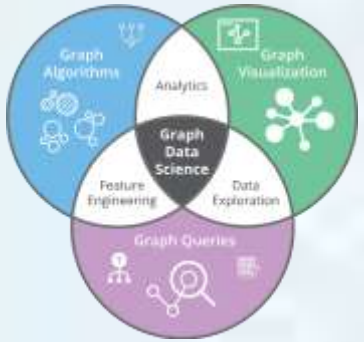


Habib University
shaping futures

CS343 Graph Data Science

Spring 2024

Introduction

Muhammad Qasim Pasta
qasim.pasta@sse.habib.edu.pk

Slides are intended to be filled during the lectures. Certain details are intentionally omitted for in-class discussions. These slides are not meant to be used as reading material.

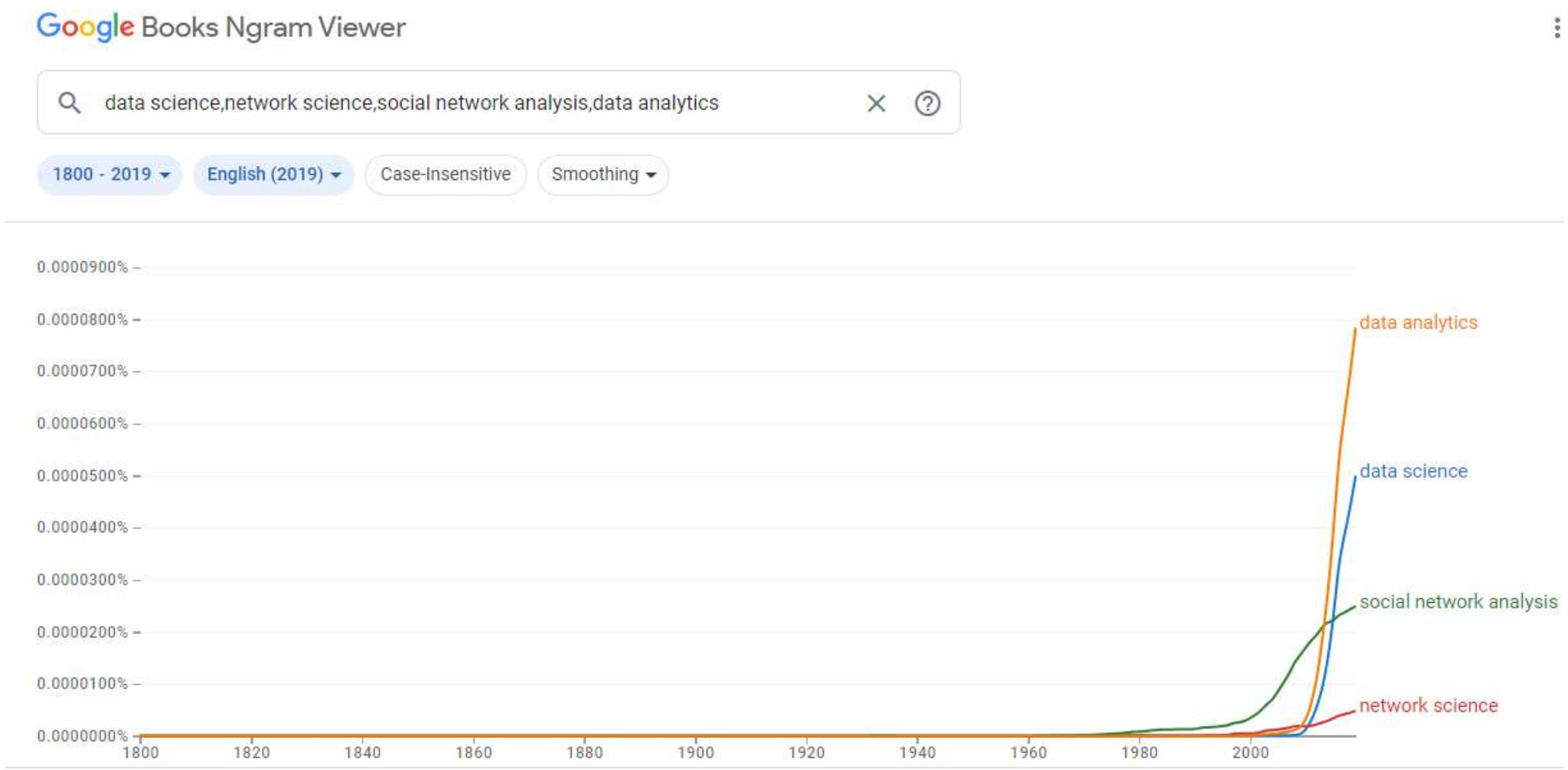
Greater Data Science

David Leigh Donho, professor of statistics at Stanford, defined following activities as part of Data Science in his paper “50 Years of Data Science”, published in 2015

- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data (Automation)
- Data Visualization and Presentation
- Data Modelling
  - Generative Modelling
  - Predictive Modelling
- Science About Data Science

