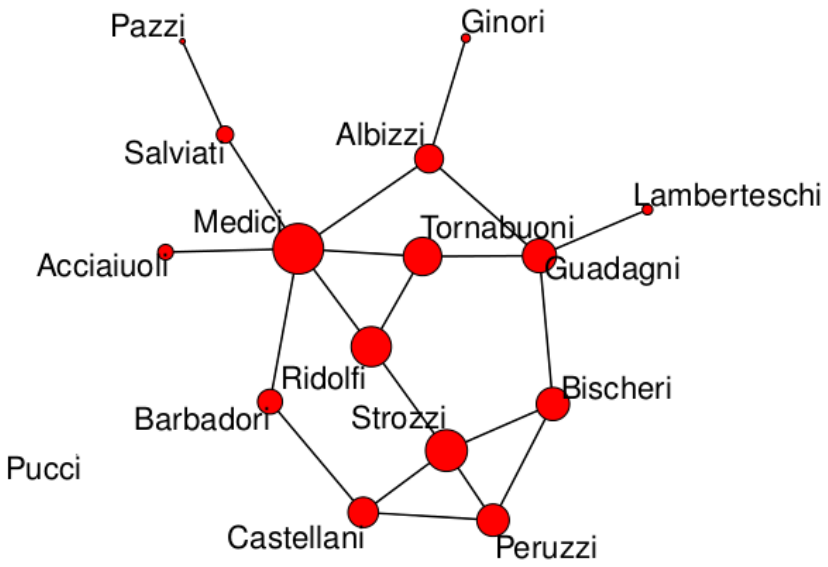
- **Eigenvector Centrality:** values of the first eigenvector of the graph adjacency matrix
  - Can be interpreted as arising from a reciprocal process in which the centrality of each actor is proportional to the sum of the centralities of those actors to whom he or she is connected
  - Measure of "coreness" in a network

- Formula:

$$e(i, Y) = \frac{1}{\lambda} \sum_j Y_{ij} e_j$$

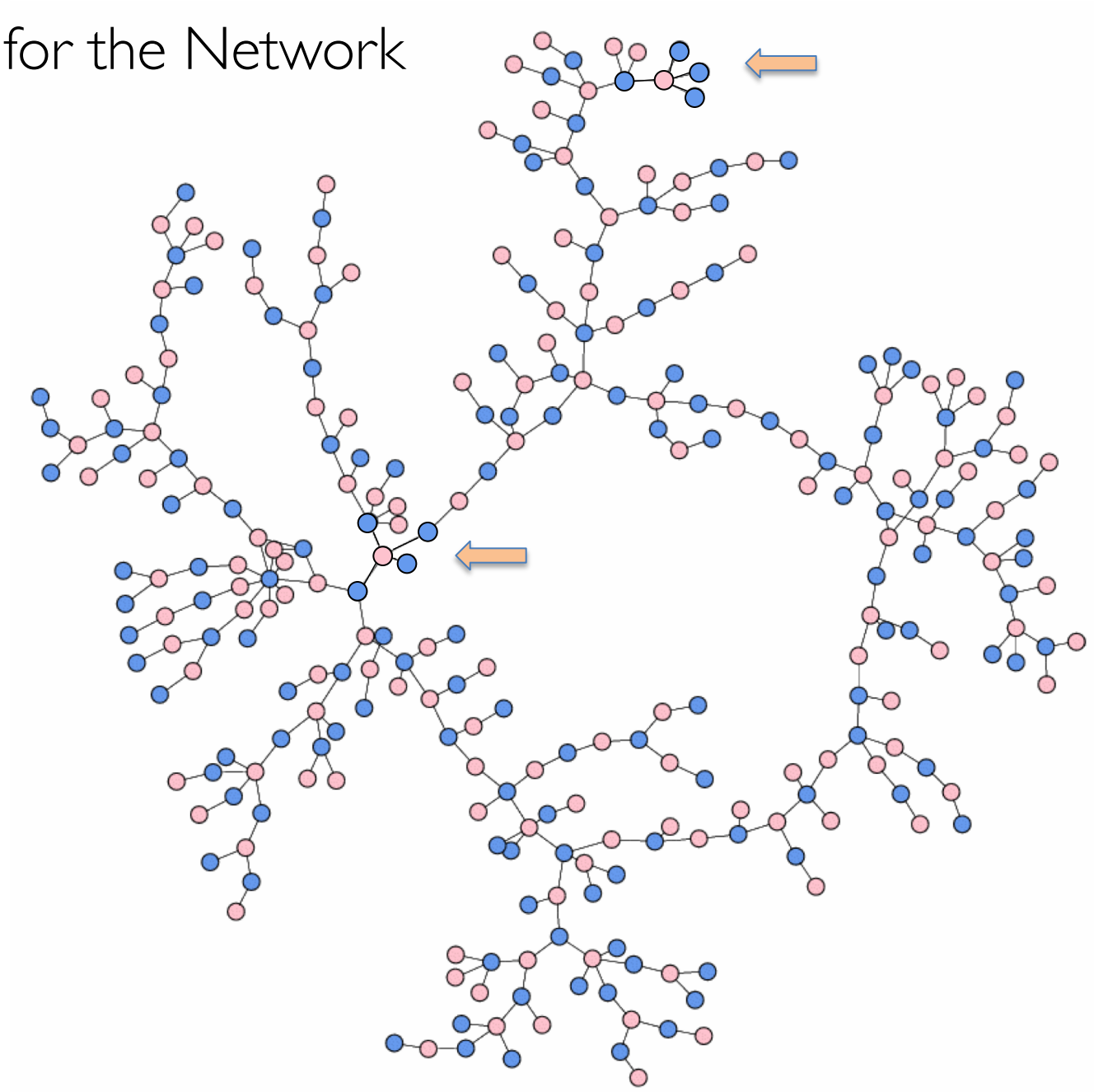
If we write this in vector notation, see it is the eigenvector equation:

