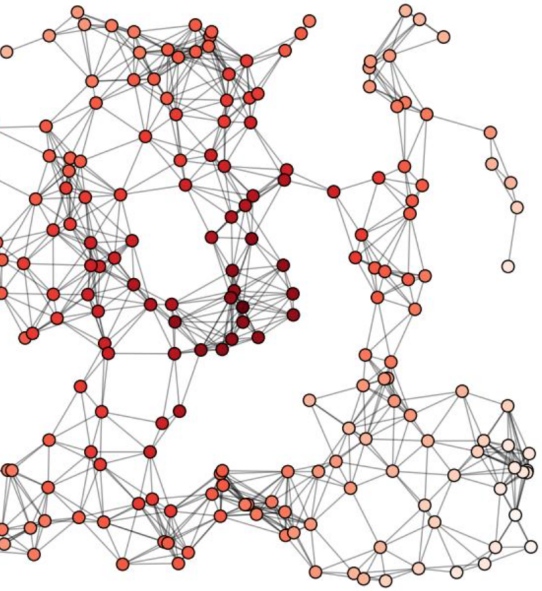


Introduction to Graph Mining

Pili Hu
http://hupili.net/

April 22, 2015
General Assembly

Speaker Bio

Pili Hu

2007-2010 UESTC -- Network Engineering
2010-2011 Baidu -- Search Engine Algorithm R/D
2011-2014 CUHK -- Decentralized Social Networks
2014-now HyperLab -- A Personal Search Engine

Co-founder of HyperLab: http://hyperlab.io/
Co-initiator of Code4HK: http://code4.hk/
Co-initiator of Open Innovation Lab: http://facebook.com/cuhkoil
Co-initiator of HKITE: http://hkite.org/

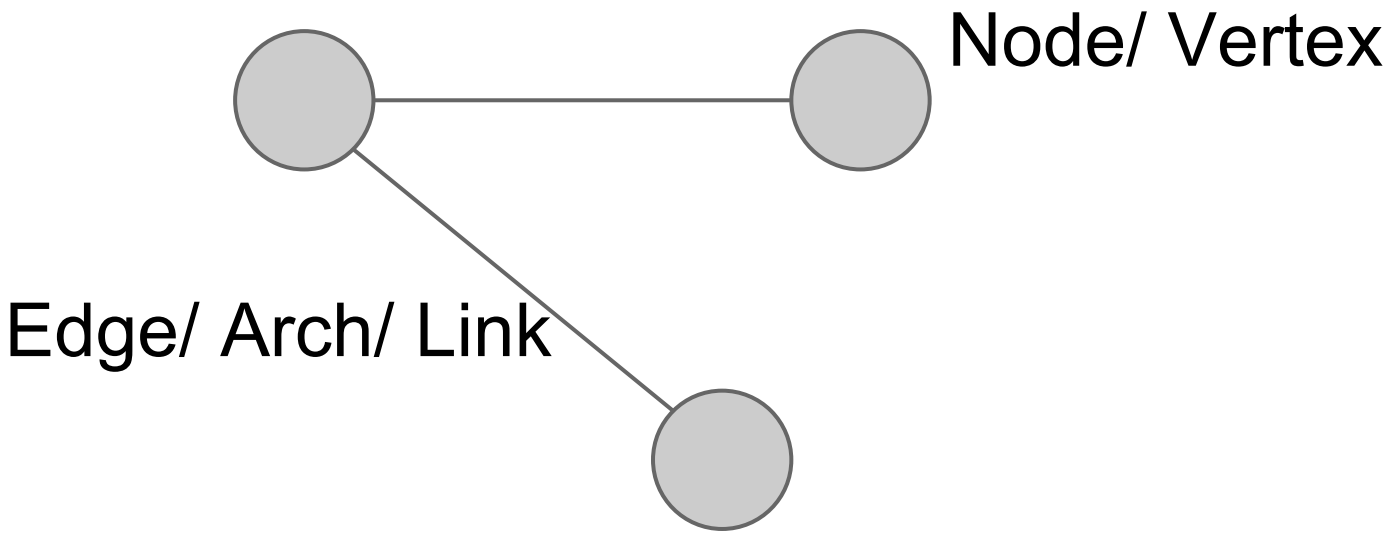
Outline

- Basics of Applied Graph Theory
- Graph operation in Python
  - networkx
- A study of LegcoHK data set via Graph tools

Basics of Applied Graph Theory

- What is a graph?
- How to represent a graph?
- Classical problems on graph
- Modern problems on graph (Applications)
- Common graph algorithms

