Project assignment
Data structure and algorithm

Patrick De Causemaecker
San Tu Pham

Contact

- Teaching assistant: San Tu Pham
  - Email: san.pham@kuleuven-kulak.com
  - Office: C744
  - Drop by or email if you have any questions!

Project Assignment

- Two papers are provided:
  - Morone, Flaviano, and Hernán A. Makse. "Influence maximization in complex networks through optimal percolation." *Nature* (2015) (along with its Supplementary Information)
  - Morone, Flaviano, et al. "Collective Influence Algorithm to find influencers via optimal percolation in massively large social media." *arXiv preprint arXiv*:1603.08273 (2016).
- Requirements:
  - Read and comprehend the algorithm from the first paper
  - Understand the implementation of the algorithm in the second one
  - Implement the algorithm based on the second paper
  - Conduct experiment on big dataset
  - Document and present your work

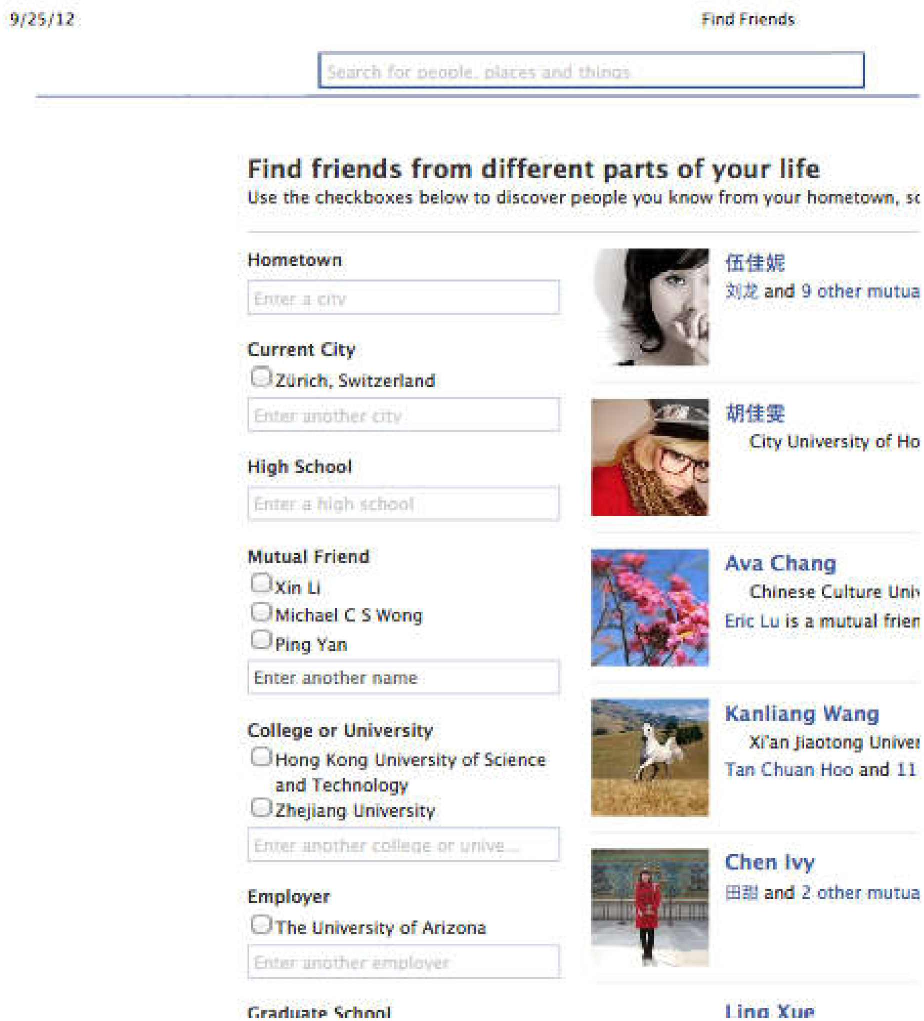
KU LEUVEN

Network science

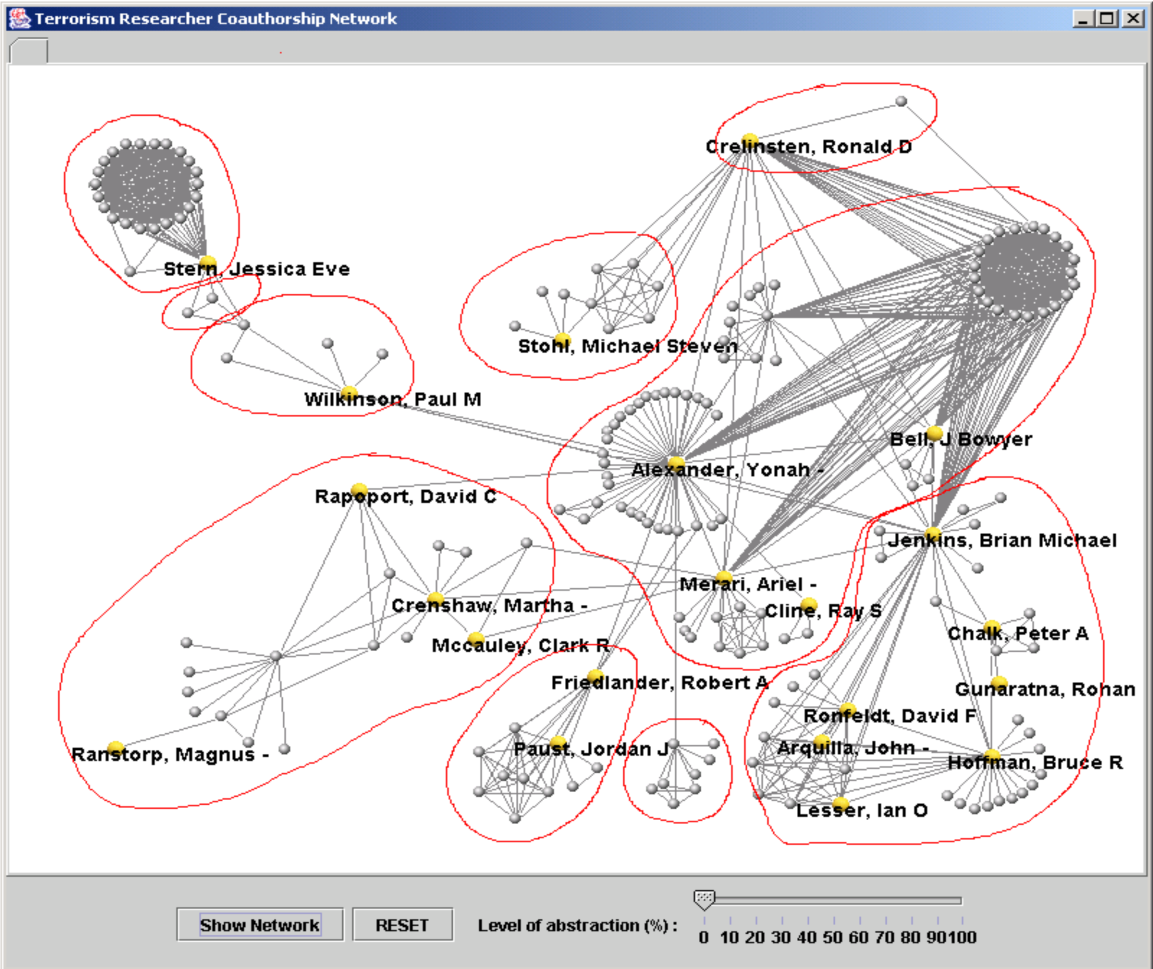
- “An academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks..”
- Applications:
  - Influencer marketing
  - Social network analysis
  - Finance: Insurance fraud detections
  - E-commerce: recommendation systems， e.g., Amazon
  - Internet Search Engine: Google PageRank
  - Law Enforcement: Crime link predictions
  - …