Image Source: WikiPedia

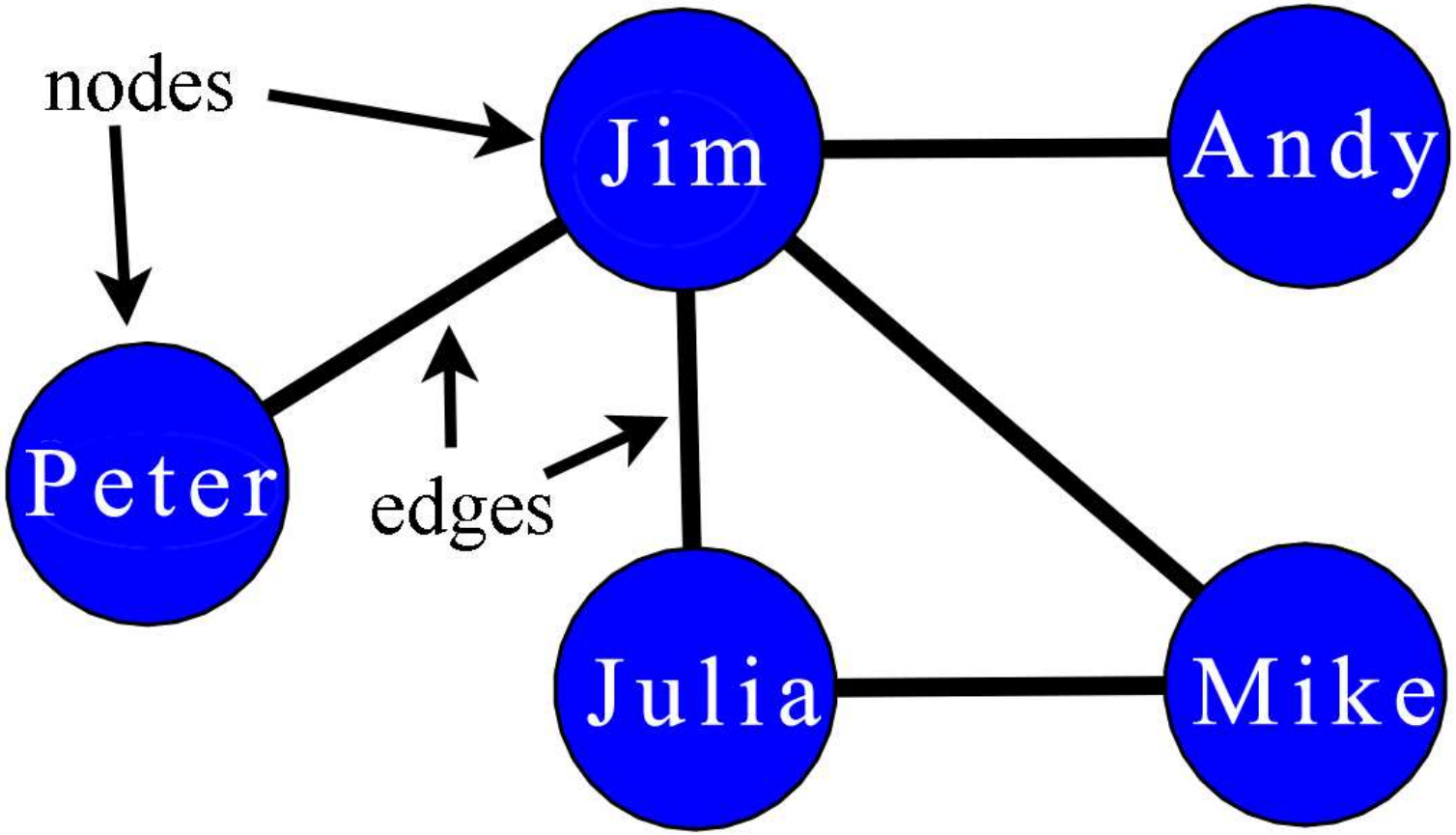
Evolution of Network Science



What is a network?

- A network is an abstract representation to model pairwise relations between objects from a certain collection.
- A network can be of tangible objects
- also possible to define a network of entities that defined in abstract space

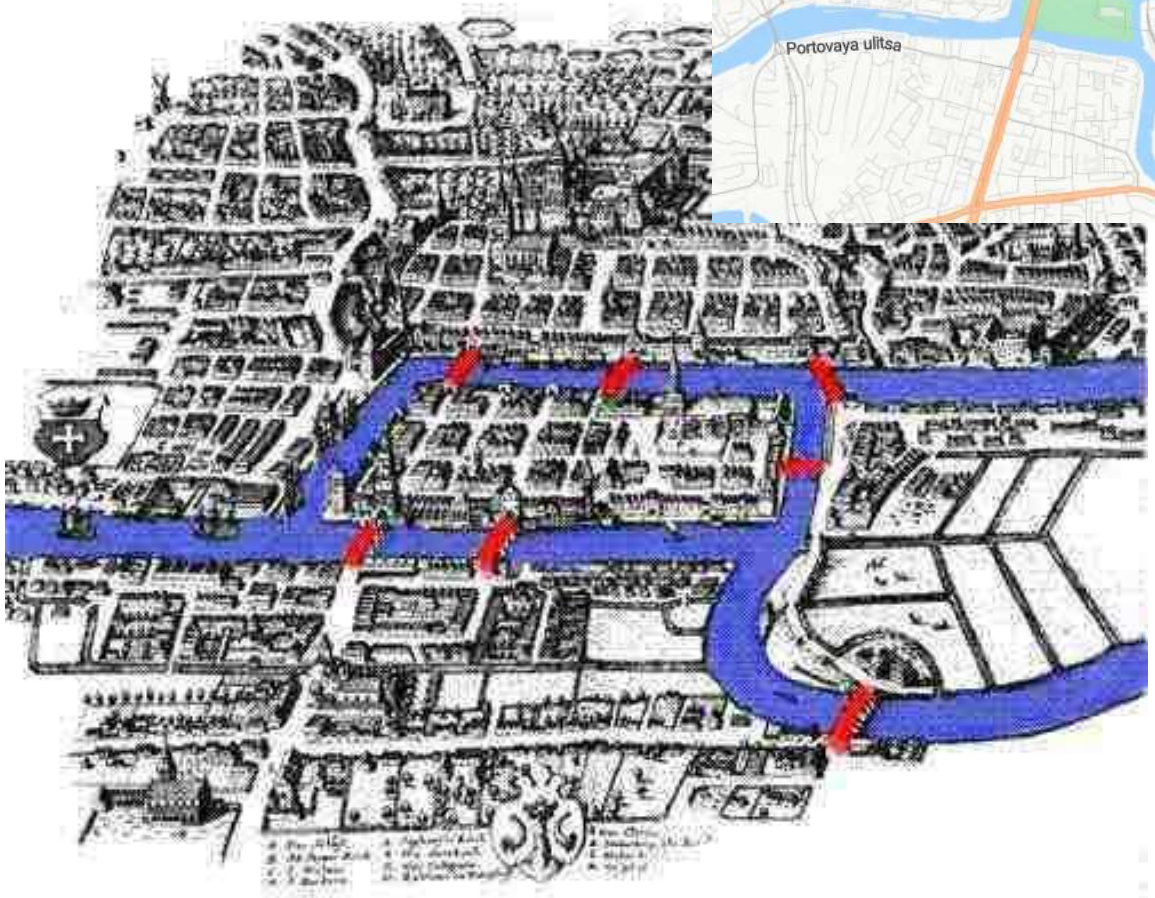
A Network?



Mathematical Structure called Graph

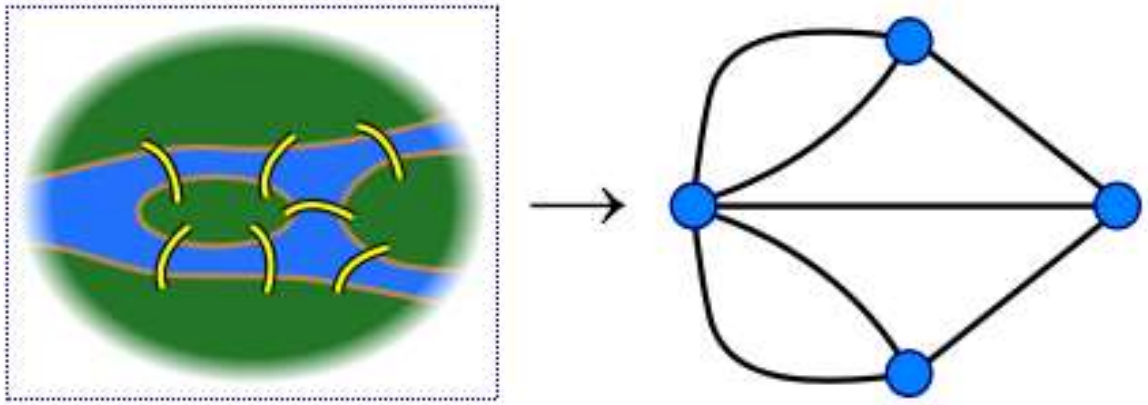
Konigsberg (Kaliningrad, Russia) bridge problem