What is a graph

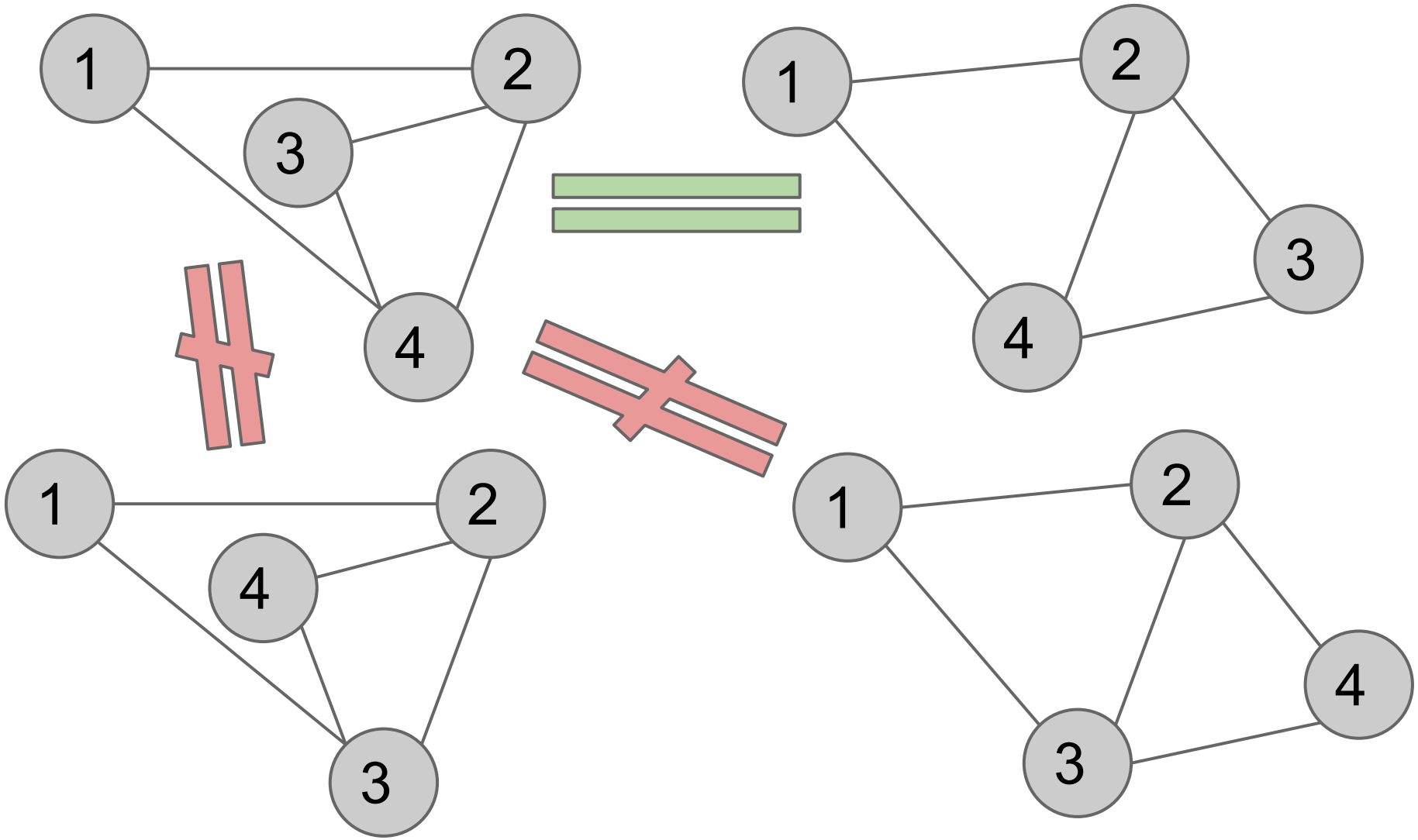


Type of Graphs

Undirected

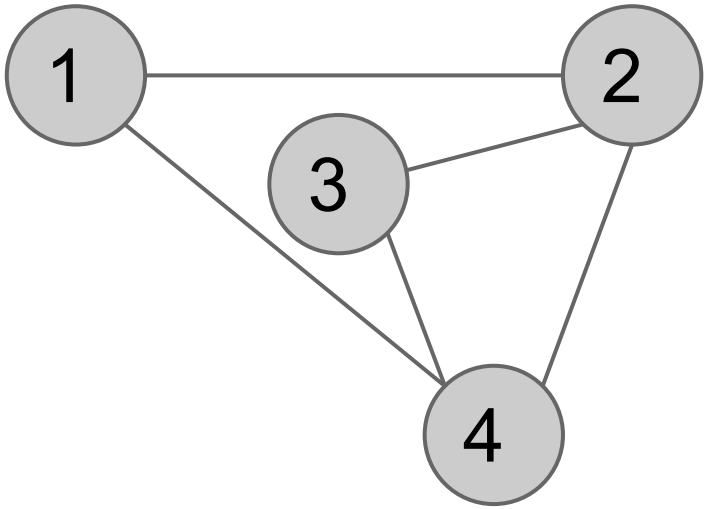
Directed

Graph cares topology only

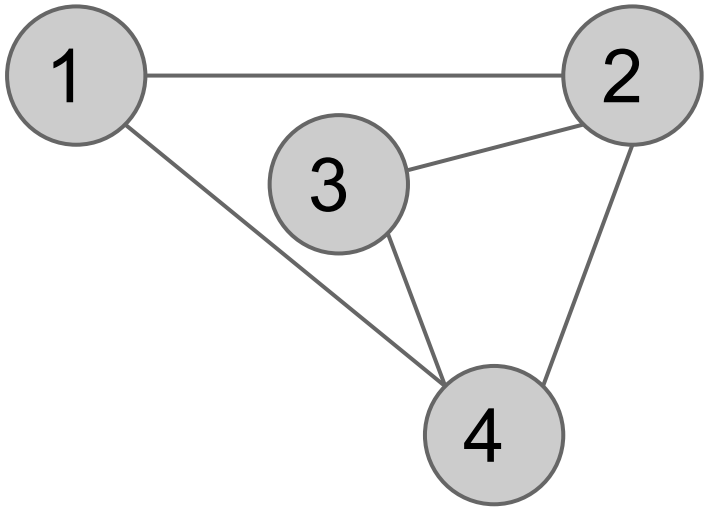


How to represent a graph

Draw a picture…

