12: Social Networks

Machine Learning and Real-world Data (MLRD)

Ann Copestake
(based on slides created by Simone Teufel)

Lent 2019

Where have we got to?

- You have now encountered two applications of ML and real-world data:
  - Sentiment Detection
  - Sequence learning for biological applications
- We will now move to the third topic: Social Networks.
- You will be given a network consisting of users and links between them.
- You will visualise this network and then write code to determine some simple statistics of the network.
- In subsequent sessions, we will use network properties in a classic ML task: **clustering**.

Overview of tasks

- Task 10: simple statistics including the **degree** of each node and the **diameter** of the network.
- Task 11: finding **gatekeeper** nodes via **betweenness centrality**.
- Task 12: using betweenness centrality of edges to split graph into **cliques** (example of clustering).
- Reading for social networks (all sessions):
  - Easley and Kleinberg for background: Chapters 1, 2, 3 and first part of Chapter 20.
  - Brandes algorithm: two papers by Brandes (links in practical notes).

Social networks

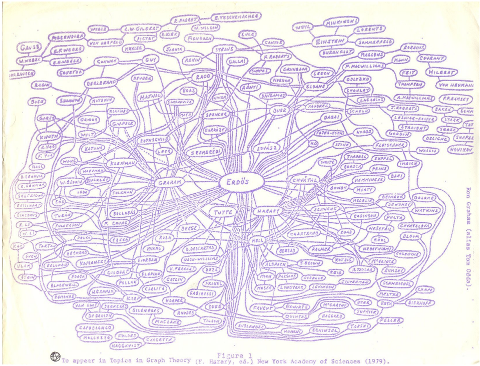
- Examples:
  - Facebook-style networks:
    - Nodes: people;
    - Links: “friend”, messages
  - Twitter-style networks:
    - Nodes: Entities/people
    - Links: “follows”, “retweets”
    - Also: research citations
- Operations on such networks
  - Which role does a node play in this network?
  - Is there a substructure in the network?
    neighbourhood areas/cliques

Some reasons to analyse social networks

- Academic investigation of human behaviour (sociology, economics etc)
- Disease transmission in epidemics.
- Modelling information flow: e.g., who sees a picture, reads a piece of news (or fake news).
- Identifying links between subcommunities, well-connected individuals:
  - recommending research papers to beginning PhD students
  - targeted advertising ...
- Lots of applications in conjunction with other approaches: e.g., sentiment analysis of tweets plus network analysis.

Erdös Number

Steps in a path between a researcher and the mathematician, Paul Erdös, counting co-authorship of papers as links.



http://oakland.edu/enp/

Network concepts illustrated by co-authorship

- Networks are modelled as graphs: undirected and unweighted here. (weight and direction for co-authorship?)
- The **degree** of a node is the number of neighbours a node has in the graph (here co-authors).
- Erdös is represented as a node of degree 509.
- The distribution of node degrees may be very skewed: American Mathematical Society data from 2004: `http://oakland.edu/enp/`
  - mean degree is 3.36
  - about 20% have degree 0 (i.e., no co-authored papers)
  - only five mathematicians had more than 200 collaborators, none beside Erdös had more than 270
- Most scientific fields can be linked to Erdös, mostly via small number of interdisciplinary authors.

Diameter and average distance of a network

- **distance** is the length of shortest path between two nodes.
- **diameter** of a graph: maximum distance between any pair of nodes.
- **small world phenomenon**, **six degrees of separation**
  - 'Chain-links': short story by Karinthy (1929): any two individuals in the world could be connected via at most 5 personal acquaintances.
  - Milgram attempted to verify experimentally (partial success).
- Natural networks tend to have closely clustered regions, connected only by a few links between them. Often these are **weak ties**.

Some important concepts for social networks

See Easley and Kleinberg (2010, Chapter 3) for full discussion:

- **giant component**: a connected component containing most of the nodes in a graph.
- **weak** and **strong ties**: the closeness of the link. e.g., two researchers co-author lots of papers together, or co-authors on one paper (with other people)? Large components often only connected by weak ties.

Bridge

bridge: an edge that connects two components which would otherwise be unconnected.

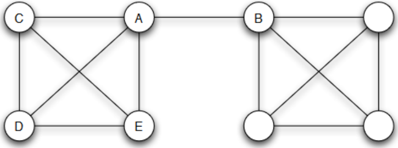
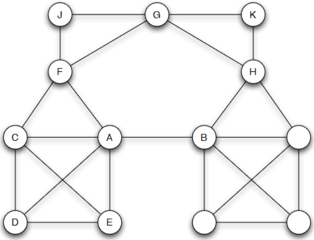


Figure 3.3 from Easley and Kleinberg (2010)

Local bridge

local bridge: an edge joining two nodes that have no other neighbours in common.



Cutting a local bridge increases the length of the shortest path between the nodes.

Figure 3.4 from Easley and Kleinberg (2010)

Triadic closure and clustering coefficient

Easley and Kleinberg (2010, p48–50)

- **triadic closure**: triangle of nodes. Thought of as a dynamic property: if A knows B and A knows C, relatively likely B and C will (get to) know each other.
- The **clustering coefficient** is a measure of the amount of triadic closure in a network.
- Clustering coefficient of a node A is the probability that two randomly selected neighbours of A are also neighbours of each other.

Small worlds (Easley and Kleinberg, Ch 20)

Is the small world phenomenon surprising?

- Short paths not surprising if links fan out enough at each step — we quickly reach everyone in the world.
- But, for small world, links must be found (to an extent) by humans: decentralized search.
- And fan out isn't what really happens: **triadic closure**.
- In fact, long weak ties are crucial.

Random graph generation for experimental investigation

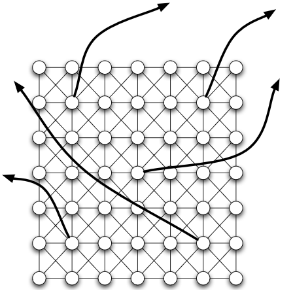


Figure 20.3 from Easley and Kleinberg (2010)

- Watts and Strogatz: randomly generated graph with triangles at close range, plus a few long random links.
- Random links generated according to inverse square of distance between nodes.
- These allow the reduction of distance to target.

Today's data

- Facebook data: combined friends list data from 10 users (**ego-networks**).
- Originally used for experiments in discovering **social circles**: e.g., family, school friends, university friends, CS department friends (contained completely in university friends).
- `http://snap.stanford.edu/data/egonets-Facebook.html`
- Also available today for the starred tick: two collaboration networks (also from SNAP).

Your task today

Task 10:

- Investigate the network using `Gephi`
  - Visualize the network with `Gephi`
  - Find network diameter
  - Visualize node degrees
  - Visualize betweenness centrality (discussed in a later lecture)
- Coding:
  - Find the degree of each node.
  - Determine the diameter of the network using a breadth-first all-pairs shortest path (APSP) algorithm.
    (More complex approaches in Algorithms course, but note there are no weights or negative edges here.)