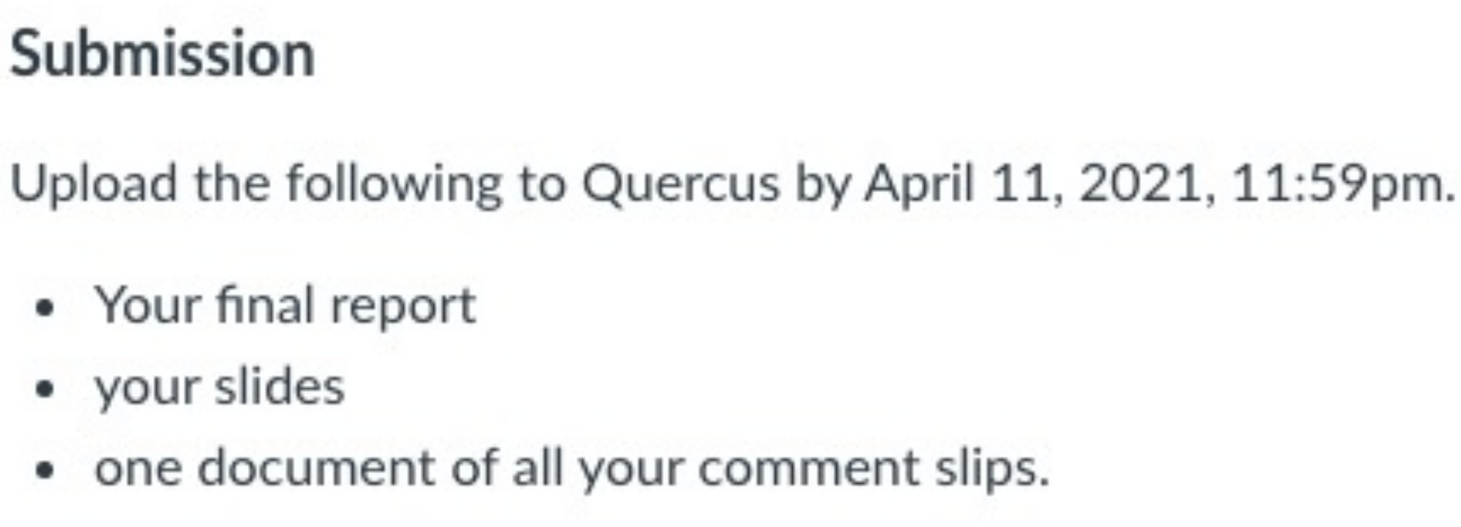
# ECE1724H S2: Empirical Software Engineering

## Social Network Analysis

The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

# about final report

Presentation 10min+3min Q&A
Comment on other’s work
merge all the comments in one file



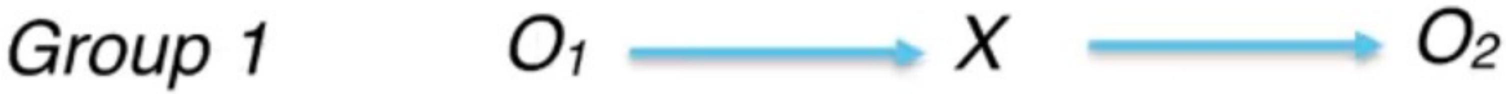
## Submission

Upload the following to Quercus by April 11, 2021, 11:59pm.

- Your final report
- your slides
- one document of all your comment slips.

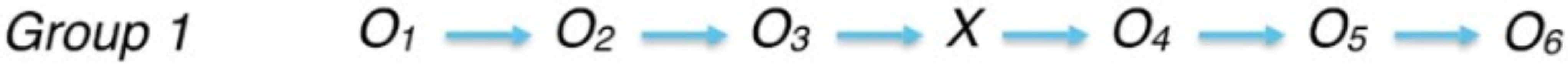
# Quasi-experiments Design

- One-group Pretest-Posttest design

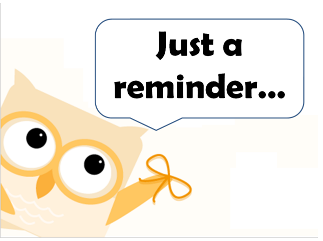


Group 1 $O_1 \rightarrow X \rightarrow O_2$

- Interrupted time series design



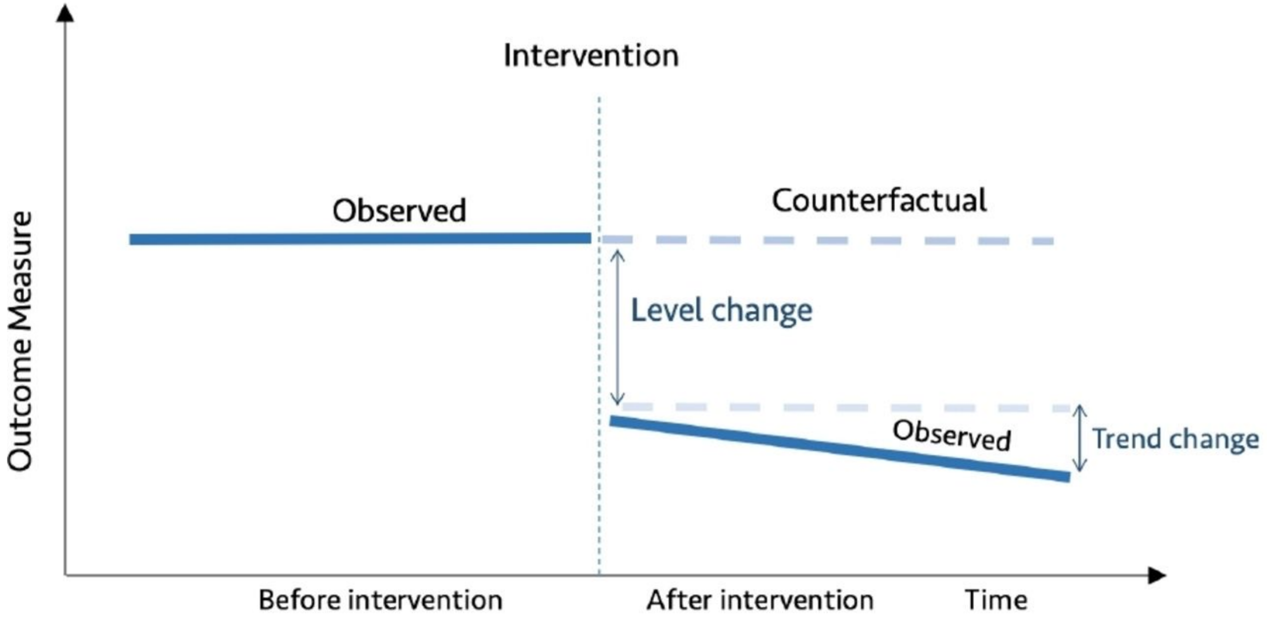
Group 1 $O_1 \rightarrow O_2 \rightarrow O_3 \rightarrow X \rightarrow O_4 \rightarrow O_5 \rightarrow O_6$



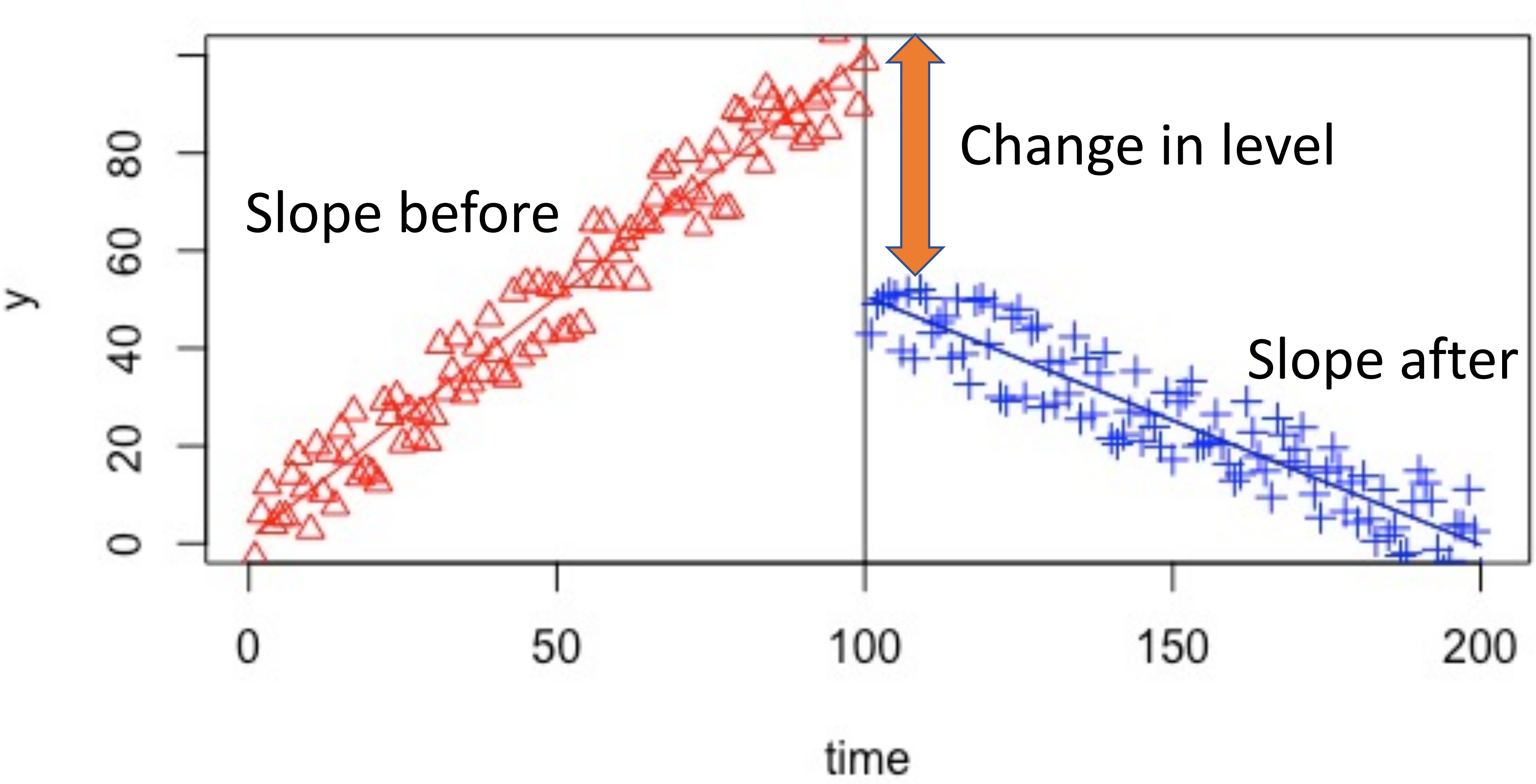
# Interrupted Time Series

Causal hypothesis –

the observations after treatment will have a different slope or level from those before treatment.



# Types of Effects - form



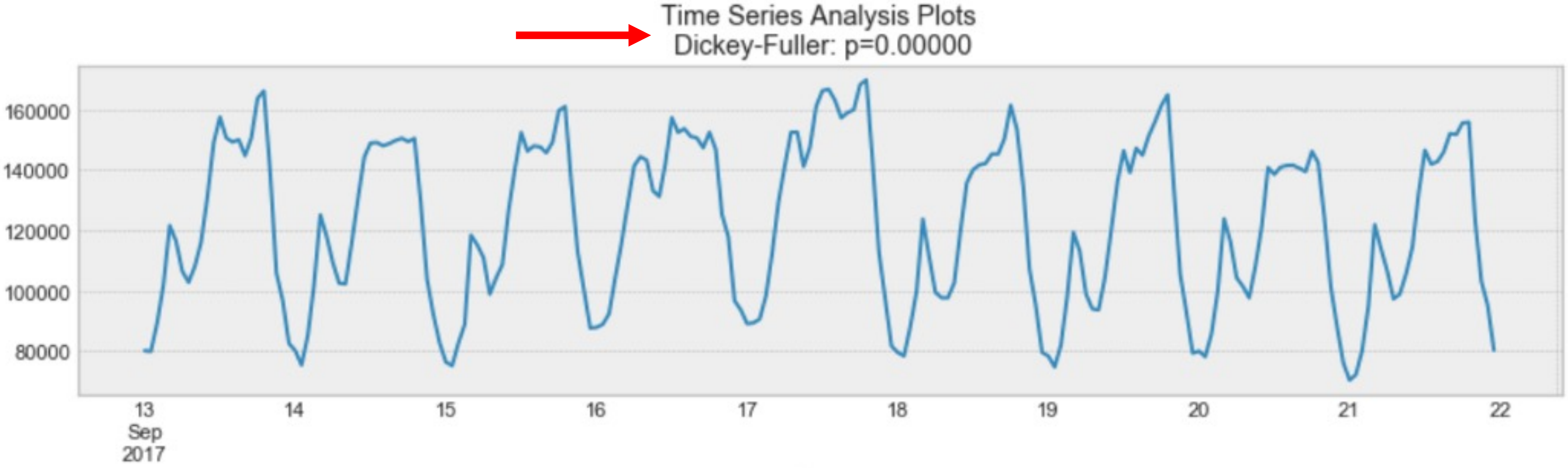
Slope before

Change in level

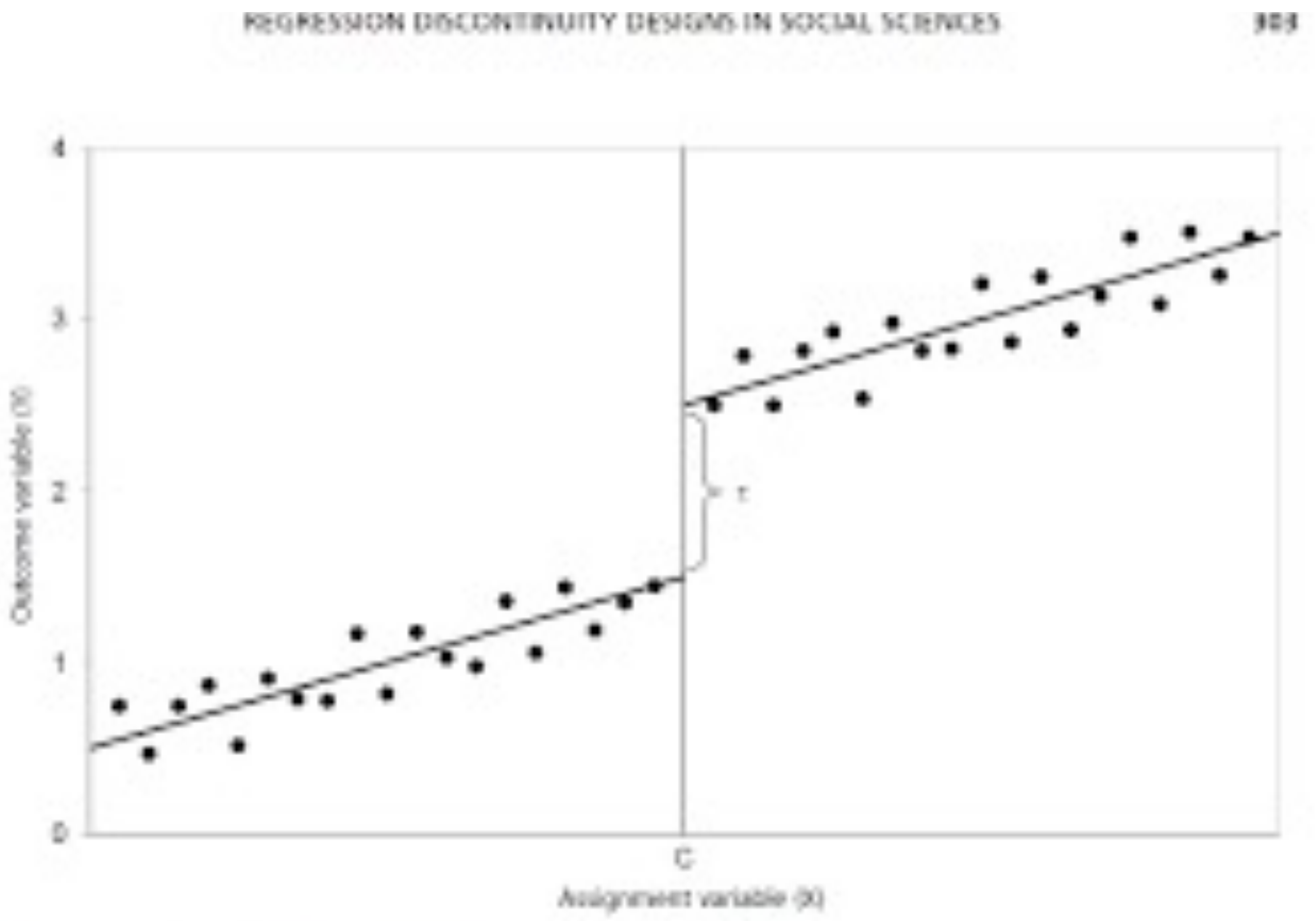
Slope after

# Stationarity

- The data has **constant mean and variance**



Time Series Analysis Plots
Dickey-Fuller: p=0.00000



REGRESSION DISCONTINUITY DESIGNS IN SOCIAL SCIENCES 303

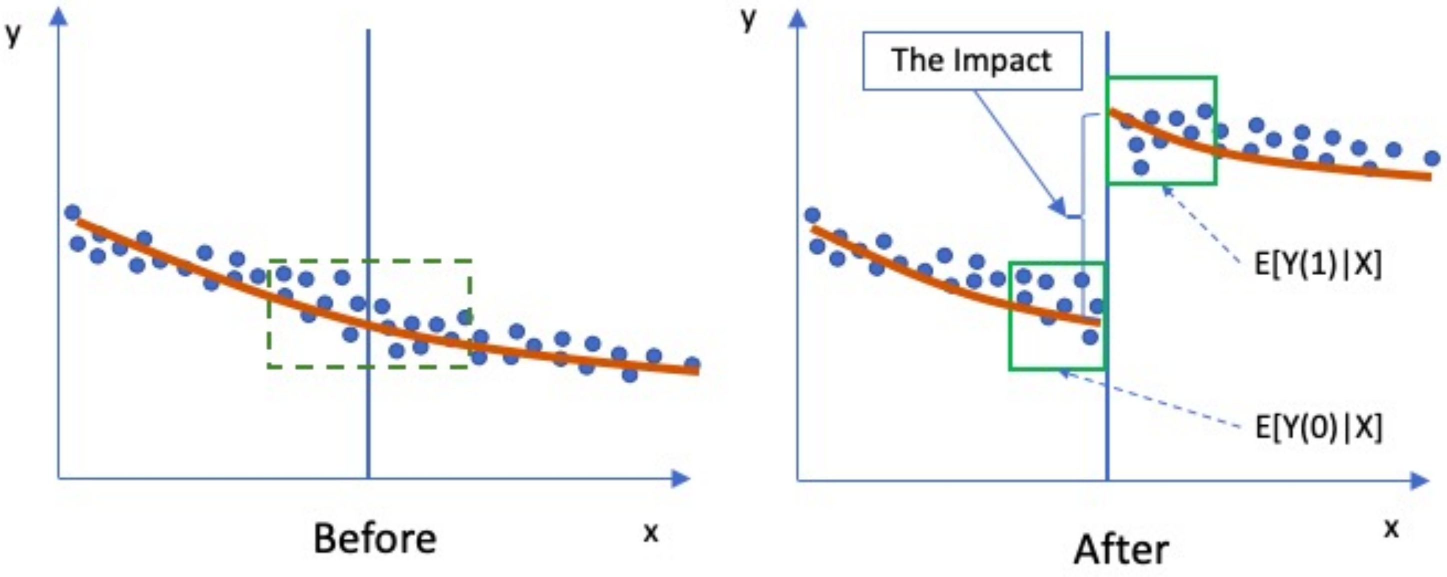
# RDD -- Regression Discontinuity Design