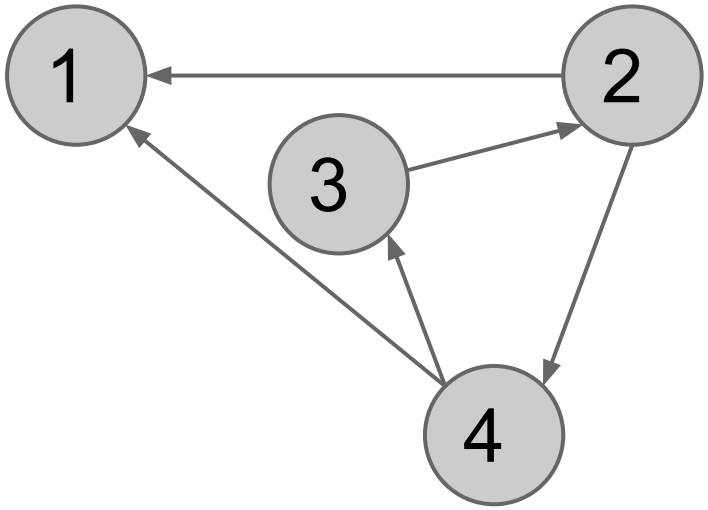


Adjacency Matrix



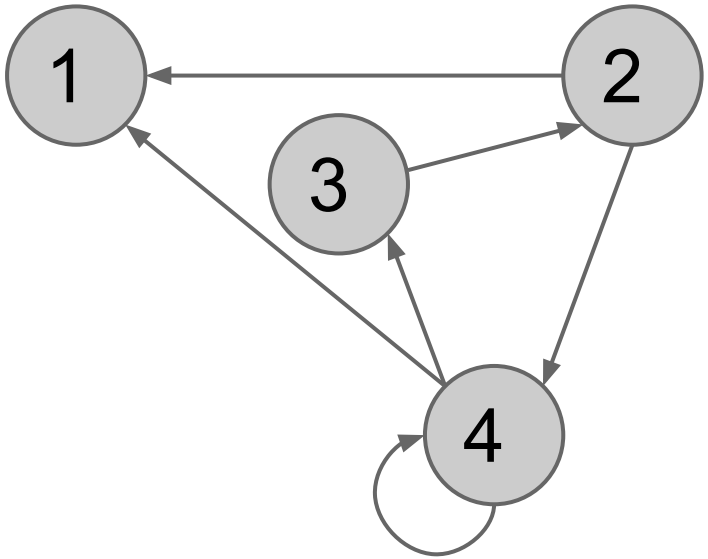
	1	2	3	4
1	0	1	0	1
2	1	0	1	1
3				
4				

Adjacency Matrix (Directed)



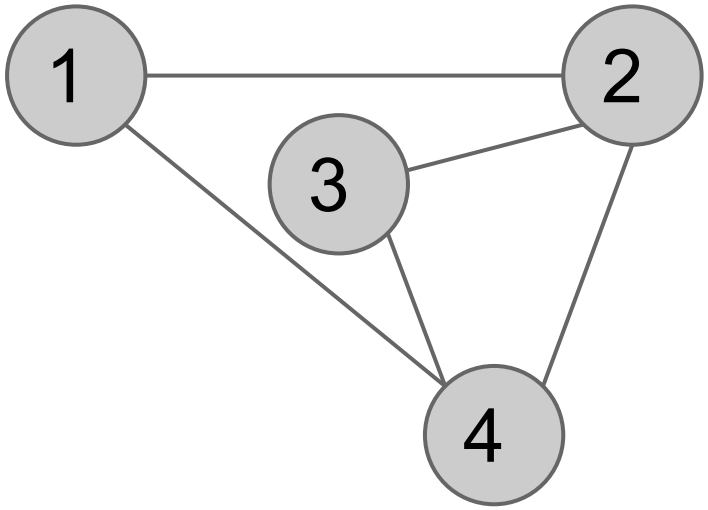
	1	2	3	4
1	0	0	0	0
2	1	0	0	1
3				
4				

Adjacency Matrix (Self-Loop)



