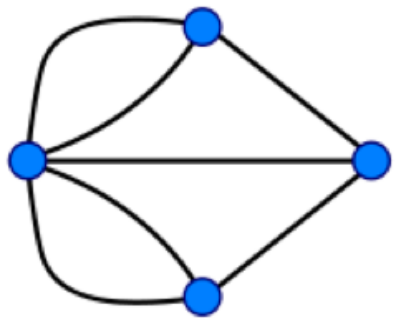
Mining Large-Scale Networks
Lecture 1

Charalampos E. Tsourakakis
Boston University
ctsourak@bu.edu

graphs: a simple model

- entities – set of vertices
- pairwise relations among vertices
  – set of edges
- can add directions, signs, weights, . . .
- This class aims to provide you with the fundamental tools to
  - model
  - analyze
  - summarize
  - leverage

  graphs in a variety of applications



graph theory

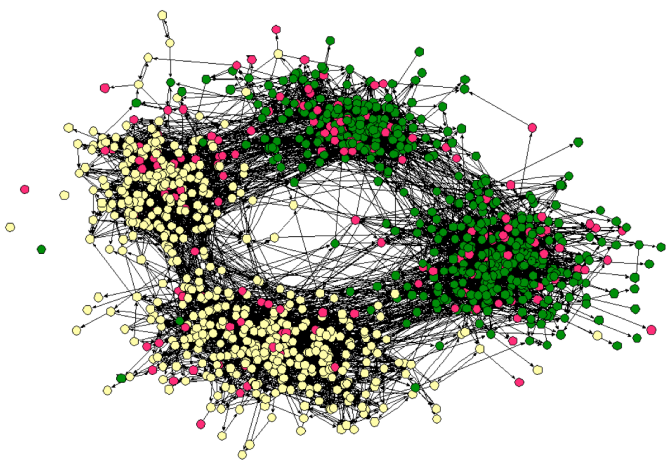
- graph theory started in the 18th century, with Leonhard Euler
  - the problem of Königsberg bridges
  - since then, graphs have been studied extensively

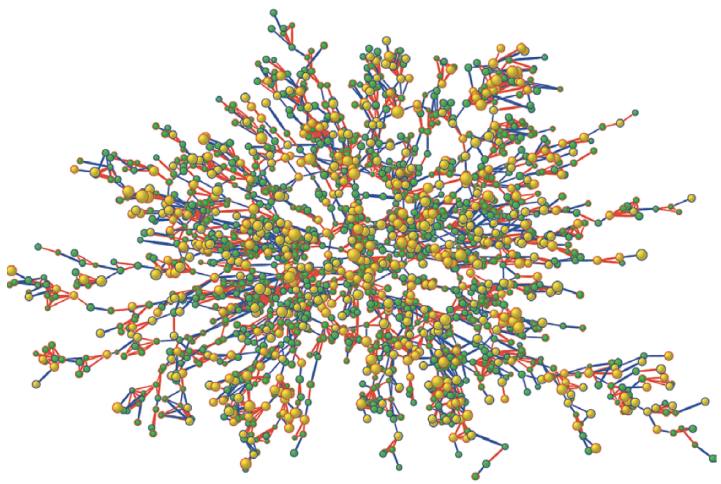
fundamental graphs

- The following unweighted, simple, undirected graphs will occur throughout our class
  - complete graph on n vertices K_n
  - path graph on n vertices P_n
  - cycle (ring) graph on n vertices C_n
  - star graph on n vertices S_n
  - The d-dimensional hypercube graph H_d

graphs and real-world datasets

- graphs datasets have been studied in the past
  e.g., networks of highways, social networks
  - usually these datasets were small
  - visual inspection can reveal a lot of information

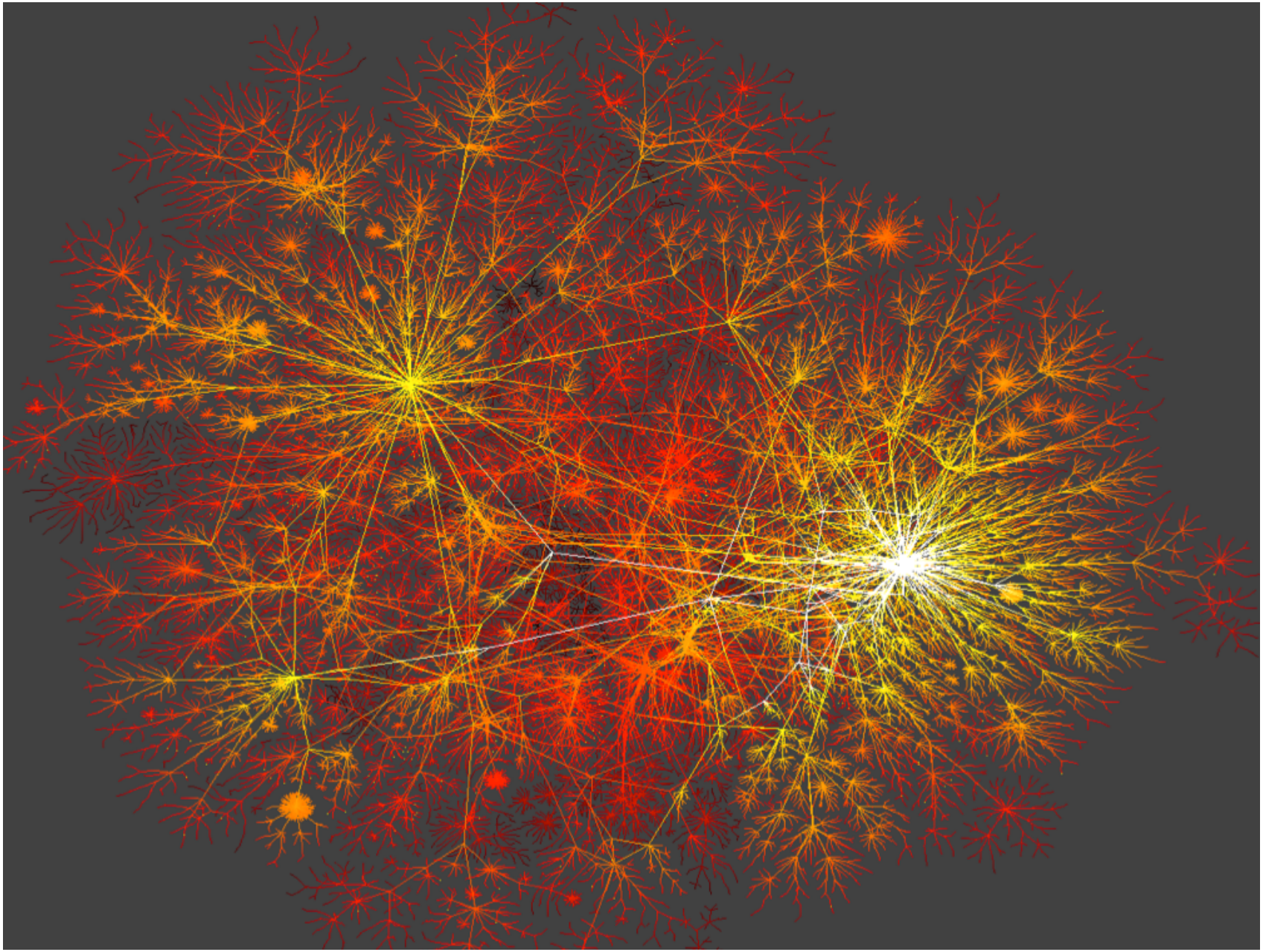




graphs and real-world datasets

- more and larger networks appear
  - products of technological advancement
    - e.g., internet, web
  - result of our ability to collect more, better-quality, and more complex data
    - e.g., gene regulatory networks
- networks of thousands, millions, or billions of nodes
  - impossible to visualize

the internet map



Traffic prediction with advanced Graph Neural Networks (DeepMind)