$$\mathbf{Yx} = \lambda \mathbf{x}$$

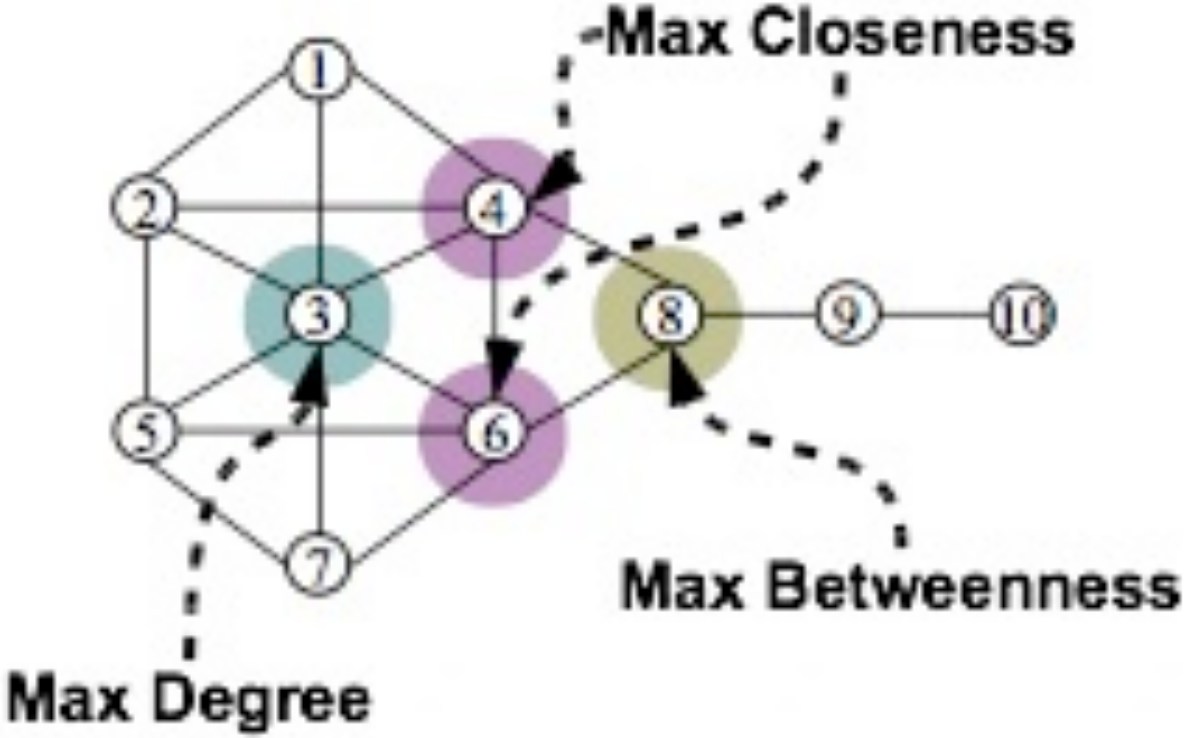


Google's PageRank is a variant of the Eigenvector centrality measure!