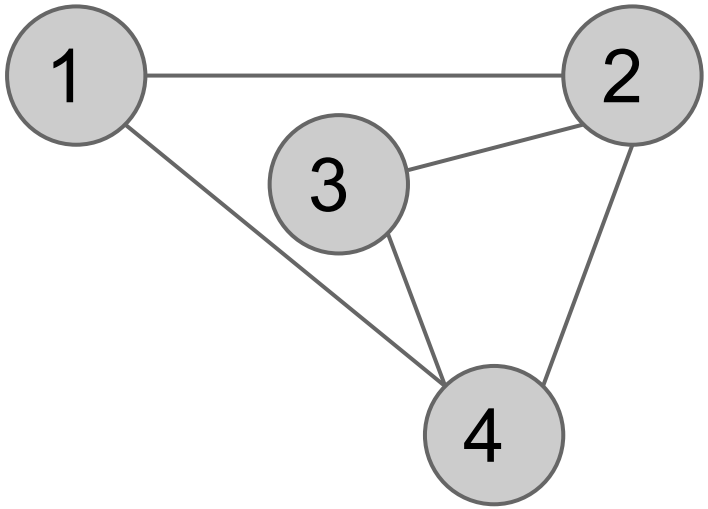
	1	2	3	4
1	0	0	0	0
2	1	0	0	1
3	0	1	0	0
4				