Influence Maximization in Social Networks

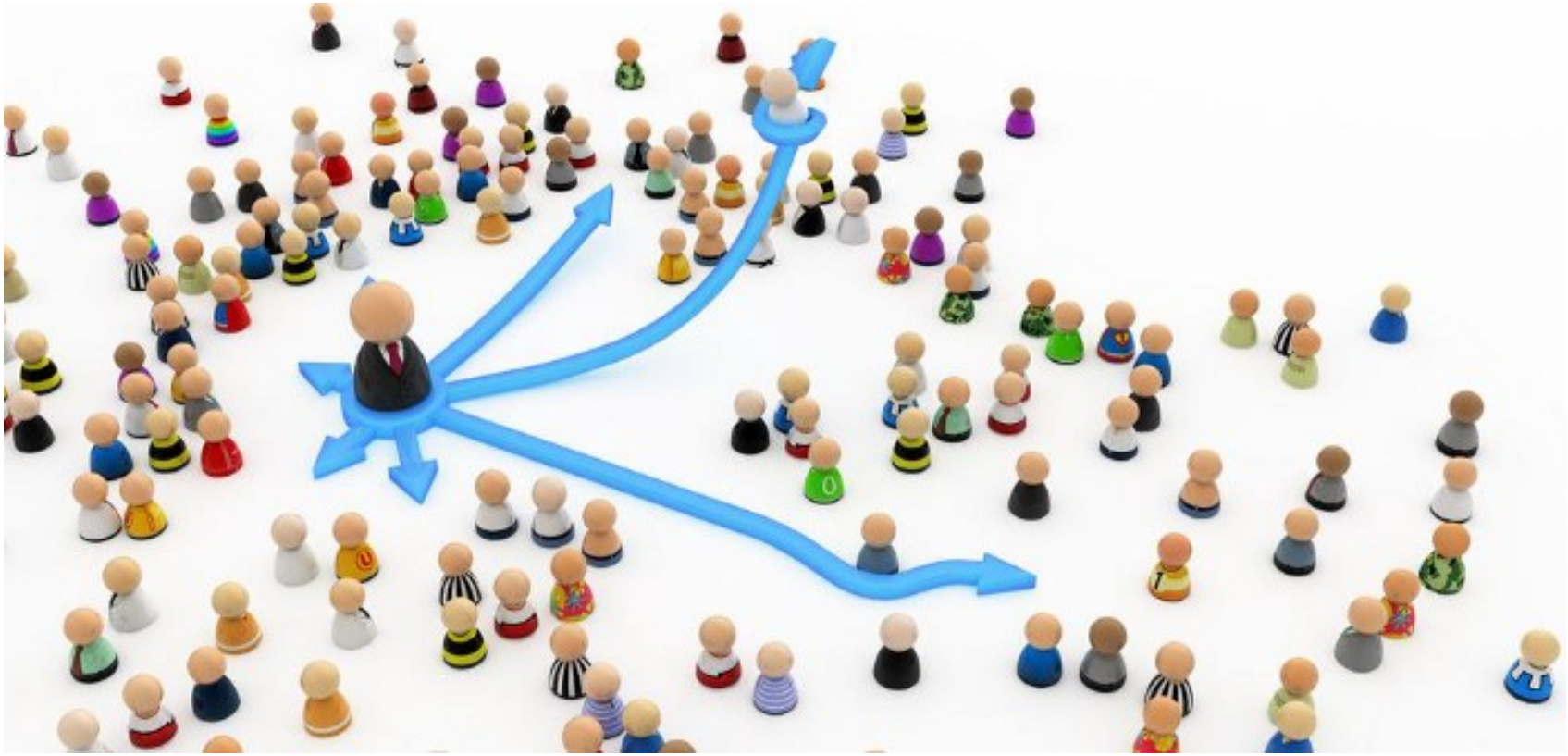
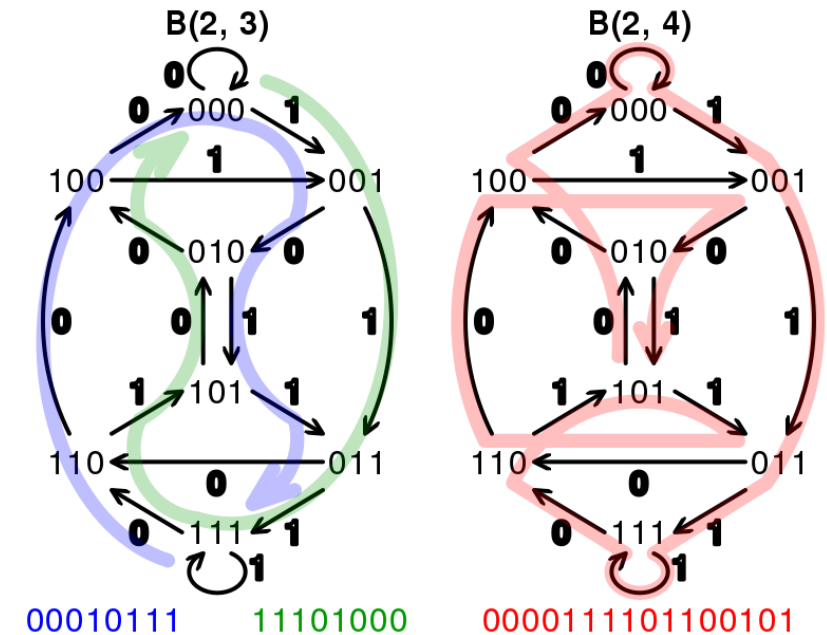


Figure source

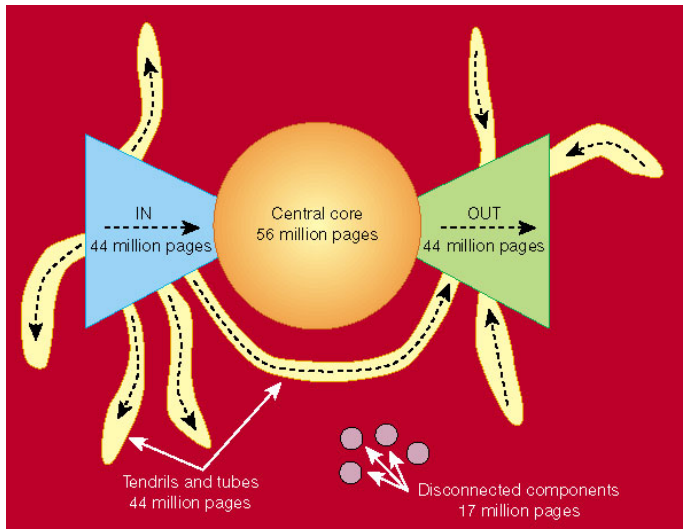
Antimoney Laundering

Locked cell phone? De Brujin Graphs!