KU LEUVEN

Network-based Business Applications



Security: Fighting Terrorism



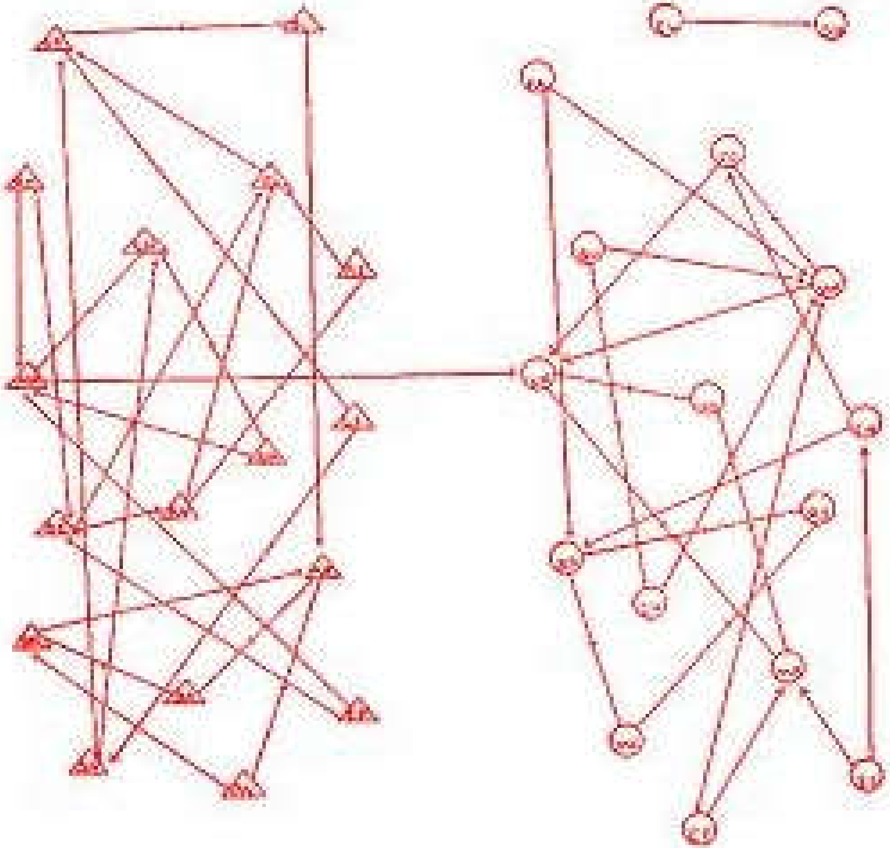
KU LEUVEN

Social Network Analysis

EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

New York Times
April 3, 1933



Epidemics: From forecasting to halting deadly viruses

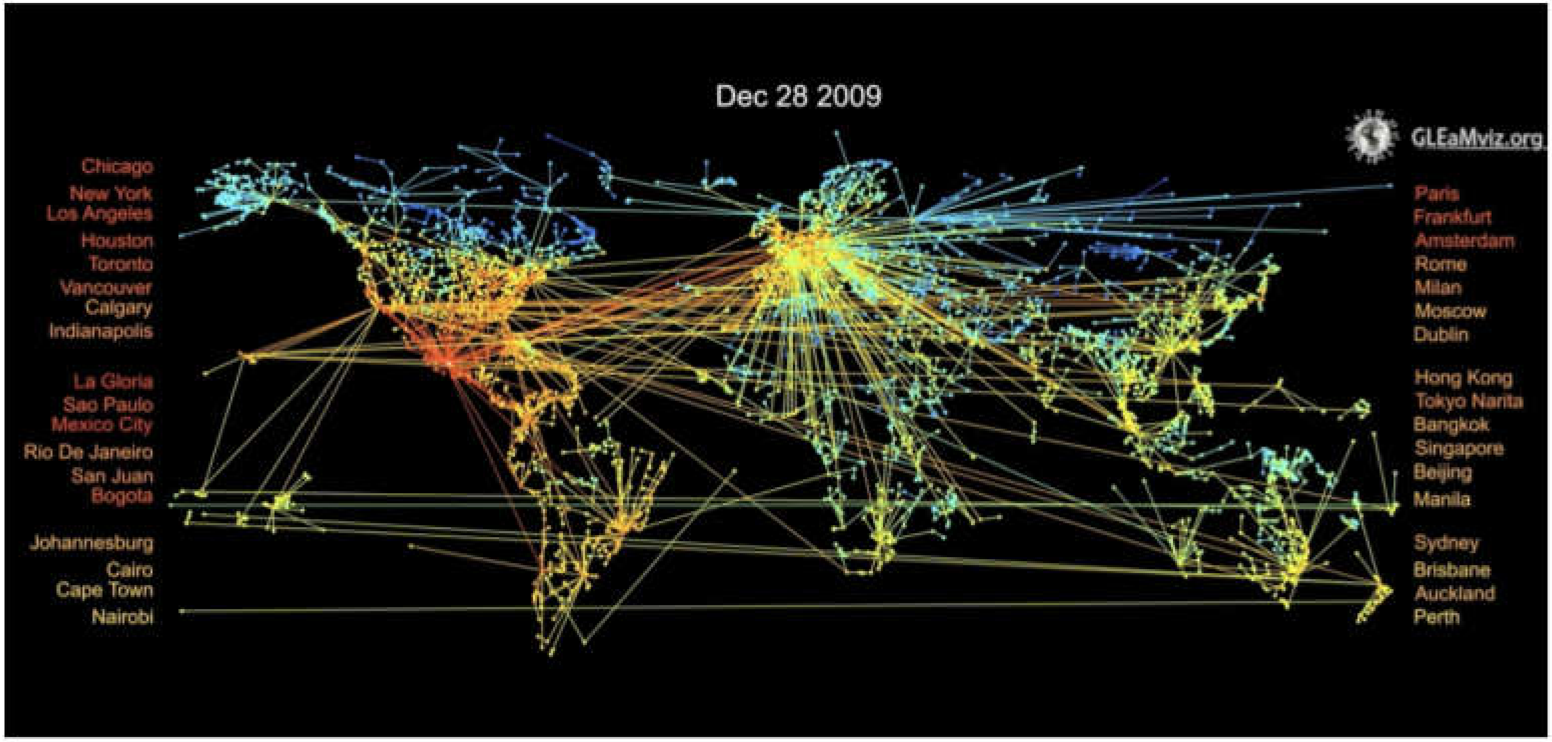


Image 1.9
Predicting the H1N1 epidemic.

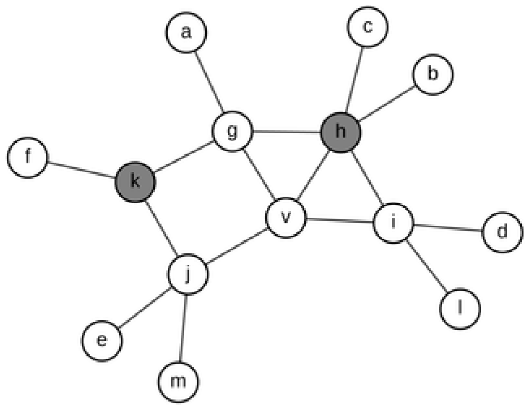
The optimal influence problem

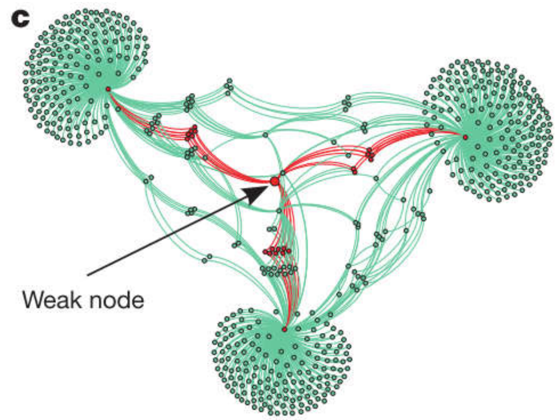
- An important problem in network science
