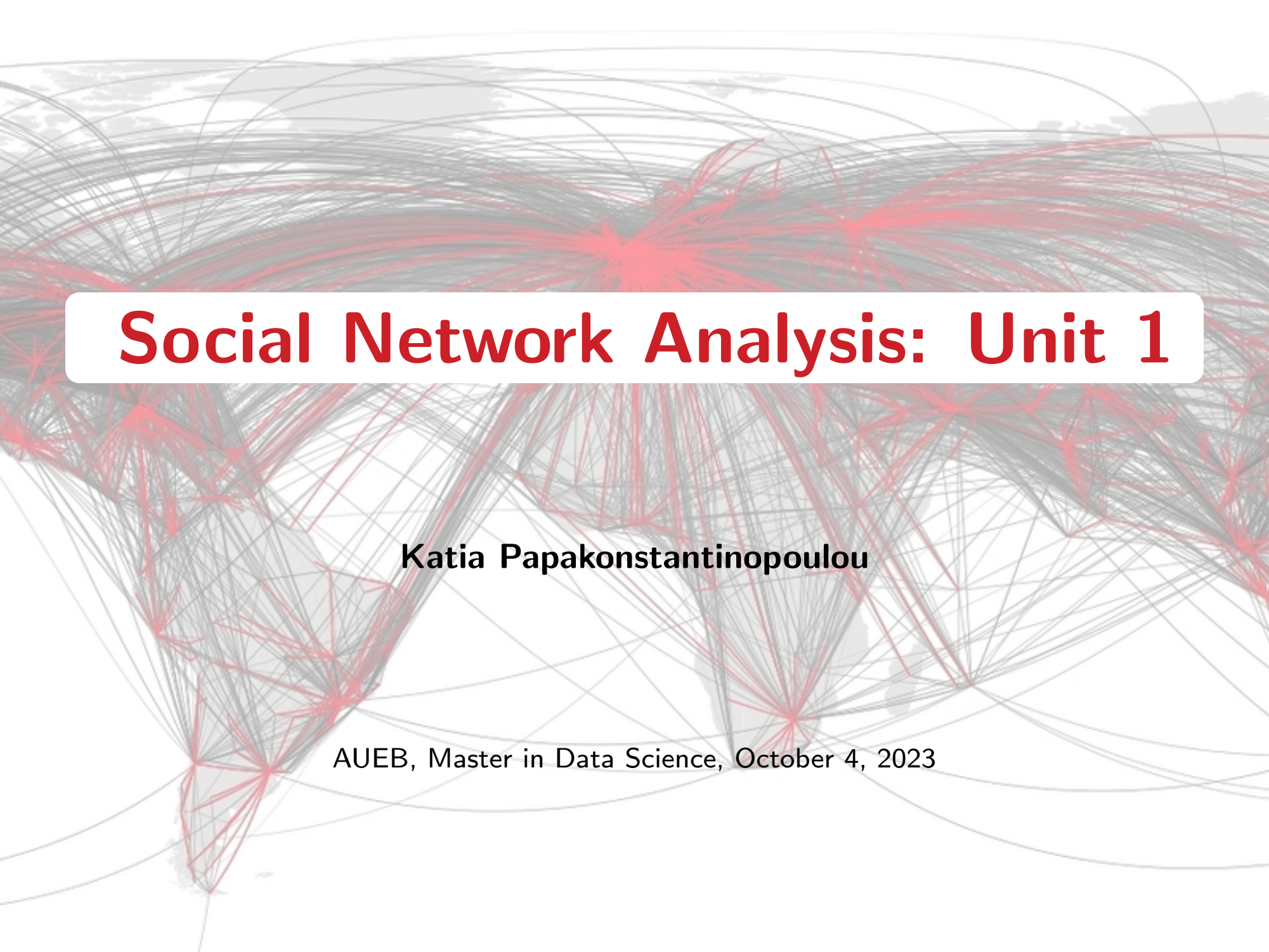


Social Network Analysis: Unit 1

Katia Papakonstantinopoulou

AUEB, Master in Data Science, October 4, 2023

Outline:

- Introduction
- Social networks as graphs
- Network metrics
- Network models
- Ranking in networks
- LAB SESSION: Gephi software

Motivation

A popular social network, e.g. Facebook, has:

- Rich information content $\rightsquigarrow$ (BIG) DATA
- Millions of individuals and organizations who create and use it $\rightsquigarrow$ ALGORITHMS
- Technology that supports it $\rightsquigarrow$ SYSTEMS

Social Network Analysis is motivated by
the interplay of Data, Algorithms and Systems

What is a network?

A network is a set of nodes interconnected by edges denoting relations.

Social networks

A social network is a structure made up of a set of social "actors" and a set of dyadic "ties" and other social interactions between actors.

Social network analysis/mining

Social network analysis (SNA) is the process of investigating social structures through the use of network and graph theories.
It characterizes networked structures in terms of nodes and the ties, edges, or links (relationships or interactions) that connect them.

We want to be able to:

- characterize network structure (identify PROPERTIES)
  - Are nodes connected through the network?
  - How far apart are they?
  - Are some nodes more important due to their position in the network?
  - Is the network composed of communities?
- model network formation (build MODELS)
  - Randomly generated networks
  - Preferential attachment
  - Small-world networks
  - Optimization, strategic network formation
- understand how network structure affects processes (design ALGORITHMS)
  - information diffusion
  - opinion formation
  - coordination/cooperation
  - resilience to attacks

Social network mining tasks with real-world examples

- Identifying influential individuals
- Recommendations of friends or products
- Identifying communities
- Information cascades, Diffusion of innovations
- Distrust/ exploring negative links

What do we hope to achieve from studying networks?

- Identify patterns and statistical properties of network data
- Design principles and models
- Understand why networks are organized the way they are, and thus predict behavior of networked systems

How do we reason about networks?

- Empirical: Study network data to find organizational principles
- Mathematical models: Graph theory and statistical models
- Algorithms for analyzing graphs

What do we study in networks?

- Structure and evolution
- Processes and dynamics

Social Networks as Graphs

Network elements: nodes, edges

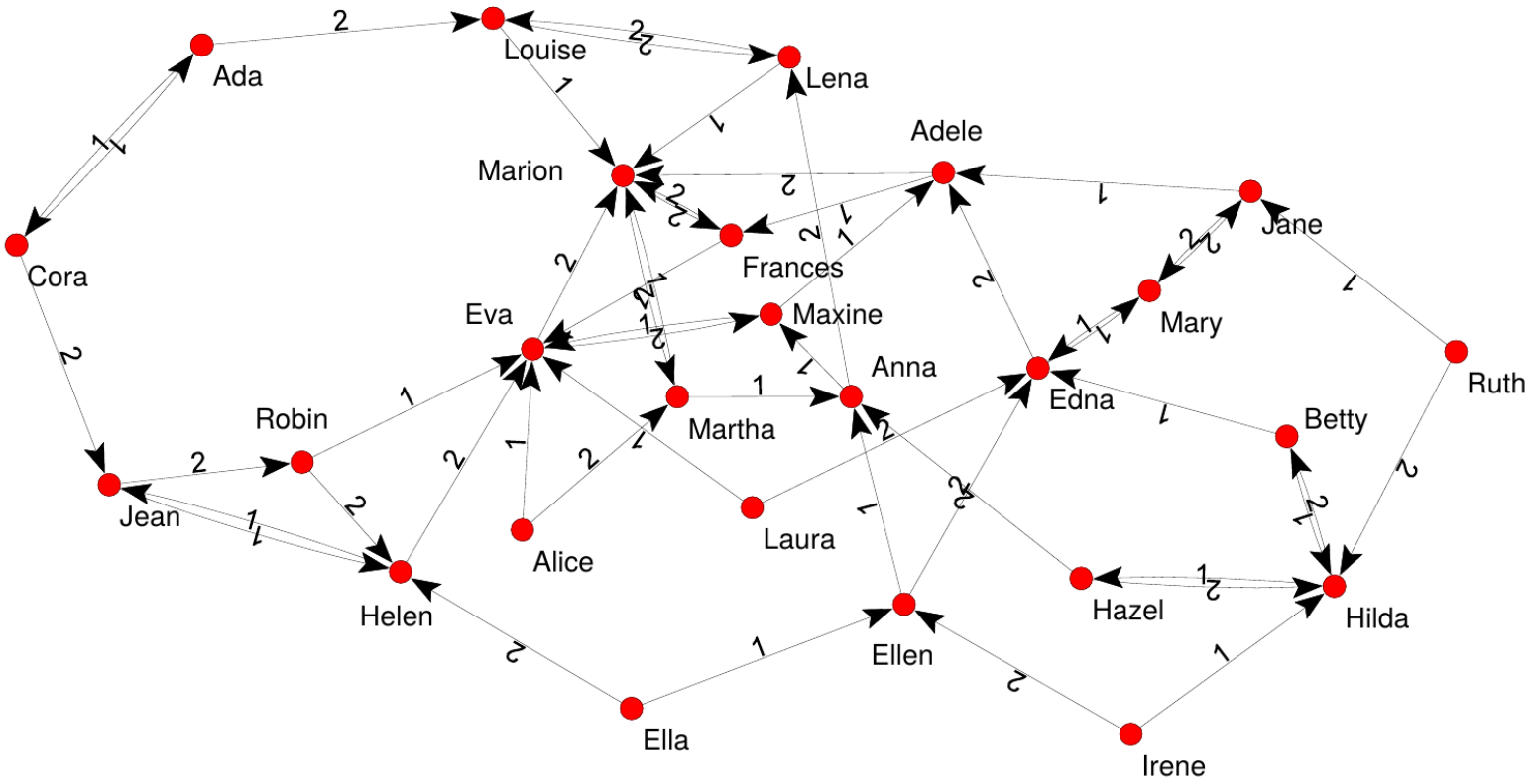
Edges denote social ties between nodes and may be:

- Directed (also called arcs, links)
  - denoted by: $A \longrightarrow B$
  - express, for example: "A likes B", "A gave a gift to B", "A is B's child", "A follows B"
- Undirected
  - denoted by: $A \longleftrightarrow B$
  - express, for example: "A and B like each other", "A and B are siblings", "A and B are co-authors"

Edges may have attributes, expressing for example:

- weight (e.g. frequency of communication)
- ranking (best friend, second best friend...)
- type (friend, relative, co-worker)
- properties depending on the structure of the rest of the graph (e.g. betweenness)

Social Networks as Graphs – Example



A graph representing a girls' school dormitory dining-table partners, 1st and 2nd choices (Moreno, The sociometry reader, 1960)

Network representation

We can represent a network using:

- the *adjacency matrix*, or
- the *edgelist*, or
- the *adjacency list*

of the corresponding graph.

Examples on the board

Network metrics

Various metrics are used to characterize networks, summarizing complex structures and behaviors.

We are interested in structural metrics not only because they summarize huge networks, but also because network structure may relate to user behavior.

Network metrics express:

- Connections: Homophily, Network Closure
- Distributions: Centrality, Density, Distance, Structural holes, Tie strength
- Segmentation: Clustering coefficient, Cohesion

Node-based metrics

Node network properties can be derived:

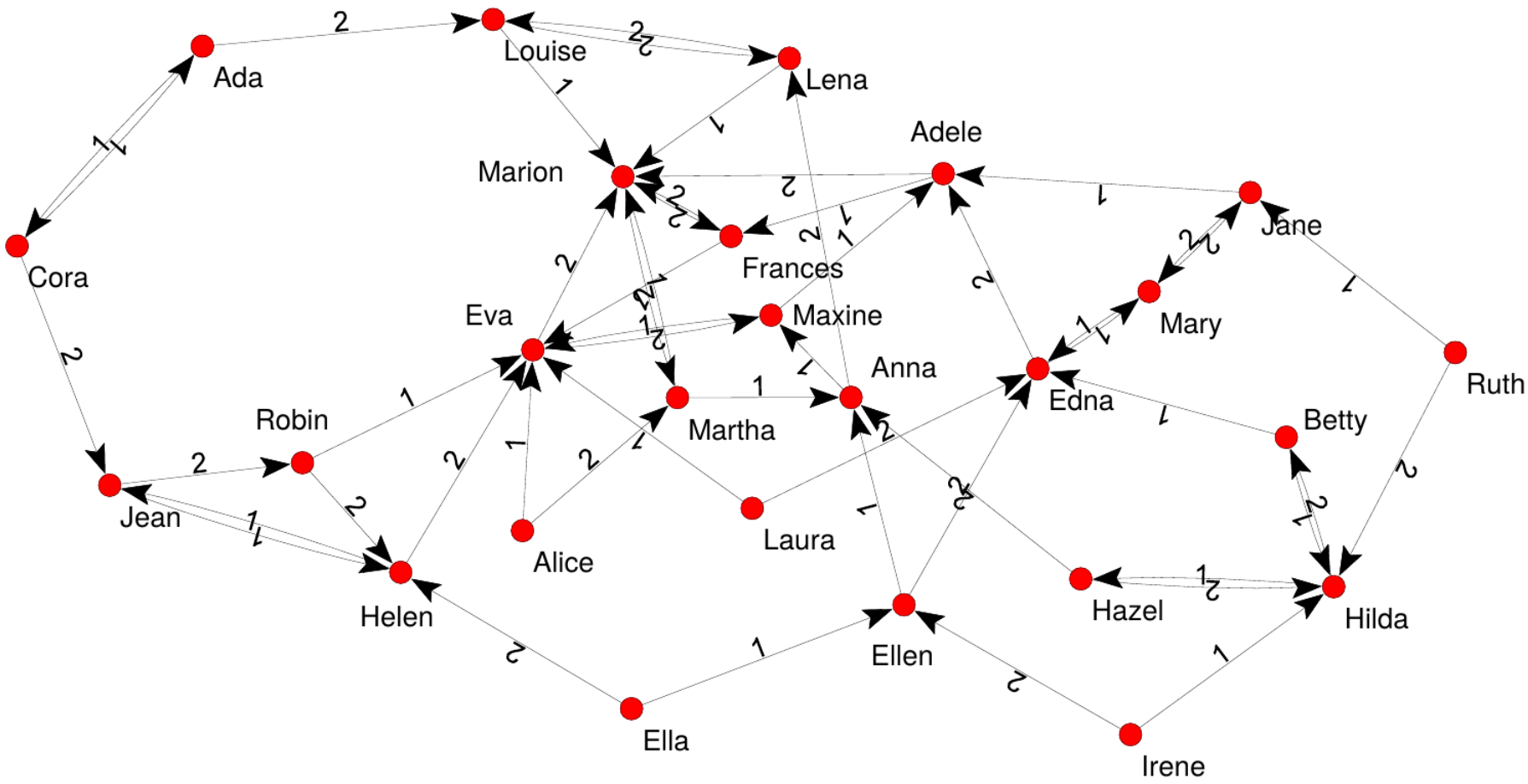
- from immediate connections
  - indegree: how many directed edges (arcs) are incident on a node
  - outdegree: how many directed edges (arcs) originate at a node
  - degree (in or out): number of edges incident on a node
- from the entire graph
  - centrality (betweenness, closeness)

Examples on the board

We will next compute the following metrics:

- degree & degree distribution
- connected components

Degree



Which node has the largest indegree?

Degree distribution

Degree distribution

The degree distribution of the nodes in a network is the frequency count of the occurrence of each node degree.

Example on the board

Diameter

Diameter

The diameter of a network is the distance between the two most distant nodes in the network.

Distance

The distance between two nodes in the network is the length of the shortest path that connects these nodes.

Example on the board

Connected components

Connected component

A connected component is a set of nodes in a graph that are linked to each other by paths.

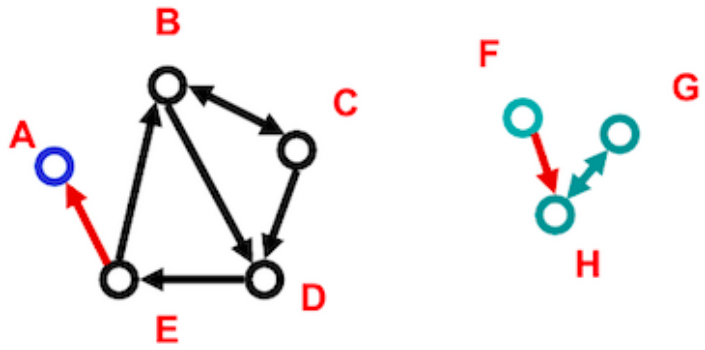
- Within a *strongly connected* component: Each node can be reached from every other node in the component by following *directed* links
- Within a *weakly connected* component: Each node can be reached from every other node in the component by following links ignoring their direction

Notes:

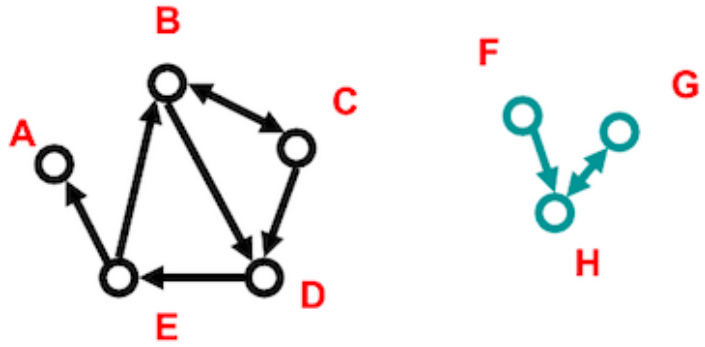
- If the largest component encompasses a significant fraction of the graph, it is called the *giant* component.
- In undirected networks one talks simply about *connected* components.

Connected components – Example

Consider the graph consisted by the set of nodes {A, B, C, D, E, F, G, H} and edges {EA, EB, BD, DE, BC, CB, CD, FH, GH, HG}.



Strongly connected components: {B,C,D,E}, {A}, {G,H}, {F}



Weakly connected components: {A,B,C,D,E}, {G,H,F}

Why do we need network models?

They allow us to:

- have simple representations of complex networks
- derive properties mathematically
- predict properties and outcomes
- test algorithms
- measure of comparison (in what ways is the real-world network different from the model? what insights can be extracted from it?)

Distinctive features of social networks

- "small-world" effect (Pool and Kochen, 1978, Milgram, 1967)
- clustering (Watts and Strogatz, 1998)
- skewed degree distribution (Albert, Barabási, Jeong, 1999)

Random graph models/ graph generators

Why do we have many graph models?

Different random graph models capture different network properties and produce different probability distributions on graphs.

- Erdös-Rényi
- Preferential Attachment
- Small-World
- Power Law
- Stochastic Block Model
- Kronecker Graphs

Erdös-Rényi: The simplest network model

Assumptions

- nodes connect at random
- network is undirected

Key parameters: p or m

- p: probability that any two nodes share an edge
- m: total number of edges in the graph

Model

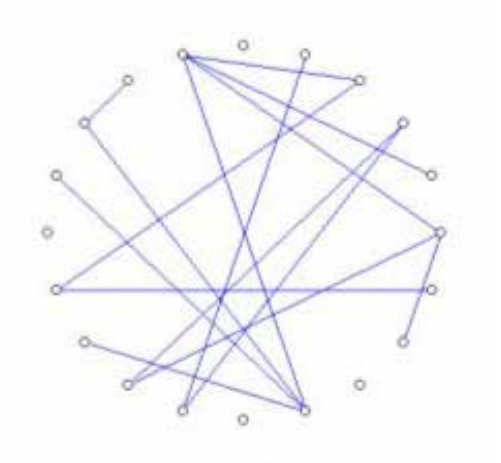
(n,p)-model: For each potential edge we flip a biased coin

- with probability p we add the edge
- with probability $1 - p$ we don't

Erdös-Rényi: Example

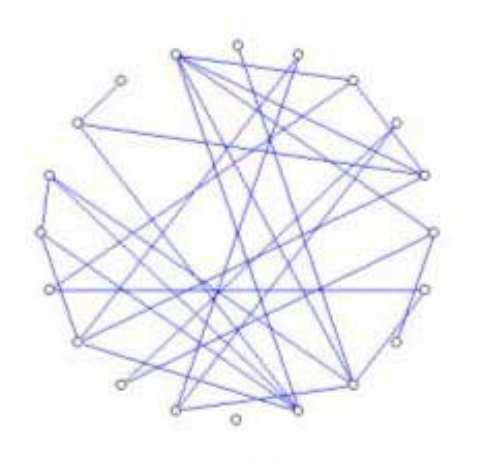
$p = 0$

(a)



$p = 0.1$

(b)



$p = 0.2$

(c)

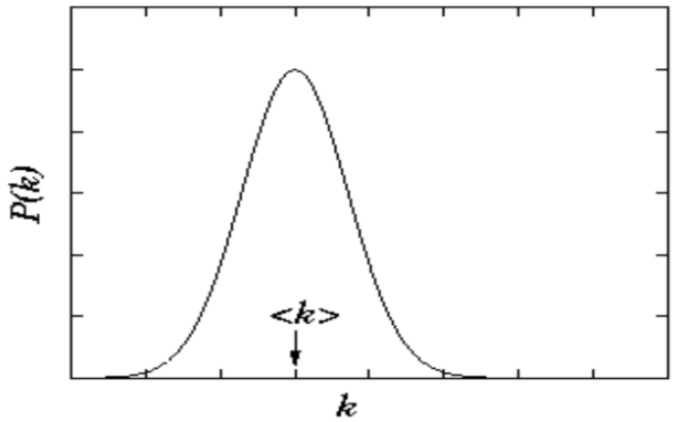
The graph becomes denser as p increases.

Erdös-Rényi: Properties

Degree distribution

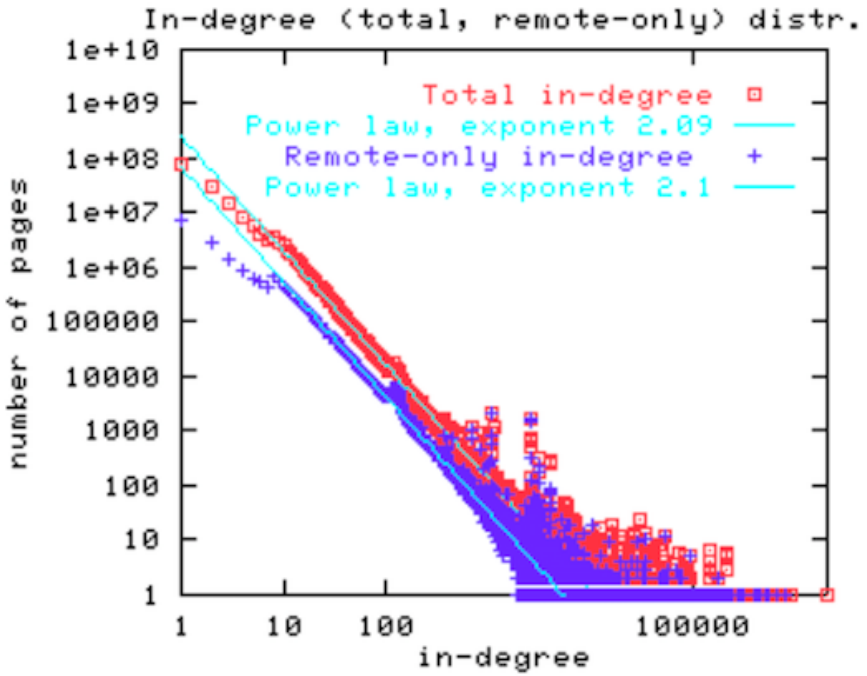
What is the probability that a node has 0, 1, 2, 3, . . . adjacent edges? (The probabilities sum to 1.)

The degrees in the graph follow the Poisson distribution:



Intuition: We don't expect large hubs in the network!

Graph models: Power Law – Observation



The fraction of Web pages that have k in-links is approximately proportional to $1/k^2$.

Graph models: Rich-get-richer

Model

- Pages are created in order, and named 1, 2, 3, . . . , N
- When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule
  - With probability p, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to this page i.
  - With probability $1 - p$, page j instead chooses a page i uniformly at random from among all earlier pages, and creates a link to the page that i points to.

Graph models: Preferential Attachment

With probability $1 - p$, page j chooses a page l with probability proportional to l's current number of in-links, and creates a link to l.

Graph models: Small World

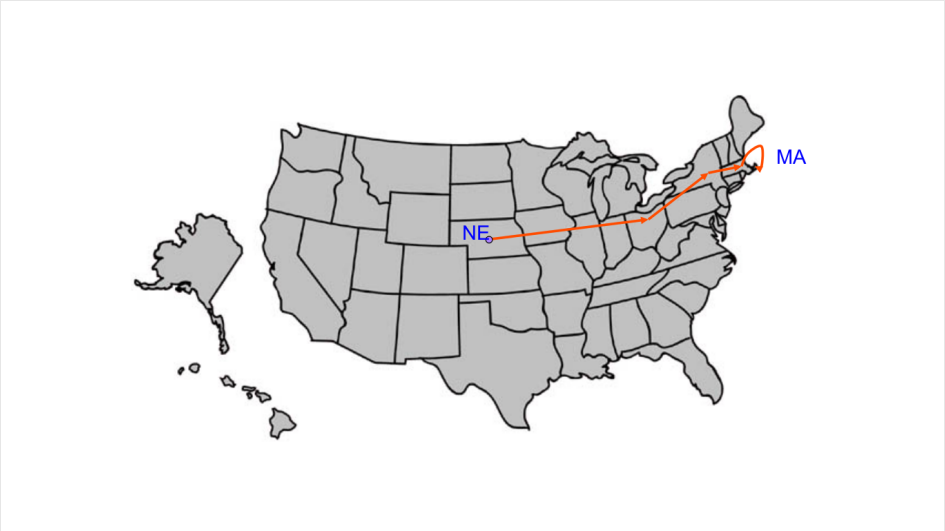
Outline:

- Small world phenomenon (Milgram's small world experiment)
- Local structure
- Small world network models
- Small world networks: why do they arise?

Small world phenomenon: Milgram's experiment

Instructions

Given a target individual (stockbroker in Boston), pass the message to a person you correspond with who is "closest" to the target.



Outcome

- 20% of initiated chains reached target
- average chain length = 6.5, aka: *"Six degrees of separation"*

Small worlds and networking

What does it mean to be 1, 2, 3 hops apart on Facebook, Twitter, LinkedIn, Google Plus?

Observations on small world networks

Small world networks exhibit:

- high clustering (i.e., my friends' friends tend to be my friends)
- low average shortest path

Generating small world graphs: Watts-Strogatz model

Watts-Strogatz model

Select a fraction p of edges
Reposition on of their endpoints

Add a fraction p of additional edges leaving underlying lattice intact

The model disallows self-edges and multiple edges.

How do search engines determine how to rank web pages?

Using automated methods that look at the Web link structure.

2 Approaches:

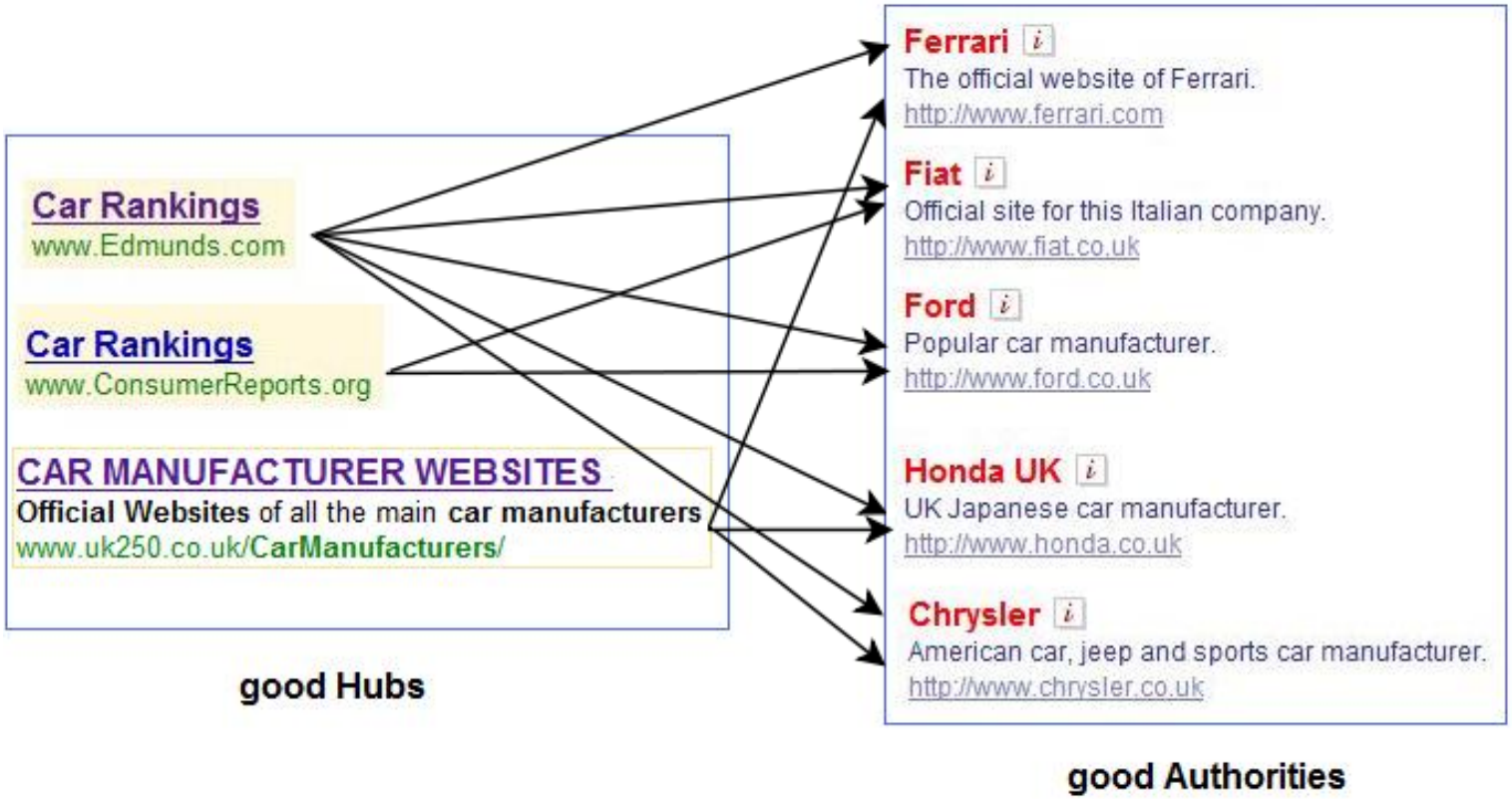
- Using Hubs and Authorities
- Using PageRank

Link Analysis Using Hubs and Authorities

Approach:

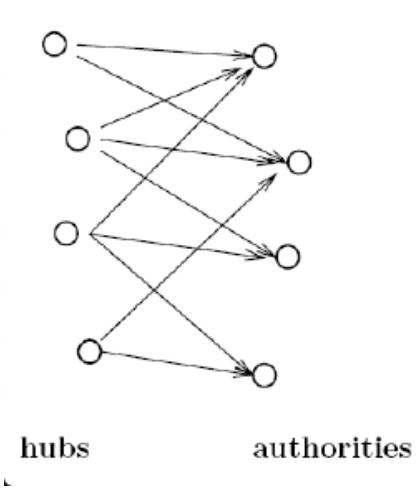
- Voting by In-Links (i.e., interpreting indegree as number of votes).
- List-Finding (a page's value as a list of resources relevant to some topic, is equal to the sum of the votes received by all pages for which it voted).
- Applying the principle of Repeated Improvement (giving each page's vote a weight equal to its value as a list).
- Hubs and Authorities (Authorities: the kinds of pages we were originally seeking, Hubs: the high-value lists).

Hubs and Authorities



Query: Top automobile makers

Hubs and Authorities



Hubs and Authorities

- Authority Update Rule: For each page p, update $auth(p)$ to be the sum of the hub scores of all pages that point to it.
- Hub Update Rule: For each page p, update $hub(p)$ to be the sum of the authority scores of all pages that it points to.

The normalized values converge to limits as the number of steps of the Repeated Improvement goes to infinity (the results stabilize so that continued improvement leads to smaller and smaller changes in the values we observe).

These equilibrium values reflect a balance between hubs and authorities: The authority score of page i is proportional to the hub scores of the pages that point to i, and the hub score of i is proportional to the authority scores of the pages i points to.

PageRank

In various settings on the WWW, endorsement passes directly from one prominent page to another (i.e., a page is important if it is cited by other important pages)

Most popular measure of importance: PageRank

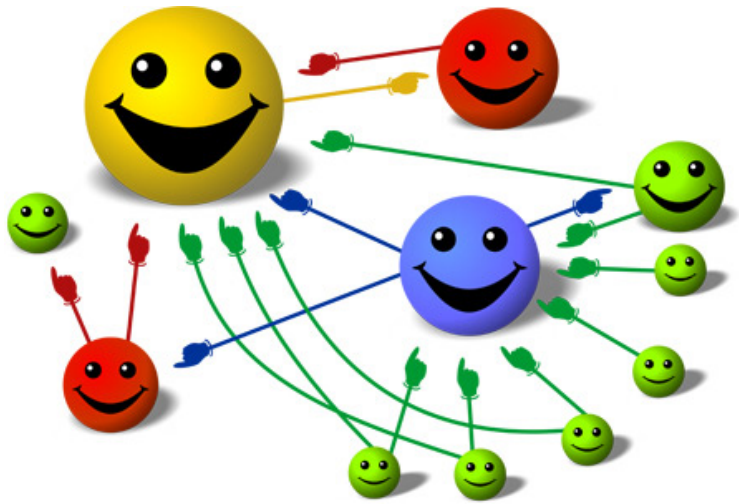
- The PageRank of each web page expresses the probability that a random web-surfer reaches this page.
- Intuition behind PageRank: We start with simple voting based on in-links, and refine it using the Principle of Repeated Improvement.

Formal description of PageRank

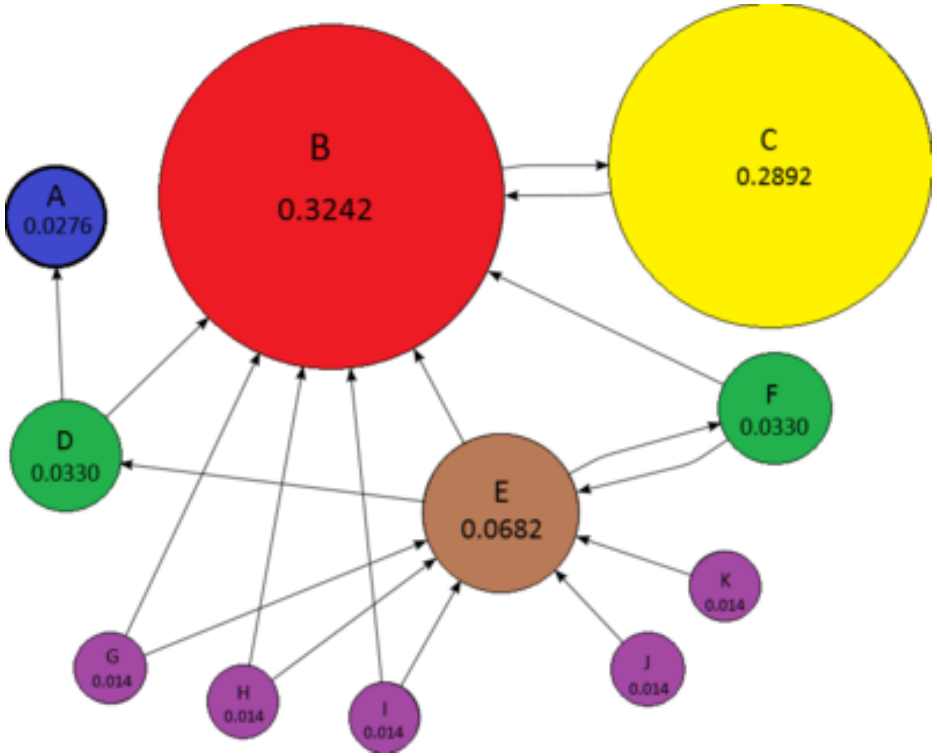
PageRank of node i in a graph of n nodes is defined as

$$\pi_i = \frac{1-c}{n} + c \sum_{k:k\to i} \frac{\pi_k}{d_k},$$

where c is the damping factor (usually set around 0.85).



PageRank


A
0.0276
B
0.3242
C
0.2892
D
0.0330
E
0.0682
F
0.0330
G
0.014
H
0.014
I
0.014
J
0.014
K
0.014

LAB SESSION: Gephi

Gephi: graph visualization and exploration software

- Presentation of main functionality
- Practice

Unit 1: Recap

We discussed the following aspects:

- Introduction
- Social networks as graphs
- Network metrics
- Network models
- Ranking in networks
- LAB SESSION: Gephi software