Applications in genome assembly too!

the Web graph – aggregation levels



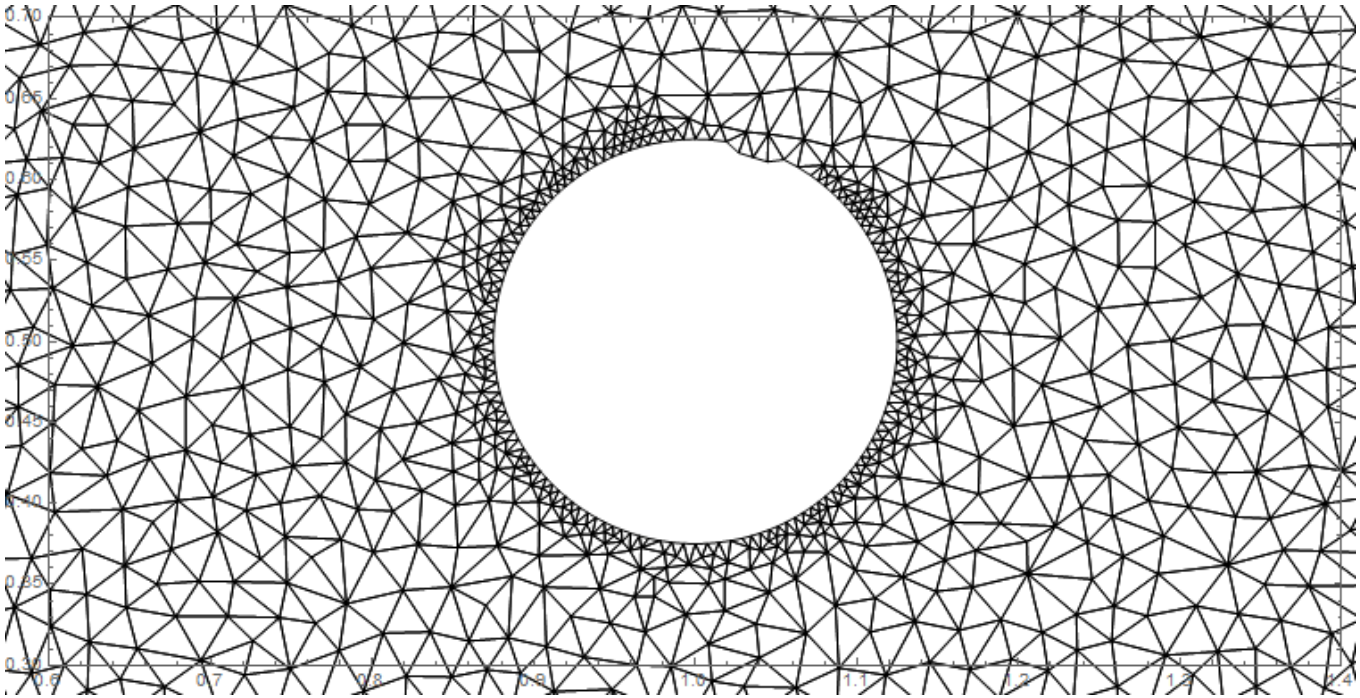
- Page graph
- Subdomain graphs
- PLD graphs

  Remark. Common Crawl project: largest publicly available crawl at the time - 3.5 billion pages, 128 billion links, 43 million pay-level domains

the human brain

Brain can be seen as a graph with $\sim 10^{11}$ neurons (nodes) and $\sim 10^{15}$ synapses (edges)

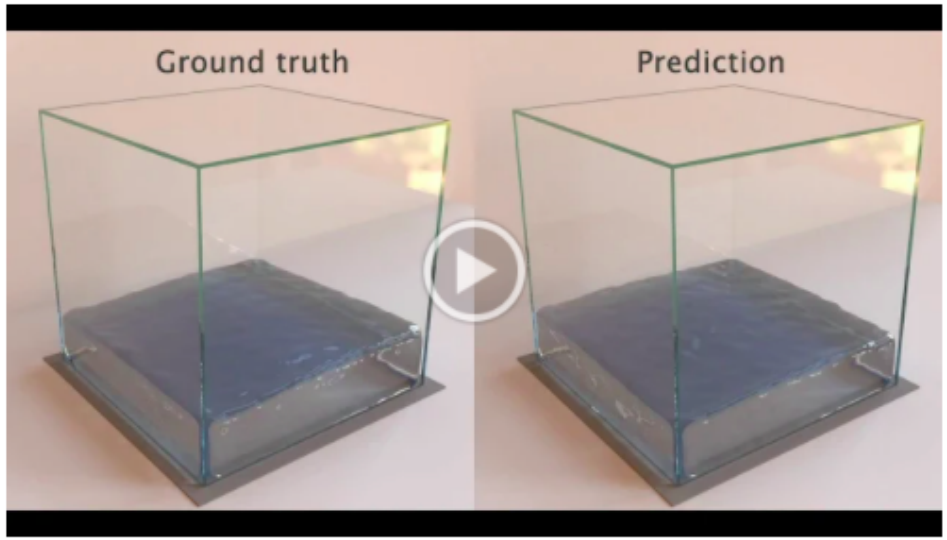
finite element meshes



Finite element meshes are used to study natural phenomena (e.g., heat diffusion) and simulate physical systems.

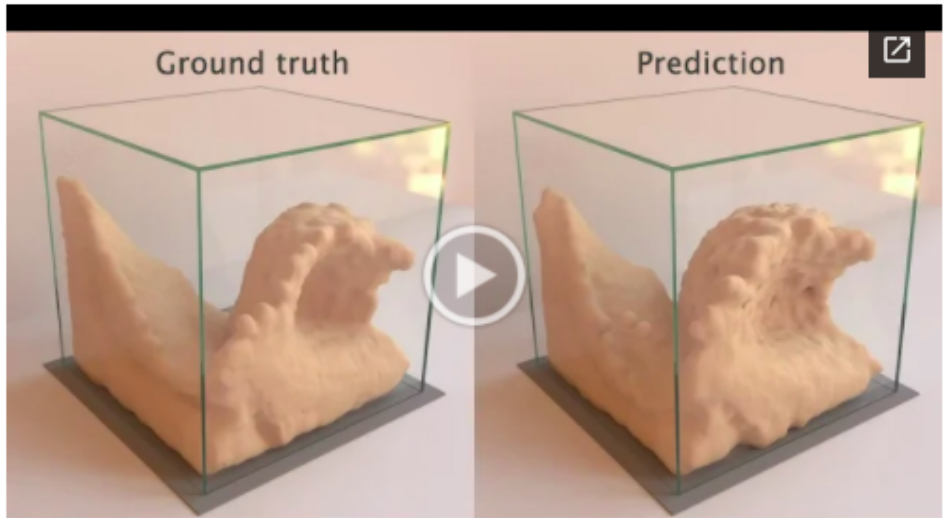
A continuous body (e.g., building, plate) is discretized into discrete elements that are interconnected.