# RDD – Basic structure

Assigns units to conditions **on the basis of a cutoff score on an assignment variable**

| | | | |
|---|---|---|---|
| $O_A$ | $C$ | $X$ | $O_2$ |
| $O_A$ | $C$ | | $O_2$ |

- $O_A$ -- pre-assignment measure of the assignment variable
- C – cutoff score



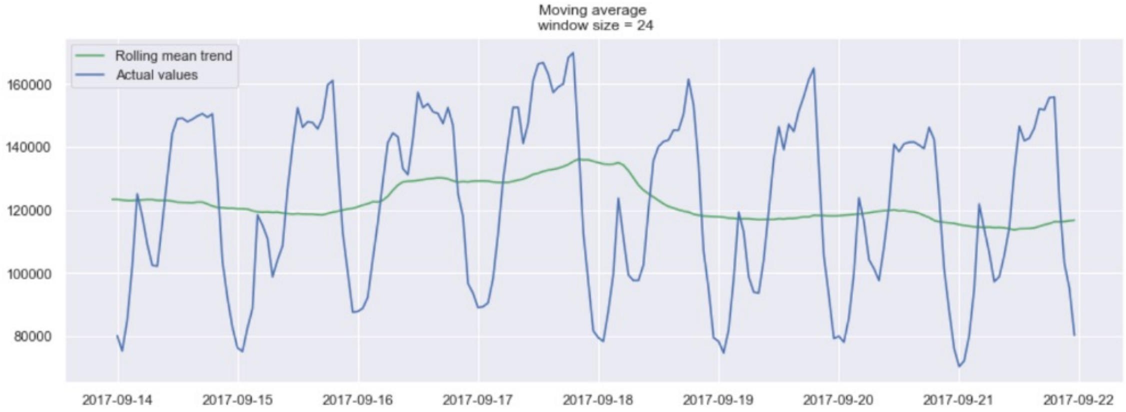
# Autoregression (AR)

- A statistical model is said to be autoregressive if it predicts future values based on previous values.
- ***AR(parameter)*** -- parameter is the number of independent variables or the count of past values considered for forecasting.

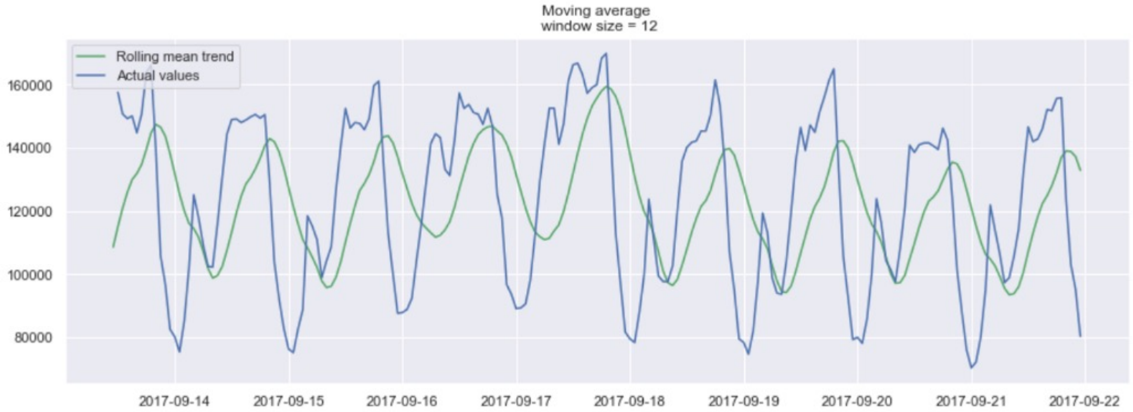
  E.G., AR (n), n = [0,1,2,...]

# Moving Average (MA)

- Statement: the next observation is the mean of all past observations.



24h window



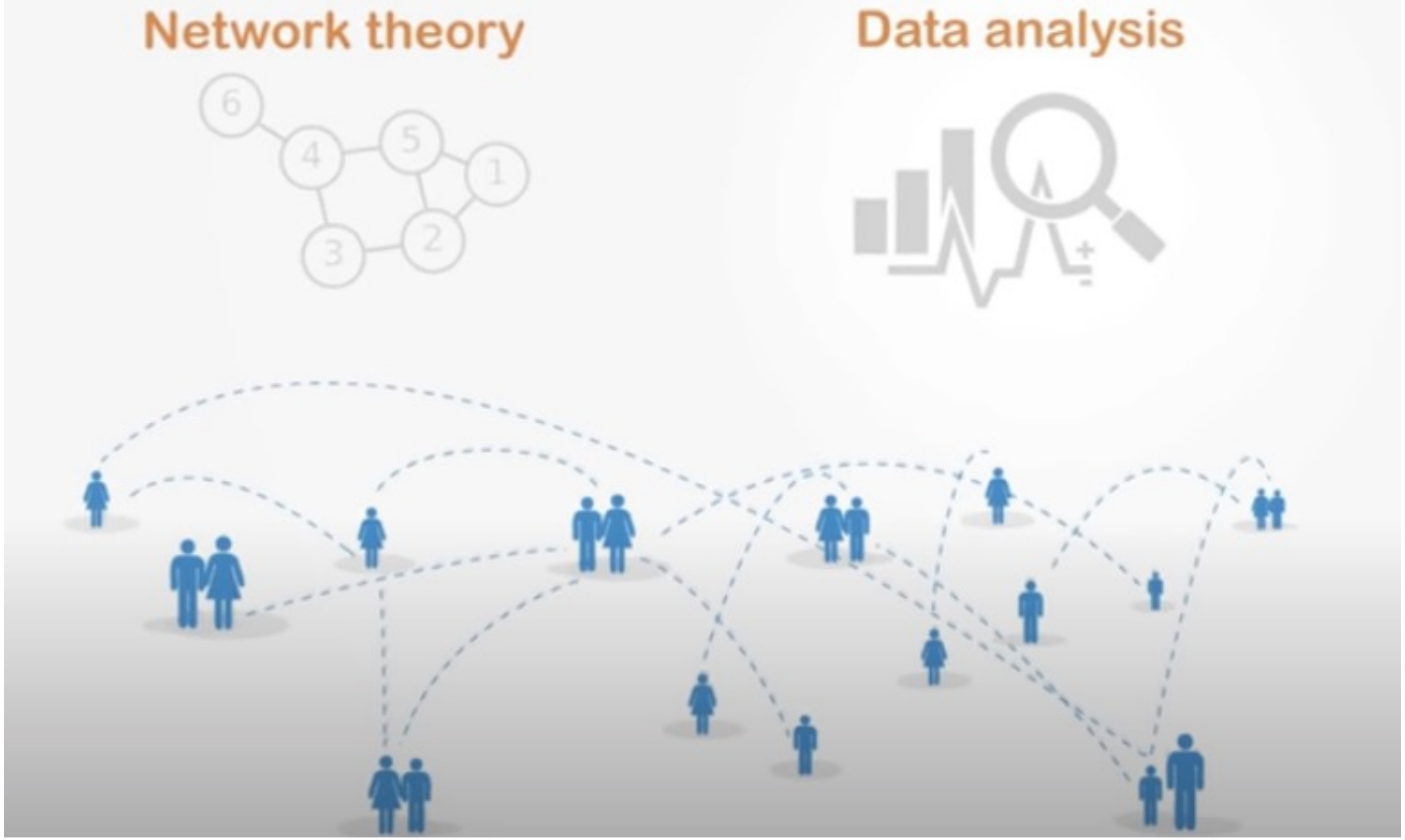
12h window

# Agenda for Today

- Paper reading presentation
- Social network analysis

- Kenmei, B., Antoniol, G., & Di Penta, M. (2008). Trend analysis and issue prediction in large-scale open source systems. Software Maintenance and Reengineering, 2008. CSMR 2008. 12th European Conference on (pp. 73-82). IEEE.
- Jadidi, M., Karimi, F., & Wagner, C. (2017). Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Male and Female Computer Scientists. arXiv preprint arXiv:1704.05801.

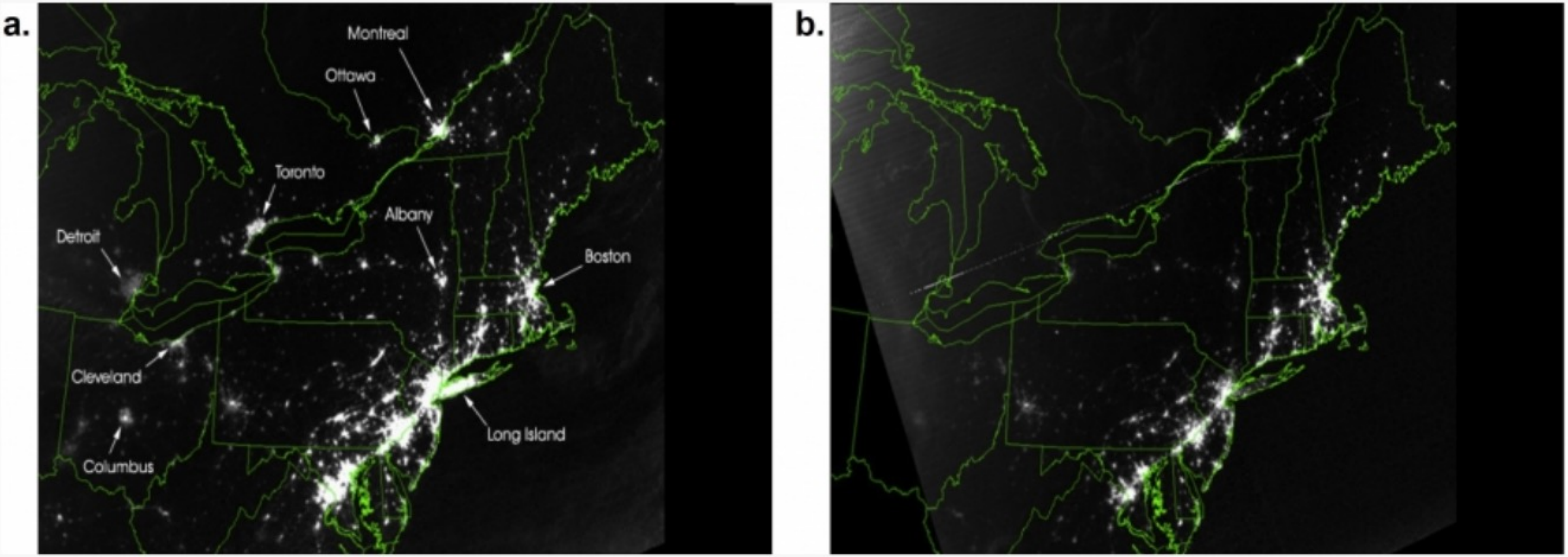
# Social Network Analysis Overview



# Agenda for Today

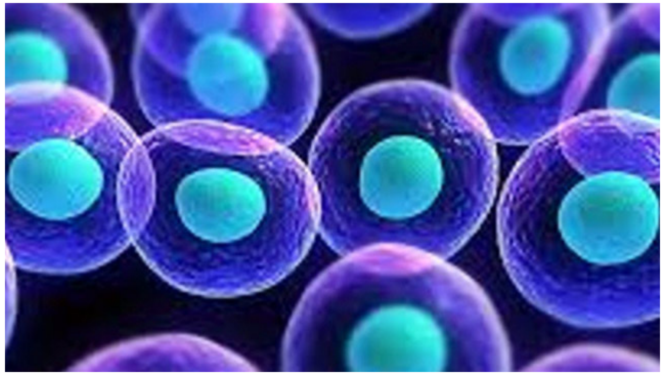
- Paper reading presentation
- ➡ Network analysis
- Social network analysis

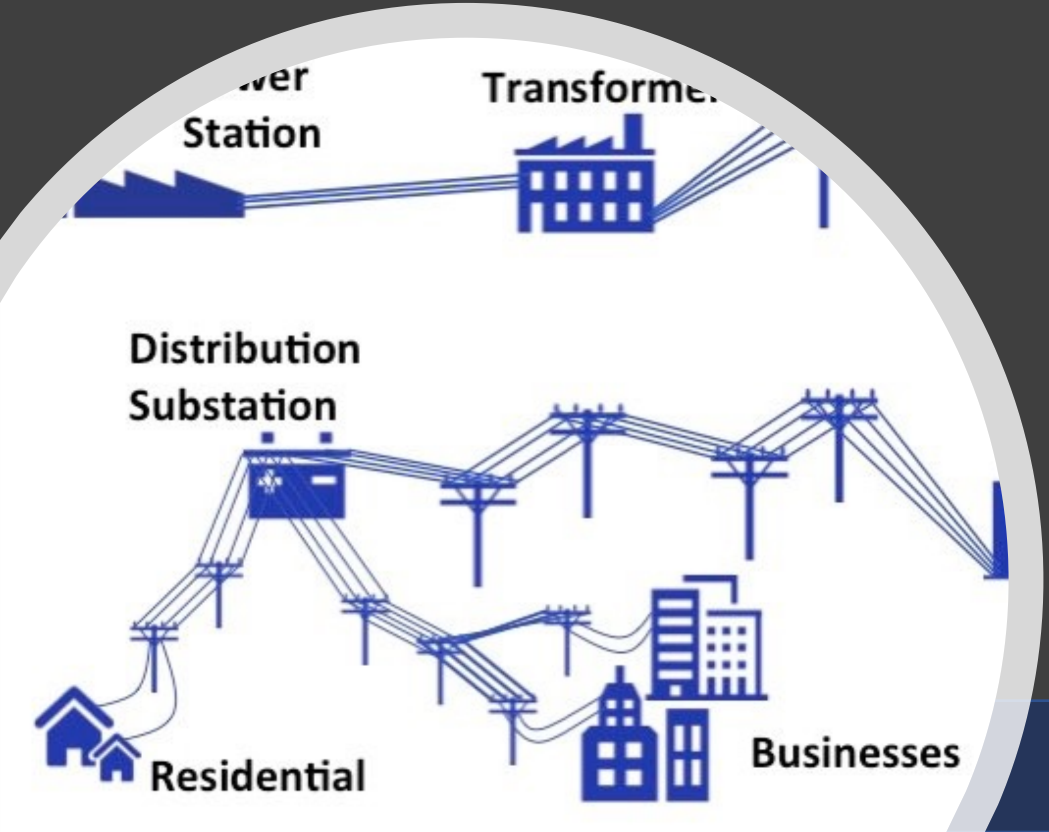
# US Northeast on August 14, 2003, before and after the blackout



# Cascading Failures in Complex Systems

- internet DDOS (routers)
- 2009-2011 financial meltdown
- money supply of terrorist orgs
- cancer cells





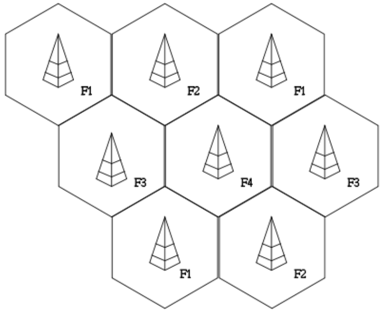
# Learned from the Northeast Blackout

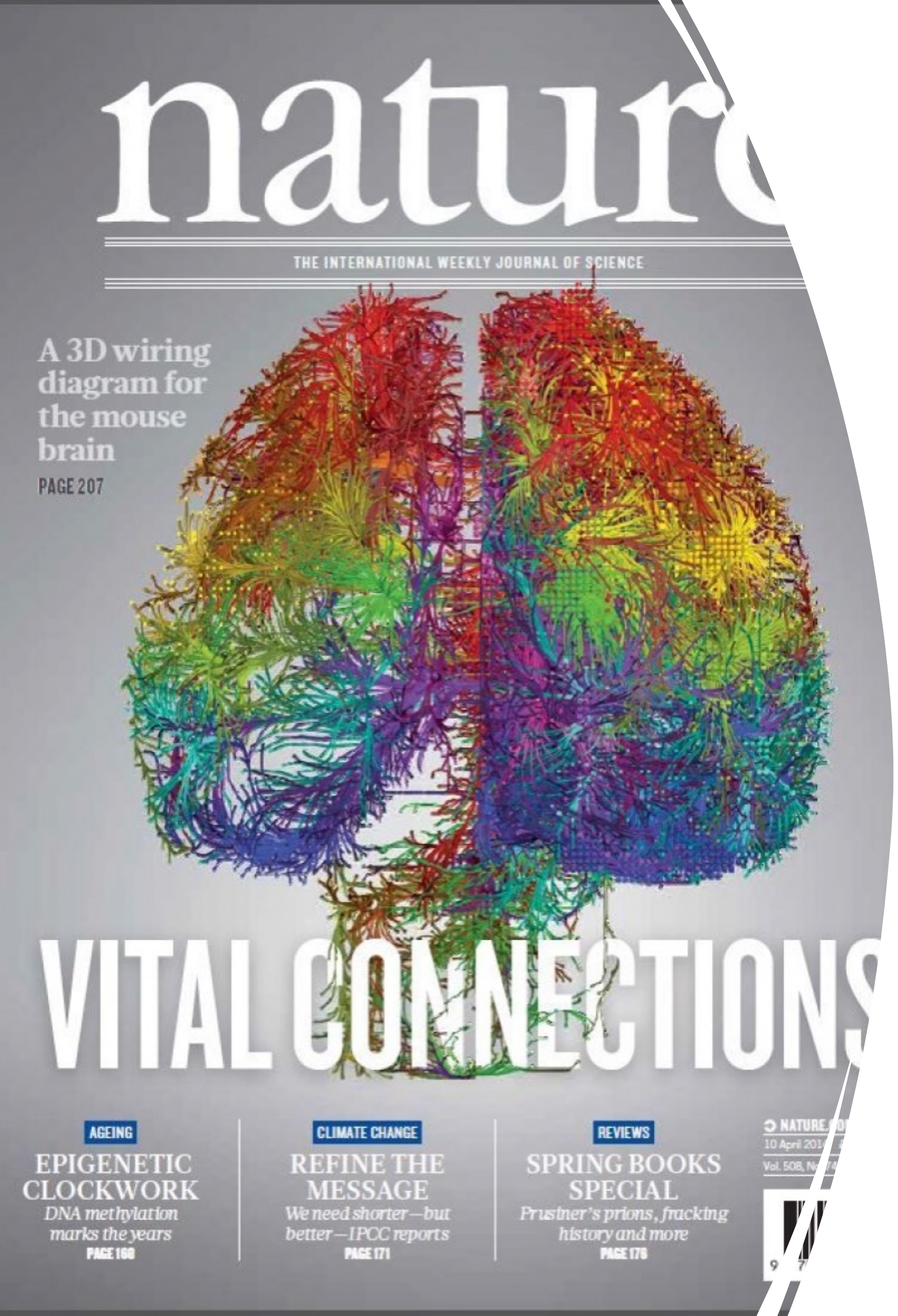
- understand the structure of the network
- model the dynamical processes
- uncover how the interplay between the network structure and dynamics affects the robustness of the whole system
- *vulnerability due to interconnectivity*