- “The most influential nodes are the ones forming the minimal set that guarantees a global connection of the network” - this minimal set is called “optimal influencers” of the network
- The optimal influence problem: Localizing the minimal set of influencers
    - First introduced in the context of viral marketing
    - NP hard
- The authors map the problem onto optimal percolation in random networks

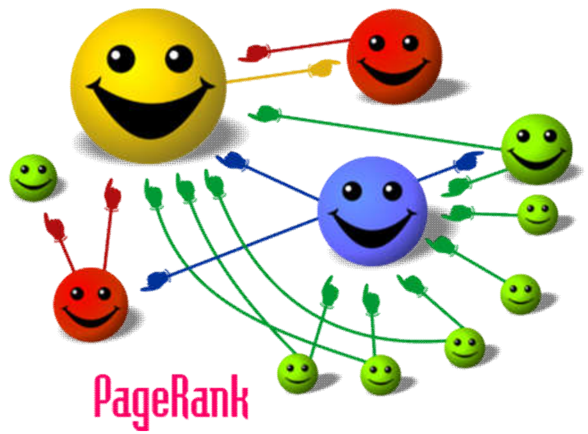
KU LEUVEN

Heuristics to identify influential spreaders

- High degree: nodes are ranked by degree, and sequentially removed starting from the node of highest degree