Simulating physics with graph neural networks



Water-3D

- 14k particles
- 800 steps
- ground truth simulator: SPH

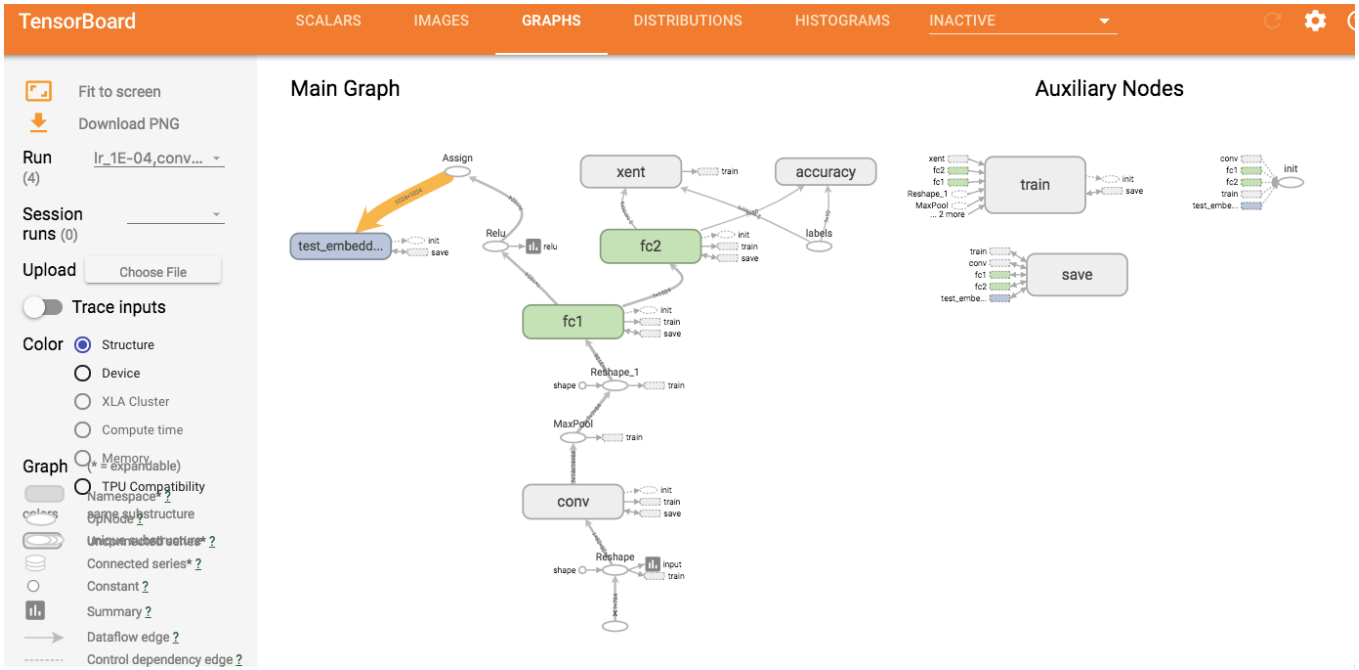


Sand-3D

- 19k particles
- 400 steps
- ground truth simulator: MPM

Source: Learning to simulate complex physics with graph networks

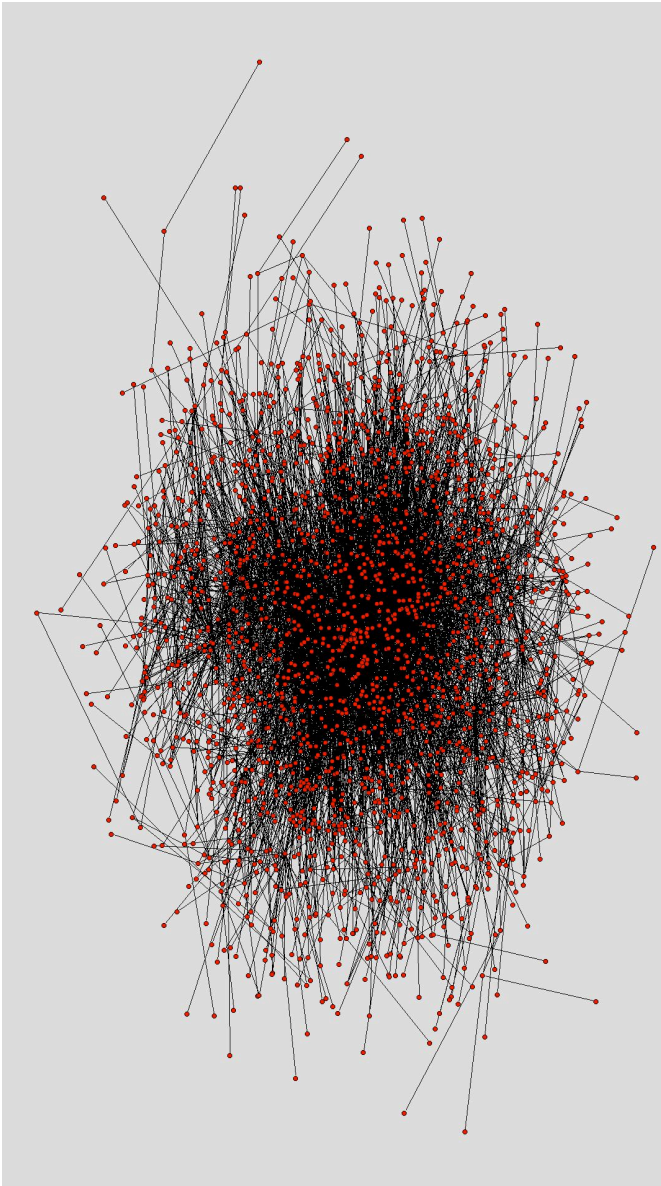
execution programs



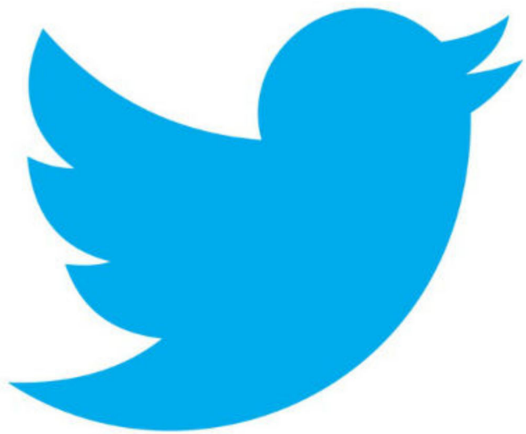
Tensorflow produces graphs, which define the program to be executed.

social networks

- links denote a social interaction
  - networks of acquaintances
  - collaboration networks
    - actor networks
    - co-authorship networks
    - director networks
  - phone-call networks
  - e-mail networks
  - IM networks
  - sexual networks



online social networks



Social interactions take place online as well.

Online social networks span a base of billions of users

Epidemics and social networks

RESEARCH ARTICLE

Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza

Simon Cauchemez, Achuyt Bhattarai, Tiffany L. Marchbanks, Ryan P. Fagan, Stephen Ostroff, Neil M. Ferguson, David Swerdlow, and the Pennsylvania H1N1 working group

PNAS February 15, 2011 108 (7) 2825-2830; https://doi.org/10.1073/pnas.1008895108

Edited by David Cox, Nuffield College, Oxford, United Kingdom, and approved December 22, 2010 (received for review June 22, 2010)