- if the seven bridges of the city of Königsberg over the river Preger
- can all be traversed in a single trip without doubling back?
- additional requirement that the trip ends in the same place it began
- Give birth to graph theory



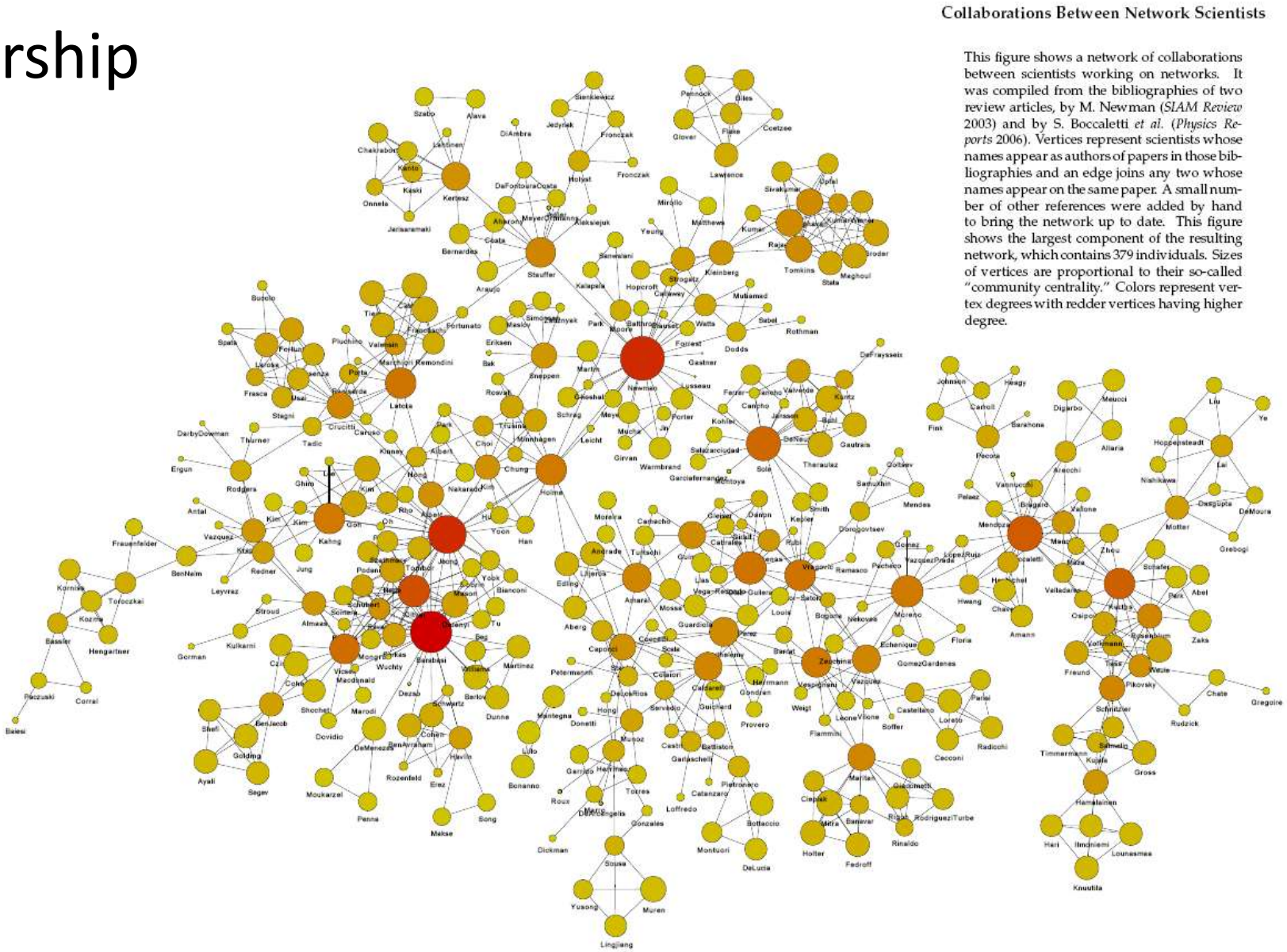
Solution as graph!

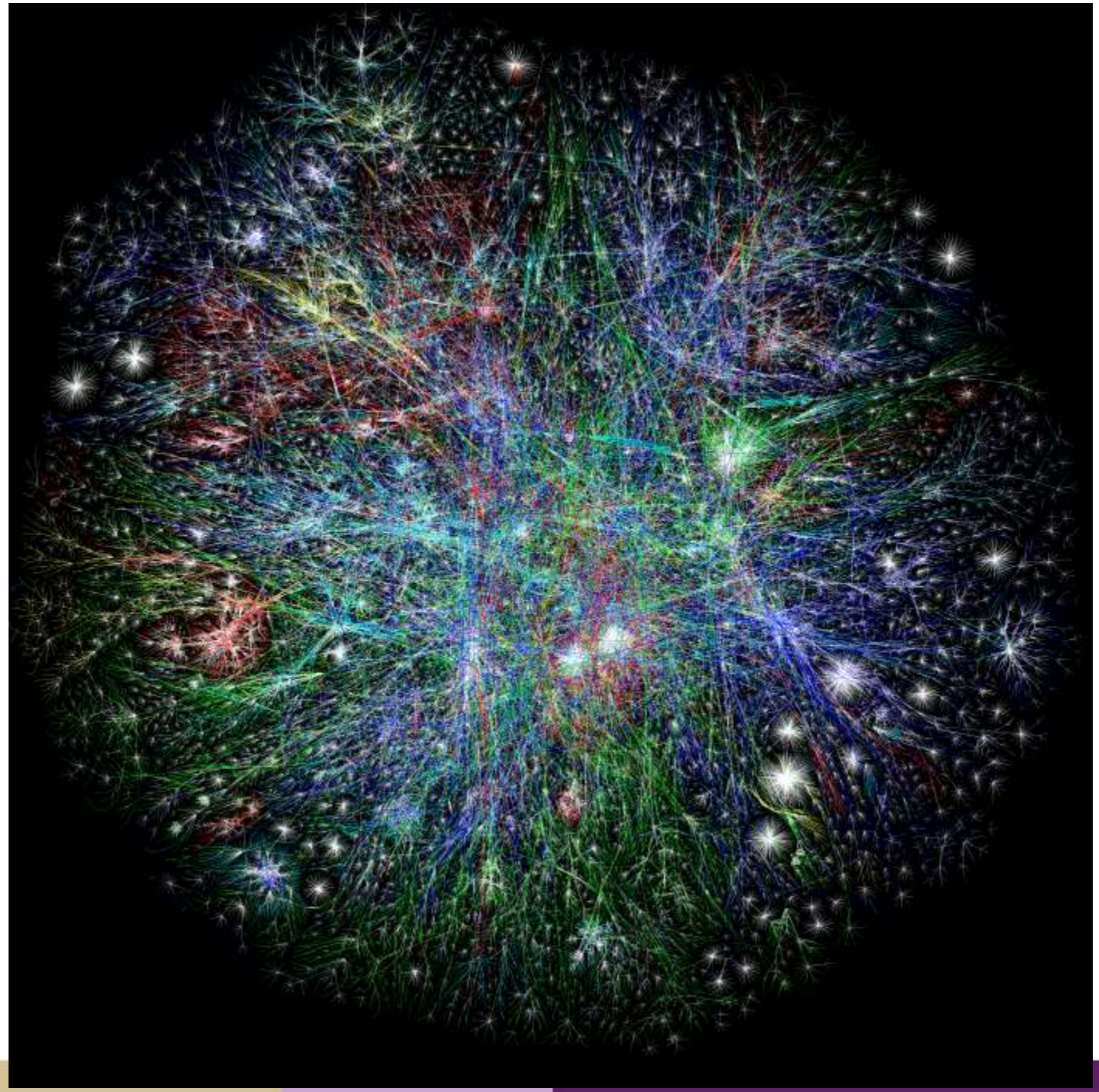
- Swiss mathematician - Leonhard Euler
- considering each piece of island as dot
- each bridge connecting two island as a line between two dots
- a graph of dots (vertices or nodes) and lines (edges or links).