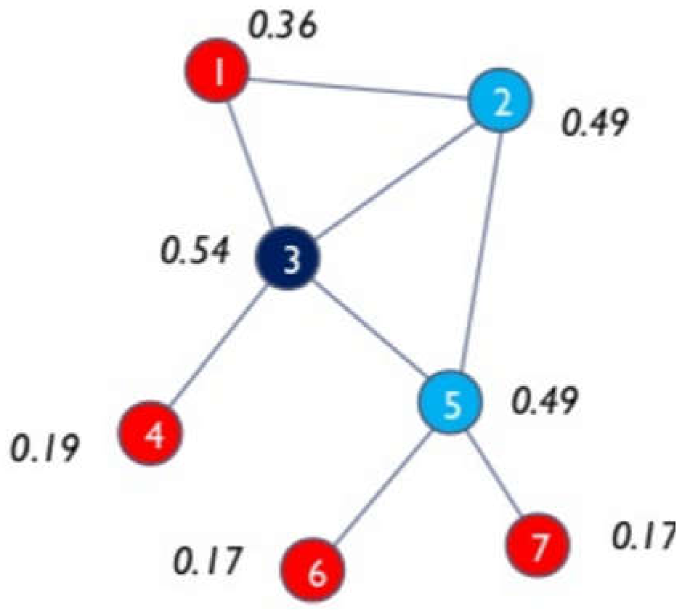




- Google PageRank: sites that get linked more are considered reputable, and, linking to other websites, they pass that reputation along

Heuristics to identify influential spreaders (cont.)

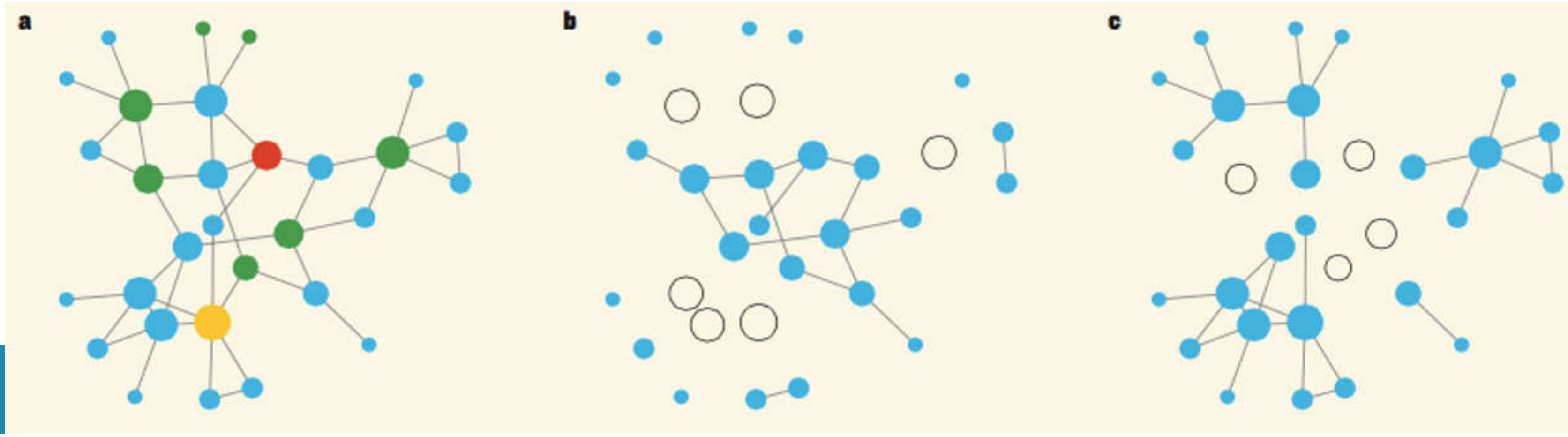
- Eigen vector Centrality (EC):
  - A node's eigenvector centrality is proportional to the sum of the eigenvector centralities of all nodes directly connected to it *.
  - EC is the eigenvector corresponding to the largest eigenvalue of the adjacency matrix. Node rank is the corresponding entry of the eigenvector.



*Giorgos Cheliotis - NUS

Influence maximization in complex networks through optimal percolation

- Optimal influence problem:
  - Localizing the minimal set of influencers.
  - (informal): find the minimal set of nodes, which, if removed, would break down the network into many disconnected pieces.
- The natural measure of influence: the size of the largest (giant) connected component as the influencers are removed from the network.