Article | Figures & SI | Info & Metrics | PDF

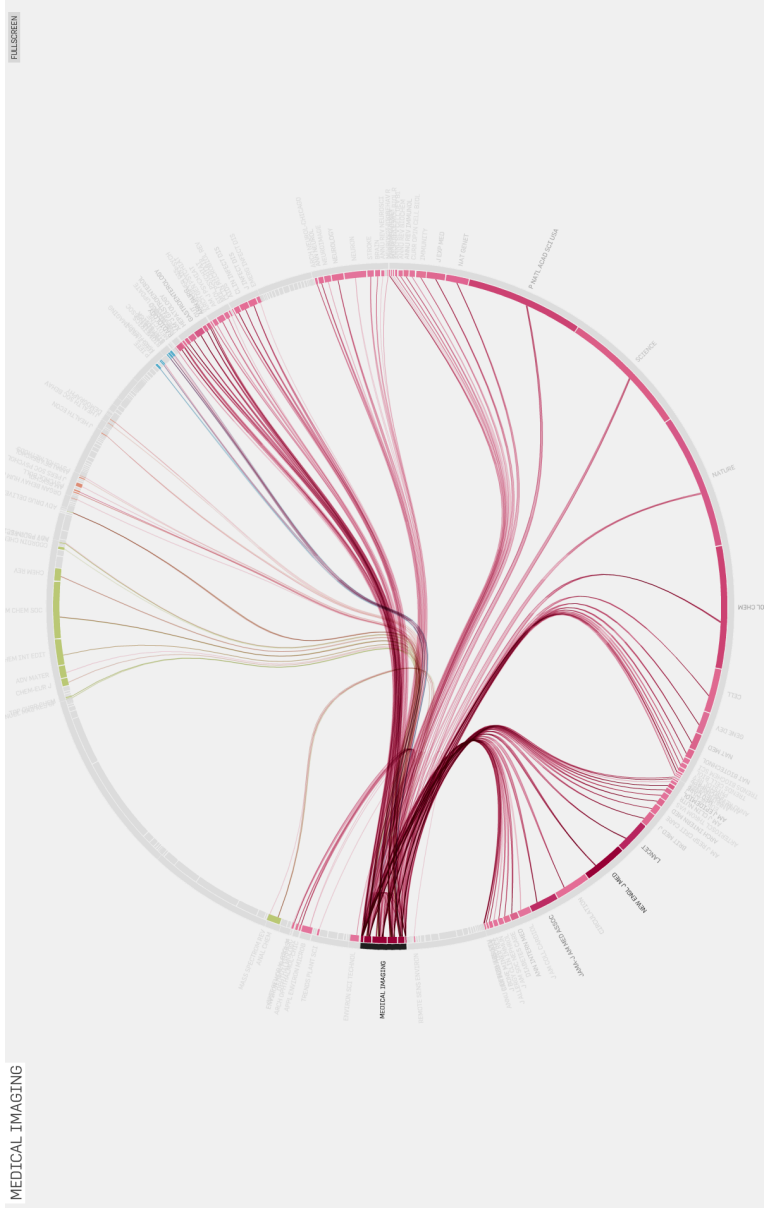
Abstract

Evaluating the impact of different social networks on the spread of respiratory diseases has been limited by a lack of detailed data on transmission outside the household setting as well as appropriate statistical methods. Here, from data collected during a H1N1 pandemic (pdm) influenza outbreak that started in an elementary school and spread in a semirural community in Pennsylvania, we quantify how transmission of influenza is

Link to PNAS paper

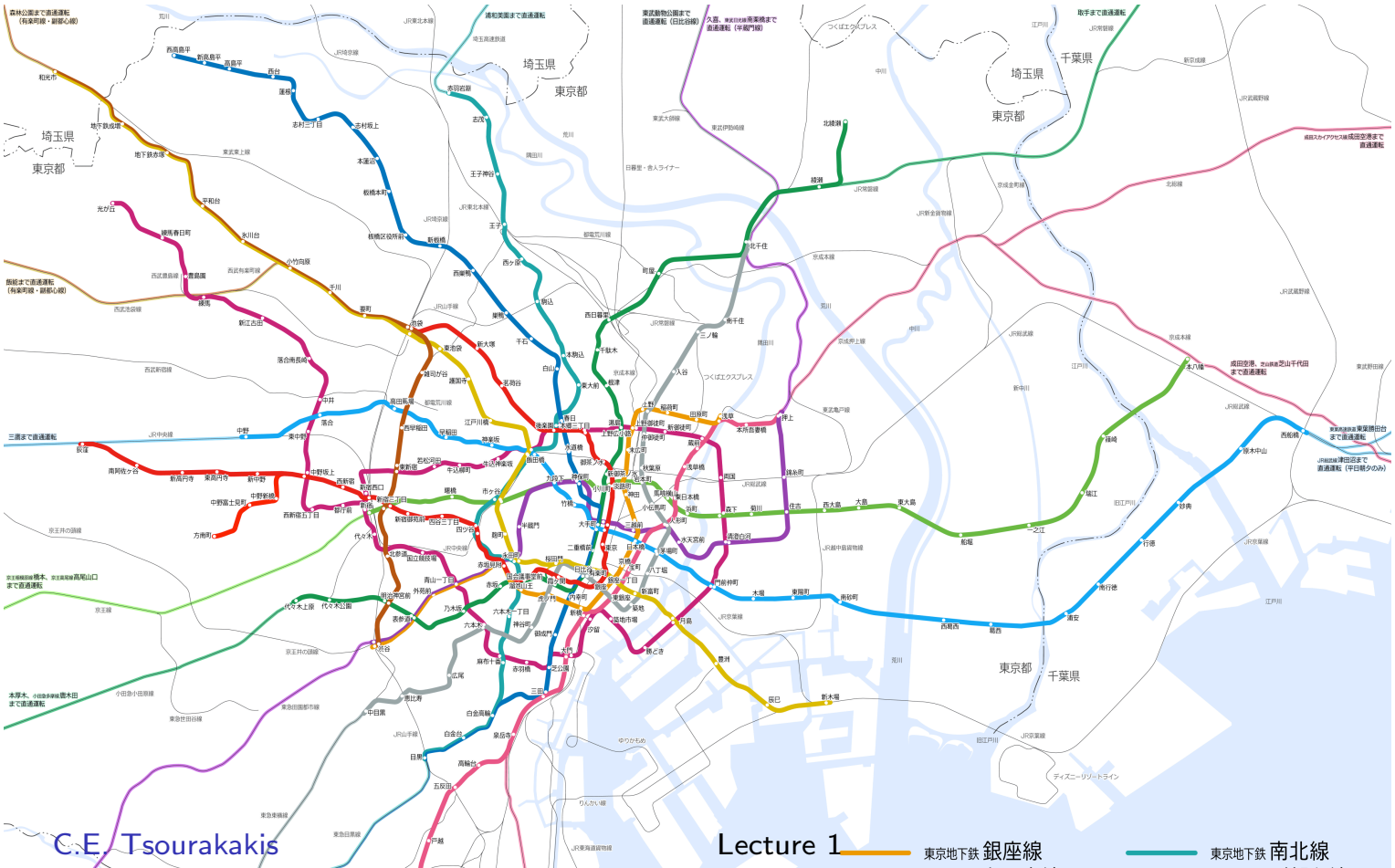
knowledge and information networks

- nodes store information, links associate information
  - citation network (directed acyclic)
  - the web (directed)
  - peer-to-peer networks
  - word networks
  - networks of trust
  - software graphs
  - bluetooth networks
  - home page/blog networks

