

12: Social Networks

Machine Learning and Real-world Data (MLRD)

Paula Buttery (slides by Simone Teufel)

Where have we got to?

- You have now encountered two applications of ML and real-world data:
 - Sentiment classification
 - 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](#).

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

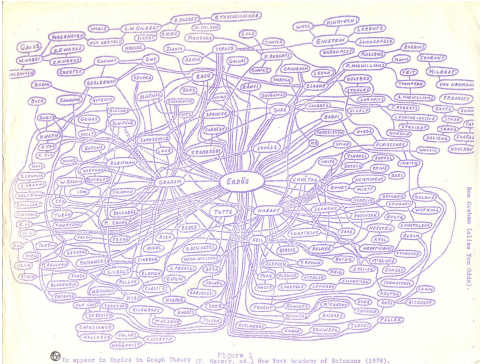
- Academic investigation of human behaviour (sociology, economics etc)
- Disease transmission in epidemics.
- Modelling information flow: e.g., who sees a picture, reads a piece of news (or fake news).
- Identifying links between subcommunities, well-connected individuals:
 - novelty
 - targeted advertising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

Diameter and average distance of a network

- Networks are modelled as graphs: undirected and unweighted here.
- **distance** is the length of shortest path between two nodes.
- **diameter** of a graph: maximum distance between any pair of nodes.

Erdős Number

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



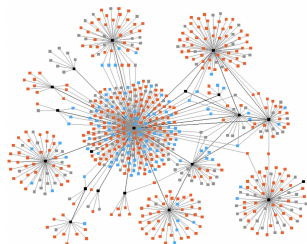
<http://oakland.edu/enp/>

Degree of a node

- The **degree** of a node is the number of neighbours a node has in the graph.
- For instance: Erdős had 509 coauthors, so is represented as a node of degree 509.
- The distribution of node degrees may be very skewed.
e.g.:
 - 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 Erdős had more than 270

Small world Phenomenon

- Natural networks often have the small world property.
- A small world network is one in which most nodes are not neighbours of each other, but the neighbours of any given node are likely to be neighbours of each other
- In other words, they have closely clustered regions, connected only by a few links between them.
- Most nodes can be reached from each other by a small number of hops.



Six Degrees of Separation

- But how many hops?
 - 'Chain-links': short story by Karinthy (1929): any two individuals in the world could be connected via at most 5 personal acquaintances.
 - Milgram (1967) attempted to verify experimentally (partial success)

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 ties**: socially distant ties, infrequent interaction (acquaintances) vs. **strong links**: close friends and family.
- The links that keep giant components together are often only weak ties.
- **bridge**: an edge that connects two components which would otherwise be unconnected.
- **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.

Triadic closure and clustering coefficient

Easley and Kleinberg (2010, p48–50)

- **triadic closure**: if A knows B and A knows C, relatively likely B and C will (get to) know each other.
- The **global clustering coefficient** is a measure of the amount of triadic closure in a network.
- It is the number of closed triads over the total number of triads (both open and closed).
- Clustering coefficient of a node A is the probability that two randomly selected neighbours of A are also neighbours of each other.

Random networks (Easley and Kleinberg, Ch 20)

- Is the small world phenomenon surprising?
- Very, if we assume random links overall.
- Not, if we create clusters with high **triadic closure** and long weak ties.
- Watts and Strogatz: randomly generated graph with high triadic closure at close range, plus a few long random links.
- Small world: also non-obvious that the links can be found (to an extent) by humans: decentralized search.
- Random links generated according to inverse square of distance between nodes.
- Allow 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.)