# Networks at the Heart of Complex Systems

*“I think the next century will be the century of complexity.”*

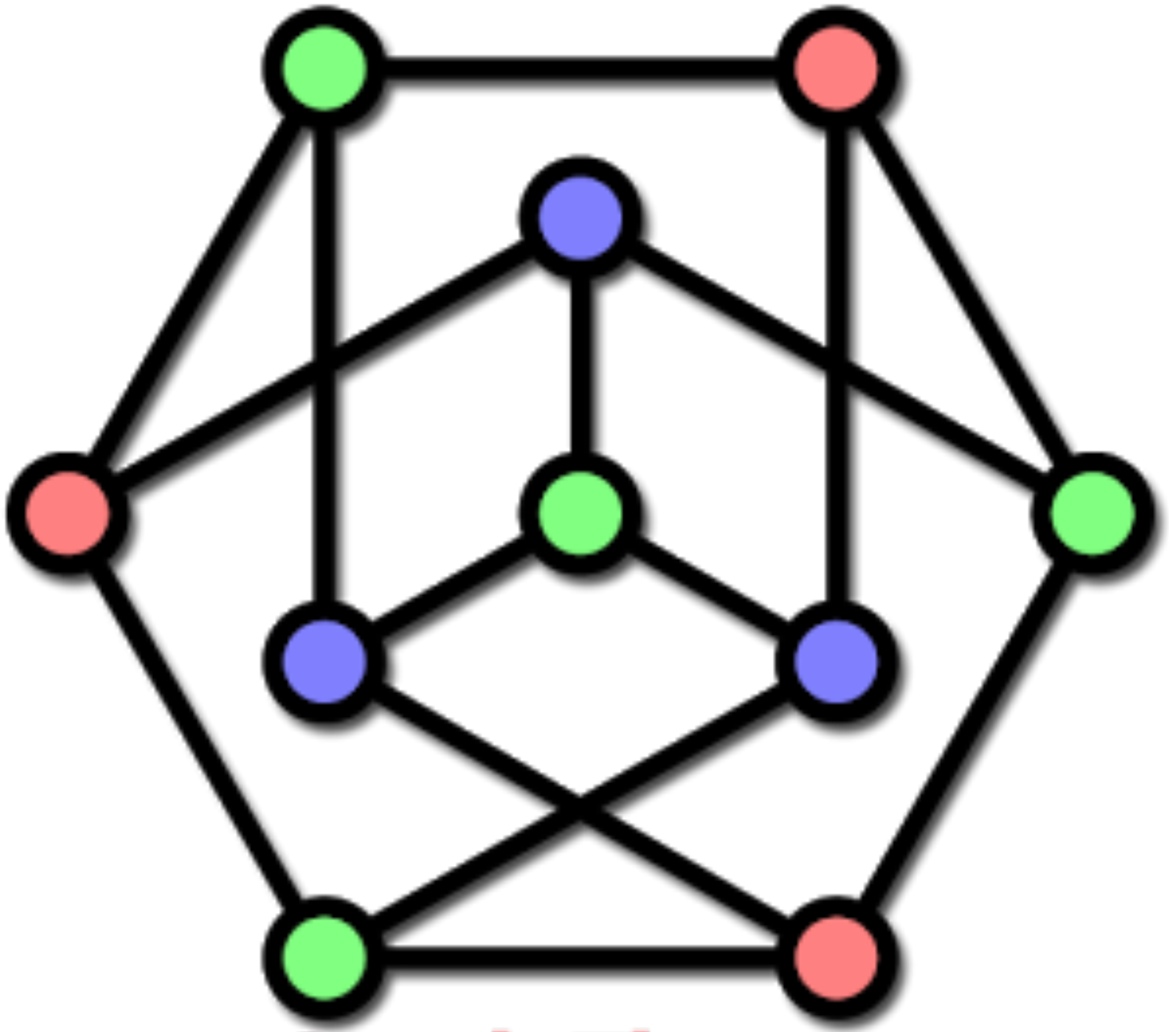
*---Stephen Hawking*



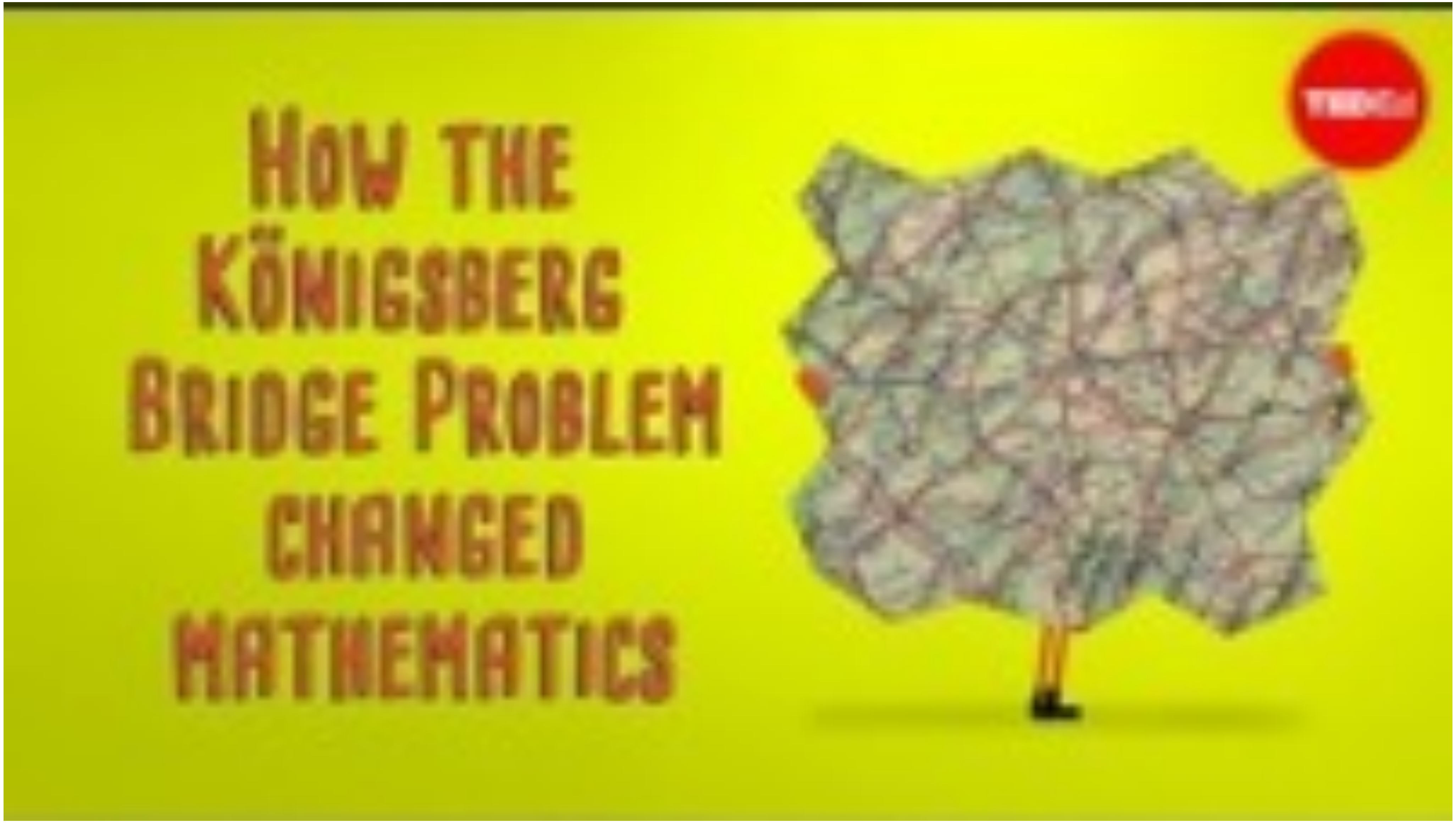

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
A 3D wiring diagram for the mouse brain
PAGE 207
VITAL CONNECTIONS
AGEING
EPIGENETIC CLOCKWORK
DNA methylation marks the years
PAGE 168
CLIMATE CHANGE
REFINE THE MESSAGE
We need shorter—but better—IPCC reports
PAGE 171
REVIEWS
SPRING BOOKS SPECIAL
Prusiner's prions, fracking history and more
PAGE 176

# The characteristics of Network Science

- Interdisciplinary Nature
- Empirical, Data Driven Nature
- Quantitative and Mathematical Nature
- Computational Nature

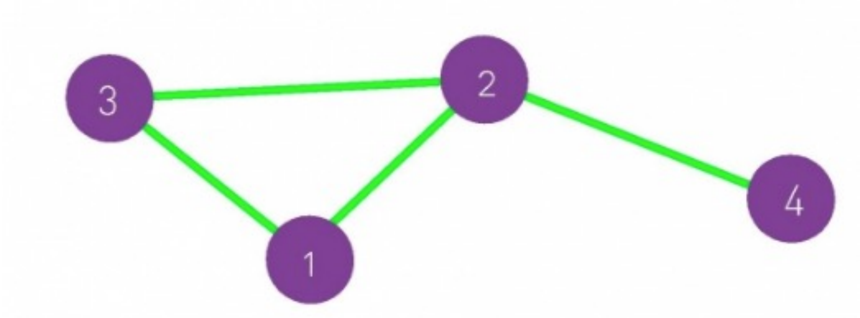


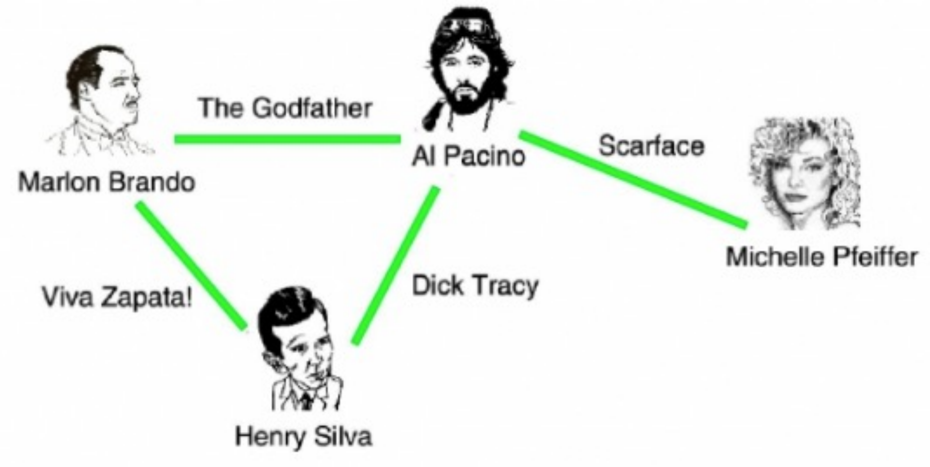
# Graph Theory

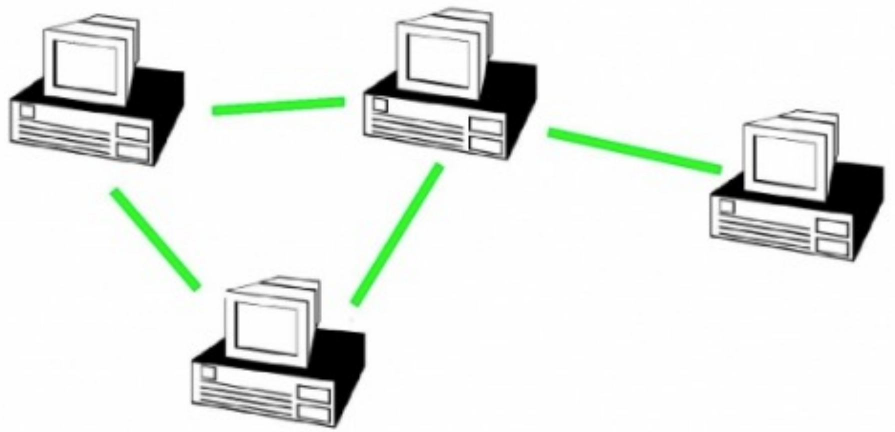

How the Königsberg Bridge Problem changed mathematics
TED-Ed

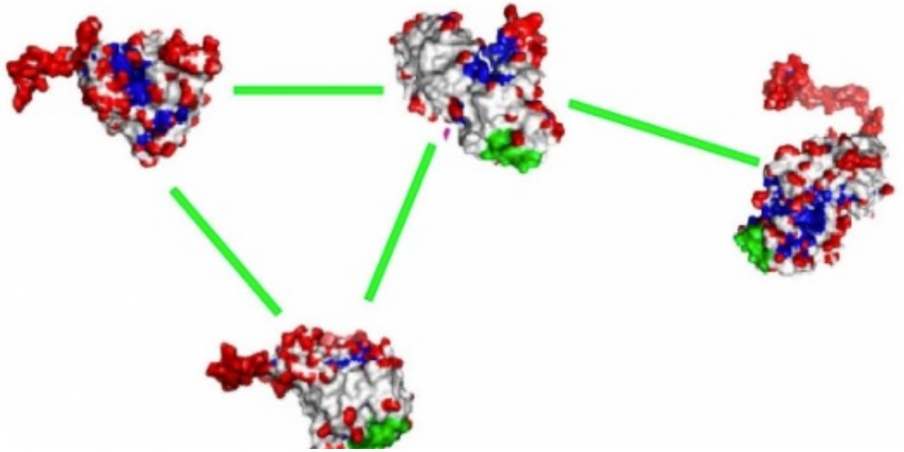
# Networks and Graphs

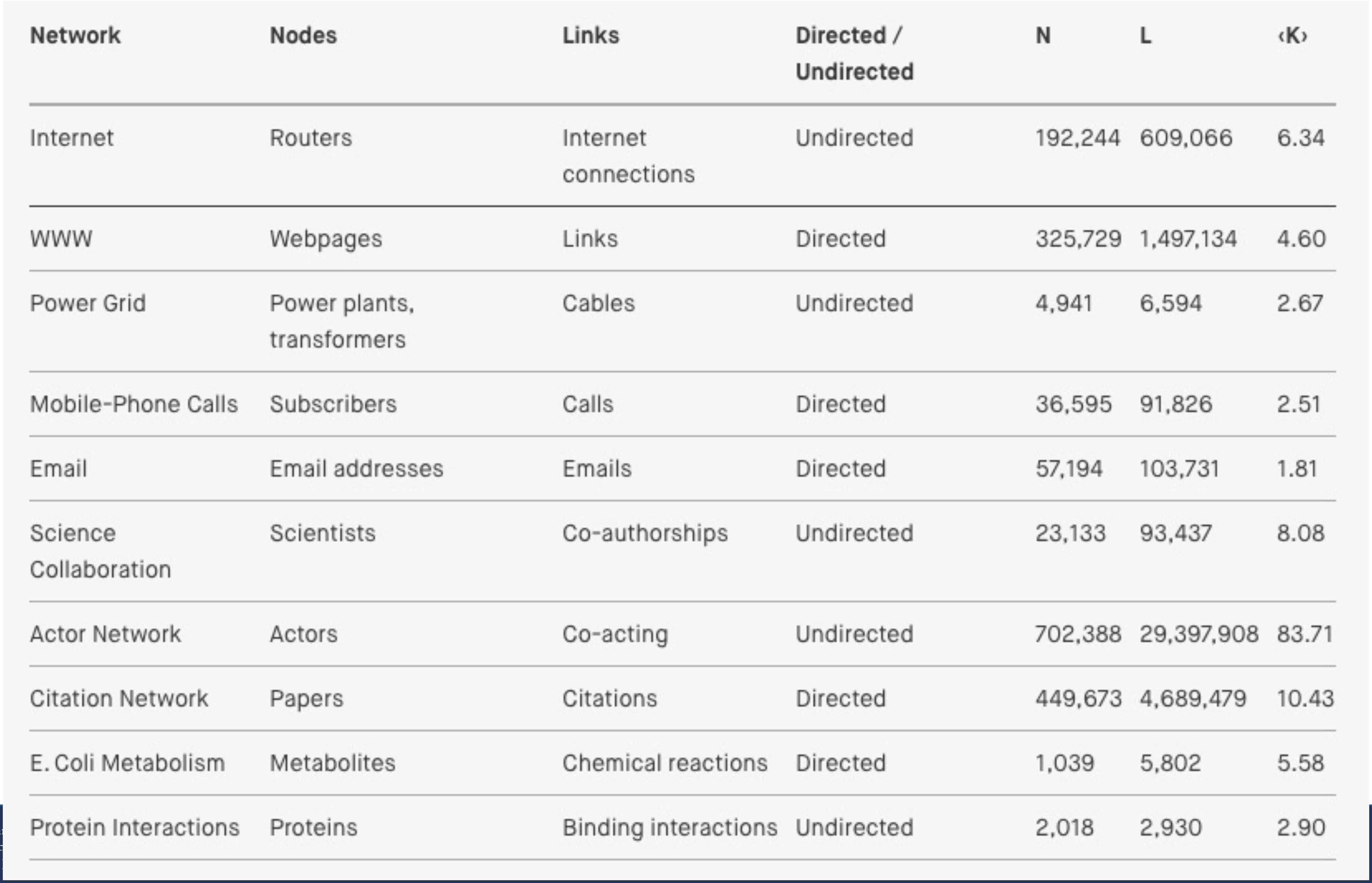
- *nodes* or *vertices*
- *links* or *edges*
- *directed* or *undirected*
- *degree*