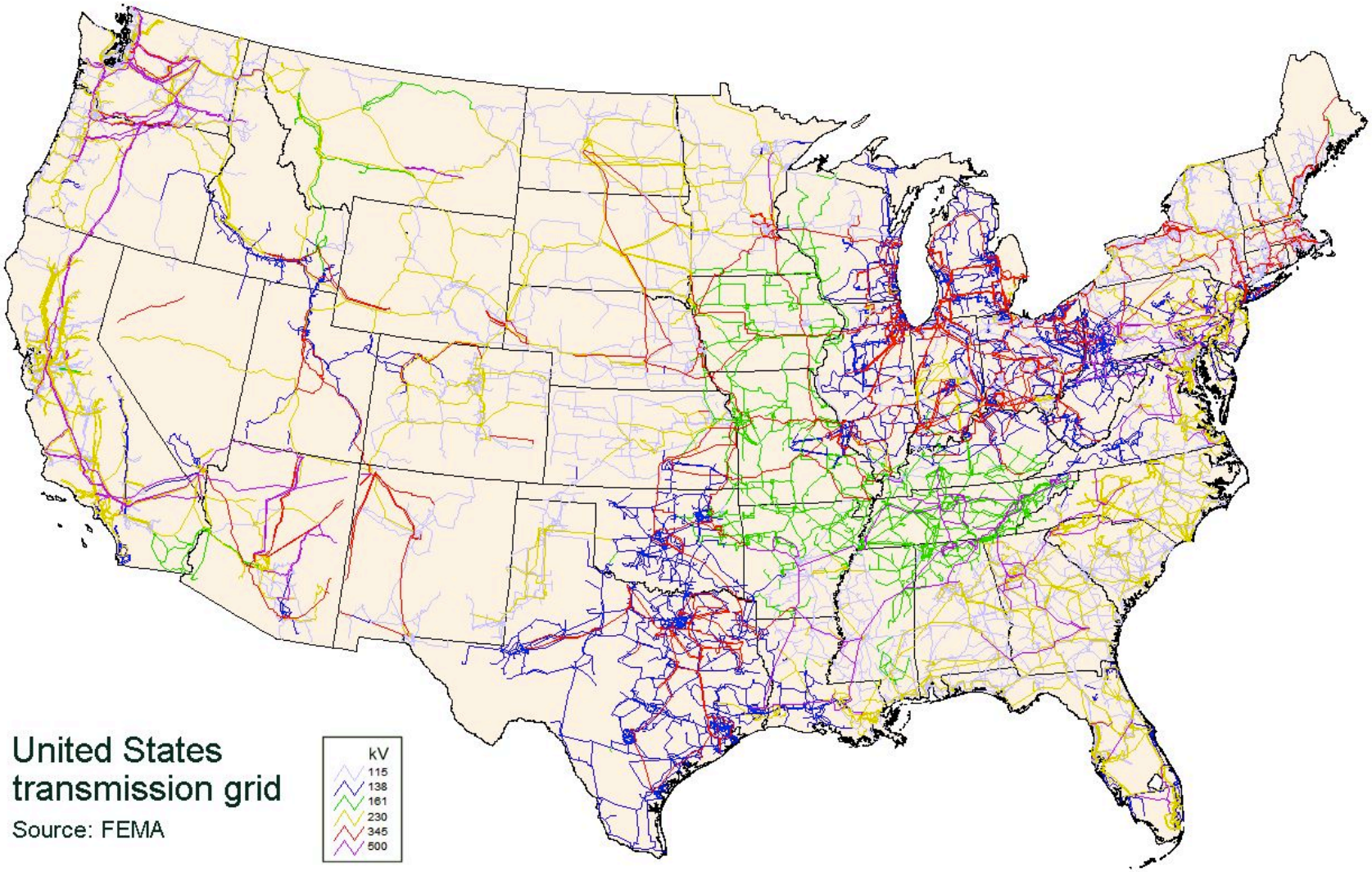


technological networks

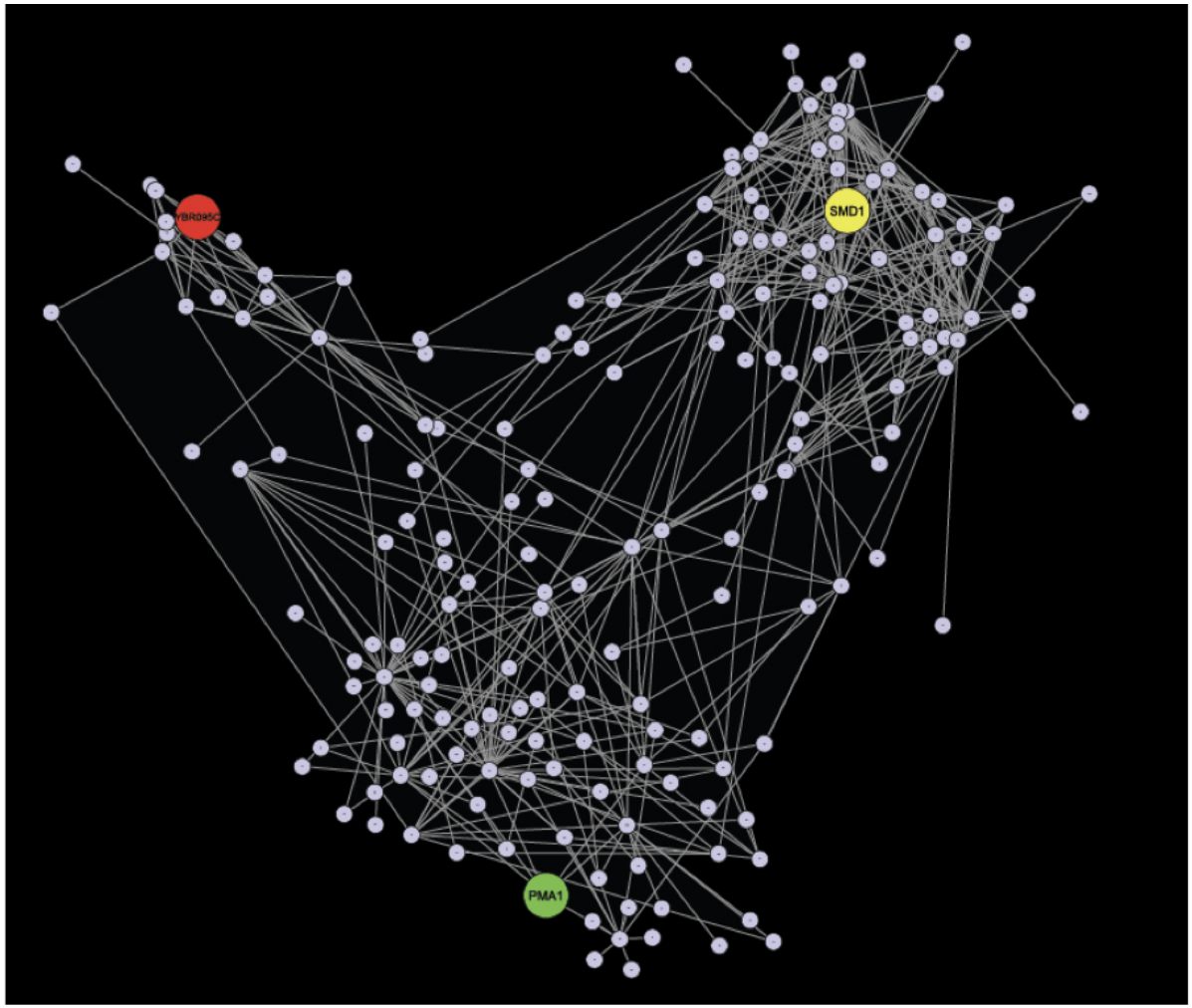
- networks built for distribution of a commodity
  - the internet, power grids, telephone networks
  - airline networks, transportation networks



