CSCC46 – Social and Information Networks
University of Toronto Scarborough, Fall 2019

Introduction to Networks

Networks are a language for representing, describing, and understanding interconnected systems. Describes complex data

- Shared vocabulary between fields (CS, finance, tech, social science, etc.)
- Data analysis and availability

Node Classification – predict the type of a given node
Link Prediction – predict whether two nodes are linked
Community Detection – identify densely linked clusters of nodes
Social Influence/Propagation – predict common pathways
Network Similarity – measure similarities between nodes and networks

---

A **network** is a collection of objects where some pairs of objects are connected by links

Objects – Nodes, vertices
Interaction – links, edges
System – network, graph

Networks refer to real-life systems (network, node, link)
Graphs are the mathematical representation of a network (Graph, vertex, edge)

Connected Component – two vertices joined by a path
Disconnected Graph – made up of two or more connected components
Bridge Edge – edge that if erased, the graph becomes disconnected
Articulation Point – node that if erased, the graph becomes disconnected
Strongly Connected Directed Graph – has a path from each node to every other node || **Weakly … -** if we disregard edge directions
In(V) – the set of nodes that can reach the node V, itself included
Out(V) – the set of nodes that can be reached by the node V, itself included
2 types of directed graphs:

- **Strongly Connected Graph –** any node can reach any node via a directed path
- **Directed Acyclic Graph (DAG) –** has no cycles (if u can reach v, v cannot reach u)

Strongly Connected Component (SCC) – 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
- Every directed graph is a **DAG** on its **SCC**s

Structure of web is that there is a single giant SCC. It only takes 1 page from one giant SCC to dual link to combine – **bowtie structure**

Measuring Networks and Models

We can represent networks two ways:

- **Edge List –** [(a,b), (b,c), (a,c)]
- **Adjacency List –** {a:[b,c], b:[a]}
- **Adjacency Matrix –** 1s where connected, 0s where not

There are **undirected** and **directed graphs** but also **unweighted** and **weighted** graphs. We would use weight instead of 1 for AdjMatrix
Bipartite Graph – a graph whos nodes can be divided into two disjoint sets (U, V) such that every link connects a node in U to one in V
Most real world networks are SPARSE

Node Degree – the number of edges adjacent to node i

- **In-degree –** the number of nodes pointing to i
- **Out-degree –** the number of nodes I points to
- **Degree –** sum of in-and-out degrees

Degree Distribution – Probability that a randomly chosen node has degree k
Clustering Coefficient – probability that a random pair of friends are connected – $C_i = (e_i) / (k * k-1)$ e=edges between neighbors, k=degree of node. Undirected counts as 2.
Path is a sequence of nodes in which each node is linked to the next one
Distance - between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
Diameter: the maximum distance between any pair of nodes in a graph

---

Erdos-Renyi Random Graph Model – $G_{n,p}$ – a undirected graph on n nodes and each edge (u,v) appears i.i.d. w/probability p

- Degree distribution is **binomial**
- Clustering coefficient is very small size

Network Structure

Triadic Closure - If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future
Triadic Closure means high **Clustering Coefficient**
Strong Triadic Closure Property – two strong ties imply a third edge

---

Span – the span of an edge is the distance of the edge endpoints if the edge is deleted
Bridge Edge – if removed, disconnects graph (span = inf)
Local Bridge – edge of span > 2 (any edge that doesn't close a triangle)
Weak ties have access to different parts of the network! Access to other sources and other kinds of information
Strong ties have redundant information
Edge Overlap – the number of shared neighbors divided by the union of neighbors

Community Detection – assembling nodes into logical groups based on common characteristics:

- Start with every node in the same cluster and break apart at "weak links" ("**divisive** clustering")
- Start with every node in its own "community" and join communities that are close together ("**agglomerative** clustering")

The **betweenness** of an edge is how many (fractional) shortest paths travel through it
Use **Girvan-Neuman** community detection algorithm to find hierarchical decompositions of networks

Signed Networks and Phenomena

Triads:

- Structural balance (stability) applies:
  - +++ = all friends
  - + - - = enemy of friend is my enemy
- Weak structural balance – allow mutual enemies (- - -)
- Incomplete graphs
  - Local view: Balance-able (if you can fill in slots to balance)
  - Global view: divide the graphs into two coalitions

Graph is **balanced** if and only if it contains **no cycle with an odd number of negative edges**

Homophily – birds of a feather flock together. Refers to the tendency for people to have (non-negative) ties with people who are similar to themselves in socially significant ways

Six Degrees of Separation and Network Searching

Average Path Length for real networks **are** like random graphs
Watts-Strogatz Small World Model – start with low-dimensional lattice, introduce randomness (shortcuts), add/remove edges to remote parts of lattice with probability p.

Regular Network – high clustering, high diameter
Small-World Network – high clustering, low diameter
Random Network – low clustering, low diameter
Intuition: It takes a lot of randomness to ruin the clustering, but a very small amount to create shortcuts

Decentralized Search – node only knows location of its friends and the target t, but doesn't know any other links

- nodes will act greedily with respect to geography: always pass the message to their neighbour who is geographically closest to t
- **Search Time –** number of steps taken to reach T
  - **Searchable – Search time is in $O(\log n)^B$**
    - **Kleinberg's Model –** $O(\log n^2)$ - searchable
  - **Not Searchable – Search time is in $O(n^a)$**
    - **Watts-Strogatz:** $O(n^{2/3})$ – not searchable

Kleinberg's model - nodes know their neighbors, each node has one random long-range link (following geography)

Power Laws, Inequality, and Unpredictability

Degree distributions are **not Gaussian – they are Heavy Tailed** (most volume at the tail end, right side)
Power law: p(x) varies with x^{-a}
Network Resilience – how does a networks connectivity change as nodes get removed?

- Random failures. **Real networks** are more resilient
- Targeted attacks (e.g. lowest degree). $\mathbf{G_{np}}$ is more resilient

Power Laws can arise from the **rich getting richer –** from the feedback introduced by correlated events

PageRank and Node Centrality

Hubs: pages that are "lists" of links that link to good stuff

Hub Update Rule: For each page p, update hub(p) to be the sum of the authority scores of all pages that it points to

Authorities: pages that are good, authoritative… and linked to by good hubs

Authority Update Rule: For each page p, update auth(p) to be the sum of the hub scores of all pages that point to it

Hub-Authority Update:

- Initialize all scores to 1
- Apply Authority Update rule
- Apply Hub Update Rule
- Normalize

PageRank – model that ranks page as important if it has more links

- Each link's vote is proportional to importance of its source page

PageRank Algorithm:

1. Initialize all nodes with 1/n PageRank
2. Perform k PageRank updates:

Basic PageRank Update Rule: Each page divides its current PageRank equally across its outgoing links. New PageRank is the sum of PR you receive.
PageRank Issue – circuits can cause PageRank to pool. We can fix this using **Scaled pagerank –** only divide a fraction s of PR among outgoing links, rest gets spread evenly over all nodes. Usually s is [0.8-0.9]
Random restarts – jumps to random node with probability 1-s (scaled pagerank)

Game Theory

		Your Partner	
		Presentation	*Exam*
You	*Presentation*	90, 90	86, 92
	Exam	92, 86	88, 88

Networks – Interconnected Structure
Game Theory – interconnected behavior

Player – the people that are involved in the scenario
Strategies – choices that can be made
Payoff – result/win/loss as a function of everyone's strategies
Payoff Matrix – matrix summarizing the payoffs of individual player strategies (see above)
A game **G** is a tuple – **(P, S, O)** set of players, set of strategies for players, and for every outcome, a payoff for each player
Rationality – every player wants to maximize payoffs and succeeds in doing so

Strictly Dominant Strategy – a strategy that is better than all other options regardless of what other players do.
Best Response – if other player plays T, then the best thing I can do is play S
Strict Best Response – if the best response is BETTER (not better or equal to) than all other responses to strategy T
A **dominant strategy** for P1 is a strategy that is a **best** response every strategy by P2
A **strict dominant strategy** for P1 is a strategy that is a strict best response every strategy by P2
Dominant strategies don't always exist!
Nash Equilibrium - Even when there are no dominant strategies, we should expect players to use strategies that are best responses to each other
Coordination game - all the players care about is playing the same strategy
Multiple Equilibria – what happens when there are multiple equilibria? Focal points – social norms, etc. help decide
Anti-coordination games – battle of the sexes. Unclear what will happen. (e.g. payoff is 1,5 and 5,1, but both are likely?)
Mixed Strategies – corresponds to a choice of mixture probabilities between 'pure' strategies