Accounting for the Network

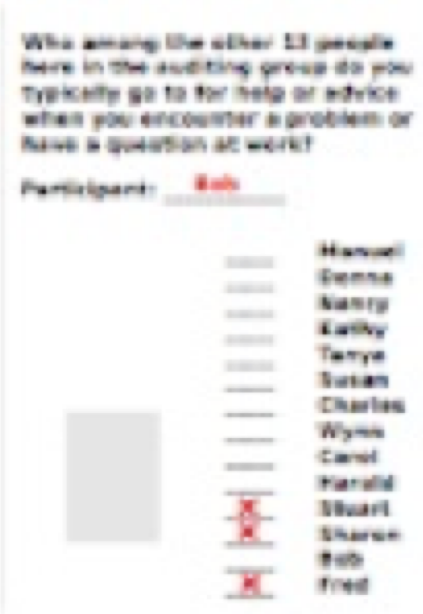


Comparing Major Centrality Measures



Adjacency Matrix

- Network data can be represented by adjacency matrices
  - Vertices on rows and columns
  - $A_{ij} = 1$ if i sends a tie to j



	Manuel	Donna	Nancy	Kathy	Tanya	Susan	Charles	Wynn	Carol	Harold	Stuart	Sharon	Bob	Fred
Manuel	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Donna	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Nancy	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Kathy	0	1	1	0	1	0	0	0	0	0	0	0	0	0
Tanya	0	1	1	1	0	0	0	0	0	0	0	0	0	0
Susan	0	1	0	1	1	0	0	0	0	0	0	0	0	0
Charles	1	0	1	0	0	0	0	0	0	0	1	0	0	0
Wynn	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Carol	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Harold	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Stuart	1	0	1	0	0	0	1	0	0	0	0	0	0	0
Sharon	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Bob	0	0	0	0	0	0	0	0	0	0	1	1	0	1
Fred	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Matrix Operations

Scalar multiplication:

$$\lambda \mathbf{A} = \lambda \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} = \begin{pmatrix} \lambda A_{11} & \lambda A_{12} & \lambda A_{13} \\ \lambda A_{21} & \lambda A_{22} & \lambda A_{23} \\ \lambda A_{31} & \lambda A_{32} & \lambda A_{33} \end{pmatrix}$$

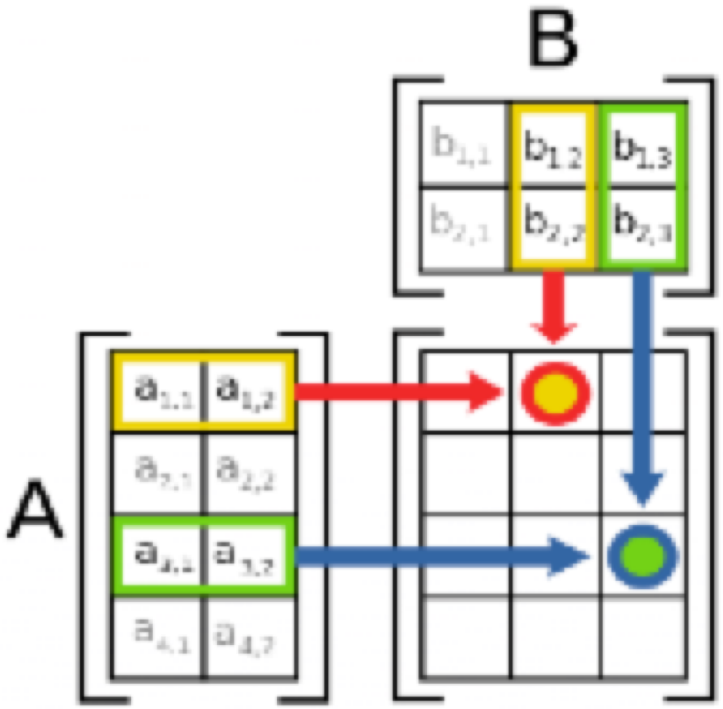
Matrix product:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix}$$

$$\mathbf{AB} = \begin{pmatrix} (\mathbf{AB})_{11} & (\mathbf{AB})_{12} & (\mathbf{AB})_{13} \\ (\mathbf{AB})_{21} & (\mathbf{AB})_{22} & (\mathbf{AB})_{23} \\ (\mathbf{AB})_{31} & (\mathbf{AB})_{32} & (\mathbf{AB})_{33} \end{pmatrix}$$

where $(\mathbf{AB})_{ij} = \sum_{k=1}^{3} A_{ik} B_{kj}$.