US power grid



interactome

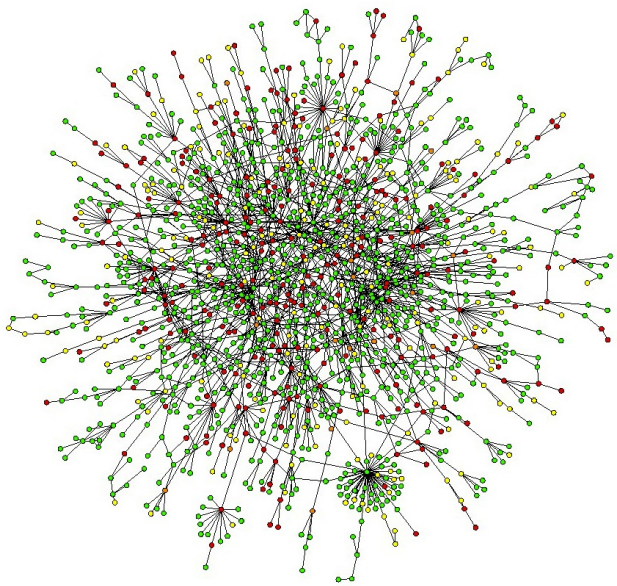


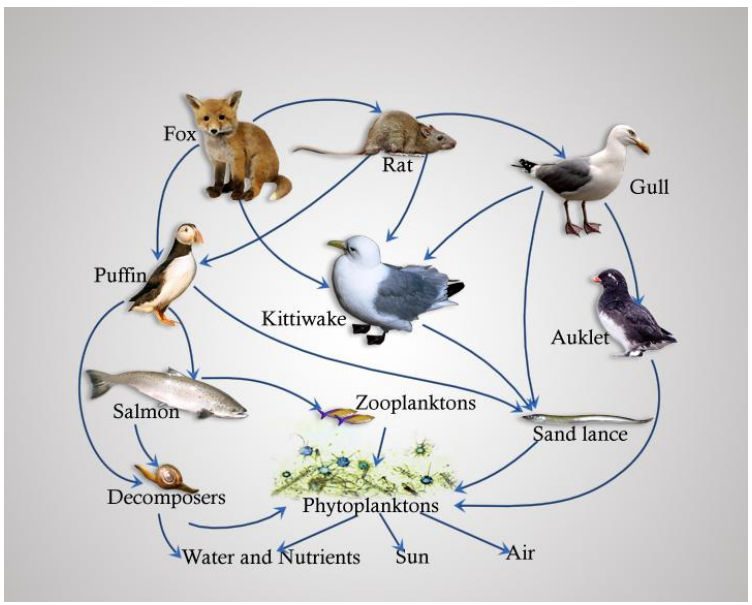
Protein-protein interaction network

Source: "Uncovering Biological Network Function via Graphlet Degree Signatures" by Milenkovi and Przulj,

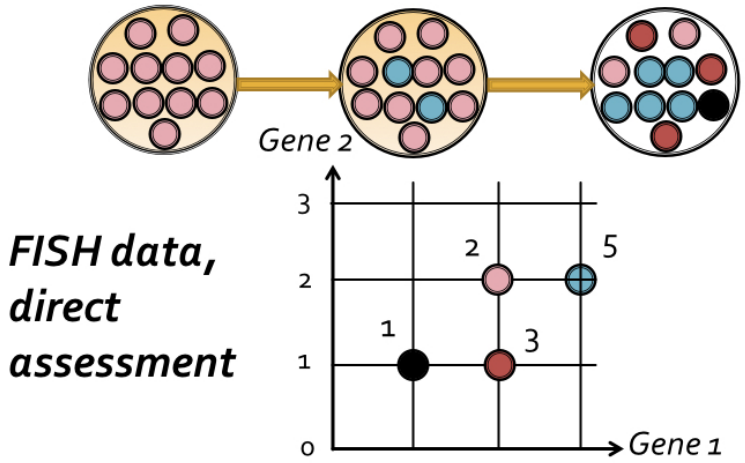
biological networks

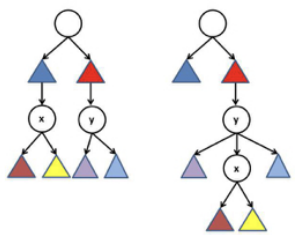
- biological systems represented as networks
  - protein-protein interaction networks
  - gene regulation networks
  - gene co-expression networks
  - metabolic pathways
  - the food web
  - neural networks





tumor phylogenies





How does cancer progress?

Thyroid functionality

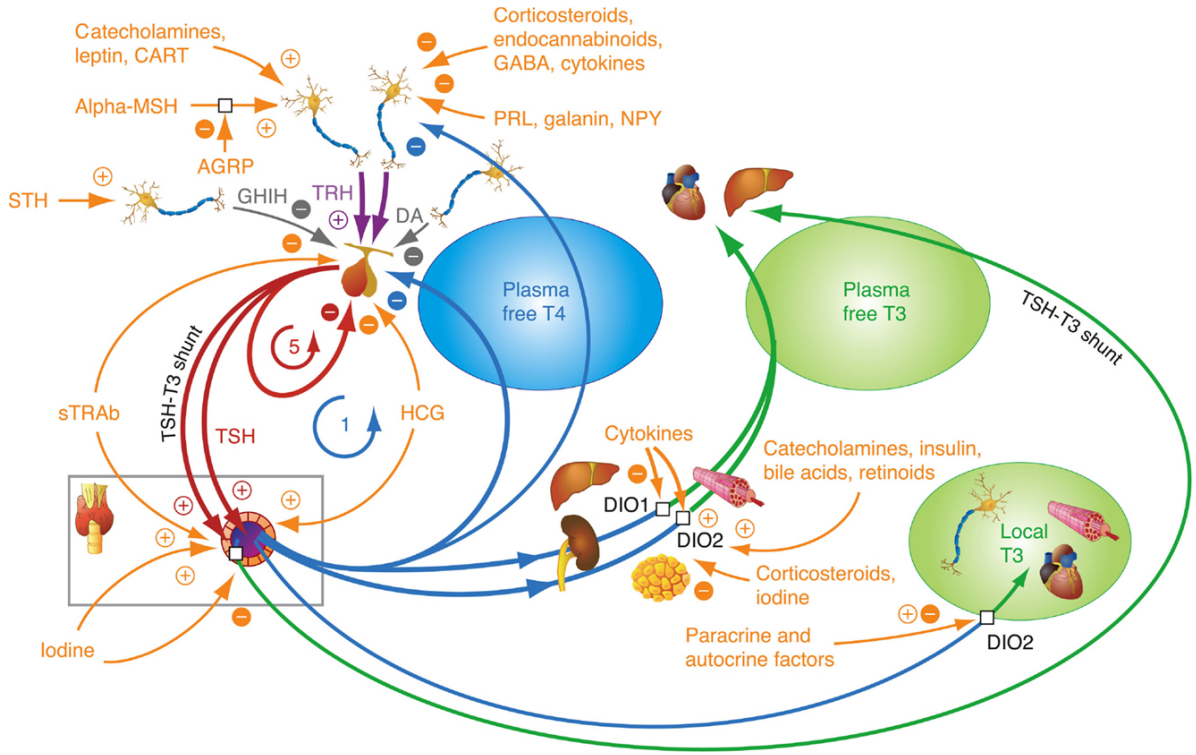
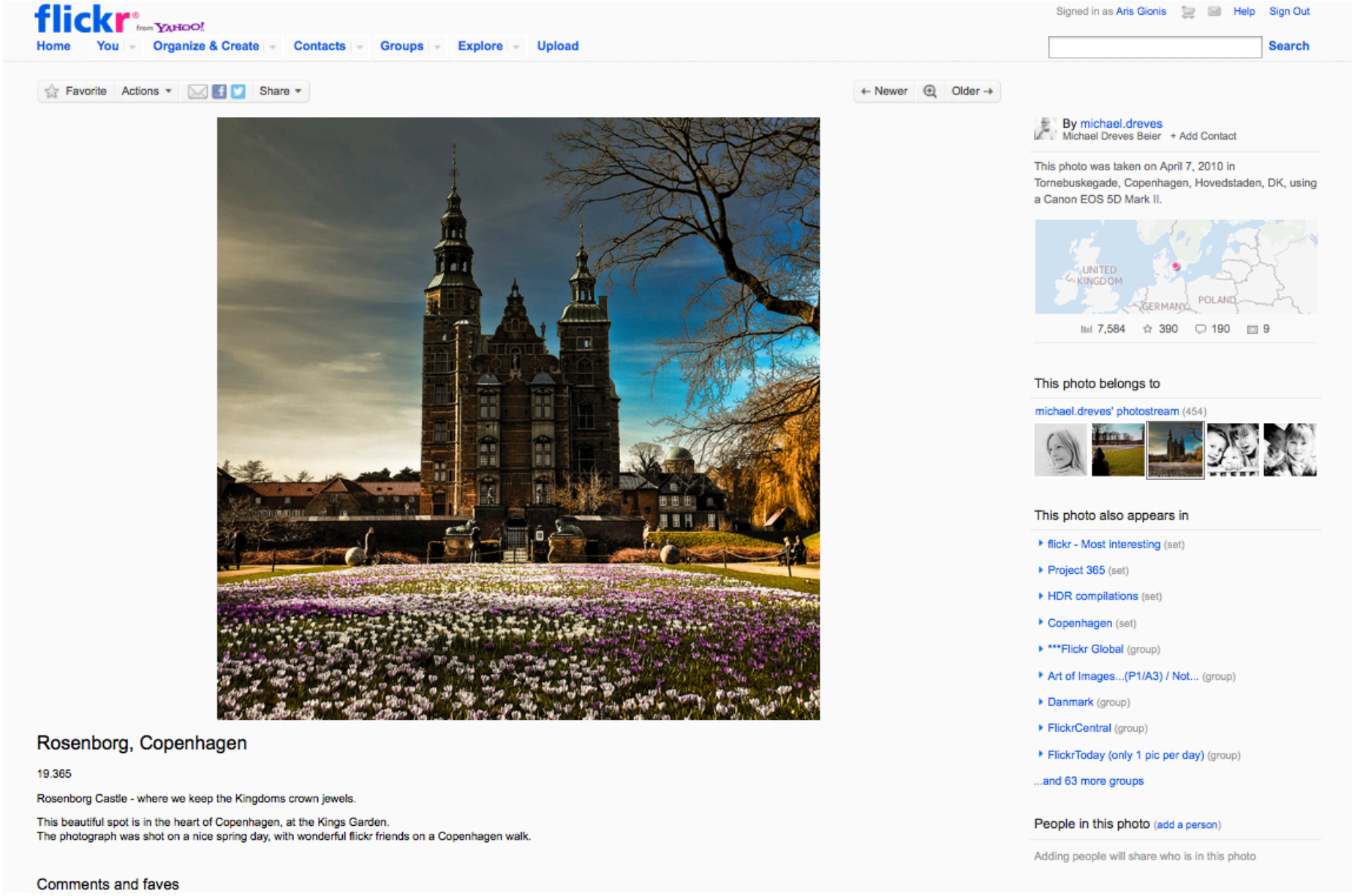