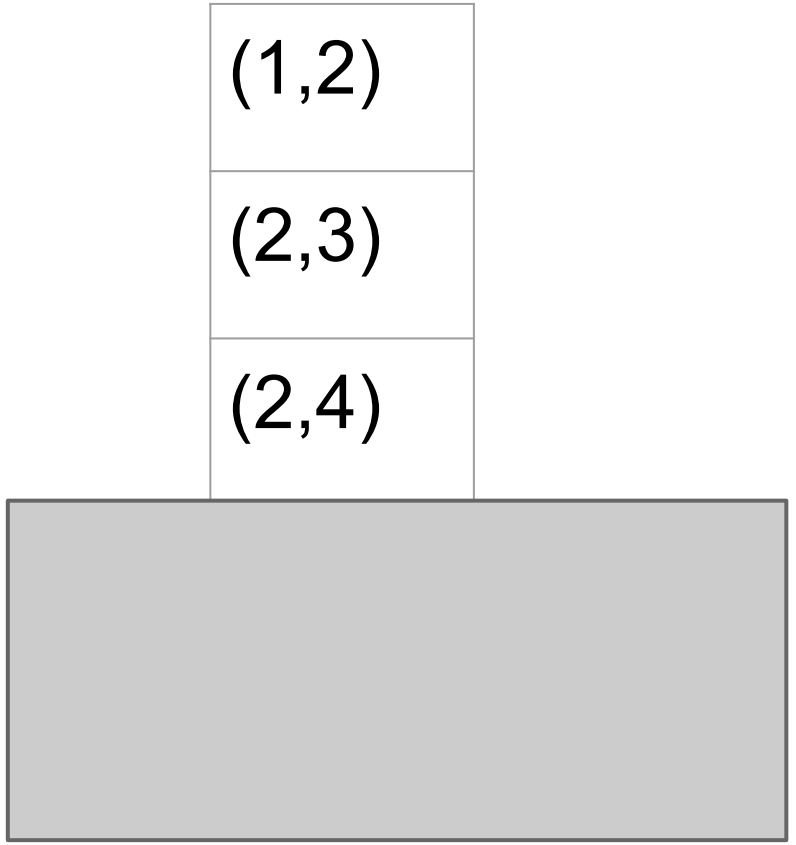
Adjacency List



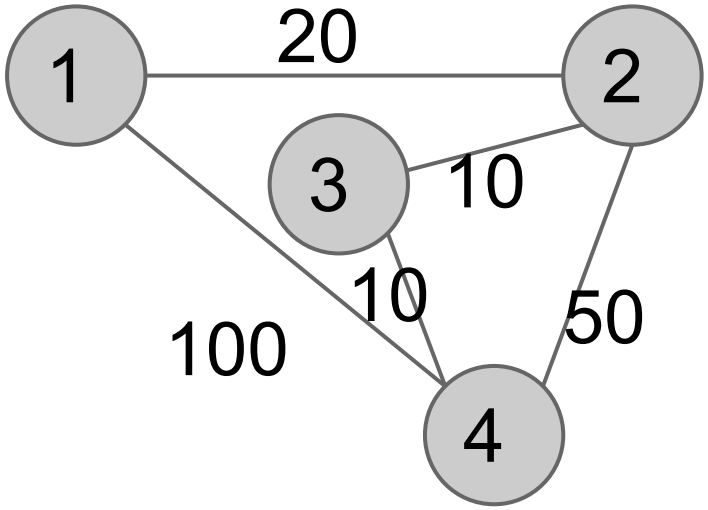
1	2, 3, 4
2	1, 3, 4
3	
4	

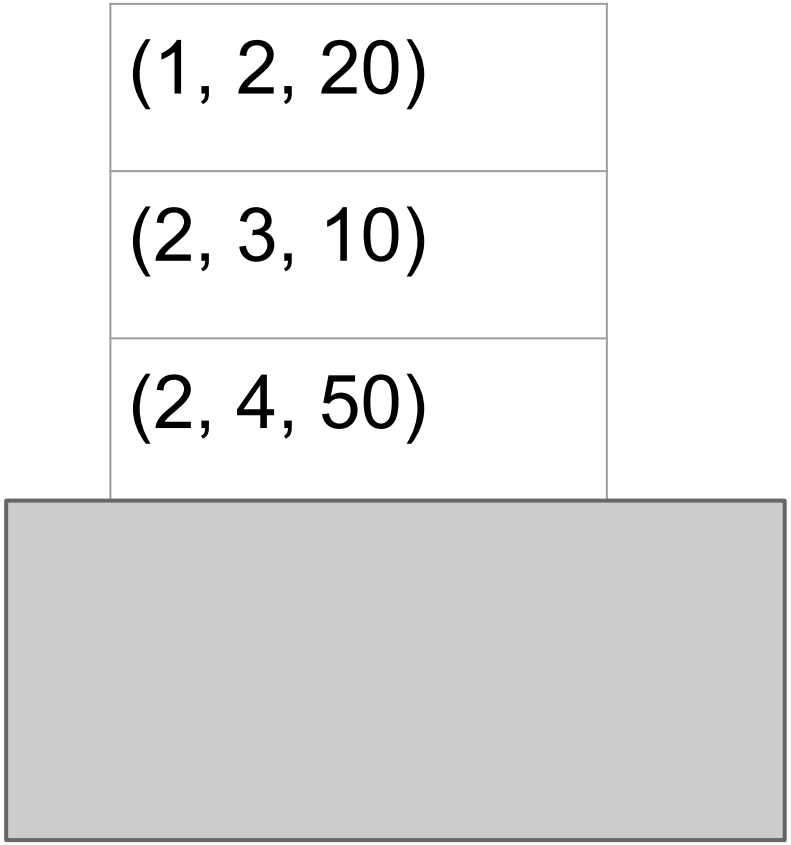
Edge List





Edge List (Weighted)





Vertex List

Why don’t we need it?

Some questions

- How do you represent weighted graph in adjacency matrix/ adjacency list?

A matter of representation

For efficiency of basic operations.

- Does A follow B?
- Who are B's friends in the network?
- What are the distribution of relationships?
  - (Edge sampling)

