Thanks to Jure Leskovec, Stanford and Panayiotis Tsaparas, Univ. of Ioannina for material in the slides
The Future Is Big Graphs

Managing IT Professional Turnover

An Internet of Things Service Roadmap

Whose Smartphone Is It?

Q&A with ACM Computing Prize Winner David Silver
what is a network or a graph?
Network Components

- **Network** (or **Graph**) \( G(N,E) \)
  - **Objects**: nodes (vertices) \( N \)
  - **Relationships**: links (edges) \( E \)

Built on the mathematics of **graph theory**
networks are ubiquitous
World economy
Human cell
Railroads
Internet
Friends & Family
Media & Information
What do the following things have in common?
Complex systems that can be modeled as Networks!
Behind many systems there is an intricate wiring diagram, a network, that defines the interactions between the components.

We will never understand these systems unless we understand the networks behind them!
But, why should we care about networks? Why now?
Why Networks?

Universal language for describing complex data
## Networks: Why Now?

### Age and size of networks

<table>
<thead>
<tr>
<th>Node Count</th>
<th>Social Network</th>
<th>WWW</th>
<th>Social Network Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^9$ nodes</td>
<td><img src="image1" alt="Social Network" /></td>
<td><img src="image2" alt="WWW" /></td>
<td><img src="image3" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^8$ nodes</td>
<td><img src="image4" alt="Social Network" /></td>
<td><img src="image5" alt="WWW" /></td>
<td><img src="image6" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^7$ nodes</td>
<td><img src="image7" alt="Social Network" /></td>
<td><img src="image8" alt="WWW" /></td>
<td><img src="image9" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^6$ nodes</td>
<td><img src="image10" alt="Social Network" /></td>
<td><img src="image11" alt="WWW" /></td>
<td><img src="image12" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^5$ nodes</td>
<td><img src="image13" alt="Social Network" /></td>
<td><img src="image14" alt="WWW" /></td>
<td><img src="image15" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^4$ nodes</td>
<td><img src="image16" alt="Social Network" /></td>
<td><img src="image17" alt="WWW" /></td>
<td><img src="image18" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^3$ nodes</td>
<td><img src="image19" alt="Social Network" /></td>
<td><img src="image20" alt="WWW" /></td>
<td><img src="image21" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^2$ nodes</td>
<td><img src="image22" alt="Social Network" /></td>
<td><img src="image23" alt="WWW" /></td>
<td><img src="image24" alt="Social Network Sites" /></td>
</tr>
<tr>
<td>$10^1$ nodes</td>
<td><img src="image25" alt="Social Network" /></td>
<td><img src="image26" alt="WWW" /></td>
<td><img src="image27" alt="Social Network Sites" /></td>
</tr>
</tbody>
</table>

- $10^9$ years to $10^0$ years
Networks: Size Matters

- **Network data: Orders of magnitude**
  - 436-node network of email exchange at a corporate research lab [Adamic-Adar, SocNets ‘03]
  - 43,553-node network of email exchange at an university [Kossinets-Watts, Science ‘06]
  - 4.4-million-node network of declared friendships on a blogging community [Liben-Nowell et al., PNAS ‘05]
  - 240-million-node network of communication on Microsoft Messenger [Leskovec-Horvitz, WWW ‘08]
  - 800-million-node Facebook network [Backstrom et al. ‘11]
How can we *study* networks?
Network Analysis

network analysis helps to reveal the underlying dynamics of these systems, not easily observable before
what do we study in networks?
Networks: Structure & Process

- Structure and evolution
  - What is the structure of a network?
  - Why and how did it become to have such structure?

- Processes and dynamics
  - Networks provide “skeleton” for spreading of information, behavior, diseases
how do we reason about networks?
Reasoning About Networks

- **Empirical studies/properties**: Study network data to find organizational principles

- **Mathematical models**: Probabilistic, graph theory

- **Algorithms**: Methods for analyzing graphs
Properties

- Six degrees of separ.
- Power-law degrees
- Strength of weak ties
- Densif. power law, Shrinking diameter

![Graph showing power-law distribution](image1)

![Graph showing shrinking diameter](image2)

Probability, $P_k = P(X=k)$

$P_k \approx k^{-1.75}$

$\alpha = 1.2$
Models

- Erdös-Renyi model
- Small-world model
- Community model
- Cascade model
Algorithms

- Decentralized search
- Link analysis
- Link prediction
- Community detection
Map of Superpowers

Properties
- Small diameter, Edge clustering
- Scale-free
- Strength of weak ties, Core-periphery
- Densification power law, Shrinking diameters
- Information virality, reproductive number