photo-sharing site



what is the underlying graph?

- nodes: photos, tags, users, groups, albums, sets, collections, geo, query, ...
- edges: upload, belong, tag, create, join, contact, friend, family, comment, fave, search, click, ...
- also many interesting induced graphs
  - tag graph: based on photos
  - tag graph: based on users
  - user graph: based on favorites
  - user graph: based on groups

- which graph to pick — not an easy choice

recurring theme

- social media, user-generated content
- user interaction is composed by many atomic actions
  - post, comment, like, mark, join, comment, fave, thumps-up, ...
  - generates all kind of interesting graphs to mine

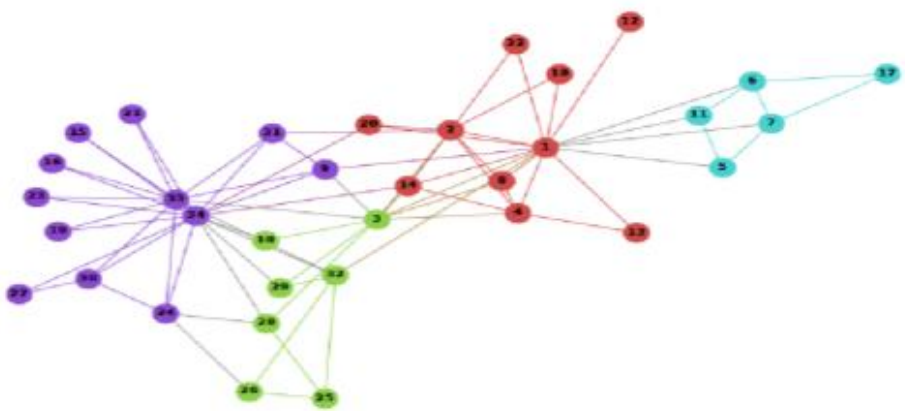
node embeddings

[Submitted on 28 Mar 2019 (v1), last revised 9 Apr 2019 (this version, v3)]

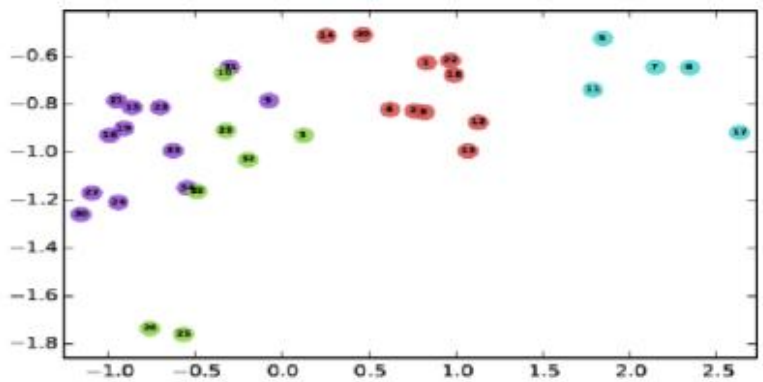
PyTorch-BigGraph: A Large-scale Graph Embedding System

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, Alex Peysakhovich

Graph embedding methods produce unsupervised node features from graphs that can then be used for a variety of machine learning tasks. Modern graphs, particularly in industrial applications, contain billions of nodes and trillions of edges, which exceeds the capability of existing embedding systems. We present PyTorch-BigGraph (PBG), an embedding system that incorporates several modifications to traditional multi-relation embedding systems that allow it to scale to graphs with billions of nodes and trillions of edges. PBG uses graph partitioning to train arbitrarily large embeddings on either a single machine or in a distributed environment. We demonstrate comparable performance with existing embedding systems on common benchmarks, while allowing for scaling to arbitrarily large graphs and parallelization on multiple machines. We train and evaluate embeddings on several large social network graphs as well as the full Freebase dataset, which contains over 100 million nodes and 2 billion edges.



Input



Output

Applications in recommender systems, node classification, link prediction . . .

now what?

- the world is full with networks and full of datasets that can be turned into networks
- what do we do with them?
  - understand their topology and measure their properties
  - study their evolution and dynamics
  - create realistic models
  - create algorithms that make use of the network structure
  - learn with and from graphs