- Every game has a mixed-strategy Nash equilibrium
  Dominant strategy? **Sometimes**.
  Pure Nash Equilibria? **Sometimes**.
  Mixed Equilibria? **Always** exists

Game Theory Applications and Network Associations

Congestion games – different paths, with variable and constant times for drivers. This is actually multiple equilibria because N number of drivers (N=2000 for example) can all be different individual drivers
Braess' paradox is the observation that adding one or more roads to a road network can end up impeding overall traffic flow through it
Price of Anarchy – the ratio between socially optimal and selfish routing

Game Theory model of Cascades
Homophily impedes diffusion
The cascade capacity of a graph G is the largest q for which some finite set S can cause a cascade

Herding – decision to be made is impacted by the choices of those who acted earlier
Cascades can be **wrong**
Cascades can be based on **very little information**
Cascades are **fragile**
Virality – person-to-person transmission, deep branching structures, infecting minds
Measuring virality;

- Depth of Cascade (susceptible to super long chain)
- Average depth of cascade (susceptible to long chain then big broadcast)
- **Average path Length between nodes –** the best way to measure virality

Contagion & Epidemics

Types of epidemic diffusions:

- **Explosive spread**
- **Slow burn**
- **Cyclical**

Modelling epidemic spread:

- First person infected, infects each of k neighbors with independent probability p. Each infected then infect k neighbors… onwards
- **Blow up –** with high contagion probability, infection spreads widely
- **Die out –** with low contagion probability, infection dies out quickly

Basic Reproductive Number R_0 – number of expected new cases caused by an individual - $R_0 = p_k$

- If $R_0 < 1$ then with probability 1 the disease dies out after finite number of steps
- If $R_0 > 1$ then with probability > 0 the disease persists by infecting at least one person each wave

Quarantine – reduce k
Improved Sanitation – reduce p

SIR epidemic model –
S – Susceptible
I – Infectious, node is infected and infects w/probability p
R – removed and no longer infects or is infectious
Percolation model – judge if each edge is infectious or not by flipping a coin
SIS model – no removed state, can keep being re-infected. can run for a very long time, cycling through targets

Simple Diffusion – become infected when someone in network is infected. Faster on small world models but slower on large world
Complex diffusion – become infected when multiple in-network infections occur. Doesn’t occur in small but slow on large world models

- Weak ties are extremely useful for simple diffusion and contagion, but they inhibit complex diffusion

Voting

Preference Relation – ranks choices in terms of preference (e.g. X>Y>Z)

- Completeness – all pairs of distinct alternatives must be ranked
- Transitive – if X > Y and Y > Z then it must follow that X > Z

Majority Rule Voting Algorithm – whoever is preferred by majority of voters wins
Condorcet Paradox – majority rule with at least three alternatives can produce a non-transitive group ranking

Borda Count – 0 for last place, 1 for 2nd last… to k-1 for being picked first (e.g. NBA MVP voting)

- Borda count always produces a complete, transitive ranking
- Gives rise to **Irrelevant Alternatives** that may influence actual ranking. What voters think of irrelevant alternative should be irrelevant to how they feel about relative ranking of other alternatives, but it isn't

1. Unanimity – there needs to be a choice
2. Independence of Irrelevant alternatives (IIA) – ordering of X and Y should only depend on X and Y, nothing else
3. Non-dictatorship (should not be what only one party thinks)

Single-Peaked preferences – voter has a distinct choice in which alternatives fall off on either side X_{s-1} and X_{s+1+} of choice X_s

- If all individual rankings are single peaked, then majority rule can be applied to all pairs of alternatives and is complete & transitive

Condorcet Jury Theorem – as the number of voters increase, the probability of choosing correct decision goes to 1