Models
- Small-world model, Erdös-Renyi model
- Preferential attachment, Copying model
- Community-affiliation Graph Model
- Microscopic model of evolving networks
- Independent cascade model, Game theoretic model, SIR

Algorithms
- Decentralized search
- PageRank, Hubs and authorities
- Community detection: Girvan-Newman, Modularity
- Link prediction, Supervised random walks
- Influence maximization, Outbreak detection, LIM
Applying Our Superpowers

- Social media analytics
- Viral marketing
Applying Our Superpowers

- Predicting epidemics: Ebola
- Drug design
examples of network studies
Networks: Social

Facebook social graph
4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]
Networks: Communication

Graph of the Internet (Autonomous Systems)
Power-law degrees [Faloutsos-Faloutsos-Faloutsos, 1999]
Robustness [Doyle-Willinger, 2005]
Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]
Seven Bridges of Königsberg

[Euler, 1735]

Return to the starting point by traveling each link of the graph once and only once.
Citation networks and Maps of science
[Börner et al., 2012]
Networks: Knowledge

Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]
Networks: Economy

Bio-tech companies

[Powell-White-Koput, 2002]
Human brain has between
~100 billion neurons, ~1,000 trillion synapses
[Sporns, 2011]
Protein-Protein Interaction Networks:
Nodes: Proteins
Edges: ‘physical’ interactions

Metabolic networks:
Nodes: Metabolites and enzymes
Edges: Chemical reactions
Web – The Lab for Humanity

The Web is a “laboratory” for understanding the pulse of humanity.
examples of network analysis impact
Networks: Impact

- **Google (Australia?)**
  Market cap: $1700 billion

- **Cisco (Greece?)**
  Market cap: $230 billion

- **Meta (Taiwan?)**
  Market cap: $770 billion
Networks: Impact

- Predicting epidemics

Real cases as of May 24, 2009
Fraction of max number of cases per cell
- < 1%
- 1% - 10%
- 10% - 50%
- 50% - 75%
- 75% - 100%

Worst case scenario
Calibration up to May 6
Average number of cases
Fraction of max number of cases per cell
May 24, 2009
- < 1%
- 1% - 10%
- 10% - 50%
- 50% - 100%
- 75% - 100%

Real
Predicted
If you want to understand the spread of diseases, *can you do it without social networks?*

If you want to understand the structure of the Web, *it is hopeless without working with the Web’s topology*

If you want to understand dissemination of news or evolution of science, *it is hopeless without considering the information networks*
EECS4414/5414 Administrivia
Logistics: Communication

- **Website**
  - [http://www.eecs.yorku.ca/~papaggel/courses/eecs4414/](http://www.eecs.yorku.ca/~papaggel/courses/eecs4414/)

- **eClass/Moodle**
  - [https://eclass.yorku.ca/course/view.php?id=88650](https://eclass.yorku.ca/course/view.php?id=88650)

- **Piazza Q&A website**
  - [https://piazza.com/yorku.ca/fall2023/eecs44145414](https://piazza.com/yorku.ca/fall2023/eecs44145414)
  - You need to register with your *yorku.ca* email
    - Please participate and help each other!

- **e-mail for personal issues**
  - papaggel@eecs.yorku.ca
Prerequisites

- **Course Prerequisites**
  - EECS-3421: Introduction to Database Systems
  - EECS-3101: Design and Analysis of Algorithms
  - MATH-2030: Elementary Probability
  - General prerequisites

- **No single topic in the course is too hard by itself**
- **But we will cover and touch upon many topics and this is what makes the course hard**

- **Good background in:**
  - Algorithms and graph theory
  - Probability and Statistics
  - Linear algebra

- **Programming:**
  - You should be able to write non-trivial programs (in Python or C)
Course Intellectual Content

- Social & Information Network Analysis
- Network Science
- Graph Theory
- Social Science & Economics
- Data Analysis
- Probability & Statistics
Topics Covered

Component I
basic graph theory, network measurements, network models

Component II
link analysis, link prediction, network ties, community detection, graph partitioning

Component III
information cascades & epidemics, graph mining, machine learning with graphs, influence maximization, connections to problems in the social sciences and economics
“Suggested” Textbooks

- Networks
- Crowds
- Markets

Reasoning about a Highly Connected World
David Easley
and
Jon Kleinberg

Mining of Massive Datasets
Jure Leskovec
Anand Rajaraman
Jeffrey David Ullman

Second Edition

+ a few more reference books
+ recent research papers on topics covered
# Coursework