	1	2	3	4
1	0	0	0	0
2	1	0	0	1
3	0	1	0	0
4	1	0	1	1

Adjacency Matrix

1	2, 3, 4
2	1, 3, 4
3	2, 4
4	1, 2, 3

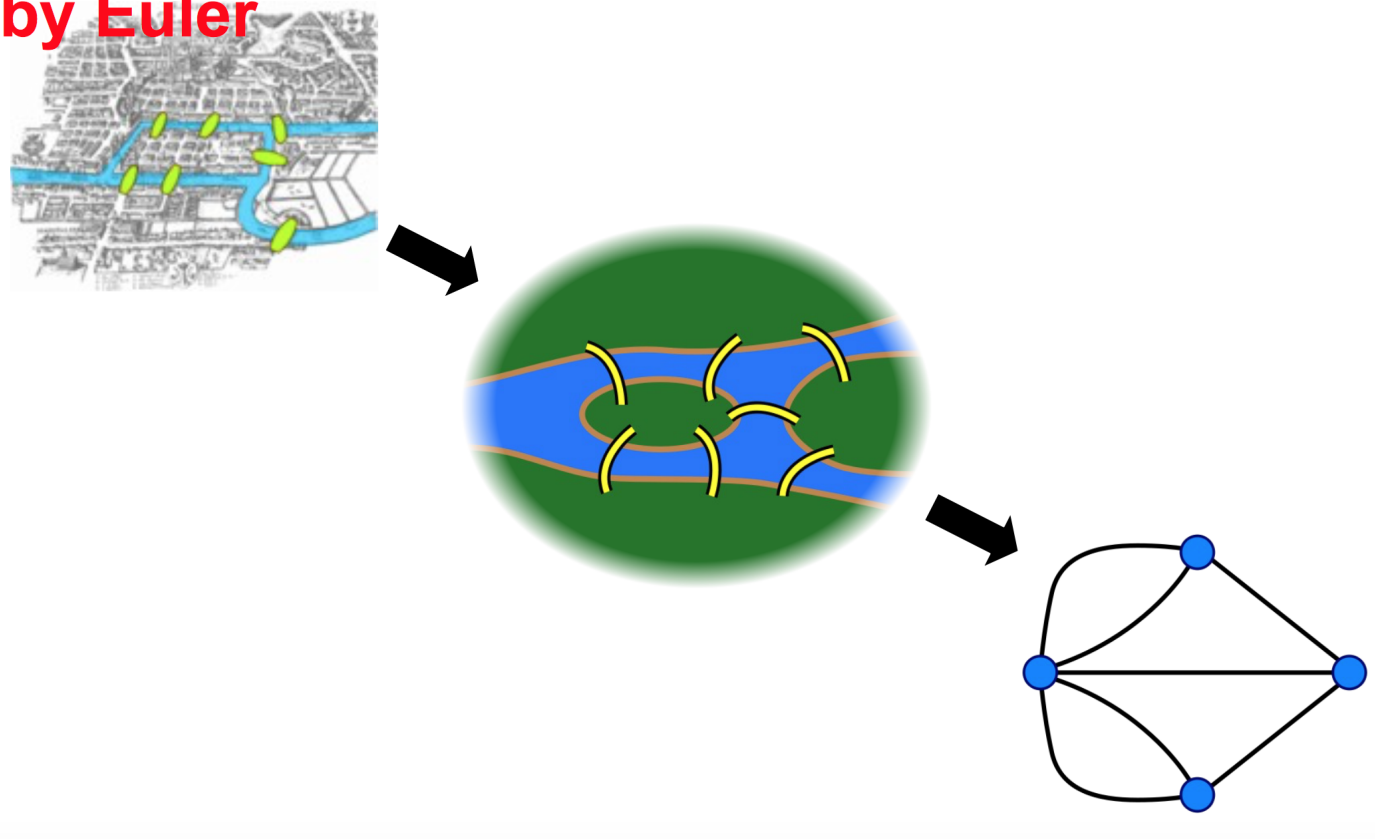
Adjacency List

(1, 2, 20)
(2, 3, 10)
(2, 4, 50)
(3, 4, 10)
(1, 4, 100)

Edge List

Classical Problems on Graphs

The Seven Bridges of Königsberg Problem by Euler



Source: Wikipedia (Königsberg)

A bunch others

- Graph colouring
  - Most famous map 4-coloring problem
- Independent Set
- Vertex/ Edge Cover
- Max Flow
- Max Matching
- …

Classical Problems

- Mathematical objects focused
- Falls in a branch called “Graph Theory”
- Usually pursue clean and exact solution
- Fun for brain training but less used in today’s “data science”

Modern Problems on Graphs

- Finding shortest paths
  - Routing Internet traffic and UPS trucks
- Finding minimum spanning trees
  - Telco laying down fiber
- Finding Max Flow
  - Airline scheduling
- Identify “special” nodes and communities
  - Breaking up terrorist cells, spread of avian flu
- Bipartite matching
  - Monster.com, Match.com
- And of course... PageRank

Source: Prof. Wing Cheong Lau, CUHK, ENGG4030

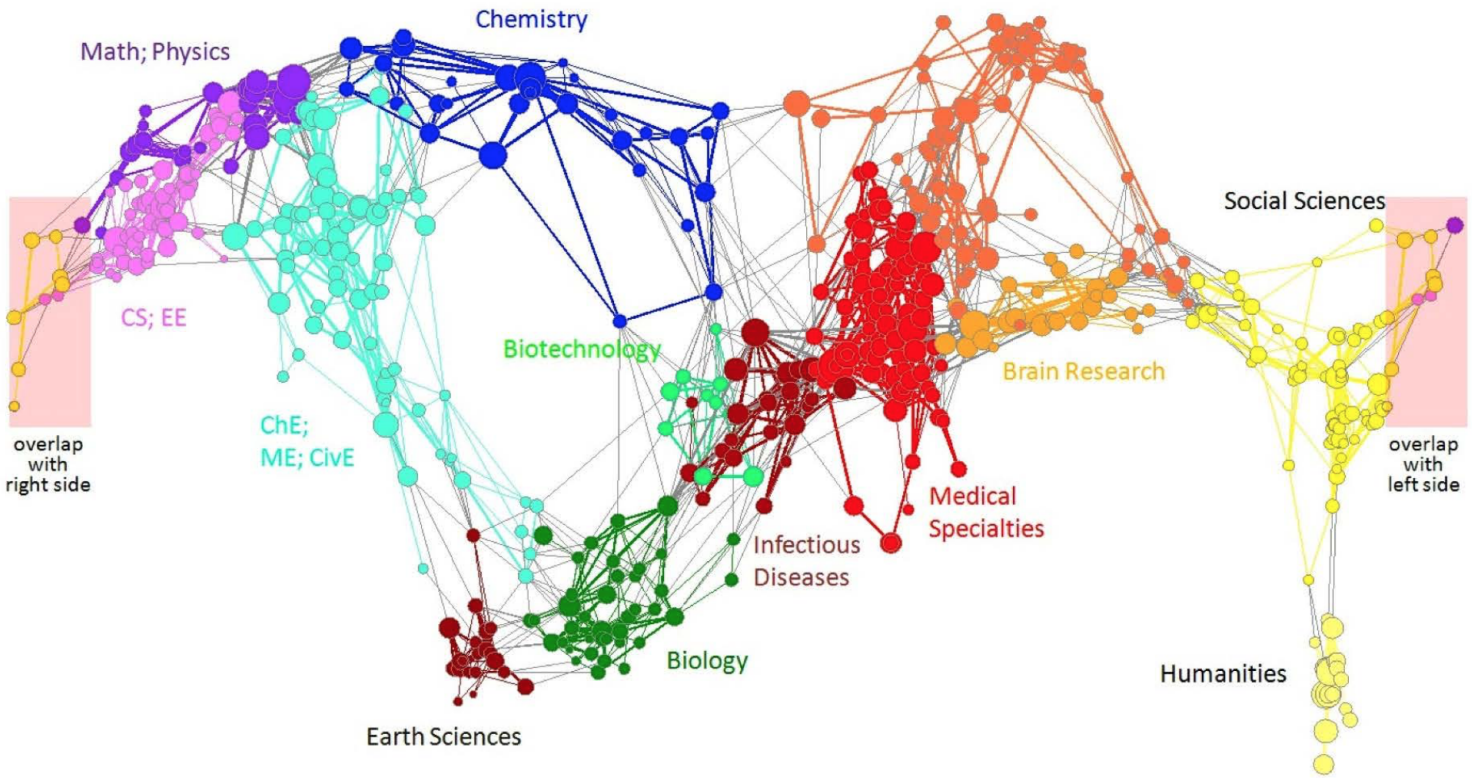
Shortest Path and Degree of Separation

Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Source: Jure Leskovec, CS224W