Matrix Multiplication



4 x 2 matrix multiplied by 2 x 3 becomes a 4 x 3 matrix

PageRank Simplified[1][2]

Suppose that page P_j has l_j links. If one of those links is to page P_i, then P_j will pass on $1/l_j$ of its importance to P_i.

The importance ranking of P_i is the sum of all the contributions made by pages linking to it:

$$I(P_i) = \sum_{P_j \in B_i} \frac{I(P_j)}{l_j}$$

where B_i is set of pages linking to P_i.

[1] http://www.ams.org/samplings/feature-column/fcarc-pagerank

[2] http://www.rose-hulman.edu/ bryan/googleFinalVersionFixed.pdf

PageRank Simplified[1][2]

Let $\mathbf{H}$ be the hyperlink matrix in which the entry in the ith row and jth column is

$$H_{ij} = \begin{cases} 1/l_j & \text{if } P_j \in B_i \\ 0 & \text{otherwise} \end{cases}$$

Let I be the vector of PageRanks.

PageRank may be expressed as:

$$I = \mathbf{H}I$$

$\mathbf{I}$ is an **eigenvector** of the matrix $\mathbf{H}$ with **eigenvalue** 1.

[1] http://www.ams.org/samplings/feature-column/fcarc-pagerank
[2] http://www.rose-hulman.edu/ bryan/googleFinalVersionFixed.pdf

Eigenfactor algorithm

$$P = \alpha H + (1 - \alpha)\, a.e^{T}$$

Matrix representing the random walk over citations

Probability of not teleporting

Cross-citation Matrix dictating the structure of the citation network

Probability of teleporting to completely new journal weighted by the number of articles in that journal

$$EF = 100 \frac{H\pi}{\sum_i [H\pi]_i}$$

Leading eigenvector of the random walk matrix P.

Normalization

West, JD et al. (2010) *College of Research Libraries*

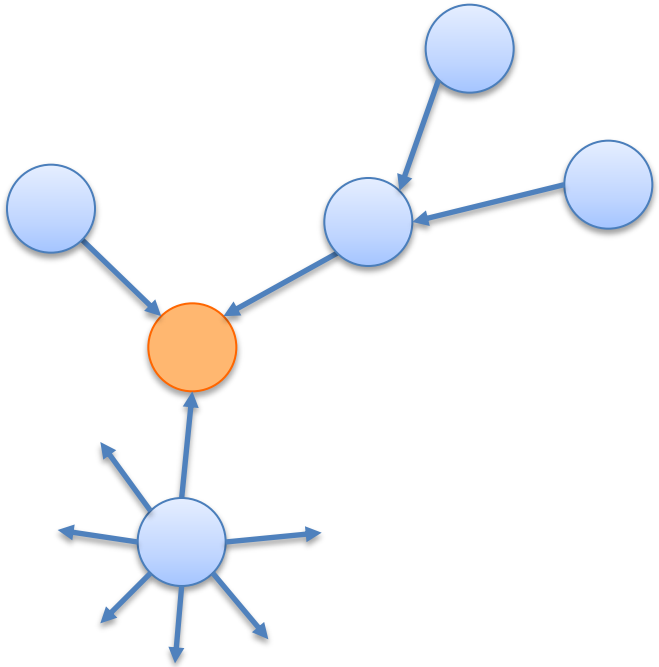
Taking into account the source of a link…

1. $r(P_i) = ?$

2. $r(P_i) = \sum_{P_j \in B_{P_i}} r(P_j)$

3. $r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{P_j}$

4. $r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{P_j}$

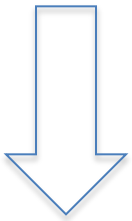


$r(P_j) = ?$

$r(P_j) = \frac{1}{n}$

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{P_j}$$

Convert to Matrix Notation



$$\pi^{(k+1)} = H\pi^{(k)}$$

Problems with the iterative process

1. Will the process continue indefinitely or will it converge?
2. Does convergence depend on the starting vector?
3. If it does converge, how many iterations can we expect until it converges?

$$\pi^{(k+1)} = H\pi^{(k)}$$

stochastic, irreducible and aperiodic

Problems with REALLY BIG networks

Homework

- Required Readings
  - Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999). [10k citations]