What else can we represent using Graphs?

Co-Authorship





Internet Routing Paths

5 million edges

Graph Colors:
Asia Pacific - Red
Europe/Middle East/Central Asia/Africa - Green
North America - Blue
Latin American and Caribbean - Yellow
RFC1918 IP Addresses - Cyan
Unknown - White

http://www.opte.org/maps/

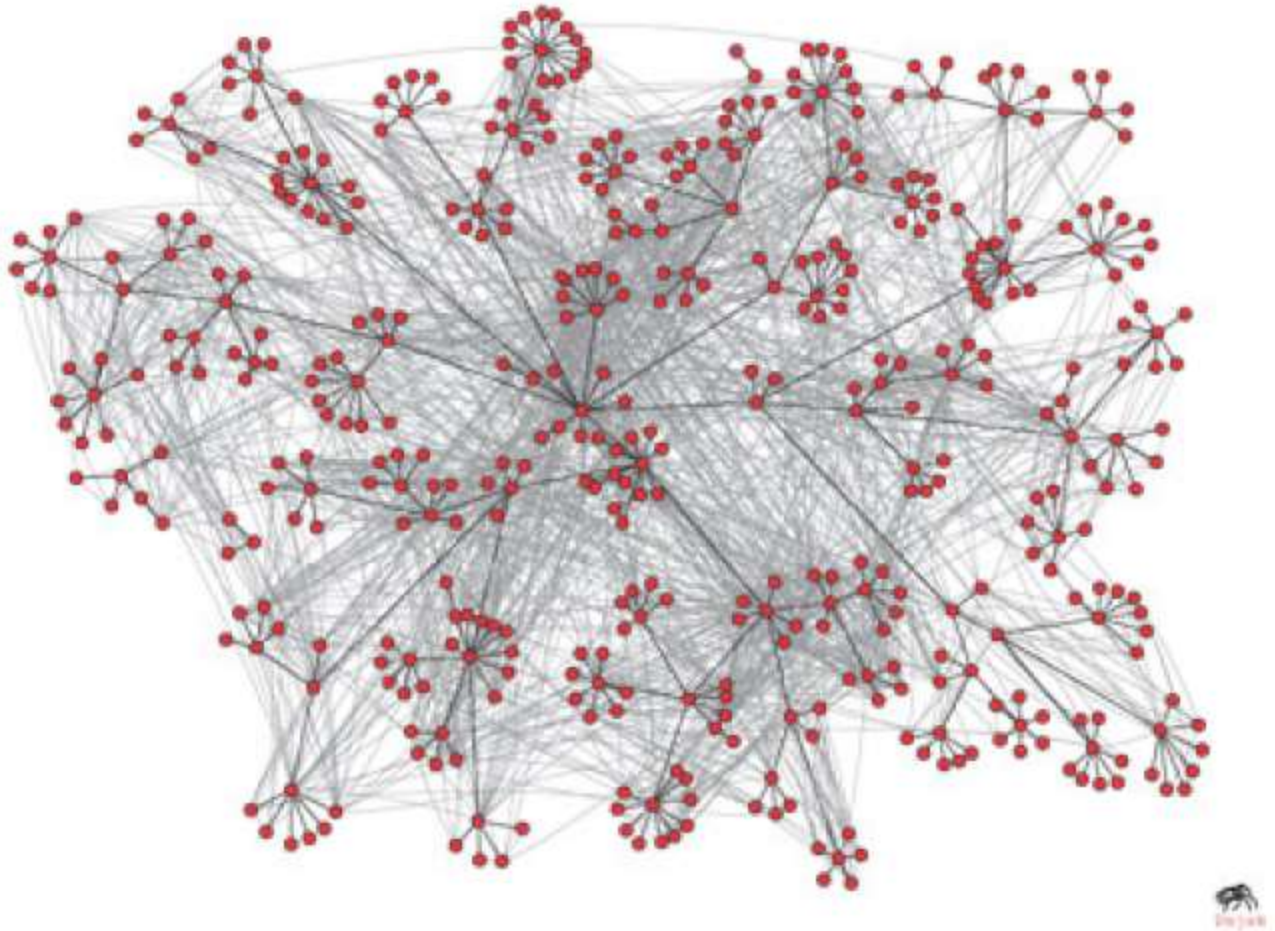
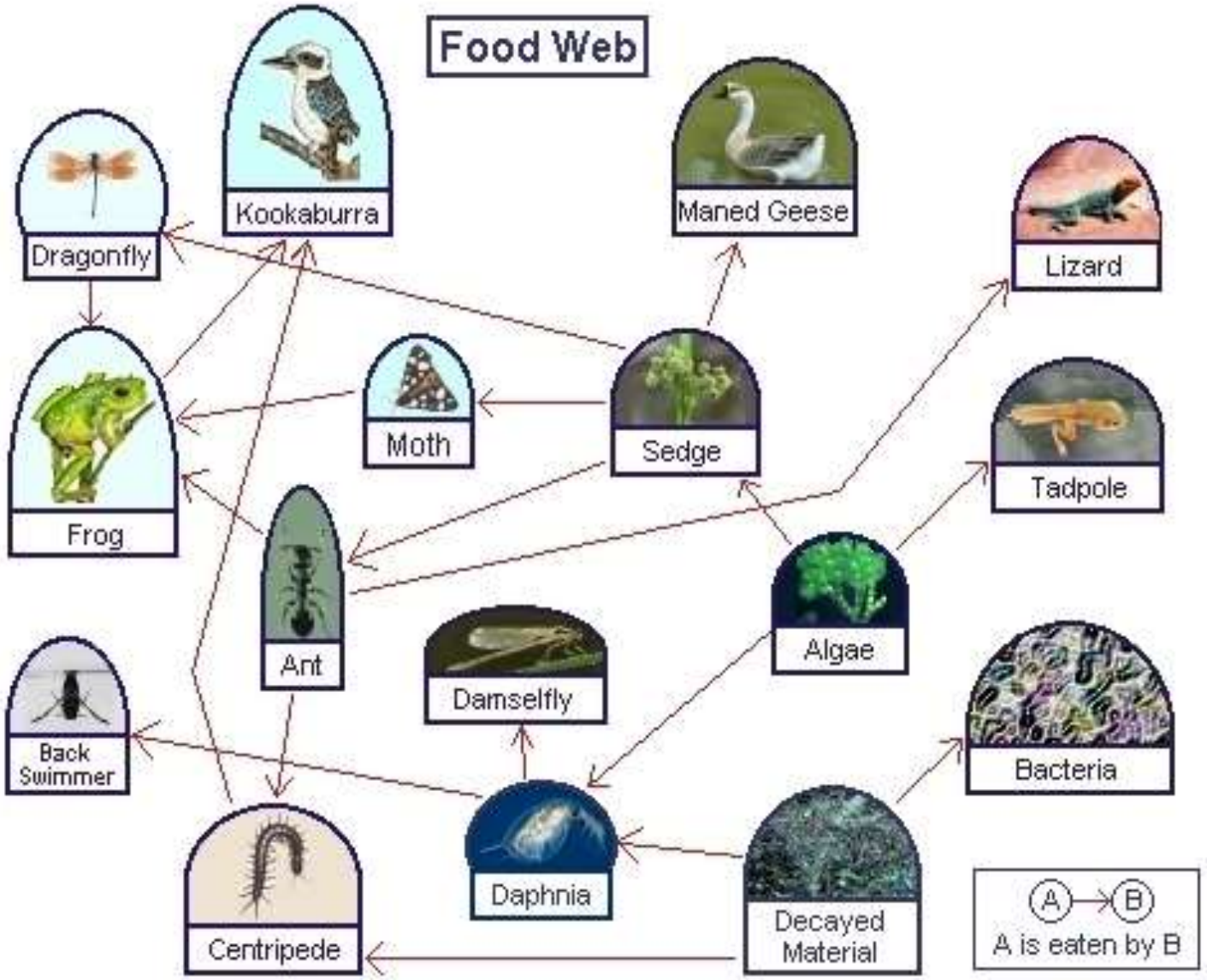
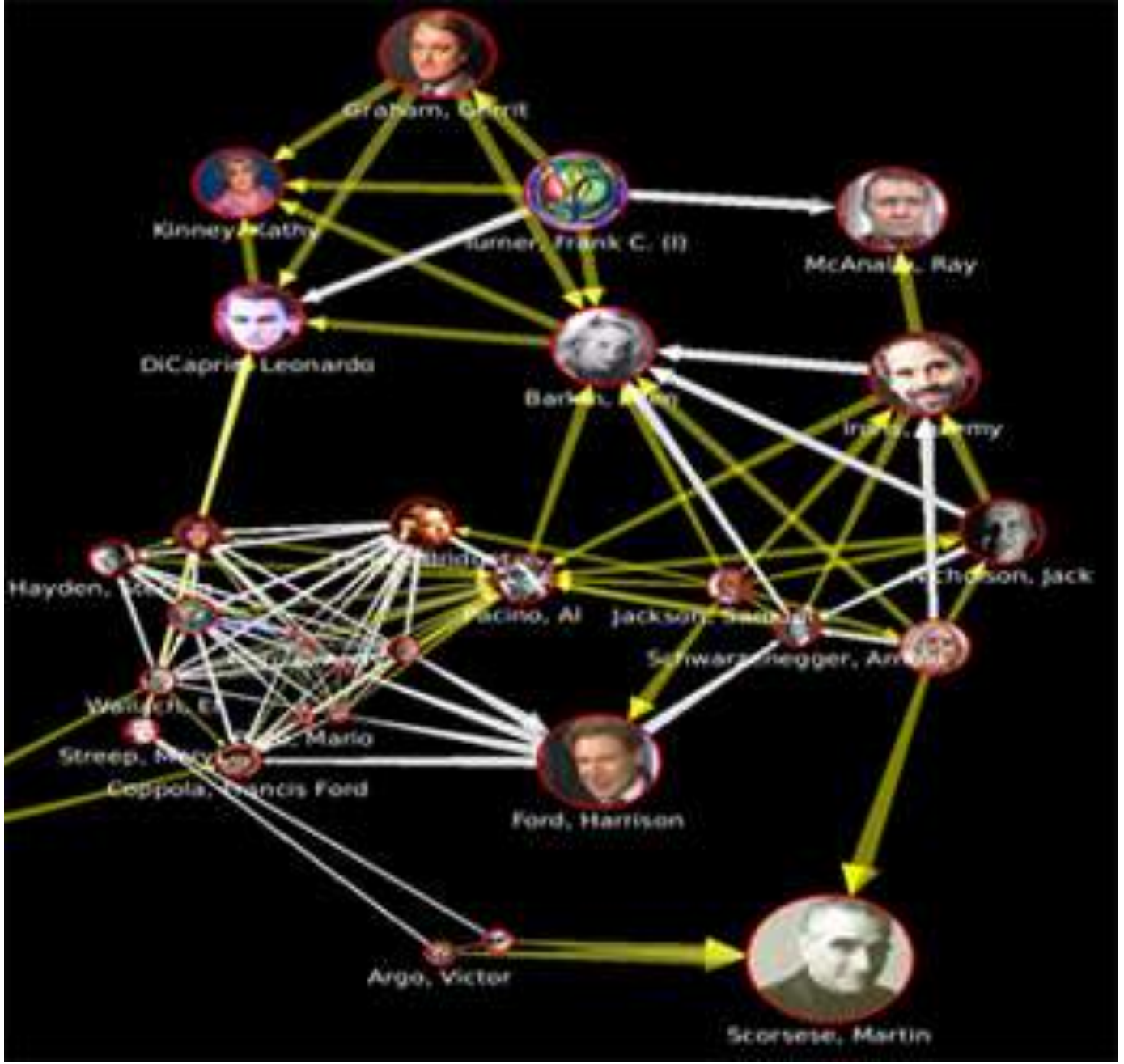


Figure 1.2: Social networks based on communication and interaction can also be constructed from the traces left by on-line data. In this case, the pattern of e-mail communication among 436 employees of Hewlett Packard Research Lab is superimposed on the official organizational hierarchy [6]. (Image from http://www-personal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

