Machine Learning with Graphs: Network Science 1/2 - Real Networks, Network Properties

Arlei Silva

Spring 2022

Why would someone want to represent a real-world system as a network? And what are the basic principles that make networks look as they do? Network science provides a common language that researchers from a diverse set of disciplines can use to pose and answer questions regarding the patterns arising from real networks and how these patterns affect the behavior of complex systems.

A Science of Networks

The terms *network* and graph are often used interchangeably. But here we will see networks as a simplified representation (or model) of a complex system. *network science* is focused on the study of patterns that emerge from real networks. It was born from the idea that networks arising from completely different applications surprisingly share common patterns. As we will see, these patterns can be conveniently described in the language of graphs. The question of whether network science is an actual science is sometimes a contentious topic. But the term has helped to bring together researchers from various disciplines such as mathematics, physics, biology, computer science, and sociology.

Real networks

We will start with an overview of the different types of real networks with representative examples.

In *social networks*, nodes/vertices represent people and edges capture social interactions (e.g. friendship, collaboration, communication).

The so-called *small-world experiment*, led by the psychologist Stanley Milgram in the '60s, is considered one of the earliest work on social networks. The goal of the experiment was to measure the average path length between a pair of people in the social network formed by acquaintances in the US. The experiment was conducted as follows, packets were mailed to random people in Omaha, Nebraska, and Wichita, Kansas. Each packet was assigned to a random target recipient in Boston, Massachusetts. Participants were asked to send the package to the target recipient if they personally knew the person, or to forward it to someone who could potentially know the target recipient. The average path found among the 64/296 packets that reached their target was close to six.

These are other examples of social networks:

- Friendship network: nodes and edges represent people and friendship relationships, respectively. This is the case for most online social networks, such as Facebook. Examples of problems in friendship networks include community detection and friend recommendation.

- Collaboration network: edges represent collaboration towards a common goal, such as writing a paper or software. Team formation is a key problem in collaboration networks.

- Opinion networks: edge have signs indicating whether nodes agree (positive) or disagree (negative) with each other on a given subject. Many problems related to opinion networks focused on predicting and controlling their dynamics.

- Other examples: contact and affiliation networks.

Technological networks are models for man-made systems designed with a specific goal, such as communication, transportation, etc.

The study of graphs was motivated by a technological network. In 1736, the mathematician Leonhard Euler was asked whether there was a route in the city of Konigsberg (now Kaliningrad, Russia) that would visit each of its seven bridges exactly once (i.e. an Eulerian path). By representing each island in the city as a node and bridges as an edge, he showed that such a route was possible iff each node that was not a start or end-point had an even degree.

Other examples of technological networks include:

- The internet: nodes are routers and edges are data connections between them. Measuring communication efficiency and robustness to failures are some of the problems studied using the Internet graph.

- Transportation network: nodes and edges represent locations and direct paths between them. Examples of problems in transportation networks include route recommendation and traffic forecasting.

- Power network: each node is a bus (generator, load, or substation) and an edge represents a power line. Power networks are useful for evaluating the robustness of power systems to cascading failures.

- Other examples: telephone, airport, sea transportation, gas, and water distribution networks.

Networks generated based on biological systems or processes are called *biological networks.* This is a diverse set of networks capturing interactions among animals, proteins, neurons, among others.

Today's artificial neural networks are inspired by biological neural networks, which have been studied for more than a century. In 1906, Santiago Ramon y Cajal and Camillo Colgi were awarded the Nobel Prize in Medicine "in recognition of their work on the structure of the nervous system". Artificial neural networks started being studied only in 1943 by Warren McCulloch and Walter Pitts. Still, all the evidence available indicates that artificial networks are very simplistic approximations of biological neural networks.

- Brain network: Nodes are neurons and edges are synapses. The technology to extract such networks (a.k.a the connectome) from brain scans is still quite limited but brain networks might enable the discovery of new treatments for brain-related diseases, such as Alzheimer's.
- Food web: Nodes are species and directed edges are predator-prey interactions. Food webs can be used to estimate the overall stability and to control the evolution of an ecosystem.
- Protein-protein interaction (PPI) network: Nodes represent proteins and edges are chemical or physical interactions. The identification of protein complexes and the prediction of function and interactions are some of the problems studied in the context of PPI networks.
- Other examples: metabolic, gene regulatory, drug interaction, brain functional connectivity, host-parasite, and mutualistic networks.

Information networks represent content/data and different types of interactions between them (e.g. citation, links, etc.).

The use of networks to represent knowledge dates back to early work in AI and expert systems. In 1989, Tim Berners-Lee started a project to manage information at CERN that was the precursor of the World Wide Web. The Web quickly became the world/ largest repository, motivating the design of mechanisms to organize and access its information. Many of these solutions were inspired by earlier work in information sciences and bibliometrics. The most successful example is *PageRank*, which applies the Web graph to measure the importance of pages.