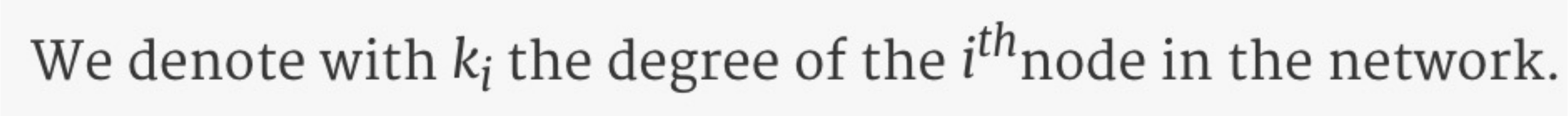




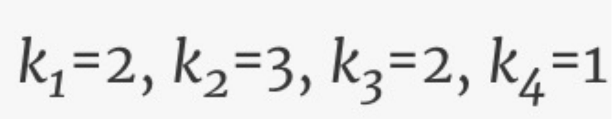


| Network | Nodes | Links | Directed / Undirected | N | L | ‹K› |
|---|---|---|---|---|---|---|
| Internet | Routers | Internet connections | Undirected | 192,244 | 609,066 | 6.34 |
| WWW | Webpages | Links | Directed | 325,729 | 1,497,134 | 4.60 |
| Power Grid | Power plants, transformers | Cables | Undirected | 4,941 | 6,594 | 2.67 |
| Mobile-Phone Calls | Subscribers | Calls | Directed | 36,595 | 91,826 | 2.51 |
| Email | Email addresses | Emails | Directed | 57,194 | 103,731 | 1.81 |
| Science Collaboration | Scientists | Co-authorships | Undirected | 23,133 | 93,437 | 8.08 |
| Actor Network | Actors | Co-acting | Undirected | 702,388 | 29,397,908 | 83.71 |
| Citation Network | Papers | Citations | Directed | 449,673 | 4,689,479 | 10.43 |
| E. Coli Metabolism | Metabolites | Chemical reactions | Directed | 1,039 | 5,802 | 5.58 |
| Protein Interactions | Proteins | Binding interactions | Undirected | 2,018 | 2,930 | 2.90 |

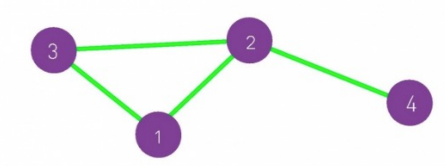
# Degree, Average Degree and Degree Distribution



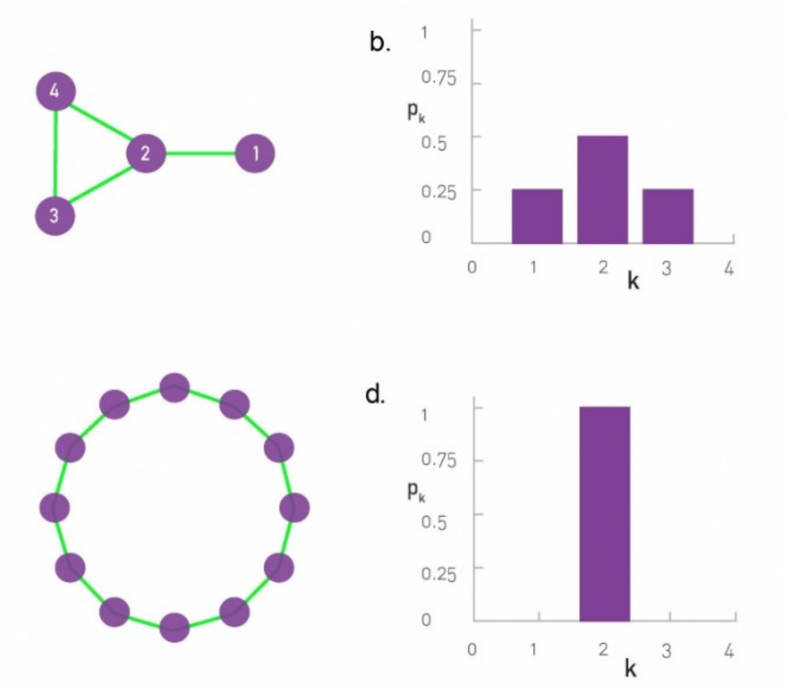
We denote with $k_i$ the degree of the $i^{th}$node in the network.



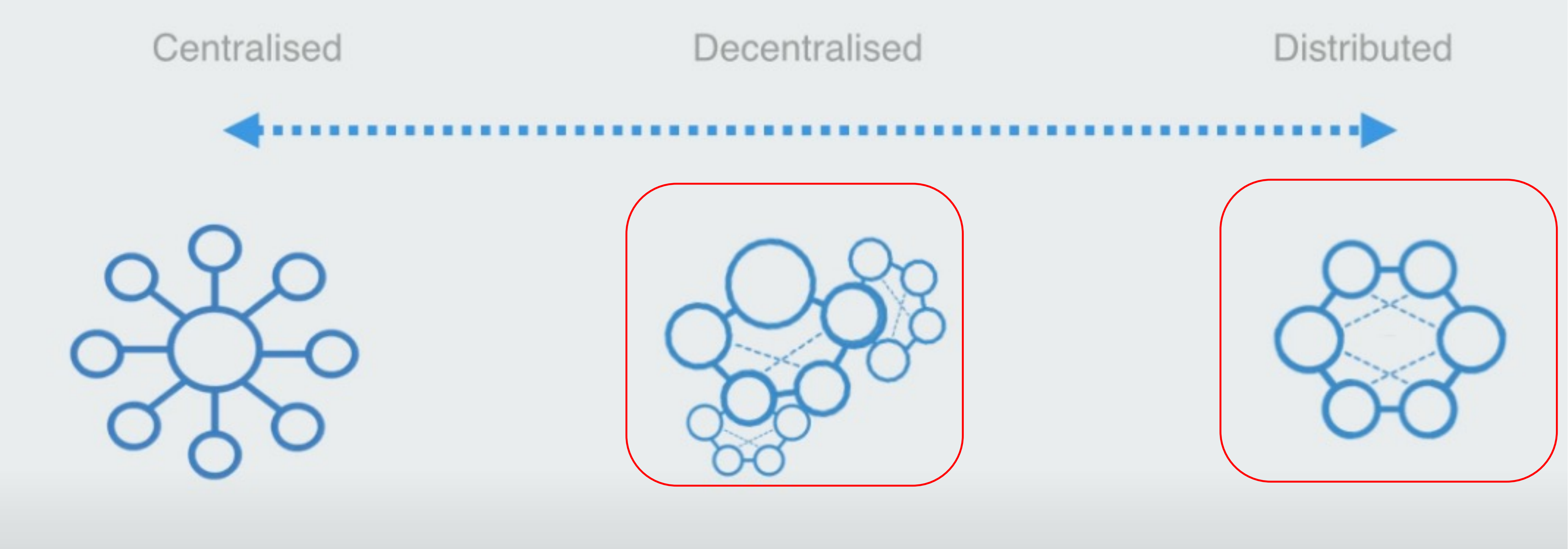
$k_1$=2, $k_2$=3, $k_3$=2, $k_4$=1



The **degree distribution**, pk, provides the probability that a randomly selected node in the network has degree k

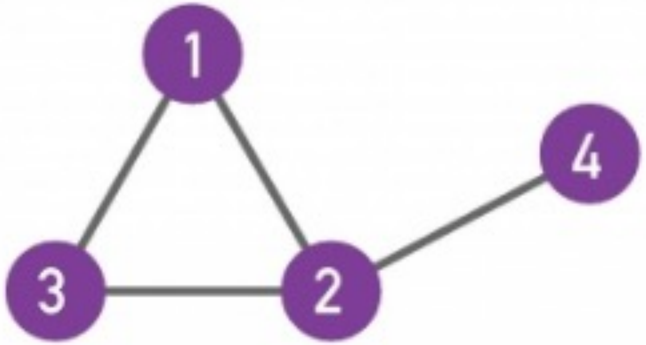


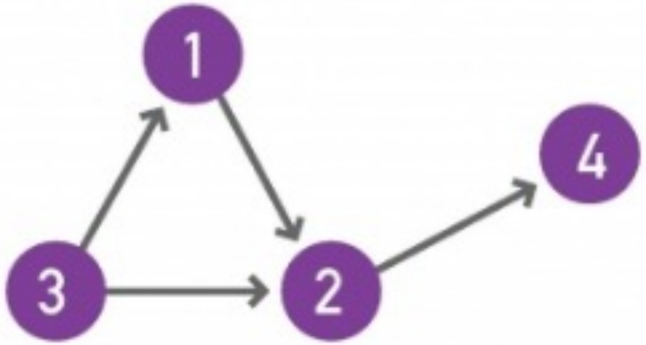
# The Spectrum of Degree Distribution



# Adjacency Matrix

- a directed network of N nodes has N rows and N columns





$$A_{ij} = \begin{matrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{matrix}$$

$$A_{ij} = \begin{matrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

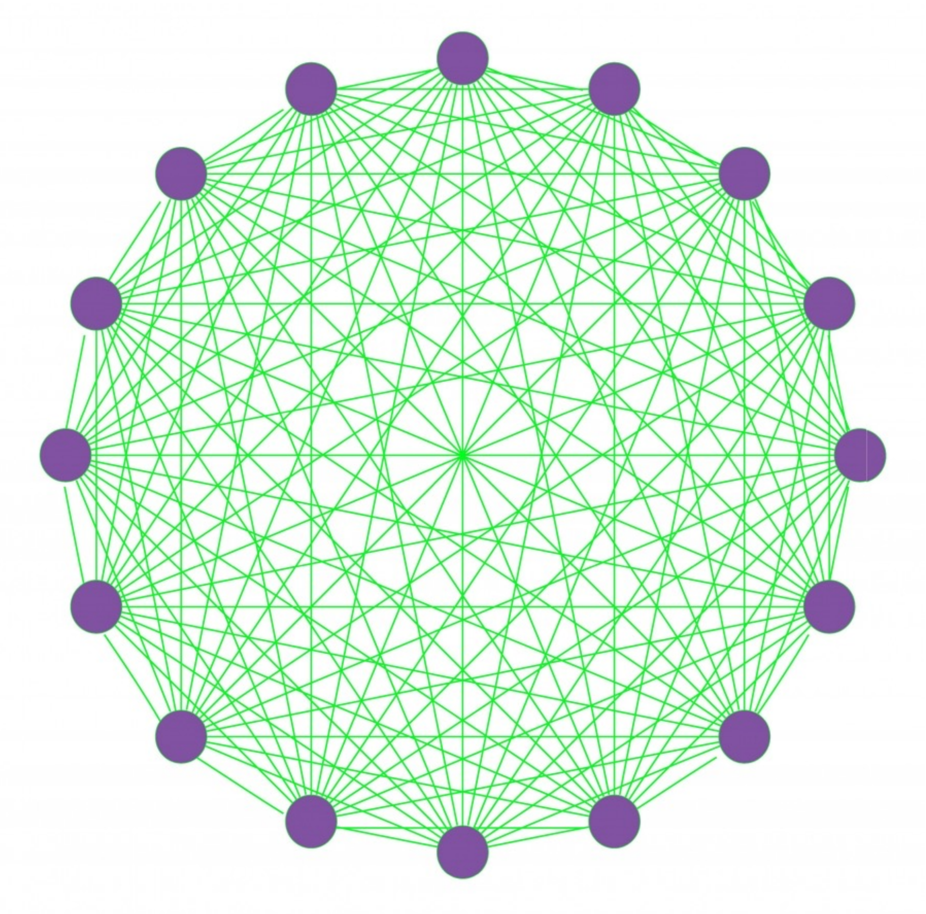
$$A_{ij} = \begin{matrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

# In real networks, the number of nodes ($N$) and links ($L$) can vary widely...

- neural network of the worm C. elegans, the only fully mapped nervous system of a living organism, has $N = 302$ neurons (nodes
- human brain is estimated to have about a hundred billion ($N \approx 10^{11}$) neurons.
- The genetic network of a human cell has about 20,000 genes as nodes;
- social network consists of seven billion individuals ($N \approx 7\times10^{9}$)
- WWW is estimated to have over a trillion web documents ($N > 10^{12}$).

# Real Networks are Sparse

Sparseness has important consequences on the way we explore and store real networks

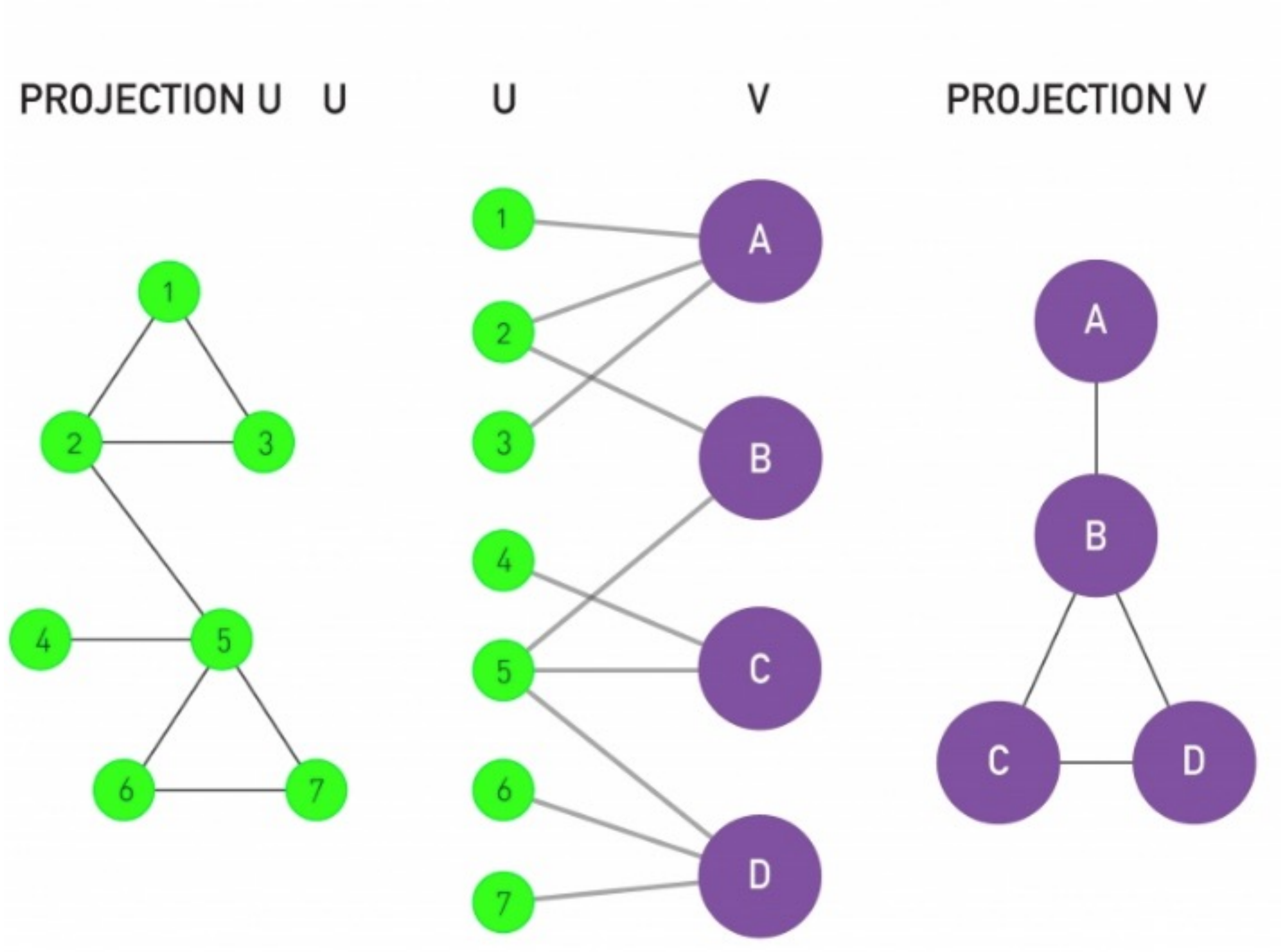


**Complete Graph (Clique)**

# Weighted Networks

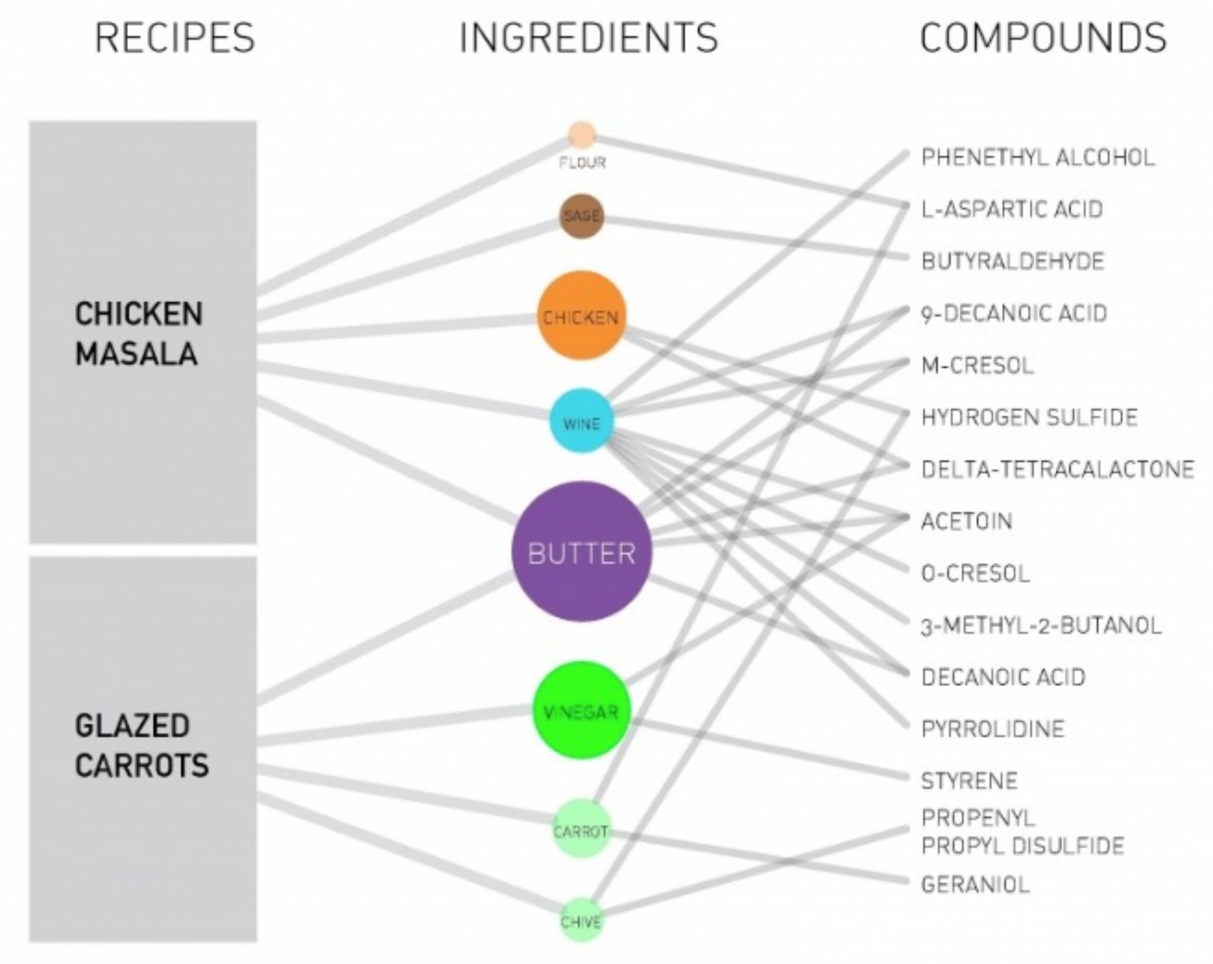
- In **mobile call networks** the weight can represent the total number of minutes two individuals talk with each other on the phone;
- on the **power grid** the weight is the amount of current flowing through a transmission line.