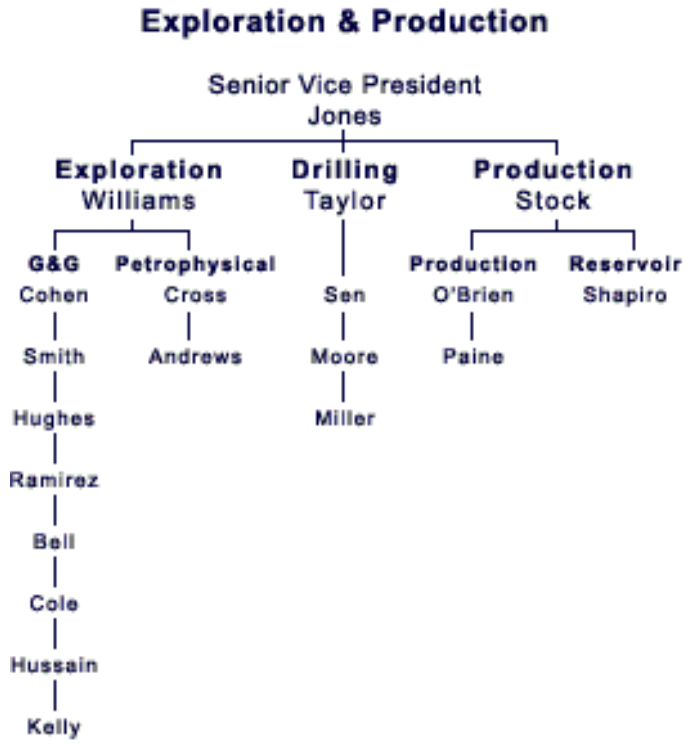

Food Web
Kookaburra
Maned Geese
Dragonfly
Lizard
Moth
Sedge
Frog
Tadpole
Ant
Algae
Damselfly
Back Swimmer
Bacteria
Daphnia
Centripede
Decayed Material
A
B
A is eaten by B



Movie Actors

*Case Study - Network in Organization

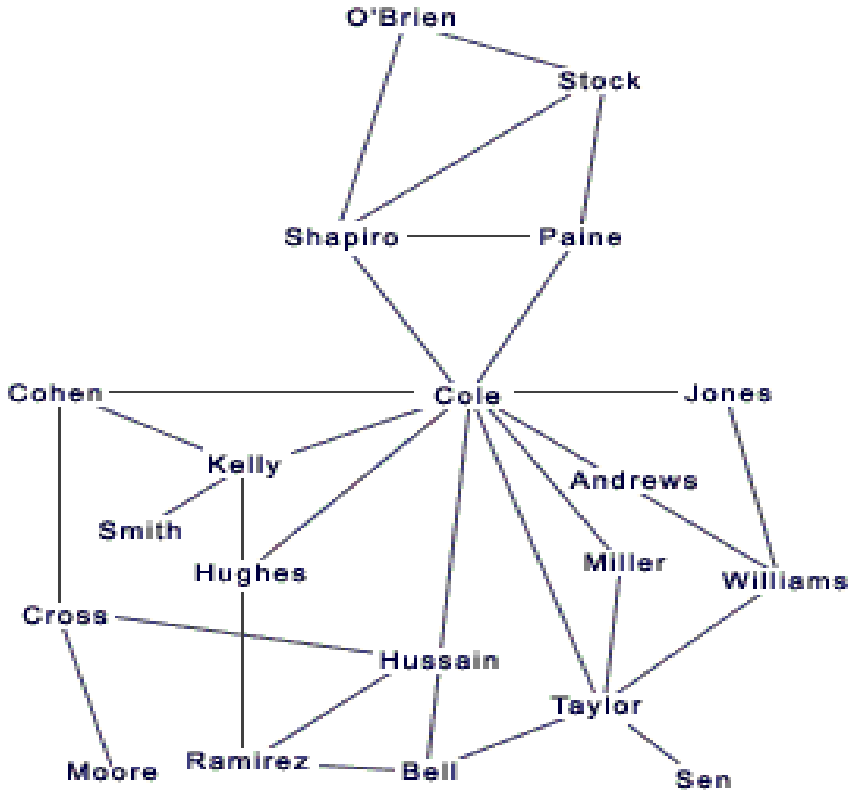
- Consider the given Organogram of an organization
- What information can we extract from this?



Source: http://www.robcross.org/network_ona.htm

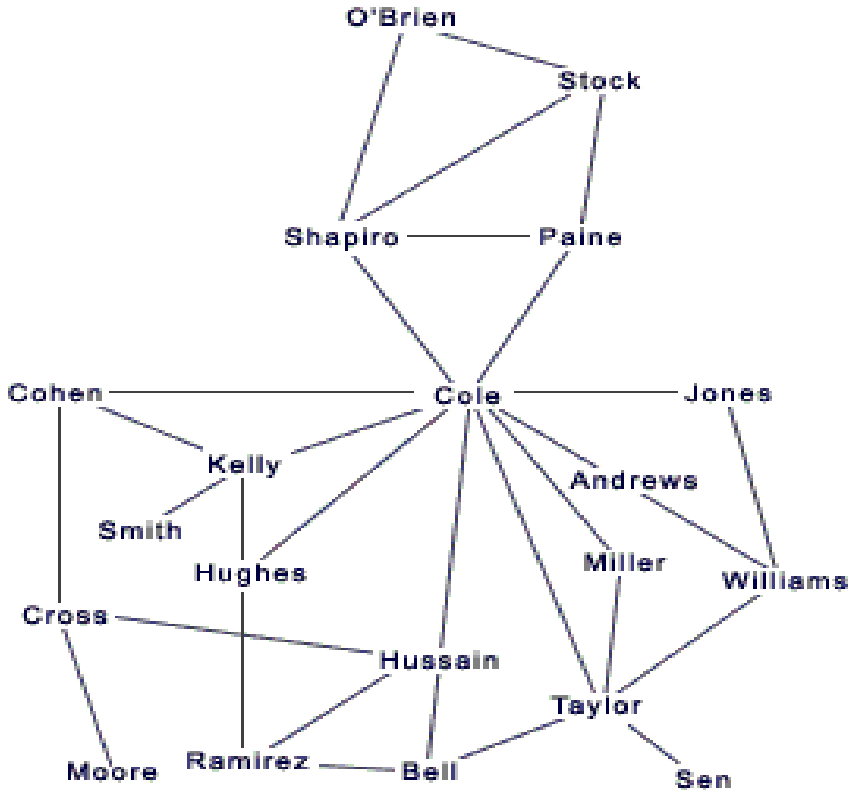
Case Study - Network in Organization

- Consider this informal network
- How did we come up with this network?