<table>
<thead>
<tr>
<th>Work</th>
<th>Weight</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Assignments</td>
<td>20%</td>
<td>A1: 10%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2: 10%</td>
</tr>
<tr>
<td>Research Project (team large project + report in research paper format)</td>
<td>40%</td>
<td>Proposal: 10%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Project milestone: 20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Final report &amp; code: 50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In-class presentation: 20%</td>
</tr>
<tr>
<td>Final Exam*</td>
<td>40%</td>
<td>Final exam grade must be &gt; 40%</td>
</tr>
</tbody>
</table>

* There will be no final exam for grad students (marking scheme to be determined):  
  - either: Project: 80%, but **more substantial** and in **teams of two**  
  - or: Project: 60% + A lecture/presentation on a relevant research topic: 20%
Course Projects

- **Substantial course project:**
  - Experimental evaluation of algorithms and models on an interesting network dataset
  - A theoretical project that considers a model, an algorithm and derives a rigorous result about it
  - Develop scalable algorithms for massive graphs

- **Performed in groups of up to 4**
  - Graduate students: team of up to 2
  - Project is the main work for the class
    - We will help with ideas, and mentoring
    - Need to start thinking about this now

- Class presentation
Network Analysis Tools

- **Highly recommend SNAP:**
  - **SNAP C++:** more challenging but more scalable
  - **SNAP.PY:** Python ease of use, most of C++ scalability
- **Other tools include:**
  - NetworkX
  - JUNG
  - iGraph
  - ...
Example Research Questions/Topics
Topics

- Measuring real networks
- Modeling the evolution of networks
- Identifying important nodes in the graph
- Finding communities in graphs
- Link prediction and recommendation
- Modeling information cascades in networks
- ...
What does a network look like?

- Measure different properties to understand the structure

degree of nodes

Triangles in the graph
Modeling Real Networks

- Real life networks are not “random”
- Can we define a model that generates graphs with statistical properties similar to those in real life?
- The rich-get-richer model

We need to accurately model the mechanisms that govern the evolution of networks (for prediction, simulations, understanding)
Is my home page as important as the Facebook page?

We need algorithms to compute the importance of nodes in a graph.

The PageRank Algorithm

- A success story of network use.

It is impossible to create a web search engine without understanding the web graph.
Given a snapshot of a social network at time $t$, we seek to accurately predict the edges that will be added to the network during the interval from time $t$ to a given future time $t'$. 

Applications
- Accelerate the growth of a social network (e.g., Facebook, LinkedIn, Twitter)
- Maximize information cascades

How do we predict future links?
What is community?

“Cohesive subgroups are subsets of actors among whom there are relatively strong, direct, intense, frequent, or positive ties.” [Wasserman & Faust ‘97]

Karate club example [W. Zachary, 1970]
How do viruses spread between individuals? How can we stop them?
How does information propagates in social and information networks? What items become viral? Who are the influencers and trend-setters?
We need models and algorithms to answer these questions

Online advertising relies heavily on online social networks and word-of-mouth marketing. There is currently need for models for understanding the spread of Covid-19 virus.
Social Media (Twitter, Facebook, Instagram) have supplanted the traditional media sources
- Information is generated and disseminated by users

Interesting problems:
- Automatically detect events using Twitter
  - Earthquake response
  - Crisis detection and management
- Sentiment mining
- Track the evolution of events: socially, geographically, over time
- ...
Research in Graph Mining

- **Current hot research topics:**
  - Graph representation learning
  - Graph neural networks
  - Graph attention mechanisms
  - Graph generative models
  - Graph classification, clustering, anomaly detection
  - Dynamic graph analysis and mining

- **Relevant research conferences**
  - Data Mining: KDD, ICDM, WSDM, WWW, ...
  - ML: ICML, NeurIPS, ECML/PKDD, ...
Starter Topic: Structure of the Web Graph
Network is a collection of objects where some pairs of objects are connected by links.

What is the structure of the network?
Components of a Network

- **Objects**: nodes, vertices
- **Interactions**: links, edges
- **System**: network, graph

$G(N,E)$
Networks or Graphs?

- **Network** often refers to real systems
  - Web, Social network, Metabolic network
  
  **Language:** Network, node, link

- **Graph** is mathematical representation of a network
  - Web graph, Social graph (a Facebook term)
  
  **Language:** Graph, vertex, edge

We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably.
Networks: Common Language

- Actor 1
  - Movie 1
- Actor 2
  - Movie 2
  - Movie 3
- Actor 3
- Actor 4
- Peter
  - friend
  - brothers
- Mary
  - co-worker
- Albert
- Tom
- Protein 1
- Protein 2
- Protein 5
- Protein 9

|N|=4
|E|=4
How to build a graph:
- What are nodes?
- What are edges?