# Bipartite Networks



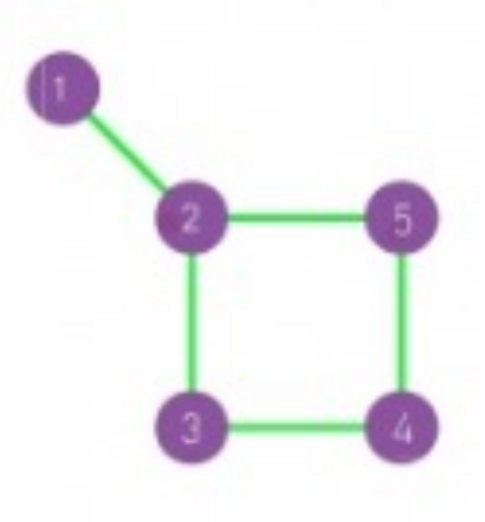
- e.g., programmers and projects on GitHub

# Multipartite Networks



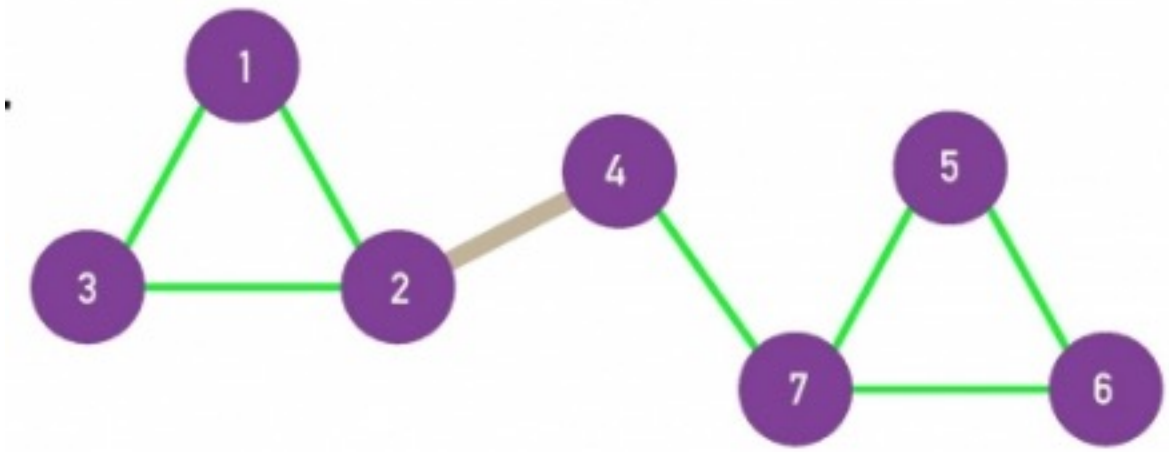
# Paths and distances

- Shortest path: between nodes i and j is the path with the fewest number of links
- The diameter of a network, denoted by $d_{max}$ is the maximum shortest path in the network



# Connectedness

- In an undirected network nodes i and j are connected if there is a path between them.
- A network is connected if all pairs of nodes in the network are connected.

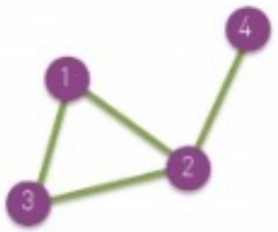


$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

# Clustering coefficient

- Def: the probability that two neighbors of a given node link to each other.
- Ci measures the network’s local link density: The more densely interconnected the neighborhood of node i, the higher is its local clustering coefficient.
  - ranges from 0 (when none of the node’s friends are friends with each other)
  - to 1 (when all of the node’s friends are friends with each other)
  - the more strongly triadic closure is operating in the neighborhood of the node, the higher the clustering coefficient will tend to be.

## a. Undirected

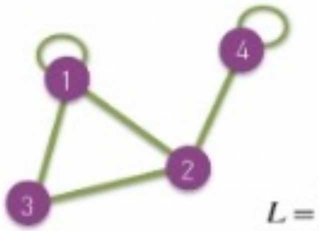


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0$ $A_{ij} = A_{ji}$

$L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij}$ $<k> = \frac{2L}{N}$

## b. Self-loops



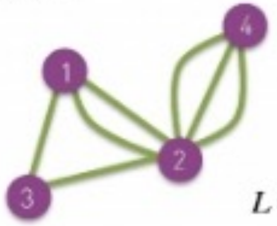
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$\exists i, A_{ii} \neq 0$ $A_{ij} = A_{ji}$

$L = \frac{1}{2} \sum_{i,j=1,i\neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}$ ?

## c. Multigraph

(undirected)

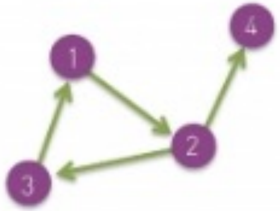


$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0$ $A_{ij} = A_{ji}$

$L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij}$ $<k> = \frac{2L}{N}$

## d. Directed



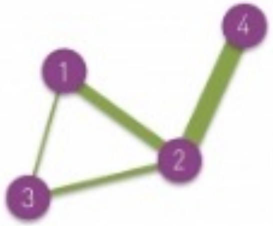
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$A_{ij} \neq A_{ji}$

$L = \sum_{i,j=1}^{N} A_{ij}$ $<k> = \frac{L}{N}$

## e. Weighted

(undirected)



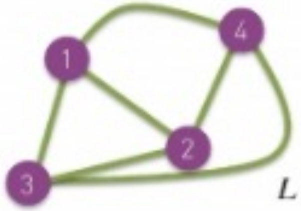
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0$ $A_{ij} = A_{ji}$

$<k> = \frac{2L}{N}$

## f. Complete Graph

(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

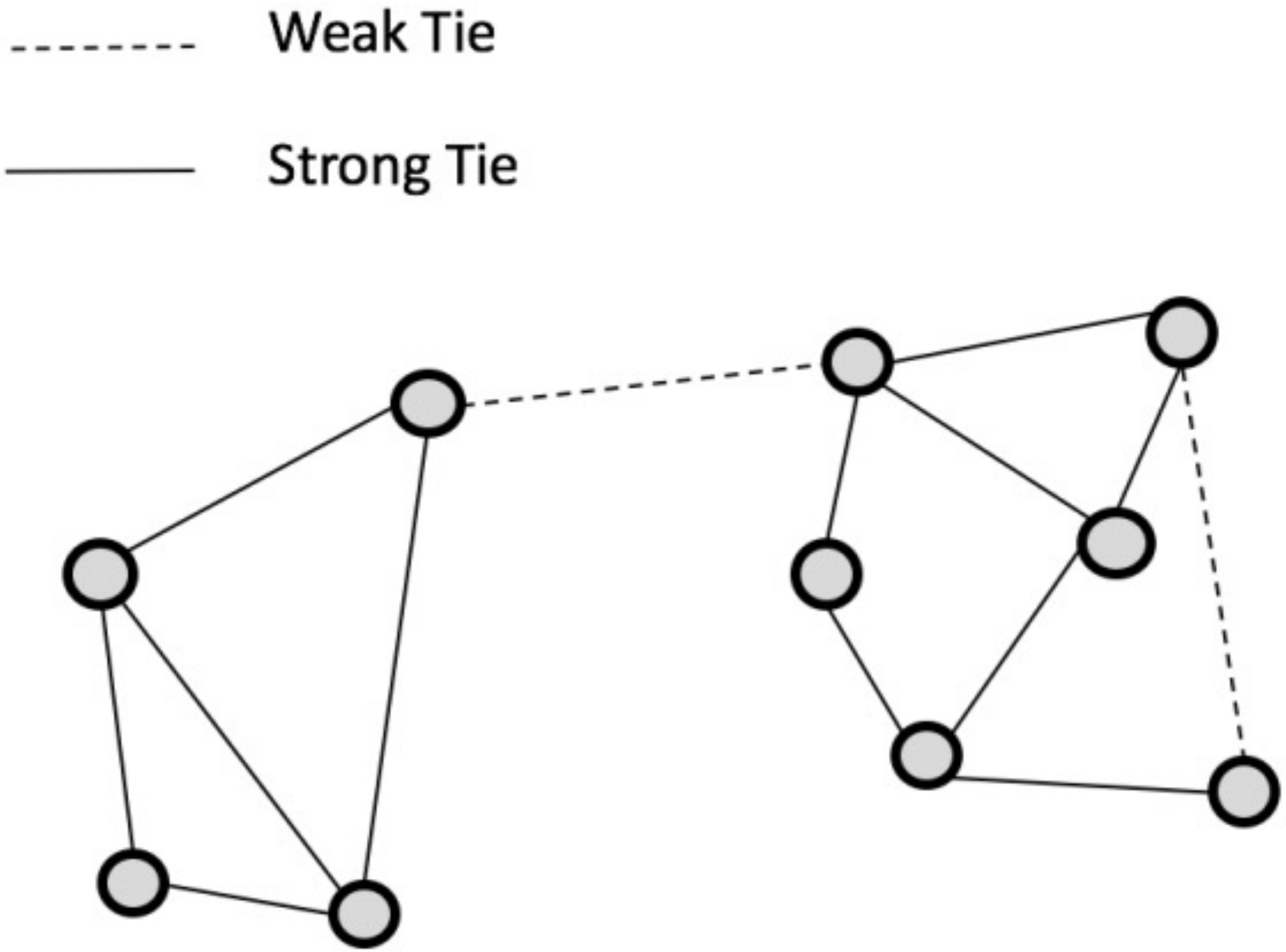
$A_{ii} = 0$ $A_{i\neq j} = 1$

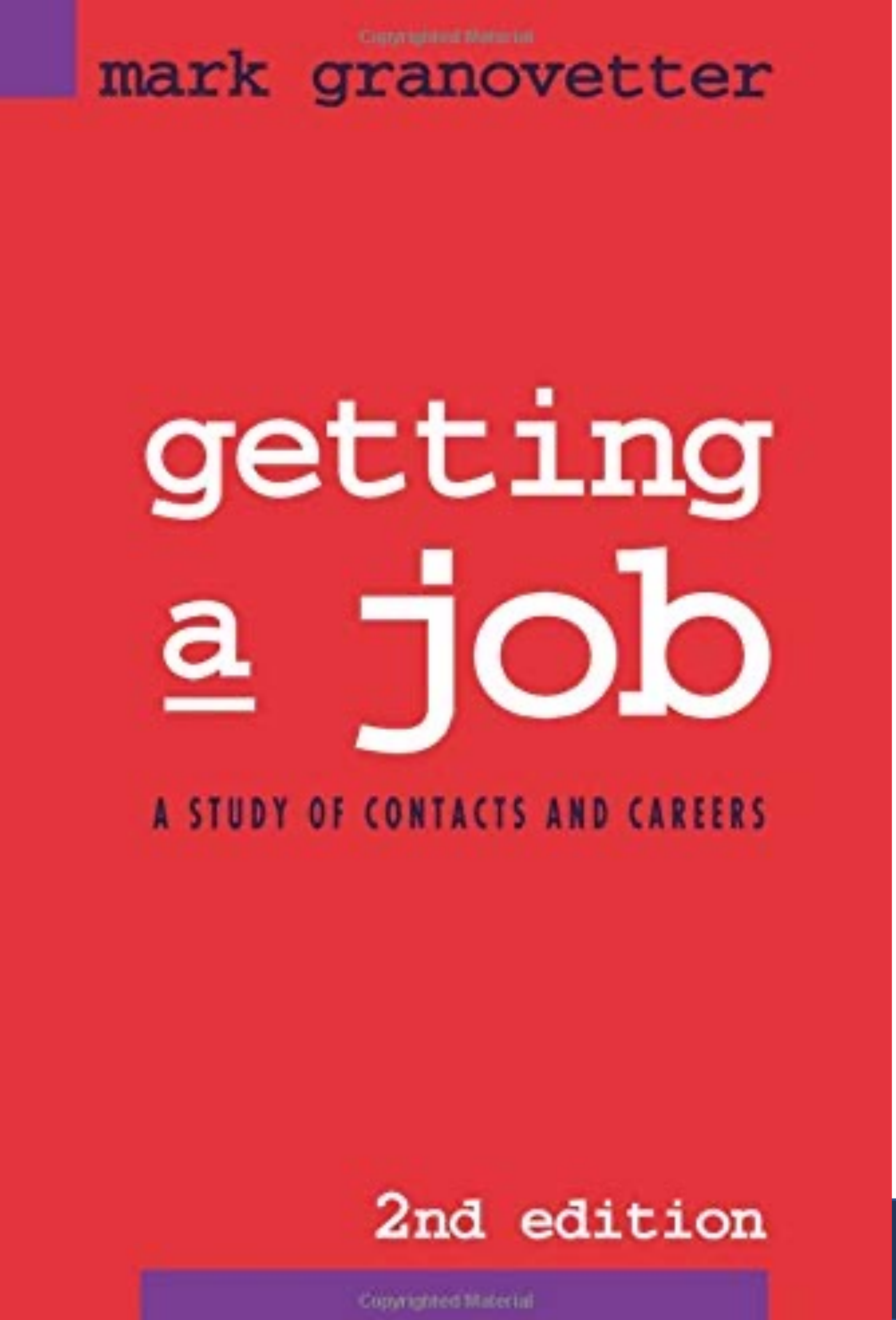
$L = L_{\max} = \frac{N(N-1)}{2}$ $<k> = N - 1$

# Agenda for Today

- Paper reading presentation
- Network analysis
- ➡ Social network analysis

# Strong and Weak Ties

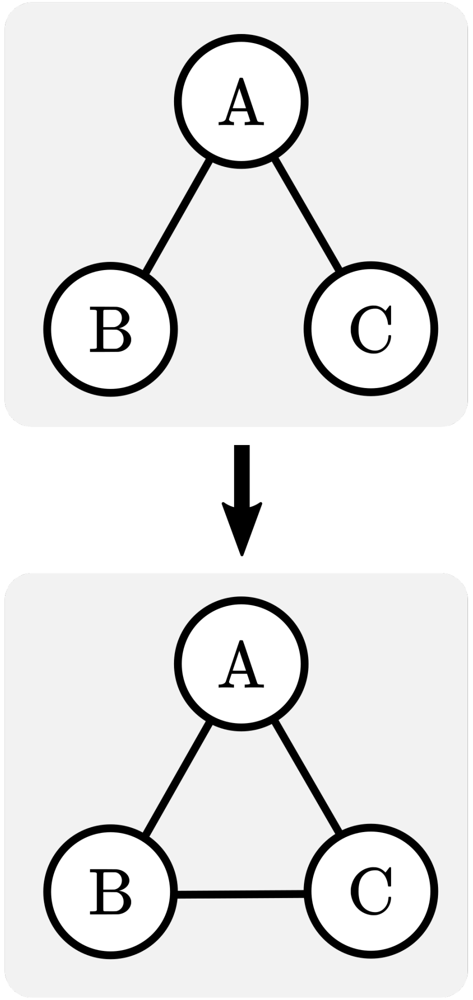




"these personal contacts were often described by interview subjects as acquaintances rather than close friends. "

Two different perspectives on distant friendships:

- structural
- interpersonal

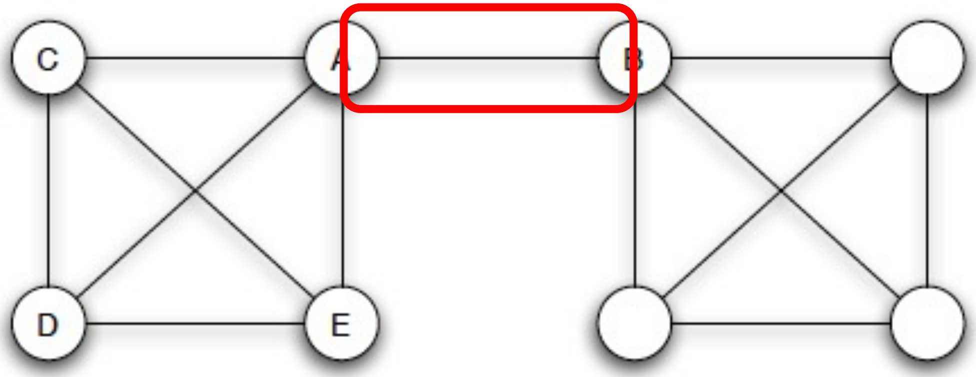


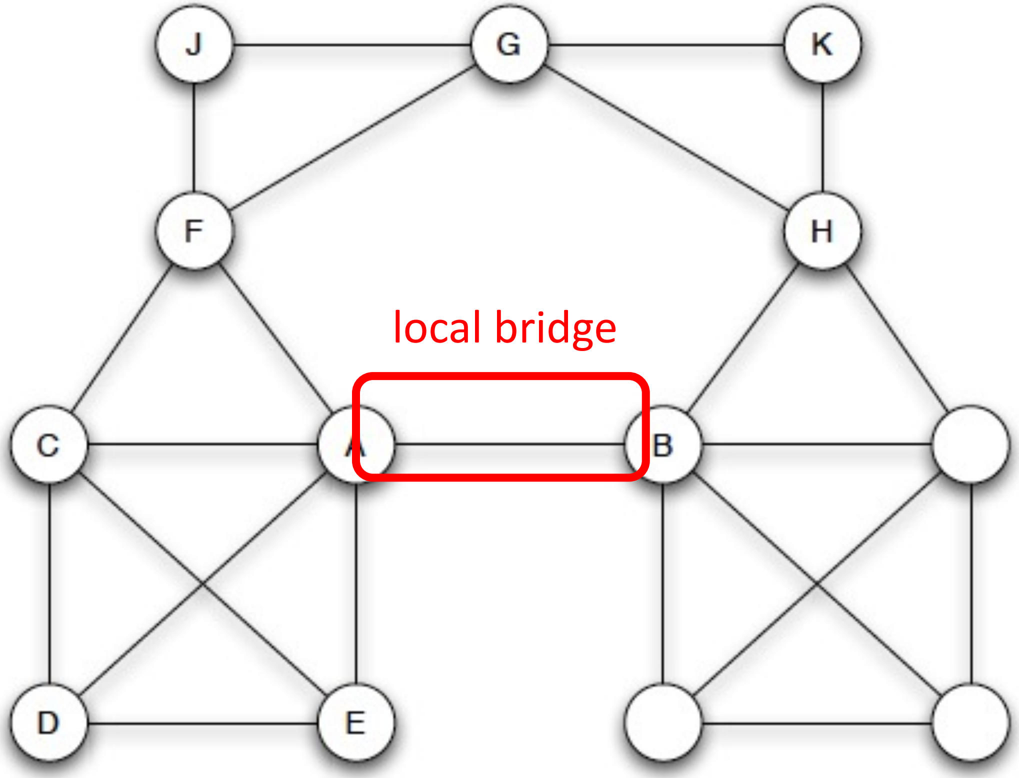
# Triadic closure

- If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future

# The Strength of Weak Ties

- Bridges and Local Bridges



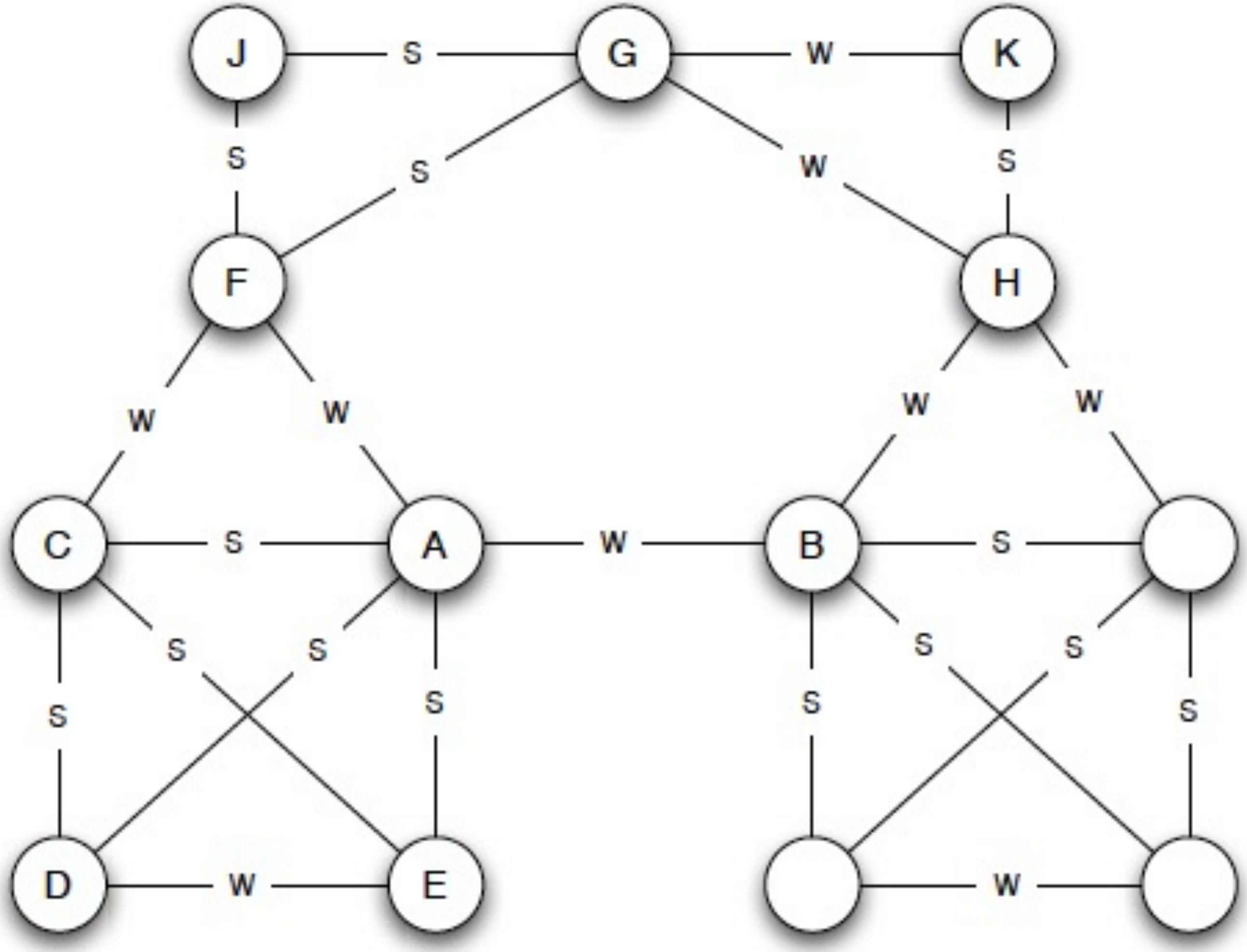


**Span** of a local bridge is the distance its endpoints would be from each other if the edge were deleted

# Strong Tie and Weak tie

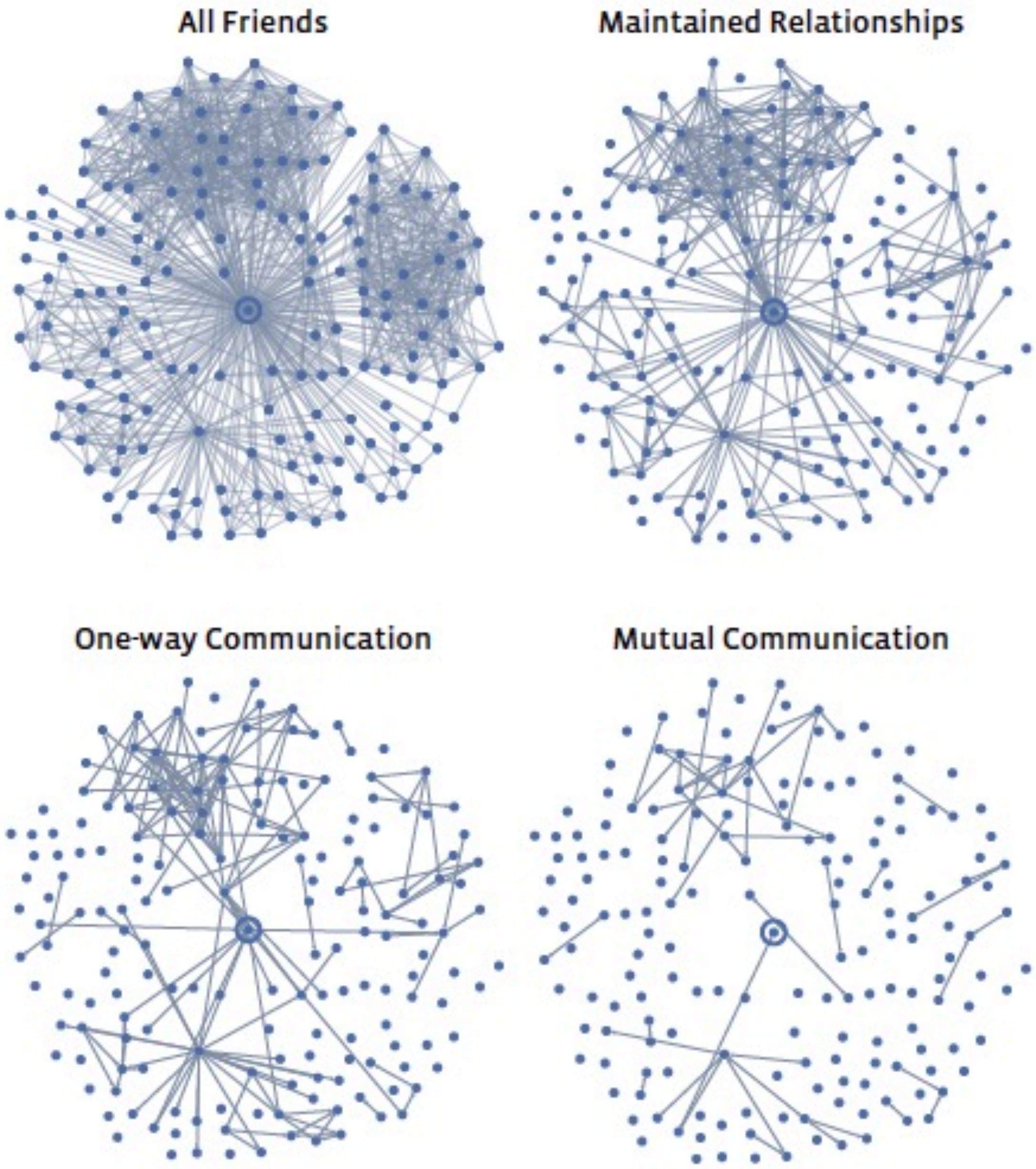
In social network

- strong ties (the stronger links, corresponding to friends),
- weak ties (the weaker links, corresponding to acquaintances)



# Four different views of a Facebook user's network neighborhood

*“even for users who report very large numbers of friends on their profile pages (on the order of 500), the number with whom they actually communicate is generally between 10 and 20”*



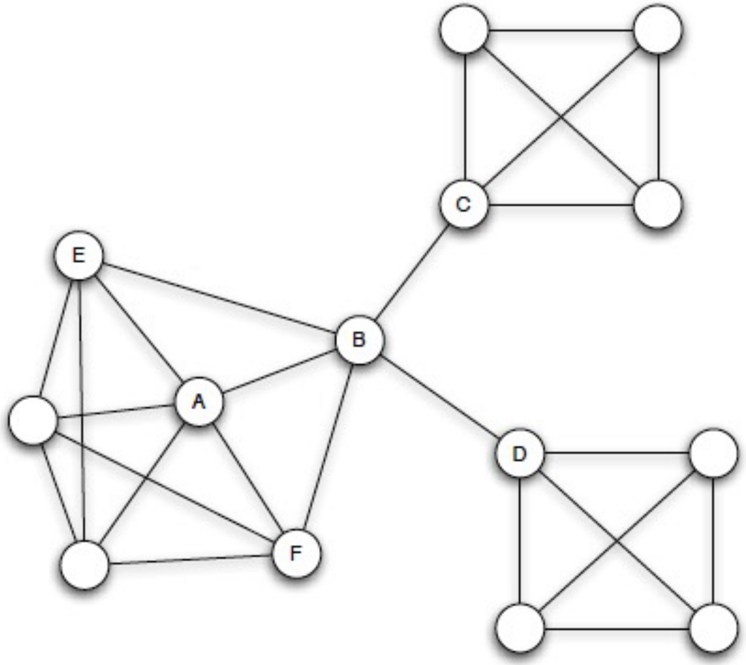
Ease of forming links

VS

Relative scarcity of strong ties

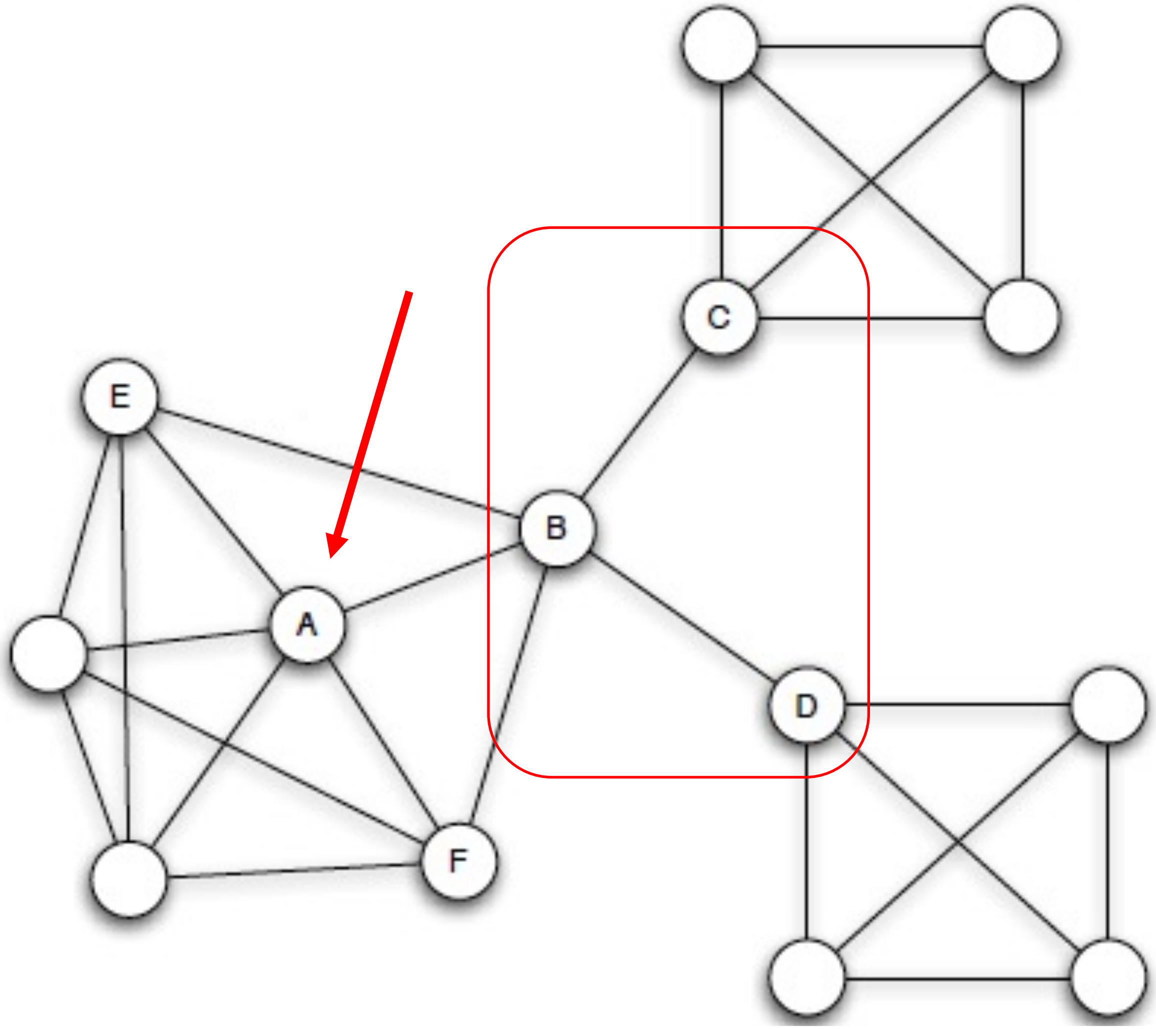
# Social networks

Few edges spanning different groups while most are surrounded by dense patterns of connections



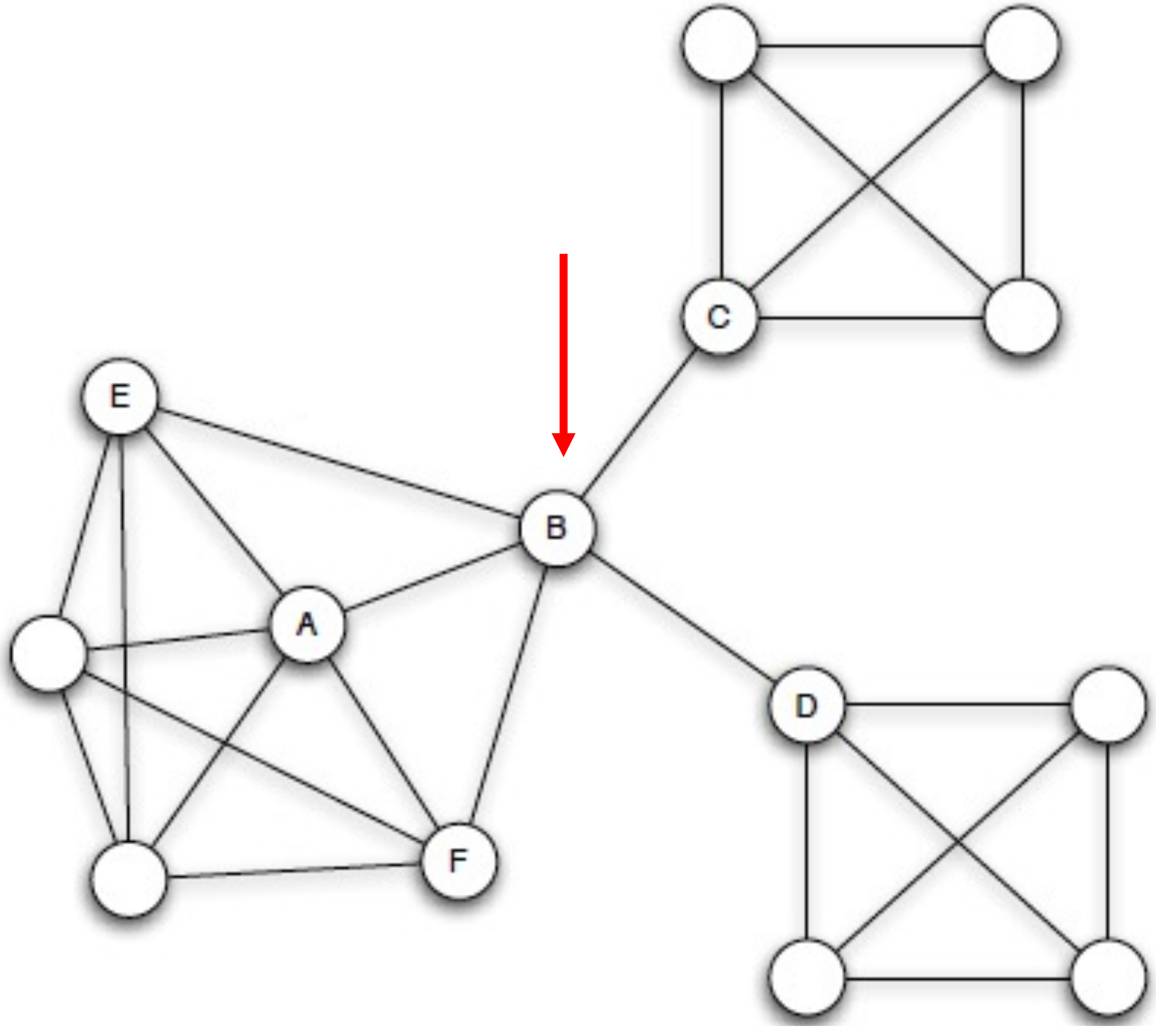
# Embeddedness

- A's set of network neighbors has been subject to considerable triadic closure;
- Embeddedness of an edge: number of common neighbors the two endpoints have
- e.g. Embeddedness (A-B) = 2
- higher trust between individuals connected by an embedded edge