- Citation network: Nodes are articles and directed edges are citations. Citation networks can be used for search, recommendation, and categorization.
- Knowledge graphs: Entities (e.g. objects, events, concepts) are represented as nodes, and semantic relationships (e.g. is a friend of, starred in, was born in) are represented as edges. Search, question-answering, and recommendation are some of the relevant problems in the context of knowledge graphs.
- Other examples: Peer-to-peer (P2P) and recommender networks.

Network Data Repositories

Network science relies on the availability of real network data. This section provides some examples of publicly available network data repositories.

General repositories:

- Stanford Network Analysis Project:
  `http://snap.stanford.edu/data/index.html`
- Network Repository:
  `https://networkrepository.com/index.php`
- UCI Network Data Repository:
  `https://networkdata.ics.uci.edu/resources.php`
- Mark Newman's datasets:
  `http://www-personal.umich.edu/~mejn/netdata/`

Biological networks:

- STRING:
  `https://string-db.org`
- BIOGRID:
  `https://thebiogrid.org`
- Human Connectome project:
  `http://www.humanconnectomeproject.org`
- The Food Web Database:
  `https://www.globalwebdb.com`
- MetaCyc:
  `https://metacyc.org`
- KEGG:
  `https://www.genome.jp/kegg/pathway.html`
- RegNetwork:
  `http://www.regnetworkweb.org`
- GRNdb:
  `http://www.grndb.com`

Technological networks:

- OpenStreetMap:
  `https://www.openstreetmap.org/`
- SciGRID:
  `http://scigrid.de`

- PyPSA-Eur:
  `https://github.com/pypsa/pypsa-eur`
- IEEE test cases:
  `https://cmte.ieee.org/pes-testfeeders/resources/`
- Water Distribution System Operations:
  `http://www.uky.edu/WDST/database.html`

Social networks:

- DBLP:
  `https://dblp.org`
- SocioPatterns:
  `http://www.sociopatterns.org/datasets/`

Information networks:

- Wikipedia:
  `https://dumps.wikimedia.org`
- Wikidata:
  `https://www.wikidata.org/wiki/Wikidata:Main\_Page`
- The GDELT Project:
  `https://www.gdeltproject.org`
- Google Knowledge Graph:
  `https://developers.google.com/knowledge-graph/`

Network Properties

Network science looks for relationships between network structure and other properties relevant to the corresponding system of interest. Here, we will present some of the key structural properties of networks. As mentioned earlier, the fact that networks from a diverse set of applications share similar structural properties is at the core of network science.

The simplest way to describe the structure of a network is *drawing* it. However, drawing networks with more than a few thousand nodes is a challenge. Modern drawing tools apply sophisticated algorithms in order to optimize one or more notions of drawing quality. Examples of quality metrics include the number of *edge crossings* and the *total length of edges*. For large graphs, one alternative is to apply graph compression/summarization. Instead, the most common approach is to compute a few statistics that will be described in the remaining of this section.

The *degree* of a vertex is its number of neighbors. As discussed earlier, the degree can also be generalized to directed and weighted graphs. The degree is an example of a *centrality*—i.e. node importance—measure. The *degree sequence* gives the degree of each vertex. Oftentimes the degree distribution or some of its moments (mean, standard deviation, etc.) is a more compact description of the degree information. As we will see in more detail, the most appropriate way to describe the distribution depends on our assumptions regarding the distribution itself.

As shown by the small-world experiment, the length of paths in the graph also provides important information about its structure. While the experiment focused on the *average shortest path*, the *diameter*—i.e. longest shortest path—is also often applied. Path-based statistics are particularly useful for measuring the efficiency of communication in a network. However, notice that computing all-pairs shortest paths takes $O(n^2 \log n + nm)$ time for sparse networks with n vertices, m edges, and $m < n^2$.[1] Thus, more efficient approximation algorithms are needed for computing path-based statistics in large graphs.

Another important property of networks is how well they can be divided into groups of nodes (a.k.a. clusters or communities). There is vast literature on network clustering metrics and algorithms. The *clustering coefficient* is the simplest of such metrics and can be defined as follows:

$$C = \frac{3 \times \# \text{ of triangles}}{\# \text{ of connected triples}}$$

where a triplet is a set of three nodes and a triangle is a clique of size 3.

References

[1] Albert-László Barabási et al. *Network Science.* Cambridge University Press, 2016.

[2] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms.* MIT press, 2009.

[3] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[4] Mark Newman. *Networks.* Oxford university press, 2018.

[1] Using the Johnson's algorithm [2].