Choice of the proper network representation of a given domain/problem determines our ability to use networks successfully:
- In some cases there is a unique, unambiguous representation
- In other cases, the representation is by no means unique
- The way you assign links will determine the nature of the question you can study
Choosing Proper Representation

- If you connect individuals that work with each other, you will explore a **professional network**
- If you connect those that have a sexual relationship, you will be exploring **sexual networks**
- If you connect scientific papers that cite each other, you will be studying the **citation network**

- If you connect all papers with the same word in the title, you will be exploring what? It is a network, nevertheless
**Undirected**
- **Links:** undirected (symmetrical, reciprocal)

**Directed**
- **Links:** directed (arcs)

**Examples:**
- Collaborations
- Friendship on Facebook

**Examples:**
- Phone calls
- Following on Twitter
Connectivity of Graphs

- **Connected (undirected) graph:**
  - Any two vertices can be joined by a path
  - A disconnected graph is made up by two or more connected components

Bridge edge: If we erase it, the graph becomes disconnected.
Articulation point: If we erase it, the graph becomes disconnected.
Connectivity of Directed Graphs

- **Strongly connected directed graph**
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)

- **Weakly connected directed graph**
  - is connected if we disregard the edge directions

Graph on the left is connected but not strongly connected (e.g., there is no way to get from F to G by following the edge directions).
Q: What does the Web “look like”?

Here is what we will do next:

- We will take a real system (i.e., the Web)
- We will represent the Web as a graph
- We will use language of graph theory to reason about the structure of the graph
- Do a computational experiment on the Web graph
- Learn something about the structure of the Web!
Q: What does the Web “look like” at a global level?

- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks

- **Side issue:** What is a node?
  - Dynamic pages created on the fly
  - “dark matter” – inaccessible database generated pages
In early days of the Web links were navigational
Today many links are transactional
The Web as a Directed Graph

I'm a student at Univ. of X
- My song lyrics

Univ. of X
- Classes
- I teach at Univ. of X

I'm applying to college
- USNews College Rankings
- USNews Featured Colleges

Networks
- Networks class blog

Blog post about college rankings
- Blog post about Company Z
What Does the Web Look Like?

- How is the Web linked?
- What is the “map” of the Web?

**Web as a directed graph** [Broder et al. 2000]:

- Given node $v$, what can $v$ reach?
- What other nodes can reach $v$?

For example:

$\text{In}(A) = \{A,B,C,E,G\}$
$\text{Out}(A) = \{A,B,C,D,F\}$
Two types of directed graphs:

- **Strongly connected:**
  - Any node can reach any node via a directed path
  
  \[ \text{In}(A) = \text{Out}(A) = \{A, B, C, D, E\} \]

- **DAG – Directed Acyclic Graph:**
  - Has no cycles: if \( u \) can reach \( v \), then \( v \) cannot reach \( u \)

Any directed graph can be expressed in terms of these two types!
**Strongly connected component (SCC)**
is a set of nodes $S$ so that:

- Every pair of nodes in $S$ can reach each other
- There is no larger set containing $S$ with this property

![Graph with strongly connected components: \{A,B,C,G\}, \{D\}, \{E\}, \{F\}](image)
There is a single giant SCC
  - That is, there won’t be two SCCs

Heuristic argument:
  - It just takes 1 page from one SCC to link to the other SCC
  - If the 2 SCCs have millions of pages the likelihood of this not happening is very very small
Broder et al., 2000:
- Altavista crawl from October 1999
  - 203 million URLs
  - 1.5 billion links
- Computer: Server with 12GB of memory

Undirected version of the Web graph:
- 91% nodes in the largest weakly conn. component
- Are hubs making the web graph connected?
  - Even if they deleted links to pages with in-degree >10
    WCC was still ≈50% of the graph
Directed version of the Web graph:

- **Largest SCC:** 28% of the nodes (56 million)
- Taking a random node \( v \)
  - \( \text{Out}(v) \approx 50\% \) (100 million)
  - \( \text{In}(v) \approx 50\% \) (100 million)

What does this tell us about the conceptual picture of the Web graph?
203 million pages, 1.5 billion links [Broder et al. 2000]
**What did We Learn/Not Learn?**

- **What did we learn:**
  - Some conceptual organization of the Web (i.e., the bowtie)

- **What did we not learn:**
  - Treats all pages as equal
    - Google’s homepage == my homepage
  - What are the most important pages
    - How many pages have $k$ in-links as a function of $k$?
      - The degree distribution: $\sim k^{-2}$
    - Link analysis ranking -- as done by search engines (PageRank)
  - Internal structure inside giant SCC
    - Clusters, implicit communities?
  - How far apart are nodes in the giant SCC:
    - Distance = # of edges in shortest path
    - $\text{Avg} = 16$  [Broder et al.]