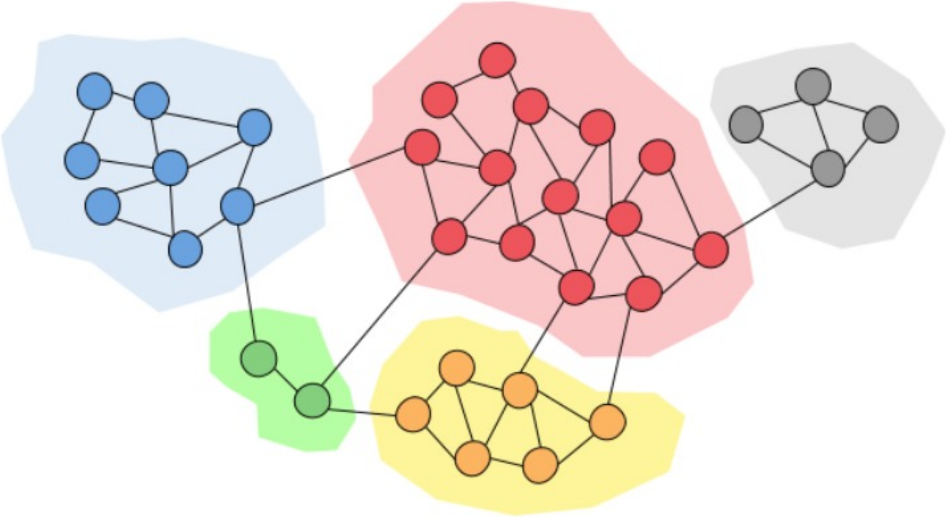
# Structural Holes

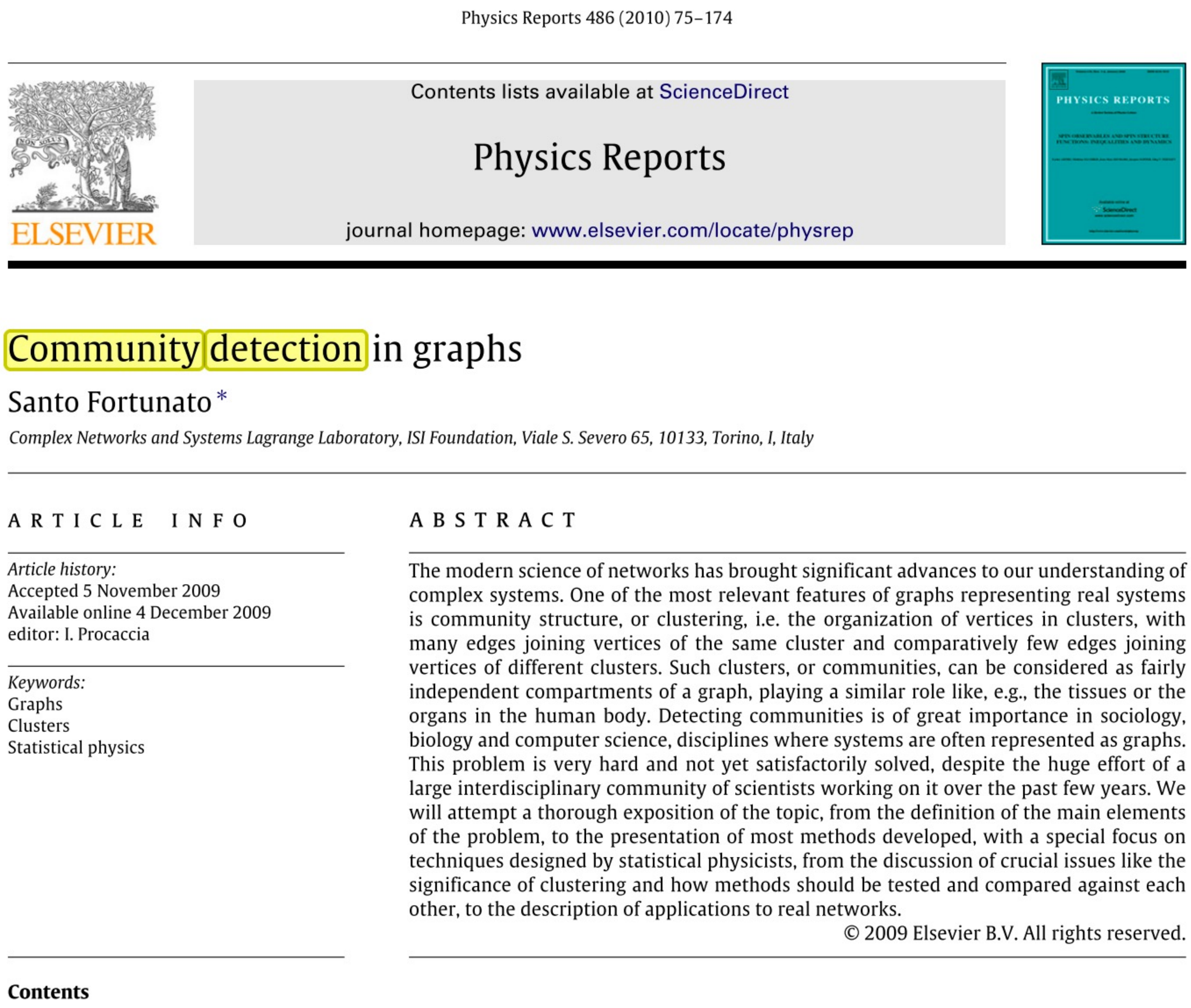
- Def: ends of multiple local bridges
- Advantages:
    - has early access to information originating in multiple, non-interacting parts of the network
    - standing at one end of a local bridge can be an amplifier for creativity
    - social "gate-keeping"



# Community

- a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities (locally dense connected subgraph)





Physics Reports 486 (2010) 75–174

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

# Community detection in graphs

Santo Fortunato*

*Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I, Italy*

ARTICLE INFO

*Article history:*
Accepted 5 November 2009
Available online 4 December 2009
editor: I. Procaccia

*Keywords:*
Graphs
Clusters
Statistical physics

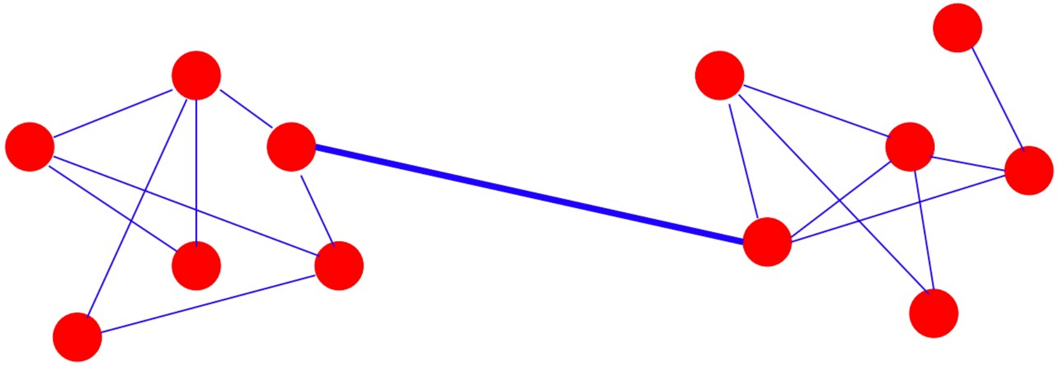
ABSTRACT

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks.

© 2009 Elsevier B.V. All rights reserved.

**Contents**

# Newman-Girvan



- Edge Centrality/Betweenness: the number of shortest paths between all node pairs that run along the edge.
- Algorithm:
  - 1. Computation of the centrality for all edges;
  - 2. Removal of edge with largest centrality: in case of ties with other edges, one of them is picked at random;
  - 3. Recalculation of centralities on the running graph;
  - 4. Iteration of the cycle from step 2

# Application scenarios

- Way, S. F., Larremore, D. B., & Clauset, A. (2016**). Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks.** In Proceedings of the 25th International Conference on World Wide Web (pp. 1169-1179). International World Wide Web Conferences Steering Committee.

# Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks

“Using comprehensive data on both **hiring outcomes and scholarly productivity for 2659 tenure-track faculty** across 205 Ph.D.-granting departments in North America, we investigate the multi-dimensional nature of gender inequality in computer science faculty hiring through a network model of the hiring process.”

Hiring outcomes are most directly affected by
(i) the relative prestige between hiring and placing institutions
(ii) the scholarly productivity of the candidates.

EXAMPLE

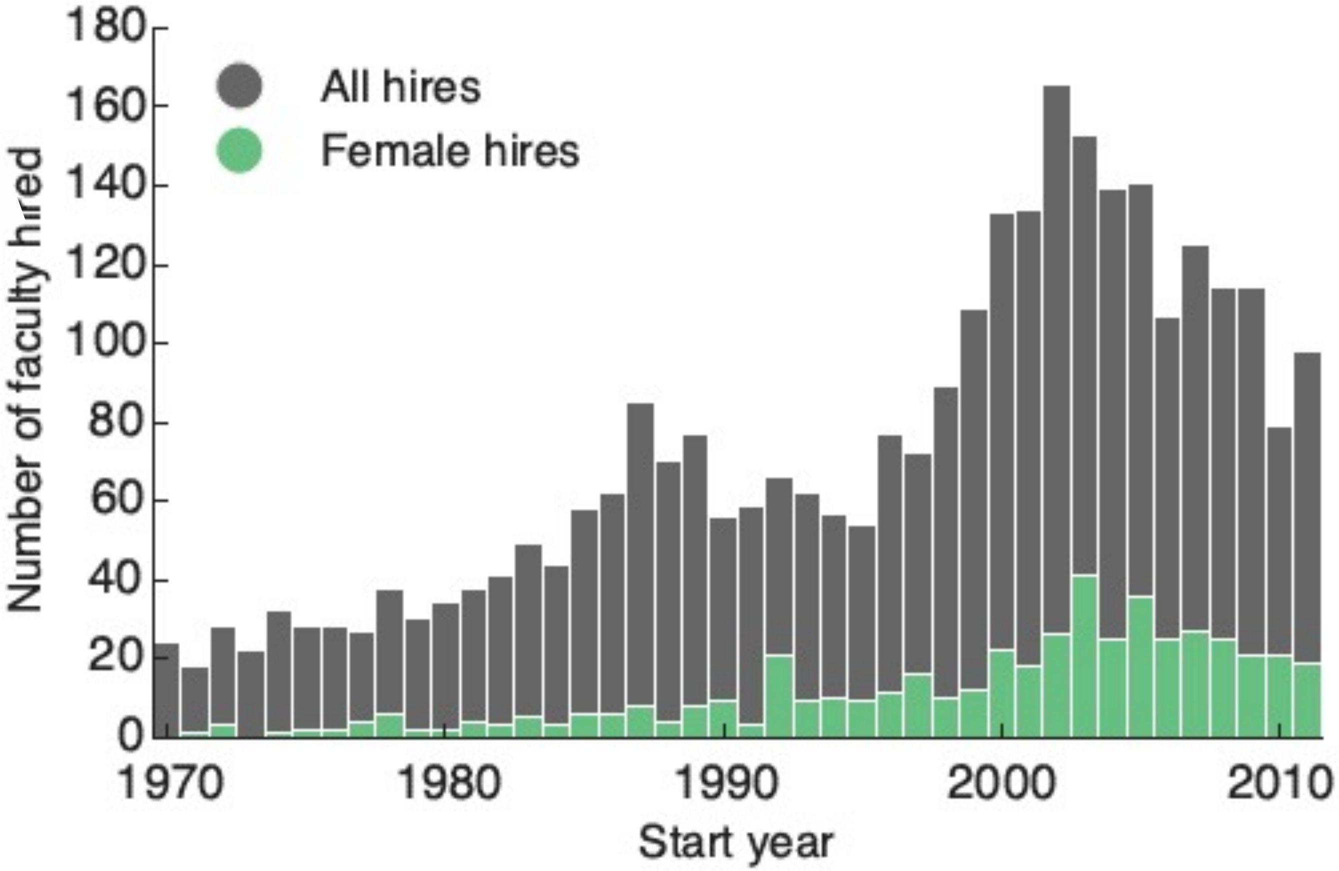


FIG. 1. For the 2659 computer science faculty in our sample (collected in 2011), the distribution of years in which they were first hired as an assistant professor.

- Network:
    - faculty hiring network:
        - a directed multigraph,
        - node is an institution
        - each Ph.D. graduate from an institution u who began as an assistant professor at v is represented by a single directed edge (u; v).
        - Each node in this network is annotated with its institution's prestige rank

- Backstrom, L., & Kleinberg, J. (2014). **Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on Facebook**. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 831-841). ACM.

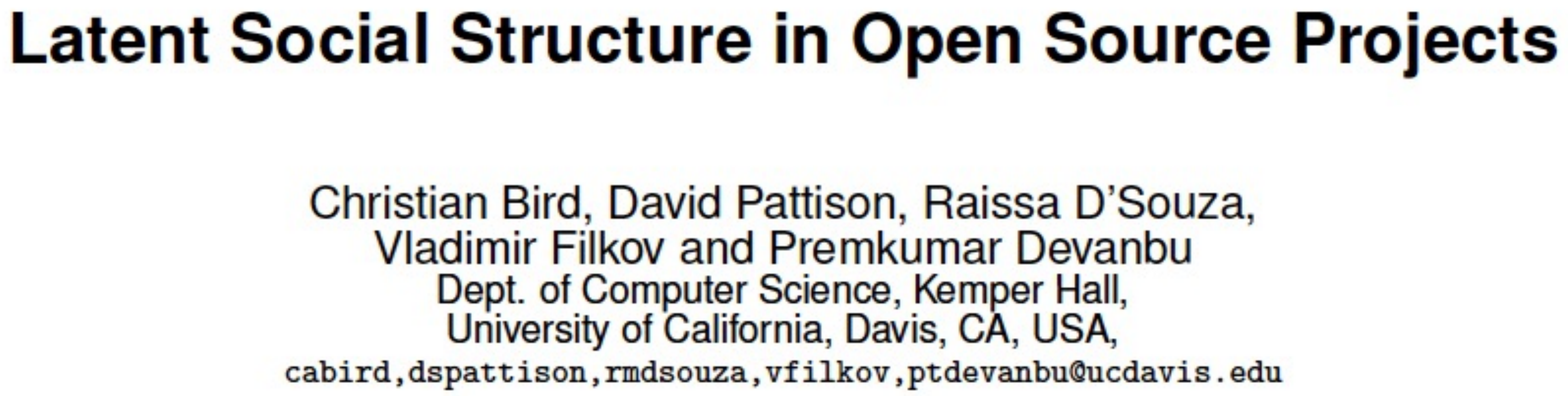
*RQ: given all the connections among a person's friends, can you recognize his or her romantic partner from the network structure alone?*

- new measure of tie strength of ‘**dispersion**’ — the extent to which two people’s mutual friends are not themselves well-connected.

- new measure of tie strength of ‘**dispersion**’ — the extent to which two people’s mutual friends are not themselves well-connected.

- dispersion measure + ML → predict romantic relation

EXAMPLE



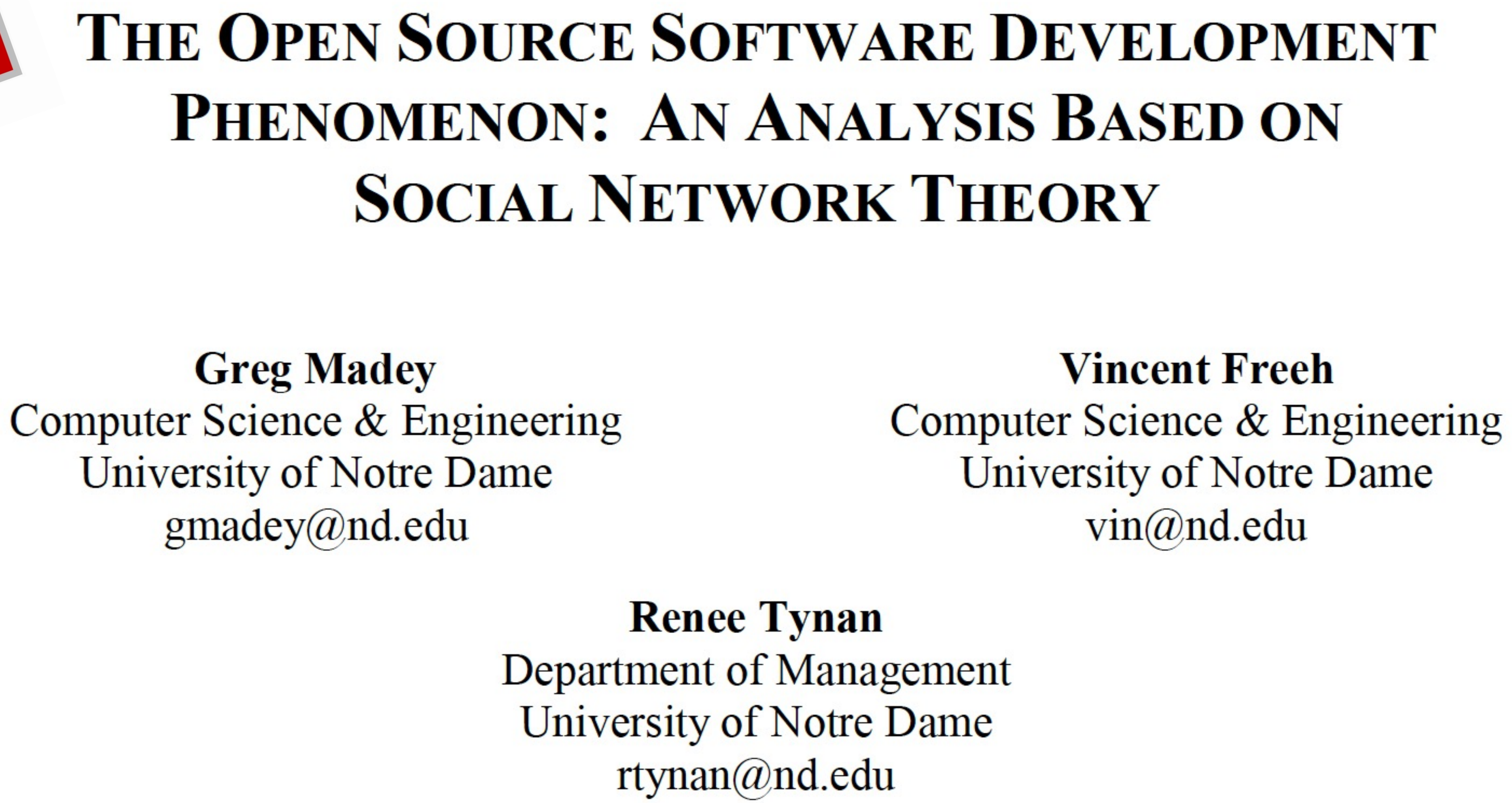
# Latent Social Structure in Open Source Projects

Christian Bird, David Pattison, Raissa D'Souza,
Vladimir Filkov and Premkumar Devanbu
Dept. of Computer Science, Kemper Hall,
University of California, Davis, CA, USA,
cabird,dspattison,rmdsouza,vfilkov,ptdevanbu@ucdavis.edu

EXAMPLE

- email social network
- Apache webserver, Ant, Python, Perl, and PostgreSQL
- ✓ (H1) - Subcommunities of participants will form in the email social networks of large open source projects and the levels of modularity will be statistically significant.
- ✓ (H2) - Social networks constructed from product-related discussions will be more modular than those relating to non-product related discussions or all discussions.
- ✓ (H3) - Pairs of developers within the same subcommunity will have more les in common than pairs of developers from different subcommunities.
- (H4) - The average directory distance between files committed to by developers in the same subcommunity will be less than similar sized groups of developers drawn different subcommunities.