Influence maximization in complex networks through optimal percolation

- The authors map the problem onto
  - Optimal percolation in random networks
  - Or: the computation of the minimal set of nodes that minimizes the largest eigenvalue of the Non-Backtracking matrix of the network.

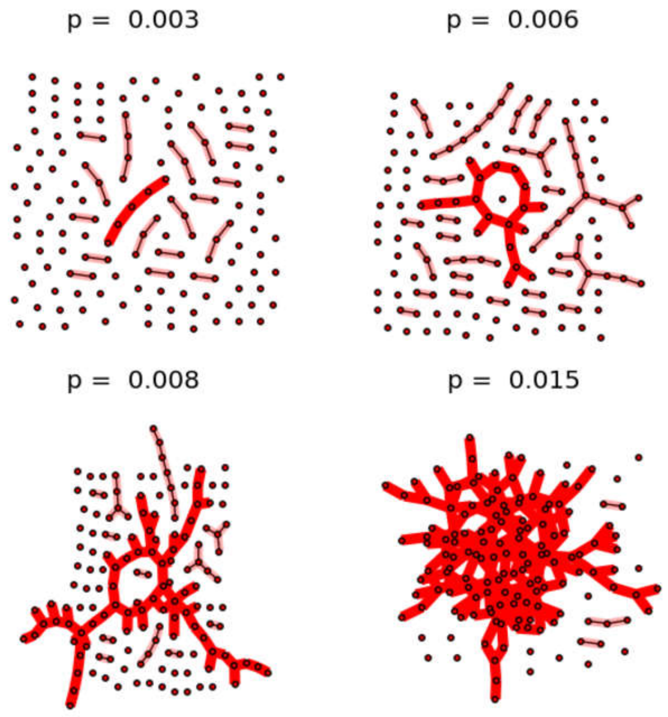
KU LEUVEN

Some definitions

- Giant component: if a network has a "giant component", that means almost every node is reachable from almost every other

Size of component	1	2	3	4	5	6	7	14	27488
# of component	2712	549	129	51	16	15	3	1	1

- Erdős–Rényi model: G(n, p) is a random graph with n vertices where each possible edge has probability p of existing
- Giant component in Erdős–Rényi model



https://networkx.github.io/documentation/networkx-1.10/examples/drawing/giant_component.html

Some definitions

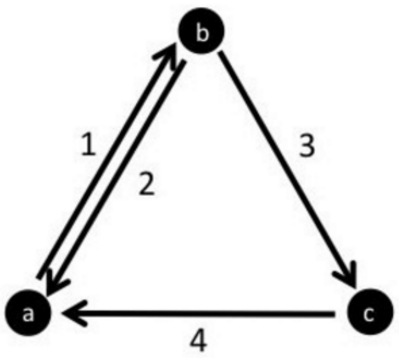
- **Percolation theory:** describes the behavior of connected clusters in a random graph.
- Percolation threshold: is the formation of long-range connectivity in random systems. Below the threshold a giant connected component does not exist; while above it, there exists a giant component of the order of system size

KU LEUVEN

Some definitions - Non-backtracking matrix

- Non-backtracking matrix (Kiichiro Hashimoto, 1989): a representation of the link structure of a network that is an alternative to the usual adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if link } j \to i \text{ exists} \\ 0 & \text{otherwise} \end{cases} \qquad B_{k\to\ell, i\to j} = \begin{cases} 1 & \text{if } j = k \text{ and } i \neq \ell \\ 0 & \text{otherwise} \end{cases}$$

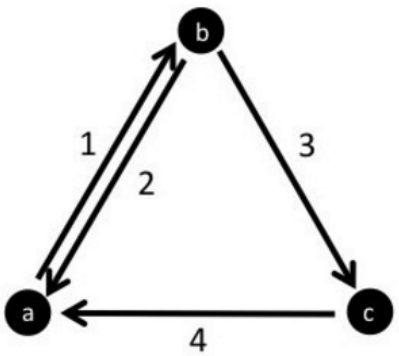


$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

https://www.quora.com/What-is-an-intuitive-explanation-of-the-Hashimoto-non-backtracking-matrix-and-its-utility-in-network-analysis

Non-backtracking matrix (cont.)

- NB matrix can be used to identify non-backtracking walks on a network (walks do not proceed from a node *i* to a node *j* only to immediately return to node *i*).
- Powers of the Hashimoto matrix generate non-backtracking walks through the network



$$B^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad B^3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

SUCCESS!