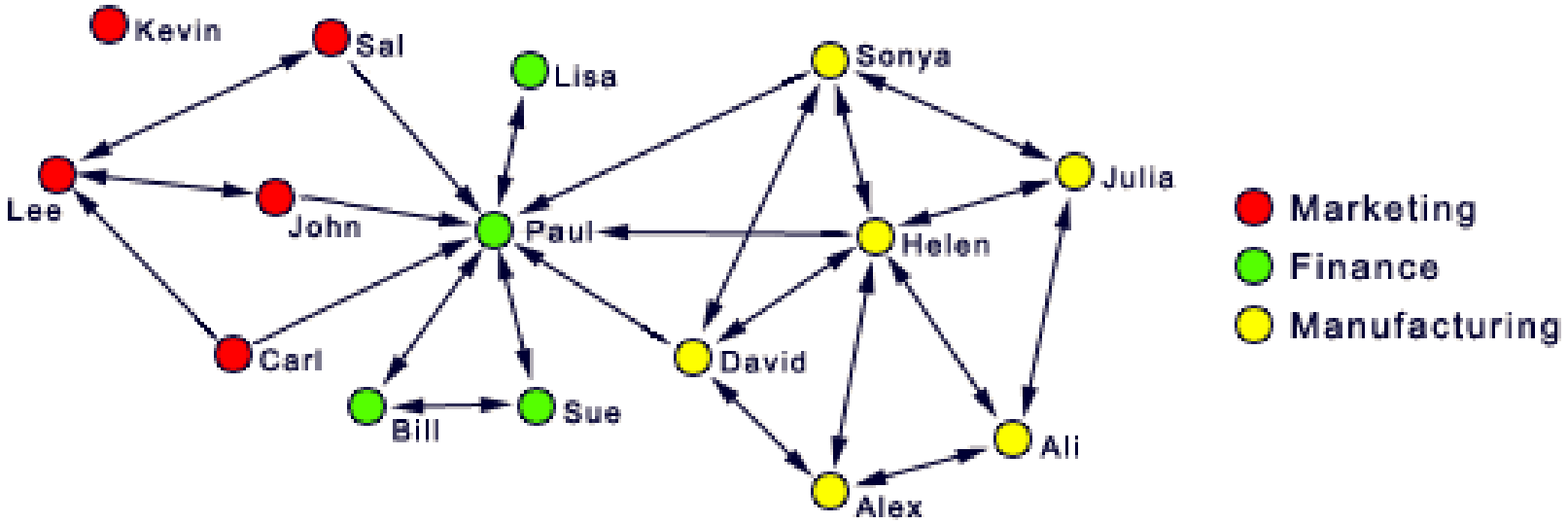


Case Study - Network in Organization

- Consider this informal network
- How did we come up with this network?
- Each edge represents communication link between two people
- What information can we extract from this network?