EXAMPLE



# The Open Source Software Development Phenomenon: An Analysis Based on Social Network Theory

**Greg Madey**
Computer Science & Engineering
University of Notre Dame
gmadey@nd.edu

**Vincent Freeh**
Computer Science & Engineering
University of Notre Dame
vin@nd.edu

**Renee Tynan**
Department of Management
University of Notre Dame
rtynan@nd.edu

EXAMPLE

# THE OPEN SOURCE SOFTWARE DEVELOPMENT PHENOMENON: AN ANALYSIS BASED ON SOCIAL NETWORK THEORY

- 33,000 open source developers
- 39,000 open source projects hosted at SOURCEFORGE
- 14 month period from January 2001 through March 2002
- Network 1:
  - each developer is a node in the network;
  - an edge exists between nodes if both developers are on the same project
- Network 2:
  - projects as nodes

EXAMPLE

# THE OPEN SOURCE SOFTWARE DEVELOPMENT PHENOMENON: AN ANALYSIS BASED ON SOCIAL NETWORK THEORY

- Clustering result:

one large cluster (6,862 developers) + the next largest cluster (55 dev) +...

- a heavily skewed distribution
- not a random network

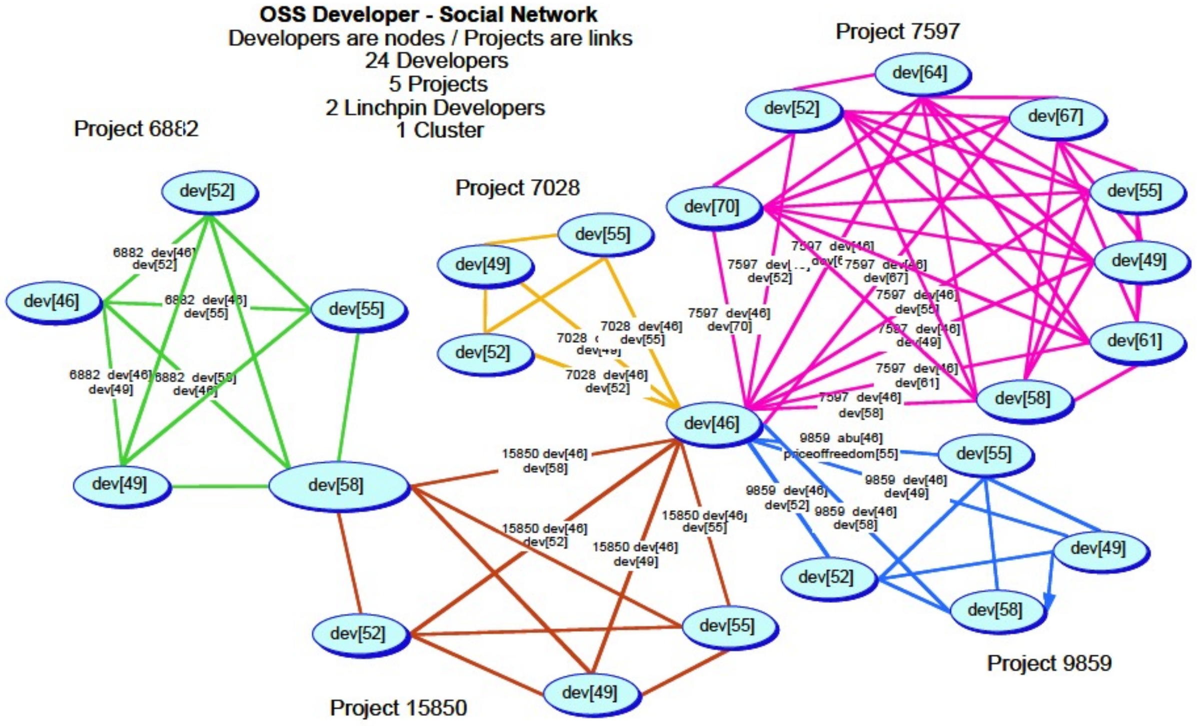
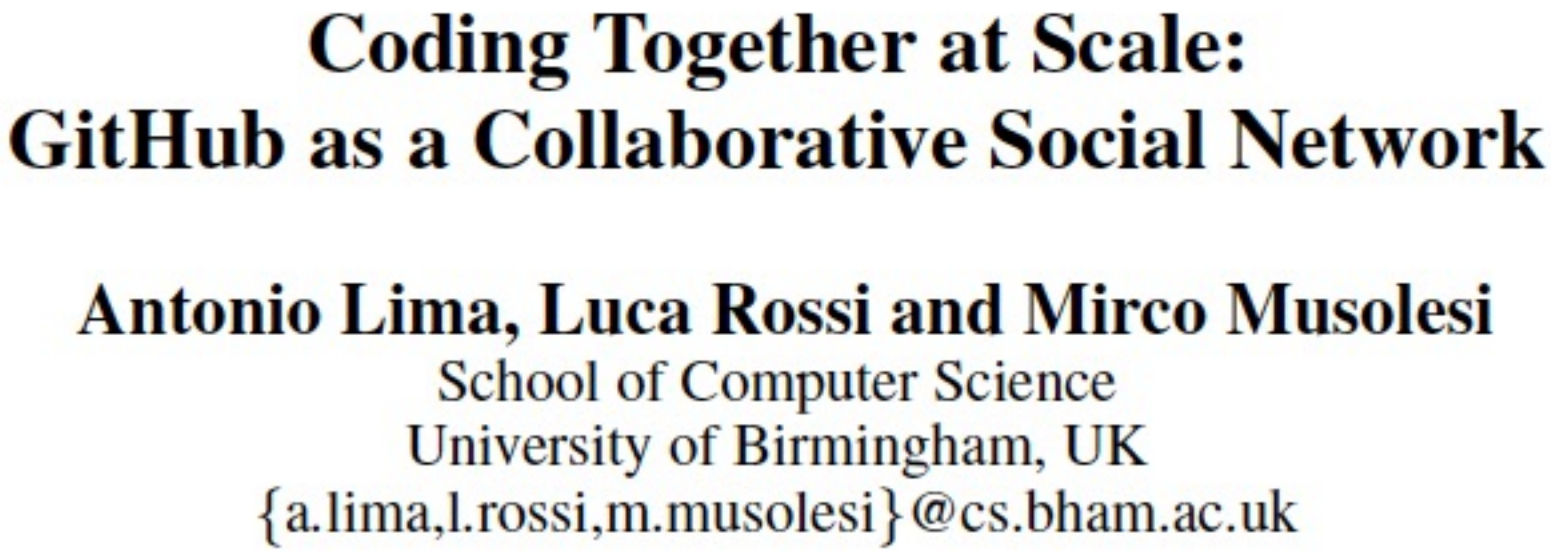


**Figure 1. Developer Social-Network, Linked by Joint Project Membership, Cluster of Size 24**

EXAMPLE

GitHub



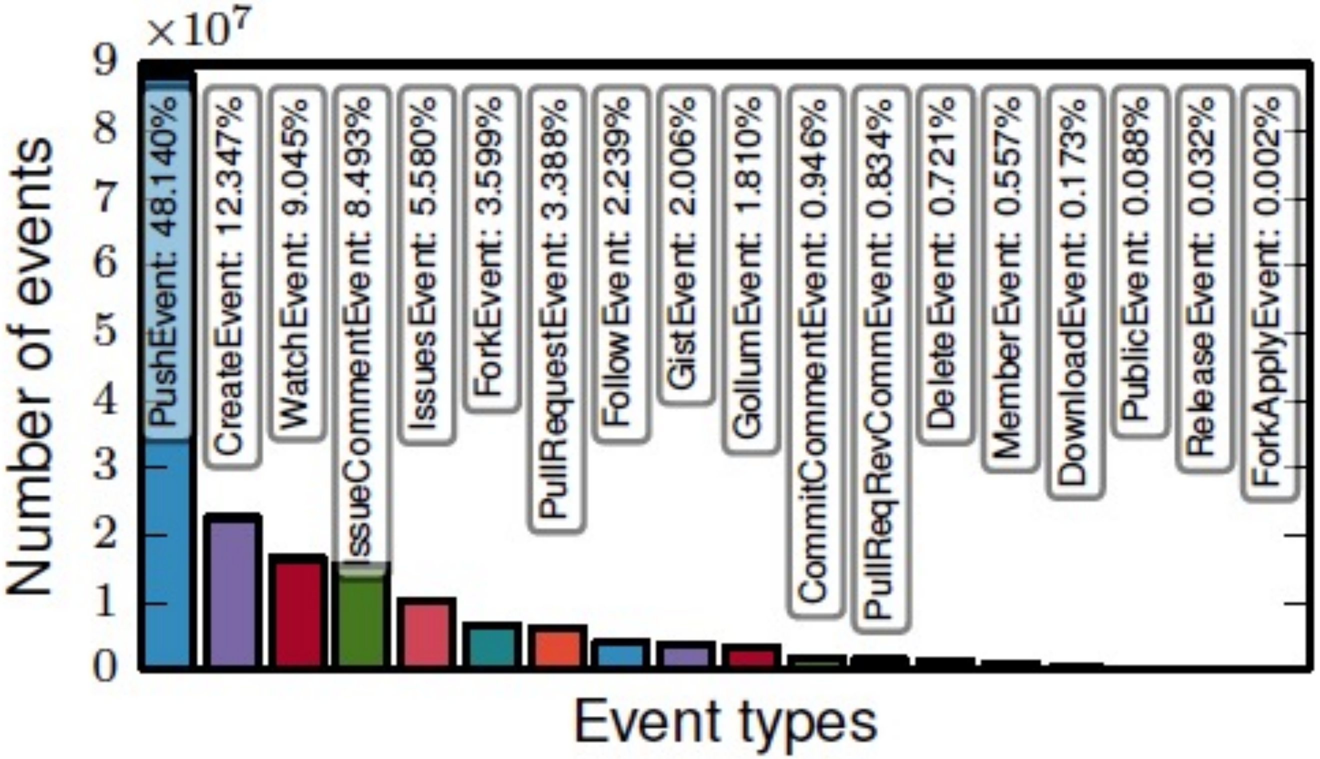
# Coding Together at Scale: GitHub as a Collaborative Social Network

**Antonio Lima, Luca Rossi and Mirco Musolesi**
School of Computer Science
University of Birmingham, UK
{a.lima,l.rossi,m.musolesi}@cs.bham.ac.uk

EXAMPLE

# GitHub as a Collaborative Social Network

- Events (18 months, between March 11, 2012 and September 11, 2013)



GitHub

EXAMPLE

# GitHub as a Collaborative Social Network

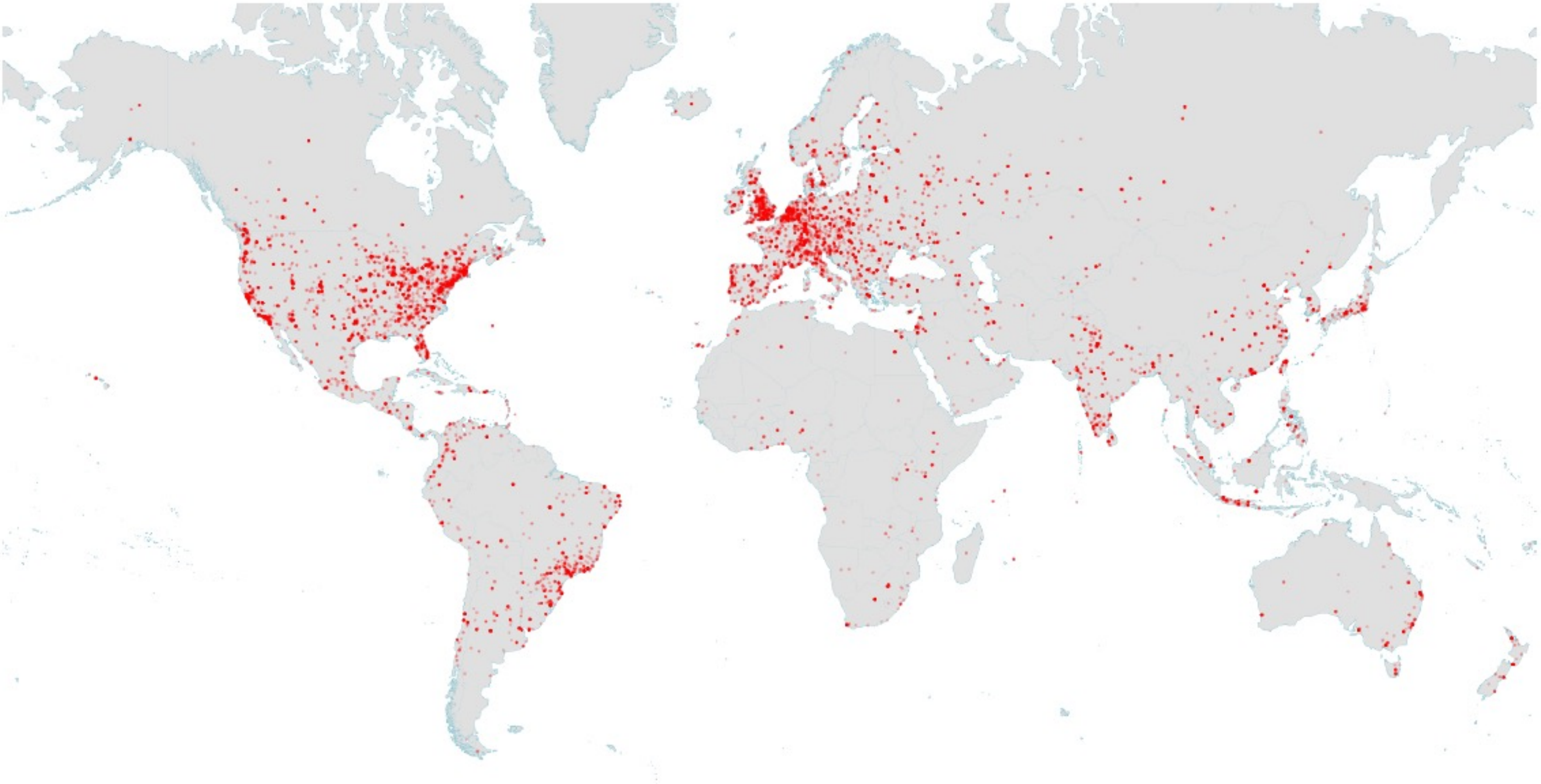
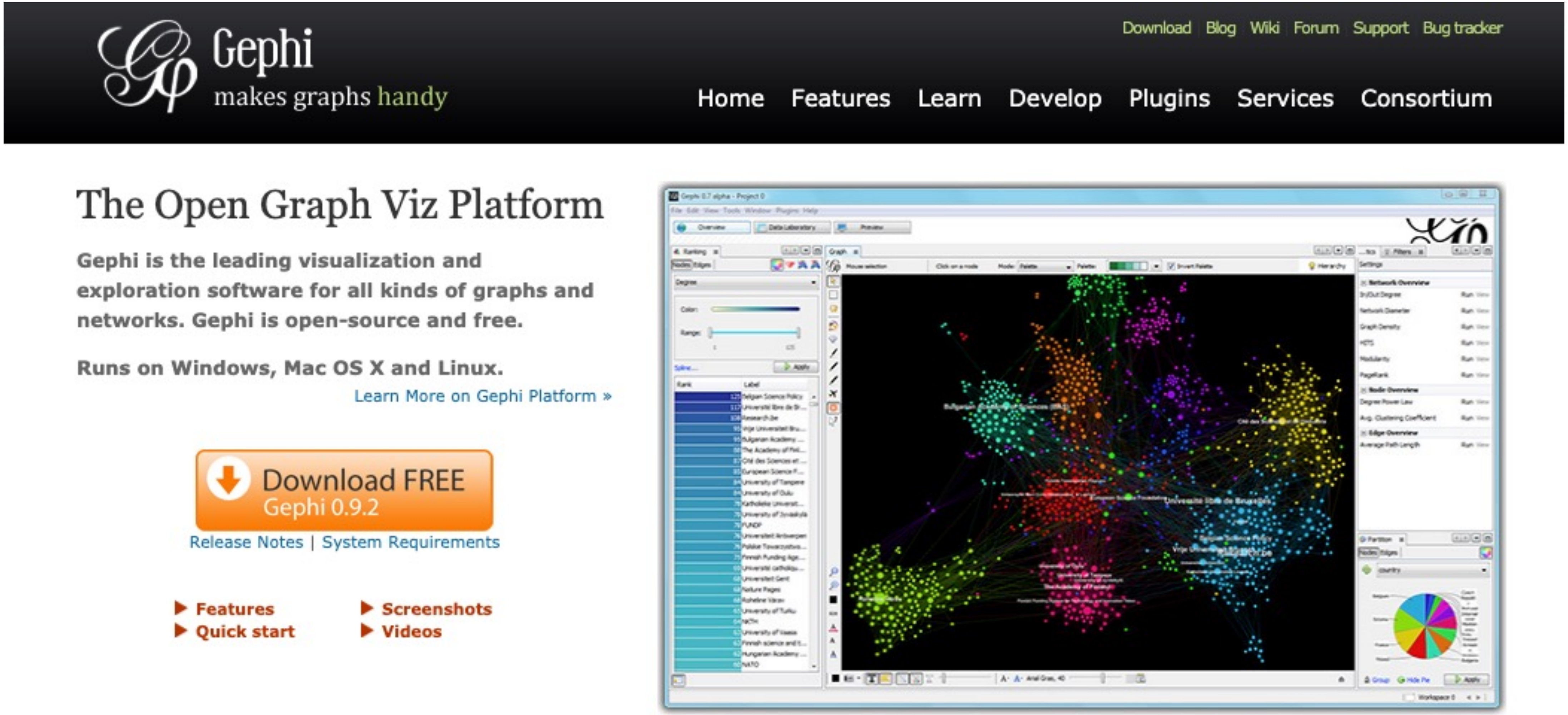


Figure 8: Distribution of GitHub users in the world. For each user, a partially transparent point is drawn on the map. The majority of users is located in North America and in Europe. The leading countries are the United States and the United Kingdom. A 15% random sample of the original distribution was used to make this figure.

# GitHub as a Collaborative Social Network

- followers graph (directed)
- collaborators graph (bipartite, repo – collaborator)
- stargazers graph (bipartite, repo – collaborator)
- contributors graph (push event)
- Observation
  - a very low reciprocity of the social ties
  - collaboration between users happens on a small fraction projects.
  - very active users do not necessarily have a large number of followers
  - impact of geography on collaboration

# Gource.io – software version control visualization



Download | Blog | Wiki | Forum | Support | Bug tracker

Gephi
makes graphs handy

Home Features Learn Develop Plugins Services Consortium

## The Open Graph Viz Platform

**Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.**

**Runs on Windows, Mac OS X and Linux.**

Learn More on Gephi Platform »

Download FREE
Gephi 0.9.2

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos