Why Networks?

Networks are a language for representing, describing, and understanding interconnected systems.
A Network!
How to understand the behaviours, decisions, beliefs, etc. of millions of people?
Social network
Educational settings: people learn and interact with each other in complex ways
EMOTIONS MAPPED
BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

1st grade  4th grade  8th grade

Moreno’s sociograms, 1934
Everything on the Internet is passed through autonomous systems (routers, etc.)
Graph of the Internet (Autonomous Systems)
In companies, people communicate and work together in large hierarchies and structures.
Email communication network
(HP Research, 436 employees)
Org chart (left) and Email communication network (right)
Microsoft, 200,000 employees [Jake Hofman, 2018]
Many diseases are transmitted socially (e.g. sexually transmitted diseases)
CHAPTER 2. GRAPHS

Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].

The subject of even the most intense gossip and scrutiny. Nevertheless, they are real: like social facts, they are invisible yet consequential macrostructures that arise as the product of individual agency.

2.3 Distance and Breadth-First Search

In addition to simply asking whether two nodes are connected by a path, it is also interesting in most settings to ask how long such a path is — in transportation, Internet communication, or the spread of news and diseases, it is often important whether something flowing through a network has to travel just a few hops or many.

To be able to talk about this notion precisely, we define the length of a path to be the number of steps it contains from beginning to end — in other words, the number of edges in the sequence that comprises it. Thus, for example, the path mit, bbn, rand, ucla in Figure 2.3 has length three, while the path mit, utah has length one. Using the notion of American high school sexual contact network (18-month period, self-identified gender)
Power is transported everywhere with interconnected stations and lines
The power grid: a network
Science is a complex system of academics working together and being influenced by each other.
Citation networks and maps of science [Börner et al., 2012]
A single person’s interactions with friends and family are a huge part of their life.
3.4. TIE STRENGTH, SOCIAL MEDIA, AND PASSIVE ENGAGEMENT

Figure 3.8: Four different views of a Facebook user’s network neighborhood, showing the structure of links corresponding respectively to all declared friendships, maintained relationships, one-way communication, and reciprocal (i.e. mutual) communication. (Image from [286].)

Notice that these three categories are not mutually exclusive — indeed, the links classified as reciprocal communication always belong to the set of links classified as one-way communication.

This stratification of links by their use lets us understand how a large set of declared friendships on a site like Facebook translates into an actual pattern of more active social interaction, corresponding approximately to the use of stronger ties. To get a sense of the relative volumes of these different kinds of interaction through an example, Figure 3.8 shows the network neighborhood of a sample Facebook user — consisting of all his friends, and all links among his friends. The picture in the upper-left shows the set of all declared friendships in this user’s profile; the other three pictures show how the set of links becomes sparser once we consider only maintained relationships, one-way communication, or reciprocal communication. An “ego network”: the neighbourhood around a single individual.
CHAPTER 1. OVERVIEW

Figure 1.3: The network of loans among financial institutions can be used to analyze the roles that different participants play in the financial system, and how the interactions among these roles affect the health of individual participants and the system as a whole. The network here is annotated in a way that reveals its dense core, according to a scheme we will encounter in Chapter 13. (Image from Bech and Atalay [50].)

Behavior and Dynamics. But the structure of the network is only a starting point. When people talk about the "connectedness" of a complex system, in general they are really talking about two related issues. One is connectedness at the level of structure — who is linked to whom — and the other is connectedness at the level of behavior — the fact that each individual's actions have implicit consequences for the outcomes of everyone in the system.

This means that in addition to a language for discussing the structure of networks, we also need a framework for reasoning about behavior and interaction in network contexts. And just as the underlying structure of a network can be complex, so too can the coupled behavior of its inhabitants. If individuals have strong incentives to achieve good outcomes, the Economy is a network: e.g. Federal funds overnight lending.
Transportation network (US only)
Political blogs prior to 2004 US Presidential election
The human brain has between 10-100 billion neurons connected to each other in complex ways.
Many, many more examples
But why should *I* care about networks?
Why study networks?

Networks are a universal language for describing complex data

Networks from science, nature, and technology are more similar than you might expect

**Shared vocabulary** between fields

- CS, finance, tech, social sciences, physics, economics, statistics, biology

**Data availability** (and computational challenges)

- Web/mobile, bio, health, medical

**Impact!**

- Social networking, social media, drug design
Why study networks?

Complex systems are all around us

- **Society** is a collection of seven billion people
- **Communications systems** link electronic devices
- **Information and knowledge** is organized and linked
- Interactions between thousands of **genes** regulate life
- Our thoughts and selves are manifested in the connections between billions of neurons in the **brain**
- **Information and diseases** spread from person to person
Why now?

Age and size of networks
Networks: Impact

Google
Market cap: $815B

Facebook
Market cap: $500B

Cisco
Market cap: $220B
Networks and Applications
Ways to Analyze Networks

**Predict** the type of a given node (*node classification*)

**Predict** whether two nodes are **linked** (*link prediction*)

**Identify** densely **linked clusters** of nodes (*community detection*)

**Predict** common **pathways** (*social influence/propagation*)

**Measure** similarity between nodes/networks (*network similarity*)
(1) Networks: Social
Application: Friend Prediction
(2) Networks: Infrastructure

The August 2003 blackout occurred on August 14, 2003, and lasted for 20 hours. The area affected included major cities such as Montreal, Ottawa, Toronto, Albany, Boston, and Long Island. The images show the network coverage before and after the blackout, with a significant decrease in illumination indicating power loss. The blackout lasted until August 15, 2003, at 9:14pm, marking the end of the disruption.
(2) Networks: Infrastructure

This illustrates two important themes of this class:

• We must **understand** how network structure affects the system

• We will **develop quantitative tools** to assess the interplay between network structure and the dynamic processes that happen on networks

• We will learn that in reality failures follow reproducible laws, and can be **quantified**, and to some extent **predicted**, using the language of network analysis
(3) Networks: Information and Knowledge

Knowledge Graphs

Heterogeneous Graphs

Multimodal Graphs
Application: Web Search

How do you go from a tiny text string to the 10 most relevant sites out of billions of pages?
(4) Networks: Online Media

Connections between political blogs
Application: Polarization on Twitter

Retweet networks: polarized (left) and unpolarized (right)
Application: Understanding Virality

Information cascades in networks
Application: Product Adoption

Invitation cascades: 60–90% of LinkedIn users signed up due to an invitation from another user

[Anderson et al., WWW ’15]
If you want to predict the spread of a disease, you need to know who is in contact with whom.

If you want to understand the structure of the Web (or serve relevant search results), you have to analyze the links.

If you want to understand the dissemination of news or the evolution of science, you have to follow the flow.
About CSCC46
Ways to Analyze Networks

What do we hope to achieve from studying networks?

**Develop** the language of interconnectedness

**Learn** the patterns and statistical properties of network data

**Understand** design principles and models of networks

**Develop** algorithmic understanding of processes in networked systems
Networks: Structure, Dynamics, Incentives

What do we study in networks?

**Structure and evolution**

What is the structure of networks?

Why and how do they come to have such structure?

How do we harness the structure to extract useful information?

**Processes and dynamics**

Networks are the “skeleton” on which information, behaviours, and diseases spread

How do information and diseases spread?

**Incentives in networks**

Behaviour is interconnected by depending on what others do

How do decisions and behaviours depend on network structure and dynamics?
Reasoning about Networks

How do we reason about networks and collective behaviour?

**Empirical analyses**: Study network data to find organizational principles

How do we measure and quantify networks?

**Mathematical models**: Graph theory and statistical models

Models allow us to understand behaviors and distinguish surprising from expected phenomena

**Algorithms for analyzing graphs**: Computer science toolkit

Overcome hard computational challenges to solve important problems
Main Tool: Graph Theory

Seven Bridges of Königsberg [Euler, 1735]

We’ll make extensive use of graph theory in this course
Main Tool: Game Theory

The mathematical theory of strategic interaction

We’ll also make extensive use of game theory in this course
(Tentative) Course Overview

**Week 1:** Course overview, Introduction to graph theory, The Web as a Network

**Week 2:** Network Representations, Affiliation, Homophily, Strong and Weak Ties, Structural Holes

**Week 3:** The Friendship Paradox, Signed Networks and Structural Balance

**Week 4:** Small Worlds, $G_n$, Decentralized Search

**Week 5:** Link Analysis & Search

**Week 6:** Power Laws and Rich-Get-Richer Phenomena

**Week 7:** Game Theory

**Week 8:** Congestion, Information Cascades, Cascading Behaviour

**Week 9:** Epidemics, Information Diffusion

**Week 10:** Voting

**Week 11:** Community Detection and Stable Matching

**Week 12:** Review
Course evaluation

35%  4 assignments

10%  2 blog posts

5%   Participation/Collegiality

50%  Final exam

Assignments are due on Mondays at 11am

Because I understand sometimes stuff happens, you get 4 "flex days": 1 flex day is a 24-hour period that you can hand in assignments late with no penalty. You can use up to 2 flex days per assignment. After that, no late assignments will be accepted.
Blog posts

• During the term, write two blog posts on the course blog about topics related to the course

• Should be fun! Pick your favourite topic and explore it in more detail

• Short essay-like posts aimed at your peers

• Commenting on others’ posts counts as participation
TAs

- Kasra Rahmani (kasra.rahmani@mail.utoronto.ca)
- George Lifchits (glifchits@cs.toronto.edu)
- Lillio Mok (lillio@cs.toronto.edu)

Very talented students! Feel free to go to them for help.
Course resources

- Course webpage (http://www.cs.toronto.edu/~ashton/csc46/)
- Quercus (course announcements, discussion, course materials)
- MarkUs (assignment submission)
  - All assignments must be submitted electronically in PDF format
- Office hours: M 2:30–3:30pm in IC 481
Textbook

- “Networks, Crowds & Markets” Easley & Kleinberg
- Available free online / reasonably-priced hardcover
- Very readable, engaging text
- Some assignment exercises from the book
Tutorials

• Time: Tuesday 10–11 and 11–12

• Required, hosted by Kasra

• Mostly working through concrete examples as a group, Q&A (group office hours), and assignment help
Questions?
First topic: Network Analysis Fundamentals and The Structure of the Web
A network is a collection of objects where some pairs of objects are connected by links.
Components of a Network

**Objects:** nodes, vertices  \[ N \]

**Interactions:** links, edges  \[ E \]

**System:** network, graph  \[ G(N,E) \]
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
A first example

The Internet in 1970
A first example

The Internet in 1970
Networks: a shared language

|N|=4
|E|=4

Protein 1 → Protein 2 → Protein 5 → Protein 9

Actor 1 → Actor 2 → Actor 3 → Actor 4

Peter → friend → Mary → co-worker → Tom

Actor 1 → Movie 1 → Actor 2 → Movie 2 → Actor 4 → Movie 3

Actor 3 → Movie 2 → Actor 4

Actor 1 → Movie 1 → Actor 2 → Movie 3

Actor 3 → Movie 2 → Actor 4

Actor 1 → Movie 1 → Actor 2 → Movie 3

Actor 3 → Movie 2 → Actor 4

Actor 1 → Movie 1 → Actor 2 → Movie 3

Actor 3 → Movie 2 → Actor 4
Choosing a Proper Representation

- **How to build a graph:**
  - What are nodes?
  - What are edges?

- The 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 a Proper Representation

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

If you connect all people with first names that share the same first letter, what are you studying?
It is a network, but is it meaningful?
Undirected and Directed Networks

**Undirected**
- **Links:** undirected (symmetrical, reciprocal)

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

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

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

- **Connected component (undirected):**
  - Any two vertices can be joined by a path
  - No superset with the same property

- A disconnected graph is made up of 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

Is this graph weakly connected?
Strongly connected?
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

It is connected but not strongly connected (e.g., there is no way to get from F to G by following the edge directions)
What is the large-scale structure of the Web?
The Structure of the Web

Q: What does the Web “look like”?
The Structure of the Web

Q: What does the Web “look like”?
The Structure of the Web

- A network!
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
The Web as a Graph

CSCC46H

CMS Department at UTSC

University of Toronto
The Web as a Graph

- In early days of the Web links were **navigational**
- Today many links are **transactional**
The Web as a Directed Graph
Other Information Networks

Citations

References in an encyclopedia
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 $v$?

$In(v) = \{w \mid w \text{ can reach } v\}$
$Out(v) = \{w \mid v \text{ can reach } w\}$

For example:
$In(A) = \{?\}$
$Out(A) = \{?\}$
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:
\[
\begin{align*}
\text{In}(A) &= \{A,B,C,E,G\} \\
\text{Out}(A) &= \{A,B,C,D,F\}
\end{align*}
\]
Directed Graphs

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 \( v \) can not reach \( u \)

Any directed graph can be expressed in terms of these two types!
Strongly Connected Component

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

What are the strongly connected components of this graph?
Strongly Connected Component

- **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

Strongly connected components of the graph: \{A,B,C,G\}, \{D\}, \{E\}, \{F\}
Strongly Connected Component

**Fact:** Every directed graph is a DAG on its SCCs

1. SCCs partitions the nodes of $G$
   - That is, each node is in exactly one SCC
2. If we build a graph $G'$ whose nodes are SCCs, and with an edge between nodes of $G'$ if there is an edge between corresponding SCCs in $G$, then $G'$ is a DAG
Strongly Connected Component

Fact: Every directed graph is a DAG on its SCCs

1. SCCs partitions the nodes of $G$
   - That is, each node is in exactly one SCC
2. If we build a graph $G'$ whose nodes are SCCs, and with an edge between nodes of $G'$ if there is an edge between corresponding SCCs in $G$, then $G'$ is a DAG.

(1) Strongly connected components of graph $G$: $\{A, B, C, G\}, \{D\}, \{E\}, \{F\}$
(2) $G'$ is a DAG:
Claim: SCCs partitions nodes of G.

This means: Each node is member of exactly 1 SCC

Proof by contradiction:

Suppose there exists a node \( v \) which is a member of two SCCs \( S \) and \( S' \)

But then \( S \cup S' \) is one large SCC!

Contradiction!
Proof of (2)

Claim: $G'$ (graph of SCCs) is a DAG.
- This means: $G'$ has no cycles

Proof by contradiction:
- Assume $G'$ is not a DAG
- Then $G'$ has a directed cycle
- Now all nodes on the cycle are mutually reachable, and all are part of the same SCC
- But then $G'$ is not a graph of connections between SCCs (SCCs are defined as maximal sets)
- Contradiction!

Now $\{A, B, C, G, E, F\}$ is a SCC!
**Graph Structure of the Web**

- **Goal:** Take a large snapshot of the Web and try to understand how its SCCs “fit together” as a DAG

- **Computational issue:**
  - Want to find a SCC containing node $v$?
  - **Observation:**
    - $\text{Out}(v)$ … nodes that can be reached from $v$
    - SCC containing $v$ is: $\text{Out}(v) \cap \text{In}(v)$
      
      $$= \text{Out}(v,G) \cap \text{Out}(\overline{v},\overline{G}),$$
      
      where $\overline{G}$ is $G$ with all edge directions flipped
Out(A) ∩ In(A) = SCC

Example:

- Out(A) = {?} 
- In(A) = {?} 

Diagram: [Graph with nodes A, B, C, D, E, F, G, H, and arrows showing the relationships between the nodes.]

- Out(A) = {?} 
- In(A) = {?}
Out(A) \cap \text{In}(A) = \text{SCC}

Example:

- Out(A) = \{A,B,D,E,F,G,H\}
- In(A) = \{A,B,C,D,E\}
- Therefore, SCC(A) = \{A,B,D,E\}
How many “big” SCCs?
Graph Structure of the Web

How many “big” SCCs?

Giant SCC1

Giant SCC2
Graph Structure of the Web

- 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

![Diagram showing two SCCs linked by a single page](image)
Structure of the Web

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 connected 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
Structure of the Web

- 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?
Bow-tie Structure of the Web

203 million pages, 1.5 billion links [Broder et al. 2000]
What did we do?

Here is what we’ve already done
- We took a real system (the Web)
- We represented the Web as a graph
- We used the language of graph theory to reason about the structure of the graph
- We did a computational experiment on the Web graph
- Learned something about the structure of the Web!
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
  - Avg = 16 [Broder et al.]
Recap

Network analysis is the language of connectedness
- Represent real-world networks from many different domains as graphs, use graph theory and algorithms to reason about them
- Social networks, information networks, knowledge networks, biological networks, etc.

Network analysis fundamentals
- Nodes, edges, paths, cycles, un/directed, connected components (weak and strong)
- Choices of representation
- Every directed graph is a DAG on its SCCs

Structure of the Web
- Looks like a bow-tie: big giant component, IN & OUT components, tendrils, disconnected components
End!

This week: sign up for MarkUs, log in to Quercus, read Ch. 1, 2.1-2.4, 13.1-13.4

Next week: Network Representations, Affiliation, Homophily, Strong and Weak Ties, Structural Holes
Read Ch. 3.1-3.3, 4.1-4.3

Contact me: ashton@cs.toronto.edu