Community Detection

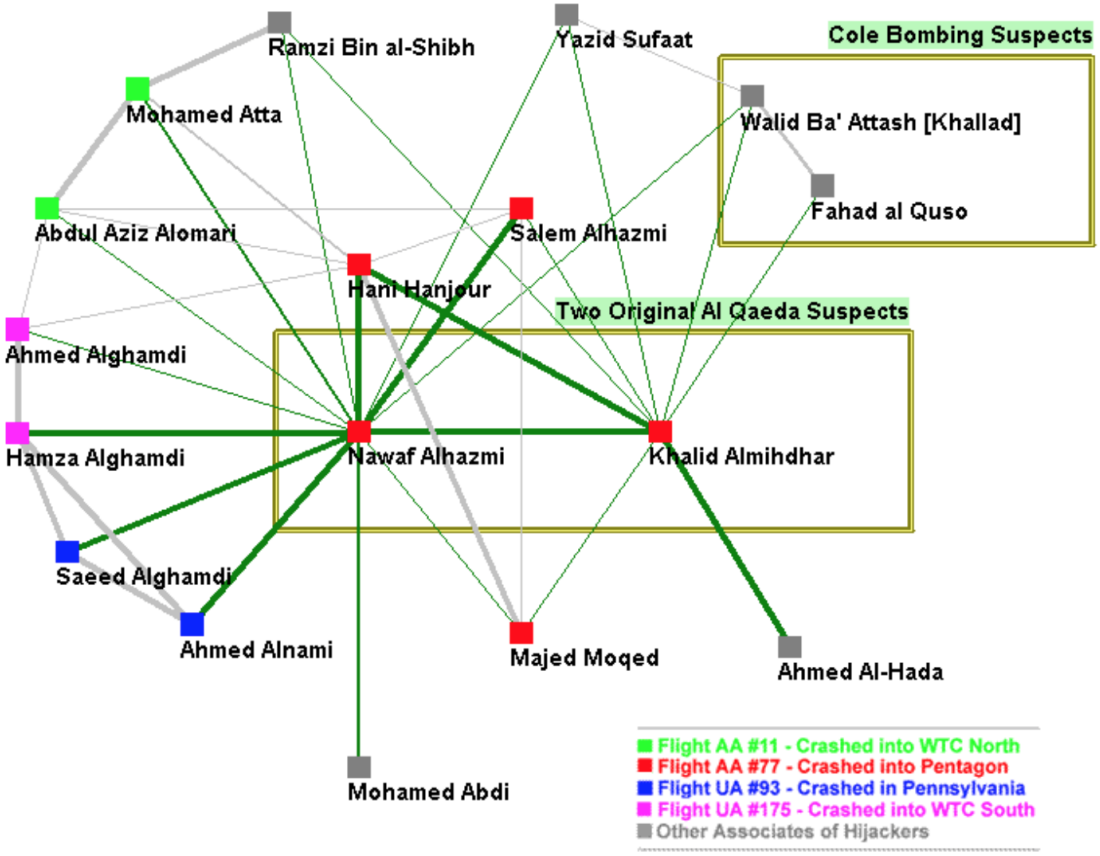


Citation networks and Maps of science

[Börner et al., 2012]