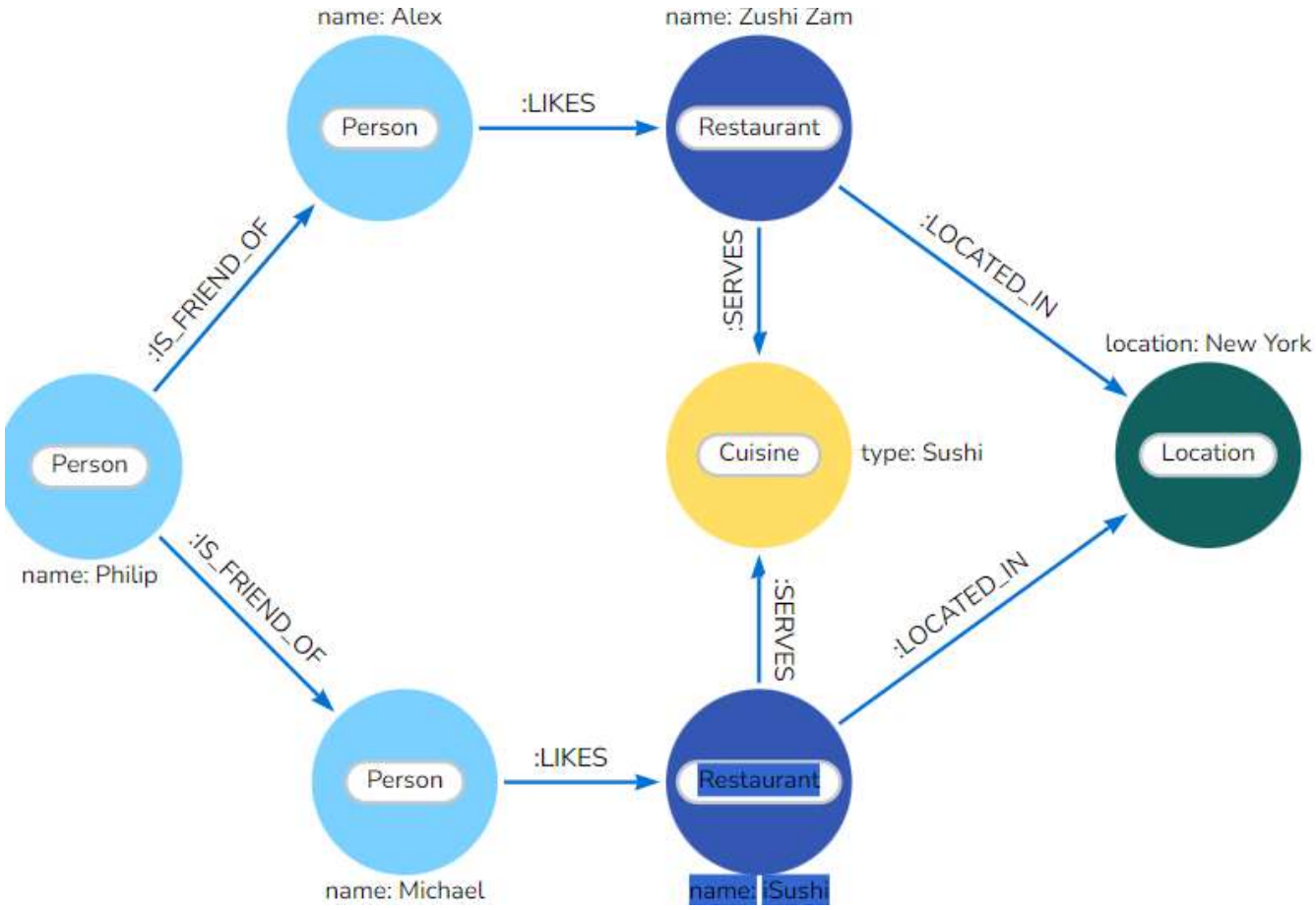
Case Study – Network in Organization



Fraud Detection

Source: Neo4j

Product Recommendation

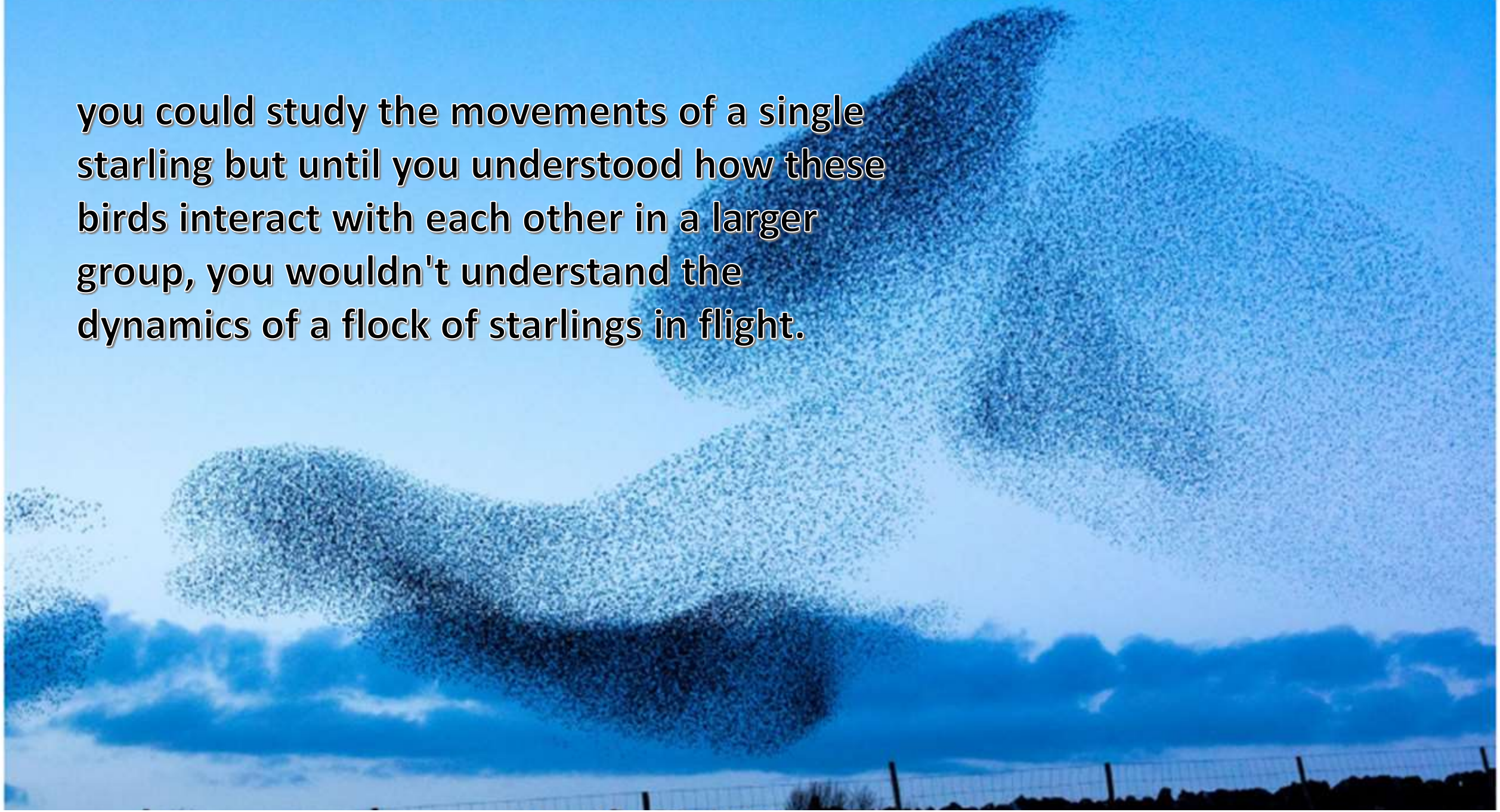


Graph Shaped Problems

- Pair wise relationship
- Self-relationship
- Path discovery
- Finding In-direct/hidden relation

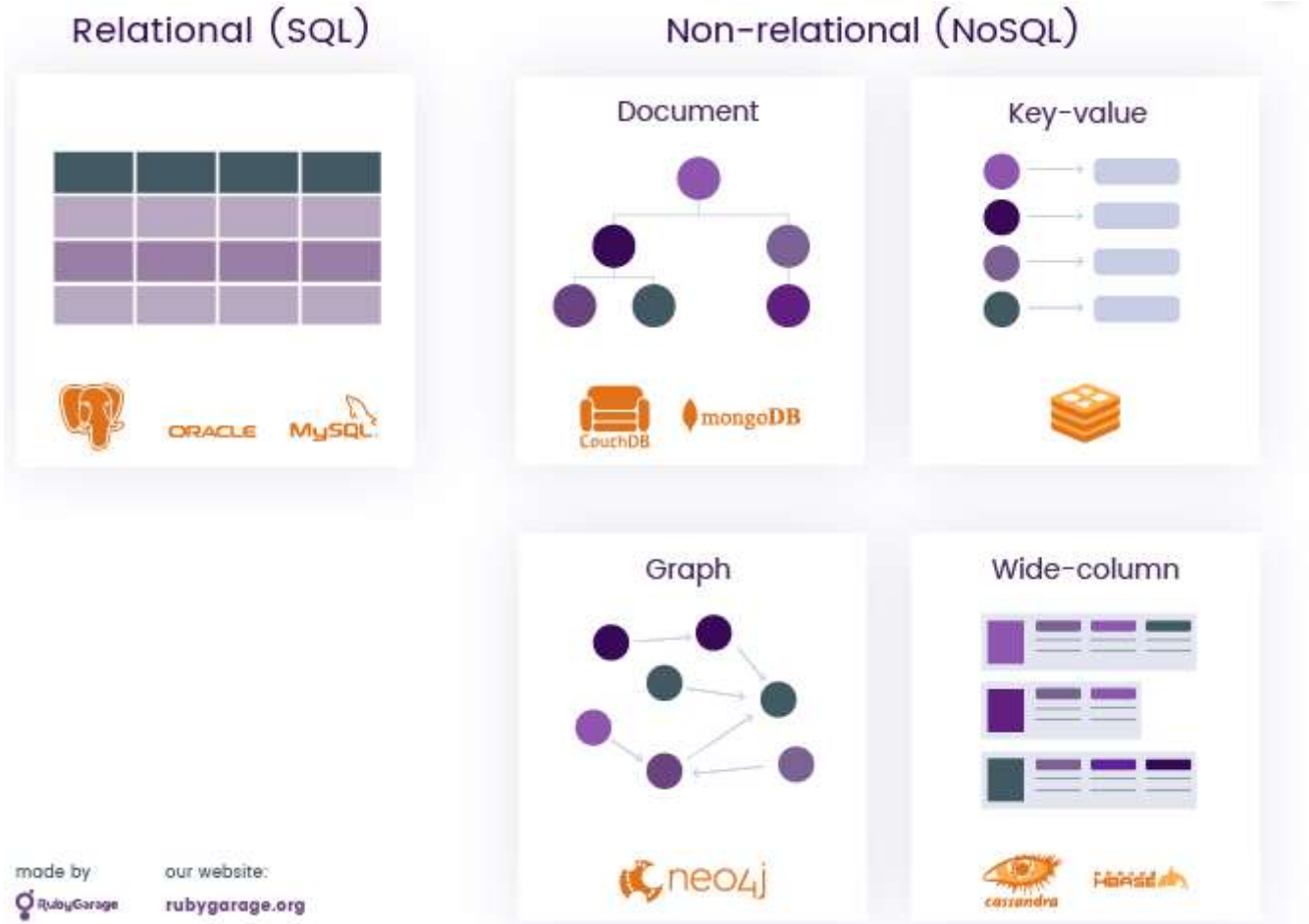
Connectedness of Data

- Big Data is used to define with five or six Vs:
  - Volume
  - Velocity
  - Variety
  - Veracity
  - Value
- Valence: *tendency of individual data to connect and overall connectedness*
  - Chemistry: combining power of an element
  - Psychology: intrinsic attractiveness of an object
  - Linguistics: number of elements a word combines



Source: A Comprehensive Guide to, Mark Needham & Amy E. Hodler

Database Systems



Naïve Graph Databases

- Graph databases can be in two forms
  - Layer over SQL
  - Implemented as Graph

- Naïve: implemented as Graph
  - Index-free adjacency: traverse relationship without using indexes
  - Local performance: Graph size does not affect the performance
  - In RDBMS: Joins is the major cause of performance issues