Source: Jure Leskovec, CS224W

Network Role Identification

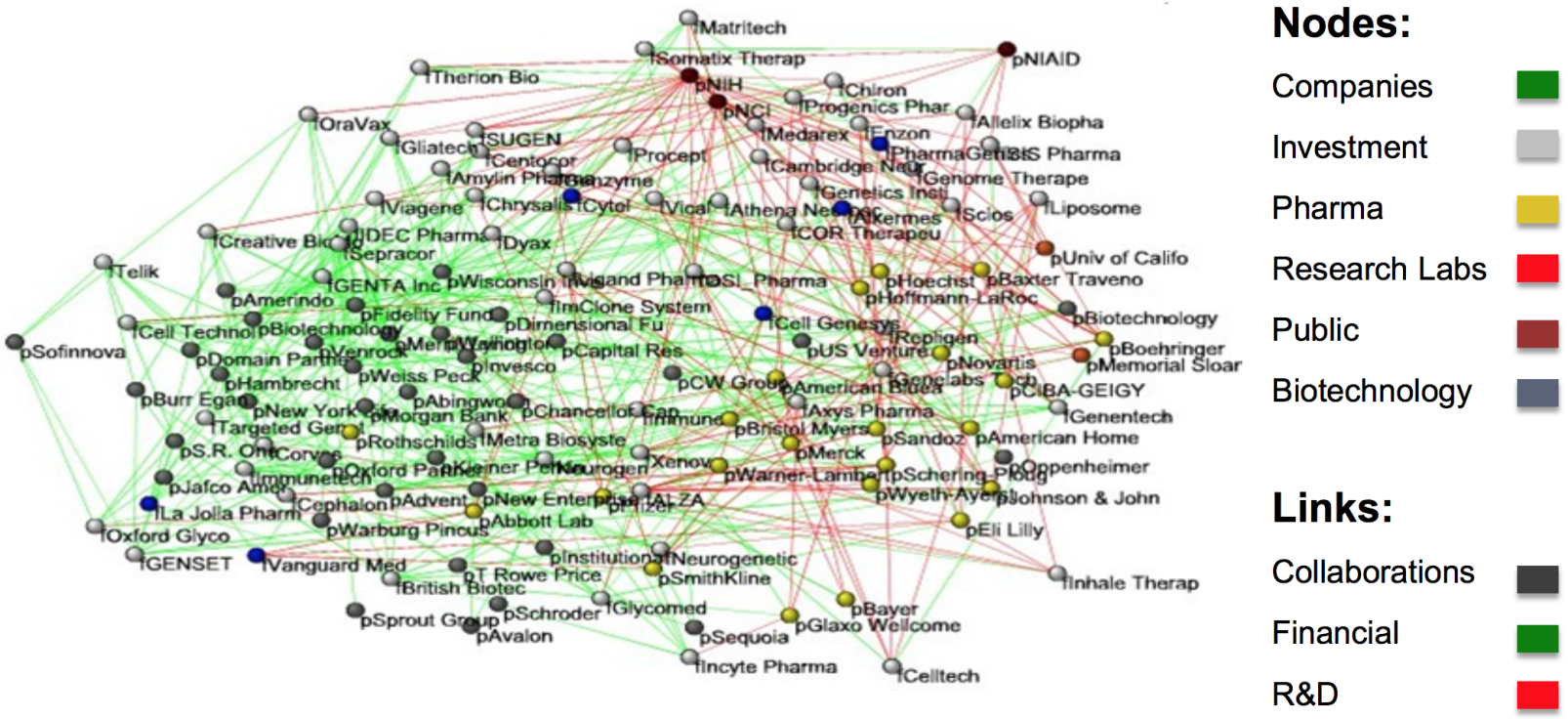


9/11 terrorist network

[Krebs, 2002]

Source: Jure Leskovec, CS224W

Graph Visualisation & Interactive Exploration



Bio-tech companies

[Powell-White-Koput, 2002]

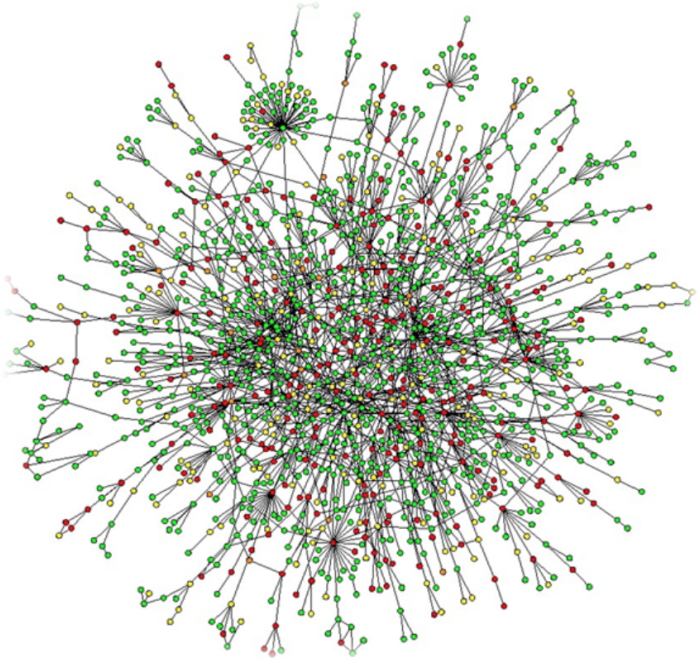
Source: Jure Leskovec, CS224W

Neuron Network
& Artificial Neuron Network

Human brain has between
10-100 billion neurons
[Sporns, 2011]

Source: Jure Leskovec, CS224W

Interaction Network



Protein-Protein Interaction Networks:

Nodes: Proteins

Edges: 'physical' interactions

Same can be applied on stocks

Source: Jure Leskovec, CS224W

Common Graph Algorithms

Will be left to the hands-on workshop.

Other learning resources

- My slides for CUHK ENGG4030 course: project.hupili.net/engg4030/
  - Graph Analysis Basics:
    - http://project.hupili.net/engg4030/t9-graph/
  - GraphLab for parallel machine learning:
    - http://project.hupili.net/engg4030/t10-bindings/
    - http://project.hupili.net/engg4030/t11-graphlab/
-

More questions to me

e@hupili.net

- https://github.com/hupili
- https://twitter.com/hupili
- https://facebook.com/hupili
- http://weibo.com/impige